Tartu Ülikool
Humanitaarteaduste ja kunstide valdkond
Eesti ja üldkeeleteaduse instituut

Kertu Saul

ESTNLTK MORFOLOOGILISE ANALÜSAATORI JA ÜHESTAJA KVALITEEDI HINDAMINE

Bakalaureusetöö

Juhendaja Siim Orasmaa

Tartu 2022

Sisukord

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud lähtudes Tartu Ülikooli eesti ja üldkeeleteaduse instituudi lõputöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Sissejuhatus

Siinses töös on fookuses EstNLTK[1] teegi koosseisus olev morfoloogiline analüsaator ja ühestaja Vabamorf[2]. Morfoloogiline analüsaator leiab sõnade algvormi, sõnaliigi, morfoloogilise struktuuri ja kategooriad ning annab neile vastavad märgendid. Ühestaja leiab tekitatud analüüsidest õige. Eesti keel on morfoloogiliselt väga rikas keel, mistõttu on morfoloogiline analüüs ja ühestamine eesti keele puhul esimene etapp igasugusele tekstitöötlusele. Seetõttu kasutatakse seda programmi näiteks õigekirjakontrollija töös, kõnesünteesis, kõrgema taseme keeletehnoloogiamoodulite implementeerimisel, kõnetuvastuses ja lingvistilises uurimistöös[3].

Viimased põhjalikumad selleteemalised uurimistööd tehti 2001. (Kaalep, Vaino 2001) ja 2008. (Veskis, Liba 2008) aastal. Leiti, et programm on mõneti ebatäpne: 2001. aasta uuringu põhjal saab sõna õige analüüsi 98% kordadest (Kaalep, Vaino 2001) ja 2008. aasta uuringu sõnul 96,23% kordadest (Veskis, Liba 2008). Kõige rohkem probleeme oli nimisõnade analüüsiga, eriti morfoloogiliselt mitmeste sõnade ja jutumärkide tuvastusega (Veskis, Liba 2008). Vahepeal on programmi närvivõrkudel põhineva morfoloogilise analüsaatori väljatöötamiseks põgusalt hinnatud (Tkachenko, Sirts 2018; Leman 2019; Milintsevich 2020; Laur jt 2020), aga põhjalikumat hindamist pole 2008. aastast saati tehtud, mistõttu tuleks programmi uuesti hinnata. See aitab leida rakenduse kitsaskohti, mida saab tulevikus edasi arendada.

Töö eesmärk on hinnata EstNLTK morfoloogilise analüsaatori ja ühestaja kvaliteeti kõige uuematel käsitsi märgendatud korpustel, et leida programmi vead, mille parandamisele peaks tulevikus keskenduma. Lisaks annab kvaliteedi hindamine programmi kasutajate parema idee, mis probleeme selle kasutamisel ette võib tulla. Eesmärgist lähtudes on sõnastatud järgmised uurimisküsimused.

[1] https://github.com/estnltk
[2] https://github.com/Filosoft/vabamorf
[3] https://www.keeletehnoloogia.ee/et/ekt-projektid/vabavaraline-morfoloogiatarkvara. Vaadatud 26.04.2022.

1. Kui hästi Vabamorf ja selle konfiguratsioonid eri tekstiliikide peal töötavad?
2. Millise kvaliteediga määrab programm lemmasid, sõnaliike ja vorme?
3. Mis on Vabamorfi tüüpvead?

Tuginedes varasematele uurimustele (Kaalep, Vaino 2001, Veskis, Liba 2008), võib oletada, et kõige rohkem probleeme on homonüümsete sõnade nominatiivi, genitiivi, partitiivi ja aditiivi eristamisel. Tekstiliikide lõikes on halvimad tulemused ilmselt netikeelsetel tekstidel.

Analüsaatori kvaliteedi hindamiseks võrreldakse tekste, mille on inimene käsitsi ja arvuti automaatselt märgendanud. Tekstid saadakse Eesti *Universal Dependencies* (de Marneffe jt 2021) puudepanga korpustest, mis jaotuvad EDT korpuseks[4] ja EWT veebitekstide korpuseks[5]. Käsitsi tehtud märgendused teisendatakse automaatmärgendusega võrdlemiseks sobivale kujule. Märgendused, mille arvuti on teinud inimesest erinevalt, liigitatakse vigadeks. Neist tehakse sagedusanalüüsid tekstiliikide lõikes.

Bakalaureusetöö on jaotatud kolmeks osaks. Esimeses osas antakse ülevaade arvutimorfoloogiast, kirjeldatakse Vabamorfi, selle kvaliteedi varasemaid uurimusi ja kuidas taolisi programme hinnatakse. Teises osas kirjeldatakse andmeid, kuidas tulemusteni jõuti ja mis hindamismeetrikaid kasutati. Kolmandas osas analüüsitakse programmi kvaliteeti eri aspektidest ja tuuakse välja rakenduse kitsaskohad.

[4] https://github.com/UniversalDependencies/UD_Estonian-EDT. Vaadatud 01.06.2022.
[5] https://github.com/UniversalDependencies/UD_Estonian-EWT. Vaadatud 01.06.2022.

1. Teooria

1.1. Arvutimorfoloogia

Morfoloogia jaguneb kaheks osaks: sõna- ja vormimoodustuseks. Sõnamoodustus keskendub liitmise või tuletamise abil täistähenduslikest sõnadest uute täistähenduslike sõnade loomisele (*elama – elu*), aga vormimoodustus näitab, kuidas ühele täistähenduslikule sõnale tunnustega uus grammatiline tähendus antakse (*elama – elan*). (Erelt jt 1995: 42) Arvutianalüüsis keskendutakse mõlemale morfoloogia tasandile (Kaalep 2014).

Sõnad koosnevad morfeemidest, mis võivad olla kas tunnused, liited või tüved. Tunnused muudavad grammatilist tähendust, nt käänet, pööret, ajavormi või arvukategooriat (*õpetaja*+***te***+***le***). Sõnas koos esinevad tunnused arvestatakse kokku sõna formatiiviks (*õpetaja*+***tele***). Liited muudavad sõna tähendust ja/või sõnaliiki, näiteks tüvest *ela* saab tekitada liite *sime* abil uue tähendusega sõna *elasime*. Tüved avavad sõna leksikaalse tähenduse. Vormimoodustuses on tüvi sõnaosa, mis eelneb tunnustele (***õpetaja***+*le*), sõnamoodustuses aga koosneb see eraldi juurmorfeemist ja liidetest (***õpe***+*ta*+*ja*). (EKK 2020: 178) Morfoloogilise analüüsi sisend on sõna ning väljund selle algvorm ja morfeemid (Viks 2000: 11–12).

Arvutimorfoloogia esimene etapp on sõnastikupõhine morfoloogia. Selles etapis leitakse morfoloogilise analüüsi väljund sõnastikke ja morfeemide liitmise eeskirju kasutades. (Muischnek jt 2012: 69) Sõnastikku kuuluvad algvormid ja nende tüved, morfoloogilised formatiivid, nende tähendused ja erandid. Algvormide ja tüvede juures on kirjeldatud nende omadusi ja seda, kuidas need erinevates sõnavormides käituvad. (Viks 2000: 13) Hea morfoloogilise analüüsi alused on niisiis küllalt mitmekesine, aga samas lihtsa esituse ja struktuuriga sõnastik ning sõnamoodustuse reegleid hästi tundev algoritm (Muischnek jt 2012: 69).

Kui põhjalikud sõnastikud aga ka ei oleks, on päris keelekasutus ikka rikkalikum. Seetõttu tuleb alati ette juhtumeid, mida sõnastikus pole kirjeldatud, sest keel on püsivas arengus. Selliste sõnade analüüsimiseks rakendatakse järgmist morfoloogilise analüüsi etappi: reeglipõhist morfoloogiat, mis määrab tundmatule sõnale grammatikareeglite põhjal analüüsid. (Muischnek jt 2012: 70)

1.1.1. Ühestamine

On sõnu ja sõnavorme, mis on sama kujuga, aga võivad erinevas kontekstis erinevaid asju tähendada. Sellistele sõnadele tekitatakse mitu analüüsi, sest kontekstita ei ole võimalik aru saada, missugune neist on õige. (Muischnek jt 2012: 71–72) Probleemi näitlustamiseks võib välja tuua eesti keele koondkorpuse[6], mille umbes 245 miljonit sõnast saab mitu analüüsi 45% (Kaalep jt 2010: 1–2).

Konteksti tuvastamiseks ja selle abil analüüside vähendamiseks kasutatakse ühestamist. Selle jaoks üritab ühestaja lokaalse konteksti järgi tuvastada, missugune analüüs on õige. Näiteks lauses *Tee on valmis* võib eeldada, et *tee* on siin lauses nimisõna, mitte tegusõna, sest sõna läheduses on juba olemas tegusõna *on*. (Muischnek jt 2012: 72)

Eesti keele jaoks on välja arendatud nii statistilisi kui ka reeglipõhiseid ühestajaid (Muischnek jt 2012: 72). Vabamorfi oma kuulub statistiliste ühestajate alla, mis põhineb Markovi varjatud mudelil. Selle jaoks koostatakse esmalt käsitsi ühestatud tekstide põhjal tabelid, kuhu kantakse märgendite esinemise tõenäosused. Nende arvutamisel arvestatakse tõenäosust, et märgend on lauses esimene ja tõenäosust, et ühele märgendile eelneb teine. Mitmese sõna puhul valib ühestaja nende tabelite põhjal välja kõige tõenäolisema märgendi. (Kaalep, Vaino 1998: 30–31, 33) Kontekstina arvestab mudel kolme lähima vasakule jääva sõna konteksti (Kaalep jt 2010: 144; Muischnek jt 2012: 72).

On aga olemas ka ühestamise viis, mis arvestab sellest palju laiemat konteksti: korpusepõhine ühestamine. Sel juhul valitakse mitme analüüsiga sõnade seast välja see

[6] https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et

analüüs, mida korpuses kõige rohkem oli. See töötab põhimõttel, et sarnastes tekstides kasutatakse sarnaseid sõnu. (Kaalep jt 2012: 85) Näiteks kui aiandustekstide korpuses saab sõna *teod* lemma analüüsideks nii *tigu* kui ka *tegu*, aga varasemalt on mitmel korral räägitud tigudest, on analüüs *tigu* tõenäolisem ja valitakse see.

1.2. Eelnevad uurimused

Eelnevalt on EstNLTK morfoloogilise analüsaatori ja ühestaja kvaliteeti hinnatud mitmel korral. 2001. ja 2008. aastal olid need puhtalt reeglipõhise morfoloogilise analüsaatori ja statistilise ühestaja[7] uurimused (Kaalep, Vaino 2001; Veskis, Liba 2008), aga 2018. aastast alates kõrvutati neid uute neurovõrkudepõhiste süsteemidega (Tkachenko, Sirts 2018; Leman 2019; Milintsevich 2020; Laur jt 2020).

Kaalepi ja Vaino 2001. aasta artikkel annab ülevaate morfoloogilise analüüsi igast sammust: lausete üksteisest lahutamine, sõnaraamatupõhine morfoloogiline analüüs, tundmatute sõnade arvamine, morfoloogiline ühestamine. Sõnaraamatupõhise analüüsiga sai selles uurimuses õige tulemuse keskmiselt umbes 98% sõnadest, sealhulgas kirjavahemärgid, sagedased lühendid ja pärisnimed ning mitmest sõnast koosnevad võõrkeelsed nimed. Haruldasi pärisnimesid, lühendeid, termineid, akronüüme, slängi jne EstNLTK selle tehnikaga aga ei analüüsi. Olenevalt teksti žanrist moodustasid sellised sõnad kuni 3% tekstist. Nende analüüsiks tuleb kasutada oletamist. See võtab arvesse sõna lõpu, silpide arvu ja sõna vormi ning vaatab, kas teatud tunnuste alusel saaks sõna kuhugi lahterdada. Näiteks kui sõna on kaks tähte pikk või sealt puudub täishäälik, võib tegu olla lühendiga. (Kaalep, Vaino 2001)

Oletamise etapis annab morfoloogiline analüüs konteksti arvestamata kõik sõna võimalikud analüüsid (Kaalep, Vaino 2001: 2). Näiteks lause *Mees peeti kinni* väljund[8] näeb välja selline:

[7] Varasemalt tuntud kui ESTMORF ja ESTYHMM (Muischnek jt 2012: 71)
[8] https://www.filosoft.ee/html_morf_et/morfoutinfo.html

(1) Mees
    mees+0 //_S_ sg n, //
    mesi+s //_S_ sg in, //
  peeti
    peet+0 //_S_ adt, sg p, //
    pida+ti //_V_ ti, //
  kinni
    kinni+0 //_D_ // (Kaalep, Vaino 2001: 2)

Võib aga juhtuda, et ka mitme analüüsi seast puudub õige variant. Selliseid juhte oli 2001. aasta seisuga 0,4%. (Kaalep, Vaino 2001: 4)

Kui analüüsid on kätte saadud, liigutakse ühestamise etappi. Seal valitakse võimalike variantide seast välja üks, mis on programmi arust kõige tõenäolisem. Näiteks jäävad näite 1 analüüsidest pärast ühestamist alles ainult õiged:

(2) Mees
    mees+0 //_S_ sg n, //
  peeti
    pida+ti //_V_ ti, //
  kinni
    kinni+0 //_D_ // (Kaalep, Vaino 2001: 5)

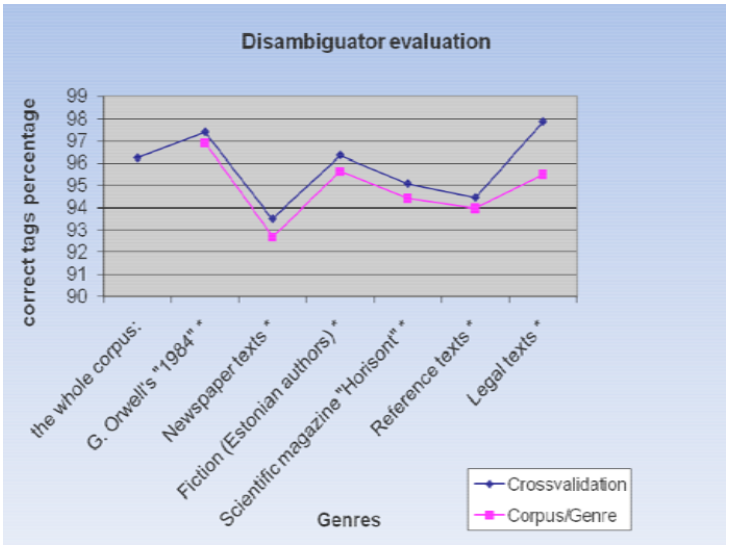
Keeruliste juhtumite puhul jäetakse vigade vältimiseks sõnale kõik analüüsid alles. Selliseid sõnu on tekstis umbes 13,5% (Kaalep, Vaino 2001: 6). Suurimad probleemsed rühmad on:

1. nud- ja tud-partitsiip, sest kontekstita on võimatu öelda, kas tegemist on tegu- või omadussõnaga
2. asesõna *ta*, mille sõnakuju ei erista nominatiivi ja genitiivi
3. tegusõna *on*, mis võib olla ainsuses ja mitmuses

4. *kui*, mis võib olenevalt kontekstist olla määr- või sidesõna
5. *mis* ja *kes*, mis võivad olla nii ainsuses kui ka mitmuses
6. homonüümid
7. *üks* ja *teine*, mis võivad olla arv- ja asesõnad (Kaalep, Vaino 2001: 6)

2008. aastal keskenduti just morfoloogilise ühestaja ESTYHMM kvaliteedi uurimisele. Selle jaoks kasutati Tartu Ülikooli morfoloogiliselt ühestatud korpust[9], kus algse morfoloogilise analüüsi tootis ESTMORF ja nende tulemuste seast valisid lingvistid õige käsitsi välja. Seejärel võrreldi seda käsitsi ühestatud korpust ESTYHMM-i väljundiga. (Veskis, Liba 2008)

Kui ühestajat treeniti vastavalt hiljem analüüsitavale žanrile, oli ühestamise täpsus keskmiselt 96.23%, kui aga kogu korpuse põhjal, siis 94.86%. Kvaliteet kõigub žanrite kaupa siiski päris korralikult. (Veskis, Liba 2008: 4)



Joonis 1. **Õige analüüsi saanud sõnade protsent žanrite kaupa** (Veskis, Liba 2008: 6)

Graafikul näitab ülemine joon žanripõhist treenimist ja alumine kogu korpuse põhist treenimist. Halvima tulemuse saavutasid ajakirjandustekstid, mis jäid mõlema lähenemise põhjal umbes 93% täpsuse juurde. (Veskis, Liba 2008)

[9] https://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=et

Eelneval hindamisel raporteeriti, et ühestamise tulemusel valitakse vale vorm 3% kordadest (Kaalep, Vaino 2001: 6), 2008. aastal aga olenevalt eelpool kirjeldatud ühestaja treeningu põhimõttest keskmiselt 3,77–5,14% kordadest (Veskis, Liba 2008: 4). Kõige suurem murelaps on seal kohal nimisõnade vormihomonüümia, kus sõna nominatiiv, genitiiv, partitiiv või aditiiv on samasugused. (Kaalep, Vaino 2001: 6) Selline sõna on näiteks *kool*, mille ainsuse genitiiv, partitiiv ja aditiiv on kirjapildiga *kooli*. Kui 2001. aastal moodustasid sellised sõnad valedest analüüsidest kolmandiku (Kaalep, Vaino 2001: 6), siis 2008. aastaks oli see arv pea kahekordistunud 61% peale (Veskis, Liba 2008: 5).

