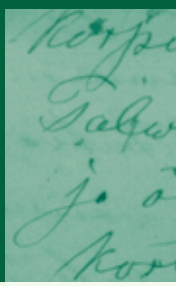
Korpused lingvistilises uurimistöös

Maarja-Liisa Pilvik, Kadri Muischnek, Liina Lindström,
Ann Veismann, Jane Padrik, Eleri Aedmaa, Jelena Kallas,
Kristina Koppel, Margit Langemets, Pärtel Lippus,
Siim Orasmaa, Peeter Tinits, Joshua Wilbur



PRAKTILINE KÄSIRAAMAT

Korpused lingvistilises uurimistöös

PRAKTILINE KÄSIRAAMAT

Korpused lingvistilises uurimistöös

PRAKTILINE KÄSIRAAMAT

Autorid:
Maarja-Liisa Pilvik, Kadri Muischnek, Liina Lindström,
Ann Veismann, Jane Padrik, Eleri Aedmaa, Jelena Kallas,
Kristina Koppel, Margit Langemets, Pärtel Lippus,
Siim Orasmaa, Peeter Tinits, Joshua Wilbur

Koostanud ja toimetanud:
Liina Lindström, Maarja-Liisa Pilvik, Ann Veismann,
Kadri Muischnek, Jane Padrik

Õpiku valmimist on toetanud:
Haridus- ja Teadusministeeriumi programm „Eestikeelsete kõrgkooliõpikute loomise toetamise põhimõtted 2018–2027“, mida koordineerib Eesti Keele Instituut;
Tartu Ülikooli eesti ja üldkeeleteaduse instituudi Kadri, Nikolai ja Gerda Rõugu Fond;
Tartu Ülikooli humanitaarteaduste ja kunstide valdkond, kirjastamistoetus;
Haridus- ja Teadusministeeriumi terminitöö programm, mida koordineerib Eesti Keele Instituut.

Koostajad ja toimetajad: Liina Lindström, Maarja-Liisa Pilvik, Ann Veismann, Kadri Muischnek, Jane Padrik

Kujundaja: Kairi Kullasepp

Retsensendid: Krista Kerge, Kadri Vider

ISBN 978-9908-57-130-0
ISBN 978-9908-57-131-7 (pdf)

Tartu Ülikooli Kirjastus, 2026

Sissejuhatus

Tänapäevase keelelise uurimistöö üks olulisemaid materjaliallikaid on keelekorpused – korrastatud ja lisaandmetega varustatud tekstikollektsioonid. Korpused annavad meile esmase info keele kasutamise kohta võrdlemisi kiiresti – me ei pea iga uut uurimust alustama andmete korjamisest, vaid saame kasutada juba olemasolevaid tekstikollektsioone. Võrreldes ajaga, mil uurimistöö tegemiseks kirjutati huvipakkuvad keelendid esmalt sedelitele, seejärel lisati sinna vajalik analüüs ja alles seejärel sai neid süvendatult analüüsima hakata, on korpuste kasutamise võimalus kogu uurimistöö kiirust väga palju muutnud: korpusest tulevad andmed on enamasti saanud juba mingi analüüsikihi (näiteks sõnaliigi, käände- või vormitunnused) ning välja on töötatud palju meetodeid, mis võimaldavad saadud andmeid ka kvantitatiivselt analüüsida.

Keelekorpused, mida hakati looma ja kasutama juba 1960ndatel, on omakorda muutnud keeleteadust ja toonud sinna nii praktilisi kui teoreetilisi teadmisi keelekasutuse ja selle varieerumise kohta. Keelekorpusi kasutatakse tänapäeval nii keele uurimiseks kui ka näiteks sõnaraamatute koostamiseks või keeletehnoloogia vahendite arendamiseks. Korpused pakuvad aga palju võimalusi ka naaberdistsipliinidele, mis mingil kujul tekstiandmeid kasutavad.

Korpustega koos on arenenud ka korpuslingvistika, millest on saanud tänapäeva keeleteaduse lahutamatu osa. Korpuslingvistikat võib käsitleda meetodite komplektina – see sisaldab akumuleerunud teadmisi ja meetodeid korpuste kasutamiseks ja nende põhjal keele (või maailma) kohta adekvaatsete järelduste tegemiseks. Selleks, et korpusi edukalt kasutada, on vaja teadmisi, mida kõike korpuste abil üldse uurida saab, mis vahendeid selleks on olemas ning milliseid meetodeid on ühe või teise uurimisülesande jaoks sobiv kasutada. Kogu see valdkond on plahvatuslikult kasvanud eriti viimastel aastakümnetel, mil digitaalseid keeleandmeid on meie ümber väga palju ja soov neid uurimistöös kasutada järjest kasvab.

Ehkki juhendmaterjale ja õpikuid on sel teemal kirjutatud väga palju, ei ole meil seni olnud süsteemset eesti keeles kirjutatud ja eesti keele jaoks kohandatud korpuslingvistika õppevara. Selle õpikuga täidame seda lünka. Kuna tegemist on esimese sellelaadse õpikuga, on see suunatud pigem algajale õppijale, kes korpuste ja korpuslingvistika maailmaga alles tutvub.

Selle õpiku suurima sihtrühmana näeme üliõpilasi, kes plaanivad korpusi kasutada oma uurimistöö jaoks. Ennekõike peame silmas keeleteaduse või laiemalt keeltega tegelevaid üliõpilasi, aga mitte ainult – meetodid sobivad mitmesuguste (digi)humanitaaria ülesannete lahendamiseks.

Õpik koosneb kuuest sisupeatükist ja näidisuurimustest. Kui sisupeatükid tutvustavad põhimõisteid ja -meetodeid ja on mõeldud pigem neile, kelle jaoks kogu korpustega töötamine on uus, siis näidisuurimused sobivad ka neile, kel on korpuste kasutamisest uurimistöös olemas juba mõningane ettekujutus ja isiklik kogemus, ent kellel on huvi või vajadus just teatud tüüpi korpuslingvistika ülesannet lahendada või soov saada laiemat pilti korpuslingvistika võimalustest. Õpikusse kaasatud näidisuurimused on valitud nii, et need esindaks võimalikult mitmekesiselt erinevate keeleteaduse suundade uurimistemaatikat ja -meetodeid ning lahendaks üht või mitut seotud uurimisküsimust eesti keele korpuste abil. Kuivõrd kogu keeleandmete maailma on viimastel aastatel mõjutanud suurte keelemudelite kättesaadavus, oleme ühe näidisuurimuse pühendanud ka neile. On aga tõsi, et kogu see maailm muutub väga kiiresti, mistõttu saame pakkuda vaid väikest sissevaadet keelemudelite kasutamisse.

Õpiku pealkirjas sisaldub väljend *praktiline käsiraamat* ja just nii olemegi seda planeerinud: selles esitatud info on praktiline, see sisaldab palju näiteid ning selles leiduvat on võimalik üsna hõlpsalt ise läbi teha. Õpiku sisupeatükid annavad veidi üldisemad soovitused korpuste kasutamise ja loomise kohta, siit leiate korpuslingvistika põhimõtted ja põhimõisted, ülevaate eesti korpustest, märgendusest, korpuste kasutamisest ja tulemuste statistilise analüüsi võimalustest ning oma korpuse loomisest. Näidisuurimuste abil püüame pakkuda võimalikult mitmekesist pilti korpusandmete kasutusvõimalustest keeleteaduse erinevate küsimuste lahendamiseks konkreetsete uurimisülesannete kaudu. Näidisuurimustele on üldjuhul lisatud ka andmed ja kood (kui kasutatud on mõnd programmeerimiskeelt), nii et neid on võimalik ise läbi teha. Õpikusse on panustanud paljud korpuste ja korpuslingvistikaga tegelevad teadlased ning see on suuresti kollektiivne töö. Nii õpiku sisupeatükkide kui näidisuurimuste juures on ära toodud autorid, kes vastavasse peatükki on panustanud.

Õpiku koostamise käigus oleme paratamatult pidanud tegelema ka terminitööga. Ehkki korpuslingvistika on juba suhteliselt pika ajalooga, on eestikeelne terminikasutus sageli ikka veel ebaühtlane ja inglise keele mõjuline. Lisaks peab korpuslingvistika terminoloogia sobima kokku üldise keeleteaduse terminoloogiaga ja ka korpuslingvistikaga tihedalt seotud naabervaldkondadega, andmeteaduse ja statistikaga. Oleme sada olulisemat korpuslingvistika terminit koondanud korpuslingvistika terminibaasi, mis on lisatud Eesti Keele Instituudi EKILexi terminibaasi ja mida saab kasutada Sõnaveebi kaudu. Sellele lisaks võib leida pikema terminite loendi ka käesoleva õpiku lõpus olevast indeksist.

Õpiku materjalid, andmed ja kood on tehtud vabalt kättesaadavaks OSF-i repositooriumis: https://osf.io/xqzsf/.

Õpik poleks valminud paljude inimeste ja institutsioonide abi ja toetuseta. Täname retsensente Krista Kerget ja Kadri Viderit kasulike märkuste ja nõuannete eest.

Sisukord

I OSA

1. Sissejuhatus lingvistilisse uurimistöösse ja korpuslingvistikasse

Ann Veismann, Kadri Muischnek, Jane Padrik

1.1. Uurimuse koostamine ja hüpoteeside püstitamine keeleteaduses

1.1.1. Induktiivne ja deduktiivne lähenemine

Enne kui hakkame uurima, mida endast kujutavad keelekorpused keeleteadusliku uurimuse andmeallikana, tuletame lühidalt meelde keeleteadusliku uurimistöö koostamise põhialused. Keeleteaduslikud uurimused võivad väga üldiselt öeldes olla kaht tüüpi. Esmalt on võimalik mingit nähtust keeles võimalikult täpselt kirjeldada, näiteks vastata küsimusele, kuidas väljendatakse eesti keeles tulevikuaega, millised on kohakäänete funktsioonid või millised on umbisikulise tegumoe kasutusviisid. Teise võimalusena saab kontrollida keele kohta tehtud oletuse või väite paikapidavust, mis seab nähtusi vastavusse, kontrollib seoseid ja seaduspärasid. Näiteks võime uurida, kas umbisikulise tegumoe kasutussagedus sõltub tekstitüübist (akadeemiline tekst *vs*. ilukirjandus *vs.* ajakirjandus). Teaduses üldiselt tehakse vahet induktiivsel ja deduktiivsel uurimusel. **Induktiivse** käsitluse puhul analüüsitakse kõigepealt vaatlusandmeid, siis interpreteeritakse neid ja kasutatakse uute hüpoteeside, teooriate ja printsiipide loomiseks ja esitamiseks, mida seejärel saab kas kinnitada või ümber lükata edasiste uurimustega. See tähendab, et liigutakse üksiknähtuste vaatlemisest ja uurimisest üldistuse poole. **Deduktiivses** uurimuses on esikohal teooriad, printsiibid, hüpoteesid, mis kõigepealt formuleeritakse või omaks võetakse. Seejärel üritatakse neid verifitseerida või falsifitseerida (kinnitada või ümber lükata) vastavate andmete toel. Liigutakse üldiselt üksikule – kõigepealt tehakse oletus üldistuse kohta ja seejärel kontrollitakse selle paikapidavust andmeid uurides. Kuigi (reaalteaduste mõjul) on levinud hoiak, et korralik teadustöö põhineb eelkõige deduktsioonil (teooriapõhiste hüpoteeside kontrollimisel) – vaatlusandmetest induktsiooni teel järelduste tegemist on võrreldud juhusliku õngitsemisega –, on tänapäevane korpuslingvistika andnud võimaluse ka teaduslikult pädevaks induktiivseks uurimistööks.

Nii induktiivse kui ka deduktiivse keeleteadusliku uuringu eelduseks on esinduslik valim tegelikke keeleandmeid, ning just seda saavad korpused pakkuda.

Kui seni on arvatud, et induktiivse uuringu puudus on kvalitatiivsed andmed, st üldistusi tehakse vähestest (juhuslikest) andmetest, mistõttu (reaal)teaduses eelistatakse deduktsiooni, siis suurte andmehulkade olemasolu on teinud võimalikuks teaduslikult veenva induktiivse uurimise. Sellises uurimisviisis, mida täpsemalt võib nimetada eksploratiivseks, selgitatakse esmalt välja keeleandmetes ilmnevad üldised tõenäolised mustrid ja seaduspärasused statistika abil. Seejärel saab nende andmete alusel püstitada täpsemaid hüpoteese, mida kas katsete või korpusuuringute andmete põhjal tehtud statistiliste mudelite abil kontrollima hakata.

1.1.2. Uurimisküsimuse esitamine, hüpoteesi püstitamine

Nii induktiivse kui ka deduktiivse uuringu aluseks on korralikult püstitatud **uurimisküsimus**. Korrektseks uuringuks tuleb esitada küsimus, mille kontrollimine on realistlik, see tähendab, et küsimus peab olema piisavalt konkreetne ja täpne. Konkreetse ja täpse küsimuse esitamine nõuab tihti põhjalikku mõtlemistööd.

Veidi lihtsam on esitada induktiivse (eksploratiivse) uurimuse küsimust, kuna see on kirjeldav, *kuidas*-küsimus. Kuid ka induktiivse uurimuse küsimus ei tohi olla liiga üldine, laialivalguv. *Kuidas*-küsimus peaks aitama siduda mingeid keelenähtusi süsteemi, mustrisse. See peaks olema täpsemalt kirjeldatav vormeli abil: Millised suhted valitsevad x ja y nähtuse vahel andmetes? Näiteks saame uurida keelendite esinemissagedusi kas tekstitüübiti või murdepiirkonniti, uurida eesti keele käänete funktsioone ja nende kasutussagedusi või uurida, milliste oleviku ajavormis esinevate verbidega eesti keeles tuleviku aega väljendatakse. Paljud sellised küsimused nõuavad korpusandmete kvalitatiivset käsitsi märgendamist, kuid võimalik on ka masinmärgendatud korpuse andmete eksploratiivne uurimine (näiteks saab suhteliselt lihtsalt teha sõnasageduste loendeid ja võtmesõnaanalüüse, vt lähemalt õpiku ptk 5 „Levinumad korpuslingvistika meetodid"). Eksploratiivse uurimuse tulemuste kohta saab teha kirjeldavat statistikat ja lihtsamaid statistilisi teste sageduserinevuste olulisuse kontrollimiseks (vt ptk 6 „Korpusandmete statistiline analüüs").

Eksploratiivse uurimuse korral ei saa püstitada hüpoteesi, sest me alles hakkame uurima, millised mustrid ja vahekorrad keelendite vahel meie uuritavas korpuses valitsevad. Alles seejärel saame uurida teist, kuid samaväärset andmekogu, püsitades hüpoteese seoste täpsema olemuse kohta, mida keerulisemate statistiliste mudelitega kontrollida.

Deduktiivse uurimuse korral püstitatakse kõigepealt teoreetilise kirjanduse ja varasemate uurimuste põhjal hüpotees selle kohta, millised seosed võiksid andmetes valitseda. Hüpoteesi püstitamiseks peab uurimisküsimuse vastus olema kontrollitav. Sellisele küsimusele vastamiseks tuleb küsimus kõigepealt sõnastada väitena ehk hüpoteesina. **Hüpotees** väidab midagi asjade seisu kohta maailmas (keeleteaduse puhul keeles). Kas esitatud küsimus on piisavalt täpne ja konkreetne,

selgub siis, kui hakkame hüpoteesi operatsionaliseerima ehk kontrollitavaks (nt mõõdetavaks) muutma.

Operatsionaliseerimine tähendabki seda, et me leiame iga hüpoteesis sisalduva mõiste jaoks mingi kontrollitava parameetri (näiteks esinemissagedus, sõnade pikkus silpides, kuid kontrollitav ja mõõdetav on ka millegi olemasolu või puudumine, st 0/1 tunnus, nagu näiteks 1. isiku asesõna kasutus koos verbiga (1) või väljajätt (0): *ma kõnnin* või *kõnnin*). Keeleteaduslikus uurimistöös on kontrollitavate tunnuste ja seoste leidmine tihti üsna keeruline. Uurimisküsimusele adekvaatselt vastamiseks on vaja rangelt, selgelt ja loogiliselt põhjendada, miks just selle nähtuse kontrollimine vastab esitatud küsimusele. Hüpoteesi püstitades on üks keele struktuuri või kasutuse aspekt see, mille esinemise iseärasuste kohta (nt sageduse, varieeruvuse kohta) me midagi teada tahame ja teine (või teised) aspekt(id) see (või need), mis võiksid esimese esinemist mõjutada. Viimased võivad olla nii keelesisesed (struktuurilised, nt aeg, kõneviis, sõnajärg, kääne, fraasi pikkus, või tähenduslikud, nt elusus, konkreetsus) kui ka keelevälised (tekstiliik, teksti autori vanus või sugu jm). See tähendab, et kontrollimine võib olla nii arvuline (mõõtmine) kui ka mittearvuline (loendamine). Arvulistest ja mittearvulistest tunnustest saab täpsemalt lugeda ptk-st 6 „Korpusandmete statistiline analüüs". Kuidas ja milliseid keeleandmeid mõõta või loendada, sõltub väga palju uurimisküsimusest. Näiteks võib tuua L. Lindströmi ja M.-L. Pilviku näidisuurimuse murrete ja keele varieerumise uurimisest, kus tahetakse teada, mis mõjutab ainsuse 1. isiku pronoomeni olemasolu või puudumist lauses (*lähen* või *ma lähen*). Selleks mõõdetakse esmalt murrete kaupa, kui sageli 1. isiku pronoomenit koos verbiga esineb (st kas murre mõjutab) ja seejärel, kas muud tegurid murde kõrval võivad ka mõjutada (nt verbi ajavorm või verbi pöördelõpu esinemine).

1.1.3. Objektiivsus, valiidsus ja reliaablus

Hea teadustöö tunnusteks on objektiivsus, valiidsus (ehk kehtivus) ja reliaablus (ehk usaldusväärsus). **Objektiivsus** tähendab seda, et tulemused ei tohi sõltuda uurija isikust ega uurimisvahenditest. See tähendab, et tuleb teha kõik selleks, et uurimistulemused poleks kallutatud soovitud tulemuse (hüpoteesi kinnitamise) poole, näiteks andmeid vastavalt valides. **Valiidsus** (ehk kehtivus) tähendab seda, et mõõtmisega saime vastuse just sellele, mida tahtsime teada, meie mõõtmismeetod peab paika, on adekvaatne sellele küsimusele vastamiseks. Tulemused käivad tõepoolest just selle kohta, mida uuriti. Valiidsus sõltub palju sellest, kui hästi ja selgelt on uuritav (keele)nähtus defineeritud, kui selgelt ja täpselt on sõnastatud eesmärk ja konkreetne hüpotees, mida on võimalik operatsionaliseerida ehk mõõdetavaks muuta. Näiteks kui meie andmeallikaks on ajakirjanduskorpus, peab ka uurimisküsimus olema esitatud ajakirjanduskeele, mitte kogu eesti keele kohta, laialt eesti keele kohta tehtud järeldus poleks valiidne.

Reliaablus (ehk usaldusväärsus, tõepärasus) näitab seda, kui tõenäoliselt saadaks uuringu kordamisel sama tulemus, see tähendab, et kasutatud meetod on stabiilne ja annab järjekindlalt sama tulemuse. Tuleb meeles pidada, et ideaalset valiidsust ja reliaablust ei ole võimalik saavutada ja uurimuse reliaablus on tugevalt seotud ka valimi (st keeleandmete) varieeruvusega, mis on keeleteaduses vägagi tavaline. Reliaablust saab kontrollida asjakohaseid andmeanalüüsi meetodeid kasutades. Meeles tuleb pidada, et esmajärjekorras tuleks kindel olla, et uuring on valiidne (vastab just sellele küsimusele, millele vastust tahtsime), alles seejärel on mõtet reliaablust hinnata.

Mõõtmise juures tuleb arvestada, milliseid andmeid kogume. Andmetüüpidest on lähemalt juttu 6. peatükis. Keeleandmeid, mis saadakse vaatluse või mõõtmise tulemusel, nimetatakse **tunnusteks** ehk **muutujateks**, vahel ka inglise keele eeskujul variaabliteks (ingl *variable*). Tunnused võivad olla nii arvulised kui ka mittearvulised. Tunnuse väärtus sõltub sellest, mida mõõdetakse, mõõtmisel antakse igale mõõdetavale objektile mõõdetava tunnuse väärtus.

Hüpoteeside püstitamisel tuleb teha vahet **sõltuval muutujal** (ehk uuritaval tunnusel) ja **sõltumatul muutujal** (ehk seletaval tunnusel). Sõltuv muutuja on see, mille varieerumine meid huvitab. Sõltumatud muutujad on need, mille mõju sõltuvale muutujale meid huvitab (nt kas 1. isiku pronoomen esineb või mitte), mille seost sõltuva muutujaga (või mõju sellele) me uuringus kontrollima hakkame (nt murre, verbi ajavorm). Näiteks võib sõltuvaks tunnuseks olla mingi varieeruv nähtus (nt *gi*-liite positsioon käändelõpu suhtes asesõnades: *kellegile* või *kellelegi*) ja seletavate tunnustena võime vaadelda käänet, tekstitüüpi, kõneliiki (jaatus-eitus), pronoomeni asukohta lauses jm, mille puhul oletame, et need võivad mõjutada *gi*-liite paiknemist, vt nt (Kängsepp 2024). Muutujate mõõtmisel on oluline, kas mõõdame **arvulisi** või **mittearvulisi** tunnuseid. Keeleandmete puhul tuleb enamasti tegemist teha mittearvuliste tunnustega (v.a foneetikas) ja sellega tuleb arvestada andmeanalüüsi meetodeid valides.

1.2. Mis on keeleandmed ja kust neid saada?

Enne kui süveneme keelekorpuste teemasse, tuleks mõelda, mis on üldse keeleandmed ja kust keeleuurija oma andmed saab.

1.2.1. Keeleandmete populatsioon ja valim

Võib öelda, et iga inimene, kes mingit keelt räägib, on keeleandmete allikas. Nn introspektiivset keeleteaduslikku uurimust tehes kasutabki keeleteadlane andmeallikana iseennast. See on võimalik, sest iga inimese teadmised keele kohta esindavad keelekogukonna ühiseid teadmisi (muidu poleks omavaheline arusaamine võimalik). Siiski on ühelt inimeselt kättesaadavad keeleandmed piiratud

ja subjektiivsed. Pole mõeldav, et üks inimene tunneks ja esindaks kõiki keele kasutusviise, ka ei ole kindla sihiga mõttetöö tulemusel keeleuurijalt kätte saadud andmed objektiivsed, sest teadvustamata keeleloome võib oluliselt erineda sellest, mida inimene arvab, kuidas ta keelt kasutab. Niisiis, kui tahame teha üldiselt kehtivaid järeldusi keele kohta, on keelt uurima asudes oluline, et andmed oleks kogutud võimalikult paljudelt keele kõnelejatelt.

Sotsiaalteadustes on olulised mõisted **populatsioon** ehk kõik teatud tunnusele vastavad uuritavad (nt Eesti elanikkond) ja **valim** ehk mingi väljavõte kogu populatsioonist. Väga oluline on, et kui me uuringuga tahame teha järeldusi populatsiooni kohta, siis valim, mida uurime, oleks esinduslik. Kui uurida kõnelejatevahelisi erinevusi (sotsiolingvistiliselt), siis ongi oluline, et kaasatud oleks populatsioonile esinduslik valim uuritava keele kõnelejaskonnast. Kuid kui uurida mingit nähtuse esinemist või varieerumist keeles (näiteks *-nud/-nd* varieerumist), siis moodustab nn populatsiooni kogu võimalik selle keele produktsioon. Keeleandmete eripära on see, et kogu populatsiooni (kogu keelt) ei ole võimalik kunagi kindlaks teha. Keel võimaldab lõpmatul hulgal kombinatsioone, keeleandmete populatsioon ei ole kindlapiiriliselt määratletav. Seepärast on oluline, et kui me väidame midagi üldiselt keele kohta, siis peavad andmed olema kogutud võimalikult erinevate parameetritega kõnelejatelt võimalikult erinevates olukordades. Teisiti öeldes, valim, mida uurida, et teha järeldusi kogu populatsiooni kohta, peab olema võimalikult **esinduslik** (võimalikult mitmekesine, võimalikult suure varieeruvusega). Samas, nagu öeldud, kuna kogu keelt ei ole võimalik kunagi kindlaks teha, siis pole ka täielik esinduslikkus keele puhul praktikas saavutatav. Seda enam tuleks kas püüelda võimalikult suure esinduslikkuse poole või olla teadlik, milliselt piiratud keeleandmete kohta uurimuse järeldused käivad. Korpuse esinduslikkusest ja tasakaalustatusest vaata pikemalt alapeatükist 1.3.1 ja ptk-st 4 „Oma korpuse loomine".

1.2.2. Keeleandmete kogumine

Keeleandmeid on võimalik koguda mitmel viisil: saab teha küsitlusi, koguda kirja pandud või suuliselt esitatud tekste, panna kõnelejaid katseolukordades keelt kasutama või hindama mingite keelendite loomulikkust või vastuvõetavust (katselisest semantikast vt (Klavan, Veismann & Jürine 2013; Jürine, Klavan & Veismann 2013). Gilquin ja Gries (2009) on näidanud, kuidas keeleandmed saab saamisviisi järgi panna loomulikkuse skaalale. Kõige vähem loomulikud keeleandmed saame katseid tehes. Eriti siis, kui katsealune peab keelega tegema midagi sellist, mida ta tavaliselt ei tee või vähemalt tavaliselt ei teadvusta seda endale sel viisil, nagu katses nõutakse (hindama, kategoriseerima vms). Kuigi katselisel teel saadud andmed on kõige vähem loomulikud, on need siiski väga vajalikud keeleandmed. Sellised katsed pakuvad väärtuslikku infot kõnelejate valikute ja hinnangute kohta keelekasutuses. Veidi loomulikumad keeleandmed saame siis, kui anname kõnelejale mingi

konkreetse ülesande kuidagi keelt kasutada, näiteks lauseid moodustada kas katse või intervjuu olukorras (näiteks võib paluda kirjeldada pilti või jutustada teatud teemal). Veel loomulikumaid andmeid saame, kui kogume keelt lindistades või tehes üleskirjutisi loomulikus keelekeskkonnas. Sellisel viisil andmete kogumine on vaevarikas ja piiratud mahuga. Tagatud on küll loomulikkus, kuid mitte esinduslikkus. Kõige loomulikumad ja esinduslikumad keeleandmed saame siis, kui kogume süstemaatiliselt kokku keelt, mida inimesed on loonud erineval otstarbel ja ilma teadmata, et keegi seda uurima hakkab – selleks saame kasutada korpusi.

Võib öelda, et tänapäevase empiirilise keeleteaduse alus on suurte ja esinduslike keelekorpuste kasutamine uurimistöö materjalina. **Keelekorpuse** all peetakse silmas digitaalset ühtses vormingus tekstikogu; tekstiks võib olla ka suulise keele üleskirjutus ning keelekorpused võivad sisaldada ka heli- ja/või videomaterjale. Suure korpuse kasutamine annab võimaluse kontrollida teooriate toel püstitatud hüpoteese statistiliste mudelite abil, aga ka piisavalt materjali, et välja selgitada keeles esinevaid seaduspärasid ja mustreid. See võimaldab uurimustel minna ka väljapoole keeleteadust, nt ühendada korpuslingvistiline analüüs inimkonna sotsiaal-kultuurilise ajaloo ja ühiskondlike teemadega (vt P. Tinitsa näidisuurimust korpuste kasutamisest digihumanitaarias).

Korpuspõhine lähenemine keelele lubab lisaks võrdlemisi lihtsale sageduse loendamisele uurida sõnavara, grammatikat, konkordantse ehk sõnade või grammatiliste üksuste esinemist kontekstis, kollokatsioone ehk sõnade koosesinemisi, keele varieerumist ja palju muud põnevat. Korpuspõhise lähenemise teeb empiirilise keeleteaduse seisukohast atraktiivseks selle kombineerimine masintöötluslike ja statistiliste meetoditega. Tänu keeletehnoloogia jõulisele arengule on võimalik läbi viia laiapõhjalisi korpuslingvistilisi uurimusi, kasutades selliseid automaatseid keeletöötlusvahendeid nagu näiteks automaatne morfoloogiline ja süntaktiline analüüs (vt ka ptk 3 „Märgendamine"). Keeleteadlase tööd hõlbustab oluliselt nii see, et korpused on juba automaatmärgendatud, kui ka see, et on võimalik luua töövoog, milles automaatne märgendamine on loodud ajutiselt andmete märgenduspõhise filtreerimise jaoks. Erinevate vahendite kvaliteet keele automaatseks töötlemiseks on loomulikult inglise keele puhul väga heal tasemel, kuid ka Eesti keeletehnoloogia areng on olnud muljetavaldav.

1.2.3. Mida ei saa korpusega uurida? Korpuslingvistika piirangud

Korpusandmete kasutamisel tuleb teadvustada ka sellise andmekogumisviisi puudusi. Esiteks on oluline arvestada, et mingi keelenäite puudumine korpusest ei anna alust väita, et seda olemas ei võiks olla või et seda kunagi ei kasutataks. Kuna korpuses puuduvad nn negatiivsed näited, siis millegi mitteolemasolu või võimatuse kohta keeles ei saa me korpuse põhjal midagi öelda. Teiseks tuleb meeles pidada, et kõik korpused on paratamatult puudulikud, nad ei saa esindada mitte kunagi kogu keelt. Korpused on paratamatult piiratud mahuga ning on tõenäoline,

et väga harva esinevaid keelendeid korpusest ei pruugi leida. Korpused esindavad neid lekte ehk allkeeli, mida sinna on kogutud, kõiki allkeeli kokku koguda ei ole võimalik. Oluline on siiski, et suurest andmehulgast (kuigi see on ebatäielik) saab üldistada mustreid ja mudeleid, mida andmete täienedes täpsustada.

Kui korpusest mingi sõna või vormi puudumine ei tähenda veel seda, et seda sõna või vormi ei kasutataks või et kasutamine oleks võimatu, siis korpusest mingi (ebahariliku) keelendivariandi üksikjuhu leidmisel võib olla mitmeid põhjuseid. Keelekasutaja võib olla kavatsuslikult tahtnud väljenduda eriliselt, see võib kanda endas mingit tähendusnüanssi või sõnumit (nt *y* kasutamine *ü* asemel). Kuid korpuses leidub ka tahtmatut varieerumist, näiteks teksti sisse lipsanud trükivigu. Üldiselt tuleb rõhutada, et keeleteadlase intuitsioon ei ole piisav otsustamaks, millised variandid on aktsepteeritavad keeleandmed ja millised mitte (ja tuleks andmetest välja arvata). Korpuspõhine keeleuurimine käsitleb korpuses leiduvaid keeleandmeid kui fakte – neid saab seletada, aga valeks tunnistada fakte ei saa. Korpusest oma andmeid kogudes ei tohiks uurija hakata trükivigu parandama, veel vähem õigekeelsusse sekkuma (nt parandama kokku- ja lahkukirjutamist). Samal ajal tuleb siiski arvestada märgendatud korpuse märgendusvigadega ja vajadusega neid käsitsi korrigeerida (vt ptk 3 „Märgendamine“).

Kui peatüki alguses oli juttu sellest, kuidas saada võimalikult loomulikke keeleandmeid, siis lõpetuseks tuleb puudutada korpusest saadud loomulike keeleandmete põhjal tehtud järelduste piire. Korpusandmed on keele uurimiseks head seetõttu, et need on vaatluslikud andmed: uurija on neutraalne vaatleja, kes teeb tähelepanekuid selle kohta, kuidas asjad on. Enamasti on need andmed sagedus-likud (kui palju milliseid sõnu või vorme esineb). Kahtlemata on sagedusel oluline roll keele õppimises, produtseerimises ja mõistmises (vt nt Divjak 2019). Siiski tuleb meeles pidada, et keel on ka sotsiaalne ja mentaalne nähtus ning lisaks sagedusele mõjutab keele loomist ja töötlust veel suur hulk tegureid, mida korpusandmed ei kajasta. See tähendab näiteks, et ainult korpusandmete põhjal järelduste tegemine selle kohta, kuidas keelt inimajus hoitakse ja töödeldakse, pole korrektne. Keele toimimise kognitiivse realistlikkuse kohta saab korpusandmete põhjal püstitada hüpoteese, mida seejärel katsete abil kontrollida. Seega tuleb olla tähelepanelik, millised keele aspektid korpuses kajastuvad ja mille kohta nende põhjal järeldusi saab teha. Kui me ei taha jääda vaatlusandmetest järelduste tegemise juurde ja püüame vastata küsimusele, kuidas keel inimajus kajastub, on metodoloogiline mitmekesisus väga oluline. Selle teema kohta vt pikemalt nt (Arppe jt 2010; Blumenthal-Dramé 2012; Kortmann 2021).

1.3. Korpused

1.3.1. Mis on korpus?

Keelekorpus on korrastatud digitaalne tekstikogu, milles on ühtlustatud tekstide esitusviis ja tekstide märgistik ning tavaliselt lisatud mingit infot märgenduse kujul. Seega on korpus tekstide kui keele kasutusnäidete kogu, kusjuures tekst võib olla nii romaan, ajaleheartikkel, blogipostitus, tekstisõnum, vestluse transkriptsioon jne. Lisaks tekstile endale sisaldab korpus enamasti ka **metaandmeid** teksti kohta (autor, ilmumiskoht ja -aasta, tekstiliigiline kuuluvus, kas tekst on tõlge või mitte jpm) ning lisainfot **märgenduse** kujul; selle kohta saab lähemalt lugeda ptk-st 3 „Märgendamine".

Kindlat piiri, millest alates muutub tekstikogu korpuseks, ei ole. Korpus võib olla ka üsna pisike tekstikogu, kuid ta peaks olema ühtses vormingus ja dokumenteeritud. Minikorpuse moodustavad näiteks ka mis tahes uurimistöö (sh tudengitöö) jaoks kogutud tekstiandmed ja on oluline silmas pidada, et neile kehtivad samad reeglid nagu suurtele ja avalikele korpustele (vt ptk 4 „Oma korpuse loomine").

Korpuse **suurust** mõõdetakse tavaliselt **tekstisõnades** ehk **sõnedes** (ingl *token*), kusjuures traditsiooniliselt loetakse nendeks ka kirjavahemärke. Korpuse suurus sõltub paljudest asjaoludest, näiteks sellest, kui palju on üldse olemas mingit tüüpi tekste selles keeles: vadjakeelseid tekste on oluliselt vähem kui eestikeelseid, eestikeelseid spordiuudiseid on ilmselt rohkem kui eestikeelseid filosoofilisi tekste, eestikeelseid 17. sajandil kirjutatud tekste jällegi vähem kui 19. sajandil kirjutatuid. 2025. aastal oli suurim eesti keele korpus 3,8 miljardi sõneline eesti keele ühendkorpus 2023 (vt ka ptk 2 „Eesti keele korpused").

Üldise **esinduslikkuse** (ka *representatiivsus*) all mõistetakse korpuslingvistikas seda, et ideaalne korpus peaks sisaldama kogu selle keele sõnavara ning kõiki võimalikke morfoloogilisi, süntaktilisi jne konstruktsioone ning sisaldama neid sama suhtelise sagedusega nagu need keelendid selles keeles esinevad. Lisaks peaks selle ideaalse korpuse tekstiliigiline koostis vastama selle keele tekstiliikidele ning nende sagedusvahekorrale selles keeles. Paraku pole selline ideaalne esinduslikkus enamasti saavutatav, aga esinduslikkuse idee ignoreerimine moonutab korpuse alusel kujunevat ettekujutust keelest. Oluline ongi teadvustada, millist keelekasutust mingi korpus esindab, millise keelekasutusvaldkonna ja ajaperioodi suhtes on ta esinduslik.

Tekstiliigi all mõistetakse teatud sotsiaalsele situatsioonile omast keelekasutusviisi (vt ka Kasik 2007: 34 jj). Pole olemas mingit „eesti keele tekstiliikide" üldaktsepteeritud loendit, vajadusest lähtuvalt võib tekstide maailma jagada laiematesse ja kitsamatesse liikidesse. Tekstiliikidena võib vaadelda nii suulist *vs.* kirjalikku keelekasutust, ilukirjandust, spordiuudiseid, toiduretsepte, printerite kasutusjuhendeid jne.

Tekstiliikide suhtes esindusliku korpuse koostamiseks on vaja kõigepealt otsustada, millist osa keelekasutusest ehk millist allkeelt peab koostatav korpus esindama. Intuitiivne vastus oleks muidugi „korpus peab esindama eesti keele kasutust tervikuna", aga selle saavutamine on praktikas võimatu, nagu juba eespool rõhutasime. Peamine põhjus on see, et enamik keelekasutust on suuline, kuid suulise keele korpuse loomine on palju aeganõudvam, keerulisem ja kallim kui kirjalikest tekstidest korpuse koostamine (vt lähemalt ptk 4 „Oma korpuse loomine"). Samuti on suulise ja kirjaliku keele ühes korpuses esitamine problemaatiline, sest suuline keel on oma olemuselt tihedalt seotud mälu ja keele töötlemisega, toimib lineaarselt (ei saa parandamiseks n-ö tagasi kerida) ning seetõttu kajastab mitte ainult keele tootmise tulemust, vaid ka selle protsessi. Suulise kõne süntaks ja sõnavalik on seetõttu kirjalikust väga erinev. Nii et selles mõttes pole üldkeele korpused rangelt võttes kunagi täielikult esinduslikud.

Kui korpuse koostaja on otsustanud, millist osa keelekasutusest või millist allkeelt koostatav ning esinduslikuna planeeritav korpus esindama hakkab, peab ta liigitama seda keelekasutust esindavad tekstid mingite tunnuste alusel tekstiliikidesse ja määrama kindlaks iga tekstiliigi hulga ja/või mõju (kui palju inimesi seda luges või kuulis) tekstide loomise perioodil ning selle järgi otsustama selle tekstiliigi osakaalu korpuses. Korpust, milles tekstiliikide osakaal vastab nende osakaalule keelekasutuses, nimetatakse **tasakaalus korpuseks**. Tähele tuleb panna, et eesti keele koondkorpuse allosaks olev tasakaalus korpus ei ole siiski samas mõttes tasakaalus, vaid seal on võrdses mahus esitatud ilukirjanduse, ajakirjanduse ja teaduskirjanduse tekstid (vt eesti keele korpustest lähemalt ptk 2 „Eesti keele korpused").

Esindusliku ja tasakaalus tekstikogu abil saame ilma lisaoperatsioone rakendamata võrrelda erinevate tekstiliikide keelekasutust kindlal ajaperioodil. Selline eesti keele korpus on näiteks baaskorpus. Selliselt koostatud korpust nimetatakse **suletud korpuseks** (ka: **staatiline** korpus), sest valmis korpusesse ei saa enam tekste juurde lisada või eemaldada, ilma et kaoks tekstiliikidevaheline tasakaal. Võib arvata, et esinduslikkust on võimalik saavutada vaid osaliselt ning pigem mõne varasema ajaperioodi suhtes, sest tänapäeval on tänu internetile juba ainuüksi kirjalike tekstide allikaid väga keeruline kirjeldada. Samas näiteks eesti vana kirjakeele korpuse varasem, 16.–17. sajandi tekste sisaldav osa on selline, kuhu on kaasatud kõik teadaolevad tekstid, seega on see suletud (rohkem selle perioodi tekste pole olemas) ja tollel perioodil kirjutatud eestikeelsete tekstide suhtes täiuslikult esinduslik korpus.

Suletud ehk staatilise korpuse vastand on **avatud** ehk dünaamiline korpus, kuhu lisatakse pidevalt uusi tekste kajastamaks keelekasutuse muutumist. Kuna sellise korpuse eesmärgiks ongi muutuste jälgimine, nimetatakse seda ka **monitorkorpuseks**. Monitorkorpused on näiteks 2. peatükis kirjeldatud RSS-i ja JSI uudisvoo korpused.

On räägitud ka korpuse **küllastatusest** (ingl *saturation*), vt nt (McEnery, Xiao & Tono 2005). Korpus on mingi keelendi suhtes küllastatud siis, kui uute tekstide lisamine korpusesse ei anna enam uut infot selle keelendi kohta. Aga muidugi võib sama korpus olla mitteküllastatud mõne teise, harvaesinevama keelendi suhtes.

1.3.2. Korpuste loomine muutuvas maailmas

Umbes viimased 20 aastat on korpusi suures osas korjatud veebist, sest sellisel viisil saab kiiresti ja odavalt koostada suure korpuse, kuid sellise korpuse tekstiliigilist koostist on raske kindlaks määrata ja see kipub olema ühekülgne, st korpus ei ole esinduslik kogu kirjaliku keelekasutuse suhtes (nt Barth & Schnell 2021: 27). Olulised probleemid **veebikorpuste** loomisel on ka masintõlkelised ja automaatselt genereeritud tekstid ning veebispämm, mille sattumist korpusesse tuleb vältida, sest need ei esinda inimese loomulikku keelekasutust, mida keeleteadlane uurida tahab. Lisaks on viimasel ajal veebi kasutamine keelematerjali allikana muutunud üha keerulisemaks: veebi „tõsisem" sisu muutub tasuliseks. On oht, et kui asutused kolivad oma kodulehed üleni sotsiaalmeediasse ning uudisteportaalid muudavad kõik postitused tasuliseks, siis ei ole veebist tekstide korjamisel enam mõtet ning tuleb tagasi pöörduda traditsiooniliste korpuse loomise meetodite juurde (Jakubíček jt 2020).

Lisaks on tekstirobotite ja suurte keelemudelite tulekuga tekkinud veelgi laiem oht, et korpustesse satuvad tekstid, mis ei esinda inimese loomulikku keelekasutust. Põhimõtteliselt võib tekkida olukord, kus me masinaga töötleme ja uurime teksti, mis on masina poolt kas täielikult kokku pandud või masina ja inimese koostöös tekkinud. Aga võib-olla see ongi tuleviku ühiskonna loomulik keelekasutus, mille evolutsiooni korpuslingvistika abiga uurida.

Samuti ei saa veebist korjata nn spontaanset kirjalikku teksti, sest inimesed peavad oma kirjalikke vestlusi sotsiaalmeediaplatvormidel, kust tekstide kogumine ilma autori nõusolekuta on ebaseaduslik. Lähemalt saab selle teema kohta lugeda artiklitest (Koppel & Kallas 2022; Suchomel & Kraus 2021).

1.3.3. Kas veebiotsing sobib korpuspäringu asendajaks?

Aeg-ajalt kasutatakse lihtsuse mõttes keelendite otsimiseks veebiotsingut, mis sobib küll esialgse info saamiseks, kuid mitte enamaks. Näiteks saab vastuse küsimusele „Kas tõesti kasutatakse sõnast *auto* mitmuse osastava vormina ka *autosi*?", kuid veebiotsinguga on võimatu vastata küsimusele „Kui suur osa sõna *auto* mitmuse osastava vormidest on kujul *autosi*?" või „Millistes kontekstides kaldutakse kasutama vormi *autosi*?".

Nimelt teostavad otsimootorid nagu Google Search või Microsoft Bing väga hästi info-otsingut, kuid sagedusandmete allikana ei ole nad usaldusväärsed. Esiteks loevad otsimootorid tulemuseks veebilehte, millel otsitav sõna esineb,

kuid ei ütle midagi selle kohta, mitu korda see sõna sellel lehel esineb, ja see pole ka nende ülesanne. Teiseks veebi otsimootorid mitte ei loenda, vaid ennustavad otsitava sagedust ja esitavad umbkaudse tulemuse. Kolmandaks ei sobi lihtne veebiotsing korpuse asemel seepärast, et teaduses on olulisel kohal uuringu korratavuse idee – kui üks uurija saab korpuse põhjal mingid arvandmed, siis peaks teine uurija sama korpuse põhjal sama uuritava nähtuse kohta saama sama tulemuse (tulemuse interpretatsioon võib muidugi olla erinev). Veebi sisu on aga pidevas muutumises ja veebiotsingud ei pruugi erineval ajal anda samu vastuseid. Lähemalt on veebi korpusena kasutamise probleemidest kirjutanud Andrew Kehoe (2020).

Suured generatiivsed keelemudelid (ingl *large language model,* LLM) on võimelised looma ehk genereerima inimese „toodetud" keelest esmapilgul eristamatut teksti, seda ka eesti keeles. Need keelemudelid on loodud ehk treenitud väga suuri tekstikorpusi kasutades. Siit kerkibki küsimus, kas kasutaja, keeleandmete otsija võiks pöörduda mõne sellisel keelemudelil põhineva juturoboti (ChatGPT, Copilot jt) poole ja paluda tal genereerida endale vajalikud keelenäited. Ent juturoboti väljundina saadud keeleandmeid ei saa siiski, vähemalt praegu, soovitada kasutada inimkeele asendajana. Põhjusi on mitu. Esiteks pole need autentsed andmed, selliselt saadud andmeid võib võrrelda andmete fabritseerimisega, mis ei ole teaduseetiliselt lubatav.

Teiseks, kuigi generatiivne tehisaru loob keelt väga inimkeele sarnasena, ei ole see siiski sama, nt ühes hiljutises uuringus (Schepens, Marx & Gagl 2023) selgus, et selliselt genereeritud teksti sõnavara on väiksem, kasutatatud on pigem keeles sagedamini esinevaid sõnu. Kuigi keeleandmed on genereeritud ennustuslikult inimeste loodud tekstide põhjal, võib tegemist olla lausetega, mis on grammatiliselt korrektsed, kuid mida vaevalt et keegi tegelikult kasutaks või mis pole konteksti või registrisse sobivad (Crosthwaite & Baisa 2023).

Lisaks on juturoboteid nn järeltreeningu käigus õpetatud vältima teatud tüüpi (solvavaid, vihaõhutavaid jne) vastuseid. Kas selle järeltreeningu käigus õppis juturobot vältima ka teatud tüüpi keelendeid, seda me ei tea. Välja on toodud ka seda, et erinevalt korpusest, mis sisaldab andmeid lausete päritolu kohta, ei ole tehisintellekti genereeritud lausete andmeallikad teada (sisendi täpne sisu) ning lisaks ei ole tehisintellekti loodu korratav (Crosthwaite & Baisa 2023). Korpuspäring on korratav ja annab korrates sama tulemuse.

Samas on see teema – suurte keelemudelite ja nendel põhinevate juturobotite kasutamine korpuslingvistikas – põnev uurimisteema, mille kohta pidevalt uusi teaduspublikatsioone ilmub. Kindlasti võib keelemudelitel põhinevatest juturobotitest abi olla andmete märgendamisel. Ka võib loota, et suured keelemudelid leiavad lingvistikas rakendust kui tööriistakomplektid, näiteks võivad nad aidata toota katseks vajalikke keelelisi stiimuleid (arvestades etteantud parameetritega, nt sõna sagedus, pikkus jm). Siiski tuleb meeles pidada, et lõplik vastutus usaldusväärse ja eetilise teaduse eest on inimesel.

1.4. Korpuste ja korpuslingvistika osa keeleteaduses

Uurimistööd alustades tasub teada, et korpuslingvistiline lähenemine ei haaku ühevõrra hästi kõigi keeleteooriatega. Erinevad keeleteadlased erinevatest keeleteaduse harudest saavad radikaalselt erinevalt aru keele olemusest ega pea tegelike mitmekesiste keeleandmete suuremahulist analüüsimist ühevõrra oluliseks. Õpikus eeldame sellistele teoreetilistele lähenemistele toetumist, mis haakuvad korpuste kasutamise põhimõtetega. Kasutuspõhise keeleteaduse üheks fundamentaalseks osaks on uurida keelt nii, nagu seda päriselus kasutatakse – just seda korpuste kasutamine võimaldab. Oluline on rõhutada, et tänu korpuslingvistikale on võimalik uurida keele aspekte, mida enne ei olnud võimalik sellisel kujul uurida, samuti on korpuste kasutamine keeleteaduses viinud nii uute keeleteooriate välja töötamiseni kui ka seniste teooriate lähtekohtade ümbervaatamiseni. Kasutuspõhisus ja korpuste kasutamine on saanud tavaliseks näiteks keele varieerumise uurimustes, kognitiivses semantikas, sotsiolingvistikas, psühholingvistikas, ohustatud keelte uurimises, digihumanitaarias jm. Omaette suureks osaks korpuslingvistilisest tööst, mis uurib keele varieerumist, on pühendatud registrite vaheliste erinevuste uurimisele (Biber 1993). Korpuslingvistilised meetodid võimaldavad uurida keelekasutust ja varieeruvust ka ajaloolisest perspektiivist. Sellise lähenemise puhul kaasnevad muidugi spetsiifilised probleemid, nagu näiteks käsikirjalise materjaliga töötamine ja selle viimine sellisele kujule, et seda oleks võimalik korpuslingvistiliste vahenditega kasutada. Võib öelda, et korpuslingvistika kui suurte tekstiandmete töötlemine on ületanud keeleteaduse piirid ja kasutusel väga mitmel pool mujal, kultuurievolutsiooni, ajaloo või tehnilise arengu uurimisel. Tekstiandmetega töötamine on seega oluline nii humanitaar- kui ka sotsiaalteadustes, aga ka laiemalt, sest andmestunud ühiskonnas on tekstiandmete kättesaadavus ja hulk järjest kasvav ning nende kasutamise metodoloogilised võimalused mitmekesised; korpuslingvistika meetodid on osa neist (vt ka Masso, Tiidenberg & Siibak 2020).

1.4.1. Korpuslingvistika mitmekesisus

Oluline on meeles pidada, et korpuslingvistilised meetodid ei moodusta ka keeleteaduses monoliitset ja ühtset kogumit, mida on lihtne defineerida. Meetodite hulk, mida korpuslingvistikas kasutatakse, on suur ja lai ning seda mitmekesisust illustreerib väga kenasti siinses õpikus kirjeldatud näidisuurimuste valik. Silmas tuleks pidada, et õpikus kajastatud meetodid on vaid üks osa meetoditest, mis tänapäeval korpuslingvistikas kasutusel on. Oleme siia valinud meetodid, millega me ise igapäevaselt kokku puutume ja mis on rahvusvahelises keeleteaduses laialdaselt kasutusel. Erinevaid meetodeid ühendab arusaam, et korpuslingvistilist uurimistööd tehes puutub uurija kokku masinloetavate tekstidega, mis on hinnatud sobivaks uurija poolt püstitatud uurimisküsimusele vastuse leidmiseks. Tekstiandmete maht, millega töötatakse, võib olla väga erinev, kuid üldiselt räägime

suurest hulgast tekstimahust, mida lihtsalt ei ole võimalik uurijal käsitsi läbi töötada – sellest ka vajadus masinloetavate tekstide järele.

Keelekorpused leiavad rakendust ka arvutilingvistikas ja keeletehnoloogias. Suured keelemudelid vajavad õppimiseks ehk treenimiseks äärmiselt suuri korpusi. Nii on levinud praktika, et keelemudeli treenimiseks kogutakse kokku nii palju andmeid kui vähegi võimalik. Samas vajab arvutilingvistika ka märgendatud erikorpusi, mis aitavad treenida vastavat ülesannet lahendavaid mudeleid või peenhäälestada suuri keelemudeleid – märgendatud nimeolemitega korpusi, süntaktiliselt märgendatud korpusi jne. Viimasel ajal on oluliseks saanud testandmestike koostamine, mille eesmärgiks on välja selgitada, milliseid ülesandeid ja millise kvaliteediga on keelemudelid võimelised ilma peenhäälestamiseta lahendama. Ka neid testandmestikke (küsimuste ja vastuste paare, tekste ja nende kokkuvõtteid, sõnade semantilise klassikuuluvuse suhtes märgendatud tekste, ka morfoloogiliselt või süntaktiliselt märgendatud tekste) saab käsitleda märgendatud erikorpustena.

Suurte keelemudelite ja nende rakendamise kohta keeleteaduslike ülesannete lahendamiseks vaata täpsemalt E. Aedmaa näidisuurimusest.

1.4.2. Korpuslingvistika eelised

Tänapäevast korpuslingvistilist uurimistööd kirjeldavad väga tabavalt kaks märksõna – kiirus ja usaldusväärsus (ingl *rapid and reliable*, McEnery & Hardie 2011: 2). See tähendab, et võrreldes inimesega, kellel kulub lugemiseks ja mõtlemiseks aega ja kes teeb vigu ning on ebajärjekindel, teeb masin seda tööd palju kiiremini ja palju usaldusväärsemalt. Kui arusaam, et masin on inimesest suurte andmemahtude töötlemisel palju kiirem, ei tekita vastuväiteid, siis võib veidi kulmu panna kergitama väide, et masinad on usaldusväärsemad kui inimesed. Jah, masinad teevad vigu, kuid nad teevad neid süstemaatiliselt (vt ka ptk 3 „Märgendamine"). Üks ja sama inimene võib ühel päeval näiteks korpusandmetest leitud sõna *tasemel* märgendada kui nimisõna alalütlevas käändes, aga teisel päeval hoopis otsustada, et tegemist on määrsõnaga. Masin märgendab iga kord sama **andmepunkti** vaid ühtviisi. Usaldusväärsuse all on siin muidugi ka silmas peetud seda, et korpused lubavad meil kiiresti ja usaldusväärselt leida infot sõnade või grammatiliste konstruktsioonide sageduste ja koosesinemiste kohta. Kui me püüaksime sagedust või koosesinemist kätte saada ilma masinat kasutamata, näiteks tugitoolis istudes ja sageduse peale mõeldes, oleksime me väga aeglased ja mitte üldsegi usaldusväärsed uurijad.

Kindlasti tasub meeles pidada, et korpuslingvistikas ei ole esikohal ainult kvantitatiivsed meetodid. Olulist rolli mängib ka andmete kvalitatiivne analüüs. Korpused lubavad meil uurida keelt, nii nagu seda päris inimesed päris elus kasutavad – me näeme, kuidas mingit sõna või grammatilist konstruktsiooni on tekstis kasutatud. Meil on võimalik analüüsida, miks mõned sõnad või grammatilised konstruktsioonid on rohkem sagedased kui teised või miks teatud sõnad esinevad

sagedasti koos või milline on see grammatiline või semantiline kontekst, mis keelelist üksust ümbritseb. Seda kõike peab uurija tegema kvalitatiivselt ja tuginedes oma varasematele keeleteoreetilistele teadmistele. Korpuslingvistika ei anna meile selgitusi, need tulevad meie kvalitatiivsest analüüsist ja teoreetilistest teadmistest.

1.4.3. Kas korpuslingvistika on teooria või meetod või midagi kolmandat?

Küsimusele „Mis on korpuslingvistika?" saab vastata väga mitmeti. Diskussioon selle üle, mis on korpuslingvistika, kas teooria, meetod, paradigma, metodoloogia vm, oli elavam korpuslingvistika algusaegadel, mil korpusandmetele toetuvat keeleteaduslikku uurimust saigi rohkem käsitleda uudse teoreetilise seisukohavõtuna – kasutuspõhise keeleuurimisena.

Küsimusega, mis on korpuslingvistika, haakub eristus **korpuspõhise** (ingl *corpus-based*) ja **korpusest ajendatud** (ingl *corpus-driven*) keeleuurimuse vahel, mis seostub induktiivse ja deduktiivse lähenemisega, millest oli juttu peatüki alguses. Korpuslingvistika kui meetod toetab korpuspõhist uurimust, mille eesmärgiks on leida kinnitust hüpoteesidele või teooriatele või neid ümber lükata või täpsustada. Korpusest ajendatud uurimuse puhul on korpusandmed ise ainuallikaks hüpoteesidele, kuidas keeles asjad on, st korpus justkui hõlmab või kehastab keeleteooriat (vt täpsemalt Tognini-Bonelli 2001: 84–85).

Tänaseks, võib öelda, on lepitud sellega, et korpuslingvistikat saabki mitmeti tõlgendada. Korpuslingvistikat võib pidada mitmesuguste meetodite koguks, mis aitab vastata mitmesuguste teoreetiliste seisukohtadega sobituvatele küsimustele, sh ka keeleteooriate paikapidavust kontrollida. Uurimisküsimusele vastuse saamiseks korpusest keeleandmete kogumine sobitub siiski paremini kognitiiv-funktsionaalse paradigma teooriatega, sest need rõhutavad loomuliku ja mitmekesise, varieeruva keelekasutuse uurimist. Lisaks võib sõnaga *korpuslingvist* tähistada nii korpuste koostamisega tegelevat keeleteadlast kui ka korpust andmete analüüsimiseks kasutavat lingvisti, ja tihti võib see olla ka sama inimene.

Siin õpikus me sellesse diskussiooni, kas korpuslingvistika on teooria või meetod, pikemalt ei lasku. Soovitame neil, keda see teema rohkem huvitab, lugeda McEnery ja Hardie (2011) õpikust peatükki 6 või tutvuda Charlotte Taylori (2008) ülevaatega. Õpikus võtame „korpus-kui-meetod" suuna ja defineerime korpuslingvistikat kui protseduuride või meetodite kogumit, mille abil saab keelt uurida (McEnery & Hardie 2011). Rõhutame, et korpuslingvistilised meetodid, mida me siin lehekülgedel kirjeldame, on vaid üks viis, kuidas uurimisküsimusele vastust leida, ning korpusandmeid on hea kombineerida ka muudel viisidel (nt katseliselt) kogutud andmetega (vt nt Klavan 2024).

1.4.4. Lühidalt korpuslingvistika ajaloost

Korpuslingvistika ajalugu on tihedalt seotud arvutite, andmebaaside ja arvutilingvistika ajalooga – suuremad korpused said võimalikuks tänu arvutimälu mahu suurenemisele, nendele suurtele korpustele sai järjest keerukamaid päringuid esitada tänu andmebaaside ja andmebaasipäringukeelte arengule ja nende automaatne analüüs ning märgendamine muutus võimalikuks tänu arvutilingvistika arengule.

Vaadates ajas tagasi, näeme muidugi, et korpuslingvistilist tööd viidi esialgu läbi paberil. Näitena ühest esimesest korpuslingvistilisest uurimusest võib tuua inglise keele grammatika, mille Charles Fries (1952) pani kokku korpuse põhjal. Enne kui korpuslingvistika ilmus keeleteaduse areenile kui metodoloogia, võis sõna *korpus* kasutada ükskõik millise *ad hoc* kokku pandud keeleteaduslike näidete kogumi kohta.

Esimeseks tõeliseks digitaalseks keelekorpuseks peetakse USA-s Browni ülikoolis koostatud Browni korpust[1], mille koostamispõhimõtted, eriti selle tekstiliigiline jaotus muutusid peagi korpuseloomise standardiks. Korpus sisaldab miljon sõna kirjaliku ameerika inglise keele tekste 15 tekstiliigist, korpuses ei ole terviktekstid, vaid igast alliktekstist on korpusesse võetud 2000-sõnaline katke (seda on nimetatud katkendikorpuseks). Hiljem märgendati Browni korpus ka morfoloogiliselt, mis inglise keele puhul tähendab peamiselt sõnaliigi märgendamist. Mõned aastad hiljem koostati Browni korpuse eeskujul briti inglise keelt sisaldav Lancasteri-Oslo/Bergeni (LOB) korpus[2]. Ka eesti keele 1980ndate aastate korpus ehk baaskorpus on koostatud Browni korpuse eeskujul (vt ptk 2.2.1 „Esimesed eesti keele korpused: esinduslikud, aga väikesed").

Kuigi Browni korpus koostati juba 1960ndate aastate teisel poolel, ilmusid esimesed põhjalikumad korpuspõhised keelekirjeldused alles 1980ndatel: 1982. aastal avaldasid W. Nelson Francis ja Henry Kučera Browni korpuse sõnade ja sõnaliikide sagedusandmed (Francis & Kučera 1982); Stig Johansson ja Knut Hofland (1989) tegid sarnase analüüsi LOB-korpuse kohta. Samuti ilmusid 1980ndate lõpupoole esimesed üksikkeelendi korpuspõhised kirjeldused nagu näiteks Sylviane Grangeri (1987) käsitlus inglise keele passiivist või paljusid korpuses märgendatavaid tunnuseid kasutav Douglas Biberi (1988) keeleliste registrite multidimensionaalne analüüs.

Samal ajal, 1980ndate teisel poolel hakati ka inglise keele suurte seletavate sõnaraamatute nagu „Collins COBUILD English language dictionary" (1987) ja „Longman dictionary of contemporary English" (1987) sõnaartikleid koostama suurte korpuste analüüsi põhjal. See tähendas muu hulgas seda, et korpuste ja

[1] https://varieng.helsinki.fi/CoRD/corpora/BROWN/

[2] https://varieng.helsinki.fi/CoRD/corpora/LOB/

eriti nende kasutajaliideste vastu tekkis kommertshuvi, mis viis Sketch Engine'i korpusanalüüsi tarkvara loomiseni (Kilgarriff jt 2004; Kilgarriff jt 2014).

Järgmine oluline teetähis korpuste koostamise ajaloos oli briti inglise keele rahvuskorpuse (British National Corpus, BNC) koostamine aastatel 1991–1995. BNC on suur (100 miljonit sõna, 10% sellest suuline keelekasutus), esinduslik ja automaatselt morfoloogiliselt märgendatud korpus. Selle ameerika inglise keele vaste on American National Corpus.

Eesti keele jaoks sellist suurt, sajamiljonilist esinduslikku korpust ei ole. 1990ndate aastate teisel poolel ja 2000ndate algul koostati küll 260 miljoni sõnaline **koondkorpus**, kuid selles on ajalehetekstid tugevalt üleesindatud.

Näeme, et arvuti mälumahu ning töökiiruse suurenedes suurenesid ka korpused ning nende kasutusvõimalused muutusid tavalisele keeleteadlasele kättesaadavamaks. Näiteks 1970ndatel võttis inglise keele sagedase sõna *when* **konkordantsiridade** (sõna(vorm) kontekstis) väljasõelumine ühe miljoni sõna suurusest Browni korpusest aega mitu tundi, 1980ndate lõpul juba ainult paar minutit (Kennedy 1998: 7). Varaste korpuste põhiline kasutusmeetod oligi konkordantsiridade leidmine korpusest, pisut hiljem lisandus kollokatsioonide tuvastamine (nende meetodite kohta vt täpsemalt ptk 5 „Levinumad korpuslingvistika meetodid").

Browni ja LOB-i korpuste koostamise peamiseks eesmärgiks oli grammatika uurimine (Hunston 2022: 18), 1980ndatel ja varastel 1990ndatel pöördus tähelepanu leksikaalsete üksuste uurimisele; mõjukas teos oli John Sinclairi 1991. aastal ilmunud „Corpus, concordance, collocation" (Sinclair 1991).

Varaste korpuseloomise projektidega kaasnesid teoreetilised arutelud korpuste koostamispõhimõtete üle, eriti esinduslikkuse definitsiooni ja selle saavutamise võimalike teede üle. Iseloomulikud teaduspublikatsioonide pealkirjad on „Corpus design criteria" (Atkins, Clear & Ostler 1992) või „Representativeness in corpus design" (Biber 1993). Hiljem, kui korpusi hakati koostama internetist automaatselt korjatud materjalist ja korpused läksid väga suureks, muutus korpuse või sellest tehtud väljavõtte esinduslikkus pigem kasutaja mureks.

Korpuslingvistika algusaegadel sisaldasid korpused põhiliselt toimetatud trükitud tekste. Suulise keele korpuse olulisus oli küll teadvustatud, aga suulise keele korpuse koostamine on aeganõudvam ja kallim kui kirjaliku keele korpuse koostamine, mistõttu suulise keele korpused olid paratamatult väiksemad. Esimese digitaalse suulise keele korpuse – Sankoffi ja Cedergreni Montreali prantsuse keele korpuse (Sankoff-Cedergren Corpus of Montréal French) – koostamist alustati 1971. aastal (Sankoff & Sankoff 1973). Sageli nimetatakse esimese suulise keele korpusena ka Londoni-Lundi suulise keele korpust[3] (Svartvik 1990), mille koostamist alustati juba aastal 1959 „Survey of English usage" projekti raames, kuid mille digiteerimiseni jõuti 1975. aastal.

[3] https://varieng.helsinki.fi/CoRD/corpora/LLC/

Digitaalsete eneseväljendus- ja suhtluskeskkondade arenedes sattusid need kõik muidugi ka korpuslingvistika huvivälja. Veebis vabalt kättesaadavate keskkondade (blogid, foorumid jm) tekstid sisalduvad suurtes veebikorpustes, privaatsete suhtluskanalite (nt Facebook või Whatsapp) vestlusi tuleb aga eraldi koguda, osutades sealjuures eraldi tähelepanu isikuandmete kaitsele ja teaduseetikale.

Veebis avaldatud tekstide hulga plahvatuslik kasv ja veebist otsimise võimaluste areng avaldasid mõju ka korpuslingvistika arengule. Tekkis kaks mõnes mõttes võistlevat suunda: korpuse koostamine veebist automaatselt tekste kogudes (Cavaglia & Kilgarriff 2001) või veebiotsingu kasutamine korpuspäringuna, kusjuures korpuseks on kogu veeb (Fletcher 2007). Teine suund paistab olevat vähempopulaarne, põhjuseks selle meetodi täielik sõltuvus mõnest suurest otsimootorist nagu Google, Bing või Yahoo! Esimese suuna abil koostatakse alates 2013. aastast eesti keele veebikorpuste sarja, mis kuulub eesti keele ühendkorpuse hulka (vt ptk 2.2.3 „Eesti keele ühendkorpused: mahukaimad eesti keele digitekstide kogud“).

Lõpetuseks

Õpiku esimeses peatükis vaatasime sissejuhatavalt keeleteadusliku uurimuse aluseid – kuidas küsimusi esitada ja andmeid koguda. Teadusliku uurimuse ülesehitamine, küsimuste esitamine, hüpoteeside püstitamine, teooriate ja meetodite kaalumine, andmete kogumine ja analüüsimine pole lihtne ja vajab õppimist ja harjutamist. Käesolevas õpikus puudutasime seda teemat ainult väga põgusalt ja soovitame sel teemal ise edasi uurida. Lisaks vaatlesime korpust kui keeleandmete kogu ja arutlesime korpuslingvistika olemuse ja ajaloo üle. Vaadates tulevikku, võime olla kindlad, et järjest olulisemaks saavad korpuslingvistikas erinevad tehnilised lahendused, mida pakuvad meile tehisintellekti (sh suured keelemudelid) ja masinõppe võimalused. Üsna kindlalt võime ennustada ka järjest eripalgelisemate korpuste loomist (nt multimodaalsed korpused, sotsiaalmeedia tekstidel põhinevad korpused). Ühtlasi oleme järjest enam andmestuvas maailmas silmitsi suurenenud eetiliste ja andmekaitsest tulenevate väljakutsetega. Korpuslingvistika mängib üha suuremat rolli kogu keeleteaduses, erinevate keeleteooriate postulaatide testimisest praktiliste rakendusteni, nt keeleõpetuses ja -õppes. Loomuliku keele töötlusel põhinevate tööriistade kasutus lihtsustab erinevaid korpuslingvistilisi protsesse, nt märgendamisprotsessi. Tehniliste võimaluste laienemine tähendab omakorda, et keeleteadlastel tuleb ennast järjest rohkem selles vallas harida. Õpiku peatükid pakuvad esimese sissevaate neisse teemadesse ja viitavad kirjandust edasilugemiseks.

Lisalugemiseks

Litosseliti, Lia (toim). 2017. *Research methods in linguistics*, 9–28, 93–118. London: Bloomsbury Academic.

McEnery, Tony & Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*, 1–24. Cambridge: Cambridge University Press.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Science 7), 1–104. Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735821.

2. Eesti keele korpused

Liina Lindström, Kadri Muischnek, Jelena Kallas,
Kristina Koppel, Pärtel Lippus, Maarja-Liisa Pilvik

Eesti keele korpusi hakati looma 1980ndate lõpus ja 1990ndate alguses. Kui esmalt alustati kirjaliku keele tekstidest korpuste koostamisega, siis alates 1990ndate teisest poolest on Eestis arendatud ka mitmeid suulise keele korpusi. Tänaseks on eesti keel korpuste osas hästi kaetud: meil on mahukad kirjakeele korpused, sh vana kirjakeele korpus ja mahukas peamiselt internetist korjatud eestikeelseid tekste sisaldav ühendkorpuste sari, mitmeid suulise keelekasutuse korpusi (sh mahukad ERR-i raadiosaadete ja taskuhäälingute korpused). Ka keeleteadus on korpuste andmeid üha rohkem uurimistööks kasutanud. Korpusi kasutatakse ka praktiliste eesmärkide täitmiseks, näiteks Eesti Keele Instituudi sõnastike loomisel toetutakse ühendkorpuste sarjale (vt K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimust sõnastike koostamise kohta) ning väljaspool keeleteadust, nt tehnoloogia arengu jälgimiseks ajalehekorpuse põhjal (vt P. Tinitsa näidisuurimust korpuste kasutamisest digihumanitaarias). Võib öelda, et eesti keel on kõnelejate arvu arvestades üks paremini korpustega kaetud keeli maailmas.

Teisalt ei tähenda see, et selles osas on kõik valmis: nii, nagu keel muutub pidevalt, on vaja korpusi täiendada uute andmetega. Samuti pöörab keeleteadus tähelepanu keele erinevatele kasutusaladele ja uurimisvõimalustele, mistõttu püsib vajadus uute korpuste järele ka edaspidi. Näiteks eesti viipekeele arendamiseks ja uurimiseks koostatakse praegu viipekeele korpust, eesti keele edukaks õpetamiseks on vaja keeleõppe seaduspärasusi uurida mitmekülgse õppijakeele korpuse põhjal, keele multimodaalse kasutamise uurimiseks, kus lisaks kõnele analüüsitakse žeste, pea-, näo- ja kulmuliigutusi, oleks kasu suurest multimodaalse keele korpusest. Kuna iga korpuse koostamine nõuab arvestataval määral tööaega, inimesi ja muid ressursse, siis on selge, et korpuste koostamine kulgeb vastavalt sellele, mis valdkondades on piisavalt uurimishuvi või arendusvajadusi ning mille jaoks on võimalik ka rahastust saada. Muidugi võib iga keeleuurija enda uurimisülesande jaoks vajadusel luua korpuse ka ise, selle kohta vt õpiku ptk-st 4 „Oma korpuse loomine".

Siin peatükis anname ülevaate sellest, milline on olnud eesti keele korpuste kujunemislugu ning millised eesti keele korpused on praeguseks olemas ja uurimistöö jaoks kasutatavad. Kuna korpustega võivad kaasneda autoriõiguse ja isikuandmete kaitsega seotud küsimused, võivad korpused olla sellest tulenevalt kas valmis

avalikuks kasutamiseks või ligipääsupiirangutega. Ligipääsupiirangud tähendavad enamasti seda, et korpust on võimalik kasutada kindlaks otstarbeks (nt lõputöö kirjutamiseks) ning kasutajaga sõlmitakse leping korpuse kasutamiseks. See piirang puudutab ennekõike suulise keele korpusi, kus isikuandmete kaitse vajadus on ilmne.

2.1. Kust korpusi leida?

Keelekorpusi saab kasutada nii spetsiaalsete korpuste **analüüsikeskkondade** kaudu kui ka neid enda arvutisse laadides. Suurtele eesti keele korpustele saab päringuid esitada Sketch Engine'i[1] ja KORP-i[2] kaudu ning osaliselt ka Keeleveebis[3]. Mitmetel spetsiifilisematel korpustel on kodulehel ka oma otsimootoriga kasutajaliides. Korpuste kasutamise enamlevinud meetoditest on lähemalt juttu ptk-s 5.

Keelekorpusi, nagu ka teisi keeleressursse, koondatakse ja hoiustatakse vastavates **andmehoidlates** ehk repositooriumides, kus leiab infot olemasolevate korpuste kohta ning saab neid ka endale alla laadida. Eesti keele ressursse on koondanud META-SHARE'i Eesti haru[4] ning European Language Grid (ELG)[5], mis koondab kõigi Euroopa Liidu keelte, sh eesti keele korpusi ja muid keeleressursse. ELG repositooriumis on üldse kokku 371 eesti keelt sisaldavat korpust, nii et oma töö jaoks sobivat ressurssi otsides tasub võtta aega selle kataloogi läbivaatamiseks. Sellistel repositooriumidel on kahjuks kalduvus ruttu vananeda: nad võivad küll olla üsna ammendavad oma loomise hetkel, aga pärast seda juurde tekkinud korpuste info ei pruugi sinna enam lisanduda. Seetõttu ei pruugi need sisaldada kõiki eesti keele korpusi. Repositooriumide ajakohase ja kättesaadavana hoidmise ülesanne on tänapäeva maailmas üha olulisem küsimus ja selle jaoks on loodud mitmeid suuri teadustaristu objekte: Euroopa üks suuremaid keeleandmeid koondavaid teadustaristuid on CLARIN (ka selle lehelt leiab eesti keele korpusi)[6]. CLARIN-i partner Eestis on alates 2025. aastast Keeleandmete Teadustaristu, mille ülesanne on üleval hoida eesti keele andmeid ja korpusi.

Alustame ülevaadet üldkeele korpustest, liikudes vanematest ja väiksematest kirjutatud keele korpustest uuemate ja suuremate poole. Seejärel tutvustame keele

---

1 https://www.sketchengine.eu

2 KORP-ist on 2025.a seisuga Eestis kasutusel vähemalt kaks varianti: https://korp.keeleressursid.ee/ ja https://korp.eki.ee/ (vt ka ptk 5.1.2 „Veebipõhised korpusanalüüsi keskkonnad").

3 https://www.keeleveeb.ee

4 https://metashare.ut.ee

5 https://live.european-language-grid.eu

6 https://www.clarin.eu/resource-families

uurimise seisukohast olulisemaid erikorpusi, mis esindavad nii kirjalikku kui suulist keelekasutust. Kokkuvõtliku tabeli eesti keele korpustest leiate selle peatüki lõpust.

2.2. Kirjakeele korpused

2.2.1. Esimesed eesti keele korpused: esinduslikud, aga väikesed

Esimese eesti keele korpusena koostati Tartu Ülikoolis 1990. aastatel klassikalise Browni korpuse[7] eeskujul 1980. aastate kirjaliku keele korpus, mida on hiljem hakatud nimetama ka **baaskorpuseks**[8]. Baaskorpuses on ühe miljoni sõna mahus tekstikatkeid 1980ndatel avaldatud tekstidest, jaotatuna kümne tekstiliigi vahel. See on klassikaline suletud esinduslik korpus, mis peegeldab aastatel 1984–1987 Eestis avaldatud algselt eesti keeles kirjutatud (mitte tõlgitud) kirjalikke tekste. Selle esinduslikkuse saavutamiseks tehti kõigepealt kindlaks, millised raamatud, ajalehed, ajakirjad jm sellel perioodil ilmusid ja kuidas nad jagunesid tekstiliikide vahel. Seejärel leiti iga tekstiliigi osakaal avaldatud tekstide koguhulgast. Korpuse tekstide jaotus vastab sellele osakaalule (tabel 2.1, vt ka Hennoste & Muischnek 2000). Igast väljavalitud tekstist läks korpusesse umbes 2000-sõnaline katkend.

Tabel 2.1. Baaskorpuses esindatud teksitüübid, nende maht ja osakaal

Valdkond	Sõnede arv	Osakaal korpuses
ajakirjandus	175 000	17,5%
dokumendid	12 000	1,2%
entsüklopeedilised teosed	20 000	2,0%
esseed ja biograafiad	90 000	9,0%
hobid ja harrastused	75 000	7,5%
ilukirjandus	250 000	25,0%
populaarteadus	150 000	15,0%
propaganda	60 000	6,0%
religioon	8000	0,8%
teadus	160 000	16,0%

[7] https://www.sketchengine.eu/brown-corpus

[8] https://cl.ut.ee/korpused/baaskorpus/1980

Baaskorpuse edasiarenduseks on 1890.–1990. aastate keelekasutust ja selle muutumisi näitlikustav läbilõikekorpus ehk **niitkorpus**[9], mis sisaldab igast kümnendist ajalehe- ja ilukirjandustekste. Niitkorpuse alamkorpused on esinduslikud valitud perioodi ajalehe- ja ilukirjanduskeele suhtes, ent mitte kogu kirjaliku keelekasutuse suhtes.

Baas- ja niitkorpus on seega suletud esinduslikud katkendikorpused, milles sisalduvatest tekstidest on olemas täielik ülevaade. Baas- ja niitkorpuse tekstid põhinevad trükitekstidel, mis on skannitud, tärktuvastatud ning käsitsi kontrollitud. Kuna tekstid digitaliseeriti käsitsi, oli olemas ka täielik kontroll märgistiku jm vormilise külje üle. Teisalt on see olnud piiranguks korpuse mahu osas.

Baas- ja niitkorpus on automaatselt morfoloogiliselt analüüsitud ja kättesaadavad korpuspäringusüsteemi KORP kaudu. Baas- ja niitkorpuse tekstid leiab ka ELG-st.

2.2.2. Teise põlvkonna eesti keele korpused: pole esinduslikud, aga nende sisu on teada

Alates 1990ndatest on eestikeelsed tekstid sündinud järjest suuremas mahus digitaalsena: ajalehtedel ja ajakirjadel on mahukad digitaalsed arhiivid, aga ka ilukirjandus- ja teadustekstid ilmuvad tänapäeval elektrooniliselt. Seetõttu hakati 1990ndate teisel poolel koguma suuremahulist **koondkorpust**[10], mis sisaldab üle 250 miljoni sõna. See sisaldab peamiselt ajalehetekste, aga ka ilukirjandust, teadustekste, seadusi, uue meedia nähtustena veebifoorumite ja jututubade tekste, Riigikogu stenogramme jm. Ajalehetekstide suure osakaalu taga on asjaolu, et ajalehtedel olid tol perioodil digitaalsed arhiivid ja nende arhiivide teisendamine korpuse formaalsele kujule võimaldas saada mõistliku aja- ja tööjõukuluga mitmekümne miljoni sõna suurusi alamkorpusi. Koondkorpust on võimalik kasutada nii Sketch Engine'i eesti keele ühendkorpuste sarja allosana (Reference Corpus) kui ka KORP-i ja Keeleveeb.ee kaudu.

Lihtsustamaks kirjaliku keelekasutuse kolme keskse tekstiliigi – ilukirjanduse, ajakirjanduse ja teaduskirjanduse – keelekasutuse võrdlemist, moodustati koondkorpuse tekstidest **tasakaalus korpuse**[11] nimeline alamkorpus, milles on nimetatud tekstiliikide tekste igaüht viie miljoni sõna mahus. Tasakaalus korpust saab kasutada Sketch Engine'is eesti keele ühendkorpuste sarja allosana (Balanced Corpus), KORP-i ja Keeleveeb.ee abil. Tuleb siiski silmas pidada, et kui 1. peatükis räägiti tasakaalustatud korpusest, siis see korpus ei ole tasakaalustatud kõigi tekstiliikide mõttes, vaid korpuse nimi osutab sellele, et valitud kolm tekstiliiki on võrdselt esindatud.

[9] https://cl.ut.ee/korpused/baaskorpus

[10] https://www.cl.ut.ee/korpused/segakorpus

[11] https://cl.ut.ee/korpused/grammatikakorpus

2.2.3. Eesti keele ühendkorpused: mahukaimad eesti keele digitekstide kogud

Eestikeelsete digitaalsete veebis avaldatud tekstide plahvatuslik kasv on loonud uued võimalused korpuste arendamiseks. Eesti Keele Instituut koostöös tarkvarafirmaga Lexical Computing Ltd. on loonud eesti keele ühendkorpuste (Estonian National Corpus) sarja, milles 2025. aasta lõpu seisuga on viis versiooni: 2013, 2017, 2019, 2021 ja 2023. Ühendkorpuste sari on mahukaim eestikeelsete digitekstide kogu, mida on uuendatud iga kahe aasta tagant. See on kättesaadav korpusanalüüsi tarkvara Sketch Engine'i kaudu ning allalaetav ELG repositooriumist või META-SHARE'i repositooriumist.

Valdava osa ühendkorpustest moodustavad veebist kogutud tekstid, mis eesti keele ühendkorpuse 2013 mahust annavad 56%, eesti keele ühendkorpuse 2017 mahust 80%, eesti keele ühendkorpuse 2019 mahust 87%, eesti keele ühendkorpuse 2021 mahust 91% ja eesti keele ühendkorpuse 2023 mahust 90%. Seega on ühendkorpus tervikuna oma olemuselt pigem veebikorpus koos selle tüüpiliste probleemidega, mille kohta saab lähemalt lugeda artiklist (Koppel & Kallas 2022).

Ühendkorpuste sari, nagu nimigi ütleb, koondab enda alla kõik kättesaadavad tänapäeva kirjaliku eesti keele kogud. Kasutaja võib päringuid teha kogu korpusest tervikuna või siis valida sellest välja endale sobivad alamkorpused, st uurida saab ka kitsama valdkonna keelekasutusi ning neid omavahel võrrelda. Ühendkorpuste sari sisaldab endas veebikorpusi, koondkorpust ja muid alamosi ehk alamkorpusi, mis on esitatud tabelis 2.2.

Tabel 2.2. Eesti keele ühendkorpuse 2023 alamkorpused (suuruse järjekorras)

Alamkorpus	Sõnede arv	Osakaal korpuses
eesti keele veebikorpus 2021 (Web 2021)	884 524 889	23,4%
eesti keele veebikorpus 2017 (Web 2017)	638 470 425	16,9%
eesti keele veebikorpus 2019 (Web 2019)	613 775 513	16,2%
eesti keele veebikorpus 2023 (Web 2023)	571 221 646	15,1%
ajamärgistatud uudisvood 2014–2023 (Timestamped 2014–2023)	344 463 562	9,1%
eesti keele veebikorpus 2013 (Web 2013)	302 950 928	8,0%
TÜ koondkorpus 1990–2008 (Reference Corpus 1990–2008)	212 424 035	5,6%
kaasaegne ilukirjandus 2000–2023 (Literature Contemporary 2000–2023)	148 559 737	3,9%

Alamkorpus	Sõnede arv	Osakaal korpuses
Vikipeedia (Wikipedia)	30 747 204	0,8%
TÜ tasakaalus korpus 1990–2008 (Balanced Corpus 1990–2008)	11 631 801	0,3%
teadustekstid (Academic Texts)	11 026 315	0,3%
Vikipeedia arutelud 2017 (Wikipedia Talk 2017)	7 571 584	0,2%
vanem ilukirjandus 1864–1945 (Literature Old 1864–1945)	7 421 569	0,2%

Ühendkorpuste loomisel poolautomaatseid meetodeid rakendades on tekstide hulgast välja jäetud masintõlkelised tekstid ja veebispämm. Samuti ei sisalda ühendkorpused sotsiaalmeediaplatvormide (nagu Facebook, Twitter / X vms) tekste ning suletud või tasuliste veebilehtede sisu. Tekstide valiku kohta vt täpsemalt (Koppel & Kallas 2022).

Ühendkorpused on morfoloogiliselt märgendatud, kasutades EstNLTK teeke (vt ka ptk 3.3 „Märgendamistööriistad"). Märgendus on automaatselt tehtud ning seda ei ole käsitsi kontrollitud, mistõttu ühendkorpuse kasutamisel peab arvestama võimalike märgendusvigadega.

Ühendkorpus võimaldab uurida eesti keelt väga mitmest perspektiivist – uurida saab nii sõnavara, morfoloogiat, süntaksit, nende varieerumist, semantikat jne. Mõned näited on toodud ka käesoleva õpiku näidisuurimustes (J. Padriku, M.-L. Pilviku, A. Veismanni näidisuurimused). Samuti kasutatakse ühendkorpust sõnaraamatute koostamisel (K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimus).

2.2.4. Erimärgendusega korpused

Erimärgendusega korpused on sellised korpused, mis on saanud mingi spetsiifilise märgenduse, sest nad on loodud mingil kindlal uurimis- või arenduseesmärgil. Sellesse alapeatükki on koondatud erimärgendusega kirjaliku keele korpused. Erimärgendusega korpused on võrreldes tavaliste kirjakeele korpustega tavaliselt mahult väiksemad, sest märgendus on lisatud käsitsi, ning sageli on need tehtud mingi kindla ülesande lahendamiseks. Kuivõrd neid on koostatud eri aegadel erinevate ülesannete lahendamiseks, siis olgu siin mainitud vaid suuremad ja olulisemad, mille kasutamise vastu võiks ka tänapäeval huvi olla.

Morfoloogiliselt käsitsi ühestatud korpus (Muischnek 2011) on umbes 500 000 sõnaline tänapäeva kirjaliku eesti keele korpus, mida saab nii tervikuna

alla laadida[12] kui ka KORP-i kaudu kasutada. Ehkki morfoloogiliselt on märgendatud pea kõik eesti keele korpused, sh suur eesti keele ühendkorpus, on selle korpuse kasutamise eelis märgenduse täpsus.

Eesti keele universaalsõltuvuste (ingl ***universal dependencies***, **UD**) **puudepanga**[13] tekstides on käsitsi märgendatud **morfoloogiline** ja **süntaktiline** info. Korpus on osa mitmekeelsest sõltuvussüntaksil põhinevast süntaktiliselt märgendatud korpuste kogust, millest avaldatakse uus versioon iga kuue kuu tagant, seda uusimat versiooni saab alla laadida UD lehelt ning sellele esitada päringuid mitme kasutajaliidese abil, mille loendi leiab samast.

Samu tekste sisaldab eesti keele piirangute grammatika (ingl *constraint grammar*) puudepank[14] (Muischnek 2015), kuid süntaktiline märgendus on teistsugune, st kasutatud on erinevaid märgendeid ja kohati eristatud ka erinevaid kategooriaid (vt ka ptk 3.2.2 „Süntaktiline märgendamine").

Paralleelkorpus on tekstikogu, milles on sama tekst mitmes keeles ja tüüpiliselt on selles laused paralleelistatud, st märgendus näitab, millised laused on üksteise tõlked. Paralleelkorpusi kasutatakse masintõlke arendamisel, tõlkeuuringutes, kontrastiivses keeleteaduses, keeletüpoloogias, leksikograafias, keeleõppes jne. Näiteks inglise-eesti/eesti-inglise õigusaktide paralleelkorpus[15] sisaldab Eesti seaduseid ja nende tõlkeid inglise keelde ning Euroopa Liidu õigusakte ja nende eestikeelseid tõlkeid.

Paralleelkorpused on enamasti tekstiliigiliselt piiratud. Kui Euroopa Liidu õigusaktide ja määruste ingliskeelsete originaalide eestikeelseid tõlkeid on palju ja need on täiesti vabalt kasutatavad, siis näiteks sloveenikeelsete tekstide eestikeelseid vasteid, mis ei oleks tõlgitud inglise keele vahendusel, on tunduvalt raskem leida. Hulgaliselt paralleelkorpusi leiab paralleelkorpuste kogust OPUS[16], ka ELG repositooriumis on eesti keele osalusega paralleelkorpusi.

OPUS-e kogust leiab näiteks mitmekeelse **subtiitrikorpuse** OpenSubtitles[17], mis sisaldab korpuseks korrastatud kujul filmide subtiitreid 94 keeles, sh eesti keeles. Kuna subtiitrid on seotud filmide või telesaadetega, mida on tõlgitud mitmesse keelde, annab see hea võimaluse neid tõlkeid omavahel võrrelda. Kuna subtiitrid on seotud kindlate kaadritega filmides, loob see võimaluse eri keelte subtiitreid omavahel aegjoondada ja nii hästi võrreldavaks teha. Subtiitrikorpust on keeleteaduses kasutatud üsna palju grammatika võrdlemisel ja keeletüpoloogias: sama kasutuskontekst võimaldab võrrelda keeltevahelisi sarnasusi ja erinevusi grammatiliste vahendite kasutamisel.

---

12 https://cl.ut.ee/korpused/morfkorpus

13 https://universaldependencies.org

14 https://github.com/EstSyntax/EDT

15 https://cl.ut.ee/korpused/paralleel

16 https://opus.nlpl.eu

17 https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles

Erimärgendusega korpusteks võib pidada veel mitmekeelseid korpusi, mis on tehtud mingil kindlal eesmärgil: näiteks ühiskonna- ja poliitikauuringuteks on koostatud suur rahvusvaheline **parlamendikõnede korpus** ParlaMint[18], mis sisaldab 29 riigi või autonoomse piirkonna parlamentides peetud kõnede stenogramme. Korpus sisaldab ka Eesti Riigikogus peetud kõnesid. Korpustes on muu hulgas märgitud näiteks kõnedele vahelesegamised või aplaus; lisaks on esitatud palju metaandmeid kõnelejate kohta (kõneleja nimi, sugu, partei, kuulumine koalitsiooni või opositsiooni jne).

2.3. Erikorpused

Kuna eelnevalt kirjeldatud korpused sisaldavad suhteliselt standardset eesti kirjakeelt, siis tuleb meeles pidada, et nende põhjal saame uurida vaid kirjalikku tänapäevast eesti keelt. Just nende tekstide jaoks on loodud ka olulised eesti keele tööriistad, nagu morfoloogia analüsaator. Lisaks sellele on aga hulgaliselt keelevariante, mis erinevad standardkeelest: näiteks erinevad murded või muud keelevariandid, vanas kirjaviisis kirjutatud tekstid või muud vanemad tekstid, mille keelekasutus erineb tänapäevast, samuti suulise kõne tekstid jm. Ka nende keelevariantide uurimiseks on loodud mitmeid korpusi. Siia hulka kuuluvad näiteks vana kirjakeele korpus, murdekorpus, teismeliste keele korpus, suulise kõne korpus, õppijakeele korpused jne. Vaatleme taolisi erikorpusi kolmes osas: esmalt mittestandardset kirjutatud keelt sisaldavaid korpusi, seejärel suulise keele korpusi ja viimaks õppijakeele korpusi.

2.3.1. Mittestandardset kirjutatud keelt sisaldavad korpused

Mittestandardset kirjutatud keelt sisaldavad korpused on näiteks vanemaid eestikeelseid kirjalikke tekste sisaldavad korpused, milles tekstid on üles kirjutatud väga varieeruvalt: need võivad olla kirjutatud kindlate reegliteta nn ebakorrapärases kirjaviisis, tänapäevasest tavast oluliselt erinevas nn vanas kirjaviisis[19] või hoopis mõnes niisuguses süsteemis, mille kirjutaja pakkus välja olemasoleva kirjaviisi täiustamiseks. Lisaks kirjaviisi varieerumisele on Eestis kuni 19. sajandi lõpuni olnud kasutusel kaks eri kirjakeelt: lõunaeesti (tartu) ja põhjaeesti (tallinna) keel. Palju varieerumist on seotud kirjutajate emakeele taustaga. Vanemate tekstide üleskirjutajad on olnud enamasti sakslased, tüüpiliselt saksa päritolu pastorid, kelle tegevus oli seotud misjonitöö ja usukuulutamisega eestlastele. Seetõttu on vanemate tekstide hulgas ülekaalus piibli jm religiooniga seotud tekstid. Tekste

[18] https://www.clarin.eu/parlamint

[19] Vt https://eki.ee/teatmik/kirjaviis/

kasutades tuleb arvestada ka saksa pastorite väga erineva eesti keele oskusega. Kuna Eesti ala on olnud murdeliselt kirev, tuleb arvestada ka piirkonnaga, kus need saksa päritolu pastorid tegutsesid – tekstides leiame palju keelejooni, mis on seotud mingi Eesti piirkonnaga (murdealaga). Lisaks on allikad gooti kirjas, halva trükikvaliteediga või käsikirjalised ning seetõttu ei pruugi kõik sõnad tekstis olla õigesti välja loetud (vt lähemalt nt Kingisepp, Prillop & Habicht 2004).

Seda olulisem on taolise korpuse puhul tekstide lemmatiseerimine ja morfoloogiline märgendamine: see aitab uurijal keelelisest kirevusest vajaliku üles leida ning sõnavormi tõlgendustöö on tema eest juba tehtud. Paraku tänapäeva keele automaatseks analüüsiks mõeldud vahendid (näiteks morfoloogia analüsaator) nende tekstide puhul ei tööta. See tähendab, et taolised korpused on koostatud ja märgendused lisatud käsitsi. Nii korpuse jaoks kasutatavate tekstide maht kui käsitsitöö seavad korpuse suurusele piirangud, st need ei saa olla nii suured kui automaatselt veebist korjatud korpused.

Seega on vanemate tekstide korpuste koostamisel ja nende kasutamisel palju piiranguid, ent sellegipoolest on need olulised allikad, kust saame teada, milline oli eesti kirjakeel näiteks 500 aastat tagasi, mil ilmusid esimesed eestikeelsed raamatud. Keeleuurimise seisukohast on need äärmiselt vajalikud allikad, mis aitavad meil mõista tänapäevase eesti keele kujunemislugu. Kaks suuremat korpust, mis sisaldavad vanemaid eestikeelseid kirjalikke tekste, on eesti vana kirjakeele korpus ja piiblitõlgete korpus (neist täpsemalt pisut allpool). Varieeruvat kirjalikku keelt võib muidugi leida ka mujalt kui vana kirjakeele korpustest. Näiteks vanemat teksti on palju Eesti Rahvusraamatukogu digitaalsete ajalehtede kogudes, mida saab kasutada ka korpusena[20] ja analüüsida korpuslingvistika töövahenditega (vt P. Tinitsa näidisuurimus). Samasugune väga varieeruvat ajaloolist keelekasutust sisaldav kollektsioon on Rahvusarhiivi vallakohtuprotokollide kollektsioon[21], kus on hulgaliselt 19. sajandi vallakohtute protokolle (nende keelelise töötlemise võimalustest vt Pilvik jt 2019, Jaanimäe 2021). Vallakohtuprotokollide nimeüksuste automaatsest tuvastamisest oleme juhtumiuuringu lisanud ka käesolevasse õpikusse (K. Muischneki ja S. Orasmaa näidisuurimus). Taolisi kollektsioone tekib mäluasutustes talletatava pärandi digitaliseerimise käigus järjest juurde ning kui need on koostatud piisavalt süstemaatiliselt, on neid võimalik korpuslingvistika vahenditega analüüsida. Varieeruv keel aga kahtlemata komplitseerib seda.

Varieeruvat kirjutatud keelt leiab muidugi üksjagu ka tänapäeval, eelkõige tänu erinevate digitaalsete platvormide ja sotsiaalmeedia võimalustele. Ennekõike digitaalsetel platvormidel toimuvat keelelist kirjustumist ning standardkeelest eemaldumist on vahel nimetatud ka destandardiseerumiseks (Kristiansen 2021). Üks selliseid olulisi kirjaliku suhtluse vorme tänapäeval on sünkroonilist ja asünkroonilist suhtlust võimaldavad sotsiaalmeediaplatvormid. Sünkroonilist

[20] https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/

[21] https://www.ra.ee/vallakohtud/

(reaalajas toimuvat) kirjalikku dialoogi või polüloogi võimaldavad nn tšätid ja jututoad. Juba 1990ndatel levinud jututubade keelekasutust on kogutud näiteks koondkorpuse uue meedia alamkorpusesse[22]. Nn tšätivestlusi (mida võimaldavad nt Meta Messenger, WhatsApp jpt platvormid) on kogutud suulise keele korpuste juurde, sest nende puhul on huvipakkuvaks peetud suulise suhtlusega sarnaseid omadusi, nagu vooruvahetus, parandusmehhanism, suhtluspartiklite kasutamine jms (vt nt Hennoste 2000). Tšätivestlusi on kogutud eesti suulise kõne korpusesse ja teismeliste keele korpusesse (vt alapeatükk 2.3.2).

Eesti vana kirjakeele korpus[23] (VAKK, Prillop 2013) sisaldab põhiosas 16. kuni 19. sajandi tekste ning võimaldab seega uurida eesti keele ajalugu kirjalike tekstimälestiste põhjal. Päris esimesed eestikeelsed tekstikatked, mis korpusesse on kaasatud, on pärit 13. sajandist, ent need on pigem lühikesed fragmendid muukeelse teksti sees. Selline on Henriku Liivimaa kroonika lause 1224–1227 „Maga magamas", mille puhul pole siiski selge, kas tegemist on eesti või hoopis liivi keelega. Esimene säilinud trükitekst, mis on korpuses, on pärit aastast 1535 (Wanradti-Koelli katekismus).

15. ja 16. sajandist on korpusesse lisatud kõik teadaolevad ja säilinud eestikeelsed tekstid (v.a nimeloendid), nii käsikirjad kui ka trükised. Tekstid on vana kirjakeele uurijate poolt märgendatud.

17. sajandist on korpusesse lisatud enamik säilinud trükitekste, sajandi esimesest poolest ja keskpaigast ka käsikirjalisi jutlusi. Suurem osa tekste on uurijate poolt märgendatud.

18. ja 19. sajandist on lisatud valik trükitekste, sest eestikeelsete tekstide hulk oli selleks ajaks juba oluliselt kasvanud. Märgendatud on osa tekstidest, kuid märgendus on automaatne ja üldiselt ei ole veel uurijate poolt üle kontrollitud.

Vana kirjakeele korpuse märgendamine on tehtud (pool)käsitsi: kasutatud on abiprogramme, mis kiirendavad korduvate sõnavormide märgendamist, ent otsuse teeb märgendaja siiski oma teadmiste põhjal. Märgendatud on metaandmed: keel, ilmumine (trükis või käsikirjaline), aasta, autor, tekstiliik (ilmalik, vaimulik, kohtuprotokoll vms), teksti pealkiri. Iga sõne on märgendatud ka morfoloogiliselt: lisatud on märksõna (lemma tänapäevastatud kujul), sõnaliik, vormiinfo ja keel.

Korpuses on 23.6.2025 seisuga kokku 3025 teksti 3 349 065 sõnega (sh vallakohtuprotokollide kollektsioon). Eestikeelseid sõnesid on 2 870 480, neist märgendatud on 1 671 545. Korpusel on oma kasutajaliides[24], mille põhjal saab teha päringuid. Vana kirjakeele tekste saab ka veebis lugeda, samuti on nende põhjal koostatud vana kirjakeele sõnastikke.

[22] https://www.cl.ut.ee/korpused/segakorpus/uusmeedia/

[23] https://vakk.ut.ee/

[24] https://vakk.ut.ee/otsi.php

Eesti piiblitõlke ajalooline konkordants[25] (Ross & Reila 2020) on andmebaas, millesse on koondatud enamik säilinud eestikeelseid piiblitõlkeid ja piiblitõlkekatkendeid kuni esimese trükipiiblini (1739). Andmebaasi eesmärk on pakkuda ülevaadet vaimuliku eesti keele kujunemisloost 17. sajandil ja 18. sajandi alguses. Andmebaas sisaldab tõlketekste ja nende põhjal koostatud märksõnastikku ning võimaldab otsinguid a) tekstide kaupa, b) kindla piiblikoha järgi, c) tänapäevastatud märksõna järgi.

Andmebaas võimaldab piiblitekstide erinevaid tõlkeid omavahel võrrelda ja selle kaudu saada aimu, kuidas piibilt on eri aegadel tõlgitud ning kuidas need tõlked on üksteist mõjutanud. Piiblitõlke ajalooline konkordants täiendab vana kirjakeele korpust, kuhu piiblitõlkeid ei ole kaasatud. Andmebaasi kogumaht on praegu umbes kaks miljonit tekstisõna.

2.3.2. Suulised korpused

Suulise keele korpused on oma olemuselt keerukamad, sest esmalt tuleb kõne kirja panna ehk **transkribeerida** ning alles seejärel on võimalik teksti uurimiseks kasutada või edasi märgendada muudest parameetritest lähtuvalt. Lisaks on suuliste korpuste puhul võimalik kõnet ja teksti **aegjoondada** nii, et transkriptsioon on kohakuti helifaili ajateljega. See võimaldab igal hetkel tagasi pöörduda vastava helifaili juurde ning huvipakkuvat kohta uuesti üle kuulata. Tänapäeval on võimalik lisaks helifailile kasutada ka videot. Enamasti nõuab see mingi lisaprogrammi kasutamist korpuse loomisel, aga ka kasutamisel. Heli ja transkriptsiooni aegjoondamiseks sobivad näiteks programmid Praat ja ELAN. Korpuste kasutamise teevad need mõnevõrra keerukamaks, seetõttu on suuliste korpuste kasutamisel kasuks vastava programmi tundmine või mõne programmeerimiskeele tundmine. Enamasti on eesti suulistel korpustel siiski ka oma kasutajaliides, mis aitab teha lihtsamaid päringuid ilma programmeerimisoskuseta.

Suulise keele korpused võivad sisaldada väga erinevat tüüpi teksti: argivestlusi, raadio- ja telesaateid, taskuhäälinguid, murdekeelt vms. Suulise keele kirjapanekuks valitud transkriptsioon võib traditsioonist ja uurimiseesmärkidest tulenevalt olla korpustes väga erinev. Suulise keele uurimise seisukohalt on oluline, et tekstid oleks kirja pandud võimalikult täpselt ning annaksid edasi kõnele iseloomulikke nähtusi. Selleks transkribeeritakse ka poolikud sõnad, suulisele kõnele iseloomulikud partiklid (*noh, jah, a, eino* jms), pausid, üneemid ehk täidetud pausid (*ee, aa*) jms. Varieeruda võib see, kuidas hääldusе varieerumist kajastada: kas transkribeerimisel on lähtutud kirjakeele traditsioonist või mitte (nt kas *lihtsalt* või *lissalt*, kas sõnaalguline väljahääldamata *h* märkida või mitte), kas märgitud on rõhku, intonatsioonimuutusi, venitusi, kokkuhääldusi, palatalisatsiooni, väldet või muud sellist, mida kirjutatud keeles tavaliselt ei kajastata. Iga korpuse puhul on need

[25] https://arhiiv.eki.ee/piibel/

valikud tehtud ja dokumenteeritud, seega tasub enne kasutamist hoolikalt tutvuda korpuste dokumentatsiooniga.

Tänapäeval on suuremate ja standardsemat keelt sisaldavate korpuste puhul sageli kasutatud automaatset **kõnetuvastust**. Paraku sõltub kõnetuvastuse kvaliteet paljudest asjaoludest – salvestuse kvaliteedist, tekstide sisust, kõnelejate kõnetempost ja artikulatsiooni selgusest, kõnelejate häälte sarnasusest (kõnelejatuvastus), ennekõike aga sellest, mis tekstidel on kõnetuvastust treenitud ja kui erinev sellest on materjal, millest korpus koosneb. Transkriptsiooni kvaliteet sõltub seetõttu ka korpuse mahust: väiksema korpuse puhul on kõnetuvastusest tulnud tekst käsitsi üle kontrollitud, ent suurte korpuste puhul (näiteks ERR-i raadiosaadete ja taskuhäälingute korpus) seda võimalust ei ole ning kasutaja peab arvestama võimalike tuvastusvigadega. Eesti korpuste koostamisel on kasutatud Tallinna Tehnikaülikooli kõnetuvastust, mida on pidevalt arendatud erinevate andmestike toel (Olev & Alumäe 2022).

Suulise keele korpuste kasutamisel on meil vaja ka salvestuste **metaandmeid** – näiteks keda, millal ja mis oludes on salvestatud, sest see võib mõjutada inimeste keelelisi valikuid. Eriti olulised on metaandmed keele varieerumise ja sotsiolingvistilise uurimise jaoks, kus kõneleja andmed (nt vanus, sugu, päritolu, haridus) on vajalikud järelduste tegemiseks selle kohta, kuidas mingid keelenähtused on levinud ja kuidas neid kasutatakse. Näiteks murdekorpuse kasutamisel on väga tähtis, et meil oleks teada kõneleja päritolu (murdeala, kihelkond), sest vastasel juhul me ei tea, mis murde kohta me järeldusi teeme. Samuti on oluline salvestusaeg, sest murrete kasutus on aja jooksul muutunud.

Korpustesse on reeglina kõnelejate ja salvestuse kohta metaandmeid kogutud, ent kuna need on isikuandmed, ei luba isikuandmete kaitse nõuded neid kergekäeliselt jagada. Isikuandmete kaitseks on suulised korpused sageli kasutuspiirangutega ja nende kasutamiseks peab olema sõlmitud leping korpuse haldajatega. Lisaks võib kogu korpuse kasutamisel olla piiranguid, sest see võib sisaldada muidki delikaatseid isikuandmeid: näiteks kõneleja häält, mis võimaldab teda tuvastada, või siis teksti sisu, mis võib kõnelejaid kahjustada, kui korpuse materjalid satuvad halbadesse kätesse. Seega on piirangud korpuste kasutamisel seotud vajadusega kõnelejat kaitsta. Samal põhjusel on avalikes otsimootorites suulise kõne tekste vähe või on need kättesaadavad vaid väikeste lõikudena: näiteks spontaanse kõne foneetilise korpuse otsimootoris on võimalik päringu tulemusena saada otsisõna koos heliga, ent selle kontekst on piiratud kahe sekundiga.

Tartu Ülikooli **suulise eesti keele korpus**[26] (SEKK, Rääbis 2013) on vestlusanalüüsi raamistikus loodud korpus, mille esmane kasutusala on olnud suulise suhtluse uurimine. Suulise keele korpust hakati koguma 1997. aastal ning see töö jätkub. Korpus koosneb salvestistest, transkriptsioonidest ja nende juurde

[26] https://keeleressursid.ee/et/220-suulise-eesti-keele-korpus

kuuluvatest põhjalikest taustakirjeldustest, mis sisaldab infot nii vestluse osaliste kui kõnesituatsiooni kohta (Hennoste 2000; Hennoste 2003; Hennoste jt 2009).

Vestlused on salvestatud loomulikes suhtluskeskkondades. Korpuse kogumist alustati kassett-diktofoniga, need salvestised on hiljem digitaliseeritud. Peagi mindi üle digisalvestamisele ja alates 2010. aastast on tehtud valdavalt videoid. Salvestus ja transkriptsioon on ilma aegjoonduseta.

Salvestiste kogumaht on 2024. a seisuga 835 tundi. Salvestised jagunevad argi- (910 salvestist) ja ametlikuks suhtluseks (4000 salvestist); 100 salvestist on nende segu või raskesti määratletavad. Suhtlusviisi järgi jagunevad vestlused vahetuks ja vahendatud suhtluseks; täpsemalt: silmast silma suhtlust on 1420 salvestist, telefonisuhtlust 3030, meediasuhtlust 550 ja arvuti vahendatud (nt Skype, Zoom) suhtlust 10 salvestist.

Suulise kõne korpuses kasutatud transkriptsiooni põhimõtted[27] pärinevad vestlusanalüüsist. Transkribeeritud on sõnad, üneemid, pealeräägimised, pausid, intonatsioonipiirid, rõhud, venitused, katkestused, naer, hääletooni muutused jm nähtused, mis on vajalikud suulise suhtluse analüüsimiseks. Materjal on transkribeeritud Wordis, tekstid on olemas ka TXT-failidena. Salvestisi on transkribeeritud umbes 2,6 miljoni sõne mahus, sellest argisuhtlust on 1 miljon sõnet ja ametlikku suhtlust 1,6 miljonit sõnet. Suhtlusviisi järgi jaotatuna on silmast silma suhtlust üle 1,5 miljoni sõne, telefonisuhtlust ja meediasuhtlust kumbagi u pool miljonit sõnet, arvuti vahendatud suhtlust 33 000 sõnet.

Korpuse otsimootor võimaldab otsida sõnavorme ja nende häädusvariante. Otsimootor ei ole avalikult ligipääsetav. Korpuse kasutamiseks tuleb individuaalset ligipääsu küsida korpuse administraatorilt ning allkirjastada konfidentsiaalsuskohustus. Väike osa korpusest (100 000 sõna) on märgendatud ja lisatud morfoloogiliselt käsitsi ühestatud korpusesse[28] (vt alapeatükk 2.2.4), mis on kättesaadav ka korpusanalüüsi keskkonnas KORP.

Tartu Ülikooli suulise ja arvutisuhtluse labor on aastast 2009 tegelenud lisaks suulise keele kogumisele ka nn netikeele tekstide kogumisega. Koostatud on tekstikogu, mis sisaldab sünkroonset suhtlust, peamiselt nn tšätivestlusi (ingl *instant messaging*). Vestlused on kogutud eri veebikeskkondadest (peamiselt MSN Messenger, Skype, Meta Messenger). Suhtlejateks on põhiliselt üliõpilased ja nende sõbrad-sugulased. 2024. aasta seisuga koosneb tekstikogu ligikaudu 900 vestlusest. Iga vestlus on varustatud taustakirjeldusega, mis on analoogiline suulise eesti keele korpuse taustakirjeldusega. Vestlused on säilitatud eri vormingutes (nt RTF, DOC, DOCX, HTML). Korpuse koostamine ja korrastamine jätkub.

Eesti keele spontaanse kõne foneetiline korpus[29] (EKSKFK, Lippus, Aare, jt 2023) on Tartu Ülikooli foneetikalaboris koostatav suulise kõne korpus, mis

[27] https://keeleressursid.ee/et/148-suulise-kone-transkriptsioon

[28] https://cl.ut.ee/korpused/morfkorpus/

[29] https://foneetikakorpus.ut.ee/ekskfk_info.html

koosneb spontaansete vestluste salvestustest ning nende transkriptsioonist ja märgendusest erinevatel lingvistilistel tasanditel. Ehkki korpus on esmajoones loodud häälduse uurimiseks, on selle kasutusala märksa universaalsem, näiteks on selle põhjal võimalik uurida lisaks veel grammatikat või suhtlust.

Korpus koosneb põhiosas spontaansetest dialoogidest kahe kõneleja vahel. Iga vestlus kestab umbes pool tundi. Salvestused on tehtud foneetikalabori vaikses salvestusruumis ja iga kõneleja on eraldi kanalis. Vähemal määral on korpuses ka monoloogilisi tekste (peamiselt loengute või ettekannete salvestused) ning kolme inimese vestlusi. Osa salvestisi on tehtud välitöödel ja auditooriumis (monoloogid). Suurem osa salvestusi on ainult helifailidena, umbes kolmandik korpusest on salvestatud ka videoga ning võimaldab uurida ka multimodaalset suhtlust.

Korpuse transkribeerimiseks ja märgendamiseks kasutatakse kõneanalüüsiprogrammi Praat, segmentimis- ja märgendusinfo salvestatakse TextGrid-vormingus. Transkriptsioon on tehtud käsitsi, seda on tehtud sõnatasandil ja häälikutasandil. Sõnatasandi transkriptsioon järgib tavaortograafiat väikeste erisustega (nt liitsõnad on märgitud plussiga, kasutatud on mõningaid lisamärgendeid). Häälikutasandil järgitakse SAMPA transkriptsiooni[30] (lihtsustatud ASCII versioon rahvusvahelisest IPA transkriptsioonist). Aegjoondusega märgendus sisaldab sõnu, häälikuid, silpe, kõnetakte, häälelaadi. Vabamorfi tarkvara abil on lisatud automaatselt ühestatud morfoloogiline analüüs.

Korpus on kogutud aastatel 2006–2023. Korpuses on kokku umbes 135 tundi kõnet ja üks miljon sõnatasandi segmenti, mille hulka kuuluvad ka näiteks pausid ja üneemid. Ilma pauside ja üneemideta morfoloogiliselt märgendatud sõnu on korpuses 629 699. Korpuses on kasutatud materjali 207 kõnelejalt üle Eesti. Kõnelejate kohta on kogutud sotsiolingvistilisi taustaandmeid (sugu, sünniaasta, sünnikoht, haridus, vanemate piirkondlik päritolu).

Korpuse kodulehel on avalik otsimootor, millega saab otsida ühe sõna piires ja vasted antakse kahesekundilises kontekstis. Korpus on kasutuspiirangutega, selle kasutamiseks tuleb sõlmida leping. Tervikkorpusele on võimalik taotleda ligipääsu repositooriumis[31]. EKSKFK võimalusi foneetika uurimiseks tutvustab selles õpikus näidisuurimuste osas P. Lippus.

ERR-i raadiosaadete ja taskuhäälingute korpused on 2023. aastal Tartu Ülikoolis koostatud suured suuliste tekstide korpused, mis on saadud veebikraapimise teel (helifailid) ning seejärel automaatselt kõnetuvastatud ja morfoloogiliselt märgendatud. Mõlemad korpused on aegjoondusega.

ERR-i korpuses[32] on Eesti Rahvusringhäälingu arhiivist saadud 53 000 raadiosaadet kogukestusega 16 047 tundi. Saated on salvestatud vahemikus 1930–2022,

[30] https://foneetikakorpus.ut.ee/ekskfk_margendamise_juhend.html#4_H%C3%A4%C3%A4likutasand

[31] https://datadoi.ee/handle/33/577

[32] https://datadoi.ee/handle/33/581

aga enamik saateid on pärit 2000. aastatest. Kokku on korpuses 110 miljonit sõna (Lippus, Alumäe, Orasmaa, Tsepelina, jt 2023).

Taskuhäälingute ehk *podcast*'ide korpuses[33] on 10 633 episoodi 184 erinevast taskuhäälingust, mis on salvestatud vahemikus 2018–2022. Salvestuste kestus kokku on 10 918 tundi, transkribeeritud tekstide maht on 85 miljonit sõna (Lippus, Alumäe, Orasmaa, Pilvik, jt 2023).

Salvestused on transkribeeritud Tallinna Tehnikaülikooli automaatse kõnetuvastusega ning tekstid on automaatselt morfoloogiliselt analüüsitud EstNLTK-ga. Korpuste kasutamisel peab kindlasti arvestama sellega, et tegemist on kõnesalvestuste automaatse transkriptsiooniga, mida ei ole kontrollitud ega parandatud ning seetõttu sisaldab transkriptsioon vigu. Samuti on morfoloogiline info lisatud automaatselt transkribeeritud tekstile ja on automaatselt ühestatud ning seetõttu võib ka see sisaldada vigu.

ERR-i raadiosaadete korpuse tekstid on kättesaadavad keskkonnas KORP[34]. Mõlemad korpused on kättesaadavad DataDOI repositooriumis, kus kasutajal on võimalik taotleda ligipääsu korpusefailidele. Mõlemad korpused on ligipääsupiirangutega, mis tagavad materjali kasutamise teaduskasutuseks ning piiravad levitamisõigusi.

TTÜ eesti laste ja noorte kõnekorpus (Meister 2015) sisaldab 309 eesti emakeelega lapse ja noore (vanuses 9–18) kõnesalvestusi. Igalt kõnelejalt on salvestatud nii etteloetud kui spontaanset kõnet. Korpuse loetud kõne osa on täies mahus märgendatud automaatselt sõna ja hääliku tasandil (kasutades TTÜ automaatse segmenteerimise tarkvara); käsitsi (programmiga Praat) on sõna ja hääliku tasandil märgendatud foneetiliselt rikkad laused, koha-, isiku- ja organisatsiooninimedega laused ning lühijutud. Korpuse spontaanse kõne osa on transkribeeritud käsitsi (programmiga Transcriber) täies mahus. Korpuse loomise eesmärk on olnud uurida laste kõne akustilis-foneetiliste tunnuste arengut, mis kaasneb vanuse ja sooga seotud anatoomiliste muutustega kõnetraktis. Vaadeldavasse vanuserühma jääb ka häälemurre. Korpust kasutatakse ka automaatse kõnetuvastuse treenimiseks.

Korpuse kõnelejad on pärit üle Eesti. Kõigilt osalejatelt on kogutud metaandmeid (vanus, sugu, kool, klass, elukoht jne). Korpuse koostamise kohta vt lähemalt (Meister & Meister 2017). Korpus on akadeemiliseks uurimistööks tasuta kasutamiseks ega ole avalikult kättesaadav. Huvi korral tuleb pöörduda TTÜ keeletehnoloogia labori poole.

TTÜ seenioride kõnekorpus sisaldab eakate (üle 60-aastaste inimeste) keelekasutust. Korpus sisaldab keskmiselt umbes pool tundi spontaanset kõnet 100 mehelt ja 100 naiselt. Spontaanset kõnet salvestati perioodil 2018–2022 intervjuu vormis, mille käigus puudutati lapsepõlvemälestusi, kooliaega, tööelu, hobisid jms.

[33] https://datadoi.ee/handle/33/585

[34] https://korp.keeleressursid.ee ja https://korp.eki.ee/

Korpuse kogumaht on umbes 95 tundi. Korpus on transkribeeritud TTÜ automaatse kõnetuvastusega. Igast tekstist umbes 20 minutit on käsitsi kontrollitud programmiga Transcriber; kokku on transkribeeritud ja kontrollitud tekstide maht umbes 70 tundi. Kõnelejate kohta on kogutud metaandmeid (vanus, sugu, haridus, võõrkeeled, elukoht lapsepõlves ja praegu) (Meister & Meister 2022).

Eakate kõnekorpus on vajalik kõnetuvastuse treenimiseks ja sotsiofoneetilisteks uuringuteks, kuid võib pakkuda huvi ka sotsiolingvistika ja kultuuriloo uurijatele. Korpus on akadeemiliseks uurimistööks tasuta kasutamiseks, aga ei ole avalikult kättesaadav. Huvi korral tuleb pöörduda TTÜ keeletehnoloogia labori poole.

Eesti teismeliste keele korpus[35] on Tartu Ülikoolis koostatav korpus, mis sisaldab 9–18-aastastelt noortelt aastatel 2020–2022 nn kodanikuteaduse meetodil kogutud suulist ja kirjalikku keelematerjali. Nelja Eesti piirkonna (Antsla, Kuressaare, Tallinna ja Tartu) teismelised (nn keelesaadikud) salvestasid pärast lühikese koolituse läbimist enda ja oma sõprade vahel peetud spontaanseid vestlusi ning jagasid neid korpuse koostajatega. Samuti jagasid nad oma kaaslastega suhtlusrakendustes (Messenger ja Discord) peetud tšätivestlusi. Kokku on materjali 131 noorelt.

Suulisi vestlusi on korpuse 2024. aasta veebruaris avaldatud versioonis (Vihman jt 2023) kokku 116, need on keskmiselt umbes tund aega pikad ning programmiga ELAN käsitsi transkribeeritud. Transkriptsioonid järgivad üldjoontes tavaortograafia reegleid, ent sisaldavad ka erimärgendeid näiteks koodivahetuse, kõnekeelsuste, naeru jm kohta. Transkriptsioon on aegjoondusega. Morfoloogiline analüüs on suulistele vestlustele lisatud automaatselt. Kokku on suulise alamkorpuse maht 91 tundi ja 741 175 sõna.

Tšätivestlusi on korpuses 110 (kokku 65 456 sõna) ning nendes on käsitsi märgendatud koodivahetuse ja lühendite kasutus, samuti on vestlustes eraldi märgendatud suhtlusrakenduste metatekstid. Tšätivestlused morfoloogilist märgendust ei sisalda. 2024. aastal käivitus uus tšätivestluste korje, mistõttu korpus täieneb.

Teismeliste keele korpus on ligipääsupiiranguga kättesaadav DataDOI repositooriumis. Ehkki korpuse tekstides esinevad isikunimed jm tundlikud andmed (nt sünnipäevad, aadressid, paroolid) on pseudonümiseeritud, tuleb materjalide kasutamiseks sõlmida konfidentsiaalsusleping ning määratleda selgelt korpuse kasutamise eesmärk ja materjalide kasutamise periood.

Eesti murrete korpus[36] (EMK) sisaldab autentseid murdetekste nii põhja- kui lõunaeesti murretest ning võimaldab võrdlevalt uurida erinevaid keelenähtusi eesti murretes. Korpus baseerub helisalvestistel, mis on tehtud põhiosas vahemikus 1957–1980 välitööde käigus. Korpus on koostatud Tartu Ülikoolis koostöös Eesti Keele Instituudiga, mille murdearhiivi on korpuses kasutatud. Helisalvestised on

[35] https://datadoi.ee/handle/33/596

[36] https://datadoi.ee/handle/33/492

litereeritud soome-ugri foneetilises transkriptsioonis ning sellest tuletatud lihtsustatud transkriptsioonis. Foneetilises transkriptsioonis tekste ja helisalvestisi säilitatakse Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis[37]. Lihtsustatud transkriptsioonis tekstid on omakorda morfoloogiliselt märgendatud. Morfoloogiline märgendamine on tehtud käsitsi, kasutades abiprogrammi Liivike. Korpus sisaldab liivi ja vadja keele alamkorpusi, mis on osalt suulised, osalt põhinevad varem välja antud tekstikogumikel (suuline keel, mis on litereeritud ja toimetatud väljaande põhimõtetest lähtudes).

EMK maht on umbes 1,2 miljonit märgendatud tekstisõna eesti murretest, millele lisandub umbes 45 000 märgendatud tekstisõna liivi keelest ning 35 000 märgendatud tekstisõna vadja keelest. Murdekorpuse morfoloogiliselt märgendatud osa saab kasutada veebipõhise otsimootori kaudu[38] ning XML-vormingus alla laadida DataDOI kaudu (Lindström, Todesk & Pilvik 2022). Sealt leiab ka korpuse ülevaate. Foneetilises ja lihtsustatud transkriptsioonis tekstide jaoks eraldi otsimootorit ei ole, ent need on kättesaadavad murdearhiivi kaudu (vt ka Lindström, Lippus & Tuisk 2019).

Murdekorpuse kõnelejate kohta käiv olulisem info on lisatud tekstide juurde. Avalikus otsimootoris on eemaldatud kõnelejate nimi, ent muid ligipääsupiiranguid ei rakendata. Repositooriumis on kõneleja nime sisaldavas versioonis rakendatud ligipääsupiiranguid. Liivi keele alamkorpus kajastub repositooriumis omaette korpusena[39], sama on plaanis vadja keele korpusega.

Eesti murrete korpuse põhjal saab uurida võrdlevalt eesti murrete grammatika ja häälduse varieerumist; selle kohta on L. Lindströmi ja M.-L. Pilviku näidisuurimus ka käesolevas õpikus.

Seto korpus[40] sisaldab uuemat seto keelt ning põhineb murdekorpusega võrreldaval andmestikul – välitööde käigus läbi viidud intervjuudel. Korpust koostatakse Tartu Ülikoolis. Korpus põhineb andmetel, mis on kogutud aastatel 2010–2023 Eestis ja Venemaal Petseri rajoonis läbi viidud välitööde käigus. Välitöödel on küsitletud palju usundi, kombestiku ja õigeusu kirikuga seotud tavade teemal, mistõttu võiks korpus huvi pakkuda ka etnoloogidele, folkloristidele ja religiooniuurijatele. Korpus on koostatud ELAN-is, kus heli, transkriptsioon ja märgendus on aegjoondusega. 2024. aasta seisuga on repositooriumisse jõudnud 36 tundi transkribeeritud ja käsitsi morfoloogiliselt märgendatud teksti (ligi 250 000 tekstisõna), ent korpuse maht kasvab veelgi. Nii transkriptsiooni kui märgenduse osas toetutakse murdekorpuse põhimõtetele, väikeste eranditega. Korpuses kasutatakse lihtsustatud transkriptsiooni[41], mis erineb pisut eesti tavaortograafiast. Seto korpus

[37] https://murdearhiiv.ut.ee/

[38] https://murre.ut.ee/

[39] https://datadoi.ee/handle/33/617

[40] https://osf.io/8WB2J/

[41] https://setko.ut.ee/lihtsustatud-transkriptsioon/

on kättesaadav murdekorpusega samas otsimootoris[42] ja OSF-i repositooriumis (Lindström, Pilvik & Todesk 2024). Korpus on terviktekstidena kasutatav vaid ligipääsupiirangutega.

2.3.3. Õppijakeele korpused

Õppijate keelt kogutakse keeleõppeprotsessi uurimiseks ja selle põhjal keeleõppe jaoks soovituste tegemiseks. Õppijakeele all mõeldakse mingit keelt teise keelena või võõrkeelena õppijate keelekasutust (L2), aga ka keelt emakeelena (esimese keelena) õppijate keelekasutust (L1). Õppijakeele hulka võib lugeda ka lastekeele, sest lapsed omandavad alles oma esimest keelt (või mitut keelt korraga). Õppijakeele uurimise abil saame teada, kuidas keeleõppija esimest või teist (mitmendat) keelt omandab, milliseid vigu ta seejuures teeb ning kuidas tema emakeel (esimene keel) uue keele omandamist mõjutab.

Õppijakeele korpused võivad sisaldada nii suulist kui kirjalikku keelekasutust. Vahel nimetatakse õppijakeele korpusi (ingl *learner corpus*) ka vahekeele korpusteks (ingl *interlanguage corpus*) või teise keele korpusteks (ingl *L2 corpus*). Õppijakeele korpuste alusel uuritakse keeleõppijate keelekasutust ja keeleoskuse arengut ning nende analüüs on sisendiks eri tüüpi keeleõppevahendite, nt õppesõnastike ja grammatikate koostamisel.

Eesti vahekeele ehk õppijakeele korpust[43] koostatakse Tallinna Ülikoolis (EVKK, vt Eslon & Metslang 2007; Eslon 2014) ja see koosneb eesti keelt teise keelena või võõrkeelena õppijate tekstidest. Korpuses on 2024. aasta alguse seisuga umbes 12 500 teksti. Korpus sisaldab valdavalt eesti keelt teise keelena õppijate kirjalikke tekste, millest suurem osa on kirjutatud eesti keele tasemeeksamitel. Üle 3000 teksti on kirjutatud eksamiväliselt, keeleõppe käigus (sh eesti keele kui teise keele olümpiaadi võistlustööd). Samuti hõlmab EVKK väikest võrdluskorpust, mis koosneb Postimehe ja Õhtulehe arvamuslugudest, ning akadeemilise eesti keele alamkorpust, kuhu kogutakse nii eesti kui ka muu emakeelega üliõpilaste töid. Kuna suurem osa eesti keelt teise keelena õppijatest on vene emakeelega, on kogutud ka venekeelsete õpilaste loovkirjutisi ning eesti emakeelega õpilaste venekeelseid kirjutisi, mis võimaldavad uurida emakeele mõju sihtkeele omandamisele. Alamkorpuste kohta saab lähemalt lugeda käsiraamatust „Eesti keele oskuse arenemine ja arendamine" (Eslon jt 2021).

Eesti vahekeele korpuse baasil on Tallinna Ülikoolis arendatud keeleõppekeskkonda ELLE[44], mis sisaldab ka korpuse analüüsivahendeid. Korpuse kasutamiseks tasubki esmalt tutvuda ELLE võimalustega.

[42] https://murre.ut.ee/

[43] https://evkk.tlu.ee/vers1/

[44] https://elle.tlu.ee/

EMMA on Tartu Ülikoolis loodud eesti keele õppijakorpus, mis sisaldab eksamitöid ja õppimisprotsessis koostatud tekste 9. ja 12. klassi õpilastelt aastatest 1999–2017. Need on valdavalt eesti keelt emakeelena (L1) õppijate tekstid. Kasutatud eksamikirjandid ja muud tekstid on algselt olnud käsikirjalised, need on ümber trükitud ja neis on märgendatud nii eksami- või tasemetööl näha olevad parandused kui ka tekstiehituslikud osad (sissejuhatused, kokkuvõtted, pealkirjad). Vigade märgendamise aluseks on võetud eksamikomisjoni loodud vigade kategoriseerimise süsteem. Korpuse koostamise kohta vt lähemalt (Sõrmus & Lepajõe 2014).

Alates 2019. aastast on koostöös Eesti Keele Instituudi ja HARNO-ga korpusesse lisatud eesti keelt teise keelena õppijate eestikeelseid tekste, samuti on laiendatud korpust nii eagrupi (on lisatud koolieelikute, 3. ja 6. klassi õpilaste tekstid) kui ka testiliigi (tasemetööd) osas. 2024. aasta seisuga on korpuses 15 953 teksti. Korpus sisaldab kokku 257 143 lauset, 3 205 290 sõna.

Kogu korpus on automaatselt lausestatud ehk jagatud lauseüksusteks ning märgendatud EstNLTK 1.4 abil. 2470 eesti keele emakeelena teksti sisaldavad lisaks vigade märgendust. Korpust on võimalik kasutada uurimis- ja teadustöö eesmärgil, ligipääs korpusele on piiratud.

Eesti lastekeele korpuse leiab rahvusvahelisest lastekeelekorpusi ühendavast keskkonnast CHILDES[45]. See sisaldab 9 eesti keele alamkorpust, mis on kogutud erinevatel aegadel ja erinevate projektide käigus. Alamkorpused on enamasti nimetatud korpuse koostaja järgi. Alamkorpused erinevad üksteisest mitmes mõõtmes: andmete kogumise viis, korpuse eesmärk, laste arv ja vanus, korpuse suurus ning hoidja ja lapse sõnade vaheline osakaal. Enamikku alamkorpusi iseloomustab pikiuurimusele omane lähenemine, kus teatud perioodi jooksul salvestatakse lapse ja hoidja(te) vestlusi ja selle kaudu uuritakse lapse keele omandamise kulgu, aga ka hoidjakeelt (lastele suunatud keelt) (Vaik & Vihman 2017). Transkribeerimisel on kasutatud CHILDES-is kasutatavat CHAT-transkribeerimissüsteemi.

TTÜ aktsendikorpus[46] võimaldab uurida eesti keele hääldust õppijakeeles. Aktsendina mõistetakse mingit keelt teise keelena kõneleja sihtkeele tüüpilisest häälдusest hälbivat kõnet, mida selle keele sünnipärased kõnelejad tajuvad võõrkeelse aktsendina. Aktsendikorpuse abil saab uurida näiteks eesti keele häälikuja vältesüsteemi omandamist, lauseintonatsiooni ning kõnerütmi. Aktsendikorpuse praktilise väljundina saab selle abil „õpetada" kõnetuvastustarkvara aktsendiga kõnet õigesti tekstiks teisendama. TTÜ aktsendikorpus sisaldab loetud lauseid ja spontaanset kõnet muu emakeelega eesti keele kõnelejatelt, kelle kõnes on võõrkeele aktsent kuuldeliselt tajutav. Korpuses on 18 erineva emakeele taustaga kõnelejarühma 187 kõnelejaga ning eesti emakeelega kontrollrühm (20 kõnelejat). Kõnelejate kohta on olemas metaandmed, mis sisaldavad mh nende eesti keele

[45] https://childes.talkbank.org/

[46] https://live.european-language-grid.eu/catalogue/corpus/14530

õppimise aega ja viise. Aktsendikorpus on osaliselt märgendatud ja segmenteeritud (sh foneetiliselt) Praatis (Meister & Meister 2012).

Korpus on tasuta kasutamiseks akadeemiliseks uurimistööks, aga ei ole avalikult kättesaadav. Huvi korral tuleb pöörduda TTÜ keeletehnoloogia labori poole.

Õpikukorpused sisaldavad õpikute tekste keeleõppija sisendkeele näitena. **Eesti keele kui teise keele õppekorpus 2022** (Koppel jt 2022) sisaldab lauseid kolmekümne neljast eesti keele kui teise keele õpikust.

Lõpetuseks

Eesti keel on oma kõnelejate arvu kohta väga hästi korpustega kaetud keel: meil on väga mitmekesiseid korpusi väga erinevate uurimiseesmärkide jaoks. Võrreldes nii mõnegi suurkeelega on nende dokumenteeritus ja kättesaadavus väga hea: olemas on nii kirjalikku kui ka suulist keelt sisaldavaid ning eri registreid katvaid korpusi, mida saab kasutada erinevatel eesmärkidel.

Teisalt on vajadus uute korpuste järele tänapäeva keeleteaduses paratamatu, sest keel ja keelekasutus muutuvad pidevalt, nagu ka olukorrad ja keskkonnad, kus keelt kasutatakse. Seetõttu on ka see ülevaade kindlasti peagi vananenud. Kui olemasolevatest korpustest ei piisa, võib oma uurimishuvi jaoks igaüks luua ise korpuse. Sellest, kuidas seda teha, räägime selle õpiku 4. peatükis.

Ülevaade kõigist selles osas käsitletud korpustest on koondatud tabelisse 2.3.

Tabel 2.3. Eesti keele korpusi

Korpus	Kirjalik/ suuline	Tekstide päritoluaeg	Suurus	Registrid	Morfoloogiline märgendus	Muu märgendus	Ligipääs
Baaskorpus	kirjalik	1980ndad	1 mln sõnet	kõik selle perioodi toimetatud kirjaliku teksti registrid	automaatne	teksti struktuur	vaba
Niitkorpus	kirjalik	1890–1990, v.a 1920ndad ja 1940ndad	3,5 mln sõnet	ajakirjandus, ilukirjandus	automaatne	teksti struktuur	vaba
Koondkorpus	kirjalik	1990–2010	250 mln sõnet	ajakirjandus (rõhuvas enamuses), ilukirjandus, teadus, populaarteadus, foorumid, kommentaarid, uudisgrupid, jututoad	automaatne	teksti struktuur	vaba
Tasakaalus korpus (koondkorpuse allosa)	kirjalik	1990–2010	15 mln sõnet	ajakirjandus, ilukirjandus, teadus	automaatne	teksti struktuur, automaatne piirangute grammatika pindsüntaks (Keeleveebis)	vaba
Ühendkorpus 2013	kirjalik	kuni 2013	563 mln sõnet	erinevad kirjalikud tekstid, sh eestikeelne veeb	automaatne	automaatne SketchEngine'i pindsüntaks	vaba
Ühendkorpus 2017	kirjalik	kuni 2017	1,1 mld sõnet	erinevad kirjalikud tekstid, sh eestikeelne veeb	automaatne	automaatne SketchEngine'i pindsüntaks	vaba

Korpus	Kirjalik/ suuline	Tekstide päritoluaeg	Suurus	Registrid	Morfoloogiline märgendus	Muu märgendus	Ligipääs
Ühendkorpus 2019	kirjalik	kuni 2019	1,5 mld sõnet	erinevad kirjalikud tekstid, sh eestikeelne veeb	automaatne	registrid, automaatne SketchEngine'i pindsüntaks	vaba
Ühendkorpus 2021	kirjalik	kuni 2021	2,4 mld sõnet	erinevad kirjalikud tekstid, sh eestikeelne veeb	automaatne	registrid, automaatne sõltuvussüntaks	vaba
Ühendkorpus 2023	kirjalik	kuni 2023	3,8 mld sõnet	erinevad kirjalikud tekstid, sh eestikeelne veeb	automaatne	registrid, automaatne SketchEngine'i pindsüntaks	vaba
Morfoloogiliselt käsitsi märgendatud korpus	kirjalik	1990–2003	0,5 mln sõnet	erinevad kirjaliku keele registrid	käsitisi	puudub	vaba
Eesti keele universaal-sõltuvuste UD puudepank	kirjalik	1990–2010	0,4 mln sõnet	erinevad kirjaliku keele registrid	käsitsi	käsitsi sõltuvussüntaks	vaba
Vana kirjakeele korpus	kirjalik	15.–19. saj	3,3 mln sõnet	vaimulikud tekstid, aja-kirjandus, ilukirjandus, õpetlikud tekstid	osal käsitsi, osal automaatne	keel, periood	vaba
Eesti piibli-tõlke ajalooline konkordants	kirjalik	17.–18. saj	2 mln sõnet	piiblitekstid	osa käsitsi lemmatiseeritud	peatükid, salmid	vaba

Korpus	Kirjalik/ suuline	Tekstide päritoluaeg	Suurus	Registrid	Morfoloogiline märgendus	Muu märgendus	Ligipääs
TÜ suulise eesti keele korpus	suuline	alates 1997	835 t salvestisi; transkribeeritud 2,6 mln sõnet	argivestlused, ametlikud vestlused	100 000 sõna käsitsi	pealerääkimised, pausid, intonatsioonipiirid, venitused, katkestused, naer, hääletooni muutused jm	piiratud
Eesti keele spontaanse kõne foneetiline korpus	suuline	2006–2023	135 t, kõnet; transkribeeritud 1 mln sõnet	vestlused, loengud, ettekanded	automaatne	aegjoondus, pausid, silbid, häälikud, kõnetaktid, häälelaad	piiratud
ERR-i raadiosaadete korpus	suuline	1930–2022	110 mln sõnet	raadiosaated	automaatne	puudub	piiratud
Taskuhäälingute korpus	suuline	2018–2022	85 mln sõnet	taskuhäälingud	automaatne	puudub	piiratud
TTÜ eesti laste ja noorte kõnekorpus	suuline	2011–2015	100 t salvestisi	loetud kõne, spontaanne kõne	puudub	puudub	piiratud
TTÜ seenioride kõnekorpus	suuline	2018–2022	95 t salvestisi	intervjuud	puudub	puudub	piiratud
Eesti teismeliste keele korpus	suuline ja kirjalik	2020–2022	suulisi vestlusi 0,7 mln sõnet; kirjalikke vestlusi 65 000 sõnet	suulised ja kirjalikud vestlused	automaatne	suulistes vestlustes pausid, koodivahetus, nimed, laulusõnad, pealkirjad; kirjalikes vestlustes koodivahetus, lühendid	piiratud

Korpus	Kirjalik/ suuline	Tekstide päritoluaeg	Suurus	Registrid	Morfoloogiline märgendus	Muu märgendus	Ligipääs
Eesti murrete korpus	suuline	1938–2010 (põhiosa 1957–1980)	1,2 mln sõnet	intervjuud	käsitsi	pausid, kokkuhäälduse, katkestused, välted, palatalisatsioon jm	osaliselt piiratud
Seto korpus	suuline	2010–2023	0,25 mln sõnet	intervjuu	käsitsi	keel, pausid, kokkuhäälduse, katkestused, välted, palatalisatsioon jm	piiratud
Eesti vahekeele ehk õppijakeele korpus	kirjalik	alates 2000	3,4 mln sõnet	teise keele õppe raames loodud tekstid, eksamitööd	automaatne	veamärgendus	piiratud
TÜ EMMA korpus	kirjalik	alates 1999	3 mln sõnet	emakeeleõppe raames loodud tekstid, eksamitööd	automaatne	veamärgendus	piiratud
Eesti lastekeele korpus	suuline	alates 1998	0,4 mln sõnet	lapsehoidja vestlused	puudub	pausid, katkestused, lühendid, hääletoon, rõhud jm	piiratud
TTÜ aktsendikorpus	suuline	alates 2011	80 t salvestisi	loetud kõne	puudub	osaliselt silbid, osaliselt häälikud	piiratud
Eesti keele kui teise keele õppekorpus 2022	kirjalik	2000–2022	35 700 lauset	õpikutekstid	automaatne	teksti struktuur	vaba

3. Märgendamine

Maarja-Liisa Pilvik, Kadri Muischnek, Pärtel Lippus, Siim Orasmaa

Ehkki omaette väärtus on ka lihtsalt suurtel tekstikogudel, on keeleteaduslikus, ent ka sellega piirnevate distsipliinide (nt etnoloogia, kirjandusteaduse, sotsiaalteaduste) uurimistöös tihtipeale rohkem kasu sellistest korpustest, mis võimaldavad uurijal kasutada lisaks **toortekstile** ehk korpuse algtekstile endale veel mingit infot teksti, selle struktuuri, sõnade vm kohta. Samuti on sellised korpused oluline ressurss mitmesuguste loomuliku keele töötluse vahendite loomiseks ja arendamiseks (nt kõnetuvastus, masintõlge, tekstide märksõnastamine ja klassifitseerimine, meelestatuse analüüs) ning need võivad toetada tõhusamat keeleõpet.

Protsessi, mille käigus varustatakse korpuse tekstid lingvistilise, temaatilise vm infoga, nimetatakse **märgendamiseks** (ingl *annotation*, vahel ka *markup*, *notation*). Märgendada võib näiteks terveid tekste, tekstides lõike, lõikudes lauseid, lausetes sõnu või sõnades häälikuid. Eristada võib struktuuri märgendust, mis puudutab teksti jagamist väiksemateks üksusteks (nt päis ja sisuosa, lõigud, laused, osalaused), ja funktsionaalset märgendust, mis kategoriseerib, üldistab ja iseloomustab teksti sisu (nt tähistab tekstisõnade sõnaliike)[1].

Lisaks sellele, mida märgendada, tuleb korpuse koostamisel mõelda ka sellele, mil moel märgendeid esitada ning kuidas märgendeid lisada. Sellest kõigest siinses peatükis räägimegi. Lisaks anname ülevaate levinumatest korpuste märgenduse liikidest ning olemasolevatest märgendustööriistadest.

[1] Vahel, eriti kvalitatiivses analüüsis, nimetatakse märgendamiseks ka (nt korpusest, andmebaasist, katsete või intervjuude käigus) kogutud andmestike **kodeerimist** ehk konkreetse uurimuse jaoks andmetest tähenduslike elementide või mustrite leidmist ja kategoriseerimist (vt ka Stefanowitsch 2020: 105–139 ja A. Veismanni näidisuurimust selles õpikus). Selles peatükis räägime põhiliselt tervete korpuste ehk tekstikogude märgendamisest, ent teistes peatükkides (nt ptk 6 „Korpusandmete statistiline analüüs“) ja näidisuurimustes katab märgendamise termin ka korpustest kogutud andmestike kodeerimist.

3.1. Märgendamise üldised põhimõtted

3.1.1. Märgendusskeemid ja märgendusvormingud

Märgendamisel kasutatakse kokkuleppelisi **märgendusskeeme** (ingl *annotation scheme*, ka *coding scheme*), mis määravad, millistele tekstiosadele millist infot lisatakse ja milliseid **märgendeid** (ingl *tag* või *label*) selleks kasutatakse. Märgendusskeem on niisiis märgendamisel rakendatavate märgendite loetelu ja nende rakendamise eeskiri. Sõltuvalt sellest, mis on korpuse loomise eesmärk ja millist tekstimaterjali korpus sisaldab, võib kasutada olemasolevaid märgendusskeeme, neid kohandada või luua päris uusi.

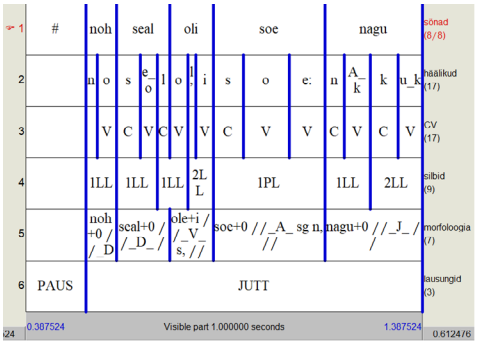
Näiteks tänapäeva eesti kirjakeele tekstides sõnavormide vormiinfo märgendamiseks on juba märgendusskeemid[2] olemas (vt ka ptk 2 „Eesti keele korpused"), neid kasutatakse erinevates eesti keele korpustes ning enamasti ei ole vajadust kasutada uues kirjakeele korpuses neist erinevat sõnaliikide jaotust või luua uusi sõnaliike ja grammatilisi kategooriaid. Samad märgendusskeemid ei pruugi aga sobida tänapäeva kirjakeelest erinevate keelekujude jaoks (suulise keele, murdekeele või vanema kirjutatud keele jaoks). Näiteks ei ole kirjakeele märgendusskeemis eraldi suhtluspartiklite kategooriat (*küll*, *siis*, *noh*, *nagu*), mis aga suulises suhtluses mängib väga olulist rolli. Osaliselt seetõttu on Eesti murrete korpuses kasutusel oma morfoloogilise info märgendamise skeem[3] (vt Lindström, Todesk & Pilvik 2022). Võib tekkida ka vajadus märkida tekstides tunnuseid, mille jaoks veel mingit üldkasutatavat märgendusskeemi olemas ei ole, nt digitaalsetes suhtluskanalites levinud lühendeid (*lol*, *ns*, *mdea*), referentsiaalsust ja viitamisahelaid (nt kellele või millele viitavad asesõnad), žeste jm-d.

Märgendusskeem mõjutab omakorda seda, milline peab olema **märgendusvorming** ehk kasutatud märgendite esitusviis: märgendusvorming peab võimaldama esitada märgendusskeemi elemente masinloetaval kujul. Märgenduse võib sealjuures lisada kas eraldi **märgenduskihtidele** (ingl *annotation layer*), mis hoitakse tekstist endast eraldi ning mis kindlate viiteseoste kaudu on seostatavad toortekstiga (kõnevoorude või sõnade kihiga), või **teksti sisse** (ingl *in-text annotation*), mispuhul lingvistiline vm info lisatakse jooksva algteksti vahele (joonis 3.1). Sama märgendusvorming võib võimaldada vahel mõlemat.

Pikalt on tekstikogude märgendamisvormingute lipulaevaks olnud XML (ingl *extensible markup language*) ning selle erinevad rakendused (näiteks TEI). XML on üldisem hierarhiline märgendusvorming, mille abil võib märgendada lisaks tekstidele ka igasugu muud erinevat infot. See koosneb noolsulgudega tähistatud märgenditest ja nende **atribuutidest**, mis võimaldavad täpsustada märgendatud elemendi mingeid omadusi. Joonisel 3.1 esitatud murdekorpuse näites on kõik

[2] https://cl.ut.ee/ressursid/morfo-systeemid/

[3] https://datadoi.ee/bitstream/handle/33/492/murdekorpus.html?sequence=15&isAllowed=y



Joonis 3.1. Näide eesti keele spontaanse kõne foneetilise korpuse märgenduskihtidest (vasakul, esitab TextGrid-vormingus faili) ja Eesti murrete korpuse tekstisisesest XML-vormingus märgendusest (paremal)

laused (tegelikult küll kõnevoorud) märgendite *<lause>* ja *</lause>* vahel. Lausete atribuudid täpsustavad märgendatud lausetes lause järjekorranumbri (*id="l1"*), kõneleja koodi (*koneleja="KJ"*) ning seose helifaili segmentidega (*helipos="3"*). Kõik lausetes olevad sõnad on omakorda märgendite *<sone>* ja *</sone>* vahel ning nende atribuudid täpsustavad lisaks sõna järjekorranumbrile ka tekstisõna lemma, sõnaliigi ja grammatilise vormi infot. Sõnadena on märgendatud murdekorpuses ka pauside, välja kuulmata kohtade ja kokkuhäälduse märgid, kommentaarid ning intervjueerija tekst. Sel juhul on tekstisõna atribuutide hulgas atribuut *meta*, mis täpsustab, mis tüüpi mittetekstilise infoga on tegu (nt „vahemärk", „intervjueerija", „kommentaar").

TEI (ingl *text encoding initiative*) on põhiliselt humanitaar- ja sotsiaalteadustes kasutatav märgendusskeemi standard, mis põhineb XML-i vormingul ning defineerib spetsiifiliselt teksti märgendamiseks vajalikud elemendid. Sealjuures on TEI märgendusskeemis terve hulk atribuute, mida saab kasutada kõikides TEI-dokumentides (nt *abbr* lühenditele või *name* pärisnimedele), ent on võimalus luua juurde ka uusi. See võimaldab eri keelte ja korpuste tekste analüüsida sarnaste tööriistadega või kuvada samadel platvormidel. Nii XML-i vormingut kui ka TEI märgendusskeemi põhimõtteid on kasutatud näiteks Eesti keele koondkorpuses[4] (joonis 3.2).

XML-vormingu põhist märgendust võib hoida samuti nii teksti sees kui ka eraldi kihtidel. Ülal toodud näited murdekorpusest ja koondkorpusest kasutavad tekstisisest märgendust. ELAN-is transkribeeritud ja märgendatud tekstid on aga vaikimisi XML-vormingu eritüüpi esindavas EAF-vormingus, milles eri tüüpi märgendusinfo on eraldi kihtidel. Korpuses kasutatav märgendusvorming võib sõltuda seega ka kasutatud tarkvarast.

[4] https://cl.ut.ee/korpused/segakorpus/

```
<teiheader>
   <filedesc>
      <titleStmt>
         <title>Capaneusi Harta</title>
         <principal>Kadri Muischnek</principal>
         <respStmt>
      <resp>teisendas SGML-kujule ja lausestas</resp>
         <name>Kristel Uiboaed</name>
         </respStmt>
      </titleStmt>
         <extent> baite: 217846; s&otilde;nu: 22821</extent>
      <publicationStmt>
      <authority>T&Uuml; arvutuslingvistika uurimisgrupp</authority>
         <pubPlace>Tartu, Liivi 2-307</pubPlace>
      <date>september 2007</date>
      <availability>
         <p>vaba kasutamiseks mittekommertsiaalsetel eesm&auml;rkidel</p>
      </availability>
       </publicationStmt>
       <sourcedesc>
         <bibl>
            <title level=m>Capaneusi Harta</title>
       </sourcedesc>
   </filedesc>
```

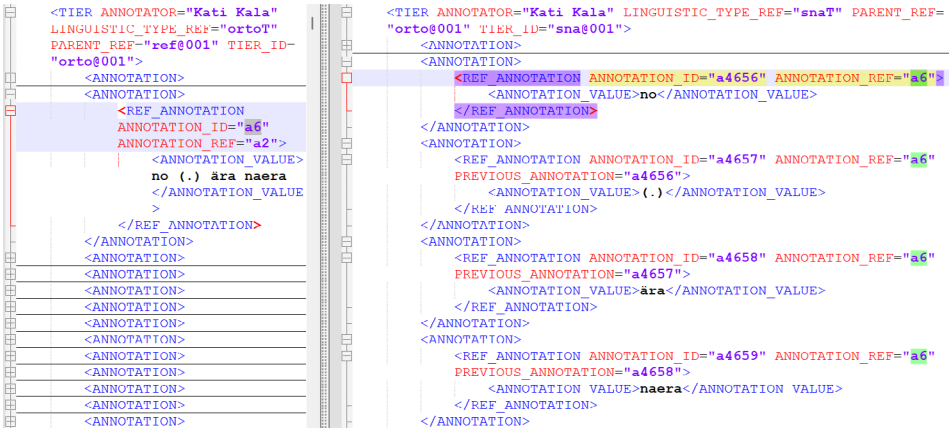
Joonis 3.2. Näide koondkorpuse ilukirjanduse alamkorpuse TEI-faili päisest

```
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION_DOCUMENT AUTHOR="" DATE="2024-02-06T12:42:24+02:00"
    FORMAT="3.0" VERSION="3.0"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocation=
    "http://www.mpi.nl/tools/elan/EAFv3.0.xsd">
    <HEADER MEDIA_FILE="" TIME_UNITS="milliseconds">
    <TIME_ORDER>
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="refT" TIER_ID="ref@001">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="ortoT" PARENT_REF="ref@001" TIER_ID="orto@001">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="orto@001" TIER_ID="sna@001">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@001" TIER_ID="lemma@001">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@001" TIER_ID="POS@001">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@001" TIER_ID="form@001">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@001" TIER_ID="ending@001">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@001" TIER_ID="clitic@001">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="keelT" PARENT_REF="ref@001" TIER_ID="keel@001">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="kommentaarT" TIER_ID="kommentaar@001">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="refT" TIER_ID="ref@002">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="ortoT" PARENT_REF="ref@002" TIER_ID="orto@002">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="orto@002" TIER_ID="sna@002">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@002" TIER_ID="lemma@002">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@002" TIER_ID="POS@002">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@002" TIER_ID="form@002">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@002" TIER_ID="ending@002">
    <TIER ANNOTATOR="" LINGUISTIC_TYPE_REF="snaT" PARENT_REF="sna@002" TIER_ID="clitic@002">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="keelT" PARENT_REF="ref@002" TIER_ID="keel@002">
    <TIER ANNOTATOR="Kati Kala" LINGUISTIC_TYPE_REF="kommentaarT" TIER_ID="kommentaar@002">
</ANNOTATION_DOCUMENT>
```

Joonis 3.3. Näide teismeliste keele korpuse EAF-vormingus faili struktuurist

Joonisel 3.3 on (lihtsustatud) näide Eesti teismeliste keele korpuse EAF-vormingus suulise suhtluse transkriptsioonist koos erinevate märgenduskihtidega. Eraldi kihtidel on mõlema kõneleja (siinses vestluses koodidega 001 ja 002) aegjoondusega kõnesegmendid (*ref*), tavaortograafias transkriptsioon kõnevoorude kaupa (*orto*), kõnevoorude üksiksõnad (*sna*), sõnade lemmad (*lemma*), sõnaliigid

(*POS*), grammatilised kategooriad (*form*), pöörde- ja käändelõpud (*ending*), kliitikud (*clitic*), muudes keeltes kui eesti (üld)keeles esinevad segmendid (*keel*) ning transkribeerija vabatekstilised kommentaarid (*kommentaar*). Eri kihid on omavahel seotud hierarhiliste pärilussuhete kaudu: *orto*-kiht on seotud aegjoondatud *ref*-kihiga, *sna*-kiht on seotud *orto*-kihiga (kindlad sõnad kuuluvad kindlasse *orto*-kihi kõnevooru), *lemma*-kiht on seotud *sna*-kihiga (igal sõnal on kindel lemma) jne. Kihtidel paiknevad omakorda konkreetsed märgendatud üksused, mis on samuti kindlate viidete kaudu seotud teiste kihtide märgendatud üksustega. Näiteks joonisel 3.4 on esitatud vasakul kõneleja (koodiga 001) *orto*-kihil paiknev kõnevoor *no (.) ära naera*, mille ID (atribuut *ANNOTATION_ID*) on „a6“. Selles esinevad üksikud sõnad (*no*, *ära*, *naera*) ja paus (*(.)*) on *sna*-kihil seotud sama vooruga *ANNOTATION_REF* atribuudi kaudu, mille väärtus on samuti „a6“ (joonisel paremal). (Vt ka J. Wilburi näidisuurimust korpuslingvistika ja ohustatud keelte kohta.)



Joonis 3.4. EAF-failis märgendatud kõnevoor (vasakul) ja selles esinevad sõnad (paremal)

Tänapäeval on XML-i põhiseid vorminguid hakanud tasapisi asendama muud vormingud, nt JSON (ingl *JavaScript object notation*), mis talletab infot arvutuslikult kompaktsemalt ning võimaldab seeläbi vähendada nii korpusefailide mahtu kui ka nende automaatseks analüüsimiseks kuluvat aega.

Joonisel 3.5 on näide ERR-i raadiosaadete korpusest, saatest „Päevakaja“ (aastal 2022). Automaatse kõnetuvastuse käigus salvestisest tuvastatud tekst on jagatud sektsioonidesse (*sections*), see kõnevoorudesse (*turns*) ning kõnevoorud on jagatud sõnadeks (*words*). Sõnavorme on omakorda märgendatud automaatselt EstNLTK morfoloogilise märgenduse tööriistadega, mille põhjal on määratud iga sõnavormi lemma (*lemma*), tüvi (*root*), pöördelõpp (*ending*), kliitik (*clitic*), sõnaliik (*postag*) ja grammatiline kategooria (*form*), kui tegemist on käänd- või pöördsõnaga. Samuti on sõnad automaatselt silbitatud (*syllables*).

```
"sections": [
  {
    "start": 11.96,
    "end": 30.64,
    "type": "speech",
    "turns": [
      {
        "start": 12.34,
        "end": 20.86,
        "speaker": "S0",
        "transcript": "Tervist on esimene jaanuar, aasta on 2022, kell sai kuus ning eetris on
        Päevakaja. Minu nimi on Johannes Voltri.",
        "words": [
          {
            "start": 12.34,
            "end": 13.06,
            "word": "tervist",
            "punctuation": "",
            "word_with_punctuation": "Tervist",
            "is_multiword": false,
            "analyses": [
              {
                "lemma": "tervist",
                "root": "tervist",
                "ending": "0",
                "clitic": "",
                "postag": "I",
                "form": ""
              }
            ],
            "syllables": [
              {
                "syllable": "ter",
                "quantity": 2,
```


Joonis 3.5. Näide ERR-i raadiosaadete korpuse JSON-vormingus failist

3.1.2. Automaatne või käsitsi märgendamine

See, mida korpuses märgendatakse, sõltub niisiis lisaks korpuse koostamise eesmärkidele praktikas paljuski ka sellest, kuidas märgendatakse: kas ja mis tööriistu selleks kasutatakse. Märgendamine võib toimuda automaatselt, käsitsi või mõlemat kombineerides, n-ö poolautomaatselt.

Automaatse märgenduse jaoks peavad olema loodud analüüsivahendid, mis tunnevad uues etteantud tekstis ära kindlad elemendid ning oskavad määrata neile märgenduse ilma, et inimene peaks protsessi oluliselt sekkuma.

Käsitsi märgendamise puhul võivad samuti olla kasutusel arvutiprogrammid, mis märgendamistööd hõlbustavad, ent otsuseid selle kohta, milliseid kategooriaid igale tekstiüksusele määrata, peab tegema algoritmi asemel inimene ise. See on nii ajalises kui ka rahalises vaates ilmselgelt oluliselt kulukam, ent annab samas enamasti ka täpsema ja nüansseerituma tulemuse. Käsitsi märgendamise oluliseks kitsaskohaks võib olla aga ebajärjekindlus, eriti eri märgendajate vahel, aga ka ühe märgendaja märgendatud tekstides.

Poolautomaatse märgendamise puhul parandab ja täiendab inimene eelnevalt automaatselt märgendatud teksti ning see võib hõlmata ka olemasolevate märgendusskeemide või -tööriistade kohandamist konkreetse korpuse materjalidele. Poolautomaatne märgendamine hõlmab tavaliselt ka **ühestamist** ehk keeleüksuse

mitme võimaliku analüüsi hulgast kontekstist lähtuvalt õige tõlgenduse väljavalimist ja vastuolude kõrvaldamist.

Suur hulk tänapäevaseid korpusi on juba nii suured, et nende käsitsi märgendamine ei oleks mõistlik ega vahel võimalikki ning märgendamiseks tuleb kasutada automaatanalüüsi vahendeid, leppides sealjuures sellega, et selle tulemusena lisatud märgenduses võib esineda ka vigu. Siiski hõlmab ka automaatse märgendamise tööriistade loomine sageli esmalt hulga korpuse tekstide käsitsi märgendamist ehk nn **treeningandmestiku** loomist. Selle treeningandmestiku pealt saab tarkvara õppida hindama tõenäosusi, millega teatud järjendid ja nende märgendid kindlas kontekstis koos esinevad. Käsitsi märgendatud treeningandmeid võib kasutada ka selleks, et testida, kui hästi mingi märgendus- või klassifitseerimistööriist oma tööga hakkama saab (vt alapeatükk 3.1.3).

Nagu 2. peatükis räägitud, märgendatakse ka erimärgendusega või tänapäeva kirjakeelest erinevat keelt sisaldavaid erikorpusi üldjuhul käsitsi, mistõttu on need korpused ka mahult väiksemad. Väiksemate või vähem uuritud keelte puhul on käsitsi või paremal juhul poolautomaatne märgendamine paratamatus, kuna nende struktuur ja grammatilised kategooriad erinevad oluliselt suurte keelte jaoks välja töötatud märgendusskeemidest, mida automaatse keeleanalüüsi tööriistades rakendatakse.

Kui eesmärgiks on koostada võimalikult kvaliteetne keeleressurss, on inimeksperdi sekkumine tihtipeale vajalik ka n-ö mittestandardse keele (nt suulise keele, murrete või veebitekstide) märgendamisel, kuna loomuliku keele automaatanalüüsi tööriistad, mis on üldjuhul treenitud kirjakeelsetel tekstidel, ei oska tingimata arvestada kirjakeele normist hälbiva keele eripäradega (nt kirjavahemärkide puudumise või ebaühtlase kasutuse, tühikute puudumise, sõnajärje muutuste, väljajättude, ortograafiliselt või ka morfoloogiliselt mittenormingupäraste sõnavormide jpm-ga). Käsitsi märgendamist tuleb sageli ette ka õppijakorpuste puhul, kus keeleõppijate loodud tekstides märgendatakse eri tüüpi vigu, selleks et mõista paremini keeleõppeprotsessi ja töötada vastavalt õppijate lähte- ja sihtkeelele välja efektiivsemaid õppemetoodikaid ja -vahendeid. Ehkki teatud tüüpi vigu, nt õigekirjavigu, on võimalik küllalt hästi tuvastada ka automaatsete tööriistadega, nt õigekirjakorrektoriga, tuleb vigade detailsemaks liigitamiseks ja muu huvipakkuva info märgendamiseks kasutada inimeksperdi abi.

3.1.3. Märgenduse täpsuse hindamine

Ehkki praktikas mitte kuigi sageli rakendatud, on heaks tavaks anda märgendatud korpuse dokumentatsioonis (vt ptk 4.6 „Metaandmed ja dokumentatsioon“) ülevaade ka märgenduse hinnangulisest täpsusest. Nagu öeldud, tuleb nii automaatse kui ka käsitsi märgendamise puhul ette vigu ja ebaühtlust, mis mõjutavad mõnevõrra koostatava korpuse usaldusväärsust. Vead võivad sealjuures tuleneda sellest, et automaatsed tööriistad ei ole sobiva keelematerjali põhjal treenitud või

et mingid märgendusklassid esinevad tekstides liiga harva; ka sellest, et käsitsi märgendamisel inimene väsib ning tähelepanu võib hajuda või sellest, et mingeid nähtusi ongi keeruline üheselt kategoriseerida.

Käsitsi märgendatud korpuste puhul võib anda teatud väikese osa tekstidest märgendamiseks korraga mitmele märgendajale ning hinnata seejärel **märgendajatevahelist kooskõla**. Kooskõlamäära mõõdikud peaksid peegeldama, kui hästi eri märgendajate otsused kokku langevad, ning tooma välja probleemseid kohti nii märgendusskeemis ja märgendusjuhistes kui ka keelematerjalis endas. Kooskõla hindamise statistilisi mõõdikuid on mitmeid (vt nt Cohen 1960; Scott 1955). K. Muischneki ja S. Orasmaa näidisuurimuses nimeüksuste märgendamisest on Hripcsaki ja Rothschildi (2005) eeskujul kasutatud **F1-skoori**, mis toetub nn täpsuse ja saagise näitajatele (ingl vastavalt *precision* ja *recall*). Märgendajatevahelise kooskõla mõõtmisel väljendab **täpsus** seda esimese märgendaja märgenduste hulka, mis langesid kokku teise märgendaja omadega; **saagis** vastupidi teise märgendaja märgenduste hulka, mis langesid kokku esimese omadega.

Automaatse märgendamise täpsuse hindamiseks märgendatakse samuti väike osa korpusest inimeksperdi poolt käsitsi. Ehkki ka käsitsi märgendaja võib teha vigu, käsitletakse seda väikest käsitsi märgendatud osa nn **kuldstandardina**, millega automaatse märgendustööriista tulemusi võrrelda. Kõige lihtsam ja tavalisem täpsuse mõõdik on automaatse märgenduse puhul **korrektsus** (ingl *accuracy*), mis näitab õigesti analüüsitud üksuste (nt sõnavormide) osakaalu kõikidest üksustest. Ka täpsus ja saagis on automaatmärgenduse hindamiseks sagedasti kasutatud mõõdikud, kui võrreldakse mitut erinevat märgenduskategooriat. Sellises kontekstis näitab täpsus, kui suur osakaal mingist automaatmärgendaja märgendatud kategooriast on märgendatud õigesti (nt kui suur osakaal automaatse analüsaatori nimisõnadeks märgendatud üksustest on märgendatud nimisõnadeks ka kuldstandardis); saagis jällegi näitab, kui suure osa mingist märgendatavast kategooriast suudab automaatmärgendaja õigesti tuvastada (nt kui suure osakaalu kuldstandardis nimisõnadeks märgendatud sõnadest on automaatne analüsaator nimisõnadeks märgendanud)[5].

3.2. Märgenduse liigid

Selles alapeatükis räägime lähemalt kõige levinumatest korpuse (funktsionaalse) märgenduse liikidest. Kaks põhilist liiki, mida kasutavad ka enamik keeletehnoloogilisi rakendusi, on **morfoloogiline** ja **süntaktiline märgendamine**. Vähem levinud on **semantiline märgendamine**. Peatükis räägime ka nimeüksuste märgendamisest ning suulise keele ja multimodaalsete korpuste märgendamisest.

[5] Vt ka P. Tinitsa näidisuurimust korpuste kasutamisest digihumanitaarias, kus täpsust, saagist ja F1-skoori on kasutatud korpuspäringu tulemuste hindamiseks.

Viimased võivad sisaldada lisaks morfoloogilisele, süntaktilisele, semantilisele ja nimeüksuste märgendusele infot ka kõnet ja multimodaalset suhtlust iseloomustavate spetsiifilisemate üksuste kohta (nt häälikud, häälelaad, žestid). Siin käsitletud märgenduse liigid ei kirjelda aga kindlasti kõiki võimalikke viise korpuse tekstide rikastamiseks. Nõnda võib märgendada ka teksti meelsust, teemasid, kasutatud keeli, pragmaatikat (nt viisakus, kõneaktid) jne.

3.2.1. Morfoloogiline märgendamine

Morfoloogilise märgendamise käigus saab iga tekstisõna info enamasti vähemalt selle sõna **sõnaliigi** (nt nimisõna, omadussõna, tegusõna) ja **lemma** kohta, kuigi neid protsesse võib käsitleda ja rakendada ka eraldi (ingl vastavalt *POS-tagging* ja *lemmatization*). Muutumatute sõnade puhul (eesti keeles nt sidesõnad, määrsõnad, kaassõnad) kattub lemma ka tekstis tegelikult esinevate sõnavormidega (nt vormi *aga* lemma on *aga*); muutuvate sõnade (käänd- ja pöördsõnade) puhul on lemma aga sõnavormi muutetunnustest puhastatud lekseem, mida kohtame sõnaraamatutes märksõnana. Näiteks sõnavormi *hobuste* lemma on *hobune* ja sõnaliik nimisõna, sõnavormi *tulnuksime* lemma on *tulema* ja sõnaliik tegusõna.

Nii sõnaliigi määramine kui ka **lemmatiseerimine** ning tegelikult pea kõik siin peatükis käsitletavad märgendusliigid eeldavad, et korpuse tekstijadad on eelnevalt **sõnestatud** (ingl *tokenization*) ehk jagatud sobivateks terviklikeks üksusteks, **sõnedeks** (ingl *token*). Tüüpiliselt kattuvad need üksused sõnadega, aga sõned võivad olla ka kirjavahemärgid, kirjavahemärkide jadad, nt *:-)*, või kombinatsioonid kirjavahemärkidest, numbritest ja sõnadest, nt *100%, m/s, v.a*, *07.06.2023*.

Olenevalt keelest ja sõnestamise tööriistast võivad sõned olla ka mitmest sõnast koosnevad üksused, mis moodustavad tähendusliku terviku (nt ingl *ice cream* või pärisnimi *New York*). Nõnda ei ole sõnestaminegi üheselt määratletav ja kindlate reeglitega protsess, vaid põhineb keelespetsiifilistel konventsioonidel ja kokkulepetel ning võiks arvestada ka konkreetse analüüsi vajadusi ja eesmärke. Märkusena olgu öeldud, et ingliskeelse terminoloogia eeskujul viidatakse sõnaga *sõne* väljaspool keeletehnoloogiat ja arvutilingvistikat sageli ka mingi tekstiüksuse (nt sõnavorm, lemma või isegi konstruktsioon) üksikule esinemisjuhule tekstis, vastandades seda selle üksuse klassi ehk tüübi (ingl *type*) esinemisele (vt M.-L. Pilviku näidisuurimust tuletuskonstruktsioonide produktiivsusest ja ptk 5.2.4.2 „Sõnavara hajuvus ja levik").

Lemmatiseerimine võimaldab uurijal hõlpsamalt leida teda huvitavat infot, eriti keeltes, millel on rikas ja kompleksne morfoloogiline süsteem ning kus ühel sõnal võib seetõttu olla kümneid või koguni sadu vorme. Nõnda võib kasutaja otsida kõiki mingi lekseemi sõnavorme korraga (nt *saime, saaksite, saavat, saada, saagu*) või leida üles sõnad konkreetsetes kontekstides (nt nimisõnad, millele eelneb vahetult omadussõna: *kuri koer*, *suur maja*). Sõnaliikide määramine omakorda võib aidata eristada päringus homonüümseid vorme (nt *või* (tegusõna), *või* (nimisõna) või *või* (sidesõna)). Nagu sõnestamisel võib ka lemmatiseerimise

ja sõnaliikide määramise puhul märgendada lisaks sõnadele ka pikemaid sõnajärjendeid, nt idioome või mitmesõnalisi ühendeid.

Lisaks lemmale ja sõnaliigile võivad olla määratud ka täpsemad kategooriad, eesti keeles on eriti olulised **morfoloogilised kategooriad**, nt arvu ja käände kategooria käändsõnadel ning aja, tegumoe, kõneviisi, arvu ja isiku kategooria pöördsõnadel. Nõnda on võimalik korpuspäringuid veelgi kitsendada. Näiteks kui meid huvitavad aluse funktsioonis esinevad arvsõnast peasõnaga hulgafraasid, saame otsida järjendeid, kus ainsuse nimetavas käändes arvsõnale järgneb ainsuse osastavas käändes nimisõna (*kolm koera* (*haukus/haukusid*), (*kogunes/kogunesid*) *üheksa meest*); kui meid huvitab konstruktsioon *pidi/peaks tehtama* (vt ptk 5.2.3 „Konkordantside koostamine grammatilise info põhjal"), saame otsida umbisikulises tegumoes tegusõna *ma*-infinitiivi, millele eelneva või järgneva kolme sõna hulgas oleks tegusõna *pidama* kindla kõneviisi lihtmineviku (*pidi*) või tingiva kõneviisi oleviku vorm (*peaks*).

