ÚLOHA ŠTÁTNEHO PROGRAMU VÝSKUMU A VÝVOJA
2003SP200280307

ZÁVEREČNÁ SPRÁVA

o riešení úlohy výskumu a vývoja

Komplexné spracovanie slovenského jazyka
a jeho elektronizácia na účely jazykovedného výskumu
tematického štátneho programu výskumu a vývoja
Aktuálne otázky rozvoja spoločnosti

1. Úvod

Úloha štátneho programu výskumu a vývoja ***Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu*** sa riešila v Jazykovednom ústave Ľ. Štúra Slovenskej akadémie vied v Bratislave od 1. 7. 2003. Pôvodný termín dokončenia úlohy 31. 12. 2005 bol na základe dodatku č. 4/2005 k zmluve o dielo predĺžený o rok, t. j. do 31. 12. 2006. Ekonomicky oprávnené náklady na riešenie úlohy ostali nezmenené – na dofinancovanie úlohy v r. 2006 sa presunula časť prostriedkov naakumulovaných z predchádzajúcich rokov.

Riešenie úlohy sa realizovalo v špecializovanom oddelení Slovenského národného korpusu JÚĽŠ SAV skladajúcom sa z programátorskej, lingvistickej a technicko-korektorskej zložky. Všetky tri zložky participovali na konkrétnych prácach vo vzájomnej súčinnosti. Keďže obsah úlohy výrazne prekračoval kapacitné možnosti ôsmich stálych zamestnancov oddelenia, na rutinných prácach pri technickom spracúvaní textov (skenovanie, rozpoznávanie a rekonštrukcia textov po skenovaní), ako aj na špecifických lingvistických prácach (anotovanie vybraných textov) sa podieľali externí spolupracovníci, prevažne študenti a absolventi odboru slovenský jazyk.

2. Ciele a harmonogram riešenia úlohy

Zmluvne dohodnutým výsledkom riešenia úlohy za celé obdobie trvania zmluvy je lingvisticky anotovaný 200-miliónový reprezentatívny korpus textov súčasného slovenského jazyka prístup-

ný verejnosti na internete a koncepcia a čiastkové riešenia elektronizácie jazykovedného výskumu na Slovensku.

K stanovenému cieľu je zmluvne požadovaný obsah riešenia:

1. Sociolingvistická analýza štýlovo-žánrovej distribúcie textov súčasného slovenského jazyka, adekvátna stratifikácia textov v Slovenskom národnom korpuse.
2. Získavanie textov na základe Zmluvy o inom použití diela.
3. Technické spracúvanie textov: prevod do jednotného formátu, selekcia grafiky, skenovanie a optické rozpoznávanie, tokenizácia, lematizácia, rozhranie na vyhľadávanie lingvistických informácií, testovanie existujúceho (zahraničného) softvéru, postupný vývoj vlastného softvéru.
4. Koncepčná príprava a realizácia lingvistického anotovania textov: morfologická a syntaktická anotácia.
5. Príprava terminologickej databázy, databázy lexikografických diel, paralelného korpusu, diachrónneho korpusu, korpusu nárečových textov, korpusu hovorených prejavov.

Plnenie celej úlohy sa členilo na etapy ohraničené kalendárnym rokom, resp. plnením čiastkových cieľov podľa nasledujúceho harmonogramu (modifikovaného dodatkom č. 4/2005 k zmluve o dielo):

10/2003 – 12/2003

Dotvorenie pracovného kolektívu, dobudovanie technického vybavenia a pracovných priestorov. Koncepcia komplexného počítačového spracovania vybraných textov súčasného slovenského jazyka a postupnej elektronizácie jazykovedného výskumu na Slovensku. Začatie zberu dát, ich štruktúrovanie na základe relevantnosti jednotlivých druhov a žánrov, vonkajšia anotácia textov. Sprístupnenie časti textov na internete.

01/2004 – 12/2004

Vývoj a testovanie vlastných programových nástrojov (lematizátor, morfologický anotátor). Dopĺňanie textov a sprístupnenie ďalších 40 miliónov slov na internete. Budovanie terminologickej databázy primárne pre potreby pracovníkov JÚĽŠ SAV a príslušných externých spolupracovníkov. Postupná tvorba databázy lexikografických diel a jej sprístupňovanie na internete. Odborné štúdie.

01/2005 – 12/2005

Testovanie ďalších programových nástrojov (syntaktický parser), začiatok ručnej syntaktickej anotácie. Vývoj morfologického analyzátora a generátora tvarov. Dopĺňanie textov do všeobecného korpusu slovenského jazyka, príprava novej verzie korpusu. Začiatok budovania paralelných korpusov. Príprava medzinárodnej konferencie.

01/2006 – 12/2006

Vyvažovanie všeobecného korpusu slovenského jazyka podľa štýlov a žánrov, rozšírenie nevyváženého korpusu o ďalšie dáta. Tvorba podkorpusov podľa jednotlivých štýlov. Príprava verzie korpusu pre vydanie na CD/DVD. Dokončenie 1. verzie morfologického generátora a analyzátora tvarov slovenského jazyka v rozsahu slovnej zásoby Krátkeho slovníka slovenského jazyka. Príprava a zverejnenie 1. verzie Slovenského závislostného korpusu (syntakticky anotovaný korpus). Dokončenie 1. verzie slovensko-ruského a slovensko-českého paralelného korpusu.

V jednotlivých etapách sa dostávali do popredia ciele, ktoré bolo možné začať plniť ako nové. Východiskovým leitmotívom, nie vždy explicitne formulovaným, bolo neustále získavanie a technické spracúvanie textov.

3. Súhrnné hodnotenie plnenia cieľov podľa zmluvy

3.1.1. S prípravou na riešenie úlohy sa začalo s istým predstihom vzhľadom na to, že začiatok úlohy štátneho programu sa plánoval skôr, ako bola napokon podpísaná zmluva. Ku koncu kalendárneho **roka 2003** sa podarilo stabilizovať pracovný kolektív v rozsahu 8 riadnych pracovných miest, ktoré sa obsadzovali na základe konkurzu z prostredia mimo JÚĽŠ SAV, iba 1 miesto získala kmeňová pracovníčka ústavu. Už vtedy boli zrejmé viaceré skutočnosti, ktoré ovplyvňovali plynulosť riešenia jednotlivých cieľov počas celého obdobia riešenia úlohy: v porovnaní s podobnými pracoviskami v zahraničí ide o kolektív čo do počtu veľmi minimálny, zložený z absolventov iných smerov ako počítačová či korpusová lingvistika, ktoré sa na Slovensku nedajú študovať ako samostatný odbor na žiadnej vysokej škole. Tento kolektív má stanovený značný rozsah špecializovaných úloh a ich plnenie si vyžaduje kvalifikáciu, ktorá, najmä v programátorskej zložke, nie je dostatočne ohodnotená. Výsledkom je značná migrácia pracovníkov, čo sa nevyhnutne odzrkadľuje na spomalení či až utlmení niektorých prác. Na plynulé plnenie všetkých úloh a výraznejšie zapojenie projektu do medzinárodného kontextu bolo potrebné doplniť oddelenie aspoň o dvoch ďalších stálych pracovníkov (programátor a lingvista-terminológ), na čo sa upozorňovalo už v r. 2004. Nové pracovné miesta sa však nepodarilo získať a oddelenie má stále iba 8 riadnych zamestnancov.

Technické vybavenie pracoviska sa na konci roka 2003 dostalo na úroveň zodpovedajúcu nárastu dát a používateľov, najmä externých používateľov korpusu. Východiskových 7 pracovných staníc a 1 server sa rozšírilo o 3 pracovné stanice, diskové pole RAID a počítače parametrizované na prácu v clustrovom režime. Ako korpusový manažér sa použil Manatee s klientom Bonito zakúpený z Fakulty informatiky Masarykovej univerzity v Brne. Začalo sa s vývojom vlastného vyhľadávacieho systému a pilotným projektom paralelného korpusu po-

čítačových termínov. Hoci sa novovybudované pracovné priestory ukazovali ako kapacitne nedostačujúce, v r. 2003 sa ich z finančných dôvodov nepodarilo rozšíriť.

Koncepcia komplexného počítačového spracovania vybraných textov súčasného slovenského jazyka a postupnej elektronizácie jazykovedného výskumu na Slovensku bola vypracovaná, prednesená a prediskutovaná na pôde JÚĽŠ SAV i na medzinárodnej konferencii a zverejnená vo viacerých publikačných výstupoch (Šimková, 2003; Šimková, 2004; presná bibliografia sa nachádza v časti 8 tejto správy). Dostupná je aj na www stránke SNK (http://korpus.juls.savba.sk/publications/index.sk.html). Súčasťou koncepcie bola stratifikácia textov v Slovenskom národnom korpuse na základe sociolingvistickej analýzy štýlovo-žánrovej distribúcie textov súčasného slovenského jazyka opretej o analýzy spojené s prípravou viaczväzkového výkladového slovníka slovenského jazyka a podobné výskumy a skúsenosti Ústavu Českého národného korpusu. Začalo sa s prípravou vlastného tagsetu (súboru značiek a pravidiel) na ručnú morfologickú anotáciu (Forróová – Garabík – Gianitsová – Horák – Šimková, 2003; http://korpus.juls.savba.sk/publications/index.sk.html).

Zber a spracúvanie dát prebiehalo podľa plánu a v súlade s kapacitnými možnosťami pracoviska z hľadiska personálneho aj technického zabezpečenia. Na internete bolo koncom roka 2003 sprístupnených takmer 170 miliónov slov (porov. prehľad verzií SNK v 3.2.2), čo umožňovalo uspokojivú prácu lexikografom i bežným záujemcom o slovenský jazyk. Boli sprístupnené v podstate všetky v tom čase spracované dáta s kompletnou bibliografickou a štýlovo-žánrovou anotáciou, ale bez výberu a vyváženosti štýlov a žánrov. Na vytvorenie reprezentatívneho korpusu nebolo k dispozícii dostatočné množstvo textov z rôznych oblastí – najviac chýbala odborná literatúra z technických a prírodovedných oblastí.

Odborná náplň stanovených cieľov bola v r. 2003 splnená aj napriek oneskorenému oficiálnemu začiatku riešenia úlohy a personálnym zmenám v riešiteľskom kolektíve. Osobitným príspevkom programátorov oddelenia SNK bolo sprístupnenie elektronickej verzie najnovšieho vydania Krátkeho slovníka slovenského jazyka na internete a tvorba a sprístupnenie paralelného korpusu počítačových termínov na stránke SNK.

3.1.2. V **r. 2004** bol Slovenský národný korpus poskytnutý používateľom v ďalšej verzii (182 mil. tokenov; porov. prehľad v 3.2.2) so spresnenou tokenizáciou a štýlovo-žánrovou anotáciou, s úspešne realizovanou lematizáciou a interne testovanou morfologickou anotáciou uskutočnenou plne automatizovane na báze softvéru z Matematicko-fyzikálnej fakulty UK v Prahe. Súčasne sa prepracúval vlastný morfologický tagset, vyvíjali a zdokonaľovali sa nástroje na anotovanie a prebiehala manuálna morfologická anotácia – zanotovalo sa takmer 100 tisíc tokenov. V rámci spoločného projektu s MFF UK Praha sa testovala možnosť konverzie medzi

českým a slovenským tagsetom, na základe ktorej by bolo možné označkovať slovenské texty slovenskými značkami.

Slovenský národný korpus bol v r. 2004 prístupný na internete v rozsahu takmer trojnásobne väčšom oproti plánu. Dopĺňanie textov a ich sprístupňovanie na internete sa však oproti predchádzajúcej etape výrazne spomalilo aj vzhľadom na náročnosť získavania a spracúvania starších textov, ktoré jestvujú len v neelektronickej podobe. Výrazným handicapom oproti plánu bola extrémna nevyváženosť korpusu v prospech publicistiky (95 %) a nemožnosť vytvoriť dostatočne veľký vyvážený korpus, pretože chýbali texty najmä z odborných žánrov.

Koncepčne sa začalo pripravovať budovanie terminologickej databázy a vytvoril sa pilotný korpus právnych textov v rozsahu 500 tis. slov. Databáza lexikografických diel naďalej obsahovala iba Krátky slovník slovenského jazyka, pretože ostatné lexikografické diela sa redakčne upravovali, resp. sa nepodarilo zmluvne dohodnúť ich sprístupnenie na internete.

3.1.3. Vecné plnenie jednotlivých cieľov v **r. 2005** sa vďaka ústretovosti objednávateľa mohlo modifikovať jednak vzhľadom na jeho možnosti v prideľovaní finančných prostriedkov, jednak sa zreálnilo podľa aktuálnej situácie v dostupnosti spracovateľných textov, ako aj v dostupnosti a kvalite existujúcich softvérov: posilnilo sa budovanie základného, primárneho korpusu a podkorpusov paralelných textov miesto korpusu nárečových textov, pristúpilo sa k vývoju vlastného morfologického analyzátora a generátora tvarov miesto vývoja korpusového manažéra. Po polročných problémoch s financovaním sa napokon pokračovalo v súlade s plánom, ktorý bol po vzájomnej dohode medzi objednávateľom a dodávateľom riešenia úlohy modifikovaný na základe dodatku č. 4/2005 k zmluve o dielo, ktorým sa plnenie úlohy predĺžilo do 31. 12. 2006 bez nároku na dodatočné finančné zdroje.

Koncom r. 2005, po intenzívnom získavaní a spracúvaní textov, bola na www stránke SNK sprístupnená nová verzia korpusu prim-2.0 v rozsahu 250 miliónov textových jednotiek s lematizáciou a automatizovanou morfologickou anotáciou. Pokračovalo sa v ručnej morfologickej anotácii (130 tisíc zanotovaných a opravených textových jednotiek; porov. v 3.2.3), začali sa práce na morfologickom analyzátore a generátore tvarov (spracovali sa vzory a paradigmy menných slovných druhov v rozsahu Krátkeho slovníka slovenského jazyka), na ručnej syntaktickej anotácii (10 tisíc zanotovaných viet; porov. v 3.2.3), ako aj na budovaní paralelných korpusov. Do databázy lexikografických diel a ďalších lingvistických zdrojov pribudli tri zväzky Bibliografie slovenských jazykovedcov, kompletné ročníky časopisu Kultúra slova od r. 1994 a obsahy všetkých čísel od r. 1967 do r. 1993, ako aj pôvodný text diela Ľ. Štúra Nauka reči Slovenskej (porov. v 3.2.4).

Členovia riešiteľského kolektívu zorganizovali medzinárodnú konferenciu Slovko 2005 – Slovanské a východoeurópske jazyky v počítačovom spracovaní, z ktorej vydali zborník. Vý-

sledky práce boli prezentované na viacerých vedeckých podujatiach doma i v zahraničí, na materiáli korpusu sa riešili viaceré výskumné lingvistické i počítačové projekty (porov. 3.2.5) vrátane dvoch diplomových prác pod vedením M. Šimkovej.

Kolektívu Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra ako pracovníkom Slovenskej akadémie vied bola v r. 2005 udelená ***Cena SAV za budovanie infraštruktúry pre vedu***.

