TARTU ÜLIKOOL

FILOSOOFIATEADUSKOND

Eesti ja üldkeeleteaduse instituut

Kaarel Veskis

PARALLEELKORPUSED ARVUTILINGVISTIKAS: LEKSIKONIDE GENEREERIMINE JA KORPUSTE VÕRDLEMINE

Magistritöö

Juhendaja dr. Heiki-Jaan Kaalep

Tartu 2007

SISUKORD

Sissejuhatus

1990-ndate aastate algus tähistab mitmeid märkimisväärseid saavutusi paralleeltekstide vaheliste vastavuste automaatse tuvastamise osas (nt Brown jt 1991; Gale, Church 1993). Edasiste aastate jooksul on esile kerkinud suur hulk paralleelistamisega seotud probleeme, kuid ka palju huvitavaid lahendusi nendele probleemidele. Samuti on tekkinud teadlikkus mitmesugustest uutest võimalustest, mida paralleelkujul keelekorpused võivad tähendada erinevate loomuliku keelega seotud eluvaldkondade jaoks.

Magistritöö üheks eesmärgiks on olla esimeseks eestikeelseks sissejuhatuseks paralleekorpuste temaatika tähtsamatesse aspektidesse ja ühtlasi anda ülevaade Tartu Ülikooli üldkeeleteaduse õppetoolis toimuva paralleelkorpuste-alase töö hetkeseisust. Teiseks eesmärgiks on anda panus Eestis toimuva leksikograafiatöö arendamisse, asetades erilist rõhku paralleelkorpustes peituva leksikaalse info automaatse esiletoomise võimaluste tutvustamisele praktiliste näidete kaudu. Töö praktilise osa tulemuseks olnud leksikonid võivad kasutust leida masintõlkerakendustes või olemasolevate leksikonide täiendamist hõlbustavate vahenditena.

Magistritöö jaguneb kuueks peatükiks. Esimesed kolm peatükki ja neljanda peatüki esimesed alaosad kujutavad endast ülevaatlikku sissejuhatust paralleelkorpuste temaatikasse. Neljanda peatüki osas 4.5 ja viiendas peatükis on ülekaalus praktilise eesti keelt puudutava paralleelkorpuste-alase töö kirjeldus.

Esimene peatükk annab ülevaate paralleelkorpuste koostamisest ja kasutamisest. Teine peatükk keskendub paralleelkorpuste peamisele rakendusvaldkonnale – korpusepõhisele masintõlkele ja tõlkeabisüsteemidele. Kolmas peatükk käsitleb paralleelkorpuste kasutus-võimalusi lingvistilises uurimistöös ja keeleõppes.

Neljas peatükk[1] tutvustab üht esialgu suhteliselt marginaalset, kuid samas perspektiivikat paralleelkorpuste rakendusvaldkonda – kakskeelsete leksikonide genereerimist. Peatükk kirjeldab ka esimest teadaolevat katset genereerida inglise-eesti leksikon automaatselt inglise-eesti paralleelkorpustest ja katse tulemuste analüüsi. Püüan siin lisaks osutada erinevatele alternatiivsetele võimalustele kasutada paralleelkorpuseid leksikograafiatöös ning kirjeldada eeldusi, millest saab lähtuda ühe keelena eesti keelt sisaldavate leksikonide genereerimiseks sobiva tarkvara loomisel tulevikus.

Viiendas peatükis[2] on kirjeldatud paralleelkorpuste võrdlemise ja hindamise alast praktilist tööd, mis kujutab endast loomulikku jätku leksikonide genereerimise teemale ning annab teisalt mitmeid vastuseid küsimustele, mida tõstatasid eesti-inglise statistilise masintõlke (Fishel jt 2007) esimesed tulemused.

Samas on tegemist uudse lähenemisega paralleelkorpuste paralleelistuskvaliteedi hindamisele – erialasest kirjandusest ei ole teada analoogseid katseid hinnata paralleelistuse kvaliteeti paralleelkorpuste võrdlemise teel poolautomaatselt.

Tahaksin tänada oma juhendajat Heiki Kaalepit, kes töö valmimisele väga olulisel määral kaasa aitas.

[1] Magistritöö neljas peatükk põhineb Eesti Rakenduslingvistika Ühingu 2007. aasta aastaraamatus (Veskis 2007) ilmunud artiklil.
[2] Viies peatükk rajaneb koos Heiki-Jaan Kaalepiga kirjutatud ja 2007. a septembris Kopenhaagenis toimuvale XI masintõlke-alasele tippnõupidamisele (MT Summit) stendiettekandena avaldamiseks esitatud artiklil.

1. Taust

1.1. Paralleelkorpused

Paralleelkorpus on korpus, mis sisaldab mingit teksti originaalkeeles ja selle tõlget teise keelde või tõlkeid teistesse keeltese. Paralleelkorpuse paralleeltekstid võivad olla ka mõne kolmanda korpusse mittekuuluva teksti tõlked. Paralleelkorpuste kasutamiseks on neid korpuseid vaja eelnevalt rohkem töödelda kui tavalisi ükskeelseid tekstikorpuseid – kahe paralleelse teksti märgendus peab olema omavahel seotud.

Tuntuimaks paralleelkorpuseks on peetud Kanada *Hansardit*[3]. See korpus koosneb Kanada parlamendidebattidest, mida avaldatakse riigi kahes ametlikus keeles – inglise ja prantsuse keeles. Väiksemate korpuste hulgast võib näiteks tuua 1 miljoni sõnalise inglise-sloveeni paralleelkorpuse[4]. Sellistest paralleelkorpustest, kus üheks keeleks on eesti keel, tuleks mainida europrojekti Multext-East[5] raames valminud paralleelkorpust, mis sisaldab George Orwelli romaani „1984" kaheksas keeles. Selle lausetasandil paralleelistatud korpuse maht on 75 000 sõnet.

On koostatud ka seaduste ja õigusaktide tekste ning tõlkeid sisaldavad inglise-eesti ja eesti-inglise paralleelkorpus (TÜPK) ja väiksemahulisemate ettevõtmistena eesti-rootsi, eesti-norra ja vene-eesti paralleelkorpused. Nii keeletehnoloogia kui ka kontrastiivne lingvistika vajab aga tingimata vähemalt paarkümmend miljonit sõna sisaldavat suurt eestikeelse osalusega paralleelkorpust; tõsise masintõlke-alase töö jaoks läheks vaja 100

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20
[4] http://nl.ijs.si/elan
[5] http://nl.ijs.si/ME/

miljoni sõna suurust paralleelkorpust (Muischnek jt 2003:25). Praegu sellises mahus paralleelkorpusi, milles oleks esindatud ka eesti keel, veel ei ole.

1.2. Paralleelistamine

1.2.1. Probleemid

Probleeme tekitab paralleelistamisel keelte erinev morfoloofiline tüpoloogia (aglutinatsioon vs fleksioon või analüütilisus) ja eesti ja indo-euroopa keelte lause-ehituslikud erinevused, näiteks eesti liitsõna väljendatakse inglise keeles tihti hoopis mitmesõnalise fraasiga.

Ometi on mujal maailmas isegi inglise ja aglutinatiivsete keelte sõnatasandil paralleelistamisel teatud edu saavutatud. Seejuures ei piisa tegelikult sõnatasandist, vaid aglutinatiivsete keelte eripärast johtuvalt tuleb inglise keele sõnadega või suuremate üksustega tihti vastavusse asetada ka aglutinatiivse (aglutineeriv-flekteeriva) keele morfeeme ja sõnaosi (Martin jt 2003).

Nimetatud lahknevus inglise ja eesti keele liitsõnade vahel esineb aga sarnaselt inglise-eesti keelepaariga ka inglise ning germaani keelte vahel. Seda lahknevust on arvestatud inglise-rootsi (Ahrenberg jt 2000a) sõnatasandi paralleelistaja loomisel – kuna tegemist on keelest sõltumatu paralleelistajaga, siis võib selle inglise-rootsi keelepaariga tehtud töö võtta üheks aluseks inglise-eesti joondamisprogrammi väljatöötamisel.

Korrektne paralleelistamine on paralleelkorpuse hilisema kasutatavuse seisukohalt kõige olulisem ja ka üksnes indo-euroopa keeli hõlmava korpuse puhul kõige töömahukam probleem. Paralleelistamist raskendab lisaks keeltevahelistele erinevustele asjaolu, et tihti sisaldavad paralleeltekstid "müra", st ühes tekstis on midagi rohkem või vähem kui temale vastavas teises tekstis, mistõttu ei saa lauseid omavahel üksüheselt kokku viia.

1.2.2. Meetodid

Paralleelistamisel ehk joondamisel eristatakse statistilisi ja lingvistilisi meetodeid, kusjuures statistilisi meetodeid peetakse tõhusamaks suuremate korpuste ja lingvistilisi meetodeid väiksemate joondamisel (Oakes, McEnery 1998). Potentsiaalselt kõige edukamaks peetakse siiski statistiliste meetodite kombineerimist lingvistiliste meetoditega ja lisaressursside (sõnastikud) kasutamisega. Statistiliste meetodite populaarsus on viimastel aastatel oluliselt tõusnud ka üldisemalt keeletehnoloogias ja arvutilingvistikas seoses infotehnoloogia kiire arengu ja järjest laiema levikuga.

Tartu Ülikooli inglise-eesti ja eesti-inglise paralleelkorpuse paralleelistamiseks kasutati Gale'i ja Churchi algoritmile toetuvat statistilist Vanilla paralleelistajat[6].

Gale'i ja Churchi algoritmil põhineb ka programmi ParaConc[7] poolautomaatne joondamisfunktsioon. Programmi ParaConc on Vanilla kõrval peetud üheks levinumaks, „vaikimisi" kasutatavaks paralleelistamisvahendiks. (Rosen 2005)

Gale'i ja Churchi algoritm eeldab, et mõlemad paralleelistatavad tekstid koosnevad võrdsest hulgast mingil kindlal moel eristatud terviktekstist väiksematest osadest, mis on juba algselt paralleelsed. Teiseks eeldab algoritm, et tõlketeksti laused on samas järjestuses kui lähteteksti laused. Kolmandaks eeldab algoritm, et tõlke pikkus sõltub originaali pikkusest. (TÜPK)

Katset on tehtud ka inglise- ja hiinakeelsete tekstide joondamisega (McEnery jt 2000). Gale'i ja Churchi algoritmi rakendamine inglisekeelsete ja ladina tähestikku transliteereritud (Pinyin) hiinakeelsete tekstide joondamiseks kukkus läbi, kuna mõlema keele vastava lause tähemärkide arvu vahel puudub sellisel juhul piisav korrelatsioon. Rööpsus on aga märksa suurem, kui tähemärkide arvu asemel võetakse arvesse sõnade arvu, sellest on ka ühe võimaliku paralleelistamismeetodi väljatöötamisel lähtutud (McEnery jt 2000). (Mõningast edu on samuti saavutatud kakskeelse leksikoni ekstraheerimisel inglise-hiina paralleelkorpusest (Fung 2000)).

[6] nl.ijs.si/telri/Vanilla
[7] http://www.athel.com/para.html

Ülevaate Gale'i ja Churchi algoritmi kasutavast ja teistest tekstide erinevatel tasemetel paralleelistamise meetoditest ja nendega seonduvatest probleemidest annab näiteks Jörg Tiedemann (Tiedemann 1997).

Lausetasandil paralleelistamise meetodid saab üldjoontes jagada kas lausete pikkusest, sõnade vastavusest või sõnade sarnasusest lähtuvateks ning hübriidmeetoditeks. (Singh, Husain 2005)

Lausete pikkusest lähtuvad meetodid põhinevad oletusel, et lähte- ja sihtkeele lausete pikkused on omavahel seotud. Lausete pikkuse tunnusena on kasutatud nii sõnade arvu (Brown jt 1991) kui ka (Gale, Church 1991) tähemärkide arvu. Mõlemad lausete pikkusest lähtuvad meetodid eeldavad, et enne automaatset lausetasandil paralleelistamist on tekstid juba lõigutasandil paralleelsed.

Sõnade vastavusest lähtuvad meetodid (nt Kay 1991) rajanevad eeldusel, et sõna ja tema tõlge teises keeles paiknevad paralleelsetes tekstides positsioonilistes vastavustes. Esialgsest sõnatasandi paralleelistusest tuletab iteratiivne algoritm tõenäolise lauseparalleelistuse, mida omakorda kasutatakse sõnatasandi täielikumaks joondamiseks.

Sõnavastavuse meetodeid arendati edasi IBMi mudel 1 raames[8] statistilise masintõlke eesmärkidel. Hübriidmeetodiks võib nimetada näiteks Robert C. Moore'i algoritmi (Moore 2002), mis toimib kahes järgus. Esimene järk baseerub lausete pikkustel ja teine IBMi 1. mudelil.

Erinevaid paralleelistusmeetodeid võrreldes on jõutud järeldusele, et paralleelistuse kvaliteet sõltub peamiselt sisendtekstide mitmesugustest omadustest: tekstidevahelistest erinevustest näiteks tabelite, jooniste või viidete osas, tekstitüübist (kas tegu on sõnasõnalise tõlkega või loomingulise tõlkega või millegi vahepealsega), keelte omavahelisest sugulusest või suguluse puudumisest, sisendtekstide mahust jne. Sõltuvalt sisendtekstide omadustest on tekstide paralleelistamiseks sobivam üks või teine paralleelistusmeetod. (Singh, Husain 2005)

[8] IBM-i mudelite kohta vt lähemalt ptk 4.4.3.

1.2.3. Tasandid

Lausetasandil paralleelistamine kui üldiselt kõige olulisem etapp võib mõnede tekstitüüpide (DVD-subtiitrid, tarkvara kasutajaliideste tõlkimisel produtseeritavad .po-failid, nummerdatud lõikudega pühakirjatekstid jms) puhul osutuda ka üsnagi lihtsalt teostatavaks, mis selliste tekstitüüpide puhul võimaldab keskenduda spetsiifiliste rakendusvõimaluste uurimisele (vt nt Simões 2004).

Lause- ja sõnatasandil tekstide joondamisele (eeskätt nn „ankurpunktidega" seostuvale G&C algoritmi edasiarendusele) on pühendatud kõige rohkem uurimistööd ja kirjutisi (nt Hofland, Johansson 1998), vähem tööd on esialgu tehtud teiste tasandite (mitmesõnalised üksused, süntaktiline ning semantilis-pragmaatiline struktuur) paralleelistamismeetodite väljatöötamiseks.

1.3. Võrdluskorpused

Lisaks paraleelkorpustele on kasutusel ka **võrdluskorpuste** *(comparable corpora)* mõiste. Võrdluskorpused võivad lisaks erinevatele keeltele sisaldada ka samakeelseid tekste (erinevates keelevariantides), mida ühendab teemavaldkond, tekstitüüp, tekstide loomise periood või mõni muu omadus.

Võrdluskorpustel on mõned kasutuseelised paralleelkorpuste ees, kuigi võrreldavate tekstide segmente ei ole võimalik omavahel paralleelkorpustele sarnaselt seostada. Tuntuim võrreldav korpus on ICE – International Corpus of English[9], mis sisaldab 20 inglise keele varianti kas inglisekeelsetest maadest nagu Suurbritannia või USA või riikidest, kus inglise keel on teiseks ametlikuks keeleks või kõrghariduse keeleks nagu näiteks Indias. Selle korpuse põhieesmärgiks on inglise keele erinevate regionaalsete variantide kontrastiivne uurimine. (Muischnek jt 2003:18)

[9] http://www.ucl.ac.uk/english-usage/ice

1.4. Paralleelkorpuste kasutamine

Nii paralleel- kui võrdluskorpustel on aga lisaks erinevate keelte või keelevariantide kontrastiivuuringutele väga palju erinevaid kasutusvõimalusi nii teoorias kui praktikas, mitmeid neist võimalustest ollakse alles avastamas. Enim levinud paralleelkorpuste rakenduslikud eesmärgid saab jagada kolme suuremasse rühma (Borin 2002:14):

(1) kontrastiivsed ja tüpoloogilised grammatikat ja leksikograafiat hõlmavad lingvistilised uurimused (vt nt Ebeling 1998);

(2) paralleelkorpuste kasutamine masintõlkesüsteemides ja tõlkeabiprogrammides (nt Melby 2000); eraldi võib nimetada mitmesuguste toodete keelelist lokaliseerimist ning internatsionaliseerimist kui masintõlke kitsama fookusega allsuunda;

(3) paralleelkorpuste kasutamine keeleõppes ja -õpetuses (nt Botley jt 2000).

Siia võiks lisada veel paralleelkorpuste põhjal sõnastike koostamis- või täiendamisvõimaluse terminite ekstraheerimise läbi (nt Bowker, Pearson 2002: 171-174)

Veidi marginaalsemate, kuid samuti huvitavate praktiliste kasutusvõimalustena võiks esile tuua mitmekeelse info-otsingu (Davis 1998) ja sõnastike käsitlemise paralleelkorpustena või paralleelkorpuste osadena, et luua uusi ja täielikumaid tõlkevahendeid (Geisler 2002) või mõnel muul eesmärgil. Sõnatähenduste ühestamiseks saab kasutada semantiliselt märgendatud paralleelkorpust (Dien 2002). Samuti on võimalik paralleelkorpusi kasutada keele lingvistiliseks analüüsiks või sünteesiks vajalike vahendite automaatseks produtseerimiseks (Kuhn 2004), erinevates keeltes tekstide automaatseks kategoriseerimiseks (Gliozzo, Strapparava 2005) jne.

Sageli on paralleelkorpustealastest uurimustest kasu praktiliste rakenduste väljatöötamisel ja vastupidi, samuti saab paralleelkorpuste erinevates kasutusvaldkondades rakendada ühtesid ja samu algoritme ja meetodeid. Nii näiteks on elektrooniliste sõnastike põhjal loodud sõnade foneetiliste transkriptsioonide paralleelkorpust analüüsides ehk häälduslikke variante automaatselt võrreldes võimalik mõõta sugulaskeelte (Müller 2005) või ka

ühe keele dialektide (Nerbonne, Heeringa 1997) omavahelist sugulusastet, selle ja teiste sarnaste analüüside tulemusi saab kasutada keeleõppes. Erinevate keelte või keelevariantide selline võrdlemine lähtub masintõlke jaoks välja töötatud algoritmidest, millega otsitakse sarnasusi sõnade ortograafilisel tasandil. Paralleelkorpused võiksid ka hõlbustada terminoloogide tööd ülevaate saamisel mingi valdkonna olemasolevast sõnavarast ning uute terminite loomisel (Trosterud 2002:120).

Lynne Bowker ja Jennifer Pearson (2002: 94-95) nimetavad kolme põhilist inimgruppi, kes peaksid olema huvitatud paralleelkorpuste kasutamisest. Need kolm gruppi on (a) keeleõppurid/-õpetajad, (b) kõik tõlkimisega kokku puutuvad inimesed ja (c) keeletehnoloogid/arvutilingvistid.

Arvutilingvistide põhihuvi on suunatud paralleelkorpuste kasutamisele treeningmaterjalina, mille põhjal välja töötada ja arendada joondamistarkvara. Mida parem tarkvara, seda suuremaid paralleelkorpuseid on võimalik koostada. Suurte paralleelkorpuste tähtsaim rakendusvaldkond on masintõlkesüsteemid.

1.5. Paralleelkonkordantsid

Esmane paralleelkorpuste kasutusfunktsioon seisneb võimaluses leida sõnade või fraaside paralleelkonkordantse erinevates keeltes (st tekstis või tekstides esinevate sõnade loendeid koos oma vahetu kontekstiga ja paralleelkontekstiga). Selleks ei ole tingimata vaja spetsiaalset tõlkeabitarkvara, kuid tõlkeabiprogrammid võimaldavad ka konkordantside leidmist, nagu ka tekstide paralleelistamist.

Tõlkeabiprogrammides esitatakse paralleelkonkordantse tõlgitava lause (lauseosa) täpse või hägusa (*fuzzy*) sarnasuse põhjal tõlkemälukorpuses olevate lausete või lauseosadega. Tõlkevastete leidmiseks tõlkemälust võib kasutada nii statistilisi meetodeid kui ka lausestruktuurianalüüsi (Gaussier jt 2000).

On olemas ka peamiselt paralleelkonkordantside esitamisele suunatud tarkvara (ParaConc) ja samuti on välja töötatud spetsiaalseid lingvistilise suunitlusega tekstitöötlusprogramme, mis sisaldavad funktsioone tekstide joondamiseks, paralleeltekstide töötlemiseks, uurimiseks ja vajaliku info ekstraheerimiseks paralleelkorpusest, näiteks TUSTEP (Tuebingen System of Text Processing Programs, Stahl 2002) või Uplug (Tiedemann 2002).

Paralleelkonkordantside leidmist muudab efektiivsemaks tekstide eelnev lausetasandist detailsem paralleelistus, mis võimaldab vastava funktsiooniga varustatud tarkvaral otsitavaid sõnu või väljendeid ülejäänud tekstist esile tõsta või ka lasta kasutajal paralleelistuses tekkinud vigu parandada (Gaussier jt 2000).

Konkordantside võrdlemisest saavad kasulikku infot nii keeleõppijad ja -õpetajad kui ka tõlkijad. Paralleelkorpust saab nõnda kasutada kakskeelse sõnaraamatu asemel või olemasolevate sõnastike täiendusena, eriti idiomaatika ja keelelise stiili osas. Erialased sõnastikud toovad harva näiteid sõnade reaalsest kasutamisest lausetes, küll aga annab paralleelkorpus võimaluse uurida sõna võimalikke kontekste erinevates keeltes. Tõlkimistöös on oluline rõhutada tõlkemälukorpusest leitavate paralleelkonkordantside rolli terminite ühtlustamisel ja tõlke järjepidevuse tagamisel.

1.6. Vorming

Maailmas ja ka Eestis on olnud üheks levinud keelekorpuste formaadiks TEI (Text Encoding Initiative), kuid nii paralleelkorpuste (Godwin-Jones 2001) kui ka tavaliste korpuste rakendatavuse seisukohalt on soovitatud hakata kasutama TEI asemel XML-formaati (*Extensible Markup Language*). XML-formaadis tekstide teisendamiseks teistsugustele kujudele on välja töötatud mitmeid spetsiaalseid meetodeid (XSLT jm) ning korpuse meta-andmestik võib XML-i puhul paikneda tekstist eraldi XML-failides. Nüüdseks on olemas XML-iga ühilduv TEI-formaat (Sperberg-McQueen, Lou Burnard

2004), mis võimaldab TEI-kujul korpuste töötlemisel kasutada olemasolevaid XML-i jaoks loodud vahendeid.

1.7. Väikesed ja vähemuskeeled

Mitmed autorid (nt Scannell 2003) on pidanud tähtsaks paralleelkorpuste rolli keeletehnoloogia arendamisel iseäranis väikeste ja vähemuskeelte jaoks. Seda võiks silmas pidada ka eesti keelele mõeldes. Väikeste keelte puhul peaks teine korpuse keel kuuluma suurte globaalsete keelte hulka, nõnda saab suures keeles olemas olevate rakenduste eeskujul lihtsamini luua rakendusi väiksemate keelte jaoks. Väikeste ja/või vähemuskeeltega paralleelkorpusi puudutavatest projektidest võiks nimetada järgmisi:

- EMILLE projekt (McEnery jt 2000) – Inglismaal elavate India päritoluga vähemusrahvuste keelte korpused ja paralleelkorpus.
- Inglisekeelse tesauruse põhjal iiri keele jaoks statistiliste meetoditega loodud tesaurus (Scannell 2003). Scannelli hilisem kirjutis[10] käsitleb Saint Louis' ülikoolis inglise-iiri paralleelkorpuse põhjal loodud tarkvara, mis seab normeeritud iiri kirjakeele sõnad vastavusse iiri keele normeerimisele eelnenud perioodi erinevate murrete vastavate sõnadega. Selline programm hõlbustab info-otsingut ja leksikograafilist tööd juhtudel, kui üks ja sama sõna võib keelesiseselt esineda mitmetel erinevatel kujudel.
- OPUS (Tiedemann, Nygaard 2004) – üle 30 miljoni sõna paralleeltekste 60 keeles (sh eesti keel), mis on saadud avatud lähtekoodiga tarkvara kasutajaliidestest ja mujalt
- STRAND (Resnik, Smith 2003) – internetist automaatselt leitud paralleeltekstide kogu, nende seas 59 dokumenti paralleelselt baski ja inglise keeles

[10] http://borel.slu.edu/pub/ccgb.pdf

- PTOLEMAIOS (Kuhn 2004) – see projekt puudutab väiksemaid keeli, mille jaoks ei ole veel välja töötatud vajalikku grammatikat ja leksikoni morfoloogilise, süntaktilise ja semantilise analüüsi ning sünteesi jaoks. (Kuhn 2004) väidab, et analüüsivahendid saab automaatselt tuletada üksnes paralleelkorpusest, mis ei ole suures osas eelnevalt lingvistiliselt märgendatud.

2. Korpusepõhised masintõlke- ja tõlkeabisüsteemid.

Masintõlke (MT) idee on juba väga vana, kuid MT statistiliste meetodite uurimist alustati 1980. aastatel. Samal perioodil algas ka tõlkeabiprogrammide levik. Praeguseks on jõutud olukorrani, kus masintõlkealane uurimine ja arendustöö ongi põhiliselt keskendunud korpustest saadava andmestiku kasutamisele masintõlkes.

Korpusepõhises masintõlkes võib üldjoontes eristada kaht põhisuunda: statistilist masintõlget (Statistical Machine Translation) ja näitepõhist masintõlget (Example-Based Machine Translation).

Statistilises MT-s saab lause- ja sõnatasandil joondatud paralleelkorpusi kasutada tõlke originaalilähedusе hindamiseks. Paralleelkorpusest saab leida, mitu korda teatud sõna, fraas või struktuur kujutub üheks või teiseks võimalikuks tõlkeks, et hinnata lause tõlkevaste adekvaatsuse tõenäosust. Parima tõlke leidmiseks võib seejärel kasutada intellektitehnikast tuntud heuristilise otsingu algoritme. Lisaks originaalilähedusele on aktsepteeritava tõlke saamiseks vaja hinnata ka tõlke soravust, mida saab teha näiteks tõenäosusliku generatiivse grammatika abil. (Koit 2003)

Paralleelkorpusi on seni kasutatud põhiliselt lähedaste sugulaskeelte vahelises MT-s, kuna korpuse paralleelistamine on sugulaskeelte puhul hõlpsam. Mittesugulaskeelte puhul peab selleks, et neist masintõlkes kasu oleks, paralleeltekstid süntaktiliselt märgendama. St, vaja on paralleelseid nn puude pankasid (vt ptk 2.2).

