Sveučilište u Zagrebu
Filozofski fakultet
Odsjek za lingvistiku

# KONTRASTIVNA ANALIZA ALATA ZA RAČUNALNO SRAVNJIVANJE REČENICA

**Diplomski rad**

Studentica:
Željana Zokić

Mentorica:
dr.sc. Ivana Simeon

Zagreb, listopad 2018.

# Sadržaj

**Sažetak**

Jezične tehnologije 21. stoljeća konstantno se razvijaju zahvaljujući uspješnoj suradnji dviju znanosti – računarstva i lingvistike. Digitalizacija tekstova koji su donedavno bili dostupni samo u papirnatome obliku dala je dodatni vjetar u leđa prikupljanju jezičnih resursa potrebnih za izgradnju jezičnih alata, i u konačnici, komercijalnih sustava i aplikacija. Jedno od zanimanja koje je značajno profitiralo od razvoja novih jezičnih tehnologija jest prevođenje. Digitalizacijom izvornih tekstova i njihovih prijevoda te izgradnjom paralelnih korpusa stvoreni su preduvjeti za razvoj sustava za računalno potpomognuto prevođenje koji se baziraju na prijevodnim memorijama. Prevođenje uz pomoć CAT (Computer-Aided Translation) alata brže je i omogućuje terminološku konzistentnost. Korištenje programa za sravnjivanje rečenica paralelnih tekstova preduvjet je za stvaranje kvalitetnih prijevodnih memorija. Danas nam je dostupan čitav niz takvih proizvoda, kako komercijalnih, tako i onih slobodnog pristupa, a njihovi razvojni inženjeri i krajnji korisnici pokušavaju pronaći optimalno rješenje.

**Ključne riječi**: jezične tehnologije, računalna lingvistika, jezični korpus, sustavi za sravnjivanje rečenica

**Abstract**

Language technologies of the 21st century are constantly evolving thanks to the successful collaboration of two sciences - computing and linguistics. The digitization of texts that were recently available only in paper form, went in favor of collecting large number of language resources needed to build language tools and, ultimately, commercial systems and applications. Translation is the profession that has surely profited the most. By digitizing original texts and their translations and building parallel corpora, the preconditions for the development of computer-assisted translation systems based on translation memories were created. Translating with CAT (Computer-Aided Translation) tools is faster and allows terminology consistency. Using text alignment programs is a prerequisite for creating quality translation memories. Today, a whole range of such products, commercial, and open source, are available for use and their developers and users are trying to find the optimal solution.

**Keywords**: language technologies, computational linguistics, language corpus, sentence alignment systems

# 1. Uvod

U današnje vrijeme sve su zastupljeniji jezični alati i aplikacije koji pronalaze korisnike među različitim generacijama i profesijama. Postalo je potpuno uobičajeno koristiti se aplikacijama koje omogućavaju glasovno diktiranje i prenošenje poruka u tekstualnom obliku izravno na računalo, pretraživanje weba obuhvaćajući različite jezike zahvaljujući integriranim automatskim prevoditeljima te, već duže vrijeme, pisanje tekstova uz pomoć sustava koji ispravlja ortografiju, tzv. *spell checkera* (provjernika pravopisa). S druge strane, velika većina korisnika vjerojatno nije svjesna da u stvaranju ovih alata sudjeluju ne samo informatičari i inženjeri, već i lingvisti. Štoviše, računalna lingvistika postala je ekonomski najproduktivnija grana lingvistike koja sa sve bržim razvojem informatike pronalazi nova polja u kojima se može angažirati.

Većina laika danas gleda na lingviste kao na načitane ljude koji se bave jezicima, izučavaju ih te ispravljaju pogreške drugih ljudi; međutim, izvrsni lingvisti danas također sudjeluju u stvaranju novih aplikacija koji se temelje na jezičnim podatcima u suradnji s informatičarima. Nove jezične tehnologije produkt su rada multidisciplinarnog tima čije su jednako bitne sastavnice informatičari i lingvisti.

Ipak, na jezične tehnologije kao produkt suradnje lingvistike i računarstva iste te znanosti gledaju s drugačije točke gledišta. Grana lingvistike koja sudjeluje u stvaranju novih jezičnih tehnologija naziva se računalna lingvistika, dok informacijske znanosti proces izgradnje sustava koji se temelje na jeziku nazivaju *Natural Language Processing* (NLP). Naime, računalna lingvistika ima zadatak da uz pomoć računala, metodološki ispravno, prikupi veliku količinu jezičnih podataka u obliku rječnika i korpusa koji će poslužiti kao sirovina za konačno stvaranje aplikacija za obradu samog jezika od strane programera. „Obradom te građe dobivaju se sekundarni podatci te se na njihovim temeljima izrađuju računalni modeli funkcioniranja pojedinih jezičnih podsustava ili, u iznimnim slučajevima, sustava u cjelini.“ (Tadić, 2003:10) S druge strane, stručnjacima iz informacijskih tehnologija jezik predstavlja još jedan oblika podataka koji se mogu obraditi, kao što su npr. i sami brojevi; međutim, zbog prirode ovakvih sustava i složenosti svakog jezika, bez dobre lingvističke potkovanosti nije moguće doći do krajnjeg proizvoda koji će pokriti sve moguće varijacije koje se mogu javiti u nekom jeziku. Kao što lingvisti i računalni stručnjaci sudjeluju u stvaranju novih jezičnih tehnologija tako i obje grane moraju sudjelovati u

njihovom evaluiranju. U konačnici je potrebno moći procijeniti koliko je neki proizvod uspio u svojoj nakani da, uz mogućnost koju nudi pojedini programski pristup, zadovolji svoje konačne korisnike uz što manje pogrešaka. Povratna informacija od samih korisnika također je bitna za daljnji razvoj i nadogradnju sustava.

## 2. Razvoj korpusne lingvistike

Povijest korpusne lingvistike vezana je za povijest kvantitativnog usmjerenja u lingvistici koji počinje već u 18. i 19. stoljeću, dakle prije računalne ere. Mnoga istraživanja na području komparativne i povijesne lingvistike temeljila su se na analizi tekstualne građe. Štoviše, jedna od prvih sastavljenih gramatika, Paninijeva, temeljena je na analizi korpusa jezika koji je korišten u Vedama. Analiza korpusa pokazala se također prikladna i u leksikografiji. Pri sastavljanju rječnika leksikografi su nastojali izvući primjere iz različitih tekstova kako bi objasnili značenje određenog pojma ili kako bi ga prikazali u najčešćem kontekstu. Poznato je tako Crudenovo istraživanje biblijskih konkordancija. (Lüdeling i Kytö, 2008:1) Ručno prikupljanje i analiza nekoć je bio mukotrpan zadatak, dok danas uz pomoć jezičnih alata možemo u kratkom roku prikazati veliki broj konkordancija, tj. riječi u kontekstu. Ako ga uspoređujemo s današnjim načinom izgradnje korpusa, nekadašnje sastavljanje predelektroničkih korpusa može nam izgledati nepotrebno, međutim da nije postojala sama svijest o potrebi prikazivanja konkordancija ni danas ne bismo imali kompleksna softverska rješenja koja nam to omogućavaju.

### *2.1. 20. stoljeće i distribucionalizam*

Duga tradicija analize lingvističkih podataka započela je tijekom devetnaestog stoljeća s historicističkim istraživanjima o promjeni i evoluciji jezika te studijama jezične tipologije Augusta Wilhema Schlegela. Međutim, ove empirijske studije imale su ozbiljna ograničenja u vidu prikupljanja, pohrane i analize podataka, nije bilo adekvatne tehnologije za proučavanje ljudskog jezika. Kasnije, u dvadesetom stoljeću, američki distribucionalizam koji su predvodili Bloomfield (1933) te, posebno, Z. Harris (1954), predlaže metodu analize, tj. distribucijsku metodu, koja polazi od promatranja i kvantifikacije te zabilježavanja frekventnosti jezičnih oblika s ciljem otkrivanja njihove unutarnje strukture. Međutim, ova je metoda također ograničena nemogućnošću pristupa kompilacijama jezičnih podataka potrebnih za njihovu daljnju analizu i klasifikaciju.

Bloomfield je lingvistiku želio predstaviti kao egzaktnu znanost koja se oslanja na prikaz lingvističkih situacija koje je moguće proučavati analizirajući stimulus i reakciju određenog ponašanja, a ne značenje koje je smatrao nedokučivim. Bloomfieldova teorija razvrstavanja jezičnih jedinica (distribucijska teorija) temelji se na zamjeni i razdiobi jezičnih jedinica. Osnovna zadaća teorije je utvrditi na koji način jezične jedinice koje imaju istu razdiobu, a međusobno su zamjenjive, mijenjaju cjelokupno značenje iskaza te ih na taj način razvrstati. Dobar primjer u hrvatskom jeziku je smoči : smoći (Smoči noge! : smoći snage) ili spavačica : spavaćica (žena koja spava : odjevni predmet) (Glovacki Bernardi, 2007:181) Kao što možemo zaključiti, bitan je kontekst odnosno situacija u kojoj se određena inačica koristi. Za istraživanje distribucija potrebna je velika količina jezične građe koja se može jednostavno obraditi; drugim riječima, potrebni su korpusi iz kojih se mogu izvući kolokacije kako bi se moglo usporediti što više stvarnih jezičnih situacija.

### *2.2 Prvi elektronički korpusi*

Uzmemo li u obzir metodu prikupljanja korpusa, povijesno gledajući, moguće je razlikovati predelektroničke i elektroničke korpuse. Do značajnijeg razvoja korpusne lingvistike došlo je 70-ih i 80-ih godina 20. stoljeća zahvaljujući olakšanom pristupu računalno pohranjenim tekstovima koje je moguće elektronički analizirati. Jedan od najznačajnijih korpusa tzv. 'prve generacije korpusa' koji je bio računalno pohranjen i čitljiv jest Brown korpus. Pionir među elektronskim korpusima sadržavao je tada impresivnih milijun riječi američkog engleskog jezika iz raznih tekstova objavljenih 1961. godine. Bio je ograničen na 500 tekstova od kojih je svaki imao po 2000 riječi. Tekstovi su se mogli podijeliti na razne žanrove, točnije njih 15. Tu su se mogli naći tekstovi informativnog, religioznog, zabavnog sadržaja. Desetljeće poslije, u ranim 70-ima, sastavljen je britanski ekvivalent Brown korpusu – LOB korpus. Sastavljanje korpusa pokrenuo je Geoffrey Leech na Sveučilištu Lancaster, međutim dovršen je u Norveškoj odakle i dolazi kratica za njegovo ime, LOB (Lancaster-Oslo-Bergen). Rađen je prema istoj 'recepturi' kao i Brown korpus, sadrži jednaki broj riječi, tekstova i tema. Široka dostupnost korpusa koji su rađeni prema istim parametrima pokazala se korisnom istraživačima koji su željeli raditi komparativne analize različitih jezičnih varijeteta. Još jedan od 'malih' korpusa koji je potrebno spomenuti britanski korpus govornog jezika LLC (London Lund Corpus of Spoken British English). To je ujedno i prvi računalno pretraživ korpus govornoga jezika koji sadrži 87 tekstova od po 5000 riječi.

