Izazovi morfosintaktičkog označavanja na primjerima španjolskog i hrvatskog jezika

Kozolić, Klara

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2020./ 2021.

Klara Kozolić

Izazovi morfosintaktičkog označavanja na primjerima španjolskog i hrvatskog jezika

Završni rad

Mentor: dr.sc. Nives Mikelić Preradović, red. prof.

Zagreb, lipanj 2021.

Izjava o akademskoj čestitosti

Izjavljujem 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.

Ovaj rad posvećujem svima koji su mi bili velika podrška tijekom preddiplomskog studija: roditeljima Željku i Renati, teti Željki, djedu Marku, Josipu, najboljim prijateljicama Celini, Ivi i Mirti, ekipi ispred knjižnice i svim ostalim prijateljima. Također sam neizmjerno zahvalna mentorici prof. Nives Mikelić Preradović na svakom prijedlogu i na strpljenju.

Sadržaj

1. Uvod

Napretkom tehnologije dolazi do njezinog korištenja u mnogim disciplinama. Lingvistika nije iznimka jer se njenim temeljnim spoznajama pridružuje informacijska tehnologija, a kao rezultat toga razvijaju se jezične tehnologije (Tadić, 2003). Postoje različite primjene jezičnih tehnologija koje olakšavaju razumijevanje jezika i ljudsku komunikaciju. Korištenjem jezičnih tehnologija, odnosno obradom prirodnog pisanog i usmenog jezika, razvilo se područje kojemu je jedan od naziva računalna obrada prirodnog jezika (Pascual, 2012). Neki od primjera njezine tekstualne uporabe su pronalaženje dokumenata u bazama podataka, izvlačenje informacija iz članaka i poruka, prevođenje dokumenata i sažimanje tekstova (Allen, 1995). Jedan od nužnih dijelova jezičnih tehnologija su korpusi čija je analiza tehnološkim napretkom znatno uznapredovala (McEnery i sur., 2019). Jedna od razina na kojoj se označavaju je *Part-of-speech* (POS) označavanje ili morfosintaktički opis (MSD). POS označavanje je „proces dodavanja odgovarajuće gramatičke kategorije riječi u tekstu koji se označava“ (McEnery i Wilson, 2001, str. 135), a MSD označavanje uz gramatičku kategoriju riječi dodaje i neka druga obilježja (npr. rod, broj, padež u morfološki složenim jezicima) (Bekavac, 2002). Zbog različitih čimbenika, dodavanje POS ili MSD oznaka riječima ne može se uvijek obavljati sa stopostotnom preciznošću. U ovom radu će se prikazati do kojih sve izazova dolazi pri procesu morfosintaktičkog označavanja. Budući da su svjetski jezici raznoliki i nemaju svi jednak stupanj flektivnosti, sintaksu ni pravopis, POS i MSD označavanje uvelike ovisi o tim razlikama. U ovom radu će se dati općeniti pregled izazova u morfosintaktičkom označavanju, a kako bi primjeri bili konkretni odabran je španjolski jezik kao manje flektivni jezik i hrvatski jezik kao visoko flektivni jezik. Iz tih će se jezika uzeti primjeri rečenica iz književnih djela, znanstvenih članaka, novinskih članaka i mrežnih rječnika žargona s različitim potencijalnim izazovima za označivače. Cilj rada je usporediti kojim se oznakama koriste pojedini označivači pri označavanju izazovnih jezičnih pojavnosti i analizirati eventualne pogreške kako bi se vidjelo odgovaraju li pregledu izazova u teoriji. POS označivači kojima će se rečenice testirati za španjolski jezik su Stanfordski POS označivač te TreeTagger označivač, a dobiveni rezultati će se usporediti. Za hrvatski jezik testiranje će se provesti na web sučelju ReLDIanno MSD označivača te će se rezultati analizirati i usporediti s onima za španjolski jezik.

2. POS ili morfosintaktičko označavanje

POS je kratica za engleski izraz *part-of-speech* koji se u modernoj lingvistici koristi kao ekvivalent za vrstu riječi (Lehmann, 2013). U obradi prirodnog jezika POS ili vrste riječi imaju vrlo važnu ulogu jer daju detaljan uvid u karakteristike riječi i njezine okoline, a za to koriste POS i MSD oznake koje se dodaju svakoj riječi unutar korpusa (Jurafsky i Martin, 2009). Danas su korpusi dostupni za mnoge svjetske jezike i sadrže stotine milijuna riječi i zbog toga je potrebno osloniti se na automatsko označavanje jer je ono ručno dugotrajan i kompliciran proces (Petkevič, 2014). Automatsko se označavanje naziva POS označavanje i sastoji se od dodavanja odgovarajuće gramatičke kategorije tj. POS oznake riječi u određenom kontekstu, a kad se riječima dodaju MSD oznake se naziva i morfosintaktičko označavanje (Agić i sur., 2013). Budući da se oznake dodaju i interpunkcijskim znakovima, nužno je da su oni odvojeni od riječi pa se prije POS ili MSD označavanja provodi tokenizacija, odnosno, „segmentacija teksta u riječi“ (Jurafsky i Martin, 2009, str. 47). Dodavanje POS oznake se sastoji od određivanja vrste riječi, dok je morfosintaktički opis tj. MSD (engl. *morphosyntactic description*) na višoj razini jer uključuje i gramatičke kategorije riječi poput roda, broja, padeža itd. (Bekavac, 2002). Za označavanje je nužno imati skupove POS ili MSD oznaka (*engl. tagsets*) (Kumawat i Jain, 2015). Skupovi POS i MSD oznaka postoje za mnoge svjetske jezike. Petrov i sur. (2011) predlažu univerzalan skup oznaka za koje smatraju da su, bar u nekom obliku, zajedničke velikom broju jezika, a radi se o sljedećim oznakama:

- NOUN (imenice),
- VERB (glagoli)
- ADJ (pridjevi)
- ADV (prilozi)
- PRON (zamjenice)
- DET (determinanti i članovi)
- ADP (adpozicije)
- NUM (brojevi)
- CONJ (veznici)
- PRT (čestice)
- '.' (interpunkcijski znakovi)
- X (ostale kategorije poput kratica i stranih riječi).

To je jedan od radova na koji se poziva projekt *Universal Dependencies* (*UD*) koji nastoji razviti međujezično označavanje u banki stabala, s ciljem da bude dosljedno za mnogo jezika (Nivre i sur., 2016), a nudi univerzalne oznake prikazane u Tablici 1:

Promjenjive riječi	**Nepromjenjive riječi**	**Drugo**
ADJ – pridjev	ADP – adpozicija	PUNCT – interpunkcija
ADV – prilog	AUX – pomoćne riječi	SYM – simbol
INTJ – usklik	CCONJ – veznici nezavisnosloženih rečenica	X – ostalo
NOUN – imenica	DET – determinant	
PROPN – vlastita imenica	NUM – broj	
VERB – glagol	PART – čestica	
	PRON – zamjenica	
	SCONJ – veznici zavisnosloženih rečenica	

Tablica1. Univerzalni skup oznaka u projektu *Universal Dependencies*
https://universaldependencies.org/u/pos/

U prvom i drugom stupcu su prikazane vrste riječi koje bi ugrubo odgovarale promjenjivim i nepromjenjivim riječima u hrvatskom, a u trećem stupcu su prikazane ostale oznake.

Kod MSD oznaka također postoji težnja za standardizacijom, a primjer toga je projekt *The MULTEXT-East* koji uključuje devet jezika od kojih je pet slavenskih, a to su bugarski, češki, hrvatski, slovenski i srpski (Erjavec i sur., 2003). Slavenski jezici su visoko flektivni pa se zato u obzir moraju uzeti puno detaljnije oznake. Erjavec i sur. (2003) zato predlažu morfosintaktičke specifikacije za pet slavenskih jezika i objašnjavaju općenitu strukturu MSD oznake, na primjer, NCMPG koja počinje s POS oznakom *N*, a atributi koji slijede su *Type:common, Gender:masculine, Number:plural i Case:genitiv*. To znači da se radi o imenici koja je po vrsti opća, rod joj je muški, broj množina, a padež genitiv. Na web stranici Multext-East Resources Version 6 (n.d.) je opisana posljednja verzija ovog projekta. To je verzija 6 koja je većim dijelom nastala u sklopu CLARIN istraživačke infrastrukture 2020. godine. Njome su nadograđene morfosintaktičke specifikacije za makedonski jezik. Nadalje, dodane su

specifikacije za albanski i za srpsko-hrvatski. Specifikacijama za srpsko-hrvatski se nastoji pokriti hrvatski, srpski, bosanski i crnogorski jezik. Također se dodaju „Damaskini“ specifikacije koje su razvijene za dijakronijski korpus balkanskih i slavenskih tekstova u razdoblju od 16. do 19. stoljeća.

Unatoč tome što još nema publikacija za verziju 6, opis sintaktičkih resursa se može pronaći u radu iz 2017. koji opisuje MULTEXT-East specifikacije. Erjavec (2017) navodi tri glavna sintaktička resursa korištena u projektu, a to su ranije opisane morfosintaktičke specifikacije, morfosintaktički leksikoni i označeni korpus „1984“. Erjavec (2017) opisuje kako se u morfosintaktičkim leksikonima nalaze potpune fleksijske paradigme odabranih lema ili oblici riječi prethodno provjereni u korpusu te kako svaka natuknica daje informacije o obliku riječi, njenoj lemi i MSD oznaci. Nadalje, dodaje kako se u morfosintaktički obilježenom korpusu nalaze riječi kojima su pridodane razriješene MSD oznake i leme. Trenutno MULTEXT-East projekt pokriva engleski, rumunjski, poljski, češki, slovački, slovenski, ruski, ukrajinski, makedonski, bugarski, latvijski, litavski, albanski, perzijski, estonski, mađarski i čečenski (Multext-East, n.d.).

Danas su razvijeni jezični alati za mnoge svjetske jezike uz pomoć kojih je, uz lematizaciju, sintaktičko parsiranje i prepoznavanje imenovanih entiteta (engl. *NER – named entry recognition*), omogućeno i POS ili MSD označavanje, a neki od njih su prikazani u Tablici 2:

Jezik	**Jezični alat**
Afrikaans	Afrikaans TnT-Tagger – POS označivač
Asamski	Assamese POS Tagger – POS označivač
Bugarski	CLaRK - odvajanje rečenica, POS označavanje, lematizacija, sintaktičko parsiranje
Češki	HMM tagger - MSD označivač
Engleski	• CLAWS – POS označivač • Stanfordski Dependency Parser – POS označavanje, sintaktičko parsiranje
Estonski	• EstNLTK – MSD označivač, NER • Vabamorf open source morphology tagger for Estonian – lematizacija, POS i MSD označivač
Finski	FinTag – POS označivač, lematizacija, NER
Grčki	ILSP Feature-based multi-tiered POS Tagger – POS označivač

Islandski	IceNLP Natural Language Processing toolkit – POS označivač, lematizacija, sintaktičko parsiranje
Latvijski	NLP-PIPE – MSD označivač, sintaktičko parsiranje, NER
Mađarski	hunpos – POS označivač
Malteški	MLSS Tagger Web Service – POS označivač
Nizozemski	• Frog – POS i MSD označivač, lematizator • Tadpole – POS i MSD označivač, lematizator, sintaktičko parsiranje
Norveški	The Oslo-Bergen tagger – MSD označivač, sintaktičko parsiranje
Njemački	• OpenNLP Part-of-Speech Tagger (German) – POS označivač • Stuttgart Dependency Parser – POS označivač, sintaktičko parsiranje
Poljski	• MorphoDiTa-based tagger for Polish language – MSD označivač • WCRFT (Wrocław CRF Tagger) – MSD označivač
Portugalski	• LX-Tagger – MSD označivač • OpenNLP Part-of-Speech Tagger (Portuguese) – POS označivač
Talijanski	Freeling – POS označivač, lematizator
Višejezični alati	• ReLDIanno (hrvatski, slovenski i srpski) – MSD označivač, NER, lematizator • Turku-neural-parser-pipeline (više od 50 jezika) – segmentacija, MSD označivač, sintaktičko parsiranje, lematizator • NCHLT Tagger (afrikaans, engleski, ndebele, xhosa, zulu, sesotho sa leboa, setswana, sesotho, siswati, tshivenda, xitsonga) – POS označivač

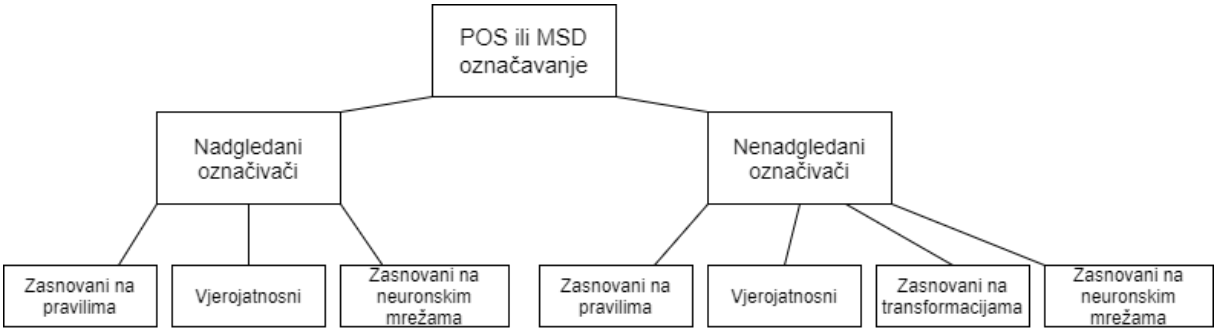
Tablica2. Primjeri jezičnih alata *https://www.clarin.eu/resource-families/tools-part-speech-tagging-and-lemmatization*

2.1. POS i MSD označivači

Jezični alati koji se koriste za dodavanje POS oznaka zovu se označivači vrsta riječi (engl. *POS taggers*) dok se oni koji dodaju MSD oznake nazivaju morfosintaktički označivači (engl. *MSD taggers*) (Tadić, 2003, str. 32). Radi se o onim dijelovima softvera koji prepoznaju riječi u tekstu te im automatski dodaju oznaku (Kumawat i Jain, 2015, str. 32). Budući da su gramatičke kategorije vrlo predvidljive, rad označivača može biti vrlo točan i detaljan te ih se smatra najvjerodostojnijim jezičnim alatom (Bekavac, 2002). Općenito, kako bi označivači označili pojavnice u korpusu, konzultiraju se s leksikonom u kojem traže riječ i ako je pronađu daju joj popis mogućih POS ili MSD oznaka (McEnery i Wilson, 2001). U korpusnoj je lingvistici

pojam leksikon ekvivalent rječničkoj bazi podataka koja „podrazumijeva pohranu leksičke građe u strojno čitljivom obliku“ (Bekavac, 2002, str. 178). Posljednji je korak u ovom općenitom prikazu razrješenje višestrukih opisa pridruženih riječima u korpusu (Tadić, 2003).

