# PARALLEELKORPUSPÕHINE TÕLKEABISÜSTEEM INTERNETIS

KATRIN TSEPELINA, KAAREL VESKIS

## Sissejuhatus

Mitmekeelsete sõnastike käsitsimeetodil kirjehaaval loomine on mõnevõrra tüütu protsess, mida on võimalik hõlbustada, genereerides olemasolevate tekstide põhjal automaatselt sõnade ja nende tõlgete loendid. Seda on kõige lihtsam teha, kasutades lähtetekstina paralleelkorpust (originaaltekst ja selle tõlge on omavahel märgendusega seotud) ning mõnd sõnade joondamise ehk paralleelistamise vahendit, mis leiab tekstis esinenud sõnade jaoks mitmesuguste algoritmide alusel paralleeltekstist oletuslikud tõlkevasted. Nõnda loodud sõnapaaride loendit saab kasutada näiteks mitmekeelses infootsingus, masintõlkes, semantilise ühestamise hõlbustamiseks jne. Samuti saab selliste sõnaloetelude abil luua ja täiendada juba olemasolevaid sõnastikke. Mõne väga kitsa valdkonna puhul võib osutuda otstarbekaks üksnes automaatselt loodava sõnastiku koostamine (vt nt Andrenucci 2007).

Automaatselt loodud sõnastik ja lähtekorpus on mõistlik koondada ühtseks tõlkeabisüsteemiks. Selline süsteem meenutab mitmekeelset otsingumootorit või tõlkemälusüsteemi, kus päringu tulemusel esitatakse otsitav sõna ka teises keeles ning lisaks esitatakse lähtekorpusest pärinevad sõna ja selle tõlkevastet sisaldavad paralleelkonkordantsid ehk näitelaused. Need annavad aimu sõna kasutamisdiapasoonist ja neid saab valikuliselt kasutada loodavas sõnastikuartiklis.

Oskusleksikograafia jaoks on väga oluline mitmekeelsus, sest terminoloogia peab olema ühtlane ja peab valitsema selgus selles osas, mis on ühe või teise termini vaste laiema levikuga keeltes. Uuringus „Eesti oskuskeelekorralduse seisund" (Erelt, Tavast 2003) väljendatakse arusaama, et keelekorpuste kasutamine on eesti keele oskusleksikograafias küsitav piisava mahu ja spetsialiseeritusega korpuste puudumise tõttu. Siiski sunnivad järjest täienevad keeleressursid ja infotehnoloogia arengud arvama, et ka eesti keele puhul haakub mitmekeelsete sõnastike koostamine nii või teisiti üha tihedamalt keelekorpustega, eelkõige omavahel tõlkelises vastavuses olevate paralleelkorpustega. Püüamegi muuhulgas selgitada, kas ja kui suurel määral konkreetne erialakeelt sisaldav paralleelkorpus koos spetsiaalse tarkvaraga võimaldab koostada või täiendada mitmekeelset erialakeelesõnastikku.

Nimetatud uuringus (Erelt, Tavast 2003) räägitakse ka eestikeelse arvutiala sõnavara arendamise olulisusest. Muuhulgas on refereeritud Ülo Jaaksoo seisukohta, et eesti keelde oleks vaja tõlkida põhilised koolides kasutatavad arvutiprogrammid, aga arvutivaldkonna ülikiire arengu tõttu ei ole võimalik kõiki programme eestindada.

Eestikeelse arvutitarkvara olemasolu ja arendamine on eesti keele positsiooni säilitamise jaoks esmatähtis ja tihedalt seotud riikliku programmi

„Eesti keele keeletehnoloogiline tugi (2006–2010)" (EKKTT) eesmärkidega. Tarkvaraalasest paralleelkorpusest või tõlkemälust genereeritud sõnastikud võimaldavad osaliselt automatiseerida tarkvara eestindamist, mistõttu on lühema ajaga võimalik eesti keelde tõlkida rohkem arvutiprogramme. Et automaatselt genereeritud sõnastik hõlbustab hoolimata osaliselt vigastest kirjetest tõlkimisprotsessi, on kasutajauuringuga tõendanud Wanwisa Khanaraksombat ja Jonas Sjöbergh (2007).

Artiklis kirjeldatud töö tulemuseks ongi selline tõlkeabisüsteem, mida saab kasutada olemasolevate arvutialaste sõnastike täiendamiseks või ka spetsiaalse tarkvarateemalise sõnastiku loomiseks. Samuti võib meie kirjeldatud meetodit kasutada tarkvara emakeelseks muutmise hõlbustamiseks. Sõnastik on internetis kasutatav, peale tõlkevastete saab sellest päringu tulemusena andmeid ka sõna reaalse kasutamise kohta (kui sagedasti on sõna tõlgitud üheks või teiseks terminiks, milline on tõlgete kontekst jne). Sõnastiku automaatseks genereerimiseks kasutasime vabavaralist tarkvarasüsteemi Uplug. Korpuse ja sõnastiku järeltöötluse ning indekseerimise käigus valminud tõlkeabisüsteemile saab päringuid esitada internetiaadressil *http://kde.teataja.ee*.

## Senistest töödest

Sõnastike automaatseks loomiseks on välja töötatud ja katsetatud erinevaid sõnaparalleelistussüsteeme (vt ülevaadet Veskis 2007).

Et meie eesmärgiks oli eesti ja inglise keele sõnatasandil joondamine, siis pakuvad meie tööle parimat võrdlusmaterjali sellised varasemad uurimused, kus on samuti joondatud keeli, mis ei ole omavahel suguluses.

Uplugi tarkvara abil on proovitud genereerida kreeka-inglise sõnastikku (Charitakis 2007). Selle käsitsi hindamisel ei arvestatud tõlkepaare, mille esinemissagedus oli alla kolme. Ülejäänud tõlkepaarid jagati viide rühma, lähtudes nende esinemissagedustest, ja kõiki rühmi vaadeldi hindamisel eraldi. Selgus, et tõlkepaaride esinemissagedus ja tõlke korrektsus on genereeritud sõnastiku puhul otseses proportsionaalses vastavuses: suurema esinemissagedusega tõlkepaaride korrektsuse protsent on suurem. Kasutatud paralleelkorpus oli suhteliselt väike, mõlemas keeles *ca* 200 000 sõna. Saadud sõnastiku alamosa, mille suurus oli 498 sõnapaari, hindasid eksperdid. Kolme ja enama sõnapaari korral oli korrektselt tõlgitud 51 %, sagedusega 11 ja enama sõnapaari korral 67 % (Charitakis 2007).

