RASPRAVE
Časopis Instituta za hrvatski jezik i jezikoslovlje 42/1 (2016.)

UDK 811.163.42'32
UDK 811.163.42-112
UDK 801.81
UDK 801.82
Pregledni rad
Rukopis primljen 16. III. 2016.
Prihvaćen za tisak 29. IV. 2016.

Ernst Hansack, Björn Hansen, Veronika Wald
Universität Regensburg
Fakultät für Sprach-, Literatur- und Kulturwissenschaften
Institut für Slavistik
Universitätsstraße 31, D-93053 Regensburg
ernst.hansack@sprachlit.uni-regensburg.de
Bjoern.Hansen@sprachlit.uni-regensburg.de
Veronika1.Wald@sprachlit.uni-regensburg.de

Marijana Horvat
Sanja Perić Gavrančić
Institut za hrvatski jezik i jezikoslovlje
Ulica Republike Austrije 16, HR-10000 Zagreb
mhorvat@ihjj.hr
speric@ihjj.hr

REGENSBURŠKI DIJAKRONIJSKI KORPUS HRVATSKOGA JEZIKA – CRODI

U članku se opisuju ciljevi i načela uspostave korpusa, metodologija izrade i dosadašnji rezultati u okviru projekta izgradnje dijakronijskoga korpusa hrvatskoga jezika, koji su za potrebe jezičnopovijesnih istraživanja pokrenuli njemački slavisti Roland Meyer (Humboldt-Universität zu Berlin) i Björn Hansen (Universität Regensburg) u suradnji s Institutom za hrvatski jezik i jezikoslovlje. Regensburški dijakronijski korpus hrvatskoga jezika u vrijeme pisanja ovoga rada još nije mrežno dostupan za javnost.

1. O hrvatskim korpusima

Korpusna lingvistika zasebna je jezikoslovna grana koja se bavi jezičnom raščlambom strojno izrađenih korpusa pisanoga ili govorenoga jezika, a korpus je skup jezičnih odsječaka (ne nužno i cijelih tekstova) odabranih i skupljenih prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak (Klobu-

čar Srbić 2008: 39). Ne namjeravajući ovdje problematizirati mogućnosti i dosege hrvatske korpusne lingvistike ni iznositi iscrpan popis hrvatskih korpusa,[1] osvrnut ćemo se na neke postojeće korpuse. Zavod za lingvistiku Filozofskoga fakulteta u Zagrebu 1996. godine pod vodstvom Marka Tadića započeo je sastavljanje *Hrvatskoga nacionalnoga korpusa* (HNK), što je ususstavljena zbirka odabranih tekstova pretežito suvremenoga hrvatskoga jezika koji pokrivaju razne medije, žanrove, stilove, područja i tematiku (v. http://www.hnk.ffzg.hr/).[2] Na istome fakultetu uspostavljen je i veliki obilježeni mrežni korpus hrWaC 2.0 Nikole Ljubešića (v. http://nlp.ffzg.hr/resources/corpora/hrwac/). U Institutu za hrvatski jezik i jezikoslovlje 2005. godine pokrenut je projekt *Hrvatska jezična riznica* (voditeljice Dunje Brozović Rončević). *Riznica* se sastoji od nekoliko potkorpusa, a u nju su uključeni i određeni dopreporodni tekstovi. Aktivna je na adresi http://riznica.ihjj.hr/, a sadašnji je voditelj Željko Jozić.

Iako se posljednjih tridesetak godina 20. stoljeća znatno radilo na računalno potpomognutim korpusima dopreporodnih tekstova hrvatskoga jezika,[3] što je rezultiralo izradom konkordancija djela nekoliko starih pisaca,[4] Hrvatska još uvijek nema jedinstven i mrežno dostupan dijakronijski korpus hrvatskoga jezika. Određeni pomaci u smjeru izgradnje takva korpusa ostvaruju se u Institutu za hrvatski jezik i jezikoslovlje, u kojemu se korpusi dopreporodnih tekstova izrađuju u okviru projekta *Starohrvatski rječnik i računalni korpus hrvatskoga jezika do konca 16. stoljeća* (voditelja Amira Kapetanovića) i projekta *Korpus hrvatskoga jezika 17., 18. i 19. stoljeća* (voditelja Jurice Budje). Unutar projekta *Dopreporodne hrvatske gramatike* (voditeljice Marijane Horvat) bit će izrađen specijalizirani korpus starih hrvatskih gramatika. Dijakronijska će perspektiva biti zastupljena i u teorijskom kognitivnolingvističkom istraživanju konceptualne metafore, metonimije, predodžbenih shema i semantičkih okvira u hrvatskome jeziku, što je predmet projekta *Repozitorij metafora hrvatskoga jezika* (voditeljice Kristine Štrkalj Despot). Korpus o kojemu će ovdje biti riječ dio je šire inicijative unutar korpusne lingvistike slavenskih jezika.

[1] Za popis hrvatskih korpusa v. npr. adresu http://www.hnk.ffzg.hr/jthj/korpusi.htm; o hrvatskim korpusima v. također Klobučar Srbić 2008: 45–50 te Štrkalj Despot i Möhrs 2015: 329–353; vidi za poljski jezik Molas 2014.

[2] O tome v. npr. Tadić 2003.

[3] U Zavodu za lingvistiku Filozofskoga fakulteta u Zagrebu pod vodstvom Milana Moguša pokrenut je projekt *Jezik Marka Marulića*, koji je 1970. godine dopunjen i preimenovan u *Kompjutorska analiza tekstova stare hrvatske književnosti*.

[4] Tako su primjerice nastale konkordancije hrvatskih tekstova M. Marulića, Karnarutićevih, Pelegrinovićevih, Lucićevih, Benetovićevih i Hektorovićevih djela. Konkordirane su također Zoranićeve *Planine*, Barakovićeva *Vila Slovinka*, Vitezovićevo *Odiljenje sigetsko*, Gundulićev *Osman*, *Razvod istarski* i dr.

