Priprema usporedivih korpusa za usporedbu

Ivana Lalli Paćelat

Odjel za interdisciplinarne, talijanske i kulturološke studije, Sveučilište Jurja Dobrile u Puli,
I. M. Ronjgova 1, 52100 Pula
ilalli@unipu.hr

Sažetak

Osim formalne usklađenosti šest korpusa prema rasponu, opsegu i strukturi, zbog prirode planirane kvantitativne analize neizostavna je i njihova usklađenost na razini POS i MSD označavanja. Budući da je ta usklađenost samo djelomično prisutna prikazuju se u ovome radu potrebni postupci svođenja postojećih oznaka pod zajedničku oznaku kako bi rezultati bili usporedivi. No, i u slučaju potpune usklađenosti skupa oznaka sa smjernicama koje propisuju standarde, neizbježno je promišljanje i usklađivanje oznaka s obzirom na razlike u poimanju i postojanju gramatičkih kategorija u pojedinim jezicima, u slučaju ovoga rada hrvatskoga i talijanskoga jezika. Nakon što su se utvrdile i prikazale razlike među korpusima, koje su proizašle iz kontrastivne analize dvaju jezika, i u skladu s time odabrane moguće zajedničke oznake za vrste riječi i druge gramatičke kategorije, pribjeglo se postupku normalizacije korpusa. Radi postizanja bolje usporedivosti rezultata na međujezičnoj razini promatrala se tako distribucija unutar zajedničkih dijelova korpusa na način da cjelinu čine samo one oznake koje su zajedničke i relevantne za ciljno istraživanje što je doprinijelo ujedno i većoj pouzdanosti rezultata.

Ovime se radom s jedne strane potvrdila važnost sustavnoga planiranja izrade skupa oznaka za vrste riječi i gramatičke kategorije za pojedini jezik u skladu s međunarodnim smjernicama koje propisuju standarde i stvaraju preduvjete usporedivosti među korpusima kako na unutarjezičnoj razini tako i na međujezičnoj razini, a s druge strane pokazalo kako se usporedba i usklađivanje MSD ili POS oznaka mogu smatrati dobrim temeljem i zanimljivim pristupom u kontrastivnoj analizi dvaju jezika.

Preparing comparable corpora for comparison

Although the six corpora included in the research were comparable with respect to size, purpose and structure, it was indispensable, due to the nature of the planned quantitative analysis, to make them comparable at the POS and MSD tagging level. Since the tagsets used to annotate the corpora were only partially compatible, several procedures were needed to convert the existing tags to a common tagset in order to have comparable results. However, also in case of full compatibility with international standards, it is inevitable to think about and to compare the tagsets because of the differences in the perception and in the existence of grammatical categories in different languages, i.e. Croatian and Italian. After the differences among the tagsets of the six corpora were identified, followed by a detailed contrastive analysis of the two languages and after the only possible common POS and MSD tagset was found, the normalization of the corpora was performed. In order to achieve better comparability of results at inter-lingual level only the distribution within the common, comparable and relevant tags were taken into account which contributed to greater reliability and accuracy of results.

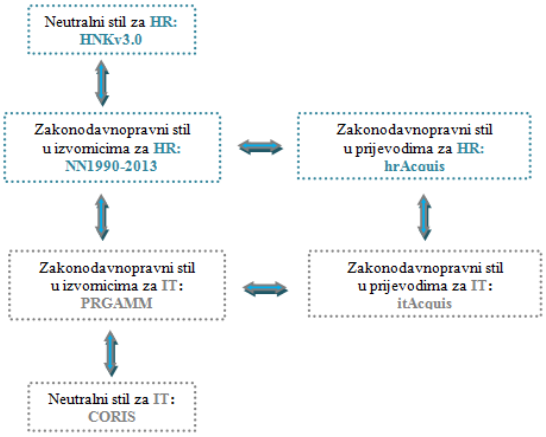
On one hand this paper confirmed the importance of systematic planning of linguistic annotation scheme for each language in accordance with guidelines which prescribe international standards and create conditions for the comparability across corpora at both inter-lingual and intra-lingual levels. On the other hand, the paper showed that comparing and analysing MSD or POS tagsets can be considered a good basis and an interesting approach for the contrastive analysis.

1 Uvod

Cilj je ovoga rada istaknuti važnost sustavnog planiranja izrade skupa oznaka za pojedini jezik u skladu sa smjernicama koje propisuju standarde i stvaraju preduvjete za usporedbu korpusa kako na unutarjezičnoj razini tako i na međujezičnoj razini. U slučaju kada takvi skupovi oznaka ne postoje, a priroda analize koja se planira provesti to zahtjeva, potrebno je uskladiti skupove oznaka korpusa koji se žele uspoređivati.

Predstavit će se u ovome radu postupci pripreme korpusa za usporedbu, koji su bili nužni kako bi se omogućila analiza registra i analiza univerzalnih obilježja prijevoda na hrvatsko-talijanskome jezičnom paru, a koje su detaljno prikazane u Lalli Paćelat (2014). Temeljni preduvjeti za analizu registra prema Biberu (1995) jesu: komparativni pristup, kvantitativna analiza i reprezentativni uzorak. Kako bi se zadovoljili navedeni uvjeti istraživanje je provedeno na šest različitih računalnih korpusa odnosno četiri vrste korpusa za oba jezika: referentni općejezični jednojezični korpusi (1) Hrvatski nacionalni korpus (HNK v 3.0) i (2) Corpus di italiano scritto (CORIS); (3) specijalizirani dvojezični usporedivi korpus (potkorpusi referentnih korpusa: NN1990-2013/PRGAMM); (4,5) jednojezični usporedivi korpusi izvornika i prijevoda na istome jeziku istoga stila (NN1990-2013/hrAcquis i PRGAMM/itAcquis) (6) i usporedni korpus hrvatskih i talijanskih prijevoda (hrAcquis/itAcquis). Više o korpusima u Tadić (2009) i Rossini Favretti i dr. (2002).

Komparativni pristup pretpostavlja usporedbu ciljnoga funkcionalnog stila s drugim stilovima. U planiranome istraživanju usporedit će se tako zakonodavnopravni stil s neutralnim stilom referentnoga korpusa koji čini skup različitih funkcionalnih stilova jednoga jezika, s istim stilom drugoga jezika te s prevedenim tekstovima istoga stila na istome jeziku. Komparativni pristup i analiza u navedenome istraživanju izgledat će kao što je prikazano na Slici 1.



Slika 1: Komparativni pristup.

U sličnim istraživanjima (Neumann, 2013; Teich, 2003; Xiao 2010 i dr.) istraživači su koristili korpuse u čijem su obilježavanju odnosno označavanju posredno ili neposredno sudjelovali, a svi su korpusi bili označeni na isti način. Suprotno njima, u navedenome istraživanju koristit će se korpusi koji su obilježeni i označeni neovisno jedni o drugima zbog čega su samoj analizi prethodile provjere usporedivosti korpusa i usklađivanja skupova oznaka. To je usklađivanje bilo potrebno kako bi se omogućilo jasnije prikazivanje rezultata, ali i kako bi se osigurala visoka razina usporedivosti među korpusima odnosno među jezicima na temelju vrsta riječi i drugih gramatičkih kategorija.

