UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA

IZDELAVA SISTEMA ZA STROJNO PREVAJANJE NA OSNOVI PRAVIL PLITKEGA PRENOSA ZA JEZIKOVNI PAR SLOVENŠČINA-MAKEDONŠČINA

MARKO TRAJKOSKI

2014
ZAKLJUČNA NALOGA
TRAJKOSKI

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna projektna naloga

Izdelava sistema za strojno prevajanje na osnovi pravil plitkega prenosa za jezikovni par slovenščina - makedonščina

(Creating a Rule-based shallow transfer machine translation system for the Slovenian-Macedonian language pair)

Ime in priimek: Marko Trajkoski
Študijski program: Računalništvo in informatika - 1. stopnja
Mentor: doc. dr. Branko Kavšek
Somentor: doc. dr. Jernej Vičič

Koper, maj 2014

Ključna dokumentacijska informacija

Ime in PRIIMEK: Marko TRAJKOSKI

Naslov zaključne naloge: Izdelava sistema za strojno prevajanje na osnovi pravil plitkega prenosa za jezikovni par slovenščina - makedonščina

Kraj: Koper

Leto: 2014

Število listov: 47 Število slik: 19 Število tabel: 3

Število prilog: 2 Število strani prilog: 7 Število referenc: 13

Mentor: doc. dr. Branko Kavšek

Somentor: doc. dr. Jernej Vičič

UDK:

Ključne besede: strojno prevajanje, strojno prevajanje sorodnih naravnih jezikov, jezikovni par slovenščina - makedonščina

Izvleček:
Zaključna projektna naloga predstavlja postavitev sistema za strojno prevajanje za jezikovni par slovenščina – makedonščina. Sistem temelji na ogrodju Apertium, katero sodi v paradigmo sistemov za strojno prevajanje na osnovi pravil plitkega prenosa, (shallow transfer RBMT). Predstavljene so metode za hitro postavitev sistema kakor tudi vsa naknadna dela in postopki, ki so bili potrebni za odpravo napak. Predstavljene so osnovne statistike izdelanih jezikovnih gradiv in rezultati vrednotenja.

Key words documentation

Name and SURNAME: Marko TRAJKOSKI

Title of final project paper: Creating a Rule-based shallow transfer machine translation system for the Slovenian-Macedonian language pair

Place: Koper

Year: 2014

Number of pages: 47 Number of figures: 19 Number of tables: 3

Number of appendices: 2 Number of appendix pages: 7 Number of references: 13

Mentor: Assist. Prof. Branko Kavšek

Co-mentor: Assist. Prof. Jernej Vičič

UDK:

Keywords: rbmt, machine translation, machine translation of related languages, language pair slovene - macedonian

Abstract:
The final project paper presents an overview of a machine translation system for the slovene and macedonian language pair. The translation system is based on Apertium's architecture, which belongs to the Shallow Parse and Transfer Rule-Based Machine Translation paradigm. Methods for speeding up the implementation of machine translation systems and an overview of all subsequent work that has been done to correct existing errors are presented. The paper also presents the basic statistics of the produced language resources and the final evaluation results.

Zahvala

Zahvaljujem se mentorju doc. dr. Branku Kavšku in somentorju doc. dr. Jerneju Vičiču za strokovno pomoč in usmeritve pri izdelavi zaključne projektne naloge.

Zahvaljujem se tudi vsem zaposlenim na Fakulteti za matematiko, naravoslovje in informacijske tehnologije.

Veliko hvala tudi moji družini za vsestransko pomoč na moji izobraževalni poti.

Hvala!

Kazalo vsebine

Seznam tabel

Seznam slik

Seznam prilog

Seznam kratic

MT - Machine Translation, strojno prevajanje
GNU - GNU's Not Unix!, projekt za izdelavo operacijskega sistema GNU
GPL - GNU General Public License, prosta licenca za programsko opremo in ostalo
XML - Extensible Markup Language, jezik za označevanje
RBMT - Rule Based Machine Translation, strojno prevajanje na osnovi pravil
LDC - Linguistic Data Consortium, konzorcij jezikovnih podatkov
WER - Word Error Rate, stopnja napačnih besed
WRR - Word Recognition Rate, stopnja prepoznavnih besed
LDC - Linguistic Data Consortium, konzorcij jezikovne podatke

1 Uvod

Postopek, pri katerem računalniški program analizira besedilo in brez človeške pomoči proizvede ciljno besedilo, se imenuje strojno prevajanje (angl. machine translation - MT). Sistemi za strojno prevajanje so precej kompleksni, vključujejo eno ali več jezične leksikone, programe za sintaktično analizo in sintezo, programe za morfološko analizo in sintezo kakor tudi druge mehanizme za avtomatizacijo prevajalskega procesa.

Ideja za strojno prevajanje obstaja že od 17. stoletja in sega vse do 50-tih in 60-tih let dvajsetega stoletja, ko je na Univerzi v Georgetownu, leta 1954 prvič javno prikazano strojno prevajanje z IBM-ovim računalnikom. S slednjim je bilo prevedeno več kot šestdeset povedi iz ruskega v angleški jezik (William John Hutchins, 2005). Takrat, ko se je pokazalo, da je problem strojnega prevajanja kompleksnejši kot se je sprva zdelo, je zanimanje za to delo nekaj let popustilo. V 80 – tih letih so se pojavili prvi komercialni prevajalni sistemi. Na prelomu tisočletja so se pojavili prvi nekomercialni oziroma brezplačni spletni strojni prevajalniki. Še vedno ne obstaja sistem, ki bi proizvedel popolno avtomatiziran in visoko kakovosten prevod.

Apertium je odprtokodna platforma za strojno prevajanje, izdana pod licenco GNU GPL. Razvita je s podporo španske vlade in Univerze v Alicanteju. Sprva je bil namenjen prevajanju med sorodnimi jeziki, potem so se njegove zmožnosti razširile tudi med manj sorodne jezikovne pare. Podatki za implementacijo novega jezikovnega para so organizirani v XML datotekah, ki so človeku relativno lahko razumljivi. Strojni prevajalni sistemi, ki so se razvili s pomočjo Apertiuma, temeljijo na pravilih plitkega prenosa.

Moja motivacija da bi začel razvijati sistem za strojno prevajanje za jezikovni par slovenščina-makedonščina sega v čas, ko sem opazil da Apertium še nima sistema za ta jezikovni par. Glede tega, sem kot študent ki prihaja iz Makedonije in študira v Sloveniji, začel razmišljati tako in ideja za razvijanje takšnega sistema mi je postajala vse bolj všeč. Po pogovoru z mentorjem in pregledu gradiv mi je slednji dal potrebne napotke in tako sem začel razvijati omenjeni strojno prevajalni sistem.

1.1 Pregled vsebine

Zaključna projektna naloga predstavlja postavitev sistema za strojno prevajanje za jezikovni par slovenščina - makedonščina. Sistem smo razvijali na ogrodju Apertium, katero sodi v paradigmo sistemov za strojno prevajanje na osnovi pravil plitkega prenosa (ang. shallow transfer RBMT), saj se je le-to izkazalo kot najprimernejšo za postavitev sistema za strojno prevajanje sorodnih jezikov.

V drugem poglavju je predstavljen pregled oziroma opis strojnega prevajanja. Podrobneje so opisani RBMT sistem, sistem RBMT s plitkim prenosom ter arhitektura prevajalne platforme Apertium, na katera smo razvijali prevajalni sistem predstavljen v tem delu. V tem poglavju so razloženi osnovni pojmi znanstvenega področja. Predstavljene so tudi osnovne značilnosti slovenskega in makedonskega jezika.

Tretje poglavje opisuje delo, katero je bilo opravljeno na enojezičnih morfoloških slovarjih izvirnega ter ciljnega jezika. Predstavljeno je tudi delo, katero je bilo opravljeno v dvojezičnemu slovarju ter nekaj osnovnih pravil strukturnega prenosa. Prav tako so tudi v tem poglavju podrobno opisane značilnosti katerim smo se še posebej posvetili.

V četrtem poglavju so opisane metode, katere so bile uporabljene za vrednotenje kakovosti prevodov ter vrednotenje kakovosti jezikovnega gradiva.

V petem poglavju so predstavljeni rezultati vrednotenja, kateri so bili doseženi pri različnih metodah testiranja.

Šesto poglavje zaključuje delo z razpravo in s smernicami za nadaljnje delo.

2 Opis Področja

V tem poglavju bodo na splošno pregledani sistemi za strojno prevajanje, prav tako bodo opisani sistemi, ki delujejo na osnovi pravil strukturnega prenosa. Pregledali bomo tudi arhitekturo platforme Apertiuma, kjer smo razvijali jezikovni par slovenščina - makedonščina. Tukaj bomo pred vsem predstavili osnovne pojme iz področja jezikovnih tehnologij, kateri so potrebni za lažje razumevanje dela.

