Ingibjörg Elsa Björnsdóttir

Vélþýðingar á íslensku og Apertium-þýðingarkerfið

1 Inngangur

Markmið máltækni er að beita tækni á náttúruleg tungumál þannig að niðurstaðan geti aukið þekkingu notandans, hvort sem það felur í sér að skilja erlent tungumál betur, lagfæra ritað mál eða nýta tungumálið á annan hátt (Martha Dís Brandt 2011:3). Vélþýðingar falla samkvæmt þessari skilgreiningu undir máltækni og þær eru í hraðri þróun í dag víða um heim (Sin-wai 2002:vii–viii).

Í þessari grein verður vélþýðingarkerfið Apertium skoðað. Það er opinn hugbúnaður sem hefur verið þróaður fyrir þýðingar af íslensku yfir á ensku (Martha Dís Brandt 2011:3). Kerfið getur einnig þýtt úr íslensku yfir á sænsku en þetta tungumálapar virðist minna þróað en þýðingarnar úr íslensku yfir á ensku. Hér verður aðallega fjallað um þýðingar úr íslensku yfir á ensku. Þýðingar úr ensku/sænsku yfir á íslensku eru ekki enn fyrir hendi.

Helsta niðurstaða úttektarinnar er að Apertium-vélþýðingarkerfið sé mjög mikilvæg atrenna að vélþýðingum fyrir íslensku, hvaða kerfi sem kunna að verða notuð í framtíðinni. Nú er ómögulegt að segja til um að hve miklu leyti tölvur geti skilið og numið t.a.m. margræðni og fyndni þegar fram líða stundir en ljóst er að bæta má nákvæmni Apertium-kerfisins, t.d. með því að stækka orðabækur og jafnvel bæta við dæmum sem kerfið gæti stuðst við. Þróun orðasafna og stórra málheilda skiptir hér lykilmáli ef takast á að auka orðaforða Apertium-kerfisins og annarra vélþýðingarkerfa fyrir íslensku. Rannsóknir

Orð og tunga 18 (2016), 131–143.

með Apertium-kerfinu geta einnig nýst í málvísindum og almennt í lýsandi þýðingafræði.

Í 2. kafla verður sagt lítillega frá uppruna og sögu vélþýðinga. Í 3. kafla eru útskýrðar helstu tegundir vélþýðinga og einkenni þeirra. Í 4. kafla er rakin saga íslenska Apertium-vélþýðingarkerfisins og í 5. kafla er fjallað ítarlegar um það. Lokaorð úttektarinnar koma fram í 6. kafla.

2 Nokkur orð um uppruna vélþýðinga

Uppruna vélþýðinga má rekja aftur til fyrri hluta 20. aldar. Frakkinn Georges Artsrouni fékk úthlutað einkaleyfi 1933 fyrir tæki sem hann kallaði vélrænan heila og gat þýtt milli tungumála með því að nota fjóra meginþætti: minni, lyklaborð til innskráningar, leitaraðferð og úttak. Leitaraðferðin var reyndar bara leit í orðabók í minninu en þetta var þó fyrsta tilraunin til einhvers sem var í ætt við vélþýðingar. Rússinn Pjotr Petrovitsj Trojanskíj fékk einkaleyfi sama ár fyrir kerfi sem notaði tvímála orðabók og úrvinnsluferli í þremur skrefum (Martha Dís Brandt 2011:7–8).

Vélþýðingar tóku kipp á nýjan leik þegar Warren Weaver, deildarforseti náttúruvísinda hjá Rockefeller Foundation í Bandaríkjunum, sendi minnisblað til hóps framúrskarandi vísindamanna árið 1948 (Trujillo 1999:4–5). Weaver hafði þegar árið 1947 hafið viðræður við vísindamenn í Bandaríkjunum og Bretlandi um hvort ekki væri nauðsynlegt að kanna hvort hin nýja tölvutækni, sem var að koma fram á sjónarsviðið, gæti nýst við þýðingar og þá einkum við að þýða úr rússnesku yfir á ensku (s.st.). Þetta var á dögum kalda stríðsins, áhugi á þróun vélþýðinga var talsverður og málefnið almennt talið brýnt. Fyrsta gerðin af vélþýðingarkerfi, sem þýddi úr rússnesku yfir á ensku, var opinberlega kynnt vísindaheiminum árið 1954. Eftir að þeim áfanga hafði verið náð hófst þróun tækni við vélþýðingar víða um heim (s.st.).

