UNIVERZA NA PRIMORSKEM
Fakulteta za matematiko, naravoslovje in informacijske tehnologije

Računalništvo in informatika – 1. stopnja

Aleš Horvat

Izdelava sistema za strojno prevajanje na osnovi pravil plitkega prenosa za jezikovni par slovenščina – hrvaščina

Zaključna projektna naloga

Mentor: doc. dr. Branko Kavšek
Somentor: mag. Jernej Vičič

Koper, 2012

UNIVERSITY OF PRIMORSKA
Faculty of Mathematics, Natural Sciences and Information Technologies

Information Sciences and Technologies – 1^{st} degree

Aleš Horvat

Production of machine translation system based on shallow transfer rules for the slovenian – croatian language pair.

Final Project Paper

Mentor: doc. dr. Branko Kavšek
Co-Mentor: mag. Jernej Vičič

Koper, 2012

Zahvala

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

Zahvaljujem se tudi svoji družini za vsestransko pomoč na moji izobraževalni poti.

Hvala!

Seznam uporabljenih kratic in simbolov

CBMT Corpus Based Machine Translation, strojno prevajanje na osnovi korpusov

EBMT Example Based Machine Translation, strojno prevajanje na osnovi primerov

GSOC Google Summer of Code, globalni program, ki študentom nudi sodelovanje v različnih odprtokodnih projektih

HTML HyperText Markup Language, označevalni jezik

LDC Linguistic Data Consortium, odprti konzorcij članov, ki se ukvarjajo z jezikovnimi tehnologijami

MSD Morphosyntactic description, oblikoskladenjska oznaka

MT Machine Translation, strojno prevajanje

RBMT Rule Based Machine Translation, strojno prevajanje na osnovi pravil

SMT Statistical Machine Translation, statistično strojno prevajanje

WER Word Error Rate, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo napake

WRR Word Recognition Rate, metrika za samodejno preverjanje kakovosti prevodov, ocenjuje stopnjo uspešnih prevodov

XML Extensible Markup Language, označevalni jezik

Povzetek

Zaključna projektna naloga predstavlja postavitev sistema za strojno prevajanje za jezikovni par slovenščina – hrvaščina. Področje strojnega prevajanja je zelo obširno, zato so na začetku dela predstavljeni osnovni pojmi ter splošen pregled nad samim področjem, ki omogočajo bralcu nadaljnje branje. Podrobneje sta opisani glavni skupini strojnega prevajanja in sicer skupina, ki temelji na osnovi pravil ter skupina, ki temelji na osnovi korpusov. V delu se osredotočimo na prevajalne sisteme, ki temeljijo na osnovi pravil, podrobneje na osnovi pravil plitkega prenosa, saj so se le-ti izkazali najboljši za postavitev prevajalnega sistema za strojno prevajanje sorodnih jezikov. Izdelani prevajalni sistem temelji na odprtokodnem ogrodju Apertium, ki sodi v paradigmo sistemov za strojno prevajanje na osnovi pravil plitkega prenosa (shallow transfer RBMT). Predstavljena je arhitektura ogrodja Apertium ter opis posameznih modulov, ki skrbijo za čim boljši prevod iz izvornega jezika v ciljni jezik. Zaključna projektna naloga prikazuje tudi vso dodatno ročno delo ter posebnosti, s katerimi smo se srečali v sklopu projetka GSOC2012. Opisuje najpogostejše napake, ki so bile odpravljene z urejanjem enojezičnih oblikoskladenjskih slovarjev izvornega ter ciljnega jezika, urejanjem dvojezičnega slovarja ter pisanjem pravil strukturnega prenosa. Pri pravilih strukturnega prenosa smo se omejili le na prvi nivo zaradi podobnosti jezikov, sicer smo pustili odprto možnost nadgraditve na drugi ter tretji nivo. V nadaljevanju so predstavljeni načini vrednotenja ter osnovne statistike izdelanih jezikovnih gradiv, skupaj z rezultati vrednotenja.

Ključne besede: strojno prevajanje, strojno prevajanje sorodnih naravnih jezikov, jezikovni par slovenščina – hrvaščina

Abstract

The final project paper presents an overview of a machine translation system for the slovene and croatian language pair. Since the field of the machine translation systems is extensive, some basic concepts and a general review of the area is presented at the beginning, allowing the reader better understanding of the paper. The main groups of machine translation systems are described as well, outlining in more details the rule-based paradigm and describing the corpora-based paradigm. In this paper we focus on machine translation systems based on transfer rules, pointing out the machine translation systems based shallow transfer rules as they have proven best for setting up machine translation systems translating related language pairs. The translation system is based on Apertium's architecture, which belongs to the Shallow Parse and Transfer Rule-Based Machine Translation paradigm. The Apertium's architecture is also outlined, describing individual modules, which take care of the optimisation of the translation from the source language to the target language. The final project paper also describes all the additional manual work that has been done during the period of the GSOC2012 project. It describes the most common errors that have been corrected primarily by editing the monolingual morphological dictionary of the the source and target languages, the bilingual dictionary and by writing structural transfer rules. The structural transfer rules have been limited to the first layer, but the possibility of implementing the second and third layer have been left open.The paper also presents the the evaluation methods, basic statistics of the produced language resources and the final evaluation results.

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

Kazalo

Slike

Tabele

Poglavje 1

Uvod

1.1 Motivacija

Ideja o strojnem prevajanju oziroma korenine strojnega prevajanja segajo vse do leta 1954, ko sta Georgetown University in IBM predstavila „The Georgetown-IBM experiment". Preizkus je temeljil na demostraciji avtomatskega prevajanja več kot 60 povedi iz ruskega jezika v angleški jezik. Sam preizkus je bil zelo uspešen, avtorji pa so trdili, da bo v roku 3 do 5 let področje strojnega prevajanja popolnoma razvito. Z gotovostjo lahko trdim, da so se avtorji s to trditvijo pošteno ušteli, saj se s tem področjem ukvarjamo še dan danes in ne glede na to, koliko časa je minilo od prve implementacije strojnega prevajanja in koliko časa ter truda so velika podjetja, kot sta Microsoft in Google, vložila v razvoj takih sistemov, smo še vedno daleč od sistema za strojno prevajanje, ki bi zadovoljil večino uporabnikov.

Apertium je odprtokodna platforma za razvoj strojnih prevajalnih sistemov, ki temeljijo na pravilih plitkega prenosa. Implementacija novega jezikovnega para je zelo enostavna. Vsi podatki so organizirani v datotekah XML, ki so človeku relativno lahko razumljive. S tako arhitekturo lahko strokovnjaki področja kasneje izboljšajo kakovost samega sistema brez večjih težav, kar predstavlja veliko prednost v samem procesu razvoja sistema.

Glavna motivacija za izdelavo prevajalnega sistema je bilo sodelovanje v mednarodnem projektu, kamor se vsako leto prijavi več kot 5.000 študentov iz vseh držav, vendar izberejo le okoli 1.000 študentov, ki jim je omogočeno delo na odprtokodnih projektih, kot je Apertium. V sklopu projekta Google Summer of Code 2011 – (GSOC2011) (Google, 2012b) sem pridobil veliko izkušenj ter podrobneje spoznal ekipo in to je bila dodatna vzpodbuda za ponovno sodelovanje ter za dodatno možnost razvoja sistema, ki bi bil uporaben širši javnosti. Skupaj z mentorji smo se pogovorili o predhodnih projektih in ugotovili, da trenutno Apertium nima dovolj dobrega sistema (produkcijske kvalitete) za jezikovni par slovenščina – hrvaščina. Pregledali smo gradiva, ki so bila izdelana v sklopu projekta GSOC2011 in prišli do zaključka, da

imamo na voljo veliko zelo kvalitetnega gradiva, ki ga lahko, z malo truda, izboljšamo in uporabimo za jezikovni par slovenščina – hrvaščina. Večina jezikovnih gradiv je bila samodejno zgrajena s pomočjo metod, predstavljenih v delu (Vičič, 2012).

1.2 Pregled vsebine

Zaključna projektna naloga predstavlja postavitev sistema za strojno prevajanje za jezikovni par slovenščina – hrvaščina. Sistem temelji na ogrodju Apertium, ki sodi v paradigmo sistemov za strojno prevajanje na osnovi pravil plitkega prenosa (shallow transfer RBMT), saj se je le-ta izkazala kot najprimernejša za postavitev sistema za strojno prevajanje sorodnih jezikov (Vičič, 2012).

Drugo poglavje prinaša pregled oziroma opis področja MT. Podrobneje predstavlja sisteme RBMT, sisteme RBMT s plitkim prenosom ter arhitekturo prevajalne platforme Apertium, ki je osnovna platforma, na kateri je izdelan prevajalni sistem, predstavljen v tem delu. Poglavje je namenjeno razlagi osnovnih pojmov znanstvenega področja, ki bralcu omogočijo nadaljnje branje.

V tretjem poglavju je opisano delo, ki je bilo opravljeno v enojezičnih oblikoskladenjskih slovarjih izvornega ter ciljnega jezika. Predstavljeno je tudi delo, ki je bilo opravljeno v dvojezičnem slovarju ter nekaj osnovnih pravil strukturnega prenosa. Podrobneje so opisane značilnosti, katerim smo se posebej posvetili.

Četrto poglavje se omejuje na vrednotenje kakovosti prevodov ter vrednotenje kakovosti jezikovnega gradiva in opisuje metode, ki so bile uporabljene za vrednotenje gradiv.

V petem poglavju so predstavljene osnovne statistike izdelanih jezikovnih gradiv ter rezultati vrednotenja, ki so bili doseženi v sklopu projekta GSOC2012 (Google, 2012c).

