FILOLOGIJA 38–39, Zagreb 2002

UDK 811.163.42:811.111'374.82
Izvorni znanstveni članak
Primljen 31. I. 2002.
Prihvaćen za tisak 20. V. 2002.

Marko Tadić*, Krešimir Šojat**
* *Odsjek za lingvistiku*
** *Zavod za lingvistiku*
Filozofski fakultet Sveučilišta u Zagrebu
Ivana Lučića 3, HR-10000 Zagreb

IDENTIFIKACIJA PRIJEVODNIH EKVIVALENATA U HRVATSKO-ENGLESKOM PARALELNOM KORPUSU

U radu se ispituju mogućnosti identificiranja prijevodnih ekvivalenata u Hrvatsko-engleskom paralelnom korpusu sravnjenu na razini rečenica koji je sastavljen u Zavodu za lingvistiku Filozofskog fakulteta u Zagrebu. Identifikacija prijevodnih ekvivalenata od po jedne riječi u svakom jeziku postignuta je generiranjem svih mogućih prijevodnih parova riječi unutar svake rečenice sravnjene u odnosu 1:1. Na te parove primijenjen je izračun statističke mjere *uzajamne obavijesnosti* (*mutual information*) koja daje podatak o statistički relevantnim supojavljivanjima riječi u takvu paru. Parovi s visokom vrijednošću uzajamne obavijesnosti predstavljaju dobre kandidate za prijevodne ekvivalente. U nastavku rada provodi se identifikacija jedinica sastavljenih od više riječi u izvornim tekstovima (hrvatskima) i traže se supojavljivanja parova riječi i u jeziku prijevoda (engleski). Tako se otkrivaju karakteristični prijevodni uzorci jer postoje parovi riječi u izvornom jeziku koji se redovito prevode istim parom riječi u ciljnome jeziku premda je u svakom jeziku vrijednost uzajamne obavijesnosti izrazito niska.
Korištenjem statističkih postupaka u paralelnim korpusima olakšava se pronalaženje kolokacija (mogućih višerječnih termina) i njihovih prijevoda. Istodobno se omogućuje uvid u odgovarajuće ko-tekstne primjere uporabe riječi u izvornom jeziku i u jeziku prijevoda. Tako priređena građa vrlo je korisna dvojezičnim leksikografima i prevoditeljima.

1. Uvod

U suvremenoj leksikografiji korištenje računalnih korpusa kao pretežnih izvora jezične građe za potvrde stvarne uporabe jezičnih jedinica postaje nezaobilazno. Za razliku od jednojezičnih korpusa koji svoju leksikografsku primjenu nalaze ponajprije u jednojezičnoj leksikografiji, u dvojezičnoj sve važniju ulogu imaju paralelni korpusi, odnosno korpusi tekstova na dvama ili više jezika (izvornih tekstova i njihovih prijevoda). Tek s pomoću tako priređenih korpusa postao je moguć korpusnolingvistički pristup višejezičnoj

građi. Svi podatci i računalni pristupi korpusnoj građi do sada prisutni u jednojezičnoj leksikografiji, kao što su abecedni i čestotni popisi riječi, uvid u kotekst, pronalaženje kolokacija, fraza, idioma, višerječnih jedinica (*multi-word units*), time su dobili dodatnu, višejezičnu dimenziju. Sad se podatci i odnosi unutar jednoga korpusa mogu promatrati i s obzirom na njegov prijevod tj. taj isti tekst iskazan u drugome jeziku. Sravnjeni (*aligned*) paralelni korpusi omogućuju to u još znatnijoj mjeri jer je usporednost takvih dokumenata iskazana eksplicitno. Ako je korpus sravnjen na razini rečenice, onda se točno zna koja je izvorna rečenica prevedena kojom ciljnom i u kakvu su one odnosu.

Osnovno je polazište ovoga rada provjera mogućnosti uporabe statističkih metoda pronalaženja prijevodnih ekvivalenata (*translational equvalents, TE*) na razini riječi i parova riječi u paralelnom korpusu tipološki relativno udaljenih jezika kao što su hrvatski i engleski.

2. Korpus

Hrvatsko-engleski paralelni korpus sastavljen je u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.[1] Riječ je o jednosmjernom paralelnom korpusu s izvornim jezikom hrvatskim i ciljnim jezikom engleskim. Tekstovi uvršteni u korpus pribavljeni su iz jednoga izvora: tjednika *Croatia Weekly* koji je izlazio u Zagrebu na engleskome jeziku od siječnja 1998. do travnja 2000. Zavodu je bio dostupan tekst u digitalnom zapisu 113 brojeva tjednika na hrvatskom i na engleskom. Korpusni parametri:

	hrvatski	engleski
članaka	4.748	4.748
rečenica	74.638	82.898
pojavnica	1.636.246	1.968.874

Tablica 1: Korpusni parametri Hrvatsko-engleskoga paralelnoga korpusa

Korpus je sravnjen na rečeničnoj razini programom *Vanilla aligner* autora Daniela Ridingsa i Pernille Danielsson (inačicom za DOS).[2] Statistika sravnjivanja:

[1] O postupcima sastavljanja korpusa vidi u Tadić 2000 i Tadić 2001.

[2] O programu Vanilla Aligner vidi u Danielsson–Ridings 1997.