POS i MSD označivači se „prema stupnju autonomije označavanja i uporabe već obilježenog korpusa mogu podijeliti na nadgledane koji na temelju korpusa izrađuju alate koji će pomoći u samom označavanju i nenadgledane“ (Bekavac, 2002, str. 177). Nenadgledani označivači za automatsko pronalaženje skupova oznaka ili transformacijskih pravila koriste napredne računalne metode (Kumawat i Jain, 2015). Daljnja podjela je „prema načinu rada prema kojem se dijele na označivače zasnovane na pravilima (engl. *rule based*) i vjerojatnosne označivače (engl. *probabilistic*)“ (Bekavac, 2002, str. 177). Kumawat i Jain (2015) dodaju i treću stavku ovoj podjeli, a to su hibridni označivači (engl. *hybrid*). Označivači zasnovani na pravilima se sastoje od skupa pisanih pravila za razrješavanje *(*engl. *disambiguation*) višestrukih morfosintaktičkih opisa, dok vjerojatnosni označivači razrješavaju višestruke opise na temelju vjerojatnosti da se neka riječ pojavi s nekom oznakom u određenom kontekstu (Jurafsky i Martin, 2009). Hibridni označivači prvo koriste vjerojatnosni račun i statistiku, a potom im se pridodaju pisana jezična pravila (Kumawat i Jain, 2015). Vjerojatnosni pristup je najzastupljeniji kod označivača, no, bez obzira na to se najčešće koristi u kombinaciji s pristupom zasnovanim na pravilima, čime se doseže visoka točnost (Bekavac, 2002). Razlog manjeg korištenja označivača zasnovanih na pravilima bi mogao biti velika količina rada i lingvističko znanje koji su na njima potrebni (Voutilainen, 1995). Posljednjih su se godina također razvili označivači zasnovani na dubokom učenju, odnosno, na povratnim neuronskim mrežama (engl. *recurrent neural networks*) (Bhonet i sur., 2018). Konkretnije, radi se o dvosmjernoj mreži s dugom kratkoročnom memorijom tj. BiLSTM (engl. *bidirectional long shorterm memory*) koja prolazi kroz nizove znakova dva puta, od početka do kraja i obrnuto (Plank i sur., 2016). Na Grafikonu 1 može se vidjeti podjela POS i MSD označivača prema autorima Kumawat i Jain (2015):



Grafikon1. Prikaz podjele POS i MSD označivača *https://doi.org/10.5120/20752-3148*

Grafikon 1 je prilagođen na hrvatski jezik na temelju grafikona u radu POS *Tagging Approaches: A comparison* autora Kumawat i Jain, 2015.

2.1.1. Označivači zasnovani na pravilima

Označivači zasnovani na pravilima predstavljaju najstarije sustave za POS označavanje, a ručno pisana pravila na kojima su zasnovani se nazivaju pravila kontekstualnih okvira (engl. *context frame rules*) (Kumawat i Jain, 2015). Arhitektura koju su ovi označivači koristili bila je na dvije razine, na prvoj razini se svakoj riječi dodavala lista mogućih oznaka uz pomoć rječnika (Jurafsky i Martin, 2009). Zatim su se na drugoj razini smanjivali popisi mogućih POS oznaka uz pomoć opširnih popisa ručno pisanih pravila za razrješavanje višestrukih morfosintaktičkih opisa (Kumawat i Jain, 2015). Brill (1992) gradi označivač zasnovan na pravilima koji je na prvoj razini pridodao riječi najvjerojatniju oznaku bez obzira na kontekst, a zatim su se pogrešno označene riječi razrješavale uz pomoć pravila. To bi značilo da će se neka riječ označiti najčešćom ili najvjerojatnijom oznakom bez obzira što, na primjer, može biti imenica i glagol, a eventualne pogreške se kasnije ispravljaju prema pravilima. Još jedan primjer ovakvog označivača je *EngCG* autora Atra Voutilainena iz 1995. koji prvotno riječi provlači kroz leksikon te vraća moguće POS oznake, a zatim na rečenicu primjenjuje ograničenja kako bi izbacio netočne POS oznake (Jurafsky i Martin, 2009, str. 137). Morfološki analizator ovog označivača ima dvije razine te broji 180 ograničenja za razrješavanje, koja na temelju konteksta odmah izbacuju netočne oznake (Samuelsson i Voutilainen, 1997). Mnogi označivači koji rade na ovakav način koriste morfološke informacije prilikom razrješavanja, a neki čak nadilaze kontekst i morfologiju pa odmah uključe pravila za velika slova ili interpunkciju (Van Guilder, 1995). Jedan od primjera onih drugih je Brillov (1992) jednostavni označivač zasnovan na pravilima što se vidi iz sljedećeg opisa:

> „Kako bi se dio riječi kojih nema u korpusu odmah točno označio označivač koristi dvije procedure. Kod prve se podrazumijeva da su sve riječi koje počinju velikim slovom vlastite imenice čak i ako ne postoje u korpusu. Kod druge se podrazumijeva da se riječi kojih nema u korpusu označava u skladu s nastavkom od 3 slova kojeg imaju druge riječi u korpusu pa bi se riječ *blahblahous* označila kao pridjev s obzirom na to da je takva oznaka najčešća za riječi koje završavaju na *-ous*.“

Iako je proces kod ovakvih označivača očigledno sporiji s obzirom na potrebno ljudsko znanje, odnosno lingviste na drugoj razini, on nije u potpunosti odbačen. Kao što je prije spomenuto, zajedno s vjerojatnosnim pristupom, pristup zasnovan na pravilima se najčešće koristi u kombinaciji. Već spomenuti autor Brill (1995) je razvio poznati označivač koji se koristi

kombinacijom ova dva pristupa (Jurafsky i Martin, 2009). Rad tog označivača će se opisati nakon pojašnjena rada vjerojatnosnih označivača.

2.1.2. Vjerojatnosni označivači

Općenito, vjerojatnost podrazumijevamo kao šansu da će se neki događaj dogoditi (Allen, 1995, str. 185). Vjerojatnosni označivači se koriste čestotom, statistikom i računom vjerojatnosti (Bekavac, 2002). Logika iza svakog od njih je pojednostavljeni pristup odabira najvjerojatnije oznake za određenu riječ (Jurafsky i Martin, 2009). Budući da kod njih nedostaje element lingvističke struke, Kumawat i Jain (2015) ističu da im je to jedan od glavnih nedostataka jer oznake ponekad odstupaju od gramatičkih pravila.

Skriveni Markovljev Model ili HMM (engl. *Hidden Markov Model*) označivač jedan je od najpoznatijih vjerojatnosnih modela za POS i MSD označavanje. Općenita logika kojom se koristi je jednaka kao i kod ostalih vjerojatnosnih označivača, a to je odabir najvjerojatnije oznake. HMM označivač se koristi za treniranje označivača i za samo označavanje, a pokreće se u sklopu računalnog programa (Agić, i sur., 2008). HMM označivači ne običavaju dodjeljivati oznake riječ po riječ, već odmah generiraju niz oznaka za cijelu rečenicu (Jurafsky i Martin, 2009). Kod HMM modela se u obzir uzimaju dvije vrste vjerojatnosti, a to su vjerojatnost niza oznaka i vjerojatnost da se riječi povežu s oznakama (Agić i sur., 2008). Prvo HMM označivač odabire niz oznaka za zadanu rečenicu, a taj niz oznaka odgovara formuli P(word | tag) * P(tag | previous n tags), a nakon toga se rečenica označava kombinacijom oznaka s najvećom vjerojatnosti (Jurafsky i Martin, 2009; Kumawat i Jain, 2015). Odnosno možemo reći da se stvaraju „dvije različite vjerojatnosne matrice koje nose naziv jezični model i od kojih se jedna zove n-gram matrica, a druga matrica emisijskih vjerojatnosti“ (Agić i sur., 2008). Kod n-gram matrica n predstavlja broj riječi koji je korišten na uzorku pa tako razlikujemo unigram (n=1), bigram (n=2) i trigram (n=3) modele (Kumawat i Jain, 2015). Bigram model koristi uvjetnu vjerojatnost da će jedna riječ iz para pratiti drugu, a trigram model koristi uvjetnu vjerojatnost jedne od tri riječi uzimajući u obzir dvije riječi koje su joj prethodile (Allen, 1995). Matrica emisijskih vjerojatnosti nam daje informacije o vjerojatnosti emisije riječi kad dobiju oznake (Agić i sur., 2008). Kako bi se odabrao najvjerojatniji niz oznaka za svaku rečenicu imajući na umu niz riječi u njoj, koristi se Viterbi algoritam (Jurafsky i Martin, 2009).

> „Viterbi algoritam računa vrijednosti dva niza (*engl. array*) za koje su potrebni broj leksičkih kategorija (N) i broj riječi u rečenici (T). Prvi od njih, SEQSCORE (n, t) bilježi vjerojatnost za najbolji mogući slijed (engl. *sequence),* sve do pozicije t koja završava u posljednjoj leksičkoj

kategoriji L_n. Drugi od njih, BACKPTR, će za svaku kategoriju na svakoj poziciji naznačiti koja prethodna kategorija čini najbolji slijed na poziciji t-1.“ (Allen, 1995, str. 202).

2.1.3. Označivač zasnovan na transformaciji

Već spomenuti Brill označivač ili označivač zasnovan na transformaciji (engl. *transformation-based tagger*) ima značajke obje ranije opisane vrste označivača jer koristi pravila pri razrješavanju, a ona su automatski pronađena u označenom korpusu (Jurafsky i Martin, 2009). Ovaj označivač radi uz pomoć algoritma koji se zove transformacijsko učenje upravljano greškama (engl. *transformation-based error-driven learning*) i koji radi na način da neoznačeni tekst prvo prođe kroz interpreter početnog stanja, a potom se uspoređuje s ručno označenim korpusom (Brill, 1995, str. 545). Pravilima iz predloška s pravilima se prilažu pogrešno označene riječi, a pravilo prema kojem ima najmanje pogrešaka postaje dio naučenih pravila (Kumawat i Jain, 2015). Rad ovog označivača bi se prema Jurafskyju i Martinu (2009) mogao podijeliti u tri faze:

1. Dodjeljivanje najvjerojatnije oznake riječima u neoznačenom tekstu.
2. Odabir transformacije koja najtočnije označava nakon pregleda svih mogućnosti.
3. Ponovno označavanje u skladu s prethodnim odabirom.

2.1.4. Označivači zasnovani na povratnim neuronskim mrežama

Osnova ovih označivača je duboko učenje, a ono se definira kao „tehnika strojnog učenja kojom se računala i uređaji podučavaju logičkom funkcioniranju“ (Chatterjee, 2020, para. 1). Neki od njegovih modela su se počeli primjenjivati u morfosintaktičkom označavanju s rezultatima od 97% točnosti (Mujtaba, 2020, para. 10). Jedan od njih su povratne neuronske mreže, tj. RNN (engl. *recurrent neural networks*). Općenito, RNN funkcionira na sljedeći način:

> „RNN je funkcija koja učitava n vektora x1,…xn, i rezultira izlaznim vektorom hn koji se odnosi na cijeli niz x1,…xn. Vektor hn se kasnije unosi u neki klasifikator ili RNN-ove u hijerarhijskim modelima. Cijela se mreža zajednički uvježbava kako bi mogla započeti s predviđanjem na način da skriveni prikaz uči važne informacije iz niza.“ (Plank i sur., 2016).

Problem kod RNN-ova predstavljaju nestajanje i eksplozije gradijenata (engl. *gradient vanishing and exploding*) u praksi, unatoč tome što su u teoriji sposobni učiti i udaljene ovisnosti (engl. *long-distance dependencies*) (Ma i Hovy, 2016). Kako bi se taj problem riješio, posebno je osmišljena RNN arhitektura, a to su LSTM (engl. *long shorterm memory*) ćelije (Dayanand, 2020; Plank i sur. 2016). Općeniti izgled LSTM ćelije podrazumijeva troja vrata s pomoću kojih se nadziru informacije koje je potrebno zaboraviti ili preusmjeriti za daljnji rad (Ma i Hovy, 2016). Nadogradnja na RNN koja ulazni niz čita 2 puta, s lijeva na desno i obrnuto

se naziva dvosmjerna povratna neuronska mreža tj. biRNN (engl. *bidirectional recurrent neural network*), a njezina paralela zasnovana na LSTM-ima su dvosmjerni LSTM-ovi tj. BiLSTM (Plank i sur., 2016). Upravo se BiLSTM-ima koriste neki od dosad razvijenih označivača. Prije njihovog rada potrebno je podijeliti podatke na riječi i oznake, pri čemu ulazni niz X označava riječi, a izlazni niz Y oznake (Dayanand, 2020). Riječi se pojavljuju u obliku vektora (engl. *vectors*) koji uzimaju u obzir sličnosti i razlike između riječi u različitim dimenzijama čime se omogućava istovremeno uvježbavanje zadataka s malom i velikom količinom podataka (Ling i sur., 2015). Takvi se modeli nazivaju vektorski prikazi riječi (engl. *word embedding*) i „svakoj riječi u korpusu pridodaju vektor u semantičkom prostoru" (Cao i Rei, 2016, str. 18). Kod testiranja morfosintaktičkog označivača Ling i sur. (2016) osmišljavaju poseban model za generiranje tih vektorskih prikaza riječi ili ih kod nenadgledanih modela generiraju iz tablica (engl. *word lookup table*). Općenito, BiLSTM kao unos uzima niz riječi te kroz njega prolazi od početka do kraja i obrnuto, dobivajući tako i prošla i buduća stanja koja se kombiniraju u izlaz, a POS oznaka se predviđa u linearnom klasifikatoru koji tu kombinaciju uzima kao ulaz (Ling i sur., 2015; Bohnet i sur. 2018; Aggarwal, 2019).

2.2. Izazovi POS i morfosintaktičkog označavanja

Izazovi kod morfosintaktičkog označavanja mogli bi se podijeliti u tri kategorije koje navode Torbar i sur. (2020), a to su višeznačnost (*engl. ambiguity*) zbog koje bi riječi mogle imati više oznaka, veličina skupa oznaka i označavanje nepoznatih riječi kojih nema u korpusu ili za koje ne postoje unaprijed definirana pravila.

