TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Ida Maria Orula

Teisendus eesti keele vana ja uue kirjaviisi vahel lõplike muunduritega

Bakalaureusetöö (9 EAP)

Juhendaja: Heiki-Jaan Kaalep

Tartu 2019

Teisendus eesti keele vana ja uue kirjaviisi vahel lõplike muunduritega

Lühikokkuvõte:

Kaasajal on aktuaalne kõiksugu kirjalike allikate automatiseeritud analüüs. Analüüsiks kasutatavad infotehnoloogilised vahendid on aga edukalt rakendatavad üksnes sõnadele, mis on morfoloogilisel tasemel vastavuses tänapäevaste õigekirja normidega. Seega tuleb ajaloolisi tekste esmalt normaliseerida. Probleemile võib läheneda kahest suunast. Ühest küljest võib vanas kirjaviisis tekstid täielikult kaasaega tuua, asendades kõik vanapärased sõnavormid nende kaasaegsete vastetega. Nii on tekstid kaasaegsetele automaatanalüüsitehnoloogiatele arusaadavad ning ka inimestele kergesti loetavad, kuid kaduma läheb oluline info kunagise keelekasutuse kohta. Teine võimalus on muuta olemasolevaid keelt analüüsivaid vahendeid selliselt, et need oskaks ära tunda ka vanas kirjaviisis sõnavorme. Bakalaureusetöös kasutatakse mõlemat lähenemist. Võttes aluseks 1739. aasta Piibli teksti, kirjutatakse uus ortograafiamuundur, mis vastendab vanas kirjaviisis sõnu nende tänapäevaste kujudega. Lisaks kohandatakse olemasolevat eesti keele morfoloogiamuundurit vana kirjakeelega, et oleks võimalik säilitada sõnade tollane kuju, seostades neid siiski kaasaegsete sõnavormidega. Töös antakse ka näpunäiteid, kuidas alustatud lahendust tulevikus edasi arendama peaks.

Võtmesõnad:

Lõplikud muundurid, vana kirjaviis, morfoloogia.

CERCS: P175 Informaatika, süsteemiteooria

Mapping Between Old and New Estonian Orthography Using Finite State Transducers

Abstract:

Nowadays it is common to analyse all kinds of written sources automatically. However, the necessary technologies are only applicable to words that follow the morphological rules of the modern language. Therefore, it is necessary to normalize historical texts that are written using the old Estonian orthography. This problem may approached from two different angles. On the one hand, it is possible to convert all old Estonian orthography forms to their modern counterparts. This would make the texts easy to understand for both the automated analysis

technologies and also for the people who are not so familiar with the old Estonian orthography. However, valuable information about how the language has changed, would be lost. The second approach is to adapt the current technologies to make them recognize the old word forms. In this thesis, both solutions are used. The author creates a new orthographic transducer that maps old word forms from the 1739 Bible translation to their modern forms. In addition, an existing morphological analyser of the Estonian language is modified, to allow it to recognize old Estonian orthography word forms. The author also gives suggestions for future developments of the created system.

Keywords:

Finite state transducers, Estonian old ortography, morphology.

CERCS: P175 Informatics, systems theory

Sisukord

Sissejuhatus

Keel kui inimkonna suhtlusvahend on algusaegadest peale olnud pidevas arengus ja muutumises. Ajaloolise keelekasutuse uurimine aitab paremini mõista ka kaasaegset keelt - seega lihtsustab eesti keele vana kirjaviisi analüüs nii praeguste kui ka tulevaste ajaloolaste ning keeleteadlaste tööd. Tänapäeval on väga levinud tekstide analüüsimine infotehnoloogiliste lahendustega, kuid kaasaja keele jaoks loodud vahendid ei ole otseselt rakendatavad vanas kirjaviisis tekstidele (Pilvik, et al., 2019). Automatiseeritud analüüsi üheks eelduseks on, et sõnad oleks morfoloogilisel tasemel vastavuses tänapäevaste õigekirjanormidega. Seetõttu on vaja ajaloolisi tekste enne edasist analüüsi normaliseerida. Näiteks 19. sajandi vallakohtuprotokollide normaliseerimist on käsitletud 2019. aasta Eesti Rakenduslingvistika Ühingu aastaraamatus ilmunud artiklis „Möistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine“ (Pilvik, et al., 2019).

Bakalaureusetöö peamine eesmärk on alustada lõpliku muunduri koostamist, mis võimaldaks teisendada sõnu eesti keele vanast kirjaviisist uude. Koostatava muunduri aluseks on 1739. aasta Piibli tõlge. Kuna vana kirjaviis on väga paljude erandite ja erijoontega, ei ole töö bakalaureusetöö tulemus veel piisav vanas kirjaviisis tekstide laiaulatuslikuks normaliseerimiseks, kuid tehtud töö näitab kätte suuna ja annab ette aluspõhja, mida tulevikus edasi arendada. Töös kirjeldatud lähenemist saab kasutada ka muude keeles ilmnevate variatsioonide puhul - näiteks murdetekstide või slängil, lühenditel ja võõrlaenudel põhineva internetisuhtlusest tulnud keelekasutuse normaliseerimiseks. Internetikeele normaliseerimist on kirjeldatud näiteks artiklis Heiki-Jaan Kaalepi, Kadri Muischneki ja Raul Sireli artiklis „Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile“ (Muischnek, Kaalep, & Sirel, Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile, 2011).

Lisaks tutvustatakse töös ka juba olemasolevat eesti keele morfoloogiat analüüsivat muundurit ning kohandatakse seda vanas kirjaviisis tekstide analüüsimise jaoks sobivamaks. Seega on töö teiseks eesmärgiks näidata, et nimetatud morfoloogiamuundurit ei pea kasutama üksnes musta

kasti (ingl *black box*) põhimõttel, vaid seda saab kohandada vastavalt lahendatavale ülesandele ja hetkevajadustele.

Töö on jaotatud 3 peatükiks. Esimeses peatükis selgitatakse, mis on lõplikud muundurid ja tutvustatakse konkreetses töös kasutatavat muundurite koostamise tehnoloogiat ja süntaksit. Lisaks tutvustatakse ka olemasolevat eesti keele morfoloogia muundurit. Antakse lühike ülevaade selle struktuurist ja kasutusvõimalustest ning kirjeldatakse, kuidas aitab see koos töö käigus valminud muunduriga vana kirjakeelt analüüsida. Kirjeldatakse ka eesti keele vana kirjaviisi ja tutvustatakse vana kirjakeele korpust.

Teises peatükis kirjeldatakse esmalt, kust leiti reeglite kirjutamisel aluseks võetud seaduspärad. Kirjeldatakse eeltöötlust, mida tehti 1739. a Piibli tekstiga, et seal teisendusreeglitele aluseks ja selgituseks näitesõnu leida. Selgitatakse ka, miks on keele uurimise seisukohast oluline ja mõistlik, et osad teisendused viidi sisse just eesti keele muundurisse, selle asemel, et neid kõiki reeglite abil uues muunduris väljendada. Selgitatakse reeglite kirjutamise protsessi ning tuuakse näiteid kirjutatud reeglitest. Seejärel kirjeldatakse eesti keele muunduris tehtud modifikatsioone ning tuuakse ka nende kohta selgitavaid näiteid.

Kolmandas peatükis kirjeldatakse muundurite testandmeid ja testimise protsessi. Esitatakse testimise tulemused ja analüüsitakse, mida testimise käigus saadud info põhjal muundurite juures muuta tuleks.

Esimeses lisas on toodud Tarmo Vaino skript eesti keele morfoloogiamuunduri ülesseadmiseks. Teises lisas on töö käigus valminud muundurireeglid tervikuna koos näitesõnadega. Kolmandas lisas on erinevate töös kasutatud muundurite väljundite võrdlemiseks esitatud katkend 1739. aasta Piiblist algsel kujul ja iga muunduriga teisendatult.

1. Vana kirjaviis ja lõplikud muundurid

Käesolev peatükk viib lugeja lähemalt kurssi töö taustaga. Esimeses alapeatükis kirjeldatakse eesti keele vana kirjaviisi. Tuuakse välja vana kirjaviisi põhilised erinevused kaasajal kehtivatest kirjakeele reeglitest ning tutvustatakse ka vana kirjakeele korpust, kus vanu kirjalikke tekste tänapäeval hõlpsasti leida võib. Teises alapeatükis selgitatakse, mis on lõplikud muundurid ning selgitatakse, miks nende kasutamine kõnealuse probleemi lahendamisel efektiivne on. Tutvustatakse konkreetses töös kasutatavate muundurite koostamise tehnoloogiat ja reeglite süntaksit. Kirjeldatakse, mil viisil on võimalik muundureid sõnade vastendamiseks kasutada – tuuakse näiteid käsurea käskudest ja nende väljunditest. Lisaks tutvustatakse olemasolevat eesti keelt analüüsivat muundurit. Kirjeldatakse selle struktuuri ja kasutusvõimalusi ning selgitatakse, mis on selle muunduri funktsioon bakalaureusetöös.

1.1 Vana kirjaviis ja vana kirjakeele korpus

Kirjaviis on „Eesti keele käsiraamatus“ (Erelt, Erelt, & Ross, Eesti keele käsiraamat: Ortograafia, 2007) defineeritud kui „ajalooliselt kujunenud õigekirjatava“. Eesti keeles on aegade jooksul kasutusel olnud kolm kirjaviisi: ladina ja alamsaksa keelte ortograafiatel põhinev korrapäratu kirjaviis, ülemsaksa ortograafial põhinev vana kirjaviis ning soome ortograafial põhinev uus kirjaviis, mis on aluseks ka kaasaegsele eesti kirjakeelele (Erelt, Erelt, & Ross, Eesti keele käsiraamat: Ortograafia, 2007).

Järgnevalt antakse vana kirjaviisi eripäradest kokkuvõtlik ülevaade Tartu Ülikooli dotsendi Valve-Liivi Kingissepa artikli „Eesti keele esimestest kirjapanekutest ja kirjaviisidest“ (Kingissepp, 2001) abil. Kuna korrapäratu kirjaviis mõjus võõrapäraselt ja kirjakeel jäi eestlastele arusaadavast keelest kaugeks, võtsid Bengt Gottfried Forselius ning Johann Hornung ette kirjaviisi ühtlustamise ja lihtsustamise, pannes sellega aluse vanale kirjaviisile. Kirjakeele rahvale arusaadavamaks muutmiseks soovitas Forselius loobuda võõrtähtedest (*c, f, q, y, x, z*), ning võttis kasutusele eraldi tähemärgid täpitähtede tähistamiseks. Kindel süsteem oli olemas vokaalide pikkuste märkimiseks. Pikka vokaali märgiti lahtises silbis ühe- ning kinnises silbis

kahekordsena (nt „saama“, „uus“, „saatko“). Endiselt kasutati ka korrapäratu kirjaviisi ajal juurdunud tava, mille kohaselt märgiti rõhulise silbi vokaali pikkust talle järgneva konsonandi korduste arvuga. Konsonantide pikkusest oli seega kirjapildis raske aru saada, kuna näiteks sõna „warras“ võis tähendada nii „varras“ kui ka „varas“. Ühesilbiliste sõnade lõpus olevaid konsonante kirjutati ühe tähega (nt „lukk“ asemel „luk“). Oluline erinevus kaasaja kirjakeelega võrreldes on ka *õ*-hääliku puudumine. *Õ*-häälik võeti kasutusele alles uues kirjaviisis Otto Wilhelm Masingu soovitusel. Vana kirjaviis jäi püsima kuni 19. sajandi esimese pooleni ning selles on kirjutatud ka 1715. aasta Uus Testament ja 1739. aasta Piibel.

Lisaks morfoloogiale erineb tollane keelekuju kaasaegsest ka näiteks lauseehituse tõttu – olgu põhjuseks siis mõne võõrkeele mõju või lihtsalt tänapäevasest keelekasutusest puuduv väljend. Arnold Kask on oma teoses „Eesti kirjakeele ajaloost I“ kirjutanud: „On ilmne, et tolleaegsed kirjamehed mõtlesid saksa keeles, tahtsid aga kirjutada eesti keeles, asendades saksakeelsed sõnad vastavate eestikeelsetega.“ (Kask, 1970, lk 9). Seega ei ole alati võimalik vana ja uue kirjakeele lauseid kohakuti tõsta ning sõnahaaval seostada. Hea näide keelekasutuse erinevusest on siinkohal lausekatkend 1790. aastast pärinevas Nicolaus von Hagemeisteri tekstist „Lühhikenne öppetus ma-rahwale” (VAKK, 2013) „...kes muido wissiste surma sisse olleksid iänud...”. On selge, et väljendit „surma sisse jääma” tänapäeval enam ei kasutata ja seega algteksti ning selle kaasaegse ümberkirjutuse vahel otseseid paralleele tõmmata ei õnnestu.

Vanas kirjaviisis kirjutatud tekstidega saab kõige mugavamalt tutvuda vana kirjakeele korpuses, mis on kättesaadav järgmisel lingil: http://vakk.ut.ee/. Korpusena mõistetakse „Eesti märksõnastiku“ (Eesti Rahvusraamatukogu, Tartu Ülikooli Raamatukogu, 2019) järgi tänapäeval elektrooniliselt hoiustatavat ning töödeldavat struktureeritud kindlat liiki tekstikogumit. Eesti vana kirjakeele korpusesse on koondatud kõik 15. ja 16. sajandist teadaolevalt säilinud eestikeelsed trüki- ja käsikirjatekstid, enamik 17. sajandist säilinud trükitud tekste ning valik 18. ja 19. sajandi trükitekste (Prillop, Vana kirjakeele korpus: Avaleht, 2013). Kokku on korpuses 1 736 240 eestikeelset sõne ja tekstid on osaliselt märgendatud, lihtsustamaks neist kaasaegses eesti keeles info leidmist ja tekstide sisu mõistmist (Prillop, Vana kirjakeele korpus: Tekstid, 2013). Korpuse märgendamine kujutab endast korpusesse kuuluvatele tekstidele täiendava informatsiooni lisamist. Lisatud informatsioon võib esitada näiteks tekstis leiduvate sõnade morfoloogilist, süntaktilist või semantilist analüüsi või ka kirjeldada teksti ülesehitust, märkides ära lausepiirid, tabelid, pealkirjad ja muud seesugused teksti komponendid (Muischnek, Keelekorpused – sama mitmekesised, 2015).

Järgmises alapeatükis tutvustatakse lõplikke muundureid ning selgitatakse, kuidas neid keele morfoloogilise ja ortograafilise analüüsi juures kasutada saab.

1.2 Lõplikud muundurid

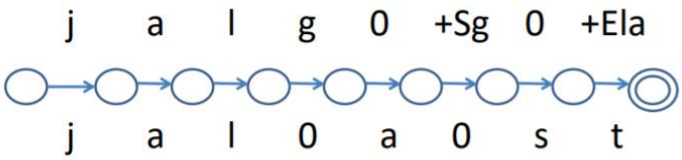
Lõplikku muundurit on Kenneth R. Beesley'i ja Lauri Karttuneni raamatus „Finite State Morphology" (Beesley & Karttunen, 2003) kirjeldatud kui abstraktset automaati, mis loob relatsiooni kahe regulaarse keele vahel. Tavalised lõplikud automaadid tunnevad ära ühte konkreetsesse regulaarsesse keelde kuuluvaid sõnesid. Lõplikud muundurid aktsepteerivad aga kahte erinevat regulaarset keelt ning suudavad lisaks keelte äratundmisele ka neisse kuuluvaid sõnesid omavahel vastendada. Seega vastab igale ühest keelest tulnud sõnele üks või mitu teise keelde kuuluvat sõne ja vastupidi.

Jaak Pruulmann-Vengerfeldt on oma magistritöös kirjeldanud lõplikel muunduritel põhinevat eesti keele morfoloogiasüsteemi ja toonud välja asjaolu, et selle näol pole tegemist mitte programmiga, vaid pigem kergesti mõistetaval kujul kirjapandud abstraktse keelekirjeldusega (Pruulmann-Vengerfeldt, 2010, lk 15). Seetõttu on lõplike muundurite koostamine ja kasutamine kasutajasõbralikum ka keeleteadlase taustaga inimeste jaoks, kellel erinevate programmide ja algoritmidega vähem kokkupuuteid on. See on oluline, kuna keelemudelite koostamisel on oluline analüüsitavat keelt sügavamalt mõista. Üksnes IT-taustaga inimene võib luua valesid seoseid ja teatud aspekte kas liigselt lihtsustada või vastupidi ebavajalikult keeruliselt kirja panna.

