Automatsko označavanje i analiza teksta za LARSP hrvatskog jezika

Tot, Bruno

Master's thesis / Diplomski rad

2019

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:378134

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2025-03-21**

Sveučilište u Zagrebu
Filozofski fakultet
University of Zagreb
Faculty of Humanities and Social Sciences

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb Faculty of Humanities and Social Sciences

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

Sveučilište u Zagrebu

Filozofski fakultet

Odsjek za informacijske i komunikacijske znanosti

Odsjek za fonetiku

Bruno Tot

AUTOMATSKO OZNAČAVANJE I ANALIZA TEKSTA ZA LARSP HRVATSKOG JEZIKA

Diplomski rad

Zagreb, studeni 2019.

Sveučilište u Zagrebu

Filozofski fakultet

Odsjek za informacijske i komunikacijske znanosti

Odsjek za fonetiku

Bruno Tot

AUTOMATSKO OZNAČAVANJE I ANALIZA TEKSTA ZA LARSP HRVATSKOG JEZIKA

Diplomski rad

Mentori:

prof. dr. sc. Nikolaj Lazić

prof. dr. sc. Vesna Mildner

Zagreb, studeni 2019.

PODACI O AUTORU

Ime i prezime: Bruno Tot

Datum i mjesto rođenja: 14.07.1993., Sinj, Republika Hrvatska

Studijske grupe i godina upisa: Fonetika (znanstveni smjer) i Informacijske i komunikacijske znanosti (smjer informatika: istraživačka), ak. god. 2016/2017

Lokalni matični broj studenta: 028245

PODACI O RADU

Naslov rada na hrvatskome jeziku:
Automatsko označavanje i analiza teksta za LARSP hrvatskog jezika

Naslov rada na engleskome jeziku:
Automatic text markup and analysis for LARSP of Croatian language

Broj stranica:

Broj priloga:

Datum predaje rada:

Sastav povjerenstva koje je rad ocijenilo i pred kojim je rad obranjen:

1. doc. dr. sc. Diana Tomić

2. prof. dr. sc. Vesna Mildner

3. prof. dr. sc. Nikolaj Lazić

Datum obrane rada:

Broj ECTS bodova: 30 (15 na studiju fonetike + 15 na studiju informacijskih znanosti)

Ocjena:

Potpis članova povjerenstva:

1. ----------
2. ----------
3. ----------

IZJAVA O AUTORSTVU DIPLOMSKOGA RADA

Ovim potvrđujem da sam osobno napisao diplomski rad pod naslovom

Automatsko označavanje i analiza teksta za LARSP hrvatskog jezika

i da sam njegov autor.

Svi dijelovi rada, podaci ili ideje koje su u radu citirane ili se temelje na drugim izvorima (mrežni izvori, udžbenici, knjige, znanstveni, stručni članci i sl.) u radu su jasno označeni kao takvi te su navedeni u popisu literature.

Bruno Tot

_____

Zahvale

Ovim putem zahvaljujem Elizabeti Kantoci na pomoći oko oblikovanja ovog rada, a još više na suradnji na neobjavljenom radu koji je poslužio kao temelj ovoga rada, Terezi Sente na izradi testnih materijala te savjetima koji su uvelike pomogli oko stvaranja algoritma za analizu hrvatskoga jezika, mentorima na mentoriranju te svima koji su na ovaj ili onaj način doprinijeli ovom radu, a nisu imenovani u ovoj zahvali.

SADRŽAJPODACI O AUTORU **3**
1. UVOD **2**
2. O LARSP-u **3**
2.1. Metode 3
2.2. Obrazac 3
2.3. Procedura 6
2.4. Orijentiranost na gramatiku 7
3. LARSPTool – ALAT ZA ANALIZU TEKSTA PREMA LARSP-U **8**
3.1. Temelj i predradnje 8
3.1.1. Eksperimentalni sustav označavanja teksta 9
3.1.2. Kako je program funkcionirao? 10
3.2. Pripremne radnje 12
3.3. LARSPTool 12
3.3.1. Priprema teksta za analizu 12
3.3.2. Rad u alatu 13
3.3.3.1. Stvaranje novog projekta 15
3.3.3.2. Glavni prozor 21
4. ZAKLJUČAK **28**
5. REFERENCIJE **29**
SAŽETAK **31**
ABSTRACT **31**

UVOD

Govor je jedan od ključnih načina ljudske komunikacije. Govor je komunikacijski kanal putem kojeg jezik postaje više od niza riječi. Intonacijom i intenzitetom može se prenijeti još više informacija nego samim riječima. Emocije, stanja i pravo značenje izrečenog manifestiraju se u suprasegmentalnim obilježjima govora. Kao i svaka druga ljudska vještina, govor i jezik se razvijaju od rođenja - prvo slušanjem, a onda i repliciranjem. Baš kao i svaka druga ljudska karakteristika, podložni su raznim poremećajima i oštećenjima. Neki od govorno-jezičnih poremećaja se mogu rehabilitirati zbog čega je ključno detektirati ih u što ranijoj dobi dok se govor i jezik razvijaju kako bi rehabilitacija bila što učinkovitija. U tu svrhu su se razvile brojne metode dijagnosticiranja govorno-jezičnih poremećaja i njihovog rehabilitiranja. Jedna od tih metoda je i LARSP koji objedinjuje nekoliko kliničkih postupaka kako bi se osigurala precizna dijagnostika i uspješna rehabilitacija dječjeg govora i jezika. Za točnu dijagnostiku potrebna je detaljna analiza. Ona se za LARSP vrši na transkribiranom dječjem govoru, snimljenom u propisanim kontekstima i uz pomoć standardiziranih materijala. U svrhu ubrzanja i olakšanja analize stvoren je alat LARSPTool koji osim brže analize većeg korpusa, uvodi i novi standard i jednoličnost prikaza podataka što znatno pridonosi učinkovitosti ove metode. Sam protokol i alat opisani su u daljnjem tekstu.

2. O LARSP-u

2.1. Metode

LARSP je lingvistički protokol koji objedinjuje tri vrste kliničkih postupaka: procjenu (engl. *assessmnet*), rehabilitaciju (engl. *remediation*) i probir tj. praćenje (engl. *screening*) kako bi se dobio što detaljniji uvid u gramatiku[1] dječjeg spontanog govora u određenoj dobi. Crystal (1979) navodi da se prvi klinički postupak-procjena odnosi na određivanje djetetovog jezičnog profila na temelju analize jezičnih struktura koje ono koristi (npr. koje vrste i oblike riječi dijete koristi, koje vrste rečenica koristi, poklapaju li se gramatičke kategorije sintagmi u rečenicama i slično). Ako se procijeni da se djetetov jezični status ne podudara s njegovom dobi, primjenjuje se drugi klinički postupak - rehabilitacija, odnosno terapija. Treći postupak (engl. *screening*) uključuje praćenje djetetovog jezičnog statusa uslijed terapije da bi vidjeli je li došlo do napretka i u kojim područjima, treba li promijeniti metodu terapije, čemu treba posvetiti pažnju itd. (Crystal, 1979). LARSP je, prema Crystalu (1979), prvi lingvistički protokol koji objedinjuje ta tri klinička postupka. Do tada su procedure podrazumijevale samo jednu od procedura pa nisu mogle pružiti sveobuhvatni opis djetetovog jezičnog stanja. Procedure koje se koriste samo metodom procjene (npr. Reynellov test ili ITPA test) pružaju mnogo podataka o izražavanju i razumijevanju nekog djeteta, ali one ne daju konkretni razlog terapije ni smjer u kojem bi terapija trebala ići (Crystal, 1976). Procedure koje koriste samo metodu rehabilitacije dat će puno smjernica za terapiju, ali neće poslužiti za procjenu ili praćenje te neće pružiti uvid u stupanj jezičnog razvoja pojedinog djeteta (Crystal, 1976).

2.2. Obrazac

Da bi protokol funkcionirao, odnosno sve tri metode tog protokola, važno je imati dobru bazu na temelju koje će se vršiti procjena. Takva baza treba imati dvije bitne karakteristike: a) mora pružati sveobuhvatan opis djetetovog gramatičkog *outputa* u bilo kojoj fazi procjene, terapije i praćenja i b) mora osigurati dosljednu terapijsku metodologiju (Crystal i sur., 1989). Prva karakteristika znači da ispitivač mora znati koje jezične strukture dijete još nije savladalo,

[1] Crystal i sur.(1979) pod gramatikom podrazumijevaju sintaksu i morfologiju

a trebalo je s obzirom na svoju dob. Važno je pratiti tijek rehabilitacije kako bismo znali kakav je napredak, što bi trebalo promijeniti u rehabilitaciji, možemo li što ignorirati, a na što trebamo obratiti pozornost. Druga karakteristika znači da nakon završene procjene i određene terapije, rehabilitator mora znati kako organizirati terapiju te na koji način i kojim redoslijedom uvoditi nove strukture. Također, rehabilitator mora znati kako reagirati ako korisnik ne usvaja strukture na način koji je zamišljen itd. (Crystal i sur., 1989).

Baza s ranije opisanim karakteristikama naziva se LARSP obrazac (slika 1). U počecima je takav obrazac bio napravljen samo za engleski jezik, ali je ubrzo izrađen i za neke druge jezike kojima je obrazac za engleski jezik poslužio kao primjer. Novoizrađen je i obrazac za hrvatski (Mildner i sur., 2019). LARSP obrazac sastoji se od 2 dijela. Prvi dio podijeljen je na 4 kategorije (A, B, C i D) i odnosi se na interakciju ispitivača i ispitanika. „A“ dio sadrži broj neanaliziranih rečenica i rečenica koje nisu pogodne za analizu. U „B“ dijelu bilježi se broj odgovora koje je potaknuo ispitivač (engl. *responses*) i njihov oblik (eliptični, skraćeni, nelogični odgovor...). U „C“ dijelu bilježi se broj spontanih odgovora (onih koji nisu odgovoreni samo jednom rečenicom nego su prošireni). „D“ dio obrasca sadrži ispitivačevu reakciju na odgovor ispitanika (potvrda točnog odgovora, direktno ispravljanje netočnog odgovora, ponavljanje netočnog odgovora kako bi se potaknulo samostalno ispravljanje itd.). (Crystal, 1979).