2. CroDi u kontekstu DFG projekta *Korpuslinguistik und diachrone Syntax: Subjektkasus, Finitheit und Kongruenz*

Izgradnja Regensburškoga dijakronijskog korpusa hrvatskoga jezika – CroDi pokrenuta je u okviru istraživačkoga projekta *Korpuslinguistik und diachrone Syntax: Subjektkasus, Finitheit und Kongruenz in slavischen Sprachen* (*Korpusna lingvistika i dijakronijska sintaksa: padež subjekta, finitnost i kongruencija u slavenskim jezicima*), koji se odvijao u razdoblju od 2012. do 2016. uz potporu Njemačke istraživačke zajednice (HA 2659/1-2, *Deutsche Forschungsgemeinschaft* – DFG). Projekt se sastoji od dvaju potprojekata koji se provode na Humboldtovu sveučilištu u Berlinu i na Sveučilištu u Regensburgu.[5] Riječ je o nastavku projekta *Korpuslinguistik und diachrone Syntax: Zur Grammatikalisierung peripherer Subjekte in slavischen Sprachen* (HA 2659/1-1) (*Korpusna lingvistika i dijakronijska sintaksa: Gramatikalizacija perifernih subjekata u slavenskim jezicima*), koji je financijski podupro DFG (2008. – 2011.). Tijekom toga projekta utemeljen je Regensburški ruski dijakronijski korpus (RRuDi). Taj korpus sadržava mnoge važne pisane dokaze iz povijesti ruskoga jezika, odnosno staroslavenskoga i srednjoistočnoslavenskoga, sve do kraja 18. stoljeća. Korpus je djelomično lematiziran, a u njemu su označene i morfosintaktičke kategorije.[6] Osim toga, u okviru istoga projekta nastao je i Regensburški poljski dijakronijski korpus (PolDi). Taj je korpus oblikovan s pomoću digitaliziranih dokumenata koje je izradila skupina sa Sveučilišta u *Göttingenu* pod vodstvom Gerda Hentschela te na temelju elektroničkih izdanja Instituta za poljski jezik Akademije znanosti u Krakovu. Korpus je javno dostupan. Grafijski je ujednačen, eksperimentalno lematiziran i označen *tagerom* (označivačem) za suvremeni poljski jezik.[7] Za povijesnojezična je istraživanja važan i višejezični paralelni korpus PROIEL (http://proiel.github.io/) koji uključuje i morfosintaktički anotirani staroslavenski *Codex Marianus glagoliticus*, koji su proučavali i Vatroslav Jagić i Josip Hamm.

Nastavak projekta bavi se pitanjem odnosa nominativa kao padeža subjekta i njegove kongruencije s finitnim predikatom. Taj se odnos često određuje kao

[5] Glavni i odgovorni voditelj projekta je Roland Meyer (HU Berlin), a potprojekt u Regensburgu vodi Björn Hansen. Suradnici na projektu u Njemačkoj jesu Ernst Hansack, Olesia Lazarenko, Iryna Parkhomenko, Veronika Wald, Uliana Yazhinova te Aleksej Tikhonov, a međunarodni su partneri Institut za hrvatski jezik i jezikoslovlje iz Zagreba (Marijana Horvat, Sanja Perić Gavrančić), Universität Tromsø, Norveška (Hanne Eckhoff), Universität Uppsala, Švedska (Ingrid Maier), Ukrajinska nacionalna akademija znanosti (Jevgenija Karpilovs'ka, Olesia Lazarenko), Taras-Ševčenko, Sveučilište u Kijevu, Ukrajina (Oksana Nika, Natalia Darchuk) kao i Iwan-Franko, Sveučilište Žytomyr, Ukrajina (W. M. Mojsijenko).

[6] Vidjeti na mrežnoj stranici http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi.

[7] Vidjeti na http://rhssl1.uni-regensburg.de/SlavKo/korpus/poldi.

aksiom rečeničnoga ustrojstva. Međutim, potvrđeni su slučajevi, i to ne samo u slavenskim jezicima, koji odstupaju od toga načela te su stoga osobito informativni za teoriju padeža. Cilj je projekta objasniti nastanak i razvoj takvih nekanonskih („perifernih") potvrda padeža za subjekt u povijesti slavenskih jezika. Pritom će se primjenjivati i razvijati suvremene metode korpusne lingvistike. Istraživanje se nastavlja na prvu fazu projekta, u kojoj je primarno bilo pitanje odnosa ostvarenih i neostvarenih subjekata. Polazeći od dosad sekundarno obrađenih subjekata u dativu, sada bi se trebala produbiti međusobna povezanost problematike padeža i finitnosti.[8]

U regensburškom se potprojektu podloga za jezičnu usporedbu proširuje i na hrvatski jezik, koji s obzirom na temu projekta pokazuje osobito važne pojavnosti u modalnim konstrukcijama. Hrvatske modalne konstrukcije dopuštaju izražavanje subjekta ili u nominativu (npr. *Zato mi*$_{nom}$ *trebamo otići do Ministarstva kulture*.) ili u dativu (npr. *Ovu stranicu valja nam*$_{dat}$ *gledati kao komercijalnu*.). Osim toga, potvrđena je varijacija u kongruenciji predikata sa subjektom (npr. *Svi kandidati*$_{pl}$ *treba*$_{sg}$ *da su dosegli*$_{pl}$ *doba puberteta*.). Projekt će pridonijeti slavističkim i općelingvističkim istraživanjima koja se bave temom subjekta, finitnosti i kongruencije te modalnosti. Za uspostavu dijakronijskoga korpusa hrvatskoga jezika – CroDi nužni su postojeći resursi i iskustva iz prve faze projekta. Uspostavom digitaliziranih, obilježenih izvora (korpusa) s otvorenim mrežnim pristupom, kao i dosljednom provedbom kvantitativnih metoda, trebao bi se postići metodološki napredak u dijakronijskim korpusnolingvističkim istraživanjima slavenskih jezika i šire.[9] Stoga je izgradnja CroDi korpusa dio šire inicijative u okviru korpusne lingvistike slavenskih jezika, zbog čega CroDi sa spomenutim korpusima RRudi, PolDi i upravo uspostavljenim ukrajinskim dijakronijskim korpusom ima neka zajednička obilježja.

[8] Iz teksta prijave projekta: „Nicht nur in slavischen Sprachen gibt es jedoch Phänomene, die von diesem Standardfall abweichen und die daher für die Theorie der Subjekteigenschaften und der Kasuslizensierung besonders aufschlussreich sind. Das Projekt hat zum Ziel, die Entstehung und Entwicklung solcher nicht-kanonischer ("peripherer") Instanzen von Subjektkasus in der Geschichte slavischer Sprachen aufzuklären. Dabei werden aktuelle korpuslinguistische Methoden eingesetzt und weiterentwickelt. Das Vorhaben schließt an die erste Förderungsphase des Projekts an, in der die overte vs. Nullrealisierung von Subjekten im Vordergrund gestanden hatte. Ausgehend von den bisher sekundär bearbeiteten sog. Dativ-Subjekten, soll jetzt der Gesamtzusammenhang der Kasus-/Finitheitsproblematik vertieft werden."

[9] Iz teksta prijave projekta: „Durch die Realisierung digitalisierter, annotierter Quellen (Korpora) mit öffentlichem Online-Zugang sowie durch die konsequente Umsetzung quantitativer Verfahren soll zudem ein Methodenfortschritt für die diachrone Linguistik in den slavischen Sprachen und darüber hinaus erreicht werden."