Mikil bjartsýni ríkti við upphaf vélþýðinga en því skeiði lauk árið 1966 þegar svokölluð ALPAC-skýrsla var kynnt í Bandaríkjunum. ALPAC-nefndin var nefnd sjö vísindamanna undir forystu Johns R. Pierce frá Bell Telephone Laboratories. Nefndin var skipuð árið 1964 af Bandaríkjastjórn til að meta þann árangur sem náðst hefði á sviði máltækni og var sjónum sérstaklega beint að vélþýðingum. Niðurstöður skýrslunnar voru þær að vélþýðingar væru of dýrar og

myndu í raun aldrei borga sig og í kjölfarið var opinbert fé til rannsókna á vélþýðingum í Bandaríkjunum verulega skorið niður (Hutchins 2003:131–135).

Þegar ALPAC-skýrslan var rituð hafði tölvutæknin hins vegar ekki þróast nægilega mikið til þess að tölvur gætu ráðið við vélþýðingar. Þeir sem sátu í ALPAC-nefndinni sáu ekki fyrir að einkatölvur kæmu brátt á markaðinn, þeir sáu ekki fyrir tilkomu Netsins og veraldarvefsins og allar þær milljónir megabæta sem síðar yrðu notaðar við vinnslu á hverri sekúndu.

Eftir því sem tölvutækninni fleygði fram varð áhugaverðara að þróa vélþýðingar. Tilraunir með þær hófust hins vegar ekki af fullri alvöru fyrr en eftir 1970. Árið 1978 hófst EUROTRA-vélþýðingarverkefnið innan Efnahagsbandalags Evrópu sem síðar varð ESB. Þótt markmið verkefnisins næðust ekki að öllu leyti hefur það haft varanleg áhrif á þróun vélþýðinga. Vélþýðingar byggðar á upplýsingum og gervigreind komu fram árið 1983 í Bandaríkjunum (Trujillo 1999:5). Frá 1980 hefur þróun í gervigreind og vélþýðingum verið mjög hröð (s.st.). Vélstuddar þýðingar verða æ algengari innan tæknilega þýðingageirans. Til dæmis hefur SDL, sem er þýðingarfyrirtæki í eigu Microsoft, þróað vélþýðingarkerfi sem kallast BeGlobal og er hægt að nota innan TRADOS Studio 2015-þýðingaforritsins. Þetta kerfi, sem er byggt á tölfræðilegum vélþýðingum, þýðir texta afar hratt á milli útbreiddustu tungumála heimsins, svo sem ensku, frönsku, þýsku, spænsku og arabísku. Árangur BeGlobal-kerfisins kemur á óvart varðandi orðaforða þótt vissulega þurfi að prófarkalesa textann og laga villur (*sdl.com*).

Í marga áratugi hafa vélþýðingar verið í stöðugri notkun innan ESB. SYSTRAN-vélþýðingarkerfið þýðir nú 60–70% texta í höfuðstöðvum Þýðingamiðstöðvar ESB í Lúxemborg. SYSTRAN hefur einnig verið notað hjá bandaríska hernum þar sem það þýðir m.a. vísindagreinar af rússnesku yfir á ensku.

Google translate-þýðingarkerfið hefur tekið miklum framförum á undanförnum árum. Það er tölfræðilegt þýðingarkerfi sem lærir smám saman af mistökum sínum og batnar stöðugt við meiri og meiri notkun.

Vélþýðingar verða æ brýnna viðfangsefni eftir því sem gervigreind fleygir fram og snjallsímar og vefsíður verða mikilvægari tæki til að afla upplýsinga. Nútímamaðurinn vill hafa aðgang að upplýsingum á Netinu á öllum tímum sólarhrings og helst lesa allt á móðurmáli sínu. Því er ljóst að vélþýðingar þurfa að koma til skjalanna að einhverju

leyti eigi að vera hægt að mæta kröfum nútímans um upplýsingar og umsvifalausar þýðingar (e. *instant translation*) allan sólarhringinn.