0:1	250	0,35%
1:0	19	0,03%
1:1	58788	83,52%
1:2	9374	13,32%
2:1	1529	2,17%
2:2	432	0,61%
ukupno	70392	100,00%

Tablica 2: Statistika sravnjenja Hrvatsko-engleskoga paralelnoga korpusa

Iz tablice 2 razvidno je da je više od 83% hrvatskih rečenica prevedeno jednom engleskom rečenicom. Taj je podatak iznimno bitan za daljnje istraživanje predstavljeno u ovom radu.

```
*** Länk: 1 – 1 ***
<BODY> <DIV0 type="MAIN"> <HEAD type="NA"> <S id="CW0101998031902O1hr.S1"> Do 1.
kolovoza zabranjeni skupovi u istoènoj Slavoniji </S> </HEAD> .EOS
<BODY> <DIV0 type="MAIN"> <HEAD type="NA"> <S id="CW0101998031902O1en.S1">
POLITICAL RALLIES IN EASTERN SLAVONIA BANNED UNITL AUGUST 1 </S> </HEAD> .EOS

.EOP

*** Länk: 2 – 1 ***
<HEAD type="PN"> <S id="GW010199803190201hr.S2"> Vlada je ocijenila kako je provođenje
mirne reintegracije Podunavlja jedan od poglavitih interesa Hrvatske </S> <S
id="GW010199803190201hr.S3"> Stoga, treba izbjeæi svaki èin koji bi mogao dovesti do
naru šavanja reda i sigurnosti ljudi </S> </HEAD> .EOS
<HEAD type="PN"> <S id="GW010199803190201en.S2"> The Government has assessed that the
implementation of peaceful reintegration in Eastern Slavonia is one of Croatia's priority interests,
therefore, any act that might endanger order and public safety should be avoided </S> </HEAD>
.EOS

.EOP

*** Länk: 1 – 1 ***
<P> <S id="GW010199803190201hr.S4"> Vlada Republike Hrvatske obvezala je ministra
unutarnjih poslova da svojim obveznim naputkom nalož i policijskim upravama na području istočne
Slavonije da sukladno odredbi Zakona o javnom okupljanju zabrane do 1. kolovoza 1998. sva javna
okupljanja. </S> .EOS
<P> <S id="GW010199803190201en.S4"> The Croatian Government has charged the Interior
Minister with the task of issuing a compulsory order instructing local police departments in Eastern
Slavonia to ban all public assemblies until August 1, in accordance with the relevant provision of
the Public Assembly Act. </S> .EOS

*** Länk: 1 – 1 ***
<S id="GW010199803190201hr.S5"> Ta odredba zapravo se odnosi na zabranu organiziranja
skupova politièkih stranaka. </S> </P> .EOS
<S id="GW010199803190201en.S4"> The provision actually refers to a ban on theorganization of
gatherings by political parties. </S> <P> .EOS

.EOP
```

Slika 1: Primjer sravnjivanja korpusa programom Vanilla Aligner.

2.1. Potkorpus

Kako bi obradba čitavoga korpusa metodom koja se iznosi u 5.1. i 6.1. rezultirala iznimno velikom količinom podataka te bi dovela do »kombinatorne eksplozije«, donesena je odluka da se istraživanje obavi ne na cijelome korpusu već na njegovom dijelu. Odabrana je sedma stranica svih 113 brojeva *Croatia Weeklyja* na kojoj su bili objavljivani tekstovi iz gospodarske rubrike. Potkorpus je tako odabran zbog očuvanja terminološke dosljednosti unutar iste rubrike i zbog mogućnosti provjere statističkih metoda na stvarnom tekstu tj. tekstu koji uključuje stanovit broj imena i brojeva.

	hrvatski	engleski
članaka	404	404
rečenica	8.420	9.373
pojavnica	195.510	234.365

Tablica 3: Korpusni parametri odabranoga potkorpusa Hrvatsko-engleskoga paralelnoga korpusa

Ukupan broj sravnjenja u potkorpusu iznosi 8187, a od toga je 6786 rečenica sravnjeno u obliku 1:1. U tih 6786 rečenica nalazi se 202.081 hrvatskih i 241.376 engleskih pojavnica i to je uzorak nad kojim je obavljeno istraživanje.

3. Cilj

Cilj je istraživanja bio pronaći moguće prijevodne ekvivalente (TE) koji se pojavljuju u rečenicama sravnjenim u obliku 1:1. Problem se teorijski može razbiti na potprobleme koji se potom rješavaju u odvojenim koracima:

- pronaći TE_{11} tj. prijevodne ekvivalente oblika:
  1 hrvatska riječ : 1 engleska riječ
- pronaći TE_{22} tj. prijevodne ekvivalente oblika:
  2 hrvatske riječi : 2 engleske riječi[3]
- ...
- pronaći TE_{xy} tj. prijevodne ekvivalente oblika:
  x hrvatskih riječi : y engleskih riječi (gdje je $x <=> y$)

U ovom radu ograničili smo se samo na prva dva slučaja, tj. proučavane su mogućnosti identifikacije samo za TE_{11} i TE_{22}.

[3] U ovom su radu uzimani samo izravni parovi riječi, tj. parovi riječi koje su se u tekstu pojavile neposredno jedna iza druge. Takav je par odabran samo unutar rečenica, tj. unutar njega se nije mogla pojaviti rečenična granica.

