TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Käbi Suvi

PÕHJASAAMI-EESTI REEGLIPÕHINE MASINTÕLGE PROGRAMMIGA APERTIUM

Bakalaureusetöö

Juhendajad PhD Francis Morton Tyers ja PhD Heiki-Jaan Kaalep

Tartu 2017

SISUKORD

SISSEJUHATUS

Käesolev bakalaureusetöö käsitleb põhjasaami-eesti reeglipõhise masintõlke süsteemi loomist programmiga Apertium. Antud teema on oluline, kuna autori teada ei ole varasemalt loodud põhjasaami-eesti reeglipõhist masintõlkesüsteemi ning soome-ugri keeletehnoloogia koostöö seisukohalt on antud süsteemi ning keelepaari arendamine olulisel kohal.

Bakalaureusetöö eesmärgiks on luua algne põhjasaami-eesti reeglipõhine masintõlkesüsteem. Kitsamaks eesmärgiks antud töös on luua süsteemile Apertium sobivad reeglid, et tõlkekvaliteet põhjasaami-eesti suunal oleks võimalikult hea ning et tõlgitava teksti sisust oleks võimalik aru saada. Ühtlasi on käesoleva töö eesmärgiks reeglipõhise masintõlkesüsteemi loomisprotsessi võimalikult hästi dokumenteerida, et ka edaspidi saaks antud töös välja toodud reegleid rakendada.

Uurimus paigutub põhjasaami keele töötlemisel Apertiumiga laiemasse konteksti – nimelt varasemalt on sama süsteemiga arendatud põhjasaami keele tõlkimist nii norra ja soome kui ka teiste saami keelte suunas. Lisaks on algeline süsteem olemas ka põhjasaami – saksa, baski ja hispaania keelte suunal.

Töös on kolm sisulist peatükki. Esimeses teoreetilises peatükis käsitletakse põhjasaami ja eesti keeli ning tuuakse lühidalt välja keelte sarnasused. Suurem osa peatükist keskendub põhjasaami ja eesti keelte erinevuste väljatoomisele, kuna just erinevused on need, mida reeglipõhise masintõlkesüsteemi loomisel tuleb arvesse võtta. Põhjasaami ja eesti keelte erinevusi on loengumaterjalides käsitlenud Jukka Mettovaara (2014a, 2014b, 2017) ning saami keelte struktuuri on põhjalikult uurinud ka Pekka Sammallahti (1998), kellele antud töös ka suuresti tuginetakse.

Teises samuti teoreetilises peatükis peatutakse pikemalt masintõlkesüsteemil Apertium. Kuna süsteem Apertium töötab n-ö konveier-meetodil, mille ühe etapi väljund on järgmise etapi sisendiks, siis on peatükis lühidalt kirjeldatud Apertiumi süsteemi erinevaid protsesse, mis on välja toodud samas järjekorras, kui reaalses programmis. Põhiliselt on süsteemi kirjeldamisel toetutud Apertiumi ametlikule dokumentatsioonile

(Forcada jt. 2010). Lisaks on peatüki lõpus kirjeldatud ka erinevaid sõnaraamatuid, mida süsteemi töötamiseks vaja läheb.

Kolmandas ja ühtlasi viimases peatükis on kirjeldatud autori läbi viidud praktilise töö protsessi programmiga Apertium. Pikemalt on peatutud nii kakskeelse sõnaraamatu genereerimise protsessil kui ka transfeerireeglite kirjutamisel, sest just nendele osadele keskenduti praktilises osas rohkem. Kolmanda peatüki teises pooles on välja toodud ka reaalsed tulemused, mis praktilise töö käigus saavutati ning ka mõtted programmi edasi arendamiseks. Välja on toodud ka näited sellest, kuidas süsteem hetkel tõlgib põhjasaami teksti eesti keelde.

1 KEELED

Antud uurimuses käsitletavateks keelteks on põhjasaami ja eesti keeled, mis mõlemad kuuluvad uurali keelkonda. Kuigi keeled on omavahel suguluses, ei tähenda keelte kuulumine samasse keelkonda seda, et need vastastikku arusaadavad on ja seega pelgalt kuulates eestlased põhjasaami keelest aru ei saa. Siiski on eesti keelel ja põhjasaami keelel nii mõndagi ühist nii sõnavaras kui ka grammatikas. (Ermakov 2011: 5)

Põhjasaami keel kuulub läänesaami keelte hulka. Põhjasaami keelt räägitakse Norra, Rootsi ja Soome põhjaosas ja keele kõnelejaskonna killustatuse tõttu on raske hinnata selle suurust – kõnelejate arv varieerub erinevate allikate kohaselt 30 000 – 40 000 rääkija (Ethnologue(b)) ja 15 000–20 000 rääkija (Sami information centre) vahel. Tähtsaimateks kontaktkeelteks on norra, rootsi ja soome keeled, mis peegelduvad ka põhjasaami keele sõnavaras. Siiski on põhjasaami keeles rohkem laensõnu norra ja rootsi kui soome keelest. (Marjomaa 2012: 7-8) Põhjasaami keel jaguneb omakorda kolmeks murderühmaks: meresaami, mida räägitakse Tromsö kandis; Finnmarki saami ehk sisemurre, mida räägitakse Põhja-Soomes ja Tornio saami, mida räägitakse Soome, Rootsi ja Norra piirialadel (Sammallahti 1998: 9).

Eesti keel on Eesti Vabariigi ametlikuks keeleks ja selle kõnelejaid on hinnanguliselt 1,1 miljonit. Areaalselt jaguneb eesti keel lõunaeesti murreteks ja põhjaeesti murreteks ja lisaks loetakse eraldiseisvaks ka Kirde-Eesti rannikumurret. Antud uurimuses on kasutatud normeeritud Eesti kirjakeelt, mis on ajalooliselt välja kujunenud põhjaeesti ehk Tallinna keelest, kuid mis tänapäeval sarnaneb kõige enam põhjaeesti keskmurdele. (EKK 2007: 25-30)

Reeglipõhise masintõlke süsteemi tarvis on tunda mõlemat keelt ja seega on järgnevalt välja toodud eesti ja põhjasaami keelte suurimad sarnasused ja erinevused, mida tuleb süsteemi üles ehitades arvesse võtta. Esmalt on räägitud just sarnasustest, kuivõrd erinevused on tõlkesüsteemide jaoks olulisemad, siis nendele on pööratud ka rohkem tähelepanu.

1.1 Põhjasaami ja eesti keele ühisjooni

Nagu juba eelpool mainitud, siis nii põhjasaami keel kui ka eesti keel on mõlemad pärit uurali algkeelest ning täpsemate määratluste järgi on mõlemad keeled soome-permi keeled. Olles seega siiski üsna lähedast päritolu on eesti ja põhjasaami keeltel üsna suur osa sõnavarast ühist päritolu: näiteks eesti ’*kala*’ ja põhjasaami ’*guolli*’või eesti ’*jõgi*’ ja põhjasaami ’*johka*’ (Ermakov 2011: 141). Samuti kasutavad nii põhjasaami (Sammallahti 1998: 95) kui ka eesti (Ethnologue(a)) keeled sõnajärge SVO.

Samuti sarnaselt eesti keelele puudub ka saami keeles substantiividel sugu, kaassõnadest eelistatakse tagasõnu eessõnadele, substantiive käänatakse mitmes käändes ja verbe pööratakse ajas, isikus ja kõneviisis (Mettovaara 2014a: 3). Kui soome keeles on palatalisatsioon indo-euroopa keelte mõjul kadunud, siis nii eesti kui ka saami keeltes on palatalisatsioon säilinud algsel kujul (Sammallahti 1998: 2).

Üheks oluliseks sarnasuseks põhjasaami ja eesti keele vahel on astmevaheldus. Häälikuloolise arengu tulemusena võivad sõnatüved erinevates vormides esineda erineval kujul, näiteks eesti keeles *vedama* : *vean*. Eesti keeles eristatakse kaht suur tüvevahelduse liiki: astmevaheldus ja lõpuvaheldus, millest esimene puudutab tüve sisehäälikuid ja väldet ja teine puudutab tüve lõpuhäälikuid. Astmevaheldus eesti keeles väljendubki selles, et erinevad tüvekujud võivad erineda oma välte või sisehäälikute poolest või ka mõlema poolest korraga nt ’*tõbi : tõve*’. (EKK 2007: 208-209)

Saami keeles kasutusel olev astmevaheldus on laiahaardelisem kui eesti keele puhul – nimelt kehtib astmevaheldus peaaegu iga konsonandi ja konsonantühendi puhul (Mettovaara 2014a: 3). Ka saami keeles eristatakse kolme astet: esimest väldet tähistatakse tüve lõpus ühe tähega ning teist ja kolmandat väldet tähistatakse enamasti kahe või kolme tähega, näiteks *go**đđ**i, go**đđ**it, go**đá**n* ’kuduma’, *vuo**dj**a, vuo**j**a* 'või' ja *gea**đg**i, gea**đgg**it* ’kivi’. (Baal jt. 2012:173-201).

1.2 Põhjasaami ja eesti keele erinevusi

Antud alapeatükis on välja toodud põhjasaami keele suurimad erinevused võrreldes eesti keelega, mis on relevantsed ka masintõlkesüsteemi üles ehitamisel. Välja on toodud suuremad erinevused, mida on ka põhjalikumalt käsitletud – keelte vahel on siiski ka väiksemaid ja mitte nii olulisi erinevusi: näiteks soome keelele sarnaselt saab ka põhjasaami keeles moodustada kas-küsimusi eraldi liitepartikli -go abil (Oahpa! Syntax) ja keelte vahel on ka mõningasi häälduserinevusi (Sammallahti 1998: 40-54).

