EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 3, 355–372

KAKSKEELSETE LEKSIKONIDE GENEREERIMINE PARALLEELKORPUSE BAASIL

Kaarel Veskis

Ülevaade. Paralleelkorpustel ja võrreldavatel korpustel on lisaks erinevate keelte või keelevariantide kontrastiivuuringutele ka mitmeid teisi kasutusvõimalusi nii teoorias kui praktikas, mitmeid neist võimalustest ollakse alles avastamas. Üks huvitavamaid suundi on sõnastike koostamine või täiendamine tõlkevastete ekstraheerimise läbi. Püüan järgnevalt anda ülevaate senisest tööst selles valdkonnas, pidades silmas eesti keelt puudutavate praktiliste rakenduste väljatöötamise võimalust tulevikus. Samuti kirjeldan katseid esialgse lahendusena kasutada inglise-eesti ja eesti-inglise erialasõnastiku lähtematerjali genereerimiseks väheseid keeleressursse (lisaks paralleelkorpusele) vajavat vabavarana saadaval olevat tarkvara.

Võtmesõnad: korpuslingvistika, arvutisõnastikud, kakskeelne leksikograafia, keeletöötlus, oskussõnastikud, eesti keel, inglise keel

1. Sissejuhatus

1.1. Paralleelkorpuste rakendused

Paralleelkorpus on korpus, mis sisaldab mingit teksti originaalkeeles ja selle tõlget teise keelde või tõlkeid teistesse keeltesse. 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.

1990. 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 eluvaldkondade jaoks.

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).

Veidi marginaalsemate, kuid seda huvitavamate praktiliste kasutusviisidena võiks sõnastike ekstraheerimise kõrval esile tuua mitmekeelse infootsingu (Davis 1998) ja sõnastike käsitlemise paralleelkorpustena või paralleelkorpuste osadena, et luua uusi ja täielikumaid tõlkevahendeid (Geisler 2002), või mõnel muul eesmärgil. 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.

1.2. Automaatselt loodavate leksikonide vajalikkus

Tuleviku seisukohalt on keeletehnoloogias üks kahest põhisuunast, millele kogu maailmas tähelepanu pööratakse, uute lähenemiste leidmine arvutis loetavate sõnastike loomiseks kas korpuste või muude allikate toel (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 luua üksnes leksikograafidele, terminoloogidele või tõlkesüsteemidele abiks olevat toormaterjali. Erialased tekstid sobivad leksikoni ekstraheerimiseks näiteks kirjanduslikest tekstidest paremini 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 muutes 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,[1] 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.

1.3. Tekstide paralleelistamine

Oskus kasutada levinud tõlkeabiprogrammides (nt Trados, MultiTrans) või korpuseanalüüsiprogrammides (Wordsmith) sisalduvaid terminiekstraheerimismooduleid (või nende moodulite tööprintsiipide tundmine) võib lisaks terminispetsialistidele suuresti abiks olla nii tõlkijatele, keeleõppijatele kui ka kõikidele tudengitele ning teadlastele, kes soovivad oma erialast terminoloogiat tundma ja kasutama õppida mõnes võõrkeeles. Lihtsat meetodit terminiandmebaasi moodustamiseks korpustest tuletatud sagedussõnastike ja võtmesõnade abil ilma spetsiaalse tarkvarata kirjeldavad Lynne Bowker ja Jennifer Pearson (2002: 137–164). Kakskeelse sõnastiku genereerimiseks on aga lisaks mõlema keele terminiandmebaasile vaja ka vastavuses olevad terminid paralleelistada.

Korrektne paralleelistamine on paralleelkorpuse hilisema kasutatavuse seisukohalt kõige olulisem ja ka üksnes indo-euroopa keeli hõlmava korpuse puhul kõige töömahukam probleem. Paralleelistamist raskendab asjaolu, et tihti sisaldavad paralleeltekstid "müra", s.t ühes tekstis on midagi rohkem või vähem kui temale vastavas teises tekstis, mistõttu ei saa lauseid omavahel üksüheselt kokku viia. Siiski on automaatne lausetasandil paralleelistamine võimalik rohkem kui 90-protsendilise täpsusega. Fraasi- ja sõnatasandil paralleelistamine on aga tunduvalt keerulisem. Sõnade järjekord lauses sõltub keelest ja mitmetes keeltes (sh eesti keeles) on sõnade järjekord üsna vaba. Leksikaalsete üksuste piire on palju raskem tuvastada kui lausepiire. Ometi on ka sõnadevaheliste tõlkevastavuste automaatne tuvastamine teataval määral võimalik ja sellel põhinebki kakskeelsete leksikonide genereerimine.

