Tartu Ülikool
Humanitaarteaduste ja kunstide valdkond
Eesti ja üldkeeleteaduse instituut

Laura Grant

Eesti keele niitkorpuse allkorpuste automaatne morfoloogiline analüüs ja ühestamine

Bakalaureusetöö

Juhendaja Kadri Muischnek

Tartu 2019

Sisukord

Sissejuhatus

Huvi ajalooliste tekstide vastu on ajaga aina suurenemas. Tänapäeval on veel paljud tekstid digitaliseerimata ning selle tõttu ei saa neist otsinguid teostada või tuleb seda teha käsitsi, mis on aeganõudev. Elektroonilisel kujul esinevatest ajaloolistest tekstidest on samuti informatsiooni kättesaamine keeruline, kuna need erinevad tänapäevasest keelekasutusest suurel määral ning seega ei anna tänapäeva keelele mõeldud tööriistad ka adekvaatseid tulemusi. (Pettersson 2016: 13)

Tänapäeva tekstidele mõeldud tööriistade mittetõhusa töötamise taga on mitmeid põhjuseid, millest on Michael Piotrowski (2012) enda raamatus „Natural Language Processing for Historical Texts" rääkinud. Esiteks on takistuseks vanade tekstide õigekiri, mis erineb sellest, millega tänapäeval ollakse harjunud. (Piotrowski 2012: 11) Vanemast eesti keelest võib näiteks tuua sõna *enämb,* mis jääb tänapäeva keele jaoks mõeldud tööriistadele arusaamatuks, kuid *enam* oleks arusaadav. Piotrowski (2012) toob veel välja, et tekstide tärktuvastamise (ehk OCR-i) käigus võib tekkida vigu, mis muudavad sõnu arusaamatuks. Peale selle võib tekstides kasutusel olla näiteks murdesõnu, sõnu mida tänapäeval ei kasutata või mille tähendus on muutunud. (Piotrowski 2012: 11–23)

Peale Piotrowski (2012) on samale teemale keskendunud ka Eva Pettersson (2016) enda doktoriväitekirjas „Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction", kus autor katsetas erinevaid normaliseerimise, lingvistilise analüüsi ja informatsiooni eraldamise meetodeid vanemat keelekasutust sisaldavate tekstide analüüsil.

Vanade vallakohtuprotokollide digitaliseerimisest ilmus selle aasta Rakenduslingvistide Ühingu Aastaraamatus artikkel „Möistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine", mille autoriteks on Maarja-Liisa Pilvik, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa ja Tõnis Türna. Selles artiklis kirjeldati 1866.–1890. aastate

vallakohtuprotokollidest digitaalse ressursi loomist, sealjuures arvestati keele mitmekesisusega. Kasutati EstNLTK vahendite komplektist automaatset morfoloogiaanalüsaatorit ja nimetuvastust. (Pilvik jt 2019)

Tänapäeva keele jaoks loodud tööriistadest ei piisa, et lemmatiseerimisel ja morfoloogilisel analüüsil häid tulemusi saada. Selleks, et seda parandada, on vaja eelnevalt tekste tänapäevase keele sarnasemaks töödelda. (Pilvik jt 2019: 139–140)

Selle bakalaureusetöö eesmärk ongi eesti kirjakeele niitkorpuse 1890.–1910. aastate allkorpuste tekstide morfoloogiline märgendamine. Selleks, et seda automaatselt teha, tuleb välja töötada tänapäeva kirjakeele normist hälbivate sõnade lemmatiseerimise põhimõtted. Peale selle on eesmärgiks koostada lisaleksikon, et parema tulemusega morfoloogilist analüüsi teostada ning viimaseks hinnata tulemuse kvaliteeti. Töö tulemusena loodud lisaleksikoni ja morfoloogiliselt märgendatud tekste on võimalik alla laadida Google Drive'ist[1].

Töö koosneb kolmest suuremast peatükist. Esimene kirjeldab eesti kirjakeele arengut ning 1890.–1910. aastatel loodud tekstide ortograafiat ja keelekasutust mõjutavaid tegureid. Teises peatükis kirjeldatakse töös kasutatavaid materjale ja meetodeid. Ning viimases peatükis antakse ülevaade töö eesmärkide täitmisest ning analüüsitakse tulemusi.

[1] Loodud lisaleksikon ja morfoloogiliselt märgedatud tekstid.
https://drive.google.com/file/d/1MapemjVJHEL8NqddYVZmJwAQ7U66J2em/view

1. Eesti kirjakeele ajalugu

Käesolevas töös kasutatud korpusetekstide publitseerimisaeg ulatub kuni 1890. aastateni. See on aga väga pikk aeg ning arvestades keeles toimuvaid muudatusi, ei saa kindlalt väita, et tänapäeva kirjakeeles kasutatakse samu sõnu või keelereegleid. Kirjakeel on mõjutatud erinevate aspektide poolt ning need ulatuvad üsna kaugele ajas tagasi.

Vanimad eestikeelsed kirjapanekud pärinevad 1220. aastatest, milleks olid esialgu koha- ja isikunimed, üksiksõnad ning lühikesed laused. Eesti kirjakeele alguseks loetakse 16. sajandi algust, mil pandi kirja esimesed eestikeelsed tekstid. (Raag 2008: 28)

Eesti keele arengu teeb omapäraseks see, et see on kujunenud kohalike murrakute ning murrete põhjal, mille baasil kujunes ühine rahvakeel, kuid seda eraldi Põhja-Eesti ja Lõuna-Eesti aladel. Keskusteks olid vastavalt Tallinn ja Tartu. Seega oli 16. sajandil, kirjakeele algusajal, Eestis kaks kirjakeelt: tallinna ja tartu keel. Kuigi kirjakeel sai alguse nende kahe keele põhjal, arenesid edasi ka üksikmurded ning see kajastub ka veel 19. sajandi tekstides. (Kask 1970: 175–176)

1.1. Ühise eesti kirjakeele loomine

Põhjus, miks eesti kirjakeel endiselt eraldi tallinna ja tartu keelena esines, oli see, et enamjaolt kasutasid murdekeelt suhtlemisvahendina talupojad, kes ei saanud vabalt liigelda. Kirjakeeles kasutasid seda sakslased, kelle keelekasutus oli vigane. See muutus aga 19. sajandi esimesel poolel, kui hakati tunnistama ühtse kirjakeele vajalikkust. (Kask 1970: 176)

Samal ajal toimus suur muutus eesti keele- ja kirjameeste seas. Seni pöörasid eesti keelele tähelepanu peamiselt sakslased, näiteks August Wilhelm Hupel ja Gustav Adolph Oldekop. Peagi ühinesid nendega ka Otto Wilhelm Masing ja Kristjan Jaak Peterson. Suuremat huvi eesti keele vastu näitas välja Pärnu Eliisabeti eesti koguduse pastor

Johann Heinrich Rosenplänter, andes aastatel 1813–1832 välja 20 raamatumahus ajakirja, mis kandis pealkirja „Beiträge zur genaurern Kenntniss der ehstnischen Sprache“. Tõlkes tähendab see „Lisandusi eesti keele lähemaks tundmaõppimiseks“. Ajakirja peamine eesmärk oli Eesti kirikuõpetajate eesti keele oskuse parandamine, kaudsem eesmärk aga eesti kirjakeele arendamine. (Raag 2008: 41–42) Rosenplänter kirjutas enda ajakirjas, et keel saab tugevaks kujuneda vaid ühe murde baasil ning kuna Eestis kasutatakse rohkem tallinna murret, siis peaks see jääma kirjakeeleks (Kask: 1970: 178).

