

# Slovensko-hrvatski paralelni korpus

Vesna Požgaj Hadži\*, Marko Tadić\*\*

\*Odsjek za slavenske jezike i književnosti, Filozofski fakultet, Sveučilište u Ljubljani  
Aškerčeva 2, 1000 Ljubljana, Slovenija  
[vesna.pozgaj-hadzi@ff.uni-lj.si](mailto:vesna.pozgaj-hadzi@ff.uni-lj.si)

\*\*Odsjek za lingvistiku, Filozofski fakultet, Sveučilište u Zagrebu  
Ivana Lučića 3, 10000 Zagreb, Hrvatska  
[marko.tadic@ffzg.hr](mailto:marko.tadic@ffzg.hr)

## Abstract

The Slovene-Croatian Parallel Corpus is a project that was launched together with about twenty others within the framework of bilateral academic and research co-operation between Slovenia and Croatia. The corpus will contain one million words (500,000 for each language) and will comprise texts of various functional styles – both Croatian originals and their Slovene translations and vice versa. We are choosing texts that were created after 1990 in order to represent the languages to be contrasted in their present state. The need for the compilation of this type of corpus was initiated by various fields simultaneously: the teaching of both languages at university level, translation studies, lexicographic research, contrastive studies, and also by the translators of these languages themselves. Particularly the latter as well as other language professionals will be able to use the parallel corpus as an additional source of information, even more valuable in view of the fact that at present we do not have a proper Croatian-Slovene (Slovene-Croatian) dictionary. This contribution is trying to show preliminary statistical results of text alignment illustrated on the test sample which includes eight bilateral agreements between both states.

## 1. Uvod

*Slovensko-hrvatski paralelni korpus* (Ministarstvo znanosti i tehnologije R. Slovenije: J6-7802-0581-99 i Ministarstvo znanosti i tehnologije R. Hrvatske: 130821) jedan je od projekata s područja humanističkih znanosti u okviru Sporazuma o bilateralnoj slovensko-hrvatskoj suradnji u području znanosti i tehnologije za 1999., 2000. i 2001. godinu. "Težinu" projektu *Slovensko-hrvatski paralelni korpus* daje činjenica da je jedan od 28 prihvaćenih projekata (od ukupno 70 prijavljenih) koji je odobrio zajednički slovensko-hrvatski odbor krajem lipnja 1999. u trajanju od 2 godine. Riječ je zapravo o drugom dvojezičnom korpusu za slovenski jezik (prvi je Slovensko-engleski korpus *Elan*, usp. Erjavec, 1999c). Rad na projektu započeo je početkom šk. god. 1999/2000. a završetak se očekuje s polovicom 2001.

Kao partneri projekta pojavljuju se Filozofski fakulteti u Ljubljani i Zagrebu. Projekt vode V. Požgaj Hadži i M. Tadić; u njemu s hrvatske strane sudjeluju I. Pranjaković, V. Muhvić-Dimanovski, B. Bekavac, S. Fulgosi, K. Šojat; na slovenskoj strani na projektu rade V. Gorjanc, A. Skubic i Š. Vintar.

## 2. Cilj i svrha projekta

*Slovensko-hrvatski paralelni korpus* bit će ishodištem za suvremena kontrastivna istraživanja dvaju genetski srodnih i susjednih jezika koji su uspostavljanjem novih država doživjeli određene preinake, između ostaloga i zbog promjene državnoga statusa. Ciljevi su projekta ovi:

- sastaviti usporedni korpus (parallel) slovenskih i hrvatskih originala te odgovarajućih prijevoda (obostrani prijevodi),
- korpus sravniti (*align*) na razini rečeničnih prijevodnih ekvivalenata,

- omogućiti pristup korpusu ponajprije na Internetu putem WWW servisa.

Istraživanja korpusne lingvistike, koja posljednjih desetljeća nadmašuju ostale lingvističke discipline, osobito u području leksikografije (jednojezičnih, višejezičnih i paralelnih korpusa kojima pripada i naš projekt) temelj su za niz proučavanja. Rezultati našega projekta našli bi svoju primjenu u:

- a) kontrastivnim proučavanjima (strukturnim i tipološkim) slovenskoga i hrvatskoga jezika na svim jezičnim razinama.
- b) leksikografskim proučavanjima
  - zbog zastarjelih slovensko-hrvatskih (srpskohrvatskih) rječnika (Jurančić 1986, 1989)<sup>1</sup> projekt je zapravo temelj i poticaj za dvojezičnu leksikografiju. Slovensko-hrvatski rječnik i obratno danas postaje nužnim normativnim priručnikom, čega su svjesni i mnogi izdavači.
  - slovensko-hrvatski paralelni korpus također je temelj za različita leksikografska i leksikološka proučavanja, npr. proučavanja "lažnih" prijatelja (rječnik slovensko-hrvatskih homonima), proučavanja terminologije (terminološki rječnici), proučavanja kolokacija itd.
- c) znanosti o prevođenju
  - izrađeni korpus, objavljen na Internetu, bit će dragocjen priručnik studentima za vježbe iz prevođenja (slovenski-hrvatski i obratno) na Filozofskim fakultetima u Ljubljani i Zagrebu, naročito zbog aktualnosti tekstova koje će korpus sadržavati. Naime, korpus obuhvaća tekstove

<sup>1</sup> O nedostacima Jurančićevih rječnika v. Požgaj Hadži (1998).



tehničku dokumentaciju i priručnike (promet, kulinarstvo, farmaceutika, elektrotehnika...) te turističke brošure.

### 3.2. Obrada

Nad dijelom sakupljene građe probno je obavljena cijela obrada, koja se sastoji od konverzije teksta u jedinstven XML zapis, priređivanja XML dokumenata za sravnjivanje i konačno sam postupak sravnjivanja. Konverzija teksta iz polaznih oblika zapisa (najčešće MS Word 97) obavljena je programom 2XML,<sup>3</sup> koji je razvijen u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu u sklopu rada na Hrvatskome nacionalnom korpusu. Smjer je konverzije najčešće bio iz RTF u XML.

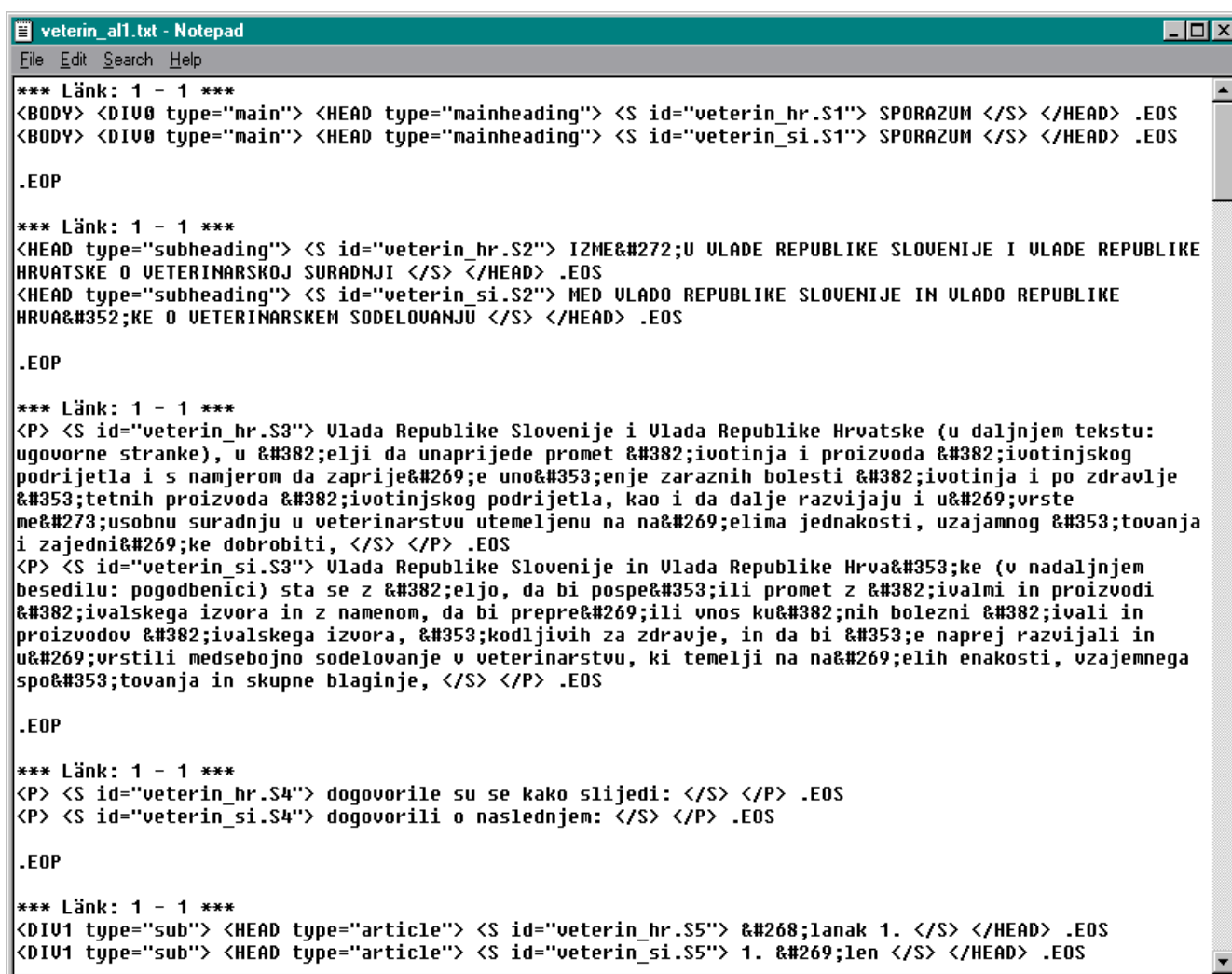
Nakon konverzije u XML sama se obrada dokumenta sastojala od segmentacije na rečenice <s> uz dodavanje ID oznaka rečenicama te provjere istovjetnosti broja odlomaka <p> kao pripreme za sravnjivanje.