Eesti keel on morfoloogiliselt väga rikas, võimaldades märgendada ohtralt kategooriaid ja nende kombinatsioone. Ühes korpuses kasutatud märgendid ei esinda aga ühtainsat ja objektiivselt ainuõiget kategoriseerimisvõimalust, vaid sõltuvad eeskätt **kokkulepetest** märgendamisskeemis ja grammatikakirjeldustes üldisemalt. Seetõttu tuleks mis tahes korpuse koostamisel anda võimalikult põhjalik ülevaade korpuse märgendamise tavadest ja kokkulepetest (vt ptk 4 „Oma korpuse loomine") ning korpuse kasutamisel nende kokkulepetega tutvuda.

Selles, et üks ja sama sõnavorm võib kontekstist olenevalt kuuluda mitmesse erinevasse sõnaliiki (nt *hullus seisus* – omadussõna, *see on täielik hullus* – nimisõna, *inimmass hullus rahutustel* – tegusõna), pole midagi ebatavalist, ent vahel võib sõna mitmeti tõlgendada ka ühes ja samas kontekstis. Üks sage otsustuskoht puudutab näiteks sõnu, mis on grammatiseerumise või leksikaliseerumise protsessi eri faasides. Nii võiks vorme *osas* (*hinnete osas*) ja *ajal* (*etenduse ajal*) analüüsida kas kaassõnadeks või nimisõna käändevormideks; samamoodi on kohati üsna ebaselged piirid nimisõnade ja määrsõnade (nt *kukkus maha, on abielus*), tegusõnade ja omadussõnade (nt *katkenud side*, *kaetud laud*), määrsõnade ja partiklite (nt *siis*, *küll*, *alles*) või ka nimi- ja omadussõnade vahel (nt *hull*, *vana*). Leksikaliseerunud üksused, nagu *parata*, *peksa* (*andma*) või *plehku* (*panema*), samuti genitiivatribuudid (nt *eesti keel*) võivad omakorda olla korpuses märgendatud kas päris omaette sõnaklassidena või olla määratud mõne üldisema sõnaklassi alla. Sõnaliikide piirialal olevate nähtuste kohta saab lähemalt lugeda teosest „Eesti keele sõnamuutmine" (Viht & Habicht 2019).

Kohati võib keeruline olla ka grammatiliste kategooriate määramine, eriti mittestandardsete või ajalooliste keelevariantide puhul, mis ei järgi tingimata tänapäeva kirjutatud standardkeele reegleid (näiteks ühildumisel ja rektsioonisuhetes). Nõnda esineb näiteks vana kirjakeele korpuse[6] tekstides sageli märgend *sg.gen.part.*, mis

[6] https://vakk.ut.ee/

väljendab seda, et arvestades perioodi ja teksti kirjutaja üldist keelekasutust ei ole sõnavormi ega konteksti põhjal võimalik täie kindlusega otsustada, kas tegemist on ainsuse omastava (*gen.*) või osastava (*part.*) vormiga (nt Georg Mülleri jutlusest 28.12.1600: *se sama piddab sẽ sinatze Lapsukeße iures armu ninck röhmu leüdma*).

3.2.2. Süntaktiline märgendamine

Süntaktiline märgendamine esitab lause struktuuri ja/või lauset moodustavate sõnade süntaktilisi funktsioone ehk väljendab, millist rolli sõna või lauseosa täidab lause struktuuris ja kuidas see suhestub teiste lauseosadega. See eeldab üldiselt, et eelnevalt on märgendatud vähemalt sõnaliigid, eesti keele puhul ka muu morfoloogiline info.

Süntaktiline märgendamine jaguneb laias laastus kolmeks: pindsüntaktiline märgendamine, fraasistruktuuri märgendamine ja sõltuvusstruktuuri märgendamine.

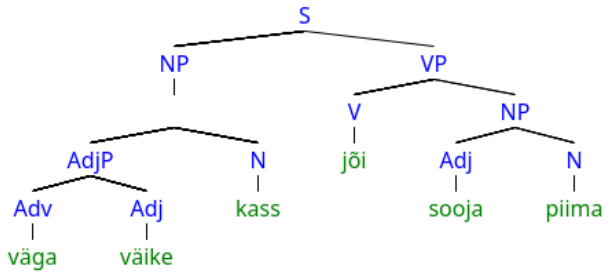
Pindsüntaktilise märgendamise puhul lisatakse igale tekstisõnale tema süntaktilise funktsiooni märgend (vt joonis 3.6), kuid ei moodustata lause struktuuri kujutavat nn süntaksipuud ehk hargmikku, mis näitaks, kuidas lauseosad on omavahel seotud ja kuidas need moodustavad lause ülesehituse.

<Väga> väga L0 D cap **@ADVL**
<väike> väike L0 A pos sg nom **@AN>**
<kass> kass L0 S com sg nom **@SUBJ**
<jõi> jooma Lb V main indic impf ps3 sg ps af **@FMV**
<sooja> soe L0 A pos sg part **@AN>**
<piima> piim L0 S com sg part **@OBJ**
<.> . Z Fst CLB #7->7

Joonis 3.6. Lause *Väga väike kass jõi sooja piima* pindsüntaktiline märgendus koos morfoloogilise märgendusega, eesti keele piirangute grammatika (Müürisep 2000) vorming

Nii fraasistruktuuri kui ka sõltuvusstruktuuri märgendamisel moodustatakse iga lause jaoks hargmik ehk **süntaksipuu**. Seetõttu nimetatakse selliselt märgendatud korpusi **puudepankadeks** (ingl *treebanks*).

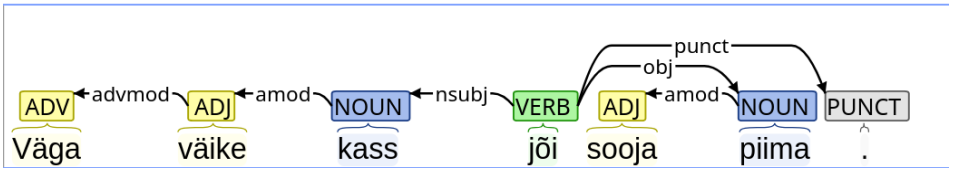
Fraasistruktuuri märgendamine põhineb fraasistruktuuripõhistel süntaksiteooriatel, mille põhiideeks on ettekujutus lausest kui hierarhilisest struktuurist, mille sõlmedeks on eri tüüpi fraasid (nt verbifraas, nimisõnafraas, omadussõnafraas). Lause fraasistruktuuri märgendamisel näidataksegi ära, kuidas sõnad fraase moodustavad ja kuidas neist fraasidest moodustub lause, moodustatakse nn **fraasipuu** (joonis 3.7).



Joonis 3.7. Lause *Väga väike kass jõi sooja piima* fraasistruktuuripuu

Sellest, mis on fraas ja kuidas seda määratakse, saab lähemalt lugeda näiteks koguteosest „Eesti keele süntaks" (Erelt & Metslang 2017: 54–56). Keeleteaduses on palju erinevaid fraasistruktuuripõhiseid süntaksiteooriaid (nt ingl *head-driven phrase structure grammar,* HPSG, või *lexical-functional grammar*, LFG). Puudepanga märgendus võib põhineda mingil kindlal süntaksiteoorial (nt LinGO Redwoods Treebank[7] põhineb HPSG-l), aga võib esitada ka nn baas-fraasistruktuuri (nt Penn Treebank[8]).

Sõltuvussüntaktilisel märgendamisel näidatakse samuti ära lauset moodustavate sõnade vahelised suhted ja moodustatakse nn **sõltuvuspuu** (joonis 3.8), ent lausestruktuuri esitamisel ei kasutata vahesõlmi (fraase) ning süntaktilised suhted on fraaside asemel tekstisõnade vahel. Need süntaktilised suhted on ebasümmeetrilised sõltuvussuhted: üks sõna on peasõna, teine laiend (ka ülemus-alluv; ingl *head-dependent*, *governor-dependent*), sealjuures võib ühel sõnal olla mitu alluvat/laiendit, aga ainult üks ülemus/pea. Sõnadevahelisel suhtel võib olla nimi, mis näitab laiendi süntaktilist funktsiooni. Üldiselt arvatakse, et sõltuvusesitus sobib vaba sõnajärjega keelte (nagu eesti keel) jaoks paremini kui fraasiesitus (Jurafsky & Martin 2025: ptk 18).



Joonis 3.8. Lause *Väga väike kass jõi sooja piima* sõltuvuspuu

[7] https://delph-in.github.io/docs/garage/RedwoodsTop/

[8] https://catalog.ldc.upenn.edu/LDC99T42

Keeleteaduses on olemas ka erinevaid sõltuvuspõhiseid süntaksiteooriaid, kuid sõltuvuspuude pankades kasutatakse tavaliselt üldistatud sõltuvussüntaktilist lähenemist.

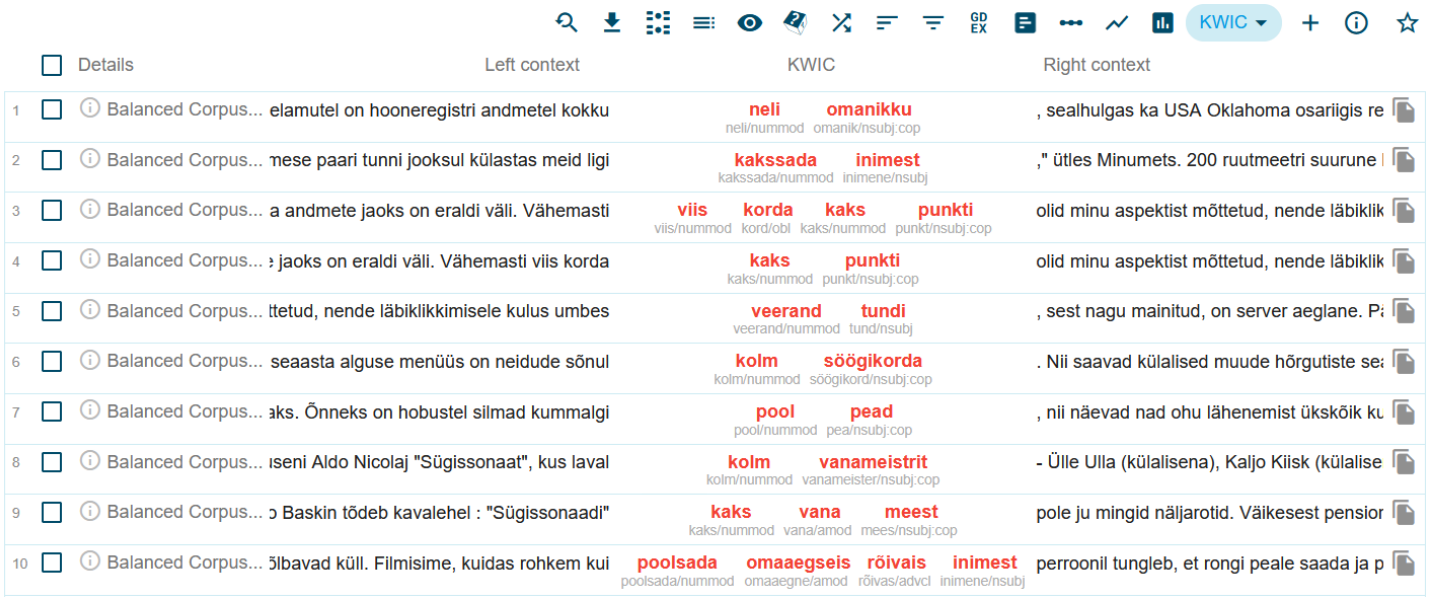
Süntaktiliselt märgendatud korpuste arv ja keelte arv, mille jaoks on olemas süntaktiliselt märgendatud korpus, on viimase 10–15 aasta jooksul kiiresti kasvanud. Korpuste märgendamisel domineeris ajalooliselt fraasistruktuuriesitus, kuid viimasel ajal on arendatud just sõltuvuspuude panku, eelkõige UD (ingl *universal dependencies*)[9] (De Marneffe jt 2021) sõltuvuspuude pankade kollektsiooni raames. UD puudepankade kollektsioon sisaldas 2025. aastal umbes 340 puudepanka rohkem kui 180 keele jaoks. Eesti keele puudepanku on UD kollektsioonis kaks: standardkirjakeelt esindav EDT (Estonian Dependency Treebank) sisaldas 2025. aastal 440 000 sõna ja veebitekstide puudepank EWT (Estonian Web Treebank) u 90 000 sõna. Neid puudepanku saab UD kodulehelt alla laadida ja neile saab esitada päringuid kasutajaliideste kaudu, mille viited leiab samuti UD kodulehelt.

Puudepankadest saab infot sagedaste süntaktiliste mustrite kohta, nt milliseid sihilisi verbe kasutatakse harva ilma sihitiseta; milliseid verbe kasutatakse harva alusega jne. Kõik UD puudepangad on märgendatud samu märgendeid ja märgendusskeeme kasutades ja nii on see sobiv materjal keeletüpoloogiliste uurimuste jaoks, vaatamata sellele, et UD esmane eesmärk on keelest sõltumatu masinõppepõhise süntaksianalüüsi tarkvara arendamine.

Puudepankade kasvava populaarsuse ja UD eduloo taga on ka arvutilingvistika praktilised vajadused: puudepanku kasutatakse masinõppel põhinevate süntaksianalüsaatorite ehk **parserite** õpetamiseks ehk treenimiseks. Selliste parseritega saab automaatselt märgendada suuri korpusi. Inglise keeles kasutatakse selliste automaatselt märgendatud puudepankade jaoks terminit *parsebank*. Nii saavad lähiajal automaatselt süntaktilise struktuuri suhtes märgendatud korpused ilmselt sama levinuks kui praegu on automaatselt morfoloogiliselt märgendatud korpused. Näiteks ka 2021. aasta eesti keele ühendkorpus on UD puudepankade abiga märgendatud, kasutades EstNLTK Stanza mudeleid[10]. Nii saame otsida näiteks aluse funktsioonis arvsõnalisi hulgafraase (*kolm meest (läheb/lähevad)*) mitte ainult sõnaliikide ja morfoloogilise info põhjal, vaid ka süntaktiliste rollide ja fraasi osaliste omavaheliste sõltuvussuhete kaudu: joonisel 3.9 on näidatud Sketch Engine'i konkordantsiotsingu `[features="sg_n"&tag="N"][tag!="[ZV]"]{,3}[syn_rel="nsubj.*"& features="sg_p" & tag="S"]` tulemus. Päringus ütleme, et ainsuse nimetavas käändes arvsõnale peab järgnema subjekti rollis ainsuse osastavas käändes nimisõna, kusjuures nende vahele võib jääda veel kuni kolm sõna (välja arvatud kirjavahemärgid ja tegusõnad).

[9] https://universaldependencies.org/

[10] https://github.com/estnltk/estnltk/tree/main/tutorials/nlp_pipeline/C_syntax



	Details	Left context	KWIC	Right context
1	Balanced Corpus...	elamutel on hooneregistri andmetel kokku	**neli** (neli/nummod) **omanikku** (omanik/nsubj:cop)	, sealhulgas ka USA Oklahoma osariigis re
2	Balanced Corpus...	nese paari tunni jooksul külastas meid ligi	**kakssada** (kakssada/nummod) **inimest** (inimene/nsubj)	," ütles Minumets. 200 ruutmeetri suurune
3	Balanced Corpus...	a andmete jaoks on eraldi väli. Vähemasti	**viis** (viis/nummod) **korda** (kord/obl) **kaks** (kaks/nummod) **punkti** (punkt/nsubj:cop)	olid minu aspektist mõttetud, nende läbiklik
4	Balanced Corpus...	ɨ jaoks on eraldi väli. Vähemasti viis korda	**kaks** (kaks/nummod) **punkti** (punkt/nsubj:cop)	olid minu aspektist mõttetud, nende läbiklik
5	Balanced Corpus...	ttetud, nende läbiklikkimisele kulus umbes	**veerand** (veerand/nummod) **tundi** (tund/nsubj)	, sest nagu mainitud, on server aeglane. P
6	Balanced Corpus...	seaasta alguse menüüs on neidude sõnul	**kolm** (kolm/nummod) **söögikorda** (söögikord/nsubj:cop)	. Nii saavad külalised muude hõrgutiste se
7	Balanced Corpus...	aks. Õnneks on hobustel silmad kummalgi	**pool** (pool/nummod) **pead** (pea/nsubj:cop)	, nii näevad nad ohu lähenemist ükskõik ku
8	Balanced Corpus...	ıseni Aldo Nicolaj "Sügissonaat", kus laval	**kolm** (kolm/nummod) **vanameistrit** (vanameister/nsubj:cop)	- Ülle Ulla (külalisena), Kaljo Kiisk (külalise
9	Balanced Corpus...	ɔ Baskin tõdeb kavalehel : "Sügissonaadi"	**kaks** (kaks/nummod) **vana** (vana/amod) **meest** (mees/nsubj:cop)	pole ju mingid näljarotid. Väikesest pensior
10	Balanced Corpus...	ɔlbavad küll. Filmisime, kuidas rohkem kui	**poolsada** (poolsada/nummod) **omaaegseis** (omaaegne/amod) **rõivais** (rõivas/advcl) **inimest** (inimene/nsubj)	perroonil tungleb, et rongi peale saada ja p

Joonis 3.9. Aluse funktsioonis esinevate hulgafraaside otsing eesti keele ühendkorpusest 2021 (Sketch Engine)

3.2.3. Semantiline märgendamine

Morfoloogilisele ja süntaktilisele märgendamisele võib järgneda semantiline analüüs, näiteks kui on vaja leida kontekstist lähtuvalt sõnavormidele sobiv tähendus (semantiline ühestamine) või selgitada välja lause tähendus. Semantiliselt märgendatud korpus võib tähendada aga ka hoopis muud kui lihtsalt sõnatähenduste osas ühestatud korpust, näiteks võib sellega viidata korpusele, kus on eristatud semantilisi rolle (agent, patsient, stiimul, instrument jne), konkreetseid infoeralduse seisukohast relevantseid tähenduskategooriaid (nt ajaväljendid, nimeüksused, aadressid), ülem- ja alammõisteid, sõnade tähendusklasse või entiteetidevahelisi suhteid. Joonisel 3.10 on näide USAS-i[11] (ingl *UCREL semantic analysis system*) märgendusskeemi põhjal semantiliselt märgendatud lausest, kus iga sõna (või mitmesõnaline üksus) on saanud kindla tähenduskategooria. USAS on algselt arendatud inglise keele analüüsiks, ent sama märgendusskeemi kasutavad märgendajad on praeguseks olemas ka hiina, hollandi, soome, prantsuse, itaalia, portugali, hispaania, indoneesia ja kõmri keele jaoks. Märgendusskeemi moodustavad 21 suuremat valdkonda (nt O – asjade, materjalide, esemetega seotu, X – psühholoogilised seisundid ja protsessid, B – keha ja indiviidiga seotu, E – emotsioonid), mis jagunevad omakorda 232 väiksemaks kategooriaks[12].

```
0000001 002  -----   -----
0000003 010  JJ      Colourless              O4.3
0000003 020  JJ      green                   O4.3 W5 L3 X9.1-
0000003 030  NN2     ideas                   X4.1
0000003 040  VV0     sleep                   B1 H4/N5
0000003 050  RR      furiously               E3- X5.2+
0000003 051  .       .
```

Joonis 3.10. Näide USAS-skeemi semantilise märgenduse väljundist[13]

Kuna paljud semantiliselt märgendatud korpused on koostatud eesmärgiga treenida loomuliku keele töötluse mudeleid erinevatel **infootsingu eesmärkidel**, võib semantiline märgendus olla ka väga spetsiifiline. Üks semantilise märgenduse liike, millel on rakenduslik funktsioon, on nimeüksuste märgendamine. Nimeüksuste märgendamise mõte on luua materjal analüüsimaks, mis liiki nimesid või milliseid isikuid, organisatsioone vm-d korpuses mainitakse, eriti just mingil viisil seostatuna, nt sama sündmuse kirjelduse juures. **Nimeüksuste automaatne tuvastamine**

[11] https://ucrel.lancs.ac.uk/usas/

[12] https://ucrel.lancs.ac.uk/usas/semtags.txt

[13] https://ucrel-api.lancaster.ac.uk/usas/tagger.html

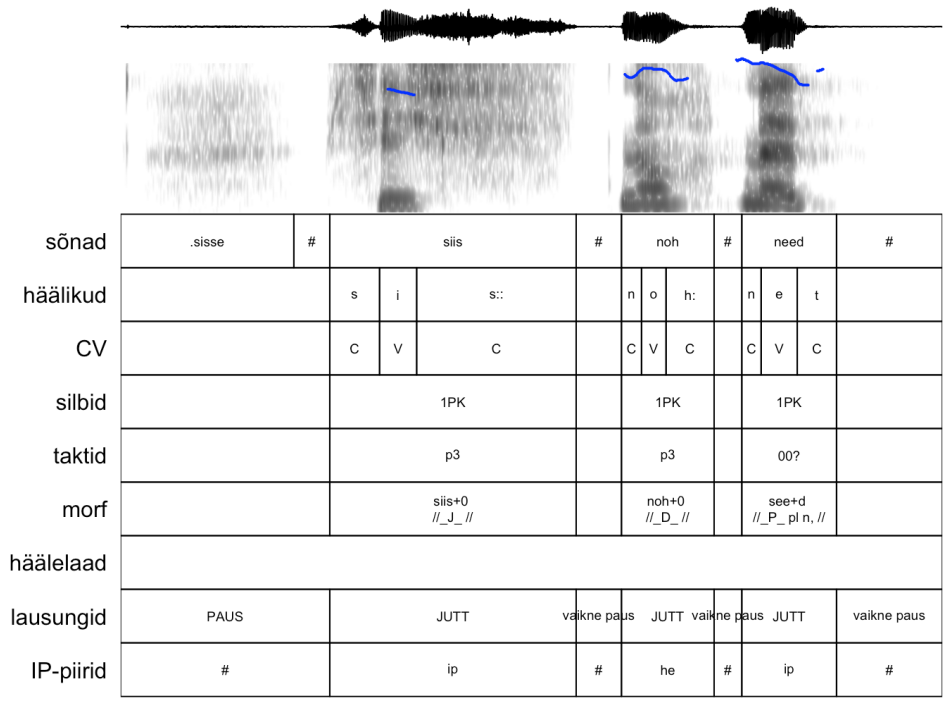
(ingl *named entity recognition*, NER) on oluline teema arvutilingvistikas, vt lähemalt K. Muischneki ja S. Orasmaa näidisuurimust nimeüksuste märgendamisest 19. sajandi vallakohtu protokollides, kus on kirjeldatud ka nimeüksuste käsitsi märgendamise protsessi ja selle ohukohti.

3.2.4. Kõnekorpuste märgendamine

On korpusi, mis kasutavad küll suulise keele materjale, ent märgivad suulisele keelele omaseid nähtusi kas ainult osaliselt või üldse mitte. Eesti murrete korpuse ja TÜ suulise kõne korpuse (vt ptk 2.3.2 „Suulised korpused") transkriptsioonid järgivad keele kirjapanekul küll kõne eripärasid (nt kokkuhäældusi, venitusi, lühendamisi, rõhke, välteid, murdekorpuse soome-ugri foneetilises transkriptsioonis ka häälikute kvaliteeti), ent puuduvad eraldi märgenduskihid prosoodia või muude suulise kõne oluliste omaduste või kategooriate märkimiseks (v.a suhtluspartiklite sõnaliigi kasutamine murdekorpuses). Eesti teismeliste keele korpuse suuliste vestluste transkriptsioonides on märgitud naeru, teatud kõnekeelsusi (nt *suht[suhteliselt]*, *aint[ainult]*), võõrkeelsete sõnade häældust (nt *õu_mai_gaad[_oh_my_god_]*) ning kohati ka eesti üldkeelest lahknevat häældust (nt *ö* häældamine *õ* asemel), ent täpsemat infot näiteks artikulatsiooni, intonatsiooni, hingamise jm kohta samuti märgendatud ei ole. Selliste korpuste koostamise eesmärk on enamasti luua tekstikogu, mis esindaks võimalikult loomulikku spontaanset keelekasutust ning mille abil oleks võimalik uurida ja kirjeldada, mille poolest erineb näiteks suuline suhtlus kirjalikust suhtlusest (sõnavara, sõnajärg, lausete pikkus, normingupäraste ja -vastaste vormide kasutus vms) või mis iseloomustab teatud kitsama keelekasutajate rühma (nt mingi murde kõnelejad, teismelised) loomulikku keelekasutust.

Osa suulise keele korpusi keskenduvad aga just suulisele keelekasutusele spetsiifilise info märgendamisele. Näiteks eesti keele spontaanse kõne foneetilises korpuses on transkriptsioon ise küll tavaortograafias (lisatud on infot häälelaadi, nt kärina, kähina või sosina kohta, samuti on transkriptsioonikihile märgitud pausid ja täidetud pausid, nt *.mm*, *.ee*, naer, matsutused, kõhatused jm-d mittelingvistiliste tasandite nähtused, mis kõneleja kõnega kaasnevad), ent nagu näha jooniselt 3.11, on arvukatel muudel kihtidel märgendatud lisaks ka aegjoondusega **häälikud**, nende kvaliteet ja pikkus, häälikuklassid (konsonandid, vokaalid), **silbid** ja nende tüübid (lühike/pikk, kinnine/lahtine), **rõhutaktid** ja välted, **lausungid** ning **intonatsioonifraasid** (vt märgenduspõhimõtteid lähemalt foneetikakorpuse kodulehelt[14]). Samuti on foneetikakorpuse sõnadele lisatud automaatne morfoloogiline analüüs.

[14] https://foneetikakorpus.ut.ee/ekskfk_margendus.html



Joonis 3.11. Näide foneetikakorpuse faili märgendusest programmis Praat

Sedavõrd põhjalik märgendus võimaldab uurida väga erinevaid suulise kõne ja laiemalt suulise suhtlusega seotud aspekte. Nagu nimigi ütleb, on korpus mõeldud eeskätt foneetiliste nähtuste uurimiseks, näiteks välte kirjeldamiseks kindla silbistruktuuriga sõnade häälikukestuste, vokaalikvaliteedi ja põhitooniliikumise kaudu, nagu seda on tehtud käesolevas õpikus P. Lippuse näidisuurimuses eesti keele väldetest. Samas on võimalik selliselt märgendatud korpuse põhjal analüüsida suhtlusdünaamikat, kõne planeerimist ja soravust jpm-d.

3.2.5. Multimodaalsete korpuste märgendamine

Multimodaalsed korpused sisaldavad, nagu nimigi viitab, korraga eri tüüpi suhtlusviise, näiteks lisaks kõnele või kirjutatud tekstile ka visuaalseid või muid mitteverbaalseid suhtlusvorme. Sellised korpused võimaldavad uurida, kuidas eri modaalsused omavahel suhestuvad ning tähendusi ja suhtumisi edastada aitavad. Ehkki sagedamini mõeldakse multimodaalsete korpuste all **audiovisuaalseid korpusi**, kus lisaks helile on salvestatud ka videopilti, on multimodaalsed korpused

näiteks ka viipekeelekorpused, mis artikulatoorset kõnet tüüpiliselt ei sisalda, või tšätikorpused, mille materjalides toimub suhtlus lisaks tekstile ka piltide, *gif*'ide, videote, häälsõnumite jm-de vahendite kaudu. Vastavalt multimodaalse korpuse tüübile võib olla vajadus märgendada seega väga erinevaid asju. Peatume siin põgusalt vaid audiovisuaalsetel korpustel, kus lisaks eelmises alapeatükis kirjeldatud suulise kõne joontele on märgendatud ka kommunikatiivseid kehaliigutusi, nagu žestid, näoilmed, peanoogutused või kulmuliigutused.

Kehaliigutuste puhul võib märgendada näiteks liigutuste **faase**: kui mingit liigutust või žesti teeme, eelneb sellele enamasti ettevalmistav faas (nt tõstame käed puhkeasendist) ning järgneb tagasiliikumise faas (nt viime käed puhkeasendisse tagasi). Samuti märgendatakse tavaliselt liigutuste **suunda** (vasakule, paremale, üles, alla, ette, taha) ja **amplituudi/ulatust** (väike liigutus, keskmine liigutus, suur liigutus).

Liigutusi võib aga märgendada ka semantiliselt, märkides ära liigutuse **tüübi** või **tähenduse**. Liigutuse tüüpidest eristatakse sageli metafoorilisi, deiktilisi, ikoonilisi, embleemseid, pragmaatilisi, interaktiivseid, vestlust liigendavaid ja kõne rõhkudega kaasnevaid liigutusi. Kui tõstame nimetissõrme püsti, võib see tähendada seda, et palume endale kõnevooru või tahame kaaskõneleja tähelepanu näiteks vastuargumendi esitamiseks või millegi selgitamiseks. Samuti võime nimetissõrme üleval hoides paluda kellelgi pisut oodata. Kui aga viibutame nimetissõrme edasi-tagasi, võime hoopis väljendada manitsust või kutsuda kaaskõnelejaid üles ettevaatusele. Teisalt võib nimetissõrme liigutamine kaasneda ka hulkade väljendamisega („üks") või loetelude esitamisega („esiteks") ning see võib mõistagi toimida ka lihtsalt deiktilise viitamise vahendina („see, millele osutan"). Kõnelejad kasutavad niisiis kommunikatiivseid kehaliigutusi erinevatel eesmärkidel. Žestid võivad osutada asukohtadele või sellele, kuidas miski ruumis paikneb, kui suur või mis kujuga miski on; need võivad toimida kommentaarina samaaegselt kõneldule, edastades suhtumisi, hinnanguid ja emotsioone; samuti võidakse žeste kasutada kuulajaga ühise konteksti loomiseks, kutsudes kuulajat kõneleja öelduga suhestuma. Näoilmed toimivad samuti mitte lihtsalt emotsioonide väljendajana, vaid ka intersubjektiivsuse loomise vahendina. Sealjuures tasub tähele panna, et eri kultuurides võivad samad liigutused ja ilmed tähendada erinevaid asju.

3.3. Märgendamistööriistad

Eestikeelsete tekstide töötlemiseks on olulisim Pythoni **teekide** kogu **EstNLTK**[15] (Estonian Natural Language Toolkit (Laur jt 2020)), mis on saanud oma nime NLTK[16] eeskujul. EstNLTK keeleanalüüsi töövoog võimaldab lisada tekstile eri-

[15] https://estnltk.github.io/

[16] https://www.nltk.org/

nevaid märgenduskihte. Automaatse **segmenteerimise** ehk **üksustamise** abil on võimalik tekst jagada lõikudeks, lauseteks, osalauseteks ja sõnedeks. Lingvistilise analüüsi vahenditest on kesksel kohal morfoloogiline analüüs, mille puhul on võimalik kasutada erinevaid märgendusskeeme ning ühestamise strateegiaid; morfoloogia töövahendid võimaldavad ka morfoloogilist sünteesi (sõnavormide genereerimist vastavalt etteantud tunnustele), automaatsilbitamist ning sõnade õigekirjakontrolli. EstNLTK sisaldab sõltuvussüntaktilise analüüsi vahendeid: on olemas nii Stanzal kui ka UDPipe'il põhinevad neuroparserite mudelid, mis märgendavad lausete süntaktilist struktuuri (lausepuid) UD (ingl *universal dependencies*) printsiipidest lähtuvalt, kasutades sisendina EstNLTK sõnestust ja morfoloogilist analüüsi. Info eraldamise vahenditest on teegis olemas nimeolemite ja ajaväljendite tuvastajad, ajaväljendite puhul ka kalendrilise semantika määraja. Semantika tööriistadest sisaldab EstNLTK eesti Wordneti, mille abil saab uurida sõnadevahelisi semantilisi seoseid (sünonüümsus, ülem- ja alammõisted), aga ka tehisnärvivõrkudel põhinevaid keelemudeleid: word2vec-i ja Berti mudelite abil saab omistada sõnadele nn tähendusvektorid (ingl *embeddings*), mida saab kasutada sõnade ja lausete semantilisel võrdlemisel ning tunnustena masinõppes. Lisaks keeleanalüüsi töövoole on EstNLTK-s vahendid ka märgendustega opereerimiseks (nt märgenduste grupeerimine ja selektiivne väljavõtete tegemine), visualiseerimiseks (Jupyter Notebooki keskkonnas) ning teisendamiseks erinevatele andmekujudele. Märgendatud tekste on võimalik täiendada metaandmetega ja salvestada (nt andmebaasi või JSON-i failidena) ning hiljem taastada ja edasi töödelda. EstNLTK ingliskeelsed kasutusjuhendid on vabalt kättesaadavad Jupyteri märkmike[17] kujul, eestikeelsed kasutusjuhendid on koondatud Tartu Ülikoolis loetavale kursusele Eesti keele töötlus Pythonis[18] (HVEE.04.004).

UDPipe'i töövoog[19] (ingl *pipeline*) lausestab ja sõnestab sisendteksti ning analüüsib selle morfoloogiliselt ja süntaktiliselt vastavalt UD märgendusskeemile. Veebiteenus võimaldab faile märgendamiseks üles laadida, kusjuures vastavaid suvandeid kasutades saab lasta märgendada ka eelnevalt lausestatud ja sõnestatud ja morfoloogiliselt märgendatud tekste; morfoloogiline märgendus peab muidugi vastama UD märgendusskeemile. Nagu eespool märgitud, saab UDPipe'i süntaksianalüüsi kasutada ka EstNLTK abil.

Sama märgenduse võimaldab väljastada ka **Stanza**[20] töövoog, mis on samuti integreeritud ka EstNLTK-sse. Erinevalt UDPipe'ist ei ole Stanzal veebiliidest, vaid see tuleb kasutamiseks enda arvutisse paigaldada.

---

17 https://github.com/estnltk/estnltk/tree/main/tutorials

18 https://github.com/d009/EstNLP

19 https://lindat.mff.cuni.cz/services/udpipe/

20 https://stanfordnlp.github.io/stanza/index.html

Ka eesti keele morfoloogia töövahendite komplekti **Vabamorf**[21] saab soovi korral kasutada eraldiseisvana, väljaspool EstNLTK-d. Vabamorfis on eesti keele morfoloogia analüsaator ja ühestaja ning morfoloogiline süntesaator, st sõnavormide genereerija.

Kui soovitakse lihtsalt asendada korpuses kõik sõnavormid nende lemmade ehk sõnaraamatuvormidega või saada kiiresti kätte ka sõnade vormiinfo, siis lihtsaim tööriist selle jaoks on **Tartu Ülikooli raamatukogu lemmatiseerimise ja morfoloogilise analüüsi veebiteenus**[22]. Teenuses saab sisendteksti kas kohapeal kirjutada või failina üles laadida.

Nimeüksuste automaatseks märgendamiseks eestikeelses tekstis on olemas kaks närvivõrkudel põhinevat mudelit, millest **EstBertNER**[23] märgendab isiku-, koha- ja organisatsiooninimesid ja **EstBertNER_v2**[24] eristab 11 nimekategooriat (isikud, kohad, organisatsioonid, geopoliitilised üksused, sündmused, tiitlid, tooted; aga ka rahaühikud, ajaväljendid, kuupäevad, protsendid), millest osa pole mitte nimed, vaid muud infoeralduse seisukohast olulised üksused. Lisaks on olemas K. Muischneki ja S. Orasmaa näidisuurimuses kirjeldatud nimeüksuste märgendaja 19. sajandi vallakohtuprotokollide tekstide jaoks[25], mis tunneb nt isikunimed tekstis ära 96% täpsusega ja mis võib olla rakendatav ka teiste sama perioodi tekstiliikide peal, aga sellega töötamiseks on vajalik programmeerimisoskus.

Suurte keelemudelite (GPT, Llama jt) kasutusvõimalused korpusandmete märgendamisel on väga kiiresti arenev valdkond. Katsetatud on peamiselt ingliskeelsete tekstide märgendamisega, näiteks märgendas parim keelemudel Danni Yu jt uuringu kohaselt (Yu jt 2024) ingliskeelses tekstis kõneakte 84-protsendilise korrektsusega. Petter Törnberg (2023) sedastas, et ChatGPT-4 liigitab Twitteri sõnumeid USA vabariiklaste *vs.* demokraatide partei esindaja kirjutatuks paremini kui inimmärgendaja. Suuri keelemudeleid ja neil põhinevaid vestlusroboteid ei saa aga andmete märgendamiseks kasutada ilma neile põhjalikku juhendit ehk viipa (ingl *prompt*) ja märgendusnäiteid andmata ning tasub katsetada erinevalt sõnastatud juhenditega mõistmaks, millise viiba alusel annab vestlusrobot vastuseks parima tulemuse. Ülesande õige sõnastamine suure keelemudeli jaoks on saanud omaette oskuseks, veebiotsinguks sobivad ingliskeelsed märksõnad *prompting*, *prompt engineering* või *instruction tuning*. Samas tuleb meeles pidada, et suured keelemudelid on siiski „mustad kastid", st pole võimalik täielikult aru saada, miks nad teevad just selliseid märgendamisvalikuid, nagu nad teevad. Samuti ei ütle nad ilma vastavate juhisteta kunagi, et ei saanud ülesandest aru või et nende meelest on juhend liiga ebamäärane või vastuoluline.

---

21 https://github.com/Filosoft/vabamorf/blob/master/LOEMIND.md

22 https://tekstianalyys.utlib.ut.ee/index.html#

23 https://huggingface.co/tartuNLP/EstBERT_NER

24 https://huggingface.co/tartuNLP/EstBERT_NER_v2

25 https://github.com/soras/vk_ner_lrec_2022

Ka multimodaalsete korpuste jaoks relevantseid **füüsilisi liikumisi** on lisaks käsitsi märgendamisele võimalik teatud määral märgendada automaatanalüüsi tööriistadega. Näiteks videopildilt kehapunktide (pea, näo osad, õlad, küünarnukid, käed, sõrmed, jalad jne) paiknemise tuvastamiseks on võimalik kasutada OpenPose'i[26] ja Mediapipe'i[27] tarkvara, ainult näo osade (suu, silmad, kulmud, nina) tuvastamiseks OpenFace'i[28] tarkvara. Nendest andmetest on võimalik mõõta näiteks liigutuste kiirust ja ulatust. Klassifitseeritud žestide automaatseks märgendamiseks on võimalik treenida nt DeepLabCuti[29] või Nova[30] tarkvaraga käsitsi märgendatud andmete põhjal oma mudel.

Paljusid keelenähtusi tuleb siiski märgendada **käsitsi** ning ka selleks on olemas palju abivahendeid, viiteid märgendamistööriistadele on kogutud vastavale veebilehele[31]. Ühed sellised paljude funktsionaalsustega märgendustööriistad on näiteks INCEpTION[32] ja brat[33]. Mõlemad pakuvad ka märgendatud andmete analüüsivahendeid. Tasub meeles pidada, et sellise töövahendi abil tehtud märgendus tuleb mõne ülesande lahendamiseks tööriistast eksportida ja seega tasub uurida, milliseid funktsionaalsusi ja ekspordivorminguid tööriist toetab.

Lõpetuseks

Õpiku kolmandas peatükis andsime ülevaate märgendamisest – protsessist, mille käigus varustatakse tekstikorpused lisainfoga nende struktuuri, sõnade või muude omaduste kohta. Muu hulgas kirjeldasime märgendamise üldiseid põhimõtteid, märgenduse erinevaid liike (nt morfoloogiline, süntaktiline, semantiline) ning eesti keele jaoks kasulikke märgendamistööriistu (nt EstNLTK, UDPipe, Stanza, Vabamorf). Multimodaalsete korpuste märgendamine hõlmab lisaks tekstile ka visuaalsete ja mitteverbaalsete suhtlusvahendite märgendamist, näiteks žeste ja näoilmeid. Erineva lisainfoga varustatud korpused on väärtuslikud loomuliku keele töötluse süsteemide, nagu kõnetuvastuse, masintõlke, tekstide märksõnastamise ja meelestatuse analüüsi väljatöötamiseks ja täiustamiseks. Uurija seisukohast on märgendatud korpused väärtuslikud seetõttu, et need võimaldavad rikkalikumat ja struktureeritumat keeleteaduslikku analüüsi. Võimalik on näiteks

[26] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[27] https://github.com/google-ai-edge/mediapipe

[28] https://github.com/TadasBaltrusaitis/OpenFace

[29] https://github.com/DeepLabCut

[30] https://github.com/hcmlab/nova

[31] https://corpus-analysis.com/tag/annotation

[32] https://inception-project.github.io/

[33] https://brat.nlplab.org/

sõnaliikide, süntaktiliste seoste, semantiliste ja isegi pragmaatiliste aspektide automatiseeritud analüüs. Eesti keele puhul tuleb silmas pidada, et meie keele rikkast ja komplekssest morfoloogilisest süsteemist tulenevalt on näiteks lemmatiseerimine ja morfoloogiline märgendamine keerukam kui analüütiliste keelte (nt inglise keel) puhul. Sageli vajame kohandatud märgendusskeeme või poolautomaatseid meetodeid, kus inimese sekkumine protsessi aitab täpsust parandada.

Lisalugemiseks

Newman, John & Christopher Cox. 2020. Corpus annotation. Magali Paquot & Stefan Th. Gries (toim), *A Practical Handbook of Corpus Linguistics*, 25–48. Cham: Springer. https://doi.org/10.1007/978-3-030-46216-1.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7), 38–45. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.3735822.

4. Oma korpuse loomine

Maarja-Liisa Pilvik, Kadri Muischnek, Kristina Koppel, Jelena Kallas, Pärtel Lippus

4.1. Miks oma korpust luua?

Nagu nägime 2. peatükis, peaksid eesti keele uurijad, kasutajad ja õppijad olema üpris rahul, kuna eesti keelt sisaldavaid korpusi on olemas väga erisuguseid: meil on suured üldkeele korpused ja väiksemad erikorpused; on suulise keele korpused ja kirjaliku keele korpused; korpuste materjalid sisaldavad erisuguseid keelekasutussituatsioone ja erinevate keelekasutajate keelt. Ometi alati olemasolevatest korpustest ei piisa ning tekib vajadus luua nende kõrvale päris uus keelekorpus. Sellised uued (tüüpiliselt väikesed) korpuseprojektid koguvadki enamasti kas kindla kõnelejaskonna keelt või kitsama valdkonna tekste. Näiteks võib meil olla soov uurida üle 60-aastaste inimeste sõnumivahetust, kokandusblogijate keelt või parteide valimisprogrammide keelt. Samuti võime mõelda oma korpuse loomisele siis, kui meid huvitab, kuidas on rääkinud või kirjutanud mõni konkreetne autor või kuidas on räägitud ja kirjutatud mingil konkreetsel ajaperioodil. Vahel tuleneb oma korpuse loomise vajadus ka sellest, et tahame uurida kindlat nähtust, näiteks dialoogide ja monoloogide alustamist, kehaosametafoore aastapäevakõnedes või väidete pehmendamise võtteid (ingl *hedging*) üliõpilastöödes. Kõikidel sellistel juhtudel ei pruugi olemasolevad korpused sisaldada huvipakkuvaid tekste kas piisaval määral, piisavalt esinduslikult või kujul, mis võimaldaks huvipakkuvat hõlpsasti üles leida.

Selles peatükis kirjeldame, mis põhimõtteid uue korpuse koostamisel silmas pidada, kust ja kuidas koguda tekste, millised on korpuste kogumise ja avalikustamisega seotud eetilised ja õiguslikud aspektid, miks on oluline pühendada aega ka korpuse koostamise dokumenteerimisele ja metaandmetele ning milliseid praktilisi küsimusi võib korpuse loomisel ette tulla.

4.2. Korpuse koostamise põhimõtted

Oma korpuse koostamine, mis hõlmab tekstide kogumist, struktureerimist, metaandmete kogumist ja korrastamist ning vajadusel ka tekstide struktuurilist ja funktsionaalset märgendamist (vt ptk 3 „Märgendamine"), on küllaltki

töömahukas protsess ning nõuab hoolikat läbimõtlemist ja otsuste tegemist mitmes etapis.

Haruharva on meil kasutada kõik mingi keelevariandi, keelekasutussituatsiooni või keelekasutajate rühma keeleandmed. Seetõttu esindavad nii juba olemasolevad kui ka loodavad korpused üldjuhul ainult väikest osa kogu (sama tüüpi) keelekasutusest. Korpusesse saab niisiis kaasata vaid kindlatel põhimõtetel koostatud valimi, mille põhjal saaks teha üldistusi terve populatsiooni kohta. Nagu õpiku 1. peatükis mainitud, on keeleandmete puhul populatsiooni enamasti võimatu täpselt kindlaks teha. Ometi mõjutab populatsiooni määratlemine nii koostatava korpuse **suurust, esinduslikkust kui ka tasakaalustatust** (vt ptk 1.3.1 „Mis on korpus?“).

Erikorpused, mille eesmärgiks on pakkuda materjali selleks, et vastata mingit kindlat teemat või keelekasutussituatsiooni puudutavatele uurimisküsimustele (nt spordiuudiste korpus, teismeliste tšätikorpus või õpikukorpused), esindavad enamasti mingit väiksemat, spetsiifilist populatsiooni ning on seetõttu ka väiksemad kui suured üldkeele korpused. Korpuse suurust tingib ja piirab kindlasti ka see, kui palju vastavaid tekste on üldse võimalik koguda, millise infoga tekste tahetakse varustada, st mille suhtes märgendada ja millised tööriistad selleks olemas on. Mida enam on protsessi võimalik automatiseerida, seda suurem (ehkki mitte tingimata esinduslikum) saab olla ka korpus. Nii erineb inimressursi kasutamise vajadus märkimisväärselt sellise korpuse puhul, mis sisaldab veebilehtedelt kogutud standardset kirjakeelset teksti, millele lisatakse ainult automaatne morfoloogiline märgendus, ja korpuse puhul, kus kehvades salvestustingimustes salvestatud suulist keelt tuleb käsitsi transkribeerida ja spetsiifiliste semantiliste kategooriate alusel märgendada. Võimalikult suur korpus ei peaks niisiis olema oma korpuse koostamisel põhiline siht.

Koostatav korpus peaks olema võimalikult **esinduslik** ja **tasakaalustatud** selle valdkonna autorite, allikate, allteemade jm suhtes. Kui korpus on väga üldine ja selle eesmärk on olla näiteks väljavõte kogu eesti keele kõnelejaskonna keelekasutusest, võiks esinduslik ideaalkorpus sisaldada nii kirjutatud kui ka suulist keelt, eri keelekasutussituatsioone, eri vanuserühmade keelt jne. Kui korpuse eesmärk on pakkuda materjali tänapäeva akadeemilise kirjutamise uurimiseks, peaks esinduslik korpus sisaldama tekste eri valdkondadest, eri tüüpi teaduspublikatsioonidest ja erineva akadeemilise ettevalmistusega kirjutajatelt. Kui korpus on koostatud ühe konkreetse valdkonna kliendisuhtlusroboti treenimiseks, peaks esinduslik korpus sisaldama võimalikult palju selle valdkonna näiteid eri tüüpi kliendisuhtlusolukordadest (nt infopäringud, kaebused, palved, teavitused, avaldused), võttes arvesse ka klientide profiili (nt pikaajalised *vs*. uued kliendid, ettevõtted *vs*. eraisikud jne). Esinduslikkus on põhiline, mis eristab korpust juhuslikust tekstikogust. Kui eesmärgiks on ka korpuse tasakaalustatus, tuleb lisaks määrata kindlaks, millistes proportsioonides eri tüüpi tekstid kirjeldatavas populatsioonis võiksid esineda. Sageli saab ülaltoodud põhimõtete järgimist kontrollida muidugi alles siis, kui korpus on juba mingil kujul valmis. Seetõttu võib korpuse koostamisele läheneda

ka tsükliliselt, kohandades vastavalt empiirilistele andmetele oma teoreetilisi eeldusi (nt korpuse aluseks oleva populatsiooni kohta) ja vastupidi – püüdes korpuse andmete tegelikku jaotumist viia vastavusse (lõdvendatud) teoreetiliste eeldustega. Alati (nt üliõpilastöö raames) ei olegi muidugi võimalik korpuse esinduslikkuse ja tasakaalustatuse tagamisele väga palju aega kulutada ning seda olulisem on, et korpuse **dokumentatsioonis** (millest tuleb pikemalt juttu alapeatükis 4.6) oleks selgelt kirjas, kuidas ja milliste kriteeriumite alusel korpust moodustavad tekstid välja valiti, et ka korpuse teised kasutajad saaksid oma uurimisküsimustest lähtuvalt teha teadlikke otsuseid.

Kui populatsiooni määratlemine ning esinduslikkuse ja tasakaalustatuse küsimused on pigem teoreetilist laadi, siis valimi koostamise juures on ka palju praktilisi otsustuskohti. Esmalt tuleb valimi moodustamisel mõelda välja, mis on need **üksused**, mida korpusesse kaasata. Kui võtta korpusesse katkendite asemel täistekste (nt terved ajaleheartiklid või kogu ilukirjandusteos, suulise keele puhul terve intervjuu või vestluse litereering), säilib võimalikult palju kontekstuaalset infot ja korpuse kasutusvõimalused on laiemad (näiteks võib korpust kasutada ka diskursusanalüüsiks ja tekstilingvistilises uurimistöös). Teisalt võib täistekstide kaasamine olla keerukas autoriõiguste tõttu (vt ka alapeatükk 4.5), samuti peab täistekstide analüüsimisel silmas pidama, et tekstide pikkus võib oluliselt varieeruda ning mingite tekstide eripärad võivad hakata domineerima, eriti väikeste korpuste puhul. Kui kasutada täistekstide asemel katkendeid, tuleb jällegi otsustada, mitmesõnalisi katkendeid kasutatakse (nt 1000 sõna, 5000 sõna, 10 000 sõna) ja kust neid katkendeid võetakse (teksti algusest, keskelt, lõpust või iga teksti puhul juhuslikust kohast). Katkendikorpus on näiteks eesti keele baaskorpus (vt ptk 2.2.1 „Esimesed eesti keele korpused: esinduslikud, aga väikesed"), kuhu on kaasatud u 2000-sõnalised katkendid tekstide juhuslikult valitud osadest. Mida spetsiifilisem on tekstitüüp, millest korpust koostatakse, seda rohkem rolli katkendikohtade valik mängib, kuna eri tekstiosade funktsioonid võivad olla erinevad: näiteks akadeemilisi tekste mõjutavad kindlad kirjutamistraditsioonid, mille tõttu on sissejuhatuse, materjali tutvustuse ja analüüsitulemuste esitamise keelekasutus küllalt erinev.

Kui tahta jagada korpust omakorda alamosadeks (nt ilukirjanduskorpus ja ajakirjanduskorpus), võiksid alamkorpuste tekstid olla võimalikult **homogeensed**. Näiteks kui koostada üldist ilukirjanduskorpust, ei ole ilmselt hea kaasata sinna romaanide ja jutustuste hulka paari üksikut luuletust. Vahel räägitakse tekstide homogeensusest ka sisemise tasakaalustatuse võtmes: iga sama pikkusega tekstilõik või -katkend võiks ideaalis sisaldada enam-vähem samas proportsioonis uuritavaid üksusi (nt lekseeme, grammatilisi kategooriaid vms). Kui erinevused tekstide vahel on väga suured, võib see kaasa tuua selle, et korpuse eri osi kasutades jõutakse märkimisväärselt erinevate tulemusteni (näiteks sõnastatistika puhul). Selliseid sagedusjaotusi saab aga kasutada pigem korpuse homogeensuse diagnostikaks või korpuse tekstide/katkendite valiku hõlbustamiseks, päris korpuse koostamise algetapis on sisemise, leksikaalse ja grammatilise tasakaalustatuse tagamine küllalt

keerukas ja vahest ka vastuoluline samm, kui eesmärgiks on luua näiteks paljusid eri keelekasutussituatsioone ja registreid esindav tekstikorpus.

Kui on selge, milliseid, kui suuri ja kui palju üksuseid korpusesse võtta, tuleb järgmiseks tegeleda tekstide tegeliku kogumisega. Harva on võimalik kätte saada **kõikne valim** ehk (peaaegu) kõik populatsiooni võimalikud esindajad (nt kõik A. H. Tammsaare tekstid). Sagedamini peame leppima sellega, et saame kasutada vaid osa kõigist võimalikest keeleandmetest selleks, et kirjeldada tervet populatsiooni. Üks võimalus korpusesse materjali valida on võtta soovitud üksustest **juhuvalim**. Sellisel juhul on igal üksusel (nt tekstikatkendil) võrdne võimalus valimisse sattuda. Mida rohkem mingit tekstitüüpi või keelenähtust populatsioonis esineb, seda rohkem satub seda juhuvalimiga ka korpusesse, harvad nähtused võivad aga lihtsa juhuvalimiga korpusest välja jääda. Kui uurime kindlat tekstitüüpi või kõnelejate rühma, tähendab juhuvalimi võtmine, et kõnelejatelt/kirjutajatelt, kellelt on üldiselt rohkem üksusi, satub ka valimisse rohkem materjali. Lisaks täiesti juhuslikule valimile võib kasutada ka **juhuslikku kihtvalimit**, mispuhul jagatakse populatsioon kõigepealt teatud tunnuste alusel n-ö kihtideks, määratakse iga kihi (hinnanguline) osakaal[1] kogu populatsioonis ning võetakse seejärel juhuvaliku teel igast kihist kindel arv tekste nii, et terves korpuses säiliksid populatsiooni eeldatavad proportsioonid. Kui lähtuda keelest tervikuna, võib kihte moodustada näiteks tekstiliikide, žanride või registrite põhjal; kui lähtuda kas keeleloomest või keelest arusaamisest, võib kihtide moodustamiseks kasutada ka demograafilisi näitajaid, nagu kõneleja vanus, sugu, elukoht või haridustase. Mõnel puhul tahame muidugi sihilikult luua hoopis sellist korpust, mille alamkorpuste proportsioonid ei oleks vastavuses nende proportsioonidega populatsioonis. Selliseks korpuseks on näiteks TÜ tasakaalus korpus (vt ptk 2.2.2 „Teise põlvkonna eesti keele korpused: pole esinduslikud, aga nende sisu on teada"), mis sisaldab enam-vähem võrdsel hulgal sõnu nii ajakirjandus-, ilukirjandus- kui ka teadustekstidest ning mis võimaldab seetõttu paremini hinnata tekstiliikide vahelisi erinevusi (nt sõnaliikide proportsioonides või teatud vormelite kasutuses). Esinduslikkuse seisukohast halvem variant korpusesse tekstide kaasamiseks on kasutada **mugavusvalimit** ehk valida korpusesse sellised üksused, mida on kõige lihtsam või odavam kätte saada. Mõneti võib mugavusvalimiks pidada näiteks tasuta sisuga veebiportaalide tekstide automaatset kogumist (nn veebikraapimist, vt alapeatükk 4.3). Samavõrd probleemaatiline on aga esinduslikkuse mõttes ka **sihipärane valim**, mille puhul korpuse koostaja üksi otsustab, millised tekstid populatsiooni kõige paremini iseloomustavad, ning kureerib hoolikalt korpuse sisu.

[1] On küllalt keeruline, kui mitte võimatu, objektiivselt määrata populatsiooni kihtide täpseid osakaale. Samuti ei ole tekstiliikide piirid sugugi üheselt defineeritavad. Seetõttu saab siin olla tegemist ainult hinnanguga sellele, milline üldine populatsioon võiks välja näha ja kui palju tekste mingilt kihilt valimisse kaasata.

4.3. Andmete kogumine ja korrastamine

Andmete kogumise viis ja vahendid sõltuvad eelkõige materjalist, mida soovitakse koguda. Laiemalt võime jagada koostatavad korpused kirjalikeks ja suulisteks korpusteks.

Kirjalike korpuste koostamine on üldjuhul lihtsam, kiirem ja odavam kui suulise korpuse koostamine[2]. Selleks võib tänapäeval ära kasutada näiteks nii olemasolevaid dokumendikogusid kui ka veebilehti.

Dokumendikogud võivad omakorda sisaldada juba **digitaalseid tekste** (nt TXT-, PDF-, DOCX-vormingus) või **mittedigitaalseid tekste**. Esimesel juhul saame kasutada korpuse koostamiseks ja korrastamiseks kohe olemasoleva korpusanalüüsi tarkvara võimalusi (nt AntConc või Sketch Engine, vt ka alapeatükk 4.3.1); teisel juhul kaasneb korpuse koostamisega aga ka tekstide digiteerimine. Seda kirjeldame täpsemalt alapeatükis 4.3.2.

Veebilehtedest korpuse koostamine tähendab, et võime koguda väga palju materjali, ent peame silmas pidama, et see on raskesti kontrollitav ja pidevalt muutuv. Samuti võib suur hulk olulist materjali (nt perioodikat) jääda maksumüüri taha või võib veebilehe omanik keelata selle allalaadimise ja edasijagamise.

Internetis leiduva materjali allalaadimiseks saab kasutada näiteks Sketch Engine'i vastavat funktsionaalsust, millest tuleb juttu alapeatükis 4.3.1. Sketch Engine on tasuline tarkvara ning selle kasutamiseks peab ostma eraisik või institutsioon (nt kõrgkool) kasutuslitsentsi. Ehkki akadeemiliseks kasutamiseks ei ole litsentside hinnad kuigi kõrged, arvestades kõiki erinevaid korpusi ja analüüsitööriistu, mida keskkond kasutada võimaldab, leidub konkreetseteks ülesanneteks sellele ka tasuta alternatiive. Nii saab ka tekstide kogumiseks kasutada vabavaralisi Pythoni või R-i pakette (Pythonis nt `Beautiful Soup`[3], R-is `rvest`[4]). Samuti on olemas üldisemaid, mitte spetsiaalselt keeleanalüüsiks mõeldud valmisrakendusi, mis etteantud veebilehtede sisu dokumentidesse koguvad (nt ParseHub[5] või Google Chrome'i laiendus Web Scraper[6]). See, kas üldse ja milliseid automaatseid sisukogujaid mingi veebileht oma sisu kallale lubab, on tavaliselt kirjas veebilehe failis nimega *robots.txt*[7].

---

[2] Sellest hoolimata olgu ära märgitud praktiline tõsiasi, et andmete kogumine, korrastamine ja märgendamine võtab alati rohkem aega, kui selleks planeeritud. Enamasti võib esialgselt planeeritud ajakulu tegeliku ajakulu leidmiseks korrutada kahega.

[3] https://www.crummy.com/software/BeautifulSoup/

[4] https://rvest.tidyverse.org/

[5] https://help.parsehub.com/hc/en-us

[6] https://chromewebstore.google.com/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn?pli=1

[7] Vt nt https://ut.ee/robots.txt

Nagu öeldud, on **suuliste korpuste** koostamine aja- ja ressursimahukam, seda ka juhul, kui mingit osa korpuse loomisest on võimalik automatiseerida. Veidi hõlpsam on koostada korpust juba olemasolevatest salvestistest (nt arhiivisalvestised, YouTube'i videod, taskuhäälingud, raadiosaated), mille puhul tuleb lisaks materjali valiku põhimõtetele panna paika ka see, mil moel salvestisi transkribeerida. Kui korpuse koostamine hõlmab ka salvestamist, lisanduvad etappide hulka salvestusseadmete soetamine, kõnelejate värbamine, kõnelemissituatsiooni määratlemine jpm-d. Täpsemalt on kõnekorpuste koostamisest juttu alapeatükis 4.3.3.

4.3.1. Kirjalikud korpused digitaalsetest tekstidest

Kui tekstid, millest tahame korpust luua, on juba kuskil digitaalsel ja masinloetaval kujul olemas, saab neist tänapäeval olemasolevate digitööriistade abil küllalt hõlpsalt koostada uue tekstikorpuse. Enne korpuse loomist on vaja siiski läbi mõelda, kuidas tekstid kokku koguda, milliseid metaandmeid on vaja säilitada, kas dokumente tuleks puhastada, kas tekste tuleks normaliseerida (st nende kirjaviisi ühtlustada, vt alapeatükk 4.3.2.4), kas dokumendid sisaldavad isikuandmeid, milline info tuleks eemaldada (nt autori kontaktandmed), millises märgistikus tekstid on jne. Korpuse failide ettevalmistamisest on põgusalt juttu alapeatükis 4.4.

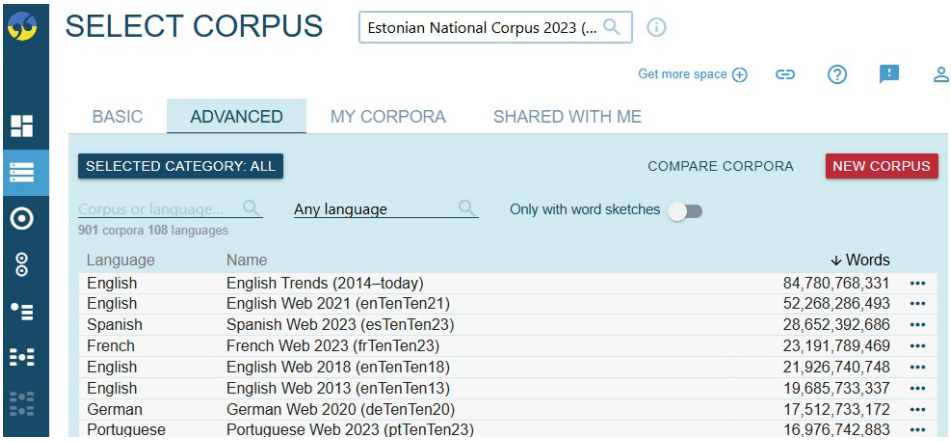
Kirjeldame siin digitaalsetest tekstidest oma korpuse loomise võimalusi korpusanalüüsi tarkvaraga Sketch Engine, kus seda saab teha kolmel viisil: 1) sisu automaatselt veebist alla laadides[8] (nn veebikraapimine), 2) oma faile üles laadides[9], 3) mõlemat meetodit kombineerides. (Sketch Engine'i kohta loe lisaks peatükist 5.1.2 „Veebipõhised korpusanalüüsi keskkonnad" ja K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimusest korpuste kasutamisest sõnastike koostamisel). Kui oma korpus on juba eelnevalt loodud, saab sinna samadel meetoditel tekste juurde lisada.

Sketch Engine'is hoiustatud andmed ei ole avalikud. Tekstid, mis sinna korpuse jaoks üles (või veebist alla) laaditakse, salvestatakse isiklikule kontole, kuhu teised kasutajad ligi ei pääse. Soovi korral saab oma korpust jagada teiste Sketch Engine'i kasutajatega.

Oletame, et soovime luua jalgpalliteemalise korpuse. Selle loomist saab alustada kolmest kohast: 1) otse Sketch Engine'i töölaualt, sinine nupp (*NEW CORPUS*) asub paremal üleval; 2) vasakult korpuse valimise menüüst, punane nupp asub paremal üleval (vt joonis 4.1) või 3) korpuse otsinguribal (*create corpus*).

[8] https://www.youtube.com/watch?v=VjHC4lMop-s

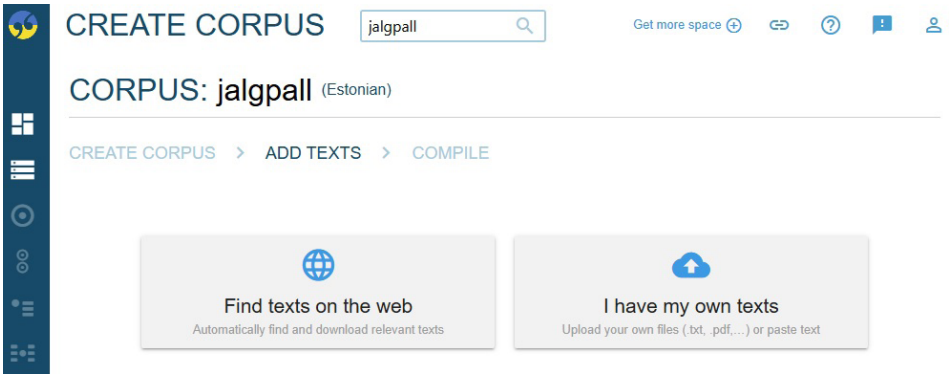
[9] https://www.youtube.com/watch?v=gMicxJAS024



Joonis 4.1. Uue korpuse loomise nupp korpuse valimise menüüs

Korpuse loomisel tuleks anda oma korpusele nimi (nt *jalgpall*) ning valida keel, milles korpuse tekstid on kirjutatud. Soovi korral võib lisada ka korpuse lühikirjelduse. Keele valik määrab ära, milliseid tööriistu saab Sketch Engine'is märgendamiseks ja hilisemaks analüüsiks kasutada. Eesti (standard)keele puhul saab kasutada kõiki samu tööriistu, mida ka näiteks ühendkorpuse puhul, samuti saab lisada kõikidele tekstisõnadele automaatse morfoloogilise märgenduse. Kui tahta luua korpust keele jaoks, mida Sketch Engine ei toeta, siis õpetuse selleks leiab kodulehelt[10].

Järgmise sammuna tuleb otsustada, kas luua oma korpust sisu veebist alla laadides (*Find texts on the web*) või oma tekstifaile üles laadides (*I have my own texts*, vt joonis 4.2).



Joonis 4.2. Korpuse sisu alla- või üleslaadimine

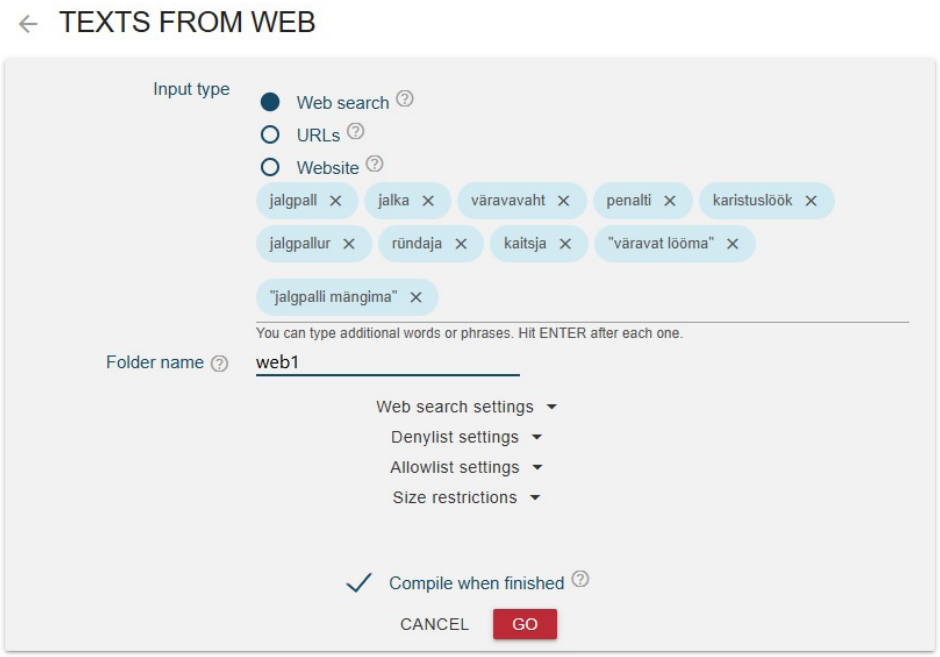
[10] https://www.sketchengine.eu/guide/create-a-corpus-from-the-web/#toggle-id-6

4.3.1.1. Oma korpuse loomine veebist

Sketch Engine'isse on integreeritud korpuse loomise tööriist, mis kasutab WebBootCaT[11] tehnoloogiat. Selle abil on võimalik automaatselt luua tekstikorpus veebilehtedelt alla laaditud keeleandmetest. Enne korpusesse lisamist kogutud andmed puhastatakse: näiteks eemaldatakse duplikaadid ehk identsed tekstid ning kõrvaldatakse mittetekstiline materjal (nt pildid ja tabelid).

Kasutaja saab määrata, millist sisu tuleks veebilehtedelt alla laadida. Seda saab teha kolmel viisil:

1. Tekstide kogumist saab alustada **võtmesõnade** abil (*Web search*). Võtmesõnad on tüüpsõnad, mis uuritavat teemat võiksid määratleda (nt saab jalgpalliteemalise korpuse kogumiseks kasutada võtmesõnu *jalgpall*, *väravavaht*, *penalti*, *karistuslöök*, *nurgalöök*, *jalgpallur* jmt). Minimaalselt saab ette anda kolm võtmesõna ning tavaliselt piisab kuni 20 võtmesõnast, seejuures sobivad võtmesõnadeks ka mitmesõnalised ühendid ja pärisnimed.
2. Tekste võib koguda, esitades **nimekirja veebiaadressidest** ehk URL-idest, millelt tuleks dokumendid alla laadida (*URLs*).
3. Alla saab laadida ka **ühe konkreetse veebilehe** kogu sisu (*Website*).



Joonis 4.3. Oma korpuse loomine võtmesõnade abil

[11] https://www.sketchengine.eu/wp-content/uploads/2015/03/WebBootCaT_web_tool_2006.pdf

Vajutades punasel nupul *GO*, hakkab veebisisu otsimine ja allalaadimine pihta (joonis 4.3). Kui korpus on kogutud, automaatselt puhastatud, lemmatiseeritud ja märgendatud, ilmub teade, et korpus on kasutamiseks valmis. Eri keelte puhul kasutatakse erinevaid märgendustööriistu, eesti keele puhul kasutab Sketch Engine hetkel vaikimisi EstNLTK versiooni 1.6[12].

Mida rohkem võtmesõnu lisada, seda suurema korpuse saab, ent kõik tulemused ei pruugi sel juhul olla korpuse teema seisukohast relevantsed. Täpsemaks otsinguks võib rakendada ka lisakriteeriume, näiteks teha mõned võtmesõnad kohustuslikuks (*Allowlist settings*) või vastupidi, keelata mingite konkreetsete sõnade esinemine ära (*Denylist settings*). Korpuse mahu suurendamiseks võib sama protseduuri korrata mitu korda. Sketch Engine tagab siinjuures, et ühtegi lehekülge ei lisataks mitu korda.

4.3.1.2. Oma korpuse loomine failidest

Kui tekstid, millest korpust koostada tahetakse, ei paikne veebis, vaid digitaalsel kujul failidena, saab need Sketch Engine'isse üles laadida. Sellistes failides võivad olla näiteks tekstid, mida tudengid on kirjutanud, teatud artisti laulusõnad, mingilt institutsioonilt või firmalt saadud meilid, enda kirjutatav romaan, suhtlusrakendustes peetud vestlused või kasvõi Tammsaare kogutud teosed. Ehkki Sketch Engine'isse üles laaditud tekstid salvestatakse kasutajakontole ning on vaikimisi nähtavad ainult kasutajale endale, tuleb, nagu alati kuskile veebikeskkonda faile üles laadides, arvestada sellega, et tekste võivad siiski näha ka muud inimesed, näiteks veebikeskkonna haldajad. Samuti ei ole kunagi päris kindel, kas pärast seda, kui oleme oma korpuse veebikeskkonnast kustutanud, jäävad meie tekstid endiselt mingisse veebiserverisse mõneks ajaks alles. Seetõttu peaks enne korpuse loomist tegema kindlaks, kas meil on üldse õigus tekste jagada ning kas need sisaldavad mingit tundlikku infot, mis tuleks esmalt eemaldada (vt ka alapeatükk 4.5).

Oma failidest korpuse loomiseks tuleb valida pärast korpusele nime andmist ja keele määramist variant *I have my own texts* („mul on oma tekstid") ning lohistada failid vastavasse aknasse (joonis 4.2). Failid võivad olla DOC-, DOCX-, HTM-, HTML-, TEI-, TMX-, TXT-, VERT-, XML- või PDF-vormingus; paralleelkorpuste ja mitmekeelsete puhul XLS-, XLSX-, TMX-, XLS-/XLIFF- või ODS-vormingus. Paralleelkorpuste kogumisest ja joondatud tekstidest saab täpsemalt lugeda Sketch Engine'i lehelt[13]. Ühte korpusesse saab üles laadida ka erinevates vormingutes faile ning mitu dokumenti korraga, kui kasutada ZIP- või tar.gz-vormingus faile. Sealjuures eirab Sketch Engine korpuse koostamisel taas kõiksugu failides olevaid pilte jm mitteteksti.

Oma korpuse loomisel Sketch Engine'is on mahupiirangud. Maksimaalselt saab üles laadida kuni 100 faili, kusjuures ükski fail ei tohi olla suurem kui 500 MB. Kõikides oma loodud korpustes saab kokku olla kuni miljon sõna. Kui on vajadus

[12] https://github.com/estnltk/estnltk/tree/version_1.6

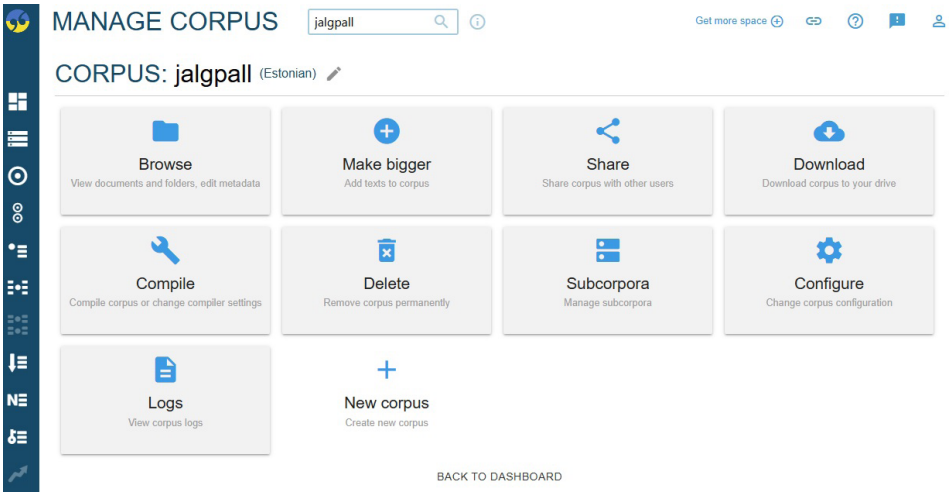
[13] https://www.sketchengine.eu/guide/setting-up-parallel-corpora/

luua suuremaid korpusi, saab seda teha osadena, st koostada esmalt väiksem korpus, laadida see alla (vt järgmist alaosa), kustutada korpus Sketch Engine'ist ära, luua uus väiksem korpus jne.

4.3.1.3. Loodud korpuse kasutamine

Loodud korpus ilmub Sketch Engine'i korpuste loetellu, ehkki on nähtav ainult selle koostajale või nendele Sketch Engine'i kasutajatele, kellega koostaja on oma korpust jaganud. Korpust saab otsida otsinguaknast või näha seda oma kontole sisse logides viimati avatud korpuste nimekirjas.

Võtmesõnade ja terminite funktsioon (*Keywords and Term Extraction*[14]) võimaldab oma korpuse sisu võrrelda mõne suurema referentskorpusega ning teha sel viisil selgeks, millised sõnad ja terminid (milleks Sketch Engine's nimetatakse igasuguseid oodatust sagedasemaid sõnaühendeid, vt lähemalt ptk 5.2.6 „Võtmesõnad") on sinu korpuses sagedad. Selle loendi abil saab samuti valida välja uued võtmesõnad, millega oma korpuse mahtu edasi suurendada. Joonis 4.4 näitab võimalusi, mida enda loodud korpusega teha: korpuse dokumente saab eraldi vaadata ja ära kustutada (*Browse*), korpust saab suurendada (*Make bigger*), jagada teiste kasutajatega (*Share*), alla laadida (*Download*), jaotada alamkorpusteks (*Subcorpus*), aga ka kustutada (*Delete*).



Joonis 4.4. Korpuse haldamise vahendid

Soovitame lisaks vaadata Sketch Engine'i õppevideoid nende YouTube'i kanalil[15].

[14] https://www.sketchengine.eu/guide/keywords-and-term-extraction/

[15] https://www.youtube.com/channel/UCo2fn2_SNxCikCSAFCBcWBw

4.3.2. Kirjalikud korpused mittedigitaalsetest tekstidest

Digitaalsete andmete hulk, millest korpust koostada ja mille kaudu keelt uurida, kasvab hüppeliselt iga aastaga. Nii avalik meedia kui ka isiklik suhtlus on kolinud suuresti digikanalitesse, ka trükitud materjalid valmivad üldjuhul esmalt digitaalselt. Sellegipoolest ei ole meid huvitavad tekstid alati digitaalsel kujul kättesaadavad. Sellised tekstid võivad olla vanemad raamatud või ajalehed, käsikirjad, arhiividokumendid ja muud ajaloolised andmekogud, mis on olulised keele, kultuuri, ajaloo ja ühiskonna uurimise allikad. Keeleteadlastele näiteks pakuvad vanemad tekstid võimaluse pääseda ligi varasematele keelekujudele, võimaldavad hinnata keeles juba toimunud ja toimumas olevaid muutusi, jälgida kirjakeele väljakujunemist ja arengut ja palju muud. Kirjandusteadlastele võivad omakorda pakkuda huvi vanemad autorid, teosed, nende keel ja teemad, stiilid ja omavahelised mõjutused.

Mittedigitaalsed tekstid, mida uurida, ei pea aga tingimata olema vanad. Näiteks võime tunda huvi kõiksugu pisitekstide vastu, nagu teavitussildid avalike asutuste ustel ja akendel või kooliõpilaste spikrid, mille digitaalkujul originaalidele (juhul, kui need üldse kunagi on olemas olnud) puudub meil ligipääs.

Kõikide selliste tekstide uurimine korpuslingvistika meetoditega nõuab, et viiksime füüsilisel kujul tekstid esmalt **masinloetavale** kujule, mis lubab teksti arvutiga töödelda. Oluline ongi ehk rõhutada, et ehkki vastandame siin digitaalseid ja mittedigitaalseid tekste, on tekstide suuremahuliseks analüüsiks tegelikult oluline just nende masinloetavus. Viimase paarikümne aasta jooksul on nii maailmas kui ka Eestis pidevalt kasvanud kultuuripärandi ülespildistamine ja digiteerimine (vt ka P. Tinitsa näidisuurimust korpuste kasutamisest digihumanitaarias). Skannitud või digikaameraga pildistatud pildifailid on oma olemuselt küll digitaalsed, aga sageli mitte masinloetavad, kuna arvuti jaoks ei ole pildifailidele jäädvustatud tekst muud kui hulk teatud väärtusega piksleid. Seega on vaja meetodeid, millega saada piltidelt kätte meid huvitav tekst.

4.3.2.1. Trükitekstide digiteerimine

Tekstide digiteerimise esimene etapp on tavaliselt füüsilise teksti ülespildistamine või skannimine. Selleks, et tekste oleks üldse võimalik edasi töödelda, on väga oluline, et loodud failid oleksid kõrge kvaliteediga. Ehkki skannida saab tänapäeval ka telefoniga, on suuremate tekstikogude skannimisel mõistlik kasutada võimsamaid skannereid, mis võimaldavad skannida korraga või ridamisi palju dokumente ning tagavad kujutiste parema kvaliteedi.

Skannitud pildifailid on tavaliselt **rasterkujul**, levinumad rasterfailide vormingud on JPEG, PNG ja TIFF. Selliste failide puhul on oluliseks kvaliteedi näitajaks nende **lahutusvõime** ehk resolutsioon, mida väljendatakse digitaalkujutiste puhul kas kuva suuruse (nt 800 × 800 pikslit) või pikslitihedusena *ppi* (ingl *points per inch*), skannimisel aga tavaliselt punktitihedusena *dpi* (ingl *dots per inch*). ***Dpi***

iseloomustab seda, kui palju trükitud „tindipunkte" suudab skanner ühe tolli kohta salvestada. Selleks, et säilitada pildifailidelt olulisi detaile, nt fondi suurust, teksti vormistust, käekirja elemente, on vaja piisavalt suurt lahutusvõimet, et pildifailis sisse suumides detailid ei kaoks. Tuleb aga meeles pidada, et mida suurem on rasterfailide lahutusvõime, seda rohkem ruumi nad salvestusseadmel võtavad. Seega tuleb leida kompromiss piisava lahutusvõime ja pildifaili suuruse vahel. Sobiv lahutusvõime sõltub originaaldokumendi suurusest ja detailsusest: mida rohkem ja väiksemat teksti dokument sisaldab, seda suurem võiks olla lahutusvõime. Tekstidokumentide digiteerimise puhul on standardsoovituseks vähemalt 300 *dpi*-d.

4.3.2.2. Tekstituvastus

Skannitud tekstid on niisiis digitaalsed, aga pikka aega ei olnud tekst skannitud dokumentides masinloetav. Tänapäevased skannerid võimaldavad dokumente skannida juba koos automaatse tekstituvastusega (PDF-vormingus). Digiteeritud pildifailidelt automaatset teksti tuvastamist ja masinloetavale kujule viimist nimetatakse **tärktuvastuseks** (ingl *optical character recognition*, OCR). Tärktuvastus töötab kõige paremini trükitekstidel ning võimaldab teha suure hulga tekste korraga masinloetavaks.

Enamasti sisaldavad tärktuvastatud tekstid siiski üksjagu vigu (vt P. Tinitsa näidisuurimust korpuste kasutamisest digihumanitaarias), olgugi et tärktuvastustarkvara on ajaga läinud järjest paremaks ja täpsemaks. Joonisel 4.5 on näide Rahvusraamatukogu DIGAR-i digiarhiivis leiduvast digiteeritud ja tärktuvastatud tutvumiskuulutusest, mis ilmus ajalehes Postimees 1934. aastal. DIGAR võimaldab ligipääsu suurele hulgale digiteeritud väljaannetele (sh e-raamatutele, ajakirjadele, ajalehtedele). DIGAR-i Eesti Artiklite portaalis[16] (DEA) saab omakorda teha päringuid kõigist Eestis ilmunud või välismaal eesti keeles avaldatud digitaalselt loodud või digiteeritud ja tärktuvastatud ajalehtedest (alates 1811. aastast) ning muudest perioodikaväljaannetest (alates 2017. aastast). Otseligipääsu tekstidele saab taotleda Rahvusraamatukogu Digilabori kaudu[17].

DEA tärktuvastatud tekstid on toimetamata (v.a autorite nimed, artiklite pealkirjad ja pildiallkirjad). Tärktuvastud tekste on võimalik aga ka käsitsi või automaatselt **järeltoimetada**, näiteks luues sõnastiku lubatud sõnavormidest, mis tekstides võivad esineda. Kuna käsitsi järeltoimetamine (nagu korpuste koostamisel igasugune käsitsi töö) võib olla küllaltki ajamahukas protsess, on seda tehtud ka n-ö **ühisloome** või rahvahanke (ingl *crowdsourcing*) meetodil, kasutades selleks vabatahtlikke huvilisi. Samuti on hakatud katsetama **suurte keelemudelite** (ingl *large language models*) võimalusi tärktuvastatud tekstide järeltoimetamiseks (nt Boros jt 2024).

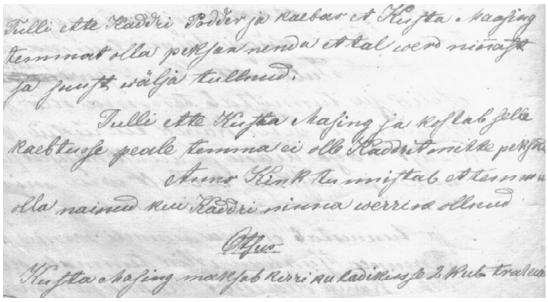
[16] https://dea.digar.ee/

[17] https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/

Joonis 4.5. Tärktuvastatud kuulutusi ajalehes Postimees nr 18, 19. jaanuar 1934 (DIGAR[18])

4.3.2.3. Käsikirjade digiteerimine

Ühisloomet on kasutatud ka käsikirjaliste tekstide masinloetavaks tegemisel. Näiteks on Rahvusarhiivil[19] mitu sellist ühisloomeprojekti, mille käigus on vabatahtlikud aidanud digiteerida muu hulgas 19. sajandi käsikirjalisi vallakohtuprotokolle ning vabadussõjas osalenud sõdurite andmetega säilikuid, laiendades nõnda oluliselt arhiivimaterjalide päringuvõimalusi.



Tulli ette Kaddri Podder ja kaebas et Kusta Maasing temmat olla peksan nenda et tal werd ninnast ja suust wälja tullnud.

Tulli ette Kusta Masing ja kostab selle kaebtusse peale temma ei olle Kaddrit mitte peksanud.

Anno Kink tunnistab et temma olla nainud kui Kaddri ninna werrine ollnud.

Otsus. Kusta Masing maksab kirriku ladikusse 2 Rubla trahwi.

Joonis 4.6. Ühisloome käigus digiteeritud Pornuse valla kohtuprotokoll nr 5, 19. juuli 1868 (Rahvusarhiiv[20], EAA.4710.1.8)

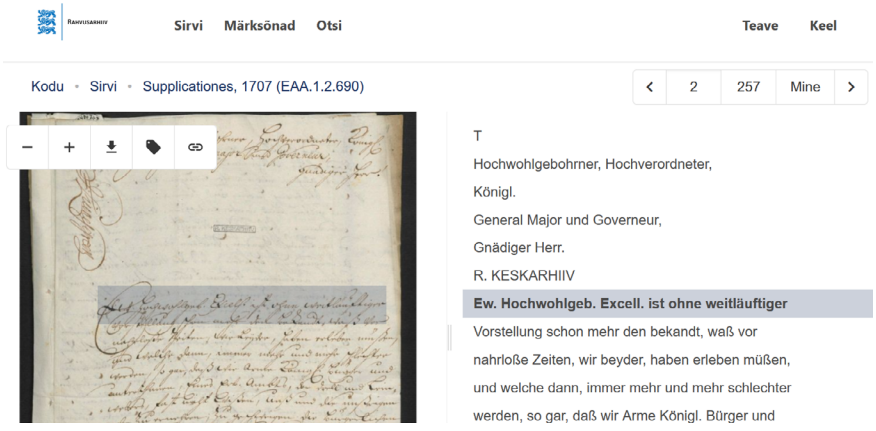
[18] https://dea.digar.ee/?a=d&d=postimeesew19340119.2.58.6&srpos=16&e=-------et-25postimeesew-1--txt-txIN%7ctxTI%7ctxAU%7ctxTA-tutvub----------

[19] https://www.ra.ee/kulastajale/uhisloome/

[20] https://www.ra.ee/vallakohtud/index.php/record/view?id=2298&ru=6NnJ3u

Loomulikult sisaldavad selliselt digiteeritud tekstid siiski ka vigu, kuna vanade arhiiviallikate lugemine ja mõistmine nõuab sisestajalt erinevaid oskusi ja kogemusi (näiteks vanemate kirjaviiside ja eri keelekujude tundmist, aga ka tehnilist täpsust ja juhendite lugemise oskust).

Käsikirjade digiteerimiseks võib kasutada ka tärktuvastust, ent tavaliste OCR-tarkvarade tuvastustäpsus on käsikirjadega väiksem kui trükiteksti puhul, eriti juhul, kui käekiri erineb oluliselt trükitekstides kasutatud kirjatüüpidest. Seetõttu on olemas eraldi programmid, mis on mõeldud spetsiaalselt **käsikirjaliste tekstide tuvastamiseks**. Neist üks tuntumaid on Transkribus[21]. Programmis tuleb esmalt hulk tekste üles laadida ning need käsitsi digiteerida (alustuseks võib tekstidel kasutada ka mõnd tarkvara pakutavatest olemasolevatest tuvastusmudelitest ning selle väljundit käsitsi järeltoimetada). Selliste algsete pildifailide ja käsitsi digiteeritud dokumentide paaride (nn treeningmaterjali) põhjal treenitakse masinõppe mudel, mis õpib, kuidas piltidelt tuvastatud struktuurid ette antud tähejärjenditega kokku käivad, ning püüab seejärel ülejäänud dokumentidest õpitu põhjal ise tekstijärjendeid ära tunda. Mida parem on digiteeritud dokumentide kvaliteet, mida selgem on käekiri, mida suurem ja ühetaolisem on treeningmaterjal ning mida paremini ülejäänud dokumentides kasutatud käekiri ja muu vormistus treeningmaterjaliga sarnaneb, seda parema tulemuse automaatsel tekstituvastusel saab. Mida varieeruvam on käekiri, seda enam tekste tuleb treeningmaterjali lisada. Lisaks tekstituvastusele võimaldab programm tekste ka märgendada, teha digiteeritud tekstidest päringuid ning oma digiteeritud tekstikorpuse ka alla laadida (nt DOCX-, PDF-, TXT- või XML-vormingus). Transkribusega digiteeritud allikakogusid leiab ka näiteks Rahvusarhiivi lehelt[22].



Joonis 4.7. Näide Transkribusega digiteeritud palvekirjast Eestimaa rootsiaegsele kindralkubernerile (Rahvusarhiiv[23], EAA.1.2.690)

[21] https://www.transkribus.org/

[22] https://rahvusarhiiv.transkribus.eu/

[23] https://rahvusarhiiv.transkribus.eu/document/513039/pages/2

4.3.2.4. Tekstide normaliseerimine ja märgendamine

Tekstituvastusele võib järgneda vajadus tekste normaliseerida ja märgendada. **Normaliseerimise** (vt lähemalt alapeatükk 4.4) eesmärgiks võib olla lihtsalt sõnakujude varieerumise vähendamine (nt *üles* ja *ülesse*, *viit* ja *viite*, *tegelt* ja *tegelikult, vana* ja *wana*), selleks et võimaldada teksti sisust lihtsamate päringute tegemist. Teinekord võib normaliseerimise vajaduse tingida aga soov kasutada tekstikogu analüüsiks tänapäeva keelel treenitud automaatse analüüsi vahendeid. Näiteks 19. sajandi eestikeelsetes tekstides on keele kirjapanemise traditsioon veel täielikult välja kujunemata, mistõttu sisaldavad tekstid palju varieerumist. Nõnda sõltus vallakohtuprotokollide tekstide kirjapilt palju näiteks kirjapanijast ja tema taustast, aga ka uue ja vana kirjaviisi võistlemisest, põhja- ja lõunaeesti keele erinevustest või murdelistest eripäradest. Selliste suurte tekstikogude käsitsi **märgendamine**, näiteks sõnavormidele grammatilise info lisamine oleks väga aeganõudev, automaatanalüüsi vahendite kasutamiseks aga oleks vaja, et tekstid näeksid rohkem tänapäeva keele moodi välja. Märgendamine võib tähendada aga ka dokumentide metaandmete ja struktuuri märgendamist (vt ptk 3 „Märgendamine"), muu hulgas võib olla kasulik ära märkida, kus algavad ja lõppevad algse dokumendi tekstiread ja leheküljed, kus paiknevad tabelid jms.

4.3.3. Kõnekorpuse loomine

Nagu korpusi üldiselt, on ka kõnekorpusi väga erinevaid. Samuti saab kõnekorpusi koostada erinevatel viisidel. Valikud sõltuvad siinjuures paljuski sellest, mis laadi kõnesalvestusi kasutatakse:

- Kas salvestused tehakse ise või on need juba olemas?
- Kas korpus sisaldab spontaanset või loetud kõnet?
- Kas korpus sisaldab ainult heli või ka videot?

4.3.3.1. Salvestamine

Kui korpuse loomist peab alustama salvestamisest, siis mõistagi tasuks salvestused teha võimalikult hea kvaliteediga. Kõnematerjali salvestamiseks on tehniliselt parimad tingimused salvestusstuudios. Kui eesmärk on koguda spontaanset kõnet välitööde tingimustes, on soovitav kasutada kvaliteetset **diktofoni** (väga laialt on kasutusel nt Zoom H2n). Kui vähegi võimalik, võiks vältida salvestamiseks telefoni kasutamist.

Hea oleks kasutada ka eraldi **mikrofone**: selleks, et mitme osalejaga vestluses oleks iga kõneleja jutt selgelt eristatav ka pealerääkimiste korral, oleks parim lahendus kasutada pea- või lipsunõelamikrofone, aga see võib osutuda keeruliseks, sest selline seadmete komplekt on küllaltki suur ja kaasas kandmiseks kohmakas ning kallis. Kui kasutada diktofoni või üht suuremat mikrofoni, siis võimalusel võiks

selle panna statiivile. Kui panna see kõneleja ees olevale lauale, kanduvad näiteks sõrmedega lauale trummeldamine või lauajalgade müksamine ja muud kolksud ja müksud mikrofoni üle ja on salvestuses palju lärmakamad, kui kohapeal tunduda võib.

Salvestamiskeskkonna valik sõltub mõnevõrra sellest, mida ja kuidas on plaanis edasi uurida. Peab tegema kompromissi helikvaliteedi ja situatsiooni loomulikkuse vahel. Helikvaliteedi nimel oleks hea välistada igasugune taustamüra: valida vaikne ruum, panna kinni muusika või taustal mängiv televiisor, mürisev arvuti, tiksuv kell jms. Kui situatsiooni spontaansus selle all ei kannata, siis võiks paluda kõnelejatel mitte kolistada lauanõudega ja istuda paigal, mitte toimetada ringi. Hea oleks valida ruum, kus on raamaturiiuleid, pehmet mööblit ja tekstiile, sest lagedas ruumis tekib tugev kaja. Kui salvestada õues, siis tuleb kindlasti kasutada mikrofoni tuulekaitset, sest ka keskmiselt tuulise ilmaga võib salvestus olla täis tuulekohinat, kui tuul peaks otse mikrofoni puhuma.

Loetud kõnet on kõige mugavam salvestada otse arvutisse tarkvaraga, mis kuvab loetava teksti ekraanile ja salvestab korraga kuvatud laused või lõigud eraldi failidesse ja tekitab korrastatud nimedega failisüsteemi. Selline programm on näiteks SpeechRecorder[24] (vt ka EKI Salvestajat[25]). Lisaks tarkvarale on vaja ka kvaliteetset mikrofoni ja helikaarti. Mikrofoni ja helikaardina saab kasutada ka diktofoni, mida on võimalik USB-liidese kaudu arvutiga ühendada.

Heli kvaliteeti silmas pidades oleks kõige parem salvestada heli WAV-vormingus 16 biti ja 44,1 kHz kvantimissagedusega. Selle kvaliteediga salvestus võtab salvestusseadmel ruumi umbes 5 MB minuti kohta. Kui faili maht on probleem ja kvaliteet ei ole nii oluline, siis võib salvestada ka MP3- või muus pakitud vormingus. Pakitud audiofaili vormingud (levinumad MP3 ja AAC) tihendavad nutikalt helifaili mahtu, eemaldades ja koondades helisignaalist komponente, mis tõenäoselt jäävad inimkõrvale märkamata. Seda nimetatakse kadudega pakkimiseks (ingl *lossy data compression*), mis tähendab, et algset signaali taastada ei ole võimalik. Seetõttu oleks parem vähemalt esmalt salvestada pakkimata WAV-vormingus ja hiljem vajadusel faile väiksemaks konvertida. Nagu öeldud, võiks võimalusel vältida telefoniga salvestamist, aga kui seda siiski teha tuleb, siis võiks vaadata üle salvestusseaded. Vanemates telefonides on sageli olnud kasutusel AMR-vorming, mis on väga tugevasti pakitud.

Kui on plaanis uurida kõne multimodaalseid aspekte, siis oleks vaja salvestada ka **videopilti**. Video salvestamine on salvestuskoha mõttes üksjagu nõudlikum, sest kui seda teha välitööde kontekstis, siis võib jääda salvestuses näha inimesi, esemeid või ruume, mida ei soovita avalikult näidata. Helikvaliteedi osas kehtivad samad reeglid nagu ainult heli salvestamise puhul: kuna õues võib olla tuuline, tuleks kasutada mikrofonil tuulekaitset ning parema tulemuse saab siis, kui

[24] https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

[25] https://koneveeb.ee/minu-haal/

igal kõnelejal on oma mikrofon. Kui videokaameral ei ole välise mikrofoni jaoks sisendit, siis on võimalik kasutada heli salvestamiseks eraldi salvestajat ja pärast heli- ja videosalvestus omavahel joondada. Kui salvestada mitme kaameraga ja/või heli eraldi seadmega, tasuks kasutada filmiklappi või teha kätega plaks salvestuse alguses, mille järgi pärast salvestused sünkroniseerida. Pikema salvestuse puhul võiks sünkroniseerimise plakse korrata salvestuse kestel (nt iga 15 tagant), sest kaamera kaadrisagedus ei pruugi olla päris ühtlane ja pikema salvestuse korral võib eri salvestusseadmete vahel tekkida küllaltki suur nihe. Kui kõnelejaid on mitu, siis tasuks kasutada mitut kaamerat. Kindlasti tuleks kasutada kaamera statiivi. Kaamera võiks paigutada kõneleja suhtes nii, et otsevaates oleks kaadris nähtaval vähemalt kõneleja ülakeha: pea ja käed (ka siis, kui ta neid liigutab). Võimaluse korral võiks kasutada mitut kaamerat erinevatest nurkadest. Kui kõnelejaid on mitu, siis võiks iga kõneleja olla kaetud otsevaates ühe kaameraga ja võimaluse korral võiks üks kaamera jälgida ka teisi kõnelejaid külgvaates või üldplaanis. Need detailid sõltuvad muidugi sellest, mis laadi uurimistöö jaoks korpust koostatakse: välitöödel murdekeelt salvestades ei huvita meid tõenäoliselt näo, kulmude või käte liikumine sel määral, kui multimodaalset suhtlust uurides.

Video salvestuskvaliteedi puhul peab arvestama seda, et failimahud on päris suured. Kõige levinum failivorming on MP4 ning pakkimata vormingut ei saa siin soovitada seetõttu, et sellised failid ei mahu hästi kuhugi ära. Minimaalselt võiks kasutada HD-vormingut, ehkki tänapäeva standardite järgi võib seda pidada juba liiga madalaks (parem oleks 4K). Enamik tavakasutuse kaameraid salvestavad vaikimisi ~30 kaadrit sekundis (ingl *frames per second* ehk *fps*), kuigi see tähendab, et kaadrid on umbes 33 ms vahega ja kõnes jõuab selle ajaga nii mõndagi juhtuda. Võimaluse korral võiks salvestada näiteks kaadrisagedusega 60 *fps*. Oluline on ka see, et ruum oleks piisavalt valgustatud selleks, et säriaeg oleks võimalikult väike. Vastasel juhul jääb pilt kiiremate liigutuste ajal udune.

4.3.3.2. Transkribeerimine

See, mida salvestatud (heli)materjaliga edasi teha, sõltub paljuski uurimisküsimustest ja mõnevõrra ka materjali olemusest. Mõnel juhul piisab lihtsalt teksti välja kirjutamisest ehk litereerimisest, teisel juhul oleks vaja tekst lisaks helifailiga ajaliselt joondada ning märkida täpsemalt hääldust või mingit muud täiendavat infot. Teksti litereerimist moel, milles lisatakse rohkem infot ka hääldus või muude kõne aspektide kohta, nimetatakse tavaliselt **transkribeerimiseks**, ent sageli kasutatakse kaht mõistet sünonüümsena.

Spontaanse kõne transkribeerimiseks võiks proovida esimese sammuna **automaatset kõnetuvastust**. Eestikeelse teksti tuvastamiseks saab kasutada TTÜ kõnetehnoloogia labori rakendust Tekstiks[26]. Rakendus võimaldab transkribeeritud teksti hõlpsalt järeltoimetada ning annab väljundi kas lihtsas SRT-vormingus

[26] https://tekstiks.ee/

subtiitritena, TRS-vormingus, kus tekst on korrastatud ajatemplitega kõnevoorudeks, või JSON-vormingus, kus tekst on kõnevoorude, lausete ja sõnade kaupa varustatud ajatemplitega, lisaks on selles kõnelejatuvastuselt saadud info kõnelejate kohta. Alla saab laadida ka DOCX-vormingus transkriptsiooni. Kõnetuvastussüsteemi on võimalik ka enda arvutisse üles seada[27], ent see nõuab pisut tehnilisi oskusi. Ehkki korpuste loomiseks vähem relevantne, saab reaalajas subtiitrite loomiseks kasutada samal kõnetuvastusmudelil põhinevat veebirakendust[28] või Google Chrome'i laiendust[29].

Automaatset kõnetuvastust võib olla vajalik ka **järeltoimetada** ehk parandada transkriptsioonides valesti tuvastatud sõnu. Nagu tärktuvastust, võib ka automaatse kõnetuvastuse väljundit järeltoimetada kas käsitsi või kasutades reeglipõhist, loendipõhist või statistilist automaattoimetamist. Mõnel juhul võib olla järeltoimetamisest lihtsam ja kiirem **käsitsi transkribeerida**. Seda näiteks siis, kui salvestus on halva kvaliteediga, kui salvestuses on palju pealerääkimisi või kui salvestatud keelekasutus erineb millegi poolest üldkeele standardist (näiteks räägitakse murdes, kasutatakse palju koodivahetust vmt). Käsitsi transkribeerimiseks või järeltoimetamiseks võib kasutada olemasolevat tarkvara. Kui tekst peaks olema **aegjoondatud** ehk helifaili ajajoonega seotud (igal transkribeeritud tekstilõigul on helifailis kindel algus- ja lõpuaeg), siis sobib kasutada näiteks programme ELAN[30] (vt J. Wilburi näidisuurimus korpuslingvistika kasutamisest ohustatud keelte uurimisel) või Praat[31]. Kui salvestusel on ka video, siis sobivad ELAN või Nova[32]. Nimetatud tarkvara sobib hästi ka tekstide märgendamiseks, kuna lubab lisada lisaks transkriptsioonikihile veel erinevaid märgenduskihte, kus võib märgendada näiteks häälikuid, võõrkeelte kasutust, naeru, žeste ja palju muud, mis suulise kõne uurimisel vajalikuks võib osutuda.

Automaatne kõnetuvastus väljastab tavaortograafias transkriptsiooni, ent korpustes võivad olla kasutusel eri **transkriptsioonisüsteemid**. Lisaks tavaortograafiale võib kasutada hääldust täpsemalt kirjeldavaid transkriptsioonisüsteeme, nt SAMPA (ingl *speech assessment methods phonetic alphabet*), IPA (ingl *international phonetic alphabet*) või soome-ugri keelte foneetiline transkriptsioon FUT (ingl *Finno-Ugric transcription*), mida kasutatakse traditsiooniliselt eesti murrete kirjapanekuks. Vestlusanalüüsi raamistikus kasutatakse omakorda tavaliselt Jeffersoni transkriptsiooni, mida on eri keeltele kohandatud (nagu nt TÜ suulise kõne

[27] https://github.com/taltechnlp/est-asr-pipeline

[28] https://eestiasr.vercel.app/

[29] https://chromewebstore.google.com/detail/estonian-asr-captions-gen/mkpaijfcijkdihjcipkfadinmmjmkadk?utm_source=item-share-cb&pli=1

[30] https://archive.mpi.nl/tla/elan

[31] https://www.fon.hum.uva.nl/praat/

[32] https://github.com/hcmlab/nova

korpuses[33]). Transkriptsioonisüsteemi valik sõltub sellest, mida ja kui täpselt on tarvis märkida.

Loetud kõne puhul on tavaliselt tekst eelnevalt olemas ning pärast salvestust on vaja see kõnega kokku viia. Seda saab teha käsitsi, ent eesti keele jaoks saab kasutada ka TTÜ **autosegmenteerijat**[34] (ingl *forced aligner*). Kui tegemist ei ole eesti keelega, siis võib proovida WebMAUS-i[35], kus on eri keeltele (sh eesti) treenitud mudelid. Autosegmenteerija töötab automaatse kõnetuvastuse mudeli baasil ning selle sisendiks on tavaortograafias tekst. Mõlemad nimetatud autosegmenteerijad annavad väljundiks segmentatsiooni Praati TextGrid-vormingus, kus on tavaortograafias kiht sõnapiiridega ning SAMPA transkriptsioonis kiht häälikupiiridega. Peab arvestama, et autosegmenteerija lähtub ortograafilisest kirjaviisist ja nagu ingliskeelne termin *forced alinger* osutab, siis pannakse teksti põhjal oodatud segmendid helilainega jõuga kokku. Selle tulemusel märgitakse suvaliselt (WebMAUS) või ei märgita üldse (TTÜ) 1) nähtusi, mida tavaortograafia ei kajasta (välde, palatalisatsioon), ning 2) juhtumeid, kus heli ja tekst ei lähe kokku (näiteks kui esineb reduktsiooni või kui on loetud valesti).

4.4. Praktilised aspektid: vormingud, märgistik, normaliseerimine

Heaks tavaks on hoida oma korpust operatsioonisüsteemist sõltumatus ja avatud **dokumendivormingus**. See tähendab, et korpuse kasutamiseks ei ole vajalik eelnevalt osta mingit eraldi tarkvara ning kasutatud vorming peaks erisuguste rakendustega kergesti ühilduma. Sellised vormingud on näiteks TXT ja XML (ja selle erinevad rakendused, nt TEI või EAF). Üha enam kasutatakse korpustes ka JSON-vormingut, eeskätt seetõttu, et see on XML-iga võrreldes kompaktsem, st võtab vähem ruumi ning on kiiremini töödeldav. Kui korpust pole plaanis märgendada, piisab täiesti TXT-vormingust. DOC ja selle analoogid korpuse vorminguks enamasti ei sobi.

Ebasobivas vormingus tekste saab **teisendada** sobivasse vormingusse kas käsitsi, kui korpusefaile on vähe, kasutades mõnd teisendajat, nt AntFileConverter[36] või pdf2txt[37], või hoopis programmeerimiskeelte (nt Python ja R) võimalusi[38].

---

33 https://keeleressursid.ee/et/148-suulise-kone-transkriptsioon

34 https://bark.cs.taltech.ee/autosegment2/

35 https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

36 https://www.laurenceanthony.net/software/antfileconverter/

37 https://www.pdf2txt.com/

38 PDF-vormingust on kaks tüüpi: digitaalselt loodud „tõelised" PDF-failid ja skannimise teel loodud „pildilised" PDF-failid. Viimaseid ei saa eelpool nimetatud tarkvaraga teisendada, vaid

Kindlasti võiks aga vältida veebipõhiseid teisendajaid, kuhu peab oma korpusefaile üles laadima, juhul kui tekstid sisaldavad isikuandmeid vm-d tundlikku infot (vt ka alapeatükki 4.5), kuna sellisel juhul puudub kontroll selle üle, millistesse serveritesse meie failid satuvad, kes nendele ligi pääsevad ja kes neid edasi võivad jagada.

Samuti tuleb kontrollida tekstide **märgistikku** ehk kodeeringut (ingl *character encoding*). Märgistik määrab, millist märki mingi arvutimälus olev number tähistab. Kuna arvutite ajaloo jooksul on olnud kasutusel erinevaid märgistikke, võib sama bait või baitide järjend[39] tekstis välja paista eri (tähe)märgina, olenevalt sellest, millises märgistikus programm arvab selle kodeeritud olevat. Tänapäeval levinuim märgistik on UTF-8 (ingl *universal coded character set + transformation format – 8-bit*), mis on üks Unicode'i standardeid, st enamik tekste luuakse tänapäeval selles märgistikus ning seda eeldavad ka korpuste töötlemise ja analüüsi vahendid. Probleemid märgistikuga kerkivad pigem varasemate tekstide puhul, siinkohal mõeldakse „varasema" all ka umbes 20 aastat vanu tekste. Nii leiab 2000. aasta 19. detsembri Õhtulehe veebiversioonis pealkirja „Šokolaad saab euromõõdud" või sama lehe 18. mai 1999. aasta numbrist pealkirja „Garaa©iomanik maadleb lukuga", kus on selgelt näha, et tekst on mingis muus kodeeringus, kui veebilehe esitaja selle arvab olevat. Märgistiku kontrollimiseks ja vahetamiseks on veebis mitmeid rakendusi; samuti saab seda teha lihttekstiga töötamiseks mõeldud tekstiredaktorites (Windowsis nt EditPadLite või Notepad++, UNIX-is/Linuxis Notepadqq, Maci jaoks on samaväärsed programmid Brackets, Sublime Text ja Textmate); paljudes programmeerimiskeeltes on olemas ka kodeeringu vahetamise käsud (nt *iconv* UNIX-i/Linuxi keskkonnas[40]).

Kui oma korpuse tekstide keel erineb suurel määral tänapäeva eesti normkirjakeelest ja korpusele soovitakse rakendada automaattöötluse vahendeid, siis tasub kaaluda **tekstide normaliseerimist**, st tekstide sõna-sõnalist tõlkimist tänapäeva kirjakeelde. Seda „tõlget" saab siis näiteks automaatselt lemmatiseerida, morfoloogiliselt ja süntaktiliselt märgendada jne. Kuna on palju erinevaid viise, kuidas keel saab tänapäeva normist erineda (vanem keelekasutus, vana kirjaviis, murdekeel, erinevate veebikogukondade kirjaviis ja keelekasutus jne), tuleks luua iga sellise keelevariandi jaoks oma automaatne normaliseerija. Kui aga uuritava korpuse keel ja kirjaviis erineb tänapäevasest vähesel määral ja erinevused on süstemaatilised, piisab tegelikult väheste reeglite komplektist, et automaatanalüüsi tulemusi palju paremaks muuta. Nii piisaks 19. sajandi lõpu tekstide puhul tähemärkide *w* ja *ß* asendamisest vastavalt *v* ja *s(s)*-ga, väikestest reeglistikest, nt mineviku kolmanda isiku mitmuse verbivormide *ivad*-lõpu asendamiseks *id*-lõpuga (*tulivad* → *tulid*), sõnade *auu* ja *nõuu* vormide viimisest kujule *au* ja *nõu*, et tulemuseks olev tekst

selleks tuleb kasutada tärktuvastust ehk OCR-tarkvara (vt alapeatükki 4.3.2.2).

[39] Bait on arvuti infoühik, koosneb kaheksast bitist. UTF-8-s esitatakse osa tähemärke (ASCII) ühe baidina, osa (nt eesti täpitähed) kahe baidina.

[40] https://www.tecmint.com/convert-files-to-utf-8-encoding-in-linux/

oleks juba paremini analüüsitav tänapäeva keele jaoks mõeldud keeletehnoloogia vahenditega. Selliseid asendusi saab teha nii lihttekstiredaktorite asendusfunktsioonide (ingl *replace*) abil kui ka **skriptidega** ehk mingis programmeerimiskeeles kirjutatud käsujadadega. Viimasel juhul tekib automaatselt dokumentatsioon sellest, mida millega asendati. Ükshaaval asendusi tehes tuleks aga kindlasti kõik tehtud asendused eraldi ka dokumentatsiooni kirja panna (vt alapeatükk 4.6).

Normaliseerides tuleb muidugi hoolt kanda selle eest, et ka teksti originaalkuju säiliks. Automaatsel normaliseerimisel käsitletakse normaliseeritud sõnakuju lihtsalt ühe märgenduskihina, tekstiredaktori asendusfunktsiooni abil normaliseerimisel tuleb enne normaliseerimist teha failist koopia.

Näide. Allpool on esitatud Räpina kihelkonna Kahkva vallakohtu protokolli digiteering (trelliga # algavad read) ja iga lause „tõlge" tänapäeva kirjaviisi. Plussmärgid tähistavad sõnapiire, mis tänapäeval puuduksid, st sõnaühendeid, mida tänapäeval kirjutataks liitsõnadena. Selle teksti normaliseeris inimene, mitte masin ja märk ¤ tähistab sõnavormi, mida pole osatud normaliseerida.

```
#Kahkwa kohtumajan sel 24 Märzil 1869.
Kahkva kohtumajas sel 24 märtsil 1869.

#Kaiwas Ewa Lepland, et temma Wirksu kõrtsi mehhe Karl
Naruskowi käest 30 rahha saada ollu ja sel 17 Märzil
mõisast tullu om Ewa Wirksu kõrtsi sisse lännu ja üts
toob wina wõtnu ja ütelnu: nüüd jääb se win to wõlla
eest ja om wälja tullu. Sis om Karli omma poijaga perra
tullu ja Ewad rindu piddu kisknuwa ming läbbi om Ewa
sõlg ärra murtus saanu.
Kaebas Eva Lepland, et temal Virksu kõrtsi+ +mehe Karl
Naruskovi käest 30 raha saada olnud ja sel 17 märtsil
mõisast tulnud on Eva Virksu kõrtsi sisse läinud ja
üks toop viina võtnud ja öelnud: nüüd jääb see viin too
võla eest ja on välja tulnud. Siis on Karli oma pojaga
perra tulnud ja Evat rindu+ +pidi kiskunud mis+ +läbi
on Eva sõlg ära murtud saanud.

#Ette tulli tunnistaja Wido Kübbar ja ütel, et Ewa om
tobi wina wõtnu ja ütelnu Karlile: nüüd olleme tassa.
Sis om Ewa wälja lännu , sis om Karli perra lännu ja
Ewa pääst rätti kisknu, sis om temma ütelnu: ärge tehke
se assi lät halwaste. Ja Jaan Mit om se wina mes Ewa
om wõtnu jagganu.
Ette tuli tunnistaja Vido Kübar ja ütles, et Eva on
toobi viina võtnud ja öelnud Karlile: nüüd oleme tasa.
```

```
Siis on Eva välja läinud , siis on Karli perra läinud
ja Eva peast rätti kiskunud, siis on tema öelnud: ärge
tehke see asi läheb halvasti. Ja Jaan Mitt on see viina
mis Eva on võtnud jaganud.

#Kohhus mõist, et Karli peab 1 Rubla trahwi waeste
ladiku masma ja Ewa 34 tunnil türma, et rismise modu
on wina wõtnu ja kõrtsi rahwale ärra jodnu. Ja et Ewa
ütel, et temal olles rahha särgi karmanin ollu, ei wõi
se perrast õige olla, et särgi karman katski om, kos
mitte rahha panda ei wõi.
Kohus mõistis, et Karli peab 1 rubla trahvi vaeste+
+laadikusse maksma ja Eva 34 tunniks türmi, et rii-
sumise moodi on viina võtnud ja kõrtsi+ +rahvale ära
jootnud. Ja et Eva ütles, et temal oleks raha särgi
karmanis olnud, ei või see+ +pärast õige olla, et särgi
karman katki on, kus mitte raha panna ei või.

#Ent sai se mõistetetu Rubla Ewale antus, et tema ko
ütelb, et temma sõlg olles ka klopi litsutu.
Ent sai see mõistetud rubla Evale antud, et tema ¤
ütleb, et tema sõlg oleks ka ¤ litsutud.
```

4.5. Eetilised ja õiguslikud aspektid

Korpuse koostamisel puutume kokku nii autoriõigusi kui ka isikuandmete töötlemist puudutavate küsimustega.

Kui kogume korpusesse suulisi või kirjalikke tekste, siis piiravad **autoriõigused** seda, kui palju või mis kujul võime oma korpust teistele kättesaadavaks teha. Vastavalt autoriõiguse seaduse[41] §-le 19 on andmekaeve ja teadustöö erandi raames lubatud tekste kopeerida ja töödelda. See tähendab, et võime oma isiklikuks teadustööks näiteks internetis olevaid materjale ilma täiendavate litsentsilepinguteta kasutada. See õigus ei laiene aga automaatselt korpuse avaldamisele avalikus otsimootoris või avatud ligipääsuga repositooriumis.

Kirjalike tekstide puhul kuuluvad õigused teksti autorile ja olenevalt lepingust võivad kuuluda ka kirjastajale. Heli- või videosalvestiste puhul on osapooli tõenäoliselt rohkem: näiteks taskuhäälingu või raadiosalvestuse puhul on peale kõneleja/esitaja autoriõigus ka heli- ja/või video režissööril. Samuti võidakse salvestuses

[41] https://www.riigiteataja.ee/akt/129062022016?leiaKehtiv

esitada muusikat, millel on omakorda esitaja, autor, kirjastaja ja muid osapooli, kellel käest peaks küsima luba juhul, kui neile kuuluvaid teoseid edasi soovime levitada. Seega tuleks leida kompromiss ühelt poolt teaduses nõutava andmete avalikkuse ja teiselt poolt autorite õigusi kaitsvate seaduste vahel.

Teine osa piirangutest on seotud **isikuandmete kaitsega** ja see puudutab eriti suulisi korpusi. Isikuandmete hulka kuuluvad tüüpiliselt näiteks nimi, aadress, meiliaadress, sünniaeg, aga ka terviseandmed, välimus ja hääl ning nende töötlemiseks peab olema õiguslik alus. Selleks aluseks võib olla isiku teadlik nõusolek või erandjuhul avalik huvi. Keeleteaduslikku uurimistööd on võimalik põhjendada avaliku huviga (vähemalt on eesti keele uurimine Eestis avalik huvi, kui lähtuda Eesti Vabariigi põhiseadusest). Sellele võib tugineda siis, kui me kasutame juba olemasolevaid kättesaadavaid salvestisi. Kui salvestada ise kõnekorpust, siis peab keelejuhtide käest salvestamiseks ja andmete töötlemiseks kindlasti luba küsima. Kuna inimese hääl võimaldab isikut identifitseerida, siis loetakse kõnesalvestisi eriliigilisteks isikuandmeteks. Tavaliselt kogutakse kõne salvestuse juurde ka sotsiodemograafilisi andmeid, nagu kõneleja sugu, vanus, haridus, emakeel, piirkondlik ja etniline päritolu jne. Selliste andmete säilitamiseks ja töötlemiseks peab küsima isiku nõusolekut. Kõige kindlam on selleks koostada ning allkirjastada kirjalik informeeritud nõusoleku dokument[42], kus oleks täpselt kirjeldatud kes, mis andmeid ja mille jaoks kogub, kuidas neid andmeid töödeldakse, kes ja mis tingimustel andmetele ligi pääseb ning kuidas ja kui kaua ja kuidas andmeid säilitatakse.

Eetilisi küsimusi isikuandmete töötlemises tuleb kõnekorpuste puhul ette ka kõne sisuga seoses. Ehkki koostame korpusi tavaliselt keeleuurimiseks, siis võib keelematerjal ise sisaldada väga erinevat sisu. Foneetiliste uurimuste jaoks piisab sageli sellest, et palume keelejuhil ette lugeda mingeid sõnu, lauseid või tekstilõike, mis on konstrueeritud sisaldama meie uuritavaid lingvistilisi üksusi. Kui aga kogume spontaanset kõnet, siis võib selles olla ka **delikaatset infot**: inimesed räägivad oma eraelust, kolleegidest, pereliikmetest ja suhtlusringkonnast, mainides teiste inimeste nimesid jms. Sellisel juhul on heaks tavaks kasutada transkriptsioonides **pseudonümiseerimist** ehk kasutada asendusnimesid või -koode (välja arvatud juhtudel, kui viidatakse avaliku elu tegelastele, nt tuntud poliitikutele või näitlejatele), seda nii isikunimede kui ka hüüdnimede, aadresside, sünnipäevade jm-de isiku identifitseerimist võimaldavate tunnuste puhul (nt kodule lähim bussipeatus, rahvatantsurühma või jalgpallimeeskonna nimi, mille liige ollakse, vmt). Pseudonümiseerimine on eriti oluline juhul, kui tahame korpuse materjale ka avalikustada ning kasutada näiteid oma uurimistöös ja ettekannetes. Pseudonümiseeritud andmetest **anonüümseteks** muutuvad andmed sel hetkel, kui hävitame võtmefailid, mis võimaldaksid pseudonüüme algsete isikuandmetega kokku viia (**anonümiseerimine**).

[42] vt nt https://ut.ee/sites/default/files/inline-files/informeerimise_ja_teadliku_nousoleku_vormi_juhis_1.pdf

4.6. Metaandmed ja dokumentatsioon

Tundub loomulik, et loodud korpuse põhiline väärtus seisneb selles, et oleme keelematerjali kogunud, süstematiseerinud ja hoolikalt märgendanud ning loonud seeläbi võimaluse uurida ja seletada, kuidas me keelt mingis situatsioonis kasutame, kuidas mõtleme, hääldame ja suhtleme. Selleks, et loodud korpus oleks aga päriselt kasulik ressurss, on aga samavõrd oluline korpuse loomise protsessi põhjalikult dokumenteerida ja varustada korpusesse kogutud keelematerjal ka metaandmetega.

Metaandmed[43] on struktureeritud ja kindlas vormingus andmed, mis annavad infot korpuse ja selle sisu kohta. Korpustekste iseloomustavad olulised metaandmed on nende suurus (sõnades, baitides), päritolu (nt veebiaadress), nende algne vorm (suuline, kirjalik), loomisaeg, autor või kõneleja ning kui võimalik ja põhjendatud, siis ka teda iseloomustavad andmed (nt nimi, vanus, sugu, haridus, päritolu). Kui koostatud korpus sisaldab mitmesse tekstiliiki kuuluvaid tekste, siis tuleks iga teksti kohta ära märkida selle tekstiliigiline kuuluvus. Lisaks üksiktekste iseloomustavatele metaandmetele peaks olema olemas metaandmed ja/või dokumentatsioon ka kogu korpuse kohta: millal ja mis viisil see on koostatud, millised nähtused ja kuidas on märgendatud, milliseid märgendeid selleks kasutatud.

See, et metaandmed peavad olema struktureeritud, ei tähenda, et tingimata peab järgima mingit standardit, vaid seda, et metaandmed esitatakse tüüpiliselt tunnuse-väärtuse (ingl *attribute-value*) paarina. Tänu sellele on suuremate korpuste metaandmed ka masinloetavad.

> **Näide.** Allpool on esitatud META-SHARE'i repositooriumis[44] olevad metaandmed eesti keele morfoloogiliselt käsitsi ühestatud korpuse kohta[45] (vt ka ptk 2.2.4 „Eesti keele ühendkorpused: mahukaimad eesti keele digitekstide kogud").
>
> ```
> doi:10.15155/TY.001A
> Corpus type: Monolingual text corpus; Manually anno-
> tated corpus
> Availability: Available - Restricted Use
> Licence CLARIN ACA - NC
> Restrictions: Academic - Non Commercial Use, Attribution
> Distribution Access/Medium: Downloadable
> User Nature: Academic
> ```

[43] Vt ka Tartu Ülikooli raamatukogu koostatud materjale lehel https://sisu.ut.ee/andmehaldus/dokumenteerimine-ja-metaandmed/.

[44] https://metashare.ut.ee/

[45] https://metashare.ut.ee/repository/browse/corpus-of-morphologically-disambiguated-estonian-texts/f8547b0ca0d311eebb4773db10791bcf844d4793099449b894d809fa49d8b4a7/

```
Languages: Estonian
Language Script: Latin
Linguality type: Monolingual
Size: 513 000 Words; 128 Files; 300,000 Tokens
```

Sama korpuse vabatekstilist dokumentatsiooni saab lugeda lehelt https://cl.ut.ee/korpused/morfkorpus/.

Erinevus metaandmete ja **dokumentatsiooni** vahel seisneb selles, et metaandmed on struktureeritud andmed kindlas vormingus, dokumentatsioon aga nn vabatekstiline kirjeldus. Kui korpuse märgendus on tehtud automaatselt, oleks nii korpuse looja kui kasutajate seisukohalt väga kasulik, kui dokumentatsioonis oleks ka märgenduse kvaliteedi hinnang. Selleks tuleks mingi osa korpuse märgendusest käsitsi üle kontrollida ja tulemus dokumenteerida: mitu % märgendatud keelenditest on saanud vale märgendi ja millised on sagedasemad vead. See võtab küll aega, aga on korpuse kasutajale väga tarvilik teadmine.

Metaandmeid ja dokumentatsiooni oleks hea talletada ja avalikustada korpusega koos. Andmeid koos metaandmetega võib säilitada ja jagada repositooriumides ehk **andmehoidlates**, kust info korpuse olemasolu ja sisu kohta võib jõuda laiema ringi huvilisteni. Üks selline andmevaramu on näiteks DataDOI[46].

Lõpetuseks

Õpiku neljandas peatükis keskendusime oma korpuse loomisele. Kuigi eesti keele kohta on olemas suur hulk väga erinevaid korpusi (nagu nägime peatükis 2 „Eesti keele korpused"), võib meil uurijatena olla soov või vajadus luua oma spetsiifiline korpus lähtuvalt meie uurimisküsimusest või uuritavast keelenähtusest. Üks kõige levinumaid põhjuseid oma korpuse loomiseks on soov uurida spetsiifilise sihtrühma keelekasutust või spetsiifilist kasutuskonteksti. Peatükis kirjeldame korpuse koostamise põhimõtteid ning rõhutame korpuse loomise protsessi põhjaliku dokumenteerimise vajalikkust. Oluline on tagada korpuse esinduslikkus ja tasakaalustatus uuritava keelenähtuse suhtes. Korpuse koostamise puhul tuleb hoolega silmas pidada eetilisi ja õiguslikke aspekte, nagu autoriõigused ja isikuandmete kaitse, eriti suuliste ja tundlikku infot sisaldavate tekstide puhul. Kokkuvõtvalt peame tõdema, et oma korpuse koostamine on võrdlemisi keeruline protsess, mis hõlmab andmete kogumist ja korrastamist, metaandmete lisamist ja vajadusel märgendamist. Seetõttu tasub eelnevalt välja selgitada, kas ehk saab vajaliku info kätte juba olemasolevatest korpustest. Kui aga uurimistöö nõuab spetsiifilist andmestikku, võib oma korpuse loomine olla ainus võimalus.

[46] https://datadoi.ee

Lisalugemiseks

Crawford, William J. & Eniko Csomay. 2016. *Doing corpus linguistics*. New York/London: Routledge, 73–93.

O'Keeffe, Anne & Michael J. McCarthy (toim.). 2022. *The Routledge handbook of corpus linguistics*. 2nd edition. London/New York: Routledge, 11–88.

Zeldes, Amir. 2020. Corpus architecture. Magali Paquot & Stefan Th. Gries (toim.), *A practical handbook of corpus linguistics*. Cham: Springer, 49–76. https://doi.org/10.1007/978-3-030-46216-1_3.

Ädel, Annelie. 2020. Corpus compilation. Magali Paquot & Stefan Th. Gries (toim.), *A practical handbook of corpus linguistics*. Cham: Springer, 3–24. https://doi.org/10.1007/978-3-030-46216-1_1.

5. Levinumad korpuslingvistika meetodid

Liina Lindström

Nagu nägime eelmistest peatükkidest, on tänapäeva korpused väga eriilmelised, seda nii suuruse, eesmärgi, vormingute, märgenduspõhimõtete ja -kihtide kui ka kõige muu poolest, mistõttu ei ole ühte ja lihtsat viisi, kuidas korpusi kasutada. Selle asemel on hulgaliselt erinevaid keskkondi ja programme, mille abil korpustest vajalikku infot kätte saada. Lisaks võime keeleuurimisel oma tähelepanu suunata väga erinevat tüüpi keelenditele – sõnadele, morfeemidele, lausetele, konstruktsioonidele, kõnevoorudele, häälikutele, žestidele või muudele tähenduslikele üksustele, mille analüüsimiseks on samuti mitmeid erinevaid vahendeid ja meetodeid.

Korpusandmete analüüs võimaldab teha järeldusi selle kohta, mis kontekstis, kui sageli ja miks mingit keelelist vahendit kasutatakse, kuidas see kombineerub muude keelenditega või mis tähendust või pragmaatilist funktsiooni need keelendid kannavad. Korpuste kasutamise suur eelis on, et saame lisaks kvalitatiivsetele andmetele (kuidas keelendit mingis konkreetses ümbruses kasutatakse) teada ka keelendite esinemissageduse, mis võimaldab meil hinnata nende levikut ja kasutusdünaamikat, analüüsida varieerumist ja teha muid järeldusi, mis on kasutussagedusega seotud.

Selles peatükis anname sissejuhatava ülevaate, mis vahendeid korpuste kasutamiseks on loodud ja millised on kõige tüüpilisemad analüüsimeetodid, mida korpuslingvistikas kasutatakse. Peatükis ei anna me niisiis täielikku ülevaadet kõigist korpuse analüüsi võimalustest, vaid keskendume korpusanalüüsi vahendites enam levinud ja lihtsamatele meetoditele.

5.1. Korpusanalüüsi vahendid

Tänapäeva korpused on reeglina nii suured, et pole mõeldav, et uurija suudaks neid läbi lugeda ja käsitsi huvipakkuvat välja otsida – sestap on vaja vahendeid, mis võimaldaksid teha süsteemseid päringuid ja pakkuda erinevaid analüüsivõimalusi. Spetsiaalne korpusanalüüsi tarkvara ja veebipõhised kasutajaliidesed pakuvad nii mitmekesiseid päringu tegemise võimalusi kui ka esmast statistikat (kasutussagedus, sagedusloendid jms). Seda kõike saab teha ka programmeerimiskeelte abil

(näiteks R, Python), ent nende kasutamine nõuab juba mõningast programmeerimisoskust. Programmeerimiskeelte suureks plussiks on aga kahtlemata see, et analüüsiviisid ja -meetodid on vähem piiratud ja võimalused arenevad pidevalt; samuti võimaldavad need teha keerukamaid statistilisi analüüse (vt ptk 6 „Korpusandmete statistiline analüüs").

Teeme järgnevalt lühikese ülevaate enamlevinud analüüsivahenditest ja -keskkondadest. Tuleb muidugi meeles pidada, et see valdkond areneb väga kiiresti – siin mainitud tarkvara ja kasutajaliidesed võivad mõne aasta pärast olla asendatud juba millegi tõhusamaga. Mainime seetõttu neid, mis on seni ajaproovile hästi vastu pidanud ning millel on taga pidev arendustöö. Põhjalikumaid nimekirju korpusanalüüsi tööriistadest leiab näiteks CLARIN-i võrgustiku lehelt[1] või lehelt corpus-analysis.com[2].

5.1.1. Korpusanalüüsi tarkvara

Korpusanalüüsiks on olemas spetsiaalseid programme. Siin loetletud tarkvara eeldab, et korpus on olemas failide kujul.

AntConc[3] on vabavaraline mitmekesine korpusanalüüsi tarkvara, mille on loonud Lawrence Anthony. See võimaldab teha konkordantside päringuid, koostada sagedusloendeid, sõnamitmikke, analüüsida märksõnu ja kollokatsioone. AntConci tootegruppi kuulub veel mitmeid programme, mis aitavad lahendada korpustega seotud ülesandeid, näiteks lisada tekstidele sõnaliigi märgendeid (TagAnt), analüüsida sotsiaalmeedia andmeid (FireAnt) jne, ent eesti keele tugi neil toodetel seni puudub. AntConc sobib hästi eelkõige väiksemahulise korpuse analüüsiks, millel ei ole märgendust, ehkki võimaldab töötada ka märgendatud tekstidega. Selliseks korpuseks võib olla näiteks hulk ilukirjandustekste või meediatekste, mis on kättesaadavad lihtteksti kujul (nt TXT-vormingus) või ka vormindatud tekstina (nt DOCX- või PDF-vormingus), ühe failina või paljude sarnasel viisil ette valmistatud failidena. AntConciga sarnaseid funktsioone pakuvad teisedki programmid, mis sageli ei ole aga vabavaralised (nt WordSmith Tools[4]).

ELAN[5] on vabavaraline audio- ja videomaterjalide transkribeerimis- ja märgendusvahend, mis võimaldab lisada eri märgenduskihte koos aegjoondusega. ELAN-is on ka korpuspäringu funktsioonid nii ühest kui ka mitmest tekstist korraga. ELAN-it on kasutatud näiteks J. Wilburi näidisuurimuses ohustatud keelte uurimisest.

---

1 https://www.clarin.eu/resource-families/corpus-query-tools

2 https://corpus-analysis.com/

3 https://www.laurenceanthony.net/software/antconc/

4 https://www.lexically.net/wordsmith/

5 https://archive.mpi.nl/tla/elan

Praat[6] on spetsiaalselt foneetiliseks analüüsiks loodud vabavaraline programm, millega on võimalik helisalvestisi transkribeerida ja märgendada, kusjuures eri märgenduskihte on võimalik lisada koos aegjoondusega. Samuti on sellega on võimalik teha erinevaid akustilisi analüüse (nt mõõta formantide või põhitooni sagedusi). Praati skriptimis- ehk käsujada kirjutamise võimaluste abil saab huvipakkuvaid andmeid ka failidest otsida ja töödelda. Praati on P. Lippus kasutanud eesti keele väldete näidisuurimuses.

5.1.2. Veebipõhised korpusanalüüsi keskkonnad

Paljude korpuste jaoks on loodud veebipõhised kasutajaliidesed, mis pakuvad erinevaid päringu- ja analüüsivõimalusi. Korpuste kasutajaliidesed on aja jooksul arenenud konkordantside (st kasutusnäidete) leidmise vahendist keerukaid kollokatiivseid, grammatilisi ja semantilisi mustreid tuvastavateks tööriistadeks. Siin tutvustame tuntumaid kasutajaliideseid, mis sisaldavad ühe või mitme keele erinevaid korpusi.

Sketch Engine
Sketch Engine[7] on korpusanalüüsi tööriist, mille abil on võimalik kasutada paljude eri keelte eri tüüpi korpusi. Sketch Engine on algselt loodud sõnaraamatute koostajate abivahendiks ning leksikaalse uurimistöö tegemiseks, ent tänaseks on tegemist universaalse korpusanalüüsi keskkonnaga, mis võimaldab lahendada eri tüüpi ülesandeid, nt uurida sõnade või konstruktsioonide esinemissagedusi ja kontekste ning teha semantilisi ja süntaktilisi analüüse. Sketch Engine'i kaudu saab muu hulgas kasutada üle 20 eesti keele korpuse, nt varasemat koondkorpust ja tasakaalus korpust ning eestikeelsel veebil põhinevaid veebikorpusi (Web 2013, 2017, 2019, 2021, 2023 väljavõtted), mis sisalduvad ka Eesti keele ühendkorpuste sarjas (Koppel & Kallas 2022, vt ka ptk 2 „Eesti keele korpused"). Sketch Engine'is on võimalik kasutada ka väiksemaid korpusi (nt leiame sealt eesti lastekeele korpuse), ning luua päris oma korpus (vt ptk 4 „Oma korpuse loomine"). Mitmekülgsed päringu- ja analüüsivõimalused teevad sellest tööriista, mis sobib erinevat tüüpi ülesannete lahendamiseks. Sketch Engine'i peamiseks miinuseks on selle kommertshuvid – selle kasutamiseks peab ostma litsentsi ehk kasutusloa. Mitmed Eesti teadusasutused on ligipääsu ostnud ning Eesti Keele Instituut teeb Sketch Engine'iga koostööd uute veebikorpuste loomiseks ja majutamiseks. Tasuta konto saab keskkonnas teha 30 päevaks, ent see ei taga ligipääsu kogu ühendkorpusele (küll aga näiteks veebikorpustele).

[6] https://www.fon.hum.uva.nl/praat/

[7] https://www.sketchengine.eu/

Sketch Engine'il on olemas eeskätt keeleõppele suunatud lihtsustatud tasuta versioon **SkELL**[8], millega saab kasutada Sketch Engine'i põhilisi funktsioone ehk konkordantsi, sõnavisandite ja tesauruse tööriistu. SkELL on olemas ka eesti keele jaoks. Selles kasutatakse andmetena mitte ühendkorpust, vaid eesti keele veebilausete korpust (mis on ühendkorpusest heade näitelausete filtriga ehk GDEX-iga välja valitud). Samuti on Sketch Engine'ist tasuta, ent piiratud funktsioonidega alla laaditav versioon **NoSketch Engine**[9], mille suurimaks puuduseks Sketch Engine'i kõrval on see, et see on vaid korpusanalüüsi tarkvara ega anna ligipääsu juba olemas olevatele korpustele. Seega sarnaneb see pigem AntConcile ja muule sarnasele tarkvarale. Selleks, et põhilisi Sketch Engine'i tööriistu oleks võimalik tasuta kasutada juba olemasolevate korpuste analüüsimiseks, haldab CLARIN-i võrgustik ka sellist **NoSketch Engine'i veebiversiooni**[10], mille kaudu saab ligi üle 260 korpusele 40 keeles (sh eesti keele ühendkorpusele 2021).

KORP

KORP (rts *konkordanser och ordfrekvensanalyser för språkforsning*) on veebipõhine korpusanalüüsi tööriist, mida kasutatakse erinevate korpuste majutamiseks ning neist päringute tegemiseks. See on tasuta vahend, mida arendab Rootsi keeletehnoloogia, keeleteaduse ja keeleandmete infrastruktuuri haldaja Språkbanken ('Keelepank') ning seda kasutavad ka mõned muud nn keelepangad, nt Soome keelepank[11]. Eestis on KORP-i hallanud Eesti Keeleressursside Keskus[12] ning selles paiknevad näiteks koondkorpuse, etTenTen2013, ERR-i raadiosaadete korpuse jpt korpuste materjalid. Olenevalt korpuse märgendusest on võimalik teha päringuid nii sõnavormide, grammatilise info kui ka metaandmete põhjal, sealjuures saab kasutada ka regulaarvaldisi ning otsida korraga mitme sõnavormi või sõnavormi omaduse põhjal. Lisaks konkordantsidele võimaldab KORP analüüsida ka päritud keelendite sagedusi. KORP-i on arendanud ka Eesti Keele Instituut[13].

5.1.3. Programmeerimiskeeled korpusanalüüsiks

Programmeerimiskeeled võimaldavad korpusandmeid töödelda ja analüüsida väga mitmekesiselt, samuti tehakse tänapäeval statistiline analüüs tavaliselt mõne programmeerimiskeele abil. Ka programmeerimiskeelte kasutamine korpusanalüüsiks eeldab üldjuhul, et korpus on failidena kättesaadav (välja arvatud juhtudel, kui

8 https://skell.sketchengine.eu/

9 https://www.sketchengine.eu/nosketch-engine/

10 https://www.clarin.si/ske/#open

11 https://korp.csc.fi/

12 https://korp.keeleressursid.ee

13 https://korp.eki.ee/

korpus kogutakse näiteks veebikraapimise teel, korpusele saab ligi rakendusliideste ehk API-de kaudu või sisaldub ligipääs korpustele juba mõnes programmeerimiskeele korpusanalüüsi paketis).

Shelli skriptid
Shelli ('kest') skriptide all peame siin silmas käsujadasid ehk **skripte**, mida saab kindlate käsuinterpretaatorite (nt Bash või sh) abil kasutada UNIX-il põhinevas operatsioonisüsteemis või täpsemalt pigem operatsioonisüsteemide perekonnas, kuhu kuuluvad näiteks MacOS, iOS, Android, Linux, BSD, Solaris jt. Shelli skriptid olid korpuslingvistika algusperioodi peamine töövahend, sest nende abil saab suurtest tekstimassiividest hõlpsalt huvipakkuvaid sõnesid (märgijärjendeid) välja otsida (käsk *grep*), asendada (käsk *sed*) jne. Shelli skripte saab edukalt kasutada ka praegu, kas siis UNIX-il põhinevasse serverisse sisse logides või oma arvutis: näiteks MacOS-is on selleks programm Terminal. Windowsi operatsioonisüsteem ei põhine UNIX-il, mistõttu selles ei saa shelli skripte jooksutada ilma lisatarkvara installimata. Shelli skriptide kasutamiseks Windowsis võib installida näiteks Windowsi alamsüsteemi Linuxi jaoks (WSL)[14]. Shelli skriptid sobivad hästi tekstiandmete töötlemiseks, päringute tegemiseks või sageduste kokkulugemiseks ning need võimaldavad kasutada **regulaaravaldisi** (vt alapeatükk 5.2.2). Shelli skriptidega ei saa siiski sooritada keerukamaid programmeerimisülesandeid, mistõttu kasutatakse tänapäeval selle asemel rohkem programmeerimiskeeli R ja Python.

R
R on programmeerimiskeel, mis on algselt loodud ennekõike statistiliseks analüüsiks ja andmete visualiseerimiseks, ent selle kasutusvõimalused laienevad pidevalt. R-is on mitmeid **pakette**, mis võimaldavad lahendada eri tüüpi ülesandeid, näiteks analüüsida tekste, ruumiandmeid, pilte jne. R on populaarne korpuslingvistika töövahend nii andmete puhastamiseks, korrastamiseks, statistiliseks analüüsiks kui ka andmete visualiseerimiseks. See on tasuta, avatud lähtekoodiga programmeerimiskeel ja sellel on aktiivne kogukond, kes R-i kasutusvõimalusi pidevalt edasi arendab. R-i kasutatakse tavaliselt RStudio keskkonnas, mis võimaldab kasutajatel R-i koodi lihtsamalt kirjutada, vigu eemaldada ning kasutada lugematuid laiendusi, mis andmeanalüüsi hõlpsamaks teevad. R-i saab kasutada ka andmete veebist kogumiseks. Erinevate tegevuste jaoks on R-is loodud palju pakette, mida on kasutatud ka selle õpiku näidisuurimustes (nt tekstide töötlemiseks sobivad paketid `tidytext` ja `tidyverse`, visualiseerimiseks `ggplot2`, mitmekülgseks korpuslingvistiliseks analüüsiks sobib pakett `quanteda`, statistilise analüüsi jaoks `stats` ja veel palju pakette, vt näidisuurimusi ja õpiku ptk 6 „Korpusandmete statistiline analüüs").

[14] https://learn.microsoft.com/en-us/windows/wsl/

Python
Python on samuti avatud lähtekoodiga programmeerimiskeel, millel on mitmeotstarbeline kasutusvaldkond. Korpuslingvistikas kasutatakse Pythonit selleks, et automatiseerida korpuste analüüsi protsessi. See hõlmab tekstide töötluse, sõnestamise, lemmatiseerimise, grammatilise analüüsi, semantilise analüüsi ja muid tekstianalüüsi meetodeid. Pythonit kasutatakse laialt **loomuliku keele töötlusel** (ingl *natural language processing,* NLP), mille jaoks on loodud mitmeid **teeke** ehk pakettide, funktsioonide ja failide kogusid (ka *koodiraamatukogusid*, vt Sügis jt 2024), näiteks `NLTK`, `TextBlob`, `spaCy` jt. Eesti keele töötluseks on vajalik teek `EstNLTK` (vt lähemalt ptk 3.3 „Märgendamistööriistad"). Need pakuvad erinevaid tööriistu ja funktsioone loomuliku keele töötluse ülesannete jaoks, nagu tekstide klassifitseerimine, meelestatuse analüüs, sõnaliikide märgendamine, lemmatiseerimine ja palju muud. Pythonit kasutatakse ka korpuste loomiseks, nt teksti kogumiseks veebist või dokumentidest.

Tänapäeval on üha enam võimalik eri programmeerimiskeelte võimalusi omavahel kombineerida. Näiteks on Pythoni loomuliku keele töötluse teekidest loodud R-is kasutatavaid versioone (nt Pythoni pakett `Spacy` on R-is pakett `spacyR`), samuti on võimalik Pythoni koodi jooksutada R-i koodiga vaheldumisi (näiteks RStudios R-i paketiga `reticulate`, Pythonis paketiga `rpy2`).

5.2. Korpuslingvistika lihtsamad analüüsimeetodid

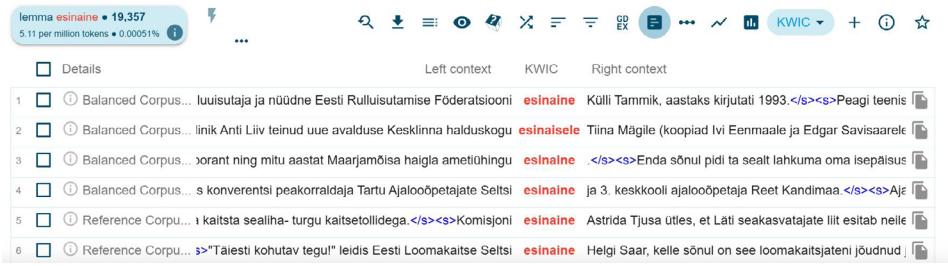
Tekstide analüüsil on terve rida küllaltki standardseid meetodeid, mida korpuslingvistikas laialt kasutatakse. Need on näiteks konkordantsid, sõnasagedused, kollokatsioonid ja sõnamitmikud. Paljud neist on juba sisse ehitatud korpusanalüüsi tööriistadesse, nagu AntConc ja Sketch Engine. Siinkohal anname nende kasutamisest lühiülevaate.

5.2.1. Konkordants ja KWIC

Konkordants (ingl *concordance*) on korpuspäringu tulemus, mille abil analüüsitakse sõnade kasutamise konteksti ja sagedust. Konkordants on sisuliselt sõna või märgijärjendi kasutusnäide. Konkordantsi väljundit, kus otsitav sõna või märgijärjend ehk **märksõna** on visuaalselt kuvatud rea keskele nii, et lausekontekst jääb mõlemale poole märksõna, nimetatakse ka **KWIC**-iks (ingl *keyword in context*, märksõna kontekstis, vt joonis 5.1).

Konkordantse saab teha sõna mingi tekstis esineva vormi (nt *esinaised*) või märgijärjendi põhjal (nt *esi, esinai*), ent otsitavaks võib olenevalt korpuse märgendusest ja päringusüsteemist olla ka sõna algvorm ehk lemma (*esinaine*), korpuses märgendatud (grammatiline) kategooria (nt nimisõna mitmuse nimetavas käändes) või mitmesõnaline konstruktsioon (näiteks *osa inimesi*, *sai tehtud*).

Sõltuvalt korpuse märgendamisest võib otsitavaid kategooriad olla veelgi, näiteks võib meil olla kasutada süntaktiline märgendus ning saame kombineerida kõiki eelpool mainitud variante. Päringut saab teha niisiis enam-vähem kõige põhjal, mis tekstis esineb regulaarselt või millele on korpuse tegemise käigus lisatud märgend. Konkordantsipäringut kasutades võime otsida tekstide seest huvipakkuvaid sõnu või märgijärjendeid koos lausekontekstiga või ilma, selleks et kasutusjuhte hiljem analüüsida, võrrelda, kokku lugeda vms. Konkordantside koostamisel saab tõhusamaks otsinguks kasutada ka regulaaravaldisi (vt alapeatükk 5.2.2), mis võimaldavad leida keerukamaid mustreid ja üldistusi, mida lihtpäringuga tabada ei ole võimalik.



Joonis 5.1. Lemma *esinaine* konkordants (KWIC) Eesti keele ühendkorpuses 2023. Selle sõna tähendusmuutuste kohta tehtud korpusuuringut vt (Kaukonen 2023a)

Konkordantsi põhjal saab hinnata näiteks sõnade tähendust ja nende muutumist, mis võib olla oluline info sõnaraamatute tegijatele ja sõnavara uurijatele (vt nt K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimust korpuste kasutamisest sõnastike koostamisel), grammatiliste kategooriate (nt käänete) kasutusfunktsioone, või hoopis seda, kuidas inimesed maailma näevad või kuidas maailm meie ümber on muutunud. Näiteks kultuuriga seotud sõnade tähendus või kasutussagedus muutub koos vastava valdkonnaga ja seega on keele kaudu võimalik uurida ka maailmas toimuvaid muutusi ja nende kajastamist (vt nt P. Tinitsa näidisuurimust elektri saabumisest Eesti aladele).

Päringuid tehes tuleb silmas pidada, et korpuslauseid võidakse kuvada vaikimisi mingis kindlas järjekorras (nt alamkorpuste kaupa või kronoloogiliselt). Kui uurimistöö jaoks on vaja juhuslikke näiteid, saab kasutada korpustööriista vastavat **juhuslikustamise** funktsiooni või tagada näidete juhuslikkus muul viisil. Vastasel korral võib juhtuda, et valim sisaldab tekste ühest allikast, ühest ajaperioodist või ühest registrist ning pole seega esinduslik kogu (alam)korpuse suhtes.

5.2.2. Regulaaravaldised

Konkordantside otsimisel on abiks **regulaaravaldised** (ingl *regular expressions,* lühendatult ka *regex, regexp*), mida kasutatakse tekstist teatud regulaaravaldises määratletud tingimustele vastava tekstimustri (märgijada) leidmiseks. Regulaaravaldis on formaalses keeles üleskirjutatud valem, mis kirjeldab teatavat sõnede klassi. Sõne all mõtleme siin mistahes sümbolitejärjendit, mis võib, aga ei pea kattuma sõna või sõnaosaga. Regulaaravaldisest võib mõelda kui kontrolleeskirjast, mida rakendatakse (mingile) sõnele; sõne kas siis vastab regulaaravaldisele või mitte.

Näiteks päring `esinai[ns][eit]` leiab üles sõna *esinaine* käändevormid: *esinaine, esinaise, esinaist, esinaisi* jne. See päring koosneb **literaalsetest sümbolitest**, mis esindavad iseennast, ja **metasümbolitest**, mis esindavad midagi muud: literaalseteks sümboliteks on siin päritava sõna alguseossa kuuluvad sümbolid `esinai`, metasümboliks on nurksulud, mis vastavad ühele sümbolile nurksulgudes antud valikust: `[ns]` tähistab, et järjendis on selles positsioonis kas *n* või *s*; `[eit]` tähistab, et selles positsioonis (peale *n*-i või *s*-i) esineb kas *e, i* või *t*.

Regulaaravaldise koostamisel tuleb läbi mõelda, 1) mis kujul otsitav tekstiosa esineda võib ning mis variandid selles esinevad (eesti keeles näiteks käände- ja pöördevormid, tüvevaheldused); 2) kui palju otsitavast märgijärjendist on vajalik regulaaravaldise tõhusaks kasutamiseks. Alati ei ole vaja kirjeldada sõnet algusest lõpuni, vaid piisab selle mingist kriitilisest osast. Näiteks sõna *mees* otsimiseks peame kindlasti arvestama tüvemuutustega käänamisel: *mees, mehe, meest, mehele*, *mehi, meestele*, ent kuna sellega on tüveosa kirjeldatud ka muude käändevormide jaoks, piisab päringuks selle formaliseerimisest: `me[eh][esi]`.

Regulaaravaldised on võimas vahend keeleandmetega töötamisel ja eriti kasulik märgendamata teksti kasutamisel, kus käänamine-pööramine ja sõnade tüve- ja lõpuvaheldused komplitseerivad huvipakkuvate sõnade tekstist otsimist. Regulaaravaldiste kasutamist raskendab see, et nende koostamine nõuab mõningast harjutamist. Järgnevalt on esitatud olulisemad regulaaravaldistes kasutatavad metasümbolid (tabel 5.1). Regulaaravaldiste harjutamiseks on veebis mitmeid treening- ja testimiskeskkondi, mida õppimisel saab kasutada[15].

[15] Vt näiteks https://regexone.com/ või https://regex101.com/

Tabel 5.1. Regulaaravaldiste metasümbolid

	Sümbol	Tähistab
Sisu- ja valiku-sümbolid	`.`	punkt: suvaline üks sümbol
	`\`	tagurpidi kaldkriips: teeb talle järgneva metasümboli literaalseks sümboliks, nt `\.`, otsib punkti ja koma järjendit (võtab punktilt metasümboli staatuse)
	`(me)`	grupeerivad sulud: kogu sulgudes olev üks sümbol või sümbolijärjend, nt `(me){1,3}` viitab järjenditele *me*, *meme*, *mememe*
	`[aeiou]`	üks kantsulgudes esitatud sümbolitest
	`[^aeiou]`	mistahes sümbol, mis ei ole üks kantsulgudes esitatud sümbolitest
	`[[:alpha:]]`	üks tähestiku täht
	`[a-z]`	üks (inglise) tähestiku väiketäht
	`[A-Z]`	üks (inglise) tähestiku suurtäht
	`[a-zõäöü]`	üks (inglise) tähestiku väiketäht või *õ*, *ä*, *ö*, *ü*
	`[A-ZÕÄÖÜ]`	üks (inglise) tähestiku väiketäht või *Õ*, *Ä*, *Ö*, *Ü*
	`[[:digit:]]`	üks mistahes number
	`[0-9]`	üks mistahes number
	`[[:punct:]]`	üks mistahes kirjavahemärk
	`|`	„või“, püstkriipsust vasakule või paremale jääv järjend, nt päring `mees|naine` leiab vasteks nii järjendi *mees* kui ka *naine*
Kvantorid	`*`	tärn: eelnev sümbol esineb null kuni lõpmatu arv kordi, nt avaldisele `lap*` vastavad *la*, *lap*, *lapp*, *lappp* jne Avaldisele `.*` vastab mistahes sümbolite järjend mistahes pikkusega (sh mitte midagi)
	`+`	pluss: eelnev sümbol esineb üks või enam korda, nt avaldisele `ka+` vastavad *ka*, *kaa*, *kaaa*, *kaaaaa*, aga mitte lihtsalt *k*
	`?`	küsimärk: eelnev sümbol esineb null või üks korda, nt avaldisele `ka?` vastavad ainult *k* ja *ka*

	Sümbol	Tähistab
Kvantorid	**`{n,m}`**	Looksulgudes saab defineerida, mitu korda eelnev sümbol või märgijada võib esineda (minimaalselt ja maksimaalselt), nt avaldis **`ai{2,4}`** leiab *aii*, *aiii* ja *aiiii*. Järjendite otsimisel rakendatakse kvantoreid järjendit sisaldavate sulgude järele, nt **`(ai){1,3}`** leiab järjendid *ai*, *aiai*, *aiaiai*
	`{n}`	leiab täpselt *n* eelnevat sümbolit, nt **`[0-9]{7}`** leiab järjendid, kus esineb järjest seitse numbrit (nagu näiteks telefoninumbrites)
	`{n,}`	leiab vähemalt *n* eelnevat sümbolit, nt **`[bcdfghjklmnprstv]{3,}`** leiab vähemalt kolme konsonandi ühendid
Positsioonisümbolid	**`$`**	dollarimärk: märgib rea lõppu
	`^`	„katus“: märgib rea algust (kui pole nurksulgude sees), nt **`^[0-9]`** leiab numbriga algavad read, **`[^0-9]`** leiab aga ühe mistahes mittenumbri

Regulaaravaldisi saab kasutada programmeerimiskeeltes, paljudes veebipõhistes korpuste kasutajaliidestes (sh Sketch Engine'is ja KORP-is), Notepad++-is jm tekstitöötlustarkvarades, spetsiaalsetes korpuspäringutarkvarades (nt AntConc) jne. Mõnes neist on vajalik regulaaravaldiste kasutamise võimalus eraldi sisse lülitada. Spetsiifilisemates veebipõhistes otsimootorites võib nende kasutamine olla siiski piiratud, näiteks saab kasutada vaid osa metasümbolitest. Kõige vabamalt saab regulaaravaldisi kasutada programmeerimiskeeltes, ehkki eri keeltes võivad mingid regulaaravaldiste elemendid käituda pisut erinevalt.

5.2.3. Konkordantside koostamine grammatilise info põhjal

Konkordantse saab moodustada ka sõnast pikemate järjendite põhjal. Kui soovime uurida mingit grammatilist konstruktsiooni, võime teha korpuspäringu, mis arvestaks sõnast suuremat konteksti, näiteks otsida kaht sõna, mis samas lauses koos esinevad. Seda on võimalik teha nii märgendamata kui märgendatud tekstide põhjal. Märgendamata tekstide puhul on sellisel juhul üsna vältimatu kasutada regulaaravaldisi, sest kõiki võimalikke järgnevusi ei suuda me üldjuhul ühte päringusse muul viisil panna. Näiteks soovides otsida *saama*-verbi ja *tud*-kesksõna ühendeid nagu *saab tehtud*, s*ai käidud*, tuleb arvestada nii *saama*-verbi erinevate vormidega (*saan, sai, saadi* jne) ja eri verbide *tud*-kesksõna vormidega (*käidud,*

tehtud, oldud jne), peale selle tuleb silmas pidada, et nende kahe sõna vahel võib olla veel mingeid sõnu (*sai toad korda tehtud*). Selline regulaaravaldis võib minna üsna keerukaks (näiteks üks võimalus: `sa[ai][ndbmtv]?.* [a-zõäöü][td]ud`) ja anda päringu vastuseks hulgaliselt selliseid ridu, mis küll vastavad regulaaravaldisele, ent ei esinda soovitud konstruktsiooni. Seepärast on morfoloogiliselt märgendatud teksti põhjal päringut hõlpsam kokku panna, ent ka siin võib vahel olla vaja koostada regulaaravaldisi. Tuntumates korpusanalüüsi tööriistades (Sketch Engine, KORP) või korpuste kasutajaliidestes saab märgendatud tekstide põhjal päringu koostamiseks kasutada ka spetsiaalseid korpuspäringukeeli.

Korpuspäringukeel CQL (ingl *corpus query language*) ja selle eelkäija CQP (ingl *corpus query processor*) on formaalsed päringukeeled korpustest komplekssete grammatiliste või leksikaalsete konstruktsioonide ja mustrite leidmiseks. CQL-i päringud võivad olla väga spetsiifilised, võimaldades uurijatel teha täpseid päringuid vastavalt uurimisvajadustele. Näiteks võib päringu abil hinnata, kui sageli teatud sõna ilmub mingi teise sõna läheduses või kui tihti esinevad teatud grammatilised konstruktsioonid.

Korpuspäringukeeled võivad erineda sõltuvalt kasutatavast korpusest või platvormist. Erinevatel eesmärkidel või alustel loodud tekstikorpustel võib olla erinev struktuur, vorming või metaandmed. Seetõttu võivad erinevad korpused või platvormid vajada erinevaid päringukeeli. Siin tutvustame mõningaid Sketch Engine'is kasutatava korpuspäringukeele[16] funktsioone ja päringuvõimalusi (vt tabel 5.2).

Iga sõnet (sh kirjavahemärke, arve jm mittesõnu) CQL-i avaldises tähistatakse kantsulgudega []. Nende sees saab täpsustada erinevaid **atribuute**, mille väärtused märgitakse jutumärkide vahele. Millised on atribuutide võimalikud väärtused, see oleneb päritava korpuse märgendusest. Sketch Engine's leiab selle *CORPUS INFO* nupu alt. Eesti keele ühendkorpuses kasutatavate morfoloogiliste märgendite täieliku nimekirja leiab samuti Sketch Engine'i lehelt[17].

Tabel 5.2. CQL-korpuspäringukeele olulisemad elemendid

Sõne ja selle atribuudid	Tähendus	Näide
`[]`	mistahes sõne	`[][][]` otsib kolme mistahes järjestikust sõnet
`word`	sõne, sõnavorm (tõstutundlik)	`[word="kallas"]` otsib sõnet *kallas*
`lc`	sõne, sõnavorm (tõstutundetu)	`[lc="kallas"]` otsib sõnesid *kallas*, *Kallas*, *KALLAS* jne

[16] https://www.sketchengine.eu/documentation/corpus-querying/

[17] https://www.sketchengine.eu/estonian-filosoft-part-of-speech-tagset

Sõne ja selle atribuudid	Tähendus	Näide
lemma	lemma (tõstutundlik)	**[lemma="kallas"]** otsib kõiki lemma *kallas* vorme (nt *kalda*, *kaldale*, *kallastelt*)
lemma_lc	lemma (tõstutundetu)	**[lemma_lc="kallas"]** otsib kõiki lemmade *kallas*, *Kallas*, *KALLAS* jne vorme (nt *kalda*, *Kallase*)
tag	sõnaliik	**[tag="U"]** leiab kõik ülivõrde vormid
features	grammatiline info	**[features="pl_p"]** leiab kõik mitmuse osastava (partitiivi) vormid (eri grammatilisi kategooriad eraldab alakriips)

Täpsustada saab ka mitut omadust, kasutades operaatoreid `&` („ja“) või `|` („või“), näiteks eesti keele ühendkorpuse puhul saab mitme omaduse põhjal otsida järgmiselt:

- `[tag="S" & features="pl_p"]` otsib kõiki nimisõnade (*S*) mitmuse (*pl*) partitiivi (*p*) vorme;
- `[tag="V" & (features="b" | features="s")]` otsib kõiki verbide (*V*) ainsuse 3. isiku oleviku (*b*) ja mineviku (*s*) vorme;
- `[(lemma="tarvis" | lemma="vaja") & tag="D"]` otsib kõik *tarvis* ja *vaja* kasutusjuhud, kus need on adverbi (*D*) funktsioonis (aga mitte nt kaassõnana).

Atribuudi väärtust saab ka eitada, kasutades hüüumärki (`!`):

- `[tag!="S" & features="pl_p"]` otsib kõiki mitmuse partitiivi vorme, mille sõnaliik ei ole nimisõna.

Mitme sõne järjendit pärides saab iga sõne puhul täpsustada tema atribuute eraldi:

- `[word="kõige"][tag="U"]` otsib järjendeid, kus on sõne *kõige*, millele järgneb ülivõrde vorm (nt *kõige suurim*).

Sõnede puhul saab kasutada ka kvantoreid, mis on märgitud loogelistes sulgudes `{}` selle sõne järel, mille esinemiskordade arvu see täpsustab:

- `[tag="N" & features="sg_n"][tag="A" & features="sg_p"]{,2}[tag="S" & features="sg_p"]` otsib arvsõnalise (*N*) põhisõnaga ainsuse (*sg*) nimetavas (*n*) käändes hulgafraase, kus põhisõna ja ainsuse (*sg*) partitiivis (*p*) nimisõnalise (*S*) laiendi vahel võib

olla veel kuni kaks ainsuse partitiivis omadussõna (*A*), nt *kolm roosi, kolm pikka roosi, kolm pikka punast roosi*.

Sõne atribuutide kirjeldamiseks võib kasutada ka regulaaravaldisi, näiteks

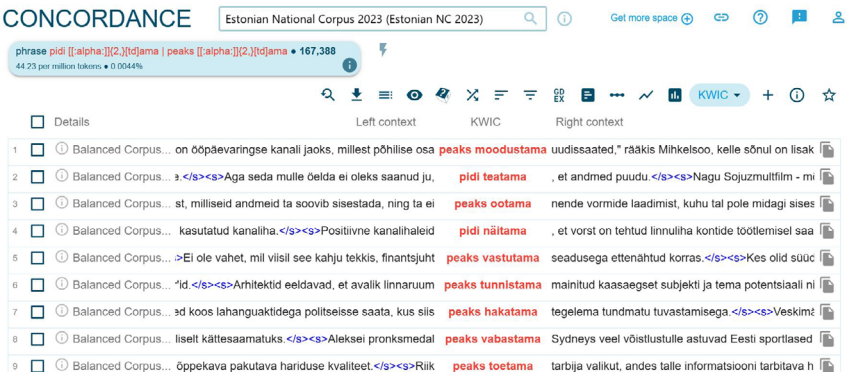
- `[lemma="muut.*"]` leiab kõik sõnad, mille lemma algab järjendiga *muut* (nt *muutma, muutus, muutlik* jne)
- `[tag="N" & features="sg_n"][tag="[ACUP]" & features="sg_p"]{,2}[tag="S" & features="sg_p"]` otsib arvsõnalise peasõnaga ainsuse nimetavas käändes hulgafraase, kus põhisõna ja laiendi vahel võib olla veel kuni kaks ainsuse partitiivis omadussõna (sh nende kesk- ja ülivõrde vormid) või asesõna, nt *kolm paksu sakslast*. Selle päringu vastuseks saadud konkordantsiridu illustreerib joonis 5.2.

Joonis 5.2. Päringu `[tag="N" & features="sg_n"][tag="[ACUP]" & features="sg_p"]{,2}[tag="S" & features="sg_p"]` tulemuseks saadud konkordantsiread Sketch Engine'is

> **Näide: konstruktsioon *pidama* + *Vtama*.** Vaatleme lähemalt eesti keele üht vähe käsitlust leidnud umbisikulist (impersonaalset) konstruktsiooni, mis koosneb *pidama*-verbi 3. isiku vormist ning verbi impersonaali *ma*-infinitiivist (nt *pidi tehtama*, *peaks antama*). Konstruktsiooni on eesti keeles vähe kirjeldatud ning on huvitav ka seetõttu, et *ma*-infinitiivi impersonaalivormi me muudes konstruktsioonides ei kohtagi. Näites vaatame, kuidas saab sellist konstruktsiooni otsida 1) märgendamata korpusest lihtteksti põhjal, 2) märgendatud korpusest CQL-i abil. Mõlemad päringunäited on koostatud Sketch Engine'is eesti keele ühendkorpuse 2023. aasta versiooni põhjal, et eri meetoditel saadud tulemusi oleks võimalik võrrelda. Eesmärk on selgitada välja, mis verbidega (*tama*-vormis) seda konstruktsiooni tüüpiliselt kasutatakse.
>
> **Märgendamata lihtteksti** põhjal otsides peame läbi mõtlema märgijada, mis on kogu konstruktsioonis ühine. Ühelt poolt on selleks *pidama*-verb, teiselt poolt verbi lõpus paiknev *-tama* või *-dama*. Samas ei tea me vähemalt esialgu, millised *pidama*-verbi vormid konstruktsioonis esineda võivad, ent oma keeletajule

toetudes teame, et kindlasti võivad seal esineda ainsuse 3. isiku vormid (kindlasti *pidi* ja *peaks*). Võime selle põhjal teha esialgse päringu ning seejärel seda tasapisi laiendada, et välja selgitada, mis variante veel kasutusel on.

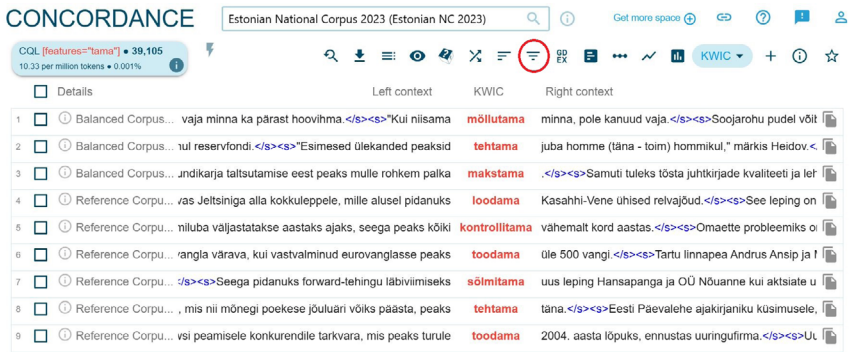
Päringu tegemiseks kasutame käesolevas näites Sketch Engine'i konkordantsi vahendit, kus valime päringutüübiks fraasiotsingu (*phrase*), mis võimaldab teha ühest sõnast suuremaid otsinguid märgijada põhjal ning rakendada regulaaravaldisi. Koostame regulaaravaldise, kus *pidama*-verbi variantideks on *pidi* või *peaks* ning sellele järgneb mingi sõna, mille lõpus on järjend *tama* või *dama* ning selle ees on vähemalt kaks tähestiku tähte: `pidi [[:alpha:]]{2,}[td]ama|peaks [[:alpha:]]{2,}[td]ama` (vt joonis 5.3). Mõnes teises päringusüsteemis, nt AntConcis saaks avaldist lühendada nii, et *pidama* variandid oleks grupeeritud sulgudesse ja eraldatud püstkriipsuga, mis tähistab sõna „või": `(pidi|peaks) [[:alpha:]]{2,}[td]ama`



Joonis 5.3. Konstruktsiooni *pidama* + *Vtama* fraasiotsingu tulemused Sketch Engine'is regulaaravaldise `pidi [[:alpha:]]{2,}[td]ama|peaks [[:alpha:]]{2,}[td]ama` põhjal

Tulemusest näeme, et see päring leiab otsitava küll üles, ent tulemusi on palju ja otsitavat konstruktsiooni selle hulgas vähe (joonisel 5.3 esitatud esimeses üheksas reas vaid üks: *peaks hakatama*). Selle põhjuseks on asjaolu, et kausatiivse sufiksiga *-ta-*/*-da-* verbe on eesti keeles väga palju ja need esinevad tekstis sageli (nt *moodustama*, *teatama*, *ootama*, *näitama*, *elavdama*), sellest tekkivaid *-tama* ja *-dama* järjendeid ei ole pelgalt märgijada põhjal võimalik impersonaalsetest kasutustest eristada. Seega tuleks tulemusi edasi piirata, ent mingit selget vormilist pidepunkti meil selleks ei ole. Jääb üle otsitavat konstruktsiooni sisaldavad laused käsitsi läbi vaadata, ent nende hulk on väga suur (167 388 vastet). Lisaks ei ole me arvestanud praegu muude võimalike *pidama*-verbi vormidega ning võimalusega, et pidama ja teise verbi vahel on veel sõnu, või et *pidama* ja teine verb esinevad lauses teises järjekorras (nt *tehtama pidi*).

Morfoloogiliselt märgendatud tekstide põhjal saame korpuspäringukeele abil kasutada infot sõnaliigi ja vormi kohta. *ma*-infinitiivi impersonaalivormi tähistab märgend *tama*, nii et võime teha kõigepealt päringu selle põhjal ja seejärel vaadata, kus ja mis vormis on kontekstis *pidama*-verb. Selleks valime päringutüübiks CQL ning sõne atribuudina lisame nurksulgude vahele otsitava vormi märgendi `[features="tama"]`, milles *features* osutab, et otsime morfoloogiliste märgendite hulgast. Selle otsingu tulemusena saame vastuseks üle nelja korra vähem ridu, mille hulgas on vähemalt esmapilgul ka õigete päringuvastete hulk oluliselt suurem kui lihttekstil põhinevas päringus (vt joonis 5.4).