Drugi dio obrasca odnosi se na analizu djetetovih rečenica, sintagmi i riječi. Taj dio obrasca podijeljen je na sedam faza, a svaka faza označava dob djeteta i traje pola godine, odnosno jednu godinu u kasnijim fazama. Prva faza obuhvaća razdoblje od 9 do 18 mjeseci starosti, druga faza (1;6 – 2;0), treća faza (2;0 – 2;6), četvrta faza (2;6 – 3;0), peta faza (3;0 – 3;6), šesta faza (3;6 – 4;6), sedma faza (4;6 +). U prvih pet faza analiziraju se 4 elementa: 1) funkcionalna vrsta rečenice, 2) sintaksa rečenice, 3) analiza sintagmi u rečenici, 4) vrsta i oblik riječi. U šestoj fazi navedene su samo konstrukcije i najčešće pogreške koje se pojavljuju u toj dobi, a sedma faza odnosi se na diskurs i stil. Za svaku fazu u obrascu su navedene jezične strukture koje bi trebale biti usvojene u toj dobi, a one su dobivene analiziranjem velikog broja dječjih spontanih govora u svim fazama prilikom izrade LARSP obrasca za engleski jezik (Crystal, 1979). Jezični profil djeteta dobiva se tako da se prebroje sve gramatičke strukture koje se pojave u njegovom spontanom govoru tijekom snimanja i usporede sa strukturama koje su na obrascu navedene kao očekivane za njegovu dob. Ako se većina proizvedenih struktura poklapa s očekivanima – jezični razvoj je uredan.

1. dio

2. dio

Name | Age | Sample date | Type

A Unanalysed: 1 Unintelligible; 2 Symbolic Noise; 3 Deviant

Problematic: 1 Incomplete; 2 Ambiguous; 3 Stereotypes

B Responses

Stimulus Type	Totals	Repetitions	Normal Response: Major: Elliptical 1	Elliptical 2	Elliptical 3+	Reduced	Full	Minor	Abnormal: Structural	∅	Problems
Questions											
Others											

C Spontaneous

D Reactions: General; Structural; ∅; Other; Problems

Stage I (0;9–1;6): Minor — Responses; Vocatives; Other; Problems. Major — Comm. 'V'; Quest. 'Q'; Statement 'V', 'N'; Other; Problems

Stage II (1;6–2;0): Conn.; Clause: VX; QX; SV; AX; SO; VO; SC; VC; Neg X; Other. Phrase: DN; VV; Adj N; V part; NN; Int X; PrN; Other. Word: -ing; pl; -ed; -en; 3s; gen

Stage III (2;0–2;6): X + S:NP; X + V:VP; X + C:NP; X + O:NP; X + A:AP. Clause: VXY; let XY; do XY; QXY; VS(X); SVC; SVO; SVA; VCA; VOA; VO_dO; Neg XY; Other. Phrase: D Adj N; Adj Adj N; Pr DN; Pron; Cop; Aux_0^M; Other. Word: n't; 'cop; 'aux

Stage IV (2;6–3;0): XY + S:NP; XY + V:VP; XY + C:NP; XY + O:NP; XY + A:AP. Clause: + S; VXY +; QVS; QXY +; VS(X) + 1; tag; SVOA; SVCA; SVO_dO; SVOC; AAXY; Other. Phrase: NP Pr NP; Pr D Adj N; cX; Xc X; Neg V; Neg X; 2 Aux; Other. Word: -est; -er; -ly

Stage V (3;0–3;6): and; c; s; Other. Coord.; Other. Coord.; Other. Coord. 1; 1 +; Subord. A 1; 1 +; S; C; O; Comparative. Postmod. clause 1; 1 +; Postmod. phrase 1 +. (+) (−)

Stage VI (3;6–4;6): NP: Initiator; Coord. VP: Complex. Clause: Passive; Complement; how; what. Conn.: and; c; s. Clause: Element ∅; ⇄; Concord. Phrase: NP: D; Pr; Pron; D∅; Pr∅; D⇄; Pr⇄. VP: Aux^M; $Aux^{\neq}$; Cop; ∅. Word: N; V; neg; ~neg. Other; Ambiguous

Stage VII (4;6+): Discourse: A Connectivity; Comment Clause; Emphatic Order; n; there; Other. Syntactic Comprehension; Style

Total No. Sentences | Mean No. Sentences Per Turn | Mean Sentence Length

1 D. Crystal, P. Fletcher, M. Garman, 1989, revised

1 2 3 4

Slika 1. LARSP obrazac za engleski jezik (preuzeto iz Crystal, 1989), dijelovi označeni brojevima objašnjeni su u poglavlju „LARSP obrazac“

2.3. Postupak

Kako bismo odredili jezični profil djeteta, potrebno je proći kroz proceduru koja se sastoji od sedam stadija:

1) prikupljanje uzorka
2) transkripcija
3) gramatička analiza
4) profiliranje
5) interpretacija
6) i 7) tijek i ciljevi terapije (Crystal, 1989).

1) Autori savjetuju prikupljanje uzorka u dva dijela od kojih svaki traje 15 minuta. U prvom dijelu vodi se spontani dijalog s djetetom o nekoj trenutnoj situaciji (npr. igra), a u drugom dijelu događa se vođeni dijalog o nekoj prošloj situaciji ili se dječji odgovori potiču slikama koje treba opisati ili objasniti.

2) Nakon što je materijal snimljen, sve izrečeno potrebno je transkribirati. Transkribirati se treba svaka djetetova i ispitivačeva rečenica, a ako dijete puno koristi „govor tijela“ potrebno je i to navesti u didaskalijama, kako bismo olakšali kasniju analizu (npr. događaji oko djeteta koji su prouzročili određenu reakciju).

3) Od ukupnog broja rečenica, izbacuju se rečenice nepogodne za analizu (nedovršene, otpjevane i sl.), a zatim se ostale rečenice sintaktički i morfološki analiziraju te se broji koliko je puta koja struktura iskorištena.

4) Dobiveni se rezultati uspoređuju s podatcima za određenu dob na LARSP obrascu.

5) Donosi se odluka o tome je li djetetov jezični razvoj uredan i, ako nije, zahtjeva li rehabilitaciju.

6) i 7) S obzirom na djetetov profil se određuju ciljevi i smjer rehabilitacije.

2.4. Orijentiranost na gramatiku

Crystal i sur. (1989) strukturu jezika podijelili su u skladu s podjelom većine lingvističkih teorija na: izgovor (u koji se ubrajaju fonetika i fonologija), gramatiku (morfologija i sintaksa) i značenje (semantika). LARSP je orijentiran isključivo na gramatiku, za razliku od lingvističkih procesa prije LARSP-a koji su se najčešće bavili semantikom. Autori smatraju da bi bez gramatike jezik bio „nedosljedna mješavina riječi i zvukova“ (1989: 14) te da je jedino na temelju gramatike moguće napraviti dosljedan obrazac za analizu jezika koji se može koristiti u rehabilitacijske svrhe. To je i za očekivati s obzirom na to da je semantika podložnija interpretacijama (koje ovise o kontekstu, regionalnoj i socijalnoj pripadnosti itd.) te bi bilo teško pronaći korpus riječi koje bi služile za određivanje nečije jezične sposobnosti. Gramatika je određenija pravilima te se na temelju gramatičke analize mogu donijeti nedvosmisleni zaključci.

3. LARSPTool – ALAT ZA ANALIZU TEKSTA PREMA LARSP-U

3.1. Temelj i predradnje

Ovaj rad, kao i sam alat, temelji se na prethodno napisanom, neobjavljenom istraživačkom seminarskom radu iz 2016. (Tot, B.; Kantoci, E., 2016) Za potrebe tog rada snimljeno je i analizirano dvoje trogodišnje djece. Zbog veličine korpusa i želje autora za njegovom cjelokupnom analizom smišljen je sustav ručnog označavanja teksta te je izrađen i jednostavan program za njegovo prikupljanje, sortiranje i jednostavnu analizu. U nastavku slijedi opis tada izrađenog programa koji je preteča LARSPToola.

Sav snimljeni materijal bio je detaljno transkribiran i transkripti su bili pažljivo oblikovani. Sljedeći korak je bila analiza teksta i prikupljanje relevantnih podataka o broju rečenica, njihovoj vrsti i strukturi. Također je bilo potrebno odrediti vrstu svake riječi te njezin rod, broj, padež i ostala svojstva te sve navedene podatke sortirati po govorniku i u odgovarajuće kategorije radi lakše međusobne usporedbe govornika. Uzimajući u obzir sam opseg istraživanja, a samim time i količinu materijala za analizu, počeli su se razmatrati načini ubrzanja postupka bez gubitka kvalitete i točnosti. Budući da je sam postupak za engleski jezik već automatiziran zahvaljujući razvijenom programu koji automatski profilira tekst (Long, 2012), istovjetno rješenje činilo se najlogičnijim i u ovom slučaju. Međutim, kako je LARSP za hrvatski jezik tada još bio u izuzetno ranoj fazi (Mildner i sur., 2019), gotovo rješenje nažalost nije bilo dostupno. Stoga je bilo potrebno krenuti s izradom programa prilagođenog za hrvatski jezik. Ograničavajući faktori bili su vremenski rok za provedbu istraživanja i obradu podataka, ali i iskustvo timskog programera. Iz tih razloga za izradu samog programa bio je odabran *python* kao programski jezik. *Python* je omogućio relativno brzu izradu i dobre rezultate jer je i inače pogodan za obradu kako teksta, tako i jezika.

3.1.1. Eksperimentalni sustav označavanja teksta

Prije negoli se tekst mogao učitati u program bilo je potrebno obaviti nekoliko pripremnih koraka. Prvo je trebalo osmisliti prikladan sustav označavanja teksta. Bilo je iznimno važno da oznake budu što intuitivnije i lako čitljive golim okom kako bi se lakše mogle ispraviti eventualne pogreške nastale u samom procesu označavanja. Nadalje, rani testovi pokazali su da je bilo potrebno napraviti distinkciju između određenih cjelina – poglavito između rečeničnih oznaka i oznaka za pojedine riječi. Sustav je također, ako je bilo moguće, trebao omogućiti i ponavljanje istih oznaka u različitim kontekstima koje bi poprimale različito značenje ovisno o kontekstu tako da se izbjegne preklapanje i prepisivanje podataka prilikom prikupljanja i pohrane. Time bi se omogućilo korištenje čak i jednoslovnih kratica za označavanje (npr. i za imenicu i instrumental) što bi doprinijelo ubrzavanju samog postupka. Važno je napomenuti da se u ovom slučaju označavanje teksta odnosilo na ručni rad.

