Vojko Gorjanc, Nataša Logar
Ljubljana

UDK 81'322:811.163.6'276.6:355

OD SPLOŠNIH DO SPECIALIZIRANIH KORPUSOV – NAČELA GRADNJE GLEDE NA NJIHOV NAMEN

Prispevek glede na značilnosti različnih vrst korpusov predstavlja izhodiščna načela njihove gradnje. Ta so povezana predvsem s specifikacijo korpusa, strojno in programsko opremo, zajemom in označevanjem korpusnih dokumentov, procesiranjem gradiva in končno oblikovanostjo korpusa. Z vidika naštetega je predstavljen 5,5-milijonski korpus slovenskih vojaških besedil, ki kljub ozkemu namenu, zaradi katerega je bil narejen, v primerjavi z *Vojaškim slovarjem*, ki je izšel leta 2002, izkazuje številne nove vojaške termine.

korpusno jezikoslovje, splošni korpus, specializirani korpus, načela za gradnjo korpusov, korpus slovenskih vojaških besedil

The article presents basic corpus design criteria according to the characteristics of different types of corpora. These criteria are concerned with the specifications of the corpus, hardware and software, data capture and mark-up of the corpus document, corpus processing, and the final design of the corpus. In relation to these criteria, a 5.5-milion word corpus of Slovene military texts is presented. Although this corpus has been designed for a very narrow purpose, many new military terms can be extracted from it if we compare it with the new *Vojaški slovar* (Military Vocabulary) printed in 2002.

corpus linguistics, general language corpus, specialized corpus, corpus design criteria, corpus of Slovene military texts

1 Uvod

Z gradnjo korpusov smo v slovenskem prostoru v devetdesetih letih prejšnjega stoletja dobili osnovno izhodišče za jezikoslovne študije, zasnovane strogo empirično, na podlagi podatkov o besedilnih realizacijah. V vsakem jezikovnem okolju prav jezikovni viri, med njimi posebej korpusi, pomenijo osnovo jezikovne infrastrukture, zato so razmisleki o njihovi gradnji vedno aktualni. Če se je o njih precej razpravljalo predvsem ob pripravi referenčnega korpusa FIDA in nekaterih vzporednih korpusov, pa je bilo sorazmerno malo diskusije o specializiranih korpusih. To vrzel skuša vsaj deloma zapolniti pričujoči članek.

2 Splošna načela gradnje korpusov

Pri načelih gradnje korpusov je najprej treba pripraviti okvirni načrt gradnje, ki zajema tako jezikoslovne kot nejezikoslovne premisleke in odločitve ter tudi

razmislek o povsem tehničnih rešitvah. V osnovi bi jih lahko strnili v naslednje sklope (Atkins et al. 1992: 2):

I. specifikacija korpusa in njegova oblika,
II. strojna in programska oprema,
III. zajem besedil in označevanje korpusnih dokumentov,
IV. procesiranje zbranega gradiva,
V. končna oblikovanost korpusa in povratne informacije v zvezi z njim.

V največji meri so jezikoslovni premisleki v zvezi s korpusom vezani na prvo točko, torej specifikacijo korpusa in njegovo obliko. Temeljni premislek se tiče tipa korpusa (Kennedy 1998: 19–23), ki ga želimo graditi; ta za seboj potegne temeljne odločitve, in sicer predvsem v zvezi s časovnim zajemanjem besedil (sinhroni ali diahroni korpus) ter premislek o zajemu besedil glede na prenosnik (pisni ali govorni korpus). Izhodiščni jezikoslovni premislek pa je vezan tudi na določitev parametrov za uravnoteženost korpusnih podatkov na eni (Biber 1993: 243) ter njihovo jezikoslovno označenostjo v korpusu na drugi strani (Atkins et al. 1992: 7–8).

Da bi z gradnjo sploh lahko začeli, je potrebna tehnična podpora, ki mora od samega začetka slediti zahtevam tako glede strojne kot programske opreme ter biti sposobna oblikovati orodja za procesiranje zbranega gradiva. Prav pri procesiranju podatkov se je treba odločati tako, da jezikovnim podatkom zagotovimo čim večjo uporabnost, izmenjavo ter trajnost, kar v zadnjem času omogočajo standardi za prenos in zapis jezikovnih podatkov (Erjavec 1998: 119).

Čeprav se razmislek v zvezi s postopki zajemanja besedil zdi dokaj trivialen, pa so se korpusi prav na tem nivoju velikokrat znašli pred nerešljivo težavo: kako sploh organizirati zbiranje besedil ter prepričati besedilodajalce, da za namene korpusa svoja besedila odstopijo. Prav zaradi nepredvideno zapletenih postopkov se je pri mnogih korpusih njihova gradnja precej zavlekla, tako da se danes vsi zavedajo zahtevnosti in zamudnosti zbiranja besedil (Atkins et al. 1992: 3).

S pridobivanjem besedil je povezano še eno temeljno vprašanje, ki ga mora vsak resno zastavljen korpusni projekt rešiti pred začetkom gradnje, tj. zagotavljanje varovanja avtorskih pravic. Potrebno je poznavanje področja varovanja avtorskih pravic, in sicer tako na mednarodni kot državni ravni ter v skladu s tem oblikovanje ustreznih rešitev (Atkins et al. 1992: 4). Prav izkušnje pri korpusih prve generacije, ki vprašanja avtorskih pravic niso zadovoljivo rešile, tako da danes tovrstnih podatkov sploh ni mogoče uporabljati, so korpuse druge generacije prisilile v razmislek ter iskanje ustreznih rešitev.

Pri končni obliki korpusa je z vseh vidikov, tako jezikoslovnih kot nejezikoslovnih, smiselno spremljati odzive na rešitve, jih sistematično obdelati in razmisleke v zvezi z gradnjo revidirati ter tako pri njegovi nadgradnji dosegati večjo kakovost ter prijaznost do uporabnikov.