2.1 Sistemi za strojno prevajanje

Obstaja več različnih metod oziroma načinov izdelave strojno prevajalnih sistemov. Sistemi za strojno prevajanje so razdeljeni v tri skupine in sicer: sistemi za strojno prevajanje kateri temeljijo na osnovi pravil (ang. rule – based machine trenslation RBMT), sistemi strojnega prevajanja kateri temeljijo na podlagi podatkov (ang. data – driven MT) in še hibridni sistemi za strojno prevajanje. V nadaljevanju bodo na kratko opisani vsi ti sistemi kakor tudi njihovi podsistemi. Podrobneje si bomo ogledali skupino prevajalnih sistemov, kateri delujejo na osnovi pravil strukturnega prenosa.

2.1.1 Strojno prevajanje na podlagi podatkov

Strojno prevajanje na podlagi podatkov je korpusno, kar pomeni da že obstoječe prevode iz korpusa uporabljamo za ustvarjanje novih prevodov. Zato takšnemu strojnemu prevajanju pravimo prevajanje na osnovi korpusov (ang. Corpus-Based - CBMT). Sistemi za strojno prevajanje na podlagi podatkov so razdeljeni v dve fazi in sicer: faza učenja, v kateri se pripravi množica referenčnih prevodov in faza prevajanja, v kateri te množice referenčnih prevodov nastopajo kot osnova za prevode novih besed. Skupina sistemov za strojno prevajanje na podlagi podatkov je razdeljena na dve podskupini:

- Sistemi statističnega strojnega prevajanja (ang. Statistical Machine Translation - SMT)
- Sistemi strojnega prevajanja na osnovi primerov (Example Based Machine Translation - EBMT)

Primeri za takšne sisteme so: Google Translate, Egzpt toolkit, Moses, Genpar toolkit.

2.1.2 Hibridni sistemi za strojno prevajanje

Hibridni sistemi lahko zavzemajo dobre strani statističnih prevajalnih sistemov, ki temeljijo na pravilih strukturnega prenosa. Glede na to kako uporabljajo katero metodo, se hibridni sistemi delijo na dve skupini:

- Najprej je metoda na osnovi pravil, potem statistična metoda (ang. Rules post-processed by statistics): besedilo se najprej prevede na osnovi pravil, potem se besedilo popravi in prilagodi s statistično metodo.
- Statistična metoda, katero uravnava metoda na osnovi pravil (ang. Statistics guided by rules): najprej se uporabljajo pravila strukturnega prenosa za pripravo besedila z namenom boljšega rezultata pri prevajanju s statistično metodo, potem zopet uporabljamo pravila strukturnega prenosa za normalizacijo prevoda.

2.1.3 Strojno prevajanje na osnovi pravil

Sistemi za strojno prevajanje na osnovi pravil (ang. Rule-Based MT) uporabljajo zbirke pravil strukturnega prenosa. Izhodiščno besedilo se najprej analizira na oblikoslovni ravni. Pomanjkljivost takšnih sistemov je v tem, da bolj kot je sistem izpopolnjen, težje ga je razširiti in nadgraditi, to je tako ker so za razvijanje takšnih sistemov potrebna kompleksna pravila in jih je težje dodajati v že obstoječo in obsežno bazo medsebojno odvisnih pravil. Način izdelave vseh RBMT sistemov je približno enak, način zapisa pravil pa se razlikuje od sistema do sistema. Skupini RBMT sistemov pripada večina današnjih komercialnih prevajalnih sistemov, kot so Apertium, Systran, Prompt,...

RBMT sistemi vzamejo izvorno besedilo in ga najprej skladenjsko označijo ter skladenjsko razčlenijo, potem se izdeluje predstavitev vhodnega besedila v obliki drevesa skladenjske izpeljave (ang. parse-tree) (Vičič, 2012). Ta predstavitev se potem še dodatno abstrahira. Abstraktna predstavitev vhodnega besedila služi zato, da bi se s prenosom prevedel v podobno predstavitev ciljnega jezika kateri sistem uporabi kot osnovo za tvorjenje besedila v ciljnem jeziku. Metoda katera uporablja vmesni jezik, transferna metoda in metoda na osnovi slovarja sta metodi kateri delujeta na osnovi pravil.

2.1.4 Strojno prevajanje na osnovi pravil plitkega prenosa

Pristopi prevajanja na osnovi pravil plitkega prenosa (ang. shallow transfer rule – based translation) ne prevzemajo znanja za besedilo. Analiza tako izvornega kot tudi ciljnega jezika je omejena na obliko-skladenjske oznake. Večina sistemov za strojno prevajanje na osnovi plitkega prenosa so zgrajeni na podoben način, tipičen sistem je razdeljen na več komponent oziroma modulov, slednji so podrobneje opisani v naslednjih podtočkah:

- Obliko-skladenjska analiza: uporablja se enojezični slovar izvirnega jezika, kateri vsebuje obliko-skladenjske oznake za vsako besedo izvirnega besedila, katere bi jih besedna oblika lahko imela.
- Razdvoumljanje - MSD označevalec (ang. disambiguation): poskuša ugotoviti smisel besede iz vsebinske značilnosti, torej izbere najverjetnejšo obliko-skladenjsko oznako za posamezno besedo vhodnega besedila glede na okolico. Modul temelji na označevalniku obliko-skladenjskih oznak ali pa na omejenih slovnicah (ang. constraint grammar).
- Strukturni prenos: uporablja plitka pravila za leksikalni in strukturni prenos, da bi označeno besedilo izvirnega jezika prenesel v ciljni jezik. To naredi s pomočjo pravil strukturnega prenosa iz dvojezičnega prevajalnega stroja.
- Obliko-skladenjska sinteza: uporablja enojezični slovar ciljnega jezika, kateri vsebuje obliko-skladenjske oznake. Nadomešča označeno besedilo izvirnega jezika z dejanskimi besednimi oblikami v ciljnem jeziku.

Na sliki 2.1 so prikazane vse posamezne komponente, oziroma moduli tipičnega sistema za strojno prevajanje na osnovi plitkega prenosa, katerega smo zgoraj podrobno opisali.

Slika 2.1: Komponente oziroma moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Arhitektura temelji na sistemu, predlaganem v (Hajič et al., 2003) in pozneje uporabljena tudi v (Corbi-Bellot et al., 2005).

2.1.5 Platforma za strojno prevajanje Apertium

Apertium je odprtokodna platforma za strojno prevajanje na osnovi pravil, katera je na začetku bila namenjena prevajanju med sorodnimi jeziki, vendar so se njene zmožnosti razširile tudi na manj sorodne jezikovne pare (Apertium, 2010). Izdana je pod licence GNU GPL. Apertium je nastal kot eden izmed sistemov za strojno prevajanje na

projektu OpenTrad, kateri je bil razvit s podporo Španske vlade. Za implementacijo novih jezikovnih parov moramo razviti jezikovne podatke (slovarji, pravila) v dobro definirani XML obliki.

Jezikovni podatki Aperituma trenutno podpirajo aragonske jezike, danski, angleški, islandski, italijanski, makedonski, norveški, okcitanske jezike, portugalski, rumunski, slovenski, srbski, hrvaški, španski, švedski, velški in še nekaj drugih jezikov. Pri razvijanju Aperiuma prav tako pomagajo podjetja kot so, Prompsit Language Engineering, Imaxin Software in Eleke Ingeniaritza Linguistika.

Strojno prevajalni sistemi, kateri so razviti z pomočjo Apertiuma, temeljijo na pravilih plitkega prenosa. Uporabljajo končne pretvornike za svojo transformacijo kakor tudi skrit model Markova za označevanje izgovarjanja, oziroma ugotavljanja besednih vrst.

3 Metodologija

V naslednjem poglavju bom predstavil vse značilnosti, katerim sem se posebej posvetil v posameznih komponentah prevajalnega sistema.

Predstavljene so samodejno zgrajena gradiva in metode, katere so omogočile izdelavo gradiv želene kakovosti kakor tudi predelave oziroma izdelave novih pravil na že obstoječa gradiva za uporabo v novem sistemu. Obliko-skladenjsko označevanje gradiv v obeh jezikih je zelo različno in to je bil razlog za dodatno delo.

3.1 Enojezični slovar izvirnega jezika - slovenščina

Kot smo že omenili, način izdelave pri vseh RBMT sistemih je precej enak, zato ni bilo veliko dodatnega dela pri izdelavi enojezičnega slovarja za oba jezika. Edino kar je bilo potrebno je bil le kratek pregled in vnos novega gradiva, v veliki večini je bilo le to dodajanje novih besed, kljub temu da sta si jezika slovenščina in makedonščina slovnično zelo različna. Kot osnovo sem uporabljal že obstoječi in zgrajen enojezični slovar iz slovenščine, kateri je bil razvit za potrebe prevajalnega sistema za jezikovni par srbski in hrvaški jezik - slovenski jezik. Obliko-skladenjske oznake besednih oblik so enake, saj je ta slovar že zelo dobro zgrajen in ni bilo potrebno dodajanja novih. V nadaljevanju je podrobneje predstavljeno delo, katero je bilo opravljeno na posameznih besednih vrstah, oziroma skupin besed.