Beata B. Megyesi ja Bengt Dahlqvist (2007) kirjeldavad Uplugi abil genereeritud rootsi-türgi sõnastikku. Kasutatud paralleelkorpus hõlmas 150 000 rootsi ja 130 000 türgi sõna, korpuse tekstid analüüsiti morfoloogiliselt, tulemsõnastiku täpsuseks oli 69 %.

Sumithra Velupillai ja Hercules Dalianis (2008) genereerisid Uplugi abil skandinaavia keelte ja soome keele sõnastikke, kasutades väga väikest paralleelkorpust. Rootsi-taani, rootsi-norra ja taani-norra sõnastikud sisaldasid 93 % ulatuses korrektseid tõlkevasteid, kuid soome-rootsi, soome-taani ja soome-norra sõnastikud olid ainult 67,4 % ulatuses korrektsed. Tõenäoliselt näitab see, et sugulaskeeli on omavahel tunduvalt lihtsam sõnatasandil automaatselt joondada.

H. Xing ja X. Zhang (2008) genereerisid Uplugi abil hiina-inglise sõnastiku, hinnates saadud sõnastiku täpsuseks umbes 74 %.

Uplugi abil loodud kitsa valdkonna väga väikese paralleelkorpuse põhjal loodud sõnastikku on rakendatud psühholoogia- ja psühhoteraapiateemalise meditsiiniportaali küsimuse-vastuse süsteemi inglis- ja rootsikeelses osas (Andrenucci 2007).

Uplugiga sarnaseid meetodeid on kasutanud sõnastike genereerimisel näiteks Eduardo Cendejas jt (2009), rakendades sõnastiku kvaliteedi parandamiseks huvitava lisavõimalusena ka süntaktilist märgendamist ja semantilisi tähendusvälju.

Inglise-eesti sõnastikku prooviti Uplugi abil juba genereerida ühes varasemas katses (Veskis 2007). Tookord oli lähtekorpuseks Tartu Ülikooli inglise-eesti paralleelkorpuse ja JRC-Acquis' paralleelkorpuse ühend ja saadud sõnastiku täpsuseks saadi juhuslikult valitud 50 kirje kontrollimisel 60 %. Selles artiklis kirjeldatavast tööst erinevalt ei kasutatud siis korpuse paralleelistamisel lemmatiseerimist ega morfoloogilist märgendamist. Kasutatud paralleelkorpuse paralleelistus oli võrreldes KDE korpusega märksa kehvema kvaliteediga ning korpuse suurem maht tekitas probleeme (nt väga pikk andmete töötlemise aeg, vead väljundis). Tookordne katse piirdus üksnes sõnastiku loomisega, nüüd on loodud aga ka sõnastiku baasil toimiv tõlkeabisüsteem, mis hõlmab peale sõnastiku ka lähtekorpuse näitelauseid.

Paralleelkorpuse kasutajaliidesed, mis võimaldavad leida paralleelkonkordantse, on kas tõlkemälusüsteemidesse integreerituna, eraldi programmidena (nt ParaConc) või internetiteenustena (nt TransSearch) kujunemas professionaalsete tõlkijate ja keeleuurijate asendamatuks abivahendiks (vt nt Simard ja Macklovitch 2005).

## Kasutatud korpused

Sõnastiku genereerimise algmaterjalina kasutasime kahte korpust: 1) KDE tarkvara dokumentatsioon eesti ja inglise keeles (*http://websvn.kde.org/trunk/l10n-kde4/et/docmessages*; edaspidi: dokumentatsioonikorpus); 2) KDE tarkvara kasutajaliidese ja rakenduste tekstilised osad eesti ja inglise keeles (*http://websvn.kde.org/trunk/l10n-kde4/et/docmessages*; edaspidi: tarkvarakorpus). Mõlema korpuse ingliskeelse osa on tõlkinud eesti keelde vabatahtlikud entusiastid Hasso Tepperi koordineerimisel.

Need kaks paralleelkorpust ühendasime, mille tulemusel saime 190 000 paralleelüksusega korpuse: inglise keeles 1,5 miljonit sõna, eesti keeles 1,1 miljonit sõna. Korpuse joondamiseks kasutasime tarkvara Hunalign (*http://mokk.bme.hu/resources/hunalign*) ja Maligna (*http://sourceforge.net/projects/align*).

## Uplug

Sõnade joondamiseks ja lähtekorpuse eeltöötlemiseks kasutasime Uplugi (Tiedemann 2003a). Uplug on rootsi teadlaste väljatöötatud vabavaraline tarkvarasüsteem, millesse on ühendatud erinevad statistilisi meetodeid kasutavad joondamismoodulid, sh GIZA++ (*http://www.fjoch.com/GIZA++.html*), kuid sõnade joondamisel on võimalik kasutada ka lingvistilist infot.

Sõnastike genereerimiseks kasutasime sõnade joondamisel Uplugi (versioon 0.2.0d) vaikimisi seadistust (vt Tiedemann 2003a). Iga seost saab defineerida tõenäosusena, mis iseloomustab sõna või sõnarühma ja paralleelteksti sõna või sõnarühma omavahelist seotust. Sõnade joondamisel võetakse põhiliselt arvesse staatilisi seoseid, mille väärtus jääb konstantseks iga sõnapaari puhul, ja dünaamilisi seoseid, mida korrigeeritakse joondamise käigus. Staatiliste seoste hulka kuuluvad koosesinemuse koefitsiendid (Dice'i koefitsient, vt Tiedemann 1999), sõnesarnasuse koefitsiendid (LCSR, Melamed 1995) ja GIZA++ seosed (Och, Ney 2003), mis lähtuvad IBM-i teadlaste statistilise masintõlke paradigmast (Brown jt 1993). Dünaamilised seosed on morfoloogiliste märgendite esinemismustrid, fraaside liigid ja sõnade suhtelised asukohad lauses. Sõnade joondamist alustatakse staatiliste seoste abil. Saadud paralleelistusi kasutatakse seejärel treeningmaterjalina, mille alusel tekitatakse dünaamilised seosed. Dünaamilisi seoseid kasutatakse joonduse parandamiseks (Tiedemann 2003b).