1.2. Uue kirjaviisi loomine

Aja jooksul tekkis vajadus eestikeelsete oskussõnade järele. Esimesena lõi eestikeelseid oskussõnu Otto Wilhelm Masing, võttes kasutusele kirjeldavat laadi liitsõnu. (Raag 2008: 43) Näiteks *taevatundja* 'astronoom', *wallemõtlemine* 'eelarvamus', *poliktundja* 'võhik', *keeldut kaup* 'salakaup' (Kask 1970: 137–138). Peale Masingu lõi oskussõnu ka Rosenplänter. Muuhulgas kirjutas ta kaks muusikaõpikut, milles mõned tema loodud oskussõnad on kasutusel tänapäevalgi: *noot, pool noot, oktaw.* (Raag 2008: 44)

19. sajandi esimesel poolel hakati tähelepanu pöörama eesti keele kirjaviisile. Alates 17. sajandi lõpust olid kasutusel põhimõtted, mida praegu nimetatakse vanaks kirjaviisiks, kuid sellega polnud võimalik kõiki eesti keele häälikuid üles märkida. Seda märkasid sakslased Anton Thor Helle ja August Wilhelm Hupel. Nad tõid näiteks sõnad *hunt, noal* (*= nõel*), *öe* (*= õe*)*, sanna* (*= sõna*). Järelikult panid nad tähele peenendatud kaashäälikuid ja *õ*-häälikut. Samad tähelepanekud tegi ka Otto Wilhelm Masing ja aastatel 1820, 1824 ja 1827 avaldas ta sel teemal kolm saksakeelset brošüüri ning tutvustas eestlastele enda ideid ajalehes „Marahwa Näddala-Leht“. Tema ettepanekutest hakati kasutama vaid *õ*-tähte. (Raag 2008: 45–47)

Vana kirjaviis ja tartu keel hakkasid 19. sajandi teisel poolel hääbuma. Eesti kirjakeel täienes ilukirjandusteostega (F. R. Kreutzwald, L. Koidula, E. Vilde jt), mis mõjutasid eesti

kirjakeele normi. Hakati avaldama eestikeelseid erialaõpikuid, mille teemadeks olid näiteks eesti keele foneetika (Weske 1879) ja eesti keele grammatika (Hermann 1884). Eesti keeles õpetati mitmetes kihelkonnakoolides ning seda oli võimalik õppida õpetajate seminaris ja Tartu ülikoolis. Peale selle jõudis eesti keel ka teatrilavadele ning poliitikaski kasutati seda. (Laanekask 2004: 36–38)

Ajakirjas „Beiträge“ pakkus välja anonüümseks jäänud „A.“, kelleks arvatakse olevat soomlane Adolf Ivar Arwidsson, et kasutusele tuleks võtta uus, soomepärane kirjaviis. Selle põhimõte on, et lühikesed häälikud kirjutatakse ühe ning pikad häälikud kahe tähemärgiga. (Raag 2008: 47) Algul oli selle kirjaviisi pooldajaid vähe, kuid näiteks F. R. Kreutzwald oli üks neist, kes seda kaitses. Aastal 1872 pidas Jakob Hurt Eesti Kirjameeste Seltsi koosolekul ettekande, pärast mida hakati pooldama uut kirjaviisi. Seda hakati ka J. V. Jannseni ajalehes „Eesti Postimees“ kasutama. (Laanekask 2004: 38–39)

Lisaks uue kirjaviisi kasutuselevõtule otsustas selts veel viis keeleotsust vastu võtta. Ühendi *ea* asemel võeti kasutusele *ää.* Näiteks *pea* asemel tuli kasutada *pää* ja *hea* asemel *hää*. Hurt põhjendas enda ettekandes seda nii, et *ää* on vanem ning laiemalt levinud ning soome keeleski kasutati seda. Peagi tekitas see aga vastuseisu ning sellest loobuti lõplikult aastal 1953. (Raag 2008: 69) Siin töös uuritaval kirjakeele perioodil kasutati *ää*-d võrdlemisi palju, seega saab järeldada, et see otsus võeti esialgu hästi vastu.

Järgmiseks otsustati kasutusele võtta *sivad* lihtmineviku 3. isiku lõpu tunnuse *sid* asemel. Näiteks mitte *nemad olid ja elasid*, vaid *nemad olivad ja elasivad*. Pikema lõpuga taheti eristada ainsuse 2. pööret ja mitmuse 3. pööret (*sina* elasid – *nemad elasivad*) ning üheks argumendiks oli ka, et see on vana Eesti vorm ja selle tõttu tuleks seda kasutada. Siiski otsustati 1910. aastal *sivad* lõpust loobuda. (Raag 2008: 69) Siin töös kasutatavad eesti kirjakeele korpuse tekstid on *sivad* kasutuselevõtust tugevalt mõjutatud.

Järgmised keeleotsused on aga seni käibel. Esiteks otsustati kasutada oleviku 3. pöördes sama astet, mis 1. ja 2. pöördes. Näiteks vormide tema *luge* ja *nemad lugevad* asemel kasutatakse vorme *loeb* ja *loevad*. Teisena võeti kasutusele *da-* ja *ta*-liiteliste verbide umbisikulises tegumoes pikemad *tatakse, tatav, tatud* vormid. Näiteks *armastatakse, armastati, armastatav, armastatud*, mitte *armastakse, armasti, armastav, armastud*. Viimaseks püsima jäänud otsuseks on *h*-hääliku säilitamine sõna alguses. Seega oli õige kirjutada *irm* ja *äbi asemel hirm* ja *häbi*. (Raag 2008: 69–70) Siin töös keskendutud korpusetekstide uurimise põhjal võib öelda, et 1872. aastal vastu võetud otsustest viimased, mis tänapäevani kasutusel on, võeti väga hästi vastu, kuna vanu vorme enam ei esinenud.

1.3. Uue kirjaviisi täiendamine

Uus kirjaviis võeti üldiselt omaks, kuid sellele vaatamata kerkisid üles uued ettepanekud, kuidas kirjakeelt veelgi parandada. Mihkel Weske avaldas raamatu „Eesti keele healte õpetus ja kirjutamise wiis“ (1879), kus ta tegi ettepaneku eristada teist ja kolmandat väldet. Tema arvates tuleks kirjutada kolmanda välte häälikuid kolmekordse tähega: (*sõida*) *linnna*, (*astu*) *saaani*. Weske arvates peaks veel kahekordse tähega kirjutama kolmanda välte sõna diftongi teist osa: *lauda* (*juures*), (*istu*) *lauuda*. (Weske 1879: 3–5)

Õigekirja reegleid täpsustati veelgi. Juhan Kurrik avaldas 1886. aasta veebruarikuus ajalehtedes väitluse „Üleüldiselt pruugitav kirjaviis“, kus ta kutsus keeleõpetajaid, toimetajaid ja kirjanikke üles tema ettepanekute kohta arvamust avaldama. Tema peamine otsust oli loobuda Weske väldete märkimise süsteemist. Lisaks pakkus ta välja, et sõnalõpulised ülipikad konsonandid tuleks märkida kahekordse tähega: *kott, tamm, kätt*. Varem kirjutati need ühe tähega: *kot, tam, kät*. (Raag 2008: 74)

Uue kirjaviisi küsimustega hakati tegelema ka kirikuringkondades, kuna sooviti välja anda uues kirjaviisis piibel. Selleks, et keeleküsimuste üle nõu pidada, loodi Kirikukomisjon, mis tuli läbirääkimiste jaoks 26. ja 27. juunil 1886. aastal Tartus kokku.

