KUIDAS MÄRKSÕNASTADA VANU EESTIKEELSEID TEKSTE?

KÜLLI PRILLOP

1. Sissejuhatus

Tartu Ülikooli eesti keele õppetoolis on käsil eesti vana kirjakeele korpuse loomine. Pelgalt tekstide arvutisse sisestamisest ja Internetti ülesriputamisest ei oleks kirjakeele ajaloo uurijaile kuigivõrd abi. Materjal oleks küll kõigile huvilistele kättesaadav, kuid ilma tekste läbi lugemata vajalikku infot leida üldjuhul ikkagi ei õnnestuks.

Esiteks seetõttu, et meie kirjakeele algusaegadel polnud veel kujunenud üldkehtivaid õigekirjanorme. Üht ja sama häälikukombinatsiooni võidi edasi anda mitut moodi. Näiteks sõna *tuul* esineb Georg Mülleri jutlustes (1600–1606)[1] variantidena *tuhl*, *twl*, *thul* ja *tul*. Sõna *uskuma* nõrgaastmelist tüve on ta kirjutanud *vssu-*, *uβu-* ja *vβu-*, Joachim Rossihnius on oma kirikukäsiraamatutes (1632)[2] aga eelistanud variante *ussu-* ja *usu-*. Sageli on vanad kirjakujud päris ettearvamatud, nt Mülleril *holkidde* 'õlgede', *hoischkab* 'hõiskab', Turu käsikirjas (XVII saj I pool) *ahakett* 'ohakad'.[3]

Teiseks seetõttu, et ennustamatuks võib osutuda ka morfoloogiliste vormide moodustusviis: Mülleril nt *anda ~ andada*, *istwat ~ istuwat* 'istuvad', *hüppas ~ hüppis*, Rossihniusel *minnenut ~ minnut ~ lahenut ~ lennut* 'läinud'.

Kolmandaks seetõttu, et vanades tekstides on sõnu, mida tänapäeval enam ei tunta, nt Rossihniusel *hagema* 'koguma', *hüüs* 'vara', *paimendama* 'hoidma'. Trükiveadki pole sugugi haruldased, nt Rossihniusel *minck* 'ning', Wanradt-Koelli katekismuses (1535) *Noer* 'koer'.[4]

Vana kirjakeele korpusest info leidmist hõlbustaks, kui meil oleks programm, mis suudaks vanu tekste korrektselt analüüsida, või kui tekstid oleks märksõnastatud ja grammatiliselt märgendatud, s.t iga sõna kohta oleks öeldud, mis on tema tänapäevane algvorm, sõnaliik, mis vormis ta on jms. Olen vanade tekstide märksõnastamiseks välja töötanud arvutiprogrammi "Vakker".

2. Tänapäeva kirjakeele ja vana kirjakeele analüüsi erinevused

Arvutuslingvistikas on kasutusel mõisted *lemmatiseerimine, morfoloogiline analüüs* ja *ühestamine*. Lemmatiseerija leiab sõna kõik võimalikud algvormid, morfoloogiaanalüsaator leiab, mis vormis sõna võib olla (nt *peeti* võib olla ainsuse osastav sõnast *peet* või umbisikulise tegumoe lihtminevik sõnast

[1] K. Habicht, V.-L. Kingisepp, U. Pirso, K. Prillop, Georg Mülleri jutluste sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 12. Tartu, 2000.

[2] V.-L. Kingisepp, K. Habicht, K. Prillop, Joachim Rossihniuse kirikumanuaalide leksika. Tartu Ülikooli eesti keele õppetooli toimetised 22. Tartu, 2002.

[3] H. Tennasilm, Turu käsikirja leksika. Tartu, 2002. (Bakalaureusetöö käsikiri Tartu Ülikooli eesti keele õppetoolis.)

[4] E. Ehasalu, K. Habicht, V.-L. Kingisepp, J. Peebo, Eesti keele vanimad tekstid ja sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 6. Tartu, 1997.

pidama). Morfoloogilise ühestaja ülesanne on leitud variantidest õige, s.t konteksti sobiv välja valida. Automaatsel morfoloogilisel ühestajal on mõtet ainult siis, kui on olemas automaatne morfoloogiaanalüsaator.

Automaatse analüsaatori loomiseks on vaja keele formaliseeritud kirjeldust või tekstikorpust, mille alusel see kirjeldus luua. Vana eesti kirjakeele morfoloogiliselt märgendatud korpust meil veel ei ole, formaliseeritud morfoloogiakirjeldusest rääkimata. Me ei tea päris täpselt sedagi, kui palju vana kirjakeel tegelikult tänapäevasest kirjakeelest erineb. Tänapäeva kirjakeele jaoks on morfoloogiaanalüsaatoreid ja ühestajaid rohkem kui üks[5] ja kõigepealt tulekski kindlaks teha, kas nende abil saaks märgendada ka vanu tekste.

Testimiseks kasutasin ESTMORF-i. ESTMORF võrdleb jooksvas tekstis olevaid sõnavorme sõnastikus olevate lekseemide kombinatsioonidega.[6] Sõnastiku aluseks on Ülle Viksi "Väikese vormisõnastiku" elektrooniline versioon. Kõigepealt analüüsisin ESTMORF-iga Heinrich Stahli "Leyen Spiegeli"[7] esimest 4200 eestikeelset sõna. Õige kirjeldus leiti 21%-l tekstisõnadest (sh need, millele saadi mitu vastust).

