Tartu Ülikool

Filosoofiateaduskond

Eesti ja üldkeeleteaduse instituut

Arvutilingvistika eriala

Riin Kirt

Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning „Eesti keele seletava sõnaraamatu“ märksõnaloendi võrdlus

Magistritöö

Juhendaja Kadri Muischnek

Tartu 2013

Sisukord

Tabelite ja jooniste sisu

Tabelid

Joonised

Sissejuhatus

Sõnasageduste uurimine on läbi kvantitatiivse keeleuurimise ajaloo olnud oluline. Sõnasageduse loendid muutusid populaarseks enne, kui kvantitatiivset lingvistikat hakati käsitlema omaette uurimisharuna. Tänapäevalgi on sõnastatistika uurimine empiirilise lingvistika alades, näiteks korpuslingvistikas populaarne teema. (Popescu 2009: v) Käesoleva rakenduslingvistika valdkonda kuuluva magistritöö eesmärgiks on esiteks, luua uus eesti keele ressurss – sõnasagedusloendid Tasakaalus korpuse (TK) põhjal ja teiseks, hinnata korpuse sõnavaralist rikkust ja tasakaalustatust (või mitte-tasakaalustatust), võrreldes korpuse sagedusloendit valikuliselt „Eesti keele seletava sõnaraamatu“ märksõnaloendiga. Töö võrdluse osas on lisaks Tasakaalus korpusele kaasatud korpuse materjalina ka osa „Eesti keele koondkorpusest“ (KK), mille kohta töös kasutatakse nimetust Koondkorpus.

Magistritöö käsitleb järgmiste tööde tegemist:

- uue iseseisva keeleressursi loomist, nimelt eesti keele sõnasagedusloendite koostamist Tasakaalus korpuse põhjal ja selleks vajalikke eeltöid;
- korpuse sagedusloendi valikulist võrdlust „Eesti keele seletava sõnaraamatu“ märksõnaloendiga (edaspidi kasutatakse ka lühendit EKSS).

Magistritöö koosneb sissejuhatusest, kuuest peatükist, kokkuvõttest, kirjanduse loetelust ja ingliskeelsest resümeest ning lisadest. Töö on struktureeritud järgnevalt: esimeses peatükis antakse lühiülevaade eesti kirjakeele põhjal koostatud sagedusloenditest, sagedusloendite kasutamisest ja koostamise võimalustest. Teises peatükis seletatakse lahti töö seisukohast olulised keeleteaduslikud mõisted. Kolmandas peatükis keskendutakse sõnavara jagunemisele. Esmalt antakse ülevaade Zipfi seadusest kui üht sõnavara jaotumist iseloomustavast suurusest ning järgnevates alajaotustes kirjeldatakse

sõnavara jagunemist nii tarvitusulatuse kui ka funktsionaalstiilide alusel. Neljandas peatükis tutvustatakse töös kasutatud andmekogusid: peatüki esimeses alajaotuses Koondkorpust ja Tasakaalus korpust ning teises EKSS'i. Viiendas peatükis kirjeldatakse sagedusloendite koostamiseks vajaliku materjali ettevalmistamist ning antakse statistiline ülevaade korpuses sisalduvast materjalist ja koostatakse erinevad kasutajatele vajalikuks osutuda võivad eesti keele sagedusloendid. Peatüki lõpus kõrvutatakse koostatud sagedussõnastikku „Eesti kirjakeele sagedussõnastikust" (Kaalep, Muischnek 2002) välja jäänud sagedasemate sõnade loendiga. Viimase osa tööst moodustab kuues peatükk, milles võrreldakse valikuliselt korpuse sagedusloendeid EKSS'i märksõnaloendiga. Valikulisest võrdlusest sõnaraamatuga leitakse nii korpuses kui ka sõnaraamatus esinevad verbid ja lihtnimisõnad ning loendid sõnaraamatus olevast ja korpusest puuduvast ning vastupidi – korpuses olevast, kuid seejuures sõnaraamatust puuduvast sõnavarast.

Töö praktiliseks väärtuseks olevad sagedusloendid ja võrdlusest saadud sõnade loendid koos seletusega on esitatud töö lisas.

1. Sagedussõnastikud

1.1. Sagedussõnastikud Eestis

Eestis hakati sõnavarastatistikaga tegelema 1960. aastatel, kui Juhan Tuldava juhtimisel moodustati keelestatistika uurimisrühm. Erinevaid sõnavarastatistilisi ülevaateid avaldati Tartu Ülikooli kogumiksarjades „Linguistica“, „Töid keelestatistika alalt“, „Keelestatistika“, viimases avaldati ka tugeva teoreetilise taustaga eesti keele sagedussõnastik (Kaasik jt 1977), mis baseerus 100000-sõnalisel 1960. aastate proosakirjanduse autori kõne tekstidel. (Kasik 2011: 210)

2001. aastal koostati Eesti Keele Instituudis terviklikku tekstikorpust kattev grammatiline sagedussõnastik „Seadusetekstide grammatiline sagedussõnastik“. Sagedussõnastiku lähtematerjaliks oli kümnest seadusest koosnev tekstikorpus (Riigi Teataja tekstid aastatest 1996–1997, kokku üle 100 000 sõnavormi): kõik sõnavormid analüüsiti kõigepealt morfoloogiliselt, seejärel ühestati mitmesed tulemused. Nii alfabeedi kui ka sageduse järgi järjestati kaks sõnastikku: lemmade sõnastik (4385 eri sõna) ja sõnavormide sõnastik (11 556 eri sõnavormi). (Viks, Hein: 2001)

Aastal 2002 koostati eesti keele sagedussõnaraamat „Eesti kirjakeele sagedussõnastik“ (Kaalep, Muischnek 2002), mis on 1990. aastate ilukirjanduse ja ajakirjanduse 1miljoni sõna suuruse korpuse põhjal peaaegu täisautomaatselt arvutatud lemmatiseeritud sõnade sagedusloend. Mõlema korpuses kasutatud tekstiklassi maht oli ümmarguselt pool miljonit sõna. Ilukirjanduse tekstidena kasutati eesti keele 90ndate aastate ilukirjanduse allkorpuse tekste, iga väljavõtte pikkus oli 2000 sõna. Ajalehtedest kasutati terviknumbreid, mitte 2000-sõnalisi katkeid ning lisaks 90ndate aastate ajakirjanduse allkorpuse materjalile kasutati ka ajalehtede internetiarhiivide tekste. (Kaalep, Muischnek 2002: 9) Kuna sagedussõnastiku eesmärgiks oli esitada tavalisi eesti keele sõnu, siis sõna sõnastikku lisamise üheks mõõdupuuks oli selle esinemine mõlemas kasutatud tekstiklassis kokku vähemalt viis korda (Kaalep, Muischnek 2002: 10). Eraldi sagedusloend esitati ka sagedasematest ühest või teisest korpuse osast

(ajakirjandus, ilukirjandus) puudunud ja seega kogu korpuse sagedusloendist välja jäänud sõnadest. Nimetatud loend on vaatluse all ka peatükis 5.3.

Käesolevas töös koostatakse eesti keele sagedussõnastik (ptk 5.), mis baseerub 15 miljoni sõna suurusel Tasakaalus korpusel (vt ptk 4.1.). Tasakaalus korpuse tekstid on terviktekstid, st sinna pole valitud tekstikatkeid erinevatest tekstidest. Sõnasageduste arvutamiseks on töös kasutatud tekstisõne absoluutsagedust, mis on sõne tekstisiseste esinemiste summa. Peatükis 1.3. tutvustatakse ka mõningaid teisi sagedussõnastike koostamise viise.

Sagedussõnastik on oluline keeleressurss, kuid universaalset kõigi ülesannete tarbeks sobilikku sagedussõnaraamatut pole võimalik luua. Sagedusloendid koostatakse enamasti lähtuvalt lahendamist vajavast probleemist. Näiteks kui on tarvis koostada sagedussõnastik füüsika valdkonda kuuluvatest sõnadest, siis peab selle aluseks olev korpus koosnema vastavatest tekstidest. Samas on ka üldkeele sõnasagedusloend väärtuslik keeleressurss.

Järgnevalt sagedussõnastike kasutamisvõimalustest.

1.2. Sagedussõnastike kasutamine

Sagedussõnastikud on populaarsed nii teoreetilistel kui ka praktilistel kaalutlustel. Nad pakuvad kiiret pilguheitu keelesõnavarasse ja seeläbi on võimalik leida nii keele põhisõnavara (kõige sagedasemad sõnad) kui ka keele perifeeria alasse kuuluv sõnavara (sagedusloendi lõpus olevad sõnad). (Hlavácová 2006: 273)

Keeleteadlased koostavad sagedusloendite põhjal katseid ning uurimusi nii leksikoloogias, semantikas, psühholingvistikas, morfoloogias kui ka teistel aladel ning arvutilingvistid kasutavad sagedusloendeid keeletehnoloogias loomuliku keele rakendustes.

Sõnasagedus on pidepunktiks ka leksikograafidele, otsustamaks sõna sõnaraamatusse kuulumise või mittekuulumise üle. Seega sagedusloendid on suureks abiks keele(õppe) materjalide koostajatele: omades eelteadmisi sõnade kasutamisest, on võimalik õppematerjalides keskenduda just reaalses elus tarvitusel olevatele sõnadele ning seeläbi pakkuda kasutajatele paremaid materjale. Keeleõppijatel on võimalik tänu teadmisele sõnade sagedusest täiendada või omandada keele sõnavara alustades just enamkasutatavatest sõnadest. Lähtuvalt eelnevast on võimalik ka õpetajatel kontrollida, kui hästi õpilased kindla sageduspiirkonna sõnu tunnevad. (Wfd)

Sõnasagedus on sõna omadus, mis on vaadeldav puhtas (st analüüsimata tekstis), tänu sellele on ta kättesaadav ka mittelingvistidele. Automaatseid sõnasageduse tulemusi kasutatakse näiteks ka tüpograafias, stenograafias, psühholoogias, psühhiaatrias, krüptograafias ja tarkvara loomisel. (Popescu 2009: 1)

Üldine sagedussõnaraamat võimaldab, nagu eelnevalt mainitud, heita pilku keelesõnavarasse ja on üheks mõõdupuuks/abivahendiks uurimisülesannete lahendamisel. Näiteks on „Eesti kirjakeele sagedussõnastikku“ (Kaalep, Muischnek 2002) kasutatud eesti keele põhisõnavara sõnastiku koostamisel, umbes 4000 sõnastiku märksõnadest 90% kuulub sagedussõnastiku järgi 10000 sagedasema märksõna hulka ning 71% kõige sagedasema 3000 sõna hulka. Lisaks sellele on 2000 sagedasema sõna hulka kuuluvad sõnad varustatud sagedusinfoga, mis aitab kasutajal otsustada, milliseid

sõnu esmajoones õppida. (Kallas, Tuulik 2011: 66-67) Kui nimetatud artiklis kasutati sagedussõnastikku põhisõnavara sõnastiku loomisel, siis artiklis „Sõnavara loomulik rikkus haritud keeleoskaja tekstides“ (Pajupuu jt 2009) hinnati sagedussõnastiku abil sõnavara raskust, hindamisel lähtuti seisukohast, et sagedamini esinevaid sõnu teatakse paremini kui harvaesinevaid. Kõrgtaseme eesti keele eksami edukalt sooritanud kohalike venelaste sõnavara ja eesti keelt emakeelena rääkivate kõrgharidusega mittefiloloogide sõnavara raskuse hindamiseks võrreldi sõnu „Eesti kirjakeele sagedussõnastikuga“ (Kaalep, Muischnek 2002), millest 10000 sagedasemat sõna moodustasid keele põhisõnavara, nendest 3000 kõige sagedasemat moodustasid tavalise sõnavara ning väljapoole 10000 sõna piire jäävad sõnad loeti harvaesineva sõnavara alla. (Pajupuu jt 2009: 190) Sarnast jaotust kasutati C-1 taseme eesti keele oskuse hindamiseks (Kerge 2008) ning 2010. aastal (Ehala jt 2010) hinnati sarnase jaotuse abil kõrgkooli üliõpilaste eesti keele oskuse taset. Artiklis „Eesti keele kasutusvariandid: korpustest tulenev käändevormide võrdlev analüüs“ (Elson, Matsak 2009: 80) on sõnade sagedusandmed võetud “Eesti kirjakeele sagedussõnastikust” (Kaalep, Muischnek 2002). Sagedussõnastikke on kasutatud ka üliõpilastöödes, nt Sirje Rammo magistritöös „Eesti keele õpik täiskasvanud õppijale“ (Rammo 2010: 19) on õpikusse sõnavara valimisel arvestatud „Eesti kirjakeele sagedussõnastikus“ (Kaalep, Muischnek 2002) esitatud sõnade esinemise sagedustega ja Merike Hüti bakalaureusetöös „Kakskeelsete eelkooliealiste laste grammatilised oskused: kolme juhtumi kirjeldus“ (Hütt 2012) kasutati sagedussõnastikku uurimisaluste sõnade leidmiseks.

Järgnevas peatükis kirjeldatakse lähemalt sagedusloendite koostamise viise.

1.3. Sagedussõnastike koostamise viisid

Käesolevas töös on sagedusloendi tegemisel sarnaselt „Eesti kirjakeele sagedussõnastikuga“ (Kaalep, Muischnek 2002) kasutatud sõnade absoluutsagedust, see tähendab, et kokku on arvutatud kõik sõna tekstis esinemise sagedused. See sageduste arvutamise viis on aga tugevalt seotud korpuse tekstivalikuga. Sagedussõnaraamatute tegemisel tuleks tähelepanu pöörata ka suuliste ja kirjalike tekstide osakaalule allikmaterjalis, kasutatud Tasakaalus korpus ei sisalda suulise keele tekste, seega on sagedussõnastikus esindatud vaid valikulise kirjaliku keele sagedus. Suulise keele kaasamine sagedusloendite tegemisse mõjutaks saadud tulemusi, näiteks on *British National Corpus*'e (BNC) põhjal koostatud suulise ja kirjaliku keele sõnasageduste võrdlusest selgunud, et 50-st kõige sagedasemast kirjaliku keele teksti sõnadest vaid 33 olid ühised suulise keele viiekümne sagedasema sõnaga. (Leech jt 2001: xi) See tähendab, et ka käesolevas töös koostatud sagedusloendid erineksid märgatavalt, kui kasutatud korpus sisaldaks ka suulise keele tekste.

Sagedusloendite koostamise meetodid sõltuvad suuresti sellest, mida sagedusloenditega soovitakse teha, „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) eesmärgiks oli esitada tavalisemate eesti keele sõnade loend. Kasutati lähenemist, mille kohaselt mõlemas loendite aluseks olnud tekstiklassis teatud arvul esinenud sõna on „tavalisem“, kui sõna, mis esineb arvukalt küll ühes tekstiklassis, kuid samas puudub teisest. Sagedusloendi koostamisel arvati sagedusloendisse sõna, mis esines korraga nii aja- kui ka ilukirjanduse tekstis ja neis kahes kokku vähemalt viis korda (Kaalep, Muischnek 2002: 10). Sõna tavalisus on kõrgelt korreleerunud sõna sagedusega. Tavalisemad sõnad kalduvad tekstides sagedamini esinema ja vastupidi – harva esinevad sõnad pole nii tavalised. Kuid siiski tavalisus ja sagedus pole omavahel sünonüümsed, keelekasutajale tavalisena tunduv sõna võib olla madala esinemissagedusega ja vastupidi. Näiteks on „Eesti kirjakeele sagedussõnastikust“ (Kaalep, Muischnek 2002) välja jäänud tavalisena tunduv eesti keele sõna 'kägu'. (Kaalep, Muischnek 2004: 57)

Kuidas saavutada seda, et sagedussõnaraamatu andmed iseloomustaksid ka sõna „tavalisust“ mitte ainult sõna sagedust – kirjeldab Jaroslava Hlaváčová artiklis „New Approach to Frequency Dictionaries - Czech Example“ (Hlavácová 2006). Tšehhi keele sagedusloendite koostamisel kasutati sõna absoluutsageduste (kõik tekstis esinenud vormid kokku liidetuna) arvutamise asemel sõnade järjestamist keskmise vähendatud sageduse (*ARF= average reduced frequency*) väärtuse alusel. See tähendab, et just sõna ARF'i väärtust, mitte sõna absoluutsagedust kasutati enamikku loenditesse, kaasa arvatud kõige sagedasema 50 000 tavalise sõna loendisse sõnade valimisel. (Čermák, Křen 2005: 3) Alusmaterjalina kasutati *Czech National Corpus*'t, mis sisaldab 100 miljonit sõnavormi, mis jaotuvad kolme tekstiklassi vahel järgnevalt: 60% ajakirjandustekste, 25% teadustekste ja 15% ilukirjandustekste. Vähendatud sageduse saamiseks jagati korpus positsioonideks (alates esimesest sõnast *1* kuni viimase sõnani *N*), igal positsioonil asus üks sõna. Seejärel arvutati selle konkreetse sõna, mille vähendatud sagedust sooviti leida, sagedus kogu korpuses (*f*) ning vastavalt saadud tulemusele jagati korpus *f* osadeks. Kui korpus oleks ühtlaselt jaotunud, siis sisaldaks iga jaotuse osa ainult ühte aluseks võetud sõna, tavaliselt sisaldab aga mõni korpuse osa rohkem sõnu ja teises ei pruugi seda üldse esineda. Vähendatud sagedusena võeti arvesse nende jaotuste arv, milles sõna esines vähemalt ühe korra. Lähenemise puuduseks on see, et näiteks korpuses vaid ühes väikeses tekstis 2 korda esinenud sõna vähendatud sageduse väärtus sõltub sellest, kas korpuse poolitamisel (*f=2*) jäävad mõlemad korpuses esinemise juhud samasse osasse või poolitatakse tekst sõnade esinemise vahelt. Seega sõna vähendatud sageduseks võib sõltuvalt korpuse poolitamise kohast saada kas 1 või 2. Selle kitsaskoha lahendamiseks kasutatakse keskmiselt vähendatud sagedust. Keskmiselt vähendatud sagedus erineb vähendatud sagedusest selle poolest, et korpuse positsioonideks jagamisel kujutatakse korpust ette sektorina. Sektori puhul on võimalik nihutada korpuse osadeks jagamise alguspunkti ning lõplikuks sagedusmõõduks võetakse keskmine aritmeetiline sagedus kõikidest võimalikest alguspunktidest arvutatud sagedustest. Mida ühtlasemalt on sõna korpuses jaotunud, seda väiksem on sõna erinevus absoluutse ja keskmiselt vähendatud sageduse vahel ja vastupidi. (Hlavácová 2006: 374-375)

ARF-i põhjal koostatud nimekirjad erinevad tavalistest sagedussõnastikest peamiselt selle poolest, et erialaterminite ja pärisnimede, mis esinevad vaid vähestes korpuse tekstides, sagedused langevad märgatavalt, samas kui ühtlaselt jaotunud sõnade puhul ARF'i ja sõnasageduste väärtused ei erine üksteisest nii drastiliselt. (Čermák, Křen 2005: 3)

Keskmiselt vähendatud sageduste abil on võimalik uurida sama absoluutsageduse väärtuse saanud sõnade tavalisust (vt Tabel 1), näiteks tšehhikeelsed sõnad *molekulovy* (molekulaarne) ja *nahromadit* (kuhjuma) said mõlemad CNC korpuse põhjal koostatud sagedusloendis absoluutsageduseks 223, keskmiselt vähendatud sagedused aga erinesid üksteisest märgatavalt (Hlavácová 2006: 377).