3.1.4. Začiatkom **r. 2006** bola sprístupnená nová verzia všeobecného, základného korpusu prim-2.1 v rozsahu 300 miliónov textových jednotiek s prvou verziou ručne morfologicky anotovaného podkorpusu (r-mak-1.0) a automatizovanou morfologickou anotáciou celého korpusu slovenskými značkami. Tak ako pri predchádzajúcich verziách sa realizovala aj automatizovaná anotácia a dezambiguácia pomocou českého softvéru a na báze českého tagsetu, ktorá ostáva k dispozícii na komparáciu. Koncom roku 2006 sa dokončila nová verzia korpusu prim-3.0 v rozsahu vyše 350 mil. textových jednotiek. Vyvážený korpus z poslednej verzie bude distribuovaný aj na CD/DVD nosičoch najmä na výučbové využitie.

V r. 2006 bolo oslovených a zazmluvnených 100 nových poskytovateľov textov (porov. 3.2.1), spolu ich je takmer 400. Vďaka rozsiahlemu skenovaniu a upravovaniu do elektronickej podoby tých textov, ktoré existujú iba v papierovej forme, pribudlo takmer 60 tisíc strán textov z obdobia r. 1955 až 1993. Pri čítaní a opravovaní skenovaných textov pomáhalo priemerne mesačne 30 študentov a iných externých spolupracovníkov. Veľký dôraz sa kládol na texty umeleckého a odborného zamerania, aby sa doplnili chýbajúce druhy textov a dosiahla lepšia vyváženosť korpusu. Celkové zväčšenie korpusu a významný nárast podielu nepublicistických textov umožnil vyhovieť požiadavkám špecifických výskumov a prác a od verzie prim-2.1 sa pripravujú samostatné podkorpusy publicistických, umeleckých a odborných textov, osobitne aj podkorpus originálnych slovenských umeleckých textov. Posledná verzia vyváženého korpusu v rozsahu takmer 200 mil. tokenov má distribúciu štýlov v pomere 33,330 % publicistické : 33,344 % odborné : 33,312 % umelecké : 0,014 % iné texty (porov. 3.2.2).

Systematicky a intenzívne pokračovala ručná morfologická anotácia vybraných textov korpusu a tvorba trénovacieho podkorpusu. Podkorpus r-mak-1.0, sprístupnený začiatkom roka 2006, obsahoval 322 600 textových jednotiek, v súčasnosti je k dispozícii nová verzia r-mak-2.0 v rozsahu 511 534 textových jednotiek. Na ručnej morfologickej anotácii sa podieľalo 6 študentiek a absolventiek odboru slovenský jazyk z FiF UK Bratislava. Pokračovalo sa aj s prípravou Slovenského závislostného korpusu – syntakticky anotovaného podkorpusu. Anotujú sa texty, ktoré prešli ručnou morfologickou anotáciou, každý text dvoma rôznymi anotátorkami. Na syntaktickej anotácii sa podieľalo priemerne 6 anotátoriek, zanotovaných je dohromady takmer 70 tisíc viet, z toho takmer 50 tisíc sa anotovalo v r. 2006. V nastupujúcej

fáze opráv sa zatiaľ urobili prvé technické opravy (korigovanie parsovania a nezanotovaných uzlov), ďalšie kroky budú vyžadovať väčšie zaangažovanie programátorskej zložky.

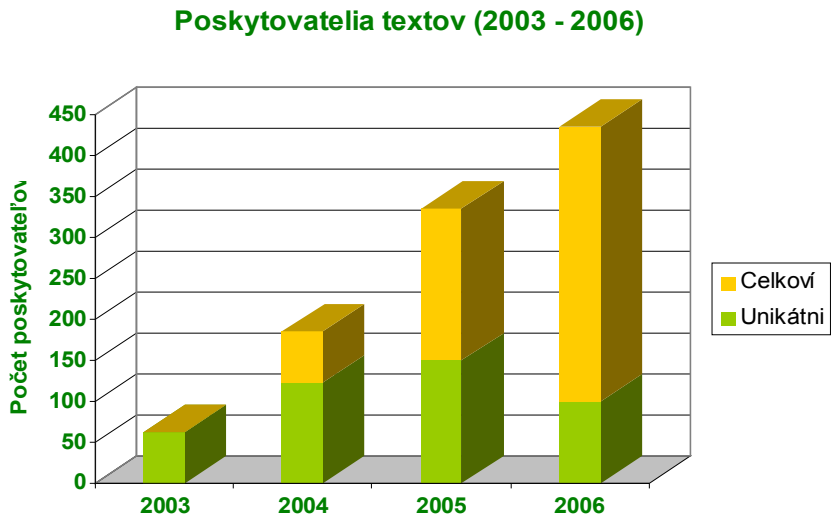
V r. 2005 sa začalo s vývojom vlastného morfologického analyzátora a generátora tvarov slovenského jazyka, do ktorého slovníka vzorov a paradigiem sa v priebehu roka 2006 dopracovali všetky slová z Krátkeho slovníka slovenského jazyka a doplnili sa najfrekventovanejšie slová z korpusu. Celkovo je spracovaných 56 269 slov s paradigmami v rámci 1 365 vzorov. Boli vypracované skripty na automatickú kontrolu doplňaných slov a vzorov, algoritmy na analýzu ľahko predikovateľného ohýbania (negácie, superlatívy) a pokračuje sa v tvorbe algoritmov pre číslovky (porov. príspevky R. Garabíka v druhej časti príloh).

V r. 2006 sa zintenzívnili práce aj na paralelných korpusoch, predovšetkým slovensko-ruskom, slovensko-francúzskom a slovensko-českom. Okrem zhromažďovania textov sa vypracovala infraštruktúra na efektívnu tvorbu paralelných korpusov a korpusový manažér pre paralelné korpusy. Koncom roka bol vytvorený paralelný rusko-slovenský korpus v rozsahu približne 900 000 slov (pre každý jazyk) a francúzsko-slovenský paralelný korpus v rozsahu vo francúzskej časti 315 599 tokenov, 13 004 viet a v slovenskej časti 194 478 tokenov, 12 286 viet. Nepomer medzi počtom tokenov a viet vo francúzskych a slovenských textoch je spôsobený pravdepodobne rozdielnou syntaktickou a morfologickou stavbou oboch jazykov (napr. vo francúzštine používanie členov pri substantívach, konjugácia slovies atď.). Obidva paralelné korpusy sú sprístupnené na internete.

Pracovisko SNK začalo s prípravou vlastnej Slovenskej terminologickej databázy zameranej na oblasť práva, ekonomiky a vybraných technických vied. V r. 2006 sa v rámci prvej fázy projektu vypracoval návrh terminologického záznamu a návrh softvérovej štruktúry databázy s cvičnými údajmi. Veľa práce sa vykonalo pri skenovaní, rekonštruovaní, spracovaní a sprístupňovaní lingvistických zdrojov – pribudli všetky súčasné časopisy JÚĽŠ SAV, vybrané monografie (napr. akademická Morfológia slovenského jazyka, 1966) a zborníky, ako aj 5 zväzkov zo šesťzväzkového Slovníka slovenského jazyka (1959 – 1968).

3.2.1. Poskytovatelia textov

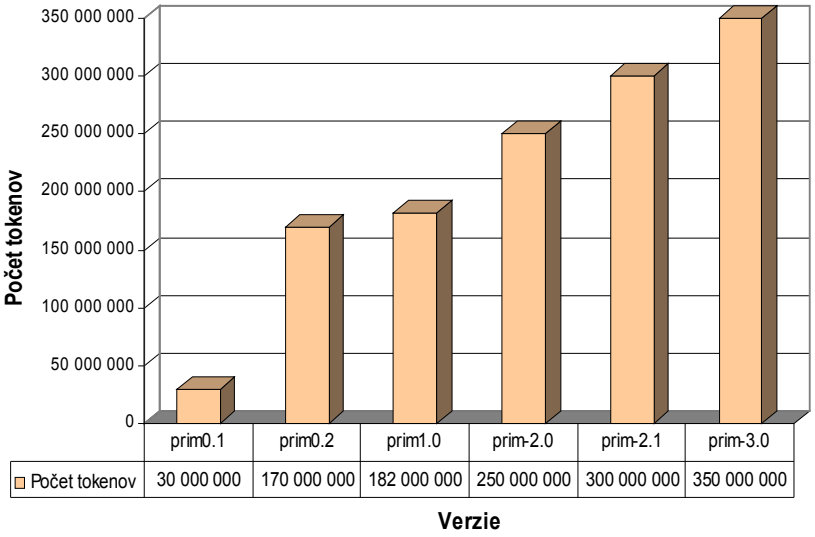
Grafické znázornenie počtu poskytovateľov textov, ktorí súhlasili so začlenením svojich textov do databázy SNK na základe Licenčnej zmluvy (elektronické verzie textov) alebo Vyhlásenia (na vyhotovenie technickej rozmnoženiny tlačeného textu), podľa jednotlivých rokov. Zoznam všetkých 435 poskytovateľov sa nachádza na stránke SNK (http://korpus.juls.savba.sk/contributors/index.sk.html), zoznam poskytovateľov textov zazmluvnených v r. 2006 je v prvej časti príloh k tejto správe.

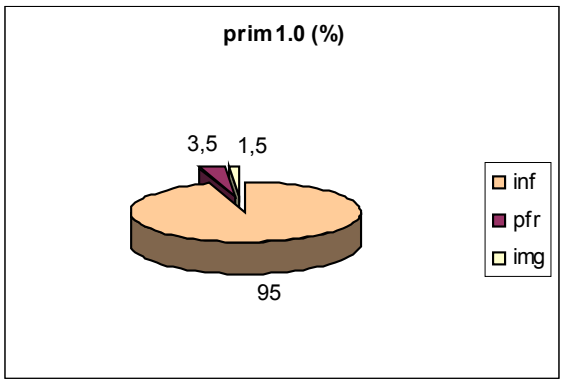


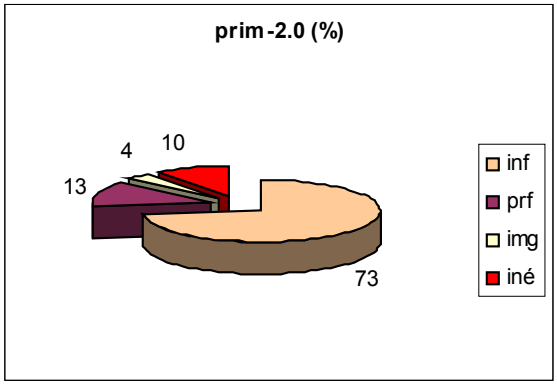
	2003	2004	2005	2006
Celkoví	0	62	185	335
Unikátni	62	123	150	100
Spolu	**62**	**185**	**335**	**435**

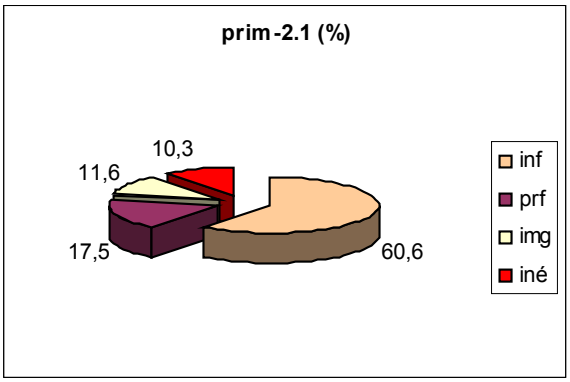
3.2.2. Verzie a podkorpusy Slovenského národného korpusu

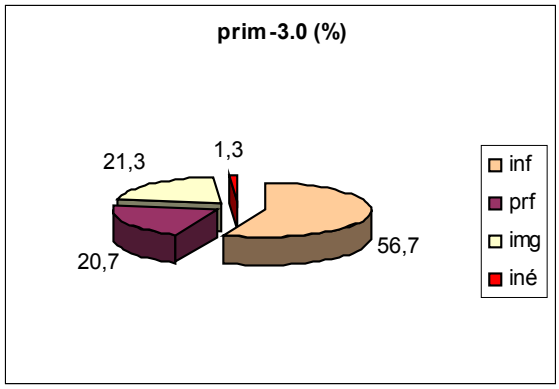
Grafické znázornenie veľkosti a štýlovej distribúcie korpusu podľa jednotlivých rokov.











Hlavný korpus je primárny, základný korpus (prim) obsahujúci všetky zmluvne získané a spracované písané texty od roku 1955 do súčasnosti. Verzia hlavného korpusu je nové „vydanie“ celého korpusu. Slovenský národný korpus sa sprístupňoval v novej verzii vždy po relevantnom prírastku textov a zmene alebo skvalitnení anotácie. Nová verzia obsahuje aj texty z predchádzajúcej verzie:

prim0.1 → sprístupnený v auguste 2003, 30 mil. tokenov
prim0.2 → sprístupnený koncom r. 2003, 170 mil. tokenov
prim1 → sprístupnený v r. 2004, 182 mil. tokenov
štýlovo extrémne nevyvážený (95 % inf : 3,5 % img : 1,5 % prf)
lematizovaný, automatizovane morfologicky anotovaný pomocou českého softvéru a na základe českého tagsetu
prim-2.0 → sprístupnený v r. 2005, 250 mil. tokenov
štýlovo vyváženejší (73 % inf : 13 % img : 4 % prf : 10 % iné)
lematizovaný, automatizovane morfologicky anotovaný pomocou českého softvéru a na základe českého tagsetu
prim-2.1 → sprístupnený začiatkom r. 2006, 300 mil. tokenov
štýlovo opäť vyváženejší (60,6 % inf : 17,5 % img : 11,6 % prf : 10,3 % iné)
lematizovaný, automatizovane morfologicky anotovaný pomocou nemeckého softvéru na základe slovenského tagsetu
prim-3.0 → sprístupnený začiatkom r. 2007, 350 mil. tokenov
štýlovo opäť vyváženejší (56,74 % inf : 20,69 % img : 21,32 % prf : 1,25 % iné)

lematizovaný, automatizovane morfologicky anotovaný pomocou nemeckého softvéru na základe slovenského tagsetu a morfologického analyzátora tvarov

Špecifickými časťami z hlavného, primárneho korpusu, tvorenými podľa udelenej licencie a štýlov, sú jednotlivé podkorpusy:

public → podkorpus textov verejne prístupný prostredníctvom siete internet na vedecko-výskumné, učebné a iné nekomerčné ciele na základe licenčnej zmluvy alebo vyhlásenia medzi poskytovateľom textu a Jazykovedným ústavom Ľ. Štúra SAV; tvorí sa pri každej verzii

juls → podkorpus textov prístupný prostredníctvom siete intranet len na pôde JÚĽŠ SAV obsahujúci navyše texty poskytnuté len na takéto interné použitie; tvorí sa pri každej verzii

snk → podkorpus textov prístupný prostredníctvom siete intranet len v rámci oddelenia SNK obsahujúci navyše texty poskytnuté len na úzke interné použitie; rozdiel oproti podkorpusu public predstavuje zvyčajne 10 – 15 mil. tokenov; tvorí sa pri každej verzii ako najväčší, východiskový súbor textov

all → tvorí sa pri každej verzii, obsahuje všetky texty

sane → tvorí sa pri každej verzii, neobsahuje texty nezodpovedajúce niektorým kritériám (texty s nesprávnou diakritikou, texty spred r. 1955, t. j. nespadajúce do vymedzeného obdobia súčasného jazyka, texty zahraničných Slovákov, ako aj lingvistické texty, ktoré sú sústredené v osobitnej položke Lingvistické zdroje)

Pri každej verzii sa teda tvoria podkorpusy public-all, public-sane, juls-all, juls-sane, snk-all, snk-sane.