## Korpuse eeltöötlus

Peale Uplugi kasutasime korpuse eeltöötluseks UNIX-i ja Perli skripte ja morfoloogilisi analüsaatoreid, korpuse eestikeelse osa lemmatiseerimiseks ja morfoloogiliseks analüüsiks tarkvara Estmorf (Kaalep, Vaino 2000) ning inglise osa analüüsisime Connexori tarkvaraga (*http://www.connexor.eu/technology/machinese/demo*).

Kogu korpuse teisendasime selliseks kujuks, kus iga sõnaga oli liidetud selle sõna lemma ja tekstis esinenud sõnavormi morfoloogiline märgend. Lemmatiseerimata paralleelkorpuse põhjal saab genereerida sõnastiku, kus iga sõnavorm on eraldi oma teise keele vastega vastavusse viidud, mis on aga üldjuhul üleliigne (näide 1):

(1)
| | |
|---|---|
| sünnipäeva | birthday |
| sünnipäevad | birthdays |
| sünnipäevade | birthdays |
| sünnipäev | birthdate |
| sünnipäev | birthday |
| sünnipäevi | birthdays |

Näites 1 toodud sõnastikufragment peaks taanduma paralleelkorpuse lemmatiseerimise korral kahele kirjele (näide 2):

(2)
| | |
|---|---|
| sünnipäev | birthdate |
| sünnipäev | birthday |

Sõnade morfoloogilise märgendamise eesmärk oli anda Uplugile joondamist parandavat lisainfot, et tugevdada eesti ja inglise sama sõnaliiki sõnade seoseid.

## Genereeritud sõnastikud

Paralleelkujul korpusest oli võimalik genereerida sõnapaaride sagedustega (esimene tulp) ja tõenäosustega (viimane tulp) sõnastikufailid (näide 3):

(3)
| | | |
|---|---|---|
| 9546 ja \| J | and \| J | 0,20714299 |
| 7400 või \| J | or \| J | 0,17058817 |
| 5581 ole \| V | be \| V | 0,17102984 |
| 4355 see \| P | this \| P | 0,15122719 |
| 4279 fail \| SSG | file \| SSG | 0,22292496 |
| 4255 ei \| VNEG | not \| VNEG | 0,16313266 |
| 3443 uus \| A | new \| A | 0,22313541 |
| 3084 kõik \| P | all \| P | 0,15615674 |
| 3042 kasuta \| V | use \| V | 0,16462792 |
| … | | |

Kokku genereerisime paralleelkorpuse põhjal kuus erinevat sõnastikuversiooni (ja igast versioonist inglise-eesti ja eesti-inglise variandi):

1) sõnastik esialgsel kujul, töötlemata paralleelkorpuse põhjal;

2) sõnastik lemmatiseeritud paralleelkorpuse põhjal;

3) sõnastik lemmatiseeritud korpuse põhjal, kus igale lemmale on lisatud sõnaliigi märgend;

4) sõnastik lemmatiseeritud ja sõnaliigi märgenditega varustatud korpuse põhjal, kasutades parema kvaliteedi saavutamiseks abivahendina EKI inglise-eesti lemmatiseeritud sõnastikku;

5) eelmise töötlusjärgu sõnastik, aga lähtekorpuses on ühtlustatud morfoloogilist infot ja sealt on kustutatud inglise keele artiklid (*a, an, the*);

6) eelmise töötlusjärgu sõnastik, aga lähtekorpusest on kustutatud osa eessõnadest (*with, to, on, of, in, by, for, at, from*), millele eesti keeles vastavad enamasti käändelõpud. Suur hulk eelmise töötlusjärgu sõnastikus leidunud sisulistest vigadest tundus olevat seotud nende eessõnadega sarnaselt Uplugiga loodud rootsi-türgi sõnastikuga, mille puhul rootsi eessõnu ei õnnestunud ühendada türgi keele eessõnadeta vastetega (Megyesi, Dahlqvist 2007).

Uplug võimaldab sõnade joondamise tulemuse parandamiseks kasutada joondamisfaasis abisõnastikku. Inglise-eesti sõnastiku genereerimisel kasutasime abisõnastikuna elektroonilist EKI inglise-eesti sõnastikku (*http://www.eki.ee/dict/inglise*).

## Genereeritud sõnastike hindamine tõenäosuste alusel

Et genereeritud sõnastikud sisaldasid mitmesuguseid süstemaatiliselt esinevaid vigu, siis kustutasime sõnastike järeltöötlusena sõnastikust kõik mitmesõnalisi üksusi sisaldavad tõlkepaarid, mitmesuguseid mittealfabeetilisi sümboleid sisaldavad vigased tõlkepaarid jms. Tõlkeabisüsteemiga liitmisel ei ole mitmesõnalised üksused olulised, sest ühesõnalise päringu tulemusel esitatakse näitelausetes ka sõnade kontekst.

Pärast süstemaatiliselt esinenud vigade ja mitmesõnaliste üksuste automaatset kustutamist sisaldas viimase (kuuenda) töötlusjärgu eesti-inglise

sõnastik 21 459 kirjet ja inglise-eesti sõnastik 22 125 kirjet. Juhusliku 200-kirjelise fragmendi täpsuseks saime selle manuaalsel hindamisel eesti-inglise sõnastiku puhul *ca* 89 % ning inglise-eesti sõnastiku puhul *ca* 96 %.

Võrdlemisel kasutasime sõnastikke sellisel kujul, millest ei olnud tõenäosuste alusel tõlkepaare välja filtreeritud, küll aga oli kustutatud ühe korra esinenud tõlkepaarid, mitmesõnalisi üksusi sisaldavad tõlkepaarid ja mitmesuguseid mittealfabeetilisi sümboleid sisaldavad vigased tõlkepaarid.

Genereeritud sõnastike kõrvutamisel ilmnes, et iga korpuse töötlemise etapiga oli saavutatud vastava korpuseversiooni põhjal genereeritud sõnastiku kvaliteedi tõus.

Mõne etapi puhul oli tulemuse paranemine selge ja etteaimatav. Näiteks oli varasemalt tõestatud, et baaskorpuse lemmatiseerimine parandab Uplugiga genereeritava sõnastiku kvaliteeti (eriti saagist, mõnevõrra ka täpsust) inglise-rootsi keelepaari puhul (Strömbäck 2005), ja seetõttu võis oletada, et sama kehtib ka inglise-eesti keelepaari korral.