4. Dosadašnji radovi

S pojavom velikih količina višejezičnoga teksta u digitalnom obliku pojavila su se i istraživanja temeljena na proučavanju paralelnih korpusa. Identifikacija prijevodnih ekvivalenata jedno je od područja na kojem se proučavanje podataka iz paralelnih korpusa pokazalo iznimno korisnim. Ako se podatak o identificiranome TE unese natrag u paralelni korpus, onda se često u literaturi (npr. Tiedemann 1999, Ahrenberg i dr. 1998) govori i o sravnjivanju riječi (*word alignment*). Identificiranje TE sastoji se zapravo od prolaženja prijevodnih podudarnosti na dvije različite razine:

1. razina pojedinačnih riječi (*single-word units*) = pronalaženje odnosa W_{L1}:W_{L2}
2. razina višerječnih jedinica (*multi-word units*) = pronalaženje odnosa W_{1L1}-W_{2L1} i W_{1L2}-W_{2L2} te potom pronalaženje W_{1L1}-W_{2L1} : W_{1L2}-W_{2L2}

Druga razina uključuje identifikaciju višerječnih jedinica bilo u izvornom, bilo u ciljnom jeziku, a tek potom i uspostavljanje TE među takvim jedinicama.

Sustavi za identifikaciju TE mogu za cilj imati pronalaženje što većeg broja parova (npr. Ahrenberg i dr. 1998, Melamed 1999) ili mogu biti usmjereni na što veću točnost identificiranih parova što zapravo vodi stvaranju dvojezičnih leksikona (npr. Tiedemann 1998 i 1999). Pozornost se, nadalje, može usmjeriti na samo jednu posebnu vrstu parova – npr. termine (Dagan & Church 1994) ili kolokacije (Smadja i dr. 1996).

Melamed (2000:227) tvrdi da se »gotovo svi sustavi sa [statističkim] funkcijama sličnosti, koji se mogu naći u literaturi, temelje na nekakvom modelu supojavljivanja uz primjenu nekakva lingvistički uvjetovana filtra«. Dakle, uz statističku obradu korpusnih podataka, rabe se i lingvistički filtri (najčešće na morfološkoj razini ili razini jednostavnih sintaktičkih konstrukcija) koji ograničuju ulazni ili izlazni skup podataka. Takvi se filtri mogu pojaviti bilo pri pronalaženju kolokacija u jednom od jezika, bilo pri uspostavljanju TE.

Uz gotovo isključivo statistički utemeljene pristupe (npr. Gale & Church 1991, Smadja i dr. 1996, Dagan i dr. 1999) mogući su i sustavi koji se služe nekim oblikom lingvističkoga znanja pri samom prikupljanju podataka iz korpusa i za samu njihovu obradu (npr. Daille 1996, Hatzivassiloglou 1996, Jacquemin 2001).

Metoda koja se iznosi u ovom radu u cijelosti se oslanja na statistički pristup bez uporabe ikakva lingvističkoga filtra prije ili poslije obrade. Moguće primjene filtara razmatrat će se u budućim radovima.

5. Pronalaženje TE_{11}

5.1. Metoda

Postupak pronalaženja TE_{11} koji je razvijen i primijenjen u ovom istraživanju može se prikazati s pomoću slijeda potpostupaka:

1. generirati sve moguće parove pojavnica iz rečenica u sravnjenju 1:1. Riječ je o jednostavnom Kartezijevu umnošku dviju rečenica obavljenom na razini riječi:

	hr rečenica	en rečenica	hr-en parovi
1. riječ	a	w	aw, ax, ay, az
2. riječ	b	x	bw, bx, by, bz
3. riječ	c	y	cw, cx, cy, cz
4. riječ		z	

Tablica 4: Shema generiranja mogućih TE_{11} parova

Tako bismo npr. za sravnjene rečenice *Ivan jede jabuku.* i *John eats an apple.* dobili parove mogućih TE_{11}: *Ivan:John, Ivan:eats, Ivan:an, Ivan:apple, jede:John, jede:eats, jede:an, jede:apple, jabuku:John, jabuku:eats, jabuku:an, jabuku:apple.*

2. na tako dobivene parove primijeniti izračun statističke mjere koja svojom vrijednošću otkriva parove koji su dobri kandidati za stvarni TE;
3. poredati parove prema izračunatoj vrijednosti i odabrati stvarne TE za leksikografsku uporabu.

U našem slučaju korištena je statistička mjera *uzajamna obavijesnost* (*mutual information, MI*), koja se često pojavljuje i u dosadašnjim radovima s toga područja. Manning i Schütze (1999:178) daju detaljnu definiciju MI. Odatle je preuzeta i formula za izračun MI (preciznije: *pointwise mutual information*) u ovome radu koja glasi $MI = \log_2 \frac{P(x,y)}{P(x)P(y)}$. Uz stanovit oprez pri uporabi[4], ponajprije pri niskoučestalim jezičnim jedinicama, daje nadasve upotrebljive rezultate.

[4] Vidi u McEnery i dr. 1997:222 i u bilješci 6, o derivatu MI tzv. MI^3. Vidi u Smadja–McKeown–Hatzivassiloglou 1996:8–14 o ostalim definicijama MI kao i o mogućim problemima vezanim uz njezinu uporabu. O problemima također vidi u Manning–Schütze 1999:181 kao i u Gale–Church 1991.