1.2.1 Ortograafia

Esimene käsitletav suurem erinevus keelte vahel on tähestik. Täna kasutusel olev põhjasaami tähestik on välja arenenud 1832. aastal Rasmus Raski poolt välja pakutud tähestikust (Omniglot). Raski põhiline eesmärk tähestikku luues oli iga erineva heli kohta jätta seda tähistama ühe tähe. Siiski arenesid Soomes, Rootsis ja Norras erinevad tähestikud, kuni need aastal 1979 ühtlustati (Mettovaara 2014a: 5). Põhjasaami tähestikus on eesti tähestikust rohkem tähti ja häälikuid ning järgnevalt on välja toodud ka võõr- ja laensõnades kasutatavad tähed:

Aa, Áá, Bb, Cc, Čč, Dd, Đđ, Ee, Ff, Gg, Hh, Ii, Jj, Kk, Ll, Mm, Nn, Ŋŋ, Oo, Pp, [Qq], Rr, Ss, Šš, Tt, Ŧŧ, Uu, Vv, [Ww], [Xx], [Yy], Zz, Žž, [Ææ (Ää)], [Øø (Öö)], [Åå] (Nickel 1994)

Järgnevalt on selgitatud tundmatute tähtedega tähistatud helisid:

1) **Áá** põhjasaami läänemurretes: /aa/, idamurretes: /a/, /ä/
2) **Đđ** /ð/ heliline dentaalspirant nagu ingl. *this*
3) **Ŋŋ** /ŋ/ velaarnasaal nagu sõnas *kang*
4) **Ŧŧ** /θ/ e helitu dentaalspirant nagu ingl. *think*
5) Kandiliste sulgude sees on ära toodud võõrsõnades ja laensõnades kasutatavad tähed. Näiteks **Yy** /ü/, kasutatakse võõrsõnades: *fysihkka* ’füüsika’ (Mettovaara 2014a: 5)

Põhjasaami keeles ei ole pikk vokaal kuidagi kirjalikult ära märgitud välja arvatud juba välja toodud pikk *a* ehk *á*. Samuti erineb põhjasaami klusiilide märkimise viis eesti keele omast – kui eesti keeles märgitakse neid *p, t, k* siis põhjasaami ortograafias märgitakse *b, d, g* näiteks *kala : guolli*. (Johnson jt. 2017: 116)

1.2.2 Substantiivid

Põhjasaami keeles on vähem käändeid kui eesti keeles. Näiteks puuduvad põhjasaami keelest väliskohakäänded (Sammallahti 1998: 2). Põhjasaami keeles on olenevalt käsitlusest kuus või seitse käänet – nimelt käsitletakse genitiivi ja akusatiivi tihti genitiiv-akusatiivina, kuna enamike nimisõnade ja asesõnade genitiiv ja akusatiiv on ainsuses identsed (Sammallahti 1998: 63).

Nominatiivi (ain. -Ø, mitm. -*t*) kasutatakse põhjasaamisaami keeles enamasti aluse ja öeldistäite käändena. Seda kasutatakse ka omaja- ja olemasolulauses omatava või olemasoleva asja käändena. Lisaks on ka kindlad verbid, millega tuleb kasutada nominatiivi: näiteks *šaddat* 'millekski saama' ja *orrut* 'paistma, välja nägema'. (Mettovaara 2014b; Sammallahti 1998: 65-70).

Nagu eesti keele omastav kääne on ka põhjasaami keeles genitiiv-akusatiiv (ain. -Ø, mitm. -*id*) omaja kääne. Lisaks kasutatakse just genitiiv-akusatiivi ees- ja tagasõnadega ja arvsõnadega, kui arv on suurem kui üks (1). Genitiiv-akusatiiv põhjasaami keeles on sihitise käändeks ja infiniitsete verbivormide aluse käändeks. (Mettovaara 2014b; Sammallahti 1998: 65-70) Eesti keelele sarnaselt puudub põhjasaami keeles genitiivi ainsuses käändelõpp (Mettovaara 2017).

Illatiivi (ain. -*i*, mitm. -*ide*) kasutatakse põhjasaami keeles sarnaselt eesti keelele sihtikoha käändena. Siiski väliskohakäänete puudumise tõttu on sellel laiem kasutusviis kui eesti keeles. Näiteks märgitakse põhjasaami keeles illatiiviga ka vastuvõtjat või

saajat, tegevuse lõpp-punkti või tähtaega ja passiiviverbi tegijat. Lisaks on ka hulganisti verbe, mille rektsioon nõuab samuti illatiivi: näiteks *liikot* ’meeldima’ ja *báhcit* ’jääma’. (Mettovaara 2014b, 2017; Sammallahti 1998: 65-70)

Lokatiiviga (ain. *-s*, mitm. *-in*) tähistatakse põhjasaami keeles peamiselt asukohta, lähtekohta, positsiooni ja omajat. Lisaks kasutatakse seda ka siis, kui tahetakse väljendada kellegi psühholoogilist või kehalist seisundit, kogejat või valdajat. Lokatiiviga tähistatakse ka tegijat siis, kui tahetakse rõhutada, et tegevus on toimunud kogemata: ***Máhtes*** *gahčai lássa láhttái* ’Mahttel kukkus klaas põrandale’. Kasutatakse ka näiteks verbiga ’kartma’ *ballat*. (Mettovaara 2014b; Sammallahti 1998: 65-70) Põhjasaami keele lokatiiv võib vastata vähemalt neljale eesti keele käändele – inessiiv, elatiiv, adessiiv ja ablatiiv (Mettovaara 2017).

Komitatiiv (ain. *-in*, mitm. *-iguin*) väljendab ka põhjasaami keeles seda, et kelle või millega koos ollakse. Samuti eesti keelele sarnaselt kasutatakse seda vahendi ja sõiduki väljendamiseks. Kuna põhjasaami keeles on olemas duaal, siis kasutatakse komitatiivi ka selleks, kui on vaja täpsustada kahest isikust teine: *Boahtibeahttigo doai* ***Iŋggáin*** *fárrui?* ’Kas sa ja Inga tulete kaasa?’. (Mettovaara 2014b; Sammallahti 1998: 65-70)

Essiiviga (ain./mitm. *-n*) tähistatakse põhjasaami keeles öeldistäidet ja määrust, mis väljendab hetkeseisundit. Lisaks kasutatakse essiivi ka seal, kus eesti keeles kasutatakse translatiivi – milleks või missuguseks alus saab. Essiiv toimib põhjasaami keeles ka ajamäärusena siis, kui tegemist on inimese elu ajastutega, näiteks: lapsepõlv, noorus jne. Määrused, mis kujutavad ilma või loodusega seotud tingimusi, on samuti essiivis, näiteks: pimedus, vihm. (Mettovaara 2014b; Sammallahti 1998: 65-70)

Põhjasaami keeles on käändeparadigmas olulisel kohal ka sõnatüvi ja substantiivi jalad. Kui jala mõistet kasutatakse eesti keeleruumis peamiselt värsimõõdu iseloomustamiseks, siis põhjasaami grammatikas on sellel laiem kasutus. Jala all on käesolevas töös mõeldud silpide kogumit, millest esimene on rõhuline ning järgnevad on rõhutud (ENE 2006). Nimelt määrab substantiivi viimane jalg ära selle, kas

nominatiivis ja nominatiivi mitmuses on sõnad tugevas või nõrgas astmes. Näiteks substantiivid, mille nominatiivi mitmuses on kaks silpi, on nominatiivi ainsuses tugevas astmes ja nominatiivi mitmuses nõrgas astmes. (Oahpa! Nouns)

1.2.3 Possessiivsufiksid

Soome keelele sarnaselt on ka põhjasaami keeles olemas possessiivsufiksid, mis näitavad kuuluvust. Siiski on possessiivsufiksite näol põhjasaami keeles tegemist märksa keerulisema morfoloogilise süsteemiga kui näiteks soome keeles. (Janda, Antonsen 2016: 332) Järgnevalt on välja toodud nimetava ainsuse näitel (*guos'si* 'külaline') possessiivsufiksid, mis on paksus kirjas märgitud:

	Sg	Du	Pl
1.	*guos'sá**n***	*guos'sá**me***	*guos'sá**met***
2.	*guos'sá**t***	*guos'sá**de***	*guos'sá**det***
3.	*guos'si**s***	*guos'si**ska***	*guos'si**set***

(Sammallahti 1998: 65)

L. Janda ja L. Antonseni 2016. aastal tehtud uurimistöö põhjal selgub, et põhjasaami possessiivsufiksite kasutus on tasapisi hakanud muutuma – autorid on leidnud, et possessiivsufiks hakkab asenduma refleksiivse genitiivse pronoomeniga. Tõenäoliselt on põhjus põhjasaamide sotsiolingvistilises situatsioonis, kus kohalikud täielikult oma emakeelses keskkonnas üles kasvada ei saa. (Janda, Antonsen 2016: 360)

1.2.4 Verbid

Põhjasaami keeles on olemas nii isikuline kui ka umbisikuline tegumood, kuid umbisikulist tegumoodi käsitletakse verbituletisena. Lisaks on veel kasutusel ka kindel kõneviis, tingiv kõneviis, käskiv kõneviis ja eesti keelele võõras potentsiaalne kõneviis.

Tegusõnu saab pöörata ainsuses, mitmuses ja duaalis, mida samuti eesti keeles ei esine. (Sammallahti 1998: 76) Astmevaheldus kehtib laialdaselt ka tegusõnade puhul ja sealgi tuleb otsustamiseks jälgida sõnatüve jalga (Oahpa! Verbs). Lisaks jällegi soome keelele sarnaselt on põhjasaami keeles olemas ka pöörduv eitusverb, mis pöördub erinevate isikute puhul, kuid ei pöördu ajas (Mettovaara 2014a: 5).

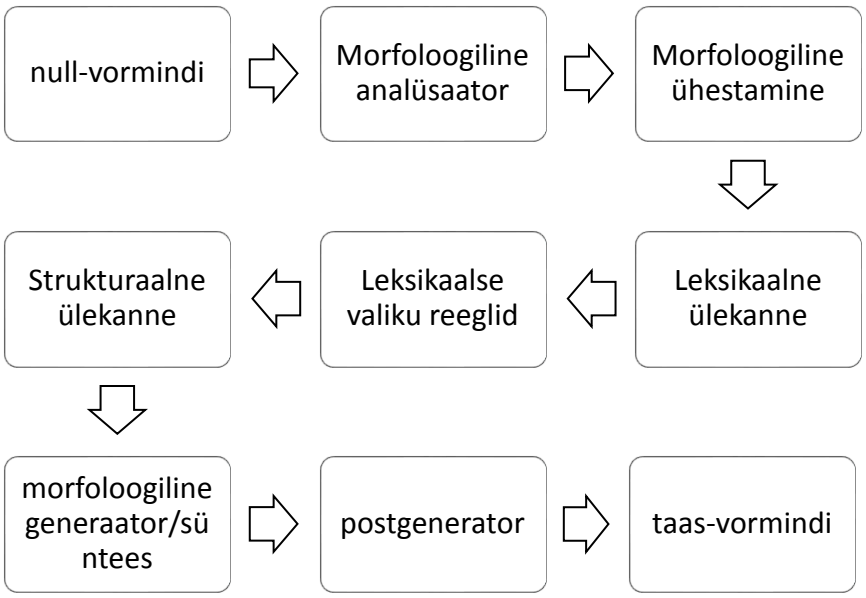
2 SÜSTEEM APERTIUM

Apertium on tasuta avatud lähtekoodiga reeglipõhise masintõlke tarkvara süsteem, mis sisaldab endas masintõlkesüsteemi ülesehitamiseks vajalikke andmeid. Esialgu hakati platvormi arendama just suunaga sugulaskeelte tõlkimisele, kuid praeguseks on seda laiendatud ka keelepaaridele, kus keeled on pärit erinevatest keeleperekondadest. Apertiumi süsteem põhineb suuresti andmetel laiahaardelisel kasutusel – morfoloogilised sõnastikud, grammatikareeglid, paralleelsed tekstid jpm. (Forcada jt 2011: 128)