Esimene pähetulev idee, kuidas saadud tulemust parandada, on viia vanad sõnad tänapäevasele ortograafilisele kujule. Koostasin selleks teisendusreeglid (nt $x \rightarrow ks$, $Vh \rightarrow VV$, $w \rightarrow v$, $uw \rightarrow u$ jne, kokku 210 reeglit).[8] Kasu oli märgatav, nüüd analüüsis ESTMORF õigesti juba 71% eespool nimetatud tekstist. Võimalik, et täpsemaid teisendusreegleid kasutades õnnestuks saada veelgi parem tulemus.

Heinrich Stahli raamat on põhjaeestikeelne. Joachim Rossihniuse lõunaeestikeelsete kirikumanuaalidega[9] ESTMORF nii hästi hakkama ei saanud: õige kirjeldus leiti 62%-l sõnedest (testitav tekstikatke oli samuti 4200-sõnaline, teisendusreegleid oli 216). Abi võiks olla ESTMORF-i sõnastiku täiendamisest vanade ja murdeliste sõnadega. Kuni meil pole vanade sõnade sõnastikku, ei saa me arvutile kuidagi selgeks teha, et *tao* on muutumatu sõna (*kui tao* 'kuidas'), aga mitte sõna *taguma* vorm; et *emmis* on *emmis* 'kuni', mitte *emis* jne. Tänapäeva kirjakeele jaoks mõeldud ESTMORF-i kasutamine andis

---

[5] Morfoloogiaanalüsaatorite kohta vt: H.-J. K a a l e p, Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus 1998, nr 1, lk 22–29; Ü. V i k s, Eesti keele avatud morfoloogiamudel. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu, 2000, lk 9–36; H. U i b o, Kahetasemeline morfoloogiamudel ja eesti keel. – Keel ja Kirjandus 1998, nr 1, lk 13–21. Internetis kättesaadavad morfoloogiaanalüsaatorid: *http://www.eki.ee/tarkvara/analyys/* ja *http://www.filosoft.ee/html_morf_et/*. Ühestajate kohta vt: H.-J. K a a l e p, T. V a i n o, Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. – Keel ja Kirjandus 1998, nr 1, lk 30–36; T. P u o l a k a i n e n, Eesti keele reeglipõhise morfoloogilise ühestamise probleemseid kohti. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu, 2000, lk 73–85; T. R o o s m a a, M. K o i t, K. M u i s c h n e k, K. M ü ü r i s e p, T. P u o l a k a i n e n, H. U i b o, Eesti keele arvutigrammatika: Mis on tehtud ja kuidas edasi? – Keel ja Kirjandus 2003, nr 3, lk 192–209.

[6] H.-J. K a a l e p, Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus 1998, nr 1, lk 22–29.

[7] H. S t a h l, Leyen Spiegel. Revall, 1641.

[8] Reeglid koostasin paarisaja esimese sõne põhjal ja seejärel üldistasin neid. Nt kui esimeste lehekülgede põhjal selgus, et *a* tuleks vahel asendada *aa*-ga ja *e ee*-ga, siis lisasin ka reeglid $o \rightarrow oo$, $u \rightarrow uu$, $i \rightarrow ii$.

[9] J. R o s s i h n i u s, Südestnische Uebersetzung des Lutherischen Katechismus, der Sonntags-Evangelien und -Episteln und der Leidengeschichte Jesu nebst einem Anhang in das Südestnische übersetzter Kirchenlieder und Stücke der Agende mit einer Einleitung von Wilhelm Reiman. Herausgegeben von der Gelehrten Estnischen Gesellschaft. Jurjew (Dorpat), 1898. (Tegemist on taastrükiga, originaal ilmus 1632.) Vt ka *http://www.murre.ut.ee/vakkur/Korpused/koond.htm*.

üllatavalt hea tulemuse. Kuid kas on mõnd meetodit, mida rakendades saaks automaatselt analüüsitud veelgi suurema osa tekstist?

Kõige lihtsam võimalus tekste morfoloogiliselt analüüsida on kasutada tabelit, mille esimeses veerus on sõnavormid, teises neile vastavad algvormid ja kolmandas grammatiline info:

lähen	minema	verb, kindel kõneviis, olevik, ainsus, 1. pööre
läheb	minema	verb, kindel kõneviis, olevik, ainsus, 3. pööre

Mingi sõna analüüsi leidmiseks tuleb see sõna tabeli esimesest veerust üles otsida. Sama rea teises veerus on sobiv algvorm, kolmandas vormiinfo. Kui tahaksime sel viisil analüüsida tänapäevases eesti keeles teksti, võiks sõnavormide leksikoni loomiseks kasutada mõnd olemasolevat sõnastikku ja genereerida sellest kõikvõimalikud sõnavormid. Kuid 35 000 algvormist saaksime umbes 1,2 miljonit sõnavormi. Tuletiste ja liitsõnade lisamine viiks sõnade hulga miljarditesse. Niisugune lahendus ei ole otstarbekas, programm töötaks liiga aeglaselt.[10] Vana kirjakeele puhul pole selline lahendus ka teoreetiliselt võimalik: meil lihtsalt ei ole sõnastikku, mille põhjal vorme genereerida. Samuti pole meil vormide genereerimise algoritmi.

