

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

DOKTORSKA DISERTACIJA

**Hitra postavitev prevajalnih sistemov
na osnovi pravil za sorodne naravne
jezike**

Jernej Vičič

Mentor: prof. dr. Igor Kononenko
Somentor: doc. dr. Tomaž Erjavec

Ljubljana, 2011

Zahvala

Hvala!

Mojim.

Kazalo

Povzetek	1
Abstract	2
1 Uvod	3
1.1 Motivacija	3
1.2 Obstoječi sistemi strojnega prevajanja sorodnih jezikov	4
1.3 Pregled vsebine	5
2 Pregled področja	7
2.1 Osnovni pojmi	7
2.1.1 Upogibna morfologija	7
2.1.1.1 Označevanje POS - Part Of Speech	7
2.1.1.2 Paradigma	7
2.1.1.3 Lema	8
2.1.1.4 Krn	9
2.1.2 Drevo izpeljav	9
2.1.3 Plitka analiza in plitki transfer besedil	10
2.1.4 Morfem	10
2.1.5 Morfološka analiza besedil	10
2.1.6 Pravila transferja na osnovi regularnih izrazov	10
2.1.7 Statistični model jezika	11
2.1.8 Končno urejanje - post-editing	11
2.2 Slovanski jeziki	11
2.3 Podobnosti slovanskih jezikov pri prevajanju (free rides)	13
2.3.1 Tipološka podobnost	13
2.3.2 Skladenjska podobnost	14
2.3.3 Morfološka podobnost	15
2.3.4 Leksikalna podobnost	17

KAZALO

2.3.5	Lažni prijatelji	17
3	Sistemi za strojno prevajanje	19
3.1	Razdelitev	19
3.1.1	Statistično strojno prevajanje - SMT	20
3.1.2	Statistično strojno prevajanje z razčlenjevanjem - SMT by parsing	21
3.1.3	Strojno prevajanje na osnovi primerov - EBMT	21
3.1.4	Strojno prevajanje na osnovi pravil - RBMT	22
3.1.5	Strojno prevajanje na osnovi pravil plitkega prenosa ter plitke analize - shallow parsing and transfer RBMT	22
3.1.6	Strojno prevajanje sorodnih naravnih jezikov ter nesorodnih jezikov v ozko omejenih domenah	23
3.2	Strojno prevajanje na osnovi pravil plitkega prenosa ter plitke analize - shallow parsing and transfer RBMT	24
3.3	Orodja za postavitev prevajalnih sistemov	25
3.4	Apertium - odprtokodno ogrodje za prevajalni sistem sorodnih jezikov	26
3.4.1	Arhitektura ogrodja Apertium	27
3.4.2	Predlagana spremenjena arhitektura	30
3.4.3	Pregled uporabljenih podatkovnih tipov	32
4	Morfologija in leksikoni	35
4.1	Enojezični morfološko označen slovar	35
4.2	Dvojezični slovar	36
4.3	Uporabljena učna gradiva	39
4.4	Metode	40
4.4.1	Izdelava enojezičnih slovarjev izvirnega in ciljnega jezika z morfološkimi oznakami	40
4.4.1.1	Izdelava paradigem	40
4.4.2	Izdelava dvojezičnih prevajalnih slovarjev	42
4.4.3	Izdelava statističnega jezikovnega modela ciljnega jezika	44
4.4.4	Modeliranje morfoloških oznak izvirnega jezika	45
5	Pravila transferja	46
5.1	Pravila regularnih izrazov	47
5.2	Apertiumov format pravil	47
5.3	Samodejna izdelava pravil	48
5.3.1	Izdelava pravil za plitvi prenos na osnovi regularnih izrazov	48
5.3.2	Opis metode	48

5.3.2.1	Izdelava pravil	50
5.3.3	Izbira najboljših pravil	51
5.3.4	Izdelava pravil na osnovi regularnih izrazov za izražanje lokalnega ujemanja morfoloških kategorij	51
5.3.5	Primeri uporabe pravil za lokalno ujemanje morfoloških kategorij	52
6	Metodologije vrednotenja sistemov in rezultati vrednotenj	56
6.1	Evalvacija sistemov za strojno prevajanje	56
6.1.1	Samodejne metode	56
6.1.1.1	Metrika BLEU	56
6.1.1.2	Metrika METEOR	57
6.1.2	Metode, ki vključujejo posege strokovnjakov	58
6.1.2.1	Utežena Levenshteinova razdalja	58
6.1.2.2	Evalvacija po smernicah LDC	58
6.2	Rezultati	59
6.2.1	Opis sistemov	60
6.2.2	Izbrane evalvacijske metrike	60
6.2.2.1	Samodejna objektivna evalvacija z metriko METEOR	61
6.2.2.2	Evalvacija z metodo, ki vključuje posege strokovnjakov na podlagi utežene Levenshteinove razdalje	61
6.2.2.3	Evalvacija z metodo, ki vključuje posege strokovnjakov na podlagi smernic (LDC, 2005)	63
7	Prevajanje na osnovi dreves izpeljave	66
7.1	Osnove	66
7.2	Metoda	67
7.2.1	Učenje povezav med oznakami in drevesi	67
7.2.2	Prevajanje	68
8	Razprava in nadaljnje delo	70
8.1	Prispevki k znanosti	70
8.1.1	Umestitev pričakovanih prispevkov k znanosti	71
8.1.2	Metoda za statistično strojno prevajanje z drevesi izpeljave za manj uporabljane jezike (less-used languages)	71
8.1.3	Metoda za samodejno označevanje paradigem	71

8.1.4	Samodejno luščenje paradigem za visoko pregibne jezike ter izdelava pripadajočih leksikonov	73
8.1.5	Ocenjevanje pravil za strukturni	73
8.1.5.1	Raziskava možnih uporab ocenjevanja pravil	73
8.1.5.2	Raziskava algoritmov za izbiro pravil	74
8.1.5.3	Izdelava metrike za ocenjevanje pravil	74
8.1.6	Hitra izdelava prevajalnega sistema na osnovi RBMT za so- rodne jezike	74
8.2	Prevajalni sistem GUAT	74
8.3	Nadaljnje delo	75
A	Prva priloga	76
A.1	Pravila transferja	76
A.2	Primeri prevodov	87
A.2.1	Dobri prevodi	87
A.2.2	Napake	87
	Seznam slik	88
	Seznam tabel	91
	Literatura	92

Seznam uporabljenih kratic in simbolov

POS Part Of Speech, označba besedne vrste, prve kategorije MSD

MSD Morfosintaktični deskriptor, oznaka morfoloških in sintaktičnih lastnosti besede

MT Machine Translation, strojno prevajanje

SMT Statistical Machine Translation, statistično strojno prevajanje

EBMT Example Based Machine Translation, strojno prevajanje na osnovi primerov

RBMT Rule Based Machine Translation, strojno prevajanje na osnovi pravil

CBMT Corpus Based Machine Translation, strojno prevajanje na osnovi korpusov

HMT Hybrid Machine Translation, hibridno (mešano) strojno prevajanje

PBMT Phrase-Based Machine Translation, strojno prevajanje na osnovi fraz

WER Word Error Rate

WRR Word Recognition Rate

LDC Linguistic Data Consortium

NIST National Institute of Standards and Technology

BLEU Bilingual Evaluation Understudy

METEOR Metric for Evaluation of Translation with Explicit ORdering

WFST Weighed Finite State Transducers

GNU GNU Not Unix

GPL GNU General Public License

LGPL GNU Lesser General Public License

XML Extensible Markup Language

Povzetek

Delo predstavlja pregled strojnega prevajanja naravnih jezikov, osredotoča se predvsem na paradigmo sistemov za strojno prevajanje na osnovi pravil plitkega prenosa, ki so najprimernejši za postavitev sistemov za strojno prevajanje sorodnih jezikov. Predstavljene so metode za hitro izdelavo vseh gradiv, ki so potrebna za postavitev takšnih sistemov. Metode so preizkušene na postavitvah delujočih prototipov prevajalnih sistemov, za vsak prototip je bila izvedena tudi evalvacija kakovosti prevodov.

Ključne besede:

strojno prevajanje, strojno prevajanje sorodnih naravnih jezikov, tehnologije hitrih postavitvev sistemov

Abstract

The work presents an overview of the machine translation of natural languages, focusing in particular on the paradigm of machine translation systems based on shallow transfer rules, which are most suitable for the installation of machine translation systems of related languages. Method for rapid development of all materials needed for the installation of such systems are presented. The methods are tested on working prototypes of translation systems. The translation quality evaluation was carried out for each prototype.

Key words:

machine translation, machine translation of related languages, speeding up the implementation of machine translation systems

Poglavje 1

Uvod

1.1 Motivacija

Sorodnost naravnih jezikov ene tipološke skupine (in včasih celo jezikov različnih tipoloških skupin, primer poskusa prevajalnega sistema za jezikovni par češčina - litovščina, opisan v Hajič et al. (2003)), omogoča lažje in natančnejše prevajanje ter omogoča uporabo enostavnejših metod, ki ne bi bile dovolj dobre za uporabo v prevajalnih sistemih nesorodnih jezikovnih parov. Uporaba preprostejših metod ne pomeni slabše kakovosti prevodov, veliko napak sistemov za prevajanje izvira ravno iz napak v analizi (parsing) izvornih povedi. Seštevanje napak v analizi, prenosu in generiranju pri sistemih za strojno prevajanje na osnovi pravil s klasično arhitekturo pogosto prinese slabše rezultate kot uporaba enostavnih metod plitke analize in prenosa.

Ena od glavnih ovir, ki upočasnjujejo proces razvoja prevajalnih sistemov na osnovi pravil je obseg človeškega dela, ki je nujno za oblikovanje pravil slovnice in slovarjev. Tudi sistem, ki so namenjeni prevajanju sorodnih jezikov se soočajo s tem problemom. Takšni sistemi navadno uporabljajo poenostavljenih arhitekturo in izkoriščajo podobnost jezikov z uporabo plitve slovnice in pravil prenosa, ampak tudi ta pravila zahtevajo veliko napora. To delo predstavlja pregled metod, ki omogočajo samodejno ustvarjanje vseh gradiv, ki so potrebna za postavitve sistema za prevajanje naravnih sorodnih jezikov.

1.2 Obstoječi sistemi strojnega prevajanja sorodnih jezikov

RUSLAN (Hajič, 1987), je prvi sistem za strojno prevajanje sorodnih jezikov. Prevajalni par sistema je bil češčina - ruščina. Sistem je uporabljal globoko sintaktično analizo (deep syntactical analysis) in . Uporaba je bila omejena na prevajanje uporabniških navodil.

Česilko (Hajič et al., 2000), je sistem za strojno prevajanje sorodnih jezikov češčine in slovaščine, arhitektura osnovne različice je bila enostavna, slovarji z direktni prevodi lem ena-na-ena z leksikalnim transferjem brez dodatnih pravil. Sistem je izkoriščal veliko podobnost jezikovnega para. Kasneje je bil sistem izpopolnjen (Łukasz Dębowski et al., 2002), dodan mu je bil tudi nov jezikovi par, češčina-poljščina.

Osnovna arhitektura sistema:

- morfološko označevanje izvirnega besedila
- dvojezični slovarji
- morfološka sinteza v ciljno besedilo

Plitvi so predlagali ter ga implementirali na poskusni različici sistema v (Hajič et al., 2003). Osnovna arhitektura spremenjenega sistema:

- morfološka analiza izvirnega besedila
- morfološko razdvoumljanje
- leksikalni/morfološki
- morfološka sinteza v ciljno besedilo

Natančnost prevodov z metodo WRR (Vogel et al., 2000) je okrog 90% za jezikovni par češčina-slovaščina ter 71,4% za jezikovni par češčina-poljščina.

GUAT (Vičič, 2009), je sistem plitkega transferja temelječ na ogrodju Apertium. Sistem podpira jezikovni par slovenščine-srbščina.

PONS (Dyvik, 1995), partial translation between related languages. Sistem za strojno prevajanje sorodnih skandinavskih jezikov; jezikovni par: norveščina-švedščina. Zanimiva lastnost sistema je, da ne uporablja morfološke analize, slovar izvirnega jezika je hranil vse besedne oblike. Sistem uporablja delno analizo izvornih povedi, povedi razdeli na kose ter manjše enote analizira.

T4F (Ahrenberg in Holmqvist, 2005), tokenization, Tagging, Transfer Transposition and Filtering. Avtorji sistema trdijo, da za strukturno sorodne jezike, abstrakna skladišna analiza ni potrebna oziroma prinaša slabe rezultate.

Turkijski jeziki (Altintas in Cicekli, 2002), avtorji so postavili sistem za strojno prevajanje sorodnih turkijskih jezikov, prevajalni par turščina - krimijska tatarščina (Crimean tatar). Avtorji trdijo, da za jezike s skupno zgodovino in podobno kulturo, ne potrebujemo semantične analize. Sistem se osredotoča na razlike na morfološki ravni.

Keltski jeziki (Scannell, 2006), sistem za strojno prevajanje med irščino (Irish) ter škotsko gelščino (Scottish Gaelic) je predstavljen v . Jezika sta si slovnično sorodna, saj imata skupnega prednika - srednjo irščino (Middle Irish).

Apertium (Corbi-Bellot et al., 2005), zbirka orodij za postavitev prevajalnih sistemov za sorodne jezike, širše je predstavljen v Razdelku 3.4. Apertium je bil najprej mišljen kot orodje za postavitev sistemov za strojno prevajanje sorodnih romanskih jezikov, tako so nastali tudi prvi jezikovni pari katalonščina - španščina, španščina - portugalščina in katalonščina - portugalščina.

1.3 Pregled vsebine

Poglavje 2 predstavlja pregled raziskovalnega področja ter razlage osnovnih pojmov znanstvenega področja, ki bralcu približajo znanstveno področje ter omogočijo nadaljnje branje. Poglavje 3 predstavlja eno od možnih razdelitev strojnega prevajanja z opisom posameznih paradigem strojnega prevajanja. Poseben poudarek je posvečen prevajanju sorodnih jezikov oziroma nesorodnih jezikov v omejenih domenah. Predstavljeno je ogrodje Apertium, ki predstavlja osnovno platformo za večino implementacij metod, ki so predstavljene v tem delu. Poglavje 4 predstavlja morfološko označene slovarje, enojezične in večjezične. Predstavljene so metode za samodejno izdelavo morfološko označenih slovarjev, ki so uporabljeni v strojnih prevajalnih sistemih. Poglavje 5 predstavlja pravila transferja, ki pri strojnih

prevajalnih sistemih na osnovi pravil omogočajo opisovanje razlik med jezikovnima paroma. Poglavje 6 predstavlja osnove evalvacije sistemov za strojno prevajanje, predstavljene so uporabljene metrike ter metodologije evalvacije. V zadnjem delu poglavja so predstavljeni rezultati evalvacij posameznih sistemov zgrajenih na osnovi metod, predstavljenih v Poglavjih 4 ter 5. Poglavje 7 predstavlja metodo, ki omogoča izdelavo sistema za strojno prevajanje na osnovi dreves izpeljave za manj uporabljane jezike oziroma za jezike, ki nimajo izdelanega skladiščno označenega dvojezičnega korpusa. Poglavje 8 zaključuje delo z razpravo in s smernicami za nadaljnje delo. V Prilogi A so prikazani primeri pravil transferja, pravil za ujemanje morfoloških kategorij bližnjih besed ter primeri prevodov predstavljenih sistemov.

Poglavje 2

Pregled področja

2.1 Osnovni pojmi

2.1.1 Upogibna morfologija

Upogibno morfologijo (inflectional morphology) po (Janda, 2007) le težko umestimo v enotno področje, umestimo jo na mejo med besedoslovje (leksiko) ter skladnjo (sintakso) jezika. V večini jezikov označuje relacije med osebo, številom, sklonom, spolom, časom in drugimi lastnostmi.

2.1.1.1 Označevanje POS - Part Of Speech

Označevanje Part-of-speech - POS, imenovano tudi slovnično označevanje je proces označevanja posamezne besede v besedilu s primerno oznako POS upoštevajoč definicijo besede ter tudi njeno okolico v besedilu (povezava z okoliškimi besedami). Oznake POS predstavljajo oznake besednih vrst, z označevanjem POS pa opisujemo proces označevanja morfoloških ter v nekaterih primerih tudi morfosintaktičnih značilk kot so oznake MSD definirane v (Erjavec, 2004).

Označevanje POS je težji problem kot samo uporaba seznama besed z ustreznimi oznakami POS, saj velik del besed predstavlja različne oznake POS, odvisno od uporabe v besedilu. Največji problem označevanja POS je odpravljanje dvoumnosti (disambiguation), izbiranje najprimernejše oznake v odvisnosti od konteksta v primeru več možnih oznak.

2.1.1.2 Paradigma

Paradigme, v našem primeru pregibne paradigme (inflectional morphology paradigm), so večdimenzionalne, potencialno rekurzivne matrike, ki so določene z obli-

koslovnimi značilnostmi besednih oblik in obrazil. Teoretični status pregibnih paradigem je kontroveržno razpravljan v morfološki teoriji. (Lieber, 1992), na primer, je trdi, da so paradigme so le skupine, podobno kot seznam povezanih stavkov. Tudi v (Halle in Alec, 1993) so pregibne paradigme predstavljene brez teoretičnega statusa. V večini drugih okvirov, na primer (Wurzel, 1987) ali (Spencer, 1991), pa pregibna paradigma opredeljuje sklop pregibanih besednih oblik za vsak leksem, ki sodi v neko skladiščno kategorijo.

Za potrebe pričujočega dela zadošča če opis razred elementov s podobnostmi, v našem primeru bodo elementi besede oziroma besedne zveze (words, phrases and phrase chunks). Predstavljajo razrede elementov, s katerimi rokujemo po enotnem načelu.

Oglejmo si še primer: če bi besede razdelili le na besedne vrste, bi težko ravnali z njimi, na primer samostalniki. Porazdelitev samostalnikov na sklanjatvene vzorce kot jih predstavlja (Toporišič, 2000) omogoča izdelavo relativno majhne zbirke pravil, ki omogočajo rokovanje z veliko množico besed (besede razdelimo na krne ter s končnicami izbiramo ostale jezikovne kategorije, kot so spol, število, sklon). Povezava med posamezno besedno obliko ter njeno paradigmo poteka prek osnovne besedne oblike - leme. Tak način opisa besed prinaša poseben problem samodejnega označevanja parov lema-paradigma, ki je opisan v razdelku 4.4.1.1.

2.1.1.2.1 Samodejno luščenje paradigem Ena od možnih delitev metod za luščenje morfologije iz besedil je deljenje metod glede na gradivo; na metode, ki delujejo na izvornem besedilu ter na metode, ki zahtevajo (morfološko) označeno besedilo. Primeri metod za samodejno luščenje paradigem iz neoznačenega korpusa:

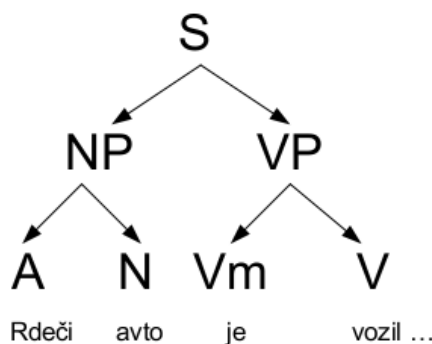
- (Sagot, 2005)
- (Goldsmith, 2001)
- (Creutz, 2006)

Primeri metod za samodejno luščenje paradigem iz označenega korpusa:

- (Erjavec in Džeroski, 2004)
- (Vičič, 2007a)

2.1.1.3 Lema

Lema v morfologiji, lemma in morphology, predstavlja kanonično obliko nekega leksema. Leksem se v tem kontekstu nanaša na sklop vseh oblik neke besede, ki



Slika 2.1: Drevo izpeljav za stavek *Rdeči avto je vozil*. S - stavek, N - samostalnik(noun), V - glagol(verb), Vm - pomožni glagol(modal verb), NP - samostalniška fraza(noun-phrase), VP - glagolska fraza(verb phrase).

imajo enak pomen, lema pa se nanaša na besedo, ki je izbrana zato, da predstavlja leksem. Proces za določanje leme se imenuje lematizacija.

2.1.1.4 Krn

Krn v morfologiji, stem in morphology, predstavlja korensko obliko besede. Krnjenje (stemming) je jezikovno odvisen postopek, pri katerem poskušamo najti niz znakov, imenujemo ga krn, ki lahko predstavlja vse oblike neke besede in istočasno to besedo loči od vseh ostalih. Pogosto, vendar ne nujno, krn ustreza korenu besede. Krnjenje je še posebej pomembno pri avtomatskem indeksiranju besedil v jezikih z bogato morfologijo, kakršna je tudi slovenščine.

2.1.2 Drevo izpeljav

Drevo izpeljav (parse tree ali concrete syntax tree) je urejeno drevo s korenem, ki predstavlja sintaktično strukturo niza glede na neko (formalno) slovnico. Sestavljeno je iz korena, vrhnje vozlišče, vej drevesa, notranja vozlišča ter listov, končna vozlišča. Pri drevesu izpeljav so notranja vozlišča označena z ne-terminalnimi simboli (non-terminals) slovnice, listi drevesa, končna vozlišča, pa s terminalnimi simboli slovnice. Drevesa izpeljav lahko izdelamo za povedi naravnih jezikov glede na slovnice teh jezikov. Slika 2.1 kaže primer jezikoslovnega drevesa izpeljav za enostaven slovenski stavek *"Rdeči avto je vozil"*.

2.1.3 Plitka analiza in plitki transfer besedil

Plitka analiza besedil (shallow parsing, tudi chunking ter "light parsing"), je analiza povedi, ki določa osnovne gradnike povedi (samostalnške skupine, glagoli, glagolske skupine itd), vendar pa ne navaja notranjo gradnikov, niti njihove vloge v glavni povedi.

Plitki je naravno nadaljevanje plitke analize pri postavitvi prevajalnega sistema, saj omogoča prenos gradnikov plitke analize izvirne povedi v gradnike ciljnega jezika. Plitki je sestavljen iz prenosa posameznih besed, ponavadi v lematizirani obliki (glej razdelek 2.1.1.3), ter pravil plitkega transferja, ki služijo za opis sprememb med izvirnim in ciljnim jezikom. Ta pravila so omejena na lokalne spremembe, ponavadi ujemanje soležnih besed v leksikalnih kategorijah ter spremembo lokalnega vrstnega reda besed. Primeri pravil so prikazani v Prilogi A.

2.1.4 Morfem

Morfem je v jezikovnem sistemu najmanjša enota besede s samostojnim pomenom. Na izrazni ravni morfeme sestavljajo fonemi (najmanjše razločevalne enote jezikovnega sistema), v pisni obliki pa so morfemi sestavljeni iz grafemov (najmanjših enot pisnega jezika).

2.1.5 Morfološka analiza besedil

Morfološka analiza je proces razdeljevanja besed na njihove morfeme (pomenske enote). Morfološka analiza je bistvena komponenta aplikacij jezikovnih tehnologij, uporabna pri odkrivanju pravopisnih napak (spelling error correction), strojnem prevajanju in drugih problemih. Izvajanje polne morfološke analize besedila običajno zahteva delitve besed na morfeme ter analizo interakcije teh morfemov, ki določajo skladenjske razrede besednih oblik kot celote. Kompleksnost morfološke analize se močno razlikuje med naravnimi jeziki, vendar velja relativno težek problem že v relativno preprostih primerih, kot je angleščina.

2.1.6 Pravila transferja na osnovi regularnih izrazov

Plitki strukturni transfer (shallow structural) omogoča premostitev slovničnih razlik obravnavanega jezikovnega para. Temelji na tehnologiji končnih avtomatov za odkrivanje vzorcev leksikalnih enot (morfosintaktično označenih kosov besedila ali fraz) fiksne dolžine, ki zahtevajo posebno obdelavo glede na slovnične razlike med jezikoma (na primer: spremembe v spolu, sklonu ali številu za zagotovitev ujemanja v ciljnim jeziku).

2.1.7 Statistični model jezika

Statistični model jezika dodeljuje verjetnost zaporedju besed s pomočjo verjetnostne porazdelitve. Jezikovno modeliranje se uporablja v mnogih primerih jezikovnih tehnologij, kot so prepoznavanje govora, strojno prevajanje, označevanje POS, razčlenjevanje in priklic informacij (information retrieval). V članku je uporabljen preprost jezikovni model, ki temelji na tri-gramih.