Za sravnjivanje je korištena DOS inačica programa Vanilla Aligner (Danielsson & Ridings, 1997), koja se pokazala posve dostatnom za potrebe manjega paralelnog korpusa kao što je ovaj slovensko-hrvatski. Kako sam program i nema dovoljno fleksibilan ulaz, XML dokumenti morali su mu se prilagoditi. U tom su procesu međutim zadržane XML oznake koje su i same ušle u postupak sravnjivanja. Time je program dobio više redundantnih podataka i mogao je obaviti sravnjivanje na kvalitetniji način.

Probni uzorak sačinjava osam međudržavnih sporazuma i ugovora između Republike Slovenije i Republike Hrvatske.<sup>4</sup> Sami se XML dokumenti mogu vidjeti na slici 1.

### 3.4. Rezultati

Preliminarna je statistika probnoga uzorka po XML elementima iznesena u Tablici 2.



```
*** Länk: 1 - 1 ***
<BODY> <DIIV0 type="main"> <HEAD type="mainheading"> <S id="veterin_hr.S1"> SPORAZUM </S> </HEAD> .EOS
<BODY> <DIIV0 type="main"> <HEAD type="mainheading"> <S id="veterin_si.S1"> SPORAZUM </S> </HEAD> .EOS
.EOP

*** Länk: 1 - 1 ***
<HEAD type="subheading"> <S id="veterin_hr.S2"> IZME&#272;U ULADE REPUBLIKE SLOVENIJE I ULADE REPUBLIKE
HRUATSKE O VETERINARSKOJ SURADNJI </S> </HEAD> .EOS
<HEAD type="subheading"> <S id="veterin_si.S2"> MED ULADO REPUBLIKE SLOVENIJE IN ULADO REPUBLIKE
HRUA&#352;KE O VETERINARSKEM SODELOVANJU </S> </HEAD> .EOS
.EOP

*** Länk: 1 - 1 ***
<P> <S id="veterin_hr.S3"> Vlada Republike Slovenije i Vlada Republike Hrvatske (u daljnjem tekstu:
ugovorne stranke), u &#382;elji da unaprijede promet &#382;ivotinja i proizvoda &#382;ivotinjskog
podrijetla i s namjerom da zaprije&#269;e uno&#353;enje zaraznih bolesti &#382;ivotinja i po zdravlje
&#353;tetnih proizvoda &#382;ivotinjskog podrijetla, kao i da dalje razvijaju i u&#269;vrste
me&#273;usobnu suradnju u veterinarstvu utemeljenu na na&#269;elima jednakosti, uzajamnog &#353;tovanja
i zajedni&#269;ke dobrobiti, </S> </P> .EOS
<P> <S id="veterin_si.S3"> Vlada Republike Slovenije in Vlada Republike Hrva&#353;ke (v nadaljnjem
besedilu: pogodbeni&#277;i) sta se z &#382;eljo, da bi pospe&#353;ili promet z &#382;ivalmi in proizvodi
&#382;ivalskega izvora in z namenom, da bi prepre&#269;ili vnos ku&#382;nih bolezn&#382;ivali in
proizvodov &#382;ivalskega izvora, &#353;kodljivih za zdravje, in da bi &#353;e naprej razvijali in
u&#269;vrstili medsebojno sodelovanje v veterinarstvu, ki temelji na na&#269;elih enakosti, uzajemnega
spo&#353;tovanja in skupne blaginje, </S> </P> .EOS
.EOP

*** Länk: 1 - 1 ***
<P> <S id="veterin_hr.S4"> dogovorile su se kako slijedi: </S> </P> .EOS
<P> <S id="veterin_si.S4"> dogovorili o naslednjem: </S> </P> .EOS
.EOP

*** Länk: 1 - 1 ***
<DIIV1 type="sub"> <HEAD type="article"> <S id="veterin_hr.S5"> &#268;lanak 1. </S> </HEAD> .EOS
<DIIV1 type="sub"> <HEAD type="article"> <S id="veterin_si.S5"> 1. &#269;len </S> </HEAD> .EOS
```