2 Obilježavanje hrvatskih i talijanskih korpusa

Od ključne je važnosti prilikom sastavljanja korpusa njegova usklađenost sa svjetskim standardima za obilježavanje. Korpusi koji će se koristiti u navedenome istraživanju usporedivi su na najvećoj mogućoj razini: prema rasponu, opsegu, i što je jako važno za ovaj rad i prema strukturi. Tipologija tekstova, njihov omjer i druge važne smjernice za izradu korpusa, kako za HNK-a, tako i za CORIS, usklađene su koliko je u danim uvjetima bilo moguće s preporukama EAGLES-a (1996). O strukturi HNK-a u Tadić (1996), (1998), (2003), a CORIS-a u Rossini Favretti (2000).

Međutim, osim temeljnih pretpostavki usklađenosti potrebnih radi uspoređivanja korpusa, zbog prirode analize koja se planira provesti, neizostavna je i usklađenost na razini POS i MSD obilježavanja odnosno označavanja. Ta je usklađenost samo djelomično prisutna, kao što će se u nastavku rada i predstaviti.

POS i MSD označavanje hrvatskih tekstova, kako onih u HNKv3.0, naravno onda i potkorpusu NN1990–2013, tako i onih u hrvatskome prijevodnom korpusu hrAcquis, obavljeno je prema MULTEXT-East (v4.0) specifikaciji (Erjavec, 2004).

Za talijanski je jezik razrađeno niz preporuka za obilježavanje korpusa u okviru EAGLES projekta (Monachini, 1995). Isto tako, nekoliko je istraživačkih skupina radilo na POS označavanju te je svaka skupina razradila svoj način označavanja i svoj skup oznaka. Više o usporedbi skupa oznaka za vrste riječi u talijanskome jeziku u Bernardi i suradnici (2005), (2006), Tamburini i suradnici (2008), Venturi (2009) i dr. Tamburini i suradnici (2008: 97) ističu kako postoji načelna suglasnost oko oznaka za osnovne vrste riječi, dok značajne razlike postoje u kriterijima za dodjeljivanje oznaka za podvrste riječi. Neusuglašenost u korištenju oznaka za POS i MSD označavanje talijanskih korpusa znatno otežava njihovu međusobnu usporedbu, a onda i usporedbu s korpusima na drugim jezicima. S obzirom na to da je MULTEXT preporuka standardiziranoga sustava za označavanje gramatičkih kategorija bila razrađena u suradnji s EAGLES inicijativom iz 1996., a da je MULTEXT-East bio samo „srednjo- i istočnoeuropski odvjetak“ toga projekta (Tadić, 2003: 107) te da su i sastavljači korpusa navodili prihvaćanje smjernica EAGLES-a očekivalo se da neće biti razlike barem u oznakama temeljnih gramatičkih kategorija za talijanski i za hrvatski jezik. Iako se skup oznaka izrađenih za označavanje talijanskoga referentnoga korpusa naziva „EAGLES-like tagset for Italian“ (Tamburini, 2000, Monachini, 1995) oznake su nešto drugačije od onih standardiziranih. Tamburini (2000) ističe da je skup oznaka primijenjen u CORIS-u u skladu sa standardima EAGLES-a razrađenih za talijanski jezik u Monachini (1995). S obzirom na već spomenutu neusuglašenost i neslaganje oko označavanja vrsta riječi za talijanski jezik autor je uzeo u obzir preporuke i primjere navedene u on-line inačici rječnika De Maura iz 2007. Skup oznaka korištenih za označavanje itAcquisa sastavljen je po uzoru na skup oznaka za španjolski jezik (Prokopidis et al., 2012) i sličniji je skupu oznaka za hrvatski jezik odnosno bliži je smjernicama EAGLES-a.

Najveća razlika između talijanskih i hrvatskih korpusa koji se rabe u planiranome istraživanju odnosi se na razinu označavanja. Dok su talijanski korpusi označeni samo na razini vrste riječi (POS), hrvatski su korpusi označeni na način da je svakoj gramatičkoj kategoriji dodijeljena i vrijednost. Stoga svaka pojavnica u hrvatskim korpusima ima morfosintaktički opis (MSD), dok takav opis nedostaje kod talijanskoga referentnog i specijaliziranog korpusa, a samo djelomično postoji kod prijevodnoga korpusa.

Ako se uzme u obzir do sada spomenuto, odnosno da svaka pojavnica u talijanskim korpusima ima oznaku o vrsti riječi, a da pojavnice u hrvatskim korpusima imaju oznake koje, osim podataka o vrsti riječi, donose i vrijednosti relevantnih morfosintaktičkih kategorija te da su same oznake, nazivi za iste gramatičke kategorije, različiti, jasno je da je u uspoređivanju hrvatskih i talijanskih prvi korak bio stvaranje preduvjeta za morfosintaktičku usporedbu hrvatskih i talijanskih korpusa.

3 Usklađivanje korpusa

3.1 Svođenje na zajedničke oznake

Prilikom uspoređivanja svih šest korpusa koristit će se samo one oznake koje su zajedničke svim korpusima, dok će se detaljnije analize prikazati samo onda kada se budu analizirali korpusi samo na jednome jeziku. Kako bi i na razini vrsta riječi hrvatski i talijanski korpusi bili usporedivi, bilo je potrebno uskladiti oznake i svesti ih pod zajednički nazivnik prvenstveno radi jasnijega prikazivanja rezultata. U odabiru zajedničkih oznaka odlučilo se prikloniti standardiziranome skupu oznaka prema MULTEXT-East (v3.0) (Tadić iz 1998. u Erjavec, 2004) preporuci.

Za svaku će se vrstu riječi prikazati u nastavku svođenje na zajedničku oznaku. Valja podsjetiti da sva tri hrvatska korpusa koriste isti skup oznaka, pa će se u nastavku rada i prikazima navoditi samo oznake u HNK-u, a vrijedit će i za korpus NN1990-2013 i hrAcquis. Što se tiče talijanskih korpusa, uspoređivat će se dva skupa oznaka, jedan koji se koristi u CORIS-u, koji vrijedi i za potkorpus PRGAMM te drugi koji se koristi za itAcquis.

3.1.1 Imenice i prilozi

Svođenje oznaka imenica pod isti nazivnik nije predstavljalo problem. U svim korpusima razlikuju se opće i vlastite imenice, kao što je to prikazano u Tablici 1.

itAcquis			CORIS	HNK	Odabrana oznaka i objašnjenje	
S			N	N	N	imenica
Ss	Sp	Sn	NN	Nc	Nc	opća imenica
SP			NN_P	Np	Np	vlastita imenica
SWs	SWp	SWn				
B			ADV	R	R	prilog
				Qr		
BN				Qz		

Tablica 1: Prijedlog zajedničkih oznaka za imenice i priloge.

