Slovenski nacionalni korpus idejni osnutek projekta

Primož Jakopin

JEZIKOSLOVNI ZAPISKI 7 • 2001 • 1–2

1. UVOD

Vsenarodni korpus besedil v slovenskem jeziku je naloga, ki čaka na izvedbo že skoraj deset let (možna angleška prevoda sta *Slovenian National Corpus* in *Bank of Slovenian*). Taki korpusi, ki obsegajo kar najbolj popoln izbor pisanih, pa tudi govorjenih besedil, spadajo že nekaj časa med osnovna izobraževalna in raziskovalna orodja nekega jezika (npr. Sinclair 1992). Uporabni so

- pri gradnji slovarjev: osnovnega slovarja jezika, terminoloških slovarjev, frazeološkega slovarja, slovarja sinonimov, lematizacijskega slovarja, dvojezičnih slovarjev
- za spremljanje stanja v jeziku
- pri ugotavljanje neologizmov
- za preverjanje jezikoslovnih hipotez
- pri izdelavi orodij za strojno prevajanje
- v literarni teoriji
- pri pouku jezika
- pri razvedrilu (sestavljanje in reševanje križank, ugibanje besed)
- za enciklopedično rabo, kot vir podatkov o vseh področjih dejavnosti nekega naroda.

Potem ko so bili besedilni korpusi v začetku devetdesetih let zgrajeni v tehnološko naprednejših, predvsem zahodnih okoljih, npr. Britanski nacionalni korpus (100 milijonov besed, od tega 2 milijona s preverjenimi oblikoslovnimi oznakami) so jih v drugi polovici prejšnjega desetletja dobili tudi manjši narodi iz naše jezikovne skupine. Primera sta Češki nacionalni korpus (100 milijonov besed) in Hrvaški nacionalni korpus (30 milijonov besed).

2. SEDANJE STANJE

V tem času so pri nas nastale 3 večje spletne besedilne zbirke (Jakopin 2000) od katerih sta dve, Korpus FIDA (Gorjanc 1999) in BESEDA (Jakopin 2000a) organizirani na način tujih besedilnih korpusov (začetek delovanja obeh v letu 1999); tretja, Zbirka slovenskih literarnih besedil Mirana Hladnika (Hladnik 1995) pa je bila na splet postavljena že nekaj let prej.

2.1 FIDA

Korpus FIDA obsega 100 milijonov besed v glavnem časopisnega jezika zadnjih nekaj let, namenjen je predvsem za interno rabo glavnih ustanoviteljic (priprava slovarjev založbe DZS, koordinatorja projekta, izdelava črkovalnikov podjetja AMEBIS, uporaba pri pouku slovenske slovnice na Filozofski fakulteti v Ljubljani) in ni javno dostopen. Korpus FIDA je bil v celoti financiran s strani obeh komercialnih partnerjev (DZS in AMEBIS).

2.2 BESEDA

Korpus BESEDA oziroma Nova BESEDA deluje v okviru Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (ISJ), trenutno je v njem 50 milijonov besed časopisnega (DELO) in leposlovnega jezika (48 avtorjev), ki so poleg konkordančnika na razpolago tudi v obliki frekvenčnih abecednega in odzadnjega slovarja besednih oblik. Besedila v njem so prečiščena in označena do ravni povedi, pribl. milijon besed pa ima preverjene oblikoslovne oznake, tako da je mogoče iskanje tudi po slovničnih kategorijah in izrazih. Korpus vsebuje med drugim celoten opus Cirila Kosmača, v pripravi pa je razširitev na 75 milijonov besed in dopolnitev z opusom Ivana Cankarja. Korpus ni namenjen samo leksikografskemu delu na Inštitutu ampak tudi za najširšo rabo (npr. Grzybek 2000), je javno dostopen in v slabih dveh letih je doživel preko 18.000 obiskov. BESEDA je bila postavljena z zelo skromnimi davkoplačevalskimi sredstvi predvsem iz rednega dela v okviru ISJ.

