Načelo večjezičnosti ali večjezični korpus iz manjše množice dvojezičnih

Jasna Belc, Miran Željko

Vlada Republike Slovenije
Generalni sekretariat Vlade Republike Slovenije
Služba za prevajanje, tolmačenje, redakcijo in terminologijo
Gregorčičeva 20, 1000 LJUBLJANA, Slovenija
Tel.: +386 1 478 25 44; faks: +386 1 478 15 62
e-naslov: jasna.belc@gov.si, miran.zeljko@gov.si

Povzetek

Članek prikazuje zamisel, kako nastane večjezični korpus na podlagi več dvojezičnih in njegovo uresničitev. Prototipni model je zamišljen na 3 korpusih s tremi različnimi jeziki, ki vsebujejo kot eno od sestavin slovenščino. Uresničitev zamisli kot zastavljeni projekt je na dosegu nekaj pretvorbnih programov in uporabi arhitekture terminološke zbirke kot lupine, ki daje zavetje večjezični korpusni zbirki (4-jezični), njen spletni prikaz pa se uresničuje z uporabo konkordančnega programa, ki omogoča iskanje poljubnih delcev segmentne celote (za preverjanje prevodnih ustreznic) znotraj 4 mogočih jezikov. Projekt je prav tako zanimiv zaradi priprav Slovenije na predsedovanje EU v drugi polovici leta 2008.

The Multilinguality Principle or a Multilingual Corpus Derived from Several Bilingual Corpora

The article presents the idea how we can build a multilingual corpus from the various bilingual corpora and its realisation. The conception of the prototype model is carried out from three bilingual corpora containing three different languages where the other component is Slovene. The idea has been carried out as a project which is attainable within the scope of some transformational programmes and the use of architecture of the terminological database as a shell for hosting the new multilingual corpus collection (4-lingual), its web representation is realised with the concordancer that allows for the searching of any part of the whole segments (to find the translation matches) within 4 available languages. The project is interesting also within the scope of the preparations of Slovenia to the EU presidency in the second half of 2008.

1. Uvod

Cilj tega projekta so jezikovne tehnologije, ki so usmerjene v izdelavo produktov uporabne vrednosti, ki so uporabni pri iskanju večjezičnih informacij bodisi zaradi običajnega prevajanja, iskanja strokovnega izrazja v več jezikih hkrati, kot podlaga za strojno prevajanje in za raznovrstne raziskave, ki lahko temeljijo ali pa se dopolnjujejo s preverjanjem na več jezikih ipd.

Glavni cilj projekta, ki že poteka, obsega:

– pripravo več dvojezičnih korpusov,
– pripravo omejenega štirijezičnega korpusa, nastalega iz 3 dvojezičnih.

Dosedanji produkti:

– terminološka zbirka s temi poglavitnimi lastnostmi: urejena in strukturirana po vsebini in obliki (metajezikovno označevanje), uporabna in splošno dostopna na spletnih straneh;
– korpusna zbirka – doslej dvojezična, v pripravi za objavo pa še več dvojezičnih zbirk, v katerih je eden od jezikov slovenščina, drugi pa (za zdaj) trije pomembni evropski in svetovni jeziki.

2. Uresničitev projekta

Uporaba obstoječih orodij nemške znamke Trados nam omogoča naslednje osnovne operacije:

– poravnavanje besedil v dvojezični različici z orodjem WinAlign;
– prevajanje in urejanje prevodnih zbirk v pomnilniku prevodov z orodjem Translator's Workbench;
– urejanje terminološke ali kakšne druge zbirke v okviru ali arhitekturi, ki služi kot ogrodje ali lupina za vnos, shranjevanje različno strukturiranih večjezičnih ali večaspektualnih podatkov glede na zamisel sestavljavca take podatkovne zbirke, hkrati tudi shranjevanje besedilnih (črkovno-številčnih podatkov), grafičnih ali kakšnih drugih podatkov – z orodjem MultiTerm.