2.1. Statistiline masintõlge

Statistiline masintõlge rajaneb nn müranivooga kanali *(noisy channel)* meetodil, mis võimaldab kasutada mitmeid informatsiooniteooriast, side, kommunikatsiooni, raadio, kõnetuvastuse jm valdkondadest pärit algoritme. Näiteks tõlkides inglise keelest eesti keelde tuleb leida selline eestikeelne lause, mille puhul tõenäosus, et see eestikeelne lause on inglisekeelse lause tõlge, on suurim. Selle tõenäosuse välja-arvutamiseks Bayesi valemi abil on tarvis teada ka tõenäosust, et mingi lause üldse eesti keeles võib esineda. Need tõenäosused saaks leida piisavalt suure paralleelkorpuse alusel. (Muischnek jt 2003:53-54)

IBM-i uurimiskeskuses valmis 1990. aastate algul statistikapõhine MT süsteem Candide inglise keelest prantsuse keelde tõlkimiseks. Statistilised andmed selle süsteemi jaoks saadi kakskeelsest Hansardi korpusest, mis sisaldab parlamendidebattide üleskirjutusi. Süsteemi edasiarendamise võimalusena nähakse lingvistiliste ja statistiliste meetodite kombineerimist: kasutada traditsioonilisi reeglipõhiseid meetodeid lausete morfoloogiliseks ja süntaktiliseks analüüsiks ja genereerimiseks ning statistilisi meetodeid sõnatähenduste ühestamiseks ja sõnavalikuks. (Koit 2003)

MT-süsteemis kasutatav paralleelkorpus ei pea alati olema lausete kujul täistekst, vaid võib olla näiteks ka ainult noomenifraase ja nende tõlkeid sisaldav paralleelkorpus. Philipp Koehn (2003) näitab, et kui integreerida tänapäevastesse statistilise masintõlke süsteemidesse eraldi noomenifraase tõlkiv allsüsteem, siis on lõpptulemus senistest saavutustest parem.

Hiljuti on katset tehtud ka eesti-inglise statistilise masintõlkega (vt Fishel jt 2007). Seejuures kasutati samu paralleelkorpuseid, mille võrdlus on esitatud käesoleva töö viiendas peatükis – TÜ paralleelkorpus ja JRC-Acquis' paralleelkorpuse inglise-eesti osa. Fraasitabelite koostamiseks ja dekodeerimiseks kasutati Mosest[11] – statistilist, keelest

[11] http://www.statmt.org/moses/

sõltumatut, fraasipõhist masintõlkesüsteemi. Sõnatasandi vastavused saadi GIZA++ (Och, Ney 2000) abil[12].

2.2. Puude pangad

Prahas asuvas Karli ülikoolis väljatöötatud tšehhi-inglise masintõlkesüsteem (Cmejrek jt 2003), kasutab lisaks inglisekeelsele korpusele ja elektroonilistele sõnastikele ka tšehhi-inglise paralleelset, lause lingvistilist tähendust esitavat puude panka (puude pangaks nimetatakse süntaktiliselt annoteeritud lausete kogumit, kus märgendus on käsitsi üle kontrollitud; Karli ülikooli projektis kasutati nn sõltuvuspuid – Dependency Based Machine Translation). See puude pank hõlmab inimestest tõlkijate poolt spetsiaalselt projekti tarbeks inglise keelest tšehhi keelde tõlgitud 11 000 lausepaari. Tõlkimisel püüti säilitada originaallausete struktuuri, et hiljem saada võimalikult häid automaattõlkeid.

Eesti keele osalusel on koostatud nn Sofie puude pank projekti Nordic Treebank Network raames. See korpus hõlmab Jostein Gaarderi romaani "Sofie maailm" kahe esimese peatüki tõlkeid kuude põhja-euroopa keelde, mis on süntaktiliselt märgendatud. Sofie puude panga märgendamisskeemiks valiti märgendamisskeem VISL[13], mis kombineerib fraasistruktuuri- ja sõltuvuspuu head omadused. Puude kombineerimise standardiks valiti TIGER XML formaat, mis võimaldab paralleelpuudepanga ükskeelsete osade töötlemiseks kasutada redigeerimisvahendit *Annotate* ning päringu- ja visualiseerimisvahendit *TigerSearch* (Nivre jt 2005).

Puude pankade vajalikkust nii keeletehnoloogias kui ka lingvistilises uurimistöös on esile toonud mitmed autorid, vt näiteks (Volk, Samuelsson 2004) või (Abeillé 2003). Nii puude pangad kui ka paralleelkorpused on arvutuslingvistikas viimasel ajal saanud väga populaarseks uurimisaineks, ent nende kahe valdkonna kombinatsiooni, paralleelsete puude pankade vastu on huvi tuntud märksa vähem.

[12] Vt lähemalt GIZA++ kohta käesoleva töö ptk 4.4.3 ja samuti (Muischnek 2006)
[13] http://visl.sdu.dk/

Siiski on paralleelpuudepankade tähtsus väga suur nii masintõlkesüsteemide arendamise kui paralleelistamisprogrammide töö parandamise, samuti komparatiivse lingvistika seisukohalt. Paralleelpuudepankasid saab tavalistest paralleelkorpustest paremini rakendada ka masintõlkeprogrammi töö tulemuse automaatseks hindamiseks, kuna lisaks paralleelvastete leidmisele saab tõlgete hindamisel kasutada ka puude pankadest saadavat morfosüntaktilist teavet. Seni on paralleelpuudepankasid loodud aga väga vähe ja olemasolevatel paralleelpuudepankadel on mitmeid puudusi – enamasti hõlmavad nad ainult spetsiifilise ainevaldkonna tekste ja pole fraasi- ning sõnatasandil joondatud. (Volk, Samuelsson 2004; Uchimoto jt 2004)

2.3. Tõlkemälusüsteemid

Tõlkemälusüsteemide osakaal professionaalsete tõlkijate poolt kasutatavate abivahendite hulgas on viimastel aastatel kasvanud, kuna infotehnoloogia arengud annavad tõlkemälu kasutamiseks järjest paremaid võimalusi. Traditsiooniline automaattõlge (kus väljundkeele lause produtseeritakse automaatselt keelereeglite ja leksikoni abil) aga õigustab end siiani vaid teatud üksikute juhtumite korral (sisend on kas eeltoimetatud, kujutab endast mingit piiratud ainevaldkonda või on väljund üksnes teksti sisu mõistmiseks vajalik toortõlge). Reeglipõhised MT-süsteemid ei oska olla loovad, arvestada piisavalt konteksti ega lahendada korrektselt semantilisi mitmesusi. Arvatakse, et ideaalses MT-süsteemis oleksid omavahel kombineeritud lingvistiline analüüs empiirilise andmestiku (korpused) kasutamisega.

Tõlkemäluprogrammides kasutamiseks on võimalik paralleelkorpusest lihtsal teel moodustada tõlkemälu, kuhu programmi kasutaja saab tõlketöö käigus omapoolseid täiendusi lisada. Arvatakse, et tõlkemälu tõlkimisel tarvitavad ja töö käigus täiendavad ettevõtted hakkavad tulevikus oma tõlketöö läbi tekkinud paralleelkorpust käsitlema kommertsiaalse produktina – potentsiaalseteks ostjateks oleksid sama tegevusvaldkonda

jagavad ettevõtted, kelle tõlkeprogrammides vastav tõlkemälu veel puudub (Bowker, Pearson 2002: 96).

Üks tuntumaid tõlkeabiprogrammide tootjaid SDL International pakub tarkvarapaketi Trados Freelance osana ka paralleeltekstide joondamisprogrammi Winalign, mis tekstide joondamisel arvestab nii tekstide struktuurilisi kui ka sisulisi omadusi ning võimaldab kasutajal joondamistulemusi korrigeerida. Samasuguseid paralleelistamisvahendeid on lansseerinud ka teised konkureerivad firmad.

Kommertseesmärkidel toodetud joondamisprogrammide tööprintsiibid sarnanevad vägagi arvutilingvistide poolt lingvistiliste uuringute tarbeks välja töötatud joondamis-programmide tööpõhimõtetele, kuid kommertsjoondajad on kasutajasõbralikumad, sisaldades kasutajaliideseid automaatse paralleelistamise tulemuste käsitsi kinnitamiseks või parandamiseks (Bowker, Pearson 2002: 102).

Joondamisalgoritmide arendamise tulemusel loodetakse tõlkeabiprogramme tõhustada selles suunas, et oleks hõlpsam tõlgitavatele keeleüksustele ainult osaliselt sarnanevaid tõlkemäluvasteid tõlketöös kasutada (Gaussier jt 2000). Samuti nähakse tõlkeabi-programmide tulevikku paralleelkonkordantside genereerimise protsessi kombineerimises terminoloogiliste andmebaaside kasutusega ja tõlkemälu lausete osade paralleelistamisega (Gaussier jt 2000).

Martin Volk (2005) on veendunud, et tuleviku tõlkeabiprogrammid põhinevad internetis leiduvatel keeleressurssidel. Üht võimalust internetist automaatselt paralleeltekstide leid-miseks on kirjeldanud Philip Resnik (1999). Nõnda leitud paralleeltekstide laused saaks pärast automaatset statistiliste meetoditega hindamist (kus saab võtta arvesse ka näiteks mõlemas keeles HTML-dokumentide struktuuri sarnasusi) lisada tõlkeprogrammi mällu.

Teksti tõlkevaste leidmiseks internetist saaks kasutada ka mõnd automaatset MT-süsteemi. On võimalik, et suured paralleelkorpused ei leiagi otseselt tuleviku tõlke-programmides rakendust ja selle asemel integreeritakse tõlkesüsteemidesse otsisüsteemid, mis otsivad internetist konkreetsele tõlgitavale tekstile sarnaseid tekste ning seejärel nende tekstide tõlkeid. Kui selline tõlkesüsteem leiab ühele lausele mitu erinevat tõlkevarianti, siis

võib statistilise MT eeskujul automaatselt hinnata tõlkevariantide kvaliteeti statistiliste keelemudelite abil. Kõigepealt otsiks see süsteem terviklausete täpseid ja nn hägusaid vasteid (*fuzzy matches*), jätkates osalausete ja fraasidega. Kui suuremaid üksuseid ei õnnestu leida, siis võib tõlkesüsteem lause tõlkimiseks kasutada muid võimalusi (nt lingvistilisel analüüsil põhinevat sihtlause automaatset genereerimist) või abistada tõlkijat sõnatasandil, otsides sõnadele vasteid onlain-sõnastikest. (Volk 2005)

Internetist paralleeltekstide leidmisest keerulisemaks ja palju tähtsamaks küsimuseks osutub tegelikult kogutud ja lause- või fraasitasandil joondatud tekstide tõlkekvaliteedi automaatne hindamine. Tõlkeprogrammi mälu peaks hõlmama ainult häid ja adekvaatseid tõlkeid. (Volk 2005)

Järgnevalt iseloomustan lühidalt paralleelkorpustel põhinevaid näitepõhiseid tõlkesüsteeme, mis kujutavad endast tegelikult tõlkemäluprogrammide edasiarendusi.

2.4. Näitepõhine masintõlge

Näitepõhine masintõlge ehk analoogtõlge hõlmab tihti ka reeglipäraseid ning statistilisi masintõlkemeetodeid. Näitepõhise hübriidsüsteemi ja tõlkemäluprogrammi tööprintsiip ei erine siiski üksteisest seni, kuni tõlkesüsteemi poolt kasutatavas paralleelkorpuses leidub enam-vähem täpne vaste tõlgitavale lausele ning selle lause tõlge teise keelde. Kui sobivat lauset ei leita, siis otsib näitepõhine süsteem tõlke genereerimiseks väiksemaid lauseüksusi, mida töödeldakse statistiliste ja/või lingvistiliste reeglite alusel, aga primaarseks jäävad lauseosade tõlkevastete leidmisel endiselt tõlkenäited. (Carl, Way 2003: xix)

Mõned olemasolevad näitepõhised süsteemid kasutavad paralleelkorpusi vahetult tõlkeprotsessi käigus, kuid tõlkimiseks vajalik teave võidakse tõlkesüsteemi poolt korpusest ammutada ka eelnevalt spetsiaalse õppemooduli poolt. Näitepõhise tõlkesüsteemi poolt nõnda automaatselt genereeritud tõlkemallid sarnanevad reeglipõhises masintõlkes

kasutatavate lingvistide poolt loodud ülekandereeglitega, kuid on siiski põhimõtteliselt nendest erinevad (Carl, Way 2003: xx).

Tõlkeprotsessi käigus paralleelsete tõlkenäidete poole pöörduvad süsteemid on üldiselt seda efektiivsemad ja parema tõlkekvaliteediga, mida suurem on tõlkenäidete korpus; mida suurem on võimalus, et tõlke saamiseks ei ole vaja kasutada grammatilisi kirjeldusi ja keerulisi reegleid, seda suurema eelise saavutavad vahetult tõlkenäiteid kasutavad süsteemid ülejäänud tõlkesüsteemide ees (Carl, Way 2003: xxvi). Loomulikult aitaks tõlke kvaliteedile kaasa kasutatava korpuse sõnatasandil paralleelistatus.

Eiichihiro Sumita (2003) kavandatud näitepõhine MT-süsteem kasutab lausetasandil joondatud kakskeelset korpust, kakskeelset tõlkesõnastikku ja mõlema keele tesaurust. Tesauruse abil mõõdetakse kõigepealt sisendlause ja paralleelkorpuse lähtekeele lausete sõnade semantilist kaugust, mille alusel leitakse sisendlausele semantiliselt kõige lähedasemad näitelaused. Kõige lähedasema näitelause ja selle teise keele vaste kattuva osa vahel luuakse jooksvalt tõlkemall. Kui genereeritakse mitu tõlkemalli (kuna mitu tõlkenäidet olid sisendist samal semantilisel kaugusel), siis valitakse kas kõige sagedasem tõlkemall või kasutatakse tõlkemalli valimiseks sõnasageduste summeerimist või juhuslikku valikuprintsiipi. Sisendlause sõnad, mida tõlkemall ei hõlma, tõlgitakse automaatselt sõnastiku abil ja nõnda saadakse väljundlause teises keeles.

Selline küllalt lihtne tõlkesüsteem on ebaefektiivne pikemate lausete korral, kuna sarnase lause leidmise tõenäosus näitelausete korpusest on väiksem. Katsetades seda meetodit tõlkimaks jaapani keelest inglise keelde, saadi ometi 80 % korral lausetest rahuldav tulemus. Ebaefektiivsust pikkade lausete korral saab vähendada, kui lasta enne tõlkeprotsessi sisend- ja näitelaused automaatselt osadeks jagada. (Sumita 2003)

3. Paralleelkorpuste kasutamine lingvistilistes uurimustes ja keeleõppes

3.1. Lingvistilised ja tõlketeoreetilised uuringud

Paralleelkorpused kujutavad endast kasulikku uurimisainest kontrastiivsele lingvistikale ning tõlketeooriale, pakkudes neile distsipliinidele omavahelisi kokkupuutepunkte ja võimaldades lingvistikal ja tõlketeoorial seeläbi senisest ulatuslikumalt üksteist täiendada.

Paralleelkorpuste abil on võimalik allutada tõlkijate loomingulised lähenemised erinevate keelekonstruktsioonide tõlkimisel statistilisele analüüsile, mille tulemused võivad pakkuda ka lingvistidele huvitavat infot vastavate keelte ülesehituse kohta. Seejuures tuleb esmalt piiritleda sagedasemad nendest juhtudest, mil ühe keele lause või lauseosa on teises keeles edasi antud võrreldes lähtekeelega erineval viisil. Sellised juhtumid võivad seisneda tüüpilistes väljajättudes, osalausete alistustüübi muutustes jne.

Seejärel tuleb määratleda teatav *interlingua,* milles esitatakse tõlgitavate keeleüksuste või mõistete sarnasused ja erinevused vastavas keelepaaris, kusjuures tõlkeüksuste maht võib olenevalt keeltest ja uurimiseesmärkidest suuresti varieeruda (Siin kasutatav *interlingua* mõiste on analoogne reeglipõhises masintõlkes mitmete keelepaaride vahelist tõlkimist hõlbustama pidava tehisliku üleminekukeelega.) Mitte-ekvivalentsete tõlkeüksuste kaardistamisel võime lähtuda mitmetest olemasolevatest teoreetilistest raamistikest. Teine võimalus on koondada tõlkeüksuste võrdlusandmed kontrastiivsesse mitmekeelsele *wordnet*ile sarnanevasse andmebaasi. (Salkie 2002)

Keele ja tõlkimisprotsessi olemuse valgustamiseks sobivad üksikutest keelepaaridest paremini mitmeid erinevaid keeli hõlmavad paralleel- või võrreldavad korpused. Mitme-

keelsete korpuste puhul saab huvitavaid järeldusi teha näiteks ka erinevate lähtekeeltega tõlgete võrdlemisest ühe ja sama keele siseselt.

Samas võime ka ainult ükskeelset korpust uurides saada konkreetsed vastused mitmetele tõlkimisega (nagu ka keeleõpetamisega) seotud küsimustele – näiteks samakeelsete tõlgitud ja originaaltekstide võrdlemisel selgub, mis määral tõlkijad suudavad vältida võõrapäraste lausekonsktruktsioonide ja sõnade põhjendamatut ülevõtmist tõlgitavast keelest.

Selliste uuringute tulemuste kinnitamiseks on aga vajalik uurimisse kaasata ka vastavad tekstid selles keeles, millest tõlgiti, ning soovitavalt ka teisi keeli, millesse samad tekstid veel tõlgitud on – nõnda saame teada, kas originaali ja tõlke süntaktilised või leksikaalsed erinevused on mingis osas võrreldes teiste keeltega märkimisväärsed. (Johansson 2002)

Nii nentis Martin Gellerstam (1996: 59) rootsi verbi *tillbringa* ülemäärast tarvitust rootsikeelsetes tekstides, mis olid inglise keelest tõlgitud, ja oletas, et tegu on inglise verbi *spend* liiga sagedase otsetõlkega (sõna *spend* tähendussfäär inglise keeles on laiem kui *tillbringa* oma rootsi keeles).

Stig Johansson (2002) kontrollis seda oletust inglise-norra paralleelkorpuse peal, kuhu on lisatud inglisekeelsete tekstide tõlkeid saksa, hollandi ja portugali keeles. Johansson leidis, et ühelt poolt inglise ja teiselt poolt norra ja saksa keel kujutavad aja möödumist erinevalt: inglise keel tõlgendab aja möödumist üldjuhul kui aja kulutamist (*consume*) või möödasaatmist, norra ja saksa keeles eelistatakse aga rääkida sündmuste kestmisest (*duration of an event*). Tõlkimisel kaldutakse aga tõepoolest kasutama inglise keelega analoogseid keelekonstruktsioone, mida tähistab ka *tillbringa* kasutamine tõlgetes. (Johansson 2002)

3.2. Paralleelkorpused keeleõppes

Nagu juba mainitud, paralleelkorpust saab kasutada kakskeelse sõnaraamatu asemel või olemasolevate sõnastike täiendusena, kui sealt tekstiotsingu abil esile tuua vajalikud konkordantsid. Tegelikult on sõnaseletuste ja grammatilise infoga varustatud paralleel-

tekstid aga üks traditsioonilisi keeleõppe vorme näiteks antiikkeelte puhul, arvuti abi paralleelkorpustest keeleõppimiseks vajaliku info ammutamisel võib pidada selle traditsioonilise keeleõppe vormi moodsaks täienduseks. Et traditsiooniline paralleeltekstidest õppimine on populaarne ka tänapäeval, sellest annavad tunnistust mitmete nimekate kirjastajate (Penguin, Harvard, Random House, Reclam, Mercier jt) poolt välja antavad paralleeltekstid (Nerbonne 2000: 3). Paralleeltekste on seostatud nn "tõlkemeetodiga" keeleõppes ning sellel meetodil on olnud nii apologeete kui ka kritiseerijaid; tõlkemeetodit on kritiseeritud "kommunikatiivse pädevuse" arendamise vajalikkuse seisukohast (Widdowson 1990: 117ff).

Ometi ei saa eitada, et keele (mitte ainult kõnekeele) võimalikult täielikuks omandamiseks ning tõlkijate koolitamisel pakuvad paralleeltekstid emakeelena teist keelt kõnelevale keeleõppijale asendamatut abimaterjali. Paralleeltekstid võimaldavad nii koos korpuseanalüüsitarkvaraga kui ka ilma selleta keeleõppijale ligipääsu autentsele keelekasutusele koos võõrkeelsete lausete tõlgetega emakeelde ja see võimalus on sageli olnud õppimist hoogustavaks teguriks. Kui aga tegemist on juba elektroonilisel kujul paralleelkorpusega, siis osutub kvaliteetse korpuse kasutamiseks mõeldud tarkvara olemasolu korpuse abil õppimisel hädavajalikuks – ilma spetsiaalse arvutiprogrammita tabab selliseid õppeprojekte tõenäoliselt ebaedu (Nerbonne 2000: 6).

Väidetavalt (Nerbonne 2000) osutuvad paralleelkorpused keeleõppijale kõige kasulikumaks leksikaalsete sõltuvuste võrdlemise osas ning sõnade või väljendite erikeelsete konkordantside esiletoomine peaks kujutama endast soovitavalt üht osa multifunktsionaalsest keeleõppetarkvarast (nagu GLOSSER – rahvusvaheline projekt, milles oli osaline ka Eesti). Korpuse lekseemid – mitte sõnavormid – peaksid olema indekseeritud, et otsinguga leitaks ka sama sõna teistsuguse tüvekujuga vormid. Korpus peaks olema lausetasandil paralleelistatud. (Nerbonne 2000)

4. Leksikoni loomine paralleelkorpuse baasil

4.1. Automaatselt loodavate leksikonide vajalikkus

Tuleviku seisukohalt on uute lähenemiste leidmine kas korpustest või muudest allikatest arvutis loetavate sõnastike loomiseks keeletehnoloogias üks kahest põhisuunast, millele kogu maailmas tähelepanu pööratakse (Muischnek jt 2003). Paralleelkorpustest automaatselt genereeritud sõnastikud leiavad rakendust nii inim- kui masintõlkes, keeleõppes ja ka mujal, näiteks sellise spetsiifilise arvutitehnoloogilise ülesande hõlbustajana, nagu seda on semantiline ühestamine (Ide jt 2002).

Esialgu saab rääkida ainult poolautomaatsest mitmekeelsete erialasõnastike koostamisest paralleelkorpuste baasil, sest ilma inimkorrektori parandusteta saab sellisel viisil koostada üksnes leksikograafidele, terminoloogidele või tõlkesüsteemidele abiks olevat toormaterjali. Erialased tekstid sobivad leksikoni ekstraheerimiseks paremini näiteks kirjanduslikest tekstidest seetõttu, et erialased terminid tähistavad enamasti kindlapiirilisi mõisteid, millele leidub kindel ja järjepidev vaste paralleelteksti sõnade hulgas (Bowker, Pearson 2002: 171–174, 220; Fung 2000). Tõlke järjepidevusest sõltub suuresti ka leksikoni genereeriva programmi võime luua korrektseid seoseid sõnade vahel.

Isegi kui keele erialase valdkonna jaoks on juba olemas kakskeelne sõnastik, siis on üldjuhul paralleelkorpuse põhjal võimalik automaatselt genereerida olemasolevast tunduvalt mahukam sõnastik. Tänapäeva uut tüüpi korpusepõhistes õppijasõnastikes esitatakse lisaks tavapärasele leksikaalsele ja grammatilisele infole ka teavet tähenduste piirangute, kollokatsioonide, grammatiliste mallide, stiili, registri ja kasutussageduse kohta ning luuakse seosed sõna struktuuri, kasutamise ja tähenduse vahel (Kitsnik 2006: 96). Korpusest automaatselt ekstraheeritud sõnastikud on üheks sellise teabe allikaks, võimaldades

muuhulgas ka parandada ja täiendada olemasolevaid sõnastikke nii uute kirjete osas kui ka näiteks muuta sõnatähenduste hierarhiat genereeritud sõnastikus sisalduva korpusest ammutatud sagedusinfo põhjal.

Olgugi et kaasaegne keeletehnoloogia võimaldab automaatselt genereerida vaid üsna suure veaprotsendiga sõnastikke isegi erialakeele puhul, on siiski tekstiressursside olemasolul mahukate ja osaliselt vigaste sõnastike loomine enamasti vajalik mitmetel põhjustel.

Näitepõhise tõlkesüsteemi jaoks on genereeritava leksikoni juures veaprotsendist olulisem kirjete arv, kuna tõlgitava üksuse vastetest korpuses peab tõlke õnnestumiseks vaid ühel olema korrektne paralleelistus (Brown 1997: 6). Suurem maht, mis tähendab saagise eelistamist täpsusele[14], on ka terminoloogi, tõlkija või sõnaraamatu koostaja seisukohalt parem lahendus. Lihtsam, kui otsida tekstist ise programmi poolt leidmata jäänud tõlkevasteid, on programmi poolt välja pakutud valesid märksõnakandidaate eraldada korrektsetest tõlgetest või kustutada.

4.2. Paralleelkorpuste ja võrdluskorpuste põhjal leksikonide genereerimise meetodid

Leksikoni genereerimine seisneb paralleelkorpuses üksteise tõlgeteks olevate sõnade või väljendite tuvastamises ja ekstraheerimises. Seda saab aga teha mitmel erineval moel.

Erinevad lähenemised tõlkevastete automaatsele ekstraheerimisele jagunevad kahte põhikategooriasse: nn „hüpoteesi kontrollimise" ehk heuristilised meetodid (nt Smadja jt 1996) ja estimeerivad meetodid (nt Hiemstra 1997).

[14] Saagise all on siin mõeldud leksikoni jaoks ekstraheeritud korrektsete seoste arvu suhet kõigi paralleelkorpuses sisalduvate ja omavahel tõlkelises seoses olevate sõnade või väljendite arvuga. Paralleelistuse saagise mõõtmiseks kasutatakse ka nn kuldstandardeid. Sellisel juhul kujutab saagis endast korrektsete seoste arvu suhet seoste arvuga korrektselt paralleelistatud alamkorpuses.
Täpsus on korrektsete tõlkevastavuste hulk leksikonis, võrrelduna kogu leksikoni mahuga.

„Hüpoteesi kontrollimine" tähendab tõlkevastete kandidaatide loendi genereerimist. Kandidaadid allutatakse statistilisele analüüsile, mis peab näitama, kas tegemist on tegelike tõlkevastetega või mitte.

IBM-i teadlaste ideedest (Brown jt 1990)[15] lähtuvast statistilise masintõlke paradigmast inspireeritud **estimeerivad meetodid** põhinevad tõenäosusliku bitekst-mudeli loomisel, mis võimaldab tõlkevasteid hinnata mitte ainult eraldi, vaid ka rühmadesse jaotatuna.