2.3 Projekt TRUBAR

Slovenska Narodna in univerzitetna knjižnica (NUK) zbira in hrani, kot obvezni izvod, vso tiskano produkcijo v našem jeziku, ki vsebuje tudi publikacije, izdane v elektronski obliki (npr. Slovar slovenskega knjižnega jezika na CD ROM-u). Sredi devetdesetih let je bila v NUKu nekaj časa predmet razprave ideja, na kratko imenovana Projekt TRUBAR, o tem, da bi poleg izvodov publikacij v papirni obliki zbirali in hranili tudi elektronske predloge, iz katerih ta dela nastanejo. Zadeva se je nekako ustavila pri spremembah Zakona o obveznem izvodu, ki bi bile potrebne za dosego tega cilja.

3. PREDLOG IZVEDBE

3.1 Obseg

Če je prvi besedilni korpus, Brownov korpus (Kučera 1967) obsegal le milijon besed in je bil sestavljen iz 500 odlomkov besedil po 2000 besed, so korpusi devetdesetih let praviloma stokrat večji, še vedno pa sestavljeni iz odlomkov. Korpusi, ki so trenutno v pripravi, primer je Ameriški nacionalni korpus (Macleod idr. 2000), segajo znatno čez te meje, se pa še vedno ne upajo odločneje približati idealu. Idealen korpus nekega jezika bi namreč najprej moral vsebovati vse publikacije, objavljene v tem jeziku, potem pa še vse ostalo, osebna sporočila, govorjene vire in drugo.

Zmožnosti sodobnih računalnikov postajajo tako velike, da bi bilo tehnično že zdaj povsem izvedljivo, elektronske kopije tekoče slovenske tiskane produkcije

sproti vključevati v korpus. Letni prirast slovenskih knjižnih del je pribl. 4000 enot (velikostnega reda 50 milijonov besed), periodičnih publikacij, časopisov in revij pa približno za petkrat več (velikostni red 250 milijonov besed, v časopisu DELO je letno npr. objavljenih okoli 20 milijonov besed). Skupaj velikostnega reda 300 milijonov besed letno, kar je sicer veliko, a še vedno le desetina diskovne kapacitete povprečnega danes prodanega namiznega računalnika (2 GB proti 20 GB).

Tako zastavljen Slovenski nacionalni korpus bi imel izreden splošnouporaben enciklopedični pomen, ki ga sedanji svetovni korpusi nimajo.

3.2 NUK in ISJ

Drugo, predvsem organizacijsko vprašanje je seveda, kako vse to zbrati, obdelati, shraniti in dati na uporabo za proučevanje in poučevanje slovenskega jezika v obsegu in na način, ki ne bi krnila avtorskih pravic njihovih lastnikov. Kot naravna partnerja v ta kontekst sodita Narodna in univerzitetna knjižnica ter Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Prva kot matična knjižnica za zbiranje in hranjenje vsega slovenskega besedilnega gradiva, drugi kot osrednja znanstvena ustanova za pripravo slovarskih in drugih jezikovnih virov v slovenskem jeziku, z že delujočim spletnim besedilnim korpusom.

Partnerja, tudi če vzamemo ZRC SAZU širše, sta prostorsko v neposredni bližini, raziskovalci obeh se poznajo in srečujejo ob mnogih priložnostih, med obema hišama, ki sta v zadnjih letih napravili pomembne korake pri uporabi računalnikov in posodobitvi ustrezne opreme, pa je tudi zelo hitra internetna povezava. Sinergijski učinek podjetja, kakršno je Slovenski nacionalni korpus, bi bil za obe ustanovi nedvomno velik.

3.3 Postavitev korpusa

Organizacijske in vsebinske priprave na postavitev Slovenskega nacionalnega korpusa bi bilo mogoče opraviti v nekaj letih - podrobnejši časovni okvir je podan v posebnem razdelku. Najprej bi bilo potrebno skleniti dogovor o sodelovanju pri projektu med obema partnerjema, NUK in ISJ (oz. ZRC SAZU), ki bi opredelil tudi sodelujoče delavce obeh strani, udeleženo programsko in strojno opremo. V tem obdobju bi začeli z obveščanjem glavnih potencialnih donatorjev besedil o projektu in nadgrajevali obstoječi korpus (ISJ) z besedili, dobljenimi na ta način kot prostovoljni prispevek.