3. Ciljevi i načela uspostave dijakronijskoga korpusa hrvatskoga jezika – CroDi

Oblikovanje dijakronijskoga korpusa hrvatskoga jezika temelji se na modelu po kojem su uspostavljeni regensburški korpusi za ruski (RRudi) i poljski jezik (PolDi). U članku *The construction and application of diachronic Slavonic corpora in linguistic research – RRuDi* (*Russian*) *and PolDi* (*Polish*) Roland Meyer (2012) obrazlaže načela i metodološke pretpostavke uspostave spomenutih korpusa, polazeći od činjenice da različiti ciljevi istraživanja zahtijevaju i različite koncepcije oblikovanja dijakronijskih korpusa. Meyer ističe da je pristup usmjeren dokumentiranju i očuvanju tekstova u cjelini, poput digitalnih biblioteka *Monumenta Germaniae Historica* (v. http://www.dmgh.de) i *Manuscriptorium* (v. http://www.manuscriptorium.com), namijenjen poglavito filološkoj analizi na svim razinama. S druge strane, korpusnolingvistički pristup ima za cilj dokumentiranje poglavito onih aspekata pisanih povijesnih izvora koji mogu biti lingvistički relevantni. Radi reprezentativnosti umjesto cjelovitih tekstova unose se odsječci odnosno ogledni dijelovi teksta. Potonja metoda upotrijebljena je za uspostavu dijakronijskoga dijela Helsinškoga korpusa engleskoga jezika (www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/)[10], koji je prema Meyerovu mišljenju dobar model i za dijakronijske korpuse slavenskih jezika. Helsinški je korpus zamišljen kao reprezentativni presjek pisanoga jezika određenoga razdoblja uključujući pritom tekstove poput pravnih i sudskih spisa, priručnika, znanstvenih rasprava, govora, dnevnika, privatne i službene korespondencije i slično. Pritom se ne izbjegavaju intervencije kakve ne dopušta filološki pristup tekstu, a to je primjerice izostavljanje odlomaka pisanih drugim jezikom ili stihova u proznom tekstu (Kytö 1996.)[11]. Broj pojavnica iz tekstova pojedinih razdoblja nije zadan u jednakim omjerima, kratki su tekstovi uključeni u cijelosti, a izvadci iz dužih tekstova katkad znatno variraju u obujmu (odsječci iz tekstova ranijega razdoblja sadržavaju oko 10 000 pojavnica, a oni iz kasnijih razdoblja oko 2500 do 5000 pojavnica). Korpus se uglavnom temelji na sekundarnim izdanjima, rjeđe na izvornim tekstovima, a njihov je odabir uvjetovan nekolikim kriterijima: kronološkim, regionalnim, sociolingvističkim i žanrovskim (Kytö i Rissanen 1993: 7)[12]. Na pitanje može li se tako oblikovan korpus smatrati uistinu reprezentativnim presjekom jezika određenoga razdoblja nema pouzdanoga odgovora, smatra Meyer, jer je nemoguće sa sigurnošću utvrditi opseg recepcije starih tekstova. Repre-

[10] Kytö 1996 (citirano prema Meyer 2012: 224).
[11] Citirano prema Meyer 2012: 226.
[12] Vidi prethodnu bilješku.

zentativnost dovodi u pitanje i činjenica da su neki od pisanih izvora sačuvani samo spletom povijesnih okolnosti, a neki zauvijek izgubljeni[13]. Ipak, varijacija različitih žanrova svakako je poželjna i s filološkoga i statističkoga aspekta te u tom smislu Helsinški korpus svakako može poslužiti kao dobar model, ističe Meyer (2012: 226).

Na opisanim su načelima uzorkovanja jezičnopovijesne građe većim dijelom utemeljeni neki od postojećih korpusa slavenskih jezika: dijakronijski dio Češkoga nacionalnoga korpusa (https://ucnk.ff.cuni.cz/)[14], Regensburški ruski dijakronijski korpus (RRuDi) i Regensburški poljski dijakronijski korpus (PolDi). Ista je koncepcija, kojom bi se trebalo omogućiti istraživanje lingvističkih, poglavito (morfo)sintaktičkih osobitosti hrvatskoga jezika u povijesnom razvoju, primijenjena u izgradnji Regensburškoga dijakronijskog korpusa hrvatskoga jezika.

4. O tekstovima

Dijakronijski korpus hrvatskoga jezika – CroDi obuhvaća tekstove od 16. do 19. stoljeća. Pritom se nastojalo izabrati ne samo tekstove autora relevantnih za pojedino razdoblje nego i one koji su široj zajednici manje poznati. Zastupljeni su tekstovi pisani i štokavskom i čakavskom i kajkavskom stilizacijom književnoga jezika, kao i tronarječnim hibridom. S obzirom na to da i u povijesnojezičnim istraživanjima treba voditi računa o funkcionalnoj raslojenosti jezika, pri izboru se osobita pozornost pridavala i žanrovskoj razgranatosti postojećih tekstova, što znači da su osim književnoumjetničkih uključeni i znanstveni, popularno-znanstveni te pravno-administrativni tekstovi. Za dio tekstova korištene su već objavljene transkripcije, podvrgnute provjeri i kritičkomu čitanju, dok je dio transkribiran prema izvorniku. Tekstovi su priređeni u Odjelu za povijest hrvatskoga jezika i povijesnu leksikografiju Instituta za hrvatski jezik i jezikoslovlje. U skladu s koncepcijom izrade korpusa koja ne uvjetuje cjelovitost tekstova neki su uvršteni odabranim odsječcima. Valja naglasiti da su od priređenih tekstova dosad u korpus uključeni ovi (ukupno 306 883 riječi i 820 317 pojavnica):[15]

[13] O pitanju reprezentativnosti Meyer upućuje na Claridge 2008 i Biber 1993.

[14] Meyer 2012: 224 upućuje na Kučera 2002 te ističe da ČNK u današnjem obliku istovremeno nudi pristup i cjelovitim tekstovima izvornika.

[15] Popis tekstova sadržava i odgovarajuće podatke na engleskome jeziku koji će biti sastavni dio popratnih obavijesti uz CroDi korpus (o tome više u nastavku članka).

