original scientific article
received: 2014-12-05

DOI 10.19233/ASHS.2016.57

JEZIKOVNI VIRI ZA PREVAJALNE SISTEME

Jernej VIČIČ
Univerza na Primorskem, Inštitut Andreja Marušiča, Muzejski trg 2, 6000 Koper
e-mail: jernej.vicic@upr.si

IZVLEČEK

V članku so predstavljeni jezikovni viri sistema za strojno prevajanje naravnih jezikov jezikovnega para slovenščina – hrvaščina. Prikazan je sistem za strojno prevajanje z vsemi pripadajočimi jezikovnimi viri. Za vsak vir je opisana metoda izdelave, tako osnovna samodejna metoda kot načini ročnega čiščenja izdelanih virov. Evalvacija je imela dva osnovna cilja: evalvacija kakovosti prevodov prevajalnega sistema ter evalvacija velikosti in kakovosti posameznih jezikovnih virov. Vsa opisana gradiva in tudi celoten sistem so prosto dostopni.

Ključne besede: strojno prevajanje naravnih jezikov, oblikoskladenjski slovar, pravilo prevoda, paradigma, lema

MATERIALI LINGUISTICI PRODOTTI PER SISTEMI DI TRADUZIONE AUTOMATICA

SINTESI

Nell'articolo viene presentato il materiale linguistico per il sistema di traduzione automatica di linguaggi naturali per la coppia di lingue sloveno-croato. Si presenta il sistema di traduzione automatica affiancato dalla completa documentazione linguistica inerente. Viene inoltre descritta la modalità di realizzazione del materiale stesso, data non soltanto dalla metodologia di base automatica, ma anche dai procedimenti manuali di selezione del materiale prodotto. La valutazione ha avuto due obiettivi fondamentali: la valutazione della qualità delle traduzioni ottenute mediante il sistema di traduzione, e la valutazione dell'ampiezza e della qualità del materiale linguistico. Tutta la documentazione descritta nonché l'intero sistema sono liberamente accessibili.

Parole chiave: traduzione automatica di linguaggi naturali, vocabolario morfosintattico, vocabolario, regole di traduzione, paradigma, lemma

UVOD

Razlogov za postavitev prevajalnega sistema za opisani jezikovni par je več. Omeniti jih velja vsaj nekaj.

Dejstvo je, da sta ekonomiji držav, v okviru teh pa še zlasti turizem na območju slovenske in hrvaške Istre (to velja pravzaprav za vse obmejne regije (SVLR, 2006), v zadnjem času vedno bolj povezani. Poleg tega pa nas povezujejo skupne zgodovinske, gospodarske, kulturne in družbene značilnosti.

Izhodišče za izdelavo prevajalnega sistema za opisani jezikovni par je med drugim tudi načelo Sveta Evrope o jezikovni raznolikosti (Jagland in Vassiliou, 2011), prevajalni sistem pa oblikovan na primer za potrebe gospodarstva na eni ter čezmejnih projektov s področja kulturne in naravne dediščine na drugi strani. Še zlasti pa nas je pri postavljanju prevajalnega sistema vodilo dejstvo, da ga pri svojem raziskovanju ter medkulturnem in čezmejnem sodelovanju lahko koristno izrabijo prav raziskovalci različnih področij tako naravoslovja in družboslovja kot tudi humanistike. Kot primer naj navedem prevajalni sistem kot pripomoček pri snovanju in vodenju različnih bilateralnih in mednarodnih projektov (npr. Interreg). Prav tako se lahko tovrstna orodja uporabljajo na področju izobraževanja.

Glede na to, da imajo mlajše generacije, generacije po razpadu Jugoslavije, pri medsebojni komunikaciji jezikovne težave, lahko takšna orodja presežejo razmejenost med državama ter nedvomno olajšajo komunikacijo med sorodnima jezikoma, še posebej na obmejnih območjih, kjer je stikov veliko več in, posledično, komunikacija pogostejša.

Gradiva, ki sestavljajo prevajalni sistem in so opisana v 3. razdelku, se lahko uporabijo na področjih humanistike ter kulturne in naravne dediščine, in sicer pri izdelavi področnih glosarjev in terminoloških slovarjev, zgrajenih s pomočjo oblikoskladenjskih slovarjev obeh jezikov, ter dvojezičnih terminoloških slovarjev, zgrajenih s pomočjo dvojezičnega slovarja prevajalnega sistema.

Članek je strukturiran takole: prvi razdelek predstavlja raziskovalno domeno in predstavi osnovne pojme. Drugi razdelek prek predstavitve osnovnih pojmov uvede bralca v raziskovalno domeno. Sledi opis posameznih jezikovnih gradiv in metodologije samodejne izdelave ter ročnega popravljanja teh gradiv v tretjem razdelku. Četrti razdelek predstavlja metode ter rezultate evalvacije jezikovnih gradiv in prevajalnega sistema. Članek se zaključi s predstavitvijo načinov dostopnosti prevajalnega sistema in jezikovnih gradiv ter opisom smernic za nadaljnje delo.

Strojno prevajanje

Pregled strojnega prevajanja (Sanchez-Martnez et al., 2007) deli področje na dve skupini: prevajanje s pomočjo pravil (*Rule-Based* – RBMT) in prevajanje na osnovi korpusov (*Corpus-Based* – CBMT).

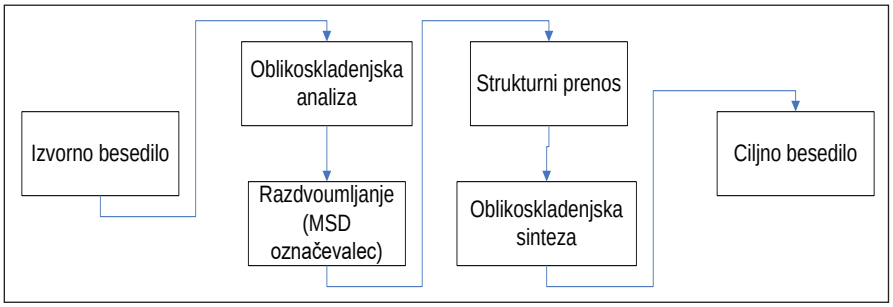
- RBMT obsega sisteme in metode za prevajanje s pomočjo zbirke pravil. Način zapisa pravil se med sistemi razlikuje, veže pa jih skupno dejstvo, da je postavitev takšnega sistema dolgotrajno opravilo. Primeri sistemov: Presis,[1] Systran,[2] Promt,[3] Apertium.[4]
- CBMT obsega sisteme, ki sledijo naslednjemu vzoru: pripravljena je množica referenčnih prevodov, ki so analizirani in prevedeni v modele prevajalnega sistema po načelih, ki določajo prevajalni sistem (faza učenja). Ti modeli služijo kot osnova za poznejše prevode neznanih povedi (faza prevajanja). Najbolj razširjena paradigma med sistemi CBMT je statistično strojno prevajanje (*Statistical Machine Translation* – SMT) Primeri sistemov: Google Translate v osnovni obliki[5] (Och, 2006), Moses (Koehn et al., 2007).
- Hibridni sistemi predstavljajo mešanico obeh pristopov. Osnova takšnih sistemov sodi v eno od predstavljenih paradigem in je oplemenitena z metodami druge paradigme. Primeri sistemov: Google Translate za izbrane jezikovne pare (Och, 2006), Microsoft Bing.[6]

Strojno prevajanje in slovenščina

Pregled spleta (2. 5. 2016) ponuja izbiro naslednjih prevajalnih sistemov, v katerih v jezikovnih parih nastopa tudi slovenščino (sistemi so urejeni po abecednem vrstnem redu):

- **Bing Translator** je hibridni sistem za strojno prevajanje naravnih jezikov. Sistem temelji na statističnem strojnem prevajalniku, ki uporablja tudi pravila, odvisna od jezika, ter določeno mero analize izvornega besedila. Microsoft ta sistem definira kot »jezikovno obveščeno statistično strojno prevajanje« (*Linguistically informed statistical machine translation*). Sistem je v osnovi statistični sistem za strojno prevajanje na osnovi fraz, ki vključuje jezikovno odvisno analizo besedila, drevesa odvisnosti (dependency trees) ter drevesa izpeljave (parse trees) in pravila za poravnavo besed (word alignment rules) za generalizacijo naučenih fraz.