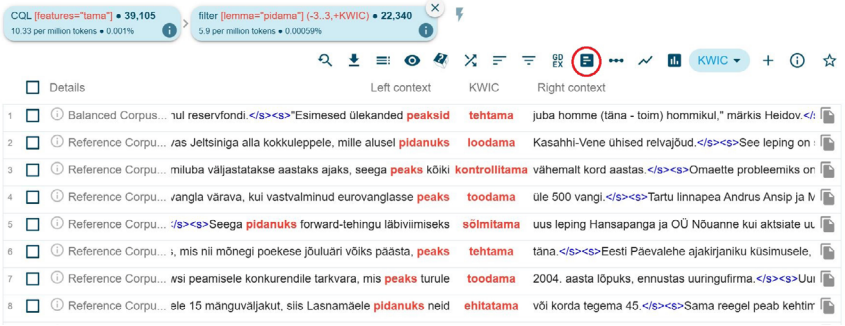


Joonis 5.4. CQL-päringu `[features="tama"]` tulemused Sketch Engine'is, punase ringiga on märgitud filtreerimistööriist

Järgmise sammuna filtreerime (*Filter*-tööriist märgitud joonisel 5.4 punase ringiga) tulemustest välja need konkordantsiread, kus ees- või tagakontekstis on kasutatud *pidama*-verbi (lemma põhjal, vt joonis 5.5).

Joonis 5.5. Filtri rakendamine päringutulemustele Sketch Engine'is

Tulemuseks saame konkordantsi, kus vasteid on veelgi vähem, täpsemalt 22 340 (joonis 5.6). Selle põhjal võime koostada juba konstruktsioonis esinevate verbide sagedusloendi, rakendades funktsiooni *Frequency* (joonisel 5.6 märgitud punase ringiga).



Joonis 5.6. Impersonaali *ma*-infinitiivi ja *pidama*-verbi ühendid peale filtri rakendamist, punase ringiga on märgitud sageduste leidmise tööriist

Saame teada, et kõige sagedasemad verbid, mis esinevad *pidama* + *Vtama* konstruktsioonis, on *tegema*, *andma*, *võtma*, *panema*, *viima*, *panema*, *valima* jne. Sageduselt kaheksas on ka *googlema*, mis tõenäoliselt on süstemaatiline märgendusviga, sest ei esine tegelikult selles konstruktsioonis (tegelikult kasutatakse kujul *googeldama*, mis sisaldab *dama*-järjendit, ent pole impersonaali *ma*-infinitiivi vorm).

Alternatiivne viis sama ülesannet lahendada on esitada päring, kus oleks arvestatud *pidama*-verbi lemmaga, *tama*-vormiga ning sellega, et need ei pruugi paikneda lauses järjestikku. Lisaks tasuks kohe arvestada ka sellega, et *pidama*-verb võib asuda nii *tama*-vormi ees kui järel. Selleks koostame CQL-päringu, mis arvestab mõlema sõnajärjega ning võimalusega, et *pidama*-verbi ja *tama*-vormi vahel võib olla 0–3 sõnet, aga mitte kirjavahemärke (`tag!="Z"`), et välistada vasted, kus *pidama* ja *tama*-vorm paikneksid erinevates (osa)lausetes:

```
([lemma="pidama"][tag!="Z"]{,3}[features="tama"])|
([features="tama"][tag!="Z"]{,3}[lemma="pidama"])
```

Ülejäänud analüüsisammud on üldjoontes samad. Loomulikult on võimalik see andmestik ka alla laadida, lisada sellele lisainfot ehk andmeid kodeerida (nt Excelis) ning teha põhjalikum kvantitatiivne analüüs (nt Excelis, R-is või mujal). Selles näites tulid selgelt esile märgendatud korpuse eelised märgendamata korpuste ees: kui uurida grammatilisi konstruktsioone, siis märgendus aitab meil huvipakkuvat oluliselt täpsemini üles leida.

5.2.4. Sõnavara analüüs: sagedusloendid

Korpuslingvistika üks peamisi eeliseid on võimalus arvesse võtta ja tõlgendada sagedusinfot: sageduse kaudu võime hinnata sõnade, häälikute, grammatiliste kategooriate, mitmesõnaliste ühendite vms esinemise tavalisust kas keeles üldiselt (võimalikult suure materjalihulga põhjal) või siis teatud perioodil, teatud registris või murdes, teatud tüüpi tekstides, teatud omadustega kõnelejatel. Sagedus ütleb meile, mis on keeles tavaline ja mis ebatavaline, tüüpiline või ebatüüpiline ning võimaldab näiteks registreid, murdeid, perioode ja indiviide omavahel võrrelda.

Korpuslingvistika üks tuntumaid tehnikaid on **sagedusloendite** koostamine. Sagedusloendeid kasutatakse sageli tekstide esmaseks kvantitatiivseks iseloomustamiseks. Sagedusloendist näeme, mis on selle teksti(kogu) kõige sagedasemad sõnad (lemmad või sõnavormid) ning võime võrrelda tulemusi mõne teise korpuse/tekstikoguga, et näha, kas nende vahel on olulisi erinevusi. Sagedusloendeid võib luua nii sõnede, lemmade kui ka sõnamitmike (vt alapeatükk 5.2.4.4) põhjal. Keeles on sagedased need sõnad, millel on grammatilisi funktsioone ja mille tähendus on seetõttu üldine või polüseemne. Sõnavara puhul on eriti sagedased sidesõnad (nt *ja, et, kui*), asesõnad (nt *see, ta*), üldise tähendusega verbid (nt *olema, saama*), mitmed adverbid (nt *siin, seal, jälle, vist, ka*) jne. Neid sõnu ühendab omavahel see, et nad on vajalikud sisusõnade omavaheliste suhete määratlemiseks, lausete ja tekstide kokkusidumiseks, hinnangute ja hoiakute edasiandmiseks jms. Ka näiteks nii komplitseeritud nähtust nagu ühe kirjaniku stiil on võimalik kirjeldada sageduste kaudu: eri kirjutajad eelistavad erinevaid sõnu või sõnavorme ning just need kõige sagedasemad sõnad võivad oma grammatilise ja süntaktilise funktsiooni kaudu paljastada nii mõndagi autorite iseloomulike keeleliste harjumuste kohta. See tähelepanek on aluseks näiteks **stilomeetriale** ehk tekstide kvantitatiivsele analüüsile, mis hindab ja võrdleb tekstide stiili sageduspõhiste mõõdikute alusel ja võimaldab nõnda piisava hulga materjali olemasolul isegi teksti autorit tuvastada (vt ülevaadet Šeļa 2021).

Leksikaalsed sõnad (millel ei ole tekstiehituse seisukohalt olulist rolli) esinevad tekstides oluliselt harvemini. Tekstides vaid korra esinevaid sõnu nimetatakse ka ***hapax legomena***'ks[18] ehk **ainukiteks**. Õigupoolest esinebki enamik sõnu tekstides vaid korra või paar. Näiteks 100 miljonit sõnet sisaldavas Briti rahvuskorpuses (British National Corpus) moodustavad ainukid üle poole kõigist sõnadest, kaks korda esinevad sõnad 13% ning vaid 5% kõigist sõnadest esinevad vähemalt 100 korda (Brezina 2018: 44). Selline jaotus – väike hulk väga sagedasi sõnu ning suur hulk harva esinevaid sõnu – on loomuliku keele tekstidele väga iseloomulik. Seda nimetatakse ka **Zipfi jaotuseks** (või ka Zipfi seaduseks) selle esmakordse kirjeldaja, ameerika lingvisti George Kingsley Zipfi (1902–1950) järgi. Zipfi seadus ütleb, et võrreldes sagedusloendi esimese sõnaga on teise sõna esinemissagedus umbes ½

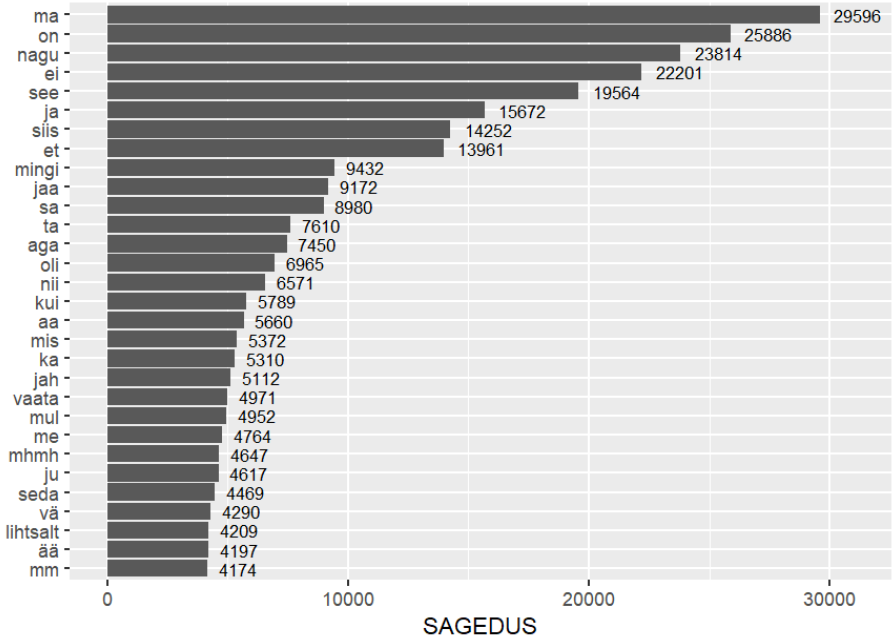
[18] *Hapax legomena* on mitmuslik vorm fraasist *hapax legomenon*, mis tuleb kreeka keelest ja tähendab 'korra öeldud'.

esimese sõna sagedusest, kolmas sõna ⅓ esimese sõna sagedusest jne (Brezina 2018: 44). Tuleb muidugi meeles pidada, et Zipfi seadus on üldistus ning iga korpuse tegelik sagedusloend võib sellest mõnevõrra erineda, vt näiteks teismeliste keele korpuse 30 kõige sagedasema sõne ja lemma jaotumist joonistel 5.7 ja 5.8.

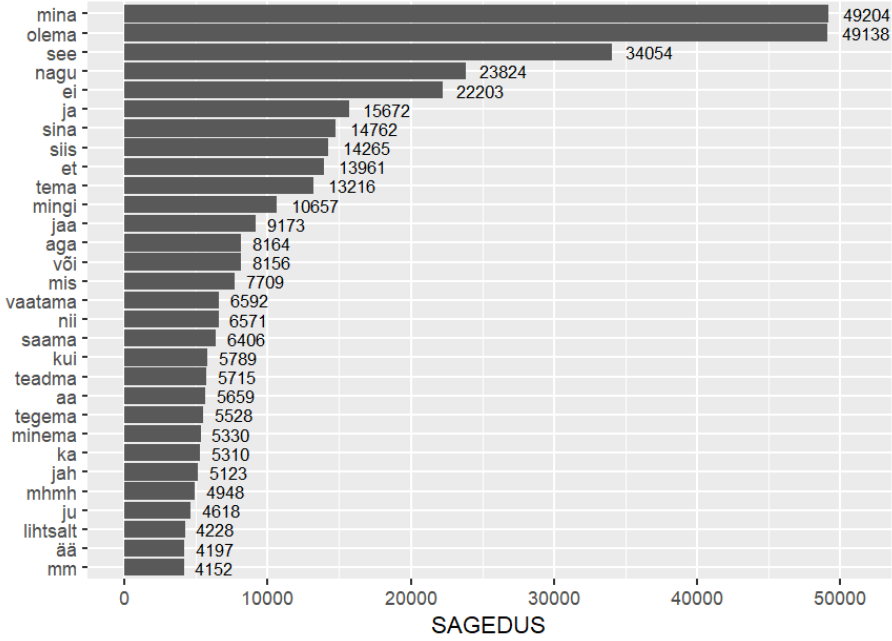
Sagedusloendi kõige sagedasemad sõnad on meile oluliseks infoallikaks siis, kui soovime süstemaatiliselt võrrelda näiteks eri tekstide, perioodide, murrete, registrite või isikute stiili. Ent kuna kõige sagedasemad sõnad on korpustes enamasti samad, võivad need analüüsi ka hoopis segada – näiteks juhul, kui soovime korpusest leida üles need sõnad, mis iseloomustavad kõige paremini tekstide temaatikat või sisu. Seepärast võib uurijal olla soov kõige sagedasemad funktsioonisõnad analüüsist hoopis eemaldada ja keskenduda selle asemel sagedamatele sisusõnadele. Sagedaste sõnade eemaldamiseks kasutatakse **stoppsõnade loendit**, mille võib koostada ise või kasutada mõnd olemasolevat loendit (näiteks AntConci töövahendite hulgas on paljudele keeltele sellised loendid juba olemas). Üks eesti ilukirjanduse põhjal koostatud stoppsõnade loend on leitav DataDOI repositooriumist (Uiboaed 2018). Tuleb siiski rõhutada, et erinevate ülesannete puhul võib vaja minna erinevaid stoppsõnade loendeid, seega tuleks nende kasutamine ja koostamine igal konkreetsel juhul hoolikalt läbi mõelda. Võib mõelda sellele, kas tarvis on üldist või valdkonnaspetsiifilist loendit ning mida eemaldatud stoppsõnadega loendiga edasi tehakse (näiteks kas mingitel elementidel, nagu eitus või intensiivistavad sõnad (nt *väga, hästi, liiga*), on edasises analüüsis tähendust kandev roll või mitte).

Sagedusloendeid võib moodustada nii **sõnede** kui ka **lemmade** põhjal. Sõnesageduse puhul loeb arvuti kokku täpselt samal kujul esinevad tähejärjendid, st lihtteksti põhjal moodustatud sagedusloend käsitleb iga vormi eraldi sõnana. Sel juhul ei eristata sama lekseemi erinevaid tähendusi (näiteks *hiir* looma ja arvutiosa tähenduses), eri lekseemide kokkulangevaid vorme (näiteks *tee* võib olla jook, käimiseks ettevalmistatud pinnaseriba või *tegema*-verbi käskiva kõneviisi ainsuse 2. isiku vorm) ega sama lekseemi kokkulangevaid muutevorme (nn vormihomonüümia, nt *tuba* võib olla nii ainsuse nimetava kui osastava käände vorm, *andsid* on *andma*-verbi lihtmineviku ainsuse 2. isiku ja mitmuse 3. isiku vorm). Seega võib olla mõttekas enne sagedusloendi koostamist tekstid **lemmatiseerida** ehk viia iga sõna tekstis tagasi selle põhivormi kujule. Ka Sketch Engine võimaldab sagedusloendeid koostada nii lemmade kui ka sõnede põhjal. Joonisel 5.7 on kujutatud teismeliste keele korpuse (Vihman jt 2023) 30 kõige sagedasemat sõnet ja joonisel 5.8 30 kõige sagedasemat lemmat. Kuivõrd tegemist on suulise keelega, on sagedusloendi tipus lisaks muudele funktsioonisõnadele suhtluspartiklid (*jah, mhmh, vä, vaata*) ja üneemid ehk täidetud pausid (nt *aa, ää, mm*).

Sagedusloendid (nagu ka sõnamitmikud, kollokatsioonid ja võtmesõnade analüüs) koostatakse enamasti nii, et suurtähed muudetakse väiketähtedeks, et mitte eristada lausealgulise suure tähega kirjutatud sõnu nende väiketähelistest variantidest. Praktikas toob see enamasti kaasa selle, et ka nimed muudetakse väiketäheliseks (näiteks loetakse samaks sõnaks üldnimi *kallas* ja pärisnimi *Kallas*). Samuti ei kaasata analüüsi kirjavahemärke.



Joonis 5.7. Teismeliste keele korpuse 30 kõige sagedasemat sõnet



Joonis 5.8. Teismeliste keele korpuse 30 kõige sagedasemat lemmat

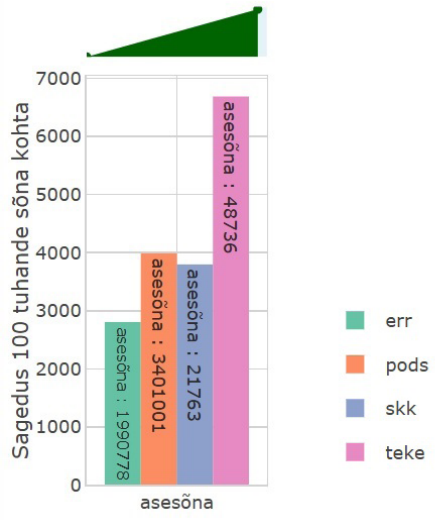
5.2.4.1. Sagedusloendi põhjal keele uurimine

Sagedusloendid võimaldavad meil võrrelda eri korpuste sagedasemat sõnavara. Me võime soovida uurida kõige sagedasemat sõnavara eri registrites (nt meediatekstides, teadustekstides või ilukirjandustekstides). Näiteks on uuritud suhtumist väljendavate partiklite levikut eesti keele registrites ning leitud, et ehkki sõnad nagu *vist, kindlasti, äkki, tegelikult, võibolla, muidugi, ehk, ilmselt, lihtsalt, tõesti* esinevad eri registrites (suulises argi- ja ametivestluses, kirjalikus tšätikeeles, veebikommentaarides, ilukirjanduses, ajakirjanduses, akadeemilistes kirjutistes) sageli, on nende esinemissagedus registriti üsna erinev. Kõige enam kasutatakse partikleid dialoogilistes vestlustes, seejuures tšätivestlustes enam kui suulistes argi- või ametivestlustes. Argivestlustes on kõige sagedasemad partiklid *vist*, *tegelikult* ja *lihtsalt*; need esinevad kõige sagedamate hulgas ka tšätivestlustes ja ametisuhtluses. Kui ajakirjanduses on nelja kõige sagedama hulgas *kindlasti, tegelikult, lihtsalt, muidugi*, siis ilukirjanduses on esinelik muidu sama, ent *kindlasti* asemel on seal *võibolla*. Akadeemilistes kirjutistes kasutatakse partikleid kõige harvem ja neist sagedasemad on *ilmselt, kindlasti, lihtsalt, tegelikult*. Partiklite esinemissagedust mõjutavad enim vajadus osutada teate tõsikindlusele või vastupidi, kahtlusele ja ebamäärasusele, mis erinevates registrites erineval määral avalduvad, samuti kasutatakse neid teate pehmendamiseks või hinnangu andmiseks (Metslang jt 2024). Seega täidavad taolised pragmaatilised partiklid tekstides olulist rolli. See käib laiemalt kõigi funktsioonisõnade kohta: nende kaudu võime võrrelda ühe keele eri registreid, murdeid, aga ka eri perioodide keelekasutust, näiteks 19. sajandi ilukirjanduse sagedasemaid sõnu tänapäeva ilukirjandusega.

Sagedusloendite põhjal saab võrrelda ka erinevate sotsiaalsete rühmade keelekasutust, näiteks teismeliste sagedasemaid sõnu täiskasvanutega või meeste ja naiste vahelist keelekasutust. Eelmises alapeatükis 5.2.4 nägime, et teismeliste keeles on sagedusloendi tipus 1. isiku pronoomenid (*ma, mina*, vt joonised 5.7 ja 5.8). Võime küsida, kas 1. isiku asesõna niivõrd sage kasutamine on iseloomulik ainult teismeliste keelele või suulisele keelele laiemalt? Selleks peaksime võrdlema teismeliste keele korpuse sagedusandmeid muude suulise keele korpuste vastavate andmetega. Selleks võib koostada korpusest ise sagedusloendi, ent praegusel juhul on meil kasutada teismeliste korpuse, spontaanse kõne foneetilise korpuse, ERR-i ja taskuhäälingute korpuse põhjal koostatud andmestikud (Lippus, Lõo, jt 2024) ja nende juurde kuuluv visualiseerimisvahend[19], kust saame otsida ka lemmade kaupa. Selle põhjal näeme, et teismeliste keele korpuses on *mina* kasutus oluliselt sagedasem kui ülejäänud suulistes korpustes (arvutatuna 100 000 sõne kohta, vt joonis 5.9). Täpsuse huvides olgu öeldud, et kuna eesti keele morfoloogia analüsaator analüüsib sõnu *me* ja *meie* asesõna *mina* mitmusevormideks, hõlmab siin lemma *mina* ka sõna *meie* vorme. Teismeliste keele korpuse ja spontaanse kõne foneetilise korpuse sagedasemad sõnavormid on esitatud ka tabelis 5.3, kust näeme, et teismeliste keeles on

[19] https://suulinekeel.ut.ee

ma sagedusloendis esikohal, ent täiskasvanute keelt sisaldavas foneetilises korpuses alles 6. kohal. Seega on 1. isiku asesõna teismeliste keeles tõepoolest oluliselt sagedasem kui täiskasvanute suulises keeles. Miks see nii, on küsimus omaette – teismeliste keelt peetakse emotsionaalseks ning tõenäoliselt on sage enesele viitamine sellega seotud. Muide, sama on leitud ka naiste ja meeste kõnet võrreldes – naiste keelekasutust on peetud emotsionaalsemaks, selles on enam endale viitamist, intensiivistavaid adverbe (eesti keeles *väga, eriti* vms), emotsioonidele viitavaid verbe, aga ka kõhklusmarkereid (*ma arvan, vist, äkki*) kui meeste keelekasutuses. See teadmine on meil seni peamiselt inglise keele põhjal tehtud uuringutest (Newman jt 2008), eesti keele uurimine on selles osas alles algusjärgus, ent on siiski üks uurimus meeste ja naiste keelekasutuse erinevustest taskuhäälingute korpuse põhjal, millest selgus, et naised kasutavad meestest sagedamini omadussõnu, adverbe, partikleid, intensiivistajaid, viisakuskonstruktsioone (sh tingivat kõneviisi ja diskursusmarkereid nagu *ma usun, et*) ning deminutiivi, samas kui meestel oli sõnavara mitmekesisus (TTR, vt alapeatükk 5.2.4.3) veidi suurem kui naistel (Kriuchkova 2025).



Joonis 5.9. Lemma *mina* esinemissagedus 100 000 sõne kohta ERR-i korpuses (joonisel *err*), taskuhäälingute korpuses (*pods*), spontaanse kõne foneetilises korpuses (*skk*) ja teismeliste keele korpuses (*teke*)

Mingit sõnavara osa võib uurida ka selleks, et hinnata ühiskonnas valitsevaid sotsiaalseid või võimusuhteid. Näiteks Elisabeth Kaukonen (2023b) on uurinud ühendkorpusesse kuuluva veebikorpuse 2021 põhjal *onu* ja *tädi* sisaldavate liitsõnade kasutust eesti keeles ja võrrelnud sagedusandmeid muid sugu väljendavate nimisõnadega (*mees, naine, poiss, tüdruk*), et välja selgitada, miks neid sõnu

lisaks otseselt sugulusele viitamisele veel inimesele viidates kasutatakse (nt *politsei-onu, koristajatädi*) ning mida selliste sõnadega tekstis tehakse. Ta võrdles *onu-* ja *tädi*-lõpuliste liitsõnade **sõnesagedust** ja **tüübisagedust**: sõnesagedus ütleb meile, kui palju soole viitavaid liitsõnu, mille teine komponent on *mees, naine, poiss, tüdruk, onu* või *tädi* korpuses esines, tüübisagedus seda, kui palju erinevaid lekseeme (antud juhul erinevaid liitsõnu) nendega moodustatakse, st kui **produktiivne** selline sõnamoodustus on (vt ka M.-L. Pilviku näidisuurimust tuletusliidete produktiivsuse kohta) ning milline on mees- ja naissoole viitavate sõnade omavaheline jaotumine. Selgus, et paarides *mees-naine* ja *poiss-tüdruk* oli nii mehele viitavate sõnade sõne- kui ka tüübisagedus oluliselt kõrgem kui naisele viitavatel sõnadel. Paari *onu* ja *tädi* puhul olid nii sõne- kui ka tüübisagedus aga kõrgemad hoopis naissoole viitava *tädi* puhul. Kõrge sõnesagedus viitab sellele, et mingit malli (nt *tädi*-lõpulisi sõnu) kasutatakse sageli, kõrge tüübisagedus aga sellele, et sageli ei kasutata mitte ainult üht kindlat vormi (nt *kokatädi*), vaid malliga moodustatakse palju erinevaid sõnu. Kaukoneni kvalitatiivsest analüüsist selgus, et *tädi*-lõpulisi liitsõnu kasutatakse tihti mitteametlikes ametinimetustes (*kokatädi, koristajatädi*), eriti lastega seotud ametites (*kasvatajatädi*), aga sageli ka halvustamiseks.

Sõnede sageduse hindamisel võib kasutada erinevaid mõõdikuid. Kõige lihtsam neist on **absoluutsagedus** – see on sõne kõigi esinemisjuhtude arv korpuses või vaadeldud korpuseosas. Absoluutsageduste põhjal on enamasti aga raske korpusi või korpuseosi omavahel võrrelda, sest korpuste suurused võivad olla väga erinevad. Erineva suurusega korpuste võrdlemiseks kasutatakse seetõttu enamasti **suhtelist sagedust**. Selleks jagatakse absoluutsagedused läbi korpuse suurusega. Suhtelised sagedused jäävad alati 0 ja 1 vahele. Neid saab omakorda teisendada **normaliseeritud sagedusteks** ehk arvutada ümber vastavalt mingile normaliseerimisbaasile (nt 1000, 10 000 või 1 000 000 sõne kohta):

Normaliseeritud sagedus = nähtuse absoluutsagedus korpuses ÷ sõnade koguarv korpuses × baas

Baasi suurus oleneb korpuse suurusest. Suure korpuse puhul on tavaline, et sõnade esinemist arvutatakse ühe miljoni sõne kohta. Väiksema korpuse puhul on sobivam arvutada näiteks 10 000 või 1000 sõne kohta. Eespool joonisel 5.9 on sagedused normaliseeritud 100 000 sõne kohta, sest ERR-i ja taskuhäälingute korpuse mahud on väga suured (vt ptk 2.3.2 „Suulised korpused"). Kui võrrelda omavahel mitmeid (alam)korpusi, võib olla mõistlik normaliseerida sagedused nii, et baasiks on (alam) korpuste keskmine sõnede arv. Normaliseerimise puhul tuleb meeles pidada, et suhteline sagedus on vaid korpuste võrdlemiseks ning tulemuste raporteerimisel tuleb alati esitada ka absoluutsagedus (näiteks sõna *nagu* esineb teismeliste keele korpuses 10 000 sõne kohta 326,75 korda, kokku 23 814 korda, vt tabel 5.3).

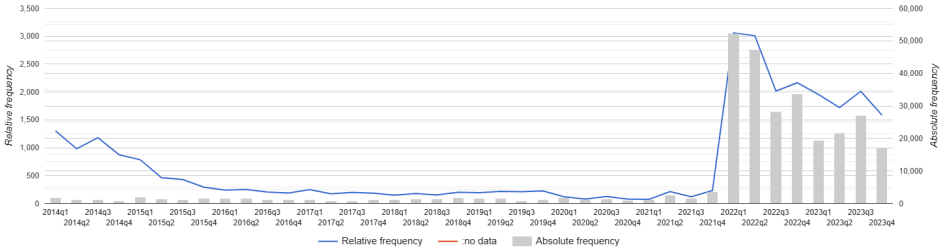
Tabelis 5.3. ongi näitena toodud 20 kõige sagedasemat sõnet, nende absoluut- ja 10 000 sõna baasil normaliseeritud sagedused kahes suulise kõne korpuses:

teismeliste keele korpuses (Vihman jt 2023) ja spontaanse kõne foneetilises korpuses (Lippus, Aare, jt 2023). Ehkki kaks korpust on erineva suurusega, näeme normaliseeritud sageduste põhjal näiteks seda, et vorm *on* on mõlemas korpuses sama sage, ent *nagu* on teismeliste suulises kõnes oluliselt sagedasem kui täiskasvanute kõnes.

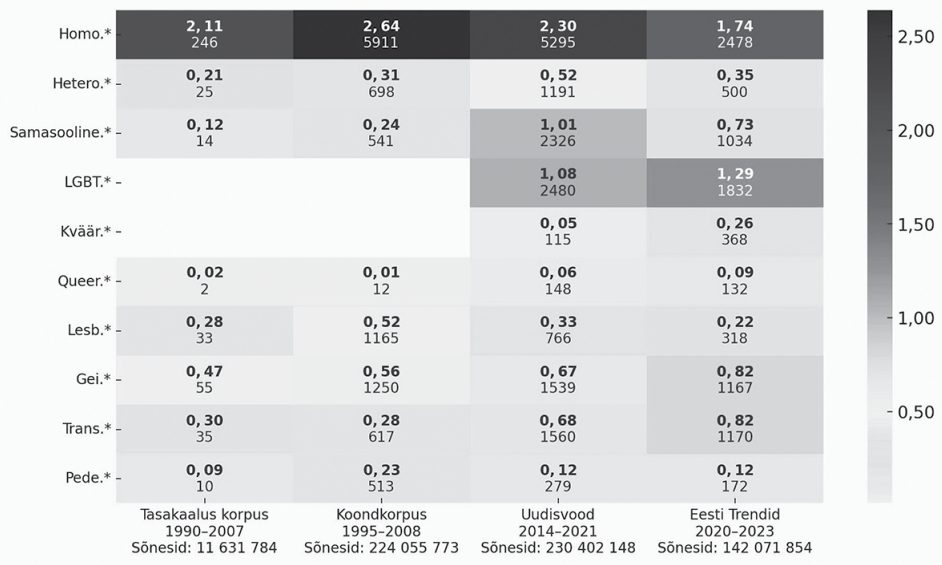
Tabel 5.3. Sagedasemad sõned teismeliste keele korpuses ja spontaanse kõne foneetilises korpuses (absoluut- ja normaliseeritud sagedus)

Teismeliste keele korpus			**Spontaanse kõne foneetiline korpus**		
Sõne	**Abs. sagedus**	**Norm. sagedus**	**Sõne**	**Abs. sagedus**	**Norm. sagedus**
ma	29 596	406,08	*et*	21 934	382,51
on	25 886	355,18	*on*	20 368	355,20
nagu	23 814	326,75	*ja*	19 088	332,88
ei	22 201	304,62	*ei*	15 464	269,68
see	19 564	268,44	*see*	15 358	267,83
ja	15 672	215,03	*ma*	12 161	212,08
siis	14 252	195,55	*siis*	11 065	192,97
et	13 961	191,56	*aga*	9280	161,84
mingi	9432	129,42	*jah*	8768	152,91
jaa	9172	125,85	*oli*	7141	124,53
sa	8980	123,21	*noh*	6298	109,83
ta	7610	104,42	*seal*	5744	100,17
aga	7450	102,22	*ka*	5692	99,26
oli	6965	95,57	*ta*	5328	92,92
nii	6571	90,16	*no*	5151	89,83
kui	5789	79,43	*nagu*	4976	86,78
aa	5660	77,66	*kui*	4860	84,76
mis	5372	73,71	*mis*	4278	74,61
ka	5310	72,86	*nii*	4263	74,34
jah	5112	70,14	*sa*	4256	74,22

Sõnade suhtelist sagedust saab kasutada ka sõna kasutamissageduse hindamiseks mingil ajaperioodil ning siduda seda maailmas toimuvaga, st siduda korpusanalüüsi tulemuse keeleväliste teguritega ning käsitleda tekstist leitut kui peegeldust maailmas toimuvast (vt ka P. Tinitsa näidisuurimust elektri saabumisest Eesti aladele). Näiteks pärisnime *Ukraina* suhteline sagedus ühendkorpuse 2023. aasta versiooni ajatempliga tekstides (joonis 5.10) kajastab hästi seda, et tekstides kirjutati palju Venemaa sõjast Ukrainas 2014. aastal (rünnakud Krimmis ja Donbassis) ja hiljem täiemahulisest sissetungist 2022. aastal.



Joonis 5.10. Sõna *Ukraina* absoluutne (hallid tulbad) ja suhteline (sinine joon) sagedus (normaliseeritud miljoni sõne kohta) ajatempliga tekstides 2014–2023 (Sketch Engine)



Joonis 5.11. LGBT+ sõnavara kasutussageduse muutumine (normaliseeritud 100 000 sõne kohta ja absoluutsagedus) (Kuusik 2024: 322)

Aet Kuusik (2024) on uurinud eesti LGBT+ sõnavara muutumist ühendkorpuse 2021 põhjal, võrreldes sõnade absoluut- (tabelis alumine arv) ja normaliseeritud sagedust (ülemine arv) korpuse erinevast ajaperioodist pärit tekstides (joonis 5.11).

Lisaks on alates 2014. aastast kasutusse tulnud sõnad *LGBT* ja *kväär* ja sagenenud sõna *samasooline* kasutus. Muutused näitavad, kuidas ajakirjanduses LGBT-teemadest räägitakse ning kuidas on muutunud ühiskondlikud hoiakud: näiteks sõna *samasooline* esilekerkimine on seostatav tsiviilpartnerluse ja abieluvõrdsuse diskussiooniga ning *LGBT* on asendanud vähemalt osaliselt sõna *homo-(seksuaalne)* (Kuusik 2024).

5.2.4.2. Sõnavara hajuvus ja levik

Kui kõige tavalisemad ja sagedasemad grammatilised sõnad (*nagu, olema, ja, et* jms) on korpuse tekstides suhteliselt ühtlaselt jaotunud (st kasutatakse kõigis tekstides, sest on tekstiehituslikult vajalikud), siis osa sõnu koonduvad ainult mingisse osasse tekstidesse ning mujal esinevad harva. See koondumine võib sõltuda teemast, žanrist/registrist, autorist, kirjutamisajast vms. Näiteks sõna *koogivorm* on seotud ennekõike retseptidega ja esineb seal sageli, ent muudes kontekstides esineb harva. Sõnavara analüüsides võib seega lisaks sagedusele olla vaja arvesse võtta ka sõnavara **hajuvust** korpuses.

Üks viis, kuidas hajuvust hinnata, on vaadata sõnavara esinemist korpuseosades või tekstifailides, millest korpus koosneb. Selle mõõdikuks on kasutatud **levikut** (ingl *range,* R), mis ütleb meile, kui mitmes korpuseosas (või tekstifailis, millest korpus koosneb) vaadeldav sõna esineb. See mõõdik ei võta arvesse korpuseosade mahtu. Näiteks kui sõna *koogivorm* esineb korpuses 8 tekstifailis, siis levik R(*koogivorm*) = 8. Sketch Engine'i sagedusloendis väljendab levikut mõõdik DOCF (ingl *document frequency*). Levikut hinnatakse vahel ka protsentuaalselt (**suhteline levik**, R%): kui mitu tekstifaili (korpuseosa) kõigist tekstifailidest (korpuseosadest) vaadeldavat sõna sisaldab. Kui korpus koosneb näiteks 123 tekstist ja koogivorm esineb neist 8 tekstis, siis suhteline levik R%(*koogivorm*) = 8 / 123 × 100 = 6,5%. Sketch Engine'is esindab seda mõõdik Relative DOCF.

Lisaks levikule ja suhtelisele levikule kasutatakse sõnavara hajuvuse hindamiseks veel mitmeid meetodeid. Näiteks võib kasutada **standardhälvet** (ingl *standard deviation*), mis näitab, kui palju sõna suhteline sagedus erinevates korpuseosades erineb selle sõna keskmisest suhtelisest sagedusest korpuses (vt ka ptk 6 „Korpusandmete statistiline analüüs"). Erinevate sõnade leviku võrdlemiseks korpuses sobivad veel mõõdikud nagu CV (ingl *coefficient of variation*), Juillandi D ning DP (ingl *deviation of proportions*). Kõigi nende mõõdikute kohta vaata lähemalt nt (Brezina 2018: 46–53).

Selleks, et välja selgitada sõnad, mis on korpuses kõige olulisemad – näiteks keeleõppe seisukohast – on seega vaja ühtaegu hinnata nii nende sagedust kui hajuvust. **Keskmise vähendatud sageduse** mõõdik (ingl *average reduced frequency,* ARF) kombineeribki sagedust ja hajuvust ning arvutab nende põhjal välja väärtuse,

millega hinnata sõna **tavalisust** kogu korpuses. ARF-i arvutamiseks on vaja teada sõne absoluutsagedust, korpuse suurust (sõnede arvu) ning sõne esinemispositsioone (esinemisjärjekorda) korpuses. See arvutab sõne esinemissageduse ümber nii, et käsitleb üksteisele lähedal asuvaid sama sõne esinemusi ühena ja vähendab selle alusel esinemissagedust, mis võib olla võimendatud mingi ühe teksti või korpuse alamosa eripäradest. Mida lähemal on ARF-i väärtus sõne esinemise absoluutsagedusele, seda ühtlasemalt on sõne korpuses levinud, ning mida suurem ARF-i väärtus on, seda tavalisem sõne korpuses on. ARF on integreeritud ka Sketch Engine'i sagedusloendisse (tabel 5.4). ARF-i arvutamise kohta vt täpsemalt (Savický & Hlavácová 2002; Brezina 2018: 54–56).

Järgnevas tabelis 5.4 on esitatud ühendkorpuse 2023 10 kõige sagedasemat lemmat, nende absoluutsagedus, normaliseeritud sagedus (1 miljoni sõne kohta), levik (DOCF), suhteline levik (Relative DOCF) ja keskmine vähendatud sagedus (ARF). Nagu näeme, pole isegi kõige sagedasemad sõnad levinud kõigis korpuse dokumentides (*olema* on kasutatud vaid 78,77% tekstidest). Võib oletada, et see on seotud väga lühikeste tekstidega korpuses, ent vajaks loomulikult lähemat analüüsi.

Tabel 5.4. Ühendkorpuse 2023 10 kõige sagedasemat lemmat, nende absoluutsagedus, normaliseeritud sagedus (1 miljoni sõne kohta), levik, suhteline levik ja keskmine vähendatud sagedus

Jrk. nr.	Lemma	Absoluutsagedus	Normaliseeritud sagedus	Levik (DOCF)	Suhteline levik (Relative DOCF)	Keskmine vähendatud sagedus (ARF)
1	*olema*	144 702 425	38 213,86	12 011 892	78,77%	93 459 800
2	*ja*	96 415 942	25 462,09	11 573 334	75,90%	62 061 984
3	*see*	62 283 206	16 448,11	8 461 142	55,49%	36 834 784
4	*et*	40 233 700	10 625,15	6 830 048	44,79%	23 274 376
5	*mina*	39 558 533	10 446,85	5 240 080	34,36%	17 366 472
6	*ei*	35 235 846	9305,29	6 369 427	41,77%	19 734 166
7	*tema*	32 986 776	8711,34	5 541 947	36,34%	15 450 645
8	*kui*	29 453 461	7778,24	6 721 755	44,08%	17 546 144
9	*mis*	28 662 207	7569,28	7 126 902	46,74%	17 075 230
10	*ka*	25 583 947	6756,36	6 703 407	43,96%	15 219 550

5.2.4.3. Sõnavara mitmekesisus

Sageli hinnatakse sagedusandmete abil ka sõnavara mitmekesisust. Nagu sagedusloenditest kergesti võib näha, on tekstides kõige sagedasemad enam-vähem samad sõnad (enamasti grammatilised, tekstiehituslikud sõnad), samas kui täistähenduslikke lekseeme esineb harvem. Selleks, et mõõta, kas mingis tekstis või korpuses kasutatakse laia valikut erinevaid sõnu või korratakse hoopis pidevalt samu sõnu, võime kasutada erinevaid sõnavara mitmekesisuse statistikuid.

Kõige lihtsam sõnavara mitmekesisuse statistik on lihtne **tüübi- ja sõnesageduse suhe** (edaspidi tüübi-sõne suhe, ingl *type-token ratio*, TTR). Inglise keele kui suhteliselt vaese morfoloogiaga keele puhul on tüübisagedusena kasutatud tavaliselt erinevate sõnede hulka, mis tekstis esineb, ning sõnesagedusena nende kõiki esinemiskordi. Eesti keele kui morfoloogiliselt rikka keele puhul on tüübisagedusena sobivam kasutada tekstis või korpuses esinevate erinevate lemmade koguarvu ning sõnesagedusena nende kõiki esinemiskordi mistahes vormis. Tüübi-sõne suhte leidmiseks jagatakse tüübisagedus sõnede kogusagedusega.

> **Näide.** Lauses *Just keele abil on inimesed võimelised korda saatma suuri tegusid, sest keel on üks mõtlemise põhialuseid* on lemmade põhjal arvutatult sõnesagedus 16, tüübisagedus aga 14 (sest kaks lemmat – *keel* ja *olema* – korduvad). TTR selles lühikeses tekstis on niisiis 14/16 ehk 0,875.

Tüübi-sõne suhe iseloomustab teksti sõnavara mitmekesisust. Kui näiteks ühe teksti TTR on 0,8 ja teise enam-vähem sama pika teksti TTR 0,93, siis näitab see, et teine tekst on leksikaalselt mitmekesisem. Tüübi-sõne suhet saab kasutada seega teksti sõnavaralise mitmekesisuse hindamisel, ent teksti asemel võib sõnavara mitmekesisust hinnata ka mingis keelelises konstruktsioonis. Seda kasutatakse näiteks morfoloogilise produktiivsuse hindamisel: tüübisageduseks võib sel juhul olla näiteks erinevate *lt*-liiteliste määrsõnade arv ning sõnesageduseks *lt*-liiteliste määrsõnade tekstis esinemise kogusagedus (vt lähemalt M.-L. Pilviku morfoloogilise produktiivsuse näidisuurimust).

Tüübi-sõne suhte kasutamine on siiski piiratud, sest see on väga tundlik teksti pikkuse osas: mida pikem tekst, seda rohkem erinevaid sõnu seal on kasutatud, seetõttu sobib see vaid sarnase pikkusega tekstide sõnavara mitmekesisuse hindamiseks. Erineva pikkusega tekstide võrdlemiseks võib kasutada **standardiseeritud tüübi-sõne suhet** (ingl *standardized type-token ratio,* STTR; tuntud ka nimetuse all *mean segmental type-token ratio,* MSTTR). Selles on erineva pikkusega tekstide probleem lahendatud nii, et tekstid on tükeldatud ühesuuruse pikkusega lõikudeks (näiteks 1000 sõna), iga lõigu kohta arvutatakse välja oma TTR ning seejärel arvutatakse välja keskmine TTR. Tekstide viimane lõik, mis on teistest erineva pikkusega, jäetakse keskmise arvutamisest välja.

Kasutatakse ka kolmandat mõõdikut – **liikuva keskmisega arvutatud tüübi-sõne suhet** (ingl *moving average type-token ratio,* MATTR), mis mõõdab samuti

ühepikkuste lõikude keskmisi tüübi-sõne suhteid, ent need lõigud on osaliselt kattuvad: MATTR kasutab tekstilõikude moodustamiseks kindla suurusega akent, mis teksti peal nõnda liigub, et osa eelmisest lõigust satub ka järgmisse lõiku. Hiljem leitakse jällegi kõikide lõikude tüübi-sõne suhete keskmised.

5.2.4.4. Sõnamitmikud ehk n-grammid

Sagedusloendeid on võimalik teha ka sõnast suuremate üksuste põhjal, näiteks sõnamitmike põhjal. **Sõnamitmikeks** ehk **n-grammideks** nimetatakse vahetult kõrvuti asetsevate sõnede ühendeid. Sõnamitmikke koostatakse tavaliselt kahest sõnast (sõnakaksikud ehk bigrammid) või kolmest sõnast (sõnakolmikud ehk trigrammid), aga neid võib koostada ka näiteks neljast sõnast. Ka ühest sõnast koosnevat üksust võib sõnamitmikute osaks lugeda, neid nimetatakse unigrammideks. Sõnamitmikud moodustatakse järjestikku esinevate sõnede põhjal nii, et tekstis loetakse kokku iga sõne talle eelneva ja/või järgneva sõnega, ning neist paaridest või kolmikutest moodustatakse vastavalt bi- või trigrammide sagedusloend.

> **Näide.** Lausetest *See on tavaline tekst. See on ilus tekst.* saame järgmised sõnade bigrammid: *see on, on tavaline, tavaline tekst, tekst see, see on, on ilus, ilus tekst*. Neid sageduse põhjal kokku võttes näeme, et bigramm *see on* esines kahel korral, ülejäänud sõnapaarid ühel korral.

Sõnamitmikud iseloomustavad seega sõnade koosesinemust ja võimaldavad leida näiteks rohkem või vähem kinnistunud üksusi või mitmesõnalisi mõisteid, mis konkreetses korpuses või korpuseosas esilduvad. Sõnamitmike sageduse kaudu (nagu ka üksiksõnade sageduse kaudu) on võimalik iseloomustada registri või autori stiili, võrrelda eri ajaperioodide või registrite keelekasutust või esilduvaid märksõnu. Näiteks Google Booksi n-grammide tööriista[20] kaudu (mille hulka kuuluvad ka unigrammid ehk üksiksõnad) võib hinnata kultuuriliste nähtuste või tehnika arengut ajas, võrdle nt bigrammide *climate change, green energy, global warming* kasutamist. (Tuleb siiski meeles pidada, et Google Books ei ole representatiivne ega tasakaalus korpus, vaid pidevalt muutuv kollektsioon, mis võib sisaldada ka raamatute kordusväljaandeid, mis teeb tulemuste interpreteerimise keerukaks.)

Sõnamitmikke saab teha mitmel tasandil: sõne- (nt *see on*), lemma- (*see olema*) või sõnaliigi järjendina (asesõna-verb). Sõnetasand võimaldab hästi üles leida kinnistunud idioome või muid leksikaliseeruvaid üksusi, lemma ja sõnaliigi tasand võimaldavad aga hinnata tüüpilisi lausekonstruktsioone. Keeleõppes kasutatakse sõnamitmikke näiteks erineva taustaga keeleõppijate emakeele mõju uurimiseks teise keele omandamisel, sest sageli mõjutavad emakeeles kinnistunud lausekonstruktsioonid ka õppijakeele lausestruktuuri ning bi- või trigrammid võimaldavad

[20] https://books.google.com/ngrams/

seda esile tuua. Ka tõlke lähtekeele mõju hindamisel on võimalik n-gramme kasutada. Sageli rakendatakse selleks puhtleksikaalsete n-grammide asemel süntaktilisi n-gramme (st lauseliikmete või sõnaliikide n-gramme). Süntaktiliste n-grammide rakendusvõimaluste kohta teise keele omandamise ning tõlkimise uurimisel saab lugeda nt (Ivaska & Bernardini 2020).

Eesti keele jaoks on olemas mõningaid valmisandmestikke, mida omavahel või oma korpuse n-grammidega võib võrrelda. Kättesaadavad on näiteks tasakaalus korpuse ilukirjandus-, ajakirjandus- ja teadustekstide n-grammide sagedusloendid[21]; eesti uuema ilukirjanduse sagedusloendid ja sõnamitmikud (Raudvere & Uiboaed 2018) ning suuliste korpuste andmestikud (Lippus, Lõo, jt 2024), vt ka visualiseerimisrakendust[22]. Suulise kõne foneetilise korpuse ja teismeliste keele korpuse suuliste tekstide kõige sagedasemad bigrammid on esitatud ka tabelis 5.5. Need on koostatud sõnede põhjal, ent neid saab koostada loomulikult ka lemmade põhjal. Sõnamitmike koostamise funktsioonid on olemas enamikus korpusanalüüsi programmides (nt AntConc) ja Sketch Engine'is, samuti mitmetes R-i pakettides (nt pakettides `tidytext`, `quanteda`, `ngram`) ja Pythonis (nt teekides `NLTK`, `TextBlob`).

Tabel 5.5. Sagedasemad bigrammid Eesti teismeliste keele korpuses ja spontaanse kõne foneetilises korpuses (sõnede põhjal).

Teismeliste keele korpus		**Spontaanse kõne foneetiline korpus**	
Bigramm	**Sagedus**	**Bigramm**	**Sagedus**
ma ei	7311	ja siis	3306
ja siis	6652	ma ei	3242
see on	5662	see on	3195
ei tea	3618	ei ole	2721
ei ole	2866	ei tea	1965
on ju	2355	et see	1366
siis ma	2188	see oli	1219
mul on	2181	on see	996
on nagu	2084	on nagu	993
ma olen	1970	nii et	984

[21] https://cl.ut.ee/ressursid/mitmikud/

[22] http://suulinekeel.ut.ee

Teismeliste keele korpus		Spontaanse kõne foneetiline korpus	
Bigramm	**Sagedus**	**Bigramm**	**Sagedus**
see oli	1835	ja ja	975
et ma	1417	seal on	848
on see	1307	ta on	830
ta on	1235	et ma	776
et nagu	1196	et et	751
ei saa	1185	ei saa	743
nagu ma	1177	siis ma	721
ma olin	1171	see et	708
jaa jaa	1089	ei olnud	691
on nii	1056	siis on	689

5.2.5. Kollokatsioonid

Alati ei paikne tekstis koos esinema kalduvad sõnaühendid kindlas järjekorras kõrvuti (nagu sõnamitmike puhul), vaid võivad üksteisest veidi kaugemal asuda või sõnajärje varieeruvuse tõttu paikneda üksteise ees või järel, nii et oluliste sõnade vahele jääb veel mõni sõna. Sellegipoolest kalduvad nad esinema üksteise läheduses ja see pole päris juhuslik. Sellised on näiteks tugiverbiühendid (milles tähendust kannab nimisõna(fraas) ning üldise tähendusega verb seob selle lausestruktuuri ja annab edasi peamiselt grammatilist infot, nt *tööd tegema, kõnet pidama*) või ühend- ja väljendverbid (nt *üle jääma*, *aru saama*, *silmas pidama*). Nende ühendite osised paiknevad tekstis üksteisele lähedal, ent mitte tingimata kõrvuti ega ka mitte samas järjekorras; võrdle näiteks *aru saama* osiste paiknemist järgmistes lausetes: *Teisest inimesest aru saada on vahel väga raske. Sain temast suurepäraselt aru*. Selliseid sõnade olulisi koosesinemisi tekstis nimetatakse **kollokatsioonideks** (ingl *collocation*).

Kollokatsioone on defineeritud mitmeti, ent siinkohal käsitleme kollokatsioone kui sõnaühendeid, mida moodustavad sõnad esinevad tekstides koos sagedamini, kui võiks eeldada nende eraldi esinemise sagedustest. Kollokatsioonid koosnevad **põhjast** (mingi vaadeldav sõna) ja temaga koos esinema kalduvatest sõnadest, mida nimetatakse ka tema **kollokaatideks**.

Anatol Stefanowitsch (2020: 215–216) on nimetanud kolm suuremat põhjuste rühma, mille tõttu sõnad tekstides koos esinema kalduvad. Kõigepealt võivad

koosesinemisel olla grammatilised põhjused: mingid grammatilised elemendid esinevad lauses teatud kombinatsioonis. Näiteks määratleja ja nimisõna esinevad koos (*tol ajal, see aasta*); mitmeosalised verbivormid moodustavad koos tervikliku konstruktsiooni ning esinevad seetõttu üksteise lähedal (*on teinud, pidi tegema*), siia hulka kuuluvad ka eespool mainitud tugi-, ühend- ja väljendverbid. Teiseks võivad kollokatsioonid olla seotud tähendusega: näiteks verbi *jooma* sagedasimad kollokaadid on eesti keele ühendkorpuses 2023 *kohv, õlu* ja *vein*, st see vedelik (ja mitte näiteks tahked objektid), mida juuakse. Kolmas suurem rühm on maailmateadmistel põhinevad kollokatsioonid. Selle rühma kohta on näide joonisel 5.12, kus on esitatud sõna *Ukraina* sagedasemad kollokaadid eesti keele ühendkorpuse 2023 põhjal. Näeme, et see kajastab hästi 2022. aastal alanud laiapõhjalist Venemaa sissetungi ja seda, kuidas sellest kirjutatakse. Kollokatsioonid paljastavad ka seda, kuidas me puhttekstiliselt asjadest tavaliselt räägime: näiteks sõna *kohv* üks sagedasemaid kollokaate on *tee*, *gei* kõige sagedasem kollokaat on *lesbi*, sest tekstiliselt esinevad need sõnad sageli üksteise lähedal (*teed või kohvi, geid ja lesbid*).

		Lemma	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓	
1	☐	sõda	35,539	524,863	188.05	8.66	9.97	•••
2	☐	Venemaa	53,864	1,268,361	231.17	7.99	9.86	•••
3	☐	vägi	12,896	238,690	113.21	8.33	8.92	•••
4	☐	president	17,306	945,363	130.35	6.77	8.49	•••
5	☐	Valgevene	7,919	120,552	88.76	8.62	8.43	•••
6	☐	relvajõud	6,816	39,188	82.48	10.02	8.37	•••
7	☐	Ukraina	12,397	599,310	110.44	6.95	8.36	•••
8	☐	sõjapõgenik	6,590	21,750	81.13	10.82	8.36	•••
9	☐	Vene	10,345	440,002	100.99	7.13	8.30	•••
10	☐	Volodõmõr	6,179	9,648	78.59	11.90	8.30	•••
11	☐	sissetung	6,113	26,735	78.13	10.41	8.24	•••
12	☐	sõjaline	6,055	134,273	77.52	8.07	8.01	•••
13	☐	agressioon	4,998	24,508	70.64	10.25	7.96	•••
14	☐	Krimm	5,222	56,129	72.13	9.12	7.95	•••
15	☐	sõdur	5,869	178,318	76.22	7.62	7.89	•••

Joonis 5.12. Sõna *Ukraina* kõige tugevamad kollokaadid ÜK 2023 põhjal (Sketch Engine)

Kollokatsioonid võivad anda infot inimeste tüüpiliste mõttemustrite kohta, näiteks keeleliste stereotüüpide kohta. Liisi Piits uuris oma doktoritöös (2015) inimest tähistavate nimisõnade sagedasemaid kollokatsioone eesti keele koondkorpuses ning leidis nende põhjal hulgaliselt keeles kajastuvaid soostereotüüpe. Näiteks sõna *poiss* kõige sagedasemad kollokaatverbid on peamiselt aktiivset füüsilist tegevust tähistavad verbid ja agressiivset tegevust tähistavad verbid nagu *võitlema, tegutsema, tapma, lõhkuma, ehitama, kaklema* jne, sõna *tüdruk* kõige sagedasemad kollokaatverbid on aga *naeratama, itsitama, kiljuma, raputama, maksma, nentima* jne, seega peamiselt emotsionaalset käitumist, vaimset tegevust, suhtlemist ja verbaalset tegevust väljendavad verbid (Piits 2015: 114–116). Need kollokatsioonid põhinevad eesti keele koondkorpuse tasakaalus korpuse ajakirjandustekstidel, mis on pärit aastatest 1990–2001. Võib oletada, et kuna selle ajaga võrreldes on ühiskond oluliselt muutunud, on muutunud ka tekstid – ja koos sellega arvatavasti ka nende sõnade kollokaatide koosseis.

Kollokatsioonide omandamine on olulisel kohal võõrkeeleõppes. Näiteks inglise keelt võõrkeelena rääkijale võib olla raske otsustada, kas eesti väljendi *ettekannet pidama* vastena peaks inglise keeles kasutama sõnaühendit *make a presentation* või *give a presentation* või on mõlemad väljendid ühtmoodi sobivad. Eesti keele võõrkeelena õppija ei pruugi aga teada, et *näidet tuuakse*, mitte *ei anta*, või siis näiteks *kartuleid võetakse*, aga *marju korjatakse*. Eesti keele õppija tarvis on Eesti Keele Instituudis loodud Kollokatsioonisõnaraamat[23]. Sõnaraamat koostati automaatselt suurte eesti keele korpuste baasil Sketch Engine'i tarkvara abil, täpsemalt saab sellest lugeda artiklist „Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel" (Kallas, Koppel & Tuulik 2015) ja K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimusest.

Nagu öeldud, ei pea kollokatsiooni moodustavad sõnad asuma tekstis tingimata kõrvuti, vaid neid tuvastatakse nn kollokatsiooniaknas (st kindla ulatusega lõigus, ideaalis sama (osa)lause piires). Kui tekstis pole osalausepiire märgendatud, siis on tavalisim kollokatsiooniaken neli sõna (st kollokatsiooni moodustavate sõnade vahele võib jääda maksimaalselt 3 sõna). Kollokatsioone analüüsides on vaja seega läbi mõelda, kust me kollokatsioone otsime ja kui suur see aken peaks olema. Näiteks kui soovime uurida nimisõna juurde kuuluvaid täiendeid, siis tüüpiliselt paiknevad need vahetult nimisõna ees; seega peaks aken olema pigem väike ja paiknema otsitavast sõnast vasakul. Tugi-, ühend- ja väljendverbide puhul võib aga oluline komponent paikneda mõlemal pool otsitavat verbi ning kollokatsiooniaken võib olla üsna suur (nt *lööma* ja *kinni* ühendid: *ust pauguga kinni lüües, lõi ukse suure pauguga kinni*).

Tuleb arvestada, et kollokatsiooniakna abil koostatud sõnapaaride puhul on võimalus sagedaste juhuslike ühendite tekkeks väga suur. Seetõttu kasutatakse kollokatsioonide tugevuse hindamiseks mitmeid statistilisi **seose tugevuse**

[23] http://www.eki.ee/dict/kol/index.cgi?Q

mõõdikuid (ingl *association measures*), mis mõõdavad sõnade koosesinemust veidi erinevatest aspektidest.

Kollokatsioone võib hinnata ainult nende koosesinemise sageduse põhjal, ent see ei ole parim viis, sest toob esile vaid need kollokatsioonid, mis tekstides üldiselt sageli esinevad, aga mitte need, mis esinevad küll harvemini, ent järjekindlalt koos. Seose tugevuse mõõdikud aitavadki mõõta, kui tugevalt on sõna ja tema kollokaadid omavahel seotud, arvestades nii sõnade koos kui ka eraldi esinemise sagedust. Kuna seosed sõnade ja nende kollokaatide vahel võivad olla väga erinevad, on ka kollokatsioonide mõõtmiseks arendatud erinevaid mõõdikuid, ent neist enamiku arvutamiseks kasutatakse nelja põhilist näitajat: kahe sõna koos esinemise sagedus, 1. sõna üldine esinemissagedus, 2. sõna üldine esinemissagedus ning kogu korpuse suurus. Kolme viimase põhjal on võimalik välja arvutada ka kahe sõna koos esinemise oodatud/teoreetiline sagedus, mis väljendab sõnade koosesinemise eeldatud sagedust olukorras, kui nende vahel ei oleks mingit seost ning need satuksid lihtsalt juhuslikult aeg-ajalt lähestikku esinema tänu nende üldisele kasutussagedusele (vt ptk 6 „Korpusandmete statistiline analüüs“ ja tabel 6.4). Näiteks kui teame, et sõnaühendi *kange kohv* sagedus eesti keele ühendkorpuses 2023 on (selles näites täpselt selles kindlas järjekorras ja täpselt üksteise kõrval) 2204, sõna *kohv* esinemissagedus kokku on 240 631, sõna *kange* esinemissagedus kokku on 78 073 ja korpuses on kokku umbes 3 000 000 000 sõna, siis *kange kohv* järjendi oodatud esinemissagedus oleks 240 631 × 78 073 / 3 000 000 000 = 6,3. Tegelik esinemissagedus (2204) on aga oodatust sadu kordi suurem.

Kollokatsioonide leidmise mõõdikud jagunevad laias plaanis kahte rühma: ühed neist väljendavad kahe sõna vahelise **seose tugevust** ja võrdlevad tegelikku koosesinemise sagedust oodatud koosesinemise sagedusega; teised mõõdikud hindavad ka **seose statistilist olulisust** ehk seda, kui tõenäoline on, et sõnade koosesinemine on juhuslik (vt ptk 6 „Korpusandmete statistiline analüüs“). Erinevalt teisest rühmast kipuvad seose tugevuse mõõdikud esile tooma kollokaate, mis on huvipakkuva sõnaga küll tugevalt seotud, ent muidu korpuses võrdlemisi harvad. Üldise ülevaate levinumatest mõõdikutest annab tabel 5.6 (vt ka Brezina 2018: 66–79).

Tabel 5.6. Ülevaade levinumatest kollokatsioonide mõõdikutest

Mõõdik	Tüüp	Sümmeetrilisus	Iseloomustus	Puudused
MS (minimaalne tundlikkus, ingl *minimum sensitivity*)	seose tugevuse mõõdik	sümmeetriline	tugevate bigrammide tuvastamiseks; ei ole väga tundlik korpuse suuruse suhtes	võib ületähtsustada väga harva esinevaid sõnu
MI (ingl *mutual information*)	seose tugevuse mõõdik	sümmeetriline	tugevate ja eksklusiivsete kollokatsioonide tuvastamiseks (nt idioomid, terminid)	võib ületähtsustada väga harva esinevaid sõnu (soovitatav kasutada koos vähima sageduse lävendiga); ei too hästi esile väga sagedasi kollokatsioone
MI2 (väikeste sageduste suhtes kohandatud MI)	seose tugevuse mõõdik	sümmeetriline	tugevate ja eksklusiivsete kollokatsioonide leidmiseks (nt idioomid, terminid); vähem tundlik väga harvade sõnade suhtes	ei pruugi hästi sobida ei väga harvade ega väga sagedaste kollokatsioonide tuvastamiseks
logDice	seose tugevuse mõõdik	sümmeetriline	tugevate ja eksklusiivsete kollokatsioonide tuvastamiseks (nt idioomid, terminid); vähem tundlik väga harvade sõnade suhtes	ei pruugi hästi sobida ei väga harvade ega väga sagedaste kollokatsioonide tuvastamiseks
ΔP (*Delta P*)	seose tugevuse mõõdik	asümmeetriline	ühesuunaliste seoste tuvastamiseks (nt grammatiliste suhete puhul)	võib olla keeruline tõlgendada

Mõõdik	Tüüp	Sümmeet-rilisus	Iseloomustus	Puudused
z-skoor	statistilise olulisuse mõõdik	sümmeetriline	tugevate ja eksklusiivsete kollokatsioonide tuvastamiseks	võib ületähtsustada madala sagedusega sõnu; toetub normaaljaotusele, mis ei ole sõnasageduste puhul enamasti kohane
t-skoor	statistilise olulisuse mõõdik	sümmeetriline	sagedaste kollokatsioonide tuvastamiseks	võib ületähtsustada kõrge sagedusega koosesinemisi ja mitte tuvastada harvaesinevaid, aga väga tugevaid kollokatsioone
log-tõepära (ingl *log-likelihood*), G^2	statistilise olulisuse mõõdik	sümmeetriline	üks usaldusväärsemaid mõõdikuid kollokatsioonide tuvastamiseks	võib ületähtsustada kõrge sagedusega koosesinemisi, eriti väga suurtes korpustes

Mõned seose tugevuse mõõdikud, nt MI, rõhutavad kollokatsioonisuhte välistavust, eelistades neid kollokaate, mis esinevad peaaegu alati vaadeldava sõnaga koos, isegi kui seda juhtub vaid korra-paar kogu korpuses (nt *lööma* kollokaadid *lokku* ja *hingekella*, mis esinevad harva ja pea alati ainult koos sõnaga *lööma*). Teised mõõdikud, nt logDice ja MI2, eelistavad selliseid kollokaate, mis esinevad peaasjalikult üksteisega koos, ent ei pruugi olla harvad (nt logDice'i põhjal on omadussõna *jändrik* sagedasimad kollokaadid *mänd, rannamänd, rabamänd ja okslik*; sõna *lahus* kõige tugevam kollokaat on *füsioloogiline*).

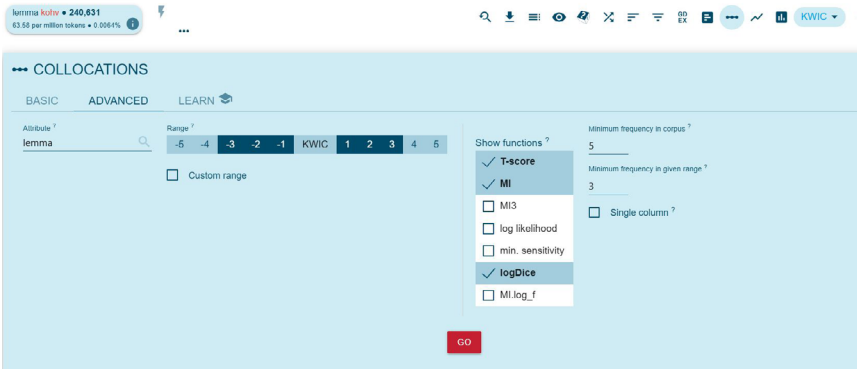
Enamik mõõdikuid, mida kasutatakse, eeldavad sõnade vastastikust tõmbumist ning seetõttu nimetatakse neid **sümmeetrilisteks** mõõdikuteks. Samas on teada, et inimese teadvuses ei ole seos kahe sõna vahel alati sümmeetriline ning sümmeetrilised mõõdikud ei tuvasta, kas esimene sõna on abiks teise ennustamisel või vastupidi. Näiteks ühend *lahku minema* sõnade vastastikune tõmbumine ei ole sümmeetriline: kui osalauses on *lahku*, siis on seal väga suure tõenäosusega ka *minema* ehk *minema* esinemine on ennustatav *lahku* järgi, kuid mitte vastupidi (Aedmaa 2014: 26). **Asümmeetrilised** mõõdikud selliste suunaliste seoste

tuvastamiseks on näiteks lihtne tinglik tõenäosus (ingl *conditional probability*) ja ΔP (Gries 2013).

Eesti keeles on kollokatsioonide mõõdikuid ühendverbide leidmiseks süstemaatiliselt võrrelnud Eleri Aedmaa oma magistritöös (2014). Ta katsetas erinevaid sümmeetrilisi ja asümmeetrilisi mõõdikuid selleks, et tasakaalus korpuse kolmest allkeelest (ilukirjandus, ajakirjandus, teaduskirjandus) üles leida ühendverbid (*ära minema*, *üles leidma* jms) ning võrdles mõõdikute efektiivsust. Vaadeldud mõõdikutest osutus selle ülesande täitmiseks kõige sobivamaks sümmeetriline ja suhteliselt lihtne olulisust mõõtev statistik t-skoor (ingl *t-score*). Mõõdikute tulemusi ühendverbide tuvastamisel mõjutas korpuse suurus: t-skoori ja enamiku teiste vaadeldud mõõdikute tulemused paranevad korpuse suuruse kasvades. Vastupidise efekti annab korpuse mahu kasvamine juba ülal mainitud vastastikuse informatsiooni väärtuse ehk MI-mõõdiku tulemustele.

Paljudes korpusanalüüsi programmides ja keskkondades (nt AntConc, Sketch Engine) on osa neist mõõdikutest n-ö sisse ehitatud ning neid on võimalik katsetada ja vahetada, ent seejuures on oluline teada, et mõõdikud tõstavad esile erinevaid kollokaate ning seega sobivad erinevate ülesannete jaoks. Kui oma uurimuses kollokatsioonide analüüsi kasutada, on seega väga oluline selgeks teha ja raporteerida, mis mõõdikut on kasutatud ja miks.

> **Näide.** Vaatame, millised on sõna *kohv* kõige tüüpilisemad kollokatsioonid eesti keele ühendkorpuse 2023 põhjal. Eesmärk on välja selgitada, kuidas kohvi eesti keeles iseloomustatakse. Selle ülesande lahendamiseks kasutame Sketch Engine'it. Kollokatsioonide leidmiseks on Sketch Engine'is kaks võimalust.
>
> 1. Alustame sellest, et teeme konkordantsipäringu sõnaga *kohv* (lemma põhjal) ning seejärel valime tulemuste lehe tööriistaribalt ikooni *Collocations*. Avanenud lehel on sakil *Advanced* võimalik valida kuvatavad kollokatsioonide leidmise mõõdikud (t-skoor, MI, MI3, log-tõepära, minimaalne tundlikkus, logDice ja MI.log_f). Võime valida katsetamiseks kõik mõõdikud ning hiljem nende tulemusi omavahel võrrelda. Lisaks saame valida kollokatsiooniakna suurust (*Range*) ning täpsustada, mis tasandi kollokaate otsime (*Attribute*). Valime atribuudiks *lemma*, mis sobib leksikaalsete kollokatsioonide leidmiseks kõige paremini (vt joonis 5.13).



Joonis 5.13. Kollokatsioonide leidmine konkordantside põhjal

Tulemuste lehel saame mõõdikute põhjal tulemusi võrrelda (mõõdikutel klõpsates). Lisaks on esitatud sõna *kohv* ja kollokaatide koosesinemissagedused ning kollokaatide eraldiesinemissagedused. Näeme, et MI annab tugevaimateks kollokaatideks tõepoolest põhiliselt harvaesinevaid sõnu (sageli pärisnimed, nt *ColdBrew*, *Kively*, või ka trükivead, nt *jöome*, *Nautimisekstimemore*), aga ka nt väga tugevalt ainult sõnaga *kohv* kollokeeruvad kollokaadi *eeljahvatatud* ja *külmpruulitud*; teiste mõõdikute puhul ei ole erinevused nii suured. Vaikimisi reastab Sketch Engine tulemused logDice mõõdiku järgi (vt tabel 5.6), mis üldiselt kipub eelistama sõnaühendeid, mille sõnad üksikuna ei ole korpuses väga sagedased. Järgmisena võime valida siit välja need tulemused, mis on ülesande püstituse seisukohalt olulised, st iseloomustavad kohvi: *kohv* on *kange*, *jahvatatud*, *lahustuv*, *kuum*, *maitsev*, *must* jne. Aga kollokatsioonidest tuleb välja ka see, kuidas kohvi tarbitakse: kollokatsioonide eesotsas on *tass* ja *tassike*, *jooma* ja *rüüpama*, *valmistama* ja *keetma* ning kõik see, mida kohvi juurde tarbitakse: *kook*, *piim*, *suhkur*, *saiake* jms. Lisaks on kollokaatide hulgas sõnad, mis viitavad sellele, mida kohvi alternatiivina pakutakse: *kohv* või *tee*, *kohv* või *kakao*. Seega saame kollokatsioonide kaudu üpris palju teada kohvijoomise traditsioonide kohta. Seda, kui paljud sagedusjärjestuses olevatest kollokaatidest meile vaadeldava sõna kohta veel midagi olulist räägivad, tuleb otsustada uurijal endal, seda ei ole võimalik mõõdikute abil täpselt öelda. Analüüsis ei saa märkamata jätta ka neid kollokaate, mis esmapilgul tunduvad veidrad. Miks on sagedusloendis kõrgetel positsioonidel näiteks *Hillar* ja *joon*? Selles näites on *Hillar* seotud nimega *Hillar Kohv* (omaaegne ajalehtede kirjasaatja Pärnumaalt), *joon* on tõenäoliselt aga verbi *jooma* valesti tuvastatud vorm.

Tulemuste raporteerimisel tuleks kindlasti esitada statistik, mille alusel on tulemused saadud, kollokatsiooniakna suurus ning kollokaadi tüüp (lemma). Kollokaatide tabelis võiks esitada kollokaadi positsiooni, koos- ja eraldiesinemissageduse ning seose tugevust väljendava(te) mõõdiku(te) väärtuse(d).

Tuleb arvestada, et see uuring kajastab kogu ühendkorpuses 2023 leiduvat ja kirjeldab üpris hästi eestlaste tüüpilist kohvijoomiskultuuri. Teist tüüpi kollokatsioonide analüüsil võib olla mõttekas aga võrrelda kollokatsioone erinevate korpusosade või eri ajaperioodide lõikes, et analüüsida keelelist, kultuurilist või ühiskondlikku muutust, mis keelekasutuses kajastub.

Tabel 5.6. Sõna kohv 25 sagedasemat kollokaati, järjestatud logDice mõõdiku alusel. Tabelis on antud ka muude mõõdikute tulemused

Kollokaat	Koosesinemissagedus	Eraldiesinemissagedus	T-skoor	MI	logDice	MI3	log-tõepära	MS	MI.log f
tass	13 208	77 444	114,88	11,39	10,41	38,77	185 264,35	0,05	108,07
jooma	25 360	391 352	159,09	9,99	10,36	39,25	305 100,76	0,06	101,34
valmistamine	5766	233 808	75,74	8,60	8,64	33,59	57 517,92	0,02	74,47
joomine	3446	60 476	58,64	9,81	8,55	33,31	40 218,15	0,01	79,89
kange	3063	77 589	55,26	9,28	8,30	32,44	33 442,61	0,01	74,48
kohv	4514	237 562	66,96	8,22	8,27	32,50	42 632,13	0,02	69,20
tee	21 280	2 085 080	144,97	7,33	8,23	36,08	175 975,50	0,01	73,01
jahvatatud	2326	15 652	48,21	11,19	8,22	33,56	31 820,99	0,01	86,75
kook	3396	140 400	58,12	8,57	8,19	32,03	33 709,71	0,01	69,69
piim	3744	208 339	60,97	8,14	8,09	31,88	34 927,81	0,02	67,00
keetma	2497	66 324	49,89	9,21	8,06	31,78	27 015,81	0,01	72,05
lahustuv	2022	10 281	44,95	11,59	8,04	33,56	28 902,47	0,01	88,26
tassike	1912	7090	43,72	12,05	7,98	33,85	28 702,07	0,01	91,06
rüüpama	1949	19 022	44,12	10,65	7,94	32,51	25 113,43	0,01	80,71
kuum	3495	261 599	58,84	7,72	7,83	31,26	30 521,07	0,01	62,95
Hillar	1646	19 813	40,54	10,35	7,69	31,72	20 483,18	0,01	76,67
suhkur	2770	203 891	52,38	7,74	7,67	30,61	24 274,41	0,01	61,35
kõrvale	4120	429 999	63,76	7,24	7,65	31,25	33 249,90	0,01	60,23
joon	2766	215 918	52,33	7,65	7,63	30,52	23 913,65	0,01	60,67
maitse	3502	370 442	58,78	7,22	7,55	30,76	28 159,53	0,01	58,89
võileib	1593	40 380	39,85	9,28	7,54	30,55	17 380,79	0,01	68,41
saiake	1295	18 510	35,95	10,10	7,36	30,78	15 650,85	0,01	72,41
kakao	1312	22 192	36,18	9,86	7,35	30,58	15 400,14	0,01	70,80
maitsev	1882	147 752	43,17	7,65	7,31	29,40	16 242,57	0,01	57,66
must	4349	678 620	65,29	6,66	7,28	30,83	31 618,13	0,01	55,76

2. Teine viis kollokatsioonide leidmiseks Sketch Engine'ist on kasutada sõnavisandite funktsiooni *Word Sketch*. See on mõeldud ennekõike sõnaraamatute tegijatele, kel on vaja kiiresti välja selgitada sõna kõige tüüpilisemad kasutusviisid, ent mitte analüüsida neid nii põhjalikult, kui muidu teadustöös vajalik. Word Sketchi väljundis on tulemused grupeeritud selle järgi, mis lauseliikmete või sõnaliikidega need peamiselt koos esinevad ning lisaks antakse infot selle kohta, millistes alamkorpustes mingid kasutusviisid kõige enam esile tulevad. Word Sketch kasutab mõõdikuna logDice statistikut. Tulemused on järjestatud seose tugevuse alusel. Selle kohta vt ka K. Koppeli, J. Kallase ja M. Langemetsa näidisuurimust.

Kui kollokatsioonid on leksikaalsete üksuste koosesinemised, siis grammatika uurimisel analüüsitakse sageli hoopis kollostruktsioone (ingl *collostruction*). **Kollostruktsioonid** keskenduvad sõnade ja grammatiliste konstruktsioonide vahelistele seostele: kui tugevalt on leksikaalsed üksused seotud kindlate grammatiliste konstruktsioonidega (Stefanowitsch 2020: 270). Kollostruktuurilise analüüsi kohta loe täpsemalt J. Padriku näidisuurimusest.

5.2.6. Võtmesõnad

Sageli ei huvita meid mitte sõnade üldsagedused, vaid see, millised sõnad vaadeldavas tekstis või korpuses sageduse poolest eriti esile tõusevad ning iseloomustavad just seda teksti või korpust eriti kujukalt, võrreldes mingi muu („tavalise") korpusega. **Võtmesõnade analüüs** aitab esile tuua neid sõnu, mis on kas sisu, žanri või stiili poolest olulised vaatluse all oleva teksti või korpuse iseloomustamiseks. Võtmesõnad on seega suhtelised ning sõltuvad sellest, mis korpust me millega võrdleme.

Võtmesõnad (ingl *keywords*) on seega need sõnad, mille sagedus **fookuskorpuses** (korpuses, mis meid huvitab) on oluliselt erinev nende sagedusest **referentskorpuses** (korpuses, millega me fookuskorpust võrdleme). Positiivsed võtmesõnad on need võtmesõnad, mis esinevad fookuskorpuses oluliselt sagedamini kui referentskorpuses. Negatiivsed võtmesõnad esinevad fookuskorpuses oluliselt harvemini kui referentskorpuses. Vahel räägitakse ka lukksõnadest (ingl *lockwords*), mis esinevad fookuskorpuses ja referentskorpuses võrdselt sageli ning seetõttu ei paku olulist informatsiooni fookuskorpuse kohta. Võtmesõnad võivad olla ka mitmesõnalised üksused – nimetame neid siinkohal **mitmesõnalisteks võtmesõnadeks**.

Võtmesõnade analüüsimiseks on oluline läbi mõelda, missugust referentskorpust valida, sest sellest sõltub võtmesõnade informatiivsus. Referentskorpus peaks olema fookuskorpusega mingis mõttes sarnane. Näiteks võib 1930ndate aastate ilukirjanduskeelt võrrelda 1930ndate aastate ajalehtedes kasutatud keelega või hoopis tänapäeva ilukirjanduskeelega – need võrdlused osutavad küll erinevatele aspektidele 1930ndate ilukirjanduskeele juures, ent on kõnekamad kui võrdlus näiteks

tänapäeva ajalehekeelega. Teisalt on soovitatud, et referentskorpus võiks olla võimalikult suur (mahult suurem või vähemalt võrdne fookuskorpusega). Hästi sobib referentskorpuseks ka üldkeele korpus, sest nii tõusevad fookuskorpuse eripärad hästi esile. Praktikas võib fookuskorpust võrrelda ka mitme referentskorpusega ning vaadata, millised võtmesõnad eri referentskorpustega võrreldes esile tõusevad.

Ka võtmesõnade leidmiseks kasutatakse mitmeid mõõdikuid. Levinuim on log-tõepära (ingl *log-likelihood*), mis hindab seda, kui tõenäoline on, et sõna sageduserinevus kahes korpuses on juhuslik. Võtmesõnade analüüs on samuti vahendina sisse ehitatud mitmesse korpuslingvistika töövahendisse, näiteks AntConci, kus on võimalik katsetada erinevaid mõõdikuid ning leida nii positiivseid kui negatiivseid võtmesõnu.

Võtmesõnade analüüs on ka Sketch Engine'i töövahendite hulgas. Seal on võimalik analüüsida nii võtmesõnu kui mitmesõnalisi termineid, mõõdikuna kasutatakse seal aga lihtsamat võtmesuse skoori, mis toetub sõnade suhtelisele sagedusele fookus- (F) ja referentskorpuses (R):

$$võtmesuse\ skoor = \frac{suhteline\ sagedus\ (F) + 1}{suhteline\ sagedus\ (R) + 1}$$

Tabel 5.7. esitab ühendkorpuse 2023 akadeemiliste tekstide alamkorpuse mitmesõnalised võtmesõnad, referentskorpuseks on ühendkorpus 2021. Näeme, et lisaks keeleteaduse terminitele (mille esildumine tuleneb alamkorpuse aluseks olevate avaliku ligipääsuga eestikeelsete teadusajakirjade valikust) tõusevad võtmesõnadena esile ühelt poolt kirjastuste nimed, mis tulevad arvatavasti viidete loenditest, ning teiselt poolt akadeemilises kirjutamises kasutatavad väljendid (*siinne artikkel, vt tabel, artikli eesmärk* jne).

Tabel 5.7. Mitmesõnalised võtmesõnad akadeemilistes tekstides eesti keele ühendkorpuse 2023 akadeemiliste tekstide alamkorpuse põhjal (Sketch Engine)

Mitmesõnaline võtmesõna	Sagedus fookus-korpuses	Sagedus referents-korpuses	Suhteline sagedus fookus-korpuses	Suhteline sagedus referents-korpuses	Võtmesuse skoor
tartu ülikooli kirjastus	663	1833	60,12888	0,62232	37,68
eesti rakendus-lingvistika	451	633	40,90215	0,21491	34,49
ühingu aastaraamat	436	559	39,54177	0,18979	34,08
siinne artikkel	546	1702	49,51790	0,57784	32,02

Mitmesõnaline võtmesõna	Sagedus fookus-korpuses	Sagedus referents-korpuses	Suhteline sagedus fookus-korpuses	Suhteline sagedus referents-korpuses	Võtmesuse skoor
vt tabel	663	2769	60,12888	0,94010	31,51
siinne uurimus	360	389	32,64917	0,13207	29,72
suuline ajalugu	371	582	33,64678	0,19759	28,93
kaasav haridus	653	3526	59,22196	1,19711	27,41
eesti keele sihtasutus	470	1897	42,62530	0,64405	26,54
keele sihtasutus	484	2419	43,89499	0,82127	24,65
rakenduslingvistika ühing	288	459	26,11933	0,15583	23,46
keele omandamine	431	2256	39,08831	0,76593	22,70
artikli eesmärk	385	1728	34,91647	0,58667	22,64

Lõpetuseks

Selles peatükis tutvustati mitmeid korpuslingvistikas sageli kasutatavaid meetodeid ja lähenemisi. Suur osa neist on seotud sõnavara analüüsiga, mis pakub huvi lisaks keeleteadlastele ka neile, kes uurivad keele kaudu kultuuri ja ühiskonda. Peatükis tutvustasime ka seda, kuidas läheneda grammatikale ja kuidas kasutada morfoloogiliselt märgendatud tekste.

Peatükis on selgitused ja näited nende analüüsimeetodite kohta, mis on mingil moel kasutatavad laialt levinud korpusanalüüsi tarkvara abil. R-i või Pythonit (ja nende erinevaid pakette ja teeke) kasutades kasvab võimalike analüüsivahendite ja -meetodite hulk, ent paraku nõuavad need kasutajalt enam teadmisi ja oskusi ning nende tutvustamine ei ole algajatele suunatud õpikus mahu tõttu võimalik. Lisaks muutub see maailm kiiresti. Soovitame otsida mujalt võimalusi nende meetoditega tutvumiseks.

Peatükist nägime, et korpuslingvistikas tehakse palju järeldusi sageduse põhjal, ent oluline on mõelda läbi, kas parasjagu on informatiivsem absoluutne, suhteline või hoopis normaliseeritud sagedus. Tõdesime ka, et eri mõõdikud, mida kasutatakse kollokatsioonide ja võtmesõnade leidmiseks, võivad viia erinevate tulemuste ja järeldusteni.

Peatükis oli fookuses ennekõike sagedusinfo, ent korpusandmete põhjal on võimalik teha ka kvalitatiivset uurimust: näiteks suulise kõne diskursuspartiklite sagedusandmetele lisaks on võimalik lähemalt analüüsida iga partikli kasutuskonteksti ja selle kaudu otsustada partikli funktsiooni üle lauses; võimalik on uurida sagedaste verbide tähendusi ja konstruktsioone, mida need moodustavad jne. Omaette meetodiks on kujunenud korpuslingvistika sagedusandmete ja (kriitilise) diskursusanalüüsi sidumiseks – inglise keeles *corpus-assisted discourse studies* ehk CADS. See meetod on kasulik siis, kui korpusest tulevad sõnad vajavad laiemat interpretatsiooni, need kannavad edasi suhtumisi ja sildistavad inimrühmi või nähtusi. Selle meetodi puhul on oluline süveneda ka sellesse, missugustest tekstidest konkordantsiread tulevad, mis on nende kontekst ja mis maailmavaadet need kannavad. CADS-i kohta vt nt (Baker 2023).

Statistika ja sagedusandmete interpretatsiooni rollist korpuslingvistikas räägime põhjalikumalt raamatu järgmises, 6. peatükis. Statistiliste meetodite rakendamist saab jälgida ka näidisuurimustest.

Lisalugemiseks

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Evert, Stefan. 2009. Corpora and collocations. Anke Lüdeling & Merja Kytö (toim), *Corpus Linguistics*, 1212–1248. Berlin / New York: De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.1212.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next .... *International Journal of Corpus Linguistics* 18(1). 137–166. https://doi.org/10.1075/ijcl.18.1.09gri.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7), 211–259. Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735821.

6. Korpusandmete statistiline analüüs

Maarja-Liisa Pilvik

Selles peatükis kirjeldame keeleandmete statistilist analüüsi. Sõna *statistiline* viitab sellele, et korpuses esinevatest keeleandmetest on vaja esmalt tuletada mõõtmise või loendamise teel arvandmed. Näiteks võime mõõta suulise korpuse vestlustes kõnelejate kõnetempot või formantide väärtusi, samuti võime lugeda kokku tekstides esinevaid sõnu, grammatilisi kategooriaid vms. Arvandmete põhjal saame seejärel anda ülevaate mingitest tüüpilistest väärtustest ja koosesinemistest ning tuvastada ja hinnata keelekasutuses esinevaid mustreid ja seoseid. Ehkki keelt saab uurida ka ilma statistilise analüüsita, on statistiliste meetodite kasutamine loomulik osa empiiriliste andmete analüüsiprotsessist. Seetõttu sobib see hästi ka korpuslingvistika tööriistakasti ning aitab suurtes andmehulkades uurida seaduspärasid, mida palja silmaga ei pruugiks näha.

Ka eelmistes peatükkides kirjeldatud meetodid võivad hõlmata statistilist analüüsi. Näiteks põhineb kollokatsioonide leidmine sõnade eraldi ja koos esinemiste sageduste statistilisel võrdlemisel (vt peatükk 5.2.5 „Kollokatsioonid“). Sellisel juhul hindame kahe või enama sõna vahelist seost: kui hästi aitab ühe sõna teadmine ennustada teise sõna esinemist tema läheduses? Seoste olemasolu, suunda ja tugevust võime hinnata aga ka muude keelenähtuste ja neid või nende kasutust iseloomustavate tunnuste puhul. Näiteks võime uurida korpuses sõnade pikkuse ja sageduse vahelist seost, teismeliste keelekasutuses ingliskeelsete sõnade osakaalu sõltumist kõneleja vanusest ja soost või hoopis vestluse teemast, eri sõnaliikide esinemissageduste seost tekstitüübi ja registriga jne. Seetõttu räägime selles peatükis statistilisest analüüsist põhjalikumalt. Olgu öeldud, et ehkki siin on statistilise analüüsi näitlikustamiseks kasutatud programmeerimiskeelt ja tarkvara R (v4.3.2, R Core Team 2023) ja selle laiendust RStudio (v2023.12.1, Posit team 2024), on võimalik enamikku siin kirjeldatud meetodeid rakendada ka muudes programmides (nt SAS, SPSS, STATA või ka Excel) ja programmeerimiskeeltes (nt Python). Analüüside kordamise hõlbustamiseks on jooniste ja muude väljundite ees esitatud ka nende saamiseks kasutatud R-i kood.

Selles peatükis kirjeldatud statistilised meetodid eeldavad, et meil on kasutada mingi (näiteks korpusest kogutud) **andmestik**. Selline andmestik on tüüpiliselt struktureeritud tabelkujul, igal real uuritava keelenähtuse kasutusjuht ehk **vaatlus** (ingl *observation*) ning igas tulbas seda kasutusjuhtu iseloomustavad **tunnused**

ehk **muutujad** (ingl *variable*, vahel ka *attribute* või *feature*). See, mis täpselt on üks vaatlus, sõltub muidugi uuritavast nähtusest ja uurimisküsimusest. Kui uurime, millest sõltub sõnaalgulise /*h*/ hääldamine või mittehääldamine, on vaatluseks iga *h*-häälikuga algava sõna üks kasutusjuht korpuses ja tunnusteks näiteks *h* häälda-mine / hääldamata jätmine, suhtlusolukord, milles konkreetne sõna on öeldud, sõna rõhulisus, see, kas sõna on nimi või mitte, *h*-le eelnev ja järgnev häälik, aga ka kõneleja omadused, nt kõneleja vanus, sugu, päritolu, või haridus (vt nt Teras 2019). Kui uurime aga, millest sõltub erinevate kõnelejate keskmine kõnetempo, võib vaatluseks olla iga eraldi kõneleja, tunnusteks tema keskmine kõnetempo (nt silpe sekundis), sugu ja vanus ning vestluspartneri sugu, vanus ja kõnetempo (vt nt Lippus, Pilvik, jt 2024). Siinjuures saame eristada **uuritavat tunnust** (ingl *dependent variable*, *response (variable)* või *outcome (variable)*), mis on mingi vaatluste omadus, mis uurijale parajasti huvi pakub (nt *h* hääldub/ei hääldu; iga kõneleja keskmine silpide arv sekundis), ja **seletavaid tunnuseid** (ingl *independent variable, explanatory variable* või *predictor (variable)*), mille abil uuritava omaduse muutumist või varieerumist seletatakse.

Nii uuritavaid kui ka seletavaid tunnuseid võib nende tüübi järgi omakorda jagada **arvulisteks** (nt pikkus, tempo, sagedus, vanus) ja mittearvulisteks ehk **kategoorilisteks tunnusteks** (nt suhtlusolukord, žanr, kõneleja sugu, kääne, isik, kõneviis). Arvuliste tunnuste väärtustega saab teha kõiksugu matemaatilisi tehteid (nt liitmine, lahutamine, keskmiste arvutamine), kategoorilisi tunnuseid saab aga ainult omavahel võrrelda (nt „A“ ei ole „B“) ja kokku lugeda (nt „A“ esineb 5 korda, „B“ 20 korda). Kategooriliste tunnuste all eristatakse omakorda **järjestustunnuseid** (nt keeleoskus või haridustase), mille puhul saame küll öelda, et üks väärtus on suurem/rohkem kui teine, ent pole selge, kas väärtuste vahe on ühesugune (nt kas A1 ja A2 keeleoskuse tasemete vahel on sama palju keeleoskuse erinevust kui A2 ja B1 keeleoskuse tasemete vahel). Ka kõiksugu küsitluste kategooriatega skaalavastused (nt „ei nõustu üldse“, „pigem ei nõustu“, „pigem nõustun“, „nõustun täielikult“) on järjestustunnused, ehkki vahel teisendatakse neid praktilistel põhjustel ka arvudeks (nt $1 < 2 < 3 < 4$). Tunnuse tüüp on oluline, kuna sellest sõltub, milliseid statistilisi näitajaid ja teste saame kasutada.

Tabelis 6.1 on toodud näide Lippus jt (2024) kõnetempo varieerumist käsitlevas artiklis kasutatud eesti spontaanse kõne foneetilise korpuse (EKSKFK, Lippus, Aare, jt 2023) andmestiku esimesest 6 reast, kuhu korpuse metaandmete põhjal on lisatud ka kõnelejate haridustase. Igal andmestiku real on ühe kõneleja andmed ühes vestluses: tema vanus, sugu, haridustase, kõneleja kõnesoleku aeg vestluse ajal (kestus sekundites) ning kogu vestluse jooksul öeldud silpide arv. Kui mõni kõneleja on osalenud mitmes salvestuses koos erinevate kaaskõnelejatega (nagu nt KJ2), on ta andmestikus esindatud mitu korda.

Tabel 6.1. Esimesed 6 rida EKSKFK-st kogutud kõnetempo varieerumise andmestikust

	fail	koneleja	vanus	sugu	haridus	kestus	silpide_arv
1	fon_1	KJ1	26	M	kõrg	538,76	1686
2	fon_1	KJ2	22	N	kõrg	848,78	3146
3	fon_2	KJ3	36	N	kõrg	1127,28	4522
4	fon_2	KJ2	22	N	kõrg	312,09	1337
5	fon_3	KJ4	30	N	kõrg	1960,21	8843
6	fon_3	KJ5	27	N	kõrg	357,85	1694

Tabelis on arvulised tunnused *vanus* (loendusandmed), *kestus* (mõõtmisandmed) ja *silpide_arv* (loendusandmed), kategoorilised ltunnused *fail*, *koneleja*, *sugu* ning järjestustunnusena käsitletav *haridus* (keskharidus on rohkem haridust kui põhiharidus, kõrgharidus on rohkem haridust kui keskharidus). Silpide arvu ja kõnesoleku kestuse põhjal saab omakorda tuletada uue suhtarvulise tunnuse kõnetempo hindamiseks: mitu silpi sekundis kõneleja keskmiselt vestluse ajal ütleb.

Sageli räägitakse üldkeeles statistikast ja statistilisest analüüsist **kirjeldava statistika** tähenduses. Kirjeldava statistika eesmärk on anda ülevaade andmete jaotumisest ja tüüpilistest väärtustest, ent kirjeldav statistika ei võimalda üldjuhul teha väga laiaulatuslikke järeldusi uuritava nähtuse kohta üldisemalt ega kinnitada või ümber lükata hüpoteese. Seda seetõttu, et nii korpusuuringuid kui ka katseid ja küsitlusi tehes saame enamasti andmeid ainult väikese osa (**valimi**) kohta kogu keelest, selle kõnelejatest, mingist kõnelejarühmast, keelekasutussituatsioonist, lausetüübist vm-st üldkogumist (**populatsioonist**), mille kohta tervikuna midagi öelda tahaksime (vt ka peatükke 1.2.1 „Keeleandmete populatsioon ja valim“ ja 4.2 „Korpuse koostamise põhimõtted“). Kirjeldava statistika abil saab iseloomustada niisiis näiteks mingit nähtust konkreetses korpusest kogutud andmestikus või paremal juhul kogu korpuses, aga mitte üldkogumis, välja arvatud harvadel juhtudel, kui meil on kasutada andmed kogu populatsiooni kohta (nt kõik Tammsaare ilmunud teosed, kõik kunagi kirjutatud eesti keele õpikud vmt).

Järeldav statistika koondab enda alla meetodeid, mis võimaldavad järeldada valimi põhjal midagi ka populatsiooni kohta üldisemalt ning hinnata sealjuures seda, kui kindlad me mingis üldistuses saame olla. Järeldava statistika abil saame niisiis teatud eksimisvõimalusega teha ennustusi ja oletusi ka meie valimist väljapoole jäävate sama populatsiooni esindajate kohta.

Järgnevates alapeatükkides kirjeldame esmalt valimi iseloomustamiseks sobivaid tüüpilisi kirjeldava statistika meetodeid ja mõõdikuid ning räägime ka sellest, kuidas eri tüüpi andmeid visualiseerida. Seejärel tegeleme levinumate järeldava

statistika meetoditega, mis võimaldavad uurida tunnustevahelisi seoseid ning teha valimi põhjal üldistusi populatsiooni kohta.

Näidisandmestikena kasutame lisaks juba eespool esitatud kõnetempo varieerumise andmestikule avaldatud ja vabalt kättesaadavat andmestikku idaseto eesja tagaeituse varieerumise kohta (Pilvik, Lindström & Plado 2021). Andmestiku jaoks on kasutatud osaliselt Eesti murrete korpuses ja osaliselt seto korpuses olevaid vestlusi, mis on lindistatud aastatel 2010–2013 idaseto kõnelejatega praeguse Venemaa aladele jäävates kunagistes seto külades. Andmestikus on märgendatud hulk tunnuseid, mis aitavad uurida, millest sõltub see, kas kõnelejad kasutavad n-ö eestipärast eeseitust (nt *ei olõ*) või hoopis seto keelele omast tagaeitust (nt *olõ-õi*). Uurimistulemused on avaldatud publikatsioonides (Lindström, Pilvik & Plado 2021; Pilvik, Plado & Lindström 2021). Võrdluseks vaatame põgusalt ka selle õpiku näidisuurimuses (Lindström & Pilvik) kasutatud 1. isiku asesõna väljendamise andmestikku (nt *ma tulen* või *tulen*).

Andmestike kasutamiseks tuleb need R-i sisse lugeda. Kõnetempo andmestik on allalaaditav DataDOI repositooriumist[1]. Kuna andmestik on salvestatud R-i andmevormingus (RDA-laiendiga), saab selle R-i lugeda otse veebilingi kaudu, kasutades järgmisi koodiridu[2].

```
> link_konetempo <- "https://datadoi.ee/bitstream/handle/33/592/fonkorp_
globaalne_konetempo.Rda?sequence=23&isAllowed=y" # andmestiku veebilink
> load(url(link_konetempo)) # laadime veebilingilt andmestiku (laaditakse
vaikimisi nimega "fonkorp2")
> fonkorp2$konetempo <- fonkorp2$silpide_arv/fonkorp2$kestus # loome uue tulba
"konetempo"
> str(fonkorp2) # vaatame, millised tunnused on andmestikus
'data.frame':  164 obs. of  25 variables:
 $ fail                    : Factor w/ 82 levels "fon_1","fon_2",..: 1 1 2 2
3 3 4 4 5 5 ...
 $ koneleja                : Factor w/ 150 levels "KJ3","KJ4","KJ1",..: 3 5 1
5 2 4 3 6 7 8 ...
 $ faili_kogukestus        : num  1535 1535 1572 1572 2415 ...
 $ vanus                   : int  26 22 36 22 30 27 26 29 23 23 ...
 $ sugu                    : Factor w/ 2 levels "M","N": 1 2 2 2 2 2 1 1 2 2
...
 $ silpide_arv             : int  1686 3146 4522 1337 8843 1694 2855 5979 2277
11278 ...
 $ sonade_arv              : int  912 1731 2538 793 5162 978 1610 3527 1262
7096 ...
 $ poolikute_arv           : int  8 10 18 7 47 12 16 26 21 107 ...
 $ venituste_arv           : int  82 53 122 50 318 76 119 165 94 340 ...
 $ korduste_arv            : int  51 26 134 7 140 36 56 42 8 101 ...
```

[1] https://datadoi.ee/handle/33/592

[2] R-i loetud andmestiku tulpadele saab viidata $ märkide abil. Siinses näites on meil andmestik nimega „fonkorp2“, kuhu tekitame uue tulba „konetempo“ olemasoleva tulba „silpide_arv“ väärtuste jagamisel olemasoleva tulba „kestus“ väärtustega. Trellidega # märgitakse R-is kommentaare jm teksti, mis ei ole ette nähtud koodina lugemiseks.

```
 $ üneemide_arv           : int  62 44 122 22 333 38 90 130 15 227 ...
 $ disf_arv               : int  141 89 274 64 505 124 191 233 123 548 ...
 $ kestus                 : num  539 849 1127 312 1960 ...
 $ art_kestus             : num  322 544 891 241 1522 ...
 $ põhitoon               : num  99.9 228.5 236 226.7 176.8 ...
 $ kaaskoneleja           : Factor w/ 150 levels "KJ3","KJ4","KJ1",..: 5 3 5
1 4 2 6 3 8 7 ...
 $ kaaskoneleja_sugu      : Factor w/ 2 levels "M","N": 2 1 2 2 2 2 1 1 2 2
...
 $ kaaskoneleja_vanus     : int  22 26 22 36 27 30 29 26 23 23 ...
 $ vanusevahe             : num  -4 4 -14 14 -3 3 3 -3 0 0 ...
 $ kaaskoneleja_kestus    : num  849 539 312 1127 358 ...
 $ kaaskoneleja_silpide_arv: num  3146 1686 1337 4522 1694 ...
 $ kaaskoneleja_sonade_arv : num  1731 912 793 2538 978 ...
 $ kaaskoneleja_disf_arv  : num  89 141 64 274 124 505 233 191 548 123 ...
 $ kaaskoneleja_põhitoon  : num  228.5 99.9 226.7 236 188.6 ...
 $ konetempo              : num  3.13 3.71 4.01 4.28 4.51 ...
```

Idaseto eituse andmestikku saab alla laadida OSF-i repositooriumist[3]. Tegemist on XLSX-vormingus tabeliga, mille R-i lugemiseks tuleb kasutada mõne lisapaketi, nt `readxl` (Wickham & Bryan 2023) abi. Paketid on erinevate funktsionaalsuste komplektid, mis täiendavad R-i põhifunktsioone. Muu hulgas on olemas paketid, mis võimaldavad R-i lugeda erinevates vormingutes andmestikke (nt XLSX-, SPSS-i, SAS-i, STATA, JSON-i või XML-i faile).

```
> install.packages("readxl") # installime XLSX-faili lugemiseks paketi
> library(readxl) # laadime kõik paketi readxl funktsioonid
> eitus <- read_excel("ISE_eitus_2010ndad.xlsx") # loeme andmestiku R-i
> str(eitus) # vaatame, millised tunnused on andmestikus
tibble [1,083 × 13] (S3: tbl_df/tbl/data.frame)
 $ Lause      : chr [1:1083] "meil olõ=õss mobiili ka võtta" "ll' kunagi aig
es'i=algu nigu õt saa=ass nigu" "aga nüüd nüüd nüüd nooril olõ ess ma ol'l'i"
"ma=tli=t (--) noorõq ti ti=i mõista m tandsigi mõista ei" ...
 $ EITUSSÕNA  : chr [1:1083] "es" "es" "es" "ei" ...
 $ VERBIVORM  : chr [1:1083] "cng" "cng" "cng" "cng" ...
 $ ASEND      : chr [1:1083] "tagaeitus" "tagaeitus" "tagaeitus" "eeseitus"
...
 $ TOPELTEITUS: chr [1:1083] "ei" "ei" "ei" "ei" ...
 $ LEMMA      : chr [1:1083] "olema" "saama" "olema" "mõistma" ...
 $ VERBITÜÜP  : chr [1:1083] "olemisverb" "modaalverb" "olemisverb"
"kognitsiooniverb" ...
 $ EELMINE    : chr [1:1083] "0" "taga" "ees" "taga" ...
 $ ISIK       : chr [1:1083] "3" "0" "3" "2" ...
 $ SUBJEKT    : chr [1:1083] "muu" "ei" "ei" "nom" ...
 $ KÕNELEJA   : chr [1:1083] "KJ4" "KJ4" "KJ4" "KJ4" ...
 $ SUGU       : chr [1:1083] "M" "M" "M" "M" ...
 $ AASTA      : num [1:1083] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010
...
```

[3] https://osf.io/gcft7/

Asesõna väljendamise andmestik L. Lindströmi ja M.-L. Pilviku siinse õpiku näidisuurimusest on tavalises CSV-vormingus (ingl *comma-separated values*), ehkki välju/tulpasid eristab tabelis koma asemel tabulaator ehk tabeldusklahv. CSV-failide R-i lugemiseks pole lisapakette tarvis installida.

```
> isik1 <- read.delim("sg1_andmestik_puhastatud.csv", sep = "\t", encoding
= "UTF-8", quote = "") # loeme andmestiku R-i, öeldes et tulpasid eraldab
tabulaator, kodeering on UTF-8 ning jutumärke loetakse sisse jutumärkidena
> str(isik1) # vaatame, millised tunnused on andmestikus
'data.frame':  16200 obs. of  16 variables:
 $ sone            : chr  "ol'in" "käisin" "lugesin" "`ütlen" ...
 $ lemma           : chr  "olema" "käima" "lugema" "ütlema" ...
 $ pron            : chr  "jah" "ei" "jah" "jah" ...
 $ lopp            : chr  "lõpuga" "lõpuga" "lõpuga" "lõpuga" ...
 $ eelmise_vorm    : chr  "verb1sg" "verb1sg" "pron1sg_muu" "pron1sg_muu" ...
 $ kaugus_eelmisest: int  184 9 3 283 152 5 65 2 17 52 ...
 $ eelmine_verb    : chr  "erinev" "erinev" "erinev" "erinev" ...
 $ eelmine_proniga : chr  "jah" "jah" "ei" "jah" ...
 $ aeg             : chr  "ipf" "ipf" "ipf" "pr" ...
 $ murre           : chr  "Alutaguse" "Alutaguse" "Alutaguse" "Alutaguse" ...
 $ khk             : chr  "Iis" "Iis" "Iis" "Iis" ...
 $ KJ_id           : chr  "KJ1" "KJ1" "KJ1" "KJ1" ...
 $ KJ_vanus        : int  75 75 75 75 75 75 75 75 75 75 ...
 $ KJ_synniaasta   : int  1887 1887 1887 1887 1887 1887 1887 1887 1887 1887
...
 $ KJ_sugu         : chr  "M" "M" "M" "M" ...
 $ fail            : chr  "Alutaguse_Iis_EMH0426.xml" "Alutaguse_Iis_EMH0426.
xml" "Alutaguse_Iis_EMH0426.xml" "Alutaguse_Iis_EMH0426.xml" ...
```

6.1. Kirjeldav ehk deskriptiivne statistika

Ees- ja tagaeituse andmestikus on peaaegu kõik tunnused kategoorilised. Seega saame arvandmeid põhiliselt tunnuste kategooriaid kokku lugedes, kokkuloetud sageduste esitamiseks sobivad hästi **sagedustabelid**. Arvulisi andmeid saab aga ka mõõtes. Sellisel juhul esitatakse andmeid alati mingites mõõtühikutes (nt millisekundites, hertsides, aastates). Korpuses esinevate sõnade puhul võime neid näiteks kokku lugeda, aga võime ka mõõta nende mingit omadust (nt pikkust tähemärkides, pikkust silpides, kestust millisekundites jne). Arvuliste andmete tüüpilisi väärtusi iseloomustavad kirjeldavas statistikas **aritmeetiline keskmine** ja **mediaan**, samuti kuuluvad kirjeldava statistika näitajate alla väärtuste ulatus ehk **haare** ja väärtuste hajuvuse mõõdikud, nagu **standardhälve**. Eriti asjakohased on need mõõtmisandmete puhul, kuid neid võib kasutada ka loendusandmete iseloomustamiseks.

6.1.1. Sagedustabelid

Kokkuloetavate andmete ehk loendusandmete (nt sõnade, bigrammide, grammatiliste vm kategooriate esinemissageduste) jaotumist saab kõige hõlpsamalt näidata **sagedustabelites**. Vaatame idaseto andmestikus, kui palju esineb andmetes ees- ja kui palju tagaeitust.

```
> table(eitus$ASEND)
 eeseitus tagaeitus
      269       814
```

See, kas sagedustabel on horisontaalne (nagu ülalolevas näites) või vertikaalne, ei ole iseenesest oluline. Vahel, eriti juhtudel, kus on palju kategooriaid (nt sõnaloendites), võib olla lihtsam jälgida vertikaalseid tabeleid, kus kategooriad on ühes tulbas ning nende sagedused andmestikus teises.

```
> data.frame(table(eitus$ASEND))
       Var1 Freq
1  eeseitus  269
2 tagaeitus  814
```

Sagedustabeli põhjal võib leida mingi tunnuse suurima sagedusega väärtuse ehk **moodi**. Näiteks on eitussõna paiknemist kodeeriva tunnuse *ASEND* mood *tagaeitus*, kuna see esineb eeseitusest sagedamini. Moodi võib leida ka järjestustunnuste ja arvuliste tunnuste puhul, kuna ka nende väärtuste esinemiskordi saab kokku lugeda. Näiteks on idaseto eituse andmestikus arvuline tunnus *AASTA*, mis väljendab kõnelejaga tehtud intervjuu salvestamise aastat. Selle mood on *2010*, mis tähendab, et aastast 2010 on andmestikus kõige rohkem vaatlusi.

```
> names(which.max(table(eitus$ASEND))) # kategoorilise tunnuse mood
[1] "tagaeitus"
> names(which.max(table(eitus$AASTA))) # arvulise tunnuse mood
[1] "2010"
```

Tabelites võime vaadelda ka mingite tunnuste või nähtuste koosesinemiste sagedusi, kasutades nn **risttabeleid** (ingl *contingency table*), mille ridades on ühe andmestiku tunnuse väärtused, tulpades teise tunnuse väärtused ning lahtrites vastavate väärtuste koosesinemise sagedused. Vaatame näiteks, kuidas ees- ja tagaeitus jagunevad vastavalt sellele, kas kasutatud on oleviku eitussõna *ei* (nt *ei olõ*) või mineviku eitussõna *es* (nt *es olõ* 'ei olnud').

```
> table(eitus$ASEND, eitus$EITUSSÕNA)
             ei  es
  eeseitus  201  68
  tagaeitus 413 401
```

Lisaks võib esitada tabelites ka nn **marginaal- ehk ääresagedusi**, milles on liidetud kokku kõik vastaval real või vastavas tulbas olevad sagedused.

```
> addmargins(table(eitus$ASEND, eitus$EITUSSÕNA))
              ei   es  Sum
  eeseitus   201   68  269
  tagaeitus  413  401  814
  Sum        614  469 1083
```

Sageli on risttabelites absoluutsagedustest informatiivsemad **suhtelised sagedused** (vt ka ptk 5.2.4 „Sõnavara analüüs: sagedusloendid"), seda eriti juhul, kui tabeli ridade või tulpade summad on väga erinevad. Suhteliste sageduste leidmiseks tuleb jagada iga tabeli lahtri väärtus läbi vastava rea või tulba marginaalsageduse või tabeli lahtrite kogusummaga, olenevalt sellest, milliste kategooriate suhtes lahtrite sagedusi tahetakse võrrelda. Suhtelised sagedused jäävad alati 0 ja 1 vahele. Kui need 100-ga läbi korrutada, saame protsendid. Kui korrutada need läbi mingi muu normaliseerimise baasiga (nt 1000, 10 000, 100 000), saame normaliseeritud sagedused (vt ka ptk 5.2.4.1 „Sagedusloendi põhjal keele uurimine").

```
> prop.table(table(eitus$ASEND, eitus$EITUSSÕNA)) # lahtrite summa 1
                    ei         es
  eeseitus  0.18559557 0.06278855
  tagaeitus 0.38134811 0.37026777

> prop.table(table(eitus$ASEND, eitus$EITUSSÕNA), margin = 1) # iga rea summa
1 ("Kuidas sõltub eitussõnade kasutus ees- ja tagaeitusest?")
                   ei        es
  eeseitus  0.7472119 0.2527881
  tagaeitus 0.5073710 0.4926290

> prop.table(table(eitus$ASEND, eitus$EITUSSÕNA), margin = 2) # iga tulba
summa 1 ("Kuidas sõltub ees- ja tagaeituse kasutus eitussõnast?")
                   ei        es
  eeseitus  0.3273616 0.1449893
  tagaeitus 0.6726384 0.8550107
```

6.1.2. Haare

Haare (ka variatsiooniulatus, ingl *range*[4]) iseloomustab arvandmete suurima ja vähima väärtuse erinevust. Suurem haare näitab, et andmete suurima ja vähima väärtuse vahe on suur, ent ei ütle midagi selle kohta, kuidas nende vahel jaotuvad kõik ülejäänud väärtused või milline on andmete kõige tüüpilisem väärtus.

```
> range(fonkorp2$konetempo) # leiame vähima ja suurima väärtuse
[1] 2.519344 6.758682
> max(fonkorp2$konetempo)-min(fonkorp2$konetempo) # leiame haarde
[1] 4.239338
```

EKSKFK kõnetempo andmestikus ulatub kõnelejate kõnetempo näiteks 2,52 silbist sekundis (kõige aeglasem kõneleja) 6,76 silbini sekundis (kõige kiirem kõneleja). Kõnetempo haare ehk variatsiooniulatus on 4,24 silpi.

6.1.3. Aritmeetiline keskmine ja standardhälve

Üks põhilisi arvandmete tüüpilisi väärtusi näitav mõõdik on **aritmeetiline keskmine**, mille leidmiseks jagatakse kõikide väärtuste summa läbi vaatluste arvuga. Vaatame näiteks kõnetempo andmestikust esimese 6 kõneleja kõnetempot.

```
> head(fonkorp2$konetempo)
[1] 3.129401 3.706497 4.011412 4.284050 4.511244 4.733828
```

Esimese 6 kõneleja keskmine kõnetempo on seega

$$\frac{3{,}129401 + 3{,}706497 + 4{,}011412 + 4{,}284050 + 4{,}511244 + 4{,}733828}{6} = 4{,}062739 \text{ silpi sekundis.}$$

Keskmine annab meile ühe konkreetse väärtuse, millega oma andmeid ülevaatlikult iseloomustada. See ei arvesta aga sellega, kui hajus on andmestik ehk kui palju iga üksik väärtus sellest keskväärtusest erineb. Näiteks keskmise kõnetempoga 4 silpi sekundis võivad olla väga erinevad kõnelejate rühmad (tabel 6.2).

[4] Pane tähele, et ptk-s 5.2.4.2 „Sõnavara hajuvus ja levik“ on sama ingliskeelne sõna teises tähenduses tõlgitud *levikuks*.

Tabel 6.2. Kolm võimalikku kuue kõneleja rühma, kelle keskmised kõnetempod on 4 silpi sekundis

Rühm	K1	K2	K3	K4	K5	K6	Keskmine
a	2	2	2	2	2	14	4
b	1,5	2,5	4	4	5,5	6,5	4
c	4	4	4	4	4	4	4

On selge, et need hulgad on olemuslikult väga erinevad. Esimeses rühmas (a) on mitu keskväärtusest väiksemat väärtust ja üks erandlikult suur väärtus. Siin ei ole aritmeetiline keskmine üksi kuigi informatiivne mõõdik, kuna ei iseloomusta hästi ühtki meie kuuest kõnelejast. Selline jaotus on enamasti omane sõnasagedustele korpustes: on suur hulk väga madala sagedusega sõnu ja üksikud väga kõrge sagedusega sõnad (vt ptk 5.2.4 „Sõnavara analüüs: sagedusloendid"). Teises rühmas (b) hajuvad üksikud väärtused keskväärtusest ühtlaselt mõlemale poole. Sellise jaotuse puhul aitab keskväärtuse teadmine pisut enam andmeid iseloomustada, kuna esindab tõepoolest mingit tüüpilist trendi. Viimases rühmas (c) on kõik väärtused keskväärtusega võrdsed. Sellisel juhul esindab keskväärtus ideaalselt kõiki rühma kõnelejaid.

Selleks, et **hajuvust** ehk üksikute väärtuste erinevust keskmisest arvesse võtta, esitatakse sageli koos aritmeetilise keskmisega ka **standardhälve** (ingl *standard deviation*). Standardhälve on *keskmine erinevus keskmisest* ning selle leidmiseks

1) lahutatakse igast üksikust väärtusest kogu hulga keskväärtus (ehk leitakse iga väärtuse hälve);
2) saadud tulemused võetakse ruutu, et saada lahti miinusmärkidest juhtudel, kus hulga keskväärtus on tegelikust väärtusest suurem;
3) saadud arvud liidetakse kokku;
4) saadud summa jagatakse läbi vaatluste arvuga, kui kasutada on kogu populatsiooni andmed (seda juhtub korpusandmete puhul harva), või vaatluste arvust ühe võrra väiksema arvuga, kui kasutada on ainult valimi andmed. Saadud väärtust nimetatakse **dispersiooniks** (ingl *variance*);
5) saadud tulemusest (ehk dispersioonist) võetakse ruutjuur. Standardhälve ongi ruutjuur dispersioonist.

Näiteks tabelis 6.2. toodud kõnelejate rühma *a* puhul {2,2,2,2,2,14}:

$$\sqrt{\frac{(2-4)^2+(2-4)^2+(2-4)^2+(2-4)^2+(2-4)^2+(14-4)^2}{6-1}} = \sqrt{\frac{4+4+4+4+4+100}{6-1}} = \sqrt{\frac{120}{5}} = 4{,}9.$$

Mida väiksem on standardhälve, seda sarnasemad on üksikud vaatlused keskmiselt keskväärtusele, ning mida suurem on standardhälve, seda rohkem väärtused keskmisest erinevad. Standardhälvet väljendatakse samades ühikutes, milles on ka mõõtmis- või loendusandmed (sellepärast kasutatakse seda sageli dispersiooni asemel). Näiteks esimese ülal toodud rühma (a) puhul erinevad rühma väärtused keskmisest ühe sekundi jooksul öeldud silpide arvust (4 silpi) keskmiselt 4,9 silbi võrra. Viimase rühma (c) puhul oleks standardhälve aga 0, sest kõikide kõnelejate kõnetempo on keskmisega võrdne.

Kuna sarnaselt aritmeetilise keskmisega on ka standardhälve lihtsalt mingi *keskmine*, on see tundlik andmestikus esinevate erandlikult suurte või väikeste väärtuste suhtes. Seetõttu võib selliste nn **erindite** (ingl *outlier* või *extreme value*) esinemisel saada suure standardhälbe isegi siis, kui suurem osa andmetest on tegelikult küllaltki homogeensed. Toodud näites on standardhälve isegi suurem kui aritmeetiline keskmine ise, mis tähendab, et kui eeldaksime, et andmed hajuvad ühtlaselt mõlemale poole keskmist, võiks meie andmestikus olla ka kõnelejaid, kes kõnelevad ühes sekundis negatiivse arvu silpe (4 – 4,9 = –0,9). Teame aga, et nii tegelikult olla ei saa, mistõttu võime järeldada, et andmetes on tõenäoliselt suuri erindeid (antud juhul hüpoteetiline kõneleja, kes suudab öelda sekundi jooksul tervelt 14 silpi).

Vaatame nüüd, milline on keskmine kõnetempo ja selle standardhälve terves EKSKFK kõnetempo valimis tegelikult.

```
> mean(fonkorp2$konetempo) # keskmine
[1] 4.738833
> sd(fonkorp2$konetempo) # standardhälve
[1] 0.7126207
```

Täiskasvanute keskmine kõnetempo on seega 4,74 silpi sekundis, kusjuures keskmiselt erineb üksikute kõnelejate kõnetempo sellest 0,71 silbi võrra.

6.1.4. Mediaan

Mediaan on teine tüüpilise väärtuse esindaja, mis erinevalt aritmeetilisest keskmisest on erandlikult suurte või väikeste väärtuste suhtes vähem tundlik. Definitsiooni järgi on mediaan kõige väiksemast kõige suurema väärtuseni järjestatud

arvurea keskmine väärtus, nii et mediaanist mõlemale poole jääb sama palju väärtusi. Näiteks ülal toodud kolme rühma (a, b ja c) mediaanid oleksid vastavalt 2, 4 ja 4. Kuna igas rühmas on paarisarv ehk 6 väärtust, kujuneb mediaan kahe keskmise väärtuse aritmeetilise keskmise kaudu (vastavalt (2 + 2) / 2 = 2, (4 + 4) / 2 = 4, (4 + 4) / 2 = 4). Ehkki mediaan ei ole mõjutatud andmetes esinevatest erandlikult suurtest ja erandlikult väikestest väärtustest, ei võta seegi arvesse andmete üldist jaotust. Näiteks hulkadel {5, 5, 5} ja {1, 5, 10} on sama mediaan.

Vaatame ka EKSKFK kõnetempo andmestiku kõnelejate kõnetempo mediaani.

```
> median(fonkorp2$konetempo)
[1] 4.740078
```

Näeme, et mediaan (4,74 silpi sekundis) on aritmeetilisele keskmisele (4,7388) väga sarnane, mis viitab, et tegemist võib olla sümmeetrilise jaotusega, kus väärtused hajuvad keskmisest ühtlaselt mõlemale poole. Kui mediaan on oluliselt väiksem kui aritmeetiline keskmine, on andmestikus tavaliselt üksikud teistega võrreldes erandlikult suured väärtused. Kui mediaan on oluliselt suurem kui aritmeetiline keskmine, on andmestikus jällegi üksikud teistega võrreldes erandlikult väikesed väärtused.

Mediaan sobibki aritmeetilisest keskmisest paremini kirjeldama jaotusi, mis on kas suurte või väikeste väärtuste suunas kaldu (näiteks palgad, korpuslingvistikas sõnasagedused), samuti arvudena kujutatud järjestusskaala (nt hinnangute) tüüpilisi väärtusi. Alati võib esitada aga korraga ka nii aritmeetilise keskmise koos standardhälbega kui ka mediaani. Jaotustest ja nende sümmeetrilisusest saab hea ülevaate erinevate graafikute abil, millest järgmiseks räägimegi.

6.1.5. Graafikud

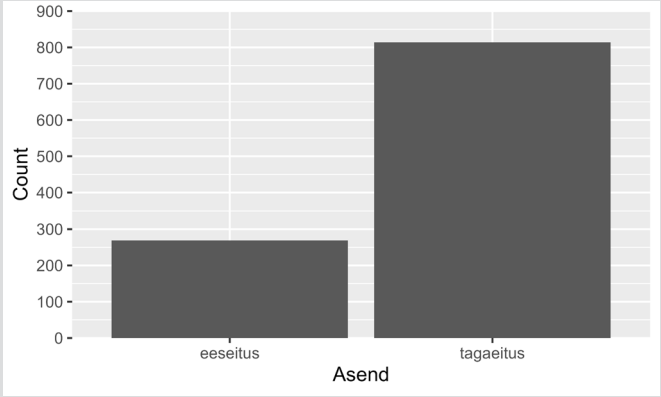
Lisaks sagedustabelitele ja nn paiknemiskarakteristikutele (mediaan ja keskmine) saab anda hea ülevaate andmetest ka jooniste ja graafikute abil. Graafikutüübi valimisel tuleb silmas pidada, mis tüüpi andmeid graafikul kuvatakse ning kas korraga kuvatakse üht või mitut tunnust. Ühe tunnuse kuvamisel illustreerime lihtsalt mingi tunnuse jaotust, mitme tunnuse kuvamise eesmärk on enamasti näidata seoseid tunnuste vahel või nende puudumist. Järgnevalt kirjeldame ülevaatlikult nelja tüüpilisemat graafikutüüpi ja nende kasutuskontekste. Kasutame graafikute tegemiseks R-i paketti `ggplot2` (Wickham 2016) ning graafikute tegemise hõlbustamiseks lisaks ka paketti `ggblanket` (Hodge 2024). Selleks, et järgmistes alapeatükkides kasutatud kood töötaks, tuleb esmalt need paketid installida ja laadida.

```
> install.packages("ggplot2")# installime paketi ggplot2
> install.packages("ggblanket")# installime paketi ggblanket
> library(ggplot2) # laadime paketi ggplot2 funktsioonid
> library(ggblanket) # laadime paketi ggblanket funktsioonid
```

6.1.5.1. Tulpdiagramm

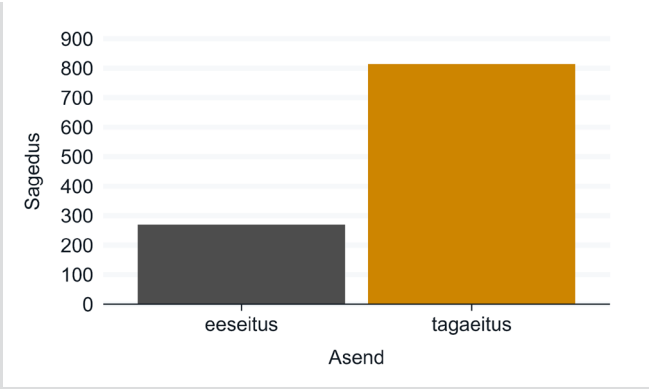
Tulpdiagramm (ingl *bar plot*) sobib **mittearvuliste** ehk kategooriliste tunnuste väärtuste esitamiseks ning näitab eri kategooriate **sagedusi**. Sisuliselt on tegemist sagedustabeli visualiseerimisega, kus tabeli kategooriad on ühel graafiku teljel tulpadeks ning nende kategooriate esinemissagedused määravad teisel teljel ära tulba kõrguse. Näiteks võime eespool toodud sagedustabelit idaseto ees- ja tagaeituse esinemise kohta kuvada tulpdiagrammina.

```
> gg_bar(data = eitus, x = ASEND) # kuvame tulpdiagrammi x-teljel andmestikust
"eitus" tunnust "ASEND" (ja selle väärtuste esinemissagedusi y-teljel)
```



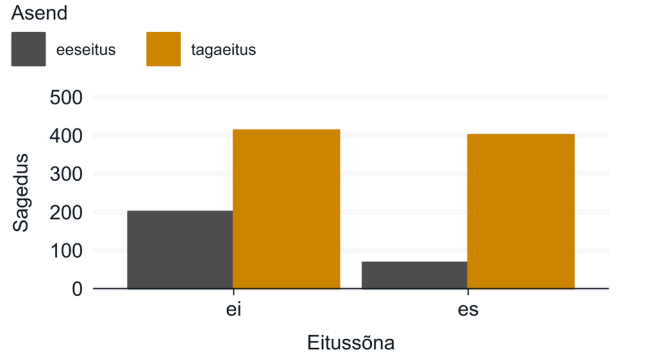
Vaikimisi tehtav graafik on üpris lihtne, ent võime seda täiendada veel erinevate elementidega, näiteks värvida tulbad eri värvi, muuta telgede pealkirju, graafiku üldist kujundust jpm.

```
> gg_bar(data = eitus, x = ASEND,
       fill = c("grey30", "orange3"), # tulbad halliks ja oranžiks
       y_label = "Sagedus", # y-telje pealkirjaks "Sagedus"
       mode = light_mode_t()) # graafiku üldine kujundus hele, legend üleval
```



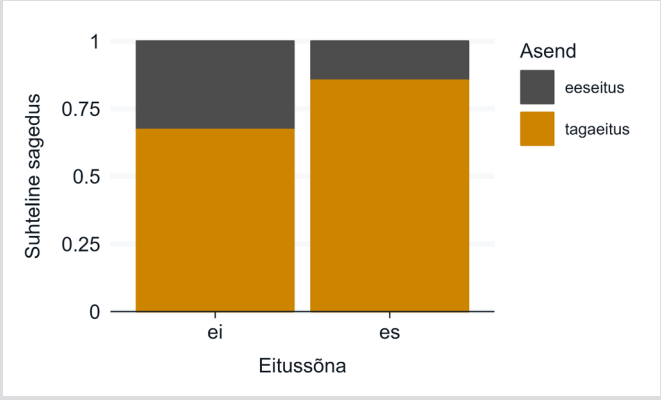
Tulpdiagrammiga võib korraga visualiseerida ka kaht kategoorilist tunnust ehk kuvada **kahe mittearvulise tunnuse vahelist seost**. Sellisel juhul on sisult tegemist risttabeli visualiseerimisega, kus ühe tunnuse kategooriaid kuvatakse tavaliselt x-teljel ning teise tunnuse kategooriaid erineva värvi või mustriga. Tulpasid võib sealjuures näidata n-ö virnastatult (ingl *stacked*) või üksteise kõrval (ingl *dodged* või *grouped*). Näiteks võime tulpdiagrammil vaadata, kuidas ees- ja tagaeitus jagunevad vastavalt sellele, kas kasutatud on oleviku eitussõna *ei* või mineviku eitussõna *es*.

```
> gg_bar(data = eitus, x = EITUSSÕNA, col = ASEND, # x-teljel tunnus
"EITUSSÕNA", tulpade värv vastavalt tunnuse "ASEND" kategooriatele
       col_palette = c("grey30", "orange3"), # värvid hall ja oranž
       mode = light_mode_t(),
       y_label = "Sagedus",
       position = "dodge") # absoluutsagedused, kategooriad kõrvuti
```



Samuti võib absoluutsageduste asemel kuvada suhtelisi sagedusi, mispuhul kõik tulbad on ühekõrgused ning võrrelda saab ühe tunnuse suhtelist jaotumist teise tunnuse kategooriates, olenemata sellest, kui palju kumbagi teise tunnuse kategooriat andmestikus esineb.

```
> gg_bar(data = eitus, x = EITUSSÕNA, col = ASEND,
       col_palette = c("grey30", "orange3"),
       mode = light_mode_r(), # graafiku üldine kujundus hele, legend paremal
       y_label = "Suhteline sagedus",  # y-telje pealkirjaks "Suhteline
sagedus"
       position = "fill") # suhtelised sagedused, kategooriad virnastatud
```

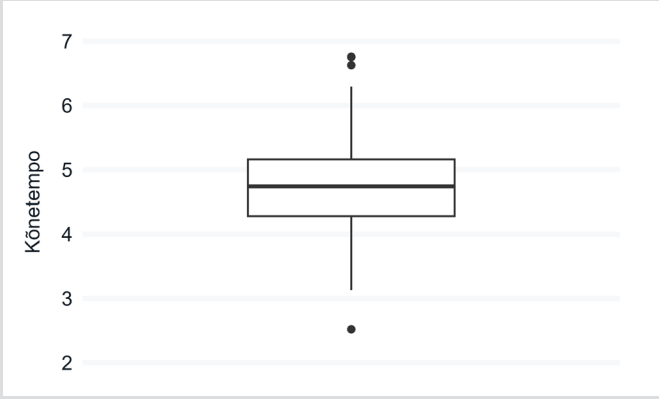


Näeme graafikutelt, et oleviku eitussõnaga *ei* kasutatakse eeseitust rohkem kui mineviku eitussõnaga *es*. Seda, kas nähtav erinevus on ka statistiliselt oluline, saame uurida statistilise testiga, mida kirjeldame alapeatükis 6.2.1.1.

6.1.5.2. Karpdiagramm

Karpdiagramm (ingl *boxplot*) sobib **arvuliste** andmete esitamiseks ning näitab arvuliste väärtuste **jaotumist**. Karpdiagramm koosneb 1) karbist endast, mille ulatus kuvab vahemikku, kuhu jääb 50% kõikidest arvulise tunnuse väärtustest, 2) karbi keskel olevast mediaanväärtust märkivast joonest ning 3) „vurrudest", mis kuvavad vahemikku, kuhu jäävad kas kõik tunnuse väärtused või valdav osa arvulise tunnuse väärtusi. Juhtudel, kus arvulistes andmetes on ka erandlikult suuri või väikeseid väärtusi ehk erindeid (ingl *outliers*), kuvatakse need väljaspool vurrude piire eraldi punktikestena. Kuvame kõnetempo jaotumist EKSKFK kõnetempo andmestikus karpdiagrammina.

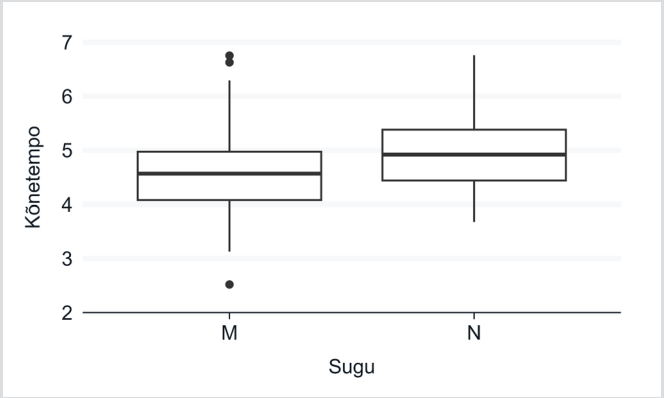
```
> gg_boxplot(data = fonkorp2, y = konetempo, # y-teljel tunnus "konetempo"
               y_label = "Kõnetempo",
               x_label = "", # x-teljel ei ole pealkirja
               mode = light_mode_t()) +
       scale_x_discrete() # vaikimisi arvuline x-telg kategooriliseks
x-teljeks
```



Näeme, et kõnetempo mediaani kuvav paks must joon jääb y-teljel umbes 4,7 silbi kanti ühe sekundi kohta. Poolte ehk 50% kõnelejate kõnetempo jääb kasti piiridesse ehk 4,3 ja 5,2 silbi vahele. Enamiku kõnelejate kõnetempo jääb kasti ümbritsevate vurrude ehk 3,1 ja 6,2 silbi vahele ning lisaks on andmestikus üks eriti aeglane kõneleja, kes jõuab sekundis öelda keskmiselt alla 3 silbi, ning üksikud eriti kiired kõnelejad, kes jõuavad sekundis öelda keskmiselt üle 6,5 silbi.

Karpdiagrammi võib kasutada ka selleks, et kuvada arvuliste andmete jaotumist erinevates rühmades ehk teisisõnu kuvada **ühe arvulise ja ühe kategoorilise tunnuse vahelist seost**. Näiteks võib meid huvitada, kas mehed ja naised räägivad sama kiiresti või on ühtede kõne teistest keskmiselt tempokam.

```
> gg_boxplot(data = fonkorp2, y = konetempo, x = sugu, # y-teljel tunnus
"konetempo", x-teljel "sugu"
               y_label = "Kõnetempo",
               mode = light_mode_t())
```

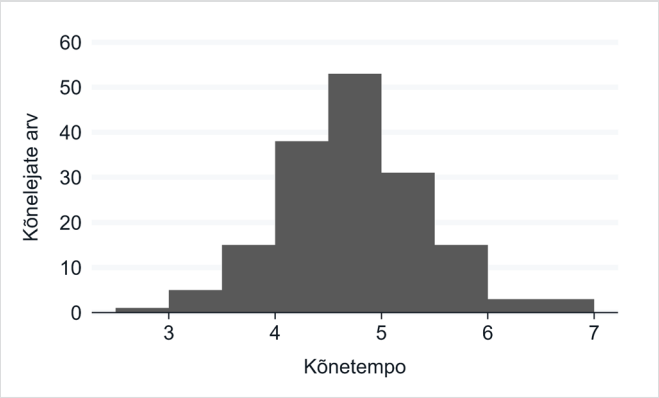


Näeme, et andmestikus on naiste kõnetempo keskmiselt tõepoolest meeste omast natuke kiirem: nii naiste kõnetempo mediaan (paks must joon), 50% vaatlustest kui ka vurrud on meeste kõnetempo omadest kõrgemal. Seda, kas nähtav erinevus on ka statistiliselt oluline, saame uurida statistilise testiga, mida kirjeldame alapeatükis 6.2.1.2.

6.1.5.3. Histogramm

Histogramm (ingl *histogram*) näitab samuti **arvuliste** andmete **jaotumist**, ent teeb seda intervallide kaudu: iga histogrammi tulp tähistab kindla suurusega intervalli ehk väärtuste vahemikku ning tulba kõrgus näitab sellesse intervalli kuuluvate arvuliste väärtuste hulka.

```
> gg_histogram(data = fonkorp2, x = konetempo, # x-teljel kõnetempo
(intervalli jäävate väärtuste esinemissagedus y-teljel)
               y_label = "Kõnelejate arv",
               x_label = "Kõnetempo",
               binwidth = 0.5, # ühe x-telje intervalli suurus 0,5 ühikut
               boundary = 0, # x-telje intervallide algus 0
               mode = light_mode_t())
```

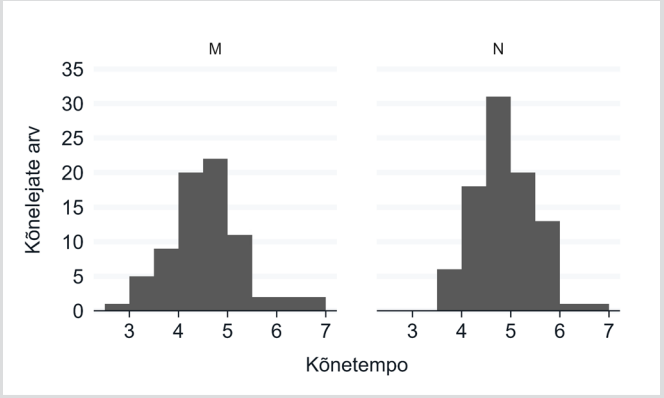


Erinevalt joonistatud karpdiagrammist on meid huvitav arvuline tunnus nüüd mitte y-, vaid hoopis x-teljel, ehkki seda saab hõlpsalt graafiku funktsioonis muuta, kui sümbol *x* vahetada *y* vastu ja vastupidi. Tegelikud kõnelejate kõnetempo väärtused on siin jagatud iga 0,5 silbi tagant kokku üheksasse vahemikku ehk intervalli. Oma kõnetempolt mingisse vahemikku jäävate kõnelejate arv on esitatud y-teljel. Näeme, et alla 3 silbi sekundis räägib andmestikus ainult üks kõneleja. Kiirusega 3–3,5 silpi sekundis räägib 5 kõnelejat, 3,5–4 silpi sekundis 15 kõnelejat jne. Näeme jooniselt ka seda, et kõnetempo varieerub üldse umbes 2,5 silbist sekundis kuni 7 silbini sekundis, kusjuures tegemist näib olevat sümmeetrilise jaotusega: kõige rohkem (üle 50) on aritmeetilise keskmise (4,7 silpi sekundis) ümber jäävaid kõnetempo väärtusi ning ülejäänud väärtused hajuvad enam-vähem ühtlaselt mõlemale poole keskmist. Sellist sümmeetrilist jaotust nimetatakse ka **normaaljaotuseks** (vt alapeatükki 6.2.1.2). Histogramm võib sageli välja näha nagu tulpdiagramm, ent kaht graafikutüüpi eristab eeskätt see, et tulpdiagrammil väljendab iga tulp mingi diskreetse kategooria sagedust, histogrammil aga mingisse vahemikku jäävate arvuliste väärtuste sagedust.

Ka histogrammil võib põhimõtteliselt kontrollida, kuidas **arvuline tunnus** on jaotunud **erinevates rühmades**, ehkki karpdiagrammil tulevad rühmadevahelised erinevused välja paremini. Näiteks kui teeme nii meeste kui ka naiste kohta eraldi histogrammid, näeme, et naiste kõnetempo on jaotunud x-teljel veidi suuremate väärtuste ümber kui meeste kõnetempo, ent sellest hoolimata on mõlemad jaotused üldjoontes sümmeetrilised.

```
> gg_histogram(data = fonkorp2, x = konetempo,
               facet = sugu, # tunnuse "sugu" kategooriad eraldi paneelidel
               y_label = "Kõnelejate arv",
               x_label = "Kõnetempo",
               binwidth = 0.5,
```

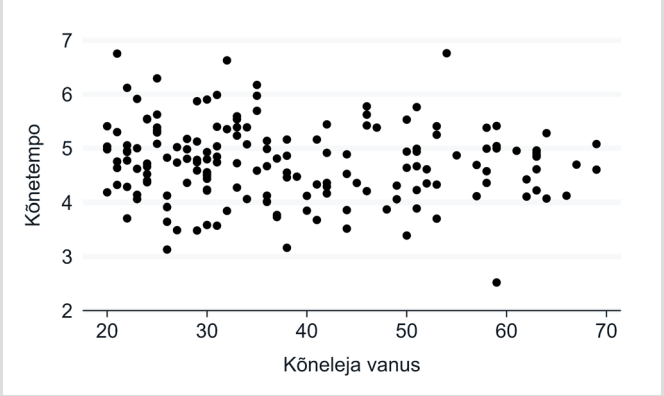
```
        boundary = 0,
        mode = light_mode_t())
```



6.1.5.4. Hajuvusdiagramm

Hajuvusdiagramm (ingl *scatter plot*) sobib kõige paremini **kahe arvulise tunnuse vahelise seose** kuvamiseks ning võimaldab visuaalselt hinnata, kas kahe arvulise tunnuse väärtused kasvavad või kahanevad sama- või erisuunaliselt või hoopis üksteisest sõltumatult. Näiteks võib meid kõnetempo puhul huvitada, kas vananedes kõnetempo aeglustub. Kuna nii vanus kui ka kõnetempo on arvulised tunnused, saab seost kuvada kõige paremini just hajuvusdiagrammil.

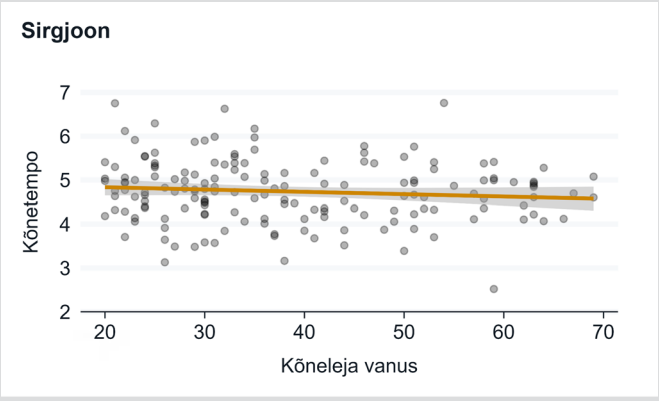
```
> gg_point(data = fonkorp2, x = vanus, y = konetempo, # x-teljel tunnus
"vanus", y-teljel tunnus "konetempo"
       x_label = "Kõneleja vanus",
       y_label = "Kõnetempo",
       mode = light_mode_t())
```



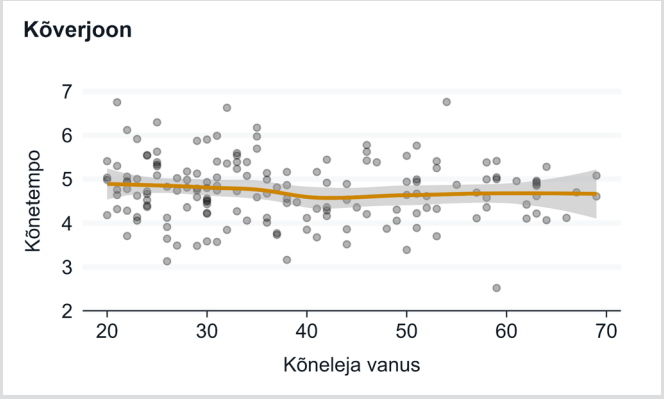
Iga punktike graafikul esindab üht kõnelejat, kelle vanust kujutab punktikese paiknemine x-telje suhtes ning kõnetempot paiknemine y-telje suhtes. Samade omadustega kõnelejate punktikesed (nt sama vanad kõnelejad, kes ütlevad keskmiselt sekundi jooksul täpselt sama palju silpe) jäävad graafikul üksteise alla, mistõttu võib vahel jääda ekslikult mulje, et andmestikus on tegelikust vähem vaatlusi. Näeme, et ehkki vanemate kõnelejate puhul varieerub kõnetempo vähem kui nooremate puhul (vaatlused ei paikne y-teljel üksteisest samavõrd laiali), ei saa joonise põhjal väita, et täiskasvanud kõnelejate kõnetempo vanusega oluliselt kiiremaks või aeglasemaks muutuks. Samuti märkame, et üle 50-aastaste kõnelejate hulgas on üks väga kiire ja üks väga aeglane kõneleja.

Hajuvusdiagrammile lisatakse sageli ka üldist tendentsi illustreeriv nn **trendijoon** (ingl *trend line*), mis ühendab omavahel ennustatud y-telje tunnuse väärtused iga x-telje tunnuse väärtuse kohta ning mis võib olla kas kõiki punkte võimalikult lähedalt läbiv sirgjoon või kõverjoon, mis järgib lähemalt kujutatava suhte tegelikku kuju. Viimast on kasulik vaadata selleks, et tuvastada, kas seos kahe tunnuse vahel on ikka kogu skaala ulatuses lineaarne ehk sirgjooneline ning kas seos on ühesuunaline ehk monotoonne (vt alapeatükki 6.2.1.3). Teeme punktid argumendi *alpha* abil ka veidi läbipaistvamaks, et näeksime paremini kattuvaid, täpselt samade omadustega kõnelejate vaatlusi (täiesti läbipaistva punkti *alpha* väärtus oleks 0, täiesti läbipaistmatu punkti *alpha* väärtus 1).

```
> gg_point(data = fonkorp2, x = vanus, y = konetempo,
       x_label = "Kõneleja vanus",
       y_label = "Kõnetempo",
       alpha = 0.3, # punktid läbipaistvamad
       title = "Sirgjoon", # joonise pealkiri "Sirgjoon"
       mode = light_mode_t()) +
    geom_smooth(method = "lm", col = "orange3") # punktipilve läbib (lineaarse
regressioonimudeli) sirge trendijoon
```



```
> gg_point(data = fonkorp2, x = vanus, y = konetempo,
       x_label = "Kõneleja vanus",
       y_label = "Kõnetempo",
       alpha = 0.3, # punktid läbipaistvamad
       title = "Kõverjoon", # joonise pealkiri "Kõverjoon"
       mode = light_mode_t()) +
    geom_smooth(method = "loess", col = "orange3") # punktipilve läbib
(lokaalse regressioonimudeli) kõver trendijoon
```



Kui kahe tunnuse vahel ei oleks mingit seost, oleks ennustatud keskmist kõnetempot kujutav lineaarne trendijoon horisontaalselt täiesti sirge, kõverjoon võib sellisel puhul ka siksakitada ebaregulaarselt üles-alla. Ülemiselt graafikult näeme, et sirge trendijoon on õige pisut paremale allapoole kaldu, mis viitab sellele, et kui vanus (x-teljel) kasvab, siis kõnetempo (y-teljel) õige pisut aeglustub (ühes sekundis öeldud keskmine silpide arv jääb veidi väiksemaks). Seda, kas sedavõrd väike kalle näitab statistiliselt olulist seost või mitte, saame jällegi kontrollida statistilise testiga, millest räägime alapeatükis 6.2.1.3.

Kirjeldav statistika ning erinevad graafikud aitavad niisiis saada andmetest esmast ülevaadet, aga ka luua hüpoteese selle kohta, milliseid seoseid andmetes võiks esineda. Kirjeldavate statistikute ning sobiva graafikutüübi valikul saame lähtuda eeskätt vaadeldavate tunnuste hulgast (üks tunnus, kaks tunnust või rohkem tunnuseid) ja tüübist (arvulised või kategoorilised tunnused).

6.2. Järeldav ehk inferentsiaalne statistika

Järeldava statistika eesmärk on üldistada meie valimi andmete põhjal midagi ka populatsiooni kohta, näiteks üldisemalt mingi keelenähtuse kasutuse, keeleliste valikute või keele kõnelejate kohta. Selleks saab lihtsamate seoste analüüsimisel

kasutada statistilisi teste, mis võimaldavad hinnata tõenäosust, et seos, mida oma andmetes täheldame (näiteks „naised räägivad kiiremini kui mehed"), on tegelikult meie valimi juhuslik eripära ega ole üldistatav tervele populatsioonile. Siin kirjeldatud statistiliste testide kasutamist nimetatakse vahel ka **ühetunnuseliseks analüüsiks**, kuna korraga hinnatakse ainult ühe tunnuse (nt sugu) mõju teise tunnuse (nt kõnetempo) väärtuste varieerumisele. Sealjuures võib, aga ei pruugi täpsustada, kumb tunnustest meid rohkem huvitab ehk teisisõnu: mis on uuritav tunnus ja mis on seletav tunnus. Vahetegemine pole oluline, kui seos toimib mõlematpidi: nt sõnapikkus saab sõltuda sõnasagedusest (sagedamad sõnad kalduvad kasutuses lühenema), aga põhimõtteliselt võib sõnasagedus sõltuda ka sõnapikkusest (lühemaid sõnu on mugavam hääldada; enamik sagedasti vaja minevaid funktsioonisõnu on lühikesed). Kui mõju saab olla aga ainult ühesuunaline, on uuritava ja seletava tunnuse eristamine asjakohane. Näiteks sugu (seletav tunnus) võib mõjutada seda, kui kiiresti keegi räägib (uuritav tunnus), ent kõnetempo ei saa mõjutada seda, mis soost inimene on. Lapse vanus (seletav tunnus) mõjutab seda, kui suur on lapse selleks hetkeks omandatud sõnavara (uuritav tunnus), aga omandatud sõnavara ei saa mõjutada seda, kui vana laps on.[5] Sealjuures ei tähenda see, et kahe tunnuse vahel on statistiline korrelatsioon, tingimata ka seda, et tegemist on põhjusliku seosega: naiseks olemine ei põhjusta kiiret kõnetempot, ehkki paljude kõnelejate andmete põhjalt üldistades ilmneb tendents, et naiskõnelejate kõnetempo on keskmiselt veidi kiirem kui meeskõnelejate oma. Kõnetempot võivad mõjutada erinevad muud tegurid, millega analüüsis ei ole pruugitud arvestada, nt kõne afektiivsus (emotsionaalne kõne on kiirem), teemade keerukus ja tuttavus, kõhklemine, vanus, vestluspartneri kõnetempo jpm ning neist mõne mõju võib avalduda ka soo kaudu.

Kui seosed on komplekssemad ning korraga mängib keeleliste valikute tegemisel rolli mitu erinevat tegurit, saab andmete analüüsimiseks kasutada **mitmetunnuselist analüüsi**. Mitmetunnuselise analüüsi meetodite hulka kuuluvad kõiksugu statistilised mudelid, mis võimaldavad uurida, kuidas meid huvitav keelenähtus (uuritav tunnus), nt sõnapikkus, kõnetempo, ühe või teise samatähendusliku keelendi või sõnade järjekorra valik, sõltub teistest lingvistilistest, kontekstuaalsetest või muudest teguritest (seletavad tunnused). Või teisisõnu: kas ja kui hästi aitab muude tegurite väärtuste teadmine ennustada meid huvitava keelenähtuse väärtusi.[6] Näiteks võime uurida, kuidas kõnetempo varieerub, kui võtame korraga

[5] Pane tähele, et uuritava ja seletava tunnuse eristust saab markeerida keeleliselt vähemalt kahe eri konstruktsiooniga: võime öelda sihitu verbiga, et Y **sõltub** X-ist, või sihilise verbiga, et X **mõjutab** Y-t. Mõlemal juhul on Y uuritav ja X seletav tunnus.

[6] Väljaspool järeldavat statistikat võib mitmetunnuseliseks analüüsiks nimetada ka selliste, eksploratiivsete statistiliste meetodite kasutamist, mis ei uuri üheainsa tunnuse sõltumist teistest tunnustest ega testi statistilisi hüpoteese, vaid püüavad paljude tunnuste koosesinemiste sageduste põhjal tuvastada andmetes peituvaid mustreid (nt klasteranalüüs, peakomponentanalüüs

arvesse nii kõneleja vanust, sugu, haridustaset kui ka kaaskõneleja kõnetempot, vanust, sugu ja haridustaset.

Korpusuuringutes on viimaste kümnendite jooksul pöördutud üha enam mitmetunnuselise analüüsi poole ning seda põhjusega. Korpusandmed esindavad (vähemalt ideaalis) loomulikku keelekasutust, ent seda väga erinevates keelekasutussituatsioonides, erinevatel ajahetkedel, erinevates kohtades, erinevatelt kõnelejatelt jne, mistõttu on vaatluste kasutuskonteksti oluliselt raskem kontrollida kui näiteks katsete puhul. Seetõttu võib enamasti eeldada, et korpusandmetest peegelduvaid kõneleja keelelisi valikuid suunab korraga ja koosmõjus mitu erinevat tegurit ning ühetunnuselises analüüsis võivad seetõttu mingid olulised seosed märkamata jätta. Siiski räägime siin esmalt just ühetunnuselise analüüsi meetoditest, kuna neid on ka algajal võrdlemisi lihtne kasutada, nende tulemused on kergemini tõlgendatavad ning nende abil on võimalik testida konkreetseid hüpoteese. Ühtlasi on siin käsitletud statistilisi teste võimalik hõlpsalt läbi viia paljudes erinevates programmides, sealhulgas ka näiteks Excelis. Ühetunnuselise analüüsi tulemused võivad olla heaks sisendiks mitmetunnuselise analüüsi tegemisele: kui oleme välja selgitanud, et andmetes esinev seos seletava tunnuse ja uuritava tunnuse vahel ei ole suure tõenäosusega juhuslik, saame edasi kontrollida seda, kas see seos jääb püsima, kui võtame samal ajal arvesse ka muid kontekstuaalseid tegureid. Näiteks kui meie andmetest selgub ühetunnuselise analüüsi käigus, et kõneleja kõnetempot mõjutavad nii kõneleja vanus kui ka sugu, siis mitmetunnuselise analüüsi käigus võime avastada, et soo efekt muutub ebaoluliseks. See võib juhtuda näiteks siis, kui korpuses on meeskõnelejad naiskõnelejatest oluliselt vanemad või nooremad, kuna sel juhul peegeldub ka soo tunnuse mõjus tegelikult hoopis kõneleja vanuse mõju.

Seoste analüüsimise testid ja mudelid, mida siinses õpikus kirjeldame, kuuluvad nn **nullhüpoteesi olulisuse testimise meetodi** alla (ingl *null hypothesis significance testing*). **Nullhüpotees** (*H0*) on vaikimisi eeldus, et tunnustevaheline seos populatsioonis puudub, näiteks „sõnasagedus ei mõjuta sõnapikkust", „sugu ei mõjuta kõnetempot", „eri käänete suhtelised kasutussagedused ei sõltu žanrist" jne, ning mis tahes seosed, mida oma valimis täheldame, on puhtalt selle valimi eripärast tulenevad juhuslikud kokkusattumused. **Alternatiivhüpotees** (*H1*) vastandub nullhüpoteesile ning väidab, et seosed, mida valimis täheldame, ei ole juhuslikud ning tunnuste vahel on olemas seos ka populatsioonis. Alternatiivhüpotees on sisuline hüpotees, mida enamasti tegelikult andmete põhjal testida soovime (nt „sõnasagedus mõjutab sõnapikkust", „sugu mõjutab kõnetempot", „eri käänete suhtelised kasutussagedused sõltuvad žanrist"). Selles peatükis kirjeldatud testidega saab aga olemasolevate andmete põhjal ümber lükata ainult nullhüpoteesi, alternatiivhüpoteesi kehtimist järeldatakse sellest, kui nullhüpotees ei kehti (sellest ka meetodi nimes *nullhüpoteesi* olulisuse testimine).

---

või korrespondentsanalüüs). Sellist analüüsi on rakendatud ka siinses õpikus A. Veismanni käitumisprofiilide analüüsi käsitlevas näidisuurimuses.

Nullhüpoteesi ümberlükkamiseks kasutavad statistilised testid mingi valimi põhjal arvutatud **teststatistikut** (nt erinevad korrelatsioonikordajad, hii-ruut-statistik või t-statistik, vt alapeatükki 6.2.1). Teststatistik on arvuline suurus, mis väljendab seda, kui hästi on valimi andmed testitava nullhüpoteesiga kooskõlas. Kui teststatistiku väärtus on nullile väga lähedal, siis pole enamiku testide puhul piisavalt alust nullhüpoteesi („seost ei ole / seos on juhuslik") hüljata. Kui aga teststatistik on nullist oluliselt suurem või väiksem, saab lugeda valimis esineva seose statistiliselt poluliseks. Võib siiski juhtuda, et leitud statistiku väärtus tuleneb puhtalt sellest, mil moel oleme koostanud oma valimi ja millised vaatlused valimisse on parasjagu sattunud, ning kui uurimust mõne muu populatsioonist juhuslikult võetud valimiga kordaksime, oleks statistiku väärtus hoopis teistsugune. Sellise stsenaariumi hindamiseks saab statistiku väärtuse, andmete hulga ja jaotumise põhjal arvutada ka nn **olulisuse tõenäosuse** ehk **p-väärtuse**, mis väljendab tõenäosust saada populatsioonist sellist valimit ja valimi andmete põhjal arvutatud samasugust (või nullist veelgi erinevamat) statistiku väärtust *juhul, kui kehtib nullhüpotees* ja populatsioonis tegelikult tunnuste vahel seost ei ole. Nagu kõik tõenäosused, võib p-väärtus jääda vahemikku nullist üheni. Mida lähemal on p-väärtus ühele, seda suurem on tõenäosus, et valimi andmetest nähtuv seos on saadud juhuslikult ja seost tunnuste vahel populatsioonis ei ole. Näiteks kui testi p-väärtus on 0,5, oleks 50-protsendiline tõenäosus, et samasuguseid või suuremaidki statistiku väärtusi võiksime saada ka nullhüpoteesi kehtides, ning kui väidaksime sellisel juhul, et populatsioonis on tunnuste vahel tõepoolest statistiliselt oluline seos, võiksime võrdselt niihästi eksida kui ka mitte eksida. Mida lähemal on p-väärtus aga nullile, seda väiksem on tõenäosus, et näeksime samasugust või veelgi tugevamat tunnustevahelist seost oma valimis puhtalt juhuse läbi ning seda väiksem on ka eksimisvõimalus, kui seose olemasolu kinnitame. Sisuliselt väljendab olulisuse tõenäosus, kui palju tõendeid on meil nullhüpoteesi vastu: mida rohkem, seda väiksem on p-väärtus. Tunnustevahelisi seoseid testivate statistiliste testide puhul loodame enamasti näha **võimalikult väikest p-väärtust**, ehkki peab meeles pidama, et p-väärtuse tõlgendamisel tuleb alati lähtuda konkreetse testi nullhüpoteesist.

Teatud eksimisvõimaluse võime endale nullhüpoteesi ümberlükkamisel siiski ka jätta. Seda lubatud eksimistõenäosust nimetatakse **olulisuse nivooks**, mida väljendatakse kreeka tähega α. Olulisuse nivoo määrab ära, kui suurt p-väärtust maksimaalselt lubame, selleks et lugeda seose olemasolu statistiliselt oluliseks. Olulisuse nivoo 0,5 ehk 50-protsendiline lubatud eksimistõenäosus ei oleks ilmselt kuigi hea mõte, kui soovime teha valimis nähtu põhjal mingeid üldistusi kogu populatsioonile. Tavaliselt määratakse humanitaar- ja sotsiaalteadustes olulisuse nivooks 0,05, ent sõltuvalt andmetest ja uurimuse riskitaluvusest võib selle määrata ka suuremaks või väiksemaks. Nullhüpoteesi saame hüljata niisiis sel juhul, kui testi p-väärtus on väiksem kui meie seatud olulisuse nivoo ehk lubatud eksimismäär nullhüpoteesi ümberlükkamisel (nt 0,05 ehk 5% või 0,01 ehk 1%). Väites

sealjuures, et tunnuste vahel on seos ja see pole vaid selle konkreetse valimi juhuslik eripära, oleme valmis selleks, et meil on võimalus teatud määral ka eksida, kuid see eksimise tõenäosus jääb alla 5%.

Siinkohal on oluline mainida, et p-väärtus ei saa meile kunagi öelda seda, kas nullhüpotees tegelikult vastab tõele või mitte, veelgi vähem seda, kas alternatiivne hüpotees vastab tõele või mitte. Saame p-väärtuse abil hinnata vaid tõenäosust, et konkreetse uurimuse andmetest leitud seos on juhuslik ja kehtib ainult meie valimis. Mida väiksem see tõenäosus on, seda kindlamad võime olla oma üldistuses. Sealjuures kasvab tõenäosus saada statistilises testis väiksem p-väärtus, kui kasutada on rohkem andmeid.

p-väärtus aitab väljendada niisiis seda, kui kindlad oma väites saame olla, ent määramatust saab väljendada ka nn **usaldusvahemike** kaudu. Valimi põhjal leitud statistik on üks konkreetne väärtus, usaldusvahemik aga intervall, mis kirjeldab, millised hinnatava parameetri väärtused on vaadeldud andmetega kooskõlas. Näiteks kui oleme arvutanud valimist kahe rühma (naiste ja meeste) kõnetempo keskväärtuste erinevuseks 0,36 (silpi sekundis), siis 95% usaldusvahemik 0,14–0,57 näitab, et ükski vahemikus 0,14–0,57 paiknevatest võimalikest keskväärtuste erinevustest ei ole meie andmetega 5% olulisuse nivool vastuolus; kui me kordaksime seejuures sama valimi võtmise ja usaldusvahemike arvutamise protseduuri 100 korda, siis ligikaudu 95 valimis 100st kataks saadud usaldusvahemik ka rühmade tegeliku keskmise erinevuse populatsioonis. Statistilistes testides on usalduvahemik ja p-väärtus omavahel seotud: kui usaldusvahemik katab nullhüpoteesi kehtivuse korral oodatud statistiku väärtuse (nt 0), on ka testi p-väärtus suur ning seost ei saa pidada statistiliselt oluliseks. Usaldusvahemikku võib leida aga ka muudele statistilistele näitajatele (nt mõju suuruse mõõdikutele), mille juurde tingimata p-väärtust ei arvutata (vt lähemalt Wallis 2021).

Siinses ülevaatlikus peatükis keskendume niisiis nullhüpoteesi olulisust testivatele statistilistele meetoditele. Seda eelkõige seetõttu, et nende kasutamine on laiemalt levinud ning algajale mõnevõrra hõlpsam ja nii olemasolevate programmide kui ka abimaterjalide tõttu ligipääsetavam. Need meetodid esindavad nn **sagedusstatistika** (ingl *frequentist statistics*) koolkonda, mis lähtub järelduste tegemisel ainult andmetest, eeldab, et andmete saamise protsess on korratav (võime võtta samal moel uue juhusliku valimi) ning hindab p-väärtuste kaudu seda, kui palju tõendeid on meie andmetes nullhüpoteesi vastu. Olgu siiski märgitud, et on olemas ka teine, nn **Bayesi statistika** koolkond, mis erineb nii filosoofiliselt kui ka metoodiliselt sagedusstatistikast põhiliselt seeläbi, et tegeleb nullhüpoteesi testimise asemel otse sisuliste hüpoteeside testimisega, võimaldab hüpoteeside testimisse kaasata lisaks andmetele ka uurija varasemaid teadmisi ja eeldusi ega kasuta statistilise määramatuse ja ebakindluse väljendamiseks p-väärtusi, vaid tõenäosusjaotusi. Bayesi statistika põhimõtete ja meetodite rakendamise kohta lingvistikas võib lähemalt lugeda näiteks artiklist (Nicenboim & Vasishth 2016).

6.2.1. Ühetunnuseline seoste analüüs: statistilised testid

Järgnevalt vaatame pisut lähemalt nelja lihtsamat statistilist testi, mida on nii keeleteaduses kui ka mujal laialdaselt kasutatud: hii-ruut-test, t-test, U-test ning Pearsoni või Spearmani korrelatsioonikordaja statistilise olulisuse test. Teste eristab esiteks see, **mis tüüpi tunnuste vaheliste seoste** olulisust nendega saab testida (nt kategooriline-kategooriline, kategooriline-arvuline või arvuline-arvuline, vt tabelit 6.3), ning teiseks see, **millised eeldused** tunnuste jaotuste, kasutada olevate andmete hulga, kirjeldatava seose suuna vm suhtes peavad nende kasutamiseks täidetud olema.

Tabel 6.3. Graafikutüübi ja statistilise testi valimine vastavalt tunnuste tüübile

1. tunnus	2. tunnus	Näide	Graafikutüüp	Test
kategooriline	kategooriline	kas idaseto eitussõna asukoht sõltub eitussõna ajavormist?	tulpdiagramm	hii-ruut-test
arvuline	kategooriline	kas täiskasvanute kõnetempo sõltub soost?	karpdiagramm, histogramm	t-test, U-test
arvuline	arvuline	kas täiskasvanute kõnetempo sõltub vanusest?	hajuvusdiagramm	korrelatsioonitest

6.2.1.1. Kahe kategoorilise tunnuse vahelised seosed: hii-ruut-test

Hii-ruut-test, täpsemalt hii-ruudu sõltumatuse test (ingl *chi-squared test for independence*) on üks korpuslingvistikas enim kasutatud statistilisi teste, mida kasutatakse **kahe kategoorilise tunnuse vahelise seose hindamiseks**. Sealjuures võib kummalgi tunnusel olla kaks või enam kategooriat. Näiteks kui deskriptiivse statistika alapeatükis 6.1.1 vaatasime risttabelite abil, kuidas idaseto eitussõna asukoht (*ei ole* või *olõ-õi*) ja eitussõna oleviku või mineviku vormi (*ei* või *es*) kasutussagedused jaotuvad, siis hii-ruut-testiga saame testida, kas nähtud erinevus kasutusprotsentides (eeseitust *ei*-ga 33% ja *es*-iga 14%, tagaeitust *ei*-ga 67% ja *es*-iga 86%) on ka statistiliselt oluline või võib see olla lihtsalt valimi juhuslik eripära.

Kui statistiliselt oluline seos on olemas, tähendab see seda, et ühe tunnuse kategooria teadmine aitab ennustada teise tunnuse kategooriat ja vastupidi. Näiteks kui teame, et tegemist on oleviku eitava vormiga, ootame, et seal esineks eeseitust rohkem kui siis, kui tegemist on mineviku eitava vormiga. Või kui teame, et tegemist on toimetamata (*vs*. toimetatud) tekstidega, ootame, et seal esineks rohkem kirjakeele normist hälbivaid vorme (nt *kellegil vs. kellelgi*, vt Kängsepp 2024); kui

teame, et nimisõna puhul on tegemist pärisnimega (*vs*. üldnimega), ootame, et selle sisseütlev kääne oleks tõenäolisemalt moodustatud pika sisseütleva vormiga (nt *Tartusse vs. Tartu*, vt Siiman 2016). Test põhineb kahe kategoorilise tunnuse kategooriate koosesinemissagedusi sisaldava risttabeli tegelike ehk vaadeldud sageduste ja teoreetiliste ehk oodatud sageduste võrdlemisel. **Tegelikud sagedused** (ingl *observed frequencies*) on väärtused, mis risttabeli lahtrites päriselt esinevad, st tunnuste kategooriate kombinatsioonide esinemissagedused. **Oodatud sagedused** (ingl *expected frequencies*) on aga teoreetilised sagedused, mida saaks lahtritesse tuletada ainult ridade ja tulpade marginaalsageduste ehk rea- ja tulbasummade põhjal (vt ka ptk 5.2.5 „Kollokatsioonid" ja ptk 6.1.1 „Sagedustabelid"): iga lahtri oodatud sageduse saab vastava rea ja vastava tulba marginaalsageduste korrutise läbijagamisel kogu tabeli summa ehk vaatluste arvuga. Need väljendavad kategooriate koosesinemissageduste teoreetilist jaotumist nullhüpoteesi korral ehk juhul, kui kahe vaadeldava tunnuse erinevad väärtused esineksid koos täiesti juhuslikult ning nende kasutuses ei esineks mingeid mustreid ega seoseid.

Tabel 6.4. Eeseituse ja tagaeituse ning eitussõnade *ei* ja *es* koosesinemise tegelikud ja oodatud sagedused

	***ei* tegelik**	*ei* oodatud	***es* tegelik**	*es* oodatud	**Kokku**
eeseitus	201	(614 × 269) / 1083 = 152,5078	68	(469 × 269) / 1083 = 116,4922	269
tagaeitus	413	(614 × 814) / 1083 = 461,4922	401	(469 × 814) / 1083 = 352,5078	814
Kokku	614		469		1083

Tabelisse 6.4 on märgitud alapeatükis 6.1.1 koostatud risttabeli põhjal eitussõna asukoha ja eitussõna ajavormi koosesinemise tegelikud sagedused ning nende põhjal on leitud vastavate koosesinemiste oodatud sagedused. Näiteks eitussõna *ei* oodatud sageduse eeseituse kontekstis (*ei olõ*) saame, kui korrutame *ei* esinemise üldsageduse 614 eeseituse üldsagedusega 269 ning jagame saadud korrutise 165 166 läbi kõikide vaatluste arvuga 1083. Oodatud sagedus 152,5078 on väiksem kui tegelik sagedus 201, mis tähendab, et eitussõnaga *ei* kasutatakse eeseitust oodatust sagedamini.

Kõiki risttabeli lahtrite tegelikke sagedusi võrreldaksegi nende oodatud sagedustega ning arvutatakse erinevuste põhjal hii-ruut-statistik: mida erinevamad on tegelikud sagedused nullhüpoteesi kehtimise korral oodatud sagedustest, seda suurem nullist on ka **hii-ruut-statistik**, ent statistiku absoluutväärtus sõltub palju ka sellest, mis on kõikide lahtrite sageduste summa (ehk mitu vaatlust meil kokku

kasutada on). Seetõttu ei maksa statistiku absoluutväärtust eri suuruses andmestike puhul võrrelda. Hii-ruut-statistiku (lihtsustatud) valem on järgmine:

$$X^2 = \sum \frac{(tegelik - oodatud)^2}{oodatud}$$

Eitussõna asukoha ja ajavormi tabeli põhjal on hii-ruut-statistiku väärtus niisiis

$$X^2 = \frac{(201 - 152{,}5078)^2}{152{,}5078} + \frac{(413 - 461{,}4922)^2}{461{,}4922} + \frac{(68 - 116{,}4922)^2}{116{,}4922} + \frac{(401 - 352{,}5078)^2}{352{,}5078} = 47{,}37085$$

Selleks, et saada teada, kas leitud statistik on nullist *oluliselt* erinev, võrreldakse seda hii-ruut-jaotuse nn **kriitiliste väärtustega** kindla **vabadusastmete arvu** (ingl *degrees of freedom*) ja olulisuse nivoo juures (vt nt Stefanowitsch 2020: 447). Kui statistik on suurem kui kriitiline väärtus, siis saab seose lugeda statistiliselt oluliseks. Hii-ruut-testis tuleneb vabadusastmete arv võrreldavate kategooriate hulgast ning see leitakse nõnda, et lahutatakse ridade ja tulpade arvust 1 ning korrutatakse omavahel saadud väärtused. 2×2 risttabeli puhul oleks vabadusastmete arv niisiis 1: (2 – 1) × (2 – 1) = 1. Vabadusastmed iseloomustavad siin nende risttabeli lahtrite arvu, mille väärtusi saab vabalt varieerida, ilma et ridade ja tulpade marginaalsagedused või üldine sageduste summa muutuks. 2×2 risttabelis on selliseid lahtreid ainult 1, sest kohe, kui oleme ühte lahtrisse mingi uue, vabalt valitud sageduse kirjutanud, on teiste lahtrite sagedused marginaalsageduste põhjal juba ette kindlaks määratud. Kui vaatleme eitussõna asukoha ja eitussõna ajavormi risttabelit, mille marginaalsagedused on meil teada, saame muuta suvaliselt vaid ühe lahtri väärtust (näiteks kirjutame olemasoleva sageduse 201 asemele sageduse 152), misjärel peame muutma automaatselt väärtusi ka kõikides teistes lahtrites ega saa neid enam vabalt varieerida (tabel 6.5).