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

Poglavje 2

Opis področja

Pričujoče poglavje predstavlja splošen pregled nad sistemi za strojno prevajanje. Podrobneje opisuje sisteme, ki delujejo na osnovi pravil strukturnega prenosa ter arhitekturo prevajalne platforme Apertium, na katerem je bil zgrajen jezikovni par. Poglavje je namenjeno predvsem predstavitvi osnovnih pojmov s področja jezikovnih tehnologij.

2.1 Sistemi za strojno prevajanje

Strojno prevajanje, SP (Machine translation – MT), je področje računalniškega jezikoslovja, ki preiskuje uporabo programske opreme oziroma računalniških sistemov za prevajanje besedila ali govora iz enega naravnega jezika v drugi jezik.

Na osnovni ravni, SP preprosto nadomešča besede enega jezika z besedami drugega jezika (word for word translation), vendar je kakovost takega sistema slaba, saj za dober prevod potrebujemo razumevanje oziroma analizo celotnih stavkov in celo večjih sklopov ter smiselno izbiro besed v ciljnem jeziku. V ta namen so nastala različna področja oziroma različni načini izdelave strojnih prevajalnih sistemov. V prihodnjih razdelkih bodo na kratko opisani sistemi za prevajanje na osnovi korpusov (Corpus-Based – CBMT) ter hibridni prevajalni sistemi, podrobneje pa si bomo ogledali skupino prevajalnih sistemov, ki delujejo na osnovi pravil strukturnega prenosa.

2.1.1 Splošno področje strojnega prevajanja

Področje strojnega prevajanja se, po klasifikaciji avtorja (Sanchez-Martnez et al., 2007), deli na dve glavni skupini: prevajanje na osnovi pravil (Rule-Based – RBMT) ter prevajanje na osnovi korpusov (Corpus-Based – CBMT). Mnogi avtorji, npr. (Vičič, 2012) in (Dorr et al., 1999) dodajajo še tretjo skupino:

- Sistemi tipa RBMT – sistemi za strojno prevajanje s pomočjo zbirk pravil. Ti sistemi besedilo najprej oblikoskladenjsko označijo in skladenjsko razčlenijo, nato pa aplicirajo pravila ter generirajo izhodno besedilo. Večina današnjih prevajalnih sistemov sodi v to skupino. Slaba lastnost sistemov RBMT je predvsem velika časovna ter stroškovna zahtevnost postavitve sistema. Primeri sistemov: Apertium (Apertium, 2010), Promt (Promt, 2010).

- Sistemi tipa CBMT – sistemi za strojno prevajanje na osnovi korpusov. Ti sistemi so razdeljeni na dve fazi in sicer na fazo učenja, v kateri se pripravi množico referenčnih prevodov, ki so analizirani in prevedeni v modele prevajalnega sistema po načelih, ki jih le-ti določajo ter na fazo prevajanja, v kateri modeli prevajalnega sistema nastopijo kot osnova za prevode novih, neznanih besed. Sisteme te skupine se deli na dve podskupini:

  - sistemi statističnega strojnega prevajanja (Statistical Machine Translation – SMT (Al-Onaizan et al., 1999)),
  - sistemi strojnega prevajanja na osnovi primerov (Example Based Machine Translation – EBMT (Nagao, 1984)).

  Primeri sistemov: Google Translate (Och, 2006), Moses (Koehn et al., 2007), Egypt toolkit (EGYPT, 2007) in Genpar toolkit (GenPar, 2010).

- Hibridno strojno prevajanje zajema hibridne sisteme za strojno prevajanje. Ti sistemi združujejo prednosti korpusnih metod ter sistemov, ki temeljijo na pravilih prenosa. Pristopi hibridnih sistemov se delijo na dve skupini in sicer:

  - Rules post-processed by statistics: Prevod se najprej zgenerira z uporabo pravil strukturnega prenosa, nato pa se le-tega poskusi popraviti s pomočjo statističnih modelov prevajalnega sistema.
  - Statistics guided by rules: Pravila strukturnega prenosa so uporabljena za pre-procesiranje podatkov z namenom boljšega rezultata pri naknadni obdelavi, kjer se uporabi statistične modele prevajalnega sistema. Pravila so nato ponovno uporabljena za post-procesiranje rezultata, ki ga poda statistični model, saj na tak način dosežemo normalizacijo.

Vsaka od opisanih skupin prevajalnih sistemov ima prednosti in slabosti, vendar se moramo osredotočiti na kakovost prevajalnega sistema ter fleksibilnost in enostavnost pri naknadnem dodajanju vsebine oziroma odpravljanju napak. V tem primeru pridejo do izraza prevajalni sistemi na osnovi pravil, saj vsebujejo kar nekaj prednosti, kot sta natančnost sledljivosti prevajalnih postopkov ter enostavno dopolnjevanje (Forcada, 2006).

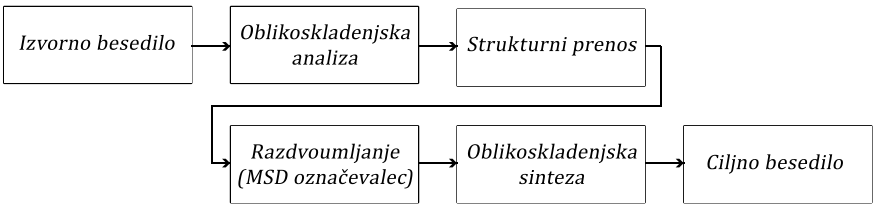
2.1.2 Strojno prevajanje na osnovi pravil

Strojno prevajanje na osnovi pravil (Rule-Based Machine Translation – RBMT) zajema sisteme in metode za prevajanje, ki uporabljajo zbirke pravil strukturnega prenosa. Vsi sistemi RBMT so si v procesu izdelave zelo podobni, saj je le-ta zelo dolgotrajno ter drago opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, kot so: Promt (Promt, 2010), Apertium (Apertium, 2010) in ostali.

Sistemi za prevajanje na osnovi pravil besedilo najprej oblikoskladenjsko označijo in skladenjsko razčlenijo, nato pa izdelajo analizo vhodnega besedila v obliki skladenjskega drevesa izpeljave (Vičič, 2012). Nad analiziranim izvornim besedilom aplicirajo pravila, s katerimi dosežejo abstraktno predstavitev vhodnega besedila. Predstavitev se nato uporabi kot osnovo pri tvorbi izhodnega besedila v ciljnem jeziku.

2.1.3 Strojno prevajanje na osnovi pravil plitkega prenosa

Sistemi strojnega prevajanja na osnovi pravil plitkega prenosa (shallow transfer rule based machine translation) temeljijo na enostavni arhitekturi, pri čemer je analiza tako izvornega kot ciljnega jezika omejena na oblikoskladenjske oznake. Na sliki 2.1 si lahko ogledamo arhitekturo, ki jo uporablja večina sistemov za strojno prevajanje na osnovi plitkega prenosa.



Slika 2.1: Komponente oziroma moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Prikazana arhitektura je bila najprej predstavljena v (Hajič et al., 2000) in pozneje uporabljena v (Corbi-Bellot et al., 2005).

Posamezne komponente oziroma moduli tipičnega sistema za strojno prevajanje na osnovni plitkega prenosa, prikazani na sliki 2.1, so podrobneje opisani v naslednjem poglavju, kjer si bomo tudi podrobneje ogledali ogrodje odprtokodnega prevajalnega sistema Apertium.

2.1.4 Ogrodje za postavitev sistemov za strojno prevajanje – Apertium

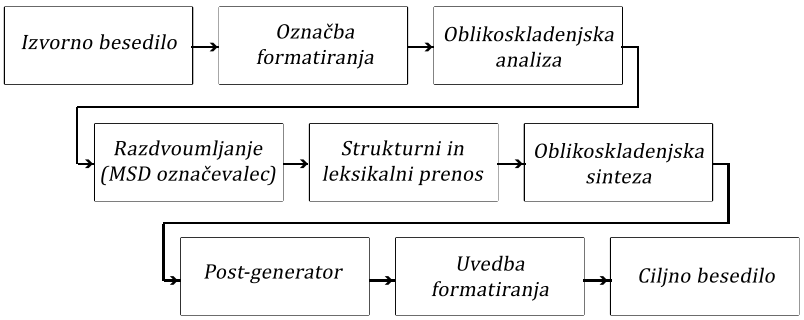
Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev prevajalnega sistema za sorodne jezike, ki temelji na osnovi pravil plitkega

prenosa (shallow transfer) (Sanchez-Martinez in Ney, 2006). Uvršča se med sisteme za samodejno prevajanje naravnih jezikov, razdeljen je na pet osnovnih faz:

- označevanje neprevajalnih razdelkov,
- leksikalni prenos,
- odpravljanje dvoumnosti (disambiguation),
- strukturni prenos,
- prevod posameznih besed in besednih zvez.

S pomočjo niza pravil, se v zadnji fazi odpravijo vse manjše napake oziroma pomanjkljivosti, ki se pojavijo v predhodnih fazah. Pravila strukturnega prenosa temeljijo na regularnih izrazih ter na besednih in fraznih slovarjih.

Arhitektura ogrodja Apertium je zelo podobno arhitekturi, ki je predstavljena na sliki 2.1, katera predstavlja module oziroma komponente tipičnega sistema za strojno prevajanje na osnovi plitkega prenosa. Slika 2.2 prikazuje ogrodje odprtokodnega sistema Apertium.



Slika 2.2: Komponente oziroma moduli odprtokodnega ogrodja Apertium. Za razliko od tipičnega ogrodja prevajalnih sistemov na osnovi plitkih pravil, Apertium dodaja še module, ki služijo označevanju delov besedila, ki se ne prevedejo ter modul za končno urejanje (post-editing) prevodov.