Faze, potrebne za uresničitev projekta:

1) Poravnave (vzporeditve) dvojezičnih besedil, od katerih je eden od jezikov slovenščina, omogočijo nastanek dvojezičnega pomnilnika, v katerega se po zamišljeni zgradbeni predlogi uvozijo rezultati dobljenih poravnav. Poravnave z orodjem WinAlign lahko sicer po vnosu ustreznih preverjenih besedil, ki si medsebojno ustrezajo glede na vmesno dejavnost prevajanja med takima besediloma (ki je običajno enosmerna, npr. iz slovenščine v francoščino ali iz angleščine v slovenščino), potekajo samodejno, vendar je zaradi standardnih algoritmov takih poravnalnih programov bolje s sodelovanjem človeka omogočiti preverjeno poravnavo, ki zagotavlja kakovost poravnanih besedil, tj., da si po dva in dva segmenta iz različnih jezikov dejansko ustrezata po prevodu. Deloma to zagotavlja sama segmentacija besedil v vsakem posameznem jeziku, ki je vključena v program poravnave *(alignment programme),* vendar pa sredstva, na katera se opira sama poravnava, po jezikih niso enako razporejena. Gre za neke vrste pravopisna interpunkcijska znamenja (ločila), ki pa jih različni jeziki različno uporabljajo. Poleg ustaljenih jezikovnih interpunkcijskih znamenj v elektronskih besedilih najdemo še druga znamenja (npr. presledek, konec

vrstice, konec odstavka, alinejni zamik ali tabulator ipd.), ta protistavimo v dveh različnih jezikovnih besedilih, ki tako razpadeta na segmente. Nastali segmenti niso vedno enako dolgi, kakor tudi ne velja vedno – čeprav je pri določeni vrsti strokovnih besedil zaželeno prevajanje z upoštevanjem enakega zaključka bistvenih besedilnih delov, kot so npr. stavki, ki se končujejo običajno s piko (.), alineje, ki se zaključujejo običajno s prehodom v novo vrstico (¶ ipd.), včasih tudi z vejico (ki pa tu ni pomembna), uvajalni stavki, ki se končujejo z dvopičjem (:), členitev besedila na širše ali ožje stavčne ali besednozvezne enote, ki se končujejo pogosto s podpičjem (;) – da bi si dvojezični segmenti ustrezali homomorfno. Poravnalni algoritmi lahko povezujejo kvečjemu dva segmenta z nasprotnim edinim segmentom (v drugem jeziku) ali nasprotno, razporeditev načina vzporeditve pa je bolj ali manj prepuščena preračunavanju s statističnimi metodami v samem programu. Program poravnave omogoča uporabniku le nekaj izbir, npr. izbiro poravnave po odstavkih kot segmentnih delih ali pa izbiro »stavčnih segmentov«, opisanih ob rabi segmentacijsko-interpunkcijskih sredstev.

2) »Pretvorba« iz poravnav v pomnilnik prevodov: po izbranih oblikovnih in vsebinskih nastavitvah v 'praznem' pomnilniku prevodov orodja Translator's Workbench (Database Setup in Project Settings) v pomnilnik uvozimo vse želene poravnave, ki smo jih uresničili s poravnalnikom dvojezičnih besedil. Pomnilnike ločimo glede na prevodni jezikovni par (npr. angleško-slovenski, slovensko-francoski ali nemško-slovenski ipd.), dobimo vsaj tri različne pomnilnike prevodov, ki jih pozneje pretvorimo v korpus zaradi objave in javnega dostopa na spletu.