Teine võimalus oleks võtta aluseks mingi hulk tekste, teha nende alusel sõnavormide leksikon ja loota, et saadud sõnavormide hulk katab piisavalt suure osa ka tundmatute tekstide sõnavarast. Heiki-Jaan Kaalep on kirjeldanud üht sellist katset.[11] Leksikoni aluseks võeti 450 000 sõna ulatuses teksti tänapäeva eesti kirjakeele korpusest,[12] sõnastikku saadi 118 000 kirjet. Selle abil analüüsiti George Orwelli "1984" eestikeelset tõlget.[13] Tulemused on järgnevas tabelis.

Tekstis sõnu kokku	80 000
Neist oli leksikonis	86%
Neist puudus leksikonist	14%

On raske oletada, kas vana kirjakeele puhul oleks need näitajad paremad või halvemad: vanade tekstide sõnavara on piiratum, kuid varieerumist on rohkem. J. Rossihniuse kirikumanuaalides kulub poole teksti katmiseks alla 90 sagedama sõnavormi, G. Orwelli romaanis aga üle 300. Kuid vaid 47% Georg Mülleri jutlusetekstide sõnedest esineb ka J. Rossihniusel. Heinrich Stahli "Leyen Spiegeli" eespool mainitud katkendis esinevatest sõnedest on Georg Müller oma jutlustes kasutanud 41%. Seega võib oletada, et kirjeldatud meetod sobib vana kirjakeele tekstide analüüsimiseks vaid sel juhul, kui leksikon on tehtud sama teksti põhjal, mida analüüsima hakatakse. Ja kuidagi peab ju morfoloogiliselt analüüsima ka neid tekste, mis leksikoni aluseks võetakse.

Morfoloogilise vormi saab paljudel juhtudel üheselt kindlaks määrata sõna lõpu järgi. Tõsi, ainult sel juhul, kui eelnevalt on teada sõnaliik ja algvorm. Näiteks *punast* on seestütleva käände vorm, kui lemmaks on *puna*, aga osastava vorm, kui lemmaks on *punane*; *määrata* on infinitiiv, kui tegemist on verbiga, aga ilmaütlev, kui noomeniga. Kui *sse*-lõpulise sõna algvorm on *ne*-lõpuline,

[10] H.-J. K a a l e p, Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. Dissertationes philologiae estonicae Universitatis Tartuensis 7. Tartu: Tartu Ülikooli Kirjastus, 1999, lk 21.

[11] H.-J. K a a l e p, Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös, lk 21–23.

[12] *http://www.cl.ut.ee/ee/corpusb/tykk.html*.

[13] *http://www.cl.ut.ee/ee/1984/*.

võib vanade tekstide puhul tegemist olla hoopis omastava, mitte sisseütleva käändega (nt *suhre waisusse sees*). Kui lemma on teada, ei peeta vormi *raud d*-lõpu tõttu osastava ega mitmuse vormiks. Konteksti arvestades saaks reegleid veelgi täpsustada. Niisugune analüsaator on vana kirjakeele jaoks üsna lihtsa vaevaga tehtav. Katsetasin jällegi "Leyen Spiegeli" esimese 4200 eestikeelse sõnega. Selles tekstilõigus on muutumatuid sõnu 27%, käändsõnu 52% ja pöördsõnu 21% (samamoodi jagunevad sõnaliigiliselt ka J. Rossihniuse kirikumanuaalid ja võib oletada, et umbes sama jaotus kehtib teistegi tekstide puhul). Ainult sõnalõppe, märksõnu ja sõnaliike arvestades leiti õige vormikirjeldus 98%-l tekstisõnadest, ühele sõnele pakuti maksimaalselt kolm erinevat analüüsi (üks ja õige analüüs leiti 65%-l tekstisõnadest). Tulemus on parem kui ESTMORF-i kasutades saadu.

Erinevalt tänapäevastest võiks vana kirjakeele tekste analüüsida kahes etapis: kõigepealt leida algvorm ja sõnaliik ning alles seejärel, kui need on õigeks tunnistatud, otsustada, millise grammatilise vormiga on tegu. Järgnevas keskendungi just märksõnastamisele.[14] Vana kirjakeele tekste saab märksõnastada kas käsitsi või poolautomaatselt. (Poolautomaatseiks nimetan kõiki neid programme, mille töösse inimene kuidagi sekkub.) Automaatne märksõnastamine on mõeldamatu, kuni meil pole piisavalt suurt sõnastikku. 100%-line täpsus pole saavutatav ka sõnastiku abil.

3. Poolautomaatsed märksõnastajad

3.1. Sõneloendi märksõnastamine. Esmapilgul lihtsaim näib olevat teha tekstide põhjal sõneloend. Siis saaks iga sõne järele kirjutada tema algvormi ja saadud tabeli abil teksti automaatselt märksõnastada. Ajaline võit tundub olevat suur, näiteks *olema* esineb väga sageli, lemma tuleb aga kirjutada ainult igale vormile eraldi, mitte igale esinemusele eraldi. (Georg Mülleri jutlustes esineb *olema* kokku 4073 korda, erinevaid kujusid on tal sellest sõnast aga ainult 33, s.t märksõna tuleks sisestada 33, aga mitte 4073 korda.)