1 Presis: http://presis.amebis.si/.
2 Systran: http://www.systransoft.com/.
3 ProMT: http://www.online-translator.com/.
4 Apertium: http://www.apertium.org/.
5 Google Translate: https://translate.google.com/.
6 Microsoft Bin translator: https://www.bing.com/translator.



Slika 1: Moduli tipičnega sistema za strojno prevajanje na osnovi pravil plitkega prenosa

- **Google Translate** je za jezikovne pare s slovenščino tipičen predstavnik sistemov statističnega strojnega prevajanja (SMT). Prevodi se izvajajo prek jezika, v tem primeru angleščine, kar pomeni, da se izvorno besedilo najprej prevede v angleščino in šele nato v ciljni jezik. Poleg te omejitve Google Translate ne uporablja dodatnih metod, ki temeljijo na pravilih, ki jih uporablja za nekatere jezikovne pare (Vičič in Kuboň, 2015).
- **iTranslate4.eu** je Evropski projekt z istim imenom (http://www.itranslate4.eu/) in željo povečati zaupanje v strojno prevajanje. Končna naloga tega sistema je postavitev spletnega portala za prevajanje med evropskimi jeziki. Portal uporablja različne prevajalne sisteme in izbiro sistemov tudi prikaže. Portal za prevode ponuja več predlogov, ki jih sestavi z različnimi prevajalnimi sistemi. Slovenščino podpira prevajalni sistem Presis1 podjetja Amebis, ki se po potrebi kombinira še z drugimi prevajalniki za druge jezike.
- **Presis** podjetja Amebis (Romih in Holožan, 2002) je bil prvi sistem za strojno prevajanje, ki je med prevajalnimi jezikovnimi pari vseboval slovenski jezik. Sistem sodi v paradigmo strojnih prevajalnih sistemov na osnovi pravil (*Rule-Based Machine Translation* – RBMT). Presis razčleni vsako poved v izvornem jeziku na slovnične komponente, kot so osebek, predmet, povedek in atributi ustreznih semantičnih kategorij. Na osnovi analiziranega izvornega besedila izbere pripravljena pravila, ki omogočajo prevod analiziranih komponent v ciljni jezik, nato sintetizira poved v ciljnem jeziku.
- **Prevajalni sistem Guat**[7] (Vičič, 2012) (ime je dobil po majhnih ribah *Gobiidae*, ki živijo tudi v slovenskem morju) je bil zgrajen med razvojem metod, prikazanih v poglavju Metodologija. Sistem podpira jezikovna para slovenščina – srbščina in slovenščina – hrvaščina. Metode so bile preverjene prek več iteracij (sistematične napake so bile popravljene in popravki so vključeni v osnovno ogrodje). Posebnosti jezikovnih parov so: jeziki so zelo pregibni, oblikoslovno in derivacijsko bogati. Visoka stopnja pregibnosti zahteva oblikoskladenjsko analizo izvornega jezika in, posledično, oblikoskladenjsko sintezo v končni fazi v ciljnem jeziku, čeprav so si jeziki sorodni.

Strojno prevajanje na osnovi pravil plitkega prenosa

Sistemi strojnega prevajanja s pravili plitkega prenosa (*shallow transfer rule based machine translation*) v večini primerov uporabljajo enostavno arhitekturo, pri čemer je analiza izvornega jezika omejena na oblikoskladenjske oznake. Arhitektura, ki jo uporablja večina sistemov za strojno prevajanje naravnih jezikov na osnovi pravil plitkega prenosa in plitke sinteze, je prikazana na sliki 1. Ta arhitektura je bila najprej predstavljena v (Hajič et al., 2000) in pozneje uporabljena tudi v ogrodju Apertium (Corbi-Bellot et al., 2005).

Opis posameznih modulov prevajalnega sistema, kot so prikazani na sliki 1:

- Oblikoskladenjska analiza (morphosyntactic analysis) izvornega besedila vsaki besedi pripiše vse možne oblikoskladenjske oznake, ki bi jih ta besedna oblika lahko imela.
- Razdvoumljanje (disambiguation) služi za izbiro najverjetnejše oznake za posamezno besedo glede na njeno okolico.
- Strukturni prenos s pomočjo pravil in dobesednih prevodov prenese označeno besedilo v ciljni jezik.
- Oblikoskladenjska sinteza nadomesti oblikoskladenjsko označeno besedilo z dejanskimi besednimi oblikami v ciljnem jeziku.

Moduli so natančneje opisani v 5. razdelku, in sicer na primeru ogrodja Apertium (Corbi-Bellot et al., 2005).

Apertium – odprtokodno ogrodje za prevajalni sistem sorodnih jezikov

Apertium (Corbi-Bellot et al., 2005) je odprtokodno ogrodje za postavitev samodejnega prevajalnega sis-

7 Prevajalnik GUAT: http://jt.upr.si/guat.

tema za sorodne jezike tipa plitkega prenosa (*shallow transfer*) (Sanchez-Martinez in Ney, 2006). Predstavlja ogrodje, ki omogoča prevajanje med sorodnimi jeziki s pomočjo pravil. Uvršča se med sisteme za samodejno prevajanje naravnih jezikov na osnovi pravil plitkega prenosa (*shallow-transfer* RBMT). Prevajanje je razdeljeno na pet osnovnih faz:

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

Arhitektura ogrodja Apertium je predstavljena na sliki 1.

METODOLOGIJA

V naslednjih razdelkih so opisana vsa jezikovna gradiva, ki jih potrebujemo za postavitev sistema za strojno prevajanje sorodnih jezikov z ogrodjem Apertium. Opisani so tudi postopki samodejne izdelave gradiv in najpomembnejše napake v njih, ki so bile ročno odpravljene.

Nabor oblikoskladenjskih oznak

V postopku oblikoskladenjskega označevanja, v literaturi pogosto predstavljenega tudi kot označevanje z oblikoskladenjskimi oznakami – MSD (morphosyntactic descriptions), so posameznim besedam v besedilu pripisane oznake, upoštevajoč besedni razred (ali: besednovrstno kategorijo) in tudi njeno okolico v besedilu. V slovenskih korpusih so standardne oznake MSD po dveh virih oblikoskladenjskih specifikacij:

- projekt JOS (Erjavec et al., 2010a), same specifikacije so predstavljene v Erjavec, 2010b,
- projekt MULTEXT(-East) (Dimitrova et al., 1998), ki temeljijo na delu skupine EAGLES (Calzolari in Monachini, 1996).

Oboje določajo strukturo in vsebino veljavnih oblikoskladenjskih oznak ali MSD-jev.

Nabor oblikoskladenjskih oznak ogrodja Apertium je prirejen za uporabo v dokumentih v formatu XML. Oznake so sestavljene iz posameznih oznak, ki jih lepimo skupaj (konkateniramo). Vrstni red ne spremeni kategorij in lastnosti posamezne oznake, a je pri prevajanju še vedno pomemben. Primeri oznak s slovenskimi prevodi so predstavljeni v tabeli 1.

Oblikoskladenjski slovar

Oblikoskladenjski slovar združuje vse besedne oblike, ki spadajo v isto pregibno skupino, v razrede z osnovno obliko – lemo. Nadalje, te razrede oziroma skupine druži v paradigme, razrede, ki združujejo vse leme, ki se spreminjajo po istih pravilih glede na oblikoskladenjske oznake.

Oblikoskladenjski slovar, ki ga uporablja Apertium, lahko pa bi takšne slovarje z manjšimi spremembami uporabljali tudi drugi prevajalni sistemi ali pa jezikovno gnane aplikacije, temelji na lemah, ki so zbrane v paradigmah. Posamezna paradigma združuje vse leme, ki se

Tabela 1: Razlaga značk in atributov zapisa oblikoskladenjskih oznak v formatu Apertium.

oznaka	opis
⟨n⟩	samostalnik
⟨nom⟩	imenovalnik
⟨gen⟩	rodilnik
⟨m⟩	moški spol
⟨f⟩	ženski spol
⟨nt⟩	srednji spol
⟨sg⟩	ednina
⟨pl⟩	množina
⟨du⟩	dvojina
⟨vblex⟩	glavni glagol
⟨vbser⟩	pomožni glagol
⟨adj⟩	pridevnik
⟨adv⟩	prislov

```
<e lm="cepljen">
     <i>cepljen</i>
      <par n="veplen/__adj"/>
</e>
<e lm="procesija">
     <i>procesij</i>
     <par n="og/a__n"/>
</e>
…
<e lm="cerkev">
     <i>cerk</i>
     <par n="cerk/ev__n"/>
</e>
lema: cerkev
krn: cerk
paradigma: cerk/ev__n
```

Slika 2: Del zapisov v enojezičnem slovarju. Lema je zapisana v atributu lm značke e, nato sledi krn ter značka par, ki označuje paradigmo. Zapis cerkev je predstavljen z lemo, krnom ter paradigmo.

spreminjajo po istih pravilih glede na oblikoskladenjske oznake.