Pruulmann-Vengerfeldti töös (Pruulmann-Vengerfeldt, 2010, lk 15) kirjeldatakse, et lõplikel muunduritel morfoloogiamudel on regulaarne relatsioon pind- ja sõnastikuesituse vahel, kus pindesituse hulka kuuluvad kõik eesti keeles kasutatavad sõnakujud ning sõnastikuesituse hulka sõnade tüved ja konkreetse sõnavormi grammatiline info. Näiteks sõnavormi „aastaid" võib muundurite süsteem tõlgendada nii käändevormina sõnast „aasta" kui ka sõnade „aas" ja „tai" liitsõnana. Selleks, et saada analüüsi tulemusena vaid lihtsõnu, tuleb muunduritesse eraldi piiranguid kirjutada.

Sõnede vastendamise selgitavaks näiteks esitatakse siinkohal joonis Heiki-Jaan Kaalepi loenguslaididelt (Kaalep, Morfoloogiline analüüs: lõplikud muundurid, 2017). Joonisel on

näha, kuidas vastendab muundur sõnatüve „jalg“ ning grammatilise info (sõnastikuesitus) sellele keeles kasutusel oleva sõnakujuga (pindesitus).



Joonis 1. Näide lõplikust muundurist

Tulenevalt lõplike muundurite võimekusest erinevate sõnede vahel seoseid luua, on muundurid sobivad ka **v**anas kirjaviisis tekstide normaliseerimisel. Vanas kirjaviisis kirjutati mitmeid sõnu uue kirjaviisiga sarnaselt, kuid siiski mitte päris identselt – just nendest erinevustest võimaldavadki lõplike muunduritena kirjapandud reeglid üle saada, ühendades omavahel sõnade vana ja uue kirjapildi.

1.2.1 Lõplikud muundurid bakalaureusetöös

Töö kirjutamisel on põhiliseks eeskujuks võetud Helsingi ülikooli arvutilingvistika professori Kimmo Mati Koskenniemi ning Pirkko Kuutti töö „Indexing Old Literary Finnish text” (Koskenniemi & Kuutti, 2017), milles teisendatakse sarnaste meetoditega kaasaegsesse kirjaviisi 1642. aasta soomekeelse Piibli teksti. Lisaks on lõplike muundurite reeglite kirjutamisel abiks Kenneth R. Beesley'i ja Lauri Karttuneni raamat „Finite State Morphology” (Beesley & Karttunen, 2003).

Järgnevalt tutvustatakse bakalaureusetöös kasutatud ja loodud muundurite koostamise loogikat ja süntaksit. Muunduri koostamiseks vajalikud teisendusreeglid kirjutati *.xfscript* laiendiga faili, millest tehti Linuxi käsurealt kasutatava *hfst-xfst* kaudu *hfst*-tüüpi[1] muundur. Siinkohal antakse ülevaade reeglites kasutatud süntaksist:

[1] *Helsinki Finite-State transducer* on Helsingi ülikoolis arendatud tehnoloogia kaalutud ja kaalumata lõplike muundurite koostamiseks. http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml

- $\rightarrow$ tähistab kindlat vastendamist. Selle sümboliga määratud teisendused toimuvad alati. Näiteks $a \rightarrow b$ puhul asendatakse iga sümbol a alati sümboliga b.
- $(\rightarrow)$ tähistab valikulist vastendamist. Selle sümboliga määratud teisendus võib toimuda või mitte toimuda. Näiteks $a\ (\rightarrow) b$ puhul võib sümbol a asenduda sümboliga b või jääda samaks.
- _ tähistab teisenduse konteksti määrates vaatluse all olevat sümbolit.
- || eraldab teisenduse ja selle konteksti. Näiteks $[a \rightarrow b \,||\, c \_ d]$ väljendab, et a asendub sümboliga b iga kord, kui sümbol jääb c ja d vahele.
- .#. märgib sõna algust ja lõppu. Näiteks võime väljendada, et a-le vastab b iga kord, kui sümbol on sõna viimane täht (ehk sellele järgneb sõna lõpp) $[a \rightarrow b \,||\, \_ .\#.]$
- $::\ x$ kus x on hetkel suvalise täisarvu tähenduses, tähistab teisendusele määratud kaalu. Näiteks $[a \rightarrow b :: 2]$ puhul on iga sümboli a asendamine sümboliga b kaaluga 2. Kaaludest räägitakse lähemalt järgmises lõigus.
- ! tähistab kommentaari.
- $define$ tähistab teisendusreegli defineerimist. Näiteks $define\ a2b\ [a \rightarrow b]$; loob teisendusreegli *a2b*, mis asendab iga *a* sümboliga *b*.
- $.o.$ tähistab kompositsiooni, mille abil saab kokku panna erinevaid teisendusreegleid/muundureid. Näiteks kui luua juurde $define\ e2f\ [e \rightarrow f]$; ning teha seejärel $define\ kokku\ a2b\ .o.\ e2f$;, on tulemuseks *kokku*, mis vastendab *a b*-ga ning *e f*-ga.
- $regex$ abil määratakse, millised teisendusreeglid muundurisse kuuluvad. Näiteks $regex\ kokku$; puhul pannakse muundur kokku reeglitest *a2b* ja *e2f*.
- ; tuleb kirjutada iga kirjutatud reegli lõppu.
- {} vahele saab kirjutada mitmest sümbolist koosnevaid sümboleid. Näiteks $\{ae\} \rightarrow \{ee\}$ vastendaks iga ühendi *ae* ühendiga *ee*.

Iga muundur pakub teisendusreeglite abil sisendsõnale sobivaid väljundsõnesid. Kuna kõigi teisendusreeglite kaal on vaikimisi 0, siis on esialgu ka kõigi pakutud väljundsõnede kaal 0. Kui on teada, et mõnede teisendusreegli muutused on tõenäolisemad kui teiste reeglite omad, siis tasub lisada reeglitele kaalud. Iga väljundsõna kaaluks saab temani jõudmiseks kasutatud teisendusreeglite kaalude summa. Mida väiksem on väljundsõna kaal, seda tõenäolisemalt on tegemist õige variandiga, seega pakutakse väikseima kaaluga sõnu eespool. Kui ka iga sümboli vahetamise kaaluks määrata 1, liiguvad rohkem teisendusi läbinud sõnad võimalike väljundite

hulgas tahapoole. Tihti on juba sellisest kaalumisest abi, kuid kui osata teisenduste tõenäosusi ligikaudu hinnata, saab täpsemate kaalude määramisega väljundite hulka veelgi kohandada.

Kaalusid võib ühe konkreetse teisenduse toimumisele või mittetoimumisele määrata mitmel viisil. Kui vaja on kaal määrata üksnes teisenduse toimumisele, saab selle kirja panna valikulise vastendamisega. $[a\ (\rightarrow)\ b :: 1]$; väljendab, et a asendumine b-ga on kaaluga 1 ning samaks jäämine vaikimisi kaaluga 0. Pikemalt võib selle kirja panna nii: $[a \rightarrow a,\ a \rightarrow b :: 1]$;. Bakalaureusetöös on eelistatud just teist kirjapanekuviisi, kuna see võimaldab hõlpsamini mõlema variandi kaale jälgida ja vajadusel muuta.

Muunduri koostamisest ja kasutamisest tuuakse siinkohal väike näide. Joonisel 2 on näha võimalik *.xfscript* faili sisu. Joonisel 3 on näidatud failist Linuxi käsureal muunduri tegemine. Joonisel 4 on toodud näide muunduri kasutamisest. Võimalikust üheksast väljundist on esitatud neli esimest. Nende puhul tuleb selgelt välja, kuidas on rakendunud teisendustele määratud kaalud. Esimese variandi kaal on 0, kuna teisendusi ei tehtud. Teise ja kolmanda variandi kaal on 2, kuna kummaski asendati üks *e* sümbol *a*-ga ning neljanda kaal on 3, kuna *e* asendati *a*-ga ning *a* omakorda *aa*-ga.

```
! e asendumine a-ga on kaaluga 1, samaks jäämine kaaluga 0
define reegel1 [e (->) a::1] ;
! a asendumine aaga on kaaluga 2, samaks jäämine kaaluga 1
define reegel2 [a -> a::1, a -> {aa}::2 ;
regex reegel1 .o. reegel2 ;
```

Joonis 2. Näide võimalikust *.xfscript* faili sisust

```
$ hfst-xfst
hfst[0]: source näide.xfscript
  Defined
  ’reegel1’
  Defined
  ’reegel2’
? bytes. 2 states, 9 arcs, ? paths
hfst[1]: save näitemuundur.hfst
```

Joonis 3. Näide muunduri kompileerimisest

```
$ echo ’tere’ | hfst-lookup näitemuundur.hfst
> tere      tere  0,000000
tere        tare  2,000000
tere        tera  2,000000
tere        taare 3,000000
```

Joonis 4. Näide muunduri kasutamisest väljundsõnade leidmiseks

Kirjeldatud põhimõtete järgi pandi töös kokku vana kirjaviisi ortograafiat kaasaja kirjaviisile vastavaks teisendav muundur, mida nimetatakse töös edaspidi ortograafiamuunduriks. Järgnevalt tutvustatakse lühidalt juba olemasolevat eesti keelt analüüsivat muundurit, millele ortograafiamuundur töö käigus ka juurde liidetakse.

1.2.2 Eesti keele morfoloogiamuunduri tutvustus

Eesti keeles on väga palju võimalusi sõnadest nii tuletiste kui ka liitsõnade moodustamiseks ja seetõttu on erinevaid sõnavorme liiga palju, et neid kõiki eraldi sõnastikuna kirja panna – selleks et otsustada, kas etteantud sõna on korrektne eestikeelne sõna, on abi sõnamoodustusreeglitest (Pruulmann-Vengerfeldt, 2010, lk 5). See tõsiasi toetab teooriat, et eesti keele morfoloogilist analüüsi on mõistlik rakendada lõplike muundurite põhiselt.

Artiklis „Estonian Morphology in the Giella Infrastructure“ (Kaalep, Moshagen, & Trosterund, Estonian Morphology in the Giella Infrastructure, 2018) kirjeldatakse eesti keele morfoloogilise analüsaatori koostamist Tromsø Ülikoolis arendatud Giella infrastruktuuris[2]. Nimetatud muunduritel põhineva morfoanalüsaatori[3] koostamise ühe ajendina on välja toodud, et kuigi eesti keele jaoks on olemas vabavaraline morfoloogiline analüsaator Filosoft[4], võib eeldada, et selle muutmine ei ole kuigi kasutajasõbralik, kuna viimasel ajal pole koodile edasiarendusi tehtud. Seega otsustati esitada eesti keele morfoloogia kirjeldus muunduritena *lexc*, *twol* ja *xfst* failidena. Giella infrastruktuur on kindlaksmääratud struktuuriga kataloogipuu, kuhu on

[2] https://victorio.uit.no/langtech/trunk/
[3] https://victorio.uit.no/langtech/trunk/experiment-langs/est/
[4] https://github.com/Filosoft/vabamorf, http://www.eki.ee/tarkvara/, https://github. com/jjpp/plamk

võimalik lisada teavet erinevate keelte kohta. Olemas on palju konkreetsest keelest sõltumatuid toiminguid, mis võimaldavad iga lisatud keele andmete põhjal vastava keele muundureid koostada. Konkreetselt eesti keelt puudutav osa Giella infrastruktuuris hõlmab endas leksikoni, kus on kümneid tuhandeid liht- ja liitsõnu ning tuletisi.

Artiklis on näiteks on toodud ühe sõnastikumuunduri (ingl *lexical transducer*) ja morfofonoloogilise muunduri kompositsioon, mis vastendab sõna „kott“ omastava käände vormiga „koti“. Siinkohal esitatakse vastav joonis:

```
Lexicon transducer:            k o t t +N +Sg +Gen
                               k o t t  >  i  +WeakGrade
Morphophonological transducer: k o t t  >  i  +WeakGrade
                               k o t 0  0  i  0
```

Joonis 5. Sõnastiku- morfofonoloogilise muunduri kompositsioon (Kaalep, Moshagen, & Trosterund, Estonian Morphology in the Giella Infrastructure, 2018)

Sõnastikumuundurite on kirjeldatud *lexc* failides ning morfofonoloogilised muundurid *twolc* failides (Kaalep, Moshagen, & Trosterund, Estonian Morphology in the Giella Infrastructure, 2018).

Järgnevalt tutvustatakse eesti keele morfoloogiamuunduri komponente ja omadusi, mis bakalaureusetöö kontekstis olulised on. Kõik töö käigus muudetud failid asuvad kaustas *experiment-langs/est/src/morphology*. Seal on eraldi kaustad liidete (*affixes*) ja sõnatüvede (*stems*) jaoks, milles kummaski on omakorda *lexc* failid erinevate sõnatüüpide ja nähtuste kirjeldamiseks. Näiteks on *stems* kaustas failid *adverbs.lexc* määrsõnade, *nouns.lexc* nimisõnade ja *verbs.lexc* tegusõnade jaoks. Failis *nouns.lexs* on toodud eestikeelsed nimisõnad koos infoga nende käänamise kohta ning neile määratud kaaluga. Kaalud tuletati sagedussõnastiku[5] põhjal nii, et kaaluklass on seda suurem, mida väiksem on sõna sagedus, ning sõnadele, mida sagedussõnastikus ei leidu, omistati maksimaalne sagedussõnastikust tuletatud kaal (H-J. Kaalep eravestluses, 2019). Näiteks sõna „hõige“ on failis esitatud sellisena: $hõige + N$: $hˇõiK2e$ $PINGE$ " weight: 9 " ;. Sümbol *K2* abil märgitakse, et seal võib toimuda häälikumuutus (nimetavas käändes „hõi**g**e“, omastavas „hõi**k**e“). Märksõna *PINGE* tähistab, et

[5] https://www.cl.ut.ee/ressursid/sagedused1/failid/lemma_kahanevas.txt

käänamine toimub sarnaselt sõnaga „pinge“ ning *weight: 9* abil on märgitud sõna kaal. Määrsõnade, tegusõnade ja muude sõnaliikidega on vastavates failides toimitud sarnaselt.

Kausta *affixes* puhul on bakalaureusetöö kontekstis olulisim fail *verbs.lexc,* kus on esitatud info tegusõnade erinevate pöörete ja aegade vormide kohta. Sarnaselt muutuvaid sõnu käsitletakse koos. Siinkohal tuuakse näide leksikonist, mis vastab sõnale „minema“.

```
LEXICON MINEMA    ! only minema

                      :min%{dbl%}%> A_INFINITIVE ;       ! minna
                      :min%{dbl%} IMPERS_A ;     ! minnakse, mindud
                      :mine SUPINE_V ;           ! minema
                      :min%{W%} IMPERATIVE ;     ! mingu
                      :läi%{W%} NU ;             ! läinud
                      :läk%>s IND_PAST ;         ! läks, läksin etc
            +Pers+Prs:lähe%{W%} IND_PRS ;        ! lähen, läheb, ...
            +Pers+Prs:lähe%{W%}%> CONDITIONAL ;  ! läheks, ...
   @R.Part.One@@P.Part.Bad@ MINEMA2 ;

LEXICON MINEMA2
   +Pers+Prs+Imprt+Sg2:mine%{W%}%> GI ;          ! mine
+Pers+Prs+Ind+Pl1+Aff%+Use%/Rare:läh%{W%}%>me GI ;  ! lähme more colloquial...
```

Joonis 6. Väljavõte kausta *affixes* failist *verbs.lexc*

Joonisel 6 on näha, et failis on väljendatud verbi erinevaid pöördevorme. Erinevad vormid ja ajad (nt *A_INFINITIVE, IND_PAST)* on samuti esitatud eraldi leksikonidena.