Tabel 6.5. 1 vabadusaste ehk vabalt muudetav sagedus 2x2 risttabelite puhul

	ei	***es***	**Kokku**
eeseitus	**201 → 152**	68 → 117	269
tagaeitus	413 → 462	401 → 352	814
Kokku	614	469	1083

2×3 risttabelite puhul oleks vabadusastmete arv juba (2 – 1) × (3 – 1) = 2, kuna vabalt muudetavaid lahtreid on ühe võrra enam. Mida suurem on vabadusastmete arv, seda suurem peab olema ka hii-ruut-statistiku väärtus selleks, et kahe tunnuse vaheline seos oleks sama olulisuse nivoo korral (nt 0,05) statistiliselt oluline. Ühe

vabadusastmega 2×2 risttabeli puhul ja olulisuse nivool 0,05 on hii-ruut-jaotuse kriitiline väärtus 3,84, selleks et väita kahe tunnuse vahel statistiliselt olulise seose olemasolu. Meie hii-ruut-statistiku väärtus 47,37 on sellest kõvasti suurem, seega saame nullhüpoteesi („seos on juhuslik“) hüljata.

Statistikaprogrammid väljastavad hii-ruut-testiga ka teststatistiku põhjal arvutatud p-väärtuse, mis erinevalt kriitiliste väärtuste tabelist väljendab konkreetset tõenäosust, et saaksime nullist samavõrd või enamgi erineva statistiku väärtuse, juhul kui nullhüpotees kehtib: kui p-väärtus on väiksem kui seatud olulisuse nivoo, võib olulisusenivool määratud eksimisvõimalusega väita, et valimis leitud seos kahe tunnuse vahel on tõenäoliselt olemas ka populatsioonis. Sealjuures esitab R väga väikeseid p-väärtusi nende **standardkujul**. Näiteks alloleva testi p-väärtus on kujul 5,875e-12, mis tähendab, et väärtus 5,875 tuleb läbi korrutada väärtusega 10^{-12}. Teisiti öeldes peaks liigutama kümnendmurru koma 12 võrra vasakule, saades p-väärtuseks 0,000000000005875. Tehtud testi p-väärtus on palju väiksem kui seatud olulisuse nivoo 0,05, mistõttu võime nullhüpoteesi hüljata.

```
> chisq.test(table(eitus$ASEND, eitus$EITUSSÕNA), correct = FALSE) # teeme
hii-ruut-testi, et testida eitussõna asendi ja ajavormi seose statistilist
olulisust

        Pearson's Chi-squared test

data:  eitus$ASEND and eitus$EITUSSÕNA
X-squared = 47.371, df = 1, p-value = 5.875e-12
```

Hii-ruut-test ütleb lihtsalt, kui suur on hii-ruut-statistiku väärtus, ning kui suur on tõenäosus, et leiaksime valimi põhjal nii suure või isegi suurema hii-ruut-statistiku juhul, kui tegelikult kehtib nullhüpotees. Lihtsustatult: kas kahe tunnuse vahel on statistiliselt oluline seos, mida saab teatud kindlusega üldistada ka populatsioonile? Test ei ütle aga midagi võrreldavat selle kohta, **kui tugev see seos on** või missugune see seos on (näiteks millise eitussõnaga on eeseitus tavalisem). Statistiliselt olulise seose kahe tunnuse vahel võib leida ka siis, kui seos on väga nõrk, aga andmeid on palju. Samamoodi võib võrdlemisi tugev seos olla statistiliselt ebaoluline näiteks sellepärast, et andmeid on liiga vähe. Seetõttu võib hii-ruut-testi tulemusi täiendada mõne kategoorilistele tunnustele sobiva seosekordajaga, näiteks **Craméri *V***[7] seosekordajaga (Cramér 1946; Ben-Shachar jt 2023), mis arvestab ka vaatluste arvu ja tabeli suurusega, jääb alati 0 ja 1 vahele ning on tänu sellele erinevalt hii-ruut-statistikust võrreldav ka erinevate andmestike puhul. Craméri *V*-d võib seega tõlgendada kui ühe kategoorilise tunnuse varieerumise **mõju suurust** teise kategoorilise tunnuse varieerumisele[8]. Seosekordaja leidmiseks kasutatakse hii-ruut-statistikut

[7] Vahel räägitakse 2×2 tabelite puhul *V* asemel ka ϕ-kordajast, mida arvutatakse samamoodi.

[8] Teiste seose tugevust indikeerivate näitajate kohta vaata näiteks (Brezina 2018: 115–116).

(χ^2), vaatluste arvu (N) ning risttabeli lühema külje pikkust (k), millest on lahutatud 1 (k-1). Seetõttu on seda võimalik välja arvutada ka käsitsi, kasutades järgmist valemit:

$$V = \sqrt{\frac{X^2}{N(k-1)}}$$

```
> test <- chisq.test(table(eitus$ASEND, eitus$EITUSSÕNA), correct = FALSE) #
salvestame hii-ruut-testi tulemuse objekti "test"
> x2 <- unname(test$statistic) # võtame välja hii-ruut-statistiku väärtuse
> k <- min(dim(test$observed)) # leiame risttabeli lühema külje pikkuse
> N <- sum(test$observed) # leiame risttabeli vaatluste arvu
> sqrt(x2/(N*(k-1))) # arvutame Craméri V
[1] 0.2091419
```

Craméri V-d võib aga arvutada ka lisapakettide abil. Näiteks on paketis `effectsize` (Ben-Shachar, Lüdecke & Makowski 2020) vastav funktsioon *cramers_v()*.

```
> install.packages("effectsize") # installime paketi effectsize
> library(effectsize) # laadime paketi effectsize funktsioonid
> cramers_v(table(eitus$ASEND, eitus$EITUSSÕNA)) # leiame Craméri V
Cramer's V (adj.) |       95% CI
--------------------------------
0.21              | [0.16, 1.00]
```

Craméri V tõlgendamisel võib lähtuda Coheni (1988) pakutud skaalast (tabel 6.6), millel V väärtuse tõlgendus on seotud testi vabadusastmete arvuga: mida suurem on vabadusastmete arv, seda väiksemad on keskmise ja suure mõju lävendid.

Tabel 6.6. Craméri V tõlgendamise skaala

Vabadusastmete arv	Väike mõju (alates)	Keskmine mõju (alates)	Tugev mõju (alates)
1	0,10	0,30	0,50
2	0,07	0,21	0,35
3	0,06	0,17	0,29
4	0,05	0,15	0,25
5	0,05	0,13	0,22

Testi tulemuste raporteerimisel esitatakse tavaliselt nii hii-ruut-statistiku väärtus, vabadusastmete arv kui ka testi p-väärtus. Hea oleks lisada ka mingi mõju suuruse kordaja. Eriti väikese p-väärtuse puhul esitatakse tavaliselt konkreetse väärtuse asemel mingi lävend, millest allapoole p-väärtus jääb. Näiteks „Seos eitussõna asukoha ja eitussõna ajavormi vahel on statistiliselt oluline ($\chi^2 = 47{,}371$, $df = 1$, $p < 0{,}001$), ent nõrk ($V = 0{,}21$)". Samuti võib testi tulemuse ja seose tugevuse märkida ära kohe risttabeli pealkirja või allkirja.

Kui teame nüüd, et nullhüpotees ei kehti ja kahe tunnuse vahel on statistiliselt oluline seos ning et selle seose tugevus on pigem nõrk, siis selleks, et näha lähemalt, **missugune kahe tunnuse vaheline seos on**, võib vaadelda testi nn **standardiseeritud jääke** (Agresti 2013: 80–81), mis jagavad tegelike ja oodatud sageduste vahed läbi vastavate oodatud sageduste rea ja tulba summade suhtes kohandatud standardvigadega. Standardiseeritud jäägid (ingl *standardized residuals*, täpsemalt *adjusted standardized residuals*) näitavad, millised kategooriate kombinatsioonid esinevad oodatust oluliselt harvemini või sagedamini koos. Kui olulisuse nivoo on seatud 0,05 peale, näitavad jäägid, mille absoluut- ehk märgist sõltumatu väärtus on suurem kui 1,96 (~2), et vastava lahtri tegelik sagedus ei sobi hästi kokku nullhüpoteesiga.

```
> test <- chisq.test(table(eitus$ASEND, eitus$EITUSSÕNA), correct = FALSE) #
salvestame hii-ruut-testi tulemuse objekti "test"
> test$stdres # võtame testi tulemusest välja standardiseeritud jäägid
                  ei        es
  eeseitus   6.882642 -6.882642
  tagaeitus -6.882642  6.882642
```

Meie näite puhul on kõik standardiseeritud jääkide absoluutväärtused kahest suuremad. Seega saame öelda, et eeseitus esineb oodatust oluliselt sagedamini oleviku eitussõnaga *ei* (*ei olõ*) ning oodatust oluliselt harvemini mineviku eitussõnaga *es*; peegelpildina esineb tagaeitus oleviku eitussõnaga oodatust oluliselt harvem ja mineviku eitussõnaga oodatust oluliselt sagedamini (*olõ-õs*)[9]. Sageli, eriti suuremate kui 2×2 tabelite puhul, võib aga juhtuda, et oodatust erinevad sagedused on näiteks ainult ühes või kahes kategoorias, mis mõjutavad nõnda ka enim hii-ruut-testi statistiliselt olulist tulemust.

Hii-ruut-testi kasutamisel on mõned **eeldused**:

- Klassikalises statistikas viidatakse sageli nn **Cochrani reeglile** (Cochran 1952, 1954, vt ka Kroonenberg & Verbeek 2018), mille põhjal peaks

[9] Olgu märgitud, et standardiseeritud jääkide põhjal ei saa väita, et eeseitus esineb pigem oleviku eitussõnaga *ei* ja tagaeitus pigem mineviku eitussõnaga *es*. Tagaeitus on tavalisem mõlema eitussõna vormiga, ent oleviku eitussõnaga on selle osakaal pisut väiksem ning see erinevus mineviku eitussõnaga võrreldes on sealjuures statistiliselt oluline.

hii-ruut-testi tegemisel suuremate kui 2×2 tabelitega vähemalt 20% oodatud sagedustest olema suuremad kui 5 ja kõik oodatud sagedused vähemalt 1. Vahel esitatakse nendele tingimustele siiski ka mööndusi. Näiteks Stefanowitsch (2020: 177) ei maini oodatud sagedusi üldse, vaid ütleb, et kuni 25% tegelikest sagedustest võivad olla 5st väiksemad, ent ükski tegelik sagedus ei tohiks siiski olla 0. Ka Wallis (2021: 117) peab Cochrani reeglit liiga konservatiivseks ning lubab nii nulle tegelike sageduste hulka kui ka 5st väiksemaid väärtusi oodatud sageduste hulka. Võib kaaluda ka teiste sarnaste statistiliste testide kasutamist, mis ei ole väikeste sageduste suhtes tundlikud. Sellised testid on näiteks **Fisheri täpne test** (ingl *Fisher's exact test*) või **G-test**, ehkki viimast peetakse väikeste sageduste puhul hii-ruut-testist oluliselt ebatäpsemaks (Agresti 2013: 77). Väikeste oodatud sagedustega 2×2 risttabelites aitab hii-ruut-testi veidi konservatiivsemaks teha ka nn Yatesi pidevusparanduse kasutamine, mis vähendab tegelike ja oodatud sageduste absoluuterinevust igas risttabeli lahtris 0,5 võrra.

```
> chisq.test(table(eitus$ASEND, eitus$EITUSSÕNA), correct = TRUE) #
hii-ruut-test Yatesi pidevusparandusega

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(eitus$ASEND, eitus$EITUSSÕNA)
X-squared = 46.399, df = 1, p-value = 9.647e-12

> fisher.test(table(eitus$ASEND, eitus$EITUSSÕNA)) # Fisheri täpne test

        Fisher's Exact Test for Count Data

data:  table(eitus$ASEND, eitus$EITUSSÕNA)
p-value = 2.482e-12
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.092037 3.963710
sample estimates:
odds ratio
         2.867294
```

- Hii-ruut-testi ei ole soovitatav kasutada siis, kui uuritavatel tunnustel on palju erinevaid väärtusi/kategooriaid, kuna ühelt poolt kasvab selle tagajärjel samuti oht, et andmed hajuvad kategooriate vahel ning tegelike sageduste hulka satub palju väikesi väärtusi või suisa nulle, ning teiselt poolt kasvab suurte tabelite puhul märkimisväärselt tõenäosus leida kuskilt lahtritest tegelike ja oodatud sageduste vahel mõni statistiliselt oluline erinevus, ehkki ühe tunnuse mõju teisele võib tervikuna olla väike. Suurte risttabelite puhul võib võimalusel kaaluda teatud kategooriate väljajätmist või ühendamist (nt madalamate sagedustega või mingitel alustel

sarnasemad kategooriad ühendada kategooriaks „muu"), ehkki selle käigus kaotame ära osa olemasolevast infost ning raskendame mõnevõrra ka tulemuste tõlgendamist.

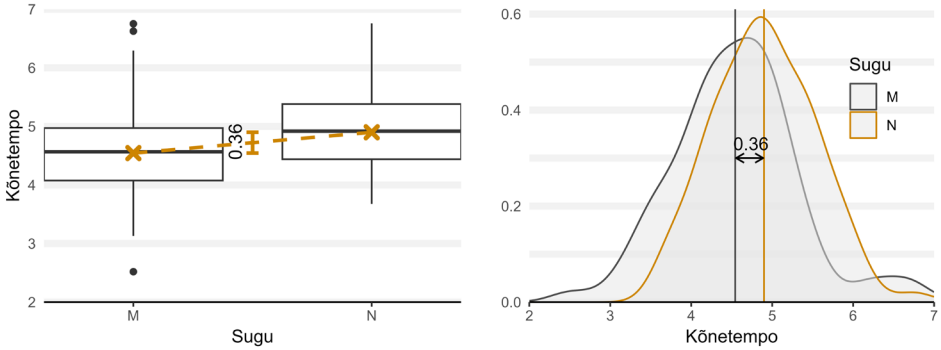
- Hii-ruut-test, nagu õigupoolest teisedki siin kirjeldatud statistilised testid, eeldab, et **kõik vaatlused andmestikus on üksteisest sõltumatud**. Korpuslingvistilises uurimuses tähendaks see näiteks seda, et kõik mingi nähtuse kasutusjuhud on pärit erinevatelt kõnelejatelt selleks, et ühegi kõneleja idiolekt tulemusi ei kallutaks. See tingimus ei ole korpuslingvistikas enamasti täidetud, kuna tihtipeale ei ole võimalik teada, kes on mingi üksiku kasutusjuhu autor, seda eriti näiteks veebist korjatud suurte tekstikorpuste puhul. Väiksemate erikorpuste puhul võib korpus olla aga nii väike, et kui võtta sealt andmestikku igalt autorilt või kõnelejalt ainult üks kasutusjuht, on andmestik järelduste tegemiseks liiga piiratud. Nõnda on korpuslingvistilistes uurimustes üldjuhul kasutusel andmestikud, milles mõni kõneleja või tekst panustab palju vaatlusi, suur hulk kõnelejaid aga ainult üksikuid vaatlusi. Ehkki see on korpuslingvistikas praktiline paratamatus, peame alati teadvustama riski, et populatsiooni keelekasutuse kirjeldamisel võime esitada veidi kallutatud pilti, mida mõjutavad need, kelle teksti rohkem valimisse satub. Tänapäeval pakuvad ühe võimaliku lahenduse probleemile uuemad statistilised mudelid, nn **segamõjudega mudelid** (vt alapeatükk 6.2.2.1.3), mis võimaldavad arvesse võtta ka seda, et valimisse sattunud vaatlused võivad olla kõnelejati või tekstiti rühmitunud (vt ka P. Lippuse näidisuurimust eesti keele väldetest ning L. Lindströmi ja M.-L. Pilviku näidisuurimust 1. isiku asesõna väljendamisest eesti murretes).

6.2.1.2. Arvulise ja kategoorilise tunnuse vahelised seosed: t-test ja U-test

Arvulise tunnuse ja kategoorilise tunnuse vahelise seose uurimisel saame võrrelda seda, kas arvulise tunnuse keskväärtused erinevates kategoorilise tunnuse kategooriates/rühmades on sarnased või erinevad, ning hinnata, kas see täheldatud erinevus on statistiliselt oluline või sattunud meie valimisse puhtalt juhuse läbi. Arvulise tunnuse keskväärtusi **kahes rühmas** saab võrrelda t-testi või U-testi abil, sõltuvalt sellest, mis tüüpi arvulise tunnusega on tegemist, kuidas tunnuse väärtused kahes rühmas jaotunud on, ja sellest, kui palju andmeid meil kasutada on. Rühmad võivad korpuslingvistilistes uurimustes olla siinjuures näiteks kaks korpust, kaks žanri/tekstitüüpi/registrit, kaks kõnelejate rühma või ka kaks grammatilist kategooriat.

Kirjeldava statistika peatükis vaatasime, kuidas keskmine kõnetempo erineb meeste ja naiste hulgas ning nägime karpdiagrammilt, et naiskõnelejad räägivad keskmiselt pisut kiiremini kui meeskõnelejad. Statistilise testi abil saame nüüd kontrollida, kas see erinevus on ka statistiliselt oluline.

T-test võrdleb mingi arvulise tunnuse aritmeetilisi keskmisi kahes rühmas. Näiteks võime tahta testida, kas joonisel 6.1 näidatud keskmine erinevus meeste ja naiste kõnetempos (0,36 silpi sekundis) on statistiliselt oluline või mitte.

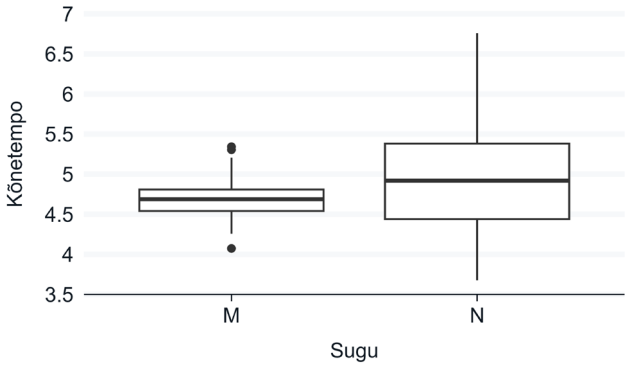


Joonis 6.1. Mees- ja naiskõnelejate kõnetempode jaotumine EKSKFK-s karpdiagrammil (vasakul, rühmade keskmised märgitud *x*-idega) ja tihedusgraafikul (paremal, rühmade keskmised märgitud vertikaaljoontega).

T-testis leitakse rühmadevahelise erinevuse hindamiseks **t-statistik**. Nagu ka hii-ruut-testis, saame teha kindlaks, kas statistiku väärtus sellise vabadusastmete arvu ja olulisuse nivoo juures on nullist oluliselt erinev: seda võib teha nii statistiku väärtuse võrdlemise põhjal t-jaotuse kriitiliste väärtustega (vt Stefanowitsch 2020: 451) kui ka p-väärtuse põhjal. T-testi puhul leitakse **vabadusastmete arv** kahe rühma võrdlemisel nõnda, et liidetakse kummagi rühma vaatluste arvud kokku ning lahutatakse saadud tulemusest võrreldavate rühmade arv ehk 2. Kõnetempo andmestikus on 74 vaatlust mees- ja 90 vaatlust naiskõnelejatelt, seega on vabadus-astmete arv 74 + 90 – 2 = 162. Sisuliselt näitab vabadusastmete arv ka siin vabalt varieeritava info hulka: kui kummagi rühma keskväärtus ei tohi muutuda, saame rühmades vabalt muuta peaaegu kõikide vaatluste väärtusi, välja arvatud viimase vaatluse oma, mille väärtus tuleneb automaatselt sellest, millised väärtused on kõi-kidel teistel vaatlustel. Vastasel juhul rühma keskväärtus muutuks. Nullhüpoteesi („rühmade keskmised ei erine“) saame hüljata jällegi siis, kui t-statistiku väärtus vabadusastmete arvu ja olulisuse nivoo korral on suurem kui jaotusega määratud kriitiline väärtus või kui testi p-väärtus on väiksem kui meie seatud olulisuse nivoo.

T-testist on erinevaid versioone sõltuvalt sellest, kas 1) kaks kategoorilise tun-nuse rühma on üksteisest sõltumatud või sõltuvad ning 2) kas rühmade hajuvus nende rühmade keskmistest on sarnane või mitte. **Rühmade sõltumatus** tähendab seda, et vaatlused ühes rühmas ei ole kuidagi süsteemselt seotud vaatlustega teises rühmas. Näiteks meie andmestikus ei ole meeste ja naiste rühmas samu kõnele-jaid (sest ükski kõneleja ei ole korpuses märgitud mõlema soo esindajaks). Kui

rühmades on aga andmed samadelt kõnelejatelt erinevatel ajahetkedel või erinevates situatsioonides, on tegemist sõltuvate valimitega ja nn paarisvõrdlusega. Tüüpiliselt kohtab selliseid andmeid kõiksugu enne-ja-pärast-uuringutes; korpuslingvistikas tuleb neid ette aga näiteks siis, kui võrdleme tekste ja nende tõlkeid või ühtede ja samade kõnelejate/kirjutajate tekste eri keelekasutussituatsioonides või eri aegadel. Sellisel juhul tuleks kasutada t-testi nn paarisversiooni (ingl *paired t-test*). Olenemata rühmade sõltumatusest eeldab t-test (täpsemalt Studenti t-test) vaikimisi, et **hajuvus keskmisest on mõlemas rühmas sarnane**. Kui aga on olukord, kus ühes rühmas on arvulise tunnuse väärtused rühma keskmisele väga sarnased, teises aga keskmisest väga erinevad, tuleks kasutada **Welchi t-testi** versiooni. Meie algandmestikus näib kõnetempo olevat joonise 6.1 põhjal nii meeskui ka naiskõnelejatel sarnase hajuvusega. Kui aga rühmad oleksid erineva hajuvusega, võiks andmete jaotumine näha välja selline nagu joonisel 6.2 (meeskõnelejad räägivad omavahel oluliselt sarnasema kiirusega kui naiskõnelejad omavahel).



Joonis 6.2. Hüpoteetiline mees- ja naiskõnelejate kõnetempode jaotumine erineva hajuvusega rühmade korral

Lisaks visuaalsele vaatlusele saab ka rühmade hajuvuse sarnasust kontrollida statistilise testiga, näiteks F-testiga (R-is funktsioon *var.test()*), mille nullhüpotees on, et kahe rühma hajuvus on võrdne.

```
> var.test(konetempo ~ sugu, data = fonkorp2) # rühmade võrdse hajuvuse test

        F test to compare two variances

data:  konetempo by sugu
F = 1.5071, num df = 73, denom df = 89, p-value = 0.0651
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

```
 0.9745578 2.3534922
sample estimates:
ratio of variances
          1.507135
```

Kui F-testi p-väärtus on suurem kui seatud olulisuse nivoo 0,05 (nagu antud juhul), saame jääda nullhüpoteesi juurde ja teha võrdset hajuvust eeldava Studenti t-testi. Kui p-väärtus on väike, peame tegema Welchi t-testi.

```
> t.test(konetempo ~ sugu, data = fonkorp2, var.equal = TRUE) # võrdse
hajuvuse puhul Studenti t-test (var.equal = TRUE)

       Two Sample t-test

data:  konetempo by sugu
t = -3.2765, df = 162, p-value = 0.001286
alternative hypothesis: true difference in means between group M and group N
is not equal to 0
95 percent confidence interval:
 -0.5704365 -0.1414133
sample estimates:
mean in group M mean in group N
       4.543509        4.899434

> t.test(konetempo ~ sugu, data = fonkorp2, var.equal = FALSE) # erineva
hajuvuse puhul Welchi t-test (var.equal = FALSE)

       Welch Two Sample t-test

data:  konetempo by sugu
t = -3.2121, df = 140.16, p-value = 0.001635
alternative hypothesis: true difference in means between group M and group N
is not equal to 0
95 percent confidence interval:
 -0.5749964 -0.1368534
sample estimates:
mean in group M mean in group N
       4.543509        4.899434
```

Testi tulemuste raporteerimisel esitatakse tavaliselt t-statistiku väärtus (sulgudes vabadusastmete arv) ja testi p-väärtus, näiteks „Mees- ja naiskõnelejate keskmise kõnetempo vahel on statistiliselt oluline erinevus ($t(162) = -3{,}28$, $p = 0{,}001$)“. Nagu hii-ruut-testi tulemust täiendas Craméri *V* seosekordaja, sobib ka siin lisaks väljendada ühe tunnuse **mõju suurust** teisele tunnusele. T-testi puhul kasutatakse mõju suuruse hindamiseks sageli **Coheni *d*** või **Hedgesi *g*** statistikut, mis sobib paremini väikeste valimite puhul.

```
> library(effectsize) # laadime paketi effectsize funktsioonid
> cohens_d(konetempo ~ sugu, data = fonkorp2, pooled_sd = TRUE) # Coheni d
võrdse hajuvusega rühmadega (pooled_sd = TRUE)
Cohen's d |         95% CI
-------------------------
-0.51     | [-0.83, -0.20]

- Estimated using pooled SD.

> cohens_d(konetempo ~ sugu, data = fonkorp2, pooled_sd = FALSE) # Coheni d
ebavõrdse hajuvusega rühmadega (pooled_sd = FALSE)
Cohen's d |         95% CI
-------------------------
-0.51     | [-0.82, -0.19]

- Estimated using un-pooled SD.

> hedges_g(konetempo ~ sugu, data = fonkorp2, pooled_sd = TRUE) # Hedgesi g
väikeste valimite jaoks võrdse hajuvusega rühmadega (pooled_sd = TRUE)
Hedges' g |         95% CI
-------------------------
-0.51     | [-0.82, -0.20]

- Estimated using pooled SD.

> hedges_g(konetempo ~ sugu, data = fonkorp2, pooled_sd = FALSE) # Hedgesi g
väikeste valimite jaoks ebavõrdse hajuvusega rühmadega (pooled_sd = FALSE)
Hedges' g |            95% CI
-------------------------
-0.51  | [-0.82, -0.19]

- Estimated using un-pooled SD.
```

Praegusel juhul näivad kõigil neljal moel arvutatud mõju suuruse hinnangud ühesugused, kuna erinevused tulevad ette alles kümnendmurru kaugema järgu kümnendkohtades ning usaldusvahemikes. Statistiku tõlgendamiseks võib kasutada Coheni (1988) pakutud skaalat (tabel 6.7).

Tabel 6.7. Coheni *d* tõlgendamise skaala vastavalt *d* absoluutväärtusele

$\|d\| < 0{,}2$	väga nõrk
$0{,}2 \leq \|d\| < 0{,}5$	nõrk
$0{,}5 \leq \|d\| < 0{,}8$	keskmine
$\|d\| \geq 0{,}8$	tugev

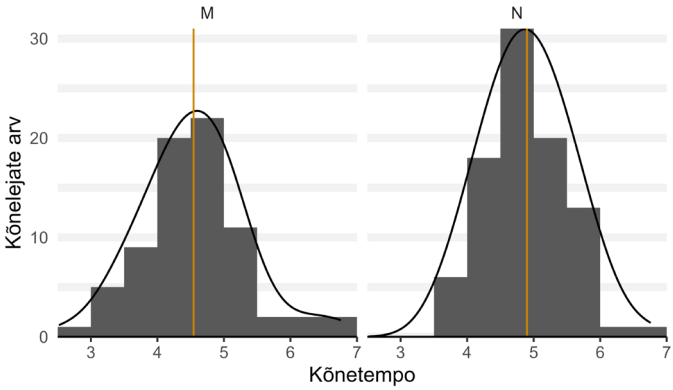
Ka t-testi kasutamisel on omad **eeldused**:

- Esiteks eeldab (kahe valimiga) test, et **kõik vaatlused on üksteisest sõltumatud**. Arvulise tunnuse puhul on selle eelduse täitmiseks võimalik näiteks kõikide kõnelejate/tekstide üksikvaatlused keskmistada, nii et igalt kõnelejalt või igast tekstist oleks vaatlusena/reana andmestikus ainult üks keskmine näitaja. Kui seda ei ole võimalik teha, on ka siin abiks **segamõjudega mudelid** (vt alapeatükk 6.2.2.1.3). Vaadeldud kõnetempo andmestikus on tegelikult vaatluste sõltumatuse nõue rikutud, kuna kõneleja, kes on osalenud mitmes erinevas vestluses, on esindatud andmestikus mitmel real. Seega võiksime tegelikult t-testi tegemiseks leida iga andmestikus oleva unikaalse kõneleja keskmise kõnetempo kõikide vestluste peale, milles ta osales.

```
> konetempo_yld <- aggregate(konetempo ~ koneleja + sugu, fonkorp2,
mean) # teeme uue andmestiku, milles keskmistame kõnetempo iga
andmestiku kõneleja (ja tema soo) kohta
> gg_boxplot(data = konetempo_yld, x = sugu, y = konetempo, y_label
= "Kõnetempo", mode = light_mode_t()) # kuvame uue andmestiku
keskmistatud kõnetempot joonisel
> var.test(konetempo ~ sugu, data = konetempo_yld) # kontrollime
kõnetempo hajuvust rühmades
> t.test(konetempo ~ sugu, data = konetempo_yld, var.equal = TRUE) #
testime kõnetempo keskmiste erinevust

        Two Sample t-test

data:  konetempo by sugu
t = -3.015, df = 137, p-value = 0.003063
alternative hypothesis: true difference in means between group M and
group N is not equal to 0
95 percent confidence interval:
 -0.5693190 -0.1183193
sample estimates:
mean in group M mean in group N
       4.608052        4.951871
```

- Teiseks eeldab t-testi tegemine, et arvuline tunnus (nt kõnetempo) on populatsioonis mõlemas rühmas **normaaljaotusega**. Normaaljaotus tähendab, et kõige rohkem esineb andmetes aritmeetilise keskmise ümber koonduvaid väärtusi ning ülejäänud väärtused hajuvad ühtlaselt mõlemale poole keskmist. Ühtmoodi vähe on normaaljaotusega väärtuste hulgas väga väikeseid ja väga suuri väärtusi. Jaotus on seega sümmeetriline ja nii-öelda kellukese kujuga ning normaaljaotuse seaduspärade tõttu saame aritmeetilise keskmise ja standardhälbe põhjal määrata ka kõikide teiste vaatluste ligikaudsed väärtused, isegi kui me neid täpselt ei tea.

Kui vaatame histogrammidelt (joonis 6.3) EKSKFK andmestiku meeste ja naiste kõnetempot, näeme, et siin on tegemist üpris sümmeetriliste jaotustega: kõige rohkem on keskmise kõnetempoga kõnelejaid (mida joonisel märgib oranž vertikaaljoon) ning sellest aeglasemaid ja kiiremaid kõnelejaid on seda vähem, mida suurem erinevus keskmisest kõnetempost on.



Joonis 6.3. Mees- ja naiskõnelejate kõnetempo normaaljaotusele vastavuse visuaalne kontrollimine

Lisaks visuaalsele vaatlusele on olemas ka statistilisi teste, mille abil normaaljaotuse eeldust kontrollida. Üks sagedamini kasutatud teste on **Shapiro-Wilki test**, mille nullhüpotees on, et valim on pärit normaaljaotusest. Seda nullhüpoteesi on testi tõlgendamisel oluline silmas pidada, kuna erinevalt seose olulisuse testidest loodame siin jääda nullhüpoteesi juurde ja näha testi tulemusena seega p-väärtust, mis oleks piisavalt suur selleks, et poleks alust nullhüpoteesi hüljata.

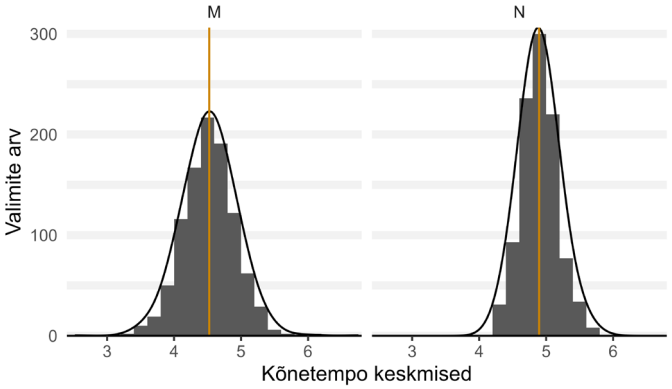
```
> shapiro.test(fonkorp2$konetempo) # kas kõnetempo üldiselt on
normaaljaotusega?

        Shapiro-Wilk normality test

data:  fonkorp2$konetempo
W = 0.99362, p-value = 0.6928

> tapply(fonkorp2$konetempo, fonkorp2$sugu, shapiro.test) # kas kõnetempo on
normaaljaotusega mõlemas rühmas (siin rühmas M ja rühmas N)?
$M
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97753, p-value = 0.2101
```

```
$N

       Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98898, p-value = 0.656
```

Kuna nii meeste kui ka naiste kõnetempo puhul saame p-väärtused, mis on märkimisväärselt suuremad kui vaikimisi seatud olulisuse nivoo 0,05, pole piisavalt tõendeid selle kohta, et andmed ei oleks normaaljaotusega.

Ehkki t-test on paljudes valdkondades levinud statistiline test, võib selle kasutamist korpus- või laiemalt keeleandmete analüüsiks pärssida tõsiasi, et **loomuliku keele andmetest tuletatud arvandmed** (nt sõnade pikkus ja sagedus või viitamiskaugused), aga ka kõiksugu metaandmed, nagu kõnelejate vanus, teoste ilmumisaastad jm, **on harva normaaljaotusega**. Mida sellisel juhul teha?

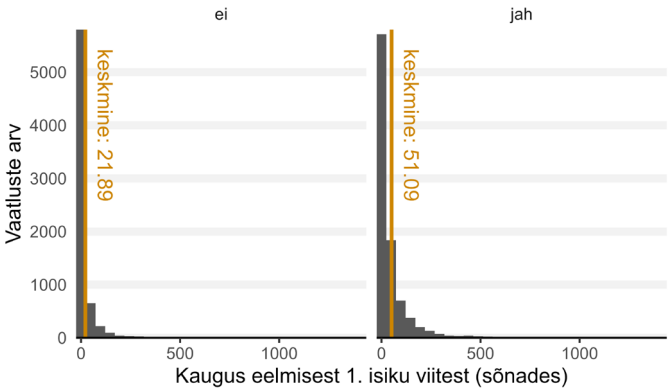
Esiteks võib püüda lihtsalt **valimit suurendada**. Normaaljaotuse nõue on rangem väikestes valimites, kus ühes rühmas on 30 vaatlust või vähem, korpusandmestikud on aga tavaliselt sellest oluliselt suuremad. Valimi suurendamisel muutub normaaljaotuse nõue vähem oluliseks tänu nn **tsentraalsele piirteoreemile** (ingl *central limit theorem*), mille järgi piisavalt paljude üksikute valimite aritmeetilised keskmised hakkavad lähenema normaaljaotusele, ehkki ükski valim ise ei pruugi normaaljaotusega olla. Näiteks kui võtame oma 164 vaatlusega kõnetempo andmestikust 1000 väiksemat valimit, milles igaühes on ainult 10 algandmetest juhuslikult valitud kõneleja kõnetempo, ning arvutame iga valimi meeste ja naiste keskmised kõnetempod, siis näeme, et need eri valimite keskmised on samuti normaaljaotusega. Jooniselt 6.4 näeme, et 1000st valimist u 200 ringis on neid, kus 10 kõneleja puhul jääb meeste keskmine kõnetempo u 4,5 silbi juurde sekundis. Naiste kõnetempo keskmised on juhuslikes valimites vähem hajuvad: u 300 valimis 1000st jääb 10 kõnelejaga valimis naiste kõnetempo keskmine 4,9 silbi kanti sekundis. Need väärtused on praktiliselt identsed terve andmestiku tegelike kõnetempo keskmistega (meestel 4,54 ja naistel 4,89 silpi sekundis).



Joonis 6.4. 1000 juhusliku valimi keskmiste kõnetempode jaotus

Kuigi uurime enamasti ainult ühtainsat, mitte mitut valimit, aitab tsentraalne piirteoreem meid seeläbi, et suurem uuritav valim katab rohkem populatsiooni andmepunkte ja nõnda ka erinevaid võimalikke väikeseid juhuslikke valimeid. Seetõttu on suuremate valimite põhjal arvutatud keskmine lähemal tegelikule populatsiooni keskmisele, samamoodi ka keskmiste erinevus. Seda hoolimata sellest, et ühele või teisele poole kaldu olevate jaotuste puhul ei pruugi aritmeetiline keskmine väga hästi kirjeldada individuaalselt ühtki populatsiooni liiget.

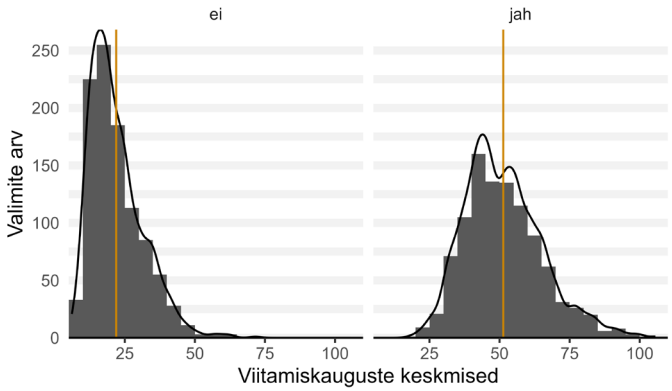
Vaatame nüüd võrdluseks L. Lindströmi ja M.-L. Pilviku siinse õpiku näidisuurimuse 1. isiku asesõna väljendamise andmestikust viitamiskauguse tunnust, mis väljendab, kui kaugel vaadeldavast 1. isiku verbivormist (nt *tahan*) on sõnades mõõdetuna eelmine lähim viide 1. isikule. Võiks oletada, et mida lähemal on



Joonis 6.5. Kaugus eelmisest 1. isiku viitest asesõnata (*ei*) ja asesõnaga (*jah*) 1. isiku verbivormide puhul (nt *tulen vs. ma tulen*)

eelmine viide, seda vähem peavad kõnelejad vajalikuks eksplitsiitselt subjektpronoomenit (*mina*/*ma*) väljendada.

Jooniselt 6.5 näeme, et tegemist on tugevalt ebasümmeetrilise jaotusega: enamik eelmisi viiteid 1. isikule on vaadeldavale vormile suhteliselt lähedal (umbes kuni 10 sõna kaugusel), ent harvadel juhtudel on eelmise ja vaadeldava viite vahel väga suur hulk sõnu (kõige äärmuslikumal juhul tervelt 1415 sõna). Sellise jaotusega arvuliste tunnuste puhul paigutub aritmeetiline keskmine suurema osa andmete väärtustest oluliselt kõrgemale. Võttes aga 16 200-st andmestiku vaatlusest jällegi 1000 väiksemat valimit, milles igaühes 100 juhuslikult valitud rida algandmestikust, näeme, et tänu tsentraalsele piirteoreemile jaotuvad nende valimite keskmised viitamiskaugused pronoomenita ja pronoomeniga vaatluste puhul küllaltki sümmeetriliselt, hoolimata sellest, et viitamiskauguse jaotus ise oli äärmiselt ebasümmeetriline (joonis 6.6).



Joonis 6.6. 1000 juhusliku valimi keskmiste viitamiskauguste jaotus asesõnata (*ei*) ja asesõnaga (*jah*) 1. isiku verbivormide puhul

Siiski ei pruugi nii kaldu olevate andmetega kahe rühma aritmeetiliste keskmiste võrdlemine olla keeleteaduslikus mõttes kuigi informatiivne. Aritmeetilised keskmised on väga mõjutatud erandlikult suurtest või väikestest väärtustest, ent viitamiskauguse puhul, mis peaks kõnelema inimese töömälu rollist keeleliste valikute tegemisel, huvitaksid meid tõenäoliselt pigem erinevused just skaala madalamas otsas.

Normaaljaotuse eelduse rikkumisel võib proovida arvandmeid niisiis hoopis mingil moel **teisendada**. Nn paremale kaldu oleva jaotuse puhul, kus meil on palju väikeseid ja vähe suuri väärtusi (nagu nt sõnasagedused või eelnevas näites toodud viitamiskaugused), on tüüpiliseks teisendamise võtteks logaritmimine. **Logaritmimine** on astendamise pöördtehe ning teisendab tegelikud arvud mingi kokkuleppelise arvu ehk *aluse* (nt 2 või 10) astmeteks (tabel 6.8).

Tabel 6.8. Tegelikud väärtused ja nende logaritmitud väärtused vastavalt astendamise alusele

Tegelik väärtus	Astendamise alus	Logaritmitud väärtus (aste)	Kontroll
1	2	**0**	2^0 = 1
4	2	**2**	2^2 = 4
10	2	**3,321928**	2^3,321928 = 10
100	2	**6,643856**	2^6,643856 = 100
1	10	**0**	10^0 = 1
4	10	**0,60206**	10^0,60206 = 4
10	10	**1**	10^1 = 10
100	10	**2**	10^2 = 100

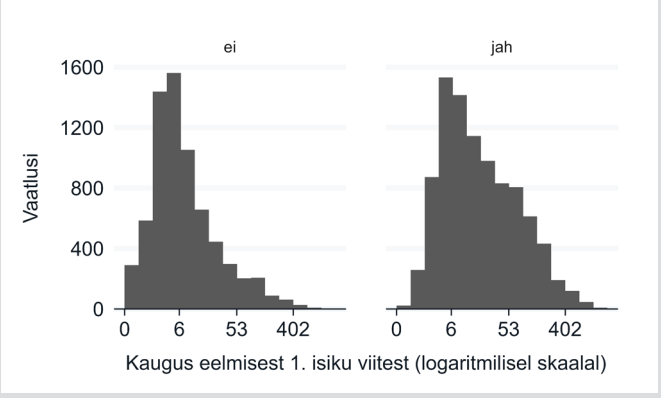
Sageli kasutatakse logaritmimisel ka naturaallogaritmi *ln*, mispuhul aluseks on nn Euleri arv *e*, mille ligikaudne väärtus on 2,71828. Arvud 1–4 saaks teisendada naturaallogaritmi abil seega arvu 2,71828 vastavateks astmeteks: 0, 0,6931472, 1,098612 ja 1,386294.

Logaritmimine vähendab seega erinevusi suurte väärtuste vahel ning toob enam esile erinevusi väiksemate väärtuste vahel, sealjuures on logaritmitud väärtused seda väiksemad, mida suurem on astendamise alus (vrd tabelis 6.8 logaritmitud väärtusi alusel 2 ja 10). Näiteks on arv 100 kümme korda suurem kui arv 10, ent kui teisendada mõlemad aluse 10 astmeteks, siis on nende vahe vaid kahekordne, kuna 2 on kaks korda suurem kui 1. Paljusid korpuslingvistilistes uurimustes levinud arvulisi näitajaid, näiteks sagedusi, viitekaugusi, aga ka näiteks hertsiskaalat (vt P. Lippuse näidisuurimust eesti keele väldetest), tajutaksegi pigem logaritmilisel skaalal: kui eelmine viide samale isikule oli ühes kontekstis 2 sõna kaugusel ja teises 22 sõna kaugusel, on see erinevus kõneleja töömälu jaoks olulisem kui erinevus 102 ja 122 sõna vahel. Olgu öeldud, et logaritmida saab ainult positiivseid, st nullist suuremaid arve. Kui arvulise tunnuse väärtuste hulgas on ka nulle või negatiivseid arve, võib kõikidele väärtustele enne logaritmimist liita mingi konstandi, mis muudaks kõik väärtused nullist suuremaks: näiteks –10 + **11** = 1, –5 + **11** = 6, 0 + **11** = 11, 4 + **11** = 15, 5 + **11** = 16.

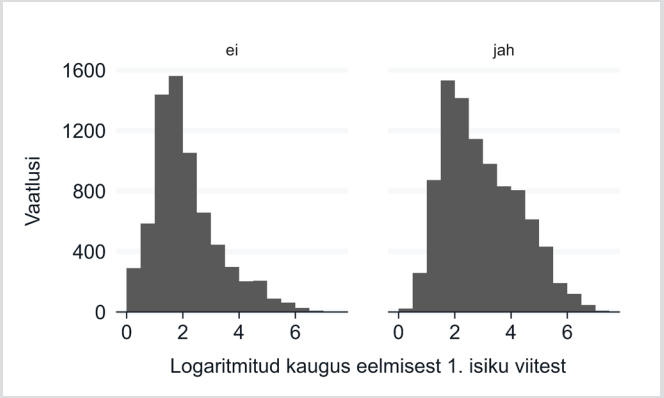
Kui logaritmime ülal näiteks toodud viitamiskaugusi, näeme, et erinevused suurte väärtuste vahel muutuvad väiksemaks ning jaotus muutub oluliselt sümmeetrilisemaks. Hõlpsamaks võrdlemiseks on esimesel alumisel graafikul kuvatud tegelikud viitamiskaugused (sõnades) *logaritmilisel skaalal* ning teisel graafikul tegelike viitamiskauguste *logaritmitud väärtused*, erinevus kahe graafiku vahel seisneb x-teljel kuvatavates väärtustes. Kuna viitamiskauguste seas on ka nulle (eelmine viide on vahetult vaadeldava vormi ees), peame logaritmimisel liitma

kõikidele algsetele viitamiskaugustele arvu 1 (R-i koodis *log1p*). T-testi tehes peaksime võrdlema tegelike viitamiskauguste asemel nüüd logaritmitud viitamiskaugusi asesõnaga ja asesõnata kasutusjuhtude puhul.

```
> # kuvame tegelikke viitamiskaugusi logaritmilisel skaalal
> gg_histogram(data = isik1, x = kaugus_eelmisest,
               facet = pron,
               y_label = "Vaatlusi",
               x_label = "Kaugus eelmisest 1. isiku viitest (logaritmilisel
skaalal)",
               binwidth = 0.5,
               boundary = 0,
               mode = light_mode_t(),
               x_transform = "log1p", # kuvame tegelikke viitamiskaugusi
logaritmilisel skaalal
               y_breaks = seq(0, 1600, 400), # y-telje intervallid
               y_labels = seq(0, 1600, 400), # y-telje intervallide sildid
               x_breaks = floor(exp(seq(0,6,2))-1)) # x-telje intervallid
```




```
> # kuvame logaritmitud viitamiskaugusi
> isik1$kaugus_eelmisest_log1p <- log1p(isik1$kaugus_eelmisest) # loome
logaritmitud kaugustega tunnuse
> gg_histogram(data = isik1, x = kaugus_eelmisest_log1p, # kuvame logaritmitud
viitamiskaugusi
               facet = pron,
               y_label = "Vaatlusi",
               x_label = "Logaritmitud kaugus eelmisest 1. isiku viitest",
               binwidth = 0.5,
               boundary = 0,
               mode = light_mode_t(),
               y_labels = seq(0, 1600, 400))
```

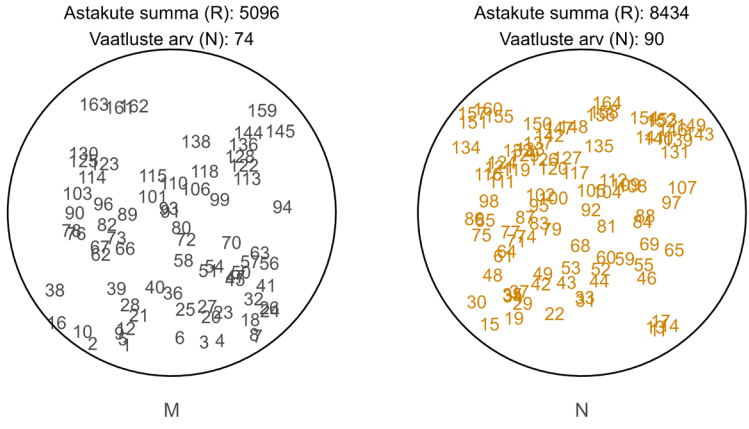


Kolmandaks võime normaaljaotuse eelduse rikkumisel kasutada mõnd sellist testi, mis normaaljaotuse eeldust ei sea. Selliseks testiks on näiteks sõltumatute rühmadega Manni-Whitney **U-test** (ingl *Mann-Whitney U-test*), mida tuntakse ka nime all Wilcoxoni astaksummatest (ingl *Wilcoxon rank sum test*). Kui tegemist on paarisvõrdlusega (ingl *paired*), saab kasutada Wilcoxoni astakmärgitesti (ingl *Wilcoxon signed rank test*).

U-test on sõltumatute rühmadega t-testi mitteparameetriline alternatiiv. See ei võrdle mitte kahe rühma aritmeetilisi keskmisi, vaid järjekorranumbritel ehk astakutel (ingl *rank*) põhinevaid rühmadevahelisi erinevusi. Erinevalt t-testist ei eelda U-test niisiis, et andmed pärinevad normaaljaotusega populatsioonist, mida saab kirjeldada kindlate parameetrite alusel, nagu aritmeetiline keskmine ja standardhälve. Selliseid meetodeid, mis ei eelda populatsioonilt mingit kindlat statistilist jaotust (nt normaaljaotust või ka hii-ruut-jaotust), nimetataksegi **mitteparameetrilisteks**.

U-testi nullhüpotees on, et tõenäosus, et juhuslikult ühest rühmast valitud väärtus (nt juhusliku naiskõneleja kõnetempo) on suurem kui teisest rühmast juhuslikult valitud väärtus (juhusliku meeskõneleja kõnetempo), on sama suur kui tõenäosus, et see on väiksem. See viitab sellele, et mingi vaatluse emba-kumba rühma kuulumine ei aita ennustada tema arvulise tunnuse väärtust. Test põhineb arvulise tunnuse väärtuste järjestamisel kasvavas järjekorras nõnda, et igale vaatlusele määratakse järjekorranumber ehk astak. Kõige väiksema astaku väärtus on 1 ja kõige suurem astak võrdub valimi kõikide vaatluste arvuga. Kui kahe rühma vaatlused paiknevad sellises järjestatud arvureas läbisegi, ei erine rühmad teineteisest arvulise tunnuse jaotuse poolest ja kui mõlemas rühmas on sama palju vaatlusi, ei tohiks ühe rühma astakute summa olla sellisel juhul teise rühma omast oluliselt suurem ega väiksem. Kui kaks rühma on aga teineteisest täiesti erinevad, paiknevad järjestatud arvurea kõrgemas otsas ainult ühe rühma vaatlused ning madalamas otsas teise rühma vaatlused ning teadmine, kumba rühma mingi vaatlus kuulub, aitab ennustada ka vaatluse huvipakkuva arvulise tunnuse väärtust.

```
> fonkorp2$astak <- rank(fonkorp2$konetempo) # lisame vaatlustele astakud
> head(fonkorp2[order(fonkorp2$konetempo), c("sugu", "konetempo", "astak")],
10) # 10 aeglaseimat kõnelejat
    sugu konetempo astak
154    M  2.519344     1
1      M  3.129401     2
89     M  3.162008     3
18     M  3.389018     4
8      M  3.482338     5
13     M  3.487966     6
144    M  3.516548     7
145    M  3.570801     8
20     M  3.583910     9
7      M  3.643304    10

> tail(fonkorp2[order(fonkorp2$konetempo), c("sugu", "konetempo", "astak")],
10) # 10 kiireimat kõnelejat
    sugu konetempo astak
113    N  5.901199   155
9      N  5.913105   156
160    N  5.971259   157
149    N  5.986507   158
164    M  6.117915   159
64     N  6.171709   160
60     M  6.293195   161
106    M  6.626139   162
43     M  6.751879   163
105    N  6.758682   164
```



Joonis 6.7. Mees- ja naiskõnelejate keskmiste kõnetempode järjestamisel saadud astakud ehk järjekorranumbrid ja astakute summad

Kõnetempo andmestiku puhul näeme, et 10 kõige aeglasema kõneleja hulgas on ainult meeskõnelejad ning 10 kõige kiirema kõneleja hulgas on rohkem naiskõnelejaid. Seega ei paikne mees- ja naiskõnelejate kõnetempo väärtused meie järjestatud tabelis täiesti juhuslikult läbisegi ning võiksime oletada, et kõneleja soo ja kõnetempo vahel on seos.

U-testi teststatistik U (või W) leitakse astakute summa (R) ning rühmade vaatluste arvu põhjal (N). Kõnetempo andmestiku vastavad andmed on esitatud joonisel 6.7.

Statistiku arvutamiseks kasutatakse alltoodud valemeid (N_1 on esimese rühma vaatluste arv, N_2 teise rühma vaatluste arv, R_1 esimese rühma astakute summa, R_2 teise rühma astakute summa) ning valitakse väärtus, mis on väiksem.

$$U_1 = (N_1 \times N_2) + \frac{N_1 \times (N_1 + 1)}{2} - R_1$$

$$U_2 = (N_1 \times N_2) + \frac{N_2 \times (N_2 + 1)}{2} - R_2$$

Praegusel juhul saame leida niisiis mees- ja naiskõnelejate rühmade U väärtused ning valida neist vähima.

$$U_M = (74 \times 90) + \frac{74 \times (74 + 1)}{2} - 5096 = 4339$$

$$U_N = (74 \times 90) + \frac{90 \times (90 + 1)}{2} - 8434 = 2321$$

Selleks, et saada teada, kas vähima statistiku väärtus (siin $U = 2321$) väljendab statistiliselt olulist seost, võrreldakse seda väikeste valimite puhul (mõlema rühma $n < 25$) taaskord kriitiliste väärtuste tabeliga (vt Stefanowitsch 2020: 449–450). Suuremate valimite puhul saab U-statistiku põhjal arvutada z-statistiku väärtuse ning võrrelda seda normaaljaotuse kriitiliste väärtustega. Samuti saame hinnata seose olulisust p-väärtuse abil.

```
> wilcox.test(konetempo ~ sugu, data = fonkorp2) # teeme U-testi

        Wilcoxon rank sum test with continuity correction

data:  konetempo by sugu
W = 2321, p-value = 0.0008603
alternative hypothesis: true location shift is not equal to 0
```

Kuna U-test tegeleb järjestatud arvureaga, sobib see hästi ka **järjestusskaala** tunnustele, nt küsimustikes sageli kasutatud Likerti skaalale, millel arvud (nt 1–5 või 1–7) väljendavad tegelikult mingeid järjestatud kategooriaid, mitte aga klassikalisi arvulisi väärtusi, nagu loendus- või mõõtmisandmed. Nagu nimi ütleb, saab järjestusskaala väärtusi küll väiksemast suuremani järjestada ning leida ka järjestatud variatsioonirea keskmise väärtuse ehk mediaani, ent matemaatilised tehted, nagu liitmine-lahutamine, korrutamine-jagamine või ka aritmeetilise keskmise leidmine, ei ole järjestusskaala väärtustega tingimata tähenduslikud. Näiteks vastused 1 („üldse ei nõustu") ja 2 („pigem ei nõustu") ei anna kokku vastust 3 („nii ja naa"), samuti ei tähenda vastus 5 („täiesti nõus") viis korda rohkem nõustumist kui vastus 1 („üldse ei nõustu"). Järjestusskaalal võib kujutada teisigi hierarhilisi tunnuseid, näiteks elusust või abstraktsust/konkreetsust. U-testiga on võimalik hinnata, kas ühes rühmas täheldatud väärtused paiknevad järjestusskaalal pigem kõrgemal või madalamal kui teises rühmas.

Ka U-testi võiks täiendada mõne **mõju suurust** kirjeldava seosekordajaga. Selleks sobivad näiteks **Glassi astakutega biseriaalne korrelatsioonikordaja** (ingl *Glass' rank biserial correlation coefficient*), mida sõltumatute rühmade puhul nimetatakse ka **Cliffi deltaks** (vt nt selles õpikus M.-L. Pilviku näidisuurimust *lt*- ja *sti*-liidete produktiivsusest).

```
> library(effectsize) # laadime paketi effectsize funktsioonid
> rank_biserial(konetempo ~ sugu, data = fonkorp2) # leiame
korrelatsioonikordaja
r (rank biserial) |         95% CI
---------------------------------
-0.30             | [-0.45, -0.13]
```

Coheni (1988) esitatud skaala järgi saab kordaja väärtust tõlgendada vastavalt tabelis 6.9 esitatule.

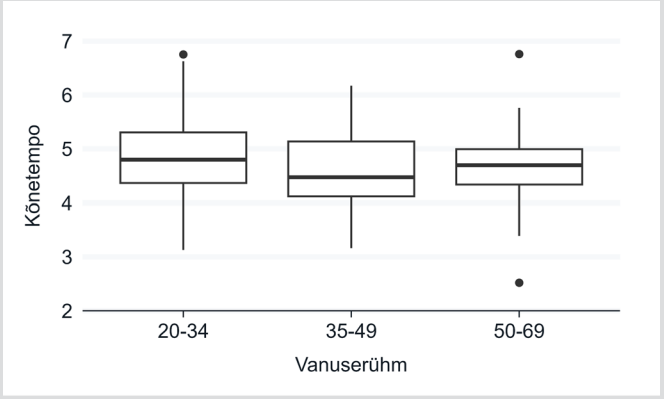
Tabel 6.9. Astakutega biseriaalse korrelatsioonikordaja *r* tõlgendamise skaala

$r < 0{,}1$	väga nõrk
$0{,}1 \leq r < 0{,}3$	nõrk
$0{,}3 \leq r < 0{,}5$	keskmine
$r \geq 0{,}5$	tugev

T-testi ja U-testiga võime võrrelda keskmisi ainult kahes rühmas, st rühmitaval kategoorilisel tunnusel saab olla ainult kaks erinevat väärtust (nt *mees*/*naine*,

pronoomeniga/*pronoomenita*). Kui **kategoorilisel tunnusel on enam kui kaks erinevat väärtust**, võib kasutada **ANOVA-testi** (ingl *analysis of variance*) ehk dispersioonanalüüsi, kui arvulise tunnuse jaotus rühmades läheneb normaaljaotusele[10], või **Kruskali-Wallise testi**, kui ei lähene ja meil on liiga vähe vaatlusi, et võiksime normaaljaotuse eeldust leevendada. Jagame näiteks kõnetempo andmestiku kõnelejad vanuse põhjal kolme vanuserühma – 20–34-aastased, 35–49-aastased ja 50–69-aastased – ning kontrollime, kas kõnetempo keskmiste erinevus eri vanuserühmades on statistiliselt oluline.

```
> # jagame kõnelejad vanuse põhjal 3 vanuserühma
> fonkorp2$vanuseryhm <- cut(fonkorp2$vanus, breaks = c(0, 34, 49, 69), labels
= c("20-34", "35-49", "50-69"))

> # karpdiagramm kõnetempo jaotumisest eri vanuserühmades
> gg_boxplot(data = fonkorp2, x = vanuseryhm, y = konetempo,
                x_label = "Vanuserühm",
                y_label = "Kõnetempo",
                mode = light_mode_t())
```




```
> tapply(fonkorp2$konetempo, fonkorp2$vanuseryhm, shapiro.test) #
normaaljaotuse kontroll rühmades
$`20-34`

       Shapiro-Wilk normality test

data:  X[[i]]
W = 0.99094, p-value = 0.8508


$`35-49`
```

[10] Õigupoolest tuleks ANOVA puhul kontrollida mitte niivõrd arvulise tunnuse enda normaaljaotust kui nn *jääkide* (ingl *residuals*) vastavust normaaljaotusele, millest on lähemalt juttu järgmises alapeatükis.

```
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97874, p-value = 0.6279


$`50-69`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.94883, p-value = 0.05404

> bartlett.test(konetempo ~ vanuseryhm, data = fonkorp2) # rühmade hajuvuse
sarnasuse kontroll

        Bartlett test of homogeneity of variances

data:  konetempo by vanuseryhm
Bartlett's K-squared = 0.17327, df = 2, p-value = 0.917

> oneway.test(konetempo ~ vanuseryhm, data = fonkorp2, var.equal = TRUE) #
parameetriline ANOVA-test sarnase hajuvusega rühmade puhul (var.equal = TRUE)

        One-way analysis of means

data:  konetempo and vanuseryhm
F = 1.6275, num df = 2, denom df = 161, p-value = 0.1996

> kruskal.test(konetempo ~ vanuseryhm, data = fonkorp2) # mitteparameetriline
Kruskali-Wallise test

        Kruskal-Wallis rank sum test

data:  konetempo by vanuseryhm
Kruskal-Wallis chi-squared = 3.6467, df = 2, p-value = 0.1615
```

Nii parameetrilise ANOVA kui ka mitteparameetrilise Kruskali-Wallise test ütlevad meile kõnetempo andmestiku puhul, et pole piisavalt tõendeid selle kohta, et eri vanuserühmade kõnetempode keskmised oleksid statistiliselt oluliselt erinevad (testide p-väärtused on suuremad kui 0,05). Kui aga statistiliselt oluline erinevus oleks olemas, näeksime testide põhjal ainult seda, et vähemalt kaks rühma on üksteisest oluliselt erinevad, ent ei näeks seda, millised need rühmad on (näiteks kas kõige nooremad erineksid kõige vanematest, kõige nooremad keskmistest, keskmised kõige vanematest või kõik rühmad üksteisest).

Erinevuste väljaselgitamiseks on võimalus teha ka iga rühmade paariga eraldi t-test või U-test (nt 3 rühma puhul teeksime 3 testi, 4 rühma puhul 6 testi jne). Sellisel juhul tuleb arvesse võtta, et teeme kahe tunnuse vahelise seose leidmiseks samade andmetega mitu võrdlust ehk testime sisuliselt mitut eraldi hüpoteesi. Seetõttu kumuleeruvad eksimismäärad, mida nullhüpoteesi ümberlükkamisel iga

paari võrdluses lubame (nn **mitmese testimise probleem**)[11]. Näiteks kui oleme seadnud olulisuse nivooks 0,05 ja võrdleme kolme vanuserühma, siis väites, et seos kahe tunnuse vahel ei ole juhuslik, lubame endale iga rühmade paari võrdluses viieprotsendilist eksimismäära. Ent tõenäosus, et teeme sealjuures vea **vähemalt ühe** hüpoteesipaari korral (nn katseviisiline viga, ingl *family-wise error rate*), on tegelikult $1 - (1 - 0{,}05)^3 = 0{,}14$ ehk 14%. Nelja rühma võrdlemise puhul kasvaks see 26% peale $(1 - (1 - 0{,}05)^6)$ jne. Üks levinud viis katseviisilise veaga arvestamiseks on kasutada **Bonferroni parandust**, mispuhul jagatakse olulisuse nivoo alfa väärtused läbi võrdluste/testide arvuga. Näiteks kui võrdleme omavahel kolme rühma, oleks iga rühmade paari võrdluse lubatud veamäär 0,05 / 3 = 0,0167 ning olulist seost väidaksime ainult siis, kui iga testi p-väärtus jääb alla selle nivoo. Bonferroni parandust peetakse sageli aga liialt konservatiivseks, eriti juhul, kui võrreldakse paljusid rühmi, kuna rühmade arvuga läbi jagades muutub alfa väärtus sellisel juhul väga väikeseks ning suureneb tõenäosus magada maha tegelikult olulisi seoseid. Paljude rühmade puhul võib kasutada seega näiteks **Holmi-Bonferroni parandust**, mis järjestab kõik paariviisilised võrdlused kasvavalt nende testi p-väärtuste järjekorras ning kahandab iga järgmise võrdluse jaoks rühmade arvu, millega alfa väärtust läbi jagada. Näiteks kolme rühma puhul peaks esimese, kõige väiksema p-väärtusega võrdluse p-väärtus olema väiksem kui 0,05 / 3 = 0,0167, teise võrdluse oma väiksem kui 0,05 / 2 = 0,025, kolmanda oma väiksem kui 0,05 / 1 = 0,05. Võrdleme näiteks kõnetempo keskmiste erinevusi, võrreldes korraga ainult kaht vanuserühma.

```
> t.test(konetempo ~ vanuseryhm, data = fonkorp2[fonkorp2$vanuseryhm != "50-
69",], var.equal = TRUE) # ainult 20-34 ja 35-49

       Two Sample t-test

data:  konetempo by vanuseryhm
t = 1.7294, df = 119, p-value = 0.08633 # p peaks olema < 0.0167
alternative hypothesis: true difference in means between group 20-34 and group
35-49 is not equal to 0
95 percent confidence interval:
 -0.03463742  0.51246882
sample estimates:
mean in group 20-34 mean in group 35-49
           4.833987            4.595072

> t.test(konetempo ~ vanuseryhm, data = fonkorp2[fonkorp2$vanuseryhm != "35-
49",], var.equal = TRUE) # ainult 20-34 ja 50-69

       Two Sample t-test

data:  konetempo by vanuseryhm
t = 1.0082, df = 121, p-value = 0.3153 # p peaks olema < 0.025
```

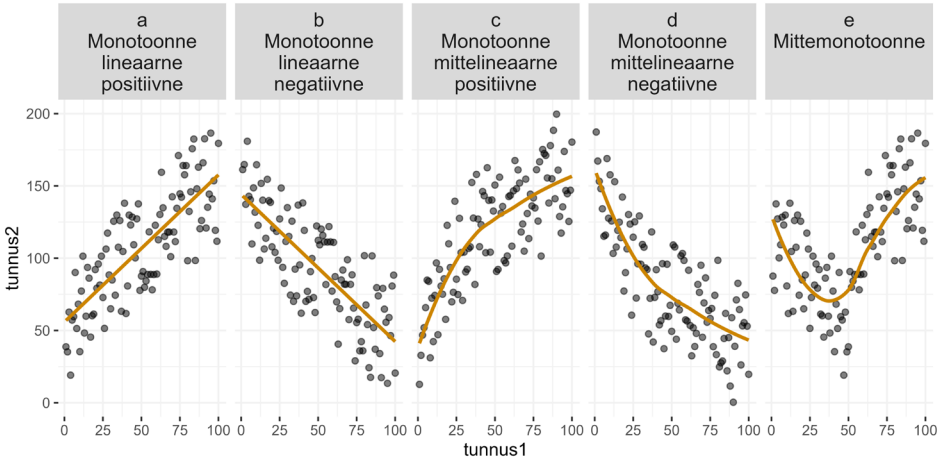
[11] Sama probleemiga puutume kokku siis, kui testime uuritava tunnuse seotust samas valimis mitme erineva seletava tunnusega ja teeme selleks mitu statistilist testi.

```
alternative hypothesis: true difference in means between group 20-34 and group
50-69 is not equal to 0
95 percent confidence interval:
 -0.1301869  0.4004043
sample estimates:
mean in group 20-34 mean in group 50-69
          4.833987            4.698879

> t.test(konetempo ~ vanuseryhm, data = fonkorp2[fonkorp2$vanuseryhm != "20-
34",], var.equal = TRUE) # ainult 35-49 ja 50-69

       Two Sample t-test

data:  konetempo by vanuseryhm
t = -0.68145, df = 82, p-value = 0.4975 # p peaks olema < 0.05
alternative hypothesis: true difference in means between group 35-49 and group
50-69 is not equal to 0
95 percent confidence interval:
 -0.4068467  0.1992327
sample estimates:
mean in group 35-49 mean in group 50-69
          4.595072            4.698879
```

Näeme, et statistiliselt olulisi erinevusi kõnetempo keskmistes ei ole ühegi vanuserühmade paari vahel. Seega saame kinnitust kolme rühma võrrelnud ANOVA-testi ja Kruskali-Wallise testi tulemustele.

6.2.1.3. Kahe arvulise tunnuse vahelised seosed: Pearsoni ja Spearmani korrelatsioonikordajad

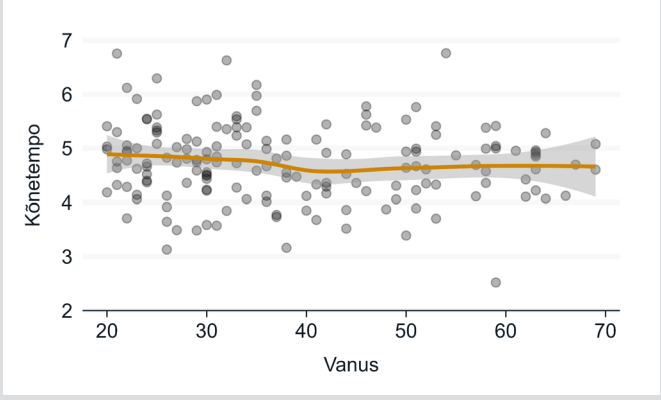
Kahe arvulise tunnuse vahelise seose hindamiseks kasutatakse enamasti kas parameetrilist Pearsoni r või mitteparameetrilist Spearmani ρ korrelatsioonikordajat, vastavalt sellele, milline on arvuliste tunnuste jaotus (normaaljaotus või mittenormaaljaotus) ja millise kujuga on kahe tunnuse vaheline seos (lineaarne või mittelineaarne). Lineaarse ehk sirgjoonelise seose puhul muutuvad mõlema tunnuse väärtused ühtlaselt kindlas suunas kogu skaala ulatuses (joonisel 6.8 paneelid *a* ja *b*), mittelineaarse seose puhul aga võib ühes skaala otsas olla seos tugevam, teises nõrgem (paneelid *c* ja *d*). Mõlemad korrelatsioonikordajad eeldavad siiski, et seos kahe arvulise tunnuse vahel on monotoonne ehk ühesuunaline, mistõttu pole need tähenduslikud näiteks juhul, kui mingis skaala ulatuses on seos negatiivne ehk kahanev, teises aga positiivne ehk kasvav (paneel *e*).



Joonis 6.8. Kahe arvulise tunnuse vahelise seose võimalikke kujusid

Eelmises alapeatükis saime teada, et vanuserühmade vahel ei ole kõnetempos olulisi erinevusi. Vaatame aga nüüd, kuidas mõjutab kõnetempot vanus pideva, mitte rühmitava tunnusena.

```
> gg_point(data = fonkorp2, x = vanus, y = konetempo,
      y_label = "Kõnetempo",
      alpha = 0.3, # punktide läbipaistvus
      size = 2, # punktide suurus
      mode = light_mode_t()) +
    geom_smooth(method = "loess", color = "orange3") # oranž kõver trendijoon
```



Näeme jooniselt, et tegemist on suhteliselt lineaarse monotoonse seosega, ent kuna punktiparvest läbi tõmmatud trendijoon on peaaegu horisontaalne, ei tundu vanuse ja kõnetempo vahel väga tugevat seost olevat. Leiame selle kinnituseks või ümberlükkamiseks ka korrelatsioonikordaja.

```
> cor(fonkorp2$vanus, fonkorp2$konetempo, method = "pearson") # Pearsoni
korrelatsioonikordaja r
[1] -0.1005938
```

Korrelatsioonikordaja on arvuline väärtus, mis näitab korrelatsiooni ehk seose 1) olemasolu, 2) tugevust ja 3) suunda, aga mitte statistilist olulisust. Kordaja väärtus jääb -1 ja 1 vahele. Kui kordaja väärtus on 0, siis seos puudub; kui kordaja väärtus on 1, siis on tegemist ideaalse positiivse seosega: iga ühikulise kasvu kohta ühes tunnuses toimub kindla suurusega kasv ka teises tunnuses; kui kordaja väärtus on -1, on tegemist ideaalse negatiivse seosega; iga ühikulise kasvu kohta ühes tunnuses toimub kindla suurusega kahanemine teises tunnuses. Selliseid ideaalseid korrelatsioone tavaliselt keeleandmetes ei kohta. Seose tugevuse hindamisel lähtutakse sotsiaal- ja humanitaarteadustes sageli taaskord Coheni (1988) skaalast (tabel 6.10).

Tabel 6.10. Pearsoni korrelatsioonikordaja *r* tõlgendamise skaala

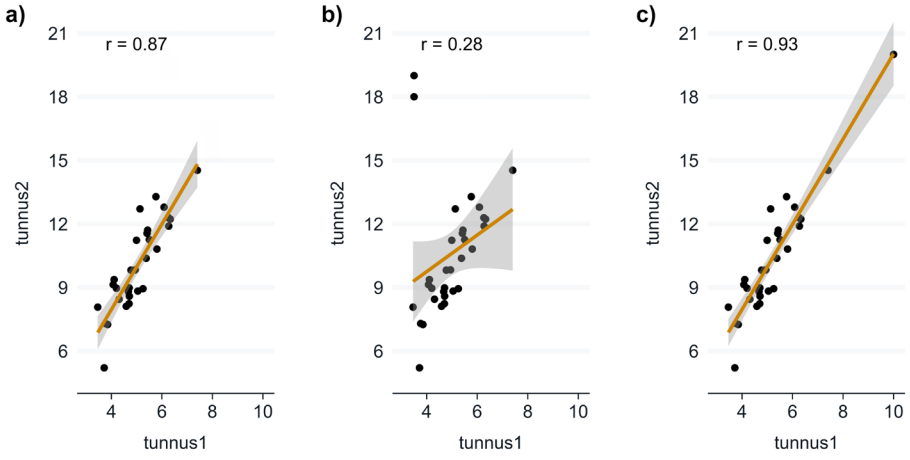
$\|r\| < 0{,}1$	väga nõrk
$0{,}1 \leq \|r\| < 0{,}3$	nõrk
$0{,}3 \leq \|r\| < 0{,}5$	keskmine
$0{,}5 \leq \|r\| < 0{,}7$	tugev
$0{,}7 \leq \|r\|$	väga tugev

Vanuse ja kõnetempo vahel on niisiis (väga) nõrk negatiivne seos: kui vanus kasvab, siis kõnetempo õige pisut aeglustub.

Pearsoni kordaja *r* sobib hästi juhul, kui mõlemad arvulised tunnused on normaaljaotusega ning kui nendevaheline seos on sirgjooneline ehk lineaarne. Jooniselt juba nägime, et seos vanuse ja kõnetempo vahel oli lineaarne. Vaatame nüüd ka tunnuste normaaljaotust.

```
> shapiro.test(fonkorp2$konetempo) # kõnetempo normaaljaotuse kontrollimine

        Shapiro-Wilk normality test

data:  fonkorp2$konetempo
W = 0.99362, p-value = 0.6928 # p > 0.05 (normaaljaotus)

> shapiro.test(fonkorp2$vanus) # vanuse normaaljaotuse kontrollimine

        Shapiro-Wilk normality test

data:  fonkorp2$vanus
W = 0.92688, p-value = 2.197e-07 # p < 0.05 (ei ole normaaljaotus)
```

Jällegi ei ole normaaljaotuse nõue nii range, kui valim on suurem, ent Pearsoni kordaja on **väga tundlik ebaharilike väärtuste suhtes**. Seda seetõttu, et kordaja leidmiseks püütakse läbi punktide tõmmata selline sirge, mis läbiks kõiki punkte võimalikult lähedalt, aga jääks samal ajal sirgeks. Sirge väljendab sisuliselt ennustusi: kui on teada ühe tunnuse väärtus, siis sirge pealt leiaksime sellele vastava teise tunnuse ennustatud väärtuse. Kui mõni vaatlus on aga teistest väga erinev, mõjutab see tugevalt seda, millise kaldega sirge on tõmmatud (joonis 6.9).

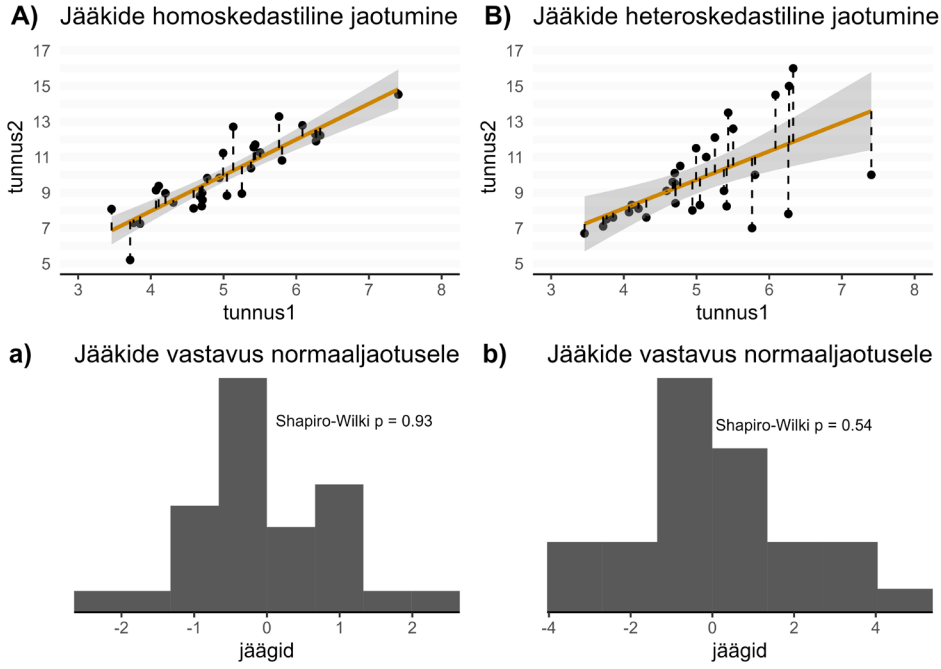


Joonis 6.9. Korrelatsioonikordaja muutus vastavalt erandlike koosesinemiste (*b*) või erandlikult suurte või väikeste vaatluste (*c*) olemasolule

Jooniselt 6.9 näeme, et kui lisame paneelil *a* kuvatud 30-le juhuslikult genereeritud vaatlusele kaks vaatlust, mille väärtuste koosesinemine on teistega võrreldes erandlik (paneel *b*), muutub läbi punktiparve tõmmatud sirge kalle oluliselt laugemaks

ning korrelatsioon muutub palju nõrgemaks. Lisades aga vaatlustele ebaharilikult suure *tunnus1* ja *tunnus2* väärtustega vaatluse (paneel *c*), mis samas sobitub hästi trendijoonele, sirge kalle väga palju ei muutu, ent korrelatsioon läheb tugevamaks. Koondmõiste **erindid** alla loeme tavaliselt ühe arvulise tunnuse vaatlemisel selle ebaharilikult suured/väikesed väärtused; kahe arvulise tunnuse vaatlemisel aga võime erinditeks pidada ka erandlikke tunnusekombinatsioone.

Tunnuste enda vastavusest normaaljaotusele on seega olulisem, et andmetes ei oleks erindeid ning et üksikute vaatluste kaugused punktidest läbi tõmmatud sirgest (neid erinevusi nimetatakse **jääkideks**) oleksid normaaljaotusega ega koonduks ühes sirge lõigus sirgele oluliselt lähemale kui mõnes teises lõigus (joonis 6.10). Vastasel juhul saame küll korrelatsioonikordajat arvutada, ent see ei ole kuigi tähenduslik, kuna iseloomustab kahe tunnuse vahelist seost heal juhul ainult väikese osa vaatluste jaoks ning viga, mis sirge ennustusega kaasneb, ei oleks täiesti juhuslik.

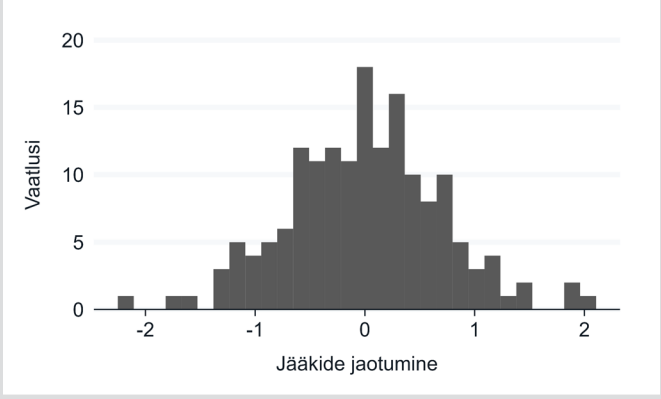


Joonis 6.10. Juhuslikult, homoskedastiliselt jaotunud jäägid (*A*), mustriliselt, heteroskedastiliselt jaotunud jäägid (*B*) ning vastavate jääkide vastavus normaaljaotusele (*a*, *b*)

Joonisel 6.10 on vasakpoolsetel paneelidel (*A* ja *a*) kuvatud juhuslikult jaotunud jäägid ning parempoolsetel (*B* ja *b*) ebaühtlaselt, mustriliselt jaotunud jäägid.

Ühtlaselt ja juhuslikult (ehk **homoskedastiliselt**) jaotunud jääkide puhul ei sõltu vaatluse ühe arvulise tunnuse väärtuse erinevus trendijoonest sellest, milline on teise arvulise tunnuse väärtus. Ebaühtlaselt ja mustriliselt (ehk heteroskedastiliselt) jaotunud jääkide puhul aga sõltub, näiteks joonise 6.10 paneelil *B* lähevad jäägid seda suuremaks, mida suuremaks läheb *tunnus1* väärtus. Kahe tunnuse vahelist seost kujutav trendijoon kirjeldab nõnda hästi vaatlusi, mille *tunnus1* väärtus on väike, aga halvasti neid, mille *tunnus1* väärtus on suur. Mõlemal puhul näeme aga, et jäägid on normaaljaotusega (histogrammid on enam-vähem sümmeetrilised ning Shapiro-Wilki testi p-väärtused on > 0,05). Vaatame nüüd ka kõnetempo ja vanuse vahelise suhte jääke. Jääkide leidmiseks peame tegema õigupoolest esmalt lineaarse regressioonimudeli, millest on lähemalt juttu alapeatükis 6.2.2.1. Homoskedastilisuse testimiseks kasutame paketti `performance` (Lüdecke jt 2021).

```
> konet_van.lm <- lm(konetempo ~ vanus, data = fonkorp2) # teeme lineaarse
mudeli
> konet_van_jaagid <- data.frame(jaagid = residuals(konet_van.lm)) # leiame
mudeli jäägid
> # vaatame histogrammil jääkide jaotumist
> gg_histogram(data = konet_van_jaagid, x = jaagid,
               x_label = "Jääkide jaotumine",
               y_label = "Vaatlusi",
               mode = light_mode_t())
```




```
> shapiro.test(konet_van_jaagid$jaagid) # testime jääkide normaaljaotust

        Shapiro-Wilk normality test

data:  konet_van_jaagid$jaagid
W = 0.99573, p-value = 0.9221 # p > 0.05 (on normaaljaotus)

> install.packages("performance") # installime paketi performance
> library(performance) # laadime paketi performance funktsioonid
> check_heteroscedasticity(konet_van.lm) # kontrollime jääkide
homoskedastilisust
OK: Error variance appears to be homoscedastic (p = 0.455).
```

Näeme, et võime tõepoolest Pearsoni kordajat kasutada, kuna 1) seos vanuse ja kõnetempo vahel on enam-vähem monotoonne ja lineaarne (ehkki nõrk), 2) jäägid on normaaljaotusega ja 3) jäägid on homoskedastiliselt jaotunud. Andmetes esineb küll meeskõnelejate hulgas üksikuid erandlikult suuri või väikesi väärtusi (vt joonist 6.1), mis võivad kordaja väärtust mõjutada. Vahel võib olla mõistlik sellised vaatlused andmete hulgast eemaldada, eriti kui tegemist on näiteks mõõtmisvigadega. Alati ei ole see aga põhjendatud, kuna erandlikud väärtused võivad olla ka loomulik osa varieerumisest ning nende eemaldamine teeks valimi populatsiooniga võrreldes kunstlikult homogeensemaks. Praegusel juhul ei ole head põhjust kiireid ja aeglaseid kõnelejaid andmetest välja jätta, ent võime võrrelda korrelatsioonikordaja väärtusi koos erinditega ja ilma erinditeta ning raporteerida mõlemat väärtust. Alloleva koodiploki väljundist näeme, et kui erindid andmestikust välja jätta, läheb korrelatsioon kahe tunnuse vahel veel nõrgemaks.

```
> out <- boxplot.stats(fonkorp2$konetempo)$out # leiame erandlikud väärtused
> fonkorp2_erinditeta <- fonkorp2[!fonkorp2$konetempo %in% out,] # teeme uue
andmestiku ilma erinditeta
> cor(fonkorp2_erinditeta$vanus, fonkorp2_erinditeta$konetempo, method =
"pearson") # leiame korrelatsioonikordaja ilma erinditeta andmestikus
[1] -0.07053582
```

Kui parameetriliste kordajate ja testide eeldused on rikutud, võib kasutada mitteparameetrilist kordajat või testi. **Spearmani astakkorrelatsioonikordaja *ρ*** sobibki juhtudel, kui arvuliste tunnuste jaotus on ebasümmeetriline, kui seos on küll ühesuunaline, aga mitte lineaarne, või kui andmetes on erindeid, mida ei saa eemaldada. Spearmani kordaja on mitteparameetriline seosekordaja, kuna põhineb sarnaselt U-testiga tegelike arvuliste väärtuste asemel järjestatud arvurea astakutel ega eelda, et tunnused või jäägid oleksid populatsioonis normaaljaotusega. Seetõttu sobib see ka järjestusskaala tunnuste vahelise korrelatsiooni hindamiseks. Hindame nüüd kõnetempo ja vanuse seost mitteparameetrilise Spearmani kordajaga.

```
> cor(fonkorp2$vanus, fonkorp2$konetempo, method = "spearman") # leiame
Spearmani korrelatsioonikordaja
[1] -0.1000425
```

Ehkki Spearmani *ρ* kui mitteparameetriline kordaja tundub jaotuste kohta käivate eelduste puudumise tõttu justkui turvalisem valik, jääb selle väärtus sageli Pearsoni *r* seosekordajast väiksemaks. Nõnda on eelduste täitmisel ja lineaarse seose puhul Pearsoni kordaja täpsem ja tugevam seose tugevuse mõõdik.

Nii parameetriline Pearsoni kordaja kui ka mitteparameetriline Spearmani kordaja näitavad küll ära seose tugevuse ja suuna (seos on kas positiivne või

negatiivne), ent ei ütle, kas see seos on ka statistiliselt oluline. Seda, kas korrelatsioonikordaja väärtus on nullist **statistiliselt oluliselt** erinev, saab kontrollida **korrelatsioonitestiga**. Korrelatsioonitesti nullhüpotees on, et korrelatsioonikordaja väärtus on 0 ehk seos kahe tunnuse vahel puudub.

```
> cor.test(fonkorp2$vanus, fonkorp2$konetempo, method = "pearson") # testime
Pearsoni kordaja olulisust

        Pearson's product-moment correlation

data:  fonkorp2$vanus and fonkorp2$konetempo
t = -1.2869, df = 162, p-value = 0.2 # p > 0.05 (ei ole statistiliselt
oluline)
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.24998984  0.05348055
sample estimates:
       cor
-0.1005938

> cor.test(fonkorp2$vanus, fonkorp2$konetempo, method = "spearman") # testime
Spearmani kordaja olulisust

        Spearman's rank correlation rho

data:  fonkorp2$vanus and fonkorp2$konetempo
S = 808674, p-value = 0.2025 # p > 0.05 (ei ole statistiliselt oluline)
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1000425

Warning message:
In cor.test.default(fonkorp2$vanus, fonkorp2$konetempo, method = "spearman") :
  Cannot compute exact p-value with ties
```

Kui saame Spearmani kordaja olulisust testides ülaloleva hoiatuse, tähendab see seda, et emb-kumb arvuline tunnus sisaldab kaht või enamat samasuguse väärtusega vaatlust (näiteks on meil mitu 22-aastast kõnelejat) ning nende järjestamisel tekivad viigiseisud ja astakud lähevad jagamisele. Näiteks kuuele kõnelejale, kelle vanused on 20, 20, 21, 22, 22, 22, peaksime määrama astakud 1,5, 1,5 (sest (1 + 2) / 2 = 1,5), 3, 5, 5, 5 (sest (4 + 5 + 6) / 3 = 5). See omakorda ei võimalda p-väärtust välja arvutada täpselt, vaid lähendusmeetodiga, mis ei tähenda aga, et test ei oleks usaldusväärne.

Korrelatsioonitesti raporteerimisel on heaks tavaks raporteerida nii seose suund, tugevus kui ka olulisus, samuti statistiku väärtus ja vabadusastmete arv. Näiteks „Pearsoni korrelatsioonikordaja põhjal on kõnetempo ja vanuse vahel negatiivne, ent nõrk ja statistiliselt ebaoluline seos (r = –0,1, 95% usaldusvahemik [–0,25, 0,05], $t(162)$ = –1,29, p = 0,2)“.

6.2.2. Mitmetunnuseline seoste analüüs: statistilised mudelid

Statistilised mudelid on kasulikud eeskätt juhul, kui tahame uurida ja kirjeldada enam kui kahe tunnuse vahelisi seoseid. Need laiendavad statistilise analüüsi võimalusi, võimaldades uuritava tunnuse väärtuste seletamisel võtta korraga arvesse mitme lingvistilise, kontekstuaalse või sotsiodemograafilise tunnuse mõjusid. Siinses alapeatükis räägime põgusalt just sellistest mudelitest, mis kirjeldavad **ühe uuritava tunnuse** väärtuste muutumist vastavalt **mitme seletava tunnuse** väärtuste kombinatsioonile. Kirjeldame vaid paari tavalisemat statistiliste mudelite klassi, mis on ka Eesti keeleteaduses laiemalt levinud. Ehkki sobiva mudeli valimine ja tõlgendamine on seda hõlpsam, mida rohkem erinevate mudelite tööpõhimõtteid ja eeldusi mõista, on nende rakendamine ja uurimistöös kasutamine võimalik ka ilma nende aluseks olevatesse matemaatilistesse seaduspäradesse ja algoritmidesse süübimata. Siin anname mõne näite statistiliste mudelite praktilisest rakendamisest ja tõlgendamisest keeleteaduslikus uurimistöös.

6.2.2.1. Regressioonimudelid

Regressioonimudelid on statistiliste meetodite pere, mis sobitavad uuritava ja seletava(te) tunnus(t)e vahelise suhte kirjeldamiseks andmetele kindla matemaatilise mudeli, üritades sealjuures minimeerida erinevusi mudeli ennustuste ja tegelike valimi andmete vahel. Mudeli põhiliseks väljundiks on koefitsiendid, mis väljendavad kvantitatiivselt iga seletava tunnuse mõju uuritava tunnuse väärtuste varieerumisele.

6.2.2.1.1. Lineaarne regressioon

Arvulise uuritava tunnuse puhul kasutatakse sageli **lineaarset regressiooni** (ingl *linear regression*), mis võimaldab ennustada arvulise uuritava tunnuse keskväärtust sõltuvalt seletavate tunnuste väärtuste kombinatsioonist (vt P. Lippuse näidisuurimust). Näiteks võime uurida lineaarse regressiooni abil, kuidas mõjutavad kõneleja kõnetempot korraga tema vanus ja sugu: kas vanemad inimesed räägivad aeglasemalt kui noored ning kas naiste ja meeste vahel on mingeid erinevusi?

Kõige lihtsamal kujul esineb lineaarse regressiooni mudel kujul $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, kus Y tähistab uuritava tunnuse keskväärtust, X_1 ühe seletava tunnuse väärtust, β_0 on **vabaliige** (ingl *intercept*), β_1 on regressioonikordaja ehk **kalle** (ingl *slope*) ning ε on mudeli juhuslik viga. Vabaliige väljendab uuritava tunnuse väärtust seletava tunnuse **referentsväärtuse** korral. Arvulistel seletavatel tunnustel on referentsväärtuseks tavaliselt 0, kategoorilistel seletavatel tunnustel on referentsväärtuseks ehk baastasemeks tähestikulises järjekorras kõige esimene väärtus. Vabaliige on mudeli ennustuste alguspunkt. Kalle väljendab seda, kui palju kasvab või kahaneb uuritava tunnuse ennustatud keskväärtus iga ühikulise muutuse kohta seletava tunnuse väärtuses. Sellise mudeli graafik on sirge joon (sellest ka nimetus *lineaarne*).

Uurime näiteks lineaarse regressioonimudeli abil, kuidas täiskasvanud inimeste kõnetempo (uuritav tunnus ehk *Y*) sõltub vanusest (seletav tunnus ehk *X*).

```
> mudel1.lm <- lm(konetempo ~ vanus, data = fonkorp2) # teeme lineaarse
mudeli, mida R-is saab kirja panna kujul y ~ x
> summary(mudel1.lm) # vaatame mudeli väljundit

Call:
lm(formula = konetempo ~ vanus, data = fonkorp2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.11083 -0.45392 -0.00184  0.42351  2.10214

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  4.941324   0.166863  29.613 <0.0000000000000002 ***
vanus       -0.005274   0.004098  -1.287                 0.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7112 on 162 degrees of freedom
Multiple R-squared:  0.01012, Adjusted R-squared:  0.004009
F-statistic: 1.656 on 1 and 162 DF,  p-value: 0.2
```

Mudeli väljundi **koefitsientide tabelist** (*Coefficients*) näeme, et vabaliikme (*(Intercept)*) hinnang tulbas *Estimate* on 4,941324, mis tähendab, et kõnelejale, kelle vanus on 0 aastat, ennustab mudel kõnetempoks 4,941 silpi sekundis. Tegemist on niisiis praegusel juhul täiesti hüpoteetilise kontekstiga, mis on lihtsalt mudeli ülejäänud hinnangute lähtekohaks. Vabaliikme p-väärtus tulbas *Pr(>|t|)* on väiksem kui 0,05, mis tähendab, et vabaliikme koefitsient on nullist oluliselt erinev. Järgmiselt koefitsientide tabeli realt näeme, et iga lisanduva aasta kohta vanuse tunnuses kahaneb kõnetempo keskmiselt 0,005274 silbi võrra sekundis. 1-aastased kõneleksid mudeli ennustuste kohaselt seega 4,941 – 0,005×1 = 4,936 silpi sekundis, 2-aastased 4,941 – 0,005×2 = 4,931 silpi sekundis jne. Kõnetempo muutumist kirjeldab selle mudeli järgi seega kõige paremini võrrand

$$kõnetempo = 4{,}941324 - 0{,}005274 \times vanus$$

Vanuse mõju aga ei ole statistiliselt oluline, kuna selle p-väärtuse tulbas on 0,2, mis on seatud olulisuse nivoost (0,05) kõrgem ega luba seega hüljata nullhüpoteesi, mille kohaselt vanuse koefitsient populatsioonis ei erine nullist ja valimist nähtuv võimalik seos kõnetempo ja vanuse vahel on vaid juhuslik.

Lineaarse regressioonimudeli väljundi viimaselt realt leiame ka mudeli kui terviku p-väärtuse, mis ainult ühe seletava tunnusega mudeli puhul on sama, mis koefitsiendi p-väärtus. Eelviimasel real on esitatud seletatud varieerumise osakaal

(ingl *multiple R-squared*) ning sama näitaja, mis on kohandatud mudelisse kaasatud tunnuste arvu suhtes (ingl *adjusted R-squared*). Viimane karistab mudelit liigse ja ebavajaliku komplekssuse eest ning on seega kasulikum mudeli headuse mõõdik. Vanus seletab kõnetempo varieerumisest nõnda vaid 0,4% ning mudel ei kirjelda tegelikult statistiliselt olulisi seoseid.

Kui seletavaks tunnuseks on ainult üks arvuline tunnus, leiame lineaarse regressiooniga tegelikult lihtsalt Pearsoni korrelatsioonikordaja ja testime selle statistilist olulisust (võrdle mudeli ja eelmises alapeatükis tehtud Pearsoni korrelatsioonitesti t-statistikut ja p-väärtust). Kui seletavaks tunnuseks on ainult üks rühmitav kategooriline tunnus, teeb lineaarne regressioon sisuliselt sama asja, mida t-test või ANOVA, sõltuvalt kategoorilise tunnuse erinevate väärtuste/kategooriate arvust. Näiteks võime uurida, kuidas kõnetempo sõltub kõneleja soost.

```
> mudel2.lm <- lm(konetempo ~ sugu, data = fonkorp2) # teeme lineaarse mudeli
> summary(mudel2.lm) # vaatame mudeli väljundit

Call:
lm(formula = konetempo ~ sugu, data = fonkorp2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02416 -0.47602  0.01983  0.45911  2.20837

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  4.54351    0.08047  56.461 < 0.0000000000000002 ***
suguN        0.35592    0.10863   3.277              0.00129 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6922 on 162 degrees of freedom
Multiple R-squared:  0.06215, Adjusted R-squared:  0.05636
F-statistic: 10.74 on 1 and 162 DF,  p-value: 0.001286
```

Mudeli väljundi koefitsientide tabelis näeme sedapuhku mitte lihtsalt tunnuse nime *sugu*, vaid *suguN*. Seda seetõttu, et kategooriliste seletavate tunnuste puhul on vabaliikme kontekstis üks kategoorilise tunnuse kahest või enamast tasemest. Sool on meie andmestikus vaid kaks võimalikku väärtust: *M* ja *N*. Vabaliikme konteksti valitakse neist vaikimisi tähestikulises järjekorras esimene, antud juhul *M*. Vabaliige väljendab nõnda meeskõnelejate ennustatud keskmist kõnetempot (4,54 silpi sekundis) ja kalle ehk soo regressioonikordaja real *suguN* muutust kõnetempos, kui vaatleme meeste asemel naisi: naised räägivad meestest keskmiselt 0,35592 silpi sekundis kiiremalt ehk 4,54351 + 0,35592 = 4,89943 silpi sekundis. p-väärtuse tulbas on väärtus 0,00129, mis on väiksem kui olulisuse nivoo 0,05. Seega on mees- ja naiskõnelejate kõnetempode erinevus statistiliselt oluline. Kui võrdleme koefitsientide tabelis real *suguN* t-statistikut ja p-väärtust varem tehtud

t-testi väljundiga (alapeatükk 6.2.1.2), leiame ühesugused väärtused. Viimaselt realt näeme, et ka mudel tervikuna on statistiliselt oluline (mudeli p-väärtus on 0,001), ning eelviimane rida ütleb, et kõneleja sooga mudel selgitab kõnetempo varieerumisest umbes 5,6%.

Nagu nägime, võiksime ühe seletava tunnusega mudeli asemel teha ka Pearsoni korrelatsioonitesti või t-testi. Lineaarsel mudelil kui parameetrilisel mudelil on parameetriliste testidega samad eeldused, mida me siin ruumi kokkuhoiu mõttes uuesti ei hinda (vt alapeatükk 6.2.1). Mudelid on aga kasulikumad just siis, kui soovime vaadelda mitme seletava tunnuse mõju korraga. Seetõttu lisame nüüd mudelisse korraga nii vanuse kui ka soo. Seletavate tunnuste peamõjud eraldame mudelis plussmärgiga.

```
> mudel3.lm <- lm(konetempo ~ vanus + sugu, data = fonkorp2) # teeme lineaarse
mudeli
> summary(mudel3.lm) # vaatame mudeli väljundit

Call:
lm(formula = konetempo ~ vanus + sugu, data = fonkorp2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90002 -0.46358  0.01197  0.41593  2.11045

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  4.764150   0.170273  27.979 < 0.0000000000000002 ***
vanus       -0.005844   0.003978  -1.469              0.14382
suguN        0.362736   0.108342   3.348              0.00101 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6898 on 161 degrees of freedom
Multiple R-squared:  0.07455,    Adjusted R-squared:  0.06306
F-statistic: 6.485 on 2 and 161 DF,  p-value: 0.001956
```

Kahe seletava tunnusega mudeli vabaliige väljendab nüüd 0-aastaste meeskõnelejate ennustatud keskmist kõnetempot, mis on 4,764 silpi sekundis. Mõlemad seletava tunnuse koefitsiendid tabelis (*vanus* ja *suguN*) väljendavad muutust uuritavas tunnuses, kui teise seletava tunnuse tase hoitakse selle referentstasemel. Näeme, et mudel on tervikuna oluline (p = 0,001956), ent seletavatest tunnustest on mudelis statistiliselt oluline vaid mees- ja naiskõnelejate erinevus. Vanus kõnetempot ennustada ei aita. Mudel seletab umbes 6,3% kõnetempo varieerumisest.

Tuleb silmas pidada, et kuigi rohkemate seletavate tunnustega mudel võib küll tõsta veidi mudeliga seletatud varieerumise protsenti, ei ole mudeli keerukamaks tegemine seletavate tunnuste mõju suurust arvestades alati õigustatud. Põhimõtteliselt võiksime kirjeldada ju kõik oma andmestiku kasutusjuhud seletavate tunnuste kaudu ära kuni kõige pisema detailini, ent kõiki neid nüansse mudelisse

kaasates riskime mudeli **ülesobitamisega** (ingl *overfitting*): mudel suudab suurepäraselt ära kirjeldada selle, mis toimub meie valimis, aga ei pruugi sugugi sobida mõne teise valimi kirjeldamiseks ega iseloomusta seega tõenäoliselt väga hästi ka seda, mis toimub tegelikult populatsioonis. Üksteisest ühe seletava tunnuse võrra erinevaid mudeleid saame R-is võrrelda funktsiooniga *anova()*.

```
> anova(mudel1.lm, mudel3.lm) # kas keerukam mudel "mudel3" on parem kui
"mudel1"?
Analysis of Variance Table

Model 1: konetempo ~ vanus
Model 2: konetempo ~ vanus + sugu
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    162 81.938
2    161 76.605  1    5.3335 11.209 0.001013 ** # p < 0.05 (kolmas mudel on
esimesest oluliselt parem, komplekssem mudel on õigustatud)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(mudel2.lm, mudel3.lm) # kas keerukam mudel3 "mudel3" on parem kui
"mudel2"?
Analysis of Variance Table

Model 1: konetempo ~ sugu
Model 2: konetempo ~ vanus + sugu
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    162 77.631
2    161 76.605  1    1.0266 2.1576 0.1438 # p > 0.05 (kolmas mudel ei ole
teisest oluliselt parem, komplekssem mudel ei ole õigustatud)
```

Kui võrdleme vanuse ja sooga mudelit esmalt ainult vanusega mudeliga, näeme, et soo lisamine vanusele aitab kõnetempot oluliselt paremini ennustada kui ainult vanus. Vanuse ja sooga mudel ei ole aga oluliselt parem ainult sooga mudelist, seega võiksime eelistada lihtsamat, vähemate tunnustega mudelit, kus kõnetempot aitab ennustada ainult sugu (*mudel2.lm*). See aga, kas hoida statistiliselt ebaolulisi tunnuseid mudelis või mitte, sõltub valitud mudeldamisstrateegiast. **Konfirmatoorse ehk kinnitava mudeldamisstrateegia** puhul on meil varasemate uurimuste või muude tähelepanekute põhjal valitud seletavad tunnused, mille mõju kohta uuritavale tunnusele on meil kindlad hüpoteesid. Seega testime regressioonimudeliga enda teoreetilist mudelit ning ka tunnused, mis osutuvad ebaoluliseks, annavad selle teoreetilise mudeli kohta olulist informatsiooni. **Eksploratiivse ehk uuriva mudeldamisstrateegia** puhul meil aga selged hüpoteesid puuduvad. Sel juhul püüame leida optimaalset mudelit, mis seletaks uuritava tunnuse varieerumist kõige paremini, aga samas võimalikult lihtsalt. Eksploratiivse mudeldamisstrateegia puhul on mõistlik ebaolulised tunnused mudelist välja jätta.

Regressioonimudelite tulemusi raporteeritakse uurimustes eri viisidel. Üsna levinud on ülal nähtud koefitsientide tabelite esitamine koos mudeli

regressioonikordajate, nende standardvigade ja p-väärtuste ning mudeli üldise headuse ja statistilise olulisuse näitajatega (tabel 6.11).

Tabel 6.11. Keskmist kõnetempot ennustava lineaarse regressioonimudeli parameetrite hinnangud (mudeli kohandatud $R^2 = 0{,}063$, $p = 0{,}002$)

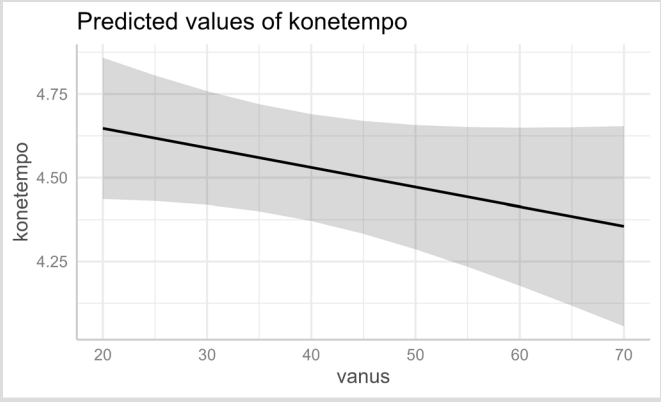
	Hinnang	Standardviga	t-statistik	p-väärtus
(Vabaliige)	4,7642	0,1703	27,979	< 0,001
vanus	–0,0058	0,0040	–1,469	0,144
sugu [N]	0,3627	0,1083	3,348	0,001

Teine viis on raporteerida mudeli parameetreid tekstina. R-is on olemas levinumate testide ja mudelite kohta teksti koostamiseks ka oma pakett `report` (Makowski jt 2023), mis küll paraku kuvab tulemusi vaid inglise keeles.

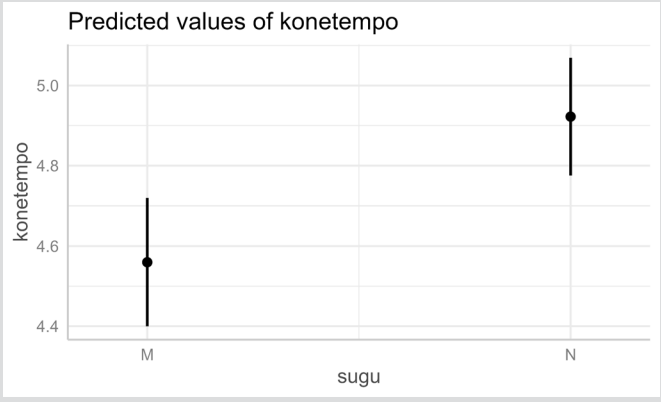
```
> install.packages("report") # installime paketi report
> library(report) # laadime paketi report funktsioonid
> report_text(mudel3.lm)
We fitted a linear model (estimated using OLS) to predict konetempo with vanus
and sugu (formula: konetempo ~ vanus + sugu). The model
explains a statistically significant and weak proportion of variance (R2 =
0.07, F(2, 161) = 6.48, p = 0.002, adj. R2 = 0.06). The model's
intercept, corresponding to vanus = 0 and sugu = M, is at 4.76 (95% CI [4.43,
5.10], t(161) = 27.98, p < .001). Within this model:

  - The effect of vanus is statistically non-significant and negative (beta =
-5.84e-03, 95% CI [-0.01, 2.01e-03], t(161) = -1.47, p = 0.144;
Std. beta = -0.11, 95% CI [-0.26, 0.04])
  - The effect of sugu [N] is statistically significant and positive (beta =
0.36, 95% CI [0.15, 0.58], t(161) = 3.35, p = 0.001; Std. beta =
0.51, 95% CI [0.21, 0.81])

Standardized parameters were obtained by fitting the model on a standardized
version of the dataset. 95% Confidence Intervals (CIs) and
p-values were computed using a Wald t-distribution approximation.
```

Kolmas ja ehk veidi lugejasõbralikum viis on kuvada seletavate tunnuste (koos) mõjusid joonistel, näiteks kasutades paketi `ggeffects` (Lüdecke 2018) võimalusi. Jooniste y-teljel on vaikimisi kuvatud uuritava tunnuse ennustatud väärtuste skaala ning x-teljel seletava tunnuse erinevad väärtused.

```
> install.packages("ggeffects") # installime paketi ggeffects
> library(ggeffects) # laadime paketi ggeffects funktsioonid
> plot(ggpredict(mudel3.lm, terms = "vanus")) # joonistame vanuse mõju graafiku
```




```
> plot(ggpredict(mudel3.lm, terms = "sugu")) # joonistame soo mõju graafiku
```



6.2.2.1.2. Logistiline regressioon

Uuritava kategoorilise tunnuse puhul on keeleteaduses levinud **logistilise regressiooni** mudelid (ingl *logistic regression*). Sõltuvalt sellest, kas uuritaval tunnusel on kaks või enam väärtust/kategooriat, nimetatakse logistilise regressiooni mudeleid kas binomiaalseks (ingl *binomial*) või multinomiaalseks (ingl *multinomial*). Erinevalt lineaarsest regressioonist ei ennusta logistilise regressiooni mudel uuritava tunnuse keskväärtust samades ühikutes, milles uuritav tunnus on mõõdetud. Logistiline regressioon ennustab hoopis ühe uuritava tunnuse väärtuse/kategooria esinemise **tõenäosust** erinevates kontekstides vastavalt sellele, millised on seletavate tunnuste väärtused. Sealjuures on tõenäosused väljendatud **logaritmitud**

šanssidena. Šanss väljendab mingi sündmuse toimumise ja mittetoimumise tõenäosuste suhet. Näiteks kui sündmuse toimumise tõenäosus on 0,8, on selle mittetoimumise tõenäosus järelikult 0,2 ning sündmuse toimumise šanss 0,8 / 0,2 = 4. Mida suurem on sündmuse toimumise tõenäosus, seda suurem on ka sündmuse toimumise šanss. Kui aga sündmuse toimumise tõenäosus on väiksem kui selle mittetoimumise tõenäosus, jääb šanss alla ühe, ent ei lange kunagi alla nulli: nt 0,2 / 0,8 = 0,25; 0,0000000000001 / 0,9999999999999 = 0,0000000000001; 0 / 1 = 0. Šansside logaritmimine võimaldab altotsast nulliga piiratud šansse väljendada pideval skaalal miinus lõpmatusest pluss lõpmatuseni (vt tabel 6.12).

Tabel 6.12. Logaritmitud šansi, šansi ja tõenäosuse suhe

Logaritmitud šanss *log(P)*	**Šanss *P*** $P = \frac{p}{(1-p)}$	**Tõenäosus *p*** $p = \frac{P}{(1+P)}$
– ∞ (miinus lõpmatus)	0,0000000	0
–2,1972246	0,1111111	0,1
–1,3862944	0,2500000	0,2
–0,8472979	0,4285714	0,3
–0,4054651	0,6666667	0,4
0,0000000	**1,0000000**	**0,5**
0,4054651	1,5000000	0,6
0,8472979	2,3333333	0,7
1,3862944	4,0000000	0,8
2,1972246	9,0000000	0,9
∞ (lõpmatus)	∞ (lõpmatus)	1

Sündmuse toimumiseks võime keeleteaduses pidada mingi keelelise konstruktsiooni, sõna või häälduvariandi kasutamist ja mittetoimumiseks seega mingi muu konstruktsiooni, sõna või häälduvariandi kasutamist. Näiteks võib sündmuse toimumine olla tagaeituse kasutamine (*vs.* eeseituse kasutamine), liitmineviku kasutamine (*vs.* lihtmineviku kasutamine), SVO sõnajärje kasutamine (*vs.* muu sõnajärje kasutamine), *h* hääldamine sõna alguses (*vs.* mittehääldamine). Samuti võib sündmuse toimumisena tõlgendada mingi sotsiolingvistilise tunnuse esinemist, näiteks meeskõneleja (*vs.* naiskõneleja), vanem kõneleja (*vs.* noorem kõneleja), madalamalt haritud kõneleja (*vs.* kõrgemalt haritud kõneleja).

Vaatame idaseto eituse andmestiku põhjal, kuidas ennustab regressioonimudel ees- või tagaeituse (nt *ei olõ vs. olõ-õi*) kasutamist vastavalt 1) eitussõnale (oleviku eitussõna *ei* või mineviku eitussõna *es*), 2) eelmisele vestluses kasutatud eituskonstruktsioonile (kas varem viimati kasutatud ees- või tagaeitust), 3) mõlemale korraga. Logistiline regressioonimudel valib **uuritava tunnuse** (praegusel juhul *ASEND*) ühe taseme, mille esinemise tõenäosust see logaritmitud šansi kaudu ennustab. Vaikimisi valitakse **tähestikulises järjekorras tagapool olev tase**. Kui uuritava tunnuse *ASEND* võimalikud väärtused on *eeseitus* ja *tagaeitus*, ennustab mudel järelikult taseme *tagaeitus* esinemise tõenäosust. Alapeatükis 6.1.1 nägime, et tagaeitus on idasetos palju tavalisem kui eestipärane eeseitus, nõnda ennustame siin uuritava tunnuse sagedama klassi esinemise tõenäosust.

Enne mudeli tegemist võiksime aga andmestikust välja jätta vaatlused, kus eelmise eituskonstruktsiooni vormi ei olnud võimalik hinnata (näiteks oli tegemist kõige esimese eituskonstruktsiooniga vestluses). Samuti peame teisendama uuritava kategoorilise tunnuse faktortunnuseks. Faktortunnus on kategooriline tunnus, millel on piiratud arv kindlaks määratud järjekorraga kategooriaid ehk tasemeid.

```
> eitus_alam <- eitus[eitus$EELMINE != 0,] # jätame vaatlusi välja
> eitus_alam$ASEND <- factor(eitus_alam$ASEND) # teisendame kategoorilise
tunnuse faktortunnuseks

> mudel1.glm <- glm(ASEND ~ EITUSSÕNA, data = eitus_alam, family = binomial) #
teeme logistilise regressiooni mudeli
> summary(mudel1.glm) # vaatame mudeli väljundit

Call:
glm(formula = ASEND ~ EITUSSÕNA, family = binomial, data = eitus_alam)

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)    0.72082    0.08708   8.277 < 0.0000000000000002 ***
EITUSSÕNAes    1.02325    0.15767   6.490       0.000000000086 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1187.7  on 1055  degrees of freedom
Residual deviance: 1141.8  on 1054  degrees of freedom
AIC: 1145.8

Number of Fisher Scoring iterations: 4
```

Logistilise regressioonimudeli väljund näeb välja sarnane lineaarse mudeli omale, ent selles ei näidata mudeli üldist statistilist olulisust ega seletatud varieerumise protsenti. Nagu öeldud, erineb ka see, mida väljendatakse koefitsientide tabeli hinnangute (*Estimate*) tulbas. Vabaliige *(Intercept)* väljendab siin tagaeituse esinemise ennustatud **logaritmitud šansse** mingis seletavate tunnuste referentskontekstis.

Nagu lineaarsegi regressiooni puhul on vabaliikme vaikimisi kontekstis arvuliste seletavate tunnuste väärtus 0 ning kategooriliste seletavate tunnuste puhul nende tähestikulises järjekorras esimene tase. Nõnda väljendab vabaliikme väärtus koefitsientide tabeli hinnangute tulbas tagaeituse esinemise logaritmitud šansse juhul, kui eitussõna on oleviku eitussõna *ei*. Eitussõna mõju hinnang real *EITUSSÕNAes* väljendab aga **muutust nendes logaritmitud šanssides**, kui vaatleme oleviku eitussõna asemel hoopis mineviku eitussõna *es*. Logaritmitud šansside väärtus võib varieeruda miinus lõpmatusest lõpmatuseni ning see on teisendatav tõenäosuseks, mis võib varieeruda nullist üheni (vt tabel 6.12). Õigupoolest on aga logistilise regressioonimudeli väljundis tarvis pöörata põhiliselt tähelepanu sellele, kas koefitsiendid on positiivsed või negatiivsed. Positiivsete koefitsientide puhul sündmuse toimumise ehk tagaeituse kasutamise tõenäosus ja šanss kasvavad, negatiivsete puhul kahanevad.

Meie mudelis, kus ainsaks eitussõna asukohta seletavaks tunnuseks on eitussõna vorm, on vabaliikme väärtus 0,72082. See on positiivne arv, mistõttu saame järeldada, et vabaliikme referentskontekstis, kus eitussõna on *ei*, on tagaeituse kasutamine tõenäolisem kui selle mittekasutamine (ehk eeseituse kasutamine). Vabaliikme väärtuse, mis väljendab tegaeituse kasutamise logaritmitud šanssi, võime teisendada ka tavaliseks šansiks ja tõenäosuseks.

```
> exp(0.72082) # teisendame logaritmitud šansi tavaliseks šansiks
[1] 2.056119 # šanss kasutada tagaeitust, kui eitussõna on "ei"

> plogis(0.72082) # teisendame logaritmitud šansi tõenäosuseks
[1] 0.6727876 # tõenäosus kasutada tagaeitust, kui eitussõna on "ei"
```

Seletavate tunnuste koefitsiendid väljendavad **muutust** sündmuse toimumise šanssides võrreldes vabaliikme kontekstiga. *EITUSSÕNAes* koefitsiendi väärtus 1,02325 on samuti positiivne, mis viitab sellele, et juhul, kui eitussõna on mineviku eitussõna *es*, kasvab tagaeituse kasutamise šanss veelgi. Selleks, et teada saada, mis see šanss ja tõenäosus täpselt on, peame liitma seletava tunnuse koefitsiendile ka vabaliikme koefitsiendi.

```
> exp(1.02325) # mitu korda kasvab šanss tagaeitust kasutada, kui eitussõna
on "es"?
[1] 2.782222
> exp(0.72082+1.02325) # milline on šanss tagaeitust kasutada, kui eitussõna
on "es"?
[1] 5.720579
> plogis(0.72082+1.02325) # milline on tõenäosus tagaeitust kasutada, kui
eitussõna on "es"?
[1] 0.8512033
```

Olevikust rääkides on niisiis tagaeituse kasutamise ennustatud tõenäosus 0,67, minevikust rääkides aga koguni 0,85. See erinevus tagaeituse kasutamise tõenäosuses on ka statistiliselt oluline (koefitsiendi *EITUSSÕNAes* p-väärtus on < 0,05).

Teeme nüüd ka mudeli, kus eitussõna asukohta ennustab eitussõna paiknemine eelmises kõneleja kasutatud eituskonstruktsioonis, et testida, kas kõnelejad kalduvad taaskasutama juba kord aktiveeritud grammatilisi struktuure.

```
> mudel2.glm <- glm(ASEND ~ EELMINE, data = eitus_alam, family = binomial) #
teeme logistilise mudeli
> summary(mudel2.glm) # vaatame mudeli väljundit

Call:
glm(formula = ASEND ~ EELMINE, family = binomial, data = eitus_alam)

Coefficients:
             Estimate Std. Error z value               Pr(>|z|)
(Intercept)   0.08829    0.12139   0.727                  0.467
EELMINEtaga   1.49084    0.15406   9.677 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1187.7  on 1055  degrees of freedom
Residual deviance: 1093.7  on 1054  degrees of freedom
AIC: 1097.7

Number of Fisher Scoring iterations: 3
```

Selle mudeli puhul väljendab vabaliige nüüd konteksti, kus kõneleja on eelnevalt kasutanud eeseitust. Vabaliikme koefitsient on endiselt positiivne (ehkki vaid napilt üle nulli), millest võime järeldada, et ka sellises kontekstis kasutatakse pigem tagaeitust. Vabaliikme koefitsient ei ole aga statistiliselt oluline (p-väärtus on 0,467), mis tähendab, et ees- ja tagaeituse kasutamise tõenäosuses ei ole tegelikult sellises kontekstis statistiliselt olulist erinevust. Juhul, kui eelmises eituskonstruktsioonis kasutati aga tagaeitust, kasvab oluliselt ka tagaeituse uuesti kasutamise tõenäosus (koefitsient on 1,49084, p-väärtus < 0,05). Järgmiseks teeme mudeli, kuhu lisame nii eitussõna kui ka eelmise eitussõna asukoha korraga.

```
> mudel3.glm <- glm(ASEND ~ EITUSSÕNA + EELMINE, data = eitus_alam, family =
binomial) # teeme logistilise mudeli
> summary(mudel3.glm) # vaatame mudeli väljundit

Call:
glm(formula = ASEND ~ EITUSSÕNA + EELMINE, family = binomial,
      data = eitus_alam)

Coefficients:
```

```
               Estimate Std. Error z value                Pr(>|z|)
(Intercept)    -0.2327     0.1351  -1.723                   0.085 .
EITUSSÕNAes     0.9566     0.1641   5.831          0.00000000552 ***
EELMINEtaga     1.4428     0.1573   9.171 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1187.7  on 1055  degrees of freedom
Residual deviance: 1057.3  on 1053  degrees of freedom
AIC: 1063.3

Number of Fisher Scoring iterations: 4
```

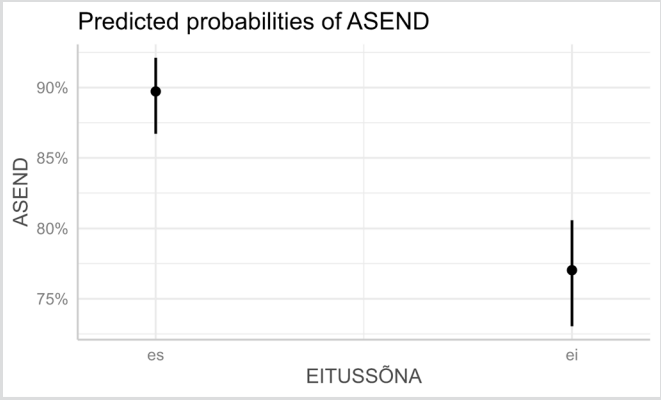
Kahe seletava tunnusega mudelis väljendab vabaliige nüüd logaritmitud šansi kaudu tagaeituse kasutamise tõenäosust kontekstis, kus eitussõna on oleviku eitussõna *ei* ning eelmine kasutatud edasiviidis oli eeseitus. Jällegi on vabaliikme koefitsient statistiliselt ebaoluline, mistõttu võime öelda, et sellises kontekstis on sama suur tõenäosus kasutada kas ees- või tagaeitust. Nagu näeme, tõstavad tagaeituse kasutamise tõenäosust oluliselt nii mineviku eitussõna *es* kasutamine kui ka see, kui eelnevalt kasutatud edasiviidis on tagaeitus juba aktiveeritud.

Võrdleme igaks juhuks keerulisemat, kahe seletava tunnusega mudelit ka lihtsamatega, kus on korraga ainult üks seletav tunnus.

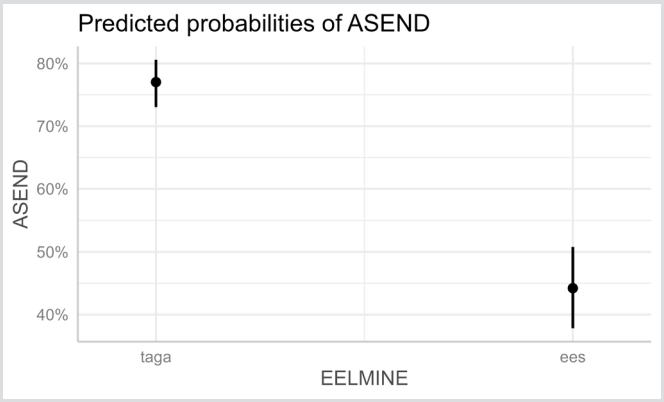
```
> anova(mudel1.glm, mudel3.glm) # kas "mudel3" on parem kui "mudel1"?
Analysis of Deviance Table

Model 1: ASEND ~ EITUSSÕNA
Model 2: ASEND ~ EITUSSÕNA + EELMINE
  Resid. Df Resid. Dev Df Deviance             Pr(>Chi)
1      1054     1141.8
2      1053     1057.3  1   84.533    < 0.00000000000000022 *** # on küll (p
< 0.05)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(mudel2.glm, mudel3.glm) # kas "mudel3" on parem kui "mudel2"?
Analysis of Deviance Table

Model 1: ASEND ~ EELMINE
Model 2: ASEND ~ EITUSSÕNA + EELMINE
  Resid. Df Resid. Dev Df Deviance      Pr(>Chi)
1      1054     1093.7
2      1053     1057.3  1   36.382    0.000000001622 *** # on küll (p < 0.05)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Näeme niisiis, et mudelisse tasub jätta seletavate tunnustena nii eitussõna kui ka eelmise eituskonstruktsiooni sõnajärje. Logistilise regressiooni mudeleid saab raporteerida sarnaselt lineaarsetele mudelitele, näiteks esitades mudeli koefitsientide tabeli, kirjutades tulemused lahti vabatekstina või kasutades jooniseid. Laseme näiteks R-il kirjutada tekstina mudeli ülevaate ning visualiseerime mudelis olevaid mõjusid. Graafikute y-teljel väljendatakse protsentidena tõenäosust kasutada tagaeitust vastavalt x-teljel kuvatud seletava tunnuse erinevatele väärtustele.

```
> library(report) # laadime paketi report funktsioonid
> report_text(mudel3.glm) # raporteerime mudeli tulemused
We fitted a logistic model (estimated using ML) to predict ASEND with EITUSSÕNA
and EELMINE (formula: ASEND ~ EITUSSÕNA + EELMINE). The model's
explanatory power is weak (Tjur's R2 = 0.13). The model's intercept,
corresponding to EITUSSÕNA = ei and EELMINE = ees, is at -0.23 (95% CI
[-0.50, 0.03], p = 0.085). Within this model:

  - The effect of EITUSSÕNA [es] is statistically significant and positive
(beta = 0.96, 95% CI [0.64, 1.28], p < .001; Std. beta = 0.96, 95% CI
[0.64, 1.28])
  - The effect of EELMINE [taga] is statistically significant and positive
(beta = 1.44, 95% CI [1.14, 1.75], p < .001; Std. beta = 1.44, 95% CI
[1.14, 1.75])

Standardized parameters were obtained by fitting the model on a standardized
version of the dataset. 95% Confidence Intervals (CIs) and
p-values were computed using a Wald z-distribution approximation.

> library(ggeffects) # laadime paketi ggeffects funktsioonid
> plot(ggpredict(mudel3.glm, terms = "EITUSSÕNA")) # joonistame eitussõna mõju
graafiku
```




```
> plot(ggpredict(mudel3.glm, terms = "EELMINE")) # joonistame eelmise
eituskonstruktsiooni mõju graafiku
```



Erinevalt lineaarsest regressioonist ei väljasta logistilise regressiooni mudel näitajaid, mis aitaksid hinnata mudeli kui terviku headust. R-i abil mudeli tulemusi ülal tekstina raporteerides väljastatakse küll Tjuri R^2, mida saab kasutada erinevate mudelite omavaheliseks võrdlemiseks. Sagedamini hinnatakse logistilise regressiooni mudelite headust aga klassifitseerimistäpsuse (ingl *classification accuracy*) ja C-indeksi (ingl *concordance index*, vahel viidatakse sisuliselt samale mõõdikule ka lühendiga AUC ehk ingl *area under the ROC curve*) abil. **Klassifitseerimistäpsus** (ka *õigsus*, vt Sügis jt 2024: 116, või *korrektsus*, vt ptk 3.1.3 „Märgenduse täpsuse hindamine“) väljendab mudeli õigete ennustuste osakaalu („Mitmele protsendile vaatlustest ennustab mudel õige uuritava tunnuse klassi?“). **C-indeks** väljendab mudeli eristusvõimet ehk võimet järjestada vaatlusi vastavalt nende ennustatud tõenäosusele olla klassifitseeritud sündmuse toimumiseks. C-indeksit saab leida vaid juhul, kui uuritaval tunnusel on ainult kaks klassi. Mõlemad näitajad varieeruvad nullist üheni. Mida lähemal ühele, seda paremini mudel uuritava tunnuse varieerumist seletab.

```
> # arvutame klassifitseerimistäpsuse
> ennustatud_klassid <- ifelse(mudel3.glm$fitted.values > 0.5, "tagaeitus",
"eeseitus") # leiame mudeli ennustused (kui ennustus on > 0.5, määrame
ennustatud klassiks "tagaeitus", vastasel juhul "eeseitus"
> tegelikud_klassid <- eitus_alam$ASEND # küsime andmestikust uuritava tunnuse
tulpa (tegelikke klasse)
> ennustuste_tabel <- table(ennustatud_klassid, tegelikud_klassid) # teeme
ennustustest ja tegelikest klassidest risttabeli
> sum(diag(ennustuste_tabel))/sum(ennustuste_tabel) # leiame
klassifitseerimistäpsuse ehk õigesti klassifitseeritud juhtude osakaalu
[1] 0.7670455
```

Mudeli klassifitseerimistäpsus on ümardatult 0,77. See tähendab, et mudel ennustab 77%-le vaatlustest õige eitussõna asukoha, aga 23%-l juhtudest ennustab valesti (nt klassifitseerib tagaeituseks mingi kasutusjuhu, mille tegelik klass on „eeseitus", või vastupidi). Kui ennustaksime ilma mis tahes mudeliteta kõikidele vaatlustele alati lihtsalt sagedamat klassi ja klassifitseeriksime kõik andmestiku vaatlused tagaeituseks, oleks meie klassifitseerimistäpsus 75%, kuna 75% kõikidest eituskonstruktsiooni kasutusjuhtudest andmestikus ongi päriselt tagaeituse kasutusjuhud. See aga tähendab, et meie kahe seletava tunnusega mudel parandab meie klassifitseerimisvõimet vaid kahe protsendipunkti võrra, mis ei ole eriti hea tulemus.

```
> # arvutame C-indeksi
> install.packages("pROC") # installime paketi pROC
> library(pROC) # laadime paketi pROC funktsioonid
> auc(eitus_alam$ASEND, mudel3.glm$fitted.values) # leiame C-indeksi/AUC
väärtuse

Setting levels: control = eeseitus, case = tagaeitus
Setting direction: controls < cases
Area under the curve: 0.7138
```

Mudeli C-indeksi väärtus on 0,71, mis väljendab rahuldavat eristusvõimet: kui mudelile anda ette andmestikust juhuslikult valitud eeseituse kasutusjuht ja juhuslikult valitud tagaeituse kasutusjuht, siis 71%-l kõikidest sellistest võimalikest paaridest ennustab mudel tagaeituse kasutusjuhule kõrgema tõenäosuse olla tagaeitus kui eeseituse kasutusjuhule. C-indeksi tõlgendamise skaala (vt Hosmer, Lemeshow & Sturdivant 2013) on esitatud tabelis 6.13.

Tabel 6.13. C-indeksi/AUC tõlgendamise skaala

$C < 0{,}5$	väga halb eristusvõime, mudel ennustab klasse tagurpidi
$C = 0{,}5$	mudel ei suuda klasse eristada
$0{,}5 > C < 0{,}7$	kehv eristusvõime
$0{,}7 \leq C < 0{,}8$	rahuldav eristusvõime
$0{,}8 \leq C < 0{,}9$	hea eristusvõime
$\geq 0{,}9$	suurepärane eristusvõime

6.2.2.1.3. Segamõjudega mudelid

Kõik tavalised regressioonimudelid eeldavad, et vaatlused on üksteisest sõltumatud (nt iga vaatlus on pärit erinevalt kõnelejalt). See eeldus on aga nii kõnetempo kui ka eituse andmestiku puhul rikutud. Kui arvulise uuritava tunnuse puhul võime teatud juhtudel andmeid mingi rühmitava tunnuse (nt kõneleja) põhjal keskmistada, nagu tegime ka ühetunnuselise analüüsi peatükis, siis kategoorilise uuritava tunnusega seda teha ei saa. Siin tulevad appi **segamõjudega mudelid** (ingl *mixed-effects models*). Segamõjudega mudelid on statistilised mudelid, mis sisaldavad nii fikseeritud kui ka juhuslikke mõjusid. **Fikseeritud mõjud** (ingl *fixed effects*) on need seletavad tunnused, mis kehtivad terve populatsiooni jaoks ning neil on andmete kodeerimisskeemis mingid kindlaks määratud väärtused. Uurijat huvitavad enamasti just fikseeritud mõjud, näiteks vanuse ja soo mõju kõnetempole, mida eeldame kehtivat terves populatsioonis, või eitussõna ajavormi ja juba aktiveeritud eituskonstruktsiooni mõju tagaeituse kasutamisele. **Juhuslikud mõjud** (ingl *random effects*) tulenevad valimi võtmise eripärast ning kirjeldavad tegureid, mille alusel vaatlused on andmestikus mingil moel rühmitunud. Juhuslikud mõjud enamasti uurijale omaette huvi ei paku, ent nendega on oluline arvestada. Näiteks kõnetempo andmestikus rühmituvad vaatlused kõnelejati, kuna samalt kõnelejalt võib olla mitu vaatlust ning kui neid vaatlusi on palju, võib kõneleja individuaalne eripära kallutada üldisi hinnanguid sellele, kuidas sugu või vanus kõnetempot mõjutavad. Eitussõna andmestikus on vaatlused lisaks kõnelejale rühmitunud aga ka tegusõnati: teatud sagedased tegusõnad kalduvad kinnistuma ees- või tagaeitusega (nt *ei tiiä*, *olõ-õi*). Seega uurides eitussõna ja eelmise kasutatud eituskonstruktsiooni mõju, tahame, et hinnangud fikseeritud mõjudele populatsioonis ei peegeldaks tegelikult üksikute sagedaste verbide tendentse esineda ees- või tagaeitusega ega üksikute kõnelejate individuaalseid eelistusi.

Juhuslikud mõjud jagatakse omakorda juhuslikeks vabaliikmeteks (ingl *random intercept*) ja juhuslikeks kalleteks (ingl *random slope*). **Juhuslik vabaliige** aitab arvestada lihtsalt sellega, et iga vaatlusi koondava rühma jaoks on vabaliikme hinnang veidi erinev (näiteks iga kõneleja räägib pisut omamoodi). **Juhuslik kalle** (ka *juhuslik tõus*, ingl *random slope*) aga võimaldab populatsiooni tasandi fikseeritud mõjude hindamisel arvestada sellega, et muutus fikseeritud mõjus võib uuritava tunnuse ennustatud väärtust mõjutada erinevates rühmades veidi eri moel (näiteks mingite tegusõnade jaoks võib olla eelnevalt kasutatud eituskonstruktsiooni sõnajärje mõju väiksem või eitussõna ajavormi mõju vastupidine).

Teeme segamõjudega mudelite näitlikustamiseks R-i pakettidega `lme4` (Bates jt 2015) ja `lmerTest` (Kuznetsova, Brockhoff & Christensen 2017) esmalt lineaarse segamõjudega mudeli, kus lisame vanuse ja soo peamõjudele kõnetempot ennustavasse mudelisse ka kõneleja juhusliku vabaliikme.

```
> install.packages("lme4") # installime paketi lme4
> install.packages("lmerTest") # installime paketi lmerTest
> library(lme4) # laadime paketi lme4 funktsioonid
> library(lmerTest) # laadime paketi lmerTest funktsioonid

> mudel1.lmer <- lmer(konetempo ~ vanus + sugu + (1|koneleja), data
= fonkorp2) # teeme lineaarse segamõjudega mudeli kõneleja juhusliku
vabaliikmega
> summary(mudel1.lmer) # vaatame mudeli väljundit

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: konetempo ~ vanus + sugu + (1 | koneleja)
   Data: fonkorp2

REML criterion at convergence: 322.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.82571 -0.35079 -0.03052  0.32057  1.65089

Random effects:
 Groups   Name        Variance Std.Dev.
 koneleja (Intercept) 0.3387   0.5820
 Residual             0.1087   0.3298
Number of obs: 164, groups:  koneleja, 139

Fixed effects:
             Estimate  Std. Error         df  t value             Pr(>|t|)
(Intercept)  4.936830    0.174956 149.308515   28.217 < 0.0000000000000002 ***
vanus       -0.008769    0.004080 151.538533   -2.149              0.03319 *
suguN        0.348772    0.112628 137.717708    3.097              0.00237 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
       (Intr) vanus
vanus -0.885
suguN -0.320 -0.018
```

Juhuslike mõjude tabelis (*Random effects*) väljendab real (*Intercept*) tulba *Variance* väärtus seda, kui palju varieeruvust vabaliikme väärtustes juhusliku mõju rühmades on. Mida suurem on see väärtus, seda tugevamalt on vaatlused juhusliku mõju alusel rühmitunud. Praegusel juhul on kõnelejate individuaalsete eelistustega arvestamine oluline, kuna varieeruvus kõnelejate vahel on üpris suur. Real *Residual* nähtav *Variance* väärtus väljendab aga seda, kui palju varieeruvust jääb veel seletamata pärast seda, kui oleme kõnelejatevaheliste erinevustega arvestanud. Praegusel juhul võib väärtus väljendada näiteks seda, et sama kõneleja võib ka eri vestlustes vastavalt teemadele ja vestluskaaslasele oma kõnetempot varieerida. **Fikseeritud mõjude koefitsientide tabelist** näeme, et kui anname mudelile infot selle kohta, millised vaatlused kõnelejati rühmituvad, muutub võrreldes tavalise lineaarse mudeliga peatükis 6.2.2.1.1 lisaks soole statistiliselt oluliseks ka vanuse efekt, ehkki see on endiselt väga nõrk.

Lõpetuseks teeme ka logistilise segamõjudega mudeli, kus võtame tagaeituse esinemise ennustamisel lisaks eitussõnale ja eelmisele kasutatud eituskonstruktsioonile arvesse ka seda, et eri kõnelejatel ja tegusõnadel võivad olla ees- ja tagaeituse kasutamisel erinevad eelistused (juhuslikud vabaliikmed). Seejärel testime, kas oleks alust arvestada mudelis ka sellega, kuidas eitussõna mõju tagaeituse kasutusele võib varieeruda vastavalt kõneleja individuaalsetele eelistustele (juhuslik kalle).

```
> mudel1.glmer <- glmer(ASEND ~ EITUSSÕNA + EELMINE + (1|LEMMA) +
(1|KÕNELEJA), data = eitus_alam, family = binomial) # teeme logistilise
segamõjudega mudeli 2 juhusliku vabaliikmega
> mudel2.glmer <- glmer(ASEND ~ EITUSSÕNA + EELMINE + (1|LEMMA) +
(1+EITUSSÕNA|KÕNELEJA), data = eitus_alam, family = binomial) # teeme
logistilise segamõjudega mudeli 2 juhusliku vabaliikmega ja 1 juhusliku
kaldega

> anova(mudel1.glmer, mudel2.glmer) # kas juhuslik kalle teeb mudelit
paremaks?

Data: eitus_alam
Models:
mudel1.glmer: ASEND ~ EITUSSÕNA + EELMINE + (1 | LEMMA) + (1 | KÕNELEJA)
mudel2.glmer: ASEND ~ EITUSSÕNA + EELMINE + (1 | LEMMA) + (1 + EITUSSÕNA |
KÕNELEJA)
             npar    AIC     BIC  logLik deviance  Chisq Df Pr(>Chisq)
mudel1.glmer    5 972.45  997.26 -481.22   962.45
mudel2.glmer    7 975.98 1010.72 -480.99   961.98 0.4689  2      0.791 # ei
(p > 0.05)


> summary(mudel1.glmer) # vaatame ainult juhuslike vabaliikmetega mudeli
väljundit

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: ASEND ~ EITUSSÕNA + EELMINE + (1 | LEMMA) + (1 | KÕNELEJA)
   Data: eitus_alam

     AIC       BIC   logLik   deviance  df.resid
   972.4       997.3   -481.2     962.4      1051

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.4196  0.0234  0.3061  0.4889  2.1740

Random effects:
 Groups    Name        Variance Std.Dev.
 LEMMA     (Intercept) 0.6469   0.8043
 KÕNELEJA  (Intercept) 0.4029   0.6347
Number of obs: 1056, groups:  LEMMA, 83; KÕNELEJA, 8

Fixed effects:
            Estimate  Std. Error  z value      Pr(>|z|)
(Intercept)   0.1557      0.3285    0.474      0.635591
EITUSSÕNAes   0.7087      0.1984    3.572      0.000354 ***
```

```
EELMINEtaga   1.0994      0.1825    6.024  0.0000000017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
             (Intr) EITUSS
EITUSSÕNAes -0.263
EELMINEtaga -0.381 -0.036
```

Näeme, et võrreldes mudeliga, kuhu oleme kaasanud ainult verbilemma ja kõneleja tunnuse juhuslikud vabaliikmed, ei anna verbilemma põhjal eitussõna mõju modifitseerimine oluliselt midagi mudeli seletusvõimele juurde. Seega võiksime valida pigem lihtsama, ilma juhusliku kaldeta mudeli (*mudel1.glmer*). Selle mudeli väljundist näeme, et vaatlused rühmituvad arvestataval määral nii lemma kui ka kõneleja põhjal; samuti näeme, et nii eitussõna kui ka eelmine eituskonstruktsioon on peamõjudena ka pärast juhusliku rühmitumisega arvestamist endiselt statistiliselt olulised tunnused: mineviku eitussõna *es* kasutamine tõstab tagaeituse kasutamise šanssi ja tõenäosust, samuti tõuseb tagaeituse kasutamise tõenäosus märkimisväärselt, kui ka eelmises kasutatud eituskonstruktsioonis paiknes eitussõna tegusõna järel.

Regressioonimudelite pere on väga suur, võimaldades valida sobiva mudeli uuritava tunnuse tüübist ja jaotusest lähtuvalt. Kui uuritavaks tunnuseks on mingid loendusandmed (st väärtused ei saa olla nullist väiksemad), kasutatakse lineaarse regressiooni asemel sageli ka **Poissoni regressiooni**; kui uuritavaks tunnuseks on mingi järjestustunnus (näiteks Likerti skaala vastused), võib kasutada näiteks **ordinaalset regressiooni**. Kui segamõjudega mudelid tulevad appi juhul, kui vaatlused ei ole sõltumatud, on kõikidel regressioonimudelitel veel terve hulk kindlaid **eeldusi**, mille rikkumisel ei ole mudelite tulemused enam ühtviisi sisukad. Näiteks on lineaarse regressiooni (nagu Pearsoni korrelatsioonikordajagi) üks eeldusi, et mudeli ennustuste ja tegelike arvuliste väärtuste erinevused (nn mudeli jäägid, ingl *residuals*) oleksid normaaljaotusega ning et suhe arvuliste seletavate tunnuste ning arvulise uuritava tunnuse vahel oleks lineaarne. Poissoni regressioon jällegi eeldab, et uuritav tunnus (loendusandmed) on Poissoni jaotusega. Ordinaalne regressioon omakorda eeldab, et seletava tunnuse mõju on sama kõigi järjestatud kategooriate vahel. Seetõttu võib juhul, kui mudeli kasutamise eeldused ei ole täidetud, pöörduda hoopis mõne teistsuguse mudelitüübi juurde, näiteks otsustuspuude juurde, mida kirjeldame järgmises alapeatükis.

6.2.2.2. Otsustuspuud

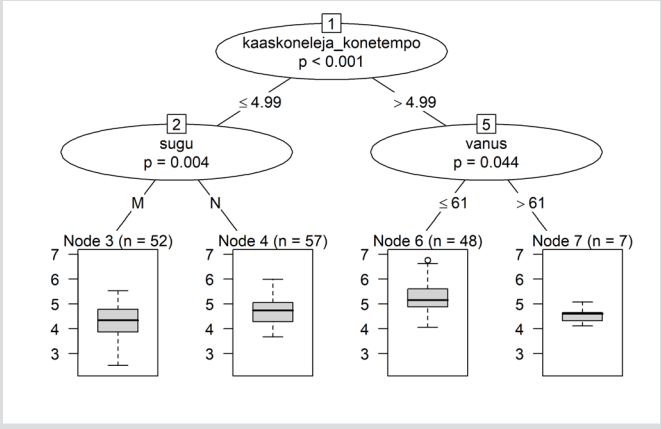
Otsustuspuud (ingl *decision trees*) on statistiliste mudelite klass, mille tööpõhimõte seisneb andmestiku tegelike vaatluste korduvas jagamises kahte rühma sõltuvalt sellest, millised on vaatluste seletavate tunnuste väärtused. Esmalt leitakse uuritava

tunnusega (nt kõnetempo) kõige tugevamalt seotud seletav tunnus (nt sugu) ning jagatakse kõik vaatlused selle tunnuse väärtuste põhjal kaheks (nt meeskõnelejad ja naiskõnelejad). Järgmiseks leitakse kummaski tekkinud rühmas uuesti uuritava tunnusega kõige tugevamalt seotud seletav tunnus ning jagatakse kummagi rühma vaatlused jälle kaheks, kusjuures kummaski rühmas võib jagunemine toimuda erineva seletava tunnuse alusel (näiteks meeskõnelejate puhul vanuse, naiskõnelejate puhul kaaskõneleja soo põhjal). Selline vaatluste kaheksjagamine üha väiksematesse ja spetsiifilisematesse rühmadesse jätkub senikaua, kuni tekkinud rühmades ei leidu enam ühtki uuritava tunnusega oluliselt seotud seletavat tunnust. Jagunemise tulemusena määratakse iga andmestiku vaatlus selle seletavate tunnuse väärtuste põhjal mingisse omavahel sarnaste vaatluste rühma ning sõltuvalt sellest, kas uuritav tunnus on kategooriline tunnus või arvuline tunnus, määratakse iga rühma iseloomustama vastavalt kas uuritava tunnuse üks klass või keskväärtus. Selliste tekkinud klasside põhjal saab ennustada, milline võiks olla iga uue, seni nägemata vaatluse uuritava tunnuse väärtus, kui teame tema seletavate tunnuste väärtusi. Regressioonimudelitest erinevad otsustuspuud seeläbi, et ei püüa sobitada andmetele mingit olemasolevat matemaatilist mudelit, vaid jõuavad sobiva mudelini andmete enda kaudu. Need on intuitiivselt võrdlemisi lihtsasti tõlgendatavad ning võimaldavad visualiseerida ka komplekssseid tunnustevahelisi interaktsioone ehk koosmõjusid.

Otsustuspuude algoritmid erinevad selle poolest, mille alusel vaatluste jagunemiseks tunnuseid valitakse. Siin vaatleme **tingimuslikke otsustuspuid** (ingl *conditional inference trees*), mis kasutavad võimalike hargnemiskohtade hindamiseks p-väärtusi. Teeme siin näitlikustamiseks kaks tingimuslikku otsustuspuud: esimeses ennustame kõnetempot kõneleja soo ja vanuse põhjal, ent lisame mudelisse veel tunnuseid, näiteks kaaskõneleja soo, vanuse ja kõnetempo; teises ennustame idaseto andmete põhjal tagaeituse kasutamist vastavalt eitussõna ajavormile (*ei* või *es*) ja eitussõna asukohale eelmises kasutatud eituskonstruktsioonis (verbivormi ees või taga) ning lisame seletavate tunnustena juurde veel verbitüübi (kognitsiooni-, liikumis-, modaal-, olemis- või muu verb) ning verbi isikuvormi (1., 2., 3. isik, üldisik või impersonaal). R-i pakett `partykit` (Hothorn, Hornik & Zeileis 2006; Hothorn & Zeileis 2015), mida siin otsustuspuude tegemiseks kasutame, nõuab ka, et kõik kategoorilised tunnused oleksid mudeli tarvis teisendatud faktortunnusteks. DataDOI-st laaditud kõnetempo andmestikus on kõneleja ja kaaskõneleja sugu juba faktorid, eituse andmestikus peame tunnused ise faktoriteks teisendama. Sõltuvalt sellest, kas uuritav tunnus on arvuline või kategooriline, nimetatakse otsustuspuid vahel ka vastavalt regressiooni- ja klassifitseerimispuuks.

```
> install.packages("partykit") # installime paketi partykit
> library(partykit) # laadime paketi partykit funktsioonid

> fonkorp2$kaaskoneleja_konetempo <- fonkorp2$kaaskoneleja_silpide_arv/
fonkorp2$kaaskoneleja_kestus # loome kaaskõneleja kõnetempo tunnuse
```

```
> konetempo_puu <- ctree(konetempo ~ vanus + sugu + kaaskoneleja_vanus
+ kaaskoneleja_sugu + kaaskoneleja_konetempo, data = fonkorp2) # teeme
otsustuspuu mudeli
> konetempo_puu # kuvame puumudelit tekstina konsoolis

Model formula:
konetempo ~ vanus + sugu + kaaskoneleja_vanus + kaaskoneleja_sugu +
    kaaskoneleja_konetempo

Fitted party:
[1] root
|   [2] kaaskoneleja_konetempo <= 4.99049
|   |   [3] sugu in M: 4.304 (n = 52, err = 21.3)
|   |   [4] sugu in N: 4.714 (n = 57, err = 18.8)
|   [5] kaaskoneleja_konetempo > 4.99049
|   |   [6] vanus <= 61: 5.269 (n = 48, err = 18.3)
|   |   [7] vanus > 61: 4.535 (n = 7, err = 0.6)

Number of inner nodes:    3
Number of terminal nodes: 4

> plot(konetempo_puu) # kuvame puumudelit joonisel
```



Otsustuspuu mudel valib kõikidest sisendiks antud seletavatest tunnustest esmalt kõige tugevamalt uuritava tunnuse varieerumisega seotud tunnuse ning jagab selle põhjal vaatlused kahte rühma. Kõnetempo andmestiku puhul osutub kõige tugevamaks seletavaks tunnuseks kaaskõneleja kõnetempo (sõlm 1). Puu vasakpoolsesse harusse jagatakse andmestiku vaatlused, mille puhul kaaskõneleja kõnetempo väärtused on 4,99 silpi sekundis või alla selle („aeglased kaaskõnelejad"), ning parempoolsesse harusse vaatlused, mille puhul kaaskõneleja kõnetempo väärtused on üle 4,99 silbi sekundis („kiired kaaskõnelejad"). Kui kaaskõneleja kõnetempo on pigem aeglane, on aeglasem ka kõneleja kõnetempo ja vastupidi. Kuna kaaskõneleja kõnetempo on arvuline tunnus, katsetab mudel kahe rühma tekitamiseks palju

erinevaid väärtusi, ent valib väärtuse 4,99 kui võimalikest optimaalsema. Aeglasemate kaaskõnelejatega vesteldes mängib omakorda rolli ka sugu (sõlm 2): meeskõnelejate kõnetempo, keda vaadeldavas rühmas on 52 (sõlm 3), on keskmiselt aeglasem kui naiskõnelejate oma, keda vaadeldavas rühmas on 57 (sõlm 4). Kui mudeli väljund joonisel võimaldab vaadelda andmete üldist jaotumist karpdiagrammil, siis konsooli trükitav mudeli väljund näitab, et aeglasemate kaaskõnelejatega kõnelevate meeskõnelejate keskmine kõnetempo on 4,304 ning aeglasemate kaaskõnelejatega kõnelevate naiskõnelejate keskmine kõnetempo 4,714 silpi sekundis. Kui kaaskõneleja räägib aga pigem kiiresti, muutub soo asemel oluliseks hoopis kõneleja vanus (sõlm 5): kõneledes kiirete kaaskõnelejatega on 61-aastaste või nooremate kõnelejate kõnetempo keskmiselt 5,269 silpi sekundis (sõlm 6), üle 61-aastastel aga 4,535 silpi sekundis (sõlm 7), ehkki selliseid vaatlusi on andmestikus kõigest 7. Võime seega öelda, et andmestiku kõige vanemate kõnelejate puhul ei toimu kaaskõneleja kiire kõnetempoga sarnast kohaldumist nagu noorematel kõnelejatel. Sisuliselt ei tee otsustuspuu midagi muud kui kuvab meie uuritava tunnuse jaotumist tegelikus valimis vastavalt sellega statistiliselt oluliselt seotud seletavatele tunnustele. Saadud tegelike vaatluste rühmade (sõlmed 3, 4, 6 ja 7) keskväärtusi saab aga kasutada uute vastavate omadustega vaatluste kõnetempo väärtuste ennustamiseks.

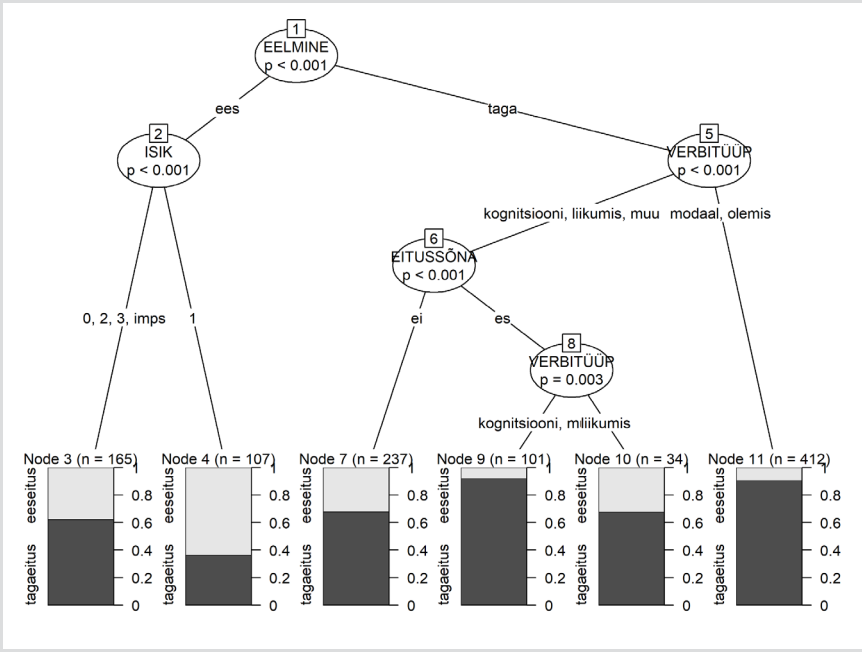
Arvulise uuritava tunnuse keskväärtusi ennustava regressioonipuu headust ehk sobivust andmetele saab hinnata, korreleerides vaatluste tegelikke kõnetempo väärtusi ja neile ennustatud kõnetempo väärtusi. Selleks saame kasutada peatükis 6.2.1.2 käsitletud Pearsoni korrelatsioonikordajat: mida lähemal on korrelatsioonikordaja 1-le, seda täpsemalt mudel kõnetempo väärtusi suudab ennustada. Kui võtta korrelatsioonikordaja väärtus ruutu, saame leida R^2 mõõdiku, mida kasutatakse ka lineaarsete mudelite headuse hindamisel (vt alapeatükk 6.2.2.1.1).

```
> cor(fonkorp2$konetempo, predict(konetempo_puu), method = "pearson") # leiame
Pearsoni korrelatsioonikordaja
[1] 0.5345886
> cor(fonkorp2$konetempo, predict(konetempo_puu), method = "pearson")^2 #
leiame R2
[1] 0.285785
```

Vaatame teise näitena ka kategoorilise uuritava tunnusega otsustuspuud (nn klassifitseerimispuud), kus ennustame tagaeituse kasutamist.

```
> eitus_alam[] <- lapply(eitus_alam, function(tulp) if(is.character(tulp))
as.factor(tulp) else tulp) # muudame kõik andmestiku kategoorilised tunnused
korraga faktoriteks
> levels(eitus_alam$VERBITÜÜP) <- gsub("verb", "", levels(eitus_
alam$VERBITÜÜP)) # lühendame verbitüübi nimetusi

> eituse_puu <- ctree(ASEND ~ EITUSSÕNA + EELMINE + VERBITÜÜP + ISIK, data =
eitus_alam) # teeme otsustuspuu mudeli
```

```
> eituse_puu # kuvame puumudelit tekstina konsoolis

Model formula:
ASEND ~ EITUSSÕNA + EELMINE + VERBITÜÜP + ISIK

Fitted party:
[1] root
|   [2] EELMINE in ees
|   |   [3] ISIK in 0, 2, 3, imps: tagaeitus (n = 165, err = 37.6%)
|   |   [4] ISIK in 1: eeseitus (n = 107, err = 36.4%)
|   [5] EELMINE in taga
|   |   [6] VERBITÜÜP in kognitsiooni, liikumis, muu
|   |   |   [7] EITUSSÕNA in ei: tagaeitus (n = 237, err = 32.1%)
|   |   |   [8] EITUSSÕNA in es
|   |   |   |   [9] VERBITÜÜP in kognitsiooni, muu: tagaeitus (n = 101, err =
7.9%)
|   |   |   |   [10] VERBITÜÜP in liikumis: tagaeitus (n = 34, err = 32.4%)
|   |   [11] VERBITÜÜP in modaal, olemis: tagaeitus (n = 412, err = 9.5%)

Number of inner nodes:    5
Number of terminal nodes: 6

> plot(eituse_puu, gp = gpar(fontsize = 10)) # kuvame puumudelit joonisel
```



Esimese asjana näeme puumudeli jooniselt, et kategoorilise uuritava tunnusega klassifitseerimispuu lõpusõlmedes ei näidata mitte karpdiagramme, vaid tulpdiagramme, mis väljendavad uuritava tunnuse väärtuste (antud juhul *eeseitus* ja

tagaeitus) osakaalusid teatud omadustega vaatluste rühmas. Kõige olulisemaks tunnuseks, mille põhjal vaatlusi klassidesse jagada, hindab puumudel eelmise kasutatud eituskonstruktsiooni sõnajärge (sõlm 1): kui viimati aktiveeritud eituskonstruktsioon oli eeseitus, on tagaeituse kasutamise tõenäosus üldiselt madalam kui siis, kui viimati kasutati samuti tagaeitust. Kui eelmine eituskonstruktsioon oli eeseitus, mängib aga järgmiseks rolli isik, kellest räägitakse (sõlm 2): ainsa vaatluste rühma, kus eeseitus on tagaeitusest tõenäolisem, moodustavad 1. isiku eitatud verbivormid (nt *maq ei tiiä*), millele on ka vahetult eelnenud eeseitus (sõlm 4); teiste isikute puhul domineerib siiski tagaeitus (sõlm 3), ehkki ka selles vaatluste rühmas on omajagu eeseituse kasutusjuhte. Kui eelmine eituskonstruktsioon oli aga tagaeitus (joonise parempoolne haru), mängivad rolli verbitüüp (sõlmed 5 ja 8) ja eitussõna ajavorm (sõlm 6). Mudeli põhjal on niisiis kõige tõenäolisem kasutada tagaeitust siis, kui 1) eelmine eituskonstruktsioon oli samuti tagaeitus, 2) verb on kas modaal- või olemisverb (sõlm 11) või 3) verb on kognitsiooniverb minevikus (sõlm 9). Puumudeli konsooli trükitud väljundist näeme ka igale vaatluste rühmale ennustatud uuritava tunnuse klassi (*eeseitus* või *tagaeitus*), vastavalt sellele, kumma väärtuse osakaal rühmas on suurem.

Klassifitseerimispuu headust saab sarnaselt logistilise regressiooniga hinnata klassifitseerimistäpsuse ja C-indeksi kaudu, kusjuures viimast saab arvutada ainult juhul, kui uuritaval kategoorilisel tunnusel on ainult kaks klassi.

```
> # arvutame klassifitseerimistäpsuse
> ennustatud_klassid <- predict(eituse_puu) # leiame mudeli ennustused
> tegelikud_klassid <- eitus_alam$ASEND # küsime uuritava tunnuse tegelikke
klasse
> ennustuste_tabel <- table(ennustatud_klassid, tegelikud_klassid) # teeme
ennustustest ja tegelikest klassidest risttabeli
> sum(diag(ennustuste_tabel))/sum(ennustuste_tabel) # leiame
klassifitseerimistäpsuse
[1] 0.7774621

> # leiame C-indeksi
> library(pROC)
> auc(tegelikud_klassid, predict(eituse_puu, type = "prob")[,2])

Setting levels: control = eeseitus, case = tagaeitus
Setting direction: controls < cases
Area under the curve: 0.7454
```

Nägime niisiis, et otsustuspuu on küllalt lihtne ja hõlbus viis visualiseerida ja analüüsida seletavate tunnuste koosmõjusid uuritava tunnuse varieerumisele ning kuna neil puuduvad eeldused andmete jaotuste kohta, on need sobivaks mitteparameetriliseks alternatiiviks regressioonimudelitele, kui viimaste eeldused on tõsiselt rikutud. Koosmõjusid saab muidugi mudeldada ka regressioonimudelitega (vt nt L. Lindströmi & M.-L. Pilviku näidisuurimust), ent kuna erinevalt otsustuspuudest

toimub nende tõlgendamine n-ö globaalselt (kõikide vaatluste jaoks, hoides samal ajal muude tunnuste väärtused konstantsena), mitte lokaalselt (ainult teatud omadustega vaatluste rühma jaoks), võib mitmete eri tunnuste vaheliste interaktsioonide lisamine teha regressioonimudeli tõlgendamise väga keerukaks. Ehkki otsustuspuud ei ole parameetrilised mudelid, eeldavad needki tegelikult, et vaatlused on üksteisest sõltumatud. Seetõttu tuleb nende kasutamisel näiteks kõnelejati või muude tunnuste alusel rühmitunud andmetega teadvustada, et saadud tulemused võivad olla kallutatud ning peegeldada populatsioonitasandi mõjude asemel üksikute rühmade (nt kõnelejate) käitumismustreid.

Lõpetuseks

Õpiku kuuendas peatükis andsime ülevaate, kuidas korpusest kogutud keeleandmetest saab tuvastada meid huvitavat keelenähtust kirjeldavaid mustreid ja seoseid. Statistiliste meetodite kasutamine on loomulik osa empiiriliste andmete analüüsiprotsessist. Kuna korpuslingvistilises uurimistöös räägime üldiselt pigem suurtest andmehulkadest, siis aitavad just statistilised meetodid leida andmetest seaduspärasusi, mida intuitiivselt või palja silmaga ei pruugi näha. Üldistatult võime rääkida kahest laiemast suunast statistikas – kirjeldav ehk deskriptiivne statistika ja järeldav ehk inferentsiaalne statistika. Kirjeldav statistika aitab iseloomustada valimi andmete jaotust ja tüüpilisi väärtusi (nt haare, aritmeetiline keskmine, mediaan). Kirjeldava statistika meetodite abil saame iseloomustada mingit nähtust konkreetses andmestikus, aga üldjuhul ei võimalda need meetodid teha väga laiaulatuslikke järeldusi uuritava nähtuse kohta kogu populatsioonis või ümber lükata hüpoteese. Järeldava statistika meetodid hõlmavad aga statistilisi teste (nt hii-ruut-test, t-test, U-test, korrelatsioonikordajad) ja statistilisi mudeleid (nt regressioonimudelid, otsustuspuud), mis lubavad valimi põhjal teha teatud kindlusega järeldusi ka populatsiooni kohta. Korpusandmete statistilise analüüsi läbiviimiseks saab kasutada erinevaid tarkvarasid ja programmeerimiskeeli; siinses õpikus keskendume nii sisupeatükkides kui ka näidisuurimustes peamiselt programmeerimiskeelele R.

Statistiliste meetodite kasutamine korpuslingvistilises uurimistöös on äärmiselt oluline ja vajalik, kuid alustava uurija jaoks võib sobiva lähenemisviisi valimine ning selle piirangute mõistmine osutuda keeruliseks. Erinevat tüüpi andmed (nt arvulised *vs*. kategoorilised tunnused) tingivad erinevate statistiliste meetodite kasutamise. Tarkvarad ja programmeerimiskeeled ei anna tavaliselt märku, kui valitud test või mõõdik ei sobi konkreetsele andmestikule, mistõttu lasub vastutus õige meetodi valiku eest uurijal endal. Täiendavaks väljakutseks on aru saada, millised eeldused peavad valitud statistilise meetodi puhul olema täidetud – näiteks kas andmed on normaaljaotusega, üksteisest sõltumatud või kas valimi maht on piisav. Seega on mistahes korpuslingvistilise analüüsi eelduseks see, et uurija saaks

aru oma andmestikust ja uurimisküsimusest: kust ja kuidas andmed on kogutud ning millist tüüpi tunnustega on tegemist. Selles peatükis anti mõned soovitused ja suunised, mis aitavad orienteeruda selliste küsimuste ja otsustuskohtade keerukas rägastikus.

Tasub rõhutada, et iga statistiline mudel on paratamatult lihtsustus – üldistus sellest, kuidas mingi keeleline nähtus keelekasutuses tegelikult toimib. Mudelid on võimelised tuvastama ja kirjeldama üldisi tendentse ja tüüpilisi, sagedasemaid mustreid, kuid ei suuda haarata kogu keelelise varieeruvuse ulatust, sealhulgas haruldasi või ebatüüpilisi kasutusjuhte. Oluline on mõista, et mudeli tulemused on otseselt seotud sisendandmetega – see, mida mudel „näeb" ja mille põhjal ta järeldusi teeb. Seetõttu on kriitilise tähtsusega see, et andmete kogumine, eeltöötlus ja kodeerimine oleksid läbimõeldud ja metoodiliselt põhjendatud. Vaid sel juhul on võimalik jõuda tõenduspõhiste ja usaldusväärsete järeldusteni.

Lisalugemiseks

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Gries, Stefan Th. 2021. *Statistics for linguistics with R: A practical introduction*. 3. tr. Berlin / Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110718256.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.195.

Paquot, Magali & Stefan Th. Gries (toim). 2020. *A practical handbook of corpus linguistics*, 473–646. Cham: Springer. https://doi.org/10.1007/978-3-030-46216-1.

Sügis, Elena, Ardi Tampuu, Anna Aljanaki, Mark Fišel & Meelis Kull. 2024. *Praktiline andmeteadus: kõrgkooliõpik*. Tartu: Tartu Ülikooli arvutiteaduse instituut. https://hdl.handle.net/10062/106497.

Wallis, Sean. 2021. *Statistics in corpus linguistics research: A new approach*. New York / Oxon: Routledge.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge. https://doi.org/10.4324/9781315165547.

Kirjandus

Aedmaa, Eleri. 2014. Sõnadevahelise seose tugevuse mõõtmise statistilised meetodid ühendverbide tuvastamisel. Magistritöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut, käsikiri. http://hdl.handle.net/10062/44260.

Agresti, Alan. 2013. *Categorical data analysis* (Wiley Series in Probability and Statistics 792). 3. tr. Hoboken, NJ: John Wiley & Sons.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert & Arne Zeschel. 2010. Cognitive corpus linguistics: Five points of debate on current theory and methodology. *Corpora* 5(1). 1–27. https://doi.org/10.3366/cor.2010.0001.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1). 1–16. https://doi.org/10.1093/llc/7.1.1.

Baker, Paul. 2023. *Using corpora in discourse analysis* (Bloomsbury Discourse). 2. tr. London / New York / Oxford / New Delhi / Sydney: Bloomsbury Academic. https://doi.org/10.5040/9781350083783.

Barth, Danielle & Stefan Schnell. 2021. *Understanding corpus linguistics*. London: Routledge. https://doi.org/10.4324/9780429269035.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Ben-Shachar, Mattan S., Daniel Lüdecke & Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software* 5(56). 2815. https://doi.org/10.21105/joss.02815.

Ben-Shachar, Mattan S., Indrajeet Patil, Rémi Thériault, Brenton M. Wiernik & Daniel Lüdecke. 2023. Phi, Fei, Fo, Fum: Effect sizes for categorical data that use the chi-squared statistic. *Mathematics* 11. 1982. https://doi.org/10.3390/math11091982.

Biber, Douglas. 1988. *Variation across speech and writing*. 1. tr. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257. https://doi.org/10.1093/llc/8.4.243.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*. Walter de Gruyter.

Boros, Emanuela, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer & Frédéric Kaplan. 2024. Post-correction of historical text transcripts with large language models: An exploratory study. Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva & Stan Szpakowicz (toim), *Proceedings of the 8th*

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 133–159. St. Julians, Malta: Association for Computational Linguistics. https://aclanthology.org/2024.latechclfl-1.14.

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Cavaglia, Gabriela & Adam Kilgarriff. 2001. Corpora from the Web. *Information Technology Research Institute Technical Report Series*. https://www.kilgarriff.co.uk/Publications/2001-CavagliaKilg-CLUK.pdf.

Cochran, William G. 1952. The χ2 test of goodness of fit. *Annals of Mathematical Statistics* 23. 315–345.

Cochran, William G. 1954. Some methods for strengthening the common χ2 tests. *Biometrics* 10. 417–451.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20. 37–46. https://doi.org/10.1177/001316446002000104

Cohen, Jacob. 1988. *Statistical power analysis for the behavioural sciences*. 2. tr. New York: Routledge.

Cramér, Harald. 1946. *Mathematical methods of statistics*. Princeton: Princeton University Press.

Crosthwaite, Peter & Vit Baisa. 2023. Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics* 3(3). 100066. https://doi.org/10.1016/j.acorp.2023.100066.

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. https://doi.org/10.1162/coli_a_00402.

Divjak, Dagmar. 2019. *Frequency in language: Memory, attention and learning*. Cambridge: Cambridge University Press.

Erelt, Mati & Helle Metslang (toim). 2017. *Eesti keele süntaks* (Eesti keele varamu 3). Tartu: Tartu Ülikooli Kirjastus. http://hdl.handle.net/10062/70510.

Eslon, Pille. 2014. Eesti vahekeele korpus. *Keel ja Kirjandus* 6. 436–451. https://doi.org/10.54013/kk679a3.

Eslon, Pille, Annekatrin Kaivapalu, Katre Õim, Mare Kitsnik, Olga Gaitšenja & Kais Allkivi-Metsoja. 2021. *Eesti keele oskuse arenemine ja arendamine. Kirjalik õppijakeel*. Tallinn: EKSA.

Eslon, Pille & Helena Metslang. 2007. Õppijakeel ja eesti vahekeele korpus. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 3. 99–116. https://doi.org/10.5128/ERYa3.07.

Fletcher, William H. 2007. Concordancing the web: Promise and problems, tools and techniques. *Corpus Linguistics and the Web* (Language and Computers 59), 25–45. Brill.