„Određena su sljedeća pravila označavanja:

1. Rečenične oznake se nalaze prije prve riječi rečenice, unutar dviju okomitih crta - ||.
2. Oznake riječi nalaze se neposredno prije svake riječi i pišu se unutar oznaka - <>, nakon čega bez razmaka mora slijediti riječ.
3. Riječi se označuju tako da se prvo pišu oznake vrste riječi pa tek onda svojstva (rod, broj...) i taj se redoslijed mora poštovati.
4. Oznake istog reda (vrsta i svojstvo su dva različita reda) međusobno se odvajaju zarezom bez razmaka.
5. Redovi oznaka obavezno se međusobno odvajaju spojnicom, bez razmaka (vrsta-svojstvo).
6. Oznake su zasad proizvoljne, ali moraju sadržavati samo mala slova i/ili brojeve i ne smiju sadržavati dijakritike, interpunkciju ili bilo koje znakove izvan ASCII kodne liste.
7. Privremeno, govornik se označuje samo jednom po bloku, samo na početku teksta, jednim velikim slovom (ASCII) nakon kojeg slijedi dvotočka i obavezan jednostruki razmak.
8. Ako riječ nema značajnih svojstava, pri označavanju svojstava obavezno nakon spojnice odmah ide kosa crta - /.

9. Oznaka praznog svojstva (/) je dopuštena samo na prvom mjestu, umjesto prve oznake.“ (Kantoci, Tot, 2016)

3.1.2. Kako je program funkcionirao?

Prva, eksperimentalna inačica programa sastojala se od 6 funkcija: 3 glavne koje su vršile samo prikupljanje i sortiranje i 3 koje su radile statistiku prema različitim načelima i zadanim kriterijima. Budući da je u ovom slučaju tekst bio transkribiran u Wordov dokument, program je bio prilagođen za čitanje .docx datoteka koristeći „python-docx“ modul. Ta se odluka pokazala olakšavajućom. Naime, svaki dio teksta bio je zapisan u zaseban odlomak (tj. blok kako ga se naziva i u daljnjem tekstu) što se pokazalo kao savršena struktura za ovo istraživanje jer jedan blok sadrži više rečenica. Blokovi u kojima prema prethodno navedenim pravilima nije bio naveden govornik zanemareni su pri prolasku programa kroz datoteku. Svi blokovi koji su imali govornika bili su dodani u listu za daljnju obradu, osim blokova u kojima je govornik bila sama ispitivačica. Ta prva faza obrade pokazala je da se ukupno radilo o 538 valjanih blokova za analizu. Po završetku prikupljanja blokova inicijalizirale su se strukture u koje su se podaci pohranjivali u drugoj fazi. Pohrana se vršila po principu ključ-vrijednost što je omogućavalo relativno jednostavno pretraživanje i izradu statistike prikupljenih rezultata. Potom se izvršavala primarna funkcija koja je iterirala redom kroz novonastalu listu blokova. Za svaki blok prvo se prema pravilima odredio govornik te se, ako taj govornik prethodno nije već viđen, stvorila nova kategorija u strukturi gdje su se prikupljale pune rečenice. Nakon određivanja govornika slijedile su rečenice. U programu, rečenica je počinjala rečeničnom oznakom i završavala novom rečeničnom oznakom ili krajem bloka. Treba naglasiti da ako se svaka sastavnica složene rečenice označila vlastitom oznakom, složena rečenica bi bila rastavljena i svaka označena sastavnica bi se tretirala kao nova rečenica, ali zbog sustava označavanja to ne bi stvaralo problem pri izradi statistike. Isto tako, kako bi se izbjeglo prethodno spomenuto preklapanje, svakoj rečenici je dodijeljen i kontrolni broj koji bi se povećao za jedan svaki puta kada bi program našao novu rečenicu. Ključ pri pohrani je bila sama rečenična oznaka uz koju je stajao kontrolni broj, a vrijednost ključa bila je sama rečenica.

Nakon pohrane rečenice, prije nego bi se prešlo na slijedeću, program bi pozvao drugu glavnu funkciju – funkciju koja je analizirala pojedine riječi. Ta funkcija je „vidjela“ isključivo riječi koje su imale oznake ispred sebe što je bilo korisno ako su se iz izračuna željele izbaciti

nerazumljive riječi ili nešto slično. Prvo se prema pravilima stvorila privremena lista riječi onim redom kojim su bile u rečenici. Zatim se za svaku riječ vršila provjera za složena vremena. Program je provjeravao postoje li u oznakama ključne riječi kao što su pomoćni glagol, perfekt, futur, itd. koje su ukazivale na neko složeno vrijeme (futur, perfekt... jer se sastoje od više riječi). Da bi provjera vratila pozitivan rezultat, morale su postojati barem dvije riječi s odgovarajućim komplementarnim oznakama. U slučaju da je rezultat provjere bio negativan, vršilo se standardno sortiranje. Ponovo se prolazilo kroz listu riječi te bi se prema prethodno navedenim načelima odvajale oznake vrste, svojstva i same riječi. Vrste i svojstva su se pohranjivala svaka u svoju strukturu pod kategorijom govornika koji je bio preuzet iz nadređene funkcije. Vrste i svojstva uz kontrolni broj rečenice činile su ključeve svojih struktura, a vrijednosti su im bile same riječi. S druge strane, ako se prvom provjerom ipak utvrdila prisutnost složenog vremena, izvršavala bi se i posljednja glavna funkcija kojoj je zadaća bila izdvojiti sve komponente složenog glagolskog vremena i pravilno ih grupirati. Funkcija je ponovo prolazila kroz listu riječi i kad bi naišla na odgovarajuću riječ, zabilježila bi njen indeks (poziciju) u listi te ju dodala u vlastitu strukturu, jednaku onima gdje su se pohranjivali ostali podaci, a koja se još zove i rječnik. Ključ za tu riječ je bila njena pozicija, a vrijednost sama riječ. Svojstva riječi i njena vrsta su se pohranile u rječnik pod zasebnim ključevima. Kada bi se prikupile sve riječi, proslijedile bi se spojene u cjelinu nadređenoj funkciji koja bi ih zatim pohranila u rječnike po vrsti i svojstvu zajedno s rečeničnim kontrolnim brojem uz koji se dodao i dodatan kontrolni broj specifičan samo za složena vremena i koji tada nije imao posebnu svrhu osim osiguravanja jedinstvenost podatka i lakšu uočljivost pri pretraživanju. Osim samih riječi, funkcija je svojoj nadređenici također proslijedila i listu pozicija odakle su bile prikupljene komponente. To je bilo ključno jer se standardna provjera vršila obavezno bez obzira na ishod provjere na složena vremena pa se na taj način izbjeglo ponavljanje riječi jer su se riječi s tim indeksima preskakale. Rezultati su pokazali da je na ovaj način bilo analizirano 635 rečenica. Podaci su se revidirali i ručno korigirali zbog faktora ljudske pogreške pri označavanju velike količine teksta i zbog korištenja tada još uvijek eksperimentalnog i novog sustava.

Kada bi se primarna obrada završila, na raspolaganju su bile 3 funkcije za prebrojavanje rezultata. Prva funkcija je računala ukupan broj pojava zadanih pojmova. To je prvenstveno bila testna funkcija i nije bila od koristi pri izradi statistike. Druga funkcija je prebrojavala rezultate po govorniku i ukupno na temelju ključnih pojmova i zadanog rječnika za pretraživanje (posebno su se mogle pretraživati rečenice, vrste riječi i svojstva). Posljednja

funkcija vršila je brojanje po govorniku, ali samih riječi (za razliku od prethodne koja je pretraživala ključeve rječnika).

3.2. Pripremne radnje

Kako bi alat mogao samostalno određivati vrste riječi bilo je potrebno sastaviti rječnik sa što više riječi i podacima o njihovoj vrsti i drugim svojstvima. U tu svrhu je izrađena aplikacija koja na temelju zadane liste riječi prikuplja podatke o njima. Lista ulaznih riječi se sastojala od skupljenih različnica iz raznih dječjih priča, bajki te titlova dječjih animiranih filmova. Time se dobila lista od 81 330 riječi s kojima se djeca različitih uzrasta najčešće susreću. Aplikacija je svaku riječ pretražila na Hrvatskom jezičnom portalu[2] te prikupila sve relevantne podatke o njoj te ih pohranila u bazu. Zbog neujednačenosti i nedosljednosti prikaza podataka na Portalu bilo je potrebno izraditi dodatne algoritme za korekciju glagolskih vremena i sl. Svi prikupljeni podaci su pohranjeni u bazu koja služi kao rječnik glavnom alatu. Nakon testova određen je dizajn baze u kojem je svaki oblik riječi novi unos u odgovarajućoj tablici. Također je svakoj riječi dodijeljen i identifikacijski broj kako bi se omogućio unos homonima, a izbjeglo multipliciranje drugih riječi. Preliminarni rječnik se tako sastoji od preko 600 000 riječi. Svaka riječ uz sebe ima i niz oznaka koje predstavljaju njena svojstva, a oznake su lako čitljive ljudima jer su jednoznačne za svaku vrstu riječi.

Treba napomenuti da su brojevi i zamjenice ručno uneseni u bazu zbog prevelike neujednačenosti i kompleksnosti njihovih prikaza na Portalu. Također, glagolska vremena, načini i sl. su svrstani pod isto svojstvo u bazi zbog praktičnosti.

3.3. LARSPTool

3.3.1. Priprema teksta za analizu

Kako bi LARSPTool mogao pravilno analizirati tekst, potrebno ga je transkribirati na točno određen način, ovisno o formatu datoteke.

[2] Dostupno na: http://hjp.znanje.hr/index.php?show=search (24.09.2019.).

U Microsoft Word dokumentima, blok teksta predstavlja jedan odlomak. Svaki odlomak predstavlja jedan niz rečenica koje je izgovorio jedan govornik koji je označen na početku odlomka. Blokovi teksta koji nemaju označenog govornika se preskaču.

Ovaj blok teksta će biti preskočen. Nema označenog govornika pa se smatra opisnim tekstom.

A: A ovaj blok?

Be: Taj i ovaj blok će biti obrađeni jer imaju označene govornike. Sve rečenice unutar ovih dvaju blokova će biti obrađene!

Međutim, ovaj blok će biti preskočen.

Baš kao i ovaj.