Višeznačnost je sastavni dio svakog jezika. Iako postoji više definicija ovog koncepta, općenito, „za izraz kažemo da je višeznačan ako ima više od jednog značenja" (Gillon, 1990). Jednako tako, za područje obrade prirodnog jezika Jurafsky i Martin (2009, str. 4) navode kako je „neki unos višeznačan ako postoji više alternativnih lingvističkih struktura koje se mogu sagraditi za njega". Postoji više podjela koncepta višeznačnosti, a jedan od najosnovnijih je na leksičku višeznačnost i strukturalnu višeznačnost koja se još dijeli na sintaktičku, klasnu i govornu višeznačnost (Stageberg, 1978, citirano u Kadłub, 2017). Drugi autori opisuju još jedan način njezine podjele na najvišoj razini pa tako Sennet (2016) navodi i pragmatičnu višeznačnost koja ovisi o kontekstu. Leksička višeznačnost se odnosi na homonimiju i polisemiju, dok strukturalnu dvosmislenost uvjetuje red riječi u rečenici (Kadłub, 2017). Ljudima je leksička višeznačnost jedna od lakših za razriješiti jer je dovoljno jednostavno promišljanje (Sennet, 2016). Bitan aspekt leksičke višeznačnosti za morfosintaktičko označavanje su homonimi koji

predstavljaju izazov jer, ne samo da mogu imati različito značenje, nego i, kako navodi Sennet (2016), mogu biti i potpuno druga vrsta riječi. Prema Tadiću (2003, str. 126) postoje dvije vrste homonimije, a to su unutarnja kod koje riječi imaju iste oznake iako predstavljaju različite oblike iste leme i vanjska kod koje ista riječ može imati više različitih oznaka jer predstavlja oblike više različitih lema. Isti autor navodi primjer oblika *gledatelji* koji može imati oznake NCMPD, NCMPL, NCMPI iako se radi o istoj lemi *gledatelj* i primjer oblika *cijene* koji istovremeno može biti oblik imenice *cijena* i treće lice množine glagola *cijeniti*.

Sljedeća je vrsta višeznačnosti sintaktička, a do nje dolazi kad red riječi u rečenici bude takav da ju je moguće shvatiti na više načina (Kadłub, 2017). Budući da se radi o shvaćanju značenja rečenice, ono nije toliko relevantno za morfosintaktičko označavanje, a ako bi neka riječ bila višeznačna Jurafsky i Martin (2009) navode kako bi se tom problemu pristupilo razrješavanjem smisla riječi (engl. *word sense disambiguition*). Sljedeća vrsta strukturalne višeznačnosti je klasna i odnosi se na to da određene riječi mogu pripadati raznim vrstama riječi (Kadłub, 2017). Prema tome, klasna je višeznačnosti vrlo bitan problem kod morfosintaktičkog označavanja jer je potrebno dati jasnu i točnu oznaku riječi koja, na primjer, može biti i imenica i glagol ovisno o kontekstu. Zadnja se višeznačnost iz ove podjele, govorna višeznačnost, odnosi na govor pa kao takva nije relevantna za morfosintaktičko označavanje. Prethodne se podjele vrlo često koriste u lingvistici i u filozofiji jezika, a jednu od podjela za svrhe morfosintaktičkog označavanja daju Quecedo i sur. (2020):

> „POS višeznačnost je ona prema kojoj riječ može imati jednu ili više sintaktičkih uloga. Višeznačnosti leme uključuje više mogućih lema površinskog oblika riječi. Do morfološke višeznačnosti dolazi kad površinski oblik riječi ima više mogućih morfosintaktičih oznaka. Višeznačnost smisla riječi nastupa kada jedna lema ima više različitih značenja."

Iz opisa označivača i njihove podjele iz prethodnog poglavlja jasno je da je jedna od njihovih zadaća razrješenje višeznačnosti. Također se podrazumijeva da će kod MSD oznaka ona prijeći na više razine od same vrste riječi poput razine padeža. Na razinama semantike i sintakse je ostalo još dosta problema poput opisa sintakse u zavisno složenim rečenicama ili određenja značenja u semantici, a algoritmi za njihovo kodiranje su vrlo kompleksni (Petkevič, 2014). Španjolski i hrvatski su jezici s različitim sintaksama te različitim stupnjevima fleksije i zato je važno pogledati s kojim se izazovima potrebno suočiti pri morfosintaktičkom označavanju svakog od njih.

Drugi od nabrojanih izazova je veličina skupa oznaka koja, kako navode Torbar i sur. (2020), može biti previše detaljna što može dovesti do toga da se miješaju slične oznake ili preopćenita,

bez dovoljno morfosintaktičkih informacija za dodavanje točne oznake. Dodatan problem kod samih oznaka mogu biti višerječni leksemi (engl. *multi-part words*) koji su sintaktički i semantički neodvojivi pa se kod pojedinih označivača gledaju kao jedna riječ (Jurafsky i Martin, 2009). Za slučajeve sintaktički neodvojivih leksema postoje posebni leksikoni, uz pomoć kojih im se dodjeljuju oznake iz posebnih skupova (McEnery i Wilson, 2001). Slavenski jezici za različite oblike fleksije u sklopu MULTEXT-East specifikacija imaju i preko tisuću oznaka (Divjak, i sur., 2017). Uzimajući to u obzir, hrvatski jezik ima puno veći skup oznaka od španjolskog jezika.

Treći od nabrojanih izazova predstavljaju nepoznate riječi. Prirodni je jezik vrlo široka raspona pa se u leksikonima može nalaziti i na stotine tisuća riječi, ali naravno da ih nije sve uvijek moguće obuhvatiti (McEnery i Wilson, 2001). Nadalje, jezik se stalno razvija pa nastaju različite novotvorenice poput, kako navode Jurafsky i Martin (2009, str. 158), „akronima, vlastitih i općih imena te glagola“ pa je potrebno imati način da se nepoznatim riječima doda oznaka. Većina se označivača koristi „metodom pogađanja na temelju sufiksa, velikog početnog slova nepoznatih riječi ili njihove raspodjele po vrstama riječi, dok se oni napredniji bave i određenjem na temelju oznaka okoline ili prefiksa“ (Orphanos i Christodoulakis, 1999, str. 134).

Veliki izazov također predstavljaju resursno ograničeni jezici (engl. *low-resource languages*) čiji je glavni nedostatak nepostojanje većih ručno obilježenih korpusa, bez obzira na to što se radi o jezicima s mnogo govornika (Agić i sur., 2016).

2.2.1. Izazovi u španjolskom jeziku

Iako španjolski jezik nije visoko flektivni jezik poput slavenskih jezika, u njemu postoji određena razina fleksije. Prvi tip fleksije koji postoji je glagolska fleksija odnosno konjugacija (*špa. flexión verbal*), a prema gramatici Gómeza Torrega (2005), glagoli u španjolskom se sastoje od korijena i flektivnih morfema koji nam daju informacije o licu, broju, vremenu, načinu i vidu. Glagol je jedina vrsta riječi u španjolskom jeziku koja ima fleksiju vremena, vida i broja, a kao i u ostalim romanskim jezicima ne nosi informacije o rodu (RAE, 2010). Druga vrsta fleksije je imenička fleksija odnosno deklinacija (*špa. flexión nominal*) koja prema Hrvatskoj enciklopediji (n.d.) podrazumijeva „oblično i značenjsko mijenjanje osnova sklonjivih imenskih po broju, rodu i padežu radi naznake njihove uloge u rečenici i njihovih veza s ostalim riječima“. Budući da u španjolskom jeziku nema padeža, ona se odnosi samo na rod i broj. Neki od primjera istopisnosti u španjolskom jeziku koje daju Quecedo i sur. (2020) su *vino* (imenica ili 3. lice jednine prošlog svršenog vremena glagola *venir*), *parecer* (glagol ili

imenica), *fui* (1. lice jednine glagola *ir* i *ser* u svršenom prošlom vremenu) ili *sobre* (prepozicija, imenica). Svi bi ovi primjeri odgovarali vanjskoj istopisnosti spomenutoj u prethodnom odjeljku jer se radi o oblicima različitih lema. Uz ranije navedene općenite izazove kod morfosintaktičkog označavanja, španjolski jezik ima i svoje specifičnosti u odnosu na hrvatski. U svom radu Parra Escartín i Martínez (2015) ističu sljedeće izazove na koje su naišli kod nekoliko označivača koje su analizirali: skraćeni oblici riječi, enklitike spojene s glagolima, imena, datumi i složene prepozicije. Dok ih u hrvatskom jeziku nema, u španjolskom postoje članovi. „Član je determinator koji stoji uz imenicu ili imensku skupinu“ (Hrvatska enciklopedija, n.d.). Skraćeni oblici riječi se sastoje upravo od člana, odnosno od određenog člana za muški rod *el* koji je spojen s prepozicijom *a* ili *de* koja mu prethodi u oblik *al* odnosno *del* (Parra Escartín i Martínez, 2015). „Enklitike su riječi bez svog naglaska koje slijede naglašenu riječ ili drugu enklitiku i s njima čini naglašenu cjelinu“ (Hrvatski jezični portal, n.d.). Čest je primjer enklitike u hrvatskom pomoćni glagol biti kao u primjeru *bila sam* i on se bez obzira na istu naglašenu cjelinu piše odvojeno. U španjolskom je situacija nešto drugačija pa se tako enklitike spajaju s glagolima s kojima čine naglasnu cjelinu. Radi se o direktnom spajanju zamjenica bez spojnice ili crtice, na primjer, glagolu *enviar* možemo dodati zamjenice *me* za indirektni objekt i *lo* za direktni objekt (Parra Escartín i Martínez, 2015). Dakle, kod oba od ovih primjera se naizgled čini kako je u pitanju jedna riječ, ali zapravo se radi o dvije i stoga je jedan od izazova za morfosintaktički označivač kako to naznačiti.

Ranije je naveden problem višerječnih leksema kojeg navode Jurafsky i Martin (2009), a primjer koji postoji u španjolskom, ali ne i u hrvatskom jeziku su glagolske perifraze. „Glagolske perifraze su sintaktičke konstrukcije od dva ili više glagola od kojih je jedan pomoćni, a drugi glavni te čine jedan predikat“ (Gomez Torrego, 2005). Budući da čine jedan predikat odnosno da su sintaktički i semantički neodvojivi, izazov za morfosintaktički označivač leži u tome hoće li se glagolsku perifrazu označiti kao jednu ili kao više riječi.

2.2.2. Izazovi u hrvatskom jeziku

U hrvatskom jeziku postoje i deklinacija i konjugacija. Deklinacija je detaljnija od španjolske jer se uz rod i broj oblici imenica također sklanjaju po padežu. Rod, broj i padež u hrvatskom nemaju posebni izraz već su izraženi jedinstvenim nastavkom (Barić i sur., 1997) Konjugacija glagola je također kompleksnija jer se unutar nje, za razliku od španjolskog, podrazumijeva i rod. Iako visoki stupanj fleksije, odnosno raznolikosti u flektivnim nastavcima u hrvatskom jeziku može predstavljati izazov za morfosintaktičko označavanje, Divjak i sur. (2017) ističu kako upravo to može biti pozitivno za slavenske jezike jer je moguće relativno precizno

pretpostaviti POS kategoriju iz njenih nastavaka u odnosu na germanske ili romanske jezike. U svom radu o morfosintaktičkom označavanju u hrvatskom i srpskom jeziku Agić i sur. (2013) kao neke od izazova pri označavanju navode:

> "U hrvatskom su se najčešće zamjene događale između pridjeva i imenica (npr. kod imena države, *Hrvatska* i *hrvatski*), imenica i glagola, pridjeva i priloga te imenica i priloga. […] Također je do problema došlo s homografima poput *strana,* miješanjem priloga s imenicama (*godinama)*, označavanjem veznika jer neke riječi također mogu imati funkciju veznika (*kako ili kada*), miješanjem imenica i glagola kako zbog homografije tako zbog imenica koje završavaju na *-lo* za srednji, *-la* za ženski rod ili na *-ti* poput infinitiva. Na morfosintaktičkoj razini je do problema dolazilo sa istim sufiksima koji označavaju različiti padež te zamjene u kategorijama roda i broja, najčešće između muškog i srednjeg roda."

U njihovim zapažanjima možemo primijetiti razne primjere vanjske istopisnosti o kojoj govori Tadić (2003) pa bi riječ *strana* mogla imati oznaku NCFSN ako se radi o imenici ili AGPFSNY ako se radi o pridjevu. Zbog postojanja padeža u hrvatskom jeziku, istopisnost će nekada biti njima uvjetovana što nije slučaj u španjolskom. Primjer unutarnje istopisnosti koji navodi Tadić (2003) su lokativ, dativ i instrumental množine pa tako riječ *gledateljima* može imati čak tri oznake: NCMPL, NCMPD i NCMPI. Osim ovih padeža istovjetan lik često imaju akuzativ i genitiv pa su Agić i sur. (2013) uočili da riječi poput *pobjednika* i *kandidata* često dobiju pogrešnu oznaku: NCMSAY umjesto NCMSG i obrnuto. Dodatan problem u hrvatskom može biti što neke riječi istovremeno mogu imati i vanjsku i unutarnju istopisnost kako navodi Tadić (2003) na primjeru lika *cijene* koji može biti u nominativu, akuzativu i vokativu množine te genitivu jednine, a isto tako može biti u trećem licu množine prezenta glagola *cijeniti*.

Još jedan od problema za slavenske jezike može predstavljati njihova veća fleksibilnost za red riječi u rečenici zato što se njime mogu povećati moguća mjesta subjekta i objekta u odnosu na glagol te broj mogućih oblika riječi (Divjak i sur., 2017). U hrvatskom jeziku postoji mnoštvo višerječnih leksema, odnosno semantički i sintaktički neodvojivih izraza, a jedan od primjera su višerječna imena poput *Sjedinjene Američke Države* (Kolokacijska baza hrvatskoga jezika, n.d.).

3. Opis odabranih morfosintaktičkih označivača

Kako bi se bolje razumjelo s kojim skupovima oznaka radi i na kojim korpusima je uvježban svaki od odabranih jezičnih alata u nastavku se opisuju njihove specifikacije i odabrani pristup svakom od njih.

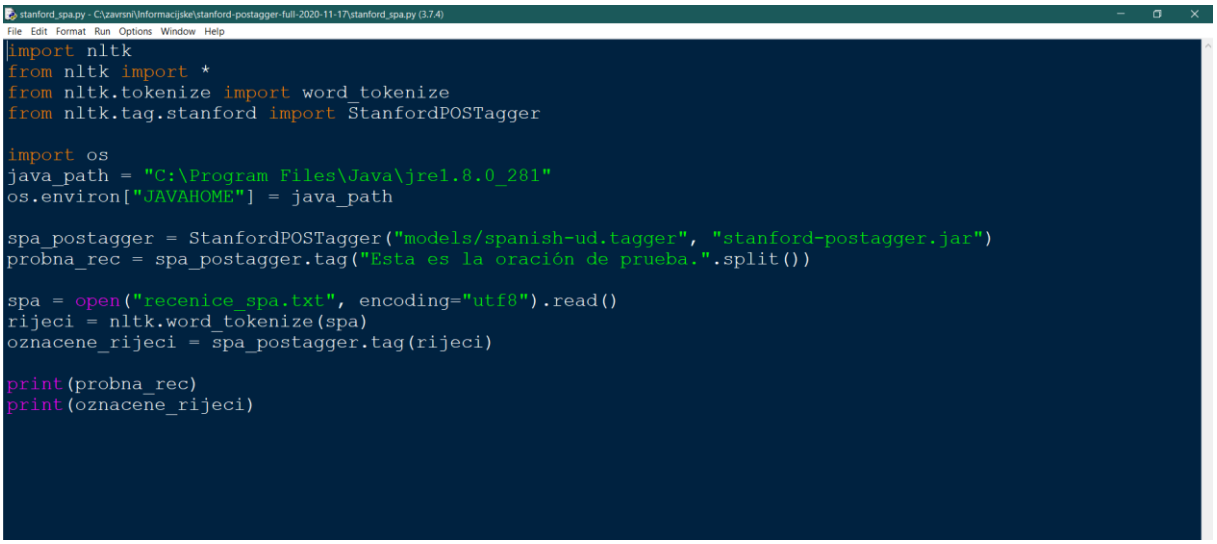
3.1. Stanfordski POS označivač

Stanfordski morfosintaktički označivač postoji za engleski, kineski, arapski, francuski, njemački i španjolski jezik, a napisala ga je Kristina Toutanova, dok su ga Dan Klein, Christopher Manning, William Morgan, Anna Rafferty, Michel Galley i John Bauer poboljšali i proširili (The Stanford Natural Language Processing Group, n.d., para. 1). Radi se o log-linearnom označivaču, a rad takvog označivača opisuju Toutanova i Manning (2000):

> „Označivač iz teksta uči log-linearni model koji se zasniva na uvjetnoj vjerojatnosti, a koristi se metodom maksimalne entropije. Taj model nekoj oznaci t u skupu mogućih oznaka T pridodaje određenu vjerojatnost uzimajući u obzir riječ i njen kontekst h koji je najčešće definiran kao niz riječi i oznaka koje im prethode."