Sel viisil töötab nt R. J. C. Watti loodud programm "Concordance",[15] mis koostab teksti(de) põhjal sõnesagedusloendi ja näitab seda tööakna vasakus servas. Kui klõpsata mõnel loendis oleval sõnal, näidatakse selle sõna kõik esinemused (kuvatava konteksti pikkust saab muuta). Vaadata saab ka kogu teksti, mõnel sõnal klõpsates näidatakse jällegi selle sõna kõiki esinemusi. Lemmatiseerimiseks avatakse uus aken, kuhu tuleb hakata sõneloendist sõnesid hiire abil lohistama, ühe märksõna kõik vormid ühte kohta. Nii valmib sõnastik, mille abil saab tekste lemmatiseerida. Lemmatiseerimine tähendab selles programmis, et sõneloendi sõnad järjestatakse nii, et üksteisele järgnevad sama sõna erinevad vormid. Oluliseks puuduseks on, et homonüüme ei ole võimalik eristada, s.t sama sõne võib küll mitme erineva märksõna alla panna, kuid ei saa öelda, millistel konkreetsetel juhtudel üks või teine märksõna tegelikult sobib.

Sõneloendi märksõnastaja on ka minu enese loodud programm "Mollerus", mida on vana kirjakeele korpust tehes kasutatud alates 1997. aastast. Töö käigus on ilmnenud, et kirjeldatud meetod on veaohtlik: isegi kogenud uurija ei pruugi märgata, et *olema* võib mõnes lauses olla substantiiv, et *liiwa* võib olla *leib*, et *Leib* võib olla saksa tsitaatsõna ('keha'), et *kena* võib olla hoopiski *käänama* jne. Vigade vältimiseks peaks iga sõne kõik esinemused läbi vaa-

[14] Siinses käsitluses hõlmab märksõnastamine ka sõnaliigi määramist.

[15] "Concordance" on saadaval aadressilt *http://www.rjcw.freeserve.co.uk* (tasuta katsetamisperiood 30 päeva).

tama, mis tähendab, et ajalist võitu tegelikult ei olegi. Pigem on ajakulu isegi suurem kui siis, kui jooksvasse teksti iga sõna järele selle lemma lisada. Viimasel juhul peab iga lauset lugema ainult ühe korra, kuid esimesel juhul nii mitu korda, kui mitu sõna selles lauses on. Muidugi võiks märksõnastatud tekstid pärast üle kontrollida, kuid milleks neist siis üldse eelnevalt sõneloendit teha?

3.2. Jooksva teksti märksõnastamine. Jooksva teksti märksõnastamine on sobiv just siis, kui nagunii peab kõik tekstid algusest lõpuni läbi lugema. Muidugi ei tähenda see, et sagedastele sõnedele tuleks lemma iga kord uuesti sisestada. Kui mingi sõne on korra juba esinenud, saab järgmistele samasugustele lemma lisada automaatselt ja vajadusel peab saama seda ka muuta.

Selline programm on näiteks ”LinguaSy Power Concordancer”.[16] Programmi märgendusdialoogis näidatakse tekstisõna ja selle konteksti. Sõnale saab lisada suvalise märgendi ja saab valida, kas see märgend lisatakse ainult sellele sõnale, faili kõikidele sellistele sõnadele või kõikide failide kõigile sellistele sõnadele. Viimasel juhul lisatakse sõna koos oma märgendiga nn universaalsesse loendisse. Teksti juba lisatud märgendite ja universaalse loendi põhjal saab ülejäänud faili automaatselt analüüsida.

Niisugune märksõnastamisviis tundus katsetamisel küllalt mugav ja kiire, kuid polnud siiski päris veatu. Suurimaks probleemiks on keeles esinev polüseemia ja homonüümia. Iga märksõna juures peab hoolikalt mõtlema, kas see saab esineda ka mõnes muus tähenduses ja kas neid tähendusi oleks vaja eristada. Kui jah, siis peab sõnale lisama ka tähendusseletuse. Abiks oleks mitmetähenduslike märksõnade loend, mida võiks kontrollida automaatselt.

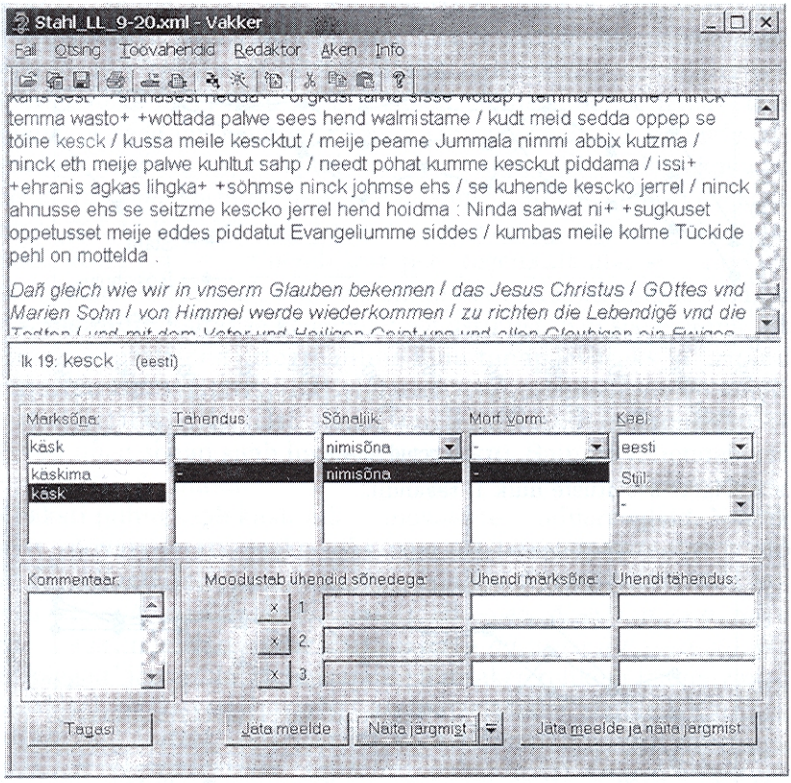
4. ”Vakker”, vanade eestikeelsete tekstide märksõnastaja

Eelmises punktis on tähelepanu pööratud vaid kõige lihtsamatele poolautomaatse märksõnastamise võimalustele. Tegelikkuses peaks kindlasti ära kasutama ka tänapäeva keele analüsaatorid ja võimalikult palju sellest, mida me juba teame vana kirjakeele kohta. Just seda olen püüdnud teha ”Vakkerit” programmeerides.