(Raag 2008: 74–75) Vastu võetud otsused näitavad hästi seda, millises seisus oli keel sel ajal, kui kirjutati tekstid, mida siin töös analüüsitakse.

Vastu võetud otsused

Järgmisena kirjeldatakse 1886. aastal vastu võetud uue kirjaviisi otsuseid, mille protokolli pani kirja Hurt (1886) ning avaldas selle „Postimehe“ juulikuu esimeses numbris. Kõigepealt tuuakse välja otsused, mis on muutumatuna tänase päevani säilinud ning seejärel need, mida on täiendatud (Raag 2008: 77). Enim pööratakse tähelepanud nendele otsustele, mis mõjutasid 1890.–1910. aastate tekste.

Tänapäevani muutumatud otsused on järgmised:

- pikkade ja ülipikkade häälikute mitte eristamine (*soola* (magu) ja *soola* (panema), *laulu* (viis) ja *laulu* (laulma)), erandiks on sulghäälikud (*rada, ratas, rattad*);
- teise vokaali ees pikk i, ü ja u tuleb kirjutada kahe tähega (*luua – luud*);
- *ää* asemel *ea* kirjutamine (*pää – pea; hää – hea; säädma – seadma*);
- ühendite *üi, ie, uo, üö, õe* asemel tuleb kirjutada *üü, ee, oo, öö* ja *õõ* (*nüid – nüüd*; *süök – söök*);
- sõnalõpuliste konsonantide kahekordsete tähtedega märkimine (*kep – kepp*; *tam – tam*m);
- *ki*-liite kirjutamine helitute häälikute järele ja *-gi* heliliste häälikute järele (*siiski, tammgi*);
- nimede suure algustähega kirjutamine;
- sõnade *sada*, *sõda* ja *koda* ainsuse omastava vorm tuleb kirjutada *saja, sõja*, *koja* (mitte *saa, sõa, koa*). (Hurt 1886: 1–2)

Hiljem muudetud otsused:

- sõnad *talitama, amet, ometi, seni, kuni* tuleb kirjutada kahekordse konsonandiga (*tallitama, ammet*);
- sõnad *auu* ja *nõuu* tuleb nimetavas käändes kirjutada *au* ja *nõu*, teistes käänetes kahesilbiliselt – *auust*, *nõuuga*;

- kasutada adverbides *ste*-liidet *-sti* asemel (*ilusaste, hoolsaste*). (Hurt 1886: 1–2)

Uue kirjaviisi otsused kajastuvad ka siin töös kasutatavates eesti kirjakeele korpuse tekstides. Kui võrrelda eri kümnendeid, siis saab vaadelda, kuidas on otsused kirjakeelde vastu võetud, või vastupidi – kuidas neid aina vähem kasutati.

1.4. Venestuse algus

Siin töös analüüsitavates tekstides esineb ka 19. sajandi lõpus toimunud venestuse mõjusid. Raag (2008) kirjutab oma raamatus, et aastal 1864 alustati Poolas ja Leedus venestamisega. Selleks keelustati neis riikides poola ja leedu keele kasutamine. Poolas vallandati 14 000 ametnikku. Peale selle ei tohtinud Leedus ladina tähestikku kasutades raamatuid trükkida, vaid need tuli trükkida vene tähtedega. Need keelud kehtisid kuni aastani 1904. (Raag 2008: 82)

Aastal 1870 avaldas Riia Vaimuliku Seminari õpetaja Peeter Mihkelson eestlastele mõeldud vene keele õpiku. See tundus esmapilgul tavaline, sisaldades endas vene tähestikku, vene häälduse õpetust ja lugemipalu koos tõlgetega. Lõpus oli kaks eestikeelset teksti, mis olid trükitud vene tähtedega. Arvatakse, et Mihkelson püüdis selle õpiku kirjutamisega alustada üleminekut vene tähestikule või ta täitis sellega lihtsalt ametivõimude käsku. (Raag 2008: 82–83)

Venestus algaski aastatel 1882–1883, kui senaator Nikolai Manassein läks Liivi- ja Kuramaale revisjoni tegema. Selle käigus kutsuti rahvast üles enda murekohtadest rääkima ning umbes 44 000 inimest seda võimalust ka kasutas. Kirjades nõuti näiteks aadli ja kirikuõpetajate mõju kaotamist ja suuremaid õigusi eesti keelele. Senaator esitas tsaarile selle kohta aruande, kus ta kirjutas, et mõisnike võim on liiga suur ning seda kasutati venestuse elluviimiseks. (Raag 2008: 83)

Venestuse mõjusid said tunda kõik. Ametiasutustes hakati senise saksa keele asemel kasutama vene keelt ning töötajateks said seal umbkeelsed venelased. Kõik, kes vene keelt ei osanud, pidid ametist lahkuma. Aastast 1887 õpetati koolis ainult vene keeles,

lapsed ei tohtinud ka omavahel eesti keeles rääkida. Erandiks olid luteri koolid, kus õpetati siiski kahte ainet eesti keeles – usuõpetust ja emakeelt. (Raag 2008: 84)

Venestuse mõjutused

Venestus on ka eesti keelele mõju avaldanud. Vene keelest on kasutusele võetud näiteks sõnad *türann, labürint, süsteem* ja *tsüklon.* Esialgu oli nende kirjakuju küll veidi teistsugune, kuna neid kirjutati venepäraselt. Võõrapärane joon oli, et *ü* ja *ö* asemel kirjutati *i* ja *e*. Näited vastavalt: *tirann, labirint, sisteema* ja *tsiklon.* Tänapäevane vorm tuleneb sellest, et 20. sajandi alguses hakati võõrsõnade kirjutamist ühtlustama. (Raag 2008: 85–86)

Peale selle tuli eesti keelde ka mitmeid vene laene. Näiteks *kultuura, ragulka* ja *sutt* (veidi) (Raag 2008: 86). Ka eesti kirjakeele korpuses esineb vene laene, näiteks sõna *kabak*, mida EKSS-i andmetel kasutati seda vanasti kõrtsi tähenduses. Välja paistis ka sõna *sobor*, mis oli analüüsitud tekstides kasutuses Venemaa kiriku- või riigitegelaste koosoleku tähenduses (VSL).