Mõlemal lähenemisel on oma plussid ja miinused. Allpool vaatluse alla tulevatest leksikoni genereerimise vahenditest võib PWA-d pidada heuristiliste meetodite ja Giza++ estimeerivate meetodite paradigmasse kuuluvaks. (Tufiş, Barbu 2001: 156, Tiedemann 2003)

Varasemad lähenemised kakskeelse leksikoni ekstraheerimisele (K-vec algoritm, DK-vec algoritm jmt) põhinesid nn „ankurpunktidest" lähtuvatel lausetasandil paralleelistamise meetoditel, mida kombineeriti leksikaalse koosesinemuse analüüsiga (nt Church jt 1991). Heuristilised meetodid sõnade joondamisel lähtuvadki üldiselt ideest, et tuleb leida sõnapaar, mis esineb koos märgatavalt sagedamini kui seda võiks lubada juhus *(t-score*, Dice'i koefitsient jm).

Lisaks koosesinemuse analüüsile on üldkasutatav ka sõnesarnasuse mõõtmine leidmaks omavahel etümoloogiliselt seotud sõnu. Sõnesarnasuse mõõdupuu näiteks on LCSR (*longest common sub-sequence ratio*), mis tähendab kahe sõne pikima ühise tähejärgnevuse ja sõnepaari pikema sõne pikkuse suhet. Sõnesarnasuse mõõtmine eeldab, et mõlemal keelel on sarnane tähestik ja et etümoloogilises suguluses olevate sõnade kirjapiltide vahel eksisteerib arvestatav sarnasus. Tõlkepaaride ekstraheerimiseks sobivaid sõnesarnasuse mõõtmise meetodeid on võrrelnud Lars Borin (1998).

Sõnade joondamise abivahendina saab kasutada ka olemasolevaid elektroonilisi kaks-keelseid sõnastikke. Seejuures sõltub selliste lisaressursside mõju ulatus sõnaparalleelistuse tulemustele suuresti konkreetse sõnastiku temaatilisest sobivusest paralleelkorpuse temaatikaga ning samuti sõnastiku mahust. Mõnede tekstide puhul võib joondamisel abi

[15] IBM-i mudelite kohta vt lähemalt ptk 4.4.3.

olla ka tekstide vorminduse võrdlemisest. Lisaks saab korpusevälise ressursina kasutada ka mitmesugust keelespetsiifilist informatsiooni sõnajärje reeglite, süntaktiliste suhete jms kohta, kuid selliste andmete kombineerimises statistiliste meetoditega ei ole esialgu märkimisväärset edu saavutatud. (Tiedemann 2003)

Kuna võrdluskorpusi on lihtsam koostada kui paralleelkorpusi, siis on püütud välja töötada ka abivahendeid sõnavastavuste leidmiseks joondamata võrdluskorpustes (nt Fung 2000).

Kui on tegemist sugulaskeeltega, siis saab sõnadevahelise vastavuse osaliselt tuletada sõnatüvede sarnasusest, mittesugulaskeelte puhul tuleb vastavuste leidmiseks kasutada statistilisi meetodeid ja/või infot sõnade konteksti ja paiknemise ning nende süntaktilise rolli kohta.

Kasutades info-otsingule sarnaseid meetodeid, võib võrdluskorpuste abiga leida vasted paralleelkorpuste põhjal või traditsioonilisel meetodil loodud leksikonides puuduvatele sõnadele. Nõnda võib osutuda võimalikuks näiteks olemasolevate elektrooniliste sõnastike pidev automaatne täiendamine uute, varem mitte eksisteerinud sõnadega või tuletistega. (Bowker, Pearson 2002: 171–174, 220; Fung 2000).

Tüüpiliselt seisneb sõnavastavuste leidmine võrdluskorpustest lähte- ja sihtkeele kõigi sõnade sagedasemate kollokatsioonide põhjal moodustatud kontekstvektorite võrldemises, kusjuures kontekstvektorite tõlkimiseks kasutatakse olemasolevaid kakskeelseid sõnastikke (Déjean jt 2002: 1). Paremate tulemuste saavutamiseks on kontekstvektor-tõlkimist vaja kombineerida keeleliste lisaressursside (mitmekeelsed tesaurused vms) abi kasutavate meetoditega.

4.3. Grammatilise analüüsi osa sõnastike genereerimisel

Eric Gaussier jt (2000: 254–255) toovad välja kolm faasi, milles sõnade või sõnaühendite tõlkepaaride ekstraheerimine paralleelkorpusest üldjuhul seisneb: tõlgitavate üksuste tuvastamine ning filtreerimine igas keeles eraldi, millele järgneb seoste leidmine tuvastatud

üksuste vahel statistiliste algoritmide abil ning lõpuks leksikoni genereerimine omavahel seostatud sõnade loendi põhjal. Vajalike üksuste tuvastamine võib toimuda ka dünaamiliselt joondamisprotsessi käigus.

Kui kasutada sõnaühendite tuvastamiseks lingvistilisi meetodeid, siis on siinjuures eelduseks, et kõigi keelte (mõlema keele) lausete morfosüntaktiline analüüs toimub piisavalt heal tasemel ja sarnasel moel, mis alati ei ole võimalik. Samuti ei ole väga lihtne kindlaks teha esimeses faasis leitud grammatiliste mallide keeltevahelisi vastavusi, eriti juhul, kui soovitakse leksikoniga hõlmata ka mitmesõnalisi sõnaühendeid.

Viimast probleemi on püütud lahendada erinevatel viisidel. Üheks väljapääsuks on igasuguste grammatiliste korrelatsioonide leidmine keelepaaride vahel enne terminitele või soovitavatele üksustele vastavate grammatiliste mallide tuvastamist. Teine võimalus on kõigepealt leida grammatilise analüüsi abil kõik üksused ainult esimeses keeles olevates tekstides ja seejärel leida nende üksuste tõlked teises keeles. Kui aga eesmärgiks on luua kakskeelne ja kahesuunaline oskussõnastik, siis tuleks esmalt tuvastada terminid mõlemas keeles ja seejärel need terminid paralleelistada.

Grammatilise analüüsi abil leitud tõlgitavate keeleüksuste piire saab täpsustada ka paralleeltekstide põhjal loodud statistiliste tõlkemudelite abil – nõnda osutub reaalseks näiteks ka kõigi püsiühendite ekstraheerimine paralleeltekstidest (Melamed 2001).

Võimalikud on ka keerulisemad algoritmid, mille korral sõnade või sõnaühendite grammatilise struktuuri vaatlemise abil parandataks joondamisel tehtud vigu, ning vastupidi, joondamismalle hinnates täpsustataks vajalike üksuste tuvastamist (Gaussier jt 2000). Allpool pööran aga põhitähelepanu grammatilist analüüsi minimaalselt kasutavatele rakendustele, mille puhul nii sõnade või sõnaühendite tuvastamine kui ka paralleelistamine toimub suuremas osas statistiliste meetoditega.

Sõnastiku automaatsel genereerimisel paralleelkorpustest tuleb vahet teha eeldefineeritud terminite ekstraheerimise ja laiema tähendusega sõnastikugenereerimise vahel (mida võidakse siiski rakendada ka erialaste tekstide põhjal). Arvutilingvistikas määratletakse tehnilisi termineid sageli kitsalt teatud kindlatele morfosüntaktilistele tunnustele vastavate

noomenifraasidena (Blank 2000: 240). Selliste kindlate omadustega noomenifraaside automaatne tuvastamine nõuab korpuselt kindlasti süntaktilist märgendust ja leksikoni genereerimine peab sellisel juhul toetuma konkreetse keele grammatilisele kirjeldusele. Seega moodustab keelest sõltumatute meetodite puhul selline kitsatähenduslik terminoloogia vaid osa korpustest genereeritavast leksikonist, kuna ekstraheeritavatele üksustele ei saa morfosüntaktilise info puudumisel kehtestada sõnaliigist vms lähtuvaid piiranguid.

Esialgu puudub eesti keele jaoks tarkvara, mis teostaks paralleelistamist ja sõnastiku genereerimist lähtudes eesti keele grammatilisest struktuurist ja selle struktuuri vastavustest mõne teise keele struktuuriga. Keelest sõltumatud meetodid tähendavad muuhulgas ka seda, et esialgsed tekstist ekstraheeritud üksused, mille omavahelise joondamise kaudu saadakse lõpuks sõnastik, peavad olema leitud põhiliselt statistiliste meetoditega. Mitmeid uurimusi, millest saab lähtuda kakskeelse leksikoni genereerimisel, on aga tehtud ka ükskeelsest korpusest statistiliste vahenditega mitmesõnaliste üksuste tuvastamiseks (Tiedemann 2003).

Kui lähtuda leksikoni võetavate üksuste joondamisel keele grammatikast, siis tekib järgmine probleem: kuigi sagedasti esinevad sõnaühendite grammatiliste mallide vastavused erinevate keelte vahel on tuvastatavad, esineb siiski suhteliselt palju mitmesusi ja kõrvalekaldeid reeglitest.

4.4. Sõnastiku genereerimise praktilised võimalused

4.4.1. Eeldused ja eeltöötlus

Mitmed kommertstarkvara tootjad (Xerox, Ahead Software, SensoLogic, SDL International) pakuvad mõne oma tarkvarapaketi osana ka terminisõnastiku automaatse ekstraheerimise võimalust etteantud tekstide põhjal, kuid nendel programmidel puudub esialgu eesti keele tugi ning kommertstarkvara poolt kasutatavad meetodid ei võimalda töö kohaldamist teiste keelte tarbeks. Seetõttu on esimeseks loogiliseks sammuks teel eestikeelse osalusega sõnastiku automaatse genereerimise poole keelest sõltumatute

meetodite katsetamine vabavara abil või eesti keelt toetava leksikonigenereerimistarkvara loomine.

Mida tuleks silmas pidada, kui on kavas välja töötada eriotstarbelist leksikograafiatööle orienteeritud tarkvara, mis genereerib paralleelkorpuse põhjal muuhulgas loodavasse sõnastikku sobivaid sõna- või väljendipaare?

Enamik leksikoni ekstraheerimise alastest töödest põhineb kindlatel eeldustel, mida tuleks arvesse võtta eriti juhul, kui tarkvara loomisel soovitakse alustada lihtsamatest algoritmidest. Ükski nendest postulaatidest ei kehti tegelikkuses sajaprotsendiliselt, kuid erandid ei põhjusta nii suurt langust tulemuste kvaliteedis, et eelduste rakendamine poleks põhjendatud. Need eeldused on:

a) mitmetähenduslikku leksikaalset üksust kasutatakse ühe ja sama teksti siseselt ainult ühes kindlas tähenduses;

b) tõlkeüksuste paari kuuluvate sõnade sõnaliigid peavad omavahel sobima, st näiteks verbile ühes keeles võib vastata üksnes verb või mõni teine sõnaliik, mis on tunnistatud võimeliseks täitma tõlkes verbi funktsiooni – ka see reegel ei ole muidugi tegelikkuses absoluutne;

c) tõlkevastete kandidaatide seas on tõenäolisemad tõlkevasted need, millesse kuuluvate sõnade suhteline asend lauses on üksteisele lähedasem;

d) tõlkepaari ühe poole leksikaalsele üksusele vastab maksimaalselt üks leksikaalset üksust tõlkepaari teisel poolel.

Esimene eeldus vastab paraku seda vähem tõele, mida vähemkasutatava sõna või sõnaühendiga on tegu ja ka suur hulk oskussõnu tõlgitakse teise keelde mitmel erineval moel isegi sama teksti siseselt. Ingeborg Blank (2000) leidis prantsuse-saksa näidiskorpuse varal tehtud katse abil, et 5–15 % terminitest on sellised, millel on teises keeles rohkem kui üks tõlkevaste (Blank 2000: 246–247).

Ühe termini tõlkevasted võivad olla aga erineva grammatilise struktuuriga. Nii näiteks esineb sõna *sihtliikmesriik* vastena TÜ inglise-eesti paralleelkorpuses kahel korral *the Member State of destination*, aga ühel korral ka *the destination Member State*; sõna

lähetuskoht esineb inglise keeles kaheksal korral kui *place of dispatch*, kahel korral kui *place of destination* ning kahel korral vastab noomenifraasile hoopis umbisikuline verbivorm: *(the products) are dispatched from*. Selliste mittevastavuste võimalikkust tuleb arvesse võtta nii terminite automaatsel piiritlemisel kui ka joondamisfaasis.

Ka viimast, üks-ühele vastavuse eeldust on leksikonide genereerimisel küll laialdaselt kasutatud, kuid ka see eeldus tekitab siiski suhteliselt palju ebakorrektseid tõlkeid, kui ühe keele liitsõnale vastab teises keeles mitmesõnaline väljend. Niisiis on see nn „1:1-kaardistuse hüpotees" inglise-eesti keelepaari korral küsitav: eesti keele liitsõnale vastab inglise (või ka näiteks prantsuse) keeles tavaliselt mitmesõnaline üksus. I. Blank (2000: 247) toob näite selle kohta, kuidas saksa keele liitsõnaline termin (*Einspruchs-beschwerdeverfahren)* esineb vastavates prantsusekeelsetes tekstides järjepidevalt kujul *procédure de recours engagée à l'éncontre d'une décision rendue sur opposition* (kompleksne noomenifraas).

Seda probleemi on täheldanud Pim van der Eijk (1993), kelle uurimus põhineb inglise-hollandi korpusel, Lars Ahrenberg jt (1998), kes tegelesid inglise-rootsi korpusega ning samuti I. Blank (2000), tuginedes saksa-inglise-prantsuse korpusele, jt. Keelespetsiifilise eel- (automaatne segmenteerimine) ja järeltöötluse (osaliste tõlgete filtreerimine) abil on 1:1-kaardistusest tulenevad probleemid vähemalt osaliselt ületatavad (Tufiş, Barbu 2001: 157). Parema tulemuse saamiseks tuleks siiski vähemalt eesti keele puhul üks-ühele paralleelistusega ühendada mitmesõnaliste üksuste tuvastamine ja joondamine.

Paralleelkorpuse eeltöötlus võiks lisaks leksikaalsete üksuste segmenteerimisele eesti keele puhul hõlmata ka Kadri Muischneki (2006) poolt käsitletud inglise-eesti masintõlke kvaliteedi parandamiseks sobivaid meetodeid – ühendverbide restruktureerimist ning liitsõnade osadeks jaotamist. Selline eeltöötlus eeldab aga omakorda morfoloogilist analüüsi, mis tähendab küll eemaldumist esialgsest keeltevahelise portatiivsuse printsiibist, kuid võimaldab tõlkevasteid ekstraheerida tunduvalt lihtsamalt.

4.4.2. Lihtsa leksikograafilise abivahendi kavand

Morfoloogiline analüüs lubaks tarkvara kavandamisel lähtuda kõigepealt ainult ühest sõnaliigist, näiteks verbidest. Sellisel juhul peaks programm esmalt morfoloogilise analüsaatori abil kõik valitud sõnaliigi esindajad sisendkorpuses tuvastama ja lemmatiseerima ja korpuse paralleelistama lausetasandil. Paralleelistamiseks võib kasutada mõnd Gale'i ja Churchi algoritmi modifikatsiooni (nt Davis jt 1995), mille puhul on võimalik enne joondamist laused filtreerida, teostades joondamise ainult punktuatsiooni, pärisnimede vms põhjal.

Seejärel võiks programm luua nimekirja kõigist võimalikest valitud sõnaliiki kuuluvatest keeltevahelistest sõnapaaridest, mis ei ületa oma joondamisüksuse piire. Selline loend võib endast juba kujutada arvestatavat abimaterjali leksikograafile, kuid tulemuse parandamiseks peaks programm lisaks võrdlema kõigi korpuse valitud sõnaliiki kuuluvate keeltevaheliste sõnapaaride elementide omavahelist ortograafilist sarnasust ja koosesinemise tõenäosust (väljendatuna näiteks seosetugevuse üldtuntud mõõdu, Dice'i koefitsiendi või mõne selle variandina). Kuna tegemist on sõnapaaridega, siis saab seejuures rakendada kollokatsioonide analüüsimisel kasutatavaid meetodeid (Brew, McKelvie 1996: 48).

Igale sõnapaarile omistaks programm nende näitajate alusel arvulise märgendi, mis iseloomustab tõenäosust, et tegemist on vastastikuste tõlgetega. Kasutajale esitataks edasiseks töötluseks üksnes need sõnapaarid, millega vastavusse seatud arv ületab kasutaja seatud lävendi. Nõnda saab tekstist automaatselt esile tuua väidetavalt 30 % kõigist tekstis leiduvatest korrektsetest tõlgetest, täpsusega 90 % (Brew, McKelvie 1996: 51). Kui tõenäosuse arvutamisel lähtuda üksnes ortograafilisest sarnasusest, siis on võimalik esile tuua ka potentsiaalsed eksitavad valepaarid (nn *faux amis*) – kirjapildilt üksteisele sarnanevad, kuid tähenduselt erinevad sõnad. (Brew, McKelvie 1996)

4.4.3. Poolautomaatsed vahendid

Kui tarkvara väljatöötamine ei ole mingil põhjusel võimalik või otstarbekas, siis on sõnastiku genereerimiseks võimalik kasutada ka vabavara.[16] Kõige lihtsam viis sõnade käsitsi joondamiseks või ka paralleelteksti põhjal sõnastiku toormaterjali loomiseks on kasutada selleks mõnd graafilist joondamisvahendit, mis võimaldab biteksti üksteisega vastavuses olevad sõnad hõlpsasti omavahel ühendada ja genereerida automaatselt omavahel seostatud sõnade loendi.

Sellise joondamisvahendi näide on Rebecca Hwa ja Nitin Madnani poolt välja töötatud Java-programm[17], mis genereerib loendi kasutaja poolt arvutihiire abil seostatud sõnadest. Katse selle programmi abil paralleelistada sõnatasandil TÜ inglise-eesti paralleelkorpusest juhuslikult valitud lause andis järgneva tulemuse (1). Toon kõigepealt ära kasutatud lause inglise ja eesti keeles ja seejärel selle lause paralleelistusele vastava väljundi.

(1)

a) The representatives agree on the desirability of acceptance of the WTO Agreement by all participants in the Uruguay Round of Multilateral Trade Negotiations (hereinafter referred to as "participants") with a view to its entry into force by 1 January 1995, or as early as possible thereafter.

b) Esindajad lepivad kokku, et on soovitav, et kõik mitmepoolsete kaubandusläbirääkimiste Uruguay voorus osalejad (edaspidi "osalejad") tunnustaksid WTO lepingut, et jõustada see 1. jaanuaril 1995 või niipea kui võimalik pärast seda.

1 1 (The, Esindajad)

2 1 (representatives, Esindajad)

[16] Vabavarana saadaval olevatest sõnatasandil paralleelistamist võimaldavatest tarkvarapakettidest ülevaate saamiseks vt http://www.cse.unt.edu/~rada/wa/#softwareWA (21.08.2006)
[17] Vt http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm (21.08.2006)

3 2 (agree, lepivad)

3 3 (agree, kokku,)

5 6 (the, soovitav,)

6 6 (desirability, soovitav,)

8 16 (acceptance, tunnustaksid)

18 11 (Uruguay, Uruguay)

16 12 (in, voorus)

17 12 (the, voorus)

19 12 (Round, voorus)

15 13 (participants, osalejad)

14 8 (all, kõik)

7 7 (of, et)

11 17 (WTO, WTO)

10 18 (the, lepingut,)

12 18 (Agreement, lepingut,)

21 9 (Multilateral, mitmepoolsete)

22 10 (Trade, kaubandusläbirääkimiste)

23 10 (Negotiations, kaubandusläbirääkimiste)

24 14 ((hereinafter, (edaspidi)

28 15 ("participants"), "osalejad"))

36 20 (force, jõustada)

35 20 (into, jõustada)

33 21 (its, see)

38 22 (1, 1.)

39 23 (January, jaanuaril)

40 24 (1995,, 1995)

41 25 (or, või)

43 26 (early, niipea)

45 28 (possible, võimalik)

46 29 (thereafter., pärast)

46 30 (thereafter., seda.)

Selline väljund on saadud püüdes seostada omavahel kõik leksikaalsed üksused, mis üldse mingit vastastikust seost omavad. Inglise ja eesti keele morfoloogilistest erinevustest tingituna ei sobi selline tulemus eriti hästi leksikograafiatööks, kuid loomulikult saab leksikograafist kasutaja juba väljundi genereerimisele eelnevalt märgistada ainult potentsiaalselt vajalikud seosed, jättes välja korduvad sõnad, numbrid, mitmesõnalised üksused jms. Sellisel juhul võiks väljund (2) välja näha umbes niisugune:

(2)

2 1 (representatives, Esindajad)

6 6 (desirability, soovitav,)

8 16 (acceptance, tunnustaksid)

19 12 (Round, voorus)

15 13 (participants, osalejad)

14 8 (all, kõik)

12 18 (Agreement, lepingut,)

21 9 (Multilateral, mitmepoolsete)

23 10 (Negotiations, kaubandusläbirääkimiste)

24 14 ((hereinafter, (edaspidi)

39 23 (January, jaanuaril)

41 25 (or, või)

43 26 (early, niipea)

45 28 (possible, võimalik)

46 29 (thereafter., pärast)

Sellise „käsitsi" sõnaparalleelistamise eelis on potentsiaalselt maksimaalne täpsus, kuid miinusteks muidugi täisautomaatsetele lähenemistele vastanduvalt märksa suuremad vajadused inimtööjõu osas ja esialgu ka võimaluse puudumine mitmesõnaliste üksuste joondamiseks.[18]

Järgmine samm täisautomaatse sõnastikugenereerimise suunas on selline poolautomaatne protsess, mille puhul tarkvara poolt teostatud sõnaparalleelistuse tulemused vaadatakse inimkasutaja poolt üle ja vajadusel parandatakse. See on võimalik näiteks kasutades Chris Callison-Burchi poolt loodud graafilist abivahendit[19], mille sisendiks on vabavarana saadaval oleva tarkvarapaketi Giza++ (Och, Ney 2000) poolt sõnatasandil paralleelistatud paralleeltekst. Kasutajale kuvatakse paralleelistus maatrikstabelina, kus on hõlpsasti võimalik parandusi teha.

Giza++[20] tööprintsiibid hõlmavad sarnaselt suurema osaga statistilise masintõlke rakendustest IBM-i uurijate mudeleid, mida tutvustasid Peter F. Brown jt (1993).[21]

[18] On küll olemas ka samalaadne joondamisvahend, mis võimaldab lisaks sõnadele ka fraaside joondamist, kuid selle kasutamine leksikograafilistel eesmärkidel nõuaks väljundi lisatöötlust – vt http://www.isi.edu/~hdaume/HandAlign/ (03.08.2006)
[19] Vt http://demo.linearb.co.uk:8080/sandbox/start.jsp (21.08.2006)
[20] Vt http://www.fjoch.com/GIZA++.html (21.08.2006)
[21] Näide Giza++ sõnaparalleelistustest koos Giza++ sisendiks olnud TÜ korpuse alamosaga on allalaaditav aadressilt www.teataja.ee/leksikonid.zip.

IBM-i mudel 1 leiab sõnadevahelised vastavused lausetasandil joondatud bitekstist sõnade koosesinemise alusel, alustades ühtlustatud tõlkevastetõenäosustest. Mudel 2 lisab sellele lihtsale tõlkemudelile positsioonilised parameetrid ning mudel 3 nn viljakusparameetrid. Viljakusparameetritega tuuakse esile mõnede sõnade kalduvus olla tõlkelises ühenduses tõenäolisemalt ühe- või mitmesõnalise vastega, mis sisaldab teatud arvu sõnu. Mudel 4 hõlmab meetodeid mitmesõnaliste üksuste tuvastamiseks biteksti põhjal genereeritud sõnaklasside võrdlemise abil lauses. Mudeliga 5 on püütud parandada eelmiste mudelite töö käigus esile tulnud vigasid. (Brown jt 1993)

4.4.4. Täisautomaatne leksikoni genereerimine tarkvarapaketi PWA abil

Rootsi teadlaste projekti Plug raames välja töötatud kahe joondamisrakenduse – Linköping Word Aligner (LWA) ja Uppsala Word Aligner (UWA) – loomisel ja arendamisel on muuhulgas arvesse võetud eespool kirjeldatud sõnade üksühese vastavuse eeldusega seotud probleeme. Mõlemad süsteemid kasutavad tekstide sõnatasandil joondamiseks võrdlemisi vähe keelespetsiifilist infot ning on seetõttu üpris lihtsalt rakendatavad ka eestikeelse osalusega sõnastike koostamiseks; mõlemad süsteemid on koostatud programmeerimiskeeles Perl ning on internetist tasuta allalaaditavad.

LWA ja UWA on integreeritud paralleelkorpuste töötlemiseks loodud laiema funktsionaalsusega tarkvaraplatvormi nimega Uplug (Tiedemann 2002). Projekti Plug käigus loodud süsteemide peamisteks rakendusvõimalusteks on peetud (Sågvall Hein 2002) erinevate masintõlkeliikide täiendamist tõlkeinfo ja leksikonidega ning samuti inimtõlkijate abistamist leksikonide loomise läbi. Vaatlen järgnevalt mõlemat joondamisprogrammi ja nende kasutusvõimalusi lähemalt.

UWA sisendiks on lause- või fraasitasandil joondatud bitekst. Operatsioonisüsteemi Windows jaoks praegu saadaval olev PWA (Plug Word Aligner – tarkvarapakett, mis ühendab endas UWA ja LWA) versioon on eelseadistatud rootsi-, inglise- ja saksakeelse sisendteksti jaoks, kuid seadistusi on võimalik kohaldada ka teistele keeltele.

Sisendtekst jagatakse programmi poolt kas juba olemasoleva (kasutaja poolt lisatud) lihtsavormilise sõnastiku või teksti enese põhjal ühe- või mitmesõnalisteks üksusteks. Sõnastiku puudumisel arvestatakse selle etapi juures tähejärjendite sagedusi ja pikkust, samuti sõnatüüpe ja punktuatsiooni.

Järgnevalt püütakse mõlema keele vastavad üksused omavahel kokku viia. See protsess algab nn „kindlate juhtumite" (nt kui korpuse paralleelistatud segment kujutab endast vaid ühesõnalist elementi) eristamisest. Seejärel hindab süsteem sõnede omavahelist sarnasust ja suhtelist paiknemist tekstis (arvestades konteksti) ja märgib tõlkevastete kandidaatidena ära eelnevalt seatud lävendid ületanud sõnepaarid. Eraldi üritatakse seostada vähesagedasi sõnesid. Viimaks toimub tõlkevastete automaatne filtreerimine, millele võib järgneda tulemuste „käsitsi" korrigeerimine.

