Uključivanje korpusa latinskih tekstova CroALa u bazu znanja latinskog jezika LiLa

Soldo, Petar

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA KLASIČNU FILOLOGIJU
SMJER: LATINSKI JEZIK, NASTAVNIČKI
Ak. god. 2020./2021.

Petar Soldo

Uključivanje korpusa latinskih tekstova *CroALa* u bazu znanja latinskog jezika *LiLa*

Diplomski rad

Mentor: dr. sc. Neven Jovanović, red. prof.

Zagreb, svibanj 2021.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Veliko hvala mojem mentoru, prof. Nevenu Jovanoviću, na pomoći i vodstvu tijekom izrade ovog rada. Hvala mu i što me uveo u svijet digitalne humanistike i za njega zainteresirao. Uz to, njemu i svim ostalim nastavnicima Odsjeka za klasičnu filologiju, hvala što su mi otkrili sve bogatstvo hrvatskog latiniteta.

Grazie mille ai colleghi di Milano, specialmente a docenti Marco Passarotti e Favio Cecchini, per grande aiuto e pazienza durante il progetto. Grazie soprattutto per lo sforzo fatto per sormontare gli ostacoli comunicativi.

Na kraju, hvala svim prijateljima i obitelji na nemaloj podršci tijekom pisanja ovog rada i tijekom čitavog studija.

1. Sadržaj

2. Uvod

Velik broj današnjih istraživanja u klasičnoj filologiji (i humanističkim znanostima općenito) koristi mogućnosti raznih digitalnih i računalnih alata i resursa koji su nam uglavnom lako i besplatno dostupni. Neki od najčešćih su korpusi, rječnici, alati za obradu prirodnog jezika i napredno pretraživanje teksta. Budući da ta pomagala, osim što olakšavaju proučavanje poznatih problema, omogućuju i postavljanje sasvim novih istraživačkih pitanja i to za velik broj disciplina, ne čudi što je njihov broj s vremenom narastao do impozantnih veličina. Navedimo samo kako projekt *Corpus Corporum*, koji nastoji na jednom mjestu okupiti čim veći broj korpusa latinskih i grčkih tekstova, broji preko 150 milijuna riječi iz preko 20 korpusa (Passarotti, i dr., 2020., str. 182.), a to je samo djelić dostupnih tekstualnih resursa.

Budući da ti resursi nisu međusobno povezani, što smanjuje njihovu iskoristivost, pokrenut je projekt *LiLa: Linking Latin*, kojeg provodi institut *CIRSCE* milanskog sveučilišta *Università Cattolica del Sacro Cuore*, a čiji je cilj spajanje raznih dostupnih resursa i alata kako bi se omogućilo njihovo čim bolje i lakše korištenje. Jedan od takvih resursa je i korpus tekstova hrvatskih latinista *CroALa: Croatiae auctores Latini*, Filozofskog fakulteta Sveučilišta u Zagrebu. Ovaj rad bavit će se uključivanjem *CroALa*-e, kao tekstualnog resursa, s *LiLa*-om.

U tu svrhu dogovorena je suradnja između članova dvaju sveučilišta. Ta se suradnja trebala ostvariti u obliku *Erasmus+* stručne prakse tijekom koje se autor ovog rada kao student diplomskog studija trebao pridružiti timu u Milanu u razdoblju od ožujka do lipnja 2020. i ondje surađivati na projektu. Plan se izjalovio zbog situacije nastale pandemijom bolesti COVID-19 te se zbog toga odvija u nešto sporijem tempu. U vrijeme pisanja ovog rada projekt još uvijek nije dovršen, a ovaj će se rad baviti dokumentiranjem zadataka obavljenih do travnja 2021.

Najprije ćemo pružiti pregled najvažnijih informacija o *CroALa*-i i *LiLa*-i te iznijeti opći plan suradnje i pregled obavljenih zadataka. Spajanje je zamišljeno u dvije faze: lematizacija *CroALa*-e i spajanje s *LiLa*-om. Najprije prikazujemo odabir reprezentativnog uzorka tekstova iz korpusa, koji će služiti za testiranje lematizatora i tagera. Opisat ćemo pripremu i oblikovanje tekstova za proces tokenizacije te opisati neke od problema koji su se pojavili prilikom obavljanja tog zadatka. Slijede prikaz postupka tokenizacije korpusa, dilema s kojima smo se tijekom njega susreli i prikaz rezultata. Nakon toga prikazujemo proces, probleme i rezultate automatske lematizacije i tagiranja korpusa te neke elemente procjene uspješnosti lematizatora i tagera korištenih na tekstovima.

U radu će se, osim spomenutih, naći još dva segmenta – najava budućih radova na projektu i primjena ovog projekta u nastavi.

U konačnici možemo, kao rezultat projekta, očekivati lematiziranu i gramatički tagiranu *CroALa*-u. Tako označen korpus važan je prvenstveno jer omogućava naprednije pretraživanje teksta (primjerice možemo vrlo jednostavno pretražiti sve oblike nekog glagola ili istražiti koliko se često neko glagolsko vrijeme pojavljuje u tekstu), a samim time i postavljanje kompleksnijih istraživačkih pitanja.

Ovaj je rad, dakle, svojevrsni izvještaj o onome što je do sad u sklopu projekta učinjeno. Cilj je prikazati ne samo pojedine postupke i rezultate projekta, već i njihovu motivaciju ili misaoni proces koji stajao iza svakog od njih. S jedne strane, namjera je takvog pristupa omogućiti čim bolje razumijevanje onoga što je učinjeno kako bi se lakše mogla provjeriti valjanost svakog dijela rada te ukazati na moguće pogreške. S druge strane, detaljna objašnjenja omogućuju eventualno ponavljanje ili korištenje pojedinih postupaka u nekim budućim istraživanjima.

Motivacija za prihvaćanje ovog posla i odabir upravo ove teme proizlazi iz činjenice da se radi o području znanosti koje je vrlo aktualno i trenutno se intenzivno razvija, a nije pretjerano zastupljeno u formalnom obrazovanju klasičnih filologa. Osim toga, dodatna vrijednost rada na ovom projektu je što pruža priliku za sudjelovanje u razvitku i stvaranju pomagala koji mogu poslužiti za daljnja istraživanja svima zainteresiranima za proučavanje tekstova hrvatskih latinista.

3. *CroALa* i *LiLa*

Prije nego što krenemo u opisivanje postupka uključivanja korpusa *CroALa* u bazu znanja *LiLa*, opisat ćemo njihovo porijeklo i namjenu.

3.1. CroALa

CroALa (Croatiae auctores Latini)[1] mrežni je korpus tekstova hrvatskih latinista. Kako navode autori to je „znanstvena i slobodno dostupna digitalna zbirka s međunarodnom recenzijom" koja „okuplja tekstove hrvatskih latinista te autora povezanih s Hrvatskom" (Jovanović, i dr., 2014.). Raspon tekstova seže od srednjeg vijeka (najstariji tekst je epitaf kraljice Jelene iz 976.) sve do najnovijeg doba (najnovija je zbirka latinskih pjesama Ivana Goluba iz 1984.). Jovanović (2020.) u repozitoriju projekta na *GitHubu* navodi kako je 24. veljače 2019. *CroALa* sadržavala 562 dokumenta i 5 892 803 riječi.

Spomenuli smo (a i naslovom sugerirali) da je *CroALa* korpus. Kako bismo spriječili nejasnoće, treba objasniti o kakvoj se vrsti korpusa radi, budući da se korpusom mogu nazivati međusobno značajno različite zbirke tekstova. Vodeći se pregledima korpusne lingvistike McEnery i Wilson (2001.) te McEnery i Hardie (2012.), pokušat ćemo odrediti najvažnija obilježja *CroALa*-e. Neki su korpusi konačne veličine, odnosno postoji zadana količina riječi koju smiju sadržavati i kada se ta granica dosegne, korpusu se prestaju dodavati novi tekstovi. *CroALa* je, s druge strane, zamišljena kao zbirka tekstova koja može neprestano rasti i stalno prihvaća nova izdanja.[2] Nadalje, odrednica je suvremenih korpusa da su strojno čitljivi, a takav je slučaj i s *CroALa*-om. Tekstovi su zapisani kao XML[3] dokumenti sukladno TEI[4] smjernicama za označavanje teksta, o čemu će više riječi biti kasnije. Sljedeće su važno obilježje anotacije, odnosno jezične informacije dodane korpusu. Naime, *CroALa*-i, za razliku od mnogih drugih korpusa, nedostaju jezične anotacije – ona nije niti gramatički tagirana[5] (engl. *POS tag*), niti lematizirana[6]. Dokumenti *CroALa*-e ipak sadrže neke metapodatke o tekstu, primjerice podatke o autoru, godini izdanja, priređivaču, vrsti teksta itd.

[1] http://croala.ffzg.unizg.hr (pristupljeno 12. siječnja 2021.)
[2] U spomenutim pregledima korpusne lingvistike ovakav bi se korpus nazivao engleskim terminom *monitor corpus*.
[3] Za više o XML-u vidi https://www.w3.org/standards/xml/core (pristupljeno 12. siječnja 2021.).
[4] Za više o TEI-u vidi https://tei-c.org/guidelines/ (pristupljeno 12. siječnja 2021.)
[5] „gramatičko tagiranje - pridruživanje oznake za vrstu riječi pojavnicama u korpusu" (prema http://ihjj.hr/mreznik/page/pojmovnik/6/, pristupljeno 12. siječnja 2021.).
[6] "lematiziranje - uspostava kanonskoga oblika pojavnice" (prema http://ihjj.hr/mreznik/page/pojmovnik/6/, pristupljeno 12. siječnja 2021.).

Svakako je još važno napomenuti da *CroALa*-u ne treba smatrati tipičnim lingvističkim korpusom kojem bi jedina namjena bila provođenje lingvističkog istraživanja. Svrha *CroALa*-e je dokumentiranje djela hrvatskih latinista i omogućavanje slobodnog pristupa svima koji ih imaju potrebu ili želju čitati. Ovo je bitno za shvaćanje obilježja *CroALa*-e, a time ćemo moći objasniti i neke izazove s kojima smo se susreli tijekom ovog projekta.

3.2. LiLa

LiLa: Linking Latin[7] baza je znanja latinskog jezika nastala u sklopu istoimenog projekta, čiji je cilj „povezati (...) obilje lingvističkih resursa i alata za obradu prirodnog jezika (NLP)[8] koji su razvijeni do sad, kako bi se premostio jaz između neobrađenih jezičnih podataka, NLP-a i opisa znanja“ (Passaroti, i dr., 2019.). U tom se članku navodi kako je projekt potaknut činjenicom da je, unatoč postojanju velikog broja raznih (računalnih) jezičnih resursa, njihova iskoristivost i interoperabilnost mala zbog značajnih razlika u formatima kojima se koriste.

LiLa koristi infrastrukturu temeljenu na određenim standardima semantičkog weba (eng. *Semantic Web*)[9] i jezikoslovnih otvorenih povezanih podataka (eng. *Linked Linguistic Open Data (LLOD)*)[10]. Za zapisivanje podataka koristi se jezik RDF (eng. *Resource Description Framework*)[11], a za njihovo pretraživanje predviđen je upitni jezik (engl. *query language*) SPARQL[12].

S lingvističke strane, *LiLa* se temelji na leksiku i kao ishodišnu točku postavlja lemu[13]. Naime, leme se nalaze u leksičkim resursima (poput rječnika ili tezaurusa) u obliku rječničkih natuknica. Pronalazimo ih i u tekstualnim resursima (npr. korpusi) u kojima leme, u određenom obliku, sačinjavaju pojavnice[14] (engl. *token*). Na kraju, svaku pojavnicu u tekstu obrađuje i odabrani NLP alat, povezujući je s lemom. Na taj način lema spaja sve komponente *LiLa*-e (Passaroti, i dr., 2019.).

[7] https://lila-erc.eu (pristupljeno 30. siječnja 2021.)
[8] U radu ćemo koristiti standardnu kraticu NLP prema engleskom *Natural Language Processing*.
[9] Za više o semantičkom webu vidi https://www.w3.org/standards/semanticweb/ (pristupljeno 30. siječnja 2021.)
[10] Za više o jezikoslovnim otvorenim povezanim podatcima vidi http://www.linguistic-lod.org/ (pristupljeno 30. siječnja 2021.)
[11] Za više o RDF-u vidi https://www.w3.org/TR/rdf11-primer/ (pristupljeno 30. siječnja 2021.)
[12] Za više o SPARQL-u vidi https://www.w3.org/TR/sparql11-query/ (pristupljeno 30. siječnja 2021.)
[13] „lema (*engl*. lemma) kanonski oblik riječi (u morfologiji i leksikografiji), kanonski oblik pojavnice (u korpusnome jezikoslovlju), tagirana vrijednost“ (prema http://ihjj.hr/mreznik/page/pojmovnik/6/, pristupljeno 30. siječnja 2021.).
[14] „pojavnica – sve što se nalazi između dva znaka koja služe kao graničnici (svako individualno pojavljivanje); svaka pojava jezične jedinice u korpusu, na razini riječi svaki oblik uključen u leksem“ (prema http://ihjj.hr/mreznik/page/pojmovnik/6/, pristupljeno 30. siječnja 2021.).

3.3. Motivacija, ciljevi i tijek suradnje

Budući da *LiLa* pokušava okupiti i povezati čim više resursa za latinski jezik, ne čudi interes za uključivanjem *CroALa*-e među tekstualne resurse. S obzirom na to da je *CroALa* nelematizirana, da sadrži obilje tekstova kasnijeg latiniteta te da su izdanja tih tekstova različitih oblika i grafija, za očekivati je da će se pojaviti izazovi prilikom spajanja spomenutih sustava. Istraživanje takvih problema i pronalaženje njihovih rješenja važno je ne samo kako bismo uključili *CroALa*-u, već i kako bismo olakšali potencijalno uključivanje ostalih nelematiziranih korpusa. *CroALa* je za *LiLa*-u, dakle, interesantna ne samo zbog kvantitete, već i zbog kvalitete. S druge strane, ova će suradnja *CroALa*-i, osim lematizacije, osigurati i pristup ostalim resursima koji su uključeni u *LiLa*-u. Time će se povećati mogućnosti istraživanja i korištenja *CroALa*-e.

Uključivanje *CroALa*-e u *LiLa*-u zamišljeno je u dvije faze – prva faza odnosi se na lematizaciju *CroALa*-e, a druga na uključivanje *CroALa*-e u *LiLa*-u. Lematizacija je nužna jer se korpusi u *LiLa*-u spajaju putem lemā. Proces lematizacije zamišljen je tako da se najprije odabere reprezentativni uzorak korpusa koji se zatim obradi pomoću više dostupnih lematizatora i tagera te se rezultat evaluira kako bi se pronašao idealni alat. Kada su lematizator i tager odabrani, na čitavom se uzorku provjerava njihova uspješnost. Za drugu fazu, odnosno za uključivanje u *LiLa*-u, predviđeno je da se najprije provjeri kakva je povezanost lemā *CroALa*-e s *LiLa*-inim lemama. Zatim treba pojavnice *CroALa*-e spojiti u RDF triplete (vidi poglavlje 5.) s lemama *LiLa*-e, a na kraju treba dodati metapodatke.

Kao što je spomenuto u uvodu, projekt nije dovršen, već je odrađen samo začetni dio prve faze, odnosno na odabranom uzorku korpusa testiran je niz alata za lematizaciju i tagiranje, a obrada rezultata još je u tijeku. Stoga ćemo u nastavku rada prikazati kako je tekao proces lematizacije i što je sve u sklopu projekta učinjeno.

4. Lematizacija *CroALa*-e

U ovom ćemo dijelu rada detaljnije opisati pojedine segmente prve faze projekta, odnosno lematizacije *CroALa*-e. Najprije ćemo prikazati odabir reprezentativnog uzorka i njegovu pripremu za obradu, zatim ćemo opisati proces automatske lematizacije i tagiranja, a na kraju ćemo prezentirati evaluaciju i odabir idealnog alata. Općeniti je cilj ovog koraka pripremiti *CroALa*-u za spajanje s *LiLa*-om.

4.1. Odabir uzorka

Kao što smo već spominjali, kako bi se *CroALa*-u moglo uključiti u *LiLa*-u, potrebno je tekstove u njoj lematizirati. Postoji više lematizatora za latinski jezik pa je potrebno izabrati onaj koji će optimalno obaviti posao. Da bi smo vidjeli koji nam od programa najviše odgovara, treba obraditi tekstove pomoću više njih i zatim provjeriti kvalitetu rezultata. Budući da je kontrola točnosti lematizatora posao koji obavlja čovjek, riječ po riječ, bilo bi izrazito vremenski zahtjevno tako pregledati čitav korpus. Iz tog je razloga potrebno stvoriti reprezentativan uzorak korpusa na kojem će se moći testirati i potom odabrati optimalni lematizator.

Prema Manningu i Schützeu (1999., str. 119.) uzorak je reprezentativan „ako ono što otkrijemo o uzorku vrijedi i za opću populaciju“. Drugim riječima, treba odabrati takav podskup tekstova *CroALa*-e za koji će rezultati lematizacije pojedinim lematizatorima biti u čim većoj mjeri jednaki rezultatima koje bismo dobili lematizacijom čitavog korpusa. Dakle, osobine našeg uzorka moraju odgovarati osobinama čitavog korpusa. Kako bismo to mogli postići, trebamo najprije dobro proučiti kako izgleda *CroALa* i koje su njene osobine.

Već smo spominjali da *CroALa* sadrži tekstove čiji vremenski raspon iznosi gotovo tisuću godina, no to nije jedini faktor koji pridonosi raznolikosti korpusa. U korpusu ćemo naći djela koja pripadaju raznim književnim rodovima i vrstama – epistolografska i historiografska djela, duže i kraće pjesme, epove, epigrame, elegije, govore itd. S druge strane, nisu svi tekstovi koji čine *CroALa*-u nužno samo književna djela, već ćemo naći druge vrste tekstova poput epitafa (npr. *Epitaf kraljice Jelene*), pravnih tekstova (npr. *Supetarski kartular*) ili čak rječnika (npr. mali hrvatsko-latinski rječnik Bartula Đurđevića). Budući da svaka od tih formi (književnih i neknjiževnih) u nekoj mjeri utječe na jezična obilježja teksta svojim uzusima ili zakonitostima, svakako je za očekivati da će žanrovska raznolikost utjecati na raznolikost jezika. U *CroALa*-i je zastupljeno preko 250 autora (ne računajući tekstove čiji nam autori nisu poznati), koji zasigurno nisu svi pisali na isti način, pa i različiti stilovi pisanja pridonose raznolikosti korpusa. Još jedna činjenica o *CroALa*-i koju treba uzeti u obzir kada razmišljamo o reprezentativnom

uzorku su različiti načini na koji su tekstovi dospjeli u zbirku. Naime, kako navode urednici, „neki od tekstova uključenih u zbirku *CroALa* nastali su jednostavnim digitaliziranjem, skeniranjem ili prijepisom starijih izdanja, izvornih ili kasnijih znanstvenih, dok su drugi tekstovi plod modernoga filološkog rada i digitalne inačice suvremenih kritičkih izdanja“ (Jovanović, i dr., 2014.). To što su tekstovi potekli iz različitih vrsta izdanja nam je važno jer i to može utjecati na osobine teksta. Primjerice, urednici mogu odlučiti ostaviti izvornu grafiju teksta ili tekst mogu „standardizirati“. Mogu birati između čuvanja teksta čim bližim izvorniku ili približavanja teksta suvremenom čitatelju. Sve to važno je u nekoj mjeri prikazati u našem uzorku.

Faktori raznolikosti koje smo naveli bili su glavna misao vodilja prilikom odabira tekstova za uzorak, ali postoje i drugi kriteriji koji su oblikovali njegov izgled. Što je sve služilo kao motivacija kod probira tekstova, najlakše je objasniti primjerom, stoga ćemo sada objasniti koji su tekstovi i zašto odabrani za uzorak. Preporučena veličina uzorka (prema onome što su savjetovali kolege iz Milana) je otprilike 50 000 pojavnica, a odlučeno je i da će se koristiti čitavi tekstovi, a ne izvadci. I ta je zadana veličina i forma uzroka utjecala na odabir tekstova.

