Integrisano okruženje za pripremu paralelizovanog korpusa

Ivan Obradović, Ranka Stanković, Miloš Utvić

Ivan Obradović - Ranka Stanković – Miloš Utvić (Beograd)

Integrisano okruženje za pripremu paralelizovanog korpusa

Razvoj paralelizovanih korpusa zahteva pripremu paralelnih tekstova za njihovu integraciju u paralelizovani korpus. Reč je o jednom kompleksnom zadatku koji se može rešiti na različite načine, i koji mora da se odvija u nekoliko koraka. U ovom radu najpre je iznet postupak pripreme paralelnih tekstova za paralelizovani korpus koji se koristi u Grupi za jezičke tehnologije Univerziteta u Beogradu. Potom je dat kratak pregled programa (XAlign, Concordancier, WS4LR), odnosno softverskih alata koji se pri tome koriste. Nedostatak udobnog okruženja sa grafičkim korisničkim interfejsom, koje bi u sebi objedinilo sve ove programe, motivisalo je Grupu za jezičke tehnologije da pristupi razvoju jednog integrisanog okruženja za pripremu paralelizovanog korpusa, pod imenom ACIDE. U radu su detaljno opisane osnovne karakteristike i način funkcionisanja okruženja ACIDE, razvijenog kako bi se korisnicima olakšala priprema tekstova za korpus.

1. Uvod

Paralelni tekst ili bitekst se najčešće formira od dva teksta od kojih je jedan originalan, a drugi je nastao njegovim prevođenjem. Radi se dakle o dva teksta iste sadržine, najčešće na dva različita jezika. Treba, međutim, napomenuti da paralelni tekstovi mogu postojati i na istom jeziku, recimo kada se radi o dva različita prevoda istog književnog dela. Paralelni tekstovi se mogu sastojati i od više tekstova, na više jezika, pri čemu je bitno da su svi tekstovi iste sadržine, dakle potekli od istog originalnog teksta. Paralelni tekstovi se mogu upariti, tako što će se uspostaviti veze između odgovarajućih elemenata jednog i drugog teksta. To uparivanje je moguće na različitim nivoima (paragrafa, rečenice, reči) a rezultat uparivanja je paralelizovani tekst (Gale and Church, 1993). Veće kolekcije paralelizovanih tekstova nazivaju se paralelizovanim korpusima. Podrazumeva se da se tekstovi u paralelizovanim korpusima u elektronskom obliku, odnosno da se mogu obrađivati uz pomoć računarske tehnologije. Paralelizovani korpusi se mogu koristiti u istraživanjima iz oblasti dvojezične odnosno višejezične leksikografije, za učenje stranog jezika, za prevođenje, za lingvistička istraživanja ili uporedna izučavanja dva ili više jezika, itd.

Treba napomenuti da pored korpusa paralelenih odnosno paralelizovanih tekstova, postoje i uporedni korpusi (*comparable corpora*) koje čine tekstovi koji nisu iste, ali su slične sadržine (bave se istom tematikom), pripadaju istom žanru i približno su istog obima (npr. korpus sportskih članaka iz srpskih i hrvatskih novina, ili pravna akta na srpskom i bošnjačkom jeziku) (Dimitrova et al., 1998). Uporedni korpusi mogu, ali ne moraju biti višejezični. Ova vrsta korpusa je ređa i ovde će nadalje biti reči samo o paralelnim, odnosno paralelizovanim korpusima.

U ovom radu će biti razmotreni problemi koji prate pripremu paralelnih tekstova i kreiranje paralelizovanih korpusa, kao i pristup rešavanju ovih problema razvijen u Grupi za jezičke tehnologije Univerziteta u Beogradu. Radi se, zapravo, o metodologiji i tehničkim rešenjima koja se u okviru Grupe koriste za kreiranje paralelizovanih korpusa.

2. Kreiranje paralelizovanog korpusa

U Grupi za jezičke tehnologije postupak pripreme tekstova i kreiranje paralelizovanog korpusa odvijaju se u sledećih nekoliko faza:

- priprema i segmentacija teksta na jedinice uparivanja (segmente)
- uparivanje segmenata
- vizuelizacija paralelizovanih tekstova, kontrola i korekcije uparivanja

- generisanje TMX formata paralelizovanog teksta
- razlaganje TMX formata na pojedinačne tekstove u XML formatu
- vertikalizacija pojedinačnih tekstova
- kreiranje korpusa

Prvi korak obuhvata pripremu i segmentaciju teksta, odnosno razlaganje teksta na manje jedinice (paragrafe, rečenice, reči) i obeležavanje jedinica teksta. Samo obeležavanje se vrši korišćenjem XML (eXtensible Markup Language) obeležja[1], u skladu sa preporukama TEI (Text Encoding Initiative) konzorcijuma[2]. U praksi se to uglavnom svodi na obeležavanje paragrafa i rečenica. Taj posao se može automatizovati korišćenjem konačnih transduktora, ali greške su neizbežne, te je neophodna i ručna intervencija. U tu svrhu se mogu koristiti razni editori (npr. Emacs, XML Spy, oXygen itd) koji omogućavaju ne samo uređivanje XML tekstova, već i proveru njihove dobre formiranosti, kao i validaciju u odnosu na određen DTD (Document Type Definition), odnosno shemu[3].

Drugi korak, uparivanje tekstova, koji se još naziva i paralelizacija tekstova definitivno zauzima ključno mesto u celokupnom procesu pripreme paralelizovanog korpusa. Zadatak paralelizacije tekstova se svodi na to da se utvrdi koji elementi jednog teksta predstavljaju prevod odgovarajućih elemenata drugog teksta. Uspostavljanje veza između odgovarajućih (prevodnih) delova paralelnih tekstova nije nimalo jednostavno i može se ostvariti na nekoliko nivoa u zavisnosti od toga koje jedinice teksta (paragrafi, rečenice, reči) se uparuju.