3) »Pretvorba iz pomnilnika v korpus: ne glede na prejšnjo smer prevajanja iz pomnilnika prevodov izvozimo njegovo vsebino in jo z ustreznim programom za predstavitev na spletu (pretvorba v html zapis ipd.) prikažemo na spletnih straneh. Take pretvorbe so bile prikazane že v prejšnjih objavah M. Željka ob nastanku prvega korpusa besedil, zajetih pri programu prevajanja zakonodaje EU iz angleščine v slovenščino (glej vir 7). Ob postavitvi več dvojezičnih korpusov pa nastane vprašanje, ali jih lahko med seboj povežemo. Na voljo imamo namreč po en skupen sestavnik vseh dvojezičnih korpusov, tj. slovenski del prevodov oz polovico vsakega jezikovnega para. Kaže pa, da pri obstoječih pomnilnikih prevodov ali dodanih dvojezičnih korpusih taka pretvorba ni mogoča, ker so segmenti zelo različni in samo označevanje z vsebinskimi atributi brez oznake zaporednega segmenta znotraj nekega besedila oz. v samem korpusu, pomnilniku ali dvojezični zbirki ni mogoče. S pretvorbo v »večjezično zbirko« pa bi dobili želeni »večjezični korpus«.

4) Kaj storiti? Ker so si izvozne in uvozne datoteke iz dveh (če ne celo vseh treh) orodij Tradosovega paketa za prevajanje in urejanje terminologije zelo podobne (vsebujejo namreč enote, ki so lahko sestavni del kake standardno kodirane xml datoteke), jih lahko združimo oz. pretvorimo eno v drugo, le da dobimo iz WinAlignovega poravnalnika in Translator's Workbencha samo dvojezične prevodne enote. Tu nam na pomoč priskoči kot ogrodje ali lupina, namenjena strukturiranemu shranjevanju podatkov, Tradosova aplikacija Multiterm. Potrebno je še nekaj pretvorb, preverjanja rezultatov, prilagajanja lastnim potrebam in zamislim in rezultati prototipne večjezične prevodne zbirke oz. korpusa so lahko pred nami.

Podobnosti med izvozom iz obeh Tradosovih orodij si lahko ogledamo na kratkem izseku segmentov, ki jih omogočajo vpisi v »terminološko zbirko« Multiterm ali prevodne enote iz Translator's Workbencha. Za vnos v štirijezični korpus prevodov (v lupini Multiterma) je treba pripraviti ustrezen program pretvorb med obema zapisoma v izvoznih datotekah (xml ali sorodnih vrst) s preambulo in ustreznim zaključkom.

Segmenti iz Multiterma:

```
**
<Creation Date>25.07.2001 - 18:04:00
<Created By>super
<Change Date>25.07.2001 - 18:04:00
<Changed By>super
<Entry Class>1
<Graphic>
<Entry Number>20251
<Subject>informatics AUL
<Subj>informatika
<SourceDoc&Lang>Sklep Sveta 92/242/EGS
<EN>standardization activities
<SL>dejavnosti standardiziranja
<Reliability>4
<DE>Normungstätigkeiten
<FR>activités de normalisation
<TermRef>uvod priloge
**
<Creation Date>25.07.2001 - 18:04:00
<Created By>super
<Change Date>25.07.2001 - 18:04:00
<Changed By>super
<Entry Class>1
<Graphic>
<Entry Number>20244
<Subject>informatics AUL
<Subj>informatika
<SourceDoc&Lang>Sklep Sveta 92/242/EGS
<EN>security of information systems
<SL>varnost informacijskih sistemov
<Reliability>4
<DE>Sicherheit von Informationssystemen
<FR>sécurité des systmes d'information
<TermRef>preambula
**
<Creation Date>25.07.2001 - 18:04:00
<Created By>super
<Change Date>25.07.2001 - 18:04:00
<Changed By>super
<Entry Class>1
<Graphic>
<Entry Number>20245
<Subject>informatics AUL
<Subj>informatika
<SourceDoc&Lang>Sklep Sveta 92/242/EGS
<EN>information market
<SL>informacijski trg
<Reliability>4
<DE>Informationsmarkt
<FR>marché de l'information
<TermRef>preambula
**
```

<Creation Date>16.07.1998 - 22:39:50
<Created By>super
<Change Date>10.09.2004 - 08:25:43
<Changed By>super
<Entry Class>1
<Graphic>
<Entry Number>2157
<Subj>informatika
<Subject>informatics AUL
<EN>data bank
<SL>banka podatkov
<FR>banque des données
<DE>Databank
<TermRef>Evropski sporazum, Ur. l. 44, 1997
**