4.5. Kakskeelse leksikoni genereerimine PWA abil paralleelkorpuste põhjal

4.5.1. Uppsala Word Aligner

Järgnevalt kirjeldan katset luua UWA abil lähtematerjal erialasõnastiku koostamiseks või täiendamiseks, kasutades selleks mahukat paralleelkorpust (730 880 paralleelset ühest või mitmest lausest või (ala)pealkirjast koosnevat lõiku, 24 169 586 sõnet).

TÜ inglise-eesti paralleelkorpuse (TÜPK) korpuse ühendasin leksikoni koostamise otstarbeks JRC-Acquis' mitmekeelse paralleelkorpuse[22] inglise-eesti alamkorpusega, kuna mõlemad korpused sisaldavad Euroopa Liidu seadusandlusega seotud tekste. Mõlemad korpused on samuti paralleelistatud Vanilla paralleelistaja[23] abil.

[22] Vt http://langtech.jrc.it/JRC-Acquis.html (21.08.2006)
[23] Vt htt://nl.ijs.si/telri/Vanilla/ (21.08.2006)

JRC-Acquis' korpus on alternatiivina paralleelistatud ka HunAligniga[24]. Kuigi HunAlign osutus võrdluses Vanillaga (vt osa 5.3.5) oluliselt täpsemaks paralleelistus-vahendiks, oli TÜ korpusega leksikoni genereerimise eesmärgil ühendatavaks alam-korpuseks siiski Vanillaga paralleelistatud JRC-Acquis' korpuse inglise-eesti osa, kuna paralleelistusmeetodite võrdluse tulemused ei olnud leksikoni genereerimisel veel teada.

TÜ korpus oli algselt kujul, kus eesti ja inglisekeelsed üksused paiknevad vaheldumisi ja on üksteisest eristatud keele nimetust sisaldavate märgenditega. JRC-Acquis' korpus oli algselt mitmesugust märgendust sisaldaval TEI-kujul, jaotatuna paljudesse failidesse. Et muuta need korpused PWA-le „arusaadavaks", tuli need korpused teisendada sellisele kujule, mille korral mõlema keele kõik üksused on ümber tõstetud kahte eraldi faili ja on tähistatud omavahel vastavuses olevate numbriliste tähistega. Selleks kasutasin osaliselt Camelia Ignat' poolt loodud Perli-programmi[25].

Katse[26] käigus genereeritud inglise-eesti õiguskeele leksikon sisaldab 130 865 märksõna (vrd EKI inglise-eesti elektrooniline sõnastik – u. 86 000 märksõna, Eesti Õiguskeele Keskuse terminibaas ESTERM 57 829 märksõna) ja koos tõlkevastetega 482 571 sõnet. Sisendina kasutatud koondkorpus hõlmas 730 880 paralleelset ühest või mitmest lausest või (ala)pealkirjast koosnevat üksusepaari. Lisaks sisendkorpusele hõlmas programmi kasutaja-poolne sisend ka väiksemahulist morfoloogiainfofaili eesti keele tüüpiliste sõnalõppude ja ebareeglipäraste verbide kohta ning ka lähtesõnastikku, milleks otsustasin valida EKI inglise-eesti sõnastiku[27]. EKI sõnastik tuli UWA jaoks teisendada sobivale kujule, mis tähendas põhiliselt mitmesugust lühendamist – mittesobivate kirjete, sulgudes asuvate märkuste, liigsete vastete jms automaatset kustutamist UNIXI vahenditega.

---

[24] Vt http://mokk.bme.hu/resources/hunalign
[25] Vt http://wt.jrc.it/lt/Acquis/JRC-Acquis.2.2/alignments/index.html (21.08.2006)
[26] Nii UWA kui LWA abil genereeritud leksikonid on allalaaditavad aadressil www.teataja.ee/leksikonid.zip (29.08.2006). Lisaks koondkorpusest ekstraheeritud leksikonidele leiab sealt ka üksnes TÜ korpuse põhjal UWA-ga genereeritud sõnastiku ning näite Giza++ sõnaparalleelistustest koos Giza++ sisendiks olnud TÜ korpuse alamosaga.
[27] Vt http://www.eki.ee/dict/inglise/ (21.08.2006)

Järgnevalt genereeritud eesti-inglise leksikoni sisendkorpuseks olid üksnes TÜ paralleelkorpuse inglise-eesti ja eesti-inglise osad.[28] See leksikon sisaldab kokku 97 300 märksõna ja koos tõlkevastetega 318 025 sõnet.

Esitan juhusliku parandusteta fragmendi (3) UWA poolt TÜ inglise-eesti paralleelkorpuse ja JRC-Acquis' mitmekeelse paralleelkorpuse inglise-eesti alamkorpuse põhjal genereeritud leksikonist, mis ei ole esitatud UWA graafilise liidese vahendusel, vaid tavalise tekstifailikatkendina:

(3)

{reserve officer}
{
1X:reservohvitseri
}
{reserve officer candidates}
{
1X:reservohvitserikandidaat
}
{reserve officer courses}
{
1X:reservohvitserikursusel
}
{reserve positions and}

[28] Fragment üksnes TÜ paralleelkorpuse põhjal genereeritud eesti-inglise leksikonist on ära toodud käesoleva töö esimeses lisas.

{

2X:reservipositsioonid ja

}

{reserve power system}

{

1X:reservelektrisüsteem

}

{reserve ratio}

{

3X:reservibaasi

2X:reservimäära

}

{reserve service}

{

1X:reservteenistus

Nagu näitest (3) paistab, on väljundsõnastikus ka eesti keele puhul üsna edukalt lahenenud juhtumid, mille puhul eestikeelsele sõnale vastab inglise keeles mitmesõnaline üksus. Küll aga hakkavad silma rohked keeltevahelistest morfoloogilistest erinevustest tingitud ebatäpsused[29], mis eksisteerivad samamoodi eesti-inglise keelepaari jaoks nagu ka rootsi ja inglise keelte jaoks (UWA on loodud arvestades rootsi keele eripärasid, kuid mitmeid väljundi vigasid ei ole suudetud ka rootsi keele puhul esialgu parandada). Samuti

[29] Seda probleemi võib tõlgendada ka andmehõreduse efektist lähtudes, nagu tehti näiteks inglise-eesti statistilise masintõlke tulemuste hindamisel (vt Fishel jt 2007).

on probleemiks ka inglise-eesti statistilise masintõlke puhul mainitud (Fishel jt 2007) inglise ja eesti keele erinev sõnajärjekord.

Kuna UWA paneb tõlkevastavuste leidmisel suurt rõhku sõnesarnasusele, siis tulenevad sellest inglise ja eesti keele mittesuguluse tõttu paljud vead ja osalised vastavused väljundis, näiteks:

(4)

{selling}
{
1X:Lepingus
1X:Selleks
1X:eelkõige
2X:lepingu
3X:müümine
10X:selle
1X:selles
1X:sellest
1X:selline
2X:sellise
1X:sellisel
1X:selliselt
2X:selliste
}

Anna Sågvall Hein (2002: 74) väidab, et erinevast morfoloogiast tulenevad ebatäpsused lahendaks rootsi keele puhul olemasoleva ülekandepõhise tõlkesüsteemi (Multra) liitmine UWA-ga. Eesti keele jaoks sellist tõlkesüsteemi kahjuks esialgu ei leidu. Vähene keelespetsiifiline morfoloogiainfo ei võimalda UWA-l teostada automaatset lemmatiseerimist, kuid oletan, et tõenäoliselt parandaks väljundsõnastiku korrektsust siiski sõnastiku genereerimisele eelnev sisendkorpuse lemmatiseerimine muude vahenditega.

Tulemust parandaks kindlasti JRC korpuse (sisendkorpuse) automaatse paralleelistuse manuaalne korrigeerimine, samuti süsteemi poolt kasutatavate statistiliste meetmete osaline laiendamine sisendkorpuselt internetile andmehõreduse efekti kahandamise eesmärgil. St üks süsteemi alam-moodul võiks olla ühendatud mõne interneti (mitmekeelse) otsimootoriga, millele võib esitada samu päringuid kui sisendkorpuselegi.

PWA-d puudutavatest publikatsioonidest ei selgu, kas süsteem kasutab rootsi keele puhul morfoloogilist infot ka selleks, et parema joondamistulemuse eesmärgil liitsõnad eelnevalt koostisosadeks liigendada (nagu seda on mitmel puhul tehtud saksa-inglise sõnaparalleelistusel – nt (Déjean jt 2002: 6)).

Genereeritud leksikoni saaks täiustada ka mitmesuguse automaatse järeltöötlusega, mis võiks tegelikult olla ekstraheerimistarkvarasse integreeritud. Näiteks oleks soovitatav automaatselt kustutada kõik üksnes mittealfabeetilisi sümboleid sisaldavad märksõnad, kui soovitud tulemus kujutab endast loomuliku keele leksikoni.

Kasutaja poolt lisatud morfoloogiainfo põhjal saaks sõnastiku genereerinud moodul kustutada ka ühe ja sama märksõna erinevaid muutelõppe sisaldavatest, kuid muus osas sarnastest tõlkevastetest ebavajalikud. Samuti ei oleks vaja kohelda erinevate leksikaalsete üksustena identseid sõnu, mille algustähed on lausesisesest positsioonist tingituna erineva suurusega.

A. Sågvall Hein näeb süsteemi arendusvõimalustena lisaks selle täiendamisele reeglipõhiste meetoditega veel sõnastiku märksõnade lemmatiseerimist ning viidete lisamist genereeritud sõnastikukirjetelt kontekstidele korpuses (Sågvall Hein 2002: 74).

4.5.2. Linköping Word Aligner

LWA loomisel on toetutud Pascale Fungi ja Kenneth Churchi (1994) ning Dan Melamedi (1997) joondamisalastele uurimustele. LWA hõlmab nagu UWA-gi põhiliselt statistilisi meetodeid. Algoritm on iteratiivne – tõlkevasted genereeritakse biteksti põhjal, siis kustutatakse genereeritud sõnapaarid bitekstist ja korratakse tsüklit. Statistilistel tõenäosustel põhinevat baasalgoritmi täiendavad neli lisamoodulit ning programmi kasutajaliides võimaldab seadistada mitmeid parameetreid (lisatestid, lävendid, tsüklite arv).

Esimene moodul alustab tööd sõnade jagamisega etteantud info põhjal mitmesugustesse kategooriatesse ja alamkategooriatesse: relevantseteks ja irrelevantseteks (irrelevantsena on määratletud näiteks inglise keele abiverb *do,* millel enamasti puudub üksühene vaste teises keeles), suletud ja avatud sõnaklassi sõnadeks – avatud sõnaklassi kuuluvaid sõnu saab edaspidi joondada vaid teiste sama sõnaklassi sõnadega ja suletud sõnaklassi sõnu saab vastavalt joondada ainult suletud sõnaklassi sõnadega. Suletud sõnaklassi sõnad jagatakse järgnevalt veel alamkategooriatesse. Kategooriate põhjal toimub iteratiivne paralleelistamine.

Morfoloogiamoodul tunneb vastava keele sufiksiloendi järgi ära ühe sõna erinevad vormid. Kui leksikaalsete üksuste paar (X, Y) on mõne eelneva tsükli käigus tunnistatud tõlkevastavuses olevaks, siis otsib moodul teisi kandidaatpaare, mille esimene element on X ja teine element Z, nõnda et leiduvad ka sõned W, F ja G, mille puhul Y = WF ja Z = WG ning F ja G sisalduvad sufiksiloendi ühes ja samas paradigmas. Kui leitakse mitu erinevat üksust, mida saab tähistada sümboliga Z ja mille sufiksid kuuluvad erinevatesse paradigmadesse, siis valitakse neist suurima sagedusega paradigma.

Järgmine moodul tegeleb mitmesõnaliste üksuste paralleelistamisega, kasutades selleks samasuguseid võtteid kui ühesõnaliste üksuste puhulgi. Mitmesõnaliste üksuste kogum koosneb süsteemi poolt automaatselt leitud üksustest ning eelnevalt lisatud keelespetsiifilistest kollokatsioonidest. Lõpuks parandatakse paralleelistust veel vastavalt

joondatavate segmentide suhtelisele asendile tekstides. LWA miinuseks on UWA-ga genereeritavatele leksikonidele sarnase sagedusinfo puudumine väljundist.

Sisendkorpuse mahupiirangu tõttu kasutasin LWA-ga katselise sõnastikumaterjali genereerimiseks TÜ paralleelkorpuse 10 000 paralleelsest lõigust koosnevat alamkorpust. Toon siin ära fragmendi[30] (5) tulemuseks saadud leksikonist, mis sisaldas 8516 kirjet:

(5)

kinnitades	*confirming*
kinnitades	*reaffirming*
kinnitatud	*affixed*
kinnitava	*assurance*
kinnitavad	*reaffirm*
kirikute	*churches*
kirjalik	*written*
kirjalike	*written*
kirjalikke	*written*
kirjaliku	*written*
kirjalikult	*writing*
kirjalikust	*written*

4.5.3. Tulemuste hindamine: ARCADE ja PWA

Joondamisprogrammide võrdleva hindamise standardiseerimisprojekti ARCADE (Véronis, Langlais 2000) raames on kindlaks tehtud, et parimate sõnadevahelise parallee-

[30] Pikem fragment on esitatud magistritöö lisas 2.

listamise süsteemide täpsus ja saagis ulatuvad umbes 75 protsendini (Véronis, Langlais 2000: 386). Seejuures erineb tulemus sõnaliigiti, ulatudes 94 protsendini adjektiivide puhul ja ainult 60–70 protsendini verbide puhul (samas).

Nii hea tulemuse saavutab joondamissüsteem aga üksnes rohke keelespetsiifilise info kasutamisega bitekstide analüüsil. Kuna UWA ja LWA loomisel on järgitud keeltevahelise portatiivsuse eesmärki, siis on nende programmide töötulemuste hindamisel saadav täpsus ja saagis oluliselt väiksem. Projekti ARCADE juhendite järgi arvutatud rootsi-inglise näidiskorpusest ekstraheeritud leksikoni täpsus on UWA poolt teostatud joondamise puhul üksnes 42,2 % ja saagis 37 %; LWA puhul on samad näitajad vastavalt 51 % ja 41,3 %.

Mitmesõnaliste üksuste korral on saagise arvutamine tunduvalt keerulisem kui sellise vastavuse korral, mille puhul ühele sõnele vastab alati ainult üks sõne. On väga raske hinnata terves korpuses sisalduvate mitmesõnaliste üksuste koguhulka ja seetõttu peaks mitmesõnalisi üksusi kontrolli kaasates teostama arvutuse väga väikese tekstinäite põhjal.

ARCADE-projektis esitatud meetodid ei võimalda hinnata seda, kui hästi süsteem jagab biteksti lähtepoole üksused korrektseteks ühe- või mitmesõnalisteks üksusteks, vaid keskenduvad süsteemi poolt pakutud sihtkeele sõnade võrdlemisele etaloniga ehk nn „kuldstandardiga" (Ahrenberg jt 2000b: 4). Seetõttu on PWA autorid välja arendanud uue kontrollmeetodi (samas), mis erinevalt ARCADE-st võtab arvesse ka mitmesõnalisteks üksusteks jaotamist nii lähte- kui sihtkeele siseselt ja samuti erinevaid juhtumeid, mille korral mitmesõnaliste üksuste joondamist saab lugeda korrektseks vaid osaliselt. Sama rootsi-inglise näidiskorpuse põhjal süsteemiga UWA genereeritud leksikoni täpsus on selle meetodi kohaselt 71,8 % ja saagis 37,4 % (LWA-ga genereeritud leksikoni puhul on need näitajad vastavalt 71,9 % ja 42,6 %).

4.5.4. Eesti-inglise paralleelkorpuste põhjal genereeritud leksikonide hindamisest

PWA tarkvarapakett sisaldab muuhulgas moodulit joondamistulemuste automaatseks hindamiseks nii ARCADE kui ka PWA meetodil. Mooduli kasutamine eeldab PWA nn

kuldstandardi formaadis võrdlusfaili olemasolu vastava keelepaari jaoks, kus korrektsed vastavused on registreeritud koos mitmesuguse juurdekuuluva infoga ning näitelausetega. Eesti-inglise keelepaari tarbeks selline kuldstandard esialgu puudub. PWA kuldstandardifailide loomiseks on kasutatud samuti tarkvaraplatvormi Uplug kuulunud programmi Plug Link Annotator (PLA), kuid kahjuks ei ühildu PLA enam aastal 2006 kasutatavate operatsioonisüsteemide ja muu vajaliku tarkvaraga (Java) ega ole seetõttu kasutuskõlblik. Seega tuleb standardfaili loomiseks leida mõni alternatiivne poolautomaatne võimalus või teostada kontroll käsitsi.

Ka PWA hindamismeetodi tulemused varieeruvad üsna palju sõltuvalt tekstide žanrist ja kuldstandardi koostamise kriteeriumitest – st sõltuvalt sellest, millise sagedusega ja funktsiooniga sõnu ja millistes proportsioonides valitakse nende näidete hulka, mille alusel paralleelistustulemusi hinnatakse (Ahrenberg jt 2000b: 6).

Rootsi-inglise korpusest pärineva kolme erineva tekstikategooria UWA-süsteemi joondamistulemuste hindamine PWA-meetodil andis kõige kõrgema tulemuse (täpsus: 81,26 %; saagis: 64, 47 %) tehnilise sisuga tekstide puhul ja kõige madalama tulemuse (täpsus: 69,04 %; saagis: 41,44 %) poliitiliste tekstide puhul (samas). Vahepealse tulemuse saavutas ilukirjanduse alamkorpus (samas). Võib eeldada, et nende kolme tekstikategooria hulgast asetub TÜ eesti-inglise õigusakte ja seadusi sisaldav paralleelkorpus žanriliselt kõige lähemale poliitika-alastele tekstidele ning oodatav tulemus leksikoni täpsuse ja saagise hindamisel peaks olema võrreldav tulemusega, mille nimetatud katse käigus andis see tekstikategooria.

See oletus pidas tulemuste kontrollimisel vähemalt osaliselt paika: UWA-ga TÜ paralleelkorpuse ja JRC-Acquis' korpuse põhjal genereeritud inglise-eesti leksikonist juhuslikult valitud 50 kirje kontrollimisel sain selle alamosa täpsuseks 60 %. UWA-ga ainult TÜ paralleelkorpuse põhjal genereeritud eesti-inglise leksikonist juhuslikult valitud 50 kirje kontrollimisel sain selle alamosa täpsuseks 87 %. Samasugune kontroll LWA-ga

TÜ eesti-inglise paralleelkorpuse põhjal genereeritud leksikoni 50-kirjelise alamosa peal andis täpsuseks 61 %.[31]

Paremad tulemused eesti-inglise leksikoni kontrollimisel on seletatavad TÜ paralleelkorpuse parema paralleelistusega (võrreldes JRC-Acquis' korpuse Vanilla versiooniga) ja sellega, et sisendkorpuse suure mahu tõttu ei suutnud programm inglise-eesti leksikoni genereerimisel läbi viia viimast, automaatse filtreerimise etappi.

Saagise hindamine ilma kuldstandardfailita oleks korpuse suure mahu tõttu keerulisem ülesanne ja väikese alamosa põhjal saadud hinnangud ei pruugi olla adekvaatsed, seetõttu saagise arvutamisest esialgu loobusin.

Uplugi tarkvara abil on proovitud genereerida ka kreeka-inglise leksikoni (Charitakis 2007), mille manuaalsel hindamisel otsustati hindamisprotsessi mitte kaasata tõlkepaare, mille esinemissagedus oli alla kolme. Ülejäänud tõlkepaarid jagati viide gruppi sõltuvalt nende esinemissagedusest ja kõiki gruppe vaadeldi hindamisel eraldi. Leiti, et tõlkepaaride esinemissagedus ja tõlke korrektsus on genereeritud leksikoni puhul otseses proportsionaalses vastavuses – suurema esinemissagedusega tõlkepaaride korrektsuse protsent on suurem. (Charitakis 2007)

Sama tendentsi kinnitab ka Uplugiga genereeritud inglise-eesti leksikonide vaatlus. Kontrollisin kümmet juhuslikku tõlkepaari UWA-ga TÜ paralleelkorpuse ja JRC-Acquis' korpuse põhjal genereeritud inglise-eesti leksikonis, mille esinemissagedus oli 3. Nendest 5 (50 %) olid ebakorrektsed. Seejärel kontrollisin kümmet juhuslikku tõlkepaari esinemissagedusega 20. Nende seas leidus ainult üks tõlkepaar, mida võis lugeda ebakorrektseks.

Projekti Plug meetoditele väga sarnast lähenemist leksikoni genereerimisele esindavad Dan Tufiş ja Ana-Maria Barbu (2001), kes kasutasid leksikoni koostamiseks europrojekti

[31] Täpsuse arvutamisel lugesin leksikoni kirje korrektseks (1 punkt), kui tõlkevastete seas leidus vähemalt üks täielikult märksõnale sisuliselt vastav tõlkevaste ja osaliselt korrektseks (0,5 punkti), kui tõlkevastete seas leidus vähemalt üks osaliselt märksõnale sisuliselt vastav tõlkevaste. Täpsuseks lugesin kirjete arvu (50) jagatud summeeritud punktide arvuga.

Multext-East[32] raames valminud paralleelkorpust, mis sisaldab George Orwelli romaani „1984" kaheksas, sh eesti keeles.

D. Tufiş ja A.-M. Barbu kirjeldavad muuhulgas ka väikese eesti-inglise sõnastiku ekstraheerimist ning tulemuste kontrollimist. Kontroll teostati „kuldstandardi" puudumisel käsitsi ja eesti-inglise sõnastiku hindamisel saadi selle täpsuseks 96,2 % ja saagiseks 57,9 %. Paremaid tulemusi võrreldes Plug-projektiga võib seletada mitmesuguse eeltöötlusega, mida oli rakendatud MULTEXT-East korpusele ja millele toetus tõlkevastekandidaatide loendi ekstraheerimine: lisaks lausetasandil paralleelistusele ka leksikaalsete üksuste segmenteerimine, morfoloogiline ühestamine ja lemmatiseerimine.

Sõnatasandil paralleelistamist saab teostada ka statistilise masintõlke rakenduste abil, eespool mainisin tarkvarapaketti Giza++. TÜ inglise-eesti korpuse paralleelistamise tulemuste kohta Giza++ abil vt K. Muischnek (2006).

4.5.5. Võrdlus ESTERMiga

Eesti Õigustõlke Keskus on koostanud tõlgitud Euroopa Liidu ja Eesti Vabariigi õigusaktide põhjal terminibaasi ESTERM[33], mis umbes neljandiku ulatuses sisaldab ka fraase. Lisaks õiguskeelele sisaldab ESTERM sarnaselt paralleelkorpuste põhjal genereeritud leksikonidega termineid ka paljudest teistest valdkondadest, mille käsitlemine seadusandlikes tekstides on vajalik.

Üks reaalsemaid inglise-eesti paralleelkorpuste põhjal genereeritud leksikonide kasutusvõimalusi näib olevat ESTERMi terminibaasi täiendamine uute terminitega leksikonidest saadud info põhjal. Kontrollimaks seda hüpoteesi, esitasin ESTERMI terminibaasile päringu kahekümne viie UWA-ga TÜ paralleelkorpuse ja JRC-Acquis' korpuse põhjal genereeritud leksikonist vabalt valitud ja leksikonis korrektset tõlget omava termini kohta.

[32] Vt http://nl.ijs.si/ME/ (21.08.2006)
[33] http://mt.legaltext.ee/esterm/ (29.05.2007)

Kaheteistkümnel juhul kahekümne viiest esines leksikonist valitud märksõna koos oma tõlkevastega täpselt samasugusel kujul ka ESTERMIS. Üheteistkümnel juhul puudus leksikoni kirje täielikult ESTERMist. ESTERMist puuduvad kirjed olid järgmised:

1) rewinder (‘ümberkerija’)
2) active biocides (‘aktiivbiotsiidid’)
3) post-release (‘keskkonda viimise järgne’)
4) allocation method (‘eraldamismeetod’)
5) pyraclostrobin (‘püraklostrobiin’)
6) pyrasolidon (‘pürasolidoon’)
7) textile committee (‘tekstiilikomitee’)
8) textile pulp (‘tekstiilimass’)
9) piston rod (‘kolvivarras’)
10) efficiency indicator (‘toimearv’)
11) terminator (‘terminaator’)

Ühel juhul ei sisaldanud ESTERM täpsel kujul inglisekeelset märksõna *(‘bioaccumulated’*), kuid sisaldas selle eestikeelset vastet (‘*bioakumuleeruv*’ inglisekeelse tõlkega ‘*bioaccumulative*’). Samuti ei olnud ESTERMis märksõna ‘*semiconductor product*’ (‘*pooljuhttoode*’), kuigi hägusa otsingu meetodil on võimalik terminibaasist leida pikem termin ‘*topography of semiconductor products*’ (‘*pooljuhttoodete topoloogia’*), mille koostisosaks on leksikonist valitud termin.

Kuna päringutest rohkem kui 50 % korral puudus ESTERMis leksikonis esinenud termin täpselt samasugusel kujul, siis leidis kinnitust oletus, et on võimalik ESTERMi terminibaasi täiendamine uute terminitega leksikonidest saadud info põhjal.

5. Paralleelkorpuste võrdlemine ja paralleelistuse kvaliteedi hindamine[34]

Kõige olulisem kvaliteedikriteerium on paralleelkorpuse puhul see, et paralleelüksuse koosseisus olev tõlge vastaks samas üksuses olevale originaalile. Juhul, kui paralleelkorpus luuakse (pool)automaatselt, on sellise nõude 100%-line täitmine pea võimatu. Ka paralleelistuste korrektsuste automaatne hindamine on väga raske, sest tegemist on tähenduste omavahelise vastavuse hindamisega.