Tabel 1. **Keskmiselt vähendatud sõnasagedus** *(Hlavácová, 2006, lk 377)***.**

Sõna	**Tõlge**	**ARF järjestus**	**ARF**	**Sagedus järjestus**	**Sagedus**	**Ilu(%)**	**Tea(%)**	**Aja(%)**
Molekulovy	molekulaarne	37502	22	18959	223	0	99	1
Nahromadit	kuhjuma	14970	136	18915	223	34	43	23

(Originaaltabelit on muudetud (inglisekeelne tõlge on asendatud eestikeelse tõlkega)).

Nii on sõna *kuhjuma* keskmiselt vähendatud sageduse alusel sagedasem kui sõna *molekulaarne*. Sageduse erinevus paistab välja ka, vaadeldes sõna esinemist kolmes tekstiklassis: *kuhjuma* on enam vähem ühtlaselt esindatud aja-, ilu- ja teaduskirjanduse tekstides, seevastu sõna *molekulaarne* on sagedane just teaduskirjanduse valdkonda kuuluvatele tekstidele. (Hlavácová 2006: 375)

Sageduse jaotumust suure korpuse erinevate žanrite vahel kirjeldab Hanhong Li keele põhivara sõnastiku koostamisest rääkivas artiklis „Word Frequency Distribution for Electronic Learner's Dictionaries" (Li 2010). Artiklis tõstatati küsimus, kumb sageduse arvutamise meetod on parem õppesõnaraamatute koostamisel, kas absoluutne või tekstižanrite siseselt koostatud jagatud sagedust arvestav sagedus. Töö eesmärgiks oli:

1) uurida jagatud sageduse meetodi kasutamist põhisõnavara valimisel,

2) koostada elektrooniline sõnaraamat, milles sagedus on jaotatud tekstižanrite kaupa.

Sagedusloendi koostamise aluseks oleva korpuse saamiseks valiti BNC 70 erinevast tekstižanrist välja 14 (9 kirjalikku ja 5 suulist), millest igaühest omakorda valiti korpusesse umbes 1 miljoni sõna suurune osa. Jaotatud sageduse arvutamiseks kasutati Carroll'i (Carroll 1970) statistilist meetodit *Um,* mis on kasutuskoefitsient erinevate tekstižanrite vahel jagatud sageduse märkimiseks; ta arvestab nii sõna hajuvuse kui ka üldsagedusega. Artiklis kõrvutati absoluutsageduse ja jagatud sageduse alusel koostatud sagedussõnastikud ning selgus, et need sõnad, mis on ühtlaselt jaotunud erinevates tekstžanrites, peaksid suurema tõenäosusega kuuluma põhisõnavarasse kui need, mis on sagedased vaid absoluutsageduse järgi. (Li 2010, 7)

Kuna sagedussõnastikes võivad ainult mõnes üksikus tekstis või ainult ühes tekstiklassis sageli esinevad sõnad sattuda sagedusloendi tippu, siis selle probleemi lahendamiseks on välja pakutud erinevaid eespool kirjeldatud lahendusi. Mõne kirjeldatud meetodi rakendamine sõnasageduste arvutamiseks Tasakaalus korpuse põhjal võiks olla käesoleva töö üks edasiarendusi.

Järgnevas peatükis seletatakse lahti keeleteaduslikud mõisted, millega töös kokku puututakse või mis on töö selguse seisukohast olulised.

2. Keeleteaduslikud mõisted

Peatükis esitatud keeleteaduslikke mõisteid on defineeritud peamiselt „Eesti keele grammatika“ (edaspidi EKG) ja „Eesti keele käsiraamatu“ (edaspidi EKK) abil. Magistritöö kuulub keeleteaduse harusse leksikoloogia ehk sõnavaraõpetus, mis uurib sõna ja sõnavara. Leksikoloogia harudeks on sõnasemantika, etümoloogia, fraseoloogia, onomastika, leksikostatistika ja leksikograafia. (EKK 2000: L 41-L 63)

Sagedussõnastike koostamiseks vajaliku alusmaterjali saamiseks oli oluline teisendada korpusematerjal sobilikule kujule, siinkohal puututi kokku morfoloogia mõistega. EKG kohaselt on **morfoloogia** ehk **vormiõpetus** grammatika osa, mis uurib grammatiliste tähenduste realiseerumist sõnavormides ning sõnavormide grammatiliste tähenduste äratundmist. (EKG 1995: §16) Samas on mõiste *morfoloogia* kasutuses ka laiemas tähenduses - morfoloogia valdkonda kuulub keelesüsteemi osa, mis puudutab tüvimorfeemide ja liide- ning tunnusmorfeemide omavahelist kombineeritust. Viimasel juhul hõlmab ta siis nii eelnevalt defineeritud vormiõpetust kui ka **sõnamoodus**tust, mis uurib sõnade moodustamist tüvimorfeemide omavahelise liitmise teel (liitsõnad nt *kassi+poeg*) või liitemorfeemide liitmise teel tüvimorfeemidele (tuletised, *kass/ilik*). (EKK 2000: SJ 11)

Morfoloogia põhiüksus on morfeem, mis on keelesüsteemi väikseim potentsiaalselt tähenduslik osa. Vastavalt sellele, kas morfeem kannab leksikaalset või grammatilist tähendust, jaguneb ta kaheks: **tüvimorfeem** ehk **tüvi** ja **tunnusmorfeem** ehk **morfoloogiline tunnus**. Morfeemidest pannakse kokku sõnad ja sõnavormid. **Sõna** (leksikaalne sõna ehk lekseem) on sama tüve alusel moodustatud sõnavormide kogum, mis võib koosneda nii ainult ühest kui ka mitmest tüvimorfeemist või hoopis tüvimorfeemi(de)st ja juurde lisatud tunnusemorfeemi(de)st. (EKG 1995: §90)

Töös koostatakse eraldi nii sõnavormide kui ka lemmade sagedusloendid. **Sõnavorm** on grammatiline sõna (iga grammatiline vorm), milles mingi leksikaalne sõna esineb, nt *kassi, kassidele, kassist* on kõik sõnavormid sõnast *kass*. Sõnavormi konkreetse esinemisjuhu kohta tekstis kasutatakse terminit **sõne**. **Lemma** ehk **sõnaraamatuvorm**

ehk **algvorm** on ühe sõnavormi esindaja sõnaraamatus. (EKK 2000: L 2) Eesti keele sõnaraamatutes on noomeni sõnaraamatuvorm ainsuse nimetav ja verbide sõnaraamatuvorm tavaliselt *ma*-infinitiiv. (Karlsson 2002: 215)

Keelesüsteemi eri tasanditel (morfoloogia, süntaks ja semantika) osutuvad olulisteks sõna erinevad omadused. Nimetatud keelekirjelduse tasandeid aitab omavahel siduda sõnade klassifitseerimine sõnaliigiti. **Sõnaliik** ehk **sõnakategooria** on leksikaalgrammatiline klass, millesse jaotuvad sõnad nii, et kindlalt grammatilist tähendust väljendavale sõnavormile vastab kindel süntaktiline funktsioon ning sellele omakorda kindel leksikaalne tähendus. (EKG 1995: §2) Sõnaliik lisatakse sõna tähtsaima morfosüntaktilise omadusena sõna leksikaalsesse esitusse ehk kirjesse. Sõnaliigid jagunevad avatud ja suletud klassideks. Avatud klassid on substantiivid, adjektiivid, verbid ja adverbid, suletud klassi moodustavad adpositsioonid, konjunktsioonid, pronoomenid jm. (Karlsson 2002: 218) Esimestesse klassidesse kuuluvad sõnad on mitmetähenduslikud semantilised morfeemid, seevastu viimastesse klassidesse kuuluvad sõnad on grammatilised morfeemid, mille arvukus erinevalt avatud klassi kuuluvatest sõnadest on piiritletud ning väga kergesti neid keelde juurde ei teki. (Karlsson 2002: 218)

Sõnade jaotamisel liikidesse tuleks tähelepanu pöörata ka sõna prototüüpsele esinemisele, sest leidub piiripealseid ning isegi lahendamatuid juhtumeid, mil mitmetähenduslik sõna võib kuuluda korraga erinevatesse sõnaliikidesse. Eesti keele sõnavara jaguneb kolme suuremasse morfoloogilisse klassi: käändsõnad, pöördsõnad ja muutumatud sõnad. (Karlsson 2002: 218)

„Uus keeleüksus võib keelde tulla mitmel viisil: uue sõnana – eesti keeles enamasti liitsõnana –, aga ka olemasoleva sõna uue tähendusena – polüseemiana.“ (Langemets 2010: 13) Seega tuleks ka sagedusloendites sisalduva materjali uurimisel arvestada polüseemiaga, mille kohaselt sõnad võivad olla mitmetähenduslikud. Lisaks polüseemiale on vajalik tähelepanu pöörata ka homonüümiale. Eesti keele käsiraamatu definitsiooni kohaselt on homonüümid samakujulised, aga erineva tähendusega sõnad. Seejuures peegeldub „samakujulisus“ kolmel erineval moel:

1. sama häälduse- ja kirjakuju, nt *tint* ‘*kala*’ ja *tint* ‘*kirjutusvedelik*’,
2. sama hääldusekuju (homofoon),
3. sama kirjakuju (homograaf). (EKK 2000: L 37)

Sagedusloendite juures oleks olulised just esimesse ja kolmandasse rühma kuuluvad homonüümid, homofoonid on erineva kirjapildiga ja seega segadust kirjalikus korpuses ei tekita. Kasutatud keelekorpustes pole sõna homonüümid ja polüseemsed sõnad eristatud, seega selles töös on samakujulised lemmad või siis sõnavormid esitatud ühe üksusena.

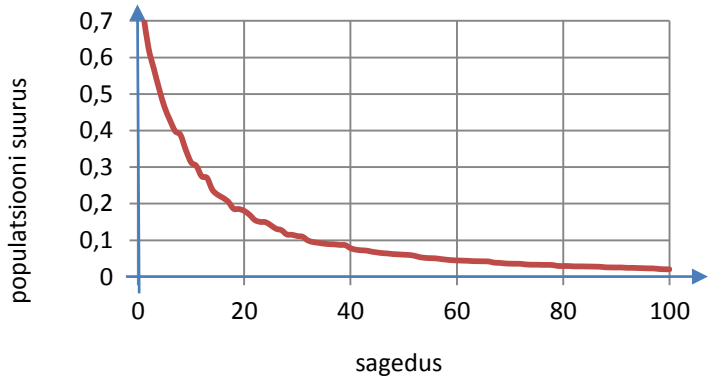
Järgmises peatükis lähemalt sellest, mis sõnade sagedusjaotuse taga peitub. Esmalt antakse ülevaade Zipf’i seadusest, kui ühest sõnade jaotust iseloomustavast suurusest ning seejärel on vaatluse all sõnavara jagunemine tarvitusulatuse ja stiili järgi. Vastavat sõnavara jagunemise liigendust kasutatakse töö võrdluse osas (ptk 7).

3. Sõnavara jagunemine

3.1. Zipfi seadus

Peaaegu kõik sõnade sageduse loendamise ülesanded on seotud Zipfi seadusega. Zipfi seadus (nimetatud Harvardi keeleteaduste professori George Kingsley Zipfi järgi) väljendab kõiki loomulikke keeli puudutavat universaalset omadust, mille kohaselt koosneb tekst peamiselt väikesest hulgast sageli korduvatest sõnadest ja suurest hulgast väikese tekstisisese esinemissagedusega sõnadest. Sõnede hierarhiline organisatsioon on seotud „minimaalse jõupingutuse“ printsiibiga, see tähendab, valides korduvalt alternatiivide seast, otsustab inimene sageli mingi väikese hulga kasuks, seevastu suuremat hulka alternatiividest kasutab ta väga harva. (Kaasik jt 1977: 154-155)

Zipfi seadus mudeldab erinevate objektide esinemissagedust mingis kogumis, selle kohaselt sagedusloendis suvalisel kohal i asuv sõna omab $1/i^0$ kordset sagedust kõige sagedasema sõna sagedusest. See tähendab, et kõige sagedasema sõna esinemissagedus on kaks korda nii kõrge kui sageduselt teisel sõnal ja samasugune jaotus läheb sageduselt järgnevatele sõnadele üle, nii on näiteks kõige sagedasem sõna korpuses sada korda sagedasem kui sajandal kohal asuv sõna. Väike arv kõige sagedamini esinenud sõnu katab suure osa teksti sõnavarast. Sagedase ja madala sagedusega sõna korpuses esinemise arvude vahel on suured erinevused. (Kilgarriff 1997: 137)



Joonis 1. ***Tüüpiline Zipfi-seaduse jaotumus***

Y-telg märgib sõna esinemise sagedust korpuses, X-telg märgib sama esinemissagedusega sõnade arvu (populatsiooni suurust).

Esitatud Zipfi kurvi joonise kohaselt umbes pooled korpuse sõnad esinevad seal vaid ühe korra, keelestatistikas nimetatakse neid sõnu *hapax legomena (*kreeka keelest *miski, mida on öeldud üks kord)*. Ühekordselt korpuses esinevate sõnade suur sagedus tähendab seda, et harv on juhus, mil korpusega töötades märgatakse seal ühte konkreetset korpuses ühekordselt esinevat sõna, kuid on väga tavaline, et märgatakse korpuses mõnda seal vaid üks kord esinevat sõna (wired). Mida väiksem on sagedus, seda rohkem sõnu selle sagedusega esineb ja vastupidi, mida suurem on sagedus, seda vähem sõnu selle sagedusega tekstis leidub. (Kaasik jt 1977: 157)

Arvo Krikmann (Krikman 2004: 118-120) (Krikman 1997: 194) on seda seadust nimetanud „Zipfi needuseks“, mille kohaselt keele ainestik jaguneb „kontiinumiks või skaalaks, mille ühes otsas paikneb väike hulk ülisuure korduvusega üksusi, teises otsas väga suur hulk minimaalse (sagedamini üheainsa) korduvusega üksusi, ja nende vahel kontiinum keskmise kordavusega üksustest, mida endid on valimis samuti keskmine hulk. Zipfi seadus on lingvistikas üks domineerivamaid jaotuslikke tunnuseid, näiteks keeleõppijatele ja -õpetajatele tähendab see, et tõhusaks keeleõppeks on otstarbekas arvestada sõnasagedusega, väikese sagedusega sõnad haaravad küll suure osa leksikonist, kuid samas vajab keeleõppija eelkõige väikest hulka korpuses sagedamini esinevaid sõnu. (Nation 2001: 16)

Järgnevalt sõnavara jaotusest. Eesti keele käsiraamatu (EKK 2000: L 22- L 30) kohaselt võib keele sõnavara jaotada tarvitusulatuse ja stiili järgi, viimasel juhul nii funktsionaaltunnuste kui ka stiilivarjundite alusel. Sarnast jaotust on kasutatud ka EKSS’i märksõnaloendi märgistamisel. EKSS’is kasutatakse **stiiliregistri ja kasutusala** märgendeid, nendega on sõnaraamatus märgendatud sõnu või nende üksiktähendusi, mis üldiselt ei kuulu üldkeelde (EKSS). Kokku on EKSS’is kasutatud 66 erinevat stiili ja kasutusala märgendit, töö lisades on esitatud nimetatud EKSS’i märgendite tabelid (EKSS). Esmalt sõnade jagunemisest tarvitusulatuse alusel.

3.2. Tarvitusulatus

Tarvitusulatuse poolest jaguneb keele sõnavara kaheks: **üld-** ja **oskussõnavara**ks. Neist esimeseks nimetatakse sõnavara, mida vajab iga keelekasutaja igapäevases suhtluses. Üldsõnavara tuumosaks on **põhisõnavara**, mis koosneb keele kõige sagedamini kasutatavast ja seetõttu kõige vajalikumast osast, just põhisõnavara omandamine on oluline võõrkeele õppimisel. Põhisõnavarasse kuuluvad sõnad pole sõltuvuses suhtlussituatsioonist, keeletarvitajate taustast ega kõneainesest. (EKK 2000: L 22)

Teise suurema osa moodustab oskussõnavara, mis on oskussõnade ehk terminite kogum. Kui üldsõnavara on igale keelekasutajale tuttav, siis oskussõnavarast vajab keeletarvitaja ainult väikest osa, just seda, millega ta erialaliselt kokku puutub. Kuna üld- ja oskussõnavara pole üksteisest jäigalt eraldatavad, siis saadakse oskussõnavarasse uusi mõisteid üldsõnavarast ning vastupidises suunas võivad erialaterminid muutuda üldkeelesõnavarasse kuuluvaks. (EKK 2000: L 22) EKSS'i loendites on esitatud 54 kasutusala märgendit, järgnevas tabelis on väljavõte töö lisades esitatud täispikast tabelist.

Tabel 2. **EKSS'i kasutusala märgendid** *(http://www.eki.ee/dict/ekss/ekss.html).*

Lühend	**Seletus**
AIAND	aiandus
AJ	ajalugu
ANAT	anatoomia
BOT	botaanika
EHIT	ehitusala

3.3. Funktsionaalstiilid

Funktsionaalstiilide alusel jagatakse keele sõnavara järgmistesse kihtidesse: **ilukirjanduskeel**e, **ajakirjanduse** ja **teaduskeel**e sõnavara, lisaks veel ametliku stiili sõnavara ja argikeelne sõnavara (EKK 2000: L 30) Käesolevas töös on vaatluse all ajakirjanduse, ilukirjanduse ja teaduskirjanduse sõnavara, kuna töös kasutatav Tasakaalus korpus jaguneb nimetatud tekstiklasside vahel.

Ajakirjanduse sõnavara ehk publitsistliku sõnavara värving on kuivavõitu informatsioonilisus, näiteks on oma stiilivärvingu poolest ajakirjanduse sõnad järgnevad: *arenguabi, kodanikkond, kodanikualgatus, pantvangistama, sundparteistama, olme, relvistu, sõjard*. (EKK 2000: L 30) Ajakirjandus vahendab erinevate eluvaldkondade ning oskuskeelte sõnavara. Lisaks iseloomustab ajakirjanduskeelt suur tsitaatsõnade, toorlaenude ning piltlike väljendite kasutamine. Ajakirjanduse kaudu tuleb keelde uut sõnavara ning ajakirjandus aitab kaasa uute terminite kodunemisele keeles. Teisalt kordab ta ka pidevalt väga suurt osa keele põhisõnavarast. (Kasik 2003)

Ilukirjanduse sõnavara all mõeldakse värvilisi sõnu, millele on omased just selle stiili tunnused. Sõnade eesmärgiks on lugeja esteetilis-emotsionaalne mõjutamine. Ilukirjandussõnavara alla kuulab ka luulekeelne sõnavara. Sõnavara näited: *ast, pild, kuratlus, piirjoonestuma, uimastuma, mässutsema, hämmelgas*. (EKK 2000: L 30)

Teaduskeele sõnavara väljendab maailma teaduslikku tunnetust, seetõttu on ta valdavalt rangelt täpne ja ratsionaalne. Teaduskeel sisaldab ohtralt oskussõnavara, lisaks veel iseloomulikke sõnu, nt *eeldada, tõestada, käsitleda, seoses, järelikult*. (EKK 2000: L 30)

Sõnavara jaguneb suhtlussituatsioonile ja -eesmärgile vastava keelekasutusviisi ehk stiili alusel. Stiili tunnusteks on iseärasused morfoloogias, sõnamoodustuses, sõnavaras, lausestuses, kujundikasutuses, teksti liigenduses ja teistes keele valdkondades. Stiilivärvinguga keelendid on markeeritud (erilised, ebatavalised), näiteks sõnad raisk ja sitapea on stiilivärvingult vulgaarsed,

ärikas ja *pihtapanema* argikeelsed, *sinitaevas* ja *kuldkiharad* poeetilised. Markeeritud sõnad tekitavad väljaspool oma kasutusala stiilimõra, mida markeerimata (tavapäraste sõnade) sõnade kasutamisel ei teki. Piir markeerimata ja markeeritud sõnade vahel ei ole püsiv, nii võib algselt neutraalsena mõjunud sõna aja jooksul omandada näiteks vulgaarse stiilivärvingu. (EKK: L 29)

Töö korpuse ja EKSS'i võrdlust käsitlevas osas kasutati EKSS'is esitatud 12 stiilimärgendit, Tabelis 3 on väljavõte töö lisades esitatud täispikast tabelist.