Najjednostavniji nivo paralelizacije je nivo *paragrafa* (Erjavec 2002, Steinberger et al., 2006). Na tom nivou uparivanje je relativno jednostavno, i tekstovi se u najvećoj meri pravilno uparuju. Međutim, već na tom nivou, pogotovo kad su u pitanju književni tekstovi, mogu da se pojave određeni problemi. Naime, može se desiti da u jednom od tekstova nedostaje jedan ili više paragrafa, a isto tako i da dva paragrafa u jednom tekstu u drugom tekstu budu spojena u jedan.

Nivo *rečenice* predstavlja sledeći nivo paralelizacije. Tekstovi paralelizovani na nivou rečenica predstavljaju osnovne jezičke resurse za uporedna istraživanja više jezika. Na ovom nivou, problem uparivanja postaje mnogo ozbiljniji i teži. Pre svega, mogu se pojaviti slični problemi kao kod uparivanja paragrafa. Pojedine rečenice mogu nedostajati, a jednoj rečenici u jednom tekstu mogu odgovarati dve, tri, pa čak i osam rečenica u drugom tekstu. No pored toga, tokom prevođenja može doći i do izmena u strukturi teksta, kao što je promena u redosledu rečenica, koje drastično otežavaju uparivanje.

Ako se izuzme uparivanje na nivou morfema, nivo *reči* je definitivno najteži kada je u pitanju problem uparivanja. Tu se, po pravilu, javljaju svi pomenuti problemi: pojedine reči nedostaju, redosled reči nije očuvan, nedeljivoj grupi reči u jednom tekstu odgovaraju nesusedne reči u drugom tekstu.

Zbog svih ovih problema, automatska paralelizacija tekstova nije nimalo jednostavan zadatak i razvijene su različite metode za njegovo rešavanje. Sve te metode mogu se svrstati u tri grupe:

---

[1] http://www.w3.org/XML

[2] http://www.tei-c.org

[3] http://www.w3schools.com/dtd/default.asp

- **statističke** (zasnovane na broju reči ili karaktera)
- **leksičke** (koje se primenjuju na prethodno leksički obrađene tekstove i koriste eksterne lingvističke resurse, npr. rečnike)
- **leksičko/statističke** (kombinuju dobre strane prethodne dve grupe metoda)

Koja god od pomenutih metoda da se koristi, proces paralelizacije ne može do kraja da se automatizuje, odnosno ne završava se automatskom paralelizacijom. Neophodno je da se izvrši provera da li je u automatskoj paralelizaciji tekstova došlo do grešaka, kao i da se eventualne greške isprave. Zbog toga se u trećem koraku vrši vizuelizacije paralelizovanog teksta, i na osnovu toga kontrola i eventualna korekcija uparivanja.

Četvrti korak je generisanje TMX formata paralelizovanog teksta. Naime, Grupa za jezičke tehnologije Univerziteta u Beogradu se opredelila da za paralelizovane tekstove koristi TMX (Translation Memory eXchange) format, koji je razvila LISA (Localisation Industry Standards Association)[4], organizacija koja se bavi razvojem standarda za softverske alate za prevođenje i lokalizaciju, kao što su Trados, SDLX, Déjà Vu i drugi. TMX predstavlja XML standard za skladištenje tzv. prevodilačkih memorija (*translation memory - TM*) i njihovu razmenu između pojedinačnih aplikacija. Kako prevodilačke memorije predstavljaju zbirke odrednica u kojima je izvorni tekst povezan sa odgovarajućim prevodom na jednom ili više ciljnih jezika, ovakav format je pogodan i za paralelizovane tekstove koji takođe nastaju prevođenjem sa jednog jezika na drugi. U TMX formatu se koriste ISO standardi za datume, vreme, i oznake država i jezika. Trenutna verzija standarda je TMX 1.4b, mada je nedavno publikovan i predlog verzije 2.0.

Problem sa TMX formatom je taj što polazni dokumenti nisu očuvani kao kompaktne celine. Naime, TMX dokument se sastoji iz tzv. jedinica prevođenja (*translation unit - tu*), pri čemu svaka od njih sadrži uparene segmente paralelnih tekstova, koji se nazivaju varijantama jedinice prevođenja (*translation unit variant - tuv*). Međutim, programski paket koji se koristi za kreiranje paralelizovanog korpusa zahteva da se za svaki od jezika paralelnih tekstova kreira poseban korpus, između kojih se onda uspostavljaju posebne veze koje omogućavaju paralelnu pretragu. Zbog toga ulazni tekstovi za paralelizovani korpus na pojedinačnim jezicima moraju da budu kompaktni i razdvojeni, tj. svaki tekst mora biti u posebnoj datoteci, u XML formatu sa obeleženim jedinicama prevođenja. Razlaganje TMX dokumenta na XML dokumente za svaki pojedinačni jezik, uz čuvanje informacija o jedinicama prevođenja, obavlja se u petom koraku.

Programski paket koji se koristi za kreiranje paralelnog korpusa takođe zahteva i da ulazni tekstovi budu vertikalizovani. Vertikalizovan tekst je tekst kod kog se u jednom redu nalazi samo jedan token, gde se pod tokenom podrazumeva reč, broj, znak interpunkcije, XML-obeležje (npr. <p> kao oznaka za početak pasusa), XML-komentar (proizvoljan tekst između ‘<!--’ i ‘-->’) i sl. Stoga se tekstovi razloženi u prethodnom koraku na XML dokumente u šestom koraku vertikalizuju.

Svi koraci su podređeni poslednjem, sedmom koraku, u kome se vertikalizovani XML dokumenti sa očuvanim jedinicama prevođenja predaju programskom paketu koji omogućava kreiranje korpusa sa morfološkom i strukturnom anotacijom, indeksiranje tekstova, kao i njihovu efikasnu pretragu korišćenjem regularnih izraza. Pošto svaka varijanta teksta na pojedinačnim jezicima ima isti broj jedinica prevođenja, programski paket nema problema da uspostavi odgovarajuće veze.