Podrobnejši opis posameznih modulov, ki so prikazani na sliki 2.2.

Označba formatiranja (De-formatter). Modul v izvornem besedilu označi dele besed, ki jih ostali prevajalni moduli ignorirajo. Posledično lahko sistem prevaja tudi besedila, ki vsebujejo urejevalne oznake, npr. oznake jezikov HTML in XML.

Oblikoskladenjska analiza (Morphological analyzer). Modul vsaki besedi izvornega besedila pripiše vse možne oblikoskladenjske oznake, ki bi jih ta besedna oblika lahko imela. Posledica je seveda dvoumen izhod, saj vsaka beseda nosi informacije o vseh možnih oblikoskladenjskih analizah besede. Modul je omejen na besede in besedne zveze, ki so zapisane v enojezičnem slovarju izvornega jezika, ki vsebuje oblikoskladenjske oznake. Pomembno je omeniti, da se analiza posameznih besed izvede samostojno, brez vpliva okolice.

Razdvoumljanje z označevanjem MSD (Morphosyntactic descriptions – oblikoskladenjski označevalec). Modul izbere najverjetnejšo oznako izmed vseh možnih oblikoskladenjskih oznak, ki jih določi oblikoskladenjska analiza za vhodno besedno vrsto. Odloči se glede na njeno okolico. Modul temelji na označevalniku oblikoskladenjskih oznak ali pa na omejenih slovnicah (constraint grammar) (Sánchez-Martínez et al., 2008).

Strukturni in leksikalni prenos (Structural transfer). Modul, s pomočjo pravil strukturnega prenosa in dobesednih prevodov, ki se nahajajo v dvojezičnem prevajalnem slovarju, prenese označeno besedilo v ciljni jezik. Pri tem uporablja pravila za leksikalni in plitki strukturni prenos, sami prevodi v obliki lem (slovarskih gesel) pa temeljijo na osnovi dvojezičnega slovarja.

Oblikoskladenjska sinteza (Morphological analyzer). Modul nadomesti oblikoskladenjsko označeno besedilo z dejanskimi besednimi oblikami v ciljnem jeziku. Pri tem uporablja enojezični slovar ciljnega jezika, ki vsebuje oblikoskladenjske oznake.

Končno urejanje (Post-generator). Modul služi odpravljanju manjših napak oziroma pomanjkljivosti, ki se pojavijo v predhodnih fazah. Uvaja tudi posebnosti ciljnega jezika ter združuje besedne zveze.

Uvedba formatiranja (Re-formatter). Modul vstavi v besedilo odseke, ki jih je začetni modul izbral in označil kot besedilo, ki se ne prevaja. Največkrat so to besedila, ki vsebujejo urejevalne oznake, npr. oznake jezikov HTML in XML.

Podrobnejši opis odprtokodnega ogrodja za postavitev prevajalnega sistema za sorodne jezike – Apertium, je dosegljiv na strežniku Apertium (Sourceforge)[1].

[1]Apertium (Sourceforge): `http://sourceforge.net/projects/apertium`

2.2 Jezikovni par slovenščina – hrvaščina

2.2.1 Zgodovina

Republika Slovenija ima približno 2.050.000 prebivalcev, nahaja se na skrajnem jugu Srednje Evrope. Na zahodu meji na Italijo, na severu na Avstrijo, na severovzhodu na Madžarsko, na vzhodu in jugu pa na Hrvaško.

Večji del 20. stoletja je bila Slovenija del Jugoslavije, države na zahodnem delu Balkana, ki je bila ustanovljena z združitvijo Slovencev, Hrvatov, Srbov ter kraljevine Srbije. Slovenski, srbski, hrvaški ter bosanski jezik spadajo v skupino južnoslovanskih jezikov, ki so bili večinoma uporabljeni v nekdanji Jugoslaviji. Našteti jeziki imajo skupne korenine in skupno zgodovinsko okolje, saj so bili govorjeni in učeni v isti državi. Žal imajo današnje mlajše generacije težave pri medsebojni komunikaciji, zato je tudi velik interes za izdelavo takega prevajalnega sistema. Vsi ti jeziki so si zelo podobni, vendar se močno razlikujejo od ostalih jezikov, kot so angleščina, arabščina, španščina, kitajščina in francoščina, zato je pri uporabi obstoječih prevajalnih sistemov, nujno potrebna tudi človeška revizija za odpravo nastalih napak.

Po statistikah sodeč (pridobljenih na Wikipediji (Wikipedia, 2012a)), veliko število slovenskega prebivalstva (okoli 113.000 oseb) govori srbski, hrvaški, bosanski ali črnogorski jezik kot svoj materni jezik. Po vsem svetu pa vse te jezike govori več kot 17 milijonov ljudi, kar predstavlja zelo veliko število potencialnih uporabnikov prevajalnega sistema.

2.2.2 Pomembnost podobnosti jezikov

Pri izdelavi sistemov za strojno prevajanje sorodnih jezikov je zelo pomembno, koliko se izbrana jezika medseboj pokrivata oziroma koliko sta si podobna. Delo (Vičič, 2012) opisuje kategorije oziroma nivoje, na katere je smiselno razdeliti podobnosti jezikov. Nivoje ločimo na sledeče podobnosti: tipološko, oblikoslovno, skladenjsko in leksikalno. V tem delu ne bomo posvečali večje pozornosti omenjenim kategorijam.

Poglavje 3

Metodologija

Za postavitev delujočega prevajalnega sistema moramo izdelati vsa jezikovna gradiva, kot so predstavljena v razdelku 2.1.4. V nadaljevanju so predstavljena samodejno zgrajena gradiva, ki so bila izdelana z metodami, predstavljenimi v delu (Vičič, 2012) ter obstoječa gradiva, ki so zahtevala predelavo oziroma izdelavo novih pravil za uporabo v novem sistemu. Gradiva so dostopna na strežniku Apertium (Sourceforge)[1]. Razlog za dodatno delo je predvsem v razliki v oblikoskladenjskem označevanju gradiv v obeh jezikih, ki ga je potrebno poenotiti.

Pričujoče poglavje predstavlja vse značilnosti, katerim smo se posebej posvetili v posameznih komponentah prevajalnega sistema.

3.1 Enojezični slovar izvornega jezika – slovenščine

Nabor oblikoskladenjskih oznak obeh jezikov se je pokrival v večini primerov, tako je bil potreben le kratek pregled in vnos novega gradiva, ki ni bil prisoten v slovenskem enojezičnem slovarju. V veliki večini je to bilo dodajanje novih oblikoskladenjskih oznak ter besednih oblik, nekaj pa tudi dodajanja novih besed. V nadaljevanju je podrobneje predstavljeno delo, ki je bilo opravljeno na posameznih besednih vrstah oziroma skupinah besed.

3.1.1 Samostalniške besede

Samostalniške besede se delijo na dve glavni skupini in sicer na obča in lastna imena. Obča imena niso zahtevala dodatnega dela, lastna imena pa so zahtevala veliko bolj pregledno kategorizacijo. Razdelili smo jih v tri kategorije in sicer na imena, priimke ter imena krajev. Imenom in priimkom smo dodali tudi oblikoskladenjske oznake ter oblike za svojilno pridevniško obliko, saj se je izkazalo, da trenutna različica slovenskega enojezičnega slovarja ni vsebovala teh oblik, ki pa se velikokrat pojavijo v besedilih.

[1] Apertium (Sourceforge): `http://sourceforge.net/projects/apertium`

Dodajanje svojilne pridevniške oblike smo morali izvesti ročno, saj imajo različne skupine imen oziroma priimkov, različne svojilne oblike. Problem smo rešili z uvedbo nove paradigme, ki vsebuje privzete svojilne končnice, za vsako skupino imen oziroma priimkov pa smo morali ročno dodati vmesni člen oziroma končnico, ki pripada imenu oziroma priimku v svojilni obliki, s sledečimi lastnostmi: osnovnik, moški spol, ednina, imenovalnik. Slika 3.1 prikazuje primer svojilnih pridevniških oblik za imeni Mark in Špela.

Mark		***Špela***	
Mark-ov	*(I)*	*Špel-in*	*(I)*
Mark-ov-ega	*(D)*	*Špel-in-ega*	*(D)*
Mark-ov-emu	*(R)*	*Špel-in-emu*	*(R)*
Mark-ov-ega	*(T-Ž)*	*Špel-in-ega*	*(T-Ž)*
Mark-ov	*(T-N)*	*Špel-in*	*(T-N)*
Mark-ov-em	*(M)*	*Špel-in-em*	*(M)*
Mark-ov-im	*(O)*	*Špel-in-im*	*(O)*

Slika 3.1: Primer svojilnih pridevniških oblik za imena oziroma priimke v ednini moškega spola. Prikazani sta dve tožilni obliki, ki sta odvisni od lastnosti živosti, ki jo nosi naslednja beseda – samostalnik.

Iz primera na sliki 3.1 vidimo strukturo svojilne pridevniške oblike. Imena oziroma priimki so razdeljeni na tri člene in sicer:

- krn imena oziroma priimka → **Mark** oziroma **Špel**,
- končnica svojilne prid. oblike (osn., m., ed., im.) → **ov** oziroma **in**,
- privzete svojilne prid. končnice, ki veljajo za vsa imena in priimke.

```
/      ->    <moški_živ><ednina><imenovalnik>
- ega  ->    <moški_živ><ednina><dajalnik>
- emu  ->    <moški_živ><ednina><rodilnik>
- ega  ->    <moški_živ><ednina><tožilnik>
/      ->    <moški_neživ><ednina><tožilnik>
- em   ->    <moški_živ><ednina><mestnik>
- im   ->    <moški_živ><ednina><orodnik>
```

Slika 3.2: Paradigma s privzetimi svojilnimi pridevniškimi končnicami. Tožilnik, ki označuje neživost ima posebno oblikoskladenjsko oznako.