Francis, Nelson W. & Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
Fries, Charles C. 1952. *The structure of English*. New York: Harcourt Brace.
Gilquin, Gaëtanelle & Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26. https://doi.org/10.1515/CLLT.2009.001.
Granger, Sylviane. 1987. *The* be+past participle *construction in spoken English: With special emphasis on the passive* (North-Holland Linguistics Series 49). Amsterdam: Elsevier Science Publishers.
Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next ... *International Journal of Corpus Linguistics* 18(1). 137–166. https://doi.org/10.1075/ijcl.18.1.09gri.
Hennoste, Tiit. 2000. Suulise eesti keele uurimine: transkriptsioon, taust ja korpus. *Keel ja Kirjandus* 2. 91–106.
Hennoste, Tiit. 2003. Suulise eesti keele uurimine: korpus. *Keel ja Kirjandus* 7. 481–500.
Hennoste, Tiit, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis & Krista Strandson. 2009. Suulise eesti keele korpus ja inimese suhtlus arvutiga. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 5. 111–130. https://doi.org/10.5128/ERYa5.07.
Hennoste, Tiit & Kadri Muischnek. 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. *Arvutuslingvistikalt inimesele* (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1), 183–317. Tartu: Tartu Ülikooli Kirjastus.
Hodge, David. 2024. *ggblanket: Simplify ggplot2 visualisation*. https://CRAN.Rproject.org/package=ggblanket.
Hosmer, David W., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied logistic regression*. 3. tr. Hoboken, NJ: John Wiley & Sons.
Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674. https://doi.org/10.1198/106186006X133933.
Hothorn, Torsten & Achim Zeileis. 2015. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research* 16. 3905–3909.
Hripcsak, George & Adam S. Rothschild. 2005. Agreement, the F-Measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association* 12(3). 296–298. https://doi.org/10.1197/jamia.M1733.
Hunston, Susan. 2022. *Corpora in applied linguistics*. 2. tr. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108616218.
Ivaska, Ilmari & Silvia Bernardini. 2020. Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics* 43(1). 33–57. https://doi.org/10.1017/S0332586520000013.

Jaanimäe, Gerth. 2021. Ajalooliste tekstide normaliseerimine. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 17. 47–59. https://doi.org/10.5128/ERYa17.03.

Jakubíček, Miloš, Vojtěch Kovář, Pavel Rychlý & Vit Suchomel. 2020. Current challenges in web corpus building. Adrien Barbaresi, Felix Bildhauer, Roland Schäfer & Egon Stemle (toim), *Proceedings of the 12th Web as Corpus Workshop*, 1–4. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.wac-1.1.

Johansson, Stig & Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar: Tag combinations and word combinations*. Oxford: Clarendon Press.

Jurafsky, Dan & James H. Martin. 2025. Speech and language processing. https://web.stanford.edu/~jurafsky/slp3/.

Jürine, Anni, Jane Klavan & Ann Veismann. 2013. Katseline semantika: planeerimine ja teostus. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 9. 85–100. https://doi.org/10.5128/ERYa9.06.

Kallas, Jelena, Kristina Koppel & Maria Tuulik. 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 11. 75–94. https://doi.org/10.5128/ERYa11.05.

Kasik, Reet. 2007. *Sissejuhatus tekstiõpetusse*. Tartu: Tartu Ülikooli Kirjastus.

Kaukonen, Elisabeth. 2023a. Kes on esinaine? *Keel ja Kirjandus* 66(3). 328–336. https://doi.org/10.54013/kk783a4.

Kaukonen, Elisabeth. 2023b. Cleaning aunts and police uncles in action. Unveiling gender dynamics in Estonian compound words. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 14(3). 137–171. https://doi.org/10.12697/jeful.2023.14.3.05.

Kehoe, Andrew. 2020. Web corpora. Magali Paquot & Stefan Th. Gries (toim), *A practical handbook of corpus linguistics*, 329–351. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_15.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. New York: Routledge.

Kilgarriff, Adam, David Tugwell, Pavel Rychlý & Pavel Smrz. 2004. The Sketch Engine. Geoffrey Williams & Sandra Vessier (toim), *Proceedings of the 11th EURALEX International Congress*, 105–115. Lorient, France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1(1). 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Kingisepp, Valve-Liivi, Külli Prillop & Külli Habicht. 2004. Eesti vana kirjakeele korpus: mis tehtud, mis teoksil. *Keel ja Kirjandus* 4. 272–280.

Klavan, Jane. 2024. *The making and breaking of classification models in linguistics: A multimethod perspective on constructional alternations*. Berlin / Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110668469.

Klavan, Jane, Ann Veismann & Anni Jürine. 2013. Katselised meetodid tähenduse uurimisel. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 4(1). 17–34. https://doi.org/10.12697/jeful.2013.4.1.02.

Koppel, Kristina & Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 18. 207–228. https://doi.org/10.5128/ERYa18.12.

Koppel, Kristina, Jelena Kallas, Katrin Tsepelina & Piret Laanesaar. 2022. Eesti keele kui teise keele õppekorpus 2022. https://doi.org/10.15155/3-00-0000-0000-0000-08C06L.

Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. *Linguistics* 59(5). 1207–1226. https://doi.org/10.1515/ling-2019-0046.

Kristiansen, Tore. 2021. Destandardization. Wendy Ayres-Bennett & John Bellamy (toim), *The Cambridge Handbook of Language Standardization*, 667–690. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108559249.002

Kriuchkova, Sofia. 2025. Erinevused eesti keelt kõnelevate meeste ja naiste suulises keelekasutuses. Magistritöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut, käsikiri. https://hdl.handle.net/10062/111466.

Kroonenberg, Pieter M. & Albert Verbeek. 2018. The tale of Cochran's rule: My contingency table has so many expected values smaller than 5, what am I to do? *The American Statistician* 72(2). 175–183. https://doi.org/10.1080/00031305.2017.1286260.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. https://doi.org/10.18637/jss.v082.i13.

Kuusik, Aet. 2024. Kuidas suhtub Eesti LGBT kogukond sõnasse *kväär*? *Keel ja Kirjandus* 4. 315–339. https://doi.org/10.54013/kk796a1.

Kängsepp, Annika. 2024. Käändevormide varieerumine ja seda mõjutavad tegurid kirjalikus keeles indefiniitpronoomeni *keegi* näitel. *Keel ja Kirjandus* 11. 1016–1037. https://doi.org/10.54013/kk803a3.

Laur, Sven, Siim Orasmaa, Dage Särg & Paul Tammo. 2020. EstNLTK 1.6: Remastered Estonian NLP pipeline. Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, jt (toim), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7152–7160. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.884.

Lindström, Liina, Pärtel Lippus & Tuuli Tuisk. 2019. The online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages and the Corpus of Estonian Dialects. Sofia Björklöf & Santra Jantunen (toim), *Multilingual Finnic – Language Contact and Change* (Uralica Helsingiensia 14), 327–350. Helsinki: Suomalais-Ugrilainen Seura. https://doi.org/10.33341/uh.85040.

Lindström, Liina, Maarja-Liisa Pilvik & Helen Plado. 2021. Variation in negation in Seto. *Studies in Language* 45(3). 557–597. https://doi.org/10.1075/sl.19063.lin.

Lindström, Liina, Maarja-Liisa Pilvik & Triin Todesk. 2024. SetKo: interdistsiplinaarne seto korpus / Interdisciplinary Corpus of Seto. https://doi.org/10.17605/OSF.IO/8WB2J.

Lindström, Liina, Triin Todesk & Maarja-Liisa Pilvik. 2022. Eesti murrete korpus. https://doi.org/10.23673/re-365.

Lippus, Pärtel, Kätlin Aare, Anton Malmi, Tuuli Tuisk & Pire Teras. 2023. Phonetic corpus of Estonian spontaneous speech v1.3. Institute of Estonian and General Linguistics, University of Tartu. https://doi.org/10.23673/re-438.

Lippus, Pärtel, Tanel Alumäe, Siim Orasmaa, Maarja-Liisa Pilvik & Liina Lindström. 2023. Eesti taskuhäälingukorpus. https://doi.org/10.23673/RE-445.

Lippus, Pärtel, Tanel Alumäe, Siim Orasmaa, Katrin Tsepelina & Liina Lindström. 2023. Eesti Rahvusringhäälingu raadiosaadete korpus. https://doi.org/10.23673/RE-441.

Lippus, Pärtel, Kaidi Lõo, Anton Malmi & Maarja-Liisa Pilvik. 2024. Suuline eesti keel arvudes. Sagedusandmestikud. Tartu Ülikool, eesti ja üldkeeleteaduse instituut. https://doi.org/10.23673/RE-463.

Lippus, Pärtel, Maarja-Liisa Pilvik, Kaidi Lõo & Liina Lindström. 2024. Kõnetempo ja -soravuse varieerumine eesti keeles. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 20. 149–163. https://doi.org/10.5128/ERYa20.09.

Lüdecke, Daniel. 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software* 3(26). 772. https://doi.org/10.21105/joss.00772.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner & Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* 6(60). 3139. https://doi.org/10.21105/joss.03139.

Makowski, Dominique, Daniel Lüdecke, Indrajeet Patil, Rémi Thériault, Mattan S. Ben-Shachar & Brenton M. Wiernik. 2023. Automated results reporting as a practical tool to improve reproducibility and methodological best practices adoption. *CRAN*. https://easystats.github.io/report/.

Masso, Anu, Katrin Tiidenberg & Andra Siibak (toim). 2020. *Kuidas mõista andmestunud maailma? Metodoloogiline teejuht*. Tallinn: Tallinna Ülikooli kirjastus.

McEnery, Tony, Richard Xiao & Yukio Tono. 2005. *Corpus-based language studies: An advanced resource book* (Routledge Applied Linguistics). London / New York: Routledge.

McEnery, Tony & Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Meister, Einar. 2015. Corpus of adolescent speech / Lastekõne korpus. https://doi.org/10.15155/9-00-0000-0000-0000-0006EL.
Meister, Einar & Lya Meister. 2017. Eesti laste kõne I: Põhitooni akustiline analüüs. *Keel ja Kirjandus* 7. 518–533. https://doi.org/10.54013/kk716a2.
Meister, Einar & Lya Meister. 2022. Estonian elderly speech corpus – Design, collection and preliminary acoustic analysis. *Baltic Journal of Modern Computing* 10(3). https://doi.org/10.22364/bjmc.2022.10.3.09.
Meister, Lya & Einar Meister. 2012. Aktsendikorpus ja võõrkeele aktsendi uurimine. *Keel ja Kirjandus* 8–9. 696–714. https://doi.org/10.54013/kk658a10.
Metslang, Helle, Külli Habicht, Tiit Hennoste, Kirsi Laanesoo-Kalk, Külli Prillop, Andriela Rääbis & Carl Eric Simmul. 2024. (Inter)subjectivity in Estonian registers. *Journal of Uralic Linguistics* 3(2). 119–157. https://doi.org/10.1075/jul.00028.met.
Muischnek, Kadri. 2011. Corpus of morphologically disambiguated Estonian texts / Morfoloogiliselt ühestatud korpus. https://doi.org/10.15155/1-00-0000-0000-0000-00085L.
Muischnek, Kadri. 2015. Estonian treebank / Eesti keele puudepank. https://doi.org/10.15155/1-00-0000-0000-0000-00089L.
Müürisep, Kaili. 2000. *Eesti keele arvutigrammatika: süntaks* (Dissertationes mathematicae Universitatis Tartuensis 22). Tartu: Tartu Ülikooli Kirjastus.
Newman, Matthew L., Carla J. Groom, Lori D. Handelman & James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3). 211–236. https://doi.org/10.1080/01638530802073712.
Nicenboim, Bruno & Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational ideas—Part II. *Language and Linguistics Compass* 10(11). 591–613. https://doi.org/10.1111/lnc3.12207.
Olev, Aivo & Tanel Alumäe. 2022. Estonian speech recognition and transcription editing service. *Baltic Journal of Modern Computing* 10(3). 409–421.
Piits, Liisi. 2015. *Sagedamate inimest tähistavate sõnade kollokatsioonid eesti keeles* (Dissertationes linguisticae Universitatis Tartuensis 23). Tartu: Tartu Ülikooli Kirjastus.
Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa & Tõnis Türna. 2019. *Möistus sai kuulotedu*: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 15. 139–158. https://doi.org/10.5128/ERYa15.08.
Pilvik, Maarja-Liisa, Liina Lindström & Helen Plado. 2021. Murded, varieerumine ja korpusandmed: eitussõna paiknemine võru ja seto eituslausetes. Lisamaterjalid. OSF. https://doi.org/10.17605/OSF.IO/GCFT7.
Pilvik, Maarja-Liisa, Helen Plado & Liina Lindström. 2021. Murded, varieerumine ja korpusandmed. Eitussõna paiknemine võru ja seto eituslausetes. *Keel ja Kirjandus* 8–9. 771–796. https://doi.org/10.54013/kk764a7.

Posit team. 2024. RStudio: Integrated development environment for R. Boston, MA: Posit Software. https://www.posit.co/.

Prillop, Külli. 2013. Corpus of old written Estonian / Vana kirjakeele korpus. https://doi.org/10.15155/1-00-0000-0000-0000-00075L.

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raudvere, Uku & Kristel Uiboaed. 2018. Uuema eesti ilukirjanduse mitmikute loendid. UTLIB andmed. https://doi.org/10.15155/re-8.

Ross, Kristiina & Heiki Reila. 2020. Concordance of the historical Estonian Bible translation / Eesti piiblitõlke ajalooline konkordants. https://doi.org/10.15155/3-00-0000-0000-0000-08238L.

Rääbis, Andriela. 2013. Corpus of spoken Estonian / Suulise keele korpus. https://doi.org/10.15155/1-00-0000-0000-0000-00077L.

Sankoff, David & Gillian Sankoff. 1973. Sample survey methods and computer-assisted analysis in the study of grammatical variation. Regna Darnell (toim), *Canadian Languages in Their Social Context*, 7–64. Edmonton, AB: Linguistic Research.

Savický, Petr & Jaroslava Hlavácová. 2002. Measures of word commonness. *Journal of Quantitative Linguistics* 9(3). 215–231. https://doi.org/10.1076/jqul.9.3.215.14124.

Schepens, Job, Nicole Marx & Benjamin Gagl. 2023. Can we utilize large language models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers. https://doi.org/10.31234/osf.io/gm9b6.

Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19. 321–325. https://doi.org/10.1086/266577.

Siiman, Ann. 2016. Ainsuse sisseütleva vormi valiku seos morfosüntaktiliste ja semantiliste tunnustega – materjali ning meetodi sobivus korpusanalüüsiks. *Emakeele Seltsi aastaraamat* 61 (2015). 207–232. https://doi.org/10.3176/esa61.10.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7). Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735821.

Suchomel, Vít & Jan Kraus. 2021. Website properties in relation to the quality of text extracted for web corpora. Aleš Horák, Pavel Rychlý & Adam Rambousek (toim), *RASLAN 2021: Proceedings of Recent Advances in Slavonic Natural Language Processing*, 167–175. Tribun EU.

Svartvik, Jan. 1990. *The London-Lund corpus of spoken English*. Lund: Lund University Press.

Sõrmus, Kadri & Kersti Lepajõe. 2014. Eesti keele kui emakeele õppija tekstikorpus EMMA. *Philologia Estonica Tallinnensis* 16. 205–225.
Sügis, Elena, Ardi Tampuu, Anna Aljanaki, Mark Fišel & Meelis Kull. 2024. *Praktiline andmeteadus. Kõrgkooliõpik*. Tartu: Tartu Ülikooli arvutiteaduse instituut. https://hdl.handle.net/10062/106497.
Šeļa, Artjoms. 2021. Erinevused, kaugused ja sõrmejäljed. Stilomeetria ja mitmemõõtmelise tekstianalüüsi alused. *Keel ja Kirjandus* 8–9. 696–717. https://doi.org/10.54013/kk764a3.
Taylor, Charlotte. 2008. What is corpus linguistics? What the data says. *ICAME Journal* 32. 179–200.
Teras, Pire. 2019. Sõnaalgulise /h/ häältdus samadel kõnelejatel avalikus ja argisuhtluses. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 10(1). 211–231. https://doi.org/10.12697/jeful.2019.10.1.11.
Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work* (Studies in Corpus Linguistics 6). Amsterdam / Philadelphia: John Benjamins Publishing Company. https://benjamins.com/catalog/scl.6.
Törnberg, Petter. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. arXiv. http://arxiv.org/abs/2304.06588.
Uiboaed, Kristel. 2018. Eesti keele stoppsõnad / Estonian stop words. https://doi.org/10.15155/RE-48.
Vaik, Kristiina & Virve-Anneli Vihman. 2017. Eesti lastekeele korpuse morfoloogilise märgendamise kitsaskohtadest. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 13. 205–221. https://doi.org/10.5128/ERYa13.13.
Vihman, Virve-Anneli, Maarja-Liisa Pilvik, Aive Mandel, Annika Kängsepp, Mari Aigro, Kadri Koreinik, Kristiina Praakli & Liina Lindström. 2023. Estonian teen language corpus. https://doi.org/10.23673/RE-455.
Viht, Annika & Külli Habicht. 2019. *Eesti keele sõnamuutmine* (Eesti keele varamu 4). Tartu: Tartu Ülikooli Kirjastus. http://hdl.handle.net/10062/76416.
Wallis, Sean. 2021. *Statistics in corpus linguistics research: A new approach*. New York / Oxon: Routledge.
Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org.
Wickham, Hadley & Jennifer Bryan. 2023. *readxl: Read Excel files*. https://CRAN.R-project.org/package=readxl.
Yu, Danni, Luyang Li, Hang Su & Matteo Fuoli. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics* 29(4). 534–651. https://doi.org/10.1075/ijcl.23087.yu.

II OSA: NÄIDISUURIMUSED

Korpused ja sõnastikud

Kristina Koppel, Jelena Kallas, Margit Langemets

Lühikokkuvõte

Siinne näidisuurimus on mõeldud neile, kes soovivad luua oma sõnastikku või kes soovivad lähemalt teada saada, kuidas korpuse põhjal sõnastikke koostatakse. Sõnastike koostamisel on üheks põhimõtteks kirjeldada tegelikku keelekasutust, mille tagab tekstikorpustele toetumine. Korpuste põhjal saab analüüsida näiteks sõnade esinemissagedust ja leksikaalgrammatilist käitumist, sõnakujude varieerumist ning tähenduste muutumise ulatust ja iseloomu nii diakrooniliselt (varasema aja tekstides) kui ka sünkrooniliselt (tänapäeva tekstides). Siin kirjeldame samm-sammult ühe konkreetse korpuspõhise sõnastiku – „Eesti keele naabersõnad 2019“[1] – koostamise põhietappe. See oli üks esimesi eesti keele sõnastikke, mille andmebaas loodi korpusandmete põhjal automaatselt. Selleks kasutati korpusanalüüsi tarkvara Sketch Engine[2] ja sõnastikusüsteemi EELex[3], keeleliste andmete allikas oli eesti keele ühendkorpus 2013. Anname lühiülevaate meetoditest, mis võimaldasid automaatselt tuvastada järgmisi sõnastikuüksusi: märksõnad, kollokatsioonid ehk naabersõnad ja näitelaused.

1. Sissejuhatus: sõnastiku koostamise etapid

Igasugune sõnastikuprojekt koosneb mitmest etapist, mis tuleb juba enne projekti algust läbi mõelda. Esimene etapp on läbi mõelda sõnastiku kontseptsioon tervikuna. Paika tuleb panna sõnastiku tüüp (nt kas üks-, kaks- või mitmekeelne; üld- või oskuskeele sõnastik; kollokatsiooni- või sünonüümisõnastik), määrata sihtgrupp (kas mõeldud keeleõppijale, kõigile keelehuvilistele, teatud valdkonna spetsialistidele), otsida rahastus, värvata meeskond, koostada ajakava, otsida keeleandmete allikaid (korpused, varem koostatud sõnastikud, arhiivid; kui

[1] https://arhiiv.eki.ee/dict/kol/

[2] https://www.sketchengine.eu/

[3] https://eelex.eki.ee/

olemasolevatest korpustest enda sõnastiku koostamiseks sobivat ei leia, tuleb see ise luua, märgendada ja korpusanalüüsi tarkvarasse paigutada (vt ptk 4 „Oma korpuse loomine") ning otsustada, kus see avaldada (veebis või paberil; eraldi veebilehel või sõnastikuportaali osana). Samuti on oluline tarkvara valik: kuidas andmeid analüüsida (korpusanalüüsi tarkvara) ning talletada (sõnastikusüsteemid).

Eesti keele jaoks on olemas väga palju eri tüüpi korpusi (vt ptk 2 „Eesti keele korpused") ja nende töötlemiseks sobivaid **korpusanalüüsi tarkvarasid** (vt ptk 5.1 „Korpusanalüüsi vahendid"). Enne korpuste tulekut kasutati keeleallikana sedelkartoteeke, mis sisaldasid käsitsi üles kirjutatud sõnasedeleid. Praegugi kasutatakse sedelkartoteeke teatud tüüpi sõnastike, näiteks „Eesti murrete sõnaraamatu"[4] koostamisel. Arvutite tulekuga loodi esimesed korpused, mis sisaldasid nii algselt paberil avaldatud ja hiljem digiteeritud tekste kui ka juba masinloetaval kujul olnud tekste. Sellisena on loodud eesti keele koondkorpus, mida alates 1990ndatest kasutati (sedelkartoteekide kõrval) Eesti Keele Instituudis (EKI) näiteks „Eesti keele seletava sõnaraamatu"[5] (EKSS 1988–2007, 2. trükk EKSS 2009) koostamisel. Praegu on mahukaim eestikeelsete digitekstide kogu eesti keele ühendkorpuste sari (Koppel & Kallas 2022), mis on olnud lähtealuseks tänapäeva sõnastike, näiteks EKI ühendsõnastiku (Langemets jt 2021) koostamisel, aga ka keele uurimisel, vt nt (Paet & Risberg 2021; Vainik, Paulsen & Lohk 2021; Veismann 2021; Risberg 2024).

Korpusanalüüsi tarkvara võimaldab küll tuvastada sõnastiku koostamiseks vajalikku infot, kuid tegemist on siiski alusandmetega, mida tuleb hiljem sõnastikusüsteemis (järel)toimetada. **Sõnastikusüsteem** ongi leksikograafilise töö teine oluline tarkvaraline komponent, sest sõnastike koostamine lihtsamates tekstitöötlus- (nt Microsoft Word) või tabelarvutusprogrammides (nt Microsoft Excel) toob kaasa väga palju ebatäpsust ja ebaühtlust. Sõnastikusüsteemi kasutamine tagab, et andmed on esitatud struktureeritud kujul ehk kindlas vormingus (nt XML, relatsiooniline andmebaas), mis võimaldab edaspidi andmeid taaskasutada ja linkida teiste rakendustega. Sõnastikusüsteem võimaldab teha lihtsamaid ja keerulisemaid struktuuripõhiseid päringuid, sortida päringutulemusi, kontrollida ristviiteid, jälgida sõnastikutöö edenemist, luua ja hallata loendeid, aga ka andmeid eksportida. Tänapäeva sõnastikusüsteemidel on andmevahetuse tagamiseks olemas ka API (ingl *application programming interface*) ehk programmiliides, mille abil üks programm teise käest veebi kaudu andmeid pärib.

Sõnastiku koostamise teine etapp sisaldab andmete korpuspõhist töötlemist ja analüüsi, mis sõltuvalt sõnastiku tüübist hõlmab endas märksõnastiku loomist, sageduste ja sõnaliikide tuvastamist, sõnatähenduste uurimist, kollokatsioonide, terminite, tõlkevastete, sünonüümide tuvastust, näitelausete valikut jm-d. Tehnoloogilise võimekuse korral saab eelnevalt loetletud korpusanalüüsi tulemusi

[4] https://www.eki.ee/dict/ems/

[5] https://www.eki.ee/dict/ekss/

automaatselt korpusest sõnastikusüsteemi üle kanda ning neid seal järeltoimetada (loe lähemalt Jakubíček jt 2018). Suurte keelemudelite tulekuga tehakse aina rohkem katseid rakendada sõnastiku sisu loomisel ka tehisintellekti abi (loe lähemalt De Schryver 2023). Tehisintellekti rakendatakse ka eesti leksikograafias, loe lähemalt E. Aedmaa näidisuurimusest.

Kolmas etapp on sõnastiku koostamine ja toimetamine, mis hõlmab automaatselt tuvastatud korpusinfo toimetamist sõnastikusüsteemis, aga ka eri tüüpi info käsitsi sisestamist sõnastikusüsteemi andmeväljadele. Näiteks ei kasutata EKI ühendsõnastiku (Langemets jt 2021) koostamisel veel korpusanalüüsi võimalusi seletuste automaatsel tuvastamisel ja grammatilise info lisamisel. Neid sisestavad leksikograafid sõnastikusüsteemi käsitsi.

Viimane etapp on sõnastiku avaldamine kas omaette veebilehel või sõnastikuportaali osana. Paberil tänapäeva sõnastikke enam üldjuhul ei avaldata – nii annab 2025. aasta seisuga Eesti Keele Instituut, kus Eestis sõnastikke peamiselt koostatakse, paberil välja veel vaid üksikuid sõnastikke, näiteks õigekeelsussõnaraamatut ÕS[6], „Eesti keele murrete sõnaraamatu" vihikuid ning piirkondlike murdesõnastike sarja[7]. Alates 2019. aastast koondatakse vähehaaval kogu sõnastikuinfo sõnastikuportaali Sõnaveeb[8] (Koppel jt 2019). Sõnaveebi sünni taga oli idee pakkuda keelekasutajale ühe päringuakna kaudu kogu infot, mis sõna kohta teada on. 2025. aasta seisuga on Sõnaveeb sõnastikuinfo koduks rohkem kui 160 andmebaasile, sh üld- ja oskuskeelesõnastikele.

Siinses näidisuurimuses kirjeldame samm-sammult korpuspõhise sõnastikuprojekti „Eesti keele naabersõnad 2019"[9] (Kallas, Koppel & Tuulik 2015) koostamise põhietappe. Projekti alguses oli plaanitud sõnastik avaldada omaette veebilehel, kuid Sõnaveebi tulekuga ilmus see naabersõnade plokina ka EKI ühendsõnastiku osana (vt joonist 1). Sõnastiku allikana kasutasime eesti keele ühendkorpust 2013, mis projekti alustamise hetkel (2014) oli mahukaim eesti keele korpus. Korpusandmete analüüsiks ja sõnastiku andmebaasi automaatseks loomiseks kasutasime korpusanalüüsi tarkvara Sketch Engine erinevaid funktsioone, sõnastikusüsteemina kasutasime EKI-s loodud veebipõhist programmi EELex.

Järgnevalt kirjeldame naabersõnade sõnastiku koostamist alates andmete kogumisest, töötlemisest ja analüüsist kuni andmebaasi loomise ja sõnastiku ilmumiseni.

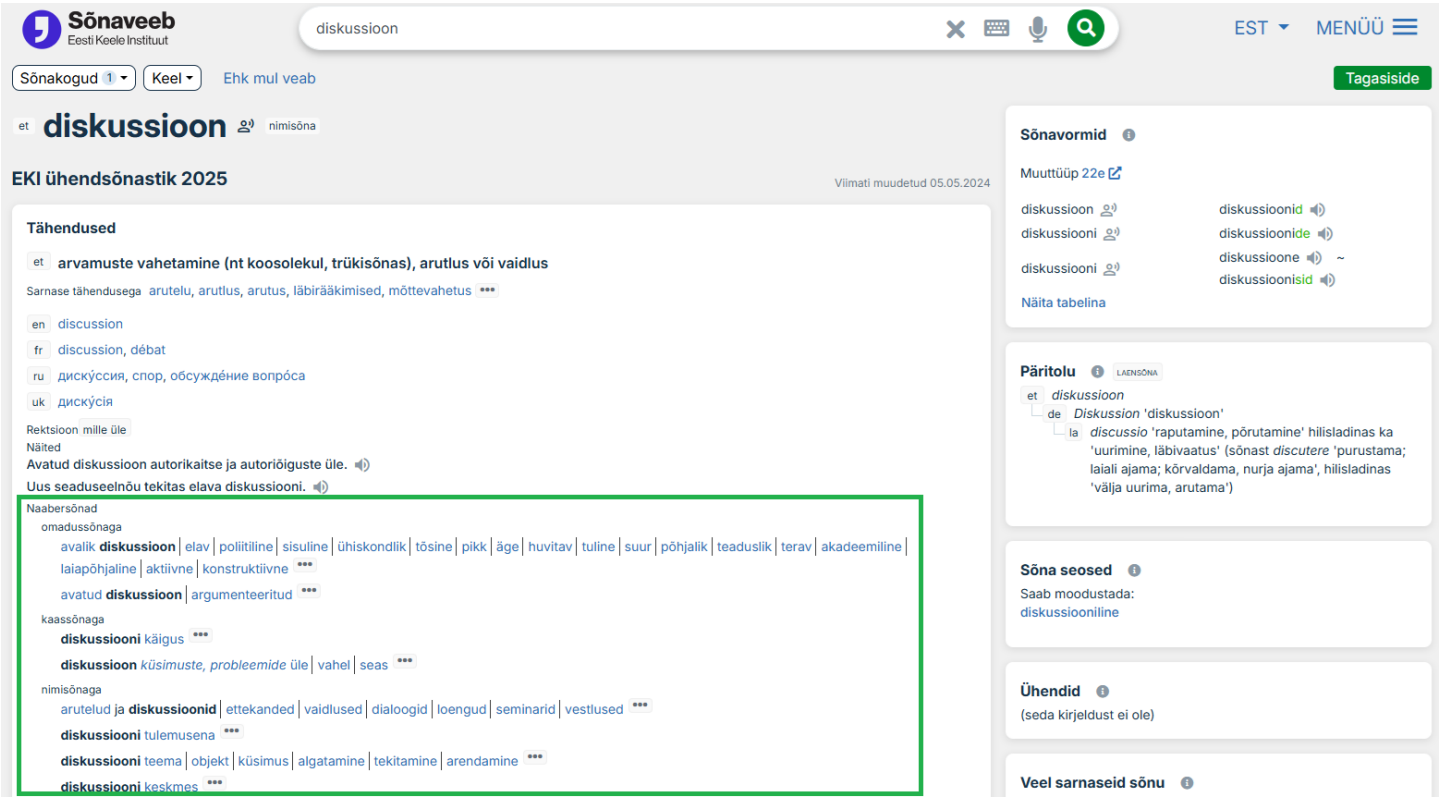
---

6 https://sonaveeb.ee/os

7 https://eki.ee/keeleinfo/sonastikud/

8 https://sonaveeb.ee/

9 *Naabersõnad* on termini *kollokatsioon* omakeelne vaste, mille võtsime kasutusele „Eesti keele naabersõnade 2019" sõnastiku avalikustamisel termini läbipaistvuse parandamiseks. Siinses näidisuurimuses kasutame naabersõnade asemel siiski terminit *kollokatsioon* (nt *ere päike* ja *kuum päike*), mille moodustajad on vastavalt kollokatsiooni põhi (*päike*) ja kollokaat (*ere*, *kuum*).



Joonis 1. Sõna *diskussioon* naabersõnad EKI ühendsõnastikus

2. Keeleandmete leksikograafiline analüüs Sketch Engine'is

Tänapäeval on korpusandmed igasuguse leksikograafilise andmebaasi allikas. Suurte korpusandmete analüüsimiseks on sõnastiku koostajatel kasutada spetsiaalsed tööriistad, mille hulgas on ka spetsiaalselt sõnastike koostamist toetav tasuline süsteem Sketch Engine (Kilgarriff jt 2004; Kilgarriff jt 2014), mis on Euroopa sõnastike koostajate seas laialdaselt kasutusel. See sisaldab muude keelte korpuste kõrval (2025. aasta seisuga umbes 110 keelt ja 900 korpust) ka eesti keele ühendkorpuste sarja (Koppel & Kallas 2022).

Sketch Engine'i funktsioonide toel saab analüüsida, tuvastada ja/või genereerida:

- märksõnaloendeid ning sõnade ja sõnaühendite esinemissagedust (sõnaloendi funktsioon *Wordlist*, *N-gram*);
- grammatilist kasutusinfot, näiteks mis sõnaliiki sõna kuulub, kas see esineb ainsuses ja/või mitmuses, kas kõikides käänetes või ainult ühes kindlas käändevormis, kas eitavas või jaatavas kõnes, kui sagedad on selle sõnavormid jmt (konkordantsifunktsioon *Concordance*);
- tähendust (konkordantsifunktsioon *Concordance*) ja tähendusjaotust (tähendusvihje funktsioon *Word Sense Induction*);
- kollokatsioone ning nende žanrilist ja temaatilist kuuluvust (sõnavisandi funktsioon *Word Sketch*);
- näitelauseid (konkordantsifunktsioon *Concordance*, heade näitelausete filter GDEX);
- leksikaal-semantilisi seoseid: sünonüüme, antonüüme jm sarnaseid sõnu (sarnaseid sõnu tuvastav funktsioon *Thesaurus*, Sketch Engine'ist eraldiseisev sõna vektoresitus *Embedding Viewer*[10]);
- oskussõnu (võtmesõnade ja terminite tuvastamise funktsioon *Keywords and Term extraction*);
- tõlkevasteid (paralleelkonkordantsi funktsioon *Parallel Concordance*);
- sõna kasutusmuutusi ajas (trendide funktsioon *Trends*, ajajoon *Timeline*).

Sketch Engine'i uusim funktsioon on ühekliki-sõnaraamat *OneClick Dictionary*[11] (Jakubíček jt 2018; Stemle, Abel & Lyding 2019; Jakubíček, Kovář & Rambousek 2022), mis võimaldab automaatselt luua sõnastiku andmebaasi. Ühekliki-sõnaraamatu funktsiooni abil saab ettemääratud sõnastiku üksused (nt märksõnad, kollokatsioonid, näitelaused) kanda otse korpusanalüüsi tarkvarast sõnastikusüsteemi, kus sõnastiku koostaja automaatselt loodud sisu kontrollib, puhastab ja toimetab. Nii on loodud näiteks sloveeni keele leksikograafiline andmebaas[12] ja suur

[10] https://embeddings.sketchengine.eu/

[11] https://www.sketchengine.eu/guide/

[12] http://eng.slovenscina.eu/spletni-slovar

hollandi keele sõnaraamat[13]. Sketch Engine'i põhifunktsioone ja nende eesti keele mooduleid on põhjalikult kirjeldatud (Kallas, Tuulik & Jürviste 2012; Kallas 2013; Kallas, Koppel & Tuulik 2015; Kallas, Suchomel & Khokholova 2017; Koppel 2020).

Naabersõnade sõnastiku andmebaasi automaatsel loomisel kasutasime selliseid Sketch Engine'i funktsioone nagu *Wordlist* ja *Word Sketch* ning heade näitelausete filtrit GDEX, mida alljärgnevalt kirjeldame.

2.1. Märksõnastiku automaatne koostamine

Märksõnastiku suurus sõltub sõnastiku sihtgrupist. Näiteks sisaldab eesti keele kui teise keele õppijale suunatud „Eesti keele põhisõnavara sõnastik"[14] (2014) 5000 märksõna, keeleteadlastele ja keelehuvilistele suunatud „Eesti keele sõnaperede"[15] (2012) sõnaraamat 9000 märksõna, kõigile eesti keele kasutajatele mõeldud „Eesti õigekeelsussõnaraamat ÕS 2025"[16] ligikaudu 60 000 märksõna ning EKI ühendsõnastik[17] ligikaudu 175 000 eesti keele märksõna (2025. aasta seisuga).

Märksõnastiku korpuspõhisel koostamisel on oluline leida sobiva sisu ja suurusega korpus. Näiteks kui tegemist on oskuskeele sõnastikuga, siis on märksõnastiku koostamise eeltingimus vastava valdkonnakorpuse olemasolu. Aga ka üldsõnastike puhul tasub arvesse võtta seda, mis tekste korpus sisaldab – näiteks kas seal on piisavas mahus ilukirjandust, suulist keelt või sisaldab korpus üksnes internetikeelt ja tarbetekste.

Naabersõnade sõnastiku andmebaasi automaatsel loomisel kasutasime tollal suurimat eesti keele korpust, milleks oli eesti keele ühendkorpus 2013 (563 mln sõnet). See sisaldas 2013. aastal kogutud veebikorpust ja Tartu ülikoolis koostatud eesti keele koondkorpust[18] ja tasakaalus korpust[19].

Märksõnastiku koostamiseks kasutasime Sketch Engine'i sõnaloendi funktsiooni *Wordlist*. Selle abil saab loendeid luua näiteks lemmade, sõnavormide, sõnaliigi, teatud grammatiliste tunnuste alusel ja regulaaravaldiste abil. Neid kriteeriume saab ka omavahel kombineerida (vt joonist 2). Mittesõnalist materjali (numbrid, kirjavahemärgid) saab loenditest automaatselt välja lülitada. Lisaks saab piirata sõnade (minimaalset ja maksimaalset) esinemissagedust, mis aitab loendist välja jätta näiteks väga madala sagedusega sõnad.

[13] https://anw.ivdnt.org/search

[14] https://arhiiv.eki.ee/dict/psv/

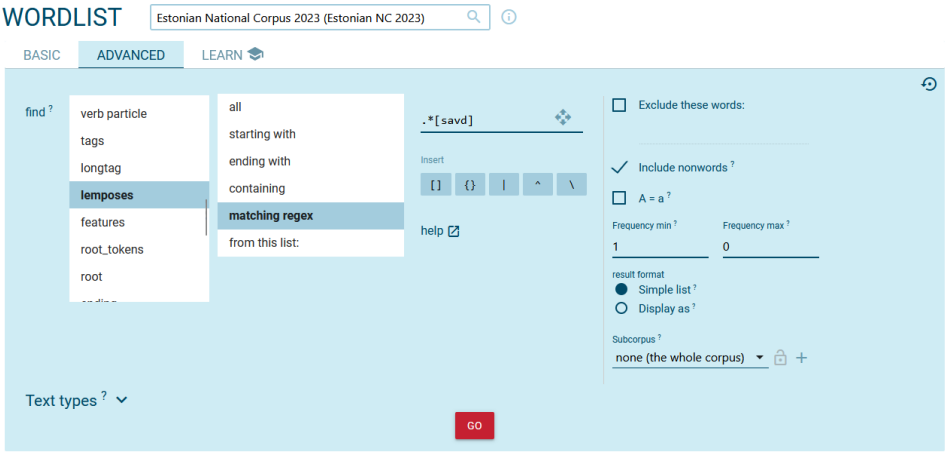
[15] https://arhiiv.eki.ee/dict/sp/

[16] https://sonaveeb.ee/os

[17] https://sonaveeb.ee/

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

[19] https://cl.ut.ee/korpused/grammatikakorpus/



Joonis 2. Sketch Engine'i sõnaloendite funktsiooni päringuaken

Joonisel 2 on kuvatõmmis sõnaloendite funktsiooni päringuaknast Sketch Engine'is, mille seadistus vastab naabersõnade sõnastiku märksõnastiku kokkupanemise aluseks olevatele parameetritele. Sagedasemate sisusõnade leidmiseks kasutasime otsinguatribuudina lemma ja sõnaliigi kombinatsiooni (*lemposes*). Regulaaravaldisega (*matching regex*) määrasime, et loendisse tulevad ainult nimisõnad (*s*), omadussõnad (*a*), määrsõnad (*d*) ja tegusõnad (*v*) (päring `.*[savd]`). Seejuures jäi sõna minimaalsageduseks (*Frequency min*) 1. Allikana kasutasime kogu korpust (*the whole corpus*). Seejärel laadisime päringu tulemused CSV-vormingus alla[20]. Kuna korpuse põhjal automaatselt loodud sõnaloend sisaldas trükivigu, märgenduse vigadest tingitud müra jm, tuli see käsitsi üle kontrollida. Näiteks eemaldasime sagedusloendist vigaseid sõnavariante (*mänedzher*, *shokk*), lühendeid (*eek*, *eur*, *toim*), pärisnimesid (*Lõunaleht*, *Suurhall*), sõnatüvest lahku kirjutatud liitsõna järelosiseid (*sugune*, *keelne*), vigaseid liitsõnu (nt *lapseoode* (valesti lemmatiseeritud määrsõna *lapseootel*), *minumeel* (ekslikult kokkukirjutatud ja valesti lemmatiseeritud ühend *minu meelest*)) ja termineid (*süsinikdioksiid*). Käsitsi üle kontrollitud loendist (u 11 000 sagedasemat sõna) saigi naabersõnade sõnastiku märksõnastiku alus. Koostamise käigus lisasid leksikograafid käsitsi automaatselt tuvastatud tegusõnadele (nt *lülitama*, *paistma*, *andma*) lisaks nende sagedasemaid ühend- ja väljendverbe (nt *sisse lülitama*, *välja lülitama*, *silma paistma*, *andeks andma*).

Järgmine etapp oli määrata parameetrid märksõnadele kollokatsioonide ja näitelausete tuvastamiseks.

[20] Tulemusi saab alla laadida veel TXT-, XML- ja PDF-vormingus.

2.2. Kollokatsioonide tuvastamine

Kollokatsioon on sõnaühend, mis moodustub sõnadest (kollokatsiooni põhjast ja kollokaadist), mis esinevad keeles sageli koos ja mille tähendus on teda moodustavate sõnade tähenduste summa (nt *ere*/*kuum päike*; *konn krooksub*/*hüppab*; *peole minema*/*jõudma*). Kollokatsiooni võib defineerida mitmeti (loe lähemalt ptk 5.2.5 „Kollokatsioonid"), kuid traditsioonilises leksikograafias (sh siin peatükis) vastandatakse kollokatsiooni püsiühendile, mida inimese mälus talletatakse tervikuna ja mille tähendus ei ole teda moodustavate sõnade (püsiühendi komponentide) tähenduste summa. Püsiühendite alla kuuluvad näiteks väljendverbid (*päikest võtma*, *rahas suplema*, *ämbrisse astuma*), ühendverbid (*ette ütlema*, *alla käima*) ja kinnisväljendid (*vesi ahjus*, *au ja uhkus*), mis vajavad sõnastikes omaette artiklit koos seletus(t)e, näidete, sünonüümide ja muuga.

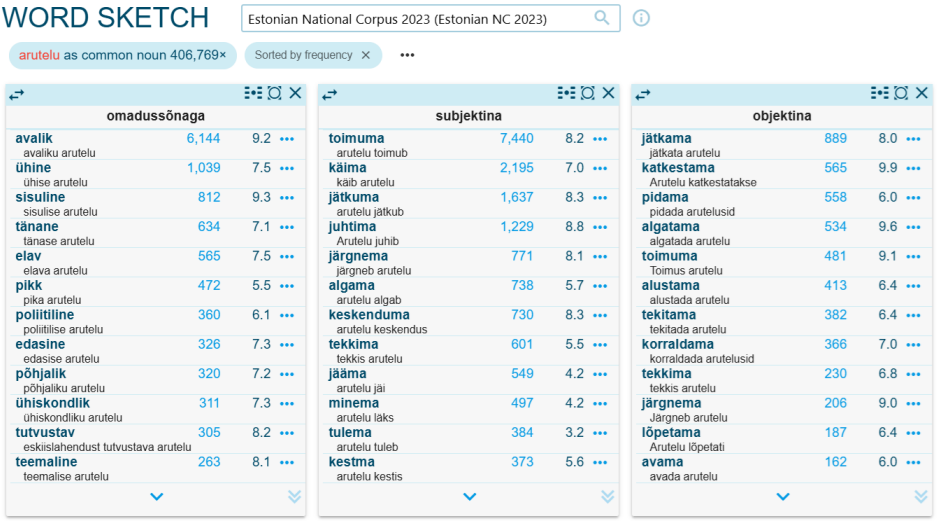
Kollokatsioonide tundmine on vajalik, et loomulikult ja ladusalt rääkida ning kirjutada. Emakeeles on kollokaatide valik pigem intuitiivne, kuid keeleõppijad nende olemasolust sageli teadlikud ei ole, mistõttu võivad nad ühe sõna naabriks valida sellise paarilise, mis nende emakeeles konteksti sobib, aga mis õpitavas keeles kõlab kummaliselt. Näiteks öeldakse eesti keeles *kange kohv* ja *lahja kohv*, inglise keeles hoopis *tugev kohv* ehk *strong coffee* ja *nõrk kohv* ehk *weak coffee*. Seega on keeleõppijal kasulik keelt õppida n-ö valmis tükkide kaupa, et tema keelekasutus oleks võimalikult sarnane emakeelse kõneleja omaga.

Varasemates eesti keele sõnastikes ei olnud traditsiooni kollokatsioone süstematiseeritult rühmade kaupa välja tuua, samas leidus neid mitmete sõnastike näitelausete hulgas. Näiteks „Eesti keele seletavas sõnaraamatus" (2009) on sõna *pidu* juures näitena esitatud tavapärased ühendid (kollokatsioonid) *pidu pidama, korraldama* ning „Eesti õigekeelsussõnaraamatus ÕS 2018" näide *Peame pidu, korraldame v teeme peo*. Selle lünga täitmiseks koostatigi „Eesti keele naabersõnade 2019" sõnastik, mis ilmus nii omaette veebilehena kui ka EKI ühendsõnastiku osana Sõnaveebis, kus kollokatsioonid on esitatud naabersõnade plokis süstemaatiliselt rühmade kaupa ning omakorda varustatud näidetega.

Kollokatsioonide tuvastamiseks kasutasime Sketch Engine'i sõnavisandi funktsiooni *Word Sketch* (Kilgarriff jt 2004). Sõnavisandi funktsiooni kasutamine eeldab keelespetsiifilist sõnavisandite grammatikat Word Sketch Grammar[21], mis määrab, milliseid kollokatsioonirühmi programm otsima hakkab. Eesti keele sõnavisandite grammatika aluseks on Jelena Kallase (2013) välja töötatud eesti keele kollokatsioonide tüpoloogia, vt koondtabelit (Kallas, Koppel & Tuulik 2015: 81–82), mis on omakorda koostatud eesti keele grammatika käsitluste alusel (Rätsep 1978; Tauli 1980; Erelt jt 1993; Erelt, Erelt & Ross 1997; Kerge 2001). Eesti keele sõnavisandite grammatika versioonis 2.1 (loe lähemalt Koppel & Kallas 2022) on kokku 113 reeglit.

[21] https://www.sketchengine.eu/my_keywords/word-sketch-grammar/

Joonisel 3 on näha nimisõna *arutelu* kolme kollokatsioonirühma, kus *arutelu* esineb a) koos omadussõnaga (*avalik arutelu*), b) subjektina koos tegusõnaga (*arutelu toimub*), c) objektina koos tegusõnaga (*arutelu jätkama*). Kollokaadid on esitatud sageduse järjekorras.



Joonis 3. Nimisõna *arutelu* sõnavisandi väljavõte eesti keele ühendkorpusest 2023

Sõnavisandi statistiline analüüs näitab, et sõna *arutelu* (kollokatsiooni põhi) esineb kõige sagedamini koos omadussõnadega *avalik*, *ühine* ja *sisuline* ning tegusõnadega *toimuma*, *käima*, *jätkuma* ja *jätkama*, *katkestama*, *pidama* (kollokaadid). Joonisel 3 on iga kollokaadi juures näha, mis vormis see kõige sagedamini esineb, nt *arutelu toimub*, *arutelu katkestatakse* jne. Näeme ka märgenduse vigadest põhjustatud vigu kollokaadirühma määramisel: *arutelu* on ekslikult määratud verbi *juhtima* subjektiks ja verbide *toimuma, tekkima* ja *järgnema* objektiks.

2.3. Näitelausete tuvastamine

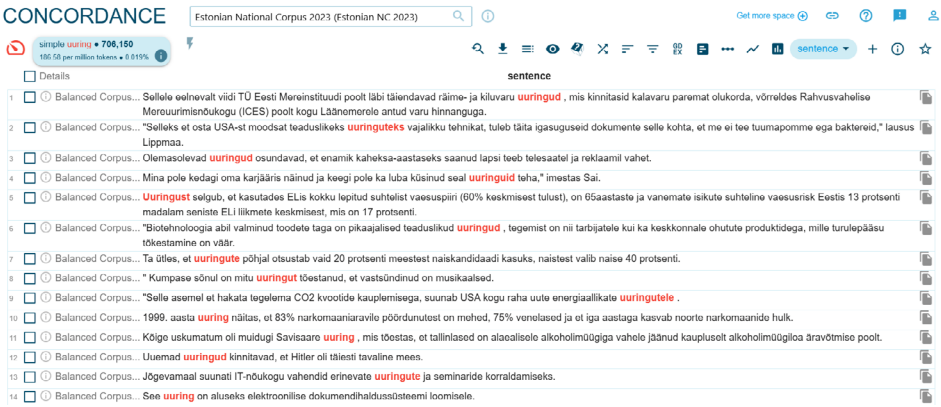
Näitelause on sõnaartikli oluline osa. See aitab sõna seletust paremini mõista ning illustreerib märksõna kasutust nii sisu kui ka vormi osas. Näitelause tõestab, et sõna keeles eksisteerib, andes aimu sõna ümbrusest nii süntaksi, kollokatsioonide kui ka stiili mõttes. Hea näide peaks vahendama ühelt poolt vaatlusaluse sõna tüüpilist, teiselt poolt aga eripärast tähendust.

Näitelause võib olla kas autentne (nt võetud otse korpusest) või kunstlik ehk sõnastiku koostaja koostatud. Autentne lause võib omakorda olla kas täiesti

autentne või autentse lause lühendatud versioon. Tänapäeva leksikograafias kasutatakse näitelausete allikana aina mahukamaid tekstikorpusi.

Näitelausete tuvastamiseks on Sketch Engine'is kaks võimalust: 1) konkordantsi funktsioon *Concordance*, 2) nn heade näitelausete filter GDEX (*Good Dictionary Examples*) (Kilgarriff jt 2008; Kosem jt 2019). GDEX-i filtri loomise algne eesmärk oli eelkõige vähendada sõnastiku koostajate ajakulu näitelausete valimisel. Selle keskmes on reeglipõhine valem, mis valib määratud parameetrite põhjal kõikidest korpuses esinevatest lausetest välja need, mis oma struktuuri ja sisu poolest leksikograafiliseks analüüsiks kõige paremini sobivad. Need parameetrid ütlevad näiteks seda, et lause peab algama suure tähega, lõppema lauselõpumärgiga ja sisaldama pöördelist verbivormi (eesti keele parameetrite kohta loe lähemalt Koppel 2017; Koppel 2020). GDEX hindab igat korpuslauset ning reastab väljundina neist loendi n-ö paremuse järjekorras, kus parimad näitelause kandidaadid on nimekirja eesotsas.

Järgnevalt kirjeldame näitelausete automaatset tuvastust eesti keele ühendkorpusest 2023 sõna *uuring* näitel. Konkordantsiridade lugemiseks on Sketch Engine'is kaks vaadet – KWIC-vaade ehk märksõna kontekstis (*keyword in context*) ning lause vaade (*sentence*). Näitelausete valimise eesmärgil on konkordantsiridu ilmselt mugavam lugeda lause vaate abil (vt joonist 4).

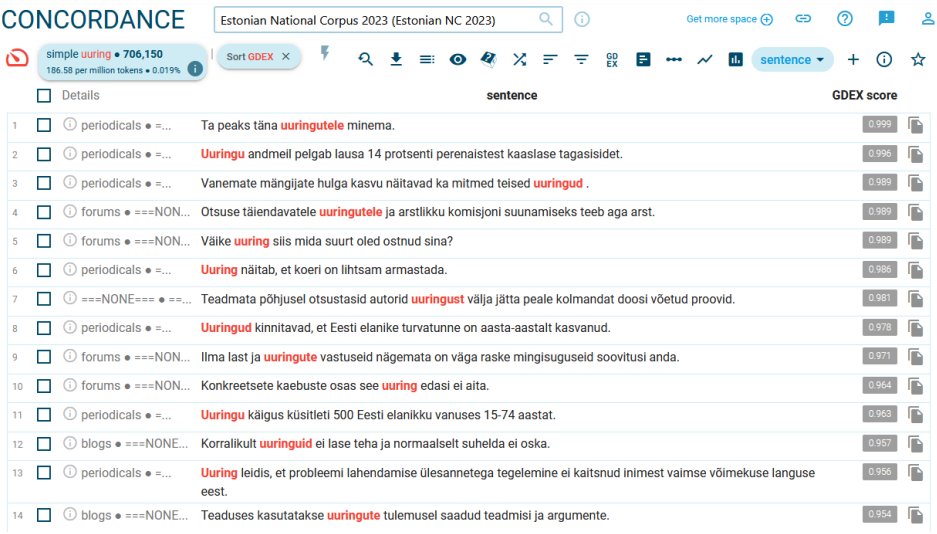


Joonis 4. Märksõna *uuring* konkordantsiread

Jooniselt 4 on näha, et korpuslaused võivad olla küllalt pikad. Selliste lausete läbi lugemiseks võib sõnastiku koostajal kuluda väga palju aega, enne kui ta leiab sealt sellise, mida sobiks otse sõnastiku näitelauseks võtta. Lisaks on oluline märkida, et kui GDEX-i filtrit pole sisse lülitatud (nagu joonisel 4), siis kuvatakse korpuslauseid vaikimisi alati alamkorpuste järjekorras, st vanematest tekstidest esimesena. Kuna ühendkorpuses on vanimad alamkorpused alati tasakaalus korpus (*Balanced*

Corpus) ja koondkorpus (*Reference Corpus*), pärinevad esimesed näited just sealt. Kui uurimistöö jaoks on vaja **juhuslikke näiteid**, tuleks selle saamiseks kasutada funktsioone juhuvalim (*Get a random sample*) või ümberpaigutus (*Shuffle lines*). Juhuvalim vähendab kuvatavate korpuslausete arvu, säilitades samal ajal valimi representatiivsuse – laused valitakse juhuslikult kõigist korpuse osadest, kusjuures lausete jaotus on lähedane üldkogumi jaotusega. Ümberpaigutuse funktsioon muudab samuti lausete järjekorda juhuslikult, kuid ei säilita valimi representatiivsust. Mõlemad funktsioonid annavad alati sama tulemuse. See on taotluslik, selleks et erinevad kasutajad (nt õppejõud ja üliõpilased) saaksid samu samme pidi jõuda täpselt sama tulemuseni. Teistsuguse valimi saamiseks tuleks kasutada teistsugust arvu (nt valida valimisse 200 lause asemel 201 lauset).

Selleks, et Sketch Engine pakuks automaatselt lühemaid ning süntaktiliselt ja grammatiliselt lihtsamaid näitelause kandidaate, tuleb sisse lülitada GDEX-i filter konkordantsiridade kohal (vt joonis 4). See võtab korpusest vaikimisi 300 juhuslikku lausetja hindab nende sobivust sõnastiku näitelauseks eesti keele jaoks välja töötatud hea näitelause parameetrite abil[22]. Soovi korral saab valimit suurendada või vähendada, samuti võib sisse lülitada GDEX-i skoori, mida kuvatakse kasutajale konkordantside järel paremas veerus (vt joonis 5). GDEX-i skoor on arvuline väärtus, mis jääb 0 ja 1 vahele ning see illustreerib seda, kui hästi korpuslaused hea näitelause parameetritele vastavad – mida kõrgem skoor, seda parem (potentsiaalne) näitelause kandidaat.

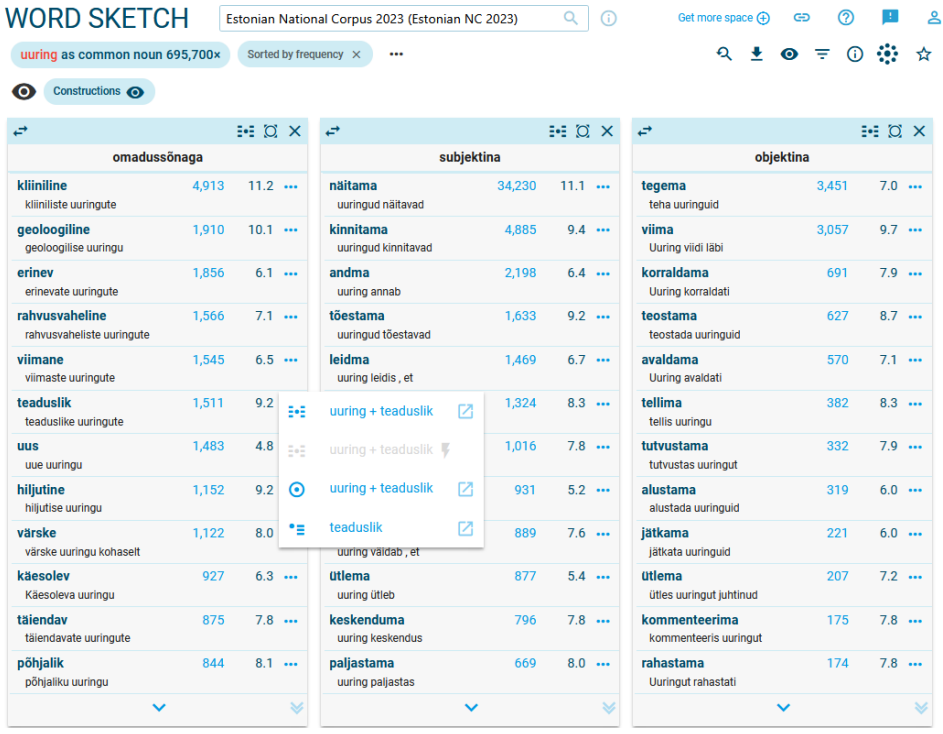


Joonis 5. GDEX-i filtriga valitud *uuringu* korpuslaused

[22] 2025. a seisuga kasutatakse Sketch Engine'is lausete hindamiseks vaikimisi GDEX-i eesti keele mooduli versiooni 1.4, mida kirjeldatakse lähemalt artiklis (Koppel 2017).

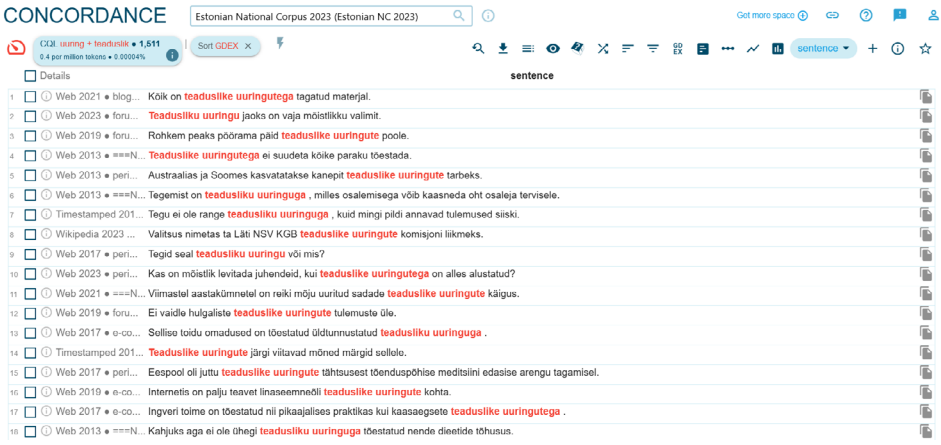
Jooniselt 5 on näha, et GDEX-i filtri abil pakutud korpuslaused on oluliselt lühemad ning grammatiliselt ja süntaktiliselt vähemkeerukad.

GDEX-i filtri saab sisse lülitada ka sõnavisandite lehel. Sealt tuleb esmalt valida üks kollokatsioon, näiteks *teaduslik uuring*, mida illustreerivaid näiteid saab näha konkordantsidena, vajutades kollokatsiooni järele olevale kolmele punktile (vt joonis 6).



Joonis 6. Sõna *uuring* sõnavisand

Sõnavisandite kaudu konkordantsiridu pärides kuvatakse need vaikimisi samuti alamkorpuste järjekorras, st vanematest allikatest esimesena. Ka siin saab sisse lülitada GDEX-i filtri, mis valib korpusest juhuvalimi lausetest, mille vastamist hea näitelause parameetritele omakorda GDEX-i skooriga hinnatakse (vt joonis 7).



Joonis 7. GDEX-i filtri abil valitud näited kollokatsioonile *teaduslik uuring*

Eelnevalt kirjeldatud kolme Sketch Engine'i funktsiooni kasutasime naabersõnade sõnastiku andmebaasi automaatsel loomisel, mida kirjeldame järgmises peatükis.

3. Andmebaasi loomine sõnastikusüsteemis

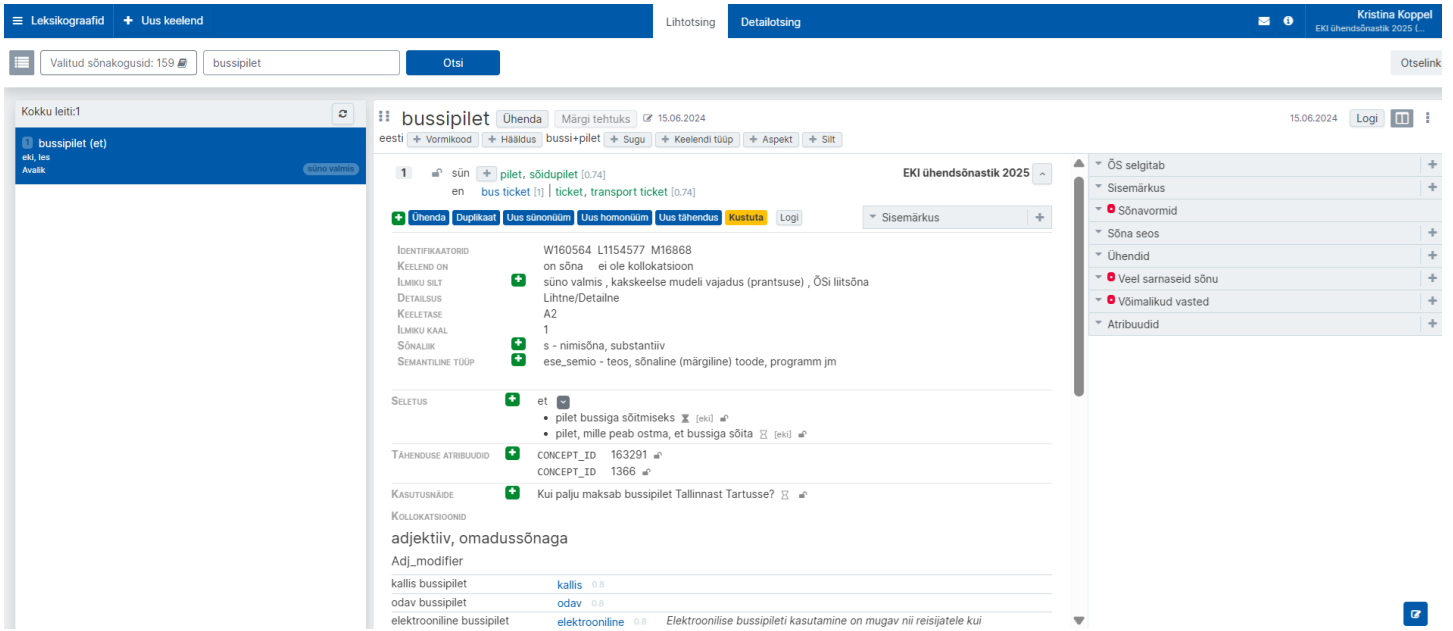
EKI-s on üldsõnastike jaoks kasutusel olnud kaks sõnastiku- (ja terminibaasi) süsteemi: EELex (Langemets, Loopmann & Viks 2006; Jürviste jt 2011) ja Ekilex (Tavast jt 2018; Langemets jt 2021). Oskuskeelesõnastikke on koostatud peale Ekilexi veel Multitermis[23]. Kuna EELex põhines Microsofti Internet Exploreri veebibrauseril, mida enam ei toetata, on 2019. aastast põhiliselt kasutusel olnud Ekilex (vt joonis 8)[24]. Ekilexi saavad kasutajakonto luua kõik keelehuvilised, seda kas lihtsalt andmete vaatamise või oma andmebaasi loomise ja avaldamise eesmärgil (Ekilexis koostatud sõnastikud avaldatakse keeleportaalis Sõnaveeb).

Euroopas on laialdaselt kasutusel ka vabavaraline sõnastikusüsteem Lexonomy[25] (Měchura 2017; Jakubíček jt 2018), kus saab samuti oma sõnastikke koostada ja avaldada, kuid kui tegemist on eesti keelt sisaldava andmebaasiga, soovitame kindlasti kasutada Ekilexi. Kuna Ekilex koondab infot sõnade ja terminite kohta paljudest sõnakogudest ja terminibaasidest, võimaldab see oma andmeid

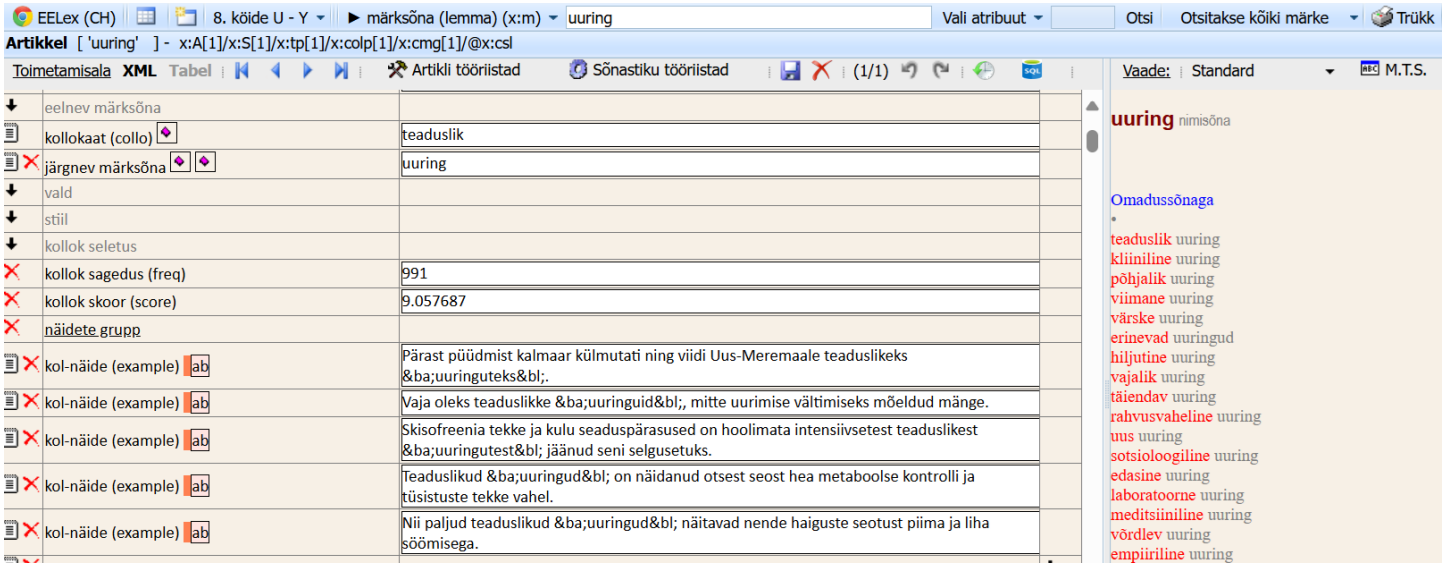
[23] https://www.trados.com/products/multiterm-desktop/

[24] EELexis koostatakse 2025. a seisuga veel näiteks „Eesti murrete sõnaraamatut", „Akadeemilist etümoloogiasõnaraamatut", „Saksa laenude sõnaraamatut" ja „Eesti perekonnanimeraamatut".

[25] https://www.lexonomy.eu/



Joonis 8. Märksõna *bussipilet* sõnastikusüsteemis Ekilex



Joonis 9. Märksõna *uuring* sõnastikusüsteemis EELex

võrrelda teiste üld- ja oskuskeelesõnastikes esitatud andmetega ning andmete esitust ühtlustada. Iga autor või töörühm täiendab ühiskasutatavaid andmeid oma infoga, näiteks lisab uusi sõnu või termineid, sünonüüme, tõlkevasteid või ajakohastab terminikirjeid. Ekilex aitab vähendada sõnakogude andmete dubleerimist, koostamisel tekkivaid püsivigu ning lihtsustab tehtu avalikustamist.

Naabersõnade sõnastiku andmebaasi loomise hetkel oli EKI-s sõnastikusüsteemina kasutusel veel XML-põhine EELex, nii võtsime vajalikud andmed Sketch Engine'ist XML-faili kujul. Märksõnastiku moodustasid sõnaloendi funktsiooni *Wordlist* abil tuvastatud sagedasemad nimisõnad, omadussõnad, määrsõnad ja tegusõnad (nagu kirjeldasime jaotises 2.1). Märksõnade kollokatsioonirühmad tuvastasime sõnavisandi funktsiooni *Word Sketch* abil (vt 2.2) ning iga kollokatsiooni jaoks võtsime korpusest viis GDEX-i filtriga valitud korpuslauset (vt 2.3). Automaatselt loodud andmebaas sisaldas 10 939 märksõna, 82 678 kollokatsioonirühma, 493 971 kollokaati ning 2 469 855 näitelauset. Järgmise etapina järeltoimetasid ja täiendasid sõnastiku koostajad seda automaatselt loodud sisu juba sõnastikusüsteemis EELex.

Joonisel 9 on kuvatõmmis sõnastikusüsteemist EELex, kus on avatud märksõna *uuring*. Paremal pool on sõnastiku veebivaade ning vasakul pool toimetamisala, kus sõnastiku koostaja sai automaatselt loodud sisu toimetada: kollokatsioone lisada, parandada, kustutada ning autentseid korpuslauseid valida. Toimetamisalal on näha ka kollokatsiooni *teaduslik uuring* jaoks eesti keele ühendkorpusest 2013 tuvastatud viit autentset näitelauset. Iga kollokatsioonirühma jaoks valisime välja vähemalt ühe illustreeriva näite, mida vajadusel toimetasime.

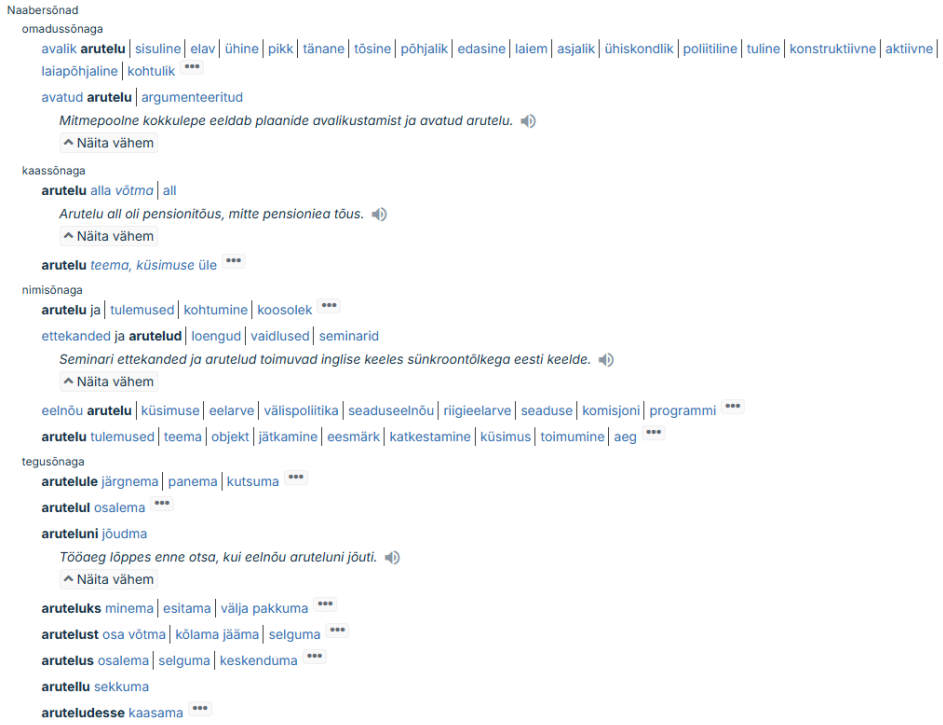
4. Sõnastiku avaldamine

Tänapäeva sõnastikusüsteemid võimaldavad reeglina esitada andmebaasi sisu otse veebisõnastikuna. Ekilexi väliskasutajaliides on Sõnaveeb, kus agregeeritud kujul näidatakse eri sõnakogude artikleid. EELexil seda funktsionaalsust ei olnud, mistõttu pidi tekitama iga sõnakogu jaoks eraldi andmebaasi ning avaldama selle omaette veebilehel. Kui naabersõnade sõnastiku automaatselt loodud andmebaas EElexis toimetatud sai, ilmus see nii omaette veebilehel[26] kui ka EKI ühendsõnastiku osana Sõnaveebis (vt nt EKI ühendsõnastikus *arutelu*[27] juures rubriiki „Naabersõnad", täiel kujul avatuna vt joonis 10).

Jooniselt 10 on näha, et sõnaartiklis *arutelu* on neli kollokatsioonirühma (omadussõnaga, kaassõnaga, nimisõnaga, tegusõnaga) ja et igat kollokatsioonirühma illustreerib vähemalt üks näitelause. Näitelaused on esialgu peidus, et kasutaja saaks esmalt tervikliku pildi kõikidest sõnapaaridest. Näitelausete nägemiseks tuleb vajutada kollokatsioonide rida lõpetavale kolmele punktile („Näita rohkem").

[26] https://arhiiv.eki.ee/dict/kol/

[27] https://sonaveeb.ee/search/unif/dlall/dsall/arutelu/1/est



Joonis 10. Märksõna *arutelu* kollokatsiooniplokk koos näitelausetega EKI ühendsõnastikus

Kokkuvõte

Näidisuurimuses kirjeldasime ühe konkreetse korpuspõhise sõnastiku – „Eesti keele naabersõnad 2019" – koostamise põhietappe: 1) sõnastikuprojekti planeerimist, sh korpusanalüüsi tarkvara, sõnastikusüsteemi ja alusandmete, eelkõige korpuse valikut, 2) korpusandmete analüüsi (märksõnastiku loomist, kollokatsioonide ja näitelausete tuvastamist), 3) andmebaasi loomist ja toimetamist sõnastikusüsteemis ning 4) sõnastiku avalikustamist.

Korpusanalüüsi tarkvara ja sõnastikusüsteemid on muutnud sõnastike koostamise kordades kiiremaks ning tänu korpusandmetele on sõnastike sisu usaldusväärne. Leksikograafia on väga kiiresti arenev valdkond ning kaasaegsed korpusanalüüsi meetodid võimaldavad juba praegu erinevaid sõnastikuüksusi automaatselt tuvastada: peale märksõnade, kollokatsioonide ja näitelausete veel ka näiteks definitsioone, sünonüüme, antonüüme, oskussõnu ja tõlkevasteid. Praegu oleme niisiis jõudnud automaatse leksikograafia ajajärku, mida siinne näidisuurimus ka näitlikustas.

Kirjandus

De Schryver, Gilles-Maurice. 2023. Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography* 36(4). 355–387. https://doi.org/10.1093/ijl/ecad021.

Erelt, Mati, Tiiu Erelt & Kristiina Ross. 1997. *Eesti keele käsiraamat*. Tallinn: Eesti Keele Sihtasutus.

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

Jakubíček, Miloš, Vojtěch Kovář & Adam Rambousek. 2022. *D4.7 Evaluation and assessment of methods for automatic drafting*. https://elex.is/wp-content/uploads/ELEXIS_D4_7_Evaluation_and_assessment_of_methods_for_automatic_drafting_of_lexicographic_resources.pdf.

Jakubíček, Miloš, Michal Měchura, Vojtech Kovář & Pavel Rychlý. 2018. Practical post-editing lexicography with Lexonomy and Sketch Engine. Jaka Čibej, Vojko Gorjanc, Iztok Kosem & Simon Krek (toim), *XVIII EURALEX International Congress: Lexicography in Global Contexts*, 65–67. Ljubljana: Ljubljana University Press, Faculty of Arts.

Jürviste, Madis, Jelena Kallas, Margit Langemets, Maria Tuulik & Ülle Viks. 2011. Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. Iztok Kosem & Karmen Kosem (toim), *eLexicography in the 21st Century: New Applications for New Users, Proceedings of eLex*, 106–112. https://elex2011.trojina.si/elex2011_proceedings.pdf.

Kallas, Jelena. 2013. *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias* (Humanitaarteaduste dissertatsioonid 32). Tallinn: Tallinna Ülikool.

Kallas, Jelena, Kristina Koppel & Maria Tuulik. 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 11. 75–94. https://doi.org/10.5128/ERYa11.05.

Kallas, Jelena, Vit Suchomel & Maria Khokholova. 2017. Automated identification of domain preferences of collocations. Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek & Vít Baisa (toim), *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference*, 309–320. Brno: Lexical Computing CZ s.r.o.

Kallas, Jelena, Maria Tuulik & Madis Jürviste. 2012. Leksikograafilise tarkvara Sketch Engine eesti keele moodul. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 3(2). 57–77. https://doi.org/10.12697/jeful.2012.3.2.03.

Kerge, Krista. 2001. *Eesti süntaks keeleõppe praktikule: käsiraamat*. Tallinn: TEA Kirjastus.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell & Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX International Congress*, 425–432. Documenta Universitaria Barcelona, Spain.

Kilgarriff, Adam, David Tugwell, Pavel Rychlý & Pavel Smrz. 2004. The Sketch Engine. Geoffrey Williams & Sandra Vessier (toim), *Proceedings of the 11th EURALEX International Congress*, 105–115. Lorient, France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Koppel, Kristina. 2017. Heade näitelausete automaattuvastamine eesti keele õppesõnastike jaoks. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 13. 53–71. https://doi.org/10.5128/ERYa13.04.

Koppel, Kristina. 2020. *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* (Dissertationes linguisticae Universitatis Tartuensis 38). Tartu: Tartu Ülikooli Kirjastus.

Koppel, Kristina & Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 18. 207–228. https://doi.org/10.5128/ERYa18.12.

Koppel, Kristina, Arvi Tavast, Margit Langemets & Jelena Kallas. 2019. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. Iztok Kosem & Tanara Zingano Kuhn (toim), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference. 1-3 October 2019, Sintra, Portugal*, 434–452. Brno: Lexical Computing CZ, s.r.o.

Kosem, Iztok, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit & Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. *International Journal of Lexicography* 32(2). 119–137. https://doi.org/10.1093/ijl/ecy014.

Langemets, Margit, Kristina Koppel, Jelena Kallas & Arvi Tavast. 2021. Sõnastikukogust keeleportaaliks. *Keel ja Kirjandus* 8–9. 755–770. https://doi.org/10.54013/kk764a6.

Langemets, Margit, Andres Loopmann & Ülle Viks. 2006. The IEL dictionary management system of Estonian. Gilles-Maurice De Schryver (toim), *DWS 2006: Proceedings of the 20 Fourth International Workshop on Dictionary Writing Systems. Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System. Turin, 5th September*, 11–16. Turin: University of Turin.

Měchura, Michal Boleslav. 2017. Introducing Lexonomy: An open-source dictionary writing and publishing system. Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek & Vít Baisa (toim), *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, 19–21. Brno: Lexical Computing CZ s.r.o.

Paet, Tiina & Lydia Risberg. 2021. Võõrsõnade tähendussoovitused ja nende esitus üldkeele sõnaraamatus. *Keel ja Kirjandus* 11. 965–984. https://doi.org/10.54013/kk767a2.

Risberg, Lydia. 2024. *Sõnatähendused ja sõnaraamat. Kasutuspõhine sisend eesti keelekorraldusele* (Dissertationes philologiae estonicae Universitatis Tartuensis 52). Tartu: Tartu Ülikooli Kirjastus.

Rätsep, Huno. 1978. *Eesti keele lihtlausete tüübid*. Tallinn: Valgus.

Stemle, Egon W., Andrea Abel & Verena Lyding. 2019. Language varieties meet one-click dictionary. Iztok Kosem & Tanara Zingano Kuhn (toim), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal.*, 537–546. Brno: Lexical Computing CZ, sro.

Tauli, Valter. 1980. *Eesti keele grammatika II. Lauseõpetus*. Uppsala: Uppsala: Finsk-ugriska institutionen, Uppsala University.

Tavast, Arvi, Margit Langemets, Jelena Kallas & Kristina Koppel. 2018. Unified data modelling for presenting lexical data: The case of EKILEX. Jaka Čibej, Simon Krek, Vojko Gorjanc & Iztok Kosem (toim), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, 749–761. Ljubljana, Slovenia: Ljubljana University Press, Faculty of Arts.

Vainik, Ene, Geda Paulsen & Ahti Lohk. 2021. Käändevormist sõnaks: mida näitab sagedus? *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 17. 285–307. https://doi.org/10.5128/ERYa17.16.

Veismann, Ann. 2021. Kas elliptiline (nimisõnata) kaassõnafraas on olemas? *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 12(1). 433–456. https://doi.org/10.12697/jeful.2021.12.1.12.

Korpuslingvistika ja ohustatud keeled

Joshua Wilbur

Lühikokkuvõte

Keeli, mille puhul on oht, et neid lähitulevikus enam ei räägita, nimetatakse ohustatud keelteks. Tavaliselt on ohustatuse põhjuseks see, et keele kõnelejad lähevad üle teisele keelele ega anna enam keelt edasi noorematele põlvkondadele. Kuna keeleteadus kui distsipliin püüab mõista, kuidas keel töötab, pole ohustatud keeled mitte ainult võrdväärsed ja õigustatud keeleteaduse uurimisobjektid (võrreldes suuremate, stabiilsete keeltega), vaid ideaalis nad peaksidki olema keeleteaduslike uurimuste osa, sealhulgas ka korpuslingvistiliste uurimuste osa. Seda eriti seetõttu, et ohustatud keelte korpus koosneb tavaliselt suulise registri tekstidest ja on seega keelekasutuse iseäranis esinduslik osa. Samal ajal seisab ohustatud keelte korpuslingvistikaga tegelemine silmitsi keerukate ülesannetega, kuna ohustatud keelte korpus (kui see üldse on olemas) on sageli piiratud suuruse ja katvusega ning tihti pole isegi selle koostamine toimunud järjekindlate põhimõtete alusel. Piiratusele vaatamata võib ohustatud keelte korpuslingvistika siiski olla tulemuslik, seda eriti juhul, kui olla teadlik võimalikest probleemidest. Ohustatud keelte tekste hoitakse sageli tarkvaras ELAN, mida kasutatakse keelte dokumenteerimisel ja mis toimib ka korpusliidesena. ELAN-i töö selgitamiseks ja ohustatud keelte korpusuuringu näiteks on järgnevalt toodud näidisuurimus kriitiliselt ohustatud Pite saami keelest. Uurimus käsitleb subjektiga viidatud entiteedi elusust, mis määrab verbimorfoloogias duaali kasutamist[1].

1. Ohustatud keelte korpuslingvistika

Keelt peetakse **ohustatuks**, kui on oht, et seda tulevikus enam ei räägita, sest nooremad põlvkonnad ei omanda keelt enam esimese keelena[2]. Kõige sagedamini muutub keel ohustatuks, kui täiskasvanud emakeelsed kõnelejad valivad lastega suhtlemiseks teise, tavaliselt domineeriva enamuse keele. Selline kogukonna laiem

[1] Uurimuse andmestiku leiab õpiku repositooriumist https://osf.io/xqzsf/.

[2] Austin ja Sallabank (2011) on üks täielikumaid teadusallikaid ohustatud keelte kohta.

keelevahetus (ingl *language shift*) toimub tavaliselt enamuse keele või keelte (näiteks ametliku riigi- või piirkondliku keele) domineerimise tõttu. Keelevahetus võib olla varjatud, näiteks siis, kui enamuse keelt peetakse praktilisemaks valikuks, mis pakub lastele tulevikus rohkem ja paremaid võimalusi. Kuid keelevahetus võib olla ka avalikult agressiivne, näiteks kui domineeriv keelerühm keelab inimestel kasutada nende vähemuskeelt[3].

Selleks, et mõista inimkeele olemust tegeliku keelekasutuse kaudu (mis on mis tahes korpuslingvistilise uurimise peamine eesmärk), tuleks kõiki inimkeeli pidada võrdselt põhjendatud ja asjakohasteks uurimisobjektideks; see tähendab, et uurida ei tule mitte ainult maailma mõnda ülekaalukalt domineerivat enamuse keelt, vaid ka juba väljasurnud keeli ja kõige väiksemaid, kõige ohustatumaid keeli, mille kõnelejad on alles vaid mõned – nii nagu seda tehakse käesolevas peatükis. Kuigi ohustatud keeli uuritakse vähem, võib nende uurimine oluliselt panustada keeltevaheliste üldistuste tegemisse. Just ohustatud keeled võivad olla eriti huvitava ja haruldase struktuuri ja kasutusviisiga, mis rõhutab inimkeelte mitmekesisust ja võimalikke keerukusi. See tähendab, et kuigi uurimisolukord on tavalisest keerulisem, saab korpuslingvistika ainult kasu ohustatud keelte uurimisest, nagu selgub ka käesolevast peatükist.

Kuna ohustatud keelte kõnelejaid on harilikult suhteliselt vähe, moodustavad nende kõnelejad tavaliselt kogukonna, mis on ühiskonnas vähemuses. Seetõttu puudub ohustatud keeltel sama suur poliitiline, praktiline ja finantsiline toetus kui enamuse keeltel. Sellest tuleneb kaks probleemi, mis muudavad korpuslingvistilise uurimise ohustatud keele puhul teistsuguseks kui suure keele korral: 1) kättesaadavate tekstide hulk ja 2) märgenduse kvaliteet. Siin peatükis näidatakse ja arutatakse, kuidas need need kaks tegurit võivad korpuslingvistilist uuringut mõjutada – nendega tuleb arvestada mitmes uurimistsükli punktis, alates korpuse loomisest, hüpoteeside testimisest ja tulemuste analüüsimisest. Lisaks käsitleme peatükis selliseid praktilisi aspekte nagu töötamine piiratud ja suulise registri andmetega, mis on saadud väikeselt kõnelejate rühmalt. Samuti seda, mil määral saab ELAN-it, mis on tavaliselt transkribeerimis- ja märgendustööriist, kasutada korpustööriistana. Et illustreerida mõnda neist raskustest, kuid ka näidata, kuidas korpusuuringud on täiesti võimalikud isegi väga ohustatud keelte puhul, esitatakse juhtumiuuring Pite saami keelest (Rootsis kõneldav Uurali keel), kus subjekti elusus määrab verbi morfoloogias duaali vormi.

1.1. Ohustatud keelte korpusi iseloomustavad omadused

Kuigi ohustatud keele andmetel tehtava korpuslingvistika eesmärgid, uurimisküsimused ja üldine kontseptsioon ei erine teoreetiliselt sellest, kuidas uuritakse korpuste abil suuri ja stabiilseid keeli, on ometi mitmeid aspekte, mis on iseloomulikud

[3] Leidub ka teisi põhjusi, vt ülevaadet (Austin & Sallabank 2011: 5–6).

just ohustatud keelte korpustele ja mis muudavad korpuslingvistilise uurimistöö eriti keeruliseks. Tuleb muidugi meeles pidada, et iga keel ja iga korpus on siiski lõppkokkuvõttes ainulaadne ega pruugi kõikidele nendele omadustele vastata.

Ohustatud keelt räägib tavaliselt väike inimrühm (kõnelejate arv tuhandetes, sadades või mõnikord isegi vähem), seega pole juhus, et ohustatud keelte korpused on tavaliselt üsna **väikesed** – kui mitte muul põhjusel, siis seetõttu, et tekstide[4] tootmiseks saadaolevate kõnelejate hulk on lihtsalt väiksem kui suurte, stabiilsete, enamuse keelte puhul. Lisaks saavad ohustatud keeled sageli vähem poliitilist toetust, mis võib kaudselt mõjutada loodud tekstide hulka – see tähendab, et korpuse jaoks on keeleallikaid raskem leida. Ohustatud keeli räägitakse sageli kogukondades, kus selle keele kasutus on kas peamiselt või isegi ainult suuline, nii et keele salvestamine korpusesse lisamiseks (sõltumata sellest, kas see on kirjalik või suuline) on suhteliselt uus ettevõtmine. Praeguste tehnoloogiliste võimaluste juures peavad tekstid siiski olema korpusesse lisamiseks mingil moel kirjalikul kujul esitatavad (teisisõnu, transkribeeritavad). Suulise registri tekstide korpusesse lisamine, eriti kui ortograafilist standardit polegi olemas, nõuab märkimisväärset täiendavat tööd. Nende ja teiste põhjuste tõttu on ohustatud keelte korpused tavaliselt üsna väikesed ja palju väiksemad kui stabiilsete enamuskeelte korpused.

Ressursside, tekstide ja sageli isegi tööjõu (st korpusi loovate lingvistide ja/või kogukonna liikmete) puudumise tõttu koosnevad sellised korpused tihti kõigist olemasolevatest tekstidest, et maksimeerida korpuse suurust. See tähendab, et puhtalt vajadusest lähtuvalt juhib korpuse loomist mõte, et on parem koguda **kõike ja kõikvõimalikku**. Kuigi korpus, mis on täiesti esinduslik, on tõenäoliselt saavutamatu isegi suurte ja stabiilsete enamuskeelte puhul[5], muudab selline oludest lähtuv koostamine esinduslikkuse ja tasakaalustatuse saavutamise ohustatud keelte korpuste jaoks veelgi võimatumaks (vt nende mõistete kohta ka õpiku ptk 1.3.1).

Nagu eespool mainitud, on ohustatud keeled sageli kasutusel kas peamiselt või ainult **suulises vormis**, erinevalt suurtest ja stabiilsetest enamuskeeltest, millel on pikaajaline kirjaliku keele traditsioon. See tähendab, et sageli ei pruugi kirjalikke tekste olla palju ja ortograafiline standard võib olla kas alles hiljuti tekkinud, veel arendamisel ja/või üsna muutuv. Kuid vähemalt praeguste tehnoloogiliste võimaluste juures sõltub korpuslingvistika ikka veel suuresti, kui mitte ainult, keele töötlemisest kirjalikul kujul lineaarsete tähemärkidena tekstifailis (isegi kui korpus esindab suulist keelt, nagu see on heli-/videosalvestuse transkriptsiooni korral).

See tähendab, et ohustatud keele korpuse loomine võib olla eriti aeganõudev ja tüütu, sest esmalt võib olla vajalik töötada välja kirjaliku transkriptsiooni standard selleks, et keelesalvestusi järjekindlalt transkribeerida. Sellised suulise keele korpused on autentne keeleressurss, kuna valdav osa inimkeelest on just nimelt

---

[4] Siin on mõeldud „teksti" kõige laiemas võimalikus tähenduses, viidates mis tahes keelelise sündmuse salvestusele, mitte ainult kirjalikele tekstidele.

[5] Vt näiteks selgitust esinduslikkuse saavutamatuse kohta (Stefanowitsch 2020: 28–36).

suuline. Kuid suulise keele esitamisel kirjalikul kujul on oma raskused isegi nendes keeltes, millel on pikk kirjalik traditsioon. Need probleemid on ohustatud keelte puhul veelgi teravamad siin mainitud põhjustel.

Ohustatud keelte suuruse ja olukorra tõttu **puuduvad sageli ka digitaalsed keeletööriistad**, mis võiksid korpuse loomist kiirendada. Loomuliku keele töötlemise (ingl *natural language processing*, NLP) tööriistad, mida kasutatakse korpusandmete (pool)automaatseks märgendamiseks või heliandmetest transkriptsioonide loomiseks, on üsna levinud ja need muutuvad maailma suurimate keelte jaoks üha täpsemaks, kuid ohustatud keelte jaoks on need tavaliselt minimaalsed või täiesti olematud. Selliste tööriistade puudumisel on korpuse loomine aeglasem ja tülikam ettevõtmine.

Keeleandmete kasutamise eetiliste ja õiguslike küsimuste poolest ei ole ohustatud ja muude keelte vahel teoreetiliselt mingeid erinevusi. Siiski on mõned praktilised aspektid, mis võivad muuta ohustatud keele korpuslingvistika nendes küsimustes natuke keerulisemaks. Esiteks räägitakse ohustatud keeli sageli koloniaalses või postkoloniaalses kontekstis (kolonialismil on oluline roll ohustatuse tekkimisel), samas kui nende keelte korpuslingvistiline uurimine leiab sageli aset enamuse kultuuris. Potentsiaalsed konfliktid ja usaldamatus, mis tekivad sellisest kontekstist, nõuavad erilist tundlikkust ja tähelepanu, eriti seoses **lubade ja piirangutega** tekstide kasutamisel andmetena, sealhulgas ka uurimistulemuste avaldamisel[6]. Kui ohustatud keele kogukond on väike, on keeleandmete anonümiseerimine keeruline ka seetõttu, et kõnelejad on äratuntavad hääle järgi (audiosalvestiste puhul), pildi järgi (videosalvestiste puhul) ja isegi info järgi, mis selgub tekstist endast. Seda silmas pidades on eriti oluline kindel olla, et teil uurijana on õigus pääseda ligi, salvestada ja avaldada oma tulemustes andmeid, mis võivad olla tundlikud või isiklikud[7].

Vaatamata nendele väljakutsetele on ohustatud keelte korpuslingvistika uurimine siiski täiesti teostatav ettevõtmine. Ohustatud keeled võivad pakkuda nii tõendeid haruldaste või ebatavaliste keelestruktuuride kohta kui ka kinnitada teatud universaalseid tendentse keeles, igal juhul täiendab ohustatud keelte mõistmine veelgi meie teadmisi inimkeelest. Lisaks annavad ohustatud keele andmed teadmisi ainulaadse kultuuri kohta, mis võib olla samamoodi ohustatud nagu keel. Nii võib ohustatud keele korpus pakkuda huvi ka teistele humanitaarteaduste distsipliinidele ning olla kasulik ressurss keelekogukondadele endile. Ohustatud keelte

[6] Carroll jt (2021) on esitanud C.A.R.E. raamistiku põliskultuuridest pärinevate andmetega töötamiseks, mis võib sobida ka ohustatud keele korpusandmete jaoks.

[7] Euroopa Liidus reguleerib isikuandmete töötlemist GDPR (vt https://gdpr-info.eu/); see kehtib alati, kui mõni andmete pakkumise, kogumise või salvestamisega seotud osapool on ELis. Isikuandmetega seotud korpuste koostamise kohta vt lisaks õpiku ptk 4.5 „Eetilised ja õiguslikud aspektid".

korpused on sageli rikkalikud allikad mitmekeelsuse, keele struktuurilise kulumise (ingl *language attrition*) ja keelevahetuse uuringuteks.

1.2. ELAN kui korpustööriist ohustatud keelte jaoks

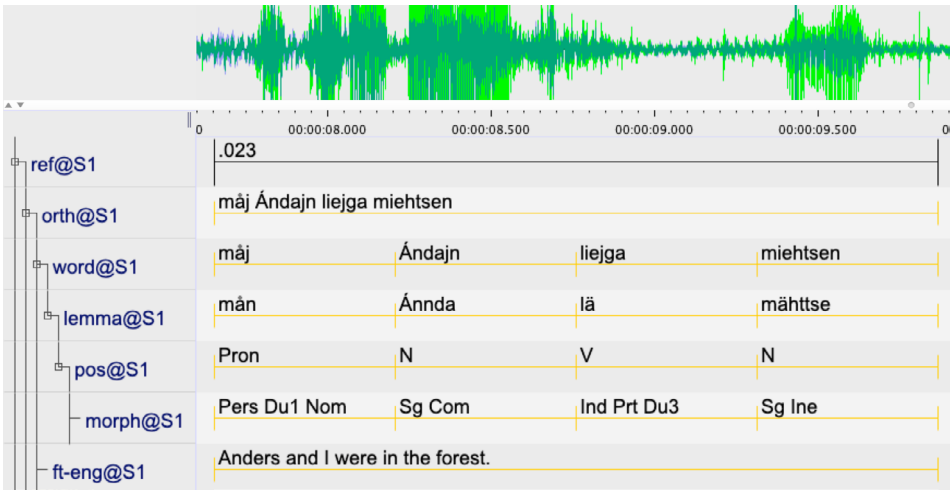
Viimase kolme aastakümne jooksul on **keelte dokumenteerimine** (ingl *documentary linguistics*) kujunenud keeleteaduse valdkonnaks, mis spetsialiseerub ohustatud keelte dokumenteerimise teooriale ja metoodikale ning kasutab kaasaegseid digitaaltehnoloogiaid, vt nt (Himmelmann 2006; Woodbury 2011). Üheks selle valdkonna levinud kõrvalsaaduseks on ohustatud keele kõikide transkribeeritud salvestiste arhiveerimine, mida saab kasutada korpusena. Üks spetsiaalselt ohustatud keelte heli- ja videosalvestuste transkribeerimiseks ja märgendamiseks välja töötatud tööriist on ELAN (ELAN (Version 6.4) 2022; Sloetjes & Wittenburg 2008). Tänu võimalusterohkele otsinguliidesele ja võimele korraga otsida mitmest märgendusfailist, saab seda kasutada ka korpustööriistana. Kuigi ELAN ei ole mõeldud eelkõige korpustööriistaks ja selle otsinguliidesed on (hoolimata nende näilisest keerukusest) oluliselt piiratud, saab ELAN-it kasutada vähemalt mõningal määral andmete ettevalmistamiseks ja edasise uurimise jaoks väljasõelumiseks.

ELAN on multimeedia keeleannotatsiooni programm, mis võimaldab ajaliselt sünkroniseeritud teksti märgendamist heli- ja/või videomeediafailidele ühe rakenduse sees. See on tasuta tarkvara, mille on välja töötanud Max Plancki psühholingvistika instituudi tehniline osakond, ja seda saab ELAN-i kodulehelt alla laadida[8]. ELAN loodigi lingvistidele, kes töötavad mittetekstipõhiste keeleandmetega, ja selle arendamist jätkatakse nende vajadusi silmas pidades. ELAN-i märgendusfailid on lihttekstifailid XML-vormingus faililaiendiga .EAF; XML-failid seetõttu, et need sobivad sisendiks paljudele rakendustele ning need ühilduvad täielikult Unicode'i standardiga.

ELAN-is saab märgendust ja transkriptsiooni korraldada hierarhiliselt, mis võimaldab ELAN-i faile kohandada vastavalt konkreetse uurija või uurimisprojekti vajadustele. Kuna aga märgenduste hierarhia jaoks ei ole kehtestatud kindlat standardit, võib mitme erineva ELAN-i korpuse kättesaadavaks tegemine, võrreldavaks muutmine ja otsitavaks tegemine olla tõsine tehniline peavalu.

Sellise hierarhia näidet võib näha joonisel 1, mis on ekraanikuva Pite saami dokumenteerimisprojekti ELAN-i failist. Kogu Pite saami korpus on korraldatud sellise struktuuri järgi, mida võib näha joonise vasakul küljel, koos näitelause transkriptsiooni ja muude märgendustega (siin: *måj Ándajn liejga miehtsen* 'meie Ándaga olime metsas'), mis asuvad helilaine kujutise all.

[8] https://archive.mpi.nl/tla/elan



Joonis 1. Ekraanikuva, mis näitab hierarhilise kihtide struktuuri rakendamist ELAN-is märgendatud Pite saami keele lausungile

Erinevaid märgenduse tasandeid nimetatakse **kihtideks** (ingl *tiers*)[9]. Lausungi kõnelejat tähistab tekst, mis järgneb @-märgile iga kihi nimetuses; näites on kõneleja pseudonümiseeritud kui *S1*. Kiht *ref* on kõrgeim ja seda kasutatakse igale lausungile unikaalse viitenumbri andmiseks[10]. Esimene alluv kiht on *orth*, mis sisaldab lausungi transkriptsiooni standardse Pite saami ortograafia järgi[11]. Ortograafiline transkriptsioon on üksustatud järgmisel madalamal kihil (*word*), ja iga sõnavorm on edasi määratud lemmaks alluval *lemma*-kihil. Sõnaliik on märgitud alluval *pos*-kihil ja kõik asjakohased morfoloogilised kategooriad on näidatud veelgi alumisel *morph*-kihil[12]. Lõpuks antakse lausungi vaba tõlge kihil *ft-eng*, mis allub *orth*-kihile (samal tasemel kui *word*-kiht). See näide esitab ainult väikse hulga kihtidest, mida kõik Pite saami ELAN-i failid sisaldavad; võib olla veel mitmesuguseid kihte, mis on samuti paigutatud sobivalt hierarhiasse.

[9] Pane tähele, et Pite saami korpuses on peamiselt leksikaalsed ja morfosüntaktilised märgenduskihid, kuid see on tingitud ainult uurimistöö tüübist, mille jaoks korpus oli mõeldud; ELAN ei sea piiranguid andmetüübile, mida märgendustasandid võivad esindada.

[10] Nende viitenumbritega saab üheselt määratleda konkreetse asukoha salvestuses, kui on tarvis lausungile viidata kuskil mujal.

[11] Pite saami ortograafia sai esmakordselt ametlikult tunnustatuks (millega sai sellest standard) augustis 2019, vt (Steggo jt 2019).

[12] Märgendid, mis sisalduvad *word*-, *lemma*-, *pos*- ja *morph*-kihtidel, on Pite saami korpuse jaoks saadud kõik automaatselt, kasutades väljaspool ELAN-it skriptimispõhiseid töövooge; vt (Gerstenberger jt 2016; Gerstenberger jt 2017).

Lisaks ELAN-i ilmselgele kasulikkusele märgendustööriistana muudavad mitmekesised **otsingufunktsioonid** selle potentsiaalselt kasulikuks ka korpustööriistana, eriti kuna otsingud on võimalikud mitmes ELAN-i failis (st kogu korpuses) korraga. Lisaks toetavad ELAN-i otsingupäringud regulaaravaldisi, neid saab piirata nii, et otsitakse ainult teatud osadest struktureeritud kihtide hierarhias, ja otsida saab isegi mustreid märgenduste ajalisi ja (teatud määral ka) struktuurilisi suhteid arvestades. Näiteks joonisel 1 esitatud kihtide struktuuri vaadates näeme, et on võimalik otsida kõiki esimese isiku asesõnu duaalis, otsides märgendit *Du1* kihil *morph*, mis samal ajal on märgendiga *Pron* kihil *pos*. Selleks tuleb kasutada ELAN-is akent *Multiple Layer Search* (mitmetasandiline otsing), mida saab avada menüüst *Structured Search Multiple eaf…*. Hoolimata keerukast välimusest on otsinguliidesel siiski ka mõningaid olulisi puudusi, mis puudutavad seda, kuidas see pääseb ligi hierarhilisele märgendusele, nagu on kirjeldatud alajaotises 2.2[13].

2. Juhtumiuuring: elusus ja duaal Pite saami keeles

Pite saami keel[14] on kriitiliselt ohustatud uurali keel, mida tänapäeval räägib vaid väike arv inimesi, kes pärinevad Arjeplogi omavalitsuse piirkonnast Rootsi Lapimaal (kaasa arvatud alad Norras). Keel ja selle kõnelejad on olnud tihedas kontaktis põhjagermaani kultuuride ja keeltega rohkem kui sajandi jooksul, nii et mitme põlvkonna vältel on oldud Pite saami keele ja kohaliku Rootsi murde (tuntud kui *arjeplogsmål*) enam-vähem kakskeelsed kõnelejad. Domineerivate põhjagermaani keelte ja ühiskondade nii avaliku kui ka kaudse surve tõttu läksid Pite saami kõnelejad Norras mitu aastakümmet tagasi üle norra keelele, samal ajal valisid paljud vanemad Rootsi poolel 20. sajandi jooksul oma lastega rääkimiseks ainult rootsi keele. See põhjustas peaaegu täieliku katkestuse põlvkondadevahelises keele edasiandmises, nii et tänapäeval räägib Pite saami keelt emakeelena ainult umbes 35 inimest ja enamik neist on vanemad kui 60. Neid fakte arvestades on ilmne, et Pite saami keel on ohustatud. UNESCO keele elujõulisuse ja ohustatuse aruandes toodud kriteeriumide järgi on Pite saami keel „kriitiliselt ohustatud" (UNESCO Ad Hoc Expert Group on Endangered Languages 2003), mis on vaid ühe sammu kaugusel „väljasuremisest".

Käesoleva uurimuse jaoks kasutatud suulise keele andmed pärinevad salvestustest, mis on kogutud erinevates keele dokumenteerimise projektides autori poolt aastatel 2008 kuni 2019 ja mõnest vanemast spontaanse kõne transkriptsioonist aastatest 1921 ja 1929. Uurimuse läbiviimise ajal (detsember 2022) sisaldas Pite saami spontaanse kõne korpus kokku 33 740 sõnet; uurimuse jaoks kasutatud

[13] Põhjalikuma arutluse ELAN-i otsingufunktsioonidest ja puudustest leiab (Wilbur 2019).

[14] ISO 639-3 kood: sje, Glottokood: pite1240.

andmed pärinevad korpuse alamosast, mis sisaldab selle läbiviimiseks vajalikku märgendust.

Salvestised on transkribeeritud ametliku Pite saami ortograafia järgi; edasine märgendus sisaldab üksustatud ja lemmatiseeritud esitust ortograafilisest esitusest, sõnaliigi ja morfoloogia andmeid ning lausungite vabu tõlkeid inglise keelde (või mõnikord rootsi keelde) vastavalt märgendushierarhiale, mida on kirjeldatud üleval jaotises 1.2 ja illustreeritud joonisel 1; seal võib olla lisaks ka teisi kihte muud tüüpi andmetega.

2.1. Duaal Pite saami keeles

Pite saami keele morfoloogiline arvukategooria on huvitav, sest lisaks ainsusele ja mitmusele on selles ka **duaal**, kui viidatavaid entiteete on täpselt kaks. Võimalike nimisõnafraaside (NP) hulgas saavad duaalis olla ainult isikulised ja enesekohased asesõnad, ent duaal avaldub ka finiitsetel verbivormidel, mis ühilduvad subjektiga arvus, sealhulgas duaalis. Näites (1) ühildub duaalis isikuline asesõna duaalis verbiga; siin on duaalis isikuline asesõna *såj*, mis viitab abielupaarile, samal ajal kui verb *inijga* 'omama/hoidma' ühildub arvus subjektiga.

(1)	*såj*	*inijga*	*tjuohte*	*buhtsujt*	*ikktij*[15]
	såj	ini-jga	tjuohte	buhtsu-jt	ikktij
	3DU.NOM	omama-3DU.PST	tuhat	põhjapõder-ACC.PL	koos

'neil (kahel) oli koos tuhat põhjapõtra'

Sellest näitest võib näha, et duaalsus ilmneb vaid isikuliste ja enesekohaste asesõnade puhul, ning seega võib tekkida küsimus, kas duaali kasutamine verbidel võib olla piiratud ainult teatud tüüpi subjektidega, millega nad ühilduvad. Lagercrantz (1926: 79) vihjab, et duaali kasutatakse Pite saami keeles ainult inimeste puhul, kuid selle väite toetuseks ei ole tehtud korpusuuringut, lisaks võib Pite saami keel olla 100 aasta jooksul pärast Lagercrantzi välitöid muutunud. Seda, et duaali kasutatakse ainult inimeste kohta, on väidetud ka teiste saami keelte kohta (kuigi pole selge, kui palju sellest uurimistööst on korpusel põhinev), seega ei tohiks olla üllatav, kui see kehtib ka Pite saami keele puhul. Näiteks lähedalt suguluses olevas Lule saami keeles iseloomustab duaal „kahte inimest või elusolendit" (Spiik 1989: 31), Inari saami keeles „elusolendid tingivad täieliku ühildumise ja elutud osalise ühildumise" (Toivonen 2007: 231). Kejonen (2017) annab ülevaate duaalist kõigis saami keeltes, tuginedes sekundaarkirjandusele. Seepärast vaatame lähemalt korpusandmeid.

[15] Näide pärineb lausungist sje19210000a-lagercrantz1957a-426.027 Pite saami korpuses.

2.2. Operatsionaliseerimine ja analüüs

Uuringu läbiviimiseks on vaja leida sellised verbid, mida siin nimetame „duaali verbideks", ja subjektid, millega nad korpuses ühilduvad. See nõuab operatsioonilise definitsioonide loomist (vt operatsionaliseerimise kohta ptk 1.1.2 „Uurimisküsimuse esitamine, hüpoteesi püstitamine" ja Stefanowitsch 2020: ptk 3.2), et selgelt ja üheselt määratleda, mida me otsime (st milliseid konstruktsioone) selles konkreetses korpuses, kasutades meie käsutuses olevat spetsiifilist korpusliidest. ELAN-i otsinguliidese efektiivseks kasutamiseks on vajalik mõista konkreetse ELAN-i korpuse kihtide struktuuri, kuna igal ELAN-i korpusel või isegi failil on tavaliselt oma struktuur.

Spontaanse suulise keele olemuse tõttu ei vasta lausungi üksused alati täielikult süntaktilistele (osa)lausetele (valestart, eneseparandused, katkestused jne on vaid mõned põhjused, miks see nii võib olla). Kihtide hierarhia kõrgeimal tasandil (juurkihil) on suuline keel Pite saami ELAN-i korpuses jagatud nii, et need vastavad ligikaudu fraasitasandi intonatsiooniüksustele, millest igaühele on antud viitenumber *ref*-kihil ja mis on transkribeeritud *orth*-kihil (nagu eespool kirjeldatud). Need üksused vastavad ligikaudu klausi tasandi süntaktilistele üksustele (nagu finiitse verbiga osalause, selle argumendid ja laiendid, sealhulgas võimalikud alistatud laused ja mittefiniitsed üksused), kuid see suhe on sageli üsna nõrk. See on spontaanses suulises keeles tavaline. Pite saami korpuses käsitletakse neid lausungiüksusi praktilistel eesmärkidel klausidena, mis aitavad piiritleda otsingu ulatust üksikute duaaliverbide kasutusjuhtude leidmiseks. Lisaks aktsepteerime korpuses esitatud märgendeid korrektsete Pite saami keelestruktuuride esitustena, pannes tähele, et tegemist on automaatmärgendusega, kasutades lõplike muundurite (ingl *finite state transducer*) ja kitsenduste grammatika (ingl *constraint grammar*) infrastruktuuri ja skripte (Gerstenberger jt 2016; Gerstenberger jt 2017). Neid kahte asjaolu arvestades saame esitada järgmised operatsioonilised definitsioonid:

Duaaliverb: finiitne verbivorm, mis:

- on märgistatud kui *V* (verb) kihil *pos@<speaker>*;
- on märgistatud *Du1, Du2* või *Du3* (duaal ja 1., 2. või 3. isik) osana morfoloogilistest glossidest kihil *morph@<speaker>*.

Subjekt: sellise osalause grammatiline subjekt, mis sisaldab duaaliverbi ja mida saab kindlaks teha järgmiselt:

- see on nimisõnafraas nominatiivis nimisõnaga või nominatiivis asesõna;
- asesõna väljajätu[16] korral tuletatakse subjekt kontekstuaalsete vihjete põhjal, näiteks vahetult eelneva(te) lausungi(te) subjekti põhjal; dialoogides

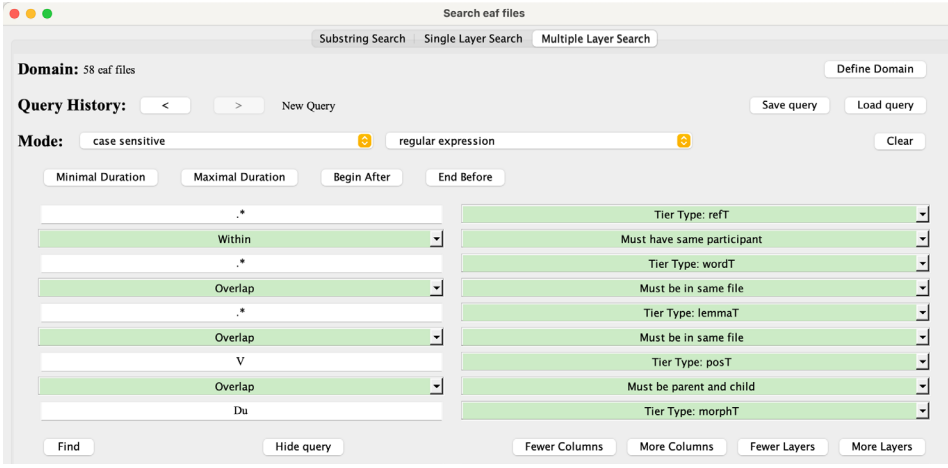
[16] See tähendab, et subjekti ei märgita otseselt isikulise asesõnaga, vaid isik selgub pöördelõpust.

võib seda järeldada näiteks esimese ja teise isiku vahetusest, kui kõneleja vahetub.

Subjekti tüüp: subjekti reaalse maailma referendi elususe väärtus:

- *hum*: inimeste jaoks;
- *animHigh*: suuremate loomade jaoks (nagu koerad või põhjapõdrad);
- *animLow*: teiste, väiksemate loomade jaoks (nagu putukad);
- *inanim*: elutute asjade jaoks (nagu kivid või autod).

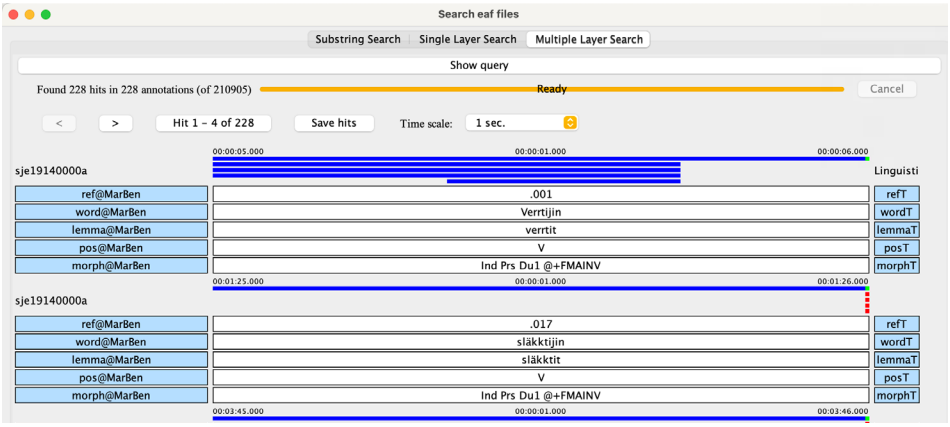
Nende operatsiooniliste definitsioonide põhjal saab ELAN-i rakenduses[17] seadistada *Multiple Layer Search* akna duaaliverbide näidete leidmiseks, otsides *V* kasutusjuhte sõnaliigi kihil (Pite saami korpuses on selleks kihitüübiks *posT*), ja *Du* kasutusjuhte morfoloogia kihil (Pite saami korpuses *morphT*). Joonis 2 esitab ekraanikuva sellest, milline näeb otsing välja ELAN-i versioonis 6.4 MacOS-is. Pange tähele, et otsingukriteeriumid on seatud tõstutundlikuks ja kasutavad regulaaravaldisi. Kuna see korpus on kavandatud nii, et kihitüüp (*Tier Type*) *posT* kehtib ainult kõigile *pos@<kõneleja>*-kihtidele ja kihitüüp *morphT* kehtib ainult kõigile *morph@<kõneleja>*-kihtidele, saab neid kasutada sõnaliigi (piirates otsingut *V*-ga) ja morfoloogia (otsides *Du*) otsinguvihjetena.



Joonis 2. Ekraanikuva, mis näitab üht võimalikku viisi, kuidas seadistada *Multiple Layer Search* otsing ELAN-is, et leida duaaliverbide näiteid Pite saami korpuses

[17] Seda saab avada menüüst *Structured Search Multiple eaf...*. Pane tähele, et esmalt tuleb valida korpus (st otsingu jaoks laaditud ELAN-i failide kogum), kasutades nuppu *Define Domain*.

Selle päringu käivitamine Pite saami spontaanse kõne korpuses annab 235 tulemust; kaks juhuslikku tulemust on näidatud joonisel 3, nagu neid kuvatakse ELAN-i konkordantsi vaates[18]. Sellest vaatest saame tulemused salvestada CSV-failina (kasutades eraldajana tabelduslahvi ehk tabulaatorit <*tab*>), klõpsates nupul *Save hits*. Seda faili saab seejärel kasutada andmestiku esialgse põhjana, kui see importida näiteks tabelarvutusprogrammi nagu Google Sheets või Microsoft Excel või kasutades mõnd programmeerimiskeelt (nt Python või R).



Joonis 3. Ekraanikuva mis näitab kaht juhuslikku tulemust ELAN-i *Multiple Layer Search* aknas, otsides duaalverbide näiteid Pite saami korpusest; kuvatud ELAN-i konkordantsivaates (*concordance view*)

Nüüd, kui oleme leidnud kõik potentsiaalsed duaaliverbide juhud ja loonud andmestiku jaoks tabeli, on järgmine samm **andmete puhastamine**. Kuigi esialgne ELAN-i tulemuste komplekt sisaldas 235 potentsiaalset duaaliverbi juhtu, tuleb esmalt eemaldada juhud, mis on sattunud andmete hulka mitmeti tõlgendatavuse tõttu (peamiselt esimese isiku oleviku duaali ja kolmanda isiku mitmuse mineviku vahel, kui tegelikult on tegemist viimasega), mitte-emakeelsete kõnelejate juhud ning valed transkriptsioonid või muud valepositiivsed tulemused; tulemuseks jääb meile 111 tõelist duaaliverbi juhtu. Seejärel saame määrata ja kategoriseerida iga kasutusjuhu grammatilise subjekti. Nagu näha, on ELAN-i otsingu funktsionaalsus

[18] Pane tähele, et kui lülitada (parema hiireklõpsuga) see vaatele *Show Frequency view (by frequency)*, kuvatakse tabamused sageduse järgi, kuigi see ei ole selle andmekogu jaoks mõttekas; kuid see võib olla kasulik muud tüüpi uurimistööde jaoks, näiteks kollokatsioonide uurimisel.

sellise otsingu jaoks, mille sihtmärgiks on morfosüntaktilised omadused, üsna piiratud[19].

Kuigi otsingut saaks laiendada nii, et see hõlmaks algset Pite saami transkriptsiooni ja isegi iga lause vabu tõlkeid (lisades otsinguliideses kihte nii, et otsing haaraks alati kõike, kus esineb kihitüüp *orthT* lausungis ja *ft-langT* tõlkes), ei ole võimalik kaasata *ainult* teisi märgendeid kihtidel *pos@<kõneleja>* või *morph@<kõneleja>*, mis kuuluvad *samasse* lausungisse. Teisisõnu ei ole võimalik otsingutulemusi nii piirata, et leiaks ainult ühe lausungi alla kuuluvaid tabamusi. Neil, kes pole Pite saami keeles piisavalt vilunud ega suuda seetõttu transkriptsioonist välja lugeda asjakohast informatsiooni grammatilise subjekti kohta, tuleb kõik tabamused ükshaaval läbi vaadata, et välja sõeluda edaspidi vajalikke vaatlusi (subjekti väljajätu korral on see niikuinii vajalik, kuna on vaja leida kontekstuaalset infot teistest diskursuse lausungitest). Seetõttu tuleb edasine andmete ekstraktimine andmekogumi täiendamiseks teha käsitsi.

Selleks on võimalik klõpsata tabamusel ELAN-i otsingutulemuste aknas, mis avab asjakohase faili konkreetse tabamuse asukohas. See on tüütu töö ja tuleb teha ükshaaval iga tabamuse vaatamiseks, toimides näiteks järgmiselt:

1. Ava päringuvaste vastavas ELAN-i failis, klõpsates vastaval tabamusel nimekirjas (vt joonis 3).
2. Määra, kas see on tõepoolest duaaliverbi näide või tuleks see andmestikust eemaldada, kuna mingil põhjusel on tegemist valepositiivse tulemusega.
3. Määra, mis on subjekt, vaadates lausungit ennast või subjekti väljajätu korral ümbritsevat diskursust.
4. Klassifitseeri subjekti tüüp vastavalt eespool toodud operatsioonilisele definitsioonile.
5. Vajadusel lisa märkmeid.

Iga tabamuse jaoks määratud subjekti tüübi saab seejärel koos muude märkmetega lisada CSV-faili (iga uus andmetüüp uude veergu), nagu esitatakse valitud arvutustabeli rakenduses. Lisaks on pseudonümiseeritud salvestiste failinimed (näiteks *t01*, *t02* jne) ja kõnelejate nimed (näiteks *s01*, *s02* jne)[20]; lisatud on metaandmed iga kõneleja sünniaasta ja soo kohta, juhuks kui need faktorid võiksid andmetega seotud olla. Joonisel 4 esitatud ekraanikuva illustreerib, milline see välja võiks näha.

[19] Nagu varasemas allmärkuses mainitud, on varem antud probleemist detailsem ülevaade, sest ELAN ei ole ideaalne korpusotsingu tööriist (Wilbur 2019: 102). ELAN-i otsinguliides on kasulik korpusuuringuteks, kus on tarvis leida konkordantse ja teha sagedusloendeid või kus tuntakse huvi märgendatud üksuste kestuste vastu.

[20] Väikeste keelekogukondadega tegelemisel, kus inimesed tunnevad sageli kõiki teisi kõnelejaid isegi siis, kui nende nimed on pseudonümiseeritud, rakendatakse seda anonüümsuse lisatasandit isikuandmete paremaks kaitsmiseks.

	A	B	C	D	E	F	G	H	I
1	**file**	**ref-no**	**speaker**	**gender**	**dob**	**lemma**	**morph**	**subj-type**	**notes**
2	t01	.007	s07	f	1877	állget	Ind Prt Du3	hum	Stállo included
3	t02	.014	s07	f	1877	vuällget	Ind Prs Du3	hum	
4	t02	.015	s07	f	1877	vuällget	Ind Prs Du3	hum	
5	t02	.016	s07	f	1877	ådtjot	Ind Prs Du2	hum	
6	t02	.017	s07	f	1877	vuällget	Ind Prt Du3	hum	
7	t02	.019	s07	f	1877	lä	Ind Prt Du3	hum	
8	t02	.019	s07	f	1877	lä	Ind Prt Du3	hum	

Joonis 4. Ekraanikuva, mis näitab mõnda rida pseudonümiseeritud CSV-andmekogumist (kasutades Google Sheetsi rakendust)

Kokku 22 suulise keele salvestusest Pite saami kõnekeele korpuses sai võtta puhastatud andmestikku 111 duaalivormis verbi ja neile vastavate subjektide elususe näidet. Kokku oli 21 teksti; neist kuusteist salvestati aastatel 2008–2010, 2013–2015, 2017 ja 2019 ning viis neist on olemas ainult transkriptsioonidena, mis on kogutud aastatel 1921 ja 1929. Kõnelejatest neli olid naised ja viis mehed, 78 andmepunkti pärinesid naistelt ja 33 meestelt. Metaandmete kokkuvõte on esitatud tabelis 1.

Tabel 1. Metaandmed (pseudonümiseeritud) kõnelejate kohta ja duaalis verbivormide arv

Kõneleja	**Sugu**	**Sünniaasta**	**Näiteid andmestikus**	**Teksti(de)s**
s01	m	1945	11	t10
s02	m	1929	10	t09
s03	m	1977	10	t06, t15, t18, t19, t20
s04	m	1946	1	t21
s05	m	1880	1	t05
s06	n	1954	45	t07, t10, t11, t13, t14, t16, t17
s07	n	1877	27	t01, t02, t03, t04
s08	n	1927	4	t12
s09	n	1936	2	t08

3. Tulemused ja arutelu

Seda andmekogumit vaadates on pilt väga selge ja toetab meie oletust, et Pite saami duaali märkimine on oluline vaid siis, kui grammatiline subjekt viitab elusolendile. 111 duaalimarkeriga verbi näitest ühildusid kõik peale ühe subjektidega, mis olid mitte ainult elus, vaid inimesed. Ainus potentsiaalselt piiripealne juhtum oli üks duaalse subjekti näide, mis sisaldas saami müütilist kuju Stállot; siiski on Stállo füüsiliselt väga inimesesarnane olend (kuigi natuke suurem kui inimesed ja tavaliselt väga kohmakas ja rumal) ja selgelt elus. Andmetes ei leidunud ühtegi juhtu loomadest või elututest asjadest. Selle põhjal näitavad korpuse andmed selgelt, et duaalimorfoloogia esineb ainult siis, kui duaalimärgistusega verbi subjekt viitab inimesele.

Tõendite puudumine ei tõesta põhimõtteliselt siiski, et midagi ei saa olemas olla; see kehtib eriti siis, kui andmekogu on väike, nagu praeguses näidisuuringus, ja nagu tihti on olukord ohustatud või muude väheste ressurssidega keelte puhul. Olukorda saab parandada, suurendades korpuse mahtu, kui see on võimalik. Korpuslingvistika alternatiivina võib välitööde ajal küsitledes soovitud keelenähtust esile kutsuda, et testida korpusuuringu paikapidavust, kuigi esilekutsumisel (ingl *elicitation*) ja muudel kõneleja hinnangule toetuvatel ülesannetel on nii teoreetilisi kui ka praktilisi puudusi. Kui keelel pole enam kõnelejaid, kelle abi keeleteadlane saab kasutada, võivad arhiivid olla ainus võimalus tekstide arvu suurendamiseks.

Kuigi elususe küsimusele oleme saanud selge vastuse, on uurimist väärt ka andmekogu teised küljed. Näiteks võib Pite saami keeles leida duaalimorfoloogiat finiitverbi indikatiivi oleviku- ja minevikuvormides (kaasa arvatud küsimused ja eitus[21]), imperatiivis ja potentsiaalis[22]; indikatiivi ja potentsiaali puhul muutub verb ka esimese, teise või kolmanda isiku järgi, samas kui imperatiiv muutub ainult teise isiku järgi. See tähendab, et iga verbi lemma kohta on kümme erinevat duaalivormi. Andmekogus esines vähemalt üks kõigist võimalikest duaali kõneviisi- või ajavormidest, välja arvatud potentsiaal, mille kohta polnud ühtegi näidet, nagu on näha tabelist 2.

[21] Pane tähele, et andmete märgendamine põhineb Giellatechno standarditel (vt http://giellalt.github.io/lang-sje/), mille kohaselt märgendatakse tegusõna eitusvormid omaette kõneviisi kategooriana (märgistatud kui *Neg*); selle uuringu jaoks määrati tegusõna eitavad vormid indikatiivi või imperatiivi kõneviisi kategooriatesse ja neid ei peetud omaette kõneviisiks.

[22] Potentsiaali kõneviisi kohta Pite saami keeles vt lähemalt (Wilbur 2014: 152–153).

Tabel 2. Uurimuse andmestiku absoluutsagedused võimalike verbivormide kohta, mis sisaldavad duaali (IND – indikatiiv, PRS – olevik, PRT – minevik, IMP – imperatiiv, POT – potentsiaal)

		Isik			
Kõneviis	**Aeg**	**1.**	**2.**	**3.**	**Kokku**
Ind	Prs	45	1	2	48
	Prt	19	1	42	62
Imp		-	1	-	1
Pot		0	0	0	0
Kokku		64	3	44	111

Lisaks võib andmete mõistmiseks vaadelda ka muid tegureid, näiteks tekstilingvistilisi tegureid (nt žanr, teksti pikkus). Kuna meie andmekogu sisaldab metaandmeid kõneleja soo ja vanuse kohta, on kaalumist väärt ka nende sotsiolingvistiliste tegurite kaasamine analüüsi. Näiteks pärines kokku 78 juhtu naiskõnelejatelt ja 33 meeskõnelejatelt. Siiski on oluline meeles pidada korpusvalimi väikest suurust ja andmete puudumist selle kohta, kui palju iga kõneleja väljendeid korpus sisaldab. Samamoodi ei saa kõnelejate vanuselisest jaotusest palju kindlaks teha. Lõppkokkuvõttes on võimatu teha mingeid mõistlikke järeldusi soo või vanuse olulisuse kohta, välja arvatud asjaolu, et mõlemast soost kõnelejad võivad kasutada duaaliverbe ja et duaaliverbe kasutatakse praegu samamoodi nagu sada aastat tagasi.

Kokkuvõte

Kokkuvõttes on ohustatud keelte korpuslingvistiline uurimine, nagu eespool toodud näidisuurimus näitab, selgelt oluline tegevus. Ohustatud keelte korpusi peetakse sageli eriti huvitavaks, kuna need esindavad kõneldud keelt. Siiski on neil korpustel rida piiranguid, mis puudutavad eriti korpuse suurust ja märgendatuse määra, mis ei mõjuta mitte ainult seda, millist uurimistööd on võimalik läbi viia, vaid võivad olla segavad asjaolud ka tulemuste analüüsimisel ja selgitamisel. See on peamiselt tingitud sellest, et sellised korpused on vähem esinduslikud ja/või tasakaalustatud kui suuremate, hästi väljakujunenud keelte korpused.

Lisaks on ohustatud keelte korpused tänapäeval sageli koostatud ELAN-is. Kuigi see on suurepärane valik mitme meediumi üheaegseks esitamiseks ning märgenduste hierarhiliseks struktureerimiseks, on ELAN-i päringufunktsioon niivõrd piiratud, et keerukamad otsingud pole võimalikud (nagu on selgitatud jaotises 2.2).

Peamine lahendus eelnevale probleemile on korpuse mahu suurendamine ja märgenduste parandamine, kuid enamiku ohustatud keelte puhul on see ebarealistlik ressursside ja kõnelejate puudumise tõttu. Selle asemel peavad uurijad lihtsalt mõistma piiranguid ja andma endast parima, et maksimaalselt ära kasutada seda, mis neil on, isegi kui see ei jõua kunagi lähedale suurtel andmetel põhinevatele korpusuuringutele, mis on võimalikud suuremates keeltes.

ELAN-i korpustööriistana kasutamise tehniliste probleemide korral võib kasutada alternatiivseid tööriistu, nagu Sketch Engine või KORP. Kuna ELAN-i failid on struktureeritud XML-failid, on võimalik neid automaatselt teisendada teistesse vormingutesse, kuid see nõuab mõningasi tehnilisi oskusi.

Lühendid

Acc	akusatiiv
animHigh	kõrge elusus
animLow	madal elusus
Du	kaksus e duaal
hum	inimene
Imp	käskiv kõneviis e imperatiiv
inanim	mitte-elus
Ind	kindel kõneviis e indikatiiv
Ine	seesütlev e inessiiv
Nom	nimetav e nominatiiv
Pers	isikuline asesõna e personaalpronoomen
Pl	mitmus e pluural
Pot	potentsiaali kõneviis, potentsiaal
Prs	oleviku ajavorm
Pst (Prt)	mineviku (preteeritumi) ajavorm
Sg	ainsus e singular

Kirjandus

Austin, Peter K. & Julia Sallabank (toim). 2011. *The Cambridge handbook of endangered languages* (Cambridge handbooks in language and linguistics). Cambridge: Cambridge University Press.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell & Shelley Stall. 2021. Operationalizing the CARE and FAIR principles for indigenous data futures. *Scientific Data* 8(1). 108. https://doi.org/10.1038/s41597-021-00892-0.

ELAN (Version 6.4). 2022. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler & Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology* 4. 29–47. https://doi.org/10.3384/nejlt.2000-1533.1643.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler & Joshua Wilbur. 2017. Instant annotations – Applying NLP methods to the annotation of spoken language documentation corpora. Francis M. Tyers, Michael Rießler, Tommi A. Pirinen & Trond Trosterud (toim), *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, 25–36. St. Petersburg, Russia: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0604.

Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? Jost Gippert, Ulrike Mosel & Nikolaus Himmelmann (toim), *Essentials of Language Documentation*, 1–30. Berlin / New York: Mouton de Gruyter.

Kejonen, Olle. 2017. *Dual number in the North Saami dialect of Ofoten and Sør-Troms*. Uppsala: Uppsala universitet, Department of Modern Languages, Finno-Ugric Languages. Magistritöö. http://uu.diva-portal.org/smash/get/diva2:1106012/FULLTEXT01.pdf.

Lagercrantz, Eliel. 1926. *Sprachlehre des Westlappischen nach der Mundart von Arjeplog* (Suomalais-ugrilaisen Seuran Toimituksia 55). Helsinki: Suomalais-Ugrilainen Seura.

Sloetjes, Han & Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf.

Spiik, Nils Eric. 1989. *Lulesamisk grammatik*. 2. tr. Jokkmokk: Sameskolstyrelsen.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7). Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735821.

Steggo, Peter, Inger Fjällås, Ole Henrik Magga & Bruce Morén-Duolljá. 2019. *Pitesamisk ortografi*. Arjeplog: Sámi Giellagáldu.

Toivonen, Ida. 2007. Verbal agreement in Inari Saami. Ida Toivonen & Diane Nelson (toim), *Saami Linguistics*, 227–258. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.288.09toi.

UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. *Language vitality and endangerment*. Paris: UNESCO.

Wilbur, Joshua. 2014. *A grammar of Pite Saami* (Studies in Diversity Linguistics 5). Berlin: Language Science Press. https://doi.org/10.17169/FUDOCS_document_000000020749.

Wilbur, Joshua. 2019. ELAN as a search engine for hierarchically structured, tagged corpora. *International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2019)*, 90–103. Tartu: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-0308.

Woodbury, Anthony C. 2011. Language documentation. Peter K. Austin & Julia Sallabank (toim), *The Cambridge Handbook of Endangered Languages*, 159–186. Cambridge: Cambridge University Press.

Korpuslingvistika rakendamisest digihumanitaarias: elektri saabumine Eesti aladele 20. sajandi algul

Peeter Tinits

Lühikokkuvõte

Korpuslingvistilised meetodid on leidnud rakendust ka laiemalt humanitaarias, nt ajaloos, kirjandusteaduses või kultuuriteaduses. Arvutuslikud meetodid aitavad uurida keele kõrval ka tekstide sisu. Sellist lähenemist on arendatud digihumanitaaria nime all. Näiteks on võimalik märksõnade korpuses esinemise kaudu uurida tehnoloogia levikut. Selles peatükis uurime elektri tulekut Eesti aladele 20. sajandi algul, toetudes Eesti Rahvusraamatukogu digiteeritud ajalehekogule. Selleks otsime kõigepealt elektriga seotud märksõnu ja sõnaühendeid, siis leiame nende seast erinevad tooted ja nähtused ning vaatame lähemalt nende esinemise aega ja ümbritsevat konteksti. Nii saab esialgse ülevaate elektrimootori, elektripliidi ja elektrilokkide leviku kohta ühiskonnas ning täpsemalt, millal ja mis kontekstis need nähtused esile kerkisid. Peatükis tutvustatakse läbivalt põhimõtteid, mida sellise uuringu koostamisel järgida, ja avatakse nende tausta.

1. Sissejuhatus

Suured tekstikorpused lubavad täiendada ja laiendada keeleteaduse uurimisseisu. Need avavad uusi võimalusi ka teistele valdkondadele humanitaar- ja sotsiaalteadustes. Traditsiooniliselt on humanitaarteadused olnud suhteliselt tekstikesksed – ajalugu, kirjandusteadus, kultuuriteadus, filosoofia tuginevad oma uuringutes paljuski tekstide lugemisele ja tõlgendamisele. Moodustades tekstidest suurema korpuse, mida saab digimeetoditega analüüsida, on võimalik vastata mõnele küsimusele ajaloos või kirjandusteaduses süsteemsemalt (nt millised autorid olid 1920ndate aastate esseistikas kõige tuntumad?) või vastata küsimustele, millele muidu raske hinnangut anda (nt milline oli suhtumine keskkonnaprobleemidesse 20. sajandi algul?). Loomulikult tuleb igas sellises uuringus siiski arvestada

kontekstiga ja lisaks automaatsele analüüsile tuleb põhjalikult tutvuda materjali endaga.

See peatükk esitab näidisuuringu, kus analüüsime tekstide sisu, toetudes korpuslingvistika meetoditele. Sellist lähenemist on arendatud **digihumanitaaria** märksõna all. Digihumanitaarial on mitmeid harusid – selles peatükis võib seda mõista kui arvutuslike meetodite kasutamist humanitaaria valdkonna küsimuste lahendamisel. Peatükis käime läbi sellise uurimuse tüüpilised sammud: 1) küsimuse püstitamine, 2) asjakohaste tekstide leidmine, 3) tulemuste määratlemine, 4) analüüs. Kasutame peatükis iga sammu juures võimalikult lihtsaid vahendeid, ent muidugi on igas etapis võimalik kasutada tehniliselt keerulisemaid vahendeid töövoo automatiseerimiseks või täpsemate tulemuste saamiseks. Siin kirjeldatud sammudega saab koondada süstemaatilise ülevaate tekstide sisust teatud aspektist.

Uuringu läbiviimiseks vajalikud sammud sõltuvad alati selle eesmärkidest ning nende läbimõtlemine on oluline osa uurimistööst. Teine ja mitte vähem oluline pool korpuspõhises uuringus on selle tehniline teostus. Selle peatüki analüüsid on tehtud R-is (versioon 4.3.1, R Core Team 2023), kasutades ka Rahvusraamatukogus loodud paketti tekstide ligipääsuks ja analüüsiks (vt ligipääsu infot DEA tekstidele Rahvusraamatukogu digilaboris[1]). Uuringu läbiviimiseks tehtud tabelid ja kasutatud kood on kättesaadav õpiku koodihoidlas[2]. Selle abil on võimalik analüüsi korrata või teha uurimuse eeskujul oma analüüs.

2. Uuringu taust

Digimeetodite kasutusvõimalused sõltuvad meie uurimisküsimustest ning mõne ülesande jaoks on nad kergemini kohandatavad kui teise. Rakendamiseks on tarvis, et meid huvitavat infot oleks tekstides kajastatud ja et me oskaksime välja pakkuda viisi, kuidas seda sealt süstemaatiliselt kätte saada. Näiteks kui meid huvitab autorite populaarsus 20. sajandi alguse esseistikas, siis võime kokku lugeda kõikide autorite mainimised neis tekstides ja vaadata, kes domineerib. Võrdlemiseks võime siin võtta mainimiste korrad, viidatud teosed, neid autoreid käsitlevate tekstiosade pikkused või autorite paiknemised koosesinemiste võrgustikus. Valik sõltub sellest, mida me tähtsaks peame.

Uurides suhtumisi keskkonnaprobleemidesse, peaksime kõigepealt üles leidma looduskeskkonda puudutavad tekstilõigud 20. sajandi algul. Tol ajal ei räägitud looduskeskkonnast *keskkonna* mõiste kaudu, mistõttu peaksime otsima erinevaid loodusega seotud sõnu ja nende esinemiskontekste. Kui sobivate **märksõnade** abil on tekstid leitud, peame välja mõtlema, kuidas võiksid hoiakud looduse suhtes olla tekstides esindatud. Võiksime näiteks vaadata loodust tähistavate sõnade ümbruses

[1] https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/

[2] https://osf.io/xqzsf/

kasutatud omadussõnu või lauseanalüüsi põhjal täpsemalt looduse kohta käivaid omadussõnu. Samamoodi võiksime püüda klassifitseerida looduse mainimiskontekste selle järgi, kas seal on juttu looduses esinevatest probleemidest, looduse pakutud ressurssidest, looduse kaitsmise soovist, ohtlikust loodusest või ilusast loodusest. Ühel või teisel viisil peame siduma leiud meie taustateadmistega, kuidas võiksid hoiakud looduse suhtes olla tekstides esindatud. Üldise reeglina tasub siin mõelda, kas me ise oskame teksti lugedes sealt vajaliku info kätte saada ning kas saaksime sama info kuidagi digivahendite abil kätte.

Üks valdkondi, kus on korpuspõhist analüüsi kasutatud, on **tehnoloogia ja uute toodete leviku uurimine**. Nende kohta on informatsioon tekstides enamasti sarnasel viisil esitatud. Uutest tehnoloogiatest on juttu ajakirjanduses: neil on uudisväärtus ja nende levitajad tahavad neid nähtavaks teha. Neil on väga tihti kindel nimi, mis aitab nendega seotud tekste üles leida. Kui tooted jõuavad laia kasutusse, tehakse neile ulatuslikke reklaamikampaaniaid ja neile võib tekkida ajalehe reklaamikülgedel ka järelturg (nt „müüa kasutatud jalgratas"). Tehnoloogiatest rääkimist ei saa otse üle kanda nende levikule, tekste tuleb ikka mõista nende kontekstis. Näiteks viimasel ajal räägitakse eesti ajakirjanduses palju tuumaenergiast, mis ei tähenda, et Eestisse oleks tekkinud tuumaelektrijaam või et see siia üldse tekib. Küll aga võib reklaamikülgedel näha näiteks mikrolaineahjusid just sel ajal, kui need olid saamas populaarseks. Kui see on olemas juba igas kodus, ei ole seda enam vaja reklaamida. Nõukogude-aegsetes lehtedes aga ei reklaamitud eriti midagi ja leviku hindamiseks oleks vaja hoopis teisi allikaid.

Tuginedes ajalehtede reklaamidele, on näiteks uuritud viljapeksumasinate levikut Inglismaal 19. sajandi algul. Tol ajal oli see üpris uus tehnoloogia, mille reklaamimiseks toodi ajalehtedes tihti esile inimesi, kes olid juba selle masina hankinud. Kasutades reklaamides toodud asukohainfot, saab nõnda ülevaate ka masinate enda levikust (Caprettini & Voth 2020). 20. sajandi Hollandi ajalehtedes avaldatud reklaamide põhjal on uuritud sigarettide päritolumaid. Sigaretireklaamides öeldi tihti, kas tegemist on Ameerika või Egiptuse päritolu sigarettidega. Lähestikku paiknevate sõnade kaudu oli võimalik jälgida, millised sigaretid olid 20. sajandi jooksul eri aegadel Hollandis populaarsed. Uuring näitas päritolumaade muutumist koos riikidevaheliste diplomaatiliste suhete muutumisega (Wevers 2022).

Selles peatükis võtame näidisena ette **elektri leviku Eesti alal**. Elekter on ligi 150 aasta jooksul mõjutanud põhjalikult meie eluviisi ja muutnud meid ümbritsevaid igapäevaseid esemeid. Meie ajakirjandusse on jõudnud nii Thomas Edisoni esimesed katsetused elektriga, esimesed elektriautod 1920ndatest kui ka viimastel aastatel pidevalt kõikuva elektrihinna ülevaated. Me võime selle suure tekstikogu põhjal proovida välja selgitada, kuidas ja millal üldse elekter Eesti aladele jõudis ja mil viisil, mis toodete vahendusel jõudis ta meie igapäevaellu. Läheneme nendele küsimustele suurte tekstikorpuste kaudu, otsides elektriga seotud teemasid ja esemeid. Tegemist on induktiivse või eksploratiivse uuringuga (vt õpiku ptk 1.1.1), kus uurimistöö toob korpusmaterjalide põhjal välja trende ühiskonnas.

3. Uuringu lähtekohad

3.1. Küsimuse püstitamine

Iga uuringu lähtekohaks on uurimisküsimus: see, mida me soovime teada saada. Teadustöö puhul on ka oluline selle info uudsus – me ei saa vastust, küsides eksperdilt või vaadates õpikust, vaid vastuse võib saada just uuringu läbiviimise kaudu. Kui teada võiks tahta kõike maailmas, siis teaduslik uurimisküsimus on enamasti kompromiss meie võimaluste ja soovide vahel: me peame mõtlema sellele, kuidas me võiksime vastuse teada saada, ning kohandama oma küsimuse nii, et see oleks korraga vastatav ja meile huvi pakkuv. Digimeetodeid kasutades peame mõtlema ka vahenditele, millega uuringut saaksime teostada.

Selles näidisuuringus võtame vaatluse alla elektriga seotud tooted ja nähtused selle leviku algusaegadest. Püstitame kolm küsimust, mida on võimalik korpusanalüüsi meetoditega lahendada:

1. Millised elektriga seotud esemed ja nähtused said Eesti aladel igapäevaelu osaks?
2. Kuidas toimus nende levik ajas? Millal hakati neist rääkima ja kas oli erineva intensiivsusega perioode?
3. Mis kontekstides neist räägiti? Mida saame järeldada esemete ja nähtuste kohta neid puudutavate tekstilõikude põhjal?

Me saame nendele küsimustele läheneda **märksõnaotsingu** kaudu. Nimelt on olnud suhteliselt tavaline täiendada elektriga seotud toote nime eesliitega *elektri-* või omadussõnaga *elektriline*. Seda on eriti oluline olnud markeerida just tehnoloogia leviku esimestel sammudel – näiteks uhiuut tehnoloogiat võib rõhutatult märgistada kui *elektriline habemeajaja*. Kui juba enamik kohvimasinaid on elektrilised või tekib uus kitsama tähendusega sõna nagu *pardel*, ei ole enam tarvis mainida, et see on elektriline. Seega kui nähtus on saanud tavaliseks, võib ta osutuda meie meetodile kättesaamatuks. Käesolevas uuringus saame seega ette võtta hulga sõnu ja sõnapaare, mis sisaldavad endas ühelt poolt viidet elektrile ja teiselt poolt viidet mõnele nähtusele või tootele (nt *elektrijõud*, *elektrivalgus*, *elektrijaam*, *elektrililine habemeajaja*), ja vaadata, kuidas nad on tekstides esitatud.

3.2. Materjalid

Lisaks küsimusele ja lähenemisele on meil vaja andmeid, millele toetuda. Kui me tahame midagi öelda ühiskonnas levinud teemade kohta Eesti aladel pika aja jooksul, siis peaksime leidma tekstikogud, mis seda kajastavad. Samaaegselt peaksime ka veenduma, et materjalid on läbi meie vaadeldava perioodi suhteliselt samasugused. Selles uurimuses võtame aluseks **digiteeritud eesti ajalehed**. Need on

kunagi paberkandjal ilmunud ajalehed, mis on nüüd üles pildistatud ja üle viidud digitaalsesse vormingusse.

Viimase paari kümnendi jooksul on säilitamise eesmärgil Eestis digiteeritud suur hulk trükipärandit. Vanade ajalehtede digiteerimisega alustati aastal 2002 ja praeguseks on oluline osa neist digiteeritud ja kättesaadavaks tehtud eri mäluasutuste lehtedel. Ajalehtedest on kõige suurem kollektsioon Eesti Rahvusraamatukogul (edaspidi RaRa), kus on hetkel digiteeritud üle 7 miljoni lehekülje ajalehti ja perioodikat. Neid materjale on võimalik uurida lähemalt DIGAR-i Eesti artiklite portaalis (edaspidi DEA)[3] ning nende töötluseks andmestikuna on rahvusraamatukogu digilaboris[4] loodud mõned abivahendid. Seal on võimalik saada ülevaadet kollektsiooni sisust (millised ilmunud lehed on digiteeritud ja millisesse arhiivi on nad paigutatud) ning uurida lähemalt selle sisu erinevate parameetrite kaudu. Lisaks on võimalik seal ligi saada ajalehtede täistekstidele, mille põhjal on võimalik luua oma tekstikorpus (vt ptk 4 „Oma korpuse loomine").

Digiteeritud tekst on oma eripäradega: need tekstid on enamasti paberkujult üles pildistatud ning piltide põhjal on omakorda loodud masinloetavad tekstid. Piltidelt teksti tuvastamist nimetatakse **tärktuvastuseks** (tähemärkide ehk tärkide tuvastamine, ingl *optical character recognition*, OCR). Lisaks tähemärkide leidmisele on tarvis need ka omavahel (väljaande tekstid võivad olla kujunduse tõttu fragmenteeritud) ja ajalehe metaandmetega siduda. See, kui hästi on informatsioon teksti kohta koondatud ja esitatud, sõltub andmete hoiustamisest ja digiteerijate prioriteetidest. Ajalehed koosnevad artiklitest, mis võivad ulatuda üle mitme lehekülje või paikneda lehekülje eri osades. Digiteeritud ajalehtede tekstist terviklike artiklite loomine nõuab digiteerijailt palju käsitööd ning seetõttu on paljud lehed digiteeritud ilma lehekülgedelt teksti artikliteks koondamata. Tehnoloogia arenedes muutuvad need etapid aga aina lihtsamaks ning automaatsed lahendused aina võimekamaks ja täpsemaks.

Digiteeritud teksti iseärasuseks on ka see, et digiteerimise käigus on enamasti tekkinud vigu sõnade tuvastamisel. Tehnoloogia edenedes saame neid probleeme üha paremini automaatselt lahendada, aga praegu peame uurimisel võimalike vigadega arvestama.

Käesolevas uuringus on kasutatud DEA kollektsiooni. Need on korpuspõhiste analüüside võimaldamiseks tehtud täistekstidena kättesaadavaks, juhul kui kasutusõigused on seda lubanud. Neile saab hetkel ligi Jupyter Notebooki lahenduse kaudu, kus on võimalik töötlemiseks välja võtta sobiv alamhulk selles (vt juhendit digilabori lehelt[5]).

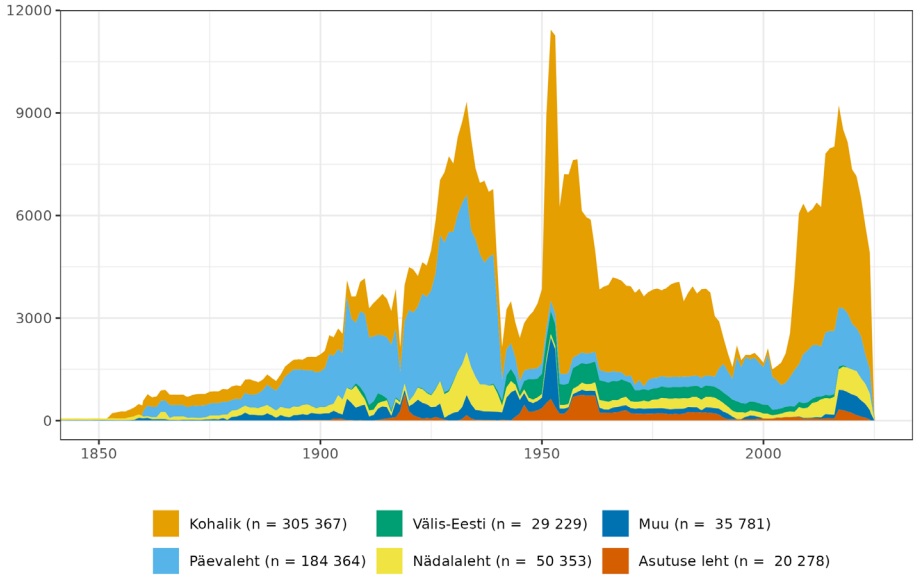
Joonisel 1 on kujutatud DEA tekstikogu sisu terves ulatuses. Sealt näeme, et kollektsiooni maht eri ajaperioodidel on üsna erinev. Enne 20. sajandit on maht

[3] https://dea.digar.ee/

[4] https://digilab.rara.ee/

[5] https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/

väiksem seepärast, et eestikeelseid ja Eestis ilmunud ajalehti ei olnud siis väga palju. Rohkelt on allikaid ajavahemikust 1920–1940, neid on peetud digiteerimisel oluliseks ja samal ajal on nad piisavalt vanad, et mitte autoriõigustega probleeme tekitada. II maailmasõja aegsest ja järgsest perioodist on lehti vähem, osaliselt seepärast, et sõda ja sellele järgnenud okupatsioon ajasid kirjastamismaastiku segi. Nõukogude ajast on kollektsioonis palju kohalikke lehti, ent olulisel kohal on ka Välis-Eesti ajakirjandus. 1990ndad ei ole veel kuigi hästi digiarhiividesse jõudnud. Alates umbes 2005. aastast on arhiiv üpris mahukas, sest tekstid sündisid juba digitaalselt.

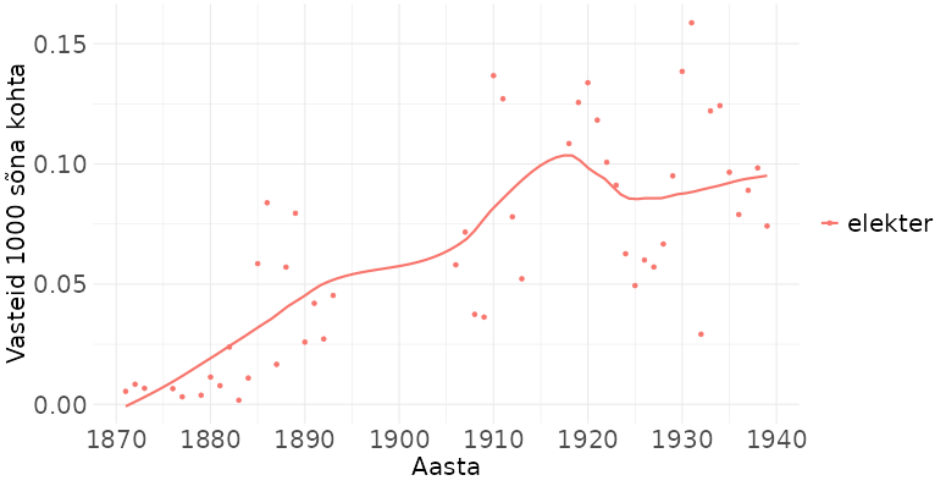


Joonis 1. Rahvusraamatukogu DEA korpuse sisu (seisuga 01.11.2023)

3.3. Valimi loomine

Materjale ajakirjanduskorpusest valides on oluline 1) **ajaperiood** – et materjalid kataks meile olulist ajavahemikku, 2) materjalide **esinduslikkus** – kui me tahame ajalehetekstide kaudu mõista teatud ühiskonnagruppide maailmapilti, peame tagama, et meie valitud materjal saaks neid esindada, 3) materjalide **võrreldavus** – kui me käsitleme pikemat ajaperioodi, siis on oluline, et meie korpuse sisu püsiks sarnane ja analüüsi tulemused ei oleks tingitud allikate vahetumisest (näiteks kohalikes ajalehtedes võivad olla kajastatud hoopis teistlaadi teemad kui kultuurilehtedes). Kui meid huvitab elektritoodete levik Eesti ühiskonnas, siis me saame lähtuda

ajalisel piiritlemisel mõnest tähtsündmusest, nt 1879. a leiutas Edison elektrilise lambipirni, 1907. a asutati Eestis esimene avalik elektrijaam, 1920–1930ndatel aastatel hakati elektrit kasutama majapidamistes laiemalt. Oma teadmiste kontrollimiseks võime vaadata RaRa sõnamitmike loendaja (vt tööriista digilabori lehel[6]) kaudu elektrile viitavate sõnaühendite sagedust DEA näidiskorpuses. Korpus näitab, et elektrist tõesti hakati rääkima 1880ndatel ning iseäranis sagedasti räägiti sellest 1910ndatel ja 1930ndatel aastatel. Oma juhtumiuuringu piirideks võime seega valida näiteks 1880–1940. Elektrist räägiti ka pärast 1940ndaid, aga siin võime panna uurimise jaoks piiri, kuna vaatluse all on piisavalt pikk ajaperiood. Pealegi tõi nõukogude aeg suuri muutusi sellesse, millest ja kuidas räägiti.



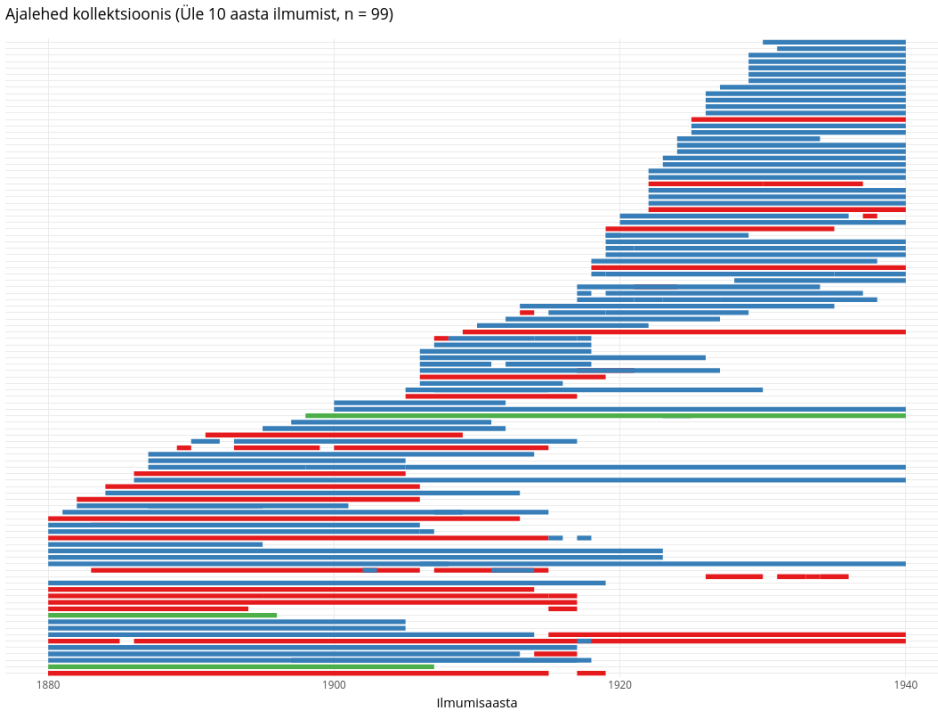
Joonis 2. Lemma *elekter* otsing RaRa sõnamitmike loendajas

Valitud ajavahemikust on andmestikus olemas hulk digiteeritud lehti, mille põhjal võib saada ülevaate elektri esinemisest Eesti meediaruumis. Selleks et ajalehtede digiteerimise seisust ülevaadet saada, on RaRa loonud interaktiivse tööriista[7], mis aitab leida eri arhiividest kindla vaatlusperioodi digiteeritud materjalid ja valida välja allikad konkreetse uurimisküsimuse lahendmiseks. Joonisel 3 on toodud väljavõte 1880–1940 ilmunud ajalehtedest, mille kogu ilmumisperiood oli vähemalt 10 aastat. Joonisel on värviga märgitud ajalehtede paiknemine erinevates arhiivides, sinine märgib RaRa DEA arhiivi, roheline Tartu Ülikooli raamatukogu arhiivi ja punasega on tähistatud digiteerimata ajalehed. Kaaludes võimalikke

[6] https://digilab.rara.ee/tooriistad/sonamitmikud-ajalehtedes/

[7] https://digilab.rara.ee/tooriistad/digiteeritud-ajalehed-eestis/

allikaid, on võimalik kujundada hea alus analüüsideks, mis 1) katab tervet ajaperioodi, 2) on piisavalt esinduslik ja 3) püsiv läbi terve perioodi.

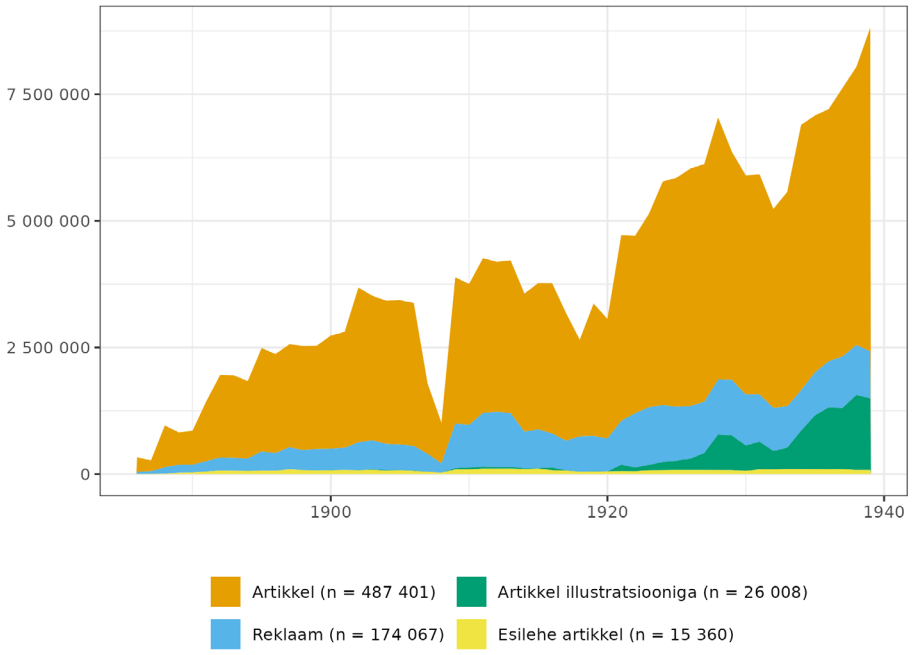


Joonis 3. Väljavõte digilabori digiteeritud ajalehtedest Eestis. Kujutatud on vähemalt 10 aastat ilmunud lehed perioodil 1880–1940. Iga joon tähistab ühe ajalehe ilmumisaega, värviga on märgitud digiarhiiv: sinised jooned viitavad RaRa DEA arhiivile, rohelised Tartu Ülikooli raamatukogu digiarhiivile ja punased jooned digiteerimata ajalehtedele

Paljud lehed ilmusid üpris lühikest aega ja mitmed olid ainult kohaliku ulatusega. Kui meil vahetub korpuses kogu aeg tekstiallikas, siis meil on raske öelda, kas kasvanud tootereklaamide hulk võib olla seotud toote levikuga või sellega, et korpusesse on sattunud väljaanne, mis sisaldab palju reklaame. Korpuse sisu tasakaalustamiseks võiksime välja valida paar suuremat lehte, mis kataks kogu Eestit ja oleksid ilmunud võimalikult pika aja vältel.

Sobiv väljaanne on ajaleht Postimees, mis ilmus püsivalt terves ajavahemikus ja võiks pakkuda korpuses stabiilsuse. Me ei saa välistada olulisi muudatusi ajalehe žanrilises koostises (kui palju sisust moodustasid lühiuudised, arvamusartiklid või reklaam), aga ühe väljaande piires võime eeldada, et žanriline koostis on suhteliselt stabiilne. Postimehe artiklid algavad kollektsioonis aastast 1886, nii me võime

võtta uuringu aluseks vahemikus 1886–1940 ilmunud ajalehenumbrid. Postimehe žanriline koostis sel perioodil on esitatud joonisel 4.



Joonis 4. Postimehe numbrite sisu žanriline jaotus 1886–1940. Kõrgus näitab sõnade hulka aastas artiklitüübi kohta

4. Uuringu läbiviimine

4.1. Otsing

Me soovime leida oma tekstikorpusest elektriga seotud tooteid ja nähtusi perioodist 1886–1940. Sobivate tekstilõikude leidmiseks kasutame märksõnaotsingut – anname arvutile sõna või sõnaosa, mis sisaldub tekstilõikudes, mis meile olulised on (vt ka õpiku ptk 5.2.1 „Konkordants ja KWIC“). Siin uuringus teeme seda R-is (versioon 4.3.1), kasutades RaRa paketti tekstide ligipääsuks ja analüüsiks[8]. Uuringus kasutatud koodi on võimalik teksti kõrval jälgida koodihoidlas[9].

[8] https://digilab.rara.ee/tooriistad/ligipaas-dea-tekstidele/

[9] https://osf.io/xqzsf/

Päringutulemuste hindamisel on olulisteks mõisteteks **saagis** (ingl *recall*) ja **täpsus** (ingl *precision*, vt ptk 3.1.3 „Märgenduse täpsuse hindamine"). Need iseloomustavad meie otsingu tulemusi sellega võrreldes, mida me soovisime leida. Loodud otsingus jääb enamasti mingi osa korpuses esinevast asjakohasest informatsioonist kõrvale (seda iseloomustab mõiste *saagis*) ja meie leidude seas on enamasti ka mõned tekstijupid, mis tegelikult ei sisalda meile vajalikku informatsiooni (seda iseloomustab mõiste *täpsus*).

Näiteks otsides märksõnaga *elektri*, võib meil jääda leidmata hulk tooteid, mille puhul peetakse loomulikuks, et see toimib elektriga (nt tänapäeval tänavavalgustus või mobiiltelefonid) ning seetõttu võib meie saagis olla nt 75% võimalikust ehk 0,75. Kui me aga kasutame otsisõnana *mootor*it, võib tulemusest ainult 20% olla elektrimootorid (ja mitte aurumootorid või bensiinimootorid), siis on meie otsingu täpsus 20% ehk 0,2. Otsinguid koostades tasub otsida tasakaalu täpsuse ja saagise vahel. Selleks kasutatakse vahel **F1-skoori**, mis on täpsuse ja saagise harmooniline keskmine. Saagise 0,75 ja täpsuse 0,2 harmooniline keskmine on 0,32. Saagise 0,75 ja täpsuse 0,9 harmooniline keskmine on 0,81. Üldiselt on harmooniline keskmine lähemal paari madalamale väärtusele kui aritmeetiline keskmine oleks.

Pole päris kindlat reeglit, mis on infootsingu juures hea F1-skoor, kuna see sõltub väga küsimuse iseloomust. Rusikareeglina võib pakkuda, et F1-skoor 0,9 võiks olla päris hea enamike probleemide juures, väärtuse 0,5–0,8 vahel tasuks mõelda, kas loodud otsing töötab püstitatud küsimuse kontekstis piisavalt hästi, ning väiksema väärtuse puhul tasub mõelda, kas otsingut saab parandada. Iga otsingu juures ei pea seda küll välja arvutama, tihti võib selle hindamine keeruline olla. Sellele temaatikale tasub aga mõelda otsingut koostades ja uurijana tuleb teadlik olla, kuidas märksõnade valik võib tulemusi mõjutada: millised kasutusjuhud võivad olla otsingutulemuste seast puudu ning millised vasted ei pruugi meie tulemuste hulka sobida. (Vt ka K. Muischneki ja S. Orasmaa näidisuurimust nimeüksuste märgendamisest, milles F1-skoori on kasutatud märgendajatevahelise kooskõla hindamiseks.)

Kasutades otsingutes üldisemaid termineid (nt *keskkond*), saame mitmekesisemaid tulemusi, aga võime leidude hulka saada ka neid, millest me tol hetkel ei huvitunud (nt *looduskeskkonna* kõrval ka *koolikeskkond* või *majanduskeskkond*). Kasutades otsingutes spetsiifilisemaid otsingusõnu (nt *looduskaitsjad* või *looduskeskkonna kaitse*), saame täpsemaid tulemusi, aga meie otsingu hulgast võib jääda välja hulk tulemusi, mis on küll temaatiliselt seotud, aga ei kasuta täpselt sama fraasi.

See probleem on veel suurem digiteeritud ja vanemate tekstide puhul. Tundes valdkonda, võib tänapäeva tekstides olla lihtne ära arvata suurt hulka märksõnu, aga tegeledes vanemate tekstidega, peab arvestama muutustega keeles ja kirjaviisis: meile tuttavad nähtused võivad olla väljendatud teisiti. Digiteeritud tekstide puhul on sagedad ka digiteerimisel tekkinud vead, kus mõni täht sõnas võib olla läinud

vahetusse sarnase väljanägemisega tähe või sümboli(te)ga. Seetõttu soovitatakse kasutada üldisemaid fraase ning teha otsingut paindlikumaks nii, et tulemuste hulka sobituksid ka mõned digiteerimisvigadega sõnad. Üldine juhtmõte siin on, et peame arvestama ja eeldama, et korpused võivad meid üllatada nii sisult, kirjaviisidelt kui ka kvaliteedilt (Nicholson 2012; Greenfield 2013).

4.2. *Elektri* otsimine

Selles peatükis on eesmärk leida tekstist elektriga seotud tooteid. Koostame selleks märksõnaotsingu. Kui me otsime sõna *elekter*, siis jäävad tulemustest välja käändevormid, kui otsime *elektri*, on vaja eraldi otsida nimetavas käändes *elekter*. Lihtsustamiseks saame otsida *elekt*, mis võib samas laiendada võimalike valevastete hulka. Tärktuvastusel võivad olla tekkinud valevormid *elektar* või *elektn*, mida see päring kaasab. Päring hõlmab ka liitsõnu (nt *hüdroelekter*, *elektrijaam*). Küll aga jäävad päringutulemustest välja vormid, kus on vigu järjendi *elekt* sees, näiteks *l* on asendatud *1*-ga (*e1ekt*). Samuti jäävad kõrvale sõnad, kus on viide elektrile, aga mõnel teisel viisil, näiteks tüvedest *vägi* või *säde* moodustatuna.

Selle päringuga leiame kokku 56 457 vastet 30 262 tekstist. Ülevaate saamiseks koondame järjendit *elekt* sisaldavad sõnad ühte tabelisse. Selleks sõnestame (ingl *tokenization,* vt ka ptk 3.2.1 „Morfoloogiline märgendamine") oma teksti sõnadeks ja võtame välja kõik sõnad, mis sisaldavad *elekt*. Üle kõigi leidude on erinevaid esinemiskujusid palju – 6155. Neist 6155 järjendit *elekt* sisaldavast sõnast ainult 367 esineb vähemalt 10 korda, samas moodustavad nad kokku 85% kõigist sõnade esinemisjuhtudest (47 711 korda). Seetõttu saame üpris hea ülevaate leidudest ka ainult sagedamini esinevate vormide põhjal.