2. Materjalid ja meetod

Käesolevas bakalaureusetöös on kasutatud eesti kirjakeele niitkorpuse[2] 1890.–1910. aastate tekste, mis sisaldavad kokku 1 144 940 sõna. Eesti kirjakeele korpus sisaldab eelkõige ilu- ja ajakirjandustekste, sest nendest koosneb eesti kultuuris tekstide põhimass, mille põhjal kehtestatakse normingute põhikohti. Peale selle kogutakse neist ka kirjakeele näiteid. (Hennoste, Muischnek 2000: 189)

Korpuses olevate ilukirjandustekstide tuumaks on realistlik ilukirjandus. Ajakirjandustekstides aga puuduvad grammatilised ja leksikaalsed erijooned, seega võib viimaseid pidada oma keelekasutuse poolest kirjakeele tüüpilisimaks esindajaks. Põhjus, miks ilukirjandustekstid keelekasutust edasi ei anna, on see, et need sõltuvad teksti autorist ja kirjandusvoolust. Ent ajakirjandustekstide keelekasutus muutub kiiresti, sest need kajastavad ühiskonnamuutuseid. (Hennoste, Muischnek 2000: 189–190)

Morfoloogilise märgendamise teostamiseks tuleb tekstid morfoloogiliselt analüüsida ning seejärel ühestada (Pilvik jt. 2019: 148). Morfoloogilist analüüsi teostatakse selles töös UNIX-i keskkonnas OÜ Filosofti reeglipõhise analüsaatoriga *etana,* mis määrab iga sõna algvormi, sõnaliigi ning grammatilise vormi nimetuse (ESTMORF). Saadud tulemust ühestatakse *etyhh*'iga, mille tööpõhimõte on statistiline, mistõttu see ei kohane tundmatute sõnavormidega. Kasutatavad programmid on Vabamorfi[3] morfoloogilise analüsaatori ja ühestaja UNIX-i käsurea versioonid. (Heiki-Jaan Kaalep 2019: isiklik suhtlus) Kuna tegemist on 1890.–1910. aastate tekstidega, siis pole automaatne analüüs täpne. Selle parandamiseks tuleb tekstid eelnevalt töödelda tänapäevase keelekasutusele sarnasemaks, ehk normaliseerida. Tundmatuks jäänud sõnad tuleb aga lisada lisaleksikoni, mille abil saab anda soovitud analüüsi programmi *etana* leksikonis puuduvate sõnade vormidele. Enne leksikoni koostamist teostatakse *etana* abil

[2] Eesti Kirjakeele Korpus. https://www.cl.ut.ee/korpused/baaskorpus/
[3] Vabamorf. https://github.com/Filosoft/vabamorf

1890.–1910. aastate tekstide morfoloogiline analüüs ilma oletamiseta, et luua tundmatute sõnade loend, mille abil luua lisaleksikon. Järgmise sammuna analüüsitakse tekstid uuesti *etana*ga, kasutades valminud lisaleksikoni ning morfoloogilist ühestajat *etyhh*. Rakendatakse ka oletajat, mis annab seni tundmatuks jäänud sõnadele analüüsi. Morfoloogiliselt märgendatud korpuse loomise viimane etapp on märgenduse kvaliteedi hindamine. Kõikidest protsessi osadest räägitakse lähemalt peatükis 3.

2.1. Tekstikorpused

Tekstikorpus on elektrooniline tekstikogu, mis on koostatud kindlatel eesmärkidel ja konkreetsete printsiipide alusel, et see aitaks iseloomustada keele seisundit (Hennoste, Muischnek 2000: 185). Korpustel on kolm erinevat põlvkonda. Esimesi korpuseid koostati sel ajal, kui arvutimälu oli piiratud, mistõttu tuli hoolikalt kaaluda, milliseid tekste sinna lisada. Tähtis oli läbi mõelda, mis eesmärgid korpusel olema hakkavad. Teise põlvkonna korpused on tuhandeid miljoneid sõnu sisaldavad korpused, mille puhul polnud enam oluline, et iga tekstiklass oleks võrdselt esindatud. Kolmanda põlvkonna korpus koosneb internetitekstidest, näiteks foorumipostitused ja kommentaariumid, mis on internetist automaatselt korpusesse salvestatud. (Muischnek 2015: 37–38)

2.2. Eesti kirjakeele korpus

Siin töös kasutatakse esimese põlvkonna korpust, mille tekstid on digitaliseeritud käsitsi, arvutisse trükkimise teel. Tekstide valikul on Tiit Hennoste lähtunud järgmistest kriteeriumitest:

- tekstid on ainult ametlikud või avalikud;
- tekstid on ainult emakeelsed – tõlketekstid on välja jäetud;
- tekstid on kirjalikul kujul ning lugemiseks määratud (välja on jäetud tekstid, mis on kirjalikud, kuid kuulamiseks määratud, ning mis on määratud kuulamiseks, kuid mis pole kirja pandud);
- ainult trükitud tekstid;

- ettevalmistatud ning redigeeritud tekstid (välja on jäetud spontaansed tekstid);
- ainult proosatekstid (välja on jäetud luuletekstid);
- ainult täisealiste autorite kirjutatud tekstid;
- vaid Eestis ringelnud tekstid (välja on jäetud väliseestlaste kogukondades ringelnud tekstid);
- ainult tekstide esmatrükid (Hennoste, Muischnek 2000: 186–187).

3. Morfoloogiline analüüs ja ühestamine

Selle bakalaureusetöö eesmärk on eesti keele niitkorpuse 1890.–1910. aastate allkorpuste morfoloogiline märgendamine. Kuna tegemist on ajalooliste tekstidega, siis tänapäeva kirjakeele analüüsiks mõeldud morfoloogiaanalüsaator ei tule sellega ise piisavalt hästi toime ning selleks, et adekvaatseid tulemusi saada, tuleb tekste normaliseerida ning luua lisaleksikon, kuhu on lisatud analüsaatorile tundmatud sõnad.

See peatükk koosneb neljast osast, milles kolmes esimeses antakse ülevaade eesti kirjakeele korpuse tekstide märgendamisprotsessist ning viimases hinnatakse tulemuse kvaliteeti.

3.1. Normaliseerimine

Kuna ajalooliste tekstide sõnavara ning kirjaviis erinevad tänapäeva keelekasutusest, siis ei anna tänapäeva tekstidele loodud tööriistad piisavalt häid tulemusi. Selleks, et seda parandada, on vaja ajaloolised tekstid enne töötlema hakkamist normaliseerida. (Piotrowski 2012: 69) Siinkohal illustreeriks normaliseerimise vajalikkust järgmised näited, mis on eesti kirjakeele korpuse 1900. ja 1910. aasta allkorpuste ajakirjandustekstidest pärinevad laused (EKK).

(1) AJA1900\aja0001 Miks see *kõik nõnda* on ?

(2) AJA1910\pl0039 Jällegi *waenlaße* aeroplan .

Neist näidetest jääb morfoloogianalüsaatorile tundmatuks üsna mitu sõnavormi. Nendeks on *kõik, nõnda, waenlaße.* Nende näidetega on põhilised normaliseerimist vajavad kohad esindatud.