vyv → tvorí sa od verzie prim1, štýlovo vyvážený podkorpus

prim1-vyv → 12 mil. tokenov 60 % inf : 20 % pfr : 20 % img

prim-2.0-vyv → 56 mil. tokenov 60 % inf : 20 % pfr : 20 % img

prim-2.1-vyv → 54 mil. tokenov 60 % inf : 20 % pfr : 20 % img

prim-3.0-vyv → 200 mil. tokenov 33,330 % inf : 33,344 % pfr : 33,312 % img : 0,014 % iné

inf → tvorí sa od verzie prim-2.1, podkorpus obsahujúci výlučne publicistické texty

prf → tvorí sa od verzie prim-2.1, podkorpus obsahujúci výlučne odborné texty

img → tvorí sa od verzie prim-2.1, podkorpus obsahujúci výlučne umelecké texty

skimg → tvorí sa od verzie prim-2.1, podkorpus obsahujúci výlučne originálne, pôvodné slovenské umelecké texty

Osobitným podkorpusom je ručne morfologicky anotovaný podkorpus, ktorý nie je fixovaný na verziu primárneho korpusu (má vlastné číslovanie):

r-mak-1.0 → sprístupnený začiatkom r. 2006, 322 600 tokenov

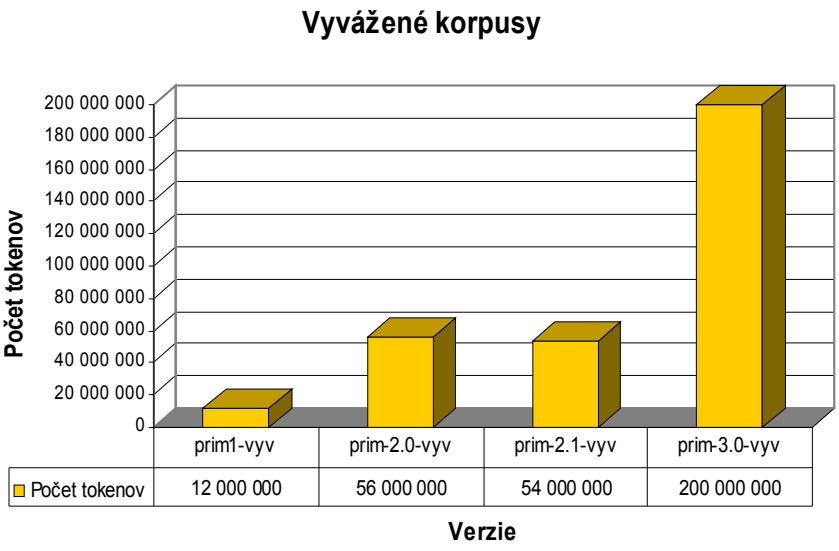
r-mak-2.0 → sprístupnený začiatkom r. 2007, 511 534 tokenov

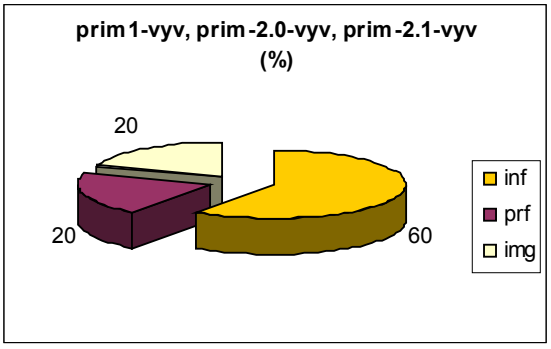
Špecifické podkorpusy sú paralelné korpusy:

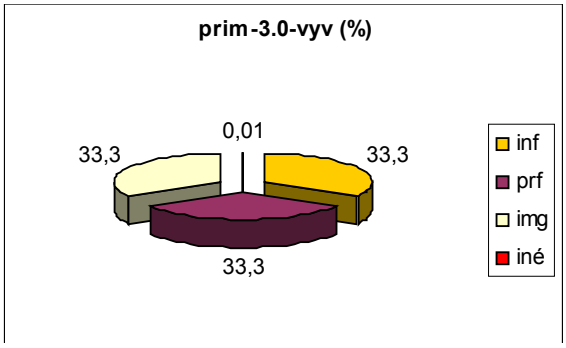
pcct → paralelný korpus počítačových termínov, sprístupnený v r. 2003

parus → paralelný rusko-slovenský korpus, sprístupnený koncom r. 2006, 900 000 tokenov v každom jazyku

frask → paralelný francúzsko-slovenský korpus, sprístupnený koncom r. 2006, 315 599 tokenov vo francúzskej časti, 194 478 tokenov v slovenskej časti

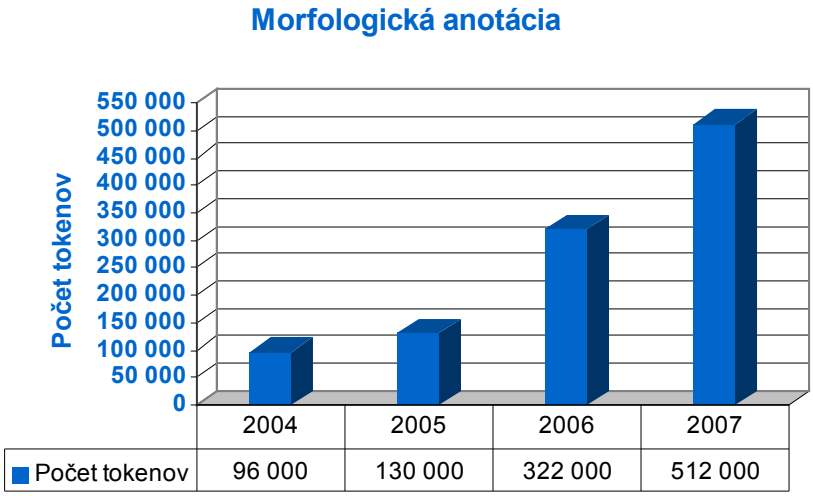


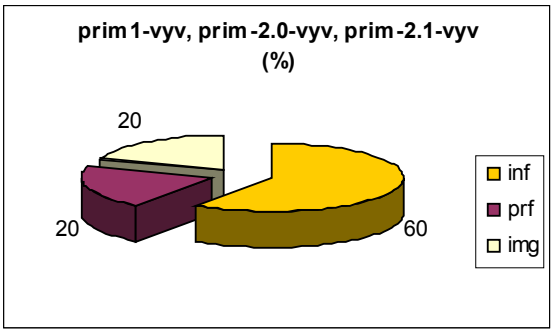


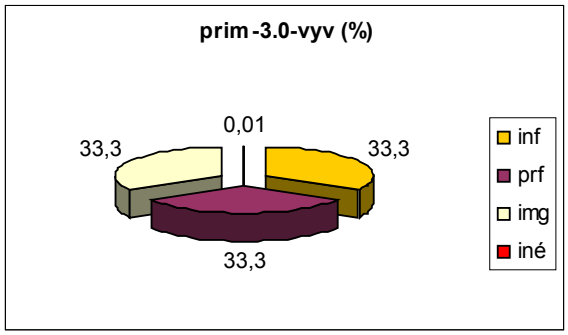


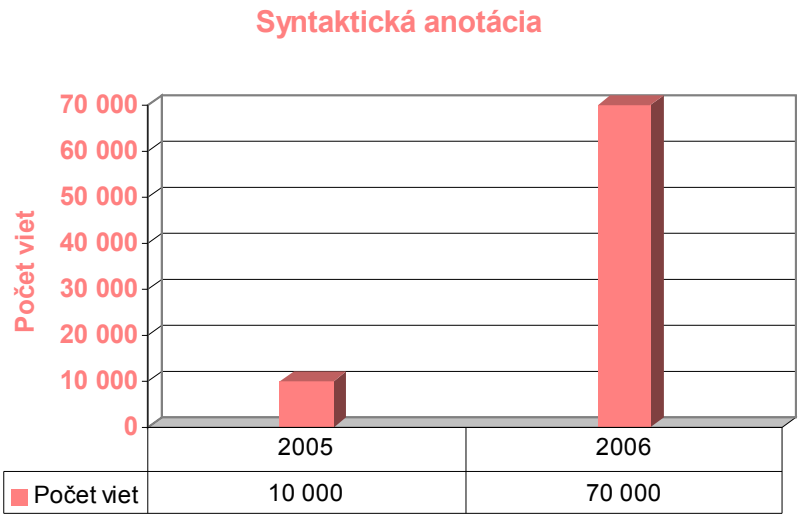
3.2.3. Ručná lingvistická anotácia vybraných textov

Grafické znázornenie množstva ručne morfologicky anotovaných tokenov a syntakticky anotovaných viet podľa jednotlivých rokov.









3.2.4. Lingvistické zdroje

Lingvistické zdroje na stránku SNK, ale aj JÚĽŠ SAV pribúdali takto:

2003 → Krátky slovník slovenského jazyka. 4., doplnené a upravené vydanie. Redigovali: J. Kačala – M. Pisárčiková – M. Považaj. Bratislava: Veda 2003.

2005 → Dvonč, L.: Slovenskí jazykovedci. Súborná personálna bibliografia slovenských slovakistov a slavistov. 1. (1925 – 1975). Martin, Matica slovenská 1987. 1388 s.

2. (1976 – 1985). Bratislava: Veda 1997. 700 s.
3. (1986 – 1995). Bratislava: Veda 1998. 728 s.
4. (1996 – 2000). Bratislava: Veda 2003. 505 s.

→ Kultúra slova. Texty ročníkov 1994 – 2004, obsahy ročníkov 1967 – 1992.

→ Štúr, L.: Nauka reči Slovenskej. V Prešporku 1846.

2006 → Slovenská reč. Texty ročníkov 1997 – 2005.

→ Jazykovedný časopis. Texty ročníkov 1997 – 2004.

→ Morfológia slovenského jazyka. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1966. 896 s.

→ Horecký, J. – Buzássyová, K. – Bosák, J. a kol.: Dynamika slovnej zásoby súčasnej slovenčiny. Bratislava: Veda 1989. 436 s.

→ Sociolinguistica Slovaca. 2, 3, 4, 5. 1996, 1997, 1999, 2000.

→ Majtán, M.: Názvy obcí Slovenskej republiky (Vývin v rokoch 1773 – 1997). Bratislava: Veda 1998. 600 s.

→ Slovník slovenského jazyka. Red. Š. Peciar. Bratislava: Vydavateľstvo SAV 1959 – 1968. 5 zv. zo 6.

→ Varia. Zborníky z kolokvií mladých jazykovedcov. I – VIII.

3.2.5. Spolupráca na grantoch, diplomové práce

Pracovníci oddelenia JÚĽŠ SAV sa podieľali a podieľajú na riešení nasledujúcich grantov:

2004 – 2006 **Morfosyntaktická analýza Slovenského národného korpusu.** Grant VEGA č. 1/1432/04. Zodpovedná riešiteľka doc. PhDr. Miloslava Sokolová, CSc., mim. prof., Filozofická fakulta Prešovskej univerzity Prešov

2004 – 2005 **Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov.** Grant v rámci medzivládnej dohody poskytovaný prostredníctvom ministerstiev školstva spolupracujúcich krajín. Partnerské pracovisko: Ústav formálnej a aplikovanej lingvistiky MFF UK Praha

2006 – 2008 **Explanačno-deskriptívna morfológia súčasnej slovenčiny.** Grant VEGA č. 2/6119/6. Zodpovedný riešiteľ prof. PhDr. Juraj Dolník, DrSc., JÚĽŠ SAV – FiF UK Bratislava

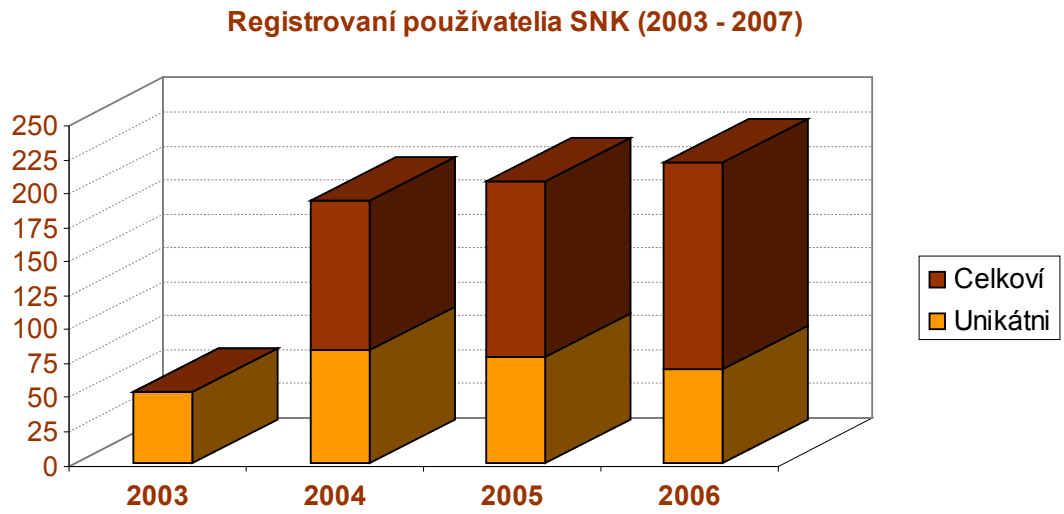
M. Šimková viedla v r. 2004 – 2005 dve diplomové práce, ktoré boli úspešne obhájené na Filozofickej fakulte Univerzity Komenského v Bratislave:

Cifrová, I.: Valencia adjektív v počítačovom spracovaní prirodzeného jazyka. Bratislava: Filozofická fakulta 2005. 62 s. + prílohy

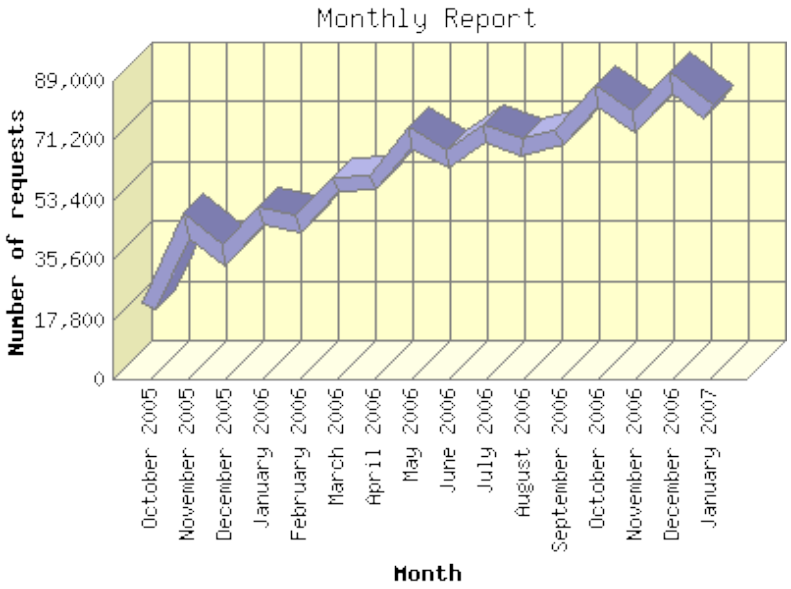
Majchráková, D.: Vyčleňovanie lexikalizovaných spojení pomocou štatistických nástrojov. Bratislava: Filozofická fakulta 2005. 63 s.