2.1.8 Končno urejanje - post-editing

Končno urejanje, postediting, je postopek za izboljšanje strojno ustvarjenih prevodov s čim manj ročnega dela, vključuje popravljanje strojno izdelanih prevodov za zagotovitev ravni kakovosti, ki je bila dogovorjena pred med stranko prevajalcem. Določene ponovljive operacije lahko avtomatiziramo oziroma vključimo v same strojne prevajalnike.

2.2 Slovanski jeziki

Slovanski jeziki so velika jezikovna družina v srednji in vzhodni Evropi ter na Balkanu in v delu Azije. Največji slovanski jezik je ruščina, ki mu sledi ukrajinščina. Slovanski jeziki imajo bogato konjugacijo in večina izmed njih, razen bolgarščine in makedonščine, ima bogato sklanjanje samostalnikov.

Beloruščina je jezik, ki ga v Belorusiji približno 7 milijonov ljudi. Spada v skupino vzhodno-slovanskih jezikov in je soroden ruščini in ukrajinščini.

Bolgarščina je jezik, ki ga govori približno 7 milijonov ljudi v Bolgariji. Velja za poseben jezik med slovanskimi jeziki, saj je izgubil sklanjatve samostalnikov. Sorodni jezik bolgarščini je makedonščina.

Bosanščina, hrvaščina, črnogorščina in srbščina (BCS) ; Te jezike govorijo na območju nekdanje Jugoslavije. Sodijo v južno-slovanski narečni kontinuum. V preteklosti je bil za te jezike uporabljen skupni izraz srbo-hrvaščina. So sorodni slovenščini na severo-zahodu ter bolgarščini in makedonščini na jugo-vzhodu. Skupaj imajo ti jeziki okoli 16 milijonov govorcev.

Češčina je zahodno-slovanski jezik, ki ga govori približno 10 milijonov ljudi v Češki republiki. Zaradi zgodovinskih okoliščin obstajata dve varianti jezika: literarna ter pogovorna, med variantama obstajajo velike razlike. Poleg ruščine je češčina slovanski jezik z največ računalniškimi jezikovnimi viri in orodji.

Kašubščina, tudi Cassubian, pomorjanščina, je jezik, ki ga na severu Poljske govori približno 50,000 ljudi. Vsi Kašubi so dvojezični (Poljska). Jezik je soroden slovinskemu jeziku (tudi v Severni Poljski) izumrl je na začetku 20. stoletja.

Makedonščina je južnoslovanski jezik, ki ga uporablja približno 1,5 milijona ljudi v Makedoniji makedonske manjšine v Albaniji, Bolgariji in Egejski Makedoniji (današnja Grčija). Sorodna je bolgarščini.

Poljščina je zahodno-slovanski jezik, ki ga uporablja približno 38 milijonov ljudi na Poljskem in nacionalne manjšine v Belorusiji, Republiki Češki, Litvi in Ukrajini.

Ruščina je vzhodno-slovanski jezik, ki ga uporablja približno 150 milijonov ljudi v Rusiji in nekdanjih sovjetskih republikah. Je slovanski jezik z največ govorcei. Ena od zanimivih lastnosti ruščine je pomanjkanje pomožnega glagola v pretekliku.

Slovaščina je zahodno-slovanski jezik s približno 4,5 milijona govorcei. Je del češko-slovaškega narečnega kontinuuma (Townsend in Janda, 2003) in je soroden s češčino, razlike so predvsem na fonetični ravni.

Slovenščina je južno-slovanski jezik, ki ga govorijo v Sloveniji in narodne manjšine na Koroškem (Avstrija) in v Benečiji (Italija) in na Madžarskem. Uporablja ga približno 1,8 milijona govorcev. Jezik je ohranil nekaj starih značilnosti, kot na primer dvojino.

Spodnje lužiška srbščina je jezik, ki ga govori približno 15 milijonov ljudi v nemški deželi Spodnja Lužica. Kot obrobna narečja, ohranja veliko starih jezikovnih značilnosti, kot so dvojina in jedrnata preteklih časov (aorist in imperfect). Po drugi strani pa je bil pod močnim vplivom okolja, nemškega jezika.

Ukrajnščina je vzhodno-slovanski jezik, ki ga uporablja približno 37 milijonov ljudi v Ukrajini. Podobno kot poljščina, uporablja pasivni pretekli deležnik.

Zgornje lužiška srbščina Zgornje lužiško srbščino govori približno 35.000 ljudi v nemški deželi Zgornja Lužica. Za ta jezik veljajo podobne lastnosti kot za Gornje lužiško srbščino, ki je opisana tem razdelku.

Starocerkvena slovanščina je izumrl jezik, v katerem so bila napisana najstarejša slovanska besedila. Temelji na srednjeveškem narečju makedonske metropole Solun in je bil verski jezik Velike Moravske. Jezik je dobro dokumentiran, obstajajo pisane slovnice ter slovarji. Večina modernejših jezikov, je z leti izgubljala posamezne značilnosti, ki so v tem jeziku ohranjene, tako bi lahko, z določenimi omejitvami, ta jezik uporabili kot univerzalni slovanski jezik za strojno prevajanje.

Polabščina je izumrl zahodno-slovanski jezik, ki je bil uporabljan v severno-vzhodni Nemčiji, natančneje med spodnjo ter srednjo Labo na zahodu ter spodnjo Odro na vzhodu. Izumrl je v 18. stoletju. Polabščina je sorodna kašubščini ter lužiški srbščini.

2.3 Podobnosti slovanskih jezikov pri prevajanju (free rides)

Izkušnje iz področja strojnega prevajanja med sorodnimi jeziki (Homola, 2010) kažejo, da je uporabno razdeliti podobnosti jezikov na kategorije (nivoje) ujemanja. Ločimo tipološko, morfološko, skladenjsko in leksikalno podobnost. V nadaljevanju sledi pregled posameznih kategorij iz vidika strojnega prevajanja.

2.3.1 Tipološka podobnost

Prva kategorija podobnosti je, za namene strojnega prevajanja, najpomembnejša. Če sta jezika prevajanega jezikovnega para iz različnih tipoloških skupin, je prevajanje oteženo. Funkcije, kot so besedni vrstni red, obstoj oziroma ne-obstoj členov, različni sistemi časov in podobne razlike, predstavljajo najhujše prepreke strojnega prevajanja.

Oglejmo si primer slovenščine in makedonščine kot jezikov, ki pripadata isti jezikovni družini, vendar se tipološko razlikujeta. Podoben primer bi lahko predstavili tudi za češčino ali srbščino ter makedonščino. Oba jezika imata bogato pregibanje glagolov ter visoko stopnjo prostosti besednega vrstnega reda, zato ni treba spreminjati vrstnega reda glagolov. Po drugi strani pa je makedonščina praktično ne pozna sklanjanja samostalnikov. Primera 2.1 in 2.3 pomenita približno isto "Moj brat je bral knjigo".

- (2.1) *Moj brat je bral knjigo.*
 PRO,1,M,SG,NOM N,M,SG,NOM VERB,C,PRES,3,SG
bral knjigo.
 VERB,M,PART,PAST,SG,M N,F,SG,ACC.
 “Moj brat je bral knjigo.”
 “My brother read a book.”

- (2.2) Брат ми читаше книга.
 N,M,SG PRO,2,M,SG,DAT VERB,M,PAST,3,SG N,F,SG
 “Brat mi čitaše kniga.”
 “My brother read a book.”

- (2.3) *Knjigo je bral moj brat.*
 N,F,SG,ACC VERB,C,PRES,3,SG VERB,M,PART,PAST,SG,M
moj brat.
 PRO,1,M,SG,NOM N,M,SG,NOM
 “Knjigo je bral moj brat.”
 “The book has been read by my brother.”

- (2.4) Книгата ја читаше брат ми.
 N,F,SG PRO,2,F,SG,ACC VERB,M,PAST,3,SG N,M,SG PRO,1,SG,DAT
 “Knigata ja čitaše brat mi.”
 “The book has been read by my brother.”

Zaradi skoraj prostega besednega vrstnega reda je pomen v obeh primerih enak, v angleščini bi tako tvorili pasivno obliko: *My brother read a book* in *The book has been read by my brother*. V makedonščini ostaja besedni vrstni red nespremenjen glede na slovenščino.

2.3.2 Skladenjska podobnost

Skladenjska podobnost je predvsem pomembna v povezavi z glagoli. Razlike v glagolski valenci negativno vplivajo na kakovost prevoda, saj razlike v valencah glagolov zahtevajo uporabo valenčnih slovarjev izvirnega ter ciljnega jezika v fazi transferja. Izdelava takšnih leksikonov je zapletena in predvsem draga. Razlike v skladenjskih strukturah manjših sestavin, kot so samostalniške in predložne fraze, niso tako pomembne. Analiza takšnih struktur je možna s pomočjo plitve skladenjske analize in sprememba skladenjske strukture ciljne povedi je lokalnega značaja.

Za sorodne jezike po navadi velja, da se besedni vrstni med prevodom ne spreminja. Obstajajo tudi izjeme kot kaže Primeru 2.5, pri prevodih povedi v prihodnjiku med slovenščino, srbsščino in hrvaščino, se besedni vrstni red spremeni.

(2.5)

Jaz se bom oblekel.(SLO)

Ja ću da se obučem.(SR)

Ja ću se obući.(CR)

2.3.3 Morfološka podobnost

Morfološka podobnost pomeni podobno strukturo morfološke hierarhije in paradigem kot na primer podobnosti v sistem sklonov, podobnosti pri spreganju glagolov itd. Slovanski jeziki, z izjemo makedonščine in bolgarščine, si delijo podobne sisteme sklonov ter spreganja. Razlike v morfologiji lahko razmeroma zlahka odpravimo z izkoriščanjem polnih morfoloških modulov za oba jezika jezikovnega para. Podobne morfološki sistemi lajšajo fazo transferja. Na primer, večina slovanskih jezikov, razen bolgarščine in makedonščine, uporablja 6-7 sklonov.

Nekaj problemov povzročajo sintetične forme, ki zahtevajo analitične konstrukte v drugih jezikih. Tak je primer prihodnjika med slovenščino in srbsščino. Primer 2.6 kaže prevod slovenske povedi v prihodnjiku v srbsko poved. Pomožni glagol *biti* v prihodnjiku pri prevodu spremeni lemo v *hteti* ter čas v sedanjik, glavni glagol, v tem primeru kupiti, pri prevodu iz preteklika preide v nedoločnik.

(2.6) *Jutri bom kupil*
 jutri-ADV biti-VERB,C,FUT,1,SG kupiti-VERB,M,PAST,M,SG
kolo.
 kolo-N,NT,SG,NOM
 "Jutri bom kupil kolo." (SLO)

Sutra ću kupiti
 sutra-ADV hteti-VERB,C,PRES,1,SG kupiti-VERB,M,INF
kolo.
 kolo-N,NT,SG,NOM
 "Sutra ću kupiti kolo." (SR)