Paradigma na sliki 3.2 prikazuje elemente moškega spola v ednini. Opazimo lahko posebnost in sicer dve različni tožilni obliki. Ti dve obliki sta posledici lastnosti živosti ali neživosti, ki jo nosi naslednja beseda – samostalnik, ki se nahaja za lastnim imenom oziroma priimkom. Izjema velja samo za ednino moškega spola. Primer si lahko ogledamo na sliki 3.3.

*Pogledal sem **Markov avtomobil**. (Neživ)*
*Pogledal sem **Markovega kameleona**. (Živ)*

Slika 3.3: Primer tožilne oblike z živim oziroma neživim samostalnikom. Svojilna pridevniška oblika lastnega imena oziroma priimka je odvisna od lastnosti živosti, ki jo nosi samostalnik.

3.1.2 Pridevniki in prislovi

Slovenska slovnica (Toporišič, 2000) navaja, da pridevnike in prislove stopnjujemo štiri-stopenjsko in sicer kot osnovnik, primernik, presežnik ter elativ. Enako pravilo velja v hrvaški slovnici. V ta namen smo izdelali različne paradigme, ki pokrivajo vse štiri osnovne oblike. Primer štiri-stopenjskega prislova prikazuje slika 3.4.

Osnovnik:	*poceni*
Primernik:	*ceneje, cenejše*
Presežnik:	*najceneje, najcenejše*
Elativ:	*prepoceni*

Slika 3.4: Primer vseh štirih stopenj za prislov poceni. Različni prislovi oziroma pridevniki imajo lahko samo eno stopnjo, lahko pa tudi različne kombinacije vseh štirih stopenj.

Paziti smo morali tudi na besede, za katere obstaja samo osnovnik oziroma za katere obstajajo različne kombinacije vseh štirih oblik. Naštete lastnosti veljajo tako za prislove kot za pridevnike. Za lažje generiranje pridevniških oblik smo paradigme osnovnih oblik vezali na sekundarne paradigme, ki vsebujejo še podatke, kot so spol, število, sklon ter določnost. Za razliko od pridevnikov, prislovov ne označujemo z dodatnimi podatki, ampak samo s stopnjo. Ugotovili smo, da trenutne oblikoskladenjske oznake označujejo vse potrebne informacije za jezikovni par, zato je bilo potrebno le dodati manjkajoče besede.

3.1.3 Glagoli

S pomočjo metod za hitro postavitev prevajalnih sistemov (Vičič, 2012), je generirano gradivo že vsebovalo glagolske paradigme, ki so vsebovale oblike za glagole v nedoločniku, namenilniku ter povedniku. Poleg naštetih oblik smo potrebovali še velelnik, deležnik na –n/–t ter deležnik na –l. Za lažje generiranje manjkajočih oblik smo ustvarili štiri privzete paradigme in jih s spremembo vmesnega člena dodali v vse glagolske paradigme. Na sliki 3.5 si

lahko ogledamo nekaj primerov vseh glagolskih oblik glagola *iskati*.

Nedoločnik:	*iskati*
Namenilnik:	*iskat*
Povednik (sed):	*iščem, iščeš, išče, ...*
Velelnik (sed):	*išči, iščiva, iščita, iščimo, iščite*
Deležnik na –n/–t:	*iskan, iskana, iskano, ...*
Deležnik na –l:	*iskal, iskala, iskalo, ...*

Slika 3.5: Glagol iskati v vseh glagoslkih oblikah, ki so bile potrebne za vzpostavitev prevajalnega sistema.

Poleg manjkajočih glagolskih oblik je bilo potrebno določiti tudi glagolski vid, ki označuje ali je glagol dovršen, nedovršen ali oboje. Izkazalo se je, da za jezikovni par slovenščina – hrvaščina potrebujemo tudi podatke o glagolski prehodnosti. V ta namen smo podvojili vse glagolske paradigme ter jih označili z obema oblikoskladenjskima oznakama za označevanje glagolske prehodnosti. Potrebovali smo obe oznaki, ker je glagol lahko prehoden, neprehoden ali kar oboje hkrati. Dodali smo tudi oblikoskladenjske oznake ter oblike povratnih glagolov. Na sliki 3.6 si lahko ogledamo primere za glagolski vid ter glagolsko prehodnost.

Glagolski vid
Dovršen: *Mark je pristopil k izpitu.*
Nedovršen: *Špela uči v osnovni šoli.*

Glagolska prehodnost
Prehodni: *Mark izpolnjuje anketo.*
Neprehodni: *Špela spi.*

Slika 3.6: Glagoslki vid in glagolska prehodnost prikazana na konkretnih primerih.

3.1.4 Števniki

Števnike je bilo potrebno kategorizirati v tri skupine, saj so lahko glavni, drugi ali vrstilni. Dodani so bili manjkajoči števniki od 1 do 100 v vseh treh oblikah. Vsakemu številniku seveda pripada tudi paradigma, ki določa število, spol in sklon.

3.1.5 Zaimki

Pri zaimkih ni bilo večjih posebnosti. Pregledati je bilo le potrebno, če so pravilno kategorizirani, saj so lahko: osebni, svojilni, oziralnostni, vprašalni, kazalni, celostni, povratni ali nedoločnostni.

3.1.6 Predlogi

Tudi pri predlogih ni bilo večjih posebnosti. Vsak predlog je bilo potrebno označiti z oblikoskladenjsko oznako za sklon, v katerem se lahko uporabi. Označevanje sklona je bilo potrebno zaradi enojezičnega slovarja ciljnega jezika, saj so v le-tem imeli predlogi omenjeno oblikoskladenjsko oznako.

3.1.7 Ostalo

V enojezičnem slovarju izvornega jezika imamo še veznike, členke, medmete ter kratice. Naštete besedne vrste ne potrebujejo dodatnih oznak, ampak samo oblikoskladenjsko oznako, ki označuje besedno vrsto.

3.2 Enojezični slovar ciljnega jezika – hrvaščine

Enojezični slovar ciljnega jezika, ki je nastal v okviru projekta GSOC 2011, je bil prevzet iz jezikovnih gradiv prevajalnega sistema srbščina in hrvaščina – makedonščina[2]. Potrebno je bilo le dodati prevode slovenskih besed, ki niso bile prisotne v tem slovarju. V veliki večini je zadostoval našim potrebam. Gradiva so dostopna na strežniku Apertium (Sourceforge)[3].

3.3 Dvojezični prevajalni slovar

Dvojezični prevajalni slovar vsebuje besede enojezičnih slovarjev ter ustrezne prevode le-teh, z vsemi ustreznimi oblikoskladenjskimi oznakami. Zgradili smo ga samodejno, z uporabo sistema Google Translate (Google, 2012a), napake in manjkajoči prevodi so bili ročno popravljeni. Ugotovili smo, da ima Google Translate zelo velike težave s prevajanjem pridevnikov, prislovov in predlogov, za razliko od teh treh besednih vrst pa zelo dobro prevaja samostalnike in glagole.

Pri sami izdelavi dvojezičnega prevajalnega slovarja smo naleteli na veliko težav, ki so v nadaljevanju podrobneje opisane.

3.3.1 Poenotenje oblikoskladenjskih oznak izvornega ter ciljnega enojezičnega slovarja

Za pravilno prevajanje moramo poskrbeti, da se oblikoskladenjske oznake izvornega ter ciljnega slovarja pokrivajo, tudi vrstni red je v tem primeru zelo pomemben. Vrstni red po navadi rešujemo s pravili strukturnega prenosa, same spremembe oziroma razlike oblikoskladenjskih oznak med izvornim in ciljnim jezikom pa moramo rešiti v dvojezičnem slovarju.

[2]Projekt Apertium (sh-mk): `http://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-sh-mk//`

[3]Apertium (Sourceforge): `http://sourceforge.net/projects/apertium`

V dvojezičnem slovarju lahko omenjene spremembe oziroma razlike rešimo na dva načina. Imamo primer, ko moramo narediti $1:1$ spremembo in primer, ko moramo narediti $1:n$ ali $n:1$ spremembo. Oba tipa sprememb lahko rešimo na zelo enostaven način in sicer z zamenjavo oblikoskladenjske oznake izvornega ter ciljnega jezika v vnosu, vendar taka rešitev lahko hitro nasiči število vnosov v dvojezičnem slovarju. Primer zamenjave oblikoskladenjske oznake $(1:1)$:

- **stol** <samostalnik><**moški**> ⇒ **miza** <samostalnik><**ženski**>
- **godina** <samostalnik><**ženski**> ⇒ **leto** <samostalnik><**srednji**>

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

- **gospoda** <sam><m><**mn**> ⇒ **gospoda** <sam><m><**ed**>
- **gospoda** <sam><m><**mn**> ⇒ **gospoda** <sam><m><**dv**>
- **gospoda** <sam><m><**mn**> ⇒ **gospoda** <sam><m><**mn**>

V takem primeru smo primorani dodati tri različne vnose, s katerimi nakazujemo, da se vse oblike iz slovenskega jezika, torej ednina, dvojina in množina, prevedejo v množino v hrvaškem jeziku. Da se izognemo neprimerni rešitvi, lahko dodamo paradigmo, ki skrbi za prenos $1:n$ ali $n:1$ določene oblikoskladenjske oznake. Primer paradigme si lahko ogledamo na sliki 3.7.

