Branko Tošović
Univerzitet „Karl Franc“
Filozofski fakultet
Institut za slavistiku
Grac, Austrija

BIBLID: 1450–5061, 18 (2014), pp. 207–214
УДК 811.163.4’362:004.9
прегледни научни рад
примљено 15. 03. 2014.
прихваћено за штампу 1. 10. 2014.

AUTOMATSKO KODIRANJE POMOĆU MORFOGENERATORA

U radu se prezentira program MorfoGenetrator za automatsko kodiranje morfoloških oblika srpskog, hrvatskog i bošnjačkog jezika u cilju dobijanja kompletnih paradigmi potrebnih u edukaciji i za pretraživanja u Gralis-Korpusu po gramatičkim parametrima.

Ključne riječi: MorfoGenerator, Gralis, koprus, morfologija, obilježavanje, kodiranje

MorfoGenerator predstavlja online analitičko-sintetički sistem za morfološko obilježavanje riječi radi generisanja paradigmi i pretraživanja po kodnim pozicijama u Gralis-Korpusu – višejezičkom paralelnom korpusu za proučavanje slovenskih jezika utemeljenom 2007. na Univerzitetu „Karl Franc“ u Gracu (Gralis-Korpus-www).

Korpus je namijenjen prvenstveno za kontrastivno proučavanje slovenskih jezika. Prema vrsti obrade on spada u anotirane korpuse. Predviđene su sljedeće vrste obilježavanja: metajezička, morfološka, sintaksička, semantička, stilistička (dvije od njih – metajezička i morfološka već postoje).

Postoje dva morfološka pristupa vrstama riječi: (1) klasični, tradicionalni, zasnovan na izdvajanju opštih tipova promjene i (2) morfogeneratorski, zasnovan na utvrđivanju pravila za obrazovanje paradigmi sa svim međusobnim odnosima, alternacijama i varijantama (dubletima, tripletima…). Prednost klasičnog pristupa (deduktivnog: opšte → pojedinačno) jeste u tome što se sve lekseme u okviru jedne vrste riječi morfološki raščlanjuju na minimalni broj vrsta (recimo, imenice na pet vrsta, glagoli na osam vrsta i sl.), ali mu je mana što se na osnovu takvog ograničenog broja tipa ne mogu automatski proizvesti paradigme svih leksema date vrste riječi. Za potrebe morfološkog obilježavanja elektronskog koprusa i edukaciju (da bi se dobili potpuni, konačni i kompletni obrasci za sve glagole) neophodno je izdvojiti minimalan broj morfoloških pravila. Prednost generatorskog pristupa je sveobuhvatnost i preciznost, jer se na osnovu njega mogu generisati paradigme za sve lekseme koje pripadaju datoj vrsti riječi. Morfogeneratorski pristup je zasnovan na indukciji (pojedinačno → opšte: svaka riječ se pojedinačno obrađuje paradigmatski, a zatim uvrštava u određeno pravilo). Njegova je vrijednost u tome što daje korektne i kompletne paradigme, ali mu je nedostatak što je broj pravila znatno veći u odnosu na broj tipova u klasičnoj tipologiji (međutim, u daljoj obradi može da se izvrši kategorijalno sužavanje na što manji broj opštih pravila). Ali se zato, za razliku od klasičnog (tradicionalnog) pristupa, može iskoristiti za automatsko morfosintaksičko obilježavanje i pretraživanje.

Jedan od najvećih problema u automatskom generisanju oblika čine mnogobrojne alternacije. Rad na MorfoGeneratoru je pokazao da postoji monoalternacije, bialternacije i polialternacije. Monoalternacije dolaze u slučajevima pojave samo jednog oblika sa glasovnom promjenom (recimo od ***bi**ti* je ***bij**em* u 1. licu jednine prezenta,