Govor je transkribiran i prozodijski označen te sadržava različite načine izražavanja kao što su spontani i pripremljeni govor. LLC je također uvršten u ICAME, jednako kao i LOB. ICAME je kratica za International Computer Archive of Modern English i uspostavljen je upravo zbog problema s vlasničkim pravima na određene tekstove s kojima se susreo LOB tijekom pripreme. U arhiv, tj. bazu postepeno se uključivalo sve više tekstova koji su se poslije pokazali kao izvrstan izvor za daljnja istraživanja. (Lüdeling i Kytö, 2008:38) Britanski lingvist John Sinclair započeo je novu eru korpusne lingvistike kritičkim osvrtom na dosadašnje milijunske korpuse i novim idejama za stvaranje još veće tekstualne baze. Smatrao je da su korpusi poput Brown korpusa premali za ozbiljnija istraživanja i da je pojavnost različitih riječi u njima nedostatna. Zalagao se sa iskorištavanje punog potencijala digitalizacije tekstova koje je omogućavala nova tehnologija 80-ih godina. Sinclair je bio začetnik Birmingham korpusa, prvog korpusa koji je prešao dotadašnji maksimum od milijun riječi. Birmingham korpus postao je baza za COBUILD projekt kojemu je cilj bio stvaranje opsežnog korpusa engleskog jezika u koji će se neprestano moći dodavati novi tekstovi. Bank of English je rezultat tog projekta koji danas sadrži 650 milijuna riječi.[1] Drugi 'mega' korpus koji je potrebno spomenuti je BNC (British National Corpus) koji je prvi premašio 100 milijuna riječi. Za razliku od BoE, on je konačan i raspoloživ svima koji ga žele nabaviti. BoE je dostupan samo istraživačima povezanim s COBUILD-om.

### *2.3. Paralelni korpusi*

Dvojezični paralelni korpusi također su se razvijali paralelno s jednojezičnim korpusima. Prvi povijesni primjer jednog takvog korpusa bio je kamen iz Rosette s tekstovima pisanim na grčkom i egipatskom jeziku te na trima različitim pismima. Današnji razvoj paralelnih korpusa prati potrebe modernog društva za što boljim prijevodima i očuvanjem jezične raznolikosti. Najbolji primjer možemo vidjeti u izgradnji paralelnog korpusa na temelju tekstova pravne stečevine Europske unije koji sadrži tekstove na 22 službena jezika EU i koji se neprestano nadopunjuje. Štoviše, nove države kandidati za ulazak u EU moraju prevesti oko 12 milijuna riječi na svoj nacionalni jezik prije postajanja punopravnim članicama. Paralelni korpusi pokazali su se iznimno korisnima upravo kod prijevoda pravnih tekstova jer ubrzavaju sam proces i omogućavaju

[1] Collins, Cobuild

dosljednost korištenjem već ovjerenih termina. Osim u prevođenju, paralelni korpusi korisni su i za lingvistička istraživanja.

# 3. Podjela jezičnih tehnologija

Uzimajući u obzir aspekt jezika koji obrađuju, jezične se tehnologije mogu podijeliti na one koje analiziraju govor te one koje analiziraju pisani jezik. Sama distinkcija međutim nije uvijek tako jasna. Možemo uzeti za primjer aplikacije za automatsko prevođenje govora u pisani jezik. U tom slučaju, te su dvije jezične razine međuovisne te se isprepleću.

Prije svega, za razvoj pojedinih tehnologija potrebno je imati bazu, tj. potrebne jezične resurse koji služe kao temelj za naknadni razvoj jezičnih tehnologija. Proces stvaranja koji vodi od teorije do konačnog proizvoda dostupnog za tržište prolazi kroz više faza, te jezične tehnologije sukladno tome možemo podijeliti na (Tadić, 2003:27):

- Jezične resurse,
- Jezične alate,
- Jezične aplikacije / komercijalne proizvode.

## *3.1. Jezični resursi*

Jezični su resursi računalno pribavljene, pohranjene i podržane zbirke jezičnih podataka, a sastoje se ponajprije od korpusa, a potom i od rječnika. (Tadić, 2003:28). Korpus je zbirka tekstova nekog jezika (ukoliko se radi o jednojezičnom korpusu) sastavljena po određenom kriteriju. Oni mogu biti računalno pretraživi ili u *offline* obliku. Korpusi su temelj lingvističkog istraživanja koji služe za razvijanje automatizacije lingvističkih alata. Treba uzeti u obzir da svaka zbirka tekstova ujedno ne čini i korpus ukoliko ne zadovoljava niz odrednica među kojima su dosljednost dizajna, uzorkovanje prema strogim lingvističkim kriterijima te pristup izvorima tekstova. Digitalizacijom tekstova i pretragom uz pomoć računala, također je postalo mnogo lakše sastavljati nove rječnike i korpuse te i na taj način stvarati bazu resursa za daljnja istraživanja. Kako bismo mogli iskoristiti puni potencijal korpusa za razvoj aplikacija o kojima će biti riječi naknadno, potrebno je da on bude kodiran i označen, tj. da sadržava dodatne lingvističke informacije koje se mogu interpretirati i obraditi uz pomoć sustava za obradu jezika, bilo u govornom ili tekstualnom obliku. Korpus može biti označen na više lingvističkih razina, pri čemu je označavanje najuspješnije na

morfološkoj i sintaktičkoj, dok su one složenije poput semantičke ili stilističke rjeđe zastupljene. (Garside et al, 2013:5). Posljednjih desetljeća, dobrim dijelom zbog vjetra u leđa uzrokovanog digitalizacijom tekstova, razvila se perspektivna grana u lingvistici – korpusna lingvistika, koja ima za cilj prikupljanje što kvalitetnijeg korpusa nekog jezika koji će taj jezik moći detaljno opisati i pozicionirati u svijetu informacijskih tehnologija.

S obzirom na vrstu, korpuse možemo podijeliti na jednojezične i višejezične usporedive ili paralelne korpuse (McEnery, prema Mitkov, 2003:450). Jednojezični korpus, kao što smo već naveli, predstavlja skup tekstnih odsječaka nekog jezika odabranih prema strogim i eksplicitnim kriterijima. Usporedivi korpusi sadrže tekstualnu građu na različitim jezicima prikupljenu prema jednakim parametrima i unutar jednakog tematskog okvira, dok su paralelni korpusi sadržajno jednaki tekstovi prevedeni iz izvornog jezika u ciljani jezik. Svaki od navedenih korpusa koristan je u različitim lingvističkim disciplinama.

### *3.2. Jezični alati*

Za razvoj računalnog potpomognutih prijevodnih sustava o kojima će biti riječ u daljnjem radu od najveće su važnosti paralelni korpusi koji predstavljaju osnovnu građu, 'sirovinu' iz koje se nakon označavanja korpusa putem jezičnih alata, i sravnjivanja prijevodnih ekvivalenata, dobivaju toliko vrijedne prijevodne memorije. Osim korpusa u tekstualnom obliku postoje i korpusi u govornom obliku. Takvi su korpusi kompleksni i, kako bi što kvalitetnije poslužili svrsi lingvističkog istraživanja (npr. dijalektalna ili prozodijska istraživanja) ili uporabi u sklopu računalnih tehnologija (npr. glasovno biranje), potrebno je da sadrže pisani transkript, ali i audio zapise. Jedno bez drugoga ne ide, jer audio snimke nije lako pretraživati zbog specifičnosti njihova formata, dok nam pisani zapisi ne omogućuju uvid u razlike u prozodiji koje su također bitne i koje se jedino mogu zabilježiti u audio formatu i jedino na taj način istražiti (McEnery, prema Mitkov, 2003:451).

Kako bismo lakše istraživali i iskoristili prednosti digitalnih korpusa, potrebno je iste i označiti. Možemo reći da je označavanje neka vrsta obogaćivanja sadržaja korpusa i njegovo nadopunjavanje različitim lingvističkim podatcima, na različitim razinama, što je ujedno i korak naprijed prema razvoju komercijalnih proizvoda nastalih u suradnji lingvistike i računarstva. Razvoj jezičnih alata zapravo je faza u kojoj lingvistika i računarstvo djeluju u najintenzivnijoj

suradnji. Označavanje se ne radi ručno, iako je i to moguće, ali bi zbog velike količine tekstova bilo neefikasno, ili bi trajalo jako dugo. Koriste se računalni programi koji rade automatsko označavanje i koji su napravljeni u suradnji s lingvistima i njihovim znanjem. Upravo u ovom segmentu vidimo prednosti razvoja računarstva koje je prije svega ubrzalo sam proces istraživanja i razvoja. Nastali alati, naravno, nisu nepogrešivi zbog kompleksnosti jezičnog sustava, te je stoga potrebna ljudska evaluacija.

Zahvaljujući klasifikaciji i interpretaciji teksta korpusa, moguće ga je i preciznije istraživati. Samo označavanje može biti izvantekstualno i strukturalno. Izvantekstualnim označavanjem tekstovima dodjeljujemo opće distinktivne informacije poput naslova, autora, datuma izdavanja, itd. dok je strukturalno označavanje moguće na više razina (Tadić, 2003:29):

1. glasovnoj/fonemskoj/grafemskoj razini

2. razini riječi (morfologija, leksikografija)

3. sintaktičkoj razini

4. semantičkoj razini

5. pragmatičkoj razini

Najčešće se označavanje odvija na morfološkoj i sintaktičkoj razini. Na morfološkoj razini koriste se alati za označavanje vrsta riječi (*POS taggers*), morfosintaktičke kategorije riječi (*MSD taggers*), npr. rod, broj i padež te lematizatori koji pridružuju riječ njenom kanonskom obliku koji najčešće odgovara natuknici u rječniku. Na sintaktičku se razinu prelazi tek nakon detaljne obrade na morfološkoj razini riječi. Alati koji se koriste za sintaktičku analizu mogu biti *parseri* ili razdjelnici. Razdjelnici razdjeljuju dijelove rečenice na lako prepoznatljive dijelove kao što su skupovi ili fraze dok *parseri* idu dublje u rečeničnu analizu oslanjajući se na formalne gramatike i mogu se podijeliti na:

1. plitke – određuju odnose ovisnosti dijelova u rečenici

2. duboke – rade sintaktičku analizu do razine leksičkih unosaka

3. robusne – zabilježavaju i neovjerene kombinacije riječi te ih analiziraju

Korpus koji je morfološki i sintaktički označen i spreman za pretraživanje nazivamo bankom stabala (*tree-bank*).

Za razvoj komercijalnih proizvoda za prevođenje također su nam bitni i sustavi za označavanje na semantičkoj razini koji označavaju semantičke uloge u rečenici kao što su agens, pacijens, itd.

Na samome početku, metode označavanja korpusa nisu bile unificirane i trebalo ih je uzeti sa zadrškom. Stoga Leech (Garside et al., 2013:6) objavljuje svojih 7 smjernica prema kojim se treba ravnati prilikom označavanja korpusa:

- Oznake treba moći ukloniti iz teksta, a da se pri tome ne promijeni njegov izvorni oblik, tj. da se ni na koji način ne utječe na njegovu strukturu.
- Oznake treba moći izvući iz samog teksta i odvojiti kako bi se mogle proučavati neovisno o tekstu.
- Potrebno je objaviti pravila prema kojima se odvija proces označavanja kako bi ih se svi istraživači mogli pridržavati.
- Mora biti jasno tko je i na koji način označio korpus.
- Potrebno je biti svjestan da oznake interpretacije nisu 100% sigurne niti su sustavi nepogrešivi.
- Oznake se moraju temeljiti na opće poznatim i neutralnim principima.
- Ne postoji niti jedan sustav označavanja koji se a priori može uzeti kao standardni.

Što se tiče zadnje točke Leechovih naputaka, potrebno ih je smjestiti u kontekst u kojem su napisane, a to je 1993. godina kada se još nisu usvojila TEI (*Text Encoding Initiative*) pravila koja danas predstavljaju standardni vodič pri označavanju i segmentiranju tekstova.

Također postoje i sustavi koji kombiniraju označavanje i analizu tekstova na različitim razinama na temelju čega nastaju sustavi za prevođenje ili sustavi za računalno potpomognuto učenje jezika.