---

[4] http://www.lisa.org/tmx

3. Softver za paralelizaciju

Sa ciljem što efikasnijeg postupka kreiranja paralelnog korpusa u svakom od navedenih koraka koristi se odgovarajući program, odnosno softverski alat. Time su gotovo svi koraci automatizovani, izuzev prvog (priprema i segmentacija) i trećeg (kontrola i korekcija uparivanja), koji su poluautomatski.

Postoji više različitih softverskih rešnja za paralelizaciju tekstova, a Grupa za jezičke tehnologije Univerziteta u Beogradu opredelila se za programske pakete XAlign i Concordancier razvijene u okviru laboratorije Loria u Francuskoj[5]. Prvi programski paket obavlja automatsku paralelizaciju tekstova koristeći statističke metode, dok drugi program omogućava pretragu paralelnih konkordanci korišćenjem regularnih izraza, kao i kontrolu i korekciju uparivanja. Svi ovi programi su kreirani korišćenjem programskog jezika JAVA, te se mogu koristiti nezavisno od toga kojim operativnim sistemom korisnik raspolaže. Za razliku od programa Concordancier koji ima grafički korisnički interfejs, programi paketa XAlign su napravljeni za korišćenje iz komandne linije. Stoga za korisnike koji su naviknuti na grafički korisnički interfejs i ne snalaze se sa batch datotekama ili skriptovima tipa Makefile, oni nisu nimalo jednostavni za korišćenje. Programi za automatsku paralelizaciju kao ulaz očekuju XML datoteke a kao izlaz takođe proizvode XML datoteke.

Za konverziju rezultata paralelizacije dobijenih korišćenjem programskih paketa XAlign i Concordancier u TMX format i dodatne mogućnosti vizuelizacije i korekcije paralelizovanog teksta koriste se odgovarajući moduli u okviru integrisanog okruženja WS4LR (WorkStation for Lexical Resources) razvijenog u okviru Grupe za jezičke tehnologije (Krstev et al., 2006).

U svrhu konverzije TMX formata u vertikalizovan format, razvijena su dva modula. Prvi modul razlaže TMX dokument na XML dokumente za svaki pojedinačni jezik tako da su očuvane informacije o jedinicama prevođenja. Drugi modul potom vrši konverziju tako dobijenih dokumenata u vertikalizovan tekst.

U poslednjem koraku koristi se programski paket IMS Corpus Workbench (CWB), razvijen na Univerzitetu u Štutgartu[6] koji omogućava kreiranje korpusa sa morfološkom i strukturnom anotacijom, indeksiranje tekstova i njihovu efikasnu pretragu korišćenjem regularnih izraza (Christ and Schulze, 1996).

4. Integrisano razvojno okruženje za paralelizovane korpuse

Nedostatak udobnog okruženja sa grafičkim korisničkim interfejsom, koje bi u sebi objedinilo sve pojedinačne softverske komponente koje se koriste u raznim fazama pripreme paralelnih tekstova za paralelizovane korpuse, predstavljao je glavnu motivaciju da Grupa za jezičke tehnologije Univerziteta u Beogradu razvije sopstveno integrisano razvojno okruženje za paralelizovane korpuse ACIDE (Aligned Corpora Integrated Development Environment). Ovo okruženje, između ostalog, obezbeđuje grafički korisnički interfejs (GUI) za:

- paralelizaciju
- vizuelizaciju paralelizovanog teksta, kontrolu i korekciju
- generisanje datoteka u TMX formatu

---

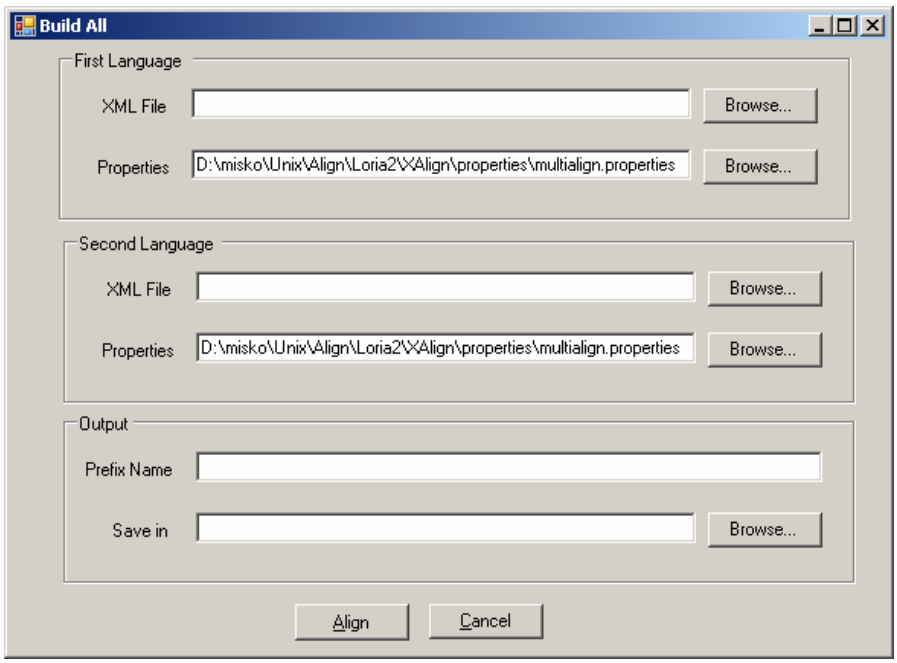
[5] http://led.loria.fr/outils/ALIGN/align.html

[6] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench

- razlaganje datoteka u TMX formatu na datoteke pojedinačnih jezika
- vertikalizaciju teksta

Sve navedene funkcije su dostupne preko menija Alignment, Tools i TMX.

Meni Alignment obezbeđuje GUI za programske pakete za paralelizaciju laboratorije Loria. Pojedinačne stavke u meniju omogućavaju korišćenje svakog od programa iz pomenutih paketa ponaosob, ali postoji i opcija Build All kojom se funkcije svih programa objedinjuju (slika 1).