5.2. Rezultati

Obrada potkorpusa opisanoga u 2.1. prema metodi definiranoj u 5.1. donijela je sljedeće rezultate:

- 202.081 hrvatskih i 241.376 engleskih pojavnica iz 6.786 rečenica dalo je 26.166 hrvatskih i 13.234 engleskih različnica te 13.222 hrvatskih lema i 10.403 engleske leme[5];
- generiranjem parova dobiveno je ukupno 4.819.953 parova pojavnica, od toga 1.944.377 različitih parova pojavnica, a od njih 1.407.727 različitih parova lema;
- međutim, od gotovo milijun i pol različitih parova lema svega ih je 262.858 imalo čestotu veću od 2. Prag je postavljen na toj razini te su u daljnje vrednovanje rezultata ušli samo parovi lema s čestotom većom od 2.

čestota	parova	točnih TE	%
>9	18	18	100,0
>8	32	32	100,0
>7	53	53	100,0
>6	99	91	91,9
>5	169	146	86,3
>4	293	235	80,2
>3	534	413	77,3
>2	1081	825	77,0

Tablica 5: Rezultati vrednovanja parova kandidata za stvarni TE_{11}

Iz tablice 5 uočljivo je da s padom čestote para pada i točnost (*precision*) para kao pravog TE, ali raste i odziv (*recall*), tj. sve je veći broj pronađenih mogućih parova.

[5] U Tadić–Fulgosi–Šojat (u tisku) pokazano je kako su izračuni MI na lematiziranom korpusu davali u prosjeku 4,5% posto točnije rezultate od izračuna MI na različnicama tj. na nelematiziranom korpusu. Stoga je i ovdje čitava obrada obavljena na lematiziranom uzorku tj. potkorpusu. Sam je postupak lematizacije obavljen poluautomatski, uporabom lemarija dobivenog pri sastavljanju *Hrvatskoga čestotnog rječnika* (Moguš–Bratanić–Tadić 1999) i probne inačice imeničnoga dijela *Hrvatskoga morfološkoga leksikona*, koji se također sastavlja u Zavodu za lingvistiku.

L1	L1	L2	L2	L12	L12	MI_L
željezara	9	željezara	9	željezara željezara	8	9,3884957118
lokomotiva	13	locomotive	9	lokomotiva locomotive	11	9,3174126138
seminar	9	seminar	12	seminar seminar	10	9,2953863074
kaznen	12	criminal	9	kaznen criminal	10	9,2953863074
vegeta	11	vegeta	11	vegeta vegeta	11	9,2689140961
crnogorski	11	montenegrin	8	crnogorski montenegrin	8	9,2689140961
ivo	10	ivo	11	ivo ivo	10	9,2689140961
bedekovčina	10	bedekovčina	10	bedekovčina bedekovčina	9	9,2544145264
kw	10	kw	9	kw kw	8	9,2364926184
17,5	10	17.5	9	17,5 17.5	8	9,2364926184
1,9	12	1.9	12	1,9 1.9	12	9,143383214
benetton	10	benetton	12	benetton benetton	10	9,143383214
2,9	12	2.9	12	2,9 2.9	12	9,143383214
2,4	12	2.4	10	2,4 2.4	10	9,143383214
hektar	12	hectare	8	hektar hectare	8	9,143383214
MW	11	mw	11	MW mw	10	9,1314105723
kulturan	11	cultural	10	kulturan cultural	9	9,1169110026
55	11	55	10	55 55	9	9,1169110026
7,5	11	7,5	10	7,5 7.5	9	9,1169110026
libor	11	libor	10	libor libor	9	9,1169110026
španjolski	17	spanish	8	španjolski spanish	11	9,1003144921
disciplina	9	discipline	11	disciplina discipline	8	9,0989890946
lužavac	9	lužavec	11	lužavac lužavec	8	9,0989890946
kalogjera	10	kalogjera	10	kalogjera kalogjera	8	9,0844895249
anketa	10	survey	13	anketa survey	10	9,0279059966
billa	9	billa	13	billa billa	9	9,0279059966
2,7	13	2.7	12	2,7 2.7	12	9,0279059966
produktivnost	8	productivity	13	produktivnost productivity	8	9,0279059966
0,8	13	0.8	13	0,8 0.8	13	9,0279059966
recesija	13	recession	13	recesija recession	13	9,0279059966
cigareta	11	cigarette	13	cigareta cigarette	11	9,0279059966
bouygues	12	bouygues	11	bouygues bouygues	10	9,0058796902

Slika 2: Izvadak iz popisa parova-kandidata za TE_{11} poredanog po padajućoj vrijednosi MI

5.3. Problemi

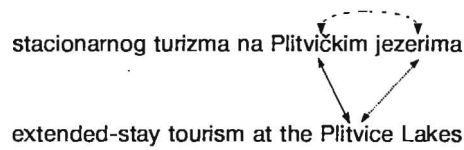
Tako obavljeno istraživanje i dobiveni rezultati nesumnjivo nukaju na stanovita pitanja koja pak zahtijevaju odgovore.

Jesu li svi pronađeni parovi stvarni TE? Kriterij za vrednovanje bio je znanje oba jezika u hrvatskih izvornih govornika koji su istodobno izvrsni znalci engleskoga. Strogo metodološki gledano, kako tu nije bilo dodatne provjere ispravnosti rezultata, moglo bi se tvrditi da su oni podložni individualnoj varijaciji znanja ocjenjivača.