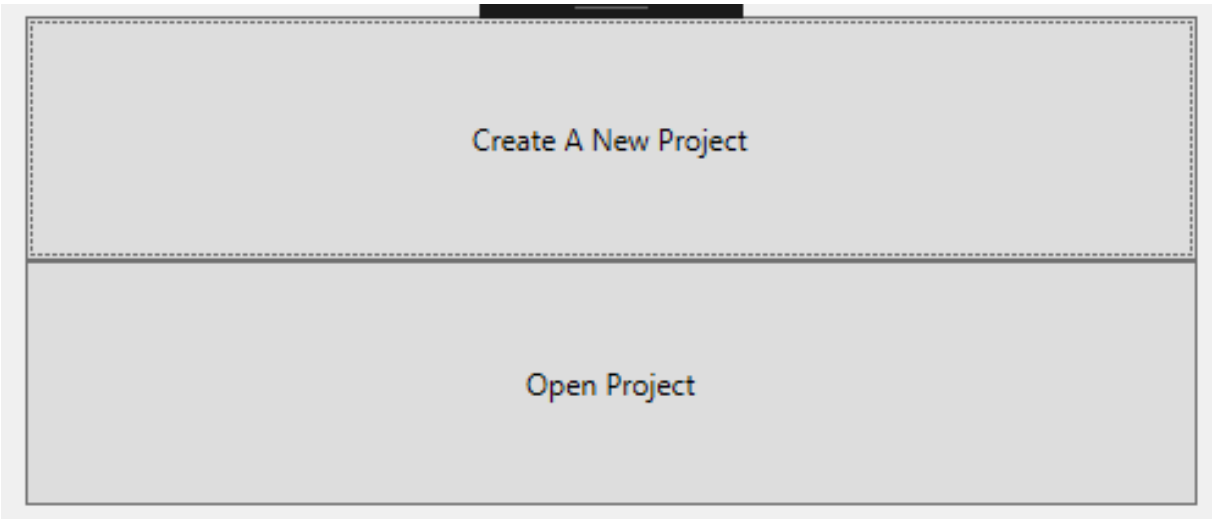
Slika 2. Primjer Transkribiranog Microsoft Word dokumenta

Iz danog primjera se može primijetiti da je drugi govornik označen dvoslovnom oznakom i to velikim i malim slovima. U odnosu na prethodno opisani eksperimentalni program, LARSPTool podržava velika i mala slova te brojeve, a oznake mogu biti bilo koje duljine. Važno je samo da se neposredno nakon oznake nalazi dvotočka nakon koje slijedi razmak.

Za obične tekstualne datoteke (.txt) vrijede slična pravila. Jedina razlika je što svaki blok teksta mora biti u svom redu. Treba primijetiti da dani primjer prikazuje neoznačeni dokument. Naime, alat u vrijeme pisanja ovog rada još uvijek ne podržava prethodno označene dokumente iako će ta značajka biti uskoro dodana sa znatnim poboljšanjima u odnosu na eksperimentalni program.

3.3.2. Rad u alatu

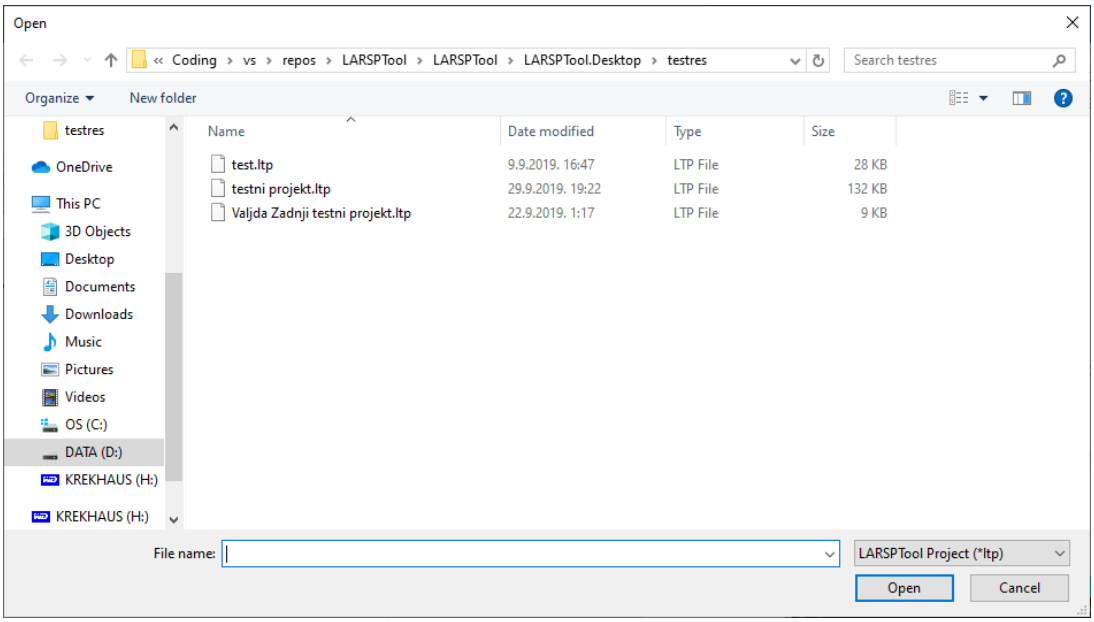
Kada korisnik pokrene aplikaciju, pojavit će se prozor s dvjema mogućnostima.



Slika 3. Prvi prozor LARSPTool-a

Kao što je vidljivo na slici 3., korisnik može stvoriti novi projekt ili otvoriti već postojeći projekt.

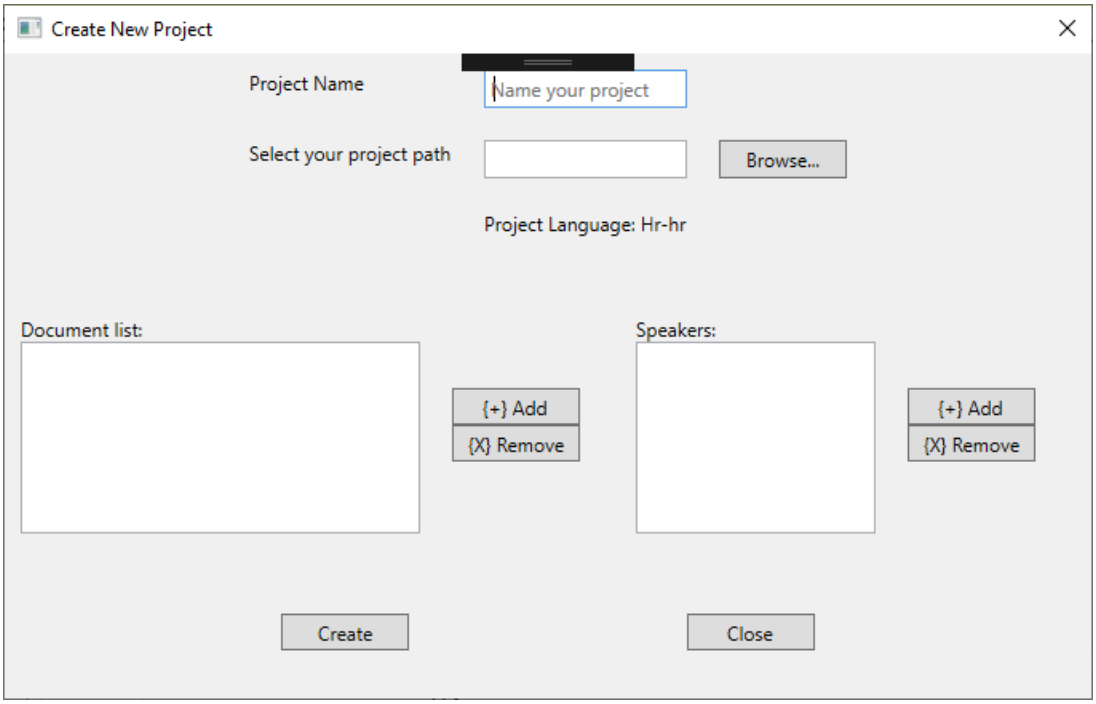
Odabirom opcije „Open Project“ otvara se novi prozor u kojem korisnik može potražiti .ltp datoteku koju želi otvoriti. Odabirom datoteke otvara se glavni prozor o kojem će biti riječi kasnije u tekstu. Ako korisnik ne odabere datoteku nego odustane, aplikacija se zatvara.



Slika 4. Prozor za otvaranje postojećeg projekta

3.3.3.1. Stvaranje novog projekta

Odabirom opcije „Create A New Project“ otvara se novi prozor u kojem korisnik postavlja potrebne elemente projekta.

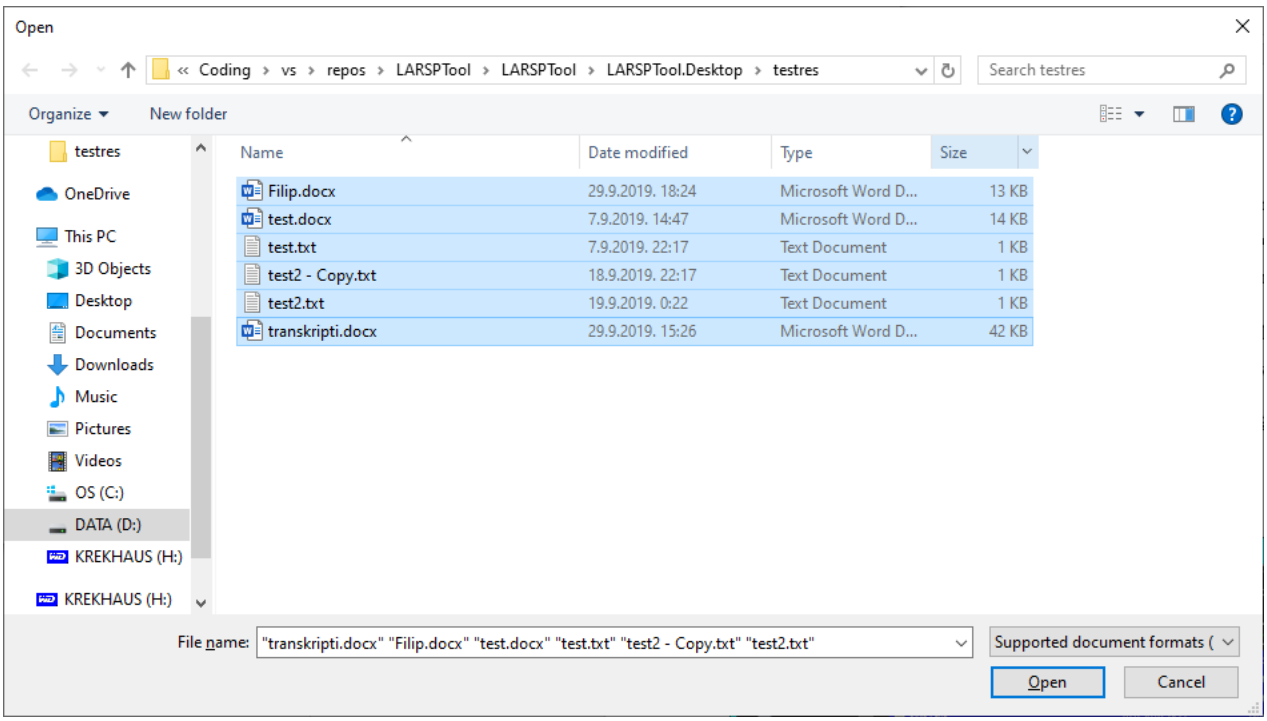


Slika 5. Prozor za stvaranje novog projekta

Sva polja i liste moraju biti popunjeni. U prva dva polja korisnik upisuje naziv projekta te putanju gdje će se datoteka pohranjivati. Kako bi lakše pronašao putanju, korisnik može kliknuti gumb „Browse…“ kako bi otvorio dijalog za odabir mape.