Svi dosadašnji modeli za španjolski jezik koriste se UD Univerzalnim skupom oznaka, a trenirani su na AnCora 3.0 korpusu i The DEFT Spanish Treebank korpusu (The Stanford Natural Language Processing Group, n.d.). AnCora korpus se sastoji od španjolskog i katalonskog korpusa, a svaki od njih sadrži po petsto tisuća riječi (Taulé i sur., 2008). DEFT korpus sadrži kompletnu banku stabala iz novina za španjolski jezik i banku stabala s foruma za raspravu za hispanskoamerički španjolski i radi se o otprilike 5000 rečenica (The Stanford Natural Language Processing Group, n.d.).

Odabrani pristup ovom označivaču je putem sučelja programskog jezika Python putem NLTK (engl. *Natural Language Tool Kit*) platforme. NLTK platforma je otvorena za korištenje i sadrži podatke, dokumentaciju i opsežan softver, a između svih funkcionalnosti za obradu prirodnog jezika pruža i sučelja za provođenje morfosintaktičkog označavanja, sintaktičkog parsanja i klasifikacije teksta (Bird i sur., 2009). Na Slici 1 se može vidjeti kod koji je korišten u Python-u za označavanje španjolskih rečenica:

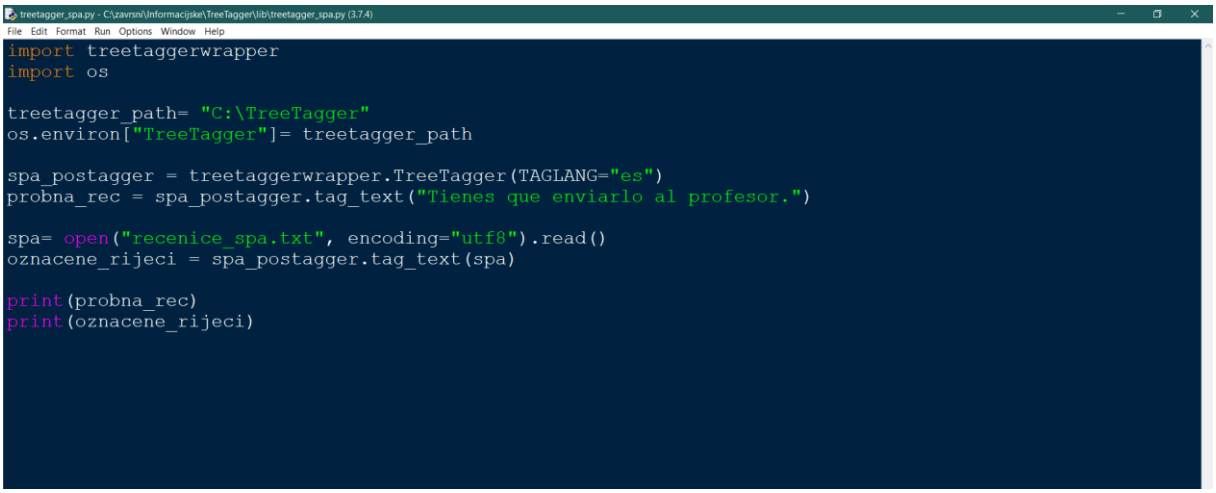


Slika1. Python kod za pokretanje Stanfordskog POS označivača

Kod je prilagođen prema uputama za pokretanje označivača u knjizi *Natural language processing with Python* autora Stevena Birda, Ewana Kleina i Edwarda Lopera (str. 179) i prema tutorijalu autorice Sabine Bartsch na web stranici *linguisticsweb.org*.

3.2. TreeTagger označivač

Označivač TreeTagger je razvio Helmut Schmid u sklopu projekta TC na Institutu za računalnu lingvistiku Sveučilišta u Stuttgartu, a koristi se za mnogo jezika, a neki od njih + su njemački, engleski, francuski, talijanski, danski, švedski, norveški, nizozemski, španjolski, bugarski i ruski (Schmid, n.d.). Model kojim se označivač koristi je ranije opisan u radu Skriveni Makrovljev model (Schmid, 1995). Skup oznaka koji je korišten za španjolski jezik u TreeTagger-u je poprilično opširan, a neki od primjera koji to pokazuju su zasebne oznake za infinitiv, gerund, particip i konjugirane oblike za dva glagola *biti* u španjolskom *(*špa. *ser i estar)* ili vrlo specifične oznake za pojedine vrste veznika u španjolskom. Ovom označivaču se također pristupilo putem sučelja programskog jezika Python, ali nije korištena NLTK platforma. Korišten je modul *treetaggerwrapper 2.3* kojim se TreeTagger omotava (engl. *wrap*) u klase u Python-u čime je omogućeno uzastopno označavanje nekoliko tekstova (Pointal, n.d.). Korišteni kod je prikazan na Slici 2:



Slika2. Python kod za pokretanje TreeTagger označivača

Kod je prilagođen prema dokumentaciji modula *treetaggerwrapper 2.3* autora Laurenta Pointala.

3.3. ReLDIanno označivač

RELDIanno označivač se koristi za hrvatski, slovenski i srpski jezik, a može mu se pristupiti putem web sučelja ili putem programskog jezika Python (Ljubešić i Erjavec, 2016). Označivač je vjerojatnosni i koristi se implementacijom CFR-ova (engl. *Conditional Random Fields*) (Ljubešić i sur., 2016). U općenitim crtama, „CRF-ovima se kombiniraju prednosti diskriminatorne klasifikacije i grafičkog modeliranja, odnosno kombiniraju se mogućnost kompaktnog modeliranja raznolikih izlaza *y* i mogućnost utjecaja na velik broj unesenih značajki *x* za predviđanje“ (Sutton i McCallum, 2011, str. 269). Kod ReLDIanno MSD označivača se pri klasifikaciji diskriminira između 2 skupa značajki od kojih jedan ima one koje dokazano rade pri procesu označavanja, a drugi je eksperimentalni (Ljubešić i Erjavec, 2016). Skup oznaka koji se koristi je predložen u sklopu *The MULTEXT-East* projekta opisanog u prvom poglavlju. Radi se o vrlo opširnom skupu oznaka kojim se nastoje obuhvatiti sve fleksijske promjene i obilježja riječi. Primjeri nekih oznaka su NCMSAN (N=imenica. C=opća, M=muški rod, S=jednina, A=akuzativ, N=neživo), AGPMSNY (A=pridjev, G=opći, P=pozitiv, M=muški rod, S=jednina, N=nominativ, Y=živo), PP1-SN (P=zamjenica, P=osobna, 1=1. lice, S=jednina, N=nominativ), SG (adpozicija u genitivu), CC (veznik nezavisnosložene rečenice) itd. (Ljubešić i Erjavec, 2016). Označivač nudi usluge „morfosintaktičkog označavanja, lematizacije, prepoznavanja imenovanih entiteta i ovisnosnog parsanja“ (Ljubešić i Erjavec, 2016). Pristup koji je odabran ReLDIanno označivaču je putem web aplikacije. Kako bi se pokrenuo morfosintaktički označivač, potrebno je u polje unijeti tekst ili prenijeti datoteku s računala te pod *Function* imati odabranu opciju *Tag* (Ljubešić i Erjavec, 2016).

4. Analiza označivača

U nastavku će se analizirati dobiveni rezultati označivača u španjolskim i hrvatskim rečenicama, kako bi se vidjelo koje su oznake dali riječima i je li bilo eventualnih pogrešaka. Rečenice su prikupljene iz književnih djela, akademskih članaka, novinskih članaka i mrežnih žargonskih rječnika. Cilj odabira rečenica iz više različitih izvora je označivačima dati više eventualnih izazova. Iako su korpusi mali, nastoje obuhvatiti što više jezičnih pojavnosti za analizu ranije opisanih izazova u morfosintaktičkom označavanju, a ne za analizu kvalitete samih označivača, za što bi bio potreban daleko opsežniji korpus.

4.1. Korpusi

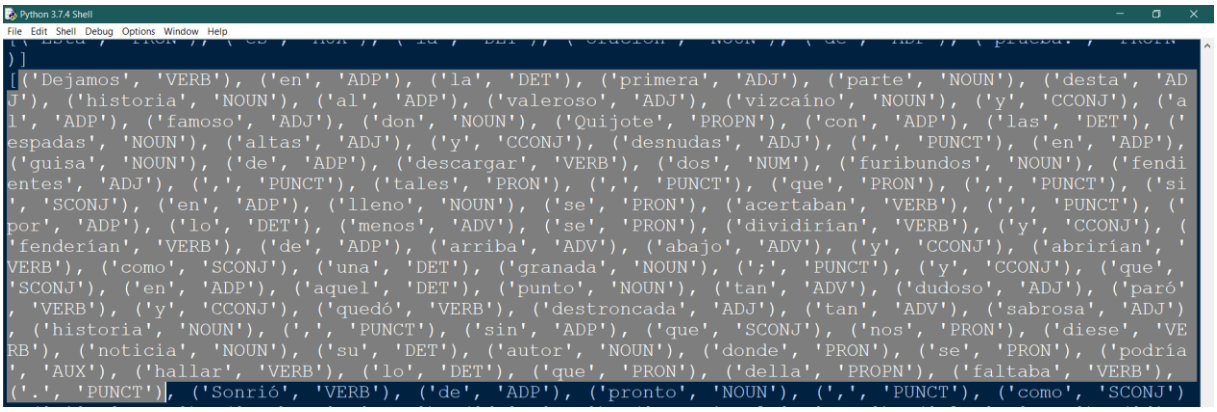
Korpus za španjolski jezik je pohranjen u tekstualnu datoteku *recenice_spa.txt*. U označivačima za španjolski su se iz književnih djela koristile rečenice iz romana *Don Quijote* Miguela de Cervantesa i *Posljednje večeri s Terezom* Juana Marsea. Rečenice su odabrane nasumično iz djela po kriteriju da se radi o jednoj duljoj, dvije ili tri rečenice. Odabrana su dva romana kako bi se obuhvatio stariji književni izraz u romanu *Don Quijote* i moderniji u romanu *Posljednje večeri s Terezom*. Rečenice su također preuzete iz znanstvenog članka *Sobre el tabú, el tabú lingüístico y su estado de la cuestión* autorice Anette Calvo Shadid. Kao novinski članak je odabran *Un recurso al Constitucional suspende el ingreso en prisión de 10 condenados por el asalto al centro cultural Blanquerna* portala *El País*. Iz ovih su izvora rečenice odabrane nasumično, po kriteriju da se radi o jednoj duljoj, dvije ili tri rečenice. Rečenice iz mrežnog rječnika žargona *AsíHablamos* odabrane su kao primjer kolokvijalnog izraza. Rečenice su odabrane nasumično iz rječnika po kriteriju od 5 riječi s obzirom na to da su primjeri za svaku riječ kraći.

Korpus za hrvatski jezik je pohranjen u tekstualnu datoteku *recenice_hrv.txt*. U označivaču za hrvatski su se iz književnih djela koristile rečenice iz romana *Povratak Filipa Latinovicza* Miroslava Krleže i pripovijetke *Prijan Lovro* Augusta Šenoe. Odabrana su dva romana kako bi se obuhvatio stariji književni izraz u pripovijetci *Prijan Lovro* i izraz svojstven Miroslavu Krleži u romanu *Povratak Filipa Latinovicza*. Iz znanstvenog članka koristile su se rečenice iz članka *Odnos strukturalne semantike prema kognitivnoj* autorice Ide Raffaelli. U označivaču za hrvatski su se iz novinskih članka koristile rečenice iz članka *Za projekte na područjima naseljenima nacionalnim manjinama 20 milijuna kuna* s portala *N1*. Iz ovih su izvora rečenice odabrane nasumično po kriteriju da se radi o jednoj duljoj, dvije ili tri rečenice. Također su se koristile rečenice iz mrežnog rječnika žargona *Žargonaut* kao primjeri kolokvijalnog govora.

Rečenice su odabrane nasumično iz rječnika po kriteriju od 5 riječi s obzirom na to da su primjeri za svaku riječ kraći.