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.

[1] Saagise all on siinses artiklis mõeldud kõigi paralleelkorpuses sisalduvate ja omavahel tõlkelises seoses olevate sõnade või väljendite arvu suhet leksikoni jaoks ekstraheeritud korrektsete seoste arvuga. Täpsus on korrektsete tõlkevastavuste hulk leksikonis, võrrelduna kogu leksikoni mahuga.

2. Sõnastike genereerimine paralleelkorpustest

2.1. Sõnastike genereerimise meetodid

Erinevad lähenemised tõlkevastete automaatsele ekstraheerimisele jagunevad kahte põhikategooriasse: nn "hüpoteesi kontrollimise" ehk heuristilised meetodid (nt Smadja jt 1996) ja estimeerivad meetodid (nt Hiemstra 1997). 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)[2] 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 (ingl *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.

2.2. 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.

[2] IBM-i (International Business Machines) mudelite kohta vt alaptk 3.3.

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). Käesolevas artiklis pööran aga edaspidi 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.

3. Sõnastiku genereerimise praktilised võimalused

3.1. Eeldused ja eeltöötlus

Mitmed kommertstarkvara tootjad (Xerox, Ahead Software, SensoLogic, TRADOS) 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, s.t näiteks verbile ühes keeles võib vastata üksnes verb või mõni teine sõnaliik, mis on tunnistatud võimeliseks täitma tõlkes verbi funktsiooni –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 leksikaalne üksus 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 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 Tartu Ülikooli (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*. I. Blank (2000: 247) toob näite selle kohta, kuidas saksa keele liitsõnaline termin *Einspruchsbeschwerdeverfahren* esineb vastavates prantsusekeelsetes tekstides järjepidevalt kujul *procédure de recours engagée à l'éncontre d'une décision rendue sur opposition* (kompleksne noomenifraas). 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. 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.

3.2. 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 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).

3.3. Poolautomaatsed vahendid

Kui tarkvara väljatöötamine ei ole mingil põhjusel võimalik või otstarbekas, siis on sõnastiku genereerimiseks võimalik kasutada ka vabavara.[3] Kõige lihtsam viis sõnade käsitsi joondamiseks või ka paralleelteksti põhjal sõnastiku toormaterjali

[3] Vabavarana saadaval olevatest sõnatasandil paralleelistamist võimaldavatest tarkvarapakettidest ülevaate saamiseks vt http://www.cse.unt.edu/~rada/wa/#softwareWA (21.08.2006).

loomiseks on kasutada selleks mõnd graafilist joondamisvahendit, mis võimaldab biteksti üksteisega vastavuses olevad sõnad hõlpsasti omavahel ühendada ja genereerida automaatselt omavahel seostatud sõnade loendi.[4]

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[5], 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++[6] tööprintsiibid hõlmavad sarnaselt suurema osaga statistilise masintõlke rakendustest IBM-i uurijate mudeleid, mida tutvustasid Peter F. Brown jt (1993). Statistiline masintõlge rajaneb nn müranivooga kanali (ingl *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 ingliskeelse lause tõlge, on suurim. Selle tõenäosuse väljaarvutamiseks Bayesi valemi abil on tarvis teada ka tõenäosust, et mingi lause üldse eesti keeles võib esineda. Need tõenäosused saab leida piisavalt suure paralleelkorpuse alusel (Muischnek jt 2003: 53–54).

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).

3.4. 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.

[4] Vt nt http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm (21.08.2006)
[5] Vt http://demo.linearb.co.uk:8080/sandbox/start.jsp (21.08.2006)
[6] Vt http://www.fjoch.com/GIZA++.html (21.08.2006)

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[7] 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. Kakskeelse leksikoni genereerimine PWA abil TÜ paralleelkorpuse põhjal

4.1. 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[8] ühendasin leksikoni koostamise otstarbeks JRC-Acquis' mitmekeelse paralleelkorpuse[9] inglise-eesti alamosaga, kuna mõlemad korpused sisaldavad Euroopa Liidu seadusandlusega seotud tekste. Mõlemad korpused on samuti paralleelistatud Vanilla paralleelistaja[10] abil.

TÜ korpus oli algselt kujul, kus eesti ja ingliskeelsed 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 korpused teisendada kujule, milles mõlema keele üksused on ümber tõstetud kahte eraldi faili ja on tähistatud omavahel vastavuses olevate numbriliste tähistega. Selleks kasutasin osaliselt Camelia Ignat' poolt loodud Perli-programmi[11].