Kuna enne UTF-8 kodeeringu kasutuselevõttu polnud võimalik tagada kõikide täpitähtede ning *š* ja *ž* tähe korrektset kuvamist kõigis brauserites, tuli need arvutisse sisestada SGML-olemitena[4]. Kuna SGML-olemid esinevad ka uuritavas korpusetekstides,

[4] SGML-olemid. https://www.cl.ut.ee/korpused/segakorpus/olemid/index.php?lang=et

tuleb need nüüd teisendada UTF-8 kodeeringus märkideks. Näitelause 2 põhjal *õ -> õ.* Seega näiteks sõnavormist *kõik* saab *kõik ja nõnda* teisendatakse sõnaks *nõnda*. Veel, näites 2 on näha, et on kasutatud *w* ja *ß*-tähte, mis tuleb teisendada vastavalt *v* ja *s*-ks. Tekstis leiduvad üleliigsed analüüsile mittekuuluvad sümbolid (kõik kirjavahemärgid, jutumärgid, valuutamärgid, kriipsud) eemaldatakse samuti, kuna nende analüüs pole lisaleksikoni loomiseks vajalik.

Järgmine samm on tundmatute sõnade jaoks lisaleksikoni loomine. Selleks, et teada saada, millised sõnad tundmatuks jäävad, tuleb normaliseeritud tekstid ilma oletamiseta morfoloogiliselt analüüsida. Enne seda peab tekstid viima sellisele kujule, et analüsaator analüüsiks vaid vajalikke osi. Rea alguses olevatele allikaviidetele tuleb automaatselt ümber kirjutada märgendid *<ignore>* ja *</ignore>*. Samuti tuleb iga lause algusesse ja lõppu automaatselt kirjutada märgendid *<s>* ja *</s>*. Ebavajalikuks osutusid ka teksti sisestamisel väljajäetud tekstiosi märkivad <gap> märgendid. Kõige selle tulemusel peaks eelmised näited 1 ja 2 enne morfoloogilise analüüsi teostamist olema sellisel kujul:

(3) <ignore> AJA1900\aja0001 </ignore> <s> Miks see *kõik nõnda* on </s>

(4) <ignore> AJA1910\pl0039 </ignore> <s> Jällegi *vaenlase* aeroplan </s>

Järgmisena tuleb kasutades morfoloogianalüsaatorit *etana* tekstid analüüsida ning iga allkorpuse tundmatutest sõnadest koostada sagedusloend, mille alusel neid lisaleksikoni lisada. Seda kirjeldatakse lähemalt järgmises alapeatükis.

3.2. Lisaleksikoni loomine

Siin bakalaureusetöös tehtav morfoloogiline analüüs on leksikonipõhine. Mis tähendab, et seni analüsaatorile tundmatuks jäänud sõnad on ette antud sõnastikuna (Pilvik jt 2019: 148). Sõnastikku lisati kolm või rohkem korda esinenud analüsaatorile tundmatud sõnavormid selle põhimõtte alusel, et nende lemma ehk algvorm oleks võimalikult muutmata kujul, kuid siiski tänapäevases keeles arusaadav (Pettersson 2016: 49–50). Parema tulemuse nimel tehti seda käsitsi. Järgnevalt tuuakse koos näidetega välja peamist tüüpi sõnad, mis sõnastikku lisati. Siin esinevates näidetes ja ühtlasi ka lisaleksikonis on esimene sõna sellisel kujul, nagu see tekstis oli ning sellele järgneb selle sõna lemma+lõpp, sõnaliik ja grammatilise kategooria märgend sellisena, nagu see esitati lisaleksikonis (ESTMORF).

Sõnavormile lemma määramisel lähtuti põhimõttest, et kui sõnavormi on võimalik analüüsida tänapäeva keeles olemasoleva sõna ortograafilise või häälduslikuna variandina, siis nii ka tehti. Siia rühma kuuluvad vanast kirjaviisist või häälduslikust variandist tulenev kahekodsete tähtede kasutamine, ühendi *ea* asemel *ää* kasutamine, kohanimede kirjutamine, *au* ja *nõu* kirjutamine, ühendite *üü, ee, oo, öö, õõ* asemel *üi, ie, uo, üö* ja *õe* kirjutamine, *ste*-liite asemel *sti*-liite kasutamine.

Teise rühma kuuluvad sõnad, mille tuletusliide (või selle puudumine) on erinev tänapäeva keeles kasutatavast. Sellisel juhul määrati lemma vastavalt kasutatud liitele. Muutmata jäeti näiteks *line*-liite abil moodustatud sõnad, mille asemel tänapäeval kasutatakse *lik*-liidet. Lisaleksikoni lisati ka need sõnavormid, mis olid vana kirjakeele lihtmineviku 3. isiku tunnusega *sivad*.

Esmalt tuuakse välja need tekstisõnad, millel lähtuti tänapäevasest ortograafilisest variandist. Vanast kirjaviisist tingituna kasutati kahekordset tähte teisiti. Korpuses oli kahekordse tähega kirjutatud sõnu, mida tänapäeval kirjutatakse ühe tähega ja vastupidi. Varieeruvused esinevad, kuna sellel perioodil, mil kirjutati siin töös uuritavad

tekstid, toimusid olulised keelevaidlused ning keele kasutajal polnud kerge neis orienteeruda (Raag 2008).

(5) hunikusse hunnik+sse //_S_ sg ill, //

(6) kommisjon komisjon+0 //_S_ sg n, //

(7) dessatini tessatin+0 //_S_ adt, sg g, sg p, //

(8) lotterii loterii+0 //_S_ sg g, sg n, //

Kohanimede kirjutamine. Siin lähtuti Eva Petterssoni (2016) põhimõttest, et kohanimed kirjutatakse nii, nagu neid tänapäevalgi kirjutatakse (Pettersson 2016: 50).

(9) Daani Taani+0 //_H_ adt, // Taani+0 //_H_ sg g, sg n, //

(10) Stokholmis Stockholm+s //_H_ sg in, //

Järgmiseks mõned selle töö esimeses peatükis kirjeldatud 1886. aastal vastu võetud keeleotsuste mõjutused. Esimeseks sõnad, millel kirjutati ühendi *ea* asemel *ää.*

(11) sääduse seadus+0 //_S_ sg g, //

(12) pääminister pea_minister+0 //_S_ sg n, //

Sõnade *au* ja *nõu* kirjutamine. Algselt kirjutati need kahekordse *u*-ga, et viidata nende kahesilbilisele häälduselе (Raag 2008: 75).

(13) auuhind au_hind+0 //_S_ sg n, //

(14) nõuukogu nõu_kogu+0 //_S_ sg g, sg n, sg p, //

Adverbides *sti*-liite asemel *-ste* kasutamine. Seda otsust hiljem muudeti, kuid see mõjutas uuritavat kirjakeele perioodi siiski.

(15) auusaste ausasti+0 //_D_ //

(16) halvaste halvasti+0 //_D_ //

Sõnad, mille kirjutamise kohta võeti aastal 1886 vastu otsus teisiti kirjutada, kuid vana vorm polnud veel uuritavatest tekstidest välja juurdunud. Varem kasutatud diftongiga variant oli omane põhjaeesti keskmurde häälduselе (Raag 2008: 75).

(17) püiab püüd+b //_V_ b, //

(18) rõemsa rõõmus+0 //_A_ sg g, //

Peale selle esines veel hulk sõnu, mille ortograafiline variant oli tänapäevasest erinev.