*bi**j**eš* u 2. licu jednine, *bi**j**e* u 3. licu jednine itd.). Bialternacije su dubletne, dvostruke alternacije koje se javljaju kada postoje dva različita gramatička lika u istoj paradigmatskoj jedinici (padežu, licu, stepenu komparacije), od kojih oba mogu biti neutralna ili je jedan od njih stilski obilježen (kao razgovorni, regionalni, dijalekatski, žargonski...). Tako od glagola *moći* imamo u 1. i 2. licu jednine sadašnjeg vremena: *mo**g**u*, *možeš.* Polialternacija podrazumijeva pojavu trostruke ili višestruke alternacije u istoj paradigmatskoj jedinici. Ona može biti, ter-, kvatro-, kvinto-... alternacija. Teralternacija (trostruko, tripletno variranje) nastaje u slučajevima postojanja triju varijanta istog oblika u istoj paradigmatskoj jedinici. Npr. u 1. licu jednine sadašnjeg vremena od glagola *i**sk**ati* imamo *i**št**em*, *i**šć**em*, *i**sk**am*. Za kvatroalternaciju (četvorostruko variranje) nemamo primjere, ali postoje slučajevi kada u istoj paradigmatskoj jedinici postoje četiri varijantska oblika, ali samo sa dvije alternacije, recimo u 1. licu prezenta glagola *ht*(*j*)*eti* sa odričnim i potvrdnim oblicima: *ho**ć**u*, ***ć**u*, *ne**ć**u*, *ht*(*j*)*e**dn**em*. Kvintoalternacija (petorostruka alternacija) nastaje u slučajevima kada se u istom paradigmatskoj jedinici pojavljuje pet varijanata istog oblika. Takav je slučaj sa pomoćnim glagolom *biti*/*jesam* u 1. licu jedn. sad. vr.: *jesam*, *sam*, *nisam*, *nijesam*, *budem*. Ovdje bi spadao i glagol *ht*(*j*)*eti* ukoliko bismo kao jednu paradigmu računali njegove (a) odrične i neodrične, (b) ekavske i ijekavske likove. Nisu nam poznati slučajevi formalne raznovrsnosti veće od pet u okviru iste paradigmatske jedinice.

Po tome da li dolaze u istom ili različitom glagolskom obliku razlikujemo: 1) morfološki homogene alternacije – glasovne promjene u istom glagolskom obliku, 2) morfološki heterogene alternacije – glasovne promjene u različitim glagolskim oblicima, npr. infinitivno- prezentsku: *moći – mo**g**u*.

Paradigmatski alternacija može biti kompletna i parcijalna. U kompletnoj ona se javlja u čitavoj paradigmi: *naći – na**đ**em*, *na**đ**eš*, *na**đ**e*; *na**đ**emo*, *na**đ**ete*, *na**đ**u*, dok u djelimičnoj zahvata samo određeni segment paradigme (npr. jedninu ili množinu) ili njen elemenat (recimo, pojedina lica ili padeže), npr. 2. i 3. licu jednine i 1. i 2. licu množine: ***mogu***, *možeš*, *može*; *možemo*, *možete*, ***mogu.***

Prema tipu kategorije postoje **a)** intrakorelacione alternacije – glasovne promjene u okviru iste glagolske kategorije (npr. vremena), iste glagolske potkategorije (npr. prošlih vremena) i iste glagolske kategorijalne jedinice (npr. aorista) sa dvije mogućnosti – u formi glasovne promjene u različitim licima, npr. prezenta: *vući*: *vučem*, *vučeš*, *vuče – vučemo*, *vučete*, *vu**k**u*, glasovne promjene u istom licu (dubletni oblici), recimo aorista: oni *re**k**oše*, *re**kn**uše*, **b)** interkorelacione alternacije – glasovne promjene u okviru iste kategorije (npr. vremena), iste potkategorije (npr. prošlih vremena), ali različitih kategorijalnih jedinica (npr. perfekta, aorista i imperfekta), recimo: one su *re**k**le*, one *re**k**oše*, one *re**c**ijahu*, **c)** suprakorelacione alternacije – glasovne promjene u okviru iste kategorije (npr. vremena), različite potkategorije (npr. sadašnjih i prošlih vremena) i različitih kategorijalnih jedinica (npr. prezenta i perfekta (*di**gn**em*, *di**gn**eš – di**ga**o*, *di**g**la*), **d)** superkorelacione alternacije – glasovne promjene u okviru oblika različitih kategorija (npr. indikativa i imperativa): infinitiva i prezenta: *ht**je**ti – ho**ć**u*, vremena i glagolskog pridjeva radnog (*di**gn**em*, *di**gn**eš – di**ga**o*, *di**g**la*), vremena i glagolskog načina (*pe**c**i – pe**č**en*) i sl., **e)** ekstrakorelacione alternacije – glasovne promjene u okviru različitih vrsta riječi (npr. palatalizacija u konjugaciji i deklinaciji: *pe**ć**i – pe**c**i* : *ru**k**a – ru**c**i*).