Problematični nisu bili ni prilozi za koje prema skupovima oznaka nije predviđena daljnja podjela na vrste priloga. Valja napomenuti da u hrvatskim korpusima postoji daljnja podjela na vrste priloga, koja međutim nije dokumentirana u specifikaciji za hrvatski jezik MULTEXT-East-a. Pod oznaku priloga (R) u hrvatskim korpusima uključit će se i dio pojavnica s oznakom čestica, kao što prikazuje Tablica 1. S obzirom na to da u talijanskim korpusima ne postoji posebna oznaka za čestice budući da u tradicionalnim talijanskim gramatikama čestice (tal. *particelle*) nemaju status vrste riječi, smatralo se to dobrim rješenjem za dio hrvatskih čestica koje se u talijanskome jeziku smatraju prilozima. O spornome statusu talijanskih čestica mnogo se pisalo, a čak i o njihovome odnosu prema česticama u hrvatskome jeziku, primjerice u Tekavčić (1989), i Jernej (1990). Pojavnice koje će priključiti prilozima u hrvatskim korpusima imaju oznake Qr i Qz. Riječ je o potvrdnim odnosno niječnim česticama koje se u talijanskome jeziku tradicionalno svrstavaju među priloge (tal. *avverbi di affermazione/negazione*)[1].

3.1.2 Glagoli

Suprotno imenicama i prilozima, svođenje oznaka glagola na zajedničku nije bilo jednako zahvalno, kao što je prikazano u Tablici 2.

itAcquis	CORIS	HNK	Odabrana oznaka i objašnjenje	
V	V	V	V	glagol
VA	V_ESSERE	Va	Va	pomoćni glagol
	V_AVERE	Vc		
VM	V_MOD	Vm	Vm	glagolski oblik

Tablica 2: Prijedlog zajedničkih oznaka za glagole.

[1] Valja napomenuti da se stavovi talijanskih lingvista razilaze se po pitanju opravdanosti svrstavanja jesno-potvrdnih čestica pod priloge.

Zbog različitih podjela unutar svakoga korpusa bilo je moguće svesti na samo dvije zajedničke oznake. S obzirom na to da se u oba jezika podjela na pomoćne i kopulativne glagole uglavnom podudara, odlučilo se za zajedničku oznaku (Va). Treba međutim imati na umu nekoliko mogućih razlika. Pomoćni su glagoli oni koji služe za tvorbu složenih glagolskih oblika, a u hrvatskome su jeziku to *biti* i *htjeti*, dok su u talijanskome to *essere* i *avere* (Silić i Pranjković, 2005: 185; Dardano i Trifone, 2003: 200). Valja spomenuti da se u talijanskome jeziku i glagoli *venire, andare, dare* i *stare* pojavljuju u ulozi pomoćnih glagola u tvorbi određenih glagolskih oblika. Pod kopulativnim glagolima podrazumijeva se glagol *biti* odnosno *essere* (usp.Težak i Babić, 1994: 198 i dr.), no postoje i drugi glagoli, koji vrše istu funkciju, no nekad se nazivaju polukopulativni (Silić i Pranjković, 2005: 269) ili kopulativni glagoli (Dardano i Trifone, 2003: 192), međutim u korpusima koji se koriste u navedenome istraživanju nisu označeni kao kopulativni.

Isto tako u hrvatskome jeziku ne postoje glagolski oblici združeni s nenaglašenim zamjenicama (poput *dillo, prendertelo, andarci, affittasi* i sl.), a glagolski pridjevi i prilozi ne čine posebnu kategoriju već su svrstani pod oznaku glavnih glagola (Vm). Potrebno je napomenuti da u talijanskome jeziku postoje *participio presente* i *participio passato*. S vremenom je particip prezenta (npr. *amante, vincente, studente*) izgubio glagolska svojstva te se danas smatra uglavnom pridjevom ili imenicom (Dardano i Trifone, 2003: 245). Svoju glagolsku vrijednost zadržao je samo u administrativnome funkcionalnom stilu, ističu Dardano i Trifone (2003: 246). Glagolski pridjev radni i trpni ostvaruju se u talijanskome jeziku istim oblikom, participom prošlim, što znači da ih nije moguće odvajati samo prema obliku. To onemogućava analizu distribucije pasivnih glagolskih oblika u talijanskim korpusima budući da nemaju posebno označen podatak o radnom ili trpnom obliku, dok ga hrvatski korpusi imaju. S druge strane talijanski *gerundio* ima dva oblika koji odgovaraju hrvatskim oblicima glagolskih priloga. Glagolski prilog radni odgovara tako talijanskome *gerundio presente*, dok glagolski prilog prošli odgovara talijanskome *gerundio passato* (Dardano i Trifone, 2003: 246). Dok bi se iz hrvatskih korpusa i itAcquisa glagolski pridjevi i glagolski prilozi mogli izvući i pretragama s MSD odnosno POS oznakama, takvu pretragu nije moguće obaviti za CORIS gdje su označeni samo glagolski pridjevi. U CORIS-u je moguća samo djelomična pretraga pomoću regularnih izraza.

Jedino što je moguće, jest, kao što to predočuje Tablica 2, razlikovanje pomoćnih (Va) i glavnih glagola (Vm). Valja pojasniti da se pod oznaku glavnih glagola (Vm) ne misli samo na samoznačne glagole, jer uključuju i modalne, fazne i perifrazne glagole koji su suznačni. Iako bi potpuno odvajanje suznačnih od samoznačnih glagola bilo zanimljivo, takva pretraga s oznakama nije ostvariva za svih šest korpusa. Moguće je, međutim, da već i odvajanje glavnih i pomoćnih glagola može biti pokazateljem neke tendencije, koju je potrebno detaljnije obraditi.

3.1.3 Zamjenice i pridjevi

Kategorije zamjenica i pridjeva vrlo se različito poimaju u hrvatskome i talijanskome jeziku zbog čega je svođenje spomenutih kategorija pod istu oznaku bilo vrlo zahtjevno, kao što to pokazuje Tablica 3.

itACQUIS			CORIS	HNK	Odabrana oznaka i objašnjenje	
A			ADJ	A	A	pridjev
As	Ap	An	ADJ	Af	Af	opisni pridjev
			~~ADJ_NUM~~	~~Ao~~	~~Ao~~	~~brojevni pridje~~
APs	APp	APn	ADJ_POS	Ps	Psx	posvojna zamj./pridjev (it)/povratno-posvojna zamj. (hr)
			PRON_POS	Px		
P			PRON	P	P	zamjenica
PE			PRON_PER	Pp	Ppx	lična zamj./ povratna zamj. (hr)
PC				Px		
PD			PRON_DIM	Pd	Pd	pokazna zamj. /pridjev (it)
DD			ADJ_DIM			
PI			PRON_IND	Pi	Piqr	neodređena zamj./pridjev (it)/upitna zamj./pridjev (it)/odnosna zamj./pridjev (it)
DI			ADJ_IND			
PQ			PRON_IES	Pq		
DQ			ADJ_IES			
PR	DR		PRON_REL	Pr		

Tablica 3: Prijedlog zajedničkih oznaka za pridjeve i zamjenice.