(19) prinz prints+0 //_S_ sg n, //

(20) klischee klišee+0 //_S_ sg g, sg n, //

Järgmiseks kirjeldatakse seda rühma, mille tuletusliite kasutamine oli erinev tänapäeva omast. Vaatamata sellele, et tänapäeval on sarnane sõna olemas, määrati nendele sõnadele see lemma, millega need tekstis esinesid.

(21) liblik liblik+0 //_S_ sg n, //

(22) tubak tubak+0 //_S_ sg n, //

(23) demokratline demo_kraatline+0 //_A_ sg n, //

(24) hariduslise haridusline+0 //_A_ sg g, //

(25) aitasivad aita+sivad //_V_ sid, //

(26) lükkasivad lükka+sivad //_V_ sid, //

Peale eelnimetatud põhimõtete, otsustati lühendid lisaleksikoni lisada lahti kirjutatuna, et need oleks informatiivsemad.

(27) hra härra+0 //_S_ sg g, sg n, //

(28) cand kandidaat+0 //_S_ sg n, //

(29) snt sent+0 //_S_ adt, sg p, //

Lisaleksikoni loomise juures tuleb hea tulemuse tagamiseks jälgida, et sõnavormid ei saaks üleliigseid analüüse. Näiteks vanas kirjakeeles esinenud sõna *sääl*, mis tänapäeval on adverb *seal*, sisestatakse morfoloogilisse analüsaatorisse tänapäevase vormina ning selle väljund on järgmine.

(30) seal
seal+0 //_D_ //
siga+l //_A_ sg ad, //
siga+l //_S_ sg ad, //

Tulemusest on näha, et adverb *seal* on saanud veel ka nimisõna ja omadussõna alalütleva analüüsi. Seda sellepärast, et tänapäeval on sõna *seal* mitmetähenduslik. Seega siinkohal tuleks alles jätta vaid esimene analüüs, kuna vajalik on ainult adverbi *sääl* analüüs, ning kuna *sääl* ei kasutatud vanas kirjakeeles mitmetähenduslikuna, siis see ei saa ka muud tüüpi sõna olla. Käsitsi üle kontrollides sellised vead eemaldatakse.

Selliseid põhimõtteid silmas pidades lisati kõikide 1890.–1910. aastate allkorpuste tundmatud sõnad lisaleksikoni. Kokku saadi 1914 sõna, millega on võimalik tutvuda alla

laadides fail Google Drive'ist[5]. Samal aadressil asub ka lõplik versioon morfoloogiliselt märgendatud tekstidest, millest räägitakse lähemalt järgmises peatükis.

3.3. Korpuse morfoloogiliselt märgendatud lõpliku versiooni loomine

Lõpliku versiooni loomiseks ühendati algne korpusetekst korpuse teisendatud versiooniga. Algse korpuseteksti puhul tehti vaid SGML-olemite teisendus UTF-8 kodeeringu märkideks. Teisendatud versiooni tarvis teostati tekstidele peatükis 3.1 kirjeldatud normaliseerimine ning seejärel tehti uus analüüs kasutades lisaleksikoni, morfoloogilist ühestajat ja tundmatute sõnade oletajat. Lõpliku tulemuse saamiseks võeti algsest korpusetekstist (SGML-olemite teisendustega) sõnad sellisel kujul nagu nad tekstis esinesid ning liideti need kokku teise versiooni analüüsidega. Selle tulemusel esinevad peatükis 3.1 toodud näited 1 ja 2 lõplikus verisoonis sellisel kujul:

(31) Miks miks+0 //_D_ //
see see+0 //_P_ sg n, //
kõik kõik+0 //_P_ pl n, // kõik+0 //_P_ sg n, //
nõnda nõnda+0 //_D_ //
on ole+0 //_V_ b, // ole+0 //_V_ vad, //
? ? //_Z_ //

(32) Jällegi jällegi+0 //_D_ //
waenlase vaenlane+0 //_S_ sg g, //
aeroplan aeroplan+0 //_S_ sg n, //
. . //_Z_ //

Näidetes 31 ja 32 on näha, et analüüs on üldiselt edukalt teostatud. Olemid on teisendatud UTF-8 märkideks ning lemmades puuduvad *w* ja *ß.* Puuduseks võib märkida selle, et sõnavormi *aeroplaan* lemma peaks olema *aero_plaan*. Järgmises alapeatükis analüüsitakse tekkinud vigu lähemalt.

[5] Loodud lisaleksikon ja märgendatud korpustekstid.
https://drive.google.com/file/d/1MapemjVJHEL8NqddYVZmJwAQ7U66J2em/view?usp=sharing

3.4. Tulemuse hindamine

Eesti keele niitkorpuse tekstide analüüsiks kasutati morfoloogianalüsaatorit *etana* ja ühestajat *etyhh.* Kuna suur osa tundmatutest sõnavormidest esines vaid üks kuni kaks korda, siis ei pidanud töö autor mõistlikuks neid käsitsi lisaleksikoni lisada. Seega rakendati oletajat, mis määras nende analüüsi ise, kuid seda kasutades võivad analüüsi tegemisel sisse tulla vead.

Tabel 1 esitab informatsiooni iga kümnendi aja- ja ilukirjandusteksti suuruse, kontrollitud sõnade hulga ning vigade arvu kohta. Esimesel real on märgitud iga allkorpuse sõnade koguarv. Teisel real on kontrollimiseks võetud sõnade arv. Igast allkorpusest võeti analüüsimiseks 250 sõna, ümardatuna täislauseteni. Järgmised read kajastavad seda, kui täpne oli morfoloogiline analüüs lisaleksikoni ja ühestamisega. Kolmandal real on õigete analüüside arv, sealhulgas ka mitmesed analüüsid mille hulgas on õige. Järgmisel real on need analüüsid, mis olid mitmesed ning mille hulgas polnud õiget analüüsi. Viimaseks on täiesti vale analüüsi saanud sõnad.

Tabelis 1 on näha, et kontrollitud sõnade arv oli 1520 ning neist vale analüüsi sai 65 sõna ehk 4,3% kontrollitud sõnadest. Õigete analüüside protsent on seega 95,7%, mille hulka arvestati ka analüüsid, mis jäid mitmeseks, kuid mille hulgas oli ka õige analüüs. Kui arvestada õigeteks analüüsideks ainult üheseid õigeid analüüse, siis on nende osakaal 85,6%. Võrreldes tulemust tundmatute sõnade osakaaluga tänapäeva kirjakeele normile vastavates tekstides, milleks on 2,58% ja neist 86,97% on suure algustähega ehk enamik on pärisnimed (Pilvik jt 2019: 148–149), siis võib öelda, et siin töös saavutatud tulemus on hea. Aga kuna kontrolliti väikest hulka lauseid, siis selleks, et korrektsemaid tulemusi näha, tuleks võtta kontrollimiseks suurem hulk sõnu (Pettersson 2016: 74–75).