3.1.1 Samostalniške besede

Samostalniške besede so razdeljene na dve glavni skupini, in sicer na splošna in lastna imena. Lastna imena so bolj zanimiva za nas, ker zahtevajo veliko večjo pregledno kategorizacijo. Delijo se v tri kategorije in sicer: imena, priimki in imena krajev. Ena izmed slovničnih razlik med obema jezikoma je, da moramo pri slovenščini tako pri imenu kakor tudi pri priimku dodajati obliko-skladenjske oznake ter oblike za svojilno pridevniško obliko.

Različne skupine imen oziroma priimkov imajo različne svojilne oblike. Problem je rešen z uvedbo paradigem, katera vsebujejo prevzete končnice. Za vsako skupino imen oziroma priimkov je dodan še vmesni člen oziroma končnica, katera pripada imenu ali priimku v svojilni obliki z naslednjimi lastnostmi: osnovnik, moški spol, ednina,

imenovalnik. Na sliki 3.1 je prikazan primer svojilnih pridevniških oblik za imena Branko in Marija.

Brank		**Marij**	
Brank – ov	(I)	Marij – in	(I)
Brank – ov – ega	(D)	Marij – in – ega	(D)
Brank – ov – emu	(R)	Marij – in – emu	(R)
Brank – ov – ega	(T-Ž)	Marij – in – ega	(T-Ž)
Brank – ov	(T-N)	Marij – in	(T-N)
Brank – ov – em	(M)	Marij – in – em	(M)
Brank – ov – im	(O)	Marij – in – im	(O)

Slika 3.1: Primer svojilne pridevniške oblike za imena oziroma priimke.

Iz primera na sliki 3.1 je razvidno, da so imena oziroma priimki razdeljeni v tri člene in sicer:

- Jedro imena oziroma priimka → **Brank** ali **Marij**
- Končnica svojilne pridevniške oblike (os., m., ed., im.) → **ov** ali **in**
- Privzete svojilne pridevniške končnice, ki veljajo za vsa imena in priimke:

```
<e><p><l></l>      <r><s n="ma"/><s n="sg"/><s n="nom"/></r></p></e>
<e><p><l>ega</l>   <r><s n="ma"/><s n="sg"/><s n="gen"/></r></p></e>
<e><p><l>emu</l>   <r><s n="ma"/><s n="sg"/><s n="dat"/></r></p></e>
<e><p><l>ega</l>   <r><s n="ma"/><s n="sg"/><s n="acc"/></r></p></e>
<e><p><l></l>      <r><s n="mi"/><s n="sg"/><s n="acc"/></r></p></e>
<e><p><l>em</l>    <r><s n="ma"/><s n="sg"/><s n="loc"/></r></p></e>
<e><p><l>im</l>    <r><s n="ma"/><s n="sg"/><s n="ins"/></r></p></e>

<e><p><l>a</l>     <r><s n="ma"/><s n="du"/><s n="nom"/></r></p></e>
<e><p><l>ih</l>    <r><s n="ma"/><s n="du"/><s n="gen"/></r></p></e>
<e><p><l>ima</l>   <r><s n="ma"/><s n="du"/><s n="dat"/></r></p></e>
<e><p><l>a</l>     <r><s n="ma"/><s n="du"/><s n="acc"/></r></p></e>
<e><p><l>ih</l>    <r><s n="ma"/><s n="du"/><s n="loc"/></r></p></e>
<e><p><l>ima</l>   <r><s n="ma"/><s n="du"/><s n="ins"/></r></p></e>

<e><p><l>i</l>     <r><s n="ma"/><s n="pl"/><s n="nom"/></r></p></e>
<e><p><l>ih</l>    <r><s n="ma"/><s n="pl"/><s n="gen"/></r></p></e>
<e><p><l>im</l>    <r><s n="ma"/><s n="pl"/><s n="dat"/></r></p></e>
<e><p><l>e</l>     <r><s n="ma"/><s n="pl"/><s n="acc"/></r></p></e>
<e><p><l>ih</l>    <r><s n="ma"/><s n="pl"/><s n="loc"/></r></p></e>
```

Slika 3.2: Paradigma s privzetimi svojilnimi pridevniškimi končnicami.

Na sliki 3.2 lahko si ogledamo paradigmo za elemente moškega spola v ednini, dvojini in množini. Zakaj je za nas toliko zanimiva prav ta paradigma? To je zaradi tega, ker imamo edino za moški spol ednino v dveh tožilnih oblikah, ti dve obliki sta posledici lastnosti živosti ali neživosti, katero nosi naslednja beseda - samostalnik, kateri se nahaja po lastnem imenu oziroma priimku. Primer živosti kakor tudi neživosti si lahko ogledamo na sliki 3.3.

Vzel sem ***Brankov računalnik.*** *(neživ)*
Vzel sem ***Brankovega psa.*** *(živ)*

Slika 3.3: Primer živosti ter neživosti.

3.1.2 Pridevniki in prislovi

V slovenski slovnici se pridevniki in prislovi stopnjujejo štiri-stopenjsko in sicer, kot osnovnik, primernik, presežnik in elativ. Glede na to, da se v makedonski slovnici pridevniki in prislovi stopnjujejo tri-stopenjsko smo morali konstruirati novo paradigmo katera ustreza slovenskemu elativu, vendar o tem bomo govorili podrobneje v nadaljevanju poglavja. Na sliki 3.4 si lahko ogledamo primer štiri-stopenjskega prislova.

Osnovik:	*svežo*
Primernik:	*svežeje, svežejše*
Presežnik:	*najsvežeje, najsvežejše*
Elativ:	*presvežo*

Slika 3.4: Prislov **svežo** v vseh štirih oblikah.

Potrebno je bilo paziti tudi na besede, za katere obstaja samo osnovnik oziroma za katere obstajajo različne kombinacije vseh štirih oblik. Lastnosti so enake tako za prislove kakor tud za pridevnike. Pri pridevnikih so paradigme osnovnih oblik vezane tudi za sekundarne paradigme, katere vsebujejo še podatke kot so na primer spol, število, sklon ter določenost, za razliko od pridevnikov, prislovi niso vezani za sekundarne paradigme, ampak samo za stopnjo in to je edina razlika.

3.1.3 Glagoli

Glagolske paradigme vsebujejo oblike za glagole v nedoločniku, namenilniku, povedniku, velelniku, deležniku na -n / -t kakor in deležniku na -l. Poleg glagolskih oblik je potrebno določiti tudi glagolsko vrsto, katera označuje ali je glagol končan ali nedokončan ali pa eno in drugo, potrebno je določiti tudi glagolsko prehodnost katera zazna ali je glagol prehoden, neprehoden ali hkrati vse skupaj in še obliko povratnih glagolov. Generirano gradivo je že vsebovalo vse te glagolske paradigme in ni bilo potrebno veliko dodatnega dela. Problem je bil v veliki slovnični razliki med jezikoma slovenščina in makedonščina in vse te paradigme enojezičnega slovarja izvirnega jezika je bilo potrebno povezati z ustrezno paradigmo enojezičnega slovarja ciljnega jezika, in vse skupaj implementirati v dvojezični slovar. Na sliki 3.5 si lahko ogledamo nekaj primerov glagolskih oblik, primer za glagolsko vrsto in še primer za glagolsko prehodnost.

Glagolske oblike
Nedoločnik: *igrati*
Namenilnik: *igrat*
Povednik (sed): *igram, igraš, igra, ...*
Velelnik (sed): *igraj, igrava, igrata, igramo, igrate*
Deležnik na -n/-t: *igran, igrana, igrano, ...*
Deležnik na -l: *igral, igrala, igralo, ...*

Glagolski vid
Dovršen: *Marija je prebrala knjigo.*
Nedovršen: *Miha dela v avtopralnici.*

Glagolska prehodnost
Prehodni: *Marija izpolnjuje anketo.*
Neprehodni: *Miha spi.*

Slika 3.5: Primeri za glagolske oblike, glagolski vid in glagolsko prehodnost.

3.1.4 Števniki

Števniki so kategorizirani v tri skupine: glavni, drugi in vrstilni. Vsakemu števniku pripada paradigma, katera določa število, spol kakor tudi sklon.

3.1.5 Zaimki

Zaimki se delijo na: osebne, svojilne, oziralne, vprašalne, kazalne, celotne, povratne in nedoločne. Pri zaimkih ni bilo potrebno veliko dela, poleg podrobnega pregleda ali so le ti pravilno kategorizirani.

3.1.6 Predlogi

V slovenski slovnici je vsak predlog označen z obliko-skladenjsko oznako za sklon, v katerem se lahko uporabi. V enojezičnemu slovarju izvirnega jezika označevanje sklona ni bilo potrebno zaradi enojezičnega slovarja ciljnega jezika, saj se v makedonski slovnici ne uporablja sklon.