NAME	TITEL	YEAR	WORDS	TOKENS	GENRE	Type of style	Dialect
Marko Marulić	Od naslidovan'ja Isukarstova i od pogarjen'ja tašćin segasvitnjih	1500	46.198	225.232	religious prose (devotional treatise)	literary	Chakavian
	Dijalozi Grgura pape (izbor)	1513	17.750	40.131	religious prose	literary	Chakavian
Petar Zoranić	Planine	written in 1536 (published in 1569)	31.880	78.360	allegorical pastoral novel	literary	Chakavian
Petar Hektorović	Mikša Pelegrinović	1556	989	2.532	epistles and funeral poems	literary	Chakavian
Matija Divković	Sto čudesa aliti zlamen'ja blažene i slavne Bogorodice, Divice Marije	1611	31.645	81.184	religious prose	literary	Shtokavian
Bartol Kašić	Predgovor Ritualu rimskom	1636	1.019	2740	prologue	literary style with elements of scientific style	Shtokavian
Bartol Kašić	Od nasledovan'ja gospodina našega Jezusa, duševno i prizamjerno	1641	5.708	13.734	religious prose (devotional treatise)	literary	Shtokavian
Jakov Mikalja	Od ortografije jezika slovinskoga ili načina od pisan'ja	1649	2.231	5.578	orthography treatise	scientific	Shtokavian

Jakov Mikalja	Gramatika talijanska ukratko	1649	2.555	6.528	grammar handbook	scientific	Shtokavian
Pavel Češković	Žalosno govorenje	1690	3.903	9.706	funeral sermon	literary	Kajkavian
	Dubrovačke oporuke	17.–18. century	12.138	29.152	last wills and testaments	administrative style with elements of colloquial style	Shtokavian
Andrija Kačić Miošić	Razgovor ugodni naroda slovinskoga	1756	37.589	45.829	epic poems and narrative chronicles	literary	Shtokavian
Blaž Tadijanović	Svašta po malo ili kratko složenje i mena i riči u ilirski i njemački jezik (izbor)	1761	2.936	7.924	language handbook	popular scientific	Shtokavian
Hilarion Gašparoti	Prodeka II	1761	3.491	17.168	funeral sermon	literary	Kajkavian
Hilarion Gašparoti	Prodeka III	1761	7.194	17.364	funeral sermon	literary	Kajkavian
Adam Baltazar Krčelić	Najvredneše stalnosti pelda	1767	8.968	21.532	funeral sermon	literary	Kajkavian
Adam Baltazar Krčelić	Dužnosti spunjenje	1772	5.071	12.164	funeral sermon	literary	Kajkavian
Joannis Baptist Lalangue	Medicina ruralis iliti vračtva ladanjska	1776	16.122	36.158	folk medicine recipes (so-called “ljekaruša”)	popular scientific	Kajkavian
Juraj Maljevec (kapucin Gregur)	Prigodnica Mariji Tereziji	1781	8.348	19.478	funeral sermon	literary	Kajkavian
Tituš Brezovački	Sveti Aleksi	1786	13.880	17.195	drama	literary	Kajkavian

Tituš Brezovački	Matijaš grabancijaš dijak	1804	15.559	38.068	comedy	literary	Kajkavian
Tituš Brezovački	Diogeneš	1805	24.690	59.192	comedy	literary	Kajkavian
	Luičeva ljekaruša	1746.	7.019	33.368	folk medicine recipes (so-called "ljekaruša")	popular scientific	Kajkavian

5. Uspostava CroDi korpusa

5.1. Priređivanje tekstova

Hrvatski su tekstovi priređeni u Odjelu za povijest hrvatskoga jezika i povijesnu leksikografiju Instituta za hrvatski jezik i jezikoslovlje u Zagrebu.[16] Priređeni su tekstovi slani u Institut za slavistiku Sveučilišta u Regensburgu na daljnju računalnu obradu. Prvo je u svim tekstovima obrisano formatiranje te su tekstovi pohranjeni u OpenOfficeu. Pritom je trebalo obratiti pozornost na sve izvorne formate tekstova te ih ponovno uspostaviti ako su se *obrisali*. Stoga je bilo nužno novu formatiranu inačicu teksta usporediti s izvornim tekstom i prema potrebi je ispraviti. To, primjerice, znači da se i u odt-formatu moraju pojaviti kosa slova na mjestima na kojima su ona u izvornom tekstu.[17]

Nakon što je tekst pregledan, izdvaja se naslov i uvrštava u tablicu[18] (vidi dolje).

Meta: Eigenschaft	**Meta: Wert**
title	Od naslidovan'ja Isukarstova i od pogarjen'ja tašćin segasvitnjih. De imitatione Christi (The Imitation of Christ).
ofauthor	Marko Marulić
dateFrom	
dateTo	
date	1500

[16] Velik dio tekstova temelji se na građi priređenoj i obrađenoj u okviru završenoga projekta MZOS-a *Tekstologija hrvatske pisane baštine* (br. projekta 212-2120920-0894) voditeljice Marijane Horvat (od 1. ožujka 2008. do konca 2013.), ali je korpus dopunjen i djelima izvan spomenutoga projekta. Stoga su u priređivanju tekstova sudjelovali: Marijana Horvat (voditeljica), Sanja Perić Gavrančić, Barbara Štebih Golub, Ivana Lovrić Jović, Martina Kramarić, Vuk-Tadija Barbarić, Željka Brlobaš, Ivana Klinčić, Vladimira Rezo, Martina Horvat i Ivo-Pavao Jazbec.

[17] Za taj dio posla bila je zadužena Veronika Wald.

[18] Tablicu je osmislio Roland Meyer.

publicationStmt	neobjavljena transkripcija Marijane Horvat; Institut za hrvatski jezik i jezikoslovlje, Odjel za povijest hrvatskoga jezika i povijesnu leksikografiju; unpublished transcription (transcription: Marijana Horvat; Institute of Croatian Language and Linguistics, Department of Croatian Language History and Historical Lexicography).
editor	Veronika Wald
editingDate	18.02.15
keywords	Chakavian, literary
genre	Religious prose (devotional treatise)

Kao što se može zaključiti uvidom u tablicu, u nju se unose i drugi metapodatci: autor[19] teksta, godina nastanka odnosno objavljivanja teksta. Uvrštena je i mogućnost upisa vremenskoga razdoblja u kojem je tekst nastao („date from" i „date to"). Ostali relevantni metapodatci jesu: informacija o izdavaču, ime osobe koja je priredila tekst[20] te datum kad je tekst obrađen. Osim toga, postoji i mogućnost navođenja ključnih riječi te određivanja žanra kojemu tekst pripada. Za hrvatske je tekstove iznimno važan podatak o stilizaciji književnoga jezika jer se tekstovi jezično razlikuju prema stilizacijama (čakavska, štokavska, kajkavska, hibrid). Svi su metapodatci, osim naslova teksta i informacija o izdavaču, pisani engleskim jezikom, a imena autora uvijek izvorno.