Tölfræðilegar vélþýðingar hafa einkum náð flugi eftir því sem gríðarlegum tvímála málheildum hefur verið komið á laggirnar og að sjálfsögðu er hægt að nota Netið sjálft sem upplýsingagrunn fyrir vélþýðingar. Vélþýðingar, sem byggjast á dæmum, hafa einnig batnað og ýmis blönduð kerfi eru nú í athugun enda eru vélþýðingar orðnar frjór vettvangur fyrir rannsóknir á sviði þýðinga og málfræði og hvað varðar möguleika véla til að læra og öðlast gervigreind.

3 Mismunandi tegundir vélþýðinga

3.1 Hvað eru vélþýðingar?

Vélþýðing er það að tölva er notuð til að þýða tölvutækan texta af einu máli á annað. Tilgangi vélþýðinga er oft skipt í tvennt: annars vegar að fá fram grófþýðingu sem getur verið slök vélþýðing en skiljanleg t.d. sérfræðingi; hins vegar að fá fram birtingarhæfan texta.

Til eru þrjár megintegundir vélþýðinga (sjá Sin-wai 2002):

> Tölfræðilegar vélþýðingar (e. *statistical machine translation, SMT*)
>
> Vélþýðingar með regluaðferð (e. *rule-based machine translation, RBMT*)
>
> Vélþýðingar sem byggjast á dæmum (e. *example-based machine translation, EBMT*)

Að auki eru til ýmis blönduð kerfi og mikil gróska er í þróun margra blandaðra kerfa.

3.2 Tölfræðilegar vélþýðingar

Tölfræðileg vélþýðingarkerfi læra að þýða með því að greina gríðarlega stórar tvímála málheildir. Setningunum er fyrst skipt niður í setningarhluta eða orð og síðan er leitað tölfræðilega að sambærilegum setningum sem hafa þegar verið þýddar. Þannig kemur í ljós hvaða þýðing setningarinnar er líklegust. Enn fremur getur tölvan greint hið þýdda efni og metið hvaða orð eru algengust og því líklegust sem

þýðing. Til þess að tölfræðileg þýðingarkerfi geti náð árangri verða tvímála málheildir að geyma hundruð milljóna orða. Sem dæmi má nefna að The Cambridge International Corpus-málheildin telur núna um það bil einn milljarð orða (O'Keefe og McCarthy (ritstj.) 2010:5). Þess vegna henta tölfræðileg þýðingarkerfi oft ekki fyrir önnur málsamfélög en þau sem hafa aðgang að stórum málheildum.

Fyrstu tölfræðilegu vélþýðingarnar voru kynntar til sögunnar af hópi vísindamanna hjá IBM á árunum 1980–1990 (Goutte, Cancedda, Dymetman og Foster (ritstj.) 2009:2). Eftir það hafa tölfræðilegar vélþýðingar tekið miklum framförum, einnig í blönduðum kerfum.

3.3 Vélþýðingar með regluaðferð

Í vélþýðingum með regluaðferð eru setningafræðireglur, orðmyndunarreglur og beygingarreglur upprunamálsins skráðar inn í þau tölvuforrit eða einingar sem annast vélþýðinguna.

Vélþýðingarkerfið notar málfræðireglurnar ásamt tvímála orðabókum, tveimur eða fleiri. Orðabækur í vélþýðingum líkjast hefðbundnum orðabókum. Þær innihalda uppflettimyndir orða og stundum einnig stakar setningar eða setningarhluta úr upprunamálinu. Orðabækurnar geyma auk þess þýðingar orðanna yfir á markmálið. Öll orð í orðabókunum eru flokkuð málfræðilega.

Vélþýðingarhugbúnaðurinn greinir hverja einstaka setningu sem skal þýða og setur inn merki sem tengist sérhverju orði til að auðkenna setningafræðilega eiginleika þess (t.a.m. orðflokk og setningarhluta). Oft eru einnig skráðar upplýsingar um beygingu. Vélþýðingarkerfið leitar síðan að þýðingu þessara vélgreindu og merktu orða í þeim orðabókum sem tölvan hefur aðgang að. Að lokum raðar tölvan orðunum aftur saman í setningu og notast þar við setningafræðireglur markmálsins til þess að setningarnar verði sem réttastar, málfræðilega. Vönduð orðabók eða tvímála málheild þarf að vera tiltæk fyrir hvert tungumálapar sem þýða skal.