Lisaks nimisõnadele tekitasid 2008. aastal palju probleeme pärisnimed ja jutumärgid. Neist olid jutumärgid märgendaja kõige levinum viga. See vale analüüs seisnes lihtsas programmiveas, kus rakendus ei tundnud ühte spetsiifilist jutumärkide stiili ära, mistõttu oli ajakirjandustekstide ühestamise kvaliteet võrdlemisi madal. (Veskis, Liba 2008: 5)

Järgmised hindamised toimuvad alles kümme aastat hiljem, aga teistsuguses kontekstis. Alates 2018. aastast on tehtud mitmeid töid tehisnärvivõrkudel põhineva morfoloogilise analüsaatori arendamiseks, mille käigus võrreldi *Universal Dependencies* (edaspidi UD) korpust Vabamorfiga. Kuna uurimuste tegelik eesmärk oli aga tehisnärvivõrkudel põhineva süsteemi arendamine, pole neis Vabamorfi hindamisele erilist rõhku pandud. (Tkachenko ja Sirts 2018; Leman 2019; Milintsevich 2020)

2020. aastal võrreldi neurosüsteeme kasutavat mitmekeelset loomuliku keele analüüsi vahendit StanfordNLPd (Qi jt 2018) ja EstNLTK uut (v1.6.4b) ja vana versiooni (1.4.1). Seal hinnati umbes 10 miljoni sõnaga korpuse peal juhuslikku väljavõtet märgendajate erinevustest. Kui süsteemid sõna samamoodi märgendasid, jäeti see analüüsist välja, kuigi kummagi süsteemi märgendus ei pruukinud õige olla. Kõigi viie tekstiliigi seast valiti juhuslikult välja 20 erineva analüüsi saanud sõna ja need vaadati hiljem käsitsi üle. (Laur jt 2020: 7156–7157)

Leiti, et uus EstNLTK 1.6 versioon on parem nii StanfordNLP-st kui ka oma vanast versioonist nii sõnade ja lausete segmenteerimisel kui ka morfoloogilisel analüüsil. Peamiselt tekkis neurovõrkudel põhineval StanfordNLP-l probleeme lemmatiseerimisega, mis EstNLTK-le seevastu raskusi ei valmistanud. Lisaks sellele uuriti eri ühestajate tööd sõnaliikide ja vormide peal ja leiti, et EstNLTK tavalise ühestaja ja EstNLTK uue korpusepõhise ühestaja väljundid olid väga sarnased, aga korpusepõhine oli siiski täpsem. Kui tavalise ühestajaga sai sajast sõnast õige analüüsi 37, siis korpusepõhisega 63 sõna. (Laur jt 2020: 7157–7158)

1.3. Kategooriasüsteemide üldised erinevused

Töös hinnatakse EstNLTK teeki kuuluva Vabamorfi tööd. Selleks aga, et Vabamorfi automaatmärgendust oleks millegagi võrrelda, on vaja andmestikku, kus päris inimesed on tekstile märgendused lisanud. Seda kutsutakse kuldstandardiks, sest eeldatakse, et inimesel on alati õigus. Taoline andmestik on olemas, aga see on UD süsteemis, kus märgendite kuju erineb mitmes aspektis Vabamorfi omas. Selles peatükis tuuaksegi need erinevused välja. Kategooriate võrdlevaid tabeleid saab vaadata lisas 1.

Verbivormidel on UD süsteemis iga verbi kirjeldava aspekti (kõneviis, aeg, tegumood, isik, arv, verbivorm) kohta eraldi märgend (Müürisep 2021), Vabamorfi süsteemis iseloomustab märgend aga ainult verbi lõppu (Kaalep 1996: 55–58). Näiteks saab kindla kõneviisi oleviku 3. isiku ainsuse aktiivi jaatavas kõnes sõna *loeb* UD (a) ja Vabamorfi (b) süsteemides järgmised märgendid:

a) Mood=Ind, Tense=Pres, Person=3, Number=Sing, Voice=Act, VerbForm=Fin
b) b

See teeb teisendused keeruliseks, sest mitu erineva tähendusega verbi kasutavad sama lõppu, nii et palju informatsiooni läheb kaduma. Selles aspektis on UD märgendus palju täpsem.

Asesõnades on UD korpuses eraldi märgendid erinevate pronoomeni tüüpide jaoks: personaal-, refleksiiv-, possesiiv-, retsiprook-, demonstratiiv-, interrogatiiv-relatiiv-, determinatiiv-, indefiniit- ja terviklikkust näitav pronoomen (Müürisep 2021). Vabamorfi süsteemis seevastu erinevaid pronoomeni tüüpe ei eristata, vaid kõik saavad ainult pronoomeni sõnaliigi märgendi (Kaalep 1996: 49). See tähendab, et teisendused tehakse ainult sõnaliigi põhjal.

Käänete puhul on teisendused lihtsad, sest mõlemas süsteemis on märgendid 15 käändele. Nende hulgas on eesti keelde ametlikult kuuluvad 14 käänet ning lisaks aditiiv ehk lühike illatiiv. Kuigi mõlemad süsteemid saavad käänete lühendid nende rahvusvahelistest nimedest, on nad seda teinud erinevalt. Näiteks UD süsteemis on genitiiv *Gen*, aditiiv *Add* ja inessiiv *Ine*, Vabamorfis aga vastavalt *g*, *adt* ja *in*. (Müürisep 2021; Kaalep 1996: 51)

Sõnaliikide teisendustega esineb rohkem raskusi. On kolm klassi sõnu, mis UD süsteemis olemas on, aga Vabamorfi omas mitte: alistavad sidesõnad (*SCONJ*), sümbolid (*SYB*) ja muu (*X*) (Müürisep 2021; Kaalep 1996: 49). Sümbolite alla kuuluvad emotikonid, matemaatilised operaatorid, valuutamärgid, meili- ja veebiaadressid jms. Muu märgendit kasutatakse mõnel koodivahetuse juhul. (Müürisep 2021)

On ka üks sõnaliik, mis on Vabamorfi süsteemis oma märgendi saanud, aga UD süsteemis mitte: genitiivatribuut (Kaalep 1996: 49; Müürisep 2021). Selle teisenduse peamine probleem on see, et genitiivatribuut võib olla nii nimisõna, asesõna kui ka omadussõna (Erelt jt 1993: 120), mistõttu pole see üheselt mingi teise kategooria alla liigutatav.

Mitut sõnaliiki iseloomustatakse Vabamorfi süsteemis ühe märgendina (Kaalep 1996: 49), aga UD omas kahe eri tulbas paikneva märgendi kombinatsioonina (Müürisep 2021). Nendeks on eri võrdes omadussõnad, lühendid ja põhi- ning järgarvud. Näiteks saab ülivõrdes omadussõna UD süsteemis märgenditeks *ADJ* ja *Degree=Sup*, aga Vabamorfis ainult *U*.

1.4. Hindamismeetrikad

On erinevaid viise morfoloogiliste analüsaatorite hindamiseks. Kõige levinumad neist on korrektsus, saagis ja täpsus. (van Halteren 1999: 81) Korrektsust on võimalik mitut moodi defineerida (van Halteren 1999: 82; Olson, Delen 2008: 138). Van Halteneri esitatud definitsiooni järgi mõõdab korrektsus, kui palju sõnu said õige märgendi. (van Halteren 1999: 82)

$$korrektsus = \frac{\text{õige märgendusega sõnad}}{\text{kõik sõnad}}$$

Seda kasutatakse pigem selliste süsteemide hindamiseks, kus on vähene või olematu mitmesus ehk üks sõna saab ainult ühe analüüsi. Kui korrektsust kasutatakse mitmesustega süsteemide hindamiseks, esitatakse lisaks sellele ka keskmine analüüside arv sõna kohta. Mitmesustega sõnade hindamiseks sobivad aga paremini saagis ja täpsus. (van Halteren 1999:82) Nende arvutamiseks jaotatakse analüüsid nelja erinevasse klassi:

a) True positive (TP) - sõnal peab mingi analüüs olema ja see on sinna ka lisatud
b) False positive (FP) - sõnale määrati automaatselt analüüs, mida seal tegelikult ei tohiks olla. Siin loetakse kokku kõik üleliigsed analüüsid ehk juhud, kui:
    i) sõnal peab olema analüüs, aga automaatne analüüs on vale
    ii) sõnal ei pea olema analüüsi, aga talle antakse automaatselt analüüs
c) True negative (TN) - sõnal ei pea analüüsi olema ja automaatselt jäetigi see määramata.
d) False negative (FN) - sõnale jäeti õige analüüs määramata. Kui sõnal on mitu valet analüüsi, läheb arvesse ainult esimene. (Faaß jt 2010: 805)

Nende kategooriate põhjal koostatakse korrektsuse, täpsuse ja saagise valemid (Olson, Delen 2008: 138):

$$täpsus = \frac{TP}{TP + FP}$$

$$saagis = \frac{TP}{TP + FN}$$

$$korrektsus = \frac{TP + TN}{TP + TN + FP + FN}$$

Valemitest avaldub, et täpsust mõjutavad üleliigsed variandid ehk tavaliselt kehtib tendents, et mida madalam on täpsus, seda rohkem on analüsaator sõnale analüüse alles jätnud. Saagist aga üleliigsed variandid ei mõjuta, sest seal otsitakse, kas kõigi analüüside seas oli vähemalt üks õige. Seetõttu on kõrge saagise saavutamiseks süsteemil mõttekas jätta sisse ka need variandid, mille õigsuses ta päris kindel ei ole. Nende kahe meetrikate kombineerimisel saab hea ülevaate analüüside arvu ja nende õigsuse suhtest, mida tehakse F1-skoori abil. (Derczynski 2016: 262)

$$F1\text{-}skoor = 2 * \frac{täpsus * saagis}{täpsus + saagis}$$ (Olson, Delen 2008: 138)

Varasemates EstNLTK morfoloogilise analüsaatori uuringutes on kasutatud erinevaid hindamismeetrikaid. EstNLTK esimeses uurimuses on hindamismeetrika mainimata jäetud (Kaalep, Vaino 2001). Veskis ja Liba (Veskis, Liba 2008) kasutasid EstNLTK ühestaja hindamiseks ainult korrektsust. Tkachenko ja Sirts on Vabamorfi väljundi hindamisel kasutanud korrektsust, aga neuromudelite jaoks F1-skoori (Tkachenko, Sirts 2018). Lemmatiseerimisele keskenduvates Milintsevichi (Milintsevich 2020) ja Lemani (Leman 2019) töödes kasutas Milintsevich korrektsust (Milintsevich 2020: 25) ja Leman täpsust (Leman 2019: 31).

Siinkohal on aga oluline täpsustada, et arvutamise valem oli esitatud ainult Milintsevichi lemmatiseerimist uurivas töös, kus kasutati järgmist valemit (Milintsevich 2020: 25):

$$korrektsus = \frac{õigete\ analüüside\ arv}{lemmade\ koguarv}$$

See definitsioon läheb aga paremini kokku eelnevalt mainitud van Halteneri korrektsuse definitsiooniga (van Halteren 1999: 82) kui sellega, mis siin töös kasutatakse (Olson, Delen 2008: 138). Teises lemmatiseerimist käsitlevas töös oli kõiki tulemusi täpsuseks kutsutud, kuigi vähemalt ühe etapi hindamise kirjeldusest avaldus, et tegemist on pigem saagise kui täpsusega (Leman 2019: 31). Ülejäänud varasemate tööde (Kaalep, Vaino 2001; Veskis, Liba 2008; Tkachenko, Sirts 2018; Laur jt 2020) tulemusi vaadates jäi mulje, et ka neis on kasutatud arvutamiseks van Halteneri korrektsuse definitsiooni (van Halteren 1999: 82), mis läheb tegelikult väga hästi kokku selle töö saagisega (Olson, Delen 2008: 138).

1.5. Vabamorf EstNLTK osana

EstNLTK-s on morfoloogiline analüüs ja ühestamine implementeeritud kui märgendajad, mida saab (vajalike eelmärgendustega) sisendtekstidel rakendada. Need jaotuvad suurematekseks ja väiksemateks alamosadeks. Suuremad on VabamorfTagger ja VabamorfCorpusTagger, mille tekstide töötlemise etapid on küll samad, aga neid rakendatakse erinevalt. Kui VabamorfTagger võtab sisendiks ühe tekstiobjeki ja kasutab tavalist statistilist ühestamist, tahab VabamorfCorpusTagger sisendiks hulka sarnaseid tekste ehk ühestab tekstid korpusepõhiselt (vt peatükk 1.1.1 *Ühestamine*) (EstNLTK dokumentatsioon B7b).

VabamorfTagger ja VabamorfCorpusTagger hõlmavad enda all teisi märgendajad, mis jaotuvad omakorda väiksemateks osadeks. Kõiki märgendajaid ja nende osi on võimalik sisse ja välja lülitada ning selle abil omavahel kombineerida. Järgnevalt on loetletud kõik märgendajad ja nende osad, mille sisse- ja väljalülitamine on võimaldatud kaldkirjas esitatud lippude abil.

1. VabamorfAnalyser (Laur jt 2020: 7154)
    a. *guess* – tavaline oletamine. Kui sõna ei esine sõnastikus ja seda ei suudeta ka liitsõnana analüüsida, kasutatakse oletamist. (EstNLTK dokumentatsioon A1)

    b. *propername* – pärisnimede oletamine. Esisuurtähega sõnadele lisatakse pärisnime analüüs. (EstNLTK dokumentatsioon A1)
2. *slang_lex* – slängisõnastik. Täiendab Vabamorfi tavalist sõnastikku slängisõnadega. (EstNLTK dokumentatsioon A1)
3. PostMorphAnalysisTagger (Laur jt 2020: 7154)
    a. *use_postanalysis* – kasuta järelparandusi. Parandab numbritega sõnu, liitühendite sõnaliiki (nt initsiaalidega nimed, emotikonid jne) ja takistab mõningate sõnade ühestamist. (EstNLTK dokumentatsioon A1)
4. CorpusBasedMorphDisambiguator (Laur jt 2020: 7154)
    a. *predisambiguate* – eelühestamine. Ühestab enne tavalist ühestamist pärisnimesid. (EstNLTK dokumentatsioon A1)
    b. *postdisambiguate* – järelühestamine. Ühestab pärast tavalist ühestamist mitmeseks jäänud sõnu. (EstNLTK dokumentatsioon A1)
    c. *disamb_compound_words* – liitsõnade ühestamine. (EstNLTK dokumentatsioon B7b)
5. VabamorfDisambiguator (Laur jt 2020: 7154)
    a. *disambiguate* - ühestamine. Mitme analüüsiga sõnale jäetakse alles ainult konteksti sobivad analüüsid. (EstNLTK dokumentatsioon A1)
6. MorphAnalysisReorderer (EstNLTK dokumentatsioon B6)
    a. *use_reorderer* - sorteerib mitmeseks jäänud sõnade analüüse vastavalt UD korpuse sagedustele, mille tulemusel liigutatakse kõige tõenäolisem analüüs esimeseks. (EstNLTK dokumentatsioon A1)

Neid on võimalik mitmel moel kombineerida. Näiteks saab morfoloogiliseks analüüsiks kasutada ainult *guess*'i või koos VabamorfDisambiguatorist *disambiguate*'i ja CorpusBasedMorphDisambiguatorist *predisambiguate*'i jne. Vaikeväärtusena rakendatakse morfoloogilises analüüsis VabamorfAnalyserit ja VabamorfDisambiguatorit. (EstNLTK dokumentatsioon A1)

2. Meetod

2.1. Programmeerimiskeskkond

Programmeerimiskeelena kasutasin Pythonit (van Rossum, Drake 2009), sest just selles keeles on kirjutatud EstNLTK teek (Laur jt 2020), mille kvaliteeti ma analüüsin. Kasutasin Pyhtoni versiooni 3.6.10 (van Rossum, Drake 2009), ja EstNLTK versiooni 1.6.9b0 (Laur jt 2020). Koodi kirjutasin ma Jupyter Notebook'i keskkonnas, mis võimaldab koodi vahele ka tekstielemente lisada. See teeb koodi paremini loetavamaks. (Kluyver jt 2016) Bakalaureusetöö raames koostatud koodile saab ligi lingilt https://github.com/wertepure/Vabamorfi-hindamine.