V tem času bi se poskusili na Ministrstvu za šolstvo, znanost in šport (MŠZŠ) in Ministrstvu za informacijsko družbo (MID) dogovoriti o načinu vključitve projekta v ustaljene tokove financiranja, ter kako pospešiti sprejem zakona o elektronski kopiji obveznega izvoda. Obenem bi za začetek skušali tudi izposlovati dogovor o obvezni donaciji vseh besedil (npr. učbenikov), sofinanciranih s strani MŠZŠ v korpus. Dograjevali in izpopolnjevali bi programsko opremo za uporabo in delovanje korpusa.

Vzporedno bi ustanovili konzorcij sponzorjev in skušali zanj pridobiti najpomembnejše založniške in časopisne hiše ter najpomembnejša podjetja s področja prodaje računalnikov in programske opreme (npr. IBM Slovenija, Microsoft Slovenija) ter izdelave programske opreme (npr. Hermes Softlab).

Ko bi bili vsi trije pogoji (financiranje s strani MŠZŠ, Zakon o elektronskem

obveznem izvodu, konzorcij sponzorjev) izpolnjeni, bi ustanovili organizacijsko enoto Slovenski nacionalni korpus in ta bi začel normalno poslovati.

3.4 Naslovne domene

Da bi bil zagotovljen kar najenostavnejši dostop do posameznih delov korpusa in njihovo kombiniranje, obenem pa omogočeno tudi oblikovanje podkorpusov za posebne namene, ki bi se jih dalo pri iskanju obravnavati ločeno, bi bila vsa v korpus vključena dela opremljena s kratkimi govorečimi oznakami.

Primeri so npr. oznaka D_99X30, kjer pomeni D časopis DELO, 99X30 pa oznako za izdajo dne 30. oktobra 1999 ali SL_KC_ kjer pomeni SL slovensko leposlovje, KC pa oznako za vsa dela Cirila Kosmača ali I_ kot oznako del v korpusu, namenjenih potrebam sekcij Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (naprej razčlenjenih v I_L_, I_E_, I_D_, I_T_ in I_H_).

3.5 Drugi udeleženci projekta

Že kmalu po začetni fazi projekta bi bilo treba k sodelovanju pritegniti še nekaj drugih ustanov, tako glede uporabe korpusa kot reševanja strokovnih vprašanj. Tu je najprej mišljena osrednja visokošolska izobraževalna ustanova za slovenski jezik in književnost, potem pa, predvsem glede govorjenega dela korpusa in s tem povezanih nalog, nosilci razvoja z govorom povezanih tehnologij.

Umestitev korpusa v mednarodni prostor terja navezavo stikov in sodelovanje s sorodnimi, predvsem evropskimi projekti in jezikovnimi viri, kakršni so Češki nacionalni korpus, Ameriški nacionalni korpus, Projekt vseeevropskih jezikovnih virov (npr. TELRI) in ustrezni raziskovalni arhiv, primer je Tractor.

Za uspešno delovanje in širitev korpusa bi bil potreben velik finančni napor, ki bi terjal vključitev konzorcija sponzorjev (zgled je Ameriški nacionalni korpus). Tu so mišljeni tako velike založniške in časopisne hiše, ki bi z zgodnjo vključitvijo v projekt zelo povečale svojo prepoznavnost in vidnost svojih del (šolska primera sta npr. Enciklopedija Slovenije založbe Mladinska knjiga in DELO, osrednji slovenski dnevnik) kot tudi drugi sponzorji zunaj založniškega in časopisnega okolja, ki bi z vključitvijo v projekt utrdili svoje mesto v slovenskem prostoru.