Morfološka anotacija za Gralis-Korpus vrši se pomoću analitičko-sintetičkog sistema „MorfoGenerator“. Prilikom pretraživanja u korpusu njeni se rezultati prika-

zuju u obliku podvučenih riječi. Ako se klikne na bilo koju od njih, otvara se prozor sa paradigmom i drugom informacijom.

Morfološko anotiranje započeto je aprila 2008. u okviru projekta „Razlike između srpskog, hrvatskog i bosanskog/bošnjačkog jezika" (FWF-Projekt, P19158-G03, 2006–2010) austrijskog Fonda za podršku naučnih istraživanja, a nastavljeno u okviru projekta „Gralis-Lexikarium", koji je finansirala Pokrajinska vlada Štajerske (2008–2013). Lingvistički sistem anotiranja i proces generisanja razradio je Branko Tošović, a programski paket na bazi lingvističkog modela razvila Olga Lehner.

Budući da Gralis-Korpus obuhvata niz jezika, bilo je neophodno izabrati ono gramatičko obilježavanje koje bi najviše odgovaralo za njihovo konstrastiranje i proučavanje. Koncepciji Korpusa najbliže je bilo Multext-East kodiranje (Multilingual Texts and Corpora for Eastern and Central European Languages – multilingual dataset for language engineering research and development: MultiText East-www), razrađeno 2004. od strane grupe autora na čelu sa Tomažem Erjavcem budući da je bilo prilagođeno za većinu slovenskih jezika. Ono se odnosi na klasične vrsta riječi, skraćenice i tzv. rezidual (Residual).[1]

Za pretraživanje u Gralis-Korpusu na osnovu gramatičke anotacije koristi se CQP-sintaksa, koja nudi široke i raznorodne mogućnosti. Pretraživanje se može vršiti i prema leksičkim spojevima. Ono može da se ograniči na dvije ili više riječi (do devet). Uparivanje jezika, izbor širine konteksta, izbor spiska riječi ili spiska rečenica sa datom riječju, a takođe formata (HTML ili Excell tabele) vrši se u posebnoj maski. Morfosintaksička anotacija može da se otvori (pomoću kvačice) ili da se sakrije.

Gramatičko kodiranje za Gralis-Koprus sastoji se od sljedećih pozicija: 1. vrsta riječi, 2. podtip vrste riječi, 3. glagolski način, 4. vrijeme, 5. lice, 6. broj, 7. rod, 8. dijateza, 9. negacija, 10. određenost, 11. refleksivnost, 12. padež, 13. kategorija živog, 14. klitika, 15. vid, 16. etikecija, 18. akcenat, 19. način vršenja glagolske radnje, 20. leksičko-semantička grupa, 21. spojivost (rekcija, kongruencija), 22. ekspresija, 23. funkcionalni stil, 24. analitički oblik, 25. destrukcija, 26. derivacija, 27. broj pravila, 28. gramatički tip generisanja oblika.

Na osnovu opšteg modela stvorena je konkretna shema sa 20 pozicija koja obuhvata sve vrste riječi i njihove kategorije. Za oznaku taksona koriste se u svakoj poziciji mala slova (samo za ukazivanje na lice i pravilo daju se brojke). Pošto ima više pozicija nega slova, neke grafeme se ponavljaju, ali ne nastaje zbrka budući da prvu poziciju uvijek zauzima vrsta riječi, što onemogućuje pojavu dvosmislenosti:

1. **Vrsta riječi n** – imenice, **v** – glagoli, **a** – pridjevi, **p** – zamjenice, **r** – prilozi, **s** – prijedlozi, **c** – veznici, **m** – brojevi, **i** – uzvici, **q** – riječce, **y** – skraćenice