2.2. Andmestik

Töös kasutatakse kahte korpust: EDT ehk Estonian Dependencies Treebank korpust ja EWT ehk Estonian Web Treebank korpust. Korpuste ja nende alamosade suurused on välja toodud tabelis 1.

Korpuste failid on jaotatud kolmeks alamosaks: *dev*, *test* ja *train*. *Dev* on kõige väiksem ja mõeldud arenduse jaoks, nii et ka mina kasutasin seda katsetamiseks. *Train* alamosa on kõige suurem ja seda kasutatakse muidu tarkvara treenimiseks. *Test* on kontrollgrupp, mille alusel katsetatakse, kui hästi programm töötab. Selle bakalaureusetöö raames pole aga vaja selliseid eristusi teha, nii et vaatluse alla võetakse kogu korpus. Igas alamosas on tekst dokumentide haaval lauseteks lahti võetud.

EDT korpuses jagunevad tekstid 3 žanri vahel: ajakirjandus, ilukirjandus ja teadustekstid. Neid kuvatakse lühenditega *aja*, *ilu* ja *tea*. Ilukirjandustekstide puhul on teksti id-s välja toodud vaid teksti žanr ja autor, näiteks *ilu_orlau*. Ajakirjandus ja teadustekstidel on teksti id-s välja toodud žanr, väljaande nimi ja avaldamise aeg, näiteks *aja_pm20001004*. Väljaande nimi võib olla pikalt välja kirjutatud (nt *eesti_arst*) või ka lühendiga esindatud (nt *ee* → *Eesti Ekspress*, *epl* → *Eesti Päevaleht*, *pm* →

Postimees). Avaldamise aeg võib piirduda vaid aastaga, aga ka aasta ja kuu või täpse kuupäevaga. Pikema kuupäeva puhul kirjutatakse see kokku, nt *20001004* ehk 4. oktoober 2000. aastal või *200009* ehk september 2000. aastal. Mõnel juhul on kuupäevale lisatud ka väljaande number, näiteks *aja_ee_199920*, kus 1999 tähistab aastat ja 20 väljaande numbrit. (Kaili Müürisep 2015)

Lisaks esinevad EDT korpuses tekstid *arborest*, *arborest-dev* ja *arborest-test*, mis ühegi žanri alla hästi ei jaotu. Nende sisu on võetud Huno Rätsepa lihtlausete korpusest, mis koosneb Huno Rätsepa raamatust „Eesti keele lihtlausete tüübid“ võetud näitelausetest (Bick jt 2004; Rätsep 1978). Kuna need moodustavad korpusest väga väikese osa, neid pole võimalik ühegi žanri alla jaotada ja need on suhteliselt lihtsalt analüüsitavad, olen otsustanud need oma tööst välja jätta.

EWT korpusega on lood EDT-ga võrreldes keerulisemad. Ka see korpus jaotub kolmeks suuremaks osaks, mis omakorda jaotuvad väiksemateks üksusteks:

1) vanem EWT osa, milles on erinevad täpsustamata uue meedia žanrid
    a) *ewtb1*
    b) *ewtb2*
2) foorumid
    a) *hfoorum*
    b) *tfoorum*
    c) *scifoorum*
    d) *sfoorum*
3) kommentaariumid
    a) *kom_jurgenstein*
    b) *kom_pandeemia*

Kahjuks pole kuskil dokumenteeritud, mis vahet on korpustel *ewtb1* ja *ewtb2*. Küll aga on teada, et *ewtb* korpuses on nii blogid, foorumid kui ka muud veebitekstid (Muischnek jt 2019: 23). Jääb aga selgusetuks, mida tähendavad nendega koos esinevad

numbritejadad (nt *ewtb1_112998*, *ewtb2_000035*), sest need ei tundu viitavat ei kuupäevale ega kellaajale. Võimalik, et need on veebilehtede id-d etTenTeni korpusest (Muischnek 2016), sest numbrijadad klapivad seal olevate failinimedega (Müürisep 2018). Samuti pole dokumenteeritud, mida need tähed foorumite ees tähendavad. Pilgust korpusele võin tuletada, et *hfoorum* on hariduse, *tfoorum* tehnika, *scifoorum* teaduse ja *sfoorum* seente kohta. Korpusesse piiludes sai ka selgeks, et *kom_jurgenstein* on Toomas Jürgensteini arvamusloo "Toomas Jürgenstein koroonakriisist: meie lõputu küünlapäev" all olev Delfi kommentaarium. Järelikult käib ka see kogumik tegelikult pandeemia alla. Ma ei oska küll öelda, miks see korpuses *kom_pandeemiast* eraldatud oli.

Tabel 1. **Korpuste üldstatistika**

Tekstitüüp	Sõnu	Sõnu kirjavahemärkideta	Lauseid	Dokumente
EDT korpus	428319	357737	29695	36
Ajakirjandus	266864	224762	18690	19
Ilukirjandus	67695	54225	5522	10
Teadustekstid	93760	78750	5483	7
EWT korpus	72698	60346	5863	41
Määramatud netitekstid	27304	23056	1715	32
Foorumid	32698	26616	2783	7
Kommentaariumid	12696	10674	1365	2
Mõlemad korpused kokku	501017	418083	35558	77

2.3. UD märgendussüsteemi teisendus Vabamorfi kujule

Eelnimetatud andmestik on UD märgendussüsteemis CoNLL-U formaadis[10], mis tuli morfoloogilise analüsaatori tööga võrdlemiseks Vabamorfi märgendussüsteemi kujule viia. Esimese sammuna koostati kategooriasüsteeme võrdlevad tabelid, kus on iga morfoloogia aspekti kohta välja toodud UD märgend ja Vabamorfi märgend, mida nende näitamiseks kasutatakse. Seal on eraldi tabelid sõnaliigi, käänete, arvu, asesõnade, verbi aspektide ja verbivormide kohta. Tabelid asuvad lisas 1.

Kuna UD ja Vabamorfi märgendid ei lähe üksteisega täielikult kokku, pidin teisendusel tegema teatud muudatusi. UD süsteemis esinevad alistava sidesõna märgendiga sõnad paigutati sidesõna märgendi alla ja sümbolid ühendati kirjavahemärkide märgendiga. Sõnad märgendiga *muu*, mille alla kuuluvad võõrkeelsed sõnad, eemaldati korpusest, kuna sellele puudub Vabamorfi süsteemis lähedane vaste (Müürisep 2021; Kaalep 1996: 49).

Vabamorfi süsteemis esinev genitiivatribuut ei läinud täpselt kokku ühegi UD sõnaliigiga (Müürisep 2021; Kaalep 1996: 49), mistõttu ei teisendatud seda käsitsi, vaid kasutati UD süsteemi välja *xpos*, milles on määratletud Vabamorfi süsteemi sõnaliik (Müürisep 2021). Seda välja ei kasutatud iga sõnaliigi teisenduses, sest puudub info selle kohta, kuidas see tekitatud on. Peale teisendust vaadati kuldstandardis genitiivatribuudi analüüsi saanud 689 sõna ka käsitsi üle ja leiti, et märgendid olid õiged.

Vormidest esindavad Vabamorfi *neg gem* (nt *ärgem*) ja *neg me* (nt *ärme*) märgendiga sõnu UD märgendussüsteemis täpselt samasugused märgendid. Korpust uurides tuli aga välja, et tegelikult ühtegi sõna, millel peaks *neg me* märgend olema, ei ole. Olid ainult need, millel peaks *neg gem* olema. Seepärast muudeti kõik märgendikombinatsiooniga *Mood=Imp + Tense=Pres + Person=1 + Number=Plur + Voice=Act + Polarity=Neg + VerbForm=Fin* sõnad Vabamorfi märgendiks *neg gem*.

[10] https://pypi.org/project/conllu/

Mitmesus esines ka tingiva kõneviisi lühikestes ja pikkades vormides, mis UD süsteemis saavad samad märgendid, aga Vabamorfis erinevad. Algselt otsustati teisendada kõik pikaks vormiks, sest seda on hiljem lihtsam vajadusel lühikeseks muuta. Siiski ilmnes korpuse ülevaatamisel, et ei esine sõnu, mille vorm oleks lühike *ks*, aga mis said märgendi *ksin*, *ksid*, *ksime* või *ksite*. Mitmesus *nuks* ja *nuksin*, *nuksid*, *nuksime*, *nuksite* vahel ei olnud samuti probleem, sest selliseid sõnu korpuses ei esine, mis pika vormi märgendi oleks saanud.

Pärast erijuhtude lahendamist kirjutati tabelite põhjal kood, mis UD süsteemi märgenduse Vabamorfi kujule viis ja siis .json formaati salvestas. Koodi ei kirjutatud otsast peale ise, vaid kasutati alusena koodi *ambiguous-morph-reordering* repositooriumist[11]. Selle koodi eesmärk oli luua leksikonid morfoloogilisel analüüsil mitmeseks jäänud sõnade ümber järjestamiseks analüüside sageduste alusel ja pärast hinnata eri leksikonide tööd. Selle jaoks kasutati samuti UD korpusest pärit tekste, mille märgendused Vabamorfi süsteemile ümber tehti. Küll aga oli sealt nii mõnigi teisendus puudu. Niisiis kasutasin seda koodi enda koodi põhjana ja tegin sinna vastavalt enda vajadustele muudatused. Peamiselt seisis see üleliigse eemaldamises ja teisenduste lisamises.

2.4. Hindamismeetrikate valik

Hindamismeetrikad valiti selle põhjal, et tulemusi oleks võimalik eelnevate töödega võrrelda ja et need vastaksid valdkonna standarditele. Nagu varem peatükis 1.4 mainiti, kasutati varasematel hindamistel suure tõenäosusega saagisele sarnast valemit, kuigi seda kutsuti neis töödes täpsuseks või korrektsuseks (Kaalep, Vaino 2001; Veskis, Liba 2008; Tkachenko, Sirts 2018; Leman 2019: 31; Milintsevich 2020: 25; Laur jt 2020). Morfoloogiliste analüsaatorite hindamise nõudeid välja toovas artiklis on soovitatud raporteerida korrektsuse asemel saagist ja täpsust. See peaks aitama eri süsteeme paremini võrrelda ja seega andma kasutajale võimalus valida süsteem, mis vastab kõige paremini tema vajadustele. (Faaß jt 2010: 804–805) Sellele lisaks on autorid arvutanud

[11] https://github.com/estnltk/ambiguous-morph-reordering

F-skoori (Faaß jt 2010: 807). Neid nõudeid on varasemalt järgitud näiteks jaapani keele morfoloogilise analüsaatori hindamisel, kus eristati korrektsust, saagist, täpsust ja F1-skoore (Den jt 2008: 1023). Lisaks on mitmeste analüüsidega materjali hindamiseks van Haltereni sõnul mõistlikum täpsuse ja saagise kasutamine (van Halteren 1999:82) ning need annavad ka analüsaatorist kasutajatele terviklikuma pildi (Faaß jt 2010: 804–805). Nendel põhjustel olen otsustanud ka oma töös tulemuste esitamiseks täpsust ja saagist kasutada. Kasutan saagist ka eelnevate uurimuste tulemuste selle töö omadega võrdlemiseks, kuigi sajaprotsendiliselt ei saa kindel olla, et kasutatud valem nies täpselt sama oli. Seal, kus täpsus ja saagis üksteisest väga palju ei erine, kasutan tulemuste näitamiseks F1-skoori.

Vabamorfi lemmadest on võimalik välja jätta võrdusmärk, plussmärk ja alakriips. Võrdusmärk eraldab lemmas sufiksit, plussmärk lõppu ja alakriips liitsõnade osi (Filosoft). Kuna kuldstandardi lemmades võrdusmärki ega plussmärki ei esine, olen otsustanud need Vabamorfi lemmadest eemaldada. Alakriipsuga on lood aga keerulisemad, kuna liitsõnu osi eemaldav alakriips esineb nii kuldstandardis kui ka Vabamorfis. Varasemalt on Vabamorfi lemmatiseerimist hinnatud bakalaureusetöös alakriipsud lemmadelt eemaldatud (Leman 2019: 31), aga magistritöös sisse jäetud (Milintsevich 2020: 28). Uurisin ka ise korpusest, kui paljudel sõnadel on alakriipsu asukoht lemmatiseerimise tulemusi mõjutanud, ja leidsin, alakriipsude välja jätmine parandas liitsõnade lemmatiseerimise tulemusi 1,5% võrra. Neist 1,35% olid kuldstandardi liitsõnad, mis Vabamorfis polnud liitsõna analüüsi saanud ehk neist puudus alakriips. 62% nendest sõnadest olid pärisnimed ja 26% nimisõnad, enamasti meditsiiniterminid. Kuna alakriipsude väljajätmine ei muuda oluliselt liitsõnade lemmatiseerimise tulemust ja üle 50% puhul on probleem spetsiifilise sõnaliigiga ehk liitsõna piiride määramise probleem pole laialdane üle igat tüüpi sõnade, olen otsustanud siinses töös alakriipsud sisse jätta. Pärisnime lemmade juures esitan siiski liitsõnapiiri probleemi näitlustamiseks tulemused ka ilma alakriipsuta.

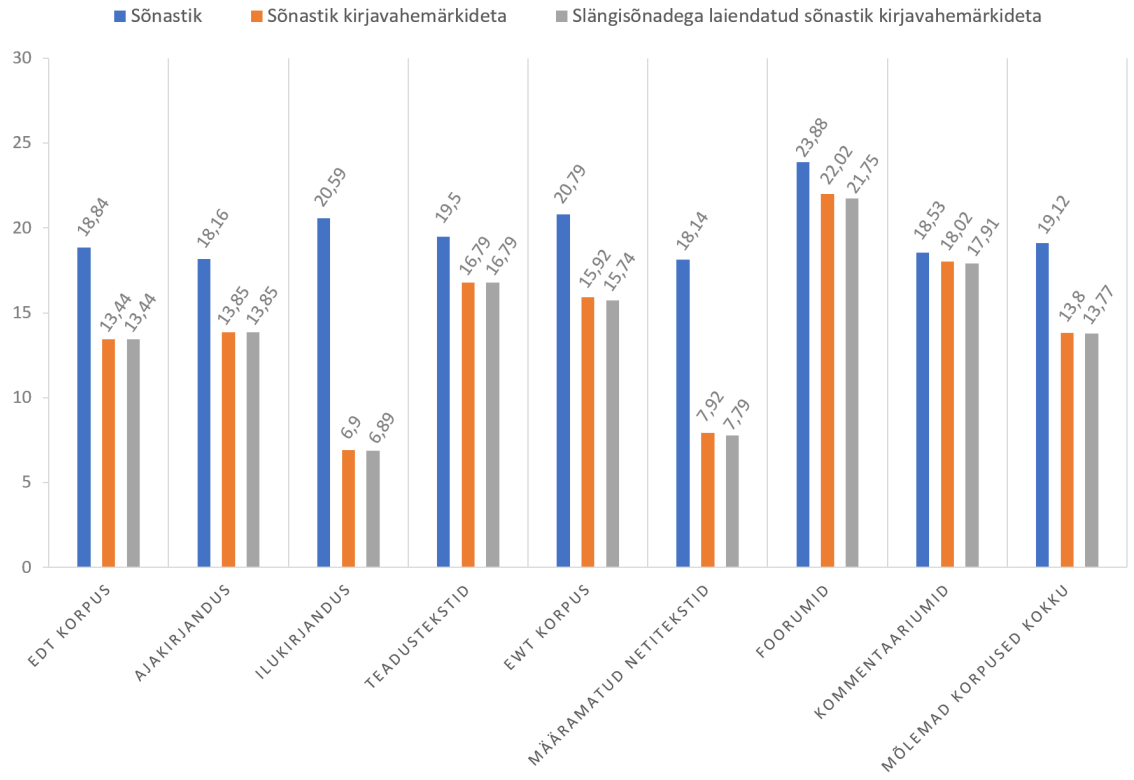
Lisaks on küsimus, kas hindamistest peaks kirjavahemärgid välja jätma või mitte. Kirjavahemärke on tavaliselt lihtne analüüsida, kuna nende analüüs on harva mitmene ja

nende lemma on üldiselt täpselt samasugune, nagu see teksti sees esineb. Nende omaduste tõttu tõstavad kirjavahemärgid harilikult tulemusi. Siiski saab ka vaielda, et kuna kirjavahemärkidel on süntaktiline tähendus ja need võivad ka näiteks netitektides saada mitmese analüüsi ja olla seal üldiselt keerulisemalt analüüsitavad, on nende sissejätmine põhjendatud. (van Halteren 1999:90) Selle dilemma lahendamiseks otsustan ma kirjavahemärkide analüüsimise üle eelnevate Vabamorfi uuringute vaates, et nende ja siinse töö tulemused oleks võimalikult hästi võrreldavad. Kuna üheski eelnevas hindamises (Kaalep, Vaino 2001; Veskis, Liba 2008; Tkachenko ja Sirts 2018; Leman 2019; Milintsevich 2020; Laur jt 2020) pole kirjavahemärke välja jäetud, olen otsustanud uuringute võrdlemise eesmärgil need ka oma töös sisse jätta. Peab aga märkima, et kuna kirjavahemärgid moodustavad kogu korpusest 16,53% ehk umbes ühe kuuendiku, hakkab see suure tõenäosusega tulemusi tõstma. Selle mõju näitlustamiseks tuuakse peatükkides 3.1 ja 3.6 välja ka ilma kirjavahemärkideta tulemus.