3.1.7 Ostalo

Obstajajo še nekatere druge besedne vrste kot so: vezniki, členki, medmeti kakor tudi kratice. Te besedne vrste ne potrebujejo dodatnih oznak, ampak samo oblikoskladenjsko oznako, katera označuje besedno vrsto.

3.2 Enojezični slovar ciljnega jezika - makedonščina

Podobno kot enojezični slovar izvirnega jezika, smo pri enojezičnemu slovarju ciljnega jezika vzeli kot osnovo že obstoječi slovar iz jezikovnih gradiv prevajalnega sistema srbskega in hrvaškega jezika - makedonski jezik. Potrebno je bilo samo dodati prevode slovenskih besed, katere niso bile prisotne v tem slovarju. Prevzete paradigme v veliki večini so zadoščale našim potrebam. V nadaljevanju je podrobneje predstavljen postopek iskanja in dodajanja prevodov manjkajočih besed.

3.2.1 Iskanje in dodajanje prevode manjkajočih besed

Zaradi lažjega pregleda, sem se odločil graditi oba enojezična slovarja vzporedno. Pri grajenju enojezičnega slovarja ciljnega jezika je bilo potrebno samodejno dodajanje prevodov nekaterih slovenskih besed. Postopek iskanja in dodajanja manjkajočih prevodov besed zgleda tako:

1. Prevedemo besedo iz slovenskega jezika v makedonski jezik
2. Preverimo ali beseda že obstaja v prevzetem slovarju
3. Če beseda obstaja v prevzetem slovarju:

(a) Preverimo ali je beseda povezana z pravilno paradigmo

(b) Če je beseda povezana z pravilno paradigmo jo dodamo v nov slovar ciljnega jezika

(c) Če beseda ni povezana z pravilno paradigmo, poiščemo pravilno paradigmo, prepišemo besedo z pravilno paradigmo in jo dodamo v nov slovar ciljnega jezika

4. Če beseda ne obstaja v privzetem slovarju:

(a) Poiščemo paradigmo, katera naj bi bila pravilna za skupino besed kateri pripada naša beseda

(b) Povežemo našo besedo z paradigmo, katero smo že izbrali in jo dodamo v nov slovar ciljnega jezika

Obstoječi slovar, katerega smo uporabljali kot osnovo je bil zelo dobro zgrajen in prevzete paradigme so prekrivale vse potrebne skupine besed, tako da, ni bilo potrebno dodajati novih paradigem.

3.3 Dvojezični prevajalni slovar

Dvojezični prevajalni slovar (znan tudi kot bidix ali transfer-leksikon) vsebuje prevod med obema jezikoma, torej vsebuje besede enojezičnih slovarjev in ustrezne prevode z vsemi ustreznimi obliko-skladenjskimi oznakami. Jezikovni par slovenski jezik - makedonski jezik ni bil razvit na platformi Apertiuma, zato smo ga morali samodejno razviti od začetka. Pri grajenju dvojezičnega slovarja sem kot pomoč uporabljal sistem Google Translate kakor tudi dvojezične obstoječe prevajalne slovarje, napake in manjkajoče prevode pa sem pripravili ročno.

Pri samem razvijanju dvojezičnega prevajalnega slovarja sem naletel na velike težave, in to zaradi tega, ker nisem imeli že zgrajen dvojezični prevajalni slovar za ta jezikovni par, da bi ga izkoristili kot osnovo pa tudi zaradi prevelike slovnične razlike med obema jezikoma, slovenščina in makedonščina. V nadaljevanju so podrobneje opisane težave in rešitve le teh.

3.3.1 Poenotenje oblikoskladenjskih oznak izvornega ter ciljnega enojezičnega slovarja

Preden začnemo z vstavljanjem in prevajanjem besed v dvojezičnem prevajalnem slovarju moramo pregledati oba enojezična slovarja. Napisati moramo izbor pravil po katerih se bodo reševale vse spremembe oziroma razlike obliko-skladenjskih oznak med

izvirnim in ciljnim jezikom. Prav tako je pomemben tudi vrstni red obliko-skladenjskih oznak. Najprej moramo poskrbeti, da se obliko-skladenjske oznake izvirnega in ciljnega jezika pokrivajo, vrstni red pa rešujemo z pravili strukturnega prenosa, za vse drugo uporabljamo nabor pravil katera smo napisali.

Odvisno od sprememb oziroma razlik obliko-skladenjskih oznak, lahko slednje rešujemo na dva načina: Prvi način je, kadar moramo narediti *1:1* spremembo in na drugi način kadar moramo narediti *1:n* ali *n:1* spremembo. Način, kako lahko rešimo oba tipa sprememb je dokaj enostaven in sicer, zamenjamo obliko-skladenjske oznake izvirnega in ciljnega jezika v vnosu. Primer zamenjave obliko-skladenjske oznake (*1:1*):

- **ustroj** <samostalnik> <**moški**> ⇒ **структура**(struktura) <samostalnik> <**ženski**>
- **barvilo** <samostalni> <**srednji**> ⇒ **боја**(boja) <samostalnik> <**ženski**>

Predstavljena rešitev je optimalna, vendar lahko postane zelo neprimerna v primeru *1:n* ali *n:1* spremembi:

- **gospoda** <sam> <m> <**ed**> ⇒ **господа** (gospoda) <sam> <m> <**mn**>
- **gospoda** <sam> <m> <**dv**> ⇒ **господа** (gospoda) <sam> <m> <**mn**>
- **gospoda** <sam> <m> <**mn**> ⇒ **господа** (gospoda) <sam> <m> <**mn**>

V tem primeru smo prisiljeni dodati paradigmo, katera bi poskrbela za prenos *1:n* ali *n:1* določene obliko-skladenjske oznake. Kakor lahko vidimo v primeru za besedo gospoda se vse oblike iz slovenskega jezika, torej ednina, dvojina in množina, prevedejo v množino v makedonskem jeziku. Da bi se izognili dodajanju trem različnim vnosom za isto besedo, dodajamo zgoraj omejeno paradigmo. Primer paradigme si lahko ogledamo na sliki 3.6.

```
<pardef n="sgpl_pl">
    <e r="LR"><p><l><s n="sg"/></l><r><s n="pl"/></r></p></e>
    <e r="LR"><p><l><s n="du"/></l><r><s n="pl"/></r></p></e>
    <e><p><l><s n="pl"/></l><r><s n="pl"/></r></p></e>
</pardef>
```

Slika 3.6: Paradigma za odpravljanje nepotrebnih vnosov v dvojezičnem slovarju.

Iz slike 3.6 je razvidno, da se paradigma imenuje **sgpl_pl**. Tukaj smo dodajali oznako **LR**, ta oznaka označuje v kateri smeri so dovoljenje izjeme, torej naša oznaka **LR** označuje, da se izjema upošteva pri prevajanju iz slovenskega jezika v makedonski jezik. Ko imamo paradigmo že pripravljeno, sam vnos v dvojezični slovar zgleda tako:

- **gospoda** <sam> <m> ⇒ **господа** (gospoda) <sam> <m> <par n=”sgpl_pl”/>

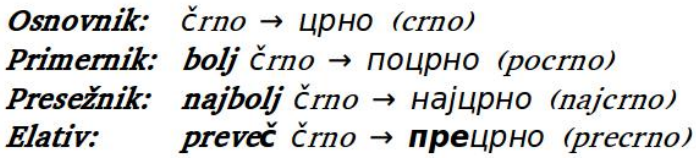
V zgornjem primeru lahko opazimo, da nam paradigma omogoča, da se izognemo podvojevanju v dvojezičnem slovarju, torej vnašanje novih vnosov in odpravljanje napak nam paradigme zelo olajšajo.

3.3.2 Stopnjevanje pridevnikov ter prislovov

Prva večja težava, na katero sem naletel je bila razlika v obliko-skladenjskem označevanju stopenj. To je zato, ker se v makedonskem jeziku prislovi in pridevniki ne stopnjujejo štiri-stopenjsko in sem moral zato z uporabo dodatne paradigme, katera določa preslikavo obliko-skladenjskih oznak rešiti ta problem. Primer obliko-skladenjskih oznak zgleda kot sledi:

- Osnovnik: <adv> ⇒ <adv>
- Primernik: <adv> <comp> ⇒ <adv> <comp>
- Presežnik: <adv> <sup> ⇒ <adv> <sup>
- Elativ: <adv> <ela> ⇒ <adv> <ssup>

Najprej je bilo potrebno pred prislove in pridevnike v ciljnem jeziku dodati besedo pre (cyr. пре) kot predpona in tako smo naredili še eno stopnjo v makedonščini, katera ustreza slovenskemu elativu. Prav tako so nastale težave pri prevajanju besed, katere niso imele enakega števila oziroma istih stopenj v izvirnem kakor tudi v ciljnem jeziku. Težave so bile rešene tako, da smo pred prislove in pridevnike izvirnega jezika dodali besedo bolj, najbolj ali preveč, odvisno od manjkajoče oblike. Primer prevajanja iz slovenskega jezika v makedonski jezik je prikazan na sliki 3.7.