Joonisel 1 on näha ”Vakkeri” põhiline tööaken. Akna ülemises osas paikneb jooksev tekst. Sõna, mille analüüsimine parasjagu käsil, on arvuti ekraanil punane. Akna keskmises osas asub info, mis on selle sõna kohta faili juba salvestatud (sõna *kesck* kohta pole veel salvestatud muud, kui et ta on eestikeelses tekstis). Akna alumine osa on info sisestamiseks (lemma, tähendus, sõnaliik jm), samas näeb ka soovitusi. Sõna *kesck* lemmaks pakub programm kaht varianti: *käskima* ja *käsk*. Kui muuta lemmat, muutuvad automaatselt ka pakutavad tähendused ja sõnaliik. Järgmise tekstisõna saab valida kas ülemisest aknast hiireklõpsuga või alumises aknas vastavale nupule klõpsates. On võimalik liikuda lihtsalt järgmisele sõnale, järgmisele eestikeelsele sõnale, järgmisele samasugusele sõnale, järgmisele märksõnastamata sõnale või mingitele muudele kriteeriumidele vastavale sõnale.

Soovituste leidmiseks on kolm tasandit. Igaüks saab ise valida, millist/milliseid neist ta kasutada soovib.

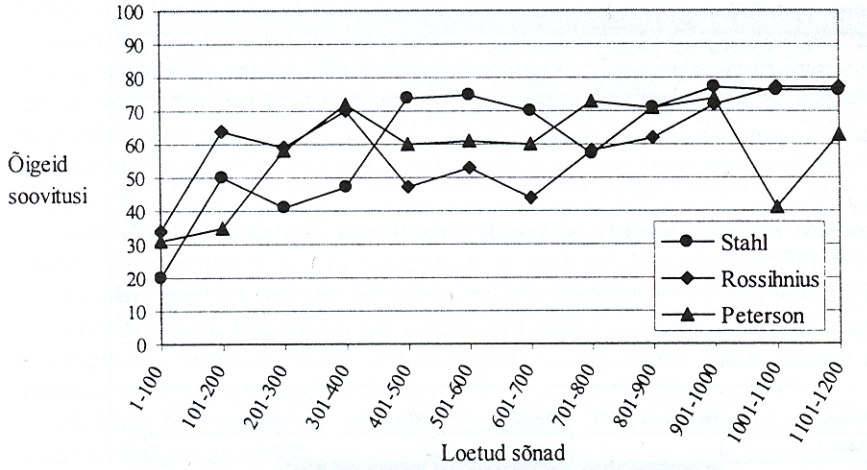
[16] Selle programmi saab aadressilt *http://www.free-esl.com/call/software.asp?swIndex=6*.



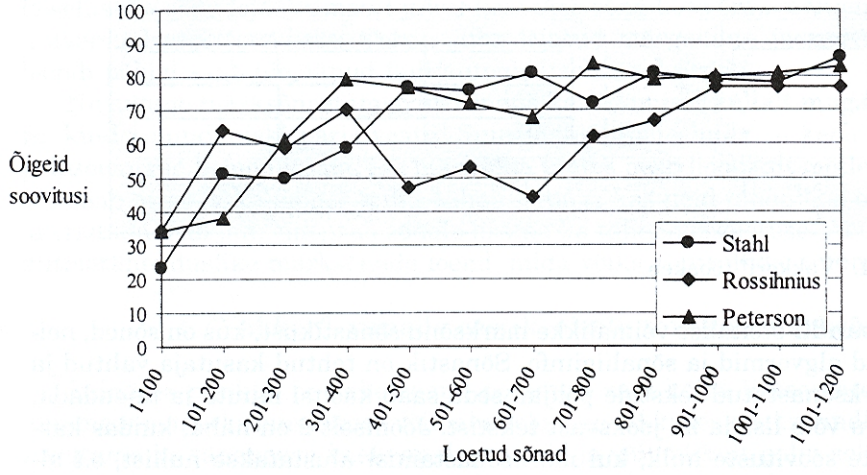
Joonis 1. "Vakkeri" tööaken.

1. tasandil otsitakse võimalikke märksõnu sõnastikust, kus on sõned, neile sobivad algvormid ja sõnaliigiinfo. Sõnastik on tehtud kasutaja valitud ja juba märksõnastatud tekstide põhjal, seda saab käsitsi muuta ja täiendada. Uusi sõnu võib lisada ka jooksvast tekstist. Jooniselt 2 on näha, kuidas kasvab õigete soovituste hulk, kui märksõnastamist alustatakse nullist, s.t alguses pole sõnastikus ühtegi sõna: iga märksõna, mis sisestatakse, lisatakse ka sõnastikku.[17] Nii täieneb sõnastik vastavalt sellele, kui palju tekstist on loetud. Esimest sadat sõna lugedes pakutakse õige lemma vaid 20–35%-l sõnadest. Kui loetud on juba 900 sõna, on õigete soovituste hulk vähemalt H. Stahli ja J. Rossihniuse tekstide puhul üle 70%. Hiljem õigete soovituste hulk enam märgatavalt ei suurene, näiteks pärast 4000 sõna "Leyen Spiegeli" lugemist on neid ikka alla 80%. On selge, et kui sõnastik on väike, pole 1. tasandist kuigivõrd abi. Mida rikkalikum on teksti sõnavara või mida varieeruvam on ortograafia, seda kehvem on tulemus.

