KGB (korpus govorjenih besedil) v slovenščini

Marko Stabej,[1] Primož Vitez[2]

(1) Oddelek za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, Ljubljana
marko.stabej@guest.arnes.si
(2) Oddelek za romanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, Ljubljana
primoz.vitez@guest.arnes.si

Povzetek

Prispevek predstavlja grob osnutek korpusa govorjenih besedil (KGB) v slovenščini kot nujno potrebnega temeljnega jezikovnega vira v slovenskem prostoru. Oblikovanje KGB bo prispevalo k kvalitetnejšim analizam slovenskega govorjenega jezika, po drugi strani pa tudi k razvoju govornih jezikovnih tehnologij. V prvi fazi bo KGB zajel predvsem govor v medijih.

Abstract

The paper presents the draft of the spoken corpus of Slovene language (KGB). The spoken corpus is a very important part of the national linguistic infrastructure. KGB would provide the means to enhance the linguistic study of the Slovene spoken language and contribute to the further development of Slovene speech technologies. In the first phase the KGB would consist of balanced media speech.

1. Uvod

Oblikovanje korpusa govorjenih besedil v slovenščini je nujno potrebno zaradi več razlogov. Vse bolj prevladujoči empirični pogled na raziskovanje jezika in iz tega izhajajoče jeziko(slo)vne aplikacije terjajo obsežno in dobro oblikovano infrastrukturo jezikovnih virov. Taka infrastruktura mora biti zasnovana in oblikovana v sodelovanju med jezikoslovjem in raziskovalnim ter aplikativnim področjem jezikovnih tehnologij.

V Sloveniji se je na področju jezikovnih tehnologij že oblikovalo precej raziskovalnih in aplikativnih jedrnih skupin z dragocenim znanjem in izkušnjami (prim. Vintar, ur. 1999, Erjavec in Gros, ur. 1998, Kačič, ur. 1998). Obstaja tudi že nekaj slovenskih jezikovnih virov, primerljivih z infrastrukturo nekaterih svetovnih (in evropskih) jezikov; slovenščina je vključena tudi v nekatere večjezikovne vire. Za jezikovno načrtovanje knjižnega jezika oziroma javnega sporazumevanja je najbolj potrebni referenčni korpus nekega jezika; v okviru projekta FIDA je bil oblikovan prvi referenčni korpus slovenskega jezika (Gorjanc 1999; http://www.fida.net) Toda korpus Fida je predvsem pisni korpus, torej večinoma zajema izhodiščno pisna (natisnjena in elektronsko objavljena) besedila (Gorjanc 1999, Stabej 1998); delež transkribiranih govorjenih besedil je sorazmerno majhen.

Zakaj je poleg referenčnega pisnega korpusa potreben tudi korpus transkibiranih govorjenih besedil? Iz zgodovine opisnih in normodajnih jezikoslovnih pristopov večine standardnih jezikov se da razbrati dolgo kulturno in statusno prevlado jezika pisnih besedil. Prestižna pisna besedila (sprva religiozna, nato dolgo časa umetniška) so bila temelj za opis in normiranje standardnega jezika. Govor oz. jezik govorjenih besedil je šele pozno postal legitimen predmet jezikoslovnega raziskovanja (razen pri nekaterih vejah jezikoslovja, npr. dialektologiji, kjer pa gre samo za posebej določen ozek segment govora). Zaradi psiholoških oz. kognitivnih in komunikacijskih razlogov, ki so danes v precejšnji meri pojasnjeni, je govor veljal za manj popolno obliko sporazumevanja, torej tudi manj prestižno, zaradi česar je bil tudi neprimeren kot vir analize oz. spoznavanja jezika. Govorjeni jezik je pravzaprav šele z razvojem elektronskih medijev (radio in televizija) postajal prestižnejši in deloma tudi normotvoren; toda prav zaradi zagotavljanja prestižnosti oz. legitimnosti je medijski govor pomenil sprva predvsem oralizacijo pisnih besedil.

Po drugi strani je seveda popolnoma očitno, da je govorno sporazumevanje v marsičem primarnejše (tako v jezikovno- kot v osebnorazvojnem smislu) in bistveno pogostejše od pisnega sporazumevanja. Govorjeni jezik zaznamujejo lastnosti, v marsičem temeljno drugačne od jezika pisnih besedil. Nekateri raziskovalci zato celo domnevajo obstoj dveh vzporednih temeljnih slovnic (oz. jezikovnih zmožnosti) nekega jezika, govornega in pisnega. Tej hipotezi je sicer v celoti težko pritrditi, saj se zdi, da različne oblike človeškega jezikovnega vedenja vendarle ustrezneje pojasnjuje izhodiščno enotna jezikovna zmožnost. Toda dejstvo je, da je analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna.