Seetõttu jäetakse suurte paralleelkorpuste puhul hindamise küsimus sageli üldse tõstmata: nt OPUS (Tiedemann, Nygaard 2004), Europarl (Koehn 2002), tšehhi-inglise paralleelkorpus (Bojar, Žabokrtský 2006). Mõne korpuse dokumentatsioon sisaldab küll hoiatust, et paralleelistus on tehtud automaatselt või öeldakse lausa, et paralleelistustes on vigu. Ainus viis korpuse paralleelistuse kvaliteeti hinnata on teha seda käsitsi, väiksema osavalimi peal, mis on aga äärmiselt töömahukas, vt nt. (Samy jt 2006; Singh, Husain 2005).[35]

Et peatükis 5.5. kirjeldatud leksikoni genereerimise ja masintõlkealaste eksperimentide tulemusi parandada, tuleks olemasolevad eesti keele osalusega paralleelkorpused ühendada nõnda, et loodavast koondkorpusest välja jätta osalised kattuvused ja võimalikult palju vigaselt paralleelistatud osi. Enne korpuste ühendamist on vaja ühtlasi hinnata korpuste

[34] Viies (ja osaliselt ka kuues) peatükk rajaneb koos Heiki-Jaan Kaalepiga kirjutatud ja 2007. a septembris Kopenhaagenis toimuvale XI masintõlke-alasele tippnõupidamisele (MT Summit) stendiettekandena avaldamiseks esitatud artiklil. Suurem osa siin kirjeldatud praktilisest tööst on minu poolt läbi viidud. Heiki-Jaan Kaalepilt pärinevad mitmed ideed ja soovitused, millest praktilise töö teostamisel lähtusin. Samuti on viiendas peatükis osaliselt kasutatud nimetatud artiklis esitatud töö kirjeldust H.-J. Kaalepi sõnastuses.

[35] Paralleelistuskvaliteedi hindamist puudutavad küsimused kattuvad osaliselt korpustest automaatselt genereeritud mitmekeelsete leksikonide kvaliteedi hindamise problemaatikaga, millest oli lähemalt juttu peatükis 5.5.4.

paralleelistuste kvaliteeti, et teha kindlaks, millist korpust kattuvate osade kaasamisel eelistada. Järgnevalt kirjeldan katset hinnata paralleelkorpuste kvaliteeti poolautomaatselt võrdlemismeetodil. Erialasest kirjandusest ei ole teada uurimistööd, mille käigus oleks varem paralleelkorpusi nende kvaliteedi hindamise eesmärgil võrreldud.

5.1. Korpuste kirjeldused

5.1.1. Maht

Kaks teadaolevalt kõige mahukamat inglise-eesti paralleelkorpust on Tartu Ülikooli inglise-eesti ja eesti-inglise õigustekstide paralleelkorpus (TÜPK) ja Ispras loodud JRC-Acquis'[36] paralleelkorpuse inglise-eesti alamkorpus.

TÜ korpuses on kaks osa, mida eristab lähte-keel. Esiteks eesti seadused ja nende tõlked inglise keelde**,** 150 000 paralleelüksust (lauset või loendi elementi) 400 tekstis. Eesti keeles 1,7 miljonit sõnet , inglise keeles 2,9 miljonit sõnet. Teiseks Euroopa Liidu õigusaktid ja nende eestikeelsed tõlked, 280 000 paralleelüksust (lauset või loendi elementi) 4000 tekstis. Eesti keeles 3,3 miljonit sõnet, inglise keeles 4,9 miljonit sõnet.

JRC-Acquis' korpuse paralleelistatud versioonis on eesti keeles 4,6 miljonit sõnet, inglise keeles 6,8 miljonit sõnet. JRC-Acquis' korpus hõlmab samuti Euroopa Liidu seadusandlikke tekste kokku 21 Euroopa keeles, kusjuures keskmiselt on keele kohta 8,8 miljonit sõnet 7600 tekstis. JRC-Acquis' korpuse paralleelistatud versioonis on eesti keeles 4,6 miljonit sõnet, inglise keeles 6,8 miljonit sõnet. Korpuse paralleelistamata versioonis on korpuste suurused eesti ja inglise keele osas vastavalt 7,2 miljonit ja inglise keeles 9,9 miljonit sõnet. Seega sisaldab paralleelistamata korpuse inglise-eesti alamosa oluliselt roh-kem sõnu kui paralleelistatud versioon.

JRC-Acquis' korpuse paralleelistamata osa sisaldab 236 796 (eesti osa) + 193 260 (inglise osa) paralleelüksust (kokku 430 056 paralleelüksust). TÜ korpuse paralleelistamata

[36] (http://langtech.jrc.it/JRC-Acquis.html

inglise-eesti osa (ilma eesti-inglise osata) hõlmab 176 921 (eesti osa) + 746 219 (inglise osa) paralleelüksust (kokku 923 140 lõiku). Paralleelüksuste suurem arv TÜ korpuses on ilmselt tingitud korpuse paralleelistamata toorversiooni liigendusvigadest (laused olid jaotunud mitme lõigu vahel).

JRC-Acquis' korpuse kõigi keelepaaride paralleelistamiseks saab kasutada kahte paralleelistajat: Vanilla ja HunAlign[37]. JRC-Acquis' korpuses on iga keelepaari kohta keskmiselt 269 148 paralleelüksust. Inglise-eesti keelepaari kohta on Vanilla paralleelistuses 295 189 paralleelset üksust ja HunAligni paralleelistuses 301 647 paralleelset üksust - keskmisest keelepaarist Ispra korpuses on inglise-eesti osa mõlemal juhul suurem.

	TÜ korpus		**JRC-Acquis' korpus**		
paralleelistus	lausetasandil paralleelistatud	lõigutasandil paralleelistatud[38]	paralleelistamata kujul	JRC Hunalign	JRC Vanilla
Sõnede hulk (milj)	12,7	8,2	16	11,5	11,5
paralleelüksuseid	436 043	230 580	430 056	301 647	295 189

Tabel 1. Korpuste suurused

JRC Hunalign – JRC-Acquis' korpuse inglise-eesti alamosa paralleelistatuna Hunaligniga

JRC Vanilla – JRC-Acquis' korpuse inglise-eesti alamosa paralleelistatuna Vanillaga

5.1.2. Paralleelistused

Mõlema korpuse paralleelistamisel on kasutatud keelest sõltumatuid meetodeid. TÜ korpuse paralleelistamisel kasutati Vanillat[39]. Seejuures järgiti sama strateegiat mida Europarli korpuse loomisel (Koehn 2002): kui paralleelistatavate tekstiosade formaalsed struktuurid olid liiga erinevad, siis jäeti vastavad osad korpusest hoopis välja.

[37] http://mokk.bme.hu/resources/hunalign
[38] Ilma eesti-inglise osata
[39] http://nl.ijs.si/telri/Vanilla/

Paralleelistamine toimus kolmes etapis: esiteks peatükkide ja lisade paralleelistamine, seejärel lõikude ja lõpuks lausete. Seejuures kasutati igal etapil kontrollimiseks (mitte paralleelistamiseks) kõige lihtsamaid ankurpunkte: peatükkide, paragrahvide, loendite nummerdust.

Juhul, kui paralleelsetes tekstides paistis olevat erinev arv paragrahve, artikleid või nummerdatud loendite elemente; või kui vastavate järjekorranumbritega elemendid ei olnud kohakuti, siis vastavaid tekste paralleelkorpusesse ei võetud. Põhjuseks oli oletus, et selliste tekstide formaalne struktuur on liiga erinev, et neid antud lihtsa meetodiga töödeldes oleks tulemused usaldusväärsed.

TÜ korpuses võivad eesti- ja ingliskeelsed laused olla omavahel 1-1, 1-2 või 2-1 vastavuses. Muud vastavused on korpusest välja jäetud. (TÜPK)

HunAligni töö jaotub kolme faasi, millest esimene on paralleeltekstide segmentide esialgne üks-ühele paralleelistus, teine faas paralleeltekstide põhjal automaatne leksikoni moodustamine ja kolmas faas paralleelistamine leksikoni abil. (Steinberger jt 2006)

Ei ole päris selge, kas JRC-Acquis' korpuse HunAligni paralleelistusfail on saadud kõiki kolme faasi läbides, kuid tõenäoliselt on siiski kasutatud ka automaatselt genereeritud leksikoni abi, kuna paralleelistusfailis sisalduvate vastavuste seas on ka mitmeid selliseid, kus ühele lõigumärgenditega tähistatud segmendile vastab teises keeles mitu segmenti, näiteks:

```
<link type="2-1" xtargets="7 8;6">
<s1><p>THE COUNCIL OF THE EUROPEAN</p><p>COMMUNITIES,</p></s1>
<s2>EUROOPA ÜHENDUSTE NÕUKOGU,</s2>
</link>
```

Erinevalt TÜ korpusest ei luba JRC-Acquis' korpuse dokumentatsioon väita, et sealt oleks paralleelistusraskuste tõttu midagi puudu. Esinevad ka 0-1 vastavused.

Et TÜ korpuse avalik versioon on paralleelistatud lausetasandil, JRC-Acquis' aga lõigutasandil, siis ei saa võrdluses kasutada TÜ korpuse avalikku versiooni. Selle asemel kasutasin korpuse autoritelt saadud TÜ korpuse versiooni, millel oli paralleelistamise viimane etapp tegemata ja milles olid seega paralleelistatud ainult lõigud.

5.1.3. Paralleelistusvigadele viitavad tunnused

Püüdsin leida korpuste paralleelistusvigu esile toovaid tunnuseid esmalt juhuslikke faile läbi vaadates ja üksteisega võrreldes. Selle tegevuse käigus selgusid järgmised üldisemad tunnused.

TÜ korpuses on kokku 40 086 juhtumit, mil ühe lõiguga on paralleelistatud täpselt sama sisuga teine lõik (vahe on ainult märgendites: <inglise> vs <eesti>). Näiteks:

<inglise> La cantidad exportada no debe superar </inglise>
<eesti> La cantidad exportada no debe superar </eesti>

Enamasti on sellistel juhtudel tegemist mõne kolmanda keele tsitaatidega (mitte inglise ega eesti keeles), mis tähendab, et selliste vastavuste rohkus ei viita mitte vigadele, vaid pigem paralleelistuse korrektsusele. JRC-Acquis' korpuse Vanilla versioonis on selliseid juhtumeid vaid 1260 (HunAligni paralleelistuses 1443), mis võivad olla tingitud JRC-Acquis' korpuse viletsamast paralleelistusest.

Nende lõigupaaride otsing tõi aga välja, et kohati on lõigupaaride kordumine TÜ korpuses tulenenud ka eesti keele paralleelistamisest eesti keelega, seda vähemalt ühel juhul terve faili ulatuses.

JRC-Acquis' korpuse inglise-eesti osa HunAligniga tehtud paralleelistus sisaldab 15 532 juhtumit, mille puhul eestikeelse jaotusega <s2> on paralleelistatud täiesti tühi (inglisekeelset osa tähistav) segment <s1>. Vanillaga tehtud paralleelistuses on <s1> tühi 10 721 juhul. Märgendile '<s2>' vastav jaotus on tühi HunAligni paralleelistuses 2788

juhul, Vanilla paralleelistuses 1475 juhul. TÜ lõppkorpuses analoogseid tühje segmente ei leidu, aga võrdluses kasutatavas lõigutasandil paralleelistatud korpuses on kokku 1453 tühja segmenti.

Korpuse loomist käsitlevas artiklis (Steinberger jt 2006) väidetakse, et kahe paralleelistaja võrdlevat hindamist pole veel teostatud. Vanilla joondamistulemusi on küll võrreldud teiste joondamismeetodite abil saadud tulemustega (Rosen 2005). Alexandr Roseni poolt kirjeldatava eksperimendi tulemusel oli Vanillaga paralleelistamisel võrreldes teiste meetoditega kõrgem saagis (kõigi kasutatud tekstide puhul) ning täpsus tekstide puhul, mis ei sisaldanud „müra" (nt väljajätud, tekstide struktuurilised erinevused või eeltöötlusel tehtud vead). (samas)

Tühjade jaotuste tunduvalt suurem hulk HunAligni paralleelistuses tundub pealiskaudsel vaatlusel olevat argument Vanilla kasuks. Siiski ei saa järeldada tühjade segmentide suuremast arvust paralleelistuse suuremat korrektsust.

JRC-Acquis' korpuse Vanilla paralleelistuse kahjuks räägib asjaolu, et JRC korpuse tegijad ei võtnud korpuse paralleelistamisel Vanillaga arvesse nn "kindlaid" (*reliable*) seoseid paralleeltekstide vahel (näiteks nummerdatud lõigud). HunAligni algoritm aga vaatleb eraldi ka paralleelteksti segmentides sisalduvaid numbrilisi sarnasusi. (Steinberger jt 2006)

Korpuste kattuvate alamosade inglise ja eesti osade mahtu (baitides) failiti võrreldes selgus, et suhteliselt on JRC Vanilla paralleelistuses kahe keele vahelised mahulised erinevused failiti palju suuremad kui TÜ korpuses. See tähendab, et tõenäoliselt on TÜ korpuse paralleelistus korrektsem.

5.2. Eeltöö

Esimene ülesanne kahe korpuse võrdlemisel oli kindlaks teha, millised tekstid esinevad mõlemas korpuses, sest nende võrdlemine võimaldab korpuse hindamist automatiseerida ning annab kõige rohkem infot korpuste tegemise protseduuri mõjust tulemusele.

Selliste tekstide leidmine ei ole üldjuhul lihtne ülesanne, sest tekstide samasuse üle otsustamine on keeruline. Samasisulised tekstid võivad olla erinevalt kujundatud, mõned osad võivad ühes tekstis puududa või olla teistsuguses järjestuses. Näiteks seadusandluse puhul võib kuupäev olla teksti alguses või lõpus, mõni lisa võib hoopis puududa jms.

Teiselt poolt esineb ka juhtumeid, mil näivalt väga sarnased tekstid on tegelikult erinevad, nt. juhul, kui uus seadusandlik akt kordab peaaegu sõna-sõnalt mõnda varasemat, või sisaldab sellest suuremahulisi tsitaate. Seega dokumentide automaatne võrdlemine, ka juhul, kui kasutatakse ligikaudset võrdlemist (Levenšteini distantsi), ei pruugi anda nende samasuse kohta adekvaatseid hinnanguid.

Õnneks on JRC-Acquis' korpuse failides kirjas CELEXi kood – Euroopa Liidu seadusandlike dokumentide identifitseerimiseks kasutatav kood. Kõigil EL tõlgetel on sama CELEXi kood mis originaalil. Nii JRC-Acquis' kui ka TÜ korpuse paralleelistatud kujust sai automaatselt teha sellise, kus iga paralleeltekst on omaette fail, kusjuures failis on kirjas ka dokumendi CELEXi kood[40]. Seega oli võimalik kokku viia JRC-Acquis' korpuse ja TÜ korpuse failid, millel on sama CELEXi kood ehk mis peaksid sisaldama sama sisuga ja ühesuguse paralleelistusega tekste.[41] Selliseid failipaare oli u 2000 ja korpuste edasisel võrdlemisel kasutasingi just neid.

Need failipaarid viisin efektiivsema võrdluse huvides omavahel võimalikult sarnasele kujule. JRC-Acquis' korpuse alamosa teisendasin TÜ korpusega sarnasele SGML-kujule,

[40] Kasutatud skript Unixi-laadses Windowsi-keskkonnas Cygwin: *csplit vanilla_jrc_originaal.xml /.*div.type=.*/ {*}*

[41] Lisaks on suur osa mõlema korpuse tekste tähistatud ka teksti **teemavaldkonna** EUROVOC-koodiga. Nende koodide abil tekstide klassifitseerimisest võib abi olla näiteks eestikeelsete terminoloogiasõnastike täiendamisel.

ühtlustasin mõlema alamosa märgendusskeemi. Et failide võrdlust hõlbustada, nimetasin kõik failid automaatselt nendele vastava CELEXi koodiga.[42]

5.3. Võrdlemine

5.3.1. Metoodika ja algoritm

Võrdlesin omavahel paarikaupa kolme paralleelistatud versiooni eelnimetatud 2000 failist: JRC-Acquis' Vanilla ja HunAligni versioone (edaspidi *JRC Vanilla* ja *JRC HunAlign*) ning TÜ versiooni. Kõigi võrdlemiste metoodika oli ühesugune.

Pärast JRC-Acquis' ja TÜ korpuste märgenduse, täpitähtede kodeeringu ja tühimärkide (tühik, reavahetus, tabulatsioon) ühtlustamist selgus, et tekstid sisaldavad ikka sisu mõttes mitteolulisi erinevusi, nt. mittestandardsel viisil esitatud täpitähed, erinevate sulgude ja kirjavahemärkide kasutamine, erinevused numbrite ja arvude esitusviisis, ebaühtlane suurtähe kasutus või erinevad viisid tähistada tekstidest välja jäetud osi.

Samuti sisaldas mõni fail ühes korpuses rohkem teksti kui sama Celexi koodiga tähistatud fail teises korpuses. Mõnikord oli see tingitud korpuste eri allikatest pärinevate lähtefailide erinevusest, kuid seda võis tingida ka TÜ korpuse paralleelistamismeetod, mis teatud juhtudel raskesti paralleelistatavad osad lihtsalt eemaldas.

Samuti selgus, et korpused sisaldavad lahknevusi ka eestikeelsete sünonüümiliste sõnade ja väljendite kasutamise osas, mis viitab sellele, et tegemist on lihtsalt erinevate tõlgetega või teksti toimetamise erinevate etappidega. Ka need juhtumid kuuluvad sisu mõttes mitteoluliste erinevuste hulka, kui eesmärgiks on paralleelistuse õigsuse kontroll.

---

[42] Selleks kasutatud skript:

```
for file in `ls`
do
mv "$file" `cat $file | grep '<div type=.body' | sed 's/<div type=.body. *n=.//' | sed 's/\" select=\"en et\">//'`
done
```

Tavaline sõnede võrdlemine ei erista juhtumeid, kus lõigud on üksteisest ainult formaalselt, mitte sisuliselt erinevad (tänu kirjavahemärkide või suurtähtede kasutusele) nendest juhtumistest, mil tegemist on hoopis erinevate lõikudega. Paralleelistamise korrektsuse kontrollimisel oleks selline eristamine aga oluline. Lahenduseks on Levešteini distantsi kasutamine ja teatud mittevastavuse protsendi lubamine, mille puhul võib lõike lugeda veel ühesugusteks.

Eesmärgiks oli esiteks välja selgitada, millises ulatuses on failid mõlemas korpuses ühtemoodi paralleelistatud. Ühtemoodi paralleelistused on tõenäoliselt korrektsed, samas kui erinevatest paralleelistustest on vähemalt üks vale. Teine etapp oleks erinevuste analüüs.

Ülaltoodu valguses tuli eri korpuste tekstide võrdlemiseks kasutada järgmist algoritmi:

1. Võrdle tekstide ingliskeelseid originaalosi ja leia ühisosa - kokkulangevate lõikude arv A. (Ei olnud ühtegi teksti, mille originaalversioonid mõlemas korpuses oleksid 100% kokku langenud. Kokkulangevate lõikude protsent oli vahemikus 0…99).

0-1 vastavusi võrdlemisel ignoreerisin, kuid hindamisel lugesin nad mittekorrektsete vastavuste hulka.

Tekstide võrdlemiseks kasutasin Unixi utiliiti *diff*, mis teatavate parameetritega kasutades annab väljundina loendi täht-tähelt kokkulangevatest ridadest ja ridadest, mis mingis ulatuses erinevad. Neid erinevaid ridu võrdlesin omakorda Levenšteini distantsi[43] kasutades, et eristada tegelikke paralleelistuserinevusi juhuslikest vormistuserinevustest.

2. Võrdle tekstide paralleelversioone, kus inglise lõik ja tema tõlge moodustavad ühe terviku, ja leia kokkulangevate vastavuste arv B.

Lähtetekstide ühisosa A leidmisel oli mittekokkulangevate märkide lubatud protsent 2 ja ühesuguste vastavuste B leidmisel 1 (kui lubatud protsent on A ja B leidmisel sama, siis võib juhtuda, et inglise originaalis on mõne rea erinevus failiti suurem kui lubatud protsent,

[43] Levenšteini distants ehk teisenduskaugus on vähim teisendusoperatsioonide arv, mis tuleb teha selleks, et muuta üks sõne teiseks.

tõlkega koos aga väiksem, ja tekib loomuvastane olukord, et õigesti paralleelistatud lõike on rohkem kui ühesuguseid lõike lähtetekstis ehk B > A)

3. Leia paralleelistuse sarnasus C=B/A, s.t. ühesuguselt paralleelistatud lõikude ja ühesuguste originaal-lõikude suhe. C peab olema 0 ja 1 vahel (incl.).

C ei näita, kummas failis on paralleelistus parem ja kummas halvem, ta näitab ainult kahe faili kattuva osa erinevust.

C võtsin indikaatoriks, mille järgi valida, milliseid faile käsitsi läbi vaadata. Oletasin, et mingid esialgu tundmatud tegurid on paralleelistamise algoritme mõjutanud. Neid tegureid võis olla rohkem kui üks ja nende mõju eri meetoditele võis olla erineva tugevusega, seetõttu kontrollisin tekste, mis jäid C erinevatesse vahemikesse.

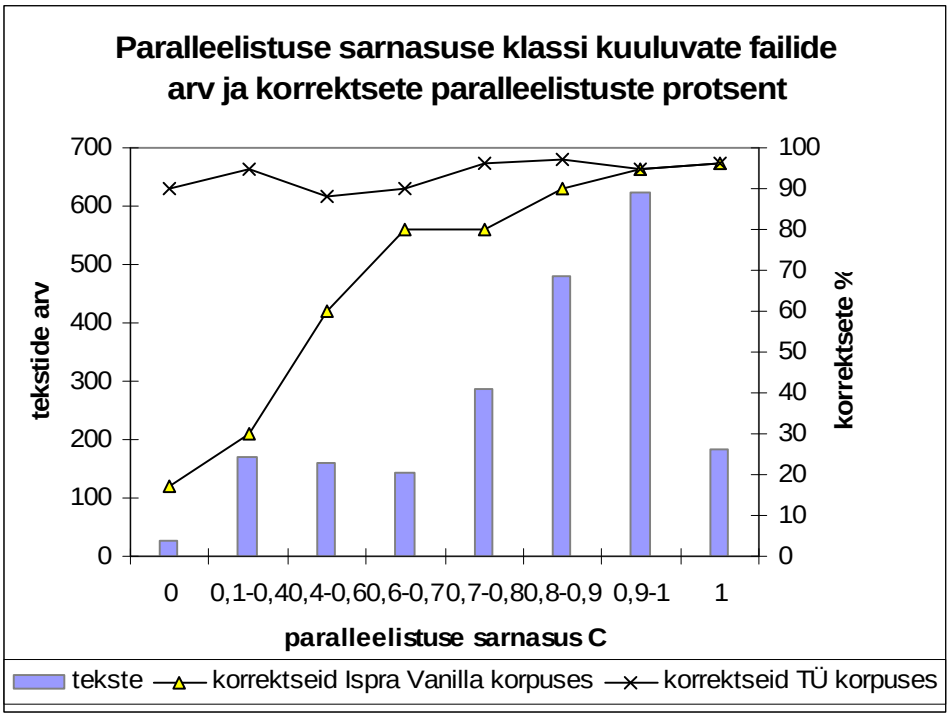
5.3.2. JRC Vanilla versiooni ja TÜ korpuse võrdlus

Jagasin failid nende paralleelistuse sarnasus C suuruse järgi 8 gruppi (vt joonist 1). Vaatasin läbi vähemalt 5% igast C vahemikku jäävatest failidest, et saada vastus küsimusele: milline on TÜ ja JRC-Acquis' korpuses korrektselt paralleelistatud lõikude protsent? Lisaks C-le on oluliseks indikaatoriks identsete ridade protsent failipaariti, mis näitab, kui suurel määral C-d antud faili puhul usaldada saab. Identsete ridade protsenti võtsin arvesse üldistuste tegemisel nende failide kohta, mis jäid läbi vaatamata.

Koostasin läbivaadatud failide põhjal tabeli korrektsete vastavuste protsendi hinnanguliste väärtustega. Iga faili korrektsete vastavuste protsendi algväärtuseks seadsin vastava failigrupi hinnangulise keskmise, välja arvatud läbivaadatud failid, mille puhul märkisin tabelisse täpse protsendi. Kuna täheldasin, et tühjad segmendid tähistavad paralleelistusvigu, siis korrigeerisin seda hinnangulist algväärtust automaatselt väiksemaks sõltuvalt tühjade segmentide protsendist antud failis. Kontrollisin hinnanguliste väärtuste keskmist failigrupiti läbivaadatud failide vastavate keskmistega, erinevuste korral korrigeerisin hinnangulisi väärtuseid vajalikus suunas.

Joonis 1 iseloomustab korrektsete vastavuste protsenti mõlemas korpuses, sõltuvalt failide paralleelistuste sarnasusest C. Näeme, et TÜ ja JRC Vanilla paralleelistuse erinevuse

korral on TÜ korpuse paralleelistus korrektsem. JRC-Acquis' korpuse alaosa korrektsete paralleelistuste protsent on üsnagi suures korrelatsioonis C väärtusega, TÜ korpuste korrektsete paralleelistuste protsenti mõjutab C vähem. Antud joonisel on esitatud keskmised statistilised väärtused, üksikute failide puhul võib siiski juhtuda, et JRC Vanilla paralleelistus on korrektsem.



Joonis 1.

Tuginedes käsitsi läbi viidud võrdlemisele võib väita, et TÜ korpuses tervikuna on korrektsete paralleelistuste protsent ca 95 ja JRC-Acquis' korpuse Vanilla paralleelistuses ca 84.

5.3.3. 0-vastavuste protsent

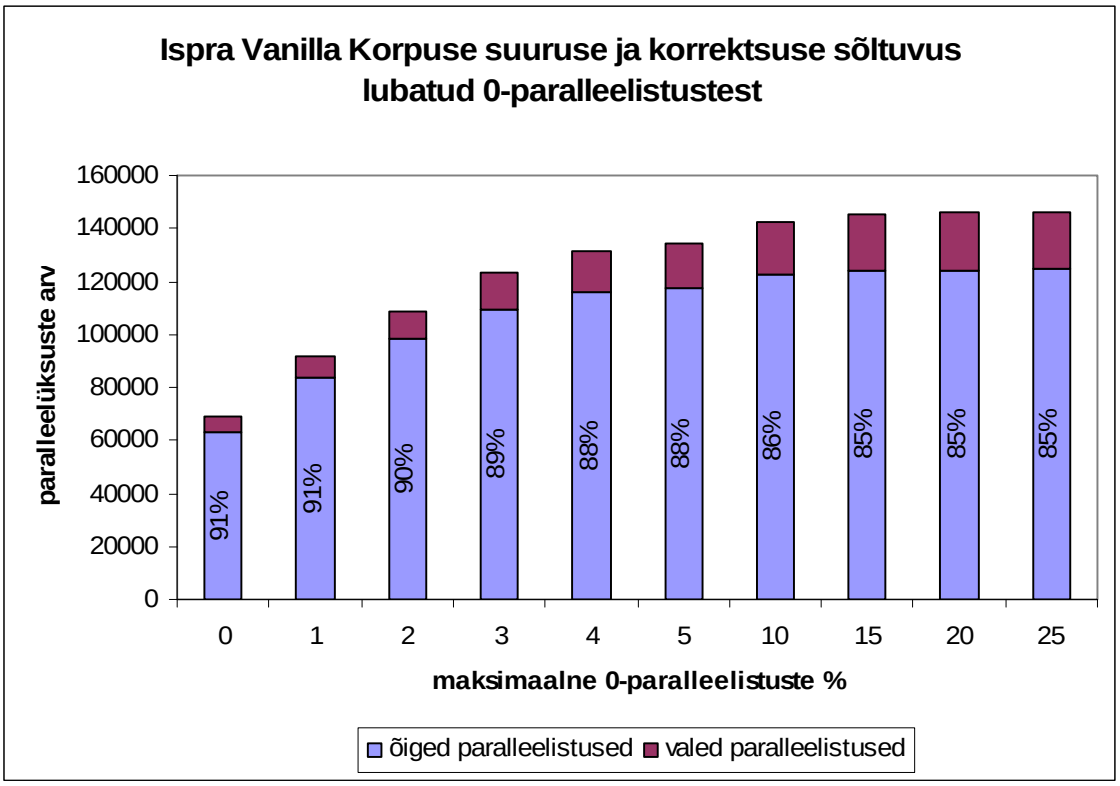
Kas on mingit viisi kindlaks teha teksti paralleelistamise korrektsust ilma teda teise korpuse tekstiga võrdlemata? Sellise tunnuse leidmine võimaldaks korpuste ühendamisel, mil vastav tekst teises korpuses puudub, vigased tekstid välja jätta.