Razlike, kot je prikazana na Primeru 2.6, rešujemo s pomočjo pravil za plitvi prenos kot je prikazano na Sliki 2.2.

```

<rule>
  <pattern>
    <pattern-item n="vbserfti"/>
    <pattern-item n="vblex"/>
  </pattern>
  <action>
    <let>
      <clip pos="1" side="tl" part="lemh"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="temps"/>
      <lit-tag v="pres"/>
    </let>
    <let>
      <clip pos="2" side="tl" part="temps"/>
      <lit-tag v="inf"/>
    </let>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_vbser"/>
        <clip pos="1" side="tl" part="temps"/>
        <clip pos="1" side="tl" part="persona"/>
        <clip pos="1" side="tl" part="nbr"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_vblex"/>
        <clip pos="2" side="tl" part="temps"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika 2.2: pravilo za prevod dela povedi v prihodnjiku. Pomožni glagol biti v prihodnjiku pri prevodu spremeni lemo v hteti ter čas v sedanjik, glavni glagol pri prevodu iz preteklika preide v nedoločnik.

2.3.4 Leksikalna podobnost

Leksikalna podobnost ne pomeni, da morata imeti besednjaka obeh jezikov skupen izvor, da morajo besede izvirati iz istega korena. Kar je pomembno za strojno prevajanje, je semantično ujemanje besed, po možnosti ena-na-ena, torej za vsako izvorno lemo naj obstaja le ena ciljna lema in obratno.

Leksikalna podobnost je z vidika strojnega prevajanja najmanj pomembna oziroma leksikalne razlike enostavno premoščamo pri prevajanju z uporabo glosarjev ter splošnih slovarjev.

Kljub temu pa lahko obstaja potreba po razširitvi dvojezičnih slovarjev z morfološkimi podatki. Oglejmo si takšen primer na jezikovnem paru slovenščine-srbščina, kjer obstaja nekaj samostalnikov, ki so različnih spolov v obeh jezikih. Primer 2.7 kaže spremembo spola, iz srednjega v moški, pri prevodu besede *okno* v srbsko besedo *prozor*, v obeh jezikih se pridevnik ujema s samostalnikom v spolu.

(2.7) *Odperto* *okno*.
 odperto-ADJ,NT okno-N,NT
 "Odperto okno." (SLO)

Otvoren *prozor*.
 otvoren-ADJ,M prozor-N,M
 "Otvoren prozor." (SR)

To razliko je mogoče popraviti med leksikalnim transferjem. V tej fazi ciljna lema zamenja mesto izvorne leme ter obdrži ostale morfološke oznake. Oznako za spol zamenjamo z ustrezno oznako ciljne leme. Takšen popravek pa povzroči neujemanje spremenjenega samostalnika z okoliškimi besedami, v večini slovanskih jezikov se sosednja samostalnik in pridevnik ujemata v spolu sklonu in številu, v nekaterih primerih tudi v drugih morfoloških kategorijah. Problem rešujemo s pravili za lokalno ujemanje, ki popravljajo porušena lokalna ujemanja v morfoloških kategorijah. pravilo lokalnega ujemanja samostalnika in pridevnika je predstavljeno na Sliki 5.4 v Razdelku 5.3.4

2.3.5 Lažni prijatelji

Lažni prijatelji so podobne besede v različnih jezikih, ki imajo različne pomene. Problem lažnih prijateljev je še posebej izrazen pri sorodnih jezikih, predvsem pri površnih poznavalcih jezikovnih parov. Lažni prijatelji lahko povzročajo težave pri

učenju tujih jezikov, predvsem jezikov sorodnih materinemu jeziku. Lažnih prijateljev je veliko med slovanskimi jeziki. Pri izdelavi sistemov za strojno prevajanje moramo biti pazljivi pri uporabi metod, ki slonijo le na podobnosti besed, za gradnjo dvojezičnih prevajalnih leksikonov, saj le-te pogosto pridelajo pare lažnih prijateljev.

Tabela 2.1: Primeri lažnih prijateljev, podobnih besed v različnih jezikih z različnimi pomeni.

primer	pomen	jezik	primer	pomen	jezik
tanjši	tanjši	SLO	tanji	cenejši	PL
najlepši	najlepši	SLO	nejlepši	najboljši	CS

Poglavje 3

Sistemi za strojno prevajanje

Strojno prevajanje, Machine translation - MT, predstavlja vsako uporabo računalnikov kot pripomočkov za prevajanje besedil iz enega naravnega jezika v drugi (EAMT, 2010). V tem delu bomo obravnavali le strojno prevajanje naravnih jezikov brez uporabnikovega sodelovanja (full-fledged translation of natural languages with no user intervention).

3.1 Razdelitev

Možni sodobni pregled strojnega prevajanja (Sanchez-Martnez et al., 2007) deli področje na dve skupini: prevajanje s pomočjo pravil (Rule-Based - RBMT) ter prevajanje na osnovi korpusov (Corpus-Based - CBMT).

- RBMT obsega sisteme in metode za prevajanje s pomočjo zbirke pravil. Način zapisa pravil se razlikuje med sistemi, veže pa jih skupno dejstvo, da je postavitve takšnega sistema dolgotrajno opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, čeprav se pri gradnji poslužujejo nekaterih manj standardnih prijemov. Primeri sistemov: Systran (Systran, 2010), (Prompt, 2010), (Apertium, 2010).
- CBMT obsega sisteme, ki sledijo naslednjemu vzoru: pripravljena je množica referenčnih prevodov, ki so analizirani in prevedeni v lasten zapis prevajalnega sistema po načelih, ki določajo prevajalni sistem (faza učenja). Ti zapisi služijo kot osnova za poznejše prevode neznanih povedi (faza prevajanja). Sisteme te paradigme delimo na dve večji podskupini: na sisteme statističnega strojnega prevajanja, SMT (Al-Onaizan et al., 1999) in (Burbank et al., 2005) ter na sisteme 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).

- Hibridni sistemi predstavljajo mešanico obeh pristopov, osnova takšnih sistemov sodi v eno od predstavljenih paradig in je oplemeniten z metodami druge paradigme.

Ta delitev je več kot le teoretična, saj kar nekaj sistemov, ki se danes vsakodnevno uporabljajo, sodi v eno od obeh opisanih kategorij. Hibridni sistemi poskušajo z uporabo mešanih prijemov izboljšati kakovost oziroma odpraviti pomanjkljivosti osnovnih sistemov. Začetni prevajalni sistemi so bili postavljeni kot zbirke pravil, saj se je dostopnost elektronskih gradiv povečala šele v zadnjem času. Vseeno pa sistemov na osnovi pravil ne smemo zanemarjati, saj vsebujejo kar nekaj prednosti kot sta natančna sledljivost prevajalnih postopkov in enostavno dopolnjevanje (Forcada, 2006). Sistemi, temelječi na metodah RBMT, dosegajo visoke rezultate tudi na račun visokih stroškov postavitve (Arnold, 2003). Sistemi, temelječi na metodah CBMT, kot so sistemi statističnega strojnega prevajanja, Statistical Machine Translation - SMT (Brown et al., 1993), ter EBMT (Nagao, 1984), omogočajo hitro postavitve prevajalnih sistemov ob predpostavki, da so dosegljivi veliki dvojezični korpusi, kar pa ni vedno res, predvsem za manj uporabljane jezike (Forcada, 2006). Prevajanje poteka na več ravneh, večina avtorjev tako predstavlja prevajalne sisteme kot skupek več modelov (Brown et al., 1993), (Sanchez-Martnez et al., 2007) in (Burbank et al., 2005).

3.1.1 Statistično strojno prevajanje - SMT

Že od nekdaj je poskušal človek opisati jezik s pomočjo pravil, prvi primeri segajo vsaj 2000 let nazaj. Pri opisovanju večine naravnih jezikov s strogimi pravili pa se pojavi kup problemov. Naravni jezik je preveč kompleksna ter živa tvorba in pravila za opisovanje so preveč kompleksna, če jih je sploh mogoče vsa zapisati. Že v začetku tega stoletja so prišli strokovnjaki do tega zaključka, "All grammars leak", (vse gramatike puščajo) (Sapir, 1921).

Natančno določanje pravil jezika, ukleščanje v stroge okvire pravil, ni obrodilo sadov, potrebujemo bolj ohlapne omejitve, ki upoštevajo tudi ustvarjalnost pri uporabi jezika.

Namesto razdeljevanja stavkov po slovničnih pravilih iščemo splošne vzorce, ki se porajajo pri uporabi jezika. Glavno orodje za iskanje takšnih vzorcev je štetje raznovrstnih objektov, bolj strokovno izraženo statistika. Od tod izvira tudi ime statistično strojno prevajanje.

Statistično strojno prevajanje, Statistical Machine Translation - SMT, je osnovano na parametričnih statističnih modelih, ki so naučeni na poravnanih dvojezičnih korpusih (učnih primerih).

3.1.2 Statistično strojno prevajanje z razčlenjevanjem - SMT by parsing

Snovalci sistemov statističnega strojnega prevajanja, Statistical Machine Translation (SMT), vedno pogosteje uporabljajo modele temelječe na drevesnih strukturah in slovnica (Eng et al., 2003), (Koehn et al., 2003), tree structured translation models, zaradi vedno trdnejše zavesti, da lahko večji napredek prinese le globlje razumevanje med modelom in predmetom, ki ga model opisuje.

(Melamed, 2004a) predlaga zmanjšanje konceptualne kompleksnosti prevajalnih modelov, temelječih na drevesih, predlaga tudi novo ime za področje: Statistical Machine Translation by Parsing (SMTbyP). GenPar (Burbank et al., 2005) je popoln sistem za postavitve prevajalnega sistema po načelih: (Melamed, 2004a) in (Melamed, 2004b).

Prvi pogoj za sistem SMTbyP je vzporedni, povedno poravnan in skladenjsko označen dvojezični korpus ter enojezični skladenjsko označen korpus (Melamed, 2004a). Primer skladenjsko označenega korpusa je (Marcus et al., 1993).

Osnovni sistem SMTbyP je sestavljen iz dveh faz: učne ter prevajalne faze.

Prvi, učni del, uporablja sintaktični razpoznavnik (analizator), kot na primer (Collins, 2003), (Charniak, 2000), ki je bil vnaprej naučen na dvojezičnem skladenjsko označenem korpusu (Marcus et al., 1993). Vsaka poved iz izvirnega in ciljnega dela korpusa je razčlenjena; rezultat so pari skladenjsko označenih, soležnih (prevodov) izvornih ter ciljnih povedi. Naslednji korak sestavlja hierarhične poravnave med drevesi izpeljave izvornih ter ciljnih povedi. Model statistične poravnave besed (Brown et al., 1993) ali (Wu, 2005) je uporabljen za modeliranje povezav (prevodov) med besedami v korpusu. Učni podatki so shranjeni za poznejšo uporabo v prevajalnem delu.

V drugem, prevajalnem delu, sistem sestavi skladenjsko drevo vhodne povedi v izvornem jeziku (povedi, ki jo sistem prevaja), izdelava primerno skladenjsko drevo v ciljnem jeziku na osnovi naučenih podatkov iz prve faze ter zamenja izvirne besede s ciljnim besedami s pomočjo modela statistične poravnave besed.

3.1.3 Strojno prevajanje na osnovi primerov - EBMT

Strojno prevajanje na osnovi primerov, Example-based Machine Translation (EBMT), (Nagao, 1984) je pristop h strojnem prevajanju, ki temelji na vzporednih dvojezič-

nih korpusih. V bistvu je prevajanje po analogiji. Sistemi EBMT razbijajo dele besedila na manjše enote ter iščejo že poznane dele za prevod, te dele ponovno združujejo v končen izdelek. Če primerjamo s prevajanjem človeka, naj bi prav ta princip najbližje odražal dejanski način prevajanja pri človeku. Prevajanje ne poteka z globoko jezikoslovno analizo izvornih besedil, besedila se razdelijo na manjše enote do te mere, da so posamezni kosi že poznani (examples), te dele prevedemo po analogiji (že videnem) ter sestavimo končni izdelek.

Sistemi za strojno prevajanje na osnovi primerov pri učenju kodirajo abstrahirano znanje v obliki primerov s spremenljivkami.

3.1.4 Strojno prevajanje na osnovi pravil - RBMT

Strojno prevajanje na osnovi pravil, Rule-Based Machine Translation (RBMT), obsega sisteme in metode za prevajanje s pomočjo zbirke pravil. Način zapisa pravil se razlikuje med sistemi, veže pa jih skupno dejstvo, da je postavitve takšnega sistema dolgotrajno opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, čeprav pri gradnji uporabljajo nekatere manj standardne prijeme. Primeri sistemov: Systran (Systran, 2010), (Promt, 2010), (Apertium, 2010).

Sistemi te paradigme izvirno besedilo najprej morfološko ter skladenjsko analizirajo ter izdelajo predstavitev vhodnega besedila, po navadi v obliki skladenjskega drevesa izpeljave. Ta predstavitev se še dodatno abstrahira s poudarkom na zahtevah strojnega prevajanja. Proces transferja prevede abstraktno predstavitev vhodnega besedila v izvornem jeziku v podobno predstavitev v ciljnem jeziku, to predstavitev sistem uporabi kot osnovo za generacijo besedila v ciljnem jeziku, v bistvu uporabi inverzne metode prvega dela na ciljnem jeziku.

3.1.5 Strojno prevajanje na osnovi pravil plitkega prenosa ter plitke analize - shallow parsing and transfer RBMT

Sistemi strojnega prevajanja s plitvim prenosom, shallow transfer machine translation, v večini primerov uporabljajo enostavno arhitekturo, kjer je analiza izvornega jezika omejena na morfologijo. Sistemi večinoma uporabljajo plitko analizo (Homola in Kuboň, 2008a).

Večina sistemov za prevajanje sorodnih jezikov temelji na strojnem prevajanju s pravili plitke analize, kot je pokazano v (Homola et al., 2009). Metode popolne slovnične analize ne dosegajo dovolj dobrih rezultatov za uporabo v sistemih za prevajanje sorodnih jezikov, njihova stopnja napak je višja kot pa prednosti, ki jih omogoča tak način razpoznavanja izvornih besedil.

3.1.6 Strojno prevajanje sorodnih naravnih jezikov ter nesorodnih jezikov v ozko omejenih domenah

Samodejno prevajanje naravnih jezikov visoke kakovosti, Fully Automatic High Quality Machine Translation, predstavlja hudo prepreko (EAMT, 2010), saj so jeziki spreminjajoče se tvorbe, ki jih težko ukalupljamo. Izdelava modelov, ki dovolj dobro opisujejo prevajanje poljubnih besedil med poljubnimi jeziki zahteva ogromna sredstva ter ogromno časa. Probleme poskušamo omejevati s pomočjo poenostavljanj. Primeri takšnih poenostavitev prevajalnih problemov vključujejo:

- prevajanje s slabimi in nenatančnimi prevodi;
- prevajanje sorodnih jezikov
- prevajanje nesorodnih jezikov v ozko omejeni domeni

Prva možnost zveni kot popolnoma neuporabna, vendar imajo takšni sistemi uporabo, uporabljajo se kot priročno, pogosto edino, orodje za razumevanje gradiva. Njihova izdelava je enostavnejša, obstaja kar nekaj kakovostnih orodij ter metod, ki temeljijo na metodah statističnega strojnega prevajanja (Razdelek 3.1.1), ki omogočajo enostavno postavitve takšnih prevajalnih sistemov kot so orodja opisana v delih (Al-Onaizan et al., 1999) in (Burbank et al., 2005). Kakovost takšnih sistemov je zelo odvisna od velikosti učnih gradiv, velikost poravnanih dvojezičnih korpusov naj bi segala vsaj v desetine milijonov besed (Och, 2006), večina sistemov temelječih na teh metodah je tako namenjena večinoma razumevanju besedila.

Prevajanje sorodnih jezikov predstavlja poenostavitev problema prevajanja tujih si jezikovnih parov oziroma popolnoma prostega prevajanja naravnih jezikov z omejitvijo razlik predvsem v strukturi povedi. Seveda pa podobnost ni omejena le na strukturo povedi, izraža se na vseh jezikovnih ravneh, povzeto po (Toporišič, 2000): glasoslovje, morfologija, besedoslovje, oblikoslovje, skladnja, so možne ravni podobnosti, ki lajšajo gradnjo prevajalnih sistemov.

Prevajanje nesorodnih jezikov v ozko omejeni domeni lahko v marsičem enačimo s prevajanjem sorodnih jezikov, saj ozko omejene domene po navadi prinašajo omejen besedni zaklad, omejen slog pisanja in podobno. Tako lahko mnoge metode, ki so zasnovane za prevajanje sorodnih jezikov, uporabimo tudi v primeru prevajanja nesorodnih jezikov v ozko omejenih domenah.

Ena od metod, ki omogočajo relativno dobre rezultate prevodov sorodnih jezikov je strojnega prevajanja na osnovi pravil plitkega prenosa - shallow transfer RBMT. Metoda ima že dolgo zgodovino in je bila uspešno uporabljena v mnogih prevajalnih sistemih, od katerih je najbolj znan Apertium (Corbi-Bellot et al., 2005).

Večina sistemov za prevajanje sorodnih jezikov temelji na strojnem prevajanju s pravili plitke analize, kot je prikazano v (Homola et al., 2009).

3.2 Strojno prevajanje na osnovi pravil plitkega prenosa ter plitke analize - shallow parsing and transfer RBMT

Sistemi strojnega prevajanja s plitvim prenosom, shallow transfer machine translation, v večini primerov uporabljajo enostavno arhitekturo, kjer je analiza izvirnega jezika omejena na morfologijo. Večina sistemov za prevajanje sorodnih jezikov temelji na strojnem prevajanju s pravili plitke analize, kot je pokazano v (Homola et al., 2009). Metode popolne slovnične analize ne dosegajo dovolj dobrih rezultatov za uporabo v sistemih za prevajanje sorodnih jezikov, njihova stopnja napak je višja kot pa prednosti, ki jih omogoča tak način razpoznavanja izvirnih besedil. Arhitektura, ki jo uporablja večina sistemov za strojno prevajanje naravnih jezikov na osnovi pravil plitkega prenosa ter plitke sinteze je prikazana na Sliki 3.1.



Slika 3.1: Moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Ta arhitektura je bila najprej predstavljena v (Hajič et al., 2000) in pozneje uporabljena tudi v (Corbi-Bellot et al., 2005)

Opis posameznih modulov prevajalnega sistema kot so prikazani na Sliki 3.1:

- Morfološka analiza izvirnega besedila vsaki besedi pripiše vse možne morfološke oznake, ki bi jih ta besedna oblika lahko imela.
- Razdvoumljanje (disambiguation) služi za izbiro najverjetnejše oznake za posamezno besedo glede na njeno okolico.

- Strukturni transfer s pomočjo pravil ter dobesednih prevodov prenese označeno besedilo v ciljni jezik.
- Morfološka sinteza nadomesti morfološko označeno besedilo z dejanskimi besednimi oblikami v ciljnem jeziku.

Moduli so natančneje opisani v Razdelku 3.4.1, kjer so opisani na primeru ogrodja Apertium.

3.3 Orodja za postavitev prevajalnih sistemov

Apertium Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev samodejnega prevajalnega sistema za sorodne jezike. Obširneje je predstavljeno v Razdelku 3.4, saj je bilo uporabljeno pri snovanju in preizkušanju večine metod, predstavljenih v tem delu. Apertium je licenciran pod licenco GNU Lesser General Public License (LGPL) (GNU, 2010).

GenPar GenPar (Burbank et al., 2005), je zbirka orodij za raziskave posplošenega razpoznavanja (generalized parsing), predvsem strojnega prevajanja z razpoznavanjem. Ta zbirka ponuja arhitekturo, načrt ter implementacijo sistema za statistično strojno prevajanje z razpoznavanjem (Statistical Machine Translation by Parsing SMTbyP).

Zbirka je prosto dostopna v okviru licence GPL (GNU, 2010) kar pomeni, da je poleg vseh programov prosto dostopna tudi izvorna koda. Pripravljena so že testna okolja s primeri učnih ter testnih podatkov. Zbirka vsebuje vse programe za takojšnje preizkušanje sistemov ali implementacijo lastnih idej v že pripravljenem ogrodju. GenPar je licenciran pod licenco GNU General Public License (GPL) (GNU, 2010).

Egypt Na poletni delavnici leta 1999 na JHU (John Hopkins University) so po vzoru (Brown et al., 1994) izdelali zbirko orodij, ki omogočajo postavitev popolnega sistema za statistično strojno prevajanje, osnovanega na dvojezičnih vzporednih korpusih. Zbirko so poimenovali Egypt. Pri snovanju delavnice so si zadali pet osnovnih ciljev (vse so dosegli):

1. Postavitev zbirke orodij za statistično strojno prevajanje. Zbirka naj bo dosegljiva vsej raziskovalni srenji. Sestavljena naj bo iz orodij za pripravo korpusov, orodij za dvojezično učenje (postavitev parametričnih modelov) ter orodij za takojšnje dekodiranje besedil.

2. Postavitev češko-angleškega sistema za prevajanje besedil na osnovi izdelanih orodij.
3. Osnovno testiranje sistema na osnovi objektivnih mer (statistično modeliranje težavnosti).
4. Izboljšanje osnovnih rezultatov z uporabo morfoloških in sintaktičnih prevajalnikov.
5. V zadnjih dneh delavnice naj bi postavili prevajalni sistem za nek nov jezik v enem samem dnevu (potrditev enostavnosti uporabe orodij).

Moses Moses (Koehn et al., 2007) je v zadnjem času najbolj uporabljano ogrodje za postavitev sistemov za statistično strojno prevajanja. Glavne lastnosti:

- Dva tipa prevajalnih modelov (translation models): na osnovi fraz, v resnici kosov besedila, (phrase-based) ter na snovi dreves (tree-based).
- Omogoča integracijo, do določene mere, eksplicitnega jezikovnega znanja na nivoju besed.
- Omogoča podporo za integracijo orodij z dvoumnimi izhodi, kot so morfološki analizatorji ter razpoznavalniki govora.
- Podpira velike jezikovne modele.

Moses je licenciran pod licenco GNU Lesser General Public License (LGPL) (GNU, 2010).

3.4 Apertium - odprtokodno ogrodje za prevajalni sistem sorodnih jezikov

Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev samodejnega prevajalnega sistema za sorodne jezike tipa Shallow transfer (Sanchez-Martinez in Ney, 2006). Predstavlja ogrodje, ki s pomočjo pravil omogoča prevajanje med sorodnimi jeziki. Uvršča se med sisteme za samodejno prevajanje naravnih jezikov na osnovi pravil plitkega transferja (shallow-transfer RBMT). Prevajanje je razdeljeno na pet osnovnih faz:

- označevanje ne-prevajanih razdelkov

- leksikalni transfer
- odpravljanje dvoumnosti (disambiguation)
- strukturni
- dejanski prevod posameznih besed ter besednih fraz

Zadnja faza predstavlja odpravljanje pomanjkljivosti ostalih faz in predstavlja niz pravil, ki odpravijo manjše napake pri prevajanju posebnosti.

Pravila, ki omogočajo prevajanje, temeljijo na regularnih jezikih, ki jih je enostavno pretvoriti v končne avtomate (transduktorje končnih stanj - finite state transducers) ter na besednih in fraznih slovarjih (enojezičnih ter večjezičnih). Prevajanje med sorodnimi jeziki izvira prav iz načina pravil, ki ne omogočajo dobrega opisa poljubnih prevajalnih konstruktov. Več o tem je opisano v Razdelkih 5.1 in 2.1.3.

Za leksikalni prenos so opisi s pomočjo regularnih jezikov oziroma iz teh izvedeni stohastični regularni modeli Skriti Markovski Model, Hidden Markov Model - HMM, predstavljen v (Welch, 2003), ali uteženi končni transduktorji, Weighed Finite State Transducers - WFST definirani v (Kornai, 1999) in (Roche in Schabes, 1997), dovolj močno orodje (Melamed, 2004a).

Strukturni nivo je kompleksnejši, nekaterih konstruktov oziroma njihovih prevodov ne moremo opisati z regularnimi jeziki. Mnogi avtorji (Melamed, 2004b), (Eng et al., 2003), (Koehn et al., 2003) predlagajo uporabo modelov, ki temeljijo vsaj na drevesnih strukturah. Tako je Apertium namenjen predvsem za sorodne jezike.

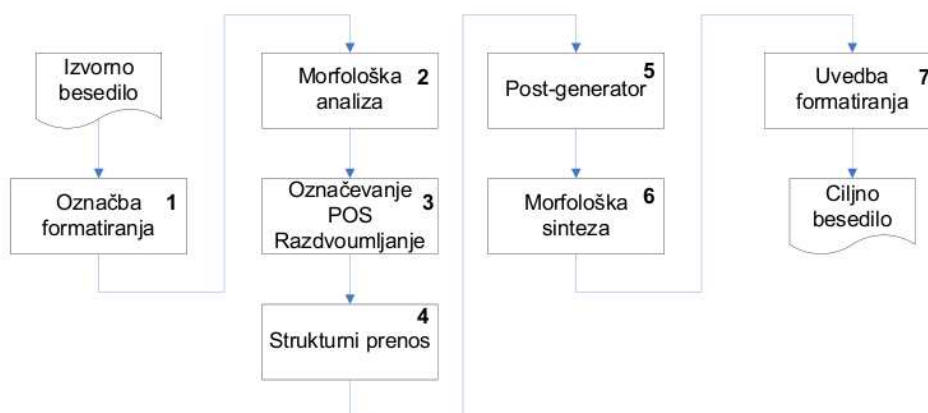
Apertium je licenciran pod licenco GNU Lesser General Public License (LGPL) (GNU, 2010).

3.4.1 Arhitektura ogrodja Apertium

Arhitektura ogrodja Apertium posnema arhitekturo predstavljeno na Sliki 3.1, ki predstavlja tipično razporeditev modulov sistemov za strojno prevajanje s plitkim prenosom.

Slika 3.2 predstavlja arhitekturo ogrodja Apertium.

Morfološka analiza izvornega besedila poišče vse možne oznake za posamezne besede. Disambiguacija izbere najverjetnejše oznake za posamezne besede glede na njeno okolico, strukturni transfer služi dejanskemu prenosu izvornega besedila v ciljni jezik, morfološka sinteza poišče ustrezne besede v ciljnem jeziku glede na prevedene morfološko označene dele besedila. Modul post-generator služi za odpravljanje napak, uvajanje posebnosti ciljnega jezika ter za združevanje besednih zvez. Sledi ponovna uvedba označevanja. Moduli so natančneje predstavljeni v Razdelku 3.4.2.



Slika 3.2: Arhitektura ogrodja Apertium: poleg osnovnih modulov, ki služijo za osnovno prevajanje in so prikazani na Sliki 3.1, Apertium dodaja še module za označevanje delov besedila, ki se ne prevajajo ter modul za končno urejanje (post-editing) prevodov.

Sledijo opisi posameznih modulov prevajalnega sistema kot so prikazani na Sliki 3.2, moduli so označeni s števili

Modul 1, označba formatiranja (De-formatter) v izvornem besedilu posebej označi dele besedila, ki jih ostali prevajalni moduli ignorirajo. Tako lahko sistem prevaja tudi besedilo z urejevalnimi oznakami kot so oznake jezikov HTML ali XML. Za posebej prirejene sisteme lahko modul označuje tudi dele besedila, ki se ne prevajajo.

Modul 2, morfološka analiza (Morphological analyzer) vsaki besedi izvornega besedila pripiše vse možne morfološke oznake, ki bi jih ta besedna oblika lahko imela. Modul podpira besede in besedne zveze, ki so zapisane v morfološko označenem enojezičnem slovarju izvornega jezika. Vsaka beseda oziroma besedna zveza je obdelana samostojno, brez vpliva okolice. Morfološka analiza vsaki besedi pripiše vse njene možne oznake, kar pomeni, da je izhod tega modula dvoumen. Primer delovanja je prikazan na Sliki 3.3.

Modul 3: razdvoumljanje s označevanjem POS (POS tagger) služi za izbiro najverjetnejše oznake za posamezno besedo glede na njeno okolico. Sistem uporablja stohastični označevalnik morfoloških oznak, ki izbere najverjetnejšo izbiro

```

Danes je lepo vreme
Danes
    Danes<adv>
je
    biti<vbser><pres><p3><sg>
    jesti<vblex><pres><p3><sg>
    prpers<prn><subj><p3><f><sg><gen>
lepo
    lep<adv>
    lep<adj><f><sg><acc><pos>
    lep<adj><f><sg><ins><pos>
    lep<adj><nt><sg><acc><pos>
    lep<adj><nt><sg><nom><pos>
vreme
    vreme<n><nt><sg><acc>
    vreme<n><nt><sg><nom>
.

```

Slika 3.3: Morfološka analiza stavka "Danes je lepo vreme.". Besede izvirne povedi so označene z vsemi možnimi ustreznimi morfološkimi oznakami iz slovarja. Najprej je zapisana besedna oblika, sledijo vse možne oznake za to besedno obliko. Za besedno obliko *lepo* je možnih pet različnih množic oznak.

izmed ponujenih izbir prejšnjega modula, modula za morfološko analizo. Osnovni označevalnik (Sánchez-Martínez et al., 2008) temelji na samodejni metodi učenja označevanja oznak POS na neoznačenem korpusu. Po trditvah avtorjev je primeren za vse evropske jezike, kakovost označevanja pa je le delno primerljiva z najboljšimi označevalci oziroma z označevalci naučenimi na označenih in pregledanih korpusih. Rezultat tega modula je po ena izbira morfološkega označevalca za vsako besedo izvirnega besedila.

Modul 4: strukturni transfer-prenos (Structural transfer) s pomočjo pravil ter dobresednih prevodov prenese označeno besedilo v ciljni jezik. Pravila strukturnega transferja temeljijo na regularnih izrazih in se osredotočajo na ujemanja med morfološkimi oznakami sosednjih besed, na lokalni besedni vrstni red ter na ostale razlike jezikovnega para, ki jih lahko opišemo z lokalnim kontekstom. Obširneje so pravila predstavljena v Poglavju 5.

Dobesedni prevodi posameznih besed v lematizirani obliki temeljijo na osnovi dvojezičnega slovarja.

Modul 5: končno urejanje (Post-generator) služi za odpravljanje sistemskih napak prejšnjih modulov, uvajanje posebnosti ciljnega jezika (uporaba diakritikov, posebnosti pri imenih, ...) ter za združevanje besednih zvez.

Modul 6: morfološka sinteza (Morphological analyzer) nadomesti morfološko označeno besedilo, ki je že prevedeno v ciljni jezik, z dejanskimi besednimi oblikami v ciljnem jeziku. Sam modul uporablja enake podatkovne strukture (enojezični morfološko označen slovar) kot Modul 2, modul za morfološko analizo, zamenjana je le smer uporabe.

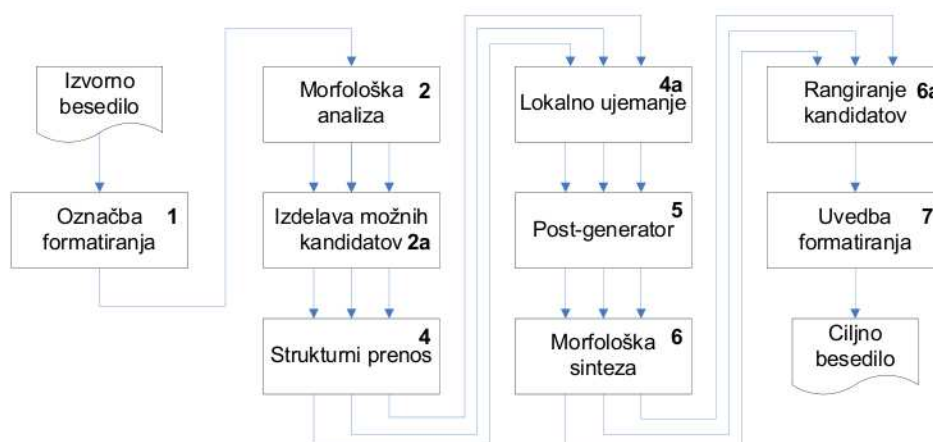
Modul 7: ponovna uvedba formatiranja (Re-formatter) odseke besedila, ki jih je začetni modul izbral in označil kot besedilo, ki se ne prevaja, ponavadi so to oznake urejanja besedila (HTML, XML in ostale), zadnji modul prevajalne verige ponovno postavi v besedilo.

3.4.2 Predlagana spremenjena arhitektura

Rezultati predstavljeni v (Homola in Kuboň, 2008b) kažejo, da sprememba arhitekture brez uporabe označevalnika POS v začetnih fazah prevajanja ter z uvedbo stohastičnega razvrščevalnika prevodov (stochastic ranker) na koncu prevajalne verige, prinaša izboljšano kakovost prevodov v primerjavi z osnovno arhitekturo predstavljeno na Sliki 3.1 ter v (Corbi-Bellot et al., 2005). Odločili smo se za izbiro spremenjene arhitekture, kot je prikazana na Sliki 3.4.

Opis dodatnih modulov prevajalnega sistema kot so prikazani na Sliki 3.4. Opisani so le novi moduli, originalni moduli so opisani v Razdelku 3.4. Moduli so označeni s števili:

Modul 2a: izdelava možnih kandidatov (Multiple candidate selector) izdelava vse možne kandidate za prevode na osnovi dvoumnega označevanja morfološkega analizatorja. Ta modul predstavlja zamenjavo za modul za razdvoumljanje, saj namesto izbire najverjetnejšega kandidata za prevode izdelava vse možne kandidate, kombinatorično eksplozijo števila možnih kandidotov nadzoruje s pravili za ujemanje med morfološkimi oznakami sosednjih besed izvirnega jezika. Izhod modula je množica kandidatov za prevode.



Slika 3.4: Moduli predlaganega (spremenjenega) sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Arhitektura temelji na sistemu predlaganem v (Corbi-Bellot et al., 2005; Hajič et al., 2003) brez uporabe sistema za razdvajanje na osnovi označevalnika POS in uporabo vseh kandidatov za prevode do zadnjih faz prevajalne verige ter z dodatkom modula za izbiro najboljšega prevoda (Ranker).

Modul 4a: lokalno ujemanje (Local agreement) modul skrbi za ujemanje bližnjih besed v morfoloških oznakah. Pravila tega modula odpravljajo napake, ki jih lahko povzročijo pravila za strukturalni transfer, še posebej samodejno zgrajena pravila. Primer 3.1 kaže takšno ujemanje samostalnika in pridevnika. Rdeče drevo, *crveno drvo*, pridevnik *crveno* se ujema s samostalnikom *drvo* v spolu, številu ter sklonu.

(3.1) *crven* *drvo*
 ADJ,NT,SG,ACC,POS N,NT,SG,ACC,POS
 "crveno drvo"

Modul 6a: rangiranje kandidatov (Stochastic ranker) izbere najverjetnejši prevod iz množice prevedenih kandidatov za prevode. Osnova za ugotavljanje kakovosti prevoda je statistični jezikovni model ciljnega jezika.

Najpomembnejši razlogi za izbiro spremenjene arhitekture so:

- Izdelava označevalnika POS, še posebej dobre kakovosti, ni enostavna naloga. Poseben problem predstavljajo morfološko bogati jeziki, mednje sodi

tudi slovenščine. Eden izmed najlažjih načinov je učenje stohastičnih označevalcev, ki temelji na algoritmu HMM Welch (2003). Nekatere dele te naloge lahko avtomatiziramo z uporabo nenadzorovanih ali delno nadzorovanih učnih metod, kot je (Brants, 2000), vendar še vedno ostaja dovolj dela z izbiro novega niza oznak, izdelavo označenega učnega korpusa, preskušanjem korpusa in na koncu izvedbe samega učnega procesa.

- Stopnja kakovosti označevanja današnjih najboljših, state-of-the-art, označevalnikov POS za visoko pregibne jezike, kot sta (Hajič, 2000) in (Erjavec, 2006) je relativno nizka, v primerjavi s kakovostjo označevalnikov POS za analitične jezike, kot je angleški jezik, in tudi v primerjavi s splošno kakovostjo prevajalnih sistemov za sorodne jezike.
- Po arhitekturi predstavljeni na Sliki 3.1 modul za morfološko razdvoumljanje sledi modulu za morfološko analizo izvirnega jezika. Napake stohastičnih metod modula za razdvoumljanje, ki uporablja označevalnik POS, so tako povzročene v zgodnjih fazah prevoda in povzročajo več težav kot napake poznejših faz prevajalnega procesa.
- Več kandidatov za prevode omogoča izbor najboljših kandidatov v zaključni fazi, ko so zbrani vsi razpoložljivi podatki za prevod.

3.4.3 Pregled uporabljenih podatkovnih tipov

Spisek podatkovnih tipov, ki jih potrebujemo za vse module spremenjenega, prevajalnega sistema (novi podatkovni tipi, ki jih zahteva spremenjena arhitektura so označeni z zvezdico - *):

Enojezični slovar izvirnega jezika z morfološkimi oznakami se uporablja za morfološko analizo izvirnega besedila. Enojezični slovar izvirnega jezika je uporabljen pri plitvi analizi (morfološki analizi) besedil v modulu za morfološko analizo (morphological analyser module), označenem s številko 2 na Sliki 3.2 in na Sliki 3.4.

Enojezični slovar ciljnega jezika z morfološkimi oznakami se uporablja za morfološko sintezo ciljnega besedila. Enojezični slovar ciljnega jezika je uporabljen pri sintezi prevedenega besedila v ciljni jezik v modulu za morfološko sintezo (morphological generator module), označenem s številko 5 na Sliki 3.2 in na Sliki 3.4. Ta dva modula uporabljata iste slovarje, kar omogoča hitro postavitev dvosmernega prevajalnega sistema.

Dvojezični prevajalni slovar se uporablja za dobesedno prevajanje v lematizirani obliki ter prevajanje fraz v lematizirani obliki. Modul, ki uporablja dvojezični slovar in pravila plitkega prenosa je modul za strukturni transfer (structural transfer module), označen s številko 4 na Sliki 3.2 na Sliki 3.4.

Pravila plitkega prenosa na osnovi regularnih izrazov se uporabljajo pravila plitkega prenosa so uporabljena za opis morfoloških ter skladenjskih razlik lokalnega obsega med jezikoma prevajanega jezikovnega para kot so lokalno ujemanje besed v morfoloških kategorijah ter lokalni vrstni red besed. Natančneje so pravila predstavljena v Poglavju 5.

Pravila na osnovi regularnih izrazov za končno urejanje (post-editing) uporablja modul za končno urejanje (post-generator), označen s številko 5 na Sliki 3.4. Pravila služijo za odpravljanje napak, uvajanje posebnosti ciljnega jezika ter za združevanje besednih zvez.

* **Pravila na osnovi regularnih izrazov za izražanje lokalnega ujemanja kategorij izvirnega jezika** uporablja modul za izbiro množice kandidatov za prevod (multiple candidate selector), označen s številko 2a na Sliki 3.4. Modul uporablja isto tehnologijo kot modul za strukturni (Structural module) s pravili za lokalno ujemanje morfoloških kategorij, ki so bila naučena na izvirnem jeziku. To metodo modul uporablja kot heuristiko za omejevanje eksplozije števila možnih hipotez za prevode dvoumne morfološke analize izvirnega besedila.

* **Pravila na osnovi regularnih izrazov za izražanje lokalnega ujemanja kategorij ciljnega jezika** so uporabljena v modulu za iskanje lokalnega ujemanja (local agreement module), označen s številko 4a na Sliki 3.4. Uporabljajo se za odpravo napak, ki jih povzročijo pravila plitkega prenosa v fazi transferja. Delovanje tega modula je v osnovi enako delovanju modula za strukturni transfer, drugačna so le pravila, ki popravljajo napake lokalnega ujemanja morfoloških kategorij, delovanje modula pa je preseljeno na izhod modula za strukturni transfer.

* **Statistični jezikovni model ciljnega jezika**¹ je uporabljan pri modulu za rangiranje prevodov (Ranker module), označen s številko 6a na Sliki 3.4. Ta modul

¹Uporabljen v modulu za stohastično rangiranje končnih prevodov, ki predstavlja razširitev osnovne arhitekture po (Hajič et al., 2003) ter je bil opisan v (Homola in Kuboň, 2008b)

izbere najboljši prevod iz seznama možnih kandidatov za prevode, ki so jih sestavili prejšnji moduli, modul deluje na principu stohastičnega jezikovnega modela ciljnega jezika, torej izbira prevode, ki so najbolj verjetni v ciljnem jeziku.

* **Jezikovni model morfoloških oznak (POS) izvirnega jezika**² uporablja modul za izbiro množice kandidatov za prevod (multiple candidate selector), označen s številko 2a na Sliki 3.4. Poleg predstavljene heuristike, uporablja še enako tehnologijo kot modul za rangiranje prevodov, vendar s pomočjo jezikovnega modela, naučenega na morfoloških oznakah korpusa v izvirnem jeziku. Metoda je natančneje predstavljena v (Homola et al., 2009).

Za samodejno pridobivanje podatkov za vsak tip predstavljenih podatkov z zgornjega spiska je bila izdelana nova metoda oziroma je bila uporabljena že znana metoda za samodejno izdelavo podatkov tega tipa. Metode so natančneje predstavljene v Razdelku 4.4.

Več popolnoma delujočih sistemov smo izdelali s pomočjo opisanih metod. Kakovost prevodov je predstavljena v Poglavju 6.

²Uporabljen v modulu za izbiro množice kandidatov za prevod (Multiple candidates selector module), ki predstavlja razširitev osnovne arhitekture po (Hajič et al., 2003) ter je bil opisan v (Homola et al., 2009)

Poglavje 4

Morfologija in leksikoni

Sistemi za strojno prevajanje na osnovi pravil plitkega transferja enostavne metode analize, transferja ter sinteze pri izdelavi prevodov iz izvirnega v ciljni jezik. Sama analiza izvirnih povedi temelji na morfološkem označevanju, ki se izvaja s pomočjo morfološko označenih enojezičnih slovarjev. Transfer uporablja pravila plitkega transferja ter dvojezični slovar izvirnega in ciljnega jezika. Pri analizi je uporabljen enojezični slovar ciljnega jezika, ki je sestavljen na enak način kot slovar uporabljen pri analizi, le uporablja se ga za sintezo besed in ne označevanje izvirnih besed.

4.1 Enojezični morfološko označen slovar

Morfološko označen slovar, ki ga uporablja Apertium, vendar bi z majhnimi spremembami takšne slovarje uporabljali tudi drugi prevajalni sistemi, temelji na lemah, ki so zbrane v paradigmah. Posamezna paradigma združuje vse leme, ki se spreminjajo po istih pravilih glede na morfološke oznake. Slika 4.1 kaže primere lem ter njihovo članstvo v paradigmah. Lema je predstavljena s svojim imenom (ime leme), krnom (najdaljšim delom, ki je skupen vsem besednim oblikam leme ter z imenom paradigme, kjer so opisana vsa pravila sprememb glede na morfološke kategorije). Posamezen zapis v slovarju je predstavljen z XML oznako *e*, atribut te oznake *lm* predstavlja ime leme, gnezdena oznaka *i* predstavlja krn besede, oznaka *par* pa ime paradigme.

Primer za lemo *cerkev* je predstavljen na Sliki 4.1.

Posamezna gesla enojezičnega slovarja so združena v morfološke paradigme, kot so definirane v (Spencer, 1991). Morfološke paradigme predstavljajo razrede lem, ki se na isti način spreminjajo (glede na vse možne besedne oblike). Z drugimi besedami, vsebujejo vse leme, katerih vse besedne oblike se spreminjajo na enak način za vse morfološke oznake (Erjavec, 2004).