Apertiumi arendajate eesmärgiks on olnud vähemuskeelte ja vähem kasutatavate keelte toetamine (Tyers 2015a) ja seega on ka antud töö autor leidnud, et põhjasaami-eesti masintõlke arendamiseks võiks kasutada just seda süsteemi. Süsteem ise on üles ehitatud sõna-sõnalise tõlkesüsteemi peale ja see kasutab eraldiseisvatest programmidest kokku pandud n-ö konveierit, kus eelneva programmi väljund on järgneva sisend.(Tyers 2015a)

Reeglipõhine masintõlge on ajalooliselt esimene lähenemine masintõlkele. Reeglipõhine masintõlge koosneb kahest suuremast komponendist: reeglid, mis arvestavad keelte erinevuseid, ja leksikonid, kus asuvad morfoloogiline, süntaktiline ja semantiline informatsioon. Kuna nii leksikonid kui ka reeglid on kirjutatud programmeerijate poolt, siis on reeglipõhine masintõlge nii aja kui ka ressursside poolest väga kulukas meetod tõlkesüsteemi ehitamiseks. (A.-L. Lagrada jt. 2009: 217-218)

2.1 Süsteemi ülesehitus



Joonis 1. **Apertiumi moodulite järjestus**

Nagu juba eelpool mainitud, siis koosneb Apertium erinevatest iseseisvatest programmidest, mis on kokku pandud üheks süsteemiks ja mille ühe osa väljund on järgmise sisend (Joonis 1). Järgnevalt seletatakse põhjalikumalt, mida iga osa eraldi teeb. Mõistete tõlkimiseks on kasutatud Heikki Vallaste koostatud e-teatmiku, mis on kättesaadav aadressil www.vallaste.ee (Vallaste 2017) ning puuduva vaste korral on ka autori poolt pakutud välja eestikeelseid vasteid.

2.1.1 Null-vormindi ja taas-vormindi

Null-vormindi eraldab tõlgitava teksti ja formaadiga kaasneva info, mis eemaldatakse – näiteks eemaldatakse RTF, HTML vms. märgendus (Forcada jt 2010: 6). Kõik formaadiga seonduv kapseldatakse eraldajate vahele ja suuremad tekstiblokid kirjutatakse ümber ajutisteks failideks. Kapseldatud formaadiplokkideks võivad olla: 1)

mitte-tühjad formaadiplokid, mis sisaldavad sisenddokumendi formaadi informatsiooni, 2) plokid, mis sisaldavad viidet kolmandale failile, 3) tühjad formaadiplokid tehisinformatsiooniga, mille ette süsteem genereerib kirjavahemärgi. (Forcada jt 2010: 111-112)

Plokkide loomise reeglid on järgmised: 1) kõik, mis ei ole osa tõlgitavast tekstist, peab olema kapseldatud formaadiplokkidesse, 2) kõrvuti ei tohi olla kaks mitte-tühja formaadiplokki – kui on, siis need ühendatakse üheks suureks formaadiplokiks, 3) tühjale formaadiplokile peab eelnema mitte-tühi formaadiplokk või faili lõpp. (Forcada jt 2010: 112) Pärast programmi töö lõppu saadetakse tekst viimase sammuna taas-vormindisse, mis taastab sisendteksti originaalse formaadi (Forcada jt 2010: 9).

Käesoleva töö raames null-vormindit ega taas-vormindit ei rakendata, kuna see on relevantne ainult siis, kui tegu on tõepoolest lõppkasutaja programmiga.

2.1.2 Morfoloogiline analüsaator

Morfoloogiline analüsaator segmenteerib teksti tekstisõnadeks ning seejärel pakub igale tekstisõnale leksikaalse vormi, mis sisaldab lemmat, leksikaalset kategooriat (substantiiv, verb, adjektiiv jne) ja morfoloogilist informatsiooni (Forcada jt 2011: 130). Morfoloogiline analüsaator analüüsib sõna vormi ning sellele vastavalt annab ka infot sõna struktuuri, arvu, käände, pöörde, sõnaliigi jpm kohta. Kui analüsaator ei suuda üheselt sõna analüüsida, siis annab see vasteks vähemalt kaks erinevat leksikaalset vormi (Forcada jt 2010: 7).

2.1.3 Morfoloogiline ühestamine

Kui morfoloogiline analüsaator on pakkunud välja mitu leksikaalset vormi ühele tekstisõnale, siis toimub järgmise sammuna morfoloogiline ühestamine. Ühestaja valib lause konteksti põhjal mitu analüüsi saanud sõnadele ühe analüüsi. Antud etapp

reeglipõhises masintõlkes on kõige rohkemate tõlkevigade põhjustajaks – kui ühestaja valib vale analüüsi, siis tuleb ka vale tõlge. (Forcada jt 2010: 8)

Ühestajat treenitakse sisendkeele korpuse peal, kuid lisaks rakendatakse sellele ka piiranguid, mis keelavad või kohustavad kindlate sõnavormide järjest esinemise. Näiteks hispaania keeles ei saa prepositsioonile järgneda personaalses vormis verb ja seega aitavad sellised reeglid ühestaja tööd lihtsustada. (Forcada jt 2010: 56-57)

2.1.4 Leksikaalse ülekande moodul

Leksikaalse ülekande moodul genereeritakse automaatselt kakskeelsest sõnastikust. Sõnastikus vastab igale sisendkeelsele sõnavormile vaid üks väljundkeelne vaste. Ühendverbe käsitletakse ühe üksusena, millele vastab samuti üks väljundkeelne sõna. (Forcada jt 2010: 8)

2.1.5 Leksikaalse valiku reeglite moodul

Leksikaalse valiku reeglite moodul lisati Apertiumi süsteemi 2007. aastal, kui kakskeelsesse sõnastikku sai sisendkeelsele sõnale lisama hakata mitu väljundkeelset vastet. Leksikaalse valiku reeglite moodul valib konteksti arvestades kõige sobivama väljundkeelse vaste. (Forcada jt 2010: 23,36) Platvorm on arendatud kiireks, et tuhanded sõnad sekunditega tõlkida, lihtsaks ja iseseisvaks, et seda saaks arendada ka ilma andmete või suure paralleelkorpuse olemasoluta (Tyers jt, 2012: 214).

Apertiumi süsteemis on kaks võimalust, kuidas olukorda, mil sõnal on mitu võimalikku vastet, käsitletakse. Esiteks, olukorras, kus väljundkeelsed vasted on sünonüümid ning ühe või teise eelistamine ei viiks tõlkeveani, saab lingvist ise valida sobivaima lemma lisades teistele lemmadele piirangud. Teiseks, olukorras, kus vale väljundkeelne vaste võib viia tõlkeveani, antakse lingvistile võimalus märgendada erinevad lemmad nii, et

leksikaalse valiku reeglid statistilisel põhimõttel suudavad valida konteksti arvestades õige lemma. (Forcada jt 2010 : 66-67)

2.1.6 Strukturaalse ülekande moodul

Strukturaalse ülekande moodul praegusel kujul loodi samuti 2007. aastal, kui Apertiumit hakati kasutama keelte tõlkimiseks, mis ei ole samast keeleperest pärit. Uus ja edasi arendatud moodul koosneb kolmest astmest: leksikaalne aste, sõnajärjendite aste ja sõnajärjendite järjendite aste. (Forcada jt 2010: 73-77)

Strukturaalse ülekande moodul analüüsib sõnajärjendeid ja/või fraase, mis vajavad spetsiaalset tähelepanu tulenevalt sisendkeele ja väljundkeele grammatilistest erinevustest. Näiteks muudetakse just selles moodulis vajadusel sõnajärge, sugu, arvu jne. Moodul genereeritakse failist, kuhu on kirjutatud iga üksikjuhtumi kohta konkreetsed reeglid, kuidas väljundkeeles sõnajärjend välja peaks nägema (Forcada jt 2010: 8-9).

2.1.7 Morfoloogiline generaator/süntees ja postgeneraator

Morfoloogiline generaator moodustatakse automaatselt morfoloogilise sõnastiku (vt. 2.2.1) alusel. Morfoloogiline generaator annab igale väljundkeelsele leksikaalsele sõnavormile vasteks tekstisõna. (Forcada jt 2010: 9)

Postgeneraator tegutseb samuti vaid väljundkeelse tulemiga ning selle mooduli ülesandeks on vaadata üle ja teha ortograafilist korrektuuri. Moodul genereeritakse automaatselt teisendusreeglite failist, mis on ülesehituselt sarnane eelmainitud sõnastikele. (Forcada jt 2010: 9) Ortograafiline korrektuur seisneb näiteks üleliigsete tühikute kustutamises, punktuatsiooni parandamises jpm (Forcada jt 2010: 41-43). Ka postgeneraatorit antud töös ehitatud programmi raames ei kasutata ega arendata.

2.2 Sõnastikud

Järgnevalt on lühidalt kirjeldatud sõnastikke, mida programm Apertium oma toimimiseks vajab.