U tablici ćemo prikazati kojih je 13 tekstova odabrano, zajedno s autorima tih tekstova, godinama nastanka ili izdanja, žanrovima te brojem riječi. Tekstovi su poredani kronološkim redom, od najstarijeg prema najnovijem. Podaci koji se nalaze u tablici dobiveni su pretraživanjem metapodataka koji se nalaze u XML datotekama u kojima su tekstovi spremljeni. Popis spomenutih datoteka i njihov izvor nalaze se u Prilog 1. Popis datoteka korištenih za dobivanje uzorka korpusa

Autor	Naziv djela	Godina	Žanr	Broj riječi
Auctores varii	*Jura sancti Petri de Gomai*	1080	prosa - charta	6348
Modruški, Nikola	*Nauicula Petri*	1463	prosa - tractatus; prosa - epistula	6607
Šižgorić, Juraj	*Prosopopeya*	1469	poesis - elegia	1000
Marulić, Marko	*Carmina Latina (excerpt 008)*	1477	poesis	4530[15]
Šižgorić, Juraj	*Odae de apostolis*	1487	poesis - oda; prosa oratio - epistula	2640
Bunić, Jakov	*De raptu Cerberi*	1490	poesis - epica	6750
Crijević Tuberon, Ludovik	*Commentariolus de origine et incremento urbis Rhacusanae*	1520	prosa oratio – historia	5404
Andreis, Franjo Trankvil	*Epistolae ad Thomam Nadasdinum*	1532	prosa oratio - epistula	5603
Beneša, Damjan	*Epicedion in morte Jacobi Boni*	1534	poesis - carmen; poesis - epicedion	1707[16]
Gradić, Stjepan	*Oratio de eligendo Summo Pontifice*	1667	prosa - oratio	2879
Bošković, Ruđer	*Ecloga recitata in publico Arcadum consessu*	1753	poesis - ecloga	3023
Kunić, Rajmund	*Ex Graeco Homeri Hymnus ad Cererem*	1794	poesis - hymnus; versio	3992
Milašinović, Franjo	*Viator Zagorianus Jožko Hranjec*	1850	poesis - carmen macaronicum	1282
			UKUPNO RIJEČI:	51 765

Tablica 1. Popis tekstova koji sačinjavaju uzorak CroALa-e za ispitivanje lematizatora

[15] U XML-u ovog dokumenta navodi se samo ukupni broj riječi svih Marulićevih *Carmina Latina*, a ne samo ovog osmog dijela koji mi koristimo, pa je broj riječi preuzet iz druge datoteke u kojoj su popisani brojevi riječi, a koji je dostupan na https://github.com/nevenjovanovic/croatiae-auctores-latini-textus/blob/master/croala-wordcounts.xml (pristupljeno 13. veljače 2021.)

[16] U XML datoteci ovog teksta nije bio dostupan broj riječi pa je on preuzet sa stranice http://croala.ffzg.unizg.hr/cgi-bin/getobject.pl?c.223:0.croala (pristupljeno 13. veljače 2021.) iz drugog dokumenta koji sadrži ranije izdanje istog teksta, ali se broj riječi provjereno poklapa.

Objasnimo ukratko motivaciju za uključivanje pojedinih tekstova u uzorak.

Prvi tekst na popisu, *Supetarski kartular* ili *Jura sancti Petri de Gomai*, odabran je kao predstavnik dvije kategorije zanimljive za naš uzorak – radi se o neknjiževnom tekstu (u pitanju je kopijalna knjiga) te pripada razdoblju srednjeg vijeka (Novak & Skok, 1952.). Tekst Nikole Modruškog, *Nauicula Petri*¸ primjer je proze ranog humanizma, a kao primjer poezije istog razdoblja nalazimo Šižgorićeve *Prosopopeyu* i *Odae de apostolis* (potonje su interesantne i jer sadrže prozni uvod). Njima se pridružuju i Marulićeve *Carmina Latina*, pjesme zanimljive zbog raznolike metrike, te *De raptu Cerberi*, kao primjer humanističkog epa. Ovi tekstovi odabrani su ne samo kako bi pokrili period 15. stoljeća i obuhvatili razne žanrove, već i jer su im izdanja raznolika. Primjerice, u nekim od ovih tekstova latinski diftong *ae* zapisuje se digrafom *ae* (npr. *Odae*), dok se kod ostalih koristi tzv. *e caudata*, odnosno grafem *ę* (npr. *Prosopopeya*). Neki od njih razlikuju pri pisanju grafeme *u* i *v* (npr. *Odae*)¸ dok se kod nekih koristi samo grafem *u*, odnosno *V* (npr. *Nauicula Petri*). Neka od njih sadrže komentare urednika (kod *Supetarskog kartulara* čak se radi o komentarima na hrvatskom), dok su neka „čisti tekst". Nešto su kasniji tekstovi Crijevićev *De origine et incremento urbis Rhacusanae*¸ odabran kao primjer historiografske proze, i *Epistolae* Franje Trankvila Andreisa, koje služe kao predstavnik epistolarne proze. Benešin *Epicedion* za Bunića odabran je kao primjer prigodne poezije, a zanimljivo nam je i njegovo izdanje jer se u njemu nalaze popunjene lakune te su sačuvane neke pravopisne greške u latinskom tekstu, poput *eciam* umjesto *etiam*. S relativno velikim vremenskim odmakom, sljedeći je, kao predstavnik govorā uzet Gradićev *Oratio de eligendo Summo Pontifice*. Boškovićeva *Ecloga* interesantna je, osim zbog vremena i žanra, i zbog svojevrsne dijaloške forme. Na kraju nalazimo dva jezično zanimljiva teksta. Prvi, Kunićev *Hymnus ad Cererem*, prijevod je s grčkog na latinski te sadrži i nekoliko grčkih riječi, a drugi, *Viator Zagorianus Jožko Hranjec* Franje Milašinovića, interesantan nam je jer je primjer makaronske poezije, koja spaja leksik vernakulara s morfologijom i sintaksom latinskog[17]. Ta su dva teksta uzeta, među ostalim, da provjerimo kako se pojedini lematizatori i tageri ponašaju na takvim tekstovima.

Osim kvalitativne analize uzorka, možemo napraviti i kratku kvantitativnu analizu. Pogledamo li raspodjelu djela po vremenskim periodima, uočit ćemo da najviše djelā potječe iz perioda između 1451. i 1550. godine (ukupno 48,5%) te da je sljedeći najzastupljeniji period druga polovica 18. st. (s 11,02% ukupnih djelā). Situacija s vremenskom raspodjelom u uzorku je vrlo

[17] makaronsko pjesništvo. *Hrvatska enciklopedija, mrežno izdanje.* Leksikografski zavod Miroslav Krleža, 2021. Pristupljeno 13. veljače 2021. <http://www.enciklopedija.hr/Natuknica.aspx?ID=38238>.

slična – i u njemu je najčešći period razdoblje između 1451. i 1550. (pripada mu 61,54% djelā) te je opet drugi period po zastupljenosti druga polovica 18. st. (15, 38%). Možemo usporediti i broj djelā po žanrovima. Vidjet ćemo da u korpusu poeziji pripada 40,28% djelā, prozi 53,9%, a „miješanih“[18] djelā je 6,01%. U uzorku je stvar malo drugačija jer na poeziju otpada 53, 85% djelā, na prozu 38,46%, a miješanih je 7,69%. Dakle, po pitanju žanra uzorak nije sasvim vjerna preslika korpusa, ali ipak daje prihvatljivu okvirnu sliku.

Podaci o vremenskoj i žanrovskoj zastupljenosti dobiveni su dohvaćanjem i obradom podataka iz XML datoteka, a uz to su provjereni i broj riječi po žanrovima i po periodima. Nakon dohvaćanja, podatci su ručno provjeravani i popravljani. Detaljnija analiza ovih podataka (koja je u jednoj mjeri odrađena) mogla bi biti zanimljiva i važna za projekt u budućnosti i za proučavanje *CroALa*-e općenito, ali bismo njenim prikazom izašli van okvira ovog rada i odveć skrenuli s teme. Iz toga razloga, *XQuery* upiti i rezultati koji su korišteni nisu izneseni u ovom radu, ali su dostupni na upit autoru.

Važno je napomenuti još dvije stvari. Prva je napomena da je vjerojatno postajala određena latentna pristranost prema tekstovima koje je autor rada poznavao od prije. Druga stvar koju treba imati na umu je da ovaj uzorak predstavlja trenutno stanje korpusa, ali treba se sjetiti i da se radi o korpusu koji je predviđen da raste i prima nove tekstove. Stoga će u budućnosti, ako bude potrebe za razmatranjem osobina korpusa, trebati pratiti promjene i novosti u korpusu.

4.2. Priprema uzorka za lematizaciju i tagiranje

Nakon što je uzorak odabran, treba ga pripremiti za lematizaciju. Ta priprema podrazumijeva da najprije dobijemo „čisti tekst“ koji ćemo zatim tokenizirati. Premda se taj zadatak čini vrlo jednostavnim, postoje određeni izazovi koje treba svladati. Tekstovi koji nam trebaju nalaze se spremljeni u XML datotekama, stoga ćemo prije nego što opišemo postupak i zahtjeve pripreme uzorka, ukratko opisati strukturu tih datoteka.

4.2.1. Struktura XML datoteka *CroALa*-e

XML (engl. *Extensible Markup Language*) je označiteljski jezik (engl. *markup language*), odnosno format namijenjen zapisivanju strukturiranih podataka koji podatke zapisuje u obliku teksta (Bagić Babac & Kušek, 2011.). Osnovna je gradivna jedinica XML dokumenta element

[18] Pod „miješana djela“ misli se na djela koja su u svojim XML datotekama označena i kao proza i kao poezija, najčešće jer dio teksta pripada jednoj, a dio teksta drugoj književnoj vrsti.

koji sadrži podatke koje želimo pohraniti, a sastoji se od otvarajuće oznake, sadržaja i zatvarajuće oznake. Ovako bi izgledao jedan XML element:

```
<title>Epigrammatum liber III, versio electronica</title>
```

Imena oznaka su proizvoljna i njima bilježimo metapodatke, odnosno koju vrstu podatka smo zapisali u sadržaju. Elemente možemo i „ugnijezditi“ jedne u druge, a svakom elementu možemo dodati i atribut kojim ćemo prenijeti dodatne informacije o sadržaju. Ovo je primjer ugniježđenog elementa koji sadrži i atribut:

```
<author ref="benes01">
    <persName xml:lang="hr">Beneša, Damjan</persName>
    <date>1476-1539</date>
</author>
```

Velika je prednost XML-a što je vrlo često ne samo strojno čitljiv, već i prilično razumljiv ljudima.

Tekstovi *CroALa*-e pohranjeni su u obliku XML dokumenata koji strukturom prate TEI (engl. *Text Encoding Initiative*) standard. TEI standard propisuje način zapisa nekog teksta u XML formatu kako bi se olakšala računalna obrada tog teksta, a to postiže zadanom shemom dokumenta i zadanim skupom XML oznaka (TEI Consortium, 2021.).

Dva su osnovna dijela svakog dokumenta u *CroALa*-i – podatci o tekstu, pohranjeni u element `<teiHeader>`, i sam tekst koji se nalazi u elementu `<text>`. Oba ova elementa imaju u sebi ugniježđene druge elemente koji preciziraju o kakvom se sadržaju radi. U elementu `<teiHeader>` na taj su način označeni podatci o djelu, autoru, izdanju, broju riječi, žanru i sl. Potkategorije elementa `<text>` mogu nam ukazivati na dijelove teksta poput naslova, uvoda, poglavlja, redaka itd., ali se njima povremeno i obilježava značenje ili sadržaj nekog dijela teksta. Primjerice, u nekim dokumentima će dijelovi teksta koji se odnose na ime nekog mjesta biti omeđeni oznakom `<placeName>`, dok će komentari priređivača teksta biti označeni oznakom `<note>`.

Budući da nam za lematizaciju trebaju tekstovi, proveden je proces njihovog izvlačenja iz XML dokumenata, koji ćemo sada i opisati.

4.2.2. Dohvaćanje i priprema tekstova iz datoteka

Da bismo dobili tekstove u obliku koji nam je potreban za lematizaciju, potrebno ih je na primjeren način dohvatiti iz datoteka. Za taj smo zadatak koristili upitni jezik *XQuery*, koji služi za odabir i transformaciju dokumenata u XML-u (Bagić Babac & Kušek, 2011., str. 36.). Pomoću *XQueryja* možemo, dakle, odabrati samo onaj dio dokumenta koji sadrži tekst. To smo napravili pomoću kratke *XQuery* skripte koja se nalazi u Prilog 2. *XQuery* skripta za dohvaćanje tekstova Svaki je tekst uzorka spremljen u zasebnu tekstualnu datoteku. Međutim, ovdje posao nije gotov. Naime, zbog strukture dokumenata i načina lematizacije, ovako dobivene tekstove treba dodatno „pročistiti“.

Za lematizaciju nam je potreban „čisti“ latinski tekst, odnosno tekst koji ne sadrži nikakve naknadne dodatke ili nešto što ne bismo smatrali dijelom izvornog teksta. Ovdje ubrajamo dodatke poput kritičkog aparata, komentara, numeracije i sl. Premda bi odabir „čistog“ teksta inače mogao predstavljati kompleksno tekstološko pitanje, u našem slučaju situaciju pojednostavljuje to što imamo jasnu ideju i cilj zbog kojih uređujemo tekst. Za nas je važno što lematizatori, zbog načina na koji su građeni, optimalno djeluju na koherentnom latinskom tekstu. To znači da moramo prirediti „destilirani“ tekst koji se sastoji od niza koherentnih rečenica. U tu svrhu iz tekstova su uklonjene numeracije i slične tekstualne oznake, znakovi nastali greškom prilikom digitalne obrade (poput grešaka nastalih pri optičkom prepoznavanju znakova, dohvaćanja teksta iz XML datoteka i sl.), priređivački komentari i kritički aparat. Da bi tekst bio čim koherentniji, ostavljeni su naknadni ispravci teksta i dopune lakuna, kao i eventualni originalni komentari autora (s time da su preseljeni na kraj teksta kako ne bi prekidali rečenice). Ponegdje su tekstovi dopunjeni znakovima koji su „ispali“ negdje u procesu digitalne obrade ili su dodani radi sustavnosti i lakše lematizacije (npr. nadodani rečenični znakovi ondje gdje nedostaju).

U našem slučaju ovaj proces naknadnog uređivanja uzorka odrađen je ručno, odnosno tekstovi su prepravljeni nakon što smo ih izvukli iz XML datoteka. Međutim, kada bismo pokušali na isti način „pročistiti“ čitavu *CroALa*-u (a što će nam u slučaju njenog čitavog spajanja s *LiLa*-om biti i potrebno), shvatili bismo da se radi o izrazito vremenski zahtjevnom poslu. Stoga se postavlja pitanje – postoji li mogućnost barem djelomične automatizacije ovog posla?

Ovdje se trebamo sjetiti ranije spomenutih potkategorija elementa `<text>`. Spomenuli smo da se pomoću njih mogu dodatno naznačiti pojedini dijelovi teksta, što je u dokumentima *CroALa*-e i učinjeno. Svi gore navedeni segmenti, poput komentara ili kritičkih aparat, uglavnom se

nalaze u zasebnim elementima. Tako ćemo u dokumentima naići na elemente <note>, <supply>, <app> ili <rdg> koji uglavnom naznačuju nešto što nije dio „izvornog“ teksta.[19] Čini se da je situacija idealna za automatizaciju – iz odabira treba samo izostaviti ove dijelove teksta (što je tehnički vrlo lako izvedivo). Ipak, zbog izbora elemenata i načina na koji su ti elementi korišteni u *CroALa*-i, jednostavno isključivanje ne bi nam dalo dobre rezultate.

Naime, nailazimo na dva problema pri pregledu elemenata ugniježđenih u element <text>. Jedan je problem što nisu svi nama „nepoželjni“ dijelovi teksta nužno odvojeni zasebnim elementom, a drugi je činjenica da bismo ponekad željeli i zadržati sadržaj nekog od gore navedenih elemenata. Situaciju je lakše shvatiti ako se pokaže na nekoliko primjera.

U dokumentu u kojem je pohranjen *Supetarski kartular* se, primjerice, hrvatski komentar priređivača Skoka i Novaka nalaze u jednakom elementu <p> kao i sav ostali izvorni latinski tekst. To znači da ni po čemu ne bismo mogli selektirati samo latinski tekst. Kad već spominjemo ovaj dokument možemo spomenuti i da smo kod njega naišli na dijelove teksta koji su, izgleda, plod digitalizacijskog lapsusa, pa je, primjerice, bilo potrebno ispraviti pojavnice poput „comparaui87"“ jer su sadržavale neželjene znakove.

Još jedan problem uočit ćemo usporedimo li korištenje elementa <note> u Kunićevom *Ex Graeco Homeri Hymnus ad Cererem* i Benešinom *Epicedionu*. Naime, u prvom se dokumentu u elementu <note> nalaze izvorni autorovi komentari, koji se tiču njegova prevođenja s grčkog, dok se u drugom dokumentu, također u elementu <note>, nalaze naknadni komentari i kritički aparat priređivača. Stoga, ako želimo ostaviti samo izvorne komentare autora, ali izostaviti naknadne komentare priređivača, ne možemo primijeniti jednaku naredbu („izostavi element <note>“) u oba teksta.

Kod Beneše možemo pronaći još jednu prepreku za potpunu automatizaciju. Naime, pri korištenju elementa <supplied>, koji se ovdje koristi za popunjavanje lakuna, jednako se tretiraju slučajevi gdje se dodaje čitava riječ ili pojedino slovo u riječi. Tako ćemo naići na slučaj:

```
<supplied>quanta</supplied>,
```

ali i na slučaj

```
m<supplied>o</supplied>rab<supplied>o</supplied>r.
```

[19] Svi elementi koji se koriste u *CroALa*-inim dokumentima u skladu su s TEI standardom pa se tako upute za njihovo korištenje i njihovo „značenje“ mogu pronaći u smjernicama koje je TEI objavio (TEI Consortium, 2021.).

Ovo je problem, jer će se prilikom upita riječ „morabor“ rastaviti i dobit ćemo rezultat „m o rab o r“. Budući da se elementi korišteni za naknadno dodavanje slova i riječi ni po čemu ne razlikuju, ne možemo kod dohvaćanja teksta napraviti zahtjev koji bi glasio „ako <supplied> sadrži dio riječi, spoji ga s okolnim slovima, a ako sadrži čitavu riječ ostavi ga kakav je“.

Pogledajmo i još jedan slučaj, u kojem nam je automatizacija otežana. Kod Bunićeva *De raptu Cerberi*, na mjestima je, umjesto jedinstvenog teksta, ostavljena mogućnost odabira između više čitanja (odnosno odabira između više rukopisa) te se oni nalaze u posebnom elementu <app>. Za razliku od kritičkog aparata kakav se nalazi kod Modruškog ili Beneše, gdje se u elementu <note> nalazi alternativno čitanje, ali je priređivač odabrao jedno za tekst, u ovom dokumentu nijednom čitanju nije dan primat. Drugim riječima, sva su čitanja odijeljena u element i, kada bismo ih samo ignorirali, dobili bismo nepotpun tekst. Ovako izgleda jedan takav slučaj:

```
<l n="2.53">Si durum <app>
          <rdg wit="#R1">flegetonta</rdg>
          <rdg wit="#Sn">Phlegethonta</rdg>
          <rdg wit="#R2 #Gl">Phlegetonta</rdg>
      </app> domas legesque deorum</l>.
```

Vidimo da su oznake za različite rukopise naznačene kao atributi elemenata (wit=""). Međutim, nisu uvijek navedeni svi rukopisi te mogu doći u raznim kombinacijama.

Posljednja dva slučaja, s elementima <supplied> i <app>, čak dopuštaju određena tehnička rješenja za (djelomičnu) automatizaciju, ali bi se ona morala raditi na razini individualnog dokumenta.