Võrdlesime abisõnastikuga ja abisõnastikuta genereeritud eesti-inglise leksikone. Kõige silmatorkavamaks erinevuseks oli, et abisõnastikku kasutades joondatakse tegelikud tõlkevasted üldiselt sagedamini õigesti kui abisõnastikku kasutamata. Näiteks abisõnastikku kasutades joondati sõna *teadma* 134 juhul oma ingliskeelse vastega *know*, aga ilma abisõnastikuta ainult 73 juhul.

Ka näitas võrdlus, et inglise eessõnade paralleelkorpusest kustutamine osutus sõnastiku genereerimise seisukohalt tõepoolest kasulikuks.

Visuaalsel võrdlemisel oli näha, et ka teiste korpuse eeltöötluse sammudega oli saavutatud tulemsõnastiku kvaliteedi paranemine. Et seda arusaama arvuliselt kinnitada, moodustasime sõnastiku, mis hõlmas üksnes neid tõlkepaare, mis olid ühised kõigile kuuele genereeritud sõnastikuversioonile. Tõlkepaaride unikaalsed tunnused võimaldasid seejuures kokku viia kõigi korpuse töötlusjärkude põhjal genereeritud sõnastike tõlkepaarid, mis kohati võisid olla üsna erineval kujul (nt lemmatiseeritud *vs*. lemmatiseerimata jne), kuid sisuliselt samad.

Sellesse ühisossa kuuluvate ja mittekuuluvate tõlkepaaride võrdlus kinnitas meie oletust, et ühisossa kuuluvates tõlkepaarides on sõnadel suurem sisuline vastavus kui ühisossa mittekuuluvates tõlkepaarides. Seega on võimalik seda ühisosa teatavate mööndustega vaadelda nn kuldstandardina, korrektsete tõlkepaaride loendina, mida sageli kasutatakse sarnaselt infootsingus levinud meetoditega ka automaatselt genereeritud sõnastike kvaliteedi hindamiseks.

Selgus, et Uplugiga igale tõlkepaarile omistatud tõlkepaari korrektsust iseloomustav lõpptõenäosus on tõepoolest korrelatsioonis ka genereeritud inglise-eesti ja eesti-inglise sõnastikes esinevate tõlkepaaride sisulise korrektsusega: mida suurem on tõenäosust näitav arv, seda kindlamini on tegemist õige tõlkepaariga. Seega oletasime, et saame genereeritud sõnastikuversioone omavahel võrrelda nõnda, et summeerime iga sõnastiku tõlkepaaride tõenäosused, jagame summad tõlkepaaride arvuga ja võrdleme omavahel saadud keskmisi tõenäosusi. Seejuures võtsime iga sõnastiku puhul tõenäosuste summeerimisel arvesse ainult neid tõlkepaare, mis kuulusid kõigi töötlusjärkude sõnastikuversioonide ühisossa. Inglise-eesti sõnastike puhul oli ühisosa suurus 111 728 ja eesti-inglise sõnastike puhul 133 137 tõlkepaari.

Selles ja järgnevates võrdlustes kasutasime ilma meiepoolse järeltöötluseta sõnastikke täpselt Uplugi genereeritud kujul. Nende sõnastike mahud ja tulemuseks saadud ühisosade keskmised tõenäosused on järgmised (vt tabelit 1).

Tabel 1.

**Genereeritud sõnastike võrdlus ühisossa kuuluvate tõlkepaaride keskmiste tõenäosuste alusel**

| Sõnastiku genereerimisel kasutatud korpuse töötlusjärk | Keelesuund | Kirjeid sõnastikus | Sõnastiku kirjetest ühisosas (%) | Ühisosa keskmine tõenäosus |
|---|---|---|---|---|
| I (esialgsel kujul) | inglise-eesti | 85 6244 | 13,05 | 0,112 |
| | eesti-inglise | 80 4955 | 13,88 | 0,110 |
| II (lemmatiseeritud) | inglise-eesti | 700 872 | 15,94 | 0,144 |
| | eesti-inglise | 671 194 | 16,65 | 0,139 |
| III (lisaks sõnaliigid) | inglise-eesti | 648 746 | 17,22 | 0,146 |
| | eesti-inglise | 641 641 | 17,41 | 0,142 |
| IV (lisaks abisõnastik) | inglise-eesti | 646 226 | 17,29 | 0,151 |
| | eesti-inglise | 639 865 | 17,46 | 0,147 |
| V (lisaks ilma artikliteta) | inglise-eesti | 618 808 | 18,06 | 0,153 |
| | eesti-inglise | 615 266 | 18,16 | 0,149 |
| VI (lisaks ilma eessõnadeta) | inglise-eesti | 590 052 | 18,94 | 0,154 |
| | eesti-inglise | 588 429 | 18,99 | 0,150 |

Nagu tabelist näha, suureneb keskmine tõenäosus tõepoolest iga korpuse töötlusjärguga. Seejuures suureneb keskmine tõenäosus hüppeliselt pärast korpuse lemmatiseerimist, mis näib seega olevat kvaliteedi seisukohalt kõige olulisem samm. Täiendav katse näitas, et sõnastike ühisosa saab tõenäoliselt võrdluses indikaatorina arvesse võtta ka kirjete protsentuaalse osatähtsuse seisukohalt. Nii eesti-inglise kui ka inglise-eesti sõnastikus on igas töötlusjärgus eelmisega võrreldes üldiselt protsentuaalselt rohkem ühisosasse kuuluvaid tõlkepaare.

## Genereeritud sõnastike omavaheline võrdlus IT-sõnastike abil

Et genereeritud sõnastike lähtematerjal on mitmel moel seotud eelkõige arvutitarkvaraga, siis valisime võrdluseks välja ühe üldsõnastiku ja neli IT-sõnastikku. Need olid EKI inglise-eesti sõnastik *(http://www.eki.ee/dict/inglise/)*,

Heikki Vallaste „E-teatmik" (*http://vallaste.ee/*), „Arvutikasutaja sõnastik" (Hanson, Tavast 2005), „Arvutisõnastik" (Liikane, Kesa 2006) ning IT terministandardi projekti (1998–2001) sõnastik (*http://www.keeleveeb.ee/dict/speciality/itstandard/*). Viimased kolm sõnastikku on internetis kasutatavad Keeleveebi kaudu (*http://www.keeleveeb.ee/*).