Slika 2 prikazuje primere lem in njihovo članstvo v paradigmah. Lema je predstavljena s svojim imenom (ime leme), krnom, najdaljšim delom, ki je skupen vsem njenim besednim oblikam, in z imenom paradigme, v kateri so opisana vsa pravila sprememb glede na oblikoskladenjske kategorije.

Primer za lemo *cerkev* je predstavljen na sliki 2. Posamezna gesla enojezičnega slovarja so združena v oblikoskladenjske paradigme, kot so definirane v (Spencer, 1991). Oblikoskladenjske paradigme vsebujejo vse leme, katerih besedne oblike se spreminjajo na enak način za vse oblikoskladenjske oznake (oznake MSD).

Slika 3 prikazuje primer paradigme za ženski samostalnik v slovenščini. Uporaba paradigme omogoča izdelavo kompaktnejšega zapisa podatkov. Za paradigmo *cerk/ev__samostalnik* v slovenskem jeziku velja: vsi samostalniki prve ženske sklanjatve paradigme *-ev*, kot so cerkev, breskev, podkev, se sklanjajo po istem vzorcu in jih združimo v isto paradigmo. Enostavno pravilo določa spremembo besede iz imenovalnika v rodilnik s spremembo končnice iz cerkev v cerkve, torej eno pravilo tako zadošča za celo skupino besed in ne le za en osamljeni primer.

Posamezen zapis v slovarju je predstavljen z oznako XML *e*, atribut te oznake *lm* predstavlja ime leme, gnezdena oznaka *i* krn besede, oznaka *par* pa ime paradigme.

Tabela 2: Razlaga značk in atributov zapisa slovarjev v formatu Apertium

oznaka	opis
⟨pardef⟩	definicija paradigme
⟨e⟩	element for entry – zapis v slovarju in paradigmi
⟨p⟩	string pair – par nizov
⟨par⟩	reference to paradigm – povezava na paradigmo
⟨re⟩	reference to regular expression – povezava na regularni izraz
⟨s⟩	reference to regular symbol – povezava na simbole
⟨i⟩	oblikoskladenjskih oznak reference to identity transduction – način za zapis para nizov z isto vsebino
⟨l⟩	left part – leva stran zapisa besedila s slovničnimi simboli
⟨r⟩	right part – desna stran zapisa besedila s slovničnimi simboli
⟨lm⟩	Lema
atribut	Opis
n	dejanska vsebina značke ásñ

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

Slika 3: Del paradigme za samostalnike ženskega spola v slovenščini. Tipični predstavnik je lema cerkev. Končnica -ev se spreminja v skladu z različnimi MSD-ji. Značke so obširneje predstavljene v Tabeli 2

```
lema: cerkev
krn: cerk
primeri besednih oblik:
besedna oblika: cerkev
pripona: ev
MSD: samostalnik ženski spol ednina imenovalnik
besedna oblika: cerkvah
pripona: vah
MSD: samostalnik ženski spol ednina+mestnik
```

Slika 4: Del paradigme cerk-ev. Lema: cerkev, krn: cerk, dve besedni obliki cerkev in cerkvah

Pri indoevropskih jezikih, ki večinoma uporabljajo konkatenativno oblikoslovje,[8] besedne oblike določajo menjave obrazil, najpogosteje 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* ter z dodajanjem pripone *-ho* moški ali srednji spol imenovalnik ali tožilnik.

Samodejna izdelava enojezičnih oblikoskladenjskih slovarjev izvornega in ciljnega jezika

Iz oblikoskladenjsko označenega in lematiziranega korpusa najprej izluščimo vse besedne oblike ter jih združimo po lemah. Lahko bi uporabili poljuben oblikoskladenjsko označen korpus, uporabili smo poravnani del korpusa MULTEXT-EAST (Dimitrova et al., 1998), ki ga sestavlja roman 1984 (Orwell, 1949) predvsem zaradi dostopnosti. V okviru tega projekta je nastal tudi leksikon, ki pa ga nismo uporabili zaradi možnih licenčnih težav, poleg tega pa nam metoda omogoča širjenje leksikona z dodatnimi korpusi.

Leme z enakimi spremembami družimo v paradigme, kar nam omogoča sestavljanje manjkajočih besednih oblik. Vsaka paradigma 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 zapisana oblikoskladenjska oznaka po (Erjavec, 2010b).

Metoda je bila predstavljena v članku (Vičič, 2009). Primer paradigme je prikazan na sliki 4.

Paradigme izdelamo z naslednjim algoritmom: 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: dve paradigmi združimo v eno, če pripadata isti besedni vrsti (prva kategorija MSD) in če se noben par zapisov ne izključuje. Dva zapisa se izključujeta, če imata enako oznako MSD in različna obrazila, kot kaže primer na sliki 5. Vsaka paradigma ima shranjen celoten seznam vseh lem, ki jo sestavljajo; ta seznam pri združevanju vsebuje leme obeh paradigem.

```
lema: cerkev
krn: cerk
 besedna oblika: cerkev
 pripona: ev
 MSD: samostalnik ženski
spol ednina imenovalnik
 lema: ana
 krn: an
 besedna oblika: ana
 pripona: a
 MSD: samostalnik ženski
spol ednina imenovalnik
ev != a
```

Slika 5: Končnici besednih oblik z isto oznako MSD se ne ujemata, kar pomeni, da paradigem ne združimo

Oblikoskladenjski slovarji izvornega in ciljnega jezika so bili zgrajeni s pomočjo paradigem; leme z manjkajočimi besednimi oblikami v originalnih slovarjih so bile dopolnjene, velikost končnega slovarja je bila približno dvajsetkrat večja od začetnega (Vičič, 2009).

Ročna predelava enojezičnega slovarja

Ročni pregled je bil zastavljen metodično: vsako besedno vrsto smo obravnavali ločeno in poskušali odkriti sistematske napake. Posebej smo se lotili odprave napak slovarja zaradi napak v izvornih učnih gradivih.

Glagoli so imeli določene že vse potrebne oznake, ki jih potrebujemo pri prevajanju v našem sistemu: namenilnik, povednik, velelnik, deležnik na -n/-t ter deležnik na -l. Poleg osnovnih oblik je bil določen tudi glagolski vid. Samodejna metoda ni upoštevala podatkov o glagolski prehodnosti. Vse glagolske paradigme so bile podvojene, tako da smo lahko označili obe oblikoskladenjski oznaki za glagolsko prehodnost. Z ročnim označevanjem smo za vsako lemo posebej določili pravilne oznake.

V slovenščini pridevnike in prislove stopnjujemo tristopenjsko, in sicer kot osnovnik, primernik, presežnik, ter dvostopenjsko kot osnovnik in elativ (Toporišič, 2000). Dopolnjene so bile paradigme, ki pokrivajo vse štiri osnovne oblike; z ročnim označevanjem so bile označene leme, za katere obstaja samo osnovnik oziroma različne kombinacije vseh štirih oblik. Za lažje generiranje pridevniških oblik so bile paradigme osnovnih oblik vezane na sekundarne paradigme, ki vsebujejo še oznake, kot so

8 Besede so sestavljene iz več združenih (*concatenated*) morfemov.

```
<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="adv"/></l>
  <r>oko<s n="adv"/></r>
</p></e>
<e><p>
  <l>okolina<s n="n"/><s n="f"/></l>
  <r>prilika<s n="n"/><s n="f"/></r>
</p></e>
```

Slika 6: Primeri dvojezičnih prevodov lem iz slovenščine v hrvaščino. Značke so obširneje predstavljene v Tabeli 2

spol, število, sklon in določnost. Uporabljene oblikoskladenjske oznake označujejo vse potrebne informacije za prevajanje jezikovnega para, razen opisanih pomanjkljivosti. Treba je bilo le dodati manjkajoče leme.