2. **Podtip vrste riječi –** I m e n i c e: **c** – zajedničke, **p** – vlastite, **m** – gradivne, **l** – zbirne. G l a g o l i: **m** – sa punim leksičkim značenjem, **a** – pomoćni, **o** – modalni, **c** – kopulativni, **b** – bazni. P r i d j e v i: **f** – opisni, **r** – odnosni, **m** – gradivni, **s** – prisvojni, **o** – redni, **m** – količinski. Z a m j e n i c e : **p** – lične, **d** – pokazne, **i** – neodređene, **s** – prisvojne, **q** – upitne, **r** – odnosne, **x** – povratne, **z** – odrične, **g** – opšte, **y** – upitno-odnosne, **j** – određene, **t** – pokazno-odnosne. P r i l o z i: **g** – opšti, **z** – odrični, **a** – pridjevski, **v** – glagolski, **q** – upitni. P r i j e d l o z i: **p** – prepozitivni, **t** – postpozitivni. V e z n i c i : **c** – naporedni, **s** – zavisni. B r o j e v i: **c** – glavni, **o** – redni, **m**

[1] Pod time se podrazumijeva ostatak, količina koja ostaje nakon nekog procesa, događaja ili neke radnje.

– zbirni, **l** – multiplikativni, **s** – posebni. R i j e č c e: **z** – odrične, **q** – upitne, **o** – modalne, **r** – potvrdne. S k r a ć e n i c e: **n** – imenske, **r** – priloške

3. **Tip oblika –** G l a g o l i: **i** – indikativ, **m** – imperativ, **c** – kondicional/potencijal 1, **h** – kondicional/potencijal 2, **n** – infinitiv, **p** – particip, **g** – glagolski prilog 1 (sadašnjeg vremena), **w** – glagolski prilog 2 (prošlog vremena), **u** – supin, **t** – prelazni, **q** – citirani, **s** – hipotetički. P r i d j e v i: komparacija – **p**: pozitiv, **c** – komparativ, **s** – superlativ. P r i l o z i: komparacija – **p**: pozitiv, **c** – komparativ, **s** – superlativ, **e** – elativ

4. **Vrijeme**: **p** – prezent, **i** – imperfekt, **f** – futur I Sr (srpski), **w** – futur I Hr (hrvatski), **z** – futur I Sr/Hr (srpski i hrvatski), **q** – futur II, **s** – perfekt, **l** – pluskvamperfekt 1, **t** – pluskvamperfekt 2, **a** – aorist

5. **Lice**: **1** – prvo, **2** – drugo, 3 – treće

6. **Broj**: **s** – jednina, **p** – množina, **d** – dual, **l** – zbirnost

7. **Rod**: **m** – muški, **f** – ženski, **n** – srednji, **l** – opšti

8. **Dijateza**: **a** – aktiv, **p** – pasiv

9. **Negacija**: **n** – da, **y** – ne

10. **Određenost**: **n** – da, **y** – ne

11. **Refleksivnost**: **n** – da, **y** – ne

12. **Padež**: **n** – nominativ, **g** – genitiv, **d** – dativ, **a** – akuzativ, **v** – vokativ, **i** – instrumental, **l** – lokativ

13. **Kategorija živog:** – da, **y** – ne

14. **Klitika**: **n** – da, **y** – ne

15. **Vid**: **p** – nesvršeni, **e** – svršeni, **b** – dvostruki

16. **Etikecija**: **n** – da, **y** – ne

17. **Prelaznost**: **n** – da, **y** – ne

18. **Destrukcija**: **n** – da, **y** – ne

19. **Derivacija**: **s** – nemotivirani, **c** – složeni.

20. **Broj pravila**: 01, 02, 03…

Nepopunjene pozicije su za sada: 11. refleksivnost, 16. etikecija, 18. destrukcija, 19. derivacija. Od 28 planiranih pozicija nedostaju sljedeće: 18. akcenat (koje se daje u Akcentarijumu[2]), 19. način vršenja glagolske radnje, 20. leksičko-semantička grupa, 21. spojivost (rekcija, kongruencija), 22. ekspresija, 23. funkcionalni stil, 24. analitički oblik.

Pomoću MorpfoGeneratora može se dobiti svaka paradigma, što je posebno važno u procesu obrazovanja. Međujezičke razlike u frekvenciji upotrebe ne određuju se paušalno, u opštim formulacijama, već objektivno pomoću korpusnih.