2.2.1 Morfoloogilised sõnastikud

Morfoloogiline ülekanne viiakse läbi Helsingis arendatud lõpliku muunduri tehnoloogiat (HFST) kasutades. Antud tehnoloogiat on laialdaselt kasutatud just uurali keelte puhul ning see on kättesaadav vaba litsensi all. Mõlema keele morfoloogiat rakendatakse lexc laiendiga failis ja mõlema keele morfofonoloogiat rakendatakse twolc laendiga failis. Üht ja sama morfoloogilist kirjeldust kasutatakse nii morfoloogilise analüüsi kui ka morfoloogilise generaatori moodulites. (Johnson jt. 2017: 117)

2.2.2 Kakskeelsed sõnastikud

Kakskeelset sõnastikku vajab leksikaalse ülekande moodul, selleks et leida igale sisendkeele leksikaalsele vormile väljundkeelne vaste. Taaskord saab kakskeelsest sõnastikku kasutada kahel eesmärgil: olenevalt tõlkesuunast loetakse sõnastikku kas vasakult paremale või vastupidi. Kakskeelsed sõnastikud sisaldavad enamasti informatsiooni vaid lemma ja sõnaliigi kohta. (Forcada jt 2010: 22-23)

2.3 Varasemad käsitlused põhjasaami keele tõlkimisest Apertiumiga

Nagu juba eelnevalt mainitud, siis varasemalt ei ole arendatud Apertiumi süsteemi suunal põhjasaami-eesti. Siiski varasemalt on tegeletud juba põhjasaami keelega süsteemselt Apertiumi süsteemis: arendatud on põhjasaami-norra ja põhjasaami-soome tõlkesüsteeme ning lisaks on arendatud ka erinevate saami keelte vahelist tõlget (põhjasaami - lule saami). Põhilised põhjused just nende keelepaaride arendamiseks on tõenäoliselt olnud piirkondlikud põhjused – kuna saamid elavad nii Norras kui ka Soomes siis ongi just need keelepaarid sattunud fookusesse.

Põhjasaami-norra keelepaari arendamist Apertiumis on arendanud põhiliselt Trond Trosterud (2012) koostöös Kevin Unhammeriga. Põhjasaami – lule saami keelepaari tõlkimist on põhjalikumalt käsitlenud Lene Antonsen, Trond Trosterud ja Francis M. Tyers (2016) ja samuti lisaks eelmainitutele ka Linda Wiechetek (Tyers jt. 2009).

Põhjasaami keele kasutamist masintõlkes käsitleti põhjalikult ka 2017. aasta maikuus toimuval 21. põhjamaade arvutilingvistika konverentsil, kus tutvustati lähemalt põhjasaami-soome keelepaari arendamist Apertiumiga (Johnson jt. 2017). Lisaks tutvustati konverentsil Apertiumis arendatavat projekti, mille käigus käsitletakse põhjasaami keelt ülekandekeelena tõlkides suunal soome/norra – põhjasaami – lõunasaami/lule saami/inari saami keel (Antonsen jt. 2017).

3 PÕHJASAAMI-EESTI REEGLIPÕHINE MASINTÕLGE

Järgnevas peatükis on antud ülevaade praktilisest tehtud tööst programmiga Apertium. Praktiline osa koosnes esmalt programmi installeerimisest ja seejärel kakskeelse sõnastiku automaatsest genereerimisest ning reeglite kirjutamisest, millest mõlemast tuleb alapeatükkides lähemalt juttu. Morfoloogiliste sõnastike loomist antud töös ei käsitleta, kuna autor sai nii põhjasaami kui ka eesti morfoloogilise sõnastiku juba olemasolevatest töötavatest süsteemidest. Välja on toodud ka praktilise töö tulemused ja samuti võimalused ning mõtted süsteemi täiendamiseks.

3.1 Programmi installeerimine

Programmi installeerimiseks seati esmalt operatsioonisüsteemi Windows üles Virtualbox Linux'i operatsioonisüsteemiga. Selleks, et Apertiumit Virtualbox'is üles seada, kasutati Tarmo Vaino soome-eesti masintõlke tarbeks loodud skrpiti, mida kohandati põhjasaami-eesti keelepaari jaoks (Lisa 1). Siiski ei rakendatud skriptist kõiki osi, kuna juba eelnevalt oli seatud üles Virtualbox, mis sisaldas Apertiumi algosiseid ning seega tuli kompileerida vaid põhjasaami ja eesti keeled ja ka põhjasaami-eesti keelepaar.

3.2 Kakskeelne sõnastik

Toimiva reeglipõhise masintõlke jaoks on vaja kakskeelse sõnastiku olemasolu. Põhjasaami-eesti suunal oli juba varasemast olemas ka pisikene kakskeelne sõnastik, mis oli autori poolt eelnevalt käsitsi loodud. Kuna aga toimiva masintõlke süsteemi loomiseks on vaja mahukamat sõnastikku, siis selleks, et vältida selle käsitsi sisestamist, otsustati kasutada *linux*'i tekstitöötlusvahendeid, et saaks automaatselt ja kiiremini arvestatava suurusega sõnastiku. Uue mahukama sõnastiku loomiseks on seega

kasutatud kahte olemasolevat sõnastikku: põhjasaami-soome sõnastik, kus on kokku 19 781 rida ja soome-eesti sõnastikku, kus on kokku 17 685 rida. Näitelaused 1 ja 2 illustreerivad põhjasaami-soome ja soome-eesti sõnastike kujusid.

(1) <e> <l>rehket<s n="n"/></l><r>lasku<s n="n"/></r> </e>

(2) <e> <l>lasku<s n="n"/></l><r>arve<s n="n"/></r> </e>

Kuna vaadeldavad sõnastikud olid veidi erinevalt vormistatud, siis tuli esmalt need vormistada sarnaseks, st. et käsuga, mis on välja toodud näites nr 3 sai eemaldatud rea algusest üleliigsed tühikud, mis soome-eesti sõnastikus esinesid. Kuna sõnastike alguses on ka palju definitsioone, mida sõnastike ühildamisel kasutada ei saanud, siis 4. näites välja toodud käsuga sai sõnastikest kätte vaid sinna lisatud sõnade kirjed.

Edasi tuli viia failid veelgi sarnasemale kujule, st tuli sõnastikele anda sama järjestus – põhjasaami-soome sõnastik viidi kujule soome-põhjasaami. Selle protsessi jaoks kasutati käsku, mis vahetas ühel real ära sõnede asukohad (näide 5).

(3) tr -s ’ ’

(4) grep '^ <e> <l>[^<]'

(5) sed 's/\(^.*<l>\)\(.*\)\(<\/l><r>\)\(.*\)\(<\/r>.*$\)/\1\4\3\2\5/'

Kui mõlemad sõnastikud olid samale suunale viidud, siis edasi kasutati käsku *join*. Kuna käsk *join* ühendab faile veergude kaupa ja vaikimisi on veeru eraldajaks tühik, siis selleks, et oleks mugavam grupeerida, defineeriti uueks veeru eraldajaks ’@’. Näitest 6 võib näha, milline oli rida failist enne käsku *join*, kuhu on lisatud juba veergude eraldamiseks märgis ’@’. Puhastatud ja märgendatud väljundid kirjutati uutesse

failidesse, et saaks edaspidi nendega lihtsamalt töötada. Uued failid ka sorteeriti tähestikulises järjekorras, sest vastasel juhul *join* käsk ei töötaks.

(6) <e> <l>@aavistus<s n="n"/>@</l><r>@kahtlus<s n="n"/>@</r> </e>

(7) join -1 2 -2 2 -t @ -e ### -o 1.1 1.4 1.3 2.4 1.5 eraldajad-sort-sme.txt eraldajad-sort-est.txt

Join käsu jaoks on vaja kasutada mitmeid erinevaid parameetreid ja lipukesi. Töös kasutatavat käsku näeb näites nr. 7. Esmalt parameetrid -1 ja -2, mis tähistavad vastavalt esimest ja teist faili. Lipukese järel olev number (nt -1 2) tähistab välja, mille järgi faile ühendatakse. Lipuke -t laseb määrata väljade jaoks eraldaja ning nagu juba eelnevalt öeldud, siis määrati veergude eraldajaks '@'. Lipuke -e näitab märgistusega ### ka ridu, mida kokku ei pandud. Viimaks lipuke -o, mis laseb määrata, milliseid välju väljundis näha peaks olema.

Peale *join* käsu kasutama õppimist ilmnes, et faile oleks saanud ühendada ka lihtsamini ning, kuna väljade eraldaja saab ise määrata kuhu vajadust on, siis ei oleks pidanud vahetama põhjasaami-soome sõnastiku suunda. Nimelt oleks lihtsalt saanud käsule öelda, et mitmenda välja järgi faile ühendada tuleb ning seejärel oleks juba sama tulemuseni jõudnud.

Peale *join* käsu kasutamist olevas põhjasaami-eesti sõnastikus on 5641 rida, mille hulka on lisatud ka 242 eelnevalt käsitsi sisestatud rida, mida juba eelpool mainitud on. Pea pooled sõnadest, mis põhjasaami-eesti sõnastikku genereeriti on nimisõnad. See on tingitud kindlasti suurest hulgast nimisõnadest, mis olid juba olemas nii põhjasaami-soome kui ka soome-eesti sõnastikus. Peale nimisõnade on uues sõnastikus enim verbe.

Kuna põhjasaami-soome sõnastikus oli väga suur hulk pärisnimesi – lausa 42,8% põhjasaami-soome sõnastiku suurusest ehk 7 578 pärisnime – siis oli see ka suurim osa, mis jäi automaatselt põhjasaami-eesti sõnastikku sisestamata, kuna soome-eesti sõnastikus on pärisnimesi vaid 400 ringis.

3.3 Transfeerireeglid

Hoolimata sellest, et põhjasaami ja eesti keeled on pärit samast keeleperekonnast on need siiski struktuurilt erinevad, mis tähendab, et suur osa praktilisest tööst toimus Apertiumi strukturaalse ülekande moodulis. Nagu juba eelpool mainitud, siis strukturaalse ülekande moodul töötab toetudes failidele, kuhu on kirjutatud transfeerireeglid, kuidas sisendkeelt väljundkeeleks tõlkida.

Paljude reeglite kirjutamisel saadi abi nii põhjasaami-soome transfeerireeglitest kui ka soome-eesti transfeerireeglitest. Siiski, reegleid soome-eesti masintõlkesüsteemist üle võttes ja rakendades neid põhjasaami-eesti suunal, tuli arvesse võtta ka soome keele erinevusi eesti keelest. Näiteks reegel, mis teisendab valitud juhtudel illatiivi translatiiviks, eesti keelele ei kehti (nt fin ’*puhumme venäjä**ksi***’ est ’*räägime vene keele**s***’). Käesoleva töö käigus valminud transfeerireeglite põhiline ülesanne on teisendada morfoloogilisi kategooriaid, sest see on süsteemi jaoks põhiline võti, mille põhjal on juba lihtne õige vorm moodustada.

3.3.1 Käändsõnad

Esimene, mille kohta reegleid kirjutati, oli lokatiivi kääne, mis põhjasaami keeles esineb, kuid eesti keeles puudub. Põhiliselt muutub lokatiiv eesti keelde tõlkides inessiiviks ja elatiiviks, kuid see võib olenevalt kontekstist muutuda ka adessiiviks (Mettovaara 2017). Kuna ka soome keeles puudub lokatiiv, siis kasutati ära põhjasaami-soome juba olemasolevaid reegleid, mida veidi modifitseeriti ning rakendati põhjasaami-eesti tõlkesüsteemile.