Ostavimo li pitanje automatizacije po strani i fokusiramo li se na činjenicu da je dugoročna želja ovog projekta priključiti čitavu *CroALa*-u *LiLa*-i, trebamo na umu imati jednu važnu činjenicu – dokumenti koji čine *CroALa*-u međusobno se mogu značajno razlikovati u strukturi. To znači da će biti potrebno smisliti rješenja za gore navedene probleme u načinu zapisa, koji vrlo vjerojatno nisu jedini s kojima bismo se susreli pri pregledavanju čitavog korpusa.

Premda je jedna mogućnost da jednostavno ručno uredimo svaki tekst, trebamo se sjetiti da *CroALa* ima gotovo 500 tekstova i preko 5 000 000 riječi, pa bi taj posao iziskivao golemu količinu vremena.

Jedno je od mogućih rješenja, koje je djelomično provedeno tijekom ovog projekta, mijenjanje ili dodavanje novih informacija u postojeće XML dokumente. Tako je prof. Jovanović, u dokumentima korištenima za uzorak, prepravio elemente `<note>` koji sadrže komentare na način da im je dodao atribut `@ana`. Taj atribut poprima vrijednost „`authorial`" ako je komentar napisao sam autor teksta, a komentar „`editorial`" ako komentar pripada uredniku ili priređivaču teksta. Elementu `<supplied>` dodan je atribut `@scope`, koji poprima vrijednost „`incipit`" ako je dodan početak riječi, vrijednost „`medium`" ako je nadopunjena sredina, vrijednost „`finis`" ako je dodan kraj riječi, a vrijednost „`verbum`" ako je dodana čitava riječ. Tekstovi koji su uređeni na taj način dostupni su na poveznici https://github.com/nevenjovanovic/croatiae-auctores-latini-textus/tree/master/subset (zadnje pristupljeno 6. ožujka 2021.).

Uočavanje ovakvih pojava u strukturi dokumenata koji čine korpus, važno je kako za naš projekt, tako i za *CroALa*-u u cjelini, ali može služiti i kao dobar poticaj za razmatranje problema digitalizacije znanstvenih izdanja tekstova.

4.2.3. Tokenizacija uzorka

Jednom kada smo dobili „čisti" tekst možemo pristupiti tokenizaciji. Bekavac (2001.) tokenizaciju objašnjava kao „dovođenje korpusa u stanje u kojem su sve riječi-pojavnice identificirane i eksplicitno obilježene". U našem slučaju to bi značilo da ćemo sve riječi, tj. pojavnice iz korpusa prikazati u tekstualnoj datoteci, na način da svaka pojavnica bude u zasebnom redu, a da se između svake rečenice nalazi prazan red. Uz tokenizaciju pojavnica, bit će nam potrebna i rečenična tokenizacija u kojoj će u zasebni red biti stavljena svaka rečenica (ovaj put bez praznina između redova). Ovako oblikovani tekstovi nužni su nam za lematizaciju.

Opet se nalazimo pred zadatkom koji se možda čini banalnim, ali u sebi krije teorijski problem – ako trebamo rastaviti tekst na riječi, kako ćemo definirati riječ? Slično pitanje možemo postaviti i u slučaju rastavljanja na rečenice, odnosno možemo se pitati što je granica rečenice. Problem manjka jasnog odgovora na pitanje „što je riječ"? nije novost u lingvistici (Bender, 2013.). U našem slučaj taj se problem konkretizirao u, primjerice, dilemi s enklitikama, odnosno pitanju trebamo li enklitike poput *-que* i *-ue* tretirati kao zasebne riječi pa ih tokenizirati ili ih treba ostaviti uz riječ na koju se naslanjaju. Srećom, u mogućnosti smo preuzeti praksu kolega iz Milana pa je odlučeno da se enklitike neće zasebno tokenizirati budući da je to do sad bila češća praksa u *LiLa*-i. Opravdanje za takav pristup je da bi odvajanje enklitika moglo dodatno

zakomplicirati posao, a prilikom lematizacije nema toliko ključnu ulogu, koliku bi imalo prilikom, primjerice, sintaktičke analize korpusa. Naravno, ako se pokaže potreba, enklitike je moguće odvojiti i naknadno.

Kod rečenica nas muči problem koje ćemo interpunkcije uzeti kao graničnike. Naime, jasno je da ćemo točku ili upitnik smatrati krajem rečenice, ali postavlja se pitanje što raditi u slučaju dvotočke, točke sa zarezom ili crtice. Budući da ovi znakovi ponekad mogu rečenicu dijeliti na dva dijela od kojih bi svaki mogao funkcionirati kao zasebna rečenica, legitimno je pitati se treba li i njih uzeti kao granicu rečenice prilikom rečenične tokenizacije. I ovdje smo se poveli za praksom uvriježenom u ranijim *LiLa*-inim projektima pa je su za granicu rečenice odabrani samo „snažnije" interpunkcije, odnosno točka, upitnik i uskličnik. Razlog za ovakvu odluku je na iskustvu temeljeno uvjerenje da dijelovi rečenice odijeljeni „slabijim" interpunkcijama (dvotočka, točka sa zarezom i crtica) u najvećem broju slučajeva ipak čine cjelinu.

Tokenizacija pojavnica, odnosno riječi, provedena je pomoću programa napisanog u programskom jeziku *Perl*[20], čiji su autori Helmut Schmid i Serge Sharoff, a koja se može preuzeti putem poveznice http://corpus.leeds.ac.uk/tools/utf8-tokenize.pl (preuzeto 28. svibnja 2020.).[21] Ovaj je tokenizator inače integralni dio alata za tagiranje i lematizaciju *TreeTagger*[22], ali se u ovoj fazi koristi odvojeno. S obzirom na to da tokenizator kao kraj rečenice prepoznaje, među ostalim, točku, bilo je potrebno sastaviti i popis kratica koje završavaju točkom, kako ih program ne bi neispravno označio kao kraj rečenice. Popis kratica koji je korišten prilikom tokenizacije nalazi se u Prilog 3. Popis kratica korištenih u tokenizaciji Taj smo popis u obliku tekstualne datoteke aktivirali prilikom pokretanja skripte, a uz njega je kao parametar odabrana i opcija „`-i`" koja naznačuje da se tokenizira talijanski jezik, jer je ta opcija potrebna prilikom tokenizacije latinskog. Dobiveni su rezultati spremljeni u obliku tekstualnih datoteka.

Tokenizacija rečenica obavljena je pomoću kratke skripte napisane u programskom jeziku *Python*, koristeći knjižnicu *NLTK* (engl. *Natural Language Toolkit*). Skripta se nalazi u Prilog 4. *Python* skripta za rastavljanje rečenica Dobiveni su rezultati u obliku tekstualnih datoteka, u kojima se svaka rečenica nalazi u zasebnom retku. Rezultati su potom ručno provjereni i mjestimično ispravljeni.

[20] Vidi https://www.perl.org/ (pristupljeno 13. ožujka 2021).
[21] Navedena poveznica direktno vodi na preuzimanje skripte, a za nešto više informacija vidi http://corpus.leeds.ac.uk/tools/ (pristupljeno 13. ožujka 2021).
[22] Vidi https://www.cis.lmu.de/~schmid/tools/TreeTagger/ (pristupljeno 13. ožujka 2021).

Rezultat tokenizacije je 49 879 tokena, ne računajući interpunkcijske znakove. S ovako uređenim tekstovima mogli smo pristupiti lematizaciji.

4.3. Obrada uzorka algoritamima za lematizaciju i tagiranje

Cilj svih gore navedenih pripremnih radnji je lematizacija i tagiranje uzorka, koja je nužni preduvjet za uključivanje *CroALa*-e u *LiLa*-u. U trenutku pisanja rada suradnja i projekt na ovom su koraku, nažalost, stali, a nastavak se može očekivati u narednom vremenu.

Ranije smo spomenuli da je lematizacija nužna jer se korpusi u *LiLa*-u povezuju putem leme (v. str. 5). Ipak, osim lemā, pojavnicama su pridruženi i gramatički, odnosno POS tagovi (engl. *POS*, tj. *part-of-speech tags*)[23], budući da se POS tagovi u nekim slučajevima koriste za odabir ispravne leme, ali i jer se tagovi također bilježe u bazi znanja. Podsjetimo, POS tagiranje znači određivanje kojoj vrsti riječi pojavnica pripada.

Prije no što krenemo u opis lematizacije i tagiranja uzorka, treba kratko spomenuti i osvrnuti se na jedan rad koji je svojom temom vrlo blizak ovom radu, pogotovo u pitanjima obrađenim u ovom poglavlju. Naime, tijekom pisanja ovog rada nastao je i diplomski rad Federice Gamba (2020.) koji se bavi uključivanjem novog tekstualnog resursa, kasnoantičke komedije *Querolus,* u *LiLa*-u. Neki su Gambini koraci u lematizaciji i tagiranju slični ili identični postupcima iz ovog rada. Ipak postoje stanovite razlike između ova dva rada, pogotovo u broju i vrsti algoritama koji su korišteni te u problemima koji proizlaze iz osobina teksta na kojem se lematizacija provodi. Ključno je što je u spomenutom diplomskom radu korišten jedinstven tekst, dok se kod nas radi o uključivanju čitavog korpusa, sastavljenog od različitih tekstova.

U ovom ćemo poglavlju predstavit tri skupine podataka: algoritme, korpuse i skupine tagova. Premda su sva tri tipa podataka međusobno povezana i često neodjeljiva, radi preglednosti su smješteni u zasebna potpoglavlja. Ukratko ćemo opisati njihove glavne karakteristike.

4.3.1. Lematizatori i tageri korišteni za označavanje uzorka

Na našem uzorku isprobano je ukupno 9 algoritama za automatsku lematizaciju i tagiranje, od kojih 4 obavljaju samo POS tagiranje, 2 samo lematizaciju, a 3 algoritma obavljaju i POS tagiranje i lematizaciju. Dva spomenuta algoritma-lematizatora[24] dolaze spremna za korištenje, bez potrebe za dodatnim modifikacijama ili pripremama, dok su ostali algoritmi iziskivali treniranje na unaprijed anotiranim korpusima. Naime, radi se o algoritmima koji koriste

[23] U ostatku rada umjesto izraza *gramatičko tagiranje* koristit će se engleska varijanta *POS tagiranje* kako bi se smanjila mogućnost zabune te olakšalo uspoređivanje i istraživanje, jer taj oblik prevladava u korištenoj literaturi.
[24] Radi se o *LEMLAT*-u i *LatMor*-u.

nadzirano učenje (engl. *supervised learning*), odnosno algoritmima koji na unaprijed označenim podacima uče kako novom ulaznom podatku pridodati ispravan izlazni podatak (Jurafsky & Martin, 2020., str. 56.). Produkt treniranja algoritma na nekom skupu podataka nazivamo modelom. U ovom je projektu korištena su ukupno 34 modela.

Ovo su algoritmi za lematizaciju i POS tagiranje, navedeni onim redom kojim se pojavljuju u tablicama s rezultatima. Svi korpusi koje spomenemo bit će detaljnije objašnjeni nešto kasnije, a sada se samo navode u kontekstu treniranja pojedinog algoritma.

1. *CLTK*

Classical Language Toolkit (CLTK) je knjižnica za programski jezik *Python* koja omogućuje NLP obradu starih, klasičnih i srednjovjekovnih jezika Europe i Azije (Johnson, i dr., 2014.-2021.). Jedan od jezika s najrazvijenijom podrškom je upravo latinski, za koji *CLTK* nudi niz opcija – među ostalima to su čitanje korpusa, ujednačavanje grafema (j/i, v/u), prozodijska analiza te, za nas važno, POS tagiranje. Naš je uzorak tagiran trima dostupnim tagerima – *backoff, TnT* i *CRF* tagerom. Svaki od njih koristi drugačiju metodu određivanja POS tagova ulaznim pojavnicama. *Backoff* tager je bayesovski, *TnT* koristi skriveni Markovljev model (engl. *HMM; Hidden Markov Model*), a *CRF* koristi model naziva *Conditional Random Field*.[25] Objašnjenje za svaki od modela donose Jurafsky i Martin (2020., str. 56. i 155.-167.). Svi tageri trenirani su na korpusu *Ancient Latin Dependency Treebank v. 1.7*.[26]

2. *LEMLAT*

LEMLAT 3.0 je algoritam za lematizaciju i morfološku analizu latinskog jezika, čija se leksička baza temelji na trima rječnicima latinskog (Georges & Georges, 1913.-1918.; Glare, 1982.; Gradenwitz, 1904.) kojima su kasnije pridodani Forcellinijev *Onomasticon* (1940.) i Du Cangeov (1883.-1887.) *Glossarium Mediae et Infimae Latinitatis* (Passarotti, Ruffolo, Cecchini, Litta, & Budassi, 2018.). Kada *LEMLAT*-u kao ulaznu informaciju damo neki oblik riječi, on nam vrati lemu i morfološka obilježja i leme i oblika. Važno je znati da *LEMLAT* ne uzima u obzir kontekst tokena koji smo unijeli, već analizira samo oblik. U tom smislu kao izlaznu informaciju može dati više lema, odnosno sve one kojima oblik može pripadati. Iz tog razloga *LEMLAT* ćemo koristiti samo kao lematizator, a POS tagove ćemo promatrati jedino kako bismo razlikovali homografne leme.

[25] Vidi https://legacy.cltk.org/en/latest/latin.html#pos-tagging (pristupljeno 20. ožujka 2021).
[26] Vidi https://github.com/cltk/latin_treebank_perseus (pristupljeno 20. ožujka 2021).

3. *LaPOS*

LaPOS (engl. *Lookahead Part of Speech Tagger*) je POS tager načinjen u programskom jeziku *C++* koji koristi tzv. *lookahead* mehanizam kako bi poboljšao tzv. *„history-based“* pristup određivanju POS tagova (Tsuruoka, Miyao, & Kazama, 2011.). Ovaj je algoritam pokazao visoku efikasnost u POS tagiranju latinskog u provedenim istraživanjima (vor der Brück, Eger, & Mehler, 2015.; Eger, Gleim, & Mehler, 2016.). Za potrebe tagiranja našeg uzorka koristili smo *LaPOS* treniran na označenim korpusima *ITTB-UD-23*, *ITTB-UD-26*, *LLCT-UD-26*, *PROIEL-UD* i *Perseus-UD*, a rezultati su POS tagovi.

4. *LatMor*

LatMor predstavlja još jedan algoritam koji omogućuje lematizaciju i pruža morfosintaktički opis danog oblika, a može i generirati oblike određene riječi te prikazivati prozodijske vrijednosti vokala (Springmann, Schmid, & Najock, 2016.). Autori navode kako je leksička baza *Berlin Latin Lexicon*, koji je nastao u ranijem projektu spajanjem više rječnika i leksikona. Kao jedna od prednosti ovog algoritma navodi se širok spektar tekstova koje uspješno analizira, a posebno je naglašena efikasnost i na klasičnim i na srednjovjekovnim tekstovima. Kao i kod *LEMLAT*-a, ni *LatMor* ne uzima u obzir kontekst, već analizira oblik dajući nam sve moguće leme i morfosintaktičke opise.[27] Stoga ćemo i kod *LatMor*-a u rezultatima promatrati samo leme, a ne i POS tagove.

5. *MarMoT*

MarMoT je morfološki tager koji koristi prilagođeni CRF model za određivanje POS tagova i morfoloških obilježja tokena (Müller, Schmid, & Schütze, 2013.). Prema riječima i ispitivanjima autora, ovaj se model odlikuje značajnim povećanjem brzine tagiranja i učenja, a rezultate prikazuje u skraćenom *CoNLL09* formatu.[28] U našem slučaju, *MarMoT* je treniran na korpusima *Capitularia*, *ITTB-UD-23*, *ITTB-UD-26*, *LLCT-UD-26*, *PROIEL-UD* i *Perseus-UD*, a rezultati su POS-tagovi i leme.

6. *NLTK*

Natural Languages Toolkit (NLTK) još je jedna knjižnica za programski jezik *Python*, koja sadrži razne module, podatke i sl. namijenjene raznim zadatcima iz područja prirodne obrade jezika (Bird, Loper, & Klein, 2009.). Konceptualno sličan spominjanom *CLTK*-u, nudi razne

[27] Za primjer vidi https://www.cis.uni-muenchen.de/~schmid/tools/LatMor/ (pristupljeno 20. ožujka 2021).
[28] Za više o *CoNLL09* formatu vidi https://ufal.mff.cuni.cz/conll2009-st/task-description.html (pristupljeno 20. ožujka 2021).

načine kojima možemo pristupiti NLP problemima. Od svih mogućnosti (pristup korpusima, tokenizacija, parsiranje itd.) nama je najvažnija mogućnost POS tagiranja. Kao i kod *CLTK*-a, i ovdje smo koristili tri verzije POS tagera, koji se međusobno razlikuju u statističkim modelima koje koriste za tagiranje. Prvi je *backoff* tager, zatim *CRF* i na kraju *TnT* tager (na temelju kojih su i nastali istoimeni *CLTK*-ovi tageri). Za naše potrebe trenirani su na korpusu *ITTB-UD-23*, a i ovdje su nam izlazni podatci samo POS tagovi.

7. *RDRPOSTagger*

RDRPOSTagger (engl. *Ripple Down Rules-based Part-Of-Speech Tagger*) tager je koji omogućuje POS tagiranje i morfološku analizu koji se temelji na tzv. *ripple down rules* načinu učenja u kojem se pravila za analizu pohranjuju u obliku stabla te omogućavaju ispravljanje starih i dodavanje novih pravila (Nguyen, Nguyen, Pham, & Pham, 2015.). Autori tvrde kako su glavne odlike ovog algoritma jednostavnost korištenja, točnost i brzina. Za naš je korpus korišten *RDRPOSTagger* treniran na korpusima *ITTB-UD*, *ITTB-UD-23*, *ITTB-UD-26*, *LLCT-UD-26* i *PROIEL-UD*, a korištena je i opcija tagiranja pomoću unaprijed treniranog modela[29] (engl. *pre-trained*) za latinski.

8. *TreeTagger*

TreeTagger jedan je od najstarijih tagera korištenih u ovom radu i često se upotrebljava prilikom mjerenja efikasnosti novijih modela, a koristi metodu stabla odlučivanja (engl. *decision tree*) kako bi ispravno odredio POS tag i lemu (Schmid, 1994.). U svrhu POS tagiranja i lematizacije našeg korpusa korišteni su modeli trenirani na korpusima *OMNIA*, *ITTB-UD-26*, *LLCT-UD-26*, *PROIEL-UD* i *Perseus-UD* te model koji je unaprijed treniran[30] za označavanje latinskog jezika.

9. *UDPipe*

UDPipe 2.0 je tzv. *pipeline* (lanac zadataka koji se obavljaju jedan za drugim) za tokenizaciju rečenica i riječi, POS tagiranje, lematizaciju te ovisnosno parsiranje. (Straka, 2018.). Ovaj algoritam za učenje koristi neuronske mreže, a usko je povezan s projektom *Universal Dependencies*[31] te je u nekoliko natjecanja NLP alatā ostvario izvrsne rezultate[32]. *UDPipe* za

[29] Unaprijed trenirani modeli također su trenirani na UD korpusima, premda autoru ovog rada nije jasan točan postupak tog treninga, a za više vidi https://github.com/datquocnguyen/RDRPOSTagger/tree/master/Models (pristupljeno 20. ožujka 2021).
[30] O unaprijed dostupnom modelu vidi https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Latin-parameter-file-readme (pristupljeno 20. ožujka 2021).
[31] Vidi https://universaldependencies.org/ (pristupljeno 20. ožujka 2021).
[32] Vidi https://ufal.mff.cuni.cz/udpipe/2 (pristupljeno 20. ožujka 2021).

treniranje zahtijeva tekstove u *CoNLL-U* formatu, a za naše je potrebe treniran na korpusima *ITTB-UD-23*, *ITTB-UD-26*, *LLCT-UD-26*, *PROIEL-UD* i *Perseus-UD* te smo kao rezultate dobili POS tagove i leme.

Kompletan proces tagiranja i lematizacije, na temelju dostavljenih podataka, proveo je dr. sc. Flavio Massimiliano Cecchini, znanstveni suradnik pri sveučilištu *Università Cattolica del Sacro Cuore*, koji je gotove rezultate elektronskim putem poslao autoru ovog rada.