Tabel 3. **EKSS'i stiiliregistri märgendid** *(http://www.eki.ee/dict/ekss/ekss.html)*

Lühend	**Seletus**
HLV	halvustav
HRV	harvaesinev
IROON	irooniline
LASTEK	lastekeelne
PILTL	piltlik

Järgnevalt töös kasutatud andmekogudest.

4. Töös kasutatud andmekogud

Sagedus on üks paljudest sõna omadustest, kuid see ei ole sõna implitsiitne omadus, mida saaks mõõta, järgides mõnda kindlat definitsiooni. Sagedus määratakse, lugedes kokku kõik sõna esinemised näidistekstis. Seega on sagedus suhteline väärtus, mis muutub tekstide vahetudes ja selle populatsiooniväärtust (väärtus üldkogumis, sõnasageduste puhul sagedus kogu keeles) pole võimalik kindlaks määrata, sest keeles ei esine tõelisi populatsioone. (Orlov jt 1982, viidatud Popescu, 2009: 1) Sagedussõnastike puhul tuleb meeles pidada, et tegemist on sõna sagedusega just selles kasutatud materjalis - saadud sagedusloend peegeldab sõna sagedust konkreetses tekstis, mitte kogu keeles. Sõnasagedus tekstis sõltub teksti pikkusest, teemast, autorist, stiilist ja teistestki parameetritest. Kogu keele sõnasageduste arvutamiseks oleks tarvis moodustada koondkogum absoluutselt kõigist keeles olevatest tekstidest (nii suulistest kui ka kirjalikest). Kuna see aga ei ole võimalik, siis kasutatakse sageduste arvutamiseks teksti valimeid - keelekorpusi. (Hlavácová 2006: 373) Käesoleva töö 5. peatükis koostatud sagedusloendid põhinevad Tasakaalus korpusel.

Peatükis 6 võrreldakse korpuse põhjal koostatud sagedusloendeid EKSS'i märksõnaloendiga. Sõnavara võrdlemisel kasutatakse lisaks Tasakaalus korpusele ka suuremat korpust – Koondkorpust. Järgnevalt kirjeldatakse esmalt kasutatud tekstikorpusi ja neis sisalduvat materjali ning seejärel antakse ülevaade EKSS'ist.

4.1. Koondkorpus ja Tasakaalus korpus

Sagedussõnastik on täpselt nii hea kui korpus, mille põhjal ta tehtud on. Mida suurem on korpus, seda usaldusväärsemaid andmeid keele kohta saame, siiski pole korpuse maht ainukene tulemusi mõjutav karakteristik. Määrav on korpuse struktuur, see tähendab, et tuleb tähelepanu pöörata väiksematele koostisosadele, millest suur korpus on moodustatud. (Hlavácová 2006: 373) Näiteks, koostades sagedussõnastikku ainult teaduskirjanduse tekstide põhjal, erineksid tulemused märgatavalt teadus-, ilu- ja ajakirjanduse tekste koondava korpuse põhjal koostatud sagedusloendist. Samuti mõjutab tulemusi see, millistest tekstidest on näiteks seesama teaduskirjanduse korpus kokku pandud (nt meditsiinitekstid vs keeleteaduse tekstid). Sagedusloendite koostamiseks on vaja representatiivset korpust, mis oleks koostatud erinevatest keele sõnavara katvatest tekstidest. (Hlavácová 2006: 373) **Representatiivsus** tähendab, et korpuse tekstide valim peaks esindama kogu keelt. (McEnery, Andrew 1997: 178) Sagedusloendi alusmaterjaliks kasutatakse umbes 15 miljoni sõnalist Tasakaalus korpust, mis on mõeldud kirjaliku keelekasutuse kolme tähtsama tekstiklassi - ilukirjanduse, ajakirjanduse ja teaduskeele võrdlemiseks. (TK) Nimetatud kolm osa on küllaltki suured ja teatud määral homogeensed, koos peaksid nad esindama Eesti kirjakeele kesksemat osa. Proportsioonidelt jaguneb Tasakaalus korpus võrdselt kolme põhilise kirjakeele tekstiliigi vahel:

1. ajakirjandus umbes 5 miljonit sõna;
2. ilukirjandus umbes 5 miljonit sõna;
3. teaduskirjandus umbes 5 miljonit sõna.

Küll aga, nagu eelnevalt mainitud, jääks Tasakaalus korpuse maht liiga väikeseks võrdluses sõnaraamatu sõnavaraga, seega on töö EKSS'i sõnavaraga võrdlemise osas (ptk 6) kasutatud „Eesti keele koondkorpust", mis on eesti keele kirjalike tekstide kogu. Töös kasutatav nimetuse „Koondkorpuse" all mõeldakse „Eesti keele koondkorpuse" allkorpustest moodustatud korpust, mis koosneb järgnevatest osistest (KK):

- Eesti ilukirjandus 1990- (5,6 miljonit sõna)
- ajaleht Postimees (lehenumbrid 27.11.1995 - 10.10.2000; 1760 numbrit 88 600 artikliga, kokku 32,9 miljonit sõna)
- ajaleht Eesti Ekspress (lehenumbrid 09.08.1996 - 29.11.2001, kokku 7,5 miljonit sõna)
- ajaleht Eesti Päevaleht (lehenumbrid 18.10.1995 - 31.10.2007; (4065 numbrit 366862 artikliga), kokku 87,9 miljonit sõna)
- ajaleht Maaleht (2001-2004, 4.3 miljonit sõna)
- ajaleht SL Õhtuleht (1997-2007, 45.5 miljonit sõna)
- ajakiri Horisont (1996 - 2003, 260 000 sõna)
- ajakiri Luup (1996 - 2002, 1,9 miljonit sõna)
- ajakiri Kroonika (2001 - 2003, 600 tuhat sõna)
- ajakiri Eesti Arst 2002 - 2004 (ca 0,7 miljonit sõna)
- ajakiri Arvutitehnika ja Andmetöötlus (1999 - 2005. 625 tuhat sõna)
- ajakiri Agraarteadus (2001 - 2006. 298 tuhat sõna)
- Mitmesugused teadusartiklid (ca 1,3 miljonit sõna)
- Doktoritööd (0,5 miljonit sõna).

Kui sagedust kasutatakse tavalisuse mõõdupuuna, siis on oluline, et kasutatavad tekstid oleksid homogeensed. Kui korpus koosneb väga erinevatest tekstiklassidest, siis jääb selgusetuks, mida sagedusloendid tegelikult näitavad (Kaalep, Muischnek 2004: 57), seetõttu on selle töö raames „Eesti keele koondkorpusest“ välja jäetud seal olevad uue meedia, seaduste ja Riigikogu stenogrammide tekstid (Uus meedia (ca 22 miljonit sõna), Eesti ja Euroopa seadused vastavalt ca 1,8 miljonit ja 10 miljonit sõna, Riigikogu stenogrammid aastatest 1995-2001 ca 13 miljonit sõna).

Kuuendas peatükis võrreldakse Koondkorpuse ja Tasakaalus korpuse põhjal koosatud sagedusloendeid „Eesti keele seletava sõnaraamatu“ märksõnaloendiga, mille ülevaatlik kirjeldus on järgnevas alapeatükis.

4.2. Eesti keele seletav sõnaraamat

Eesti Keele Instituudis koostatud „Eesti keele seletav sõnaraamat" (EKSS) on seletussõnaraamatu 2., täiendatud ja parandatud trükk. EKSS on deskriptiivne sõnaraamat, mis registreerib, kuidas kirjeldatud sõnu keeles kasutatud on. (Kasik 2011: 203) Sõnaraamatu koostamise vajaduste katmiseks loetakse optimaalseks piiriks umbes 100 miljoni sõna suurust tekstikorpust, sellise mahuga eesti keele tekstikorpust on kasutatud ka EKSS'i 2. trüki toimetamisel. Varasemalt kasutati sõnaraamatu koostamisel peamiselt ilukirjanduse näiteid sisaldavat sedelikogu, kuid praegu kasutatav elektrooniline tekstikorpus sisaldab põhiliselt ajakirjanduskeelt. Samas valdav osa seletussõnaraamatu esimese trüki – "Eesti kirjakeele seletussõnaraamatu" sõnadest on säilitatud, uued sõnad ja tähendused on olemasolevaga liidetud, äärmiselt vähe märksõnu on eemaldatud. (EKSS)

Eesti keelde tekkis palju uusi sõnu 1990. aastatel, mil keelt mõjutasid nii inglise kui ka soome keel. Uute sõnade tekkele aitasid kaasa ka muutunud majandussüsteem ja seadusandlus, arvutimaailm ning mitmesugused täiesti uued nähtused. (Voll 2009: 8) Sõnaraamatu viimasesse trükki on juurde lisatud ligi 4000 uut sõna, sealhulgas infotehnoloogia, majanduse, meditsiini, juura, tehnika, botaanika ja kokanduse valdkonda kuuluvaid sõnu. Samuti on lisatud ka uusi üldkeele sõnu, mis kirjeldavad uuema aja eluolu, näiteks *kaalujälgija*, *spaa*, *m-maksed*. EKSS on eesti kirjakeele kõige suurem varamu, sisaldades ligi 150 000 märksõna. (EKSS)

Töös võrreldakse valikuliselt korpuse sõnavara EKSS'i märksõnaloendiga, kasutatakse märksõnade **põhiloendit** ning arvesse on võetud ka **uudissõnade** loendis olevaid sõnu. Loendite täpsemad kirjeldused antakse kuuendas peatükis, milles kirjeldatakse ka EKSS'i märksõnaloendite kasutamist ja korpuseloenditega ühele kujule viimist.

Enne EKSS'i märksõnaloendi ja korpuse sagedussõnastiku võrdluse osa juurde asumist on vajalik Tasakaalus korpuse põhjal sagedussõnastiku koostamine.

5. Tasakaalus korpusel põhinev sagedussõnastik

Peatükki on koondatud sagedussõnastiku koostamise alla käivad ülesanded, alustades materjali ettevalmistamisest kuni tehtud sagedusloendite kirjelduseni. Peatükk jaguneb kolmeks alapeatükiks ja kokkuvõtteks.

Esimeses alapeatükis töödeldakse sagedusloendite aluseks olev korpus sobilikule kujule; sagedusloendi koostamiseks oli tarvis korpust, milles igale sõnavormile vastaks kindel lemma. Kuna aga kasutatavas Tasakaalus korpuses leidus sõnu, millele pakuti mitu lemma analüüsi, siis oli vajalik korpuse materjali töötlus.

Teises alapeatükis kasutatakse eelnenud peatükis töödeldud materjali ja antakse statistiline ülevaade korpuses leiduvast sõnavarast. Selgitatakse sagedusloendi koostamise põhimõtted ja kirjeldatakse sagedussõnastiku tegemisel arvesse võetud materjali ja ülesandeid, millega kokku puututi ning esitatakse koostatud sagedusloendite nimekiri.

Kolmandas alapeatükis võrreldakse koostatud sagedusloendeid „Eesti kirjakeele sagedussõnastikus“ esitatud loendiga, mis koosneb sajast sagedasemast sõnastikust välja jäänud sõnast (Kaalep, Muischnek 2002: 200-201).

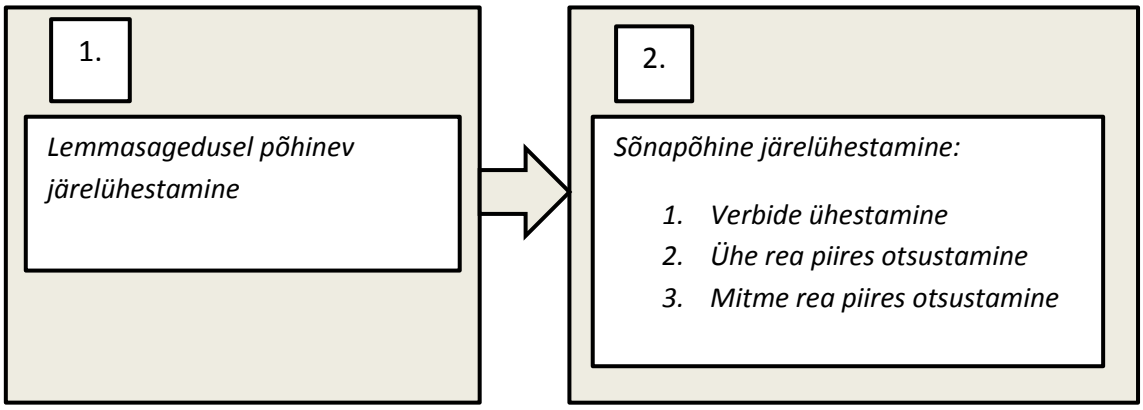
5.1. Korpuse materjali töötlus – järelühestamine

Magistritöö üheks eesmärgiks on eesti keele lemmade ja sõnavormide sagedusloendite koostamine. Sõnavormide sagedusloendi saamiseks liidetakse kokku kõik sõnavormi esinemiskorrad tekstis, kuid lemmade sagedusloendi saamiseks on esmalt tarvis sõnavormid lemmatiseerida. Lemmatiseerimisel taandatakse sõnavormid ühele põhivormile ehk algvormile (Karlsson 2002: 216). Lemmade leidmiseks kasutatakse morfoloogilist analüsaatorit, mis määrab sõnavormi põhjal selle struktuuri, sõnaliigi ja käände või pöörde. (Kaalep 1997: 20) Sõnaliigid, mida analüsaator eristab on esitatud leheküljel http://www.filosoft.ee/html_morf_et/morfoutinfo.html#1. Morfoloogiline analüsaator ei lahenda aga sõnade mitmesuse probleemi, nt analüüsides sõna *mees* pakub analüsaator analüüsideks kahte lemmat *mees* (*mees+0 //_S_ sg n, //*) ja *mesi* (*mesi+s //_S_ sg in, //*). Selleks, et teada, kumba lemmadest on konkreetses kontekstis kasutatud, on tarvis teksti ühestada. Morfoloogiline ühestamine seisneb morfoloogiliselt analüüsitud teksti igale sõnale antud kõigi võimalike morfoloogiliste analüüside hulgast konkreetsesse konteksti sobilikuma analüüsi valimises. (Kaalep 1999: 26)

On võimalik, et ühestamise käigus jääb endiselt järele mitmese analüüsiga ridu, sest ühestamisprogrammil ei ole piisavalt infot otsustamaks, milline morfoloogilise analüsaatori poolt mitmesele sõnavormile pakutud analüüside hulgast on selles kontekstis õige. Näiteks kasutatud korpuses oli ühestamata jäänud partitsiip *kohanud (koha+nud //_V_ nud, // koha=nud+0 //_A_ // koha=nud+0 //_A_ sg n, // koha=nud+d //_A_ pl n, // kohta+nud //_V_ nud, //)*. Kuid kuna lemmade sagedusloendi saamiseks oli oluline, et korpuses vastaks igale sõnavormile üks lemma, siis oli vaja juba ühestatud teksti veelkord ühestada. Kirjeldatud protsessi nimetatakse järelühestamiseks.

Antud töös kasutatavat morfoloogilist järelühestajat rakendati morfoloogilise analüsaatori ESTMORF (Kaalep 1999) väljundile, mida oli juba morfoloogiliselt ühestatud programmiga t3mesta. ESTMORF on OÜ Filosoft poolt Tartu Ülikoolis „Väikese vormisõnastiku“ (Viks 1992) põhjal välja töötatud morfoloogiline analüsaator, mille koosseisus on ka tundmatute sõnade mõistataja. T3mesta on Markovi peitmudelil põhinev trigrammidega statistiline ühestaja. (Kaalep, Vaino 1998) Järelühestamisel

rakendati mitmeseks jäänud analüüsiga ridadele ühestamisreegleid ning seeläbi jäeti pakutud lemmade hulgast alles parim vaste kasutatavaks analüüsiks. Tehtud järelühestamine jagunes kahte suuremasse etappi (vt joonis 2).



Joonis 2. **Järelühestamise ülesehitus**

Esmalt kasutati lemmasagedusel põhinevat järelühestamist, pärast selle läbimist rakendati saadud väljundil järelühestamise teist osa – sõnapõhist järelühestamist – mis lähtuvalt ühestamise objektist või valitud meetodist jagunes kolmeks etapiks: verbide ühestamine, ühe rea piires otsustamine ja mitme rea piires otsustamine. Järgnevalt lähemalt järelühestamisest, alustades lemmasagedusel põhineva järelühestamise kirjeldusega.

5.1.1. Lemmasagedusel põhinev järelühestamine

Järelühestamise esimeses osas kasutati lemmasagedusel põhinevat järelühestamist, rakendatud meetodist on lähemalt kirjutatud artiklis „A trival method for choosing the right lemma“. (Kaalep jt 2012) Artiklis välja toodud lemmade mitmesuse tüübid jagunevad nelja suuremasse gruppi: (Kaalep jt 2012: 83-84)

1. erinevate lemmade homonüümsed sõnavormid, nt sõnavormi *lõi* lemmaks võis olla nii *looma* kui ka *lööma*,
2. tundmatutele sõnadele oletatud lemmad, selle all mõeldakse analüsaatori leksikonist puuduvaid sõnu, mille lemma ei tule mitte morfoloogilise analüsaatori leksikonist, vaid mille võimalikud lemmad ja grammatilised kategooriad morfoloogiline analüsaator oletab sõnavormi väliskuju põhjal. Näiteks võiks (morfoloogilise analüsaatori oletamismooduli arvates) sõnavormi *isolaadid* lemmaks olla nii *isolaat* kui ka *isolaad*.
3. vokaal- või konsonantlõpulised lemma mitmesused, nt sõnavormi *Liisile* algvormiks võis olla nii *Liis* kui ka *Liisi*,
4. nimetavas käändes paralleelvormidega sõnad, see tähendab, et õigeks analüüsiks sobivad tegelikult kõik analüüsimisel pakutud lemmad, siia alla kuuluvad näiteks sõnavormi *mandri* paarid *manner* ja *mander* ning sõnavormi *talveks* võimalikud lemmad *tali* ja *talv*. Mõlemad nimetatud variantidest on korrektsed analüüsid, kuid teksti autor võis eelistada just ühte neist.