4.2. Rezultati i usporedba španjolskih označivača

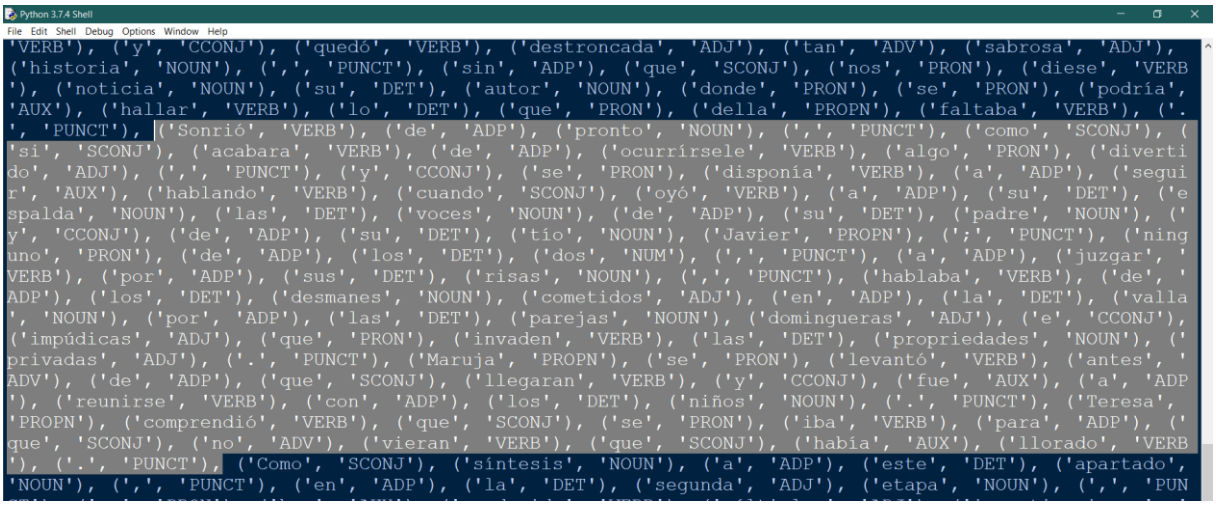
Prvo će se analizirati Stanfordski POS označivač. Kao što je ranije navedeno ovaj se označivač koristi Univerzalnim skupom oznaka projekta *Universal Dependencies* i izlaz koji je dobiven u Python-u odgovara tom skupu. Prije samog korpusa za označavanje, isprobala se jedna rečenica za provjeru rada označivača. Ona neće činiti predmet analize. U prvoj rečenici je dobiveni izlaz prikazan na Slici 3:



Slika3. Rezultat Stanfordskog POS označivača na rečenici iz romana *Don Quijote*

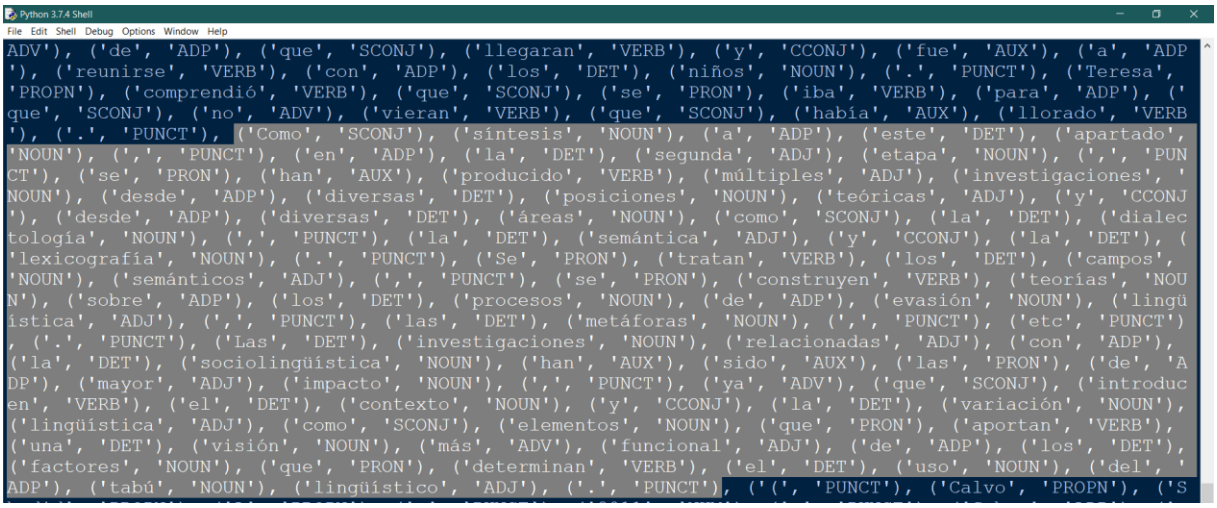
Prva zanimljivost uočena je kod riječi *desta* koja je označena kao pridjev. Ne radi se o pridjevu, već o spoju prepozicije *de* i pokaznog determinanta *esta.* Taj se spoj prema rječniku instituta *Real Academia Española (RAE)* (2014) više ne koristi u španjolskom jeziku. Uzimajući u obzir mali broj oznaka s kojima radi ovaj označivač, ne iznenađuje da je ovom spoju pridodana oznaka ADJ za pridjev jer se nalazi prije imenice *historia* koju pobliže označava. Ranije je kod izazova u označavanju navedeno spajanje španjolskih prepozicija s članom. U Stanfordskom POS označivaču je spoju prepozicije *a* s članom *el* dodana oznaka ADP za adpoziciju. Dodatan problem bi mogla stvarati NLTK platforma jer su Talamé i sur., (2019) u svom istraživanju opazili kako kod tokeniziranja program nije odvajao riječi od interpunkcijskih znakova koji su joj prethodili. Budući da se radi o interpunkciji koja je „zalijepljena“ na riječ, moguće je da po istoj analogiji skraćeni oblici riječi nisu dobro tokenizirani. Nadalje, označivač je riječ *lleno* u ovom slučaju dobro označio kao imenicu iako je njena puno češća uporaba kao pridjev (hrv. *pun*). Još je jedna zanimljivost da je označivač odvojeno označio priložne lokucije *de arriba abajo* i *por lo menos* iako se radi o sintaktički i semantički neodvojivim izrazima. Nivre i sur. (2016) navode kako se višerječni izrazi kod UD univerzalnog skupa oznaka označavaju posebnim ovisnosnim vezama, a ne povezivanjem više tokena u jedan. Kod riječi *como* je došlo

do pogreške jer se u ovom slučaju ne radi o vezniku zavisnosložene rečenice već o prilogu pa bi točna oznaka bila ADV. Riječ *della* je označena kao vlastito ime PROPN, no ovdje se ponovo radi o skraćenom obliku riječi. Skraćeni oblik se sastoji od prepozicije *de* i zamjenice *ella*, koji prema rječniku RAE (2014) više nije u jezičnoj uporabi. Iako se prema objašnjenju oznake PROPN ona ponekad može odnositi na zamjenice i to u slučaju ako se pojavljuju kao višerječno vlastito ime koje ima funkciju vlastite imenice (Nivre, i sur., 2016), u ovom kontekstu se osobna zamjenica *ella* odnosi na imenicu *historia*, koja je opća imenica.



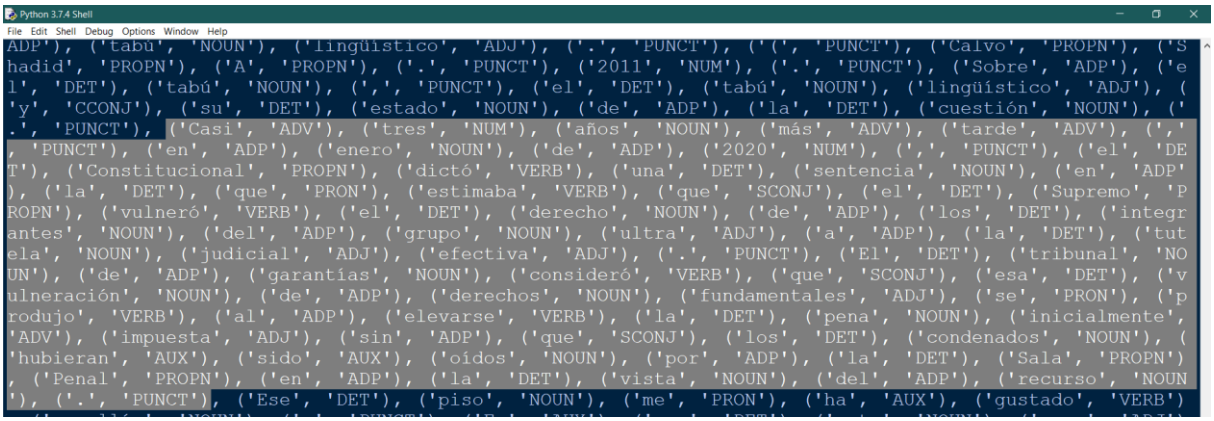
Slika4. Rezultat Stanfordskog POS označivača na rečenicama iz romana *Posljednje večeri s Terezom*

Rečenice na Slici 4 pripadaju romanu *Posljednje večeri s Terezom*. Izazov s enklitikama spojenima direktno na glagol koji spominju Parra Escartín i Martínez (2015) je u Stanfordskom POS označivaču riješen tako da se lik *ocurrírsele* jednostavno gledao kao glagol bez obzira na zamjenicu *le* za indirektni objekt koja je na njega „zalijepljena“. Također je moguće da je došlo do problema prilikom tokenizacije kao u prethodnom paragrafu. Isti je slučaj s povratnim glagolom *reunirse* kod kojeg je zamjenica *se* u infinitivu vizualno spojena s glagolom *reunir*. U ovom je primjeru također bila glagolska perifraza *seguir hablando* koja je dobro označena jer čini predikatnu cjelinu pa zato glagol *seguir* koji je ovdje pomoćni ima oznaku AUX, a ne VERB. Kao i u prethodnom primjeru i ovdje su riječi višerječnog izraza *a juzgar por* odnosno prepozicijske lokucije označene zasebnim oznakama, koje su točne za svaku riječ. U španjolskom su neki veznici zavisnosloženih rečenica također višerječni pa su tako cijeli izrazi *antes de que* i *para que* zapravo veznici.



Slika5. Rezultat Stanfordskog POS označivača na rečenicama iz znanstvenog članka *Sobre el tabú, el tabú lingüístico y su estado de la cuestión*

Rečenice na Slici 5 pripadaju članku *Sobre el tabú, el tabú lingüístico y su estado de la cuestión.* Ono što je odmah uočljivo je riječ *como* koja je u prvoj rečenici oba puta i u trećoj jednom označena sa SCONJ, ali nije veznik zavisne rečenice, već je prijedlog čiji je hrvatski ekvivalent *kao* pa bi odgovarajuća oznaka bila ADP. Nadalje, iako je za pretpostaviti da će kratica *etc* biti označena s X kao ostalo, označivač joj je dao oznaku PUNT za interpunkciju. U ovom se primjeru pojavljuje još jedan veznik kojeg tvore dvije riječi, a to je uzročni veznik *ya que* kod kojeg je samo *que* označen kao veznik zavisne rečenice dok je riječi *ya* pridodana točna gramatička kategorija ako bi je se gledalo izolirano, a to je ADV tj. prilog.



Slika6. Rezultat Stanfordskog POS označivača na rečenicama iz članka s portala *EL PAÍS*

Slika 6 prikazuje rezultat označivača na rečenicama iz članka s portala *EL PAÍS*. Riječ *oídos* je pogrešno označena kao imenica *unutrašnje uho*, ali zapravo se radi o participu glagola *oír.* U ostatku teksta nisu zamijećene dodatne pogreške uzrokovane homonimijom. Ovakvom

rezultatu bi moglo pridonijeti što se AnCora korpus, jedan od korpusa na kojima je treniran označivač, većinom sastoji od novinskih članaka (Taulé i sur., 2008).

Slika7. Rezultat Stanfordskog POS označivača na rečenicama iz mrežnog rječnika žargona *AsíHablamos*

Na Slici 7 mogu se vidjeti rezultati na rečenicama iz mrežnog rječnika žargona *AsíHablamos.* U španjolskom jeziku se subjekt podudara s glagolom tj. predikatom (RAE, 2010, str. 3967) pa će u rečenici *Ese piso me gusta mogollón*, riječ *piso* (3. lice jednine) biti subjekt jer se podudara s *gusta* (3. lice jednine), a zamjenica *me* indirektni objekt. Zbog toga *mogollón* ovdje nikako ne može biti imenica u funkciji subjekta, već se radi o prilogu, iako je u nekim kontekstima i imenica. *Mogollón* je kolokvijalni izraz koji se u ovom kontekstu koristi umjesto priloga *mucho* (hrv. *puno*). Zanimljivo je vidjeti kako je riječ *guay* (hrv. *kul, fora*) u svakom od primjera različito označena. U prva dva primjera se radi o pridjevu i stoga je oznaka VERB kod *es guay* pogrešna, a ADJ ispravna. Kao posljedica toga i glagol *es* je krivo označen kao pomoćni glagol (AUX) jer je označivač mislio da je *guay* glagol. U trećoj je rečenici označeno kao imenica (NOUN), ali se radi o prilogu (ADV). U ovim se primjerima također vidi kako ¿ nije označen kao interpunkcija, a radi se o znaku kojim se označava početak upitne rečenice. Ovu su pojedinost NLTK platforme opazili Talamé i sur., (2019) u svom radu jer prilikom tokeniziranja program nije odvajao riječ od interpunkcijskih znakova ispred nje.

Slijedi analiza TreeTagger označivača. Kao što je ranije navedeno ovaj označivač ima dosta detaljniji skup oznaka i izlaz koji je dobiven u Python-u odgovara tom skupu, a također su navedene i leme svake pojavnice. Prije samog korpusa za označavanje, isprobala se jedna rečenica za provjeru rada označivača. Ona neće činiti predmet analize. U prvoj rečenici je dobiveni izlaz na Slici 8:

Slika8. Rezultat TreeTagger označivača na rečenici iz romana *Don Quijote*

Prva zanimljivost je također uočena kod riječi *desta* za koju je rečeno je je spoj prijedloga *de* i zamjenice *esta* koji se više ne koristi. TreeTagger ju je označio oznakom VLfin, odnosno kao konjugirani glagol. S obzirom na to da ovaj označivač koristi skup oznaka u kojem su zasebno navedene oznake za najpoznatije spojeve riječi *al* i *del* koji su navedeni kao izazovi u ranijem poglavlju, ne iznenađuje da ga je ovaj zastarjeli oblik zbunio. Moguće je pretpostaviti da bi TreeTagger također imao posebnu oznaku za taj skraćeni oblik da se nastavio koristiti. Skraćeni oblik *al* prijedloga *a* i člana *el* je ovaj označivač označio posebnom oznakom PAL. Pojavnica *guisa* je ovdje zbog vanjske istopisnosti pogrešno označena kao konjugirani glagol VLfin (3. lice jednine glagola *guisar*), a zapravo se radi o imenici *način.* Još je jedna zanimljivost uočena kod načina na koji ovaj označivač označava glagolsku perifrazu *quedó destroncada* kod kojeg lik *quedó* označava kao konjugirani glagol (VLfin) dok lik *destroncada* označava kao particip (VLadj).

Slika9. ***Rezultat TreeTagger označivača na rečenicama iz romana Posljednje Večeri s Terezom***

Na Slici 9 prikazan je izlaz označivača na primjerima iz romana *Posljednje Večeri s Terezom*. Ono što je označivač ovdje potpuno pogrešno označio je lik *ocurrírsele* koji predstavlja izazov zbog zamjenice *le* spojene direktno na infinitiv povratnog glagola *ocurrirse.* On je ovdje označen kao opća imenica (NC). Takav rezultat je iznenađujući s obzirom na to da u skupu oznaka postoje oznake za glagole s enklitikama u gerundu (VCLIger), indikativu (VCLIinf) i finitnom obliku (VCLIfin) namijenjene rješavanju ovog problema. Pretpostavljeno je da je označivač zbunio povratni glagol i da bi se nepovratni glagol u infinitivu s enklitikom za direktni objekt označio oznakom VCLIinf. Ta je pretpostavka i isprobana s glagolom *enviar* (hrv. poslati) koji kao primjer za ovaj izazov navode Parra Escartín i Martínez (2015). Glagol nije povratan i može nositi direktni objekt kao enklitiku. Umjesto testne rečenice u program je unesena rečenica *Tienes que enviarlo al profesor*. (hrv. *Moraš to poslati profesoru*.). TreeTagger je liku *enviarlo* dao pretpostavljenu oznaku VCLinf, a to se može vidjeti na Slici 10:

Slika10. Rezultat TreeTagger označivača na liku *enviarlo*

Nadalje, još je jedna od zanimljivosti da je obje riječi u višerječnom kondicionalnom vezniku *como si* TreeTagger označio kao veznike zavisne rečenice (CSUBX), dok kod vremenskog veznika *antes de que* to nije bio slučaj pa je riječ *antes* označena kao prilog (ADV), a riječ *de* kao prijedlog (PREP). Radi se o točnim oznakama, ako te riječi gledamo izvan veznika.