Tabel 1. Kõige levinumad vasted märksõnaotsingule. Jrk näitab järjekorranumbrit sagedusloendis, N näitab esinemiskordade arvu

Jrk	Vaste	N	Jrk	Vaste	N
1	*elektri*	13735	26	*elektrita*	236
2	*elektrijaama*	3676	27	*elektrijaamast*	227
3	*elekter*	3263	28	*elektriliini*	221
4	*elektriga*	1937	29	*elektrivoolu*	216
5	*elektrivalg*	1839	30	*elektrienergia*	189
6	*elektrijaam*	1203	31	*elektrivalgustus*	183
7	*elektriwalg*	1151	32	*elektrijaamale*	168
8	*elektriv*	1060	33	*elektrimootorid*	153

Jrk	Vaste	N	Jrk	Vaste	N
9	*elektrit*	1037	34	*elektrilamp*	149
10	*elektrotehnika*	763	35	*elektrist*	146
11	*elektriwoolu*	720	36	*selektsiooni*	146
12	*elektriwalgustusega*	652	37	*elektriraudtee*	140
13	*elektriwalgustuse*	506	38	*elektrivalgus*	137
14	*elektriwalgustus*	484	39	*elektrilambid*	134
15	*elektrivalgust*	446	40	*elektrivalgusega*	134
16	*elektr*	363	41	*elektrijaamade*	130
17	*elektriw*	316	42	*elektrijõu*	129
18	*elektrimootor*	311	43	*elektrienergiat*	124
19	*elektrijaamas*	307	44	*elektrifitseerimine*	122
20	*elektro*	307	45	*elektriwool*	115
21	*elektrivalgustusega*	306	46	*elektriliinide*	111
22	*elekt*	298	47	*elektriwalgus*	111
23	*elektriwalgust*	293	48	*castelekt*	110
24	*elektrifitseerimise*	281	49	*elektrivarustusega*	109
25	*elektrijõul*	255	50	*elektrimootori*	108

Esimese asjana peaksime vaatama tulemusi kriitiliselt: kas leitud sõnad vastavad meie ootustele ning on tõepoolest seotud elektriga? Tabelis 1 on kujutatud 50 levinumat vastet, koodihoidlas on näha terve tabel. Näeme, et levinumad sõnad on ootuspäraselt *elektri*, *elektrijaama*, *elekter* ja *elektriga*. Levinumate sõnade seas on ka mõned, mille puhul tähendus pole kindel ning mida võiksime kontrollida, näiteks *elektrivalg* ja *elektriv*. Kontrollides näeme, et nad viitavad elektrivalgustusele ja on kas lühendid või on pikk sõna läinud digiteerimise käigus tükkideks – seega võib pidada tulemuste jaoks sobivaks. Vaadates tulemustes edasi näeme, et 146 korda esineb nende seas *selektsiooni* ning tulemuste seas paistab ka kummaline *castelekt* 110 korda ja *lastelekt* 98 korda. Kontrollimisel selgub, et tegemist on tüüpveaga: tihti on *leht* loetud digiteerimisel kui *lekt*, mistõttu on sattunud meie otsingusse ka *lastelekt*, *pilkelekt* ja muud säärast. Need otsingusse sattunud vormid eemaldame oma edasisest analüüsist; selleks teeme tabelisse uue tulba ja märgime

sinna, kui sõna ei vasta meie ootustele. Kõrvale jätame siin näiteks elektronid, mis puutuvad füüsikasse üldisemalt, ja näidendi „Elektra".

Nende tulemuste põhjal võime üldistada, et sobivate sõnade puhul on *elekt* enamasti sõna alguses, ja võime kõrvale jätta kõik sõnad, kus see nii ei ole. Igaks juhuks võime siiski teha loetelu kõigist sõnadest meie tabelis, kus ei ole *elekt* sõna alguses. Neid on kokku 544, mille kiire ülevaatamine ei käi ka üle jõu. Saame korrastada ka muud süstemaatilised varieeruvused, nt *v* ja *w* vaheldumise.

Arvestades nende leitud variantidega, saame moodustada reeglid, mille alusel välistada otsingust neile lisatingimustele mitte vastavad vormid. Tingimustele mitte vastavaid leide oli kokku 1519 esinemisjuhtu ning parandatud otsingu tulemuseks on nüüd 54 938 leidu. Oluline on tähele panna, et enne otsingutulemustesse vaatamist ei saanud me teada kõiki variante elektriga seotud sõnadest ja ka sellega seostamata sõnadest. Oleksime võinud kõrvale jätta *selekt*, aga siis oleks meil kõrvale jäänud ka näiteks *väliselekter* ja *walguselekter*. Alustades alguses üldisema otsinguga saab katta suurem hulga erinevaid vorme ehk parema saagise. Otsingu tulemusi analüüsides saab aga järk-järgult parandada ka otsingu täpsust.

4.3. Elektriga seotud nähtused

Vaadeldes nähtuste ja toodete levikut püüame aru saada, milliste toodete ja nähtustega tegemist oli. Selle jaoks tuleks meil need leiud märgendada ja jaotada tüüpidesse. Märgendamise kohta on rohkem juttu õpiku ptk-s 3. Siinkohal mõtleme märgendamisel all seda, et mõtleme läbi, millised tüübid on materjalis olulised, ning lisame selle info igale kasutusjuhule. See, mis kategooriaid on määrata, tuleneb tihti uurimisala traditsioonidest ja varasematest uuringutest. Mõne vähemuuritud teema puhul tuleb aga tüübid andmetele tuginedes uurijal endal välja kujundada. Elektritooted 20. sajandi algul on just sedalaadi küsimus.

Paigutame kõik elektriga seotud sõnade leiud tabelisse ning märgime igaühe kõrvale, mis tüüpi eseme või nähtusega tegemist on. Praegusel juhul püüame andmestikus määratleda eri tüüpi esemeid ja nähtusi.

Esmalt paigutame kokku ühe nähtuse kohta kasutatud erinevad terminid (näiteks *elektrijõud* ja *elektrivägi*), nende erinevad käändevormid (*elektrijõuga* ja *elektrijõu*) ja ühtlustame sagedased digiteerimisvead (kui *elektri* asemel on *elektn*). Seejärel määrame iga sõna tüübi. Meil on elektrit sisaldavaid sõnu kokku 6155. Märgendamisel võime teha mõningaid praktilisi lihtsustusi ja märgendada levinumad tüübid esimesena. Märgendame sõnad 32 kategooriasse: näiteks koondame *elektrijaamad* ja *elektrivabrikud* kokku *jaama* mõiste alla ning *elektriraudtee* ja *elektrirongi rongi* mõiste alla.

Eraldi tasub mõelda meid huvitavatele üksustele; meile vajalik info toote või nähtuse sisu kohta ei pruugi paikneda pelgalt ühes sõnas. Kui me vaatame leidude kontekste lähemalt, siis leiame sealt nii *elektrimootori*, *elektri-mootori* kui ka *elektri mootori*. See tähendab, et sõnadevaheline tühik ei pruugi siin eraldada veel

mõisteid. Kui soovime käsitleda *elektri-mootorit* ühe üksusena, on soovitav sõnestamisel mitte sidekriipse kõrvale jätta. Laiendame vastavalt ka oma otsingut siin.

	A	B	C	D
27	elektrivalgusega	184		valgus
28	elektrijaamale	168		jaam
29	elektrimootorid	154		mootor
30	elektrilamp	148		valgus
31	elektrivool	144		infra
32	elektrist	144		yld
33	elektriraudtee	135		rong
34	elektrilambid	134		valgus
35	elektrijaamade	130		jaam
36	elektrijõu	127		jõud
37	elektri-aurulokke	124		lokid
38	elektrienergiat	123		yld
39	elektrifitseerimine	122		elektrifitseerimise
40	elektrivalguse	109		valgus
41	elektrivarustusega	109		infra
42	elektrimootori	108		mootor
43	elektriliinide	106		infra
44	elektrimootorit	103		mootor

Joonis 5. Väljavõte märgenduste tabelist. Nähtavad tulbad on sõna, esinemiste arv, kommentaarid (joonisel tühi tulp) ja määratud tüüp

Kui meil on olemas arusaam tüüpidest, mida püüame leida, saame protsessi veidi automatiseerida, näiteks kasutada regulaaravaldisi (vt ptk 5.2.2 „Regulaaravaldised“). Näiteks *elektromagneti* leidmiseks võime kasutada regulaaravaldist `magne(e)?[td]`, mis hõlmab nii magneteid kui magneetilist kui ka näiteks vormi *elektromagneedi*. Loeme selle faili töövoosse sisse ning kasutame seda, et märgendada ka haruldasemad tüübid.

Kui paljude leidude puhul on lihtne aru saada nende tähendusest sõna enda järgi, siis mõned leiud võivad meis tekitada küsimusi ja arusaamiseks peame vaatlema sõnu nende esinemiskontekstis. Näiteks mis on *elektrimalg*, *elektriv* ja *elektrivalg*? Teksti vaadates näeme, et need viitavad elektrivalgustusele, kas lühendatud kujul või tuvastusveana (*w* on tuvastatud *m*-ina).

Märgendamise põhjal moodustame uue tabeli, kus iga leiu kõrval on selle tüüp ning metaandmed ilmumisaja, artiklitüübi ja muu sellise kohta. Nüüd, kui meil on andmed märgendatud, saame neid analüüsida ja vaadata, mis trende seal näeme.

5. Analüüs

Regulaaravaldiste abil sai kategoriseeritud 50 454 leidu 54 666st ehk 92%. Kokku saime 36 tüüpi, milles vähemalt üks vorm esines korpuses vähemalt 10 korda. Kolm levinumat tüüpi olid üldine (28%), elektrivalgus (24%) ja elektrijaam (13%). Järgmised kolm rääkisid elektrist suuremas ühiskondlikus plaanis: infrastruktuur (elektriliinid, elektrivõrgud jne, 8%), elektrotehnika ja elektromehaanika valdkond (2%) ja elektrifitseerimine (2%). Ajalehtedes niisiis käsitleti rohkem elektri tootmist, elektri levikut ühiskonnas ja ka elektrit valdkonnana. Toodetest räägiti vähem.

Sagedasimad elektripõhised tooted olid *elektrimootor* (1109 korda), *elektrirong* (807 korda), *elektrilokid* (609 korda), *elektrikell* (318 korda), *elektriteater* (191 korda) ja *elektritriikraud* (190 korda). Märgendatud said veel *elektripliit, elektritramm, elektritool, elektridünamo, elektriorel, elektriahi*, mida esines vähem. Sirvides märgendamata tüüpe, leiame veel mõned tooted – *elektrikraana, elektrivann, elektriküünlad, elektrikeetja, elektrijuukselõikamismasin, elektrilatern, elektritaskulamp* ja *elektriauto*. Nende esinemiskorrad ei olnud piisavalt sagedased ning jäävad uuringust välja.

5.1. Sageduste muutused ajas

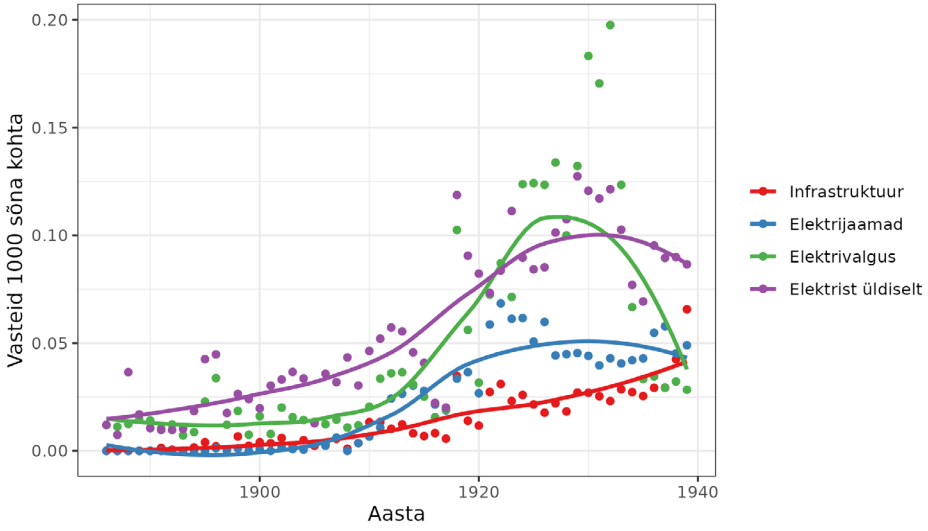
Sõnade sagedust kasutatakse tihti mõne teema või arutelu leviku hindamiseks. Näiteks on seeläbi mõõdetud mõne kuulsuse nagu Che Guevara või Marilyn Monroe tuntust või siis hoopis jäätise, pitsa ja *sushi* levikut ingliskeelsetes raamatutes – neid otsinguid saab praegu teha Google Ngramsi otsingus (Michel jt 2011). Sarnaselt on tekstide kaudu mõõdetud mõnede tehnoloogiate, nagu telegraaf või raadio, levikut ning suurte haiguspuhangute toimumist (Lansdall-Welfare jt 2017). Suured tekstikorpused võimaldavad mõõta seda, kuidas ja kui palju ajalehtedes mingeid teemasid mingil perioodil käsitleti ning kui palju need seeläbi ühiskondlikku tähelepanu pälvisid. Siin on ka näiteks esile toodud üldisi trende sellest, kuidas ilukirjandusteosed muutusid vähem positiivseks läbi aja (Morin & Acerbi 2017) või teadustööd rohkem positiivseks (Vinkers, Tijdink & Otte 2015). Trende hinnates on oluline säilitada kriitilist pilku ning mõelda läbi, mida sageduse muutumine tähendab (nt kui mikrolaineahjudest enam ei räägita, ei pruugi see tähendada, et neid enam ei kasutata).

Sagedusmuutusi hinnates on tähtis, et korpuse osad oleksid omavahel võrreldavad – kas žanride mahult või läbi eri perioodide. Mõningaid parameetreid on lihtne ühtlustada – kui muutub tekstide maht, saame kasutada **suhtelise sageduse mõõdikuid**, mis arvestavad korpuse suurusega (vt ptk 5.2.4.1 „Sagedusloendi põhjal keele uurimine"). Keerulisem on, kui ajalehte lisanduvad uued žanrid – näiteks kultuuriuudised ja spordisektsioon. Sel juhul väheneb kõigi teiste žanride osakaal ja näiteks tehnikauudiseid või reklaami võib kokku olla lehes suhteliselt vähem, kuigi nende teemade juures muutust ei toimunudki. Kui me moodustame ise korpuse, on võimalik neid osakaale täpselt jälgida. Suuremate tekstikorpuste puhul

võib see olla keerulisem. Näiteks on Google Booksi korpust kritiseeritud just sellest lähtuvalt – kuna tekstide valikut ei ole kerge jälgida ning korpuse žanriline koostis muutub ajas, siis ei saa mitmeid analüüse selle põhjal teha (Pechenick, Danforth & Dodds 2015).

Selles uurimuses saame sagedusi vaadata otsingutulemuste kaupa või tekstižanride kaupa. Nimelt on meie korpuses eristatud tavalised artiklid ja reklaamtekstid. Korpuses on kokku 174 067 reklaamkülge ja 513 008 tavalist artiklit; reklaamid moodustavad korpusest 25%. Otsingutulemuste seas on 15 607 reklaami ja 13 530 tavalist artiklit – reklaamid moodustavad tervelt 54%. Otsitud elektrialased tooted on selgelt rohkem esindatud reklaamides kui tavalistes artiklites. Elektriga seotud nähtusi leidsime 9% reklaamidest, aga 2,6% tavalistest artiklitest. Saame seda vaadata ka läbi aja: elektritoodete reklaamimine kasvas perioodi jooksul sujuvalt, 19. sajandil oli see 2,5% ja vähem, sajandi algul juba 5% ümber ja alates 1920ndatest tõusis see juba 10%ni või rohkem. Reklaamküljed sisaldavad väga tihti mitmeid reklaame, mistõttu ei saa öelda, et 10% reklaamitud toodetest olid elektriga seotud, aga need said kindlasti üha suuremaks osaks inimeste igapäevaelust.

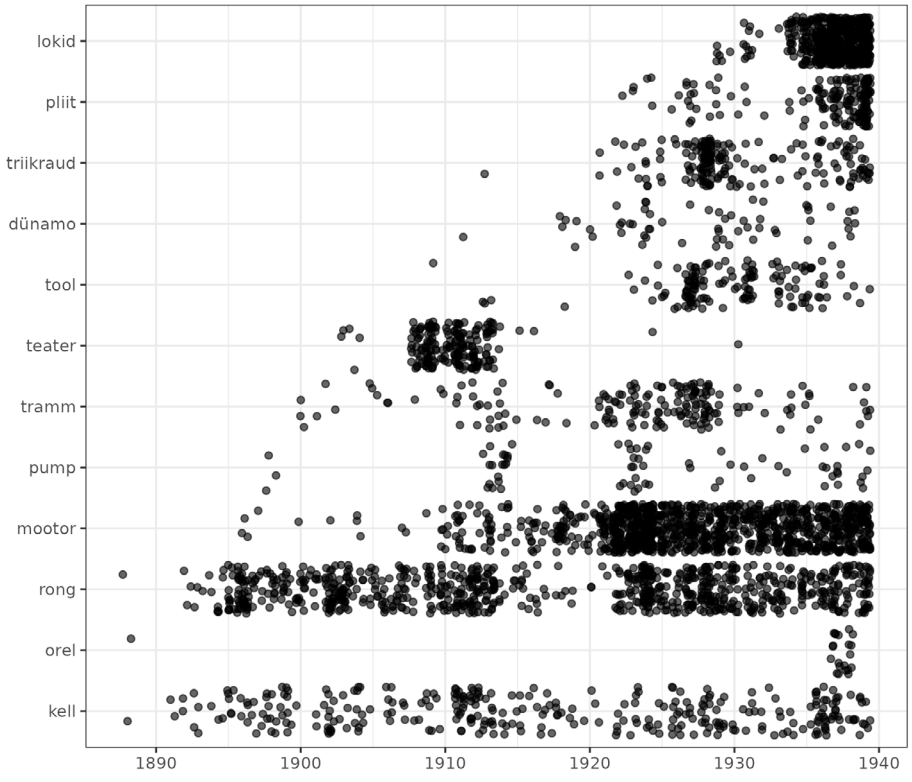
Kui leiud on suhteliselt sagedased, võime võrrelda nende osakaalu korpuses. Vaatame lähemalt sagedasemate tüüpide levikut (joonis 6). Need on 1) elekter üldiselt, 2) elektrivalgus, 3) elektrijaam ja 4) elektri infrastruktuur. Nende tüüpide esinemine 1000 sõne kohta on kujutatud joonisel 6. Kõigi nende tüüpide kasutussagedus kasvas perioodi jooksul: elekter üldiselt 0,02 sõnelt 0,1 sõneni, elektrivalgus umbes samas vahemikus, elektrijaamad ja elektriga seotud infrastruktuur



Joonis 6. Nelja nähtusetüübi – elekter üldiselt, elektrivalgus, elektrijaamad ja elektri infrastruktuur – kasutussagedused 1000 sõne kohta Postimehes 1886–1940

nullist 0,05 sõneni 1000 sõne kohta. Elektril üldiselt ja infrastruktuuril on kasv sujuv, elektrijaamadel ja elektrivalgusel toimub 1920ndate algul hüppeline kasv, 1930ndate teises pooles räägitakse elektrivalgusest vähem, võimalik, et see on siis juba tavaliseks saanud ja ei vaja eraldi rõhutamist.

Kui leiud on haruldased, võivad meid huvitada esmamainimised ja see, kas neist räägiti üldjoontes palju või vähe. Selle vaatlemiseks oleme leiud paigutanud joonisele 7, näidates iga esinemiskorda punktina. Punktid on paigutatud veidi hajutatult, et me näeksime, kui palju punkte ühele aastale asetub. Graafikule on valitud 12 levinumat toodet ja eset esmamainimise järjekorras – elektrilisest kellast elektrilokkideni. Näeme ka siin mõningaid trende: elektridünamost ja elektripumbast räägiti läbivalt suhteliselt harva, elektrilokkide, elektripliidi ja elektrimootori puhul räägitakse neist alguses harva ja seejärel mingil hetkel juba massiliselt – eeldatavasti siis, kui tooted on jõudnud reklaamidesse.



Joonis 7. Levinud elektriga seotud toodete mainimiskorrad. Iga punkt tähistab üht mainimiskorda, need on hajutatud vertikaalteljel, et näidata ajalist jaotumist. Tooted on järjestatud esmamainimise järjekorras

Näeme ka huvitavaid muutusi perioodi jooksul – näiteks elektriteatrist räägitakse lühikest aega väga palju aastail 1909–1912. Sõnaga *elektriteater* viidati toona kinole või kinoprojektorile ning peagi oli see nimetus asendunud *kino*ga. Pliit ja elektrilokid mõlemad saavad populaarseks 1930ndate teisel poolel, aga elektripliidist oli juttu juba terve 1920ndate vältel. Triikraual näeme suurt kampaaniat 1920ndate teisel poolel – see võib olla seotud mõne konkreetse reklaamikampaaniaga. Elektripumbast räägitakse mõnel aastal palju ja siis on jälle pikk paus – võimalik, et need ei olnud toona nii levinud ja neid kasutati ainult mõne suurema õnnetuse puhul. Kõige sujuvam kasv tundub olevat elektrimootoritel – esimestest mainimistest 1890ndate lõpus saavad nad järk-järgult tavalisemaks ja on 1920ndateks juba masstooted. Elektrimootorite puhul on huvitav võrrelda nende levikut teistsuguste mootoritega.

Tulemuste vaatlemisel võime uurida tekste ka lähemalt, näiteks selleks, et välja selgitada, kas elektrioreli esmamainimisel mõeldi tõesti elektriorelit või on see valesti määratletud. Sama ka pumba ja triikrauaga. Selleks saame võtta otsingust välja leidude originaalkontekstid. Näeme, et need on tõesti õiged leiud ja omamoodi huvitavad. 1888. aastal mainitakse elektrilist orelit seoses selle leiutamisega. 1898. aastal räägitakse elektrilisest pumbast aga juba üpris igapäevases võtmes – ilmselt ei olnud sellest varem põhjust ajalehes rääkida, kuigi nähtus oli juba levinud. 1913. aastal on tegemist varase elektritriikraua reklaamiga, kuid massilisema leviku sai see alles 1920ndatel. Huvitav on ka elektrikell – sellest on olnud juttu perioodi jooksul läbivalt, aga aeg tundub varajane elektrooniliste käekellade jaoks. Kas oskad arvata, millega võiks olla tegemist?

5.2. Koosesinevad sõnad

Tekstides esinevate märksõnade puhul võib osutuda tähenduslikuks nende esinemiskontekst. Märksõna vahetust ümbrusest saab uurija aru, millega seoses märksõnast räägiti, ja nii jälgida teatud teemade kerkimist ja kadumist märksõnade ümber. Näiteks reklaamkuulutustes võib leida sõnu *ostan, müün, vahetan, eurot, krooni* jms. Nii saame vaadelda, kui palju räägiti ja kirjutati elektrist reklaamides ja tavalistes ajaleheartiklites.

Lisaks on iseloomulik, et sarnase tähendusega sõnad esinevad sarnastes kontekstides, st neid ümbritsevad tihti samad sõnad. Sellele on üles ehitatud mitmeid meetodeid sõnade tähendussarnasuste tuvastamiseks. Näiteks saab korpuses jälgida sõnade *loodus* ja *keskkond* kasutuse sarnastumist või sõna *propaganda* muutumist informatsiooni levitamisest spetsiifilisemalt poliitilise ja ideoloogilise mõjutustegevuse kirjeldamiseks.

Võime kasutada märksõnu ümbritsevaid kontekste kahel eesmärgil: selleks, et välja selgitada, millistele objektidele märksõnad täpsemalt viitavad, ning selleks, et paremini mõista, kuidas neid objekte toona kasutati. Võtame ette varem toodud 12 elektriga seotud toodet või nähtust ning otsime välja märksõnale eelneva

ja järgneva 100 tähemärgi laiuse konteksti. Nende põhjal saame teha ühe lihtsa sagedusanalüüsi (vt ptk 5.2.4 „Sõnavara analüüs: sagedusloendid"), kui vaatame, millised sõnad esinevad kõige sagedamini valitud toodete kontekstis. Suures osas on need lihtsalt eesti keeles sagedased sõnad. Sisulisema tulemuse saamiseks võime rakendada **stoppsõnade** nimekirja, st sõnade nimekirja, mille analüüsist eemaldame. Selliseid nimekirju on erinevaid, aga neid võib ka ise oma uuringu jaoks luua või kohandada. Siin kasutame üht eestikeelse ilukirjanduse põhjal koostatud nimekirja (Uiboaed 2018).

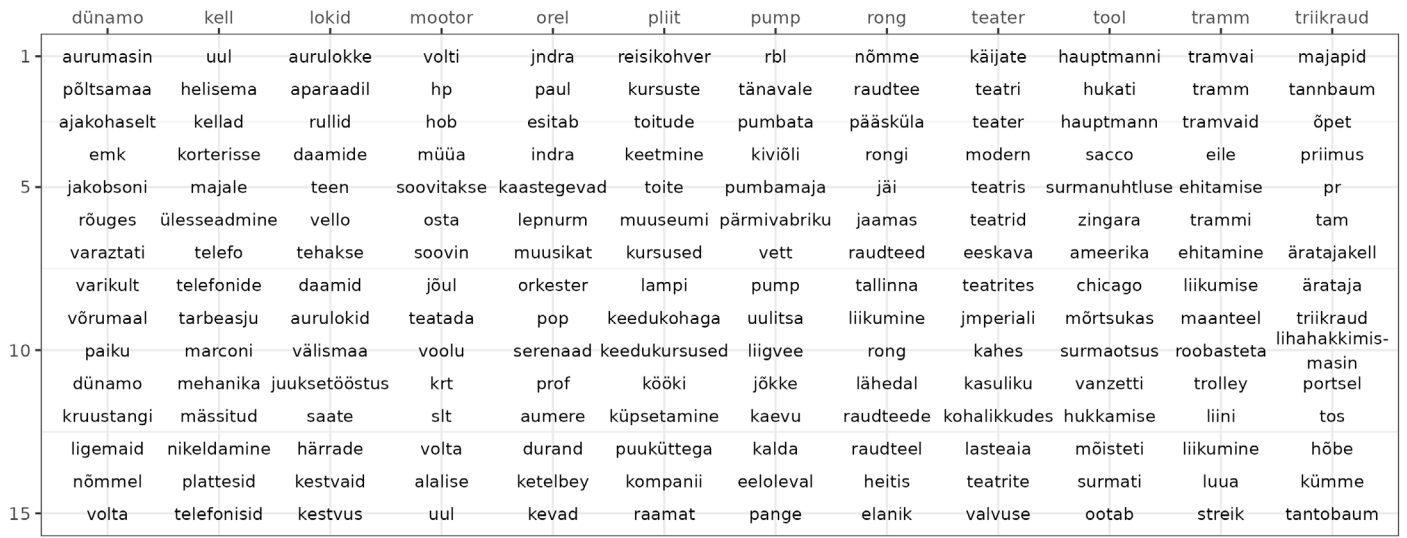
Joonisel 8 on kujutatud iga leitud tüübiga sagedasti koos esinevad sõnad, mis ei olnud meie stoppsõnade nimekirjas. Näeme seal ootuspäraseid seoseid. Elektrirong on seotud raudtee ja kohanimedega, elektripump vee ja tuletõrjega, ja elektrilokid rullide ja daamidega. Näeme ka veidi huvitavamaid seoseid: elektrimootoreid osteti, müüdi ja vahetati, elektrilokkidest rääkimisel oli oluline märksõna *välismaa* ja elektriipliit seostub reisikohvri ja kursustega. Mitmel pool korduvad lühendid *kr, nr, jne, tn, tän*. Neid sõnu ei ole stoppsõnade nimekirja kaasatud, aga selles korpuses on need nii tavalised, et võiksime need samuti välja jätta.

Järgmiseks kasutame võtmesõnade analüüsi (vt ptk 5.2.6 „Võtmesõnad"). **Võtmesõnade analüüs** püüab esile tuua sõnu, mis on iseloomulikud ühele tekstile või korpusele (fookuskorpusele) võrreldes teiste tekstidega või korpustega (referentskorpusega). Selles analüüsis kasutame võtmesõnade leidmiseks tf-idf-mõõdikut (ingl *term frequency – inverse document frequency*, *tf-idf*, vt (Manning, Raghavan & Schütze 2008: 108–110). Mõõdiku nimi viitab selle sisule: me jagame sõna sageduse selles dokumendis dokumentide hulgaga, kus seda sõna leidub. Sõnade puhul, mis esinevad paljudes tekstides, jagame nende sageduse läbi suurema numbriga ning tf-idf-mõõdiku järgi saavad suurema väärtuse sõnad, mis on vaid vähestes tekstides. Meil on võimalik võrrelda tootetüüpe teineteisega: millised sõnad iseloomustavad just seda toodet võrreldes teiste toodetega.

Joonisel 9 on toodud igale tootetüübile iseloomulikud sõnad, mis aitavad mõista tootetüüpide sisu. Elektripumbaga seoses ei räägita enam linnast või tööst ega ka Peterburist, vaid esile on tõusnud sõnad *liigvee, jõkke, kaevu, kalda*. Elektripliidi puhul on esile tulnud nii keetmine kui küpsetamine ja näiteks *keedukursused*. Selgub ka elektrikella taust: esimesed sõnad on seotud korterite ja majadega: *uul, korterisse, majale, ülesseadmine*. Võib arvata, et tegu ei ole veel elektrilise käekellaga, vaid hoopis uksekellaga. Näeme ka, et elektrimootoreid soovitakse jätkuvalt osta, müüa ja vahetada – teistel toodetel nii palju järelturgu ei olnud. Elektri(auru) lokid on seotud juuksetööstusega, aga mängu on tulnud ka härrad.

Tüüpide sisu täpsemaks mõistmiseks tasub meil vaadata uuesti sõnade esinemiskontekste. Nüüd, kui meil on antud mõned kaasnevad märksõnad, saame otsida, mis kontekstides esinesid koos *pliit* ja *kursused* või *aurulokid, daamid* ja *härrad*. See annab põnevat infot, näiteks et Tartu elektrijaam korraldas elektripliidi tulekul selle kasutamiseks kursusi (näide 1). Kuigi elektrilokid on suunatud daamidele ja härrasid kutsutakse juuksurisse, siis ajalehest saame teada, et oli ka vähemalt

Joonis 8. Analüüsis leitud tootetüüpidega sagedasti koosesinevad sõnad



	dünamo	kell	lokid	mootor	orel	pliit	pump	rong	teater	tool	tramm	triikraud
1	aurumasin	uul	aurulokke	volti	jndra	reisikohver	rbl	nõmme	käijate	hauptmanni	tramvai	majapid
	põltsamaa	helisema	aparaadil	hp	paul	kursuste	tänavale	raudtee	teatri	hukati	tramm	tannbaum
	ajakohaselt	kellad	rullid	hob	esitab	toitude	pumbata	pääsküla	teater	hauptmann	tramvaid	õpet
	emk	korterisse	daamide	müüa	indra	keetmine	kiviõli	rongi	modern	sacco	eile	priimus
5	jakobsoni	majale	teen	soovitakse	kaastegevad	toite	pumbamaja	jäi	teatris	surmanuhtluse	ehitamise	pr
	rõuges	ülesseadmine	vello	osta	lepnurm	muuseumi	pärmivabriku	jaamas	teatrid	zingara	trammi	tam
	varaztati	telefo	tehakse	soovin	muusikat	kursused	vett	raudteed	eeskava	ameerika	ehitamine	äratajakell
	varikult	telefonide	daamid	jõul	orkester	lampi	pump	tallinna	teatrites	chicago	liikumise	ärataja
	võrumaal	tarbeasju	aurulokid	teatada	pop	keedukohaga	uulitsa	liikumine	jmperiali	mõrtsukas	maanteel	triikraud
10	paiku	marconi	välismaa	voolu	serenaad	keedukursused	liigvee	rong	kahes	surmaotsus	roobasteta	lihahakkimis-masin
	dünamo	mehanika	juuksetööstus	krt	prof	kööki	jõkke	lähedal	kasuliku	vanzetti	trolley	portsel
	kruustangi	mässitud	saate	slt	aumere	küpsetamine	kaevu	raudteede	kohalikkudes	hukkamise	liini	tos
	ligemaid	nikeldamine	härrade	volta	durand	puuküttega	kalda	raudteel	lasteaia	mõisteti	liikumine	hõbe
	nõmmel	plattesid	kestvaid	alalise	ketelbey	kompanii	eeloleval	heitis	teatrite	surmati	luua	kümme
15	volta	telefonisid	kestvus	uul	kevad	raamat	pange	elanik	valvuse	ootab	streik	tantobaum

Joonis 9. Analüüsis leitud tootetüüpidega seotud võtmesõnad (*tf-idf*)

üks härra, kes 1935. aastal juba kolmandat aastat lasi endale elektrilokke teha (näide 2). *Juuksetööstused* on aga juuksurid ja neid oli tegutsemas Eestis mitmeid.

(1) „Tartu linna elektrijaama korraldusel toimuvad juba pikemat aega keedukursused elekterpliitidel. kus selgitatakse otstarbekat ja kokkuhoidlikku keetmist ning praadimist elektri abil." (16. aprill 1939)
(2) „Ligp. daamid! Kõige parem isiklik jõulukink on teile elegantsed elektriaurulokid mida võite omada H. PAURSON'i Ht V daamide ja härrade juuksetööstuses / Aleksandri tän 78-a. Töö korralik, hind odav ja kokkuleppel." (10. detsember 1935)

6. Automatiseerimise võimalused

Selles peatükis kasutasime juhtumiuuringu tegemiseks suhteliselt lihtsaid vahendeid. Otsingusõna leidmisel toetusime eelteadmistele ja ootustele korpuse kohta, otsingu tulemuste seast said huvipakkuvad tüübid leitud lihtsa vaatluse teel ning tulemused määratletud regulaaravaldiste abil loodud reeglitega. Tekstijuppide dateerimiseks toetusime andmestikus olevatele metaandmetele, kus olid eraldatud reklaamtekstid tavalistest artiklitest.

Kõiki uuringu samme saaks teha ka suuremal määral lisavahenditele ja digimeetoditele toetudes. Näiteks otsingusõnade nimekirja saaks täiendada tehnikasõnastike abil või tulemuste põhjal nimekirja järk-järgult laiendades. Võimalik oleks kasutada tähendusvektoreid, et leida kontekstide alusel elektrile sarnaseid sõnu ja vigase digiteerimise tõttu esinevaid vorme (Hämäläinen & Hengchen 2019). Nii saaks esialgset märksõnaloendit andmepõhiselt laiendada. Leidude määratlemisel oleks võimalik toetuda automaatlahendustele, näiteks Levenshteini kaugustele sõnakujude võrdlemisel või teemamudelitele kontekstidevaheliste sarnasuste automaatsel leidmisel. Ka analüüsil saaks toetuda masinõppe meetoditele eri tüüpi kontekstide võrdlemisel, näiteks eristamaks, kui palju elektrimootoriga seotud artiklitest on seotud reklaamiga ja kui palju räägitakse sellest üldisemalt või kui palju elektripumba mainimistest räägivad õnnetusjuhtumitest ja sellest, kus need toimunud on. Lisavõimalusi avab siin üha arenev tehisintellekt, mis pakub mitmesuguseid võimalusi tekstide loovaks märgendamiseks (Karjus 2023).

Kokkuvõte

Juhtumiuuringu tulemusena saime ülevaate peamistest elektriga seotud nähtustest ja toodetest, millest räägiti eesti ajakirjanduses 1886–1940, tuginedes digiteeritud Postimehe väljaannetele. Saime näha, kuidas elektrist räägiti vaadeldud perioodi

jooksul üha rohkem ning kuidas esimeseks oluliseks teemaks oli valgustus. Alates 1910. aastast sai oluliseks teemaks elektrijaamad ja sellest ajast alates räägiti üha rohkem elektri infrastruktuurist. Nägime, kuidas elektrilisest rongist, kellast ja orelist räägiti juba enne 1890ndaid ja esimesed kaks said ühiskonna elu osaks, kolmas jõudis Eestisse alles 1930ndate teisel poolel. Nägime, kuidas elektrimootorid said väga sagedaseks 1920ndatel ja elektrilokid koos elektripliidiga 1930ndate teisel poolel. Enne nende massilist levikut mainiti neid tooteid vähem. Tuleb välja ka *elektriteatri* mainimiste tihe periood 1910. aasta ümber, pärast seda sai tavaliseks sõna *kino*. Triikraudu reklaamiti sageli 1920ndate lõpul, võib arvata, et pärast seda oli see toode ühiskonnas juba tuttav ja levinud.

Kirjandus

Caprettini, Bruno & Hans-Joachim Voth. 2020. Rage against the machines: Labor-saving technology and unrest in industrializing England. *American Economic Review: Insights* 2(3). 305–320. https://doi.org/10.1257/aeri.20190385.

Greenfield, Patricia M. 2013. The changing psychology of culture from 1800 through 2000. *Psychological Science* 24(9). 1722–1731. https://doi.org/10.1177/0956797613479387.

Hämäläinen, Mika & Simon Hengchen. 2019. From the Paft to the Fiiture: A fully automatic NMT and word embeddings method for OCR post-correction. *Proceedings - Natural Language Processing in a Deep Learning World*, 431–436. Incoma Ltd., Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-056-4_051.

Karjus, Andres. 2023. Machine-assisted mixed methods: Augmenting humanities and social sciences with artificial intelligence. https://doi.org/10.48550/ARXIV.2309.14379.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team & Nello Cristianini. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences* 114(4). E457–E465. https://doi.org/10.1073/pnas.1606380114.

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2008. *Introduction to information retrieval*. 1. tr. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809071.

Michel, Jean-Baptiste, Yan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, jt. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182. https://doi.org/10.1126/science.1199644.

Morin, Olivier & Alberto Acerbi. 2017. Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. *Cognition and Emotion* 31(8). 1663–1675. https://doi.org/10.1080/02699931.2016.1260528.

Nicholson, Bob. 2012. Counting culture; or, How to read Victorian newspapers from a distance. *Journal of Victorian Culture* 17(2). 238–246. https://doi.org/10.1080/13555502.2012.683331.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* 10(10). e0137041. https://doi.org/10.1371/journal.pone.0137041.

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Uiboaed, Kristel. 2018. Eesti keele stoppsõnad / Estonian stop words. https://doi.org/10.15155/RE-48.

Vinkers, Christiaan H., Joeri K. Tijdink & Willem M. Otte. 2015. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *BMJ* h6467. https://doi.org/10.1136/bmj.h6467.

Wevers, Melvin. 2022. Mining historical advertisements in digitised newspapers. Estelle Bunout, Maud Ehrmann & Frédéric Clavert (toim), *Digitised Newspapers – A New Eldorado for Historians?* (Studies in Digital History and Hermeneutics 3), 227–252. Berlin / Boston: Walter de Gruyter GmbH. https://doi.org/10.1515/9783110729214-011.

Konstruktsioonide produktiivsus: *lt*- ja *sti*-tuletusliite võrdlus veebikorpuse tekstide põhjal

Maarja-Liisa Pilvik

Lühikokkuvõte

Selles uurimuses uurime *lt*- ja *sti*-tuletuskonstruktsioonide produktiivsust, et hinnata kasutusandmete põhjal, kui palju erineb kahe malli potentsiaal aidata luua keelde uusi sõnu. Materjalina kasutame eesti keele ühendkorpuse 2021. aasta veebitekstide alamkorpuse VERT-vormingus faile. Failidest korjame uuritavaid sõnu sõnaliigimärgendi ning lemma lõpujärjendi põhjal, kasutades Pythoni skripte. Materjali analüüsiks ja visualiseerimiseks kasutame R-i. Produktiivsuse hindamiseks kasutame mõõdikuid, mis toetuvad tüübi-, sõne- ja üks kord esinevate tüüpide sagedustele ning eeldavad seetõttu ka korpusandmete puhastamist käsitsi või automaatsete meetoditega. Lisaks produktiivsuse mõõdikute arvutamisele võrdleme ka samast tüvest tuletatud *lt*-ja *sti*-lõpulisi määrsõnu ning uurime, kuidas tüve kasutussagedus tuletistes korreleerub tüve kasutussagedusega tuletusaluseks olevas omadussõnas.

1. Sissejuhatus

Siinses näidisuurimuses tegeleme kahe määrasõnatuletuse liite – *-lt* (nt *kaunilt*) ja *-sti* (nt *kaunisti*) – näitel sõnamoodustusmallide produktiivsuse uurimisega. **Produktiivsuse** mõistet on keeleteaduses defineeritud mitmel eri moel (vt nt Bauer 2001: 11–32; Barðdal 2008: 9–19), ent enamasti kasutatakse mõistet selleks, et eristada avatud ja suletud liikmesusega keelelisi konstruktsioone. Produktiivsed mallid võivad hõlpsalt laieneda ning nende abil on võimalik luua keelde vajadusel uusi keelelisi üksusi. Samas võidakse produktiivseks pidada ka selliseid malle, millega on moodustatud küll suhteliselt vähe erinevaid, kuid samas semantiliselt väga sarnaste omadustega vorme (Barðdal 2008: 91). Selles uurimuses tegeleme põhiliselt produktiivsuse laiendava omadusega ning võrdleme kahe tuletuskonstruktsiooni kasutusest ilmnevat potentsiaali luua keelde uusi sõnu. Ehkki sagedamini on

produktiivsusele viidatud just tuletusmorfoloogia puhul, on seda mingite abstraktsete konstruktsioonimallide avatuse hindamisel kasutatud ka liitsõnamoodustuse, muutemorfoloogia ja isegi süntaktiliste konstruktsioonide analüüsimisel. Seepärast eelistame siin rääkida just *konstruktsioonide*, mitte *protsesside*, *reeglite*, *liidete*, *vormide* või *sõnade* produktiivsusest (vt ka Booij 2018).

Nii *lt*- kui ka *sti*-liidet on peetud produktiivseteks tuletusliideteks, mis võimaldavad moodustada omadussõnadest reeglipäraselt määrsõnu, mis vastaksid küsimuse *Milline?* asemel küsimusele *Kuidas?*. Liidete abil moodustatakse seega põhiliselt viisimäärsõnu (nt *rõõmsalt* ja *rõõmsasti*). „Eesti keele käsiraamatu" (Erelt, Erelt & Ross 2020: 372) põhjal liituvad mõlemad liited üldjuhul omadussõna omastava käände tüvele (näide 1); *-lt* võib liituda lisaks ka kõikide kesksõnade tüvedele (näited 2–5), *-sti* aga ainult teatud *tav*-kesksõnadele, millest tuletatud määrsõnad väljendavad tõenäosust ja ebakindlust (näide 5). Samuti moodustatakse *sti*-liitelisi määrsõnu harva komplekssetest omadussõnadest (Kasik 2015: 384). Eeskätt selliste vormimoodustuslike piirangute, aga ilmselt ka vormiökonoomsuse taotluse tõttu on *sti*-liidet peetud tänapäeva keeles vähem sagedaseks ja vähem produktiivseks liiteks.

(1) *hõlbus* → *hõlpsa* → *hõlpsalt*, *hõlpsasti*
(2) *armunud* → *armunult*, ?*armunusti*
(3) *võidetud* → *võidetult*, ?*võidetusti*
(4) *kõlav* → *kõlavalt*, ?*kõlavasti*
(5) *kuuldav* → *kuuldavalt*, *kuuldavasti*

Lisaks vormilistele moodustuspiirangutele võib mingite tuletuskonstruktsioonide produktiivsust piirata (sealjuures tihtipeale korraga) veel terve hulk muid tegureid, näiteks

1) **semantiline reeglipärasus**: kas tuletuskonstruktsioon lisab tuletusalustele sõnadele regulaarselt sarnase lisatähenduse või mitte. Kui tuletuskonstruktsiooniga saab moodustada palju eri tähendusklassidesse kuuluvaid sõnu, võib see konstruktsiooni produktiivsust aja jooksul vähendada, kuna vormi ja tähenduse seos keelekasutaja mälus nii selgelt ei kinnistu;
2) **semantilised piirangud**: kas tuletuskonstruktsiooni saab kasutada ainult kindlat tüüpi tähendusklassi kuuluvate sõnadega. Semantika piirab mõnevõrra ka selles näidisuurimuses vaadeldavaid tuletusliiteid: omadussõnadest, mis ei seostu sündmusega (nt *kilone*), tuletatakse harvem ka määrsõnu (*kiloselt*);
3) **võistlevate sõnavormide või konstruktsioonide hulk**: kas mingi tuletis on asendatav teiste sõnade või konstruktsioonidega. Mida vähem on konstruktsioonil konkurente, seda tõenäolisemalt võib teda vaja minna uute keelendite moodustamiseks;

4) konstruktsiooni väljendatud mõistete või funktsioonide **pragmaatiline vajalikkus**: kui konstruktsiooni läheb vaja harva või väga spetsiifilistes keelekasutuse situatsioonides, jääb konstruktsiooni produktiivsus madalamaks.

Viimasest kahest punktist rääkides eristatakse vahel ka moodustus- ja kasutusproduktiivsust: mingi konstruktsioon võib olla moodustuslikust küljest potentsiaalselt väga produktiivne, ent see potentsiaal ei realiseeru keele kasutuses pragmaatilistest teguritest tingituna. Kasutuspõhise lähenemise puhul aga ei saa neid kaht väga rangelt lahus hoida: keeleliste struktuuride paindlikkus kujuneb ja areneb keele kasutuses, mitte sellest sõltumatult, ja seega võib ka pragmaatilisi vajadusi pidada mingi konstruktsiooni produktiivsust pärssivateks või soodustavateks teguriteks.

Määrsõnatuletuse kasutusmustreid kirjakeele korpuses võib mõjutada ka keelekorralduslik vaade ühe või teise liitega määrsõnade sobilikkusele. Näiteks on eesti keelehooldeallikates taunitud „tarbetute" *lt*-lõpuliste määrsõnade moodustamist olemasolevate aja-, koha-, viisi- jm määruste kasutamise asemel (nt *kohe* → *kohene* → *koheselt* pro *kohe*, *vastupidi* → *vastupidine* → *vastupidiselt* pro *vastupidi*, *järk-järgult* → *järkjärguline* → *järkjärguliselt* pro *järk-järgult*, vt Mäearu 2000). Johannes Aavik (1936: 3) on soovitanud ökonoomsuse põhimõttel liita pikematele sõnadele *lt*-liide (eriti *ne*-, *line*-, *lik*-, *ik*-lõpulistele omadussõnadele, samuti *v*-, *nud*- ja *tud*-kesksõnadele), lühematele *sti*-liide. Kuna vanemaks peetava *sti*-liite rolli oli *lt*-liide juba 1930ndatel jõuliselt üle võtmas, soovitas Aavik aga võimalusel eelistada soome keele eeskujul justnimelt *sti*-liidet (nt *kiiresti, julgesti, selgesti, õigesti*, mitte *kiirelt, julgelt, selgelt, õigelt*), et „teda päästa". Annika Küngas (2013) on näidanud, et vanas kirjakeeles oli *sti*- ja *lt*-liidete kasutuse suhe tõepoolest tänapäeva keelele vastupidine ning *-sti* määrsõnades oluliselt sagedasem kui *-lt*.

On näidatud, et produktiivsus on seotud selgelt sageduse ja keele töötlemisega ajus. Tuletusliidetega, mis on produktiivsed, esineb keelekorpustes oluliselt vähem väga kõrge sagedusega tuletisi ja palju madala sagedusega tuletisi. See omakorda võib tähendada, et selliste sõnade töötlemisel aktiveeritakse konkreetse leksikaalse sõnavormi (nt *lihtsasti*) tähenduse asemel mälus pigem konstruktsiooni abstraktne tähendus (nt $[[x]_{ADJ}$ *-sti*$]_{ADV} \leftrightarrow [x_{SEM}$ omadusega sündmusviis$]_{SEM}$), mis aitab sõna ja selle tähendust ära tunda. Väga sagedaste tuletiste puhul aga on tõenäolisem vormide töötlemine tervikliku, iseseisva funktsionaalse üksusena ning selliste tuletistega võib kaasneda ka nende tähenduse või kasutuse muutumine. Nii on juhtunud näiteks adverbidega *tegelikult*, *ilmselt*, *põhimõtteliselt*, *kindlasti* ja *vähemasti*, mida lisaks viisimäärusele kasutatakse ka modaalpartiklite või diskursusemarkeritena (Küngas 2013; Valdmets 2013)

Kõikide nende erinevate aspektide põhjal võib eeldada, et produktiivsus ei ole mitte kategooriline tunnus (mingi konstruktsioon kas on produktiivne või mitte), vaid et tegemist on pigem skalaarse omadusega (mõni konstruktsioon on produktiivsem kui mõni teine), mis võib avalduda keelekasutuses eri nurkade alt, muu hulgas ka kasutussageduse kaudu. Sageli suudab küll keele kõneleja ka

intuitiivselt öelda, millised tuletuskonstruktsioonid on produktiivsemad ja millised pigem mitte, ent selline tunnetuspõhine hinnang ei võimalda vastata küsimusele, *kui palju* üks konstruktsioon teisest produktiivsem on. Seda aitavad hinnata kõiksugu nimetamiskatsed (nt *loetle etteantud aja jooksul võimalikult palju* lt-*liitega sõnu!*) ja mõistmiskatsed (nt *otsusta, kas tegemist on eesti keeles olemas oleva sõnaga või mitte!*), aga ka suured tekstikorpused. Produktiivsuse empiiriliseks hindamiseks on välja töötatud palju sageduspõhiseid mõõdikuid, mis võimaldavad keelekorpuse andmeid kasutades hinnata konstruktsioonide produktiivsuse *väljundit* erinevates registrites, tekstitüüpides, eri ajaperioodidel jne. Eesti keele sõnamoodustust ja kitsamalt tuletussüsteemi on kvantitatiivsete meetoditega vähe uuritud. Selle põhjuseks võib olla ühelt poolt olnud materjalide vähene kättesaadavus ja kvantitatiivsete analüüsimeetodite keerukaks või ebaotstarbekaks pidamine. Teiselt poolt raskendab kvantitatiivset analüüsi kindlasti see, et suur osa eesti keele tuletussüsteemist on väga kompleksne, tuues sageli kaasa muutusi nii sõnatüvedes kui ka tuletusliidete vormis. Samuti võivad tuletusliited olla mitmetähenduslikud ning mõjutada erinevaid sõnaliike. Näiteks *us*-liitega võib tuletada nimisõnu nii tegusõnadest (*kaklema* → *kaklus*) kui ka omadussõnadest (*raske* → *raskus*). Siiski leidub ka eesti keele kohta uurimusi, mis hindavad kvantitatiivselt muu hulgas ka morfoloogilist produktiivsust (vt nt Kerge 2002; Kerge 2003), samuti on siinses näidisuurimuses rakendatud mõõdikuid kasutatud eesti keele süntaktiliste konstruktsioonide produktiivsuse hindamisel (Muischnek & Sahkai 2010).

2. Produktiivsuse mõõdikud

Kasutame selles uurimuses intuitiivselt võrdlemisi lihtsasti tõlgendatavaid mõõdikuid, mille kasutuselevõtt keeleteaduses seondub eeskätt Harald Baayeni ja tema kolleegide tööga hollandi ja inglise keele tuletusmorfoloogia uurimisel 1990ndatel (Baayen & Lieber 1991; Baayen 1992; Baayen 1994; Baayen 1996; Baayen & Renouf 1996; Baayen 2009), ehkki sarnaseid arvutuslikke hinnanguid leiab ka varasematest töödest. Kolm põhilist mõõdikut, mida on produktiivsuse korpuspõhisel uurimisel kasutatud, on nn realiseerunud produktiivsus (ingl *realized productivity*, ka *extent of use*), potentsiaalne produktiivsus (ingl *potential productivity*, ka *category-conditioned productivity* või *expansion rate*) ja laiendav produktiivsus (ingl *expanding productivity,* ka *hapax-conditioned productivity*). Lisaks on kasulik hinnata nn tüübi- ja sõnesageduse suhet (ingl *type-token ratio* ehk TTR, vt ptk 5.2.4.3 „Sõnavara mitmekesisus"), mis väljendab konstruktsiooniga väljendatud sõnavara suurust. Produktiivsuse hindamise mõõdikuid on õigupoolest veel terve hulk (vt nt Zeldes 2012; Heede & Lauwers 2023).

Realiseerunud produktiivsus $\boldsymbol{V(C, N)}$ on tegelikult lihtsalt ühe konstruktsiooni (***C***) tüüpide (lemmade või konkreetsete sõnavormide) arv korpuses, milles on mingi kindel arv (***N***) sõnu. Kui konstruktsiooni esindab korpuses palju

erinevaid tüüpe, võib eeldada, et see konstruktsioon on 1) struktuuriliselt avatud, 2) oluliste semantiliste piiranguteta, 3) pragmaatiliselt kasulik. Tüüpide arvu puuduseks on peetud aga seda, et see kõneleb produktiivsusest ainult minevikus ega aita alati ennustada, kui palju uusi sõnu selle konstruktsiooniga võiks edaspidi juurde tekkida. Näiteks inglise keele deverbaalne tuletusliide *-ment* (nt *achievement*, *movement*) on tüüpide arvu järgi otsustades väga produktiivne, ent uute sõnade moodustamiseks seda enam ei kasutata (Baayen & Lieber 1991; Bauer 2001).

Realiseerunud produktiivsuse mõõdikut täiendab **tüübi- ja sõnesageduse suhe** $\boldsymbol{V}(\boldsymbol{C}, \boldsymbol{N})/\boldsymbol{N}(\boldsymbol{C}, \boldsymbol{N})$, mis suhestab konstruktsiooni tüüpide arvu ($\boldsymbol{V}(\boldsymbol{C})$) konstruktsiooni kõikide esinemiste arvuga ($\boldsymbol{N}(\boldsymbol{C})$) korpuses, milles on $\boldsymbol{N}$ sõna. Mõne liitega sõnu võib esineda korpuses kokku küll väga palju, ent lähemal vaatlusel võib selguda, et suurema osa esinemiskordadest katavad ainult paar üksikut vormi. Teisel puhul võib konstruktsiooni esineda oluliselt väiksem arv kordi, ent esinemiskorrad jaotuvad ühtlasemalt eri tüüpide vahel. Mida suurem on tüübi-sõne suhe, seda rohkem *erinevaid* sõnu mingi konstruktsiooniga kasutatakse ning seda produktiivsemaks võib konstruktsiooni pidada. Kui tüübi-sõne suhe on 1, on iga konstruktsiooni esinemiskord korpuses eri tüübiga. Sarnaselt realiseerunud produktiivsusega on mõõdiku miinuseks see, et see peegeldab põhiliselt mineviku produktiivsust ega ole otseselt tõlgendatav tõenäosusena, millega konstruktsiooni abil keelde uusi sõnu võiks toota. Samuti on mõlemad mõõdikud tundlikud sõnede arvu suhtes, kaldudes väikese sõnede (*Nc*) hulga puhul produktiivsust või sõnavara ulatust ülehindama.

Potentsiaalne produktiivsus $\boldsymbol{V}(\mathbf{1}, \boldsymbol{C}, \boldsymbol{N})/\boldsymbol{N}(\boldsymbol{C}, \boldsymbol{N})$ on konstruktsiooni ainult üks kord esinevate tüüpide ($\boldsymbol{V}(\mathbf{1}, \boldsymbol{C})$) osakaal kõigist konstruktsiooni esinemiskordadest ($\boldsymbol{N}(\boldsymbol{C})$) korpuses, milles on $\boldsymbol{N}$ sõna. Ainult ühe korra esinevat tüüpi nimetatakse kreeka keelest tuletatud nimega *hapax legomenon* ('üks kord öeldud'), mida siinses õpikus kutsume eestikeelse terminiga *ainuk* (vt ka õpiku ptk 5.2.4 „Sõnavara analüüs"). Potentsiaalne produktiivsus on realiseerunud produktiivsuse kõrval üks enim kasutatud produktiivsuse empiirilisi mõõdikuid. Kui morfoloogiline konstruktsioon on oluliste semantiliste ja moodustuspiiranguteta ja samas piisavalt regulaarse tähendusega, saab selle abil hõlpsalt luua sõnu, mida varem pole kohatud. Piisavalt suures korpuses võib selliste moodustusprotsesside väljundiks pidada sõnu, mis esinevad seal ainult üks kord (ehkki põhimõtteliselt ka enam kui ühe korra, ent siiski harva esinevaid sõnu). Mida suurem on selliste sõnade osakaal kõikidest konstruktsiooni kasutuskordadest, seda tõenäolisemalt saab konstruktsiooni abil sõnavara laiendada ka edaspidi, väljaspool olemasolevat korpust. Potentsiaalset produktiivsust saab seega tõlgendada ka kui *tõenäosust*, millega järgmine korpusesse lisatav sama konstruktsiooni esindav sõna on selline sõna, mida me pole varem korpuses kohanud.

Nii realiseerunud produktiivsuse, tüübi-sõne suhte ja potentsiaalse produktiivsuse kriitikana on esitatud seda, et need kipuvad alahindama sagedamini esinevate

konstruktsioonide produktiivsust, sest mida rohkem sõnu oleme juba näinud, seda väiksemaks jääb tõenäosus, et moodustatakse vorm, mida varem pole ette tulnud. Selle aluseks on asjaolu, et keelekasutajate pragmaatilised vajadused võivad ka väga produktiivsete konstruktsioonide puhul ühel hetkel ammenduda ja vajalikud mõisted on juba olemasoleva sõnavaraga kaetud.

Neljas mõõdik, **laiendav produktiivsus** ***V*(1, *C*, *N*)/*V*(1, *N*)** väljendab konstruktsiooni ainult üks kord esinevate tüüpide (***V*(1, *C*)**) osakaalu kõigist korpuses üks kord esinevatest tüüpidest (***V*(1)**) korpuses, milles on ***N*** sõna. Laiendav produktiivsus kombineerib omavahel potentsiaali laieneda ja pragmaatilise kasulikkuse: mida suurema osa ainult ühe korra esinevatest sõnadest mingi konstruktsioon hõlmab, seda suurem on tema potentsiaal sõnavara laiendada ja seda mitte ainult tänu oma struktuurilisele ja/või semantilisele avatusele, vaid ka sellepärast, et konstruktsiooniga loodavaid mõisteid läheb keeles sagedamini vaja. Laiendavat produktiivsust saab niisiis tõlgendada kui *tõenäosust*, millega järgmine korpusesse lisatav varem korpuses mitte kohatud sõna esindab just uuritavat konstruktsiooni (nt on *lt*-lõpuline määrsõna). Ühes ja samas korpuses, kus kõikide üks kord esinevate sõnade koguarv eri sufiksite võrdlemisel nimetajas ei muutu, võib laiendavat produktiivsust hinnata põhimõtteliselt ka lihtsalt ainukite arvu põhjal. Sel juhul kaob vajadus kõikide sadade miljonite korpuse sõnade esinemissagedusi kokku lugeda, ent samas ei saa laiendavat produktiivsust ka enam tõenäosusena tõlgendada.

Suurte korpuste puhul on pakutud, et ainukite ehk üks kord esinevate vormide kokkulugemisele lisaks või koguni selle asemel võiks konstruktsiooni produktiivsuse hindamisel kaasata ka kaks ja kolm korda esinevad sõnad (vastavalt *dis legomenon* ja *tris legomenon*). Seda seetõttu, et üks kord esinevate sõnade rühm võib kõige tõenäolisemalt sisaldada trükivigu ning nende kasutamiseks peaks sõnade nimekirju käsitsi hoolikalt puhastama.

Mõistagi ei oleks nende produktiivsuse mõõdikutega kuigi mõistlik võrrelda näiteks eri sõnaklasside sõnamoodustust, kuna nimisõnu kasutatakse keeles rohkem kui tegusõnu, tegusõnu omakorda rohkem kui omadussõnu jne. Seega võib keele üldisest struktuurist tulenev pragmaatiline nõudlus produktiivsuse kvantitatiivset väljundit oluliselt moonutada. Küll aga on näidatud, et sarnaseid funktsioone täitvate konstruktsioonide võrdlemisel annavad kirjeldatud mõõdikud intuitiivselt küllalt usutavaid tulemusi.

3. *lt*- ja *sti*-tuletusmallide produktiivsus korpuses

Kasutame näidisuurimuses eesti keele ühendkorpuse 2021. aasta versiooni (vt Koppel & Kallas 2022). Selle korpusefaile saab alla laadida META-SHARE'i kaudu[1]. Failid on üles laaditud alamkorpuste kaupa, kusjuures suuremad alamkorpused võivad olla jagatud veel omakorda eri osadeks. Üles laaditud korpuse suurus kokku on hinnanguliselt 2,4 miljardit sõna. Siin kasutame ainult ÜK 2021. aasta veebikorpusi (Web2021), mille suuruseks on umbes 741 miljonit märgendatud sõna. Korpuse failid on VERT-vormingus, mis tähendab seda, et iga sõne paikneb eraldi real („vertikaalselt") ning selle järel on muu seda sõnet iseloomustav info (nt sõnaliik, grammatilised kategooriad, tunnused, lõpud). Eraldi ridadel on märgitud ka dokumentide, lõikude, lausete ja osalausete algused ja lõpud (joonis 1).

Morfoloogilise produktiivsuse uurimiseks on korpusefailide kasutamine kasulik, kuna

1) produktiivsuse mõõdikud on tõlgendatavad kindla suurusega korpuse kontekstis, ent uurimisülesandest lähtuvalt võib loendatavate üksuste hulka olla vajalik muuta (nt mitte lugeda sõnedeks kirjavahemärke, arvestada ainult kindlaid sõnaliike, käsitleda liitsõnu eraldi);
2) see võimaldab loendada jooksvalt ka tekstisõnu ja joonistada selle põhjal nn produktiivsuse kõveraid (vt allpool jooniseid 3 ja 4), mis näitavad, kuidas mingi konstruktsiooni produktiivsuse kvantitatiivne väljund muutub, kui mõõdame seda eri suurusega korpuse osade põhjal;
3) see võimaldab materjali vajadusel kontrollitult juhuslikustada. Loomuliku keele tekstides ei esine sõnad juhuslikus järjekorras, vaid iga kasutatud sõna tingib mingil määral selle, mis sõna või sõnaklass talle võib järgneda. Nii enamik produktiivsuse definitsioonidest kui ka siinses uurimuses kasutatud produktiivsuse mõõdikud aga eeldavad justnimelt teatud juhuslikkust: 1) kui peaksime lisama korpusesse mingi tuletuskonstruktsiooniga ühe sõna, siis kas see oleks täiesti uus sõna või mingi sõna, mida oleme korpuses juba näinud?; 2) kui peaksime lisama korpusesse lihtsalt ühe uue sõna, mida me pole korpuses varem näinud, siis kui tõenäoliselt oleks see sõna moodustatud just konkreetse tuletuskonstruktsiooniga? Kumbki küsimus ei küsi lisainfot selle kohta, missuguseid muid sõnu oleme korpuses hiljuti näinud või millisesse konteksti uut sõna otsime;
4) see võimaldab valida korpusest uurimiseks täpselt vajaliku suurusega tüki. Eri suurusega (alam)korpusi või žanre võiks võrrelda samas suurusjärgus sõnade hulga põhjal, kuna vastasel juhul võidakse mingi konstruktsiooni produktiivsust väiksemas korpuses ülehinnata;

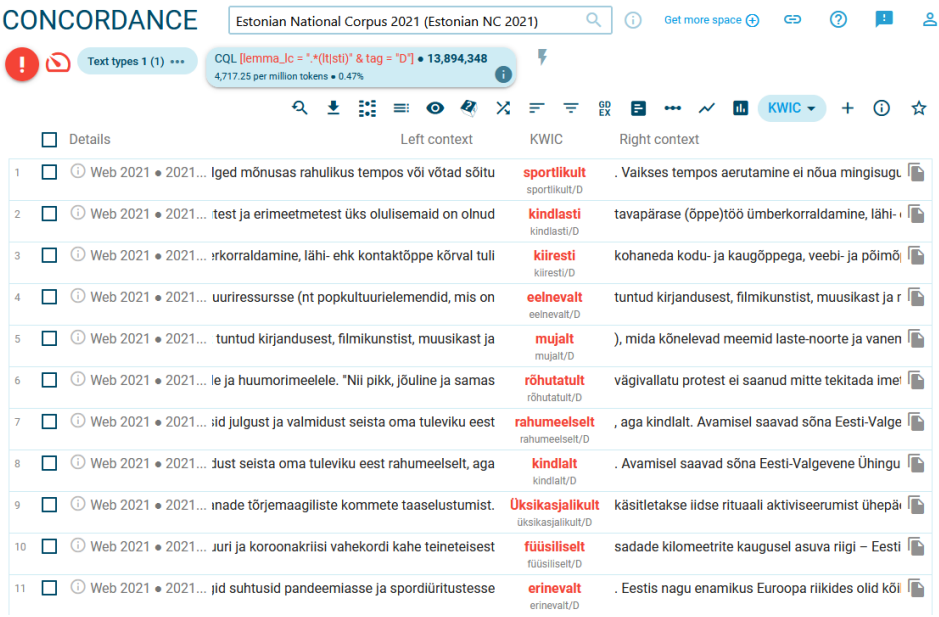
[1] https://metashare.ut.ee/repository/browse/estonian-national-corpus-2021-vert/4547c7bea0d411eebb4773db10791bcfd961b8c70b544966800142b04f957a86/

```
<doc id="1245780" src="Web 2021" topic="travel & tourism" topic_src="site list" title="Arles – Lõuna-Prantsusmaa üks turismipärlitest - ReisiGuru.ee" length="1k-5k" crawl_date="2021-06-04 02:18" lang_diff="0.60" ip="217.146.69.5" domain_distance="0" url="https://reisi.guru/lugu/504" enc_meta="utf-8" enc_chared="utf-8" lang="estonian" lang_scores="estonian: 1994.52, finnish: 679.82, english: 489.18, arabic: 0.00, chinese: 0.00, danish: 465.52, french: 499.52, german: 448.77, hindi: 0.00, italian: 401.16, japanese: 0.00, korean: 0.00, polish: 461.86, portuguese: 395.60, russian: 0.00, spanish: 412.39, swedish: 472.50">
<p langdiff="0.99">
<s>
2014.03.07  N.? 20140307-n  ? 20140307  20140307  0   card_<?>_digit          1 2 nummod  08:44 08:44 N ? root
08:44 N.? 08:44-n ? 08:44 08:44 0   card_<?>_digit          2 0 root
- Z --z   - -       Dsh       3 4 punct ReisiGuru.ee  ReisiGuru.ee  H sg_n  conj
ReisiGuru.ee  H.sg.n  ReisiGuru.ee-h  sg_n  ReisiGuru.ee  ReisiGuru.ee  0   prop_sg_nom         4 2 conj  08:44 08:44 N ? root
</s>
</p>
<p langdiff="0.69">
<s>
See P.sg.n  see-p sg_n  see see 0   sg_nom    dem     1 6 nsubj:cop kliima  kliima  S sg_n  root
on  V.b olema-v b ole ole 0   mod_indic_pres_ps3_sg_ps_af     fin Intr  2 6 cop kliima  kliima  S sg_n  root
aastaringselt D aastaringselt-d   aasta ringselt  aasta_ringselt  0            3 4 advmod  nauditav  nauditav  A sg_n  amod
nauditav  A.sg.n  nauditav-a  sg_n  nauditav  nauditav  0   pos_sg_nom          4 6 amod  kliima  kliima  S sg_n  root
vahemereline  A.sg.n  vahemereline-a  sg_n  vahe mereline vahe_mere=line  0   pos_sg_nom          5 6 amod  kliima  kliima  S sg_n  root
kliima  S.sg.n  kliima-s  sg_n  kliima  kliima  0   com_sg_nom          6 0 root
<g/>
, Z ,-z   , ,       Com       7 9 punct rannad  rand  S pl_n  conj
kaunid  A.pl.n  kaunis-a  pl_n  kaunis  kaunis  d   pos_pl_nom          8 9 amod  rannad  rand  S pl_n  conj
rannad  S.pl.n  rand-s  pl_n  rand  rand  d   com_pl_nom          9 6 conj  kliima  kliima  S sg_n  root
<g/>
, Z ,-z   , ,       Com       10  15  punct säilinud  säilima V nud conj
palju D palju-d   palju palju 0            11  12  advmod  ajaloolist  ajalooline  A sg_p  nsubj
ajaloolist  A.sg.p  ajalooline-a  sg_p  aja looline aja_looline t   pos_sg_part         12  15  nsubj säilinud  säilima V nud conj
imetlusväärselt D imetlusväärselt-d   imetlus väärselt  imetlus_väärse=lt 0            13  14  advmod  hästi hästi D   advmod
hästi D hästi-d   hästi hästi 0            14  15  advmod  säilinud  säilima V nud conj
säilinud  V.nud säilima-v nud säili säili nud   mod_indic_impf_ps_neg     fin Intr  15  6 conj  kliima  kliima  S sg_n  root
arhitektuuri  S.sg.g  arhitektuur-s sg_g  arhitektuur arhitektuur 0   com_sg_gen          16  15  obl säilinud  säilima V nud conj
ja  J ja-j    ja  ja  0   sub_crd        17  18  cc  huvitav huvitav A sg_n  conj
huvitav A.sg.n  huvitav-a sg_n  huvitav huvitav 0   pos_sg_nom          18  16  conj  arhitektuuri  arhitektuur S sg_g  obl
<g/>
..... Z .....-z   ..... .....       Ell       19  6 punct kliima  kliima  S sg_n  root
</s>
</p>
```

Joonis 1. Näide eesti keele ühendkorpuse VERT-vormingus failist

5) see võimaldab kaasata kõiki mingi konstruktsiooni esinemisjuhte, paljudel korpuste kasutajaliidestel (sh Sketch Engine'il) on aga tihtipeale andmete kuvamis- ja allalaadimispiirangud (vt nt J. Padriku näidisuurimust kollostruktuurilisest analüüsist).

Enne korpusefailidest otsimiseks skripti kirjutamist oleks aga kasulik esmalt nt Sketch Engine'i kaudu vaadata, kui palju vasteid võiksime leida ning millist märgendust failides kohata (joonis 2).



Joonis 2. Eesti keele ühendkorpuse Web 2021 alamkorpuses esinevad *lt*- ja *sti*-lõpulised määrsõnad Sketch Engine'i konkordantsiotsingus

Nagu öeldud, kasutame selles näidisuurimuses kättesaadavaid korpusefaile. Kõike, mis seondub suurte korpuste töötlemisega ja keele automaatanalüüsi vahendite kasutamisega, teeme siin Pythoni koodiga (versioon 3.9). Kõike, mis puudutab kogutud andmete analüüsimist ja visualiseerimist, aga R-is (versioon 4.3.2, R Core Team 2023). Nõnda kasutame ära kummagi programmeerimiskeele tugevusi: Python on suurte andmete töötlemisel kiirem, R analüüsimisel ja visualiseerimisel paindlikum.

3.1. Sõnade kogumine korpusest

Esmalt **võtame korpusest välja iga tuletusliite kohta kõikide seda tuletusliidet kandvate sõnavormide lemmad** ehk algvormid. Määrsõnade kui muutumatute sõnade sõnavorm ja lemma on ühesugused (nt *raskelt* : *raskelt*), käänd- ja pöördsõnade puhul aga erinevad (nt *paberite* : *paber*, *tulime* : *tulema*). Sealjuures salvestame korpuse materjali järjest läbi käies, mitu muud sõna peab korpuses vahepeal esinema, et tuletuskonstruktsiooni uuesti vaja läheks. See tähendab, et iga korpuses vastutulev sõna saab sisuliselt järjekorranumbri. Selleks, et päringutulemuste hulka satuks võimalikult vähe ebasoovitavaid tulemusi, saame päringus täpsustada ka sõnaliiki ja otsida ainult määrsõnadeks märgendatud sõnu (nt *töö läks raskelt*). Nõnda jätame potentsiaalseid *lt*-tuletisi uurides andmestikust välja sõnad, mis on tegelikult pärisnimed (nt *Anvelt*) ja muu juhusliku müra. Kui otsiksime lemma asemel sõnavormi põhjal, võimaldaks sõnaliigi täpsustamine jätta välja ka käändsõnade alaltütleva käände vormid (nt *tulin ära liiga raskelt* töölt). Päris puhast andmestikku siiski kindlasti nii kätte ei saa, kuna ühendkorpuse tekstid on märgendatud automaatselt, mistõttu vahel võib nt vorm *raskelt* olla analüüsitud ekslikult määrsõnaks, vahel ekslikult omadussõna alaltütleva käände vormiks. Kuna konteksti läbivaatamine läheks sellise uurimistöö puhul väga töömahukaks, peame siin leppima teatud vormihomonüümiast tingitud eksimisvõimalusega.

Salvestame skriptiga kogutava info ka iga korpuse žanri kohta eraldi (blogid, foorumipostitused, ajakirjandus, e-kaubandus, Vikipeedia, teadustekstid, ilukirjandus), et vajadusel oleks võimalik tulemusi ka registrite lõikes vaadelda, ehkki selles uurimuses me seda ei tee. Samad konstruktsioonid võivad näidata eri registrites erinevat (kasutus)produktiivsust, kuna vajadus nendega loodavate mõistete järele ei pruugi igas registris olla samasugune, kasutusel võivad olla teised sarnase tähendusega konstruktsioonid jne (vt Pilvik 2021).

Kasutame lemmade otsimiseks Pythoni skripti *main.py*, mis käib läbi kõik kokku pakitud korpusefailid, ning selle sees omakorda skripti *kogu_lemmasid.py*, mis otsib igast korpusefailist üles vajaliku info ning kirjutab leitud lemmad, nende järjekorranumbrid ning kogu (alam)korpuse sõnade arvu eraldi failidesse.

Saame ühtekokku korpusest 11 116 719 *lt*-lõpulise ja 2 777 626 *sti*-lõpulise määrsõna kasutusjuhtu. Kui vaatame esimest 10 korpusest kogutud *lt*- ja *sti*-tuletist (tabel 1), näeme esiteks, et *lt*-tuletisi tuleb korpuses ette oluliselt sagedamini (esimesed 10 vormi leiame korpuse esimese 1900 sõna hulgast), ning teiseks, et *lt*-tuletiste hulgas on rohkem erinevaid sõnavorme, samas kui *sti*-tuletiste hulgas kipuvad korduma samad vormid (nt *kindlasti*, *kiiresti*).

Tabel 1. Esimesed 10 */t-* ja *sti*-lõpulist määrsõna korpuses koos nende esinemise järjekorranumbriga; alakriipsud märgivad liitsõna-, võrdusmärgid tuletuspiire

***lt*-lõpuline sõne**	***lt*-lõpulise sõne järjekorranumber**	***sti*-lõpuline sõne**	***sti*-lõpulise sõne järjekorranumber**
sportliku=lt	39	*kindlasti*	666
eelnevalt	745	*kiiresti*	676
mujalt	751	*kindlasti*	2382
rõhutatult	929	*kindlasti*	2497
rahu_meelselt	1065	*kindlasti*	3270
kindlalt	1067	*tõesti*	4024
üksik_asjalikult	1596	*enamasti*	4847
füüsiliselt	1660	*kiiresti*	5051
erinevalt	1676	*uuesti*	5382
julgelt	1867	*hästi*	5471

3.2. Andmestike puhastamine

Teise sammuna peaksime kogutud **andmestikke puhastama**. Puhastamine on vaadeldavate produktiivsuse mõõdikutega vajalik samm eeskätt sellepärast, et mõõdikud toetuvad suuresti ainult üks kord esinevatele sõnavormidele (nn ainukitele). Üks kord esinevad sõnavormid on aga just see sõnade rühm, millest suur hulk tekib trüki- ja kirjavigadest, mistõttu hakkaksid ebasoovitavad vormid mõõdikuid oluliselt mõjutama. Lisaks trüki- ja kirjavigadele peaksime puhastamise käigus välja jätma ka määrsõnad, mis ei ole omadussõnadest tuletatud (nt *mujalt*, *sealt*, *nimelt*, *kõigepealt*, *varsti*, *tõesti*) ning kõiksugu onomatopoeetilised sõnad (nt *klopsti*, *vulpsti*). Selleks, et olla kindlad, et analüüsime justnimelt sõnatuletuse produktiivsust ja mitte liitsõnamoodustuse oma, peaksime õigupoolest vaatlema ainult liitsõnade järelosiseid. Nõnda esindaksid näiteks *rõõmsalt* ja *mega_rõõmsalt* üht ja sama tüüpi. Tuleb aga arvestada, et samamoodi saab vormist *rahu_meelselt* vorm *meelselt*, vormist *era_kordselt* vorm *kordselt*, vormist *iga_päevaselt* vorm *päevaselt* jne, mispuhul liitsõna esiosiste kustutamine kaotab ära ka olulise osa sõna tähendusest ning selle loomise ja kasutamise pragmaatilisest motivatsioonist.

Kõige parema tulemuse annab muidugi jällegi andmestike **käsitsi** ülevaatamine ja parandamine, olgugi et suurte andmestike puhul jääb andmestikku paratamatult vigu ja ebaühtlust. Lisaks on käsitsi parandamisel alati oht materjali oma subjektiivsetest hinnangutest ja teadmistest tulenevalt vääriti tõlgendada, sest kontrollija ei tea lihtsalt kõiki keeles kasutusel olevaid sõnu (näiteks kas eesti keeles on olemas

sõnad *jõngalt* või *kasuistiliselt*?) ning ka keelekasutajate loovust ilma kontekstita hinnata on keeruline. Ka praegusel juhul ei saa andmestiku puhastamisel automaatselt välja visata sõnu, mis ei tundu viisimäärsõnana kuigi harilikud (nt *intervjueeritavalt*, *jämeduselt*), sest lausekontekstita ei ole võimalik kindlalt öelda, et neid ei ole just selles funktsioonis kasutatud.

Üldiselt on käsitsi puhastamist mõistlikum teha tüüpide, mitte sõnede kaupa: koostada kõikidest kogutud sõnavormidest unikaalsete vormide loend ning märkida iga vormi kohta selle parandatud vorm (kui on vaja parandada) või info selle kohta, kui mõne vormi peaks loendist välja jätma (tabel 2). Pärast tüüpide ülevaatamist saab kõik ühesugused sõned sama tüübi põhjal korraga ära parandada.

Tabel 2. Näide tuletiste käsitsi parandamise tabelist

***lt*-lõpuline tüüp**	***lt*-lõpulise tüübi viimane osis**	***lt*-lõpulise tüübi parandatud viimane osis**	**Kommentaar**
avaöikult	*avaöikult*	*avalikult*	
isukalt	*isukalt*	*isukalt*	
konjugeeru=nult	*konjugeerunult*	*konjugeerunult*	
aja_mahulise=lt	*mahuliselt*	*mahuliselt*	
muusika_meelse=lt	*meelselt*	*meelselt*	
Armeenia-mängult	*mängult*	*mängult*	välja
sama_pingsa=lt	*pingsalt*	*pingsalt*	
tegemilt	*tegemilt*	*tegemilt*	välja
vahi_torni=likult	*tornilikult*	*tornilikult*	
kramp_tõsise=lt	*tõsiselt*	*tõsiselt*	

Nii suure korpuse puhul on käsitsi kontrollimine aga tülikas ja väga aeganõudev, kuna peame läbi vaatama 30 639 unikaalset *lt*-lõpulist lemmat ja 380 *sti*-lõpulist lemmat. Seetõttu võiksime põhimõtteliselt püüda andmestikke puhastada ka **automaatselt**. Selleks võib kasutada vähemalt kaht viisi. Esiteks: kuna vaadeldavate *lt*- ja *sti*-lõpuliste määrsõnade kohta peaks leiduma ka vastav omadussõna (nt *avalik* → *avalikult*), võime iga lemma puhul kustutada sõnavormi lõpust liite ära ning lasta eesti keele **morfoloogia analüsaatoril** (vt ptk 3.3) ilma oletamiseta hinnata, kas saadud tüvi on analüüsitav kui mõne omadussõna või selle võrdevormi ainsuse omastava vorm (nt *avaliku-*, *raske-*, *kiire-*, *võimsama-*) või omadussõna

funktsioonis partitsiibi ainsuse vorm (nt *eksinu-*, *seotu-*). Teatud määrsõnade lemmadele, näiteks *hästi*, mille puhul tuletuse aluseks olev omadussõna on teistsuguse tüvega, määrame skriptis (*vordle_omadussona_omastavaga.py*) vastava omadussõna (*hea*) käsitsi. Sõnavormid, mille puhul omadussõna analüüsi ei pakuta (nt *tõe-*, *nime-*), jätame analüüsist välja.

Sellisel moel kogutud ja puhastatud andmestikest võiksid jääda välja näiteks kõiksugu trüki- ja kirjavigadega vormid (nt *korralilult*, *lihtssalt*, *tõelsielt*), mitte omadussõnadest tuletatud määrsõnad (nt *mujalt*, *sealt*, *püsti*, *alasti*) ning käändsõnade muutevormid (nt *eestilt*). Samas jäävad välja aga ka paljud analüsaatorile tundmatud kõnekeelsused (*lebolt*), veebikeelele iseloomulikud graafilised kirjakujud (nt *6nnelikult*, *tookalt*) ja võõrkeelsetest omadussõnadest tuletatud vormid (nt *fancymalt*, *leanilt*). Sellisel moel automaatne puhastamine ei toimi aga kunagi vigadeta ning EstNLTK morfoloogia analüsaator võib pakkuda ka ilma oletamise parameetri sisselülitamiseta omadussõna analüüse ka paljudele vigastele sõnavormidele ja vormidele, mis tegelikult omadussõna analüüsi ei tohiks saada (nt *abosuluutse-* → *abosuluutne*, *sea-* → *siga*, vt tabel 3).

Tabel 3. Näide morfoloogia analüsaatori abil automaatselt puhastatud *lt*-lõpulistest vormidest

***lt*-lõpuline sõne**	***lt*-lõpulise sõne järjekorranumber**	***lt*-lõpulise sõne tuletusalus**	***lt*-lõpulise sõne viimase osise tuletusalus**
abosuluutselt	406 333 478	*abosuluutne*	*abosuluutne*
ajutiselt	169 482 216	*ajutine*	*ajutine*
kuri_kuulsalt	330 795 052	*kurikuulus*	*kuulus*
majanduslikult	104 139 844	*majanduslik*	*majanduslik*
osaliselt	372 686 596	*osaline*	*osaline*
peamiselt	714 062 067	*peamine*	*peamine*
sealt	318 737 310	*siga*	*siga*
tahtlikult	12 824 911	*tahtlik*	*tahtlik*
tunduvalt	661 951 020	*tunduv*	*tunduv*
õrnalt	569 618 422	*õrn*	*õrn*

Teine võimalus andmestikku **automaatselt** puhastada oleks võrrelda ilma tuletusliiteta vorme (mis peaksid vastama omadussõnade ainsuse omastava vormidele)

mingi (**digitaalse**) **sõnastiku** paradigma kirjetega, kui neid on võimalik kätte saada. Nii võime iga korpusest kogutud sõne puhul lõigata sõne lõpust ära *lt*- või *sti*-järjendi ning kontrollida, kas järelejäänud vorm (nt *raske-*, *kiire-*, *sea-*) on sõnastiku paradigmades märgitud mõne omadussõna ainsuse omastava vormiks (tabel 4). Sellise lahenduse kitsaskoht seisneb jällegi selles, et sõnastikud ei pruugi sisaldada produktiivsete mallide põhjal moodustatud spontaansete tuletiste tüvesid (nt *grilliva-*, *asimovliku-*, *hoiatavama-*, *tõdemusliku-*). Samuti võib erineda see, kuidas ja kuhu on korpuses ja sõnastikus märgitud liitsõnapiire.

Tabel 4. Näide EKI ühendsõnastiku 2021 abil automaatselt puhastatud *lt*-lõpulistest vormidest (Ekilexi API abil sõnastiku andmeid on kogunud Mari Aigro)

***lt*-lõpuline sõne**	***lt*-lõpulise sõne järjekorranumber**	***lt*-lõpulise sõne viimane osis**	***lt*-lõpulise sõne viimase osise liiteta tüvi**
juriidiliselt	623 793 604	*juriidiliselt*	*juriidilise*
jätkuvalt	365 515 716	*jätkuvalt*	*jätkuva*
lihtsalt	540 430 915	*lihtsalt*	*lihtsa*
lühidalt	260 177 094	*lühidalt*	*lühida*
otseselt	710 616 491	*otseselt*	*otsese*
pidevalt	247 429 708	*pidevalt*	*pideva*
sealt	543 932 070	*sealt*	*sea*
ära_tuntavalt	239 069 779	*tuntavalt*	*tuntava*
umbes-täpselt	440 831 665	*täpselt*	*täpse*
visuaalselt	260 134 031	*visuaalselt*	*visuaalse*

Tabelis 5 on esitatud produktiivsuse arvutamisel kasutatavad loendusandmed algsetes korpusest kogutud (puhastamata) andmetes ning kolmel eri meetodil puhastatud andmetes. Nagu näeme, on andmete puhastamisel väga suur roll, kuna vastasel juhul saavad tüüpide ja ainukite arvule toetuvad produktiivsuse hinnangud kõvasti moonutatud. Samuti on suured erinevused eri puhastamismeetodite vahel: võrreldes käsitsi puhastamisega jääb morfoloogia analüsaatoriga puhastamisel sisse oluliselt rohkem, sõnastikupõhise puhastamisega oluliselt vähem vorme.

Tabel 5. Eesti keele ühendkorpuse veebikorpuste alamosast kogutud *lt*- ja *sti*-tuletiste sõnede, tüüpide ja ainukite arv vastavalt andmestiku puhastamise meetodile

	Liide	Algsetes andmetes	Pärast käsitsi puhastamist (% algsetest andmetest)	Pärast morfana-lüsaatoriga puhastamist (% algsetest andmetest)	Pärast sõnastikuga puhastamist (% algsetest andmetest)
Sõnede arv	*-lt*	11 116 719	9 323 101 (83,9%)	9 480 962 (85,3%)	8 831 252 (79,4%)
	-sti	2 777 626	2 349 329 (84,6%)	2 270 831 (81,8%)	1 626 563 (58,6%)
Tüüpide arv	*-lt*	30 639	14 126 (46,1%)	21 931 (71,6%)	5367 (17,5%)
	-sti	380	174 (45,8%)	241 (63,4%)	140 (36,8%)
Ainukite arv	*-lt*	16 588	5065 (30,5%)	10 490 (63,2%)	612 (3,69%)
	-sti	86	14 (16,3%)	61 (70,9%)	8 (9,3%)

Kuna automaatne puhastamine ei anna kõnealuse teema puhul kõige usaldusväärsemaid tulemusi, kasutame selles miniuurimuses edaspidi sõnade käsitsi parandatud nimekirju.

3.3. Produktiivsuse analüüsimine

Lõpuks saame **analüüsida puhastatud sõnade nimekirju**. Selleks, et arvutada eespool kirjeldatud produktiivsuse mõõdikuid, peame lugema kokku, 1) kui palju tuletiste kasutusjuhte ehk sõnesid esines (*Nc*), 2) kui palju esines kummagi tuletusliitega tüüpe ehk unikaalseid vorme (*V*) ja 3) kui paljud tüüpidest esinesid ainult ühe korra (*Nc1*). Lisaks nendele nn *hapax legomenon*'idele loeme kokku ka kaks (*Nc2*) ja kolm korda (*Nc3*) esinenud tüüpide arvu ehk *dis* ja *tris legomenon*'id. Seda selleks, et vähendada veidi ainukite eristaatust ning võtta arvesse üldisemalt suure korpuse seisukohast haruldasi sõnu.

Üks võimalus mingite konstruktsioonide produktiivsust hinnata on arvutada nimetatud sagedused ja produktiivsuse mõõdikute väärtused terve korpuse kohta korraga, st seisus, kus oleme näinud kõiki korpuse sõnu. Sellisel juhul saame iga võrreldava konstruktsiooni iga mõõdiku kohta ainult ühe väärtuse (tabel 6).

Tabel 6. *lt*- ja *sti*-konstruktsioonide sõnede arv (*Nc*), tüüpide arv (*V*), üks, kaks ja kolm korda esinevad tüüpide arvud (vastavalt *Nc1*, *Nc2* ja *Nc3*)

Liide	Nc	V	Nc1	Nc2	Nc3
-lt	9 323 101	14 126	5065	1662	833
-sti	2 349 329	174	14	9	3

Näeme, et ehkki *lt*-lõpulisi määrsõnu on kasutatud korpuses umbes neli korda rohkem kui *sti*-lõpulisi määrsõnu (*Nc*), esineb erinevaid *lt*-liitelisi tüüpe (*V*) korpuses tervelt 81 korda rohkem kui erinevaid *sti*-lõpulisi tüüpe. Niisiis on *lt*-konstruktsiooni **realiseerunud produktiivsus**, mis mõõdab konstruktsiooni abil moodustatud sõnavara suurust, oluliselt kõrgem kui *sti*-konstruktsiooni oma. Kui vaadata *sti*-sõnu lähemalt (tabel 7), moodustavad enamiku lemmade arvust sõnade *kindlasti*, *hästi*, *täiesti, uuesti*, *kiiresti* ja *enamasti* kasutusjuhud, mis ei ole enamasti kasutusel viisimäärustena, vaid kas modaaladverbina (*kindlasti*), intensiivistajana (*hästi*, *täiesti*) või ajamäärusena (*uuesti*, *enamasti*). Ka *lt*-lõpuliste määrsõnade sagedustabeli tipus troonivad modaaladverbid ja diskursusemarkerid (*lihtsalt*, *tegelikult*, *vähemalt*, *ilmselt*), ent nende osakaal kõikidest vaadeldava liitega lemmadest ei ole nii suur kui *sti*-liitelistel sõnadel.

Tabel 7. Kümme sagedaimat *lt*- ja *sti*-lõpulist määrsõna

***lt*-lemma**	**Sagedus**	**% kõigist *lt*-lemmadest**	***sti*-lemma**	**Sagedus**	**% kõigist *sti*-lemmadest**
lihtsalt	636 647	6,83	*kindlasti*	470 896	20,04
tegelikult	409 743	4,39	*hästi*	464 904	19,79
vähemalt	363 180	3,90	*täiesti*	264 375	11,25
ilmselt	269 686	2,89	*uuesti*	170 140	7,24
täpselt	217 473	2,33	*kiiresti*	165 607	7,05
tavaliselt	157 049	1,68	*enamasti*	105 048	4,47
piisavalt	156 686	1,68	*ilusasti*	70 658	3,01
loomulikult	136 656	1,47	*kenasti*	63 032	2,68
pidevalt	129 779	1,39	*kõvasti*	60 509	2,58
üldiselt	112 731	1,21	*valesti*	51 656	2,20

Realiseerunud produktiivsuse erinevuses peegeldub küllap ka asjaolu, et ehkki mõlemal tuletuskonstruktsioonil on sarnased semantilised piirangud, ei ole *lt*-tuletistel morfoloogilisi moodustuspiiranguid ning neid on moodustatud omajagu ka kesksõna vormidest. Samuti võivad *lt*-tuletised tunduda teatud struktuuriga (nt eesti keeles sagedaimate *lik*- ja *ne*-lõpuliste) omadussõnade puhul loomulikumad, ehkki millegi loomulikkust saab korpuse materjali põhjal hinnata vaid spekulatiivselt ning vormide vastuvõetavuse väljaselgitamise jaoks sobivad paremini hinnangukatsed. Erinevus realiseerunud produktiivsuses võib viidata aga näiteks ka sellele, et lühemad *lt*-tuletised on võimalikud *sti*-tuletiste tähendusväljad üle võtnud ning viisimääruste jaoks, mille puhul *sti*-tuletise moodustamine olnuks küll võimalik, oli vastav *lt*-tuletis juba olemas ning piisavalt sagedasti kasutusel.

See, miks näeme korpuses nii vähe (erinevaid) *sti*-liitelisi määrsõnu, võib tuleneda aga ka palju proosalisematest teguritest, näiteks sellest, et morfoloogia analüsaator ei tunne harvemaid (produktiivselt moodustatud) sõnavorme määrsõnadena ära ning määrab neile mõne muu sõnaliigi (harilikult nimisõna või pärisnime) analüüsi, mida me määrsõnu otsiva skriptiga seetõttu kätte ei saa. Näiteks leiame ühendkorpusest Sketch Engine'i kaudu otsides (CQL-päringuga `[lemma_lc=".*sti" & tag!="D"]`) esinemisjuhte ka vormidest *imelikusti*, *tavalisesti*, *avalikusti*, *tõsisesti* või *turvalisesti*, ent need on analüüsitud nimisõnadeks. Ka *kindlasti* on tervelt 1537 juhul saanud mõne muu analüüsi kui määrsõna oma, samamoodi on käändsõnadeks analüüsitud terve hulk mittestandardseid vorme, nagu *töesti*, *hasti*, *h2sti*, *kövasti* jms. Eriti suurte korpuste automaatselt kogutud ja märgendatud materjalidega töötades, mille puhul ei ole selget ülevaadet, milliseid otsusi sõnavormide klassifitseerimisel on tehtud, tuleb seega alati arvesse võtta, et mingi osa infot võib olla läinud kaduma.

Ehkki tüübisagedusel on oluline roll mingi (tuletus)konstruktsiooni tähenduse kujunemisel ning see on võrdlemisi intuitiivne produktiivsuse näitaja, ei saa selle põhjal siiski hinnata, kui tõenäoliselt konstruktsiooni ka tulevikus sõnavara rikastamiseks edasi kasutatakse. Selleks saab kasutada kaht teist, n-ö tulevikku vaatavat produktiivsuse mõõdikut.

Nagu varem öeldud, väljendab **laiendav produktiivsus** tõenäosust, et mis tahes järgmise korpusesse lisanduva täiesti uue sõna jaoks kasutatakse just kõnealust tuletuskonstruktsiooni. Üldiselt leitakse laiendav produktiivsus konstruktsiooni ainukite ja kogu korpuse ainukite arvu suhtena. Kuna võrdleme aga tuletuskonstruktsioonide produktiivsust samas korpuses, on kogu korpuse ainukite arv mõlema tuletuskonstruktsiooni puhul sama ning seega võime laiendavat produktiivsust võrrelda siin lihtsalt üks kord esinevate *lt*- ja *sti*-tuletiste arvu põhjal (*Nc1*). Nõnda väldime küllaltki ressursimahukat protsessi, kus peaksime oma väga suures korpuses esinevate sõnade kõiki esinemiskordi kokku lugema ja sellest vaid ühe korra esinevate sõnade hulga välja võtma (rääkimata sellest, kui tahame arvestada ka sellega, et üks kord esinevate sõnade hulgas on palju vigaseid vorme). Ka laiendava produktiivsuse põhjal on *lt*-konstruktsioon oluliselt produktiivsem, mis tähendab, et tõenäosus, et järgmine korpusesse lisanduv sõna, mida seal varem ei

ole esinenud või mida on esinenud ainult paar korda, on just *lt*-lõpuline määrsõna, on üle 372 korra suurem kui tõenäosus, et see on *sti*-lõpuline määrsõna (vt tabel 6).

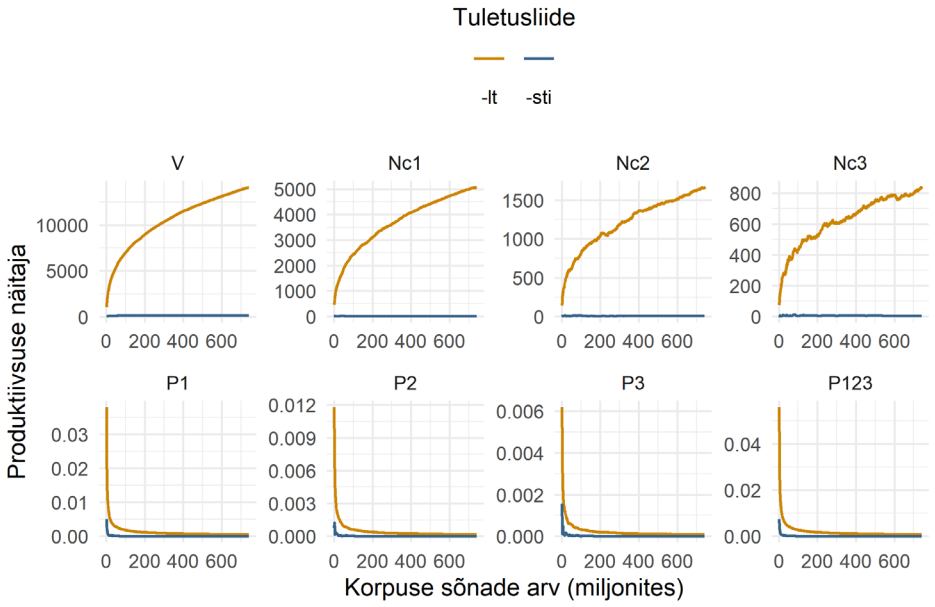
Nende sageduste põhjal saame arvutada ka **tüübi-sõne suhte** (TTR) ning **potentsiaalse produktiivsuse** näitajad (tabel 8), mis suhestavad harvaesinevate tüüpide arvu kõikide konstruktsiooni sõnede arvuga ning väljendavad seeläbi tõenäosust, et mis tahes järgmine *lt*- või *sti*-liitega moodustatud sõna on selline sõna, mida me pole korpuses varem kohanud. Mõlema mõõdiku väärtus on seda suurem, mida vähem on sõnede hulgas väga sagedasi tüüpe ning mida enam on erinevaid harva kasutatud tüüpe. Kui aga suurema osa sõnede hulgast hõlmavad vaid paari üksiku sõna väga sagedased kasutusjuhud, on mõõdikute väärtus väiksem.

Tabel 8. *lt*- ja *sti*-konstruktsioonide tüübi-sõne suhe (*TTR*) ning potentsiaalse produktiivsuse näitajad, kui arvestada ainult üks kord esinevaid tüüpe (*P1*), ainult kaks korda esinevaid tüüpe (*P2*), ainult kolm korda esinevaid tüüpe (*P3*) ja üks kuni kolm korda esinevaid tüüpe korraga (*P123*).

Liide	TTR	P1	P2	P3	P123
-lt	0,001515	0,000543	0,000178	0,000089	0,000811
-sti	0,000074	0,000006	0,000004	0,000001	0,000011

Näeme, et *lt*-liide on tõepoolest *sti*-liitest kõikide produktiivsuse näitajate poolest oluliselt produktiivsem ning et üks kord esinevaid tüüpe on mõlema liite puhul rohkem kui kaks ja kolm korda esinevaid tüüpe, mistõttu on ka *P2* ja *P3* väiksemad kui *P1*.

Teine viis produktiivsust hinnata on jagada korpus ühesuurusteks, nt miljonist sõnast koosnevateks osadeks ning arvutada mõõdikud kumulatiivselt iga osa (pluss sellele eelnenud osade) kohta eraldi, justkui käiksime sõnahaaval korpust läbi ning hindaksime iga miljoni sõna järel, kui palju erinevaid ja üks, kaks või kolm korda esinevaid tüüpe selleks hetkeks oleme kohanud. Kuna salvestasime andmestike kogumisel ka iga tuletise järjekorranumbri terve korpuse sõnade hulgas, on selle põhjal võimalik määrata, millised tuletised jäävad korpuse tekste läbi käies esimese miljoni nähtud sõna hulka, millised esimese kahe miljoni hulka ja nii edasi. Nõnda saame arvutada produktiivsuse mõõdikutele terve hulga erinevaid väärtusi ja nende näitajate põhjal joonistada nn produktiivsuse kasvu- või kahanemiskõveraid (joonis 3).

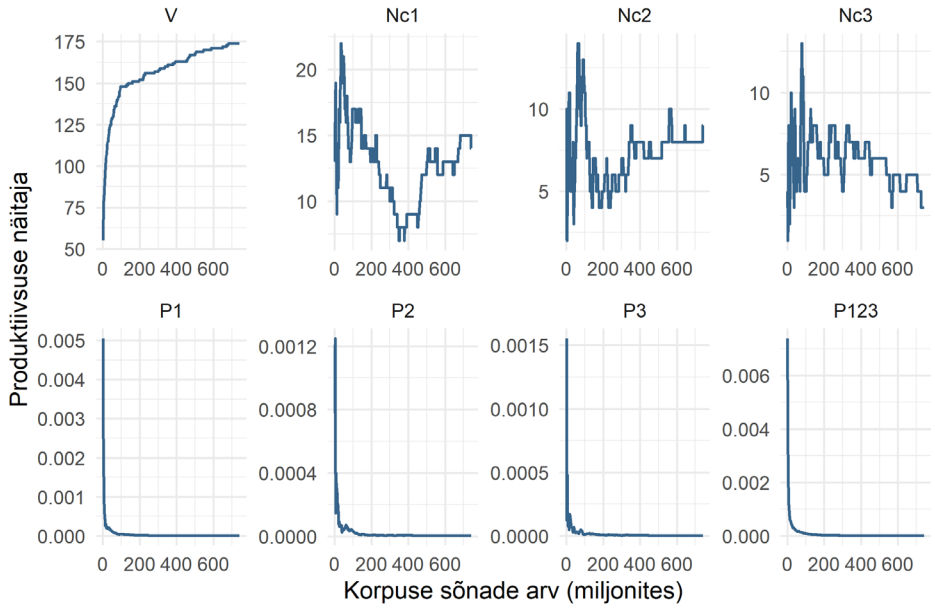


Joonis 3. *lt*- ja *sti*-konstruktsioonide produktiivsuse näitajate kasvu- ja kahanemiskõverad

Jooniselt näeme, et *lt*-konstruktsioonide tüüpide arv (*V*) ning 1, 2 ja 3 korda esinevate tüüpide arv (*Nc1, Nc2, Nc3*) kasvab järsult korpuse esimeste tekstide põhjal ning seejärel stabiilselt korpuse üldise sõnade arvu suurenedes. See on tavaline produktiivsete mallide muster: varem kohtamata (või ainult korra või paar kohatud) sõnu näeme korpuse esimeste sõnade hulgas rohkesti, seejärel hakkavad sõnad korduma ning uusi tuletisi tuleb ette harvem, ent stabiilselt. See tähendab, et konstruktsioone ei esinda ainult üksikud sagedased vormid, mis tulevad ette juba korpuse esimestes tekstides, vaid tuletusmalle on läinud uute tuletiste moodustamise jaoks tarvis kõikides korpuse alamosades. Potentsiaalse produktiivsuse kõverad (*P1, P2, P3, P123*) on jällegi vastupidised, kuna tegu on tõenäosusega: tõenäosus, et kohtame mõnd varem nägemata *lt*-tuletist on suur, kui oleme läbi vaadanud ainult väikese hulga korpuse sõnu, ent mida rohkem sõnu oleme juba näinud, seda väiksemaks läheb ka potentsiaalne produktiivsus. Ehkki uusi tüüpe moodustatakse korpuses pidevalt juurde, langeb vajadus uute moodustiste järele niisiis järsult juba pärast esimest paari miljonit sõna. Pärast seda kasutatakse niisiis sagedamini juba olemasolevaid tuletisi ning sõnavara rikastamise vajadus tekib harvemini. Mida laugem on sealjuures potentsiaalse produktiivsuse kurv, seda produktiivsemaks võib konstruktsiooni pidada.

Kuna *-lt* on nii palju sagedasem ja produktiivsem, on sellega võrreldes *sti*-konstruktsiooni kõverad joonistatud väga väikeste väärtuste põhjal ning

paistavad peaaegu sirged. Kui vaadata aga *sti*-konstruktsiooni eraldi, näeme mõnevõrra sarnaseid trende, ent kuna skaalad on väiksemad, on ka igasugu kõikumised nähtavamad (joonis 4).

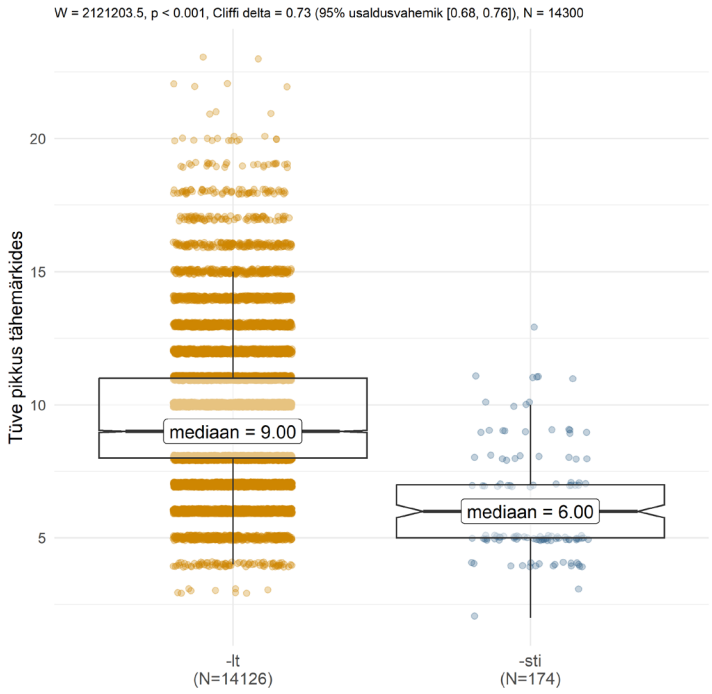


Joonis 4. *sti*-konstruktsioonide produktiivsuse näitajate kasvu- ja kahanemiskõverad

Joonis 4 viitab sellele, et *sti*-liite (kasutus)produktiivsus on oluliselt piiratum kui *lt*-liite oma: põhiosa tüüpidest (*V*) ilmub esimese 100 miljoni sõna jooksul ning sealt edasi tuleb uusi, varem nägemata *sti*-liitelisi sõnu ette harva. Ainult üks, kaks ja kolm korda esinevate tüüpide kõverad (*Nc1, Nc2, Nc3*) ei ole ühesuunalised ning kõiguvad üles-alla: näiteks ainukite kõver langeb, kuni oleme läbi käinud 400 miljonit korpuse sõna, ent hakkab siis jälle kasvama. Kuna aga *y*-teljel nähtav skaala on väga piiratud (jäädes alati alla 25 tüübi), võib selliseid kõikumisi pidada ka juhuslikeks ning kui venitaksime skaalat suuremaks (nt 0st 100ni), näeksime vaid sirgeid jooni. Potentsiaalse produktiivsuse negatiivsed kõverad räägivad sama lugu nagu *lt*-liitegi puhul (mida enam korpuse sõnu oleme näinud, seda väiksemaks jääb tõenäosus, et korpusesse lisanduv *sti*-tuletis oleks sõna, mida me varem korpuses pole kohanud), ent kõverad on väga järsud: see tähendab, et *sti*-konstruktsioon ammendab enda võimalikud tüübid suhteliselt kiiresti ning uusi, produktiivselt moodustatud tuletisi luuakse üliharva.

3.4. Tuletuskonstruktsioonide võrdlemine

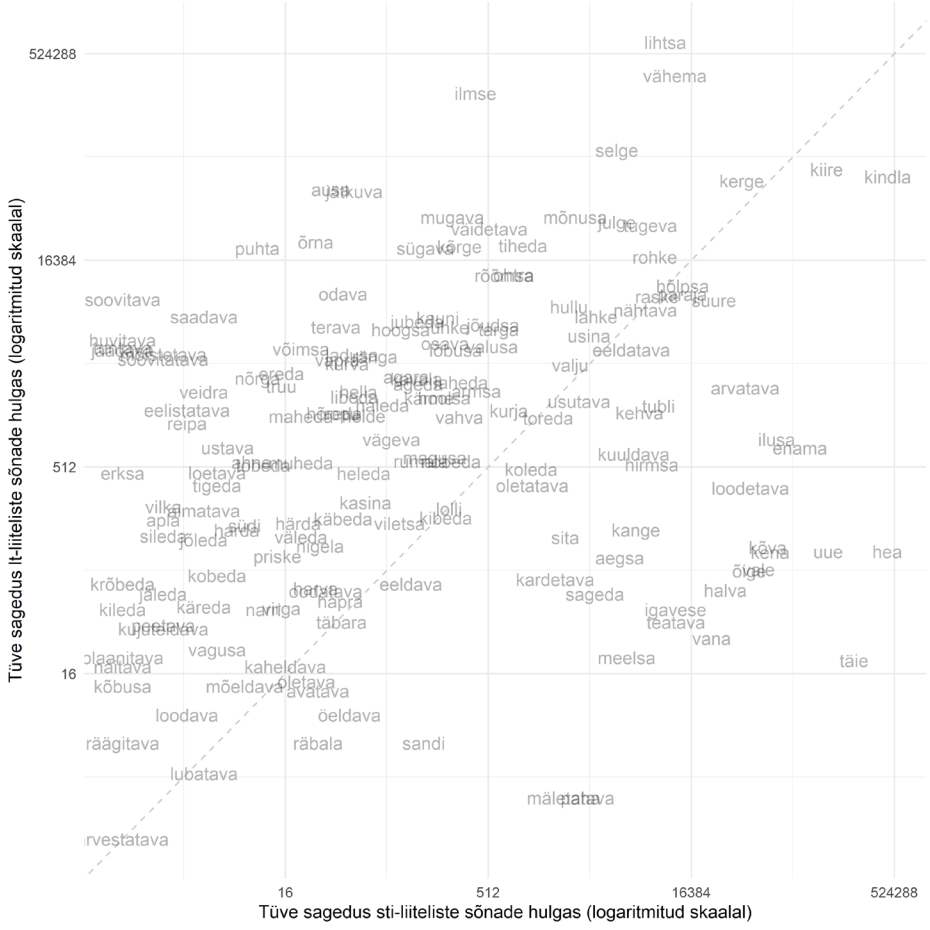
Miks me selliseid erinevusi kahe liite produktiivsuses näeme? Põhjuseid selleks võib olla mitu. Üheks neist on kindlasti piirangud sobivatele alussõnadele: *lt*-konstruktsiooni saab moodustada kõikidest kesksõnadest ning *-lt* liitub hõlpsasti ka komplekssetele omadussõnadele, eriti sagedasti (*li*)*ne*- ja *lik*-liitelistele omadussõnadele. Viimast illustreerib kaudselt ka näiteks erinevus *lt*- ja *sti*-määrsõnade tüvede keskmises pikkuses (joonis 5): *lt*-liidetega sõnade tüved on keskmiselt oluliselt pikemad (mediaan on 9 tähemärki) kui *sti*-liidetega sõnade tüved (mediaan on 6 tähemärki). Kuna *-sti* lisab võrreldes *lt*-liitega sõnale ühe silbi juurde (vrd *hõlpsalt* ja *hõlpsasti*), võib liite harvema ja piiratuma kasutuse taga olla ka vormiökonoomsuse taotlus.



Joonis 5. *lt*- ja *sti*-konstruktsioonide tüvepikkuste jaotused karpdiagrammil. Manni-Whitney U-testi põhjal on erinevus kahe konstruktsiooni vahel statistiliselt oluline (W = 2121203,5, p < 0,001) ja suur (Cliffi delta = 0,73 [0,68, 0,76])

Teine tegur, mis kasutusproduktiivsust võib piirata, on nn tüübiblokeerimine (ingl *type blocking*, vt nt Bauer 2001: 136–138), mispuhul lähedaste tähenduste väljendamist võimaldavate ning seetõttu omavahel võistlevate konstruktsioonide hulgast

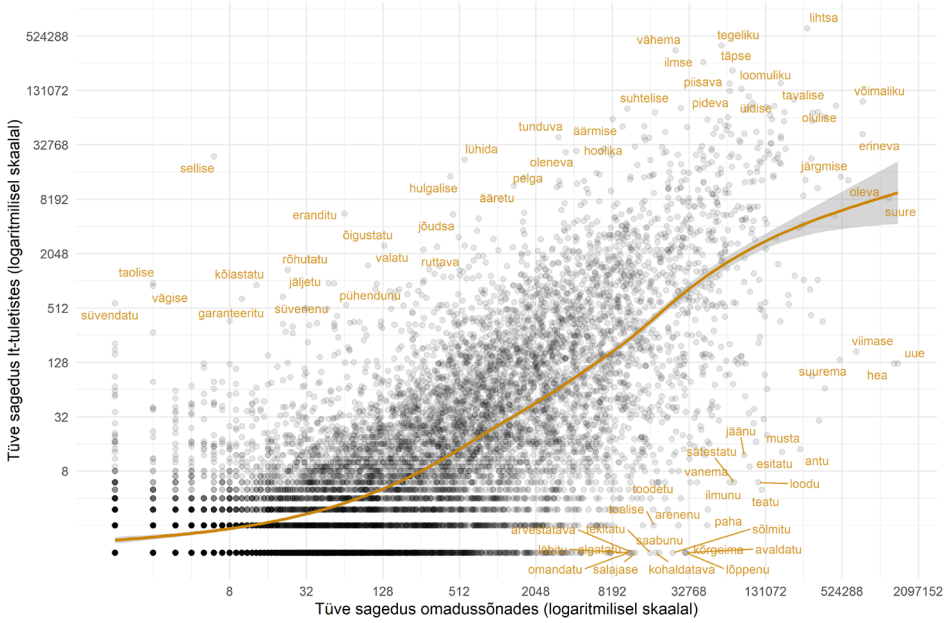
võib ühe kasutamine tõrjuda teise kasutust samas tähenduses. Tähendustele ilma kontekstita mõistagi kuigi hästi ligi ei pääse. Küll aga on näidatud, et vähemasti tuletuskonstruktsioonide puhul kipuvad väga kõrge sagedusega tuletised esinema mingis muus tähenduses kui seda on tuletuskonstruktsiooni põhitähendus. Kuna sellised sõnad on kasutuses rohkem automatiseerunud ega nõua sõna tähenduse protsessimiseks mingi abstraktsema konstruktsioonitähenduse kasutamist (nt teadmist, et *lt*-liitega võib moodustada viisi märkivaid määrsõnu), muutub konstruktsiooni skemaatiline tähendus ja produktiivsus seda nõrgemaks, mida rohkem selliseid väga sagedasi, idiosünkraatiliselt käituvaid lekseeme tuletuskonstruktsiooni esindajate hulgas on. Siin võiksime vaadelda eraldi nende tüvede sagedusi, millega on andmestikus moodustatud tuletisi mõlema liitega (joonis 6).



Joonis 6. Tüvede esinemise sagedus *lt*- ja *sti*-liitega (kuvatud on ainult tüvesid, mis esinevad vähemalt ühe korra mõlemas tuletuskonstruktsioonis)

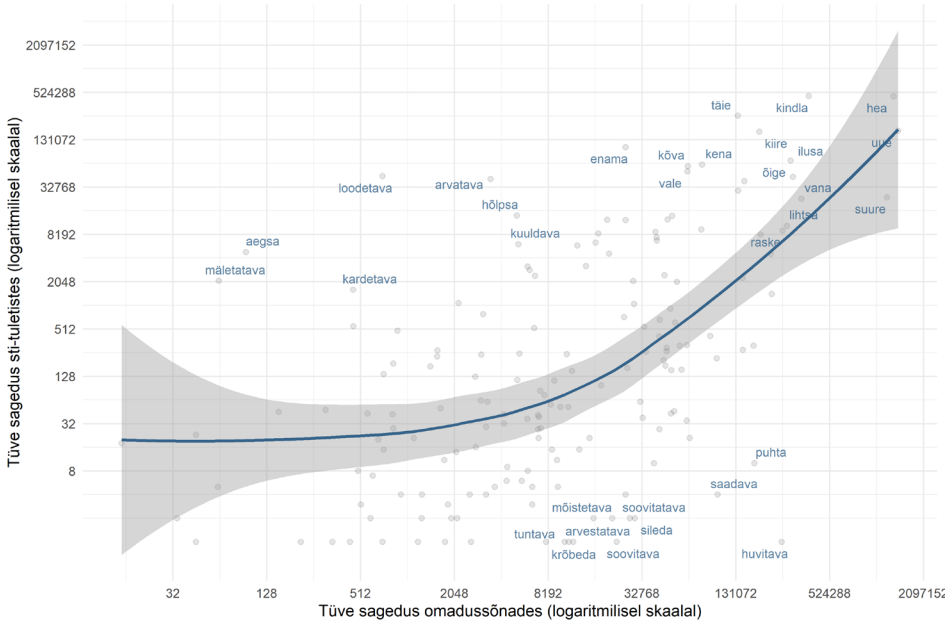
Näeme esiteks, et jagatavad tüved on tüüpiliselt kas lihttüved või moodustatud *tav*-keskkõnast, ning teiseks, et suurema osa ühiste tüvede puhul on nende kasutus *lt*-tuletiste hulgas sagedasem kui *sti*-tuletiste hulgas (enamik tüvesid jääb hallist katkendjoonest ülespoole). Nõnda võib ühe tuletise kõrge kasutussagedus pärssida teise kasutamist, näiteks mitmed partitsiipidest tüved (*soovitava-, huvitava-, mõistetava-, jäädava-*) on väga sagedasti kasutusel *lt*-lõpuliste määrsõnadena, aga väga harva *sti*-lõpulistena. Suure osa katkendjoone lähedusse ning pigem sagedusspektri alumisse ja keskossa jäävate tüvede puhul on siiski märgatav sarnane trend: mida sagedasem on tüvi *lt*-tuletiste hulgas, seda sagedasem on see ka *sti*-tuletiste hulgas. Sellised sõnad (nt *kurjalt/kurjasti*, *viletsalt/viletsasti, virgalt/virgasti, rohkelt/rohkesti, kergelt/kergesti*) on suure tõenäosusega niisiis produktiivselt moodustatud tuletised, mille tähendusel on suur ühisosa kõnealuste tuletuskonstruktsioonide üldise viisi näitava tähendusega. On ka rühm tüvesid, mille kasutus on sagedasem hoopis *sti*-tuletiste hulgas (nt *mäletatavasti*, *loodetavasti, teatavasti, kardetavasti, valesti, kõvasti, kenasti, uuesti, pahasti*, *vanasti, meelsasti, õigesti, täiesti, enamasti*, *hästi/heasti*). Paljud neist sõnadest ei ole tüüpiliselt kasutusel viisimäärustena, vaid intensiivistajate (*hästi, täiesti, kangesti*), ajamääruste (*uuesti*, *enamasti*, *vanasti*, *aegsasti*, *sagedasti*) ja modaaladverbidena (*teatavasti*, *meelsasti*, *mäletatavasti*, *loodetavasti*).

Võime ka võrrelda tuletise tüve ja eeldatava tuletusaluse sõna esinemissagedust korpuses. Produktiivse moodustuse puhul võiksime eeldada, et tuletise (nt *banaalselt*) suhteline sagedus korpuses korreleerub tuletusaluse sõna (*banaalne*) suhtelise sagedusega, sest mõisteid läheb tarvis samal määral, lihtsalt erinevates süntaktilistes kontekstides. Ebaproduktiivse moodustuse puhul aga võib ühe vormi suhteline sagedus olla oluliselt kõrgem seetõttu, et sõnavormi ja tuletuskonstruktsiooni üldine tähendus on üksteisest lahknenud (vt nt Pilvik 2021). Kuna salvestasime korpusefailidest *lt-* ja *sti*-lõpulisi lemmasid kogudes ka kõik ette tulnud omadussõna analüüsiga lemmad, võime morfoloogia süntesaatoriga sedapuhku lasta hoopis sünteesida iga omadussõna ainsuse omastava vormi (nt *banaalne* → *banaalse*, skript *synteesi_omastava_vorme.py*) ning võrrelda selle põhjal, kuidas tüve kasutussagedus tuletistes korreleerub tüve kasutussagedusega omadussõnades (joonised 7 ja 8). Kuvades sageduste suhet hajuvusgraafikul, võime omakorda vaadelda lähemalt mingit huvipakkuvat sagedusspektri osa. Joonistel 7 ja 8 on eraldi välja toodud näiteks tüved, mis esinevad sagedasti tuletistes, aga harvemini omadussõnades (x-telje madalam sagedusspekter graafiku ülemises vasakpoolses osas), sagedasti omadussõnades, aga harva tuletistes (y-telje madalam sagedusspekter graafiku alumises parempoolses osas) või sagedasti mõlemas kontekstis (x- ja y-telje kõrge sagedusspekter graafiku ülemises parempoolses osas).



Joonis 7. Tüvede esinemissagedused *lt*-tuletistes ja omadussõnades. Trendijoon kuvab üldistatud aditiivse mudeli (ingl *generalized additive models*, Wood 2011; 2017) ennustust. Mudeli põhjal seletab omadussõna logaritmitud sagedus 57,97% tüve esinemissagedusest *lt*-tuletises

Näeme esiteks, et *lt*-tuletiste puhul on tugev korrelatsioon selle vahel, kui sagedasti keelekasutajatel mõnd omadussõna vaja läheb, ning selle vahel, kui sagedasti läheb seda omadust tarvis sündmusviisi tähistamiseks. See viitab selgelt *lt*-konstruktsiooni produktiivsusele ja avatusele, kuna keelekasutaja saab semantiliselt reeglipärasel viisil omadussõnadest moodustada vajadusel viisimäärusi. Sõnade hulgas, mis jäävad üldisest trendijoonest eemale, leiame palju partitsiipe. Osa neist on tavalised omadussõnades, aga mitte *lt*-tuletistes (nt *avaldatud*, *arenenud*, *sõlmitud*, *arvestatav*, *algatatud, antud, teatud*), osa tavalised *lt*-tuletistes, aga mitte omadussõnades (nt *süvendatult*, *õigustatult*, *pühendunult*, *rõhutatult*, *garanteeritult*), osa, eeskätt leksikaliseerunud *v*-keskSõnad, on jällegi tavalised mõlemas (nt *tunduv/tunduvalt*, *olenev/olenevalt*, *pidev/pidevalt*, *piisav/piisavalt*). Trendijoonest eemal leiame ka kõrge kasutussagedusega modaalpartiklid ja diskursusmarkerid (nt *tegelikult*, *ilmselt*, *üldiselt*, *lihtsalt*, *täpselt*), mille tähendus ja funktsioon lahkneb kasutuses sageli tuletuskonstruktsiooni üldisest sündmusviisi markeerivast tähendusest.



Joonis 8. Tüvede esinemissagedused *sti*-tuletistes ja omadussõnades. Trendijoon kuvab üldistatud aditiivse mudeli (ingl *generalized additive models*, Wood 2011; 2017) ennustust. Mudeli põhjal seletab omadussõna logaritmitud sagedus 31,83% tüve esinemissagedusest *sti*-tuletistes

Ehkki *sti*-tuletisi on andmetes oluliselt vähem, näeme ka nende puhul tegelikult positiivset korrelatsiooni omadussõna (logaritmitud) sageduse ning tüve (logaritmitud) kasutussageduse vahel tuletistes. Seega võib eeldada, et ka suur osa *sti*-tuletistest on moodustatud produktiivse malli järgi ning nende tähendus ja funktsioon ei hälbi konstruktsiooni üldisest sündmusviisi märkivast tähendusest. Seega võib veidi lihtsustades öelda, et kui *lt*-tuletuse produktiivsust toetab suur tüüpide arv ning reeglipärane moodustamine sagedastest komplekssetest (nt *lik-* ja (*li*)*ne*-lõpulistest) omadussõnadest ja partitsiipidest, samas kui *lt*-lõpuline vorm võiks vormihomonüümia tõttu alaltütleva käändega konstruktsiooni produktiivsust pigem pärssida, siis *sti*-tuletuskonstruktsiooni produktiivsust hoiab tüüpide vähesusest hoolimata alal semantiline reeglipärasus ja analoogia (vt Barðdal 2008: 91).

Sellesse lühiuurimusse sõnatähenduste põhjalikum ja üksikasjalikum analüüs paraku ei mahu ning ilma kontekstita oleks see ka üksjagu spekulatiivne. Nii tuletuskonstruktsioonide omavaheliste ja tuletiste ning tuletusaluste sõnade semantiliste vahekordade eritlus võimaldaks aga põhjalikumat sissevaadet nii määrsõnatuletuse produktiivsuse erinevatesse aspektidesse kui ka üksiksõnade grammatiseerumise astmesse. Lisaks on veel terve hulk viise, kuidas uurimuse teemaga võiks edasi minna. Kuna tegemist on tuletistega, oleks huvitav produktiivsuse

hindamisel arvesse võtta ka nende üldist sõnapere suurust (nt *kiire, kiirelt*, *kiiresti, kiiremini*, *kiirendama*, *kiirenema*, *kiirustama* jne) või sõnaperesse kuuluvate sõnade sagedust. Samuti oleks põhjalikumas uurimuses tarbeline võrrelda eri tuletiste kasutuskontekstidest tuletatud tähendusvektoreid (ingl *embeddings*), et objektiivsemalt tuvastada, millised samatüvelised *lt*- ja *sti*-tuletised on kasutusel sarnastes kontekstides (= sarnastes tähendustes) ja millised erinevates.

Kokkuvõte

Juhtumiuuringus võrdlesime kahe eesti keele tuletuskonstruktsiooni – *lt*- ja *sti*-konstruktsiooni – produktiivsust, kasutades eesti keele ühendkorpuse 2021. aasta veebi alamkorpusest kogutud sagedusandmeid. Tuletisi otsisime lemma vormi lõpujärjendi (*-lt* või *-sti*) ning määrsõna sõnaliigimärgendi järgi, et jätta välja kõiksugu muude sõnaliikide *lt*- või *sti*-lõpulised vormid. Andmete kogumiseks korpuse korpusefailidest kasutasime Pythoni skripte ning kogutud andmete analüüsimiseks ja visualiseerimiseks R-i võimalusi. Produktiivsuse kvantitatiivseks hindamiseks rakendasime teiste keelte morfoloogiauurimustes (ja üha rohkem ka süntaktiliste konstruktsioonide uurimustes) laialdaselt kasutatud mõõdikuid: realiseerunud produktiivsus (ehk tüüpide arv), tüübi- ja sõnesageduse suhe, laiendav produktiivsus ning potentsiaalne produktiivsus. Kuna mõõdikud toetuvad erinevate tüüpide ja ainult üks kord esinevate tüüpide arvule ning on seega tundlikud korpustes ette tulevatele kirja- ja trükivigadele, võrdlesime ka korpusandmestike käsitsi ja automaatse puhastamise tulemusi. Materjali käsitsi puhastamine on ajakulukas, ent annab kõige usaldusväärsemaid tulemusi ning võimaldab ka analüüsitavasse materjali rohkem sisse süübida, samas kui erinevad automaatse puhastamise meetodid on kiiremad, ent võivad relevantsete vormide arvu tugevalt üle- või alahinnata.

Uurimuses saime kinnitust emakeelse kõneleja intuitsioonile, et *lt*-liide on oluliselt produktiivsem kui *sti*-liide, mis nähtub nii erinevate kasutusel olevate lekseemide arvus kui ka tõenäosuses, millega sõnavara on kummagi liite abil võimalik edasi rikastada. *sti*-liite produktiivsust pärssivate põhjustena võib välja tuua nii moodustuslikud piirangud (tuletisi moodustatakse harva komplekssetest omadussõnadest ja muudest partitsiipidest peale *tav*-partitsiibi), võimaliku vormiökonoomsuse taotluse (*-sti* lisab sõnale ühe silbi, samas kui *-lt* ei lisa) kui ka tüübiblokeerimise (sagedasti kasutusel olevad *lt*-lõpulised viisimäärused võivad piirata vastavate *sti*-lõpuliste vormide levikut). Teisalt nägime adjektiivide ja tuletiste tüvede kasutussagedusi võrreldes, et mõlema konstruktsiooni puhul on positiivne korrelatsioon selle vahel, kui sagedasti mõnd omadust tähistatakse, ning selle vahel, kui sagedasti moodustatakse selle omadusega sündmusviisi tähistavaid adverbe. Seetõttu võib mõlemat pidada produktiivseks moodustusmalliks, ent kasutuses piiravad kahe konstruktsiooni produktiivsust erinevad tegurid. Samadest

tüvedest moodustatud *lt*- ja *sti*-lõpuliste sõnade sagedusi võrreldes nägime ka, et sõnad, mille tüve sagedus *lt*-lõpulistes sõnades korreleerub tüve sagedusega *sti*-lõpulistes sõnades, esindavad tõenäolisemalt tuletuskonstruktsioonide viisi väljendavat põhitähendust ning on moodustatud produktiivselt, samas kui sõnad, mille tüve suhteline kasutussagedus ühe liitega on teise omast oluliselt suurem, esinevad tõenäoliselt muus tähenduses või funktsioonis, nt ajamääruse, intensiivistaja või pragmaatilise partiklina.

Näidisuurimuse valmimist on toetanud EKKD projekt „Eesti keele morfosüntaktiline varieerumine" (2024–2027).

Kirjandus

Aavik, Johannes. 1936. *Eesti õigekeelsuse õpik ja grammatika*. Tartu: Noor-Eesti. http://hdl.handle.net/10062/29254.

Baayen, R. Harald. 1992. A quantitative aspects of morphological productivity. Geert E. Booij & Jaap Van Marle (toim), *Yearbook of Morphology 1991*, 109–149. Dordrecht: Springer. https://doi.org/10.1007/978-94-011-2516-1_8.

Baayen, R. Harald. 1994. Productivity in language production. *Language and Cognitive Processes* 9(3). 447–469.

Baayen, R. Harald. 1996. The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics* 22(4). 455–480.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. Anke Lüdeling & Merja Kytö (toim), *Corpus Linguistics. An International Handbook*, 900–919. Berlin: Mouton de Gruyter.

Baayen, R. Harald & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29. 801–834.

Baayen, R. Harald & Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language* 72. 69–96.

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic* (Constructional Approaches to Language 8). Amsterdam / Philadelphia: John Benjamins. https://doi.org/10.1075/cal.8.

Bauer, Laurie. 2001. *Morphological productivity* (Cambridge Studies in Linguistics 95). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486210.

Booij, Geert E. 2018. *The construction of words: Advances in Construction Morphology* (Studies in Morphology 4). Cham: Springer.

Erelt, Mati, Tiiu Erelt & Kristiina Ross. 2020. *Eesti keele käsiraamat*. Uuendatud tr. Tallinn: Eesti Keele Sihtasutus.

Heede, Margot Van den & Peter Lauwers. 2023. Syntactic productivity under the microscope: The lexical and semantic openness of Dutch minimizing

constructions. *Folia Linguistica* 57(3). 723–761. https://doi.org/10.1515/flin-2023-2028.
Kasik, Reet. 2015. *Sõnamoodustus* (Eesti keele varamu 1). Tartu: Tartu Ülikooli Kirjastus. http://hdl.handle.net/10062/50084.
Kerge, Krista. 2002. Kirjakeele kasutusvaldkondade süntaktiline keerukus. Reet Kasik (toim), *Tekstid ja taustad: artikleid tekstianalüüsist* (Tartu Ülikooli eesti keele õppetooli toimetised 23), 29–46. Tartu: Tartu Ülikooli Kirjastus.
Kerge, Krista. 2003. *Keele variatiivsus ja* mine*-tuletus allkeelte süntaktilise keerukuse tegurina* (Tallinna Pedagoogikaülikooli humanitaarteaduste dissertatsioonid 10). Tallinn: TPÜ Kirjastus.
Koppel, Kristina & Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 18. 207–228. https://doi.org/10.5128/ERYa18.12.
Küngas, Annika. 2013. *lt-* ja *sti*-liiteliste adverbide kasutuse muutumine eesti kirjakeeles. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 4(3). 73–92. https://doi.org/10.12697/jeful.2013.4.3.04.
Muischnek, Kadri & Sahkai. 2010. Liitpredikaadid leksikoni-grammatika kontiinumil: konstruktsioonide produktiivsusest verbiga *minema* moodustatud liitpredikaatide näitel. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 2. 295–316.
Mäearu, Sirje. 2000. *lt*-liitelised määrsõnad. *Keelenõuanne soovitab* 2. https://keeleabi.eki.ee/artiklid2/lt.html.
Pilvik, Maarja-Liisa. 2021. *Action nouns in a constructional network: A corpus-based investigation of the productivity and functions of the deverbal suffix* -mine *in five different registers of Estonian* (Dissertationes philologiae estonicae Universitatis Tartuensis 48). Tartu: University of Tartu Press.
R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Zeldes, Amir. 2012. *Productivity in argument selection: From morphology to syntax*. Berlin / Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110303919.
Valdmets, Annika. 2013. Modal particles, discourse markers, and adverbs with *lt*-suffix in Estonian. Liesbeth Degand, Bert Cornillie & Paola Pietrandrea (toim), *Discourse Markers and Modal Particles: Categorization and Description*, 107–132. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.234.
Wood, Simon N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* 73(1). 3–36.
Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. 2. tr. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.

Kollostruktuuriline analüüs

Jane Padrik

Lühikokkuvõte

Selles peatükis tutvustatakse kahte korpuslingvistilist meetodit, mida tuntakse üldnime all *kollostruktuuriline analüüs* ja mille abil saab korpusandmete põhjal uurida keele grammatilist struktuuri. Näidisuurimusena on kasutatud kohakäände ja kohakaassõna paralleelset kasutust eesti keeles, nt *Raamat on laual vs. Raamat on laua peal*. Peatüki sissejuhatavas osas tutvustatakse kollostruktuurilist analüüsi, seejärel kirjeldatakse kahe meetodi – lihtne kollekseemanalüüs ja distinktiivne kollekseemanalüüs – rakendamist. Peatüki kokkuvõtvas osas arutletakse, milleks see meetod hea on ja milleks ta nii hea ei ole.

1. Sissejuhatus: mis on kollostruktuuriline analüüs ja milleks seda kasutada?

Kollostruktuuriline analüüs on meetod, millega saab uurida keele grammatilist struktuuri, kasutades selleks korpuslingvistilisi vahendeid. See meetod võimaldab uurida seoseid sõnade ja konstruktsioonide vahel. **Konstruktsioon** tähendab siinkohal vormi ja tähenduse kooslust, mille komponendid ei ole teineteisest lahutatavad. Näiteks võib tuua konstruktsiooni, mis koosneb alalütlevas käändes tegijast, modaalsest verbist ning *da*-tegevusnime vormis tegusõnast: *Tal tuleb ennast käsile võtta* (Penjam 2009). Keeleteadlastena võib meid huvitada, millised on sagedased nimi- või asesõnad, mis esinevad selles konstruktsioonis alalütlevas käändes või millised on sagedased verbid, mis esinevad selles konstruktsioonis *da*-tegevusnime vormis. Kollostruktuurilise analüüsi eesmärgiks on kirjeldada grammatiliste konstruktsioonide tähendust ja seeläbi näidata, mis ulatuses mingid kindlad sõnad on kindlate konstruktsioonidega lõimitud. Peatüki sissejuhatav osa annab ülevaate kollostruktuurilise analüüsi kujunemisloost ja sellega seonduvatest keeleteoreetilistest küsimustest, mis on olulised korpuslingvistika seisukohast laiemalt. Peatüki teine osa keskendub meetodi selgitusele, kasutades näidisuurimusena alalütleva käände ja kaassõna *peal* paralleelset kasutust eesti keeles. Peatüki võtame lühidalt

kokku paari mõttega kollostruktuurilise analüüsi kitsaskohtadest ja edasistest võimalikest arendustest.

Kollostruktuurilise analüüsi tõid keeleteaduse maastikule Anatol Stefanowitsch ja Stefan Gries, kes on mitmes uurimuses selle lähenemise põhimõtteid lahti kirjutanud ja selgitanud. Nendest olulisemad uurimused on Stefanowitsch & Gries (2003); Gries & Stefanowitsch (2004); Stefanowitsch & Gries (2005); Stefanowitsch & Gries (2009); Stefanowitsch (2013); Stefanowitsch (2014); Gries (2019). Nendest lähtub ka siinne õpiku peatükk. Lisaks annavad lugejasõbraliku ülevaate kollostruktuurilisest analüüsist Gilquin (2013) ja Hilpert (2012; 2014). Eesti keeles on kollostruktuurilisest analüüsist pikemalt kirjutanud Kristel Uiboaed (2013). Kollostruktuurilise analüüsi meetodid taanduvad ühele korpuslingvistikas enimlevinud eeldusele – distributiivsele hüpoteesile (ingl *distributional hypothesis*). Siinpuhul viidatakse tihti Firthi kuulsale väljaütlemisele „You shall know a word by the company it keeps" (Firth 1968: 179) või Harrise mööndusele „[d]ifference in meaning correlates with difference in distribution" (Harris 1954).

Kollostruktuuriline analüüs on katusterminiks tervele meetodite kogumile – lihtne kollekseemanalüüs, distinktiivne kollekseemanalüüs ja koosvarieerumise kollekseemanalüüs. **Lihtsa kollekseemanalüüsiga** (ingl *simple collexeme analysis*) saab uurida konstruktsiooni ja selle konstruktsiooni kindlas positsioonis esinevate sõnade vahelisi seoseid. **Distinktiivne kollekseemanalüüs** (ingl *distinctive collexeme analysis*) võimaldab uurida sõnade seoseid kahe funktsionaalselt seotud konstruktsiooniga. **Koosvarieerumise kollekseemanalüüsi** (ingl *co-varying collexeme analysis*) raskuskeskmeks on seosed kindla konstruktsiooni erinevates positsioonides esinevate sõnade vahel. Siinses peatükis tuleb täpsemalt juttu kahest esimesest meetodist – lihtsast kollekseemanalüüsist ja distinktiivsest kollekseemanalüüsist. Koosvarieerumise kollekseemanalüüsi meetodi kirjeldust ja näidisuurimusi võib vaadata lähemalt Griesi ja Stefanowitschi artiklitest, nt (Stefanowitsch & Gries 2005; Stefanowitsch 2013).

Kollostruktuurilise analüüsi keskseks mõisteks on konstruktsiooni **kollekseemid** ehk sõnad, mis on seotud grammatiliste konstruktsioonidega. Meetodi väljatöötamise taustal on põhimõtted, mis on pikalt olnud kasutusel kollokatsioonide ja grammatiliste mustrite analüüsimisel. Mis eristab kollostruktuurilist analüüsi tavapärasest kollokatsioonide uurimisest on see, et selle asemel, et võrrelda kahe sõna koosesinemise sagedust nende individuaalse esinemisega ülejäänud korpuses (vt ptk 5.2.5 „Kollokatsioonid"), võrreldakse sõnade individuaalset esinemist kindla grammatilise struktuuri kindlas positsioonis. Kollostruktuurilise meetodi eripära seisneb veel selles, et see järgib rangeid kvantitatiivse teadustöö põhimõtteid. Kuigi Stefanowitsch (2020: 270) on ise öelnud, et see lähenemine on pigem induktiivne (vt ptk 1.1.1 „Induktiivne ja deduktiivne lähenemine"), siis võib seda kasutada ka teatud grammatiliste konstruktsioonide kohta käivate hüpoteeside testimiseks.

Tavapärane korpuslingvistiline uurimus keskendub konkordantsiridadele, mida keeleuurija oma uurimisküsimusest lähtuvalt analüüsib. Paraku tuleb

keeleteaduses laiemalt ette juhtusid, kus korpusi kasutatakse selleks, et välja noppida uurimuse kontekstis hästi töötavaid näiteid (ingl *cherry picking*). See on viinud tõdemuseni, et korpused on alakasutatud – keeleteadlastel on üpris pikalt võtnud aega, et ära tabada korpuste täispotentsiaal keeleteadusliku uurimistöö kontekstis. Kui aga võtta süstemaatilisem lähenemine, on võimalik läbi viia metodoloogiliselt rangemate printsiipidega kvantitatiivset korpusanalüüsi. Just see soov on olnud kollostruktuurilise analüüsi meetodi väljatöötamise taustal.

Algselt oli eesmärgiks laiendada juba olemasolevat uurimistööd kollokatsioonide uurimises leksikaal-grammatilistele suhetele. Grammatika uurimine korpuslingvistiliste meetoditega ei ole aga niisama lihtne ja sellise uurimusega kaasneb rida probleeme. Nii näiteks on üheks probleemikohaks uuritava grammatilise struktuuri operatsionaliseerimine või defineerimine – millist konkreetset keelekasutust pidada just uuritava grammatilise struktuuri näiteks. Mida keerulisem (nt abstraktsem, raskesti hoomatavam) on grammatiline struktuur, mida uuritakse, seda keerulisem on seda korpuslingvistilises võtmes operatsionaliseerida (vt ka õpiku ptk 1.1.2 „Uurimisküsimuse esitamine, hüpoteesi püstitamine"). Teiseks laiemaks probleemkohaks on välja mõelda, kuidas teha korpuspäringuid nii, et saaksime vajalikud grammatilised struktuurid korpusest kätte. Morfoloogiliselt markeeritud ja lihtsamate grammatiliste struktuuride puhul on seda lihtsam teha. Näitena keerulisest grammatilisest konstruktsioonist, mida korpusest kätte saada, on transitiivne verb. Siin teeb asja keeruliseks see, et paljud nimisõnafraasid, mis verbile järgnevad, ei ole sihitised, nt *Jüri magas terve päeva*. Lisaks on sage, et grammatika uurimise juurde korpuslingvistiliste vahenditega kuulub ajamahukas käsitöö – andmete puhastamine.

Juba kirjeldatud kahe probleemi valguses on selge, et grammatika uurimine on korpuslingvistide jaoks keerukam kui sõnavara uurimine. Seega on loomulik, et paljud grammatikauurijad on korpustega tegutsemisel lähtunud sõnakesksest lähenemisest. Sõnakesksus tähendab siin seda, et sõnatasandi kollokatsioonide uurimise loogika on üle kantud grammatika uurimisse. Kollostruktuurilist lähenemist võib pidada struktuuritundlikuks lähenemiseks selles mõttes, et meid huvitab sõnade kasutus mingis kindlas grammatilises struktuuris, millel on kindel sõnade järjekord ja mis esinevad teineteise suhtes mingis kindlas grammatilises positsioonis, nt omadussõnad suhtes nimisõnadega jne. Kollostruktuurilise lähenemise loogika seisneb selles, et üks tunnus koosneb sõnavarast (tunnuse väärtusteks on individuaalsed sõnad) ja teine tunnus koosneb mingist grammatilise struktuuri aspektist. Selline sõnade kaudu lähenemine lihtsustab grammatika korpuslingvistilist uurimist.

Kollostruktuuriliste meetodite väljakujunemisel on olulist rolli mänginud **konstruktsioonigrammatika** areng (Fillmore, Kay & O'Connor 1988; Goldberg 1995; Croft 2001; Goldberg 2005). Konstruktsioonigrammatika on keeleteaduslik teooria, mille peamine seisukoht on, et keel on konstruktsioonide kogum ja iga konstruktsioon on vormi ja tähenduse ühend. Konstruktsioonigrammatika uurib

keeles esinevaid konstruktsioone ning nendevahelisi seoseid ja erinevusi. Kuigi kollostruktuuriline analüüs ei piirdu vaid konstruktsioonigrammatikaga ja seda analüüsimeetodit võib kasutada ükskõik millises keeleteooria raamistikus, mis tegeleb mingilgi määral keelekasutusega, on just argumentstruktuurikonstruktsioonid olnud kollostruktuuriliste analüüside algseks materjaliks. Kollostruktuuriline analüüs lähtub korpuslingvistika loogikast, et sõnade ja grammatiliste üksuste vahelist suhet saab uurida, kasutades teatud seose tugevuse mõõdikuid – grammatiline struktuur on lihtsalt järjekordne tingimus, mille võtmes jälgida leksikaalsete üksuste koosesinemist. On arvatud, et keelekõnelejad teevad keelelise sisendi kohta alateadlikult statistilist analüüsi ja et statistilised seosed, mis andmetes avalduvad, peegelduvad psühholingvistiliste seostena keelekasutajate peas (Stefanowitsch & Gries 2003: 236–237; Gries & Stefanowitsch 2004: 123; Stefanowitsch 2006: 74).

2. Eesti keele alalütlev kääne ja kaassõna *peal*

Järgnevalt näitlikustame kollostruktuurilise lähenemise kahte meetodit, kasutades kahte väliskoha väljendamiseks kasutatud konstruktsiooni eesti keeles – nimisõna alalütlevas käändes (näide 1: *laual*) ja sellega paralleelselt kasutatav kaassõna *peal*, millele eelneb nimisõna omastavas käändes (näide 2: *laua peal*).

(1) *Raamat on laual.*
(2) *Raamat on laua peal.*

Uurimuse laiem taust on kasutuspõhine keeleteadus, mille üks põhialustest on nn sünonüümia vältimise printsiip (ingl *the principle of no synonymy*): kui kaks üksust keeles vormiliselt erinevad, siis erinevad need üksused ka tähenduse poolest (Goldberg 1995: 3; 67). Olgugi, et eesti keeles saame öelda nii *Raamat on laual* kui ka *Raamat on laua peal*, võime sünonüümia vältimise printsiibist lähtuvalt eeldada, et need kaks lauset ei tähenda täpselt sama asja. Keeleteadlast võib huvitada, miks valib keelekasutaja mõnikord käändega ja teinekord kaassõnaga konstruktsiooni. Kollostruktuuriline analüüs on üks nendest korpuslingvistilistest meetoditest, mis aitab seda välja selgitada. Meie uurimisküsimuseks on seega „Millised sõnad esinevad sagedasti alalütlevas käändes ja millised sõnad kaassõnaga *peal* ja seda just asukoha väljendamisel?“. Eeldame, et nimisõnad, mis esinevad alalütlevas käändes, viitavad sagedamini suurtele ja liikumatutele objektidele (nt tänav) ja kaassõnaga *peal* koosesinevad nimisõnad väiksematele ja liikuvatele või liigutatavatele objektidele (nt kapp). Seega kasutame kollostruktuurilist meetodit deduktiivselt (vt ptk 1.1.1) – meil on varasemate uurimuste (nt Klavan 2012) põhjal kujunenud välja hüpotees, mida soovime korpuslingvistilise meetodiga kontrollida.

Sellise uurimismaterjali korpuslingvistilise analüüsi puhul on esimeseks probleemiks see, kuidas nende kahe konstruktsiooni esinemised korpusest kätte saada. Kui kaassõna *peal* konstruktsioonide väljavõtmisega korpusest ei ole väga palju

vaeva, kuna tegemist on muutumatu sõnaga, siis käändekonstruktsiooni väljavõtmiseks on vaja morfoloogiliselt märgendatud korpust. Näidisuurimuse andmed pärinevad eesti keele ühendkorpusest (ÜK 2021, suurus 2,4 miljardit sõna). Ühendkorpused on mahult suurimad eesti keele korpused, mis koosnevad suures osas veebist kogutud tekstidest (vt õpiku ptk 2.2.3). Korpuspäringute tegemiseks saame kasutada korpusanalüüsi tarkvara Sketch Engine veebiliidest. ÜK 2021 on sõnestatud, lausestatud, osalausestatud, lemmatiseeritud ja morfoloogiliselt ning sõltuvussüntaktiliselt märgendatud (Koppel & Kallas 2022: 208). Siinse näidisuurimuse kontekstis on just morfoloogiline märgendamine kriitilise tähtsusega – see võimaldab korpusest Sketch Engine'i päringukeelt (CQL) kasutades välja võtta alalütleva käände kasutused (vt ka ptk 5.2.3. „Konkordantside koostamine grammatilise info põhjal"). Konkreetsed päringud, mida näidisuurimuses Sketch Engine'i kaudu on kasutatud, on toodud välja vastava kollostruktuurilise meetodi seletuse juures. Uurimuse kollostrukuurilised analüüsid on tehtud R-iga (R Core Team 2023), kasutades Stefan Griesi välja töötatud koodiridu. Juhised selleks leiab Stefan Griesi kodulehelt[1].

Enne veel, kui liigume kollostruktuurilise analüüsi enda juurde, on vaja selgitada üht olulist nüanssi, mis teeb näidisuurimuse korpusvalimi kokkupanemise keeruliseks. Nimelt ei ole kaks uuritavat konstruktsiooni paralleelselt kasutusel kõigis tähendustes. Peale kohatähenduse tarvitatakse alalütlevat käänet ka aega väljendavate nimisõnadega, nt *sel nädalal, tol hetkel, eelmisel aastal*. Veel võivad alalütlevas käändes esinevad sõnad väljendada „spetsiifilisi, kontekstuaalselt avalduvaid tähendusi" (Erelt jt 1993: 54–55): väliskohakäänded, sh alalütlev kääne, kuuluvad kohustusliku laiendi vormina mitmesse rektsioonistruktuuri, mistõttu vastavad nad paljudel juhtudel indoeuroopa keelte daativi funktsioonidele (Vainik 1995: 162).

„Eesti grammatika" (Metslang jt 2023: 156) loetleb järgnevad alalütleva käände funktsioonid:

a) asukoht, sh asukohana mõistetud sündmus: *Traktor töötab põllul*; *Üliõpilased on loengul*;
b) seisund: *Nägu on naerul*;
c) toimumisaeg: *õhtul*, *kolmapäeval*;
d) valdaja, kogeja või muu tegevussubjekt: *See ununes mul täielikult*;
e) vahend: *Jaan mängis kitarril ühe loo. Sõitsin jalgrattal tööle*;
f) viis: *valjul häälel*.

Lisaks võivad alalütlevas käändes olla sõltuvusmäärus ja tingimusmäärus (Metslang jt 2023: 156).

Loetletud funktsioonidest kasutatakse kaassõna *peal* alalütleva käände paralleelvariandina vaid lokatiivse tähendusega koha (a) ja vahendi (e) väljendamiseks.

[1] https://www.stgries.info/teaching/groningen/index.html

Siinses näidisuurimuses ei ole rakendatud andmestiku käsitsi kodeerimist, et välja sõeluda just need kasutusjuhud, kus alalütlev kääne ja kaassõna *peal* on koha või vahendi tähenduses. Selline semantiline analüüs ei ole paraku korpuslingvistika seisukohast automaatselt tehtav ja nõuaks palju töötunde, et kogu materjal käsitsi läbi käia. Seega on valimi moodustamisel võetud arvesse kõiki alalütleva käände ja kaassõna *peal* funktsioone.

2.1. Lihtne kollekseemanalüüs

Kollostruktuurilise analüüsi kõige esimene ja lihtsam variant püüab välja selgitada statistilisi seoseid, mis esinevad konstruktsiooni kindla positsiooni ja selles positsioonis esinevate sõnade vahel. Nii näiteks võime vaadelda ainult alalütlevas käändes nimisõnu ning küsida, millised nimisõnad on tugevalt alalütleva käändega seotud. Sellises uurimuses on kaks nominaalset tunnust – alalütlevas käändes nimisõna lekseem ja kääne. Eeldame, et kõige sagedamini esinevad alalütlevas käändes nimisõnad, mis väljendavad kohta, aega ja tegevussubjekti.

Lihtsast kollekseemanalüüsist annavad hea ülevaate Stefanowitsch ja Gries (2003), veidi hilisema põhjaliku meetodi kirjelduse on esitanud Stefanowitsch (2013). Selleks, et välja selgitada statistiline seos leksikaalse üksuse l_i, mis kuulub sõnaliiki *L* (nt üksus *laud* sõnaliigist *nimisõna*), ja konstruktsiooni *c* vahel, mis kuulub konstruktsioonide klassi *C* (nt alalütlev kääne käändekonstruktsioonide klassist), on vaja teada nelja esinemissagedust:

(i) leksikaalse üksuse $\boldsymbol{l_i}$ (nt *laud*) sagedus konstruktsioonis ***c*** (nt *laual*);
(ii) leksikaalse üksuse $\boldsymbol{l_i}$ sagedus teistes konstruktsioonides, mis kuuluvad klassi ***C*** (nt *lauale*, *laud*, *lauda* jne, aga mitte *laual*);
(iii) teiste leksikaalsete üksuste sagedus sõnaliigist ***L***, mis esinevad konstruktsioonis ***c*** (nt *toolil*, *hobusel* jne, aga mitte *laual*);
(iv) teiste leksikaalsete üksuste sagedus sõnaliigist ***L***, mis esinevad teistes klassi ***C*** konstruktsioonides (nt *toolile*, *hobust* jne, aga mitte *toolil*, *hobusel*).

Tabelis 1 on näitena toodud kõik neli sagedust nimisõna *laud* esinemise kohta alalütleva käände konstruktsioonis (ÜK 2021).

Tabel 1. Lihtne kollekseemanalüüs: nimisõna *laud* alalütlevas käändes

	laud	**Teised nimisõnad**	**Kokku**
alalütlev kääne	54 357	44 741 555	44 795 912
teised käänded	375 374	717 132 323	717 507 697
Kokku	429 731	761 873 878	762 303 609

Sellise sagedusinfo põhjal saab välja arvutada seose tugevuse (st kas leksikaalne üksus l_i on konstruktsioonis *c* oodatust vähem või rohkem sage), kasutades standardseid seose tugevuse hindamise statistilise analüüsi protseduure, nt logaritmilise tõepärasuhte kriitiliste väärtuste põhjal saadud *p*-väärtust. Siinkohal tasub tähele panna, et kollostruktuuriline analüüs kui kvantitatiivne korpuslingvistiline meetod eeldab mõningast baasteadmist lihtsamatest statistilistest meetoditest (nende kohta vt ptk 6 „Korpusandmete statistiline analüüs"). Keeleteaduses on mitmeid häid statistika käsiraamatuid, kust saab huvi korral ennast selles vallas täiendada (Levshina 2015; Brezina 2018; Winter 2019; Gries 2021).

Tabelis 1 toodud sagedused ainult ühe leksikaalse üksuse kohta (*laud*) on ÜK 2021 korpusest Sketch Engine'i konkordantsiotsingu kaudu leitud järgmiselt:

1. lahter „Kokku & Kokku" (väärtus 762 303 609) ehk kõik nimisõnad korpuses on leitud CQL-päringuga `[tag="S"]`;
2. lahter „Kokku & *laud*" (väärtus 429 731) on leitud CQL-päringuga `[lemma="laud" & tag="S"]`;
3. lahter „alalütlev kääne & Kokku" (väärtus 44 795 912) on leitud kahe päringu tulemused kokku liites: alalütlev kääne ainsuses `[tag="S" & features="sg_ad"]` + alalütlev kääne mitmuses `[tag="S" & features="pl_ad"]`[2];
4. lahter „alalütlev kääne & *laud*" (väärtus 54 357) on leitud kahe päringu tulemused kokku liites: *laual* `[lemma="laud" & tag="S" & features="sg_ad"]` + *laudadel* `[lemma="laud" & tag="S" & features="pl_ad"]`[3];
5. lahter „teised käänded & *laud*" (väärtus 375 374) on leitud lahutades lahtri „Kokku & *laud*" väärtusest lahtri „alalütlev kääne & *laud*" väärtuse (429 731 – 54 357 = 375 374);
6. lahter „alalütlev kääne & teised nimisõnad" (väärtus 44 741 555) on leitud lahutades lahtri „alalütlev kääne & Kokku" väärtusest lahtri „alalütlev kääne & *laud*" väärtuse (44 795 912 – 54 357 = 44 741 555);
7. lahter „Kokku & teised nimisõnad" (väärtus 761 873 878) on leitud lahutades lahtri „Kokku & Kokku" väärtusest lahtri „Kokku & *laud*" väärtuse (762 303 609 – 429 731 = 761 873 878);
8. lahter „teised käänded & Kokku" (väärtus 717 507 697) on leitud lahutades lahtri „Kokku & Kokku" väärtusest lahtri „alalütlev kääne & Kokku" väärtuse (762 303 609 – 44 795 912 = 717 507 697);

[2] Sama tulemuse saab kätte ka ühe päringuga, kasutades atribuudi *features* asemel atribuuti *case*: `[tag="S" & case="ad"]`.

[3] Sama tulemuse saab kätte ka ühe päringuga, kasutades atribuudi *features* asemel atribuuti *case*: `[lemma="laud" & tag="S" & case="ad"]`.

9. lahter „teised käänded & teised nimisõnad“ (väärtus 717 132 323) on leitud lahutades lahtri „Kokku & teised nimisõnad“ väärtusest lahtri „alalütlev kääne & teised nimisõnad“ väärtuse (761 873 878 – 44 741 555 = 717 132 323).

Kui selline sageduste arvutamise protseduur on korratud kõigi leksikaalsete üksustega, mis esinevad uuritavas korpuses konstruktsioonis *c* (nt *laud*, *tool*, *hobune* jne), saab igale leksikaalsele üksusele arvutada seose tugevuse hinnangu, järjestada kollekseemid seose tugevuse järgi ning analüüsida tulemusi uurimisküsimusest lähtuvalt.

Siinse uurimuse kollekseemanalüüs on läbi viidud juhindudes Stefan Griesi koodiridadest (Gries 2024). Selleks, et analüüsi teostada, on kõigepealt vaja ette valmistada andmestik, mis on sisendiks tarkvaraprogrammiga R tehtavale andmeanalüüsile. Andmestikus on vaja esitada eraldi ridadel iga leksikaalse üksuse kohta kahes tulbas kaks sagedust – nimisõna kogusagedus korpuses (nt lahtri „Kokku & *laud*“ väärtus) ja nimisõna sagedus uuritavas konstruktsioonis (nt lahtri „alalütlev kääne & *laud*“ väärtus). Kuna keeles esinevate nimisõnade kogusagedus ja alalütleva käände kogusagedus on liialt suured, oleme siinses uurimuses praktilistel kaalutlustel kasutanud lihtsustatud lähenemist ning võtnud uuritavaks konstruktsiooniks alalütleva käände ainult ainsuslike nimisõnadega. Lisapiirangu uurimusele seab Sketch Engine'i veebiliidese standardlitsents, mis võimaldab alla laadida vaid 1000 sagedasemat sõna.

Sammud andmestiku koostamiseks:

1. Konkordantsiotsing ÜK 2021 korpusest Sketch Engine'i kasutajaliidese kaudu, et leida kõik alalütleva käände kasutused ainsuslike nimisõnadega (CQL-korpuspäringukeele kasutamise kohta vt ptk 5.2.3): `[tag="S" & features="sg_ad"]`;
2. 1000 sagedama alalütlevas käändes esineva ainsusliku nimisõna loendi allalaadimine Sketch Engine'i tööriista *Frequency* kaudu (arvutatud 10 miljoni sõna põhjal);
3. Konkordantsiotsing ÜK 2021 korpusest Sketch Engine'i kasutajaliidese kaudu, et leida kõik nimisõnad (mis tahes käändes ja arvus): `[tag="S"]`;
4. 1000 sagedama ainsusliku nimisõna loendi allalaadimine Sketch Engine'i tööriista *Frequency* kaudu (arvutatud 10 miljoni sõna põhjal).

Sammud kollekseemanalüüsi läbiviimiseks tarkvaraprogrammiga R:

- Laadi alla ja installi tarkvaraprogramm R ning selle laiendus RStudio.
- Käivita RStudio, kopeeri järgmine rida programmi konsooliaknasse (*Console*) ja vajuta *Enter*-klahvi:
  ```
  source("https://www.stgries.info/teaching/groningen/coll.analysis.r")
  ```

- Programmi käivitamiseks kirjuta konsooliaknasse järgmine rida ja vajuta *Enter*-klahvi:
  `coll.analysis()`
- Järgmiseks trükitakse RStudio konsooliaknasse laaditud programmijupi kirjeldus ning palutakse jätkamiseks vajutada *Enter*-klahvi. Kollekseemanalüüsi tegemiseks tee edasi järgmised valikud:
  - analüüs, mida soovid teha: 1 (*collocation/collexeme analysis*)
  - sõna/konstruktsioon, mida uurid: adessiiv
  - korpuse suurus (siinse uurimuse puhul ÜK 2021 suurus): 2410296919
  - kas soovid Fisheri-Yatesi täpse testi tulemusi ('jah' või 'ei'): *no*
  - laadi üles õpiku repositooriumist leitav sisendfail: *collex.txt*

Programm viib nende sätete põhjal läbi analüüsi ning kirjutab väljundi eraldi faili.

Siinse näidisuurimuse lihtsa kollekseemanalüüsi tulemused on kokkuvõtlikult esitatud tabelis 2. Tavapäraselt keskendutakse kollekseemanalüüsi puhul vaid kõige olulisematele kollekseemidele. Tabelis 2 on välja toodud 40 statistiliselt kõige olulisemat kollekseemi, mis esinevad nimisõna positsioonis alalütleva käände konstruktsioonis. Tabeli teine veerg näitab seose tugevust konstruktsiooni ja lekseemi vahel. Seos, mis on tugevam kui 1,3 näitab, et nimisõna ja konstruktsiooni vahel on seos, mis on statistiliselt oluline ($p < 0{,}05$) (Hilpert 2006: 245; vt ka Gries, Hampe & Schönefeld 2005). Mida tugevamalt on konstruktsioon ja nimisõna seotud, seda suurem on seose tugevuse näitaja.

Tabel 2. 40 statistiliselt kõige olulisemat kollekseemi, mis esinevad ainsusliku nimisõna positsioonis alalütleva käände konstruktsioonis

Kollekseem	Seose tugevus	Kollekseem	Seose tugevus
1. üldjuht	4,79	21. suvi	4,04
2. vahendus	4,77	22. alus	4,03
3. hommik	4,48	23. talv	3,98
4. laupäev	4,38	24. juht	3,95
5. hetk	4,34	25. välismaa	3,85
6. kolmapäev	4,34	26. territoorium	3,80
7. neljapäev	4,33	27. viis	3,80
8. teisipäev	4,32	28. kuju	3,79
9. õhtu	4,29	29. istung	3,69
10. määr	4,28	30. lava	3,68

Kollekseem	Seose tugevus	Kollekseem	Seose tugevus
11. tasand	4,26	31. tagajärg	3,68
12. pühapäev	4,25	32. korrus	3,68
13. reede	4,23	33. aadress	3,67
14. sügis	4,23	34. komme	3,61
15. mood	4,20	35. koduleht	3,61
16. esmaspäev	4,20	36. mai	3,59
17. kevad	4,14	37. november	3,56
18. nädalavahetus	4,13	38. väljak	3,55
19. hinnang	4,11	39. august	3,55
20. tänapäev	4,10	40. veebruar	3,51

Sarnaselt alalütleva käändega läbiviidud kollekseemanalüüsile võime läbi teha ka kollekseemanalüüsi kaassõnaga *peal*. Analüüsi protseduur on sama, mis ülalpool kirjeldatud. Selle analüüsi puhul on kasutatud andmestikku *collex_2.txt*.

Tabel 3. 40 statistiliselt kõige olulisemat kollekseemi, mis esinevad ainsuliku nimisõna positsioonis kaassõna *peal* konstruktsioonis

Kollekseem	Seose tugevus	Kollekseem	Seose tugevus
1. silm	5,49	21. mägi	4,06
2. maa	5,45	22. lava	4,02
3. jää	5,11	23. õlg	3,90
4. laud	5,05	24. voodi	3,86
5. nurk	5,00	25. serv	3,81
6. põld	4,95	26. plaat	3,80
7. õu	4,91	27. uks	3,77
8. koht	4,86	28. aken	3,76
9. piir	4,64	29. linn	3,74
10. paber	4,58	30. pakk	3,74

Kollekseem	Seose tugevus	Kollekseem	Seose tugevus
11. kapp	4,57	31. kivi	3,68
12. tool	4,49	32. pilt	3,66
13. kõht	4,46	33. kaas	3,64
14. tee	4,42	34. joon	3,64
15. vesi	4,35	35. pööre	3,58
16. külg	4,34	36. sild	3,58
17. sein	4,29	37. selg	3,57
18. plats	4,26	38. nina	3,56
19. maja	4,20	39. jõgi	3,55
20. pool	4,19	40. kael	3,55

Lisaks sellele, et tabelites 2 ja 3 toodud nimekirjad annavad vastuse keeleteadlase uurimisküsimustele, on loetelu kasulik ka praktilistel kaalutlustel. Näiteks saab loetelu kasutada nii leksikograafias kui ka keeleõppes. Tabelites 2 ja 3 toodud nimisõnade loetelu aitab nii leksikograafil kui keeleõpetajal otsustada, milliseid alalütlevas käändes või kaassõnaga *peal* esinevaid nimisõnu pidada tervikuteks ja millele seetõttu sõnaraamatus või sõnavara õpetamisel rohkem tähelepanu pöörata. Keeleteadlasele pakub tabelites 2 ja 3 toodud nimekiri huvi teoreetilisest vaatepunktist. Näiteks kõrvutades alalütleva käände kollekseeme seesütleva käände kollekseemidega saame järeldusi teha eesti keele kohaväljendite uurimiseks – kas eelistame eesti keeles öelda *hoovis* või *hoovil*. Kollekseemanalüüsi võlu peitub selles, et see aitab üksikute koosesinemiste tasandilt liikuda üldisemale tasandile ja uurida, kas nimisõnad, mis esinevad sagedasti koos alalütleva käändega, moodustavad semantilisi tähendusrühmi. Nii näiteks kinnitab nimisõnade loend tabelis 2 seda, et nimisõnad, mis esinevad oodatust enam alalütlevas käändes, väljendavad enamasti lokatiivse tähendusega aega (nt *üldjuhul*, *hommikul*, *laupäeval*). Sagedased on ka nimisõnade adverbilaadsed kasutused (nt *vahendusel, määral, tasandil*). Oluliselt vähem esineb esimese neljakümne sagedasema alalütleva käände kollekseemi hulgas koha tähendusega nimisõnu; need, mis esinevad, viitavad pigem suurtele ja liikumatutele objektidele (nt *välismaal, territooriumil, laval*). See tulemus lisab kinnitust uurimuses püstitatud hüpoteesile. Konstruktsioonigrammatiku jaoks oleks siit samm edasi arutlemine selle üle, mida konstruktsiooni nimisõna positsioonis esinevad lekseemid ütlevad laiemalt konstruktsiooni enda tähenduse kohta.

2.2. Distinktiivne kollekseemanalüüs

Distinktiivne kollekseemanalüüs on algse, lihtsa kollekseemanalüüsi edasiarendus, mille eesmärgiks on leida **erinevusi** seosetugevustes, mis esinevad kahe omavahel seotud konstruktsiooni mingi kindla positsiooni ja selles positsioonis esinevate sõnade vahel. Seda meetodit on sagedasti rakendatud inglise keele süntaktiliste alternatsioonide, nt daativi alternatsiooni uurimiseks (nt *Mary gave the book to John* 'Mari andis raamatu Juhanile' *vs*. *Mary gave John the book* 'Mari andis Juhanile raamatu'). Siinses näidisuurimuses võrdlen aga omavahel alalütleva käände ja kaassõna *peal* konstruktsioone. Uurimisküsimuse võib sõnastada järgmiselt: Millised nimisõnad esinevad sagedasti alalütlevas käändes ja millised sõnad kaassõnaga *peal*? Selline lähenemine võimaldab uurida, kas kaks omavahel seotud konstruktsiooni erinevad tähenduse poolest. Eeldan, et nimisõnad, mis esinevad alalütlevas käändes, viitavad sagedamini abstraktsetele (nt aasta) ning suurtele ja liikumatutele objektidele (nt tänav) ja kaassõna *peal* väiksematele ja liikuvatele või liigutatavatele objektidele (nt kapp). Kui see nii on, peaks seda peegeldama ka käände ja kaassõnaga esinevate nimisõnade omavaheline võrdlus.

Distinktiivse kollekseemanalüüsi töötasid välja Gries ja Stefanowitsch (2004). Eesmärgiks on võrrelda kõigi sõnade sagedusi, mis mingis konstruktsioonis esinevad, nende sõnade sagedustega teistes, võrreldavates konstruktsioonides. Sellest tulenevalt hõlmavad keelekasutajate teadmised informatsiooni sellest, millised sõnad võrreldavate konstruktsioonide lõikes kõige paremini eristuvad, ilma et oleks oluline, kas need sõnad on üldse nende konstruktsioonidega tugevalt seotud. Selline teadmine võib keelekasutajale kasulik olla, kui tal on vaja eristada peeneid tähendusnüansse pealtnäha samatähenduslike konstruktsioonide vahel. Uurijal lasub vastutus võrreldavate konstruktsioonide valimisel.

Sarnaselt lihtsa kollekseemanalüüsiga läheb statistiliste seoste tugevuste väljaselgitamiseks vaja sagedustest koosnevat risttabelit. Eesmärgiks on välja selgitada, kui tugevalt on seotud üks kindel leksikaalne üksus l_i, mis kuulub sõnaliiki L (nt üksus *laud* sõnaliigist *nimisõna*), konstruktsioonidega c ja d, mis kuuluvad konstruktsioonide klassi C (nt nimisõna alalütlevas käändes ja nimisõna + kaassõna *peal* lokatiivse väliskoha konstruktsioonide klassist). Selleks on meil tarvis teada järgmisi sagedusi:

(i) leksikaalse üksuse $\boldsymbol{l_i}$ (nt *laud*) sagedus konstruktsioonis $\boldsymbol{c}$ (nt *laual*);
(ii) leksikaalse üksuse $\boldsymbol{l_i}$ sagedus konstruktsioonis $\boldsymbol{d}$ (nt *laua peal*);
(iii) teiste leksikaalsete üksuste sagedus sõnaliigist $\boldsymbol{L}$, mis esinevad konstruktsioonis $\boldsymbol{c}$ (nt *toolil, hobusel* jne, aga mitte *laual*);
(iv) teiste leksikaalsete üksuste sagedus sõnaliigist $\boldsymbol{L}$, mis esinevad konstruktsioonis $\boldsymbol{d}$ (nt *tooli peal*, *hobuse peal* jne, aga mitte *laua peal*).

Sellise sagedusinfo põhjal saab välja arvutada seose tugevuse ja suuna (st kas leksikaalne üksus l_i on oodatust vähem või rohkem sage konstruktsioonis c või d), kasutades standardseid seosetugevuse hindamise statistilise analüüsi protseduure, nt logaritmilise tõepärasuhte kriitiliste väärtuste põhjal saadud p-väärtust.