U već spominjanom diplomskom radu, Gamba (2020., str. 15.) je koristila *UDPipe*, *TreeTagger*, *MarMoT*, *CLTK* te *Collatinus*, koji u našem radu nije korišten. Osim toga, Gamba je *MarMoT* trenirala samo na korpusu *Capitularia* i nije koristila *TreeTaggerov* unaprijed dostupan model za latinski.

4.3.2. Korpusi korišteni za treniranje lematizatora i tagera

Objasnili smo kako je većinu korištenih lematizatora i tagera nužno prije korištenja trenirati na određenom označenom korpusu. Ukratko ćemo predstaviti svaki od navedenih korpusa[33] i opisati njihove osobine. Jedna od ključnih osobina korpusa je i skup tagova (engl. *tagset*) koji je korišten za njegovo označavanje. U ovom će poglavlju biti samo navedeno koji je skup oznaka korišten za pojedini korpus, a skupove ćemo detaljnije objasniti u idućem poglavlju, kako bismo ostvarili bolju preglednost rada. Ako zanemarimo njihove različite verzije, za treniranje naših tagera i lematizatora korišteno je ukupno 6 korpusa.

ITTB-UD, *ITTB-UD-23* i *ITTB-UD-26*[34] tri su verzije istog korpusa, odnosno korpusa *Index Thomisticus Treebank (ITTB)*, koji sadrži tekstove Tome Akvinskog i s njime povezanih autora, pisanih srednjovjekovnim latinskim. Taj je korpus preveden iz originalnog načina anotacije u način sukladan principima projekta *Universal Dependencies* (Cecchini, Passarotti, Marongiu, & Zeman, 2019.), što znači da se u trenutnoj verziji koristi tzv. *UD (Universal Dependencies) tagsetom*[35], o kojem će više riječi biti kasnije.

LLCT-UD-26[36] je najnovija verzija korpusa *Late Latin Charter Treebanks (LLCT)*, koji sadrži 521 pravni dokument (lat. *cartulae*) na ranosrednjovjekovnom latinskom s područja Italije (Cecchini, Korkiakangas, & Passarotti, 2020.). Tekstovi se po svojim jezičnim obilježjima

[33] Većina navedenih korpusa u stvari su banke stabala (engl. *treebank*), odnosno „parsirani tekstni korpusi koji imaju označenu sintaktičku i semantičku strukturu rečenica“ (http://ihjj.hr/mreznik/page/pojmovnik/6/). Međutim, kako bi se izbjegla konfuzija, nazivat ćemo ih jednostavno korpusima, budući da nam glavna odrednica banke stabala, sintaktičko parsiranje, i onako trenutno nije važno.
[34] Vidi https://universaldependencies.org/treebanks/la_ittb/ (pristupljeno 27. ožujka 2021).
[35] Vidi https://universaldependencies.org/u/pos/ (pristupljeno 27. ožujka 2021).
[36] Vidi https://universaldependencies.org/treebanks/la_llct/ (pristupljeno 27. ožujka 2021).

razlikuju od klasičnog, ali i od tipično srednjovjekovnog latinskog. Kao i *ITTB*, i *LLCT* je doživio pretvorbu u *UD tagset*, a UD verzija korištena je za treniranje modela.

PROIEL-UD[37] je korpus latinskih tekstova preuzetih iz banke stabala starih indoeuropskih jezika, koja je nastala u sklopu projekta *Pragmatic Resources in Old Indo-European Languages (PROIEL)*, koji se bavio paralelnim istraživanjem prijevoda Biblije na nekoliko indoeuropskih jezika (Haug & Jøhndal, 2008.). Latinski dio korpusa sadrži većinu *Vulgate*, izbor iz Cezarovih *Galskih ratova*, Ciceronovih *Pisma Atiku* i Paladijevog *Poljodjelstva* te prvu knjigu Ciceronovog *O dužnostima*. I kod ovog je korpusa za treniranje korištena verzija prevedena u *UD tagset*.

PerseusUD[38], odnosno *Universal Dependencies Latin Treebank*, korpus je sastavljen od odabranih latinskih dijelova korpusa *Ancient Greek and Latin Dependency Treebank 2.1.* Ovaj korpus sadrži dijelove Augustovih *Res Gestae,* Ciceronova *Protiv Katiline*, Jeronimove *Vulgate*, Vergilijeve *Eneide*, Ovidijevih *Metamorfoza*, Petronijeva *Satirikona*, Fedrovih *Ezopskih basni*, Propercijevih *Elegija*, Salustijeva *Rata s Katilinom*, Svetonijevih *Života careva* i Tacitove *Povijesti*. I ovaj je korpus nastao pretvorbom ranije verzije u *UD tagset*.

Maločas spomenuti *Ancient Greek and Latin Dependency Treebank (AGLDT)*, u dvije je verzije korišten za treniranje naših modela – u verziji 1.7. za *CLTK*, a u verziji 2.1. za ostale, u obliku korpusa *PerseusUD*. Ključna je razlika između ove dvije verzije istog korpusa u načinu tagiranja. Kod verzije korištene za *CLTK*, ne koristi se *UD tagset*, već zaseban skup tagova.[39]

Capitularia su korpus sastavljen od anotiranih srednjovjekovnih merovinških i karolinških isprava na latinskom jeziku. Dio je to projekta *Computational Historical Semantics*[40], a za nas je važno da ni ovaj korpus ne koristi *UD tagset*, već zaseban skup tagova *CHSTS*.

U sklopu projekta *OMNIA* (fr. *Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins*) koji se bavio razvijanjem digitalnih alata za srednjovjekovni latinski (Bon, 2011.), nastao je anotirani korpus korišten za treniranje *TreeTaggera* kako bi se dobio model za POS tagiranje i lematizaciju tekstova na srednjovjekovnom latinskom. Dobiveni model iskorišten je na našem uzorku, a i kod ovog je korpusa važna činjenica da koristi zaseban skup tagova, različit od do sad spominjanih.[41]

[37] Vidi https://universaldependencies.org/treebanks/la_proiel/ (pristupljeno 27. ožujka 2021).
[38] Vidi https://universaldependencies.org/treebanks/la_perseus/ (pristupljeno 27. ožujka 2021).
[39] Vidi https://github.com/PerseusDL/treebank_data/tree/master/v1/latin (pristupljeno 27. ožujka 2021).
[40] Vidi https://www.comphistsem.org/ (pristupljeno 27. ožujka 2021).
[41] Za popis korištenih tagova, listu lema i pregled anotiranih korpusa vidi https://glossaria.eu/outils/lemmatisation/ (pristupljeno 27. ožujka 2021)

Na kraju treba još zasebno spomenuti i tekstove korištenje za dobivanje *TreeTaggerova* unaprijed dostupnog modela za latinski.[42] Premda se radi o dijelovima već spomenutih korpusa (*PROIEL*, *Perseus* i *ITTB*), za njihovo je označavanje korišten drugačiji skup tagova[43] pa ih je važno izdvojiti.

U svom radu Gamba koristi uglavnom iste korpuse za treniranje modela, a jedini koji ne upotrebljava je *OMNIA*.

4.3.3. Skupovi tagova korišteni u automatskom tagiranju

Kao što se može uočiti iz njihova opisa, korpusi korišteni za treniranje tagera označeni su s ukupno 5 različitih skupova tagova. Informacije o skupovima tagova vrlo su važne, pogotovo kad dođemo do pitanja usporedbe različitih modela za tagiranje, jer ne postoji savršeno preklapanje između sadržaja različitih skupova.

Prvi i u našem radu najkorišteniji skup tagova s kojim se susrećemo je više puta spominjani *UD tagset*[44]. Radi se o skupu tagova koji se koristi za označavanje banaka stabala u sklopu projekta *Universal Dependencies*[45] (Nivre, i dr., 2018.), a temelji se na kombinaciji nekoliko ranijih skupova tagova (Zeman, 2018., str. 32.). Projekt *Universal Dependencies* smišljen je s ciljem provođenja paralelnih istraživanja na više različitih jezika na konzistentan način, a u skladu s time i *UD tagset* zamišljen je kako bi omogućio označavanje različitih jezika istim skupom tagova. Ukupno je 17 različitih tagova u *UD tagsetu*, koje ćemo sada samo kratko navesti i u pojednostavljenom obliku opisati što označuju, a podrobnije ćemo objasniti kako se koriste u potpoglavlju 4.4.1. Radi veće preglednost tagove prikazujemo u tablici.

POS tag	Što označuje	POS tag	Što označuje
/ADJ	pridjev	/PART	čestica
/ADP	prijedlog	/PRON	zamjenica
/ADV	prilog	/PROPN	vlastita imenica
/AUX	pomoćni glagol	/PUNCT	interpunkcija
/CCONJ	usporedni veznik	/SCONJ	zavisni veznik
/DET	determinator	/SYM	simbol
/INTJ	usklik	/VERB	glagol
/NOUN	imenica	/X	ostalo
/NUM	broj		

Tablica 2. Pojednostavljeni kratki opis UD tagseta

[42] Vidi fusnotu 30.
[43] Vidi https://www.cis.lmu.de/~schmid/tools/TreeTagger/data/Lamap-Tagset (pristupljeno 27. ožujka 2021).
[44] Vidi fusnotu 35.
[45] Vidi fusnotu 31.

Važno je napomenuti da spomenuti UD korpusi ne koriste sve naveden tagove, niti međusobno jednake podskupe tagova. Pogledom u njihovu usporednu dokumentaciju[46] ustanovit ćemo da ITTB i LLCT koriste 15 tagova (ne koriste /INTJ i /SYM), PROIEL koristi 14 tagova (ne koristi /PART, /SYM i /PUNCT), a Perseus 12 tagova (ne koristi /PROPN, /DET, /AUX, /PART i /SYM).

Skup POS tagova korišten u *Ancient Greek and Latin Dependency Treebanku*, kojim se služe *CLTK*-ovi tageri, zapravo je dio „šifre" kojim se u spomenutom korpusu označuju morfološka obilježja riječi (ukupno 9 obilježja). Za označavanje našeg korpusa korišteno je ukupno 12 tagova, koji su popisani i ukratko opisani u tablici.

POS tag	Što označuje	POS tag	Što označuje
/A	pridjev	/P	zamjenica
/C	veznik	/R	prijedlog
/D	prilog	/T	particip
/E	usklik	/U	interpunkcija
/M	broj	/V	glagol
/N	imenica	/X	ostalo

Tablica 3. Krati opis skupa tagova korištenog u CLTK-u, tj. AGLDT-u

Treba napomenuti da se ovaj sustav tagova koristio u ranijim verzijama *AGLDT*-a, na kojima je *CLTK* treniran, a da novije verzije koriste nešto drugačiji sustav. Prepreku pri točnijem opisu načina na koji se tagovi koriste predstavlja manjak jasne dokumentacije o pravilima POS tagiranja korpusa. Primjerice, dostupni popis POS tagova[47] predviđa i tag /I, sa značenjem *„exclamation"*, premda u našem korpusu ta oznaka nije dodijeljena niti jednom tokenu. Jedan od važnijih detalja je i korištenje zasebnog POS taga za participe. Za oblik koji ne poznaju *CLTK* tageri umjesto taga vraćaju rezultat /Unk.

Srednjovjekovni korpus *Capitularia* koristi skup oznaka *CHSTS*, tj. *Computational Historical Semantics TagSet* (Gleim, i dr., 2019., str. 14.)[48]. Pri označavanju našeg korpusa pojavljuje se 17 tagova iz ovog skupa, koje ćemo prikazati u tablici zajedno s njihovim jednostavnim opisom.

[46] Vidi https://universaldependencies.org/treebanks/la-comparison.html#morphology (pristupljeno 27. ožujka 2021).
[47] Vidi fusnotu 39.
[48] U navedenom radu u popisu tagova nedostaje tag /ORD, pa za bolji i pregledniji pregled tagova vidi vor der Brück, Eger i Mehler (2015., str. 107.).

POS tag	Što označuje	POS tag	Što označuje
/ADJ	pridjev	/NP	antroponim
/ADV	prilog	/NUM	glavni broj
/AP	prijedlog	/ORD	redni broj
/CON	veznik	/PRO	zamjenica
/DIST	dijelni broj	/PTC	čestica
/FM	strani materijal	/V	glagol
/ITJ	usklik	/XY	ostalo
/NE	vlastita imenica	/$	interpunkcija
/NN	imenica		

Tablica 4. Kratak opis skupa tagova CHSTS korištenog u korpusu Capitularia.

Zanimljivo je da ovaj skup tagova razlikuje vlastite imenice i antroponime, te da ima mogućnost označavanja tri različite vrste brojeva.

Projekt *OMNIA* koristi 12 tagova za označavanje tokena. Premda je na web stranicama projekta dostupan popis tagova[49], nisu navedena njihova objašnjenja, a nisu uspjeli ni pokušaji pronalaska dodatne literature o ovom skupu tagova. Popis se nalazi u tablici, uz objašnjenja koja se temelje na pregledu načina uporabe pojedinih tagova u originalnom korpusu za tagiranje.

POS tag	Što označuje	POS tag	Što označuje
/ADV	prilog	/PRE	prijedlog
/CON	veznik	/PRO	zamjenica
/INT	usklik	/QLF	pridjev
/NAM	vlastito ime	/SUB	imenica
/NUM	broj	/VBE	glagol
/PON	interpunkcija (zarez)	/SENT	interpunkcija (točka)

Tablica 5. Kratki opis tagova korištenih u projektu OMNIA

Sudeći prema imenima tagova, vjerojatno se radi o originalno francuskom skupu tagova (što je logično i s obzirom na to da se radi o francuskom projektu), a zanimljivo je primijetiti da se tagovi za zarez i točku razlikuju.

Posljednji skup tagova je tzv. *Lamap*[50], koji je korišten pri treniranju *TreeTaggerova* unaprijed dostupnog modela za latinski, kojeg je priredila Gabriele Brandolini. Sadrži ukupno 22 taga koja su prikazana u tablici.

[49] Vidi https://www.glossaria.eu/sources/treetagger/classes.txt (pristupljeno 27. ožujka 2021).
[50] Vidi https://www.cis.lmu.de/~schmid/tools/TreeTagger/data/Lamap-Tagset (pristupljeno 27. ožujka 2021).

POS tag	Što označuje	POS tag	Što označuje
/ABBR	kratica	/INT	usklik
/ADJ	pridjev (i redni broj)	/N	imenica
/ADV	prilog	/NPR	vlastita imenica
/CC	usporedni veznik	/POSS	posvojna zamjenica
/CLI	enklitika	/PREP	prijedlog
/CS	zavisni veznik	/PRON	(osobna) zamjenica
/DIMOS	pokazna zamjenica	/PUN	interpunkcija
/ESSE	glagol biti	/REL	odnosna zamjenica
/EXCL	usklik	/SENT	interpunkcija
/FW	strani izraz	/SYM	simbol
/INDEF	neodređena zamjenica	/V	glagol

Tablica 6. Kratki opis TreeTaggerova skupa tagova Lamap

Ono što se najprije ističe je brojnost tagova u ovom skupu. Velika se granularnost primjećuje kod zamjenica, koje su, za razliku od ostalih skupova, razvrstane u zasebne tagove po vrsti.

Kao što možemo vidjeti, skupovi tagova međusobno se razlikuju po veličini i sadržaju, što uvelike otežava njihovo uspoređivanje. Za više o tome vidi potpoglavlje 4.4.3.

Gamba se u svojem radu susreće s ukupno 4 skupa tagova, od koja su 3 ista kao naša. Naime, autorica koristi i tager *Collatinus*. koji upotrebljava zaseban skup tagova.

4.3.4. Rezultati tagiranja i lematizacije

Rezultati tokenizacije i/ili lematizacije pojedinih modela najprije su spremani u odvojene datoteke, a kasnije su spojeni u zajedničku tablicu, koju smo u projektu nazvali *Tabula Synoptica*. Rezultati za svaki tekst spremljeni su u zasebnu *Tabula Synoptica*. Sve se tablice nalaze u CSV formatu. Ukupni broj tokena, u svim tekstovima, iznosi 58 185 (uključujući i interpunkcijske znakove), a svakom je tokenu pridruženo po 36 POS tagova i/ili lema.

Primjer prikaza jedne rečenice iz *Tabula Synoptica* Andreisovih *Epistolae* nalazi se u Tablici 7 (dolje). Ovdje možemo na primjeru pokazati ono što smo do sad objasnili o procesu lematizacije i tagiranja. U prvom su stupcu nazivi pojedinih algoritama/modela korištenih za obradu tokena, a u prvom se retku nalaze tokeni iz teksta. Svaki algoritam/model daje nam rezultat za koji pretpostavlja da je točan, a koji može biti u obliku leme, u obliku POS taga ili u obliku i jednog i drugog. Kod modela koji nam kao rezultat daju i lemu i POS tag, zapis ima oblik „lema/POStag“.

Pogledamo li rezultate *LEMLAT*-a i *LatMor*-a, primijetit ćemo da nam kao rezultate daju sve moguće leme za dani oblik i da *LEMLAT* za token „.“ (tj. točku) ostavlja prazno mjesto, a *LatMor* znak „/“. Budući da su *LEMLAT* i *LatMor* samo lematizatori koji primarno lematiziraju na temelju svoje leksičke baze (a POS tagove koriste samo za razlučivanje homografnih lema),

ne uzimaju u obzir kontekst u kojem se token nalazi, a kada se susretnu s tokenom koji ne mogu spojiti s lemom iz baze, jednostavno ga ne označe. Promotrimo li pažljivo POS tagove, vidjet ćemo da se tagovi koje su kao rezultat ponudili modeli *CLTK* (sva 3), *MarMoT-Capitularia*, *OMNIA* i *TreeTagger-TreeTagger-latin* razlikuju i međusobno i od tagova ostalih modela, koji svi redom koriste *UD tagset*. Razlog tomu je upravo činjenica da su trenirani na korpusima označenim različitim skupovima tagova (vidi 4.3.3.).

Treba napomenuti i da *TreeTagger*, zbog načina odlučivanja, može ponuditi više lema u jednom polju, jer ponekad predlaže sve „vjerojatne“ leme. Još jedan detalj oko lematizacije tiče se modela *MarMoT-Capitularia*, koji je, iz za sad nejasnih razloga, trima pojavnicama dodijelio sasvim krive leme, u krivom formatu zapisa. Radi se o pojavnicama *Cererem* (kod Kunića), *Par* (kod Marulića) i *rempublicam* (kod Crijevića). Njihove leme nisu zapisane u skladu s lematizacijskom shemom koju *MarMoT* inače koristi.