Slika 3.7: Primer prevajanja prislova iz slovenskega jezika v makedonski jezik z manjkajočimi oblikami.

Da bi dosegli omenjeno stopnjevanje smo v dvojezičnemu slovarju dodali vnos za vsako stopnjo posebej. Problem z manjkajočimi stopnjami sem rešil tako, da sem dodal dodatne obliko-skladenjske oznake katere pravilo strukturnega prenosa zazna, ter pravilno doda besedo bolj, najbolj ali preveč. Na sliki 3.8 je dan primer vnosov za prevod *črno* → *црно* (crno).

```
<e><p><l>črn<s n="adj"/></l><r>црн<s n="adj"/></r></p><par n="only_pst2null"/></e>
<e><p><l>črn<s n="adj"/><s n="add_comp"/></l><r>црн<s n="adj"/><s n="comp"/></r></p></e>
<e><p><l>črn<s n="adj"/><s n="add_sup"/></l><r>црн<s n="adj"/><s n="sup"/></r></p></e>
<e><p><l>črn<s n="adj"/><s n="add_ela"/></l><r>црн<s n="adj"/><s n="ssup"/></r></p></e>
```

Slika 3.8: Dvojezični vnosi za manjkajoče stopnje z dodatnimi oblikoskladenjskimi oznakami.

Iz slike 3.8 lahko opazimo dodatne obliko-skladenjske oznake **add_comp**, **add_sup** ter **add_ela**, kateri so potrebni za pravilno prevajanje manjkajočih stopenj. Opazimo lahko paradigmo z imenom **only_pst2null**, katera poskrbi, da se obliko-skladenjske oznake v osnovniku pravilno prevedejo.

3.4 Pravila strukturnega prenosa

V tem poglavju bomo strmeli za izjemnimi pravili strukturnega prenosa. Obstajajo tri nivoji pravil strukturnega prenosa. Glede tega, da je struktura obeh jezikov precej

različna smo imeli precej izjem pri pravilih strukturnega prenosa. Za lažje razumevanje bom predstavil, tako osnovna kakor tudi zahtevnejša pravila z konkretnimi primeri.

3.4.1 Urejanje vrstnega reda obliko-skladenjskih oznak izvornega ter ciljnega enojezičnega slovarja s pomočjo makrojev.

V enem od predhodnih poglavij, točneje v podpoglavju 3.3.1 sem omenil, da je vrstni red obliko-skladenjskih oznak izvirnega ter ciljnega enojezičnega slovarja zelo pomemben za pravilno prevajanje besed. Da bi dosegli vrstni red prekrivanja obliko-skladenjskih oznak izvirnega in ciljnega slovarja, uvedli smo nekaj makro-jev. S pomočjo makro-jev lahko definiramo vrstni red obliko-skladenjskih oznak. Na spodnji sliki si lahko ogledamo primer makro-ja, kateri ureja vrstni red obliko-skladenjskih oznak samostalniškim besedam.

```
<def-macro n="urediSamostalnik" npar="1">
 <let>
  <clip pos="1" side="tl" part="whole"/>
  <concat>
    <clip pos="1" side="tl" part="lema"/>
    <clip pos="1" side="tl" part="samostalnik"/>
    <clip pos="1" side="tl" part="spol"/>
    <clip pos="1" side="tl" part="število"/>
    <clip pos="1" side="tl" part="sklon"/>
  </concat>
 </let>
</def-macro>
```

Slika 3.9: Primer makrota, ki ureja vrstni red oblikoskladenjskih oznak samostalniškim besedam.

Izhod ter vrstni red obliko-skladenjskih oznak:

- Beseda <samostalnik> <spol> <število> <sklon>

Izhod na konkretnem primeru:

- Inštruktor <samostalnik> <moški> <ednina> <imenovalnik>

Glede tega, da v makedonskem jeziku ne uporabljamo sklone, oznaka za sklon ostane prazna.

3.4.2 Stopnjevanje pridevnikov ter prislovov z manjkajočimi oblikami

V podpoglavju 3.3.2 sem opozoril na manjkajoče stopnjevanje pridevnikov ter prislovov in opisali pristop za njihovo rešitev. Rešitev, katero sem opisal je bila v dvojezičnemu slovarju, v nadaljevanju pa bom opisal pravilo strukturnega prenosa katero poskrbi, da se ta rešitev pravilno izvede.

Dodatne obliko-skladenjske oznake (add_comp, add_sup ter add_ela), katere sem dodal v dvojezični slovar opozarjajo, da moramo dodati dodatne besede. Pravilo strukturnega prenosa katero je predstavljeno na sliki 3.10, skrbi za pravilno razvrstitev dodatnih besed.

```
<rule comment="Prislovi">
<pattern>
 <pattern-item n="prislovi"/>
</pattern>
<action>
<choose>
 <when>
  <test>
   <and>
    <equal><clip pos="1" side="sl" part="stopnjaSL"/><lit-tag v="add_ela"/></equal>
    <equal><clip pos="1" side="tl" part="stopnjaMK"/><lit-tag v="ssup"/></equal>
   </and>
  </test>
  <out>
   <lu>
    <lit v="veliko"/>
    <lit-tag v="adv.sint.ela"/>
   </lu>
   <b/>
   <lu>
    <clip-pos="1" side="tl" part="lema"/>
    <clip-pos="1" side="tl" part="prislov"/>
    <lit-tag v="sint"/>
   </lu>
  </out>
 </when>
...
```

Slika 3.10: Pravilo strukturnega prenosa, ki ureja prislove.

Prikazano pravilo na sliki 3.10 je zelo enostavno, skrbi pa za prislove kateri so v elativu. Najprej pravilo sprejme en vzorec in sicer prislove, potem z označbo <equal>, preveri ali je obliko-skladenjska oznaka izvirnega jezika enaka *add_ela*, ter za tem preveri ali je obliko-skladenjska oznaka ciljnega jezika enaka *ssup*. V primeru ko sta pogoja izpolnjena, pravilo odda v izhodu dve besedi, in sicer:

- preveč ⇒ veliko <prislov> <sint> <ela>
- prislov ⇒ prislov <prislov> <sint>

3.4.3 Specifična pravila s primeri

Do sedaj smo si ogledali makro, s katerim urejamo vrstni red obliko-skladenjskih oznak in pravilo strukturnega prenosa, z katerim smo dodajali besede bolj, najbolj ali preveč, manjkajočim stopnjam pridevniških in prislovnih besed. Sedaj si bomo ogledali še eno specifično pravilo strukturnega prenosa.

Kot primer sem vzel pravilo strukturnega prenosa katero ureja vzorce biti + poljubni glagol z lastnostjo L-participle. Primer pravila si lahko ogledamo na sliki 3.11.

```
<rule comment="Glagol biti + Glagol L-participle">
<pattern>
  <pattern-item n="glagol_biti"/>
  <pattern-item n="glagol_L-participle"/>
</pattern>
<action>
  <out>
   <lu>
    <clip pos="1" side="tl" part="lema"/>
    <clip pos="1" side="tl" part="glagol_biti"/>
    <clip pos="1" side="tl" part="oblika"/>
    <clip pos="1" side="tl" part="oseba"/>
    <clip pos="1" side="tl" part="število"/>
   </lu>
   <b/>
   <lu>
    <clip pos="2" side="tl" part="lema"/>
    <clip pos="2" side="tl" part="glagol"/>
    <clip pos="2" side="tl" part="glagolska_dovršnost"/>
    <clip pos="2" side="tl" part="glagolska_prehodnost"/>
    <clip pos="2" side="tl" part="spol"/>
    <clip pos="2" side="tl" part="število"/>
   </lu>
  </out>
</action>
</rule>
```

Slika 3.11: Pravilo strukturnega prenosa, ki ureja vzorca biti + poljubni glagol z lastnostjo L-participle.

Iz slike 3. 11 je razvidno, da pravilo pri vhodu prejme dva vzorca, glagol biti in poljubni glagol z lastnostjo *L-participle*. Na izhodu v drugi leksikalni enoti lahko opazimo, da smo opustili oznako katera določa obliko glagola. Obliko smo namenoma opustili, saj v makedonskem enojezičnem slovarju nismo dodatno označili deležnika na

-l. V nadaljevanju si lahko ogledamo vpliv pravila strukturnega prenosa.

- bi igral
  - biti <glagol> <sed> <1. os> <ed>
  - igrati <glagol> <nedov> <nepreh> <L-part> <m> <ed >
- би играл (bi igral)
  - би(bi) <glagol> <sed> <1. os> <ed>
  - игра(igra) <glagol> <nedov> <nepreh> <m> <ed>

4 Vrednotenje

V naslednjem poglavju bom predstavil vrednotenje sistemov za strojno prevajanje. Vrednotenje oziroma evalvacija je zelo pomembna. Na podlagi vrednotenja lahko določimo kakovost prevajalnega sistema v praksi. Obstaja več različnih metod vrednotenja in vse dajejo različne rezultate. Za katero metodo vrednotenja naj bi se mi odločili? Vrednotenje je zelo subjektivno in kompleksno, zato še vedno ni določena univerzalna metoda ocenjevanja. Poglejmo si najprej različne kriterije ocenjevanja, potem pa se bomo vrnili na te metode vrednotenja, katere smo izbrali kot najbolj ustrezne za naš prevajalni sistem.