3. Tulemused ja analüüs

3.1. Tundmatud sõnad

Tundmatute sõnade alla arvestatakse sõnad, millele pole sõnastikupõhise analüüsi etapis ühtegi analüüsi antud. Töös uuritakse Vabamorfi tavalise ja slängisõnastikuga laiendatud sõnastiku tulemusi[12]. Nende uurimine näitab, kui suur osa sõnadest on sõnastikest puudu ja kas neid peaks tulevikus täiendama. Samuti avaldub tulemustest, kui palju kasu on slängisõnastiku kasutamisest. See annab näiteks netikeele uurijatele teadmise, kui palju nende tulemused võiks slängisõnastiku kasutamisest mõjutatud olla, ilma seda ise katsetamata. Tulemused esitatakse nii kirjavahemärkidega kui ilma, sest kirjavahemärke sõnastikes ei ole ehk kõik kirjavahemärgid jäävad analüüsita.



Joonis 2. **Tundmatute sõnade protsent sõnade koguarvust oletamist kasutamata**

[12] Märgendajana kasutati VabamorfAnalyserit, mille lipud *guess* ja *propername* olid välja lülitatud (vt peatükk 1.5 *Vabamorf EstNLTK osana*).

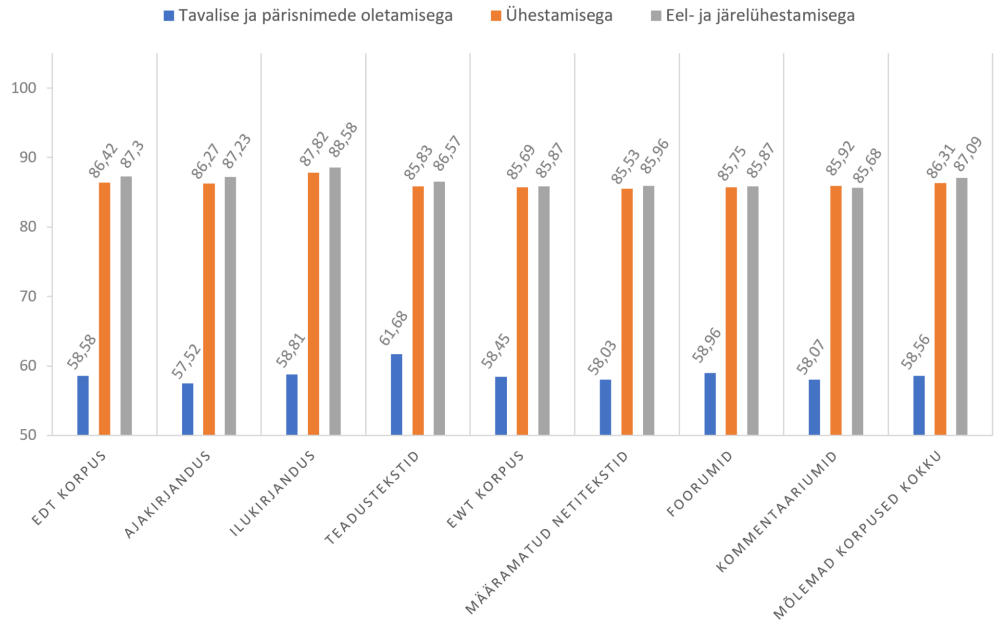
Joonisel 2 avaldub, et kirjavahemärgid moodustavad tundmatutest sõnadest tavaliselt väikese osa. See tendents ei kehti ainult ilukirjanduse ja määramatute netitekstide puhul, kus moodustavad kirjavahemärgid tundmatutest sõnadest vastavalt 66,5 ja 57,4 protsenti. Neis on ka kirjavahemärke arvestamata kõige vähem tuvastamata sõnu: ilukirjanduses 6,9% ja määramatutes netitekstides vaid protsent rohkem. Keskmiselt oli tuvastamine programmile raskem siiski EWT korpuse netikeele puhul, kus jäi märgenditeta kirjavahemärke arvesse võttes 20,79% ja ilma nendeta umbes 16% sõnadest, mida on 1,75–2,48% rohkem kui EDT tavaliste kirjakeelsete tekstide puhul. See tendents ei avaldu aga iga tekstiliigi lõikes, sest näiteks määramatutel netitekstidel, mis kuuluvad EWT korpuse alla, jäi märgendita kaks korda vähem sõnu kui teadustekstidel, mis EDT korpuses on. Teadustekstide suur tundmatute arv tuleneb terminitest, mida nende harulduse tõttu sõnastikes ei esine (nt *klopidogreel*, *demüelinisatsioon*, *lakunaarne*). Kokku jäi aga tavalise sõnastikuga tuvastamata 19,12% kirjavahemärkidega ja 13,8% kirjavahemärkideta sõnadest.

2001. aastal tehtud uurimuse põhjal jäi tundmatuks maksimaalselt 3% sõnadest (Kaalep, Vaino 2001: 3). Artiklis pole aga kahjuks mainitud, mis korpuse põhjal need tulemused saadi, mistõttu ei saa kindlalt öelda, et programmi sõnade tuvastamise võimekus oleks langenud. Siiski on siinsed tulemused võrreldavad 2020. aastal tehtud uurimuse tulemustega, kus kasutati samuti EDT korpust ja leiti, et ilma oletamisteta jääb analüüsita umbes 19% sõnadest (Milintsevich 2020: 22, 28), mis on väga lähedane siinse töö EDT korpuse kirjavahemärkidega 18,84% tulemustele. Sellest võib järeldada, et kahe aastaga pole EstNLTK sõnastikupõhine analüüs paranenud.

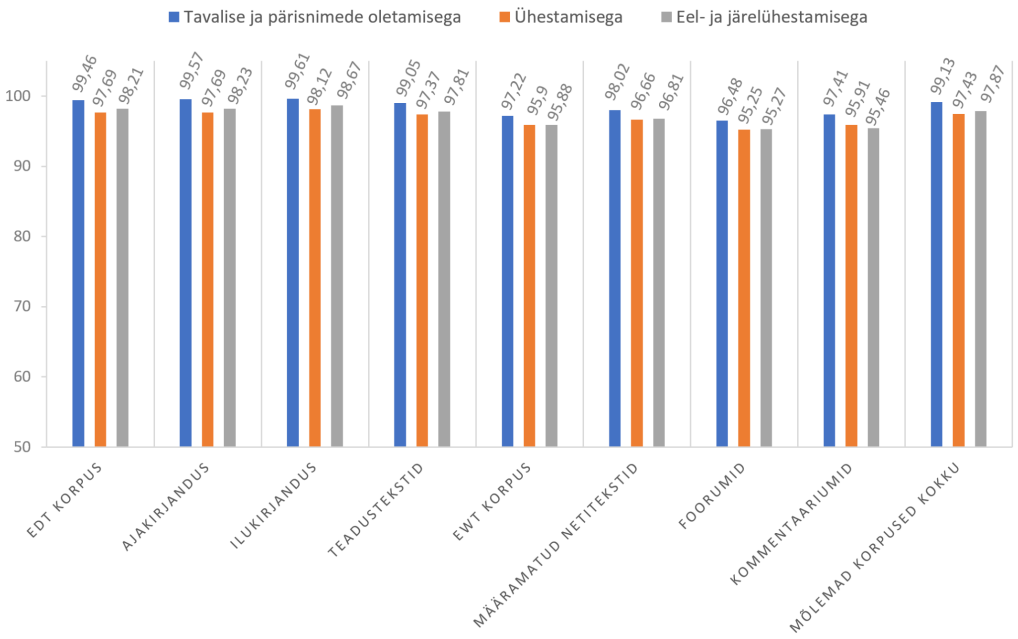
Üldiselt parandas slängisõnastiku lisamine tulemusi oodatavalt just EWT korpuse puhul, kus sai selle abil märgendi 0,18% tundmatutest sõnadest. Siiski vähendas see tundmatute arvu ka EDT korpuses, kus ilukirjandustekstides langes tundmatute sõnade arv 0,01% ja toorstatistikat vaadates vähenes see arv ka teistes EDT tekstitüüpides, kuigi liiga vähe, et protsendipunktina avalduda. Sellest saab järeldada, et slängisõnastiku lisamisest on kasu kõigi tekstiliikide lõikes.

3.2. Lemmade analüüsi kvaliteet

Algvormi õige tuvastamine on oluline infootsingus. Näiteks kui kasutaja otsib otsingumootorist sõna *pidu*, võiks ta saada vastuseks ka tekstid, kus see esineb vormis *peole* või *pittu* jne. Kui aga sõna *peole* on lemmatiseeritud sõnaks *pihk*, mitte *pidu*, ei väljasta otsingumootor seda teksti kasutajale.



Joonis 3. **Lemma analüüside kuldstandardiga ühildumise täpsus protsentides**



Joonis 4. **Lemma analüüside kuldstandardiga ühildumise saagis protsentides**

Oodatavalt on joonisel 3 kõige suurem täpsus eel- ja järelühestamisega, mis juhul jäetakse sõnale kõige vähem analüüse alles, ja joonisel 4 kõige suurem saagis oletamistega, aga ilma ühestamiseta, sest selle konfiguratsiooniga on sõnal kõige rohkem analüüse. Saagisest avaldub, et oletamistega puudub mõlema korpuse peale analüüside seast õige lemma ainult 0,87% sõnadest. Ühestamiste väiksemast saagisest on näha, et ühestamiste lisamine kaotab ära 1,26–1,7% õigetest analüüsidest. Siiski muutub õigete analüüside ja analüüside arvu vahekord ühestamise tulemusel tunduvalt paremaks, tõustes keskmiselt 27,75% tavalise ja 28,53% eel- ja järelühestamise lisamisega (vt joonis 3).

Enamikes tekstides on tavalisele ühestamisele eel- ja järelühestamise lisamine saagist tõstnud, mis tähendab, et kuigi eel- ja järelühestamise tulemusel jääb alles vähem analüüse, eemaldati neist tavalise ühestamisega võrreldes väiksem hulk õigeid. Ainus erand on siin kommentaariumid ja nende tulemusel EWT korpus üldiselt, kus eel- ja järelühestamise lisamisel saagis langes. See tähendab, et kommentaariumite analüüsimisel on eel- ja järelühestamisest pigem kahju kui kasu. Selle põhjus on ilmselt see, et eelühestamine on spetsiifiliselt pärisnimede jaoks, kus eeldatakse, et pärisnimed algavad suure algustähega. Netikeeles ei ole see aga alati nii (vt peatükk 3.3.1. *Pärisnimede lemmad*). Lisaks eelühestamise probleemidele on üle kõigi konfiguratsioonide EWT korpuse saagis ja täpsus EDT korpusest keskmiselt 1,79–2,33% madalamad ehk analüsaatorile valmistab netitekstides esinevate sõnade lemmatiseerimine rohkem raskusi kui tavalise kirjakeele sõnade lemmatiseerimine.

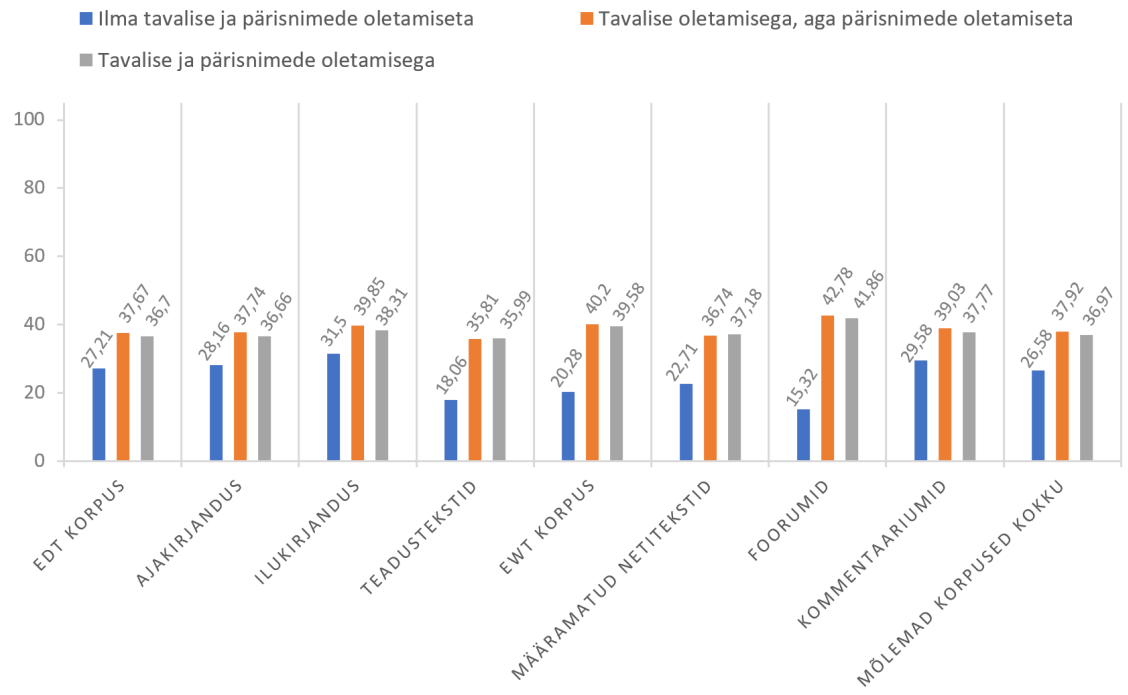
Uuriti ka, missugused sõnad tavaliselt vale lemma analüüsi saavad. Leiti, et faktoreid on mitmeid. Esiteks on paljud suurtähega algavad sõnad saanud pärisnime analüüsi. Need on tavaliselt meditsiiniterminid, nt *lipohüalinoos*, *hemianopsia*, aga ka mõned tavalisemad sõnad, nt *hilisstaadium*, *uudistetoimetus*. Liitsõnalisi meditsiinitermineid ei oska analüsaator ka õigesti osadeks jagada, eelistades liitsõnaosana mõnda tavalisemat sõna. Näiteks jaotatakse liitsõna *atsetüülkoliin* osadeks *atsetüülko_liin*. Samuti ei oska programm seda alati tavalistel sõnadel, kuid siis on need tavaliselt käändes: näiteks

jaotatakse sõnad *reviiri*, *dzunglieluviisilt* ja *rasvhapete* osadeks *rev_iiri*, *dzunglieluvii_silt* ja *rasvhap_pete*.

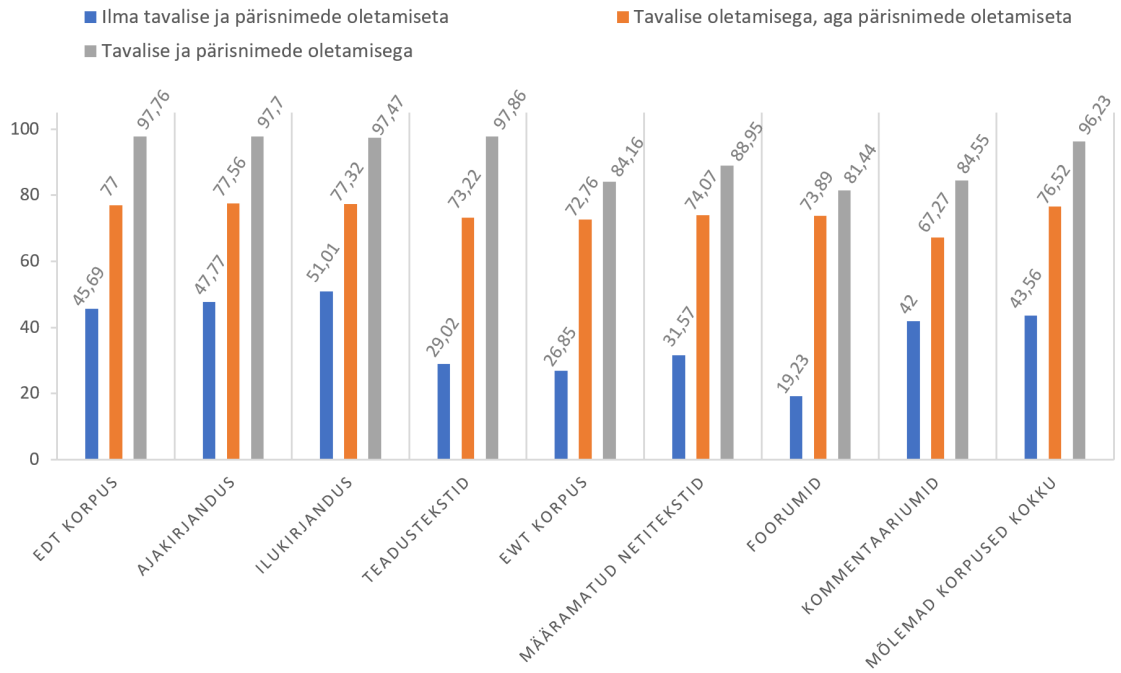
Välja paistsid aga ka mõned sõnad, mille analüsaator oli tegelikult õigesti liitsõnaks analüüsinud, aga kuldstandard ei olnud. Sellised olid näiteks *nikotiini_rikas*, *põgenike_paat*, *kodu_käijate_uurija*, mis kuldstandardis alakriipsuga eemaldatud ei olnud.