Kostir þýðingarkerfa með regluaðferð eru þeir helstir að þau halda sig nær frumtextanum við þýðingar. Tölfræðileg þýðingarkerfi hafa aftur á móti tilhneigingu til að leggja of mikla áherslu á að þýðingin verði þjál (Forcada o.fl. 2011:128). Það er einnig auðveldara að ritstýra og prófarkalesa texta úr þýðingarkerfum með regluaðferð en úr hinum tölfræðilegu þar sem þeir eru aðeins nær frumtextanum og kerfið hefur ekki reynt að gera textann þjálli. Að lokum má nefna að sérfræðingar eiga auðveldara með að finna og leiðrétta villur í þýð-

ingarkerfum með regluaðferð (Forcada o.fl. 2011:129). Einnig þurfa þýðingarkerfi með regluaðferð almennt smærri málheildir en tölfræðileg þýðingarkerfi. Tölfræðileg þýðingarkerfi þurfa stórar málheildir eins og t.d. The British National Corpus sem er um 100 milljónir orða. Þetta getur í sumum tilvikum orðið til þess að þýðingarkerfi með regluaðferð henta betur fyrir tungumál lítilla málsvæða.

3.4 Vélþýðingar sem byggjast á dæmum

Það er engin tilviljun að vélþýðingar byggðar á dæmum komu fram á sjónarsviðið um 1980 eða á sama tíma og fyrstu þýðingarminnin voru að líta dagsins ljós (Somers 1999:114–115). Þýðingarminni er gagnvirkt tæki fyrir mann sem þýðir. Dæmavélþýðingar eru aftur á móti í eðli sínu sjálfvirk þýðingartækni eða sjálfvirk aðferðafræði (Somers 1999:115).

Meginþættir dæmavélþýðinga eru:

- að bera setningar eða setningarhluta saman við gagnagrunn með raunverulegum dæmum
- að auðkenna setningar eða setningarhluta sem eiga saman
- að setja aftur saman (sjá hér fyrir neðan) setningar eða setningarhluta til að skila þýddum texta á markmálinu

Dæmavélþýðingar byggjast á stórum málheildum. Tvímála hliðruð (e. *parallel*) málheild er það sem til þarf. Þegar fundin hefur verið málheild sem hentar til verksins þarf að bera kennsl á þær setningar eða setningarhluta sem eiga í raun saman (Somers 1999:118–119). Reynslan hefur sýnt að skipta þarf setningum að einhverju leyti niður í hluta til að kerfið verði nógu nákvæmt. Hins vegar má ekki greina setningarnar svo mikið niður að samhengi týnist. Hér er því um að ræða ákveðið jafnvægi sem þarf að reyna að finna og skilgreina í hvert skipti sem dæmavélþýðingarkerfi er sett upp. Að lokum vaknar sú spurning hversu stór gagnagrunnurinn þurfi að vera, þ.e. hve mörg dæmi þurfi að vera aðgengileg í gagnagrunninum. Dæmavélþýðingarkerfi verða almennt betri eftir því sem dæmin eru fleiri en það á aðeins við upp að ákveðnum þröskuldi. Þegar honum hefur verið náð virðist kerfið ekki batna meira þótt fleiri dæmum sé bætt við enda eru þá orðnir svo margir þýðingarkostir í boði að það þarf í raun að fara aftur yfir textann (Somers 1999:119). Hins vegar getur verið mjög skilvirkt að

bæta dæmavélþýðingarkerfi með mjög stóran gagnagrunn við aðra tegund af þýðingarkerfi, svo sem tölfræðilegt vélþýðingarkerfi eða vélþýðingarkerfi með regluaðferð.