Paradigma = "pl_sgpl"

<množina>	*<->*	*<množina>*
RL: <množina>	*<->*	*<dvojina>*
RL: <množina>	*<->*	*<ednina>*

Slika 3.7: Paradigma za odpravljanje nepotrebnih vnosov v dvojezičnem slovarju. Oznaka RL označuje prevajalno smer, v kateri se upošteva zapisana izjema.

Iz slike 3.7 lahko razberemo, da se paradigma imenuje *pl_ sgpl*. Potrebno je dodati tudi oznako, ki označuje, v kateri smeri so dovoljene izjeme. V našem primeru imemo oznako *RL*, le-ta označuje, da se izjema upošteva pri prevajanju iz slovenskega jezika v hrvaški jezik. Sedaj, ko imamo pripravljeno paradigmo, sam vnos v dvojezičnem slovarju izgleda tako:

- **gospoda** <sam><m> ⇒ **gospoda** <sam><m><par n="pl_sgpl"/>

Paradigma nam torej omogoči, da se izognemo podvojevanju vnosov v dvojezičnem slovarju. Na tak način je vnašanje novih vnosov ter odpravljanje napak, veliko lažje in preglednejše. Reševanje vrstnega reda oblikoskladenjskih oznak je predstavljeno v razdelku 3.5.

3.3.2 Stopnjevanje pridevnikov ter prislovov

Prva težava, na katero smo naleteli, je bila razlika v oblikoskladenjskem označevanju stopenj. Problem smo rešili z uporabo dodatne paradigme, ki določa preslikavo oblikoskladenjskih oznak:

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

Naslednja težava je nastala pri prevajanju besed, ki niso imele enakega števila oziroma istih stopenj v izvornem ter ciljnem jeziku. Kot je bilo omenjeno v razdelku 3.1.2, v obeh jezikih prislove in pridevnike stopnjujemo štiri-stopenjsko in sicer kot osnovnik, primernik, presežnik in elativ. Težavo smo rešili na tak način, da smo pred prislove in pridevnike dodali besedo bolj, najbolj ali preveč, odvisno od manjkajoče oblike. Primer prevajanja iz hrvaškega jezika v slovenski jezik si lahko ogledamo na sliki 3.8.

Osnovnik: *bijelo -> belo*
Primernik: *bjelije -> bolj belo*
Presežnik: *najbjelije -> najbolj belo*
Elativ: *prebijelo-> preveč belo*

Slika 3.8: Primer prevajanja prislova iz hrvaškega jezika v slovenski jezik z manjkajočimi oblikami. Manjkajočim oblikam dodamo besedo bolj, najbolj ali preveč.

Do omenjene rešitve smo prišli tako, da smo v dvojezičnem slovarju dodali vnos za vsako stopnjo posebej, pri manjkajočih stopnjah pa še oblikoskladenjsko oznako, po kateri pravilo strukturnega prenosa zazna, da mora dodati besedo bolj, najbolj ali preveč. Na sliki 3.9 si lahko ogledamo primer vnosov za prevod $bijel \rightarrow bel$, opazimo lahko tudi dodatne oblikoskladenjske oznake *add_comp, add_sup* ter *add_ela*, ki so potrebne za pravilno prevajanje manjkajočih stopenj ter dodatno paradigmo z imenom *only_pst2null*, ki poskrbi, da se oblikoskladenjske oznake v osnovniku pravilno preslikajo.

```
bijel <pridevnik>          <->  bel <pridevnik><par n="only_pst2null>
bijel <pridevnik><comp>    <->  bel <pridevnik><add_comp>
bijel <pridevnik><sup>     <->  bel <pridevnik><add_sup>
bijel <pridevnik><ssup>    <->  bel <pridevnik><add_ela>
```

Slika 3.9: Dvojezični vnosi za manjkajoče stopnje z dodatnimi oblikoskladenjskimi oznakami.

3.3.3 Glagolski prislovi

V enojezičnem slovarju ciljnega jezika – hrvaščine, so bili prisotni tudi glagolski prislovi, ki se v slovenskem jeziku prevedejo v načinovne prislove s končnicami -oč/-eč/-e/-aje. Težave smo imeli z glagolskimi prislovi, ki nimajo ustreznega prevoda v slovenskem jeziku, zato je bilo potrebno le-te prevesti v pridevnike z naslednjimi lastnostmi: moški spol, ednina, imenovalnik.

- Glagolski prislovi s primernim prevodom
  - viseći ⇒ viseč,
  - čekajući ⇒ čakajoč,
  - djelujući ⇒ delujoč.
- Glagolski prislovi s prevodom v pridevnik (m., ed., im.)
  - porazeći ⇒ poražen,
  - poštujući ⇒ spoštovan,
  - kupujući ⇒ kupljen.

3.4 Označevalnik oblikoskladenjskih oznak

V modulu za razdvoumljanje smo uporabili dve tehniki. Pri prevajanju v smeri hrvaščina – slovenščina so bila uporabljena pravila slovnice z omejitvami (constraint grammar), ki so bila izdelana v projektu Apertium (sh-mk)[4], v okviru projekta GSOC2011 (Google, 2012b). Za prevajalno smer slovenščina – hrvaščina smo izbrali tehniko, ki kot osnovo uporablja oblikoskladenjski označevalnik, ki smo ga povzeli iz sistema (Horvat in Vičič, 2012). Izdelan je bil s pomočjo nenadzorovane metode (Sánchez-Martínez et al., 2008), za učenje smo uporabili del dvojezičnega korpusa OPUS (Tiedemann, 2009). Uporabili smo povedi dela prevodov namizja KDE ter dela podnapisov, v skupni dolžini 50 milijonov besed.

3.5 Pravila strukturnega prenosa

Trenutno poglavje je namenjeno izjemoma pravilom strukturnega prenosa. Pravila prenosa so zaradi možnosti večje fleksibilnosti pri zaznavanju besed ali stavkov, razdeljena v tri nivoje. Omejili smo se le na prvi nivo, saj je struktura obeh jezikov jezikovnega para zelo podobna. Za lažje razumevanje bodo predstavljena tako osnovna, kot tudi zahtevnejša pravila, s

[4]Projekt Apertium (sh-mk): `http://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-sh-mk//`

konkretnimi primeri. Opomba: pravila so napisana za prevajanje iz hrvaškega jezika v slovenski jezik, torej je v opisanih primerih hrvaščina izvorni jezik, slovenščina pa ciljni jezik. Struktura pravila je razdeljena na dva dela in sicer na *vzorec*, ki določa eno ali več vhodnih besednih vrst ter *akcijo*, kjer se zapisane spremembe izvedejo.

3.5.1 Urejanje vrstnega reda oblikoskladenjskih oznak izvornega ter ciljnega enojezičnega slovarja s pomočjo makrov

Za dosego pravilnega prevoda, je pri prevodih potrebno poskrbeti za pravilen vrstni red obklikoskladenjskih oznak izvornega ter ciljnega jezika, kot je predstavljeno v razdelku 3.3.1. Vrstni red oblikoskladenjskih oznak preuredimo z uvedbo makra. Na sliki 3.10 si lahko ogledamo primer makra, ki ureja vrstni red oblikoskladenjskih 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 3.10: Primer makra, ki ureja vrstni red oblikoskladenjskih oznak samostalniškim besedam.

Izhod ter vrstni red oblikoskladenjskih oznak:

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

Izhod na konkretnem primeru:

- Avtomobil <samostalnik><moški><ednina><imenovalnik>

3.5.2 Stopnjevanje pridevnikov ter prislovov z manjkajočimi oblikami

V razdelku 3.3.2 smo opisali pristop, ki je potreben za reševanje manjkajočih stopenj pridevnikov in prislovov. Rešitev smo opisali v dvojezičnem slovarju, sedaj pa se bomo poglobili v pravilo, ki poskrbi, da se le-ta tudi pravilno izvede.

V dvojezičnem slovarju smo dodali dodatne oblikoskladenjske oznake *add_comp*, *add_sup* ter *add_ela*, ki nas v pravilu opozorijo, da moramo dodati dodatne besede. Primer pravila si lahko ogledamo na sliki 3.11.

```
<rule comment="Prislovi">
<pattern>
   <pattern-item n="prislovi"/>
</pattern>
<action>
<choose>
   <when>
    <test>
     <and>
      <equal><clip pos="1" side="sl" part="stopnjaHR"/><lit-tag v="ssup"/></equal>
      <equal><clip pos="1" side="tl" part="stopnjaSL"/><lit-tag v="add_ela"/></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 3.11: Pravilo strukturnega prenosa, ki ureja prislove. V primeru, da sta pogoja izpolnjena, se na izhod doda besedo preveč.

Pravilo sprejme en vzorec in sicer prislove. Z značko <equal> preverimo, če je oblikoskladenjska oznaka izvornega jezika enaka *ssup*, nato pa preverimo, če je oblikoskladenjska oznaka ciljnega jezika enaka *add_ela*. V primeru, da sta pogoja izpolnjena, oddamo na izhod dve besedi in sicer:

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

Za primer, predstavljen v razdelku 3.3.2, bi se rezultat glasil: *Preveč bel.*

3.5.3 Primeri pravil

Do sedaj smo si ogledali makro, s katerim smo urejali vrstni red oblikoskladenjskih oznak in pravilo, s katerim smo dodajali besede bolj, najbolj ali preveč manjkajočim stopnjam pridevniških ter prislovnih besed, sedaj pa si bomo pogledali dve specifični pravili.

Primer prvega pravila strukturnega prenosa si lahko ogledamo na sliki 3.12.

```
<rule comment="Glagol biti + Glagol delNaL">
<pattern>
   <pattern-item n="glagol_biti"/>
   <pattern-item n="glagol_delNaL"/>
</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 3.12: Pravilo strukturnega prenosa, ki ureja vzorca biti in poljubni glagol (deležnik na –l).