3.2.1. Pärisnime lemmade analüüsi kvaliteet

Järgnevalt uuritakse, kui hästi töötab Vabamorfi lemmatiseerija ilma ühestamiseta pärisnimede peal. Lemmatiseerija leiab käändes või pöördes sõna algvormi. See võib nimede puhul aga ka inimesele keeruline ülesanne olla. Näiteks lauses *Kuldmedal läks seekord Mäele* on kontekstita keeruline öelda, kas nime algvorm on *Mägi* nagu tõkkejooksja Rasmus Mägi või *Mäe* nagu maadleja Epp Mäe. Õige pärisnime tuvastamine on tähtis infootsingus ja info eraldamise ülesannetes.



Joonis 5. **Pärisnime lemmade kuldstandardiga ühildumise täpsus protsentides (%) liitsõnu osi eraldava alakriipsuga (ilma ühestamiseta)**



Joonis 6. **Pärisnime lemmade kuldstandardiga ühildumise saagis protsentides (%) liitsõnu osi eraldava alakriipsuga (ilma ühestamiseta)**

Joonisel 6 saagise protsente vaadates avaldub, et kui oletamistega ei leidunud EDT korpuse analüüside seas õiget pärisnime lemmat 2,24 protsendil pärisnimedest, siis EWT korpuse tulemused on sellest 7 korda madalamad ehk pärisnimedest puudus õige lemma 15,84% kordadest. Neist said kõige madalama saagise foorumid, kus tuvastati mõlema oletamisega vaid 81,44% pärisnimedest. Tuvastamata sõnu uurides tundub probleemi allikas olevat kasutajanimed, mis ei jälgi tavalisi nime tunnuseid: need algavad tihtipeale väikese algustähega (*olevipoeg*) ja võivad sisaldada numbreid (*Ceus12345*) ning lühendeid (*PC_man*).

Huvitaval kombel on joonisel 5 täpsus ilma mõlema oletamiseta madalam kui oletamiste lisamisega. Nimelt mõjutavad täpsust üleliigsed variandid, ehk tavaliselt on täpsus seda madalam, mida rohkem analüüse sõnal on. Siit see aga välja ei joonistu, sest analüüse tekitatakse oletamiste lisamisega aina juurde. Neist tulemustest saab seetõttu järeldada, et oletamiste lisamisega tekib rohkem õigeid variante kui üleliigseid. Siiski ei muutu üleliigsete ja õigete analüüside vahekord oletamiste lisamisega nii palju

paremaks, kui õige analüüsiga sõnade arv. Näiteks mõlema korpuse peale parandab oletamise lisamine täpsust 11,34% võrra, aga saagist lausa 32,96% ehk peaaegu kolm korda rohkem.

On ka näha, et tavalisele oletamisele pärisnimede oletamise lisamine vähendab analüsaatori täpsust, kuigi vähem kui protsendi võrra. Pärisnimede oletamise lisamisest on siiski suur kasu, sest pärisnimede oletamine tekitab sõnale peaaegu 20% õigeid analüüse juurde. Kokku parandab sõnastikupõhisele morfoloogilisele analüüsile oletamiste lisamine saagist rohkem kui kahekordselt.

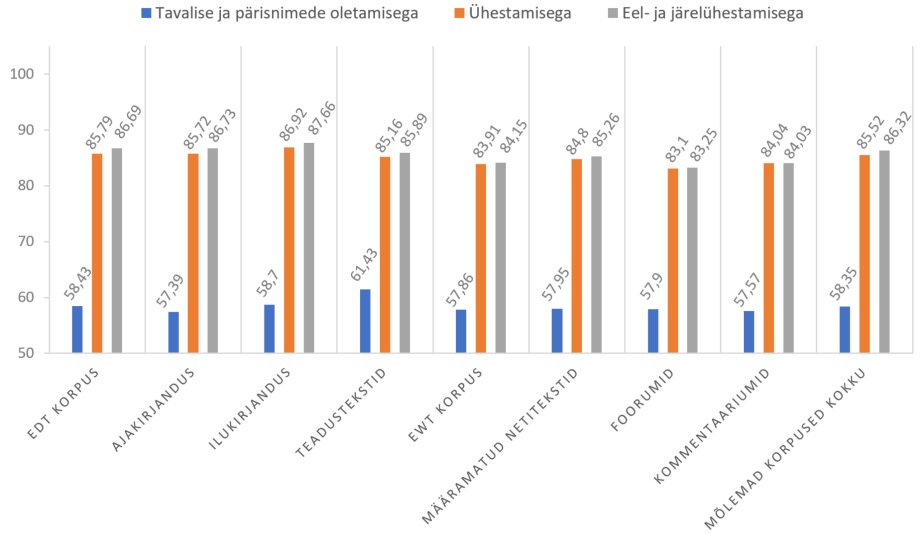
Kuna liitsõnapiiri oli analüsaatoril just pärisnimede jaoks kõige raskem määrata (vt peatükk 2.4. *Hindamismeetrikate valik*), uurisin, kuidas mõjutab tulemusi liitsõnade osi eraldava alakriipsu eemaldamine (vt lisa 2 *Pärisnime lemmad alakriipsuta*). See küll tõstis skoore, aga väga vähe. Täpsus ei tõusnud kordagi üle ühe protsendi ja saagis tegi seda ainult foorumites. Kuna isegi pärisnimede seas, mille lemmade ühildumist kuldstandardiga mõjutas liitsõnapiir sõnaliikide seast kõige rohkem, ei tekita liitsõna piiri eemaldamine tulemustes olulisi erinevusi, võib järeldada, et morfoloogilise analüsaatori liitsõnalisuse määramine töötab hästi. Siiski on võimalus, et UD korpuses tekitati liitsõnapiir automaatselt ja neid parandati seejärel minimaalselt. Selle tulemusel võib ka kuldstandardi liitsõnapiirides vigu olla (vt peatükk 3.3 *Lemmade analüüsi kvaliteet*), mis juhul ei pruugi eelnev järeldus kehtida.

Pärisnimede lemma määramisega eksitakse kõige sagedamini sõnadega, mis on ainsuse nominatiivis, genitiivis, inessiivis või partitiivis. Seal esines peamiselt kolm probleemi. Esimene on võõrapärased nimed, kust kas eemaldatakse lõpust täishäälik (*Time* → *Tim*), analüüsitakse nime liitsõnana (*Persepolis* → *Perse_polis*) või tavalise sõnana (*Sean* → *sead*, *Mulan* → *mula*). Teine on pärisnimed, mis olid originaaltekstis kirjutatud väikese tähega ega saanud seetõttu pärisnime analüüsi (*kalevipoeg*, *carmen*, *elisa*). See on peamiselt probleem netitekstides, kus ortograafiareegleid nii tulihingeliselt ei järgita. Kolmas on teadmatus, kas nime algvorm peab olema genitiivis või nominatiivis. Siin eksib analüsaator mõlemat pidi: nimest *Mustamäe* saab *Mustamägi*, nimest *Kivirähk* aga

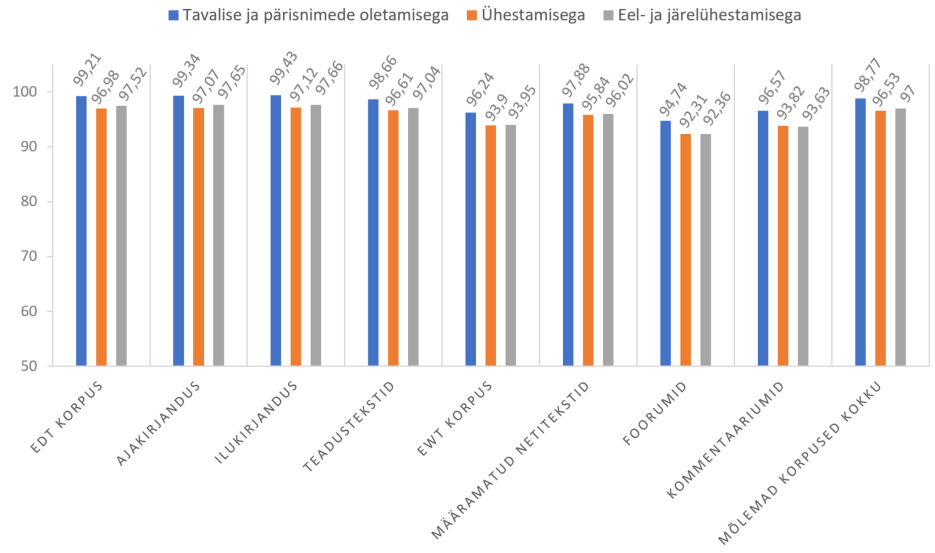
Kivirähu. Üks väga huvitav tendents jäi veel silma: ne-lõpulisele ainsuse genitiivis isikunimele eelistab analüsaator tekitada soomepärase algvormi. Näiteks ainsuse genitiivis nimele *Ruitlase* määrati algvormiks *Ruitlanen*, vormile *Haapsallase* *Haapsallanen* ja leedu nimevorm *Taupomasise* muutus vormiks *Taupomasinen*.

3.3. Sõnaliikide analüüsi kvaliteet

Sõnaliigi õige tuvastus on tähtis lingvistilises uurimistöös. Seda kasutatakse süntaksi uurimisel ja sellest on kasu ka rakenduste, nt grammatikakorrektori loomisel.



Joonis 7. **Sõnaliigi analüüside kuldstandardiga ühildumise täpsus protsentides**



Joonis 8. **Sõnaliigi analüüside kuldstandardiga ühildumise saagis protsentides**

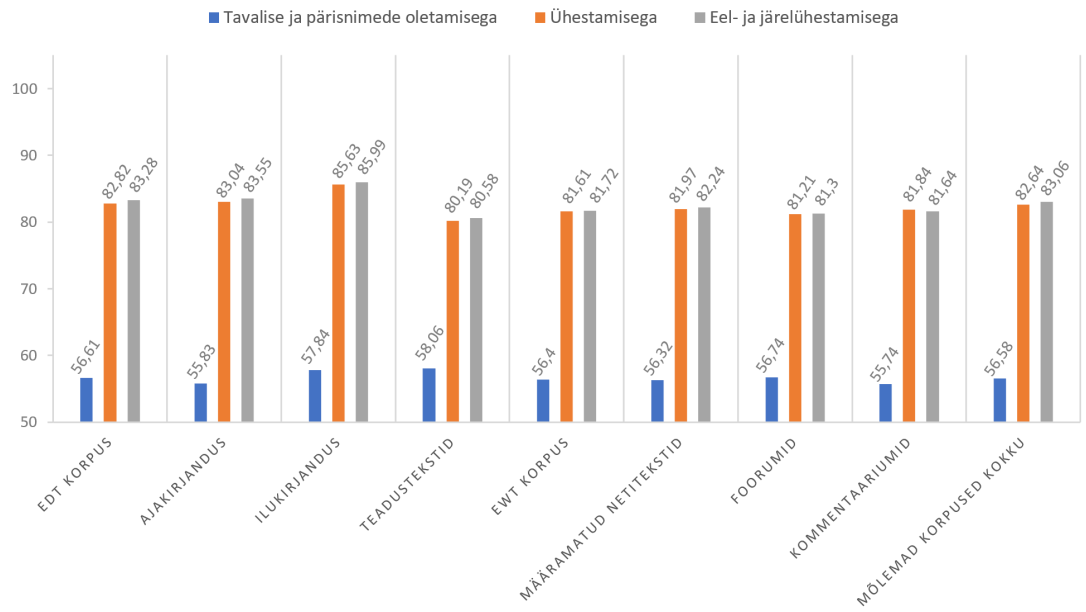
Nagu lemmade juureski on täpsuse ja saagise muutumise tendentsid konfiguratsioonide lõikes oodatavad, kuid madalamad kui lemmadel. Kui oletamistega jääb joonisel 4 õige lemma analüüsita 0,87% sõnadest, siis joonisel 8 jääb õige sõnaliigi analüüsita 1,23%. Ühestamiste lisamise tulemusel läheb õige sõnaliigi analüüs kaduma keskmiselt 1,77-2,24% sõnadel, mida on rohkem kui lemmade 1,26-1,7%. Kõigest sellest saab järeldada, et analüsaatorile on õige sõnaliigi määramine keerulisem kui lemmatiseerimine.

Lemmadega oli ka sarnasusi, sest mõlemil langetas kommentaariumite puhul eel- ja järelühestamise lisamine õigete analüüside arvu, kuigi igal pool mujal see tõusis. Siin ei olnud eemaldanud see siiski nii palju õigeid analüüse, et see kogu EWT korpuse tulemust endaga kaasa kisuks, mille tõttu on kogu EWT korpuse peale eel- ja järelühestamise lisamine siiski kasulikum kui selle välja jätmine.

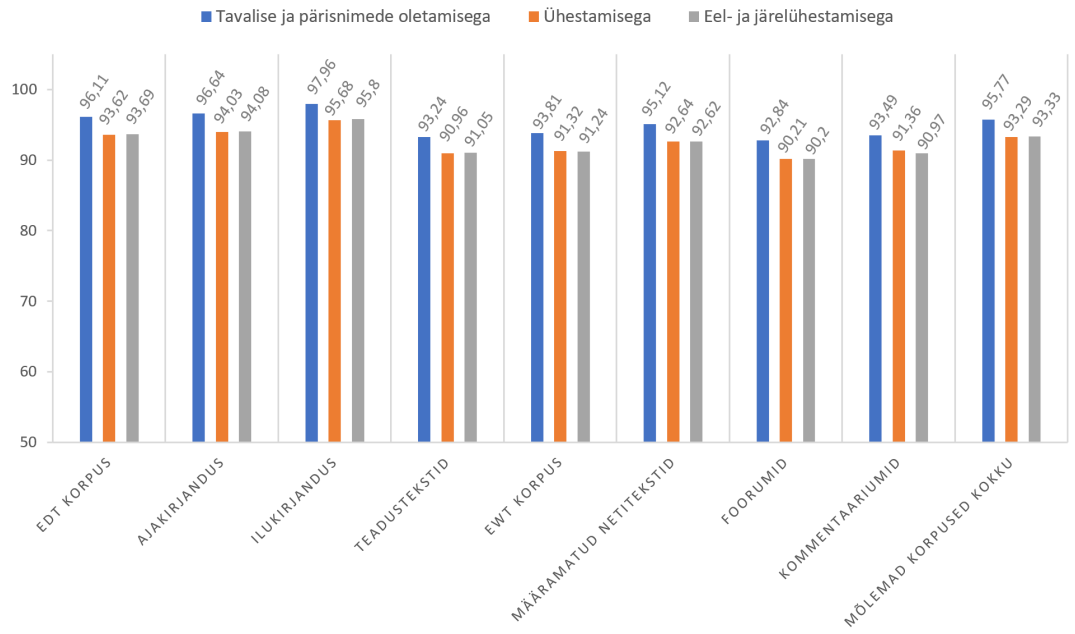
Uuriti ka, mis sõnaliigid kõige sagedamini vale analüüsi said. EWT korpuses olid nendeks pärisnimed, mille põhjustasid eelkõige väiketähelistest ja/või numbritega kasutajanimed, nt *Geenius666*, *d33m0n*, *mezik*. Mõlema korpuse peale kokku oli siiski kõige sagedam probleemiallikas määrsõna. EDT korpuses anti määrsõnadele kõige tihemini sidesõna analüüs, mis tulenes eelkõige sõna *aga* sellistest ühenditest, kus sel on rõhutav roll: *te aga mõtlete*, *jäid aga valitsema*. Lisaks tekitasid samasuguseid probleeme sõnad *pool* (*teisel pool lauda*), *kätte/käes* (*hakkab kätte maksma*) ja *sama* (*sama suur*). Ka kirjavahemärgid said tihti vale analüüsi, mis tulenes eelkõige kahest märgist: loendites kasutatav punkt · ja *%*, mis analüüsiti lühenditeks. Kui aga mitte arvestada seda, mis sõnaliigid tekstides kõige rohkem esinevad, said vale analüüsi kõige rohkem hüüdsõnad ja lühendid. Enamasti said hüüdsõnad kas nimisõna analüüsi (*jumal*, *vott*, *väk*) või suure esitähe tõttu pärisnime analüüsi (*Vat*, *Hommikust*, *Ergh*). Lühenditega oli sama lugu: saadi kas nimisõna (*WiFi*, *msnis*, *nato*), pärisnime (*de*, *la*, *SZTAKI*) või mõlema analüüsid (*TOP*, *ELU*, *RIBA*).