U Tablici 3 mogu se zamijetiti i značajne razlike u oznakama kod talijanskih korpusa.

Dok se skup oznaka u CORIS-u drži podjela tradicionalnih talijanskih gramatika, skup oznaka itAcquisa ima nešto drugačiju podjelu. Kod skupa oznaka itAcquisa kod promijenjivih riječi posebnu oznaku imaju oblici u jednini (s), množini (p) i oni koji su neutralni što se tiče broja (n), zbog čega su oznake u tablicama usporedno prikazane. Zanimljivo je zamijetiti da je ono što se u CORIS-u označava kao pokazni, neodređeni i upitni pridjev (ADJ_DIM, ADJ_IND, ADJ_IES) u itAcquisu označeno kao pokazni, neodređeni, upitni, ali i odnosni „determinante“ (DD, DI, DQ) prema uzoru na španjolski skup oznaka. Riječ je o posebnoj oznaci (D) za vrstu riječi koja uključuje i članove, što svakako ukazuje na njezin sporan status.

Iz ovoga kratkog prikaza samo nekoliko uočenih razlika u označavanju talijanskih korpusa, jasno je da odabir oznaka ovisi o teoriji na koju se oslanja te da nije u potpunosti teorijski neutralno. Ovaj je pokušaj usklađivanja oznaka dobar primjer kako nije dobro imati različite skupove POS i MSD oznaka za isti jezik jer se na taj način onemogućava unutarjezična analiza korpusa označenih različitim označivačima koji se nisu vodili istim smjernicama, a što u konačnici ograničava njihovu primjenu.

Ako se usporede samo oznake, koje donosi Tablica 3, vidljivo je da u oba jezika postoje posvojni pridjevi. Međutim, ono što se u talijanskome jeziku smatra posvojnim pridjevom u hrvatskome se jeziku smatra posvojnom zamjenicom. Definicije se, kako posvojnih zamjenica u hrvatskome, tako i talijanskih posvojnih pridjeva, podudaraju: obje kategorije iskazuju pripadnost te se određuju prema licima (usp. Silić i Pranjković, 2005: 123; Dardano i Trifone, 2003: 138). U hrvatskome jeziku smatraju se zamjenicama jer zamjenjuju posvojne pridjeve (Težak i Babić, 2005: 127).

U Primjeru 1 prenosi se rečenica koju Silić i Pranjković (2005: 123) navode kao jedan od primjera uporabe posvojnih zamjenica. Ista će se rečenica prevesti na talijanski jezik. Posvojni će pridjev odnosno posvojna zamjenica, kao istovrijednice, biti posebno istaknuti.

Tvoja profesorica dobro pjeva
La tua insegnante canta bene.

Primjer 1: Posvojne zamjenice u hrvatskome jeziku i posvojni pridjevi u talijanskome jeziku.

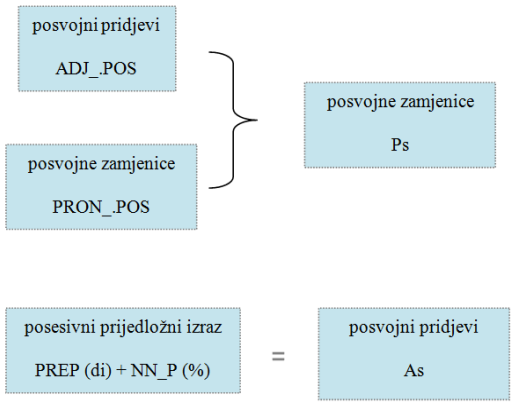
Osim oblika posvojnih zamjenica pod posvojne pridjeve u talijanskome spadaju još i pridjevi *altrui* i *proprio* (Dardano i Trifone, 2003: 139). U talijanskome jeziku posvojne zamjenice i posvojni pridjevi jednaki su po obliku, podsjeća Jernej (2005: 126). Ulogu onoga što se u hrvatskim gramatikama naziva posvojnim pridjevima, u talijanskome jeziku preuzeli su prijedložni izrazi *complemento di specificazione* točnije *complemento di specificazione di appartenenza o di specificazione possessiva* (usp. Dardano i Trifone, 2003: 139; Jernej, 2005: 290). Primjer 2 prema uzoru na primjere ponuđene u Silić i Pranjković (2005: 123) to zorno prikazuje. Njihovi će se primjeri prevesti na talijanski jezik, a posvojna zamjenica odnosno pridjev u hrvatskome i prijedložni izraz u talijanskome jeziku biti će posebno istaknuti.

Njegova sestra studira engleski.
Sua sorella studia inglese.

Petrova sestra studira engleski.
La sorella di Pietro studia inglese.

Primjer 2: Posvojna zamjenica odnosno pridjev u hrvatskome i prijedložni izraz u talijanskome jeziku.

Moguće rješenje u usporedbi talijanskih i hrvatskih korpusa sažeto je prikazano na Slici 2. Lijevi dio prikaza odnosi se na talijanski jezik, dok se desni dio prikaza odnosi na hrvatski jezik.

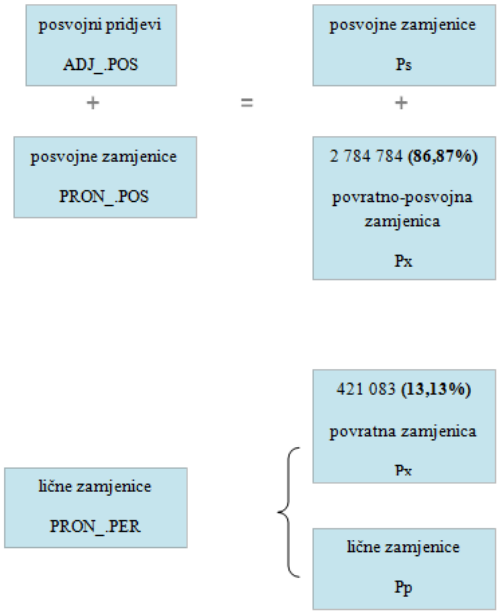


Slika 2: Mogućnosti kontrastivne analize posvojnih zamjenica odnosno pridjeva u postojećim korpusima.