Slika11. Rezultat TreeTagger označivača na rečenicama iz znanstvenog članka *Sobre el tabú, el tabú lingüístico y su estado de la cuestión*

Na Slici 11 prikazan je izlaz označivača na primjerima iz znanstvenog članka *Sobre el tabú, el tabú lingüístico y su estado de la cuestión*. Kao i Stanfordski POS označivač, TreeTagger je riječ *como* u prvoj rečenici oba puta i u trećoj jednom označio s CSUBX, ali nije veznik zavisne rečenice, već je prijedlog čiji je hrvatski ekvivalent *kao* pa bi odgovarajuća oznaka bila PREP. Kratica *etc* je također pogrešno označena kao opća imenica (NC), no to ne iznenađuje previše jer unatoč opširnom skupu oznaka, ne postoji posebna oznaka za kratice ili kategorija „ostalo“.

Slika12. Rezultat TreeTagger označivača na rečenicama iz članka s portala *El PAÍS*

Na Slici 12 prikazan je izlaz označivača na primjerima iz novinskog članka s portala EL PAÍS. Prva pogreška se javlja s nazivom za Ustavni sud *el Constitucional* koji bi trebao imati oznaku NP za vlastitu imenicu. TreeTagger ga je označio kao pridjev (ADJ) jer riječ *constitucional* (hrv. *ustavni*) to i jest, ako ne čini naziv za sud. TreeTagger je *que* u *el Constitucional dictó una sentencia en la que* označio kao veznik zavisne rečenice, ali zapravo se radi o relativnoj

zamjenici korištenoj kako se imenica *sentencia* ne bi ponavljala pa bi točna oznaka bila REL za relativne zamjenice. Riječ *oído* je također pogrešno označena kao imenica *unutrašnje uho*, ali zapravo se radi o participu glagola *oír*. Riječ *Penal* čini dio višerječnog vlastitog imena institucije *Sala Penal* (hrv. *kazneni sud*) pa bi trebala imati oznaku NP, umjesto NC.

Slika13. Rezultat TreeTagger označivača na rečenicama iz mrežnog rječnika žargona *AsíHablamos*

Na Slici 13 je prikazan izlaz označivača na rečenicama iz mrežnog rječnika žargona *AsíHablamos*. Kako je već objašnjeno kod analize Stanfordskog POS označivača, zbog subkategorizacijskog okvira glagola *gustar* riječ *mogollón* nikako ne može biti imenica u funkciji subjekta (NC), već se radi o prilogu (ADV). Također je do problema došlo s riječi *guay* koja je u sva tri primjera označena kao vlastita imenica NP, a zapravo se radi o pridjevu u prva dva primjera i stoga je oznaka ADJ ispravna. U zadnjem primjeru se radi o prilogu pa je potrebna oznaka ADV.

Na kraju, oba su morfosintaktička označivača imala problema s oblicima *desta* i *della* koji se više ne koriste u španjolskom jeziku iz djela *Don Quijote*. Dok ih je Stanfordski POS označivač označio kao ADJ (pridjev) i PROPN (vlastita imenica), Treetagger ih je označio kao VLfin (glagol u finitnom obliku). Oba od tih rješenja daleko su od stvarnih gramatičkih kategorija tih riječi. Moderne skraćene oblike riječi *al* i *del* svaki je od označivača označio na različite načine. Stanfordski POS označivač je koristio oznaku ADP kojom se označava samo gramatička kategorija prijedloga *a* i *de*. TreeTagger je koristio oznake PAL i PDEL koje su osmišljene posebno za ove skraćene oblike riječi, no nije eksplicitno naznačeno o kojim se gramatičkim

kategorijama radi. Oba označivača su također pogriješila s oznakom za riječ *como* koja je u danom kontekstu bila prepozicija, dok su je oni označili kao veznik (CSCONJ u Stanfordskom POS označivaču i CSUBX u TreeTagger označivaču).

Mnogo je razlika opaženo u oznakama iz kojih je jasno da je TreeTagger detaljniji. Prvi od primjera su oznake za interpunkcijske znakove kod kojih Stanfordski POS označivač ima oznaku PUNCT za sve njih, dok TreeTagger označivač ima oznake za svaki pojedinačno poput FS za točku, CM za zarez ili COLON za dvotočku. Stanfordski je POS označivač članove označavao oznakom DET za determinante, TreeTagger označivač koristio je oznaku ART za član. Specifičnije oznake TreeTagger označivača postoje i za brojeve, glavni broj *tres* (hrv. *tri*) je tako označen oznakom CARD, dok je u Stanfordskom POS označivaču označen s oznakom NUM. Redni broj *segunda* (hrv. *druga*) je u TreeTagger označivaču označen kao ORD, a u Stanfordskom POS označivaču označen je kao pridjev (ADJ). Razlika je također uočena kod označavanja prijedloga koje Stanfordski POS označivač označava s ADP, odnosno kao adpozicije, što je krovni termin za prepozicije i postpozicije, dok TreeTagger označivač daje oznaku PREP. TreeTagger označivač također ima posebnu oznaku za kvantifikatore, odnosno prema Hrvatskoj enciklopediji (n.d.) „oznaka za jedinicu kojom se označava količina kao odredba imena u lingvistici.“. Prema rječniku RAE (2014) oni mogu biti više vrsta riječi, a u primjeru na kojem je zamijećena oznaka QU za kvantifikator je *tales.* Toj je riječi Stanfordski POS označivač dodijelio oznaku PRON za zamjenicu, a to u danom kontekstu i jest jer zamjenjuje imenicu *fendientes*. Još je jedan primjer u kojem TreeTagger označivač detaljnije označava zamjenice, a to je oznaka DM koja se daje pokaznim zamjenicama. Nadalje, TreeTagger označivač nešto detaljnije označava i različite tipove veznika. Na primjer, veznik *que* ima posebno za njega namijenjenu oznaku CQUE, veznik *si* ima oznaku CSUBX kojom su pokriveni svi veznici zavisno složenih rečenica, dok su kod Stanfordskog POS označivača svi veznici zavisnosloženih rečenica objedinjeni oznakom SCONJ.

Razlike su opažene i kod načina označavanja glagolskih perifraza. Na primjer glagolsku perifrazu *seguir hablando* Stanfordski POS označivač označava kao AUX (seguir) i VERB (hablando), dok TreeTagger označivač koristi oznake VLinf (seguir) i VLger (hablando). Iako se ne radi o pogrešnim oznakama, čini se kako je kod Stanfordski POS označivača postignut dojam povezanosti ova dva glagola, dok se kod TreeTagger označivača može pomisliti da se nabraja glagol u infinitivu pa glagol u gerundu. Slične se razlike zapažaju u označavanju jednog od prošlih glagolskih vremena koje se sastoji od pomoćnog glagola *imati* i participa pa je Stanfordski POS označivač tako u primjeru *han producido* dao oznake AUX i VERB dok je

TreeTagger koristio oznake VHfin i VLadj. Zanimljivo je da TreeTagger ima cijeli skup oznaka za oblike pomoćnog glagola imati (špa. *haber*), a VHfin je jedna od njih koja predstavlja taj glagol u finitnom obliku. Ipak, valja napomenuti kako upravo veličina skupa oznaka kako su naglasili i Torbar i sur. (2020) predstavljala određene izazove. Na primjer kod Stanfordskog POS označivača je nekad dolazilo do nedostatka morfoloških informacija pa je tako lik *hablando* imao oznaku VERB koja ne sadrži informaciju o tome da se radi o gerundu. Također je uočeno kako niti jedan od dva označivača u korištenom skupu oznaka nikako ne kodira muški i ženski rod.

U Tablici 3 se može vidjeti pojava pogrešaka uzrokovanih izazovima navedenima u odjeljku 2.2.1.:

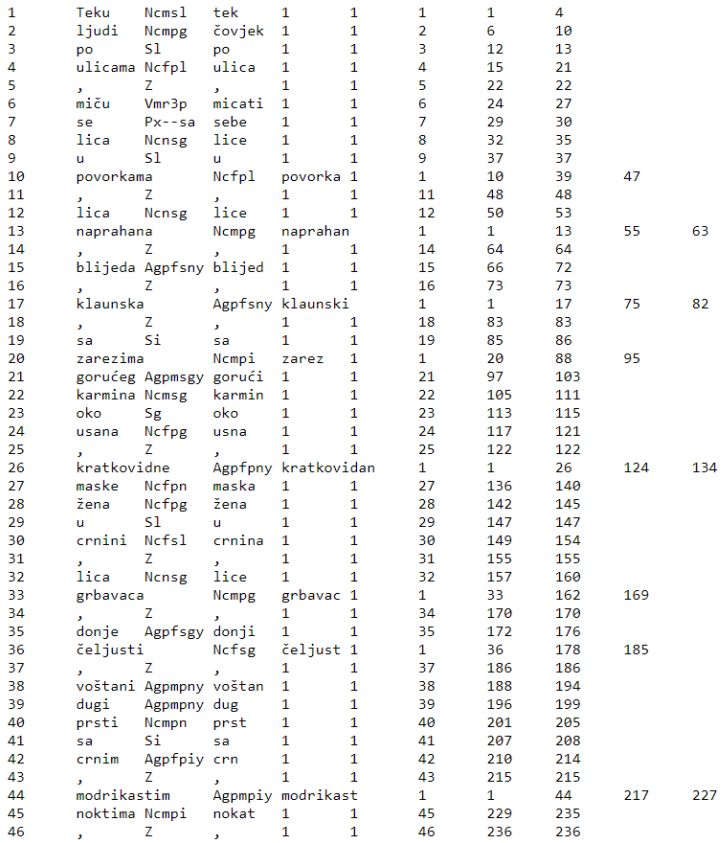
Izazovi	**Broj pogrešaka: Stanfordski POS označivač**	**Broj pogrešaka: TreeTagger**
Vanjska istopisnost	9	12
Skraćeni oblici riječi	0	1
Enklitike spojene na glagol	0	1
Glagolske perifraze	0	0

Tablica3. Prikaz broja pogrešaka na španjolskom korpusu uzrokovanih izazovima morfosintaktičkog označavanja

Višeznačnost se pokazala kao najveći izazov za označivače u korištenim primjerima što nije toliko iznenađujući podatak s obzirom da je svaki jezik inherentno višeznačan. Skraćeni oblici riječi zasigurno su prepoznat problem u morfosintaktičkom označavanju španjolskog jezika, a rješava se na različite načine. Kako je već navedeno, moguće je da odabrana platforma NLTK za pokretanje Stanfordskog POS označivača takve oblike nije odvojila prilikom tokenizacije pa se ne može reći da ih je sam označivač zanemario. Zato nisu računate kao pogreške. Iz istog razloga je tako odlučeno s enklitikama za ovaj označivač. Također, nije nužno da označivači višerječne izraze označavaju kao jednu cjelinu pa im ni to nije uzeto kao pogreška.

4.3. Rezultati hrvatskog označivača

Na web sučelje ReLDIanno označivača je prenesena datoteka *recenice_hrv.txt* te su odabrane opcije Tag + Lemmatise. Budući da je izlaz u tabličnom obliku previše dugačak kako bi se prikazao za svako djelo posebno, na Slici 14 je prikazan kao općeniti primjer:



1	Teku	Ncmsl	tek	1	1	1	1	4		
2	ljudi	Ncmpg	čovjek	1	1	2	6	10		
3	po	Sl	po	1	1	3	12	13		
4	ulicama	Ncfpl	ulica	1	1	4	15	21		
5	,	Z	,	1	1	5	22	22		
6	miču	Vmr3p	micati	1	1	6	24	27		
7	se	Px--sa	sebe	1	1	7	29	30		
8	lica	Ncnsg	lice	1	1	8	32	35		
9	u	Sl	u	1	1	9	37	37		
10	povorkama		Ncfpl	povorka	1	1	10	39	47	
11	,	Z	,	1	1	11	48	48		
12	lica	Ncnsg	lice	1	1	12	50	53		
13	naprahana		Ncmpg	naprahan		1	1	13	55	63
14	,	Z	,	1	1	14	64	64		
15	blijeda	Agpfsny	blijed	1	1	15	66	72		
16	,	Z	,	1	1	16	73	73		
17	klaunska		Agpfsny	klaunski		1	1	17	75	82
18	,	Z	,	1	1	18	83	83		
19	sa	Si	sa	1	1	19	85	86		
20	zarezima		Ncmpi	zarez	1	1	20	88	95	
21	gorućeg	Agpmsgy	gorući	1	1	21	97	103		
22	karmina	Ncmsg	karmin	1	1	22	105	111		
23	oko	Sg	oko	1	1	23	113	115		
24	usana	Ncfpg	usna	1	1	24	117	121		
25	,	Z	,	1	1	25	122	122		
26	kratkovidne		Agpfpny	kratkovidan		1	1	26	124	134
27	maske	Ncfpn	maska	1	1	27	136	140		
28	žena	Ncfpg	žena	1	1	28	142	145		
29	u	Sl	u	1	1	29	147	147		
30	crnini	Ncfsl	crnina	1	1	30	149	154		
31	,	Z	,	1	1	31	155	155		
32	lica	Ncnsg	lice	1	1	32	157	160		
33	grbavaca		Ncmpg	grbavac	1	1	33	162	169	
34	,	Z	,	1	1	34	170	170		
35	donje	Agpfsgy	donji	1	1	35	172	176		
36	čeljusti		Ncfsg	čeljust	1	1	36	178	185	
37	,	Z	,	1	1	37	186	186		
38	voštani	Agpmpny	voštan	1	1	38	188	194		
39	dugi	Agpmpny	dug	1	1	39	196	199		
40	prsti	Ncmpn	prst	1	1	40	201	205		
41	sa	Si	sa	1	1	41	207	208		
42	crnim	Agpfpiy	crn	1	1	42	210	214		
43	,	Z	,	1	1	43	215	215		
44	modrikastim		Agpmpiy	modrikast		1	1	44	217	227
45	noktima	Ncmpi	nokat	1	1	45	229	235		
46	,	Z	,	1	1	46	236	236		

Slika14. Prikaz izlaza u ReLDIanno označivaču

Prve su dvije rečenice iz romana *Povratak Filipa Latinovicza* Miroslava Krleže vrlo dugačke i sastoje se od dosta nabrajanja. Većina riječi koje se nabrajaju bi trebala biti u nominativu, no ponekad su pogrešno označene genitivom pa se tako u primjeru *miču se lica u povorkama, lica naprahana*, druga pojava lika *lica* označava s NCNSG umjesto NCNSN. Također, dolazi do grešaka u poklapanju roda pridjeva s imenicom koju označava zbog neuobičajene pozicije uzrokovane zarezom ili nabrajanjem bez glagola. To se može vidjeti u primjeru *sa crnim, modrikastim noktima* gdje se pridjev *crnim* (AGPFPIY) koji je odvojen zarezom od *modrikastim noktima*, ne podudara u rodu s imenicom *noktima* (NCMPI) dok se pridjev *modrikastim* (AGPMPIY) podudara. Isto tako, dolazi do primjera vanjske istopisnosti gdje je lik *teku* označen kao lokativ jednine imenice tek (NCMSL) umjesto kao treće lice množine glagola teći (VMR3P). Javljale su se i pogreške u jednini i množini što je vidljivo na primjeru *donje čeljusti* jer se pri nabrajanju misli na nominativ množine (AGPFPNY i NCFPN), a ne na genitiv jednine kako je označeno (AGPFSGY i NCFSG).