Empirični pristopi k raziskovanju jezika so seveda najrazličnejših vrst. Toda razvoj korpusnega jezikoslovja je v empirično perspektivo vnesel za razumevanje jezika in sporazumevanja nasploh izredno dragoceno dimenzijo avtentičnosti opazovanega gradiva. Zbiranje in urejanje korpusnih jezikovnih virov torej načeloma v ničemer ne vpliva na nastanek izhodiščnega gradiva. Vsak kasnejši poseg v gradivo, ponavadi opravljen predvsem zaradi tehnološke obdelave za primerno dostopnost in uporabnost elektronskih jezikovnih virov, pa je dokumentiran in zato transparenten. Tak pristop k oblikovanju jezikovnih virov,

ki je v marsičem diametralno nasproten intuitivnemu jezikoslovnemu opazovanju (temeljno pa se razlikuje tudi od načrtovanih in v idealnem okolju oblikovanih (analognih in digitaliziranih) govornih zbirk za potrebe jezikovnih, predvsem govornih tehnologij), pomeni nujen pogoj za kvalitativni preskok v analizi jezika. In če se korpusni pristop v raziskovanju nekega jezika omeji samo na pisni jezik, je to splošno gledano v nasprotju s samo izvorno korpusno epistemologijo. Seveda imajo tudi samo pisni korpusi svoj smisel, tako kot poligon za tehnološka vprašanja v zvezi z gradnjo in obdelavo korpusov kot tudi za določene analize in aplikacije. Toda če je idealni cilj korpusno podprtega jezikoslovja spoznavanje jezika, kot je izpričan v vseh razsežnostih sporazumevanja, je samo pisni korpus premalo.

2. Namen KGB

KGB bi pomenil uravnotežen referenčni vir za raziskovanje govorjene slovenščine. V idealnem primeru bi torej moral v ustreznih deležih zajeti vzorce vseh govornih položajev. Merila za doseganje uravnoteženosti bi morala biti oblikovana v skladu z mednarodnimi priporočili oz. standardi (npr. Eagles; URL: http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html) in z upoštevanjem slovenskih posebnosti. Seveda pa bo pri načrtovanju uravnoteženosti potrebno upoštevati tudi metodološko in tehnološko zahtevnost gradnje govornega korpusa. Preambiciozno zastavljen obseg korpusa in prerazvejana mreža zajema besedil bi kaj lahko privedla do neizvedljivosti načrta. Zato bo najbrž potrebno načrt obsega in uravnoteženosti KGB oblikovati postopno. Toda smisel KGB se vendarle izpolni samo s pogojem, da je referenčen, saj bi tako dopolnil korpus Fida in v slovenskem prostoru zagotovil kvalitetni minimum referenčnosti jezikovnih virov. Idealni cilj bi segel še dlje: združitev korpusa FIDA (kvantitativno in kvalitativno dopolnjenega) in KGB v enovit korpus. Vsaj za jezikoslovno raziskovanje in aplikacije bi bilo še bolj idealno, če tak enovit korpus ne bi bil zaključenega, temveč odprtega tipa, torej z nenehnim nadzorovanim vključevanjem novega gradiva. Toda KGB bi bil tudi v bistveno manj idealnem obsegu izjemno uporaben. Ne le za jezikoslovje, temveč tudi za govorne tehnologije, ki morajo pri razvoju sistemov za razpoznavo in sintezo govora vse bolj upoštevati in uporabljati tudi avtentično (torej nenačrtovano, neidealizirano) govorno gradivo.

3. Zajem

Premislek o zajemu besedil za KGB je torej zaznamovan po eni strani s potrebami referenčnosti in uravnoteženosti, torej z vprašanji zvrstnosti, po drugi strani pa z metodološko in tehnološko izvedljivostjo in nenazadnje tudi s finančnimi in pravnimi vprašanji.

Zasnova KGB odpira predvsem vprašanje zajema spontanega zasebnega dialoga, ki v vsakem primeru pomeni poseg v privatno sfero ali celo v avtorske pravice posameznih govorcev. Problem je rešljiv z vnaprejšnjim dogovorom s subjekti, čeprav je lahko ob prisotnosti mikrofonov in snemalne tehnike tudi v tem primeru vprašljiva spontanost govornih izvedb. Zato je v začetni fazi izdelave govornega korpusa smiselna odločitev za zajem javno govorjenih besedil. Ena glavnih tehničnih prednosti pri takšni odločitvi je lažja dostopnost besedil in akustična kakovost posnetega gradiva. Lažje pa je tudi načrtovanje reprezentativnosti po različnih kriterijih, saj je mogoče za načrt zajema uporabiti npr. programske sheme medijskih hiš. KGB bi bil torej v prvi fazi referenčni korpus slovenskega javnega govora. Šele po uspešno opravljeni prvi fazi (ki bi prinesla poleg rezultatov tudi potrebne izkušnje in znanje) bi se delo nadaljevalo z zajemom in transkripcijo spontanega nejavnega govora. Za drugo fazo bi bila potrebna bistveno bolj kompleksna tipologija zajema govornih položajev (predvsem po sociooloških in pragmatičnih, pa tudi po tematskih in drugih vidikih), pa tudi sama metodologija zajema in transkripcija bosta precej zahtevnejši.