- Hutchins in Somers (1992) navaja tri kriterije:
  - informativnost: v kolikšni meri je informacija v prevodu enaka informaciji izvornega besedila,
  - razumljivost: ali je prevod jasen in razumljiv,
  - ustreznost jezika: ali je v prevodu uporabljen jezik, primeren vsebini in sporočilu.
- Konzorcij LDC (2005) priporoča dva kriterija z izdelanimi lestvicami:
  - vsebinska ustreznost prevodov ter
  - slovnična pravilnost prevodov v ciljnem jeziku.
- Statistični pristopi: v tej skupini se nahajajo vse samodejne metode, ki omogočajo oprijemljivejše ocene. Vse metode iz te skupine primerjajo število napak različnih vrst.

V svojem projektu sem se odločil, da bom uporabljal samodejne metode vrednotenja, katere temeljijo na primerjavi števila napak različnih vrst. Napake so določene kot razlike med referenčnim prevodom in prevodom prevajalnega sistema. Ogledali si bomo dve metodi, in sicer: metodo za vrednotenje kakovosti prevodov in metodo za vrednotenje kakovosti jezikovnih gradiv. V nadaljevanju bo podrobneje predstavljena vsaka od metod posebej, v naslednjem poglavju pa bomo predstavili rezultate, katere smo dosegli z uporabo omenjenih metod.

4.1 Vrednotenje kakovosti prevodov

Pri vrednotenju kakovosti prevodov sem uporabil samodejno metodo WER (ang. word error rate), katera računa stopnjo napačnih besed. Ta metoda se najpogosteje uporablja za vrednotenje kakovosti prepoznavanja govora in strojnega sistema prevajanja, uporablja pa se tudi za primerjavo kakovosti različnih prevajalnih sistemov ali vrednotenje izboljšav znotraj določenega prevajalnega sistema. Metoda WER nam ne da nobenih informacij za naravo napak v prevajalnem sistemu, zato je potrebno še dodatno delo za identifikacijo vzroka oziroma izvora napake (ten Thije, Zeevaert, 2007).

Utežena Levenshteinova razdalja (ang. weighted Levenshtein edit-distance) predstavlja razširitev osnovne razdalje, katera šteje najmanjše število sprememb, katere moramo odpraviti med referenčnim prevodom in prevodom prevajalnega sistema, potem pa še utežimo število sprememb z dolžino povedi. Metoda WER temelji na uteženi Levenshteinovi razdalji (ten Thije, Zeevaert, 2007). Dovoljene spremembe so vstavitev, brisanje in zamenjava besed.

Vrednost WER izračunamo po naslednji metodi:

$$WER = \frac{S + D + I}{N}, \tag{4.1}$$

kjer je

- S - število substitucij,
- D - število izbrisov,
- I - število vstavkov,
- N - število besed v povedi.

Rezultat, ki ga dobimo pri uporabe WER metrike predstavlja stopnjo napake prevajalnega sistema. Če želimo dobiti rezultat kakovosti prevajalnega sistema, uporabimo različico metrika, ki opisuje stopnjo prepoznavnih besed WRR (ang. word recognition rate). WRR je razlika med popolnim prevodom ter napako sistema.

Vrednost WRR izračunamo po naslednji metodi:

$$WRR = 1 - WER, \tag{4.2}$$

4.2 Vrednotenje kakovosti jezikovnih gradiv

Pri vrednotenju kakovosti jezikovnih gradiv ugotavljamo pokritost slovarjev in njihovo kvaliteto, na tak način lahko ugotovimo ali enojezični slovarji vsebujejo vse potrebne analize besed ali ne. Vrednotenje smo izvedli s pomočjo dveh korpusov, in sicer:

- MULTEXT-EAST: večjezična zbirka jezikovnih gradiv, zapisana v standardizirani obliki (Erjavec, 2004). Trenutna zbirka vsebuje 17 srednjih ter vzhodnoevropskih jezikov. Zbirka vsebuje gradiva za oblikoskladenjske specifikacije, oblikoskladenjske leksikone kakor tudi označeni vzporedni primerjalnih in govornih korpusov.
- OPUS: zbirka besedil iz spleta. Namen projekta OPUS je pretvoriti ter uskladiti brezplačne spletne podatke z dodatnimi jezikovnimi oznakami ter le-te prostodostopno ponuditi javnosti (OPUS, 2013). Pri tej zbirki smo se omejili le na zbirko podnapisov, saj je bila že sama po sebe dovolj obširna.

Rezultati vrednotenja kakovosti jezikovnih gradiv, so prikazani v naslednjem poglavju.

5 Rezultati

V naslednjem poglavju bom predstavil osnovne statistike jezikovnih gradiv katere sem ustvaril v sklopu razvijanja prevajalnega sistema. Ogledali si bomo rezultate, katere sem dosegel z metodo testov, vrednotenje kakovosti prevodov in pokritje slovarjev in korpusov.

5.1 Pokritost slovarjev

Glede na to da sem prevajalni stroj začeli graditi samodejno od začetka in sem vse besede dodajali ročno, lahko tudi brez vsakega testiranja zaključim, da enojezični slovar izvirnega jezika - slovenščina kakor tudi enojezični slovar ciljnega jezika makedonščina vsebujeta vsak posebej okoli 1000 unikatnih besed. Vendar, ne bomo pustili vse na predvidevanjih in bomo testirali oba enojezična slovarja, ter si ogledali pokritost dvojezičnega slovarja. Najbolj zanimivo za nas v tem testiranju bi bilo število paradigem, katere smo uporabili pri razvijanju enojezičnega slovarja. V tabeli 5.1. si lahko podrobneje ogledamo pokritost slovarjev.

Tabela 5.1: Pokritost slovarjev

Slovar	Št. besed
Enojezični slovar izvornega jezika - slovenščina	1216 (3816 paradigem)
Enojezični slovar ciljnega jezika - makedonščina	1197 (474 paradigem)
Dvojezični slovar	1476

Iz tabele 5.1. je razvidno, da so tako enojezični kakor dvojezični slovarji približno enako pokriti. Zanimivo je opaziti, da je število paradigem v enojezičnem slovarju izvirnega jezika dosti večje od števila paradigem v enojezičnemu slovarju ciljnega jezika, kar pa je delno zaradi slovnične razlike med obema jezikoma, ter zaradi tega, ker smo uporabljali že obstoječ enojezični slovar z vsemi prevzetimi paradigmami. Več informacij, glede pokritosti samih besednih vrst, je predstavljeno v naslednjem podpoglavju, kjer opisujem rezultate tako imenovane testvoc metode - metode za testiranje posameznih slovarjev.

5.2 Testiranje slovarjev

Slovar sem testiral s tako imenovano metodo *testvoc* (Testvoc, 2013). Metoda najprej razširi enojezični slovar izvirnega jezika in potem testira vsako možno analizo besede skozi faze prevajalnega sistema. Testvoc uporabljamo, da bi ugotovili katera analiza besede ima pravilen prevod v enojezičnem slovarju ciljnega jezika. Simboli kateri označujejo napake so # ali @, če se beseda pravilno prevede je potem brez teh simbolov. Pomen simbolov je sledeči:

- @ - v dvojezičnem slovarju ne obstaja prevod za to besedo
- # - oblikoskladenjske oznake niso pravilno označene, torej se beseda ne prevede pravilno

Testvoc testiranje je zelo pomembno v prvi fazi razvoja jezikovnega para, saj na tak način zagotovimo, da se vse analize besed iz enojezičnih slovarjev prevedejo pravilno.

Ko dobimo pozitivni rezultati pri testvoc testiranju lahko nadaljujemo na naslednjo fazo razvoja. Pri naslednji fazi razvoja pišemo pravila strukturnega prenosa, ki urejajo večbesedne prevode oziroma stavke. V tej fazi lahko pridemo do nezaželene ali nepričakovane napake, katere niso vidne pri enobesednih prevodih. Dalje potrebujemo še dodatno testiranje prevajalnega sistema na korpusih, kjer se lahko skrite napake pojavijo.

V tabeli 5.2 si lahko ogledamo rezultate testiranja enojezičnega slovarja izvornega jezika. Rezultati prikazujejo kakovost prevajanja posameznih besed iz slovenskega jezika v makedonski jezik.

Tabela 5.2: Rezultat Testvoc (Smer: slovenščina - makedonščina)

B. vrsta	Skupno	Pravilni	Z @	Z #	%
pridevniki	16844	16844	0	0	100
samostalniki	10093	10093	0	0	100
predlogi	9445	9445	0	0	100
glagoli	3044	3044	0	0	100
zaimki	2235	2235	0	0	100
števniki	1834	1834	0	0	100
prislovi	52	52	0	0	100
prir. vez.	20	20	0	0	100
podr. vez.	7	7	0	0	100

V tabeli 5.3 si lahko ogledamo rezultate testiranja enojezičnega slovarja ciljnega jezika. Rezultati prikazujejo kakovost prevajanja posameznih besed iz makedonskega jezika v slovenski jezik.