```

<e lm="cepljen"><i>cepljen</i><par n="žveplen/___adj"/></e>
<e lm="cepljenje"><i>cepljenj</i><par n="žvižganj/e___n"/></e>
<e lm="ceremonija"><i>ceremonij</i><par n="žog/a___n"/></e>
<e lm="cerkev"><i>cerk</i><par n="cerk/ev___n"/></e>

```

lema: cerkev

krn: cerk

paradigma: cerk/ev___n

Slika 4.1: Del zapisov v enojezičnem slovarju. Lema *cerkev* je predstavljena z lemo, krnom ter paradigmo.

Slika 4.2 kaže primer paradigme za ženski samostalni in v slovenščini.

Uporaba paradigme omogoča uporabo kompaktnejšega zapisa podatkov, kot je prikazano na primeru, predstavljenem na Sliki 4.2 za lemo *cerkev* slovenskem jeziku: vsi samostalniki prve ženske sklanjatve paradigme *-ev*, kot so *cerkev*, *breškev*, *podkev*, se sklanjajo po istem vzorcu in jih združimo v isto paradigmo. Tako lahko enostavno pravilo določa spremembo besede iz imenovalnika v rodilnik s spremembo končnice iz *cerkev* v *cerkve*, torej *-ev* → *-ve*. Eno pravilo tako zadošča za celo skupino besed in ne le za en osamljen primer.

Pri indo-evropskih jezikih, ki večinoma uporabljajo konkatativno morfologijo¹, besedne oblike določajo menjave pripon ter včasih predpon. V to družino spada večina evropskih jezikov. Primer iz češčine: pridevnik *sladký* (sladek) lahko spremenimo v *nej-slad-ší-ho* (najslajši - moški ali srednji spol imenovalnik ali tožilnik), z dodajanjem pripone *nej-*, ki predstavlja presežnik, in z menjavo pripone *-ký* (komparativ) s pripono *-ší* in z dodajanjem pripone *-ho* moški ali srednji spol imenovalnik ali tožilnik).

V Tabeli 4.1 je predstavljen primer iz slovenščine za lemo *mesto*, ta lema vsebuje 18 besednih oblik, za tri števila in 6 sklonov.

4.2 Dvojezični slovar

Dvojezični slovarji temeljijo na parih izvorna lema - ciljna lema, torej na dobese-dnih prevodih lem. Primeri dvojezičnih prevodov lem iz slovenščine v srbsčino so predstavljeni na Sliki 4.3.

¹besede so sestavljene iz več zlepljenih (concatenated) morfemov. Morfemi vključujejo krne ter

```

<pardef n="cerk/ev__n">
  <e>
    <p>
      <l>
        ev
      </l>
      <r>
        ve
        <s n="n"/><s n="f"/><s n="sg"/><s n="gen"/>
      </r>
    </p>
  </e>
  <e>
    <p>
      <l>
        ev
      </l>
      <r>
        ev
        <s n="n"/><s n="f"/><s n="pl"/><s n="gen"/>
      </r>
    </p>
  </e>
  ...
</pardef>

```

Slika 4.2: Del paradigme za samostalnike ženskega spola v slovenščini. Tipični predstavnik je lema *cerkev*. Končnica *-ev* se spreminja v skladu z različnimi MSD.

Poleg samega prenosa iz izvirnega v ciljni jezik, lahko opišemo še prenos morfoloških oblik, ki se spremenijo pri samem prevodu. Primer 4.1 kaže prevod slovenske besede *okno* v srbsko besedo *prozor*, kjer se spremeni tudi spol iz srednjega v moški.

(4.1) *okno prozor*
 N,NT N,M
 “okno”
 “prozor”

predpone in pripone

Tabela 4.1: Vse besedne oblike za slovensko lemo mesto

besedna oblika	število	sklon
mest-o	ednina	imenovalnik
mest-a	ednina	rodilnik
mest-u	ednina	dajalnik
mest-o	ednina	tožilnik
mest-u	ednina	mestnik
mest-om	ednina	orodnik
mest-a	množina	imenovalnik
mest-	množina	rodilnik
mest-om	množina	dajalnik
mest-a	množina	tožilnik
mest-ih	množina	mestnik
mest-i	množina	orodnik
mest-i	dvojina	imenovalnik
mest-	dvojina	rodilnik
mest-oma	dvojina	dajalnik
mest-i	dvojina	tožilnik
mest-ih	dvojina	mestnik
mest-oma	dvojina	orodnik

```

<e><p>
  <l>okno<s n="n"/><s n="nt"/></l>
  <r>prozor<s n="n"/><s n="m"/></r>
</p></e>
<e><p>
  <l>okolica<s n="n"/><s n="f"/></l>
  <r>okolina<s n="n"/><s n="f"/></r>
</p></e>
<e><p>
  <l>okoli<s n="adp"/></l>
  <r>oko<s n="adp"/></r>
</p></e>
<e><p>
  <l>okolishčina<s n="n"/><s n="f"/></l>
  <r>prilika<s n="n"/><s n="f"/></r>
</p></e>

```

Slika 4.3: Primeri dvojezičnih prevodov lem iz slovenščine v srbščino.

4.3 Uporabljena učna gradiva

Pri večini metod, predstavljenih v Razdelku 4.4, je bil kot učna množica uporabljen večjezični poravnani korpus MULTEXT-EAST (Erjavec, 2010) in (Dimitrova et al., 1998).

MULTEXT-East² je večjezična zbirka jezikovnih gradiv, zapisana je v standardizirani obliki ter podpira velik del centralno-evropskih ter vzhodno-evropskih jezikov. Uporablja morfosintaktične označbe po vzoru EAGLES, (Leech in Wilson, 1996). Korpusni del gradiv je zapisan v standardizirani obliki v formatu XML (Bray et al., 2008), po smernicah TEI-P4 (Consortium, 2007). Gradiva sestavljajo morfosintaktične specifikacije, morfosintaktični leksikoni ter označeni, vzporedni, primerjalni ter govorni korpusi. Trenutna različica gradiv obsega 16 jezikov in je prosto dostopna za raziskovalne namene.

Primer povedi iz korpusa je na Sliki 4.4; vsaka poved je shranjena v znački *s*, atribut *id* služi za povezavo z drugimi jeziki. Vsaka beseda je shranjena v znački *w*, atribut *lemma* predstavlja lemo nesede, atribut *ana* pa morfosintaktično oznako besede. Značka *c* označuje ločila.

```
<s id="Osl.2.3.5.11">
  <w lemma="priti" ana="Vmpps-dma">Prišla</w>
  <w lemma="biti" ana="Vcip3d--n">sta</w>
  <w lemma="do" ana="Spsg">do</w>
  <w lemma="podrt" ana="Afpnsg">podrtega</w>
  <w lemma="drevo" ana="Ncnsg">drevesa</w>
  <c>,</c>
  <w lemma="o" ana="Spsl">o</w>
  <w lemma="kateri" ana="Pr-nsl----a">katerem</w>
  <w lemma="on" ana="Pp3msd--y-n">mu</w>
  <w lemma="biti" ana="Vcip3s--n">je</w>
  <w lemma="praviti" ana="Vmpps-sfa">pravila\index{pravilo}</w>
  <c>.</c>
</s>
```

Slika 4.4: Označena poved v korpusu (Erjavec, 2010).

Vzporedni del korpusa, v tem delu je najbolj uporabljan, sestavlja roman Georga Orwella "1984" (Orwell, 1949), preveden v vseh 16 jezikov. Celoten roman je morfosintaktično označen ter vzporedno poravnani na nivoju povedi s pivotnim jezikom - angleščino. Vsi prevodi so poravnani z angleškim izvirnikom.

²Korpus je dostopen na naslovu: <http://nl.ijs.si/ME/V4/>.

4.4 Metode

Vsak modul s Slike 3.4 je sestavljen iz osnovne programske opreme ter jezikovno-odvisnih podatkov. Podatki so v ogrodju Apertium strukturirani v človeku berljivem formatu XML (Bray et al., 2008). Naslednji razdelki predstavljajo opis metod za samodejno izdelavo podatkov iz Razdelka 3.4.3.

4.4.1 Izdelava enojezičnih slovarjev izvirnega in ciljnega jezika z morfološkimi oznakami

Iz morfološko označenega ter lematiziranega korpusa najprej izluščimo vse besedne oblike ter jih razdelimo po lemah. Uporabili smo korpus (Erjavec, 2010), primer označene povedi iz tega korpusa je na Sliki 4.4. Leme družimo v paradigme, kar nam omogoča sestavljanje manjkajočih besednih oblik, izdelava paradigem je razložena v Razdelku 4.4.1.1.

4.4.1.1 Izdelava paradigem

Leme z enakimi spremembami družimo v paradigme, vsaka ima naslednje elemente:

- tipična lema; iz te leme izpeljemo začetno paradigmo;
- krn; najdaljši skupni del vseh besednih oblik v lemi;
- množica vseh besednih oblik razdeljenih na krn ter obrazila, k vsaki besedni obliki je zapisan me morfo-sintaktična oznaka (MSD) po (Erjavec, 2004).

Primer paradigme je prikazan na Sliki 4.5.

Označene leksikone, zbirke besednih oblik s pripisanimi leмами ter MSD označbami izvlečemo iz označenega korpusa, kot je (Erjavec, 2010). Paradigme izdelamo z algoritmom, ki je zapisan na Sliki 4.6.

Vse besedne oblike za vsako lemo združimo v razred, ki predstavlja to lemo. Za vsak razred izdelamo paradigmo, ki vsebuje na začetku le zapise ene leme. Sledi združevanje paradigem po algoritmu s Slike 4.6: dve paradigmi združimo v eno paradigmo, če pripadata isti besedni vrsti (prva kategorija oznak MSD) in če se noben par vsi zapisi ne izključuje. Dva zapisa se izključujeta, če imata enake morfološke oznake in različna obrazila kot kaže primer na Sliki 4.7. Torej dve paradigmi družimo, če ena predstavlja popolno podmnožico druge paradigme. Vsaka ima za beleženo celotno listo vseh lem, ki jo sestavljajo.

```

lema: cerkev
krn: cerk
primeri besednih oblik:
besedna oblika: cerkev
pripona: ev
MSD: noun+feminine+singular+nominative
(samostalnik+ženski spol+ednina+imenovalnik)
besedna oblika: cerkvah
pripona: vah
MSD: noun+feminine+singular+locative
(samostalnik+ženski spol+ednina+mestnik)

```

Slika 4.5: Del paradigme *cerk-ev*. Lema: cerkev, krn: cerk, dve besedni obliki *cerkev* in *cerkvah*

```

for vsaka do
  for vse preostale paradigme do
    if vse besedne oblike z istimi morfološkimi oznakami imajo
    enake pripone then
      združi paradigmi primerjanega para
    end if
  end for
end for

```

Slika 4.6: Algoritem za gradnjo paradigem

Enojezični morfološko označeni slovarji izvirnega in ciljnega jezika so bili zgrajeni s pomočjo paradigem, leme z manjkajočimi besednimi oblikami v originalnih leksikonih so bile tako dopolnjene, velikost končnega leksikona je bila približno 20 krat večja od začetnega.

```

lema: cerkev
krn: cerk
besedna oblika: cerkev
pripona: ev
MSD: noun+feminine+singular+nominative
(samostalnik+ženski spol+ednina+imenovalnik)
lema: ana
krn: an
besedna oblika: ana
pripona: a
MSD: noun+feminine+singular+nominative
(samostalnik+ženski spol+ednina+imenovalnik)
ev != a

```

Slika 4.7: Besedni obliki se ne ujemata, kar pomeni, da paradigmi ne združimo.

4.4.2 Izdelava dvojezičnih prevajalnih slovarjev

Poseben problem pri uporabi stohastičnih modelov je v redkih podatkih (sparse data problem). Osnovni korpus ima določeno število dovolj dobro opisanih pravil ter dovolj pogosto zastopanih besed, vsebuje pa tudi velik odstotek slabo predstavljenih besed ter pravil. Z večanjem korpusa uvajamo tudi nove besede. Tako se odstotek slabo opisanih besed ter pravil ne manjša z večanjem korpusa. Problem pomanjkljivih podatkov rešujemo s pomočjo naprednih algoritmov, ki jih poskušajo zakriti. Algoritmi upoštevajo predhodno znanje o problemu, izkušnje iz sorodnih področij ali pa celo povsem tujih področij. Šumne podatke izločamo s pomočjo zakonitosti v podatkih, z izločanjem ekstremov. Paziti moramo, da pri izločanju napačnih podatkov ne pretiravamo in korpusa preveč ne "porežemo", poenostavimo.

Izdelali smo metodo poravnave lem. Model za poravnavo besed na osnovi verjetnosti (SMT word-to-word model) (Brown et al., 1993), je bil naučen na vzporednem, povedno poravnanem seznamu, ki je bil izluščen iz korpusa (Erjavec, 2010). Uporabili smo orodje GIZA++ (Och in Ney, 2003). Izvleček seznama je prikazan na Sliki 4.8. V korpusu je vsaki besedi dopisana še njena morfološko-skladenjska oznaka (MSD) ter lema. Vzporedno povedno poravnan seznam je bil izdelan tako, da so

bile za vsako besedo izluščene iz korpusa samo leme in besedne vrste, ki so prva kategorija morfološke oznake. Slika 4.8 kaže izvleček pripravljenih učnih podatkov.

```
p riti_V bit i_V do_S podrt_A drevo_N ,
o_S kater i_P on_P bit i_V praviti_V .
```

Slika 4.8: Pripravljeni učni podatki: leme in besedne vrste za vsako besedo v korpusu.

Besede v enojezičnih slovarjih so zapisane v lematizirani obliki, besedne oblike pa so zabeležene v paradigmah, natančneje je razloženo v Razdelku 4.1. Metoda poravnave lem omogoča boljše rezultate v primerjavi s poravnavo besed v korpusu, zaradi zmanjšanja prostora iskanja. Dvojezični prevajalni slovar je sestavljen iz parov *izvorna lema* - *ciljna lema*, ki omogočajo prevajanje v ciljni jezik. Metoda poravnave lem omogoča boljše rezultate v primerjavi s poravnavo besed v korpusu, zaradi zmanjšanja prostora iskanja, kot je opisano v Enačbi 4.2 in na Sliki 4.9.

Število besednih oblik v besedilu je veliko večje od števila lem, še posebej za visoko pregibne jezike kot so slovanski jeziki. Slika 4.2 kaže razliko v številu besednih oblik za isti korpus (Erjavec, 2010) v petih jezikih, treh visoko pregibnih slovanskih jezikih: slovenščine, srbsščina in češčina ter v angleškem in estonskem jeziku, ki sta uporabljena kot referenci.

Tabela 4.2: Primerjava število lem s številom besednih oblik v korpusu MULTEXT-EAST (Erjavec, 2010)

jezik	število besednih oblik	leme	razmerje ³
slovenščine	20,923	7,895	2.65
srbsščina	21,505	8,392	2.56
češčina	22,273	9,060	2.46
angleščina	11,078	7,020	1.58
estonsščina	18,853	8,679	2.17

Omejitev prostora iskanja poveča natančnost modela poravnave besed, vendar v modelu ni več informacije o besednih oblikah. To informacijo smo ohranili prek povezave s paradigmami. Oglejmo si zmanjšanje iskalnega prostora z opisano metodo:

Nekaj enostavnih definicij, ki bodo olajšale formulacijo Enačbe 4.2:

L - jezik, vse besede

E_L - leme jezika L

$E_{L(i)}$ - i^{ta} lema z vsemi besednimi oblikami

$$|L| = \sum_{i=0}^{|E_L|} E_{L(i)} \quad (4.2)$$

Preiskovalni prostor je zmanjša iz $|L|$ na $|E_L|$.

Oglejmo si primer:

S predpostavko da je roman Georga Orwella "1984" (Orwell, 1949), ki sestavlja večjezični povedno poravnani del korpusa (Erjavec, 2010), dober vzorec opazovanega jezika, v našem primeru slovenščine, dobimo vrednosti števila besednih oblik in lem kot so predstavljene na Sliki 4.9 ter Tabeli 4.2. Preiskovalni prostor se je za slovenščino v primeru te novele zmanjšal iz 20,923 besednih oblik na 7,895 lem.

Izvirni jezik $|L| = 20923$

Lematiziran jezik $|E_L| = 7895$

Slika 4.9: Zmanjšanje iskalnega prostora za slovenski jezik (relativno majhen korpus MULTTEXT-EAST (Erjavec, 2010))

4.4.3 Izdelava statističnega jezikovnega modela ciljnega jezika

Bistveni del celotnega prevajalnega sistema je statistični modul za rangiranje prevodov (Ranker). Glavni problem arhitekture, predstavljene v Razdelku 3.4.3 je, da morfološki analizator ter strukturni prenos povzročata lokalne morfološke ter strukturne dvoumnosti (ambiguities). Njihova kombinacija nato ustvari veliko število različic (hipoteze) v procesu prevajanja. Bilo bi zelo zapleteno, če sploh mogoče, ročno izdelati pravila, ki bi omogočila reševanje tovrstnih dvoumnosti. Zato je v osnovno arhitekturo dodan modul za rangiranje prevodov na osnovi stohastičnih metod, katerega cilj je izbira povedi, ki najbolj ustreza modelu ciljnega jezika; torej povedi, ki je slovnično najbolj pravilno zapisana v ciljnem jeziku. Takšno rangiranje prevodov nam samo zase ne zagotavlja, da bo izbrana poved najboljši prevod izvirne povedi, modul poišče le slovnično najbolj pravilno poved v ciljnem jeziku, predhodni moduli prevajalnega sistema pa zagotavljajo, da se pri prevodu ohrani pomen izvirne povedi.

Jezikovni model je bil naučen na korpusu sestavljenem na naključno izbranih člankih uporabljenih jezikov iz Wikipedije ⁴. Velikost korpusov je bilo približno 15

⁴(<http://cs.wikipedia.org>, <http://en.wikipedia.org>, <http://et.wikipedia.org>, <http://sr.wikipedia.org>)

milijonov besed za češčino in angleščino ter približno 7 milijonov za estonščino in srbščino.

4.4.4 Modeliranje morfoloških oznak izvirnega jezika

ujemanje morfoloških kategorij lahko modeliramo s pravili, ki temeljijo na regularnih izrazih. pravila so opisana v Razdelku 5.1. Uporabljena je bila ista oblika pravil, kot je uporabljena v Apertiumu, saj je bilo najbolj enostavno uporabiti enako tehnologijo. Samodejna izdelava teh pravil je predstavljena v Razdelku 5.3.4. Mehanizem za odkrivanje nemogočih kandidatov prevoda je prikazan v algoritmu Sliki 4.10.