[17] Siin ja edaspidi on testimiseks kasutatud kaht ebakorrapärases kirjaviisis teost (eelkõige selliste tekstide analüüsimiseks ongi "Vakker" mõeldud): Heinrich Stahli "Leyen Spiegelit" (vt viide 7) ja Joachim Rossihniuse kirikumanuaale (vt viide 9). Võrdlusmaterjali pakkumiseks on analüüsitud ka üht uuemat kirjapanekut: Kristian Jaak Petersoni päevaraamatut (*http://www.eki.ee/peterson/4.html*). Tekstid on võetud iga raamatu algusest. Eelnevalt on märgendatud võõrkeelsed lõigud, ääremärkused, pealkirjad ja liitsõnaosade piirid. Kõik K. J. Petersoni teosed avaldati aastal 2001 raamatus "IAAK. Kristian Jaak Peterson 200". Samas ilmus ka Ülle Viksi, Ene Vainiku ja Indrek Heina koostatud ligi 200-leheküljeline "K. J. Petersoni sõnastik". Sõnastiku loomise eeltööks oli tekstide märgendamine (märksõnastamine ja morfoloogiline analüüs). Selleks kasutati poolautomaatset arvutiprogrammi, mille täpset kirjeldust kahjuks pole avaldatud. (Sõnastiku eessõnast lk 237 selgub, et tegemist ei ole nn jooksva teksti analüsaatoriga nagu "Vakker".)



Joonis 2. Õigete soovituste hulk 1. tasandil.



Joonis 3. Õigete soovituste hulk 2. tasandil.

Üritasin saadud tulemust parandada teisendusreegleid appi võttes. Reeglid määraks, et *ck* = *k*, *bb* = *b*, *tz* = *ts* jne. Selgus, et autoreil on siiski omad ortograafiaeelistused, näiteks J. Rossihnius on kirjutanud küll nii *ninck* kui ka *nink* 'ning', kuid *ninck* 3209 korda, *nink* vaid 4 korda. Teisendusreeglite kasutegur jäi alla 1%, seega loobusin neist.

2. tasandil otsitakse võimalikke märksõnu kõigepealt samuti sõnastikust, aga kui otsitavat sõna ei leita, korratakse otsingut sõnalõppe arvestamata. Näiteks kui sõnastikus on olemas vorm *olen* ja ignoreeritakse lõppe *n* ja *me*, leiab programm märksõna ka vormile *oleme*. Lõpud, mida ei arvestata, saab kasutaja ise määrata. Seda võib teha jooksvalt, teksti märksõnastamise ajal. Arvestama peab, et mida rohkem lõppe ignoreeritakse, seda aeglasem on otsing. Lõppude arvestamata jätmisest oli kõige rohkem kasu K. J. Petersoni päevaraamatu analüüsimisel (tulemus paranes keskmiselt 11%), J. Rossihniuse analüüsil oli kasu vaid 1%. Kasu on seda suurem, mida reeglipärasem on teksti vormistik ja kirjaviis. Mida rohkem sõnu on sõnastikus, seda väiksemaks kasutegur muutub.

Olgu esitatud ka ignoreeritavate lõppude loend, nagu see oli pärast "Leyen Spiegeli" 4200 sõna lugemist: *a*, *ckut*, *d*, *da*, *dda*, *de*, *dt*, *e*, *i*, *idt*, *is*, *it*, *kem*,

ket, *ko*, *kut*, *l*, *le*, *lle*, *lt*, *m*, *ma*, *mb*, *me*, *mma*, *mme*, *n*, *nut*, *o*, *p*, *s*, *se*, *sime*, *simme*, *sit*, *sse*, *st*, *t*, *ta*, *te*, *tud*, *tut*, *ust*, *wat*, *x*, *xe*, *xit*. Ainult vale lemma pakuti seda loendit kasutades 1,5%-le Stahli sõnedest.

Muidugi võiks programm ise aru saada, milliseid lõppe ignoreerida. Kahjuks selgus, et seda on oodatust raskem ellu viia. Näiteks vormide *andma* ja *annap* põhjal otsustataks, et lõpud on *dma* ja *nap*, millest oleks kasu veel vaid *kandma* ja *kündma* vormide märksõnastamisel. Kui loendisse on kogunenud piisavalt lõppe, s.t kui seal on juba olemas ka *ma* ja *p*, võiks lõpust *dma* ja *nap* kustutada. Kuid sellestki täiendusest pole tegelikult abi, sest mida rohkem tekstist on märksõnastatud, seda vähem mõtet on lõppude ignoreerimisel.

Ka **3. tasandil** otsitakse kõigepealt sõnastikust. Kui see ei anna tulemust, teisendatakse tekstisõna ortograafiline kuju tänapäevaseks ja üritatakse seda analüüsida ESTMORF-i abil. Eelnev sõnastikust otsimine on vajalik, sest nii lemmatiseeritakse ka need sõnad, mida ESTMORF ei tunne (kõik tekstis ettetulevad uued sõnad saab lisada sõnastikku). Samuti vähendab see pakutavate märksõnade hulka.