Tabel 1. Sõnade arv kokku, kontrollitud sõnad, vale analüüsiga sõnad

Sõnade arv	**Allkorpus**						
	1890. aasta		**1900. aasta**		**1910. aasta**		
	Ajakirj.	**Ilukirj.**	**Ajakirj.**	**Ilukirj.**	**Ajakirj.**	**Ilukirj.**	**Kokku**
Sõnade arv korpuses	193 000	155 000	206 983	188 262	214 131	187 564	**1 144 940**
Kontrollitud sõnade arv	261 100%	246 100%	260 100%	249 100%	249 100%	255 100%	**1520** **100%**
Õige analüüs	229 87,7%	220 89,4%	210 80,8%	217 87,1%	204 81,9%	221 86,7%	**1301** **85,6%**
Mitmene analüüs koos õigega	24 9,2%	14 5,7%	35 13,5%	24 9,6%	33 13,3%	23 9,0%	**153** **10,1%**
Vale analüüs	8 3,1%	12 4,9%	15 5,8%	7 2,8%	12 4,8%	11 4,3%	**65** **4,3%**

Tabel 2 esitab informatsiooni märgendamisel tekkinud vea hulkadest ning vea tüüpidest. Esimesel real on kokku loetud analüüsid, kus on sõnale vale algvorm määratud. Neid esines 11, mis on 16,9% kõikidest vigadest. Kõikidel juhtudel oli tekkinud viga oletamise käigus.

(33) kumardawad kumarda+vad //_V_ vad, //

Näites 33 on algvormis topelt *m* puudu.

Järgmise vea tüübi alla koondati need vead, kus oli valesti määratud sõnaliik ja/või grammatiliste kategooriate märgend. Selliseid vigu oli pisut rohkem kui eelmiseid – 14 sõnas ja kõikidest vigadest 21,5% olid seda tüüpi. Sellel juhul oli näha, et vea tegi morfoloogiline ühestaja, mis valis konteksti mitte sobiva variandi.

(34) sul sina+l //_P_ sg ad, //

tänu tänu+0 //_K_ //

toowad too+vad //_V_ vad, //

lehkawad lehkav+d //_A_ pl n, //

lilled lill+d //_S_ pl n, //

Lauses 34 pole *tänu* mitte kaassõna funktsioonis, vaid tähistab nimisõna *tänu* partitiivi.

Tabel 2. Vea tüübid.

Vea tüüp	**Allkorpus**						
	1890. aasta		**1900. aasta**		**1910. aasta**		**Kokku**
	Ajakirj.	**Ilukirj.**	**Ajakirj.**	**Ilukirj.**	**Ajakirj.**	**Ilukirj.**	
Vale algvorm	0	0	4	0	3	4	**11 16,9%**
Vale sõnaliik ja/või grammatiliste kategooriate märgend	2	3	4	2	1	2	**14 21,5%**
Valed nii algvorm kui ka grammatiliste kategooriate märgend	6	8	6	4	2	2	**28 43,1%**
Vale w-> v teisendus	0	0	0	0	2	1	**3 4,6%**
Mitmene analüüs, mille hulgas pole õiget	0	1	1	0	2	0	**4 6,2%**
Trükivead	0	0	0	1	2	2	**5 7,7%**
Kokku	**8 12,3%**	**12 18,5%**	**15 23,1%**	**7 10,8%**	**12 18,5%**	**11 16,9%**	**65 100%**
Kümnendite kaupa kokku	**20 30,8%**		**22 33,8%**		**23 35,4%**		**65 100%**

Tabelis järgmisel real olevat vea tüüpi esines kõige rohkem – 28 sõnas ja see moodustas 43,1% kõikidest vigadest. Siia tüübi alla loeti need sõnad, millel olid valed nii algvorm kui ka grammatiliste kategooriate märgend.

(35) *Tahetavat Tahe=tav+t //_H_ sg p, //*
nimelt nimelt+0 //_D_ //
kevadel kevad+l //_S_ sg ad, // kevade+l //_S_ sg ad, //
iseäralist ise_äraline+t //_A_ sg p, //
" " //_Z_ //
puu-istutamise puu-istutamine+0 //_S_ sg g, //
pidu pidu+0 //_S_ sg n, //

Näites 35 on näha, et oletaja on teinud vea *Tahetavat* lemmat määrates, märkides selle nimeks, kuna see on suure algustähega. Tegelikult on aga tegu verbiga ning õige analüüs oleks sellel järgmine: *taht+tavat //_V_ tavat, //*. Näites 35 on ka teist tüüpi viga, nimelt *pidu* on saanud vale grammatilise kategooria. Selles lauses peaks see olema partitiivis, seega analüüsiga *pidu+0 //_S_ sg p, //.*

Järgmise veana käsitleti normaliseerimise käigus üleliigset *w* teisendamist *v*-ks. Neid vigu esines kõige vähem – vaid kolm (4,6% kõigist) ning kõigil esinenud juhtudel oli tegu nimedega – New York ja Wismar. Selle vältimiseks oleks pidanud *w* teisendamise käigus eirama nimesid, kus *w* ka tänapäeval on.

Eelviimane vea tüüp tabelis on mitmese analüüsiga sõnad, mille hulgas pole õiget analüüsi. Ka selle vea esinemine on väike – neljal korral, ehk 6,2% kõikidest vigadest.

(36) *Kohtuministeriumi Kohtuministerium+0 //_H_ sg g, // Kohtuministeriumi+0 //_H_ sg g, //*
juurde juurde+0 //_K_ //
on ole+0 //_V_ b, // ole+0 //_V_ vad, //
nõupidamise nõu_pidamine+0 //_S_ sg g, //
komisjon komisjon+0 //_S_ sg n, //
asutatud asuta+tud //_V_ tud, // asuta=tud+0 //_A_ // asuta=tud+0 //_A_ sg n, // asuta=tud+d //_A_ pl n, //

Näitelauses 36 on mitmene, kuid ilma õige analüüsita sõnavorm *Kohtuministeriumi*. Siin eeldas oletaja, et kuna see sõnavorm algab suure tähega, siis on tegu nimega. Õige analüüs oleks sellele nimisõnaanalüüs *kohtu_ministeerium+0 //_S_ adt, sg g, //.*

Viimasena on tabelis 2 määratud trükivead, mida esines kontrollitud lausetes viis korda, ehk kõikidest sõnadest 7,7%.

(37) mmailmast mma_ilm+st //_S_ sg el, //

Näites 37 on näha, et sõna algusesse sattunud kahekordne täht on juhuslik. Vaatamata sellele on oletaja siiski õige sõnaliigi ja grammatiliste kategooriate märgendi määranud.

Tabelis 2 olnud vigade tüübi järgi saab öelda, et üldine märgendamise tulemus on hea. Vead jagunesid kolme uuritud kümneni vahel üsna võrdselt. Enamik vigu on tekkinud selle tõttu, et oletaja ei suuda ajalooliste mõjutustega sõnu oletada. Et seda parandada, saaks normaliseerimistehnikaid veel edaspidi arendada.

Kokkuvõte

Bakalaureusetöö eesmärk oli eesti kirjakeele niitkorpuse 1890.–1910. aastate tekstide morfoloogiline märgendamine, et ajaloolistest tekstidest oleks võimalik kiiremini ja mugavamalt informatsiooni kätte saada. Selleks, et tekste automaatselt märgendada, oli eesmärgiks luua kirjakeele normist hälbivate sõnade lemmatiseerimise põhimõtted. Peale selle oli eesmärgiks koostada lisaleksikon, et parema tulemusega morfoloogilist analüüsi teostada ning viimaseks hinnata tulemuse kvaliteeti. Töö autori hinnangul kõik eesmärgid täideti.