```
izdelaj vse kandidate za prevod
izberi prvi kandidat
while obstaja kandidat do
  sprememba = uporabi pravila na kandidat
  if sprememba == kandidat then
    obdrži kandidata
  else
    zavrzi kandidata
  end if
  izberi naslednjega kandidata
end while
```

Slika 4.10: Izločitev (morebiti) vseh nemogočih kandidatov za prevode z uporabo pravil lokalnega ujemanja.

Vsa pravila, katerih regularni izraz lahko apliciramo na kandidata prevoda, se uporabijo. Če neko pravilo spremeni kandidata, pomeni, da kandidat ni idealno sestavljen in modul ga zavrže.

Poglavje 5

Pravila transferja

V splošnem, povzeto po (Hutchins in Somers, 1992), lahko delovanje sistemov RBMT opišemo kot razpoznavanje (analiziranje) izvornega besedila, ki se zaključi z izdelavo vmesne (simbolne) predstavitve. Iz te predstavitve je v naslednji fazi sintetizirano besedilo v ciljnem jeziku. Prevajalne sisteme opisanega tipa lahko razdelimo glede na vrsto vmesne predstavitve:

- na sisteme z vmesnim jezikom (interlingua);
- na sisteme s transferjem.

Sistemi z vmesnim jezikom uporabljajo eno vmesno predstavitev, kar omogoča enostavno kombiniranje jezikovnih parov, saj lahko nove prevajalne sisteme izdelamo brez novih pravil za prenos iz izvornega v ciljni jezik, za vsak jezik definiramo le pravila za prehod v vmesni jezik. Problem takšnih sistemov pa je, da je izdelava vmesnega jezika zahtevno opravilo še posebej za proste domene ter za sisteme z visoko kakovostjo prevajanja, kamor sodijo tudi sistemi za prevajanje sorodnih jezikov.

Sistemi s transferjem uporabljajo dve vmesni predstavitvi, eno za izvorni jezik ter eno za ciljni, takšni predstavitvi je lažje izdelati, takšen pristop tudi omogoča večjo fleksibilnost, vendar je potrebno izdelati pravila za vsak jezikovni par. Primer arhitekture sistemov s transferjem je prikazan na Sliki 5.1

Sistemi strojnega prevajanja s transferjem delujejo najprej izvedejo analizo izvornega besedila v vmesno predstavitev izvornega jezika, nato uporabijo leksikalne preslikave (dvojezični slovarji) ter strukturna pravila transferja za prenos v vmesno predstavitev ciljnega jezika ter nadalje v dejansko besedilo v ciljnem jeziku. Nivo analize izvornega besedila in s tem stopnja abstrakcije vmesne predstavitve besedila je odvisen od posameznega sistema za prevajanje in, posredno, od prevajanega



Slika 5.1: Splošna shema sistema za strojno prevajanje s transferjem, sistem ima dve vmesni predstavitvi besedila: izvorna ter ciljna vmesna predstavitev, med njima poteka .

jezikovnega para. Prevajanje oddaljenih jezikov, kot sta na primer angleščina in kitajščina, zahteva polno analizo (skladenjsko in pomensko). Pri prevajanju sorodnih jezikov, kot je jezikovni par slovenščine - srbsščina, pa boljše rezultate dosežemo s plitvo analizo ter posledično s plitvim transferjem in s plitvo sintezo, (Homola in Kuboň, 2008a) in (Corbi-Bellot et al., 2005).

5.1 Pravila regularnih izrazov

pravila na osnovi regularnih izrazov se uporabljajo predvsem v sistemih strojnega prevajanja s plitvim transferjem, shallow- MT. Spremembe s pomočjo teh pravil so najpogostejše povezane z leksikalnimi oblikami; tako je ponavadi vmesna predstavitev besedila za vsako besedo sestavljena iz leme, leksikalne kategorije in opisa morfoloških pregibanj. Takšna pravila so tudi omejena na lokalni kontekst.

5.2 Apertiumov format pravil

Apertiumov modul strukturnega prenosa (Structural module) uporablja tehnologijo končnih avtomatov za odkrivanje vzorcev fiksne dolžine leksikalnih enot (kosov besedila ali fraz), ki zahtevajo posebno obdelavo glede na slovnične razlike med jezikoma (na primer: spremembe v spolu, sklonu ali številu za zagotovitev ujemanja v ciljnem jeziku, sprememba vrstnega reda besed leksikalne spremembe kot na primer spremembe v predlogih, ...). Po detekciji vzorcev, ki potrebujejo spremembe, se le-te izvedejo (izhod modula so spremenjene leksikalne enote). pravila so zgrajena iz dveh delov: končnega števila elementov, ki opisuje vzorce fiksne dolžine ter dela, ki omogoča opis akcije, ki je potrebna za spremembo vzorca. Vzorci (pattern) so običajno izraženi v leksikalnih kategorijah, na Primeru 5.2, *pomožni glagol biti*

v *prihodnjiku* ter *glagol poljubne oblike*. Ukrep (action) določa, katere ukrepe je treba izvesti za ustrezen najden vzorec.

Primer pravila je predstavljen na Sliki 5.2. pravilo je sestavljeno iz dveh delov: vzorec (pattern) in ukrep (action). pravilo opisuje spremembe načina zapisa prihodnjika iz slovenščine v srbsčino. Vzorec je sestavljen iz dveh leksikalnih enot: *pomožni glagol biti v prihodnjiku* ter *glagol poljubne oblike*, ukrep pa spremeni lemo prvega glagola v *hteti*, čas prvega glagola v *sedanjik* ter čas drugega glagola v *nedoločnik*, v nadaljevanju so izpisane leksikalne kategorije za obe besedi. Več pravil je predstavljenih v Prilogi A.

5.3 Samodejna izdelava pravil

5.3.1 Izdelava pravil za plitvi prenos na osnovi regularnih izrazov

Pri snovanju preizkusa smo se omejili na samodejno izdelavo pravil plitkega prenosa. Takšna pravila so najprimernejša za prevajalne sisteme sorodnih jezikov. Samodejne metode popolne slovnične analize povedi ne dosegajo zadovoljive kakovosti in njihova vpeljava v sisteme za prevajanje sorodnih jezikov bi zmanjšala kakovost prevodov v primerjavi s prevodi z uporabo pravil plitkega prenosa (shallow rules). To dejstvo predstavljajo avtorji (Homola in Kuboň, 2008a) in (Forcada, 2006). Samodejna izdelava pravil plitkega prenosa je bila izvedena s pomočjo metode in orodja, predstavljenega v (Sanchez-Martinez in Forcada, 2009). To orodje izdeluje pravila, ki jih lahko nepredelana uporabimo v ogrodju prevajalnih sistemov temelječih na Apertiumu (Corbi-Bellot et al., 2005), za ostale sisteme moramo format pravil spremeniti, vendar je tudi ta postopek lahko popolnoma samodejen.

5.3.2 Opis metode

Metoda, predstavljena v (Sanchez-Martinez in Forcada, 2009), je osnovana na predlogah poravnave (AT - Alignment Templates), ki so bile predlagane kot izboljšava osnovnih metod statističnega strojnega prevajanja v (Och in Ney, 2004).

AT lahko definiramo kot posplošitev poravnave parov fraz¹ z uporabo besednih razredov.

¹Fraza v okviru tega dela, tudi splošneje v statističnem strojnem prevajanju, pomeni kolokacijo dveh ali več besed in ne tvori nujno pomensko zaključene enote.

```

<rule>
  <pattern>
    <pattern-item n="vbserfti"/>
    <pattern-item n="vblex"/>
  </pattern>
  <action>
    <let>
      <clip pos="1" side="tl" part="lemh"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="temps"/>
      <lit-tag v="pres"/>
    </let>
    <let>
      <clip pos="2" side="tl" part="temps"/>
      <lit-tag v="inf"/>
    </let>

    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_vbser"/>
        <clip pos="1" side="tl" part="temps"/>
        <clip pos="1" side="tl" part="persona"/>
        <clip pos="1" side="tl" part="nbr"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_vblex"/>
        <clip pos="2" side="tl" part="temps"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika 5.2: Primer pravila za strukturni . pravilo opisuje spremembe načina zapisa prihodnjika iz slovenščine v srbsščino.

AT razširimo z množico omejitev, ki nadzorujejo uporabo predlog kot pravil plitkega transferja. V ta namen:

- povezave med fraza, izvlečene iz učnih primerov, shranimo v dvojezični slovar sistema RBMT, ki omogoča reprodukcijo leksikalne vsebine pri prevajanju;
- besedni razredi so določeni z lingvističnim znanjem in ne na podlagi statistike;
- množica omejitev, ki so bile naučene iz učnih primerov, so dodane vsaki AT in omejujejo uporabo AT kot pravila transferja. Tako spremenjene AT lahko poimenujemo razširjene predloge.

Iz razširjenih predlog lahko sestavimo pravila transferja. Učenje predlog iz povredno poravnane dvojezičnega korpusa je sestavljeno iz treh faz:

1. iskanje besednih povezav med izvornim in ciljnim delom korpusa;
2. izločanje dvojezičnih parov fraz;
3. posplošitev teh dvojezičnih stavek parov fraz z uporabo besednih razredov namesto samih besed.

Uporaba razredov besed omogoča opis zamenjave vrstnega reda besed, sprememb pri sklanjanju in uporabi predlogov ter ostalih razlikah med izvornim in ciljnim jezikom oziroma njunima vmesnima predstavitevama.

5.3.2.1 Izdelava pravil

Strukturni v Apertiumu uporablja končne avtomate za iskanje vzorcev leksikalnih oblik fiksne dolžine. Pri izbiri uporablja algoritem najdaljšega ujemanja vzorca iz leve proti desni (LRLM - Left to Right Longest Match). Za izbrana pravila na vzorcu izvede ukrepe (actions) opisane v pravilih. Generično pravilo plitvega prenosa je tako sestavljeno iz vzorca leksikalnih oblik za iskanje ter iz opisa potrebnih transformacij na tem vzorcu. Primer pravila je predstavljen na Sliki 5.2.

pravilo je sestavljeno iz množice razširjenih predlog poravnave (AT) z enakim zaporedjem besednih razredov izvornega jezika, vendar z drugačnimi sekvencami besednih razredov ciljnega jezika in/ali drugačno poravnavo in/ali drugačnim naborom omejitev ciljnega jezika.

Generirana koda izvede najbolj pogosto AT, ki zadošča omejitvam ciljnega jezika.

5.3.3 Izbira najboljših pravil

Metode za samodejno in nenadzorovano izdelavo pravil po navadi izdelajo veliko množico pravil, za izbiro najboljših pravil uporabimo metode ocenjevanja pravil, ena od možnih izbir je metoda, predstavljena v (Vičič in Forcada, 2008). Metoda temelji na jezikovnem modelu ciljnega jezika.

Model ciljnega jezika kot opisan v (Bahl et al., 1989) in (Clarkson, 1997), v našem primeru jezikovni model na tri-gramih, se uporablja kot merilo za točkovanje kakovosti izdelanih kandidatov za prevode. Jezikovni model dodeli višjo vrednost, ponavadi verjetnost, zaporedjem besed, ki se večkrat pojavljajo v učni množici. Jezikovni model, za vsako poved, poskuša določiti njeno verjetnost da bi se pojavila v učnem korpusu, ne pa verjetnost, da je ta poved res prevod izvirne povedi.

Metoda vrednotenja pravil temelji na predpostavki, da model ciljnega jezika zadošča za določitev kakovosti pravil transferja, saj modul za strukturni prenos (pravila transferja) ne spremenijo pomena prevodom. Kot opisano v Razdelku 5.1, se pravila transferja največ ukvarjajo z zamenjavo vrstnega reda besed, ter ujemanjem sosednjih besed.

Z dovolj velikim testnim korpusom ovrednotimo kakovost pravil. pravila, ki katerih uporabi so prevodi boljše ocenjeni, so tudi sama boljše ocenjena in pri večih možnostih, sistem izbere pravilo z boljšo oceno.

5.3.4 Izdelava pravil na osnovi regularnih izrazov za izražanje lokalnega ujemanja morfoloških kategorij

Modul za odpravljanje napak lokalnega ujemanja morfoloških kategorij uporablja enak tip pravil kot modul za plitvi opisana v (Sanchez-Martinez in Forcada, 2009), pravila so poenostavljena. pravila odražajo le ujemanje sosednjih besed, kot je na primer ujemanje pridevnikov ter samostalnikov v sklonu, spolu ter številu. Takšna pravila je veliko lažje sestaviti kot pravila za strukturni . Metoda odkriva le lokalno ujemanje v okviru največ treh besed (z uporabo jezikovnega modela na osnovi tri-gramov), vendar bi samo z uporabo drugačnega modula lahko razširili njeno delovanje. Zahteve za uporabo metode so preprostejše kot pri metodi za izdelavo pravil strukturnega prenosa opisani v (Sanchez-Martinez in Forcada, 2009), potrebujemo le enojezični, morfološko označen korpus.

pravila lokalnega ujemanja uporabljata dva modula s Slike 3.4, modul za izbiro množice kandidatov ter modul za iskanje lokalnega ujemanja. Prvi modul uporablja pravila naučena na morfološko označenem korpusu izvornih besedil, drugi na enako označenem korpusu ciljnih besedil. Tudi pri tej metodi smo uporabili korpus (Erjavec, 2010), ki vsebuje tako izvirno kot ciljno besedilo.

Trigrame in bigrame morfoloških oznak, brez dejanskih besed, samo oznake, zgradimo iz označenega korpusa s pomočjo standardne metode za izdelavo stohastičnih jezikovnih modelov, kot je prikazana v (Homola in Kuboň, 2008b). Za vsak bigram in trigram je bilo preverjeno ujemanje morfoloških kategorij med vsemi enotami. Oznake in njihovi položaji so prosti. Kandidat za novo pravilo lokalnega ujemanja je shranjen v obliki, prikazani na Sliki 5.2, če obstajajo ujemanja morfoloških kategorij med enotami trigrama oziroma bigrama. pravilo določimo kot veljavno, če se tako opisano ujemanje pojavi na enakih trigramih v določenem številu primerov. Prag za izbiro veljavnih pravil je bil določen empirično na podlagi manjšega števila testnih primerov, metodo za boljšo določitev praga bo treba dodatno raziskati. Algoritem, ki opisuje ta postopek je prikazan na Sliki 5.3.

```

izdelaj bigrame in trigrame iz korpusa
Izberi prvega kandidata
while še obstajajo kandidati do
    ugotovi katere morfološke oznake se ujemajo
    if najdi razred z istim ujemanjem a then
        dodaj kandidata v razred a
    else
        naredi nov razred
    end if
end while
izberi razrede z več kot pragom kandidatov

```

Slika 5.3: Proces samodejne izdelave pravil iz označenega korpusa.

5.3.5 Primeri uporabe pravil za lokalno ujemanje morfoloških kategorij

Leksikalna podobnost sorodnih jezikov, obširneje predstavljena v Razdelku 2.3.4, zagotavlja, da se večina besed semantično enostavno veže na besede v ciljnem jeziku, tako je izdelava slovarjev enostavna in napake so redke. Kljub temu pa lahko obstaja potreba po razširitvi dvojezičnih slovarjev z morfološkimi podatki. Primer je opisan v Razdelku 2.3.4 ter predstavljen na Primeru 2.7. Oglejmo si takšen primer na jezikovnem paru slovenščine-srbščina, kjer obstaja nekaj samostalnikov, ki so različnih spolov v obeh jezikih. Primer 2.7 kaže spremembo spola, iz srednjega v moški, pri prevodu besede *okno* v srbsko besedo *prozor*, v obeh jezikih se pridevnik ujema s samostalnikom v spolu.

Na Sliki 5.4 je zapisano pravilo za lokalno ujemanje pridevnika in samostalnika.

```
<rule>
  <pattern>
    <pattern-item n="adj"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_adjec"/>
        <clip pos="2" side="tl" part="gen"/>
        <clip pos="2" side="tl" part="nbr"/>
        <clip pos="2" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="adj_degree"/>
        <clip pos="2" side="tl" part="adj_definitness"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_nom"/>
        <clip pos="2" side="tl" part="gen"/>
        <clip pos="2" side="tl" part="nbr"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>
```

Slika 5.4: pravilo ujemanja pridevnika in samostalnika, ki si sledita. Besedi se morata ujemati v spolu, sklonu ter številu. Pri prevajanju se spreminjajo morfološke kategorije samostalnika in ne pridevnika, zato je ujemanje vezano na samostalnik.

Pridevnik in samostalnik na Primeru 5.1 se ujemata spolu, sklonu, številu ter določnosti.

(5.1) *kolesa* *so* *vozila*
 kolo-SAM, SR, MN biti-P. GLAG, MN voziti-GLAG, SR, MN, PRET
 "kolesa so vozila" (SLO)

bicikli *su* *vozili*
 bicikl-SAM, M, MN jesam-P. GLAG, MN voziti-GLAG, M, MN, PRET
 "Bicikli su vozili" (SR)

Na Sliki 5.5 je zapisano pravilo za lokalno ujemanje samostalnika, pomožnega glagola leme *jesam* - biti ter glagola.

pravila lokalnega ujemanja so omejena na fiksno določeno okolico in ne zajamejo oddaljenih odvisnosti (long-distance relations). Primer 5.2 kaže poved, kjer s pravili lokalnega ujemanja ne moremo popraviti možnih napak. Vrinjeni stavki so poljubne dolžine in jih ne moremo opisati s pomočjo vzorcev končne dolžine kakršne uporablja Apertium. Sistemi plitkega transferja takšnih odvisnosti ne zajamejo in pri njihovi uporabi se zanašamo na podobnost jezikov, kjer naj bi bilo takšnih problemov malo.

(5.2) *Kolo, bilo je*
 kolo-SAM, SR, ED biti-P. GLAG, ED, PRET biti-P. GLAG, ED
rdeče, je vozilo...
 rdeče-PRID, SR, ED, M biti-P. GLAG, ED voziti-GLAG, SR, ED, PRET
 "Kolo, bilo je rdeče, je vozilo..." (SLO)