Drugo bi se pitanje moglo usmjeriti prema zastupljenosti brojeva tj. znamenaka i imena u popisu parova. Prevode li se imena i brojevi doista i može li se u tom slučaju govoriti o pravim TE? Razlozi za uključivanje brojeva mogu se, osim već spomenutoga razloga provjere metode na stvarnom a ne na pročišćenom tekstu, argumentirati različitošću pravopisnih pravila za brojeve u hrvatskome (decimalni zarez) i engleskome (decimalna točka). Metoda je ponudila za TE parove brojeva pisanih različitim pravopisom sa 100% točnošću te

je tako bez ikakve sadržajne analize opravdala svoju svrhu. Kod imena se u hrvatskome javljaju morfološke varijacije (razriješene postupkom lematizacije na polaznu lemu), ali i jezičnospecifični nazivi za npr. toponime kao što su Beč : Vienna; Bruxelles : Brussels itd. U tom slučaju metoda daje rezultate kao i za bilo koji drugi par pravih prijevoda te se tretman ni brojeva ni imena ne razlikuje od ostalih pojavnica.

Uočljive su, međutim, karakteristične pogreške nastale uslijed »indirektnih parova« nastalih pri uparivanju elemenata čestih kolokacija bilo u izvornom, bilo u ciljnom, ili u oba jezika kao što su *Plitvička:Plitvice, Plitvička:lakes, jezera:Plitvice, jezera:lakes*. Taj je problem poznat iz literature (npr. Melamed 2000: :227) može ga se na našem primjeru prikazati sljedećim grafikonom:



Slika 3: Primjer direktnih (puna crta) i indirektnih (točkasta crta) parova[6] i kolokacija (crtkana crta)

Bi li se nekako mogli izbjeći indirektni parovi? Bi li se to moglo postići uparivanjem kolokacija iz izvornoga jezika s kolokacijama u ciljnome jeziku? Taj pokušaj čini drugi dio istraživanja.

6. Pronalaženje TE_{22}

6.1. Metoda

Postupak pronalaženja parova kolokacija tj. parova parova (TE_{22}) zapravo je derivat TE_{11} na višoj razini kompleksnosti. Kao i TE_{11} može se prikazati s pomoću slijeda potpostupaka:

1. generirati sve moguće parove parova pojavnica iz rečenica u sravnjenju 1:1:

	hr rečenica	en rečenica	hr-en parovi parova
1. riječ	a	v	abvw, abwx, abxy, abyz
2. riječ	b	w	bcvw, bcwx, bcxy, bcyz
3. riječ	c	x	cdvw, cdwx, cdxy, cdyz
4. riječ	d	y	
5. riječ		z	

Tablica 6: Shema generiranja mogućih TE_{22} parova

[6] Grafikon je adaptiran prema Melamed 2000:227, samo su postave preuzete iz Hrvatsko-engleskoga paralelnog korpusa.

Tako bismo npr. za sravnjene rečenice *Ivan jede crvenu jabuku.* i *John eats the red apple.* dobili parove mogućih TE_{22} : *Ivan jede : John eats, Ivan jede : eats the, Ivan jede : the red, Ivan jede : red apple, jede crvenu : John eats, jede crvenu : eats the, jede crvenu : the red, jede crvenu : the apple, crvenu jabuku : John eats, crvenu jabuku : eats the, crvenu jabuku : the red, crvenu jabuku : red apple.*

2. na tako dobivene parove parova primijeniti izračun MI;
3. poredati parove prema izračunatoj vrijednosti MI i odabrati stvarne TE za leksikografsku uporabu.

Valja napomenuti da pri određivanju MI parova parova nije korišten podatak o vrijednosti MI između elemenata istojezičnoga para. Jedini parametar u formuli za MI bila je čestota para parova.

6.2. Rezultati

Obrada parova parova lema u potkorpusu definiranom u 2.1. dala je sljedeće rezultate:

- hrvatsko-hrvatskih različitih parova lema pronađeno je 76.078, a englesko-engleskih 67.322;
- generiranjem parova parova dobiveno je ukupno 4.492.171 parova parova lema, od toga 4.157.880 različitih hrvatsko-engleskih parova parova lema;
- međutim, svega je 194.185 parova parova lema imalo čestotu veću od 1. Prag je postavljen na nešto višoj razini nego kod TE_{11} jer, kao što će se vidjeti u sljedećoj tablici, postotak pogreške znatno je veći kod TE_{22};
- nadalje, parovi su poredani prema padajućoj vrijednosti izračunate MI, a prag za vrednovanje postavljen je ponovno na MI => 9. Pri vrednovanju rezultata uzimane su u obzir frekvencija para parova i frekvencija istojezičnoga para.

čestota	parova	točnih TE	%
>9	28	24	85,7
>8	46	33	71,7
>7	76	55	71,4
>6	138	88	63,7
>5	258	143	55,4
>4	446	226	50,6
>3	923	392	42,4

Tablica 7: Rezultati vrednovanja parova kandidata za stvarni TE_{22}

Iz tablice 7 uočljivo je da netočnost pada znatno brže nego kod TE_{11} te se već kod čestote >3 spušta ispod 50%. Time se sasvim izvjesno može ustvrditi da MI ne daje dobre rezultate za niskofrekventne parove.