Bakalaureusetöö annab ülevaate eesti kirjakeele ajaloost kuni perioodini, milleni siin töös keskenduti, ehk kuni 20. sajandi alguseni. Kirjeldati 19. sajandil tekkinud vajadust ühise ja uue kirjakeele järele. Sealjuures anti ülevaade sellest, milliseid keeleuuendusi võeti vastu ning kuidas aja jooksul kirjakeelt täiendati. Anti ka põgus ülevaade 19. sajandi lõpus alustatud venestusest ning venestuse mõjutustest eesti keelele.

Töö teine osa tutvustas töös kasutatavaid materjale ja meetodeid. Lähemalt räägiti eesti kirjakeele korpusest ning mis põhimõtete alusel on sinna tekstid valitud. Meetodite osa juures kirjeldati morfoloogianalüsaatorit ja ühestajat. Anti ülevaade korpusetekstide märgendamisprotsessist.

Morfoloogilise analüüsi ja ühestamise peatükk keskendus töö eesmärgile. Kirjeldati normaliseerimisvõtteid ning lisaleksikoni koostamise põhimõtteid. Viimaseks hinnati morfoloogilise märgendamise kvaliteeti. Jõuti arusaamani, et morfoloogiline märgendamine oli edukas, kuid, et tulemusi veelgi paremaks saada, tuleks normaliseerimistehnikaid arendada. Lisaks saaks siin töös püstitatud eesmärke edasi arendada, kui märgendada kogu eesti kirjakeele korpus.

Kirjandus

EKK = Eesti kirjakeele korpus. https://www.cl.ut.ee/korpused/baaskorpus/. Vaadatud 03.10.2018.

EKSS = Eesti keele seletav sõnaraamat. https://www.eki.ee/dict/ekss/. Vaadatud 16.05.2019.

ESTMORF = Morfoloogilise analüsaatori ESTMORF kasutamine. http://www.filosoft.ee/html_morf_et/morfoutinfo.html. Vaadatud 25.05.2019.

Hennoste, Tiit, Kadri Muischnek 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Toim. Tiit Hennoste. Tartu: Tartu Ülikooli kirjastus, 183–217. https://dspace.ut.ee/handle/10062/41671. Vaadatud 21.03.2019.

Hermann, Karl August 1884. Eesti keele Grammatik. Koolide ja iseõppimise tarwis kõikidele, kes Eesti keelt õigesti ja puhtasti kõnelema ja kirjutama ning sügawamalt tundma ja uurima tahawad õppida. Tartu: Wilhelm Just.

Hurt, Jakob 1886. Püha kiri pannakse uue kirjutuswiisi järele ümber. – Postimees nr 28, heinakuu 5, lk 1–2, Tartu. https://dea.digar.ee/cgi-bin/dea?a=d&d=postimeesew18860705&e=-------et-25--1--txt-txIN%7ctxTI%7ctxAU%7ctxTA----------. Vaadatud 02.05.2019.

Kaalep, Heiki-Jaan 2019. Suuliselt suhtlus (28.05).

Kask, Arnold 1970. Eesti kirjakeele ajaloost I–II. Tartu: Tartu Riiklik Ülikool.
Laanekask, Heli 2004. Eesti kirjakeele kujunemine ja kujundamine 16.–19. sajandil. Tartu: Tartu Ülikooli Kirjastus. http://dspace.ut.ee/bitstream/handle/10062/1138/Laanekask.pdf?sequence=5&isAllowed=y. Vaadatud 03.05.2019.

Muischnek, Kadri 2015. Keelekorpused – sama mitmekesised kui keel ise. – Oma Keel 1, 37–44. http://www.emakeeleselts.ee/omakeel/2015_1/OK_2015-1_05.pdf. Vaadatud 21.03.2019.

Loodud lisaleksikon ja märgendatud korpustekstid. https://drive.google.com/file/d/1MapemjVJHEL8NqddYVZmJwAQ7U66J2em/view?usp=sharing. Vaadatud 29.05.2019.

Pettersson, Eva 2016. Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction. Uppsala: Uppsala Universitet. http://uu.diva-portal.org/smash/get/diva2:885117/FULLTEXT01.pdf. Vaadatud 10.05.2019.

Pilvik jt 2019 = Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa, Tõnis Türna 2019. *Möistus sai kuulotedu*: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. – Eesti Rakenduslingvistika Ühingu aastaraamat, 15, 139–158. http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa15.08. Vaadatud 02.05.2019.

Piotrowski, Michael 2012. Natural Language Processing for Historical Texts. Germany: Morgan & Claypool Publishers.

Raag, Raimo 2008. Talurahva keelest riigikeeleks. Tartu: Atlex. http://dspace.ut.ee/bitstream/handle/10062/34490/raag_talurahvakeelest.pdf. Vaadatud 15.03.2019.

Vabamorf. https://github.com/Filosoft/vabamorf. Vaadatud 29.05.2019.

VSL = Võõrsõnade leksikon. https://www.eki.ee/dict/vsl/. Vaadatud 20.05.2019.

Weske, Mihkel 1879. Eesti keele healte õpetus ja kirjutuse wiis. Tartu: Schnakenburg. https://www.digar.ee/arhiiv/et/raamatud/14260. Vaadatud 24.05.2019.

Morphological analysis and disambiguation of the Corpus of Written Estonian. Summary

The aim of this Bachelor's thesis was to provide part-of-speech tagging for the Corpus of Written Estonian during the years 1890 to 1910. In order to do it automatically, the main part of the work was regarding the principles of lemmatization which can be used to analyze contemporary written language that is falsely written. Additionally, a user lexicon was compiled, which contains word-forms otherwise not recognized by morphological analyzer together with their lemmas, part-of-speech tags and grammatical category annotations, so the morphological analysis would be better.

The current bachelor thesis consists of three main parts. The first part introduces the relevant background information: the development of Written Estonian, with main emphasis on the second half of the 19th and the very beginning of 20th century.

The aim of this thesis is to annotate morphologically the corpora of Written Estonian dating from periods 1890–1899, 1900–1909 and 1910–1919. The second part of the thesis gives an overview of these corpora and the principles underlying their compilation.

The third part of the thesis reports on the measures undertaken for providing accurate morphological annotation of these corpora containing older Written Estonian, namely normalization and compilation of a special lexicon.

The goal of the thesis was achieved. The main contributions of this thesis are the following.

User lexicon consisting of 1914 words was created that enables to perform accurate morphological analysis of word-forms that would otherwise be not recognized by the morphological analyzer *etana.* All unknown word-forms occurring three or more times in one corpus were added to the lexicon.

Morphologically annotated versions of the corpora of Written Estonian from the periods 1890–1899, 1900–1909 and 1910–1919 were created using the aforementioned lexicon.

The correctness of morphological annotation was evaluated. 85,6% of tokens had received unambiguous correct analysis and 95,7% of the words-forms correct annotation, that could also be ambiguous. Most errors were caused by guesser, that was used to give analysis to out-of-vocabulary wordforms occurring two or one times in one corpus.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Laura Grant,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Eesti keele niitkorpuse allkorpuste automaatne morfoloogiline analüüs ja ühestamine", mille juhendaja on Kadri Muischnek, 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.

Laura Grant

30.05.2019