TOKEN	**Vale**	**felix**	**et**	**redi**	**.**
CLTK-CRF	/D	/A	/C	/V	/U
CLTK-TnT	/Unk	/A	/C	/Unk	/U
CLTK-backoff	/Unk	/A	/C	/Unk	/U
LEMLAT	ualus/N\ualeo/V	filix/N\felix/A\felix/N	et/X	redius/A\redeo/V	
LaPOS-ITTBUD23	/ADJ	/ADJ	/CCONJ	/VERB	/PUNCT
LaPOS-ITTBUD26	/ADJ	/ADJ	/CCONJ	/VERB	/PUNCT
LaPOS-LLCTUD26	/PROPN	/NOUN	/CCONJ	/VERB	/PUNCT
LaPOS-PROIELUD	/PROPN	/NOUN	/CCONJ	/VERB	/ADV
LaPOS-PerseusUD	/NOUN	/ADJ	/CCONJ	/VERB	/PUNCT
LatMor	valere/V\Valis/PN	felix/ADJ\felix/N	et/CONJ	redire/V\reire/V	/
MarMoT-Capitularia	valeo/V	felix/ADJ	et/CON	redo/V	./$.
MarMoT-ITTB-UD23	Valis/ADJ	felix/NOUN	et/CCONJ	redo/VERB	./PUNCT
MarMoT-ITTB-UD26	Valis/NOUN	felix/ADJ	et/CCONJ	redo/VERB	./PUNCT
MarMoT-LLCT-UD26	Valis/PROPN	felix/VERB	et/CCONJ	ro/VERB	./PUNCT
MarMoT-PERSEUSUD	Valis/NOUN	felix/ADJ	et/CCONJ	ro/VERB	./PUNCT
MarMoT-PROIELUD	uale/ADJ	felix/NOUN	et/CCONJ	redior/VERB	./ADV
NLTK-123	/Unknown	/ADJ	/CCONJ	/Unknown	/PUNCT
NLTK-CRF	/ADJ	/NOUN	/CCONJ	/VERB	/PUNCT
NLTK-TnT	/Unk	/ADJ	/CCONJ	/Unk	/PUNCT
OMNIA	<unknown>/SUB	felix/QLF	et/CON	redeo/VBE	./SENT
RDRPOSTagger-ITTB	/ADJ	/ADJ	/CCONJ	/VERB	/PUNCT
RDRPOSTagger-ITTB23	/ADJ	/ADJ	/CCONJ	/VERB	/PUNCT
RDRPOSTagger-ITTB26	/ADJ	/ADJ	/CCONJ	/VERB	/PUNCT
RDRPOSTagger-LLCT26	/NOUN	/VERB	/CCONJ	/PROPN	/PUNCT
RDRPOSTagger-PROIEL	/VERB	/NOUN	/CCONJ	/VERB	/PUNCT
RDRPOSTagger	/ADJ	/VERB	/CCONJ	/VERB	/PUNCT
TreeTagger-TreeTagger-latin	valeo/V	felix/ADJ	et/CC	redeo/V	./SENT
ITTB-UD26	<unknown>/ADJ	felix/ADJ	et/CCONJ	<unknown>/VERB	./PUNCT
LLCT-UD26	<unknown>/PROPN	felix/PROPN	et/CCONJ	<unknown>/PROPN	./PUNCT
PERSEUSUD	<unknown>/NOUN	felix/ADJ	et,etiam/CCONJ	<unknown>/VERB	./PUNCT
PROIELUD	<unknown>/ADJ	felix/PROPN	et/CCONJ	redeo/VERB	./PUNCT
UDpipe-ITTB23	Valis/NOUN	felix/ADJ	et/CCONJ	redior/VERB	./PUNCT
UDpipe-ITTB26	Valis/ADJ	felix/ADJ	et/CCONJ	redior/VERB	./PUNCT
UDpipe-LLCT	Valis/PROPN	felix/NOUN	et/CCONJ	ro/VERB	./PUNCT
UDpipe-PERSEUS	Valus/ADV	felix/ADJ	et/CCONJ	ro/VERB	./PUNCT
UDpipe-PROIEL	valus/ADJ	felix/PROPN	et/CCONJ	redeo/VERB	./PUNCT

Tablica 7. Primjer prikaza rezultata za rečenicu "Vale felix et redi." iz Tabula Synoptica Andreisovih Epistolae

4.4. Evaluacija rezultata i odabir optimalnog tagera i lematizatora

Cilj obrade uzorka navedenim tagerima i lematizatorima je odabir onoga koji je najtočnije označio naš uzorak. Isti taj alat trebao bi nam omogućiti i najtočniju moguću lematizaciju i tagiranje čitave *CroALa*-e. Ovaj stadij projekta još uvijek je u tijeku, a u radu ćemo prikazati rezultate koji su do sad dobiveni i obrađeni.

Da bismo otkrili koji od navedenih modela najbolje obavlja posao, moramo procijeniti koliko je koji bio uspješan u dodjeljivanju lema i POS tagova te usporediti njihove rezultate. Da bismo to napravili najprije moramo izgraditi tzv. „zlatni standard“, s kojim ćemo uspoređivati rezultate i tako mjeriti uspješnost svakog modela. Budući da se korišteni modeli međusobno razlikuju u standardima koje koriste, zbog razloga objašnjenih u prethodnom poglavlju, bit će potrebno pronaći i način da pri uspoređivanju te razlike premostimo, kako bismo osigurali čim objektivniju procjenu.

4.4.1. Zlatni standard

Zlatni standard (engl. *gold standard*) je termin kojim se u NLP-u nazivaju ručno označeni korpusi koji služe za procjenu algoritama i modela za automatsku anotaciju teksta (Wissler, Almashraee, Dagmar, & Paschke, 2014.). U našem projektu zlatni standard označuje dio uzorka koji će biti ručno tagiran i lematiziran. Ručno tagiranje može se obaviti sasvim „od nule“, odnosno ručnim dodjeljivanjem taga i leme svakoj pojavnici, ili se mogu odabrati rezultati jednog od iskorištenih modela koji se onda ručno ispravljaju kako bi se dobio željeni ispravan skup podataka. U našem je slučaju odabrana prva strategija, odnosno potpuna ručna anotacija uzorka. Dio uzorka koji ćemo tagirati treba sačinjavati oko 10% pojavnica ukupnog uzorka (tj. oko 5800 pojavnica) te biti reprezentativan za naš uzorak. Kako bi se to postiglo, nasumično se odabiru rečenice koje će se analizirati.

Za potrebe ovog rada, preliminarno je označeno 612 pojavnica, nasumično odabranih iz tekstova proporcionalno njihovoj veličini. Premda taj broj pojavnica nije dovoljan za procjenu rada uzorka, dobro je poslužio za analizu problema koji nastaju pri evaluaciji.

Kako navodi Gamba, koja također odabire anotaciju od nule, gradnja zlatnog standarda nije uopće trivijalan zadatak, kakvim bi se na prvu mogao činiti. Naime, moguće je odabrati više pristupa i načina tagiranja i lematizacije (što je vidljivo i iz 5 različitih skupova tagova s kojima smo se susreli), a svaki od pristupa podrazumijeva donošenje određenih „odluka“, odnosno prihvaćanje određenih principa prilikom analize i anotacije teksta. Neki od problema s kojima smo se susreli bit će opisani u zasebnom poglavlju, ali za ilustraciju možemo razmotriti pitanje

priloga *fuse* (obilno, opširno). Možemo ga tagirati kao prilog i pridružiti mu lemu *fuse*, što je uobičajna praksa s prilozima. Međutim, taj je prilog nastao od participa glagola *fundo, 3., fusi, fusum* (a i rječnik *Lewis & Short* ga navodi pod natuknicom tog glagola), pa bi mu se mogla dodijeliti i lema *fundo*. I jedna i druga opcija su na svoj način valjane, ali je važno kod svih ovakvih primjera uočiti da je došlo do određenog odabira, odnosno odluke, i moramo moći obrazložiti zašto neke opcije smatramo valjanima ili ih preferiramo, a neke odbacujemo. Osim lingvističkih i gramatičkih argumenata za odluku, na umu treba imati i to da će naše odluke utjecati i na procjenu uspješnosti pojedinog tagera ili lematizatora. Primjerice, odlučimo li lematizirati prilog *fuse* vlastitom lemom (tj. lemom *fuse*), oni lematizatori koji joj pridruže lemu *fundo* „pogriješit će“ u odnosu na naš zlatni standard. Drugim riječima odluke bi trebalo temeljiti i na principima za koje želimo da se primjenjuju na našem korpusu.

Premda bismo teoretski mogli izgraditi zasebni zlatni standard za svaki algoritam ili model, pazeći pritom na strategiju označavanja koja je korištena u svakome od njih, to bi bilo izrazito vremenski zahtjevno, a ne bi uvelike doprinijelo točnosti prosudbe.

Za POS tagiranje zlatnog standarda koristi se *UD tagset*¸ kao što to čini i Gamba, budući da se taj skup tagova koristi u *LiLa*-i. Taj skup tagova ukratko je naveden u Tablica 2, a ovdje ćemo detaljnije objasniti predviđenu uporabu tagova i prikazati kako je svaki tag korišten u našem radu. Uz to, spomenut ćemo i neke opće principe lematizacije koji su prihvaćeni tijekom označavanja. U ovom će popisu biti opisani samo oni slučajevi koji se pojavljuju u smanjenom dijelu uzorka od 612 pojavnica. Naravno, to znači da bi se pregledavanjem punog uzorka mogao naći još čitav niz problema i primjera relevantnih za ovaj popis, ali bismo tada vremenski i sadržajno izašli van predviđenog opsega ovog rada. To također znači i da se u budućnosti načini korištenja pojedinih tagova ili principi lematizacije mogu mijenjati, sukladno spoznajama koje nastanu kao rezultat obrade većeg broja slučajeva.

Paralelan je popis korištenih tagova te opis lematizacije moguće pratiti u radu Federice Gamba (Gamba, 2020., str. 27.-35.). Treba imati na umu da Gamba u svom radu pojavnicama dodjeljuje i tzv. univerzalna svojstva (eng. *universal features*)[51], što u našem radu nije slučaj.

Nekoliko je faktora utjecalo na principe tagiranja i lematizacije. U prvom redu ti principi ovise o znanju označivača (odnosno autora ovog rada) i njegovom shvaćanju latinskog jezika, koji se uvelike temelje na tzv. gramatici Gortan-Gorski-Pauš (GGP) (2005.), a ista je gramatika bila i prvi odabir za konzultiranje u nejasnim situacijama. Uz navedenu gramatiku, često su

[51] Vidi https://universaldependencies.org/u/feat/ (pristupljeno 17. travnja 2021.).

konzultirane i prakse korištenja *UD tagseta* pri označavanju četiri latinska korpusa (*ITTB*, *LLCT*, *PROIEL* i *Perseus*) te načini na koje je Gamba koristila pojedine tagove. Nadalje, u obzir je uzet način na koji su određene riječi i oblici označeni u rječniku *Lewis & Short* (1879.), a povremeno je konzultiran i Žepićev rječnik (2000.).[52] Ako ništa od navedenog ne bi dalo jasan ili zadovoljavajući odgovor, odgovori su potraženi u raznim drugim gramatikama latinskog jezika ili u znanstvenoj literaturi.

Prilikom lematizacije glavni je princip bio za lemu uzeti oblik koji služi kao rječnička natuknica u rječniku *Lewis & Short*, a u slučaju riječi koje pripadaju kasnijem latinitetu konzultirane su natuknice iz rječnika *Du Cange* (1883.-1887.). Ako rječnička natuknica vodi nekoj drugoj, kao lema se uzima oblik krajnje natuknice. Sukladno tome, lema za imenske riječi uzima oblik nominativa jednine[53] (i to muškog roda za zamjenice i pridjeve), a kod glagola 1. lice jednine indikativa prezenta aktivnog[54]. Lema se zapisuje standardnom grafijom, ne koristeći grafem *j* i razlikujući grafeme *v* i *u*, a u ostalim pitanjima grafije (npr. verzije riječi s *ch* i *c*) prati se *Lewis & Short*. Ovakav bi princip trebao olakšati dosljednu lematizaciju uzorka.

/ADJ je tag koji se koristi za pridjeve, odnosno riječi koje uglavnom modificiraju imenice i pobliže određuju njihova svojstva ili osobine. Kao pridjevi se zbog morfosintaktičkih osobina označuju i redni brojevi (npr. *primus*). Supstantivirani pridjevi tretiraju se kao imenice jer se često značenjem i sintaktički tako ponašaju. Primjerice *cardinalis* u značenju „kardinal“ se označuje kao imenica, a ne kao pridjev. Neki pridjevi, poput *meus*, *tuus* ili *suus* se ne označuju kao pridjevi jer se u *UD*-u smatraju determinatorima[55]. Svi se pridjevi lematiziraju u obliku pozitiva, čak i oni sa supletivnim komparativom i superlativom (npr. *plurimus* se lematizira kao *multus*).

/ADP je tag kojim se označuju adpozicije, skupina riječi koja uključuje prepozicije i postpozicije. U našem smanjenom dijelu uzorka kao /ADP su označeni prijedlozi koji su svi redom prepozicije. Kao postpozicije bismo mogli eventualno zamisliti prijedložne ablative *gratiā* i *causā*, a Gamba spominje i oblike zamjenica poput *mēcum*, koje objašnjava kao anastrofu prepozicija. Ipak, nijedan se od ovih slučajeva nije pojavio u analiziranim rečenicama.

/ADV je tag koji označuje priloge, odnosno riječi koje pobliže opisuju glagole u kategorijama vremena, mjesta, načina, smjera i sl., a mogu modificirati i pridjeve ili druge priloge. Svi se

[52] Za oba su rječnika uglavnom korištena mrežna izdanja dostupna na logeion.uchicago.edu i solr.ffzg.hr/logeion.
[53] Osim kod *pluralia tantum*.
[54] Osim ako taj oblik ne postoji.
[55] O determinatorima vidi niže, na str. 32.

prilozi tagiraju kao prilozi, bez obzira jesu li izvedeni od neke druge riječi ili nisu, ali se lematiziraju prema porijeklu. Primjerice, prilozi *fuse* i *anxie* tagirani su kao prilozi, ali su im dodane leme *fundo* i *anxius*, prema nastanku. Premda u analiziranim rečenicama nalazimo samo pozitive priloga, ostali bi se stupnjevi također lematizirali na oblik pozitiva. Tagom /ADV označena je i pojavnica koja je u stvari kratica *etc.,* budući da je na više mjesta tako označena u korpusu *ITTB*.

/AUX se koristi za riječi koje prate neke oblike glagola i izražavaju gramatička svojstva koja taj oblik nema. To su često glagoli (pa tako i u latinskom i hrvatskom), ali u nekim jezicima to mogu biti i druge riječi. U našem slučaju ovaj tag se koristi samo za oblike glagole *esse*, jer oni najčešće služe ili kao pomoćni glagol pri stvaranju određenog vremena ili glagolskog oblika (npr. pasivni oblici perfekta, pluskvamperfekta i futura II. ili perifrastične konjugacije) ili kao kopula u imenskom predikatu. S obzirom na jednu i na drugu funkciju, u slučaju da se pojave, ovdje bi se mogli ubrojiti i neki drugi glagoli, poput glagola *eo* (kod infinitiva futura pasivnog) ili *fio* (u ulozi kopule).

/CCONJ je tag kojim se označuju usporedni, tj. koordinirani veznici, kojima se povezuju riječi ili nezavisne rečenice. U slučaju veznika *ac*, treba naglasiti da se lematizira kao *atque*, prema *Lewis & Shortu*.

/DET je oznaka za tzv. determinatore, odnosno riječi koje pobliže označuju imenicu ili imensku skupinu (engl. *noun phrase*) i izražavaju na što se imenica, tj. imenska skupina referira u kontekstu. Stvoriti jasna pravila za dodjeljivanje ovog taga komplicirano je iz razloga što ova vrsta riječi u tradicionalnim gramatikama većine jezika ne postoji, a, koliko je autoru ovog teksta poznato, nijedna gramatika latinskog jezika ne bavi se jasno pitanjem determinatora. Iz tog razloga su načela za dodjeljivanje ovog taga modelirana prema praksi četiriju UD korpusa latinskog (premda ni ondje ne postoje sasvim dosljedna pravila za ovu vrstu riječi) i prema onome što je odlučila Gamba. Ova vrsta riječi pokriva dio onoga što tradicionalne gramatike smatraju zamjenicama ili pridjevima. Tako se u našem radu kao determinatori označuju posvojne zamjenice (*meus*, *tuus*, *suus*, *noster*), pokazne zamjenice (*hic, iste, ille, is, ipse, idem*), neodređene zamjenice (*quisque*), zamjenički pridjevi (*alius, solus, ullus, nullus, uterque*) i neki pridjevi koji određuju količinu poput *omnis*.

/INTJ je tag koji se koristi za usklike, odnosno riječi koje se koriste kao uzvik ili dio uzvika, obično izražavaju neku emociju i nisu sintaktički povezani s ostatkom rečenice. U ovom malenom uzorku dodijeljena je jedino uskliku *o*, koji prethodi vokativu.

/NOUN se koristi za označavanje imenica, odnosno riječi koje obično označavaju bića, stvari, pojave, mjesta, ideje itd. Treba napomenuti da vlastite imenice imaju zaseban tag, /PROPN.

/NUM je tag koji označava brojeve. U našem je slučaju ovaj broj dodijeljen samo glavnim brojevima. Problem prilikom lematizacije stvara različito pisanje brojeva, odnosno mogućnost da se pišu slovima te arapskim ili rimskim brojkama. Radi jednostavnosti odlučeno je da će se lematizirati čim sličnije moguće tokenu. U slučaju da se radi o brojevima koji se mogu deklinirati, stavlja se u nominativ muškog roda.

/PART je tag predviđen za čestice, odnosno funkcionalne riječi[56] (engl. *function words*) koje moraju biti povezane s drugim riječima kako bi dobile značenje i koje ne spadaju niti u jednu drugu skupinu riječi, a često izražavaju gramatičke kategorije negacije, načina, vremena i sl. U našem je slučaju ovaj tag dodijeljen jedino riječi *non*, ali je, naravno, moguće zamisliti i druge riječi koje trebaju dobiti ovaj tag, poput *haud*.

/PRON je tag kojim se označuju zamjenice, odnosno riječi koje zamjenjuju imenice ili imenske skupine i čije se značenje može iščitati iz konteksta. Neke od riječi koje se tradicionalno smatraju zamjenicama u našem se slučaju označuju kao determinatori. Tagom /PRON označene su, zasad, osobne zamjenice (*ego*, *tu*), odnosne zamjenice (*qui, quisquis*) i upitne zamjenice (*quis*).

/PROPN je tag koji služi za vlastite imenice, odnosno imenice koju označuju ime ili dio imena neke određene individue, mjesta ili predmeta. Osim očekivanih osobnih imena i imena gradova, pokrajina i sl., u našem su tekstu ovim tagom označene i neke riječi koje su svojim oblikom pridjevi. Radi se o etnicima (npr. *Genuenses* ili *Slovini*) te o pridjevima koji funkcioniraju kao prezimena, poput pridjeva *Nadasdinus* (latinizrani oblik prezimena Tome Nádasdyja). Iznimka su dva biblijska imena, *Symon Zelotes* i *Iacobus Iustus*, jer se ne radi o prezimenu *per se*, već o općem pridjevu koji služi kao svojevrsni „nadimak“.[57] Za leme nekih vlastitih imenica koje se ne mogu naći u rječnicima, pretpostavlja se standardni oblik (npr. za pojavnicu *Bucharim*, latinsko ime grada Bakra, uzima se lema *Bucharis*, kao nominativ jednine, premda to ime nećemo naći u korištenim rječnicima).

/PUNCT je tag za interpunkcije, tj. razgotke, odnosno znakove i grupe znakova koji ne spadaju u slova i kojima se razgraničuju rečenice i dijelovi rečenica. Radi o znakovima poput točke (.),

[56] Hrvatski termini koji bi mogli ovome odgovarati su nepunoznačne ili gramatičke riječi.
[57] Naravno, moguće je tvrditi da su upravo tako nastala i brojna prezimena pa da i ove riječi treba smatrati prezimenima, ali, uzimajući povijesni kontekst u obzir, nije odabrana ta logika.

zareza (*,*), upitnika (*?*), uskličnika (*!*), dvotočke (*:*), trotočke (...) i sl. Lema je kod interpunkcija uvijek jednaka pojavnici.

/SCONJ je tag koji se koristi za označavanje zavisnih, odnosno subordiniranih veznika, koji vežu dijelove rečenice na način da jedan dio postaje konstituent drugog. Kod nas se uvijek radi o povezivanju glavne i zavisne rečenice. Premda ih se može relativno jednoznačno odrediti prateći popis zavisnih veznika pojedinih zavisnih rečenica, treba paziti na slučajeve homografije, gdje neki veznici izgledaju kao (ili su čak po postanku) druge vrste riječi (npr. *cum* kao prijedlog ili veznik ili *ubi* kao prilog ili veznik).

/VERB je tag koji se koristi za glagole, riječi koje označuju neke radnje, događanja, stanja ili zbivanja, a često mogu biti predikati i ravnaju mnogim drugim dijelovima rečenice. Osim finitnih oblika, u našem se označavanju ovaj tag koristio i za sve nefinitne oblike, poput infinitiva, participa i gerundiva. Svi se navedeni oblici lematiziraju kao glagol od kojeg nastaju.