Z jezikoslovnega vidika je pri zajemanju besedil eden od temeljnih premislekov povezan z besedilnimi lastnostmi in njihovim vplivom na odločitev za zajemanje besedil. Nabori korpusno zanimivih besedilnih lastnosti so se oblikovali postopo-

ma, v glavnem nekako do začetka devetdesetih let (Atkins et al. 1992, Biber 1993); predstavljajo osnovo za določitev tistih, ki bodo relevantne za uravnoteženost korpusa, ter tistih, ki se jih bo korpusnim dokumentom pripisovalo glede na odločitev, katere bi bile pri uporabi korpusnih podatkov relevantne.

Nabor lastnosti korpusnih dokumentov:

Prenosnik: pisni (pisni za branje, pisni za govor), govorjeni (govorjeni za zapis)
Avtorskost: enoavtorski, večavtorski, skupni, korporacija
Število avtorjev
Ustroj dokumenta: enobesedilni, večbesedilni (časopis, revija, zbirka pesmi ...)
Besedilna predpriprava: pripravljen, na podlagi zapiskov, spontano
Medij: natisnjen (knjiga, periodika, ephemera), rokopis
Besedilna samooznaka: pesem, roman, kratka zgodba ...
Besedilna vrsta in govorni položaj: odprt niz (jezikovno in kulturno najbolj specifična lastnost)[1]
Stvarnost: stvarna : umetna literatura
Besedilni kontekst: izobraževanje, dom, delo, prosti čas ...
Jezikovna funkcija: glede na vplivanjsko vlogo
Predmetnostno področje: glede na dogovorjeni nabor
Strokovnost: strokovni, nestrokovni, poljudni
Datum izida
Besedilni status: original, ponatis
Jezik
Jezikovni status: original, prevod
Metodologija (lastnost, relevantna za korpuse strokovnih besedil)
Ime avtorja
Spol avtorja
Starost avtorja
Regijska pripadnost avtorja
Nacionalnost avtorja
Prvi jezik avtorja
Avtoriteta avtorja (lastnost, relevantna za korpuse strokovnih besedil)
Starost ciljne publike
Velikost ciljne publike
Družbeno razmerje avtorja do ciljne publike

S širjenjem računalniških korpusov se je pojavila tudi potreba po vrednotenju in razvrstitvi korpusov. Z opisom karakteristik, s katerimi lahko korpuse vrednotimo, in z definiranjem zvrsti korpusov, ki jih je med seboj smiselno razlikovati, se je ukvarjala skupina za tipologijo korpusov pri evropski iniciativi Eagles – Expert

[1] Eden od francoskih govorjenih korpusov tako v nizu govornih položajev izpostavi tudi »v trgovini s siri« http://www.people.cornell.edu/pages/adl6/french_corpus.htm; govorjeni korpus modernega francoskega jezika nastaja pod okriljem Cambridge University Press in Univerze Cornell (Ithaca, NY).

Advisory Group for Language Engineering (Erjavec 1996/97: 82). Eagles je z analizo stanja na področju gradnje korpusov povzel rešitve in jih sistemiziral, tako da se danes pri gradnji korpusov večina zgleduje prav po njih. Eagles nikoli ni želel postati korpusni standard, pač pa oblikovati le priporočila za gradnjo korpusov, a pri tem ostati odprt za različne rešitve.

3 Gradnja specializiranih korpusov

Načela oblikovanja specializiranih korpusov so izhodiščno prekrivna z načeli oblikovanja splošnih, dodatno pa jih je treba preoblikovati in dopolniti predvsem glede na namen korpusa. Oblikovanje načel korpusov strokovnih jezikov in njihova gradnja sta izhodišči za raziskave strokovnih jezikov, aktualne predvsem na področju terminologije; izjemno dinamiko znanstvenega in tehnološkega razvoja ter hiter pretok informacij lahko namreč zajamejo le aktualni korpusi.

Za razliko od splošnih referenčnih korpusov predstavljajo specializirani jezik v določeni funkciji in tako predstavljajo omejeni, glede na tip specializiranega korpusa točno določeni del jezika. Ko govorimo o korpusih strokovnega jezika, namreč vedno ne govorimo le o določenem predmetnostnem področju, kot je to običajno pri npr. terminoloških raziskavah, ampak je lahko korpus strokovnega jezika tudi veliko ožje zamejen, in sicer tako predmetnostno, npr. samo korpus s področja jedrske fizike, kot komunikacijsko, npr. korpus besedil komunikacije med kontrolnim stolpom in pilotom. Vsekakor je na začetku gradnje korpusa strokovnega jezika pomembna prav določitev predmetnostnega področja in s tem možnega nabora besedil. Pri določanju predmetnostnega področja pa se moramo zavedati njegovih mehkih mej, še posebej s težnjo po multidisciplinarnosti raziskav, ki je odprla področne povezave in naredila stranske meje v razmerju do drugega področja še mehkejše. Hkrati pa je vsaj okvirno treba določiti tudi zgornje meje besedil; do kod v splošni jezik seči z ugotavljanjem besedilne vpetosti jezikovnih elementov, značilnih za strokovno komunikacijo (Meyer, Mackintosh 1996: 260, 268–269).

Izhodiščni premisleki o načelih gradnje in načrtih za izdelavo korpusa so prekrivni; priporočila in standardi v zvezi z gradnjo korpusov veljajo tudi tu, pri korpusih strokovnih jezikov pa predvidevajo tudi posebnosti.