Bicikl, bio je
 bicikl-SAM, M, ED biti-P. GLAG, ED, PRET jesam-P. GLAG, ED
crven, je vozilo...
 crven-PRID, M, ED, M biti-P. GLAG, ED voziti-GLAG, SR, ED, PRET
 "Bicikl, bio je crven, je vozilo(NAPAKA)..." (SR)

```

<rule>
  <pattern>
    <pattern-item n="CAT__noun"/>
    <pattern-item n="CAT__vbserjesam"/>
    <pattern-item n="CAT__vblex"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lema"/>
        <clip pos="1" side="tl" part="noun"/>
        <clip pos="1" side="tl" part="noun_1"/>
        <clip pos="1" side="tl" part="noun_2"/>
        <clip pos="1" side="tl" part="noun_3"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="vbser"/>
        <clip pos="2" side="tl" part="vbser_1"/>
        <clip pos="2" side="tl" part="vbser_2"/>
        <clip pos="1" side="tl" part="sam_2"/>
        <clip pos="2" side="tl" part="vbser_4"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lemh"/>
        <clip pos="3" side="tl" part="vblex"/>
        <clip pos="3" side="tl" part="vblex_1"/>
        <clip pos="1" side="tl" part="sam_2"/>
        <clip pos="1" side="tl" part="sam_3"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika 5.5: pravilo ujemanja samostalnika, pomožnega glagola leme *jesam* - biti ter glagola. Pomožni glagol ter samostalnik se ujemata v številu, samostalnik in glagol na tretjem mestu se ujemata v spolu in številu.

Poglavje 6

Metodologije vrednotenja sistemov in rezultati vrednotenj

6.1 Evalvacija sistemov za strojno prevajanje

6.1.1 Samodejne metode

6.1.1.1 Metrika BLEU

Bilingual Evaluation Understudy - BLEU (Papineni et al., 2001) je bila prva in je še vedno najbolj razširjena metrika za evalvacijo kakovosti prevodov sistemov strojnega prevajanja. Kakovost prevodov je predstavljena kot natančnost ujemanja prevodov sistema za strojno prevajanje z referenčnimi prevodi poklicnih prevajalcev. Vrednosti so izračunane za posamezne prevedene odseke, po navadi povedi, ter povprečene za celoten testni korpus. Berljivost ter slovnična pravilnost nista upoštevani.

BLEU uporablja spremenjeno različico preciznosti (precision), ki za razred predstavlja število pravilno klasificiranih elementov (true positives), za primerjavo kandidata za prevod z enim ali več referenčnimi prevodi. Sprememba z osnovno preciznostjo naj bi poskrbela za lastnost sistemov strojnega prevajanja, ki težijo k daljšim prevodom.

Na Sliki 6.1 je predstavljen primer kandidata za prevod ter referenčni prevod.

Hruška hruška hruška .
Hruška je sladka .

Slika 6.1: Kandidat za prevod ter referenčni prevod.

Osnovno preciznost (precision) opisuje Enačba 6.1, izračunane vrednosti veljajo za primer na Sliki 6.1.

$$P = \frac{m}{w_t} = \frac{3}{3} = 1 \quad (6.1)$$

m predstavlja število besed kandidata za prevode, ki so v referenčnih prevodih in w_t število vseh besed v kandidatu za prevod. Izračunan vrednost je 1, kar bi pomenilo popoln prevod kar seveda ni res. Sprememba metrike BLEU je, da za vsako besedo v kandidatu za prevod, algoritem poišče največje število pojavitev v referenčnih prevodih m_{max} , za primer opisan na Sliki 6.1 velja $m_{max} = 1$, saj se beseda *hruška* pojavi le enkrat. Enačba opisuje spremenjeno metriko ter izračunano vrednost za primer na Sliki 6.1.

$$P = \frac{m}{w_t} = \frac{1}{3} \quad (6.2)$$

Metoda je uporabljena za n-grame do predefinirane dolžine, po navadi $n = 4$. Rezultati unigramov približno odražajo ustreznost prevodov (adequacy), koliko izvirne vsebine je preneseno v prevod. Rezultati za daljše unigrame pa opisujejo slovnično pravilnost prevoda (fluency).

Veliko avtorjev, kot na primer (Callison-Burch et al., 2006) in (Labaka et al., 2007), se strinja, da metrika BLEU sistematično zapostavlja sisteme temelječe na pravilih (RMBT) ter da ni primerna za pregibne jezike. Metrika naj bi bila uporabljana v ožjem obsegu kot doslej, predvsem za primerjanje sorodnih sistemov ter za sledenje postopnih sprememb pri gradnji sistema za strojno prevajanje.

6.1.1.2 Metrika METEOR

Metric for Evaluation of Translation with Explicit ORdering - METEOR Lavie in Agarwal (2007) temelji na harmonični sredini natančnosti ter priklica unigramov (unigram precision and recall), kjer je priklic močnejše utežen kot natančnost. Vsebuje še več metod jezikovnih tehnologij, ki niso prisotne pri ostalih samodejnih metrikah strojnega prevajanja, kot so krnjenje in ujemanje sinonimov kot pomoč pri iskanju ujemanja besed. Krnjenje je predvsem primerno za visoko pregibne jezike saj omejuje vpliv napačne uporabe pregibanja; na primer napačne uporabe sklona pri samostalniki.

6.1.2 Metode, ki vključujejo posege strokovnjakov

6.1.2.1 Utežena Levenshteinova razdalja

Metrika temelječa na uteženi Levenshteinovi razdalji (weighted Levenshtein edit-distance) (Levenshtein, 1965), poznana tudi kot Word Error Rate (WER) izračuna najmanjše število sprememb, ki jih moramo narediti za izdelavo *pravilne* povedi v ciljnem jeziku iz samodejno izdelane povedi (prevoda ocenjevanega sistema). Število sprememb še utežimo z dolžino povedi. Dovoljene spremembe so vstavitev, brisanje ter zamenjava besede.

Kot pravilen prevod po navadi pojmujeemo poved, ki popolnoma izraža pomen izvirne povedi ter je slovnično pravilno zapisana v ciljnem jeziku. Opisana metrika kaže, koliko dela moramo opraviti za izdelavo dobrega prevoda iz že izdelanega strojnega prevoda, metrika v grobem ponazarja kompleksnost opravila končnega čiščenja prevodov (post-editing task).

Izvedbo testiranja kakovosti prevodov prevajalnega sistema s pomočjo Levenshteinove razdalje sestavljajo naslednja dejanja:

- izbira testnih povedi v izvirnem jeziku;
- prevajanje testnih povedi s pomočjo testiranega sistema;
- *ročno* popraviljanje prevodov, popraviljavci upoštevajo navodilo čim manjšega števila sprememb;
- izračun utežene Levenshteinove razdalje;

Predstavljena metrika opisuje velikost napake prevajalnega sistema. pogosto želimo rezultate takšne evalvacije predstaviti kot kakovost prevajalnega sistema, takrat uporabimo različico metrike, imenovano Word Recognition Rate - WRR, ki je enostavno razlika med "popolnim"prevodom ter napako sistema, torej: $(1 - \text{WER})$.

6.1.2.2 Evalvacija po smernicah LDC

Smernice LDC (LDC, 2005) so bile predstavljene na letni delavnici evalvacije strojnega prevajanja NIST (Machine Translation Evaluation Workshop) in so najpogosteje uporabljana načela za ročno ocenjevanje kakovosti prevodov sistemov za strojno prevajanja. Pri ročnem ocenjevanju kakovosti prevodov upoštevamo dve lestvici, ki predstavljata vsebinsko ustreznost prevodov (adequacy) ter slovnično pravilnost prevodov v ciljnem jeziku (fluency).

Prva lestvica kaže kakovost prevodov, koliko izvirnega pomena se je pri prevodu ohranilo:

- 5 = Vse
- 4 = Večina
- 3 = Precej
- 2 = Malo
- 1 = Nič

Druga lestvica kaže slovnično pravilnost povedi v ciljnem jeziku. Pri prevodu v ciljni jezik velja:

- 5 = Prevod brez napak
- 4 = Dober ciljni jezik
- 3 = Ciljni jezik, kot ne-materni jezik (non-native language)
- 2 = Ciljni jezik z veliko napakami
- 1 = Nesmiselno besedilo

Ločeni lestvici za kakovost prevodov ter slovnično pravilnost sta bili izdelani ob predpostavki, da lahko tudi prevod z veliko slovničnimi napakami prikaže vso informacijo, ki je zapisana v originalu.

6.2 Rezultati

Metode, predstavljene v tem prispevku v Razdelku 4.4 se osredotočajo na gradnjo sistemov za strojno prevajanje za sorodne, morfološko bogate jezike. Poleg same uporabnosti predstavljenih metod ter evalvacije hitrosti izdelave novih prevajalnih sistemov, stremi predstavljena evalvacija k preverjanju kakovosti samodejno izdelanih podatkov za sisteme strojnega prevajanja na popolnoma funkcionalnih sistemih. Štiri popolnoma delujoči sistemi za strojno prevajanje so bili zgrajeni in ovrednoteni:

1. SL-SR, prevajalni sistem za jezikovni par slovenščine - srbsščina
2. SL-CS, prevajalni sistem za jezikovni par slovenščine - češčina
3. SL-EN, prevajalni sistem za jezikovni par slovenščine - angleščina
4. SL-ET, prevajalni sistem za jezikovni par slovenščine - estonščina

6.2.1 Opis sistemov

Sistem za prevajanje jezikovnega para slovenščina-srbščina (SL-SR) je bil zgrajen kot pilotni sistem, ki je služil za testiranje naših metod v procesu svojega razvoja. Metode, predstavljen v tem prispevku so bile pregledane v več iteracijah (sistematične napake so bile odpravljene in popravki vključeni v novo iteracijo sistema). Ta jezikovni par je bil uporabljen za preverjanje kakovosti predstavljenih metod na popolnoma funkcionalnem sistemu strojnega prevajanja. Oba jezika sta pregibno, morfološko ter derivacijsko bogata. Čeprav sta oba jezika sorodna, visoka stopnja pregibnosti obeh jezikov še vedno zahteva morfološko analizo izvirnega jezika in posledično morfološko sintezo ciljnega jezika

. Sistem za prevajanje jezikovnega para slovenščina-češčina (SL-CS) je bil izdelan tako za preverjanje uporabnosti metod, predstavljenih v Razdelku 4.4 na novem jezikovnem paru sorodnih jezikov in da bi preizkusili, kako hitro je možno izdelati nov sistem. Lastnosti tega jezikovnega para so podobne lastnostim prvega jezikovnega para (SL-SR). Sistem je izdelala ena oseba v dveh dneh na običajnem osebem računalniku ¹.

Sistema za jezikovna para SL-EN in SL-ET sta bila izdelana za oceno uporabnosti predstavljenih metod in celotne arhitekture za oddaljene jezikovne pare. Rezultati predstavljeni v Razdelkih 6.2, 6.2.2.2 in 6.2.2.3 kažejo jasno zmanjšanje kakovosti prevodov z uporabo enake metodologije in enakih učnih podatkov. Estonski jezik je bil izbran kot oddaljen pregibni jezik, angleški jezik je bila izbran kot linearni, oddaljen jezik,

6.2.2 Izbrane evalvacijske metrike

Evalvacija prevodov je bila opravljena s tremi metodami vrednotenja, vsaka od njih je podrobno opisana v Razdelku 6.1, sama uporaba pa v nadaljevanju razdelka:

1. Samodejna objektivna evalvacija z uporabo metrike METEOR (Banerjee in Lavie, 2005; Lavie in Agarwal, 2007).
2. Evalvacija z metodo, ki vključuje posege strokovnjakov na podlagi utežene Levenshteinove razdalje.
3. Evalvacija z metodo, ki vključuje posege strokovnjakov na osnovi smernic (LDC, 2005).

Metrika BLEU (Papineni et al., 2001) je najbolj razširjena metrika za evalvacijo sistemov strojnega prevajanja, vendar je nismo uporabili, saj mnogi avtorji kot na

¹prenosni računalnik z 2 GB RAM in procesorjem Intel Core2 duo.)

primer (Callison-Burch et al., 2006) in (Labaka et al., 2007) soglašajo, da BLEU sistematično zapostavlja sisteme paradigme RBMT in ni primerna za visoko pregibne jezike. Avtorji metrike METEOR (Banerjee in Lavie, 2005), (Lavie in Agarwal, 2007) navajajo, da njihova metrika rešuje večino težav metrike BLEU. Trdijo tudi, da ta metrika bolje korelira s človeško presojo.

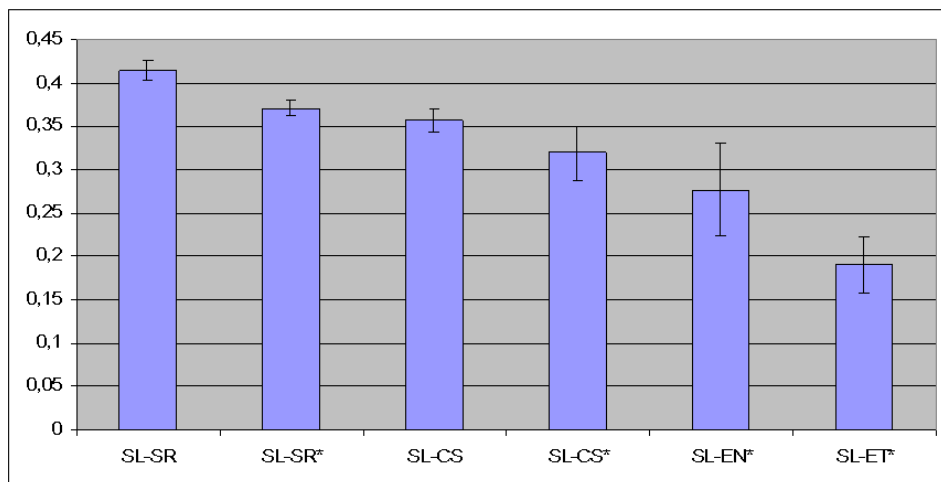
6.2.2.1 Samodejna objektivna evalvacija z metriko METEOR

Metrika METEOR je natančneje opisana v Razdelku 6.1.1.2. Uporabljena je bila javno dostopnih implementacija metrike METEOR (Lavie in Agarwal, 2007), verzija v0.6. Metrika uporablja mehanizem krnjenja kot enega izmed algoritmov za večanje korelacije s človeško oceno za visoko upogibne jezike. Uporabili smo mehanizem krnjenja, ki je stranski proizvod našega prevajanja sistem. Rezultati so predstavljeni na Sliki 6.2, števila označene z * kažejo vrednosti metrike METEOR z običajnim Porter-stem (Porter, 1980) algoritmom krnjenja, ostala števila kažejo vrednosti metrike z lastnim algoritmom krnjenja.

Večjezični vzporedni korpus (Erjavec, 2010) je bil uporabljen kot testna množica primerov s poravnanimi referenčnimi prevodi. Testni primeri niso bili uporabljeni pri gradnji sistemov. K-kratno prečno preverjanje (Kohavi, 1995) je bilo uporabljeno kot metoda za določitev generalizacije rezultatov na neodvisnih podatkih, saj je ena najprimernejših metod za majhne podatkovne nize. V našem primeru je bilo uporabljeno petkratno prečno preverjanje namesto pogostejše uporabljenega desetkratnega prečnega preverjanja, saj gradnja sistema za prevajanje ni popolnoma avtomatizirana. Korpus je bil razdeljen na pet delov, vsak del sestavljen iz približno 1.700 povedi. Pri vrednotenju je bila petina podatkov dodeljena testni množici, ostale povedi pa učni množici, postopek je bil izveden petkrat. Rezultati evalvacije so predstavljeni na Sliki 6.2.

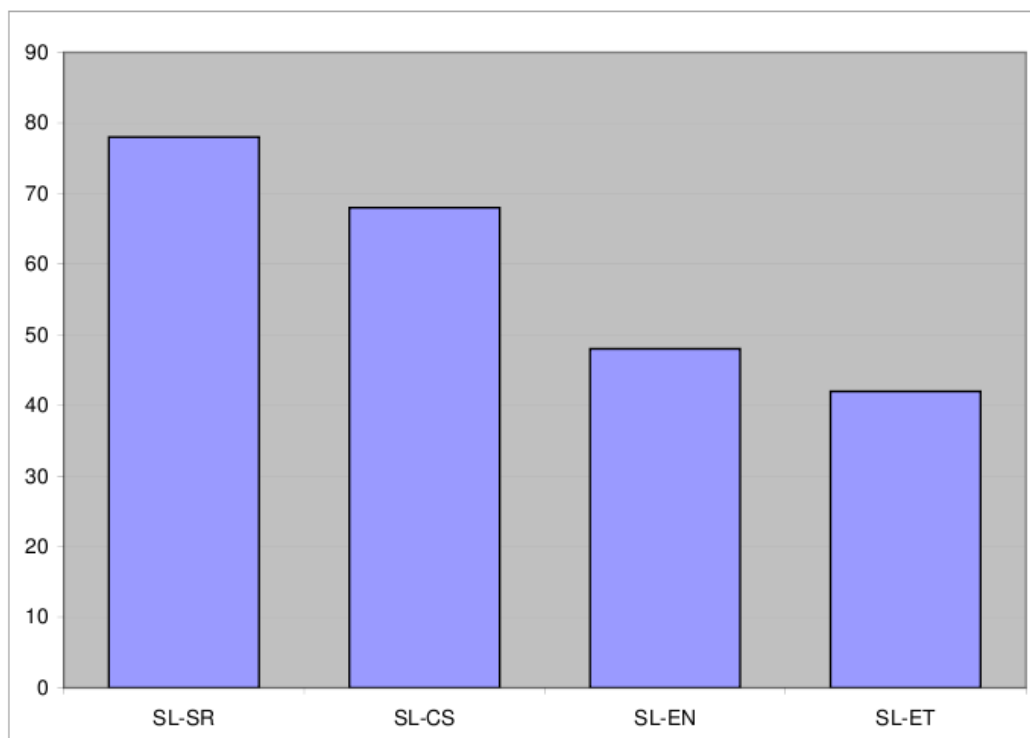
6.2.2.2 Evalvacija z metodo, ki vključuje posege strokovnjakov na podlagi utežene Levenshteinove razdalje

Utežena Levenshteinova razdalja je natančneje predstavljena v Razdelku 6.1.2.1. Iz korpusa smo naključno izbrali 200 povedi, ki niso bile del učne množice. Uporabili smo soležne povedi za vse jezike (iste testne primere, vendar v drugem jeziku). Povedi so bile prevedene s prevajalnim sistemom ter ročno popravljene. Kot *zadovoljivi prevod* štejemo prevod, ki popolnoma odraža vsebino izvirne povedi ter je slovnično popolnoma pravilno zapisan v ciljnem jeziku. Izračunali smo uteženo Levenshteinovo razdaljo med prevodi sistema ter popravljenimi prevodi. Popravljalci so sledili napotkom naj prevode popravijo s čim manj spremembami.



Slika 6.2: Rezultati evalvacije z metriko METEOR. Uporabili smo 5-kratno prečno preverjanje. Vrednosti predstavljajo povprečja 5 iteracij s standardno deviacijo. Evalvacije označene z zvezdico * predstavljajo uporabo krnjenja z algoritmom Porter-stem, ostala pa z uporabo lastnega algoritma za krnjenje.

Vrednotenja so večinoma opravljali študenti in raziskovalci, sodelujoči pri poskusu. Ocene kakovosti prevodov slovenščine in češčine sta opravila po dva ocenjevalca, ki jima je bil ciljni jezik materin jezik (native speaker). Ocene kakovosti prevodov angleščine, srbščine in estonščine, je opravil po en ocenjevalec, ki mu je bil ciljni jezik materin jezik. Rezultati so predstavljeni na Sliki 6.3 ter predstavljajo WRR, Word Recognition Rate (1 - WER), ki predstavlja kakovost sistema namesto napake sistema.



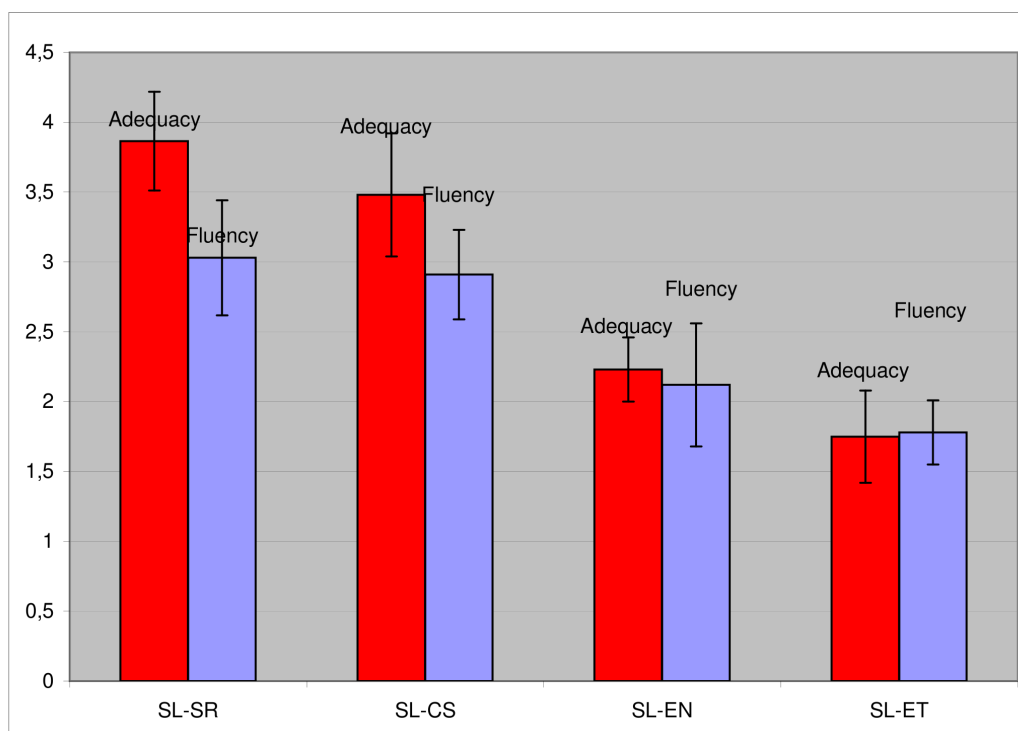
Slika 6.3: Rezultati evalvacije s pomočjo metrike Word Recognition Rate (WRR).

6.2.2.3 Evalvacija z metodo, ki vključuje posege strokovnjakov na podlagi smernic (LDC, 2005)

Metoda ročne evalvacije na podlagi smernic (LDC, 2005) je natančneje predstavljena v Razdelku 6.1.2.2. Iz korpusa smo naključno izbrali 100 povedi, ki niso bile del učne množice. Uporabili smo soležne povedi za vse jezike (iste testne primere, vendar v drugem jeziku).

Vrednotenja so večinoma opravljali študenti in raziskovalci sodelujoči pri poskusu. Ocene za slovenski jezik in češki jezik sta opravila po dva ocenjevalca, ki jima je bil ciljni jezik materni jezik (native speaker), ocene za angleški jezik, srbski jezik in za estonski jezik, je opravil po en ocenjevalec, ki mu je bil ciljni jezik materni jezik. Rezultati so predstavljeni na Sliki 6.4. Rezultati za sistema SL-SR in SL-CS so zadovoljivi, predvsem vrednosti za ustreznost prevodov, vrednosti za preostala sistema so nižje, predpostavljamo, da predvsem na račun razlikosti jezikovnih pa-

rov.



Slika 6.4: Rezultati evalvacije po smernicah (LDC, 2005). Povprečne vrednosti štirih neodvisnih ocenjevanj kažejo visoke vrednosti za vsebinsko ustreznost prevodov (adequacy) in nižje vrednosti za slovnično pravilnost.

Tabela 6.1 kaže zadovoljivo (satisfactory) (SL-CS) ter zelo visoko (very-high) (SL-SR) ujemanje med ocenjevalci (inter-rater agreement) glede na Cohenov kapa koeficient (Cohen, 1960).

Tabela 6.1: Cohenov kapa koeficient (Cohen, 1960) za sistema SL-SR in SL-CS kaže zadovoljivo (satisfactory) (SL-CS) ter zelo visoko (very-high) (SL-SR) ujemanje med ocenjevalci (inter-rater agreement).

ff	sl-sr	sl-cs
kapa	0,86	0,69
95% CI	0,70 – 0,90	0,57 – 0,81
opazovano ujemanje	0,86	0,79
pričakovano ujemanje	0,300	0,317
primeri	100	100

Poglavje 7

Prevajanje na osnovi dreves izpeljave

Statistično strojno prevajanje (SMT), kot je definirano v (Al-Onaizan et al., 1999), predstavlja eno najbolj raziskanih področij CBMT v zadnjih letih. SMT temelji na statističnih modelih, katerih parametri so izpeljani iz opazovanja dvojezičnih vzporednih korpusov. Statistično strojno prevajanje na osnovi dreves izpeljave (Statistical Machine Translation by Parsing - SMTbyP), kot je opisano v (Melamed, 2004a), predstavlja podmnožico SMT, kjer so parametri statističnih modelov naučeni pri analizi skladenjsko označenih, dvojezičnih, vzporednih korpusov.

Najpomembnejša prednost sistemov SMTbyP v primerjavi s sistemi SMT je v zmožnosti obvladovanja rekurzivnih struktur v povedih, primer so vrinjeni stavki. Pri učenju modelov ter pri samem prevajanju (uporabi modelov) so uporabljeni statistični modeli analize besedila (statistical parsing models). Večina statističnih modelov analize besedila, primera sta (Collins, 2003) in (Charniak, 2000) je naučenih na skladenjsko označenih dvojezičnih poravnanih korpusih (treebanks), primer takšnega korpusa je (Marcus et al., 1993). Manj uporabljeni jeziki (less used languages) takšnih korpusov nimajo.

7.1 Osnove

Osnovna hipoteza na kateri temelji predstavljena metoda je, da POS oznake vsebujejo dovolj skladenjske informacije, ki omogoča abstrakcijo besed iz učnega korpusa. Statistični model naučen na besedah korpusa je modeliral ločeno. Prostor iskanja se z uporabo POS oznak namesto pravih besed močno zmanjša, tako potrebujemo manj podatkov za učenje učinkovitih prevajalnih modelov. Niz sestavljen iz POS oznak lahko sestavimo iz izvirnega označenega besedila z brisanjem besed. Takšni nizi predstavljajo liste v drevesih izpeljave (z abstrahiranjem besed). SMTbyP gradi drevo izpeljave izvirne povedi ter ga poravna z drevesom izpeljave

v ciljnem jeziku. Naš pristop uporablja iste algoritme z eno razliko: niz POS oznak, izdelan iz izvirne označene povedi je poravnan z drevesom izpeljave v ciljnem jeziku. Predstavljena metoda uporablja poravnave med POS oznake izvirne povedi ter drevesom izpeljave v ciljnem jeziku. Tako lahko metodo uporabimo tudi za manj uporabljene jezike.

7.2 Metoda

Večina metod statističnega strojnega prevajanja (Brown et al., 1993), (Melamed, 2004a) je jezikovno neodvisna, prevajalne metode delujejo dvosmerno, omogočajo prevajanje iz izvirnega v ciljni jezik ter obratno. Jezikovna neodvisnost je omogočena z indukcijo prevajalnega znanja iz vzporednih podatkov brez dodatnega jezikovnega znanja.

Osnovna učna množica naše metode je prav tako poravnan vzporedni dvojezični korpus. Metoda obe predstavljeni splošnosti oziroma neodvisnosti zanemarija, saj zahteva jezik s standardizirano zbirko dreves izpeljave (treebank) kot ciljni jezik ter jezik s solidnim označevalnikom POS kot izvorni jezik. V prvo skupino sodi le peščica največjih svetovnih jezikov kot so angleščina, arabščina, kitajščina, španščina in drugi, v drugo skupino pa veliko večja množica jezikov, saj je razvoj označevalnika POS veliko manjši problem. Metoda je razdeljena na dva dela:

- Učenje povezav med oznakami POS izvornih povedi ter izdelanimi drevesi izpeljave ciljnega jezika.
- Preiskovanje naučenih primerov iz prvega dela in izdelava množice fiksne velikosti najboljših kandidatov ciljnih prevodov (n-best-set).

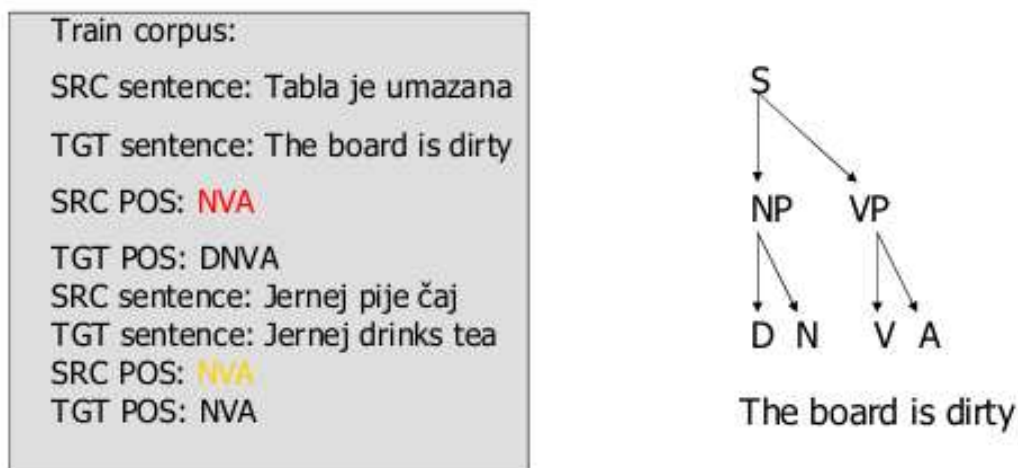
7.2.1 Učenje povezav med oznakami in drevesi

Prevajalni model je naučen na dvojezičnem vzporednem korpusu, kot (Erjavec, 2010). Korpus je sestavljen iz povedno poravnanih delov izvirnega in ciljnega jezika, kot je prikazano na Sliki 7.2.

Standardni algoritem SMTbyP gradi drevo izpeljave iz izvirne povedi ter ga poravna (poišče skupne točke) z ustreznim drevesom izpeljave v ciljnem jeziku. Besede so modelirane v posebnem statističnem modelu, uporabimo lahko praktično poljuben model poravnave posameznih besed (word-by-word alignment model).

Naš pristop je v primerjavi z osnovnim različen le v akcijah, ki vključujejo izvirno poved, saj metoda izhaja iz dejstva, da za izvorni jezik ne obstaja dovolj dobrega algoritma za skladiščno analizo povedi. Vsak par povedi iz učnega dela korpusa je

obravnavan samostojno. Ciljna poved je analizirana z analizatorjem (Collins, 2003), ki je bil naučen na skladenjsko označenem korpusu (Marcus et al., 1993); rezultat analize je drevo izpeljave z izračunano stopnjo zaupanja (confidence score). Primer drevesa izpeljave je predstavljen na Sliki 7.1.



Slika 7.1: Primer drevesa izpeljave.

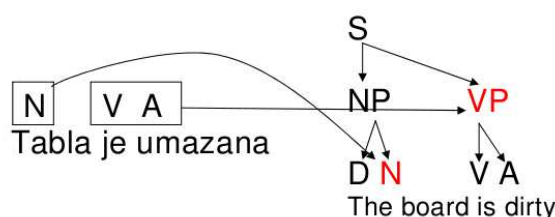
Drevesa izpeljave so sestavljena iz besed v listih, sledijo oznake POS v prvem nivoju. Oznake POS so združene v frazah, ki tvorijo preostale nivoje do vrhnjega. Vsaka beseda ima ustrezno oznako POS. Abstrakcija besed v drevesih izpeljave skoraj nič ne vpliva na količino informacij s stališča skladnje. Notranja vozlišča denotirajo simbole slovnice.

Za izvorni jezik ne obstaja skladenjski analizator; izvorna poved je morfološko označena z oznakami POS, v našem testnem primeru smo uporabili že vnaprej pripravljen ter označen korpus (Erjavec, 2010), oznake POS so bile pridobljene iz korpusa. Opisano zaporedje pare oblike razvidne iz primera prikazanega na Sliki 7.2.

Poravnave so točkovane glede na izbrana pravila, ki so uporabljena pri izdelavi poravnave (vsaka skupina pravil ima svojo težo). Končni izdelek je par prikazan na primeru b) na Sliki 7.2.

7.2.2 Prevajanje

V tej fazi se odvija prevod vhodne izvirne povedi v, po možnosti pravilno, poved v ciljnem jeziku. Vhodna poved, poved, ki jo prevajamo, je morfološko označena



Slika 7.2: Primer poravnave morfoloških oznak (POS) z drevesom izpeljave.

(oznake POS). Algoritem poišče niz oznak POS v učnih podatkih, enostavno iskanje s popolnim ujemanjem je razširjeno z iskanjem podnizov ter iskanjem podobnih nizov, podobnost je računana s pomočjo metode Levenshteinove razdalje (edit distance), (Levenshtein, 1965). Rezultati so ocenjeni v skladu z uporabljenimi iskalno metodo (metode so točkovane). Rezultat iskanja je množica najboljših n-teric.

V zadnjem koraku metode se vsak zapis samostojno uporablja za izdelavo kandidata za prevod. Besede ciljnega drevesa izpeljave so napolnjene prek poravnave z izvornimi oznakami POS ter posredno z izvornimi besedami. Izvirne besede so prevedene s pomočjo modela za neposredni prevod besed (word-by-word model).

Prevodi so točkovani s pomočjo točkovanj posameznih faz prevajalnega procesa ter pomnoženi z verjetnostjo, da kandidat za prevod sodi v jezik ciljnega jezika po statističnem jezikovnem modelu ciljnega jezika, v našem primeru smo uporabili jezikovni model (Clarkson, 1997). Najbolje točkovani kandidat je izbran kot končni prevod.

Poglavje 8

Razprava in nadaljnje delo

8.1 Prispevki k znanosti

Disertacija vsebuje izvirne prispevke k področju strojnega prevajanja naravnih jezikov na osnovi pravil plitkega prenosa. Izvirne prispevke k znanosti smo objavili v naslednjih glavnih publikacijah (Vičič in Brodnik, 2008; Homola et al., 2009; Mikolič et al., 2009; Vičič, 2009) ter v vrsti referatov, objavljenih na mednarodnih konferencah (Vičič in Erjavec, 2002; Vičič, 2007a,b; Vičič in Forcada, 2008; Vičič, 2008; Vičič et al., 2009; Homola in Vičič, 2010; Vičič in Homola, 2010). Razdelek predstavlja najpomembnejše izvirne prispevke k znanosti, uvodni del obsega njihove krajše opise. Obširneje so prispevki predstavljeni v ločenih razdelkih, ki sledijo.

1. Metoda, za prevajanje s pomočjo dreves izpeljave za jezike z omejeno podporo jezikovnih tehnologij, jezike za katere ne obstaja standardna zbirka dreves izpeljave, treebank (Marcus et al., 1993).
Metoda je bila predstavljena v (Vičič in Brodnik, 2006) ter v (Vičič in Brodnik, 2008).
2. Metoda za samodejno označbo paradigem.
Prispevek je predstavljen v (Vičič, 2007a) in (Vičič, 2007b).
3. Samodejno luščenje paradigem za visoko pregibne jezike ter izdelava pripadajočih leksikonov
4. Ocenjevanje pravil za strukturni
 - uporaba ocenjevanja pravil.

- algoritmi za izbiro pravil.
- metrike ocenjevanja pravil.

Prispevek je predstavljen v (Vičič in Forcada, 2008).

5. Hitra izdelava prevajalnega sistema na osnovi RBMT za sorodne jezike V sistemu so implementirani ter preizkušeni opisani prispevki k znanosti. Izdelani so napotki za hitro izdelavo podobnih sistemov z drugimi jezikovnimi pari. Sistem temelji na odprtokodnih tehnologijah, tudi vse predstavljene in implementirane metode so ponujene z odprtokodno licenco v okviru projekta Apertium¹.

Prispevek je predstavljen v (Vičič, 2007a).

8.1.1 Umestitev pričakovanih prispevkov k znanosti

Slika 8.1 kaže eno od možnih razdelitev paradigme strojnega prevajanja. V ospredju so metode, ki slonijo na pravilih, te metode predstavljajo osnovo članka. Predstavljeni so še alternativni sistemi, ki jih lahko v grobem združimo v dve skupini: SMT ter EBMT. Sisteme sloneče na pravilih nadalje razdelimo na sisteme popolnega razčlenjevanja kot sta (Promt, 2010) in (Systran, 2010) ter sisteme plitkega razčlenjevanja, primera takšnih sistemov sta (Corbi-Bellot et al., 2005) ter (Homola in Kuboň, 2008a).

Metode napovedane v Razdelku 8.1 so oštevilčene, na Sliki 8.1 so cifre oštevilčenih metod zapisane ob opisih posameznih delov prevajalnih sistemov.

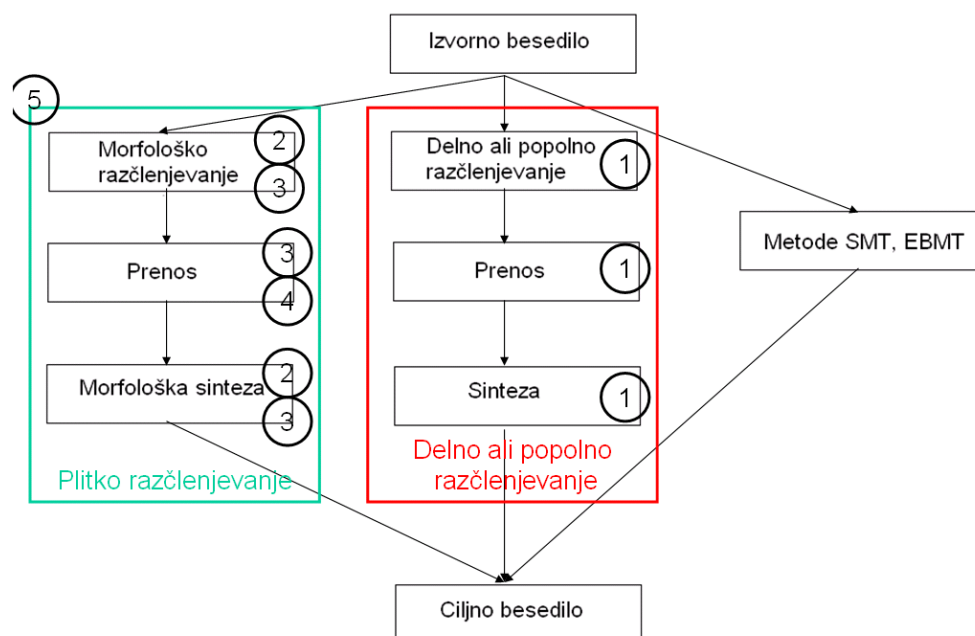
8.1.2 Metoda za statistično strojno prevajanje z drevesi izpeljave za manj uporabljane jezike (less-used languages)

Metoda omogoča izdelavo sistema za statistično strojno prevajanje z drevesi izpeljave za jezike, ki nimajo izdelane standardizirane zbirke dreves izpeljave (trebank), ki je običajna učna množica sistemov te paradigme. Metoda je obširnejše predstavljena v Poglavlju 7

8.1.3 Metoda za samodejno označevanje paradigem

Pravilna izbira sklanjatvenega vzorca omogoča enostavno sklanjanje besed z zelo omejenim številom pravil. Število sklanjatvenih vzorcev v naravnih jezikih je re-

¹Apertium: machine translation toolbox, <http://sourceforge.net/projects/apertium/index{apertium}/>



Slika 8.1: Ena od možnih razdelitev strojnega prevajanja z umestitvijo pričakovanih prispevkov k znanosti. Prispevki so predstavljeni z zaporednimi številkami.

lativno majhno in pravila lahko ročno vnesemo. Zavedati se moramo še velike množice izjem, ki jih ta metoda samo označi, prepozna jih pravilno kot izjeme. Predpogoj za metodo je dovolj dober in dovolj velik enojezični korpus. Oglejmo si primer za slovenski jezik, na primer slovenski jezik vsebuje po 4 osnovne sklanjatvene vzorce za vsak spol, 6 sklonov ter tri števila, skupaj $4 \cdot 6 \cdot 3 = 72$ enostavnih pravil, ki le zamenjujejo končnice. pravila so predstavljena v delu (Toporišič, 2000). Za slovenščino obstaja zelo kakovosten in velik po merilih, opisanih v (Rydberg-Cox et al., 2001), enojezični korpus FIDA (Erjavec et al., 1998) z nadaljevanjem FIDA (FidaPlus, 2008).

Besedi s pomočjo krnenja (Popovič in Willett, 1992) ter dodajanja končnic sklanjatvenih vzorcev sestavimo množico z dvojniki možnih sklanjanj. S preiskovanjem korpusa oziroma besed v korpusu izračunamo verjetnosti za posamezne sklanjatvene vzorce. Sklanjatveni vzorci z najvišjo oceno so dodatno preverjeni s pragovno funkcijo.

8.1.4 Samodejno luščenje paradigem za visoko pregibne jezike ter izdelava pripadajočih leksikonov

Samodejno luščenje paradigem iz označenega korpusa predstavlja poseben primer razvrščanja neznanih primerov v množice ter druženja množic. Razvit je algoritem za samodejno luščenje osnovnih paradigem na osnovi vnaprej označenega enojezičnega korpusa. paradigme so dodatno združene na osnovi podobnosti. Za preizkus metode je bil uporabljen korpus MULTEXT-EAST (Erjavec, 2010) ter pripadajoči leksikoni.

8.1.5 Ocenjevanje pravil za strukturni

Naloga raziskave ocenjevanja pravil strukturnega transferja je razdeljena na tri področja:

- raziskava možnih uporab ocenjevanja pravil;
- raziskava algoritmov za izbiro pravil;
- izdelava metrike za ocenjevanje pravil.

Vsako področje je natančneje predstavljeno v nadaljevanju.

8.1.5.1 Raziskava možnih uporab ocenjevanja pravil

Porajajo se naslednje možnosti uporabe ocenjevanja pravil, ki bodo ustrezno popravljene oziroma razširjene:

- Ocenjevanje obstoječih pravil, ki so jih ročno pripravili strokovnjaki (pravila, ki se uporabljajo v sedanjih prevajalnih sistemih, temelječih na ogrodju Apertium); to ocenjevanje nam omogoča vzpostavitev korelacije med ocenjevalnim modelom ter trenutno najboljšo možno izbiro pravil. Ta pravila so v praksi preverjena, saj jih uporabljajo pri prevajanju dnevnih izdaj časopisov. pravila, ki jih sistem oceni kot neprimerna, lahko, po pregledu strokovnjaka, ustrezno popravimo oziroma izbrišemo.
- Odkrivanje primernejših iskalnih algoritmov. V sistemu Apertium poteka izbira pravil po požrešni metodi algoritma najdaljšega možnega ujemanja iz leve proti desni (Left-to Right Longest Match Rule Selection - LRLM). Preiskovanje vseh možnih pokritij izvirne povedi (iskanje optimalnega pokritja) je zamudno.

- Ocenjevanje samodejno grajenih pravil. Metode za samodejno grajenje pravil (Sanchez-Martinez in Ney, 2006) in (Sanchez-Martinez in Forcada, 2007) izdelajo veliko število pravil, izbira najprimernejših pravil v določenih primerih je posebna domena ocenjevanja pravil.

8.1.5.2 Raziskava algoritmov za izbiro pravil

V sistemu Apertium poteka izbira pravil po požrešni metodi algoritma najdaljšega možnega ujemanja iz leve proti desni (Left-to Right Longest Match Rule Selection - LRLM). V večini primerov se ta algoritem lepo sklada s človeškim načinom tvorjenja povedi, v določenih primerih pa ta algoritem ne najde najboljše rešitve, najboljšega pokritja izvirne povedi. Odkrivanje takšnih primerov ter iskanje boljšega algoritma predstavlja poseben razdelek doktorskega dela.

8.1.5.3 Izdelava metrike za ocenjevanje pravil

V okviru doktorskega dela bo izdelana posebna metrika za ocenjevanje pravil strukturnega transferja. Preizkušena bo na ročno grajenih pravilih preizkušenega prevajalnega sistema (Forcada, 2006) ter na nepreizkušenih pravilih novega testnega sistema (Vičič in Forcada, 2008).

8.1.6 Hitra izdelava prevajalnega sistema na osnovi RBMT za sorodne jezike

Osnova sistema je Apertium, ki je predstavljen v Razdelku 3.4. Osnovni vodili snovanja sistema:

- Omogoča enostavno dodajanje novih metod ter preizkušanje njihove uporabnosti, kar omogoča enostavno in kar najbolj objektivno evalvacijo opisanih prispevkov k znanosti. Metode so bile preizkušene na dejanski uporabi in ne le v umetno ustvarjenih okoljih.
- Pri snovanju ter postavljanju sistema so bili izdelani napotki za hitro izdelavo podobnih sistemov z drugimi jezikovnimi pari.

8.2 Prevajalni sistem GUAT

Prevajalni sistem GUAT, ime je dobil po majhnih ribah Gobiidae, ki živijo tudi v slovenskem morju, je bil zgrajen med razvojem metod, prikazanih v Poglavjih 4 in

5. Sistem podpira jezikovni par slovenščino-srbščino. Metode so bile preverjene skozi več iteracij (sistematične napake so bile popravljene so vključeni v osnovno ogrodje). Ta jezikovni par je bil uporabljen za preverjanje kakovosti predstavljenih metod na popolnoma delujočem prevajalnem sistemu. Posebnosti jezikovnega para so: oba jezika sta zelo pregibna, morfološko in derivacijsko bogata. Čeprav sta jezika sorodna, visoka stopnja pregibnosti zahteva morfološko analizo izvornega jezika ter posledično morfološko sintezo v končni fazi v ciljni jezik.

Eden od najprivlačnejših razlogov za uporabo RBMT sistema za strojno prevajanje je sposobnost za strokovnjake s področja za nadaljnje izboljšanje samodejno izdelanih prevajalnih podatkov. Eksperti lahko enostavno dodajajo prevajalna previla ter popravljajo napake v morfologijah jezikov sistema ter v dvojezičnih prevajalnih slovarjih.

8.3 Nadaljnje delo

Delo predstavlja poskus združevanja večih metod za hitro postavitve prevajalnih sistemov za sorodne visoko pregibne jezike. Sistem temelji na paradigmi strojnega prevajanja na osnovi pravil plitkega prenosa, ki se je na pilotnih sistemih, predstavljenih v tem delu ter v mnogih znanstvenih člankih kot so (Corbi-Bellot et al., 2005), (Hajič et al., 2003), (Homola, 2010), (Scannell, 2006), izkazala kot najprimernejša za postavitev sistema za strojno prevajanje sorodnih jezikov. Metode so bile preizkušene na primeru samodejne izdelave prevajalnega sistema. Evalvacija kaže perspektivne rezultate, čeprav je možnost napredovanja še vedno dovolj velika.

Prav možnost izboljšave postavljenega prevajalnega sistema je ena največjih prednosti uporabljene tehnologije, saj eksplicitno zapisana pravila transferja ter slovarji omogočajo iterativno izboljševanje kakovosti prevodov.

Prikazane metode omogočajo hitrejšo izdelavo sistemov za strojno prevajanje za nove jezikovne pare in ena od možnosti nadaljevanja predstavljenega dela je postavitev pan-jugoslovanskega prevajalnega sistema, torej sistema za strojno prevajanje vsej uradnih jezikov bivše Jugoslavije, saj so si vsi ti jeziki med sabo sorodni.

Dodatek A

Prva priloga

A.1 Pravila transferja

Prikazana so pravila transferja za jezikovni par slovenščina - srbščina, smer SL → SR. Na Sliki A.1 je prikazano pravilo ujemanja pridevnika in samostalnika v sklonu, spolu in številu. Pridevniku pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe. Pravilo je predvsem uporabno pri napačnih izbirah modula za razdvoumljanje, pri sistemih z večimi možnimi kandidati za prevod olajšamo delo sistema za izbiro najboljšega prevoda (Ranker).

Na Sliki A.2 je prikazano pravilo ujemanja dveh pridevnikov in samostalnika v sklonu, spolu in številu. Pridevnikoma pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

Na Sliki A.3 je prikazano pravilo ujemanja zaimka in samostalnika v sklonu, spolu in številu, zaimku pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