3.6 Vprašanje avtorskih pravic

Besedila, shranjena v korpusu, ne bi bila spletno dostopna niti v izvirni obliki niti v celoti. Uporabljena bi bila le za izdelavo kumulativnih statističnih kazalcev jezika, kakršni so recimo seznami besed ali besednih zvez, in v konkordančnih seznamih, a tam le v obliki ožjega citata, se pravi ne več kot treh povedi - tekoče povedi, povedi pred njo in povedi za njo.

Korist za lastnika avtorskih pravic danega besedila bi bila v tem, da bi uporabnik v konkordančnem seznamu vsake poizvedbe po korpusu imel, za vsak najdeni primer posebej, kazalec na vir, iz katerega je bil črpana pojavitev iskanega jezikovnega ali informacijskega izraza (npr. *pridevnik pridevnik pridevnik samostalnik* ali *na licu mesta* ali *sonaravna uporaba gozda*) in njegovo ožje sobesedilo. Tako bi bil usmerjen naprej v tiskani izvod publikacije, ga morda tudi kupil, ali pa se, v primeru da bi lastnik avtorskih pravic besedilo ponujal na spletu v elektronski obliki, preko kazalca lahko preselil na njegovo spletno stran s plačljivim celotnim delom.

4. OKVIRNI TERMINSKI NAČRT

1. **2002**
- sporazum med ZRC SAZU in NUK o izvedbi projekta
- izbor sodelavcev
- izbor programske opreme
- postavitev zrcalne kopije obstoječega korpusa tudi na strežniku NUK
- prizadevanje za financiranje projekta s strani MŠZŠ in MID ter za sprejem ustreznih podpornih aktov na Ministrstvu za šolstvo, znanost in šport

2. **2003**
- iskanje sponzorjev
- kvalitetna dopolnitev korpusa z oblikoslovno označitvijo 5 milijonov besed s strani ISJ
- kvantitetna dopolnitev z novimi, na prostovoljni osnovi pridobljenimi besedili
- izpopolnjevanje in dopolnjevanje programske opreme
- seznanjanje medijev in javnosti o projektu

3. **2004**
- sprejem zakonodaje o obveznem elektronskem izvodu z ustreznim zavarovanjem avtorskih pravic
- odobritev financiranja s strani MŠZŠ in MID
- ustanovitev konzorcija sponzorjev
- ustanovitev organizacijske enote Slovenski nacionalni korpus

4. **2005**
- preizkus programske opreme z dnevnim dodajanjem vseh novih besedil
- postavitev hitrega strežnika s korpusom
- domača in mednarodna promocija korpusa
- začetek normalnega poslovanja Slovenskega nacionalnega korpusa.

5. SODELAVCI

V prvih fazah, na začetku projekta, bi sodelovali predvsem delavci iz vrst ISJ in NUK, kasneje pa bi za delovanje in širitev korpusa skrbela manjša, za določen čas (postavljen s strani glavnega financerja, praviloma 5 let) angažirana skupina, združena v posebni, s strani ZRC SAZU in NUK ustanovljeni organizacijski enoti:

1. vodja in koordinator, z znanji s področja računalniškega jezikoslovja, skrbi za implementacijo korpusa, sodeluje pri vseh opravilih v zvezi s projektom

2. tajnik in manager, skrbi za promocijo projekta, za stike z uporabniki, dobavitelji besedil, s financerji ter sponzorji, sodeluje pri pripravi besedil

3. oskrbnik podatkovnih zbirk in spletnih strani, z računalniškimi znanji in smislom za jezikoslovje, odlično obvlada angleški jezik

4. redaktorji, uredniki besedil, jezikoslovci slovenisti z znanjem tujih jezikov, z interesom za naravoslovje in računalništvo.

Vsi zgoraj navedeni so moškega spola, mišljena pa sta seveda enakovredno oba spola.

6. OPREMA

Sem spadata predvsem programska in strojna oprema, pri čemer je prva ključnejša in v marsičem pogojuje izbor druge.

6.1 Programska oprema

Pri izbiri ustrezne programske opreme za projekt se je potrebno najprej odločiti ali izbrati eno od že obstoječih rešitev ali iti v razvoj nove, lastne. Glede na to, da je za tako nalogo potreben obsežen softver, z velikimi stroški in dolgim časom, ki je potreben za razvoj in izpopolnitev, je očitno primernejša prva možnost, izbira že obstoječe programske opreme.