3.1 Velikost

Korpusi strokovnih jezikov v izhodišču ne gradijo na količini, ampak kakovosti besedilnih virov (Meyer, Mackintosh 1996: 268). Že zaradi količine besedilne produkcije seveda ne moremo pričakovati korpusov tako velikega obsega kot pri referenčnih, hkrati pa je velikost odvisna od tega, katero predmetno področje bo korpus zajel, saj je posledično od količine besedilne produkcije na tem področju odvisno, kakšna je sploh zgornja meja obsega korpusa, vsekakor pa velja, da je zgornja meja nedoločena, treba je torej težiti k zajetju čim večje količine dostopnih besedil (Pearson 1998: 59). Pri korpusih strokovnih jezikov je tudi popolnoma

izključena možnost besedilnega vzorčenja. Za korpus so relevantna le besedila v celoti, saj je vsebinska razpršenost besedilno popolnoma nepredvidljiva (Bowker 1995: 42–423; Meyer, Mackintosh 1996: 268); z naključnim vzorčenjem bi tako lahko izgubili informacije, ki se v besedilu lahko pojavijo tudi samo enkrat.

3.2 Sinhronost

Korpusi strokovnih jezikov težijo po zajetju čim bolj aktualnega besedilnega gradiva in sprotnem zajemanju novega; tako se v tem okviru gradijo le dinamični korpusi. Zaradi hitrega razvoja znanosti in tehnologije pa morajo biti grajeni pretočno; poleg vključevanja novega morajo namreč omogočati tudi izločanje starejšega besedilnega gradiva (Meyer, Mackintosh 1996: 269). Starejše gradivo pa tvori podkorpuse, ki so dragoceni vsaj z vidika večje besedilne informacije; pri prvih pojavitvah termina je iz besedila ponavadi lažje ugotoviti njegov pomen kot kasneje, ko se termin že ustali v terminološkem sistemu, hkrati pa podajajo informacijo tudi o starejših rešitvah, ki se mogoče niso ustalile ali pa so zaradi spremembe pojmovnega polja enostavno izginile (Bowker 1996: 43).

3.3 Uravnoteženost

Čeprav se zdi vprašanje uravnoteženosti in reprezentativnosti specializiranih korpusov v razmerju do splošnih referenčnih sorazmerno lahko rešljivo, pa vendarle v premislek ponuja niz vprašanj.

1. *Prenosnik*. Tako kot za splošne referenčne korpuse velja tudi za korpuse strokovnih jezikov klasična delitev na pisni in govorni kod, čeprav se danes s pojavitvijo elektronskih medijev vse bolj briše klasična delitev; prav na področju strokovne komunikacije je elektronski prenosnik zaradi hitrega pretoka informacij še posebej aktualen. Kljub temu je izhodiščna orientacija na pisna besedila ustrezna, saj velja, da je večina znanstvene komunikacije vezane predvsem na pisni prenosnik, ki ima še vedno prestižno vlogo, vendar jo z elektronsko komunikacijo počasi, a vztrajno izgublja. Če so referenčni korpusi druge generacije že v izhodišču predpostavljali tudi zajetje govora, pa se pri korpusih strokovnih jezikov šele v zadnjem času vse bolj uzavešča spoznanje o nujnosti upoštevanja tudi govornih besedil, saj korpusni pristop teži k popisu jezikovne realnosti, ta pa se v strokovni komunikaciji nemalokrat kaže v govoru drugače kot v zapisu. Tako se pri zajetju govora premišlja o uravnoteženosti korpusa z vidika udeležencev komunikacije v razmerjih strokovnjak : strokovnjak, strokovnjak : laik in strokovnjak : leksikograf/terminograf (Meyer, Mackintosh 1996: 272–273). Prav zajetje govora pri analizah strokovne komunikacije bi lahko prineslo novosti na ravni opisa strokovne komunikacije, podobno, kot se je to zgodilo pri referenčnih korpusih.
2. *Analiza področja*. Ko je predmetnostno polje določeno in so njegove meje vsaj okvirno določene, je potrebna analiza vseh podpodročij, ki jih zajema, saj je osnovna ideja korpusa, da optimalno pokrije celotno strukturo posameznega področja.

3. *Besedilnovrstna uravnoteženost*. S tega vidika je prav tako pomembna predhodna analiza, saj predmetnostna področja glede nabora besedilnih vrst v strokovni komunikaciji niso prekrivna. Korpus določenega strokovnega jezika mora zajeti besedila različnih jezikovnih podzvrsti, znotraj njih pa čim širšo paleto besedilnih vrst. Razumljivo je, da različna besedila prinašajo različne informacije, zato je korpus s tega vidika nujno uravnotežiti. Tako mora v osnovi pokriti tri osnovne tipe besedilnih vrst, in sicer znanstvena (znanstvena razprava, monografija; znanstvena diskusija), didaktična (učbenik, navodilo, enciklopedijski članek; pedagoška komunikacija) in poljudna (poljudnoznanstveni članek, časopisni članek) (Bowker 1996).
4. *Avtor/Urednik*. Za doseganje reprezentativnosti se v zbirko vključujejo besedila v stroki uveljavljenih avtorjev, ustreznost besedilnega gradiva lahko zagotavlja tudi urednik ali uredniški odbor uveljavljene strokovne publikacije. Ob tem je treba zagotoviti tudi z vidika avtorskosti čim širšo zastopanost in s tem izključiti individualne posebnosti (Atkins, Clear 1992), hkrati pa tudi tu skušati glede avtorskosti zajeti tako eno- in večavtorska besedila ter besedila, katerih avtorskost je vezana na institucijo (Pearson 1998: 60); količina različnih tipov je odvisna od področja, saj je na nekaterih prevladujoč enoavtorski pristop, na drugih pa večavtorski. Pomemben kriterij v okviru avtorskosti je (splošna) sprejetost besedila v strokovni javnosti; besedila, ki so bila v stroki argumentirano javno zavrnjena, namreč ne morejo predstavljati reprezentativne besedilne zbirke.
5. *Prevodnost/Neprevodnost.* Če je za angleški prostor relevantna razprava o izvornih in prevodnih besedilih ter o prvem jeziku avtorja, pa za vse druge jezike pomeni danes prav soočanje z angleščino kot vse bolj prvotnim medijem znanosti in tehnologije pomemben segment strokovne komunikacije. Angleški korpusi težijo k izvorno angleškim besedilom, na prevode pristajajo le, če se na določenem področju ne da zagotoviti ustreznih besedilnih virov (Bowker 1995: 44). Pri jezikih z manjšim številom govorcev, na področju strokovne komunikacije pa pravzaprav število govorcev sploh ni več relevantno, je situacija seveda popolnoma drugačna. Težava je sploh besedilna pokritost celotnega področja, tako da se pojmovni svet stroke ne gradi izključno z ubesediljenjem, ampak velikokrat samo s prevodnim naborom terminologije. Po drugi strani pa se angleščina sooča z novim izzivom. Če je pri referenčnih korpusih še lahko izključevala besedila avtorjev, katerih angleščina ni prvi jezik, pa je na področju strokovne komunikacije popolnoma drugače. Na tem delu človekovega življenja in ustvarjanja se vse bolj brišejo nacionalne in regionalne meje, angleški jezik pa postaja prevladujoči medij, tako da ga korpusi strokovnih jezikov morajo vključevati (Bowker 1995: 44). Čeprav se zavedajo njegove drugačnosti in prisotnosti »nenaravnih jezikovnih konstrukcij« mimo tega dejstva enostavno ne morejo.[2]