```

<!--ujemanje pridevnika in samostalnika v sklonu, spolu
in številu, samostalnik je "glavni" -->
<!--primer: rdeče drevo-->
<rule>
  <pattern>
    <pattern-item n="adj"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_adjec"/>
        <clip pos="2" side="tl" part="gen"/>
        <clip pos="2" side="tl" part="nbr"/>
        <clip pos="2" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="adj_degree"/>
        <clip pos="2" side="tl" part="adj_definitness"/>
      </lu>
      <b pos="1"/>
        <lu>
          <clip pos="2" side="tl" part="lemh"/>
          <clip pos="2" side="tl" part="a_nom"/>
          <clip pos="2" side="tl" part="gen"/>
          <clip pos="2" side="tl" part="nbr"/>
          <clip pos="2" side="tl" part="sklon"/>
        </lu>
      </out>
    </action>
  </rule>

```

Slika A.1: Pravilo: ujemanje pridevnika in samostalnika v sklonu, spolu in številu, pridevniku pripišemo iste kategorije kot jih ima samostalnik.

```

<!--ujemanje dveh pridevnikov s samostalnikom v sklonu, spolu
in številu, samostalnik je "glavni" -->
<!--primer: mlado rdeče drevo-->
<rule>
  <pattern>
    <pattern-item n="adj"/>
    <pattern-item n="adj"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_adjec"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="1" side="tl" part="adj_degree"/>
        <clip pos="1" side="tl" part="adj_definitness"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_adjec"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="2" side="tl" part="adj_degree"/>
        <clip pos="2" side="tl" part="adj_definitness"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lemh"/>
        <clip pos="3" side="tl" part="a_nom"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.2: Pravilo: ujemanje dveh pridevnikov in samostalnika v sklonu, spolu in številu, pridevnikoma pripišemo iste kategorije kot jih ima samostalnik.

```

<!--ujemanje zaimka in samostalnika -->
<!--primer: moj avto-->
<rule>
  <pattern>
    <pattern-item n="prn"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>
    <call-macro n="f_concord2">
      <with-param pos="2"/>
      <with-param pos="1"/>
    </call-macro>
    <call-macro n="f_lexicAdj">
      <with-param pos="1"/>
    </call-macro>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_prn"/>
        <clip pos="2" side="tl" part="gen"/>
        <clip pos="2" side="tl" part="nbr"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_nom"/>
        <clip pos="2" side="tl" part="gen"/>
        <clip pos="2" side="tl" part="nbr"/>
        <clip pos="2" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.3: Pravilo: ujemanje zaimka in samostalnika v sklonu spolu in številu, zaimku pripišemo iste kategorije kot jih ima samostalnik.

Na Sliki A.4 je prikazano pravilo ujemanje zaimka, pridevnika in samostalnika v sklonu spolu in številu, zaimku in pridevniku pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

Na Sliki A.5 je prikazano pravilo ujemanje zaimka, pridevnika in samostalnika v sklonu spolu in številu, zaimku in pridevniku pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

Na Sliki A.6 je prikazano pravilo za prenos delov povedi, ki tvorijo prihodnjik. V slovenščini tvorimo prihodnjik s pomožnim glagolom *biti* v prihodnjiku ter deležnikom na l. V srbsščini pa s pomožnim glagolom *hteti* ter nedoločno obliko glagola. Zaradi poenostavitve sistem namesto deležnika na l označi glagol v slovenščini kot navadni glagol v pretekliku, pravilo pa poišče poljubno obliko glagola ter jo spremeni v nedoločnik. Lema pomožnega glagola *biti* se prevede v lemo pomožnega glagola *hteti*, čas pa se spremeni v sedanjik. V komentarju je zapisan primer uporabe.

```

<!--ujemanje zaimka, pridevnika in samostalnika (pomemben
zaimek in samostalnik) -->
<!--primer: moj lepi avto-->
<rule>
  <pattern>
    <pattern-item n="prn"/>
    <pattern-item n="adj"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>
    <call-macro n="f_concord2">
      <with-param pos="1"/>
      <with-param pos="2"/>
      <with-param pos="3"/>
    </call-macro>
    <call-macro n="f_lexicAdj">
      <with-param pos="1"/>
    </call-macro>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_prn"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_adjec"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
        <clip pos="2" side="tl" part="adj_degree"/>
        <clip pos="2" side="tl" part="adj_definitness"/>
      </lu>
      <b pos="2"/>
      <lu>
        <clip pos="3" side="tl" part="lemh"/>
        <clip pos="3" side="tl" part="a_nom"/>
        <clip pos="3" side="tl" part="gen"/>
        <clip pos="3" side="tl" part="nbr"/>
        <clip pos="3" side="tl" part="sklon"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.4: Pravilo: ujemanje zaimka, pridevnika in samostalnika v sklonu spolu in številu, zaimku in pridevniku pripišemo iste kategorije kot jih ima samostalnik.

```

<!--ujemanje samostalnika in navadnega glagola v spolu
in številu -->
<!--primer avto je vozil -->
<rule>
  <pattern>
    <pattern-item n="nom"/>
    <pattern-item n="vbser"/>
    <pattern-item n="vblex"/>
  </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="a_vblex"/>
        <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.5: Pravilo: ujemanje samostalnika in navadnega glagola v spolu in številu glagolu pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.

```

    <!-- prihodnjik 1-->
    <!-- primer: kupil bom -> kupiti ću ali kupiću-->
<rule>
  <pattern>
    <pattern-item n="vblex"/>
    <pattern-item n="vbserfti"/>
  </pattern>
  <action>
    <let>
      <clip pos="2" side="tl" part="lemh"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="2" side="tl" part="temps"/>
      <lit-tag v="pres"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="temps"/>
      <lit-tag v="inf"/>
    </let>

    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_vblex"/>
        <clip pos="1" side="tl" part="temps"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="2" side="tl" part="a_vbser"/>
        <clip pos="2" side="tl" part="temps"/>
        <clip pos="2" side="tl" part="persona"/>
        <clip pos="2" side="tl" part="nbr"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.6: Pravilo: prva oblika prihodnjika.