Kada se uzme u obzir sve do sada rečeno, proizlazi da je vrlo teško odijeliti samo posvojne pridjeve jer treba uzeti u obzir da se oni u talijanskome jeziku drugačije ostvaruju, a budući da su talijanski tekstovi označeni samo na razini vrste riječi, potrebno bi bilo uzeti u obzir određeni broj prijedloga i određeni broj vlastitih imenica. Međutim, ono što se svakako može uspoređivati na temelju postojećih oznaka i korpusa jesu posvojne zamjenice u hrvatskome koje odgovaraju talijanskim posvojnim zamjenicama i posvojnim pridjevima. Kada je riječ o zamjenicama, problematična je i pripadnost povratnih zamjenica. Pod oznaku Px u hrvatskim korpusima spadaju povratna (*se, sebe*) i povratno-posvojna zamjenica (*svoj*), dok se u talijanskome jeziku povratne zamjenice svrstavaju pod lične zamjenice i broje više oblika (*mi, ti, si, ci vi si*), a povratno-posvojna zamjenica spada pod posvojne zamjenice odnosno posvojne pridjeve (*suo*). Stoga se broj pojavnica označenih s Px morao podijeliti u dvije skupine. Za podjelu se koristila sljedeća pretraga s regularnim izrazima:

1) [msd="Px.*pn.*"]
Pojašnjenje: p→personal n→nominal
2) [msd="Px.*nsa.*"]
Pojašnjenje: s→possessive a→adjectival

Kao što je vidljivo iz pretrage i same oznake upućuju na njihovu pripadnost, pa će se povratne zamjenice radi usklađivanja s talijanskim oznakama uključiti u lične zamjenice, dok će se povratno-posvojna zamjenica priključiti skupini hrvatskih posvojnih zamjenica odnosno talijanskih posvojnih zamjenica i posvojnih pridjeva u omjeru prikazanome na Slici 3.



Slika 3: Prijedlog usklađivanja oznaka povratne i povratno-posvojne zamjenice.

3.1.4 Prijedlozi, veznici i brojevi

Nakon usklađivanja oznaka zamjenica i pridjeva, prelazi se na ponešto jednostavnije usklađivanje oznaka prijedloga, veznika i brojeva, kao što prikazuje Tablica 4.

itAcquis		CORIS	HNK	Odabrana oznaka i objašnjenje	
E		PREP	Sp	Sp	prijedlog
EA		PREP_A			
Cc		CONJ_C	Ccs	Cc	veznik nezavisnosloženih rečenica
			Ccc		
Cs		CONJ_S	Css	Cs	veznik zavisnosloženih rečenica
			Csc		
NOs	NOn	ADJ_NUM	Ao	M	brojevi i brojevni pridjevi
NUM		C_NUM	M		

Tablica 4: Usklađivanje oznaka prijedloga, veznika i brojeva.

Dok se u talijanskome razlikuju jednostavni prijedlozi i prijedlozi združeni s članom, u hrvatskome postoji i podjela na jednostavne i složene prijedloge, koja ovom prilikom nije prikazana jer nije usporediva s talijanskom podjelom. Budući da talijanska podjela prijedloga ne postoji u hrvatskome jeziku, svi prijedlozi obaju jezika svedeni su pod jednu zajedničku oznaku (Sp).

Što se tiče veznika, u oba jezika razlikuju se veznici nezavisnosloženih rečenica i zavisnosloženih rečenica. Hrvatski korpusi spomenute veznike dijele još i na jednostavne i složene. Radi pojednostavljivanja smatra se dovoljnim podjela na veznike nezavisnosloženih rečenica ili konjunktore i na veznike zavisnosloženih rečenica ili subjunktore, kao što je prikazano u Tablici 4.

Usporedbom tradicionalnih podjela na vrste riječi u hrvatskim i talijanskim gramatikama uočava se razlika što se tiče statusa brojeva. Brojevi (na tal. *numerali*) se uglavnom u talijanskome jeziku svrstavaju među pridjeve (*aggettivi cardinali, ordinali e multiplicativi*; npr. Dardano i Trifone, 2003: 138). Naravno, postoje i u obliku brojevnih imenica i prijedložnih izraza, no u manjem broju, a odnose se na posebne brojeve (*numerali frazionari, distributivi i collettivi*; npr. Dardano i Trifone, 2003: 150; Faloppa, 2011 i dr.). Brojevi se u talijanskim gramatikama ne smatraju posebnom vrstom riječi (Dardano i Trifone, 2003; Lo Duca, 2011; Faloppa, 2011 i dr.), dok oni u hrvatskome uživaju status zasebne vrste riječi (Barić et al., 1997; Težak i Babić, 2005; Silić i Pranjković, 2005 i dr.). Problematiku statusa brojeva kao vrstu riječi i općenito morfoloških obilježja brojevnih riječi opširno je prikazala za hrvatski jezik Tafra (1989), (2000), (2005), a za talijanski jezik Faloppa (2011) i dr.

S obzirom na to da pod oznaku za talijanske brojevne pridjeve (ADJ_NUM) iz CORIS-a spada i dio pojavnica koje su u hrvatskim korpusima označene kao brojevi (M.*l), a ostatak pojavnica s oznakom brojeva (M.*d i M.*r) odgovara oznaci brojeva (C_NUM) u CORIS-u te da slično vrijedi i za itAcquis, u kojemu se za razliku od CORIS-a ne dodijeljuje oznaka pridjeva, bilo je važno svesti usporedive dijelove pod jednu zajedničku oznaku. Rješenje spomenutih nepodudarnosti u označavanju korpusa, ali i u poimanju brojeva u hrvatskim i talijanskim tradicionalnim gramatikama, pronađeno je, za spomenuto istraživanje, u odvajanju brojevnih pridjeva od ostalih pridjeva te u njihovom uključivanju pod zajedničku oznaku brojeva (M), kao što je detaljno prikazano u Tablici 4.

3.2 Problem člana

Prvi rezultati usporedbe omjera vrsta riječi u hrvatskome i talijanskome jeziku, pokazali su znatne razlike u distribuciji riječi u referentnim korpusima.

Prilikom promatranja rezultata bilo je važno uzeti u obzir i činjenicu da je „talijanski sustav djelomično analitički, dok je hrvatski u potpunosti sintetički, s vrlo složenim sklonidbenim sustavom i slobodnim redom riječi“ (Sočanac, 2004: 151). Ta činjenica ima kao posljedicu veliku razliku u broju prijedloga, koji su u talijanskome jeziku preuzeli i funkciju padežnih nastavaka (Sočanac, 2004: 151). Teško je, međutim, odrediti točan omjer prijedloga s takvom funkcijom u talijanskome jeziku. Još jedna bitna razlika među jezicima jest nepostojanje člana u hrvatskome jeziku. Suprotno razlici u prijedlozima, ova će se razlika, koja još u većoj mjeri utječe na usporedbu distribucije vrsta riječi u hrvatskim i talijanskim korpusima, pokušati neutralizirati. Način na koji će se to provesti i njegova opravdanost prikazat će se u nastavku rada.