[2] Podobno je tudi s splošnim jezikom, a tam ta problem angleški referenčni korpusi zaobidejo, saj ne zajemajo besedil avtorjev, ki jim angleščina ni prvi jezik.

Biber in dr. (2000: 246) opozarjajo, kako je »pomembno, da se vnaprej zavedamo, da je reprezentiranje jezika – ali le dela jezika – problematična naloga. […] Vendar pa bo to, da smo /pri gradnji korpusa/ pozorni na določena vprašanja, omogočilo čim večjo reprezentativnost korpusa glede na naše trenutno vedenje o jeziku«. Tudi Leech (1991: 27/2005: 33) je pri pojmu reprezentativnosti previden: »V praksi je korpus 'reprezentativen' do te mere, da se spoznanja, ki temeljijo na njegovi vsebini, lahko posplošijo na večji hipotetični korpus. […] Danes je treba v predpostavko o reprezentativnosti preprosto verjeti.« Gorjanc (2005b: 38) pa povzema, da se »[p]ri nekaterih korpusnih pristopih /pojem reprezentativnosti/ sploh opušča […] in se nadomešča le z okvirno mrežo kriterijev zajetja besedil v korpus in veliko količino besedilnega gradiva«.

3.4 Homogenost

Za razliko od pokritosti čim več različnih zvrsti, ki jo pričakujemo pri splošnih korpusih, je pri korpusih strokovnih jezikov aktualnejši pojem homogenosti korpusa, in sicer homogenosti »v smislu besedišča, ki ga predstavlja« (Vintar 2003: 32). Homogenost je tesno povezana z namenom korpusa. Pri gradnji korpusa za namene luščenja terminologije se je npr. pokazalo (Vintar 2003: 30–31), da je luščenje tem bolj uspešno, čim manj je v korpusu različnih besedilnih vrst, boljši pa so bili rezultati tudi pri luščenju iz besedil, med katerimi je bil manjši časovni razpon, in besedil, ki so vsebovala malo ekskurzov na druga področja.

4 Korpus vojaških besedil

V okviru Ciljnega raziskovalnega programa Konkurenčnost Slovenije 2001–2006 je bil Fakulteti za družbene vede leta 2002 odobren raziskovalni projekt z naslovom *Nadgradnja slovenskega vojaškega slovarja: večjezični vojaški slovar*.[3]

KSVJG je prva in testna različico korpusa slovenskih vojaških besedil[4] in ni – kot bo iz nadaljnjega razvidno – uravnoteženi ali homogeni korpus sodobnega slovenskega vojaškega jezika nasploh. Njegovo oblikovanje je determiniral naslednji glavni namen: izločitev okrog 1000 sodobnih vojaških terminov, ki bodo zgolj *dopolnitev* nabora terminov, ki so že zapisani v obstoječem *Vojaškem slovarju*, ki je v uredništvu T. Korošca in drugih izšel leta 2002 (dalje VS 2002).[5] Dopolnilnost že obstoječemu VS 2002 je bila torej tudi izhodišče za izbor besedil, ki so bila

[3] V5-0775, vodja dr. Tomo Korošec. Projekt sta financirala Ministrstvo za obrambo RS in Ministrstvo za šolstvo, znanost in šport RS. Pri tem – kot je jasno iz naslova – leksikografskem projektu je nastal tudi Korpus slovenskega vojaškega jezika Grizold (http://ksvjg.fdv.uni-lj.si/, dalje KSVJG).

[4] Dejansko smo se sodelavci s Fakultete za družbene vede na tem korpusu učili graditi korpus, pri čemer pa ne smemo pozabiti na dejstvo, da KSVJG ne bi nastal brez dragocenih izkušenj, ki so jih z nami delili ustvarjalci *Korpusa slovenskega jezika FIDA* (zlasti Vojko Gorjanc in Simon Krek), hkrati pa je pomembno že samo dejstvo, da je prav ta, pravkar imenovani, sicer splošni korpus v slovenskem prostoru že obstajal, kar je pomenilo, da smo imeli konkretno predstavo, kaj želimo napraviti in kako si bomo s takim jezikovnim virom lahko pomagali.

[5] Prva izdaja tega *Vojaškega slovarja* je sicer iz leta 1977, izdaja iz leta 2002 je dopolnjena in predelana.

vključena v KSVJG, seveda pa smo se zavedali dejstva, da bomo v korpusu lahko preverjali tudi pogostost že v VS 2002 zapisanih terminov, morebitno spreminjanje njihovega pomena, razvoj v besednozvezne termine itd., nenazadnje pa tudi koristnosti uporabe korpusa v pedagoškem procesu (zlasti med študenti obramboslovci).