Katse[12] 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 482 571 sõnet. Sisendina kasutatud koondkorpus hõlmas 730 880 paralleelset ühest või mitmest lausest või (ala)pealkirjast koosnevat üksusepaari. Lisaks sisendkorpusele hõlmas programmi kasutajapoolne sisend ka väiksemahulist morfoloogiainfo faili

[7] Vt http://stp.ling.uu.se/plug/pwa/index.html (21.08.2006)
[8] Vt http://www.cl.ut.ee/korpused/paralleel/index.php?lang=et (21.08.2006)
[9] Vt http://langtech.jrc.it/JRC-Acquis.html (21.08.2006)
[10] Vt http://nl.ijs.si/telri/Vanilla/ (21.08.2006)
[11] Vt http://wt.jrc.it/lt/Acquis/JRC-Acquis.2.2/alignments/index.html (21.08.2006)
[12] 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 Giza++ sõnaparalleelistused koos Giza++ sisendiks olnud TÜ korpuse alamosaga.
[13] Vt http://www.eki.ee/dict/inglise/ (21.08.2006)

eesti keele tüüpiliste sõnalõppude ja ebareeglipäraste verbide kohta ning ka lähtesõnastikku, milleks otsustasin valida EKI inglise-eesti sõnastiku[13]. 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.

Esitan juhusliku parandusteta fragmendi (1) UWA poolt TÜ inglise-eesti paralleelkorpuse ja JRC-Acquis' mitmekeelse paralleelkorpuse inglise-eesti alamosa põhjal genereeritud leksikonist, mis ei ole esitatud UWA graafilise liidese vahendusel, vaid tavalise tekstifailikatkendina:

```
(1)
{reserve officer}
{
   1X:reservohvitseri
}
{reserve officer candidates}
{
   1X:reservohvitserikandidaat
}
{reserve officer courses}
{
   1X:reservohvitserikursusel
}
{reserve positions and}
{
   2X:reservipositsioonid ja
}
{reserve power system}
{
   1X:reservelektrisüsteem
}
{reserve ratio}
{
   3X:reservibaasi
   2X:reservimäära
}
{reserve service}
{
   1X:reservteenistus
```

Nagu näitest (1) 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, mis kehtivad samamoodi eesti-inglise keelepaari puhul nagu ka rootsi ja inglise keelte puhul (UWA on loodud arvestades rootsi keele eripärasid, kuid mitmeid väljundi vigasid ei ole suudetud ka rootsi keele puhul esialgu parandada).

Kuna UWA paneb tõlkevastavuste leidmisel suurt rõhku sõnesarnasusele, siis tulenevad sellest inglise ja eesti keele mittesuguluse tõttu paljud vead ja osalised vastavused[14] väljundis, näiteks:

[14] Võib oletada, et mitmed vead on tingitud ka sellest, et sisendkorpuse suure mahu tõttu ei suutnud programm katse käigus läbi viia viimast, automaatse filtreerimise etappi.

(2)
{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 eesmärgiga kahandada andmehõreduse efekti. S.t üks süsteemi alammoodul 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 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 (2002: 74) näeb süsteemi arendusvõimalustena lisaks selle täiendamisele reeglipõhiste meetoditega veel sõnastiku märksõnade lemmatiseerimist ning viidete lisamist genereeritud sõnastikukirjetelt kontekstidele korpuses.

4.2. 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.

Sisendkorpuse mahupiirangu tõttu kasutasin LWA-ga katselise sõnastikumaterjali genereerimiseks TÜ paralleelkorpuse 10 000 paralleelsest lõigust koosnevat alamosa. Toon siin ära fragmendi tulemuseks saadud leksikonist, mis sisaldas 8516 kirjet:

(3)

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

4.3. Tulemuste hindamine: ARCADE ja PWA

Joondamisprogrammide võrdleva hindamise standardiseerimisprojekti ARCADE (Véronis, Langlais 2000) raames on kindlaks tehtud, et parimate sõnadevahelise paralleelistamise 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, Merkel jt 2000: 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 UWA süsteemiga 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%).

4.4. 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 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 2006. aastal kasutatavate operatsioonisüsteemide ja muu vajaliku tarkvaraga (Java) ega ole seetõttu kasutuskõlblik. Seega tuleb standardfaili loomiseks leida mõni alternatiivne poolautomaatne võimalus või kontrollida käsitsi.