Na Sliki A.7 je prikazano pravilo za prenos delov povedi, ki tvorijo prihodnjik s svojilnim zaimkom *si*. V slovenščini tvorimo prihodnjik s pomožnim glagolom *biti* v prihodnjiku ter deležnikom na l. V srbščini pa s pomožnim glagolom *hteti* ter nedoločno obliko glagola. Zaradi poenostavitve sistem namesto deležnika na l označi glagol v slovenščini kot navadni glagol v pretekliku, pravilo pa poišče poljubno obliko glagola ter jo spremeni v nedoločnik. Lema pomožnega glagola *biti* se prevede v lemo pomožnega glagola *hteti*, čas pa se spremeni v sedanjik. Svojilni zaimek se ne spreminja. V komentarju je zapisan primer uporabe.

Na Sliki A.8 je prikazano pravilo za prenos nedoločnika. Nedoločnik iz slovenščine v srbščino prevedemo kot členek *da*, ki mu sledi glagol v sedanjiku. Pravilo prevede vzorec navadnega glagola v poljubni obliki, ki mu sledi glagol v nedoločniku, v členek *da*, ki mu sledi glagol v istem času kot je prvi glagol iz izvirnega dela - slovenščine. V komentarju je zapisan primer uporabe.

```

<!-- prihodnjik 2 -->
<!-- primer: bom si kupil -> kupiti ću si ali kupiću si-->
<rule>
  <pattern>
    <pattern-item n="vbserfti"/>
    <pattern-item n="prnalone"/>
    <pattern-item n="vblex"/>
  </pattern>
  <action>
    <let>
      <clip pos="1" side="tl" part="lemh"/>
      <lit v="hteti"/>
    </let>
    <let>
      <clip pos="1" side="tl" part="temps"/>
      <lit-tag v="pres"/>
    </let>
    <let>
      <clip pos="3" side="tl" part="temps"/>
      <lit-tag v="inf"/>
    </let>
    <out>
      <lu>
        <clip pos="1" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_vbser"/>
        <clip pos="1" side="tl" part="temps"/>
        <clip pos="1" side="tl" part="persona"/>
        <clip pos="1" side="tl" part="nbr"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="3" side="tl" part="lemh"/>
        <clip pos="3" side="tl" part="a_vblex"/>
        <clip pos="3" side="tl" part="temps"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.7: Pravilo: druga oblika prihodnjika.

```

<!-- Namesto infinitiva v srb = da + sedanjik -->
<!-- primer: On želi delati (slo) -> On želi da radi (srb) -->
<rule>
  <pattern>
    <pattern-item n="vblex"/>
    <pattern-item n="vblexinf"/>
  </pattern>
  <action>
    <out>
      <lu>
        <clip pos="1" side="tl" part="whole"/>
      </lu>
      <b pos="1"/>
      <lu>
        <lit v="da"/>
        <lit-tag v="part"/>
      </lu>
      <b pos="1"/>
      <lu>
        <clip pos="2" side="tl" part="lemh"/>
        <clip pos="1" side="tl" part="a_vblex"/>
        <clip pos="1" side="tl" part="temps"/>
        <clip pos="1" side="tl" part="persona"/>
        <clip pos="1" side="tl" part="nbr"/>
      </lu>
    </out>
  </action>
</rule>

```

Slika A.8: Pravilo: nedoločnik.

A.2 Primeri prevodov

A.2.1 Dobri prevodi

Primeri A.2 do A.1 kažejo dobre prevode, ki ne potrebujejo dodatnih komentarjev.

(A.1) *Danes je lepo vreme.*

Danas je lepo vreme.

(A.2) *Danes je lepo vreme.*

Danas je lepo vreme.

JERNEJ

A.2.2 Napake

JERNEJ

Slike

2.1	Drevo izpeljav za stavek <i>Rdeči avto je vozil</i> . S - stavek, N - samostalnik(noun), V - glagol(verb), Vm - pomožni glagol(modal verb), NP - samostalniška fraza(noun-phrase, VP - glagolska fraza(verb phrase).	9
2.2	pravilo za prevod dela povedi v prihodnjiku. Pomožni glagol biti v prihodnjiku pri prevodu spremeni lemo v hteti ter čas v sedanjik, glavni glagol pri prevodu iz preteklika preide v nedoločnik.	16
3.1	Moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Ta arhitektura je bila najprej predstavljena v (Hajič et al., 2000) in pozneje uporabljena tudi v (Corbi-Bellot et al., 2005)	24
3.2	Arhitektura ogrodja Apertium: poleg osnovnih modulov, ki služijo za osnovno prevajanje in so prikazani na Sliki 3.1, Apertium dodaja še module za označevanje delov besedila, ki se ne prevajajo ter modul za končno urejanje (post-editing) prevodov.	28
3.3	Morfološka analiza stavka "Danes je lepo vreme.". Besede izvirne povedi so označene z vsemi možnimi ustreznimi morfološkimi oznakami iz slovarja. Najprej je zapisana besedna oblika, sledijo vse možne oznake za to besedno obliko. Za besedno obliko <i>lepo</i> je možnih pet različnih množic oznak.	29
3.4	Moduli predlaganega (spremenjenega) sistema za strojno prevajanje na osnovi pravil plitkega prenosa. Arhitektura temelji na sistemu predlaganem v (Corbi-Bellot et al., 2005; Hajič et al., 2003) brez uporabe sistema za razdvoumljanje na osnovi označevalnika POS in uporabo vseh kandidatov za prevode do zadnjih faz prevajalne verige ter z dodatkom modula za izbiro najboljšega prevoda (Ranker).	31
4.1	Del zapisov v enojezičnem slovarju. Lema <i>cerkev</i> je predstavljena z lemo, krnom ter paradigmo.	36

4.2	Del paradigme za samostalnike ženskega spola v slovenščini. Tipični predstavnik je lema <i>cerkev</i> . Končnica <i>-ev</i> se spreminja v skladu z različnimi MSD.	37
4.3	Primeri dvojezičnih prevodov lem iz slovenščine v srbsščino.	38
4.4	Označena poved v korpusu (Erjavec, 2010).	39
4.5	Del paradigme <i>cerk-ev</i> . Lema: <i>cerkev</i> , krn: <i>cerk</i> , dve besedni obliki <i>cerkev</i> in <i>cerkvah</i>	41
4.6	Algoritem za gradnjo paradigem	41
4.7	Besedni obliki se ne ujemata, kar pomeni, da paradigmi ne združimo.	42
4.8	Pripravljeni učni podatki: leme in besedne vrste za vsako besedo v korpusu.	43
4.9	Zmanjšanje iskalnega prostora za slovenski jezik (relativno majhen korpus MULTEXT-EAST (Erjavec, 2010))	44
4.10	Izločitev (morebiti) vseh nemogočih kandidatov za prevode z uporabo pravil lokalnega ujemanja.	45
5.1	Splošna shema sistema za strojno prevajanje s transferjem, sistem ima dve vmesni predstavitvi besedila: izvorna ter ciljna vmesna predstavitev, med njima poteka	47
5.2	Primer pravila za strukturni pravilo opisuje spremembe načina zapisa prihodnjika iz slovenščine v srbsščino.	49
5.3	Proces samodejne izdelave pravil iz označenega korpusa.	52
5.4	pravilo ujemanja pridevnika in samostalnika, ki si sledita. Besedi se morata ujemati v spolu, sklonu ter številu. Pri prevajanju se spreminjajo morfološke kategorije samostalnika in ne pridevnika, zato je ujemanje vezano na samostalnik.	53
5.5	pravilo ujemanja samostalnika, pomožnega glagola leme <i>jesam</i> - biti ter glagola. Pomožni glagol ter samostalnik se ujemata v številu, samostalnik in glagol na tretjem mestu se ujemata v spolu in številu.	55
6.1	Kandidat za prevod ter referenčni prevod.	56
6.2	Rezultati evalvacije z metriko METEOR. Uporabili smo 5-kratno prečno preverjanje. Vrednosti predstavljajo povprečja 5 iteracij s standardno deviacijo. Evalvacije označene z zvezdico * predstavljajo uporabo krnjenja z algoritmom Porter-stem, ostala pa z uporabo lastnega algoritma za krnjenje.	62
6.3	Rezultati evalvacije s pomočjo metrike Word Recognition Rate (WRR).	63

6.4	Rezultati evalvacije po smernicah (LDC, 2005). Povprečne vrednosti štirih neodvisnih ocenjevanj kažejo visoke vrednosti za vsebinsko ustreznost prevodov (adequacy) in nižje vrednosti za slovnično pravilnost.	64
7.1	Primer drevesa izpeljave.	68
7.2	Primer poravnave morfoloških oznak (POS) z drevesom izpeljave. .	69
8.1	Ena od možnih razdelitev strojnega prevajanja z umestitvijo pričakovanih prispevkov k znanosti. Prispevki so predstavljeni z zaporednimi številkami.	72
A.1	Pravilo: ujemanje pridevnika in samostalnika v sklonu, spolu in številu, pridevniku pripišemo iste kategorije kot jih ima samostalnik.	77
A.2	Pravilo: ujemanje dveh pridevnikov in samostalnika v sklonu, spolu in številu, pridevnikoma pripišemo iste kategorije kot jih ima samostalnik.	78
A.3	Pravilo: ujemanje zaimka in samostalnika v sklonu spolu in številu, zaimku pripišemo iste kategorije kot jih ima samostalnik.	79
A.4	Pravilo: ujemanje zaimka, pridevnika in samostalnika v sklonu spolu in številu, zaimku in pridevniku pripišemo iste kategorije kot jih ima samostalnik.	81
A.5	Pravilo: ujemanje samostalnika in navadnega glagola v spolu in številu glagolu pripišemo iste kategorije kot jih ima samostalnik. V komentarju je zapisan primer uporabe.	82
A.6	Pravilo: prva oblika prihodnjika.	83
A.7	Pravilo: druga oblika prihodnjika.	85
A.8	Pravilo: nedoločnik.	86

Tabele

2.1	Primeri lažnih prijateljev, podobnih besed v različnih jezikih z različnimi pomeni.	18
4.1	Vse besedne oblike za slovensko lemo mesto	38
4.2	Primerjava število lem s številom besednih oblik v korpusu MULTEXT-EAST (Erjavec, 2010)	43
6.1	Cohenov kapa koeficient (Cohen, 1960) za sistema SL-SR in SL-CS kaže zadovoljivo (satisfactory) (SL-CS) ter zelo visoko (very-high) (Sl-SR) ujemanje med ocenjevalci (inter-rater agreement).	65

Literatura

- Lars Ahrenberg in Maria Holmqvist. Back to the future? the case for english-swedish direct machine translation. In *Paper presented at the Conference on Recent Advances in Scandinavian Machine Translation*, 2005.
- Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Laerty, Dan Mela-med, Franz-Josef Och, David Purdy, Noah A. Smith, in David Yarowsky. Statistical machine translation, final report. Technical report, JHU, 1999.
- K. Altintas in I. Cicekli. A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, 2002.
- Apertium. Apertium: machine translation toolbox, 2010. URL <http://sourceforge.net/projects/apertium>.
- Dough Arnold. *Computers and Translation: A translator's guide*. Benjamin Translation Library, 2003.
- L.R. Bahl, P.F. Brown, P.V. de Souza, in R.L. Mercer. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7), 1989.
- S. Banerjee in A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the ACL*. Ann Arbor, Michigan, 2005.
- Thorsten Brants. TnT—a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*. Seattle, WA, 2000.
- Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, in François Yergeau. Extensible Markup Language (XML) 1.0 (Fifth Edition). Technical report, W3C, 2008.

- Peter Brown, Peter Cocke, Stephen Della Pietra, Vincent Della Pietra, Fredrik Jelinek, John Lafferty, Robert Mercer, in Paul S. Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2), 1994.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, in Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):163–311, 1993.
- A. Burbank, M. Carpuat, S. Clark, M. Dreyer, P. Fox, D. Groves, K. Hall, M. Hearne, D. Melamed, Y. Shen, B. Wellington A. Way, in D. Wu. Final report of the 2005 language engineering workshop on statistical machine translation by parsing. Technical report, JHU, <http://www.clsp.jhu.edu/ws2005/groups/statistical/documents/finalreport.pdf>, 2005. URL <http://www.clsp.jhu.edu/ws2005/groups/statistical/documents/finalreport.pdf>.
- Chris Callison-Burch, Miles Osborne, in Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of EACL*, 2006.
- Eugene Charniak. A maximum-entropy-inspired parser. In *ANLP*, pages 132–139, 2000.
- P.R. R. Rosenfeld Clarkson. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings ESCA Eurospeech*, 1997.
- Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.
- Michael Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, 2003.
- TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical report, TEI consortium, 2007.
- 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 EAMT conference*, pages 79–86, May 2005.
- Mathias Creutz. Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition. Technical report, Helsinki University of Technology, 2006.

- Ludmila Dimitrova, Nancy Ide, Vladimir Petkevič, Tomaž Erjavec, Heiki Jaan Kalep, in Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL*, pages 315–319, 1998.
- H. Dyvik. Exploiting structural similarities in machine translation. *Computers and Humanities*, 28:225–245, 1995.
- EAMT. European association for machine translation, 2010. URL <http://www.eamt.org/>.
- EGYPT. The egypt statistical machine translation toolkit, 2007. URL <http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/>.
- Katherine Eng, Alex Fraser, Daniel Gildea, Viren Jain, Zhen Jin, Shankar Kumar, Sanjeev Khudanpur, Franz Och, Dragomir Radev, Anoop Sarkar, Libin Shen, David Smith, in Kenji Yamada. Final report, syntax for statistical mt group. Technical report, JHU, 2003.
- T. Erjavec, V. Gorjanc, in M. Stabej. Korpus fida. In *Proceedings of International Multi-Conference Information Society - IS'98*, 1998.
- Tomaž Erjavec. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*, 2004.
- Tomaž Erjavec. Multilingual tokenisation, tagging, and lemmatisation with totale. In *Proceedings of the 9th INTEX/NOOJ Conference*, 2006.
- Tomaž Erjavec. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *LREC*, 2010.
- Tomaž Erjavec in Saša Džeroski. Machine learning of language structure: Lemmatising unknown slovene words. *Applied Artificial Intelligence*, 18, 2004.
- FidaPlus. The fidaplus corpus, 2008. URL <http://www.fidaplus.net>.
- Mikel L. Forcada. Open-source machine translation: an opportunity for minor languages. In *Strategies for developing machine translation for minority languages (5th SALT MIL workshop on Minority Languages)*, 2006.
- GenPar. The genpar toolkit for research on generalized parsing, 2010. URL <http://nlp.cs.nyu.edu/GenPar/>.

- GNU. Gnu general public license, 2010. URL <http://www.gnu.org/licenses/index%5Fhtml#GPL>.
- John Goldsmith. Unsupervised learning of the morphology of a natural language. In *Proceedings of ACL 2001*, 2001.
- J. Hajič. An mt system between closely related languages. In *Proceedings of the third conference of the European Chapter of the Association for Computational Linguistics*, 1987.
- J. Hajič, J. Hric, in V. Kuboň. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, 2000.
- Jan Hajič. Morphological tagging: data vs. dictionaries. In *Proceedings of the North American chapter of the Association for Computational Linguistics conference*, 2000.
- Jan Hajič, Petr Homola, in Vladislav Kuboň. A simple multilingual machine translation system. In *Proceedings of the MT Summit IX*, New Orleans, 2003.
- Morris Halle in P. Alec. *Distributed Morphology and the pieces of inflection*, pages 111–176. Cambridge, MA: MIT Press, 1993.
- Petr Homola. *Syntactic analysis in machine translation*. Studies in Computational and Theoretical Linguistics. Institute of Formal and Applied Linguistics, 2010.
- Petr Homola in Vladislav Kuboň. A method of hybrid MT for related languages. In *Proceedings of IIS*, 2008a.
- Petr Homola in Vladislav Kuboň. Improving machine translation between closely related romance languages. In *Proceedings of EAMT*, pages 72 – 77, 2008b.
- Petr Homola in Jernej Vičič. Combining MT systems effectively. In *Proceedings of the 23th International Florida-Artificial-Intelligence-Research-Society Conference (FLAIRS 2010)*, pages 198–203, Daytona Beach, Florida, USA, 2010. Florida AI Research Society, Florida AI Research Society.
- Petr Homola, Vladislav Kuboň, in Jernej Vičič. *Shallow Transfer Between Slavic Languages*, page 219–232. Academic publishing house EXIT, Warsaw, 2009.
- W. J. Hutchins in H. L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.

- Laura Janda. Inflectional morphology. In Dirk Geeraerts in Hubert Cuyckens, editors, *Handbook of Cognitive Linguistics*, pages 632–649. Oxford: Oxford U Press, 2007.
- Philipp Koehn, Franz Josef Och, in Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL*, 2003.
- 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)*, 2007.
- Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143, 1995.
- Andras Kornai. *Extended Finite State Models of Language*. Cambridge University Press, 1999.
- Gorka Labaka, Nicholas Stroppa, Andy Way, in Kepa Sarasola. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. In *Proceedings of the Machine Translation Summit XI*, pages 41–48, 2007.
- A. Lavie in A. Agarwal. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of Workshop on SMT at the ACL conference*, 2007.
- LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, LDC, 2005.
- GN Leech in A. Wilson. Eagles recommendations for the morphosyntactic annotation of corpora. Technical report, ILC-CNR, Pisa, 1996.
- V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, pages 845–848, 1965.
- Rochelle Lieber. *Deconstructing Morphology: Word Formation in Syntactic Theory*. University of Chicago Press, 1992.
- Mitchell P. Marcus, Beatrice Santorini, in Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

- I. Dan Melamed. Statistical machine translation by parsing. In *ACL*, pages 653–660, 2004a.
- I. Dan Melamed. Algorithms for syntax-aware statistical machine translation. In *TMI*, 2004b.
- Vesna Mikolič, Jernej Vičič, in Jana Volk. *Namen in metode urejanja večjezičnega korpusa turističnih besedil (TURK)*, pages 65–74. Znanstveno-raziskovalno središče, Založba Annales, 2009.
- Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and Human Intelligence*, 1984.
- Franz Josef Och. Challenges in Machine Translation. In *Proceedings of ISCSLP*, 2006.
- Franz Josef Och in Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29:19–51, 2003.
- Franz Josef Och in Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, pages 417–449, 2004.
- George Orwell. 1984. Secker and Warburg, London, 1949.
- Kishore Papineni, Salim Roukos, Todd Ward, in Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM, 2001.
- M. Popovič in P. Willett. The effectiveness of stemming for natural language access to slovene textual data. *Journal of the American Society for Information Science*, 43(5), 1992.
- M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.
- Prompt, 2010. URL <http://www.e-prompt.com/>.
- Emmanuel Roche in Yves Schabes. *Finite-State Language Processing*. MIT Press, 1997.
- Jeffrey A. Rydberg-Cox, Anne Mahoney, in Gregory Crane. Document quality indicators and corpus editions. In *JCDL*, pages 435–436, 2001.
- Benot Sagot. Automatic acquisition of a slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658, proceedings of TSD'05*, pages 156–163, 2005.

- Felipe Sanchez-Martinez in Mikel L. Forcada. Automatic induction of shallow-transfer rules for open-source machine translation. In Andy Way in Barbara Gawronska, editors, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, volume 2007:1, pages 181–190. Skovde University, September 2007. ISBN 978-91-977095-0-7.
- Felipe Sanchez-Martinez in Mikel L. Forcada. Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, 34:605–635, 2009.
- 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*, pages 756–767. Springer-Verlag, August 2006. ISBN 3-540-37334-9. 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*, pages 73–82, January 2007.
- Edward Sapir. *Language: An Introduction to the Study of Speech*. Harcourt, Brace New York:, 1921.
- K. P. Scannell. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, 2006.
- Andrew Spencer. *Morphological Theory*. Blackwell Publishing, 1991.
- Systran. Systran, 2010. URL <http://www.systran.co.uk/>.
- 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.
- Jože Toporišič. *Slovenska slovnica*. Založba Obzorja, Maribor, 2000.

- Charles E. Townsend in Laura A. Janda. *Gemeinslavisch und slavisch im vergleich. Einführung in die Entwicklung von Phonologie und Flexion.*, 2003.
- Jernej Vičič in Petr Homola. Speeding up the implementation process of a shallow transfer machine translation system. In *Proceedings of the 14th EAMT Conference*, pages 261–268, Saint Raphael, France, 2010. European Association for Machine Translation.
- Jernej Vičič, Petr Homola, in Vladislav Kuboň. A method to restrict the blow-up of hypotheses of a non-disambiguated shallow machine translation system. In *RANLP*, pages 1–8, Borovec, Bulgaria, 2009. ISBN 978-954-452-012-0.
- Jernej Vičič. Rapid development of RBMT systems for related languages. In *Translating and the computer 29 : proceedings of the twenty-ninth international conference on translating and the computer*, pages 162–1733, 2007a.
- Jernej Vičič. Rapid development of rbmt systems for related languages, a case study on language pair slovenian - serbian. In *Zb. Elektrotehn. racun. konf. ERK, zv. B*, pages 95–98, 2007b.
- Jernej Vičič. Rapid development of data for shallow transfer rbmt translation systems for highly inflective languages. In *Language technologies : proceedings of the conference*, pages 98–103, 2008.
- Jernej Vičič. *Metode hitre izdelave gradiv za prevajalne sisteme plitkega prenosa za visoko pregibne jezike*, pages 133–153. Znanstveno-raziskovalno središče, Založba Annales, 2009.
- Jernej Vičič in Andrej Brodnik. A method for statistical machine translation by parsing for less-used languages, 2006.
- Jernej Vičič in Andrej Brodnik. A method for statistical machine translation by parsing for less-used languages. *Advances in Methodology and Statistics*, 1, 2008.
- Jernej Vičič in Tomaž Erjavec. Vsak začetek je težak : avtomatsko učenje prevajanja slovenščine v angleščino. In *Language technologies, proceedings of the conference*, pages 20–27, 2002.
- Jernej Vičič in Mikel L. Forcada. Comparing greedy and optimal coverage strategies for shallow-transfer machine translation. In *Intelligent information systems XVI : proceedings of the International IIS '08 conference*, pages 307–316, 2008.

- Stephan Vogel, Franz Josef Och, in Hermann Ney. The statistical translation module in the verbmobil system. In *KONVENS*, pages 291–293, 2000.
- Lloyd R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):1–14, 2003.
- Dekai Wu. Mt model space: statistical versus compositional versus example-based machine translation. *Machine Translation*, 19(3-4):213–227, 2005.
- Wolfgang U. Wurzel. Paradigmenstrukturbedingungen: Aufbau und veränderung von flexionsparadigmen. In *Proceedings of the 7th International Conference on Historical Linguistics*, pages 629–644, 1987.
- Łukasz Dębowski, Jan Hajič, in Vladislav Kuboň. Testing the limits – adding a new language to an MT system. *The Prague Bulletin of Mathematical Linguistics*, (78):95–101, 2002. ISSN 0032-6585.

Stvarno kazalo

- algoritem, 50, 57, 67, 73, 74
- Apertium, 5, 23, 27, 28, 35, 40, 45, 71, 73, 74
- bigram, 52
- BLEU, 56, 57, 60, 61
- drevo izpeljave, 66–68
- GUAT, 4, 74
- LDC, 58
- luščenje, 8, 70, 73
- METEOR, 57, 60–62
- morfoloških, 33
- morfologija, 7, 23
- paradigma, 5, 7, 8, 15, 19, 20, 22, 35–37, 40–42, 61, 70, 71, 73
- plitki transfer, 4, 10, 26, 35, 50, 54
- podobnost, 13–15, 17, 23, 52, 54, 69
- pravilo, 1, 3, 4, 8, 10, 15–17, 19, 20, 22–24, 26, 27, 31, 33, 36, 42, 44–55, 57, 68, 70–74
- prevajalni sistem, 19, 20, 26
- prevod, 3, 4, 11, 14–17, 19, 21–23, 25, 27–39, 44, 45, 48, 51, 52, 56–64, 67–69
- slovenščina, 4, 59, 60, 76, 101
- slovenščina, 9, 17, 32, 43, 47, 52
- sorodni jezik, 5, 15, 25–27, 32, 71, 74
- srbsščina, 11–13, 17, 43, 47, 52, 59, 60, 76
- transfer, 4, 10, 22–24, 26, 27, 33, 46–51, 70, 73
- trigram, 52
- ujemanje, 10, 13, 17, 33, 45, 47, 50–57, 64, 65, 73, 74