Slika 1. Panel za objedinjeno pozivanje programa za paralelizaciju

Pri korišćenju opcije Build All, korisnik samo treba da zada ulazne tekstove u obliku XML datoteka, kao i ime i adresu izlaznih datoteka. Pritiskom na dugme Align biće generisani rezultati automatske paralelizacije i odmah prikazani u programu Concordancier gde se mogu pregledati, pretraživati i korigovati. Podrazumeva se da su ulazni tekstovi prethodno obeleženi u skladu sa preporukama TEI, pri čemu je korisniku omogućeno da specificira koja obeležja treba uzeti u obzir prilikom paralelizacije. Tako, na primer, Grupa za jezičke tehnologije Univerziteta u Beogradu koristi XML obeležja , <p> i <seg> da njima redom označi poglavlja, paragrafe, segmente (rečenice), pa se program za paralelizaciju podešava da uparivanje vrši na nivou ovih jedinica. Primera teksta u XML formatu pripremljenog za XAlign dat je u okviru.

```
<body>
    <div type='article'>
    <title>Knjige o novom alžirskom ratu</title>
    <subtitle>Grozna neprijatnost za intelektualce</subtitle>
    <author>Akram B. Elijas</author>
    <p><seg>"Ko ubija?"</seg><seg>Sećamo se da je ovo pitanje postavljeno posle jezivih pokolja koji su okrvavili, u jesen 1997, okolinu grada Alžira.</seg> <seg>U to vreme, mnogi posmatrači pominjali su odgovornost vojske.</seg><seg> Međutim, time su na sebe navukli bes gradskih vlasti, ali i francuskih intelektualaca koji su poleteli u pomoć alžirskom režimu: napravljen je niz reportaža i dokumentarnih filmova, u nameri da se dokaže kako je jedino islamizam, sa svojim oružanim grupama, kriv za takve zločine.</seg></p>
    </div>
</body>
```

Kao što je već pomenuto, metoda koju koristi XAlign je zasnovana na broju karaktera (tj. dužini segmenta). Ovakav pristup je veoma uspešan (i do 96% tačnih uparivanja). Proizvedeni rezultat su XML dokumenti sa informacijom o uparivanju. Od dva ulazna teksta, XAlign kreira tri izlazna XML dokumenta: prva dva su kopije ulaznih tekstova takve da je svakom segmentu pridružen atribut **id** čija je vrednost jedinstven identifikator – redni broj segmenta u okviru teksta; treći izlazni dokument čuva informacije o uparenim segmentima, koristeći njihove identifikatore. Rezultat paralelizacije XAlign-om ilustruje naredni primer.

Prvi izlazni dokument (tekst na srpskom, original):

```
<seg id="n94"> Sportska prognoza je igra u kojoj učesnik, popunjavanjem listića koji izdaje priređivač igre na kojem su označeni takmičarski parovi, pogađa rezultat fudbalske ili druge sportske utakmice za svaki takmičarski par, koristeći oznake predviđene pravilima igre. </seg>
```

Drugi izlazni dokument (tekst na engleskom, prevod):

```
<seg id="n98"> Sports pool is a game in which a player takes part by filling in a ticket, issued by the game organizer, with previously printed opponents in matches, e.g. soccer or other. </seg>
<seg id="n99"> The player guesses the results of the matches on the ticket for each pair using symbols defined by the rules of the game. </seg>
```

XML kodirana informacija o uparivanju (treći izlazni XML dokument):

```
<link targets="n98 n99" type="linking" id="l3" />
<link targets="x94 l3" />
```

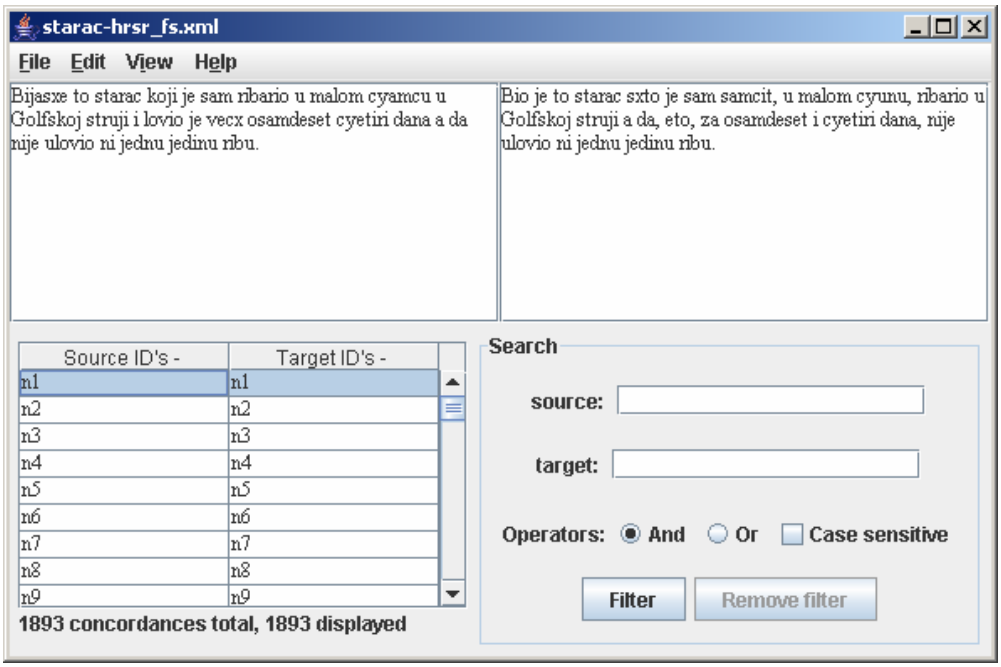
Kada je u pitanju XML kodirana informacija o uparivanju, treba imati u vidu da se u toku paralelizacije identifikatori segmenata iz jednog teksta (u ovom slučaju srpskog) preimenuju tako što se ‘n’ zameni sa ‘x’ kako bi se razlikovali od identifikatora u drugom tekstu. U ovom primeru segmenti iz engleskog teksta čiji su identifikatori n98 i n99 spojeni su, i dodeljen im je novi identifikator l3. Potom je uspostavljena veza između spojenih segmenata na engleskom sa identifikatorom l3 i segmenta na srpskom sa identifikatorom x94 (izvorno n94).