Segmenti iz Translator's Workbencha:
**
<TrU>
<CrD>18052004, 16:58:32
<CrU>SVEZ
<Att L=Stanje>Pravna redakcija
<Att L=Področje>Znanost in kultura
<Txt L=Oznaka>32002D2367
<Seg L=EN-GB>Major projects
<Seg L=SL>Glavni projekti
</TrU>
<TrU>
<CrD>18052004, 16:58:32
<CrU>SVEZ
<Att L=Stanje>Pravna redakcija
<Att L=Področje>Znanost in kultura
<Txt L=Oznaka>32002D2367
<Seg L=EN-GB>These will be subject to a formal project management approach, as follows:
<Seg L=SL>Uradni pristop projektnega upravljanja bo veljal za naslednja področja:
</TrU>
<TrU>
<CrD>18052004, 16:58:32
<CrU>SVEZ
<Att L=Stanje>Pravna redakcija
<Att L=Področje>Znanost in kultura
<Txt L=Oznaka>32002D2367
<Seg L=EN-GB>Different instruments of cooperation between national statistical organisations and Eurostat will be put in place.
<Seg L=SL>Vzpostavljeni bodo različni instrumenti sodelovanja med nacionalnimi statističnimi organizacijami in Eurostatom.
</TrU>
<TrU>
<CrD>18052004, 16:58:33
<CrU>SVEZ
<Att L=Stanje>Pravna redakcija
<Att L=Področje>Znanost in kultura
<Txt L=Oznaka>32002D2367
<Seg L=EN-GB>Quality assurance and the scientific basis of Community statistics will be the result of close cooperation between official and academic statisticians.
<Seg L=SL>Zagotavljanje kakovosti in znanstvena podlaga statistike Skupnosti bosta posledica tesnega sodelovanja med uradnimi in akademskimi statističnimi krogi.
</TrU>
<TrU>
<CrD>18052004, 16:58:34
<CrU>SVEZ
<Att L=Stanje>Pravna redakcija
<Att L=Področje>Znanost in kultura
<Txt L=Oznaka>32002D2367
<Seg L=EN-GB>Specific projects
<Seg L=SL>Specifični projekti
</TrU>
**

Slika 1

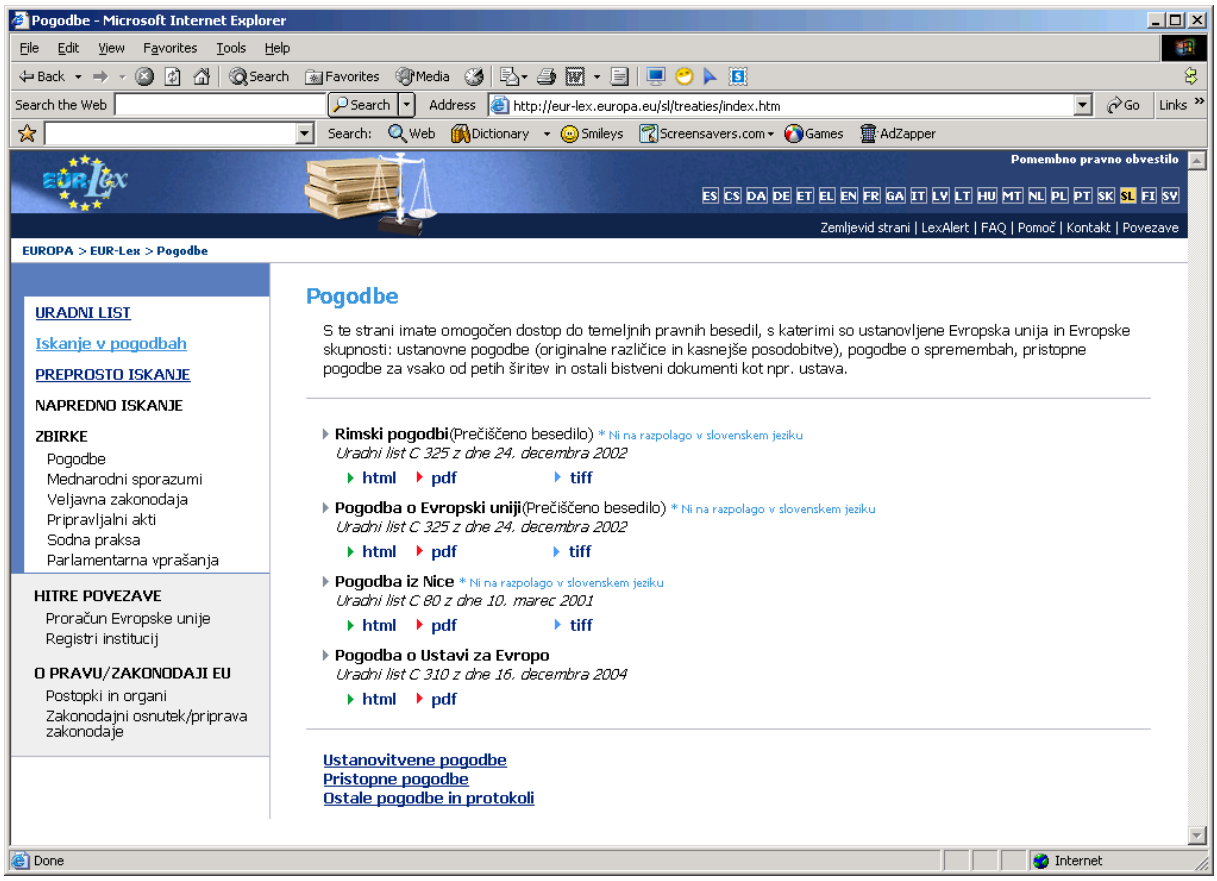
Viri za večjezični korpus:

Med slovenskimi viri, ki so primerni za vzporeditev z drugimi jeziki, so zlasti ustanovitvene in pristopne pogodbe Evropskih skupnosti, med temi tudi že ratificirane pogodbe o dveh novih članicah, ki bosta predvidoma vstopili v Evropsko unijo leta 2007, ter še neratificirana Pogodba o Ustavi za Evropo. Prevod historičnih različic teh pogodb je bil ena od obvez Slovenije za njeno polnopravno članstvo v Evropski uniji leta 2004.

Uresničitev projekta kot takega je povezana z naslednjimi koraki in predvidenimi rešitvami morebitnih težav:

a) prenos iz enega xml zapisa (izvoz dvojezičnega korpusa oz. prevodne zbirke iz Translator's Workbencha) v drug xml zapis zaradi že vgrajenih funkcij v Multitermu ne bi smel povzročati težav: dvojezične zapise iz dvojezičnih korpusov se tako lahko zaporedno prenese kot xml zapise v lupino Multiterma, ki mu vnaprej določimo izbrane atribute, npr. želene jezike in druge podatke, atrubute, ki jih vsebuje tudi dvojezična prevodna zbirka ali korpus (Področje, Stanje, Oznaka – ki se nanašajo na posamezen dokument in kot so zapisani
in poenoteni v vseh pomnilnikih prevodov);

b) zaporedni prenos (uvoz) xml zapisov dvojezičnih prevodnih zbirk v Multitermovo lupino utegne le podvojiti ali potrojiti vnos slovenskega segmenta v Multitermovo podatkovno bazo, kar se da rešiti s preprostim algoritmom iskanja in brisanja 'sinonimnih terminov', v našem primeru homografnih segmentov (stavkov ali besednih zvez). Ta poseg v samem Multitermu – glede na številne funkcije, ki omogočajo redakcijo, zlivanje enot ali zlivanje in brisanje odvečnih ali enakopisnih segmentov v programu Multiterm - ali pa ob ponovnem izvozu dobljene vsebine iz Multiterma ob ustvarjanju nove večjezične zbirke ni posebno zapleten, predvsem zaradi številnih možnosti, ki jih nudi sam Multiterm ali pa homogenost zapisa v formatu xml. Celoten potek uskladitve (poenotenja in normalizacije) novonastale večjezične zbirke zato ne predstavlja velikih težav, le preudaren razmislek in načrtovanje zaporednih korakov urejanja same zbirke, ki se jo pripravi za nov izvoz podatkovne baze v format xml, ki je podlaga vsebinskega vira podatkov, po katerih se 'sprehaja' konkordančni program med iskanjem želenih iskanih nizov (zadetkov) iz enega od jezikov v večjezičnem korpusu, postavljenem na splet.