3.2.6. Používatelia SNK

Používateľov Slovenského národného korpusu delíme na dve skupiny z hľadiska registrácie: neregistrovaní používatelia pracujú jednoduchým spôsobom vyhľadávania pomocou www rozhrania (do r. 2006 pomocou vlastného korpusového manažéra korman) s obmedzenou verziou korpusu a môžu používať aj lingvistické zdroje (návštevnosť stránky SNK a lingvistických zdrojov predstavuje priemerne 10 000 vstupov denne). Registrovaní používatelia pracujú pomocou korpusového manažéra Manatee s klientom Bonito a býva ich približne 200 v rámci kalendárneho roka (registrácia sa vždy v januári zruší a záujemcovia si ju musia obnoviť).



	2003	2004	2005	2006
Celkoví	0	110	130	152
Unikátni	53	83	78	69
Spolu	**53**	**193**	**208**	**221**



4. Vyhodnotenie plnenia opatrení prijatých na priebežných oponentských konaniach

V záznamoch z priebežných oponentských konaní k úlohe výskumu a vývoja ***Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu*** tematického štátneho programu výskumu a vývoja ***Aktuálne otázky rozvoja spoločnosti*** sa priebežne konštatovalo kvalitné plnenie odborných cieľov úlohy v súlade s vecným a časovým harmonogramom. Na poslednom priebežnom oponentskom konaní za rok 2005 bola kvalita

riešenia úlohy hodnotená dvakrát v kategórii výborne (riešenie úlohy je na excelentnej svetovej úrovni) a raz v kategórii veľmi dobre (riešenie úlohy má veľmi vysoký štandard).

Za prvé dve priebežne hodnotené obdobia (rok 2003 a prvá polovica r. 2004) členovia oponentskej rady formulovali zásadné pripomienky týkajúce sa heterogénnosti posudzovaných podkladov v oblasti vykazovania čerpania finančných prostriedkov zo štátneho rozpočtu. Riešiteľovi bolo uložené skonkretizovať čerpanie a použitie prostriedkov štátneho rozpočtu, vypracovať prehľadnú tabuľku a komentár k štruktúre čerpania finančných prostriedkov. Uložené úlohy boli splnené v stanovenom termíne, t. j. k 15. 12. 2004. Po zlepšení harmonogramu financovania úlohy zo strany objednávateľa, keď pridelené finančné prostriedky neprichádzali na účet riešiteľa koncom rozpočtového roka, ale už aspoň v treťom štvrťroku, a po prijatí samostatnej ekonómky na pracovisko riešiteľa sa uvedené problémy v podstate eliminovali. Na ďalších priebežných oponentských konaniach neboli vznesené žiadne zásadné pripomienky k čerpaniu finančných prostriedkov a jeho vykazovaniu.

Za rok 2004 sa konali dve priebežné oponentské konania, na ktoré riešiteľ predložil dve samostatné správy. Uložené vypracovanie jednej syntetickej správy bolo splnené.

Na priebežnom oponentskom konaní za prvú polovicu roka 2005 bola na žiadosť riešiteľa vznesená zásadná pripomienka k vecnému a termínovému plneniu úlohy, a to aj v súvislosti s naakumulovaním finančných prostriedkov, ktoré sa presúvali v čerpaní od r. 2003. Oponentská rada na tomto základe odporúčala posunúť termín riešenia úlohy výskumu a vývoja na rok 2006 formou dodatku č. 4 k zmluve o dielo bez nároku na navýšenie finančných prostriedkov. Dodatok s vypracovanou finančnou kalkuláciou a harmonogramom riešenia na rok 2006 bol podpísaný a úloha sa riešila do 31. 12. 2006. Odporúčanie zaradiť pokračovanie úlohy výskumu a vývoja od roku 2006 do programu Komplexné riešenie podpory a efektívneho využívania infraštruktúry výskumu a vývoja, ktoré bolo formulované už v zázname z oponentského konania za rok 2004 a opakovane aj za prvý polrok 2005, sa napokon nerealizovalo zrejme vzhľadom na zmenu koncepcie zo strany objednávateľa a gestora programov výskumu a vývoja. Ostatné odporúčania a úlohy uložené aj na poslednom oponentskom konaní za celý rok 2005 (dopĺňanie dosiahnutých výsledkov na webovú stránku SNK, jej prelinkovanie na príslušné stránky MŠ SR, poskytnutie písomného stanoviska k oponentským posudkom, konkretizovanie výsledkov a výstupov, spracovanie návrhu realizačných výstupov, sledovanie a propagovanie výstupov riešenej úlohy aj po jej ukončení po dobu 3 rokov) sa podľa druhu úlohy splnili alebo priebežne plnia.

5. Popis prác na riešení úlohy a porovnanie dosiahnutých výsledkov riešenia s plánovanými cieľmi úlohy

Úloha štátneho programu výskumu a vývoja ***Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu*** sa začala riešiť od základov ako úloha interdisciplinárneho charakteru v odbore, ktorý bol vo vyspelých krajinách (aj v susednej Českej republike či Poľskej republike) v značnom štádiu rozvoja, no na Slovensku úplne absentoval.

V prvých etapách v r. 2003 – 2004 sa vypracúvala základná koncepcia a dokumentácia korpusu (porov. štúdie v časti 8 tejto správy a dokumentáciu priloženú k správam na priebežné oponentské konanie v r. 2004, dostupnú aj na http://korpus.juls.savba.sk) počnúc vymedzením rozsahu a obsahu základného, primárneho korpusu (200 miliónov textových jednotiek písaných textov súčasného slovenského jazyka obdobia rokov 1955 – 2005 s vonkajšou, bibliografickou a štýlovo-žánrovou, a vnútornou, morfologickou a syntaktickou anotáciou) cez štruktúru a technické spracúvanie dát v korpuse vrátane tokenizácie až po podrobnejšie koncepcie jednotlivých druhov anotácií a špecializovaných súčastí korpusu. Dôležitá bola v tejto fáze úzka spolupráca s korpusovo-lingvistickými a počítačovými pracoviskami najmä v ČR (Matematicko-fyzikálna a Filozofická fakulta Univerzity Karlovej v Prahe, Fakulta informatiky Masarykovej univerzity v Brne), z ktorých viacerí špičkoví odborníci prednášali na pôde SNK JÚĽŠ SAV v rámci špecializovaných seminárov a poskytli svoje skúsenosti, ako aj počítačové nástroje na analýzu a spracúvanie prirodzeného jazyka. Jednotliví pracovníci SNK JÚĽŠ SAV mali možnosť absolvovať na týchto pracoviskách krátkodobé pracovné pobyty a pravidelne sa zúčastňovať vedeckých seminárov a konferencií.

Po vypracovaní a oponovaní koncepcie budovania Slovenského národného korpusu a jeho stratifikácie (v rámci interných oponentských konaní na pôde JÚĽŠ SAV i diskusií na medzinárodných a domácich konferenciách) sa na báze Licenčnej zmluvy o inom (nekomerčnom) použití diela podľa autorského zákona a Vyhlásenia na vyhotovenie technickej rozmnoženiny začalo so získavaním a spracúvaním textov. Postupne sa formulovali zásady technického spracovania textov (konverzie elektronických verzií na jeden formát, skenovanie a zhotovovanie technických rozmnožení starších textov, odstraňovanie grafiky a iných netextových častí a pod.), zásady tokenizácie a lematizácie (segmentácie a prvotného lingvistického spracovania) získaných textov, ako aj zásady bibliografickej a štýlovo-žánrovej anotácie. Všetky dokumenty sú dostupné na stránke SNK JÚĽŠ SAV. Na extrakciu textových dát z rôznych vstupných formátov (spočiatku najmä T602, rtf, html, MS Word formát, neskôr PDF, quark, InDesign, PageMaker) do jednotného formátu SNK (inšpirovaný jazykom XML) boli postupne vyvinuté a zdokonalené potrebné konverzné nástroje. Len v r. 2006 bolo vytvorených 838 konverzných skriptov. Na vyhľadávanie (lingvistických) informácií v korpuse sa začal vyvíjať vlastný korpusový

manažér korman, súbežne sa používal korpusový manažér Manatee s klientom Bonito zakúpený z FI MU Brno. V r. 2003 bola na internete otvorená stránka Slovenského národného korpusu (http://korpus.juls.savba.sk) so základnými informáciami o korpusoch, korpusovej lingvistike a samotnom projekte a s prvou verziou základného, primárneho korpusu v rozsahu takmer 30 miliónov textových jednotiek. V tom istom roku bola sprístupnená druhá verzia korpusu a v druhej polovici r. 2004 bola verejnosti k dispozícii tretia verzia v rozsahu 182 mil. tokenov s prepracovanou tokenizáciou a štýlovo-žánrovou anotáciou a s úspešnou lematizáciou pomocou českého softvéru (z ÚFAL MFF UK Praha). Na interné použitie bola k dispozícii aj automatizovaná morfologická anotácia a dezambiguácia pomocou nástroja z toho istého pracoviska. Vzhľadom na rozdiely v morfologickej stavbe slovenčiny a češtiny, odlišné lingvistické hodnotenia niektorých javov a technicko-inžinierske východisko českého tagsetu sa začalo s vlastnou ručnou lematizáciou a morfologickou anotáciou, pre ktorú bol v oddelení SNK vyvinutý vlastný anotačný nástroj. Posledné dve verzie korpusu, završujúce práce za rok 2005 (prim-2.1 v rozsahu 300 miliónov tokenov) a za rok 2006 (prim-3.0 350 miliónov tokenov), ktoré boli oveľa náročnejšie z hľadiska získavania a spracúvania starších textov, obsahovali okrem automatizovanej morfologickej anotácie na základe českého tagsetu aj automatizovanú anotáciu na základe slovenského tagsetu a natrénovania softvéru na ručne anotovanom podkorpuse.

Cieľ sprístupniť 200-miliónový korpus textov bol prekročený takmer dvojnásobne. Na vytvorenie, sprístupnenie a prevádzkovanie korpusu bolo nevyhnutné splniť aj ďalšie body zmluvne požadovaného obsahu riešenia (porov. body 1., 2. a 3. v úvode časti 2), t. j. vypracovať celkovú koncepciu, ako aj čiastkové zásady pre jednotlivé oblasti získavania a spracúvania textov, testovať zahraničné počítačové nástroje a vyvíjať vlastné podľa aktuálneho stavu vývoja informačných technológií a špecifík slovenského jazyka.

Kým prvé verzie korpusu obsahovali predovšetkým publicistické texty, a to predovšetkým texty z posledného desaťročia (ide o pomerne ľahko získateľné a spracovateľné dáta), postupne sa do korpusu darilo získavať texty z rôznych štýlov a žánrov, ako aj z predchádzajúcich desaťročí. Pri posledných verziách sa už dosiahol lepší pomer štýlového zastúpenia (prim-3.0 obsahuje 56,74 % publicistických, 20,69 % umeleckých a 21,32 % odborných textov) aj lepšie zastúpenie textov z obdobia rokov 1955 (dolná hranica na zaradenie textu do korpusu súčasného jazyka) až 1992 (dokedy neexistujú elektronické verzie textov). Vzhľadom na to bolo možné a malo zmysel vytvoriť pre špecifické potreby niektorých výskumov samostatné podkorpusy s publicistickými textami, odbornými a populárno-náučnými textami a s umeleckými textami – spomedzi týchto sú dokonca osobitne vyčlenené originálne (pôvodné) slovenské umelecké texty. To všetko predovšetkým vďaka intenzívnemu získavaniu príslušných textov v rámci autorského zákona (v r. 2005 bolo oslovených a zazmluvnených 140

nových poskytovateľov textov, v r. 2006 to bolo rovných 100, spolu ich je 435) a vďaka skenovaniu a manuálnemu upravovaniu do elektronickej podoby tých textov, ktoré existujú iba v papierovej forme. V r. 2005 – 2006 bolo ročne naskenovaných a spracovaných takmer 60 tisíc strán textu. Pri čítaní a opravovaní skenovaných textov pomáhalo priemerne mesačne 30 študentov a iných externých spolupracovníkov.

Podľa plánu tak pokračovalo vyvažovanie korpusu a dopĺňanie chýbajúcich druhov textov. Štýlovo vyvážený korpus z poslednej verzie dosahuje hranicu 200 miliónov tokenov a možno ho považovať za prvú verziu reprezentatívneho korpusu textov súčasného slovenského jazyka. Tento podkorpus bude distribuovaný aj na CD a DVD nosičoch predovšetkým na výučbové využitie.

Po testovaní viacerých zahraničných softvérov na lematizáciu a morfologickú anotáciu textov sa v r. 2003 začalo s teoretickou prípravou vlastného tagsetu – súboru značiek a pravidiel na ručnú morfologickú anotáciu (Forróová – Garabík – Gianitsová – Horák – Šimková, 2003; http://korpus.juls.savba.sk/publications/index.sk.html). Po začatí praktického anotovania v r. 2004 sa tagset čiastočne modifikoval (Garabík – Gianitsová – Horák – Šimková, 2004; http://korpus.juls.savba.sk/publications/index.sk.html) a do konca roka sa zanotovalo takmer 100 tisíc textových jednotiek, časť z nich bola aj opravená. Začalo sa s anotáciou románu G. Orwella 1984 a vybraných populárno-náučných a publicistických textov (wikipedia, Sme, InZine). Na ručnej morfologickej anotácii sa podieľali študenti a absolventi slovenského jazyka z viacerých fakúlt (Bratislava, Trnava, Ružomberok, Prešov). V priebehu roka 2005 sa ich počet vzhľadom na postup prác a ďalšie faktory znížil z 15 na 4, v r. 2006 ich bolo 6 z FiF UK Bratislava. Každý text sa spočiatku anotoval dvakrát rôznymi anotátorkami, po porovnaní anotácií a odstránení takto zachytených chýb nasledovala ďalšia oprava supervízorkou. Postupne, so skvalitňovaním automatizovanej anotácie, sa pristúpilo k jednotlivým anotáciám, resp. opravám automatizovane zanotovaného textu (podrobnejšie porov. Garabík – Gianitsová-Ološtiaková, 2005; Karčová – Šimková, 2006; http://korpus.juls.savba.sk/publications/index.sk.html). Prvá verzia ručne morfologicky anotovaného podkorpusu r-mak-1.0 bola sprístupnená v rozsahu 322 600 tokenov začiatkom r. 2006, druhá verzia r-mak-2.0 o rok neskôr už obsahuje 511 534 tokenov. Ručne morfologicky anotovaný korpus slúži ako materiálová báza pre výskumy v oblasti gramatiky slovenského jazyka, na výučbové využitie a ako trénovací korpus pre morfologický anotátor.

Výsledky akvizície a spracovania textov, ako aj lingvistickej anotácie sú ľahko prístupné a ďalej skúmateľné prostredníctvom korpusového manažéra Manatee s klientom Bonito (z dielne Fakulty informatiky MU v Brne), ktorého nové verzie boli postupne sprevádzkované aj na vyhľadávanie v SNK a pre poslednú verziu bolo vytvorené vlastné www rozhranie. Vývoj

korpusového manažéra korman, ktorého pilotná verzia sa používala pri prvostupňovom vyhľadávaní v textoch korpusu na internete od r. 2003, sa vzhľadom na nedostatočné programátorské kapacity na pracovisku SNK a na sprístupnenie aktualizovanej verzie českého korpusového manažéra spĺňajúcej aj všetky potreby používateľov SNK, zastavil. Miesto toho sa v súlade so špecifickým charakterom slovenského jazyka pristúpilo k vývoju vlastného morfologického analyzátora a generátora tvarov slovenského jazyka, ktorý sa uplatňuje jednak pri zlepšovaní automatizovanej morfologickej anotácie, jednak v ďalších aplikáciách (opravovacie a vyhľadávacie programy). V r. 2005 sa realizovala konverzia dát na počiatočné naplnenie morfologického analyzátora, vypracovala sa metodika ručného dopĺňania dát a do konca roka boli pre morfologický slovník spracované vzory a paradigmy menných slovných druhov v rozsahu Krátkeho slovníka slovenského jazyka. Na konci roka 2006 obsahoval morfologický analyzátor slovenského jazyka podrobné gramatické kategórie flexie 56 269 slov rozdelených do 1 365 rôznych paradigiem ohýbania s celkovým počtom 601 253 unikátnych slovných tvarov a 1 616 379 rôznych dvojíc gramatických kategórií a slov. Analyzátor pokrýva veľkú väčšinu slov v bežných slovenských textoch s výnimkou skratiek, čísloviek a vlastných mien, ktorých zaradenie do slovníka sa bude realizovať v ďalšej fáze.