Lähtudes paralleelistusalgoritmide tööpõhimõttest võiks oletada, et selliseks tunnuseks sobib 0-vastavuspiirkondade protsent: mida rohkem esineb juhtumeid, mil paralleelistusalgoritm pole originaalile või tõlkele vastavust leidnud, seda tõenäolisem on, et ka leitud vastavused on valed.

TÜ korpuse andmete põhjal arvutatuna oli kovariatsioon tühjade ridade protsendi ja korrektsete vastavuste hinnangulise protsendi vahel 0,33 ja korrelatsioon 0,02. S.t. seos puudub. Ilmselt on siin põhjuseks TÜ korpuse paralleelistamisel kasutatud algoritm, mis enne väiksemate üksuste paralleelistamise juurde asumist eemaldas eelmisel etapil loodud 0-vastavused.

JRC-Acquis' korpuses oli kovariatsioon -53,82 ja korrelatsioon -0,42. S.t. eksisteerib keskmise tugevusega seos. Seega JRC-Acquis' korpuse puhul võib 0-vastavuste protsenti kasutada paralleelistuse kvaliteedi indikaatorina.

Joonis 2 kirjeldab JRC Vanilla korpuse (õigemini selle TÜ korpusega võrreldud osa) põhjal loodavate alamkorpuste suurusi ja korrektsusi, kui seada erineva ulatusega piiranguid 0-vastavusi sisaldavate tekstide korpusesse võtmisel. 25%-st leebemad piirangud ei avalda korrektsusele ega korpuse mahule erilist mõju, sest vastavate failide arv on väike. Usutavasti kehtivad samad proportsioonid ka JRC-Acquis' korpuse selle osa kohta, mida pole TÜ korpusega võrreldud.



Joonis 2.

5.3.4. JRC HunAligni ja TÜ korpuse võrdlus

JRC HunAligni korpuse failide paralleelsuse sarnasuse määr C TÜ korpuse vastavate failidega oli keskmiselt suurem kui JRC Vanilla ja TÜ korpuse vahel.

Juhuslike failide läbivaatamisel selgus, et kui HunAlign on leidnud vastavuse 0-1, siis antud korpuses on viga nii selles vastavuses kui ka eelmises või järgmises. 95% käsitsi kontrollitud failidest oli HunAligni paralleelistus väljaspool nimetatud 0-1 veapiirkonda täiesti korrektne. Paraku on 0-1 vastavuste hulk küllaltki suur, nii et keskmine hinnanguline korrektsete paralleel-lõikude protsent antud korpuses on 94, mis on praktiliselt võrdne TÜ korpuse 95-ga.

Erinevalt JRC Vanilla versioonist ei olnud HunAligni versioonis märgatavat seost 0-vastavuste protsendi ja sama faili teiste vastavuste korrektsuse vahel.

TÜ paralleelkorpuse veebilehel[44] korpuse koostamise protseduuri kirjelduses on muuhulgas öeldud:

„Mõnikord juhtus, et paralleelistatavate üksuste arv paralleelsetes tekstides on liiga erinev (s.t. ühes tekstis on paralleelistatavaid üksusi enam kui kaks korda rohkem kui teises), et tulemus võiks olla mõistlik. Näiteks kui ühes tekstis on ühes artiklis 1 lõik ja teises 5, siis on ilmselt tegemist kas elementide (nt tabelite või viidete) ärajäämisega ühest tekstist või kujundusliku erinevusega (nt ühes tekstis on loetelu esitatud ühel real, teises aga mitmel). Sellisel juhul jäeti antud artikli lõigud paralleeltekstist lihtsalt välja, kuid teksti ülejäänud osad võeti siiski sisse." (TÜPK)

Selliseid üks-mitmele vastavusi, mis TÜ korpusest välja jäeti, esineb ka JRC Hunaligni versioonis. Kas need lõigud tuleks samuti TÜ korpusega liidetavast JRC-Acquis' korpuse osast välja jätta?

Sellele küsimusele vastamiseks vaatasin läbi JRC Hunaligni versioonis kümme juhuslikku 1-3 (st ühele inglisekeelsele lõigule on asetatud vastavusse kolm eestikeelset lõiku) vastavust. Kokku oli selliseid vastavusi 1565. Läbivaadatud vastavustest kaheksa olid sisuliselt korrektsed vastavused ja kaks osaliselt ebakorrektsed[45]. Kümnest juhuslikust 1-4 vastavusest (kokku 187) olid neli täielikult või osaliselt ebakorrektsed. Kümnest juhuslikust 1-5 vastavusest (kokku 152) olid kõik kümme osaliselt ebakorrektsed. Mitmetel juhtudel on inglisekeelse lõiguga paralleelistatud rohkem kui 5 eestikeelset lõiku.

Kümnest juhuslikust 3-1 vastavusest (kokku 180) olid kõik kümme osaliselt ebakorrektsed. Kümne juhusliku 4-1 vastavuse (kokku 42) seas leidus üks korrektne vastavus, ülejäänud vastavused olid siiski ebakorrektsed. Kümnest juhuslikust 2-2 vastavusest (kokku 22) olid kolm korrektsed, ülejäänud ebakorrektsed.Võib oletada, et ka 1-2 vastavuste veaprotsent JRC korpuses on mõnevõrra suurem võrreldes 1-1 vastavustega.

[44] http://www.cl.ut.ee/korpused/paralleel/index.php?lang=et

[45] Ebakorrektseteks nimetan paralleelistusi, mille üks pool sisaldab originaallõiku või -lõike, millele paralleelistuse teises pooles ei leidu sisulist vastet.

Korrelatsiooni üks-mitmele (1-2 ja 2-1) vastavuste sageduse ja veaprotsendi vahel on tuvastatud ka araabia-hispaania-inglise paralleelkorpuse puhul (Samy jt 2006).

5.3.5. JRC-Acquis' korpuse Vanilla ja HunAligni versioonide võrdlus

JRC-Acquis' korpuse kahe erineval meetodil paralleelistatud failide võrdlemisel kasutasin varemkirjeldatud metoodikat: algul leidsin paralleelistuse sarnasuse määra C ja seejärel võrdlesin valikuliselt faile igast C vahemikust. Erinevus oli ainult selles, et lähtetekstide originaalid langesid alati 100% kokku.

Veerand võrreldavast 2000 failipaarist olid täiesti ühesugused. Ülejäänutest oli HunAligni paralleelistus alati korrektsem kui Vanilla abil tehtu. See kinnitab JRC HunAligni versiooni ja TÜ korpuse võrdlusel tehtud hinnangut JRC HunAligni versiooni ja TÜ korpuse korrektsusetaseme sarnasuse kohta.

6. Kokkuvõte

Lausete paralleelistamise meetodite võrdlemise suhtes ollakse seisukohal, et praegu pole võimalik välja tuua ühte ja parimat meetodit (Rosen 2005; Singh and Husain 2005), mis ületaks teisi mistahes korpusel. Samuti ei ole õigustatud paralleelistusmeetodi ühe korpuse peal saadud täpsuse ja saagise ülekandmine teise korpuse iseloomustamiseks, isegi kui seal on kasutatud sama meetodit (Singh, Husain 2005).

Viiendas peatükis kirjeldatud võrdlus näitas veelkord, et ka juhul, kui sama korpuse peal kasutatakse sama meetodit (Vanilla), võivad tulemused olla radikaalselt erinevad, kui korpused on erinevalt normaliseeritud või meetodit rakendatakse veidi erineval moel. Samuti näitas võrdlus, et eri kohtades loodud, sama sisu, kuid erinevatest allikatest pärinevate ja erinevate meetoditega paralleelistatud korpusi on võimalik kasutada korpuste endi kvaliteedi hindamisel.

Võrdlemise tulemusel saab väita järgmist:

- JRC-Acquis' korpuse HunAligni versioon on palju korrektsem kui Vanilla versioon.
- JRC-Acquis' HunAligni versioonis 0-vastavus +/- üks paralleelne rida on suure tõenäosusega ebakorrektsed ja need võib eemaldada.
- Suure tõenäosusega on JRC -Acquis' HunAligni versioonis ebakorrektsed ka vastavused, kus ühe inglisekeelse lõiguga on paralleelistatud rohkem kui neli lõiku, ja need võib eemaldada. Ülejäänud üks-mitmele vastavuste veaprotsent vajab täpsemat määratlemist.
- Võrreldud korpused erinesid mitte ainult oma vastavuste ja tõlkeversioonide poolest, vaid ka lähtetekstide osas – polnud ühtegi ingliskeelset teksti, mis oleks formaalselt 100% kokku langenud oma ingliskeelse vastega teises korpuses.

- Tekstide automaatseks võrdlemiseks kasutasin vastavuste sarnasuse mõõtu C. C arvestamine võimaldas vähendada käsitsi läbivaatamise vajadust.
- Ankurpunktide arvestamine, olgu see siis paralleelistamisel (Hunalign) või vahepealse tulemuse filtreerimisel (TÜ versioon Vanillast) on isegi formaalselt lihtsate tekstide puhul oluline – vigade arv väheneb 10%, võrreldes ankrupunkte mittearvestava meetodiga (JRC-Acquis Vanilla).
- Leitud hinnangud korpustele võivad seletada samade korpuste peal tehtud (Fishel jt 2007) statistilise masintõlke eksperimentide tulemuste erinevusi.
- Leitud tunnused (tühjade ridade protsent failis JRC-Acquis Vanilla puhul, tühjade vastavuste olemasolu JRC-Acquis Hunaligni puhul) võimaldavad korpusse lisamisel eelistada usaldusväärsemaid vastavusi.
- Kirjeldatud võrdlusmeetodid peaksid olema rakendatavad ka muude korpuste puhul, mille jaoks eksisteerib sarnaseid, osalise kattuvusega paralleelkorpusi, nt sloveeni-inglise Euroopa seadusandluse korpused SVEZ-IJS ja JRC-Acquis' korpus (Erjavec 2006).

Viiendas peatükis kirjeldatud tööst selgub, et nii leksikoni genereerimiseks kui ka muudeks rakendusteks oleks otstarbekam JRC-Acquis' korpuse Vanilla versiooni asemel kasutada HunAligni versiooni ühendatuna TÜ paralleelkorpusega. Kuigi Fishel jt (2007) leidsid, et ka Vanilla versiooni ühendamine TÜ paralleelkorpusega aitab eesti-inglise statistilise masintõlke tulemusi mõnevõrra parandada, on siiski väga tõenäoline, et Vanilla versiooni asendamine HunAligni versiooniga oleks ka masintõlke puhul vajalik.

Lisaks TÜ paralleelkorpusele ja JRC-Acquis' paralleelkorpusele on koostatud veel üks tekstimahukas mitmekeelne paralleelkorpus – OPUS (Tiedemann, Nygaard 2004) – mis samuti sisaldab eestikeelset osa. Muuhulgas sisaldab OPUS 115 000 sõne ulatuses eestikeelset Euroopa Liidu seaduseandlikku teksti (ja eestikeelse tekstiga paralleelistatud inglisekeelset teksti), mille ühendamine TÜ paralleelkorpusega oleks samuti soovitatav. Kõnealuse OPUSe osa ühendamisel TÜ paralleelkorpusega on tõenäoliselt vähemalt osa-

liselt võimalik lähtuda eelnevas peatükis kirjeldatud kahe korpuse ühendamisele eelnenud tööst korpuste võrdlemisel ja hindamisel.

Arvesse võttes kahe paralleelistusmeetodi võrdlustulemusi (vt osa 5.3.5) võib oletada, et OPUSe osa ühendamisel TÜ paralleelkorpusega ei tohiks kasutada korpuse kodulehel[46] saadaval olevaid Vanillaga paralleelistatud inglise-eesti tekste. Selle asemel tuleks kasutada paralleelistamata tekste ja paralleelistus teostada HunAligni või mõne teise meetodi abil.[47]

Lähemas tulevikus oleks vaja niisiis TÜ paralleelkorpus ühendada JRC-Acquis' korpuse HunAligni versiooniga ning OPUSe inglise-eesti seadusandlikke tekste hõlmava osaga. Probleeme tekitab seejuures samasuguste CELEXi koodidega failide olemasolu mõlemas korpuses. Tuleb välja töötada automaatne meetod, mille abil iga lõigupaari eraldi "vaadeldes" toimub teatud tunnuste abil kattuvate lõikude tuvastamine ja selliste lõigupaaride puhul kas ühe või teise korpuse lõigu eelistamine. Reeglina tuleks eelistada JRC HunAligni versiooni, kuna pärast 0-vastavuste ja vigaste üks-mitmele vastavuste eemaldamist on HunAligni paralleelistus suure tõenäosusega TÜ paralleelkorpuse paralleelistusest korrektsem.

Eesti keelt hõlmavate mitmekeelsete sõnastike koostamine ja täiustamine on üks peamisi teid tulevikus eesti keele väljatõrjumise takistamiseks mitmetest eluvaldkondadest. Seetõttu osutub väga oluliseks ka magistritöö neljandas peatükis tutvustatud töö jätkamine praktilisel tasandil leksikograafide ja keeletehnoloogide poolt.

Esialgsed tulemused paralleelkorpustest leksikonide ekstraheerimise vallas jätavad veel kõvasti soovida – kontrollimisel osutus kõige suurema sisulise täpsusega genereeritud leksikoniks TÜ paralleelkorpuse põhjal genereeritud eesti-inglise leksikon, mille juhuslikult valitud 50 kirjelise fragmendi kontrollimine andis selle osa täpsuseks 87 %; teiste leksikonide täpsus jäi kontrollimisel eesti-inglise leksikonile suuresti alla. Siiski peaks kõik mitmekeelsete leksikonide koostamisega tegelevad inimesed mõtlema selle peale, kuidas

[46] http://logos.uio.no/opus/EUconst.html
[47] Samamoodi talitati OPUSe tekstide integreerimisel tšehhi-inglise korpusse (vt http://ufal.mff.cuni.cz/czeng/)

paremini ammutada automaatsete vahenditega leksikone rikastavat infot nii tava- kui paralleelkorpustest.

Leksikoni genereerimise katset PWA-ga tuleks korrata loodava ühendparalleelkorpuse peal, kuna parem paralleelistuse kvaliteet tagab kindlasti korrektsema leksikoni. Võrreldes erinevaid leksikonigenereerimismeetodeid tuleks leida inglise-eesti keelepaarile sobivaim ning sellest lähtuda spetsiaalselt sellele keelepaarile mõeldud leksikonigenereerimistarkvara loomisel.

THE ROLE OF PARALLEL CORPORA IN COMPUTATIONAL LINGUISTICS: COMPARISON OF PARALLEL CORPORA AND GENERATION OF BILINGUAL LEXICONS FROM PARALLEL CORPORA

Summary

In addition to contrastive studies of languages or language variants, parallel/comparative corpora have many other uses both in theory and practice, while the potential of some of such uses is still awaiting discovery. One of the most interesting trends involves dictionary compilation or revision by means of extracting translation equivalents.

The present master's thesis attempts a survey of what has been done, with a view to some possible practical applications to Estonian in the future. For example, a simple lexicographic device has been outlined to enable the lexicographer to generate a list of translation equivalents by using a parallel text. Also, a report of attempts is presented to generate source material for an English-Estonian Estonian-English technical dictionary using not only the parallel corpus but also some free software, which needs little additional language resources beside the corpus material.

For the time being multilingual technical dictionaries can be compiled from parallel corpora only semi-automatically, because without intervention on the part of a human proofreader the method would yield but raw material to help lexicographers, terminologists or translation systems.

There is no software that could perform alignment and dictionary generation on the basis of the Estonian grammatical structure in correlation with the structure of some other language. Although the quality of the lexicon to be generated would certainly be improved

by preliminary morphological analysis of the parallel corpus, my present attention has been focused on language independent approaches to dictionary extraction. The word aligners UWA and LWA developed by Swedish researchers within the Plug project (Tiedemann 2002) use relatively little language-specific information, which makes them easily applicable in automatic generation of dictionaries containing Estonian material.

The master's thesis describes an attempt to develop source material for a technical dictionary by means of UWA and LWA, drawing on the English-Estonian parallel corpus of the University of Tartu and the English-Estonian subsection of the JRC-Aqcuis multilingual parallel corpus. One of the dictionaries so generated contains 130 865 entries and 482 571 word forms. The precision of a random sample of 50 entries turned out to be 60%. In addition the thesis provides a survey of the working principles of the used programmes, and some suggestions on how to improve UWA results with a view to an analogous device to be possibly developed for Estonian.

The availability of partially overlapping parallel corpora for a language pair opens up opportunities for automatically comparing, evaluating and improving them. I compare and evaluate the alignment quality of two English-Estonian parallel corpora that have been created independently, but contain overlapping texts. I describe how to determine the overlapping parts and find their alignment similarities that allow us to economize on manual evaluation effort. A feature is also suggested that could be used instead of comparing and manual checking to predict the alignment correctness.

While it has been common practice to compare and evaluate alignment quality when describing alignment methods, to my knowledge this is the first time when the alignments of completed parallel corpora are compared and evaluated.

It appears that corpora that have been created independently, containing essentially the same texts from independent sources, and which have been aligned with different methods, can be used for evaluating the alignment quality of the corpora themselves.

It appeared that the corpora compared were different not only in their translation versions and alignments, but in their source text parts as well. There was not a single text,

the English part of which had completely coincided with that of the corresponding one from the other corpus.

When comparing the corpora, I used the alignment similarity measure C and this allowed me to economize on manual evaluation.

The percentage of correctly aligned paragraphs ranged from 84% in JRC-Acquis Vanilla version to 94-95% in JRC-Acquis HunAlign version and the UT corpus.

The different levels of alignment quality may explain some of the differences in results by (Fishel et al, 2007), observed in statistical MT experiments conducted on different corpora.

The method used for determining C can be used also for selecting the most trustworthy alignments from the combination of two corpora.

The features that predict the quality of alignments – the proportion of 0-alignments in case of JRC Vanilla, and the mere existence of 0-alignments in case of JRC HunAlign – allow one to select the aligned units that are more trustworthy, even in the absence of a comparable text from another corpus.

The evaluation results of JRC-Acquis corpus might be transferable to other language pairs of the corpus; this assertion needs further investigation, however.

Kirjandus

Abeillé, Anne (ed.) 2003. Building and Using Parsed Corpora. Series: Text, Speech and Language Technology, vol. 20. Dordrecht: Kluwer, 440 p.

Ahrenberg, Lars; Andersson, Mikael; Merkel, Magnus 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. – Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics. Montréal, Canada, 10–14 August 1998, 29–35.

Ahrenberg, Lars; Andersson, Mikael; Merkel, Magnus 2000a. A knowledge-lite approach to word alignment. –Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 97–116.

Ahrenberg, Lars; Merkel, Magnus; Sågvall Hein, A.; Tiedemann, Jörg 2000b. Evaluation of Word Alignment Systems. – Proceedings of LREC 2000, Athens/Greece, Volume III, 1255-1261.

Blank, Ingeborg 2000. Terminology extraction from parallel technical texts. – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 237–252.

Bojar, O.; Žabokrtský, Z. 2006. CzEng: Czech-English Parallel Corpus, Release version 0.5. Prague Bulletin of Mathematical Linguistics, 86. (in print); http://ufal.mff.cuni.cz/~zabokrtsky/papers/pbml-06-OB-ZZ.pdf (24.05.2007).

Borin, Lars 1998. Linguistics isn't always the answer: Word comparison in computational linguistics. – Proceedings of the 11th Nordic Conference on Computational Linguistics (NODALIDA), University of Copenhagen, Denmark, 140–151.

Borin, Lars 2002. …and never the twain shall meet? – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 47–59.

Botley, Simon Philip; McEnery, Anthony Mark; Wilson, Andrew (ed.) 2000. Multilingual Corpora in Teaching and Research. Amsterdam: Rodopi, 214 p.

Bowker, Lynne; Pearson, Jennifer 2002. Working with specialized language: A practical guide to using corpora. London/New York: Routledge, 242 p.

Brew; C.; McKelvie, D. 1996. Word-pair extraction for lexicography. – Proceedings of the Second International Conference on New Methods in Language Processing. K. Oazer; H. Somers, (ed.). Ankara: Bilkent University, 45–55; http://citeseer.ist.psu.edu/brew96wordpair.html (21.08.2006).

Brown, Peter F.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Frederick; Lafferty, John D.; Mercer, Robert L.; Roossin, Paul S. 1990. A statistical approach to machine translation. – Computational Linguistics, 16(2), Cambridge: MIT Press, 79–85.

Brown, Peter F.; Lai, Jennifer; Mercer, Robert L. 1991. Aligning sentences in parallel corpora. – Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 169–176.

Brown, Peter F.; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Mercer, Robert L. 1993. The mathematics of statistcal machine translation: Parameter estimation. – Computational Linguistics, 19(2), Cambridge: MIT Press, 263–311.

Brown, Ralf D. 1997. Automated dictionary extraction for „knowledge-free" example-based translation. – Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation (T MI-97); http://citeseer.ist.psu.edu/brown97automated.html (21.08.2006).

Carl, Michael; Way, Andy (ed.) 2003. Recent advantages in example-based machine translation. Dordrecht: Kluwer Academic Publishers, 520 p.; http://www.cnts.ua.ac.be/~walter/papers/2004/d04.pdf (26.05.2007).

Charitakis, Konstantinos 2007. Using parallel corpora to create a Greek-English dictionary with Uplug. – NODALIDA 2007 Conference Proceedings (to appear). Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (eds.), 212–215.

Church, Kenneth W.; Gale, William A.; Hanks, Patrick; Hindle, Donald 1991. Using statistics in lexical analysis. – Exploiting On-Line Resources to Build a Lexicon. Zernik, Uri (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, 115–164.

Cmejrek, Martin; Curin, Jan; Havelka, Jiri 2003. Treebanks in machine translation. – Proc. Of the 2nd Workshop on Treebanks and Linguistic Theories, Växjö, 209–212.

Davis, Mark; Dunning, Ted; Ogden, Bill 1995. Aligning noisy corpora. – Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics, Belfield, Dublin: University College Dublin, 67–74.

Davis, M. W. 1998. On the effective use of large parallel corpora in crosslanguage text retrieval. –, Cross-Language Information Retrieval. Grefenstette, G. (ed.). Boston: Kluwer Academic Publishers, 11-22.

Déjean, Hervé; Gaussier, Eric; Sadat, Fatia 2002. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. – Proc. of COLING, Tapei, Taiwan, 218-224; http://www.xrce.xerox.com/Publications/Attachments/2002-025/dejean.pdf (21.08.2006).

Dien, D. 2002. Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. – Proceedings of Workshop on Machine Translation in Asia, COLING-02, Taiwan, 9/2002, 26-32; http://acl.ldc.upenn.edu/W/W02/W02-1607.pdf (26.05.2007).

Ebeling, Jarle 1998. Contrastive linguistics, translation, and parallel corpora. – *Meta*, 43:4, 602–615.

van der Eijk, Pim 1993. Automating the Acquisition of Bilingual Terminology. – Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, 113–119.

Erjavec, Tomaž 2006. The English-Slovene ACQUIS corpus. – Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006. Genoa, Italy, 2138-2141; http://nl.ijs.si/svez/bib/svez-lrec06.pdf (26.05.2007).

Fishel, Mark; Kaalep, Heiki-Jaan; Muischnek, Kadri 2007. Estonian-English statistical machine translation: the first results. – Proceedings of NODALIDA. Tartu (to appear).

Fung, Pascale 2000. A statistical view on bilingual lexicon extraction – from parallel corpora to non-parallel corpora. – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 219–236; http://citeseer.ist.psu.edu/fung98statistical.html (21.08.2006).

Fung, Pascale; Church, Kenneth W. 1994. K-vec: a new approach for aligning parallel texts. – Proceedings of the 15th International Conference on Computational Linguistics. Kyoto, Japan, 1096–1102.

Gale, William; Church, Kenneth. 1993. A program for aligning sentences in bilingual corpora. – Computational Linguistics 19(1), Cambridge: MIT Press, 75–102.

Gaussier, E.; Hull, D.; Ait-Mokhtar, S. 2000. Term alignment in use: machine-aided human translation, – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 253–274.

Geisler, Christer 2002. Reversing a Swedish-English dictionary for the Internet. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 122–133.

Gellerstam, Martin 1994. Translations as a source for cross-linguistic studies. – Aijmer, Karin; Altenberg, Bengt; Johansson, Mats (ed.): Languages in contrast: papers from a symposium on text-based cross-linguistic studies, Lund: Chartwell-Bratt, 53-62.

Genereeritud leksikonid; http://www.teataja.ee/leksikonid.zip (29.08.2006).

GIZA++: Training of statistical translation models; http://www.fjoch.com/GIZA++.html (21.08.2006).

Gliozzo, A.; Strapparava, C. 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. – Proc. of the ACL Workshop on Building and Using Parallel Texts (in conjunction of ACL-05), University of Michigan, Ann Arbor, 9–16.

Godwin-Jones, Bob 2001. Emerging technologies. Tools and trends in corpora use for teaching and learning. – Language Learning & Technology . Vol. 5, No. 3, September 2001, 7–12.