4 Saga íslenska Apertium-kerfisins

Martha Dís Brandt, meistaranemi við Háskólann í Reykjavík, tók þátt í að þróa frumgerð af Apertium-kerfinu sem hafði það að markmiði að þýða á milli íslensku og ensku. Notuð voru máltæknitól sem þegar voru til í IceNLP-safni máltæknitóla sem dr. Hrafn Loftsson og Hlynur Sigurþórsson höfðu þróað. Laga þurfti IceNLP-máltæknitólin að Apertium-kerfinu og voru IceNLP-tólin gerð að opnum hugbúnaði til þess að hægt væri að setja þau sem einingar inn í Apertium-flæðið. Dr. Francis Tyers setti upp umgjörð fyrir þrjár fyrstu íslensku orðabækurnar og tók þátt í mótun og gerð flutningsreglna fyrir íslensku. Eiríkur Rögnvaldsson prófessor hefur einnig tekið þátt í verkefninu af hálfu Háskóla Íslands. Martha Dís Brandt leiðrétti síðan meira en 5000 færslur í íslensku tvímála Apertium-orðabókinni og bætti um 19.400 færslum við tvímála orðabókina. Einnig bætti hún við flutningsreglum og lagfærði kerfið að öðru leyti (Martha Dís Brandt 2011:1–2). Auk þess var búið til textasafn úr um 188.000 línum úr íslensku Wikipediu. Síðan voru gæði Apertium metin og reyndist villutíðni (e. *word error rate, WER*) 50,60% og villutíðni óháð stöðu (e. *position-independent word error rate, PER*) 40,78% (Martha Dís Brandt 2011:2). Þetta er nokkuð hærri tíðni en hjá Google translate eða Tungutorgi, vélþýðingarkerfi sem eðlisfræðingurinn Stefán Briem þróaði (*tungutorg.is*).

5 Apertium-vélþýðingarkerfið

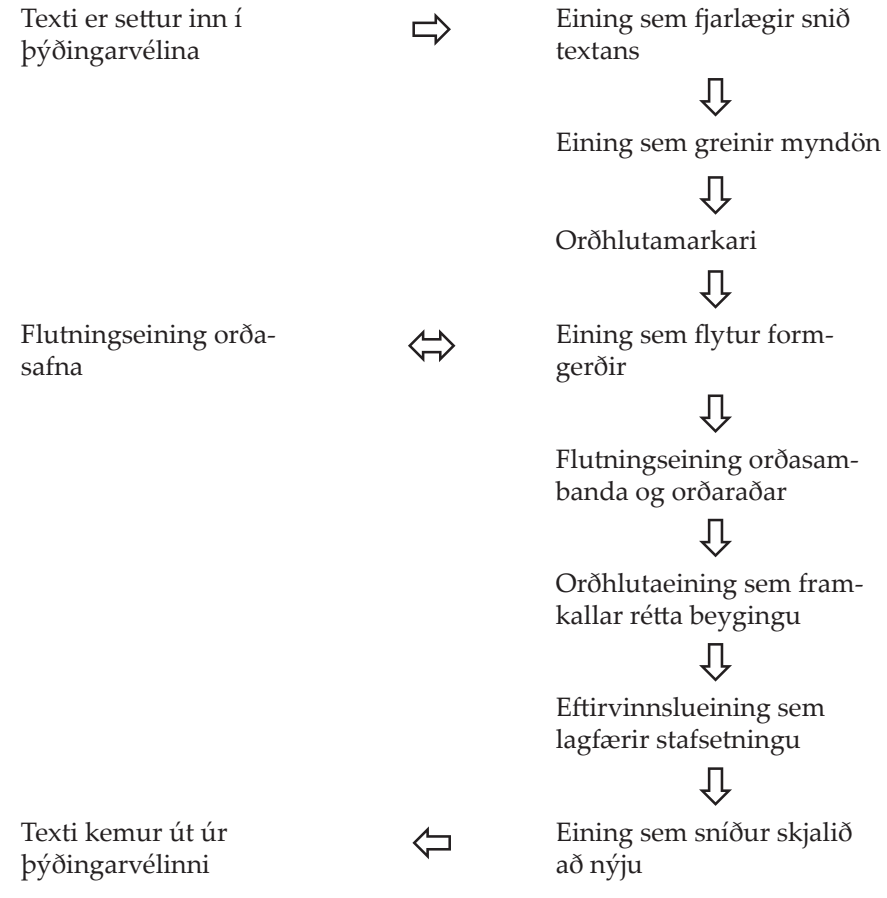
5.1 Almennt um Apertium-kerfið og tæknina sem það notar