Nakon što se metapodatci uvrste u tablicu, tekst se mora pohraniti kao ODT-datoteka, a zatim se umnaža u ODT-formatu i uvrštava u *tei_RM.odt*. Dobiveni *tei_RM.odt* predložak obrađuje se prema međunarodnim standardima obrade metapodataka obuhvaćenima projektom TEI (*Text Encoding Initiative*).[21]

U sljedećem se koraku tekst pohranjuje kao XML-datoteka, stoga u rubrici *Extras* valja odabrati odgovarajući XML filtar, a zatim u rubrici *Neu* (*novo*) slijediti naredbe *Allgemein* (*općenito*) > *Anwendung* (*uporaba*): *OpenOffice Writer* (*.odt*), ime filtra i tipa datoteke: *TEI-P5*, *Transformation* > *ExportXSLT*, učitati datoteku *odttotei.xsl* i pritisnuti *OK*.

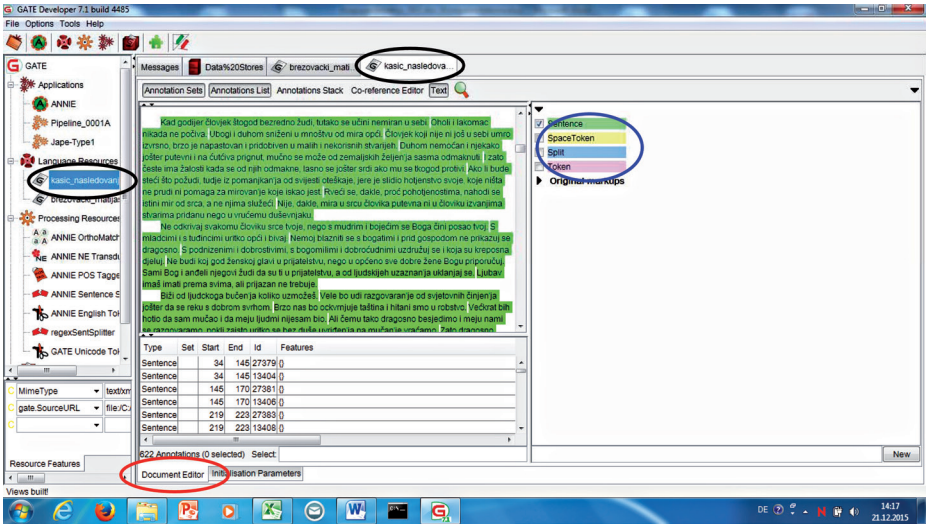
Nakon svih navedenih postupaka u rubrici *Datei* (datoteka) treba odabrati funkciju *Exportieren* i novi tip podataka (*TEI-P5* (*.xml*)) te datoteku prebaciti u odgovarajuću mapu. XML-format omogućuje da tekst bude obilježen u *GATE* (General Architecture of Text Engineering) platformi.

[19] Kada je autor nepoznat, u tablicu se unosi odrednica *anonymous*.
[20] U slučaju hrvatskih tekstova to je Veronika Wald.
[21] Tu obradu obavlja Roland Meyer (Humboldt-Universität zu Berlin).

5.2. Anotacije

Sve se anotacije koje se nalaze u korpusu mogu podijeliti u dvije skupine – one koje su automatski obilježene i one koje su ručno obilježene. Automatsko unošenje anotacija provodi se u programu *GATE* s pomoću sustava *ANNIE* (*A Nearly-New Information Extraction system*). Najvažniji su alati *Unicode Tokeniser* i *RegEx Sentence Spliter*, koji izdvajaju pojavnice i rečenice. Pri automatskom obilježavanju valja obratiti pozornost i na to da se kao kôd u odgovarajući redak unese Unicod *utf-8*. *Sentence splitter* razdvaja tekst na rečenice, a *Tokeniser* u jednostavne pojavnice poput brojeva, interpunkcije i riječi.

Razine automatskoga obilježavanja pojavljuju se s desne strane (označeno plavom bojom), kao što se vidi na slici dolje (slika 1). Ista slika prikazuje da je tekst Bartola Kašića *Od nasledovan'ja gospodina našega Jezusa, duševno i prizamjerno* (označeno crnom bojom) otvoren u prozoru programa GATE i da su obilježene sve rečenice (zazelenjeno). Ispod teksta pojavljuje se popis anotacija te se navodi ukupan broj pojavnica, u ovom slučaju rečenica, koji u obrađenom tekstu Kašićeva *Nasledovan'ja* iznosi 622 pojavnice (označeno crvenom bojom).



Slika 1. Automatske anotacije u GATE-u

Dok se raščlanjivanje teksta na pojavnice odnosno rečenice odvija automatski, morfosintaktička se anotacija provodi ručno. Podjela teksta na rečenice ne čini se dovoljno uporabljivom (za jezikoslovna istraživanja) jer se rečenica može sastojati od više iskaza te se u jezikoslovaca uobičajilo dijeliti tekst na

klauze. Poznato je da ne postoji alat koji bi omogućio automatsko označivanje *klauza* u rečenicama, stoga je taj mukotrpan i zahtjevan posao učinjen ručno.[22]

Pravila obilježavanja *klauza* razradio je Roland Meyer (HU Berlin). Pod nazivom *klauza* razumijevamo finitni glagolski oblik sa svim dopunama koje ovise o njemu. Ako se unutar veće *klauze* pojavljuje jedna manja, valja najprije obilježiti manju i tek tada onu veću koja ju obuhvaća, što je prikazano na slici 2 (tamnijom su bojom obilježene manje klauze koje se nalaze unutar veće).

Oveh kuliko potrebuješ s tobum vzemi ter, da se kuliko razveseliš, sprejti se po varašu moreš.

Slika 2. Obilježavanje *klauza*

Za *klauze* koje sadržavaju modalne konstrukcije postoje dodatne anotacije koje su osmišljene za različite tipove modalnih konstrukcija. Slika 3 prikazuje obilježavanje *klauza* s modalnim konstrukcijama, što se vidi pod oznakom *ComplementSubtype*.

Slika 3. Prikaz obilježavanja *klauza*

Sve su modalne konstrukcije uglavnom obilježene kao infinitivne (1) ili kao da-konstrukcije. Te su anotacije na slici 3 označene tamnocrvenom bojom. Pritom modalni glagol mora imati finitni oblik, a u slučaju da zavisna rečenica počinje veznikom da, subjekt modalne konstrukcije mora biti istovjetan sa subjektom zavisne rečenice, kao u primjeru br. 2.

[22] Manji dio *klauza* u GATE-u obilježile su Veronika Wald (12 168 *klauza*) i Anna Stupavsky (1170), dok je većina posla odrađena u Odjelu za povijest jezika i povijesnu leksikografiju Instituta za hrvatski jezik i jezikoslovlje. Prebacivanjem obilježenih *klauza* iz Word-formata u GATE-format bavila se Anna Stupavsky.