### *3.3. Komercijalni proizvodi*

Krajnji rezultat razvoja različitih jezičnih alata predstavljaju aplikacije i programi koji su dostupni za korištenje stručnjacima u određenom području, ali i ostalim pojedincima kojima mogu biti od koristi. Najbolji su primjer, poznat velikoj većini ljudi koja koristi računalo, *spell check*

programi koji su najčešće integrirani u *office* programe ili ih se može naći na internetu. Osnova tih programa su alati za analizu teksta na razini višeslova i riječi. Postoje gramatički provjernici koji su pak složeniji i zahtijevaju kombiniranje različitih alata koji analiziraju tekst na razini morfologije i sintakse. Osim programa koji su slobodni za korištenje i skidanje, postoje i komercijalni programi za koje je potrebno imati licenciju i koji se moraju kupiti. Takvi su sustavi za strojno prevođenje kao što je SYSTRAN ili sve popularniji sustavi za prevođenje koji se temelje na prijevodnim memorijama kao što je Trados. Među ostalim proizvodima jezičnih tehnologija valja spomenuti sustave za indeksiranje tekstova, sustave za sažimanje, sustave za crpljenje informacija i nazivlja, sustave za diktiranje… (Tadić, 2003:43). Iz nabrojanih primjera možemo najbolje vidjeti kolika je zapravo korist od razvoja jezičnih tehnologija i širina područja u kojima se koriste.

## 4. Jezične tehnologije u prevođenju

U doba komunikacije na globalnoj razini te razvoja novih tehnologija, prenošenje informacija i ideja ne bi trebalo nailaziti na nepotrebne prepreke, pa samim time ni raznolikost jezika ne bi trebala predstavljati nepremostivu barijeru. Iako je danas engleski jezik svojevrsna *lingua franca* novoga doba, s druge strane, pa možda upravo i zbog toga, razvija se potreba za lokalizacijom koja je suprotna globalizaciji engleskog jezika. Države i jezični instituti ulažu napore u očuvanje pojedinih jezika, ali ih i 'unapređuju' u vidu modernizacije u sklopu novih tehnologija. Budući da ne postoji nekakav univerzalni jezik, javlja se potreba za razvojem alata koji će nam omogućiti pristup informacijama neovisno na kojem one jeziku bile izvorno napisane ili izrečene. Upravo se tu otvara prostor za razvoj jezičnih tehnologija, preciznije, alata za strojno prevođenje, koji bi trebali prevladati te prepreke. Osnovni problem strojnog prevođenja jest priroda samoga jezika i njegova kompleksnost. Prijevodi iz izvornog jezika u ciljni nisu uvijek jednostavni i logični, stoga je potrebno tekstove prvotno označiti, i identificirati jezične kategorije. Stroj ne može imati uvid u neke izvanjezične komponente kao što su namjera ili povijesni/kulturni kontekst, što pak može vidno utjecati na kvalitetu samog prijevoda. Dvosmislene i višesmislene jezične jedinice najveći su neprijatelj sustava za prevođenje.

Potrebno je istaknuti da ne postoje savršeni sustavi za automatsko prevođenje i oni su u biti još uvijek nedostignuti cilj. Postoje sustavi za potpuno automatsko prevođenje; međutim, oni

nisu savršeni i omogućuju nam prevođenje teksta čisto na informativnoj razini kako bismo uvidjeli o čemu se o nekom tekstu radi, ali ne da bismo taj isti tekst i dali objaviti. Za to nam je potrebna ljudska intervencija, tj. editiranje i lektura teksta. U nekim slučajevima moguće je uz pomoć prethodnog uređivanja izvornog teksta postići bolji konačni rezultat automatskog prijevoda. Jedno je sigurno: razvoj jezičnih tehnologija nastavlja se, kako u akademskom, tako i u komercijalnom smislu i stručnjaci su u konstantnoj potrazi za alatima koji bi olakšali sam postupak prevođenja i prijenosa informacija.

## 5. Pristupi i metode u strojnom prevođenju

Sustavi za strojno prevođenje mogu se klasificirati na različite načine koje je potrebno pobliže objasniti kako bismo stekli uvid u osnovne koncepte i terminologiju vezanu za temu ovog rada. S obzirom na broj jezika koje obrađuju, sustave za automatsko strojno prevođenje moguće je podijeliti na dvojezične i višejezične. Dvojezični sustavi namijenjeni su prevođenju s jednog jezika na neki drugi ciljni jezik dok višejezični sustavi mogu prevoditi s više različitih jezika na veći broj jezika. S obzirom na smjer prevođenja, sustavi mogu biti jednosmjerni i višesmjerni. Dvojezični su sustavi najčešće jednosmjerni dok su višejezični sustavi često dvosmjerni. Ako uzmemo u obzir stupanj automatizacije onda sustave možemo podijeliti na (Hutchins, prema Mitkov, 2003:501):

- Sustave za automatsko strojno prevođenje uz ljudsku pomoć
- Sustave za strojno potpomognuto prevođenje

Kao što smo već spomenuli, sustavi za automatsko strojno prevođenje nisu savršeni i omogućuju prijevode na osnovnoj razini prenošenja informacije i, ako se radi o nekoj složenijoj domeni, konačan rezultat može biti jasan samo stručnjaku u tome području. S druge strane, takvi sustavi mogu proizvoditi i kvalitetom bolje prijevode ako ih prevoditelji uređuju i interveniraju, koristeći svoje jezično znanje i znanje o izvanjezičnim činjenicama. Sustavi za strojno potpomognuto prevođenje zapravo su računalni alati koji prevoditelju omogućuju brže i lakše prevođenje nudeći mu digitalne rječnike, terminološku konzistenciju i prijevodne memorije na jednome mjestu, takoreći sva dostignuća računalne tehnologije na polju jezika koja uvelike olakšavaju posao. Ti alati poznatiji su kao CAT (*Computer-Aided Translation*) alati i kombinirani u jednu platformu tvore tzv. *workstation* ili *translator workbench*.

### *5.1. Izravni sustavi*

Prema generalnoj strategiji i metodologiji sustave za prevođenje možemo podijeliti na izravne sustave i neizravne sustave. Izravni sustavi temelje se na zamjeni riječi izvornoga jezika riječima ciljnog jezika, s minimalnom sintaktičkom i semantičkom intervencijom. Takvi sustavi kao pomagalo koriste prvenstveno rječnike izvornoga jezika i ekvivalentne prijevode ciljnog jezika, tj. dvojezične rječnike. Kvaliteta ovakvih prijevoda, koji su ujedno i najraniji primjeri računalnih prijevoda, ovisi o vrsti teksta koja se prevodi, pa će jednostavniji tekstovi jasno definiranog i limitiranog vokabulara i stila biti lakše prevodivi. Također nije zanemarivo o kojem se izvornome jeziku radi i na koji se prevodi, jer ukoliko se radi o jezicima koji nisu toliko bliski, tj. srodni, potrebno je više pozadinskog lingvističkog znanja za koje onda izravni sustavi jednostavno neće biti dovoljni.

### *5.2. Neizravni sustavi*

Tu dolazimo do druge vrste sustava, a to su neizravni sustavi, koji se potom dijele na one temeljene na međujeziku i sustave transfera. U procesu prevođenja kod sustava koji se baziraju na transferu, sam proces obuhvaća tri faze. U prvoj fazi analize *parserom* se analizira izvorni tekst i segmentira na sintaktički i semantički prepoznatljive strukture. U drugoj fazi dolazi do transfera u kojem se pronalaze leksički ekvivalenti ciljanog jezika uz pomoć dvojezičnih rječnika. Također se odvija strukturalno prilagođavanje na temelju transformacijskih pravila koja opisuju relaciju dvaju jezika kako bi se riječi uklopile na ispravan način. Na kraju, u procesu generiranja novih struktura u ciljnom jeziku na temelju transfera dolazi se do prijevoda. S druge strane, sistemi koji koriste međujezik polaze od pretpostavke da je moguće prikazati jezik posrednik koji se temelji na generalnim pravilima koja su zajednička velikom broju jezika, no nikako nisu univerzalna (Hutchins, prema Mitkov, 2003:503). Taj bi jezik posrednik trebao moći prikazati bilo koji jezik i u isto vrijeme biti dostatan i jasan kako bi generirao prijevod na neki drugi ciljani jezik. Njegova ekonomičnost ujedno je i njegova prednost pred sustavom za prijenos. Dakle, kod neizravnih sustava s međujezikom, jezik posrednik je taj međustupanj koji je potreban da bi se generirao prijevod, za razliku od sustava transfera koji taj stupanj nema i kojemu su potrebni dvojezični rječnici i usporedne gramatike kao pomoć pri slaganju i prijevodu. Dodatna prednost sustava koji

koriste međujezik je orijentiranost u svakoj fazi na pojedini jezik i zbog toga im nije potrebno opširno znanje iz transformacije iz jednog jezika u drugi (Hutchins, prema Mitkov, 2003:505).

### *5.3. Sustavi temeljeni na pravilima*

Sustave također možemo podijeliti na one koji se temelje na pravilima i one koji se temelje na podatcima. Prevoditeljski sustavi koje smo spomenuli dosad (izravni, neizravni (transfer i međujezik)) spadaju u sustave koji se temelje na pravilima. Povijesno gledano, sustavi temeljeni na pravilima ujedno su i prvi koji su se počeli koristiti i razvijati i njihova je složenost rasla dodavanjem novih razina lingvističke analize, od izravnih do neizravnih. Da bi jedan takav sustav proizveo kvalitetan prijevod potreban je bogat izvor alata u obliku *parsera* i *chunkera* i resursa u obliku rječnika i gramatika iz kojih bi on crpio potrebne informacije. Također, za još preciznije prijevode potrebni su i alati za semantičku analizu koji rješavaju najveći problem kod prevođenja, a to je višeznačnost, unutar jednog jezika ili između nekog para jezika. Vrhunac sustava koji se temelje na pravilima su sustavi koji koriste tehnike umjetne inteligencije da bi se postignuli kompleksni sistemi koji se temelje na znanju (*knowledge-based interlingua systems*) (Hutchins, prema Mitkov, 2003:508).

### *5.4. Empirijski, kvantitativni sustavi*

Sustavi koji se temelje na empirijskoj, kvantitativnoj paradigmi, oprečni su sustavima koji se temelje na pravilima, i upravo su oni otvorili put razvoju alata o kojima će biti riječ u ovome radu. Njihov je razvoj započeo 90-ih godina zahvaljujući sve većoj dostupnosti sirovih podataka u obliku paralelnih korpusa iz kojih se crpe podatci, tj. već ovjereni prijevodi. Sustav funkcionira po principu pretraživanja već zabilježenih primjera ili po principu bilježenja statističke vjerojatnosti da će neka rečenica biti prevedena drugom rečenicom u ciljnome jeziku, te na temelju toga, generira prijevod. Potonji pristup ne ovisi o pravilima pojedinog jezika i usporednim gramatikama, već o statističkoj vjerojatnosti prevođenja. Mnogi će stoga nazvati ovaj pristup više matematičkim nego lingvističkim, no on se pokazao iznimno uspješnim i primjenjivim kod profesionalnih prevoditelja, ali i ostalih ljudi kojima je često potrebno 'instant' prevođenje kako bi dobili opći dojam o čemu se radi u nekom tekstu ili samo kako bi prenijeli određenu poruku. Imajući u dosegu dvije verzije jednog te istog teksta, sustav može uspoređivati rečenicu po rečenicu i utvrđivati koji par rečenica ima najveću vjerojatnost da se zamjenjuje međusobno, tj. da

se rečenica ulaznog teksta najčešće prevodi rečenicom u tekstu koji već postoji pohranjen i s kojim ga uspoređujemo. Sustav ne koristi znanje za generiranje prijevoda, već nudi prijevod samo na temelju onoga što je zabilježio kao ovjereni prijevod.