Tabelis 4 on näitena toodud kõik neli sagedust nimisõna *laud* esinemise kohta alalütleva käände ja kaassõna *peal* konstruktsioonis. Sageduste leidmiseks tehti neli CQL-korpuspäringut:

1. *laud* alalütlevas käändes (päring a: `[lemma="laud" & features="sg_ad"]`, päring b: `[lemma="laud" & features="pl_ad"]`[4]);
2. alalütleva käände kõik esinemisjuhud, sh *laual* (päring a: `[tag="S" & features="sg_ad"]`, päring b: [tag="S" & features="pl_ad"][5]);
3. *laud* kaassõnaga *peal* (päring: `[lemma="laud"][lemma="peal" & tag="K"]`);
4. kõik nimisõnad kaassõnaga *peal*, sh *laua peal* (päring: `[tag="S"] [lemma="peal" & tag="K"]`).

Ülejäänud lahtrite väärtused on leitud vastavalt arvutusloogikale, mis on lahti kirjutatud lihtsa kollekseemanalüüsi alaosas.

Tabel 4. Distinktiivne kollekseemanalüüs: nimisõna *laud* alalütlevas käändes ja kaassõnaga *peal*

	laud	**Teised nimisõnad**	**Kokku**
alalütlev kääne	54 357	44 741 555	44 795 912
kaassõna *peal*	8 243	490 315	498 558
Kokku	62 600	45 231 870	45 294 470

Kui selline sageduste arvutamise protseduur on korratud kõigi leksikaalsete üksustega, mis esinevad uuritavas korpuses konstruktsioonides *c* ja *d*, saab kollekseemid järjestada seose tugevuse järgi ja analüüsida tulemusi uurimisküsimusest lähtuvalt. Leksikaalseid üksusi, mis esinevad oodatust oluliselt sagedamini ühes või teises konstruktsioonis, nimetatakse selle konstruktsiooni **distinktiivseteks kollekseemideks**. Ühes konstruktsioonis oodatust oluliselt sagedamini esinevad leksikaalsed üksused esinevad automaatselt teises konstruktsioonis oodatust oluliselt vähem sagedamini.

Ka allpool esitatud distinktiivne kollekseemanalüüs on läbi viidud juhindudes Stefan Griesi R-i koodiridadest (Gries 2024). Analüüsi sisendandmestiku ridadel on vaja jällegi iga leksikaalse üksuse kohta esitada kahes tulbas kaks sagedust – nimisõna sagedus alalütlevas käändes (nt lahtri „alalütlev kääne + *laud*“ väärtus) ja nimisõna sagedus kaassõnaga *peal* (nt lahtri „kaassõna *peal* + *laud*“ väärtus). Näidisuurimuses andmed on võetud eesti keele ühendkorpusest (ÜK 2021, suurus

[4] Sama tulemuse saab kätte ka ühe päringuga, kasutades atribuudi *features* asemel atribuuti *case*: `[lemma="laud" & case="ad"]`.

[5] Sama tulemuse saab kätte ka ühe päringuga, kasutades atribuudi *features* asemel atribuuti *case*: `[tag="S" & case="ad"]`.

2,4 miljardit sõna). Nimisõnade valimisel võeti aluseks 1000 sagedasemat nimisõna, mis kummaski konstruktsioonis esinevad. Seejärel valiti ainult need nimisõnad, mis olid mõlemas loendis esindatud. Kokku koosneb andmestik 305-st nimisõnast. Praktilistel kaalutlustel uurime nimisõnu vaid ainsuse vormis.

Sammud andmestiku koostamiseks:

1. Konkordantsiotsing ÜK 2021 korpusest Sketch Engine'i kasutajaliidese kaudu, et leida kõik alalütleva käände kasutused ainsuslike nimisõnadega: `[lemma="LEMMA" & tag="S" & features="sg_ad"];`
2. Konkordantsiotsing ÜK 2021 korpusest Sketch Engine'i kasutajaliidese kaudu, et leida kõigi nimisõnade kasutused kaassõnaga *peal* ainsuse vormis: `[lemma="LEMMA" & tag="S" & features="sg_g"] [lemma="peal" & tag="K"]`.

Mõlema korpuspäringu puhul tuleb "LEMMA" asemel sisestada konkreetne otsitav nimisõna, nt "laud".

Sammud distinktiivse kollekseemanalüüsi tegemiseks tarkvaraprogrammiga R:

- Käivita RStudio, kopeeri järgmine rida programmi konsooliaknasse (*Console*) ja vajuta *Enter*-klahvi:
  `source("https://www.stgries.info/teaching/groningen/coll.analysis.r")`
- Programmi käivitamiseks kirjuta konsooliaknasse järgmine rida ja vajuta *Enter*-klahvi:
  `coll.analysis()`
- Järgmiseks trükitakse RStudio konsooliaknasse laaditud programmijupi kirjeldus ning palutakse jätkamiseks vajutada *Enter*-klahvi. Distinktiivse kollekseemanalüüsi tegemiseks tee edasi järgmised valikud:
  - analüüs, mida soovid teha: 2 (*(multiple) distinctive collocate/collexeme analysis*)
  - distinktiivsete kategooriate arv: 1 (kahe võrreldava konstruktsiooni puhul)
  - sisendi vorming: 2 (*Frequency table*)
  - kas soovid Fisheri-Yatesi täpse testi tulemusi: *yes*
- laadi üles õpiku repositooriumist leitav sisendfail: *dist_collex.txt* (näidisfail, mis on sisendiks distinktiivse kollekseemanalüüsi tegemiseks nimisõnade kasutuse kohta alaütlevas käändes ja kaassõnaga *peal*)

Programm viib nende sätete põhjal läbi analüüsi ning kirjutab väljundi eraldi faili (nt nimega *dist_collex_tulemused.csv*).

Distinktiivse kollekseemanalüüsi tulemused on kokkuvõtlikult esitatud tabelis 5. Esimeses veerus on välja toodud 20 statistiliselt kõige olulisemat kollekseemi, mis esinevad ainsusliku nimisõna positsioonis alalütleva käände konstruktsioonis ja kaassõna *peal* konstruktsioonis. Sulgudes on toodud absoluutsagedus (nimisõna

esinemissagedus alalütlevas käändes, millele järgneb pärast koolonit nimisõna esinemissagedus kaassõnaga *peal*). Teine veerg näitab seose tugevust konstruktsiooni ja lekseemi vahel. Seos, mis on tugevam kui 1,3 näitab, et nimisõna ja konstruktsiooni vahel on seos, mis on statistiliselt oluline (p <0,05) (Hilpert 2006: 245; vt ka Gries, Hampe & Schönefeld 2005). Mida tugevamalt on konstruktsioon ja nimisõna seotud, seda suurem on seose tugevuse näitaja.

Tabel 5. 20 statistiliselt kõige olulisemat kollekseemi, mis esinevad ainsusliku nimisõna positsioonis alalütleva käände konstruktsioonis ja kaassõna *peal* konstruktsioonis

Alalütlev kääne		**Kaassõna *peal***	
Kollekseem	**Seose tugevus**	**Kollekseem**	**Seose tugevus**
1. aasta (755761:162)	6,01	1. pea (1219:2638)	-3,45
2. aeg (402729:94)	5,81	2. hoov (916:1802)	-3,35
3. nädal (87664:47)	4,89	3. kõht (1101:2101)	-3,32
4. tase (57600:51)	4,38	4. linn (4578:7843)	-3,24
5. kuju (40771:37)	4,35	5. nurk (2953:4249)	-3,05
6. viis (53508:52)	4,28	6. raha (910:1237)	-2,98
7. aadress (38380:43)	4,13	7. voodi (1169:1313)	-2,79
8. kord (60807:87)	3,90	8. aken (1732:1868)	-2,75
9. teema (60558:100)	3,75	9. maamuna (881:922)	-2,72
10. päev (167541:307)	3,68	10. kivi (1256:1296)	-2,71
11. müük (19807:35)	3,67	11. õu (3312:3214)	-2,65
12. territoorium (20594:37)	3,66	12. selg (2145:1810)	-2,51
13. koduleht (28872:52)	3,66	13. põhi (791:670)	-2,51
14. tegemine (20052:39)	3,58	14. uks (3981:2999)	-2,40
15. öö (17037:53)	3,11	15. muru (2326:1691)	-2,36
16. alus (141093:474)	3,06	16. maa (42446:27938)	-2,36
17. riik (13774:45)	3,06	17. pulk (824:577)	-2,32
18. võistlus (14021:46)	3,05	18. jää (3827:2435)	-2,23

Alalütlev kääne		Kaassõna *peal*	
Kollekseem	**Seose tugevus**	**Kollekseem**	**Seose tugevus**
19. sait (10644:41)	2,89	19. liiv (1127:710)	-2,21
20. taust (19476:87)	2,75	20. kere (1102:666)	-2,17

Tabeli 5 põhjal saame kontrollida oma hüpoteesi. Meeldetuletuseks – meid huvitas, millised nimisõnad esinevad sagedasti alalütlevas käändes ja millised kaassõnaga *peal*. Oletasime, et nimisõnad, mis esinevad alalütlevas käändes, viitavad sagedamini abstraktsetele ning suurtele ja liikumatutele objektidele ning kaassõna *peal* esineb sagedamini koos väiksemate ja liikuvate või liigutatavate objektidega.

Tabelist 5 jääb silma, et alalütleva käände kollekseemide hulka kuulub rida abstraktseid nimisõnu, nt aja väljendamisega seotud nimisõnad *aasta, aeg, nädal, kord, päev, öö, võistlus*; nimisõnad, mis alalütlevas käändes on adverbilise tähendusega, nt *tase* (*tasemel*), *kuju* (*kujul*), *viis* (*viisil*), *alus* (*alusel*), *taust* (*taustal*); muud abstraktsema tähendusega nimisõnad, nt *aadress, teema, koduleht, riik, sait*. Omaette rühma moodustavad nimisõnad, mis on seotud tegevustega, nt *müük, tegemine.* Kaassõna *peal* kollekseemide hulka aga kuuluvad tabeli 5 järgi mitmed väiksemad ja liikuvad või liigutatavad objektid, nt *raha, voodi, aken, kivi, uks, pulk.* Teised nimisõnad, mis väljendavad kohta, on näiteks *hoov, nurk, maamuna, õu, põhi, muru, maa, jää, liiv*. Silma jääb, et tegemist on üldiselt väga lühikeste sõnadega, v.a *maamuna*. Spetsiifilise väljendi koos kaassõnaga *peal* moodustab nimisõna *linn* (*linna peal*). Omaette rühma moodustavad kaassõna *peal* kollekseemide seas kehaosad, nt *pea, kõht, selg, kere*.

Lisaks semantikale tasub tähele panna, et mõlema konstruktsiooniga sagedasti koosesinevad kollekseemid erinevad ka oma vormi poolest. 305 nimisõnast, mida uurimuses vaatlesime, liigitas kollekseemanalüüs 158 kokku alalütleva käändega; nendest 20 nimisõna on liitsõnad (nt *koduleht, maa-ala, töökoht*). Kaassõnaga *peal* esinevatest kollekseemidest (kokku 147) on vaid 4 liitsõnad (*maamuna, aknalaud, maakera, töölaud*). Veel saab välja arvutada, kui suur on keskmine tähemärkide ja silpide arv nimisõnadel, mis vastavalt alalütlevas käändes ja kaassõnaga *peal* koos esinevad. Keskmiselt on nimisõnad, mis esinevad alalütlevas käändes, 5,6 tähemärki pikad ja kahesilbilised (silbid on loetud nimisõna ainsuse nimetavas käändes). Seevastu kaassõnaga *peal* esinevad nimisõnad on keskmiselt 4,6 tähemärki pikad ja ühesilbilised.

Kokkuvõtvalt võime distinktiivse kollekseemanalüüsi põhjal järeldada, et nimisõnad, mis sagedasti esinevad koos uuritava kahe konstruktsiooniga, erinevad laias laastus oma semantikalt ja vormilt, mis viitab sellele, et need kaks konstruktsiooni – nimisõna alalütlevas käändes ja nimisõna koos kaassõnaga

peal – erinevad oma tähenduse poolest. Alalütlevat käänet kasutame ajaliste ja abstraktsemate nimisõnadega (sh adverbilises tähenduses), kaassõna *peal* konkreetselt kohatähenduste ja väiksemate objektidega (sh kehaosadega). Lisaks sellele kipuvad alalütlevas käändes esinevad nimisõnad olema pikemad (sh liitsõnad) kui kaassõnaga *peal* koos esinevad nimisõnad. Distinktiivse kollekseemanalüüsi tulemused kinnitavad varasemate korpusuuringute tulemusi (Klavan 2012; Klavan 2024), kus liitsõnalisus ja sõnade pikkus olid kaks paljudest tunnustest, mis kohakäände ja kaassõnalise konstruktsiooni paralleelset kasutust mõjutasid.

Kokkuvõte

Näidisuurimuses tutvusime kollostruktuurilise analüüsi kasutusvõimalustega grammatika uurimiseks. Eestikeelse näidisandmestiku varal kirjeldati kahe kollostruktuurilise meetodi rakendamist – lihtne kollekseemanalüüs ja distinktiivne kollekseemanalüüs. Siinsest peatükist jäi välja koosvarieerumise kollekseemanalüüs. Loodetavasti ei tekkinud lugejal aga tunnet, et oma uurimuses tulebki rakendada mõlemat või kõiki kolme kollostruktuurilist meetodit korraga. Meetodi valik tuleb alati teha vastavalt uurimisküsimusele. Suur hulk varasemaid uurimusi on näidanud, et kollostruktuurilisi meetodeid saab edukalt rakendada, otsimaks vastuseid uurimisküsimustele, mille fookuses on lingvistilised konstruktsioonid ja keelestruktuur laiemalt.

Kollostruktuuriline lähenemine on tugevalt seotud kasutuspõhise keeleteadusega, spetsiifilisemalt konstruktsioonigrammatikaga. Sellise uurimuse keskmes on grammatilised konstruktsioonid, nt eesti keele alalütlev kääne või nimisõna esinemine koos kaassõnaga *peal*. Kollostruktuuriline lähenemine võimaldab uurida, millised sõnad esinevad grammatiliste konstruktsioonide kindlates positsioonides, aidates seeläbi jõuda keeleteadlasel parema ja täpsema kirjelduseni selle konstruktsiooni olemusest. Kollostruktuurilist analüüsi on edukalt kasutatud näiteks õppijakeele uurimiseks. Gries ja Wulff (2009) kombineerisid kollostruktuurilise analüüsi tulemusi keeleteaduslike katsetega ja uurisid, kas saksa keelt emakeelena kõnelevad inglise keele õppijad talletavad inglise keele verbikonstruktsioonide mustreid (*She tried rocking the baby vs. She tried to rock the baby*) eraldi konstruktsioonidena. Gilquin (2015) ja Deshors (2017) kasutasid kollostruktuurilist analüüsi inglise keele fraasiverbide kasutuse uurimisel. Koosvarieerumise kollekseemanalüüsi põhjal tuleb väja, et süntaktilises konstruktsioonis verb-partikkel-objekt on eri verbid ja partiklid omavahel erineval määral seotud. Kollostruktuurilise analüüsi tulemused keeleõppe uurimustes on rakendatavad ka praktikas, olles sisendiks teaduspõhiste keeleõppe materjalide väljatöötamisel. Sellised materjalid (sh keeleõppe sõnaraamatud) põhinevad tegelikul keelekasutusel ja võtavad arvesse keeleõppijate emakeelest tingitud eripärasid.

Lisaks keeleõppele on kollostruktuuriline analüüs aidanud vastata keeleajaloolistele uurimisküsimustele. Hilpert (2012) uuris COHA (The Corpus of Historical

American English) korpuse põhjal inglise keele nimisõnalist konstruktsiooni *many a NOUN* (nt *Many a day will pass before this construction is properly understood*) semantilist ja stilistilist muutust viimase 200 aasta jooksul. Hilpert (2012) mainib, et kollostruktuuriline analüüs aitab uurijal välja selgitada, millised uued leksikaalsed üksused või tüübid on uuritavas konstruktsioonis kasutusele tulnud, võimaldades uurijal seeläbi jälgida grammatiseerumisprotsessi, mille käigus konstruktsioon omandab abstraktsema, grammatilisema tähenduse. Gyselinck (2018) rakendas kollostruktuurilist analüüsi sünkroonilise ja diakroonilise varieeruvuse uurimiseks, et välja selgitada konstruktsioonis rohkem või vähem eelistatud kollokatsioonilised mustrid, kirjeldada, kuidas need ajas on muutunud, ja arutleda selle üle, mida me saame keeleteadlastena leksikaalsete üksuste vahelise seose tugevuse põhjal väita nende kinnistumise kohta.

Kuigi suur enamus kollostruktuurilise analüüsi uurimusi on tehtud inglise keele andmete põhjal, siis leiab näiteid ka teiste keelte kohta. Eesti keeles on lisaks kohakäänete ja kaassõnade paralleelsele kasutusele (Klavan 2012) kollostruktuurilist analüüsi kasutatud ka näiteks eesti murrete verbikonstruktsioonide tuvastamiseks (Uiboaed 2013). Bernolet ja Colleman (2016) on distinktiivset kollekseemanalüüsi kasutanud hollandi keele daativi alternatsiooni uurimiseks. Nad jõuavad järelduseni, et üks ja sama verb võib erinevas tähenduses eelistada erinevat daativi konstruktsiooni. Sarnaselt paljude teiste uurijatega on Bernolet ja Colleman (2016) kollostruktuurilist analüüsi rakendanud koos teiste korpuslingvistiliste ja katseliste meetoditega. On huvitav tõdeda, et kollostruktuurilist analüüsi on pigem kasutatud verbikonstruktsioonide uurimiseks, nt modaalverbid, kausatiivid, ditransitiivne konstruktsioon, passiiv jne. Märgatavalt vähem on kollostruktuurilist analüüsi rakendatud käändsõnakonstruktsioonide uurimiseks. Üheks meetodi edasiarenduseks on kollostruktuurilise analüüsi rakendamine keele idiomaatilisuse uurimiseks (Flach 2021) ja distinktiivse koosvarieerumise kollekseemanalüüsi välja töötamine, mis sobib just fraseoloogia uurimiseks (Stefanowitsch & Flach 2020).

Kui peatükki lugedes tekkis tunne, et kollostruktuuriline lähenemine on aeganõudev ja proovilepanev, siis nii see tõepoolest on – näidisuurimuses kasutatud korpusvalimi koostamine ja andmete töötlemine võtab kaua aega, kuna eeldab käsitööd. Kui aga andmestik on juba kokku pandud, läheb edasine statistiline andmeanalüüs võrdlemisi lihtsalt tänu Stefan Griesi välja töötatud R-i koodidele ja põhjalikele juhistele. Lisaks mahukale tööle andmestikega on selle meetodi puhul räägitud vahel ka muudest puudustest. Teoreetilisel tasandil võib üheks kollostruktuurilise analüüsi miinuseks pidada seda, et analüüs ei võta arvesse homonüümiat ja polüseemiat. Polüseemia mittearvestamine on näha ka käesolevas näidisuurimuses, kus näiteks distinktiivse kollekseemanalüüsi puhul ei ole selge, kas nimisõna *koduleht* koosesinemine alalütleva käändega väljendab kohta (nt *Saadet saab järelkuulata raadiojaama kodulehel*) või omajat (*Raadiojaama kodulehel on ilus disain*). Homonüümia probleem tuleb alalütleva käände uurimisel samuti esile. Näiteks tuleks puhta andmestiku ettevalmistamiseks leida keeletehnoloogiline lahendus,

kuidas eristada, millal on vorm *teel* kasutatud koha tähenduses (nt *Bussi teel on palju peatusi*) ja millal on seda vormi kasutatud omaja väljendamiseks (nt *Bussi teel on imelik maitse*, st bussis pakutaval joogil). Andmeteadusliku vajakajäämise probleemistik on laiem ja selle kohta leiab palju huvitavat lugemist Schmid & Küchenhoff (2013). Lühidalt öeldes taandub diskussioon erinevate seose tugevuse mõõdikute plussidele ja miinustele ning *p*-väärtuse rakendatavusele korpuslingvistilistes uurimustes. Kollostruktuurilist analüüsi kimbutavad samad probleemid, mis kaasnevad traditsiooniliste seose tugevuse mõõdikute rakendamisega kollokatsioonide uurimisele.

Mainitud miinustest hoolimata on kollostruktuurilise meetodi rakendamine grammatika uurimisel andnud sisukaid ja huvitavaid tulemusi. Näidisuurimuses tutvustatud meetodid pakuvad keelemustrite leidmiseks ja analüüsimiseks struktureeritud kvantitatiivse lähenemise, võimaldades siduda omavahel grammatika ja tähenduse uurimise. Kollostruktuurlise analüüsi pinnalt võib kooruda uusi uurimisküsimusi ja hüpoteese, mille paikapidavuse kontrollimiseks saab rakendada juba keerulisemaid korpuslingvistilisi või katselisi meetodeid. Ükskõik millise meetodi keeleteadlane valib, on oluline enda jaoks selgeks mõelda meetodi head ja vead ning püsida järelduste tegemisel meetodi seatud piirides.

Kirjandus

Bernolet, Sarah & Timothy Colleman. 2016. Sense-based and lexeme-based alternation biases in the Dutch dative alternation. Jiyoung Yoon & Stefan Th. Gries (toim), *Corpus-based Approaches to Construction Grammar* (Constructional Approaches to Language 19), 165–198. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/cal.19.07ber.

Brezina, Vaclav. 2018. *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Deshors, Sandra C. 2017. Zooming in on verbs in the progressive: A collostructional and correspondence analysis approach. *Journal of English Linguistics* 45(3). 260–290. https://doi.org/10.1177/0075424217717589.

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

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64(3). 501–538. https://doi.org/10.2307/414531.

Firth, John R. 1968. A synopsis of linguistic theory, 1930–55. Frank R. Palmer (toim), *Papers in Linguistics 1952–59*, 168–205. Bloomington / London: Indiana University Press. https://doi.org/10.1111/j.1473-4192.2007.00164.x.

Flach, Susanne. 2021. Beyond modal idioms and modal harmony: A corpus-based analysis of gradient idiomaticity in *mod + adv* collocations. *English Language & Linguistics* 25(4). 743–765. https://doi.org/10.1017/S1360674320000301.

Gilquin, Gaëtanelle. 2013. Making sense of collostructional analysis: On the interplay between verb senses and constructions. *Constructions and Frames* 5(2). 119–142. https://doi.org/10.1075/cf.5.2.01gil.

Gilquin, Gaëtanelle. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11(1). 51–88. https://doi.org/10.1515/cllt-2014-0005.

Goldberg, Adele. 2005. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure* (Cognitive Theory of Language and Culture Series). Chicago: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/C/bo3683810.html.

Gries, Stefan Th. 2019. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3). 385–412. https://doi.org/10.1075/ijcl.00011.gri.

Gries, Stefan Th. 2021. *Statistics for linguistics with R: A practical introduction*. Berlin / Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110718256.

Gries, Stefan Th. 2024. Coll.analysis 4.1. A script for R to compute perform collostructional analyses. https://www.stgries.info/teaching/groningen/index.html.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676. https://doi.org/10.1515/cogl.2005.16.4.635.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on „alternations". *International Journal of Corpus Linguistics* 9(1). 97–129. https://doi.org/10.1075/ijcl.9.1.06gri.

Gries, Stefan Th. & Stefanie Wulff. 2009. Psycholinguistic and corpus-linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7(1). 163–186. https://doi.org/10.1075/arcl.7.07gri.

Gyselinck, Emmeline. 2018. *The role of expressivity and productivity in (re)shaping the constructional network: A corpus-based study into synchronic and diachronic variation in the intensifying fake reflexive resultative construction in 19th to 21st Century Dutch*. Ghent University. http://hdl.handle.net/1854/LU-8574274.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10(2–3). 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2). 243–256. https://doi.org/10.1515/CLLT.2006.012.

Hilpert, Martin. 2012. Diachronic collostructional analysis meets the noun phrase: Studying many a noun in COHA. Terttu Nevalainen & Elizabeth Closs Traugott (toim), *The Oxford Handbook of the History of English*, 233–244. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199922765.013.0022.

Hilpert, Martin. 2014. Collostructional analysis: Measuring associations between constructions and lexical elements. Dylan Glynn & Justyna A. Robinson (toim), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 391–404. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.15hil.

Klavan, Jane. 2012. *Evidence in linguistics: Corpus-linguistic and experimental methods for studying grammatical synonymy* (Dissertationes Linguisticae Universitatis Tartuensis 15). Tartu: University of Tartu Press.

Klavan, Jane. 2024. *The making and breaking of classification models in linguistics: A multimethod perspective on constructional alternations*. Berlin / Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110668469.

Koppel, Kristina & Jelena Kallas. 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 18. 207–228. https://doi.org/10.5128/ERYa18.12.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/z.195.

Metslang, Helle, Mati Erelt, Külli Habicht, Tiit Hennoste, Reet Kasik, Pire Teras, Annika Viht, jt. 2023. *Eesti grammatika*. (Toim) Ellen Niit, Reet Kasik, Külli Habicht, Helle Metslang & Andriela Rääbis. Tartu: Tartu Ülikooli Kirjastus. https://doi.org/10.12697/EG

Penjam, Pille. 2009. Mis on konstruktsioonigrammatika? *Oma Keel* 19. 5–12.

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Schmid, Hans-Jörg & Helmut Küchenhoff. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577. https://doi.org/10.1515/cog-2013-0018.

Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2(1). 61–77. https://doi.org/10.1515/CLLT.2006.003.

Stefanowitsch, Anatol. 2013. Collostructional analysis. Thomas Hoffmann & Graeme Trousdale (toim), *The Oxford Handbook of Construction Grammar*, 290–306. Oxford / New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.013.0016.

Stefanowitsch, Anatol. 2014. Collostructional analysis: A case study of the English *into*-causative. Thomas Herbst, Hans-Jörg Schmid & Susen Faulhaber (toim), *Constructions Collocations Patterns*, 217–238. Berlin / New York: De Gruyter Mouton. https://doi.org/10.1515/9783110356854.217.

Stefanowitsch, Anatol. 2020. *Corpus linguistics: A guide to the methodology* (Textbooks in Language Sciences 7). Berlin: Language Science Press. https://doi.org/10.5281/ZENODO.3735821.

Stefanowitsch, Anatol & Susanne Flach. 2020. Too big to fail but big enough to pay for their mistakes: A collostructional analysis of the patterns [*too* ADJ *to* V] and [ADJ *enough to* V]. Gloria Corpas Pastor & Jean-Pierre Colson (toim), *Computational Phraseology*, 247–272. John Benjamins Publishing Company. https://doi.org/10.1075/ivitra.24.13ste.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. https://doi.org/10.1075/ijcl.8.2.03ste.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43. https://doi.org/10.1515/cllt.2005.1.1.1.

Stefanowitsch, Anatol & Stefan Th. Gries. 2009. Corpora and grammar. Anke Lüdeling & Merja Kytö (toim), *Corpus Lingiustics: An International Handbook* (Handbooks of Linguistics and Communication Science 29), kd 2, 933–952. De Gruyter Mouton. https://doi.org/10.1515/9783110213881.2.933.

Uiboaed, Kristel. 2013. Kollostruktsioonilised meetodid ja konstruktsioonilise varieerumise tuvastamine. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 4(1). 185–204. https://doi.org/10.12697/jeful.2013.4.1.11.

Vainik, Ene. 1995. *Eesti keele väliskohakäänete semantika: kognitiivse grammatika vaatenurgast* (Ars grammatica). Tallinn: Eesti TA Eesti Keele Instituut.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York: Routledge. https://doi.org/10.4324/9781315165547.

Korpuspõhine semantika: käitumisprofiilide analüüs

Ann Veismann

Lühikokkuvõte

Siin peatükis saame teada, mis on käitumisprofiilide analüüs ja kuidas seda rakendada sünonüümia ja polüseemia uurimisele. Sellega seoses vaatame üle, mida ja kuidas saab semantikas korpuspõhiselt uurida ja kuidas see on seotud onomasioloogilise ja semasioloogilise vaatepunktiga tähendusele; milliseid küsimusi leksikaalses semantikas esitatakse ja milliseid vaatlusi saab semantika kohta korpusest teha. Täpsemalt vaatame, kuidas andmeid semantilise uurimuse jaoks kodeerida ja mida seejuures silmas pidada. Näidisuurimuses võtame vaatluse alla sõna *jooksma* polüseemia ja selle, kuidas aglomeratiivse hierarhilise klasteranalüüsi abil *jooksma* tähendusrühmi saab võrrelda ja kirjeldada.

1. Sissejuhatus

Võib öelda, et semantika uurimisobjekt, tähendus, on keeleteaduses kõige raskemini empiiriliselt tabatav, sest asub nii keelekõneleja meeles kui ka kõnelejate vahelises ühises teadmiste ruumis. Seetõttu on semantikas olnud suhteliselt levinud ka teoreetilised ja introspektiivsed tööd. Tähenduste uurimisele saab läheneda kahest vaatepunktist – onomasioloogiliselt ja semasioloogiliselt (Geeraerts 2010). **Onomasioloogilisest** vaatepunktist lähtudes saab küsida, kuidas mingeid nähtusi maailmas nimetatakse. Ühte ja sama objekti või nähtust saab nimetada mitmeti, nii on kõige tavalisem onomasioloogiline uurimus sünonüümide võrdlemine ja nende tähenduserinevuste kirjeldamine (täielikku sünonüümiat leidub väga harva, enamasti on kahe erineva nimetuse olemasolule võimalik leida mingi selgitus, nt sõnu *pang* ja *ämber* eristab murdeline päritolu, st keeleväline tegur, kuid tegemist võib olla ka keelesisese eristusega, nt *ilus* ja *kaunis*). **Semasioloogilisest** vaatepunktist lähtudes saab küsida, mida üks või teine sõna (või ka konstruktsioon) tähendab, lähtekoht on mingi keeleline vorm, millele saab tähendusi omistada (nt

mida tähendab *kulgema* või *jooksma*). Sõna mitmetähenduslikkuse ehk polüseemia uurimine lähtub semasioloogilisest vaatepunktist.

Tähendustest saab uurida nii üksikute sõnade tähendusi (leksikaalne semantika) kui ka grammatilisi tähendusi (nt käänete funktsioone). Tähenduste selgitamiseks on võimalik teha mitmesuguseid katseid (vt Klavan, Veismann & Jürine 2013), kuid kui oleme nõus, et tähendus selgub eelkõige kasutuses, siis saab just korpusest andmeid sõnade mitmekesise kasutuse kohta.

Korpusandmete põhjal saab tähendusi selgitada kõige üldisemalt öeldes **konteksti** abil – sõna tähendus selgub sellest, millises kontekstis teda kasutatakse. Konteksti saab korpuses analüüsida formaalsete vahenditega (automaatselt) või uurija poolt käsitsi määratud omaduste abil. Selle õpiku näidisuurimus kollostruktsioonilisest analüüsist (J. Padriku näidisuurimus) tugineb korpusandmete automaatsele analüüsile. Automaatset korpusanalüüsi kasutavad näiteks veel distributiivse semantika mudelid (nt vektorsemantika). Need põhinevad arusaamal, et sarnane kontekstiline jaotusmuster keelekorpuses võrdub sarnase tähendusega. Esimestena sõnastasid **distributiivse semantika** alusmõtteid Zellig Harris ja John R. Firth (Harris 1954; Firth 1968), lihtsustatult võib öelda, et sõna iseloomustab see, milliste kaaslastega koos ta esineb. Teisiti öeldes, sõnade jaotumuslik ehk distributsiooniline info peegeldab nende semantilist või funktsioonilist infot. Koosesinemissageduste erinevust saab esitada (ja mõõta) kaugusena (ja sarnasust lähedusena) (vt nt Bolognesi 2020: 77; Geeraerts jt 2023).

Mitmesugused semantikauurimused nõuavad aga tihti (tähenduste kohta) andmeid, mida korpusest automaatselt ei saa. Sel juhul tuleb andmeid (pool)käsitsi märgendada ehk **kodeerida** ehk annoteerida. Korpuste märgendamisest oli juttu peatükis 3, siin näidusuurimuses kasutame termineid *kodeerima* või *annoteerima* tegevuse kohta, kus korpusest võetud lausetele lisatakse käsitsi infot, mida korpuses ei ole (ja millega neid korpusesse tagasi ei panda). Tegelikkuses muidugi kasutatakse käitumisprofiili analüüsis ka andmeid, mis võivad olla saadud varasema automaatse märgendamise teel (nt morfoloogiline info). Võib rääkida ka sellest, kas tegemist on nähtavate vormimustritega (nt kollokatsioonid, käändelõpud) või kodeeritud kasutusjoontega (ingl *feature analysis*) (Glynn 2014a: 307). Järgnevalt on ühe näite põhjal juttu sellest, mida andmete käsitsi kodeerimise juures silmas pidada.

Kognitiivses semantikas võeti 2000ndatel kasutusele meetod, mida nimetati **käitumisprofiili analüüsiks** (ingl *behavioural profile analysis*, vt nt Divjak & Gries 2006; Gries & Divjak 2009). See on tüüpiline näide mahukast süntaktiliste, morfoloogiliste ja semantiliste joonte kodeerimisest, et seejärel kindlaks teha mingite keelendite tähenduslikke omadusi. Esialgu keskendus meetod eelkõige sünonüümia selgitamisele (nt Dagmar Divjak ja Stefan Th. Gries 2006 analüüsisid 'üritamist' ja 'püüdmist' tähendavate vene keele verbide sünonüümiat), laienedes seejärel ka polüseemia uurimisse (nt Gries 2006 inglise *run* 'jooksma' analüüs, Berez & Gries 2008 inglise *get* 'saama' analüüs). Kui sünonüümia uurimustes võrreldakse

uuritavate sõnade esinemiskontekste, et leida sõnadevahelisi erinevusi, siis polüseemia uurimuste eesmärgiks on välja selgitada sõna tähenduste jaotus – milliseid tähendusrühmi saab moodustada ja millised neist on üksteisele tähenduslikult lähedasemad.

Kuna andmete käsitsi kodeerimine on väga ajamahukas töö, siis üha enam püütakse seda asendada automaatse märgendamisega. Nii on ka tänapäeval sünonüümia ja polüseemia uurimises hakanud käitumisprofiili analüüsi üha enam asendama suuremate andmehulkade analüüsimist võimaldavad automaatanalüüsi meetodid (distributiivsel semantikal põhinevad mudelid, nagu word2vec või Bert (Mikolov jt 2013; Devlin jt 2019)). Siiski on meetod ise, kuigi seda ei nimetata enam alati käitumisprofiili analüüsiks, jäänud semantikauurimuste paratamatuks osaks, kuna kõiki semantilisi jooni, mida semantikauurija mingi kindla uuritava nähtuse kontekstis oluliseks peab, ei hakka ükski korpus kunagi (või lähiajal) pakkuma. Semantiliste ja muude uurijale huvipakkuvate joonte käsitsi kodeerimine on sagedasti vajalik ka morfoloogia ja süntaksi uurimustes, mis lähtuvad funktsionaalse keeleteaduse teooriatest ja mis uurivad grammatiliste nähtuste seost nende tähenduse või funktsiooniga. Nagu öeldud, ei saa korpus kunagi pakkuda keeleandmete täielikku märgendatust tunnuste osas, mis uurimuse seisukohalt olulised võivad olla.

Järgneva näidisuurimuse põhjal selgitame klassikalise käitumisprofiili analüüsi põhimõtteid. Seda võib pidada üldisemalt näiteks keeleliste tunnuste analüüsiks, mida saab rakendada mitmesugustele küsimustele vastamiseks, sh nt varieerumise analüüsiks (vt ka L. Lindströmi ja M.-L. Pilviku näidisuurimust). Eesti keele tähenduste kohta on käitumisprofiili analüüsi kasutanud näiteks Jane Padrik *peal* ja adessiivi sünonüümia selgitamiseks (Klavan 2012), Kristiina Kask sõna *seisma* polüseemia uurimiseks (Kask 2014) ja Mariann Proos sõna *nägema* polüseemia uurimiseks (Proos 2016).

Näitlikus uurimuses soovime kasutuspõhiselt kindlaks teha väga polüseemse sõna *jooksma* tähendusrühmade sarnasusi. Enne kui asume käitumisprofiili analüüsi kirjeldama, peame korraks süvenema polüseemia probleemi. Kuni 20. sajandi lõpuveerandini kirjeldati sõnade tähendusi eelkõige tarvilike ja piisavate tunnuste kogumina. Näiteks võiks sõna *hiir* kirjeldada tunnustega *+loom*, *+väike*, *+näriline*, *+pikk saba*, *-talveuni*. Seejuures pole oluline, et *hiir* võib tähendada ka arvutihiirt, millega need tunnused ei klapi. Neid kahte *hiire* tähendust käsitletaks seega homonüümidena, millel ei ole omavahel midagi ühist (kuigi võib arvata, et arvutihiir on saanud oma nime hiire kui loomaga võrdlusest, välise sarnasuse alusel).

Kui hiire puhul oleks selline käsitlus veel mõistetav ja vastuvõetav, siis keerulisem on sõnade puhul, millel on palju tähendusi ja mille kasutused varieeruvad vastavalt kontekstile kas suuremal või vähemal määral. Tuntud on filosoof Ludvig Wittgensteini näide (Wittgenstein 2005, originaalis 1953 *Philosophical Investigations*), et nii lihtsale sõnale nagu *mäng* on keeruline anda üht kindlapiirilist tähendust. Pigem saab erinevaid *mänge* kirjeldada kui perekonda, mille liikmetel on

erineval arvul sarnasusi, st tähendust võiks kirjeldada nn perekondliku sarnasuse alusel. 1970ndatest alguse saanud kognitiivne lingvistika sai inspiratsiooni samal perioodil psühholoogias tehtud kategoriseerimise uuringutest (Rosch & Mervis 1975; Rosch 1978), mis näitasid, et kategooriad (mõisted) järgivad pigem prototüübi-struktuuri, kus keskel on üks tüüpilisem liige (millel on palju sellele mõistele vastavaid tunnuseid) ja äärealadel vähemtüüpilised, millel tüüpilisi tunnuseid vähem. Näiteks kui mõelda lindude kategooria peale, siis uurimused on näidanud, et angloameerika kultuuriruumis vastab tüüpiliselt linnule punarind ja äärealadel paiknevad jaanalind ja pingviin.

Kui elusloodus es võime leida kindlapiirilisi kategooriaid (nagu lind), siis enamasti on kategooriad hägusate piiridega, st üleminek ühest kategooriast teise on tavaline. Näiteks võib kirjeldada tüüpilist taburetti, tooli ja tugitooli (Geeraerts 2010) ning mõelda, millised on nende omavahelised üleminekualad, kus objektil on mõlema kategooria omadusi (pehmus, käetoed, jalgade arv jne). Nii hakkasidki kognitiivsed keeleteadlased ka sõnade tähendustest rääkima prototüübi ja tähendusvõrgustike abil. Sõnal on üks keskne prototüüpne tähendus, millest hargnevad vähemtüüpilised, kuid erineval määral keskse tähendusega seotud tähendused. Neid seoseid saab kirjeldada erinevate tähenduse laienemise mehhanismide alusel (nt metafoor ja metonüümia).

Selline tähenduse kirjeldus aitab hästi mõista polüseemia olemust: ei ole mõeldav, et iga uue nähtuse jaoks leiutatakse keeles uus nimetus. Nähtuste kõrvutamine ja võrdlemine kergendab nende mõistmist, st tegemist pole juhusliku samakujulise nimetuse kasutamisega, nagu homonüümia korral (vrd *tee* kui jook ja *tee* kui rada). Kuigi selline sõnatähenduse käsitlus on intuitiivselt mõistetav ja teoreetiliselt hästi kirjeldatav, on tähenduste võrgustikuna analüüsimisel ka olulisi puudusi. Nimelt on keeruline empiiriliselt kindlaks teha nii prototüüpset tähendust kui ka seda, milliseid tähendusi üldse eraldiseisvateks tähendusteks pidada. Teisiti öeldes, postuleerides kontiinumi, üleminekud tähenduste vahel, on raske selles kontiinumis kindlaid piire seadma hakata.

Polüseemiat on nimetatud probleemiks, mis põhjustab keeleteadlastele suurt peavalu, kuid mille olemasolu tavaline keelekasutaja peaaegu kunagi ei märka (Cuyckens & Zawada 2001). Mittemärkamise üks põhjus on keelekasutuse kontekstuaalsus – sõna tähendust selgitab kontekst, kus teda kasutatakse. Nii on üks võimalus sõna tähenduste struktuuri selgitamiseks selle sõna kasutuskontekstide võrdlus. Just seda võimalust hakati kasutama polüseemia uurimisel, kui käitumisprofiili analüüs oli andnud sünonüümia selgitamisel häid tulemusi.

Siin vaatleme käitumisprofiili analüüsi meetodi näitena polüseemiat, kuid seda saab rakendada väga mitmesuguste semantiliste probleemide uurimisel. Näiteks Piia Taremaa on uurinud liikumisverbide semantikat, kasutades samuti sõnade kasutusjuhtude kohta suure hulga tunnuste märgendamist (Taremaa 2021). Oluline on järgnevas tähele panna eelkõige käitumisprofiili märgendamise üksikasju

ja probleemkohti. Sellele järgnev statistiline analüüs sõltub uurimisküsimusest, millele vastust tahetakse, ja võib olla väga erisugune.

Kuna sünonüümia ja polüseemia uurimustes on olnud sageli statistilise meetodina kasutusel klasteranalüüs, siis kasutame seda ka siin. Samas näiteks Glynn (2014b) on näidanud, et stabiilsema ja usaldusväärsema tulemuse annab korrespondentsanalüüs. Glynn lisab oma uurimusse sotsiolingvistilised faktorid: registri (blogi, vestlus) ja dialekti (Briti või Ameerika inglise keel), mille mõju kasutustele korrespondentsanalüüsis hästi ilmneb.

2. Andmete kogumine ja kodeerimine

Kasutusprofiili analüüs põhineb sellel, et sõna (või muu meid huvitava keelendi) kasutusjuhte kodeeritakse võimalikult paljude tunnuste alusel. Neid tunnuseid on klassikalises käitumisprofiili analüüsis nimetatud **ID-siltideks** (ingl *ID-tags*). Igal tunnusel ehk ID-sildil on tasemed, n-ö realisatsioonid (nt käändetunnuse tasemed on *nimetav*, *omastav* jne; elususe tunnuse tasemed on *elus* ja *elutu*). Kui iga kasutusjuhu kohta on tunnused kodeeritud, tekib igale sõnale n-ö **käitumisprofiil** – millises kasutuses see sõna kui sageli esineb. Kui võrdleme ühe sõna eri kasutusi, siis iseloomustab käitumisprofiil kas üksikkasutusi või kasutusrühmi (tähendusi).

Käitumisprofiili analüüsi saab jagada neljaks sammuks (Gries & Divjak 2009):

1) andmete pärimine korpusest;
2) andmete kodeerimine (käsitsi ja poolautomaatselt);
3) kontekstis koosesinemiste tabeli loomine (milline ID-sildi tase esineb kui mitu korda iga sõnaga (sünonüümia puhul) või kas iga kasutusjuhu või iga tähendusega (polüseemia puhul): tähendused toimivad teatavas mõttes sünonüümsete üksustena, nagu tegemist oleks eri sõnadega, mille erinevust ja sarnasust me määrame);
4) koosesinemiste tabeli hindamine esmalt kirjeldava statistika abil (sagedused), seejärel eksploratiivsete statistiliste meetodite rakendamine (nt klasteranalüüs). Kirjeldava statistilise analüüsi võimalustest on pikemalt juttu ka selle õpiku 6. peatükis.

Nn käitumuslike joonte kodeerimiseks tuleb esmalt võtta uuritava nähtuse (nt sõna) ja selle esinemiskonteksti kohta korpusest piisavalt suur ja esinduslik **valim**. Kontekst võiks olla vähemalt terve lause või osalause, kuid vaja võib minna ka laiemat konteksti.

Klassikalise käitumisprofiili analüüsi puhul rõhutatakse vajadust kodeerida iga lause puhul ära võimalikult suur hulk erinevaid tunnuseid, mis iseloomustavad selle lause koostisosi, sh uuritavat sõna ennast. Alles pärast esmast kõikide tunnuste analüüsi saame hakata otsustama, millised tunnused ei ole käitumisprofiilis

olulised[1]. Kodeeritavad tunnused hõlmavad enamasti võimalikult laialt keele eri tasandeid (morfoloogia, süntaks, semantika). Nagu öeldud, nimetatakse tunnuseid **ID-siltideks** ja nad on harilikult rühmitatud, hästi tavaline ongi keele tasandi järgi kategooriatesse jagamine, kuid kodeeritavaid tunnuseid saab süstematiseerida ja tasemeteks jaotada erinevalt. Lisada saab ka keelevälised tunnuseid nagu alamkorpus või register (aja- või ilukirjandus), murre vm. Jane Padrik on oma uurimustes adessiivi ja *peal*-kaassõnafraasi konstruktsioonilise sünonüümia kohta jaganud tunnused semantilisteks, süntaktilisteks ja morfosüntaktilisteks (Klavan 2012). Kask (2014) on koostanud *seisma* polüseemia analüüsimiseks rühmad *tähendus*, *agent*, *taust*, *morfoloogia*, kus on kokku 17 ID-silti, ja Proos (2016) *nägema* analüüsiks rühmad *allikas*, *semantiline*, *süntaktiline*, *morfoloogiline*, kokku samuti 17 ID-silti.

Nagu näha, võib ID-siltide jaotamine (ja ka nimetamine) olla üsna erinev, oluline on siiski, et kasutuskonteksti analüüsimisel ei piirdutaks väga vähest tüüpi tunnustega, vaid püütaks tabada nähtust mitmekülgselt. Samas tuleb arvestada uurimisküsimuse ja uuritava nähtuse eripäradega, nt liikumisverbide uurimuses on rõhk eelkõige liikumisega seotud semantilistel ja grammatilistel kategooriatel. ID-siltide koostamisel peab võtma toeks teoreetilised lähenemised, et ühtki olulist tunnust kahe silma vahele ei jääks. Käitumisprofiili analüüsi üldprintsiip on olnud võimalikult paljude tunnuste kodeerimine, näiteks on Gries (2006: 75) *run* uurimuses kodeerinud 252 ID-silti (neist 40% käsitsi), sest tema sõnul selgub alles analüüsist, mis on oluline. Siiski tuleb iga tunnust hoolega kaaluda, et mitte paisutada tunnuste arvu siltidega, mis lõppandmestikus iseloomustavad vaid üht või paari kasutusjuhtu. Näiteks jätab Gries *run* klasteranalüüsiks alles 252-st ID-sildist vaid 55 (Gries 2006: 75, 80).

Formaalseteks (süntaksi ja morfoloogia) ID-siltideks võivad olla näiteks ajavorm, kõneviis, tegumood, kõneliik, verbi vorm, transitiivsus, kääne, arv, sõnaliik, süntaktiline positsioon lauses. Semantilisteks tunnusteks on näiteks elusus, konkreetsus, loendatavus, sõna semantiline tüüp, tähendus.

Seejärel tuleb kindlaks määrata ID-siltide **tasemed**. Näiteks morfoloogiasse kuuluva sildi „kõneviis" tasemed võiks olla eesti keeles *kindel*, *tingiv*, *käskiv*, *kaudne*. ID-sildil peaks olema vähemalt kaks taset, st minimaalselt tunnuse olemasolu või puudumine (*jah*/*ei*), nt *jooksma* semantiline tunnus „sihtkoht" võib kas esineda või mitte (*Poiss jooksis kiiresti metsa* = *jah*; *Poiss jooksis kiiresti* = *ei*). Mõne ID-sildi juures on tarvis ka taset *ei kehti* või *ei saa määrata* (enamasti lühendatult NA (ingl *not applicable* või *not available*)), näiteks *jooksma* ID-silt „sihtkoha kääne", saaks taseme *ei saa määrata*, kui lauses sihtkohta üldse ei väljendata.

[1] Samas saab siiski tunnuste märgendamist teine kord kaaluda ka veidi vähem mahukana. Seda eelkõige siis, kui eelnevad uurimused ja teoreetiline kirjandus on juba näidanud teatud tunnuste suuremat asjakohasust ja teiste mitterelevantsust uurimisküsimuse jaoks.

Kui sünonüümia uurimisel on sõnad ise need, mille käitumisprofiile võrreldakse, siis polüseemia uurimisel on eraldi küsimus see, mille suhtes ID-siltide tasemete esinemissagedusi võrrelda. Põhimõtteliselt võiks olla iga lause kasutusjuht eraldi tähendus. See oleks n-ö maksimaalse eristamise tase, kus sõna iga uus kasutus on esialgu käsitletav eraldiseisva tähendusena. Sel juhul saaksime sarnaste kasutuste põhjal moodustunud rühmadele hiljem tähendused omistada, kui analüüsiksime iga rühma lauseid ja määraksime, mis neid tähenduslikult ühendab.

Teine võimalus on, et võrdlusaluseks on sõnaraamatust saadud tähendused. Sel juhul on uurimuse roll kontrollida sõnaraamatus antud tähenduste sarnasust ja erinevust, grupeerumist ID-siltide alusel – selgitada polüseemse sõna tähenduste sisemist struktuuri/süsteemi (st paiknemist tähendusvõrgustikus). Eeldus on, et lähedased tähendused käituvad kasutuses sarnaselt. Tulemusena saaksime nt sõnaraamatus esitatavaid tähendusrühmi (ja nende esitamisjärjekorda) korrigeerida. Siiski on oluline meeles pidada, et sõnaraamat pakub tähendustest leksikograafide tehtud üldistuse. Selleks, et saada täpsemalt teada, milline üldistus kujuneks kasutuskontekstide põhjal, on mõistlikum esialgu eristada pigem rohkem tähendusi. Tihti nimetatakse neid **parafraasideks** ja need kujunevad lausete analüüsimise käigus.

Näiteks Kask (2014) on kasutanud 18 parafraasi sõnale *seisma* ja Proos (2016) 12 tähendusrühma sõna *nägema* iseloomustamiseks, mõlemad on algselt toetunud sõnaraamatule ja seejärel teooria ja korpusandmete alusel oma nimekirja kohendanud. Gries (2006) kasutas inglise sõna *run* polüseemia uurimuses lauseid analüüsides 48 tähendust, mis ta oli määranud varasemate uurimuste, WordNeti[2] info ja enda intuitsiooni põhjal. Kui kasutada käitumisprofiili analüüsi polüseemia uurimiseks, siis üldiselt on mõtet tähendusi eristada pigem võimalikult peenelt, et seejärel vaadata, millised neist muude kodeeritud tunnuste alusel sarnaseks osutuvad. Lihtsamalt öeldes võiks analüüsis kasutatud semantilised rühmad olla täpsemad kui sõnaraamatu tähendusrühmad, mis on enamasti tehtud võimalikult suure üldistusastmega.

Sõnale *jooksma* annab „Eesti keele seletav sõnaraamat" EKSS[3] 5 tähendust lihtverbina, 26 tähendust 10 ühendverbi peale (*järele, kinni, kokku, läbi, maha, ringi, välja, ära, üle, ümber*) ja 19 väljendverbi, kokku 50 sõnaraamatu kirjet. EKI ühendsõnastikus[4] on tehtud suurem üldistus ja jaotatud *jooksma* tähendused neljaks, koos alatähendustega seitsmeks, kõiki ühend- ja väljendverbe eraldi käsitlemata. Kuna näidisuurimuse andmestik on suhteliselt väike, sisaldades 320 lauset, siis ei ole mõttekas tähenduseristusi väga peenelt teha. Sõnaraamatuid uurides torkab silma, et põhitähendusi ei ole väga palju, kuid sõnaga *jooksma* on mitmesuguseid ühend- ja väljendverbe. Võime oma näidisuurimuse eesmärgiks võtta välja selgitada, kas ja

[2] https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et

[3] https://arhiiv.eki.ee/dict/ekss/index.cgi?Q=jooksma&F=M

[4] https://sonaveeb.ee/search/unif/dlall/eki/jooksma/1/est

kuidas ühend- ja väljendverbid sobivad põhitähendustega. EKSS-i ja ühendsõnastiku alusel võiks koostada näiteks tabelis 1 kirjeldatud tähendusrühmad.

Tabel 1. *jooksma* näidisuurimuses kasutatud tähendusrühmad

Tähenduse kirjeldus	**Tähendus-rühm**	**Ühend- ja väljendverbide (abi)määrsõnad**
(inimeste, loomade kohta:) jalgadel kiiresti edasi liikuma, jalga jala ette tõstes *Koer jookseb õues.*	T1	
(hrl vedeliku või peeneteralise aine kohta:) voolama, valguma *Vesi jooksis kraavi.*	T2	
kiiresti või ühetasaselt liikuma, (nagu) pinda mööda libisema *Vagun jooksis rööbastelt.*	T3	
edenema, sujuma, laabuma, ladusalt toimuma; toimima, töötama *Programm jookseb masinas.*	T4	
pidevalt, pikana kulgema või toimuma, (ajas, ruumis) kaugemale, edasi liikuma *Magistraal jooksis kaugusse.*	T5	
(filmi, saate kohta:) linastuma, eetris või kinolinal olema *Mis film kinos jookseb?*	T6	
liikumise peatumine	T7	*kinni, kokku, liiva, lühisesse, tupikusse, umbe, ummikusse*
suund, siht, eesmärk	T8	*edasi, eemale, ette, järele, järgi, laiali, otsa, peale, sisse, tagasi, vastu, välja, ära, üles*
liikumistee	T9	*läbi, mööda, ringi, üle*
muu idiomaatiline kasutus	T10	*kaasa, tormi, võidu*

Näidisuurimuse mahu tõttu on välja valitud väike hulk morfoloogilisi ja süntaktilisi ID-silte, tegelikkuses saaks neid rohkem määrata. Tähendusrühma ID-sildi tasemed vastavad tabelis 1 kirjeldatud tähendustele. Verbi *jooksma* semantikas võib

lisaks pidada oluliseks, kas jooksja on elus (inimene, loom) või elutu (nt vedelik), kas tegemist on konkreetse (nt auto, inimene) või abstraktse nähtusega (nt mõte), mis tüüpi jooksjaga on tegemist. Lisaks saab määrata, kas osalauses on väljendatud jooksmise alguspunkt, jooksmise koht või trajektoor, jooksu sihtkoht või eesmärk ja see, kas on lisatud infot liikumisviisi kohta (nt *kiiresti*).

Morfosüntaktiliselt võiks pidada oluliseks seda, millise verbivormiga on tegemist (finiitvorm, *ma*-infinitiiv, *da*-infinitiiv), kas verb on kasutuses sihitu e intransitiivsena (*laps jookseb* (*õues*)) või sihilisena e transitiivsena (*jookseme 3 kilomeetrit, nina jookseb verd*), milline on verbi ajavorm (väikse andmestiku tõttu ei erista siin mineviku tüüpi), kas tegemist on liit- või ahelverbiga (jättes siiski kodeerimata, mis on teine verb (nt *hakkama, pidama*)), kas tegemist on ühend- või väljendverbiga (jättes kodeerimata, milline on (abi)määrsõna (nt *ära*, *maha*)), kas kasutatud on eitavat või jaatavat kõnet. Eesti keele verbi tähendust uurides vajab otsustamist, kas võtta uurimisandmestikku ka infinitiivid või piirduda finiitsete verbivormidega, kas uurida nii liht- kui ka liitöeldise kasutusjuhte.

Tabel 2. *jooksma* näidisuurimuse ID-sildid koos kategooriate ja tasemetega

ID-sildi kategooria	ID-silt	ID-sildi tasemed
semantika	tähendus	10 taset: T1–T10; vt täpsemalt tabelist 1
	jooksja semantiline tüüp	15 taset: aeg; andmed, info; ese; film; inimene; kehaosa; kollektiiv/institutsioon; loom; maastik; masin; protsess; teema; toru/kanal; vedelik
	jooksja elusus	2 taset: jah (elus) / ei (eluta)
	jooksja konkreetsus	2 taset: jah (konkreetne) / ei (abstraktne)
	lähtekoht	2 taset: jah / ei
	asukoht või liikumistee	2 taset: jah / ei
	sihtkoht, suund või eesmärk	2 taset: jah / ei
	liikumisviis	2 taset: jah / ei
süntaks	verbi transitiivsus	2 taset: transitiivne (sihiline) / intransitiivne (sihitu)
	liit- või ahelverb	2 taset: jah / ei
	ühend- või väljendverb	2 taset: jah / ei

ID-sildi kategooria	ID-silt	ID-sildi tasemed
morfoloogia	verbi vorm	3 taset: finiitne; *ma*-infinitiiv; *da*-infinitiiv
	aeg	2 taset: olevik; minevik
	kõneliik	2 taset: jaatav / eitav

Oma uurimust tehes tuleks kõik ID-sildid ja nende tasemed lahti kirjutada, koos näidete ja määramispõhimõtetega, sest suurt andmestikku kodeerides on väga oluline olla järjekindel. Võimaluse korral tasub kaaluda 10% andmestiku topelt-kodeerimist kellegi teise poolt, et näha, kas arusaamad siltidest ja nende tasemetest kodeerimisolukordades on ühtsed.

Kõik otsused, mida kodeerida ja milliseid tunnuse tasemeid eristada, tuleb hoolikalt dokumenteerida ja põhjendada (lähtudes kirjandusest ja uurimiseesmärgist) enne kodeerimistööle asumist. Tõsi on see, et kodeerimise käigus konkreetsemalt materjaliga tutvudes võib tulla uusi mõtteid või ka kategoriseerimisprobleeme. Sel juhul saab kodeerimisskeemi korrigeerida, kuid hoolega tehtud eeltöö ja läbimõeldud otsused aitavad vältida olukorda, kus poole töö pealt, kui hulk aega on juba kodeerimise peale kulutatud, tuleb hakata andmeid uuesti läbi vaatama ja otsuseid suurel hulgal muutma (nt tunnuseid kas välja jätma või lisama või tunnuste tasemeid muutma).

Reaalsuses tuleb tegelike andmetega kohtudes ikka välja näiteid, mida ette näha ei osanud ja mille järgi ID-silte või nende tasemeid kohandada tuleb. Eriti puudutab see semantilise kategooria ID-silte. Näiteks käesoleva uurimuse kontekstis on võimatu mõelda ette välja kõiki „jooksja tüübi" tasemeid. See tähendab, et need võib leida n-ö alt-üles (ingl *bottom-up*), materjali põhjal. Selleks tuleks kõigepealt kirja panna jooksja lemma ja seejärel koondada sarnase tähendusega lemmad tüüpidesse.

Andmestiku saamiseks tuleb korpusest pärida kõik uuritava sõna esinemised (lemmana), järjekord juhuslikustada, vajalik arv lauseid alla laadida ja arvutustabelisse kanda, iga lause eraldi real. Kui palju lauseid on piisav, sõltub uurimisküsimusest ja korpuse suurusest. Polüseemia analüüsimisel on üldiselt kasutatud vähemalt 500–1000-lauselisi valimeid. Gries (2006) analüüsis sõna *run* 'jooksma' tähenduste selgitamiseks 814 lauset, Glynn (2014b) kordas *run* uurimust 500 lausega. Kask (2014) kasutas *seisma* analüüsis 500 lauset ja Proos (2016) *nägema* uurimuses 700 lauset. Kui andmed on kantud tabelarvutusprogrammi (nt Excel), tuleb tabeli veergudele anda nimed (ID-sildid). Hea on endale samasse faili teisele lehele koostada kodeerimisjuhis, kus on kirjas, millised tasemed igal sildil võivad olla (need saab arvutustabelis ka rippmenüüks muuta) koos näidetega, millised juhud mingile tasemele vastavad. See aitab meil kodeerijana järjekindel olla.

Nr	Reference	Left	Kwic	Right	tähendusrühmad	jooksja tüüp	jooksja konkreetne/abstraktne	Jooksja elusus	Verbi transitiivsus	verbi vorm	liit- või ahelverb	ühend- või väljendverb	eitav/jaat	AEG	Lähtekoht	Sihtkoht	Asukoht/l	Kuidas?
1	Web 2023,blogspot.com	polegi nii hullu nälja, siis oled lihtsalt väsinud. </s><s> Täna ostsin ka sooja spordialuspesu, külm on väljas niisama	jooksta	. </s><s> Ma olen kunagi päris kõva koormusega spordikoolis käinud, niiet pärais hullult segast ma ka spordiga loodetavasti	T1	inimene	konkreetne	jah	intransitiivne	da-inf	ei	ei	jaatav	olevik	ei	ei	jah	jah
2	Web 2017,blogspot.com	aastat tagasi sotsiaalministeeriumi ametnik kui oma kasulastele abi otsisin ja mitte kuskilt seda ei saanud. </s><s>	jooksime	arstide, psühhiaatrite, psühholoogide vahet. </s><s> kulutasime tööpäevi ja bensiini. </s><s> psühholoog ei kuulanud meid üldse	T1	inimene	konkreetne	jah	intransitiivne	fin	ei	ei	jaatav	minevik				
4	Web 2019,kakupesa.net	, et probleem on olemas ja tudengid said tööaega juurde. </s><s> Viimasel ajal on sebimist nii palju olnud, et siin väga ringi	jooksma	ei kipu – õhtupoolikud olen hotellis istunud ja lihtsalt laiselnud (kahel õhtul tuli ka koduste kursuste	T9	inimene	konkreetne	jah	intransitiivne	ma-inf	jah	jah	eitav	olevik	ei ei	ei ei	jah jah	ei ei
5	Web 2021,blogspot.com	kuuris saab oma lumepesuga vobshe tasuta põõnata. </s><s> mis muud uudist peale selle et juhatus ML teemaga pealtnägijasse	jooksis	? </s><s> Siit vist kohe elu kaob, kui keegi oma nimel konkreetsema küsimuse esitab. </s><s> nati kallimaks läks viin poole kallimaks	T1	kollektiiv/institutsioon	konkreetne	ei	intransitiivne	fin	ei	ei	jaatav	minevik	ei	jah	ei	ei
6	Web 2013,judo.ee	ja langes välja. </s><s> Andre Seppa lennukas kommentaar võistlustele: "Võistluspäeva hommikul olin kaalust üle ja pidin	jooksma	minema. </s><s> Aga tunne oli hea ja siht oli ainult võit! </s><s> Esimesed matšid olid head ka. </s><s> Neljandas ringis olin uimane ja ei olnud	T1	inimene	konkreetne	jah	intransitiivne	ma-inf	jah	ei	jaatav	minevik	ei	ei	ei	ei
7	Web 2021,wordpress.com	umbes pool päeva suht endast väljas, isegi aru saamata, miks täpselt. </s><s> No tegi veits tigedaks küll. </s><s> Aga mis	jookseme	siis kümme. </s><s> Ongi lihtsam, saab varem õhtale. </s><s> Ja olgem ausad, ma hakkasin seda poolmaratoni natuke juba kartma ka,	T1	inimene	konkreetne	jah	transitiivne	fin	ei	ei	jaatav	olevik	ei	ei	ei	ei
8	Web 2017,sekretar.ee	, efektiivsusele ja majanduslikule kasule orienteeritud edukas mees/naine. </s><s> " Teine: "Olen vahuleaetud hobune, kes	jookseb	, kuni kukub." </s><s> Maailm rahuneks, kui piisavalt palju inimesi muudaks minajutustust, rääkides endast endale umbes nii	T1	loom	konkreetne	jah	intransitiivne	fin	ei	ei	jaatav	olevik	ei	ei	ei	ei
9	Web 2019,fastfords.ee	välja ja "tuuning" tagatulede leidmisega pole suurt edu olnud. </s><s> Pealegi ühe tagumise tule nurgast on hakanud mõra alla	jooksma	. </s><s> Toon siia siis võrlduseks pildid mõlemast autost 2005 ja siis vanem mudel sellest järeldan et tuled on sarnased kuid 1	T5	toru, kanal	konkreetne	ei	intransitiivne	ma-inf	jah	ei	jaatav	minevik	jah	jah	ei	ei

Joonis 1. Näide märgendatud andmetabelist

3. Andmete pärimine korpusest ja märgendamiseks ettevalmistamine

Võtame eesti keele ühendkorpuse 2021 (Estonian National Corpus 2021). Pärime Sketch Engine'i vahendusel korpusest *jooksma* laused (*Concordance → Advanced*: *jooksma* lemma, sõnaliik: verb; vt õpiku ptk 5.2.3 „Konkordantside koostamine grammatilise info põhjal"). Saame valida, kas soovime tulemust lausete (*Sentence*) või KWIC kujul. Viimasel juhul näidatakse meile päritud sõnale eelnevat ja järgnevat konteksti võrdselt lausepiire arvestamata. Juhuslikustame lausete järjekorra (või kasutame funktsiooni *Get a random sample*). Laadime saadud tulemuse tabelarvutusprogrammi (nt Excel). Anname tabeli tulpadele nimed (lause, ID-sildid).

Kui vajalik hulk lauseid on olemas, kategooriad, ID-sildid ja nende tasemed on paigas, saab hakata andmestikku kodeerima. Enne kodeerimist tuleb andmestik puhastada valepositiivsetest vastustest ja sellistest vastetest, mis meile huvi ei paku, näiteks kui automaatsel morfoloogilisel märgendamisel on homonüümid saanud vale lemma (nt sõna *jooksma* lihtmineviku esimese isiku vorm *jooks-i-n* on vormihomonüümiline sõna *jooma* tingiva kõneviisi oleviku esimese isiku vormiga *joo-ksi-n*) või kui lause on pooliku konteksti tõttu täisti arusaamatu (seda juhtub siiski harva, kuid vahel on konteksti mõistmiseks vajalik avada lausest pikem tulemus Sketch Engine'is ja vahel harva võib olla tarvis lause veebist üles otsida, et veel laiema konteksti põhjal tähendust mõista). Andmestiku puhastamisel võiks vältida ebasoovitavate ridade kohe ärakustutamist ning pigem tekitada andmestikku lisatulp, milles esialgu märkida ära välja jäetavad kasutusjuhud.

Oluline on tähele panna, et juba vajaliku andmestiku kokkusaamine võib teinekord olla väga mahukas töö. Näiteks võib tuua, et kui Jane Padrik võrdles adessiivi ja *peal* konstruktsioonilist sünonüümiat, siis selleks, et saada andmestikku 300 kasutusjuhtu, kus adessiiviga väljendatakse (sünonüümselt *peal*-kaassõnafraasiga) asukohta, tuli käsitsi läbi töötada 1700 korpuslauset, sest adessiivil on eesti keeles peale koha väljendamise palju teisi sagedasi funktsioone (nt omaja väljendamine) (Klavan 2012: 258).

Näidisuurimuse jaoks oleme võtnud eesti keele 2021. aasta ühendkorpusest 320 lauset (NB! selle põhjal saadud tulemusi ei saa pidada esinduslikuks) ja analüüsinud neid tabelis 1 esitatud ID-siltide ja nende tasemete alusel. Näidist märgendatud tabeli esimestest ridadest vt joonis 1.

4. Andmete analüüs

Andmete kodeerimise kontrolliks ja seejärel andmetest üldise ülevaate saamiseks on mõtet teha märgendatud andmestiku põhjal risttabeleid. Kontrollida võiks näiteks seda, ega kuskil ei ole jäänud tühja lahtrit või ega mõni ID-sildi tase pole

saanud kogu andmestikus vaid mõne üksiku märkimise (siis tuleks kaaluda sellest tasemest loobumist).

Kõige esmase ülevaate andmetest annab sageduste võrdlemine olulisemates kategooriates, nt tähendusrühmades (vt tabel 3). Näiteks saame teada, et suurem osa *jooksma* kasutustest kuulub tähendusrühma „jalgade abil kiiresti liikuma" (T1, 178 esinemist). Ühe tähendusrühma oluliselt suurem sagedus on polüseemia puhul suhteliselt tavaline. Soovitatav on vaadata ka õpiku statistikapeatükist (ptk 6) erinevaid kirjeldava statistika võimalusi.

Tabel 3. Lausete arv tähendusrühmades

Tähendusrühm	Lausete arv
T1	178
T2	25
T3	7
T4	16
T5	13
T6	4
T7	22
T8	35
T9	16
T10	4
Kokku	320

Kui kõik andmed on kontrollitud, ühtki tühja lahtrit ei esine, siis võib alustada klasteranalüüsiga.

4.1. Klasteranalüüsist veidi lähemalt

Selleks, et mõista paremini, kuidas järgnevalt kirjeldatud analüüs toimub, tuleks korraks süveneda klasteranalüüsi olemusse. **Klasteranalüüs** on eksploratiivne tehnika, mis võimaldab tuvastada mustreid suurtes ja keerulise struktuuriga andmetes. Leitud mustrite põhjal saab püstitada hüpoteese, mida teiste statistiliste meetodite abil kontrollida. Klasteranalüüsi meetodeid on mitmeid ja hierarhiline klasteranalüüs on vaid üks neist. Kõige lihtsamalt võib klastreid andmetes võrrelda

mustriga. Täiesti juhuslikult paiknevatest andmepunktidest on raske mustrit tabada, kuid kui mõnes kohas on andmepunkte rohkem, on muster tajutav.

Klasteranalüüsi aluseks on kaugused vektorruumis. Vektorid saadakse **distants- ehk kaugusmaatriksite** abil. Võtame näiteks ühe väljamõeldud andmetabeli (tabel 4), kus on kolm vaatlust/rida (V1, V2, V3) kahe tunnusega/tulbaga (T1, T2):

Tabel 4. Andmetabeli näide

	T1	T2
V1	3	5
V2	1	7
V3	4	9

Neid andmeid saame kujutada kolme vektorina kahemõõtmelises ruumis. Esimese punkti koordinaadid on [3; 5], teisel [1; 7] jne. Kui kahe- ja kolmemõõtmelist ruumi suudame ette kujutada, siis nelja ja enama mõõtme korral meie ettekujutamise võime tõrgub, kuigi matemaatiliselt on võimalikud n-mõõtmelised ruumid.

Selleks, et võrrelda vaatluste omavahelisi erinevusi (n-mõõtmelises vektorruumis), tuleb distantsmaatriksisse arvutada iga kahe vaatluse vaheline vähima erinevuse mõõt (ehk vähim kaugus vektorruumis). Kaugusi saab arvutada erinevate meetoditega, arvuliste tunnuste korral on tavaline **eukleidiline kaugus**, st kaugus, mis arvutatakse nii, nagu Pythagorase teoreemi järgi käib hüpotenuusi leidmine. Näidistabeli (tabel 4) eukleidilise meetodi distantsmaatriks (tabel 5) on selline (siin maatriksis ei esitata diagonaali, st V1 ja V1, V2 ja V2 jne vahelisi nulle):

Tabel 5. Distantsimaatriksi näide (eukleidiliste kaugustega)

	V1	V2
V2	2,828427	
V3	4,123106	3,605551

Tabelist 5 näeme, et V1 ja V2 on omavahel „lähemal“ (kaugus 2,83) kui V3 neist mõlemast (kaugus V1-st 4,12 ja V3-st 3,6). Sama näeksime visuaalselt, kui joonistaksime Tabeli 4 järgi punktid kahemõõtmelisse ruumi. Võib proovida seda teha ruudulise paberi peal: T1 annab punkti kauguse x-teljel ja T2 y-teljel. Kuid enamasti on meil tegemist väga paljude vaatluste ja suure hulga tunnustega, nii et selline visualiseerimine ei ole võimalik.

Klasterdamise meetodid saab laias laastus jagada kaheks – mitte-hierarhiline ja hierarhiline. **Mitte-hierarhilise klasterdamise** korral moodustuvad klastrid gruppidena, kus sarnasemad andmepunktid on üksteisele lähemal; **hierarhiline klasterdamine** seevastu näitab kaugussuhteid puu kujul (dendrogrammina). Kui mitte-hierarhiline klasterdus eeldab pigem arvulisi andmeid, siis hierarhilist klasterdamismeetodit saab kasutada ka segaandmete või täielikult mittearvuliste andmete korral, nagu keeleandmed enamasti on. Manning ja Schütze (1999: 500) on leidnud, et hierarhiline klasteranalüüs annab rohkem infot, samas kui mittehierarhiline klasterdus on suure andmehulga puhul mugavam.

Hierarhiline aglomeratiivne klasterdamine toimub andmete distantsmaatriksi järgi nii, et esimeses lahenduses moodustab iga vaatlus omaette üksuse, seejärel liidetakse klastriks kaks kõige lähedasemat vaatlust ja nii üha edasi, kuni „puu" saab ühe „tüve". Liitmise algoritme on erinevaid: *complete, single, average* ja *Ward,* neid võib proovida varieerida parima tulemuse saamiseks, järgnevas näidises oleme kasutanud viimast.

4.2. *jooksma* analüüs hierarhilise aglomeratiivse klasteranalüüsi abil

Viime klasteranalüüsi selles näidisuurimuses läbi rakendustarkvaraga R (R Core Team 2023). Et andmeid R-is töödelda tuleb andmetabel salvestada CSV-vormingus. Kuna meie näidistabel on üsna väike (320 rida ja 15 veergu), saame kasutada üht väga lihtsat võimalust oma andmed R-i viia – märgime need otse Exceli tabelis, kopeerime lõikepuhvrisse (*Ctrl+C*) ning loeme need otse lõikepuhvrist R-i. Kogu analüüsiks kasutatud R-i kood koos andmestikuga on leitav õpiku OSF-i repositooriumist[5]. Analüüs toetub Natalia Levshina raamatu „How to do linguistics with R" peatükile 15 (Levshina 2015).

Järgmisena tuleb koostada andmete alusel distantsmaatriks. Kõige tuttavam distantsmaatriks igapäevaelust on näiteks linnade kauguste tabel, kus linnadevahelised kaugused kilomeetrites on antud tabeli kujul nii, et iga linn esineb nii reas kui ka tulbas, tabeli ülemine ja alumine kolmnurk kordavad sama infot ja keskel moodustub diagonaal nullidest (see on iga linna kaugus iseendast). Kuidas aga saada mittearvulistest andmetest kaugusmõõte? Meie tabelis on 320 rida (lauset) ja 15 tunnust (ID-silti). Kui sooviksime võrrelda iga rida andmetes teiste ridadega, siis võiksime koostada 320 × 320 distantsmaatriksi. Kuid meie soovime võrrelda omavahel 10 tähendusrühma, seejuures, nagu sagedusandmetest nägime, ei esinda tähendusrühmi võrdne arv lauseid (ridu). Seega tuleks meil esmalt selgitada ID-siltide tasemete proportsioonid iga tähendusrühma kohta. Näiteks esimese tähendusrühma ID-siltide proportsioonid on esitatud joonisel 2.

[5] https://osf.io/xqzsf/

```
          Tähendusrühm.T1                 Tähendusrühm.T10                  Tähendusrühm.T2
              1.000000000                       0.000000000                      0.000000000
          Tähendusrühm.T3                 Tähendusrühm.T4                  Tähendusrühm.T5
              0.000000000                       0.000000000                      0.000000000
          Tähendusrühm.T6                 Tähendusrühm.T7                  Tähendusrühm.T8
              0.000000000                       0.000000000                      0.000000000
          Tähendusrühm.T9                     J_tüüp.aeg               J_tüüp.andmed, info
              0.000000000                       0.000000000                      0.000000000
               J_tüüp.ese                     J_tüüp.film                  J_tüüp.inimene
              0.000000000                       0.000000000                      0.921348315
         J_tüüp.kehaosa    J_tüüp.kollektiiv/institutsioon                    J_tüüp.loom
              0.000000000                       0.022471910                      0.056179775
         J_tüüp.maastik                     J_tüüp.masin                 J_tüüp.protsess
              0.000000000                       0.000000000                      0.000000000
          J_tüüp.teema                J_tüüp.toru, kanal                  J_tüüp.vedelik
              0.000000000                       0.000000000                      0.000000000
J_konkreetsus.abstraktne          J_konkreetsus.konkreetne                    J_elusus.ei
              0.005617978                       0.994382022                      0.016853933
         J_elusus.jah Verbi_transitiivsus.intransitiivne   Verbi_transitiivsus.transitiivne
              0.983146067                       0.893258427                      0.106741573
       Verbi.vorm.da-inf                 Verbi.vorm.fin                 Verbi.vorm.ma-inf
              0.230337079                       0.634831461                      0.134831461
   Liit..või.ahelverb.ei          Liit..või.ahelverb.jah       Ühend..või.väljendverb.ei
              0.685393258                       0.314606742                      1.000000000
Ühend..või.väljendverb.jah             Eitav.jaatav.eitav             Eitav.jaatav.jaatav
              0.000000000                       0.101123596                      0.898876404
          Aeg.minevik                      Aeg.olevik                  Lähtekoht.ei
              0.460674157                       0.539325843                      0.960674157
        Lähtekoht.jah                      Sihtkoht.ei                  Sihtkoht.jah
              0.039325843                       0.629213483                      0.370786517
   Asukoht.Liikumistee.ei          Asukoht.Liikumistee.jah                      Viis.ei
              0.764044944                       0.235955056                      0.758426966
             Viis.jah
              0.241573034
```

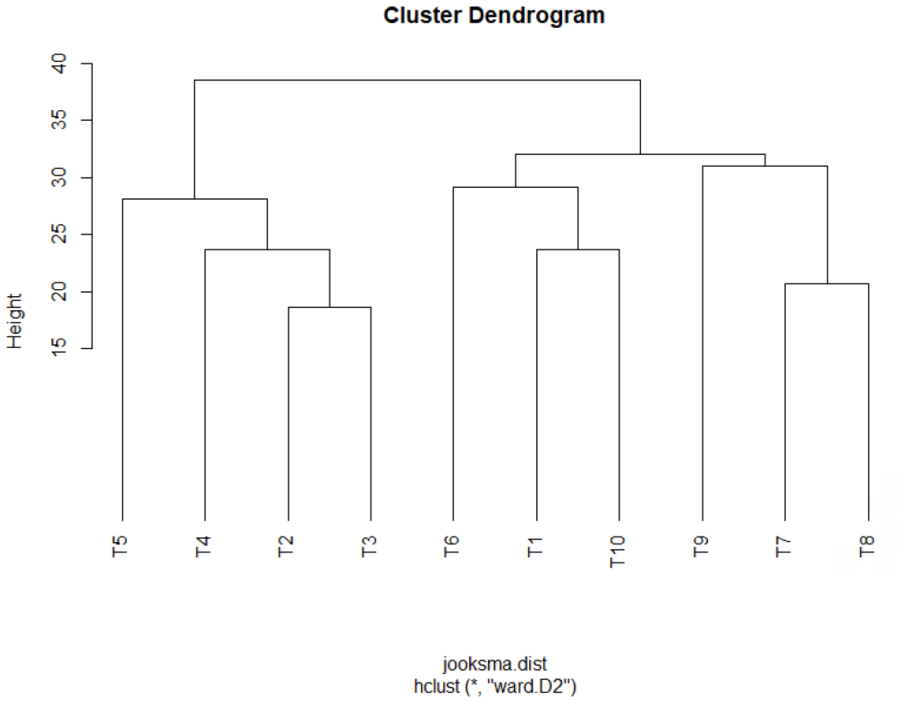
Joonis 2. Tähendusrühma T1 ID-siltide proportsioonid

ID-siltide proportsioonidega andmestikust, kus iga tähendusrühm on eraldi real, distantsmaatriksi koostamiseks on mitmeid meetodeid (eukleidiline, Manhattani, Canberra). Kuna meil on tegemist algselt mittearvuliste andmetega, siis ei ole kõik meetodid kasutatavad. Käitumisprofiili analüüsis kasutatakse sagedasti Canberra meetodit, sest käitumisprofiili vektorid võivad tihti sisaldada väga väikseid väärtusi (nullilähedased väärtused tekivad, kui tunnus esineb harva). Erinevaid meetodeid tasub tundma õppida ja ka andmete analüüsis katsetada. Canberra meetodiga koostatud distantsmaatriks *jooksma*-verbi tähendusrühmade kohta on esitatud joonisel 3.

```
        T1    T10     T2     T3     T4     T5     T6     T7     T8
T10 23.699
T2  25.048 24.080
T3  26.320 25.596 18.619
T4  26.596 26.867 22.063 22.968
T5  30.037 28.585 24.819 25.618 27.675
T6  29.374 26.308 28.983 27.827 27.081 32.533
T7  22.547 28.709 29.316 32.500 30.101 33.053 28.123
T8  25.073 30.977 28.720 23.436 26.610 30.806 28.832 20.732
T9  28.170 31.832 33.661 32.576 32.541 29.699 30.486 29.395 28.100
```

Joonis 3. *jooksma* tähendusrühmade (Canberra) distantsmaatriks

Nüüd saame distantsmaatriksi alusel teha hierarhilise aglomeratiivse klasteranalüüsi. Hierarhiline klasteranalüüs on eksploratiivne tehnika, mis leiab andmetes klastrid kodeeritud tunnuste järgi, tulemuseks on dendrogramm (vt joonis 4).



Joonis 4. *jooksma* dendrogramm

Vaatame saadud dendrogrammi lähemalt. Esmalt võib tähele panna, et laias laastus on jagunenud tähendusrühmad kaheks: T2, T3, T4 ja T5 vasakul ja T6, T1, T10, T9, T7 ja T8 paremal. Kõige esimesena on (vasakul) kokku grupeeritud T2 ja T3: 'voolamine' ja 'sujuv liikumine' (vt tähendusrühmade selgitusi tabelist 1). Järgmisel tasemel on nendega liidetud T4: 'edenema, sujuma, ladusalt toimima'. Võib näha, et nende kolme tähendusrühma ühine omadus on liikumise sujuvus. Siia lisandub T5: 'pikana kulgema, suunduma' (ka siin on tegemist katkestamata liikumisega).

Dendrogrammi paremas pooles on esimesena kokku liitunud T7 ja T8, need on ühend- ja väljendverbid – T7 rühm väljendab lõpule jõudmist: *kinni, kokku, liiva, lühisesse, tupikusse, umbe, ummikusse jooksma*, T8 väljendab suunda, sihti ja eesmärki (nt *edasi, otsa, sisse, välja, üles*). Nendega liitub ka T9: liikumisteega seotud ühendverbid (*läbi, mööda, ringi, üle*). Teise parempoolse klastri moodustavad kõigepealt T1 ja T10: 'elusolendi jalgadel liikumine' (T1) ja rühm, kus on idiomaatilised *kaasa, tormi, võidu jooksma*. Kuna viimased eeldavad elusolendi

liikumist, sobib selline grupeerumine hästi ka meie keeletajuga. Keerulisem on selgitada nendega liitunud rühma T6 ('linastuma'). Rühm on liitunud üsna kaugel ja selles on vähe lauseid (vt tabel 3). Kui meil oleks andmeid rohkem, võiks tulla teistsugused tulemused, samuti saaks siis proovida semantiliselt heterogeensemate ühendverbide rühma T9 tähendusi täpsemalt eristada.

Dendrogramm visualiseerib distantsimaatriksist saadud kaugused andmete järjest kaheks kokku ühendamise teel, kuid dendrogramm ei ütle, mitme klastrina tuleks tulemust tõlgendada, see jääb uurija otsustada andmete sisulist sarnasust hinnates. Samuti tuleb hinnata, kas dendrogramm on õnnestunud, kas jagunemised ja klasterdused sobivad kokku ka keeletajuga (vt nt Proos 2016: 38–44). Siiski on tehnikaid, mis pakuvad klastrite arvu ja aitavad klastreid valideerida (nt siluetitehnika, *bootstrapping*, vt lähemalt Levshina 2015, ptk 15). Siluetitehnika aitab leida optimaalse klastrite arvu. Silueti laius on vahemikus 0 (klastristruktuuri puudumine andmetest) kuni 1 (perfektne klastriline eristumine). Üldiselt loetakse silueti laiust, mis jääb alla 0,2, indikaatoriks sellest, et andmetes kindlat klastrilist struktuuri ei ole.

Kokkuvõte

Käitumisprofiili meetodi kõige suurem puudus on see, et töö on väga ajamahukas ja nõuab täpsust ja tähelepanu, lisaks teeb inimfaktor selle subjektiivseks. Kuna töö on mahukas, siis võib nii kirjeldatud andmehulk jääda väikseks ning seega muutub küsitavaks representatiivsus ja statistiline olulisus, tegelikult kogu töö tulemus. Näiteks näidisuurimuse andmete klastriline struktuur pole selge, kuigi tähendusrühmade kokkuliitmise järjekord on keeletaju jaoks mõttekas.

Meetodi hea küljena on välja toodud eelkõige seda, et traditsioonilist semantilist analüüsi on võimalik operatsionaliseerida ja kvantifitseerida, et seejärel hüpoteese püstitada ja neid testida. Lisaks siin esitatud klasteranalüüsile on võimalik rakendada käitumisprofiili analüüsiga saadud andmetele ka keerulisemaid statistilisi meetodeid (näiteks mitmetunnuselist analüüsi, vt ptk 6 ja L. Lindströmi ja M.-L. Pilviku keele varieerumise näidisuurimust), mis võimaldavad keele eri tasandite keerulist koostoimimist analüüsis korraga tabada. Klasteranalüüsist korpuslingvistikas võib pikemalt lugeda näiteks (Divjak & Fieller 2014; Moisl 2015; Moisl 2020).

Kirjandus

Berez, Andrea L. & Stefan Th. Gries. 2008. In defense of corpus-based methods: A behavioral profile analysis of polysemous *get* in English. *Proceedings of the 24th Northwest Linguistics Conference* 157–166. https://doi.org/10.5167/UZH-84678.

Bolognesi, Marianna. 2020. *Where words get their meaning* (Converging Evidence in Language and Communication Research (CELCR) 23). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Cuyckens, Hubert & Britta Zawada. 2001. *Polysemy in cognitive linguistics: Selected papers from the International Cognitive Linguistics Conference, Amsterdam, 1997* (Current Issues in Linguistic Theory 177). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Jill Burstein, Christy Doran & Thamar Solorio (toim), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Divjak, Dagmar & Nick Fieller. 2014. Cluster analysis: Finding structure in linguistic data. Dylan Glynn & Justyna A. Robinson (toim), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 405–441. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.16div.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2(1). 23–60. https://doi.org/10.1515/CLLT.2006.002.

Firth, John R. 1968. A synopsis of linguistic theory, 1930–55. Frank R. Palmer (toim), *Papers in Linguistics 1952–59*, 168–205. Bloomington / London: Indiana University Press. https://doi.org/10.1111/j.1473-4192.2007.00164.x.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford / New York: Oxford University Press.

Geeraerts, Dirk, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco & Michael Lang. 2023. *Lexical variation and change: A distributional semantic approach*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198890676.001.0001.

Glynn, Dylan. 2014a. Corpus methods and statistics for semantics: Techniques and tools. Dylan Glynn & Justyna A. Robinson (toim), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (Human Cognitive Processing), 307–341. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.12gly.

Glynn, Dylan. 2014b. The many uses of *run*: Corpus methods and Socio-Cognitive Semantics. Dylan Glynn & Justyna A. Robinson (toim), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 117–144. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*. *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (Trends in Linguistics Studies and Monographs 172), 57–99. Berlin / New York: Mouton de Gruyter. https://doi.org/10.1515/9783110197709.57.

Gries, Stefan Th. & Dagmar Divjak. 2009. A corpus-based approach to cognitive semantic analysis: Behavioral profiles. Vyvyan Evans & Stéphanie Pourcel (toim), *New Directions in Cognitive Linguistics* (Human Cognitive Processing), 57–75. John Benjamins Publishing Company. https://doi.org/10.1075/hcp.24.07gri.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10(2–3). 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Kask, Kristiina. 2014. *seisma*-verbi polüseemia: korpuspõhine käitumisprofiil ja klasteranalüüs. Magistritöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut, käsikiri.

Klavan, Jane. 2012. *Evidence in linguistics: Corpus-linguistic and experimental methods for studying grammatical synonymy* (Dissertationes Linguisticae Universitatis Tartuensis 15). Tartu: University of Tartu Press.

Klavan, Jane, Ann Veismann & Anni Jürine. 2013. Katselised meetodid tähenduse uurimisel. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 4(1). 17–34. https://doi.org/10.12697/jeful.2013.4.1.02.

Levshina, Natalia. 2015. *How to do linguistics with R*. Amsterdam / Philadelphia: John Benjamins Publishing Company. https://doi.org/10.1075/z.195.

Manning, Christopher & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv. https://doi.org/10.48550/arXiv.1301.3781.

Moisl, Hermann. 2015. *Cluster analysis for corpus linguistics*. Berlin / Munich / Boston: Walter de Gruyter GmbH.

Moisl, Hermann. 2020. Cluster analysis. Magali Paquot & Stefan Th. Gries (toim), *A Practical Handbook of Corpus Linguistics*, 401–434. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_18.

Proos, Mariann. 2016. Mida ütleb korpus tähenduse kohta? Käitumisprofiili analüüsi ja klasteranalüüsi meetod tajuverbi *nägema* tähenduse uurimisel. Magistritöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut, käsikiri.

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Rosch, Eleanor. 1978. Principles of categorization. Eleanor Rosch & Barbara B. Lloyd (toim), *Cognition and categorization*, 27–48. Hillsdale, NJ: Erlbaum.
Rosch, Eleanor & C. B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7. 573–605.
Taremaa, Piia. 2021. Verbs of horizontal and vertical motion: A corpus study in Estonian. *Finnish Journal of Linguistics* 34. 221–256.
Wittgenstein, Ludvig. 2005. *Filosoofilised uurimused*. Tartu: Ilmamaa.

Murded ja keele varieerumise uurimine: ainsuse 1. isikule viitamine eesti murretes

Liina Lindström, Maarja-Liisa Pilvik

Lühikokkuvõte

Juhtumiuuringus vaatleme murrete ja laiemalt varieerumise analüüsi võimalusi Eesti murrete korpuse põhjal. Tutvustame esmalt, millised on lähenemised, mis sobivad murrete korpuspõhiseks analüüsiks. Näidisuurimus käsitleb ainsuse 1. isiku verbivormide esinemist koos 1. isiku subjektpronoomeniga või ilma ning seda, millised tunnused mõjutavad subjektpronoomeni olemasolu või puudumist lauses. Peatükis näitame, kuidas andmed materjalist kätte saada ja valimit koostada, kuidas andmeid ette valmistada, puhastada ja märgendada. Rakendame sagedusele orienteeritud ühe tunnuse analüüsi ning piirangutele orienteeritud ühe tunnuse analüüsi (varieerumise analüüsi). Kasutame selleks segamõjudega logistilist regressioonanalüüsi. Andmete analüüsi viime läbi tarkvaraga R.

1. Murded ja murdekorpus

Eesti murdeuurimise traditsiooni on kuulunud murdetekstide kuuldeline kirjapanek ja alates 1960ndatest laialdane murrete salvestamine kõigepealt lintmagnetofonidega, hiljem uuemate meetoditega vastavalt tehnoloogia uuenemisele ja kättesaadavusele (kassettmagnetofonid, diktofonid, minidiskid, digitaalsed diktofonid jm-d digitaalsed salvestusvahendid). Neid materjale on pidevalt transkribeeritud, traditsiooniliselt soome-ugri foneetilises transkriptsioonisüsteemis. Materjale hoitakse peamiselt kahes arhiivis: Eesti Keele Instituudi eesti murrete ja soome-ugri keelte arhiivis (EMSUKA)[1] ja Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis (EMSA)[2]. Nende kahe kollektsiooni põhjal on alates 1998. aastast koostatud Eesti murrete korpust (EMK), mis sisaldab 1) soome-ugri foneetilises transkriptsioonis tekste; 2) lihtsustatud transkriptsioonis tekste; 3) morfoloogiliselt

[1] http://emsuka.eki.ee

[2] http://murdearhiiv.ut.ee

märgendatud tekste; 4) muid seotud materjale (andmed kõnelejate kohta jms). Murdekorpuse tekstid foneetilises transkriptsioonis leiab EMSA-st. Murdekorpuse morfoloogiliselt märgendatud osa on kasutatav avaliku otsimootori kaudu[3] ja murdekorpuse tekstid on arhiveeritud ka avalikus repositooriumis[4] (Lindström, Todesk & Pilvik 2022). Korpuse kirjelduse leiab nii repositooriumist kui ka käesoleva õpiku peatükist 2.3.2 „Suulised korpused".

Murrete uurimise eesmärk on vaadelda keele geograafilist varieerumist, näiteks kuidas häälikulised, sõnavaralised, morfoloogilised ja süntaktilised jooned on eesti keele alal levinud. Korpus võimaldab siia juurde tuua rohkem andmeid ja süsteemse lähenemise, aga ka kvantitatiivse mõõtme: korpuse põhjal võime vaadelda, mis on mingil murdealal tüüpiline (= sage) ja mis ebatüüpiline (= harv), mitte ainult seda, kas mingi joon on mingis murdes olemas või mitte (nagu on andmed esitatud „Väikeses murdesõnastikus"[5] või murdeatlastes, nt (Saareste 1955). Lisaks võimaldab murdekorpus vaadelda keelejoonte varieerumist nii ühe murdeala sees kui ka murrete vahel[6].

Murdekorpuse puhul tuleb silmas pidada seda, et tegemist on suulise keelega. See tähendab, et selles esindatud keel on spontaanne, st planeerimata, see on lineaarne ning temporaalselt piiritletud, st toimub reaalajas ja seda piiritleb kõnetempo (Auer 2009). Seetõttu sisaldab suuline tekst takerdusi, parandusi, üneeme ja muid kõne planeerimisega seotud nähtusi (Metslang jt 2023: 987 jj). Võrreldes kirjutatud keelega on kõnelejal märgatavalt vähem võimalusi oma lausungit süntaktiliselt läbi töötada ning seetõttu ei ole suulise kõne lausungid nii keerukad ja komplekssed kui kirjutatud keele laused.

Teiseks tuleb meeles pidada, et murdekorpuste maht on enamasti piiratud materjali kättesaadavuse tõttu, samuti on selliste korpuste koostamine väga ressursimahukas töö. Murdekorpus on sageli ka sisemiselt heterogeenne: murded jagunevad väiksemateks aladeks (murrakuteks) ning analüüsil tuleb alati läbi mõelda, kui suurele piirkonnale on võimalik tulemust üldistada. Kuna enamasti on murdekorpuse kasutamise eesmärk murdeid omavahel võrrelda, tuleb meeles pidada, et korpuse maht ei võimalda tavaliselt uurida keelejooni, mis esinevad suhteliselt harva, sest igast piirkonnast ei pruugi selleks olla piisavalt andmeid. Murdekorpus sobib aga hästi sagedaste nähtuste analüüsiks ja võrdlemiseks.

---

[3] https://murre.ut.ee/

[4] https://datadoi.ee/handle/33/492

[5] https://arhiiv.eki.ee/dict/vms/

[6] Õigupoolest tuleb Eesti alal rääkida kahest keelest (põhjaeesti ja lõunaeesti keel), mille murdeid analüüsime ja võrdleme. Lihtsuse huvides ei ole seda siin tehtud.

2. Lähenemised korpuspõhises dialektoloogias

Järgnevas ülevaates esitame lühikese ülevaate sellest, mis laadi uurimusi on võimalik murrete korpuse põhjal läbi viia. Võimalusi on pikemalt kirjeldatud ka mujal (Lindström & Pilvik 2018; Szmrecsanyi & Anderwald 2018).

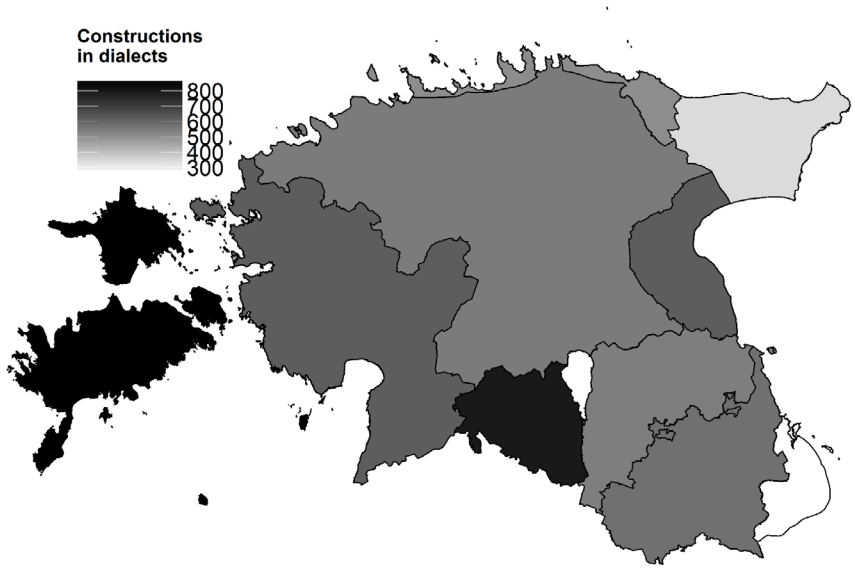
2.1. Kvalitatiivne analüüs

Murdekorpuse üks kasutusvõimalusi on üksiknäidete analüüs, eesmärgiga osutada murrete varieerumise kvalitatiivsele küljele – milliseid variante üldse murdematerjalides esineb ning mille poolest need üksteisest või normikirjakeelest erinevad, missuguseid tähendusi/funktsioone kannavad jne. Korpus sobib selleks hästi seetõttu, et see on veebi kaudu hõlpsasti kättesaadav, analüüsitud ning mitmekesiste päringuvõimalustega. Eesti murrete korpuse veebipõhine otsimootor võimaldab otsida nii sõne, lemma, sõnaliigi kui ka morfoloogilise info põhjal. Kvalitatiivse analüüsi jaoks on vaja huvipakkuvad keelendid välja otsida ja ükshaaval läbi analüüsida. Võime kvalitatiivsest analüüsist leida näiteks, millised on oleva käände (essiivi) vormid ja kasutusviisid eesti murretes (Metslang & Lindström 2017) või kuidas kasutatakse *mine*-lõpulisi teonimesid (Pilvik 2017).

2.2. Kvantitatiivne ühe joone analüüs

Kvantitatiivne ühe joone analüüs võimaldab vaadelda ühe nähtuse levikut ning selle sageduserinevusi murretes. Kvantitatiivne analüüs eeldab sageduspõhiste kvantitatiivsete meetodite rakendamist ning andmete visualiseerimist ning sellel on kaks põhilist eesmärki: andmete kirjeldamine ja andmete seletamine.

Kvantitatiivse analüüsi alla liigitub eraldi **sagedusele orienteeritud analüüs**, milles vaadeldakse eelkõige keelelise joone levikusagedust murretes. Näiteks eesti murrete põhjal on läbi viidud uurimus mineviku liitaegade (nt *on käinud, olid teinud*) esinemissageduse kohta, mis ei vaatle mitte detaile liitaegade vormistamisel, vaid liitaegade esinemissagedust üldiselt (Lindström jt 2015; Lindström jt 2019). Eriti kõrge või madala esinemissageduse puhul on põhjust küsida, kas tegemist on näiteks keelekontaktide mõjuga või mingite muude muutustega vastaval murdealal. Tulemusi on võimalik visualiseerida **sageduskaardina** (joonis 1).



Joonis 1. Liitaegade (täis- ja enneminevik) kasutussagedus murdekorpuses (Lindström jt 2019)

Sageduskaardi koostamisel tuleb arvesse võtta seda, et valitud üksuse (murde, kihelkonna) puhul võib materjali hulk, millest vaadeldavat keelejoont otsitakse, olla väga erinev ning saadud tulemused pole seetõttu võrreldavad. Selle probleemi ületamiseks kasutatakse **sageduste normaliseerimist**: absoluutsageduste asemel võrreldakse omavahel suhtelisi sagedusi, mis võtavad arvesse materjali hulka vastavast korpuseosast (vt õpiku ptk 5.2.4.1 „Sagedusloendi põhjal keele uurimine").

Tabelis 1 on esitatud joonise 1 aluseks olevad normaliseeritud sagedusandmed: iga murde kohta käiva korpuseosa maht, konstruktsiooni (liitaegade) absoluutsagedus ja konstruktsiooni normaliseeritud esinemissagedus. Normaliseerimise baasiks on kõigi murrete keskmine maht (83 431). Normaliseeritud sagedusi kujutatakse ka kaardil 1.

Tabel 1. Liitaegade (*olema* + *nud*-kesksõna konstruktsiooni) absoluut- ja normaliseeritud sagedus korpuse eri osades

Murre	Sõnesid korpuses	Konstruktsiooni absoluutsagedus	Konstruktsiooni normaliseeritud sagedus
rannamurre	51 667	316	510
idamurre	45 280	339	625
saarte murre	166 898	1723	861
keskmurre	130 086	860	552
Mulgi murre	63 516	617	810
Alutaguse murre	47 660	193	338
Seto murre	39 175	123	262
Tartu murre	65 591	428	544
läänemurre	154 400	1157	625
Võru murre	70 038	486	579

Piirangutele orienteeritud ühe joone analüüsis keskendutakse murdejoone kahe (või enama) variandi varieerumist mõjutavate piirangute ja tegurite väljaselgitamisele. Sisuliselt on tegemist sama meetodiga nagu sotsiolingvistikas kasutatud klassikaline **varieerumise analüüs**, mille algus ulatub juba 1960ndatesse ja mille alusepanijaks peetakse William Labovi. Varieerumise analüüsis uuritakse mingi keelelise üksuse ehk **muutuja** eri esinemiskujude kasutamistingimusi: millised keelevälised või keelesisesed tunnused kallutavad kõnelejaid valima üht või teist vormi (näiteks võru seesütleva käände *n*- või *h*-lõpulist vormi alal, kus mõlemad on kasutusel, vt Mets 2011). Keelelist muutujat on defineeritud kui sama asja ütlemist kahel (või enamal) erineval viisil (Labov 1972: 188). Need variandid peavad esinema samas kontekstis: näiteks samas positsioonis (*b* ~ *v* varieerumine: *kivi* ~ *kibi*) või samas funktsioonis/tähenduses (leksikaalselt: *tarvis* või *vaja*; morfoloogias: *-n, -h* või *-s* seesütleva lõpuna, nt *küläṅ* ~ *küläh* ~ *külas;* süntaksis näiteks infiniitse verbivormi valik koos verbidega *näima/paistma/tunduma*: *paistab olevat* ~ *olema*). Varieerumise analüüsis nimetatakse uuritavat nähtust niisiis **sõltuvaks muutujaks ehk uuritavaks tunnuseks**, ning analüüsi kaasatakse **sõltumatuid muutujaid ehk seletavaid tunnuseid**, mille mõju sõltuva/uuritava tunnuse variantide esinemisele üritatakse kvantitatiivse analüüsi käigus välja selgitada. Seda, millised võiksid olla sobivad seletavad tunnused, tuleb otsustada sõltuvalt uuritavast nähtusest. Murdekorpuse andmete puhul võib ühe seletava tunnusena arvesse võtta murret

(või murrakut), et välja selgitada, kui suur roll on murdest tingitud erinevustel võrreldes muude keelesiseste ja -väliste tunnustega. Põhjalikuma ülevaate varieerumise analüüsist leiab eesti keeles artiklist (Pilvik, Plado & Lindström 2021), selles näidisuurimuses esitame näite ainsuse 1. isiku asesõna kasutamisest koos samale isikule viitava verbivormiga.

2.3. Korpuspõhine dialektomeetria

Dialektomeetriliste meetoditega võrreldakse paljude üksiknähtuste alusel murrete omavahelisi keelelisi sarnasusi ning keelelisi kaugusi. See meetod kaasab korraga mitmeid erinevaid keelejooni ning võimaldab näha laiemat pilti, piltlikult öeldes puude asemel metsa (Nerbonne & Kretzschmar 2013). Dialektomeetrilist analüüsi on klassikaliselt tehtud atlasandmestike põhjal, kus võetakse arvesse mingite joonte esinemist või mitteesinemist mingil murdealal ning võrreldakse murrete lähedust ja sarnasust nii tekkivate kimpude kaudu.

Kvantitatiivne mitme joone analüüs ehk **korpuspõhine dialektomeetria** (ka *kobardialektoloogia*) hõlmab ühtaegu paljusid keelelisi jooni ning nende varieerumist murretes, ent see võtab arvesse ka nende joonte esinemissagedust murretes. Eristatakse ülalt-alla (ingl *top-down*) ja alt-üles (ingl *bottom-up*) korpuspõhist dialektomeetriat. Ülalt-alla dialektomeetria puhul on eelnevalt välja valitud analüüsi kaasatavad jooned, nende sagedused ja/või tõenäosused ning joonte põhjal arvutatud murrete lingvistilised kaugused. Alt-üles dialektomeetriline analüüs ei põhine mitte etteantud joontel, vaid korpusest mingil viisil ekstraheeritud andmetel, näiteks sõnamitmikel (bi- või trigrammidel, vt õpiku ptk 5.2.4.4 „Sõnamitmikud ehk n-grammid“) (Wolk & Szmrecsanyi 2016; Szmrecsanyi & Anderwald 2018). Heaks korpuspõhise kobardialektoloogia näiteks on Benedikt Szmrecsanyi käsitlus Briti inglise murretest (Szmrecsanyi 2013). Eestis on kobardialektoloogia meetodeid kasutanud Kristel Uiboaed eesti murrete verbiühendite uurimisel (Uiboaed 2013).

3. Uurimus 1. isiku pronoomeni kasutamisest eesti murretes

Järgnevalt vaatleme lähemalt üht näidet korpuspõhisest murdeuurimisest. Näidisuurimuseks on valitud nähtus, mis varieerub ka tänapäeva eesti keeles – ainsuse 1. isiku subjektpronoomeni olemasolu või puudumine lauses, nt *lähen* või *ma lähen*. Püstitame kaks peamist uurimisküsimust:

1) Kas pronoomeni kasutamisel koos verbiga on murrete vahel olulisi sageduserinevusi?
2) Millised keelelised ning taju ja mäluga seotud nähtused mõjutavad pronoomeni esinemist või mitteesinemist?

Tegemist on seega sagedusele orienteeritud ühe joone analüüsiga (pronoomeni olemasolu sagedus eri murretes) ning piirangutele orienteeritud ühe joone analüüsiga (varieerumist mõjutavate tegurite väljaselgitamine). Samal teemal on varem kirjutatud artiklis (Lindström jt 2009).

3.1. Valimi moodustamine

Valimi moodustamiseks otsime eesti murrete andmetest välja ainsuse 1. isiku verbivormid ja vaatame, kas nendega koos esineb asesõna või mitte. Asesõna esinemine on seega meie muutuja ehk sõltuv tunnus, ning see on kahetasandiline (sellel on kaks varianti): asesõna kas on või ei ole 1. isiku verbivormiga samas osalauses.

Uurimiseks vajamineva materjali leiame **Eesti murrete korpusest**. Selle võib välja otsida kahel erineval viisil:

1) murdekorpuse veebipõhist otsimootorit[7] kasutades või
2) repositooriumis[8] paiknevaid murdekorpuse korpusefaile kasutades.

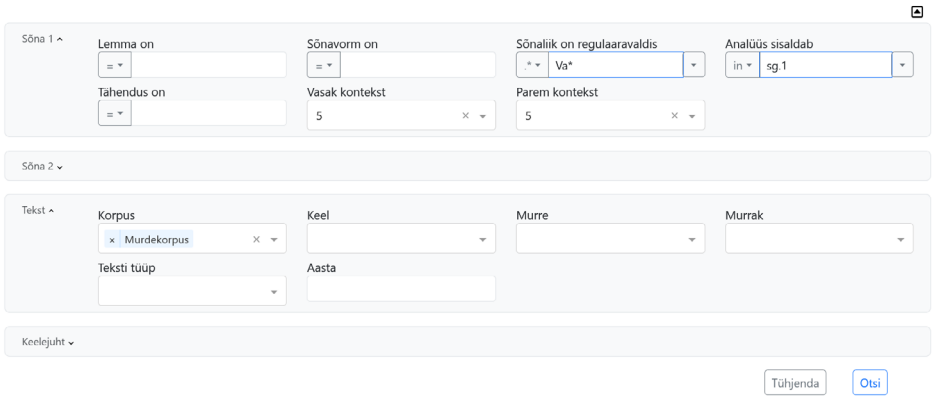
Murdekorpuse veebipõhise otsimootori kasutamine on lihtsam ja mugavam, ent see ei ole sama paindlik kui failidest vajaliku info väljaotsimine mõne skriptimiskeele abil. Vaatame esmalt veebipõhisest **otsimootorist** otsimise varianti, kuigi osa vajalikku infot otsimootoris ei kuvata (nt kõneleja isik). Samuti ei võimalda otsimootori kasutamine lisada andmestikule automaatselt tunnuseid, mille kaudu vaatlusi hiljem iseloomustada (nt kas verbivormi lõpus on pöördelõpp *-n*, kui kaugel on eelmine viide 1. isikule jms); seda kõike tuleb teha käsitsi.

Murdekorpuse kasutamiseks on esmalt vaja tutvuda selle märgendusega. Igale sõnale on lisatud lemma kirjakeelestatud kujul, sõnaliigiinfo ja morfoloogiline info. Murdekorpus on märgendatud käsitsi, mis tähendab, et andmeid tuleb vaadata kriitilise pilguga – siin võib olla erisusi ja vigu, mis tulenevad eri märgendajate tööst.

Murdekorpuses on isikulõpp märgitud jaatavate verbivormide juurde (nt märgend *pers.ind.pr.sg.1.* vastab isikulise tegumoe kindla kõneviisi oleviku ainsuse 1. isiku vormile), ent kuna eituses see enamasti puudub (nagu *ei tea, ei tunne, ei lähe*), siis eitusvormidele pole isiku kategooriat lisatud või on lisatud ainult juhul, kui isik tõesti eitusvormiga seostub (nt Tartu murdes *es annava* '(nad) ei andnud', *es anname* '(me) ei andnud'). Seetõttu on selles analüüsis vaadeldud pronoomeni koosesinemist vaid jaatavate verbivormidega. Veebipõhisest otsimootorist otsimiseks sobib päring, mis on esitatud joonisel 2. Sõnaliigi osas tasume regulaaravaldist `Va*`, mis otsib ühtaegu nii verbe (*V*) kui ka abiverbe (*Va*).

[7] https://murre.ut.ee/

[8] https://datadoi.ee/handle/33/492



Joonis 2. Ainsuse 1. isiku jaatavate vormide otsimine murdekorpuse otsimootorist

Päringu vastuseks saadud tabeli saab laadida alla ja seda mõnes tabeltöötlusprogrammis, nt Excelis edasi analüüsida.

Siinse näidisuurimuse tarvis oleme andmestiku kogunud R-i **skriptide abil** DataDOI repositooriumis paiknevatest murdekorpuse morfoloogilise märgendusega XML-failidest. Iga murde iga faili puhul otsitakse skripti abil välja kõik ainsuse 1. isiku (1SG) verbivormi sisaldavad laused (sõnaliik *V* või *Va*, vormiinfo sisaldab järjendit *sg.1*). Iga leitud lause puhul eraldatakse vastava kõneleja ja salvestuse metaandmed (ID, vanus, sünniaasta, sugu, nimi, kihelkond, murre, fail). Iga 1SG verbivormi puhul lauses võetakse välja vastava verbivormi sõne, lemma, vorm ja liik, leitakse selle vormi ees- ja tagakontekst (20 sõna kummaltki poolt, k.a pausid, poolikud sõnad ja arusaamatuks märgitud katked), leitakse vaadeldavale vormile eelnenud lähim 1SG verbivorm (ainult jaatavad vormid) ja lähim 1SG pronoomen (ükskõik mis vormis), samuti arvutatakse automaatselt nende vormide ja vaadeldava vormi vaheline kaugus sõnades. Iga 1SG verbivorm ja selle juurde kuuluv automaatselt eraldatud info pannakse päringuvastete tabelis eraldi reale. Lõpuks lisatakse automaatselt tunnus selle kohta, kas verbivormi lõpus on pöördelõpp *-n* või mitte. Kasutatud skriptid koos kommentaaridega paiknevad õpiku lisamaterjalide hulgas[9]. Eri ülesanded oleme jaganud eri skriptidesse, nii et konteksti lisamist ja viitamiskaugusi oleks võimalik põhiskriptis kasutada funktsioonidena. Nii väldime seda, et kood muutuks liiga pikaks ja kompleksseks, samuti võimaldab see koodiosade hõlpsamat taaskasutamist.

Andmestikus vastab niisiis igale 1SG verbivormi kasutusjuhule (vaatlusele) üks rida. Kuna kogu info on tabelisse lisatud automaatselt, ent nii automaatne kategoriseerimine kui ka murdekorpuse käsitsi märgendus võib sisaldada vigu, tuleks andmed esmalt üle vaadata.

[9] https://osf.io/xqzsf/

ees	sone	lemma	liik	vorm	taga	lopp	lause_id	eelmine_1sg	eelmise_1sg	eelmise_1sg	kaugus_ee	eelmine_	eelmise_1sg	eelmise_1sg	kaugus_e	hoiatus	murre	khk	fail	KJ	KJ_vanu	KJ_synnia	KJ_sugu
viimatti küsi	mõistaks	mõistma	V	pers.knd.pr.s	ega neid `pri	FALSE	l1									oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
[räägib] [tau	ol'in	olema	V	pers.ind.ipf.s	pisikkene po	TRUE	l9	mõistaks	mõistma	pers.knd.pr.s	184	ma	mina	sg.nom.	186	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
paeu `kuori (	käisin	käima	V	pers.ind.ipf.s	`mõisas karc	TRUE	l9	ol'in	olema	pers.ind.ipf.s	9	maa	mina	sg.nom.	10		Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
kodu `mõisti	lugesin	lugema	V	pers.ind.ipf.s	ära (.) mis =	TRUE	l24	käisin	käima	pers.ind.ipf.s	476	minule	mina	sg.all.	3	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
kui `jälle ei r	`ütlen	ütlema	V	pers.ind.pr.s	`varsti (...) `e	TRUE	l26	lugesin	lugema	pers.ind.ipf.s	327	minu	mina	sg.gen.	283	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
kõva `jõuga	`müisin	müüma	V	pers.ind.ipf.s	ära (.) meil c	TRUE	l76	`ütlen	ütlema	pers.ind.pr.s	816	minu	mina	sg.gen.	152	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
ei tuld (...) n	jäin	jääma	V	pers.ind.ipf.s	maha (...) ah	TRUE	l94	`müisin	müüma	pers.ind.ipf.s	384	minu	mina	sg.gen.	5		Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
siis | noo eg	öl'in	olema	V	pers.ind.ipf.s	pisikkene [n	TRUE	l100	jäin	jääma	pers.ind.ipf.s	65	minu	mina	sg.gen.	71	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
= s nied (.) n	öl'in	olema	V	pers.ind.ipf.s	pisikkene po	TRUE	l100	öl'in	olema	pers.ind.ipf.s	2	mina	mina	sg.nom.	3	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
poisikkene ja	käisin	käima	V	pers.ind.ipf.s	siis nendega	TRUE	l100	öl'in	olema	pers.ind.ipf.s	17	mina	mina	sg.nom.	18		Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
`jälle obuse	`jaksasin	jaksama	V	pers.ind.ipf.s	`käija (.) noh	TRUE	l100	käisin	käima	pers.ind.ipf.s	52	mina	mina	sg.nom.	71		Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
sie öl'i poisi	õlen	olema	V	pers.ind.pr.s	küll (...) jaa (	TRUE	l102	`jaksasin	jaksama	pers.ind.ipf.s	25	mina	mina	sg.nom.	97	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
[naerukil] or	räägin	rääkima	V	pers.ind.pr.s	selle peig+m	TRUE	l107	õlen	olema	pers.ind.pr.s	24	ma	mina	sg.nom.	5	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
(.) ja siis me	las'in	laskma	V	pers.ind.ipf.s	oma `lahti (.	TRUE	l132	räägin	rääkima	pers.ind.pr.s	684	minul	mina	sg.ad.	50	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
tia misukkes	lasen	laskma	V	pers.ind.pr.s	koa oma `lak	TRUE	l132	las'in	laskma	pers.ind.ipf.s	9	mina	mina	sg.nom.	10	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
pane ja ega	`rääkkisin	rääkima	V	pers.ind.ipf.s	sial et üks õl	TRUE	l140	lasen	laskma	pers.ind.pr.s	84	mina	mina	sg.nom.	85	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
tämа = li jäll	`ütlesin	ütlema	V	pers.ind.ipf.s	et siin onn k	TRUE	l147	`rääkkisin	rääkima	pers.ind.ipf.s	169	minule	mina	sg.all.	8	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
(...) mina `üt	`ütlen	ütlema	V	pers.ind.pr.s	et sie nüüd s	TRUE	l147	`ütlesin	ütlema	pers.ind.ipf.s	14	minu	mina	sg.gen.	4	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
arenend onn	küs'in	küsima	V	pers.ind.pr.s	et mis = teil	TRUE	l149	`ütlen	ütlema	pers.ind.pr.s	159	maa	mina	sg.nom.	148	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
= ja `jälle (.)	`ütlesin	ütlema	V	pers.ind.ipf.s	et mina põle	TRUE	l149	küs'in	küsima	pers.ind.pr.s	9	mina	mina	sg.nom.	10	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	75	1887	M
*ütles et (.)	panin	panema	V	pers.ind.ipf.s	*ninda = t (.	TRUE	l1					mina	mina	sg.nom.	33	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N
se vihk *lõig	õlin	olema	V	pers.ind.ipf.s	*ikke pali *t	TRUE	l1	panin	panema	pers.ind.ipf.s	18	mina	mina	sg.nom.	20	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N
seda+visi üle	tulin	tulema	V	pers.ind.ipf.s	sialt (.) teije	TRUE	l3	õlin	olema	pers.ind.ipf.s	79	mina	mina	sg.nom.	80		Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N
teije pagaril	*käisin	käima	V	pers.ind.ipf.s	sääld (.) *kõ	TRUE	l5	tulin	tulema	pers.ind.ipf.s	17	mina	mina	sg.nom.	26	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N
rie = pääl ja	tulin	tulema	V	pers.ind.ipf.s	*neije taga (	TRUE	l5	*käisin	käima	pers.ind.ipf.s	24	mina	mina	sg.nom.	25	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N
kuda üks *ku	*kuulin	kuulma	V	pers.ind.ipf.s	seda = ja (.)	TRUE	l5	tulin	tulema	pers.ind.ipf.s	54	mina	mina	sg.nom.	55	oli ka lähem	Alutaguse	Iis	Alutaguse_li	KJ	95	1867	N

Joonis 3. Näide skripti abil automaatselt täidetud tabelist

3.2. Andmestiku korrastamine

Püüame selles uurimuses seletada 1SG pronoomeni kasutamise varieerumist tunnuste järgi, mida on varasemates uurimustes peetud olulisteks pronoomeni kasutamist mõjutavateks teguriteks ja mille kodeerimine saaks toimuda võimalikult suures ulatuses automaatselt. Sellised tunnused on 1SG pöördelõpu esinemine, verbivormi aeg, eelmise 1SG (jaatava) verbivormi lemma, kaugus eelmisest 1SG (jaatavast) verbivormist, eelmise 1SG pronoomeni vorm, kaugus eelmisest 1SG pronoomenist, kõneleja kood, kõneleja sugu, kõneleja vanus, murre, kihelkond, faili nimi.

Lihtsustame siin pisut ülesannet ja viitamissuhete tegelikku komplekssust, vaadeldes viitamiskaugusi ainult 1SG verbivormide ja pronoomenite põhjal. Tegelikult muidugi võidakse kõnelejale viidata ka näiteks 2. isiku vormidega (eelkõige intervjueerija küsimustes). Kuna murdekorpuse tekstides ei ole intervjueerijate tekst kuidagi märgendatud, tuleks selliseid juhte otsida suuresti käsitsi. Otsingut saaks küll pisut kitsendada (nt kasutada otsimiseks nimekirja murdekorpuses märgendatud 2SG verbivormidest ja pronoomenitest), ent see ei taga kindlasti päris puhast tulemust. Samuti ei tegele me praegu selle küsimusega, kas 1SG vorm viitab sisuliselt kõnelejale või tsiteerib kõneleja mõne teise inimese teksti, kuna ka selleks tuleks tekste käsitsi üle vaadata. Kindlasti tasub põhjalikumas analüüsis pöörata tähelepanu näiteks ka lausestruktuuridele (kas verbivorm esineb pea- või kõrvallauses, kas ta on rinnastatud mõne teise 1SG verbivormiga jms), verbide semantilistele klassidele ja esinemissagedusele, ent hetkel hoiame oma analüüsi võimalikult lihtsana.

Pisut peame siiski andmeid ka praegusel juhul käsitsi korrastama. Seda võib teha mis tahes tabelarvutusprogrammis.

1) Vaatame üle **pöördelõpu** tulba (*lopp*) ja kontrollime juhtusid, kus verbivormi lõpus on kliitik *-gi/-ki*, kuna sellistel puhkudel on pöördelõpu olemasolu märgitud väärtusega *FALSE*, kuigi tegelikult võib kliitikule pöördelõpp eelneda (nt *elasingi*). Kliitikud on lemma väljale märgitud plussmärgiga (nt *elama+ki*).
2) Loome andmestikku uue tulba (nt nimega ***kommentaar***), kuhu märgime sõna *välja*, kui vormiinfo või konteksti põhjal on tegemist eituse vormiga (nt *maa=i mäletta*) või kui murdekorpuse failides on ekslikult 1. isiku vormiinfo määratud mõnele muule vormile (nt *lähäb*, *ujusime*). Samuti saame selles tulbas märkida ära, kui vormiinfo mõni muu osa on vigane (nt on oleviku asemel minevik või kindla kõneviisi asemel tingiv kõneviis).
3) Loome andmestikku tulba ***pron***, milles kodeerime uuritavat tunnust ehk 1SG pronoomeni esinemist. Kui ees- või tagakonteksti põhjal on näha, et 1SG verbivormi juurde kuulub ka 1SG subjekti pronoomen, märgime sellesse tulpa *jah*, vastasel juhul *ei*. Pronoomenite märkimist lihtsustab pisut see, et oleme andmestikku skriptiga kokku korjates märkinud tulpa *hoiatus*, kui verbist kuni 2 sõna kaugusel esineb mõni pronoomen. Kui see käib uuritava verbivormiga kokku, läheb tulpa *pron* väärtus *jah*. Kui ei käi, läheb väärtus *ei* ja sel juhul tuleb ära muuta ka info lähima eelmise 1SG

pronoomeni kohta, sest tegelikult esineb eelmine pronoomen lähemal kui automaatselt märgitud sai. Pronoomenite kodeerimisel tuleb arvestada, et pronoomen võib esineda ka verbivormi järel (nt *läksin ma*).

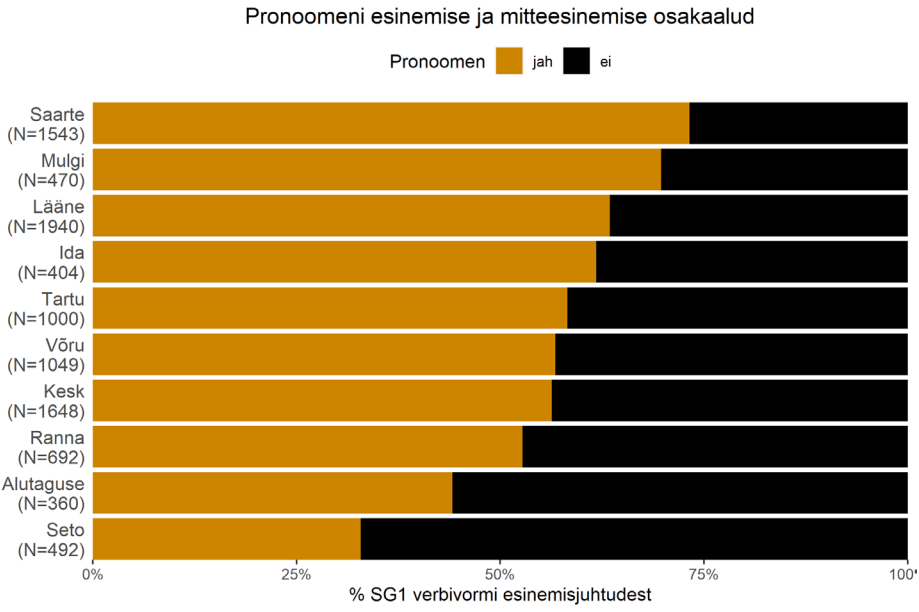
4) Vaatame üle **eelmise 1SG pronoomeni esinemisjuhud ja viitekaugused**. Kui 1SG pronoomen on vaadeldavale 1SG verbivormile lähemal kui 3 sõna, on ta arvatud verbivormi juurde kuuluvaks asesõnaks ja mitte verbivormile eelnenud asesõnaks. Selline klassifitseerimine võib aga olla kohati ebatäpne (nt lauses *mina võtsin panin* ei kuulu pronoomen *mina* vahetult verbivormi *panin* juurde, ent on sellele lähemal kui 3 sõna).

3.3. Andmete analüüs

3.3.1. Sagedusele orienteeritud ühe tunnuse analüüs

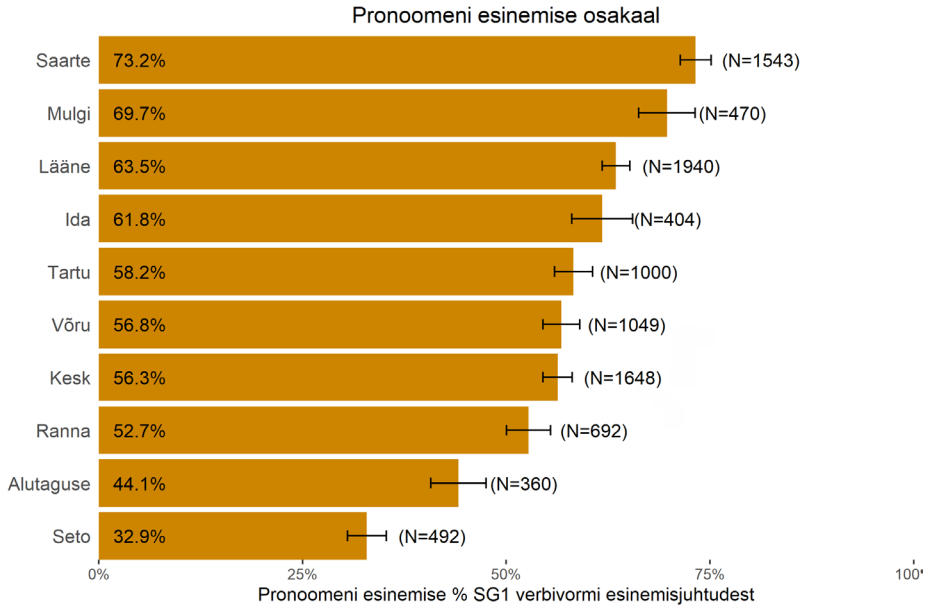
Vaatleme esmalt, kui sageli eri murretes 1SG pronoomenit verbiga koos üldse kasutatakse. Siinjuures jätame välja vaatlused, mis on pärit teistega võrreldes oluliselt hilisemast ajast (salvestatud pärast 2010. aastat).

Sageduste normaliseerimisega eelpool mainitud viisil pole siin põhjust tegeleda, kuna meid ei huvita otseselt see, kui sageli 1SG verbivormid kuskil murdes esinevad. Selle asemel vaatleme siin absoluutsageduste asemel **suhtelisi esinemissagedusi** (mitmel %-l mingi murde 1SG verbivormi kasutusjuhtudest oli kasutatud ka pronoomenit ja mitmel %-l mitte).



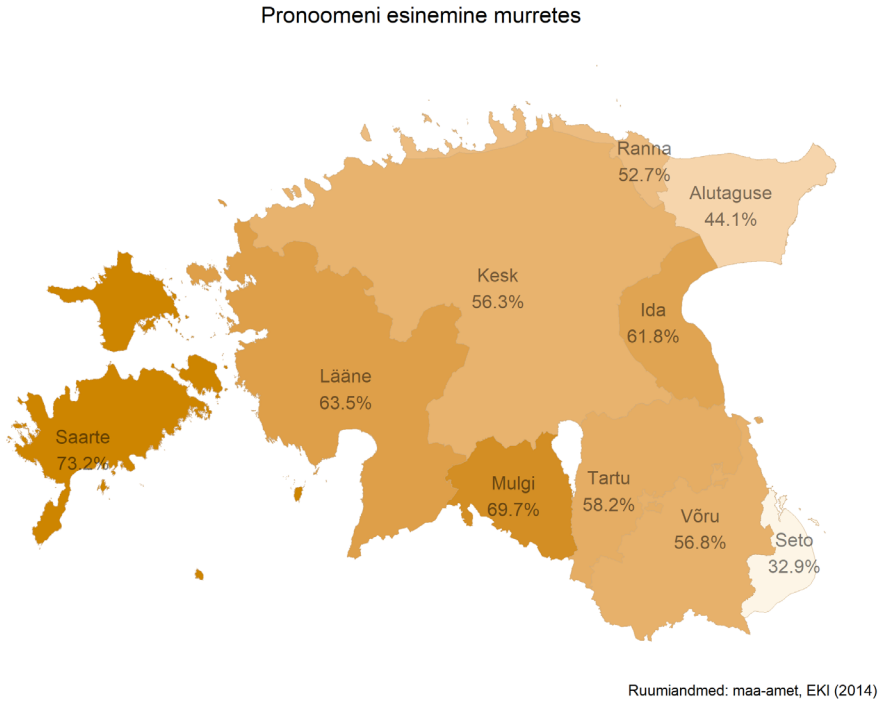
Joonis 4. 1SG verbivormide esinemine koos subjektpronoomeniga ja ilma

Võime vaadelda eraldi ka ainult pronoomeni esinemise protsente ning lisada joonisele usaldusvahemikud, mis võtavad arvesse vaatluste üldist arvu: mida vähem on kuskilt murdest vaatlusi, seda ebakindlamad võime olla arvutatud protsendi paikapidavuses selles murdes üldisemalt. Sellisel juhul võib protsentide erinevust kahe murde vahel pidada oluliseks siis, kui nende usaldusvahemikud üksteisega ei kattu. Nõnda võib näiteks saarte ja Mulgi murde vaheline erinevus tuleneda puhtalt juhusest, saarte ja läänemurde vaheline erinevus on aga statistiliselt oluline. Selgelt erinevad teistest madalama pronoomeni kasutuse poolest Seto ja Alutaguse murre, kus enamik (st üle 50%) 1SG verbivormidest esineb ilma pronoomenita.



Joonis 5. 1SG verbivormide esinemine koos subjektpronoomeniga

Kolmandaks võime suhtelisi sagedusi kuvada ka kaardi peal, et näidata paremini geograafilisi erinevusi ja üleminekuid. Selleks kasutame põhjana Eesti kihelkondade ruumiandmeid, millega ühendame oma andmestiku sagedusandmed.



Joonis 6. Pronoomeni esinemine murretes sageduskaardina

3.3.2. Piirangutele orienteeritud analüüs

Järgmiseks vaatame, mis siis ikkagi pronoomeni esinemist/mitteesinemist tingib. Siinjuures ei saa me rääkida kategoorilistest piirangutest (mingis kontekstis pronoomen *ei tohi* esineda, mingis kontekstis *peab* esinema), vaid **tõenäosuslikest eelistusmustritest** (mingis kontekstis pronoomen *pigem* esineb, mingis kontekstis *pigem* ei esine). Tunnused, mida nende eelistusmustrite analüüsil kasutame, said nimetatud juba eelmises alapeatükis. Need on ühelt poolt seotud grammatiliste, keelesiseste teguritega (nt verbi grammatiline aeg, pöördelõpu olemasolu) ning teisalt keeleväliste, mälu ja protsessimisega seotud teguritega (viitamiskaugus, 1SG viite aktiveeritus), ehkki olgu öeldud, et ka keelesisesed tegurid võivad mängida rolli mälu ja keele mentaalse töötluse juures. Lisaks on meil kasutada sotsiolingvistilised andmed kõnelejate murdetausta, soo ja vanuse kohta. Erinevaid kõnelejaid on andmestikus kokku 345.

Kasutame analüüsis tunnust „kaugus eelmisest viitest 1. isikule“ (*kaugus_eelmisest*), mille tuletame korpusest kogutud ja käsitsi parandatud tunnuste põhjal automaatselt. Määrame sellesse tulpa mistahes eelmise 1SG vormi (1SG verbi või 1SG pronoomeni) kauguse, mis vaadeldavale verbivormile parajasti lähemal

on. Kaugust mõõdame sõnades. Samuti teeme tunnuse „eelmise viite vorm" (*eelmise_vorm*), kuhu määrame väärtuse vastavalt sellele, kas kõige lähemal on 1SG verbivorm (*verb1sg*), 1SG pronoomen ainsuse nimetavas käändes (*pron1sg_nom*) või 1SG pronoomen mingis teises käändes (*pron1sg_muu*). Oletame, et kui 1. isik on mingil moel lähikontekstis aktiveeritud, võib ta hõlpsamini uuesti kordamata jääda; kui aga eelmine viide on kaugel, on tõenäolisem isikuviite pronoomeni näol võimalikult eksplitsiitseks tegemine.

Mälu ja keele mentaalse töötlusega seotult saame tuletada parandatud ja puhastatud andmestikust veel kaks lisatunnust: 1) kas korratakse lähimat eelmist 1SG verbivormi (*eelmine_verb*), 2) kas lähima eelmise 1SG verbivormiga koos esines pronoomen või mitte (*eelmine_proniga*). Nende tunnuste abil paneme referentsiaalse aktiveerituse hüpoteesiga võistlema või seda täiendama oletuse, mille kohaselt kõnelejad taaskasutavad juba kord aktiveeritud struktuure: kui aktiveeritud on pronoomeniga verbivorm, on tõenäolisem, et kõneleja taaskasutab seda uuesti; kui aktiveeritud on pronoomenita verbivorm, on tõenäolisem, et kohtame samasugust pronoomenita kasutust ka järgmise isikuviite korral. Seda nähtust on nimetatud ka **praimimiseks** (ingl *priming*, vt nt Torres Cacoullos & Travis 2014) või struktuuriliseks püsivuseks/järjekindluseks (ingl *structural persistence*, Szmrecsanyi 2005) ning see on nähtus, mida eriti suulise korpuse andmetega töötades peame arvesse võtma.

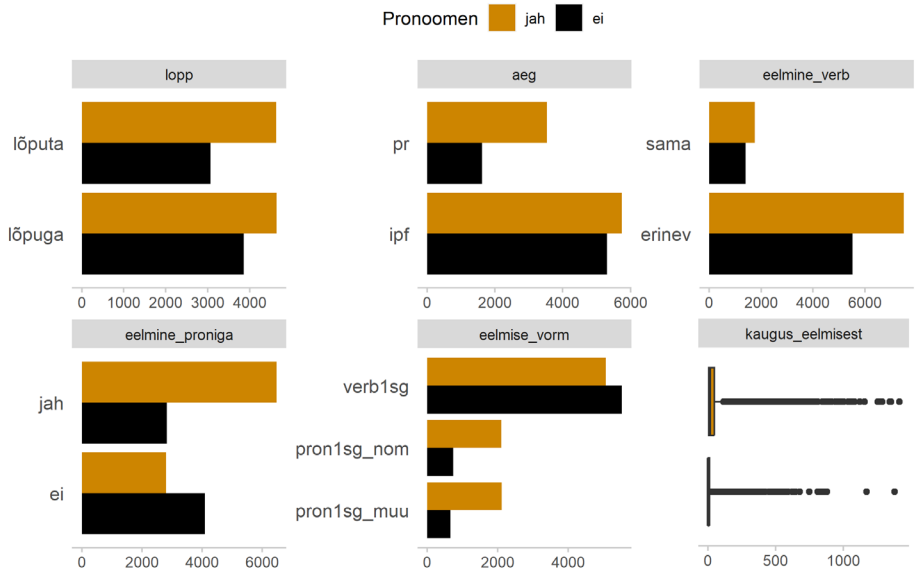
Analüüsi kaasatud tunnused on esitatud tabelis 2.

Tabel 2. Analüüsi kaasatud tunnused

Tunnus	Tunnuse kood	Väärtused	Näide
1. isiku pronoomeni esinemine	*pron*	*jah* *ei*	*miss **ma** nüüd pian tegema* *üks+kõrt *kuulen et tasane vile*
1. isiku pöördelõpu *-n* esinemine	*lopp*	*lõpuga* *lõputa*	*miss ma nüüd **pian** tegema* *nende `juures **köis'i** seal=ja*
verbivormi aeg	*aeg*	*olevik* *minevik*	*ma **lää** joosõ `tarrõ* *maq = ku mõtsa+vah'ihh **ol'l'i***
kaugus eelmisest viitest 1. isikule (sõnade arv)	*kaugus_eelmisest*	0-1045 (keskmine 38,61)	*ku mia esi <u>kirjutti</u>* *[`saatte+lehe ja][2] **saadi*** *vabrikkusse piima siit*

Tunnus	**Tunnuse kood**	**Väärtused**	**Näide**
eelmise viite vorm	*eelmise_vorm*	*verb1sg* *pron1sg_nom* *pron1sg_muu*	*kaheksa `aastane ol'in = ja kui **ma `akkasin** orjama* *mina küll ei terettänd, **`vas'tsin** aga pal'ittu vahelt* *tema vahib `mulle otsa ja **maa `lükkan***
eelmine 1. isiku verb sama	*eelmine_verb*	*sama* *erinev*	*korvi võttan `kaasa (.) noa **võttan** kaa* *ma=tsin et **nägin** küll*
eelmine 1. isiku verb pronoomeniga	*eelmine_proniga*	*jah* *ei*	*mina käin korralikkult, **tien** kalmistud alatti ülesse* *selle sügise tulin ära ja **läksin** juba `kooli*
murre	*murre*	*Alutaguse* *Ida* *Kesk* *Lääne* *Mulgi* *Ranna* *Saarte* *Seto* *Tartu* *Võru*	
kõneleja identifikaator	*KJ_id*	N = 345	
kõneleja sugu	*KJ_sugu*	*M* *N*	
kõneleja vanus	*KJ_vanus*	42–101 (keskmine 79)	
verbilekseem	*lemma*	N = 567	*ütlema, rääkima, teadma, jõudma* jne

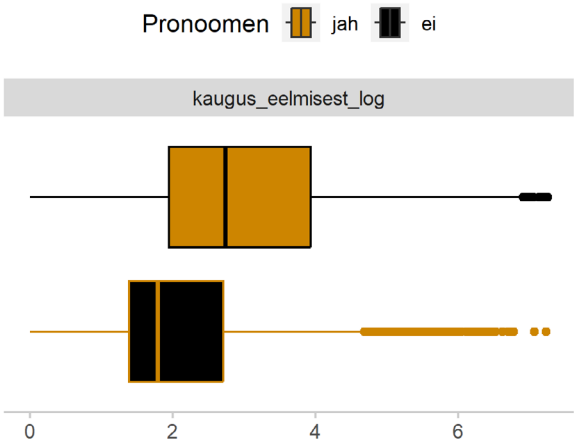
Anname esmalt ülevaate uuritava tunnuse jaotumisest lingvistiliste tunnuste põhjal (joonis 7).



Joonis 7. Pronoomeni esinemise jaotus vastavalt kodeeritud seletavatele tunnustele

Näeme jooniselt 7, et pronoomeni eksplitsiitset kasutust tundub soosivat 1) verbi olevikuvorm (*aeg* = *pr*), 2) see, kui eelmist verbi ei korrata (*eelmine_verb* = *erinev*), 3) see, kui ka viimati kasutatud verbivorm esines pronoomeniga (*eelmine_proniga* = *jah*), 4) see, kui eelmise 1SG viite jaoks ei kasutatud 1SG verbivormi, vaid hoopis pronoomenit (ükskõik mis kujul, *eelmise_vorm* = *pron1sg_nom* või *eelmise_vorm* = *pron1sg_muu*), 5) mingil määral ka see, kui verbivorm on lõputa (*lopp* = *lõputa*), ehkki nii lõputa kui ka lõpuga vormides pronoomen pigem esineb kui ei esine. Ainus arvuline seletav tunnus – kaugus lähimast eelnevast viitest 1. isikule (kas pronoomenile või verbivormile) – on aga nähtavalt väikeste väärtuste suunas kaldu: enamik viitekaugusi on väga väikesed (kuni 10 sõna kaugusel), mõned üksikud väga suured. Tõenäoliselt ei tasu ka eeldada, et suhe pronoomeni kasutustõenäosuse ja viitekauguse vahel oleks lineaarne, st et tõenäosus pronoomenit kasutada kasvaks ühtlaselt iga lisanduva sõnaga, mis kahe viiteüksuse vahele jääb. Tajume igasugu sagedusi (ja ka kaugusi) pigem logaritmiliselt, mis tähendab seda, et olulisemaks muutuvad erinevused väikeste sageduste vahel, ja vähem oluliseks erinevused suurte sageduste vahel. Näiteks tajume ilmselt tugevamalt viitesuhte aktiveerituse erinevusi kontekstides, kus ühel puhul on eelmine viide 1 sõna kaugusel, teisel puhul 10 sõna kaugusel, kui kontekstides, kus ühel puhul on eelmine viide 101, teisel puhul 110 sõna kaugusel, ehkki ühikutes on tegemist sama erinevusega. Selleks, et viitekauguse tunnust nüüd mudelis kasutada, võiksime selle tunnuse esmalt **logaritmida**. See tagab selle, et rõhutame mudelis erinevusi

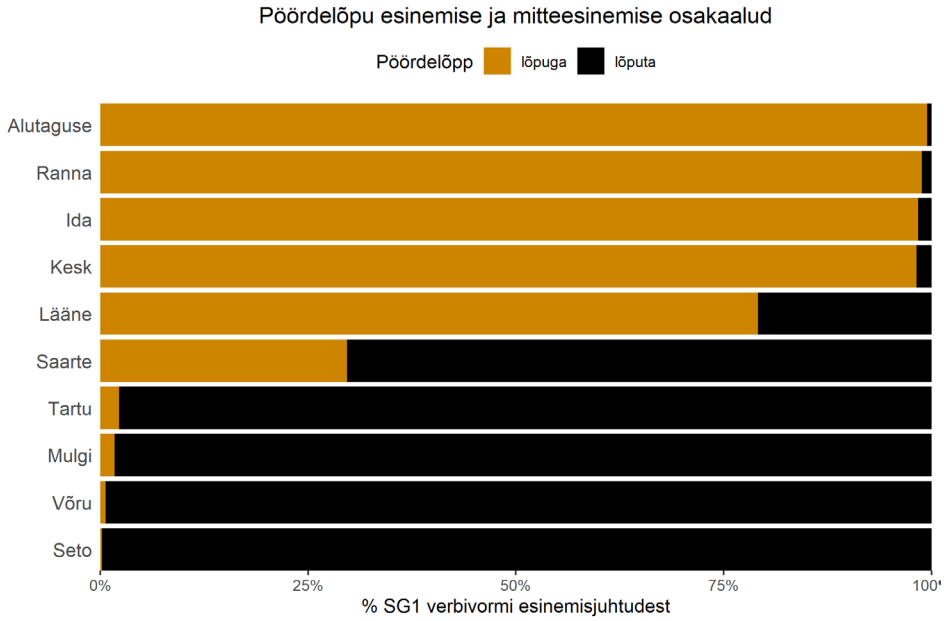
väiksemate väärtuste vahel ja taandame erinevusi suuremate väärtuste vahel (logaritmimise kohta vt lähemalt õpiku ptk 6.2.1.2).



Joonis 8. (Logaritmitud) viitekauguse jaotumine sõltuvalt uuritava tunnuse väärtustest

Logaritmitud viitekaugused jaotuvad ühtlasemalt (joonise 8 x-teljel on nüüd absoluutkauguste asemel logaritmitud väärtused), samuti näeme, et eksplitsiitse pronoomeniga verbivormid paiknevad lähimast 1SG viitest keskmiselt kaugemal kui pronoomenita verbivormid.

Jooniste põhjal saame üht-teist järeldada selle kohta, milline võib olla üksikute tunnuste seos uuritava tunnuse varieerumisega, samuti võime välja arvutada koosesinemiste osakaalud ning nende erinevust statistiliselt testida (vt ptk 6.2.1 „Ühetunnuseline seoste analüüs: statistilised testid"). Keerulisem on aga uurida seda, milline on nende tunnuste mõju siis, kui vaadelda neid korraga ja omavahelises koosmõjus. Selleks saame kasutada mitmetunnuselist varieerumise analüüsi. Enne statistilist mudeldamist võiks mõelda ka sellele, kas ja kuidas seletavad tunnused omavahel seotud võiksid olla. Murretega seoses teame näiteks seda, et murded erinevad üksteisest selle poolest, kui palju neis pöördelõppu kasutatakse: lõunaeesti murretes ainsuse 1SG pöördelõpu *-n* kasutamist peaaegu ei kohtagi, põhjaeesti murretes on see aga selgelt verbilõpuna kinnitunud. Sellel on oma keeleajaloolised põhjused: eesti keeles on *-n* sõna lõpust üldiselt kadunud, ent põhjaeesti murretes säilinud just 1. isiku pöörde lõpus, lõunaeesti murretes aga kadunud ka selles positsioonis (Lindström jt 2009). Kui nüüd pöördelõpp ilma murdeinfota mudelisse panna, läheks see teadmine kaduma. Kui aga panna mudelisse iseseisvate tunnustena korraga nii pöördelõpp kui ka murre, hakkaksid kaks tunnust seletama vähemalt osaliselt sama asja.



Joonis 9. Pöördelõpu *-n* esinemine murretes

Jooniselt 9 näeme, et pöördelõpu kasutamises esineb varieerumist sisuliselt vaid lääne- ja saarte murdes. Seega tuleks pöördelõpu puudumise või olemasolu rolli eksplitsiitse pronoomeni ennustamisel hinnata ainult koosmõjus murdealaga.

Seletame varieerumist siinses uurimuses nn **konfirmatoorse ehk kinnitava lähenemisega**, mis tähendab, et testime ühe konkreetse teoreetilise mudeli paikapidavust, mitte ei püüa andmetest pronoomeni kasutamise ennustamiseks optimaalset ja võimalikult hea seletusjõuga mudelit tuletada. Testitavas teoreetilises mudelis oletame, et pronoomeni esinemist mõjutavad

1. **verbi ajavorm**: verbi olevikuvorm tõstab pronoomeni kasutamise tõenäosust, kuna olevikus kasutatakse paljusid taju- ja suhtlusverbidega suhtlusüksusi, mis suulises suhtluses on oma sageduse tõttu küllaltki kinnistunud järjendid (nt *ma tiijä*, *ma mõtlen*, *ma arva*);
2. **murdeala ja pöördelõpu koosmõju**: saarte ja läänemurdes vähendab pöördelõpu olemasolu pronoomeni kasutamise tõenäosust, kuna isiku topeltmarkeerimine pole tingimata vajalik; mujal murretes mõju tõenäoliselt puudub, kuna pöördelõppu kas märgitakse peaaegu alati või peaaegu mitte kunagi;
3. **eelmise 1SG viite vorm koosmõjus viitamiskaugusega**: kui 1. isiku referent on kõige viimati aktiveeritud mitte verbivormi ega nominatiivis oleva 1SG pronoomeni kaudu, vaid mõnes muus käändes oleva pronoomeni

vormiga (nt *mul*), suurendab see pronoomeni kasutamise tõenäosust võrreldes teiste viitamise vormidega. Mida kaugemale eelmine viide aga jääb, seda väiksemaks jääb eelmise viite vormi mõju ning pronoomeni kasutamise tõenäosus kasvab lihtsalt koos kasvava viitamiskaugusega;

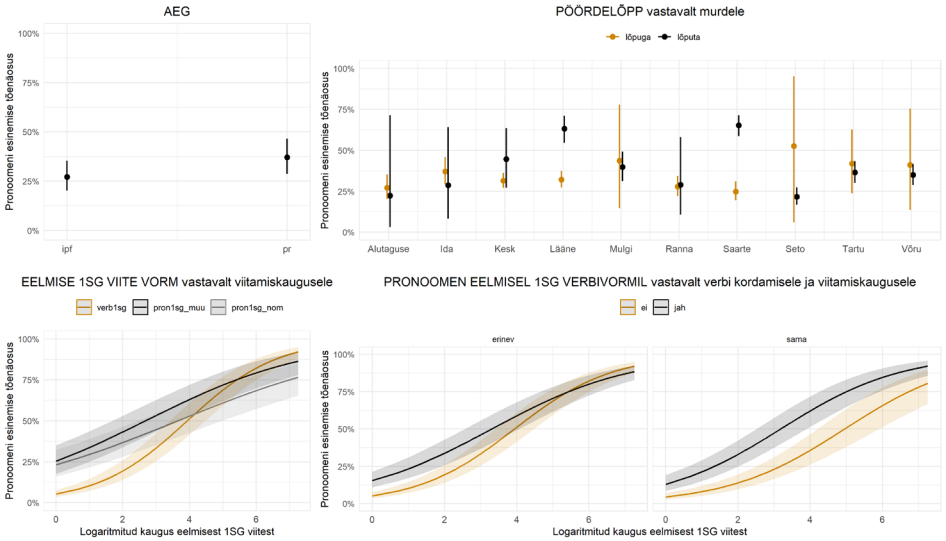
4. **pronoomeni olemasolu eelmise 1SG verbi juures koosmõjus eelmise 1SG verbi kordamise ja viitamiskaugusega**: kui eelmine 1. isiku verbivorm esines koos pronoomeniga, võiks lähtuvalt aktiveeritud mustrite taaskasutamise hüpoteesist olla pronoomeni uuesti kasutamise tõenäosus suurem. Oletame, et see tõenäosus suureneb veelgi, kui ka verb ise on sama ehk korratakse täpselt sama struktuuri. Efekt aga jällegi kaob oletatavasti viitamiskauguse kasvades.

Kasutame segamõjudega logistilist regressiooni (ingl *mixed-effects logistic regression*), mille jaoks on R-is funktsioon *glmer()* paketist `lme4` (Bates jt 2015). Mudeli formaalne esitus näeb välja selline:

```
pron ~ lopp * murre + aeg + eelmise_vorm * kaugus_
eelmisest_log + eelmine_verb * eelmine_proniga *
kaugus_eelmisest_log + (1|KJ_id) + (1|lemma)
```

Meetodiga on võimalik hinnata paljude tegurite kombineeritud mõju (nn **fikseeritud mõjusid**) uuritava tunnuse väärtuse valikule, võttes arvesse ka seda, et vaatlused võivad kuidagi grupeeruda või klasterduda. Näiteks meie andmestikus võib üks kõneleja panustada andmestikku mitu verbivormi ning kuna iga kõneleja võib oma pronoomeni kasutamise eelistuses keskmisest pisut erineda, tuleks sellega ka statistilises mudelis arvestada, selleks et saada testitavatele fikseeritud teguritele vähem kallutatud hinnangud. Samuti võiksime mudelis arvesse võtta seda, et iga verbilekseem võib käituda pisut erinevalt: mõnd verbi kasutatakse sagedamini pronoomeniga koos, mõnd sagedamini ilma. Selliseid vaatlusi klasterdavaid, üldjuhul väga paljude võimalike tasemetega tegureid nimetatakse segamudelites **juhuslikeks mõjudeks**. Valitud meetod seab siiski ka piirangud sellele, kui komplekssed teoreetilist mudelit testida saame. Põhimõtteliselt võiksime näiteks testida ka kõikide mälu ja protsessimisega seotud tegurite omavahelisi koosmõjusid või nende koosmõjusid lingvistiliste teguritega. Kõigeks selleks aga vajaksime oluliselt rohkem või teisiti jaotunud andmeid, kuna segamudeli algoritm ei suuda olemasolevate andmete põhjal nii komplekssse mudeli puhul uuritavatele parameetritele adekvaatseid hinnanguid pakkuda.

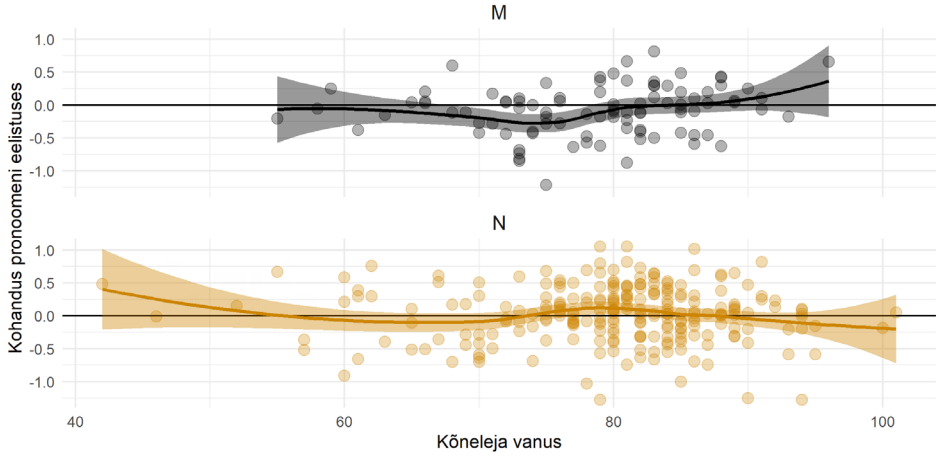
Joonis 10 visualiseerib regressioonimudeli fikseeritud mõjusid nii, et y-teljel on mudeli ennustatud pronoomeni kasutamise tõenäosus mingile vastavate tunnustega vaatlusele (hoides ülejäänud seletavate tunnuste väärtused konstantsed). Näeme, et kaks püstitatud hüpoteesi saavad kinnituse: 1) verbi oleviku ajavormis (*pr*) esineb pronoomen tõenäolisemalt kui minevikuvormis (*ipf*); 2) saarte ja läänemurdes, kus 1SG verbi pöördelõpp *-n* võib kord esineda, kord mitte, on pöördelõputa



Joonis 10. Fikseeritud mõjud regressioonimudelis

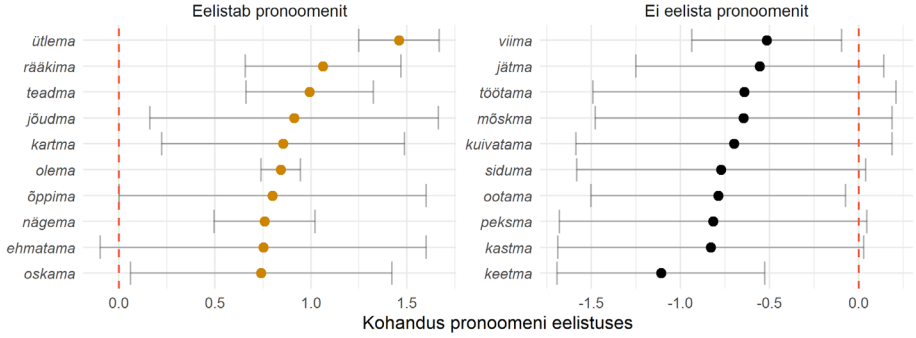
vormide puhul pronoomeni kasutamine oluliselt tõenäolisem kui pöördelõpuga vormide puhul. Seega aitab pronoomeni eksplitsiitne väljendamine neil aladel kompenseerida verbivormis puuduvat referentsiaalset infot. Teistes murretes, kus eriti varieerumist ei esine, pöördelõpu mõju aga puudub. Ülejäänud kaks hüpoteesi aga oletatud kujul täpselt kinnitust ei leia: 1) pronoomeni kasutamise tõenäosust tõstab see, kui viimati lähedal aktiveeritud 1SG viide oli 1SG pronoomen *mis tahes* vormis (sh nominatiivis). Efekt aga kaob ootuspäraselt viitamiskauguse kasvades; 2) pronoomeni kasutamise tõenäosus on mõnevõrra suurem, kui seda kasutati ka eelmise 1SG verbivormi juures. Viitamiskauguse kasvades kaob efekt juhul, kui eelmine 1SG verbivorm oli erinev, ent püsib, kui eelmine 1SG verbivorm oli sama. Tõenäoliselt võib seda seletada praimimisega – tendentsiga korrata (täpselt) sama struktuuri – või on tegemist pronoomeniga ja pronoomenita kinnistunud verbilekseemide kaudse mõjuga (nt *ma arvan* või *palun*).

Võime visualiseerida ka mudeli juhuslikke mõjusid, et näha, kas kohandused mudeli ennustustes konkreetsetele kõnelejatele ja verbidele (nn juhuslikud vabaliikmed) korreleeruvad muude tunnustega, mida nende kohta teame. Kui kohandused on positiivsed (st nullist suuremad), eelistatakse pronoomeni kasutust keskmisest enam, ning kui negatiivsed, siis keskmisest vähem.



Joonis 11. Juhuslikud vabaliikmed regressioonimudelis (kõneleja)

Jooniselt 11, mis kuvab mudeli ennustuste kohandused individuaalsetele kõnelejatele, näeme, et need on jaotunud nulli ümber võrdlemisi juhuslikult ning pronoomeni kasutuse eelistusi ei saa siduda kõneleja vanuse ega sooga (küll aga kõneleja murdetaustaga, nagu nägime jooniselt 6).



Joonis 12. Juhuslikud vabaliikmed regressioonimudelis (verbi lemma)

Joonis 12 kuvab omakorda mudeli ennustuste kohanduste põhjal 10 kõige pronoomenilembesemat ja 10 kõige vähem pronoomenilembest verbi. Näeme, et kui pronoomenit mitte-eelistavad verbid on üldjuhul harilikud tegevusverbid, siis pronoomeni kasutus on eriti ootuspärane suhtlusverbidega *ütlema* ja *rääkima*, samuti kognitiivsete verbidega *teadma, kartma, nägema, oskama*, koopula- ja abiverbiga *olema*, modaalverbiga *jõudma* ning ka verbidega *õppima* ja *ehmatama*, mille

kohandused on aga ebamäärasemad, kuna verbi esinemissagedus andmestikus on madal (sellele viitab ka lai usaldusvahemik punkthinnangu ümber).

Tuleb siiski rõhutada, et juhuslike mõjude sisulisel tõlgendamisel tuleks olla pigem ettevaatlik. Nii kõnelejate kui ka verbide jaotus andmestikus on kallutatud: on üksikud kõnelejad ja verbid, kes panustavad palju andmepunkte, ent on ka neid, kellelt on andmestikus vaid üks vaatlus. Selliseid üksikute vaatlustega „rühmi" kaldub mudel pidama pigem keskmisele tendentsile (nullile) sarnasemaks, mistõttu ei peegelda juhuslike mõjude hinnangud tingimata üksikute kõnelejate/verbide tegelikke eelistusi.

Kokkuvõte

Peatükis vaatasime, kuidas Eesti murrete korpuse andmete põhjal on võimalik uurida grammatika varieerumist. Meie uuritavaks tunnuseks (muutujaks) oli subjektpronoomeni esinemine koos ainsuse 1. isiku jaatavate verbivormidega. Eesti keele puhul teeb selle põnevaks asjaolu, et põhjaeesti murretes on 1SG pöördelõpuks eksplitsiitne *-n*, lõunaeestis see puudub ning lääne- ja saarte murdes varieerub.

Materjal koguti Eesti murrete korpusest automaatselt ning näites on minimeeritud käsitsi tehtavat tööd. Sellegipoolest tuli võimalike vigade vältimiseks automaatselt kogutud materjal mõnes kohas käsitsi üle vaadata. Esmalt vaadeldi ühetunnuselise kvantitatiivse analüüsina pronoomeni esinemist või mitteesinemist koos ainsuse 1. isiku verbivormiga normaliseeritud sageduste põhjal, et võrrelda, kui sarnased või erinevad murded selle tunnuse poolest on. Selgub, et Seto ja Alutaguse murre eelistavad pronoomenit mitte väljendada, ülejäänud murretes väljendati pronoomenit rohkem kui 50% juhtudel. Eriti sagedasti väljendati pronoomenit saarte ja Mulgi murdes. Seda, miks üks või teine murdeala eelistab pronoomenit väljendada või vastupidi, me selles peatükis ei analüüsinud, sest see vajab põhjalikumat keeleajaloo ja keelekontaktide käsitlust.

Selles peatükis oli meie murretega seotud põhiüksuseks murdeala, kuhu kogusime kõik vastavalt murdealalt pärit andmed. See lähenemine eeldab, et murded on seesmiselt piisavalt ühtsed. See ei pruugi aga alati nii olla: ka ühe murdeala sees võib mingite keelenähtuste vahel olla piirkondlikke erinevusi. Varasema 1SG pronoomeni kasutamise puhul vaadeldi põhiüksusena hoopis kihelkondi (Lindström jt 2009), ent tookord oli põhjuseks murdekorpuse ebaühtlane katvus. Siiski tasub iga nähtuse puhul mõelda, mis on optimaalne uuritav üksus (murre, murrak, küla) ning kui palju on materjali iga taolise üksuse kohta.

Järgmisena viisime läbi piirangutele orienteeritud analüüsi, milles vaatlesime, millised keele ja mäluga seotud seletavad tunnused mõjutavad enim pronoomeni väljendamist või mitteväljendamist. Vaatlesime esmalt tunnuste mõju pronoomeni kasutamisele ükshaaval. Seejärel kasutasime segamõjudega logistilist regressiooni,

kuhu kaasasime fikseeritud tunnustele lisaks juhuslikke mõjusid (kõneleja info ja verbi lemma). Üheks oluliseks tunnuseks oli *-n* lõpu olemasolu pöördelõpus, ent kuna see on seotud tugevalt murdealaga (lõunaeesti murded on ilma *n*-lõputa, põhjaeesti murded *n*-lõpuga, saarte ja läänemurdes varieerub), siis kaasasime selle tunnuse analüüsi vaid koosmõjus murdealaga. Ka viitamiskauguse kaasasime mudelisse koosmõjus teiste mälu ja protsessimisega seotud tunnustega. Tulemustest nägime, et olulisteks osutusid aeg ning murdeala ja pöördelõpu olemasolu koosmõju, samuti leidsime nii referentsiaalse kui ka struktuurilise praimimise efektid, ent nende mõju kadus, mida kaugemaks eelmine viide jäi.

Varieerumise analüüsi tulemusi on võimalik võrrelda teiste keeltega, sest see on nähtus, mis varieerub paljudes keeltes ning varieerumist mõjutavad tunnused on vähemalt osaliselt samad (Torres Cacoullos & Travis 2019). Käesoleva analüüsi eesmärk oli mõõta ennekõike murretevahelisi erinevusi. Loomulikult ei ole selle analüüsi tulemused absoluutsed ja vaieldamatud. Me saaksime analüüsi lisada veel tunnuseid, mis praegu kas polnud teada või olid töö automatiseerimise huvides kõrvale jäetud. Tunnuste valimiseks tuleks põhjalikult tutvuda varasemate uurimustega sarnasel teemal ning katsetada ka selliseid tunnuseid, mis andmeid lähemalt vaadates tunduvad olulised ja operatsionaliseeritavad. Üpris tavaline on võtta arvesse grammatilist konteksti (nt verbivormi jaatus-eitus, kõneviis vms), sõnajärge, sõnade/lausete pikkust, sõnade üldist sagedust korpuses, lause (implitsitsiitse) subjekti elusust jms, sõltuvalt uuritavast nähtusest. Praeguses andmestikus kaasasime verbi lemma juhusliku mõjuna, ent sageli rühmitatakse neid tähenduse alusel eraldi tunnuseks (nt kognitsiooniverbid, suhtlusverbid, liikumisverbid, tegevusverbid jne) ning kaasatakse mudelisse fikseeritud mõjuna.

Näidisuurimuse valmimist on toetanud EKKD projekt „Eesti keele morfosüntaktiline varieerumine" (2024–2027).

Kirjandus

Auer, Peter. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences* 31(1). 1–13. https://doi.org/10.1016/j.langsci.2007.10.004.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lindström, Liina, Mervi Kalmus, Anneliis Klaus, Liisi Bakhoff & Karl Pajusalu. 2009. Ainsuse 1. isikule viitamine eesti murretes. *Emakeele Seltsi aastaraamat* 54 (2008). 159–185.

Lindström, Liina & Maarja-Liisa Pilvik. 2018. Korpuspõhine kvantitatiivne dialektoloogia. *Keel ja Kirjandus* 8–9. 643–662. https://doi.org/10.54013/kk730a3.

Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma & Kristel Uiboaed. 2015. Mineviku liitaegade kasutusest eesti murretes keelekontaktide valguses. *Võro Instituudi toimõtisõq* 29. 39–70.

Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma & Kristel Uiboaed. 2019. On the use of perfect and pluperfect in Estonian dialects. Sofia Björklöf & Santra Jantunen (toim), *Multilingual Finnic – Language Contact and Change* (Uralica Helsingiensia 14), 155–193. Helsinki: Suomalais-Ugrilainen Seura. https://doi.org/10.33341/uh.85035.

Lindström, Liina, Triin Todesk & Maarja-Liisa Pilvik. 2022. Eesti murrete korpus. https://doi.org/10.23673/re-365.

Mets, Mari. 2011. Spoken Võro in real time. Variation of the inessive ending. *Linguistica Uralica* 47(4). 257–272. https://doi.org/10.3176/lu.2011.4.02.

Metslang, Helle, Mati Erelt, Külli Habicht, Tiit Hennoste, Reet Kasik, Pire Teras, Annika Viht, jt. 2023. *Eesti grammatika*. (Toim) Ellen Niit, Reet Kasik, Külli Habicht, Helle Metslang & Andriela Rääbis. Tartu: Tartu Ülikooli Kirjastus. https://doi.org/10.12697/EG

Metslang, Helle & Liina Lindström. 2017. The essive in Estonian. Casper de Groot (toim), *Uralic Essive and the Expression of Impermanent State* (Typological Studies in Language 119), 57–90. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.119.03met.

Nerbonne, John & William A. Jr. Kretzschmar. 2013. Dialectometry++. *Literary and Linguistic Computing* 28(1). 2–12. https://doi.org/10.1093/llc/fqs062.

Pilvik, Maarja-Liisa. 2017. Deverbal *-mine* action nominals in the Estonian dialect corpus. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 8(2). 295–326. https://doi.org/10.12697/jeful.2017.8.2.10.

Pilvik, Maarja-Liisa, Helen Plado & Liina Lindström. 2021. Murded, varieerumine ja korpusandmed. Eitussõna paiknemine võru ja seto eituslausetes. *Keel ja Kirjandus* 8–9. 771–796. https://doi.org/10.54013/kk764a7.

Saareste, Andrus. 1955. *Petit atlas des parlers estoniens / Väike eesti murdeatlas*. Uppsala: Kungliga Gustav Adolfs Akademien.

Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–149. https://doi.org/10.1515/cllt.2005.1.1.113.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Lieselotte Anderwald. 2018. Corpus-based approaches to dialect study. Charles Boberg, John Nerbonne & Dominic Watt (toim), *The Handbook of Dialectology*, 300–313. Hoboken, NJ: Wiley Blackwell. https://doi.org/10.1002/9781118827628.ch17.

Torres Cacoullos, Rena & Catherine E. Travis. 2014. Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics* 63(3). 19–34. https://doi.org/10.1016/j.pragma.2013.08.003.

Torres Cacoullos, Rena & Catherine E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692. https://doi.org/10.1515/ling-2019-0011.

Uiboaed, Kristel. 2013. *Verbiühendid eesti murretes* (Dissertationes philologiae estonicae Universitatis Tartuensis 34). Tartu: Tartu Ülikooli Kirjastus.

Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (toim), *The Future of Dialects: Selected Papers from Methods in Dialectology XV* (Language Variation 1), 225–244. Berlin: Language Science Press.

Eesti keele välted foneetika korpuse põhjal

Pärtel Lippus

Lühikokkuvõte

Siin peatükis kasutatakse eesti keele spontaanse kõne foneetilist korpust, et kirjeldada eesti keele vältesüsteemi. Korpus koosneb kõnesalvestustest (WAV-failid) ja mitmetasandilisest aegjoondusega annotatsioonist (TextGrid-failid). Juhtumiuuringus kasutatakse Praati skripte, et leida annotatsiooni põhjal üles uuritavad sõnad ning arvutada häälikute kestused. Lisaks kasutatakse helifaile, millest samuti Praati skriptide abil leitakse põhitooni ja vokaalide formantväärtused. Seejärel analüüsitakse saadud andmeid R-is. Kõnelejatevaheliste erinevuste vähendamiseks kasutatakse erinevaid normaliseerimismeetodeid. Tulemuste kirjeldamiseks, visualiseerimiseks ja väidete tõestamiseks kasutatakse lineaarseid segamudeleid.

1. Sissejuhatus

Käesolevas peatükis teeme läbi väikese välteuurimuse eesti keele spontaanse kõne foneetilise korpuse (Lippus jt 2023) andmetega. Uurimuse eeskujuks on Lippus jt (2013) sama korpuse põhjal läbi viidud uurimus. Eesti keele vältesüsteemi kirjeldamiseks analüüsime häälikute kestust, põhitooni ja vokaalide formante. Andmed kogume failidest Praati (Boersma & Weenink 2023) skriptide abil. Andmete analüüsiks kasutame R-i (R Core Team 2023).