*(1) I dostojno vele veće **ima biti** iznutra nego ča se vidi izvanka.*[23]

*(2) Zato ja scijenim i nahodim da **su dostojne** da **se štampaju**.*[24]

Ako se subjekt modalne konstrukcije i subjekt zavisne *da*-rečenice ne poklapaju, uvodi se oznaka *da_other* (označeno zelenom bojom). Primjerice, u rečenici br. 3 masno otisnuta zavisna rečenica obilježit će se oznakom *da_other*:

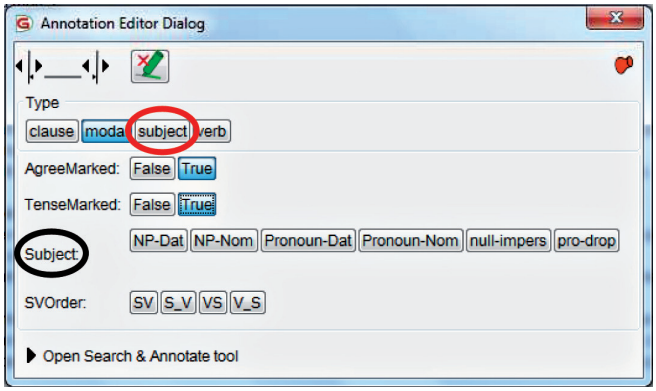
*(3) Valja **da je moja brada starija**.*[25]

Ako se modalni glagol u modalnoj konstrukciji pojavljuje u nekongruentnu obliku (infinitivu ili participu), takva se konstrukcija obilježava kao *inf_other* (označeno crnom bojom) (4). U slučaju izostanka infinitiva ili zavisne *da*-rečenice *klauza* se obilježava kao elipsa (označeno plavom bojom) (5).

*(4) Trudno [je] istino **moći suditi** dobri li duh ali ini tebe pelja na želin'je ovo ali ono, ali si još ganut od tvoga duha.*[26]

(5) Kak bi mi bilo moguče?[27]

Nakon obilježavanja *klauza* slijedi obilježavanje modalnih riječi.[28] Nakon što se obilježi modalna riječ (u izborniku *modal*), u prozoru se pojavljuju opcije obilježavanja koje su relevantne za modalnu strukturu (slika 4).



Slika 4. Obilježavanje modalnih riječi

[23] Marulić, *Od naslidovan'ja Isukarstova i od pogarjen'ja tašćin segasvitnjih*, 1500.
[24] Divković, *Sto čudesa aliti zlamen'ja blažene i slavne Bogorodice, Divice Marije*, 1611.
[25] Brezovački, *Matijaš grabancijaš dijak*, 1804.
[26] Marulić, *Od naslidovan'ja Isukarstova i od pogarjen'ja tašćin segasvitnjih*, 1500.
[27] Brezovački, *Diogeneš*, 1805.
[28] Sve anotacije modalnih konstrukcija provela je Veronika Wald.

Te su opcije: obilježavanje kongruencije (AgreeMarked), glagolskoga vremena (TenseMarked), subjekta te redoslijeda pojavljivanja subjekta i glagola (SVOrder). Pri obilježavanju kongruencije i glagolskoga vremena postoje dvije mogućnosti: potvrđenost (True) ili izostanak (False) bilo koje od spomenutih dviju kategorija.

U prvom retku (slika 4) pod opcijom *Type* ponuđena je anotacija za subjekt (označeno crvenom bojom). Anotacija *subject* pojavljuje se i u obilježavanju modalnih konstrukcija (označeno crnom bojom). Tehnički gledano, dvostruko obilježavanje nije problematično jer ako se u rečenici ne pojavljuje subjekt, onda ne postoji mjesto na kojem bismo mogli zabilježiti izostanak subjekta.

Subjekt se može obilježavati na ove načine:

NP-Dat – subjekt je imenska fraza u dativu, npr. *Svarhu svega s velicim ponižen'jem sartca i prihiljenim počtovan'jem, s punom virom i pravom misal'ju počten'ja Božjega tribi je* ***popu Božjemu*** *pristupiti na služen'je tolika posvetilišća i na tican'je i vazetje njega.*[29]

NP-Nom – subjekt je imenska fraza u nominativu, npr. *Da ča sam* ***ja****, Gospodine, da smim pristupiti k tebi*?[30]

Pronoun-Dat – subjekt je zamjenica u dativu, npr. *Zato bi se imao človjek u Bogu ustanovititi i sasma ukrijepiti neka* ***mu*** *ne bi bilo potrebno mnoga utješen'ja iskati*.[31]

Pronoun-Nom – subjekt je zamjenica u nominativu, npr. *I potom toga još bi navišćeno glavi od crikve da i* ***on*** *ne smi mu reći*.[32]

null-impers – označava izostanak subjekta koji se ne može obilježiti kao *pro-drop*, npr. *Istina jest da kada ti veliš Daj mi nož, da ti razumi(j)eš 'koji god nož', ali u njemačkomu* ***ne more se*** *tako reći, nego valja vazda nadodati das ili ein.*[33]

pro-drop – subjekt je zamjenica koja u tekstu nije potvrđena, ali se može rekonstruirati prema glagolskom obliku, npr. *Preštimanje, zahvalnost koju dužni* ***jeste*** *dati ocu, ne skratete odvetku*.[34]

Anotacija za redoslijed subjekta i glagola *SVOrder* u relevantna je prije svega za one modalne konstrukcije u kojima se pojavljuje tzv. dativni odnosno nekanonski subjekt.

Oznake se razrješavaju ovako:

SV – prvo subjekt, iza njega modalni glagol odnosno modalna riječ (nema

[29] Marulić, *Od naslidovan'ja Isukarstova i od pogarjen'ja tašćin segasvitnjih*, 1500.
[30] Dubrovačke oporuke, 17. – 18. stoljeće.
[31] Kašić, *Od nasledovan'ja gospodina našega Jezusa, duševno i prizamjerno*, 1641.
[32] *Dijalozi Grgura pape* (izbor), 1513.
[33] Tadijanović, *Svašta po malo ili kratko složenje imena i riči u ilirski i njemački jezik* (izbor), 1761.
[34] Krčelić, *Najvredneše stalnosti pelda*, 1767.

umetnutih riječi), npr. *Ne **mu bilo treba** posuditi novca koji iz svojeh krepostih zadobival je dosta*.[35]

S_V – redoslijed je jednak kao u prethodnom slučaju, ali pritom su subjekt i modalni glagol ili riječ razdvojeni drugim riječima, npr. *Moja si, veće **ti** se micati **ne valja***.[36]

VS – prvo modalni glagol, iza njega modalna riječ (nema umetnutih riječi), npr. *Kaj štimaš da bi **moguče bilo jednomu** takve ciganije delati da ne bi drugi ove ali spazili ali zeznali*.[37]

V_S – redoslijed je isti kao u prethodnom slučaju (*VS*), ali su modalni glagol ili riječ i subjekt razdvojeni drugim riječima.