Hiemstra, Djoerd 1997. Deriving a bilingual lexicon for cross-language information retrieval. – M. Heemskerk; M. Diepenhorst (eds.), Proceedings of the fourth Groningen International Information Technology Conference for Students, 21–26.

Hofland, Knut; Johansson, Stig 1998. The Translation Corpus Aligner: a program for automatic alignment of parallel texts. – Stig Johansson, Signe Oksefjell (ed.), Corpora and Cross-Linguistic Research: Theory, Method and Case Studies. Amsterdam: Rodopi. 87–100.

Hwa, Rebecca; Madnani, Nitin 2004. The UMIACS word alignment interface; http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm (21.08.2006).

Johansson, Stig 2002. Towards a multilingual corpus for contrastive analysis and translation studies. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 47–59.

Ide, Nancy; Erjavec, Tomaz; Tufiş, Dan 2002. Sense discrimination with parallel corpora. – Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and

Future Directions. ACL2002, Philadelphia, 56–60; http://www.racai.ro/~tufis/Selected_Papers/sense-discrimination.pdf (21.08.2006).

Index of /telri/Vanilla; http://nl.ijs.si/telri/Vanilla/ (21.08.2006).

Index of JRC-Acquis/alignments; http://wt.jrc.it/lt/Acquis/JRC-Acquis.2.2/alignments/index.html (24.05.2007).

Inglise-eesti-inglise sõnastik; http://www.eki.ee/dict/inglise/ (21.08.2006).

TÜPK = Inglise-eesti ja eesti-inglise paralleelkorpus; http://test.cl.ut.ee/korpused/paralleel/ (24.05.2007).

Kay, Martin 1991. Text-translation alignment. ACH/ALLC '91: „Making Connections" Conference Handbook. Tempe, Arizona.

Kitsnik, Mare 2006. Keelekorpused ja võõrkeeleõpe. – Eesti Rakenduslingvistika Ühingu aastaraamat 2 (2006). Toim Helle Metslang; Margit Langemets, Tallinn: Eesti Keele Sihtasutus, 93–1007.

Koehn, Philipp 2002. Europarl: A multilingual corpus for evaluation of machine translation. Draft, unpublished; http://people.csail.mit.edu/~koehn/publications/europarl.ps (24.05.2007).

Koehn, Philipp 2003. Noun Phrase Translation. PhD thesis, University of Southern California; http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/thesis-readable.ps (24.05.2007).

Koit, Mare 2003. Masintõlge ja kus temast kasu on? - Arvutimaailm, 2003, nr 4, 51–55; http://www.am.ee/6569 (21.08.2006).

Kuhn, Jonas 2004. Experiments in parallel-text based grammar induction. – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics: ACL 2004, 470–477.

Langlais, P., Simard, M., & Véronis, J. 1998. Methods and practical issues in evaluating alignment techniques. – Joint 17th International Conference on Computational Linguistics (COLING'98) and 36th Annual Meeting of the Association for Computational Linguistics (ACL'98), Montréal, 10-14.

Linear B. Word alignment tool; http://demo.linearb.co.uk:8080/sandbox/start.jsp (21.08.2006).

Martin, J.; Johnson, H.; Farley, B.; Maclachlan A. 2003. Aligning and using an english-inuktitut parallel corpus. – Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond. Edmonton, Canada, 115-118; http://acl.ldc.upenn.edu/W/W03/W03-0320.pdf (26.05.2007).

McEnery, A.M.; Piao, Scott; Xin, Xu 2000, Parallel alignment in English and Chinese. – Multilingual Corpora in Teaching and Research. S.P. Botley, A.M. McEnery and A. Wilson (eds.). Rodopi, Amsterdam – Atlanta, 177–191.

Melamed, Dan 1997. A word-to-word model of translational equivalence. – Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Madrid, Spain, 490–497.

Melamed, Dan 2001. Empirical methods for exploiting parallel texts. Cambridge, Massachusetts: MIT Press, 198 p.

Melby, Alan K. 2000. Sharing of translation memory databases derived from aligned parallel text. – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 347–368.

Moore Robert C. 2002. Fast and accurate sentence alignment of bilingual corpora. – Machine Translation: From Research to Real Users (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California). Heidelberg, Germany: Springer-Verlag, 135-244.

Muischnek, Kadri 2006. Improving the quality of the statistical machine translation with linguistic preprocessing: the case of particle verbs in Estonian; http://www.id.cbs.dk/~dh/ngslt/projects/muischnek_final_paper.doc (21.08.2006).

Muischnek, Kadri; Orav, Heili; Kaalep, Heiki-Jaan; Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Toim. U. Talvik. Tallinn: Eesti Keele Sihtasutus.

Multext-East home page; http://nl.ijs.si/ME/ (21.08.2006).

Müller, Karin 2005. Revealing phonological similarities between related languages from automatically generated parallel corpora. – Proceedings of the ACL Workshop on Building and Using Parallel Texts. ACL, 33–40 ; http://acl.eldoc.ub.rug.nl/mirror/W/W05/W05-0805.pdf (26.05.2007).

Nerbonne, John; Heeringa, Wilbert 1997. Measuring dialect distance phonetically. – Proceedings of the third meeting of the SIGPHON at ACL, 11–18.

Nerbonne, John 2000. Parallel texts in computer-assisted language learning. – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 354-369.

Nivre, Joakim; De Smedt, Koenraad; Volk, Martin. 2005. Treebanking in Northern Europe: A white paper. – Henrik Holmboe (ed.), Nordisk Sprogteknologi 2004. Arbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Copenhagen: Museum Tusculanums Forlag, 97–112; http://ling.uib.no/~desmedt/trepil/whitepaper/whitepaper-yearbook2004.pdf (26.05.2007).

Oakes, M.P.; McEnery, A.M. 1998. Bilingual text alignment: an overview. – McEnery, A.M, Botley, S.P and Wilson, A. (eds), Multilingual Corpora in Teaching and Research, Amsterdam: Rodopi, 1-37.

Och, Franz Josef; Ney, Hermann 2000. Improved statistical alignment models. – Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, China, 440–447.

Resnik, Philip 1999. Mining the web for bilingual text. – Proc. Of 37th Meeting of the ACL. Maryland, 527-534; http://umiacs.umd.edu/~resnik/pubs/acl99.ps.gz (26.05.2007).

Resnik, P.; Smith, N. A. 2003. The Web as a parallel corpus. – Computational Linguistics, 29(3), Cambridge: MIT Press, 349–380.

Rosen, Alexandr 2005. In search of the best method for sentence alignment in parallel texts. – Proceedings of SLOVKO 2005, the Third International Seminar on Computer Treatment of Slavic and East European Languages, VEDA. Bratislava; http://utkl.ff.cuni.cz/~rosen/public/slovko05.pdf (26.05.2007).

Salkie, Raphael 2002. How can linguists profit from parallel corpora? – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 93–109.

Samy, D.; Sandoval, A.M.; Guirao, J.M.; Alfonseca, E. 2006. Building a parallel multilingual corpus (Arabic-Spanish-English). – Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006. Genoa, Italy, 2176-2181.

Scannell, K. P. 2003. Automatic thesaurus generation for minority languages: an Irish example. – Actes de la 10e conférence TALN à Batz-sur-Mer, volume 2, 203–212.

Singh, A. K.; Husain, S. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. – Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan, Association for Computational Linguistics, 99–106.

Smadja, F.; McKeown, K. R.; Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: A statistical approach. – Computational Linguistics, 22(1), Cambridge: MIT Press, 1–38.

Software for word alignment; http://www.cse.unt.edu/~rada/wa/#softwareWA (21.08.2006).

Sperberg-McQueen, C. M.; Burnard, Lou (eds.) 2004. Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition; http://www.tei-c.org/P4X/index.html (28.05.2007).

Stahl, Peter 2002. Building and processing a multilingual corpus of parallel texts. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 47–59.

Steinberger, Ralf; Pouliquen, Bruno; Widiger, Anna; Ignat, Camelia; Erjavec, Tomaž; Tufiş, Dan; Varga, Dániel 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. – Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy; http://langtech.jrc.it/Documents/0605_LREC_JRC-Acquis_Steinberger-et-al.pdf (24.05.2007).

Sumita, Eiichihiro 2003. An Example-Based Machine Translation System Using DP-Matching Between Word Sequences. – Recent Advantages in Example-Based Machine Translation. Michael Carl and Andy Way (eds.). Dordrecht: Kluwer Academic Publishers, 189-209.

Sågvall Hein, Anna 2002. The PLUG-project. Parallel corpora in Linköping, Uppsala, Göteborg: aims and achievements. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 61–78.

The JRC-Acquis multilingual parallel corpus; http://langtech.jrc.it/JRC-Acquis.html (21.08.2006).

The PLUG Word Aligner – PWA; http://stp.ling.uu.se/plug/pwa/index.html (21.08.2006).

Tiedemann, Jörg 1997. Automatical lexicon extraction from aligned bilingual corpora. Diploma thesis, University 'Otto-von-Guericke', Magdeburg: Department of Computer Science; http://stp.ling.uu.se/~joerg/diplom/diplom.html (26.05.2007).

Tiedemann, Jörg 2002. Uplug – a modular corpus tool for parallel corpora. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 181-197.

Tiedemann, Jörg 2003. Recycling translations. Extraction of lexical data from parallel corpora and their application in natural language processing; http://stp.ling.uu.se/~joerg/phd/html/thesis_html.html (21.08.2006).

Tiedemann, J.; Nygaard, L. 2004. The OPUS corpus – parallel and free. – Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04), Lisbon, Portugal, 1183-1186; http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf (26.05.2007).

Trosterud, T. 2002. Parallel corpora as tools for investigating and developing minority languages. – Parallel corpora, parallel worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Borin, Lars (red.). Amsterdam: Rodopi, 111–122.

Tufiş, Dan; Barbu, Ana-Maria 2001. Automatic construction of translation lexicons. – Advances in Automation, Multimedia and Modern Computer Science, A Series of Reference Books and Textbooks in Electrical and Computer Engineering. V. V. Kluew; C. E. D'Attellis; N.E. Mastorakis (eds). WSEAS Press, 156–172.

Uchimoto, Kiyotaka; Zhang, Yujie; Sudo, Kiyoshi; Murata, Masaki, Sekine, Satoshi; Isahara, Hitoshi 2004. Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. – Proceedings of the Multilingual Language Resources (MLR2004); 2004, Geneva,

Switzerland;
http://cs.nyu.edu/~sekine/papers/mlr04-uchimoto.pdf (26.05.2007).

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón (2005): Parallel corpora for medium density languages. – Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria; http://www.metacarta.com/docs/ranlp05parallel.pdf (26.05.2007).

Volk, Martin; Samuelsson, Yvonne 2004. Bootstrapping parallel treebanks. – 20th International Conference on Computational Linguistics, Geneva, 29 August 2004; http://www.coli.uni-saarland.de/conf/linc-04/volk.pdf (26.05.2007).

Volk, Martin 2005. The role of the Web in machine translation. – M. Langemets, P. Penjam (eds.), Proc. Of the 2nd Baltic Conference on Human Language Technologies, Tallinn, 79-85.

Véronis, Jean; Langlais, Philippe 2000. Evaluation of parallel text alignment systems: The ARCADE project. – Parallel Text Processing: Alignment and Use of Parallel Corpora. Véronis, J. (ed.). Dordrecht: Kluwer, 369–388.

Veskis, Kaarel 2007. Kakskeelsete leksikonide genereerimine paralleelkorpuse baasil. – Eesti Rakenduslingvistika Ühingu aastaraamat 3. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Rakenduslingvistika Ühing, lk. 355–372.

Widdowson, H. G. 1990. Aspects of language teaching, Oxford, Oxford University Press, 213 p.

Lisa 1. Fragment UWA-ga TÜ paralleelkorpusest ekstraheeritud leksikonist

...

{viljapuuistandike}
{
2X:fruit tree plantations
}
{viljastamine}
{
1X:fertilisation
}
{viljastamise}
{
1X:violation
}
{viljastanud}
{
2X:fertilizing
}
{viljastumiseks}
{
1X:fertilization
}
{viljatu}
{
3X:infertile
}
{viljatut}
{
2X:infertile
}
{viljatükk}
{
1X:blemished
}
{viljelemine}
{
1X:eliminated
}
{viljelty}
{
1X:viljelty
}
{viljelusest}
{
1X:permanent abandonment
}
{vilju}
{
3X:units
}
{vilken}
{
6X:vilken
}

{vilkingid}
{
5X:wilkings
}
{vilkingite}
{
1X:wilkings
}
{vilkudes}
{
1X:slow-flashing
}
{vilkur}
{
6X:flashing lamps
}
{vill}
{
9X:wool
}
{villa}
{
1X:villa
}
{villane}
{
1X:woollen
}
{villarasvaalkoholid}
{
1X:cetostearyl
}
{villasest}
{
1X:woollen
}
{villes déterminées par}
{
1X:villes déterminées par
}
{villija}
{
6X:bottler
}
{villimine}
{
8X:bottling
1X:racking
}
{vin}
{
3X:vin
}
{vingt-neuf}
{
2X:vingt-neuf
}
{vinho}
{
2X:vinho
}
{vinifera}
{
1X:vinifera
}
{vinifera L}
{
2X:vinifera L
}
{vinification}
{
1X:vinification
}

{vinificazione}
{
1X:vinificazione
}
{vinkamiin}
{
1X:vincamine
}
{vino}
{
6X:vino
}
{vinorelbiini}
{
1X:vinorelbine
}
{vinte}
{
5X:vinte
}
{vinte e}
{
1X:ventidue
}
{vintraudset}
{
1X:a rifled barrel
}
{violetne}
{
4X:violet
}
{violetne spargel}
{
1X:violet asparagus
}
{violetne/purpurne}
{
1X:violet/purple
}
{violetne/roheline}
{
1X:violet/green
}
{violetne/roheline spargel}
{
2X:violet/green asparagus
}
{violetse}
{
3X:violet
}
{violetse spargli}
{
2X:violet asparagus
}
{violetse/rohelise}
{
1X:violet/green
}
{violetse/rohelise spargli}
{
2X:violet/green asparagus
}
{violetsed/purpursed}
{
1X:violet/purple tops
}
{violetset}
{
1X:violet
}
{violetset sparglit}
{

1X:violet asparagus
}
{violetsete/purpursete}
{
1X:violet/purple
}
{violettpunase}
{
1X:violet-red
}
{virde}
{
2X:provided
}
{virens}
{
2X:virens
}
{virgaurea}
{
1X:virgaurea
}
{virgem}
{
1X:virgem
}
{virgen}
{
1X:virgen
}
{virginiamütsiin}
{
1X:virginiamycin
}
{virginiamütsiini}
{
4X:virginiamycin
}
{virginiamütsiiniresistentne}
{
1X:virginiamycin-resistant
}
{virginiamütsiiniresistentsed}
{
1X:virginiamycin-resistant
}
{virginiana}
{
2X:virginiana
}
{virksomhed}
{
2X:virksomhed
}
{virksomheder}
{
1X:virksomheder
}
{virnastatavate}
{
1X:stackable
}
{virnastatavates}
{
5X:stackable
}
{viroidid}
{
1X:viroids
}
{viroididele}
{
1X:viroids
}

{virola}
{
1X:virola
}
{viroloogiainstituut}
{
1X:Virology
}
{viroloogiline}
{
1X:virological
}
{viroloogiliseks}
{
2X:virological
}
{viroloogiliste}
{
2X:virological
}
{virsiku}
{
1X:peel
}
{virsikud}
{
5X:Peaches
13X:peaches
}
{virsikute}
{
7X:peaches
}
{virsikutest}
{
1X:the fruit nectars
}
{virtsa}
{
2X:liquid manure
}
{virtuaalne kollokatsioon}
{
1X:virtual collocation
}
{virtuaalse kollokatsiooni}
{
1X:virtual collocation
}
{virtud}
{
2X:virtud
}
{virtud del}
{
1X:virtud del
}
{visés}
{
1X:visés
}
{visatakse}
{
7X:Discard
2X:discard
}
{viscum}
{
1X:viscum
}
{visiidid}
{
3X:visits
}

{visiiditasu}
{
1X:visit fee
1X:visit fees
}
{visiirid}
{
1X:visors
}
{visiite}
{
1X:visits
}
{visiitkaardid}
{
1X:visiting cards
}
{viskekeha}
{
1X:kinetic
}
{viski}
{
7X:whiskey
6X:whisky
}
{viskoosist}
{
1X:viscose
}
{viskoosne}
{
3X:viscous
}
{viskoossus}
{
2X:viscosity
}
{vismutsubgallaat}
{
1X:bismuth subgallate
}
{vismutsubkarbonaat}
{
1X:bismuth subcarbonate
}
{vismutsubnitraat}
{
1X:bismuth subnitrate
}
{vismutsubnitraat ja}
{
1X:bismuth subnitrate
}
{vismutsubsalitsülaat}
{
1X:bismuth subsalicylate
}
{visuaal-}
{
2X:visual
}
{visuaallennuks}
{
1X:visual flight
}
{visuaallähenemise}
{
1X:visual approach exercise
}
{visuaalne}
{
1X:Visual
1X:the visual

```
   9X:visual
}
{visuaalse}
{
   11X:visual
}
{visuaalsed}
{
   1X:visual
}
{visuaalseks}
{
   1X:visual
}
{visuaalsel}
{
   2X:visual
}
{visuaalselt}
{
   6X:visually
}
{visuaalset}
{
   3X:visual
}
{visuaalseteks}
{
   1X:visual checks
}
{visuaalsüsteem}
{
   1X:visual system
}
{visvaart}
{
   2X:visvaart
}
{vitamiin}
{
   4X:Vitamin
   4X:vitamin
}
{vitamiin B1}
{
   1X:vitamin B1
}
{vitamiin B2}
{
   1X:vitamin B2
}
{vitamiin B3}
{
   1X:vitamin B3
}
{vitamiin B6}
{
   1X:vitamin B6
}
{vitamiin C}
{
   1X:vitamin C
}
{vitamiin H}
{
   1X:vitamin H
}
{vitamiine}
{
   1X:vitamin
   12X:vitamins
}
{vitamiini}
{
```

1X:vegetable juice
2X:vegetable juices
1X:vitamin D
}
{vitamiini-}
{
2X:vitamin
}
{vitamiinid}
{
1X:Vitamins
15X:vitamins
}
{vitamiinide}
{
1X:the vitamin
6X:vitamin
27X:vitamins
}
{vitamiinidest}
{
1X:vitamins
}
{vitaminiseerimine}
{
1X:vitaminization
}
{vitelli}
{
3X:vitelli
}
{vitreous}
{
2X:vitreous
}
{vitriinides}
{
2X:showcases
}
{vitro}
{
6X:vitro
}
{vivo}
{
8X:vivo
}
{vl39}
{
1X:ISO/IEC
}
{vleesras}
{
1X:vleesras
}
{voce}
{
1X:voce
}
{voeder}
{
2X:voeder
}
{voeding}
{
2X:voeding
}
{voedselhulp}
{
1X:voedselhulp
}
{voi}
{
11X:voi

}
{voimassa}
{
8X:voimassa
}
{voimassa viisi}
{
1X:voimassa viisi
}
{voimassaolon}
{
2X:voimassaolon
}
{voiöljy}
{
11X:voioeljy
}
{voiöljyksi ja}
{
1X:voioeljyksi ja
}
{vol}
{
5X:vol
}
{volatiilsus-}
{
1X:volatility
}
{volatiilsuse}
{
1X:volatility
}
{volatiilsusega}
{
1X:volatility
}
{voldaan}
{
1X:voldaan
}
{voldi}
{
3X:volts
}
{voldik}
{
2X:folder
}
{voldise}
{
1X:a CV product
1X:voltage
}
{volditud}
{
1X:concertinafolded
}
{voldoen}
{
1X:voldoen
}
{volgende}
{
1X:volgende
}
{volgnummer}
{
1X:volgnummer
}
{volikiri}
{
1X:Authorisation document
1X:Authorisation documents

29X:authorisation
60X:authorisation document
1X:authorisation documents
}
{volikogu}
{
6X:Council
1X:Councils
259X:council
23X:councils
}
{volikogu esimehe valimiseni}
{
1X:volikogu esimehe valimiseni
}
{volikogu liikmeks}
{
1X:unelected independent
}
{volinike}
{
4X:proxies
}
{volitada teisi}
{
2X:authorise other persons
}
{volitajad}
{
1X:delegators
}
{volitamise}
{
1X:the limits
}
{volitamiseks}
{
2X:authorise a joint
}
{volitatu}
{
2X:delegator
}
{volitatud}
{
6X:Approved
18X:Authorised
1X:Obligation
20X:agencies
29X:agency
193X:approved
772X:authorised
1X:invalidate
2X:invalidated
1X:negotiated
1X:obligated
6X:obligation
4X:obligations
3X:political
1X:violated
1X:violates
9X:violation
1X:violations
}
{volitatud ametiisiku juuresolekul}
{
1X:data protection supervision
}
{volitatud importijad}
{
1X:designated importers
}
{volitatud isikud koha}
{

2X:projects receiving financial
}
{volitatud politseiasutused}
{
1X:police authorities authorised
}
{volitus}
{
2X:Authorisation
36X:authorisation
18X:mandate
1X:mandates
}
{volituse}
{
1X:clause
1X:course
3X:limits
6X:validity
}
{volitused}
{
3X:Powers
3X:appointed
6X:authorised
5X:entrusted
95X:powers
1X:unlimited
}
{volitused kestavad uue}
{
3X:a court-appointed member
}
{volitusi}
{
1X:Powers
1X:complaints
62X:powers
}
{volitusnormita}
{
1X:a provision delegating
}
{volitust}
{
1X:limits
}
{volituste}
{
3X:Committee
10X:committee
1X:committees
1X:situated
1X:virtue
}
{volituste lõppemise}
{
7X:termination
}
{volledig}
{
2X:volledig
}
{volti}
{
1X:volts
}
{voltmeetrite}
{
1X:voltmeters
}
{voluntary}
{
5X:voluntary

}
{vom}
{
1X:vom
}
{vom Antragsteller beantragter}
{
1X:vom Antragsteller beantragter
}
{von}
{
47X:von
}
{von Artikel}
{
1X:von Artikel
}
{von Lebensmittelzubereitungen}
{
1X:von Lebensmittelzubereitungen
}
{von Maschinenanlagen}
{
1X:von Maschinenanlagen
}
{vooderdatud}
{
2X:lined
}
{voodipesu}
{
3X:bed linen
}
{voolavad}
{
1X:free-flowing
}
{vooluallikate}
{
2X:power supplies
}
{vooluhulk}
{
1X:unpermitted
}
{vooluimpulss}
{
1X:impulse
}
{voolujaotusega}
{
1X:divider
}
{voolukulu}
{
10X:flowrate
2X:flowrates
}
{voolukulu Q}
{
1X:the gas meter
}
{vooluringide}
{
1X:design drawings
2X:sub-assemblies
}
{voolutatakse}
{
1X:flushed
}
{vooluvee}
{
1X:barring

}
{voor}
{
31X:voor
}
{voor Zuivel}
{
1X:voor Zuivel
}
{voor gehandicapten}
{
1X:voor gehandicapten
}
{voor het}
{
1X:voor het
}
{voor ten}
{
2X:voor ten
}
{voor uitvoer}
{
1X:voor uitvoer
}
{vooraf}
{
1X:vooraf
}
{vooraf van}
{
1X:vooraf van
}
{voorbehoud}
{
1X:voorbehoud
}
{voorbehoud dat}
{
1X:voorbehoud dat
}
{voorbehoud van bijzondere}
{
1X:voorbehoud van bijzondere
}
{vooru}
{
19X:Round
2X:round
1X:vooru
}
{vooru tulemusi ühendava}
{
2X:Embodying
1X:Embodying the Results
}
{voorus}
{
2X:course
1X:offers
}
{voorwaarden}
{
1X:voorwaarden
}
{voorzien}
{
1X:voorzien
}
{vorbehaltlich eines}
{
1X:vorbehaltlich eines
}
{vorgesehene ermaessigte Zollsaetze}

{
1X:vorgesehene ermaessigte Zollsaetze
}
{vorm}
{
43X:Form
9X:Format
4X:Forms
12X:for
113X:form
128X:format
1X:formats
6X:forms
1X:status
3X:work
}
{vorm(id}
{
2X:form(s
}
{vorme}
{
2X:Forms
3X:form
22X:forms
1X:more
}
{vormi}
{
2X:FORM
24X:Form
1X:Formal
10X:Format
1X:Forms
7X:Origin
182X:form
15X:formal
166X:format
1X:formats
11X:forms
3X:origin
3X:permit
}
{vormi CO}
{
1X:Form CO
}
{vormi V}
{
1X:Form V
}
{vormi-}
{
4X:Format
1X:form
10X:format
}
{vormid}
{
1X:Board
2X:Form
8X:Forms
4X:form
14X:format
16X:formats
47X:forms
1X:origin
2X:permit
1X:permits
2X:voting
}
{vormide}
{
2X:invoice

}
{vormides}
{
7X:forms
4X:place particularly
}
{vormiga}
{
2X:Origin
}
{vormil}
{
2X:a form
16X:form
2X:forms
}
{vormile}
{
1X:Article
1X:form set
1X:variety
}
{vormilise}
{
2X:a formal arrangement
}
{vormimiseks}
{
2X:moulding
}
{vorming}
{
11X:format
}
{vorminõuete}
{
5X:formalities
}
{vormiriietus}
{
2X:Uniform
3X:uniform
1X:uniforms
}
{vormis}
{
4X:Form
3X:a form
182X:form
2X:formal
21X:format
21X:forms
1X:visits
}
{vormis X}
{
1X:Form X
}
{vormis teabesertifikaadi INF}
{
1X:information certificate INF
}
...