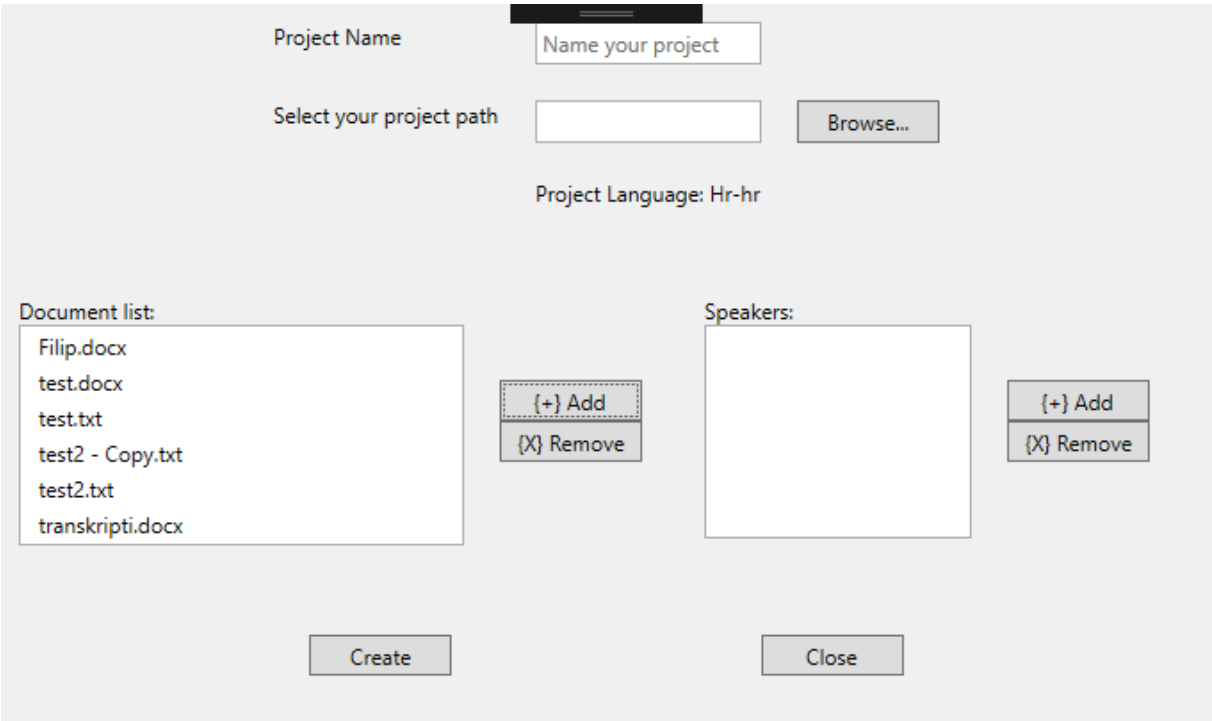
Također, u budućim verzijama korisnik će moći i odabrati jezik projekta. Trenutna verzija ima integriranu podršku za hrvatski i nije modularna pa je jezik naveden samo kao tekst.

Prva lista sadrži transkribirane dokumente koji će biti korišteni u projektu. Dokumenti se mogu dodavati klikom na gumb „Add“ koji otvara prozor za odabir podržanih datoteka. Moguće je dodati više datoteka odjednom.



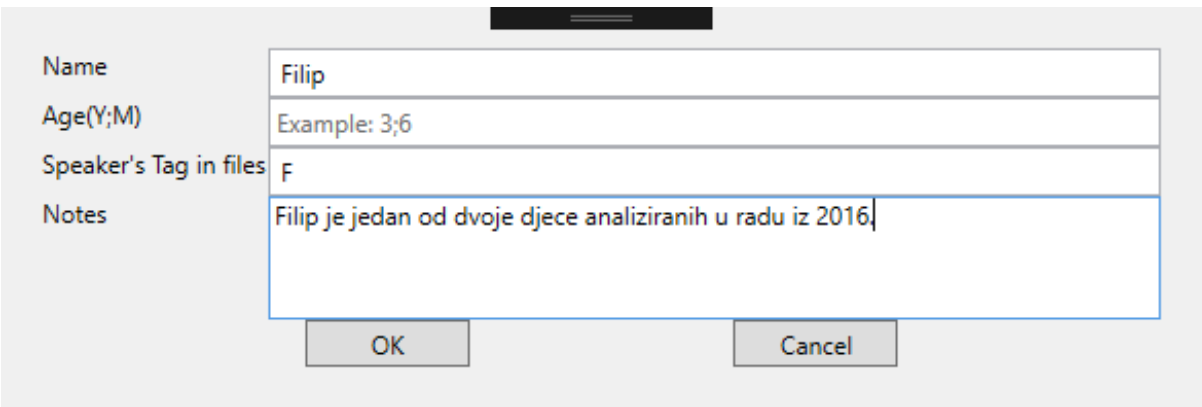
Slika 6. Dodavanje više datoteka na popis dokumenata

Po odabiru datoteka, u listi dokumenata pojavljuju se nazivi dodanih datoteka (slika 6). Dokumente se s liste briše klikom na gumb „Remove“. Dokumenti se mogu brisati samo jedan po jedan.



Slika 7. Prikaz liste s dodanim dokumentima

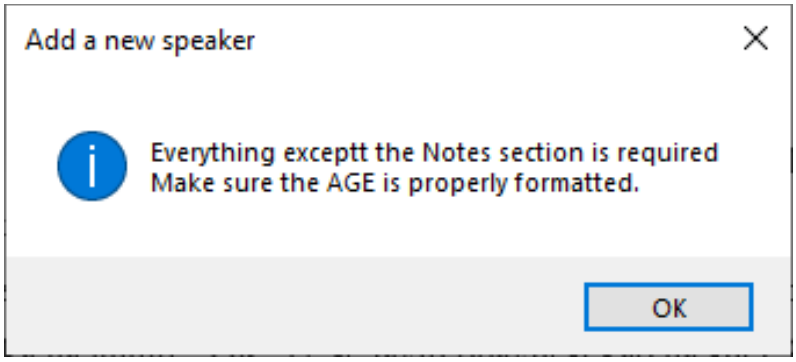
Druga lista sadrži popis svih govornika za koje korisnik želi vršiti obradu. Klikom na gumb „Add“ otvara se novi prozor za dodavanje novog korisnika.



Slika 8. Dijalog za dodavanje govornika

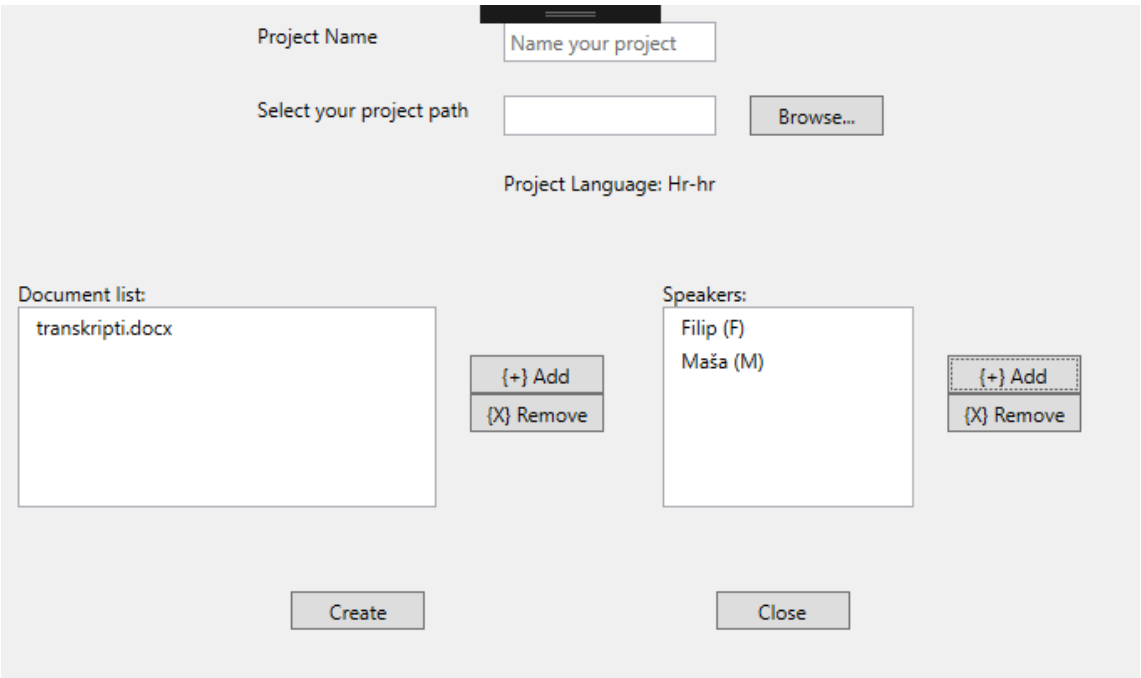
Za dodavanje novog govornika potrebno je unijeti njegovo ime, dob (onako kako se označava u LARSP-u) i oznaku u tekstu. Također je moguće dodati i bilješke ako ima potrebe za tim.

Bilješke su jedino neobavezno polje. Ako je neko od ostalih polja prazno ili je dob pogrešno upisana, prilikom klika na gumb „OK“ će se javiti obavijest kao na slici 9.



Slika 9. Greška prilikom dodavanja novog govornika

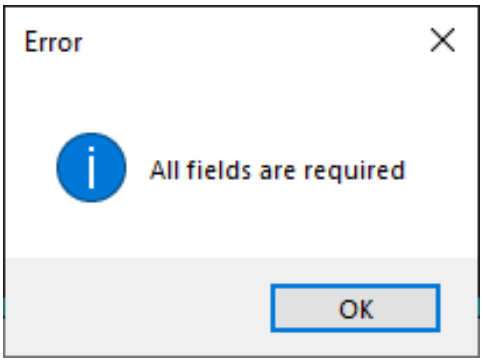
Dodani govornici će biti prikazani na listi u obliku Ime (oznaka) kao na slici 10.



Slika 10. Popis govornika u projektu

Govornici se, kao i dokumenti, s liste brišu jedan po jedan.