U MorpfoGeneratoru može se izabrati za pretragu pojavnica ili lema, kraj ili početak riječi, njen segment, određena vrstu riječi. MorpfoGenerator nudi takođe obrnuto sortiranje i podatke o frekvenciji u tri jezika (srpskom, hrvatskom, bošnjačkom). Za imenice, pridjeve i glagole, koji se odlikuju širokim sistemom promjene (u obliku deklinacije, konjugacije i komparacije), pripremljene su tri odvojene maske.

MorpfoGenerator sadrži takođe statističku informacija o svakoj vrsti riječi u Gralis-Korpusu.

Morfološka anotacija za srpski, hrvatski, bošnjački i crnogorski jezik izvršena je

[2] To je program pomoću koga se može naći akcenat bilo koje riječi (http://www-gewi.uni-graz.at/gralis-alt/php/en/Akzentarium/suche.php). Postoji direktna veza između MorfoGeneratora i Akcentarijuma.

za 100.461 riječi. Njihove paradigme generisane su pomoću 822 pravila.[3]

Postupak anotiranja izgleda ovako.

a) Priprema se spisak svih riječi u okviru jednog standardnog jezika (u slučaju Gralis anotacije to srpski, hrvatski i bošnjački).

b) Riječi se dijele na vrste riječi.

c) U okviru svake vrste riječi lekseme se objedinjavaju u poseban spisak po gramatičkom obilježju važnom za generisanje paradigme (za imenice kategorija roda, za pridjeve komparacija, za glagole vid i sl.).

d) Dobijeni spisak se razbija na podspiskove, uzimajući u obzir dopunsko gramatičko obilježje (za imenice to je kategorija živog, za glagole prelaznost itd.).

Time se završava priprema leksičke građe za njenu gramatičku obradu.

e) Vrši se analiza spiska pod **d** i utvrđuju paradigmatski markeri (padeži, lica i sl.) relevantni za generisanje kompletnih paradigmi (npr. za imenice muškog roda na suglasnik to je nominativ, genitiv, vokativ i instrumental jednine te nominativ i genitiv množine).

f) Pravi se spisak riječi sa opštim gramatičkim markerima (recimo za imenice muškog roda na suglasnik koje imaju u nominativu jednine pokretno **a**).

g) Bira se tipična riječ u toj grupi i za nju izrađuje tabela sa svih 20 pozicija i svim oblicima. Po toj riječi i broju tipa naziva se svako pravilo (npr. 174}$^{\text{mudrac}}$).

Time se završava lingvistički rad i počinje softverski.

h) Uzima se materijal dobijen u procedurama **a** – **g** i pretvara u relacionu bazu podataka MySQL. Pri tome se izrađuju dvije tabele. Jedna se sastoji od četiri stupca u kojima se nalaze leme, njihovi nepromjenljivi dijelovi, važno gramatičko obilježje (npr. za imenice kategorija za živo, za glagole prelaznost), broj lingvističkog pravila, recimo:

obnemoći	**obnemo**	n	01e
dići	**di**	y	03e
dolivati	**dol**	y	24

Osnovu druge tabele čini tabela sa kodovima, kodnim pozicijama i paradigmom za određeno pravilo, ali se razlikuje od prve novim stupcem sa nepromjenljivim segmentima riječi.

i) Ove dvije tabele pretvaraju se u format **.csv**. U novim tabelama sadržaj stubaca razdvaja se tačkom i zarezom:

šetati;še;y;45;
plesati;ple;y;48;
platiti;pla;y;90e;

Podaci se unose u MySQL u kojoj se generišu oblici i paradigme. Recimo MySQL tabela sa krajem glagola sadrži 378 pravila/tipova i više od 45.000 redova. Kao rezultat nastaje zbirna tabela za sve vrste riječi.

j) Pripremaju se dvije maske – jedna za pretragu informacije prema gramatičkoj anotaciji i druga za odražavanje rezultata pretrage.