Empirijski sustavi koji se temelje na primjerima koriste već gotove prijevode koji su pohranjeni u prijevodnim memorijama, a sam proces prevođenja sastoji se od triju faza: pronalaženja prijevodnog ekvivalenta *(matching)*, sravnjivanja dijelova koji su odgovarajući *(alignment)*, te kombiniranja *(recombination)*, kojim se generira novi prijevod. Ovakvi sustavi koji cijeli proces odrađuju sami ipak ovise o kvaliteti drugih sustava koji se temelje na pravilima jer ulazni tekst treba biti parsiran na temelju jednakih pravila kao i onaj koji je već pohranjen, a u konačnom rezultatu, zbog specifičnosti jezika koji su u kombinaciji, potrebno je konzultirati gramatike. S druge strane, moguće je tekstualnom zapisu pristupiti kao čistom zapisu slijeda znakova, no takav pristup u konačnici otežava postupak sravnjivanja i kombiniranja (Somers, prema Mitkov, 2003:514).

## 6. Prijevodne memorije

Razvojem računalnih sustava za prevođenje ipak nije nestala potreba za profesionalnim prevoditeljima. Paralelno se tako razvijaju računalni jezični alati, tzv. CAT (*Computer-Aided Translation*) alati koji prevoditeljima ubrzavaju i olakšavaju prevođenje tekstova repetitivne strukture i sadržaja. Nova dostignuća u razvoju alata za prevođenje oslanjaju se na široku dostupnost dvojezičnih paralelnih korpusa i izradu prijevodnih memorija u koje su pohranjeni već prevedeni parovi tekstova. Tekstovi koje sadrže prijevodne memorije segmentirani su na rečenice i paragrafe i međusobno sravnjeni vodeći se različitim smjernicama o čemu će naknadno biti više riječi. Osim za prijevodne memorije, paralelni korpusi pokazali su se iznimno korisnima za izradu terminoloških baza te za pretraživanje konkordancija u tekstu. Iako je sravnjivanje rečenica također jedna od faza prethodno spomenutih sustava koji se baziraju na primjeru, njihov je cilj u konačnici ponuditi gotov prijevod dok kod rada s prijevodnim memorijama upravo prevoditelji imaju posljednju riječ i odlučuju koje će prijedloge prihvatiti odnosno odbaciti. Što je najvažnije, samim time kontroliraju memoriju koja je produkt sravnjivanja korpusa, a mogu i pravovremeno riješiti situacije u kojima se javlja problem višeznačnosti. Prijevodne su memorije ušteda vremena

i novca, vidno potrebna automatizacija zbog sve većeg obujma posla, ali s ljudskim faktorom koji u konačnici editira i garantira kvalitetu prijevoda.

Potrebno je, međutim, napomenuti da je korištenje prijevodnih memorija primjereno i najefikasnije kod tekstova koji slijede određenu strukturu koja se ponavlja i u kojima se javlja vokabular koji je specifičan za neku određenu domenu te je u pravilu vrlo jasan i nedvosmislen. To su najčešće tekstovi tehničkih priručnika koji se objavljuju iz godine u godinu i čiji se sadržaj nerijetko minimalno razlikuje od prethodnika. Može se raditi o različitim nadogradnjama nekog sustava ili komercijalnog proizvoda za koje je potrebno izdati novi priručnik koji se temelji na već napisanoj prethodnoj verziji i bilo bi zaista gubljenje vremena kada bi se morao ponovo prevoditi po principu riječ po riječ i onaj dio koji je prevoditelju već poznat od prije. Korištenjem prijevodnih memorija također se osigurava terminološka dosljednost koja je bitna u takvim vrstama tekstova.

### *6.1. Segmentacija i sravnjivanje prijevodnih memorija*

Kako bismo dobili prijevodne memorije, potrebno je sravniti dokumente izvornog i ciljanog jezika, a to je moguće obaviti ručno i automatski. Ručno sravnjivanje dokumenata je proces koji podrazumijeva spajanje segmenata izvornog teksta sa segmentima sadržajno istog teksta prevedenog na drugi jezik, te stvaranje arhiva tih paralelnih korpusa tj. prijevodnih memorija. Proces se odvija u dva koraka:

- Segmentacija originalnih tekstova i prijevoda
- Povezivanje segmenata originalnog teksta s odgovarajućim segmentima prijevoda

Originalni tekst i prijevod se u postupku segmentacije dijele na manje fragmente na temelju jednakih pravila o segmentaciji čime dobivamo jasan uvid gdje završava jedan segment, a počinje drugi. Najjasnije pravilo koje možemo primijeniti u tom postupku jest da točka nakon koje slijedi razmak i veliko početno slovo označava granicu između segmenata. Ovo pravilo bez problema može segmentirati sljedeći tekst:

*Danas smo vježbali glagole. Sutra ćemo vježbati pridjeve.*

Implementacijom pravila, dobili bismo:

*Danas smo vježbali glagole.*

*Sutra ćemo vježbati pridjeve.*

Ipak, na problem bismo mogli naići u slučaju drugog primjera:

*Dr. Polić nije bio na zadnjem sastanku.*

Ako primijenimo isto pravilo, dobivamo:

*Dr.*

*Polić nije bio na zadnjem sastanku.*

Najveći broj sustava za potpomognuto prevođenje nudi mogućnost određivanja pravila pri segmentaciji što uvelike olakšava sam proces i smanjuje ovakve pogreške. Kako bismo iskoristili puni potencijal prijevodnih memorija, tekstove je potrebno segmentirati na temelju jednakih pravila jer se njihova struktura može razlikovati. Sukladno s tim razvijen je TMX (*Translation Memory eXchange*), te poslije poboljšani SRX (*Segmentation Rule eXchange*) standardni format za segmentaciju i označavanje koji se temelji na XML-u. XML se pokazao idealnim jezikom za označavanje tj. kodiranje dokumenata jer je jasan i lako čitljiv kako računalima, tako i ljudima.

Ukoliko su originali i prijevod u istom formatu te sličnih interpunkcijskih znakova, a njihovi segmenti u odnosu 1:1 (jedan segment originalnog teksta odgovara jednom segmentu prijevoda) proces sravnjivanja u pravilu je visoke preciznosti i nije potrebna veća intervencija prevoditelja. Međutim, to nije uvijek slučaj. Često se događa da se jedan segment originalnog teksta prevodi dvama segmentima (odnos 1:2) ili obrnuto, dva se segmenta prevode jednim segmentom u ciljnome tekstu (2:1). Također se može desiti da neki segment originalnog teksta u prijevodu jednostavno nedostaje (1:0) ili se u prijevodu stvaraju novi segmenti (0:1). Ručno sravnjivanje ponekad može biti zahtjevno te iziskuje temeljitu ljudsku kontrolu, tj. višestruku reviziju samog postupka. Kao posljedica toga razvile su se nove metodologije i alati za automatsko sravnjivanje dokumenata o kojima ćemo nešto više reći u nastavku.

### *6.2. Automatsko sravnjivanje*

Ručno sravnjivanje može biti učinkovito onda kada se original i prijevod ne razlikuju mnogo u formatu, paragrafima i broju rečenica, međutim može i zahtijevati mnogo vremena ukoliko to nije slučaj. Danas postoje algoritmi koji omogućuju automatsko sravnjivanje rečenica i

koji slijede standardnu proceduru segmentacije i sravnjivanja teksta. Automatsko sravnjivanje odvija se bez intervencije korisnika, a možemo razlikovati dva metodologijska pristupa automatskom sravnjivanju (Brkić et al. 2009:355):

- Na temelju dužine segmenata (rečenica ili znakova)
- Na temelju dvojezičnog rječnika
- Hibridni pristupi

Prva metodologija kreće od pretpostavke da se duži segmenti originalnog teksta u načelu prevode dužim segmentima i u ciljanom, tj. prevedenom tekstu. Nakon segmentacije teksta izračunavaju se statistički parametri koji se temelje na dužini segmenata, a najuspješnija je ona segmentacija nakon koje dobivamo ujednačenu distribuciju odnosa dužine originalnog teksta i prijevoda. Ovu su strategiju prvi primijenili Brown (1993) te Gale i Church (1993) gotovo u isto vrijeme. Druga se metodologija razvila na temelju znanja o značenju i prijevodu jedne riječi drugom riječju u ciljanom jeziku. Ako se pojedina riječ javlja u originalnom tekstu, za pretpostaviti je da će se ista riječ morati pojaviti i u prijevodu i da će je sustav pokušati pronaći i povezati segmente na temelju znanja iz rječnika. Strategiju koja se temelji na rječniku prvi su predložili Wu (1994) i Chen (1993) i pokazala se iznimno korisnom kod sravnjivanja tekstova jezika koji nisu toliko srodni, a gdje strategija koja se temelji na dužini segmenata može pokazati svoje nedostatke. (Yu et al., 2012:2)

Naravno, kombinacija obiju strategija daje najbolje rezultate, stoga možemo izdvojiti najuspješnije hibridne modele kao što su Moorova metoda te Brauneova i Fraserova metoda. Moore (2002) jer razvio hibridnu strategiju koja objedinjuje metodu koja se temelji na dužini segmenata i metodu koja se temelji na rječniku, tj. podudarnosti riječi u segmentima. Zanimljivost ove strategije jest to da ona zapravo ne zahtijeva dvojezični rječnik kao resurs na temelju kojeg se spajaju segmenti, već se sam proces odvija u dvama koracima. Na prvome stupnju odvija se automatsko sravnjivanje segmenata na temelju njihove dužine, a sustav bilježi samo one parove za koje je vjerojatnost podudarnosti vrlo visoka. Nakon prvog koraka, koji je ujedno i preduvjet za uspješnost drugog koraka, sustav automatski izrađuje vlastiti rječnik na temelju provjerenih parova, metodom statističke podudarnosti i na temelju njega sravnjuje i one segmente koji nakon prvog koraka nisu pokazali određenu razinu podudarnosti da bi bili prihvaćeni. Brauneova i Fraserova metoda zapravo je modificirana verzija Moorove metode koja drugi korak dijeli na dva stupnja. Nakon što su uz pomoć podudarnosti riječi sravnjeni segmenti koji se nalaze u odnosu 1-

1 u drugom stupnju pronalaze se segmenti koji nisu sravnjeni te se oni pridodaju već sravnjenima i tako se dobivaju kombinacije 1 – više segmenata ili više segmenata – 1. (Abdul-Rauf et al. 2012:3).

## 7. *Acquis Communautaire*

Korpus koji ćemo koristiti pri analizi alata za sravnjivanje rečenica je *Acquis Communautaire*[2], zbirka tekstova pravne stečevine Europske unije. Zakonske odredbe koje ovaj korpus obuhvaća vrijede na razini cijele Europske unije i čine njenu pravnu bit. AC uključuje izvorne norme sadržane u osnivačkim ugovorima te njihove izmjene, izvedene zakonske odredbe, deklaracije i rezolucije europskih organizacija i međunarodne ugovore potpisane od strane Unije. No, pravna stečevina nadilazi standarde i obuhvaća političke ciljeve Unije. Zemlje kandidati za ulazak u Europsku uniju moraju prihvatiti pravnu stečevinu prije pristupanja. Na taj način, kada zemlja ulazi, preuzima i integrira sve pravne odrednice u svoje nacionalno zakonodavstvo i mora ih primjenjivati od trenutka njenog punopravnog članstva. Jedan od uvjeta ulaska je također i prevođenje pravne stečevine na jezik države kandidata. AC se neprestano nadopunjuje novim tekstovima, a sve je počelo 50-ih godina prošlog stoljeća.