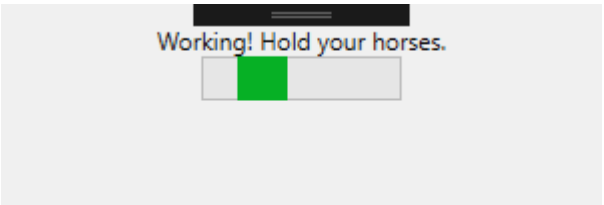
Kad su svi podaci uneseni klikom na gumb „Create“ provjeravaju se polja i ako su neka prazna pojavljuje se obavijest kao na slici 11.



Slika 11. Greška prilikom izrade novog projekta

Ako su sva polja pravilno ispunjena, klikom na gumb „Create“ pokreće se automatska predobrada. Klikom na gumb „Close“ ili zatvaranjem prozora u bilo kojem trenutku vraća korisnika na prvi dijalog.

Budući da predobrada može dugo trajati, tijekom predobrade pojavljuje se prozor koji upućuje na rad programa (Slika 12). Treba napomenuti da metoda predobrade ovisi o autoru algoritama za pojedini jezik.



Slika 12. Prozor koji upućuje na obradu u tijeku

Tijekom predobrade program gradi podatkovno stablo, određuje granice rečenica, njihove vrste i vrste riječi u rečenicama. Govornici koji su označeni u dokumentima, a nisu na popisu će biti preskočeni.

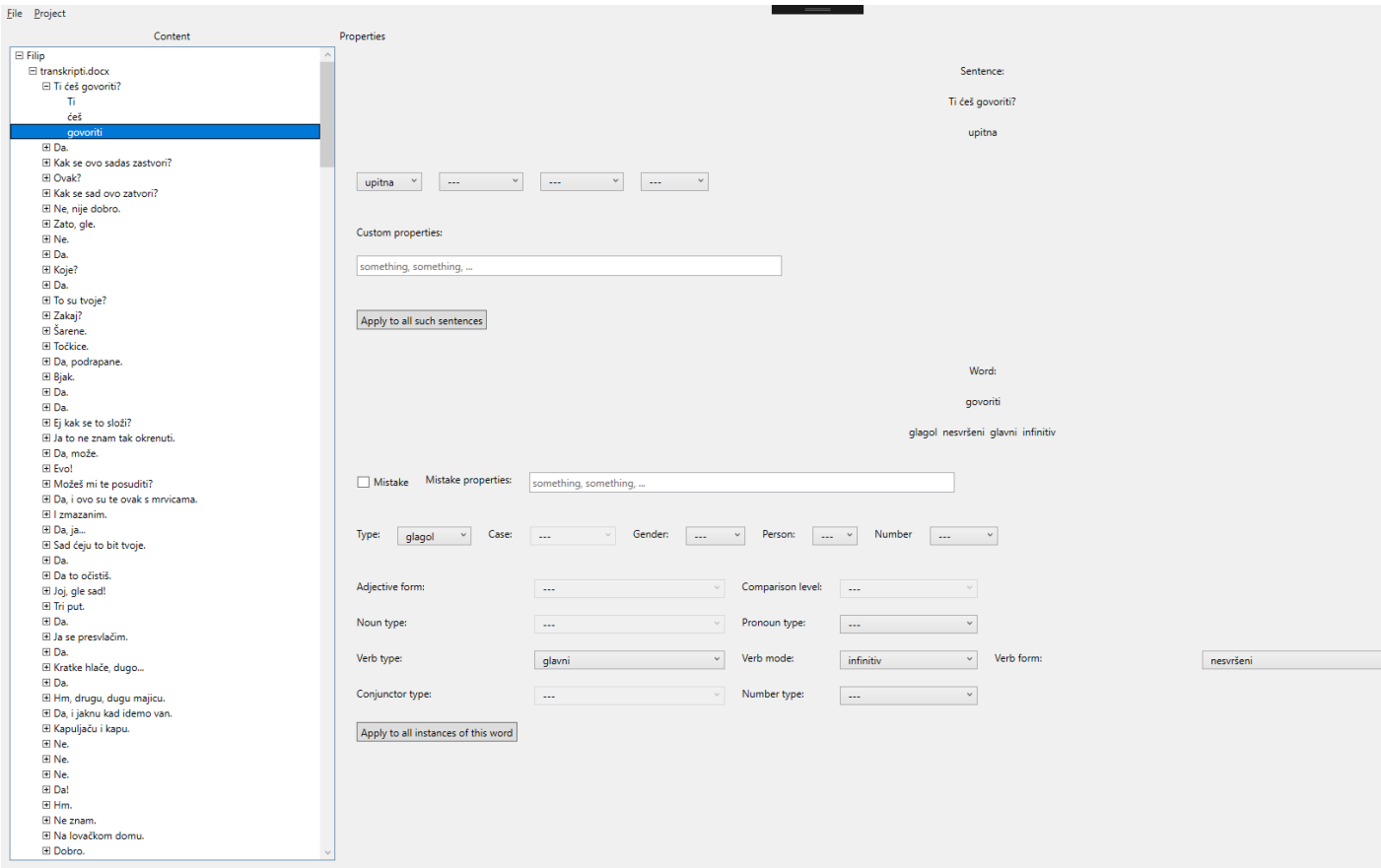
U trenutnoj verziji modula za hrvatski, program određuje samo osnovnu vrstu rečenice na temelju interpunkcije na kraju, a svojstva riječi se detektiraju rangiranjem rezultata iz baze

po sličnosti korištenjem Levenshtein Distance[3] metode. Rezultati se također rangiraju prema broju rezultata za svaku vrstu prema načelu „manje je bolje“ jer se pretpostavlja da manji broj rezultata predstavlja točniji pogodak. U sljedećoj verziji, algoritam će implementirati Croapi[4] i UDPipe[5] kako bi se povećala brzina i točnost.

[3] Levenshtein (1966).
[4] Dostupno na: https://croapi.github.io/vrste-rijeci/index.html#about (24.09.2019.)
[5] Straka et al. (2017).

Kad se predobrada završi otvara se glavni prozor (Slika 13).

3.3.3.2. Glavni prozor



Slika 13. Glavni prozor LARSPTool-a

Glavni prozor je podijeljen na dva glavna dijela:

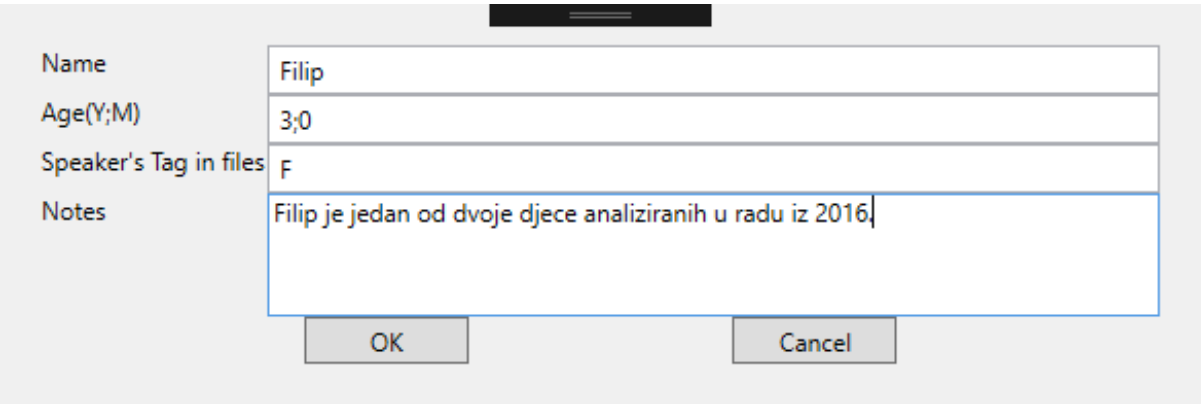
1. Podatkovno stablo
2. Dio sa svojstvima

Podatkovno stablo ima 4 razine:

- Govornik
    - Dokument u kojem je govornik pronađen
        - Rečenica iz tog dokumenta koju je govornik izgovorio
            - Riječi te rečenice

Svaki element koji ima oznaku s lijeve strane može se proširiti čime se prikazuju njegovi podelementi. Po stablu se može kretati i tipkovnicom; strelicama gore i dolje kreće se po prikazanim elementima, strelicom desno odabrani element se proširuje, a strelicom lijevo se skuplja.

Dvoklikom na govornika se otvara prozor za uređivanje podataka o tom govorniku..



Slika 14. Prozor za uređivanje podataka o govorniku

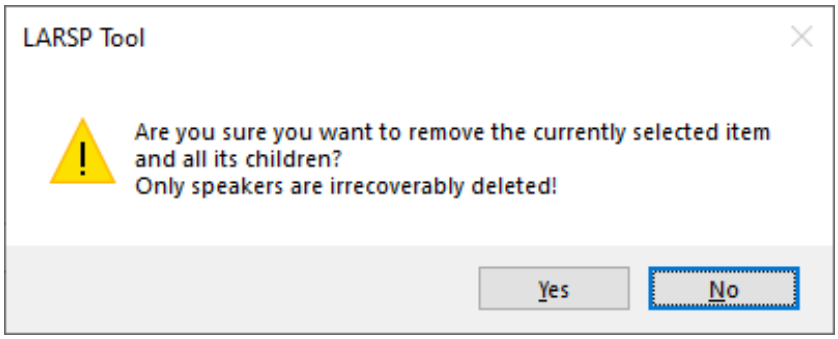
Prozor (slika 14.) je identičan onome za dodavanje novog govornika te kao i kod dodavanja ovdje vrijede ista pravila. Prozor za uređivanje se može otvoriti i odabirom opcije „Edit Speaker“ u izborniku „Project“.

U izborniku „Project“ nalaze se sve mogućnosti vezane za dodavanje, uređivanje ili brisanje elemenata stabla.

„Add Speaker“ otvara prozor za dodavanje novog govornika kao i kod stvaranja projekta. Dodavanjem govornika u projekt se ne vrši ponovna obrada teksta, već se svi dokumenti ponovo analiziraju te se govorniku dodjeljuju rečenice iz blokova u kojima je on pronađen. Sve rečenice i riječi već su obrađene kod stvaranja projekta zbog čega je postupak dodavanja govornika znatno brži.

„Add Documents“ otvara prozor za odabir datoteka. Odabrane datoteke se, kao i kod stvaranja projekta, dodaju na popis i obrađuju. Ovdje se vrši predobrada dokumenata isto kao i prije te se svakom govorniku dodaje dokument u kojem je taj govornik pronađen, zajedno s rečenicama i riječima. Ova radnja može dugo potrajati pa se opet pojavljuje indikator rada.