V júni 2005 sa začalo s ručnou syntaktickou anotáciou textov, ktoré boli predtým ručne anotované morfologicky. Východiskom pre anotáciu vybraných textov SNK sa stal manuál *Anotace na analytické rovině. Návod pro anotátory* vypracovaný v Ústave formálnej a aplikovanej lingvistiky MFF UK Praha v rámci projektu Pražského závislostného korpusu (PDT 1.0). Dôvodom na zvolenie smerovania v línii pražskej závislostnej syntaxe bola blízkosť syntaktických štruktúr oboch jazykov, blízkosť teoretických východísk na analytickej rovine a v neposlednom rade ústretovosť pražských kolegov ochotných odovzdať svoje cenné skúsenosti a poskytnúť aj kompletné softvérové vybavenie na syntaktickú anotáciu vrátane syntaktického parsera. Programátori SNK pripravili prostredie pre manuálne syntaktické anotácie a realizovali konverzie textov. Na ručnej syntaktickej anotácii sa podieľalo najskôr 11 spolupracovníkov, ku koncu roka 2005 ich bolo 6, v priebehu roka 2006 takisto 6 (išlo o študentky a absolventky odboru slovenský jazyk z Univerzity Cyrila a Metoda v Trnave, z Prešovskej univerzity, Katolíckej univerzity v Ružomberku a Univerzity Komenského v Bratislave). Anotovali sa texty, ktoré prešli ručnou morfologickou anotáciou, každý text dvoma rôznymi anotátormi. Týmto spôsobom je zanotovaných takmer 70 000 viet (jednotlivo 35 000), čo je dobré východisko na prípravu Slovenského závislostného korpusu. Uskutočnila sa prvá fáza opráv – korekcia delenia na vety a doanotovanie vynechaných uzlov.

Koncepčná príprava a realizácia lingvistického anotovania textov: morfologická a syntaktická anotácia prebehli v plnom rozsahu podľa požadovaného obsahu riešenia (bod 4 v úvo-

de časti 2). Vývojom morfologického analyzátora pre slovenčinu sa doplnil a rozšíril aj bod 3 – postupný vývoj vlastného softvéru. Plánovaný zámer sprístupniť 1. verziu Slovenského závislostného korpusu na internete do konca r. 2006 sa však vzhľadom na plnenie a prekračovanie iných úloh presunul do r. 2007.

Posledným bodom požadovaného obsahu riešenia bola príprava terminologickej databázy, databázy lexikografických diel, paralelného korpusu, diachrónneho korpusu, korpusu nárečových textov a korpusu hovorených prejavov. Už množstvo úloh sústredených v tomto jednom bode naznačuje, že ich formulovanie pri zadávaní riešenia úlohy a príprave projektu síce vychádzalo z názvu zahŕňajúceho komplexné spracovanie slovenského jazyka, ale nemohlo byť reálne splniteľné 8-členným kolektívom, ktorý mal predovšetkým vybudovať základný korpus písaných textov súčasného slovenského jazyka. Okrem databázy lexikografických diel každá z týchto úloh predstavuje objem prác, ktorý by mohol byť pre takýto kolektív predmetom samostatného projektu na niekoľko rokov. Zmluvne požadovaný obsah riešenia je v tomto bode širší aj v porovnaní s Harmonogramom riešenia úlohy (Príloha č. 2 k zmluve o dielo), kde sa diachrónny korpus ani korpus hovorených prejavov nenachádza. Budovanie korpusu nárečových textov bolo v pôvodnom harmonograme plánované na rok 2005, ale pri modifikácii harmonogramu na základe odsúhlasenia oponentskou radou a následného dodatku č. 4/2005 k zmluve o dielo sa nárečový korpus ako nerealizovateľný do nového plánu nezahrnul. Zo šiestich položiek v tomto bode požadovaného obsahu riešenia tak ostali tri: terminologická databáza, databáza lexikografických diel a paralelný korpus.

Pracovisko SNK začalo s prípravou vlastného projektu Slovenskej terminologickej databázy zameranej na oblasť práva, ekonomiky a vybraných technických vied v r. 2005, keď bola vypracovaná jej koncepcia. Absenciu takejto databázy roky kritizuje domáca prekladateľská i široká odborná verejnosť a v poslednom období ju akútne potrebujú aj prekladatelia Európskej komisie, Európskeho parlamentu, ako aj iných európskych inštitúcií. V roku 2006 sa v rámci prípravnej fázy projektu vypracoval návrh terminologického záznamu a návrh softvérovej štruktúry databázy s cvičnými údajmi. S cieľom zostaviť odborný ekonomický podkorpus bola nadviazaná spolupráca s Ekonomickou knižnicou, vydavateľstvom IURA Edition a jednotlivými autormi publikácií z oblasti ekonomických vied a práva, vybranými na základe vypracovanej bibliografie. Slovenský národný korpus bude aj ďalšími odbornými textami slúžiť ako zdroj na budovanie Slovenskej terminologickej databázy prostredníctvom automatizovanej extrakcie termínov z jednotlivých oblastí.

Databáza lexikografických diel, rozšírená o ďalšie lingvistické zdroje, nebola hlavnou úlohou riešiteľského kolektívu, no môžeme konštatovať, že u používateľov jazyka vzbudila najväčší záujem a najviac prispieva k návštevnosti stránky i oceňovaniu projektu elektronizácie

jazykovedného výskumu na Slovensku a využívaniu jeho výsledkov širokou verejnosťou. Už sprístupnenie Krátkeho slovníka slovenského jazyka v r. 2003 vzbudilo pozornosť aj v susednej Českej republike, kde niečo také nebolo možné dodnes uskutočniť, ale najmä u bežných pracovníkov so slovenským jazykom (učitelia, redaktori, ale aj zahraniční slovakisti). V r. 2005 boli sprístupnené všetky štyri zväzky bibliografií slovenských jazykovedcov od r. 1925 do r. 2000, ktoré sú značnou raritou už v knižnej podobe (nie je veľa odborov so systematickou a úplnou bibliografiou ani na Slovensku, ani vo svete), a ich digitalizácia a poskytnutie na internete uľahčuje orientáciu v otázkach slovenského jazyka i jazyka vôbec skúmaných slovenskými slovakistami a slavistami za celý čas novodobej slovenskej jazykovedy. Postupné spracúvanie a sprístupňovanie časopisov vydávaných Jazykovedným ústavom Ľ. Štúra SAV (Kultúra slova, Slovenská reč, Jazykovedný časopis) zaradilo časopiseckú produkciu ústavu medzi ostatné na internete prístupné odborné a vedecké periodiká. Osobitne Kultúru slova uvítali Slováci pracujúci so slovenským jazykom v zahraničí (lektori, prekladatelia) vzhľadom na mnohé aktuality, ktoré sa v nej publikujú (nové terminologické súpisy, nové slová, odpovede na otázky a pod.). Keď v r. 2006 pribudol do databázy Slovník slovenského jazyka (1959 – 1968) a Názvy obcí Slovenskej republiky, mnohí používatelia sa na internete (blogy) i na iných miestach veľmi pochvalne vyjadrovali o práci v JÚĽŠ SAV a slovníky zaradili medzi svoje obľúbené linky. Rozšírenie položky Lingvistické zdroje o vybrané monografie a zborníky (porov. v časti 3.2.4) ďalej prispieva k sprístupňovaniu slovenskej jazykovednej produkcie všetkým záujemcom o slovenský jazyk.

SNK sa buduje ako všeobecný jednojazyčný korpus, ale v roku 2005 sa začali systematické práce aj na paralelných korpusoch (slovensko-ruský, slovensko-francúzsky, slovensko-český). V prvej fáze išlo o získavanie textov ruských, francúzskych a českých originálov a prekladov, z ktorých sa začal budovať korpus paralelných textov najmä z oblasti umeleckej literatúry. Vzhľadom na náročnosť získavania textov na základe Licenčnej zmluvy (potrebný je súhlas na text originálu i prekladu) postupuje budovanie týchto korpusov veľmi pomaly. Vyvinuli sa nástroje potrebné na zarovnávanie textov, riešila sa otázka nástrojov na vyhľadávanie v paralelnom korpuse a využitie existujúcej morfologickej anotácie obidvoch spracovaných jazykov. Výsledkom bolo vytvorenie infraštruktúry na efektívnu tvorbu paralelných korpusov a korpusového manažéra pre paralelné korpusy. Koncom roka 2006 bol na internete sprístupnený rusko-slovenský paralelný korpus v rozsahu približne 900 000 slov (pre každý jazyk) a francúzsko-slovenský paralelný korpus v rozsahu vo francúzskej časti 315 599 tokenov, 13 004 viet a v slovenskej časti 194 478 tokenov, 12 286 viet. Česko-slovenský paralelný korpus bude sprístupnený v r. 2007 po doriešení súhlasov na všetky doteraz získané texty a po doriešení spolupráce pri lematizácii a morfologickej anotácii a dezambiguácii českých textov.

Zredukovanie posledného súboru úloh na tri umožnilo riešiteľskému kolektívu bezo zbytku splniť aj tento bod požadovaného obsahu riešenia. Slovenská terminologická databáza má za sebou prípravnú fázu, databáza lexikografických diel obsahuje s lingvistickými zdrojmi dohromady 12 základných položiek (nepočítame jednotlivé zväzky niektorých diel či čísla časopisov a zborníkov) a predstavuje vynikajúci zdroj informácií o slovenskom jazyku a jeho výskume. Obidve databázy sú dôležitými a všeobecne využiteľnými zložkami elektronizácie jazykovedného výskumu na Slovensku. Paralelné korpusy sú zatiaľ sprístupnené dva a môžu slúžiť na komparatívne výskumy, ako aj na výučbové využitie pre cudzincov alebo translatológov.

6. Použitá technická infraštruktúra

6.1. Hardware

Ku koncu roka 2006 pozostávala počítačová sieť SNK z dvoch serverov, 10 pracovných staníc pre pracovníkov SNK a 23 pracovných staníc pre externých spolupracovníkov. Na jednej pracovnej stanici je ako operačný systém použitý Windows XP Professional, na všetkých ostatných je GNU/Linux Kubuntu (breezy). Na serveri jano je GNU/Linux Debian (testing), na serveri palo Ubuntu (breezy). Dve pracovné stanice sú Mac Mini, na jednej je operačný systém MacOS X Tiger, na druhej GNU/Linux Debian (unstable).

Sieť je logicky súčasťou privátnej siete JÚĽŠ SAV s rozsahom IP čísel 192.168.0.0/255.255.255.0, pričom servery jano a palo majú prístup aj do vonkajšej siete JÚĽŠ SAV 147.213.138.0/255.255.255.0. Fyzicky je sieť SNK tvorená štruktúrovanou kabelážou s hierarchicky zapojeným väčším počtom gigabitových ethernet switchov.

K dispozícii je vybavenie na vybudovanie wi-fi siete, ktoré slúži na mobilné pripojenie prenosných počítačov a PDA do siete SNK.

Popis serverov:

jano.juls.savba.sk, Compaq Prdiant server, CPU Intel Pentium 4
1GHz, 512 MB RAM, 18 GB SCSI disk

palo.juls.savba.sk, Noname, CPU Intel Pentium 4 3 GHz, 2 GB RAM,
2×120 GB IDE HDD (zapojené ako RAID-1)

Server jano slúži ako file server pre korpusové dáta, palo je určený predovšetkým ako fileserver pre dáta užívateľov a ako server sprostredkujúci vonkajšie služby a presmerovanie služieb do vnútra privátnej siete na clustrové počítače. K serveru jano je pripojené vysokokapacitné IDE RAID pole s celkovou diskovou kapacitou 4.5 TB, ktoré je využité na ukladanie hlavných dát korpusu.

Popis služieb poskytovaných serverom jano:

- NFS a SAMBA export diskového poľa (/data)
- sekundárny DNS server pre doménu juls.savba.sk
- WWW rozhranie k prototypu korpusového manažéra
- MySQL databáza s metadátami korpusu

Popis služieb poskytovaných serverom palo:

- NFS a SAMBA export domovských užívateľských adresárov (/home)
- port forwarding do vnútornej siete na poskytovanie služby korpusového manažéra Manatee
- mail server pre adresu korpus.juls.savba.sk (postfix) + antispamová ochrana (greylisting, spamassassin)
- NIS autentifikačný server pre pracovné stanice
- WWW stránka korpusu

Cluster

Na náročné výpočtové úlohy súvisiace s konverziou dát slúži výpočtový cluster pozostávajúci z 11 počítačov (CPU Intel Pentium 4,3 GHz, 2 GB RAM, 120 GB IDE HDD) umiestnených vo vnútornej sieti. Cluster zabezpečuje pre verejnosť priamo službu korpusového manažéra (Manatee) a v rámci SNK konverzie textov.

6.2. Software

Ako korpusový manažér sa v SNK používa Manatee s grafickým klientom Bonito, ktorý okrem obvyklého hľadania podporuje regulárne výrazy, nastaviteľné atribúty textu a tokenov, bibliografickú anotáciu a autentifikáciu registrovaných užívateľov.

Manatee pracuje pod operačným systémom GNU/Linux, klient Bonito je napísaný v skriptovacom jazyku TCL/Tk a teoreticky funguje na všetkých platformách, kde funguje TCL/Tk. Prakticky má autor otestovanú funkčnosť na moderných POSIXových operačných systémoch, MacOSX a na Microsoft Windows 95/98/NT a kompatibilných, čo sprístupňuje korpus prakticky každému záujemcovi s dostatočne modernou výpočtovou technikou.

Okrem toho je na WWW stránke SNK prístupná obmedzená verzia korpusu bez nevyhnutnosti zaregistrovať sa. Táto verzia slúži na jednoduché rýchle hľadanie pomocou rozhrania ku korpusovému manažéru postaveného na báze web application frameworku Karrigell a pythonovského API k Manatee.

Pilotným projektom, ktorého výsledok je od novembra 2003 prístupný cez WWW stránku SNK, je Parallel Corpus of Computer Terms. Ide o prototyp paralelného korpusu založeného na dostupnosti prekladov z oblasti výpočtovej techniky – typicky užívateľských rozhraní k programom. Je implementovaný kompletne v programovacom jazyku Python a na ukladanie dátových štruktúr používa relačnú databázu (MySQL).

Ďalšie dva projekty sprístupnené cez WWW rozhranie na stránke SNK sú paralelné korpusy – paralelný rusko-slovenský korpus a paralelný francúzsko-slovenský korpus, ktoré ob-

sahujú po vetách zarovnané texty (prevažne beletristické) s možnosťou korpusového vyhľadávania. Korpusy používajú WWW rozhranie totožné s rozhraním primárneho korpusu, s úpravami pre možnosť paralelného vyhľadávania a zobrazovania konkordancií.