Podsjetit će se na početku da vrijednost člana većinom ostaje implicitna u hrvatskome, dok je ona u talijanskome eksplicitna (Ljubičić, 2000: 226). Talijanski određeni član nastao je od pokaznoga pridjeva, a u nekim je slučajevima do danas sačuvao pokazno značenje, koje se onda i očituje u prijevodu uporabom pokazne zamjenice u hrvatskome jeziku (Ljubičić, 2000: 179). S druge strane, Ljubičić (2000: 193–194) upozorava kako je „teško izvršiti razgraničenje između neodređenoga člana i broja, ali još je teže odijeliti broj i član od neodređenoga pridjeva“ (u hrvatskome zamjenice) te ističe kako se neodređeni član ponekad prevodi neodređenim pridjevom (u hrvatskome zamjenicom) kada član nema samo gramatičko značenje, nego i ono leksičko.

Međutim i Ljubičić (2000: 228) i Karlić (2014) slažu se da samo u slučajevima kada članovi nose i leksičko značenje te kada nose komunikacijski relevantnu informaciju koja se ne može zaključiti iz konteksta, onda i jezici, koji nemaju član, poput hrvatskoga, mogu eksplicitno izraziti vrijednost člana drugim sredstvima.

Izražava li se zaista i koliko često vrijednost člana u hrvatskome jeziku provjerilo se na odabranome uzorku korpusa itAcquisa i hrAcquisa[2]. Valja napomenuti da se u oba slučaja radi o prijevodima s istoga jezika, dakle oba prijevoda imaju isti status te jednaki utjecaj izvornika s obzirom na to da među njima ne postoji izravan prijevodni proces. Test je pokazao da u talijanskim prijevodima članovi čine od 7,7% do čak 9,2% ukupnoga broja pojavnica. Isto tako, test je pokazao da od ukupnoga broja članova u talijanskim usporednim tekstovima, maksimalno se za njih 4%, a nekad i mnogo manje, vrijednost eksplicitno iskazuje drugim sredstvima. Da se vrijednost članova u većini slučajeva ne iskazuju u hrvatskome jeziku pokazuje Primjer 3.

1) „...u vezi s kvalitativnim Ø značajkama riže...“(jrc31999R0691)
„...per quanto riguarda le caratteristiche qualitative del riso...“(jrc31999R0691)

2) „...za potvrđivanje usklađenosti Ø određenog proizvoda ili Ø obitelji proizvoda...“(jrc31999R0691)
„...se per un dato prodotto o un gruppo di prodotti determinati...“(jrc31999D0089)

3) „...je li Ø postojanje nadzornog Ø sustava tvorničke proizvodnje za koji je odgovoran proizvođač potreban i dovoljan Ø uvjet...“(jrc31999D0089)
„...l’esistenza nella fabbrica di un sistema di controllo della produzione, effettuato dal fabbricante, sia una condizione necessaria e sufficiente...“(jrc31999D0089)

4) „...Ø Uredba (EZ-a) br. 708/98 ovime se izmjenjuje i dopunjuje kako slijedi ...“(jrc31999R0691)
„...Il regolamento (CE) n. 708/98 è modificato come segue...“(jrc31999R0691)

5) „...Ø EUROPSKI PARLAMENT I Ø VIJEĆE EUROPSKE UNIJE...“(jrc32006L0012)
„...IL PARLAMENTO EUROPEO E IL CONSIGLIO DELL’UNIONE EUROPEA...“(jrc32006L0012)

Primjer 3: Primjeri neiskazivanja vrijednosti člana u hrvatskome jeziku.

Eksplicitno iskazivanje vrijednosti člana posvojnom, neodređenom i pokaznom zamjenicom prikazano je u Primjeru 4.

1) „...posljednji puta izmijenjena i dopunjena Uredbom (EZ-a) br. 2072/98, a posebno njezin članak 8. točku b...“(jrc31999R0691)
„...modificato da ultimo dal regolamento (CE) n. 2072/98 (2), in particolare l’articolo 8, lettera b...“(jrc31999R0691)

2) „...budući da je zato poželjno odrediti onaj koncept proizvoda ili obitelji...“(jrc31999D0089)
„...è opportuno definire il concetto di prodotto o di gruppo di prodotti...“(jrc31999D0089)

3) „...«proizvođač» je svaka osoba čijom aktivnošću nastaje otpad...“(jrc32006L0012)
„...«produttore»: la persona la cui attività ha prodotto rifiuti...“(jrc32006L0012)

Primjer 4: Primjeri iskazivanja vrijednosti člana u hrvatskome jeziku.

S obzirom na to da su članovi činili sveukupno 7,97% pojavnica u referentnome talijanskom korpusu odnosno 7,12% svih pojavnica u specijaliziranome korpusu i 5,26% u prijevodnome korpusu, a da članovi ne postoje u hrvatskome, odlučilo se radi usporedivosti omjera isključiti broj pojavnica s oznakom člana. Na taj način uspoređivali bi se omjeri samo postojećih vrsta riječi u oba jezika. Takva metodološka odluka, osim što nalazi opravdanje u prethodnim radovima usmjerenim na prevođenje hrvatsko-talijanskoga jezičnog para (Ljubičić, 2000; Katušić, 1981; 1982a; 1982b; 1982c), ali i onim usmjerenim na proučavanje određenosti i neodređenosti u hrvatskome jeziku (Karlić, 2014), potvrđena je i testom na usporednim tekstovima.

3.3 Normalizirana veličina korpusa

Nakon što su utvrđene i prikazane razlike među korpusima, koje su proizašle iz kontrastivne analize dvaju jezika na razini vrsta riječi i pokušaja usklađivanja oznaka korištenih u morfosintaktičkome označavanju hrvatskih i

[2] Riječ je o posebnoj vrsti usporednoga korpusa, sastavljenoj od prijevoda na dva jezika bez izvornika.

talijanskih korpusa, provjerio se njihov utjecaj na pouzdanost rezultata distribucije vrsta riječi svih šest korpusa.

Iz provjere proizlazi kako te razlike ne mijenjaju u velikoj mjeri sliku distribucije vrsta riječi pojedinoga korpusa. Međutim, radi postizanja veće usporedivosti odlučeno je promatrati distribuciju unutar zajedničkih dijelova korpusa na način da cjelinu čine samo one oznake koje su zajedničke i relevantne za ciljno istraživanje. Relativne će se frekvencije stoga izračunavati u odnosu na normaliziranu veličinu korpusa, a podrazumijevat će veličinu korpusa umanjenu za apsolutnu frekvenciju oznaka koje nisu zajedničke ili koje nisu relevantne za ovo istraživanje. Radi se o vrstama riječi koje ne postoje u oba jezika, poput člana i čestica ili o oznakama koje nisu relevantne, a nisu u jednakoj mjeri označene u korpusima poput kratica, pokrata, simbola i interpunkcijskih znakova. Postoji i vrsta riječi koja postoji i označena je u svim korpusima, no neće se uzimati u obzir u normaliziranim korpusima. Radi se o uzvicima koje se zbog više razloga odlučilo isključiti iz normaliziranih korpusa. Osim što čine jako mali dio korpusa, uzvici nisu uobičajeni u zakonodavnopravnome stilu, što je vidljivo iz njihove još manje uključenosti u specijaliziranim korpusima. Dodatni je razlog niski postotak točnosti u označavanju uzvika u specijaliziranim korpusima, neovisno o jeziku[3].