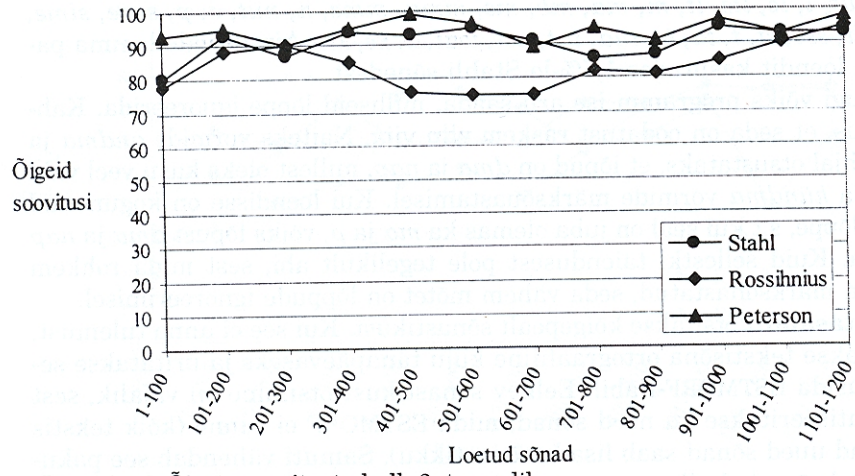
Iga teksti puhul saab kasutada erinevaid teisendusreegleid. Näiteks K. J. Petersoni päevaraamatu jaoks läks vaja 29 teisendust (*gg* → *g*, *bb* → *b*, *dd* → *d*, *ll* → *l*, *mm* → *m*, *nn* → *n*, *rr* → *r*, *hh* → *h*, *ss* → *s*, *kk* → *k*, *tt* → *t*, *pp* → *p*, *o* → *u*, *'*→ *h*, *e* → *õ*, *w* → *v*, *i* → *j*, *eä* → *ea*, *ö* → *õ*, *o* → *õ*, *nd* → *nud*, *a* → *aa*, *e* → *ee*, *i* → *ii*, *o* → *oo*, *u* → *uu*, *ä* → *ää*, *ü* → *üü*, *ö* → *öö*), H. Stahli "Leyen Spiegeli" jaoks 210. Sel tasandil pakutakse rohkem lemmavariante kui kahel eelmisel, nt "Leyen Spiegelis" keskmiselt kaks, sõnale *hoitut* koguni 11 (subst *hoitu*, adj *hoitud*, verb *hoidma*, verb *õiduma*, adj *uitu*, subst *uid*, subst *uit*, subst *oid*, verb *hoiduma*, adj *huitu*, subst *hoid*). Õigeid soovitusi on algusest peale üle 75%, K. J. Petersoni puhul koguni üle 90% (vt joonis 4).

Õigete soovituste hulk on üsna stabiilne seetõttu, et teisendusreeglid on eelnevalt kindlaks määratud. Kui reegleid lisataks ja täiendataks jooksvalt teksti märksõnastamise ajal, oleks õigete soovituste hulk alguses 30% piirimail nagu ka teiste tasandite puhul. Nii ei ole tehtud seetõttu, et H. Stahli reeglid sobivad enam-vähem samal kujul ka teistele ebakorrapärases kirjaviisis tekstidele (Rossihniuse analüüsimiseks on Stahli reeglitele lisatud kuus teisendust), Petersoni reeglid peaksid aga sobima vanas kirjaviisis tekstidele.

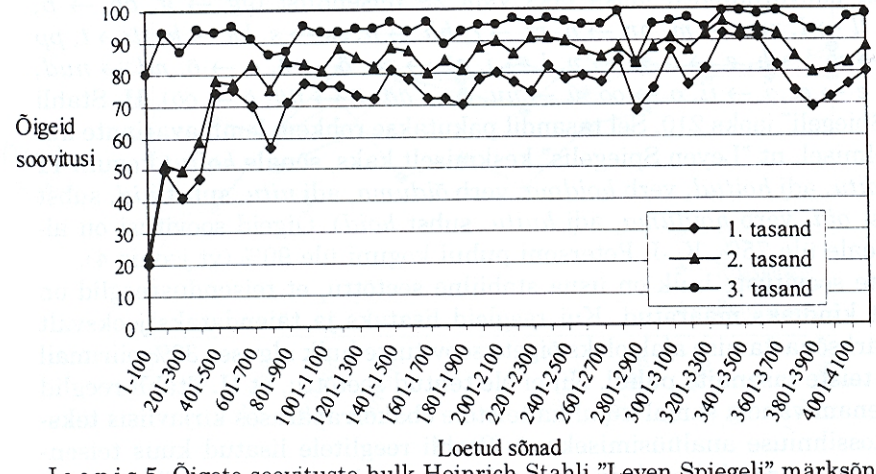
Joonis 5 kujutab erinevate tasandite võrdlust. Näha on seegi, et kui loetud on umbes 1000 sõna, õnnestub edaspidi 3. tasandil õigesti analüüsida üle 90% tekstist.

Kui sõnastiku ja ESTMORF-i abil on võimalik saada niivõrd hea tulemus, mis kasu on siis üldse 2. tasandist? Kas 2. ja 3. tasandi koos kasutamine parandaks tulemust veelgi? Põhimõtteliselt on kolm võimalust: (1) kasutada 3. tasandit ainult nende sõnade puhul, mis 2. tasandil jäid analüüsimata, (2) kasutada 2. tasandit ainult nende sõnade puhul, mis 3. tasandil jäid analüüsimata, (3) kasutada kõikide sõnade puhul nii 2. kui ka 3. tasandit.