S obzirom na to da je format koji se generiše automatskom paralelizacijom praktično nečitljiv za običnog korisnika, Loria je razvila program Concordancier kojim se omogućava

vizuelizacija uparivanja, korekcija pogrešno uparenih segmenata, kao i pretraga paralelnih konkordanci pomoću regularnih izraza.

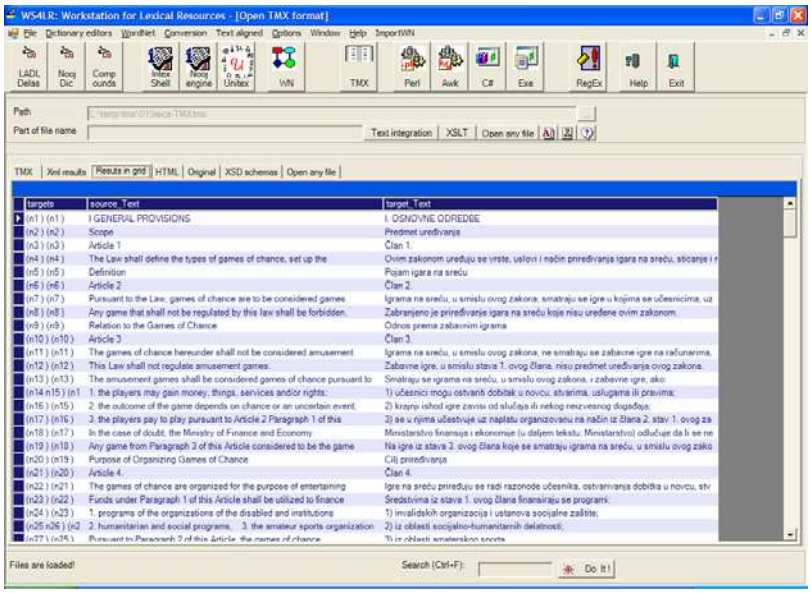
Iako je uparivanje na nivou rečenice u 90% slučajeva 1:1 (jednoj rečenici teksta u jednom jeziku odgovora jedna rečenica teksta u drugom), prilikom prevođenja, kao što je već napomenuto, moguće je i drugačije uparivanje, najčešće 1:2, 2:1, 1:3 , 3:1, 2:2, 1:0 ili 0:1. Sem toga, ponekad tokom prevođenja redosled rečenica može biti promenjen čime dolazi do unakrsne zavisnosti između rečenica. Konačno, može se desiti da posle uparivanja 1:2 ili 2:1 koje je pogrešno, dođe do "smicanja" prilikom 1:1 uparivanja velikog broja rečenica koje slede, posebno ako su približno iste dužine. Naime "smaknute" rečenice se uparuju 1:1, ali tako što se rečenica u jednom jeziku, umesto rečenice koja jeste njen prevod, uparuje sa rečenicom koja za ovom sledi. Ovake greške prilikom uparivanja moraju se ručno ispravljati, što omogućava program Concordancier.

Na slici 2 dat je izgled panela programa Concordancier u kome su prikazana dva uparena segmenta hrvatskog i srpskog prevoda romana "Starac i more", Ernesta Hemingveja.



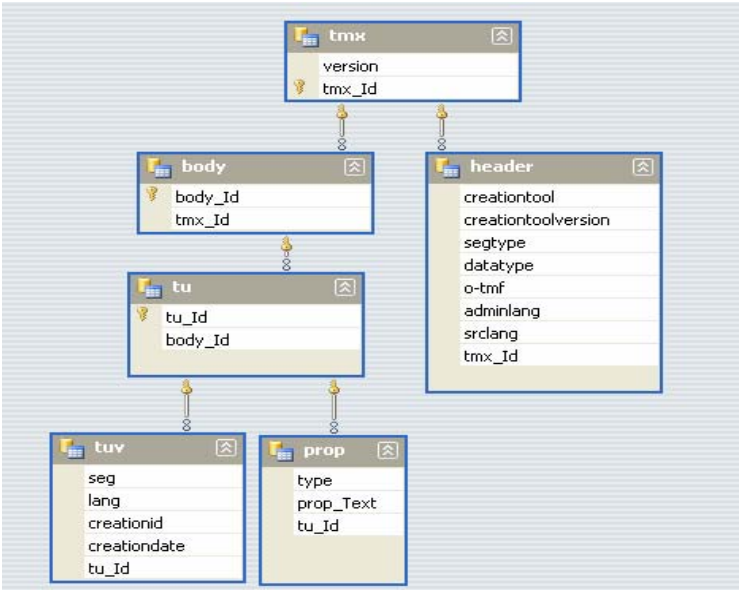
Slika 2. Uparivanje alatom Concordancier

Kako ispravljanje nekih grešaka uz pomoć programa Concordancier može biti vrlo mukotrpno, to su u okviru ACIDE okruženja, kroz module softverskog alata WS4LR, obezbeđene i dodatne mogućnosti vizuelizacije i korigovanja grešaka uparivanja, koje korisniku mogu znatno olakšati uspešno završavanje zadatka paralelizacije. Naime, integrisanjem jedinica paralelizovanih tekstova u memoriji računara kreira se interna tabelarna reprezentacija (*dataset*), pa se tabela sa uparenim rečenicama može prikazati kao na slici 3. Korisniku se potom daje mogućnost da izvrši dodatne ispravke grešaka uparivanja (kao što je "smicanje"). Sem tabelarnog prikaza, korisniku je omogućeno da generiše tekstualni zapis i XML oblik integrisanih podataka. Konačno, koristeći odgovarajuću XSLT (Extensible Stylesheet Language Transformations) transformaciju, korisnik može XML oblik transformisanih podataka prevesti u HTML i druge formate, u zavisnosti od vrste vizualizacije koja mu je potrebna.