Poleg samodejno zgrajenih gradiv smo uporabili tudi že obstoječa gradiva projekta Apertium, izdelana v okviru pilotskega prevajalnega sistema srbščina in hrvaščina ter makedonščina[9] v okviru projekta Google Summer Of Code 2011 (Google, 2012b).

Dvojezični slovar

Dvojezični slovarji temeljijo na parih *<izvorna lema – ciljna lema>* in na poravnanih besednih zvezah v lematizirani obliki, torej na dobesednih prevodih lem. Primeri dvojezičnih prevodov lem iz slovenščine v hrvaščino so predstavljeni na sliki 6.

Pri prevajanju se poleg samega prenosa iz izvornega v ciljni jezik prenesejo oziroma prevedejo tudi oblikoskladenjske oznake. Primer (1) prikazuje prevod hrvaške besede *prozor* v slovensko besedo *okno*, pri čemer se spremeni tudi spol iz moškega v srednji.

Oznaka izvorne leme se ponavadi ujema z oznako ciljne leme, še posebej pri sorodnih jezikih. Uporaba oznak omogoča razdvoumljanje lem z istim imenom in različnim pomenom. Dvojezični slovar z menjavo oznak omogoča opisovanje leksikalnih razlik med jezikoma.

Samodejna izdelava dvojezičnih prevajalnih slovarjev

Dvojezični prevajalni slovar vsebuje besede enojezičnih slovarjev ter njihove ustrezne prevode z vsemi ustreznimi oblikoskladenjskimi oznakami. Pri tem moramo paziti, da se oblikoskladenjske oznake izvornega ter ciljnega slovarja pokrivajo, v tem primeru je pomemben tudi vrstni red. V primeru nepravilnega zaporedja oblikoskladenjskih oznak se beseda ne bi pravilno prevedla.

(1)

prozor ⟨samostalnik⟩⟨**moški**⟩ ⇒ **okno** ⟨samostalnik⟩⟨**srednji**⟩
stol ⟨samostalnik⟩⟨**moški**⟩ ⇒ **miza** ⟨samostalnik⟩⟨**ženski**⟩
godina ⟨samostalnik⟩⟨**ženski**⟩ ⇒ **leto** ⟨samostalnik⟩⟨**srednji**⟩

Dvojezični prevajalni slovar lahko izdelamo iz poravnanega dvojezičnega korpusa s pomočjo statističnih metod oziroma modelov (Vičič, 2008). Poseben problem pri uporabi statističnih modelov je v redkih podatkih (*sparse data problem*) (Katz, 1987). Osnovni korpus ima določeno število dovolj dobro opisanih pravil in dovolj pogosto zastopanih besed, vsebuje pa tudi delež slabo predstavljenih besed in pravil. Z večanjem korpusa uvajamo tudi nove besede. Tako se, ob predpostavki, da se porazdelitev besed ne spremeni, odstotek slabo opisanih besed in pravil z večanjem korpusa ne manjša. Problem redkih (pomanjkljivih) podatkov rešujemo s pomočjo naprednih algoritmov, ki upoštevajo predhodno znanje o problemu, izkušnje s sorodnih domen ali pa celo povsem tujih domen. Šumne podatke izločamo s pomočjo zakonitosti v podatkih, z izločanjem ekstremov. Paziti moramo, da pri izločanju napačnih podatkov ne pretiravamo in korpusa ne »porežemo«, poenostavimo preveč.

Opisanega problema smo se lotili z dvema metodama (Vičič in Homola, 2010), ki sta predstavljeni v naslednjih razdelkih:

- *poravnava lematiziranih besed s pomočjo besednih vrst*: iskanje poravnav med lemami s pripisano oznako besedne vrste jezikovnega para učnega korpusa namesto iskanja povezav med vsemi besednimi oblikami, oznaka besedne vrste odpravlja veliko dvoumnosti, na žalost pa ne vseh;
- *razširitev dvojezičnega slovarja s podobnicami in iskanje najprimernejših paradigem v ciljnem enojezičnem slovarju*: pri tej metodi se zanašamo na podobnice; leme so prenesene v ciljni jezik brez prevoda, v ciljnem slovarju pa je novi lemi poiskana najprimernejša paradigma.

Poravnava lematiziranih besed s pomočjo oznak besednih vrst

Dvojezični prevajalni slovar je sestavljen iz parov *<izvorna lema z oznako besedne vrste – ciljna lema z oznako besedne vrste>*, ki omogočajo prevajanje v ciljni jezik. Oznake besednih vrst v dvojezičnih slovarjih

9 Projekt Apertium (sh-mk): http://sourceforge.net/p/apertium/svn/46791/tree/trunk/apertium-sh-mk/

omogočajo enostavno izogibanje dvoumnostim enako imenovanih lem različnih besednih vrst.

(2)

```
priti_SAMOSTALNIK biti_GLAGOLP
do_PREDLOG
podrt_PRIDEVNIK drevo_SAMOSTALNIK .

o_PREDLOG kateri_ZAIMEK on_ZAIMEK
biti_GLAGOLP
praviti_GLAGOL .
...
```

Besede v enojezičnih slovarjih so zapisane v lematizirani obliki, besedne oblike pa so zabeležene v paradigmah. Metoda poravnave lem omogoča boljše rezultate v primerjavi s poravnavo besed v korpusu zaradi zmanjšanja prostora iskanja (Saleh, 2009). Omejitev prostora iskanja poveča natančnost modela poravnave besed, vendar v njem ni več informacije o besednih oblikah. To informacijo smo ohranili s povezavo s paradigmami v enojezičnih slovarjih. Podobna metoda je uporabljena tudi v (Vargas-Sierra in Lindemann, 2013).

Za samo učenje poravnave lem lahko uporabimo poljuben statistični algoritem za iskanje poravnave besed v dvojezičnih, povedno poravnanih korpusih (SMT *word-to-word model*).

Uporabili smo orodje GIZA++ (Och in Ney, 2003), ki temelji na algoritmu, prikazanem v (Brown et al., 1993). Model je bil naučen na vzporednem, povedno poravnanem seznamu lem z oznakami besednih vrst, ki je bil izluščen iz korpusa 1984. Del seznama pripravljenih učnih podatkov je prikazan na primeru (2).

Razširitev dvojezičnega slovarja s podobnicami in iskanje najprimernejših paradigem v ciljnem enojezičnem slovarju

Ta metoda omogoča večanje dvojezičnega slovarja in ustrezno popravi enojezični oblikoskladenjski slovar. Metode, opisane v razdelkih poglavja Metodologija ne zagotavljajo popolne pokritosti enojezičnih slovarjev z dvojezičnim slovarjem. Leme izvornega enojezičnega slovarja, ki po izvajanju teh metod nimajo prevodov, poskušamo prevesti s pomočjo metode, ki temelji na podobnicah. Podobnice – *cognates* so besede, ki imajo skupen etimološki izvor. Pri prevajanju med dvema jezikoma so predvsem dobrodošle tiste, ki se s časom niso veliko spremenile ne v pomenu niti v obliki.

Metoda doda manjkajoče zapise v dvojezični slovar: za vsak vnos izvornega slovarja, ki nima pokritja v dvojezičnem slovarju, tj. nima ustreznega prevoda, vstavimo v dvojezični slovar nov par **izvorna lema – izvorna lema**, kar pomeni, da prevajamo lemo v enako lemo v ciljnem jeziku. Poleg same leme je novemu zapisu dodan tudi del MSD-ja; pri primeru (3) je zapisana besedna vrsta in spol. Nov vnos se v ciljni enojezični slovar doda, če nove leme z enako besedno vrsto ne najdemo v ciljnem slovarju. V ciljnem slovarju je dodana nova lema in zanjo izbrana paradigma, ki najbolj ustreza novo dodani lemi; to je paradigma, ki omogoča generiranje besednih oblik z ustreznimi MSD-ji in vsebuje najdaljšo pripono, ki ustreza novi lemi. Algoritem ponovimo še za ciljni slovar, trenutni izvorni slovar postane v drugem delu metode ciljni.