Üks oluline omadus eesti keele morfoloogiamuunduri juures, mis ka bakalaureusetöös esile kerkib, on liitsõnade moodustamise piiramine. Eesti keeles võivad liitsõnade moodustamisel osaleda vaid teatud tüüpi sõnad. Kui teostada liitsõnade moodustamine üksnes erinevate lihtsõnamuundurite korrutamisega, kasvaks muundur liiga suureks. Lõplikel muunduritel ei ole mälu – muunduri puhul ei ole üheski olekus teada, millisest olekust ta sinna parajasti jõudnud on. Liitsõnade moodustamisel oleks aga vaja meeles pidada, mis liiki komponente sõnasse juba lisatud on ja kui palju neid komponente kokku on. See probleem on lahendatud lipudiakriitikutega (ingl *flag diacritics*). (Kaalep, Moshagen, & Trosterund, Estonian Morphology in the Giella Infrastructure, 2018)

Lipudiakriitikud on olemuslikult tõeväärtuse tüüpi muutujad, mis võimaldavad muunduri töö ajal keelatud teid blokeerida, tagades sellega muunduri väiksemad mõõtmed. Liitsõnade moodustamiseks on lihtsaim viis korrutada lihtsõnamuundurit iseendaga piiramatul arvul kordi,

kuid ainult selle loogika kasutamine viiks lõpmatu tsüklini ning piiramatu arvu liitsõnakomponentideni – seega tuleb piirata tsükli läbimiste arvu. Lihtsõnamuundurile võib juurde lisada lipudiakriitiku, mille väärtus teatud olekute läbimisel muutub ja jõuab lõpuks väärtuseni, mis blokeerib järgmise võimaliku tee ning lõpetab tsükli. (Kaalep, Moshagen, & Trosterund, Estonian Morphology in the Giella Infrastructure, 2018)

Teoses „Finite State Morphology“ on tutvustatud erinevaid lipudiakriitikuid. Järgnevalt näidatakse siin kahte tüüpi lipudiakriitikuid, millega ka bakalaureusetöös kokku puututi (vt peatükk 2.4). P-tüüpi lipudiakriitikud (ingl *P* ehk *Positive (Re)Setting*) esitatakse kujul $@P.tunnus.väärtus@$ ning diakriitiku peale sattudes määratakse näidatud tunnuse väärtuseks diakriitikus väljendatud väärtus (Beesley & Karttunen, 2003, p 455). Eesti keele morfoloogiamuunduris nõutakse $@P.Part.Bad@$ abil failis *morphology/affixes/verbs.lexc*, et sõnale ei tohi enam edasisi sõnakomponente järgneda. R-tüüpi lipudiakriitiku (ingl *R* ehk *Require Test*) peale sattudes kontrollitakse, kas tunnuse väärtus on parajasti võrdne diakriitikuga määratud väärtusega ning kui ei ole, siis see tee blokeeritakse (Beesley & Karttunen, , p 456). *Verbs.lexc* failis nõutakse $@R.Part.One@$ abil, et tüvi oleks vaadeldavas sõnas esimesel kohal. Bakalaureusetöö peatükis 2.4 tuleb lipudiakriitikuid muuta, võimaldamaks vanas kirjaviisis levinud liitsõnade moodustamise struktuuri.

Morfoloogiamuundurit saab kasutada nii selliselt, et sisendsõna kohta väljendatakse sisemine info grammatilise struktuuri kohta, kui ka selliselt, et väljendataks üksnes sõna korrektne vorm koos selle kaaluga. Erinevus tuleb hästi välja joonisel 7.

```
$ echo ’minu’ | hfst-lookup morfoloogiamuundur_analüüsiga.hfst
> minu      mina+Pron+Sg+Gen+Emph   2,000000

$ echo ’minu’ | hfst-lookup morfoloogiamuundur_analüüsita.hfst
> minu      minu  2,000000
```

Joonis 7. Morfoloogiamuunduri variandid

Edaspidi nimetatakse töös autori koostatud muundurit ortograafiamuunduriks, äsjakirjeldatud eesti keele morfoloogiat analüüsivat muundurit morfoloogiamuunduriks ning nende kompositsiooni teel kokkupanekul saadud muundurit kirjaviisimuunduriks. Skript eesti keele morfoloogiamuunduri ülesseadmiseks on leitav lisadest (lisa 1). Morfoloogia- ja

kirjaviisimuunduritest on eri juhtudel kasutusel versioon, mis sõna puhul ka analüüsi väljendab või versioon, mis analüüsi ei väljenda. Järgnevas peatükis kirjeldatakse vana kirjaviisi teisendamiseks vajalikke reegleid ortograafiamuunduris ning modifikatsioone morfoloogiamuunduris.

2. Seaduspärade leidmine ja muundurireeglite kirjutamine

Siin peatükis keskendutakse bakalaureusetöö praktilisele osale. Esimeses alapeatükis kirjeldatakse 1739. aasta Piibli (kättesaadav järgnevalt aadressilt: https://www.eki.ee/piibel/index.php) tekstidest teisendusreeglite aluseks ja selgituseks sobivate näitesõnade leidmist. Teises alapeatükis selgitatakse lühidalt, milliseid eri tüüpi muutusi võib vana ja uue kirjaviisi puhul täheldada ning miks on keele uurimise seisukohast õigustatud ja oluline, et osad muutused pandi kirja teisendusreeglitena uues ortograafiamuunduris ning teised viidi sisse hoopis olemasolevasse morfoloogiamuundurisse. Kolmandas alapeatükis tutvustatakse reeglite kirjutamise protsessi ning tuuakse näiteid kirjutatud reeglitest. Olulisemate reeglite puhul on veidi pikemalt lahti seletatud ka reegli lõpliku kujuni jõudmise protsess, kuna see aitab ka reeglite olemust paremini mõista. Neljandas alapeatükis kirjeldatakse eesti keele muunduris tehtud modifikatsioone ning tuuakse ka nende kohta selgitavaid näiteid.

2.1 Piibliteksti algtöötlus sobivate näitesõnade leidmiseks

Kuigi vana kirjaviisi eripärasid on kirjeldatud mitmes allikas, on tihtipeale välja toodud vaid mõned erijooned ning näitesõnu on üldjuhul vähe. Kimmo Koskenniemi ja Pirkko Kuutti võtsid oma töös „Indexing Old Literary Finnish text“ reeglite koostamise aluseks hulga sõnu, mis esinesid korpuse tekstides vähemalt kuus korda (Koskenniemi & Kuutti, 2017). Bakalaureusetöö autor pidas suurema pildi nägemiseks ja rohkemate näitesõnade leidmiseks vajalikuks ka Piibli sõnavarast ülevaatliku listi koostamist. Saadud list järjestati sõnade esinemissageduse alusel, eeldades, et sagedaminiesinevad sõnad väljendavad ka põhilisi morfoloogilisi eripärasid, võimaldades seeläbi katta võimalikult suure osa Piibli sõnavarast.

Tekstide töötluseks ja näitesõnade leidmiseks kasutati eelkõige loomuliku keele töötluseks mõeldud teekide kogumit NLTK (Bird, Loper, & Klein, 2009) ning selle eesti keelele keskendunud versiooni EstNLTK-d (Orasmaa, Petmanson, Tkachenko, Laur, & Kaalep, 2016).

Esiteks laeti alla kogu 1739. aasta Piibli tekst ning salvestati eraldi tekstifaili. Saadud fail vaadati ridahaaval läbi ning igast reast eemaldati numbrid ja kirjavahemärgid. Allesjäänud sõned lisati üldisesse sõnade järjendisse. Kokku leiti selle töötluse tulemusel 679630 sõna. Saadud järjendi põhjal loodi NLTK teegi võimalusi kasutades sagedussõnastik (ingl. *FreqDist),* mis koosneb ennikutest, kus igale sõnale on vastavusse seatud selle esinemiste arv analüüsitavas järjendis. Sagedussõnastiku pikkust vaadates selgus, et piiblitekstides oli kokku 29807 erinevat sõnavormi.

Sõnu vaadates selgus, et paljude sõnade kirjapilt ei ole 1739. aastaga võrreldes muutunud – need sõnad ei olnud reeglite loomiseks vajalike seoste otsimise algfaasis vajalikud. Sõnad jagati kahte eraldi listi: esiteks sõnad, mille kirjapilt ka tänapäeval grammatiliselt korrektne on, ja teiseks sõnad, mida kaasajal teistmoodi kirjutatakse. Jagamiseks kasutati EstNLTK teegi meetodit *spellcheck*, mis võtab argumendiks sõne või sõnede listi ning tagastab listi sõnastikest. Igas sõnastikus on sellele vastava sõna kohta näidatud kolm tunnust:

1. õigekirjakontrolli tulemus tõeväärtusena (tõene, kui sõna on kaasaja normide kohaselt grammatiliselt korrektne, ning väär vastasel korral);
2. list, mis korrektse sõna puhul on tühi, kuid vigase puhul sisaldab soovitusi korrektsetest sõnadest, mis vigase kirjapildiga kõige sarnasemad on;
3. analüüsitav kirjapilt.

Näiteks sõna „tere“ puhul oleks väljund järgmine: [{’spelling’: True, ’suggestions’: [], ’text’: ’tere’}]. Sõna „terre“ puhul aga „[{’spelling’: False, ’suggestions’: [’tere’, ’tarre’, ’tetre’, ’tedre’], ’text’: ’terre’}]“. Sõnad, mille õigekirjakontrolli tõeväärtus oli tõene, jäeti konteksti tarbeks eraldi listis alles. Vanas kirjaviisis sõnade listi jäi alles 23536 sõna.

Edasi lühendati analüüsitavate sõnade listi sellega, et sorteeriti välja sõnad, mille muutumise reeglid olid juba teada. Esmalt asendati kõik sõnades esinevad „w“ tähed „v“ tähtedega. Seejärel eemaldati listist sõnad, millele *spellcheck* pakkus soovituseks kirjapilti, kus topeltkonsonandid oleks asendatud ühekordsetega. Näiteks polnud nende sammude tagajärjel listis enam sõna „wanna“, sest esimese muutuse tagajärjel sai sõna kirjapildiks „vanna“ ning selle kohta sisaldas *spellcheck*’i soovituste list ka sõna „vana“. Tulemuseks saadi 20642 sõna. Sõnade arv oli muidugi endiselt liiga suur, et sõnade hulka tervikuna hoomata, kuid kui kõige sagedamini esinevate ja juba teadaolevate seaduspärade tõttu muutuvad sõnad eemaldada, oli teiste reeglite märkamine juba mugavam.

Loodud näidissõnade listidest oli kasu , ning üritati leida viisi muudatuse toimumisele kindlama konteksti määramiseks. Näiteks kerkis paljudes näitesõnades esile seaduspära, et sõnades, kus tänapäeval kirjutatakse lõppu „u“, oli vanas kirjaviisis sõna viimaseks täheks „o“. Seda võib jälgida näiteks sõnade „wasto“ („vastu“), „minno“ („minu“), „paljo“ („palju“) ja „armo“ („armu“) puhul. Enne vastava reegli kirjapanekut kontrolliti, kui suur osakaal on sõnadel, mis lõppevad „o“-ga nii vanas kui ka uues kirjaviisis. Siinkohal oli abi eelpoolmainitud järjendist, kuhu korjati kokku sõnad, mille kirjapilt ei ole aja jooksul muutunud. 6271 sõna peale oli selliseid sõnu 23. Enamik neist olid nõrgeneva laadimuutusega sõnade omastava käände vormid. „Eesti keele käsiraamatu“ põhjal on laadimuutus astmemuutuse vorm, mille puhul on sõnad üksteisest erinevad „s“-i või sulghääliku olemasolu poolest (Erelt, Erelt, & Ross, Eesti keele käsiraamat: Morfoloogia, 2007). Konkreetsemalt on tegu kaoga ehk nõrgeneva laadimuutuse alaliigiga, mille puhul tugeva astme „s“ või sulghäälik sõnast lihtsalt ära jäetakse (Erelt, Erelt, & Ross, Eesti keele käsiraamat: Morfoloogia, 2007). Seega kuuluvad kirjeldatud sõnade hulka näiteks „tegu“ („teo“), „nägu“ („näo“) ja muud seesugused. Kuna selliste sõnade osakaal kõigi sõnade hulgas on üpris väike, siis ei peetud seda takistuseks antud seaduspära põhjal valikulise vastandamisega reegli defineerimisel.

2.2 Vajalike muudatuste liigitus

Muutusi, mida vanas kirjaviisis sõnades tegema peab, et neid automatiseeritult analüüsida, on mitmesuguseid. Mõnel juhul on erinevus vaid üksikus tähes: näiteks kirjutati *v* asemel vanas kirjaviisis alati *w* ning *õ* asemel on 1739. aasta Piiblis kasutusel *o* või *ö* (Kask, 1970). Lisaks esineb aga ka olukordi, kus mõne sõna kasutus on aegade jooksul lihtsalt muutunud. Heaks näiteks on siinkohal sõna „pöörama“, mis kuulus tollases keelekasutuses samasse muuttüüpi sõnaga „naerma“ (H-J. Kaalep eravestluses, 2019). Seega öeldi näiteks „pöörake“ asemel „pöörge“.

Teoreetiliselt on küll võimalik sõna muuta ka üksnes kirjapildi tasemel. Töö käigus valminud ortograafiamuunduris on juba olemas reegel, mis võimaldab vokaalide vahele jääva *g* asemel *k* kirjutada (reeglitest on lähemalt juttu järgmises alapeatükis). Juurde tuleks lisada veel reegel, mis paigutaks *r* ja *g* vahele *a*. Nii tunneks morfoloogiamuundur sõna küll ära, kuid sealjuures läheks kaduma oluline info tollase keelekasutuse suhtes. Sisuliselt ei ole siin tegemist ortograafilise muutusega vaid sõna muuttüübi muutusega. Lisaks tuleks siis kirjutada

teisendusreeglid ka kõigi teiste „pöörma“ pöördevormide kohta, mis tänapäevasest erinevad. Näiteks „pöörnud“ ja „pöörvat“. Mõistlikum lahendus on modifitseerida olemasolevat morfoloogiamuundurit selliselt, et aktsepteeritav oleks ka sõna „pöörama“ vana kirjaviisi aegne kasutus.

Seetõttu koosneski bakalaureusetöö praktiline pool kahest osast: ortograafiamuunduri loomisest ning morfoloogiamuunduri kohandamisest. Selleks, et ortograafiamuundur oleks võimalikult hästi kasutatav ka eraldiseisvana, jäeti ka sinna väljakommenteeritult sisse ka mõned levinumaid morfoloogilisi muutusi kajastavad teisendusreeglid. Selline lähenemine võimaldab muundurite kasutajal valida, mis tema jaoks antud hetkel oluline on: kas soovitakse üksnes lihtsustada ja kaasajastada mõnda vanas kirjaviisis kirjutatud teksti või on oluline saada infot ka teksti ajaloolise keelekasutuse kohta. Esimesel juhul piisab ortograafiamuunduri kasutamisest ning näiteks sõna „pöörnud“ asemel pakutakse kasutajale lihtsalt sõna „pööranud“. Morfoloogiamuundur üksi oskab kasutajale küll öelda, et sõna „pöörnud“ puhul on tegemist on verbi „pöörma“ ühe vormiga, kuid kui sõna on esitatud teisendamata ortograafiaga kujul (nt „poörma“ või „pöorma“), on tulemuseks üksnes, et sõna ei tunta ära. Täpseima tulemuse saamiseks tuleb seega kasutada ortograafiamuunduri ja morfoloogiamuunduri kokkupanekul saadud kirjaviisimuundurit tervikuna, kuna see arvestab ühteaegu nii ortograafiliste kui ka morfoloogiliste muutustega.

Järgevalt kirjeldatakse ortograafiamuunduri jaoks reeglite kirjutamist ning morfoloogiamuundurisse tehtud modifikatsioone. Alapeatükis 2.4 tuuakse ka näiteid nimetatud muundurite eraldi ja koos kasutamisest.