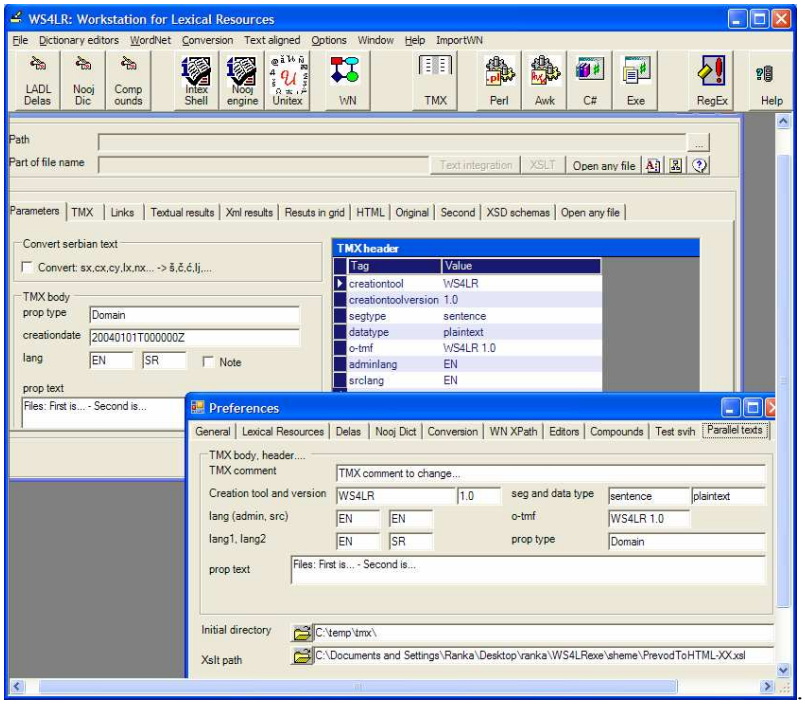
Slika 3. Tabela integrisanih jedinica paralelizovanog teksta

Posle kontrole i eventualne korekcije uparivanja ekvivalentne jedinice paralelizovanog teksta (obično po jedna ili više rečenica) se prevode u TMX dokument. Izgled TMX dokumenta koji se pritom kreira, odnosno njegova XML šema prikazana je na slici 4 kroz odgovarajući XSD (XML Schema Definition). Iz šeme se vidi da tekst u TMX formatu sadrži zaglavlje (header) i "telo" teksta (body) koje čine jedinice prevođenja (tu) sa odgovarajućim karakteristikama (prop), a unutar njih, varijante jedinica prevođenja (tuv).



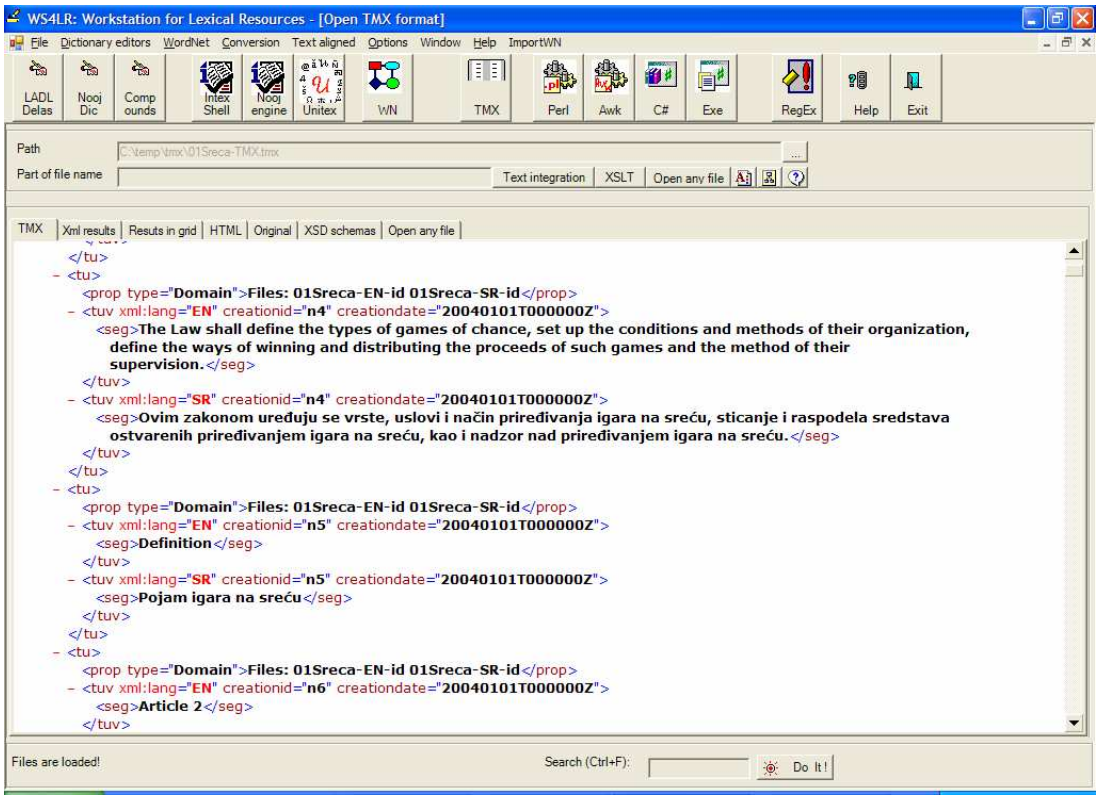
Slika 4. XSD shema TMX formata paralelizovanog teksta

Na slici 5 je prikazan panel iz WS4LR za zadavanje parametara za kreiranje dokumenta u TMX formatu na osnovu XML dokumenata koje je proizveo XAlign, a korisnik potom iskontrolisao i eventualno iskorigovao. Kreirani TMX dokument se može koristiti odmah nakon generisanja, ali se može i naknadno učitati, obrađivati i pretraživati.