Slika 2: Sravnjivanje Vanilla alignerom

### 3.3. Sravnjivanje

<sup>3</sup> Sam se program i njegovo funkcioniranje dijelom prikazuje u Tadić (2000).

<sup>4</sup> Sporazumi su bili ovi: o znanstvenoj i tehnološkoj suradnji, o veterinarskoj suradnji, o socijalnom osiguranju, o uređivanju vodnogospodarskih odnosa, o zaštiti od prirodnih i civilizacijskih katastrofa, Suradnja na području obrazovanja za

	HR	SI
odlomaka <p>	522	522
rečenica <s>	891	891
pojavnica <w>	13.549	13.307

Tablica 2: Broj <p>, <s> i <w> elemenata u probnom uzorku od 8 međudržavnih ugovora

Valja uočiti identičnost broja odlomaka i rečenica koja se pojavljuje zbog prirode pravnoga teksta. Tekstovi na oba jezika objavljeni su kao originali i usklađeni već u izvorniku. Stoga su sva strojna sravnjivanja u ovom probnom uzorku oblika 1-1. Gotovo identično ponašanje može se naći u drugim paralelnim korpusima pravnih tekstova (v. Gamper, 2000). Dapače, kad se god pri pregledu sravnjivanja naišlo na 2-1 ili 1-2 sravnjivanje, to je redovito bio znak greške u ulaznim podacima, tj. greške u modulu za segmentaciju rečenica.

Rezultat sravnjivanja Vanillom mogu se uočiti na slici 2 odnosno, slici 3 u sažetijem zapisu.

```

-- <body>
<link xtargets="veterin_hr.S1 veterin_si.S1" />
<link xtargets="veterin_hr.S2 veterin_si.S2" />
<link xtargets="veterin_hr.S3 veterin_si.S3" />
<link xtargets="veterin_hr.S4 veterin_si.S4" />
<link xtargets="veterin_hr.S5 veterin_si.S5" />
<link xtargets="veterin_hr.S6 veterin_si.S6" />
<link xtargets="veterin_hr.S7 veterin_si.S7" />
<link xtargets="veterin_hr.S8 veterin_si.S8" />
<link xtargets="veterin_hr.S9 veterin_si.S9" />
<link xtargets="veterin_hr.S10 veterin_si.S10" />
<link xtargets="veterin_hr.S11 veterin_si.S11" />
<link xtargets="veterin_hr.S12 veterin_si.S12" />
<link xtargets="veterin_hr.S13 veterin_si.S13" />
<link xtargets="veterin_hr.S14 veterin_si.S14" />
<link xtargets="veterin_hr.S15 veterin_si.S15" />
<link xtargets="veterin_hr.S16 veterin_si.S16" />
<link xtargets="veterin_hr.S17 veterin_si.S17" />
<link xtargets="veterin_hr.S18 veterin_si.S18" />
<link xtargets="veterin_hr.S19 veterin_si.S19" />
<link xtargets="veterin_hr.S20 veterin_si.S20" />
<link xtargets="veterin_hr.S21 veterin_si.S21" />
<link xtargets="veterin_hr.S22 veterin_si.S22" />
<link xtargets="veterin_hr.S23 veterin_si.S23" />
<link xtargets="veterin_hr.S24 veterin_si.S24" />
<link xtargets="veterin_hr.S25 veterin_si.S25" />
<link xtargets="veterin_hr.S26 veterin_si.S26" />
<link xtargets="veterin_hr.S27 veterin_si.S27" />
<link xtargets="veterin_hr.S28 veterin_si.S28" />
<link xtargets="veterin_hr.S29 veterin_si.S29" />
<link xtargets="veterin_hr.S30 veterin_si.S30" />
<link xtargets="veterin_hr.S31 veterin_si.S31" />
<link xtargets="veterin_hr.S32 veterin_si.S32" />
<link xtargets="veterin_hr.S33 veterin_si.S33" />
<link xtargets="veterin_hr.S34 veterin_si.S34" />
<link xtargets="veterin_hr.S35 veterin_si.S35" />
<link xtargets="veterin_hr.S36 veterin_si.S36" />
<link xtargets="veterin_hr.S37 veterin_si.S37" />
<link xtargets="veterin_hr.S38 veterin_si.S38" />
</body>