(3)

```
slovenski slovar:
lema: list
krn list
paradigma žvenket/__samostalnik

slovensko-srbski dvojezični slovar:
list samostalnik moški spol se prevaja
v
list samostalnik moški spol

srbski slovar:
lema: list
krn list
paradigma um/__samostalnik
```

Prvi del primera kaže slovensko lemo **list** z označeno paradigmo **žvenket/__samostalnik**, ki je prisotna v izvornem slovarju in nima prevoda v dvojezičnem slovarju. Drugi del primera kaže nov vnos v dvojezični slovensko-srbski slovar; vpis prevaja slovensko lemo **list** v srbsko lemo **list**. Poleg same leme je zapisan še del MSD-ja, v tem primeru še besedna vrsta (samostalnik) in spol (moški), ki bi se lahko tudi zamenjal, vendar se pri pomanjkanju dodatnih informacij zanašamo na podobnost besed. Tretji del primera opisuje nov vpis v ciljnem slovarju. Dodana je nova lema **list** in poiskana najustreznejša paradigma **um /__samostalnik**.

Ročna predelava dvojezičnega slovarja

Dvojezični slovar, ki je bil izdelan samodejno, je bil razširjen s pomočjo prevajalnega sistema Google Translate (Google, 2012a). Prevajali smo samo leme. Napake in manjkajoči prevodi so bili ročno popravljeni. Tak način uporabe sistema se je izkazal za neprimernega pri prevajanju pridevnikov, prislovov ter predlogov brez okolice (prevajali smo samo leme). Kakovost prevodov samostalnikov in glagolov je tudi zadovoljiva.

V novonastalem dvojezičnem slovarju so bile uporabljene le prevedene leme, ki so imele vnose v slovarju ciljnega jezika. Po opravljeni prvi iteraciji izdelave dvojezičnega slovarja je sledila druga iteracija enakega procesa z zamenjanima izvornim in ciljnim slovarjem (smer hr – sl). V nadaljevanju so opisane najpomembnejše težave, ki smo jih odpravljali pri gradnji dvojezičnega prevajalnega slovarja.

Prva težava, na katero smo naleteli, je bila razlika v

stopnjevanju prislovov. V obeh jezikih prislove stopnjujemo štiristopenjsko, in sicer kot osnovnik, primernik, presežnik ter elativ. Težava je nastala pri prevajanju besed, ki niso imele enakega števila oziroma istih stopenj v izvornem in ciljnem jeziku. Težavo smo rešili na tak način, da smo pred prislovi dodali besedo bolj, najbolj ali preveč, odvisno od manjkajoče oblike. Primer (4) kaže primere prevajanja iz hrvaškega v slovenski jezik:

```
<!-- primer: bom kupil-->
<rule>
 <pattern>
  <pattern-item n="pomožni glagol v prihodnjiku"/>
  <pattern-item n="glavni glagol"/>
 </pattern>
 <action>
  <let>
   <clip pos="1" side="tl" part="lema"/>
   <lit v="hteti"/>
  </let>
  <let>
   <clip pos="1" side="tl" part="oblika"/>
   <lit-tag v="deležnik"/>
  </let>
  <let>
   <clip pos="2" side="tl" part="oblika"/>
   <lit-tag v="nedoločnik"/>
  </let>
  <out>
   <lu>
    <clip pos="1" side="tl" part="lema"/>
    <clip pos="1" side="tl" part="pomožni glagol"/>
    <clip pos="1" side="tl" part="oblika"/>
    <clip pos="1" side="tl" part="oseba"/>
    <clip pos="1" side="tl" part="število"/>
   </lu>
   <b pos="1"/>
   <lu>
    <clip pos="2" side="tl" part="lema"/>
    <clip pos="2" side="tl" part="glavni glagol"/>
    <clip pos="2" side="tl" part="oblika"/>
   </lu>
  </out>
 </action>
</rule>
```

Slika 7: Primer pravila za strukturni prenos. Pravilo opisuje spremembe načina zapisa prihodnjika iz slovenščine v hrvaščino. Posamezne značke so predstavljene v Tabeli 3.

(4)

bijelo (osnovnik) – belo (osnovnik)
bjelije (primernik) – bolj belo (primernik)
najbjelije (presežnik) – najbolj belo (presežnik)

Na podobno težavo naletimo tudi pri pridevnikih. V enojezičnem slovarju ciljnega jezika hrvaščina so bila prisotna tudi deležja, ki se v slovenskem jeziku prevedejo v načinovne prislove s končnicami *-oč/-ečl-e/-aje*. Težave smo imeli z glagolskimi prislovi, ki nimajo ustreznega prevoda v slovenskem jeziku, zato jih je bilo treba prevesti v pridevnike (moški spol, ednina, imenovalnik). Primer (4) kaže glagolske prislove s primernim prevodom ter prevodi v pridevnik

(5)

Glagolski prislovi s primernim prevodom:
viseći ⇒ viseč,
čekajući ⇒ čakajoč,
Glagolski prislovi s prevodom v pridevnik:
poštujući ⇒ spoštovan,

Pravila prenosa

Apertiumov modul strukturnega prenosa (*Structural transfer module*) uporablja tehnologijo končnih avtomatov za odkrivanje vzorcev fiksne dolžine leksikalnih enot (kosov besedila ali fraz),[10] 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 ...).

Pravila so zgrajena iz dveh delov: končnega števila elementov, ki opisujejo vzorce fiksne dolžine, in dela, ki omogoča opis akcije, ki je potrebna za spremembo vzorca. Vzorec je predstavljen s sekvenco leksikalnih kategorij izvornega jezika poljubne dolžine, ločenih s presledki (*b* – blank). Na sliki 8 je vzorec oblike: *pomožni glagol v prihodnjiku* in *glavni glagol poljubne oblike*. Ukrep (*action*) določa akcije, ki naj se izvedejo nad sekvencami vzorca ter izhodni vzorec leksikalnih kategorij ciljnega jezika, ki naj se zgradi. Po detekciji vzorcev se izvedejo spremembe, ki so opisane v telesu pravila (izhod modula so spremenjene leksikalne enote).

Primer pravila je predstavljen na sliki 8. Pravilo je sestavljeno iz dveh delov: vzorec (*pattern*) in ukrep (*action*). Opisuje spremembe načina zapisa prihodnjika iz slovenščine v hrvaščino. Vzorec je sestavljen iz dveh leksikalnih

10 Fraza je v tem primeru del besedila (*chunk of text*), ki nima nujno zaključenega pomena oziroma drugačne jezikoslovne razlage za razdelitev.

Tabela 3: Razlaga oznak in atributov zapisa pravil v formatu Apertium

oznaka	Opis
⟨rule⟩	celotno pravilo
⟨pattern⟩	vsebuje eno ali več značk (pattern-item), ki definirajo
	leksikalne oblike, na katere lahko apliciramo pravilo
⟨pattern-item⟩	del vzorca, leksikalna enota
⟨action⟩	del pravila, ki opisuje ukrep, spremembo vzorca
⟨let⟩	sprememba izvornega dela
⟨clip⟩	izbere del leksikalne enote, ki ustreza atributom
⟨lit⟩	generira niz črk
⟨lit-tag⟩	generira niz črk, ki opisujejo jezikovno oznako
⟨out⟩	vsebuje vse, kar bo pravilo izpisalo
⟨lu⟩	definira vsebino celotne leksikalne enote
⟨b⟩	(blank), ločilo med leksikalnima enotama, pogosto je presledek
⟨call-macro⟩	klic makra (programske kode)
atribut	Opis
side	smer, ki jo naslavlja značka (izvorna/ciljna)
part	ime dela, ki ga naslavlja značka
n	dejanska vsebina značke ápattern-itemñ
v	dejanska vsebina značk álitñ in álit-tagñ
pos	(position), zaporedna številka leksikalne enote

enot: *pomožni glagol biti v prihodnjiku* in *glagol poljubne oblike*, ukrep pa spremeni lemo prvega glagola v *hteti*, obliko prvega glagola v *deležnik* ter obliko drugega glagola v *nedoločnik*; v nadaljevanju so v znački <*lu*> (lexical unit) izpisane leksikalne kategorije za obe besedi.

Posamezne oznake zapisa pravil so predstavljene v Tabeli ***3***.