[3] Pod pravilom se podrazumijeva matrica saodnosa između kodnih pozicija i paradigmatskih punjenja, tačnije procedura razotkrivanja sistema promjenljivosti riječi, njihovog variranja i prezentiranja u cilju generisanja svih oblika i cjelovitih paradigmi.

k) MorfoGenerator se povezuje sa Gralis-Korpusom, čime se stvaraju informacioni kanali u dva pravca – od MorfoGeneratora prema Korpusu i od Korpusa prema MorfoGeneratoru s ciljem da se bilo koji oblik naveden u MorfoGeneratoru može naći u Korpusu kako bi se dobila informacija o njegovoj frekvenciji u Korpusu i paradigma čiji je on član.

Izvršeno morfološko anotiranje ukazalo je na sljedeće zakonomjernosti. P r v o, postoje dva suprotna pravila generisanja oblika i paradigmi: a) pravila pomoću kojih se može dobiti paradigma samo za jednu riječ, b) pravila za generisanje paradigmi za stotine i na hiljade riječi. D r u g o, izdvajaju se opšta i pojedinačna pravila. Prva se tiču tipa promjene (deklinacije, konjugacije, komparacije), druga se odnose na neke gramatičke kategorije. T r e ć e, postoje strukturna, kategorijalna i interkategorijalna pravila. Strukturna pravila ukazuju na to kakva se formalna sredstva koriste za generisanje paradigmi i do kakvih alternacija dolazi na spoju nastavka i prethodne morfeme. Kategorijalna pravila pokrivaju morfološku specifičnost u okviru određene gramatičke kategorije (roda, broja, vida i sl.). Č e t v r t o, na pojavu velikog broja pravila ne utiče veliki broj nastavaka, što potvrđuje i činjenica da u generisanju imenskih paradigmi učestvuje svega 11 fleksija **(-a**, **-o**, **-e**, **-i**, **-u**, **-ø**, **-oj**, **-om**, **-em**, **-ama**, **-ima**), koje se ponavljaju u obliku sinkretizma. P e t o, broj i struktura pravila zavisi od svake vrste riječi. Najviše pravila zahtijevaju glagoli (378), zatim imenice (311), mnogo manje pridjevi (71) i zamjenice (50), a na posljednjem mjestu su brojevi (12). Međutim, ako se broj pravila dovede u vezu sa brojem leksema, dobiće se sasvim druga slika: 135 zamjenica traži 50 pravila, 198 brojeva 12, dok za 37.606 imenica treba 311, za 30.030 glagola 378, za 32.492 pridjeva 71.

Br.	Vrsta riječi	Broj riječi	%	Broj pravila	Saodnos broja riječi broja pravila
1.	Imenice	37.606	37,43	311	0,0083
2.	Pridjevi	32.492	32,34	71	0,0022
3.	Glagoli	30.030	29,89	378	0,0126
4.	Brojevi	198	0,20	12	0,0606
5.	Zamjenice	135	0,13	50	0,3704
Svega		100.461	100,00	822	0,0052

Prema odnosu broja riječi i broja pravila najniži parametar nalazimo kod pridjeva (0,0022) i imenica (0,0083), viši kod zamjenica (0,37), a najviši kod brojeva (0,6%). Za anotiranje 37.606 imenica treba imati 311 pravila (gotovo da nema imenica bez deklinacije), koja su složena zbog raznovrsnosti gramatičkih kategorija (sedam padeža, dva broja, tri roda, kategorija živog), raznorodnih alternacija i postojanja varijantnih oblika. Paradigme pridjeva (32.492) generišu se pomoću 71 pravila i obuhvataju kategorije karakteristične za imenice (rod, broj, padež) i pridjeve (komparaciju i određenost). Zamjenice čine zatvoreni skup sastavljen od 135 jedinica. Za generisanje njihove paradigme potrebno je imati 50 pravila. Druga vrsta riječi koja čini zatvoreni skup – brojevi (198) traže 12 pravila. Što se tiče glagola, njihova pravila su

najsloženija pošto obuhvataju više kategorija nego druge vrste riječi (glagolski način, vid, lice, dijatezu, rod, konjugaciju, deklinaciju, prelaznost, refleksivnost, kategoriju živog i dr.) i imaju širok sistem alternacija i varijativnosti. Stoga 30.030 glagola prikrivaju 378 pravila. Š e s t o, stvaranje i postojanje triju različitih normi za tri veoma bliska jezika – srpski, hrvatski i bošnjački (razumijevanje između njihovih nosilaca u svakodnevnoj komunikaciji je gotovo stopostotno) na istoj dijalekatskoj osnovi – štokavskoj zahtijeva da se uzimaju u obzir sve standardološke razlike, kojih nema tako mnogo ali koje ipak opterećuju morfološko obilježavanje. Zbog specifičnosti gramatička anotacija za Gralis-Korpus je unikalna (nisu nam poznati slučajevi kada anotacija obuhvata istovremeno tri jezika).