2.3 Ortograafiamuunduri teisendusreeglid

Põhiliste seaduspärade kohta, mille jaoks reeglit vaja on, saadi informatsiooni Arnold Kase teosest „Eesti kirjakeele ajaloost I“ (Kask, 1970). Igale seaduspärale otsiti kinnitust ka peatükis 2.1 kirjeldatud Piibli sõnade listist. Lisaks väljendavad reeglid ka muudatusi, millele Kase teoses tähelepanu polnud pööratud, kuid mis näitesõnu analüüsides siiski selgelt välja paistsid.

Iga kirjutatud reegli juures tuuakse ära ka näitesõnad, mis antud reegli alusel muutuvad. Enamik näitesõnu sobivad iseloomustama mitut erinevat teisendusreeglit, kuid kirjelduses keskendutakse iga sõna puhul üksnes hetkel vaatluse all olevale reeglile. Näiteks sõna „minno“ ehk „minu“ puhul tuleb asendada kahekordne konsonant ühekordsega ning lisaks asendada täht *o* tähega *u*. Kui sõna tuuakse näiteks häälikupikkuste kontekstis, siis jäetakse sõna lõpuhääliku muutus ajutiselt tähelepanuta.

Reeglitele kaalude määramisel lähtuti suhtelisest hinnangust sellele, kui tõenäoline või sagedane mõni konkreetne muutus paistis olevat. Nagu peatükis 1.2 kirjeldatud, on muunduri väljundsõne seda tõenäolisem, mida väiksem on tema kaal. Sellest lähtuvalt jäeti kõige tõenäolisematele teisendusele kaal määramata ning vähemesinevatele anti kaaluks 1 või 2. Seega mida haruldasem on mõni kirjeldatud teisendus, seda väiksema tõenäosusega on seda konkreetses sõnas vaja. Lisaks saab iga sõna seda suurema kaalu, mida rohkem muudatusi temas tehakse. Näiteks sõna „teäte“ ehk „teate“ puhul võib ortograafiamuundur eri reeglite tõttu pakkuda tulemusteks nii „teate“, „teati“ kui ka „teadi“. Kirjeldatud loogika alusel pakutaks varianti „teadi“ kõige viimasena, kuna selle kuju saamiseks oli vaja teha rohkem muudatusi (*ä* asemel *a* ja *t* asemel *d*). Erinevate muutuste esinemissagedust aitas hästi kontrollida Piibli sõnadest koostatud järjend.

Enne konkreetsete teisendusreeglite kirjutamist defineeriti reeglid ühe- ja kahekordsete vokaalide ja konsonantide väljendamiseks (*VOK, KONS, topeltVOK, topeltKONS*). Järgnevalt esitatakse vokaalide märkimiseks kirjutatud reeglid. Konsonantidega toimiti sarnaselt.

$$define\ VOK\ \ [a \mid e \mid i \mid o \mid u \mid \text{õ} \mid \text{ä} \mid \text{ö} \mid \text{ü}\ ]\ ; \tag{1}$$

$$define\ topeltVOK\ \ [\{aa\} \mid \{ee\} \mid \{ii\} \mid \{oo\} \mid \{uu\} \mid \{\text{õõ}\} \mid \{\text{ää}\} \mid \{\text{öö}\} \mid \{\text{üü}\}]\ ; \tag{2}$$

Sellisel kujul defineeritud vokaalid ja konsonandid aitasid edasiste reeglite kirjutamisel muutuste konteksti määrata – mõned häälikumuutused leiavad aset üksnes peale konsonanti, teised üksnes vokaalide vahel ja nii edasi. Tänu vokaale ja konsonante väljendavatele reeglitele ei pidanud kõiki tähti iga säärase reegli juures uuesti kirja panema, mis muudab reeglid loetavamaks.

2.3.1 Häälikupikkuste väljendamine

Esimene oluline kategooria tegeleb häälikupikkuste muutmisega. Nagu alapeatükis 1.1 mainitud, erines vanas kirjaviisis ühe- ja kahekordsete vokaalide ja konsonantide kasutamise loogika suuresti kaasaja kirjaviisist. Sellesse kategooriasse kuuluvad reeglid kolme tüüpi muutuste väljendamiseks:

1. kahekordne konsonant tuleb asendada ühekordsega (nt „emmale“, „minno“, „waggasid“, „pallutakse“ ehk „emale“, „minu“, „vagasid“, „palutakse“);
2. ühekordne konsonant tuleb asendada kahekordsega (nt „peatük“, „kül“, „kät“, „wiskümend“ ehk „peatükk“, „küll“, „kätt“, „viiskümmend“);
3. ühekordne vokaal tuleb asendada kahekordsega (nt „se“, „ramato“, „job“, „sago“ ehk „see“, „raamatu“, „joob“, „saagu“).

Kahekordse konsonandi asendamiseks ühekordsega katsetati mitmeid lähenemisi. Esmalt kirjeldati, et kui sõnas on *VOK* reegliga määratud sümboli järel järjest kaks *KONS* reegliga määratud sümbolit, siis võib teine neist asenduda nullsümboliga. Reegel näeks välja selline:

$$define\ eemaldaKONS\ \ [\ KONS\ (\rightarrow)\ 0\ ||\ VOK\ KONS\ \_\ ]\ ; \tag{3}$$

Sellisel kujul reegel ei ole aga piisavalt täpne, kuna hakkab teist konsonanti eemaldama ka erinevate konsonantide ühendi puhul. Nii pakkus muundur reegli tulemusel näiteks sõna „rahwas“ vasteks ka sõnu „ratas“ ja „rabas“. Seetõttu katsetati uut lähenemist. Ühe reegliga (4) lisati sõnasse iga *topeltKONS* abil määratud topeltkonsonandi ja vokaali vahele tähekombinatsioon, mida eestikeelsetes sõnades muidu ei esine. Seejärel kustutati nimetatud tähekombinatsiooni eest ära üks sellele eelnev konsonant (5) ning viimase reegliga (6) kustutati tähekombinatsioon ise. Seejärel pandi reeglid kokku (7).

$$define\ tähista\ [0\ (\rightarrow)\{asdf\}\ ||\ topeltKONS\ \_\ VOK]; \tag{4}$$
$$define\ eemalda\ [KONS \rightarrow\ 0\ ||\ \_\{asdf\}]; \tag{5}$$
$$define\ puhasta\ [\{asdf\} \rightarrow 0]; \tag{6}$$
$$define\ eemaldaKONS\ tähista\ .o.\ eemalda\ .o.\ puhasta\ ; \tag{7}$$

Selline lahendus töötas, kuid oli üpris aeganõudev. Seetõttu otsustati reeglis eraldi lahti kirjutada, mis tähega iga kahekordne konsonant asenduda võib. Osade häälikute puhul on ka tänapäevases kirjaviisis topeltkonsonandi kirjutamine tavapärane (näiteks hääliku *l* puhul sõnad

„tulen“ ja „tullakse“) ning osade puhul välistatud (näiteks häälik *b*). Seda tuli väljendada ka reegli kirjapanekul. Seega kirjeldati näiteks hääliku *b* jaoks ainult ühte võimalikku vastendust (väljendamaks kahekordse konsonandi asendumist ühekordsega) ning hääliku *l* jaoks kahte (kahekordne *l* võib asenduda ühekordsega või jääda kahekordseks). Kui häälikul võivad esineda mõlemad variandid, määrati kahekordseks jäämisele kaaluks 1, kuna rohkemates sõnades on siiski vaja ühekordset häälikut. Reegel (8) sai kirja järgmiselt:

$$define\ eemaldaKONS\ [\{bb\} \rightarrow\ b,\ \{dd\} \rightarrow\ d,\ \{ff\} \rightarrow\ f,\ \{gg\} \rightarrow\ g,\ \{hh\} \rightarrow\ h,\ \{jj\} \rightarrow\ j,\ \{jj\} \rightarrow\ \{jj\}::1,\ \{kk\} \rightarrow\ k,\ \{kk\} \rightarrow\ \{kk\}::1,\ \{ll\} \rightarrow\ l,\ \{ll\} \rightarrow\ \{ll\}::1,\ \{mm\} \rightarrow\ m,\ \{mm\} \rightarrow\ \{mm\}::1,\ \{nn\} \rightarrow\ n,\ \{nn\} \rightarrow \{nn\}::1,\ \{pp\} \rightarrow\ p,\ \{pp\} \rightarrow \{pp\}::1,\ \{rr\} \rightarrow\ r,\ \{rr\} \rightarrow\ \{rr\}::1,\ \{ss\} \rightarrow\ s,\ \{ss\} \rightarrow\ \{ss\}::1,\ \{tt\} \rightarrow\ t,\ \{tt\} \rightarrow\ \{tt\}::1,\ \{vv\} \rightarrow\ v,\ \{ww\} -> \ w]\ ; \quad (8)$$

Teist tüüpi teisendusreegli kirjutamisel otsustati esialgu arvestada üksnes juhtudega, kus kahekordistamist vajav konsonant asub sõna lõpus, kuna sõna keskel oli sellised olukorrad pigem erandlikud. Näiteks kirjapilti „kümmend“ esines Piiblis 1161 ning „kümend“ vaid 110 korral. Kindlama konteksti määramisega saavutatud väiksemat väljundite arvu peeti antud juhul olulisemaks erandjuhtudega arvestamisest. Lisaks ei kahekordistata reegliga nõrku sulghäälikuid ega ka näiteks *f* häälikut, kuna eestikeelsete sõnade puhul ei ole selline kirjakuju tõenäoline. Piibli sõnade listist kaashäälikuga lõppevaid sõnu uurides selgus, et enamike häälikute puhul on tavalisem olukord, kus kahekordistamist tegema ei pea. Erandiks oli häälik *p*, mille puhul olid sagedasemad just kahekordset konsonanti nõudvad sõnad (nt „sep“ ehk „sepp“ ja „noletup“ ehk „nooletupp“). Vastavalt leitud sagedustele määrati ka teisenduste kaalud. Reegel (9) sai kirja selliselt:

$$define\ lisaKONS\ [k \rightarrow\ \{kk\}::1,\ k \rightarrow\ k,\ l \rightarrow\ \{ll\}::1,\ l \rightarrow\ l,\ m \rightarrow\ \{mm\}::1,\ m \rightarrow\ m,\ n \rightarrow\ \{nn\}::1,\ n \rightarrow\ n,\ p \rightarrow\ \{pp\},\ p \rightarrow\ p::1,\ s \rightarrow\ \{ss\}::1,\ s \rightarrow\ s,\ t \rightarrow\ \{tt\}::1,\ t \rightarrow\ t\ ||\ VOK\ \_\ .\ \#.]\ ; \quad (9)$$

Ühekordse vokaali kahekordistamise reegel (10) kirjutati sama loogika alusel. Kuna see muutus toimub valdavalt vaid sõna esisilbis, kirjutati see piirang kontekstina juurde. Täpsemalt nõutakse reeglis, et kirjeldatud teisendus toimib üksnes juhul, kui vokaalile eelnevad sõna algus ja konsonant ning sellele järgneb kas konsonant või sõna lõpp (nt sõna „se“ ehk „see“ puhul). Nii vähendati taaskord pakutavate väljundite hulka. Näiteks sõna „ramato“ (ehk „raamatu“) puhul on kontekstiga reegli tulemuseks üksnes „ramato“ ja „raamato“. Konteksti määramata

pakutaks ka sõnu „raamaatoo“, „ramaato“ ja muud seesugust. Lisaks ei toimu teisendust näiteks sõna „taewas“ („taevas“) puhul, kuna *a* häälikule ei järgne konsonant. Sõnu, kus esisilbi vokaali kahekordistama ei pea, on märgatavalt rohkem – seetõttu määrati kahekordistavatele vastendamistele kaaluks 1.

$$define\ lisaVOK\ [a \rightarrow \{aa\}::1,\ a \rightarrow a,\ e \rightarrow \{ee\}::1,\ e \rightarrow e,\ i \rightarrow \{ii\}::1,\ i \rightarrow i,\ o \rightarrow \{oo\}::1,\ o \rightarrow o,\ u \rightarrow \{uu\}::1,\ u \rightarrow u,\ \text{õ} \rightarrow \{\text{õõ}\}::1,\ \text{õ} \rightarrow \text{õ},\ \text{ä} \rightarrow \{\text{ää}\}::1,\ \text{ä} \rightarrow \text{ä},\ \text{ö} \rightarrow \{\text{öö}\}::1,\ \text{ö} \rightarrow \text{ö},\ \text{ü} \rightarrow \{\text{üü}\}::1,\ \text{ü} \rightarrow \text{ü}\ ||\ .\#.KONS\_[KONS\ |\ .\#.]]\ ; \quad (10)$$

Nende kolme teisendusreegliga said kaetud eripärad, mida vana kirjaviisi juures kõige levinumalt rõhutatakse. Ometi ei ole need veel piisavad vanas kirjaviisis kirjutatud tekstide normaliseerimiseks.

2.3.2 Muud ortograafilised ja morfoloogilised muutused

Teise kategooria moodustavad reeglid selliste ortograafiliste muutuste kohta, mille puhul on kas kindlalt teada, et nad esinevad eranditult alati (*w* asemel on alati *v*), või mille esinemise puhul on võimalik määrata kindla konteksti. Kolmandasse kategooriasse kuuluvad sellised reeglid, mille aluseks olevad muutused on näitesõnadest silma jäänud, kuid mille esinemisele ei suudetud leida kindlat seaduspära või konteksti. Selle kategooria muutuste põhjalikum uurimine ja reeglite täpsustamine on kindlasti oluline samm, mida töö edasiarendamiseks tegema peab. Eriti need reeglid, mille puhul kindlat konteksti määrata ei õnnestunud, võivad mõnikord liiga agaralt töötada ning tuua sellega kaasa väga suure väljundsõnade hulga. See, millisesse bakalaureusetöös määratud kategooriasse üks või teine reegel kuulub, on kindlasti vaieldav ja tööd edasi arendades tuleb selles tõenäoliselt muudatusi teha, kuid praegune jaotus töötas töö autori meelest muundureid jooksvalt katsetades kõige paremini. Kõigi ortograafiamuundurireeglitega saab tutvuda lisades (lisa 2). Siinkohal tuuakse näidetena välja eelkõige sellised reeglid, mis vastavad otseselt või kaudselt „Eesti kirjakeele ajaloost I“ teoses kirjeldatud erijoontele.

Kõige lihtsam reegel, mis ortograafiamuunduri koostamiseks kirjutada tuli, on tingimata *w* asendumine *v*-ga (11). Kuna vanas kirjaviisis kirjutati *v* märkimiseks alati häälikut *w*, ei olnud reeglile vaja määrata mingit konteksti ega kaalu.

$$define\ w2v\ [w \rightarrow v\ ]; \quad (11)$$

Arnold Kase teosest Hornungi grammatika kohta lugedes jäid silma mitmed erisused, mis ka Piibli sõnade listis esindatud olid. Nimelt on Hornungi grammatikale iseloomulikud näiteks *de*-tunnuseline mitmuse omastav (nt „Jummalade“) ja paralleelvormid *nud*-kesksõnadest (nt „piddanud“ ja „piddand“), mitmuse kolmanda pöörde lihtminevikust („läksid“ ja „läksivad“) ja *ta*- ning *da*-liitelistest tegusõnadest („kustutatud“ ja „kustotud“) (Kask, 1970, lk 70). Piiblis vastasid nendele eripäradele näiteks järgnevad sõnad: „preestride“ („preestrite“) „kirjotud“ („kirjutatud“), „häwwitud“ („hävitatud“), „puhhastakse“ („puhastatakse“) ja „läkkitand“ („läkitanud“).