U pripovijetki *Prijan Lovro* Augusta Šenoe prva se pogreška dogodila s neprepoznavanjem lika *govorasmo*, odnosno imperfekta prvog lica množine glagola govoriti (VME1P) koji je označen kao imenica množine ženskog roda u nominativu (NCFPN). Također je čak dva puta došlo do

pogreške s označavanjem vlastitog imena Lovro. Prvi put je označeno kao RGP, a drugi put mu je pogrešno označen padež. U primjeru gledajući *ispod oka Lovru*, vlastito ime Lovro je u akuzativu (NPMSAY), a ne u dativu (NPMSD). Obrnuta stvar se dogodila s imenicom *mladoj* koja je u dativu (NCFSD), ali je bila označena u akuzativu i to u pogrešnom, srednjem rodu (NCNSA).

U znanstvenom je članku zamjenica *kojemu* bila označena lokativom (PI-NSL) iako se u danom kontekstu *tumačenje značenja kojemu su se priklonili semantičari* radi o dativu (PI-NSD). ReLDIanno označivač je na probleme naišao i u dijelu *unatoč svojem nedvosmislenom oslanjanju.* Prvo, pogrešna oznaka je dodijeljena prijedlogu *unatoč* jer iza sebe otvara mjesto imenskoj riječi u dativu (SD), a ne u genitivu (SG). Nakon toga je krivo označena zamjenica svojem kao imenica ženskog roda u genitivu (NCFSG) umjesto povratne zamjenice srednjeg roda u dativu (PX-NSD). Potom, je pridjev *nedvosmislenom* označen oznakom za lokativ (AGPNSLY), umjesto za dativ (AGPNSDY). Isto se ponavlja kod imenice *oslanjanju* koja ima oznaku NCNSL umjesto oznake NCNSD.

U novinskom članku portala *N1* i u kolokvijalnim izrazima nije bilo mnogo pogrešaka. Javile su se pogreške kod podudaranja roda pridjeva s rodom imenice koju označavaju. Ovo je vidljivo u primjeru *kulturnog, nacionalnog, jezičnog i vjerskog identiteta* u kojem su pridjevi *kulturnog*, *nacionalnog* i *jezičnog* označeni oznakama za srednji rod (AGPNSGY), umjesto za muški (AGPMSGY), kojeg je imenica *identiteta* koju označavaju. U kolokvijalnom izrazima probleme su stvarale same kolokvijalne riječi. U primjeru *On je cicija, zadnji će izvući novčanik* je imenica *cicija* pogrešno označena ženskim rodom (NCFSN) umjesto muškim (NCMSN). U primjeru *Tako krmi da ga ni topovi neće probuditi*, glagol *krmiti* tj. *biti u dubokom snu* je označen kao jednina lokativa (NCFSL) imenice *krma,* a ispravna bi oznaka bila VMR3S. Kolokvijalni glagol *nafrljiti* tj. *pojačati* je također označen kao imenica (NCFSL) u jednom primjeru, te kao pridjev (AGPMSNY) u drugom umjesto kao imperativ za drugo lice jednine (VMM2S). U primjeru *drži ga se k'o pijan plota* je došlo do dvije pogreške. Prva je kod lika *ga* koji je označen kao PP3MSA, no *držati se* iza sebe otvara mjesto imenskoj riječi u genitivu pa bi oznaka trebala biti PP3MSG.

U Tablici 4 se može vidjeti pojava pogrešaka uzrokovanih izazovima navedenima u odjeljku 2.2.2.:

Izazovi	Broj pogrešaka
Unutarnja istopisnost	16
Vanjska istopisnost	2

Tablica4. Prikaz broja pogrešaka na hrvatskom korpusu uzrokovanih homonimijom

Iz Tablice 4 se može iščitati kako je unutarnja istopisnost prouzrokovala osam puta više pogrešaka nego vanjska. Takvoj brojci govori u prilog dosta miješanja nominativa i genitiva množine te dativa i lokativa općenito.

4.4. Usporedba španjolskih morfosintaktičkih označivača s hrvatskim

Jedna od najvećih razlika je veličina skupa oznaka korištenih u pojedinom jeziku. Dok su u španjolskim označivačima korišteni skupovi oznaka Universal Dependencies projekta od 17 oznaka (Nivre, i sur., 2016) i skup od 75 oznaka u TreeTagger označivaču (TreeTagger, n.d.), u hrvatskom označivaču ReLDIanno koristi čak 933 različitih oznaka (Erjavec i Ljubešić, 2016). Ove brojke oslikavaju koliko visoko flektivni jezici moraju imati različitih oznaka kako bi se obuhvatili svi rodovi u svim padežima za imenske riječi i svi finitni oblici glagola. Ono što je također zapaženo je činjenica da kod odabranih španjolskih označivača nema oznaka za glagolska vremena ni glagolske načine. Da ih je bilo, broj oznaka u skupu bi se povećao. Oznake koje ne postoje u hrvatskom su oznake koje su posebno osmišljene u TreeTagger označivaču za spomenute skraćene oblike riječi i enklitike spojene na glagole. To su izazovi svojstveni španjolskom i unatoč tome što se zbog spajanja ili lijepljenja grafički radi o jednoj riječi, sintaktički se radi o dvije riječi sa svojim zasebnim funkcijama.

Istopisnost se kod oba jezika pokazala kao najveći od izazova u morfosintaktičkom označavanju. Najveća je razlika u tome što u hrvatskom jeziku ona ima više razina zbog postojanja padeža. Zanimljivo je da iako u oba jezika postoji rod, u španjolskim označivačima se on ne pojavljuje unutar ni jedne oznake pa tako do pogrešaka u podudaranju pridjeva i imenica u rodu nije dolazilo, dok je to bila jedna od češćih pogrešaka u hrvatskom jeziku. Čini se kako su još jedan od općenitih izazova predstavljale nepoznate riječi poput kolokvijalnih izraza *guay* na španjolskom ili *nafrljiti* na hrvatskom. Stanfordski POS označivač se koristio s nekoliko modela za pogađanje nepoznatih riječi uključujući „n-gram prefikse i sufikse za n do 4 i detektore za veliko početno slovo, crtice i brojeve“ (Toutanova i sur., 2003). Kod TreeTagger označivača riječi se pretražuju u leksikonu i ako nisu pronađene šalju se dalje u leksikon sufiksa koji riječima dodaje vjerojatnost oznake na temelju njihovih sufiksa (Schmid, 1995, 1994). ReLDIanno označivač i poznate i nepoznate riječi tretira na isti način koristeći „znanje iz

morfosintaktičkog leksikona indirektno u formi klasifikacijskih značajki kao što su token za sufikse od 1 do 4 ili token za MSD hipoteze na temelju sufiksa.“ (Erjavec i Ljubešić, 2016).

5. Zaključak

U radu su prikazani način rada i podjela POS i MSD označivača te izazovi s kojima se oni općenito susreću. Takvih izazova je mnoštvo i razlikuju se od jezika do jezika. Kako bi se te razlike prikazale, odabrani su primjeri španjolskog i hrvatskog jezika koji imaju mnoge razlike u pravopisu, sintaksi, ali i u razini fleksije.

Upravo je fleksija utjecala na razlike u izazovu koji se pojavljuje u oba jezika, a to je višeznačnost koja je inherentna svim jezicima općenito. Radi se o značajki jezika pod kojom se podrazumijeva da riječi koje se isto pišu mogu pripadati potpuno različitim vrstama riječi. U hrvatskom ona postoji i na unutarnjoj razini jer neke riječi mogu imati potpuno iste likove unutar svoje deklinacije, što je primjerice uvijek slučaj kod dativa i lokativa. Nadalje, korpusi sastavljeni od književnih tekstova, novinskih članaka, akademskih članaka te kolokvijalnih izraza neće imati isti vokabular. Testne rečenice nastojale su obuhvatiti književni izraz, starije jezične oblike koji više nisu u uporabi, formalni izraz, novinarski izraz te neformalni, razgovorni izraz. Starije riječi koje su dio pasivnog leksika kao i kolokvijalni izrazi su se pokazali kao jedan od izazova za označivače jer ti izrazi ne moraju uvijek činiti dio leksikona.

Kod španjolskih se označivača nije nužno uvijek radilo o pogreškama, već o različitim pristupima označavanju pojedinosti španjolskog jezika. Također, moguće je da su u slučaju Stanfordskog POS označivača ovisile o tokeniziranju u odabranoj NLTK zbirci.

Iako su na vrlo visokom stupnju točnosti, teško je pretpostaviti hoće li označivači ikada imati stopostotnu točnost, uzimajući u obzir da se jezik stalno mijenja te je nemoguće prikupiti baš sve riječi u leksikone i korpuse. Uz to uvijek postoje višeznačnost i nejasnoće koje odstupaju od pravila u samom jeziku. Unatoč tome, morfosintaktičko označavanje nastavlja biti jedan od temeljnih jezičnih alata za obilježavanje korpusa koji su neiscrpan izvor za usvajanje jezika i istraživanja u mnogim znanostima.

Literatura

1. Aggarwal, R. (2020, March 3). *Bi-LSTM - Raghav Aggarwal*. Medium. Dostupno na https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0
2. Agić, Ž., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., & Søgaard, A. (2016). Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*, *4*, 301–312. https://doi.org/10.1162/tacl_a_00100
3. Agić, Ž., Ljubešić, N., & Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing.* str. 48-57, Sofija, Bugarska. https://www.bib.irb.hr/638909
4. Agić, Ž., Tadić, M., & Dovedan, Z. (2008). Investigating Language Independence in HMM PoS/MSD-Tagging. *Proceedings of the 30th International Conference on Information Technology Interfaces.* str. 657-662, Cavtat/Dubrovnik, Hrvatska. Dostupno na https://www.bib.irb.hr/348726?&rad=348726
5. Alese, E. (2018, June 20). *The curious case of the vanishing & exploding gradient*. Medium. https://medium.com/learn-love-ai/the-curious-case-of-the-vanishing-exploding-gradient-bf58ec6822eb
6. Allen, J. (1995). *Natural language understanding* (2nd ed). The Benjamin/Cummings Publishing Co., Inc., USA
7. Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., Znika M. (2005). *Hrvatska gramatika*. Školska knjiga.
8. Bekavac, B. (2002). Strojno obilježavanje hrvatskih tekstova - stanje i perspektive. *Suvremena lingvistika, 53-54* (1-2), str. 173-182. Preuzeto s https://hrcak.srce.hr/16343
9. Bird, S., Klein, E., Loper, E. (2009). *Natural Language Processing with Python* (1st ed). O'Reilly Media, Inc.
10. Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., & Maynez, J. (2018). Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* str. 2642–2652, Melbourne, Australija. 10.18653/v1/P18-1246

11. Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*. str. 152–155, Trento, Italija. https://doi.org/10.3115/974499.974526
12. Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, *21*(4), str. 543-565. Preuzeto sa https://www.aclweb.org/anthology/J95-4004
13. Cao, K. & Rei, M. (2016). A Joint Model for Word Embedding and Word Morphology. *Proceedings of the 1st Workshop on Representation Learning for NLP* str. 18-26, Berlin, Njemačka. 10.18653/v1/W16-1603
14. Chatterjee, M. (2021, May 13). *Deep Learning Tutorial: What it Means and what's the role of Deep Learning*. GreatLearning Blog: Free Resources What Matters to Shape Your Career! Dostupno na https://www.mygreatlearning.com/blog/what-is-deep-learning/
15. Dayanand, T. (2020, Rujan 3). *POS Tagging Using RNN - Towards Data Science*. Medium. Dostupno na https://towardsdatascience.com/pos-tagging-using-rnn-7f08a522f849
16. Dicionario de la lengua española. (2014). cuantificador. U *Real Academia Española (RAE)*. Preuzeto 12. svibnja. 2021. s https://dle.rae.es/cuantificador?m=form
17. Dicionario de la lengua española. (2014). dello, lla. U *Real Academia Española (RAE)*. Preuzeto 8. svibnja. 2021 s https://dle.rae.es/dello?m=form
18. Dicionario de la lengua española. (2014). deste, ta. U *Real Academia Española (RAE)*. Preuzeto 8. svibnja. 2021 s https://dle.rae.es/deste
19. Divjak, D., Sharoff, S., & Erjavec, T. (2017). Slavic Corpus and Computational Linguistics. *Journal of Slavic Linguistics, 25*(2), str. 171-200. Preuzeto s https://www.jstor.org/stable/26535064
20. Erjavec, T. (2017). MULTEXT-East. U (Nancy Ide, James Pustejovsky, eds.): *Handbook of linguistic annotation.* str. 441-462. Springer. 10.1007/978-94-024-0881-2_17
21. Erjevec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M. & Vitas, D. (2003). The MULTEXT-east morphosyntactic specifications for Slavic languages. *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic.* str. 25-32, Budimpešta, Mađarska.
22. Gillon, B. (1990). Ambiguity, Generality, and Indeterminacy: Tests and Definitions. *Synthese, 85*(3), str. 391-416. Preuzeto s http://www.jstor.org/stable/20116854
23. Gómez Torrego, L. (2005), G*ramática didáctica del español.* Ediciones SM.
24. Hrvatska enciklopedija (n.d.). član. U *Hrvatska enckilopedija*. Preuzeto 21. travnja 2021. s https://www.enciklopedija.hr/Natuknica.aspx?ID=13443

25. Hrvatska enciklopedija (n.d.). član. U *Hrvatska enckilopedija*. Preuzeto 13. svibnja 2021. s https://www.enciklopedija.hr/Natuknica.aspx?ID=34869
26. Hrvatska enciklopedija (n.d.). deklinacija. U *Hrvatska enckilopedija*. Preuzeto 21. travnja 2021. s https://www.enciklopedija.hr/Natuknica.aspx?ID=14293
27. Hrvatski jezični portal (HJP) (n.d.). enklitika. U *Hrvatski jezični portal.* Preuzeto 21. travnja. 2021. s https://hjp.znanje.hr/index.php?show=search_by_id&id=fFxjURY%3D
28. Jurafsky, D., & Martin, J.H. (2009). *Speech and language processing* (2nd ed). Prentice Hall.
29. Kadłub, M. (2017). Sources of Ambiguity in Language. *Studia Anglica Resoviensia*. 14. str. 44-57. 10.15584/sar.2017.14.4
30. Kolokacijska baza hrvatskog (n.d.). O bazi. U *Kolokacijska baza hrvatskog*. Preuzeto 23. travnja 2021. s http://ihjj.hr/kolokacije/o-bazi/
31. Kumawat, D., & Jain, V. (2015). POS Tagging Approaches: A Comparison. *International Journal of Computer Applications*, *118*(6), str. 32–38. https://doi.org/10.5120/20752-3148
32. Lehmann, C. (2013). The nature of parts of speech. *Language Typology and Universals*. 66(2). 10.1524/stuf.2013.0008
33. Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., & Luís, T. (2015). Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* str. 1520-1530, Lisabon, Portugal. 10.18653/v1/D15-1176
34. Ljubešić, N. i Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia.
35. Ljubešić, N., Klubička F., Agić Ž., Jazbec, I.P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* str. 4264-4270, Portorož, Slovenia.
36. Ma, X. & Hovy, E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* str. 1064-1074, Berlin, Njemačka. 10.18653/v1/P16-1101
37. McEnery, T., & Wilson, A. (2001). *Corpus Linguistics* (2nd ed). Edinburgh University Press.

38. McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics, 39*, 74-92. doi:10.1017/S0267190519000096
39. Mujtaba, H. (2021, May 20). Part of Speech (POS) tagging with Hidden Markov Model. GreatLearning Blog: Free Resources What Matters to Shape Your Career! Dostupno na https://www.mygreatlearning.com/blog/pos-tagging/
40. MULTEXT-East. (n.d.). *Multext-East V6 "CLARIN."* Multext-East Resources. Preuzeto 23. svibnja 2021., s http://nl.ijs.si/ME/V6/
41. Nivre, J., Marneffe, M., C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation.* str. 1659–1666, Portorož, Slovenia. Dostupno na http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf
42. Orphanos, G., S., i Christodoulakis, D., N. (1999). POS Disambiguation and Unknown Word Guessing with Decision Trees. *Ninth Conference of the European Chapter of the Association for Computational Linguistics.* str. 134-141, Bergen, Norveška. Dostupno na https://www.aclweb.org/anthology/E99-1018
43. Parra Escartín, C., & Martínez A., H. (2015). Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task. *Procesamiento del Lenguaje Natural*, (54),29-36. 1135-5948. Dostupno na https://www.redalyc.org/articulo.oa?id=515751523003
44. Periñán Pascual, J. C. (2012). En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica. *Onomázein: Revista de Lingüística, Filología y Traducción*, (26), 13-48. http://hdl.handle.net/10251/45752
45. Petkevič, V. (2014). Ambiguity, language structures and corpora. *La Linguistique*, 50(2), str. 63-82. https://doi.org/10.3917/ling.502.0063
46. Petrov, S., Das, D., & McDonald, R. (2011). A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).* str. 2089–2096, Istanbul, Turska. http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

47. Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* str. 412-418, Berlin, Njemačka. https://doi.org/10.18653/v1/p16-2067
48. Pointal, L. (n.d.). treetaggerwrapper 2.3. U *The Python Package Index.* Preuzeto 26. travnja 2021. s https://pypi.org/project/treetaggerwrapper/
49. Quecedo, J. M. H., Koppatz, M. W., Furlan, G., & Yangarber, R. (2020). Neural disambiguation of lemma and part of speech in morphologically rich languages. *Proceedings of the 12th Language Resources and Evaluation Conference.* str. 3573–3582, Marseille, Francuska. Dostupno na https://www.aclweb.org/anthology/2020.lrec-1.439
50. Real Academia Española, Asociación de Academias la Lengua Española. (2010). *Nueva Gramática de la lengua española*. Espasa.
51. Samuelsson, C., Voutilainen, A., (1997). Comparing a Linguistic and a Stochastic Tagger. *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. str. 246-253, Madrid, Španjolska. 10.3115/976909.979649
52. Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
53. Schmid, H. (1995) Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
54. Schmid, H. (n.d.). TreeTagger - a part-of-speech tagger for many languages. U TreeTagger. Preuzeto 26. travnja 2021. s https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
55. Sennet, A. (2011). Ambiguity. U Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition). Metaphysics Research Lab, Stanford University. Dostupno na https://plato.Stanford.edu/archives/spr2016/entries/ambiguity/
56. Sutton, C., & McCallum, A. (2011). An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning.* 4(4), 267-373. https://doi.org/10.1561/2200000013
57. Tadić, M., (2003). *Jezične tehnologije i hrvatski jezik*. EX LIBRIS.
58. Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python. In *XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48,* Salta, Argentina.

59. Taulé, M., Martí, M. A., Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Maroko. Dostupno na: http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf
60. The Stanford Natural Language Processing Group (n.d.). CoreNLP Spanish FAQ. U *The Stanford Natural Language Processing Group.* Preuzeto 24. travnja 2021. s https://nlp.Stanford.edu/software/spanish-faq.html
61. The Stanford Natural Language Processing Group (n.d.). Stanford Log-linear Part-Of-Speech Tagger. U *The Stanford Natural Language Processing Group.* Preuzeto 24. travnja 2021. s https://nlp.Stanford.edu/software/tagger.shtml
62. Tobar, J. J. C., Solano, M. A. J., Sierra-M, L., & Cobos, C. A. L. (2020). Etiquetado de partes del discurso sobre un corpus en castellano basado en metaheurísticas. *Revista Ibérica De Sistemas e Tecnologias De Informação,* 215-228. Preuzeto s https://www.proquest.com/scholarly-journals/etiquetado-de-partes-del-discurso-sobre-un-corpus/docview/2452332522/se-2?accountid=202234
63. Toutanova, K. & Manning., C.D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, str. 63-70.
64. Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL 2003*, str. 252-259.
65. Van Guilder, L. (1995). Automated part of speech tagging: A brief overview. *Handout for LING361*. Dostupno na http://ccl.pku.edu.cn/doubtfire/NLP/Lexical_Analysis/Word_Segmentation_Tagging/POS_Tagging_Overview/POS%20Tagging%20Overview.htm
66. Voutilainen, A., (1995). A Syntax-Based Part-of-Speech Analyser. *EACL '95: Proceedings of the seventh conference on European chapter of the Association for Computational* str. 157–164, Dublin, Irska. https://doi.org/10.3115/976973.976996

Popis slika

Popis tablica

Popis grafikona

Prilozi

Prilog 1 – Programski kod u Python-u

```
import nltk
from nltk import *
from nltk.tokenize import word_tokenize
from nltk.tag.Stanfordski import StanfordskiPOSTagger

import os
java_path = "C:\Program Files\Java\jre1.8.0_281"
os.environ["JAVAHOME"] = java_path

spa_postagger = StanfordskiPOSTagger("models/spanish-ud.tagger", "Stanfordski-postagger.jar")
probna_rec = spa_postagger.tag("Esta es la oración de prueba.".split())

spa = open("recenice_spa.txt", encoding="utf8").read()
rijeci = nltk.word_tokenize(spa)
oznacene_rijeci = spa_postagger.tag(rijeci)

print(probna_rec)
print(oznacene_rijeci)

import treetaggerwrapper
import os

treetagger_path= "C:\TreeTagger"
```

```
os.environ["TreeTagger"]= treetagger_path


spa_postagger = treetaggerwrapper.TreeTagger(TAGLANG="es")

probna_rec = spa_postagger.tag_text("Quiero contarselo.")


spa= open("recenice_spa.txt", encoding="utf8").read()

oznacene_rijeci = spa_postagger.tag_text(spa)


print(probna_rec)

print(oznacene_rijeci)
```

Prilog 2 – Sadržaj datoteka recenice_spa.txt i recenice_hrv.txt na kojima su testirani označivači

Dejamos en la primera parte desta historia al valeroso vizcaíno y al famoso don Quijote con las espadas altas y desnudas, en guisa de descargar dos furibundos fendientes, tales, que, si en lleno se acertaban, por lo menos se dividirían y fenderían de arriba abajo y abrirían como una granada; y que en aquel punto tan dudoso paró y quedó destroncada tan sabrosa historia, sin que nos diese noticia su autor donde se podría hallar lo que della faltaba.

Sonrió de pronto, como si acabara de ocurrírsele algo divertido, y se disponía a seguir hablando cuando oyó a su espalda las voces de su padre y de su tío Javier; ninguno de los dos, a juzgar por sus risas, hablaba de los desmanes cometidos en la valla por las parejas domingueras e impúdicas que invaden las propriedades privadas. Maruja se levantó antes de que llegaran y fue a reunirse con los niños. Teresa comprendió que se iba para que no vieran que había llorado.

Como síntesis a este apartado, en la segunda etapa, se han producido múltiples investigaciones desde diversas posiciones teóricas y desde diversas áreas como la dialectología, la semántica y la lexicografía. Se tratan los campos semánticos, se construyen teorías sobre los procesos de evasión lingüística, las metáforas, etc. Las investigaciones relacionadas con la sociolingüística han sido las de mayor impacto, ya que introducen el contexto y la variación lingüística como elementos que aportan una visión más funcional de los factores que determinan el uso del tabú lingüístico.

Casi tres años más tarde, en enero de 2020, el Constitucional dictó una sentencia en la que estimaba que el Supremo vulneró el derecho de los integrantes del grupo ultra a la tutela judicial efectiva. El tribunal de garantías consideró que esa vulneración de derechos fundamentales se produjo al elevarse la pena inicialmente impuesta sin que los condenados hubieran sido oídos por la Sala Penal en la vista del recurso.

Ese piso me ha gustado mogollón.

Es una moto guay. Ese tío es guay; es súper simpático y generoso. Me lo estoy pasando guay en tu fiesta.

La chica se enfadó y le metió una buena hostia al chico.

Álvaro es un pijo asqueroso; ¿te has fijado cómo ha tratado al camarero?

Teku ljudi po ulicama, miču se lica u povorkama, lica naprahana, blijeda, klaunska, sa zarezima gorućeg karmina oko usana, kratkovidne maske žena u crnini, lica grbavaca, donje čeljusti, voštani dugi prsti sa crnim, modrikastim noktima, sve prilično ružno. Gadna lica, zvjerske njuške, žigosane bludom i porocima, zlobom i brigama, lica smolava i ugrijana, glave mrkvaste, gubice crnačke, zubala tvrda, oštra, mesožderska, a sve je sivo kao fotografski negativ.

Govorasmo o svem i svačem, kao ljudi kad se prvi put vide. Mene začudo hvalili, na veliku moju nepriliku. Lovro govorio mladoj sve u pol glasa u kratkim riječima, a ona mu isto tako odgovarala. Starci se kriomice gurkali gledajući ispod oka Lovru. Govorilo se slovenski, ali tako da mi se je činilo, da ti mili ljudi ne govore obično svojim jezikom.

Tumačenje značenja kojemu su se priklonili S. Ullmann i semantičari skloni semičkoj analizi udaljilo ih je od temeljnih načela Saussureova učenja. Uključivanje stvari i znanja ili misli o njoj kao relevantnih kategorija u definiranju značenja, ili drugim riječima, nemogućnost da se stvar izbaci iz semantičke analize pokazala je da strukturalna semantika unatoč svojem nedvosmislenom oslanjanju na Saussureov modernizam nije uspjela taj modernizam provesti do kraja.

Vlada je donijela i uredbu o financiranju javnih potreba nacionalnih manjina koje se ostvaruju kroz programe i projekte udruga nacionalnih manjina i financiraju iz državnog proračuna radi unaprjeđenja prava pripadnika nacionalnih manjina te zaštite i promicanja kulturnog i nacionalnog, jezičnog i vjerskog identiteta.

Koji si car, kak si se to sjetio!

On je cicija, zadnji će izvući novčanik.

Otkad je upoznao tog tipa, drži ga se k'o pijan plota!

Tako krmi da ga ni topovi neće probuditi.

Nafrlji radio, svira moja stvar! Hladno je, nafrlji tu grijalicu!

Izazovi morfosintaktičkog označavanja na primjerima španjolskog i hrvatskog jezika

Sažetak

Posljednjih se desetljeća sve brže razvijaju jezične tehnologije kao težnja da se jezik kao sredstvo ljudske komunikacije kodira. One ubrzavaju usvajanje jezika, ali i olakšavaju lingvistička, psihološka i ostala istraživanja. Vrlo važnu ulogu za razvoj kvalitetnih i točnih jezičnih tehnologija imaju korpusi koji se označavaju na različitim razinama, a jedna od njih je dodavanje POS ili MSD oznake riječi s obzirom na stupanj flektivnosti jezika. U ovom radu će se opisati POS i MSD označavanje i navesti poteškoće koji se općenito javljaju pri tom procesu. Kako bi primjeri bili konkretni odabran je španjolski jezik kao manje flektivni jezik i hrvatski jezik kao visoko flektivni jezik. Španjolski će biti testiran na primjerima Stanfordskog POS označivača i TreeTagger označivača putem sučelja programskog jezika Python. Isprobat će se rečenice na španjolskom iz raznih registara i s mogućim poteškoćama za označivače. Isti proces bit će ponovljen na web sučelju ReLDIanno označivača za hrvatski jezik. Usporedit će se dobiveni rezultati između španjolskih označivača i razlike u problemima kod označavanja manje flektivnih i visoko flektivnih jezika na primjerima španjolskog i hrvatskog.

Ključne riječi: morfosintaktičko označavanje, španjolski, hrvatski, POS oznake, MSD oznake, POS označivači

Challenges in POS/MSD tagging with Spanish and Croatian examples

Summary

During the last decades language technologies have developed at a fast rate with the purpose of coding language as a means of human communication. They are helping with the process of learning a language as well as with investigations in many fields such as linguistics and psychology. Corpora play a big role in developing high-quality and correct language technologies. They can be tagged on a few different levels, one of which is assigning POS or MSD tag to a word considering the grade of inflection in certain language. In this BA thesis POS and MSD tagging will be described as well as certain difficulties that occur during that process in general. To make the examples more concrete, the Spanish language is chosen as a language with less inflection and the Croatian language is chosen as language with high inflection. Spanish will be tested by Stanford POS Tagger and TreeTagger with the Python programming language. Spanish sentences from different registers which present possible challenges will be tested. The same process will be repeated with web interface of ReLDIanno tagger for Croatian. The results from Spanish and Croatian taggers and differences in problems between Spanish and Croatian will be compared.

Key words: morphosyntactic tagging, Spanish, Croatian, POS tags, MSD tags, POS taggers