Sissejuhatuseks teeme lühiülevaate eesti vältesüsteemist. Vältesüsteemi põhjalikuma käsitluse leiab näiteks Lippus jt (2013) sissejuhatusest või eesti keele häälduse tervikkäsitlusest (Asu, Lippus, Pajusalu, jt 2016)[1]. Vältesüsteem on kahtlemata eesti keele fonoloogilise süsteemi üks olulisemaid nähtusi, mida on ka kõige põhjalikumalt uuritud. Mis selle huvitavaks teeb, on esiteks see, et kolmese pikkusvastandusega keeli ei ole maailmas kuigi palju (eriti neid, millel oleks süsteemi kaasatud nii vokaalid kui konsonandid). Eesti keeles võib nii konsonant kui vokaal olla nii lühike, pikk kui ülipikk. Õigemini, nagu loodetavasti ka järgnev analüüs

[1] Lisaks võib terminoloogia kohta selgitusi otsida foneetika sõnastikust: https://sonaveeb.ee/ds/fon.

näitab, on eesti välde sõnatasandi (või täpsemini rõhulisest ja järgnevast rõhutust silbist koosneva kõnetakti) nähtus. **Kolme välte süsteemis** võivad väldet kanda kas rõhulise silbi vokaal (nimetame vokaalikeskseks malliks), rõhulise ja rõhutu silbi vaheline konsonant (konsonandikeskne mall) või mõlemad kombinatsioonis (vt näiteid tabelist 1). Esimeses vältes (Q1) on rõhuline vokaal ja järgnev konsonant alati lühikesed. Teises (Q2) ja kolmandas (Q3) vältes võib vokaalikeskses mallis rõhuline vokaal olla pikk või ülipikk monoftong või diftong, millele järgneb lühike konsonant. Konsonandikeskses mallis järgneb lühikesele vokaalile pikk või ülipikk konsonant või konsonantühend. Segamallis on teises ja kolmandas vältes rõhuline pikk vokaal või diftong, millele järgneb pikk konsonant või konsonantühend.

Tabel 1. Näited võimalikest fonoloogilistest struktuuridest kolmes vältes. Näitesõnad on tavaortograafia kõrval esitatud rahvusvahelises foneetilises transkriptsioonisüsteemis (IPA)

Mall		**Q1**		**Q2**		**Q3**	
Vokaal	monoftong	*sada*	sɑ.tɑ	*saada*	sɑ:.tɑ	*saada*	sɑ::.tɑ
	diftong			*paise*	pɑise	*paise*	pɑi:.se
Konsonant	geminaat			*kalla*	kɑl.lɑ	*kalla*	kɑl:.lɑ
	konsonantühend			*metsa*	met.sɑ	*metsa*	met:.sɑ
Mõlemad	monoftong + geminaat			*saate*	sɑ:t.te	*saate*	sɑ:t:.te
	diftong + geminaat			*võite*	vɤit.te	*võite*	vɤit:.te
	diftong + konsonantühend			*paista*	pɑis.tɑ	*paistma*	pɑis:t.mɑ

Teiseks teeb eesti keele välte huvitavaks see, et tegemist on kompleksse prosoodilise nähtusega, millel on seosed kõigi muude keele tasanditega, mis tähendab, et eesti keele kirjeldamisel peab vältega arvestama ka siis, kui vaatluse all on midagi muud kui välde. Näiteks eristab välde süstemaatiliselt käänet sõnatüüpides, mis on astmevahelduslikud (nt omastav *pluusi* – osastav *pluusi*) ja seetõttu võib see olla ka ainus tunnus, mis eristab täis- ja osasihitist: *Ma õmblesin pluusi* võib tähendada nii seda, et õmblesin parasjagu pluusi (Q3, osastav kääne, osasihitis), kui ka seda, et õmblesin pluusi valmis (Q2, omastav kääne, täissihitis). Teisalt võivad välte minimaalpaare või -kolmikuid moodustada täiesti erinevate tüvede eri sõnaliigi vormid, nt *kala* (nimisõna, 'vees elav kõigusoojane selgroogne') – *kalla* (nimisõna, taim ladinakeelse nimetusega *Zantedeschia*) – *kalla* (verbi *kallama* käskiva kõneviisi 2. isiku vorm).

Kuigi fonoloogiliselt võib väldet pidada rõhulise silbi omaduseks, kirjeldab eesti keele kolme välte süsteemi kõige paremini esimese ja teise silbi kestuste suhe. Fonoloogilise pikkuse kõige otsesem akustiline vaste on **kestus**. Kui võrrelda ainult rõhulise silbi (häälikute) kestust, siis esimene ja teine välde eristuvad selgelt: pikk häälik on umbes kaks korda pikem kui lühike. Teise ja kolmanda välte puhul on häälikutasandil erinevus aga väike: ülipikk häälik on ainult pisut pikem kui pikk. Samas kuigi rõhutus silbis fonoloogilised pikkusvastandused puuduvad, varieerub välteti ka rõhutu silbi kestus, mistõttu on nii Lehiste (1960) kui Liiv (1961) pakkunud välte kirjeldamiseks kõige sobivamaks mõõdikuks rõhulise ja rõhutu silbi kestuste suhet. Samuti on mõlemad kirjeldanud teist ja kolmandat väldet eristava lisatunnusena **põhitooni**, mille langus kolmandas vältes on varasem kui teises vältes. Hilisemad uurimused on näidanud, et põhitoon on ka vältetaju jaoks väga oluline tunnus, mis kolmandat väldet eristab.

Mõnevõrra varieerub hääliku pikkusest sõltuvalt ka **vokaalide kvaliteet**. Kui Lehiste (1960) seda väga selgelt ei leidnud ja Eek & Meister (1998) osutasid mõnele väikesele tendentsile, siis Lippus jt (2013) leidsid spontaanse kõne põhjal, et rõhulise silbi vokaalid olid teises ja kolmandas vältes selgemini hääldatud kui esimeses vältes ning kui rõhutu silbi vokaalid olid üldiselt redutseeritud, siis kolmanda välte rõhutu silbi vokaalid olid tugevasti redutseeritud.

Siin näidisuurimuses kontrollime, kas varasemates uurimustes leitud vältega seotud seaduspärasused kehtivad ka väikeses foneetika korpusest võetud valimis. Täpsemalt on eesmärk

1. kontrollida kestussuhteid:
   a) kas rõhulise vokaali kestused eristuvad kolmes vältes?
   b) kas rõhutu vokaali kestused eristuvad välteti?
   c) kas silbialguskonsonantide kestustes on välteti erinevusi?
2. kontrollida, kas põhitooni langus toimub kolmandas vältes varem kui esimeses ja teises vältes;
3. kontrollida, kas vokaalide kvaliteet sõltub vältest.

Neile küsimustele vastamiseks teeme läbi järgmised sammud:

1. Praati skriptiga otsime TextGrid-failidest üles huvipakkuvad sõnad ja kirjutame häälikute kestused tabelisse.
2. Avame tabeli programmis R, tutvume leitud materjaliga ja sorteerime sealt tingimustele mittevastava välja.
3. Analüüsime häälikute kestusi.
4. Leiame Praati skriptiga sõnade põhitoonikontuurid.
5. Normaliseerime R-is kõnelejatevahelist varieeruvust.
6. Analüüsime põhitooni liikumist seoses vältega.
7. Leiame Praati skriptiga vokaalide formantväärtused.
8. Normaliseerime kõnelejatevahelist varieerumist.

9. Normaliseerime vokaalikategooriate erinevused.
10. Analüüsime vokaalide kvaliteeti seoses vältega.

2. Materjal

Materjalina kasutame **eesti keele spontaanse kõne foneetilisest korpusest** (Lippus jt 2023)[2] pärit faile. Kuna tervikkorpusele on ligipääs piiratud, on selle näite jaoks valitud korpusest neli faili (kaks nais- ja kaks meeskeelejuhti), mida on võimalik avaandmetena kättesaadavaks teha. Need materjalid on kättesaadavad käesoleva õpiku repositooriumis[3].

Korpus sisaldab spontaanse kõne salvestusi, mis on märgendatud sõna, hääliku ja silbi tasandil, lisaks veel mitmeid erinevaid märgenduskihte. Spontaanne kõne on väga varieeruv ja kui me tahame uurida just nimelt seda, kuidas vaadeldavad akustilised tunnused on seotud meie uuritava tunnusega (vältega), siis peaksime võimalikult palju üritama kontrolli alla saada muid tunnuseid, mis samuti võivad akustiliste tunnuste varieerumist mõjutada. Niisiis, seame mõned tingimused sellele, milliste piirangutega materjali korpusest otsime:

- **kahesilbilised sõnad** – eesti keeles võib sõnas olla teoreetiliselt üks kuni kaksteist silpi (Asu, Lippus, Pajusalu, jt 2016: 154–156). Ühesilbilised sõnad kolme välte süsteemis ei osale. Välistame ka võimaliku sõnaisokroonia efekti – kui sõnas on rohkem silpe ja häälikuid, siis häälikud on lühema kestusega – ning piirdume ainult kahe silbi pikkuste sõnadega.
- **lahtised silbid** – teises ja kolmandas vältes võib rõhulise silbi riim koosneda pikast monoftongist või diftongist, lühikesest vokaalist ja konsonandist või ka pikast vokaalist ja konsonandist (vt tabel 1). Erinevatel kombinatsioonidel on ka mõningane mõju silbi kestusele üldiselt (Lippus & Šimko 2015). Et hoida ära võimalikku silbistruktuurist tingitud varieerumist, võtame uurimusse ainult vokaalikeskse malliga monoftongidega lahtiste silpidega sõnad (nt *sada – saada – saada*, *pole – poole – poole*).
- **välistame fraasilõpulised sõnad** – nagu paljudes keeltes, märgitakse prosoodilisi-süntaktilisi piire eesti keeles lõpupikenemisega, st fraasi lõpus häälikud pikenevad (Krull 1997). Seetõttu jätame välja sõnad, millele järgneb paus või mis on venitatud.
- **välistame sosina**, kuna tahame vaadata ka põhitooni liikumist ja sosinkõnes põhitoon puudub. Samuti võib sosin teha keerulisemaks vokaali formantide leidmise.

[2] https://foneetikakorpus.ut.ee/

[3] https://osf.io/xqzsf/

3. Praat

Kestuse, põhitooni ja formantide andmete kogumiseks kasutame vabavaralist programmi Praat (Boersma & Weenink 2023)[4]. Praatis on vahendid akustiliseks analüüsiks graafilises kasutajaliideses, mille võimalused vaatame siin põgusalt üle. Lisaks graafilisele kasutajaliidesele on Praatis ka täismahus skriptikeel, mida kasutame käesoleva peatüki andmete kogumiseks. Selle näidisuurimuse skriptid on mõnevõrra kommenteeritud, aga väga põhjaliku ja algajale sobiliku Praati skriptikeele õpetuse leiab Praati kasutusjuhendist[5]. Samuti selgitab põhjalikumalt nii akustilise foneetika aluseid kui Praati kasutamist eesti keeles akustilise foneetika meetodite õpik (Lippus 2026).

Korpusest on meil vaja helifaile (WAV-vormingus) ja helifaili juurde kuuluvaid aegjoondusega TextGrid-faile. TextGrid[6] on Praati märgendusvorming. Nagu öeldud, leiab siin peatükis analüüsitavad failid õpiku repositooriumist. Faili avamiseks valime Praati objektiaknas *Open* menüüst käsu *Read from file*.... Sama käsuga saab avada nii helifaili kui TextGrid-faili. Nüüd peaks olema failid ilmunud *Objects* loendisse. Valime mõlemad objektid ja vajutame nuppu *View & Edit*, mille peale avaneb uus aken, kus on kolm jaotust: helilaine, spektrogramm ning TextGridi objekti märgenduskihid (vt joonis 1). Selles aknas on graafilise kasutajaliidese vahenditega võimalik helilaine põhjal kõne akustilisi tunnuseid mõõta. Pikemalt me sellest siiski siin peatükis ei räägi, vaid kasutame olemasolevaid skripte.

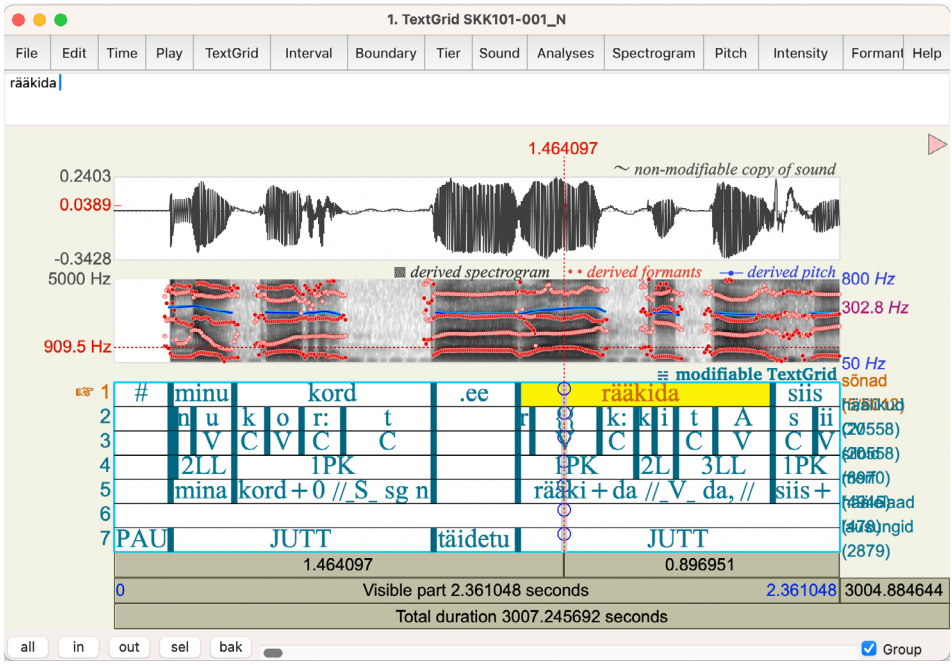
Kõigepealt kogume kestusandmed. Valime Praati objektiakna menüüst *Praat* käsu *Open Praat script*... ja avame faili *korp_opik_fonkorp_valteandmed.praat*. Skript avaneb uues skriptiaknas ja on oma olemuselt tavaline tekstifail. Enne skripti käivitamist tuleks skripti algusest üles otsida ja enda arvuti failisüsteemi arvestades ära muuta kaustade aadressid, kus asuvad korpuse failid (muutuja *kaust$*) ja kuhu kirjutatakse tulemused (*tulemusteKataloog$*). Tuleb meeles pidada, et kausta aadressi lõpus oleks kindlasti kaustaeristaja /. Seejärel valime menüüst *Run* käsu *Run*. Kui skript lõpetab, siis on tulemuste kausta tekkinud fail nimega *kestuste_tulemused.txt*.

Kuigi siin peatükis kasutame Praati, et TextGridi märgendusega helifailidest andmed kätte saada, siis on ka R-i ja Pythoni pakette, millega saab selle töö ära teha. TextGrid-failide lugemiseks ja muude Praatiga tekitatavaid objekte analüüsimiseks on R-is pakett `rPraat` (Bořil & Skarnitzl 2016). Selle paketiga töötades peab aga näiteks põhitooni- ja formantanalüüsid tegema siiski Praatis (ja kui me ei taha teha korraga tervest failist analüüse, vaid eri lõikudest erinevate sätetega, siis on see tülikas). Päris alternatiiviks Praatile on R-i pakett `phonTools` (Barreda 2023), mis võimaldab R-is teha põhitooni- ja formantanalüüsi ja palju muud (Praatis on siiski mõnevõrra laiemad võimalused täpsemalt seadeid määrata). On

[4] https://www.praat.org/

[5] https://www.fon.hum.uva.nl/praat/manual/Scripting.html

[6] https://www.fon.hum.uva.nl/praat/manual/TextGrid.html



Joonis 1. Praati toimetamisakna ekraanipilt

ka pakett `speakr` (Coretta 2024), millega saab R-is Praati skripte käivitada, aga Praatile suunatavad käsud peavad seal olema Praati skriptikeeles. Pythonis on Praati kasutamiseks teek `Parselmouth` (Jadoul, Thompson & de Boer 2018), mille kaudu saab kasutada suuremat osa Praati funktsioone Pythoni keeles.

4. Kestused

Alustame analüüsi häälikukestustest. Kuna vaatluse all on kahesilbilised lahtiste silpidega (ehk vokaalikeskse malliga) sõnad, siis on igas sõnas neli häälikut: alguskonsonant (C1), esimese silbi rõhuline vokaal (V1), teise silbi alguskonsonant (C2) ja teise silbi rõhutu vokaal (V2), mille kestusi analüüsi kaasame. R-i skript, millega käesolev analüüs on tehtud, on kättesaadav õpiku repositooriumis. Kõik kirjeldatud sammud on skriptis kommenteeritud, et analüüsi oleks võimalik ise läbi teha.

4.1. Andmete lugemine R-i ja täiendav filtreerimine

Esimese sammuna loeme Praati häälikukestuste leidmise skriptiga saadud andmed R-i ja teeme tabelis mõned kohendused:

- kodeerime kõneleja soo ja välte tunnustena, mille tasemetel on kindel järjekord. Sellega välistame analüüsis tasemete järjekorra varieerumise juhusliku esinemisjärjekorra tõttu erinevates andmestiku alamosades, mille tagajärjel võiks erinevatel joonistel olla kategooriate järjekord erinev;
- arvutame häälikute algus- ja lõpuaegade põhjal häälikukestused ja teisendame need sekunditest millisekunditeks, sest häälikukestusi on nii visuaalselt parem jälgida. Selle sammu oleks võinud teha ka Praati skriptis, aga kestuse arvutamine on väga lihtne tehe: lahutame lõpuajast algusaja. Igal juhul on mõistlik andmestikus talletada ka algus- ja lõpuajad, sest kui hiljem on analüüsi käigus tarvis midagi helifailist täpsustada või juurde mõõta, on aegade kaudu võimalik sõna failist üles otsida;
- otsime morfoloogilisest märgendusest üles sõnaliigi tähise ja paneme selle eraldi tulpa. Ka seda oleks võinud teha Praati skriptis, aga R-is on see natuke lihtsam.

Seejärel filtreerime veel andmestikust üht-teist välja, et vähendada häälikukestuste varieerumist, mis võiks olla seotud muude nähtustega kui välde. Jällegi, kõike seda saaks teha ka Praati skriptis, kuid R-iga on mõnevõrra lihtsam.

Kuna andmestikku jäi sisse ka häälikutasandil venitatuks märgitud sõnu, siis viskame nad nüüd välja. Venitus on häälikutasandil märgitud topeltkooloniga (nt sõna *kõne* oleks venitatud rõhutu silbi vokaaliga SAMPA transkriptsioonis transkribeeritud [k7ne::]). **Välja jäävad** kõik sõnad, kus mõni häälik on märgitud venitatuks.

Jätame välja sõnaliigi poolest grammatilised sõnad, mis võivad olla suure tõenäosusega redutseeritud. Andmestikku jätame alles adjektiivid (sõnaliigitähisega *A*, *C*, *U*), pärisnimed (*H*), hüüdsõnad (*I*), põhi- ja järgarvsõnad (*N*, *O*), nimisõnad (*S*) ja tegusõnad (*V*). Välja jäävad määrsõnad, sidesõnad, kaassõnad, asesõnad, lühendid jms. (Täpsemalt vaata morfoloogilise analüüsi kohta Vabamorfi kirjeldusest[7].) Tabelis 2 on loend sõnadest, mis selle tulemusena andmestikku jäid.

[7] https://www.filosoft.ee/html_morf_et/morfoutinfo.html#2

Tabel 2. Andmestikus esinevad sõnad

Välde	N	Sõnad
Q1	248	*kõne, ole, tohi, kadu, sõna, tere, kobe, mõju, näha, koha, pära, lumi, võru, kana, vahe, kogu, neli, keele, teha, mage, ole/.?, kahe, pidi, suhe, pähe, vali, müra, sõnu, nime, sada, saja, muna, kuju, sära, nõme, tuli, pere, mina, tore, koma, kena, tänu, huvi, sisu, rida, vaba, büroo, Mare/.isikunimi, Pire, vajab, kada, pole, viga, sobi, mari*
Q2	111	*keele, räägi, hääle, Mona, Lisa, tuua, näeme, joone, liigu, viisi, saame, kiire, sööme, roosa, kuule, tüübi, geeni, paari, Via, suuda, viie, viide, loome, teeme, luua, ruumi, poole, moora, Leelo*
Q3	50	*tuua, jooni, jääma, saada, poole, maali, haara, keeli, tooma, tuuma, soola, tooli, jääda, suuri, saama, luua, piisa, moodi, viia/.?, kuulu, viia*

Tabelist 2 torkavad silma mõned sõnad, mis on mitteootuspäraselt kategoriseeritud ja mille puhul peaks kaaluma, kas need tuleks veel andmestikust eemaldada. Esmavälteliste sõnade hulgas on sõna *büroo*, mis on võõrsõna rõhuga teisel silbil. See ei sobitu ülejäänud materjali hulka, kus on rõhk esisilbil. Mingil põhjusel on see TextGridil märgendatud nii, nagu teine silp oleks lühike. Võiks minna Praati, otsida sõna helifailist üles ja kontrollida, kuidas seda on hääldatud, aga selle asemel viskame käesolevas analüüsis selle sõna lihtsalt andmestikust välja.

Samuti võiks välja jätta esimese välte alla sattunud sõna *keele*. Tõenäoliselt on tegemist (fookus)rõhulise sõnale järgneva deaktsentueeritud häälduse ja pikk vokaal on redutseerunud. Kuna esmavältelisi sõnu on andmestikus nagunii rohkem kui teisi, võime need ka rahumeeli välja jätta.

Veel on näha, et andmestikku on sattunud sõna *ole*, mis ortograafias ei alga konsonandiga, aga sõnaalguskonsonandiks on märgenduses /j/. Ilmselt on see tekkinud koartikulatoorselt järjendis *ei ole* sõnade piirile. Kuna ei ole ühtegi empiirilist uurimust, mis võrdleks selliseid koartikulatsioonist tekkinud konsonante muude konsonantidega ja ei ole päris selge, kas need kestuste poolest muudest konsonantidest ei eristu, siis jätame siin need ka igaks juhuks välja.

Teisevälteliste sõnade hulgas leidub mõningaid võõrnimesid, mis ortograafia järgi on lühikese vokaaliga (*Mona, Lisa*) või sobimatu struktuuriga (*Via*), aga kuna andmed valisime häälikutasandi märgenduse põhjal, siis need on õiges kohas teisevälteliste sõnade hulgas (SAMPA transkriptsioonis [moonA], [liisA], [viijA]) ja võivad jääda andmestikku alles.

4.2. Kestusandmete analüüs

Teeme nüüd väikese kokkuvõtte, kui palju andmeid õnnestus kätte saada. Kokku on andmestikus 345 vaatlust kahelt nais- ja kahelt meeskeelejuhilt. Teeme selle põhjal risttabeli kõneleja soo ja välte esinemisjuhtudest (tabel 3). Tegemist on seega suhteliselt väikese andmestikuga, aga sellele vaatamata on iga välte kohta üle 10 vaatluse. Erinevus andmete jaotuse osas välteti on ka mõnevõrra ootuspärane: esmavältelisi sõnu on oluliselt rohkem kui kõiki muid, sest andmete otsingul seatud piirang, et kaasata analüüsi ainult lahtised silbid, piirab teise ja kolmanda vältega sõnade kaasamist oluliselt rohkem.

Tabel 3. Analüüsitavate sõnade arv välteti mees- ja naiskeelejuhtidel

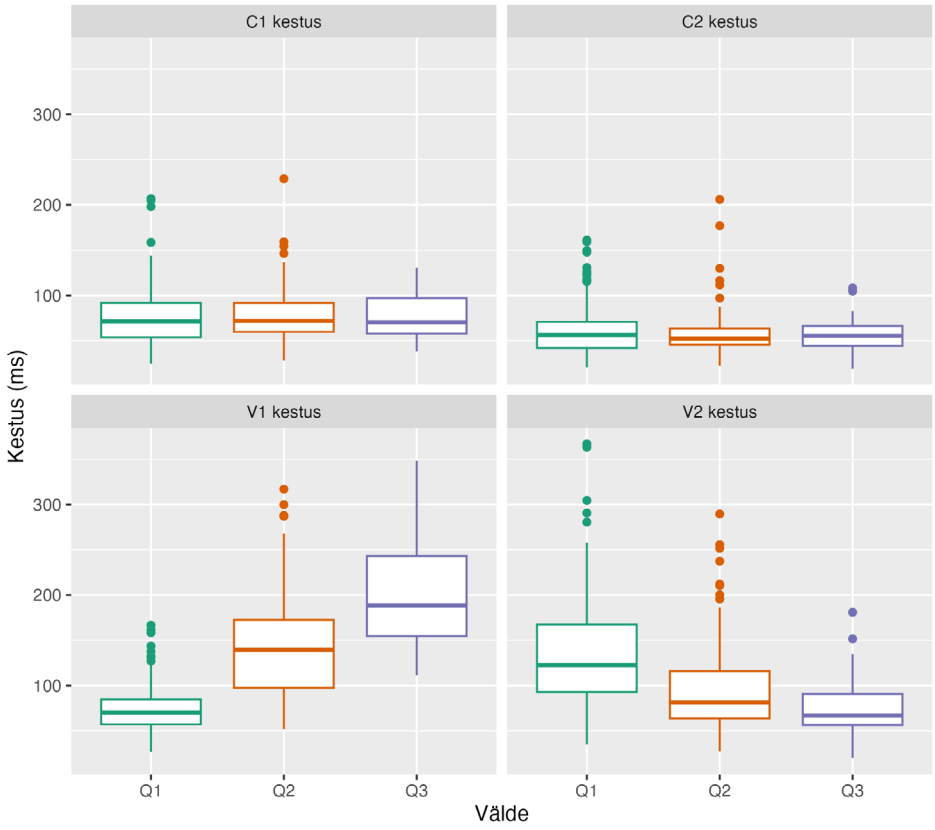
sugu	Q1	Q2	Q3
N	93	56	26
M	91	55	24

Et andmetest paremat ülevaadet saada, joonistame häälikute kestused välteti karpdiagrammina (joonis 2). Karpdiagramm näitab andmete jaotust: karbi keskel jämedam joon tähistab mediaanväärtust, karp kvartiilhaaret ja vurrud (kasti alumisest ja ülemisest küljest välja ulatuvad jooned) usaldusvahemikku; karpdiagrammi kohta leiab rohkem infot õpiku peatükist 6.1.5.2 „Karpdiagramm".

Lisaks joonisele teeme keskmiste ja standardhälvete tabeli (tabel 4). Nagu joonisel kast ja vurrud, annab standardhälve tabelis infot andmete jaotuse ja hajuvuse kohta. Kui võrdleme kolme välterühma keskmisi kestusi, siis standardhälbe piirid näitavad seda, kui selgelt rühmad eristuvad (vt ka õpiku peatükki 6.1.3 „Aritmeetiline keskmine ja standardhälve").

Tabel 4. Häälikukestuste keskmised (*kesk*) ja standardhälbed (*sd*) millisekundites

	C1		V1		C2		V2	
Välde	kesk	*sd*	kesk	*sd*	kesk	*sd*	kesk	*sd*
Q1	77	31	74	26	61	28	133	58
Q2	77	30	145	58	58	25	98	53
Q3	77	24	202	59	57	18	75	30



Joonis 2. Häälikute kestused välteti: C1 – sõna alguskonsonant, C2 – vokaalidevaheline konsonant, V1 – rõhulise silbi vokaal, V2 – rõhutu silbi vokaal

Tabelist 4 ja jooniselt 2 näeme, et konsonantide puhul on väldetevahelised erinevused väiksemad kui standardhälbed ühe välte piires, vokaalide puhul on väldetevaheline varieerumine suurem, aga standardhälbe piirides on siiski ka kattumisi. Järgnevalt analüüsime tabelis 4 ja joonisel 2 esitatud tulemusi põhjalikumalt häälikute kaupa ning kontrollime järeldava statistika abil, kas kestuse vältega seotud varieerumine on selle hääliku puhul statistiliselt oluline või mitte.

4.2.1. Rõhulise vokaali kestus

Välte kirjeldamisel huvitab meid enim rõhulise silbi vokaali kestus, sest see on vokaalikeskses mallis fonoloogiliselt välte kandja. Tabelist 4 ja jooniselt 2 võib näha, et esmavältelise vokaali kestus on keskmiselt 74 millisekundit, teises vältes vokaali kestus keskmiselt 145 millisekundit, mis tähendab, et teises vältes vokaal on keskmiselt umbes poole pikem kui esimeses vältes. Kolmandas vältes on vokaali

kestus keskmiselt 202 millisekundit ja see on teisest vältest üksjagu pikem. Kui aga arvesse võtta ka standardhälbeid, siis on välterühmade esimese vokaali kestuses mõningane kattuvus, mida näeme joonisel 2 sellest, et eri väldete kastid on osaliselt üksteisega kohakuti.

Et kontrollida, kas vokaali kestus eristab kolme väldet, testime tulemusi **lineaarse segamudeliga**, kasutades R-i paketti `lme4` (Bates jt 2015). Siin mudelis võtame uuritavaks tunnuseks hääliku kestuse ja seletavaks tunnuseks sõna välte. Peale selle lisame mudelisse ka **juhuslikud vabaliikmed** (ingl *random intercept*) keelejuhtidele ja sõnadele, kuna mõned keelejuhid võivad rääkida kiiremini kui teised ja mõningaid sõnu võidakse hääldada kiiremini kui teisi. Juhuslikke kaldeid, mis arvestaksid lisaks sellega, et seletava tunnuse ehk välte mõju võib samuti sõnuti erineda, siin mudelis lisada ei saa. Seda seetõttu, et enamik sõnavorme on alati samas vältes (nt sõna *keele* on alati Q2 või sõna *mage* on alati Q1) ja ainult väike hulk sõnu on leksikonis varieeruva vältega (vt nt Piits & Kalvik 2017). Samuti ei lisa me mudelisse juhuslikke kaldeid, mis võimaldaksid arvestada sellega, et välte mõju võib ka kõnelejati erineda (nt võivad mõne kõneleja häälikud olla esimeses vältes keskmisest lühemad, aga teises vältes keskmisest pikemad, teisel kõnelejal jällegi esimeses vältes keskmisest pikemad, aga teises vältes ainult natuke pikemad jne). Regressioonimudelite kohta saab rohkem lugeda õpiku ptk-st 6.2.2 „Mitmetunnuseline seoste analüüs: statistilised mudelid".

Kuna seletaval tunnusel *välde* on kolm taset, siis annab mudeli väljund meile kaks võrdlust baastaseme ja mõlema alternatiivse taseme vahel. Vaikimisi oleks meil baastase Q1 ja mudel esitaks võrdluse Q1 *vs.* Q2 ning Q1 *vs.* Q3. Alternatiivsete tasemete võrdlemiseks (Q2 ja Q3 vahelise erinevuse leidmiseks) peaksime tegema *post-hoc*-testi. Kuna me teame, et kõige suurem erinevus on Q1 ja Q3 vahel ja me tahame kontrollida, kas kummagi tasemega ka Q2 erinevus oluline on, võiksime määrata baastasemeks hoopis Q2. Nii saame hiljem *post-hoc*-testi vajadusest kõrvale põigata, sest lineaarse mudeli väljund annab võrdlused Q2 *vs.* Q1 ning Q2 *vs.* Q3.

Kuna uuritav tunnus on kestus ja kestusandmete jaotus on tihti paremale kaldu, siis **logaritmime** uuritava tunnuse väärtused, et paremini täita mudeli eeldusi normaaljaotuse osas (vt ka ptk 6.2 „Järeldav ehk inferentsiaalne statistika"), st paneme uuritavale tunnusele *V1_kest* ümber käsu *log()*. Pikemalt mudeli eelduste kontrollimisega siiski siin praegu ei tegele, jäägu see statistikaõpikutele.

Lineaarse segamudeli jaoks oleks nüüd tarvis aktiveerida paket `lme4` ning selle abipakett `lmerTest` (mis lisab mudeli väljundisse *p*-väärtused). Mudeli saab käsuga *lmer()* ning regressioonivalem on järgmine:

```
log (V1_kest) ~ välde + (1|fail) + (1|sõna)
```

Valemis on täpsustatud, et (logaritmitud) rõhulise silbi vokaali kestus sõltub vältest ning mudeli vabaliiget korrigeeritakse iga faili/kõneleja ning iga sõna suhtes.

Mudeli koguväljundit saame vaadata käsuga *summary()*, aga kuna kõige rohkem huvitab meid sealt välte **fikseeritud mõjude** osa, siis ruumi kokkuhoiu mõttes

esitame siin teksti sees väljundist ainult fikseeritud mõjude (ingl *fixed effects*) tabeli (tabel 5), mis antud mudeli puhul sisaldab ainult välte mõju kestusele.

Tabel 5. Rõhulise vokaali kestuste segamudeli fikseeritud mõjud

	Hinnang	Standard-viga	Vabadus-astmete arv	t-väärtus	Pr(>|t|)
(vabaliige)	4,961	0,143	3,510	34,661	<0,001
välde Q1	-0,599	0,054	103,373	-11,158	<0,001
välde Q3	0,299	0,062	168,127	4,784	<0,001

Tabeli 5 viimasest tulbast näeme, et kõik mõjud on mudeli seisukohast statistiliselt olulised ($p < 0{,}001$ tähendab, et on väiksem kui 0,1% tõenäosus, et näeksime enda andmetes väldete vahel sellist erinevust, kui tegelikult kehtiks nullhüpotees ja väldete vahel vokaalikestuses erinevusi ei oleks).

Mudeli väljundi esimeses tulbas on esitatud mudeli uuritava tunnuse (V1 kestus) hinnangulised väärtused. Esimesel real näeme mudeli **vabaliikme** hinnangulist väärtust, see on uuritava tunnuse väärtus seletava tunnuse baastaseme korral ehk V1 kestus siis, kui välde on Q2. Teistel ridadel näeme vastava taseme ja baastaseme erinevust: kui palju kestus keskmiselt muutub, kui Q2 asemel on Q1 või kui Q2 asemel on Q3. Aga kuna mudelisse sisestasime uuritava tunnuse V1 kestuse logaritmitud kujul, siis on mudeli hinnangulised väärtused samuti logaritmitud kestused. Selleks, et neid millisekunditesse tagasi saada, tuleks hinnangute väärtused kokku liita ning astendada[8]:

- Q2 (ehk vabaliikme) hinnang on 4,961 ehk $e^{4{,}961} = 142{,}8$ millisekundit;
- Q1 hinnang on 4,961 + -0,599 = 4,362 ehk $e^{4{,}362} = 78{,}5$ millisekundit;
- Q3 hinnang on 4,961 + 0,299 = 5,26 ehk $e^{5{,}26} = 192{,}5$ millisekundit.

Niisiis on mudeli hinnangul esimeses vältes rõhuline vokaal lühem kui teises vältes ja see erinevus on statistiliselt oluline ($\beta = -0{,}599$, $t = -11{,}158$, $p = <0{,}001$)[9]. Kolmandas vältes on rõhuline vokaal pikem kui teises vältes ja see erinevus on samuti statistiliselt oluline ($\beta = 0{,}299$, $t = 4{,}784$, $p = <0{,}001$).

[8] Kui logaritmimisel teisendatakse väärtused irratsionaalarvu *e* ehk Euleri arvu (mille väärtus on umbes 2,71828) astmeteks, siis logaritmitud väärtuse originaalühikutesse teisendamiseks tuleb astendada *e* logaritmitud väärtuse astmesse (nt 2,71828 astmes 4,961).

[9] Tunnuse olulisuse hinnang tuleb lineaarse segamudeli väljundist, siin tabeli 5 teiselt realt: Q2 (baastaseme) ja Q1 erinevus on -0,599, see on mudeli hinnang ja siin tähistatud kui β ning tõenäosus, et leiaksime valimis sellise erinevuse, kui tegelik erinevus tasemete vahel on 0, on väiksem kui 0,1% (ehk $p < 0{,}001$).

4.2.2. Rõhutu silbi vokaali kestus

Rõhutu silbi vokaali kestuste juures näeme tabelist 4 ja jooniselt 2 pööratud seost esimese vokaaliga: mida suurem välde, seda lühem vokaal, erinevused jäävad paarikolmekümne millisekundi piiresse. Standardhälve on kõige väiksem lühikesel kolmanda välte vokaalil. Teeme siin samasuguse segamudeli nagu rõhulise silbi vokaali kestusega, ent uuritav tunnus on sedapuhku rõhutu vokaali silbi kestus.

Tabel 6. Rõhutu vokaali kestuste segamudeli fikseeritud mõjud

	Hinnang	**Standard-viga**	**Vabadus-astmete arv**	**t-väärtus**	**Pr(>|t|)**
(vabaliige)	4,470	0,148	3,630	30,266	<0,001
välde Q1	0,338	0,062	63,194	5,488	<0,001
välde Q3	-0,244	0,076	110,726	-3,222	0,002

Mudeli väljundist tabelis 6 näeme, et esimeses vältes on rõhutu vokaal pikem kui teises vältes ja erinevus on oluline (β = 0,338, t = 5,488, p = <0,001). Kolmandas vältes on rõhutu vokaal lühem kui teises vältes ja see erinevus on samuti statistiliselt oluline (β = -0,244, t = -3,222, p = 0,002).

4.2.3. Konsonantide kestus

Varasemad uurimused (Lippus jt 2013) on näidanud, et ka silbialguskonsonantidel on välteti väikesed kestuserinevused. Tavaliselt jäävad need 10 millisekundi piiresse ja tavaliselt on konsonandid kõige lühemad teises vältes. Käesolevas andmestikus nii suuri erinevusi ei ole: tabelist 4 näeme, et sõna alguskonsonandi (C1) kestus on kõigis kolmes vältes keskmiselt sama väärtusega ja teise silbi alguskonsonandi (C2) keskmiste kestuste erinevused jäävad 5 ms piiresse.

Tabel 7. Sõna alguskonsonandi kestuste segamudeli fikseeritud mõjud

	Hinnang	**Standard-viga**	**Vabadus-astmete arv**	**t-väärtus**	**Pr(>|t|)**
(vabaliige)	4,210	0,105	5,141	40,110	<0,001
välde Q1	0,037	0,069	74,426	0,537	0,593
välde Q3	0,053	0,078	143,129	0,680	0,498

Tabel 8. Teise silbi alguskonsonandi kestuste segamudeli fikseeritud mõjud

	Hinnang	Standard-viga	Vabadus-astmete arv	t-väärtus	Pr(>|t|)
(vabaliige)	4,090	0,122	4,947	33,536	<0,001
välde Q1	0,016	0,077	108,527	0,203	0,839
välde Q3	-0,057	0,076	261,538	-0,749	0,454

Mudelid osutavad, et olulist vältega seotud varieerumist konsonantide kestustes ei ole: tabelites 7 ja 8 toodud mudelite väljundid näitavad, et ei esimese ega kolmanda välte konsonantide kestuse erinevus ei ole baastaseme ehk teise välte konsonandi väärtustest oluliselt erinev[10].

4.2.4. Kestussuhe

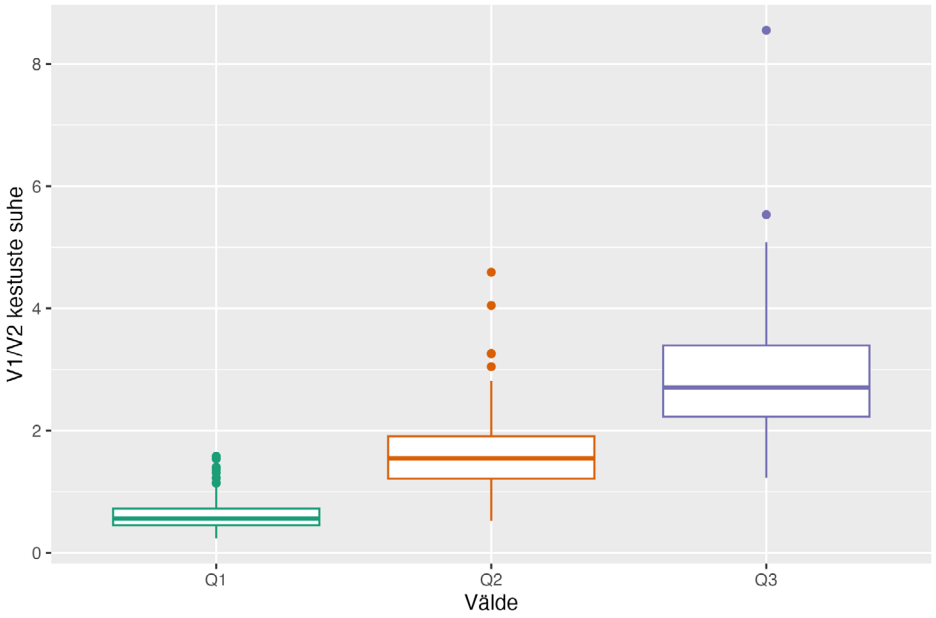
Kuna rõhulise vokaali puhul on väga selge erinevus esimese ja teise välte vahel, aga teise ja kolmanda välte erinevus on võrdlemisi väike, tasub vaadata V1/V2 kestuste suhet, mida kõige sagedamini on kasutatud välte eristuse mõõdikuna (Asu, Lippus, Pajusalu, jt 2016: 134–137). Kuigi sellest räägitakse tihti ka kui silbisuhtest, siis jäetakse enamasti arvutusest välja silpide alguskonsonandid ja arvestatakse ainult silbiriimi, mis meie andmestikus tähendab ainult vokaali (kinniste silpide puhul moodustaks riimi vokaal ja koodakonsonant või -konsonandid). Tavaliselt esitatakse kestussuhet nii, et rõhulise vokaali kestus jagatakse rõhutu vokaali kestusega (vt joonis 3).

Silbisuhete hindamiseks viidatakse kõige sagedamini Lehiste (1960) üldistatud sihtarvudele:

- esimeses vältes 2/3 (ehk 0,7),
- teises vältes 3/2 (ehk 1,5),
- kolmandas vältes 2/1 (ehk 2,0).

Lehiste esitas neid murdarvudena ning tegemist ei olnud täpsete mõõtmisandmetega, vaid mõõtmiste põhjal üldistatud kirjeldusega, mille lähedale võiksid jääda ka mõõtmisandmed. Kui võrrelda eri uurimuste tulemusi (vt nt Asu, Lippus, Pajusalu, jt 2016: 136, tabel 4.1), siis on neis üksjagu varieerumist, aga üldine ja püsiv on see, et esmavältelise sõna rõhuline silp on oluliselt lühem kui rõhutu silp (suhe <1),

[10] Tabelites 7 ja 8 näitab seda teise ja kolmanda rea viimases tulbas olev *p*-väärtus, mis on suurem kui 0,05, väljendades seda, et tõenäosus leida meie korpusvalimist tabeli hinnangu lahtrites esitatud imeväikesi, peaaegu nullilähedasi erinevusi, kui tegelikkuses kehtib nullhüpotees ja eesti keeles väldete vahel konsonandi kestustes väldetevahelisi erinevusi ei ole, on suurem kui 5%.Teisisõnu on suur tõenäosus, et näeme väga väikeseid erinevusi seetõttu, et tegelikkuses ongi erinevused väga väikesed ja välde uuritavate konsonantide kestust oluliselt ei mõjuta.



Joonis 3. Rõhulise ja rõhutu silbi vokaalide kestuste suhe

teisevältelise sõna rõhuline silp on pikem kui rõhutu silp (suhe on >1, aga <2) ning kolmandavältelise sõna rõhuline silp on väga palju pikem kui rõhutu silp (suhe >2). Siin uurimuses on keskmised silbisuhted sellised:

- esimeses vältes 0,6,
- teises vältes 1,7,
- kolmandas vältes 3.

Testime ka lineaarse segamudeliga, kas silbisuhe eristab välteid.

Tabel 9. Silbiriimide kestussuhete segamudeli fikseeritud mõjud

	Hinnang	**Standard-viga**	**Vabadus-astmete arv**	**t-väärtus**	**Pr(>|t|)**
(vabaliige)	0,471	0,082	5,853	5,757	0,001
välde Q1	-0,922	0,060	70,577	-15,311	<0,001
välde Q3	0,570	0,072	120,321	7,885	<0,001

Mudeli väljundist tabelis 9 näeme, et esimeses vältes on kestussuhe väiksem kui teises vältes ja erinevus on statistiliselt oluline (β = -0,922, t = -15,311, p = <0,001). Kolmandas vältes on kestussuhe suurem kui teises vältes ja see erinevus on samuti statistiliselt oluline (β = 0,570, t = 7,885, p = <0,001).

5. Põhitoonianalüüs

Kuigi välte primaarne tunnus on pikkus ja kestuste analüüs näitas, et rõhulise ja rõhutu silbi kestuste suhe eristab selgelt kõiki kolme väldet, siis varasemad uurimused (eriti tajukatsed) on näidanud, et põhitoonikontuur on kestussuhete kõrval oluline sekundaarne vältetunnus.

5.1. Põhitooni andmete kogumine Praatis

Põhitooni analüüsimiseks läheme uuesti Praati. Praati toimetamisaknas kuvatakse põhitoonikontuuri sinise joonena spektrogrammi peal ning selle skaala kuvatakse akna paremas servas (vt joonis 1). Põhitoonianalüüsi puhul on oluline määratleda võimalikult täpselt analüüsitav sageduste vahemik, et vältida oktavivigu. Põhitooni seadeid saab Praati toimetamisaknas määrata menüüst avanevas valikute aknas: *Pitch*: *Pitch settings…*.

Põhitooni analüüsimisel võib kergesti juhtuda, et ekslikult peab algoritm kahte täisvõnget üheks või vastupidi, mistõttu tekivad andmetesse nn **oktavivead**. See tähendab seda, et põhitooni väärtust näidatakse oktavi võrra kõrgema või madalamana, kui see tegelikult on. Veaohtu on võimalik vähendada, kui seada analüüsile täpsed piirid, mis vahemikus kõneleja põhitoon liigub ja kust algoritm seda otsima peab. Samas ei tohi piire liiga kitsalt seada, sest kui põhitooni tegelik väärtus jääb analüüsitavast vahemikust välja, siis surutakse põhitoonianalüüsile samuti peale oktaviviga, et näidata väärtust etteantud vahemikus.

Meeshääle ulatus on tavaliselt vahemikus 75–200 Hz, naishääle ulatus 150–300 Hz. Mida täpsemalt oskame konkreetse kõneleja hääleulatust määratleda, seda vähem vigu andmestikku satub. Analüüsivahemiku määramiseks peaksime tegema mõningaid vaatlusi, et teada saada, mis vahemikus kõneleja põhitoon varieerub. Et seda teha automaatselt ja parimal viisil, kasutame meetodit, mida on kirjeldanud De Looze & Hirst (2008): teeme esmalt põhitoonianalüüsi küllaltki suure analüüsivahemikuga, siis leiame sealt esimese ja kolmanda kvartiili ja seame põhitoonianalüüsi alumiseks piiriks 0,75-kordse 1. kvartiili ja ülemiseks 2-kordse 3. kvartiili.

Käivitame nüüd Praatis skripti *korp_opik_fonkorp_pohitoon.praat*. Ka siin skriptis tuleb konkreetse arvuti failisüsteemi järgi ära vahetada kaustade aadressid, kust skript faile otsib ja kuhu tulemused kirjutab. See skript eeldab, et oleme juba jooksutanud eelmist, kestuste leidmise skripti, sest siit on märgenduse järgi

sõnade otsimise osa välja jäetud ja skript võtab aluseks kestuste tabeli, kuhu põhitooni andmed juurde lisatakse. Seega peaks tulemuste kaustas juba olema fail *kestuste_tulemused.txt*. Skripti jooksutamise lõpuks on töökataloogi tekkinud uus fail *pohitooni_tulemused.txt* (mille leiad ka peatüki repositooriumist). Loeme selle faili R-i (arvestades, et Praati *--undefined--* väärtused võiks saada siin *NA*-ks) ning kohendame andmestikku samamoodi nagu kestuste puhul tegime: arvutame häälikute algus- ja lõpuaegadest kestused, eemaldame valimist valesti klassifitseeritud ja venitatud sõnad ning grammatilised sõnad.

5.2. Põhitooni väärtuste normaliseerimine

Kuna meil on andmeid erinevatelt kõnelejatelt, kelle häälekõrgus erineb, siis on vaja neid võrdlemiseks **normaliseerida**. Normaliseerimiseks ei ole kahjuks ühte head meetodit, sest selle käigus kaotame alati midagi ära algsest varieeruvusest ja tekitame andmetesse moonutusi.

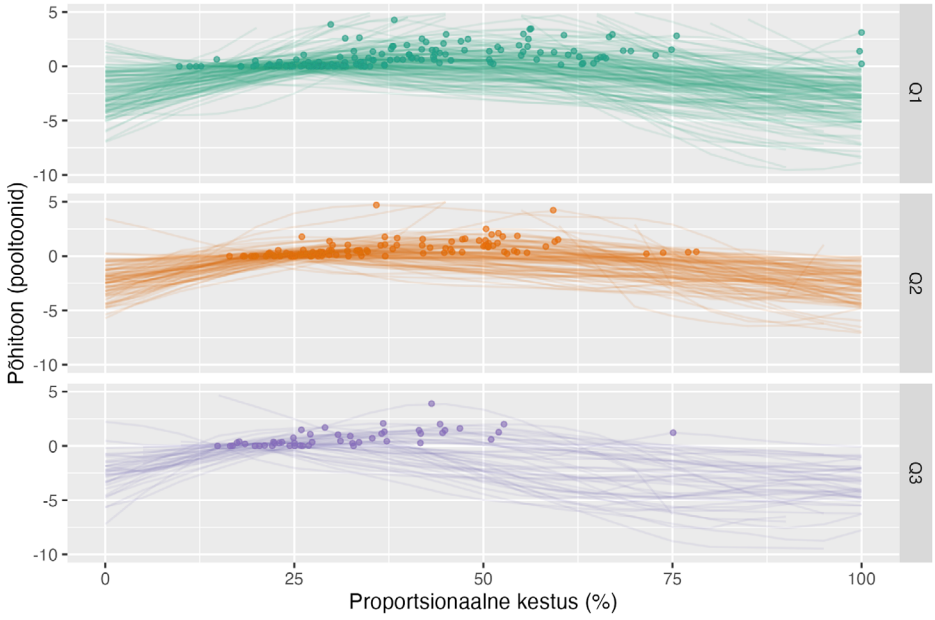
Hertsiskaala on inimese taju jaoks logaritmiline: näiteks muutust 50 hertsilt 100 hertsile tajume me sama suurena nagu muutust 200 hertsilt 400 hertsile, mistõttu kõrgemates sagedusvahemikes on ka varieerumine arvuliselt suurem. Selleks, et logaritmilisust eemaldada, on mõistlik teisendada andmed lineaarsele skaalale. Hääle põhitooni kirjeldades teisendatakse hertsides mõõdetud põhitoon tihti pooltoonideks. **Pooltooniskaala** on muusikas kasutatav intervallskaala, mis jagab oktavi 12 pooltooniks, võttes tihti nullpunktiks esimese oktavi la. Kui me tahame aga normaliseerida kõne põhitooni kõnelejate vahel, siis võime nullpunktiks võtta iga kõneleja keskmise häälekõrguse.

Siin uurimuses vaatame põhitooni liikumist üksikute sõnade piires, mistap lisaks kõnelejate individuaalsele häälekõrgusele oleks tarvis normaliseerida ka lauseintonatsioonist tingitud varieerumist – kui üldiselt algab lausung kõrgema põhitooniga ja intonatsioonifraasi jooksul põhitoon langeb (Asu, Lippus, Salveste, jt 2016), siis siin uurimuses meid ei huvita, kas sõna asus fraasi alguses, keskel või lõpus. Seetõttu teisendame iga sõna põhitooni pooltoonideks selle sõna alguspunkti suhtes. Või õigemini, kuna sõna päris algus võib olla helitu (algab näiteks klusiiliga), siis võtame normaliseerimise aluseks esimese vokaali alguse. Hertside pooltooniskaalale teisendamiseks kasutame funktsiooni *f2st()* paketist `hqmisc` (Quene 2022). Andmestikus on põhitoon mõõdetud vokaalide algusest ja lõpust ning lisaks ka kontuur kogu sõna jooksul (21-st võrdsete vahedega punktist). Pooltooniskaala teisenduseks teisendame iga sõna kontuuri V1 alguspunkti suhtes.

5.3. Põhitoonikontuuride visualiseerimine

Järgnevalt vaatame põhitoonikontuure joonistel, et nende kujust ülevaadet saada. Joonisel 4 on välja joonistatud kõikide sõnade individuaalsed põhitoonikõverad ning täpiga on tähistatud põhitooni maksimumpunkt. Kuigi kontuurid on

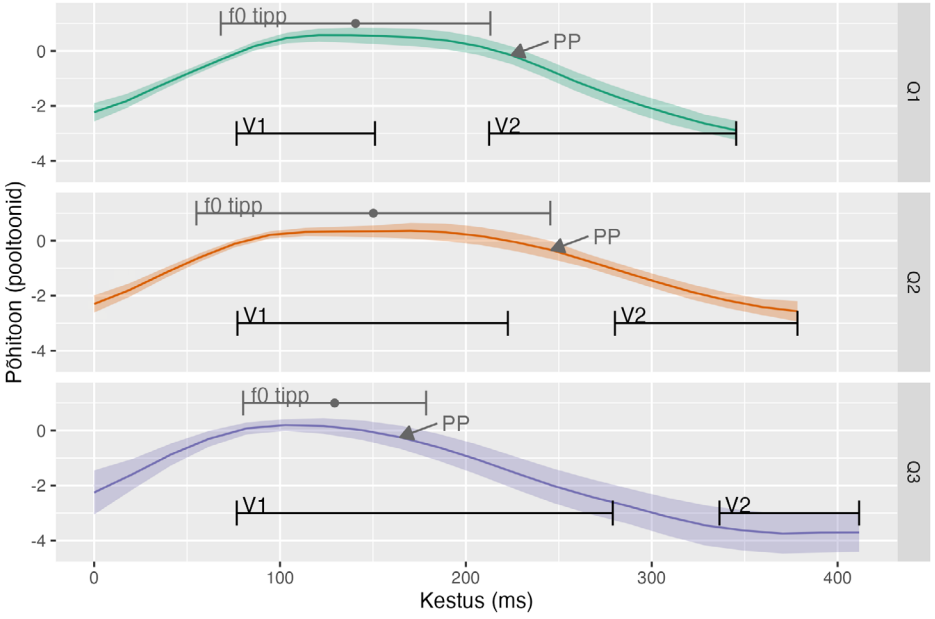
võrdlemisi sarnased kõigis väldetes, siis mõnevõrra joonistub juba välja, et kolmanda välte puhul on põhitooni tipp x-teljel natuke varasem kui esimeses ja teises vältes, kuigi varieerumine on kaunis suur. Joonisel on **kestustelg normaliseeritud** sõna kogukestuse suhtes nii, et kõik kontuurid on sama pikkusega. Seega kui võrrelda tipu asukohta absoluutteljel, võib erinevus tipu asukoha osas kolmandavältelistes sõnades ära kaduda.



Joonis 4. Üksikute sõnade põhitoonikontuurid. Põhitooni väärtused on teisendatud pooltooniskaalale V1 alguse suhtes; kestusteljel normaliseeritud aeg sõna kogukestuse suhtes; täpiga on tähistatud kontuuri maksimumpunkt

Joonisel 5 on põhitoonikontuurid skaleeritud **absoluutajateljele** ning selleks, et vähendada müra üksikute kontuuride varieerumisest (üksikute sõnade jooned joonisel 4), on joonisel kontuurid keskmistatud ja esitatud usaldusvahemiku piiridega (umbes sama, mis vurrud karpdiagrammil). Et kontuuri sõna struktuuriga paremini seostada, on kontuuride all näidatud vokaalide paiknemine. Samuti on joonisel osutatud põhitooni tipu asukoht standardhälbe piiridega. Nagu ka joonilt 4 nägime, on põhitooni tipu asukoht sõnati väga varieeruv. Lippus jt (2013) kirjeldavad seda kui kõrget platood, mis esimeses ja teises vältes püsib pea kogu esimese silbi jooksul ning kolmandas vältes esimese silbi esimeses pooles. Mõõtes tippu kui põhitooni absoluutset maksimumi võibki see sattuda võrdlemisi juhuslikku punkti kõrge platoo jooksul. Seetõttu oleks mõttekam tipu asemel kirjeldada

pöördepunkti, kust kõrge põhitoon pöördub langusele. Joonisel 5 on pöördepunkt (*PP*) tähistatud noolega ning see on siin defineeritud kui punkt, kus põhitoon on maksimumväärtusest langenud 0,5 pooltooni võrra. Nagu näeme, on esimeses ja teises vältes pöördepunktid võrdlemisi kohakuti, paiknedes teise silbi alguses, kolmandas vältes on pöördepunkt märksa varasem, asudes esimese silbi keskel. Tajukatsed on näidanud, et kolmanda välte puhul on oluline tunnus see, et põhitoon esimese silbi jooksul märgatavalt langeb (Lippus, Pajusalu & Allik 2011).

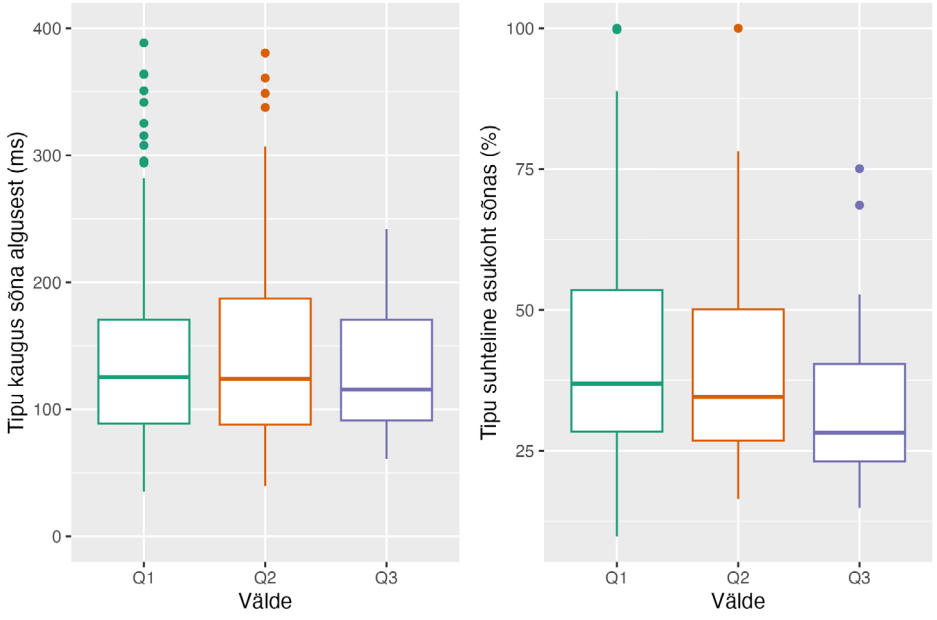


Joonis 5. Keskmised põhitoonikontuurid. Kontuuri kohal on näidatud põhitooni tipu asukoht standardhälbe piiridega ja languse algust tähistav pöördepunkt (*PP*); kontuuri all on näidatud esimese ja teise vokaali paiknemine

5.4. Põhitooni tipu kaugus välteti

Jooniselt 5 nägime, et pöördepunkt võiks välte kirjeldamiseks olla parem mõõdik kui põhitooni absoluutne tipp, aga kuna tippu on lihtsam defineerida, siis testime järgnevalt, kas tipu asukoht on ka statistiliselt oluline välte eristaja. Joonise 5 (ja ka varasemate uurimuste põhjal) võib väita, et sõna piires on kolmandas vältes põhitooni tipp varasem kui esimeses ja teises vältes. Testime põhitooni tipu suhtelist asukohta sõna piires jällegi lineaarse segamudeliga. Teeme kaks analüüsi: esmalt vaatame tipu absoluutset kaugust sõna algusest (joonisel 6 vasakul) ning seejärel

teisendame tipu kauguse sõna algusest tipu suhteliseks asukohaks esimeses silbis (joonisel 6 paremal).



Joonis 6. Põhitooni tipu kaugus sõna algusest (vasakul) ja tipu proportsionaalne asukoht sõnas (paremal)

Tabelist 10 ja joonise 6 parempoolselt paneelilt näeme, et millisekundites mõõdetud tipu absoluutkaugus sõna algusest ei ole ei esimese ja teise välte (β = -0,015, t = -0,230, p = 0,819) ega ka teise ja kolmanda välte vahel (β = 0,087, t = -0,946, p = 0,347) oluliselt erinev – keskmiselt jääb tipp esimeses vältes 140 ms, teises vältes 150 ms ja kolmandas 130 ms kaugusele, aga varieeruvus on väga suur.

Tabel 10. Põhitooni tipu absoluutkaugust kirjeldava segamudeli fikseeritud mõjud

	Hinnang	**Standard-viga**	**Vabadus-astmete arv**	**t-väärtus**	**Pr(>|t|)**
(vabaliige)	-2,025	0,118	4,215	-17,153	<0,001
välde Q1	-0,015	0,067	34,674	-0,230	0,819
välde Q3	-0,082	0,087	77,604	-0,946	0,347

Kui aga vaadata tipu suhtelist asukohta sõnas (joonis 6 paremal, tabel 11), siis näeme, et see ei ole oluliselt erinev esimese ja teise välte vahel (β = 0,048, t = 0,966, p = 0,379), aga kolmandas vältes on tipp natuke varasem kui teises vältes (β = -0,179, t = -2,562, p = 0,019).

Tabel 11. Põhitooni tipu suhtelist asukohta kirjeldava segamudeli fikseeritud mõjud

	Hinnang	**Standard-viga**	**Vabadus-astmete arv**	**t-väärtus**	**Pr(>|t|)**
(vabaliige)	3,583	0,042	4,353	86,190	<0,001
välde Q1	0,048	0,050	4,876	0,966	0,379
välde Q3	-0,179	0,070	19,710	-2,562	0,019

6. Vokaalikvaliteet

Kui varasemad eesti vokaalisüsteemi kirjeldused on leidnud, et vokaalikvaliteet vältega seoses ei varieeru (Lehiste 1960; Liiv 1962) või siis varieerub võrdlemisi vähe (Eek & Meister 1998), siis Lippus jt (2013) tulemused näitasid, et esiteks on rõhulised vokaalid teises ja kolmandas vältes siiski oluliselt perifeersemad kui esimeses vältes, teiseks on rõhutud vokaalid üldiselt redutseeritumad kui rõhulises silbis, ning kolmandaks on rõhutud vokaalid kolmandas vältes oluliselt redutseeritumad kui esimeses ja teises vältes. Vokaalikvaliteeti saab iseloomustada nende formantide järgi. Selleks viime läbi formantanalüüsi.

6.1. Formantanalüüs

Praati toimetamisaknas kuvatakse formante punaste täpikestena spektrogrammi peale ja skaala kuvatakse vasakul servas (vt joonis 1). Formantanalüüsi tulemused ja täpsus sõltuvad sellest, mis sagedusvahemikust mitut formanti otsime. See peaks olema kooskõlas kõneleja kõnetrakti pikkusega. Formantanalüüsi seadeid saab sättida menüüs *Formants → Formant settings...*. Kõige üldisemalt võiks seada formantanalüüsi laeks (ingl *formant ceiling*) meeskõnelejatel 5000 Hz ja naiskõnelejatel 5500 Hz. Kõige rohkem vigu teeb automaatne formantanalüüs tagavokaalidega /u/ ja /o/, mille keele kõrgust väljendav sagedus F1 ja keele tagapoolsust väljendav F2 on väga lähestikku. Nende vokaalide puhul aitaks automaatanalüüsi tulemusi parandada, kui kohandada seadeid igale kõnelejale ja igale vokaalikategooriale (vt nt Escudero jt 2009), kuid piirdume siin praegu ainult meeste ja naiste eristamisega.

Klassikaliselt kasutatakse vokaalikvaliteedi kirjeldamisel ***sihtväärtuse*** mõistet: igal häälikul on mingid artikulatoorsed sihid, mille poole liigutakse, ja mingi aja

jooksul hääliku algusest saavutatakse siht, mida natuke aega hoitakse, et siis edasi liikuda järgmise hääliku suunas. Võiks arvata, et häälikupiiridel on siirded ühelt häälikult teisele ning keskosas on just sellele häälikule omane stabiilne periood. Tegelikult erilist stabiilsust ei ole, pidev liikumine toimub kogu aeg. Aga kui me soovime monoftonge kirjeldada vokaaliklassiti ja mitte arvestada väga palju seda, mis konsonantide kontekstis nad on, siis oleks mõistlik mõõta formante vokaali keskosast ja piisab ühest väärtusest ühe vokaali kohta. Kui me aga tahame kirjeldada diftonge või ka just nimelt konsonandi konteksti, siis oleks mõistlik mõõta formantide kontuuri dünaamiliselt terve vokaali ulatuses.

Teeme siin lihtsamal viisil: mõõdame ühe punkti väärtused vokaali keskelt. Kui Praati toimetamisaknas vertikaalne kursoriosa panna vokaali keskele ja horisontaalne kursoriosa asetada kohakuti formantidega, võib formandi väärtust lugeda vertikaalskaalal spektrogrammist vasakul. Siin analüüsis aga mõõdame formantväärtusi skriptiga vokaali keskpunktist, mille arvutame välja TextGrid-failis märgitud häälikupiiride põhjal. Kuna formantanalüüs teeb siiski küllaltki palju vigu, siis võtame formantväärtused mitte vokaali keskpunkti ühest ajahetkest, vaid vokaali keskosa keskmised väärtused. Nii anname üksikutele hälbivatele analüüsipunktidele väiksema kaalu.

Jooksutame Praatis skripti *korp_opik_fonkorp_formandid_vokaali_keskelt.praat* ja loeme tulemuseks saadud faili *formandid_tulemused.txt* R-i.

Koondame vokaali märgendeid nii, et märgend oleks ainult ühe tähemärgiga (kaotame ära pikkuse ja lisakvaliteedimärgid) ning teisendame SAMPA transkriptsiooni IPA sümboliteks, et sümbolid näeks kenamad välja. Tabel 12 annab ülevaate sellest, kui palju erinevaid vokaale andmestikus on.

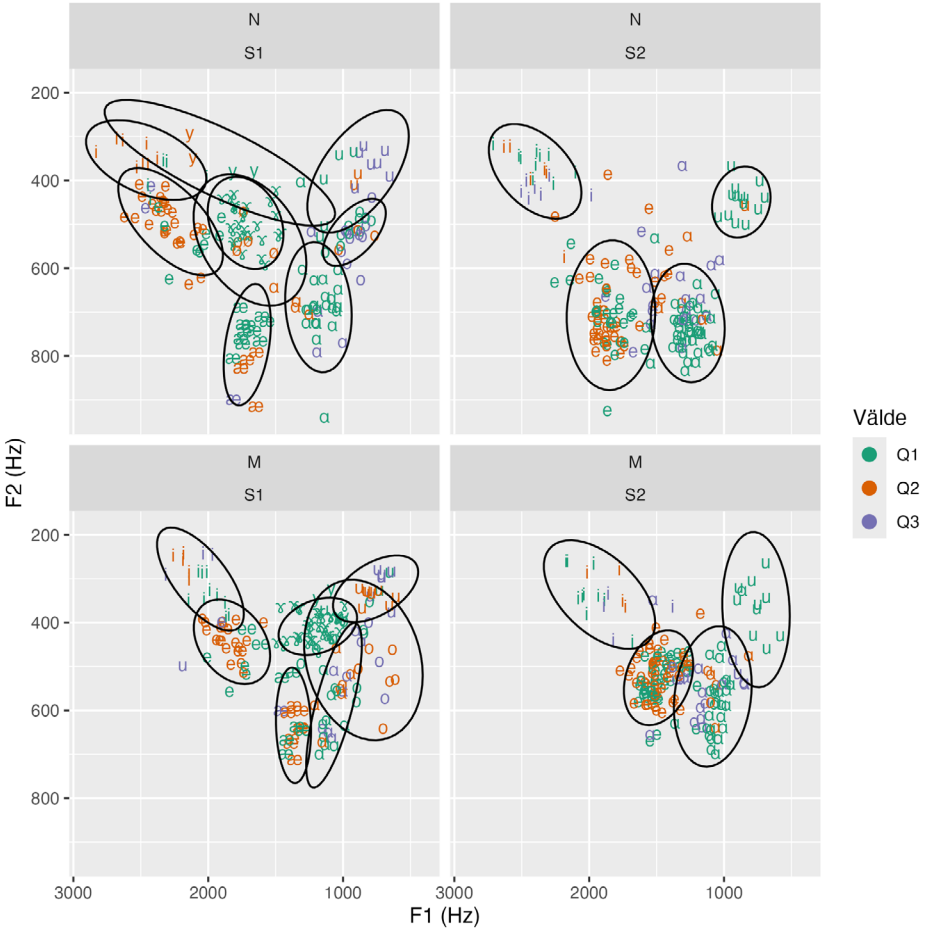
Tabel 12. Vokaalide arv

Välde	Silp	ɑ	e	i	u	o	y	æ	ʏ	ø
Q1	S1	33	17	15	8	20	5	17	69	
Q1	S2	71	68	20	24			1		
Q2	S1	11	48	13	10	10	2	13		4
Q2	S2	10	89	10	1	1				
Q3	S1	12	3	3	15	15		2		
Q3	S2	34	4	11	1					

Kuna meil on võrdlemisi väike andmestik, siis paratamatult ei esine siin kõiki vokaale kõigis vältusastmetes. Vähemsagedastest vokaalidest on andmestikus olemas /y/ ainult esma- ja teisevälteliste, /ʏ/ ainult esmavälteliste ja /ø/ ainult teisevälteliste sõnade rõhulistes silpides.

Järgsilbis on /æ/ märgitud ühes sõnas (*pähe*), aga kuna see on fonoloogiliselt /e/, siis võime selle ümber kategoriseerida.

Joonisel 7 on nüüd kujutatud vokaalid formantruumis hertsiskaalal, meeste ja naiste andmed eraldi.



Joonis 7. Mees- ja naiskõnelejate rõhuliste (S1) ja rõhutute (S2) silpide vokaalid F1-F2 formantruumis

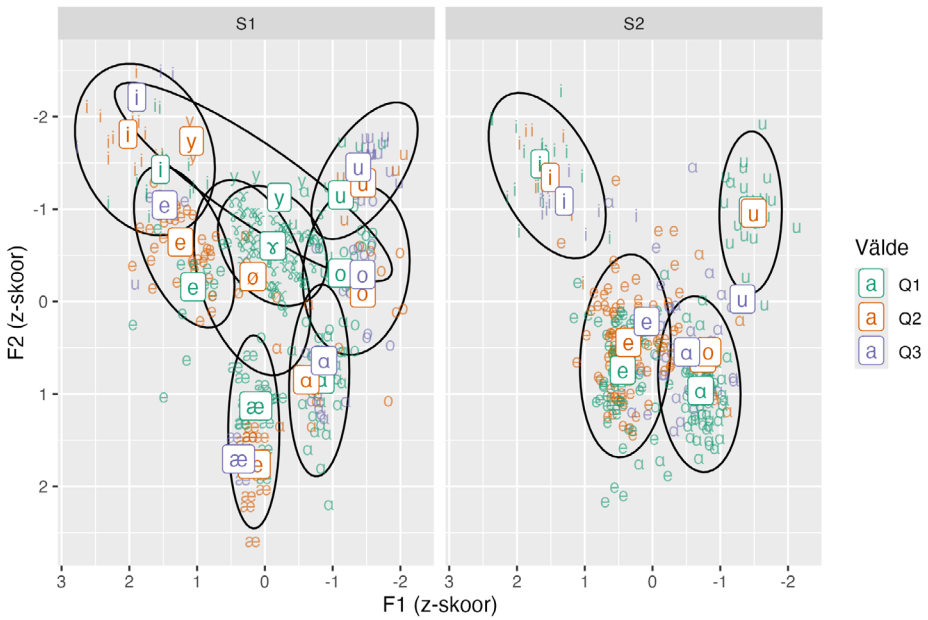
6.2. Formantväärtuste normaliseerimine

Kui põhitooniandmeid teisendatakse enamasti pooltooniskaalale, siis formantväärtuste esitamiseks on peale hertsiskaala ka teisi skaalasid, nt barkide, *mel*'ide või

ERB skaala, mis on inimtaju jaoks lineaarsed. Andmete teisendamine hertsidest barkideks muudab jaotust normaaljaotusele lähemaks, kuid ei kaota ära kõnelejatevahelisi kõnetrakti suurusest tulenevaid erinevusi.

Nagu põhitoon, varieeruvad ka formantväärtused kõnelejati. Suur erinevus on meeste ja naiste vahel, mis ei ole tingitud mitte erinevustest keelekasutuses, vaid kõnetrakti mõõtmetest. Kuna naistel on kõnetrakt veidi lühem kui meestel, siis on nende formantväärtused natuke suuremad. Aga ka samast soost kõnelejate vahel on varieerumist. Kui me tahame vaadelda vokaalide paiknemist kõneleja füsioloogilistest eripäradest sõltumata ja kujutada vokaaliruumi kõigi jaoks ühel joonisel, siis peaksime **formantandmeid normaliseerima**.

Üks normaliseerimisviis, kuidas vähendada kõnelejatevahelisi erinevusi nii, et jääks alles vokaalide omavahelised suhted vokaaliruumi sees, oleks teisendada väärtused **z-skoorideks**. Selle käigus teisendatakse kõneleja formantväärtused tema vokaaliruumi standardhälbe ühikuteks (vt allolevat valemit, kus x_i on formantväärtus, $\bar{x}$ on kõneleja formantväärtuste keskmine ning σ on standardhälve). Seda meetodit tuntakse foneetikas ka Lobanovi normaliseerimise all (Boris Lobanovi (1971) järgi, kes seda esimesena vokaalide kirjeldamisel kasutas, kuigi meetod on signaalitöötluses väga laialt kasutusel). Sellel normaliseerimisel on ainult üks häda: ei ole enam sisukat skaalat. Z-skoor kirjeldab ühe andmepunkti kaugust keskväärtusest rühma standardhälvetes. Seega on z-skooride ühikuks üks standardhälve ning väärtused võivad varieeruda umbes vahemikus -3 (standardhälvet) ja 3 (standardhälvet).



Joonis 8. Normaliseeritud vokaalide formantväärtused F1-F2 formantruumis

$$z = \frac{(x_i - \bar{x})}{\sigma}$$

Joonis 8 kujutab vokaalide normaliseeritud formantväärtusi formantruumis ilma sugu eristamata. *Z*-skoorideks teisendatud formantväärtuste puhul saab paremini võrrelda erinevate kõnelejate vokaalide paiknemist üksteise suhtes.

6.3. Vokaalide redutseerumine ja välde

Kuidas aga mõõta vokaalikvaliteedi seost vältega? Vokaalikvaliteet on kirjeldatud mitme formandi väärtusega, mis on seotud vokaalikategooriaga ja redutseerumine mõjutab eri vokaalikategooriate puhul väärtuste muutumist eri suundades: kui madal tagavokaal /ɑ/ redutseerub, siis tema F1 kahaneb ja F2 kasvab, aga kui kõrge eesvokaal /i/ redutseerub, siis tema F1 kasvab ja F2 kahaneb. Lahendus oleks üritada **normaliseerida vokaali kvaliteeti** vokaalikategooria suhtes.

Vokaalide redutseerumist üldiselt iseloomustab liikumine vokaaliruumi keskpunkti suunas. Et saada kahemõõtmelises (F1-F2) ruumis kirjeldatud andmed ühemõõtmeliseks, võiks mõõta iga vokaali kaugust vokaaliruumi keskpunktist. Selleks arvutame eukleidilise kauguse. **Eukleidiline kaugus** on otsekaugus kahe punkti vahel mitmemõõtmelises ruumis ja selle arvutamiseks kasutame Pythagorase teoreemi. Allolevat valemit vaadeldes võib kujutleda, et kui ja on vastavalt kõneleja ühe vokaali F1 ja F2 väärtus, siis ja on selle kõneleja formantruumi keskpunkti F1 ja F2 väärtused.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

Eukleidilise kauguse arvutamiseks on parem pöörduda tagasi mõõdetud väärtuste juurde ja mitte kasutada z-skoorideks teisendatud väärtusi, sest algandmetel on ühikud, mille põhjal saab hinnata muutuse suurust. Logaritmilisuse eemaldamiseks teisendame andmed hertsiskaalalt **barki skaalale**, mis arvestab paremini tajuga. Hertside barkideks teisendamiseks leiab käsu *f2bark()* paketist `hqmisc` (Quene 2022).

Nüüd arvutame iga vokaali kauguse selle kõneleja kõigi vokaalide keskmisest. See ei pruugi küll antud andmestiku puhul väga head tulemust anda, sest eesti keele vokaaliruum on ise pisut ebasümmeetriline ja meil on tasakaalustamata hulk vokaale igalt kõnelejalt, mistõttu neist keskmine võib olla mõnevõrra juhuslik. Lippus jt (2013) artiklis valiti iga kõneleja vokaaliruumi keskpunkti leidmiseks ainult nn nurgavokaalid /ɑ i u/. Vokaalide kaugus formantruumi keskpunktist välteti on näidatud joonisel 9. Jällegi testime tulemusi lineaarse segamudeliga (mille väljund rõhulise silbi vokaalidega on tabelis 13, rõhutu silbi vokaalidega tabelis 14).

Tabel 13. Rõhulise vokaali kvaliteedi segamudeli fikseeritud mõjud

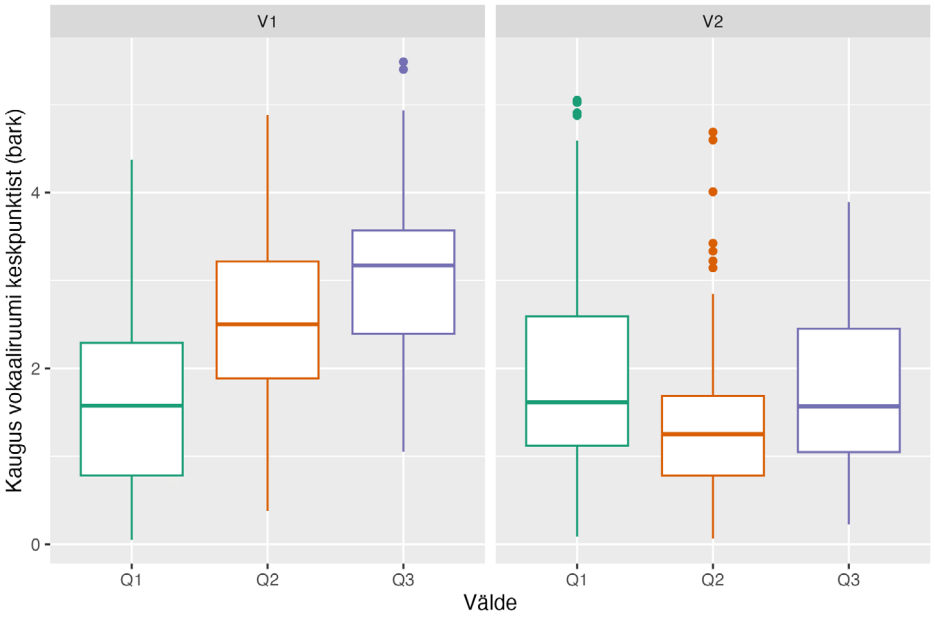
	Hinnang	Standard-viga	Vabadus-astmete arv	t-väärtus	Pr(>|t|)
(vabaliige)	2,943	0,182	45,331	16,211	<0,001
välde Q1	-0,851	0,214	112,315	-3,984	<0,001
välde Q3	0,101	0,185	322,385	0,547	0,584

Rõhulises silbis on esmavälteliste sõnade vokaalid tsentraalsemad kui teises vältes (β = -0,851, t = -3,984, p = <0,001). Erinevus teise ja kolmanda välte vahel ei ole statistiliselt oluline (β = 0,101, t = 0,547, p = 0,584).

Tabel 14. Rõhutu vokaali kvaliteedi segamudeli fikseeritud mõjud

Hinnang	Standard-viga	Vabadus-astmete arv	t-väärtus	Pr(>|t|)	
(vabaliige)	1,750	0,214	24,160	8,160	<0,001
välde Q1	0,452	0,233	112,043	1,939	0,055
välde Q3	-0,156	0,197	327,138	-0,790	0,430

Rõhutu silbi kvaliteedis ei ole siin andmestikus väldete vahel olulisi erinevusi (vt tabel 14).



Joonis 9. Rõhulise ja rõhutu silbi vokaalide kaugus formantruumi keskpunktist barkides väldete kaupa

Kokkuvõte

Siin peatükis analüüsisime kahesilbiliste sõnade häälikukestusi, põhitooni ja vokaalikvaliteeti, et kirjeldada eesti keele välteid. Andmed kogusime Praati skriptidega eesti keele spontaanse kõne foneetilisest korpusest pärit failidest. Nagu paljud varasemad uurimused on näidanud, eristab eesti keele kolme väldet kõige paremini rõhulise ja rõhutu silbi (vokaalide või silbiriimide) kestuste suhe, mis võtab kokku ühe mõõdikuna rõhulise silbi pikenemise ja rõhutu silbi lühenemise, mis vältega kaasneb.

Lisaks kestustele eristab välteid põhitooni langus, mis on absoluutajas mõõdetuna teisevältelistes sõnades veidi hilisem kui esma- ja kolmandavältelistes sõnades. Kui aga vaadata põhitooni joondumist suhtelisel ajateljel sõna piires, siis eristub esimesest ja teisest kolmas välde, kus langus toimub varem. Põhitooni tipp on esimeses vältes tihti teise silbi alguses, teises vältes esimese silbi lõpus ja kolmandas vältes esimese silbi esimeses pooles.

Vokaali kvaliteedi varieerumise seotust vältega nägime siin uurimuses ainult rõhulise silbi vokaali puhul ja tulemused olid kooskõlas varasemate tulemustega: esimeses vältes oli see teise vältega võrreldes oluliselt tsentraliseeritum, teise ja kolmanda välte vahel erinevust ei olnud. Rõhutus silbis välteerinevust ei õnnestunud

välja tuua. Peab siiski arvestama, et siin kasutatud valim on väike – ainult nelja kõneleja materjal –, mistõttu võivad tulemust rohkem mõjutada üksiku kõneleja individuaalsed eripärad, kõnesituatsioon, sõna esinemiskontekst jm faktorid, mida siin uurimuses ei kontrollitud.

Näidisuurimuse valmimist on toetanud projekt EKKD-TA6 „Liigutustest kõneni: suulise eesti keele multimodaalne analüüs" (2024–2027).

Kirjandus

Asu, Eva Liina, Pärtel Lippus, Karl Pajusalu & Pire Teras. 2016. *Eesti keele hääldus* (Eesti keele varamu 2). Tartu: Tartu Ülikooli Kirjastus. http://hdl.handle.net/10062/57960.

Asu, Eva Liina, Pärtel Lippus, Nele Salveste & Heete Sahkai. 2016. F0 declination in spontaneous Estonian: Implications for pitch-related preplanning in speech production. *Speech Prosody 2016*, 1139–1142. Boston, MA: Boston University. https://doi.org/10.21437/SpeechProsody.2016-234.

Barreda, Santiago. 2023. phonTools: Functions for phonetics in R.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Boersma, Paul & David Weenink. 2023. Praat: Doing phonetics by computer. https://www.praat.org.

Bořil, Tomáš & Radek Skarnitzl. 2016. Tools rPraat and mPraat. Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (toim), *Text, Speech, and Dialogue: Proceedings of the 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016*, 367–374. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-45510-5_42.

Coretta, Stefano. 2024. *speakr: A wrapper for the phonetic software Praat*. Manual. https://CRAN.R-project.org/package=speakr.

De Looze, Céline & Daniel Hirst. 2008. Detecting changes in key and range for the automatic modelling and coding of intonation. *Proceedings of the 4th International Conference on Speech Prosody*, 135–138. Campinas, Brazil. https://doi.org/10.21437/SpeechProsody.2008-32.

Eek, Arvo & Einar Meister. 1998. Quality of standard Estonian vowels in stressed and unstressed syllables of the feet in three distinctive quantity degrees. *Linguistica Uralica* 34(3). 226–233. https://doi.org/10.3176/lu.1998.3.11.

Escudero, Paola, Paul Boersma, Andréia Schurt Rauber & Ricardo A. H. Bion. 2009. A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America* 126(3). 1379–1393. https://doi.org/10.1121/1.3180321.

Jadoul, Yannick, Bill Thompson & Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71. 1–15. https://doi.org/10.1016/j.wocn.2018.07.001.

Krull, Diana. 1997. Prepausal lengthening in Estonian: Evidence from conversational speech. Ilse Lehiste & Jaan Ross (toim), *Estonian Prosody: Papers from a Symposium*, 136–148. Tallinn: Institute of Estonian Language.

Lehiste, Ilse. 1960. Segmental and syllabic quantity in Estonian. Thomas A. Sebeok (toim), *American Studies in Uralic Linguistics* (Uralic and Altaic Series), 21–82. Bloomington: Indiana University Publications.

Liiv, Georg. 1961. Eesti keele kolme vältusastme vokaalide kestus ja meloodiatüübid. *Keel ja Kirjandus* 7–8. 412–424, 480–490.

Liiv, Georg. 1962. On the quantity and quality of Estonian vowels of three phonological degrees of length. Antti Sovijärvi & Pentti Aalto (toim), *Proceedings of the Fourth International Congress of Phonetic Sciences*, 682–687. The Hague: Mouton & Co.

Lippus, Pärtel. 2026. *Akustilised meetodid foneetikas. Kõneanalüüs programmiga Praat*. Tartu: Tartu Ülikooli Kirjastus. https://kodu.ut.ee/~partel/akustilised-meetodid-foneetikas/.

Lippus, Pärtel, Kätlin Aare, Anton Malmi, Tuuli Tuisk & Pire Teras. 2023. Phonetic corpus of Estonian spontaneous speech v1.3. Institute of Estonian and General Linguistics, University of Tartu. https://doi.org/10.23673/re-438.

Lippus, Pärtel, Eva Liina Asu, Pire Teras & Tuuli Tuisk. 2013. Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian. *Journal of Phonetics* 41(1). 17–28. https://doi.org/10.1016/j.wocn.2012.09.005.

Lippus, Pärtel, Karl Pajusalu & Jüri Allik. 2011. The role of pitch cue in the perception of the Estonian long quantity. Sónia Frota, Gorka Elordieta & Pilar Prieto (toim), *Prosodic Categories: Production, Perception and Comprehension*, 231–242. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0137-3_10.

Lippus, Pärtel & Juraj Šimko. 2015. Segmental context effects on temporal realization of Estonian quantity. Maria Wolters, Judy Livingstone, Bernie Beattie, Rachel Smith, Mike MacMahon, Jane Stuart-Smith & Jim Scobbie (toim), *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5. Glasgow: University of Glasgow.

Lobanov, Boris M. 1971. Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America* 49(2B). 606–608. https://doi.org/10.1121/1.1912396.

Piits, Liisi & Mari-Liis Kalvik. 2017. Varieeruva vältega sõnade häälduskuuringud kõnesünteesi teenistuses. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 13. 123–140. https://doi.org/10.5128/ERYa13.08.

Quene, Hugo. 2022. *hqmisc: Miscellaneous convenience functions and dataset*. Manual. https://CRAN.R-project.org/package=hqmisc.

R Core Team. 2023. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Nimeüksuste märgendamine 19. sajandi vallakohtu protokollides

Kadri Muischnek, Siim Orasmaa

Lühikokkuvõte

Selles peatükis käime läbi nimeüksuste tuvastamise ülesande lahendamise etapid 19. sajandist pärit kõikuva kirjaviisi ja murdejoontega tekstides. Seega saavad siin kokku kaks teemat: nimeüksuste tuvastamine ja tänapäeva standardkirjakeelest erineva tekstiga töötamine. See on eelkõige rakenduslik töö: eesmärgiks on luua võimalikult hea nimeüksuste korpus masinõppepõhiste nimeüksuste tuvastamise mudelite treenimiseks.

Nimeüksuste märgendamise näitel tutvume (käsitsi) märgendamise põhimõtete ja etappidega, sh märgendajatevahelise kooskõla kui märgenduse kvaliteedi hindamise vahendiga.

Nimeüksuste suhtes märgendatud korpus on meie näites treeningandmestikuks masinõppepõhise nimeüksuste tuvastaja loomisel, aga selline käsitsi märgendatud korpus võib olla ka omaette eesmärgiks ja sellisena uurimistöö materjaliks. Uurimus paigutub digihumanitaaria ja keeletehnoloogia valdkonda.

1. Nimeüksuste tuvastamine

Nimeüksuste tuvastamine (ingl *named entity recognition* ehk NER) on automaatse infoeraldamise alamülesanne, mille eesmärgiks on tekstist nimede leidmine ja klassifitseerimine. Tüüpiliselt eristatakse **isikunimesid** (nt *Alar Karis*), **kohanimesid** (nt *Kadrina*) ja **organisatsiooninimesid** (nt *AS Eesti Energia*), aga sõltuvalt infoeraldamise eesmärkidest võib nimekategooriaid olla teisigi (nt tootenimed, sündmuste nimed jm-d).

Nimeüksuste märgenduse rakendusvõimalused on laiad, alates infootsingust ja meediamonitooringust (nt isikutega seotud dokumentide või meediakajastuse sirvimine) kuni dokumentide semantilise esituse ja analüüsini (nimedevaheliste seoste tuvastamine, nt millised isikud mingis kohas samaaegselt viibisid ja üldisemalt, millised sündmused on nimede mainimisega seotud). Keerukamate

rakenduste puhul ei piisa siiski ainult nimede tuvastamisest, vaid vaja on ka täiendavat analüüsi: 1) nimede algkujule viimist ehk **lemmatiseerimist** (nt nime *Alar Karisele* algvormiks on *Alar Karis*), 2) **nimede samaviitelisuse lahendamist** (ingl *coreference resolution*), kus koondatakse kokku ühte ja sama olemit/entiteeti mainivad nimed (nt nimekujud *Alar Karis, A. Karis* ja *president Karis* viitavad kõik president Alar Karisele) ja viitesõnad (nt *ta*, *tema* viitavad tekstis mainitud isikule).

Selleks, et nimeüksuseid saaks automaatselt tuvastada, tuleb kõigepealt luua **käsitsi märgendatud nimeüksuste korpus**, mida treeningmaterjalina kasutades saab luua masinõppepõhise tuvastaja. Eesti keele jaoks on juba olemas mitu märgendatud nimeüksustega korpust, nt Nimeüksuste korpus 2013 (Laur 2013) ja Uus Eesti nimeüksuste korpus 2021 (Sirts 2021), mis koosnevad tänapäeva normkirjakeelsetest ajalehetekstidest. Nendel korpustel väljatreenitud nimeüksuste tuvastaja näidisrakendus[1] on saadaval ka veebis.

Tuleb aga mainida, et kui automaattuvastaja on treenitud toimetatud kirjakeelel ja ainult üht tekstiliiki esindavatel tekstidel (ajalehetekstid), siis selle rakendamine treeningkorpusest oluliselt erineval keelesisendil (nt vanal kirjakeelel) ei ole väga tulemuslik. Kvaliteetse tulemuse saavutamiseks tuleb nimeüksuste tuvastaja välja treenida keelematerjalil, mis on võimalikult sarnane programmi eeldatava sisendiga. Käesolev uurimus tutvustabki ettevalmistavaid töid ja eksperimente nimeüksuste automaattuvastamiseks **19. sajandi vallakohtuprotokollidest**.

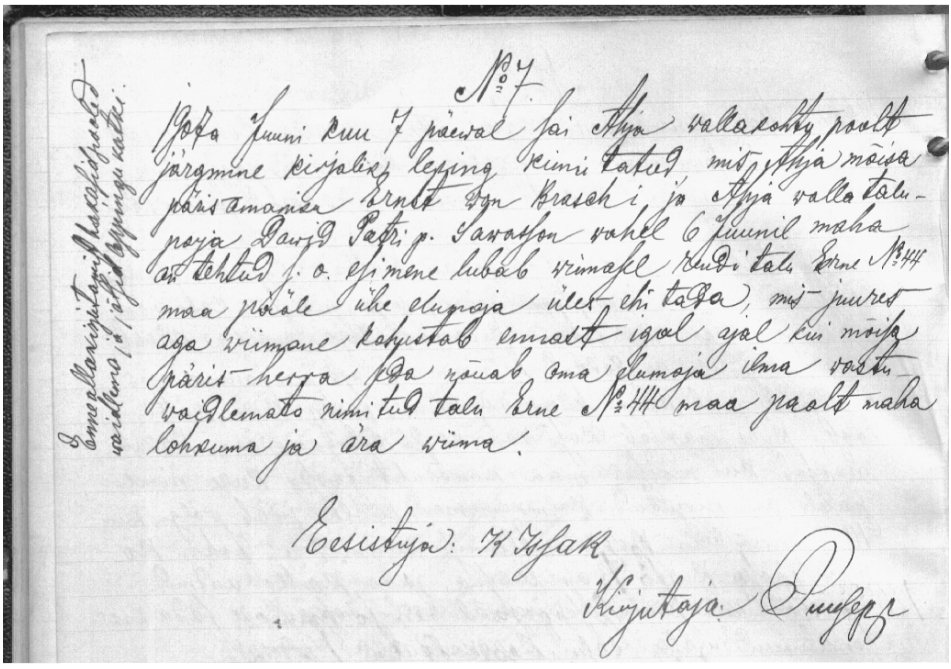
2. Materjali tutvustus ja taust

Vallakohtud ehk kogukonnakohtud olid 19. sajandi algusest kuni 1918. aastani tegutsenud talurahvakohtud, mis tegelesid talupojaseisusest isikute üleastumiste, nõuete, vaidluste ja perekonnasuhetega. Kohtumenetlus oli suuline ja eestikeelne, aga kohtuasja osapooled, sisu ja otsus pandi kirja kokkuvõtlikku kohtuprotokolli.

Rahvusarhiivis on hoiul ligi 450 kogukonnakohtu protokolli- ja lepinguraamatud. Nende digiteerimiseks on Rahvusarhiiv algatanud **ühisloomeprojekti**[2], mille käigus oli 2024. aasta juunikuu lõpuks digiteeritud u 156 500 protokolli (vt protokolli näidet jooniselt 1).

[1] https://ner.tartunlp.ai/

[2] https://www.ra.ee/vallakohtud/



PROTOKOLL

Ahja: Lepinguraamat (1904-1908)

Leidandmed	EAA.3257.1.9
Kaader	45
Daatum	07.06.1907
Protokolli number	7

1907a Juuni kuu 7 päewal sai Ahja wallakohtu poolt järgmine kirjalik leping kinnitatud, mis Ahja mõisa pärisomaniku Ernst won Braschi ja Ahja walla talupoja Dawid Peetri p. Lawasson wahel 6 Juunil maha on tehtud s.o. esimene lubab wiimase rudi talu Erne N: 44 maa pääle ühe elumaja üles ehitada misjuures aga wiimane kohustab ennast igal ajal kui mõisapärisherra seda nõuab oma elumaja ilma vastuvaidlemata nimitud talu Erne N: 44 maa pealt maha lahkuma ja ära wiima.

Küljekommentaar: Enne allakirjutamist hakkasid pooled waidlema ja jätsid leppingu katki.

Joonis 1. Käsikirjaline Ahja vallakohtu protokoll ja sisestatud protokollitekst

Protokollide tekstid on valdavalt eestikeelsed, kuigi on ka saksa- ja venekeelseid protokolle. Digiteeritud eestikeelsete protokollide korpus on väärtuslik ressurss nii talurahva ajaloo kui eesti keele ja eesti kirjakeele ajaloo uurimiseks. Kõnealused tekstid on keeleliselt ja ortograafiliselt heterogeensed: keelekasutust on mõjutanud kirjutaja haridus ja murdetaust, kogukonnakohtute tegutsemisaega jääb üleminek vanalt kirjaviisilt uuele.

Trükitud tekstides muutus uus kirjaviis valdavaks 1870. aastatel, kuid kohtuprotokollides kasutatakse vana kirjaviisi kauem ja mõnes tekstis on kasutusel kaks kirjaviisi segamini.

Näiteks Kodavere vallakohtus 1876. aastal kirja pandud protokoll[3] on valdavalt uues kirjaviisis, kuigi kohanimi *Nina* on, ilmselt harjumusest, kirjutatud vana kirjaviisi kohaselt *Ninno*:

> *Tulli ette Ninno Küla - majaperemees ja kalakaubleja Fedor Jwanow Karelin ja kaebas, et tema kaubelnud kalapüüdja Pawel Nasarowi kääst ahvenad 96 Kop puud, …*

Kuid samas protokollis paar lauset edasi on kasutatud vana kirjaviisi:

> *Pawel Nasarow, Kallapüüdja, ettekutsutud, ütles:*

Selles peatükis tutvustatava ülesande – nimeüksuste tuvastamise – seisukohalt on oluline **algustäheortograafia**, sest suur algustäht peaks andma signaali selle kohta, et sõna on tõenäoliselt nimeüksuse osa. Pärisnimed algavadki nendes tekstides tavaliselt suure algustähega, kuid varasemates tekstides ei järgita seda reeglit nii järjekindlalt. Ent suure algustähega võib protokollides, ilmselt saksa keele eeskujul, alata ka sõna, mis pole pärisnimi:

> *Tulli Sare Metsa ülle Watja Waltmann ette ja kaibas et olla, Kuddina Metsa jau seest Kaks mändi ärra Warrastat* (Tartu, Kodavere, Pala 1869)[4].

Lisaks kohtuprotokolli tekstile on korpuses olemas ka protokollide metaandmed ehk andmed andmete kohta, mis näitavad, millal ja millises vallakohtus kõnealune kohtuvaidlus toimus. Kohtu nimi võimaldab siduda teksti kindla murdealaga. See võimaldab uurida keelekasutuse ja ortograafia muutumist ajas ning mõjutatust kohalikust murdest.

[3] https://www.ra.ee/vallakohtud/index.php/record/view?id=12703&ru=6QWq93

[4] https://www.ra.ee/vallakohtud/index.php/record/view?id=21335

2.1 Milleks nimeüksuste tuvastamine vallakohtu protokollides?

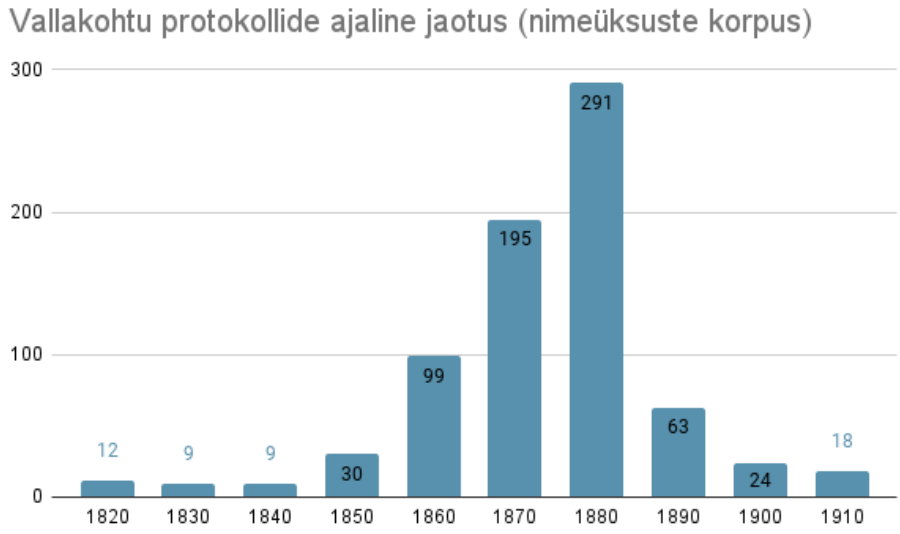
Vallakohtu protokollide automaatanalüüsi kontekstis võimaldab nimeüksuste märgendus **paindlikumat dokumentide otsimist ja sirvimist**. Näiteks mingi konkreetse isikuga seotud protokollide otsingutulemustes saab välja tuua ka teised dokumentides mainitud isikud, mis annab esmase ülevaate, milline isikute ring oli kohtuasjaga seotud, ning seeläbi on võimalik dokumente paremini detailseks (täisteksti) uurimiseks välja valida (Pilvik jt 2019). Nimeüksuste märgendus on ka esmaseks sammuks vallakohtu protokollide korpuse ühendamisel ja **linkimisel teiste sama perioodi kajastavate andmekogudega**, nt kirikuraamatute ja kõrgemate kohtute toimikutega (Lust & Türna 2021).

2.2 Valim

Kuna digiteeritud protokollide hulk on suur, ei ole mõeldav kõigi protokollide käsitsi märgendamine ning nimeüksuste käsitsi märgendamiseks tuleb esmalt moodustada alamkorpus (edaspidi: **nimeüksuste korpus**).

Alamkorpus peaks võimalikult ühtlaselt katma kõiki valdu ning kogu vaadeldavat ajaperioodi, et oleks võimalik süsteemselt uurida nimeüksuste esinemist ning testida masinõpet. Ent kuna ühisloome käigus said protokollide sisestajad vabalt valida, milliseid protokolle nad sisestasid, siis ei eksisteeri sellist ühtlast jaotust juba sisestatud protokollide kogumis – pigem peegeldab sisestatud protokollide kogum sisestajate eelistusi ja huvisid. Näiteks 2024. aasta juuni lõpu seisuga oli kõige rohkem sisestatud Tartumaa valdade protokolle (38 221), mida on oluliselt rohkem kui Harjumaa valdade protokolle (24 663). Järva, Pärnu, Saare ja Viru maakondadest on igaühest sisestatud vähem kui 10 000 protokolli. Ka sisestatud protokollide ajaline jaotus on ebaühtlane: valdav osa protokolle katavad ajavahemikku 1860–1880.

Käesolevas töös kasutasime ühtlase juhuvalimi võtmist üle terve sisestatud protokollide kogumi, seega peegelduvad sisestajate valikud ja eelistused ka nimeüksuste korpuses. Täiendava kitsendusena jätsime juhuvalikut tehes välja protokollid, mis olid lühemad kui 500 tähemärki, et välistada lühikesed sedastused stiilis *Kohtu rahwast ei olnud.* või *Kohto mõistjad koos olnud, agga ei olle ühtegi mõistetud.* Protokollide ajalisest jaotusest korpusvalimis annab ülevaate joonis 2.



Joonis 2. Protokollide ajaline jaotus perioodil 1820–1910 (nimeüksuste korpuses)

2.3 Märgendusjuhiste väljatöötamine

Ükskõik millise nähtuse käsitsi märgendamine on aeganõudev töö, seetõttu tuleb enne tööle asumist põhjalikult järele mõelda, mida ja kuidas märgendada, sh millistesse klassidesse märgendatav nähtus liigitada.

Järgnevalt kirjeldame nimeüksuste käsitsi märgendamise juhiseid ja ka nende liikide ja juhisteni viinud mõttekäike. Selle eesmärgiks on näidata, et märgendusjuhendi koostamine on omaette töö, mille jaoks tuleb oma korpuse märgendamise projektis planeerida piisavalt aega ning võib ka juhtuda, et esialgne märgendusjuhend tuleb ümber teha ja juba märgendatud tekstid ümber märgendada.

Nimeüksuste märgendamisjuhendit koostades tuli leida vastused küsimustele 1) milline on **nimeüksuse ulatus** ja 2) millistesse **liikidesse** tahame selles korpuses nimeüksused jagada. Nimeüksuse ulatuse all mõeldakse seda, kas tekstis järjestikused pärisnimed moodustavad ühe või mitu nimeüksust ning kas nimeüksuse koosseisus on ainult pärisnimi või kuulub sinna ka liigisõna.

Näiteks lause *Tartu linnas elab härra Kask* puhul võib nimedena märgendada ainult pärisnimesid *Tartu* ja *Kask*, aga informatiivsem on kaasata nimeüksuse koosseisu ka liigisõna *linnas* ja võib-olla ka tiitel *härra*. Sageli otsustatakse liigisõna kaasamine nimeüksuste liikide kaupa, näiteks võib otsustada kohanimede puhul

liigisõna kaasata, kuid isikunimede puhul tiitlit mitte. Nii oleks eelnevas lauses kaks nimeüksust *Tartu linnas* ja *Kask*.

Sellised otsused pannakse kirja märgendusjuhendisse, kuid muidugi pole enne märgendama asumist võimalik ette näha kõiki materjalis esineda võivaid erijuhte. Näiteks on 1827. aasta Rõngu vallakohtu protokollis[5] kirjas lause *Peter wottab Astuwerre Tee weeren tühja Suurekorwa Kolmekandine Nurme tük*. Selles lauses on isikunimi *Peter* ning kaks kohanime: *Astuwerre* ja *Suurekorwa. Astuwerre* puhul võime, pärast väikest kõhklust, otsustada, et liigisõna *Tee* kuulub nimeüksuse koosseisu, kuid milline osa sõnaühendist *Suurekorwa Kolmekandine Nurme tük* tuleks märgendada nimeüksuseks?

Kui tulla tagasi nimeüksuste liikide juurde, siis tuleks silmas pidada, et märgendades tekstides mingit nähtust plaaniga kasutada seda andmestikku masinõppe treeningmaterjalina, on mõistlik silmas pidada ka seda, et masinõppe abil on raske tuvastada klasse, mis esinevad tekstides harva. Seega tuleb leida mõistlik tasakaal klasside hulga ja suuruse mõttes: vähem, aga suuremad klassid sobivad hästi masinõppe jaoks, täpsem jaotus väiksematesse klassidesse on jällegi informatiivsem.

Tüüpiline nimeüksuste liigitus sisaldab klasse ***isik***, ***koht*** ja ***organisatsioon***. See jaotus on välja töötatud eelkõige tänapäeva meediatekstide märgendamiseks ja nendest info eraldamiseks. Asudes täiesti teistsugust tekstikogumit nimeüksuste suhtes märgendama, tuleb eelnevalt uurida, milliseid nimesid need tekstid sisaldavad.

Kogukonnakohtute protokollides on palju isikunimesid, need isikud on pärit oma talust/külast/vallast, nad tegutsevad nimedega maastikuobjektidel (näiteks varastavad puid kindla nimega metsast) ja lepivad kohtus kokku, milline ese kellele kuulub. Ka sellel esemel võib olla nimi, näiteks *laew „Eduard"*. Kaugelt kõige sagedasemad nendes tekstides on isikunimed, sageduselt järgmised on kohanimed.

Seega paistab lahendus esmapilgul olevat lihtne: võiksime jagada nimeüksused isikunimedeks ja kohanimedeks ning lisada kategooria *muu* kõigi ülejäänud nimeüksuste jaoks. Kuid kogukonnakohtute protokollide nimeüksuste liigituse puhul on probleemiks talude, külade ja valdade nimed – kas need on tavalised kohanimed, st kas *Kogri talu* on samasugune nimi nagu *Kogri oja*? Talude, külade ja valdade nimesid kasutatakse nendes tekstides sageli inimeste identifitseerimiseks, st nad tähistavad korraga nii kohta kui ka selle kohaga seotud inimeste gruppi – taluperet, külakogukonda, vallarahvast jne. *Kogri oja* tähistab maastikuobjekti, aga *Kogri talu* võib küll tähistada selle talu maid, kuid enamasti kasutatakse seda selle taluga seotud inimeste identifitseerimiseks: *Jaan Kogri talust* on *Kogri Jaan*. Ahja vallakohtu protokollist[6] pärinevas üsna tüüpilises näites on mainitud *Jakobit*, kes on kas *Sae* talu peremees ja/või kannab perekonnanime *Sae*: *Kaibas Perremees Sae Jakob…*

---

[5] https://www.ra.ee/vallakohtud/index.php/record/view?id=9106&ru=73eBqw

[6] https://www.ra.ee/vallakohtud/index.php/record/view?id=11300&ru=6IErjs

Paralleelselt liigisõnaga *talu* on kasutusel ka liigisõna *pere*, mis viitab otseselt taluga seotud inimeste rühmale, taluperele, nagu näiteks Väike-Kareda vallakohtu protokollis[7]: *Kohto ette astus Prohweti perre tüdruk …* Inimest võib muidugi identifitseerida ka kasutades korraga nii talu- kui perenime, nagu näiteks ühes Holstre vallakohtuprotokollis[8]: *… mis senni Otsa Jaan Piip piddanud.*

Kui kohtuistungil osaleja pole kohaliku valla elanik, mainitakse tavaliselt ära tema päritoluvald, nagu ühes Kirna vallakohtu protokollis[9]: *Kohto ette astus Laupa walla mees Hans Kilgas ja kaebas …* Ühes Avinurme vallakohtu protokollist[10] pärit näites on kohtuistungil osalejate kohta kirjas nii ees-, perekonna- kui ka isanimi, kuid ikkagi on peetud vajalikuks kirja panna ka nende päritolutalu ja -küla nimed: *Tulid ette Alekere küla Tamme talu peremees Madis Tamm Josepi p ja Liiwaku talu peremees Jüri Müür Madise poeg ja palusid…*

Kuna nendes tekstides on talude, külade, valdade jne nimede kasutus inimese identifitseerimiseks väga sage, otsustasime nende jaoks luua eraldi nimeüksuste liigi nimetusega ***koht-organisatsioon***. Algselt oletasime, et tekstilises kontekstis on siiski võimalik määrata, kas talu, küla, valla jm nime on kasutatud koha või inimrühma tähistamiseks ja nimeüksuste klassi *koht-organisatsioon* on võimalik jagada kaheks alaliigiks: koht-organisatsioon kohatähenduses ja koht-organisatsioon inimrühma tähenduses.

Märgendamise käigus juhtus aga see, mis esialgse märgendamisotsuse ja tegeliku keelematerjali kohtumisel sageli juhtub: kategooria *koht-organisatsioon* alaliikide eristamine tekstis osutus liiga problemaatiliseks ja nii otsustasime sellest loobuda. Järgnevas, Hageri vallakohtu protokollist[11] pärit näites … *minu lehmad käisid Wundri kapsas ja karduflis* … on eelnevast tekstist selge, et *Wundri* on talunimi, aga kas see talu on selles tekstis esitatud kapsaste ja kartulite omanikuna või asuvad kapsa- ja kartulipõld selle talu maa peal, on raske otsustada.

Nagu öeldud, on nendes tekstides talude nimesid sageli kasutatud inimeste identifitseerimiseks ja tegelikult ongi, eriti vanemates protokollides, piir talu- ja perenime vahel ähmane, said ju eestlased sageli oma perekonnanime selle talu nime järgi, kus nad perekonnanimede panemise ajal elasid. Seega on raske öelda, kas Aakre vallakohtu 1826. aastal kirja pandud protokolli[12] lauses *Parma Margus võttab omma Maade man veel Mulgi tühhi maa* on *Parma* talunimi või perekonnanimi või korraga mõlemat. Nii otsustasimegi, et eesnimele eelnev nii talu- kui perekonnanimena tõlgendatav nimi märgendatakse alati talunimeks, st koht-organisatsiooniks, mitte isikunime osaks.

---

7 https://www.ra.ee/vallakohtud/index.php/record/view?id=20027&ru=5fj8wA

8 https://www.ra.ee/vallakohtud/index.php/record/view?id=8697&ru=6WdENj

9 https://www.ra.ee/vallakohtud/index.php/record/view?id=23286&ru=9i47Fq

10 https://www.ra.ee/vallakohtud/index.php/record/view?id=14519

11 https://www.ra.ee/vallakohtud/index.php/record/view?id=4769

12 https://www.ra.ee/vallakohtud/index.php/record/view?id=6388&ru=5pOc9a

Isikunimed ja koht-organisatsioonide nimed ongi nendes tekstides kõige sagedasemad. „Tõelised" kohanimed märgivad tavaliselt loodusobjekte: jõgesid, järvi, metsi, karjamaid, põlde ja ka inimtekkelisi kohti, mida ei ole nendes tekstides kasutatud inimeste identifitseerimiseks, nt *Suuresadam* Emmaste valla kohtuprotokollist[13] pärinevas näites: .. *tänawuse meresöidu ajaks, mis üheksamal selle kuupäewal Suuresadamal algab,..*

Organisatsioonide nimed on nendes tekstides väike ja kindlapiiriline rühm, valdavalt on tegemist kohtute nimedega, nt Massu vallakohtu protokollis[14]: *1913 a. 21 detsembril Massu Walla kohus omas koraldawas kohtu koos olemisel,...*

Kohtuprotokollides on sageli juttu esemetest, mille omandiõiguse üle vaieldakse või kokku lepitakse. Vahel on sellisel esemel ka nimi, nt eespool mainitud *laew „Eduard"*.

Kui ei taheta teha väga väikeseid märgendatava nähtuse klasse, on tavaliselt märgendussüsteemis olemas ka märgend ***muu***, mida kasutatakse nimede puhul, mis ei mahu eelkirjeldatud kategooriate alla, aga mille puhul on siiski identifitseeritav, millega on tegu. Selliselt on kohtuprotokollides märgendatud nimega sündmused, nt *Paide Mardilaat* ning kohtupidamisel kasutatud seadusekogude nimed ja vastavad lühendid, nt *Liiwimaa talurahwa seaduse raamatu*, *Liiwl. Talor. Säd.*, *Tallorahwa seaduse ramato*, *T. S. R.*

Kõnealused kohtuprotokollid on kohati raskesti loetavad ja raskesti mõistetavad. Märgendatavate tekstide hulgas on ka selliseid, kus ka konteksti põhjal pole võimalik aru saada, millest täpselt jutt käib. Nii on raske öelda, mis on *albri Lemete* ühes 1860. aasta Kotlandi vallakohtu protokollis[15]: *ja teine olli Karrala wallast albri Lemete, mis ouest ulla alt ärra warrastanud.* Selliste juhtumite jaoks tegime eraldi kategooria ***teadmata***.

Märgendusjuhendi koostamine on iteratiivne protsess: koostatakse esialgne märgendusjuhend, märgendatakse nende järgi osa tekste, selle töö käigus tekib uusi teadmisi, küsimusi ja kahtlusi, mille põhjal täiendatakse märgendusjuhendit ja – mis väga oluline – kontrollitakse üle juba tehtud märgendus.

Parim viis märgendusjuhendit testida on see, kui selle alusel märgendab sama teksti mitu inimest ja tulemust võrreldakse. Märgenduse erinevused näitavadki kätte need aspektid, mille osas märgendusjuhendit tuleks täiendada ja täpsustada.

---

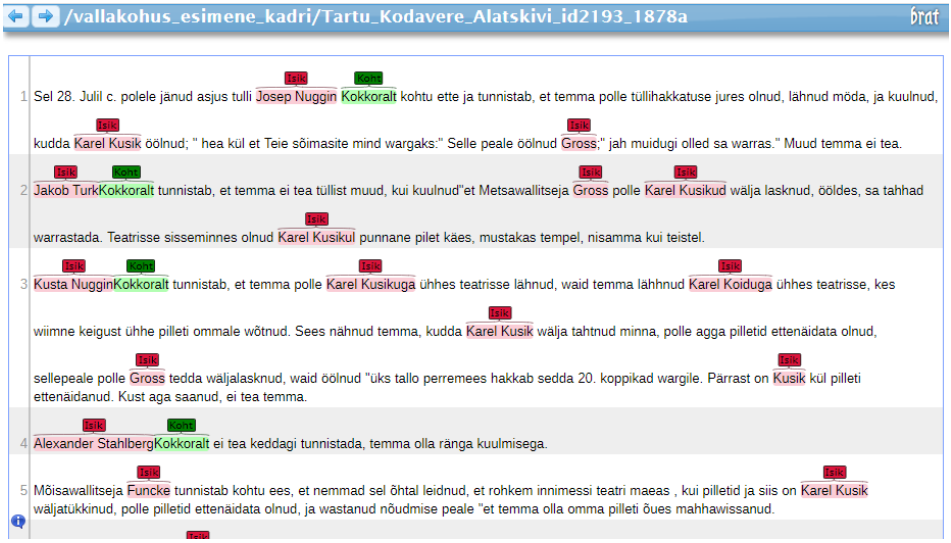
13 https://www.ra.ee/vallakohtud/index.php/record/view?id=15415&ru=5ppzgk

14 https://www.ra.ee/vallakohtud/index.php/record/view?id=446

15 https://www.ra.ee/vallakohtud/index.php/record/view?id=15249&ru=6ViCEQ

2.4 Märgendamisprotsess

Märgendamistööriistana kasutasime veebipõhist keskkonda **brat**[16] (Stenetorp jt 2012), mis võimaldab intuitiivse liidese abil märgendada nimeüksustele vastavaid fraase ning uurida ka korpuse esmast statistikat. Joonisel 3 on näha ekraanitõmmist brati keskkonnast.



Joonis 3. Ekraanitõmmis: vallakohtu protokolli Tartu_Kodavere_Alatskivi_id2193_1878a[17] märgendamine keskkonnas brat

Jaotasime märgendamisprotsessi neljaks iteratsiooniks ehk alametapiks, mille käigus märgendatavate protokollide arv varieerus vahemikus 250–500. **Alametappideks** jaotus oli vajalik, et hinnata töö mahukust ja märgendamisele kuluvat aega; iga etapi järel toimus märgendamisel esilekerkinud probleemide ja küsimuste arutamine, märgendusjuhiste täpsustamine ning märgenduste parandamine. Põhiosa protokolle märgendas lingvistika taustaga isik, kel oli ka varasemalt kogemusi vana kirjakeele tekstidega.

Märgendatud korpuse kogusuurus on 1500 dokumenti ning u 320 000 sõna; tabelis 1 on toodud märgendatud korpuse statistika nimeüksuste liikide kaupa (tabel 1).

[16] https://brat.nlplab.org/

[17] https://www.ra.ee/vallakohtud/index.php/record/view?id=2193

Tabel 1. Nimeüksuste jaotus liikide kaupa märgendatud korpuses

Nimeüksuse liik (ingl lühend)	Nimeüksuste arv	Osakaal kogu korpuses
isik (PER)	23 126	84%
koht-organisatsioon (LOC_ORG)	2 733	9,9%
koht (LOC)	1 008	3,7%
organisatsioon (ORG)	419	1,5%
ese, muu, teadmata (MISC)	254	0,9%
Kokku	27 540	100%

Nagu tabelist 1 nähtub, on nimede jaotus liigiti ebaühtlane: domineerivad isikunimed (84% kõigist nimedest), samas muude nimede sagedus on oluliselt madalam.

2.5 Märgendajatevaheline kooskõla

Käsitsi märgendamisel on oluline välja selgitada, kuivõrd hästi langevad kokku kahe sõltumatu märgendaja märgendusotsused ehk milline on **märgendajatevaheline kooskõla**. Kõrge kokkulangevus viitab märgendusjuhiste ja ülesande selgusele (ja sellele, et märgendusprotsess on hästi korratav), samas madal kokkulangevus annab märku ülesande raskusest ja/või halvast defineeritusest. Enamasti eeldatakse kokkulangevust mõõtes, et märgendajad on omas valdkonnas eksperdid ning motiveeritud märgendama. Samuti on märgendajatevaheline kooskõla heaks orientiiriks automaatmärgenduse arendamisel: eeldatavasti jääb automaatmärgenduse kvaliteet inimestevahelisele kooskõlale alla või on sellega enam-vähem võrdne.

Kooskõla mõõtmiseks leidub mitmeid statistiliselt korrigeeritud mõõdikuid (nt Coheni κ-statistik), ent nende puhul on probleemiks see, et tuleb arvutada ka n-ö negatiivsete näidete hulk, mis ei ole loomuliku keele puhul määratletav (st kuidas määrata, kui palju on n-ö mittenimefraase, mis tuleks jätta märgendamata). Alternatiiv on kasutada kooskõla leidmisel F1-skoori[18], nagu soovitavad Hripcsak ja Rothschild (2005), mille puhul negatiivsete näidete hulka leidma ei pea.

F1-skoor toetub kahele infoeraldamises laialdaselt kasutatavale mõõdikule: saagisele ja täpsusele. **Täpsus** näitab, kui suur osakaal kõigist esimese märgendaja märgendustest langesid kokku teise märgendaja omadega; mittekokkulangevaid võib tõlgendada üleliigsete märgendustena. **Saagis** näitab, kui suur osakaal kõigist teise märgendaja märgendustest langesid kokku esimene märgendaja omadega;

[18] *F1-score* (ingl), eesti keeles vt ka (Sügis jt 2024: 119)

mittekokkulangevaid saab tõlgendada puuduolevate märgendustena. **F1-skoor** on saagise ja täpsuse harmooniline keskmine, mis arvutatakse järgmise valemi järgi:

$$F1 = \frac{(2 \times saagis \times täpsus)}{(saagis + täpsus)}$$

Kuigi F1-skoor annab optimistliku hinnangu kooskõlale, st kõrgema kooskõla kui annaks κ-statistik (kui saaksime leida negatiivsete näidete hulga), siis teoorias läheneb suure negatiivsete näidete hulga korral κ-statistiku väärtus F1-skoorile ja seega saab neid mõõdikuid käsitleda sarnast tulemust andvatena (Hripcsak & Rothschild 2005).

Mõõtsime kooskõla 250 protokollil, mis olid ka teistkordselt märgendatud teise lingvisti poolt. Märgendamisel kasutati samu juhiseid, ent märgendus toimus sõltumatult, st eelmise märgendaja märgendused polnud näha.

Käesolevas töös kasutasime **bratiaa**[19] tööriista, mis võimaldab kooskõla arvutada otse bratist imporditud failide peal. Kooskõla hindamisel kasutasime n-ö ranget režiimi: kokkulangevateks lugesime ainult need märgendused (fraasid), mille piirid tekstis langesid täpselt kokku, osaline ülekattuvus loeti märgendusveaks.

Kahe märgendaja keskmiseks kooskõlaks (F1-skoor) mõõtsime 0,95, mis on üsna kõrge näitaja (skaalal, kus nullilähedased väärtused näitavad madalat kooskõla ning 1 on täielik kooskõla). Uurisime ka kooskõla nimeliikide kaupa. Suurim oli kooskõla kõige sagedasemal nimeliigil – isikunimedel – 0,98. Paremuselt teine kooskõla näitaja oli organisatsiooninimedel (0,87), millele järgnes kategooria *muu* (0,83). Organisatsiooninimede kõrge kooskõla taga oli tõenäoliselt nimeliigi suhteliselt väike varieeruvus, sest enamasti oli tegemist kõrgemate kohtuinstantside nimedega, mis olid kergesti tekstist tuvastatavad. Muude nimede kategoorias domineerisid talurahvaseadusele viitavad lühendid (nt *T.S.R = Tallorahwa seadusse ramato*), mille esildumine ongi peamine põhjus suhteliselt kõrge kooskõla taga.

Kategooria *koht-organisatsioon* kooskõlaks mõõtsime 0,81 ning kohanimede kooskõlaks 0,74. Koht-organisatsiooni nimede märgendamise kooskõlaga võib rahule jääda, arvestades kategooria keerulisust. Kohanimede suhteliselt madala kooskõla taga on tõenäoliselt nimede suur varieeruvus ning episoodilisus (st ühes protokollis mainitud n-ö päris kohad ei pruukinud mujal korduda). Kõige madalam oli kooskõla kategooriates *teadmata* (0,54) ja *ese* (0,20), mis on seletatav nende probleemsuse ning suhteliselt madala esinemissagedusega.

Kuna Rahvusarhiivi ühisloome käigus toimus samuti nimede käsitsi märgendamine, oli võimalik mõõta kooskõla ka meie märgendajate ning ühisloome märgendajate vahel. Ühisloomes oli kasutusel vaid kaks nimekategooriat – isikunimed ja kohanimed –, seega tuli kooskõla mõõtmiseks kategooriaid lihtsustada. Jätsime oma korpuses samuti alles ainult isikunimed ja kohanimed: taandasime kategooria

[19] https://github.com/kldtz/bratiaa

koht-organisatsioon kohanimedeks ning eemaldasime nimed kategooriatest *organisatsioon*, *muu* ja *teadmata*. Kooskõla mõõtmine andis tulemuseks keskmise F1-skoori 0,68, mis on oluliselt madalam meie märgendusjuhiseid järgivate märgendajate kooskõlast (0,95). See tulemus annab märku nimede märgenduse ebaühtlusest ühisloomes (mis on mõneti ootuspärane, kuna seal polnud nimede märgendamine kohustuslik) ning ühtlasi näitab nimede märgenduse korrastamise vajadust.

3. Märgendatud korpuse ettevalmistamine masinõppekatseteks

Kuigi brat on mugav tööriist nimede märgendamiseks, ei ole brati poolt väljastatav vorming (st brati väljundfail) koheselt masinõppel kasutatav. Põhjus on selles, et brati väljundfailis (.ANN-failis) on toodud indeksitega välja iga nimeüksuse täpne asukoht tekstis: milline on nime algusindeks ja milline on nime lõpuindeks (vt joonis 4), samas vajavad masinõppesüsteemid sisendiks sõnestatud teksti ning märgendusinfot sõnede kaupa.

id	nimeliik	algus	lõpp	nimeüksuse fraas
T1	Org	21	43	Kassari walla wolikogu
T2	Isik	47	59	Kustaw Wälja
T3	KO_koht	86	99	Kassari walla

Joonis 4. Väljavõte nimeüksuste märgendusest lauses *Kassari walla wolikogu ja Kustaw Wälja tegiwad ühe teisega kaupa Kassari walla Kooli maade üle* brati märgendusfailis. Iga nime kohta on välja toodud selle indeks (T1, T2, T3), nimeliik ja asukoht (algus- ja lõppindeks) ning nimeüksuse fraas. Väljavõte pärineb Kassari vallakohtu 1889. aastal kirjapandud protokollist[20]. Märgendamisel brati tööriistas kasutati *KO_koht* lühendit tähistamaks koht-organisatsiooninime

Üheks levinud masinõppel kasutatavaks vorminguks on **IOB2-vorming**, kus iga sõne puhul on välja toodud selle nimeüksuse kuuluvust (või mittekuuluvust) tähistav märgend: kas sõne kuulub nime algusesse (*B – begin*), nime sisse (*I – inside*) või on nimest väljaspool (*O – outside*); nime algusesse või sisse kuuluvate sõnade puhul lisatakse ka nimeliik (vt tabel 2).

[20] https://www.ra.ee/vallakohtud/index.php/record/view?id=20356&ru=6scL2b

Tabel 2. Lause *Kassari walla wolikogu ja Kustaw Wälja tegiwad ühe teisega kaupa Kassari walla Kooli maade üle* IOB2-vormingus esitus. Iga sõne on eraldi real ning sõnele järgneb nimeüksusesse kuuluvuse märgend. Kasutatud tähistused: B-ORG – organisatsiooninime algus, I-ORG – organisatsiooninime keskosa või lõpp, B-PER – isikunime algus, I-PER – isikunime keskosa või lõpp, B-LOC_ORG – koht-organisatsiooni nime algus, I-LOC_ORG – koht-organisatsiooni nime keskosa või lõpp ning O – väljaspool nimefraase paiknev sõna

Sõne	**IOB2 märgend**
Kassari	B-ORG
walla	I-ORG
wolikogu	I-ORG
ja	O
Kustaw	B-PER
Wälja	I-PER
tegiwad	O
ühe	O
teisega	O
kaupa	O
Kassari	B-LOC_ORG
walla	I-LOC_ORG
Kooli	O
maade	O
üle	O
.	O

Märgenduste teisendamine IOB2-vormingusse nõuab teksti **sõnestamist** (on vaja tuvastada, millised tekstiüksused on eraldiseisvad sõned) ning sõnede ja nimeüksuste kokkuviimist vastavate tekstiüksuste algus- ja lõppindeksite järgi. Kuigi teksti sõnestamiseks saime kasutada EstNLTK töövahendeid (Laur jt 2020), tuli neid kohandada konkreetse korpuse erisusi arvestama. Näiteks leidus tekstides ametikohaga kokkukirjutatud nimesid (*talumeesNikolai*, *wöörmündriJaan*), mis tuli järelparandustes tõsta lahku kaheks sõneks.

Nii sõnestuse (järel)parandamine kui ka tekstiüksuste joondamine on detailimahukad programmeerimisoskusi nõudvad ülesanded, mida me ei hakka siin

süvitsi käsitlema. Käesoleva töö jaoks lõi teisendusteks vajalikud skriptid Kristjan Poska oma bakalaureusetöös (Poska 2021), huvilised võivad uurida lähtekoodi GitHubi repositooriumis[21].

Viimane, aga küllaltki oluline samm masinõppe ettevalmistustes on andmete jagamine **treening-, valideerimis- ja testhulgaks**. Treeninghulga abil õpetatakse masinõppesüsteemi ennustama igale sõnele nimeüksuse kuuluvuse märgendit. Valideerimishulka kasutatakse süsteemile parimate õpiparameetrite[22] leidmiseks: treeninghulgal treenitakse välja mitu eri mudelit, mis kasutavad erinevaid õpiparameetrite väärtusi, ning valideerimishulgal hinnatakse, milline neist mudelitest saavutab parima tulemuse. Kui parimate õpiparameetrite väärtustega mudel on leitud, toimub selle kvaliteedi lõplik hindamine testhulgal.

Oluline on veel märkida, et hulgad on rangelt kattumatud – tekstid, mis esinevad ühes hulgas, ei esine teistes hulkades. See tagab süsteemi hindamise objektiivsuse – mudel peab treeninghulgal omandama piisavalt suure üldistusvõime, et teha täpseid ennustusi ka valideerimis- ja testhulga tekstidel, mida mudel seni kohanud pole.

Andmete jagamisel treening-, valideerimise- ja testhulgaks kasutatakse enamasti põhimõtet, et treeninghulk on proportsionaalselt suurim alamhulk andmetest, et anda süsteemile õppimiseks võimalikult palju näiteid; valideerimis- ja testhulgad on enamasti väikesed. Sageli kasutatavad jaotused on näiteks 80% andmetest treeninguks, 10% valideerimiseks ja 10% testimiseks või 60% andmetest treeninguks, 20% valideerimiseks ja 20% testimiseks. Käesoleva töö katsetes kasutati 75% tekstidest treeninguks, 8% valideerimiseks ning 17% testimiseks.

4. Masinõppekatsed

Masinõppekatseid loodud nimeüksuste korpusel kajastab detailselt artikkel (Orasmaa jt 2022).

Masinõppe-eksperimentides prooviti nn traditsioonilist masinõpet, kus süsteem õppis nimeüksuste märgendeid ennustama inimeste poolt ettemääratud õpitunnuste alusel, ning neuromudelitel põhinevat ülekandeõpet (ingl *transfer learning*), kus süsteem oli juba varasemalt eeltreenitud suurel keelekorpusel n-ö „üldtasemel keelt tundma" ning see peenhäälestati (ingl *fine-tuning*) nimeüksuseid tuvastama.

Traditsioonilise masinõppe lähenemisena kasutati Tkachenko, Petmansoni ja Lauri (2013) poolt loodud eesti keele nimeüksuste mudelit, mis treeniti uuesti välja

[21] https://github.com/pxska/bakalaureus/tree/main/experiments

[22] Näiteks neuromudelite puhul on muudetavateks õpiparameetriteks õpisamm (ingl *learning rate*), mis määrab, millises ulatuses toimub treenimisel mudeli kaalude uuendamine, ning ploki suurus (ingl *batch size*), mis määrab, kui suure ploki treeningandmeid peab mudel läbi töötlema enne mudeli uuendamist.

vallakohtu protokollide andmestikul. Mudel kasutab õpitunnuseid (ingl *features*), mis analüüsivad sõna kuju (erinevate liiki sümbolite esinemine sõnas, sh suur- ja väike tähtede esinemine), sõna morfoloogilise analüüsi tulemusi (nt lemmat, sõnaliiki), sõna esinemist suures nimeloendis (ligi 903 000 nime sisaldav loend, mis sisaldab enamuses ingliskeelseid nimesid) ning sõna teisi esinemisi samas tekstis (nt kas sõna esineb kusagil lause keskel suure algustähega). Tasub märkida, et õpitunnuste määramiseks ja eraldamiseks luuakse arvutiprogramm ning korpusest eraldatakse tuhandeid tunnuseid, mis kombinatsioonis tagavad mudeli ennustusvõime; üksikute tunnuste roll eraldiseisvana on aga väike.

Hoolimata sellest, et Tkachenko, Petmanson ja Laur (2013) lõid oma mudeli eeskätt tänapäevaste ajalehetekstide analüüsimiseks (sellest lähtuvalt toimus ka õpitunnuste valik), saavutas see pärast vallakohtu protokollide andmestikul ümbertreenimist F1-skoori tulemuseks 0,89.

Neuromudelitel põhineva ülekandeõppe eripäraks on, et õpitunnuste määramisel enam inimese abi[23] pole vaja – neuromudel on eeltreenitud suurel korpusel, kus see on õppinud eristama keele analüüsimiseks vajalikke tunnuseid. Käesolevas töös katsetasime Berti neuromudeleid (Devlin jt 2019), mis olid eeltreenitud tänapäevastel eesti kirjakeele korpustel suurusega 38 miljonit sõnet kuni 2,5 miljardit sõnet. Üldiselt võib öelda, et katsetatud mudelid andsid häid tulemusi – nimeüksusi tuvastama häälestatud mudelid saavutasid F1-skoori vahemikus 0,90–0,93. See on mõnevõrra üllatav, kuna neuromudelid olid eeltreenitud tänapäevasel keelel, mitte meie poolt analüüsitaval 19. sajandi keelel, mis polnud pealegi veel homogeenne. Võib oletada, et neuromudelid suutsid tabada mingeid üldisi nimedele omaseid keelelisi mustreid; samas ei saa öelda, et tulemus oleks ühtlaselt hea kõigi nimeliikide lõikes.

Vaadates parima süsteemi tulemusi nimeliikide lõikes, siis F1-skoor oli kõrgeim isikunimede ja organisatsiooninimede tuvastamisel (vastavalt 0,96 ja 0,95). Organisatsiooninimede puhul on üllatav, et masinõppe tulemus (F1-skoor 0,95) ületab inimmärgendajate vahelist kooskõla (0,87). Inimmärgendajate vahelisi erinevusi uurides jäi silma, et enamasti olid põhjuseks märgendusjuhiste uuendamise järel parandamata jäänud märgendused; masinõppe treeningkorpuses oli aga organisatsiooninimede märgendus ühtlustatud, mis tagas parema tulemuse.

Paremuselt kolmas oli tulemus koht-organisatsiooninimede tuvastamisel (0,81), millele järgnesid *mitmesugused nimed*[24] (0,74) ja viimasena kohanimed (0,66).

Kokkuvõttes võib öelda, et nagu märgendajatevahelise kooskõla puhul, nii mängis ka masinõppel rolli nimede sagedus: sagedasemaid nimesid oli kergem

[23] Ehk siis: Tkachenko, Petmansoni ja Lauri (2013) lähenemise puhul pidid mudeli loojad ise kirjutama arvutiprogrammi, mis eraldas sobivad õpitunnused.

[24] Kategooria *mitmesugused nimed* (*Misc* ehk ingl *miscellaneous*) sisaldas käsitsi märgendamisel eristatud kategooriaid *esemed*, *muud* ja *teadmata*, mis võeti madala sageduse tõttu kokku ühe märgendi alla.

õppida, samas kui väikese sagedusega (ja/või selgelt määratlemata) nimede puhul on raske saavutada head tulemust isegi masinõppe tippklassi kuuluvate neuromudelite abil. Kui nimeüksuste automaattuvastamist vaadelda kui abivahendit märgenduse tekitamisel ja korrastamisel ning jätta märgenduse lõplik parandamine inimesele, on praktikas suur abi juba ka sagedasi nimesid hästi tuvastavast süsteemist.

Kokkuvõte

Selles uurimuses vaadeldi nimeüksuste näitel korpuse käsitsi märgendamist masinõppe mudeli treenimise jaoks. Ülesande muudab keerukamaks, ent samas ka huvitavamaks märgendatavate tekstide vanus ja varieeruvus, sh murdelisus.

Kuigi siinses näites oli märgendamise eesmärgiks treeningandmestiku loomine masinõppe tarvis, sobib märgendamise näide eeskujuks ka lihtsalt käsitsi märgendatud korpuse loomisel, ükskõik kas märgendada soovitakse nimeüksusi, semantilisi kategooriaid või hoopis midagi kolmandat. Ikka kehtivad samad põhimõtted: mõtle läbi märgendatavate tekstide valik ja märgendatavad nähtuse liigid, koosta esialgne märgendusjuhend, rakenda seda tekstidel, täpsusta ja täiusta vajadusel märgendusjuhendit ning ära unusta juba märgendatud tekste üle kontrollimast. Lase vähemalt osa materjali märgendada kahel inimesel ning mõõda märgendajatevahelist kooskõla.

Oluline on veelkord toonitada, et sellise töö jaoks tuleb planeerida piisavalt aega – väga sageli juhtub, et märgendusjuhendit tuleb täpsustada korduvalt ja juba märgendatud tekstid tuleb siis muidugi uuesti läbi vaadata.

Näidisuurimuse valmimist on toetanud EKKD projekt EKKD-TA10 „Infoeraldus ajalooliste institutsioonide protokollide (1880–1940) näitel".

Kirjandus

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Jill Burstein, Christy Doran & Thamar Solorio (toim), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Hripcsak, George & Adam S. Rothschild. 2005. Agreement, the F-Measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association* 12(3). 296–298. https://doi.org/10.1197/jamia.M1733.

Laur, Sven. 2013. Estonian NER corpus / Nimeüksuste korpus. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-00073L.

Laur, Sven, Siim Orasmaa, Dage Särg & Paul Tammo. 2020. EstNLTK 1.6: Remastered Estonian NLP pipeline. Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, jt (toim), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7152–7160. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.884.

Lust, Kersti & Tõnis Türna. 2021. Valdade iseseisvumise raske algus: vallakirjutajad 1866–1891. *Tuna. Ajalookultuuri ajakiri* 24(3). 10–32.

Orasmaa, Siim, Kadri Muischnek, Kristjan Poska & Anna Edela. 2022. Named entity recognition in Estonian 19th century parish court records. Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, jt (toim), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5304–5313. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.568.

Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa & Tõnis Türna. 2019. *Möistus sai kuulotedu*: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 15. 139–158. https://doi.org/10.5128/ERYa15.08.

Poska, Kristjan. 2021. *Nimeolemite tuvastamine 19. sajandi vallakohtu protokollides*. Tartu: Tartu Ülikool. Bakalaureusetöö. http://hdl.handle.net/10062/74471.

Sirts, Kairit. 2021. New Estonian NER corpus; Uus Eesti nimeüksuste korpus. Center of Estonian Language Resources. https://doi.org/10.15155/1-00-0000-0000-0000-001B1L.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou & Jun'ichi Tsujii. 2012. brat: A web-based tool for NLP-assisted text annotation. Frédérique Segond (toim), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102–107. Avignon, France: Association for Computational Linguistics. https://aclanthology.org/E12-2021.

Sügis, Elena, Ardi Tampuu, Anna Aljanaki, Mark Fišel & Meelis Kull. 2024. *Praktiline andmeteadus. Kõrgkooliõpik*. Tartu: Tartu Ülikooli arvutiteaduse instituut. https://hdl.handle.net/10062/106497.

Tkachenko, Alexander, Timo Petmanson & Sven Laur. 2013. Named entity recognition in Estonian. Jakub Piskorski, Lidia Pivovarova, Hristo Tanev & Roman Yangarber (toim), *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, 78–83. Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/W13-2412.

Suured keelemudelid ja nende rakendamine keeleuurimisel

Eleri Aedmaa

Lühikokkuvõte

Keelemudelite kasutamine on erinevate tekstirobotite rakenduste kaudu tehtud lihtsamaks kui kunagi varem. Muu hulgas saab ka tehniliste oskusteta uurida seda, kas ja kuidas võiks keelemudeleid keeleteaduses rakendada; mudelite kasutamine väljaspool rakendusi eeldab vaid minimaalseid programmeerimisoskuseid. Kuna tegemist on ülikiiresti areneva valdkonnaga, teeme selles peatükis vaid lühikese sissejuhatuse suurte keelemudelite olemusse, hetkeolukorda (2025. aasta) ning rakendusprintsiipidesse.

1. Taust

Suured keelemudelid (SKM, ingl *large language models*, LLM) on tohutu suurel hulgal tekstiandmetel treenitud keelemudelid, mis on võimelised genereerima loomulikku keelt ja sooritama erinevaid ülesandeid, näiteks tõlkima, kokkuvõtteid tegema, vastama küsimustele, tootma erinevat tüüpi loomingulist sisu, kirjutama koodi jne. OpenAI arendatud SKM-il põhineva juturoboti ChatGPT[1] ilmumise järel 2022. aasta novembris on väga paljude erialade esindajad, sh humanitaarteadlased, püüdnud leida võimalusi ChatGPT ja teiste generatiivsete keelemudelite rakendamiseks oma töö- ja eraelus. SKM-id lahendavad (üllatavalt) keerulisi ülesandeid, kuid vahel jäävad imelihtsate ülesannetega hätta. Selleks, et SKM-id saavutaksid keeleteaduses meetodina usaldusväärsuse, on praegu nendega eksperimenteerimine ja nende töö hindamine väga oluline.

Suuresti ChatGPT-ga alguse saanud tehisintellektivaimustus on jätnud mulje, et SKM-id hüppasid järsku eikusagilt välja. Tegelikult on tegemist asjade loomuliku käiguga, sest arvutusvõimsus on jõudnud sellisele tasemele, et oleme suutelised välja arendama väga võimekaid mudeleid. Metodoloogiliselt liigutigi

[1] https://chatgpt.com/

traditsioonilistest statistilistest meetoditest (vt õpiku ptk 6) edasi siis, kui masinõppes võeti kasutusele esimesed tehisnärvivõrgud. Tehisnärvivõrk on inimese aju jäljendada püüdev arvutuslik arhitektuur, mis koosneb omavahel ühendatud elementidest (sõlmedest), mis omakorda asetsevad mitmetel omavahel ühendatud kihtidel (tavaliselt sisendkiht, varjatud kihid ja väljundkiht). Närvivõrkude eesmärk on andmete põhjal välja arvutada otsuseid ja ennustusi, näiteks seda, milline sõne eelnevast kontekstist lähtuvalt järgmisena esineda võiks. Jättes kõrvale tehnilised detailid, siis aastate jooksul on välja töötatud erinevat tüüpi närvivõrke, mis üha edukamalt on suutnud muu hulgas ka loomulikku keelt töödelda. Enimkasutatud närvivõrgud loomuliku keele jaoks on olnud 1980. aastatel väljatöötatud rekurrentsed närvivõrgud, näiteks pika lühiajalise mälu võrgud (ingl *long short term memory*, LSTM), mis omal ajal tõid läbimurde näiteks masintõlkes.

SKM-id nii, nagu me neid praegu tunneme, on **tehisnärvivõrgud**, mis rakendavad 2017. aastal tutvustatud transformerarhitektuuri (Vaswani jt 2017). Võrreldes neile eelnenud keelemudelitega kasutavad **transformermudelid** (nt ingl *generative pretrained transformer* ehk GPT) lisakomponendina sisemist tähelepanumehhanismi (ingl *self-attention*). See osa võimaldab mudelil keskenduda sisendi (nt lause) erinevatele osadele (nt sõnadele), nendevahelistele seostele ning hinnata sisendi osade tähtsust teiste osade suhtes. Seesuguse protsessi abil on transformermudelid võimelised varasemast täpsemalt tabama seda, mida kasutaja oma sisendiga püüab saavutada. SKM-idel on miljardeid (isegi triljoneid) mudeli käitumist ja võimekust juhtivaid **parameetreid**, mida mudel treenimisel optimeerib, et saavutada võimalikult täpne tulemus.

SKM-ide arendamisel läbitakse kaks peamist treeningetappi. Esimene etapp on mudeli **eeltreenimine**, mille tulemusena valmib (baas)mudel, mis võib olla võimeline lahendama terve rea ülesandeid, mida treeningandmetest on võimalik õppida. Paljusid turul olevaid suuri keelemudeleid nimetataksegi eeltreenitud või baasmudeliteks. Samas tuleb meeles pidada, et suur osa SKM-ide treeningandmetest on internetist kokku kraabitud ehk mudelid on üldjuhul võimelised päris hästi hakkama saama teemadel, millest internetis palju juttu on. Samal ajal ei pruugi keelemudel olla väga pädev valdkondades, mille tekstid pole avalikult internetis kättesaadavad. Seega tuleb spetsiifilisema ülesande lahendamiseks mudelit tihtipeale lisaandmetega **peenhäälestada** (ingl *fine-tuning*). Seesugune lisatreenimine võib tähendada mingi spetsiifilist valdkonda puudutavate tekstide lisamist, aga ka näiteks mingis kindlas keeles andmetega treenimist ehk baasmudelile seesuguse keele õpetamist, milles mudel väga hea ei ole. Kui võtame näiteks eesti keele, siis paljudes praegu turul olevates mudelites on eesti keel olemas, sest veebis on mingi hulk eestikeelset materjali kõigile kättesaadav. Seetõttu on see mõnevõrra juhuslik tulemus, et osad mudelid oskavad eesti keelt sellisel tasemel, et emakeelena kõneleja küll leiab vigu, kuid üldiste ülesannete lahendamisel need kasutajat ei häiri. Praegu pole veel võimalik anda head ülevaadet sellest, kui hästi eesti keelt töödelda suutvad mudelid erinevate ülesannete lahendamisel töötavad, aga tabelis

1 on ülevaade mudelitest, mille eesti keele tugi on katsetamist väärt. Siin on aga oluline piirang see, et mudelite kasutamine on piiratud – paljud praegu turul olevad mudelid on tasulised ja/või nende sisendi ja väljundi pikkus on limiteeritud.

Tabel 1. Näiteid SKM-idest, mis toetavad eesti keelt (täieneb pidevalt)

Mudelid	**Autor**	**Selgitus**
GPT-5, GPT-4o, GPT-4.1, GPT-3.5, ChatGPT, o1, o3[2]	OpenAI	Need mudelid ei ole avatud lähtekoodiga, eestikeelsete andmete hulga ja sisu kohta pole ametlikku ülevaadet.
Gemini 2.5 Pro, Gemini 2.0 Flash[3]	Google	Need mudelid ei ole avatud lähtekoodiga, eesti keel on ametlikult nimetatud toetatud keelte seas, kuid treeningandmete eestikeelse sisu ja mahu kohta puudub täpne ülevaade.
Claude Opus 4.1/4, Claude Sonnet 4/3.7, Claude Haiku 3.5/3[4]	Anthropic	Need mudelid ei ole avatud lähtekoodiga, eesti keelt pole nimetatud ametlikult toetatud keelte seas.
LLammas[5]	TartuNLP	Mudel on avatud lähtekoodiga, tegemist on LLama-2 mudeli eestikeelseks peenhäälestatud versiooniga.
EuroLLM[6]	erinevad Euroopa teadusasutused	Mudeli arendamist on toetanud Euroopa Komisjon eesmärgiga luua mudel kõikide EL-i ametlike keelte jaoks.
DeepSeek V3.1/R1	DeepSeek	Mudel on avatud lähtekoodiga, tegemist on aga Hiina päritolu mudeliga ja seega võivad rakenduda soovitused mudelit mitte kasutada.

SKM-ide rakendamine eeldab **prompti** ehk **viiba**, teisisõnu juhise või päringu sisestamist mudelile. Seesugust tegevust nimetatakse **promptimiseks**, millest antakse täpsem ülevaade allpool.

[2] https://openai.com

[3] https://deepmind.google/technologies/gemini/

[4] https://www.anthropic.com/claude

[5] https://huggingface.co/tartuNLP/Llammas

[6] https://eurollm.io/

2. Näiteid suurte keelemudelite rakendamisest humanitaarias

SKM-ide rakendamist pole veel jõutud paljude ülesannete lahendamisel uurida, kuid sellele vaatamata on juba ilmunud mõned artiklid, mis käsitlevad nende mudelite rakendamist keeleteaduses. Näiteks on proovitud semantiliste rollide määramist ChatGPT 3.5 ja Bingi juturobotiga, kus leiti, et inimestega võrreldes on Bingi võimekus vaid pisut madalam (94,5% *vs*. 92,7%), mis viitab SKM-ide heale võimekusele lahendada keeleteaduslikke ülesandeid (Yu jt 2024). Kuzman jt (2023) näitasid, et ChatGPT 3.5 on võimeline ingliskeelseid tekstiliike määrama umbes 70% täpsusega ning tõid välja, et sama mudel saavutab kiiduväärt tulemusi ka sloveeni keeles, mida SKM-ide kontekstis võib inglise keelega võrreldes pidada limiteeritud ressurssidega keeleks. Karjus (2023)[7] demonstreeris GPT-3.5 ja GPT-4 näitel, kuidas SKM-ide abil erinevaid suuremahulisi protsesse automatiseerida, mh keeleteaduses, näiteks sõnade tähendusmuutuste ja uute sõnade tuvastamisel ning teksti märgendamisel. Karjus ja Cuskley (2024) hindasid GPT-4 võimekust korpuse annoteerijana keskmisest keerukama ülesande korral.

Umbes aasta pärast ChatGPT väljaandmist ilmus ülevaatlik artikkel generatiivse tehisintellekti rakendamisest leksikograafias, kus analüüsitakse kümmet selleks ajaks ilmunud teemakohast artiklit ja ettekannet, mille põhjal sünteesitakse välja põhilised trendid ning tõdetakse, et uus ajastu leksikograafias SKM-ide rakendamise näol on alanud (De Schryver 2023). Ka Eestis on SKM-ide uurimine ja rakendamine leksikograafias hoo sisse saanud, sest Eesti Keele Instituudil on käimas (2024–2027) uurimisprojekt „Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed“[8], mis seda teemat avada püüab.

Suurte keelemudelite rakendusvõimalusi on uuritud ka **korpuslingvistika** kontekstis. Näiteks on leitud, et SKM-ide (selles uurimuses GPT-4) abiga on võimalik tõhustada korpuste lingvistilist märgendamist, sest SKM-id suudavad väga hästi tabada keele semantilisi ja pragmaatilisi aspekte (Yu jt 2024). Teisest inglise keele põhjal tehtud uurimusest selgus, et ChatGPT 3.5 on võimeline assisteerima ka sagedusloendite tegemisel ning tuvastama kollokatsioone ja erinevaid grammatilisi mustreid, samal ajal tekitas mudelile raskusi sõnade ja tekstilõikude registri/žanri tuvastamine (Uchida 2024). Lisaks on leitud, et ChatGPT suudab küll päris hästi võtmesõnu nende tähenduse alusel kategoriseerida, kuid kipub rühmad üsna üldiseks jätma; ChatGPT võimekus konkordantsanalüüsi teha hinnati halvaks, sest mudel tegi valesid järeldusi ja kippus ka sisendandmeid muutma; diskursusanalüüsi tulemused olid samuti kehvad, sest mudel ei teinud vahet otseste ja kaudsete küsimuste vahel. Kokkuvõtlikult pandi ChatGPT oskused kvalitatiivset analüüsi läbi viia küsimärgi alla ning toodi välja, et mudeliga kaasnevad ka teised probleemid, näiteks raskused tulemuste reprodutseerimisega ning eetilised küsimused (Curry, Baker & Brookes 2024).

---

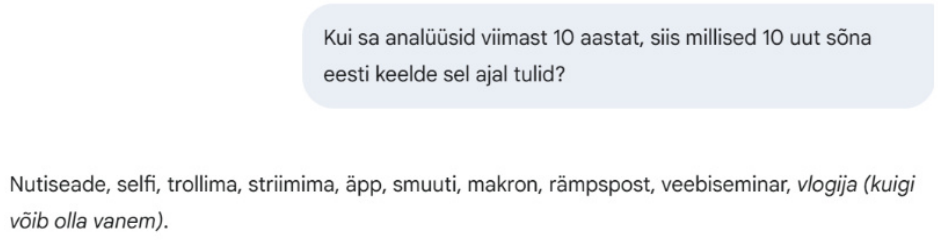
7 Artikli lõpust leiad viibad, mida juhtumiuuringutes kasutati.

8 https://www.etis.ee/Portal/Projects/Display/8d08c5e9-e052-4480-bffd-e102090df536

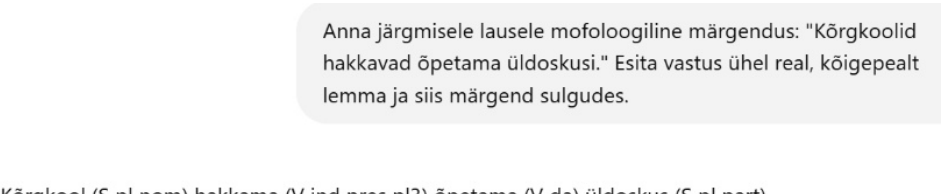
3. Rakendusviisid

Eeltreenitud SKM-ide promptimist võib mõneti vaadelda kui suurest tekstikorpusest päringute tegemist. Kasutajana aga tuleb meeles pidada, et paljudel juhtudel ei tea me täpselt, millised andmed sellises korpuses on ja kas korpuse sisu vastab ülesande püstitusele. SKM-ide rakendusviisid sõltuvad kasutaja eesmärgist, võimekusest, olemasolevatest ressurssidest ja huvist. Järgnevalt vaatleme põhilisi kasutusviise koos näidetega.

ChatGPT eeskujul on paljudel SKM-idel olemas **juturoboti** versioonid (nt Google'i Gemini ja Gemini Advanced[9], Anthropicu Claude[10], Microsofti Copilot[11] jne), mis on veebilehitseja ja/või rakenduse teel lihtsasti kõigile kasutatavad ilma tehnilisi oskusi eeldamata. See võimaldab mudelitega lihtsalt tutvust teha ning saada aimu tööpõhimõtetest ja võimekusest. Näiteks saab mudelitelt uurida uute sõnade kohta (vt joonis 1) või lasta tekste märgendada (vt joonis 2), SKM-id võivad assisteerida ka koodi kirjutamisel.



Joonis 1. Kuvatõmmis vestlusest rakendusega Gemini Advanced 2.0 Flash



Joonis 2. Kuvatõmmis vestlusest rakendusega ChatGPT 4o

[9] https://gemini.google.com/app

[10] https://claude.ai/chats

[11] https://www.bing.com/chat

Olenevalt mudelist, platvormist ja soovist mudeli kasutamise eest maksta on juturobotite abiga võimalik lahendada ka keerulisemaid ülesandeid, kui vaid fakte küsida. Näiteks on juturobotid võimelised andmeid analüüsima ja tulemusi visualiseerima nii, et kasutaja ei pea tegema muud, kui sisestama kirjelduse sellest, mida ta väljundina soovib (vt joonis 3).

Joonis 3. Pildil on ChatGPT väljund, kui viip on järgmine: „Võta aluseks kõige sagedasemad eesti keele sõnad, jäta alles ainult verbid, adjektiivid, noomenid ja adverbid, visualiseeri top 100 nii et verbid on sinised, adjektiivid punased, noomenid rohelised ja adverbid mustad. Pealkirjaks pane „Kõige sagedasemad sõnad eesti keeles“ ja lisa viide allikale.“

Näeme, et ChatGPT järgib küll viibas antud juhendeid, kuid tulemuse korrektsuse peab kasutaja siiski **kriitiliselt** üle vaatama. Näiteks viitab mudel Eesti Keele Instituudi korpusele, mida tegelikult ei eksisteeri, ning sõnapilvest on kindlasti puudu eesti keeles väga sagedasi sõnu, nt *olema* (vrd sagedusloenditega ptk-s 5). See näide tõestab, et mudelid võivad väga enesekindlalt toota väljundeid, mis tegelikult ei päde. Võimaluse korral peaks mudelile ette andma kontrollitud andmed, aga ka siis tuleks vastus üle vaadata, sest oma andmete sisestamine ei välista mudelite omaloomingut. Pane tähele, et erinevad turul olevad tekstirobotid ei pruugi sama viiba andmisel täpselt sama väljundit anda.

Kui lihtsalt kasutatavad juturobotid on heaks alguspunktiks, siis suuremate tekstimassiivide töötlemiseks võiks ühekaupa viipade andmise asemel kasutada mudeleid **rakendusliideste** (API-de) kaudu. Selline kasutusviis eeldab küll algtasemel programmeerimisoskusi, kuid võimaldab tegevusi automatiseerida (nt sama viipa korrata suure andmestiku ja/või paljude failide peal korraga) või mudel

oma töövoogu integreerida. Seejuures on paljudel juhtudel mõistlikum kasutada mudeleid, mis pole häälestatud dialoogi pidama (nt Gemini 2.0 Flash Gemini juturoboti asemel) ning on paindlikumad lahendama spetsiifilisemaid ülesandeid.

Lisaks promptimisele on mudelite rakendamisel API kaudu võimalik määrata ka mitmete **parameetrite väärtusi**. Kõige levinumad parameetrid, mida kasutaja ise mudida saab, on **mudeli valik** ja **temperatuur**, mis lubab mudeli väljundit teha rohkem või vähem juhuslikuks ja loomingulisemaks. Reeglina tähendab madalam temperatuur etteaimatavamat väljundit kui kõrgem temperatuur. Olenevalt mudelist on lisaks võimalik määrata maksimaalne sõnede (ingl *token*) arv ja teisigi seesuguseid parameetreid. **Sõned** on väiksem keeleüksus, mida mudel töötleb, ja see võib, aga ei pruugi võrduda sõnaga või sõnaosaga. Tekstide üksustamine sõnedeks ehk sõnestamine aitab mudelil keelt osadeks tükeldada ja sedakaudu tööd kiiremini ja tõhusamalt teha. Sõnede arv on ka mudelite kasutamise hinnastamise aluseks, sest tihtipeale oleneb just sisendi ja väljundi hind sellest, kui palju sõnesid need sisaldavad.

Oma andmete lisamist mudelitesse võimaldavad nii mõned juturobotite rakendused kui ka mudelite kasutamine API kaudu. Kui see vajadus tekib, siis peab esmalt läbi mõtlema selle, kas andmed sisaldavad infot, mida ei peaks/võiks kolmandate osapooltega ehk mudelite omanikega jagama. **Kui andmed on tundlikud**, siis neid ei peaks laskma töödelda mudelitel, mille üle puudub teadmine, kuhu andmed edasi liiguvad. Sellisel juhul võib olla mõistlikum rakendada mudelit, millel on avatud lähtekood ja mida saab lokaalselt jooksutada.

```
Gemini 2.0 Flash
suurendama: kasumit suurendama, müüki suurendama, efektiivsust suurendama
valima: parimat kandidaati valima, sobivaimat varianti valima, uut presidenti valima
lugema: raamatut lugema, ajalehte lugema, sõnumeid lugema
tantsima: rahvatantsu tantsima, tangot tantsima, diskol tantsima
ümber tegema: korterit ümber tegema, plaane ümber tegema, riideid ümber tegema
kirjutama: kirja kirjutama, raamatut kirjutama, blogipostitust kirjutama

Claude 3.5 Sonnet
suurendama: tootmist suurendama, kasumit suurendama, mahtu suurendama
valima: presidenti valima, esindajat valima, juhti valima
lugema: raamatut lugema, ajalehte lugema, uudiseid lugema
tantsima: valssi tantsima, koos tantsima, peol tantsima
ümber tegema: maja ümber tegema, seadust ümber tegema, teksti ümber tegema
kirjutama: raamatut kirjutama, artiklit kirjutama, luuletust kirjutama

GPT-4o
suurendama: mõju, mahtu, kiirust
valima: presidenti, ametikohta, nimekirjast
lugema: raamatut, ajalehte, kokku
tantsima: valsil, laval, muusika saatel
ümber tegema: plaani, seadust, kodutööd
kirjutama: artiklit, käsitsi, paberile
```

Joonis 4. Gemini 2.0 Flash, Claude 3.5 Sonnet ja GPT-4o väljundid, kui viip on järgmine: „Mis on eesti keeles järgmiste sõnade kolm kollokatsiooni? Formaat võiks olla järgmine: „sõna: komadega eraldatud kollokatsioonid“.“

Joonisel 4 on välja toodud kolme SKM-i tulemused, kui sisend on täpselt sama fail. See koosneb lemmadest, iga lemma eraldi real. Päring tehti keelemudelitest nende API-de kaudu, koodi modifitseeriti nii, et ta loeks sisendsõna TXT-failist ning kirjutaks vastused samuti TXT-vormingus faili. Viip oli kõikidele mudelitele täpselt sama: „Mis on eesti keeles järgmiste sõnade kolm kollokatsiooni? Formaat võiks olla järgmine: sõna: komadega eraldatud kollokatsioonid." Tulemusest on näha, et kõik mudelid saavad käsust aru ja järgivad ka etteantud vormingu nõudeid. See, et mudelid annavad väga erinevaid vastuseid, ei ole üllatav, sest nad pole (arvatavasti) treenitud samamoodi. Miks mudelid sellised vastused annavad, pole ainult selle näite põhjal seletatav.

4. Promptimine

Vaatamata sellele, kuidas SKM-e rakendatakse, on kasutaja ülesandeks sisestada oskuslik viip, mis juhendab mudelit tegema seda, mida kasutaja soovib. Välja on kujunenud mõned üldised nipid, kuidas promptida nii, et SKM-i väljund oleks kasutajat rahuldav, kuid kõige parema tulemuse tagab eksperimenteerimine erinevate viipadega.

Üldised promptimise põhitõed on järgmised:

- **Alusta lihtsalt** ning lisa prompti järk-järgult aspekte, mis on soovitava väljundi jaoks olulised.
- Tihti aitab see, kui defineerid ülesande kontekstis mudelile **rolli**, nt „oled keeleteaduse üliõpilase abiline", „mõtle nagu inimene tänavalt" jne.
- Ole **täpne ja konkreetne**, üldised viibad annavad halvema tulemuse, nt „leia tekstist väljendverbid" *vs.* „leia tekstist väljendverbid ehk sõnaühendid, mis koosnevad käändsõnast ja verbist, käändsõnu võib olla ühendis mitu, kirjuta iga tuvastatud väljendverb eraldi reale".
- Kui **väljundi formaat** on oluline, siis on tihtipeale vajalik anda ette ükskaks näidet, nt „tuvasta tekstist kõik isikute, organisatsioonide ja kohanimed" *vs*. „tuvasta tekstist kõik isikute, organisatsioonide ja kohanimed, soovitud formaat on järgmine:
  - Isikunimed: tabulaatoriga eraldatud nimekiri isikunimedest
  - Organisatsioonide nimed: tabulaatoriga eraldatud nimekiri organisatsioonide nimedest
  - Kohanimed: tabulaatoriga eraldatud nimekiri kohanimedest"
- **Kirjelda, mida sa tahad, et mudel teeks**, mitte seda, mida ta tegema ei peaks, nt „kirjuta etteantud tekstidest sisukokkuvõtted, ära ole liiga napisõnaline, aga ära kirjuta rohkem kui viis lauset" *vs.* „kirjuta etteantud tekstidest 2- kuni 5-lauselised sisukokkuvõtted".

- Proovi esialgu promptida **näiteid ette andmata** (ingl *zero-shot*), kui see ei toimi, siis anna ette paar näidet (ingl *few-shot*) ja kui ka see ei toimi, siis mõtle mudeli peenhäälestamise peale.

Pane tähele, et paljud rakenduste või API-de kaudu kättesaadavad mudelid ei ole eraldi treenitud spetsiifiliste (eesti keele teadmisi sisaldavate) andmete peal ja seega pole ka oskuslikust promptimisest palju kasu. Seesugused mudelid polegi mõeldud ülesanneteks, mis nõuavad erialateadmisi ning hea tulemuse saavutamiseks peaks arendama ja/või rakendama peenhäälestatud mudeleid.

5. Kitsaskohad

SKM-ide rakendamisel tuleb kasutajana arvesse võtta mitmeid SKM-ide rakendamise kitsaskohti, mis võivad puudutada nii mudelite kättesaadavust kui ka tulemuste sisu.

Kõik mudelid **pole kõigile kättesaadavad**. Enimkasutatavad ja praegu ka eesti keeles kõige võimekamad mudelid on loonud kommertsettevõtted, kelle mudelite kasutamine on tasuline. Kui kõrvale jätta rida juturoboteid, mis on mingi piirini tasuta ja kõigile kättesaadavad, siis mudelite rakendamine suure andmehulga analüüsimiseks vajalike rakendusliideste kaudu on üldjuhul tasuline. Vabavaraliste mudelite kasutamine nõuab reeglina väga head arvutusvõimsust, mis samuti pole paljudele inimestele kättesaadav. Lisaks erinevatele ressurssidele, mida SKM-ide rakendamine nõuab, pole kasutajatel garantiid, et täna kättesaadav kommertsmudel on samamoodi kasutatav ka homme. Lisaks sellele on mõned mudelid kättesaadavad vaid teatud regioonides.

Kommertsmudelite treenimist puudutav info **pole avalikult kättesaadav**. See tähendab, et paljudel kasutajatel pole kindlat ja usaldusväärset infot selle kohta, kuidas on mudel treenitud – missugused on mudelite treeningandmed ja kuidas nad on instrueeritud väljundit genereerima, millises ulatuses ja kuidas on mudeleid tsenseeritud, kui palju on mõeldud mudelite keskkonnamõjule ja eetilisusele. See kõik kokku teeb mudelite tulemuste lahtiseletamise ja analüüsi keerukaks.

Üldjuhul on SKM-idel nii **sisendi kui väljundi maksimaalsed pikkused** sõnedes kindlaks määratud. Erinevate mudelite seas on need väärtused varieeruvad, aga enne keelemudeli rakendamist võiks neid teada.

SKM-ide väljundid **ei ole lihtsalt korratavad**, sest SKM-id ei ole oma olemuselt 100% deterministlikud. See tähendab, et sama viiba kordamine erinevatel ajahetkedel ei taga sama vastust. Kasutajal võib olla võimalus mudeli väljundi varieeruvust (temperatuuri) muuta, aga mitte kõikidel kasutusjuhtudel.

SKM-id hallutsineerivad ehk genereerivad (tihti väga enesekindlalt) väljundit, milles leidub **valeinfot**. Seega on kasutaja ülesandeks hinnata mudeli väljundi tõepärasust. Hallutsineerimise vähendamiseks on olemas rida meetodeid (nt ingl

retrieval augmented generation, RAG), mille rakendamine aga eeldab kasutajalt häid tehnilisi oskusi või väljatöötatud, üldjuhul tasuliste lahenduste rakendamist.

Kui mudeli treeningandmestik on mingis suunas **kallutatud**, siis tulemuseks on ka kallutatud mudel. See tähendab, et kui mudel on treenitud veebimaterjalidel, mis sisaldavad näiteks rassistlikku sisu, siis ka mudel võib rassistlikku väljundit toota. Mudelite loojad küll püüavad kallutatust vähendada, kuid kõikide inimgruppide suhtes võrdselt käituvat mudelit on raske välja anda. Teadlik kasutaja peab meeles pidama, et ka promptida tuleks vastutustundlikult, et mitte osaleda kahjuliku sisu taastootmises. Mõned mudelid on võimelised genereerima väga kahjustavat sisu (nt kuidas endale või teistele inimestele liiga teha) ja seetõttu peab kasutaja lisaks enda tegevustele ka mudeli genereeritud sisu kriitiliselt hindama.

Kokkuvõte

SKM-id on juba praegu väga head ja saavad iga päevaga paremaks. Mudelite keeleteaduslike ülesannete lahendamise võimekuse hindamine on lapsekingades, mis tähendab, et nendega katsetamine kõikvõimalike ülesannete lahendamiseks on SKM-ide arengu seisukohalt väga väärtuslik, eriti kui see käib koos inimekspertide hinnangutega mudelite tulemustele. Kuigi olemasolev kirjandus annab juba vihjeid selle kohta, milline võiks olla SKM-ide roll korpuslingvistikas (nt töömahukate ülesannete nagu andmete eeltöötlus, korpuste märgendamine vm automatiseerimine), on praegu veel vara usaldusväärselt soovitada kindlaid mudeleid ja ülesandeid, kus neid igasuguse kahtluseta rakendada võiks.

Näidisuurimuse valmimist on toetanud EKKD projekt EKKD-III1 „Suurte keelemudelite rakendamine leksikograafias: uued võimalused ja väljakutsed".

Kirjandus

Curry, Niall, Paul Baker & Gavin Brookes. 2024. Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics* 4(1). 100082. https://doi.org/10.1016/j.acorp.2023.100082.

De Schryver, Gilles-Maurice. 2023. Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography* 36(4). 355–387. https://doi.org/10.1093/ijl/ecad021.

Karjus, Andres. 2023. Machine-assisted mixed methods: Augmenting humanities and social sciences with artificial intelligence. https://doi.org/10.48550/ARXIV.2309.14379.

Karjus, Andres & Christine Cuskley. 2024. Evolving linguistic divergence on polarizing social media. *Humanities and Social Sciences Communications* 11(1). 422. https://doi.org/10.1057/s41599-024-02922-9.

Kuzman, Taja, Igor Mozetič & Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction* 5(3). 1149–1175. https://doi.org/10.3390/make5030059.

Uchida, Satoru. 2024. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics* 4(1). 100089. https://doi.org/10.1016/j.acorp.2024.100089.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is All you Need. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (toim), *Advances in Neural Information Processing Systems*, kd 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Yu, Danni, Luyang Li, Hang Su & Matteo Fuoli. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics* 29(4). 534–651. https://doi.org/10.1075/ijcl.23087.yu.

Terminite loetelu