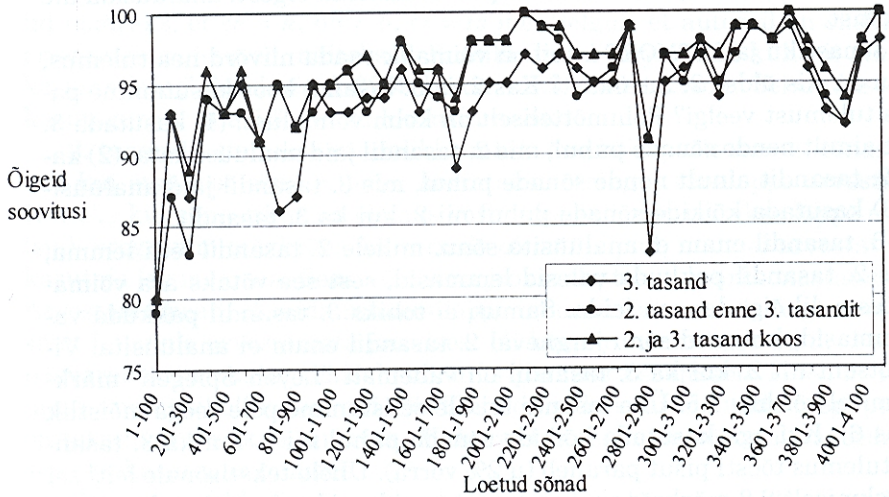
Kui 3. tasandil enam ei analüüsita sõnu, millele 2. tasandil leiti lemma, ei tohiks 2. tasandil pakkuda valesid lemmasid, sest see võtaks ära võimaluse 3. tasandil õige lemma leida. Samuti ei tohiks 3. tasandil pakkuda valesid lemmasid, kui neid sõnu järgneval 2. tasandil enam ei analüüsita. Vigade protsent nii 2. kui ka 3. tasandil oli vähemalt "Leyen Spiegeli" märksõnastamisel võrdne: 1,5. Ühe tasandi teisele eelistamine pole seega mõistlik (vt joonis 6). Kui aga kasutada kõikide sõnade puhul nii 2. kui ka 3. tasandit, siis tulemus tõesti pisut paraneb (u 2% võrra). Ühele tekstisõnale leiti sel juhul keskmiselt 2,2 märksõna (arvestamata neid, mida ei suudetud analüüsida), neist õige valimine jääb programmi kasutaja ülesandeks.



Joonis 4. Õigete soovituste hulk 3. tasandil.



Joonis 5. Õigete soovituste hulk Heinrich Stahli "Leyen Spiegeli" märksõnastamisel.



Joonis 6. Õigete soovituste hulk Heinrich Stahli "Leyen Spiegeli" märksõnastamisel (tasandite kombineerimine).

Pakutavate variantide järjekord on suvaline, sellest hoolimata oli õige variant esimene rohkem kui 80%-l juhtudest. (Kui õige märksõna soovitatakse esimesena, peab "Vakkeri" kasutaja tegema paar klahvivajutust vähem.) Soovitusi võiks püüda kuidagi targalt järjestada. Näiteks *tud*-partitsiipidele pakutakse alati vähemalt kolm algvormi varianti: esimesena *tu*-liiteline substantiiv, teisena adjektiiv ja alles kolmandana verb. Õige on enamasti just kolmas variant. Kahjuks on see ainus reegel, mis testimise käigus tuvastada õnnestus, ja ainult üks reegel ei muudaks tulemust paremaks.

5. Kokkuvõtteks

Tänapäevaste kirjakeelte automaatsed analüsaatorid tuginevad sõnastikele ja teadaolevatele või korpuste põhjal leitud reeglitele. Vana kirjakeele jaoks meil sellist algmaterjali pole, seega pole vähemalt esialgu automaatne analüüs võimalik. Arvestama peab ka seda, et vanu tekste on piiratud hulk, ainult vahel harva leitakse mõni uus dokument. Pole mõtet plaanida analüsaatorit, mille väljatöötamiseks ja testimiseks on vaja suurem osa olemasolevatest tekstidest käsitsi märgendada.

"Vakker" on poolautomaatne jooksva teksti märksõnastaja, mis on loodud vana kirjakeele korpust silmas pidades. See on esimene tõsisem katse vanimate tekstide märgendamist automatiseerida. Märksõnastamisel kasutatakse töö käigus pidevalt lisanduvaid teadmisi analüüsitava teksti ortograafia ja vormitunnuste kohta, samuti tänapäeva kirjakeele jaoks mõeldud morfoloogiaanalüsaatorit ESTMORF. "Vakker" töötab seda tulemuslikumalt, mida rohkem tekstist on juba analüüsitud. Esialgsete testide põhjal võib väita, et pärast umbes 1000 tekstisõna analüüsimist leitakse õige algvorm umbes 95%-le selle teksti sõnadest (sh need, millele leitakse mitu võimalikku märksõna). Rohkem kui 80%-l juhtudest pakutakse õige algvorm esimesena. Tegelikkuses tuleb kindlasti ette nii selliseid kirjutisi, mida analüüsitakse kehvemini, kui ka selliseid, mida analüüsitakse paremini. Kui tekstisõna kohta on teada tema märksõna (algvorm) ja sõnaliik, saab lihtsa vaevaga lisada ka vormiinfo.