Sõnastike automaatseks võrdlemiseks viisime nimetatud sõnastikud korpusest genereeritud sõnastikega sarnasele kujule, kus ühel real on ainult üks ingliskeelne termin ja selle eestikeelne vaste.

Alljärgnevate tabelitega püüame anda ülevaate võrdluses kasutatud sõnastike omavahelistest seostest arvudes väljendatuna. Perli skriptiga leitud kattuvate ja unikaalsete kirjete määrad kirjeldavad ühtlasi üsna ilmekalt olemasoleva eestikeelse IT-alase terminoloogia hetkeseisu. Samal ajal ei pretendeeri siinne võrdlus absoluutsele täpsusele: välja on jäetud pikemad seletused, mis võivad sisaldada tõlkevasteid, mis võrdluses ei kajastu.

Elektroonilised IT-sõnastikud kattuvad üksteisega meie andmete järgi suhteliselt suures ulatuses, kuid iga sõnastik sisaldab arvestataval määral ka unikaalseid kirjeid, mida teistes ei leidu. EKI inglise-eesti sõnastik sisaldab üldkeelesõnastikule kohaselt võrdlemisi vähe IT-valdkonna terminoloogiat.

Tabelitest 2 ja 3 on näha, et kõige suurem absoluutarvuline kattuvus (3587) on „Arvutisõnastiku" ja „Arvutikasutaja sõnastiku" vahel, kuid IT terministandardi projekti sõnastik koosneb tervelt 72 % ulatuses „Arvutisõnastiku" kirjetest ja ainult 23 % ulatuses „Arvutikasutaja sõnastiku" kirjetest.

Tabel 2.

**IT-sõnastike ja EKI inglise-eesti sõnastiku kattuvate kirjete arv***

| | EKI inglise-eesti sõnastik | „Arvuti-sõnastik" | „Arvuti-kasutaja sõnastik" | IT termini-standardi sõnastik | „E-teat-mik" |
|---|---|---|---|---|---|
| EKI inglise-eesti sõnastik | **136 013** | 1417 | 1119 | 482 | 654 |
| „Arvuti-sõnastik" | 1417 | **14 275** | 3587 | 2925 | 1659 |
| „Arvuti-kasutaja sõnastik" | 799 | 3587 | **5640** | 929 | 1103 |
| IT termini-standardi sõnastik | 262 | 1630 | 767 | **4068** | 469 |
| „E-teatmik" | 654 | 1659 | 1103 | 469 | **7693** |

* Poolpaksus kirjas on kirjete absoluutarvud.

Tabel 3.

**IT-sõnastike ja EKI inglise-eesti sõnastiku omavaheline kattuvus protsentuaalselt***

| | EKI inglise-eesti sõnastik | „Arvuti-sõnastik” | „Arvuti-kasutaja sõnastik” | IT termini-standardi sõnastik | „E-teat-mik” |
|---|---|---|---|---|---|
| EKI inglise-eesti sõnastik | **136 013** | 9,93 | 19,84 | 11,85 | 8,50 |
| „Arvuti-sõnastik” | 1,04 | **14 275** | 54,50 | 71,90 | 21,57 |
| „Arvuti-kasutaja sõnastik” | 0,59 | 25,13 | **5640** | 22,84 | 14,34 |
| IT termini-standardi sõnastik | 0,19 | 11,42 | 13,60 | **4068** | 6,10 |
| „E-teatmik” | 0,48 | 11,62 | 19,56 | 11,53 | **7693** |

* Poolpaksus kirjas on kirjete absoluutarvud.

Tabel 4 näitab, et protsentuaalselt kõige rohkem unikaalseid kirjeid sisaldab H. Vallaste „E-teatmik”, kuid tuleb arvesse võtta, et see sõnastik sisaldab Heikki Vallaste enda sõnul ka kaheldava väärtusega tõlkevasteid, mis on sõnastikku võetud seletaval või teiste vastete täiendamise eesmärgil.[1] Kõige rohkem teistest sõnastikest puuduvaid kirjeid on „Arvutisõnastikus”.

Kokku sisaldavad need neli sõnastikku unikaalseid kirjeid 15 513, mis on umbes poole vähem nende sõnastike kirjetest summeerituna.

Tabelist 5 selgub, et tarkvaraalasest korpusest genereeritud ja järeltöötluse läbinud sõnastik sisaldab suhteliselt palju (9437) selliseid tõlkepaare, mis puuduvad nii mahukast üldsõnastikust kui ka spetsiifilistest IT-alastest sõnastikest ja teatmikest. Automaatselt loodud sõnastiku unikaalsete tõlkepaaride tõlkekorrektsus on küll selgelt väiksem võrreldes sellesama sõnastiku ja teiste sõnastike kattuvate tõlkepaaridega, kuid sellest hoolimata arvame, et just automaatselt loodud sõnastiku unikaalsed tõlkepaarid ja neid sisaldav internetis kasutatav tõlkeabisüsteem on arvuti- või tarkvaraalase sõnastiku looja või täiendaja jaoks huvitav ja kasulik materjal.

Seejuures näib meie unikaalsete tõlkepaaride potentsiaal peituvat eeskätt teistes sõnastikes juba leiduvate märksõnade tõlkevastete täiendamises, vähem päris uute märksõnade loomises. Kui võrrelda näites esitatud automaatse sõnastiku unikaalsete tõlkepaaride loendi ingliskeelset poolt EKI

[1] Heikki Vallaste e-kiri K. Veskisele 19. II 2010.

Tabel 4.

**Unikaalsete kirjete osakaal IT-sõnastikes***

| | Kokku kirjeid | Unikaalseid kirjeid | Unikaalseid kirjeid protsentuaalselt |
|---|---|---|---|
| „Arvuti-sõnastik” | 14 275 | 7791 | 55 |
| „Arvuti-kasutaja sõnastik” | 5640 | 1651 | 29 |
| IT termini-standardi sõnastik | 4068 | 706 | 17 |
| „E-teatmik” | 7693 | 5365 | 70 |
| Kokku | 31 676 | 15 513 | 49 |

* Unikaalsed on siin need kirjed, mis puuduvad täpselt samasugusel kujul teistest tabelis nimetatud sõnastikest.

Tabel 5.