Lisa 2. Fragment LWA-ga TÜ paralleelkorpusest ekstraheeritud leksikonist

...

kahe aasta *biennial*

kahe aasta jooksul *two*

kahekolmandikulise häälteenamusega *two-thirds*

kaheksanda *eighth*

kahekümnendal *twentieth*

kahelda *doubt*

kahepoolse *bilateral*

kahepoolseid *concluded*

kahepoolsetes *double-taxation*

kahes *two*

kahest *two*

kahjum *taxable*

kahjustab *kahjustab*

kahjustada *effectiveness*

kahjustamata *clarifications*

kahjustavate *affecting*

kaht *two*

kahte *two*

kahtlemise *doubting*

kahtlust *doubt*

kahvaturoheline *pale*

kaimanisaared *cayman islands*

kaitse *protection*

kaitse *kaitse*

kaitse *defence*

kaitse kohta *protection*

kaitse kohta *community*

kaitseaspektide *security*

kaitsega *environment*

kaitseidentiteedi *esdi*

kaitseks *by*

kaitseks *protect*

kaitsemeetmeid *stringent*

kaitsepoliitika *defence policy*

kaitsepoliitikaga *framework*

kaitsepoliitilise tähendusega *implications*

kaitsepoliitilise tähendusega *implications.*

kaitsepoliitilise tähendusega otsuste *decisions*

kaitset *otherwise*

kaitsevad *defend*

kaitsmiseks *internal*

kaitsta *protect*

kaitsta *protected*

kaitsta *safeguard*

kaitstuse *protection*

kajastamiseks reflect

kaks two

kaks kaks

kalanduspoliitika policy

kalatoodete fisheries

kalatoodetelt islands

kalendriaasta calendar year

kalendriaastaks calendar

kaltsiumfluoriidi fluorine

kammitud carded

kanaari saarte canary

kanaari saartele customs

kanada canada

kanadast canada

kanded reflect

kandmise registration

kandmist activities

kantakse registered

kantud registered

kantud listed

kantud list

kantud included

kaotamise penalty

kaotamise kohta june

kaotatakse abolished

kaotatud lost

kapital capital

kapitali capital

kapitaliosa share

kapitalis capital

kapitalist capital

kapitaliturul capital

karistusmaksed periodic penalty payments

karjamaapasture

karusnahad pelts

karusnahku annex

kas governors

kas kogu auditors

kasu benefits

kasu benefit

kasuks favour

kasum hands

kasumi profits

kasumit profits distributed

kasumit profits

kasumitaotluseta intermediate

kasutab exercise

kasutab exercises

kasutada used

kasutada use

kasutades using

kasutajate users

kasutamine use

kasutamine *accordance*

kasutamisdeklaratsiooni *similar*

kasutamise *use*

kasutamise *regards*

kasutamisel *formulation*

kasutamist *use*

kasutatakse *used*

kasutatava *fine-cut*

kasutatavad *used*

kasutatavate *used*

kasutuseks *national*

kasutusele *resort*

kasutusotstarbest *exemptions*

kasvatajale *breeder*

kasvatatud *animals*

kasvatatud *animal*

kasvatatud *birds*

kasvatatud *captivity*

kasvava *growing*

kasvu *growth*

kasvupiirkonnas *situated*

katavad *communities*

kategooria *category*

kategooria ainete *transactions*

katkematu *years*

katkestanud *started*

katmiseks *cover*

katta *rise*

katteleht *binder*

kauba *goods*

kaubad *goods*

kaubandus *sustainable*

kaubandus *sub-committee*

kaubandusaspektide *aspects*

kaubanduse *trade*

kaubanduseeskirjade *regulations*

kaubandusele *trade*

kaubandusettevõtjate *traders*

kaubanduskokkuleppe *tariffs*

kaubanduslepingu *multilateral trade agreement*

kaubanduslepingu *plurilateral*

kaubanduslepingud *agreement*

kaubanduslepingud *agreements*

kaubanduslepingus *multilateral trade agreement*

kaubanduslepingus *agreement*

kaubanduslepingute *trade*

kaubanduslepingute *multilateral*

kaubanduslepingutega *covered*

kaubanduslepingutes *multilateral*

kaubandusliku *commercial*

kaubandusläbirääkimiste *trade negotiations*

kaubandusläbirääkimiste uruguay vooru round

kaubandusmeetmete trade measures

kaubandusorganisatsioon world

kaubanduspoliitika trade policy review

kaubanduspoliitika trading

kaubandussektor diminished

kaubandust other

kaubandust environmental measures

kaubandustegevus commercial presence

kaubandustegevuse commercial presence

kaubandustõkete barriers

kaubandustõkkeid barriers

kaubandusvoogude flows

kaubasaajale consignee

kaubasaatjale consignor

kaubavahetuse goods

kaudselt indirectly

kaudu hendamise

kaup goods

kaupa goods

kaupade goods

kaupade arrangements

kauplemise trade

kauplemist trade

kausside outside

kava scheme

kava plan

kavade scheme

kavade schemes

kavandamisel framing

kavandamisele framing

kavandatav proposed

kavandatavad envisaged

kavandatud designed

kavandatud enter

kavatseb important

kavatsevad intend

kavatsusest intention

keda assisted

keda whom

keelab prohibit

keelab prohibits

keelamine prohibition

keelamistprohibition

keelata prohibit

keelata prohibited

keelatakse prohibited

keelatakse prohibits

keelatud prohibited

keelavad prohibit

keelduda refuse

keeldude prohibitions

keeles ametlikus

keeltes keeles

keeltest languages

keelust prohibitions

keemilised mechanical

keemiliselt chemically

keemiliste chemical

keeratav fine-cut

keeratava fine-cut

keeratud particles

keersmaeker keersmaeker

kehtestabfalls

kehtestada establish

kehtestada made

kehtestama make

kehtestama established

kehtestamast introducing

kehtestamise last

kehtestamist restriction

kehtestamist restrictions

kehtestatakse quotas

kehtestatakse reduction

kehtestatakse reliefs

kehtestatakse forecast

kehtestatiproducers

kehtestatud established

kehtestatud monitoring

kehtestavad member

kehtetuks repealed

kehtetuks tunnistatud eec

kehtib apply

kehtib valid

kehtib paragraph

kehtinud applicable

kehtiv valid

kehtiva operating

kehtiva valid

kehtivad apply

kehtivad sitting

kehtivate force

kehtivate implemented

kehtivate implement

kehtivus validity

kehtivusaeg validity

kehtivusaega validity

kehtivuse validity

kehtivuse expired

kehtivuse expire

kehtivuspäevani validity

kehtivust validity

kel believe

kellel legal

kellel holding

kellel *lawyers*

kellel *lawyer*

kellel *issues*

kellele *whom*

keraamilisel *ge.*

kergesti *kergesti*

kerma *tarkoitettuihin*

kes *contracting*

kes *see*

kes *legal person*

kes *resident*

kes *running*

kes *exercise*

kes *return*

keskkonda *environment*

keskkonna *environment*

keskkonnakaitse *environmental*

keskkonnakaitseks *protect*

keskkonnakomitee *environment*

keskkonnameetmete *environmental*

keskkonnaotstarbeliste *system*

keskkonnapoliitika *policies*

keskkonnas *producing*

keskmine *average*

keskmise *average*

keskmise suurusega *small*

keskpanga *bank*

keskpanga *bank.*

keskpangad *central banks*

keskpankade *central banks*

kestab *years*

kestab *year*

kestus *duration*

kestust *duration*

kiechle *kiechle*

kiidab *acting*

kiiremini *rapidly*

kiiresti *soon*

kiirgusavarii *emergency*

kiitnud *by*

kilogrammi *carcases*

kilogrammi *carcase*

kilogrammi *kilograms*

kindla *fixed*

kindlaks *analysis*

kindlaks *fixed*

kindlaks *determine*

kindlaks *stipulated*

kindlaks *unanimous*

kindlaks *visa*

kindlaks määrata *determine*

kindlaks määrata *determined*

kindlaks määratud *by*

kindlaksmääramiseni *date*

kindlaksmääratud fixed

kindlatesse limited

kindlustada ensure

kindlustades access

kindlustamisele rural areas

kinni recovered

kinnipeetava withholding tax

kinnisvara property

kinnitab approved

kinnitab approve

kinnitab assurance

kinnitab lay down

kinnitades confirming

kinnitades reaffirming

kinnitatud affixed

kinnitava assurance

kinnitavad reaffirm

kirikute churches

kirjalik written

kirjalike written

kirjalikke written

kirjaliku written

kirjalikult writing

kirjalikust written

kirjas tr&uuml

kirjavahetuse letters

kirjavahetuses letters

kirjeldus description

kirjelduse description

kirjeldused content

kirjelduste particular

kirjelduste particulars

kirjutatud signed

kirjutatud forms

kirjutatud formed

kirjutavad signed

kitsed goats0

kitsede goats

kiu fibre

kiu stated

kiu state

kiu kirjeldus column'fibre

kiud fibre

kiud fibres

kiudude fibres

kiududestbacking

kiudusid fibres

kiukoostis fibre composition

kiukoostis base

kiukoostise fibre composition

kiuprotsendid fibre

kiusisaldus stated

kiusisalduse stated

kiusisaldusega contents

kiusisaldusega content

klauslile referred

klauslite clauses

kleinwalsertali austria

kleinwalsertalis austria

kliendideklaratsioon showing

kliendile declaration

klient furnish

kliima climate

kodade consisting

kodakondsus citizenship

kodakondsus nationality

kodakondsuse nationality

kodakondsuse citizenship

kodakondsuseta stateless

kodakondsust. citizenship.

kodanik nationals

kodanik national

kodanike nationals

kodanike citizens

kodanikel nationals

kodanikele nationals

kodanikud nationals

kodukordprocedure

kodukorra procedure

kodukorra. procedure.

kodukorraga rules

kodukorras rules

kogeda financing

kogemused experience

kogemuste experiences

kogemuste experience

kogemustele experiences

kogu entire

kogu toote product

kogumassist total weight

kogus quantity

kogused quantities

kogusegaslips

koguseid quantities

kogusumma burden

kogutoodangu accordance

koha office

kohal chambers

kohal chamber

kohaldada apply

kohaldada applicable

kohaldada applying

kohaldades applying

kohaldamise application

kohaldamise applications

kohaldamise id

kohaldamise applying

kohaldamisega application

kohaldamiseks purpose

kohaldamiseks purposes

kohaldamiseks application

kohaldamisel purposes

kohaldamisel purpose

kohaldamisel application

kohaldamisel throughout

kohaldamist preceding

kohaldata apply

kohaldataks applied

kohaldatakse apply

kohaldatakse apply.

kohaldatakse applied

kohaldatakse applies

kohaldatav applicable

kohaldatava applicable

kohaldatavad applicable

kohaldatavaks become

kohaldatavate applicable

kohalikud regional

kohandada adapted

kohandada adapt

kohandada adjusted

kohandamiseks adapting

kohandamist adaptation

kohandatakse adapting

kohandatakse adjusted

kohandatud adapted

kohandusi adjustments

kohapeal institutions

kohasel post

kohaselt cohesion

kohaselt according

kohaselt. management.

kohasust transactions

kohdan kohdan

kohe immediately

kohestatud other

kohestatud animal hair

kohta certain arable crops

kohta per

kohta organization

kohta investment

kohta clothing

kohta c

kohta tomatoes

kohta western

kohta kehtivad subject

kohta vastavalt control

kohta vastavalt controls

kohta. per

kohta. investment

kohtades places

kohtadesse points

kohtlemine *treatment*

kohtu *justice*

kohtu *secrecy*

kohtu *court*

kohtu *pending before*

kohtu *judicial authority*

kohtu otsusega *settled*

kohtuasja *ruling*

kohtuasjasse *instance*

kohtuistung *hearing*

kohtujurist *reasoned*

kohtujurist *reason*

kohtujuristi *advocate*

kohtujuristid *advocates-general*

kohtujuristide *advocates-general*

kohtujuristide *advocates*

kohtukulude *costs*

kohtule *case*

kohtule *appealed*

kohtule *appeal*

kohtule *document*

kohtult *decision*

kohtumiste *intervals*

kohtunik *judge*

kohtunike *judges*

kohtunikel *immune*

kohtunikku *judges*

kohtunikku *judge*

kohtuniku *judge*

kohtunikud *judges*

kohtunud ukogu

kohtuotsuse *judgment*

kohtuotsusega *decision*

kohtus *justice*

kohtus *practise*

kohtus *pending before*

kohtus *judicial authority*

kohtusekretär *registrar*

kohtusekretäri *assistant*

kohtusekretäri *registrar*

kohtusekretärile *instance*

kohtusekretärile *registrar*

kohtusse *justice*

kohtusse *court*

kohtusse *courts*

kohtusse *appear*

kohtute *judicial*

kohus *court*

kohus *instance*

kohusetundlikult *preserve*

kohustatud *obliged*

kohustub undertaken

kohustus *obligation*

kohustuse *obligation*

kohustused *mfn*

kohustused *obligations*

kohustused *responsibilities*

kohustusest *obligation*

kohustusi *obligations*

kohustusi *impartially*

kohustusi *duties*

kohustusi *responsibilities*

kohustusi *offering*

kohustusi *addition*

kohustuslikku *partly*

kohustuslikuks *destination*

kohustust *obligations*

kohustust *provide*

kohustuste *specific commitments*

kohustuste *obligations*

kohustuste *responsibilities*

kohustuste täitmiseks *obligations*

kohustuste täitmisel *performance*

kohustuste täitmisel *discharge*

kohustuste täitmist *responsibilities*

kohustuste täitmist *performance*

kohustustega *debts*

kohustustega *obligation*

kohustustega *obligations*

kohustustele *legal*

kohustustest *admission*

kohustuvad *undertake*

kohvikannusoojendajad *coffee cosy covers*

koja *chamber*

kokku *agreed*

kokku *agree*

kokku *convene*

kokku *member*

kokku *members*

kokku *set*

kokku järgmises *agrees*

kokku järgmises *agree*

kokku lepitud *agreed*

kokku leppinud *agreed*

kokkulepe *agreement*

kokkulepete *agreements*

kokkulepitud *agreed*

kokkuleppe *agreement*

kokkuleppe *agreements*

kokkuleppega *bank*

kokkuleppeid *amendment*

kokkuleppel *agreement*

kokkuleppelised *agreed allowances*

kokkuleppelist *allowances*

kokkuvõttetabelid *tables*

kokkuvõttetabelit *summary*

kokkuvõttetabelite *summary*

kollane yellow

kollektiivse collective

kolm three

kolmanda third

kolmanda isiku vastu tugineda third parties

kolmanda riigi third country

kolmandal third

kolmandal etapil third stage

kolmandas lõigus third subparagraph

kolmandast third

kolmandate third parties

kolmandate isikute third

kolmandate riikide countries

kolmandate riikide kodanike nationals

kolmandatesse export

kolmandatest riikidest countries

kolmandikku upon

kolmandiku thirds

kolmandiku third

kolmas third

kolmas by

kolme three

kolme and'four

kolme kuu jooksul pärast months

kolmekomponentsete ternary

kolmekümnendal 30th

kolmeneljandikulise three-fourths

kolmeneljandikulise fourths

kolmest three

kolmeteistkümne thirteen

koma citation

komisjon commission

komisjon komisjon

komisjoni commission

komisjoni komisjoni

komisjoni amended by regulation

komisjoni board

komisjoni n

komisjoni schoolchildren

komisjoni ettepaneku põhjal proposal

komisjoni ettepanekut whereas article

komisjoni nimel komisjoni liige oj

komisjoni nimel komisjoni liige franz fischler commission

komisjoni nimel komisjoni liige mario monti commission

komisjoni nimel komisjoni liige ren&eacute ren&eacute

komisjoniga commission

komisjonil commission

komisjonile commission

komisjonile komisjonile

komisjonile. them.

komitee committee

komitee komitee

komiteed committee

komiteede committees

komiteega committee

komiteele committee

komiteele komiteele

komitees committee

kompensatsiooni compensation

kompensatsiooni saamise tingimustele producers

kompensatsioonide producers

kompenseerimise compensation

kompenseerimiseks enforced

kompenseeriv adjustment

kompenseeriva compensatory

koncentreras koncentreras

koncentrerat koncentrerat

koncentrerat smör för koncentrerat smoer foer

konföderatsiooni annexed

konkreetse particular

konkurentsi competition

konkurentsivõime competitiveness

konkurentsiõiguse competition law

konsensus consensus

konsensuse consensus

konsensust consensus

konsultatsioone consultations

konsultatsioone consultation

konsultatsiooni consultations

konsultatsiooni consultation

konsultatsioonide consultations

konsultatsioonide consultation

konsultatsioonikava plan

konsultatsioonimenetlust procedures

konsulteerib consult

konsulteerib consults

konsulteerib consulted

konsulteerida consulted

konsulteerida relating

konsulteerides authority

konsulteerimis consultation

konsulteerimise consultation

konsulteerimist consulting

konsulteerimist consultation

konsulteerimist consultations

konsulteerimist euroopa parlamendiga consulting

konsulteeris cross-industry

konsulteeriv consulting

konsulteeriva consulting

konsulteerivad consulted

konsulteerivad consult

kontaktkomitee contact

kontekstis context

kontrastsele taustale

kontroll control

kontrollaktsiapakist participates

kontrolli checks

kontrolli control

kontrolli controls

kontrollib examine

kontrollib verify

kontrollida inspect

kontrollikoda auditors

kontrollikoda court

kontrollikoda doing

kontrollikoja reports

kontrollikoja auditors

kontrollikoja agreement between

kontrollikoja managing

kontrollikojale court

kontrollimenetlusele non-economic

kontrollimiseks bank.

kontrollimiseks. bank.

kontrollitakse performed

kontrollitakse carried

kontsentreeritud concentrated

kontsessioone free

kontsessiooni concession

kontsessioonid concessions

kontsessioonide commitments

kontsessioonide concessions

kontsessioonide concessionaires

kontsessiooniga affected

konventsioone conventions

konventsiooni convention

konventsiooni osapoolte quota

konventsiooni osapoolte quotas

konventsioonide conventions

konventsiooniga whereas

konverents conference

konverents lepib kokku agrees

konverentsi conference

konverentsidel conferences

konverentsil conference

konverentsil conferences

konverentsile conference

koodi code

koodi codes

koodidega by

koodidega i.a

koodiga code

koodiga codes

koodiga replaced

koondnomenklatuuri nomenclature

koopia copy

koopia proof

koopiad copies

koopiat issued

koopiate application

koordineerimist policies

kooritud skimmed-milk

koos together

koos tariff quota

kooskõla coherence

kooskõlas accordance

kooskõlas conformity

kooskõlas artiklisaccordance

kooskõlas käesoleva compatible

kooskõlas käesoleva lepingu accordance

kooskõlastamiseks coordination

kooskõlastamiseks. coordination

kooskõlastamist coordination

kooskõlastavad coordinate

koosneb consist

koosneb fibre

koosneb fibres

koosnev preparing

koosnev composed

koosneva composed

koosnevate judges

koosseisusitting

koosseisucomposition

koosseisus composition

koostab up

koostab prepare

koostada up

koostamise preparation

koostamisel up

koostas drew

koostatakse drawn up

koostatud up

koostatud established

koostatud esimesele

koostis composition

koostise composition

koostisega composition

koostises composition

koostisosad components

koostisosi representing

koostöö cooperation

koostööd cooperation

koostööd cooperate

koostööd member

koostööga counter

koostööga cooperation

koostöös liaison

koostöös cooperation

kopenhaageni copenhagen

kord arrangements

kord aastas once

korda *arrangements*

kordamööda *examine*

kordumatu *means*

korduvaid *chain*

korea *kong*

korra *arrangements*

korra kohaselt *uniform*

korra kohta *applying*

korra üksikasjalikud rakenduseeskirjad

piima *between*

korral *consultation*

korral *radiological*

korral *necessary*

korral *event*

korraldada *conducted*

korraldada *conduct*

korraldatakse *rights*

korraldav *authority within*

korraldava *authority*

korraldus *organization*

korraldus *gatt*

korraldusasutus *management*

korraldusasutuse *management authority*

korraldusasutused *management authorities*

korraldusasutusele *management authority*

korraldusasutusele *satisfied*

korralduse *common*

korralduskomitee *committee*

korralduskomitee *whereas*

korralduskomiteede *233/94*

korrapäraselt *periodically*

korrapäraselt *regularly*

korrektse *finds*

korrelatsioonitabel *correlation table*

korrigeerib *adjust*

korsetitoodete *corsetry*

korvausta *'ilman*

kosmeetikakotid *make-up cases*

kostjaks *proceedings*

kotid *cases*

kraasitud *combed*

kraasmed *ex*

kraasmenetletud *case*

kreeka *republic*

kreeka *greece*

kreekas *greece*

kriiside *forces*

kriisiolukordade *crisis*

kriisiolukordades *crisis*

kriisivaatluskeskus *cell*

kriisivaatluskeskuse *cell*

kriminaalasjades *criminal matters*

kriminaalõiguse *criminal*

kriteeriume *criteria*

kriteeriumid *criteria*

krokodilliliste *tanned*

kude *cotton*

kui *wound up*

kui *price*

kui *unless*

kui *authority*

kui *save*

kui *voting*

kui *member*

kui *majority*

kui *audit*

kui *persistent*

kui *used*

kui *longer*

kui *seven*

kui *derogation*

kui *response*

kui *health*

kui *whenever*

kui *kui*

kui *appeal*

kui *annexed*

kui *distributed*

kui *w*

kuivsöödaturu *oils*

kuivõrd *far*

kuivõrd *insofar*

kujul *form*

kujul *forms*

kujundamine *framing*

kujundamist *defence*

kuld *environmentally*

kulla *gold*

kullaga *gold*

kullas *gold*

kultuuride *crops*

kultuurliikide *parental*

kulud *charged*

kulud *charge*

kulud *expenditure*

kulude *expenditure*

kulusid *expenditure*

kulusid *expenses*

kulutukseen *kulutukseen*

kulutused *expenditure*

kumb *whichever*

kummipaelad *elastic*

kuna *since*

kuni *until such time*

kuni *headings*

kuni *pending*

kuni *up*

kuni *kuni*

kuningriigi *kingdom*

kuningriik *kingdom*

kunstlikult *artificially propagated*

kunstlikult paljundatud *propagated*

kunstlilled *artificial flowers*

kuritegevuse *combating*

kuritegude *offences*

kuriteo *criminal*

kursiga *monetary*

kus *registry*

kus *kus*

kus *leghold trap*

kus *situated*

kus *member*

kust *state*

kustutatakse *paragraph*

kustutatakse *paragraphs*

kustutatakse *deleted*

kustutatakse *article*

kutseorgani *member*

kutsub *invites*

kutsub *invite*

kutsuda *invited*

kutsuda *invite*

kutsuvad *invite*

kuu *months*

kuu *month*

kuud *six months*

kuud *months*

kuud *month*

kuue *six*

kuue kuu jooksul pärast *months*

kuueks *six*

kuulama *hear*

kuulata *procedure*

kuuluksid *include*

kuuluv *falling*

kuuluva *falling*

kuuluvad *fall*

kuuluvad *falling within*

kuuluvaid *export*

kuuluvaid *contaminated*

kuuluvate *heading*

kuuluvate *manufactured tobacco*

kuuluvatele *cheeses*

kuuluvatele *cheese*

kuupäev *date*

kuupäevadate

kuupäeval *date*

kuupäevani *into*

kuupäevast *date*

kuus kuud *six months*

kvalifikatsiooni *qualifications*

kvalifitseeritud acting by a qualified majority

kvalifitseeritud qualified

kvalifitseeritud häälteenamusega teha qualified majority

kvalifitseeritud häälteenamuseks qualified majority

kvaliteedi quality

kvaliteeti changing

kvantitatiivse quantitative

kvantitatiivsete quantitative

kvartali second

kvoodi volumes

kvoodi volume

kvoodi quota

kvoodiga replaced

kvoodile quota

kvoodimahu volume

kvoodimahu volumes

kvoot respect

kvootide quotas

kÕrged lepinguosalised high

käes contracting

käesolev member states

käesolev member states.

käesolev hereto

käesolev day following

käesolev act

käesolev force

käesolev k&auml

käesolev provisions

käesolev keeping

käesolev closely

käesolev määrus communities.

käesolev määrus jõustub enter

käesolev määrus jõustub euroopa ühenduste teatajas avaldamise päeval day

käesolev määrus jõustub järgmisel päeval pärast into

käesolev määrus jõustub kolmandal päeval pärast force

käesolev määrus jõustub seitsmendal päeval pärast force

käesoleva conferred upon

käesoleva provided

käesoleva s&auml

käesoleva hereto

käesoleva falling under

käesoleva act

käesoleva force

käesoleva recommendations

käesoleva k&auml

käesoleva set

käesoleva provisions

käesoleva annexes

käesoleva *lays down*

käesoleva *closely*

käesoleva *laid down*

käesoleva *representatives*

käesoleva *rulings*

käesoleva akti *subject*

käesoleva artikli kohaldamisel *purposes*

käesoleva artikli kohaldamisel *purpose*

käesoleva direktiivi mõistes *following*

käesoleva direktiiviga reguleeritavas valdkonnas *text*

käesoleva direktiiviga reguleeritavas valdkonnas *texts*

käesoleva jaotise title

käesoleva käsitusleppe *understanding*

käesoleva lepingu *treaty*

käesoleva lepingu *treaty.*

käesoleva lepingu teiste sätete *provisions*

käesoleva lõike alusel *paragraph*

käesoleva määruse *confidential information*

käesoleva määruse lisas loetletud *regulation*

käesoleva protokolli *article 1*

käesoleva põhikirja *statute*

käesolevaga *regulation*

käesolevaga *regulations*

käesolevaga *hereby*

käesoleval *question*

käesolevale *annexed*

käesolevale *party*

käesolevale direktiivile *reference*

käesolevale direktiivile *references*

käesolevale protokollile lisatud *schedule*

käesolevale protokollile lisatud *schedules*

käesolevas *by*

käesolevas *provided*

käesolevas *s&auml*

käesolevas *referred*

käesolevas *falling under*

käesolevas *set*

käesolevas *lays down*

käesolevas *questions*

käesolevas *laid down*

käesolevas lepingus *treaty*

käesolevas lõikes *paragraph*

käesolevas määruses *purposes*

käesolevas määruses ettenähtud meetmed *measures*

käesolevas määruses sätestatud meetmed *measures*

käesolevast lepingust *deriving*

käesolevast lepingust *member state*

käesolevat lepingut agreement

käesolevat lõiget paragraph

käibemaksu tobacco

käibemaksu turnover

käibemaksu-omavahendite vat

käigus pending before

käsitle national

käsitle nationals

käsitleb interim agreement

käsitleb scheduled

käsitleb specific measures

käsitletud dealt

käsitlev act concerning

käsitlev whereas

käsitleva act concerning

käsitleva exemptions

käsitleva protokolli kingdom

käsitlevad relating

käsitlevad sitlevad

käsitlevaid concerning

käsitlevas exemptions

käsitlevas services

käsitlevas section

käsitlevatwaste

käsitlevatconcerning

käsitlevate mandate

käsitlevate approximation

käsitlevates approximation

käsitsi handmade

käsituslepe understanding

käsitusleppe article

käsitusleppes understanding

kättesaadav nevertheless

kättesaadavad available

kättesaadavaid available

kättesaadavaks available

kõige most

kõigi every

kõigi liikmete membership

kõigil union

kõigil practical arrangements

kõigis president j.

kõigist &ldquo

kõik membership

kõikide copies

kõikide expansion

kõikidele liikmesriikidele member states

kõikidele liikmesriikidele member states.

kõikidest inform

kõnealune tribunal

kõnealune authority

kõnealuse question

kõnealuse representative

...