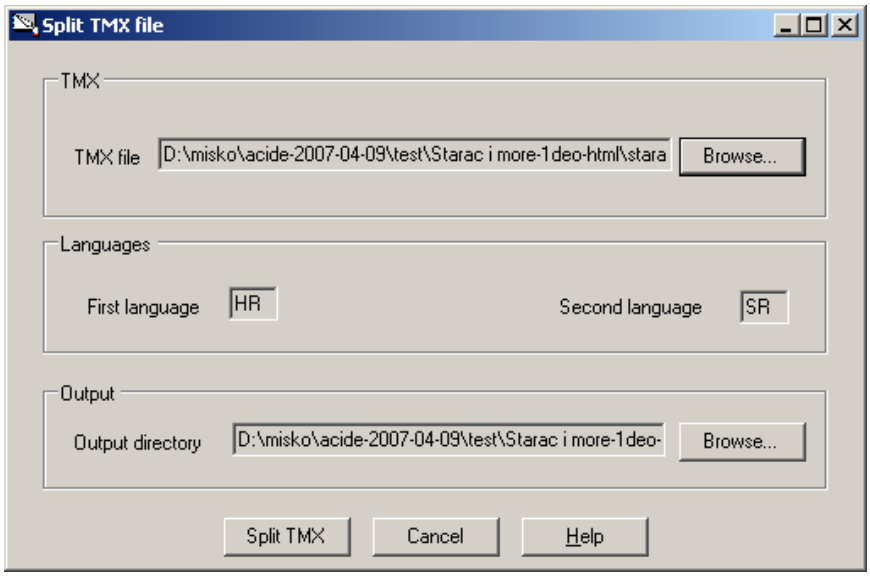


Slika 5. Parametri za generisanje TMX dokumenta

Izgled TMX dokumenta koji se dobija, odnosno izgled paralelizovanog teksta u TMX formatu koji je rezultat konverzije iz XML dokumenata koje generiše XAlign u jedinstveni TMX dokument prikazan je na slici 6. Na njemu se mogu uočiti prevodne jedinice <tu> sa svojim karakteristikama <prop> i unutar njih varijante jedinica prevođenja, sa naznačenim jezikom (lang="SR" odnosno lang= "EN") i paralelizovanim jedinicama <seg>.



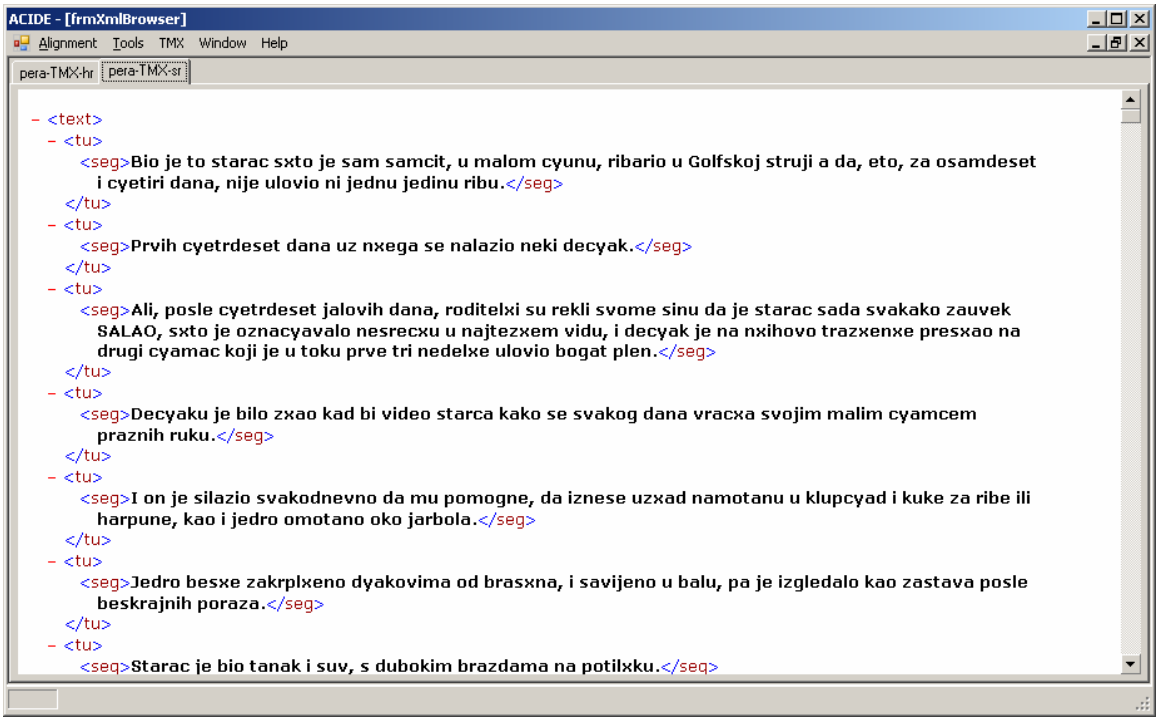
Slika 6. TMX format paralelizovanog teksta



Slika 7. TMX format paralelizovanog teksta

Razlaganje TMX formata na pojedinačne jezike, neophodno za formiranje paralelnog korpusa, u okruženju ACIDE se vrši opcijom Split TMX, koja se nalazi u okviru menija TMX (slika 7).