Pravila prenosa so skupaj z dvojezičnim slovarjem uporabljena v modulu za strukturni prenos pri dejanskem prevajanju oblikoskladenjsko označenih leksikalnih enot (po navadi besed ali besednih zvez). S pravili poskušamo opisati strukturne razlike med jezikoma, torej potrebne spremembe za pravilne prevode iz izvornega v ciljni jezik. Pravila plitkega prenosa, kot jih uporablja Apertium, naslavljajo le dele besedila končne velikosti; večina pravil naslavlja dele besedila dolžine 1, 2 ali 3 besede. Modul v izvornem besedilu poišče dele besedila, ki jih naslavlja pravilo. Pravilo na delu besedila, ki ga naslavlja, izvede akcijo in vrne spremenjeno besedilo.

Sama izbira pokritja posameznih izvornih povedi s pravili poteka po principu najdaljšega ujemanja z leve strani (LRLM – *Left-to-Right Longest Match*). Za poved v izvornem jeziku je izbrana takšna veriga pravil, da je za dele, pri katerih bi lahko uporabili več pravil, izbrano tisto, ki naslavlja daljše besedilo od leve proti desni.

Primer kaže poved »Jutri bom kupil rožo« in njen prevod; del te povedi *bom kupil* je posebej označen in naslavlja pravilo na sliki 8.

```
bom kupil
biti-gl pomožni prihod 1os edn
kupiti-gl glavni deležnik edn moški
”Jutri bom kupil rožo.” (SLO)
ću kupiti
hteti-gl pomožni sedanjik 1os edn
kupiti-gl glavni nedoločnik
”Sutra ću kupiti cvijet.” (HR)
```

Oglejmo si še delovanje pravila na primeru 4. Prva beseda pokritja, pomožni glagol v prihodnjiku, ustreza besedi *bom* iz primera, druga beseda, glavni glagol, ustreza besedi *kupil*. Pred izvajanjem samega izpisa pravilo postavi novo lemo prvi besedi *hteti* in obliko glagola v deležnik. Obliko drugega glagola spremeni v nedoločnik. Pravilo pri samem izpisu za vsako besedo le izpiše že spremenjene lastnosti v vnaprej pripravljenem vrstnem redu, kot je prikazano na primeru (6).

Ročna izdelava pravil

S pomočjo metode za samodejno izdelavo pravil in izbiro najboljših (Vičič, 2012) smo izdelali veliko število pravil, saj metoda pri tem ni uporabljala nobenih omejitev. Tako so se pravila med seboj tudi izključevala (kar pomeni, da so delovala na istih vhodnih nizih, sistem bi izbral prvo pravilo, vsa ostala pa bi bila neuporabna).

Metoda bi potrebovala še metriko za vrednotenje pravil, sama uporaba ovrednotenih pravil pa bi zahtevala tudi arhitekturno spremembo prevajalnega sistema. Ta del že presega namene tega članka.

Ostala pravila smo izdelali ročno. Pravila strukturnega prenosa so razdeljena v tri nivoje zaradi večje fleksibilnosti pri zaznavanju besed ali stavkov. Omejili smo se le na prvi nivo, saj je struktura obeh jezikov jezikovnega para zelo podobna. Opomba: pravila so napisana za prevajanje iz hrvaškega v slovenski jezik, torej je v opisanih primerih hrvaščina izvorni jezik, slovenščina pa ciljni jezik. Oglejmo si primere osnovnih in specifičnih pravil:

- Osnovna pravila, ki so potrebna za pravilno prevajanje posameznih besed ali skupin besed – usklajevanje oblikoskladenjskih oznak, so bila dodana za naslednje besedne vrste ter naslednje skupine besed: samostalnike, pridevnike, svojilne

zaimke, glagole, glagolske prislove, glagol biti, glagol imeti, glagol hoteti, predloge, veznike, števila, pridevnik + samostalnik ter svojilni zaimek + pridevnik + samostalnik itd.
- Nekaj specifičnih pravil, ki so potrebna za pravilno prevajanje skupin besed: je + glagol, se + glagol, se + ne biti (preteklik) + glagol, predlog + samostalnik, ne + glagol biti itd.

Dodanih je bilo 31 pravil prenosa.

Tabela 4: Pokritost slovarjev

Slovar	Št. slovarskih gesel (lem)
Enojezični slovar – SLV	25.923 (1.901 paradigem)
Enojezični slovar – HRV	17.330 (1.014 paradigem)
Dvojezični slovar	17.330 (slovarski vnosi)

METODOLOGIJA EVALVACIJE

Naslednji podrazdelki predstavljajo in opisujejo osnovne statistike jezikovnih gradiv, ki so bila ustvarjena v sklopu projekta. Podrobneje opisujejo tudi rezultate vrednotenja prevodov sistema.

Pokritost korpusov

Tabela 4 prikazuje število slovarskih gesel, ki jih vsebuje enojezični slovar izvornega jezika – slovenščine, število slovarskih gesel, ki jih vsebuje enojezični slovar ciljnega jezika – hrvaščine in število vnosov v dvojezičnem slovarju, natančneje, koliko slovarskih gesel ima primerne prevode v dvojezičnem slovarju. Poleg naštetih lastnosti tabela prikazuje tudi število vsebovanih paradigem v posameznem enojezičnem slovarju tako izvornega kot ciljnega jezika.

Tabela 5: Pokritost korpusov: korpus je bil razdeljen na manjše dele, za vsakega je bila izračunana pokritost, prikazano je povprečje vseh delov korpusa ter standardna deviacija

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

Tabela ***5*** predstavlja rezultate vrednotenja pokritosti (*coverage*) korpusov z jezikovnimi gradivi. Metoda je bila izvedena na dveh različnih korpusih, in sicer na korpusu MULTEXT(-East) (Erjavec, 2010a; Dimitrova et al., 1998) ter na delu korpusa OPUS (subs) (Tiedeman, 2012).

Pri korpusu OPUS smo se zaradi časovnih omejitev omejili na del zbirke podnapisov, natančne vrednosti so predstavljene v Tabeli ***5***. Vsebino omenjenih zbirk smo razdelili na intervale po 10.000 besed in jih posamezno prevedli. Na tak način smo izračunali še povprečje in standardno deviacijo. Ob predpostavki, da uporabljeni korpusi dovolj dobro predstavljajo opazovano jezikov-

Tabela 6: Rezultat testiranja z orodjem testvoc (Smer: hrvaščina – slovenščina)

B. vrsta	Skupno	Pravilni	Z @	Z #	%
Pridevniki	1.517.798	1.517.798	0	0	100
Glagoli	1.018.517	1.018.517	0	0	100
Imena	726.576	726.576	0	0	100
Samost.	135.031	135.031	0	0	100
Pom. gl.	35.112	35.112	0	0	100
Zaimki	10.683	10.683	0	0	100
Števniki	10.165	10.165	0	0	100
Prislovi	8.568	8.568	0	0	100
Predlogi	101	101	0	0	100
Kratice	56	56	0	0	100
Medmeti	49	49	0	0	100
Vezniki	71	71	0	0	100

11 Orodje testvoc je del zbirke orodij Apertium: http://wiki.apertium.org/wiki/Testvoc.

Tabela 7: Rezultat testvoc (Smer: slovenščina – hrvaščina)

B. vrsta	Skupno	Pravilni	Z @	Z #	%
Pridevniki	749.994	263.260	370.603	116.131	35.2
Glagoli	77.254	58.991	495	17.768	76.4
Imena	437.433	437.433	0	0	100
Samostalniki	72.478	72.478	0	0	100
Pom. glagoli	120	120	0	0	100
Zaimki	3.382	3.382	0	0	100
Števniki	8991	8991	0	0	100
Prislovi	7.388	4.739	1.610	1.039	64.2
Predlogi	84	84	0	0	100
Kratice	56	56	0	0	100
Medmeti	49	49	0	0	100
Vezniki	56	56	0	0	100

no domeno, nam pokritost oceni pričakovani odstotek neznanih besed pri prevodih. Standardna deviacija predstavlja mero razpršenosti podatkov.

Ob izvajanju testiranja korpus MULTEXT-EAST (Orwell) še ni vseboval hrvaškega prevoda romana 1984, tako je bilo preverjanje te prevajalne smeri s korpusom MULTEXT-EAST omejeno na izvorni jezik, slovenščino.