Na slici 4 (na sljedećoj stranici) mogu se uočiti neobično visoke vrijednosti MI (>9) između nekih parova parova kao što su npr. *posto za : percent for, 30 milijun : 30 million* itd. iako su MI vrijednosti između elemenata istojezičnoga para izrazito niske (<0). Shematski bi se taj odnos mogao prikazati ovako:

$$(W_{1L1}W_{2L1}(-MI) : W_{1L2}W_{2L2}(-MI))\,(MI=>9)$$

i on je sasvim neočekivan. Jedno o mogućih objašnjenja takva ponašanja MI moglo bi se prikazati obosmjernošću odnosa elemenata u MI. Svaki element koji ulazi u formulu zapravo govori o vjerojatnosti supojavljivanja drugoga elementa. Ta argumentacija, međutim, ne objašnjava zašto je u tim slučajevima zapravo detektiran pravi TE_{22}. Nije li možda riječ o sredstvu za detekciju karakterističnih prijevodnih obrazaca, tj. kombinacija riječi u izvornome jeziku koje se redovito prevode istim parom riječi u ciljnom jeziku bez obzira što svaki od parova u svom jeziku nije nipošto statistički relevantan. Svaki takav istojezični par izvornoga jezika postaje relevantan tek kad se supostavi s istim takvim istojezičnim parom u ciljnom jeziku i kad se u paru parova otkrije prijevodna pravilnost.

6.3. Problemi

Kao i u slučaju TE_{11}, uočljive su, međutim, karakteristične pogreške kad su parovi TE_{22} dijelovi kolokacija duljih od 2 riječi npr.:

- 2:3 *krapinski neandertalac : the Krapina Neandertal;*
- 3:2 *konferencija za novinare : press conference;*
- 3:3 *Europska monetarna unija : European monetary union;*
- 5:3 *prodaja na veliko i malo : wholesale and retail;*
- 5:4 *Hrvatska agencija za promicanje ulaganja : Croatian investment promotion agency;*
- itd.

Te bi se pogreške mogle razriješiti obradama na daljnjim razinama kompleksnosti kao što su TE_{12}, TE_{21}, TE_{23}, TE_{32}, TE_{33} itd.

Također je jedan od ozbiljnijih nedostataka te metode generiranja parova mogućih TE svojevrsno »pregeneriranje parova« koje se pojavljuje kad se jedna riječ pojavi više puta bilo u izvornom, bilo u ciljnom jeziku. To se može najlakše uočiti u slučajevima kad frekvencija para nadilazi pojedinačne frekvencije elemenata u paru. Shematski se to može prikazati tablicom 8.

P11P12	P11	P11P12_MI	P21P22	P21	P21P22_MI
19 milijun	9	0,3082685347	19 milllion	9	0,6391527362
plav zastava	11	6,9233347331	blue flag	12	8,0559203727
jahta i	9	-0,766065201	yacht and	11	-1,413529193
300 milijun	9	0,4269130312	300 million	9	0,3838956809
16 milijun	9	-0,116229294	16 million	9	0,235797042
promicanje ulaganje	9	3,4754175202	croatian investment	9	-3,621941259
promicanje ulaganje	9	3,4754175202	promotion agency	9	2,8055135752
1,8 posto	9	0,7654497094	1.8 percent	9	0,6694520257
sezonski proizvod	10	4,3778484676	seasonal product	10	4,3166347267
milijun tona	15	0,5927219242	million ton	14	0,4494192537
damir begović	10	7,0672802349	damir begović	10	6,8359546888
promicanje ulaganje	9	3,4754175202	investment promotion	10	1,6749628121
2,5 posto	10	-8,5037805	2.5 percent	9	-0,096082721
za promicanje	10	0,2736865634	promotion agency	9	2,8055135752
posto za	13	-6,223234977	percent for	10	-6,217463870
za promicanje	10	0,2736865634	croatian investment	9	-3,621941259
1,1 posto	11	0,8325639052	1.1 percent	11	0,9589586429
i voda	11	-0,869876007	and water	11	-2,413529193
30 milijun	10	-1,046686361	30 million	10	-0,697591356
18 posto	10	-0,270174200	18 percent	10	-0,026541787
za promicanje	10	0,2736865634	investment promotion	10	1,6749628121
2,8 posto	12	1,3504122101	2.8 percent	11	1,3739961421
25 milijun	12	-0,488198071	25 milllion	12	-0,417878209
35 milijun	12	0,5449687927	35 million	11	0,6349281503
o. slobodan	10	0,4246791083	free trade	12	2,6615837823
obvezan socijalan	10	4,5140265939	compulsory social	12	5,6126684973
slobodan trgovina	10	3,7834872689	free trade	12	2,6615837623
1 siječanj	12	2,8396024658	january 1	10	1,9246181882
obvezan socijalan	10	4,5140265939	and compulsory	12	-0,968438029
deutsche telekom	11	6,3804224113	deutsche telekom	11	7,1516037206
energija plin	10	3,5967459419	and water	11	-2,413529193
dolar malo	11	-1,2781171631	million less	10	-1,680912615

Slika 4. Izvadak iz popisa parova-kandidata za TE_{22} poredanog po padajućoj vrijednosti MI

	hr rečenica	en rečenica	hr-en parovi
1. riječ	a	w	aw, ax, ay, ax, az
2. riječ	b	x	bw, bx, by, bx, bz
3. riječ	a	y	aw, ax, ay, ax, az
4. riječ	c	x	cw, cx, cy, cx, cz
5. riječ		z	

Tablica 8: Shema »pregeneriranja« mogućih TE_{11} parova

Iz tablice 8 može se uočiti da je par *ax* generiran dva puta, iako se i *a* i *x* pojavljuju svega po dva puta u svojim rečenicama. Za potrebe ovoga istraživanja taj smo problem ostavili po strani jer bi znatno usložnio samo generiranje