Istraživački centar za znanost i tehnologiju Europske komisije JRC (*Joint Research Center*) zaslužan je za stvaranje danas najvećeg paralelnog korpusa koji se sastoji od 22 jezika i koji je javno dostupan za pretraživanje i korištenje u svrhu istraživanja; riječ je o JRC-Acquis paralelnome korpusu. Kao dio Europske komisije, JRC funkcionira kao referentni centar za znanost i tehnologiju u Uniji i njegova pružanje personalizirane znanstvene i tehničke podrške za razvoj, provedbu i praćenje politike EU. Jedini paralelni korpus koji se može mjeriti s JRC-Aquis korpusom je Europarl, koji sadrži transkribirane tekstove rasprava Europskog parlamenta na 21 jeziku Europske unije. Treba napomenuti da JRC-Acquis ne sadrži sve dokumente zbirke AC, već se pri njegovom sastavljanju vodilo računa da se obuhvate dokumenti koji su dostupni na najmanje 10 službenih jezika od tadašnjih 21, te najmanje 3 jezika država članica koje su se priključile Europskoj uniji 2004. godine.

[2] https://ec.europa.eu/jrc/en

JRC-Aquis sastoji se od gotovo 19.000 dokumenata po jeziku, s prosječnom veličinom od 8 do 13 milijuna riječi po jeziku. Većina je tekstova ručno klasificirana prema EUROVOC domenama, tako da korpus može biti koristan i za pretraživanje prema ključnim riječima. JRC-Acquis kodiran je u XML-u u skladu sa smjernicama za kodiranje teksta i dostupan je u HTML obliku na stranicama Europske komisije (https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis). Zbog velikog broja paralelnih tekstova na mnogim jezicima, JRC-Acquis posebno je prikladan za obavljanje svih vrsta *cross-language* istraživanja, kao i za testiranje sustava za automatsko sravnjivanje rečenica, ekstrakciju termina i klasifikaciju dokumenata (Erjavec et al., 2005:531).

Sastavljanje tako opsežnog paralelnog korpusu zahtijeva pomnu razradu koraka koji mu prethode. Prvi je korak klasifikacija dokumenata pravne stečevine (AC). Dokumenti koji su uzeti u obzir za izgradnju korpusa klasificirani su jedinstvenim CELEX kodom u kojem prvi broj označava vrstu dokumenta, sljedeća četiri godinu nastajanja dokumenta, te slova i još četiri znamenke koje kodiraju dokument. Bitno je naglasiti da verzije prijevoda svakog dokumenta imaju jednak CELEX broj. Osim tematskog određenja dokumenta, a radi lakše klasifikacije, potrebno je i urediti i samu građu dokumenta, tj. njegov tekst kako bi ga bilo moguće računalno što lakše i unificiranije obraditi. Tekstovi sa službenih stranica transformirani su iz HTML formata u XML format, kodiran prema UTF-8 zapisu. Jezik svakog teksta određen je obradom uz pomoć softvera koji identificira jezik analizom n-grama. Dijelovi teksta rastavljeni su na paragrafe čime je olakšano povezivanje prijevodnih ekvivalenata na različitim jezicima. Kako su pravni tekstovi jasno strukturirani, paragrafe je zapravo lako odijeliti i oni se u različitim jezicima uglavnom ne razlikuju. Jedinstveni dijelovi svakog pravnog dokumenta također su tijelo dokumenta, popisi i aneksi. Ti su dijelovi jasno odijeljeni kako bi se korisnicima omogućilo što lakše pretraživanje, npr. popisi imena, mjesta, datumi i referencije mogu biti korisni za testiranje sustava za prepoznavanje imena. Sravnjivanje paragrafa dokumenata na različitim jezicima rađeno je koristeći dva alata koji se temelje na različitim modelima, Vanilla aligner koji za sravnjivanje koristi Gale & Church metodu i HunAlign koji koristi kombiniranu metodu rječnika i Gale & Church metode. Ukoliko rječnik i ne postoji, HunAlign ima mogućnost korištenja najprije sravnjivanja na temelju dužine rečenica, te zatim izgradnje vlastitog rječnika uz pomoću kojeg vrši kontrolu. Vanilla aligner koristi čisto statističku metodu sravnjivanja uzimajući u obzir dužinu rečenice/paragrafa i nudi odnose među paragrafima, 1-1, 1-2, 2-1, 1-0, 0-1. Procijenjeno je da je

85% paragrafa JRC-Acquis kolekcije sravnjeno po principu 1-1 što odgovara postotku od 89% na testiranjima Gale & Church metode. Prednost HunAlign metode jest i to što omogućava razdvajanje jedne rečenice u više od dvije rečenice (Steinberger, et al., 2006:4).

# 8. Analiza alata za sravnjivanje rečenica

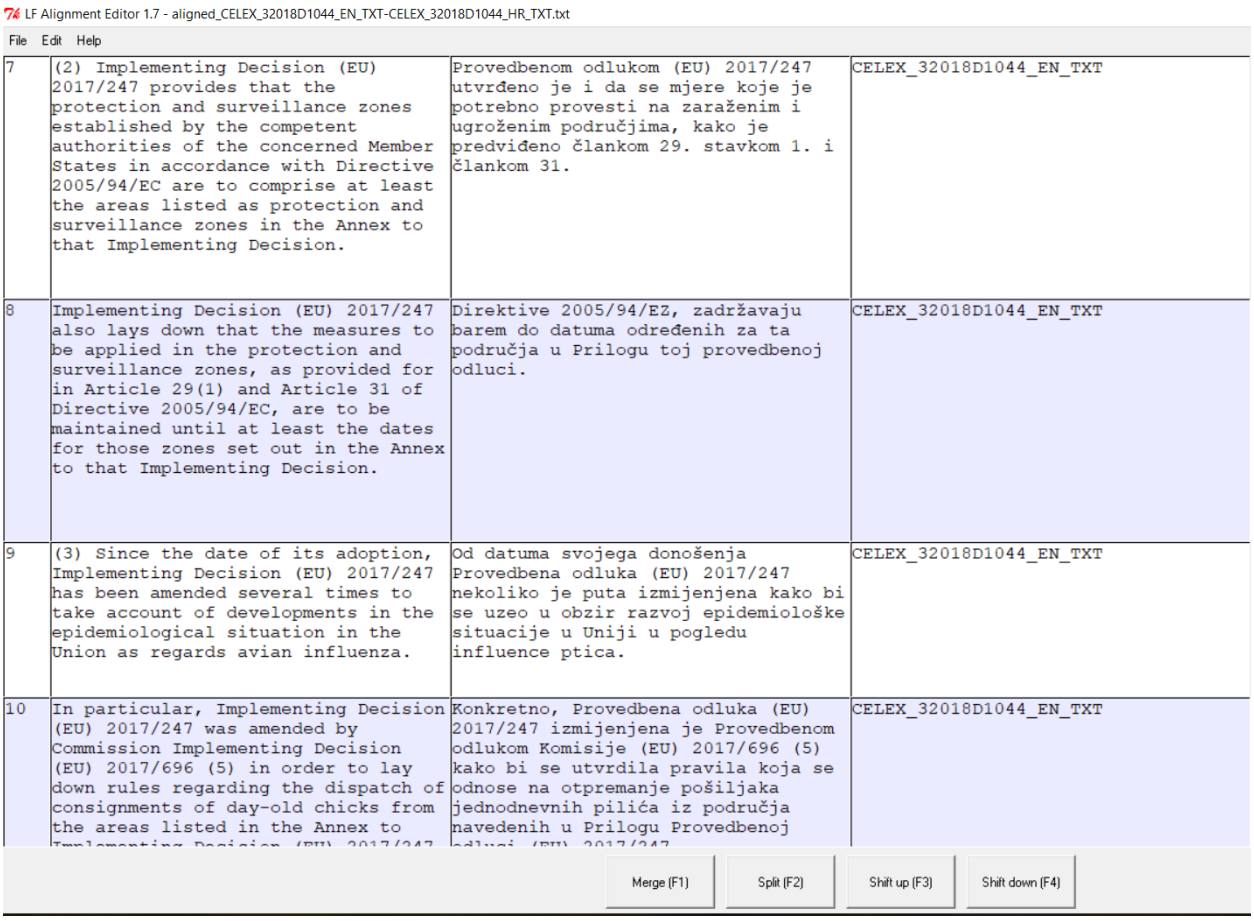
Cilj ovog rada je uspoređivanje različitih alata za sravnjivanje rečenica. Alati koje ćemo usporediti su LF aligner, Bitext2tmx i Coral. Razlog zašto smo se odlučili analizirati baš te tri aplikacije jest to što se temelje na različitim algoritmima i nude različita sučelja za korisnike. Sva tri programa su *open source* programi što znači da su dostupni za besplatno skidanje i slobodno korištenje. Zanimljivost programa Coral krije se u činjenici što je razvijen u Hrvatskoj, na Fakultetu elektrotehnike i računarstva. Na kraju ovog rada pokušat ćemo sažeto prikazati prednosti i nedostatke svakog navedenog programa.

Tekstovi koji su korišteni prilikom analize navedenih alata preuzeti su sa stranice https://eur-lex.europa.eu/homepage.html koja sadrži zakonodavne odredbe Europske unije. Na stranici Eur -Lex moguće je pretraživati tekstove pomoću CELEX broja ili jedinstvenog broja dokumenta. Također, moguće je pretraživati i registar zakonodavstva prema određenim područjima kao što su poljoprivreda, ribarstvo, oporezivanje itd. Ovaj je pristup posebno koristan ukoliko ne tražimo neki određeni dokument, već nas zanima pregled dokumenata u registru. Svaki od dokumenata moguće je otvoriti u trima formatima: HTML, PDF i u formi službenog lista. Dokumenti su, ukoliko su prevedeni, dostupni na sva 24 službena jezika Europske unije.

## *8.1. LF aligner*

Među open source programima za automatsko sravnjivanje rečenica ističe se LF aligner čiji je autor Mađar, Farkas András. LF aligner bazira se na HunAlign algoritmu što znači da koristi višesmjeran pristup sravnjivanju rečenica, koristeći rječnik ali i dužinu segmenata. Formati dokumenata koji se mogu koristiti u ovom alatu su txt, doc, docx, rtf, pdf, html, dok je kasnije moguće kao *output* generirati txt, TMX i xls ekstenzije. Ono što je zanimljivo kod ovog alata jest da ga nije potrebno instalirati, dovoljno je samo dvostrukim klikom otvoriti ikonu programa koji smo skinuli i raspakirali, tj. napravili dekompresiju. Nakon što smo otvorili program, slijedimo njegove jednostavne upute. U prvome koraku program nas pita koji tip dokumenta želimo otvoriti.

Osim gore navedenih formata, LF aligner nudi i obradu tekstova dostupnih putem URL adrese te automatsko skidanje zakonodavnih tekstova Europske unije putem CELEX broja ili broja dokumenta i godine izdanja u slučaju izvješća Europskog parlamenta. Moguće je izabrati više jezičnih kombinacija za istovremenu analizu, njih čak 99. Mi smo se odlučili za testiranje upotrijebiti engleski original i hrvatski prijevod istog teksta. S Eur-Lexa skinuli smo dokumente u PDF formatu koje LF aligner preporučuje konvertirati u txt kako bi se uklonio sadržaj zaglavlja i podnožja koji može biti problematičan kod segmentacije. Ovakav pristup nije se pokazao dobar i odmah se nakon konverzije u txt format mogla primijetiti razlika u broju segmenata na engleskom, odnosno hrvatskom jeziku. Tekst na engleskome jeziku u samome startu ima više segmenata koji se na kraju približno izjednače, no s dosta odstupanja. Najčešće se greške javljaju kod nabrajanja stavki u nekom zakonu gdje se iz nekog razloga u hrvatskoj verziji teksta brojevi u zagradama, npr. (1), (2) ispisuju nakon teksta i ne čine vezanu cjelinu s rečenicom koja slijedi nakon. U ovom je slučaju moguće prije samog procesa segmentacije intervenirati u txt generirani tekst i to ručno ispraviti pa se zapravo i tome može doskočiti, no to oduzima mnogo vremena i nama je cilj usporediti uspješnost i jednostavnost automatske segmentacije i sravnjivanja. Algoritam na kojemu je temeljen LF aligner uzima u obzir i određena jezična pravila kao što je pravilo da je novu rečenicu lako prepoznati ukoliko nakon točke dolazi veliko slovo. U nekim slučajevima to može biti problem, kao što je slučaj s rednim brojevima u hrvatskom jeziku. Primjer možemo vidjeti na Slici 1. gdje je prikazano kako su se segmenti odlomili jer je nakon rednog broja slijedila riječ Direktiva koja mora biti napisana velikim slovom.