Najstarejša besedila v KSVJG so iz leta 1985, najmlajša imajo letnico 2004 (gl. Prilogo 1). Z izjemo enega zbornika sicer javno predstavljenih referatov, ki pa (še) ni izšel, so bila vsa besedila objavljena. Manjši del besedil je tudi prevodov.

I. Večinski del KSVJG (68 % pojavnic) izvira iz besedil revije *Slovenska vojska*, in sicer smo v korpus vključili vsa besedila iz te revije (brez angleških povzetkov) od letnika 1997 do 2004. Gre za revijo, ki je začela izhajati leta 1993, danes je podnaslovljena z *Informativno vojaškostrokovno glasilo Ministrstva za obrambo RS*, izhaja pa vsakih 14 dni. Namenjena je širokemu krogu bralcev, ki so kakorkoli poklicno, službeno, študijsko ali interesno povezani s področjem obrambe, varnosti, zaščite in reševanja, z vojstvom, Slovensko vojsko, Ministrstvom za obrambo RS ipd. Njena naklada je bila v letu 2005 10.500 izvodov. Poleg strokovnih tem, ki so obravnavane na poljuden način, npr. vojaška tehnologija, vojaška psihologija, profesionalizacija vojske itd., revija vsebuje še prispevke o aktualnem dogajanju v Slovenski vojski in na ministrstvu za obrambo, o sodelovanju z Natom, vojaškozgodovinskih temah, športu v vojski, objavljeni so različni intervjuji itd. Gre torej za besedila, ki na poljuden, informativen način poročajo zlasti o stvareh, ki se tičejo vojske kot organizacije in skupnosti ljudi ter njenega vsestranskega delovanja. Avtorji besedil v *Slovenski vojski* so le v manjši meri vojaški strokovnjaki. – Iz povedanega sledi, da je v tem delu korpusa zajeta v glavnem le splošnejša, tudi v širši javnosti prisotna vojaška terminologija.

II. Na drugem mestu po deležu pojavnic (26 %) so v korpusu članki iz strokovne revije *Naša obramba* oz. *Revije Obramba*,[6] in sicer od letnika 1985 do 2002. Revija, ki izhaja kot mesečnik, je namenjena vojaškim strokovnjakom in strokovnjakom s sorodnih področij. Članke, ki so vključeni v korpus, je izbral Miroslav Ulčar, dolgoletni urednik in lektor omenjene revije, sourednik VS 2002 in velik poznavalec vojaške terminologije, in sicer na podlagi svojega védenja o tem, katera področja vojaške stroke so bila s termini v že obstoječem VS 2002 slabo zastopana. Tako so bili izbrani npr. članki s področja zgodovine orožja, ki so zajemali podrobnosti, ki dotlej v slovenskem prostoru še niso bile opisovane, zato pa tudi ne poimenovane, članki o novejši vojaški tehniki in tehnologiji, prevedeni članki o vrhunskih letalih, površinskih ladjah, podmornicah in tankovski tehnologiji, članki o taktiki modernih vojsk, specialnih silah in njihovi opremi itd. Prispevki iz *Naše obrambe*/*Revije Obramba* so v primerjavi s prispevki iz revije *Slovenska vojska* precej bolj specializirani, zato pa tudi terminološko bogatejši.

[6] Revija *Naša obramba* je začela izhajati leta 1969, leta 1991 se je preimenovala v *Revijo Obramba*.

III. Vsa druga besedila, iz katerih je v korpusu še preostalih 6 % pojavnic, so novejši vojaški priročniki, zborniki strokovno-znanstvenih prispevkov in po ena številka dveh revij (*Vojstvo* in *Bilten Slovenske vojske*). Podatkov o njihovi branosti nimamo, teh besedil tudi nismo načrtno (i)zbrali, sama po sebi so tudi dokaj raznorodna, ker pa predstavljajo le manjši del korpusa in ker gre za vojaška besedila večjega kroga avtorjev, smo jih pravzaprav brez zadržkov (tudi ob zavedanju, da gre za prvo različico nekega bodočega splošnega vojaškega korpusa) vključili – pomembno je bilo tako pri teh besedilih kot tudi pri reviji *Slovenska vojska* in *Naša obramba/Revija Obramba*, da smo jih pač lahko brezplačno dobili,[7] in to seveda v elektronski obliki.

KSVJG je v skladu s pravkar opisanimi viri torej razdeljen na tri podkorpuse: SV, NO in ZZ.

Metodologija gradnje korpusov strokovnih jezikov je – kot je zapisano zgoraj – v veliki meri prekrivna z gradnjo referenčnih korpusov, za doseganje uravnoteženosti pa so glede na specifiko strokovne komunikacije oblikovani posebni parametri zajema besedil (Gorjanc 2005a: 8). Če torej v zvezi s KSVJG premislimo izhodiščne predpostavke, ki so aktualne pri gradnji kakršnegakoli korpusa (Atkins et al. 1992: 2), ugotovimo naslednje:

a) Specifikacija korpusa in njegova oblika

KSVJG je specializirani enojezični sinhroni korpus vojaških besedil, ki so v pretežni meri iz druge polovice devetdesetih let 20. stoletja. V celoti so v njem le pisna besedila.[8] Obsega 5,5 milijonov pojavnic.