Meetod seisnes järelühestamisel morfoloogiliselt analüüsitud tekstis leiduvatele lemmasagedustele toetumisel. Võttes algoritmi lihtsustatult kokku, koostati mitmese analüüsiga jäänud sõnavormile õige lemma valimiseks sagedusloendid tema kõigist tekstis esinenud lemmavormidest. Järelühestamise tulemusel valiti sõnavormile sobilikuks analüüsiks lemma, mille esinemissagedus oli kõrgeim. (Kaalep jt 2012: 85) Näiteks sõnavorm *viigi* võib vastata ainsuse omastavas käändes kahele lemmale – *viik* ja *viig*. Samas, uurides tekstisisest materjali, selgub, et näiteks sporditekstides ei esinenud lemmat *viig*, millel tähenduseks EKSS'i järgi *'17.–19. saj. poliitilise rühmitise liige Inglismaal'*, küll aga esines lemma *viik* ühiseid vorme, näiteks *viiki*. Kasutades

sõnasagedusel põhinevat järelühestajat oli võimalik automaatselt valida sellest mitmesuse paarist sõnavormile sobilik lemma.

Saadud järelühestamise tulemusi hinnati 110000-sõnalise korpuse peal, milles esines 1670 mitmeseks jäänud rida, programm muutis 1190 rida, neist 1120 korrektselt ning 80 ebakorrektselt. Programmi täpsuseks jäi 0,94 ning saagiks 0,67. (Kaalep jt 2012: 86)

5.1.2. Sõnapõhine järelühestamine

Pärast lemmasagedusel põhineva järelühestamise rakendamist jätkus ühestamine üksikumate probleemsete kohtadega. Järelühestamise teises osas koostati *Shelli* skript, milles asendati ühel real asuv mitmeseks jäänud lemmade loend ühe vastavale sõnavormile sobiliku lemmaga, sisendiks kasutati lemma sagedusel põhineva ühestamise väljundit. Seejuures jagunes skriptiga ühestamine kolme osasse (vt ka joonis 2):

1. verbide ühestamine,
2. ühe rea piires otsustamine,
3. mitme rea piires otsustamine.

Esimese osa järelühestamisest moodustas mitmeseks jäänud verbivormide ühestamine. Mitmeseks jäänud verbivormide analüüsiridadest koostati sagedusloend, sagedusloendi koostamiseks eemaldati realt kõik peale analüüsimärgendite, nt korpuses olevast reast

kohanud koha+nud //_V_ nud, // koha=nud+0 //_A_ // koha=nud+0 //_A_ sg n, // koha=nud+d //_A_ pl n, // kohta+nud //_V_ nud, //

jäi järele _V_ nud, _A_ _A_ _A_ _V_ nud. Kõik kirjeldatud kujule viidud analüüsiread sorteeriti, seeläbi moodustusid erinevad grupid, nt sõnaga *kohanud* kuulus samasse gruppi sõnavorm *kaevelnud* (kaevelnud kaeble+nud //_V_ nud, // kaevel=nud+0 //_A_ // kaevel=nud+0 //_A_ sg n, // kaevel=nud+d //_A_ pl n, // kaevle+nud //_V_ nud, //). Sobilikud analüüsid leiti saadud gruppides pakutud variante eraldi analüüsides ning seejärel lisati *Shelli* skripti vastav rida. Näiteks sõnavormile *kohanud* sai analüüsiks

teine verbi analüüs ehk siis lemma *kohtama* ja sõnavormi *kaevelnud* analüüsiks jäi verbi esimene analüüs ehk lemma *kaeblema*. Konkreetsele sõnavormile valiti kõige tõenäolisem analüüs, konteksti ei arvestatud, st tegemist on ühestamisega leksikaalse tõenäosuse põhjal.

Analüüsiridade sagedusgruppide koostamine lihtsustas ühestamisreeglite kirjutamist, eelkõige tänu sellele, et andis ühestamist vajavatest ridadest koondülevaate. Teisalt aga võimaldas sama analüüsi saanud ridu ühestamiseks koondada. Näiteks eelnevalt kirjeldatud ühestamist vajavasse rühma kuuluvate sõnavormide hulgas oli valdavalt sobilikuks analüüsiks esimene verbi analüüs. Seega lähtuvalt sellest kirjutati *Shelli* skripti esmalt otseselt sõnavormist lähtuvad analüüsid nendele sõnadele, mis said vähemlevinud analüüsi (nt sõna *kohanud* sai teise verbi analüüsi *kohta+nud //_V_ nud,*) ja seejärel üldine ühestamisrida, mis andis ülejäänud _V_ nud, _A_ _A_ _A_ _V_ nud analüüsiga sõnavormidele lemmaks sagedamini esinenud esimese verbi lemma.

Teise osa järelühestamisest moodustas ühe rea piires otsustamine, selleks koostati esmalt sagedasemate mitmesuse paaride loendid ja uuriti nende kasutamist korpuses. Näiteks esines lihtsõnade hulgas 11240 korda mitmeseks jäänud lemmade paar *luba* ja *luga*, millest esimene on EKSS'is defineeritud järgnevalt: '*kelleltki saadud suuline v. kirjalik nõusolek millekski; õigus, voli millekski'* ning teine (*luga*) '*niisketel aladel kasvav kitsaste või redutseerunud lehtedega rohttaim'*. Mitmesust põhjustas ühine genitiivivorm *loa* (esines korpuses 6509 korda). Korpuse materjalist vastavate mitmesuse paaride ülevaatamisel ei tuvastatud õige lemmana sõna *luga*. Lemmat *luga* leidus paaril korral teistes vormides, kuid neil juhtudel vormihomonüümiat ei esinenud ning sellest lähtuvalt järelühestamist tarvis polnud, nt '*…muust ümbrusest madalam lugasid täis kasvanud lohk…*', '*Tõmbiõieline luga on teine üliharuldastest taimedest…*'. Seega korpusematerjalile toetudes osutati *luba-luga* paari puhul lemma *luba* kasuks. Tõenäosus, et korpuses esinev sõnavorm *loa* vastab lemmale *luba* oli suurem, samas muidugi jääb vearisk, et kusagil korpuses on sõnavormi *loa* lemmaks siiski *luga*.

Võttes rea piires järelühestamise ülevaatlikult kokku, oli töökäik järgnev: esmalt koostati mitmeseks jäänud analüüsiga sõnade analüüside sagedusloend, seejärel analüüsiti korpuse näiteid ning kirjutati ühestamiseks sobilik asendus *Shelli* skripti.

Muidugi, eelnevalt kirjeldatud sõnapaari *luba-luga* ühestamist võib lugeda mõnevõrra lihtsaks ülesandeks. Üks lemmadest on haruldane sõna ning kuulub pigem keele perifeeriasse, seevastu teine on tavaline üldkeele sõna. Isegi faili vaatamata, teades, et tegemist pole rangelt botaanika valdkonda kuuluva tekstiga, võib suure tõenäosusega mitmesuse paari vaadates otsustada analüüsil lemma *luba* kasuks. Küll aga pole kõigi vormihomonüümia juhtumite lahendamine nii must-valge ülesanne, nt võttes sõnavormi *kaevates*, mille lemmaks võib olla nii *kaebama* kui ka *kaevama*, seisab uurija silmitsi keerulisema ülesandega. Korpuses esinemise sageduse järgi pole võimalik otsustada, kumb lemmapaarist on parasjagu tekstis esineva sõnavormi analüüsiks. Tarvis on igal esinemisjuhul täpsemalt vaadelda konteksti, järelühestamise kolmandal etapil koostatud ühestamisreeglid arvestavad ühestamist vajava sõna eelneva või järgneva sõna analüüsiga. Näiteks sõnavormi *kaevatud* lemmade mitmusepaari korral valitakse fraasis *kohtusse kaevatud* lemmaks *kaebama* ning fraasis *välja kaevatud* valitakse lemmaks *kaevama*. Kirjeldatud mitme sõna konteksti arvestavat lähenemist kasutati ka esimeses ühestamisjärgus tekkinud ebakorrektsete ühestuste muutmiseks, nt otsustas lemmasageduse põhine järelühestaja ekslikult, et fraasis *mullu maist* on sõnavormi *maist* lemmaks *maa*; parandamiseks kirjutati reegel, milles seisis, et eelnevalt nimetatud fraasi korral, st siis kui sõnavormile *maist* eelneb määrsõna analüüsi saanud sõna *mullu*, asendatakse ühestamise eelneval järgul saadud lemma *maa* lemmaga *mai*.

Järgnevalt analüüsitakse järelühestamise tulemusi.

5.1.3. Järelühestamise kokkuvõte

Kokkuvõte järelühestamisest on tehtud Tasakaalus korpuse põhjal, milles sõnesid (ilma kirjavahemärkideta) on kokku 14438223. Enne järelühestamist oli *t3mestaga* ühestatud Tasakaalus korpuses erinevaid analüüse 16610934 (mitmesus 15%), pärast järelühestamist 15000562 (mitmesus 4%). Kokku esines korpuses erinevaid sõnavorme 997934, neist 580805 (58%) esinesid korpuses vaid ühel korral. Erinevaid sõnavorme, mille analüüs jäi mitmeseks, esines pärast järelühestamist tekstis 18996, nendest enamik said pärisnime märgendi (11940). Ülejäänud (7056) mitmeseks jäänud analüüsiga sõnade seas esines mitmuse ja ainsuse vahelist eristamatust, määrsõna ja sidesõna vahelist eristamatust, kirjavigu, võõrkeelseid sõnu ja muud, mis ei mõjutanud hilisemas järgus tehtud lemmade ja sõnavormide sagedusloendite usaldusväärsust. Kokku esines mitmeseks jäänud ridu 545002, nendest üle poole, täpsemalt 319166 moodustasid sõnavormi *on* mitmuse read (nt ole+0 //_V_ b, // ole+0 //_V_ vad, //). 187347 korral esines sõnade *kui*, *nagu*, *just*, *kuigi*, *justkui*, *otsekui* määrsõna ja sidesõna analüüsi vahelist mitmust (nt Kui kui+0 //_D_ // kui+0 //_J_).

Leidus veel 28353 pärisnimede mitmust, järgnevates näidetes on sõnavorm *Hõbemägi* mitmesõnaline liitsõnalise ja lihtsõnalise pärisnime analüüsi vahel ning sõnavormi *Tõnisted* lemmaks võib olla *Tõniste*, mille korral oleks tegemist ainsuse nominatiivi vormiga (Hõbemägi Hõbe_mägi+0 //_H_ sg n, // Hõbemägi+0 //_H_ sg n, //**;** Tõnisted Tõniste+d //_H_ pl n, // Tõnisted+0 //_H_ sg n, //). 5463 korda esines sama analüüsi, kuid erineva pakutud lemmaga variante (nt kirjaviga kellegile kelleg+le //_S_ sg all, // kellegi+le //_S_ sg all, //). 4246 korral esines ainsuse ja mitmuse vormide vahelist mitmesust (nt jõud jõud+0 //_S_ sg n, // jõud+d //_S_ pl n, //).

Lemmade sagedusloendite arvutamiseks töödeldi morfoloogiliselt järelühestatud faili veelkord. Järelühestamisel mitmeseks jäänud sõnade puhul, näiteks eelnevalt nimetatud sõnavormi *on*, millel on nii ainsuse kui ka mitmuse oleviku kolmanda isiku analüüs, on oluline sagedussõnastikus kajastada iga esinemist korpuses ühekordselt, ühesõnaga, vaja oli iga tekstisõna jaoks jätta alles ainult üks analüüs. Kui mingi sõnavormi

ühestamine ja järelühestamine polnud õnnestunud, valiti lemmade sagedusloendite tegemisel mitmeseks jäänud analüüside seast automaatselt esimene analüüs.

Korpuses on esitatud iga tekstis tühikutega eristatud sõne ise real, morfoloogiline analüüs ja ühestamine käsitlevad näiteks verbi liitvorme (*ei teinud*, *on teinud*) ning perifrastilisi verbe (*lõi kokku*, *paistis välja*) eraldi sõnadena ja sellest lähtuvalt on tekstis mitmesõnalise kombinatsioonina mõeldud ühenditel sagedusloendis esitatud kõik osad eraldi. Della Summers kirjutab, et taolised sõnade koosesinemised suurendavad sõna sagedust sõnastikus, näiteks inglise keeles on sagedane sõna *day* (*päev*), mille sagedusest osa moodustab esinemine fraasides, nt *one day* (üks päev), *some day* (ühel päeval). (Summers 1996: 5)

Järgmises peatükis koostatakse järelühestatud korpuse põhjal sagedusloendid.

5.2. Sagedusloendite koostamine

5.2.1. Sagedusloendite alusmaterjal

Pärast morfoloogilist järelühestamist, mille järel kahandati ühele sõnale pakutavate lemmade arvu, koostati sagedusloendid, mis põhinesid statistilise ühestajaga *t3mesta* morfoloogiliselt ühestatud ning seejärel reeglipõhise meetodiga järelühestatud umbes 15-miljoni sõne suuruse Tasakaalus korpuse sõnavaral. Kogu korpuse materjali põhjal sagedusloendeid siiski ei koostatud, välja otsustati jätta järgneva analüüsi saanud read:

- kirjavahemärgid (nt **. //_Z_ //**),
- pärisnimed (nt **Leonardo+0 //_H_ sg n, //**),
- lühendid (nt **nt+0 //_Y_ ?, //**),
- genitiivatribuudid (nt **vene+0 //_G_ //**),
- numbriga kirja pandud arvud (nt **69+0 //_N_ ?, //**),
- ka rooma numbrid (nt **II+0 //_O_ ?, //**).

Nimetatud read on eemaldatud automaatselt, kasutades analüüsimärgendeid. Kui mõni eelnevatest sõnaklassi liikmetest on jäänud sagedusloendisse sisse, siis selle pärast, et oli mingil analüüsi astmel saanud vale analüüsi. Automaatsel numbreid sisaldavate ridade eemaldamisel jäid sagedussõnastikust välja ka osaliselt numbriga kirjutatud sõnad näiteks *mp3, 3D, 6-aastane*.

Sagedusloendid sisaldavad ka võõrkeelseid sõnu, mis said morfoloogilise analüsaatori oletamisrežiimis vale analüüsi ja seetõttu ei eristunud omakeelsetest sõnadest. Vajadusel on võimalik neid hiljem sagedusloendit üle vaadates käsitsi eemaldada. Loenditest on aga juba eemaldatud erinevaid sümboleid, nt ß, —,¦, >, ù, ^, δ, %, û, ź, •, ≤, ÿ, ≥, $, }, φ, ε, §, µ, ÷, ¿, <, ū, ú, °, sisaldavaid ridu.

15-miljoni suurune Tasakaalus korpus jaguneb 5-miljoni sõna suurusteks osadeks ajalehetekstide, ilu- ja teaduskirjanduse vahel. Pärast kõiki eelnevalt loetletud eemaldamisi jäi algsest korpusest sagedusloendite koostamise aluseks 12878133 sõneline fail. Tabelist 4 (lk 39) on näha, et materjali kitsendamisel eemaldati ajalehtedest ja teaduskirjandusest rohkem sõnesid kui ilukirjandusest. Seda võib põhjendada sellega, et ilukirjanduse tekstides esines vähem pärisnimesid ja numbreid sisaldavaid ridu.

Eesti keel on rikka morfoloogiaga, seega sagedusloendid koostati nii lemmade kui ka sõnavormide põhjal. Enne sageduste arvutamist oli vajalik aga lemmade ja sõnavormide teisendamine morfoloogilise analüsaatori väljundi kujult sagedusloendi tegemise kujule. Lemmade puhul oli vajalik algvormi kättesaamiseks puhastada analüüs. Näiteks sõnavormile *majandusmehed* vastas lemmaanalüüs *majandus_mees+d*, koostatava lemmade sagedusloendi tarbeks eemaldati liitsõnapiiri tähistav alakriips (_) ning märgendiga *pluss* (+) eraldatud mitmuse tunnus *-d*, seega järele jäi lemma *majandusmees*. Verbide, näiteks verbivormi *kuulanud* puhul, mis sai morfoloogiliseks analüüsiks *kuula+nud*, asendati märgendile *pluss* (+) järgnev osa *ma*-tegevusnime tunnusega *-ma* ning lemmaks sai *kuulama*.

Sõnasagedusloendite koostamisel oli oluline, et sama sõnavormi erikujulised korpuses esinemisjuhud oleks koondatud ühe sageduse alla. Näiteks selleks, et sama sõnavormina – *kassile* – läheksid kirja nii sõnakujud *kassile*, *Kassile*, kui ka *KASSILE*, teisendati kogu korpusematerjal väiketäheliseks.

Kombineeritud loendi ja ainult lemmade loendi puhul tuleb tähelepanu juhtida ka märgendamisest tingitud lemmade iseärasusele, nii on näiteks sõnastikus kõik omadussõna võrded esitatud eraldi lemmadena. See tähendab, et näiteks lemma *suur* alla ei ole koondatud tema kesk- ja ülivõrre (*suurem, suurim*), sõnastikus esinevad nad kõik omaette lemmadena.

Ettevalmistatud korpusematerjali põhjal koostati erinevad kasutajatele vajalikuks osutuda võivad sagedusloendid (5.2.3.). Sõnasageduste arvutamiseks kasutati tekstisõne absoluutsagedust, mis on sõne tekstisiseste esinemiste summa. Kui korpus sisaldab erinevaid tekstiklasse ja on piisava suurusega, siis võib neid absoluutsageduste tulemusi küll usaldada, kuid siiski jääb sisse ebaühtlaselt jaotunud sõnade probleem. See tähendab, et leidub tekste, milles mõningad sõnad on ebaharilikult kuhjunud ning seeläbi võib näiteks vaid ühes tekstis arvukalt esinenud sõna saada korpuse sagedusloendis palju kõrgema väärtuse, kui ta tegeliku keelekasutuse põhjal oleks saanud. (Hlavácová, 2006, lk 374)

Lisaks sellele tuleb silmas pidada, et loendites on esitatud sõnade, mitte sõnatähenduste sagedused. Ühel sõnakujul võib olla mitu erinevat tähendust, kuid kuna kasutatud korpuses pole võimalikud sõnatähendused eristatud, siis ei arvestata sõnatähendustega ka sagedusloendite koostamisel, st näiteks sõna *tee* kõik tähendused loetakse üheks sõnaks. Järgnevalt koostatud sagedusloenditest.

5.2.2. Kogu Tasakaalus korpusel põhinev statistika

Sagedusloendite koostamise aluseks oli 12878133-sõneline fail. Selles leidus 314844 erinevat lemmat (L) ning 739682 erinevat sõnavormi (V). Allolevas tabelis on esitatud korpuses sisalduva materjali omavahelised suhtarvud nii kogu korpuse kui ka allosade kaupa. N/(L|V) näitab lemma või sõnavormi keskmist korduvust valimis, (L|V)/N näitab sõnastiku mahu suhet keskmisesse pikkusesse (sõnastiku suhtelist „rikkust“). Keele analüütilisuse astet näitab L/V. (Kaasik jt 1977: 5)

Tabel 4. **Tasakaalus korpuse suhtarvud**

Korpus	N	L	V	N/L	N/V	L/N	V/N	L/V
tasak	12878133	314844	739682	40,90	17,41	0,02	0,06	0,43
aja	4067064	140577	339119	28,93	11,99	0,03	0,08	0,42
ilu	4723832	135658	315134	34,82	14,99	0,03	0,07	0,43
tea	4087237	145175	341539	28,15	11,97	0,04	0,08	0,43

Selgub, et keskmiselt esineb üks lemma kogu korpuses 40,9 ja sõnavorm 17,41 korda. Allkorpuste sagedusi uurides selgus, et aja- ja teaduskirjanduse lemmade (~28) ja sõnavormide (~12) keskmised esinemise sagedused kattusid, kuid ilukirjanduses esinenud lemmade (34,8) ja sõnavormide (14,9) keskmine korduvus korpuses oli kõrgem. Korpuse mahu vähenemisega väheneb ka seal esinenud erinevate lemmad ja sõnavormide arv, näiteks kogumikus „Keelestatistika 2“ (Kaasik jt 1977: 5) ilukirjandusproosa lekseemide sagedussõnastiku arvutustes oli näha, et umbes 100000 sõna suuruses korpuses esines üks lemma keskmiselt 6,8 korda.