Pravilo na vhodu prejme dva vzorca, glagol biti ter poljubni glagol (deležnik na –l). Na izhodu, v drugi leksikalni enoti, oviti z značko <lu>, lahko opazimo, da smo opustili oznako, ki določa obliko glagola. Obliko smo namenoma opustili, saj v slovenskem enojezičnem slovarju nismo dodatno označili deležnika na –l. Oglejmo si vpliv pravila na primeru:

- je igrao
  - biti <glagol><sed><3.os><ed>
  - igrati <glagol><nedov><nepreh><delNaL><m><ed>
- je igral
  - biti <glagol><sed><3.os><ed>
  - igrati <glagol><nedov><nepreh><m><ed>

Drugi primer pravila strukturnega prenosa si lahko ogledamo na sliki 3.13.

```
<rule comment="Glagol + pomožni glagol Hoteti">
<pattern>
   <pattern-item n="glagol"/>
   <pattern-item n="pomožni_glagol_hoteti"/> <!-- <vbmod><clt><futI> -->
</pattern>
<action>
   <out>
     <lu>
      <clip pos="1" side="tl" part="lema"/>
      <clip pos="1" side="tl" part="glagol"/>
      <clip pos="1" side="tl" part="glagolska_dovršnost"/>
      <clip pos="1" side="tl" part="glagolska_prehodnost"/>
      <clip pos="1" side="tl" part="oblika"/>
      <clip pos="1" side="tl" part="spol"/>
      <clip pos="1" side="tl" part="število"/>
     </lu>
     <b/>
     <lu>
      <clip pos="2" side="tl" part="lema"/>
      <clip pos="2" side="tl" part="glagol_biti"/>
      <clip pos="2" side="tl" part="oblika"/>
      <clip pos="2" side="tl" part="oseba"/>
      <clip pos="2" side="tl" part="število"/>
     </lu>
   </out>
</action>
</<rule>
```

Slika 3.13: Pravilo strukturnega prenosa, ki ureja vzorca glagol in pomožni glagol hoteti z lastnostima: $naslonka, prihodnjik$ ($clitic$, $future$).

Pravilo prejme dva vzorca na vhodu, poljubni glagol in pomožni glagol hoteti z lastnostima: *naslonka, prihodnjik*. Omenjeni lastnosti sta zelo pomembni, saj se lema *hoteti* z le-temi prevede v pomožni glagol *biti* z lastnostjo *prihodnjik*, v nasprotnem primeru pa se prevede v lemo *hoteti*. Oglejmo si primer:

- Kupiti ću
  - kupiti <glagol><dov><preh><namenilnik>
  - htjeti <p.glagol><naslonka><prihodnjik><1.os><ed>
- Kupil bom
  - kupiti <glagol><dov><preh><m><ed>
  - biti <p.glagol><prihodnjik><1.os><ed>

Iz primera je razvidno, da se pomožni glagol *htjeti* prenese v pomožni glagol *biti* zaradi opisanih lastnosti *naslonka*, *prihodnjik*. Opazimo lahko tudi prenos omenjenih oblikoskladenjskih značk <naslonka><prihodnjik> v slovensko oblikoskladenjsko značko, ki označuje prihodnost, <prihodnjik>. Ostale oblikoskladenjske oznake se preslikajo brez dodatnih sprememb. Oglejmo si še primer, ko pravilo ni upoštevano:

- Ja hoću automobil.
  - jaz <zaimek><osebni><1.os><mžs><ed><imenovalnik>
  - htjeti <p.glagol><sedanjik><1.os><ed>
  - automobil <samostalnik><m><ed><imenovalnik>
- Jaz hočem avtomobil.
  - jaz <zaimek><osebni><1.os><mžs><ed><imenovalnik>
  - hoteti <p.glagol><nedov><preh><sedanjik><1.os><ed>
  - avtomobil <samostalnik><m><ed><imenovalnik>

Poglavje 4

Vrednotenje

Poglavje predstavlja metode vrednotenja sistemov za strojno prevajanje. Vrednotenje (evalvacija) je zelo pomembno, saj nam poda mero kakovosti jezikovnih parov v praksi. Težave nastopijo pri izbiri metod vrednotenja, saj različne metode podajo različne rezultate (Vičič, 2012). Evalvacija je zelo subjektivna ter kompleksna, zato univerzalna metoda ocenjevanja še ni določena. Strokovnjaki predlagajo različne kriterije ocenjevanja:

- Hutchins in Somers (Hutchins in Somers, 1992) navajata tri kriterije:
  - informativnost: v kolikšni meri prevod posreduje enake informacije kot izvirnik,
  - razumljivost: ali je prevod jasen,
  - ustreznost jezika: ali je v prevodu uporabljen jezik, primeren vsebini in sporočilu.
- Konzorcij (LDC, 2005) priporoča dva kriterija z izdelanima lestvicama:
  - vsebinska ustreznost prevodov,
  - slovnična pravilnost prevodov v ciljnem jeziku.
- Statistični pristopi; vse samodejne metode izvirajo iz te skupine in omogočajo oprijemljivejše ocene. Za vse metode te skupine je skupno, da primerjajo število napak različnih vrst. Večina teh metod ni primerna za sisteme RBMT (Callison-Burch et al., 2006).

Napake so določene kot razlike med referenčnim prevodom in prevodom prevajalnega sistema. Zaradi časovne stiske smo se odločili za izbiro le ene metode, ki omogoča samodejno vrednotenje prevodov, ki pa so bili ročno popravljeni, post-editing. Metoda uporablja metriko WER. Rezultati, ki smo jih dosegli, so predstavljeni v naslednjem poglavju.

4.1 Vrednotenje kakovosti prevodov

Vrednotenje kakovosti prevodov smo izvedli z metriko WER, kjer smo kot referenčna besedila uporabili ročno popravljene prevode vrednotenega sistema. WER je samodejna metoda, ki računa stopnjo napačnih besed (word error rate – WER). Metoda je največkrat uporabljena za vrednotenje kakovosti prepoznave govora in strojnih sistemov prevajanja. Uporablja se tudi za primerjavo kakovosti različnih prevajalnih sistemov ali za vrednotenje izboljšav znotraj določenega prevajalnega sistema. Vsekakor nam WER metoda ne poda nobenih informacij o naravi napak v prevajalnem sistemu, zato je potrebno dodatno delo za indentifikacijo vzroka oziroma izvira napake.

WER temelji na uteženi Levenshteinovi razdalji (weighted Levenshtein edit-distance) (Fu, 1982). Ta predstavlja razširitev osnovne razdalje (Levenshtein, 1965), ki šteje najmanjše število sprememb, ki jih moramo opraviti med referenčnim prevodom in prevodom prevajalnega sistema. Število sprememb še utežimo z dolžino povedi. Dovoljene spremembe so vstavitev, brisanje in zamenjava besede.

Izračun vrednosti WER za eno poved je oblike:

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

kjer je

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

Rezultat, ki nam ga poda WER metrika, predstavlja stopnjo napake prevajalnega sistema. V primeru, da želimo rezultat predstaviti kot kakovost prevajalnega sistema, uporabimo različico metrike, ki opisuje stopnjo prepoznavnih besed (word recognition rate – WRR). WRR metodo si lahko predstavimo kot razliko med popolnim prevodom ter napako sistema.

Izračun vrednosti WRR je oblike:

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

Postopek vrednotenja je sledeč:

- izberemo izvorno besedilo,
- prevedemo izvorno besedilo,
- ročno popravimo prevedeno, tako da naredimo čim manj sprememb,
- izračunamo WER med predovom in popravljenim besedilom (referenčnim prevodom).

4.2 Vrednotenje kakovosti jezikovnih gradiv

Vrednotenje kakovosti jezikovnih gradiv smo razdelili na dva dela:

- testiranje medsebojne pokritosti jezikovnih gradiv,
- testiranje pokritosti korpusov,

Namen vrednotenja je predvsem ugotoviti pokritost slovarjev ter kakovost le-teh, saj na tak način ugotovimo ali enojezični slovarji vsebujejo vse potrebne analize besed ali ne. Prvi del smo opravili z metodo testvoc (Tyers, 2012) , drugi del s pomočjo dveh korpusov. Prvi korpus, na katerem smo računali pokritost slovarjev, je MULTEXT-EAST (Dimitrova et al., 1998; Erjavec, 2010) (Orwell), ki je večjezična zbirka jezikovnih gradiv, zapisana v standardizirani obliki. Trenutna zbirka vsebuje večji del srednjih ter vzhodnoevropskih jezikov, skupaj 17. Gradiva sestavljajo oblikoskladenjske specifikacije, oblikoskladenjski leksikoni ter označeni vzporedni, primerjalni in govorni korpusi.

Drugi korpus, na katerem smo računali pokritost slovarjev, je OPUS (Tiedemann, 2012). OPUS je zbirka prevedenih besedil iz spleta, sam namen projekta je pretvoriti ter uskladiti brezplačne spletne podatke z dodatnimi jezikovnimi oznakami ter le-te prostodostopno nuditi javnosti. Pri OPUS zbirki smo se omejili le na zbirko podnapisov, ki je bila že sama po sebi dovolj obširna.

Rezultati, ki so bili doseženi z metodo vrednotenja kakovosti jezikovnih gradiv, so prikazani v naslednjem poglavju.

Poglavje 5

Rezultati

Pričujoče poglavje predstavlja in opisuje osnovne statistike jezikovnih gradiv, ki so bila ustvarjena v sklopu tekočega projekta. Podrobneje si bomo ogledali pokritost slovarjev in korpusov, rezultate, ki smo jih dosegli z metodo testvoc ter rezultate vrednotenja kakovosti prevodov.