3.4. Vormide analüüsi kvaliteet

Nii nagu sõnaliigilgi, on õigest vormi analüüsist kasu eelkõige lingvistilises uurimistöös, peamiselt süntaksi uurimisel. Seda saab ka rakendada grammatikakontrollijas. Näiteks kui kasutaja mingis kontekstis valet käänet kasutab, saab analüsaatori abiga selle tuvastada ja teha vastav soovitus.



Joonis 9. **Vormi analüüside kuldstandardiga ühildumise täpsus protsentides**



Joonis 10. **Vormi analüüside kuldstandardiga ühildumise saagis protsentides**

Joonise 10 saagisest on näha, et oletamistega ei leitud õiget vormi mõlema korpuse peale 4,23% sõnadest, mis on 3,5 korda rohkem kui sõnaliikidel (vt joonis 8) ja peaaegu 5 korda rohkem kui lemmadel (vt joonis 4). Ühestamise lisamine eemaldab neist veel 2,44–2,48% õigeid analüüse, mida on samuti rohkem kui lemmadel ja sõnaliikidel. Sellest kõigest saab järeldada, analüsaatorile valmistab just vormi leidmine kõige rohkem raskusi.

Seekord pole aga ainult kommentaariumid need, kus eel- ja järelühestamise tulemusel õigete analüüside arv langeb, vaid ka foorumid ja määramatud netitekstid. Nende tulemusel eemaldatakse tervest EWT korpusest eel- ja järelühestamise lisamisega 0,08% õigeid analüüse. Siiski on täpsusest näha, et eel- ja järelühestamisega on õigete analüüside arv kõigist analüüsidest parema tasakaaluga kui tavalisel ühestamisel.

EWT korpuse saagis on EDT korpusega võrreldes 2,3–2,45% väiksem, kuigi joonisel 8 oli see sõnaliigil 2,97–3,57% ja joonisel 4 lemmal 1,79–2,33%. Sellest võib järeldada, et analüsaatori jaoks on netitekstidel ja tavalises kirjakeeles tekstidel kõige vähem vahet lemmade puhul, keskmiselt vormi ja kõige rohkem sõnaliigiga.

Tavalise ja pärisnimede oletamisega, aga ilma ühestamiseta tekkinud vigadest moodustavad käändsõnad 48,31% ja verbid 24,89%. Neis kerkis esile kaks suuremat probleemi: negatiivsed verbid ja arvud. Negatiivsetel verbivormidel (*ei maksa*, *ei väärinud*, *ei tülitaks*) ei suuda analüsaator negatiivsust tuvastada, sest oletamise staadiumis ei vaadata konteksti ehk analüsaator ei arvesta verbi ees oleva eitussõnaga. Numbritena kirjutatud arvud (*3*, *2008*, *7,5*) saavad aga vormi analüüsiks väga tihti küsimärgi, sest analüsaator ei suuda ilma kontekstita arvu käänet ära tunda. Just selle tõttu on analüüsivigade seas kõige sagedasemad käänded ainsuse nominatiiv (*aastani **2007***), genitiiv (*alla **7,5** kraadi*) ja adessiiv (***8.** märtsil*).

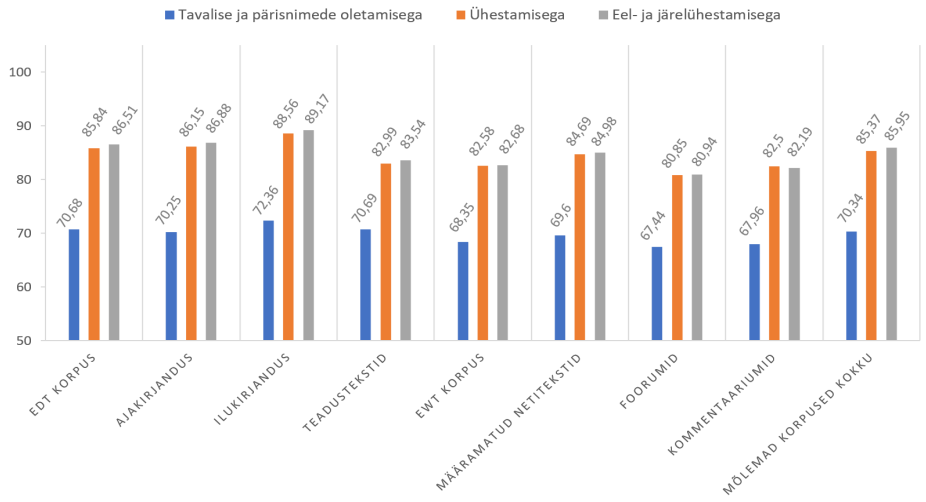
Ühestamise käigus said need probleemid parandatud, aga asendusid uutega. Neist puudutas valdav enamus ehk 82,96% käändsõnu. Omavahel läksid kõige rohkem segamini ainsuse nominatiiv, genitiiv, partitiiv ja aditiiv. Samad käänded tekitasid kõige

rohkem valesid analüüse ka 2001. ja 2008. aasta uurimustes (Kaalep, Vaino 2001: 6; Veskis, Liba 2008: 4–5). Siinses töös oli kohati probleeme ühildumise mitte äratundmisega. Näiteks analüüsiti seetõttu tihti ainsuse nominatiivi genitiivina (***Serbia*** *alustab*), genitiivi partitiivina (***arseeni*** *hinda*) või aditiivina (***sängi*** *alla*), partitiivi genitiivina (*rasket* ***noorpõlve***) või nominatiivina (***sõda*** *toetatakse*) ja aditiivi genitiivina (***ravikompleksi*** *kuulusid*). Veel tekkis probleeme eluta ja elusa objekti eristamata jätmise tõttu, mille põhjusel analüüsiti näiteks ainsuse genitiivi nominatiivina (kirjutas **muusika**) ja aditiivi genitiivina (**Tallinna** kolida).

Võrreldes käändsõnadega tekkis verbidega võrdlemisi vähe ühestamisvigu: 5,09% ehk käändsõnadest 16 korda vähem. Neist sai vale analüüsi kõige sagedamini märgend o ehk käskiva kõneviisi oleviku 2. isiku ainsuse aktiiv jaatavas kõnes (***ole*** *kuss,* ***mine*** *metsa*). See läks kõige tihemini segamini ainsuse genitiivi ja nominatiiviga, sest mõnda selle vormi sõna on võimalik ka nimisõnana analüüsida (***tule*** *tagasi*, ***lõigu*** *suurematek*s, ***lisa*** *lahusele*).

3.5. Analüüsi ja ühestamise koondtulemus

Siinsed joonised kujutavad olukorda, kus kõik väljad - lemma, sõnaliik ja vorm - peavad olema korrektsed, et sõna tulemus õigeks loetaks.

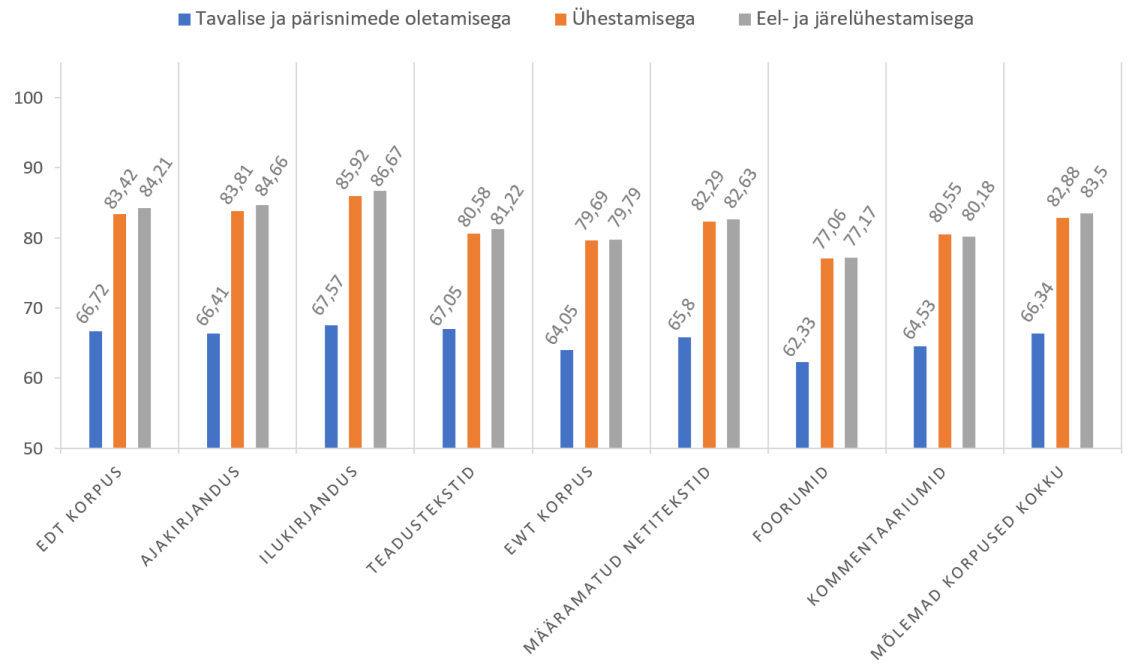


Joonis 11. **Sõna lemma, sõnaliigi ja vormi analüüside kuldstandardiga ühildumise F1-skoor protsentides kirjavahemärkidega**

Joonisel 11 on näha, et parim suhe analüüside arvu ja õigete analüüside vahel on mõlema korpuse peale kokku eel- ja järelühestamisega, kus see on 85,95%. Sellele jääb tavaline ühestamine alla ainult 0,58% ehk eel- ja järelühestamise lisamine ei paranda oluliselt tulemusi. Küll aga tekitab suure vahe oletamistele ühestamise lisamine, mille tulemusel tõuseb F1-skoor 15,03%. Eelnevas 2018. aasta uurimuses on öeldud, et vaikekonfiguratsioonis on Vabamorfi korrektsus EDT korpuse peal 79,17% (Tkachenko, Sirts 2018: 7). Siinne F1-skoor, mis arvestab peale eelnevate uurimustega paremini kokku mineva saagise ka täpsusega (vt peatükk 1.4. *Hindamismeetrikad*), on siinse töö tulemused eelnevast kõrgemad. Seda võib mõjutada aga ka asjaolud, et varasemas töös kasutati hindamiseks ainult 10% korpusest ja mitmeste analüüside puhul arvestati ainult esimesega (Tkachenko, Sirts 2018: 6, 7).

EDT korpuse tulemused on järjekordselt paremad kui EWT omad, kus F1-skoor on 2,33–3,83% madalam kui EDTs. Parima F1-skoori said ilukirjandustekstid ja halvima foorumid, mille vahe oli eel- ja järelühestamisega 8,23%. Huvitav on ka see, et ühestamiste konfiguratsioonidega sarnanevad teadustekstide F1-skoorid rohkem EWT korpuse omadele kui EDT-le, kust teadustekstid pärit on. Neid tekste ühendab suur tundmatute sõnade hulk (vt peatükk 3.1 *Tundmatud sõnad*), mille analüüsimisraskuse tõttu langevad ka skoorid. EWT korpus paistab aga silma ka sellega, et tavalisele ühestamisele eel- ja järelühestamise lisamine tõstab tulemusi ainult 0,1%, EDT korpusel seevastu 0,67%. Kommentaariumites lausa langetas selle lisamine tulemust 0,31% võrra. Järelikult on eel- ja järelühestamise lisamisest rohkem kasu tavalise kirjakeelega tekstidele ja netikeelsetele võib see lausa kahjulik olla.

Kuna kirjavahemärkidel on nii lemmat, sõnaliiki kui ka vormi tavaliselt lihtne analüüsida, uurin, kui palju nende eemaldamine tulemusi mõjutab.



Joonis 12. **Sõna lemma, sõnaliigi ja vormi analüüside kuldstandardiga ühildumise F1-skoor protsentides kirjavahemärkideta**

Kirjavahemärkide eemaldamine langetas oodatavalt skoore, olles joonisel 12 mõlema korpuse peale 2,45–4% madalam kui joonise 10 kirjavahemärkidega tulemus. Kõige rohkem mõjutas kirjavahemärkide eemaldamine oletamistega tulemusi, kus võrreldes eel- ja järelühestamisega langes F1-skoor 1,6 korda. EDT F1-skoor langes olenevalt konfiguratsioonist keskmiselt 2,3–3,96%, EWT seevastu 2,89–4,3%. See tähendab, et kirjavahemärkide eemaldamine langetab netikeele skoore rohkem kui tavalise kirjakeele omi. Selle põhjus on ilmselt selles, et kirjavahemärgid on kergesti analüüsitavad ja need moodustavad EWT korpusest 0,88% suurema osa kui EDT korpusest. Kuna netikeele emotikonid võivad koosneda kirjavahemärkidest ja seega kirjavahemärkide analüüsil probleeme tekitada, kontrollisin ka, kas Vabamorfi emotikonide analüüs ühtib kuldstandardiga. Tuli välja, et vähemalt lihtsamad emotikonid nagu naerunägu :) ja kurb nägu :(analüüsitakse kirjavahemärkidena nagu kuldstandardiski. Seetõttu võib järeldada, et kirjavahemärkide eemaldamine langetab EWT skoore rohkem kui EDT omi sellepärast, et analüsaator oskab emotikone õigesti analüüsida ja neid on EWT korpuses rohkem kui EDTs.

3.6. Korrektne ühene analüüs

Ühesed analüüsid teevad programmi kasutamise mugavamaks. Mida rohkem on üheseid analüüse, seda vähem peab kasutaja manuaalselt mitme analüüsiga sõnu läbi vaatama, et neis õige analüüs välja valida. Kui aga ühesed analüüsid õiged ei ole, peab olenevalt eesmärgist ka need manuaalselt läbi vaatama. Siin uuritaksegi, kui palju manuaalset tööd kasutaja potentsiaalselt tegema peab.

Lisaks sellele vaadatakse siin peatükis üle korpusepõhise ühestaja kvaliteet. Eri sisendkorpusteks võeti kokku EDT ja EWT alamosad ehk ajakirjandus, ilukirjandus, teadustekstid, määramatud netitekstid, foorumid ja kommentaariumid. See jaotus ei pruugi aga korpusepõhise ühestaja eeliseid väga hästi näidata, sest kuigi need on erinevad tekstitüübid, räägitakse ühe sees siiski erinevatest teemadest. Näiteks arutatakse foorumite all olevates tekstides nii hariduse, tehnika, teaduse kui ka seente kohta.

Analüüs loetakse üheseks ja korrektseks, kui sõnale anti ainult üks analüüs ja seal olid lemma, sõnaliik ja vorm kõik õiged. Tulemus arvutati peatükis 1.4 välja toodud van Halteneri korrektsuse valemiga ja korrutati protsendi saamiseks sajaga.

Tabel 2. **Korrektsete üheste analüüsidega sõnade protsent (%) sõnade koguarvust kirjavahemärkideta**

Teksti-tüüp	Oleta-mised	Ühesta-mine	Ühesta-mine ilma järel-analüüsi-deta	Eel-ühesta-mine	Järel-ühesta-mine	Eel- ja järel-ühesta-mine	Eel-, järel- ja liit-sõnade ühesta-mine	Koprusе-põhine ühesta-mine
EDT korpus	60,69	83,61	83,62	84,24	84,01	**84,49**	84,26	84,24
Ajakir-jandus	59,85	83,64	83,64	84,35	84,12	**84,62**	84,26	84,24
Ilukir-	62,85	87,23	87,24	87,78	87,49	**87,97**	87,95	87,93

jandus								
Teadus-tekstid	61,54	80,92	80,92	81,39	81,19	**81,62**	81,58	81,56
EWT korpus	57,42	79,86	79,86	79,91	**80,06**	80,05	80,01	79,98
Määra-matud neti-tekstid	58,34	81,68	81,68	81,91	81,99	**82,15**	82,08	82,04
Foorumid	56,83	78,36	78,35	78,39	78,46	**78,49**	78,44	78,43
Kommentaariumid	56,98	79,84	79,84	79,5	**79,99**	79,58	79,61	79,56
Mõlemad korpused kokku	60,22	83,07	83,07	83,61	83,44	**83,85**	83,64	83,62

Kõige rohkem korrektseid üheseid analüüse on siiski tekstipõhise eel- ja järelühestamisega, mitte uue korpusepõhise lahendusega. Selles konfiguratsioonis teeb keskmiselt rohkem tööd ära eelühestamine, mis tõstis tulemust tavalise ühestamisega võrreldes 0,54%, järelühestamine aga 0,37%. Siiski tuleb märkida, et EWT korpuses on hoopis järelühestamise tulemusel rohkem õigeid üheseid analüüse kui eelühestamisega.

Kõige suurem hüpe toimub oletamistele tavalise ühestamise lisamisega, mis tõstab korrektsete üheste analüüside arvu keskmiselt 22,85%. Sellega võrreldes paraneb tulemus järgmiste etappide lisamisega väga vähe, iga konfiguratsiooniga alla protsendi. Niisiis sõltub üheste korrektsete analüüside arv kõige rohkem ühestamisetapi lisamisest ja palju vähem ühestamise liigist.