Analüütilisuse indeks (L/V), mis näitab lemmade arvu suhet sõnavormidesse, oli sarnane nii kogu korpusel kui ka tema allosadel (0,42-0,46). Ilukirjandusproosa

lekseemide sagedussõnastiku analüütilisuse indeksiks saadi 0,48. (Kaasik, Soontak, Viilup, & Ääremaa, 1977, lk 6)

Sagedussõnastiku iseloomustamisel võetakse arvesse ka sõnastikus üks kord esinevate lemmade (L^1) ja sõnavormide (V^1) arv. Sõnastikus üks kord esinevaid lemmasid leidus Tasakaalus korpuses 173448 ning korpuse igas allosades leidus neid üle 7000.

Tabel 5. **Korpuses üks kord esinenud lemmad ja sõnavormid**

Korpus	N	L	V	L^1	V^1	L^1/L	V^1/V	L^1/N	V^1/N
tasak	12878133	314844	739682	173448	420112	0,55	0,57	0,01	0,03
aja	4067064	140577	339119	75554	197133	0,54	0,58	0,02	0,05
ilu	4723832	135658	315134	75906	184927	0,56	0,59	0,02	0,04
tea	4087237	145175	341539	73931	187357	0,51	0,55	0,02	0,05

Üks kord esinevaid sõnavorme oli kogu Tasakaalus korpuses 420112 ning allosades üle 184000. Seega, üks kord korpuses esinenud lemmade ja sõnavormide osatähtsus sõnastikus jäi 51-59% vahele – kõige rohkem esines neid ilukirjanduses ja kõige vähem teaduskirjanduses. Kogu tekstist katsid üks kord esinenud lemmad 1-2% ning üks kord esinenud sõnavormid 3-5%. „Eesti kirjakeele sagedussõnastikust“ (Kaalep, Muischnek 2002) moodustasid üks kord korpuses esinenud lemmad umbes 53% ja nad katsid 3,5% kogu tekstist. Sarnaselt eelnevatele sagedussõnastikele oli ka ilukirjandusproosa lekseemide sagedussõnastikus üks kord esinevate lemmade osatähtsus üle viiekümne protsendi (59%), kuid erinevalt teistest koostatud sagedusloenditest katsid nad koguni 9% kogu tekstist. (Kaasik jt 1977: 5)

Sõnavara kumulatiivsest teksti katmise võimest annab ülevaate Tabel 6, mis on koostatud kogu Tasakaalus korpuse sageduse kahanemise alusel järjestatud lemmade ja sõnavormide sagedusloendite põhjal. Tabelist on näha, mitu protsenti kogu korpuse sõnavarast moodustas kindel sagedusvahemik.

Tabel 6. **Lemmade ja sõnavormide kumulatiivne osakaal teksti katmisel**

Sagedusvahemik	Lemmade katvuse %	Sõnavormide katvuse %
1…10	18,8	12,9
1…20	23,6	17,0
1…50	31,3	22,6
1…100	37,9	27,9
1…250	47,9	35,9
1…500	56,6	42,3
1…1000	65,6	49,0
1…1500	70,7	53,0
1…2000	74,1	56,0
1…3000	78,6	60,2
1…5000	83,5	65,6
1…10000	88,7	72,7
1…12206	90,0	74,6
1…80446	97,5	90,0

Selgub, et esimesed kümme sõna katavad 18,8% lemmadest ja 12,9% sõnavormidest ning 90% Tasakaalus korpuse sõnavara katmiseks läheb lemmade loetelu korral tarvis üle kümne tuhande ning sõnavormide loendi tarbeks on vaja üle 80000 sagedasema sõna kaasamist. Seejuures 90% täitumisel esineb iga lemma korpuses vähemalt 63 ja iga sõnavorm vähemalt 10 korda.

Tasakaalus korpuse esimese kümne lemma tekstikatmise protsent on sarnane ka „Eesti kirjakeele sagedussõnastikus“ (Kaalep, Muischnek 2002) ja ilukirjandusproosa sagedussõnastikus, kus kümme sagedasemat lemmat katsid vastavalt 19,3 ja 18,6% kogu tekstist. Saavutamaks 90-protsendilist lemmade tekstikatvust läks tarvis alla 10000 sagedasema „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) lemma ja ilukirjandusproosa sagedussõnastikus (Kaasik jt 1977) läks 90% teksti katmiseks tarvis üle 5000 sagedasema lemma.

Tabelis 7 esitatakse sagedussõnastiku leksikaalne spekter, mis kujutab sõnasageduste jaotumist koostatud sõnastikus (Kaasik jt 1977: 6). Tabelist on näha, et alla kümne korra korpuses esinenud sagedusega sõnade koormus tekstis on nii lemmade kui ka sõnavormide põhjal pea 90 protsenti, seejuures nagu juba Tabelist 6 selgus,

moodustasid korpuses üks kord esinevad lemmad ja sõnavormid korpuse üldmahust üle 55 protsendi.

Tabel 7. **Korpuse leksikaalsed spektrid**

Sagedus korpuses	Lemma	Sõnavorm
1	55,1	56,8
2	12,8	13,5
3	6,1	6,4
4-10	12,8	12,8
>10	13,2	10,4

Kuna sagedusloendid koostati automaatselt, siis jäi sisse korpuses esinevaid kirja- ja trükivigu ning morfoloogilisel ühestamisel tekkinud probleeme; sagedusloendite kompaktsuse säilitamiseks ja ühtlasi ka sõnavarast parema ülevaate saamiseks oli mõistlik piirata esitatud sõnade arvu. Rohkem kui 10 korda esinenud sõnad katavad 90% sõnavormidest ja umbes 95% kogu lemmadest ning moodustavad korpuse leksikaalsest spektrist lemmade puhul 13,2 ja sõnavormide puhul 10,4%.

Käesolevas peatükis esitatud meetodiga koostatud sagedusloendites esitatakse kümme või enam korda korpuses esinenud lemmad ja sõnavormid. Järgnevalt lähemalt koostatud loenditest.

5.2.3. Kümme ja enam korda korpuses esinenud lemmade ja sõnavormide sagedusloendid

Töös koostati nelja erinevat tüüpi sagedusloendid, koostatud loendid on esitatud töö lisades ning osaliselt on nad kättesaadavad ka internetiaadressil *http:://www.cl.ut.ee/ressursid/sagedused1/*.

Esiteks koostati lemmade ja sõnavormide loendid nii kogu Tasakaalus korpuse kui ka selle kolme allosa (aja-, ilu- ja teaduskirjanduse) põhjal. Kaldkirjas on esitatud töö lisades olevate failide nimed. Iga fail sisaldab omakorda nelja töölehte: lemmad sorteeritult sageduse ja alfabeetilise järjestuse alusel ning sõnavormid sorteeritult sageduse ja alfabeetilise järjestuse alusel.

1_Tasak_loend - sisaldab vähemalt 10 korda kogu Tasakaalus korpuses esinenud lemmasid/sõnavorme (töölehed: *tasak_lemma_nr, tasak_lemma_alfa, tasak_sonavorm_nr, tasak_sonavorm_alfa).*

1a_Ajakirjandus - sisaldab vähemalt 10 korda Tasakaalus korpuse ajakirjanduse allosas esinenud lemmasid/sõnavorme (töölehed: *aja_lemma_nr, aja_lemma_alfa, aja_sonavorm_nr, aja_sonavorm_alfa).*

1b_Ilukirjandus - sisaldab vähemalt 10 korda Tasakaalus korpuse ilukirjanduse allosas esinenud lemmasid/sõnavorme (töölehed: *ilu_lemma_nr, ilu_lemma_alfa, ilu_sonavorm_nr, ilu_sonavorm_alfa).*

1c_Teaduskirjandus - sisaldab vähemalt 10 korda Tasakaalus korpuses teaduskirjanduse allosas esinenud lemmasid/sõnavorme (töölehed: *tea_lemma_nr, tea_lemma_alfa, tea_sonavorm_nr, tea_sonavorm_alfa).*

Järgnevalt on Tasakaalus korpuse sõnavormide sagedusloendi näitel esitatud väljavõte sageduse kahanemise alusel sorteeritud loendist (vt lisadest faili *1_Tasak_loend* tööleht *tasak_sonavorm_nr*).

Sõnavorm	**Sagedus**
krooni	9257
olema	9213
kuni	9176
kuigi	9116
hea	9114
mees	9108

Ning väljavõte alfabeetiliselt sorteeritud loendist kogu Tasakaalus korpuse lemmade sagedusloendi näitel on järgnev (vt lisadest faili *1_Tasak_loend* tööleht *tasak_lemma_alfa*):

Lemma	**Sagedus**
kalk	76
kalkulaator	65
kalkulatsioon	39
kalkuleerima	44
kalkuleerimine	17
kalkun	40

Teiseks koostati loend ***2_jagatud***, milles on näha, kuidas Tasakaalus korpuses on sõnasagedused jagatud kolme allkorpuse (aja-, ilu-, ja teaduskirjanduse) vahel. Loend koostati nii lemmade (tööleht *jagatud_lemmad_nr*) kui ka sõnavormide (tööleht *jagatud_sonavormid_nr*) kohta, mõlemas neist esitati esmalt sagedus kogu Tasakaalus korpuses ja selle järel sagedused kolmes alaosas. Väljavõte loendist (vt lisadest faili *2_jagatud* tööleht *jagatud_lemmad_nr*):

Sagedus	**Lemma**	**Aja**	**Ilu**	**Tea**
343	õppematerjal	33	4	306
343	standardsort	0	0	343
343	sentimeeter	191	129	23
343	rõivas	136	181	26
343	padi	36	302	5
343	kirjutaja	98	77	168
343	kahtlustatav	96	0	247
343	jätmine	136	39	168
343	järeltest	0	0	343

Selle loendi põhjal on võimalik uurida nii sõnavara jaotust kolme tekstiklassi vahel, tekstiklassidele ühist sõnavara kui ka sõnavara, mis on iseloomulik just kindlale Tasakaalus korpuse osale. Samuti näitab see loend, kuidas tekstis sama sageduse saanud sõnade esinemise sagedused korpuse allosade vahel on väga erinevad. Selleks, sõnasagedusloend iseloomustaks ka sõna „tavalisust“ mitte ainult sõna sagedust, oleks mõistlik kasutada näiteks peatükis 1.3. kirjeldatud keskmiselt vähendatud sageduse meetodit.

Kolmandaks koostati loend ***3_yhised*** Loend sarnaneb ülesehituse ideelt „Eesti kirjakeele sagedussõnastikuga“, milles esitati sagedusloendites ainult need sõnad, mis esinesid mõlemas koostatud korpuse alusmaterjaliks olnud tekstiklassis (aja- ja ilukirjanduses) kokku vähemalt viis korda (Kaalep, Muiscnek 2002: 10). Loendis *3_yhised* on esitatud sõnad, mis esinesid kõigis Tasakaalus korpuse kolmes tekstiklassis (aja-, ilu- ja teaduskirjanduses) ja kokku vähemalt 10 korda. Koostati nii ühiste lemmade (tööleht: *ühised_lemmad_nr*), kui ka ühiste sõnavormide loendid (tööleht: *ühised_sõnavormid_nr*). Väljavõte loendist (vt lisadest faili *3_yhised* tööleht *ühised_lemmad_nr*):

Sagedus	Lemma	Aja	Ilu	Tea
228	kärbes	54	163	11
228	eriliselt	69	99	60
228	ebaoluline	25	19	184
227	talvine	103	83	41
227	suisa	124	92	11
227	pahane	73	149	5
227	mikroorganism	8	1	218

Viimase loenditüübina koostati lemmade ja sõnavormide kombineeritud loend ***4_kombineeritud***, milles esitati lemma sagedus ja selle järel kõigi tema korpuses esinenud sõnavormide sagedused, ka see loend sorteeriti nii alfabeetiliselt (tööleht: *kombineeritud_alfa*) kui ka sageduse kahanemise alusel (tööleht: *kombineeritud_nr*). Loend annab informatsiooni selle kohta, millistest sõnavormidest lemma sagedus koosneb. Näiteks lemma *pearaputus* esineb Tasakaalus korpuse põhjal tehtud sagedusloendis 13 korda, kokku on tal korpuses 6 erinevat sõnavormi, kõige sagedasem neist ainsuse nimetavas käändes olev sõnavorm *pearaputus*.

pearaputus 13

6 pearaputus
2 pearaputusega
2 pearaputuste
1 pearaputuse
1 pearaputusi
1 pearaputustest

Kombineeritud loendist tuleb eriti selgelt välja, et sagedusloendis pole eristatud homonüüme; järgnev väljavõte korpuses 182 korda esinenud lemma *aas* sõnavormidest.

182 aas

67 aasa
42 aasal
21 aas
12 aasad
12 aasale
6 aasaga
5 aasadel
4 aasade
4 aasu
3 aasalt
…

5.2.4. Kokkuvõte korpuses esinenud lemmadest ja sõnavormidest

Kümme või enam korda korpuses esinenud lemmasid oli kokku 44141 ning sõnavorme 82945. Tabelis 8 on esitatud ka eraldi kõigis allkorpustes leidunud sõnade absoluutsagedused. Jaotuses pole arvestatud sellega, et konkreetne sõna võis puududa mõnest allkorpuse osast – kui koondsagedus ületas 10 sõna piiri, siis kaasati ta sagedusloendisse. Ajakirjanduse allosas leidus 10 või enam korda esinenud lemmasid 19995, ilukirjanduses 18613 ja kõige rohkem erinevaid lemmasid leidus teaduskirjanduses – 21643. Sõnavormide sagedused jagunesid järgnevalt: ajakirjanduses 34034, ilu- ja teaduskirjanduses vastavalt 31548 ja 38304 sõnavormi.

Tabel 8. **Kokkuvõte kümme või enam korda korpuses esinenud lemmadest ja sõnavormidest**

Lemmad/sõnavormid	kogu korpus	ühine kolmele allosale	aja	ilu	tea
10+ lemmat korpuses	44141	26171	19995	18613	21643
10+ sõnavormi korpuses	82945	53538	34034	31548	38304
ainult ühes korpuse allosas esinenud lemmade arv			1168	780	4737
ainult ühes korpuse allosas esinenud sõnavormide arv			1282	1271	5097

Koostati ka sagedusloend, kuhu valiti sõnad, mis esinesid kõigis allkorpuse osades korraga ja kokku vähemalt kümme korda. Selliseid lemmasid oli kokku 26171. Seega kirjeldatud fail oli 17970 sõna võrra väiksem failist, mis sisaldas kõiki korpuses vähemalt 10 korda esinenud sõnu, sõltumata sellest, kas nad esinesid kõigis alakorpustes või mitte. Kõigis kolmes Tasakaalus korpuse osas koos esinenud sõnavorme leidus 53538, seega 29407 sõnavormi ei olnud kõigile allosadele ühised.

Ainult ajakirjanduse tekstides leiduvaid lemmasid oli 1168, ilukirjanduses 780 ning teaduskirjanduses koguni 4737 – teaduskirjanduse sõnavara oli aja- ja ilukirjanduse

omast sedavõrd erinev. Ainult ajakirjanduse tekstides leiduvaid sõnavorme oli 1282, ilukirjanduses 1271 ning teaduskirjanduses 5097.

Millist sagedusloendit siis ikkagi kasutada, kas seda, mis sisaldab kogu Tasakaalus korpuse sagedusi või seda, mis sisaldab kolmele tekstiklassile ühiseid sagedusi? Ühest vastust on keeruline välja pakkuda. Kuna sagedusloendeid pole käsitsi puhastatud, siis loendite piiritlemine vaid kõigile korpuse allosadele ühiste sõnadega täitis mõnevõrra loendite puhastamise rolli. Näiteks sel viisil koostatud loenditest jäid välja mõningad morfoloogilisel märgendamisel ekslikult loendist välja jääva pärisnime analüüsi asemel muu analüüsi saanud sõnad, nt *piiroja*, *kandimaa*. Samuti jäid välja paljud teaduskirjanduse sõnad, (nt *ettevaatusprintsiip, proteiinisisaldus, laktatsioon*, *sõnavorm*), välja jäi ka ajakirjanduse sõnu (nt *poolkaitsja, suusaliit*, *valuutafond*) ja ilukirjanduse sõnu (nt *vahimees, valgusolend*, *kähisema*).

Kuna allkorpuste osi on kolm, siis tekkis küsimus, kuidas jagunevad omavahel sõnad, mis on ühised vaid kahele tekstiklassile. See tähendab, et kui moodustada lemmade põhjal paarid, vastavalt siis ajakirjandus ja ilukirjandus, ajakirjandus ja teaduskirjandus ning ilukirjandus ja teaduskirjandus, kuidas siis ühiste sõnade arv jaguneb. Selgus (Tabel 9) et aja- ja ilukirjanduses koos esinevaid ning samas teaduskirjandusest puuduvaid sõnu oli 5517, ajakirjanduses ja teaduse tekstides koosesinevaid sõnu oli 4552 ning teaduskirjanduses ja ilukirjanduses koosesinevaid sõnu oli vaid 1216. Ajakirjandus on lüli kahe äärmuse, ilukirjanduse ja teaduskirjanduse vahel ning, nagu eelnev tabel (Tabel 8) näitas, andis teaduskirjanduse allosa sagedusloendisse juurde kõige rohkem vaid ühte tekstiklassi kuuluvaid sõnu (sõnu, mis esinesid ainult teaduskirjanduses).

Tabel 9. ***Lemmade vaheline koosesinemine kolmes Tasakaalus korpuse allosas***

Korpuse allosa	**ilu**	**tea**
aja	5517	4552
ilu	-	1216

On arusaadav, et aja-, ilu- ja teaduskirjanduse sõnavara ongi erinev ning ühele tekstiklassile tavalist sõna ei pruugi korraga ülejäänud kahes allkorpuse osas korraga leiduda. Nii näiteks leidub ilukirjanduses lemma *sonima*, mida teistes tekstiklassides ei leidunud ja seetõttu seda tavalisena tunduvat sõna tekstiklasside ühiste sõnade sagedusloendisse ei lisata. Muidugi tuleb tähele panna, et kõik need arvud põhinevad ainult ühel, sagedusloendite aluseks olnud korpusel. Teistsugune tekstivalik ja korpuse maht võivad tulemusi muuta. Näiteks puudub ajakirjanduse allosast sõna *footon*, aga kui korpuse ajakirjanduse osa sisaldaks kvantfüüsika alast artiklit, milles kirjutatakse elektromagnetkiirguse väikseimatest osakestest, siis esineks see sõna ka korpuse ajakirjanduse osas.

Sellest, kuidas mõjutab korpuse suuruse kasvamine sagedusloendit, mis sisaldab vaid kõigile tekstiklassidele ühist sõnavara, annab ülevaate järgmises peatükis tehtav võrdlus Tasakaalus korpuse põhjal koosatud sagedusloendi ja Eesti kirjakeele sagedussõnastikust (Kaalep, Muischnek 2002) välja jäänud sagedasemate sõnade loetelu vahel.