Slika 2

3. Uporabna vrednost projekta

Prednosti takega večjezičnega »korpusa«: poravnave so narejene po »stavčnih segmentih« in ne po odstavkih, kar omogoča večjo verjetnost zadetkov v pomnilniku prevodov in tudi v korpusu. Poznavanje najpomembnejših dokumentov zakonodaje EU je bistvenega pomena za vse državljane držav članic Evropske unije. Prikaz takih segmentov v okolju različnih tujih spremnih jezikov omogoča primerjavo slovenščine s še tremi tujimi jeziki, med katerimi lahko izberemo tistega, ki ga najbolje poznamo ali potrebujemo pri svojem delu (npr. prevajanju, , tolmačenju, lektoriranju oz. jezikovni redakciji, dejavnostih v zvezi s terminologijo, stroko oz. samimi pripravami na predsedovanje Slovenije Evropski uniji). Nadaljnja uporaba večjezičnega korpusa bo pokazala, v katero smer naj se še razvijajo dodatne aplikacije, ki jih je ponudila svojim uporabnikom Služba za prevajanje, tolmačenje, redakcijo in terminologijo v Generalnem sekretariatu Republike Slovenije (GSV).

GSV s svojim terminološkim in jezikovnotehnološkim delom ponuja vedno nove rešitve, hkrati pa daje svoje produkte na voljo tudi raziskovalnim ustanovam, ki z dodatnimi funkcijami (označevanja: besednovrstnega *(Part of Speech)*, skladenjskega (besednozveznega, vsaj z določanjem jeder in ujemanj) ali pomenoslovnega (dodajanje ontologij ali pomenoslovnih abstraktnih kategorij v smislu raziskovalnega dela J. Pustejovskega in drugih jezikoslovcev), tipiziranih pomenoslovnih kategorij, ki jih uporabljajo zlasti kategorialne slovnice itd.) ob skupnem sodelovanju omogočajo nastanek vedno novih produktov, namenjenih zlasti prevajalcem in jezikoslovcem v neposredno uporabo ali tudi za nadaljnje jezikoslovne, jezikovnotehnološke ali računalniške raziskave.

In še okvirni številčni izračun:

Najpomembnejši dokumenti, ki naj bi sestavljali prototipni projekt večjezičnega korpusa s stavčno poravnanimi segmenti (ustanovitvene in pristopne pogodbe, druge pogodbe in protokoli, glej sliko 2), štejejo skupaj okrog 20 000 strani uradneg lista EU. Povprečno dobimo iz ene strani vsaj 10 prevodnih enot, ena prevodna enota pa šteje približno 30 besed. Tako je mogoče v povprečju pričakovati vsaj 5 do 6 milijonov besed v celotnem osnovnem večjezičnem korpusu z osnovnimi poravnanimi dokumenti v 4 jezikih, in ta bo v jeseni na vpogled in uporabo na spletnih straneh naše službe.

4. Pomen večjezičnega korpusa za druge stroke in uporabnike

Poleg nekaj navedenih konkretnih primerov uporabe bo imel večjezični korpus in njegov spletni konkordančnik tudi pomembnejšo pravno-politično, geografsko prepoznavalno, kulturno in izobraževalno vlogo. Na pravno-političnem in geografskem področju gre za pomembno vlogo pri širjenju načel demokratičnosti uporabe uradnega jezika, ki ga lahko primerjamo z drugimi, zlasti delovnimi jeziki institucij EU, povečanje prepoznavnosti Slovenije in slovenskega

jezika, večje poznavanje Slovenije in slovenščine s pomočjo učenja jezika za potrebe prevajanja v ustanovah EU iz slovenščine v tuje jezike za tolmače in prevajalce slovenščine (med katerimi so tudi tujci), pa tudi zaradi želje po poznavanju slovenske kulture nasploh, ne le zgolj v pravnostrokovnih krogih in pravno-političnih besedilih. Omenjeni korpus je lahko tudi vzvod za širjenje tovrstnega vedenja, kar postavlja Slovenijo v sklop uresničitev prej nezavednih želja posameznikov kot sodržavljanov Evropske unije.