Nud-kesksõnade jaoks reegli (12) kirjutamisel tuli arvestada, et sõnalõpu *nd* võib mõnikord ka muutmata jääda. Lisaks pikalt kirjutatud kesksõna vormidele on selline lõpp õige ka mõnede nimisõnade, nt „wend“ ehk „vend“ ja „and“ puhul. Sõnade listi uurides tundus, et sagedasem on siiski olukord, kus sõnalõppu muuta tuleb. Selle põhjal määrati ka kaalud. Lisaks sai piiranguks määrata, et selline muutus tohib toimuda üksnes sõna lõpus. Vastasel korral oleks muundur rakendanud reeglit ka näiteks sõna „andma“ jaoks, pakkudes võimalikuks väljundsõnaks „anudma“.

$$define\ NUD\ [\{nd\} \rightarrow\ \{nud\}::1, \{nd\} \rightarrow\ \{nd\}::2\ ||\ \_\ .\#.\ ]\ ; \tag{12}$$

Ta- ja *da*-liiteliste verbide puhul pandi reeglisse (13) kirja, et tühisõne võib asenduda silbiga *ta,* kui talle järgneb kas *tud* („kirjotud“) või *takse* („kirjotakse“) ning seejärel sõna lõpp. Kuna esines ka palju sõnu, mille puhul reeglit rakendada ei tohiks, näiteks „seätud“ („seatud“), „pattud“ („patud“) ja „kogutakse“, tuli reegel kirjutada valikulise vastendamisega. Kaalude lisamisel kirjutati valikuline vastendamine lahti kaheks teisenduseks, kusjuures *ta* silbi lisamine tähistati kaaluga 1 ja sõna samaksjäämine jäi kaaluta.

$$define\ TUD\ [0 \rightarrow\ \{ta\}::1,\ \ 0 \rightarrow 0\ ||\ \_\ [\{tud\}\ |\ \{takse\}\ |\ \{ta\}]\ .\#.\ ]\ ; \tag{13}$$

De-tunnuselise mitmuse omastava eripära kattis ära reegel (14), mille kohaselt võivad nõrgad sulghäälikud asenduda oma tugevate vastetega ja vastupidi. Reegli kirjutamise aluseks olid näiteks sõnad „preestride“ („preestrite“), „prohwetide“ („prohvetite“), „keikist“, („kõigist“), „wadage“ („vaadake“) ja „laenada“ („laenata“). See on kindlasti üks reegel, mis vajab tulevikus veel täpsustamist, kuna see ei kata päris kõiki vajalikke sulghäälikute muutumi (arvestab vaid vokaalide vahel olevatega), kuid kipub siiski ka ebavajalikes kohtades liigseid väljundsõnu genereerima.

$$define\ SULG\ [g \rightarrow k::2,\ g \rightarrow g,\ k \rightarrow g::1,\ k \rightarrow k,\ d \rightarrow t::2,\ d \rightarrow d,\ t \rightarrow d::1,\ t \rightarrow t\ ||\ VOK\_VOK]\ ; \quad (14)$$

Mineviku mitmuse kolmanda pöörde teisendamiseks määrati, et *va* silp võib sõna lõpus *i* ja *d* vahel asenduda tühisõnega. Kirjutamisel tuli arvestada ka sõnadega nagu „käiwad“ („käivad“) ja „otsiwad“ („otsivad“), mida reegel (15) muuta ei tohi. Selgus, et vähemalt Piibli tekstides oli rohkem just selliseid sõnu – seetõttu määrati muutmise toimumise kaaluks 1. Lisaks kirjutati reeglisse sisse *va* asemel *wa*, kuna see teisendusreegel paigutati enne muundurit, mis *w* tähe *v*-ks teisendaks.

$$define\ IVAD\ [\{wa\} \rightarrow 0::1,\ \{wa\} \rightarrow \{wa\}\ ||\ i\_d\ .\#.]\ ; \quad (15)$$

Arnold Kask on oma eespoolmainitud teoses keskendunud ka konkreetselt 1739. aasta Piiblile. Järgnevalt kirjeldatakse väljatoodud erijooni ja esitatakse mõned nende põhjal kirjutatud muundurireeglid. Vaid ühte tähte hõlmavatest erinevustest toob Kask välja, et *õ*-häälikut märgib piiblis kas *ö* või *o* ning vokaalidevahelist *j*-i *i*. Pikkade vokaalide asemel kasutatakse diftonge, millest osad on samamoodi kasutusel ka tänapäeval („pea“, „hea“), kuid enamik vajavad teisendamist. Sellised sõnad on näiteks „seäl“ („seal“), „voöras“ („võõras“) ja „moök“ („mõõk“). (Kask, 1970, lk 94)

Näitesõnade põhjal selgus, et *eä* võib lisaks märkida ka ühendit *ää* (nt sõnas „peästma“ ehk „päästma“) ja *oö* asemel võib esineda ka *öo.* Lisaks võivad nii *oö* kui ka *öo* tähistada lisaks *õõ*-le ka *öö*-d („noör“ ehk „nöör“). Kuna *eä* on eesti keeles küllaltki harvaesinev diftong (olemas näiteks liitsõnas „teeäär“), määrati reeglis (16) ühendi muutmata jätmise kaaluks 2. Asendumine *ea* või *ää*-ga jäeti ilma kaaluta, kuid võib olla mõistlik määrata *ää*-ga asendumise kaaluks 1, kuna seda esineb vähemate sõnade puhul. Nii *oö* kui ka *öo* puhul on sagedasem asendumine *õõ*-ga ja seda väljendati ka kaaludega. Võimalust, et *oö* või *öo* jääb muutmata, reeglites ei kajastatud, kuna see on eesti keeles äärmiselt ebatõenäoline. Siinkohal esitatakse vaid reegel *oö* muutmiseks (17), kuna *öo* jaoks on reegel analoogiline.

$$define\ eä2\ [\{eä\} \rightarrow [\{ea\}\ |\ \{ää\}],\ \{eä\} \rightarrow \{eä\}::2]\ ; \quad (16)$$

$$define\ oö2\ [\{oö\} \rightarrow \{õõ\}::1,\ \{oö\} \rightarrow \{öö\}::2]\ ; \quad (17)$$

Kask mainib muuhulgas veel ka järgsilbi *o* säilimist („kokko“ ehk „kokku“), *ste*-liitelisi adverbe („ussinaste“ ehk „usinasti“), tugevaastmelist sisseütlevat käänet („külges“ ehk „küljes“) (Kask,

1970, lk 94). Näitesõnadest selgus, et *o* tuleks mõnikord asendada *u*-ga ka sõna keskel, näiteks sõnas „koggodus“ ehk „kogudus“. Lisaks selgus, et sõna lõpus on tõenäolisem *o* asendumine *u*-ga, kuid sõna keskel pigem muutmata jäämine. Seetõttu tuli *o* hääliku muutmiseks kirjutada kaks reeglit (18, 19), et seda nähtust kaalude abil väljendada.

$$define\ o2u2\ [o \rightarrow\ u::1,\ o \rightarrow\ o\ ||\ KONS\ \_\ KONS]\ ; \tag{18}$$

$$define\ o2u\ [o \rightarrow\ u,\ o \rightarrow\ o::1\ ||\ \_\ .\#.]\ ; \tag{19}$$

Osade sõnade puhul on ka Piibli tõlkes veel säilinud selle tüve varasem kuju – nii on näiteks „veise“ asemel kirjas „weikse“ ning „pudulojused“ asemel „puddolojuksed“. Lisaks esineb mõnede sõnade varasemaid traditsioonilisi vorme: „sanna“, „seie“, „keik“ ja „leikas“ ehk „sõna“, „siia“, „kõik“ ja „lõikas“ (Kask, 1970, lk 94). Kinnitust leidis ka näitesõnadest silmajäänud ja esialgu kummaline tundunud nähtus – sõna „sõda“ nõrga astme vorm oli „söa“ (Kask, 1970, lk 95).

Sõnadega nagu „weiksed“ („veised“) ja „sörmuksed“ („sõrmused“) tegelemiseks kirjutati reegel, mis eemaldas sõna lõpus vokaali ja *sed* vahelt *k*, tähekombinatsioon „söa“ vastendati alati sõnaga „sõja“ ning lisaks kirjutati reegel, mille kohaselt võib diftongile *ei* vastata diftong *õi*. Sõna „sanna“ puhul lisati esmalt juba olemasolevale reeglile, et ka *a* häälik võib *õ*-ga asenduda, kuid näitesõnu analüüsides selgus, et tegu oli pigem siiski erandliku juhuga. Seetõttu võeti see muudatus tagasi, ning sõna jaoks kirjutati eraldi reegel (20).

$$define\ sõna\ [\{sanna\} -> \{sõna\}\ ;] \tag{20}$$

Sarnaselt toimiti ka teiste erandlike sõnadega. Eespoolmainitutest näiteks sõnaga „keik“ ehk „kõik“. Kõigi ortograafiamuunduri tarbeks kirjutatud teisendusreeglite ja nende aluseks olevate näitesõnadega saab tutvuda lisades (lisa 2). Järgnevalt näidatakse, milliseid muutuseid viidi töö käigus sisse morfoloogiamuundurisse.

2.4 Morfoloogiamuunduri modifikatsioonid

Kuna bakalaureusetöö põhirõhk oli siiski ortograafiamuunduri koostamisel, viidi morfoloogiamuundurisse sisse vaid osad töö käigus avastatud morfoloogilised erijooned. Sellega loodab autor eelkõige näidata, et kuigi morfoloogiamuunduri struktuur võib esialgu tunduda liiga keeruline, on selle muutmine ja hetkeülesandele kohandamine tegelikult võimalik ja huvilistele jõukohane.

Peatüki sissejuhatuses toodi näiteks sõna „pöörma“, mille puhul oli vana kirjaviisi ajal kasutusel kaasajast erineva muuttüübiga vorm. Lisaks sellele sõnale on näitesõnade puhul alust kahtlustada muuttüübi muutumist ka sõnade „hakkama“, „lükkama“, „hukkama“ ja „käskima“ puhul. Piiblis esineb hulganisti vorme, mida kaasajal nende sõnade puhul õigeks ei loetaks – näiteks „hakkada“, „hakkage“, „lükkago“, „käskma“ ja „käsknud“ ehk „hakata“, „hakake“, „lükaku“, „käskima“ ja „käskinud“. Nende vormide puhul paistab, et sõnad „hakkama“, „lükkama“ ja „hukkama“ käitusid vana kirjaviisi ajal sarnaselt sõnaga „leppima“ („leppida“, „leppige“ jne) ning „käskma“ ehk „käskima“ käitus sarnaselt sõnaga „laskma“ („lasknud“ jne). Kuna need vormid esinesid Piiblis järjepidevalt, otsustati lisada morfoloogiamuundurisse *morphology/stems* kausta verbide leksikoni *verbs.lexc* ka nende verbide vanapäraste muuttüüpidega variandid. Kaalud valiti samad, mis kaasaegse muuttüübiga versioonidel ja ka kaasaegse muuttüübiga variant jäeti alles. Lisatud read tähistati kommentaariga „modified“, et neid hiljem faili originaalsisust eristada. Järgnevalt on joonisel 8 toodud näide nimetatud failist peale sõna „hakkama“ vanapärase muuttüübiga lisamist.

```
hakkama=saama+V:hˇakkama=sˇaa SAAMA “weight: 11“ ;
hakkama+V:hˇakka HAKKAMA “weight: 3“ ;
hakkama+V:hˇakka LEPPIMA “weight: 3“ ;            !modified
```

Joonis 8. Faili *morphology/stems/verbs.lexc* sisu

Järgnevalt esitatakse näited morfoloogiamuunduri väljundist käsureal enne ja pärast kirjeldatud muutuse tegemist.

```
$ echo ’hakkada’ | hfst-lookup morfoloogiamuundur_analüüsiga.hfst
> hakkada   hakkada+?   inf
```

Joonis 9. Käsurea väljund enne muutust

```
$ echo ’hakkada’ | hfst-lookup morfoloogiamuundur_analüüsiga.hfst
> hakkada   hakkama+V+Inf     3,000000
```

Joonis 10. Käsurea väljund peale muutust

Erinevad liitsõnade moodustamist puudutavad muudatused tuleb tingimata viia sisse just morfoloogiamuundurisse. Vana kirjaviisiga võrreldes on liitsõnade moodustamise reeglid mõneti muutunud. Näiteks ei ole tänapäeval enam grammatiliselt korrektne öelda „ärapäästan“ või „ülesehitada“, kuid vanas kirjaviisis tekstides on sõnad nagu „ärrapeästan“, „üllesehhitan“ ja muud sarnased vormid väga tavalised. Kirjaviisimuundur suutis õigesti ära tunda küll sõnad „ärra“ („ära“) ja „peästan“ („päästan“), kuid kuna selline liitsõna pole eesti keeles lubatud, siis „ärrapeästan“ puhul õiget varianti ei pakutud. Küll aga esines ühes testimise staadiumis (kus ka ortograafiamuundur veel sõnadele väga palju võimalikke väljundeid esitas) pakutud variantide hulgas näiteks sõna „eraõppeseen“. Nimisõnadest liitsõnade moodustamise reeglid on eesti keeles vabamad ja nii tundus see sõna morfoloogiamuundurile igati sobivana. Selleks, et morfoloogiamuundur aktsepteeriks ka vana kirjaviisi päraseid liitsõnu, tuli muundurist eemaldada piirang, mis takistaks tegusõnadel sääraste liitsõnade moodustamise. Nagu peatükis 1.3 mainitud, olid need piirangud saavutatud lipudiakriitikute abil. Seega tuli keelavad märgid *morphology/affixes* kausta *verbs.lexc* failis õigetes kohtades välja kommenteerida. Järgnevalt tuuakse sellest joonise 11 üks näide. Algne rida on hüüumärgi abil välja kommenteeritud ning selle all on uus lisatud rida. Sarnaselt tuli toimida kõigi vajalike tegusõnavormide puhul, et tuntaks ära nii „ärapäästa“, „ärapäästnud“, „ärapäästate“ ja muud seesugused.

```
LEXICON A_INFINITIVE
! @R.Part.One     INF_COMP ;
INF_COMP ;                          !modified
```

Joonis 11. Faili *morphology/affixes/verbs.lexc* sisu

Oluline on mõista, et see muutus võimaldab sääraseid tegusõnu ära tunda siiski üksnes ortograafiliselt korrektsel kujul. Selgituseks tuuakse taas mõned käsureaväljundid. Joonisel 12 on toodud morfoloogiamuunduri väljundid, joonisel 13 ortograafiamuunduri väljundid ning joonisel 14 kirjaviisimuunduri väljundid (võimalikest väljunditest on iga muunduri puhul esitatud vaid mõned esimesed).

```
$ echo ’ärapäästnud’ | hfst-lookup morfoloogiamuundur_analüüsiga.hfst
> ärapäästnud     ära+Adv#päästma+V+Der/nud+A+Sg+Nom 60,000000

$ echo ’ärrapeästnud’ | hfst-lookup morfoloogiamuundur.hfst
> ärrapeästnud    ärrapeästnud+?    inf
```

Joonis 12. Morfoloogiamuunduri väljund

```
$ echo ’ärrapeästnud’ | hfst-lookup ortograafiamuundur.hfst
> ärrapeästnud    ärapeastnud       0,000000
ärrapeästnud      ärapäästnud       0,000000
ärrapeästnud      arapeastnud       1,000000
```

Joonis 13. Ortograafiamuunduri väljund

```
$ echo ’ärrapeästnud’ | hfst-lookup kirjaviisimuundur_analüüsiga.hfst
> ärrapeästnud    ära+Adv#päästma+V+Der/nud+A+Sg+Nom       60,000000
ärrapeästnud      ära+Adv#päästma+V+Der/nud+A        60,000000
```

Joonis 14. Kirjaviisimuunduri väljund

Kolmandat tüüpi muutus, mis morfoloogiamuunduris tehti, oli see, et sinna lisati erinevate nimi- ja määrsõnade vormid, mis varasemalt teisel kujul kasutusel olid. Näiteks „pitk“ tähenduses „pikk“, „seie“ tähenduses „siia“ ja „sanna“ tähenduses „sõna“. Järgnevalt tuuakse joonisel 15 näide kausta *morphology/stems* kausta *nouns.lexc* faili sisust, kuhu on lisatud, et sõna „pikk“ võib olla esitatud ka kujul „pitk“.

```
pikk+N:pˇikk%>%{pl.i%}  PIIM  “weight: 5“ ;
pikk+N:pˇitk%>%{pl.i%}  PIIM  “weight: 5“ ;     !modified
```

Joonis 15. Faili *morphology/stems/nouns.lexc* sisu

Siinkohal on hea võrrelda analüüsi väljendava ja analüüsi mitteväljendava morfoloogiamuundurit. Joonisel 16 on näha, kuidas üksnes analüüsi väljendava muunduri väljundist on näha sõna „pitkad“ seos sõnaga „pitk“. Seetõttu võib analüüsi mitteväljendava muunduri kasutamisel sellistel puhkudel sõna tähendus ebaselgeks jääda.

```
$ echo ’pitkad’ | hfst-lookup morfoloogiamuundur_analüüsiga.hfst
> pitkad    pikk+N+Pl+Nom     6,000000

$ echo ’pitkad’ | hfst-lookup morfoloogiamuundur_analüüsita.hfst
> pitkad    pitkad      6,000000
```

Joonis 16. Analüüsiga ja analüüsita morfoloogiamuunduri väljundid

Viimaste muutustena võimaldati morfoloogiamuunduril õigeteks lugeda *ivad*-lõpulisi lihtmineviku mitmuse kolmanda pöörde vorme ning lisaks *nud*-kesksõnade kaasaegsele vormile ka *nd*-lõpulisi sõnu. Järgnevalt näidatakse, kuidas võimaldati *ivad*-lõpulised pöördevormid (nt sõnas „läksivad“ ehk „läksid“). Selleks tuli failis *morphology/affixes/verbs.lexc* lisada tavapärase pöördelõpule eelneva *i*-tunnuse juurde lisada ka *iva*. Seda on kujutatud joonisel 17.

```
LEXICON IND_PAST !indicative past
!@R.Part.One@@P.Part.Bad@ IND_PAST_COMP ;
IND_PAST_COMP !modified (for compounds)
+Pers+Prt+Ind:i MARKED_PERSON ;
+Pers+Prt+Ind:iva MARKED_PERSON !modified (läksivad)
```

Joonis 17. Faili *morphology/affixes/verbs.lexc* sisu

Muutusi, mida tasub ortograafiamuundurist morfoloogiamuundurisse tuua ning muutusi, mis tuleks teha näitesõnade analüüsi tagajärjel, on veel mitmeid, kuid eelnevaga näidati ära mõned põhilised viisid morfoloogiamuunduri muutmiseks. Võimaldamaks ortograafiamuunduri eraldiseisvat kasutamist näiteks tekstide lihtsustamisel, jäeti morfoloogiamuundurisse ületoodud teisendused reeglitena väljakommenteeritult ka ortograafiamuundurisse alles. Terviklikul kujul on ortograafiamuunduri loomiseks kasutatud reeglid leitavad lisas 2.