Slika 1.

Pregledavanjem sravnjenih segmenata u programu moguće je spojiti segmente 7 i 8 koji će onda odgovarati engleskom segmentu po principu 1-1. Ukoliko pak odlučimo sravniti pravne tekstove EU koje program automatski skida ako unesemo CELEX oznaku, program radi gotovo bez pogreške na objema razinama segmentacije koje program nudi: na razini paragrafa i na razini rečenice. Do grešaka ponovo može doći kod cijepanja rečenica u slučaju kad nakon rednog broja slijedi riječ koja počinje velikim slovom. Ovaj je problem moguće riješiti stvarajući popis iznimki koje su specifične za svaki jezik i u kojem slučaju ne mora izričito doći do počinjanja nove rečenice. U engleskom jeziku najbolji su primjer kratice Mr. ili Mrs., dok su kod nas to redni brojevi. Readme.txt dokument koji se nalazi u instalacijskom folderu sadrži upute za implementaciju pravila (liste) u kod programa.

Kada smo pregledali grafički prikaz sravnjenih segmenata, možemo ga prebaciti u XCL format gdje možemo ispraviti greške koje su se možda pojavile. LF aligner radi u pozadini dok

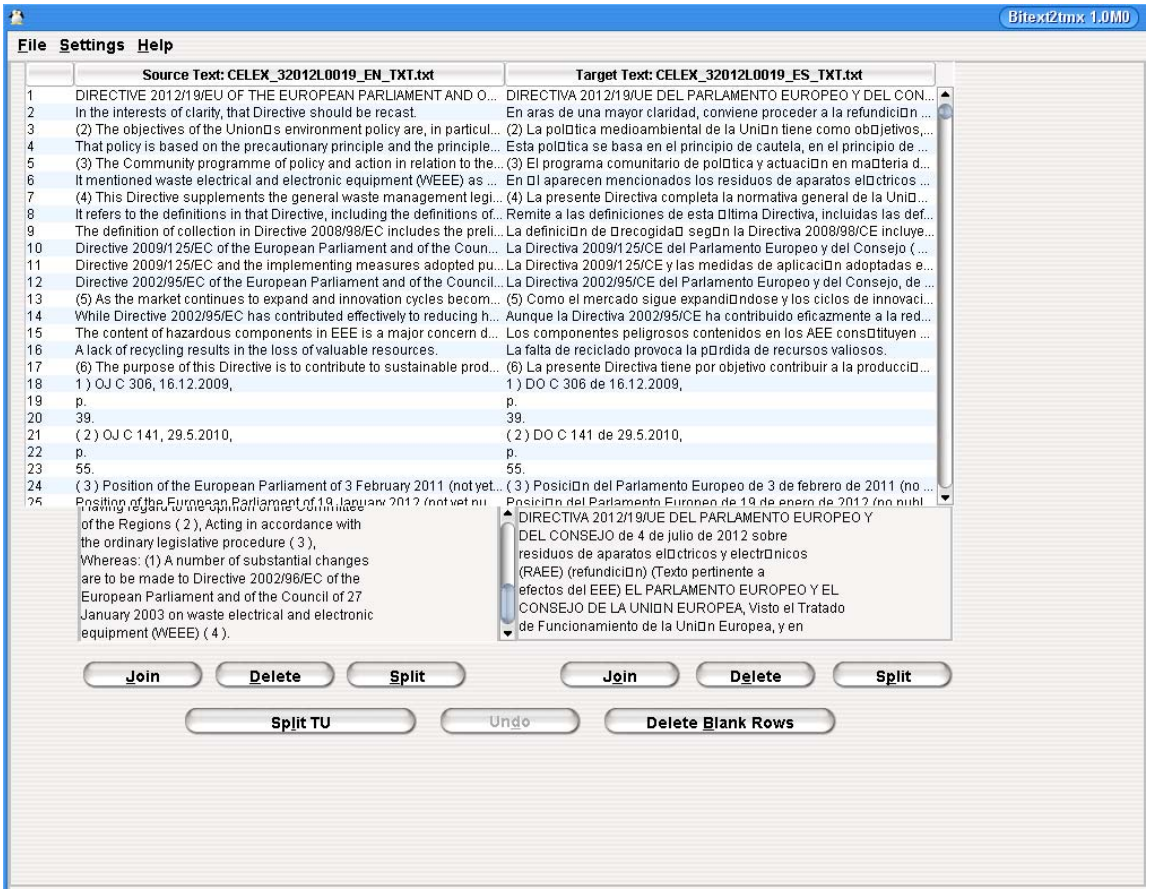
ispravljamo greške i naposljetku nas pita želimo li naš projekt spremiti u TMX formatu koji se poslije može koristiti kao prijevodna memorija u nekom od CAT alata (Computer-Assisted Translation), što je i krajnji cilj računalnog sravnjivanja dvaju identičnih tekstova na različitim jezicima.

Sučelje LF alignera vrlo je jednostavno za korištenje i nudi četiri komande: MERGE (F1), SPLIT (F2), SHIFT UP (F3) i SHIFT DOWN (F4). SPLIT i SHIFT UP su zapravo suprotne komande tako da SPLIT spaja označeni segment s donjim segmentom i lijepi ga na njegov početak dok SHIFT UP spaja označeni segment s gornjim segmentom i lijepi ga na kraj. MERGE lijepi označeni segment za onaj iznad njega i ostavlja prazno mjesto dok SHIFT DOWN premješta segment dolje i ostavlja prazno mjesto. Uz pomoć MERGE i SHIFT DOWN komande moguće se kretati i premještati segmente gore-dolje. Ako nam ostane praznih segmenata, uvijek ih je moguće ukloniti uz pomoć F5 komande (ostale prečace možemo vidjeti u izborniku Help na izbornoj traci). Komande nam se mogu učiniti zbunjujućim i potrebno je neko vrijeme dok se ne naviknemo na upravljanje u grafičkom pregledniku.

### 8.2. Bitext2tmx

Druga aplikacija koju smo odlučili analizirati je Bitext2tmx. Radi se o programu za automatsko sravnjivanje tekstova razvijenom u programskom jeziku Javi. Njegovi autori su Susana Santos i Sergio Ortiz-Rojas sa Sveučilišta u Alicanteu. Program je besplatan i može se preuzeti na sljedećem linku: *https://sourceforge.net/projects/bitext2tmx/*. Za razliku od LF alignera Bitext2tmx kao input može uzeti samo txt format dokumenta i po završetku sravnjivanja nam nudi tmx output koji je moguće iskoristiti za izgradnju prijevodnih memorija i integrirati u neki u neki od sustava za računalno prevođenje kao što su Trados ili OmegaT.

I u ovom slučaju koristili smo pravne tekstove s Eur-Lex stranice koje smo morali konvertirati u txt format budući da nisu dostupni u tom obliku. Kako Bitext2tmx nije razvijen za hrvatski jezik, odlučili smo se za jezični par engleski – španjolski. Za razliku od LF alignera koji nudi 99 jezika koje je moguće upariti, Bitext2tmx nudi 21 jezik i analizu samo dva jezična para istovremeno. Aplikacija se otvara također dvostrukim klikom na ikonu programa i na početku je potrebno odabrati jezični par koji će se procesirati. Ono što na prvi pogled uočavamo je staromodnost sučelja programa i njegove grafike.



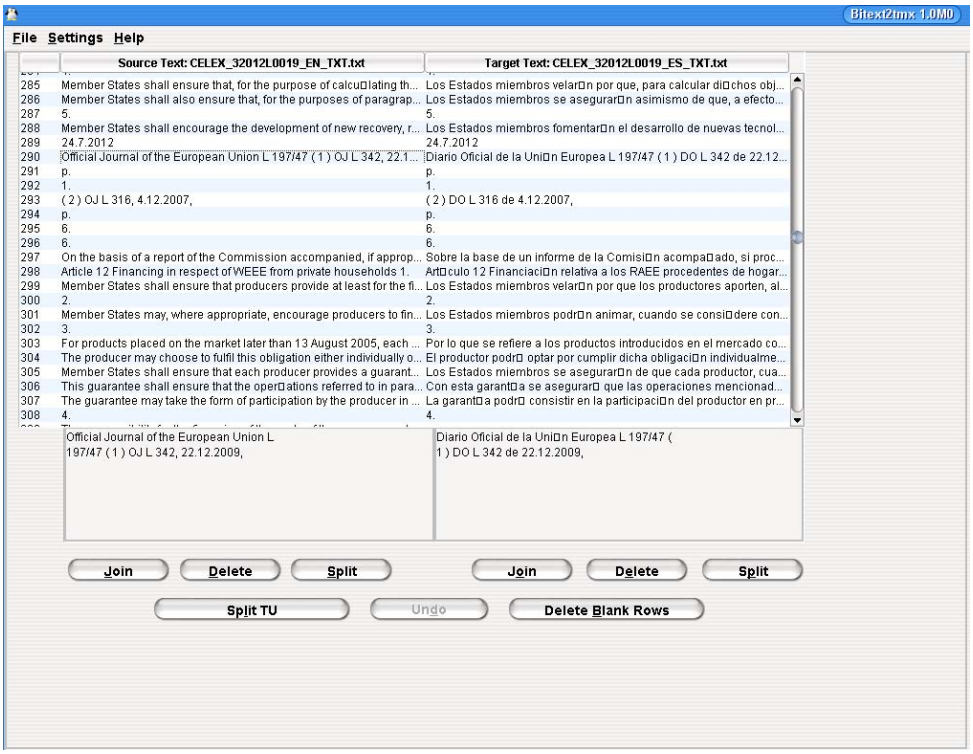

Slika 2.

Takav *demode* izgled ne čudi s obzirom na to da se radi o programu koji je nastao 2008. godine i od tada se nije značajno unaprjeđivao i nadograđivao. Zašto samo onda uzeli ovaj program u obzir? Zato što je besplatan i zato što je jedan od najpopularnijih programa za automatsko sravnjivanje koji se nude. Ukoliko krajnji korisnici nisu velike prevoditeljske kompanije koje si mogu priuštiti Trados, svakako je zanimljivo vidjeti koji su to drugi programi koji su nam na raspolaganju ako sami želimo graditi svoje prijevodne memorije na jednostavan način i poslije ih upotrijebiti kako bismo si olakšali posao prevođenja. Tmx dokumente koje dobivamo nakon sravnjivanja možemo koristiti u OmegaT programu koji je besplatan, za razliku od Trados SDL studija. Ono što nas zanima jest koji je od alata za sravnjivanje najbolji i najjednostavniji za upotrebu.

Prije analize, također je preporučljivo očistiti sadržaj zaglavlja i podnožja kako ne bi došlo do njihovog miješanja u ostatak teksta jer čisti tekst u txt formatu tretira fusnote kao da su dio rečenice koja je dio teksta i nalazi se iznad fusnote, a ne kao zaseban dio. Nakon čišćenja zaglavlja i podnožja, rezultati su bolji, no tu se LF aligner pokazao boljim jer u slučaju tekstova skinutih s

Eur-Lexa on bez problema sravnjuje i taj dio koji nam je, na kraju krajeva, također bitno prevesti i ne možemo ga izbaciti iz teksta. Pravni tekstovi sami po sebi sadrže velik broj fusnota koje nude objašnjenja i ukoliko se služimo programom koji će to odraditi automatski, uštedjeli smo još više na vremenu. Razlog zašto se LF aligner bolje snalazi kod sravnjivanja fusnota jest to što on nudi automatsko skidanje CELEX dokumenata koji su već sravnjivani putem istog programa, a i cijeli je ACQUIS sravnjivan uz pomoć HunAligna. Potrebno je ponovo istaknuti da se ni LF aligner ne snalazi najbolje s fusnotama kada pdf format pretvaramo u txt format i potrebno je napraviti preinake u txt tekstu ukoliko želimo dobiti optimalne rezultate.

