Masintõlge – kas emakeele päästerõngas?

Toomas Koitmäe, Merli Mändul
Viljandi gümnaasiumi abituriendid

Masintõlge on meile, infoühiskonna noortele, olnud kättesaadav aastast 2005, mil suurkorporatsioon Google teatas esimese, massidele tarbimiseks sobiva tarkvara – Google Translate'i loomisest. Google Translate kujutab endast vabavaralist tööriista, mis lubab tõlkida lauseid, dokumente ja isegi terveid veebilehekülgi kõigest mõne hetkega. Väga paljud meist on harjunud vajaduse korral tõlkeabi kasutama, kuid kui palju me teame sellest, kuidas üks tänapäevane masintõlge tegelikult töötab?

Tänapäeva kooliteed alustavate õpilaste seas võib kuulda arvamust, et tõlkelehekülg on nagu tuba, mis on täis pisikesi kakskeelseid päkapikke, kes töötavad meie heaks. Tegelikult on kogu süsteemi taga palju reaalsem, kuid samas keerukam maailm. Protsessi, mille käigus kõiki meie lauseid võõrkeelde pannakse, nimetatakse statistiliseks masintõlkeks. Statistiline masintõlge tähendab seda, et arvuti võrdleb meie antud teksti suure hulga olemasolevate tekstidega, leiab parima n-ö mustri ja edastab kasutajale suurima kattuvusega teksti.

Põhimõtteliselt võib arvuti uut keelt õppida samuti nagu meie – õpid ära sõnavara ja grammatikareeglid ja keel ongi selge. Paraku, nagu ka inimese jaoks, on keele õppimine raske eelkõige erandite tõttu. Ja nagu eranditest veel vähe oleks – paljudel eranditel on omakorda erandid. Kogu keelt arvutisse salvestada on sisuliselt võimatu. Seega on Google asjale lähenenud hoopis teist teed pidi – arvuti avastab need seaduspärasused ise.

Google Translate töötab läbi miljoneid ja miljoneid inimese tõlgitud tekste, näiteks raamatuid ja eri organisatsioonide (teiste hulgas ÜRO) veebilehti. Programm skaneerib sisestatud laused ja neid hakatakse võrdlema kõikide teiste tekstidega, et leida sarnasusi. Kui ühel korral on sarnasus leitud, oskab programm sedasama mudelit või mustrit kasutada ka järgmisel korral. Kui seda protsessi korratakse miljardeid ja miljardeid kordi, on lõpuks koos suur hulk mustreid, mis tähendab, et tõlkeprogramm areneb ja on saanud targemaks.

Kogu tõlke kvaliteedi määrab tegelikult tekstide hulk, sest tõenäosus, et soovitud tõlkega sarnaseid lausemalle võib leida mõnest tõlkeprogrammile aluseks olevast tekstist, kasvab iga uue tekstiga. Seega on seletatav, miks paljudes keeltes tõlkeprogrammi kvaliteet veel paljuski soovida jätab. Tõlkekvaliteedi jaoks on kindlasti olulised ka serveri maht ja kiirus, sest kui kasutaja sisestab oma teksti, peab programm suutma silmapilkselt miljardid teised tekstid läbi töötada ja kõige tõenäolisemad mustrid välja valida.

Praegu tuleb siiski tunnistada, et eesti keelest võõrkeelde tõlkimisel või vastupidi peame arvestama olulisi vajakajäämisi. Selgitamaks välja, milline on praegu statistilise masintõlke tase, koostasid siinse artikli autorid eelmisel õppeaastal kaks uurimust. Koitmäe hindas Google Translate'i tõlketaseme erinevust, kui tõlkida eestikeelseid tekste saksa või inglise keelde. Merli Mänduli töö uuris aga Google Translate'i ja Bing Translate'i, praegu kahe suurima võimsusega tõlkeprogrammi tööd.

Analüüsides tõlget inglise ja saksa keelest ning võrreldes Google'it ning Bing Translate'i, selgus, et masintõlge teeb üpris palju vigu, mis üldiselt on märgatavad ning kohati isegi veidrad või lausa naljakad. Seetõttu on need tõlkevead tunduvalt ohutumad kui inimeste tehtavad, sest need on lihtsalt äratuntavad. Seda tõestasid ka 16–30aastaste inimeste seas tehtud küsitluse tulemused. Vastajaile anti lugeda neli teksti: kaks inglise ja saksa keelest masina ning kaks inimese tõlgitud teksti. Ülesandeks oli arvata ära, milliste tekstide puhul on tegemist masintõlkega ja millised on inimese tõlgitud. Tulemuseks oli, et kõik 50 inimest vastasid õigesti. See küsimus andis kindlust – inim- ja masintõlge erinevad veel tublisti.

Võrdluses inglise ja saksa keelest tõlkimisel selgus veidi üllatavana, et Google Translate jätab tõlkimata ingliskeelsest tekstist rohkem sõnu kui saksakeelsest. Ometi võiks eeldada, et ingliskeelseid alustekste on rohkem, mistõttu peaks masintõlkel olema rohkem võimalusi eri alustekste võrrelda. Ingliskeelsest tekstist jäi tõlkimata 20 sõna ehk umbes 2,5% kogu sõnade arvust. Saksa keelest eesti keelde tõlkimisel jätab Google Translate tõlkimata või pakub vasteks ingliskeelse sõna 17 korral, mis moodustab ligikaudu 1,4% kogu tekstist.