UD tagset predviđa još i tagove /SYM i /X, koji se u našem radu još ne koriste. /SYM je tag za simbole, odnosno znakove koji nisu alfanumerički, a koji se od interpunkcija razlikuju po tome što ih se može zamijeniti riječima, /X je tag za one pojavnice koje se ne mogu smisleno povezati s niti jednim drugim tagom, a autori upozoravaju da se treba koristiti vrlo oprezno i restriktivno. Moguća uporaba taga /X u našem korpusu bila bi za strane riječi koje se ne mogu analizirati kao dio rečenice.

4.4.2. Neki problemi tagiranja i lematizacije

Kako je već navedeno, ne postoje sasvim jasne i jednoznačne upute za tagiranje i lematizaciju i moguće je na više načina riješiti pojedine probleme koji se javljaju. S obzirom na to, nerijetko se javljaju dvojbe oko korištenja pojedinih tagova i pitanja u vezi dodjeljivanja leme. Ovdje ćemo opisati neke od problema s kojima smo se susreli prilikom tagiranja i lematizacije malog dijela uzorka. Njih možemo smatrati „slabim točkama“ našeg sustava označavanja i može se očekivati da će na ovim mjestima u budućnosti doći do promjene principa.

Jedan od najvećih problema predstavljaju determinatori. Probleme djelomično uzrokuje činjenica da je determinatore općenito teško definirati i odrediti, jer nije uvijek jasna granica između determinatorā, zamjenicā i pridjevā. O tome svjedoče i naputci za označavanje tzv. pronominalnih riječi koje daje *UD tagset*[58], a i Zeman (2018.) ih u svom pregledu POS tagova navodi kao primjer kategorije riječi čije su granice vrlo nejasne. Uz to je otegotna okolnost i

[58] Vidi https://universaldependencies.org/u/overview/morphology.html#pronominal-words (pristupljeno 19. travnja 2021.).

što determinatori kao vrsta riječi u tradicionalnim gramatikama ne postoje pa je tim teže odrediti koje točno riječi pripadaju ovoj skupini. Za orijentaciju je koristan rad Olge Spevak (2014., str. 41.) u kojem promatra determinatore u sklopu istraživanja imenske skupine u latinskom, ali, kako i sama autorica tvrdi (str. 1.), ne radi se o točnom popisu već o promatranju njihovih tipičnih osobina. Kao referenca za pregled riječi koje spadaju u kategoriju determinatora može poslužiti Pinksterova *The Oxford Latin Syntax* (2015., str. 49.) koja donosi pregled i podjelu determinatora u latinskom. Međutim, Pinkster među njima navodi i pokazne zamjenice (*hic*, *iste*, *ille*), za koje tvrdi da se mogu koristi i kao zamjenice i kao determinatori. Upravo to čini možda i ključni problem s determinatorima – iste riječi ponekad se ponašaju kao zamjenice, a ponekad kao determinatori. Primjerice, *ille* može zamjenjivati konkretnu riječ (npr. u rečenici *quę res illi valde grata fuit*)[59] ili može dodatno pojašnjavati neku riječ u kontekstu (npr. *cuius voluntate hic excuditur illa nostra oratiuncula*)[60]. Pogledamo li koje su leme označene kao zamjenice, a koje kao determinatori u latinskim korpusima UD-a[61], vidjet ćemo da između dva popisa postoje neka preklapanja i da ta preklapanja nisu ista između različitih korpusa. Da su tagovi /DET i /PRON dodjeljivani istoj lemi može biti odraz upravo stava da ista lema može pripadati i jednoj i drugoj kategoriji, ovisno o kontekstu, ali može biti i posljedica nedosljednog označavanja. Mogućnost dodjeljivanja različitih tagova istoj lemi spominje i Zeman (str. 38.), a takav bi se pristup, uz još razmatranja i rasprave, mogao prihvatiti i u našem korpusu.

Iz sličnih razloga moguće je raspravljati i o opravdanosti neselektivnog označavanja participa i gerundiva kao glagola. Premda UD predviđa da se svi gerundivi i participi označuju kao glagoli kojima će se zatim dodati osobina (engl. *feature*)[62] kojom će se naznačiti da se radi o pridjevskim oblicima, neki su se od njih u značenju toliko primakli pridjevima da bi ih se tako moglo i tagirati te lematizirati. Primjerice, gerundiv *observandissimus* u *Du Cangeu* ima vlastitu natuknicu i navodi se kao *„titulus honorarius“*. Ako je u pitanju zaista oblik uvriježen u obraćanju, ima razloga razmisliti barem o dodjeljivanju vlastite leme, ako ne i taga /ADJ. Još jedan primjer, koji se ne pojavljuje u našem dijelu uzorka, ali je ilustrativan, možemo naći i u participu *sapiens*. Premda se on čak i u rječniku *Lewis & Short* navodi kao dio natuknice glagola *sapio*, jasno je iz objašnjenja da se radi o participu koji se svojim značenjem pomakao od

[59] Andreis, F. T., *Epistulae ad Thomam Nadasdinum*, 1532., versio electronica, (pismo 4. 12. 1541.), http://croala.ffzg.unizg.hr/cgi-bin/getobject.pl?c.210:5.croala.
[60] Andreis, F. T., *Epistulae ad Thomam Nadasdinum*, 1532., versio electronica, (pismo 30. 11. 1545.), http://croala.ffzg.unizg.hr/cgi-bin/getobject.pl?c.210:6.croala.
[61] Vidi https://universaldependencies.org/treebanks/la-comparison.html#morphology (pristupljeno 19. travnja 2021.).
[62] Vidi https://universaldependencies.org/u/feat/VerbForm.html#Part (pristupljeno 19. travnja 2021.).

glagola i zadobio specifičnije određenje (kako navodi *Lewis & Short*, pod utjecajem Grka dobiva značenje slično pridjevu *σοφός*). Ova bi se rasprava mogla protegnuti općenito na oblike koji su supstantivirani, odnosno treba pogledati ima li razloga i temelja tagirati ih i lematizirati kao imenice, ili ih tretirati prema izvornom obliku.

Kod nekih pridjeva postoji i pitanje korištenja taga /PROPN, koji se koristi za vlastite imenice. Premda bi se prema samom nazivu očekivalo da je ovaj tag rezerviran za imenice, uputstva UD-a navode kako se može koristiti i za „riječi imenskog sadržaja" (engl. *„nominal content word"*), u što bi se mogli uračunati i pridjevi. Kao što je već navedeno u objašnjenju taga, nije sasvim jasno bi li pridjevi u situacijama poput *Symon Zelotes* i *Iacobus Iustus* trebali biti označeni tagom /ADJ ili tagom /PROPN.

Još jedan problem čine kratice. Najjednostavniji slučaj je kada se skrati jedna riječ, jer ju tada možemo označiti tagom koju bi imala da nije skraćena i u lemi ju razriješiti u puni oblik. Takav je slučaj s kraticom *eg.* u značenju *egregius*. Nju smo tagirali tagom /ADJ i lematizirali kao *egregius*. Međutim, situacija je kompliciranija s kraticama poput *etc.* koje krate dvije ili više riječi. U ovom slučaju nije jasno kako bi kraticu trebalo označiti. U našem je korpusu ona označena kao prilog, jer je tako u nekim slučajevima označena u korpusu *ITTB*. U drugim je slučajevima, međutim označena tagom /X, što bi se trebalo maksimalno izbjegavati. Dodatni je problem kako takvu kraticu lematizirati. S obzirom na to da se ovdje radi o specifičnoj lingvističkoj situaciji, jer se kratica kao token ne može razriješiti nakon što smo ju obradili tagerima i lematizatorima, treba pronaći neko rješenje.

U slučaju riječi *quoque* i *verum* javlja se problem sukobljene literature. Naime, jedni će izvori ove riječi smatrati usporednim veznicima, dok će ih drugi klasificirati kao priloge. Tako je riječ *quoque* u gramatikama *GGP* (str. 157.) i *Menge* (2004., str. 593.) te u rječniku *Lewis & Short* navedena kao veznik, dok je u Žepićevu rječniku te gramatikama *Allen & Greenough* (1903., str. 197.) i *Bennett* (1908., str. 227.) označena kao prilog. Treba napomenuti da se u dvjema potonjim gramatikama prilozi navode u sklopu čestica. S druge strane, sve navedene gramatike navode smatraju riječ *verum* veznikom (Gortan, Gorski, & Pauš, str. 158.; Menge, Burkard, & Schauer, str. 610.; Allen & Greenough, 1903.; Bennet, 1908., str. 225.), dok je oba navedena rječnika smatraju prilogom. U našem je sustavu tagiranja odlučeno ove riječi označiti kao veznike, ali i ovdje postoji prostor za raspravu.

Na kraju možemo spomenuti i problem tagiranja koji je specifičan za naš korpus. Radi se o tekstu Franje Milašinovića *Viator Zagorianus Jožko Hranjec*, jer je riječ o makaronskoj pjesmi u kojoj se pojavljuju „čiste" hrvatske riječi te „miješane" hrvatsko-latinske riječi, odnosno

hrvatske riječi s latinskim gramatičkim nastavcima. Četiri su takve pojavnice u našem malom dijelu uzorka. To su *Ferkuljka, klafrando, kumi* i *kumeque*.[63] U ovom slučaju manji problem čine POS tagovi, jer je *Ferkuljka* ime pa dobiva tag /PROPN, *klafrando* je gerundiv pa mu se prema pravilima dodjeljuje tag /VERB, a *kumi* i *kumeque* su imenice i dobivaju tag /NOUN. Veći je problem pitanje lematizacije. Riječ *Ferkuljka* lematizirana je identičnom lemom *Ferkuljka*, jer se radi o osobnom imenu i to u nominativu. Gerundivu *klafrando* u našem je slučaju dodijeljena zamišljena latinska glagolska lema *klafro* (radi se, naravno, o 1. konjugaciji), premda je razmotrena i mogućnost lematiziranja hrvatskom lemom *klafrati*. Nešto je kompliciranija situacija s pojavnicama *kumi* i *kumeque*, jer nije sasvim jasno treba li ih lematizirati kao hrvatske riječ ili kao latinske. Naime, mogli bismo zamisliti da je oblik *kumi* vokativ makaronske riječi *kumus* i tako ga lematizirati (što je u našem slučaju učinjeno), premda bi se moglo raditi o hrvatskoj kratkoj množini riječi *kum*. S druge strane, jasno je da *kume* nije latinska riječ (jer Milašinović uredno koristi *ae* kao nastavak), već hrvatska, pa je stoga lematizirana kao *kuma*, a taj oblik slučajno odgovara i potencijalnoj latinskoj lemi.

Kao što smo već zaključili, ovaj popis problema nije iscrpan te će se tijekom označavanja kompletnog uzorka gotovo sigurno pojaviti dodatna pitanja.

4.4.3. Harmonizacija skupova tagova i evaluacija pojedinih modela

Zadnji je korak, koji u našem projektu nije proveden do kraja, usporedba rezultata pojedinih tagera i lematizatora sa zlatnim standardom i odabir najboljeg. Za mjeru uspješnosti je zasad odabrana najjednostavnija moguća metoda, odnosno postotak točno označenih riječi u odnosu na zlatni standard. Rezultati se trenutno provjeravaju djelomično ručno (uz pomoć mogućnosti programa *MS Excel*), a barem dio posla mogao bi se ubuduće automatizirati. Zbog različite metodologije, odvojeno ćemo promatrati rezultate lematizacije i rezultate tagiranja. Kao što je napominjano do sada, i ovdje otvoreni problemi i njihova rješenja podložni su promjenama i korekcijama u nastavku projekta.

Pri procjeni lematizacije, kriteriji za točnu lemu bit će nešto blaži nego za točan POS tag. Naime, zbog različitih standarda koji različiti lematizatori koriste, poput točnog oblika leme, grafije i sl., ručni pregled s blažim kriterijima omogućava veću fleksibilnost i bolju procjenu svakog modela. Primjerice, priznat ćemo svaku prihvatljivu grafiju leme (neovisno o korištenju grafema *u/v* ili *i/j*) i drugačije kanonske oblike lema (npr. *ac* umjesto *atque*). Općenito govoreći,

[63] Čitava rečenica glasi: „*Interea Ferkuljka senex klafrando perorat,/Desinite o, inquit kumi kumeque feroci/ Concertare viro*."

točna lematizacijska strategija nije presudna za povezivanje s *LiLa*-om jer postoji sustav za harmonizaciju različitih pristupa lematizaciji (Mambrini & Passarotti, 2019.). Ipak, treba moći vrlo jasno razložiti zašto se neke leme prihvaćaju, a zašto neke odbacuju. Među spomenutim algoritmima za obradu nalaze se i dva lematizatora, *LEMLAT* i *LatMor*, koji nam za svaku pojavnicu vraćaju sve moguće leme jer analiziraju oblik bez uzimanja konteksta u obzir. Kod ova će se dva lematizatora rezultat označiti kao točan samo u slučaju da su sve leme koje oni ponude zaista mogući oblici pojavnice. U slučaju da ponude lemu koja nikako ne može biti pridružena nekom obliku ili u slučaju da ne dodijele lemu, smatrat će se da su pogriješili. Sličan je slučaj i s modelima koji su trenirani uz pomoć *TreeTaggera*, koji ponekad nudi više mogućih rješenja za lemu.

Kod POS tagera su kriteriji nešto stroži, jer je lakše unaprijed odrediti točnije kriterije prihvaćanja ili ne prihvaćanja taga. Ono što predstavlja problem kod tagiranja različiti su skupovi tagova. Naime, kao što je vidljivo iz pregleda (4.3.3) skupovi sadržavaju različit broj različitih tagova, a svaki treba usporediti s našim zlatnim standardom, koji je tagiran *UD tagsetom*. Koristeći se matematičkim rječnikom, mogli bismo reći da je problem što ne možemo napraviti bijekciju između skupa tagova korištenog u našem zlatnom standardu i ostalih skupova tagova. S obzirom na to, treba provesti svojevrsnu harmonizaciju, odnosno razraditi pravila o prevođenju svakog pojedinog skupa tagova na skup tagova korišten u zlatnom standardu. kako bismo mogli vidjeti jesu li točno označili pojavnice. Detaljniji pregled ove problematike donosi Gamba (str. 55.-58.).

Pogledajmo neke od problema harmonizacije. Najjednostavniji su slučaj modeli koji koriste *UD tagset*, koji je korišten za označavanje zlatnog standarda. Međutim, ni kod njih situacija nije sasvim jasna, jer ne koriste svi sve tagove iz *UD tagseta*, niti ih koriste na isti način. Modeli trenirani na *LLCT*-u neće koristiti tagove /INTJ i /SYM. Tag /SYM se u našem korpusu ne pojavljuje, stoga nema potrebe za prijevodom, ali problem nastaje jer mi tag /INTJ koristimo. U tom će slučaju biti ključno provesti „prevođenje“, odnosno vidjeti koji se tag koristi u spomenutom korpusu u slučaju usklika. Pregledom korpusa, možemo utvrditi da se usklici označavaju tagom /ADV, pa ćemo modelima koji su trenirani na LLCT-u priznavati tag /ADV za pojavnice označene tagom /INTJ u zlatnom standardu. Ista stvar vrijedi i za modele trenirane na *ITTB*-u (verzijama 2.3 i 2.6), koji također ne koristi /INTJ i /SYM.

Još jedan problem kod *ITTB*-a počiva u njegovim različitim verzijama, koje tag /DET koriste na različit način.[64] Naime *ITTB* (verzija 2.1) koristi /DET u punom smislu toga taga (odnosno za determinatore općenito), dok verzije 2.3 i 2.6 koriste taj tag samo za protočlan *ly*. To znači da modeli koji su temeljeni na tim verzijama *de facto* nikad neće koristiti ovaj tag (budući da u našem korpusu protočlan *ly* ne postoji). Sličnu situaciju imamo i s modelima treniranim na *Perseusu*, koji tag /DET uopće ne koristi. U slučaju modela temeljenih na ta dva korpusa, pojavnice koje smo mi tagirali kao determinatore bit će tagirane ili kao pridjevi ili kao zamjenice. Ta situacija razumljiva je s obzirom na općenitu problematiku determinatora kao vrste riječi. U tim slučajevima priznat ćemo tagove /ADJ i /PRON kao točan odgovor. Isti princip prenosimo i na tag /PART koji se ne koriste *ITTB-23, ITTB-26*, *PROIEL* ni *Perseus*. Budući da oni riječi koje smo mi označili kao čestice smatraju prilozima, kod tih ćemo pojavnica priznati tag /ADV. Treba još napomenuti i da *Perseus* ne koristi tagove /PROPN, umjesto kojeg ćemo priznavati tag /NOUN, i /AUX, umjesto kojeg ćemo priznavati tag /VERB.

Navedeni principi koriste se i kod prijevoda ostalih skupova tagova u UD. Kod nekih je tagova moguće napraviti jasno preslikavanje između skupova, poput taga za imenice koji ima jasan ekvivalent u svim skupovima. Kod nekih tagova treba priznati više mogućih rješenja. Primjerice, nijedan drugi skup tagova ne prepoznaje determinatore kao zasebnu skupinu, pa kod njih priznajemo tagove za pridjeve ili zamjenice. Slično je i s veznicima, koji u nekim skupovima nisu podijeljeni na usporedne i zavisne, pa u oba slučaja priznajemo isti tag. U nekim slučajevima, pak, jedan tag iz našeg skupa obuhvaća više tagova iz nekog drugog skupa. Primjerice za tag /PUNCT ćemo kod skupa *OMNIA* priznati i /PON i /SENT. Veći problem nastaje kada u nekom skupu postoji tag za koji nemamo pravi ekvivalent u zlatnom standardu. Npr. skup *Lamap* koristi tag /ABBR za kratice i on se ne može preslikati u naš skup tagova. U tom slučaju priznat ćemo taj tag kad se radi o kratici. Prikaz preliminarnog sustava „prevođenja“ iz ostalih skupova u zlatni standard vidljiv je u Tablica 8., uz napomenu da postoji mogućnost njegove izmjene tijekom daljnjeg rada na projektu.

Općenito govoreći, rigoroznost kriterija pri procjenjivanju uspješnosti različitih modela nije jednoznačno zadana, već ovisi o željenim ishodima evaluacije. Primjerice, ako bismo željeli biti stroži, mogli bismo sve *UD tagsetove* evaluirati na jednak način, unatoč činjenice da neke tagove ne koriste ili ih koriste na različite načine (npr. slučaj s *ITTB* i tagom /DET). Na taj bismo se način osigurali da će procjena odabrati onaj model koji najviše odgovara našem načinu

[64] Vidi https://github.com/UniversalDependencies/UD_Latin-ITTB/tree/master#changelog.

tagiranja. S druge strane, možemo zamisliti i blaži način ocjenjivanja, koji u većoj mjeri u obzir uzima individualne principe tagiranja svakog modela, ali u tom slučaju riskiramo prihvaćanje i potencijalno neprihvatljivih opcija, zbog šire definicije točnog taga. Ova je rasprava posebno bitna kada se radi o automatiziranoj procjeni i usporedbi, gdje ljudska procjena pojedinih slučajeva nije moguća.

I ovdje je važno naglasiti da su odluke i principi harmonizacije preliminarni, odnosno podložni promjenama, a u ovom su radu navedeni više u svrhu ilustracije problema koji proizlaze iz procesa komparativne evaluacije različitih modela, nego da bi se dali sigurni i finalni odgovori na navedena pitanja.

Za onaj model ili algoritam koji se pokaže najuspješnijim bit će provedena provjera točnosti na čitavom uzorku (na svih 58 185 tokena). Kada se utvrdi koliko je spomenuti model uspješan, i uz moguću provjeru najčešćih grešaka, otvara se mogućnost lematizacije čitave *CroALa*-e.