Zaradi naštetih dejstev je znotraj splošne zvrsti javnega govora za potrebe govorjenega korpusa najustreznejši zajem govora v medijih, kar med drugim zanesljivo zagotavlja kvaliteto akustičnega signala, ki je osnova za fonetično analizo in transkripcijo. Govorce, ki nastopajo v govornih občilih (radio, televizija), je treba razdeliti na medijske (poklicne) govorce, katerih govorne izvedbe bomo imenovali medijski govor, in na druge govorce, ki v oddajah najpogosteje nastopajo v vlogi gostov, njihova govorna kompetenca pa je bolj ali manj odvisna od izkušenj, ki so si jih v javnem izrekanju pridobili v lastni jezikovni praksi.

Poklicni medijski govorci se pri svojem izrekanju najpogosteje znajdejo v dveh tipičnih sporočanjskih položajih. Novinarji in napovedovalci največkrat izgovarjajo vnaprej pripravljeno zapisano besedilo, zato bomo ta položaj imenovali oralizacija pisnega besedila. V kontaktnih oddajah ali oddajah z vabljenimi gosti pa so soočeni z izzivom prostega govora, pri čemer naj bi bile njihove govorne izvedbe zaznamovane s pridobljeno poklicno govorno kompetenco. Medijski govor je torej v obeh navedenih položajih govor, ki naj bo kar najzvestejši približek idealnim govornim izvedbam (idealna intenca), ob čemer se jasno zastavlja vprašanje govorne norme, njenega opisa in predlogov za njeno morebitno redefinicijo. Odstopi od idealne (predpisane) normativne izreke v teh položajih predstavljajo poglavitne prvine jezikovne realnosti. Kot element individualnih izvedb namreč hkrati pričajo o absolutni neidealnosti realiziranih govornih tvorb in s svojo spontanostjo potrjujejo smisel zasnove korpusa govorjenih besedil za slovenski jezik.

S stališča besediščne, oblikoslovne, skladenjske in besedilne analize je seveda od javnih govorjenih besedil najbolj zanimiv prosti govor (brez pisne predloge), saj se najbolj razlikuje od jezika pisnih besedil. Zato je najbrž smiselno, da ima v KGB prosti govor večinski delež.

4. Oblika

Zajeto govorno gradivo mora biti posneto na nosilce zvočnega signala, ki so najlažje uporabni za takojšnjo računalniško aplikacijo. To pomeni, da je najustreznejši nosilec elektronskega zapisa minidisk, ki je neposredno kompatibilen z računalniško tehnologijo. Zajeto govorno

gradivo mora biti sproti označeno in arhivirano po vnaprej izdelanih in standardiziranih načelih.

KGB ima v glavnem dva jasno opredeljena cilja: korpus je v prvi vrsti sestavljen iz zbirke zvočnih signalov (posnetih govorjenih besedil), ki jih je za raznovrstne obdelave (analize) kajpak treba transkribirati. Fonetične in fonološke analize zahtevajo programsko opremo, ki omogoča vizualizacijo in ustrezno označevanje akustičnega signala. Transkribirano besedilo pa po drugi strani omogoča strukturno jezikovno analizo govora (morfološko, skladenjsko, semantično) in ustvarja ustrezne razmere za raziskovanje in opredelitev slovnice slovenskega govora. Gre za jezikoslovno področje, ki je v slovenskem prostoru razmeroma novo, njegova snov pa izrazito nezadostno raziskana.

Poglavitni problem izpopolnjene izvedbe KGB je torej tehnologija transkripcije govornega gradiva. Raziskovalne ekipe si v zadnjih dveh desetletjih na različnih znanstvenih ustanovah intenzivno prizadevajo izdelovati metodologije govornih tehnologij za slovenski jezik, vendar so rezultati njihovih raziskovalnih naporov (in tako je tudi drugod po svetu) močno omejeni. Razlogi za to tičijo prav v neidealni naravi individualnih spontanih govornih tvorb in v težavah s prepoznavanjem povezav med jezikovnimi substancami (realizacija) in formami (reprezentacija). Idealno izvedena tehnologija simbolnega razpoznavanja govornega akustičnega signala bi omogočila avtomatičen transfer akustične vsebine v grafično obliko, z drugimi besedami, idealen razpoznavalnik govora in pretvarjalnik v pisno obliko bi skrajno ekonomiziral transkripcijo zvočnega gradiva KGB. Zaradi neponovljivih specifik govornih realizacij pa razpoznavalnika, ki bi enako popolno obravnaval interindividualno in intraindividualno variabilnost govora, ni. To pomeni, da bo zasnova in izdelava KGB za slovenski jezik zahtevala precejšnje število izurjenih transkriptorjev, ki bodo za svoje delo (v hipotetični primerjavi z idealnim razpoznavalnikom) sproti rabili znatne količine časa.