P1P1P2P2	P11	P1P1P2P2_MI
19 milijun 19 milllion	11	9,8479273304637
plav zastava blue flag	15	9,5908421909610
jahta i yacht and	11	9,5584207132687
300 milijun 300 million	9	9,5584207132687
16 milijun 16 million	9	9,5584207132687
promicanje ulaganje croatian investment	9	9,5584207132687
promicanje ulaganje promotion agency	9	9,5584207132687
1,8 posto 1.8 percent	9	9,5584207132687
sezonski proizvod seasonal product	11	9,5439211435736
milijun tona million ton	22	9,4735318156822
damir begović damir begović	10	9,4064176198236
promicanje ulaganje investment promotion	9	9,4064176198236
2,5 posto 2.5 percent	9	9,4064176198236
za promicanje promotion agency	9	9,4064176198236
posto za percent for	13	9,4064176198236
za promicanje croatian investment	9	9,4064176198236
1,1 posto 1.1 percent	11	9,2689140960737
i voda and water	11	9,2689140960737
30 milijun 30 million	9	9,2544145263786
18 posto 18 percent	9	9,2544145263786
za promicanje investment promotion	9	9,2544145263786
2,8 posto 2.8 percent	11	9,1433832139896
25 milijun 25 milllion	12	9,1433832139896
35 milijun 35 million	11	9,1433832139896
o. slobodan free trade	10	9,1433832139896
obvezan socijalan compulsory social	10	9,1433832139896
slobodan trgovina free trade	10	9,1433832139896
1 siječanj january 1	10	9,1433832139896
obvezan socijalan and compulsory	10	9,1433832139896
deutsche telekom deutsche telekom	10	9,1314105723237
energija plin and water	9	9,1169110026286
dolar malo million less	9	9,1169110026286

Slika 4 — nastavak.

parova. Taj je problem rješiv dodatnom provjerom frekvencije pojedinih jedinica unutar rečenice te u slučaju ponavljanja neke jezične jedinice, valja obaviti provjeru »pregeneriranja parova« u kojima se nalazi ta jedinica.

7. Daljnji smjerovi istraživanja

Daljnje primjene ove za sada isključivo statističke metode mogle bi se projicirati u nekoliko smjerova:

- generiranje ostalih mogućih parova: TE_{12}, TE_{21}, TE_{23}, TE_{32}, TE_{33}, TE_{34}, TE_{43}, TE_{44}...;
- eksperimentiranje s drukčijim pragovima čestote i vrijednosti MI;
- primjena drugih statističkih mjera: npr. Diceov koeficijent, hi-kvadratni test ili logaritamska sličnost (*Log likelihood*);

- reiteracija TE_{11} s ispuštenim već pronađenim TE_{12}, TE_{21}, TE_{22}, čime bi se zapravo uklonili »indirektni parovi«;
- uporaba lingvističkih filtara (kao što su podatak o vrsti riječi ili gramatičkoj kategoriji neke pojavnice) prije primjene statističkih mjera ili nakon toga kako bi se ograničio djelokrug njihova djelovanja na npr. samo kombinacije *pridjev + imenica* itd.

8. Zaključak

Rad je pokazao primjenu jedne od statističkih metoda za pronalaženje mogućih prijevodnih ekvivalenata na temelju sravnjenih paralelnih korpusa. Prijevodni ekvivalenti traženi su na razini jedne riječi TE_{11} kao i na razini parova riječi TE_{22}. Uočena je visoka učinkovitost MI vrijednosti za identificiranje TE_{11} uz nešto lošiju primjenljivost na TE_{22}. Također je pokazano kako statističke metode mogu pripomoći leksikografima u pronalaženju kolokacija tj. kotekstnim primjerima jezične uporabe kako u izvornom tako i u ciljnom jeziku.

Literatura

Ahrenberg, Lars, Mikael Andersson, Magnus Merkel. 1998. A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts. *Proceedings of COLING-ACL '98*, Montreal : ACL. 29–35.

Ahrenberg, Lars, Magnus Merkel, Anna Sågvall Hein, Jörg Tiedemann. 2000. Evaluation of Word Alignment Systems. *LREC2000 Proceedings: Second International Conference on Language Resources and Evaluation*. Eds. M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer. Atena, Pariz : ELRA. 1255–1261.

Armstrong, Susan, Kenneth W. Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, David Yarowsky (eds.). 1999. *Natural Language Processing Using Very Large Corpora*. Dordrecht : Kluwer.

Barnbrook, Geoff, Permilla Danielsson, Michaela Mahlberg (eds.) (u tisku). *Meaningful Texts : The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham : Birmingham University Press.

Choueka, Yaacov, Ehud S. Conley, Ido Dagan. 1998. A Comprehensive Bilingual Word Alignment System – Application to Disparate Languages. *Parallel Text Processing : Alignment and use of Translation Corpora*. Ed. J. Veronis. Dordrecht : Kluwer. 69–96.

Dagan, Ido, Kenneth W. Church. 1994. Termight : Identifying and Translating Technical Terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing ANLP-94*, ACL : Stuttgart. Str. 34–40.

Dagan, Ido, Kenneth W. Church, William A. Gale. 1999. Robust Bilingual Word Alignment for Machine Aided Translation. U zb. Armstrong i dr. 1999, 209–224.

Daille, Béatrice. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. U zb. Klavans–Resnik 1996, 49–66.

Danielsson, Pernilla, Daniel Ridings. 1997. *Practical Presentation of a »Vanilla«*

Aligner, 1999-10-11, *http://nl.ijs.si/telri/Vanilla/doc/ljubljana.html*, 2001-10-03.