3. Muundurite testimine

Kuna bakalaureusetöös keskenduti ennekõike just 1739. aasta Piiblis kasutusel olevale vanale kirjaviisile, siis on ka muunduri testimisel rõhk sellel, kui edukalt pakutakse õigeid sõnavorme just Piiblitekstist pärit sõnadele. Lisaks tahtis autor aga kontrollida, mil määral on võimalik koostatud muundurit rakendada ka muude vanas kirjaviisis tekstide normaliseerimiseks. Seetõttu valiti teiseks testhulgaks sõnu vana kirjakeele korpuse 18. sajandi ilmalikest tekstidest. Käesoleva peatüki esimeses alapeatükis kirjeldatakse testimise protsessi ning hinnatakse tulemusi kummalgi testhulgal. Teises alapeatükis võrreldakse kõigi töös kasutatud muundurite väljundeid ja analüüsitakse nende nõrkusi ja tugevusi.

3.1 Testhulkade valimine ja ortograafiamuunduri testimine

Kuna töö põhirõhk oli ortograafiamuunduri koostamisel, siis kasutati testimiseks üksnes seda muundurit. Lisaks sellele, kui tihti suudab ortograafiamuundur sisendsõnale õige vaste pakkuda, tuli ka välja selgitada, kui kõrgel positsioonil korrektne variant võimalike väljundsõnade hulgas paikneb. Selle põhjal sai lisaks teisendusreeglite toimimise kontrollimisele analüüsida ka reeglitele seatud kaalude headust. Kui õige sõnakuju on pakutud variantide seas küll olemas, kuid alles kuskil võimalike variantide listi lõpus, siis ei toimi kaalud järelikult kuigi efektiivselt ja vajavad täpsustamist.

Muunduri efektiivsust testiti kahe erineva sõnahulga peal. Esiteks võeti ette kogu 1739. aasta Piibli tõlke sõnestatud tekst ning valiti sellest välja iga tuhandes sõna. Esialgsesse loendisse sattus väga palju korduvaid ja vanast kirjaviisist saati muutmata jäänud sõnu nagu „et“, „ma“, „ja“ ja muud seesugust. Seetõttu kirjutati sõnade valimine ümber nii, et koostati nii-öelda triviaalsete sõnade list ning kui Piibli tekstist valituks osutunud sõna oli tolles listis esindatud, siis valiti tekstist hoopis sellele eelnev sõna. Sel viisil saavutati testimiseks kasutatavate sõnade suurem varieeruvus. Teise sõnade listi jaoks valiti vana kirjakeele korpusest 18. tekstid ning valiti neist samade kriteeriumite alusel näitesõnad. Järjendid teisendati hulkadeks. Piiblitekstist pärit sõnahulka jäi 460 ning korpusetekstide sõnahulka 154 sõna.

Testsõnad salvestati eraldi failidesse ning lasti muunduril pakkuda uues kirjaviisis vasteid iga failis leiduva sõna jaoks (vt näidet peatükist 1.2). Kummagi analüüsi tulemused salvestati taas eraldi failidesse, millede põhjal hinnati Pythoni koodi abil muunduri täpsust. Iga sõna puhul vaadati, mitmendana muundur sellele õiget vastet pakkus (näidet muunduri väljundist vt peatükist 1.2). Täpsemalt vaadati, kui tihti oli õige variant pakututest esimene, esimese viie ja esimese kümne seas ning kui tihti ei leidunud ootuspärast sõnakuju pakutud väljundite seas üldse. Esialgu kasutati korrektse sõnakuju leidumise hindamiseks taas EstNLTK *spellcheck* meetodit, kuid lõpuks vaadati väljund siiski ka käsitsi läbi. Tulemuste hindamisel jäeti täielikult kõrvale testsõnade hulka sattunud pärisnimed, kuna nendega tegelemine ei olnud bakalaureusetöö skoobis.

Vana kirjaviisi teisendamiseks muunduri koostamine on mahukas ülesanne ja erandlikke sõnu on palju, seega ei oleks mõeldav bakalaureusetöö koostamise käigus kõigi erijuhtude ja sõnadega arvestada. Töö eesmärk oli vaid alustada vana kirjaviisi muunduri koostamist ja näidata, kuidas on võimalik morfoloogiamuundurit muuta. Seega otsustati, et lõpliku testimise käigus leitud erandlikel sõnadel või päris uutel seaduspäradel põhinevaid reegleid enam muundurisse juurde ei kirjutata, vaid jäetakse need tulevaseks edasiarenduseks. Mõnel juhul parandati testimise tulemusel saadud info põhjal olemasolevaid reegleid, kuid ainult juhul, kui olemasoleva reegli muutmine ei paistnud liigselt suurendavat muunduri poolt pakutavate väljundite hulka.

Piiblitekstidest testiti muundurit 460 sõna peal. Nendest sõnadest 32 sattusid olema isiku- või kohanimed ja jäid seetõttu analüüsist kõrvale – seetõttu loetakse analüüsil testitud sõnade koguarvuks 428 ja sellest lähtutakse ka täpsusprotsentide arvutamisel. Esimesena pakuti õiget varianti 289 sõna puhul. Suur osa nendest sõnadest olid oma kirjapildilt ka muutmata kujul üsna lähedased ja vajasid vaid mõningaid levinumaid muutusi. Näiteks sõnad „perre“ ehk „pere“ ja „näggo“ ehk „nägu“. Sellest, et reeglitele määratud kaalud töötavad mõistlikult, andis tunnistust asjaolu, et kuigi mõne sõna puhul oli pakutud variante mitukümmend, oli õige sõnakuju nende variantide seas esimene. Näiteks kui eemaldada kaalud peatükis 1.2.1 kirjeldatud näitemuundurilt, on kõigi väljundsõnade kaal 0 ning muutmata jäetud varianti „tere“ pakutakse alles kõige viimasena. Joonisel 4 oli näha, et kaalutud reeglite puhul saime nõuda, et just muutmata jäänud variant oleks tõenäoliseim ja seetõttu pakuti varianti „tere“ esimesena. Peatükis 1.2.1 kirjeldatud muunduri puhul ei tekita see veel probleeme, sest väljundite hulk on niigi väike, aga suuremate muundurite puhul muutuks õige väljundi otsimine väga tülikaks.

Näiteks sõnale „ärrakautada“ („ärakaotada“) pakkus muundur 96 võimalikku väljundit ning sõnale „pöggeneda“ („põgeneda“) lausa 112. Sellegipoolest tuleks tulevikus leida viise reeglite täpsustamiseks ja seeläbi võimalike väljundite arvu piiramiseks.

Esimese viie variandi seas oli õige kuju 91 ning esimese kümne seas 10 sõna puhul. Sinna hulka kuulusid näiteks sõnad „ette“ ja „ikka“, mille puhul oli muutmata jäänud kuju kaal pisut suurem, kui näiteks variantide puhul, mis saadi kahekordse kaashääliku ühekordseks muutmisel. Paljude sõnade puhul oli kaalude selline määramine õigustatud, kuid tulevikus tasub uurida, kas annab muutusele kindlamat konteksti määrata. Nende kahe näitesõna puhul võib näiteks spekuleerida, et näiteks kahesilbilistes sõnades tuleks eelistada kahekordse konsonandi säilitamist. Selle oletuse kinnitamiseks tuleks aga tutvuda palju rohkemate näidissõnadega.

Tagapool kui kümne esimese seas oli õige variant 5 sõna puhul ning õiget varianti polnud pakutud väljundite hulgas 11 sõna puhul. Sõnad „wette“ („vette“) ja „mitte“ paistsid toetavat „ette“ ja „ikka“ najal püstitatud hüpoteesi. Vaadates sõnu, millele õiget varianti ei pakutud, avastati kirjutatud reeglites ka vigu või lihtsalt kahe silma vahele jäänud detaile. Näiteks ei osanud muundur pakkuda sõna „pois“ puhul õiget sõna „poiss“, kuna teisendusreegliga *ou2õu* (vt lisa 2) vastendati tähekombinatsioon *oi* alati diftongiga *õi* sõnade „woi“ („või“), „oige“ („õige“) jms tõttu. Otsustati, et vastava muutuse sisseviimine ei tohiks muunduri toimimist halvemaks muuta ega ebamõistlikult palju valesid väljundsõnu juurde tekitada – seega parandati nimetatud reeglit, kasutades diftongide *oi* ja *õi* puhul valikulist vastendamist. Sõna „ial“ ehk „iial“ ja „elevantilu“ ehk „elevandiluu“ juhtisid tähelepanu vokaali kahekordistamise reegli probleemile: vokaale kahekordistatakse üksnes konsonandiga algava sõna esisilbis, kuna just selles kontekstis on see muutus kõige levinum. Selle reegli muutmine võib aga hakata palju üleliigseid väljundeid genereerima, seetõttu jääb see ülesanne tulevikku. Lisaks leiti, et reeglid ei arvesta liitsõnadega, milles esimene komponent lõpeb *o*-ga ja teine algab vokaaliga („kässoöppetus“ ehk „käsuõpetus“ ja „armoöppetus“ ehk „armuõpetus“). „Kässoöppetus“ osutus probleemseks sõnaks ka seetõttu, et andis koguni 1920 võimalikku väljundit, mis on taaskord kindel märk sellest, et tulevikus tuleb leida veel viise reeglite täpsustamiseks ja väljundite arvu piiramiseks.

Eraldi arvestati sõnu, millele ortograafiamuundur küll ootuspärase vaste pakkus, kuid mis oleks morfoloogiamuunduri poolt tagasilükatud, kuna sõnade kasutus on kaasajaks muutunud. Selliseid sõnu oli 13 ja sinna hulka kuulusid näiteks „pidavad“ ehk „peavad“, „andid“ ehk „annid“, „misga“ ehk „millega“, „senna“ ehk „sinna“ ja „oiniktall“ ehk „oinatall“. Tekstide

lihtsustamise eesmärgil tuleb nende sõnade teisendamiseks ortograafiamuunduris eraldi reeglid kirjutada, kuid keelekasutuse info säilitamiseks tasub sõnad juurde kirjutada morfoloogiamuundurisse ja seostada need nende kaasaegsete kujudega (nagu varasemalt näidatud nt sõna „pitk“ ehk „pikk“ puhul).

Vana kirjakeele korpuse 18. sajandi tekstide 154 sõnale pakkus muundur õiget vastet esimesena 96 ning esimese viie seas 35 sõna puhul. Ka siin oli testitud sõnade pealt näha, et tuleb muuta vokaalide kahekordistamise reeglit. Näiteks jäid tundmatuks sõnad „ädika“ ehk „äädika“ ja „prukida“ ehk „pruukida“. Kaks sõna jäid tundmatuks tähekombinatsiooni *dt* tõttu, mis oleks tulnud asendada tähega *t*. Nimetatud sõnad olid „laosojadte“ ehk „lausujate“ ning „wotmadta“ ehk „võtmata“. Piibli sõnadelistist kontrolliti, et seal esines ühendit *dt* vaid liitsõnades nagu „headteggemised“ ja „ülemadteenrid“. Seega oli arusaadav, et muundurireeglites sellise juhtumiga arvestatud ei oldud.

3.2 Muundurite väljundite võrdlus

Võimaldamaks selgemini mõista erinevate töös kasutatud ja loodud muundurite kasutusvõimalusi ja piiranguid, otsustati teisendada kõigi muunduritega ühesugust sisendit ning esitada vastavad väljundid kõrvuti. Sel viisil paistavad kõige selgemini silma iga muunduri olulised iseloomulikud omadused ja eelkõige just ka töö käigus valminud ortograafiamuunduri olulisus vanas kirjaviisis tekstide analüüsimisel. Lisades (Lisa 3) on esitatud neli versiooni 1739. aasta Piiblist pärit „Esimese Moosese raamatu“ kuuendast peatükist. Esmalt on toodud tekst algselt kujul (kättesaadav järgneval aadressil: https://www.eki.ee/piibel/index.php#1Ms6) ning seejärel teisendatuna üksnes ortograafiamuunduriga, üksnes morfoloogiamuunduriga ning viimaseks kirjaviisimuunduriga. Ortograafiamuunduriga teisendamiseks kasutati ka neid reegleid, mis lõplikus versioonis välja on kommenteeritud, kuna muutused viidi sisse morfoloogiamuundurisse.

Piiblist valitud tekstikatkendist tehti esmalt sõnade järjend ning lasti see igast muundurist läbi. Muundurite väljundid salvestati tekstifailidesse, võimaldamaks nende edasist töötlust Pythoni koodi abil. Iga sõna puhul vaadati, mitmendana ta muunduri pakutud väljundite seas esines. Kui õiget varianti pakuti esimesena, kirjutati teisendatud kujul teksti üksnes see sõna. Kui õige variant oli ortograafiamuunduri puhul viie ning morfoloogia- ja kirjaviisi muunduri puhul

kolme esimese seas, siis kirjutati teisendatud teksti list pakutud variantidest kuni õige sõnani. Seega näiteks sõna „ja“ kohale kirjutati esimeses teisendatud kujul tekstis üksnes „ja“, kuna see oli ka ortograafiamuunduri esimene pakkumine. Sõna „jure“ kohale kirjutati aga „[jure, juurde]“, kuna õige variant oli muunduri väljundi seas teisel kohal. Morfoloogia- ja kirjaviisimuundurite puhul on lugemisselguse huvides kasutatud varianti, mis sõnade morfoloogilist analüüsi ei väljasta (vt peatükk 1.2.2). Kui muundur pakkus ainult ühe väljundi, mis oli vale, esitati see tekstis läbikriipsutatud kujul. Kui muundur sõna ära ei tundnud, esitati sõna kujul *[sõna+?]*. Ortograafiamuunduri ja morfoloogiamuunduri võrdluses on erinevused esitatud kapiteelkirjas.