Tabela 5.3: Rezultat Testvoc (Smer: makedonščina - slovenščina)

B. vrsta	Skupno	Pravilni	Z @	Z #	%
glagoli	480	480	0	0	100
števniki	331	331	0	0	100
zaimki	93	93	0	0	100
predlogi	2	2	0	0	100

V tabeli 5.3 lahko opazimo da nekatere besedne vrste manjkajo. To je posledica pravil prenosa iz enojezičnega slovarja ciljnega jezika, torej so besede malo drugače razvrščene kot v enojezičnem slovarja izvornega jezika. Glede na to, da sem skoraj vse od začetka samodejno razvijal sem lahko zadovoljen z dobljenimi rezultati.

5.3 Rezultati vrednotenja kakovosti prevodov

Evalvacijo kakovosti prevodov sem izvedel z uporabo WER metode, katera je podrobneje opisana v podpoglavju 4.1. Rezultati, katere sem pridobil so zelo dobri, saj moramo upoštevati dejstvo, da sem samodejno razvijal sistem od začetka in le-ta še ni dokončan, pa še to, da sem se omejil le na prvi nivo pravil strukturnega prenosa. Samodejno matriko WER sem uporabljal na manjšem testnem vzorcu, ki je bil ročno pripravljen.

6 Zaključek in nadaljnje delo

6.1 Zaključek

Ta projekt predstavlja postavitev prevajalnega sistema za jezikovni par slovenščina - makedonščina. Sistem je zgrajen na osnovi pravil plitkega prenosa. Za razvijanje sistema sem uporabljal ogrodje Apertium, ker se je le-ta izkazal kot najprimernejši za postavitev sistema za strojno prevajanje sorodnih jezikov. Predstavil sem lastnosti izvornega ter ciljnega jezika, kaj vse je potrebno za razvijanja tovrstnega projekta ter težave oziroma omejitve, katere so bile prisotne pri gradnji gradiv. Pri postavitev ogrodja sistema sem uporabljal metode za samodejno izdelavo prevajalnega sistema, katere so predstavljene v delu (Vičič, 2012). Opravljeno delo pri razvijanju sistema je bilo v večini izvedeno ročno, v manjšem obsegu pa sem si pomagal z dodatnimi skriptami.

Glede tega, da eksplicitno zapisana pravila strukturnega prenosa ter slovarji omogočajo iterativno izboljševanje kakovosti prevodov, sistem, ki je zasnovan na osnovi pravil plitkega prenosa lahko izboljšujemo in izpopolnjujemo z dodatnim ročnim pregledom prevajalnih gradiv.

Kakovost izdelanega prevajalnega sistema za jezikovni par slovenščina - makedonščina je zadovoljiva, če upoštevamo, da sem skoraj vse samodejno razvijal. Možnosti izboljšave seveda obstajajo kljub temu, da prevodi sistema že dosegajo produkcijsko kakovost. Možnost izboljšave je predvsem v dodajanju novih besed v oba enojezična slovarja ter dodajanja novih prevodov v dvojezični slovar. Vrednotenje samega sistema ni izvedeno v popolnosti, zato pripravljam obširnejše vrednotenje sistema in primerjavo z ostalimi prevajalnimi sistemi, ki vsebujejo opisani jezikovni par.

6.2 Nadaljnje delo

Razvoj jezikovnega para slovenščina - makedonščina še zdaleč ni končan. Doseženi rezultati so dokaj zadovoljivi, vendar se ne bom vstavil tukaj, saj sem se odločil, da bom nadaljeval z razvojem jezikovnega para z dodajanjem novih besed v oba enojezična slovarja ter dodajanjem novih prevodov v dvojezični slovar. Poleg dodajanja novih besed je potrebno tudi dodatno delo na označevalniku obliko-skladenjskih oznak v

oba enojezična slovarja, saj so le-te bile delno privzete iz projekta apertium-hbs-slv ter apertium-sh-mk. Glede tega, da imata Slovenščina in Makedonščina zelo velike slovnične razlike, kvaliteta prevoda še vedno ni primerljiva s kvaliteto pravil gramatike z omejitvami, ki skrbijo za prevajanje slovenskega jezika v makedonski jezik in obratno. Zato bom v prihodnosti prevzel pravila gramatike z omejitvami iz slovenskega jezika ter jih preuredil za makedonski jezik in obratno.

Načinov, kako izboljšati trenutni jezikovni par je veliko, zato se z vsemi močmi zavzemam, da bo razvoj še v nadaljnje potekal brez večjih težav in da se ne bom ustavil tukaj.

7 Literatura

Antonio M. Corbi-Bellot, Mikel L. Forcada, in Sergio Ortiz-Rojas. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In Proceedings of the EAMT conference, 79–86. HITEC e.V., May 2005.

Tomaž Erjavec. *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora* in *Proceedings of the 4. Conference on Language Resources and Evaluation, LREC'04, 1535–1538.* ELRA, 2004.

Jan Hajič, Petr Homola, in Vladislav Kubon. A simple multilingual machine translation system. In Eduard Hovy in Elliott Macklovitch, editors, Proceedings of the MT Summit IX, 157–164, New Orleans, USA, 2003. AMTA.

William John Hutchins. *The history of machine translation in a nutshell*, 2005.

William John Hutchins, Harold Somers. *An introduction to machine translation*, London, Academic press, 1992.

Vladimir Levenshtein. *Levenshtein distance. Binary codes capable of correcting deletions, insertions and reversals.* Doklady Akademii Nauk, 845–848, 1965.

Jan D. ten Thije, Ludger Zeevaert. *Linguistic distance. Receptive multilingualism: linguistic analyses, language policies, and didactic concepts.* John Benjamins Publishing Company, 2007.

Jernej Vičič. *Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike.* PhD thesis, Univerza v Ljubljani, 2012.

Apertium. A free/open-source machine translation platform. 2010
URL http://www.apertium.org/?id=whatisapertium&lang=en.

LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, LDC, 2005.

OPUS. Opus - an open source parallel corpus. (Dostopano: 2013)
URL http://opus.lingfil.uu.se/trac

Testvoc. Metode za testiranje posamičnih slovarjev. (Dostopano: 2013)
URL http://wiki.apertium.org/wiki/Testvoc

Priloge

A Pravila prenosa

Poglavje prikazuje pravila prenosa, katera uporablja prevajalni sistem Apertium. Zapisana so v XML obliki in za lažje razumevanje je zapis nekoliko prirejen. Prikazana so pravila prenosa v obe smeri, torej Makedonščina → Slovenščina, ter Slovenščina → Makedonščina.

Na naslednji sliki je prikazano prazno pravilo. Glede na to, da pravilo le prebere samostalnik in ga ponovno izpiše je v glavnini namenjeno samo za prikaz uporabe pravil.

```
<!--prazno pravilo, uporabimo samo za prikaz-->
<rule>
  <pattern>
    <pattern-item n="samostalnik"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.1: Prazno pravilo za samostalnik. Pravilo prebere samostalnik in ga izpiše na svoj izhod, torej ne opravi nobene spremembe.

Na sliki A.2 je prikazano pravilo ujemanja pridevnika v sklonu in številu. Glede na to, da v makedonščini ne obstajajo skloni in je pridevnik vedno v rodilniku to pravilo spreminja sklon pridevnika iz rodilnika v imenovalnik.

Na sliki A.3 je prikazano pravilo ujemanja dveh pridevnikov in samostalnika v sklonu in številu. Podobno kot v pravilu A.2 se oba, pridevnik in samostalnik ujemata edino v številu, med tem ko se sklon iz imenovalnika spreminja v imenovalniku.

Na sliki A.4 je prikazano pravilo ujemanja predpone in pridevnika v sklonu in številu. Lahko opazimo, da je pravilo razdeljeno na dva dela in sicer, ko imamo predpono *naj* in vse ostalo. Vse ostalo je enako kot pri obeh prejšnjih pravilih. Razlika pri predponi *naj* je zaradi tega ker je pridevnik v superlativu v isti obliki kot pri primerjalniku z edino razliko, da moramo dodati predpono *naj*.

```
<rule>
   <pattern>
     <pattern-item n="pridevnik"/>
     <pattern-item n="imenovalnik"/>
   </pattern>
   <action>
     <out>
         <lu>
           <clip pos="1" side="tl" part="lem"/>
           <clip pos="1" side="tl" part="pridevnik"/>
           <clip pos="2" side="tl" part="gen">
           <clip pos="2" side="tl" part="nbr">
         </lu>
         <b/>
         <lu>
           <clip pos="2" side="tl" part="lem"/>
           <clip pos="2" side="tl" part="imenovalnik"/>
           <clip pos="2" side="tl" part="gen"/>
           <clip pos="2" side="tl" part="nbr"/>
         </lu>
  </chunk>
     </out>
   </action>
</rule>
```

Slika A.2: Ujemanje pridevnika v sklonu in številu.