5.3. Eesti kirjakeele sagedussõnastikust välja jäänud sõnad

„Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) koostamisel sagedussõnastikku sõnavara lisamise üheks põhimõtteks oli sõna leidumine kasutatud tekstikorpuse mõlemas osas, seega pidi sõna esinema nii ilukirjanduse kui ka ajakirjanduse tekstides. Kui sõna puudus ühest tekstiklassist või esines mõlema teksti peale kokku vähem kui viis korda, siis teda sagedussõnastikku ei lisatud. (Kaalep, Muischnek 2002: 201) Eraldi sagedussõnastiku loendina esitati sajast sagedasemast korpuses vaid ühes sagedussõnastiku alusmaterjalina kasutatud tekstiklassis leiduvast sõnast koosnev loend. Järgnevalt vaadeldakse, kuidas viisteist korda suurema Tasakaalus korpuse põhjal tehtud sagedusloendis on esindatud nimetatud „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) loendisse kuuluvad sõnad.

Vaatluse alla võeti 10 sõna kummastki tekstiklassist. Tabelites 10 ja 11 on esitatud sõna ja selle esinemissagedus „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) koostamise aluseks olnud korpuses. Seejärel on lisatud sagedus Tasakaalus korpuse põhjal koostatud materjalis, jaotununa tekstiklasside vahel. Tasakaalus korpuse põhjal koostatud tekstiklassi sisestes jaotustes on märgistatud kõige sagedamini esinenud tekstiklassi lahter.

Tabel 10. **Võrdlus „Eesti kirjakeele sagedussõnastiku“ ajakirjanduse andmetega**

Sagedus 1mln	**Lemma**	**Ajakirjandus 5mln**	**Ilukirjandus 5mln**	**Teaduskirjandus 5mln**
110	liider	908	61	167
76	investeering	596	25	528
60	investor	342	4	55
57	börs	464	10	25
56	tarbija	423	19	275
50	konkurent	476	56	122
46	sätestama	226	5	1010
45	peatreener	645	2	0
43	finaal	597	24	4
42	nõunik	349	58	35

Tabel 11. **Võrdlus „Eesti kirjakeele sagedussõnastiku“ ilukirjanduse andmetega**

Sagedus 1mln	Lemma	Ajakirjandus 5mln	Ilukirjandus 5mln	Teaduskirjandus 5mln
67	pomisema	14	575	0
54	kummarduma	34	366	5
50	puuraidur	2	19	2
42	silitama	24	342	2
38	võpatama	11	383	0
34	seisatama	6	193	0
33	kuulatama	5	219	3
29	kohendama	67	218	9
29	kaamel	14	159	3
28	palat	35	70	15

Tabelist 10 selgub, et Tasakaalus korpuse sagedusloendi ilukirjanduse osast ei puudunud need sõnad, mis esinesid „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) andmete põhjal ainult ajakirjanduse tekstides, nt sõnad *liider* ja *finaal*. Tabelist 11 selgub, et Tasakaalus korpuse sagedusloendi ajakirjanduse osast ei puudunud need sõnad, mis esinesid „Eesti kirjakeele sagedussõnastiku“ (Kaalep, Muischnek 2002) andmete põhjal ainult ilukirjanduse tekstides, nt sõnad *pomisema* ja *kaamel*. Seega, kui käesolevas töös oleks sagedussõnastiku koostamisel arvestatud „Eesti kirjakeele sagedussõnastikuga“ (Kaalep, Muischnek 2002) sama põhimõttega, et sõna sõnastikku lisamiseks peab ta esinema mõlemas tekstiklassis koos vähemalt 5 korda, siis erinevalt „Eesti kirjakeele sagedussõnastikust“ (Kaalep, Muischnek 2002), lisataks need sõnad Tasakaalus korpuse põhjal koostatud sagedussõnastikku.

Samas on huvitav tähelepanek, et kuigi korpuse mahu suurenemisega on esinenud sõnade sagedus suurenenud ja mõlemasse tekstiklassi on juurde tulnud sõnu, mida seal varem ei esinenud, on sõnasageduste taga peituv informatsioon jäänud samaks. See tähendab, et miljoni sõnalise korpuse põhjal „ilukirjanduslikeks“ loetud sõnade (nende, mida ajakirjanduse tekstides ei leidunud) esinemissagedus 15 korda suurema korpuse ilukirjanduse osas on ikka kõrgem kui sama korpuse ajakirjanduse osas. Näiteks sõna *pomisema*, esineb miljonisõnalise korpuse ilukirjanduse osas 67 korda ja ajakirjanduses seda sõna seal ei leidunud, Tasakaalus korpuse materjalides leidus *pomisema* ajakirjanduse tekstides 14 korda ning ilukirjanduse tekstides 575 korda, seega sõna *pomisema* võib endiselt pidada pigem „ilukirjanduslikumaks“ sõnaks.

Võrreldes „Eesti kirjakeele sagedussõnastikuga“ (Kaalep, Muischnek 2002), lisandus Tasakaalus korpuse põhjal tehtud loenditesse ka teaduskirjanduse tekstižanr. Vaadeldes kolme tekstiklassi vahelist sageduste jaotust, oli näha, et ajakirjandus on vahepealne osa kahe äärmuse - ilu- ja teaduskirjanduse vahel. Kui eesmärgiks oleks koostada tavalisemate sõnade loend, siis oleks soovitatav tekstisõnede kuhjumise vältimiseks kasutada pigem ptk 1.3. kirjeldatud keskmiselt vähendatud sageduse meetodit või siis tekstiklasside siseselt vähendatud jaotuse meetodit.

Töö järgmine osa keskendub sagedusloendites sisalduva materjali analüüsile, mis on esitatud võrdluses EKSS'i märksõnaloendiga.

6. Sagedussõnastiku valikuline võrdlus sõnaraamatu märksõnaloendiga

Peatükis võrreldakse korpuse sagedussõnastiku loendeid valikuliselt sõnaraamatu märksõnaloendiga. Miks üldse võrdlus sõnaraamatuga? Esiteks on muidugi huvitav teada, kuidas kattub sõnaraamatu märksõnaloend korpuse materjaliga, see annab hinnangu nii korpuse kui ka sõnaraamatu leksikonile. Lisaks sellele on tänu loendite ühendamisele võimalik kiiremini leida korrektseid korpuses esinevaid, aga samas ka sealt puuduvaid sõnu. Sõnaraamatusse lisatud sõnad on korrektsed, kuid korpuses leidub palju vigu, (trüki-, kirja-, ühestamisvead). Samas on võrdluse abil võimalik leida sõnaraamatust puuduvaid sõnu ning hinnata sõnaraamatus sõnade reaalset kasutust võrdlusaluse korpuse piires. Samuti on võimalik tuvastada õigekeelsuse normist regulaarselt erinevaid muutevorme.

Pealkirjas oleva märksõna *valikuline* all mõeldakse seda, et võrdlusesse võetud materjali on piiratud. Esiteks otsustati magistritöö mahust lähtuvalt piirduda kahe sõnaklassiga: verbid ja substantiivid. Teise valikulisuse kriteeriumi tingis nimetatud sõnaklasside erinev moodustusviis. Liitsõnalised käändsõnad on avatud sõnaklass – käändsõnalisi liitsõnu võib moodustada lõputult ja kõiki neid sõnu pole otstarbekas sõnaraamatusse lisada. Liitverbid ja liitsõnalised verbituletised on aga oletatavasti suletum sõnaklass, liitverbe erinevalt liitsõnalistest käändsõnadest nii kergesti ei moodustata. (EKK 2000: SM 19) Töös otsustati vaatluse alla võtta kõik verbid ja substantiividest piirduti lihtsõnadega. Lisaks eelnevale selgus töö käigus, et sõltuvalt sõnaliigist ja sõnade arvust oli tarvilik analüüsimisel võrdlusesse võetud sõnade piiritlemine. Eelnevast lähtudes tuleb öelda, et peatükis esitatud võrdluse näol on pigem tegemist katsega võrrelda korpuse ja sõnaraamatu leksikoni, mitte põhjapaneva ja täieliku võrdlusega.

Järgnevates võrdluse peatüki alajaotustes on täpsemalt kirjeldatud võrdlemise alla võetud materjali.

6.1. Võrdluseks kasutatud materjal

Võrdluse osas kasutatud materjali on juba ülevaatlikult tutvustatud neljandas peatükis - „Töös kasutatud andmekogud“. Käesoleva peatüki alajaotuses põhjendatakse võrdluse eesmärgist lähtuvalt materjali valikut ja kirjeldatakse materjal võrdluseks ette valmistamist. Esmalt kasutatud korpusematerjalist.

6.1.1. Tasakaalus korpus ja Koondkorpus

Viiendas peatükis koostati sagedusloendid Tasakaalus korpuse põhjal, kuid nagu juba sissejuhatavas peatükis kirjutati, otsustati võrdluseks sõnaraamatu märksõnaloendiga kasutada ka Koondkorpust. Miks otsustati võrdluses kasutada mõlemat nimetatud korpust? Kokkuvõtlikult küsimusele vastates tuleb öelda, et eelkõige võib valikut põhjendada sobilikuma materjali puudumisega.

Töös koostatud sagedusloendid põhinevad 15-miljoni sõna suurusel korpusel, samas kui „Eesti keele seletava sõnaraamatu“ aluseks on olnud umbes 100 miljoni suurune tekstikorpus (vt ptk 4.2.). Sellest lähtuvalt võib öelda, et võrreldavad materjalid on koostatud liiga erineva suurusega korpuste põhjal. Üks näide suurema tekstikorpuse vajalikkusest on võrdlusest välja tulevate reaalsest kasutusest puuduvate sõnade loetelu koostamine. Tänu suurema korpuse kasutamisele ei ole saadud sõnade loetelu nii suures sõltuvuses kasutatud korpuse mahust, kui see oleks olnud väiksema korpuse põhjal.

Samas tuleb tähele panna, et kasutatud Koondkorpus on küll sõnade arvult mahukas andmekogu, kuid korpuse nõrgaks küljeks on üldine korpustes esinema kippuv probleem, nimelt tänapäeva korpused põhinevad suuresti kergesti kättesaadaval ajakirjanduskeelel. Seega, kui seda tüüpi informatsiooni leidub korpustes piisavalt, siis samas teistest tekstitüüpidest jääb vajaka. Üldise kogu keelt katva korpuse tegemiseks, mis oleks sobilik sõnaraamatu koostamiseks (ja ühtlasi ka sõnaraamatuga võrdlemiseks), oleks tarvis rikkalikku keelt sisaldavat tasakaalustatud tekstivalikut kõigist tekstitüüpidest, mitte ainult mõnest neist. (Čermák, Křen, 2005: 2) Tehtavas võrdluses annab ülevaate eesti kirjakeele tekstivalikust Tasakaalus korpus, mis jaguneb võrdselt kolme tekstiklassi vahel (aja-, ilu- ja teaduskirjandus).

Seega Koondkorpus annab võrdlusesse juurde vajalikku keelematerjali ning selle allosa – Tasakaalus korpus aitab võrdlemisel jälgida kasutatud tekstiklasside omavahelist tasakaalu.

Koondkorpuse võrdluseks kaasamisel koostati selle põhjal sarnaselt peatükis 5 koostatud Tasakaalus korpuse loenditele lemmade sagedusloend. Sellest eraldati omakorda morfoloogilisel märgendamisel verbi ja substantiivi analüüsi saanud read, viimastest lihtsõnade kättesaamiseks eemaldati loendist liitsõna piiri – alakriipsu (_) sisaldavad sõnad. Koostatud loendist sai võrdluse peatüki alusloend. Kuna kasutada saadud EKSS'i loendi kuju erines sellest mõnevõrra, siis järgnevas alapeatükis kirjeldatakse pikemalt EKSS'i loendi võrdluseks ettevalmistamist, sealhulgas ka korpuse loendiga samale kujule viimist.

6.1.2. EKSS'i märksõnaloend

Teiseks võrdluse pooleks, nagu juba nimetatud, oli EKSS'i märksõnaloend[1]. Kuna korpuse sagedusloendi ja sõnaraamatu märksõnaloendite võrdlus oli valikuline, siis tuli loendist eraldada vajalik materjal (verbid ja lihtnimisõnad). EKSS'i failis (vt Tabel 12) olid sõnaklassid osaliselt tähistatud (sealhulgas ka töös kasutatavad lihtnimisõnad), lisaks sellele oli loendis esitatud ka sõnade stiiliregistri ja kasutusvaldkondade märgendid (vt ka 3.2., 3.3.). Märgendatud faili kasutamine muutis ühelt poolt töö huvitavamaks, pakkudes pilguheitu sõnu markeerivatele tunnustele ja samas ka lihtsamaks, kuna loendist võrdluseks kasutatavate lihtnimisõnade eraldamine oli (pool)automaatselt võimalik.

Tabel 12. **Väljavõte võrdluses kasutatud EKSS'i märksõnaloendi algkujust**

Märksõna(d)	Sõnaliik	Valdkond	Stiil
mask	s	fot	piltl
maskaroon	s	kunst	-
maskeerima	-	sõj	piltl
\maskeerimis\ülikond	-	-	-
maskeering	s	-	-
maskeeritult	adv	-	-

[1] Loendid on saadud Eesti keele instituudi leksikograafiasektorist aprillis 2012, autor tänab Andres Loopmanni ja Margit Langemetsa lahke abi eest.

Verbidel, nagu ka näiteks liitnimisõnadel polnud EKSS'i failis sõnaliigi märgendit, seepärast tuli enne korpuse failidega võrdlemist teha veel mõned sammud EKSS'i loendist verbide kättesaamiseks. Esmalt ühendati sõnaraamatu loend kogu Koondkorpuse failiga, seeläbi oli võimalik eemaldada sõnaraamatu märksõnaloendist sõnad, mis olid Koondkorpuses olemas, aga kuulusid tegusõnadest erinevasse sõnaklassi. Nüüd jäid faili alles EKSS'i ja korpuseloendiga ühtivad verbid ja kõik korpusest puuduvad, kuid EKSS'is esinevad sõnad. Viimastest verbide kättesaamiseks kasutati verbe eristavat tunnust sõnastikukirjes - tegusõnad lõppevad *ma*-tegevusnimega. Lõpliku verbide nimekirja saamiseks tuli siiski loend käsitsi läbi vaadata, sest sisse jäid mõningad *ma*-lõpulised, korpusest puuduvad ja EKSS'i loendis sõnaliigi suhtes märgendamata sõnad, nt *süliema, varuema, laevaema, kinnitussumma, alglima, kaasajateema*.

Nagu eelnevalt öeldud, lihtnimisõnad olid EKSS'i loendis sõnaliigi suhtes märgendatud, kuid enne korpusefailiga võrdlemist oli vajalik ühtlustada korpuse ja sõnaraamatu liitsõnamärgendi kasutamist. Nii näiteks esines korpuses liitsõnu, millel oli morfoloogilise analüüsi käigus liitsõnapiir määramata jäänud, seepärast kontrolliti korpuse põhjal moodustatud substantiivide loendis sisalduvate sõnade lihtsõnalisust ka EKSS'i faili abil, milles sõnapiire tähistasid kaldkriips (\) ja püstkriips (|). Kontrollimiseks ühendati EKSS'i failis liitsõnamärgendi saanud read Koondkorpuse põhjal lihtsõna analüüsi omavate sõnadega ning seejärel eemaldati kattuvad sõnad edaspidi kasutatavast lihtsõnade loendist, näiteks puudus korpuse morfoloogiliselt analüüsitud versioonis liitsõnapiiri tähistav märge järgnevatel sõnadel: *jojobaõli, juhuseks, jumalakoda, jututuba*.

Lisaks eelnevalt kirjeldatule - korpusest puudus EKSS'is olev sõnapiiri märgend - leidus ka vastupidiseid juhtumeid, nt sõna *atmosfäär* jaotus korpuse morfoloogiliselt analüüsitud tekstides liitsõnaks *atmo_sfäär+0*, kuid EKSS'i märksõna kirjetes esines ta lihtsõnana. Korpusest puuduvate, kuid EKSS'is olemas olevate sõnade hulgast eemaldati need, mis esinesid korpuses liitsõnana. Seega, sarnaselt sõnale *atmosfäär* on sagedusloenditest välja jäänud ka teised sõnad, mille lihtsõna vs liitsõna märgendus kahes andmeallikas ei ühtinud.

Peale kirjeldatud sõnapiiri erinevusele tekitas segadust ka sõnaliigi märgend. EKSS'is olevate, kuid korpusest puuduvate sõnade hulka oli jäänud neid, mis olid saanud korpuses substantiivist erineva analüüsi. See tähendab, et nendel sõnadel polnud võimalustki EKSS'i loeteluga kattuda. EKSS'is kuulusid substantiivide alla nt *dvd, cd* ja *aids*, mis korpuses said analüüsiks *_Y_* ehk lühend. Need sõnad, mis korpuses olid olemas, kuid olid saanud teise analüüsi, eemaldati korpusest puuduvate sõnade loendist.

Tagasi pöördudes konkreetselt EKSS'i loendi juurde, Tabelist 12 (lk 53) näha, et EKSS'i loendi algkuju erineb korpuse sagedusloendi kujust. Lähtuvalt sellest oli järgmine etapp, millega materjali võrdluseks ette valmistamisel kokku puututi EKSS'i märksõnaloendi töötlemine korpuse põhjal koostatud sagedusloendiga samale kujule. Tehtud loendi ühtlustamise kirjeldamisel on võrdusmärgi (=) ees esitatud märksõnakuju EKSS'i märksõnaloendis ning võrdusmärgi järel on esitatud sõna teisendatult võrdluses kasutatavale kujule.

EKSS'ist on eemaldatud mitmesõnalised fraasid ning loend on sümbolitest puhastatud. Seega on EKSS'i märksõnaloenditest välja jäänud näiteks eraldi märksõnaks olnud *alla laadima* ning liitverbide liitsõnaosade vahelised piirid on eemaldatud, näiteks *taas|alustama = taasalustama* ning *\taas\esitama = taasesitama*. EKSS'i loendites on koolonitega (: :) eraldatud mitmussõnad (*plurale tantum) (*nt *lutikaline :: lutikalised*), korpuse põhjal tehtud sagedusloendis on lemmad ainsuses, siis valiti ka EKSS'i loenditesse ainsuse vorm, seega *lutikaline :: lutikalised = lutikaline*.

Deminutiivide fakultatiivne *-ne* on märksõnas esitatud ümarsulgudes, näiteks *päike(ne)*. Lisaks on sarnaselt märgendatud veel mõne märksõna osa, näiteks *eine(s)tama*. Sellistel juhtumitel on loendisse võetud mõlemad variandid ehk siis viimase näite puhul *eine(s)tama = einestama* ja *einetama* ning *päike(ne) = päike* ja *päikene.*

Kuna korpuse põhjal sagedusloendite tegemisel on numbreid sisaldavad read välja jäetud, siis on need eemaldatud ka EKSS'i loendist, nii on välja jäetud näiteks *mp3* jms. Samakujulise nimetavaga või homonüümsete märksõnade mitmekordsed esinemised EKSS'i loendis on eemaldatud, näiteks *maht* ja *maht_ = maht*.

EKSS'i märksõnaloendina on eraldi välja toodud uute sõnade loend, milles on esitatud sõnaraamatusse lisatavad uudissõnad. Kuna EKSS'i sõnade loend on esitatud mõistete kaupa, siis ühe kirjapildiga sõnadel võib olla erinev tähendus. Nii on näiteks uudissõnade loendis esitatud sõna *bambus*, mis on sõnaraamatus juba teise tähendusega esitatud. Kuna korpuse põhjal ei ole võimalik teisiti kui konkreetseid kasutusjuhte üle vaadates kindlaks teha, millist tähendust on kasutatud, siis selles töös eraldi uudissõnade loendiga ei tegeletud. Küll aga arvati uudissõnad välja peatükis 6.4.3. koostatud sõnaloendist, mis sisaldab EKSS'ist puudunud sõnu.