V spolupráci s Tauridským národným inštitútom V. I. Vernadského bol vytvorený základný korpus súčasného krymskotatárskeho jazyka v rozsahu 371 536 tokenov prevažne publicistického štýlu, v cyrilickej ortografii. V korpuse je možné vyhľadávať pomocou regulárnych výrazov, čo vzhľadom na aglutinačný charakter krymskotatárskeho jazyka dostatočne nahrádza chýbajúcu morfologickú analýzu.

Na pracovisku bol vyvinutý morfologický analyzátor slovenského jazyka s podrobnými gramatickými kategóriami flexie rozdelenými do rôznych paradigiem ohýbania, ktorý pokrýva veľkú väčšinu slov v bežných slovenských textoch. Analyzátor pracuje na princípe aplikovania postupnosti levenštejnových editovacích operácií na lemu konkrétneho slova podľa jej príslušnosti k danému vzoru. Vygenerované tvary sú uložené v cdb konštantnej databáze, ku ktorej sa dá veľmi efektívne pristupovať a vyhľadávať. API k analyzátoru je tvorené modulom v programovacom jazyku Python, poskytujúcom logiku potrebnú na vyhľadávanie slov, lem a gramatických kategórií a na analýzu niektorých morfologických javov.

Na projekt Slovenskej terminologickej databázy sú vytvorené základné programovacie prostriedky slúžiace na efektívne zapĺňanie databázy, na prezeranie a modifikácie vstupov. Databáza a jej možnosti editovania využíva MoinMoin wiki-systém s možnosťou autorizovanej editácie registrovanými užívateľmi, so systémom kontroly revízií a efektívnych kolaboračných úprav.

Syntaktická analýza používa grafický editor závislostných stromov Tred vyvinutý na MFF UK v Prahe.

7. Čerpanie finančných prostriedkov na riešenie úlohy

Komentár k tabuľkám na nasledujúcich stranách:

1. Na základe žiadosti o predĺženie zmluvy o dielo do konca roku 2006 a následnej žiadosti o presun finančných prostriedkov vo výške 1 600 tis. Sk naakumulovaných v predchádzajúcich rokoch riešenia úlohy mohlo pokračovať riešenie a financovanie úlohy štátneho programu do 31. 12. 2006. Vzhľadom na charakter prác v r. 2006 sa až 70 % týchto prostriedkov čerpalo predovšetkým na vyplatenie dohôd o vykonaní práce. Podľa plánu sa čerpali finančné prostriedky na služobné cesty v rámci Slovenska i do zahraničia, kde členovia riešiteľského kolektívu prezentovali práce vykonané na projekte alebo sa zúčastňovali vedeckých seminárov z oblasti všeobecnej, počítačovej a korpusovej lingvistiky. Z výpočtovej techniky sa dokúpili iba najnevyhnutnejšie súčasti krátkodobej spotreby.

2. Mzdy a odvody za riešiteľský kolektív v r. 2006 (1 750 + 607 tis. Sk) tak ako počas celej doby riešenia úlohy štátneho programu uhradila SAV zo svojej rozpočtovej kapitoly.

Tabuľkový prehľad čerpania financií úlohy "Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu" za rok 2006

Ekonomicky oprávnené náklady úlohy výskumu a vývoja	Plánované výdavky na celú dobu riešenia Celkom ŠR v tis. Sk	2006	
		Plánované výdavky celkom ŠR v tis. Sk	Skutočné výdavky celkom ŠR v tis. Sk
Bežné výdavky, z toho:	**7 882,-**	**1 600,-**	**1 600,-**
Mzdy,platy,odmeny a OOV	429,-	177,-	0,-
Náklady na poistenie, NÚP	0,-	0,-	0,-
Tovary a ďalšie služby z toho:	7 453,-	1 423,-	1 600,-
cestovné výdavky,	290,-	130,-	119,-
energia, voda a komunikácie	180,-	200,-	44,-
materiál a dodávky	330,-	93,-	337
dopravné	0,-	0,-	0,-
nájomné za prenájom	0,-	0,-	0,-
ostatné tovary a služby	6 653,-	1 000,-	1 100,-
Kapitálové výdavky:	**2 436,-**	**0,-**	**0,-**
nákup nehm.inv.maj.	300,-	0,-	0,-
nákup prev.str.a zar.	0,-	0,-	0,-

Z rozpočtovej kapitoly SAV:

mzdy: 1 750 000,- Sk

odvody: 607 000,- Sk

tovary a služby: 796 000,- Sk

kapitálové výdavky: 100 000,- Sk

Rekonštrukcie a modernizácie	2 136,-	0,-	0,-
Spolu	**10 318,-**	**1 600,-**	**1 600,-**

V r. 2006 poskytla SAV na riešenie úlohy aj 100 tis. Sk kapitálových prostriedkov a 796 tis. Sk na tovary a služby.

3. Cieľové odmeny ani kapitálové prostriedky MŠ SR ako objednávateľ neposkytlo ani jeden rok napriek tomu, že sú zakotvené v zmluve.

4. Zhodnotenie nákladov za celú dobu riešenia štátneho programu ukazuje, že v prvých dvoch rokoch sa najväčšia časť nákladov použila na nákup materiálu a dodávky (potrebné technické vybavenie a zariadenie priestorov na prácu). V posledných dvoch rokoch sa väčšia časť finančných prostriedkov vynakladala na vyplácanie dohôd o vykonaní práce externým spolupracovníkom.

8. Ukazovatele výstupov riešenej úlohy a jej spoločensko-hospodársky dopad

(Príloha č. 6 podľa Usmernenia Ministerstva školstva Slovenskej republiky pre oponentúry úloh výskumu a vývoja štátnych programov výskumu a vývoja)

Evidenčné číslo úlohy: 2003SP200280307	**Termín riešenia:** 1. 7. 2003 – 31. 12. 2006
Dodávateľ riešenia: Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied	
Názov štátneho programu: Aktuálne otázky rozvoja spoločnosti	
Názov úlohy: Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu	

I. **Kategória – publikácie vzniknuté v rámci riešenie úlohy**

o vedecké práce publikované v recenzovaných vedeckých časopisoch a zborníkoch (zahraničné/domáce)

GARABÍK, Radovan: Corpus Construction Tools. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Zborník z medzinárodnej vedeckej konferencie MegaLing'05. Meganom, Krym, Ukrajina 27. júna – 2. júla 2005. Red. V. P. Zacharov – S. S. Dikareva. St. Petersburg: Izdateľstvo Osipov 2005, s. 26 – 32.

GARABÍK, Radovan: Levenshtein Edit Operations as a Base for a Morphology Analyzer. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005, s. 50 – 58.

GARABÍK, Radovan: Processing XML Text with Python and Element Tree – a Practical Experience. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2005, s. 160 – 165.

GARABÍK, Radovan: Slovak morphology analyzer based on Levenshtein edit operation. In: WIKT 2006 – 1st Workshop on Intelligent and Knowledge oriented Technologies. Pro-

ceedings of the WIKT'06 conference. Bratislava 2006. V tlači.

GARABÍK, Radovan – GIANITSOVÁ-OLOŠTIAKOVÁ Lucia: Manual Morphological Annotation of the Slovak Translation of Orwell's Novel 1984 – Methods and Findings. In: Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Ed. R. Garabík. Bratislava: Veda 2005, s. 59 – 66.

GARABÍK, Radovan – ЗАХАРОВ, Виктор Павлович: Параллельный русско-словацкий корпус. In: Труды международной конференции Корпусная лингвистика. Zborník z medzinárodnej konferencie konanej v Sankt-Petersburgu 10. – 14. októbra 2006. Санкт-Петербург: Издательство С.-Петербургского университета 2006, s. 81 – 87.

GIANITSOVÁ, Lucia: Morphological Analysis of the Slovak National Corpus. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2005, s. 166 – 178.

HORÁK, A. – GIANITSOVÁ, L. – ŠIMKOVÁ, M. – ŠMOTLÁK, M. – GARABÍK, R.: Slovak National Corpus. In: Text, Speech and Dialogue. 7th International Conference TSD 2004 Proceedings. Ed. P. Sojka – I. Kopeček – K. Pala. Berlin – Heidelberg: Springer – Verlag 2004, s. 89 – 94.

MAJCHRÁKOVÁ, Daniela – KARČOVÁ, Agáta: Distribúcia prípon *-mi, -ami* v tvaroch maskulínnych substantív zakončených na konsonantické skupiny. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity 2006.

ŠIMKOVÁ, Mária: Čo je možné dozvedieť sa zo Slovenského národného korpusu. In: Čeština doma a ve světě, 2004, roč. 12, č. 3 – 4, s. 130 – 145.

ŠIMKOVÁ, Mária: Репрезентативность корпуса как лингвистическая проблема. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Zborník z medzinárodnej vedeckej konferencie MegaLing'05. Meganom, Krym, Ukrajina 27. júna – 2. júla 2005. Red. V. P. Zacharov – S. S. Dikareva. St. Petersburg: Izdateľstvo Osipov 2005, s. 130 – 139.

ŠIMKOVÁ, Mária: Slovak National Corpus – history and current situation. In: Insight into Slovak and Czech Corpus Linguistics. Ed. M. Šimková. Bratislava: Veda 2005, s. 152 – 159.

ŠIMKOVÁ, Mária: Princípy a problémy spracovania neohybných slovných druhov v ručne anotovaných textoch Slovenského národného korpusu. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity 2006.

ŠIMKOVÁ, Mária – GARABÍK, Radovan: Синтаксическая разметка в Словацком национальном корпусе In: Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006, s. 389 – 394. ISBN 5-288-04181-4

- odborné práce publikované v recenzovaných odborných časopisoch a zborníkoch (zahraničné/domáce),

DOMIN, Pavol: Niektoré technické aspekty budovania a používania korpusov. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 159 – 163.

FORRÓOVÁ, Martina – HORÁK, Alexander: Morfologická anotácia korpusu. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 174 – 186.

GARABÍK, Radovan: Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia. In: Slovenčina na začiatku 21. storočia. Ed. Mária Imrichová. Prešov: Prešovská univerzita, Fakulta humanitných a prírodných vied 2004, s. 164 – 173.)

GARABÍK, Radovan: Parallel Corpus of Computer Terms. In: Varia XIII. Zborník materiálov z XIII. kolokvia mladých jazykovedcov (Modra – Piesok 3. – 5. 12. 2003). Zost. M. Šimková – K. Gajdošová. Bratislava: Slovenská jazykovedná spoločnosť pri SAV – Jazykovedný ústav Ľ. Štúra SAV 2006, s. 80 – 88. ISBN 80-89037-04-6.

GARABÍK, Radovan: Словацкий национальный корпус. In: Труды международной конференции Корпусная лингвистика. Zborník z medzinárodnej konferencie konanej v Sankt-Petersburgu 11. – 14. októbra 2004. St. Petersburg: Izdateľstvo S.-Peterburgskogo universiteta 2004, s. 99 – 121.

GARABÍK, Radovan: Computer(ized) linguistic resources at the Ľ. Štúr Institute of Linguistics. In: Лексикографічний бюлетень 13. Київ: Інститут української мови НАН України 2006. p. 56 – 59. Proceedings of the conference Applied (computer) Linguistics. 27-28 April 2006. Kiev, Ukraine.

GARABÍK, Radovan – КУБЕДИНОВА, Ленара Шакировна: Корпус текстов крымскотатарского языка. In: Труды международной конференции MegaLing'2006 – Горизонты прикладной лингвистики и лингвистических технологий. Партенит, Ukraine St. In print.

GIANITSOVÁ, Lucia: Zamyslenie nad výučbou zámen a čísloviek pri príprave morfologickej anotácie SNK. In: Tradiční a netradiční metody a formy práce ve výuce českého jazyka na základní škole. Sborník prací z mezinárodní konference konané 19. 3. 2004 na Pedagogické fakultě UP v Olomouci. Ed. M. Polák – K. Vodrážková. Olomouc: Univerzita Palackého, 2005, s. 53 – 65.

GIANITSOVÁ, Lucia: Možnosti využitia SNK v pedagogickej praxi. In: Zborník z konferencie Učebné materiály vo vyučovaní materinského jazyka, slohu a literatúry. Banská Bystrica: Pedagogická fakulta Univerzity M. Bela 2004.

KARČOVÁ, Agáta – ŠIMKOVÁ, Mária: Морфологічна анотація текстів словацького національного корпусу. In: Лексикографічний бюлетень 13. Київ: Інститут української мови Національної академії наук України 2006. s. 71 – 76. ISBN 966-02-4002-3.

LEVICKÁ, Jana: Proces ustaľovania terminologických sústav a prínos terminologických databáz. In: Vedecká rozprava k problémom terminológie obchodu, obchodných vied a marketingu obchodu. Ed. K. Viestová. Bratislava: Ekonomická univerzita 2005, s. 22 – 29.

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

ŠIMKOVÁ, Mária: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca 33. Prednášky z XL. letnej školy slovenského jazyka a kultúry. Ed.: Jozef Mlacek – Miloslav Vojtech. Bratislava: Filozofická fakulta Univerzity Komenského 2004, s. 204 – 218.

- odborné práce publikované v nerecenzovaných odborných časopisoch a zborníkoch (zahranič-

né/domáce),

GIANITSOVÁ, Lucia: Pokus o vymedzenie syntetických a analytických tvarov v slovenčine s ohľadom na školskú prax. In: Slovenský jazyk a literatúra v škole 2003/04, č. 3 – 4.

HORÁK, Alexander: Prehľad textových korpusov slovanských jazykov. In: Varia XII. Zost. Mária Šimková. Bratislava: Slovenská jazykovedná spoločnosť pri SAV 2005, s. 20 – 28.

LEVICKÁ, Jana: Terminologie et Informatique. In: Actes du 8e Séminaire international d'études doctorales. Prešov: Filozofická fakulta Prešovskej univerzity 2004, s. 191 – 195.

LEVICKÁ, Jana: Terminologická synonymia v oblasti stavebníctva vo francúzštine a problematika ekvivalentácie. In: Odborný preklad 1. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2005, s. 117 – 133.

LEVICKÁ, Jana: Teoretické východiská budovania terminologickej databázy. In: Odborný preklad 2. Ed. J. Šoltýs. Bratislava: AnaPress/Slovenská spoločnosť prekladateľov odbornej literatúry 2006, s. 73 – 81. ISBN 80-89137-54-5.

TARABA, Ján – LEVICKÁ, Jana: Approche contextuelle du vocabulaire économique francais-slovaque. In: Philologica LIX Studia romanistica. Bratislava: Filozofická fakulta Univerzity Komenského 2005, s. 81 – 93.

ŠIMKOVÁ, Mária: Počítačové spracovanie prirodzeného jazyka a Slovenský národný korpus. In: Počítačová podpora prekladu. Zborník prednášok. Ed. Marián Smolík – Jaroslav Šoltys – František Tomášik. Bratislava: Slovenská spoločnosť prekladateľov odbornej literatúry 2003, s. 15 – 19.

o odborné knižné publikácie (zahraničné/domáce)

Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005. 248 s.