Zlatni standard	LLCT	ITTB (2.1)	ITTB (2.3&-2.6)	PROIEL	Perseus	CLTK	Lamap[65]	OMNIA	CHSTS
/ADJ	/ADJ	/ADJ	/ADJ	/ADJ	/ADJ	/A	/ADJ	/QLF	/ADJ, /ORD[66]
/ADP	/ADP	/ADP	/ADP	/ADP	/ADP	/R	/ADP	/PRE	/AP
/ADV	/ADV	/ADV	/ADV	/ADV	/ADV	/D	/ADV	/ADV	/ADV, /PTC
/AUX	/AUX	/AUX	/AUX	/AUX	/VERB	/V	/ESSE	/VBE	/V
/CCONJ	/CCONJ	/CCONJ	/CCONJ	/CCONJ	/CCONJ	/C	/CC	/CON	/CON
/DET	/DET	/DET	/ADJ, /PRON	/DET	/ADJ, /PRON	/A, /P	/DIMOS, /POSS, /INDEF, /ADJ	/QLF, /PRO	/ADJ, /PRO
/INTJ	/ADV	/INTJ	/ADV	/INTJ	/INTJ	/E	/INT, /EXCL	/INT	/ITJ
/NOUN	/NOUN	/NOUN	/NOUN	/NOUN	/NOUN	/N	/N	/SUB	/NN
/NUM	/NUM	/NUM	/NUM	/NUM	/NUM	/M	/N	/NUM	/NUM, /DIST
/PART	/PART	/PART	/ADV	/ADV	/ADV	/D	/ADV	/ADV	/ADV
/PRON	/PRON	/PRON	/PRON	/PRON	/PRON	/P	/PRON, /REL	/PRO	/PRO
/PROPN	/PROPN	/PROPN	/PROPN	/PROPN	/NOUN	/N	/NPR	/NAM	/NE. /NP
/PUNCT	/PUNCT	/PUNCT	/PUNCT	/PUNCT	/PUNCT	/U	/PUN, /SENT	/PON, /SENT	/$
/SCONJ	/SCONJ	/SCONJ	/SCONJ	/SCONJ	/SCONJ	/C	/CS	/CON	/CON
/VERB	/VERB	/VERB	/VERB	/VERB	/VERB	/V	/V	/VBE	/V
/X	/X	/X	/X	/X	/X	/X	/FW[67]	/	/XY, /FM[68]

Tablica 8. Tablični prikaz preliminarnog sustava prevođenja različitih skupova tagova u onaj korišten u zlatnom standardu

[65] U tablici nisu navedeni tagovi koji nemaju jasan ekvivalent u skupu tagova korištenom za označavanje zlatnog standarda.
[66] Budući da se u našem korpusu redni brojevi tagiraju kao pridjevi, tag /ORD se prihvaća u slučaju taga /ADJ u zlatnom standardu.
[67] Oznaka za strane riječi uzeta je kao ekvivalent tagu /X, jer je njime moguće označiti riječi iz drugih jezika koje se mogu smisleno označiti na drugačiji način.
[68] Vidi funsotu 67.

5. Uključivanje *CroALa*-e u *LiLa*-u

Nakon što se provede potpuna lematizacija uzorka (a u finalu i korpusa), može se pristupiti povezivanju s *LiLa*-om. Budući da ovaj postupak još nije započet, ukratko ćemo opisati osnovne principe povezivanja, koji općenito vrijede za uključivanje tekstualnih resursa u *LiLa*-u. Da bismo bolje razumjeli postupak uključivanja, najprije treba ukratko objasniti neke važnije koncepte u *LiLa*-inoj arhitekturi.

Kao što smo u uvodu objasnili, *LiLa* koristi lemu kao centralni koncept za povezivanje resursa. Međusobni odnosi raznih elemenata u *LiLa*-i zapisani su pomoću tzv. RDF tripleta, koji se sastoje od tri dijela – predikata, subjekta i objekta – a mogu se zamisliti i prikazati kao mrežni grafovi[69] (Passarotti, i dr., 2020.). Ukratko, u RDF tripletima se predikatom izriče odnos subjekta i objekta. Primjerice, informaciju da je riječ *terra* ženskog roda mogli bismo izraziti sljedećim tripletom: *terra* (subjekt), *imaRod* (predikat), *ženski* (objekt).

Premda se elementi tripleta teoretski mogu bilježiti na bilo koji način, standardna je praksa da svaki element ima svoj tzv. URI (engl. *Uniform Resource Identifier*), odnosno jedinstvenu identifikacijsku oznaku (Cimiano, Chiarcos, McCrae, & Gracia, 2020., str. 12.). URI bi trebao biti jedinstven, odnosno odnositi se samo na jedan entitet, i trebao bi omogućavati nedvosmislenu identifikaciju entiteta na koji se odnosi. S obzirom na to da *LiLa* prati FAIR principe (Wilkinson, i dr., 2016.), odnosno pazi na pronalažljivost, dostupnost, interoperabilnost i ponovnu upotrebljivost podataka koje objavljuje, [70] URI-ji u *LiLa*-i koriste HTTP kako bi se podatcima moglo lako pristupiti. Zbog toga shema URI-ja u *LiLa*-i ima sljedeći oblik: `http://{domena}/{tip}/{koncept}/{referenca}`. Takav se URI dodjeljuje svim resursima (npr. lemama, oblicima, vrstama riječi, korijenima riječi, etimologijama itd.), odnosima (ono što su predikati u RDF-u, npr. *imaOsnovu*, *imaRod*, *jeVrstaRiječi* itd.) i vrijednostima (npr. imenica, ženski rod, nominativ, prezent, prvo lice itd.). Primjerice, riječ *amor* ima URI `http://lila-erc.eu/data/id/lemma/88717`, dok je URI za predikat koji prikazuje da RDF objekt označava vrstu riječi RDF subjekta `http://lila-erc.eu/ontologies/lila/hasPOS`. Dakle, pomoću URI-ja bismo mogli informaciju da je riječ *amor* po vrsti riječi imenica, zapisati sljedećim tripletom:

```
<http://lila-erc.eu/data/id/lemma/88717> ;
<http://lila-erc.eu/ontologies/lila/hasPOS> ;
<http://lila-erc.eu/ontologies/lila/noun> .
```

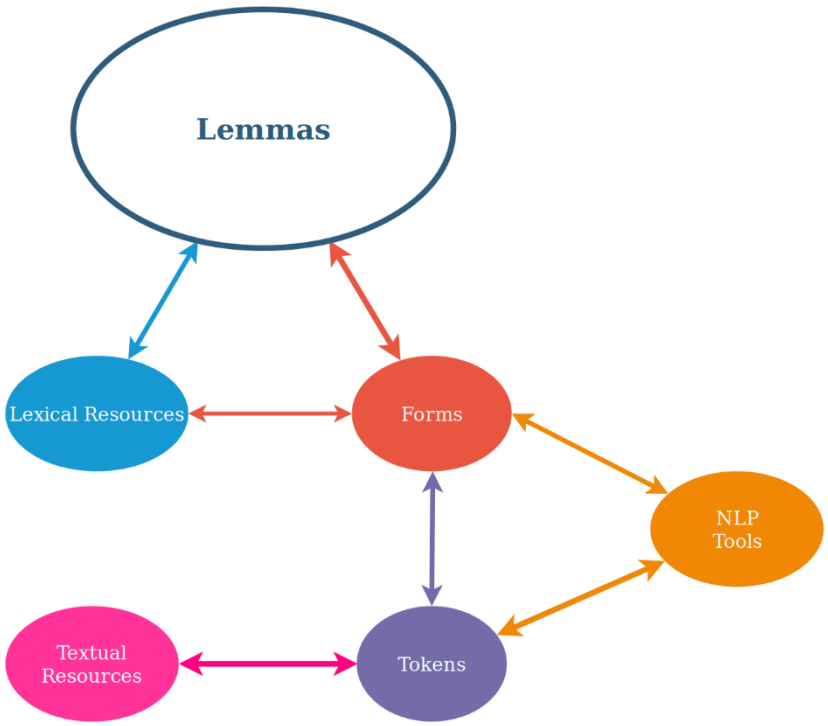
[69] Vidi https://enciklopedija.hr/natuknica.aspx?ID=70130 (pristupljeno 21. 4. 2021.).
[70] Vidi http://lib.irb.hr/web/hr/vijesti/item/2245-fair_principi.html (pristupljeno 22. 4. 2021.)

Jedna je od ključnih karakteristika *LiLa*-e da svaka lema ima jedinstveni identifikator pomoću kojeg je spojena s ostalim resursima. Na taj način, dodavanjem URI-ja svakoj pojavnici, korpusi se mogu povezati s lemama iz *LiLa*-e, a preko njih i svim drugim podatcima koji se u njoj nalaze.

Upravo je povezivanje svake pojavnice s određenom lemom sljedeći korak u uključivanju *CroALa*-e, u *LiLa*-u. Kako bismo to napravili najprije će trebati provjeriti povezanost lema iz korpusa s lemama u *LiLa*-i, odnosno vidjeti koliko je lema iz *CroALa*-e povezano s jednom lemom, dvije ili više lema ili nijednom lemom u *LiLa*-i. U slučaju povezanosti s dvije ili više leme, morat će se izvesti dodatno razlučivanje lemā, a u slučaju da lema nedostaje dodat će se u tzv. banku lemā.

Nakon toga može se pristupiti gradnji RDF tripleta kojim će se pojavnice povezati s lemama. Kako bi se taj posao mogao obaviti, svaka pojavnica mora dobiti vlastiti URI. Za tu su potrebu pripremljene tokenizirane verzije tekstova u XML-u[71], za koje se mogu generirati URI-ji. Zadnji je korak dodjeljivanje metapodataka o tekstovima (reference na tokene, autore itd.) kako bi tekstovi bili pretražljiviji.

U finalu bismo trebali dobiti verziju tekstova iz *CroALa*-e u kojoj preko svake riječi možemo pristupiti resursima iz *LiLa*-e koji su s njom povezani.



Slika 1. Pojednostavljeni prikaz načina povezivanja resursa u LiLa-i

[71] Vidi https://github.com/nevenjovanovic/croatiae-auctores-latini-textus/tree/master/subset-tokenized (pristupljeno 22. 4. 2021.)

Treba istaknuti i još jednu bitnu osobinu *LiLa*-e, koja ju čini posebno vrijednom u lematizaciji latinskih tekstova. Naime, pomoću jedinstvenih identifikatora (tj. URI-ja), *LiLa* stvara centralni popis lemā u kojem je moguće nedvosmisleno odabrati jednu između više homografnih lema. Drugim riječima, moguće je jednoznačno odrediti na koju lemu mislimo i koristiti jedinstveni identifikator u različitim projektima kako bismo se na tu lemu referirali. Usporedimo li taj sustav za lematizaciju s nekim drugim sustavima, primjerice onakvim kakav koristi platforma *Arethusa*[72] (u kojoj se homografne leme međusobno razlikuju pomoću broja dodanog lemi – npr. *dico1* i *dico2* – gdje broj označava kojoj natuknici iz rječnika *Lewis & Short* lema odgovara), vidjet ćemo da je razlikovanje lemā u *LiLa*-i jednostavnije i otvara manje prostora za zabunu. Naime, način na koji *Arethusa* razlikuje leme čini taj sustav ovisnim o pregledavanju vanjskog rječnika i ne postoji način da se direktno iz leme dođe do značenja. S druge strane, *LiLa* značenje ne definira pomoću rječnika, već je svaka lema povezana s određenim značenjem preuzetim iz leksičko-semantičke baze podataka za latinski jezik *Latin WordNet*[73] (Franzini, i dr., 2019.).

[72] Vidi https://www.perseids.org/tools/arethusa/app (pristupljeno 3. svibnja 2021.).
[73] Vidi https://latinwordnet.exeter.ac.uk/ (pristupljeno 3. svibnja 2021.)

6. Primjena projekta u nastavi

Kao što je najavljeno, ovaj rad sadrži i pedagošku komponentu. U ovom ćemo dijelu rada razmotriti mogućnosti primjene predstavljenog projekta u nastavi latinskog jezika u školama, odnosno u gimnazijskim programima.

Budući da naš projekt i ovaj rad spajaju teme iz dva područja, digitalne humanistike i hrvatske latinske književnosti, trebamo najprije provjeriti koje je njihovo mjesto u poučavanju latinskog u školama. Prema aktualnom *Kurikulumu nastavnog predmeta Latinski jezik za osnovne škole i gimnazije* (2019.) hrvatski se latinisti obrađuju i spominju jedino u klasičnim gimnazijama u 4. razredu (i kod početničkog i kod nastavljačkog programa). Teme hrvatskog latiniteta nisu predviđene za obradu u općegimnazijskom programu učenja latinskog. Ipak, učenici bi neke hrvatske latiniste trebali poznavati s nastave *Hrvatskog jezika*, jer se nalaze na popisu književnih tekstova za cjelovito čitanje ili čitanje ulomaka u *Kurikulumu nastavnog predmeta Hrvatski jezik* (2019.), premda hrvatska književnost pisana latinskim jezikom u istom dokumentu nije izričito spomenuta.[74]

Što se tiče digitalne humanistike, ona se u *Kurikulumu za Latinski jezik* izrijekom spominje u kontekstu povezivanja latinskog s drugim predmetima i međupredmetnim temama (str. 65.), kao primjer mogućnosti korištenja informacijske tehnologije. Jedini konkretniji primjer primjene digitalnih alata u nastavi nalazimo u domeni *Jezična pismenost*, gdje se kao jedna od razrada ishoda navodi da se „(učenik) služi dvojezičnim rječnicima u knjižnome i digitalnom formatu“. Prema iskustvu autora ovog rada u nastavi latinskog, digitalni formati često se stavljaju u drugi plan ili sasvim izostaju, a najpopularniji je digitalni resurs platforma *Perseus Digital Library*[75], kao izvor tekstova, ali i kao pomagalo pri njihovu čitanju.

S obzirom na takvu situaciju, prikazat ćemo tri moguća načina primjene ovog projekta (ili njegovih dijelova) u nastavi. Prvi je pristup opći pregled i predstavljanje projekta, drugi primjena rezultata projekta prilikom obrade hrvatske novolatinske književnosti, a treći primjena nekih segmenata projekta na tekstovima koji se koriste u nastavi.

Projekt učenicima možemo predstaviti u obliku klasičnog nastavnog, frontalnog tipa, u trajanju 45 minuta, u kojem je glavna metoda metoda rasprave, a učinkovitoj realizaciji sata bi u velikoj mjeri doprinijela prezentacija u ulozi vizualnog pomagala. S obzirom na ranije navedene karakteristike kurikuluma, ovaj je pristup najpogodniji za 4. razrede klasičnih gimnazija, ali je

[74] Spominje se samo obrada pojedinih književnih razdoblja, poput humanizma, renesanse, baroka itd.
[75] Vidi http://www.perseus.tufts.edu/hopper/ (pristupljeno 23. travnja 2021.).

uz prilagodbu visoko primjenjiv i na ostale razine učenja. U uvodnom dijelu sata, gdje učenike treba motivirati i uvesti u temu, može se postavljanjem pitanja i kratkom vođenom raspravom uvesti u glavne elemente projekta, odnosno hrvatski latinitet i digitalnu humanistiku. Dobre polazišne točke za ovo su pitanja „čime se danas istraživači latinskog bave?“ i „na koji način današnji istraživači latinskog istražuju jezik i književnost?“. Kada smo učeniku uveli u temu, u središnjem dijelu sata treba im predstaviti projekt. Nemoguće je prikazati sve elemente i korake projekta, a ni one koji su odabrani ne možemo prikazivati detaljno, stoga treba odabrati što želimo da učenici čuju i, nadajmo se, zapamte. Kao početne i ključne koncepte treba objasniti *CroALa*-u (kao zbirku tekstova dostupnu na internetu), *LiLa*-u (kao računalni sustav ili bazu znanja o latinskom jeziku dostupnu na internetu) i leme (kao „temeljne“ oblike riječi, slične onima koje nalazimo u npr. rječnicima). Želimo da učenici shvate da je glavni cilj zabilježiti znanje koje bi inače ostalo u glavama istraživača (a ovako zabilježeno može se koristiti za daljnja istraživanja) i dobiti zbirku tekstova koja je povezana s mnogim drugim resursima za latinski. Kako bismo to dočarali, možemo učenicima prikazati jednu rečenicu iz *CroALa*-e i na odabranim riječima pokazati kako izgleda proces spajanja s *LiLa*-om. Ključno je da se kod objašnjavanja držimo osnovnih ideja i izbjegavamo ulazak u tehnička pitanja. Učenicima se kao zadatak može zadati i da sami probaju ručno, na papiru, povezati zadane riječi s *LiLa*-om u obliku svojevrsne „mentalne mape“. Pri odabiru rečenice za prezentaciju i vježbu, treba paziti da se radi o gramatički jednostavnoj rečenici koju će učenici brzo shvatiti, ali bilo bi dobro i da se u njoj nalaze riječi na kojima se mogu pokazati izazovi tagiranja i lematizacije (npr. oblici koji se, van konteksta, mogu protumačiti kao različite vrste riječi ili oblici kod kojih pri izboru leme treba paziti na homografe). Na kraju, možemo vrlo kratko rezimirati ono što smo prikazali (primjerice sumirati na jednom slajdu ili zapisati na ploču) te pokrenuti raspravu o svrhama i iskoristivosti ovog projekta. S učenicima možemo raspraviti o istraživačkim pitanjima koje nam ovaj projekt otvara (ali i onima koje nam ne otvara!), o osmišljavanju sličnih projekata, o osobnim stavovima učenika o prikazanim temama i sl. Naravno za očekivati je i da će učenici imati mnogo pitanja, ali treba paziti da se odgovaranjem na njih ne udaljimo previše od teme. Opći je cilj ovakvog sata prikazati čime se danas mogu baviti istraživači latinskog, kako možemo koristiti računala da bismo istraživali latinski jezik i književnost te usputno prikazati veliku količinu hrvatske novolatinske književnosti koju imamo sačuvanu i lako dostupnu.

Osim predavanja (ili, dapače, uz njega) učenicima se rezultati ovog projekta mogu predstaviti kroz praktičan rad. S obzirom na to da naši rezultati obuhvaćaju samo tekstove hrvatskih latinista, ovakav je način primjene projekta u nastavi uglavnom namijenjen samo za učenike

četvrtih razreda klasičnih gimnazija. Praktičan rad bi podrazumijevao neki problemski zadatak na nekom od tekstova iz *CroALa*-e spojenih u *LiLa*-u koji učenici sami, uz prethodno objašnjenje, rješavaju. Učenici bi, pristupajući raznim resursima u *LiLa*-i putem lemā iz korpusa, mogli odgovarati na mnoga lingvistička pitanja. Primjerice, učenici bi mogli istražiti indoeuropske etimologije pojedinih riječi ili tražiti riječi istog porijekla, tražiti izvedenice pojedinih riječi ili određene tvorbene afikse i sl. Za ovakvu je vrstu zadataka posebno pogodno *LiLa*-ino sučelje *LodLive*[76] koje pruža mogućnost pristupačnog grafičkog kretanja mrežom resursa. Još jednostavniji pristup bio bi jednostavno zadati učenicima da tekst obrade koristeći se dostupnim digitalnim gramatičkim pomagalima iz *LiLa*-e.

Za ostvarivanje praktične nastave, mogu se koristiti i samo određeni segmenti ovog rada. Ovakav pristup omogućio bi rad i s učenicima iz općegimnazijskih programa učenja latinskog te s učenicima iz prvih, drugih i trećih razreda klasičnih gimnazija. U vrlo pojednostavljenom obliku moglo bi se iskoristi i u osnovnoškolskom učenju latinskog, za učenike sedmih ili osmih razreda. Primjerice, učenici bi mogli koristi *LiLa*-u za obradu tekstova iz udžbenika, umjesto novolatinskih tekstova, kako bi istražili određene jezične pojave ili jednostavno bolje i lakše razumjeli tekst. Za takav bi pristup koristan bio *LiLa*-in *Text Linker*[77], alat koji je još uvijek u razvoju, a koji služi za povezivanje teksta koji korisnik unese s lemama iz *LiLa*-e. Uz to, jedan bi nastavni zadatak mogla predstavljati i usporedba *Text Linkera* s drugim alatima za analizu latinskog teksta, poput *Collatinusa*[78].

Osim toga, neki od postupaka pripreme i lematizacije/tagiranja uzorka bi također mogli biti korišteni u nastavi. Primjerice, učenicima bi se mogao zadati zadatak da označe odabrane rečenice iz tekstova i zatim procjene uspješnost nekih od ponuđenih modela, izražavajući točnost svakog modela u postotcima. Treba paziti da se učenicima zadatak vrlo jasno objasni i demonstrira na primjeru te treba pažljivo birati rečenice koje će učenici moći svladati.