Ako pogledamo na prikaz sravnjivanja na Slici 3., možemo uočiti kako je program u redu 290 spojio zaglavlje i podnožje u isti red i onda je došlo do razdvajanja segmenata zbog pojavljivanja točki nakon kratice za stranicu i za broj, a u oba jezika ona postoji. Da bismo stvarno vidjeli što se tu dogodilo, moramo otvoriti tekstove u izvornom obliku kako ne bismo pogriješili prilikom sređivanja tog 'nereda'. Kod oznake (1) u redu 290 zapravo počinje podnožje i moramo ga razdvojiti. Specifičnost ovog programa je ta što to ne možemo napraviti odmah u retku kao u LF alignеru već nam se kad kliknemo na redak ispod teksta otvori novi prozorčić u kojem nam se taj segment detaljno prikaže i možemo označiti na kojem ga mjestu želimo prelomiti, u ovom slučaju prije oznake (1) i kliknuti na komandu SPLIT s lijeve strane izbornika, isto ponovimo s desne strane.




Slika 3.

Bitext2tmx je na ovaj način pokušao riješiti prikazivanje dugih segmenata da bi uštedio na prostoru. To bi bilo u redu da i samo sučelje nije zbijeno i slova presitna. Čak i kad u gornjem desnom kutu kliknemo na uvećanje, grafički prikaz ostaje iste veličine, samo se povećava količina sivih rubova što je vrlo nepregledno. Za spajanje segmenata dovoljno je označiti jedan segment i donji će se segment sjediniti s njim. Na kraju uređivanja preporučuje se očistiti prazna polja, koja ne služe ničemu već zauzimaju mjesta u memoriji. Projekt je lako spremiti i sačuvati u tmx formatu za izgradnju prijevodnih memorija. Potrebno je napomenuti da i ovaj program ima boljku razdvajanja u novi segment poslije svake točke iza koje slijedi veliko slovo, ali za razliku od LF alignera, ne nudi mogućnost nadogradnje programa pravilima za iznimne slučajeve. U slučaju teksta na španjolskome, Bitext2tmx ne ispisuje naglašena slova i vjerojatno nailazi na problem s definicijom UTF-8 koda.

### 8.3. Coral

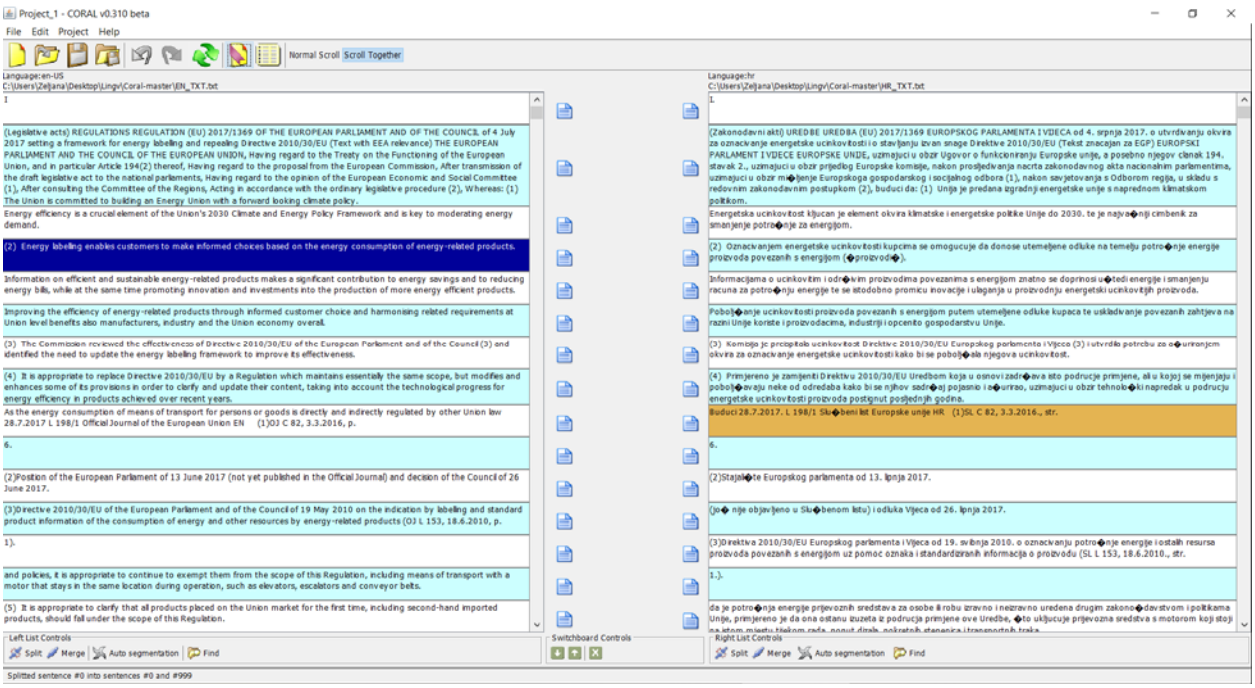
Treći program za sravnjivanje paralelnih tekstova koji ćemo proučiti i usporediti s ostalima jest Coral (CORpus ALigner). On nam je posebno zanimljiv budući da je razvijen na našem

Sveučilištu u suradnji Fakulteta elektrotehnike i računarstva i Filozofskoga fakulteta (Odsjeka za informacijske i komunikacijske znanosti i Odsjeka za lingvistiku) 2008. godine. Uzimajući u obzir glavne aktere na ovom projektu, jasno je vidljiva međuovisnost računarstva i lingvistike pri stvaranju jezičnih alata za računalno korištenje. Coral je pisan u Java programskom jeziku i može se koristiti na bilo kojem operacijskom sustavu, naravno, uz instaliran Java Runtime Environment. Moguće je ručno sravnjivanje te računalno, automatsko sravnjivanje temeljeno na Gale & Church metodi. Coral je kao i prethodni programi koje smo analizirali, besplatan i lako ga je instalirati. Rađen je po uzoru na neke druge, već postojeće alate, kako bi se pružilo optimalno rješenje.

Kao input Coral može koristiti txt i xml format, pa smo za testiranje ponovno koristili dokumente s Eur-Lexa koje smo iz pdf formata konvertirali u txt format. Nakon pokretanja programa, prvi je korak učitavanje dokumenata. Kako bismo odmah doskočili problemu razdvajanja rečenica na dvije rečenice onda kada one to nisu, kod učitavanja tekstova možemo kliknuti na opciju *Advanced* i napisati koje iznimke program može ignorirati. Ipak, potrebno je naglasiti da ova opcija vrijedi jedino ako je format učitanog teksta u xml-u. Kako je Coral je rađen za potrebe sravnjivanja i testiranja paralelnih tekstova na engleskome i hrvatskom jeziku, jedino je i moguće izabrati tu kombinaciju jezika.

Ukoliko tekst već nije odijeljen na rečenice, kod prikaza paralelnih tekstova dobit ćemo prilično dugačke dijelove teksta, no ako kliknemo na opciju *Auto segmentation*, dobit ćemo uspješno segmentirane tekstove. Budući da smo učitali tekstove u txt formatu, nismo mogli unijeti iznimke koje mogu utjecati na uspješnost razdvajanja pa tu onda ima dodatnog posla i pregledavanja. Kao i u posljednja dva primjera, potrebno je očistiti tekst od zaglavlja i podnožja prije učitavanja. Međutim, ono što bismo preporučili jest učitati ih ipak sa zaglavljem i podnožjem i jednostavno desnim klikom nakon segmentacije označiti segment i kliknuti na *Delete* element. Razlog zbog kojeg preporučujemo takav pristup jest taj što je u txt tekst povezan u jednu cjelinu i puno je teže razlučiti segmente, tj. rečenice bez konstantnog pregledavanja originalnog dokumenta u pdf formatu. Nakon segmentacije u Coralu puno ćemo brže uočiti dijelove koji nam stvaraju problem te ih ukloniti.

Na prvi pogled, program se vidno razlikuje od LF alignera i Bitext2tmx programa zbog „bogatije“ alatne trake i mogućnosti koje nudi (Slika 4.). Neka vizualna rješenja podsjećaju na Tradosov WinAlign koji je, čini se, bio uzor za grafičko korisničko sučelje (GUI).

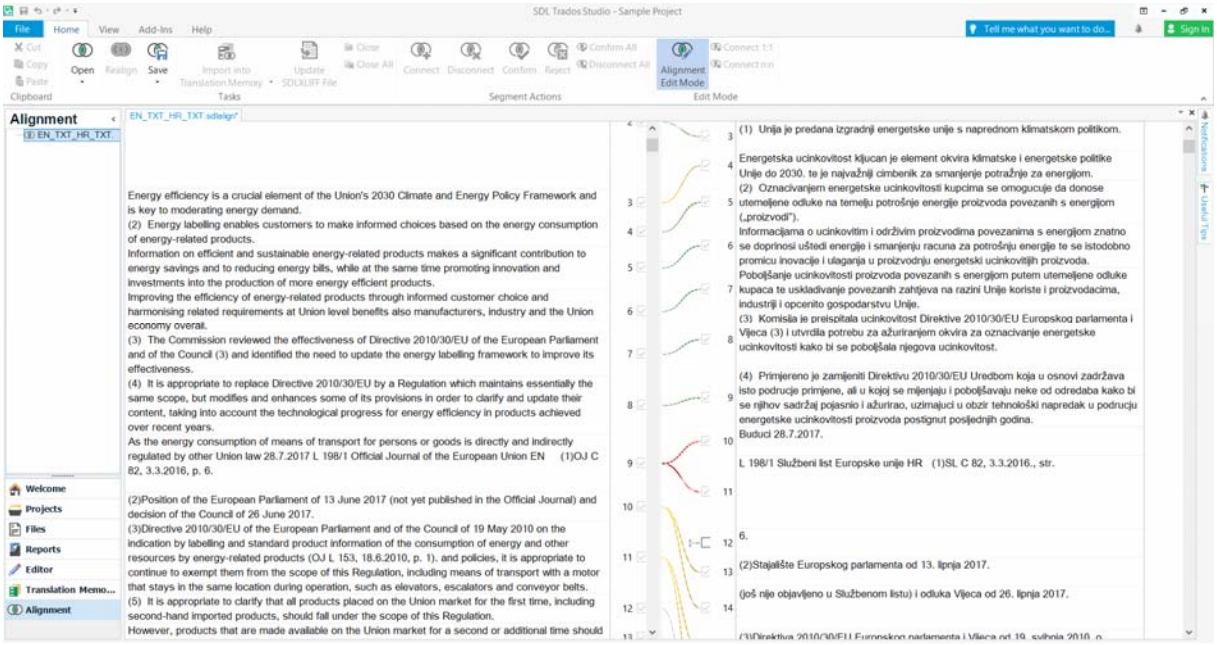


Slika 4.