II. Kategória – Výsledky chránené na základe osobitného právneho predpisu, alebo výsledky uplatnené na trhu (jedná sa o patenty, licencie, úžitkové vzory a pod.; do tejto kategórie nepatri obchodné tajomstvo a pod.);

-

III. **Kategória - Aplikované výsledky** (okrem identifikačných údajov, ako sú lokalizácia výsledku; technické parametre výsledku; ekonomické parametre výsledku; názov vlastníka výsledku - uviesť aj stručný slovný popis výstupu a súčasne aj spôsob realizácie – už uplatnenej, alebo plánovanej)

-

- softvérový produkt,

Na pracovisku Slovenského národného korpusu boli riešiteľským kolektívom vyvinuté vlastné softvérové produkty na spracovanie textov a efektívnu prácu s nimi:

– anno – nástroj na ručnú morfologickú anotáciu slovenských textov – uplatnený

– morfologický analyzátor a lematizátor – nástroj na automatizovanú lematizáciu (rozpoznávanie základných tvarov slov) a automatizovanú morfologickú anotáciu slovenských textov (priraďovanie informácií o tvare slova v texte na báze slovníka vzorov a paradigiem) –

uplatnený

– dictd – serverový softvér na efektívne vyhľadávanie v slovníkových databázach podľa protokolu RFC 2229 – pomocou neho sú na internete prístupné na prezeranie dáta v databáze lexikografických príručiek – uplatnený

– k3t – systém na efektívnu organizáciu textových dát a ich usporiadanie ako vstupných dát v rámci korpusu – uplatnený

– programy na extrakciu textových dát z rôznych vsupných súborov, formátov: T602, PDF, rtf, html, MS Word formát, quark, InDesign, PageMaker – uplatnený

– paralelný korpus počítačových výrazov – uplatnený

– paralelný korpus rusko-slovenský a francúzsko-slovenský – uplatnený

– wikiengine umožňujúci ukladanie, úpravu a efektívne fulltextové hľadanie v dokumentoch prostredníctvom http protokolu – na vyhľadávanie v bibliografii slovenských jazykovedcov, v odborných a vedeckých časopisoch JÚĽŠ SAV a v ďalších dokumentoch spracovaných a sprístupnených na internete v položke Lingvistické zdroje – uplatnený

– vylepšenie karrigell-u – umožnenie práce s textami v Unicode – uplatnený

– unicode – program na rýchly a efektívny výpis informácií o vlastnostiach znakov v Unicode – uplatnený

-

IV. Kategória - Ostatné výsledky (okrem identifikačných údajov uviesť aj stručný slovný popis výstupu)

-

- elektronické dokumenty, t.j. dokumenty vydané len vo forme čitateľnej prostredníctvom počítača, internetu a pod.,

Slovenský národný korpus vo všetkých verziách a so špecifickými súčasťami – http://korpus.juls.savba.sk

Vyvážený korpus z verzie prim-3.0 na CD/DVD nosiči s vybranými špecifickými súčasťami korpusu (ručne morfologicky anotovaný podkorpus) a sprievodnými textami.

- usporiadanie/zorganizovanie konferencie,

10. – 12. novembra 2005 zorganizovali pracovníci oddelenia Slovenského národného korpusu medzinárodnú konferenciu Slovko 2005 – Slovanské a východoeurópske jazyky v počítačovom spracovaní.

- usporiadanie/zorganizovanie výstavy,

-

V dňoch 13. 12. 2006 – 18. 2. 2007 sa v Slovenskom národnom múzeu v Bratislave konala výstava Komunikácia.2007, na ktorej bol prezentovaný aj Slovenský národný korpus. Na jej príprave spolupracovala za riešiteľský kolektív M. Šimková.

- ostatné výsledky, ktoré nie je možné zaradiť do žiadneho z vyššie uvedených druhov výsledkov (koncepcie, metodiky, štúdie a pod., ktoré riešiteľ v podobe zmluvného zabezpečenia, alebo inej

formy záväzku odovzdáva realizátorovi pre konkrétne aplikácie a využitie v hospodárskej a spoločenskej praxi, buď s okamžitým využitím alebo s perspektívou využitia v budúcich obdobiach)

Koncepcia Slovenskej terminologickej databázy

FORRÓOVÁ, Martina – GARABÍK, Radovan – GIANITSOVÁ, Lucia – HORÁK, Alexander – ŠIMKOVÁ, Mária: Návrh morfologického tagsetu SNK. In: Slovanské jazyky v počítačovom spracovaní Slovko 2003. Medzinárodná konferencia. Materiály neboli publikované. http//korpus.juls.savba.sk/publications

GARABÍK, Radovan: Some Aspects of Tokenization and Segmentation in Slovak National Corpus. In: Slovanské jazyky v počítačovom spracovaní Slovko 2003. Medzinárodná konferencia. Materiály neboli publikované. http//korpus.juls.savba.sk/publications

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

V. Kategória – Pridaná hodnota riešenej úlohy výskumu a vývoja (jej spoločensko.-hospodársky dopad)

- zoznam diplomantov a diplomových prác, ktoré sa obhájili v rámci úlohy (uviesť meno študenta, tému diplomovej práce, časové rámce a vedúceho diplomovej práce),

CIFROVÁ, Ivana: Valencia adjektív v počítačovom spracovaní prirodzeného jazyka. Bratislava: Filozofická fakulta Univerzity Komenského.

Obhájená v júni 2005.

Diplomová vedúca: Mária Šimková

MAJCHRÁKOVÁ, Daniela: Vyčleňovanie lexikalizovaných spojení pomocou štatistických nástrojov. Bratislava: Filozofická fakulta Univerzity Komenského.

Obhájená v júni 2005.

Diplomová vedúca: Mária Šimková

- zoznam PhD študentov, ktorí sa školia v rámci riešenia úlohy (uviesť meno PhD študenta, tému dizertačnej práce, časové rámce a školiteľa),

VASILIŠINOVÁ, Dorota: Tvorba terminologickej databázy slovenskej právnickej terminológie na podklade paralelného slovensko-francúzskeho a francúzsko-slovenského korpusu práva Európskej únie
Štúdium od 1. 10. 2006 do 30. 9. 2010
Školiteľ: doc. RNDr. Vladimír Petkevič, CSc.

- ďalšie projekty výskumu a vývoja, ktoré priamo súvisia s riešenou úlohou a boli predložené v rámci SR do APVT/APVV, VEGA a pod., (uviesť názov projektu, časové rámce riešenia, zdroje financovania podľa poskytovateľa a ich objem a jeho ďalšie identifikačné údaje a to, či bol vybraný na financovanie),

Morfosyntaktická analýza Slovenského národného korpusu. Grant VEGA č. 1/1432/04. Zodpovedná riešiteľka doc. PhDr. Miloslava Sokolová, CSc., mim. prof., FF PU Prešov. Doba riešenia 2004 – 2006.

Primárnou materiálovou bázou tohto projektu je Slovenský národný korpus. Viacerí riešitelia úlohy VaV sú, resp. boli spoluriešiteľmi aj tohto projektu, ktorého cieľom je príprava gramatiky na korpusovom materiáli. Recipročne tento projekt pomáhal pri precizovaní vnútornej štruktúry korpusu a viacerí študenti a doktorandi sa priamo podieľali na ručnej morfologickej anotácii vybraných textov korpusu.

Budovanie paralelných korpusov (slovensko-chorvátsky a slovensko-ruský korpus). Grant VEGA č. 2/5035/25. Zodpovedný riešiteľ RNDr. Radovan Garabík. Doba riešenia 2005 – 2007.

Projekt sa rieši v kolektíve SNK JÚĽŠ SAV v spolupráci s Katedrou matematickej lingvistiky Filologickej fakulty Sankt-Petersburgskej štátnej univerzity. Ide o tvorbu špecializovaných nástrojov na spracovanie textov dvoch jazykov a na vyhľadávanie v nich formou paralelných konkordancií (v rámci odseku alebo vety). Využívajú sa texty získané v procese riešenia úlohy štátneho programu, recipročne sa využívajú skúsenosti zahraničného partnera a lingvistické anotácie.

Explanačno-deskriptívna morfológia súčasnej slovenčiny. Grant VEGA č. 2/6119/6. Zodpovedný riešiteľ prof. PhDr. Juraj Dolník, DrSc., JÚĽŠ SAV. Doba riešenia 2006 – 2008.

Primárnou materiálovou bázou tohto projektu je Slovenský národný korpus, najmä jeho ručne morfologicky anotovaný podkorpus, ale aj výsledky automatizovanej morfologickej anotácie, ktorú budú môcť riešitelia grantu svojimi výskumami zdokonaľovať.

Slovník súčasného slovenského jazyka. Projekt zmluvne podporovaný Ministerstvom školstva SR, Ministerstvom kultúry SR a SAV. Zodpovedná riešiteľka Mgr. Alexandra Jarošová, CSc., JÚĽŠ SAV. Doba riešenia 2005 – 2012.
Primárnou materiálovou bázou tohto projektu je Slovenský národný korpus, bez ktorého by koncipovanie veľkého výkladového slovníka súčasnej slovenčiny nebolo možné.

Do Agentúry na podporu výskumu a vývoja bol podaný v r. 2005 a opakovane aj v r. 2006 projekt **Slovenská terminologická databáza (primárne v oblasti ekonómie, práva a techniky)**. Projekt nebol prijatý na financovanie.

- uviesť ďalšie konkrétne formy medzinárodnej spolupráce v rámci riešenia úlohy,

V r. 2004 – 2005 bol v spolupráci s Matematicko-fyzikální fakultou Karlovej univerzity v Prahe riešený Projekt slovensko-českej spolupráce na základe Dohody medzi vládou SR a ČR o VTS č. 130 **Využitie spoločných vlastností češtiny a slovenčiny na budovanie anotovaných národných jazykových korpusov**. V rámci neho sa realizovalo testovanie a aplikácia viacerých softvérov a metód na počítačové spracovanie prirodzeného jazyka, uskutočnilo sa viacero workshopov (v Bratislave a v Prešove), počas ktorých sa školili anotátori na ručnú syntaktickú anotáciu, a pracovné pobyty v Prahe, počas ktorých sa konzultovali jednotlivé postupy a realizovali sa konverzie vstupných dát z morfologickej aj syntaktickej anotácie.

V spolupráci s Katedrou matematickej lingvistiky Filologickej fakulty Sankt-Petersburgskej štátnej univerzity sa buduje paralelný rusko-slovenský a slovensko-ruský korpus.

- stručný popis spôsobu zabezpečenia publicity/popularizácie výsledkov riešenej úlohy, prípadne aké popularizačné aktivity sú naplánované, uviesť prípadnú internetovú stránku úlohy, ak bola vytvorená (rozsah maximálne 100 slov)

Slovenský národný korpus je so všetkými špecifickými súčasťami prístupný na stránke http://korpus.juls.savba.sk.

Členovia riešiteľského kolektívu sa pravidelne zúčastňujú relevantných domácich i zahraničných vedeckých podujatí, na ktorých prezentujú dosiahnuté výsledky a publikujú ich v príslušných zborníkoch materiálov. Publikačné výstupy, ako aj interné materiály sú dostupné na stránke SNK.

Osobitne sa prezentovali možnosti využitia korpusu v lingvistike, pedagogickej i prekladateľskej praxi na viacerých podujatiach na univerzitách v Prešove, Ružomberku, Nitre, Trnave, Bratislave, Prahe a na ďalších špeciálnych podujatiach (Studia Academica Slovaca, seminár Európske záležitosti a otázky jazykovej kultúry a terminológie pre prekladateľov zo štátnej správy, doktorandské semináre, semináre v spoločnostiach prekladateľov). Na pôde SNK bolo viacero prezentácií pre žiakov základných a stredných škôl, ako aj pre zahraničných lektorov slovenského jazyka. O SNK sa informovalo v printových (Sme, .týždeň) aj v iných médiách (televízia, rozhlas, InZine).

V r. 2006 bol pripravený CD/DVD nosič so základnou verziou korpusu predovšetkým na výučbové využitie.

- iné údaje dokumentujúce pridanú hodnotu úlohy

Kolektívu Slovenského národného korpusu Jazykovedného ústavu Ľ. Štúra ako pracovníkom Slovenskej akadémie vied bola v r. 2005 udelená ***Cena SAV za budovanie infraštruktúry pre vedu***.

9. Špecifikácia očakávaných spoločensko-ekonomických kvantitatívnych a kvalitatívnych dopadov ukončenej úlohy štátneho programu výskumu a vývoja

(Príloha č. 7 podľa Usmernenia Ministerstva školstva Slovenskej republiky pre oponentúry úloh výskumu a vývoja štátnych programov výskumu a vývoja)

Názov štátneho programu V a V: Aktuálne otázky rozvoja spoločnosti

Názov úlohy výskumu a vývoja: Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu

P.č.	Názov prínosu	S*	N*
	Zvýšenie vzdelanosti obyvateľov	X	

Okrem možnosti získať priamo z materiálu Slovenského národného korpusu informácie o systémových vlastnostiach slovenského jazyka sú prostredníctvom siete internet v položke Lingvistické zdroje verejne prístupné viaceré časopisecké, zborníkové i monografické publikácie a slovníky aj starších vydaní, ktoré inak už v podstate nie sú dostupné. Lingvistické zdroje sa využívajú vo veľkom rozsahu a osobitne ich oceňujú v školskej praxi na Slovensku i v zahraničí.			
	Zvýšenie úspešnosti zapájania sa do európskych programov výskumu a vývoja	X	
Vybudovanie Slovenského národného korpusu vrátane jeho základnej lingvistickej anotácie a špecifických súčastí (najmä paralelné korpusy) umožňuje zapojenie sa do európskych programov výskumu a vývoja (napr. Multext-East, WordNet), ako aj výraznejšie zapojenie slovenčiny do komparatívnych výskumov. Osobitným prínosom bude tvorba Slovenskej terminologickej databázy, ktorá vo svojej štruktúre počíta aj s cudzojazyčnými ekvivalentmi.			
	Zníženie zastaranosti techniky v infraštruktúre výskumu a vývoja	X	
Na vybudovanie a prevádzkovanie Slovenského národného korpusu bolo nevyhnutné zabezpečiť dostatočné hardvérové i softvérové vybavenie, ktoré je potrebné neustále inovovať v súlade s aktuálnymi trendmi v rozvoji informačných a jazykových technológií.			
	Zvýšenie integrácie výskumu a vývoja na národnej úrovni	X	
Lingvistické výskumy na Slovensku sa pred vybudovaním Slovenského národného korpusu uskutočňovali na individuálne excerpovanom materiáli, v Jazykovednom ústave Ľ. Štúra SAV bolo na požiadanie možné nahliadnuť do kartotéky. V súčasnosti má každý záujemca o výskum slovenského jazyka na Slovensku i v zahraničí k dispozícii neporovnateľne väčší objem materiálu, materiálu, ktorý je rovnaký, voľne prístupný, na ktorom sa jednotlivé výskumy dajú opakovať a verifikovať tak zistené výsledky.			
	Zvýšenie kvality výskumu a vývoja a podpora excelentnosti	X	
Slovenská jazykoveda sa na základe skúmania jazyka na materiáli Slovenského národného korpusu presúva z kategórie vedy postavenej najmä na individuálnej empírii a introspekcii do kategórie exaktných vied. Mnohé výskumy môžu prebiehať efektívnejšie a s väčšou výpovednou hodnotou. Počítačové spracovanie slovenčiny ako prirodzeného jazyka by mohlo ašpirovať na zaradenie medzi excelentné výskumy.			

10. Návrh zabezpečenia patentovo-právnej ochrany výsledkov riešenia úlohy

Napriek viacerým inovatívnym riešeniam a výstupom celkovo výsledky riešenia úlohy nespadajú pod patentovú ochranu.