Varasemalt on leitud, et ühese analüüsi saanud sõnadest sai tavalise ja pärisnimede oletamisega korrektse analüüsi 97.41%. Küll aga pole artiklis täpsustatud, kas nii lemma, sõnaliik kui ka vorm pidid õige analüüsiga olema, et kogu analüüs õigeks

loetaks, või pidi olema ainult üks neist. (Tkachenko, Sirts 2018: 3) Seetõttu ei ole võimalik öelda, kas tulemused on ajaga halvenenud, mida graafikule vaadates võiks esialgu eeldada.

Kokkuvõte

Bakalaureusetöö eesmärk oli hinnata EstNLTK morfoloogilist analüsaatorit ja ühestajat Vabamorfi ning selle konfiguratsioonide kvaliteeti eri tüüpi tekstide peal. Selle tulemusel saavad programmi kasutajad parema idee, mis probleeme võib selle kasutamisel ette tulla ja programmi tegijad näevad programmi kitsaskohti, mille parandamisele tulevikus keskenduda. Autori hinnangul see eesmärk täideti.

Vabamorfi väljundi hindamiseks koostati kood, mille abil võrreldi seda UD korpusest saadud inimeste märgendatud kuldstandardiga. Kuna Vabamorf ja UD korpus kasutavad erinevaid märgendussüsteeme, teisendati UD märgendid Vabamorfi kujule. Teooriapeatükis anti ülevaade nende süsteemide erinevustest, arvutimorfoloogiast, Vabamorfi tööpõhimõttest, selle eelnevatest hindamistest ja morfoloogiliste analüsaatorite ja ühestajate hindamismeetrikatest. Meetodiosas kirjeldati kasutatavaid korpusi, kuidas nende märgendussüsteemi teisendus Vabamorfi kujule toimus, mis hindamismeetrikate alusel tulemusi esitati ning mis programmeerimiskeelt ja -keskkonda teisenduste ja hindamiste jaoks kasutati.

Kuldstandardina kasutati UD 501017 sõnalist korpust, mis jaotub EDT ja EWT korpusteks. EDT alla kuuluvad ajakirjandus, ilukirjandus ja teadustekstid ning EWT alla määramatud netitekstid, kommentaariumid ja foorumid. Töös hinnati nende korpuste ja tekstitüüpide lõikes Vabamorfi võimekust analüüsida sõna lemmat, pärisnime lemmat, sõnaliiki ja vormi. Lisaks sellele uuriti, kui suur protsent tekstides esinevaid sõnu puudub morfoloogilise analüsaatori sõnastikust, kui palju kasu on tavalisele sõnastikule slängisõnastiku lisamisest ja kui palju üheseid korrektseid analüüse Vabamorf tekitab.

EDT korpuse tulemused olid iga hindamise lõikes EWT omadest paremad. Tekstitüüpide seast töötas analüsaator kõige paremini ilukirjanduse ja kõige halvemini foorumite peal. Ilma oletamisi kasutamata ja kirjavahemärke arvesse võtmata jääb tundmatuks keskmiselt 13,8% sõnu. Slängisõnastiku lisamisest on kasu kõigi

tekstiliikide lõikes, aga rohkem EWT korpuse puhul, kus selle lisamise tulemusena vähenes tundmatute sõnade arv 0,18%. Kõige rohkem korrektseid üheseid analüüse tekitati tekstipõhise eel- ja järelühestamise konfiguratsiooniga, mille tulemusel sai korrektse ühese analüüsi 83,85% sõnadest. Vaikekonfiguratsioonile eel- ja järelühestamise lisamine pole aga netikeelsete tekstide puhul alati soovitatav, sest kommentaariumite puhul langetab selle lisamine igas aspektis skoore.

Lemma, sõnaliigi ja vormi lõikes on Vabamorf parim lemmade määramises, kus jäi koos oletamistega, aga ilma ühestamiseta vaid 0,87% sõnadest õige lemmata. Eraldi uuriti ka pärisnime lemmasid ja leiti, et EDT korpuses jäi sama konfiguratsiooniga õige lemma analüüsita 2,24% ja EWT korpuses 15,84% pärisnimedest. EWT madala skoori põhjustasid kasutajanimed, sest need ei pruugi alata suure algustähega ja võivad sisaldada numbreid ja lühendeid. Pärisnimed oli tüüpviga ka sõnaliikide puhul, kus jäi tavalise ja pärisnimede oletamise konfiguratsiooniga õige sõnaliigi analüüsita 1,23% sõnades. Siiski oli seal pärisnimede asemel suurim murekoht määrsõnad, mis said tihti sidesõna analüüsi. Kõige rohkem raskusi valmistas Vabamorfile õige vormi määramine. Sama konfiguratsiooniga jäi õige vormita keskmiselt 4,23% sõnadest. Vormi määramine on üldiselt raskem käändsõnade puhul. Oletamistega konfiguratsioonis olid aga tüüpvigadeks peamiselt negatiivsed verbid ja arvud, mille vormi ei suudeta ilma kontekstita õigesti tuvastada. Pärast tekstide ühestamist suurenes käändsõnadega seotud probleemide arv pea kahekordselt. Seal tekitas kõige rohkem vigu neli käänet: ainsuse nominatiiv, genitiiv, partitiiv ja aditiiv. Verbidest oli suurim probleemiallikas käskiva kõneviisi oleviku 2. isiku ainsuse aktiiv jaatavas kõnes (nt ***ole*** *kuss*).

Kirjandus

Bick jt = Bick, Eckhard, Heli Uibo, Kaili Müürisep 2004. Arborest - a Growing Treebank of Estonian. – Nordisk Sprogteknologi 2004. Nordic Language Technology. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Ed. Henrik Holmboe. Copenhagen: Museum Tusculanums Forlag, 125–142; https://www.visl.sdu.dk/pdf/Bick_Uibo_Muurisep_Arborest_NorFA_yearbook_2004.pdf. Vaadatud 02.05.2022.

Den jt = Den, Yasuharu, Junpei Nakamura, Toshinobu Ogiso, Hideki Ogura 2008. A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation. – Proceedings of the Sixth International Conference on Language Resources and Evaluation. Marrakech: European Language Resources Association, 1019–1024; http://www.lrec-conf.org/proceedings/lrec2008/pdf/258_paper.pdf. Vaadatud 16.05.2022.

Derczynski, Leon 2016. Complementarity, F-score, and NLP Evaluation. – Proceedings of the International Conference on Language Resources and Evaluation. Portorož: European Language Resources Association, 261–266; https://aclanthology.org/L16-1040.pdf. Vaadatud 16.05.2022.

EKK = Erelt, Mati, Tiiu Erelt, Kristiina Ross 2020. Eesti keele käsiraamat. Uuendatud väljaanne. Tallinn: Eesti Keele Sihtasutus.

Erelt jt = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare 1993. Eesti keele grammatika II. Süntaks. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Erelt jt = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Silvi Vare, Kristiina Ross, Henno Saari, Kaja Tael, Tiiu Erelt, Ülle Viks 1995. Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

EstNLTK dokumentatsioon A1; https://nbviewer.org/github/estnltk/estnltk/blob/version_1.6/tutorials/nlp_pipeline/A_01_short_introduction_and_tutorial_for_linguists.ipynb. Vaadatud 20.05.2022.

EstNLTK dokumentatsioon B6; https://nbviewer.org/github/estnltk/estnltk/blob/version_1.6/tutorials/nlp_pipeline/B_06_morphological_analysis.ipynb. Vaadatud 20.05.2022.

EstNLTK dokumentatsioon B7b https://nbviewer.org/github/estnltk/estnltk/blob/version_1.6/tutorials/nlp_pipeline/B_07b_morph_analysis_with_corpus-based_disambiguation.ipynb. Vaadatud 20.05.2022.

Faaß jt = Faaß, Gertrud, Ulrich Heid, Helmut Schmid 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR as an Example of Morphological Evaluation. – Proceedings of the Seventh International Conference on Language Resources and Evaluation. Valletta: European Language Resources Association; http://www.lrec-conf.org/proceedings/lrec2010/pdf/409_Paper.pdf. Vaadatud 16.05.2022.

Filosoft; https://www.filosoft.ee/html_morf_et/morfoutinfo.html. Vaadatud 17.05.2022.

van Halteren, Hans 1999. Syntactic Wordclass Tagging. Dordrecht: Kluwer Academic Publishers, 82–90.

Kaalep, Heiki-Jaan 1996. ESTMORF: a morphology analyser for Estonian. – Estonian in the Changing world. Tartu: Tartu Ülikooli Kirjastus, 43–97.

Kaalep, Heiki-Jaan, Tarmo Vaino 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. – Keel ja Kirjandus 1, 30–36.

Kaalep, Heiki-Jaan, Tarmo Vaino 2001. Complete Morphological Analysis in the Linguist's Toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum Pars V. Tartu: Eesti Fennougristide Komitee, 9–16; http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf. Vaadatud 26.04.2022.

Kaalep, Heiki-Jaan 2014. Arvutimorfoloogia loeng 1. https://kodu.ut.ee/~hkaalep/arvutimorf_14/loeng1.pdf. Vaadatud 26.04.2022.

Kaalep, Heiki-Jaan 2016. Arvutimorfoloogia loeng 3. https://kodu.ut.ee/~hkaalep/arvutimorf_16/loeng3.htm. Vaadatud 26.04.2022.

Kaalep jt = Kaalep, Heiki-Jaan, Kadri Muischnek, Kristel Uiboaed, Kaarel Veskis 2010. The Estonian Reference Corpus: its composition and morphology-aware user

interface. – Vol 219: Human Language Technologies – The Baltic Perspective, 143–146; https://doi.org/10.3233/978-1-60750-641-6-143. Vaadatud 23.05.2022.

Kaalep jt = Kaalep, Heiki-Jaan, Riin Kirt, Kadri Muischnek 2012. A trivial method for choosing the right lemma. – Vol 247: Human Language Technologies – The Baltic Perspective, 82–89; doi.org/10.3233/978-1-61499-133-5-82. Vaadatud 24.05.2022.

Kluyver jt = Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Perez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damian Avila, Safia Abdalla, Carol Willing 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. – Positioning and Power in Academic Publishing: Players, Agents and Agendas. Ed. by F. Loizides, B. Schmidt, 87–90.

Laur jt = Laur, Sven, Siim Orasmaa, Dage Särg, Paul Tammo 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of the 12th Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 7152–7160; https://aclanthology.org/2020.lrec-1.884.pdf. Vaadatud 26.04.2022.

Leman, Laura Katrin 2019. Tehisnärvivõrgul põhinevate lemmatiseerijate võrdlev analüüs eesti keeles; http://hdl.handle.net/10062/64538. Vaadatud 26.04.2022.

de Marneffe jt = de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, Daniel Zeman 2021. Universal Dependencies. – Computational Linguistics, Vol 47, Issue 2, 255–308; https://doi.org/10.1162/coli_a_00402. Vaadatud 15.05.2022.

Milintsevich, Kirill 2020. Lexicon-Enhanced Neural Lemmatization for Estonian; https://comserv.cs.ut.ee/home/files/Milintsevich_Thesis.pdf?study=ATILoputoo&reference=85CBA6E56A7D5B8F1ABAC9A53934520DE957BB7B. Vaadatud 26.04.2022.

Muischnek, Kadri 2016. Eesti veeb 2013 (etTenTen) korpus, morfoloogiliselt ühestatud. – Eesti keeleressursside keskus. https://doi.org/10.15155/1-00-0000-0000-0000-0012EL. Vaadatud 24.05.2022.

Muischnek jt = Muischnek, Kadri; Mark Fišel, Heiki-Jaan Kaalep, Mare Koit, Kaili Müürisep, Heili Orav, Kadri Vare, Haldur Õim 2012. Arvutilingvistika ja

keeletehnoloogia Tartu Ülikoolis. – Emakeele Seltsi Aastaraamat. Toim. Mati Erelt, Sirje Mäearu. Tallinn: Tartu Ülikooli Kirjastus. 66−102; http://dx.doi.org/10.3176/esa57.05. Vaadatud 17.05.2022

Muischnek jt = Muischnek, Kadri, Kaili Müürisep, Dage Särg 2019. CG Roots of UD Treebank of Estonian Web Language. – Proceedings of the NoDaLiDa 2019 Workshop on Constraint Grammar - Methods, Tools and Applications. Linköping: Linköping University Electronic Press, 23−26; https://ep.liu.se/ecp/168/006/ecp19168006.pdf. Vaadatud 02.05.2022.

Müürisep, Kaili 2015. EDT korpuse dokumentatsioon; https://github.com/EstSyntax/EDT/blob/master/puudepangaallikad.pdf. Vaadatud 25.05.2022.

Müürisep, Kaili 2021. UD eesti keele dokumentatsioon; https://github.com/EstSyntax/EstUD/blob/master/EestiUDdokumentatsioon.pdf. Vaadatud 17.05.2022.

Olson, David L., Dursun Delen 2008. Advanced Data Mining Techniques. Berlin: Springer.

van Rossum, Guido, Fred L. Drake 2009. Python 3 Reference Manual. Scotts Valley: CreateSpace.

Qi jt = Qi, Peng, Timothy Dozat, Yuhao Zhang and Christopher D. Manning 2018. Universal Dependency Parsing from Scratch. – *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160-170; https://nlp.stanford.edu/pubs/qi2018universal.pdf. Vaadatud 01.06.2022.

Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. – Eesti NSV Teaduste Akadeemia Emakeele Seltsi toimetised nr. 12. Tallinn: VaIgus.

Tkachenko, Alexander, Kairit Sirts 2018. Neural Morphological Tagging for Estonian; https://arxiv.org/pdf/1810.06908. Vaadatud 26.04.2022.

Veskis, Kaarel, Erkki Liba 2008. Automatic Tagger Evaluation NGSLT: NLP Syntax Assignment Report; http://teataja.ee/veskis-liba-syntax-assignment-modified.pdf. Vaadatud 26.04.2022.

Viks, Ülle 1992. Väike vormisõnastik I: Sissejuhatus & grammatika. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Viks, Ülle 2000. Eesti keele avatud morfoloogiamudel. – Arvutilingvistikalt inimesele. Toim. Tiit Hennoste. Tartu: Tartu Ülikooli Kirjastus, 9–36; http://dspace.ut.ee/bitstream/handle/10062/41671/hennoste_arvutuslingvistkalt_ocr.pdf. Vaadatud 26.04.2022.

Evaluation of EstNLTK's morphological analyser and disambiguator

Estonian is a very morphologically rich language, which is why morphological analysis is pivotal for any NLP task in this language. The aim of this thesis is to evaluate the quality of EstNLTK's morphological analyser and disambiguator Vabamorf for the Estonian language in order to identify its standard errors and how well its configurations work on different corpora.

The author created a Python program to compare EstNLTK's automatic annotations to UD corpus' gold standard annotations. However, the two systems use different tags, which is why firstly UD tags were automatically converted to match Vabamorf's. The thesis also gives an overview of the differences these systems have, computational morphology, previous evaluations, how Vabamorf works and how morphological analysers and disambiguators are usually evaluated.

The corpus consists of 501017 words and is made up of two smaller corpora: EDT and EWT. EDT includes newspapers, fiction and scientific texts, EWT includes comment sections, forums and other web texts. Both precision, recall and F1-scores were used for evaluation.

The results of the EDT corpus exceeded those of EWT in every aspect. The best results were achieved when pre- and postdisambiguation were added to the default configuration of guessers and regular disambiguation. However, it did also lower scores on comment sections. Out of the lemma, part of speech and form categories Vabamorf is best at lemmatization and worst at finding the correct form. Most form-related problems were caused by 4 cases: singular nominative, genitive, partitive and short illative. The frequent mistakes for part of speech include adverbs, proper names, interjections and abbreviations.