„Delete Current Item“ briše trenutno odabrani element stabla. Tipkovnička kratica je tipka „Delete“. Prije brisanja traži se potvrda brisanja (Slika 15).



Slika 15. Potvrda brisanja elementa

Brisati se mogu svi elementi osim riječi. Kao što je navedeno u dijaloškom okviru, svi elementi podređeni odabranom elementu također će biti izbrisani. Može se primijetiti i da je navedeno da su samo govornici nepovratno izbrisani. To je zato što se dokumenti, rečenice i riječi zapravo ne brišu iz projekta već su označeni kao izbrisani. Ovo je temelj planirane značajke restauracije izbrisanih elemenata. Govornici se stvarno brišu iz projekta te se dodavanjem govornika s istim podacima kao izbrisani zapravo stvara novi govornik i dodaje se na dno stabla.

Desni dio glavnog prozora je, kako je već spomenuto, dio sa svojstvima. On je podijeljen vodoravno na dva dijela:

1. Svojstva rečenice
2. Svojstva riječi

Svojstva rečenice se ažuriraju kada je izabrana nova rečenica ili riječ. To omogućava uređivanje trenutne rečenice bez potrebe za dodatnim kretanjem po stablu. Za lakše snalaženje, iznad padajućih izbornika prikazana je trenutna rečenica i sva određena joj svojstva.

Svojstva rečenice određuju se pomoću padajućih izbornika. Svaki padajući izbornik predstavlja niz svojstava koja se međusobno isključuju. U ovoj verziji:

1. izjavna/upitna/usklična rečenica jer rečenica ne može biti sve troje odjednom
2. jednostavna/složena
3. neproširena/proširena/zavisna/nezavisna
4. eliptična

Svaki od izbornika može i ne mora imati izabrano jedno od svojstava što je prikazano trostrukom povlakom. Osim predodređenih svojstava, polje za pisanje naslovljeno „Custom properties“ omogućuje korisniku ručno dodavanje svojstava rečenici. Svojstva su odvojena zarezom tako da bi niz „prvo svojstvo, drugo svojstvo, treće svojstvo“ dodijelio rečenici tri nova svojstva.

Na posljetku, kako bi se olakšala obrada velike količine teksta, gumb „Apply to all such sentences“ primjenuje određena svojstva na sve rečenice jednake odabranoj.

Svojstva riječi prikazana su samo kada je odabrana riječ. Kao i kod rečenica, prikazana je trenutna riječ i njena svojstva. Prvi red sastoji se od okvira za označavanje radi li se o grešci govornika koji, ako je označen, otključava polje za upisivanje u kojem korisnik može upisati sve karakteristike greške, ali nije obavezno. Svojstva su odvojena zarezom kao i kod rečenica.

Drugi red sastoji se od niza padajućih izbornika koji sadržavaju svojstva zajednička većem broju vrsta riječi (vrsta riječi, padež, rod, lice, broj). Potom slijede redovi padajućih izbornika sa svojstvima za određenu vrstu riječi:

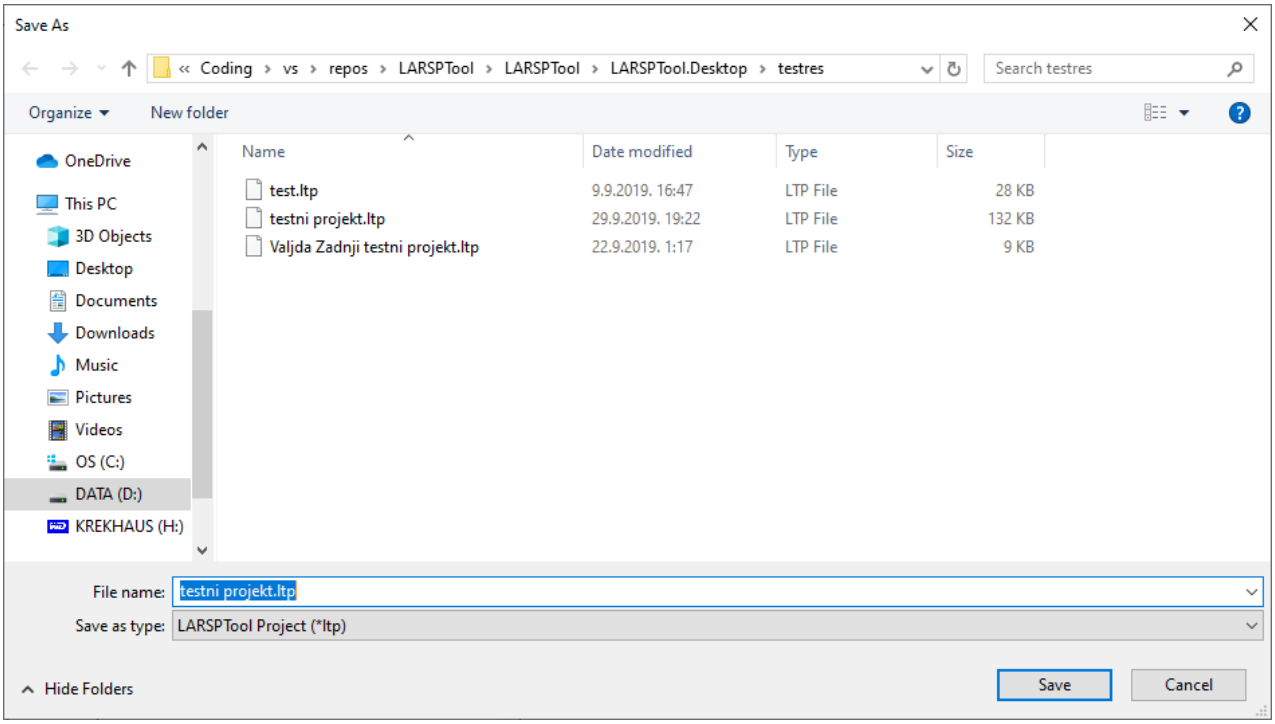
1. pridjevi (oblik i stupanj komparacije)
2. imenice i zamjenice (vrsta)
3. glagoli (vrsta, način/oblik/vrijeme, vid)
4. veznici i brojevi (vrsta).

Na kraju gumb „Apply to all instances of this word“ kao i kod rečenica primjenjuje sva određena svojstva na sve identične riječi. Vrsta riječi određuje koja svojstva se mogu odrediti. Ovisno o tome koja je vrsta odabrana, izbornici se zaključavaju i otključavaju tako da se kod priloga kao na slici 13. ne može odrediti nijedno dodatno svojstvo.

Svojstva iz izbornika dodjeljuju se čim se odabere vrijednost, a svojstva grešaka i prilagođena svojstva dodjeljuju se kada korisnik napusti polje za pisanje (kad ono izgubi fokus). Prikazi svojstava će se pravilno ažurirati kada korisnik odabere neki drugi element.

Izbornik „File“ sadrži radnje vezane za cjelokupni projekt i aplikaciju.

„Save Project“ otvara dijalog u kojem korisnik može odabrati gdje želi pohraniti i kako želi nazvati datoteku. Unaprijed će biti predloženi naziv i putanja s obzirom na parametre zadane pri stvaranju projekta. Tipkovnička kratica je „Control/Command+S“.

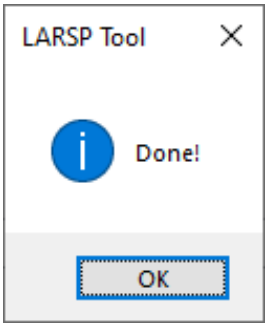


Slika 16. Dijalog za spremanje projekta

Projekt se pohranjuje u .ltp datoteku koja nije namijenjena za čitanje izvan alata. Zbog metode pohrane i korištenja brzog algoritma za kompresiju, datoteka je relativno male veličine i brzo se sprema i ponovo učitava.

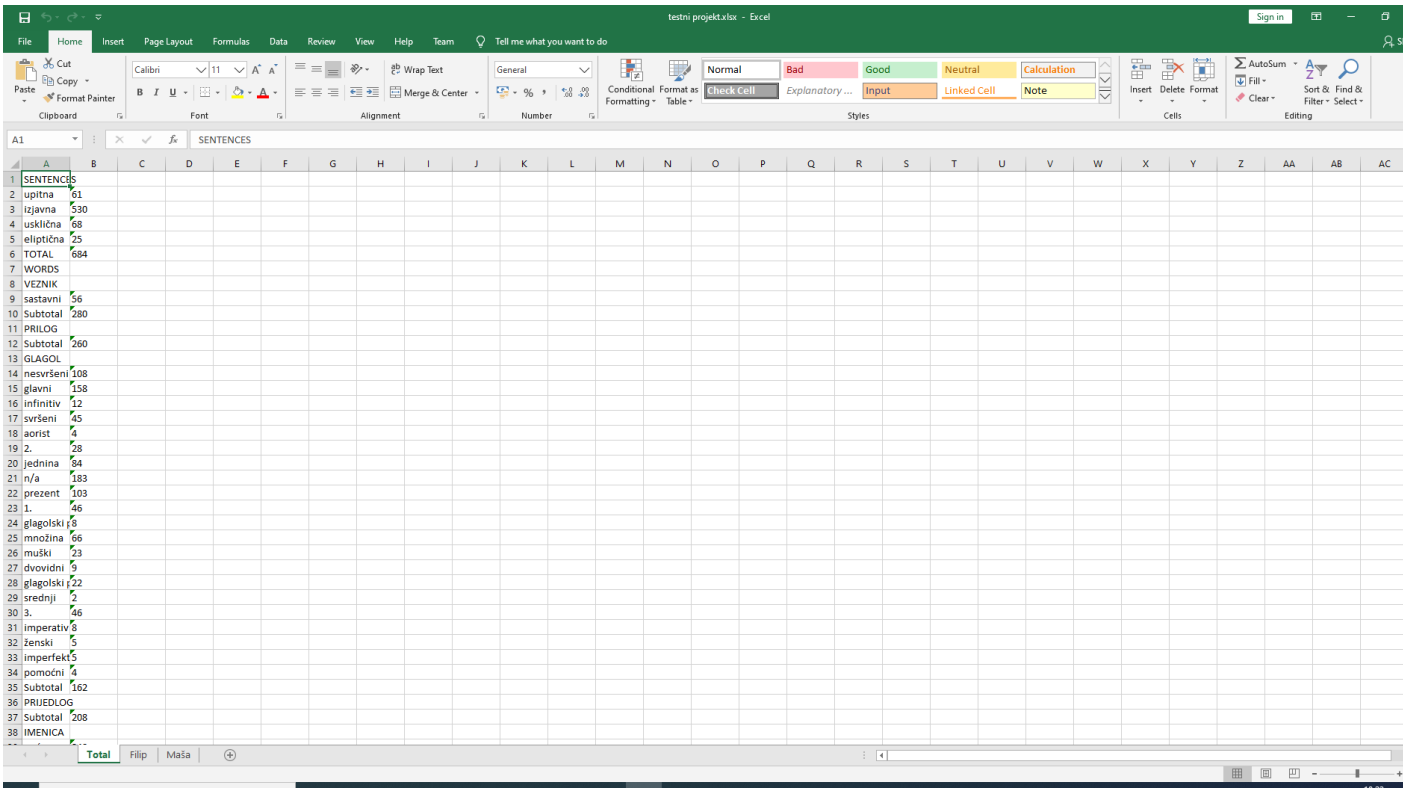
„Export Project Data“ omogućuje izvoz podataka projekta u lakše čitljiv oblik. Ova funkcija namijenjena je za završetak obrade projekta. U trenutnoj verziji podržan je samo .xlsx format. Odabirom ove funkcije otvara se dijalog za spremanje datoteke te se kao i kod spremanja projekta predlaže naziv datoteke prema nazivu projekta.