Eelnevalt mainiti, et kasutatud EKSS'i failis esitati lisaks märksõnale ja tema sõnaliigile ka tema stiiliregistri ja kasutusala märgend. Tuleb tähele panna, et sõna võib olla mitmetähenduslik ja seega tema erinevad tähendused võivad EKSS'i loendis omada ka erinevaid stiiliregistri ja kasutusala märgendeid. Näiteks sõna *lagrits*, mille esimene tähendus (EKSS'i kohaselt *'magusjuurest saadav magusaine'*) ei oma kasutusala ega stiiliregistri märgendit, kuid teine tähendus (EKSS'i kohaselt *'oravast väiksem segametsade näriline'*) on märgendatud lühendiga *ZOOL*, mis tähendab, et tegemist on zooloogia valdkonda kuuluva sõnaga. Kuna korpuse sagedusloendi põhjal on hetkel võimatu sõnade erinevate tähenduste määratlemine, siis antud töös on arvestatud sõnade, mitte tähenduste tasandit.

Viimane „mugandus" nõudis ka EKSS'i faili töötlust. Näiteks esineb EKSS'i märksõnaloendis homonüümne sõna *pill*. Juhul kui seda sõna korpuses ei esineks, siis annaks ta korpuses mitte esinenud sõnade arvule juurde ühe sõna asemel 5 sõna. Kuna korpuse tekstide põhjal on sõnade tähenduste eristamine võimatult palju aega võttev protsess, siis on sellised sõnad koondatud ühe lemma alla. Tahtmata kaotada nende kasutusala ja/või stiiliregistri märgendeid, on kõigi mõistete märgendid kinnitatud ühe lemma külge. Ning koosesinemise korral on nimekirjad sorteeritud nii, et ka koosesinemiste arv ei oleks ühel sõnakujul mitmekordne.

Sõna *pill* on saanud viis erinevat märgendit (vt Tabel 13), kolmel korral stiiliregistri märgendi (murdekeelne, piltlik ja kõnekeelne) ning kolmel korral kasutusalamärgendi

(etnograafia ja etnoloogia, farmaatsia, merendus). Järgnevalt on esitatud sõna *pill* EKSS'is esinenud stiiliregistri ja kasutusala märgendid.

- *MER 'peli' (' rõhtsa võlliga vints ankrute allalaskmiseks ja tõstmiseks')*
- *ETN 'veski võll'*
- *MURD 'pilliroog'*
- *FARM 'väike ümmargune ravimkuulike seespidiseks tarvitamiseks; tablett'*
- *PILTL 'ritsikad häälestavad, timmivad oma pille'*
- *PILTL 'nutt, kisa, hädaldamine'*
- *KÕNEK 'sõnalist teksti, muusikat jm. helisid vahendav aparaat'*
- *KÕNEK 'mingi sõidu‑ v veoriista kohta'*

Tabel 13. **Sõna *pill* kasutusala ja stiiliregistri märgendid**

Sõna	Kasutusala	Stiil
pill	-	murd
pill	-	piltl:: kõnek
pill	etn	-
pill	farm	-
pill	mer	-

Kõik erineva märgendi saanud, kuid sama sõnaraamatu märksõnaga tähistatud sõnad on korpusega võrdlemiseks koondatud ühe sõnakuju alla, millele on lisatud kõik erinevad esinenud märgendid, seega asendatakse sõna *pill* viis erinevat esinemist EKSS'i loendis ühe esinemisega, millesse on koondatud kõigi viie esinemise märgendid, uueks loendis esinemise kujuks saab järgnev: *pill etn::farm::mer murd::piltl::kõnek.*

Sarnaselt sõnale *pill* on EKSS'is esitatud ka näiteks sõna *puhetama*, millel on EKSS'i märksõnaloendis kolm homonüümset varianti (vt Tabel 14), esimene ilma lisamärgenditeta ning ülejäänud kahe puhul on vastavalt tegemist harva esineva ning murdekeelse sõnaga.

Tabel 14. **Sõna *puhetama* kasutusala ja stiiliregistri märgendid**

Sõna	Kasutusala	Stiil
puhetama	-	-
puhetama	-	hrv
puhetama	-	murd

Märgendite koondamisel ühe sõnakuju alla jääb sõna *puhetama* EKSS'i loenditesse kujul *puhetama hrv::murd*. Kuna korpuseloendite puhul on tegemist sõnade mitte mõistete sagedusloendiga, siis on õigustatud sama sõnakuju erinevate mõistete märgendite koondamine.

6.2. Erinevus EKSS'i ja korpuse sõnavaras

Enne võrdluse ülesehitust kirjeldava peatüki juurde jõudmist on tarvilik aidata lugejal mõista korpuse ja sõnaraamatu loendite sõnavara põhimõttelist erinevust. EKSS'is olevad märksõnad on kontrollitud ja teadlikult sõnaraamatusse lisatud eestikeelsed sõnad, seevastu korpuse sõnaloendid on koostatud reaalse keelekasutuse põhjal. Korpuseloendis leidub sõnu, mis ei ole tegelikult korrektsed eesti keele sõnad; korpuse tekstid sisaldavad n võõrkeelseid sõnu, morfoloogilise ühestamise probleeme ja nii õigekirja- kui ka trükivigu. Viimaste vigade liigid eristuvad üksteisest esinemissageduse poolest, just trükkimisest tekkivad vead - täh(t)e(de) puudumised ning asukoha muutumised põhjustavad vigu, mille korpuses ühekordne esinemine on sagedasem kui levinud grammatikavigadel. Ptks 3.1. anti ülevaade Zipfi seadusest, mille kohaselt madala esinemissagedusega sõnu leidub korpuses kõige rohkem. Osa neist moodustavad seega ka korpuses leiduvad vigased sõnavormid. Järgneva näite varal (vt Tabel 15) on kirjeldatud, miks otsustati analüüsist välja jätta üks kord esinevad sõnad. Illustreerimaks pisut seda pilti, mida korpuses leidub ja põhjendamaks, miks ei saa korrektseid võrdluse loendeid tekitada automaatselt, on võetud korpusest sõna *pressiesindaja* erinevad esinemise variandid, neist vaid esimene (kõige sagedasem vorm) on korrektne.

Tabel 15. **Zipfi seadus sõna *pressiesindaja* näitel**

Sagedus	***Lemma***	***Sagedus***	***Lemma***
39445	**pressiesindaja**	1	pressiseindaja
47	pressiesindja	1	pressiiendaja
19	pressesindaja	1	pressiessindaja
16	pressisesindaja	1	pressiesineja
11	pressiesndaja	1	pressiesindsaja
8	pressiesidaja	1	pressiesindajaja
5	pressiesidnaja	1	pressiesindaaja
5	pressieisndaja	1	pressiesimdaja
4	pressieesindaja	1	pressiesiindaja
4	presiesindaja	1	pressiesiesindaja
3	pressiesinadaja	1	pressiesiandja
3	pressieindaja	1	pressiensindaja
2	pressiesindjaja	1	presseesindaja
2	pressiesindadaja		
2	pressiesinaja		
2	pressieseindaja		
2	pressiendaja		

Sagedusloendis alla liikudes on näha erinevad korpuses esinevad vigased sõnakujud. Tähelepanu tuleks pöörata sellele, et vaid korra korpuses esinenud sõna *pressiesindaja* vigaseid variante oli peaaegu sama palju (13) kui selle sõna rohkem kui üks kord esinenud vigaseid variante kogu korpuses kokku (16).

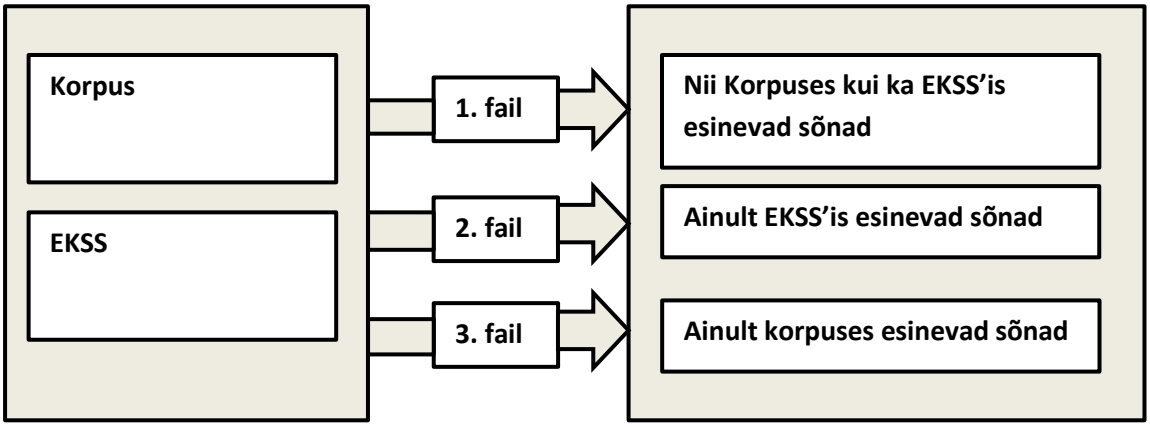
Seega kuigi korpuses olevad haruldasemad/uued sõnad võivad esineda (ja esinevadki) korpuses vaid ühe korra, leidub üks kord esinevate sõnade hulgas ka väga palju praaki, mille käsitsi läbivaatamine on kaunis üksluine ja ajakulukas. Käesolevas töös otsustati käsitsi üle vaatamist nõudvates failides piirduda **vähemalt kaks korda** esinevate sõnadega.

6.3. Võrdluse ülesehitus

Loendite võrdlemiseks kasutati Linuxi käsku *join*, mille abil ühendatakse korpuse põhjal koostatud sagedusloendid ja EKSS'i märksõnadeloendid. Võrdlemisel saadi nii korpuses ja sõnaraamatus olevate ühiste sõnade loend kui ka loendid, milles on vaid EKSS'is olevad ja korpusest puuduvad sõnad ja vastupidi, korpuses olevad ja EKSS'ist puuduvad sõnad. Võrdlemisel kasutatakse ka Tasakaalus korpuse põhjal tehtud sagedusloendeid. Nende põhjal on näha, kuidas korpuse suurus mõjutab sõnade kattumist sõnaraamatu märksõnaloendiga. See tähendab, milline on korpuses ja sõnaraamatus olevate ühiste sõnade või siis vaid ühte võrdlusobjekti kuuluvate sõnade hulk siis, kui korpusena kasutati kas siis Koondkorpuse või Tasakaalus korpuse põhjal tehtud sagedusloendeid. Teiseks oli Tasakaalus korpuse põhjal tehtud loendite abil võimalik jagada sõnad kolme Tasakaalus korpuse tekstiklassi vahel ja seeläbi leida sõnaraamatuga ühiste sõnade esinemise sagedus nii aja-, ilu- kui ka teaduskirjanduse allosas. Sõnadest, mis kuulusid vaid ühte nimetatud tekstiklassidest, joonistus tinglikult välja tekstižanri iseloomulik sõnavara. Ülevaatlikkuse mõttes jagati võrdlemise peatükk väiksemateks terviklikeks üksusteks:

1) esiteks **sõnaliikide** kaupa: analüüsi alla võetakse kahte sõnaliiki - verbi ja substantiivi klassi kuuluvad sõnad,

2) teiseks koostati mõlema sõnaliigiga **kolm** loendit - esimese neist moodustasid nii EKSS'is kui ka Koondkorpus esinevad sõnad ning ülejäänud kahes grupis on vastavalt ainult EKSS'is olevad ja Koondkorpusest puuduvad sõnad ning ainult Koondkorpuses esinevad sõnad, mis puuduvad EKSS'i märksõnaloendist.



Joonis 3. **Korpuse sagedusloendi ja sõnaraamatu märksõnaloendi võrdluse ülesehitus**

Joonis 3. seletab lahti võrdluse struktuuri. Alusmaterjaliks olevad andmed saadi EKSS'ist ja Koondkorpusest. Esmalt moodustati failid nii korpuses kui ka EKSS'is esinenud verbidest ja substantiividest, seejärel koostati ülejäänud joonisel kujutatud failid. Seega, kokku moodustati mõlema sõnaklassi peale 3+3=6 suuremat faili, mida omakorda oli võimalik jaotada väiksemateks terviklikeks üksusteks.

Läbivalt kogu võrdluse peatükis esitatakse lühikesed väljavõtted töö lisades olevatest tabelitest, mis annavad kohe teksti lugedes saada ülevaade failides sisalduvast materjalist ja aitavad mõista analüüsi tulemusi.

Järgnevalt tehtud võrdlustest ja nende tulemustest, esmalt nii korpuses kui ka EKSS'is esinenud sõnadest.

6.4. Võrdlus

6.4.1. Nii Koondkorpuses kui ka EKSS'is esinevad sõnad

Nii Koondkorpuses kui ka EKSS'is esinevad verbe oli kokku 6829. Saadud verbide nimekiri kõrvutati Tasakaalus korpuse sagedusloendiga, see tähendab, kui verb esines Tasakaalus korpuses, siis lisati juurde vastava verbi sagedus kogu Tasakaalus korpuses ja selle järel sagedused Tasakaalus korpuse kolmes osas (aja-, ilu-, teaduskirjanduses).

Tabel 16. **Nii Koondkorpuses kui ka EKSS'is esinevad verbid**

Nii Koondkorpuses kui ka EKSS'is esinevad verbid		6829	
Puuduvad Tasakaalus korpusest			1186
Esinevad Tasakaalus korpuses			5643
	Ainult ajakirjanduses	190	
	Ainult ilukirjanduses	1060	
	Ainult teaduskirjanduses	316	

6829 nii Koondkorpuses kui ka EKSS'is esinenud verbist 5643 esinesid Tasakaalus korpuses, seega 1186 verbi, mis esinesid nii Koondkorpuses kui ka EKSS'is, ei esinenud Tasakaalus korpuses (vt Tabel 16). Kolme tekstiklassi vahel verbe jagades selgus, et ainult ajakirjanduses esinevaid verbe oli 190, teaduskirjanduses 316 ning ainult ilukirjanduses leidus 1060 verbi.

Nii Koondkorpuses kui ka EKSS'is esinenud lihtsõnalisi nimisõnu oli 20038, nendest 3757 puudus Tasakaalus korpusest ning 2140 nimisõna esines seal ainult ühe korra. Substantiivide puhul oli näha, et ainult ajakirjanduses esinevaid sõnu (970) oli tunduvalt vähem kui ainult ilukirjanduses (2057) ja ainult teaduskirjanduses (1743) esinevaid sõnu. Samas, sarnaselt verbidele oli ainult ilukirjanduses leiduvad sõnu kõige rohkem (vt Tabel 17).

Tabel 17. **Nii Koondkorpuses kui ka EKSS'is esinevad substantiivid**

Koondkorpuses ja EKSS'is koos esinevad substantiivid		**20038**	
Puuduvad Tasakaalus korpusest			3757
Esinevad Tasakaalus korpuses			16281
	Ainult ajakirjanduses	970	
	Ainult ilukirjanduses	2057	
	Ainult teaduskirjanduses	1743	

Järgnevalt analüüsitakse nii korpuses kui ka sõnaraamatus esinenud sõnu, mida leidus ainult ühes korpuse osas (aja-, ilu- või teaduskirjanduse tekstides). Kuna Koondkorpus koosneb suuresti ajakirjanduse tekstidest, siis sõnade tekstiklasside jagamisel toetuti Tasakaalus korpuse sagedustele. Igast tekstiklassist esitatakse loend 20 ainult selles tekstiklassis esinenud verbi ja substantiiviga ning seejärel kirjeldatakse üldistatult žanrisisest sõnavara. Täispikad loendid on esitatud töö lisades.

Nagu eelnevalt maininud, Tasakaalus korpuses esines 190 verbi ning 970 substantiivi, mida leidus ainult ajakirjanduse tekstides.

Tabel 18. **Tasakaalus korpuse ajakirjanduse allosas esinenud verbid ja substantiivid**

Sagedus Koondkorpuses	Lemma	Sagedus Tasakaalus korpuses
474	basiilik	9
456	finišeerima	14
340	hoonestama	7
613	katamaraan	7
383	keraamik	10
124	krediteerima	5
2615	loosima	73
251	munitsipaliseerima	6
999	oligarh	11
1348	palling	41
305	pedaalima	4
965	purjetaja	22
299	spaa	6
138	sprintima	4
55	steppima	6
430	subsideerima	18
37	telefoniseerima	8
360	trampliin	12
662	triatlon	37
319	virtuoos	14

Ainult ajakirjanduse tekstides esinenud verbe iseloomustab väike esinemissagedus, vaid neli verbi esinevad Tasakaalus korpuses kümme või enam korda, ülejäänud esinevad vähem. Verbide hulka kuulub toorlaene, nt *hängima, skannima,* lisaks sellele ka *eerima*-lõpulisi sõnu, nt *demonteerima, parafeerima*. Ajakirjanduse substantiivide hulgas oli enim uuema aja nähtusi/esemeid kirjeldavaid sõnu, näiteks: *soundtrack, spaa, casting, sämpel*. Huvitaval kombel kerkisid ajakirjanduses esinevatest sõnadest esile spordivaldkonda kuuluvad sõnad, näiteks *pallur, sett, koondislane, tsenderdama*. Ilmselt ei kuulunud ilu- ja teaduskirjanduse tekstide valikusse spordivaldkonna kirjandust, samas ajalehtedes on aga spordirubriik või vähemalt päevakajaliste spordiuudiste kajastamine tavaline. Ajakirjanduse sõnavaras tõusis esile ka poliitikaga seotud sõnavara, nt *pikettima, leiborist,*

äärmuslane, kongresmen ning seoses sellega ka asukohaga seotud inimese nimetamised, näiteks *albaanlane, horvaat, norralanna, narvalane* ja rahaühikud, nt *tolar, leev, ruupia*.

Ainult teadustekstides leidus 316 verbi ja 1743 substantiivi.

Tabel 19. **Tasakaalus korpuse teaduskirjanduse allosas esinenud verbid ja substantiivid**

Sagedus Koondkorpuses	Lemma	Sagedus Tasakaalus korpuses
32	infitseerima	31
46	lokaliseeruma	28
57	inkubeerima	47
60	tüsistuma	24
63	märgendama	58
79	legitimeerima	31
81	standardima	59
124	redutseerima	90
132	indutseerima	108
159	hospitaliseerima	109
232	resistentsus	184
232	pronoomen	231
233	levimus	220
252	entiteet	233
275	menopaus	30
285	transistor	117
121	meelika	179
311	varieeruvus	271
327	laktatsioon	311

Teaduskeele sõnavara iseloomustab oskuskeele rohkus, sõnavara eristuvad selgesti aja- ja ilukirjanduse sõnavarast. Teaduskirjanduses esindatud erialaterminid on tugevalt seotud Tasakaalus korpusesse valitud tekstidega, näiteks kuna Tasakaalus korpus sisaldab nii keeleteaduse ja meditsiinialaseid tekste, siis leidub korpuses nii keeleteaduse mõisteid (nt *morfoloogia, klusiil, genitiiv)* kui ka meditsiinitermineid (nt *aneurüsm, gastriit, vaskuliit*). Teadustekstides leidub ka hulgaliselt *eerima-, eeruma*-lõpulisi sõnu.