Viri

1) Hans van Halteren (1999): Syntactic Wordclass Tagging, Kluwer Academic Publishers, Dordrecht, Boston, London.
2) Laurent Romary (2000): TMF – Terminological Markup Framework, Laboratoire LORIA, (CNRS, INRIA, Univerza v Nancyju, ISO meeting, London, 2000)
http://www.loria.fr/projets/TMF/DOC/SLIDES/TMF-ISO_pres.ppt
3) Nancy Ide, Laurent Romary: A Common Framework for Syntactic Annotation
http://acl.ldc.upenn.edu/P/P01/P01-1040.pdf
4) Petr Sgall, Jarmila Panevová, Eva Hajičová (2004): Deep Syntactic Annotation: Tectogrammatical Representation and Beyond, hlt-naacl2004, (Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting).
http://acl.ldc.upenn.edu/hlt-naacl2004/frontiers/pdf/naacl04sph.pdf
5) Ray C. Dougherty (1994): Natural Language Computing, An English Generative Grammar in Prolog, Laurence Erlbaum Associates, Publishers, Hillsdale, New Jersey, Howe, UK.
6) Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2004): Massive multilingual corpus compilation: Acquis Communautaire and totale. In Proc. of the Second Language Technology Conference. April 2004, Poznan.
7) Miran Željko (2002): Pripomočki na spletu za prevajalce zakonodaje EU. Zbornik mednarodne konference Informacijska družba 2002 – jezikovne tehnologije. Ljubljana, oktober 2002.
http://nl.ijs.si/isjt02/zbornik/sdjt02-05zeljko.pdf
8) Miran Željko, Adriana Krstič (2002): Web-based Trados Databases – an Alternative Approach. Kongres Mednarodne zveze prevajalcev. Vancouver, Kanada, avgust 2002.
9) Darja Erbič, Adriana Krstič Sedej, Jasna Belc, Nataša Zaviršek - Žorž, Nevenka Gajšek, Miran Željko) (2005): Slovenščina na spletu v dokumentih slovenske različice pravnega reda Evropske unije, terminološki zbirki in korpusu. Simpozij Obdobja 24: Razvoj slovenskega strokovnega jezika, Ljubljana, november 2005.
10) Jasna Belc (2002): Konferenci ob rob: Sodelovanje na področju terminologije in drugih sorodnih disciplin, zlasti jezikovnih tehnologij. Zbornik prispevkov s simpozija Terminologija v času globalizacije. Ljubljana, 5. in 6. junij 2003, str. 361–365.
11) Miran Željko (2003): Evroterm in Evrokorpus – terminološki slovar in korpus prevodov. Zbornik prispevkov s simpozija Terminologija v času globalizacije. Ljubljana, 5. in 6. junij 2003, str. 139–149.
12) Trados Manual for MultiTerm and Translator's Workbench, v. 7.0 (2005).
13) Tomaž Erjavec (2005): Foundational Course: Annotation of Language Resources: XML, TEI, OWL:od http://nl.ijs.si/et/teach/esslli05/esslli05-1.html
do http://nl.ijs.si/et/teach/esslli05/esslli05-5.html.
14) James Pustejovsky (2005): Type Selection and the Semantics of Local Context, Lectures at ESSLLI 2005, Edinburgh: http://www.macs.hw.ac.uk/esslli05/giveabs.php?30
15) Glyn Morrill (1994): Type Logical Grammar: Categorial Logic of Signs, Kluwer Academic, Dordrecht.