Lisa 1: Andmeväljade võrdlus

Lisa 1.1. Sõnaliigid

Sõnaliik	UD (lisandid eraldi tulbas)	Vabamorf
omadussõna (algvõrre)	ADJ Degree=Pos	A
omadussõna (keskvõrre)	ADJ Degree=Cmp	C
omadussõna (ülivõrre)	ADJ Degree=Sup	U
määrsõna	ADV	D
genitiivatribuut	-	G
pärisnimi	PROPN	H
hüüdsõna	INTJ	I
sidesõna	CCONJ	J
kaassõna	ADP	K
põhiarvsõna	NUM Card=Yes	N
järgarvsõna	NUM Ord=Yes	O
asesõna	PRON/DET	P
nimisõna	NOUN	S
tegusõna	VERB	V
abitegusõna	AUX	V
lühend	Abbr=Yes	Y
kirjavahemärk	PUNCT	Z
alistav sidesõna	SCONJ	J*
sümbol	SYM	Z*
muu	X	-

Lisa 1.2. Käänded ja arv

Kääne	UD (Case=)	Vabamorf
nominatiiv/nimetav	Nom	n
genitiiv/omastav	Gen	g
partitiiv/osastav	Par	p
illatiiv/sisseütlev	Ill	ill
aditiiv/lühike sisseütlev	Add	adt
inessiiv/seesütlev	Ine	in
elatiiv/seestütlev	Ela	el
allatiiv/alaleütlev	All	all
adessiiv/alalütlev	Ade	ad
ablatiiv/alaltütlev	Abl	abl
translatiiv/saav	Tra	tr
essiiv/olev	Ess	es
terminatiiv/rajav	Ter	ter
abessiiv/ilmaütlev	Abe	ab
komitatiiv/kaasaütlev	Com	kom
ainsus	Sing	sg
mitmus	Plur	pl

Lisa 1.3. Verbid

Vorm	UD
indikatiiv/kindel kv	Mood=Ind
imperatiiv/käskiv kv	Mood=Imp
konditsionaal/tingiv kv	Mood=Cnd

kvotatiiv/kaudne kv	Mood=Qot
olevik	Tense=Pres
minevik	Tense=Past
aktiiv/isikuline tegumood	Voice=Act
passiiv/umbisikuline tegumood	Voice=Pass
1. isik	Person=1
2. isik	Person=2
3. isik	Person=3
ainsus	Number=Sing
mitmus	Number=Plur
infinitiiv (da/vat)	VerbForm=Inf
finitiiv	VerbForm=Fin
partitsiip/kesksõna (v/tav/nud/tud)	VerbForm=Part
supiin/ma-tegevusnimi	VerbForm=Sup
konverb (des/mata/maks)	VerbForm=Conv
eitus	Polarity=Neg

Lisa 1.3.1. Verbivormid

Vabamorf	Vorm	UD
b	kindel kõneviis olevik 3. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=3 Number=Sing Voice=Act
d	kindel kõneviis olevik 2. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=2 Number=Sing Voice=Act
da	infinitiiv jaatav kõne	VerbForm=Inf
des	gerundium jaatav kõne	Verbform=Conv

ge	käskiv kõneviis olevik 2. isik mitmus aktiiv jaatav kõne	Mood=Imp Tense=Pres Person=2 Number=Plur Voice=Act VerbForm=Fin
gem	käskiv kõneviis olevik 1. isik mitmus aktiiv jaatav kõne	Mood=Imp Tense=Pres Person=1 Number=Plur Voice=Act VerbForm=Fin
gu	käskiv kõneviis olevik 3. isik mitmus aktiiv jaatav kõne	Mood=Imp Tense=Pres Person=3 Number=Plur Voice=Act VerbForm=Fin
gu	käskiv kõneviis olevik 3. isik ainsus aktiiv jaatav kõne	Mood=Imp Tense=Pres Person=3 Number=Sing Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 1. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=1 Number=Plur Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 1. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=1 Number=Sing Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 2. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=2 Number=Plur Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 2. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=2 Number=Sing Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 3. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=3 Number=Plur Voice=Act VerbForm=Fin
ks	tingiv kõneviis olevik 3. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=3 Number=Sing Voice=Act VerbForm=Fin
ksid	tingiv kõneviis olevik 2. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=2 Number=Sing Voice=Act VerbForm=Fin
ksid	tingiv kõneviis olevik 3. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=3 Number=Plur Voice=Act VerbForm=Fin
ksime	tingiv kõneviis olevik 1. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=1 Number=Plur Voice=Act VerbForm=Fin
ksin	tingiv kõneviis olevik 1. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=1 Number=Sing Voice=Act VerbForm=Fin
ksite	tingiv kõneviis olevik 2. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Pres Person=2 Number=Plur Voice=Act VerbForm=Fin
ma	supiin aktiiv jaatav kõne sisseütlev	VerbForm=Sup Voice=Act Case=Ill

maks	supiin aktiiv jaatav kõne saav	VerbForm=Sup Voice=Act Case=Tra
mas	supiin aktiiv jaatav kõne seesütlev	VerbForm=Sup Voice=Act Case=Ine
mast	supiin aktiiv jaatav kõne seestütlev	VerbForm=Sup Voice=Act Case=Ela
mata	supiin aktiiv jaatav kõne ilmaütlev	VerbForm=Sup Voice=Act Case=Abe
me	kindel kõneviis olevik 1. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=1 Number=Plur Voice=Act VerbForm=Fin
n	kindel kõneviis olevik 1. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=1 Number=Sing Voice=Act VerbForm=Fin
neg	eitav kõne	Polarity=Neg
neg ge	käskiv kõneviis olevik 2. isik mitmus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=2 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg gem	käskiv kõneviis olevik 1. isik mitmus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=1 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg gu	käskiv kõneviis olevik 3. isik mitmus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=3 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg gu	käskiv kõneviis olevik 3. isik ainsus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=3 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg gu	käskiv kõneviis olevik passiiv eitav kõne	Mood=Imp Tense=Pres Voice=Pass Polarity=Neg VerbForm=Fin
neg ks	tingiv kõneviis olevik 1. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=1 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg ks	tingiv kõneviis olevik 1. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=1 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin

neg ks	tingiv kõneviis olevik 2. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=2 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg ks	tingiv kõneviis olevik 2. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=2 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg ks	tingiv kõneviis olevik 3. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=3 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg ks	tingiv kõneviis olevik 3. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Pres Person=3 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg me	käskiv kõneviis olevik 1. isik mitmus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=1 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 1. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Past Person=1 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 1. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Past Person=1 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 2. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Past Person=2 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 2. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Past Person=2 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 3. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Past Person=3 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nud	kindel kõneviis lihtminevik 3. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Past Person=3 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin

neg nuks	tingiv kõneviis minevik 1. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=1 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nuks	tingiv kõneviis minevik 1. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=1 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg nuks	tingiv kõneviis minevik 2. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=2 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nuks	tingiv kõneviis minevik 2. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=2 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg nuks	tingiv kõneviis minevik 3. isik mitmus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=3 Number=Plur Voice=Act Polarity=Neg VerbForm=Fin
neg nuks	tingiv kõneviis minevik 3. isik ainsus aktiiv eitav kõne	Mood=Cnd Tense=Past Person=3 Number=Sing Voice=Act Polarity=Neg VerbForm=Fin
neg o	käskiv kõneviis olevik 2. isik ainsus aktiiv eitav kõne	Mood=Imp Tense=Pres Person=2 Number=Sing Voice=Act Polarity=Neg Verbform= Fin
neg o	kindel kõneviis olevik 1. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=1 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg o	kindel kõneviis olevik 1. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=1 Number=Sing Voice=Act Polarity=Neg Verbform=Fin
neg o	kindel kõneviis olevik 2. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=2 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg o	kindel kõneviis olevik 2. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=1 Number=Sing Voice=Act Polarity=Neg Verbform=Fin

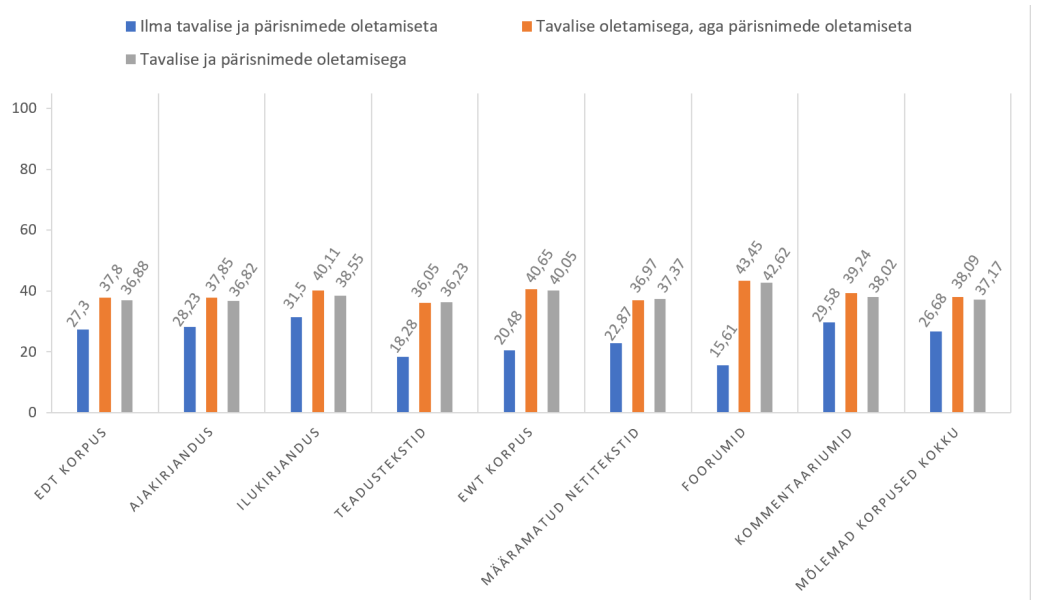
neg o	kindel kõneviis olevik 3. isik mitmus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=3 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg o	kindel kõneviis olevik 3. isik ainsus aktiiv eitav kõne	Mood=Ind Tense=Pres Person=3 Number=Sing Voice=Act Polarity=Neg Verbform=Fin
neg vat	kaudne kõneviis olevik 1. isik mitmus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=1 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg vat	kaudne kõneviis olevik 1. isik ainsus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=1 Number=Sing Voice=Act Polarity=Neg Verbform=Fin
neg tud	kesksõna minevik passiiv eitav kõne	Tense=Past Voice=Pass Polarity=Neg Verbform=Part
neg vat	kaudne kõneviis olevik 2. isik mitmus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=2 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg vat	kaudne kõneviis olevik 2. isik ainsus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=2 Number=Sing Voice=Act Polarity=Neg Verbform=Fin
neg vat	kaudne kõneviis olevik 3. isik mitmus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=3 Number=Plur Voice=Act Polarity=Neg Verbform=Fin
neg vat	kaudne kõneviis olevik 3. isik ainsus aktiiv eitav kõne	Mood=Qot Tense=Pres Person=3 Number=Sing Voice=Act Polarity=Neg Verbform=Fin
nud	kesksõna minevik aktiiv jaatav kõne	Tense=Pass Voice=Act Verbform=Part
nuks	tingiv kõneviis minevik 1. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=1 Number=Plur Voice=Act VerbForm=Fin
nuks	tingiv kõneviis minevik 1. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=1 Number=Sing Voice=Act VerbForm=Fin
nuks	tingiv kõneviis minevik 2. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=2 Number=Plur Voice=Act VerbForm=Fin

nuks	tingiv kõneviis minevik 2. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=2 Number=Sing Voice=Act VerbForm=Fin
nuks	tingiv kõneviis minevik 3. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=3 Number=Plur Voice=Act VerbForm=Fin
nuks	tingiv kõneviis minevik 3. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=3 Number=Sing Voice=Act VerbForm=Fin
nuksid	tingiv kõneviis minevik 2. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=2 Number=Sing Voice=Act VerbForm=Fin
nuksid	tingiv kõneviis minevik 3. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=3 Number=Plur Voice=Act VerbForm=Fin
nuksime	tingiv kõneviis minevik 1. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=1 Number=Plur Voice=Act VerbForm=Fin
nuksin	tingiv kõneviis minevik 1. isik ainsus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=1 Number=Sing Voice=Act VerbForm=Fin
nuksite	tingiv kõneviis minevik 2. isik mitmus aktiiv jaatav kõne	Mood=Cnd Tense=Past Person=2 Number=Plur Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 1. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=1 Number=Plur Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 1. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=1 Number=Sing Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 2. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=2 Number=Plur Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 2. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=2 Number=Sing Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 3. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=3 Number=Plur Voice=Act VerbForm=Fin
nuvat	kaudne kõneviis minevik 3. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Past Person=3 Number=Sing Voice=Act VerbForm=Fin
o	käskiv kõneviis olevik 2. isik ainsus aktiiv jaatav kõne	Mood=Imp Tense=Pres Person=2 Number=Sing Voice=Act VerbForm=Fin
s	kindel kõneviis lihtminevik 3. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=3 Number=Sing Voice=Act VerbForm=Fin

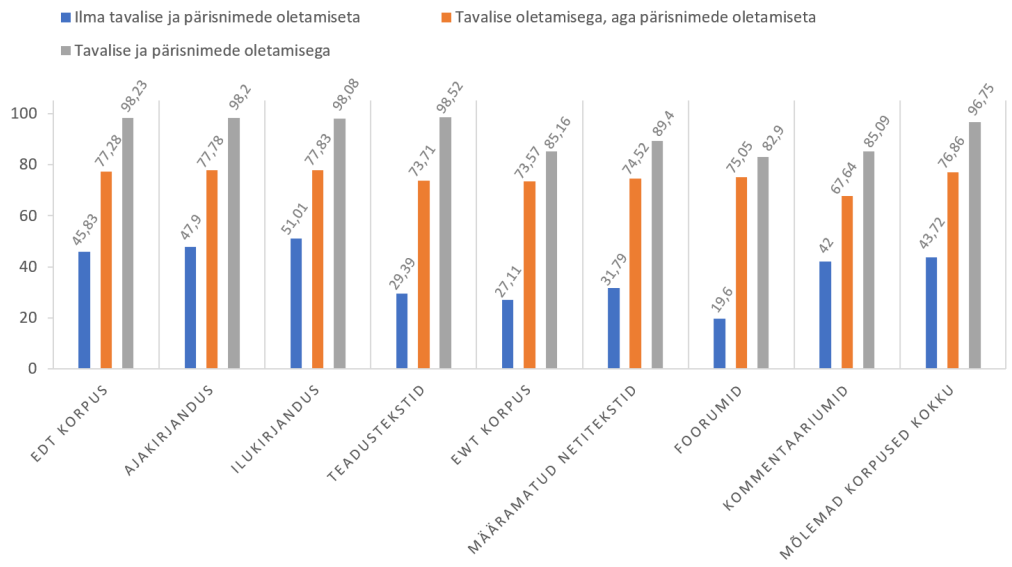
sid	kindel kõneviis lihtminevik 2. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=2 Number=Sing Voice=Act VerbForm=Fin
sid	kindel kõneviis lihtminevik 3. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=3 Number=Plur Voice=Act VerbForm=Fin
sime	kindel kõneviis lihtminevik 1. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=1 Number=Plur Voice=Act VerbForm=Fin
sin	kindel kõneviis lihtminevik 1. isik ainsus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=1 Number=Sing Voice=Act VerbForm=Fin
site	kindel kõneviis lihtminevik 2. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Past Person=2 Number=Plur Voice=Act VerbForm=Fin
ta	kindel kõneviis olevik passiiv eitav kõne	Mood=Ind Tense=Pres Voice=Pass Polarity=Neg VerbForm=Fin
tagu	käskiv kõneviis olevik passiiv jaatav kõne	Mood=Imp Tense=Pres Voice=Pass VerbForm=Fin
taks	tingiv kõneviis olevik passiiv jaatav kõne	Mood=Cnd Tense=Pres Voice=Pass VerbForm=Fin
takse	kindel kõneviis olevik passiiv jaatav kõne	Mood=Ind Tense=Pres Voice=Pass VerbForm=Fin
tama	supiin passiiv jaatav kõne	VerbForm=Sup Voice=Pass
tav	kesksõna olevik passiiv jaatav kõne	Tense=Pres Voice=Pass VerbForm=Part
tavat	kaudne kõneviis olevik passiiv jaatav kõne	Mood=Qot Tense=Pres Voice=Pass VerbForm=Fin
te	kindel kõneviis olevik 2. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=2 Number=Plur Voice=Act VerbForm=Fin
ti	kindel kõneviis lihtminevik passiiv jaatav kõne	Mood=Ind Tense=Past Voice=Pass VerbForm=Fin
tud	kesksõna minevik passiiv jaatav kõne	Tense=Past Voice=Pass VerbForm=Part
tuks	tingiv kõneviis minevik passiiv jaatav kõne	Mood=Cnd Tense=Past Voice=Pass VerbForm=Fin

tuvat	kaudne kõneviis minevik passiiv jaatav kõne	Mood=Qot Tense=Past Voice=Pass VerbForm=Fin
v	kesksõna olevik aktiiv jaatav kõne	Tense=Pres Voice=Act VerbForm=Part
vad	kindel kõneviis olevik 3. isik mitmus aktiiv jaatav kõne	Mood=Ind Tense=Pres Person=3 Number=Plur Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 1. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=1 Number=Plur Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 1. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=1 Number=Sing Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 2. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=2 Number=Plur Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 2. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=2 Number=Sing Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 3. isik mitmus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=3 Number=Plur Voice=Act VerbForm=Fin
vat	kaudne kõneviis olevik 3. isik ainsus aktiiv jaatav kõne	Mood=Qot Tense=Pres Person=3 Number=Sing Voice=Act VerbForm=Fin

Lisa 2. Pärisnime lemmade analüüsi kvaliteet alakriipsuta



Joonis 13. **Pärisnime lemmade kuldstandardiga ühildumise täpsus protsentides (%) liitsõnu osi eraldava alakriipsuta**



Joonis 14. **Pärisnime lemmade kuldstandardiga ühildumise saagis protsentides (%) liitsõnu osi eraldava alakriipsuta**

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kertu Saul,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „EstNLTK morfoloogilise analüsaatori ja ühestaja kvaliteedi hindamine“, mille juhendaja on Siim Orasmaa, 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.

Kertu Saul
02.06.2022