Apertium er opið grunnstætt (e. *shallow transfer*) vélþýðingarkerfi sem hefur reynst mjög gagnlegt fyrir smærri málsamfélög. Þýðingarkerfið var upphaflega þróað á Spáni úr tveimur kerfum, interNOSTRUM og traductor.universia.net, sem voru þróuð í Háskólanum í Alicante

(Forcada o.fl. 2011:130). Apertium var upphaflega hannað til þýðingar á milli skyldra tungumála, svo sem spænsku og katalónsku, en ræður nú við tungumál sem eru mun fjölbreyttari að gerð og eru ólík orðhlutafræðilega (*apertium.org*).

Apertium-kerfið notar ferjöld (e. *finite-state transducers*), svokölluð falin Markovslíkön, til að marka orðhluta og framkvæma böggun (e. *chunking*) til að flytja setningarhluta (sjá Forcada o.fl. 2010:6–10; sbr. einnig Sigrúnu Helgadóttur 2007).

Þýðingarvélin er sett saman úr átta eininga samtengdum hlutum (færibandi) sem notast við textaflæði (e. *text streams*) til að auðvelda greiningu og óháðar prófanir, sjá *Mynd 1*.



Mynd 1. Einingarnar átta mynda færibandið sem er grunnurinn að Apertium-vélþýðingarkerfinu.

Einingar Apertium-kerfisins eru eftirfarandi (sjá Forcada o.fl. 2010:6–10):

- Eining sem fjarlægir snið textans og aðgreinir sjálfan textann frá upplýsingum um snið (HTML o.s.frv.).
- Eining sem greinir myndön á yfirborði (e. *surface form*) textans og skilar af sér fyrir hverja einingu einni eða fleiri orðasafnsmyndum sem samanstanda af flettu, orðflokki og upplýsingum um beygingu.
- Orðhlutamarkari (e. *part-of-speech tagger*) framkvæmir eina af greiningunum á tvíræðum orðum eftir því í hvaða samhengi orðið stendur. Hann notar svokallað falið Markovslíkan.
- Flutningseining orðasafna les orðasafnsmyndir frummálsins og skilar viðeigandi orðasafnsmynd í markmálinu. Við þetta notar einingin tvímála orðabók. Sú eining, sem flytur formgerðir, sækir orðin í orðabókina.
- Flutningseining sem vinnur með orðasambönd og orðaröð.
- Orðhlutaeining sem framkallar rétta beygingu í markmáli á yfirborði, úr uppflettimynd.
- Eftirvinnslueining sem lagfærir stafsetningu í markmálinu.
- Eining sem gefur skjalinu aftur upprunalegt snið.

Apertium-kerfið þarf að hafa tiltæka einmála orðabók upprunatungumálsins fyrir sérhvert tungumálapar. Þessi einmála orðabók er notuð af einingunni sem greinir myndön. Tvímála orðabók, sem nær til upprunamálsins og markmálsins, er notuð af flutningseiningu orðasafna og að lokum er einmála orðabók markmálsins notuð af einingunni sem setur saman setningarnar. Í kerfinu notar flutningseining flutningsreglur (sjá Mörthu Dísi Brandt, Hrafn Loftsson, Hlyn Sigurþórsson og Tyers 2011:2).

Kostir Apertium-kerfisins eru fjölmargir enda er kerfið einfalt bæði í notkun og þróun. Hönnun þess byggist að mestu á einföldum og þekktum forsendum Unix-kerfa. Apertium nær þýðingarhraða sem nemur um 10.000 orðum á sekúndu á venjulegum fartölvum. Tæknin þarf því ekki stór gagnaver (Forcada o.fl. 2009:4). Af einingum Apertium-kerfisins er einingin, sem greinir myndön, e.t.v. mikilvægust af því að hún getur ekki einungis nýst til vélþýðinga heldur einnig í WordNet-viðmóti þar sem hægt er t.d. að búast við