Primjena ovog projekta u nastavi omogućava postizanje ne samo ishoda za predmet *Latinski jezik*, već i ishoda za međupredmetnu temu *Uporaba informacijske i komunikacijske tehnologije* (2019.), a otvara se i mogućnost korelacije s ostalim nastavnim predmetima, primjerice *Informatikom*, *Hrvatskim jezikom*, *Matematikom* i sl.

[76] Vidi https://lila-erc.eu/lodlive/ (pristupljeno 23. travnja 2021.).
[77] Vidi http://lila-erc.eu:8080/LiLaTextLinker/ (pristupljeno 24. travnja 2021.). Napomena: radi se o beta verziji.
[78] Vidi https://outils.biblissima.fr/en/collatinus-web/ (pristupljeno 3. svibnja 2021).

7. Zaključak

U radu je prikazana prva od dvije faze uključivanja *CroALa*-e u *LiLa*-u, u kojoj je trebalo pripremiti *CroALa*-u za povezivanje. Budući da proces pripreme još uvijek nije sasvim završen, u radu su prikazani obavljeni postupci, zajedno s izazovima i problemima koje je trebalo riješiti, i najavljeni budući radovi na projektu.

Prije uključivanja *CroALa*-u je trebalo lematizirati i obaviti gramatičko, odnosno POS tagiranje. Da bismo to mogli učiniti, trebalo je odabrati i pripremiti reprezentativne tekstove *CroALa*-e na kojima ćemo testirati razne modele za automatsku lematizaciju i tagiranje, od kojih će najuspješniji poslužiti za obradu čitavog korpusa.

Prikazan je način odabira reprezentativnog uzorka tekstova *CroALa*-e i ukratko opisan način dobivanja i obrade za to potrebnih metapodataka. Priprema tekstova za obradu uključivala je dohvaćanje tekstova iz XML datoteka i njihove tokenizacije. U opisu procesa dohvaćanja tekstova upozorili smo na problem odabira „čistog“ teksta i na raznoliku strukturu XML datoteka koja može utjecati na rezultate upita. Kod opisa tokenizacije adresiran je problem definicije riječ, odnosno pojavnice i kraja rečenice.

Kao drugi korak u prvoj fazi projekta, opisan je postupak lematizacije i tagiranja uzorka te je prikazan onaj dio evaluacije lematizatora i tagera koji je do sad odrađen. Opisani su modeli i algoritmi za lematizaciju i tagiranje (njih ukupno 37), korpusi na kojima su trenirani te skupine tagova (engl. *tagset*) koji svaki od tagera koristi. Ukratko su prikazani rezultati lematizacije i tagiranja. Sljedeći je korak bio evaluacija svakog lematizatora i tagera te odabir najuspješnijeg od njih. Za te potrebe, izgradili smo tzv. zlatni standard, odnosno ručno smo označili dio uzorka kako bismo mogli s njime usporediti rezultate svakog pojedinog alata i tako ocijeniti njegovu uspješnost. Pokazalo se da postavljanje standarda za lematizaciju i tagiranje nije trivijalan zadatak te smo, što je detaljnije moguće, pokušali opisati temeljne principe lematizacije i tagiranja. Budući da je za tagiranje korišten skup tagova *UD tagset*, prikazana je i uporaba svakog taga, kao i teorijski te praktični problemi koji su se prilikom tagiranja pojavili. Adresiran je i problem usporedbe modela koji koriste različite skupove tagova. Općenito govoreći, velik broj različitih modela i različitih skupova tagova može se pokazati korisnim jer omogućava odabir onog modela ili skupa koji najviše odgovara potrebama korisnika – ako trebamo, primjerice, označiti srednjovjekovni tekst, u kojem velik broj riječi otpada na brojeve, model treniran na srednjovjekovnim tekstovima koji koristi skup tagova u kojem se razlikuju različite vrste brojeva može nam biti iznimno koristan. No ono što često predstavlja najveći problem pri

korištenju nekog modela je nepostojeća ili teško dostupna dokumentacija o načinima korištenja skupa tagova ili nastanku modela. Upravo iz tog razloga, nemali dio ovog rada posvećen je upravo definiranju i opisivanju principa lematizacije te načina na koji su pojedini tagovi korišteni.

U završnom dijelu rada opisani su budući radovi na projektu, koji se tiču spajanja lematiziranih tekstova *CroALa*-e s *LiLa*-om. U sklopu toga u najkraćim je crtama objašnjena arhitektura i način funkcioniranja *LiLa*-e. Posljednje se poglavlje bavi prijedlozima za korištenje ovog projekta u nastavi.

Kao što je nekoliko puta o ovom radu naznačeno, mnoga su od prikazanih rješenja problema (pogotovo ona koja se tiču lematizacije i tagiranja uzorka) podložna promjenama i korekcijama. Naime, budući da proces evaluacije nije dovršen, moguće je da će se tijekom daljnjih radova javiti novi problemi ili spoznaje, koje će nam ukazati na potrebu za primjenom drugačijeg pristupa.

Na posljetku ostaje izraziti nade da će rezultati ovog rada s jedne strane poslužiti u razvijanju digitalnih alata za obradu latinskog jezika, a s druge strane koristiti svim istraživačima hrvatske novolatinske književnosti.

8. Literatura

Allen, J. H., & Greenough, J. B. (1903.). *New Latin Grammar.* Boston: Ginn & Company.

Bagić Babac, M., & Kušek, M. (2011.). *Informacija, logika i jezici: Skripta: jezici za označavanje sadržaja.* Zagreb: Fakultet elektrotehnike i računarstva. Dohvaćeno iz https://www.fer.unizg.hr/_download/repository/ILJ-2011-12-XML_v1.4.pdf

Bekavac, B. (2001.). Primjena računalnojezikoslovnih alata na hrvatske korpuse. Dohvaćeno iz http://darhiv.ffzg.unizg.hr/id/eprint/2360

Bender, E. M. (2013.). *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax.* Morgan & Claypool. doi:https://doi.org/10.2200/S00493ED1V01Y201303HLT020

Bennet, C. E. (1908.). *A Latin Grammar.* Boston: Allyn and Bacon.

Bird, S., Loper, E., & Klein, E. (2009.). *Natural Language Processing with Python.* O'Reilly Media Inc. Dohvaćeno iz https://www.nltk.org/book_1ed/

Bon, B. (2011.). OMNIA : outils et méthodes numériques pourl'interrogation et l'analyse des textes médiolatins (3). *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA*. doi:10.4000/cem.12015

Cecchini, F. M., Korkiakangas, T., & Passarotti, M. (2020.). A new Latin treebank for universal dependencies: Charters between ancient Latin and romance languages. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings* (str. 933.-942.). Marseille: European Language Resources Association (ELRA). doi:10.5281/zenodo.3830352

Cecchini, F. M., Passarotti, M., Marongiu, P., & Zeman, D. (2019.). Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (str. 27.-36.). Brussels, Belgium: Association for Computational Linguistics. doi:10.5281/zenodo.2552202

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020.). *Linguistic Linked Data. Representation, Generation and Applications.* Springer. doi:10.1007/978-3-030-30225-2

Du Cange, C. d., Henschel, L. G., Carpenter, P., Favre, L., & Adelung, J. C. (1883.-1887.). *Glossarium mediæ et infimæ latinitatis.* Niort: L. Favre. Dohvaćeno iz logeion.uchicago.edu

Eger, S., Gleim, R., & Mehler, A. (2016.). Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016* (str. 1507.-1513.). Portorož, Slovenija: European Language Resources Association. Dohvaćeno iz https://www.aclweb.org/anthology/L16-1239

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., . . . Zampedri, F. (2019.). Nunc est aestimandum. Towards an evaluation of the Latin WordNet. *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)* (str. 13.-15.). Torino: Accademia University Press. doi:10.5281/zenodo.3518774

Gamba, F. (2020.). *Including a New Textual Resource into the LiLa Knowledge Base: Lemmatization, PoS Tagging and Linking of Querolus.* Università di Pavia, Dipartimento di Studi Umanistici. (neobjavljeno).

Gleim, R., Eger, S., Mehler, A., Uslu, T., Hemati, W., Lücking, A., . . . Hoenen, A. (2019.). A practitioner's view: a survey and comparison of lemmatization and morphological tagging in German and Latin. *Journal of Language Modelling*, str. 1.-52. doi:10.15398/jlm.v7i1.205

Gortan, V., Gorski, O., & Pauš, P. (2005.). *Latinska gramatika* (12. izd.). Zagreb: Školska knjiga.

Haug, D. T., & Jøhndal, M. L. (2008.). Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (str. 27.-34.). European Language Resources Association. Dohvaćeno iz http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf

Johnson, K. P., Burns, P., Stewart, J., Cook, T., Besnier, C., & Mattingly, W. J. (2014.-2021.). CLTK: The Classical Language Toolkit. Preuzeto 7.. travnja 2021. iz https://github.com/cltk/cltk

Jovanović, N. (28. lipnja 2020.). *nevenjovanovic/croatiae-auctores-latini-textus*. Dohvaćeno iz GitHub: https://github.com/nevenjovanovic/croatiae-auctores-latini-textus

Jovanović, N., Haskell, Y., Lonza, N., Lučin, B., Marinova, E., Novaković, D., & Tunberg, T. O. (27. siječnja 2014.). *CroALa*. Preuzeto 9. siječnja 2021. iz http://croala.ffzg.unizg.hr/intro/

Jurafsky, D., & Martin, J. H. (2020.). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3. izd.). Preuzeto 7.. travnja 2021. iz https://web.stanford.edu/~jurafsky/slp3/

Kurikulum međupredmetne teme Uporaba informacijske i komunikacijske. (2019.). Zagreb: Ministarstvo znanosti i obrazovanja.

Kurikulum nastavnog predmeta Hrvatski jezik za osnovne škole i gimnazije. (2019.). Zagreb: Ministarstvo znanosti i obrazovanja. Dohvaćeno iz https://skolazazivot.hr/wp-content/uploads/2020/06/HR-OSiGM_kurikulum.pdf

Kurikulum nastavnog predmeta Latinski jezik za osnovne škole i gimnazije. (2019.). Zagreb: Ministarstvo znanosti i obrazovanja. Dohvaćeno iz https://skolazazivot.hr/wp-content/uploads/2020/06/LJ_kurikulum.pdf

Lewis, C. T., & Short, C. (1879.). *A Latin Dictionary.* Oxford: Claredon Press. Dohvaćeno iz logeion.uchicago.edu

Mambrini, F., & Passarotti, M. (2019.). Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. *LAW 2019 - 13th Linguistic Annotation Workshop, Proceedings of the Workshop* (str. 71.-80.). Firenca, Italija: Association for Computational Linguistics. doi:10.18653/v1/w19-4009

Manning, C., & Schütze, H. (1999.). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT.

McEnery, T., & Hardie, A. (2012.). *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001.). *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

Menge, H., Burkard, T., & Schauer, M. (2004.). *Lehrbuch der lateinischen Syntax und Semantik.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Müller, T., Schmid, H., & Schütze, H. (2013.). Efficient Higher-Order CRFs for Morphological Tagging. *Proceedings of the 2013 Conference on Empirical Methods in Natural*

Language Processing (str. 322.-332.). Seattle, Washington, SAD: Association for Computational Linguistics. Dohvaćeno iz https://www.aclweb.org/anthology/D13-1032/

Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., & Pham, S. B. (2015.). RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (str. 17.-20.). Gothenburg, Sweden: Association for Computational Linguistics. doi:10.3115/v1/E14-2005

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Antonsen, L., Aplonova, K., . . . Zhu, H. (2018.). Universal Dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Dohvaćeno iz http://hdl.handle.net/11234/1-2895

Novak, V., & Skok, P. (1952.). *Supetarski kartular.* Zagreb: JAZU.

Passaroti, M. C., Cecchini, F. M., Franzini, G., Litta, E., Mambrini, F., & Ruffolo, P. (2019.). The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. *CEUR Workshop Proceedings* (str. 27.-36.). Leipzig: Sveučilište u Leipzigu. doi:10.5281/zenodo.3358550

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., . . . Sprugnoli, R. (2020.). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, str. 177.-212. doi:10.4454/ssl.v58i1.277

Passarotti, M., Ruffolo, P., Cecchini, F. M., Litta, E., & Budassi, M. (2018.). LEMLAT 3.0. doi: https://doi.org/10.5281/zenodo.1492133

Pinkster, H. (2015.). *The Oxford Latin Syntax* (Svez. I. The Simple Clause). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199283613.001.000

Schmid, H. (1994.). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester, UK. Dohvaćeno iz https://www.cis.lmu.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Spevak, O. (2014.). *The Noun Phrase in Classical Latin Prose.* Leiden, Nizozemska: Brill.

Springmann, U., Schmid, H., & Najock, D. (2016.). LatMor: A Latin finite-state morphology encoding vowel quantity. *Open Linguistics, 2.*, str. 386.-392. doi:10.1515/opli-2016-0019

Straka, M. (2018.). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (str. 197.–207.). Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/K18-2020

TEI Consortium. (2021.). *TEI P5: Guidelines for electronic text encoding and interchange* (4.2.1. izd.). TEI Consortium. Preuzeto 3. ožujka 2021. iz http://www.tei-c.org/Guidelines/P5/

Tsuruoka, Y., Miyao, Y., & Kazama, J. (2011.). Learning with lookahead: Can history-based models rival globally optimized models? *CoNLL 2011 - Fifteenth Conference on Computational Natural Language Learning, Proceedings of the Conference*, str. 238.-246. Dohvaćeno iz https://www.aclweb.org/anthology/W11-0328.pdf

vor der Brück, T., Eger, S., & Mehler, A. (2015.). Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization models. *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (str. 105.-113.). Peking. Kina: Association for Computational Linguistics. doi:10.18653/v1/W15-3716

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (15.. 3. 2016.). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. doi:https://doi.org/10.1038/sdata.2016.18

Wissler, L., Almashraee, M., Dagmar, M., & Paschke, A. (2014.). The Gold Standard in Corpus Annotation. *Proceedings of the 5th IEEE Germany Students Conference 2014.* doi:10.13140/2.1.4316.3523

Zeman, D. (2018.). *The World of Tokens, Trees and Tags.* Prag: Ústav formální a aplikované lingvistiky (ÚFAL). Dohvaćeno iz https://ufal.mff.cuni.cz/books/2018-zeman

Žepić, M. (2000.). *Latinsko-hrvatski rječnik* (13. izd.). Zagreb: Školska knjiga.

9. Popis tablica i slika

Tablice

Slike

Slika 1. preuzeta je s https://lila-erc.eu/wp-content/uploads/2018/10/LilaSimplified.png (22. travnja 2021.)

10. Prilozi

Prilog 1. Popis datoteka korištenih za dobivanje uzorka korpusa

Sve su datoteke preuzete iz repozitorija 12. ožujka 2020., u trenutku u kojem je najaktualnija verzija repozitorija bila ona od 1. prosinca 2019., koja je dostupna putem poveznice https://git.io/JO7JP.

Popis korištenih datoteka:

1. aa-vv-supetarski.xml
2. andreis-f-epist-nadasd.xml
3. benesa-d_epigr03_croala5095251.croala-lat1.xml
4. boskovic-r-ecl.xml
5. bunic-j-de-r.xml
6. gradic-s-oratio.xml
7. kunic-r-hymnus-cererem.xml
8. marulus-m-carmina008.xml
9. milasin-f-viator.xml
10. modr-n-navic.xml
11. sisgor-g-odae.xml
12. sisgor-g-prosopopeya.xml
13. tubero-comm-rhac.xml

Prilog 2. *XQuery* skripta za dohvaćanje tekstova

```
(:returns raw texts from DB:)
declare namespace tei = "http://www.tei-c.org/ns/1.0";
for $results in //*:text
return $results//text()
```

Prilog 3. Popis kratica korištenih u tokenizaciji

I.	XL.	Jo.
II.	C.	P.
III.	CC.	Abb.
IIII.	CCC.	D.
V.	CCCX.	PP.
VI.	M.	ee.
VII.	DC.	Rempub.
VIII.	LXVIII.	Mich.
IX.	etc.	eg.
X.	Lyc.	Cal.
XI.	Tyt.	Xav.
XII.	PONT.	Carmelit.
XIII.	MAX.	Excalc.
XIIII.	soc.	EMINENTISS.
XV.	Soc.	LXVII.
XVI.	s.	EE.
XVII.	S.	XXXVII.
XVIII.	R.	XLVII.
IXX.	E.	XVIIII.
XX.	CAR.	XXXI.
XXV.	DD.	

Prilog 4. *Python* skripta za rastavljanje rečenica

```
#!/usr/bin/env python3

import sys

from nltk.tokenize import sent_tokenize

with open(sys.argv[1], 'r', encoding="utf-8") as text:

     text = text.read().replace('\n', ' ')

     sttext = sent_tokenize(text)

     for sent in sttext:

          print (sent.encode("utf-8"), sep="")
```

Napomena: ova skripta dijeli rečenice na temelju interpunkcija i to bez uzimanja kratica u obzir, pa rezultate treba naknadno provjeriti i po potrebi prepraviti.

Sažetak

Uključivanje korpusa latinskih tekstova *CroALa* u bazu znanja latinskog jezika *LiLa*

Unatoč postojanju velikog broja raznih digitalnih alata i resursa za obradu i istraživanje latinskog jezika, njihova je iskoristivost i interoperabilnost skromna zbog značajnih razlika u formatima kojima se koriste. Upravo je to početna premisa projekta *LiLa: Linking Latin*, koji provodi milansko sveučilište *Università Cattolica del Sacro Cuore* i čiji je cilj povezati obilje jezičnih resursa i alata razvijenih za latinski i učiniti ih međusobno iskoristivima. Ovaj rad prikazuje spajanje korpusa tekstova hrvatskih latinista *CroALa* kao tekstualnog resursa u *LiLa*-u. Projekt je zamišljen u dvije faze: lematizacija *CroALa*-e i spajanje tekstova u *LiLa*-u. Budući da projekt nije dovršen, prikazuju se samo rezultati iz dijela prve faze projekta. Najprije se opisuje odabir reprezentativnog uzorka tekstova *CroALa*-e na kojima su testirani automatski lematizatori i gramatički tageri, a zatim slijedi opis dohvaćanja tekstova iz XML datoteka i njihova tokenizacija. Potom se opisuju alati za automatsku lematizaciju i tagiranje koji su korišteni te se prikazuju neobrađeni rezultati i (preliminarne) metode komparativne evaluacije uspješnosti pojedinih alata. U završnom dijelu rada najavljuje sam proces povezivanja *CroALa*-e s *LiLa*-om, a zadnje se poglavlje bavi primjenom ovog projekta u nastavi latinskog jezika.

Ključne riječi: *CroALa*, *LiLa: Linking Latin*, korpus, hrvatski latinisti, lematizacija latinskog, gramatičko tagiranje latinskog, jezikoslovni povezani otvoreni podatci

Summary

Including the *CroALa* corpus of Latin texts into the *LiLa* knowledge base

Despite a large number of various digital tools and resources for processing and researching Latin, their usability and interoperability remains modest because they use substantially different formats. This is the main premise of the *LiLa: Linking Latin* project, developed by the *Catholic University of the Sacred Heart* in Milan. Its main goal is to link these various linguistic resources and tools developed for Latin and make them interoperable. This paper presents the inclusion of the *CroALa* corpus of texts by Croatian Latin writers into *LiLa* as a textual resource. The project is conceived in two phases: lemmatizing *CroALa* and linking the texts to *LiLa*. Because the project is still in progress, only the results of a part of the first phase are presented. Firstly, the process of choosing a representative sample of texts from *CroALa* is presented. The texts are selected for testing automatic lemmatizers and POS taggers. After that, the querying of the texts from their XML files and their tokenization are described. Then the tools used for lemmatization and tagging are addressed, together with the raw results and (preliminary) methods for performance evaluation. In the closing part of the paper, the actual linking of *CroALa* to *LiLa* is announced. The last chapter deals with the application of this project in teaching Latin in secondary schools.

Key words: *CroALa*, *LiLa: Linking Latin*, corpus, Croatian Latinists, lemmatizing Latin, POS tagging Latin, Linked Linguistic Open Data (LLOD)