Po završetku izvoza pojavljuje se obavijest o završetku (slika 17.) nakon čega se program može sigurno zatvoriti.



Slika 17. Potvrda o završetku izvoza podataka

Izvezeni dokument izgleda kao na slici 18.

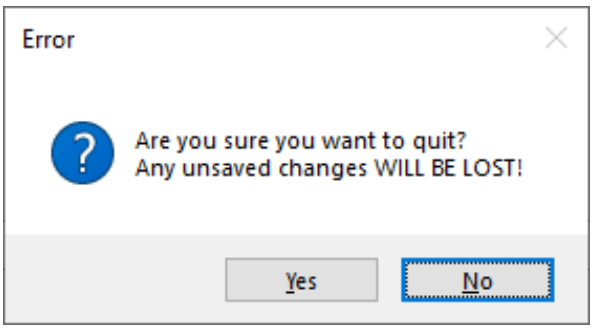


1	SENTENCES	
2	upitna	61
3	izjavna	530
4	usklična	68
5	eliptična	25
6	TOTAL	684
7	WORDS	
8	VEZNIK	
9	sastavni	56
10	Subtotal	280
11	PRILOG	
12	Subtotal	260
13	GLAGOL	
14	nesvršeni	108
15	glavni	158
16	infinitiv	12
17	svršeni	45
18	aorist	4
19	2.	28
20	jednina	84
21	n/a	183
22	prezent	103
23	1.	46
24	glagolski p	8
25	množina	66
26	muški	23
27	dvovidni	9
28	glagolski p	22
29	srednji	2
30	3.	46
31	imperativ	8
32	ženski	5
33	imperfekt	5
34	pomoćni	4
35	Subtotal	162
36	PRIJEDLOG	
37	Subtotal	208
38	IMENICA	

Uz list s ukupnim podacima, svaki govornik je na zasebnom listu koji je nazvan po njemu. Svaki odjeljak podataka označen je nazivom napisanim velikim slovima. Isto vrijedi i za zbroj svakog većeg odjeljka (rečenice, riječi, pogreške). Riječi koje su u projektu označene kao pogreške ne ubrajaju se u ostale riječi. Sve pogreške, bez obzira imaju li dodana svojstva ili ne, ubrajaju se u pogreške zbog čega se može činiti da zbroj pogrešaka ne odgovara zbroju

svih potkategorija. Svako svojstvo za rečenice, riječi i pogreške čini svoju potkategoriju tako da se riječi sa zajedničkim svojstvima ubrajaju pod sve kategorije čija svojstva sadržavaju (sve opće imenice u genitivu bit će ubrojene pod imenice, opće imenice i imenice u genitivu).

Posljednja funkcija je funkcija „Quit“ koja zatvara program. Tipkovnička kratica je „Control/Command+Q“, ali se također izvodi bilo kakvim zatvaranjem glavnog prozora. Prije zatvaranja korisnika se pita je li siguran da želi zatvoriti aplikaciju te ga se upozorava da će sve nespremljene promjene biti izgubljene (Slika 19).



Slika 19. Potvrda o zatvaranju aplikacije

Napomena: izgled izvezenog dokumenta i svih elemenata aplikacije nije nužno konačan i odražava dizajn posljednje verzije aplikacije za vrijeme pisanja ovog rada.

4. ZAKLJUČAK

Ova aplikacija će, u budućnosti, imati podršku za dodavanje bilo kojeg jezika. Može se primijeniti na bilo koji pravilno transkribirani korpus teksta. Iako je napravljena s analizom dječjeg govora na umu, zapravo se može koristiti za analizu bilo kakvog teksta. Važno je zapamtiti da se radi o alatu tj. o softveru koji je namijenjen pomaganju ljudima u obradi podataka, a ne njihovoj zamjeni. Svrha sve automatizacije jest olakšavanje i ubrzavanje obrade velike količine podataka. Do sada je, barem za hrvatski jezik, analiza bila ručna i samim time dugotrajna i podložnija ljudskoj pogrešci s povećanjem količine materijala za obradu čemu može posvjedočiti i autor ovog rada. LARSPTool znatno skraćuje vrijeme obrade povećavajući istovremeno korpus koji se može analizirati. Također, ovim alatom se pokušava uvesti standard i jednoličnost obrade teksta što bi omogućilo lakše i preciznije profiliranje i praćenje razvoja govora djeteta kroz vrijeme. Također, sabiranjem podataka prikupljenih korištenjem ovog alata mogu se pratiti trendovi dječjih govornih poremećaja i općenito razvoja govora kroz generacije. Brza i točna dijagnostika je ključ za kvalitetnu i učinkovitu rehabilitaciju.

5. REFERENCIJE

1. Apel, K., Masterson, J.J. (2004). *Jezik i govor od rođenja do 6.godine*. Lekenik: Ostvarenje d.o.o.
2. Ball, Martin J., Crystal, D., Fletcher, P. (2011). *Assessing Gramar, The Languages of LARSP*. Bristol: Multilingual Matters. Dostupno i na: http://www.davidcrystal.community.librios.com/books-and-articles (24.09.2019.)
3. Croapi. https://croapi.github.io/vrste-rijeci/index.html#about (24.09.2019.)
4. Crystal, D. (1976). *Child language, learning and linguistics*. London: Edward Arnold, 106 pp. 0 7131 5890 5; 0 7131 5891 3.
5. Crystal, D. (1979). *Working with LARSP*. London: Edward Arnold. Dostupno i na: http://www.davidcrystal.community.librios.com/books-and-articles (24.09.2019.)
6. Crystal, D. i sur. (1989). *Grammatical Analysis of Language Disability*. Drugo izdanje. London: Cole and Whurr. Dostupno i na: http://www.davidcrystal.community.librios.com/books-and-articles (12.10.2019.)
7. Hrvatski jezični portal. http://hjp.znanje.hr/index.php?show=search (24.09.2019.)
8. Kantoci, E.; Tot, B. (2016). *Primjena LARSP-a na hrvatski jezik u dobi od tri godine*. Seminarski rad, neobjavljen. Filozofski fakultet Sveučilišta u Zagrebu.
9. Levenshtein, V.I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady. 10 (8): 707–710.
10. Long, S. (2012). '*Computerized Profiling' of Clinical Language Samples and the Issue of Time*. Assessing Grammar: The Languages of LARSP. 29-42. Dostupno i na: https://pdfs.semanticscholar.org/5aed/7d789b87e63489bad807ecc243c8c02c8694.pdf (13.10.2019.)
11. Mildner, V.; Stojanovik, V., and Tomić, D. (2019). Croatian LARSP. In *Grammatical Profiles: Further Languages of LARSP* (Eds. M. J. Ball, P. Fletcher & D. Crystal), 82–119. Bristol: Multilingual Matters.

12. Straka, M.; Straková, J. (2017). *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada.
13. Žganec,T. (2015). *Primjena LARSP-a na hrvatski jezik u dobi od pet godina*. Seminarski rad, neobjavljen. Filozofski fakultet sveučilišta u Zagrebu.

SAŽETAK

Automatsko označavanje i analiza teksta za LARSP hrvatskog jezika

Govor je jedan od najznačajnijih načina ljudske komunikacije. Govor i jezik se zajedno razvijaju od rođenja – prvo slušanjem, a zatim imitacijom. Kao i sve ljudske osobine, podložni su poremećajima i oštećenjima. Neka od njih se mogu rehabilitirati zbog čega je važno što prije ih detektirati. U tu svrhu je razvijen LARSP – lingvistički protokol koji objedinjuje procjenu, rehabilitaciju i probir tj. praćenje kako bi omogućio točnu dijagnozu i uspješnu rehabilitaciju. Temelji se na gramatičkoj analizi transkribiranog dječjeg govora. Ručna analiza većeg korpusa teksta dugotrajna je i podložna ljudskim pogreškama. LARSPTool je alat razvijen kako bi olakšao taj postupak. Automatska predobrada i prikupljanje podataka znatno ubrzavaju analizu i omogućuju analizu mnogo većeg korpusa. LARSPTool također uvodi standard i jednoličnost obrade i prikaza podataka što otvara daljnje mogućnosti za praćenje trendova razvoja govora i jezika kod djece kroz generacije. Trenutna verzija ima integriranu podršku za hrvatski, a kasnije verzije bit će modularne. Također, dostupna je samo za Windows operativne sustave, ali postoji mogućnost razvoja verzija i za druge platforme.

Ključne riječi: LARSP, djeca, govor, automatska analiza, jezik

ABSTRACT

Automatic text markup and analysis for LARSP of Croatian language

Speech is one of the most significant methods of human communication. Speech and language start developing together from birth – first by listening then by imitation. Like all human characteristics it is subject to disorders and defects. Some of them can be remedied which is why it is important to detect them as soon as possible. To that end LARSP was developed – a linguistic protocol which unifies assessment, remediation and screening in order to enable correct diagnosis and successful remediation. It is based on grammatical analysis of transcribed children's speech. Manual analysis of a larger corpus of text is time-consuming and susceptible

to human error. LARSPTool is a tool developed to facilitate that process. Automatic preprocessing and data gathering significantly speed up analysis and enable analysis of a much larger corpus. LARSPTool also introduces a standard and uniformity to both data processing and presentation which opens up further possibilities of tracking children's speech and language development trends over generations. The current version has integrated support for Croatian but later versions will be modular. Also, it is only available for Windows operating systems but there is a possibility of developing versions for other platforms as well.

Keywords: LARSP, children, speech, automatic analysis, language