Näidetes 8 ja 9 väljatoodud reeglid käsitlevad lokatiivi muutumist elatiiviks, arvestades ka sõna semantilist märgendit. Taaskord semantilist märgendit arvestades tõlgitakse näites 10 lokatiiv adessiiviks. Näide 11 on kõige lihtsam lokatiivi käsitlev reegel, mis

tõlgib lokatiivi, millel muud semantilist märgendit ei ole, inessiiviks, mis on ka sagedasim kääne, milleks lokatiiv eesti keeles muutub.

(8)
```
<when>
<test><and><equal><clip pos="1" side="tl" part="a_cas"/><lit-tag v="loc"/></equal>
<or><equal><clip pos="1" side="sl" part="a_func"/><lit-tag v="@ADVL-ela→"/></equal>
<equal><clip pos="1" side="sl" part="a_func"/><lit-tag v="@←ADVL-ela"/></equal></or></and></test>
<let><clip pos="1" side="tl" part="a_cas"/><lit-tag v="ela"/></let>
</when>
```

(9)
```
<when>
<test><and><equal><clip pos="1" side="tl" part="a_cas"/><lit-tag v="loc"/></equal>
<equal><clip pos="1" side="sl" part="a_func"/><lit-tag v="@N←"/></equal></and></test>
<let><clip pos="1" side="tl" part="a_cas"/><lit-tag v="ela"/></let>
</when>
```

(10)
```
<when>
<test><and><equal><clip pos="1" side="tl" part="a_cas"/><lit-tag v="loc"/></equal>
<equal><clip pos="1" side="sl" part="x_func"/><lit-tag v="←hab→"/></equal></and></test>
<let><clip pos="1" side="tl" part="a_cas"/><lit-tag v="ade"/></let>
</when>
```

(11)
```
<when>
<test><equal><clip pos="1" side="tl" part="a_cas"/><lit-tag v="loc"/></equal></test>
<let><clip pos="1" side="tl" part="a_cas"/><lit-tag v="ine"/></let>
</when>
```

Põhjasaami kääne genitiiv-akusatiiv võib samuti nii eesti kui ka soome keelde tõlgituna olla kas genitiiv või ka partitiiv. Reeglite rakendumisel on taaskord arvesse võetud sõnade semantiline märgendus, mille üht hulka on defineeritud ’a_func_obj’, näiteks 12. näitereeglis teisendatakse akusatiiv vastavalt semantilisele märgendusele partitiiviks.

(12)
```
<when>
  <test><and><equal><clip pos="1" side="tl" part="a_cas"/><lit-tag v="acc"/></equal>
    <not><equal><clip pos="1" side="sl" part="a_func_obj"/><lit
v=""/></equal></not>
    <not><equal><clip pos="1" side="tl" part="a_pos"/><lit-tag
v="prn.pers"/></equal></not></and></test>
  <let><clip pos="1" side="tl" part="a_cas"/><lit-tag v="par"/></let>
</when>
```

Nagu alapeatükis 1.2.3 mainitud, siis soome keelele sarnaselt on põhjasaami keeles olemas possessiivsufiksid, mis liituvad nimisõnadele märkimaks objekti kuuluvust. Siiski soome keelest erinevalt saab põhjasaami keeles märkida ka duaali possessiivsufiksiga. Duaali eesti keelde tõlkides muutub see mitmuseks ja vastavalt isikule siis ka 1., 2. või 3. isiku mitmuseks. Näites 13 on välja toodud reegel, mis muudab duaali kategooriad eesti keelde tõlkimisel mitmusekategooriateks.

(13)
```
<when>
  <test><equal><clip pos="1" side="tl" part="a_px"/><lit-tag v="px1du"/></equal></test>
  <let><clip pos="1" side="tl" part="a_px"/><lit-tag v="px1pl"/></let>
</when>
<when>
  <test><equal><clip pos="1" side="tl" part="a_px"/><lit-tag v="px2du"/></equal></test>
  <let><clip pos="1" side="tl" part="a_px"/><lit-tag v="px2pl"/></let>
</when>
<when>
  <test><equal><clip pos="1" side="tl" part="a_px"/><lit-tag v="px3du"/></equal></test>
  <let><clip pos="1" side="tl" part="a_px"/><lit-tag v="px3pl"/></let>
</when>
```

3.3.2 Pöördsõnad

Selleks, et programm pöördsõnu õigesti käsitleks, tuli ka pöördsõnadele kirjutada reeglid. Esialgu kirjutati programmi väga lihtsad ja tühjad reeglid, selleks, et programm

vähemalt töötaks. Kui see samm tehtud, siis hakati uurima, kuidas reegleid edasi arendada.

Aspekt, millele eriti pöördsõnadele reegleid kirjutades tähelepanu tuli pöörata, olid kategooriad ja kategooriate teisendused. Nimelt selleks, et programm suudaks tõlgitud sõnu samasse pöördesse panna, mis sisendkeeles, tuleb ka väljundkeelsele sõnale anda samasugused ning samas järjekorras kategooriad. Näiteks, kui põhjasaami sõna esineb duaalis, siis reeglitega tuleks duaal automaatselt teisendada mitmuseks nagu juba eelnevalt ka possessiivsufiksite puhul välja toodud. Samuti on näiteks põhjasaami verbidel kategooriad erinevas järjekorras kui eesti verbikategooriatel.

(14) $ echo 'liikoba' | hfst-lookup sme-est.automorf.hfst
liikoba liikot<vblex><iv><indic><pres><p3><du> 0,000000

$ echo 'armastavad' | hfst-lookup est-sme.automorf.hfst
armastavad armastama<vblex><actv><pres><indic><p3><pl> 0,000000

Näites 14 on välja toodud esmalt põhjasaami sõna ’*liikoba*’ üks võimalik tähendus ja selle kategooriad. Samuti on näidatud eestikeelse sama sõna sama vormi kategooriad. Näitest võib näha seda, et põhjasaami verbidel on olemas märgend selle kohta, kas tegemist on transitiivse (’*tv*’) või intransitiivse (’*iv*’) verbiga, samas kui eesti keeles vastav märgend puudub. Vastupidiselt on eesti keeles ära märgitud, kas tegu on isikulise või umbisikulise tegumoega (märgend <actv>) – see märgend puudub põhjasaami verbidel. Lisaks tuleb kindlasti tähele panna kategooriate järjestust: kategooriad <indic><pres> on eesti keeles teises järjekorras kui põhjasaami keeles ning ka seda tuleb reeglitega muuta. Vastav näitereeglid ongi välja toodud näites 15, kus vahetatakse kategooriate järjekorda: reegel teisendab järjestuse <indic><pres> või <indic><pret> järjestuseks <pres><indic> või <pret><indic>.

(15) <when>
<test><equal><clip pos="1" side="sl" part="a_screeve"/><lit-tag v="indic.pres"/></equal></test>

```
        <let><clip pos="1" side="tl" part="a_screeve"/><lit-tag v="pres.indic"/></let>
      </when>
      <when>
        <test><equal><clip pos="1" side="sl" part="a_screeve"/><lit-tag v="indic.pret"/></equal></test>
        <let><clip pos="1" side="tl" part="a_screeve"/><lit-tag v="pret.indic"/></let>
      </when>
```

Näites number 16 välja toodud reegel käsitleb märgendi <actv> puudumist põhjasaami keeles. Nimelt põhjasaami keeles puudub umbisikuline tegumood sellisel kujul nagu see eksisteerib eesti keeles, vaid seda käsitletakse põhjasaami keeles verbituletisena (Mettovaara 2017). Märgend <actv> lisatakse reegli järgi igale sõnale automaatselt, kui sel sõnal on olemas ka märgend <vblex>, mis tähistab verbi. Antud reegel on kirjutatud selliseks, kuna töö autoril ei ole veel põhjalikke kogemusi reeglite kirjutamisel ning antud reegel tuleb tulevikus kindlasti ümber muuta, et iga verb ei saaks ekslikult märgendit <actv> nagu see praegusel hetkel saab.

(16)
```
      <when>
        <test><equal><clip pos="1" side="sl" part="a_actv"/><lit-tag v="vblex"/></equal></test>
        <let><clip pos="1" side="tl" part="a_actv"/><lit-tag v="vblex.actv"/></let>
      </when>
```

Lisaks tegusõnade pöördelistele kategooriatele lisati reeglitesse ka reegel tegusõna käändeliste vormide tõlkimiseks. Eraldi reegleid vajavadki need esmalt juba kategooriate erinevuse tõttu.

Näites number 17 käsitletakse taaskord duaali, mis põhjasaami keeles esineb nii possessiivsufiksite, verbide ja personaalpronoomenite näol. Järgneva reegliga muudetakse taaskord duaal lihtsa teisendusega eesti keeles mitmuseks.

(17)
```
<when>
  <test><equal><clip pos="1" side="tl" part="a_nbr"/><lit-tag v="du"/></equal></test>
  <let><clip pos="1" side="tl" part="a_nbr"/><lit-tag v="pl"/></let>
</when>
```

3.3.3 Küsipartikkel *-go*

Soome keelele sarnaselt moodustatakse ka põhjasaami keeles *kas*-küsimusi liites tegusõnadele liitepartikkel. Põhjasaami keeles on selleks partikliks *-go*. Näites 18 välja toodud reegel muudab *-go* partikliga tegusõnad masinale loetavaks ja väljundiks peale käesoleva reegli rakendamist ongi ’*tegusõna*+*go*’. Lisaks tekib partiklile ’+*go*’ juurde märgend <qst>.

(18)
```
<when>
  <test><equal><clip pos="1" side="tl" part="a_qst"/><lit-tag v="qst"/></equal></test>
  <let><clip pos="1" side="tl" part="a_qst"/><concat><lit v="+go"/><lit-tag
v="qst"/></concat></let>
</when>
```

3.4 Tulemused

Eelnevates alapeatükkides mainitud reegleid arendati süsteemi näidisteksti põhjal, mille põhjasaamikeelne versioon on ära toodud lisas nr. 2 ning lisas nr. 3 on välja toodud ka näidisteksti eestikeelne versioon. Lisas nr. 4 on välja toodud põhjasaami keelest eesti keelde tõlgitud sama tekst, milleks on kasutatud käesoleva töö käigus arendatud süsteemi.