Právna ochrana výsledkov riešenia úlohy je zabezpečená autorským zákonom – výstupy z korpusu sú chápané ako citácie podľa § 25 zákona č. 618/2003 Z. z. Prípadné ďalšie výstupy nad rámec citácií získaných z korpusu sa riešia individuálne podľa konkrétnych požiadaviek

používateľa a nakladania s dátami. Obvykle sú poskytované údaje, na ktoré je dispozičné právo, s podmienkou ďalšieho použitia výlučne na nekomerčné vedecko-výskumné a vzdelávacie ciele.

11. Návrh na vysporiadanie zostatkových materiálov a predmetov

Na financovanie riešenia úlohy štátneho programu neposkytol objednávateľ žiadne kapitálové prostriedky, ich potrebu napĺňala SAV zo svojej rozpočtovej kapitoly. Finančné prostriedky poskytnuté v kategórii tovary a služby boli zväčša vynaložené na dohody o vykonaní práce pre externých spolupracovníkov. Výpočtová technika a ďalšie časti zariadenia zakúpené v rámci bežných prostriedkov predovšetkým v prvých dvoch rokoch riešenia úlohy sú predmetmi krátkodobej spotreby. Vzhľadom na pokračovanie riešiteľského kolektívu v rozpracovaných úlohách na tom istom pracovisku je možné využitie zostatkových materiálov a predmetov samotným kolektívom.

12. Časový harmonogram realizácie výsledkov úlohy prostredníctvom realizátorov na nasledujúce tri roky

Okrem kolektívu Oddelenia súčasnej lexikológie a lexikografie JÚĽŠ SAV, ktorý koncipuje nový výkladový slovník slovenského jazyka predovšetkým na materiáli Slovenského národného korpusu (projekt má stanovenú dobu riešenia do r. 2013), a stálych registrovaných používateľov SNK (okolo 200 ročne) sa realizátorom výsledkov úlohy stáva aj samotný riešiteľský kolektív. Vzhľadom na potrebu ďalšieho fungovania SNK ako pracoviska základného výskumu a potrebu systematického budovania vedného odboru, ktorý doteraz na Slovensku chýbal a ktorý zapája Slovensko do medzinárodného kontextu vývoja jazykových a informačných technológií, bola v júni 2006 podpísaná zmluva o spolupráci medzi Ministerstvom kultúry SR, Ministerstvom školstva SR a Slovenskou akadémiou vied, ktorej predmetom je spolupráca zmluvných strán s cieľom dobudovania SNK a vytvorenia špecifických diel využívajúcich zhromaždené dáta podľa dohodnutej koncepcie. Na nasledujúce tri roky je naplánované využitie výsledkov úlohy v tomto rozsahu:

Rok 2007

Dokončenie ručnej lingvistickej anotácie: ručne morfologicky anotovaný korpus do rozsahu 1,2 mil. textových jednotiek, syntakticky anotovaný korpus do rozsahu 50 tis. jednotlivých viet. Vývoj príslušných nástrojov na automatizované spracúvanie slovenčiny ako prirodzeného

jazyka (analyzátor a generátor tvarov slovenského jazyka, morfologický lematizátor, anotátor a dezambiguátor).

Rok 2008

Vybudovanie paralelného korpusu slovensko-českého a česko-slovenského v rozsahu min. 5 mil. textových jednotiek pre potreby koncipovania prekladového slovníka a porovnávacích výskumov. Vytvorenie príslušných nástrojov na zarovnávanie textov, automatizované označkovanie oboch textov a ich efektívne využívanie.

Vytvorenie slovenského elektronického slovníka WordNet a integrácia do projektu EuroWordNet (sémantický slovník viacerých európskych jazykov).

Rok 2009

Vytvorenie Frekvenčného slovníka súčasnej slovenčiny a Retrográdneho slovníka súčasnej slovenčiny na báze korpusu písaných textov.

Prvá verzia Slovenskej terminologickej databázy.

13. Závery a odporúčania vyplývajúce z riešenia úlohy

Riešiteľský kolektív splnil úlohu vytvoriť národný korpus slovenského jazyka v dostatočne širokom rozsahu a zábere, lingvisticky anotovaný a vybavený výkonnými vyhľadávacími nástrojmi, nástrojmi na indexáciu a kategorizáciu textových dokumentov, nástrojmi na štatistickú analýzu (rôzne štatistické údaje sú priamo dostupné na stránke korpusu) a obsahujúci aj špecifické súčasti (ručne morfologicky a syntakticky anotované podkorpusy, paralelné korpusy, databázu lexikografických diel a lingvistických zdrojov). Korpus poskytuje bohatý výskumný a výučbový materiál pre oblasť lingvistiky a počítačového spracovania prirodzeného jazyka – každý rok je okolo 200 registrovaných používateľov z rôznych vedecko-výskumných domácich i zahraničných pracovísk, bežná návštevnosť najmä lingvistických zdrojov dosahuje priemerne už približne 10 000 prístupov denne. Korpus slúži ako referenčná materiálová báza na tvorbu nového 8-zväzkového slovníka súčasnej slovenčiny a na aktualizáciu praktických jazykových príručiek. V rámci grantov VEGA sa na ňom realizovali a realizujú viaceré výskumy s výstupmi v oblasti gramatiky a lexikológie slovenského jazyka. Cenným prínosom bolo aj vytvorenie širšieho okruhu spolupracovníkov, najmä spomedzi študentov a absolventov slovenského jazyka z viacerých vysokých škôl na Slovensku. Niektoré výsledky sa však mohli dosiahnuť efektívnejšie a mohli byť kvalitnejšie i výraznejšie zapojené do medzinárodného kontextu v prípade posilnenia základného kolektívu aspoň o dvoch stálych zamestnancov.

Napriek istým obmedzeniam a problémom s priebežným financovaním riešenia úlohy štátneho programu výskumu a vývoja ***Komplexné spracovanie slovenského jazyka a jeho***

elektronizácia na účely jazykovedného výskumu vytvorili pracovníci oddelenia Slovenského národného korpusu dobrý základ na systematický rozvoj odboru počítačová a korpusová lingvistika na Slovensku a na počítačové spracovanie slovenčiny ako prirodzeného jazyka. Príprava viacerých zdrojov verejne prístupných prostredníctvom počítačovej siete internet sa stala prvou fázou elektronizácie jazykovedného výskumu slovenčiny, ktorej výsledky sú dostupné nielen pre slovenské slovakistické pracoviská, ale aj pre zahraničných bádateľov a záujemcov o slovenský jazyk. Vďaka pochopeniu a ústretovosti vedenia Ministerstva kultúry SR v súvise so starostlivosťou o štátny jazyk, Ministerstva školstva SR a Slovenskej akadémie vied bude možné ďalšie pokračovanie budovania Slovenského národného korpusu a využívanie dosiahnutých výsledkov aj priamo doterajším riešiteľským kolektívom. Na vytvorenie väčšej stability a kontinuity tohto druhu výskumu a jeho pevné zaradenie do medzinárodného kontextu by bolo vhodné inštitucionalizovať samostatný odbor počítačovej a korpusovej lingvistiky na niektorej z vysokých škôl na Slovensku.

14. Záverečná karta úlohy výskumu a vývoja

Dodávateľ: Jazykovedný ústav Ľ. Štúra SAV Bratislava	**Číslo zmluvy: 346/2003**
	Číslo úlohy výskumu a vývoja: 2003SP200280307
Názov úlohy výskumu a vývoja: Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu	
Názov štátneho programu / podprogramu výskumu a vývoja: Aktuálne otázky rozvoja spoločnosti	
Zodpovedný riešiteľ: PhDr. Mária Šimková	
Zoznam pracovníkov kolektívu hlavného riešiteľa, ktorí sa podieľali na riešení úlohy (meno, priezvisko, tituly, vek, rozsah práce v človekorokoch) Pavol Domin, RNDr., PhD., 1974, 0,5 Martina Forróová, Mgr., 1977, 0,75 Radovan Garabík, RNDr., 1974, 4 Lucia Gianitsová, Mgr., 1977, 1,5 Rudolf Hladík, 1955, 5 Alexander Horák, Mgr., 1978, 2,25 Kristína Horná, 1979, 2,5 Helena Ivoríková, Mgr., 1978, 6,25 Agáta Karčová, Mgr., 1982, 1,5 Jana Levická, Mgr., 1976, 2,25 Daniela Majchráková, Mgr., 1980, 1,25 Peter Rybár, RNDr., 1974, 1 Mária Šimková, PhDr., 1963, 5 Martin Šmotlák, RNDr., PhD., 1976, 2 Tomáš Vančo, 1979, 2,5 Dorota Vasilišinová, Mgr., 1979, 1,25	
Objednávateľ: Ministerstvo školstva SR	
Doba riešenia úlohy výskumu a vývoja: 1. 7. 2003 – 31. 12. 2006	
Skutočný rozpočet projektu v Sk v členení verejné zdroje/iné zdroje – pri iných uviesť presný zdroj: Plánovaný rozpočet z prostriedkov MŠ SR bol 10 318 tis. Sk. V roku 2003 sa 193 tis. Sk vrátilo do ŠR. Skutočne minuté prostriedky z rozpočtu MŠ SR za celú dobu trvania projektu boli 10 125 tis. Sk. Z rozpočtovej kapitoly SAV bolo na riešenie úlohy poskytnutých 9 492 tis. Sk.	

Publikácie (knihy, články, prednášky a pod.) zahrňujúce výsledky úlohy výskumu a vývoja: ***Uveďte maximálne päť najvýznamnejších publikácií***	Computer Treatment of Slavic and East European Languages. Zborník z medzinárodnej vedeckej konferencie Slovko 2005. Red. R. Garabík. Bratislava: Veda 2005. 248 s.
	HORÁK, A. – GIANITSOVÁ, L. – ŠIMKOVÁ, M. – ŠMOTLÁK, M. – GARABÍK, R.: Slovak National Corpus. In: Text, Speech and Dialogue. 7th International Conference TSD 2004 Proceedings. Ed. P. Sojka – I. Kopeček – K. Pala. Berlin – Heidelberg: Springer – Verlag 2004, s. 89 – 94.
	GARABÍK, Radovan: Corpus Construction Tools. In: Труды международной конференции MegaLing'2005. Прикладная лингвистика в поиске новых путей. Zborník z medzinárodnej vedeckej konferencie MegaLing'05. Meganom, Krym, Ukrajina 27. júna – 2. júla 2005. Red. V. P. Zacharov – S. S. Dikareva. St. Petersburg: Izdateľstvo Osipov 2005, s. 26 – 32.
	ŠIMKOVÁ, Mária: Možnosti využitia Slovenského národného korpusu na štúdium slovenského jazyka. In: Studia Academica Slovaca 33. Prednášky z XL. letnej školy slovenského jazyka a kultúry. Ed.: Jozef Mlacek – Miloslav Vojtech. Bratislava: Filozofická fakulta Univerzity Komenského 2004, s. 204 – 218.
	MAJCHRÁKOVÁ, Daniela – KARČOVÁ, Agáta: Distribúcia prípon *-mi, -ami* v tvaroch maskulínnych substantív zakončených na konsonantické skupiny. In: Sondy do morfosyntaktického výskumu slovenčiny na korpusovom materiáli. Ed. M. Sokolová – M. Ivanová. Prešov: Filozofická fakulta Prešovskej univerzity 2006.
Výsledky záverečnej oponentúry:	
V čom vidíte uplatnenie výsledkov tejto úlohy výskumu a vývoja v spoločenskej/hospodárskej praxi/hlavný spoločensko-hospodársky prínos výsledkov úlohy (max. 100 slov):	Výsledky riešenia úlohy prispievajú k zexaktneniu lingvistických výskumov a sprístupňovaniu ich výsledkov širokej verejnosti. V rámci počítačového spracovania slovenčiny ako prirodzeného jazyka boli vyvinuté viaceré nástroje na spracúvanie textov veľkého rozsahu a efektívnu prácu s nimi. Ukončená fáza budovania korpusu písaných textov slovenského jazyka predstavuje dobrý základ pre ďalší rozvoj počítačovej a korpusovej lingvistiky na Slovensku. Bola vytvorená solídna materiálová báza pre potreby tvorby lexikografických a gramatických príručiek, ako aj na výučbové využitie a komparatívne výskumy.

Súhrn výsledkov riešenia úlohy výskumu a vývoja a naplnenia cieľov úlohy výskumu a vývoja v slovenskom jazyku netechnickým/laickým spôsobom (max. 300 slov)[1]:

Riešením úlohy Komplexné spracovanie slovenského jazyka a jeho elektronizácia na účely jazykovedného výskumu bol vytvorený Slovenský národný korpus ako elektronická databáza písaných textov slovenského jazyka z obdobia rokov 1955 – 2005. Základná, všeobecná verzia Slovenského národného korpusu obsahuje 350 miliónov textových jednotiek (slová, ale aj interpunkcia a iné neslovné jednotky) z rôznych štýlov a žánrov, vydavateľstiev, regiónov a pod. Samotná textová databáza slúži najmä na tvorbu slovníkov, lingvistické informácie v ňom obsiahnuté aj na tvorbu gramatických príručiek. Bežní záujemcovia o slovenský jazyk môžu v korpuse zisťovať reálne fungovanie jednotlivých jazykových prostriedkov a javov. Korpus je veľmi dobre využiteľný v školskej praxi na tvorbu cvičení pre slovenských i zahraničných študentov. Špecifické súčasti korpusu – databáza lexikografických príručiek a ďalšie lingvistické zdroje – sprístupňujú širokej verejnosti aktuálne výsledky jazykovedných výskumov. Paralelné korpusy umožňujú porovnávať systém slovenského jazyka so systémami iných jazykov. Vyvinuté počítačové nástroje umožňujú prácu s reálnym textovým materiálom veľkého rozsahu pri jeho spracúvaní i vyhľadávaní v ňom.

[1] Je potrebné uvedené napísať takým spôsobom, ktorý umožni text zverejniť na internetovej stránke gestorom štátnych programov (MŠ SR) – t.j. neuvádzať údaje, ktoré sú predmetom obchodného tajomstva a pod.

Súhrn výsledkov riešenia úlohy výskumu a vývoja a naplnenia cieľov úlohy výskumu a vývoja v anglickom jazyku netechnickým/laickým spôsobom (max. 300 slov)[2]:

Realisation of the project of the ***Integrated Computational Processing of the Slovak Language for Linguistic Research Purposes*** resulted in creation of the Slovak National Corpus, which is an electronic database of texts written in 1955 – 2005. Primary, general version of the Slovak National Corpus contains 350 million tokens (i.e. words but also punctuation and other non-word items) from different styles and genres, publishing houses, regions and so on. The textual database itself serves especially for compilation of dictionaries, linguistic information and also for creation of grammatical manuals. People interested in the Slovak language can use the Corpus in order to find out the actual functioning of different linguistic means and phenomena. The Corpus can become a very useful tool for the didactic purposes - e.g. creation of exercises for Slovak and also for foreign pupils and students. Specific parts of the Corpus – database of the lexicographical manuals and other linguistic resources – enable the general public to access the latest results of linguistic research. Parallel corpora offer an opportunity to compare the system of the Slovak language with systems of other languages. Developed software tools enable to work - process and retrieve within a genuine text material of a large extent.

Podpisom záverečnej karty úlohy výskumu a vývoja zodpovedný riešiteľ vyjadruje súhlas na zverejnenie v nej uvedených údajov.

V..... Dátum:..... Podpis zodpovedného riešiteľa:.....

[2] Je potrebné uvedené napísať takým spôsobom, ktorý umožni text zverejniť na internetovej stránke gestorom štátnych programov (MŠ SR) – t.j. neuvádzať údaje, ktoré sú predmetom obchodného tajomstva a pod.