Temeljni parameter, ki je določal homogenost za zgoraj opredeljeni namen (dopolnitev VS 2002), je bil izpolnjen z izborom člankov iz *Naše obrambe/Revije Obramba*, ta izbor pa je vodila tema prispevkov. Navedeni cilj izpolnjuje dobra četrtina korpusa oz. podkorpus NO, ki ga zvrstno lahko označimo kot poljudnoznanstvenega. Večinski delež korpusa (podkorpus SV) je namenjen preveritvi in besedilnemu opazovanju predvsem širše rabljene vojaške terminologije, saj pokriva področja vojaške stroke in slovenske vojaške organiziranosti, ki se tiče tudi splošne javnosti. Zvrstno gre tu za poljudnostrokovna in publicistična besedila. Najmanjši del korpusa (podkorpus ZZ) je nastal brez natančnejših vnaprejšnjih meril izbora in je najbolj heterogen ter zvrstno obsega tako znanstvena kot poljudnostrokovna besedila. Njegova prednost je novejša letnica izdaje (vse po letu 2000) in številčnost avtorjev ter področij, njegovo terminološko vrednost pa bo treba še analizirati.

b) Strojna in programska oprema

Skupno smo zbrali nad 4300 wordovih datotek, ki smo jih najprej prečistili (odstranili vse stile, izbrisali deljaje, slike, popravili šumnike in druge posebne znake

[7] Za sodelovanje se zahvaljujemo projektni sodelavki Mileni Sevšek Potočnik iz Službe za publicistiko Ministrstva za obrambo RS.

[8] V majhnem segmentu (*Postrojitvena pravila Slovenske vojske*) se besedila, ki so v korpusu, stalno uresničujejo tudi v govoru.

ter tiste znake, ki jih je optični čitalnik prebral napačno, itd.), vse datoteke smo nato enovito poimenovali in jih shranili v formatu .txt.

Orodja za procesiranje besedil in za delo s korpusom so za končno verzijo KSVJG izdelali v podjetju Amebis, d. o. o.

c) Zajem besedil in označevanje korpusnih dokumentov

Zbiranje besedil je potekalo približno pol leta. Avtorske pravice za vključena besedila so pogodbeno zaščitene. Le večinski delež besedil iz *Naše obrambe/Revije Obramba* je bil skeniran (594 datotek),[9] ostala besedila smo dobili v elektronski obliki. Kljub zavedanju, da so podatki o avtorstvu (pri strokovnih besedilih še posebej o avtorjevi avtoriteti), vrsti besedila (znanstveno, poljudnoznanstveno itd.), prevodnosti/neprevodnosti itd. za terminološko analizo korpusa pomembni, smo se zaradi »testnosti« korpusa in časovne stiske odločili, da v glavo dokumentov, ki so v korpusu, damo le podatek o viru (SV, NO) oziroma skupno ime skupine (ZZ) ter letnico izida.

č) Procesiranje zbranega gradiva

Pretvorbo v format XML, segmentacijo, tokenizacijo in lematizacijo korpusa je izvedel Tomaž Erjavec (Institut Jožef Stefan, Erjavec 2003).[10] Zaradi orodij za delo s korpusom, že preizkušenih na Korpusu slovenskega jezika FIDA, smo najeli podjetje Amebis, d. o. o. Danes je mogoče v KSVJG pregledovati konkordančne sezname, dobiti podatke o pogostosti ter o najpogostejših sopojavnicah, izdelan je leksikon lem in frekvenčna lista besed, konkordance lahko urejamo glede na pojavnice za ali pred konkordančnim jedrom, iskalni pogoj je lahko ena ali več (bližnjih) besed, iščemo lahko tudi z nadomestnimi znaki, dobimo ob absolutni pogostosti tudi podatke o vzajemnih vrednostih (MI, MI^3) itd. (o vsem naštetem skupaj s primeri iz korpusa FIDA gl. pri Gorjanc 2005a: 72–86). KSVJG deluje na strežniku Fakultete za družbene vede, kjer je na voljo za brezplačno uporabo z geslom za raziskovalne namene.

d) Končna oblikovanost korpusa in povratne informacije v zvezi z njim

K zgoraj že navedenim podatkom o končni obliki korpusa je treba dodati, da je bolj kot količina pojavnic pri korpusih strokovnih jezikov pomembna kakovost besedil. Za opredeljeni namen lahko pričakujemo največjo uporabnost pri tistem delu korpusa, ki je bil načrtno izbran, se pravi pri *Naši obrambi/Reviji Obramba*.

V naslednjih stopnjah bo treba KSVJG nadgraditi, pri čemer bo treba izhajati iz razširitve njegovega prvotnega namena v namen postati splošni korpus sodobnih

[9] V tej fazi je delo organiziral Janez Jug.

[10] Sodelovanje je bilo opravljeno v okviru raziskovalnega projekta *VoiceTRAN: večjezični prenosni govorni komunikator za bojevnika 21. stoletja* (M2-0019), ki sta ga financirali Ministrstvo za šolstvo, znanost in šport RS in Ministrstvo za obrambo RS, vodila pa ga je Jerneja Žganec Gros (Alpineon, d. o. o.). V okviru nadaljevanja tega projekta bo izvedeno tudi luščenje terminov, ki ga bo opravila Špela Vintar (Filozofska fakulteta; Vintar 2003).

slovenskih vojaških besedil. Kot tak bo moral (bolje) zadostiti merilom homogenosti, ki bodo temeljila na natančnejši razčlenitvi tako jezikoslovnih kot nejezikoslovnih vključevalnih parametrov: treba bo pregledati najnovejše objavljanje slovenskih vojaškostrokovnih besedil, opredeliti najaktualnejša in v zadnjem času intenzivno razvijajoča se področja, tudi sicer uravnotežiti znotrajvojaška predmetnostna področja, dobiti podatke o ciljni publiki in branosti besedil, podatke o vplivnosti avtorjev, premisliti delež prevodov (npr. Natovih dokumentov), najti ustrezno razmerje med poljudnejšimi in ozko specializiranimi besedili, premisliti meje, ki ločijo vojaško stroko od drugih strok, upoštevati razvoj Slovenske vojske in sodelovanje Slovenije v vojaških in mirovnih misijah po svetu itd. Odprto ostaja tudi vprašanje govornega podkorpusa, ki pa bi bil lahko na vojaškem področju dokaj hitro izvedljiv vsaj z naborom povelij, in vključitve besedil iz elektronskega prenosnika.