Süsteemi näidisteksti tõlkides on tulemus silmaga nähtav: teksti teema on juba arusaadav ning enamike lausete mõtteist saab samuti aimu. Siiski esineb tõlkimisel veel ka mitmeid vigu – korduvalt esinenud vead on välja toodud tabelis 1, kus selgitatakse ka

lähemalt, miks mingid kindlad vormid vigu võivad tekitada. Apertiumis märgitakse sõnastikust puuduvaid sõnu tärniga (*) ja vigaseid tõlkeid trellidega (#).

Tabel 1. **Tüüpvead**

Vea tüüp	**Näited**	**Selgitus**
Sõna puudub leksikonist	*olgobealde; *muhte	Sõna puudub kakskeelsest sõnastikust
Vale vaste leksikonis	Jah ta **kuulub** häält#	Automaatselt genereeritud leksikonis on verbi vaste, millel agent puudub
Mitu tõlget leksikonis	silmade/silmude	Leksikonis on sama algvormiga sõnad ja süsteem ei oska nende vahel valida
Lokatiiv	**#Toomas** on #väike #koerakene#; "#Rääkima **musse** kus ta on	Lokatiivi olemasolevad reeglid ei tööta õigesti
Partitiiv	Ta otsib **#Toomas#**	Olemasolev reegel, mis teisendab akusatiivi partitiiviks ei kata kõiki võimalikke partitiivi esinemise kohti eesti keeles
Kas-küsimus	On**#go**; Nägid**#go**	Süsteemis puudub reegel partikli *+go* tõlkimise kohta
Eitus	#Tema **#ei #saama**; **#ei #oskama** rääkida	Süsteemis puuduvad reeglid eituse tõlkimise kohta
Personaalpronoomen	**#tema** mängivad, ta kuulub **#tema** hästi#	Süsteemis puuduvad reeglid personaalpronoomenite tõlkimise kohta
Numeraalid	#tema **#kaks** lapsega	Süsteemis puuduvad reeglid numeraalide tõlkimise kohta
Täisminevik	#tema **on #nägema**	Süsteemis puuduvad reeglid mineviku tõlkimise kohta
Omadussõnad	**#suur #vana** puu; ta näeb **#väike** #käsi	Süsteemis puuduvad reeglid omadussõnade tõlkimise kohta

Ülaltoodud tabelist lähtuvalt saab tüüpvead jagada kolme suuremasse kategooriasse: 1. puudulikust leksikonist tingitud vead, 2. olemasolevatest reeglitest tingitud vead ja 3. reeglite puudumise tõttu esinevad vead.

Puudulikust leksikonist tingitud vead on üsna kergesti leitavad ja ka parandatavad. Nimelt, kui sõna puudub leksikonist, siis saab see automaatselt endale külge tärni (*). Ka valed vasted leksikonis on kergelt tekstist üles leitavad, sest tänu sellistele tõlgetele võibki tihti lause arusaamatuks jääda. Samuti jääb kohe silma see, kui süsteem ei oska valida mitme leksikoni tõlke vahel (*silmade/silmude*). Nagu eelnevalt mainitud, siis saab leksikoni puuduvad sõnad käsitsi sisestada, kuid see on ajamahukas töö ning kui üks tekst töötab korralikult ei tähenda see seda, et ka teised tõlgitavad tekstid sama sõnavaraga kasutaksid.

Teise kategooria vead ehk olemasolevatest reeglitest tingitud vead on antud töö kontekstis ehk keerulisimad vead, mida parandada. Erandiks on ehk *kas*-küsimuse reegli täiendamine, mida saaks teha üsna lihtsate vahenditega soome-eesti keelepaari eeskujul, kuid ka see nõuaks veidi rohkem süvenemist, kuna siis muutub ka süntaks. Lokatiivi ja genitiiv-akusatiivi reeglite täiendamine nõuaks ka suuremamahulist näidisteksti, et siis iga üksikjuhtum reeglina kirja panna.

Kolmanda kategooria vead ehk vead, mis tulenevad reeglitest, mida veel ei ole kirjutatud, on heaks teejuhiks tulevikus. Nimelt saab kategooriate kaupa edaspidi hakata süsteemi täiendama ja ülaltoodud tabelist näeb ära ka need kohad, kus oleks uusi reegleid enim vaja on.

Süsteemist on hetkeseisuga puudu reeglid kirjavahemärkide käsitlemiseks ja seega võib ka tulemis märgata palju veatähistusi ’#’. Nimelt lisatakse tähis hetkel igale kirjavahemärgile, kuigi tegelikult süsteem kirjavahemärkidega ei eksi.

Töö käigus valminud reeglite fail ning kakskeelne sõnastik ja ka uuendatud ’modes’ fail on kõik üles laetud Apertiumi repositooriumisse ehk andmehoidlasse. Konkreetne keelepaar ning eelnimetatud failid on kättesaadavad ja allalaetavad aadressil https://svn.code.sf.net/p/apertium/svn/incubator/apertium-sme-est/.

KOKKUVÕTE

Käesoleva töö põhieesmärgiks oli luua algne põhjasaami-eesti reeglipõhine tõlkesüsteem programmiga Apertium. Kitsamaks eesmärgiks oli luua nimetatud süsteem sellisel tasemel, et tõlgitud tekstist oleks võimalik aru saada. Nagu tööst selgus, siis täideti eesmärk: süsteemi näidisteksti tõlkest on võimalik väga hästi aru saada ning ka lausete mõtteist saab suures osas aru. Siiski, kui võtta natukene spetsiifilisem põhjasaamikeelne uudis ning seda antud programmi abil tõlkida, siis on juba eestikeelsest tõlkest arusaamine raskendatud.

Töö esimeses osas kirjeldati põhilisi erinevusi põhjasaami ja eesti keele vahel. Käesoleva töö teine osa andis ülevaate süsteemi Apertium toimimismehhanismidest ning tööpõhimõtetest. Lisaks on teises sisupeatükis kirjeldatud ka laiemat konteksti, kuhu antud töö paigutub põhjasaami keele töötlemise taustal.

Töö kolmandas peatükis kirjeldatakse süsteemi loomist nullist ning üheks käesoleva töö eesmärgiks oli ka praktilise töö dokumenteerimine. Esmalt kirjeldatakse põhjalikult põhjasaami-eesti kakskeelse sõnastiku automaatset genereerimist põhjasaami-soome ja soome-eesti kakskeelsete sõnastike põhjal. Seejärel kirjeldatakse transfeerireeglite kirjutamist nii käändsõnade, pöördsõnade kui ka *kas*-küsimuste näitel. Peatükis on näitena välja toodud ka mõned konkreetsed reeglid, mida on tõlkimisel rakendatud.

Kolmanda peatüki lõpus on välja toodud ka sagedasimad vead, mis esinesid süsteemi näidisteksti tõlkimisel. Vead on sisu ja põhjuse järgi klassifitseeritud kolme suuremasse kategooriasse: 1. puudulikust leksikonist tingitud vead, 2. olemasolevatest reeglitest tingitud vead ja 3. reeglite puudumise tõttu esinevad vead. Vigade kategoriseerimine aitab süsteemsemalt mõista, miks üks või teine asi väljundtekstis valesti on läinud.

Kahtlemata on käesoleva töö käigus ehitatud süsteemi võimalik üsna suures mahus edasi arendada. Tüüpvigade kategooriad loovad ka süsteemi arendamiseks hea platvormi, kust võib aimu saada sellest, milliseid reegleid esmajärjekorras arendada tuleks. Võimalusel tuleks arendada ka kakskeelset sõnastikku ning soovituslikult tuleks seda teha taas automaatselt genereerides, kuna käsitsi sõnastiku sisestamine on väga

ajamahukas ettevõtmine. Kui süsteemi näidistekst juba õigesti tõlgitud saab, siis võiks vaadata ka põhjasaamikeelsete uudiste tõlkimise suunas.

Tulevikus tuleks mõelda ka laiahaardelisema koostöö peale teadlastega, kes arendavad põhjasaami-soome tõlkesüsteemi ja soome-eesti tõlkesüsteemi: paljud probleemid, mis põhjasaami-eesti reeglipõhises masintõlkesüsteemis lahendust vajavad ning esile kerkivad, on sarnased eelmainitud keelepaaride arendamisel esilekerkivate probleemidega ja seega saab üheskoos neid efektiivsemalt lahendada.

KIRJANDUS

Antonsen, Lene, Trond Trosterud, Francis M. Tyers 2016. A North Saami to South Saami Machine Translation Prototype. Northern European Journal of Language Technology, 2016, Vol. 4, Article 2, pp 11–27.

Antonsen, Lene, Ciprian Gerstenberg, Maja Kappfjell, Sandra Nystø Rahka, Marja-Liisa Olthuis, Trond Trosterud, Francis M. Tyers 2017. Machine translation with North Saami as a pivot language. Proceedings of the 21st Nordic Conference of Computational Linguistics.

Baal, Berit Anne Bals, David Odden and Curt Rice 2012. An analysis of North Saami gradation. Phonology, 29, pp 165-212

EKK = Erelt, Mati, Tiiu Erelt, Kristiina Ross 2007. Eesti keele käsiraamat. Kolmas, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.

ENE = Värsijalg 2006. Eesti entsüklopeedia veebiversioon. http://entsyklopeedia.ee/artikkel/v%C3%A4rsijalg1 Kasutatud 19.04.2017

Ermakov, Natalia (Koost.) 2011. Uurali keelte sõnastik. Tallinn: Atlex.

Ethnologue (a). Estonian, Standard. https://www.ethnologue.com/language/ekk Kasutatud 05.03.2017.

Ethnologue (b). Sami, North. http://www.ethnologue.com/language/sme Kasutatud 03.03.2017.

Forcada, Mikel L., Boyan Ivanov Bonev, Sergio Ortiz Rojas, Juan Antonio Perez Ortiz, Gema Ramirez Sanchez, Felipe Sanchez Martinez, Carme Armentano-Oller, Marco A. Montava, Francis M. Tyers 2010. Documentation of the Open-Source Shallow-Transfer Machine Translation Platform *Apertium.* Departament de Llenguatges i Sistemes Informatics Universitat d'Alacant

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Francis M. Tyers 2011. Apertium: a free/open-source platform for rule-based machine translation. Springer Science+Business Media B.V. 2011

Janda, Laura A., Lene Antonsen 2016. The ongoing eclipse of possessive suffixes in North Saami. A case study in reduction of morphological complexity. UiT The Arctic University of Norway.

Johnson, Ryan, Tommi A. Pirinen, Tiina Puolakainen, Francis Tyers, Trond Trosterud, Kevin Unhammer 2017. North Sámi to Finnish rule-based machine translation system. Proceedings of the 21st Nordic Conference of Computational Linguistics.