```
<!--ujemanje dveh pridevnikov s samostalnikom v sklonu in
številu-->
<!--primer: убав нов автомобил (ubav nov avtomobil) →
lep nov avto-->
<rule>
   <pattern>
     <pattern-item n="pridevnik"/>
     <pattern-item n="pridevnik"/>
     <pattern-item n="samostalnik"/>
   </pattern>
   <action>
     <out>
         <lu>
           <clip pos="1" side="tl" part="lem"/>
           <clip pos="1" side="tl" part="pridevnik"/>
           <clip pos="3" side="tl" part="gen"/>
           <clip pos="3" side="tl" part="nbr"/>
         </lu>
         <b/>
         <lu>
           <clip pos="2" side="tl" part="lem"/>
           <clip pos="2" side="tl" part="pridevnik"/>
           <clip pos="3" side="tl" part="gen"/>
           <clip pos="3" side="tl" part="nbr"/>
         </lu>
         <b/>
         <lu>
           <clip pos="3" side="tl" part="lem"/>
           <clip pos="3" side="tl" part="samostalnik"/>
           <clip pos="3" side="tl" part="gen"/>
           <clip pos="3" side="tl" part="nbr"/>
         </lu>
     </out>
   </action>
</rule>
```

Slika A.3: Ujemanje dveh pridevnikov in samostalnika v sklonu in številu.

```
<!--ujemanje predpona in pridevnika v sklonu in številu-->
<!--primer:  висок (visok) → visok
             повисок (povisok) → višji
             највисок (najvisok) → najvišji-->
<rule>
   <pattern>
     <pattern-item n="predpona"/>
     <pattern-item n="pridevnik"/>
   </pattern>
   <action>
     <choose>
       <when>
         <test>
           <and>
             <equal>
               <clip pos="1" side="sl" part="lem"/>
               <lit v="наj (naj)"/>
             </equal>
             <equal>
               <clip pos="1" side="sl" part="pridevnik"/>
               <lit-tag v="adj.sup"/>
             </equal>
           </and>
         </test>
         <out>
             <lu>
               <lit v="naj"/>
             </lu>
             <lu>
               <clip pos="2" side="tl" part="lem"/>
               <lit-tag v="adj.comp"/>
               <clip pos="2" side="tl" part="gen"/>
               <clip pos="2" side="tl" part="nbr"/>
             </lu>
         </out>
       </when>
       <otherwise>
         <out>
             <lu>
               <clip pos="2" side="tl" part="lem"/>
               <clip pos="1" side="tl" part="pridevnik"/>
               <clip pos="2" side="tl" part="gen"/>
               <clip pos="2" side="tl" part="nbr"/>
             </lu>
         </out>
       </otherwise>
     </choose>
   </action>
</rule>
```

Slika A.4: Ujemanje predpona in pridevnika v sklonu in številu.

Do sedaj smo si ogledali pravila prenosa v smeri Makedonščina → Slovenščina. Naslednja pravila prenosa so v smeri Slovenščina → Makedonščina.

Na sliki A.5 je prikazano pravilo ujemanja pridevnika v sklonu in številu. Opazimo lahko, da je prikazano pravilo precej enako kot pravilo A.2 ter, da je edina razlika v tem, da deluje v nasprotni smer, torej pridevnik izgublja sklon, natančneje sklon se spreminja v rodilniku.

Na sliki A.6 je prikazano pravilo ujemanja glagolov v sklonu in številu. Podobno kot prejšnjo pravilo, tudi to pravilo izbriše sklone iz glagola, torej ga spremeni v rodilnik.

Kot smo že omenili in sedaj lahko opazimo je struktura pri vseh pravilih precej enaka.

```
<!--ujemanje pridevnika v sklonu in številu-->
<rule>
  <pattern>
    <pattern-item n="pridevnik"/>
    <pattern-item n="imenovalnik"/>
  </pattern>
  <action>
    <let>
      <clip pos="2" side="tl" part="sklon"/>
      <lit-tag v="imenovalnik"/>
    </let>
    <out>
       <lu>
   <clip pos="1" side="tl" part="lemh"/>
   <clip pos="1" side="tl" part="pridevnik"/>
   <clip pos="1" side="tl" part="gen"/>
   <clip pos="1" side="tl" part="nbr"/>
   <lit-tag v="nom.ind"/>
</lu>
<b pos="1"/>
<lu>
   <clip pos="2" side="tl" part="lemh"/>
   <clip pos="2" side="tl" part="imenovalnik"/>
   <clip pos="2" side="tl" part="gen"/>
   <clip pos="2" side="tl" part="nbr"/>
   <clip pos="2" side="tl" part="sklon"/>
   <lit-tag v="ind"/>
</lu>
    </out>
  </action>
</rule>
```

Slika A.5: Ujemanje pridevnika v sklonu in številu.

```
<!--ujemanje glagolov v sklonu in številu-->
<rule>
  <pattern>
    <pattern-item n="imenovalnik"/>
    <pattern-item n="glagol biti"/>
    <pattern-item n="glagol"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="whole"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lemh"/>
        <clip pos="3" side="tl" part="glagol"/>
        <clip pos="3" side="tl" part="temps"/>
        <clip pos="1" side="tl" part="gen"/>
        <clip pos="1" side="tl" part="nbr"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika A.6: Ujemanje glagolov v sklonu in številu.

B Primeri prevodov in težave

B.1 Primeri prevodov

Najprej si bomo ogledali nekaj dobrih prevodov kateri ne potrebujejo dodatnih komentarjev. Glede na to, da je projekt še vedno v fazi razvijanja bodo nekateri prevodi za katere smatramo da so dobri imeli napake. Predstavili bomo prevode v obe smeri. Najprej je pri vsakem primeru zapisano izvorno besedilo, nato sledi prevod.

Naslednji prevodi so v smeri Slovenščina → Makedonščina:

(B.1) Jaz sem Branko.
Јас сум Бранко. (Jas sum Branko.)

(B.2) Njihov avto je drag.
Нивниот автомобил е скап. (Nivniot avtomobil e skap.)

(B.3) Sem kupil nov lep čajnik.
Купив нов убав џајарник. (Kupiv nov ubav čajarnik.)

Naslednji prevodi so v smeri Makedonščina → Slovenščina:

(B.4) Јас сум Марија. *(Jas sum Marija.)*
Jaz sem Marija.

(B.5) Нејзиниот глас е многу убав. *(Nejziniot glas e mnogu ubav.)*
Njen glas je zelo lep.

(B.6) Колку си стар? Имам деветнаесет години. *(Kolku si star? Imam devetnaeset godini.)*
Koliko si star? Imam devetnajst let.

B.2 Težave

Največja težava pri prevajanju besedila je bila ta, da imamo premajhen fond besed. To, da je premajhen fond besed največja težava je dober znak za naš projekt glede na to, da je dodajanja besede najlažji del razvijanja tega projekta.

Naslednja izmed večjih težav je bila in še vedno je, prevajanje pridevnikov v smeri Slovenščina → Makedonščina. Kot smo že omenili v enem od poglavji, se v makedonskem jeziku pridevniki ne stopnjujejo štiri-stopenjsko, zato sem moral samodejno dodajati še eno stopnjo, da bi ustrezal slovenskemu elativu. Po mojem mnenju je problem nastal tukaj, vendar navkljub vsem mojim poskusom nisem prišel do rešitve.

Spodaj je prikazano pravilo strukturnega prenosa, katero skrbi za prevod pridevnikov v smeri Slovenščina → Makedonščina in to pravilo je po mojem mnenju odgovorno za težavo s katero se moram ukvarjati.

```
<rule comment="Pridevniki">
<pattern>
  <Pattern-item n="pridevniki"/>
</pattern>
<action>
 <choose>
   <when>
    <test>
     <and>
      <equal><clip pos="1" side="tl" part="stopnjaSL"/><lit-tag v="add_ela"/></equal>
      <equal><clip pos="1" side="tl" part="stopnjaMK"/><lit-tag v="ssup"/></equal>
     </and>
    </test>
    <out>
     <lu>
      <lit v="velik"/>
      <lit-tag v="adv.sint.ela"/>
     </lu>
     <b/>
     <lu>
      <clip-pos="1" side="tl" part="lema"/>
      <clip-pos="1" side="tl" part="pridevnik"/>
      <lit-tag v="sint"/>
     </lu>
    </out>
   </when>
...
```

Slika B.1: Pravilo strukturnega prenosa, katero ureja pridevnike.

V nasprotni smer, torej Makedonščina → Slovenščina prevajanje pridevnikov ne predstavlja problemov.

To so trenutno največje težave v mojem projektu. Obstajajo tudi nekatere manjše težave, katere pa niso toliko pomembne. Lahko rečem, da sem zadovoljen z tem, kar sem do sedaj ustvaril v svojem projektu glede na to, da je projekt še vedno v fazi razvijanja.