Sljedeća se anotacija odnosi ponovno na subjekt rečenice. Za analizu modalnih konstrukcija relevantan je podatak je li subjekt ostvaren i u kojem padežu dolazi.

Četiri su mogućnosti obilježavanja subjekta:

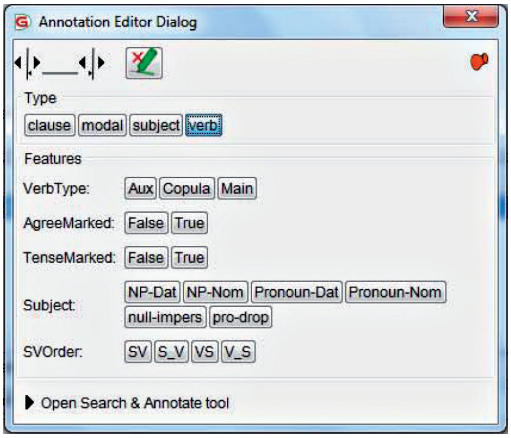
N-dat – subjekt je imenica u dativu

N-nom – subjekt je imenica u nominativu

P-dat – subjekt je zamjenica u dativu

P-nom – subjekt je zamjenica u nominativu.

Pri obilježavanju subjekta ne postoji mogućnost *pro-drop* ili *null-impers* jer tehnički nije moguće obilježiti ono što nije potvrđeno. Nadalje, anotacije za subjekt pojavljuju se također ispod opcije *modal* (slika 4) i *verb* (slika 5).

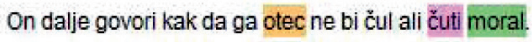


Slika 5. Obilježavanje glagola

[35] Krčelić, *Najvredneše stalnosti pelda*, 1767.
[36] Zoranić, *Planine*, 1536.
[37] Brezovački, *Diogeneš*, 1805.

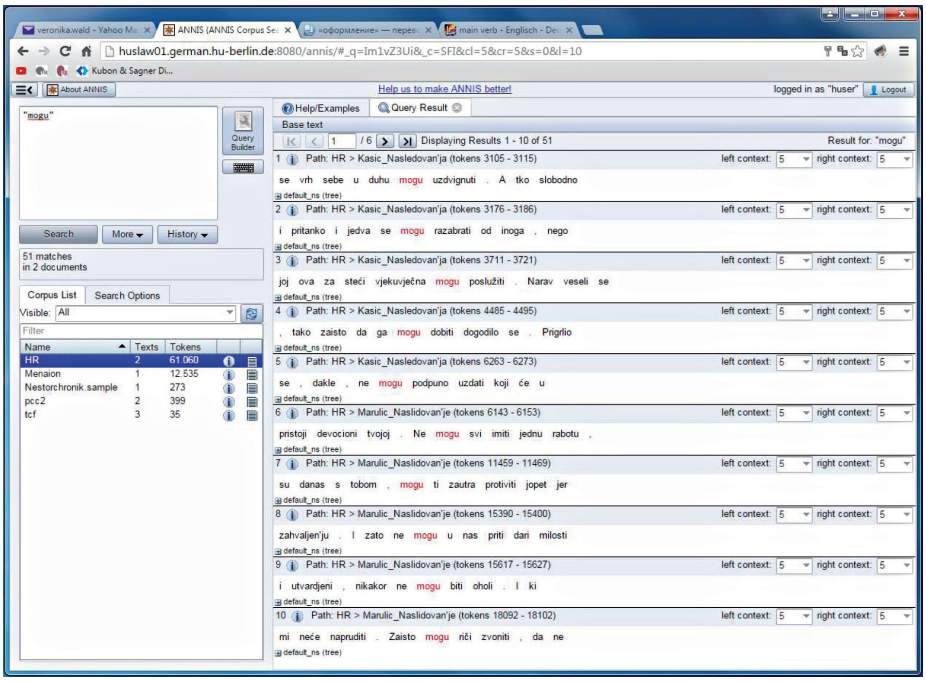
Pri obilježavanju glagoli se razvrstavaju na tri tipa i obilježavaju oznakama *Aux*, *Copula* i *Main*. Ostale anotacije odgovaraju onima koje se u shemi pojavljuju pod opcijom *modal*. To je obilježavanje kongruencije (*AgreeMarked*), glagolskoga vremena (*TenseMarked*), subjekta te redoslijeda subjekta i glagola unutar rečenice (*SVOrder*). Za obilježavanje infinitiva valja u odgovarajućim retcima (uz *AgreeMarked* i *TenseMarked*) odabrati opciju *False*. Shemu anotacija za modalne konstrukcije razradili su Björn Hansen i Veronika Wald, a u program ju je implementirao Roland Meyer.

Svaka anotacija ima svoju boju. Na slici 6 vidi se kako izgleda rečenica nakon obilježavanja svih potrebnih dijelova konstrukcije – subjekt je obilježen narančastom bojom, modalni glagol zelenom, a samoznačni ružičastom.



Slika 6. Anotacije sastavnica modalne konstrukcije (Brezovački, *Sveti Aleksi*)

Sve su anotacije s pomoću programa GATE unesene u sustav ANNIS (Annotation of Information Structure) (slika 7).



Slika 7. ANNIS

Korpus CroDi dostupan je u sustavu ANNIS[38] za sve korisnike. ANNIS ima otvoren pristup te je osobito prikladan za rukovanje lingvističkim korpusom s različitim tipovima. Stoga je Roland Meyer odabrao ANNIS zbog izuzetne kvalitete vizualizacije (građe) i relativno jednostavnih mogućnosti pretraživanja korpusa koji ima nekoliko razina.

6. Zaključak

Korpus CroDi nastao je kao jedan u nizu korpusa slavenskih jezika, čiju je koncepciju osmislio njemački slavist Roland Meyer po uzoru na Helsinški korpus engleskoga jezika. Zbog načina prezentiranja građe takav model ne omogućuje potpunu filološku analizu starih tekstova na svim razinama, nego primarno ima za cilj raščlambu morfosintaktičkih jezičnih obilježja pisanih povijesnih izvora. Prvo istraživanje na CroDi korpusu odnosi se na analizu modalnih konstrukcija, a provedeno je na Sveučilištu u Regensburgu pod vodstvom Björna Hansena. Potpunom uspostavom Regensburškoga dijakronijskog korpusa hrvatskoga jezika ubrzat će se jezičnopovijesna istraživanja te pridonijeti učinkovitijoj raščlambi starih hrvatskih tekstova. Važno je istaknuti da će se tekstovima tako osigurati međunarodna, osobito slavistička, pozornost, što će omogućiti nova (komparativna) istraživanja.