**Unikaalsete kirjete osakaal**

| | Kokku kirjeid | Unikaalseid kirjeid | Unikaalseid kirjeid protsentuaalselt |
|---|---|---|---|
| EKI inglise-eesti sõnastik | 144 224 | 139 071 | 96 |
| „Arvuti-sõnastik” | 14 275 | 7410 | 52 |
| „Arvuti-kasutaja sõnastik” | 5640 | 1446 | 26 |
| IT termini-standardi sõnastik | 4068 | 683 | 17 |
| „E-teatmik” | 7693 | 5191 | 67 |
| Korpusepõhine sõnastik pärast järeltöötlust | 14 882 | 9437 | 63 |
| Kokku | 190 782 | 163 238 | 86 |

inglise-eesti sõnastikuga, siis selgub, et kõik need sõnad esinevad ka EKI sõnastikus märksõnadena, kuid EKI sõnastikus (ega ka IT-sõnastikes) ei ole meie näite eestikeelseid vasteid.

(4) …

| | |
|---|---|
| handshake | tagasiside |
| hands-on | näitlik |
| ... | |
| handyman | käsitööline |
| handy | mugav |
| hang | hanguma |
| hangman | pooja |
| hangman | poomismäng |
| hangup | hangumine |
| hangup | lahutamine |
| han | hani |
| ... | |

Tabelist 6 ja 7 ilmneb, et kuigi iga korpusetöötluse etapiga vähenes genereeritud sõnastikus kirjete arv, suurenes siiski üldiselt iga etapiga genereeritud sõnastiku ning kõigi teiste võrdluses kasutatud sõnastike ühiste kirjete arv. Ka EKI inglise-eesti sõnastikku ja IT-sõnastikke võib vaadelda nn kuldstandardina ja seega näitab kattuvate kirjete arvu igakordne suurenemine eeskätt seda, et sõnastikes leiduvate korrektsete tõlkepaaride arv suurenes iga töötlusjärgu käigus tehtud korpuseteisenduste tulemusel, aga annab aimu ka määrast, mille võrra igast töötlusjärgust sõnastiku parandamisel kasu oli. Võrdluse tulemustest selgub, et kõige efektiivsem oli tulemsõnastiku kvaliteedi seisukohalt paralleelkorpuse lemmatiseerimine ja abisõnastiku kasutamine, mille tulemuseks oli teiste sõnastikega ühiste kirjete hüppeline kasv genereeritud sõnastikus. Korpuse ülejäänud töötlusjärgud küll parandasid mõnevõrra tulemust, kuid saavutatud efekt jäi marginaalseks. Korpuse töötlusjärkudest lahus tuleb vaadelda sõnastiku järeltöötlust, mis küll mõnevõrra kahandas meie automaatse sõnastiku ja teiste sõnastike ühiste kirjete arvu, kuid samal ajal on ebakorrektsete kirjete väljafiltreerimisega tublisti suurenenud ühiste kirjete protsentuaalne osakaal.

EKI inglise-eesti sõnastik sisaldab küll väga vähe teistes IT-sõnastikes leiduvaid kirjeid, aga tabelist 7 näeme, et korpusepõhine järeltöötluse läbinud sõnastik kattub tervelt 20 % ulatuses EKI inglise-eesti sõnastikuga ja tunduvalt vähem IT-sõnastikega. See on ühelt poolt seotud EKI inglise-eesti sõnastiku palju suurema mahuga, kuid teisalt näitab see ka, et KDE tarkvara kasutajaliideste korpus sisaldab suhteliselt palju üldkeelt. Seega võib korpusepõhine sõnastik ja selle kasutamist hõlbustav tõlkeabisüsteem üht-teist pakkuda ka üldkeelesõnavara uurijale.

Tabel 6.

**Genereeritud sõnastike ning EKI sõnastiku ja IT-sõnastike kattuvate kirjete arv**

| Sõnastiku genereerimisel kasutatud korpuse töötlusjärk | Kokku kirjeid sõnastikus | EKI sõnastik | „Arvuti-sõnastik" | „Arvuti-kasutaja sõnastik" | IT termini-standardid | „E-teatmik" |
|---|---|---|---|---|---|---|
| Kokku kirjeid sõnastikus | | 136 013 | 14 275 | 5640 | 4068 | 7693 |
| I (esialgsel kujul) | 856 244 | 4168 | 1144 | 906 | 7693 | 583 |
| II (lemmatiseeritud) | 700 872 | 5003 | 1270 | 1002 | 459 | 698 |
| III (lisaks sõnaliigid) | 648 746 | 5009 | 1275 | 998 | 457 | 695 |
| IV (lisaks abisõnastik) | 646 226 | 5266 | 1286 | 1005 | 462 | 701 |
| V (lisaks ilma artikliteta) | 618 808 | 5268 | 1286 | 1008 | 460 | 701 |
| VI (lisaks ilma eessõnadeta) | 590 052 | 5270 | 1288 | 1010 | 464 | 708 |
| VII (lisaks järeltöötlus) | 22 125 | 4389 | 920 | 738 | 367 | 476 |

## Tulemus

Loodud tõlkeabisüsteemi veebipäringu aadress on *http://kde.teataja.ee*. Süsteemis saab päringuid teha sõnastikule, mis on genereeritud kõik eelnevalt kirjeldatud töötlusjärgud läbinud paralleelkorpuse põhjal ja mis on lisaks ka järeltöödeldud. Päringusõna sisestamisel peab kasutaja määrama sisestatava sõna lähtekeele, selle järgi otsitakse sõna kas eesti-inglise või inglise-eesti sõnastikust ja konkordantse vastavast paralleelkorpusest, seejuures esitatakse suurema tõenäosusega tegelikud tõlkepaarid väiksema tõenäosusega tõlkepaaridest eespool.

Sõna ja ta oletatav vaste on näitelausetes poolpaksus kirjas. Selline sõna esiletõstmine eristab meie tõlkeabisüsteemi teistest, mis reeglina sellist võimalust ei paku.

Sõnastiku administreerimisliidese kaudu on võimalik näha mitmesugust päringustatistikat, mis võimaldab muuhulgas otsinguaega optimeerida, lisaks saab konfigureerida mitmesuguseid seadeid, näiteks muuta päringuga esitatavate tõlkevastete miinimumtõenäosust ja valida sõnastikku, millele päringuid esitatakse. Sellist statistikat saab kasutada tõlketeoreetilistes uurimustes, tõlkesüsteemi parandamisel jne (vt Simard, Macklovitch 2005).

Tabel 7.