Naslednja odločitev se nanaša na izbor domačega ali tujega softvera; obe plati imata dobre in slabe strani. Tujo tovrstno opremo je v akademske, se pravi izobraževalne in raziskovalne nepridobitne namene mogoče dobiti brezplačno, z upoštevanjem ostalih pogojev, ki morajo biti izpolnjeni (npr. operacijski sistem strežnika). V naši širši okolici je najbolj uveljavljen sistem Corpus Workbench, vir je Institut für machinelle Sprachverarbeitung iz Stuttgarta, ki je bil med drugim uporabljen pri Češkem nacionalnem korpusu in pri sicer majhnem korpusu na Odseku za inteligentne sisteme Inštituta Jožef Stefan; sistem teče na operacijskih sistemih Solaris in Linux. Domači alternativi sta sistem ASP32, uporabljen pri FIDI (avtor podjetje Amebis) in NEVA, uporabljena pri BESEDI oz. Novi BESEDI (avtor pisec teh vrstic); oba sta napisana za operacijski sistem Windows NT oz. Windows 2000.

Glede na to, da je tujo rešitev zelo težko (oz. drago) spreminjati ali prilagajati lastnim potrebam in tudi glede na to, da sta oba domača sistema že uveljavljena in preizkušena z nekajletno uporabo, ter da je na obeh sodelujočih ustanovah, ISJ in NUK, večje število delavcev, ki poznajo programsko opremo za pripravo ustreznih podatkovnih zbirk, je nedvomno ustreznejša domača rešitev.

6.2 Strojna oprema

Strojna oprema, potrebna za izvedbo projekta, je v večji meri že pri obeh partnerjih. Potreben bi bil še en zmogljiv strežnik ter osebni računalniki novih sodelavcev z ustrezno dopolnitvijo mrežnih povezav.

7. ZAKLJUČEK

Podanih je bilo nekaj misli, kako postaviti Slovenski nacionalni korpus v obliki, skladni s potrebami širše skupnosti uporabnikov, raziskovalcev in učiteljev našega jezika, s sedanjim trenutkom, sodobnimi tehnološkimi možnostmi, o rešitvi, ki bi močno prispevala k utrditvi in okrepitvi slovenske jezikovne in narodne identitete in ki bi pomenila tudi izviren prispevek k reševanju teh problemov v širšem, svetovnem merilu.

Čas je dozorel in več znamenj kaže, da utegne do realizacije projekta, take ali drugačne, priti v letu ali dveh.

Viri in literatura

GORJANC, V. (1999), Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. *Zbornik predavanj / 35. seminar slovenskega jezika, literature in kulture.* Ljubljana, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete, 47–59.

GRZYBEK, P. (2000), Pogostnostna analiza besed iz elektronskega korpusa slovenskih besedil, *Slavistična revija 48/2*, 141–157.

HLADNIK, M. (1995), Elektronski literarnovedni viri in računalniško pisanje, *Jezik in slovstvo* 40, št. 7, 243–254.

JAKOPIN, P. (1999), Slovenian National Corpus from Fiction to Reality, Predavanje na 31. kongresu American Association for the Advancement of Slavic Studies, St. Louis, MO.

JAKOPIN, P. (2000), Slovenian texts on the internet, *Zapiski*, May 2000, 7, 4–7.

JAKOPIN, P. (2000a), BESEDA – a text corpus of Slovenian, *Digital resources for the humanities, conference abstracts*, University of Sheffield, 70–72.

KUČERA, H. & Winthrop, F. (1967), *Computational Analysis of Present-Day American English*, Brown University Press, Providence, RI.

MACLEOD, C., Ide, N., Grishman, R. (2000), The American National Corpus, Standardized Resources for American English, *Proceedings of the Second Language Resources and Evaluation Conference* (LREC), Atene, 831–836.

SINCLAIR, J. (1992), *Corpus, concordance, collocation*, Oxford University Press, Oxford.