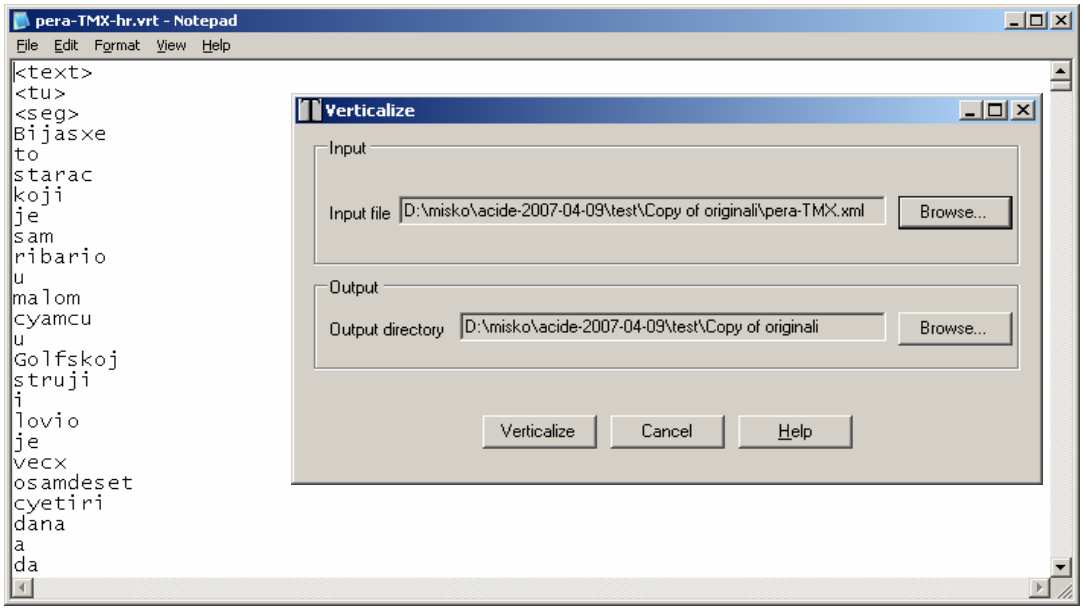
Kao rezultat razlaganja TMX dokumenta formiraju se dva XML dokumenta, gde se svaki od tekstova koji su paralelizovani ponovo nalazi u celini, ali ovoga puta sa očuvanim informacijama o jedinicama uparivanja. Izgled jednog takvog dokumenta koji sadrži paralelizovan tekst romana "Starac i more" u tzv. Aurora zapisu, u kome su slova ć, č, š ,ž ,đ, dž, lj i nj kodirana ACCII karakterima cx, cy, sx, zx, dx, dy, lx i nx, prikazan je na slici 8.



Slika 8. Razlaganje TMX dokumenta

Poslednji korak u pripremi paralelizovanog teksta za korpus je njegova vertikalizacija, koja se obavlja korišćenjem opcije Verticalize u meniju Tools. Rezultat vertikalizacije (slika

9) predstavlja osnovni resurs za kreiranje paralelnog korpusa pomoću programskog paketa IMS CWB.



Slika 9. Vertikalizacija teksta

Programsko okruženje Acide razvijeno je korišćenjem programskog jezika C#. Stoga je za ovo okruženje potrebno da računarski sistem na kome se instalira ima prethodno instaliran Microsoft .NET Framework 2.0, što znači i Windows XP SP2, kao i odgovarajuće hardverske resurse. Razlog za korišćenje ove tehnologije je jednostavnost pri programiranju korisničkog interfejsa, kao i dobra podrška za rad sa XML podacima.

5. Zaključak

Priprema paralelnih tekstova za paralelni korpus predstavlja kompleksan zadatak koji se odvija u nekoliko koraka. Veći dao ovog zadatka moguće je automatizovati, ali je intervencija korisnika u određenim fazama neophodna. Da bi se korisniku olakšalo obavljanje ovog zadatka razvijeno je integrisano razvojno okruženje za paralelizovane korpuse ACIDE koje se trenutno nalazi u fazi završnog testiranja.

Kako bi se postigao što viši stepen automaizacije, u planu je da se u ACIDE dodaju funkcije za proveru raznih vrsta grešaka, kao i za kreiranje i upoređivanje raznih vrsta statistika vezanih za paralelne tekstove (npr. uporedne liste učestanosti). Sem toga u neposrednoj budućnosti će se u interakciji sa korisnicima dodatno unaprediti korisnički interfejs, kako bi za korisnika postao još udobniji, a s tim u vezi predstoji i kreiranje detaljne dokumentacije o svim mogućnostima ovog integrisanog okruženja.

Literatura

[1] Christ, Oli and B.M. Schulze, (1996): "Ein flexibles und modulares Anfragesystem für Textcorpora". In: Feldweg, H.; Hinrichs, E. W. (eds.) (1996). Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen, Tübingen: Niemeyer, pp. 121-134.

[2] Gale William A., Kenneth W. Church (1993): “A program for aligning sentences in bilingual corpora, Computational Linguistics”, Vol. 19/1, pp. 75 – 102.

[3] Krstev Cvetana, Ranka Stanković, Duško Vitas, Ivan Obradović (2006): "WS4LR - a Worksation for Lexical Resources", in Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, May 2006, pp. 1692-1697.

[4] Dimitrova, L., T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevič, D. Tufiş (1998): "Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages", in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics *COLING-ACL '98*. Montréal, Québec, Canada, pp. 315-319.

[5] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, Daniel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, 2006.

[6] Tomaž Erjavec: Compiling and Using the IJS-ELAN Parallel Corpus. *Informatica*, 26(3), pp. 299-307, 2002.

SUMMARY

The development of aligned corpora requires a preparation of parallel texts for their integration into aligned corpora. This is a very complex task, which can be solved in different ways, and which has to be realized in several of steps. At the beginning of this paper we outline the procedure for preparation of parallel texts for aligned corpora which is being used in the Human Language Technology Group at the University of Belgrade. Texts are marked using XML tags, in accordance with the TEI (Text Encoding Initiative) consortium recommendations, and their alignment is performed at the level of paragraphs and sentences. We then give an overview of the software, namely programs (XAlign, Concordancier, WS4LR) that are used for alignment. The absence of a comfortable environment with a graphical user interface, where all these programs would be united, motivated the Human Language Technology Group to embark on the task of developing an integrated environment for the preparation of aligned corpora, under the name of ACIDE. For the construction of this environment we chose the C# programming language. Among other things, ACIDE provides a graphical user interface (GUI) for alignment and visualization of aligned texts, their control and correction, as well as generation of files in TMX format. ACIDE also enables the decomposition of TMX files into files for particular languages, and text verticalization. In order to achieve a higher level of automatization, we plan to add functions for verification and detection of different types of errors, as well as for generation and comparison of different types of statistics related to aligned texts (for example, comparative frequency lists). Finally, the creation of detailed documentation on all the possibilities of this integrated environment is also envisaged.