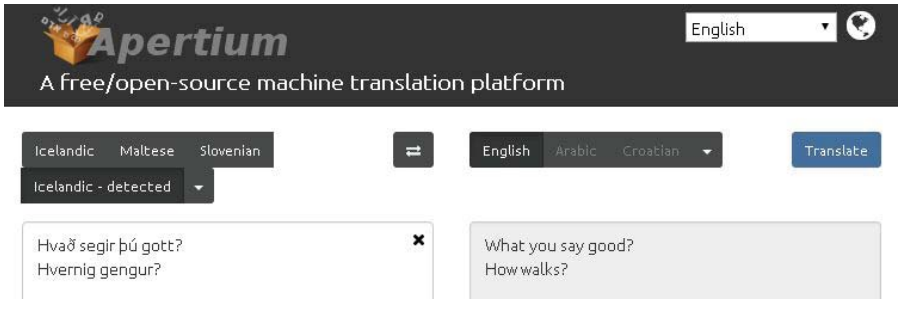
að í framtíðinni verði unnt að gera setningafræðilegan samanburð á ólíkum tungumálum í rannsóknarskyni (Pala, Bosch og Fellbaum 2008:14-16).

Framsetning málvísindalegra gagna í forritinu er þannig að flestir ættu að geta skilið hana. Um er að ræða orðhluta, setningarhluta og einföldustu þætti setningafræði. Apertium notast við Subversion (SVN) til að vista kóðann en er hins vegar grundvallað á hefðbundnum forritunarmálum eins og C++ og aðgengilegt þeim sem hafa grundvallarkunnáttu í forritunarmálum.

Apertium fylgir hugmyndafræði Unix-kerfa í flestum aðalatriðum. Apertium er gert úr mörgum smáum einingum (forritum) sem gera aðeins einn hlut í einu og framkvæma hann vel (Forcada o.fl. 2009:4). Þessar einingar eru síðan tengdar saman með textaflæði eins og áður hefur verið greint frá.

5.2 Viðmót og notkun Apertium-kerfisins

Viðmót Apertium-kerfisins er þægilegt á *apertium.org*, sjá *Mynd 2*. Íslenska er valin vinstra megin og síðan býðst að velja ensku eða sænsku hægra megin. Sjálf þýðingin er frá vinstri til hægri, þ.e. frá íslensku yfir á ensku eða sænsku og tekur aðeins nokkrar sekúndur. Notandanum verður þó fljótlega ljóst að bæta þarf íslenska Apertium-kerfið og að það skilar ekki fullkomnum texta frekar en önnur vélþýðingarkerfi. Spurningin *Hvernig gengur?* á íslensku skilar til dæmis niðurstöðunni *How walks*? á ensku. Eins og Martha Dís Brandt benti á verður að leggja frekari vinnu í kerfið til að það geti höndlað m.a. fleiryrtar sagnir, samsett orð og fleira (Martha Dís Brandt 2011:71).



Mynd 2. Viðmót Apertium-kerfisins.

Það sem vélþýðingarkerfi eiga almennt erfiðast með að vinna úr er fyndni, kaldhæðni og margræðni. Hvert orð hefur í vélþýðingarkerfinu aðeins eina merkingu. Apertium-orðabækur ráða í dag einungis við eina samsvörun fyrir hvert orð. En verið er að þróa orðabækurnar þannig að hugsanlega geti þær í framtíðinni skráð fleiri en eitt markmálsjafngildi fyrir hvert uppflettiorð í upprunamálinu (Forcada o.fl. 2011:134). Á þessu er erfitt að ráða bót þannig að erfitt er að segja til um það í dag hve langt tölvutæknin muni komast á þessu sviði.

6 Lokaorð

Í vissum skilningi má segja að Apertium-kerfið sé tilraunastofa í vélþýðingum, þ.e. opinn hugbúnaður sem allir geta tekið þátt í að þróa. Apertium er því einn besti kostur sem Íslendingar eiga nú til að rannsaka og þróa vélþýðingarkerfi fyrir íslensku. Apertium-kerfið hefur þægilegt viðmót og þýðir á örskammri stund af íslensku yfir á ensku eða sænsku. Ákveðnar takmarkanir eru þó enn á íslenska Apertium-kerfinu, t.d. er varðar orðaforða og margræðni.

Niðurstaða þessarar úttektar er sú að Apertium-kerfið sé tiltölulega einfalt í notkun og þróun og bjóði upp á marga möguleika á blandaðri notkun með öðrum kerfum, t.d. vélþýðingarkerfum sem byggjast á dæmum, auk þess sem myndangreinandi eining Apertium-kerfisins geti nýst í öðrum kerfum við samanburðarrannsóknir á tungumálum.