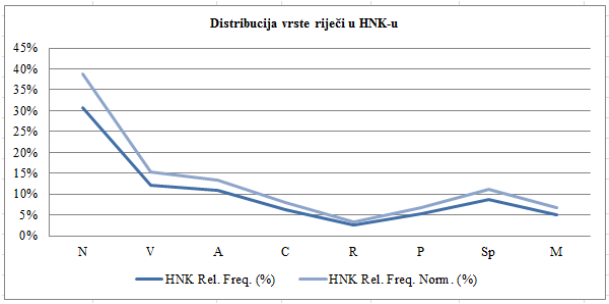
Normalizirani korpus uključivat će dakle samo apsolutnu frekvenciju sljedećih vrsta riječi odnosno oznaka: imenica (N), glagola (V), pridjeva (A), veznika (C), priloga (R), zamjenica (P), prijedloga (Sp) i brojeva (M). Tablica 5 prikazuje apsolutni broj pojavnica u izvornim oblicima korpusa, relativni broj pojavnica u normaliziranim korpusima i njihov omjer.

Korpus	Apsolutni broj pojavnica u izvornome korpusu	Apsolutni broj pojavnica u normaliz. korpusu	Omjer broja pojavnica u izvornome i normaliz. korpusu
HNK	216 812 148	177 216 713	0,8173
NN_{1990_2013}	92 363 788	70 852 385	0,7671
hrAcquis	11 209 795	8 784 240	0,7836
CORIS	130 294 347	98 300 670	0,7544
PRGAMM	9 575 784	7 325 921	0,7650
itAcquis	16 955 483	12 464 404	0,7351

Tablica 5: Apsolutni brojevi pojavnica u izvornim i u normaliziranim korpusima.

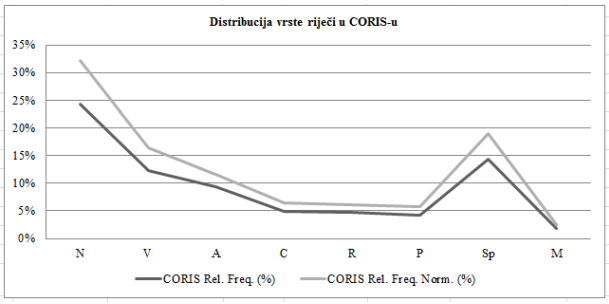
Iz Tablice 5 vidljivo je da su se korpusi smanjili za približno isti postotak. Da postotak nije u potpunosti isti i da se ne odražava jednako na svaku vrstu riječi, prikazat će u nastavku Grafikon 1 i Grafikon 2.

Grafikon 1 prikazuje distribuciju relativne frekvencije vrste riječi u referentnome korpusu hrvatskoga jezika (HNKv3.0) u njegovu izvornome obliku i distribuciju vrsta riječi u normaliziranome korpusu.



Grafikon 1: Usporedba distribucije vrste riječi u HNK-u.

U Grafikonu 2 prikazana je distribucija relativne frekvencije vrste riječi u referentnome korpusu talijanskoga jezika (CORIS) u njegovu izvornome obliku i distribuciju vrsta riječi u normaliziranome korpusu.



Grafikon 2: Usporedba distribucije vrste riječi u CORIS-u.

Kao primjer razlika u distribuciji vrsta riječi između izvornih oblika korpusa i normaliziranih korpusa prikazana je distribucija frekvencija vrsta riječi referentnih korpusa. Iako je ta razlika kod referentnih korpusa najmanja, vidljivo je, kao što to prikazuju i Grafikon 1 i Grafikon 2, da se svođenje korpusa na normaliziranu veličinu ne odražava jednako na svaku vrstu riječi te da je odluka o korištenju normaliziranih korpusa opravdana i da će doprinijeti većoj pouzdanosti rezultata.

4 Zaključak

Iz svega prikazanog u ovome radu očita je važnost i sustavno planiranje izrade skupa oznaka za vrstu riječi i ostale gramatičke kategorije za svaki pojedini jezik u skladu sa zajedničkim međunarodnim smjernicama koje propisuju standarde i stvaraju preduvjete usporedivosti među korpusima kako na unutarjezičnoj tako i na međujezičnoj razini.

Dok takav skup oznaka za hrvatski jezik postoji i dosljedno se koristi, puno je problematičnije stanje sa skupovima oznaka za talijanski jezik budući da postoje različiti skupovi oznaka. Ovaj rad jasno pokazuje kako neusklađenost oznaka, tj. postojanje različitih skupova POS i MSD oznaka za isti jezik, onemogućuje unutarjezičnu analizu korpusa označenih različitim skupovima oznaka, što u konačnici ograničava njihovu primjenu. Istovjetna primjedba vrijedi i na međujezičnoj razini i u toliko je bitno da se korpusni lingvisti pridržavaju zajedničkih međunarodnih smjernica jer se time omogućuje jednostavnija usporedivost rezultata pretrage korpusa na različitim jezicima.

[3] Pojavnice koje su označene kao uzvici u specijaliziranim, kako hrvatskim tako i talijanskim korpusima, uglavnom nisu uzvici, već se radi o kraticama, pokratama ili vrlo često stranim riječima, što ukazuje na problematičnost označavanja uzvika neovisno o jeziku i o korištenome automatskom označivaču.

No, i kada bi postojala potpuna usklađenost sa smjernicama, neizbježno je promišljanje i usklađivanje oznaka s obzirom na razlike u poimanju i postojanju gramatičkih kategorija u pojedinim jezicima. Moglo bi se zaključiti dakle da se usporedba i usklađivanje MSD ili POS oznaka mogu smatrati dobrim temeljem i zanimljivim pristupom u kontrastivnoj analizi dvaju jezika.

S druge strane, treba spomenuti da već neko vrijeme postoje pokušaji sastavljanja univerzalnih skupova oznaka za vrste riječi i druge gramatičke kategorije kao primjerice univerzalni skup oznaka za vrste riječi prikazan u Petrov i suradnici (2012: 2089), koji uključuje 12 vrsta riječi određenih na temelju analize skupova oznaka 22 jezika, među kojima se nalazi i talijanski, ali ne i hrvatski jezik ili poput onoga sastavljenog u sklopu projekta *The Universal Dependencies*[4], a koji uključuje i hrvatski jezik (Agić et al., 2015). Može se stoga i zaključiti da ako se budući sastavljači korpusa budu pridržavali međunarodnih smjernica i univerzalnih skupova oznaka za vrste riječi i druge gramatičke kategorije, usklađivanja poput ovoga prikazanog u ovome radu u budućnosti možda više neće biti potrebna.

5 Bibliografija

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richard Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Linden, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Hector Alonso Martınez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze i Daniel Zeman. 2015. *Universal dependencies 1.1.*

Eugenija Barić, Mijo Lončarić, Dragica Malić, Slavko Pavešić, Mirko Peti, Vesna Zečević i Marija Znika. 1997. *Hrvatska gramatika*. Školska knjiga, Zagreb

Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari i Fabio Tamburini. 2006. POS tagset design for Italian. U *Proceedings 5th International Conference on Language Resources and Evaluation – (LREC 2006)*, str. 1396–1401. European Language Resources Association (ELRA), Genova.