Program nam nudi mogućnost ručnog i automatskog sravnjivanja uz pomoć Gale & Curch metode. Za razliku od drugih programa, nakon osnovne segmentacije ne pretpostavlja da su i segmenti upareni. Dodatni prostor u sredini, između dvaju tekstova, olakšava nam ručno ovjeravanje ispravnih parova povezivanjem suprotnih ikonica u obliku dokumenta. U slučaju da se slažemo s automatskom segmentacijom koju je izvršio program, klikom na *Project* možemo povezati sve segmente u odnosu 1:1 kako bismo izbjegli uparivanje jedan po jedan. Kako to najčešće nije slučaj, bolje će nam poslužiti ako aktiviramo Gale & Church algoritam u istom izborniku. Nakon obrade, dobit ćemo uparene segmente po različitom principu, 1:1, 1:2 ili 2:1. Ukoliko želimo dobiti odnose 1:1, jednostavno možemo odvajati ili spajati segmente koristeći komande *Split* i *Merge*. S opcijom *Split* dvostrukim klikom jednostavno uđemo u segment i označimo mjesto gdje ga želimo razdvojiti, a s uporabom opcije *Merge* nema zabune jer, kao i u paketu Microsoft Office, uz pomoć tipke ctrl + klik označimo segmente koje želimo ujediniti. Potrebno je detaljno pregledati dokument zbog iznimki pri odvajanju rečenica. Odnos segmenata koji odmah možemo provjeriti ako pogledamo statistiku može nam puno otkriti o rezultatu segmentiranja i povezivanja prijevodnih parova. Opcija provjere statistike još je jedna od korisnih značajki koje nudi Coral, a koju druge dvije aplikacije nemaju. Nakon završenog projekta, lako ga je prenijeti u tmx format za daljnje korištenje.

## 8.4. SDLTrados

Jedan drugačiji alat za sravnjivanje rečenica koji je potrebno spomenuti jest *translation alignment tool* (nekadašnji WinAlign) koji dolazi u paketu SDL Trados Studio. SDL Trados je najpoznatiji komercijalni program za računalno potpomognuto prevođenje koji sadrži niz alata koji prevoditeljima olakšavaju proces prevođenja. Zapravo se radi o prevoditeljskoj radnoj stanici koja nudi sve potrebno na dohvat ruke. Temelji se na tehnologiji za izradu baze podataka prijevodnih memorija te također omogućuje izgradnju terminoloških baza. Translation alignment tool njegov je sastavni alat za sravnjivanje prijevoda. Alat za sravnjivanje nije besplatan i može se kupiti samo u paketu s Tradosom, no dostupna je probna verzija koja se može koristiti 30 dana i nedavno je izašla najnovija, Studio 2019 (Slika 5.).




Slika 5.

Novi Tradosov alat za sravnjivanje izgledom podsjeća na stariji Winalign, a ujedno i na Coral, no uvedene su neke vizualne i funkcionalne promjene. Nakon što smo uz pomoć algoritma spojili segmente, njihove su veze označene različitim bojama kao što je i vidljivo iz Slike 5. Različite boje različito vrednuju kvalitetu veze između sravnjenih segmenata, zelena označava najveću kvalitetu veze, a ujedno je i vjerojatnost točnosti, žuta označava srednju vjerojatnost, dok je crvena veza loša i sigurno je treba prepraviti. Ono što je malo neobično za ovu verziju

Tradosovog alata za sravnjivanje jest to što ne nudi naredbe *Split* i *Merge*. Tekstovi se mogu kopirati i lijepiti u susjedne segmente, no to je ponešto nespretno jer nam tako ostaju prazna polja koja je zatim potrebno ukloniti te si zapravo stvaramo još više posla i veću nepreglednost. Samo u *Alignment Edit Mode* izborniku možemo utjecati na veze i to označavanjem kućica. U Trados Studio možemo učitati prethodno generirane tmx datoteke u formi prijevodne memorije, no zanimljivo je da Tradosov alat za sravnjivanje nudi jedino .sdalign ekstenziju u kojoj se može spremiti prijevodna memorija. Kao takva neće biti kompatibilna s drugim programima za računalno prevođenje, ali će biti maksimalno prilagođena radu u Tradosu.

## 9. Prednosti i nedostatci

Pregledom gore navedenih programa za sravnjivanje paralelnih tekstova uvidjeli smo različite pristupe i rješenja koja su smišljena s ciljem da pomognu prevoditelju u njegovom poslu. Iako su neki programi poput Bitext2tmx i Corala već zastarjeli, što vizualno, što zbog toga što se više ne nadograđuju, oni su i dalje funkcionalni, jednostavni i što je bitno, dostupni svima na korištenje. LF aligner je vizualno također jednostavan, međutim upravo je ta jednostavnost, ako je odrađena dobro, ono što je poželjno kod ovakvih programa. LF aligner, Bitext2tmx i Coral su *open source* programi što znači da su dostupni za javno korištenje dok je Trados komercijalni program. I jedan i drugi tip programa imaju svoje prednosti i nedostatke. Najveća prednost *open source* programa jest prije svega to što je besplatan. Krajnjim korisnicima to omogućuje da isprobaju program i vide odgovara li im bez obavezivanja, što je bitno ako ga ne koriste toliko često da bi im se isplatilo kupiti neki komercijalni program. Najčešći primjeri te vrste korisnika su prevoditelji koji tek počinju prevoditi i studenti. Softveri otvorenog koda (*open source*) također omogućuju spremanje koda i interveniranje u njegov zapis, čime je korisniku dana sloboda prilagodbe programa i njegovoga poboljšavanja, naravno, ukoliko korisnik ima potrebna znanja iz programiranja. S druge strane, slobodni softveri imaju i svoje nedostatke, a to su prije svega napuštanje razvoja aplikacije od strane njegovog pokretača zbog nedostatka sredstava. U tom će se slučaju korisnik vjerojatno opet naći u potrazi za idealnim rješenjem. Zbog velikog broja korisnika, a ujedno i zbog toga što pokretaču aplikacije to vjerojatno nije jedini posao, kao ni izvor prihoda, velika je vjerojatnost da tehnička podrška neće uvijek biti dostupna. Brojni su programi uglavnom ponuđeni na engleskom jeziku, što je i dalje problem za one koji ga ne koriste, ali redovito prevode tekstove različitih jezika. Ako bismo birali između programa otvorenog koda,

onda bi izbor bili Coral i LF aligner. Coral iz razloga što nudi mnogo više opcija od preostalih dvaju programa, te preglednost i jednostavnost korištenja. Također, moguće pogreške u sravnjivanju mogu se odmah definirati na početku. Ono što je mana Corala, a nudi LF aligner, jest mnoštvo jezičnih kombinacija i vrsta dokumenata koje prihvaća. Taj softver je također više puta nadograđivan od njegove prve verzije koja je izašla. Korisničko sučelje Corala je pak najbolje prilagođeno korisnicima (*user friendly*). Tradosov alat za sravnjivanje rečenica, kad je riječ o upravljanju vezama među segmentima, pokazao se najtežim za korištenje, no na njihov novi pristup jednostavno se treba naviknuti. Između ostalog, kao i LF aligner, on prihvaća analizu dokumenata u pdf formatu gdje su se pokazali puno bolji rezultati nego kod sravnjivanja txt datoteka.

## 10. Zaključak

Ne možemo poreći promjene, ponekad i drastične, koje je sa sobom donijela globalizacija interneta i razvoj informacijskih znanosti. Te promjene uvelike su utjecale na način života ljudi kao i na njihov dosadašnji način obavljanja posla. Prevoditeljski se posao promijenio i klasični rječnici i enciklopedije u papirnatome obliku sve se češće zamjenjuju ekvivalentima u elektronskom obliku. Sam proces prevođenja ubrzan je i olakšan zbog pristupa već prevedenim tekstovima iz iste domene. Danas je neupitno i od velike važnosti imati sve izvore i alate na jednom mjestu kako bismo bili što efikasniji i kako bismo dobili što bolji prijevod. U ovome smo tekstu prikazali kako od jezičnih resursa u formi korpusa i rječnika, te alata za analizu tekstova na različitim lingvističkim razinama, dolazi do stvaranja konačnog proizvoda koji je namijenjen širokome tržištu i koji je uvelike unaprijedio jednu ljudsku djelatnost kao što je prevođenje. To je ujedno i konkretan primjer suradnje lingvistike i računarstva, tj. primjene novih tehnologija u unapređivanju ljudskog rada. Prikazom različitih rješenja za sravnjivanje paralelnih tekstova pokušali smo dati bolji uvid u neke od najpopularnijih programa koji nam se nude te navesti njihove prednosti i nedostatke. Pokazalo se da *open source* programi mogu parirati nekim 'većim' komercijalnim programima kao što je Trados, te da je hrvatski Coral najbolji od besplatnih programa koje smo analizirali i jedina mu je realna mana ta što nije dostupan i za druge kombinacije jezika. Coral je sažeo najbolje od drugih alata i ponudio zaista kvalitetan proizvod. Kao što smo već napomenuli i ranije, a kao i kod razvoja drugih programa, ne samo na području lingvistike, bitna je povratna informacija korisnika kao i njihovo mišljenje o tome što je dobro, što

loše i što bi se još moglo unaprijediti. Od digitalizacije korpusa i prvih paralelnih korpusa, ljudi su uočili njihovu vrijednost i iskoristivost za izgradnju jezičnih aplikacija odnosno komercijalnih proizvoda kojima budućnost predviđa još bolji napredak i korak bliže prema za sad nedostižnom potpuno automatskom prevođenju. Ljudska je intervencija još uvijek potrebna, ali novi su alati uvelike olakšali posao. Automatsko prevođenje neće zamijeniti ljudsko prevođenje, niti mu predstavlja prijetnju; upravo suprotno, omogućuje još veći napredak. Prevoditeljima je danas bitno da idu u korak s vremenom i da budu upoznati s novim alatima koji se nude jer su oni postali neizbježni u njihovome poslu. Pristup istraživanju lingvistike promijenio se zahvaljujući razvoju digitalnih tehnologija. Neovisno radi li se o teorijskome, računalnom ili empirijskom pristupu u istraživanju, neupitno je da će se isti u budućnosti voditi analizom podataka i da se stvara jedno interdisciplinarno polje koje otvara niz mogućnosti.

Nastavlja se istraživanje na akademskom planu, ali i u tvrtkama koje su se specijalizirale za razvoj programa za prevođenje. Vjerujemo da napredak računalnih tehnologija, lingvističke teorije i umjetne inteligencije vodi do konstantne potrage za novim alatima koji će pojedincima olakšati posao i da je smjer istraživanja koji se neće napustiti u budućnosti.

**Literatura**


Abdul-Rauf, S., Fishel, M., Lambert, P., Noubours, S., Sennrich, R. (2012). *Extrinsic evaluation of sentence alignment systems. Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*. Istanbul, 27 May 2012 - 27 May 2012, 6-10. URL: https://www.zora.uzh.ch/id/eprint/62565/ (pristupljeno 15. siječnja 2018)

Brkić, M., Seljan, S., Mikulić, B. (2009). Using Translation Memory to Speed up Translation Process. *International Conference The Future of Information Sciences,* 353-363.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. (2005). *Massive multi lingual corpus compilation: Acquis Communautaire and totale. Archives of Control Sciences,* 4, 529-540. URL: http://nl.ijs.si/janes/wp-content/uploads/2014/09/erjavecothers05.pdf (pristupljeno 7. lipnja 2018)

Garside, R., Leech, G., McEnery, T. (2013). *Corpus Annotation. Linguistic Information from Computer Text Corpora.* London, New York: Routlege

Glovacki Bernardi, Z. (2007). *Uvod u lingvistiku*. Zagreb: Školska knjiga

Lüdeling, A., Kytö, M. (2008). *Corpus Linguistics. An International Handbook.* Berlin, New York: Walter de Gruyter

Mitkov, R. (2004). *The Oxford Handbook of Computational Linguistics.* London: Oxford University Press

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). *The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages.* URL: https://arxiv.org/ftp/cs/papers/0609/0609058.pdf (pristupljeno 7. lipnja 2018)

Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex Libris

Yu, Q., Max, A., Yvon, F. (2012). *Revisiting sentence alignment algorithms for alignment visualization and evaluation.* University Paris Sud.

URL: https://perso.limsi.fr/yvon/publications/sources/Yu12revisiting.pdf (pristupljeno 3. kolovoza 2018)