Frantzi, Katerina T., Sophia Ananiadou, Junichi Tssujii. 1999. Automatic Classification of Technical Terms using the NC-value Method for Term Recognition. *Papers in Computational lexicography : COMPLEX '99*, eds. Ferenc Kiefer, Gábor Kiss, Júlia Pajzs. Budimpešta : Lingvistički institut Madžarske akademije znanosti. Str. 57–66.

Gale, William A., Kenneth W. Church. 1991. Identifying Word Correspondances in Parallel Texts. *Proceedings DARPA Speech and Natural Language Workshop*, San Mateo : Morgan Kaufmann. 152–157.

Hatzivassiloglou, Vasileios. 1996. Do We Need Linguistics When We Have Statistics? : A Comparative Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System. U zb. Klavans–Resnik 1996, 67–94.

Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, MA, London : MIT Press.

Klavans, Judith K., Philip Resnik (eds.) 1996. *The Balancing Act : Combining Symbolic and Statisticas Approaches to Language*, Cambridge, MA, London : MIT Press.

Manning, Christopher, Heinrich Schütze. 1999. *Foundations of Statistic Natural Language Processing*. Cambridge, MA, London : MIT Press.

Melamed, I. Dan. 1996. Automatic construction of clean broad-coverage translation lexicons. *Proceedings of the 2nd Conference of the Association of Machine Translation in the Americas*. Montreal. 125–134.

Melamed, I. Dan. 1999. Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics* 25:1, 107–130.

Melamed, I. Dan. 2000. Models of Translational Equvalence among Words. *Computational Linguistics* 26:2, 221–249.

Moguš, Milan, Maja Bratanić, Marko Tadić. 1999. *Hrvatski čestotni rječnik*. Zagreb : Školska knjiga *i* Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu.

Resnik, Philip, I. Dan Melamed. 1997. Semi–Automatic Acquisition of Domain-Specific Translation Lexicons. *Proceedings of the 7th ACL Conference on Applied Natural Language Processing*. Washington.

Smadja, Frank. 1993. Retrieving Collocations from Text : XTRACT. *Computational Linguistics* 19:1, 143–177.

Smadja, Frank, Kathleen R. McKeown, Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics* 22:1, 1–38.

Tadić, Marko. 2000. Building the Croatian-English Parallel Corpus. *LREC2000 Proceedings: Second International Conference on Language Resources and Evaluation*, eds. M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer. Atena, Pariz : ELRA. 523–530.

Tadić, Marko. 2001. Procedures in Building the Croatian-English Parallel Corpus. *International Journal of Corpus Linguistics : Special issue*. 2001. Str. 107–123.

Tadić, Marko, Sanja Fulgosi, Krešimir Šojat (u tisku). U zb. Barnbrook i dr. (u tisku), 195–206.

Tiedemann, Jörg. 1998. Extraction of translation equivalents from parallel corpora. *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALI98*, Kopenhagen : Centar za jezične tehnologije i Odsjek za opću i primijenjenu lingvistiku Sveučilišta u Kopenhagenu.

Tiedemann, Jörg. 1999. Word Alignment Step by Step. *Proceedings of the 12th Nordic Conference on Computational Linguistics NODALIDA99*. Trondheim : Sveučilište u Trondheimu.

Tiedemann, Jörg. 2000. Exatracting Phrasal Terms using Bitexts. *Proceedings of the Workshop on Terminology Resources and Computation within the LREC2000*, ed. Key-Sun Choi. Atena : ELRA. 57–63.

Vintar, Špela. 1999. A Lexical Analysis of the IJS–ELAN Slovene-English Parallel Corpus. *Language technologies – multilingual aspects : proceedings of the workshop within the framework of the 32nd Annual Meeting of the Societas Linguistica Europaea*, ed. Špela Vintar. Ljubljana : Filozofska fakulteta. 63–70.

Identification of translational equivalents in Croatian-English parallel corpus

Summary

The contribution is investigating the possibilities of identification of translational equivalents (TE) in Croatian-English parallel corpus aligned at the sentence level and collected in the Institute of Linguistics, Faculty of Philosophy, University of Zagreb. At the beginning the identification of TEs between single words is being accomplished by generating all possible word pairs with first word in pair from source language and second word in pair from target language. Only sentences with 1:1 alignment were included in processing. The statistical measure of Mutual Information was applied to generated pairs of words and it gave us the statictically relevant cooccurences. Pairs with high MI value are considered good TE candidates. In the second part of paper the identification of multi-word units (in this case only MWUs with 2 elements) has been achieved by applying the same statistical measure in both, source (Croatian) and target (English) language. The MI value has been applied on pairs of pairs of words giving the possible candidates of translational patterns. By high MI values it has been detected that there were pairs of words in source language, which were regularly translated with fixed pair of words in target language although the MI values for monolingual pairs in each language were extremely low. The contribution aims to show how the usage of statistical methods in parallel corpora processing can facilitate the detection of collocations (possible multi-word terms) and their TEs. At the same time the correspondent co-textual examples of word-usage is being provided in both, source and target language. This is of relevance for multilingual lexicographers as dictionary-writers and translators as the most important group of dictionary-users.

Ključne riječi: hrvatsko-engleski paralelni korpus, višerječne jedinice, prijevodni ekvivalenti, sravnjivanje riječi, uzajamna obavijesnost

Key words: Croatian-English parallel corpus, multi-word units, translational equivalents, word alignment, mutual information