CroDi će imati slobodan pristup i različite razine pretraživanja uz mogućnost kasnijih proširivanja, a bit će mrežno dostupan na poslužitelju Humboldtova sveučilišta u Berlinu od ljeta 2016. godine. Informacije o tome bit će objavljene na adresi: http://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/rund-ums-institut/regensburger-korpora/index.html.

Literatura:

Klobučar Srbić, Iva. 2008. Obol korpusne lingvistike suvremenoj leksikografiji. *Studia lexicographica* 2/3. 39–51.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Text: Coding conventions and lists of source texts.* Technical report. Universität Helsinki. http://khnt.hit.uib.no/icame/manuals/HC/ (pristupljeno 23. veljače 2016.).

[38] ANNIS je pretraživačka i vizualizacijska, višeplatformska (Linux, Mac, Windows) arhitektura otvorenoga koda (*open source*) koja se temelji na mrežnom pregledniku za složeni višeslojni lingvistički korpus s različitim vrstama anotacija. ANNIS pomaže vizualizaciji podataka iz različitih područja, kao što su sintaksa, semantika, morfologija, prozodija, referencijalnost, leksik itd. Za projekte koji rade s govornim jezikom, potrebna je također i podrška za audioanotacije i videoanotacije.

Kytö, Merja; Rissanen, Matti. 1993. *General Introduction*, chapter I. Mouton de Gruyter. Berlin – New York. 1–17.

Meyer, Roland. 2012. The construction and application of diachronic Slavonic corpora in linguistic research – RRuDi (Russian) and PolDi (Polish). *Diachrone Aspekte slavischer Sprachen*. *Slavolinguistica 16*. Hrgs. Björn Hansen. Verlag Otto Sagner. München – Berlin – Washington D.C. 223–242.

Meyer, Roland; Hansen, Björn; Hansack, Ernst. 2012. *Korpuslinguistik und diachrone Syntax: Subjektkasus, Finitheit und Kongruenz in slavischen Sprachen.* (HA 2659/1-2, DFG-Projektantrag).

Moguš, Milan. 1987. Listajući kompjutorsku konkordanciju Lucićevih djela. *Mogućnosti* 1–2. 90–98.

Molas, Jerzy. 2014. Praktyczny przewodnik po korpusie języka chorwackiego. *Praktyczny przewodnik po korpusach języków słowiańskich.* Red. Milena Hebal-Jezierska. Uniwersytet Warszawski. 82–97

Štrkalj Despot, Kristina; Möhrs, Christine. 2015. Pogled u e-leksikografiju. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 41/2. 329–353.

Tadić, Marko. 2003. *Jezične tehnologije i hrvatski jezik*. Ex libris. Zagreb.

Mrežni izvori:

Croatian web corpus (hrWaC). http://nlp.ffzg.hr/resources/corpora/hrwac/ (pristupljeno 23. veljače 2016.).

Hrvatski nacionalni korpus. http://www.hnk.ffzg.hr/ (pristupljeno 23. veljače 2016.).

Jezične tehnologije za hrvatski jezik. http://www.hnk.ffzg.hr/jthj/korpusi.htm (pristupljeno 23. veljače 2016.).

Manuscriptorium. Digital Library of Written Cultural Heritage. http://www.manuscriptorium.com (pristupljeno 4. veljače 2016.).

Monumenta Germaniae Historica. http://www.dmgh.de (pristupljeno 4. veljače 2016.).

Polish Diachronic Corpus PolDi. http://rhssl1.uni-regensburg.de/SlavKo/korpus/poldi (pristupljeno 15. veljače 2016.).

Regensburg Russian Diachronic Corpus RRuDi. http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi (pristupljeno 15. veljače 2016.).

The Helsinki Corpus of English Texts. www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/ (pristupljeno 4. veljače 2016.).

Ústav Českého národního korpusu. https://ucnk.ff.cuni.cz/ (pristupljeno 4. veljače 2016.).

Das Regensburger diachrone Korpus der kroatischen Sprache – CroDi

Zusammenfassung

In dem Artikel wird das Projekt zum Aufbau eines diachronen Korpus der kroatischen Sprache CroDi am Institut für Slavistik der Universität Regensburg in Zusammenarbeit mit dem Institut für die kroatische Sprache und Linguistik in Zagreb vorgestellt. Beschrieben werden die Zielsetzung und die Prinzipien für die Erstellung des Korpus, das methodische Vorgehen bei der Ausarbeitung der Einträge und die bisherigen Resultate. Die Idee zur Erstellung eines diachronen Korpus der kroatischen Sprache entstand im Rahmen von zwei DFG-Projekten (2008-2016) zum Thema „Korpuslinguistik und diachrone Syntax“. In Regensburg existieren bereits ein „Russisches diachrones Korpus“ (RRudi) und ein „Polnisches diachrones Korpus“ (PolDi), weitere Korpora sind im Aufbau. Die mit diesen Korpora gewonnenen Erfahrungen sind in das „Diachrone Korpus der kroatischen Sprache“ eingeflossen. Die Prinzipien für die Erstellung und Bearbeitung der Korpora hat im wesentlichen Roland Meyer in seinem Artikel „The construction and application of diachronic Slavonic corpora in linguistic research“ (2012) festgelegt. Modell und Vorbild war dabei vor allem das „Helsinki Corpus of English Texts“. Das kroatische Korpus enthält bisher Texte des 16.-19.Jh. in allen Ausprägungen des Kroatischen (Štokavisch, Čakavisch und Kajkavisch). Neben den bedeutenden Vertretern der kroatischen Literatur dieser Zeit sind auch weniger bekannte Autoren und anonyme Texte vertreten. Die Texte sind zum Teil von Hand annotiert. Die Prinzipien dazu wurden von Roland Meyer, Björn Hansen und Veronika Wald festgelegt und werden ausführlich dargestellt. Durch die Aufnahme in das Korpus werden die enthaltenen Texte international zugänglich, was weitere komparatistische Untersuchungen ermöglicht.

Ključne riječi: Regensburški dijakronijski korpus hrvatskoga jezika, korpusna lingvistika, povijesna sintaksa, modalnost, subjekt, finitnost, kongruencija

Schlüsselwörter: Regensburger Diachrones Korpus des Kroatischen CroDi, Korpuslinguistik, historische Syntax, Modalität, Subjekt, Finitheit, Kongruenz