Pokritost slovarjev

Pokritost slovarjev smo testirali z orodjem testvoc.[11] Osnovna metoda orodja: razširiti enojezični slovar izvornega jezika, nato pa testirati vsako možno besedno obliko izvornega slovarja skozi vse faze prevajalnega sistema. Na tak način ugotovimo, katera analiza besede ima pravilen prevod v enojezičnem slovarju ciljnega jezika, torej brez simbolov za oznako napak # ali @.

Pomen simbolov, ki označujejo napake:

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

V Tabeli 6 so predstavljeni rezultati testiranja enojezičnega slovarja ciljnega jezika. Rezultati prikazujejo kakovost prevajanja posameznih besed iz hrvaškega v slovenski jezik.

V Tabeli 7 so predstavljeni rezultati testiranja enojezičnega slovarja ciljnega jezika z metodo testvoc (Tyers et al., 2010). Rezultati prikazujejo kakovost prevajanja posameznih besed iz hrvaškega v slovenski jezik.

Razlika med obema smerema obstaja, ker je slovenski slovar večji, tako pokriva vse hrvaške besede, druga smer (hrvaški enojezični slovar) pa v tem projektu ni bil dopolnjen.

Vrednotenje kakovosti prevodov

Predstavljeni sistem še ni dokončan; zaradi časovne stiske smo se morali omejiti samo na prvi nivo pravil prenosa. Kljub temu smo se odločili za prvo testiranje sistema na manjšem testnem vzorcu, ki je bil ročno pripravljen: novica iz korpusa SETIMES (Tyers in Alperen, 2010), ki je bila uporabljena v vseh novih sistemih projekta GSOC2011.

Testni primeri so bili izbrani iz korpusa MULTEXT--EAST, in sicer dela, ki ni bil uporabljen kot učna množica pri samodejnih metodah. Vključili smo še skupni testni vzorec projekta Apertium Google Summer Of Code 2011 (Google, 2012b): novica iz korpusa SETIMES (Tyers in Alperen, 2010), ki je bila uporabljena v vseh novih sistemih projekta.

Pri vrednotenju prevodov je bila uporabljena metrika Human-targeted TER (HTER) (Snover et al., 2006), ki temelji na uteženi Levenshteinovi razdalji (*weighted Levenshtein edit-distance*) (Fu, 1982). Ta predstavlja razširitev osnovne Levenshteinove razdalje (Levenshtein, 1965), ki šteje najmanjše število sprememb, ki jih moramo opraviti med prevodom sistema za strojno prevajanje in referenčnim prevodom. Število sprememb še utežimo z dolžino povedi. Dovoljene spremembe so vstavitev, brisanje in zamenjava besede. Namesto referenčnih prevodov so bili pri testiranju prevedeni primeri ročno popravljeni, pri popravljanju je bilo upoštevano načelo čim manjšega števila sprememb, ki že omogoči popolnoma pravilno poved v ciljnem jeziku, ki popolnoma odraža izvorni pomen.

Vrednost na poseben način uporabljene metrike HTER je: 13,7 %.

Metrika BLEU (Papineni et al., 2001) je najbolj razširjena metrika za vrednotenje sistemov strojnega prevajanja, vendar mnogi avtorji (prim. Callison-Burch

et al., 2006; Labaka et al., 2007), soglašajo, da BLEU sistematično zapostavlja sisteme RBMT in ni primerna za visoko pregibne jezike. Metrike nismo uporabili pri testiranju predstavljenega sistema.

ZAKLJUČEK IN NADALJNJE DELO

Kakovost predstavljenega prevajalnega sistema presega raven eksperimentalnih in poskusnih storitev. Prevodi predstavljenega sistema že dosegajo kakovost, ki omogoča širšo uporabo kot zgolj le akademsko postavitev v namene preizkusa metod. O tem lahko sklepamo iz vrednotenja z metodo HTER kot tudi iz pričevanja uporabnikov, ki so sistem preizkušali. Jezikovna gradiva so zapisana v (človeku) berljivem formatu, kar omogoča relativno enostaven vnos popravkov in posledično izboljšavo kakovosti prevajanja.

Projekt Apertium je odprtokoden. Vsa izdelana gradiva so prosto dostopna z licenco GNU *Lesser General Public License* (LGPL) (GNU, 2010) na strežniku projekta.[12] Izdelan je bil tudi spletni vmesnik do »živega« prevajalnega sistema. Prevajalnik je na voljo na strežniku jezikovnih tehnologij Univerze na Primorskem.[13]

Vsi jezikovni viri bodo dostopni prek slovenske raziskovalne infrastrukture CLARIN.[14]

Za slovenščino obstajata še dva enojezična oblikoskladenjsko označena slovarja, in sicer Multext-East (Erjavec, 2010a) in Sloleks (Arhar, 2009). Z relativno majhnim vložkom bi lahko predvsem slednjega uporabili za širjenje enojezičnega slovarja, ki je bil pripravljen v tem projektu (dodajanje novih lem v primerne paradigme, ustvarjanje novih paradigem). Tehnično bi bilo takšno združevanje leksikonov možno, upoštevati pa moramo neskladne licenčne pogoje gradiv.

Poleg osnovnega namena prevajalnega sistema, prevajanja jezikovnega para, so predstavljena gradiva uporabna tudi pri mnogih drugih jezikoslovnih raziskavah in aplikacijah. Ne nazadnje lahko del gradiv uporabimo pri gradnji prevajalnega sistema za nov jezikovni par. V načrtu imamo izdelavo prevajalnega sistema za jezikovni par slovenščina – italijanščina ter dolgoročni načrt izdelave prevajalnika za sorodne južnoslovanske jezike (slovenščina, hrvaščina, srbščina, bosanščina, makedonščina).

Gradiva pa niso uporabna le v prevajalnem sistemu, oblikoskladenjsko označeni slovar in dvojezični slovar sta uporabno gradivo za jezikoslovne raziskave in tudi za izdelavo jezikoslovno gnanih aplikacij. Način dostopnosti gradiv omogoča relativno prosto uporabo, standardiziran način označevanja pa enostavno uporabo.

12 Projekt Apertium: http://www.apertium.org/.
13 Strojno prevajanje: http://jt.upr.si/mt_slo.html.
14 CLARIN: http://clarin.si.

LINGUISTIC MATERIALS FOR THE MACHINE TRANSLATION SYSTEMS

Jernej VIČIČ
University of Primorska Andrej Marušič Institute, Muzejski trg 2, 6000 Koper, Slovenia
e-mail: jernej.vicic@upr.si

SUMMARY

Rule based machine translation systems require quality language resources, such as morphologically enriched dictionaries, bilingual dictionaries and translation rules. Materials are prepared in a standardized format and are also suited for use in a multitude of applications. The article presents the methods that have been used both to build language resources as well as the extent and quality of the produced material and a fully functional machine translation system.

The paper presents linguistic materials used in a machine translation system for the language pair Slovenian – Croatian. It presents the machine translation system with the associated language materials. The presented methods include: automatic production of monolingual morphologies, bilingual translation dictionaries and translation rules.

The paper also presents the manual cleaning for each language material used in the translation system. The evaluation had two main objectives: evaluation the translation quality of the basic translation system and evaluation of the size and quality of the individual language resources. All materials and the entire translation system are freely available.

Keywords: Machine translation of natural languages, morphosyntactic dictionary, translation rule, paradigm, lemma

LITERATURA

Arhar, Š. (2009): Učni korpus SSJ in leksikon besednih oblik za slovenščino. Jezik in slovstvo, 54, 3–4, 43–56.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. & R. L. Mercer (1993): The mathematics of statistical machine translation: parameter estimation. Computational linguistics, 19, 163–311.

Callison-Burch C., Osborne, M. & P. Koehn (2006): Re-evaluating the role of BLEU in machine translation research. Proceedings of EACL, Trento, Association for Computational Linguistics, 249–256.

Calzolari, N. & M. Monachini (1996): Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: a common proposal and applications to European languages. Eagles report.

Corbi-Bellot, A. M., Forcada, M. L. & S. Ortiz-Rojas (2005): An open-source shallow-transfer machine translation engine for the Romance languages of Spain. Proceedings of the EAMT conference. Budapest, EAMT, 79–86.