Raffaella Bernardi, Andrea Bolognesi, Corrado Seidenari i Fabio Tamburini. 2005. Automatic induction of a POS tagset for Italian. U *Proceedings Australasian Language Technology Workshop -ALTW 2005*, Sydney.

Douglas Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison.* Cambridge University Press, Cambridge.

Maurizio Dardano i Pietro Trifone. 2003. *La lingua italiana: morfologia, sintassi, fonologia, formazione delle parole, lessico, nozioni di linguistica e sociolinguistica* (7. izd.). Zanichelli, Bologna.

Tomaž Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. U *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC'04, str. 1535–1538. European Language Resources Association (ELRA), Paris.

Federico Faloppa. 2011. Numerali. U R. Simone, ur., Enciclopedia dell'italiano, str. 972–974. Istituto dell'Enciclopedia Italiana, Roma.

Josip Jernej. 1990. Riflessioni sulle unità linguistiche chiamate „particelle". *Italica Belgradensia*, 3: 1–4.

Josip Jernej. 2005. *Konverzacijska talijanska gramatika za početnike i napredn*e (11. izd.). Školska knjiga, Zagreb.

Virna Karlić. 2014. *Određenost i neodređenost u srpskom i hrvatskom jeziku*. Neobjavljena doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu.

Maslina Katušić. 1981. Note preliminari sulla traduzione dell'articolo italiano. *Studia Romanica et Anglica Zagrabiensia*, XXVI (1-2): 149-158.

Maslina Katušić. 1982a. Neka razmišljanja o mogućnosti prevođenja □određenog□ i neodređenog□člana. *Strani jezici*, XI (1-2): 17–26.

Maslina Katušić. 1982b. L'articolo italiano: un problema di traduzione (I). *Studia Romanica et Anglica Zagrabiensia*, XXVII (1-2): 145–196.

Maslina Katušić. 1982c. L'articolo italiano: un problema di traduzione (II). *Studia Romanica et Anglica Zagrabiensia*, XXVIII (1-2): 111–166.

Ivana Lalli Paćelat. 2014. *Analiza zakonodavnopravnoga stila hrvatskoga i talijanskoga jezika: unutarjezična, međujezična i prijevodna perspektiva*. Neobjavljena doktorska disertacija, Filozofski fakultet Sveučilišta u Zagrebu.

Maslina Ljubičić. 2000. *Studije o prevođenju*. Hval, Zagreb.

Maria Giuseppa Lo Duca. 2011. Parti del discorso. U R. Simone, ur., *Enciclopedia dell'italiano. Istituto dell'Enciclopedia Italiana*, Roma. Dostupno na: http://www.treccani.it/enciclopedia/parti-del-discorso_(Enciclopedia_dell'Italiano)/

Monica Monachini. 1995. *ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines.* Technical report, Pisa.

Stella Neumann. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German.* De Gruyter Mouton, Boston/Berlin.

Slav Petrov, Dipanjan Das i Ryan McDonald. 2012. A universal part-of-speech tagset. U Proceedings of the 8th International Conference on Language Resources and Evaluation – (LREC 2012), str. 2089–2096. http://arxiv.org/pdf/1104.2086v1.pdf

Prokopis Prokopidis, Vassilis Papavassiliou, Antonio Toral, Marc Poch, Francesca Frontini, Francesco Rubino i Gregor Thurmair. 2012. WP-4.5: Report on the revised Corpus Acquisition & Annotation subsystem and its components. Panacea Project. http://hdl.handle.net/10230/22514

Rema Rossini Favretti. 2000. Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. U R. Rossini Favretti, ur., *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, str. 39–56. Bulzoni, Roma.

[4]Više o projektu na http://universaldependencies.org/hr/pos/index.html [posjećeno 5. rujna 2016.].

Rema Rossini Favretti, Fabio Tamburini i Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. U A. Wilson, P. Rayson, i T. McEnery, ur., *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, str. 27–38. Lincom-Europa, Munich.

Josip Silić i Ivo Pranjković. 2005. *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Školska knjiga, Zagreb.

Lelija Sočanac. 2004. *Hrvatsko-talijanski jezični dodiri: s rječnikom talijanizama u standardnome hrvatskom jeziku i dubrovačkoj dramskoj književnosti.* Nakladni zavod Globus, Zagreb.

Marko Tadić. 1996. Računalna obrada hrvatskoga i nacionalni korpus. *Suvremena lingvistika*, 22(41–42): 603–612.

Marko Tadić. 1998. Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika. *Filologija*, 30–31: 337–347.

Marko Tadić. 2003. *Jezične tehnologije za hrvatski jezik.* Exlibris, Zagreb.

Marko Tadić. 2009. New version of the Croatian National Corpus. U D. Hlaváčková, A. Horák, K. Osolsobě i P. Rychlý, ur., *After Half a Century of Slavonic Natural Language Processing*, str. 199–205. Masaryk University, Brno.

Branka Tafra. 1989. Što su brojevi? (gramatički i leksikografski problem). *Rasprave Zavoda za jezik IFF*,15: 219–237.

Branka Tafra. 2000. Morfološka obilježja brojevnih riječi. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 26: 261–275.

Branka Tafra. 2005. *Od riječi do rječnika*. Školska knjiga, Zagreb.

Fabio Tamburini, Corrado Seidenari, Andrea Bolognesi i Raffaella Bernardi. 2008. Italian Lexical-Classes Definition Using Automatic Methods. U R. Rossini Favretti, ur., *Frames, Corpora and Knowledge Representation*, str. 95–120. Bononia University Press, Bologna.

Fabio Tamburini. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. U R. Rossini Favretti, ur., *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento*, str. 57–73. Bulzoni, Roma.

Fabio Tamburini, Corrado Seidenari, Andrea Bolognesi i Raffaella Bernardi. 2008. Italian Lexical-Classes Definition Using Automatic Methods. U Rema Rossini Favretti, ur., *Frames, Corpora and Knowledge Representation*, str. 95–120. Bononia University Press, Bologna.

Elke Teich. 2003. *Cross-linguistic variation in system and text*. Mouton de Gruyter, Berlin/New York.

Pavao Tekavčić. 1989. Prema kontrastivnoj gramatici tzv. „čestica“ u hrvatskom ili srpskom jeziku i talijanskom jeziku. *Rad JAZU*, 427: 127–194.

Stjepko Težak i Stjepan Babić. 2005. *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje* (15. izd.). Školska knjiga, Zagreb.

Giulia Venturi. 2009. Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana, *Technical report TRIPLE - RTT/1*.

Richard Xiao. 2010. How different is translated Chinese from native Chinese. *International Journal of Corpus Linguistics*, 15(1): 5–35.