Lagarda, A.-L., V. Alabau, F. Casacuberta, R. Silva, E. Díaz-de-Liaño 2009. "Statistical Post-Editing of a Rule-Based Machine Translation System". Proceedings of NAACL HLT 2009: Short Papers, pages 217–220, Boulder, Colorado. Association for Computational Linguistics

Marjomaa, Marko 2012. North Sami in Norway: An Overview of a Language in Context. Working Papers in in European Language Diversity 17. Toim. Johanna Laakso. Mainz, Germany.

Mettovaara, Jukka Pekka 2014a. Oanehis álggahus davvisámegillii. Ante Aikio ja Helmi Länsmani originaalteksti põhjal eesti keelde tõlkinud ja töödelnud Jukka Pekka Mettovaara. Loengukonspekt.

Mettovaara, Jukka Pekka 2014b. Kokkuvõte käänete funktsioonidest. Loengukonspekt.

Mettovaara, Jukka Pekka 2017. Kaugemad sugulaskeeled IV. Loengukonspekt. Kasutatud 2017 kevad.

Nickel, Klaus Peter 1994. Samisk grammatikk. 2. utgave. [Karasjok]: Davvi Girji.

Oahpa! Consonant gradiation. http://oahpa.no/sme/gramm/stadieveksling.eng.html Kasutatud 03.03.2017.

Oahpa! Nouns. http://oahpa.no/sme/gramm/substantiv.eng.html Kasutatud 17.02.2017

Oahpa! Syntax. http://oahpa.no/sme/gramm/syntaks.eng.html Kasutatud 05.03.2017

Oahpa! Verbs. http://oahpa.no/sme/gramm/verb.eng.html Kasutatud 27.03.2017

http://www.omniglot.com/writing/northernsami.htm

Sami information centre. Sapmi. http://samer.se/1192 Kasutatud 04.03.2017.

Sammallahti, Pekka 1998. The Sami Languages. An Introduction. Kárášjohka: Davvi Girji.

Trosterud, Trond, Kevin Brubeck Unhammer 2012. Evaluating North Sámi to Norwegian assimilation RBMT. University of Tromsø, Kaldera språkteknologi.

Tyers, Francis M., Linda Wiechetek, Trond Trosterud 2009. Developing Prototypes for Machine Translation between Two Sámi. Procceedings of the 13th Annual Conference of the EAMT.

Tyers, Francis M., Felipe Sanchez-Martinez, Mikel L. Forcada 2012. Flexible finite-state lexical selection for rule-based machine translation. Proceedings of the 16th EAMT Conference, 28-30 May 2012, Terento, Italy. Universitat d'Alacant.

Tyers, Francis M. 2015a. Session 0a: Introduction. Loengukonspekt. https://svn.code.sf.net/p/apertium/svn/branches/courses/2015-tartu/slides/ Kasutatud 04.04.2017

Vallaste, Heikki 2017. E-teatmik www.vallaste.ee Kasutatud 7.04.2

RULE-BASED SME-EST MACHINE TRANSLATION WITH PROGRAMME APERTIUM. SUMMARY

The objective of this Bachelor's thesis was to create Northern Sami-Estonian rule-based translation system with programme Apertium. The narrower objective was to create a system what can translate on such a level that the translated text can be understood. As described above, given paper has fulfilled the set goal: translation of the model text is understandable and the main ideas of particular sentences are understandable as well. However, if we take a news article with more specific language in Northern Saami and translate it with given programme, the comprehension of the result is more difficult.

In the first chapter of the paper, there was described the main differences between Northern Saami and Estonian language. The second chapter of this paper gave an overview of the mechanics of the operation of the system in Apertium. In addition, the second chapter is also describing the broader context in which the work is positioned in Northern Saami rule-based machine translation development.

The third chapter of the thesis describes the creation of a system from scratch and fulfils one objective of this work: documentation of the practical work. Firstly, in the chapter there is described in detail the automatic creation of Northern Saami-Estonian bilingual dictionary, what is based on Northern Saami-Finnish and Finnish-Estonian bilingual dictionaries. The chapter also covers the topic of written transfer rules for nouns, verbs and *if*-questions. There are also concrete examples of the transfer rules that are applied in the system.

At the end of the third chapter, there are also brought out the most frequent errors that occurred in the system while translating the model text. Errors are classified according to the contents and the cause - they are classified to three major categories: 1. insufficient lexicon 2. defects due to the existing rules and 3. errors what occur due the lack of rules. Categorizing the errors helps to understand more systematically why one or the other thing has gone wrong in the output text.

Undoubtedly, the system built during writing the thesis can be developed further on a large scale. Error categories are also creating good platform for the further development: the categories are showing which kind of rules should be a priority in development. If possible, the bilingual dictionary should also be developed and amended and recommendedly automatically generated. If the model text gets correct translation then there is a possibility to also try and translate more specific news in Northern Saami.

In the future, the developers of Northern Saami-Estonian machine translation should also think of a more comprehensive cooperation with the scientists who develop the Northern Sami-Finnish translation system and the Finnish-Estonian translation system: many problems Northern Saami-Estonian rule-based machine translation has, are similar to the above-mentioned language pairs' problems. Developing and emerging challenges together may result with more effective results.

LISA 1. SKRIPT PÕHJASAAMI-EESTI KEELEPAARI INSTALLEERIMISEKS APERTIUMIS

```
#!/bin/bash
# Viimati töötas korra sellisena
JOBNO=$(grep -c processor /proc/cpuinfo)
#--------------------------------------------------------
# Viimati töötas korra sellisena
# 17.03.16
Get_packages()
{
  echo == Some basic staff
  sudo apt-get -y install autoconf automake libtool libsaxonb-java python-pip
  sudo apt-get -y install python-lxml python-bs4 python-unittest2
  sudo apt-get -y install libxml-twig-perl antiword xsltproc
  sudo apt-get -y install poppler-utils wget python-feedparser subversion
  sudo apt-get -y install cmake
  sudo apt-get -y install python-tidylib python3-yaml libxml-libxml-perl
  sudo apt-get -y install libtext-brew-perl
  sudo apt-get -y install gawk flex

  # bison annab õige yacc'i !!!
  sudo apt-get -y install bison

  # paneb glib.h õigesse kohta
  #sudo apt-get -y install libglib2.0-dev

  # c/c++ kompilaator & tema sõbrad
  sudo apt-get -y install g++ libicu-dev subversion cmake libboost-dev build-essential
  sudo apt-get -y install libgoogle-perftools-dev
```

```
  # Vt http://apertium.projectjj.com/apt/howto.txt
  # Alternatiiv oleks ise SVNist tõmmata, kompileerida ja installida
  echo == Get_packages_from_Apertium_repo

  # TV-{{ stable releases
  # wget http://apertium.projectjj.com/apt/install-release.sh -O - | sudo bash
  # }}
  # nightly build - seda kasutame, aga mingis konteksti oleks release sobivam?
  wget http://apertium.projectjj.com/apt/install-nightly.sh -O - | sudo bash

  sudo apt-get -y update
  sudo apt-get -y dist-upgrade
  sudo apt-get -y install apertium-all-dev cg3ide
  # sudo apt-get install apertium-dev apertium-lex-tools cg3 hfst libhfst45-dev
}

# Viimati töötas korra sellisena
# pole testitud
Update_all_packages()
{
  echo == Update_all_packages
  wget http://apertium.projectjj.com/apt/install-nightly.sh -O - | sudo bash

  sudo apt-get update
  sudo apt-get dist-upgrade
}

#-----------------------------------------------------
```

```
# Viimati töötas korra sellisena
# 17.03.16
Checkout_the_source_code_from_Giellatekno_svn()
{
  echo == Checkout_the_source_code_from_Giellatekno_svn
  mkdir $HOME/giellatekno ; pushd $HOME/giellatekno
  svn co https://victorio.uit.no/langtech/trunk/langs/sme langs/sme
  svn co https://victorio.uit.no/langtech/trunk/experiment-langs/est experiment-langs/est
  svn co https://victorio.uit.no/langtech/trunk/giella-core giella-core
  svn co https://victorio.uit.no/langtech/trunk/giella-shared giella-shared
 #svn co https://victorio.uit.no/langtech/trunk/giella-templates giella-templates
  popd

  echo == Checkout_the_source_code_from_Apertium_svn
  mkdir $HOME/apertium ; pushd $HOME/apertium
  svn co https://svn.code.sf.net/p/apertium/svn/incubator/apertium-sme-est apertium-sme-est
  popd
}

# Viimati töötas korra sellisena
# pole testitud
Update_the_source_code_from_Giellatekno_svn()
{
  echo == Update_the_source_code_from_Giellatekno_svn
  echo See jupp kirjuta juurde...
  #svn update $HOME/giellatekno/core
  #svn update $HOME/giellatekno/giella-sme
  #svn update $HOME/giellatekno/giella-est
```

```
  #svn update $HOME/apertium/apertium-sme-est
}

#-----------------------------------------------------

# Viimati töötas korra sellisena
# 17.03.16
Build_Giellatekno_components()
{
  echo == Build_Giellatekno_core_components
  pushd $HOME/giellatekno/giella-core
  ./autogen.sh
  ./configure --disable-silent-rules --prefix=/usr/local
  sudo make install
  echo "export GTCORE=$HOME/giellatekno/giella-core" >> $HOME/.profile
  . ~/.profile
  popd

  echo == Build_Giellatekno_shared_components
  pushd $HOME/giellatekno/giella-shared
  ./autogen.sh
  ./configure --disable-silent-rules --prefix=/usr/local
  sudo make install
  echo "export GIELLA_SHARED=$HOME/giellatekno/giella-shared" >>
$HOME/.profile
  . ~/.profile
  popd

  echo == Build_est_lang
  pushd $HOME/giellatekno/experiment-langs/est
```

```
  ./autogen.sh -l
  . ~/.profile
  ./configure --with-hfst --without-xfst --enable-apertium
  make V=1 -j ${JOBNO}
  popd

  echo == Build_sme_lang
  pushd $HOME/giellatekno/langs/sme
  ./autogen.sh -l
  . ~/.profile
  ./configure --with-hfst --without-xfst --enable-apertium
  make V=1 -j ${JOBNO}
  popd
}

# Viimati töötas korra sellisena
# 17.03.16
Build_Apertium_components()
{
  echo == Build_Apertium_components
  pushd $HOME/apertium/apertium-sme-est
  ./autogen.sh \
          --with-lang1=$HOME/giellatekno/langs/sme/tools/mt/apertium \
              --with-lang2=$HOME/giellatekno/experiment-
langs/est/tools/mt/apertium \
          --prefix=/usr/local
  make -j ${JOBNO}
  popd
}
```

```
#-----------------------------------------------------

# Viimati töötas korra sellisena
# pole testitud
Setup4Virt_machine_from_Tino()
{
  Update_all_packages
  Checkout_the_source_code_from_Giellatekno_svn
  Build_Giellatekno_components
  Build_Apertium_components
}

# Viimati töötas korra sellisena
# pole testitud
Setup4clean_Ubuntu_1604()
{
  Get_packages
  Checkout_the_source_code_from_Giellatekno_svn
  Build_Giellatekno_components
  Build_Apertium_components
}

# Viimati töötas korra sellisena
# pole testitud
Update()
{
  Update_all_packages
  Update_the_source_code_from_Giellatekno_svn
  Build_Giellatekno_components
```

```
  Build_Apertium_components
}

#-----------------------------------------------------

#Setup4Virt_machine_from_Tino
#Setup4clean_Ubuntu_1604
#Update

PisiTest

sudo sh -c 'echo '$0' $(date) >> '$HOSTNAME'.log'
```

LISA 2: SÜSTEEMI NÄIDISTEKST PÕHJASAAMI KEELES

1: GOS DUOMMÁ LEA?