Ka PWA hindamismeetodi tulemused varieeruvad üsna palju sõltuvalt tekstide žanrist ja kuldstandardi koostamise kriteeriumitest – s.t sõltuvalt sellest, millise sagedusega ja funktsiooniga sõnu ja millistes proportsioonides valitakse nende näidete hulka, mille alusel paralleelistustulemusi hinnatakse (Ahrenberg jt 2000: 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 poliitikaalastele 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 leksikonist juhuslikult valitud 50 kirje kontrollimisel sain selle alamosa täpsuseks 60%. Samasugune kontroll LWA-ga TÜ eesti-inglise paralleelkorpuse põhjal genereeritud leksikoni 50-kirjelise alamosa peal andis täpsuseks 61%. 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.

Projekti Plug meetoditele väga sarnast lähenemist leksikoni genereerimisele esindavad Dan Tufiş ja Ana-Maria Barbu (2001), kes kasutasid leksikoni koostamiseks europrojekti Multext-East[15] 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. Kontrolliti 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 saab paralleelistada ka statistilise masintõlke rakenduste abil, eespool mainisin tarkvarapaketti Giza++. TÜ inglise-eesti korpuse paralleelistamise tulemuste kohta Giza++ abil vt K. Muischneki kirjutist (2006).

5. Kokkuvõte

Eesti keelt hõlmavate mitmekeelsete sõnastike koostamine ja täiustamine on üks peamisi teid tulevikus eesti keele väljatõrjumise takistamiseks mitmetest eluvaldkondadest. Seetõttu osutub väga oluliseks siinses artiklis tutvustatud töö jätkamine praktilisel tasandil leksikograafide ja keeletehnoloogide poolt. Kuigi esialgsed tulemused paralleelkorpustest leksikonide ekstraheerimise vallas jätavad veel kõvasti soovida, peaks siiski kõik mitmekeelsete leksikonide koostamisega tegelevad inimesed mõtlema selle peale, kuidas paremini ammutada automaatsete vahenditega leksikoni rikastavat infot nii tava- kui paralleelkorpustest.

[15] Vt http://nl.ijs.si/ME/ (21.08.2006)

Kirjandus

Ahrenberg, Lars; Andersson, Mikael; Merkel, Magnus 1998. A simple hybrid aligner for generating lexical correspondences in parallel texts. – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 10–14 August 1998. Montréal, Canada, 29–35.

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

Ahrenberg, Lars; Merkel, Magnus; Sågvall Hein, Anna; Tiedemann, Jörg 2000. Evaluation of word alignment systems. – Proceedings of LREC 2000. Vol. III. Athens/Greece, 1255–1261.

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

Borin, Lars 2002. ...and never the twain shall meet? – Lars Borin (Ed.). Parallel Corpora, Parallel Worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Amsterdam: Rodopi, 47–59.

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

Bowker, Lynne; Pearson, Jennifer 2002. Working with Specialized Language: A Practical Guide to Using Corpora. London/New York: Routledge.

Brew; Chris; McKelvie, David 1996. Word-pair extraction for lexicography. – K. Oazer, H. Somers (Eds.). Proceedings of the Second International Conference on New Methods in Language Processing. 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), 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 statistical machine translation: Parameter estimation. – Computational Linguistics 19 (2), 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).

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

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

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. University College Dublin. Belfield, Dublin, 67–74.