Što se tiče daljeg rada, on će biti nastavljen u četiri faze. **(1)** Pošto je u periodu od 2008. do 2013. urađena morfološka anotacija za 100.461 promjenljivih vrsta riječi i za oko 11.000 nepromjenljivih (ukupno oko 112.000), predstoji provjera stvorenog sistema, posebno pravila i rezultati generisanja, korektura svega onoga što se pokazano slabim, pogrešnim, nepreciznim, nakon čega će MorfoGenerator biti otvoren svima za korišćenje. **(2)** Uslijediće rad na eliminisanju gramatičke homonimije. **(3)** Nakon toga dolazi morfološko anotiranje drugih slovenskih jezika (u saradnji sa stručnjacima iz drugih zemalja da bi se postojeći sistemi za obilježavanje pojedinih jezika modifikovali za Gralis-Korpus i, u slučaju kompatibilnosti, potpuno preuzeli). **(3)** Na kraju će **se** pristupiti izradi modela sintaksičke, semantičke i stilističke anotacije.

IZVORI I UPUTNA LITERATURA

Gralis-Korpus <http://www-gewi.uni-graz.at/gralis/korpusarium/gralis_korpus.html> 10.10.2013.

Tošović B., „Das Gralis-Korpus“, in Tošović B. (Hg.), *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen* (Graz, 2008), 724–749.

Tošović B., «Гралис-Корпус», *Wiener slawischer Almanach* (2013), 83, 89–111.

Тошович Б. «Сопоставительное изучение славянских языков при помощи многоязычного „Гралис-Корпуса“», в. Станковић Б. (ред.), *Изучавање словенских језика, књижевности и култура као инословенских и страних* (Београд: Славистичко друштво Србије, 2008), 336–340.

Бранко Тошович

АВТОМАТИЧЕСКАЯ КОДИРОВКА ПРИ ПОМОЩИ «МОРФОГЕНЕРАТОРА»

Резюме

В настоящей статье представлена онлайн программа «Морфогенератор» для автоматической кодировки материала в Гралис-Корпусе. Он является параллельным корпусом для исследования славянских языков, созданным в 2007 году в Университете им. Карла и Франца в Граце. Корпус предназначен в первую очередь для сопоставительного изучения славянских языков. По разметке он относится к аннотированным корпу-

сам. Предусмотрены следующие типы его разметки: метаязыковая, морфологическая, синтаксическая, семантическая и стилистическая, две из которых (метаязыковая и морфологическая) уже существуют. Морфологическая аннотация проведена для сербского, хорватского и бошняцкого языков при помощи аналитическо-синтетической системы „MorphoGenerator“. Работа над морфологической аннотацией началась весной 2008 года в рамках проекта „Различия между боснийским/бошняцким, сербским и хорватским языками“ (2006–2010) австрийского Фонда для поддержки научных исследований и продолжилась в рамках проекта „Lexikarium“, который финансировало Правительство Штирии (2008–2013). Грамматическая кодировка для Гралис-корпуса состоит из следующих позиций: 1. часть речи, 2. подтип части речи, 3. наклонение, 4. время, 5. лицо, 6. число, 7. род, 8. залог, 9. отрицание, 10. определенность, 11. возвратность, 12. падеж, 13. одушевленность, 14. клитика, 15. вид, 16. вежливость, 17. переходность, 18. ударение, 19. способ глагольного действия, 20. лексико-семантическая группа, 21. сочетание (управление, согласование), 22. экспрессия, 23. функциональный стиль, 24. аналитическая форма, 25. деструкция, 26. словообразование, 27. номер порождения, 28. грамматический тип порождения.

Ключевые слова: МорфоГенератор, Гралис, морфология, разметка, кодировка