Väljunditest võib selgelt jälgida iga muunduri nõrkusi ja tugevusi. Üksnes morfoloogiamuunduri abil teisendatud tekst väljendab väga ilmekalt, miks on vaja vanas kirjaviisis tekste enne analüüsimist normaliseerida – ära ei tunta ühtegi sõna, mille kirjapilt kaasaja kirjaviisist erineb. Lisaks pakutakse näiteks sõna „sured“ puhul ainult pöördevormi sõnast „surema“, kuigi tegelikult on tekstis mõeldud hoopis sõna „suured“. Ortograafiamuundur üksi on teksti teisendamises päris edukas, kuid kuna sellel puudub võime väljundeid keeleliselt analüüsida ja sõnade korrektsuse üle otsustada, satub väljundite hulgas kõrgele positsioonile ka sõnu, mida eesti keeles üldse olemas ei ole (nt sõna „inimeste“ puhul). Kui ortograafiamuunduri väljund anda sisendiks morfoloogiamuundurile (kirjaviisimuunduri tööpõhimõte), filtreerib morfoloogiamuundur väljundist välja kõik sõnad, mida eesti keeles olemas ei ole, ja seega on pakutud alternatiivide hulk väiksem ning ei sisalda sisutühje tähekombinatsioone (nt „inimesde“). Samas jääb kirjaviisimuundur sarnaselt morfoloogiamuundurile hätta pärisnimedega, kuna need ei ole korrektsed eestikeelsed sõnad ja on seega muundurile tundmatud. Ortograafiamuundur võib nimele õige teisenduse küll genereerida, ent see filtreeritakse välja. Nagu peatükis 2.4 mainitud, võib sarnane olukord tekkida ka liitsõnadega, mille moodustamis tänapäevased kirjakeelenormid enam ei võimalda. Pärisnimedega arvestamiseks tuleks lisada nimede list morfoloogiamuundurisse. Sõnade „pikk“ ja „pitk“ puhul tuleb hästi välja see, kuidas ainult ortograafiamuunduriga teisendades muutub tekst küll kaasaja kirjakeelt mõistvale inimesele selgemaks, kuid läheb kaotsi info ajaloolise keelekasutuse kohta.

Peatüki kokkuvõtteks võib öelda, et kirjaviisimuundurist on vanas kirjaviisis tekstide normaliseerimisel juba kindlasti abi, kuid selle täiendamise ja parandamisega seisab veel suur töö ees. Bakalaureusetöö kontekstis võib ortograafiamuunduri toimimisega rahule jääda.

Esimese viie pakutava väljundi hulka kuulus õige sõnakuju Piibli tekstide puhul 89% ning korpuse 18. sajandi ilmalike tekstide puhul 84% testsõnadest. Arvestades seda, et muundur pakub väljundsõnu vaid tähekombinatsioone reeglite alusel ringi tõstes, teadmata seejuures midagi eesti keelest, on testimisel saavutatud täpsusprotsendid küllaltki head.

Kokkuvõte

Bakalaureusetöö raames valminud ortograafiamuunduri kasutegur vanas kirjaviisis tekstide automatiseeritud analüüsil on lisas esitatud Piibli katkendi võrdlustest selgelt näha. Ka enamikele juhuslikult valitud testsõnadele pakkus muundur õiget kaasaja kirjakeele normidele vastavat sõnakuju. Töös kirjeldati näidete ja selgitustega ka ortograafiamuunduri reeglite kirjutamise protsessi ja muundurite testimistulemuste kirjeldamisel juhiti tähelepanu ka mõnedele kõige veaohtlikumatele kohtadele, millega muundurireeglites tulevikus tegelema peaks. Lisaks tutvustati töös ka erinevaid muutusi, mida peaks sisse viima hoopis olemasolevasse eesti keele morfoloogiat analüüsivasse muundurisse. Tutvustati morfoloogiamuunduri struktuuri ning selgitati näidete toel, kuidas ja kuhu erinevat tüüpi muutuseid sisse saab viia. Lisas esitati ka skript, mille abil morfoloogiamuundur üles seada, ja töös oli rohkelt näiteid muundurite rakendamisest..

Töö eesmärk oli luua aluspõhi vana kirjaviisi normaliseerimiseks lõplike muundurite abil ning näidata kätte suund alustatud lahenduste edasiarendamiseks. Eelneva põhjal võib öelda, et töö täitis oma eesmärgi. Loodud ortograafiamuunduri teisendusreeglid on töös terviklikul kujul olemas, seega on muunduri edasiarendamiseks vaja vaid reeglid *.xfscript* faili kopeerida, neid soovitud kujul muuta, kustutada või lisada ja täiendatud reeglifaili pealt töös kirjeldatud viisil uus muundur kokku panna. Olemasolevat morfoloogiamuundurit muudeti küll vähe, kuid tehtud muutused katsid päris mitu olulist kategooriat, andes seega ehk inspiratsiooni mitmete analoogsete muudatuste sisseviimiseks. Vähem oluline pole ka asjaolu, et töö käigus näidati, et kuigi olemasolev morfoloogiamuundur võib tunduda keeruline ja hirmutav, on see tegelikult üpris kasutajasõbralik ning sellega töötamist ei pea kartma. Töös kohandati seda üksnes vanale kirjaviisile vastavaks, kuid sarnaseid rakendusi leidub kindlasti mitme eri valdkonna puhul.

Konkreetses töös alustatud muunduri edasiarendamiseks tuleks süsteemselt tutvuda veelgi enamate näitesõnadega nii 1739. aasta Piiblist kui ka muudest vanas kirjaviisis kirjutatud tekstidest. Sel viisi leiaks kindlasti viise kirjutatud reeglite täpsustamiseks, uute reeglite kirjutamiseks ning suudetaks paremini kohandada ka morfoloogiamuundurit. Töö autoril on huvi loodud muundurite ning nende võimalike laiendustega ka tulevikus edasi tegeleda.

Viidatud kirjandus

Beesley, K. R., & Karttunen, L. (2003). *Finite State Morphology.* CSLI Publications.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc. Kasutamise kuupäev: veebruar 2019. a., allikas https://www.nltk.org/

Eesti Rahvusraamatukogu, Tartu Ülikooli Raamatukogu. (2019). *Eesti märksõnastik*. Kasutamise kuupäev: 3. märts 2019. a., allikas https://ems.elnet.ee/index.php

Erelt, M., Erelt, T., & Ross, K. (2007). *Eesti keele käsiraamat: Morfoloogia*. Kasutamise kuupäev: 4. aprill 2019. a., allikas Eesti Keele Instituut: https://www.eki.ee/books/ekk09/index.php?p=3&p1=3&fbclid=IwAR02yXUBTCauHlzfHdpeMhpepdlxm8Znvf3nI9XoYp34wsRw5cLa3jXOH2A

Erelt, M., Erelt, T., & Ross, K. (2007). *Eesti keele käsiraamat: Ortograafia*. Kasutamise kuupäev: 17. veebruar 2019. a., allikas Eesti Keele Instituut: https://www.eki.ee/books/ekk09/index.php?p=2&p1=2

Kaalep, H.-J. (6. aprill 2017. a.). Morfoloogiline analüüs: lõplikud muundurid. Kasutamise kuupäev: 8. mai 2019. a., allikas http://kodu.ut.ee/~hkaalep/markile/morfoloogia_2017_informaatikutele.pdf

Kaalep, H.-J., Moshagen, S. N., & Trosterund, T. (2018). Estonian Morphology in the Giella Infrastructure. *Human Language Technologies - The Baltic Perspective*, 47-54. (K. Muischnek, & K. Müürsepp, Toim-d) IOS Press BV. doi:doi:10.3233/978-1-61499-912-6-47

Kask, A. (1970). *Eesti kirjakeele ajaloost I.* Tartu: Tartu Riiklik Ülikool, Eesti keele kateeder.

Kingissepp, V.-L. (2001). *Eesti keele esimestest kirjapanekutest.* Kasutamise kuupäev: 28. märts 2019. a., allikas Emakeeleselts: Oma Keel: http://www.emakeeleselts.ee/omakeel/2001_1/OK_2001-1_01.pdf

Koskenniemi, K. M., & Kuutti, P. (detsember 2017. a.). *Indexing Old Literary Finnish text.* Kasutamise kuupäev: 26. märts 2019. a., allikas https://core.ac.uk/download/pdf/146449084.pdf

Koskenniemi, K., & Kuutti, P. (2017). *Indexing Old Literary Finnish text.* Kasutamise kuupäev: 13. märts 2019. a., allikas https://core.ac.uk/download/pdf/146449084.pdf

Muischnek, K. (2015). *Keelekorpused – sama mitmekesised.* Kasutamise kuupäev: 3. märts 2019. a., allikas Emakeele Selts: Oma Keel: http://www.emakeeleselts.ee/omakeel/2015_1/OK_2015-1_05.pdf

Muischnek, K., Kaalep, H.-J., & Sirel, R. (2011). Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 111-127. doi:http://dx.doi.org/10.5128/ERYa7.07

Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S., & Kaalep, H.-J. (2016, mai 23-28). EstNLTK - NLP Toolkit for Estonian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. (N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, . . . S. Piperidis, Eds.) Portorož, Sloveenia: European Language Resources Association (ELRA). Retrieved veebruar 2019, from https://estnltk.github.io/estnltk/1.4.1/

Pilvik, M.-L., Muischnek, K., Jaanimäe, G., Lindström, L., Lust, K., Orasmaa, S., & Türna, T. (2019). Möistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 139-158. doi: http://dx.doi.org/10.5128/ERYa15.08

Prillop, K. (9. jaanuar 2013. a.). *Vana kirjakeele korpus: Avaleht*. doi:https://doi.org/10.15155/TY.0005

Prillop, K. (9. jaanuar 2013. a.). *Vana kirjakeele korpus: Tekstid*. doi:10.15155/TY.0005

Pruulmann-Vengerfeldt, J. (2010). Praktiline lõplikel automaatidel põhinev eesti keele morfoloogiakirjeldus. TÜ matemaatika-informatika teaduskonna magistritöö. Tartu. Kasutamise kuupäev: 2. mai 2019. a., allikas https://cyber.ee/research/theses/jaak_pruulmann-vengerfeldt_msc.pdf

VAKK. (2013). doi:doi:10.15155/TY.0005

Lisa 1: Skript morfoloogiamuunduri ülesseadmiseks[6]

```
#!/bin/bash
# viimati töötas:
#      2018-01-17 ubuntu 16.04
#      2018-01-16 ubuntu 16.04
# sudo mount -o uid=1000,gid=1000 -t vboxsf gt-gramcheck jagatud

refresh_all_stuff()
{
    echo ==
    echo == refresh_all_stuff
    echo ==
    sudo apt-get -y update
    sudo apt-get -y dist-upgrade
}

get_basic_staff_from_Ubuntu_repos()
{
    echo ==
    echo == get_basic_staff_from_Ubuntu_repos
    echo ==
    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
    sudo apt-get -y install bison
    sudo apt-get -y install libgoogle-perftools-dev
    sudo apt-get -y install libhfst-dev libpugixml-dev libarchive-dev
    sudo apt-get -y install libcg3-dev
}

# muuhulgas vist uuemad hfst asjad
get_packages_from_Apertium_repo()
{
    echo ==
    echo == get_packages_from_Apertium_repo
    echo ==
    # unhammer@fsfe.org said:
    # On Ubuntu, you should get packages like hfst, hfst-ospell and pugixml
    # from repos, and preferably remove anything you installed to
    # /usr/local. The only thing to build from source is divvun-gramcheck.
    # You get the relevant packages using
    # wget http://apertium.projectjj.com/apt/install-release.sh -O - | sudo
bash
    wget https://apertium.projectjj.com/apt/install-nightly.sh -O - | sudo bash
    sudo apt-get update
    sudo apt-get -y install apertium-all-dev
    sudo apt-get -y install hfst-ospell-dev libhfstospell10
    sudo apt install divvun-gramcheck
 }
```

[6] Skripti autor on Tarmo Vaino

```
# vaja lausekontrollija jaoks
build_divvun_from_github_sources()
{
    echo ==
    echo == build_divvun_from_github_sources
    echo ==
    # then do a fresh checkout of divvun-gramcheck and it should build
    # *without* any changes to PKG_CONFIG_PATH or similar:
    #    git clone https://github.com/divvun/divvun-gramcheck
    #    cd divvun-gramcheck
    #    ./autogen.sh
    #    ./configure
    #    make -j2
    #    make install
    pushd ~
    svn co https://github.com/divvun/divvun-gramcheck.git
    pushd ~/divvun-gramcheck.git/trunk/
    ./autogen.sh
    ./configure
    #./configure --enable-checker    <-- HJK käes 15.01.2018 virises, et
pugixml on puudu...
    #./scripts/get-pugixml-and-build #<< seda ei tee, tuli valmistehtult
    make -j
    sudo make install
    popd
    popd
}

# eesti muunduri tegemiseks vajalikud keelest sõltumatud asjad
build_giella_core_from_github_sources()
{
    echo ==
    echo == build_giella_core_from_github_sources
    echo ==
    mkdir -p $HOME/giellatekno
    pushd $HOME/giellatekno
    svn co https://victorio.uit.no/langtech/trunk/giella-core giella-core
    pushd $HOME/giellatekno/giella-core
    ./autogen.sh
    ./configure --disable-silent-rules --prefix=/usr/local
    #make -j #<< pole vaja
    sudo make install
    echo "export GTCORE=$HOME/giellatekno/giella-core" >> $HOME/.profile
    . ~/.profile
    popd
    popd
}
```

```
# eesti muunduri tegemiseks vajalikud keelest sõltumatud asjad
build_giella_shared_from_github_sources()
{
    echo ==
    echo == build_giella_shared_from_github_sources
    echo ==
    mkdir -p $HOME/giellatekno
    pushd $HOME/giellatekno
    svn co https://victorio.uit.no/langtech/trunk/giella-shared giella-shared
    pushd $HOME/giellatekno/giella-shared
    ./autogen.sh
    ./configure --disable-silent-rules --prefix=/usr/local
    #make -j #<< pole vaja
    sudo make install
    echo "export GIELLA_SHARED=$HOME/giellatekno/giella-shared" >>
$HOME/.profile
    . ~/.profile
    popd
    popd
}

# tee eesti muundur
build_exp_lang_est_from_github_sources()
{
    echo ==
    echo == build_exp_lang_est_from_github_sources
    echo ==
    mkdir -p $HOME/giellatekno
    pushd $HOME/giellatekno
    svn co https://victorio.uit.no/langtech/trunk/experiment-langs/est
experiment-langs/est
    pushd ~/giellatekno/experiment-langs/est/
    ./autogen.sh
    # Esimesel korral või kui on vaja lisada keskkonnamuutuja GTLANG_est
    # ./autogen.sh –l
    ./configure --with-hfst --without-xfst
    # lausekontrollija puhul võiks olla: ./configure --with-hfst --without-xfst
--enable-grammarchecker --enable-alignment --enable-reversed-intersect
    # 12.10.2018 töötas ka selline asi:
    # ./configure --with-hfst --without-xfst --disable-transcriptors --enable-
spellers --enable-grammarchecker --enable-tokenisers
    # 1.02.2019 sai selle peale veateate:
    # configure: error: divvun-validate-suggest required for building grammar
checkers
    # lahenduseks oli sudo apt install divvun-gramcheck (mis on nüüd
...from_Apertium... skriptis olemas)
    make –j
    # vist ainult lausekontrollija puhul: sudo make install
    # et tekitada kataloogi modes, kus asuvad töövood
    # vist ainult lausekontrollija puhul: pushd ~/giellatekno/experiment-
langs/est/tools/grammarcheckers
    # vist ainult lausekontrollija puhul: make dev
    # vist ainult lausekontrollija puhul: popd
    popd
    popd
}
```

```
#http://wiki.apertium.org/wiki/Using_Giellatekno_Divvun_spellers_with_LibreOffice-Voikko_on_Debian

# vaja LibreOffice spelleri jaoks
get_voikko_from_Apertium_repo()
{
    echo ==
    echo == $FUNCNAME
    echo ==
    # wget http://apertium.projectjj.com/apt/install-nightly.sh
    # sudo bash install-nightly.sh
    sudo apt-get install libreoffice-voikko
}

# vaja LibreOffice spelleri jaoks
build_voikko_from_github_sources()
{
    echo ==
    echo == $FUNCNAME
    echo ==
    pushd ~
    git clone https://github.com/voikko/corevoikko/
    pushd corevoikko/libvoikko
    ./autogen.sh
    ./configure --with-dictionary-path=/usr/share/voikko:/usr/lib/voikko --
enable-hfst
    make -j
    sudo make install
    echo 'export LD_LIBRARY_PATH=/usr/local/lib:"${LD_LIBRARY_PATH}"' >>
~/.bash_profile
    echo 'export PATH=/usr/local/bin:"${PATH}"' >> ~/.bash_profile
    popd
    popd
}

test_est_grammarchecker()
{
    echo ==
    echo == test_est_grammarchecker
    echo ==
    pushd ~/giellatekno/experiment-langs/est/tools/grammarcheckers
    echo 'Ta ei ( tule.' | ./modes/estgram.mode
    popd
}

if [ -z $1 ]
then
       echo vaikimisi teeme kõike
       refresh_all_stuff
       get_basic_staff_from_Ubuntu_repos
       get_packages_from_Apertium_repo
       build_divvun_from_github_sources
       build_giella_core_from_github_sources
       build_giella_shared_from_github_sources
       build_exp_lang_est_from_github_sources
       test_est_grammarchecker

else
       echo ainult seda
       $1
fi
```