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

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.
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.
Fung, Pascale 2000. A statistical view on bilingual lexicon extraction – From Parallel Corpora to non-parallel corpora. – J. Véronis (Ed.). Parallel Text Processing: Alignment and Use of Parallel Corpora. Dordrecht: Kluwer, 219–236. http://citeseer.ist.psu.edu/fung98statistical.html (21.08.2006).
Gale, William; Church, Kenneth 1993. A program for aligning sentences in bilingual corpora. – Computational Linguistics 19 (1), 75–102.
Gaussier, E.; Hull, D.; Ait-Mokhtar, S. 2000. Term alignment in use: Machine-aided human translation. – J. Véronis (Ed.). Parallel Text Processing: Alignment and Use of Parallel Corpora. Dordrecht: Kluwer, 253–274.
Geisler, Christer 2002. Reversing a Swedish-English dictionary for the Internet. – Lars Borin (Ed.). Parallel Corpora, Parallel Worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Amsterdam: Rodopi, 122–133.
Genereeritud leksikonid. www.teataja.ee/leksikonid.zip (29.08.2006).
GIZA++. Training of statistical translation models. http://www.fjoch.com/GIZA++.html (21.08.2006).
Gliozzo, Alfio; Strapparava, Carlo 2005. Cross language text categorization by acquiring multilingual domain models from comparable corpora. – Proceedings of the ACL Workshop on Building and Using Parallel Texts (in conjunction of ACL-05). University of Michigan, Ann Arbor, 9–16.
Hwa, Rebecca; Madnani, Nitin 2004. The UMIACS Word Alignment Interface. http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm (21.08.2006).
Hiemstra, Djoerd 1997. Deriving a Bilingual Lexicon for Cross-Language Information Retrieval. – M. Heemskerk, M. Diepenhorst (Eds.). Proceedings of the 4th Groningen International Information Technology Conference for Students. University of Groningen, 21–26.
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 (21.08.2006).
Inglise-eesti-inglise sõnastik. http://www.eki.ee/dict/inglise/ (21.08.2006).
Inglise-eesti ja eesti-inglise paralleelkorpus. http://test.cl.ut.ee/korpused/paralleel/ (21.08.2006).
Kitsnik, Mare 2006. Keelekorpused ja võõrkeeleõpe. – Helle Metslang, Margit Langemets (toim.), Maria-Maren Sepper (keeletoim.). Eesti Rakenduslingvistika Ühingu aastaraamat 2. Estonian Papers in Applied Linguistics 2. Eesti Rakenduslingvistika Ühing. Tallinn: Eesti Keele Sihtasutus, 93–107.
Kuhn, Jonas 2004. Experiments in parallel-text based grammar induction. – Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. ACL, 470–477.
Linear B. Word alignment tool. http://demo.linearb.co.uk:8080/sandbox/start.jsp (21.08.2006).
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 8th 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.
Melby, Alan K. 2000. Sharing of translation memory databases derived from aligned parallel text. – J. Véronis (Ed.). Parallel Text Processing: Alignment and Use of Parallel Corpora. Dordrecht: Kluwer, 347–368.
Multext-East home page. http://nl.ijs.si/ME/ (21.08.2006).
Muischnek, Kadri; Orav, Heili; Kaalep, Heiki-Jaan; Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Haridus- ja Teadusministeerium, Eesti keelenõukogu. Tallinn: Eesti Keele Sihtasutus.
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).
Oakes, M. P.; McEnery, A. M. 1998. Bilingual text alignment: An overview. – A. M. McEnery, S. P. Botley, A. Wilson (Eds.). Multilingual Corpora in Teaching and Research. Amsterdam: Rodopi, 1–37.
Och, Franz Josef; Ney, Hermann 2000. Improved statistical alignment models. – Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hongkong, China, 440–447.
Sågvall Hein, Anna 2002. The PLUG-project. Parallel corpora in Linköping, Uppsala, Göteborg: Aims and achievements. – Lars Borin (Ed.). Parallel Corpora, Parallel Worlds. Language and Computers: Studies in Practical Linguistics nr. 43. Amsterdam: Rodopi, 61–78.
Smadja, Frank; McKeown, Kathleen R.; Hatzivassiloglou, Vasileios 1996. Translating collocations for bilingual lexicons: A statistical approach. – Computational Linguistics 22 (1), 1–38.
Software for word alignment. http://www.cse.unt.edu/~rada/wa/#softwareWA (21.08.2006).
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 2002. Uplug – a modular corpus tool for parallel corpora. – Lars Borin (Ed.). Parallel Corpora, Parallel Worlds. Language and Computers: Studies in Practical Linguistics nr. 43. 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).
Tufiş, Dan; Barbu, Ana-Maria 2001. Automatic Construction of Translation Lexicons. – V. V. Kluev, C. E. D'Attellis, N. E. Mastorakis (Eds.). Advances in Automation, Multimedia and Modern Computer Science. A Series of Reference Books and Textbooks in Electrical and Computer Engineering. WSEAS Press, 156–172.
Véronis, Jean; Langlais, Philippe 2000. Evaluation of parallel text alignment systems: The ARCADE project. – J. Véronis (Ed.). Parallel Text Processing: Alignment and Use of Parallel Corpora. Dordrecht: Kluwer, 369–388.

Kaarel Veskis töötab Tartu Ülikooli üldkeeleteaduse õppetooli juures, kus põhiülesandeks on eesti keele segakorpuse koostamine.
kaarel.veskis@ut.ee

GENERATION OF BILINGUAL LEXICONS FROM A PARALLEL CORPUS

Kaarel Veskis
University of Tartu

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 article 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, there have been reports of attempts 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 could be compiled from parallel corpuses only semiautomatically, 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 paralleling and dictionary generation on the basis of the Estonian grammatical structure and its possible points of equivalence with the structure of some other language. Although the quality of the lexicon to be generated would certainly be improved by preliminary morphological analyis of the parallel corpus, our 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 article 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 headwords and 482 571 word forms. The precision of a random sample of 50 entries turned out to be 60%. In addition the article 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.

Keywords: corpus linguistics, automatically created dictionaries, bilingual lexicography, language processing, technical dictionaries, Estonian, English