2: Duommá ja Máret leaba gárdimis. Lea buorre dálki odne, lea hui liekkas. Muhte ikte lei hui galmmas! Soai eaba sáhttán stoahkat dalle olgun. Duommá ja Máret liikoba stoahkat, soai stoahkaba álo ovttas gárdimis stuora dálu olgobealde.

3: Duommá lea unna gánddaš ja son lea guhtta jagi boaris. Unna nieiddaš lea su oappáš, son lea vihtta jagi boaris. Duommás lea unna beatnagaš, dat nai lea dál gárdimis. Beana liiko stoahkat dainna guvttiin mánáin. Beana lea hui ilolaš dál.

4: Leago Márehis maid beana? Ii, Márehis ii leat beana, sus lea bussá. Muhte bussá lea dálus, bussá oađđá.

5: Sudno eadni lea dálu siste bussáin ovttas, son geahččá láseráiggi olggos ja oaidná Duommá ja Máreha stoahkame. Duommá viehká johtilit stuora boares muora lusa, son čiehkáda Márehis. Dieđátgo manne? Máret čohkká ja sus leat gieđat čalmmiid ovddas. Son ii oainne maidege ja son lohká. Manne son dahká nu? Ja maid Duommá dahká muora luhtte?

6: Dat lea stoagus. Go Máret geargá lohkamis, son geahččá birra. Son ohcá Duommá: gosa dat manai? Oidnetgo su?

7: Máret ii dieđe gos Duommá lea. Son jearrá beatnagis: "Oidnetgo Duommá?". Muhte beana dieđusge ii máhte hállat! Máret ii oaččo vástádusa gažaldahkasis. Ii giige oaččo vástádusa, go hállet beatnagiiguin!

8: Máret geahččá eatnis láse duohken, su eadni čaibmá. Máret smiehttá, ahte dat lea oaidnán gosa Duommá lea mannan: "Muital munnje gos son lea!", son dadjá eadnásis. "Ii Máret, in sáhte muitalit dutnje!", son vástida. Vaikke son várra diehtá gos gánda lea, son ii siđa muitalit.

9: Máret vázzá hihtásit gárdima čađa. Son geahčča ain ohcat Duommá. Son geahčča beavddi vuolde ja stuoluid vuolde, muhte Duommá ii leat dieppe. Son geahčča juohke sajis, muhte ii son sáhte gávdnat Duommá.

10: De son gullá jiena, dat boahtá stuora boares muora duohken. Sáhtášiigo leat Duommá? Die dat jietna lea fas! Son guldala dárkilit. Ii leat loddi iige makkárge eará ealli. Dál son gullá dan bures. Ferte leat Duommá!

11: De son oaidná unna gieđaža nai ja go son boahtá lagabui, de son oaidná su oaivvi maid! Máret čaibmá ja dadjá: "Gávdnen du!". Soai goappašagat leaba ilus ja mannaba dállui, lea áigi borrat juoidá ja juhkat veaháš čázi!

LISA 3: SÜSTEEMI NÄIDISTEKST EESTI KEELES

1 KUS ON Jakob?

2 Jakob ja Mari on aias. Täna on väga ilus ilm, on väga soe. Aga eile oli väga külm! Nad ei saanud siis väljas mängida. Jakob ja Mari armastavad mängida, nad mängivad alati koos aias, mis asub suure maja ees.

3 Jakob on väike poiss ja ta on kuue-aastane. Väike tüdruk on tema õde, ta on viie-aastane. Jakobil on pisike koer, ka koer on aias. Koerale meeldib lastega mängida. Koer on praegu väga õnnelik.

4 Kas Maril on ka koer? Ei, Maril ei ole koera, tal on kass. Aga kass on majas, ta magab.

5 Nende ema on majas koos kassiga, ta vaatab aknast, kuidas Jakob ja Mari mängivad. Jakob jookseb kiiruga suure puu juurde, ta peidab end Mari eest. Kas sa tead miks? Mari istub ning hoiab oma käsi silmade ees. Ta ei näe midagi ja ta loetleb numbreid. Miks ta seda teeb? Ja mida teeb Jakob puu taga?

6 See on mäng. Kui Mari on numbrite loetlemise lõpetanud, vaatab ta ringi. Ta otsib Jakobi - kuhu ta läks? Kas sa nägid teda?

7 Mari ei tea, kus Jakob on. Ta küsib koeralt: " Kas sa nägid Jakobit?". Aga loomulikult ei oska koer rääkida! Seega ei saa Mari temalt ühtegi vastust oma küsimusele. Inimesed ei saa kunagi koertelt vastuseid!

8 Mari vaatab ema, kes on aknal, ta ema naerab. Mari arvab, et ta teab kuhu Jakob läks, "Ütle mulle, kuhu ta läks!" lausub ta emale. "Ei Mari, ma ei või sulle seda öelda!" vastab ta. Kuigi ta arvatavasti teab, kus Jakob on, ta ei taha seda öelda.

9 Mari jalutab aeglaselt läbi aia. Ta ikka veel üritab leida Jakobit. Ta vaatab laua alla ja toolide alla, aga Jakob ei ole seal. Ta otsib kõikjalt, kuid ta ei leia Jakobit.

10 Äkitselt kuuleb ta krabinat, see tuleb suure puu tagant. Kas see võiks olla Jakob? Jälle see krabin! Ta kuulatab tähelepanelikult. See ei ole lind ega muu loom. Ta kuuleb nüüd seda krabinat päris hästi. See peab olema Jakob!

11 Siis märkab ta väikest kätt ja kui ta jalutab lähemale, märkab ta ka Jakobi pead! Ta naerab ja ütleb: "Ma leidsin su!" Nad on mõlemad õnnelikud ja lähevad tuppa, sest on aeg midagi süüa ja juua natuke vett.

LISA 4: SÜSTEEMI NÄIDISTEKSTI TÕLGE PÕHJASAAMI-EESTI SUUNAL

*1#: KUS #TOOMAS ON#?

*2#: #Toomas ja #Mari on aias#. On hea ilm täna#, on õige soe#. *Muhte eile oli õige külm! #Tema #ei #saama mängida siis väljas#. #Toomas ja #Mari armastavad mängida#, #tema mängivad aina koos aias #suur maja *olgobealde#.

*3#: #Toomas on #väike poisike ja ta on #kuus aasta vana#. #Väike #tüdrukukene on ta õde#, ta on #viis aasta vana#. #Toomas on #väike #koerakene#, #tema *nai on nüüd aias#. Koer armastab mängida #tema #kaks lapsega#. Koer on õige rõõmus nüüd#.

*4#: On#go #Mari ka koer#? #Ei#, #Mari #ei #olema koer#, tas on kass#. *Muhte kass on majas#, kass magab#.

*5#: #Tema ema on maja #seest kassiga koos#, ta vaatab akent välja ja näeb #Toomas ja #Mari #mängima#. #Toomas jookseb ruttu #suur #vana puu #kokku#, ta peidab #Mari#. Tead#go #miks#? #Mari istub ja tal on käed silmade/silmude eest#. Ta #ei #nägema *maidege ja ta loeb#. Miks ta teeb nii#? Ja #meie #Toomas teeb puu #juures#?

*6#: #Tema on mäng#. Kui #Mari *geargá #õpingud#, ta vaatab ümber#. Ta otsib #Toomas#: kus #tema lahkus#? Nägid#go #tema#?

*7#: #Mari #ei #teadma kus #Toomas on#. Ta küsib koerast#: "Nägid#go #Toomas#?"#. *Muhte koer muidugi #ei #oskama rääkida! #Mari #ei #tohtima vastust #küsimus#. #Ei #teine #tohtima vastust#, kui räägivad koertega!

*8#: #Mari vaatab #ema #aken #taha#, ta ema naerab#. #Mari arvab#, et #tema on #nägema kus #Toomas on #lahkuma#: "#Rääkima musse kus ta on!"#, ta ütleb #ema#. "#Ei #Mari#, #ei #saama rääkida susse!"#, ta vastab#. Kuigi ta ehk teab kus poiss on#, ta #ei #soovima rääkida#.

*9#: #Mari kõnnib aeglaselt aia #läbi#. Ta katsetab üha otsida #Toomas#. Ta vaatab laua #alt ja toolide #alt#, *muhte #Toomas #ei #olema seal#. Ta vaatab #mis asendis#, *muhte #ei ta #saama kohtuda #Toomas#.

*10#: Jah ta kuulub häält#, #tema saab #suur #vana puu #taha#. #Saama#go olla #Toomas#? Seal #tema hääl on taas! Ta kuulab targalt#. #Ei #olema lind #ei *makkárge #teine loom#. Nüüd ta kuulub #tema hästi#. Peab olla #Toomas!

*11#: Jah ta näeb #väike #käsi *nai ja kui ta saab lähemale#, jah ta näeb ta pead #meie! #Mari naerab ja ütleb#: "Kohtusin #sina!"#. #Tema #mõlemad on õnnes ja lahkuvad majja#, on aeg süüa midagi ja juua vähe vett!#

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Käbi Suvi (04.06.1994),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

„PÕHJASAAMI-EESTI REEGLIPÕHINE MASINTÕLGE PROGRAMMIGA APERTIUM“,

mille juhendajad on Francis Morton Tyers ja Heiki-Jaan Kaalep,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 25.05.2017