```

Slika 3: Sažet zapis sravnjivanja u kojem se navode samo vrijednosti id-atributa za <S>-ove

## 4. Zaključak

U radu je prikazan početak rada na projektu *Slovensko-hrvatski paralelni korpus* koji je odobrilo i financijski podržalo slovensko i hrvatsko Ministarstva znanosti i tehnologije. Korpus je time omogućio suradnju stručnjaka dvaju istovrsnih fakulteta. Do sada su sakupljeni tekstovi različitih funkcionalnih stilova i različitih područja (terminološka raznovrsnost). Oni zapravo predstavljaju kompromis između očekivane uporabnosti korpusa s jedne strane i same dostupnosti tekstova u digitalnom zapisu s druge strane. Naime, posljednjih desetak godina uočljivo je nepostojanje prijevoda sa slovenskog ili na slovenski i/ili hrvatski (osobito u digitalnom obliku), što je prouzročilo značajne teškoće oko prikupljanja građe za korpus. *Slovensko-hrvatski paralelni korpus* omogućit će niz različitih jezikoslovnih istraživanja: kontrastivnih, leksikografskih, didaktičko/metodičkih, a posebice znanost o prevođenju te razvitak jezičnih tehnologija za oba jezika. Potencijalni su korisnici korpusa osim istraživača i studenti zagrebačke slovenistike i ljubljanske kroatistike te prevoditelji kojima će korpus biti dostupan putem WWW-a kao dopunski izvor informacija, osobito u situaciji kada dvojezičnih hrvatsko-slovenskih (i obratno) rječnika zapravo nema.

## 5. Literatura

- Ahrenberg, Lars; Merkel, Magnus; Ridings, Daniel; Sígvald Hein, Anna & Tiedemann, Jörg, (1999). Automatic processing of parallel corpora: A swedish perspective. (<http://numerus.ling.uu.se/~corpora/plugin/>)
- Danielsson, Pernilla & Ridings, Daniel. 1997, Practical presentation of a "vanilla" aligner, Presented at the TELRI Workshop on Alignment and Exploitation of Texts. Institute Jožef Stefan, Ljubljana (<http://svenska.gu.se/PEDANT/workshop/workshop.html>).
- Erjavec, Tomaž, 1999a. Making the ELAN Slovene/English Corpus. In Špela Vintar (Ed.), Proceedings of the workshop Language technologies — Multilingual Aspects, (pp. 23–30). Ljubljana: Department of Translation and Interpreting, Faculty of Arts, University of Ljubljana.
- Erjavec, Tomaž, 1999b. A TEI encoding of aligned corpora as translation memories. In Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99), Bergen: ACL.
- Erjavec, Tomaž, 1999c. Slovensko-angleški korpus *Elan*. *Slavistična revija*, 47, 4:515-522. (<http://nl.ijs.si/elan>)
- Gamper, Johann, 2000. "A Parallel Corpus of Italian/German Legal Texts" u: LREC2000 Proceedings, Athens 2000-05-29–2000-06-02, Pariz-Atena 2000, 531-538.
- Jurančič, Janko, 1986. Srbskohrvatsko-slovenski slovar, 3. izdaja, DZS, Ljubljana.
- Jurančič, Janko, 1989. Slovensko-srbskohrvaški slovar, 2. popravljena izdaja, DZS–Školska knjiga, Ljubljana–Zagreb.
- Požgaj Hadži, Vesna, 1998. Razumijevanje znanstvenog diskursa u studenata slovenistike. *Filologija*. knj. 30-31, 509-518.
- Tadić, Marko, 2000. "Building the Croatian-English Parallel Corpus" u: LREC2000 Proceedings, Athens 2000-05-29–2000-06-02, Pariz-Atena 2000, str. 523-530.

Tiedemann, Jörg, 1998. Parallel corpora in Linköping, Uppsala and Göteborg (PLUG). Work package 1. Department of Linguistics, Uppsala University. (<http://numerus.ling.uu.se/~corpora/plug/>)

## **6. Zahvala**

Zahvaljujemo *Službenom listu Republike Slovenije* za ustupljene digitalne oblike međdržavnih ugovora između Republike Slovenije i Republike Hrvatske.