Töö koostamise käigus hakkas aga silma veel üks huvitav asjaolu, mida algselt ei olnudki plaanis uurida. Nimelt pakkus Google Translate nii inglise kui ka saksa keelest tõlkides ootamatult vasteks tihti soomekeelseid sõnu. Põhjus võib olla eesti ja soome keele suguluses, mistõttu

masinal ilmselt eestikeelsete tekstide hulgast vaste otsimine ei õnnestu ning ta otsib seejärel mõnest teisest keelest võimalikult lähedase vaste. Inglise keelest tõlkides pakkus masintõlge kaheksa soomekeelset vastet ning saksa keelest tõlkides kolm.

Esines ka sõnale täiesti väära tähenduse andmist. Selliseid saksakeelseid sõnu leidus kümnel korral ning ingliskeelseid viiel. Ohtralt esines sõnajärjevigu, ent see on loomulik, sest eesti, inglise ja saksa keeles on lauseliikmete järjekord lauses erinev ning kuna tõlkeprogramm vaatab enamasti iga sõna eraldi, arvestamata konteksti, on vead kerged tekkima. Kõige sagedamini esile kerkinud probleem oli siiski käändsõnafraasi sobimatu ülesehitus, mis tuleneb ilmselt sellest, et tõlkeprogrammil on suuri raskusi sõnaliikide äratundmisega. Kuna eesti keeles on 14 käänet, saksa keeles neli ning inglise keeles ainult kaks, oli Google Translate tihti hädas võrreldavatest alustekstidest õige vaste leidmisega. Siinkohal ei saa öelda, et tõlge ühest keelest oleks parem kui teisest, sest masintõlge eksis mõlema puhul mitu korda. Näiteks oli masintõlkele probleemiks käändsõnafraasi osiste ühildumine arvus ning sõnade omavaheline kontekstiline seostumine, mis toob taas kord välja masintõlke miinuse – see tõlgib iga sõna eraldi, mitte teksti kui tervikut.

Viimasest tõigast tulenevalt tekkis ehk veidi ettearvatavam probleem – sünonüümide kasutamine. Mõlemast keelest tõlkides tegi Google Translate umbes sama palju vigu, andes sõnale vasteks tähenduselt küll õige, kuid konteksti sobimatu vaste, mis muutis saadud laused arusaamatuks ning tihti isegi naljakaks. Näiteks ingliskeelne sõna *poke* peaks tähendama pistmist või lööki/kolakat ja kui sobivaks tõlkeks oleks *Tahad kolakat saada?*, andis Google Translate vasteks hoopis *Ma annan sulle selline pistma!*.

Kokkuvõtvalt ei saa öelda, kas Google tõlgib paremini inglise või saksa keelest, sest vigade arv oli mõlemas tekstis ligilähedane ning olulisi erinevusi ei esinenud. Võrreldes aga Bing Translate'i ja Google Translate'i, kerkisid vead rohkem esile kui saksa ja inglise keele võrdlusel. Näiteks oli tõlkimata jäetud sõnu Google'i tõlkes 24 ning Bingi tõlkes 29. Tõlkimata jäetud sõnu üksikult sisestades tõlkis Google neist ära 87,5% ning Bing kõigest 31%. See näitab, et Google Translate on võimekam kui Bing Translate. Sõnale andis vale tähenduse Google Translate 32 korral, Bing Translate aga 41 korral. Üks huvitavamaid tõlkevigu oli Bing Translate'i tõlkes see, et nimele Heido pakuti vasteks Andrus. Ka verbi pöördeliste vormide ning nimisõnade käändevormide puhul edestas Google Translate Bingi mitmekordselt. Lisaks torkas

silma kolmekordne erinevus suure ja väikese algustähe kasutamisel. Nimelt eksis Google Translate siin ainult viiel korral, Bing Translate'i eksimuste arv oli lausa viisteist. Bingi eksimuste arvu suurendas tunduvalt lause keskel esineva nimisõna arusaamatutel põhjustel suure algustähega kirjutamine. Ka stiililiselt andis Google Translate tihti palju paremaid vasteid kui Bing Translate, kuigi mõlemal tõlkeprogrammil on suuri raskusi värvikate väljendite tõlkimisega. Näiteks on Raul Rebase kasutatud väljend *kulda pasandav eesel* Google Translate'i tõlgituna *shitting kuld on eesli* ning Bing Translate annab vasteks *eesli ja situb kulla*. Niisiis võib öelda, et kui tuleks valida Google'i ning Bing Translate'i vahel, oleks ilmselt parema tulemuse saamiseks arukas kasutada Google Translate'i. Selle tõlkeprogrammi paremuse on taganud kindlasti suurem alustekstide maht, pikem kasutamisaeg, millest tulenevalt on antud rohkem tagasisidet ning on olnud võimalusi programmi parandada.

Praeguse infoühiskonna liikmed tahavad saada infot kõikvõimalikes keeltes. Viimasel ajal on märgata suundumust, et eesti noor ei võta emakeeleõpet enam tõsiselt, sest peamised infovood liiguvad maailmas teistes keeltes ja globaalset konteksti arvestades võib tunduda mõistlikum õppida inglise, saksa, hiina, hispaania või portugali keelt. Siin võikski statistilised masintõlkeprogrammid eesti keelt tublisti aidata, sest tõlkeprogrammide arenguga kasvab ka nende kvaliteet. Kui ühel hetkel on saavutatud kõrgeim võimalik kvaliteet, ei ole meil enam vaja eesti kirjakeelele eelistada võõrkeeli, sest kogu maailma info jõuab meieni meie emakeeles. Selleks aga, et tõlkeprogrammid areneksid, peaksid ka eestlased oma panuse andma, avaldades internetis rohkem eestikeelset teksti, mis on n-ö käsitsi tõlgitud ka teistesse keeltesse. See on mõlemapoolne võit – programmi kvaliteet paraneb ja eesti keel elab edasi tugevamana kui kunagi varem.