Letnica 1985, do katere sega KSVJG, se zdi za korpus sodobnih slovenskih vojaških besedil kar skrajna, dopolnjeni del korpusa bi namreč zaradi družbenopolitičnih sprememb, ki so se močno tikale tudi slovenske vojske in vojaške stroke, lahko postavili v čas po letu 1990, besedila do leta 1990 pa bi lahko postala podkorpus.

Primer leksikalne analize terminov *bataljon* in *častnik* – primerjalno z VS 2002:

Bataljon

SSKJ: 1. vojaška enota iz več čet; 2. velika množica

KSVJG: 5.482 pojavitev, vse v terminološkem pomenu[11]

Pogoste stalne besedne zveze (zaporedno navedene po MI[3]):

a) z levim prilastkom: gardni b., artilerijski b., motorizirani b., inženirski b., pehotni b., gorski b., mehanizirani b., oklepnomehanizirani b., učni b., oklepni b., lovski b., tankovski b., raketni b., logistični b., topniški b., izvidniški b.

b) z desnim prilastkom: b. zračne obrambe – b. za mednarodno sodelovanje, b. za zveze, b. za nadzor zračnega prometa

(Izločene lastnoimenske zveze: Štajerski b., Pohorski b., Celjski b., Cankarjev b. itd.)

Kolokacije: eliten b. – poveljnik -a, poveljstvo -a, komandant -a, pripadnik -a, štab -a, premik -a, sestava -a, naloga -a, postroj -a, četa -a, popolnitev -a – b. brigade, b. vojaške policije, b. pehote – ustanoviti b., razpustiti b., ukiniti b.; poveljevati -u

Namen dopolniti geslovnik VS 2002 korpus KSVJG pri terminu *bataljon* izpolnjuje z naslednjim: izmed v KSVJG najpogostejših stalnih besednih zvez v VS 2002 ni terminov *gardni, artilerijski, gorski, mehanizirani, učni, logistični* in *topniški bataljon* ter *bataljon za mednarodno sodelovanje*, *bataljon zračne obrambe* in *bataljon za nadzor zračnega prometa*. Tudi mnogih kolokacij v VS 2002 ni: *poveljstvo, komandant, premik, sestava, naloga, postroj, popolnitev bataljona* ter *bataljon vojaške policije* in *bataljon pehote* ter še *ukiniti bataljon* in *poveljevati bataljonu*.

[11] Za primerjavo: v 100-milijonskem korpusu FIDA je pojavitev različnice *bataljon* 1.473, v veliki večini v pomenu, ki ga SSKJ navaja pod 1.

Velja poudariti, da so v VS 2002 tudi izrazi, ki jih korpus v množici najpogostejših ne izkazuje.[12]

Častnik

SSKJ: član poveljniškega vojaškega osebja; oficir[13]

KSVJG: 5.052 pojavitev[14]

Pogoste stalne besedne zveze (zaporedno navedene po MI^3):

a) z levim prilastkom: poveljujoči č., visoki č., rezervni č., štabni č., izvršilni č., obveščevalni č., dežurni č., poklicni č., sanitetni č., mornariški č.

b) z desnim prilastkom: č. vojnih enot, č. stalne sestave, č. rezervne sestave, č. generalštaba – č. za zvezo/zveze

(Izločeni pogosti lastnoimenski zvezi: Šola za častnike, Združenje slovenskih častnikov.)

Kolokacije: izkušen č., upokojen č., mlad č. – usposabljanje -ov, izobraževanje -ov, povišanje -ov, združenje -ov – status -a

Izmed pogostih stalnih besednih zvez iz KSVJG v VS 2002 manjkajo: *visoki*, *izvršilni*, *poklicni*, *sanitetni častnik* ter *častnik stalne* in *rezervne sestave*, *častnik generalštaba* pa je v VS 2002 *generalštabni častnik*. Manjkajo tudi navedene nestalne besedne zveze, čeprav je v posebnem razdelku VS 2002 registrirano *Združenje slovenskih častnikov*.

5 Sklep

Korpusi strokovnih jezikov so se pojavili precej pozneje kot splošni referenčni, zato se načela za njihovo gradnjo zares šele oblikujejo; glede splošnih postopkov gradnje tudi pri njih veljajo ista izhodišča, a se glede na to, da zajemajo le jezik v točno določeni funkciji, dopolnjujejo in na novo premišljajo. Zaradi dinamike znanstvenega in tehnološkega razvoja pa pomenijo tisto osnovo, ki bo omogočala sprotno spremljanje jezikovnega dogajanja na strokovnih področjih in delovala predvsem na ravni terminološkega usklajevanja, ki je zaradi dinamike razvoja vse težje obvladljivo; pomembni so tudi kot izjemno dragocen vir podatkov o trenutnem védenju.

Dragocenosti tovrstnih virov se pri nas nekatera področja že zavedajo, med njimi je tudi vojaška leksikografija. Prva različica korpusa slovenskih vojaških besedil KSVJG omogoča avtomatsko pridobitev številnih novih terminov, podatkov o njihovi pogostosti, o razvoju v večbesedne termine in o siceršnjem besedilnem okolju itd. Je dobra podlaga za nadaljnji premislek o izboljšavah in nadgradnji v homogen korpus sodobnih slovenskih vojaških besedil. Mnoge rešitve v njem so

[12] Verjetno se ti izrazi z manjšo pogostostjo v korpusu celo pojavijo, vendar je naš namen tu le kratka izpostavitev tistega pogostega izrazja, ki v VS 2002 manjka, se pravi izpostavitev tistega, za kar je KSVJG primarno namenjen.

[13] Normativnost je danes ravno obratna: *oficir* > *častnik*.