5.1 Pokritost slovarjev

V tabeli 5.1 si lahko ogledamo nekaj lastnosti slovarjev, ki smo jih zgradili v sklopu projekta. Natančneje si lahko ogledamo, koliko slovarskih gesel vsebuje enojezični slovar izvornega jezika – slovenščine ter koliko slovarskih gesel vsebuje enojezični slovar ciljnega jezika – hrvaščine. Pomembno je omeniti, da je bilo v sklopu tekočega projekta dodanih preko 4.000 slovarskih gesel v enojezični slovar ciljnega jezika – hrvaščine. Ogledamo si lahko tudi pokritost dvojezičnega slovarja, natančneje, koliko besed si lasti primerne prevode.

Tabela 5.1: Lastnosti slovarjev.

Slovar	Št. slovarskih gesel (lem)
Enojezični slovar izvornega jezika – slovenščine	25.923 (1.901 paradigem)
Enojezični slovar ciljnega jezika – hrvaščine	17.330 (1.014 paradigem)
Dvojezični slovar	17.330 (slovarski vnosi)

Enojezični slovar izvornega jezika je veliko obširnejši od slovarja ciljnega jezika. V sklopu projekta so bili izdelani tudi spiski besed s pripadajočimi prevodi, ki niso bili zajeti v prevajalni smeri iz slovenskega jezika v hrvaški jezik. S pomočjo spiskov bomo v prihodnosti imeli veliko manj dela z dodajanjem manjkajočih prevodov oziroma besed v enojezični slovar ciljnega jezika – hrvaščino. Več informacij glede pokritosti samih besednih vrst, je predstavljenih v razdelku 5.3, kjer opisujemo rezultate tako imenovane metode testvoc (Tyers, 2012) – metode za testiranje posameznih slovarjev.

5.2 Vrednotenje kakovosti jezikovnih gradiv – pokritost korpusov

V tabeli 5.2 si lahko ogledamo rezultate, ki smo jih dosegli z metodo vrednotenja kakovosti jezikovnih gradiv. Metodo smo izvedli na dveh različnih korpusih, podrobneje opisanih v predhodnem razdelku 4.2. Iz tabele lahko razberemo število besed, vsebovanih v vsakem korpusu posebej, doseženo povprečje pokritosti ter standardno deviacijo.

Tabela 5.2 prikazuje pokritost korpusov MULTEXT-EAST (Orwell) ter OPUS (subs).

Tabela 5.2: Pokritost korpusov.

Korpus	Št. besed	Povprečje	STDEV
MULTEXT-EAST (Orwell) SL	104.482	94,23%	0,15%
OPUS (subs) SL	2.562.969	91,72%	0,21%
OPUS (subs) HR	307.564	77,34%	0,31%

Ugotovili smo, da s trenutno verzijo enojezičnega slovarja izvornega jezika, slovenščine, pokrivamo kar 94,23% MULTEXT-EAST korpusa (Orwell) oziroma 91,72% OPUS korpusa (subs). Enojezični slovar ciljnega jezika, hrvaščine, pa pokriva 77,34% OPUS korpusa (subs).

5.3 Testiranje slovarjev

Testiranje slovarjev smo razdelili na dva dela in sicer na testiranje medsebojne pokritosti slovarjev in na pokritost korpusov.

Pokritost slovarjev smo testirali z metodo testvoc (Tyers, 2012). Metoda razširi enojezični slovar izvornega jezika, nato pa testira vsako možno analizo besede skozi vse faze prevajalnega sistema. Na tak način ugotovimo, katera analiza besede ima pravilen prevod v enojezičnem slovarju ciljnega jezika, torej brez simbolov # ali @. Pomen simbolov, ki označujejo napake, je sledeči:

- @ – beseda ne vsebuje prevoda v dvojezičnem slovarju,
- # – beseda se ne prevede pravilno – oblikoskladenjske oznake niso pravilno označene.

Testiranje enojezičnih slovarjev z metodo testvoc je zelo pomembno v prvi fazi razvoja jezikovnega para. S pozitivnimi rezultati zagotovimo, da se vse analize besed, ki so vsebovane v enojezičnih slovarjih, pravilno prevedejo.

Naslednja faza razvoja je pisanje pravil strukturnega prenosa, ki urejajo večbesedne prevode oziroma stavke, le te pa nam večkrat lahko povzročijo

nezaželene ali nepričakovane napake oziroma napake, ki niso vidne pri enobesednih prevodih. Posledično jezikovni par potrebuje dodatno testiranje prevajalnega sistema na korpusih, kjer se lahko skrite napake pojavijo.

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

Tabela 5.3: Rezultat testvoc (Smer: hrvaščina – slovenščina).

B. vrsta	Skupno	Pravilni	Z @	Z #	%
pridevniki	1.517.798	1.517.798	0	0	100
glagoli	1.018.517	1.018.517	0	0	100
imena	726.576	726.576	0	0	100
samostalniki	135.031	135.031	0	0	100
p. glagoli	35.112	35.112	0	0	100
zaimki	10.683	10.683	0	0	100
števniki	10.165	10.165	0	0	100
prislovi	8.568	8.568	0	0	100
predlogi	101	101	0	0	100
kratice	56	56	0	0	100
medmeti	49	49	0	0	100
vezniki	71	71	0	0	100

V tabeli 5.4 si lahko ogledamo rezultate testiranja enojezičnega slovarja ciljnega jezika. Rezultati prikazujejo kakovost prevajanja posameznih besed iz hrvaškega jezika v slovenski jezik.

Tabela 5.4: Rezultat testvoc (Smer: slovenščina – hrvaščina).

B. vrsta	Skupno	Pravilni	Z @	Z #	%
pridevniki	749.994	263.260	370.603	116.131	35.2
glagoli	77.254	58.991	495	17.768	76.4
imena	437.433	437.433	0	0	100
samostalniki	72.478	72.478	0	0	100
p. glagoli	120	120	0	0	100
zaimki	3.382	3.382	0	0	100
števniki	8991	8991	0	0	100
prislovi	7.388	4.739	1.610	1.039	64.2
predlogi	84	84	0	0	100
kratice	56	56	0	0	100
medmeti	49	49	0	0	100
vezniki	56	56	0	0	100

V tabeli 5.4 lahko opazimo napake, ki so posledica manjkajočih prevodov

v dvojezičnem slovarju ali nepravilne postavitve oblikoskladenjskih oznak. Napake so prisotne pri pridevnikih, glagolih ter prislovih. Pomembno je poudariti, da so le-te posledica manjkajočih besed v enojezičnem slovarju ciljnega jezika – hrvaščine. Število manjkajočih besed je razvidno v tabeli 5.2.

5.4 Rezultati vrednotenja kakovosti prevodov

Evalvacijo oziroma vrednotenje kakovosti prevodov smo izvedli z uporabo WER metode, podrobneje opisane v razdelku 4.1. Pridobljeni rezultati so zelo dobri, saj moramo upoštevati dejstvo, da sistem še ni dokončan, in da smo se omejili le na prvi nivo pravil strukturnega prenosa. Samodejno metriko WER smo uporabili na manjšem testnem vzorcu (van Gompel., 2012), ki je bil ročno pripravljen in uporabljen v vseh novih sistemih projekta GSOC2012. Vzorec je bil uporabljen v fazi razvoja, saj vsebuje enostavne povedi. Vrednost metrike WER na testnem vzorcu je 13.7% (WRR = 86.3%).

Vrednotenje smo pognali tudi na članku, ki je bil prevzet iz wikipedije (Wikipedia, 2012b). Izbrali smo članek v hrvaškem jeziku, ki opisuje Hrvaško. Osredotočili smo se na članke iz wikipedie, saj menimo, da so le-ti napisani v slogu, ki je trenutno najbolj uporabljen, zato tudi želimo doseči zelo dobre rezultate na tem področju, saj bomo na tak način pokrili veliko večino besedila hrvaškega jezika. Izbrani članek je bil ročno preveden v slovenski jezik. Vrednost metrike WER na članku je 14.8% (WRR = 85.2%).

Poglavje 6

Zaključek in nadaljnje delo

6.1 Zaključek

Delo predstavlja postavitev prevajalnega sistema za jezikovni par slovenščina – hrvaščina. Sistem temelji na skupini strojnega prevajanja na osnovi pravil plitkega prenosa, podrobneje na prevajalnem sistemu Apertium, saj se je le-ta izkazal kot najprimernejši za postavitev sistema za strojno prevajanje sorodnih jezikov (Vičič, 2012). Predstavljene so bile lastnosti izvornega ter ciljnega jezika ter težave oziroma omejitve, ki so bile prisotne pri izdelavi gradiv. Podrobneje smo si ogledali delo, ki je bilo opravljeno v sklopu projekta GSOC2012, izpostavljene so bile zanimivosti posameznih komponent, ki so zahtevale dodatno delo. Za postavitev ogrodja sistema so bile uporabljene metode za samodejno izdelavo prevajalnega sistema, predstavljene v delu (Vičič, 2012). Delo, ki je opisano v poglavju 3, je bilo v večini izvedeno ročno, v manjšem obsegu pa s pomočjo dodatnih skript, ki so bile napisane izrecno za ta projekt.

Delo predstavlja tudi rezultate za novo izdelani jezikovni par, ki so bili doseženi v sklopu projekta GSOC2012. Zavedati se moramo, da lahko sistem, zasnovan na osnovi pravil plitkega prenosa, še izpopolnimo z dodatnim ročnim pregledom prevajalnih gradiv, saj eksplicitno zapisana pravila strukturnega prenosa ter slovarji omogočajo iterativno izboljševanje kakovosti prevodov.