Málföng tungumáls nýtast ekki einungis til vélþýðinga heldur opna þau nýjar leiðir til rannsókna á málinu og á samanburði við önnur tungumál. Því er mjög mikilvægt að grundvallarmálföng séu til fyrir íslensku. Þróun málheilda og orðasafna er sérstaklega mikilvæg til að hægt sé að auka orðaforða og nákvæmni vélþýðingarkerfa fyrir íslensku í framtíðinni.

Heimildir

apertium.org (18. júlí 2015)

Forcada, Mikel L., Francis M. Tyers og Gema Ramírez Sánchez. 2009. The Apertium machine translation platform: five years on. Í: Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez og Francis M. Tyers (ritstj.). *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, bls. 3–10. Alicante.

Forcada, Mikel L., Boyan Ivanov Bonev, Sergio Ortiz Rojas, Juan Antonio Pérez Ortiz, Gema Ramírez Sánchez, Felipe Sánchez Martínez, Carme Armentano-Oller, Marco A. Montava, Francis M. Tyers. 2010. *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform* Apertium. Alicante: Universitat d'Alacant. Departament de Llenguatges i Sistemes Informatic.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez og Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25:127–144.

Goutte, Cyril, Nicola Cancedda, Marc Dymetman og George Foster (ritstj.). 2009. *Learning Machine Translation.* Cambridge/London: The MIT Press.

Hutchins, John. 2003. ALPAC: the (in)famous report. Í: S. Nirenburg, H. Somers og Y. Wilks (ritstj.). *Readings in machine translation*, bls. 131–135. Cambridge: The MIT Press.

Martha Dís Brandt. 2011. *Developing an Icelandic to English Shallow Transfer Machine Translation System.* Háskólinn í Reykjavík.

Martha Dís Brandt, Hrafn Loftsson, Hlynur Sigurþórsson og Francis M. Tyers. 2011. Apertium-IceNLP: A rule-based Icelandic to English machine translation system. Í: *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-2011).* Leuven: European Association for Machine Translation.

O'Keefe, Anne og Michael McCarthy (ritstj.). 2010. *The Routledge Handbook of Corpus Linguistics.* 2010. London / New York: Routledge.

Pala, Karel, Sonja Bosch og Christiane Fellbaum. 2008. Building resources for African languages. Í: *LREC Proceedings*. *SALTMIL Workshop*, bls. 13–18. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W10_Proceedings.pdf

sdl.com (18. júlí 2015)

Sigrún Helgadóttir. 2007. Mörkun íslensks texta. *Orð og tunga* 9:75–107.

Sin-wai, Chan. 2002. Translation and Information Technology: Machine and Machine-aided Translation in the New Century. Í: Chan Sin-wai (ritstj.). *Translation and Information Technology*, bls. vii–xii. Hong Kong: Chinese University Press.

Somers, H. 1999. Review Article: Example-based Machine Translation. *Machine Translation* 14:113–157.

Trujillo, Arturo. 1999. *Translation Engines: Techniques for Machine Translation.* Berlin/Heidelberg: Springer Verlag.

tungutorg.is (18. júlí 2015)

Lykilorð

vélþýðingar, Apertium, máltækni, íslenska

Keywords

machine translation, Apertium, language technology, Icelandic

Abstract

There has been rapid development in language technology and machine translation in recent decades. There are three main types of machine translation: statistical machine translation, rule-based machine translation, and example-based machine translation. In this article the Apertium machine translation system is discussed in particular. While Apertium was originally designed to translate between closely related languages, it can now handle languages that are much more different and variable in structure. Anyone can participate in the development of the Apertium system since it is an open source software. Thus Apertium is one of the best options available in order to research and develop a machine translation system for Icelandic. The Apertium system has an easy-to-use interface, and it translates almost instantly from Icelandic into English or Swedish. However, the system still has certain limitations as regards vocabulary and ambiguity.

Ingibjörg Elsa Björnsdóttir MA
Doktorsnemi í þýðingafræði við Háskóla Íslands
ieb@talnet.is