**Genereeritud sõnastike ning EKI inglise-eesti sõnastiku ja IT-sõnastike kattuvate kirjete protsent***

| Sõnastiku genereerimisel kasutatud korpuse töötlusjärk | Kokku kirjeid sõnastikus | EKI sõnastik | „Arvuti-sõnastik" | „Arvuti-kasutaja sõnastik" | IT termini-standardid | „E-teatmik" |
|---|---|---|---|---|---|---|
| Kokku kirjeid sõnastikus | | 136 013 | 14 275 | 5640 | 4068 | 7693 |
| I (esialgsel kujul) | 856 244 | 3,06/ 0,49 | 8,01/ 0,13 | 16,06/ 0,11 | 10,00/ 0,05 | 7,58/ 0,07 |
| II (lemmati-seeritud) | 700 872 | 3,68/ 0,71 | 8,90/ 0,18 | 17,77/ 0,14 | 11,28/ 0,07 | 9,07/ 0,10 |
| III (lisaks sõnaliigid) | 648 746 | 3,68/ 0,77 | 8,93/ 0,20 | 17,70/ 0,15 | 11,23/ 0,07 | 9,03/ 0,11 |
| IV (lisaks abisõnastik) | 646 226 | 3,87/ 0,81 | 9,01/ 0,20 | 17,82/ 0,16 | 11,36/ 0,07 | 9,11/ 0,11 |
| V (lisaks ilma artikliteta) | 618 808 | 3,87/ 0,85 | 9,01/ 0,21 | 17,87/ 0,16 | 11,31/ 0,07 | 9,11/ 0,11 |
| VI (lisaks ilma eessõnadeta) | 590 052 | 3,87/ 0,89 | 9,02/ 0,22 | 17,91/ 0,17 | 11,41/ 0,08 | 9,20/ 0,12 |
| VII (lisaks järeltöötlus) | 22 125 | 3,23/ 19,84 | 6,44/ 4,16 | 13,09/ 3,34 | 9,02/ 1,66 | 6,19/ 2,15 |

* Nt 3,23% EKI inglise-eesti sõnastiku kirjetest kattub viimase töötlusjärgu korpusest ekstraheeritud ja järeltöötluse läbinud sõnastikuga ja 19,84% korpusest ekstraheeritud ja järeltöötluse läbinud sõnastiku kirjetest langeb kokku EKI inglise-eesti sõnastikuga.

Päringu tegijal on võimalus päringu tulemust hinnata kas õigeks või valeks ja hiljem saab süsteemi kaudu info esitamisel kasutajate hinnanguid arvesse võtta.

## Kokkuvõte

Artiklis kirjeldatud töö eesmärgiks oli genereerida korpusepõhine internetis kasutatav tõlkeabisüsteem, mis ühendab tarkvarasüsteemi Uplug abil genereeritud sõnastiku selle allikaks olnud paralleelkorpusega.

KDE vabavaralise töölauakeskkonna dokumentatsiooni ja rakenduste tekstilisest osast koostatud paralleelkorpuse põhjal genereerisime kuus erinevat sõnastikuversiooni, igaühe neist paralleelkorpuse töötlemise erinevas staadiumis.

Pärast paralleelkorpuse inglise ja eesti keele omapära arvestavat automaatset eeltöötlust ja genereeritud sõnastiku spetsiifikat arvestavat automaatset järeltöötlust oli sõnastiku täpsus juhusliku 200-kirjelise fragmendi alusel hinnates eesti-inglise sõnastiku puhul *ca* 89 % ning inglise-eesti sõnastiku puhul *ca* 96 %. Madalamate tõenäosustega tõlkepaare sõnastikust kustutades oleks võimalik saavutada veelgi kõrgem täpsusaste, kuid maksimaalne täpsus ei olegi oluline, kui tegemist on tõlkeabisüsteemiga, mis võimaldab päringusõna vaadelda ta kasutamiskontekstis paralleelkonkordantsi näol.

Sõnastikuversioonide automaatseks võrdlemiseks kasutasime kõigi töötlusjärkude sõnastikuversioonide ühisosa tõenäosuste keskmist ja viit leksikograafide ja arvutiala asjatundjate käsitsi loodud inglise-eesti sõnastikku.

Genereeritud sõnastike kõrvutamisel ilmnes, et korpuse töötlemise iga etapiga saavutati vastava korpuseversiooni põhjal genereeritud sõnastiku kvaliteedi tõus. Seejuures olid tulemsõnastiku kvaliteedi parandamiseks kõige efektiivsemad sammud paralleelkorpuse lemmatiseerimine ja abisõnastiku kasutamine.

Peale sõnastiku oli internetis kasutatava tõlkeabisüsteemi loomiseks vaja viia paralleelkorpus täiendavalt sellisele kujule, mis võimaldas indekseerida kogu sõnastikus ja paralleelkorpuses sisalduva vajaliku info viietasemeliseks andmemassiiviks ning luua graafiline kasutajaliides, et saaks sõnastikule ja paralleelkorpusele interneti kaudu päringuid esitada. Selle töö tulemusena valminud tõlkeabisüsteemis esitatakse otsisõna ja tema oletatav vaste näitelausetes esiletõstetuna rasvases kirjas.

Loodud tõlkeabisüsteem sisaldab nii arvutitarkvaraga seotud kui ka üldkeeletermineid, mis teistest inglise-eesti-inglise sõnastikest puuduvad, kuid mis ometi on laialdaselt kasutusel vabavaraliste operatsioonisüsteemide ja programmide eestikeelsetes kasutajaliidestes, ning rohkelt näiteid, mis võimaldavad saada ülevaadet sõna erinevate tõlgete kontekstidest. Tõlkeabisüsteemi on võimalik täiendada analoogsete ja teiste keelevaldkondade paralleelkorpuste ja nende põhjal genereeritud sõnastikega. Praegusel kujul saavad süsteemist abi tõlkijad ja tarkvara eestindajad ning süsteemist võib kasu olla eestikeelse tarkvara terminoloogilise järjekindluse tagamisel.

Tehtud töö tulemusel saadud sõnastik võimaldab täiendada olemasolevaid arvutialaseid sõnastikke. Artiklis kirjeldatud töö erinevate etappide dokumentatsioon lihtsustab märkimisväärselt järgmiste eesti keelt ühe keelena hõlmavate sõnastike automaatset loomist.