Kakovost izdelanega prevajalnega sistema za jezikovni par slovenščina – hrvaščina je trenutno najboljša, če upoštevamo vse trenutno poznane sisteme za predstavljen jezikovni par (Vičič, 2010). Prevodi predstavljenega sistema že dosegajo produkcijsko kakovost, vseeno pa še obstaja možnost izboljšave. Ta je predvsem v dodajanju novih prevodov v dvojezični slovar, ki trenutno ne pokriva popolnoma enojezičnih slovarjev, posledično tudi dodajanje novih oziroma manjkajočih besed v enojezični slovar ciljnega jezika – hrvaščine. Vrednotenje samega sistema še ni bilo izvedeno v popolnosti,

zato pripravljamo obširnejše vrednotenje sistema in primerjavo z ostalimi prevajalnimi sistemi, ki vsebujejo opisan jezikovni par.

6.2 Nadaljnje delo

Razvoj jezikovnega para slovenščina – hrvaščina še zdaleč ni končan. Zelo smo zadovoljni z doseženimi rezultati, saj lahko sistem uvrstimo med sisteme produkcijske kakovosti, vendar bomo nadaljevali z razvojem jezikovnega para z dodajanjem manjkajočih besed v enojezični slovar ciljnega jezika – hrvaščino ter z razširitvijo samega besedišča enojezičnih slovarjev. Poleg izpopolnjevanja slovarjev je potrebno tudi dodatno delo na označevalniku oblikoskladenjskih oznak za izvorni jezik – slovenščino, saj je bil uporabljen izdelek projekta Apertium (sl-es)[1] (Horvat in Vičič, 2012), izdelan pa je bil s pomočjo nenadzorovane metode (Sánchez-Martínez et al., 2008). Čeprav smo za izdelavo označevalnika uporabili preko 50 milijonov besed iz korpusa OPUS, kakovost komponente še vedno ni primerljiva s kakovostjo pravil slovnice z omejitvami, ki skrbijo za prevajanje hrvaškega jezika v slovenski jezik. Zato smo se odločili, da bomo prevzeli pravila slovnice z omejitvami iz hrvaškega jezika ter jih preuredili za slovenski jezik.

Jezikovni par je potrebno testirati na več korpusih, saj bomo na tak način imeli večjo možnost zaznave napak, ki jih bomo rešili z dodatnimi pravili strukturnega prenosa ter zaznave manjkajočih besed. Dodatne manjkajoče besede oziroma spisek najpogostejših manjkajočih besed, ki niso vsebovane v slovarjih, bomo pridobili s pomočjo prevajanja slovenskih ter hrvaških člankov iz wikipedie.

Načinov, kako izboljšati trenutni jezikovni par je veliko, zato se z vsemi močmi zavzemamo, da bo razvoj še nadalje potekal brez večjih težav.

[1]Projekt Apertium (sl-es): `http://apertium.svn.sourceforge.net/svnroot/apertium/nursery/apertium-sl-es/`

Dodatek A

Primeri prevodov

A.1 Dobri prevodi

Primeri A.1 do A.6 prikazujejo dobre prevode, ki smo jih dosegli z uporabo izdelanega prevajalnega sistema. Prevajali smo iz hrvaškega jezika v slovenski jezik.

(A.1) *Marko ima novu igračku.*

Marko ima novo igračo.

(A.2) *Kupiti ću novi automobil.*

Kupil bom nov avtomobil.

(A.3) *Djevojčica je njegova sestra i ima pet godina.*

Deklica je njegova sestra in ima pet let.

(A.4) *Kad Marica završava s brojanjem, pogledava uokolo.*

Ko Marija preneha s štetjem, pogleda okoli.

(A.5) *Hrvatska (službeni naziv: Republika Hrvatska) je europska država, zemljopisno smještena na prijelazu iz Srednje u Jugoistočnu Europu.*

Hrvaška (uraden naziv: Republika Hrvaška) je evropska država, zemljepisno umeščena na prehodu iz Srednje v Jugovzhodno Evropo.

(A.6) *Hrvati čine 89,6% stanovništva, a najznačajnija nacionalna manjina su Srbi koji čine 4,5% stanovništva.*

Hrvati tvorijo 89,6% prebivalstva, a najpomembnejša narodna manjšina so Srbi, ki tvorijo 4,5% prebivalstva.

A.2 Prevodi, ki vsebujejo napake

Primer A.7 prikazuje napačno razdvoumljanje – *oči*. Modul, ki skrbi za ujemanje oblikoskladenjskih kategorij, je izbral napačen sklon: rodilnik namesto orodnika.

(A.7) *Marica sjedi i drži ruke ispred očiju.*

Marija sedi in drži roke pred oči.

Primer A.9 prikazuje napačno razdvoumljanje – *svojo*. Modul, ki skrbi za ujemanje oblikoskladenjskih kategorij, je izbral napačen spol: ženski namesto moškega.

(A.8) *Pavao mu je dao svoju putovnicu.*

Pavel mu je dal svojo potni list.

Primer A.9 prikazuje napačno razdvoumljanje – *prispel*. Modul, ki skrbi za ujemanje oblikoskladenjskih kategorij, je izbral napačen spol: moški namesto srednjega.

(A.9) *Uskoro je stigao zrakoplov.*

Kmalu je prispel letalo.

Literatura

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Laerty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, in David Yarowsky. Statistical Machine Translation, Final Report. Technical report, JHU, 1999.

Apertium. Apertium: machine translation toolbox, 2010. URL `http://sourceforge.net/projects/apertium`.

Chris Callison-Burch, Miles Osborne, in Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of EACL*, 249–256. Association for Computational Linguistics, 2006.

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.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevič, Tomaž Erjavec, Heiki Jaan Kaalep, in Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL*, 315–319. Association for Computational Linguistics, 1998.

Bonnie J. Dorr, Pamela W. Jordan, in John W. Benoit. A survey of current paradigms in machine translation. *Advances in Computers*, 1–68, 1999.

EGYPT. The EGYPT Statistical Machine Translation Toolkit, 2007. URL `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/`.

Tomaž Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, in Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. ELRA.

Mikel L. Forcada. Open-source machine translation: an opportunity for minor languages. In *Strategies for developing machine translation for mi-*

nority languages (5th SALTMIL workshop on Minority Languages), 1–7. Genoa, Italy, 2006.

King Sun Fu. *Syntactic Pattern Recognition and Applications*. Prentice-Hall, Englewood Cliffs, NJ, 1982.

GenPar. The GenPar Toolkit for Research on Generalized Parsing, 2010. URL `http://nlp.cs.nyu.edu/GenPar/`.

Google. The Google translator, july 2012a. URL `http://www.google.com/translate_t`.

Google. Google Summer of Code 2011, july 2012b. URL `http://www.google-melange.com/gsoc/homepage/google/gsoc2011`.

Google. Google Summer of Code 2012, july 2012c. URL `http://www.google-melange.com/gsoc/homepage/google/gsoc2012`.

J. Hajič, J. Hric, in V. Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, 7–12. Association for Computational Linguistics, 2000.

Aleš Horvat in Jernej Vičič. Strojno prevajanje med slovenščino in španščino. In *In Proceedings of the ERK*. Založba FE-FRI, 2012.

W. J. Hutchins in H. L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.

Philipp Koehn, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, in Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07)*, 177–180. Association for Computational Linguistics, 2007.

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

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

Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*, 173–180, 1984.

Franz Josef Och. Challenges in Machine Translation. In *Proceedings of the ISCSLP*, 15. Springer, 2006.

Promt, 2010. URL http://www.e-promt.com/.

Felipe Sanchez-Martinez in Hermann Ney. Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules. In Sampo Pyysalo Tapio Salakoski, Filip Ginter in Tapio Pahikkala, editors, *Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, 756–767. Springer-Verlag, August 2006. Copyright Springer-Verlag.

Felipe Sanchez-Martnez, Juan Antonio Perez-Ortiz, in Mikel L. Forcada. Integrating corpus-based and rule-based approaches in an open-source machine translation system. In Frank Van Eynde, Vincent Vandeghinste, in Ineke Schuurman, editors, *Proceedings of METIS-II Workshop: New Approaches to Machine Translation, a workshop at CLIN 17 - Computational Linguistics in the Netherlands*, 73–82, January 2007.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, in Mikel L. Forcada. Using target-language information to train part-of-speech taggers for machine translation. *Machine Translation*, 22(1-2):29–66, 2008. DOI: 10.1007/s10590-008-9044-3.

Jörg Tiedemann. News from opus - a collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, 5:237–248, 2009.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*. Založba FE-FRI, 2012.

Jože Toporišič. *Slovenska slovnica*. Založba Obzorja, Maribor, 2000.

F. M. Tyers. Data consistency and quality, september 2012. URL http://wiki.apertium.eu/index.php/Session_7:_Data_consistency_and_quality.

Maarten van Gompel. UniLang - Where is James?, september 2012. URL http://www.unilang.org/ulrview.php?res=394,387.

Jernej Vičič. Strojno prevajanje in slovenščina. In *Proceedings of the 13th International Multiconference Information Society - IS 2010*, 47–52. Institut Jožef Stefan, Institut »Jožef Stefan«, Ljubljana, 2010.

Jernej Vičič. *Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike*. PhD thesis, Univerza v Ljubljani, 2012. URL http://eprints.fri.uni-lj.si/1778/.

Wikipedia. Wikipedia, the free encyclopedia, september 2012a. URL http://en.wikipedia.org/wiki/Slovenia.

Wikipedia. Hrvatska – Wikipedia, the free encyclopedia, september 2012b. URL http://hr.wikipedia.org/wiki/Hrvatska.