[14] Za primerjavo: v kopusu FIDA se *častnik* pojavi 2461-krat.

dobre, korpus je uporaben tudi v pedagoškem procesu, nenazadnje pa ni nepomembno tudi dejstvo, da je pri njegovi izdelavi in uporabi prišlo do produktivnega medprojektnega in interdisciplinarnega sodelovanja več institucij, ki ga bo v prihodnje pri tovrstnih projektih smiselno le še krepiti.

Literatura

Atkins, Sue, Clear, Jeremy, Oster, Nicholas, 1992: Corpus Design Criteria. *Literary and Linguistics Computing* 7/1. 1–16.

Biber, Douglas, 1993: Representativeness in Corpus Design. *Literary and Linguistics Computing* 8/4. 243–257.

Biber, Douglas, Conrad, Susan, Reppen, Randi, 1998: *Corpus Linguistics: Investigating Language Structure in Use*. Cambridge: Cambridge University Press.

Bowker, Lynne, 1996: Towards a Corpus-Based Approach to Terminography. *Terminology* 3/1. 27–52.

Erjavec, Tomaž, 1998: Standardizacija zapisa jezikovnih podatkov. *Jezikovne tehnologije za slovenski jezik/Language Technologies for the Slovene Language*. Zbornik konference. Ur. T. Erjavec, J. Gros. Ljubljana: Institut Jožef Stefan. 119–123.

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

Erjavec, Tomaž, 2003: Označevanje korpusov. *Jezik in slovstvo* 3–4. 61–76.

Gorjanc, Vojko, 2005a: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Gorjanc, Vojko, 2005b: V mavrici jezikovnih podatkov. *Študije o korpusnem jezikoslovju*. Ur. V. Gorjanc, S. Krek. Ljubljana: Krtina. 173–194.

Kennedy, Graeme, 1998: *An Introduction to Corpus Linguistics*. London, New York: Longman.

Korošec, Tomo et al., 2002: *Vojaški slovar*. Ljubljana: Ministrstvo za obrambo.

Leech, Geoffrey, 1991/2005: The State of the Art in Corpus Linguistics. *English Corpus Linguistics*. Ur. K. Aijmer, B. Altenberg. London, New York: Longman. / Stanje stvari v korpusnem jezikoslovju. *Študije o korpusnem jezikoslovju*. Ur. V. Gorjanc, S. Krek. Ljubljana: Krtina. 29–57.

Meyer, Ingrid, Mackintosh, Kristen, 1996: The Corpus from a Terminographer's Viewpoint. *International Journal of Corpus Linguistics* 1/2. 257–285.

Pearson, Jennifer, 1998: *Terms in Context*. Amsterdam: John Benjamins.

Vintar, Špela, 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Spletni strani

Korpus slovenskega jezika FIDA. Http:// www.fida.net.

Korpus slovenskega vojaškega jezika Grizold. Http://ksvjg.fdv.uni-lj.si.

Priloga: Besedila, zajeta v Korpusu slovenskega vojaškega jezika Grizold.

1) Naša obramba/Revija Obramba 1985–1997, 2000, 2001, 2002 (izbor)
2) Slovenska vojska 1997–2004 (celota)
3) Druga besedila (celota):
   - Dušan Gorše: Sporazum o konvencionalnih silah v Evropi. BILTEN SLOVENSKE VOJSKE, letnik 4, št. 2, november 2002. 75–95.
   - DOKTRINA CIVILNE OBRAMBE REPUBLIKE SLOVENIJE. Ljubljana: Ministrstvo za obrambo Republike Slovenije, Sekretariat generalnega sekretarja, Služba za publicistiko, 2002.
   - Boštjan Lesjak, Vasilije Maraš, ur.: POSTROJITVENA PRAVILA SLOVENSKE VOJSKE. Ljubljana: Ministrstvo za obrambo, Center vojaških šol, 2004.
   - VOJSTVO, november 2000.
   - Karlo Nanut, ur.: VOJAŠKA ZGODOVINA 6 (zbornik). Ljubljana: Generalštab Slovenske vojske, Vojaški muzej, 2003.
   - Karlo Nanut, ur.: VOJAŠKA ZGODOVINA 7 (zbornik). Ljubljana: Generalštab Slovenske vojske, Vojaški muzej, 2004.
   - Karlo Nanut, ur.: VOJAŠKA ZGODOVINA 8 (zbornik). Ljubljana: Generalštab Slovenske vojske, Vojaški muzej, 2004.
   - Marko Unger, Radovan Lukman, Anže Rode, Iztok Beslič: TAKTIKA: Skripta. Ljubljana: Ministrstvo za obrambo, Center vojaških šol, 2004.
   - Tomaž Pörš, ur.: VOJAŠKOŠOLSKI ZBORNIK 1, Ljubljana: Generalštab Slovenske vojske, Center vojaških šol, 2003.
   - Tomaž Pörš, ur.: VOJAŠKOŠOLSKI ZBORNIK 3. Ljubljana: Generalštab Slovenske vojske, Center vojaških šol, 2004.
   - VARNOSTNO-POLITIČNI DIALOG SLOVENIJA – AVSTRIJA. Ljubljana: Ministrstvo za obrambo Republike Slovenije, Urad za obrambo politike, Center za strateške študije, 8.–9. oktober 2001. Zbornik referatov, pripravljen za tisk, vendar ni izšel.
   - Milan Frankovič, Rajko Najzer: CIVILNA OBRAMBA V REPUBLIKI SLOVENIJI: Brošura. Ljubljana: Ministrstvo za obrambo Republike Slovenije, Sektor za civilno obrambo, 2005.
   - STRATEŠKI PREGLED OBRAMBE: Strnjen povzetek. Ljubljana: Ministrstvo za obrambo Republike Slovenije, 2004.
   - VARSTVO OKOLJA V SLOVENSKI VOJSKI: Skripta. Ljubljana: Generalštab Slovenske vojske, Poveljstvo za doktrino, razvoj, izobraževanje in usposabljanje, 2004.