Dimitrova, L. et al. (1998): Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. COLING-ACL, Montréal, Association for Computational Linguistics, 315–319.

Erjavec T., Fišer, D., Krek, S. & N. Ledinek (2010): The JOS Linguistically Tagged Corpus of Slovene. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Malta, ELRA.

Erjavec, T. (2010): MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Malta, ELRA.

Fu, K. S. (1982): Syntactic Pattern Recognition and Applications. Prentice-Hall, Englewood Cliffs, NJ.

GNU (2010): GNU General Public License. http://www.gnu.org/licenses/index_html#GPL.

Google (2012a): The Google translator. http://www.google.com/translate_t.

Google (2012b): Google Summer of Code 2011. http://www.google-melange.com/gsoc/homepage/google/gsoc2011.

Hajič, J., Hric, J. & V. Kuboň (2000): Machine translation of very close languages. Proceedings of the 6th Applied Natural Language Processing Conference, Hong Kong, Association for Computational Linguistics, 7–12.

Jagland, T. & A. Vassiliou (2011): Skupna izjava Sveta Evrope in Evropske komisije. Evropska komisija, 1–2.

Katz, S. (1987): Estimation of Probabilities from Sparse Data for the Language Model. IEEE Transactions on Acoustics, Speech and Signal Processing, 35, 3, 400–401.

Koehn, P. et al. (2007): Open Source Toolkit for Statistical Machine Translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07), ACL, 177–180.

Labaka, G., Stroppa, N., Way, A. & K. Sarasola (2007): Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. Proceedings of the Machine Translation Summit XI, EAMT, 41–48.

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

Och, F. J. & H. Ney (2003): A Systematic Comparison of Various Statistical Alignment Models. Computational linguistics, 29, 19–51.

Och, F. J. (2006): Challenges in Machine Translation. In: Proceedings of the ISCSLP, Springer, 15.

Orwell, G. (1949): 1984. London, Secker and Warburg.

Papineni, K., Roukos, S., Ward, T. & W.-J. Zhu (2001): BLEU: a method for automatic evaluation of machine translation. Technical report, IBM.

Romih, M. & P. Holozan (2002): A slovenian-english translation system. V: Proceedings of the 3rd Language Technologies Conference, 167.

Saleh, I. (2009): Automatic extraction of lemma--based bilingual dictionaries for morphologically rich languages. Thesis, Georgetown University.

Sanchez-Martinez, F. & H. Ney (2006): Using Alignment Templates to Infer Shallow-Transfer Machine Translation Rules, Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing {FinTAL}, volume 4139 of Lecture Notes in Computer Science, Springer-Verlag, 756–767.

Sanchez-Martinez, F., Perez-Ortiz, J. A. & M. L. Forcada (2007): Integrating corpus-based and rule-based approaches in an open-source machine translation system, Proceedings of METIS-II Workshop: New Approaches to Machine Translation, Leuven, 73–82.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & J. Makhoul (2006): A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas, AMTA, 223–231.

Spencer, A. (1991): Morphological Theory. Oxford, Blackwell Publishing.

SVLR – Služba za lokalno samoupravo in regionalno Politiko (2006): Slovenija – Hrvaška Operativni Program.

Tiedeman, J. (2012): Parallel Data, Tools and Interfaces in OPUS, 8th International Conference on Language Resources and Evaluation (LREC'2012). Istanbul, ELRA, 1–8.

Toporišič, J. (2000): Slovenska slovnica. Maribor, Založba Obzorja.

Tyers, F. M. & M. Alperen (2010): A parallel corpus of Balkan languages, MultiLR Workshop at LREC2010, Malta.

Tyers, F. M., Sánchez-Martìnez, F., Ortiz-Rojas, S. & M. Forcada (2010): Free/open-source resources in the Apertium platform for machine translation research and development. The Prague Bulletin of Mathematical Linguistics, 93 (93), 67–76.

Vargas-Sierra, C. & D. Lindemann (2013): Bilingual Lexicography and Corpus Methods: The Example of German-Basque as Language Pair. Procedia – Social and Behavioral Sciences, 249–257.

Vičič, J. (2008): Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages. Language technologies: proceedings of the conference, Ljubljana, Institut Jožef Stefan, 98–103.

Vičič, J. (2009): Metode hitre izdelave gradiv za prevajalne sisteme plitkega prenosa za visoko pregibne jezike. V: Mikolič, V. (ur.): Jezikovni korpusi v medkulturni komunikaciji. Koper, Založba Annales, 133–153.

Vičič, J. & P. Homola (2010): Speeding up the Implementation Process of a Shallow Transfer Machine Translation System. In Proceedings of the 14th (EAMT) Conference, Saint Raphael, EAMT, 261–268.

Vičič, J. (2012): Hitra postavitev prevajalnih sistemov na osnovi pravil za sorodne naravne jezike. Doktorska disertacija. Ljubljana.

Vičič, J. & V. Kuboň (2015): A comparison of MT methods for closely related languages: A case study on Czech – Slovak and Croatian – Slovenian language pairs, Text, Speech, and Dialogue: TSD. Plzen, Springer Verlag, 216–224.

Koše dela; žena dela "žoke" (nogavice), Robidišče (foto: Jernej Šušteršič, 1951; Vir: Slovenski etnografski muzej, http://www.etno-muzej.si/sl

original scientific article
received: 2016-07-05

DOI 10.19233/ASHS.2016.58

OBELEŽJI V SPOMIN DEPORTIRANIM IZ JULIJSKE KRAJINE PO DRUGI SVETOVNI VOJNI V GORIŠKEM PARKU SPOMINA

Urška LAMPE
Inštitut Nove revije, zavod za humanistiko, Gospodinjska ulica 8, 1000 Ljubljana
e-mail: urskalampe@gmail.com

IZVLEČEK

V goriškem Parku spomina stojita dve obeležji posvečeni spominu na deportacije iz časa po drugi svetovni vojni. Na podlagi krajše zgodovinske analize dogodkov iz maja 1945, ko je prišlo do dogodkov, poznanih kot deportacije iz Julijske krajine, in zgodovinskega trenutka nastanka obeh obeležij (prvo je bilo postavljeno leta 1960, drugo pa leta 1985/86) avtorica opozarja na historično netočnost in zavajajočo sporočilnost spomenikov, predvsem drugega. Namen prispevka je tudi poudariti pomen ne samo zgodovinopisne obravnave komemoracij in spominskih obeležij, temveč predvsem natančnega poznavanja dogodkov, ki jih ti artefakti obeležujejo. Zgodovinarji morajo na s historiografskega vidika napačne interpretacije dogodkov opozoriti, saj poleg tega, da vodijo v izkrivljanje zgodovine, tudi neprestano generirajo nacionalne konflikte v obmejnem prostoru.

Ključne besede: deportacije, Gorica, Park spomina, lapidarij, nacionalni konflikti, komemoracije, 1945, 1960, 1985/86

I DUE MONUMENTI IN MEMORIA DEI DEPORTATI DALLA VENEZIA GIULIA DEL SECONDO DOPOGUERRA NEL PARCO DELLA RIMEMBRANZA DI GORIZIA

SINTESI

A Gorizia, nel Parco della Rimembranza sono collocati due monumenti lapidari in ricordo alle persone deportate nel secondo dopoguerra da parte delle autorità jugoslave. Sulla base di una breve analisi degli eventi del maggio 1945, quando si verificarono le deportazioni dalla Venezia Giulia, e del momento storico nel quale i due monumenti vennero eretti (il primo nel 1960, il secondo nel 1985/86), l'autrice del saggio richiama l'attenzione sulle imprecisioni storiche e sul messaggio fuorviante dei due monumenti, in particolare del secondo. L'intento è di sottolineare non solo l'importanza dello studio storico delle commemorazioni e dei monumenti, ma in particolare della precisa conoscenza degli eventi che questi artefatti aspirano a ricordare. Risulta dunque necessario che gli storici mettano in discussione le imprecisione delle interpretazioni storiche degli eventi che conducono non solo alla deformazione della storia, ma costantemente generano contrasti nazionali nelle zone di confine.

Parole chiave: deportazioni, Gorizia, Parco della Rimembranza, lapidario, contrasti nazionali, commemorazioni, 1945, 1960, 1985/86