Lisa 2: Ortograafiamuundurireeglid koos näidissõnadega

```
define KONS [r | t | p | s | d | f | g | h | j | k | l | v | b | n | m | w] ;
define topeltKONS [{rr} | {tt} | {pp} | {ss} | {dd} | {ff} | {gg} | {hh} | {jj} |
{kk} | {ll} | {vv} | {ww} | {bb} | {nn} | {mm}] ;
define VOK [a | e | i | o | u | õ | ä | ö | ü] ;
define topeltVOK [{aa} | {ee} | {ii} | {oo} | {uu} | {õõ} | {ää} | {öö} | {üü}] ;

!emmale, sinno, temma, waggasid, pallutakse AGA tulla, minna, sinna
define eemaldaKONS [{bb} -> b, {dd} -> d, {ff} -> f, {gg} -> g, {hh} -> h, {jj} ->
j, {jj} -> {jj}::1, {kk} -> k, {kk} -> {kk}::1, {ll} -> l, {ll} -> {ll}::1, {mm} -
> m, {mm} -> {mm}::1, {nn} -> n, {nn} -> {nn}::1, {pp} -> p, {pp} -> {pp}::1, {rr}
-> r, {rr} -> {rr}::1, {ss} -> s, {ss} -> {ss}::1, {tt} -> t, {tt} -> {tt}::1,
{vv} -> v, {ww} -> w] ;

!peatük, kül AGA moistkem, kaswagem
define lisaKONS [k -> {kk}::1, k -> k, l -> {ll}::1, l -> l, m -> {mm}::1, m -> m,
n -> {nn}::1, n -> n, p -> {pp}, p -> p::1, s -> {ss}::1, s -> s, t -> {tt}::1, t
-> t || VOK _ .#.] ;

!se, kuendmaks, job, rikidest
define lisaVOK [a -> {aa}::1, a -> a, e -> {ee}::1, e -> e, i -> {ii}::1, i -> i,
o -> {oo}::1, o-> o, u -> {uu}::1, u -> u, õ -> {õõ}::1, õ -> õ, ä -> {ää}::1, ä -
> ä, ö -> {öö}::1, ö -> ö, ü -> {üü}::1, ü -> ü || .#. KONS _ [KONS | .#.]] ;

define häälikupikkused eemaldaKONS .o. lisaKONS .o. lisaVOK ;

!läksiwad AGA käiwad, räkiwad, otsiwad, eksiwad
!define IVAD [{wa} (->) 0::1 || i _ d .#.] ; !morfoloogiamuunduris

!kirjotud, häwwitud, önnistud, ärrakautud AGA seätud, pattud, wallitsetud
!puhhastakse, kautakse, ärratakse AGA kogutakse, seatakse
define TUD [0 (->) {ta}::1 || _ [{tud} | {takse}] .#.] ;

!katsund, piddand, woind, läkkitand AGA wiskümend, ohwriand, wend
!define NUD [{nd} -> {nud}::1, {nd} -> {nd}::2 || _ .#.] ; !morfoloogiamuunduris

!hakkada, lükkada, hukkada, hakkage, lükkago
!define KATA [{kkada} -> {kata} || _ .#.] ; !morfoloogiamuunduris

!kutsnud, wotwad, rääksid, usksid, seiswad, jookswad, kutswad, süütma, aitma,
waatma
!AGA mahhalasknud, woiksid, olleksid, läksid, wöttaksid, wötma, moistma, saatma
define VOKVAHELE [0 -> [a | e | i | u]::1, 0 -> 0 || KONS _ [{nud} | {sid} | {wad}
| {ma}] .#.] ;

!kuendamal, kolmandamal, wiendamal, neljandama)
define DAMAL [{am} (->) 0 || [{kuuend} | {kolmand} | {neljand} | {viiend} |
{kuuend}] _ ] ;

!nuumweiksed, lojuksed, sörmuksed AGA uksed, vennaksed, ommaksed
define KS [k (->) 0 || VOK _ {sed}] ;

!define sõnamuutused IVAD .o. TUD .o. NUD .o. KATA .o. VOKVAHELE .o. DAMAL .o. KS;
define sõnamuutused TUD .o. VOKVAHELE .o. DAMAL .o. KS;
```

```
!suggu->soo (suggule, suggust), sou->sugu
define sugu [{suggu} (->) {soo}, {sou} (->) {sugu}] ;

!söa=sõja (söawäggi, söamehhed)
define sõja [{söa} -> {sõja}] ;

!define pikk [{pitk} -> {pikk} | {pik}] ; !morfoloogiamuunduris

!maenits->manits (maenitsus, manitsema)
define manitsus [{maenits} -> {manits}] ;

!naene, naese, naest
!define naine [{nae} -> {nai} || _ [s | n]] ; !morfoloogiamuunduris

define juurde [{jure} (->) {juurde}] ;

!külges, selgas, jalgas
define lg2 [g (->) [j | 0] || l _ VOK s] ;

!ep/es = ei (es lõunaeesti murded, ep põhjaeesti murded)
define ei [{ep} -> {ei}, {es} -> {ei} || .#. _ .#.] ;

!define siia [{seie} -> {siia} || .#. _ .#.] ; !morfoloogiamuunduris

!define sõna [{sanna} -> {sõna}] ; !morfoloogiamuunduris

define granaatõun [{kranati} -> {granaat}] ;

!define tüvemuutused sugu .o. granaatõun .o. sõja .o. pikk .o. manitsus .o. sõna
.o. naine .o. juurde .o. lg2 .o. ei .o. siia .o. sõna;
define tüvemuutused sugu .o. granaatõun .o. sõja .o. manitsus .o. juurde .o. lg2
.o. ei ;

!päwal, päwa, pääw, näwad
define äw2äe [0 -> e::1, 0 -> 0::2 || ä _ w] ;

!päiwil, päiwist
define äi2äe [i -> e || {pä} _ w] ;

!w->v alati (wiis, waest, woimus)
define w2v [w -> v] ;

!iggaweste, töeste AGA waeste
define e2i [e -> i::1, e -> e::2 || t _ .#.] ;

!keikist, wadage
define SULG [g -> k::2, g -> g, k -> g::1, k -> k, d -> t::2, d -> d, t -> d::1, t
-> t || VOK _ VOK] ;

!annud, tunnud, künnud, kannud AGA pannud, linnud
define nn2ndn [0 (->) d || n _ {nud}] ;

!wasto, paljo, koggodus
define o2u2 [o -> u::1, o -> o || KONS _ KONS] ;
define o2u [o -> u, o -> o::1 || _ .#.] ;
```

```
!poia, koia, maia, raianud AGA hoia, teie, aia
define i2j [i -> j::1, i -> i::2 || VOK _ VOK] ;

!tundiad, andia, hoidiat, kandia, hüdia, näggia, teggia, mängia
define tegijanimed [{dia} -> {dja}, {dia} -> {dia}::3, {gia} -> {gija}, {gia} ->
{gia}::3] ;

!nenda
define e2õ [e -> õ::2, e -> e || KONS _ KONS] ;

!kurbdus, kaebdus, põlgdus
define bd2b [{bd} -> b, {bd} -> {bd}::2, {gd} -> g, {gd} -> {gd}::2 || _ {us}] ;

!önsaks, tansinud AGA ainsa
define ns [0 (->) [d | t] || n _ s] ;

define kontekstiga äw2äe .o. te2de .o. äi2äe .o. w2v .o. e2i .o. SULG .o. nn2ndn
.o. o2u2 .o. o2u .o. i2j .o. tegijanimed .o. e2õ .o. bd2b .o. ns ;

!seäl, teäda, peält, peästma, heäl
define eä2 [{eä} -> [{ea} | {ää}], {eä} -> {eä}::2] ;

!tousma, nou, louna, oue, jouab, woi, woib, woim, weewoud
define ou2õu [{ou} -> {õu}, {ou} -> {oo}::3, {oi} (->) {õi}] ;

!keik, leikama, veike AGA seiswad, leiwa, weiste, seitse
define ei2 [{ei} -> {ei}, {ei} -> {õi}::1, {ei} -> {äi}::2] ;

!woörad, wööras, moöga, moötis, moöt, noör, töod
define oö2 [{oö} -> {õõ}::1, {oö} -> {öö}::2] ;
define öo2 [{öo} -> {õõ}::1, {öo} -> {öö}::2] ;

!öige, jöeks, nöel, öest, pöddema, öddedele, öppetus, körb)
define ö2õ [ö -> õ::1, ö -> ö::2] ;

!lähhäb, nääb, lähhäwäd
define ä2 [ä -> ä, ä -> [e | a]::1] ;

!öölda, röömsad
define öö2 [{öö} -> {öö}, {öö} -> [{öe} | {õõ}]::1] ;

!teud, seutud, seo, peus AGA weeupputus, walleusk
define eu2eo [{eu} -> {eo}::1, {eu} -> {eu}::2] ;

!öälus, öälattel, söäluda, nööl
define öä2õe [{öä} -> {õe}] ;

!kautama, lautab, maud, tautud, jauks AGA aus ja kaup
define au2ao [{au} -> {ao}, {au} -> {au}::1] ;

define piiramata  eä2 .o. ou2õu .o. ei2 .o. oö2 .o. öo2 .o. ö2õ .o. ä2 .o. öö2 .o.
eu2eo .o. öä2õe .o. au2ao ;

regex häälikupikkused .o. sõnamuutused .o. tüvemuutused .o. kontekstiga .o.
piiramata ;
```

Lisa 3: Piiblikatkendite võrdlus

A) Vanas kirjaviisis	B) Morfoloogiamuunduriga teisendatult
Neil päiwil ollid sured pitkad mehhed Ma peäl, ja ka pärrast sedda, kui Jumala lapsed innimeste tüttarte jure heitsid, siis töid nemmad neile lapsi ilmale; needsammad on need wäggewad, mis ammust aiast kuulsad mehhed on.	Neil *[päiwil+?] [ollid+?]* ~~sured~~ pitkad[7] *[mehhed+?]* ~~Ma~~ *[peäl+?]*, ja ka *[pärrast+?] [sedda+?]*, kui Jumala lapsed *[innimeste+?] [tüttarte+?] [jure+?]* heitsid, siis ~~töid~~ *[nemmad+?]* neile lapsi ilmale; *[needsammad+?]* on need *[wäggewad+?],* mis ammust ~~aiast~~ kuulsad *[mehhed+?]* on.
Kui Jehowa näggi, et innimeste kurjus suur olli Ma peäl, ja keik temma süddame möttette mötlemissed üsna kurjad iggapäwa: Siis kahhetses Jehowa, et ta innimest Ma peäle olli teinud, ja ta südda teggi haiget.	Kui *[Jehowa+?] [näggi+?]*, et *[innimeste+?]* kurjus suur *[olli+?]* ~~*Ma*~~ *[peäl+?]*, ja *[keik+?] [temma+?] [süddame+?] [möttette+?] [mötlemissed+?]* üsna kurjad *[iggapäwa+?]*: Siis *[kahhetses+?] [Jehowa+?]*, et ta *[innimest+?]* ~~Ma~~ *[peäle+?] [olli]* teinud, ja ta *[südda+?] [teggi+?]* haiget.
Ja Jehowa ütles: Ma tahhan innimest, mis minna ollen lonud, Ma peält ärrakautada, nihästi innimessed kui lojuksed ja romajad ja linnud, mis taewa al: sest ma kahhetsen, et ma neid ollen teinud.	Ja *[Jehowa+?]* ütles: Ma *[tahhan+?] [innimest+?]*, mis ~~minna~~ *[ollen+?] [lonud+?]*, ~~Ma~~ *[peält+?] [ärrakautada+?][nihästi+?] [innimessed+?]* kui *[lojuksed+?]* ja *[romajad+?]* ja linnud, mis *[taewa+?] [al+?]*; sest ma *[kahhetsen+?]* et ma neid *[ollen+?]* teinud.

[7] Morfoloogiamuunduris sõna „pitk“ tähenduses „pikk“, seetõttu on „pitkad“ aktsepteeritav vorm. Analüüsiga muundur väljendaks ka sõna seost sõnaga „pikk“ (vt peatükk 2.4).

C) Ortograafiamuunduriga teisendatult

Neil päevil olid [sured, suured] PIKAD mehed [Ma, Maa] peal, ja ka pärast seda, kui Jumala lapsed [INIMESTI, INIMESDE, INIMESTE] [TÜTARTI, TÜDARTI, TÜTARDE] [JURE, JUURDE] heitsid, siis tõid nemad neile lapsi ilmale; needsamad on need vägevad, mis [AMUST, AMMUST] ajast kuulsad mehed on.

Kui JEHOVA nägi, et [INIMESTI, INIMESDE, INIMESTE] kurjus suur oli [Ma, Maa] peal, ja [KEIK, KEIKK, KÕIK] tema südame [MÕTETI, MÕDETI, MÕTEDE, MÕTEDI, MÕTETE] mõtlemised üsna kurjad igapäeva: Siis kahetses JEHOVA, et ta inimest [Ma, Maa] peale oli teinud, ja ta süda TEGI haiget.

Ja JEHOVA ütles: Ma tahan inimest, mis mina olen [LONUD, LOONUD], [Ma, Maa] pealt ärakaotada, [NIHÄSTI, NIIHÄSTI, NIHESTI, NIIHÄSTI] inimesed kui lojused ja [ROMAJAD, ROOMAJAD] ja [LINUD, LIINUD, LINDNUD, LINNUD], mis taeva [AL, ALL], sest ma kahetsen, et ma neid olen teinud.

D) Kirjaviisimuunduriga teisendatult

Neil päevil olid [sured, suured] PITKAD mehed [Ma, Maa] peal, ja ka pärast seda, kui Jumala lapsed INIMESTE TÜTARDE JUURDE heitsid, siis tõid nemad neile lapsi ilmale; needsamad on need vägevad, mis AMMUST ajast kuulsad mehed on.

Kui *[JEHOWA+?]* nägi, et INIMESTE kurjus suur oli [Ma, Maa] peal, ja KÕIK tema südame MÕTETE mõtlemised üsna kurjad igapäeva: Siis kahetses *[JEHOWA+?]*, et ta inimest [Ma, Maa] peale oli teinud, ja ta süda [TEEGI, TEGI] haiget.

Ja *[JEHOWA+?]* ütles: Ma tahan inimest, mis mina olen LOONUD, [Ma, Maa] pealt ärakaotada, NIIHÄSTI inimesed kui lojused ja ROOMAJAD ja LINNUD, mis taeva ALL, sest ma kahetsen, et ma neid olen teinud.

Lisa 4: Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Ida Maria Orula**

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

Eesti vana ja uue kirjaviisi teisendus lõplike muunduritega

mille juhendaja on **Heiki-Jaan Kaalep**,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Ida Maria Orula

09.05.2019