Tako za samo transkripcijo kot za oblikovanje elektronskega korpusa transkribiranih besedil je torej potrebno oblikovati podrobna načela. Pri tem je najbolje sprejeti mednarodna priporočila in standarde (prim. Erjavec 1996/97); za transkripcijo jih bo treba šele izbrati in preizkusiti, za samo obliko KGB pa bi kazalo prevzeti rešitve pri oblikovanju korpusa Fida (Erjavec 1998). Dokler ni vsaj okvirno jasno, kako bo oblikovanje KGB potekalo (torej do prvih pravih preskusov snemanja, transkripcije in računalniške obdelave), najbrž tudi še nima smisla govoriti o sami velikosti korpusa.

5. Organizacija in izvedba

Že prva, sorazmerno lažja faza izdelave KGB bo zahtevna in nujno terja sodelovanje več partnerjev. Seveda se pri vsakem projektu postavlja vprašanje interesa, znanja, zmožnosti in časa sodelujočih ter vprašanje denarja. Omeniti je treba, da je skupina raziskovalcev s Filozofske fakultete pri Ministrstvu za znanost in tehnologijo RS v okviru razpisa za (so)financiranje znanstvenih projektov za l. 2000 prijavila projekt Jezikoslovna načela oblikovanja za gradnjo korpusov slovenskega jezika. Del projektne vsebine je bilo tudi oblikovanje načel za gradnjo govornega korpusa v slovenščini. Projekt je bil – kot skoraj vsi projekti s področja humanistike – brez obrazložitve zavrnjen. Izkušnja korpusa Fida kaže, da je za nemoteno in učinkovito delo dobrodošlo sodelovanje med raziskovalnimi in kapitalskimi partnerji. Toda vsaj na prvi pogled je razvoj korpusa govorjenih besedil za kapitalske partnerje manj privlačen. Zato bi se za izdelavo prve faze KGB morale povezati zainteresirane raziskovalne ustanove, se dogovoriti o deležu dela in poiskati vire financiranja. Najbrž je KGB v interesu tako tistih ustanov, ki se ukvarjajo z jezikoslovjem, kot tudi tistih, ki so dejavne na področju jezikovnih tehnologij. Jezikoslovje bi najbrž lahko adekvatno poskrbelo za izdelavo okvira za zajem besedil in njihovo (ročno) transkripcijo, strokovnjaki za jezikovne tehnologije pa za tehnologijo zajema, za standardizacijo in oblikovanje korpusa. Prvi obotavljivi korak je narejen: morda bi se lahko ideja o KGB jasneje oblikovala tudi v okviru kake stalne debatne skupine v Slovenskem društvu za jezikovne tehnologije (URL: http://nl.ijs.si/sdjt/).

6. Literatura

Erjavec, T. 1996/97:Računalniške zbirke besedil. *Jezik in slovstvo 2/3*, 81-95.

– – 1998: Oznake korpusa FIDA. *Uporabno jezikoslovje 6.* Tematska številka Jezikovne tehnologije. Ur. Zdravko Kačič. 85-95.

Erjavec, T., Gros, J., ur. 1998: *Jezikovne tehnologije za slovenski jezik: zbornik konference = Language technologies for the Slovene language: proceedings of the conference*. Ljubljana: Institut Jožef Stefan.

Gorjanc, V. 1999: Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. *Zbornik predavanj / 35. seminar slovenskega jezika, literature in kulture, 28. 6. - 17. 7. 1999*. Ljubljana : Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete, 47-59.

Kačič, Z. 1998: *Uporabno jezikoslovje 6*. Tematska številka Jezikovne tehnologije. Zbornik referatov z II. kongresa Društva za uporabno jezikoslovje Jezik za danes in jutri. Ljubljana: Društvo za uporabno jezikoslovje.

Stabej, M. 1998: Besedilnovrstna sestava korpusa Fida. *Uporabno jezikoslovje 6*. Tematska številka Jezikovne tehnologije. Ur. Zdravko Kačič, 96-106.

Vintar, Š., ur. 1999: *Proceedings of the Workshop Language Technologies - Multilingual Aspects* : within the framework of the 32nd Annual Meeting of the Societas Linguistica Europea, 8 - 11 July 1999, Ljubljana, Slovenia. Ljubljana : Filozofska fakulteta, Oddelek za prevajanje in tolmačenje.