## Kirjandus

A n d r e n u c c i, Andrea 2007. Creating a Bilingual Psychology Lexicon for Cross Lingual Question Answering. – A Pilot Study (*http://www.informatik.uni-trier.de/~ley/db/conf/iceis/iceis2007-2.html*). – ICEIS, nr 2, lk 129–136.

B r o w n, Peter, D e l l a P i e t r a, Stephen, D e l l a P i e t r a, Vincent, M e r c e r, Robert 1993. The Mathematics of Statistical Machine Translation Parameter Estimation [Electronic version]. – Computational Linguistics, nr 19, lk 263–311.

C e n d e j a s, Eduardo, B a r c e l ó, Grettel, G e l b u k h, Alexander, S i d o r o v, Grigori 2009. Incorporating Linguistic Information to Statistical Word-Level Alignment (*http://www.visionbib.com/bibliography/journal/cia.html*). – CIARP09, lk 387–394.

C h a r i t a k i s, Konstantinos 2007. Using Parallel Corpora to Create a Greek-English Dictionary with Uplug. – Processings of Nodalida 2007. The 16th Nordic Conference of Computational Linguistics, 25–26 May 2007 in Tartu, Estonia. Tartu, lk 212–215.

E r e l t, Tiiu, T a v a s t, Arvi 2003. Eesti oskuskeelekorralduse seisund. Tallinn: Eesti Keele Sihtasutus (*http://www.hm.ee/index.php?popup=download&id=3980*).

Hanson, Tavast 2005 = *http://keeleveeb.edu.ee/dict/speciality/aks/*

K a a l e p, Heiki-Jaan, V a i n o, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikooli Kirjastus, lk 87–99.

K h a n a r a k s o m b a t, Wanwisa, S j ö b e r g h, Jonas 2007. Developing and Evaluating a Searchable Swedish-Thai Lexicon. – Nodalida 2007 (*http://dr-hato.se/research/thailex.pdf*).

Liikane, Kesa 2006 = *http://keeleveeb.edu.ee/dict/speciality/aks/*

M e l a m e d, Dan 1995. Automatic Evaluation of Uniform Filter Cascades for Inducing N-best Translation Lexicons. – 3rd Workshop on Very Large Corpora (*www.aclweb.org/anthology/W/W95/W95-0115.pdf*).

M e g y e s i, Beata B., D a h l q v i s t, Bengt 2007. The Swedish-Turkish Parallel Corpus and Tools for its Creation. – Proceedings of NoDaLida 2007. May 24-26 2007, Tartu, Estonia (*http://dspace.utlib.ee/dspace/bitstream/10062/2564/1/reg-Megyesi-12.pdf*).

O c h, Franz Jozef, N e y, Hermann 2003. A Systematic Comparison of Various Statistical Alignment Models [Electronic version]. – Computational Linguistics 29 (*http://www.aclweb.org/anthology/J/J03/J03-1002.pdf*).

S i m a r d, Michel, M a c k l o v i t c h, Elliott 2005. Studying the Human Translation Process through the TransSearch Log-Files, in Proceedings of the AAAI Symposium on Knowledge Collection from Volunteer Contributors, Stanford, California, USA (*http://www-rali.iro.umontreal.ca/Publications/files/Simard-Macklovitch-KCVC05.pdf*).

S t r ö m b ä c k, Peter 2005. The Impact of Lemmatization in Word Alignment. Master Thesis, Department of Linguistics and Philology, Uppsala University (*stp.ling.uu.se/exarb/arch/2005_stromback.pdf*).

T i e d e m a n n, Jörg 1999. Word Alignment – Step by Step. – NODALIDA'99, the 12th Nordic Conference on Computational Linguistics (*http://portal.acm.org/citation.cfm?id=1220386*).

T i e d e m a n n, Jörg 2003a. 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. VIII 2006).

T i e d e m a n n, Jörg 2003b. Combining Clues for Word Alignment. – EACL'03, 10th Conference of the European Chapter of the ACL. ACL Press (*portal.acm.org/citation.cfm?id=1067852*).

V e l u p i l l a i, Sumithra, D a l i a n i s, Hercules 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic languages. – Proceedings of Workshop MMIES-2: Multi-Source, Multilingual Information Extraction and Summarization, held in conjunction with COLING-2008, Manchester, 23 August, 2008 (*http://www.aclweb.org/anthology-new/W/W08/W08-1403.pdf*).

Veskis, Kaarel 2007. Kakskeelsete leksikonide genereerimine paralleelkorpuse baasil. – Eesti Rakenduslingvistika Ühingu aastaraamat 3. Tallinn, lk 355–372.
Xing, Hao-chun, Zhang, Xin 2008. Using Parallel Corpora and Uplug to Create a Chinese-English Dictionary. Master Thesis, Department of Computer and Systems Sciences, KTH/Stockholm University (*http://api.ning.com/files/ZMoTcaMWUsmpzU*Ec*jfFz97NdZGGaKLzOp7Mp35dA1zB14h1EEpKH7ZRM0gD6gu4p-LjQsASM4-jFSALV6jjdMjwXLYtwTz/Using_parallel_corpora_and_Uplug_to_create_a_ChineseEnglish_dictionary_by_XingZhang.pdf%22*).

**An Online Parallel Concordancer Usable as a Translation Tool**

Keywords: corpus linguistics, specialized lexicography, automatic lexicon generation, word alignment, translation software

This article describes the creation process of an online English-Estonian parallel concordancer that includes a bilingual English-Estonian lexicon. The lexicon has been automatically extracted from a parallel corpus containing KDE software documentation and textual parts of KDE applications. Using the Uplug word alignment tool, we created six different versions of the lexicon that correspond to different preprocessing stages of the source corpus. The precision of the final versions of the lexicons was 96% for the English-Estonian lexicon and 89% for the Estonian-English lexicon. We compared these different versions of the lexicon, using the common entries of all our lexicons and other English-Estonian lexicons as gold standard. The result of our work is an online translation tool that will, as we hope, make a valuable resource for translators and software localizers.

*Katrin Tsepelina* (*b. 1982*), *University of Tartu, Department of General Linguistics, Institute of Estonian and General Linguistics, specialist, katrin.tsepelina@ut.ee*

*Kaarel Veskis* (*b. 1976*), *MA, University of Tartu, Department of General Linguistics, Institute of Estonian and General Linguistics, specialist, kaarel.veskis@ut.ee*