Võrreldes teiste tekstiklassidega leidus ilukirjanduses rohkem tekstiklassi piiresse jäävaid sõnu, ainult ilukirjanduses leidus 1060 verbi ning 2057 substantiivi. Ilukirjanduses olid just verbid selgesti eristuvad – levinud olid onomatopoeetilised verbid. Ilukirjandusele omaste nimisõnade hulgas leidus onomatopoeetilisi ja deskriptiivseid sõnu, nt *nuukse, raksatus, võpatus, lörin*. Leidus ka halvustava tähendusega sõnu, nt *lirva, tolgus* ja samas ka neile vastanduvaid emotsionaalset deminutiivsust märkivaid sõnu, nt *taadike, latseke, musuke*. Kõnekeelsustki oli teistest žanritest enam, nt *puhvaika, pinss, issu, radikas*. Samuti on ilukirjanduses talletunud aktiivsest kasutusest kadumas olevaid või juba kadunud sõnu, nt *bonne, sõbruke, körtsik*.

Tabel 20. **Tasakaalus korpuse ilukirjanduse allosas esinenud verbid ja substantiivid**

Sagedus Koondkorpuses	Lemma	Sagedus Tasakaalus korpuses
48	konspiraator	26
169	kähisema	64
110	luristama	33
47	luuletajanna	26
109	miraaž	24
100	nagisema	34
129	nurruma	39
51	nuukse	27
130	nõksatama	61
88	pilkuma	43
97	pungitama	32
123	raksatus	23
87	rinnatis	25
124	sonima	33
212	talaar	24
59	tolgus	24
114	võbisema	29
99	võdisema	31
54	võpatus	24

Loendites esitatud sõnade hulgas leidus selliseid, mis tundusid peale substantiiviks olemise kuuluvat ka teise sõnaklassi. Näiteks teaduskirjanduse sõnade hulgas leidus sõna *meelika*, mida EKSS'is esitatud tähenduses *'vanakreeka lüürilised luuletused, mida kanti ette lauldes'* korpuses ei leidunud. *Meelika* on korpuses saanud vale morfoloogilise analüüsi, märgitud substantiivi asemel on ta tegelikult korpuses kasutuses pärisnimena.

Pisteliselt saadud sagedusloendit üle vaadates leiti veel pärisnimesid, mis olid substantiivi analüüsi saanud ja seega tabelisse jäänud. Näiteks sõna *riin* tundub esmapilgul tavalise eesnimena, kuid korpuses esines ta 4 korda substantiivina. EKSS'i kohaselt on tal selleks täielik „õigus" olemas, nimelt on tal ka harva esinev tähendus *piirkond* ning murdekeelne tähendus *sirel*. Korpuses esinemisi üle vaadates selgus, et tegemist polnud korrektsete substantiivide esinemisega, nt saadi analüüsiks *riin* jägmises näites:

tippbale tippba+le //_S_ sg all, //
riinidega riin+dega //_S_ pl kom, //

On näha, et probleem on tekkinud sõna poolitamisel ja selle tulemusena on ühest sõnast moodustunud kaks ja mõlemad sõnaosad on eraldi morfoloogiliselt analüüsitud. Sarnaselt eelneva näitega kerkis esile sõna *kanter*, mis pärisnime kasutuses jookseb läbi tuntud Eesti

kettaheitja perekonnanimena. EKSS'i märksõnastikus on substantiivi seletusteks antud *'kirikumuusika õpetaja koolis, kiriku eeslaulja ja koori juhataja'*. Korpuses leidub lauseid, milles pärisnimena esinenult on *kanter* saanud substantiivi analüüsi, näiteks:

Rõõmutses rõõmutse+s //_V_ s, //
ka ka+0 //_D_ //
suurvõistluste suur_võistlus+te //_S_ pl g, //
uustulnuk uus_tulnuk+0 //_S_ sg n, //
Kanter kanter+0 //_S_ sg n, //

Korpust üle vaadates selgus aga, et oli tekkinud ka morfoloogilise ühestamise probleem, mille käigus muusikastiil *kantri* (*'USA-st pärinev suurelt osalt rahvamuusikal põhinev levimuusika liik'* (EKSS)) oli saanud analüüsiks *kanter*, nt

plaat plaat+0 //_S_ sg n, //
kõlab kõla+b //_V_ b, //
peaaegu pea_aegu+0 //_D_ //
nagu nagu+0 //_D_ // nagu+0 //_J_ //
kantri kanter+0 //_S_ sg g, //

Lisaks eelnevale leidus sõnu, mis kuuluvad mitmesse sõnaliiki, nt *ai* on nii tuntud interjektsioon kui ka tähenduses *'seeliku (v. särgi) äärispael'* (EKSS). Korpuses siiski sõna *ai* viimases tähenduses ei leidunud, substantiivi analüüsi saanud sõna esines järgnevas järjendis:

liigutamist liigutamine+t //_S_ sg p, //
ja ja+0 //_J_ //
ai'd ai+d //_S_ pl n, //
karistati karista+ti //_V_ ti, //
ühe üks+0 //_N_ sg g, //
lisanipsuga lisa_nips+ga //_S_ sg kom, //

Sagedusloendid on heaks pidekohaks korpuses esinevate probleemsete kohtade leidmiseks. Mõned sõnad ei tundu esmapilgul substantiividena „tavalised" või siis lähtuvalt eelmistest näidetest, seostuvad esmajoones nimedega ning nende substantiivi tähendus ei pruugi olla nii tuntud. Samas näiteks kajastub sõnasagedustes ka substantiivi analüüsi saanud, kuid korpuses perekonnanime tähistav sõna, mis võib tunduda täiesti tavalise substantiivina. See tähendab, et erinevalt nimest *Kanter*, mille substantiivivorm ei pruugi keelekasutajale niivõrd tuttav olla, esineb perekonnanimena näiteks lindude

nimestikku kuuluv *Luik*. Tuleb silmas pidada, et sagedusi ei saa võtta üks ühele, sageduste taga peitub rohkem informatsiooni.

Loendites esineb vale morfoloogilise analüüsi tõttu sagedusloendisse sattunud sõnu nagu näiteks eelnevalt välja toodud *meelika*, mille puhul on tegemist korrektse sõnaga, mis on esindatud ka EKSS'is, kuid mille EKSS'is esindatud tähendus pole korpuses tegelikult kajastatud. Üldist sõnasagedusloendit vaadates ei pruugikski vale analüüs silma paista, kuid korpuse sõnaliigiti analüüsimisel kerkis see esile. Töös esitatud loenditesse on need sõnad sisse jäetud just selle tõttu, et pöörata tähelepanu korpuses tegelikult leiduvale materjalile. Kogu sagedusloendit ei jõua üks inimene läbi vaadata ja põgusal ülevaatamisel ei olekski probleemkohad nähtavad. Selleks, et viga märgata, on tarvis teada, mida otsida tuleb, seetõttu on pigem kasulikum võimalikele veakohtadele tähelepanu pööramine kui osaline faili parandamine.

Järgmises peatükis juba nendest sõnaraamatu märksõnadest, mida Koondkorpuses ei leidunud.

6.4.2. Sõnad, mis esinevad EKSS'is, kuid puuduvad Koondkorpusest

Peatükki on koondatud need EKSS'i märksõnaloendis olevad sõnad, mida ei leidunud Koondkorpuses.

Tabel 21. **Korpusest puuduvad verbid ja substantiivid**

Sõnaliik		Verbid	Substantiivid
Kokku		3342	5012
Märgendita		1588	1980
Märgendiga		1754	3032
	Stiil	1388	885
	Kasutusala	352	2130
	Mõlemad märgendid	14	17

Tabelist 21 selgub, et EKSS'i märksõnaloendis leidus 3342 korpusest puuduvat verbi. Nendest üle poole ehk 1754 omasid mingit märgendit: stiiliregistri märgendit 1388 ja kasutusmärgendit 352 sõna ning 14 sõna omas mõlemat märgendit. 1588 sõna puudus korpusest ning ei omanud EKSS'i märgendit.

Koondkorpusest puuduvaid substantiive leidus EKSS'i märksõnastikus 5012, sarnaselt verbidega nendest üle poole omasid stiiliregistri või kasutusala märgendit, vastavalt 885 ja 2130 ning 17 sõnal olid mõlemad märgendid. Substantiividel esines sõnastikus verbidest rohkem kasutusala märgendeid (verbidel 352) ja vähem stiiliregistri märgendeid (verbidel 1388).

EKSS'i loendi ettevalmistamise osas oli kirjas, et deminutiivvormis sõnadest, mida loendis esitati sulgudega, näiteks sõna *päike(ne)*, võeti märksõnade võrdlemise loendisse sisse mõlemad sõnad (*päike* ja *päikene*). Sellistest sõnadest neid, mille kumbki vorm polnud korpuses esindatud, leidus 39, nt *tolake* ja *tolakene*, *topsuke* ja *topsukene*, *tossuke* ja *tossukene*, *totsike* ja *totsikene*. Ülejäänutest oli vähemalt üks või mõlemad paarilised korpuses esindatud, eraldi väljavõtet pole selle kohta töös esitatud.

Järgnevalt lähemalt märgendite taga peituvatest sõnadest. Tabelis 22 on esitatud 5 sagedasemat stiilimärgendit, mis esinesid Koondkorpusest puudunud, kuid EKSS'is leidunud sõnadel. Tabelist on näha, et kõige rohkem puudus korpusest sõnaraamatus harva esinevatena märgendatud sõnu, murde- ja kõnekeelseid sõnu ja piltlikke väljendeid ning seejärel luulekeele ning halvustava märgendi saanud sõnu. Verbide ja substantiivide sagedasemate stiiliregistri märgendite loetelud kattuvad, erinevuseks on vaid see, et sageduselt esimene ja

teine märgend on koha vahetanud – verbidel on kõige sagedasem märgend *hrv* (harva esinevad sõnad), substantiividel aga *murd* (murdesõnad).

Tabel 22. **Verbide ja substantiivide stiilimärgendid EKSS'is**

Stiil	Verbidel	Substantiividel
hrv	767	270
murd	501	381
kõnek	98	201
piltl	54	24
luulek	15	15

Sõnade puudumist korpusest võib põhjendada korpuse tekstide valikuga. Korpuse maht on piiratud ja seega ei saa korpus sisaldada kõiki kindla eriala sõnavarasse kuuluvaid tekste, näiteks puuduvad sagedusloendist tekstiilitöötluse sõnavarasse kuuluvad sõnad nagu *bordeerima- 'riietuseset vms ääristama, palistama'* (EKSS), *dekateerima- 'villast riiet aurutades v vee ja auruga töödeldes viimistlema'* (EKSS), *karustama- '(riiet) karuseks tegema'* (EKSS), *eteldama- 'kootud eseme silmuseid erilise õmblusega ühendama'* (EKSS).

Sõnad kaovad aktiivsest kasutusest, sõnaraamatusse on varasemalt lisatud sõnu, mis pole enam käibel ning on talletunud vaid vanemates tekstides. Siia hulka kuuluvad harva (hrv) esinevad sõnad (sõnaseletused EKSS'ist), nt: *ärisema* (*'urisema, lõrisema*), *põdelema* (*'põdur olema, tihti põdema'*), *põietama* (*'põienahka aknaaugu vms ette kinnitama'*) ning vananenud sõnad ja murdesõnad, näiteks: *räselema* (*'rüselema, hullama'*), *pokerdama* (*'takerdudes kõnelema'*), *hämm* (*'väike heina- v viljahunnik'*), *jõmp* (*'meeste paks poolpalitu, jopp'*), *lödastik* (*'mädamülgas, tüma koht, pehme soo'*), *vihalane* (*'vaenlane, vihamees'*), *klaun* (*'kloun'*), *paljundis* (*'paljundatud eksemplar'*) , *panila* (*'panipaik'*), *pihulane* (*'kihulane'*). Kõne-, luule- ja lastekeelsete sõnade ning piltlike ja halvustavate ning vulgaarsete väljendite ja slängi puudumist korpusest võib põhjendada suulise keele tekstide puudumisega korpusest. Korpuses leidub selliseid sõnu enamasti ilukirjanduse tekstides, mille osakaal on korpuse üldmahuga võrreldes väike, mõned näited korpusest puudunud sõnadest: *põssama* (*'varastama'*), *pökerdama* (*'saamatult v. lohakalt joonistama v kirjutama, mäkerdama'*), *apama* (*'jooma'*), *sullama* (*'end pesema, kümblema'*), *helama* (*'helisema, kõlama'*), *larask* (*'lobamokk, laralõug'*), *mirgel* (*'smirgel'*), *sähkam* (*'võmm, sähvakas'*), *soum* (*'vaht'*). Lisaks sisaldab sõnaraamat ka tehistüvesid, mis „…on kirjakeele rikastamise eesmärgil vabalt loodud uued tüved" (Metsmägi jt 2012: 16) nt *üllastuma* (*'üllaks muutuma'*), *aabe* (*'kirjatäht'*).

Järgnevas Tabelis 23 on esitatud 7 sagedasemat korpusest puuduvate verbide ja substantiivide kasutusala märgendit.

Tabel 23. **Kümme sagedasemat kasutusvaldkonna märgendit**

Sagedus	Verbide kasutusala märgendid
68	tehn
42	keem
27	med
27	keel
22	põll
17	mer
17	maj

Sagedus	Substantiivide kasutusala märgendid
199	med
188	aj
171	zool
171	bot
162	geol
149	keem
137	tehn

Verbidest olid korpuses enim puudu tehnika, keemia, meditsiini, keeleteaduse, põllumajanduse, merenduse ja majanduse valdkonda kuuluvaid sõnu. Substantiividest puudus korpusest meditsiini, ajaloo, botaanika, zooloogia, geoloogia, keemia ning tehnika kasutusala märgendiga sõnu. Illustreerimaks sõnaraamatus leiduvat materjali, esitatakse väljavõtted mõningatest sagedasemast korpusest puudunud, kuid sõnaraamatus esinenud sõnadel leidunud kasutusalamärgenditest, seletused on võetud EKSS'ist.

Meditsiini kasutusala märgendiga tähistatud ridadel esines termineid, nt *juustund* (*'tuberkuloossel kärbumisel tekkiv juustutaoline mass'*), *kätgut* (*'soolest valmistatud imenduv haavaõmblusmaterjal'*) ja haiguste nimetusi, nt *buboon* (*'lümfisõlmede muhk'*), *koksiit* (*'puusaliigesepõletik'*).

Ajaloo märgendi saanud sõnad tähistasid rahaühikuid, territooriumi märgendamise ja seisuslike klasside ning eluoluga seotud vahendeid, nt *villaan* (*'feodaalist sõltunud talupoeg keskaja Lääne-Euroopas'*), *tarantass* (*'neljarattaline reisivanker'*), *sükofant* (*'vanaaja Ateenas isik, kes elatus pealekaebamisest'*), *kreevin* (*'Lõuna-Lätis Bauska ümbruses elanud vadjalane'*).

Botaanika märgendi saanud sõnade hulgas oli taimenimetusi: nt *kohhia* (*'poolkõrbealade kitsaste lehtedega kääbuspõõsas v rohttaim'*), *ringik* (*'peam. kuusikutes ja männikutes kasvav väikese kuplikujulise poolkeraja kübaraga söödav seen'*), *serradell* (*'liblikõieline rohttaim, mille üht liiki kasvatatakse meist lõuna pool söödataimena; linnujalg'*) ning muud taimedega

seotud terminoloogia, nt *zoohoor* (*'loomlevija (taim, vili v eos)'*), *teloom* (*'taimede ürgseks põhiorganiks peetav lihtsa steeliga varretaoline moodustis (teloomiteoorias)'*).

Lisaks märgendi saanud sõnadele puudus korpusest ka sõnu, mis olid stiiliregistri või kasutusala märgendiga markeerimata, failis olevaid sõnu pisteliselt EKSS'i sõnaraamatuga kõrvutades leiti, et mingil põhjusel oli kasutada saadud EKSS'i failides mõningad märgendid puudu, mis sõnaraamatu internetiversioonis olid nähtavad. Näiteks sõnal *hästitama* (*VAN 'parastama'*) ei olnud võrdluses kasutatud failis märgendit *VAN*, samamoodi puudus märgend sõnalt *telefoonima* (*VAN 'helistama, telefoneerima'*). Seega on EKSS'is esinenud ja samas korpuses puudunud sõnade hulgas märgendamata sõnu tegelikult vähem, kui töös esitatud statistika näitab. Järgnevalt mõned näited märgendamata EKSS'i märksõnadest, mis võiksid olla märgendatud. Märgistamata on verb *faulitsema* (*'fauli kasutama'*), kuid samas märksõna *faul* (*'määrustevastane ebaaus võte v toiming'*) on saanud sõnaraamatus kasutusmärgendi *SPORT*. Sarnaselt eelneva näitega, on sõna *jalutelema* märgendamata, kuid samas tema definitsiooniks olev märksõna *jalutlema* (*'edasi-tagasi, ringi jalutama'*) on saanud märgendi *HRV*.

Järgmises peatükis viimasest koostatud võrdluse failist, analüüsitakse sõnu, mis kuulusid korpuse põhjal tehtud sõnavaraloendisse, kuid puudusid sõnaraamatu märksõnaloendist.

6.4.3. Sõnad, mis esinevad Koondkorpuses, kuid puuduvad EKSS'ist

Clarence Barnhart'i andmetel jõuab inglise igapäevakeelde igal aastal umbes 800 uut sõna, millest sõnaraamatutes leiab koha ligikaudu 500. (Landau 2001: 202) Keelekorpused peegeldavad keelekasutust ja on heaks allikmaterjaliks uute kasutusele võetud sõnade leidmiseks. Käesolevas peatükis analüüsitakse Koondkorpuses leidunud verbe ja substantiive, mida EKSS'i märksõnaloendis ei leidunud.

Võrdlusest tuli välja, et 9050 verbi ja 277989 substantiivi ei leidnud vastet sõnaraamatu märksõnaloendist. Saadud arvandmete puhul tuli silmas pidada, et korpus on koostatud reaalses elus kasutatavatest tekstidest ning paratamatult sisaldas materjal ka vigaseid sõnu, mida ei tohigi sõnaraamatus esineda, seega osutus tarvilikuks faili ülevaatamine. Peatükis 6.2. kirjutati, et enim vigaseid sõnu leidus korpuses vaid üks kord leidunud sõnade hulgas ning käesolevas töös otsustati üle vaadata vaid kaks või enam korda korpuses esinenud sõnad. Kuna aga substantiivide fail osutus käsitsi läbivaatamiseks ikka liiga mahukaks, siis otsustati selle puhul piirduda vaid iga kümnenda ning sarnaselt verbidele, korpuses vähemalt kaks korda esinenud sõnaga. Seega nimisõnade 277989 sõna suurusest failist saadi 27798 sõnaline fail, millest pärast korpuses üks kord esinenud sõnade (18784 sõna) eemaldamist jäi käsitsi üle vaatamiseks järele 9014 sõna ning verbide 9050 pikkusest nimistust eemaldati 6305 üks kord korpuses esinenud sõna, ülejäänud 2745 verbist koosnev fail vaadati käsitsi üle.

Failide ülevaatamisel eemaldati trüki- ja õigekirjavigu, võõrkeelseid sõnu ja vale analüüsi saanud ridu, automaatselt eemaldati ka sidekriipsuga kirjutatud sõnad. Nii eemaldati nt õ asemel ö-ga, ü asemel y-iga, š asemel sh-ga ja ž asemel zh-ga kirjutatud sõnad ning sidekriipsuga kirjutatud sõnadest eemaldati, nt *k-rautama, või-maldama, üle-jääma* ning teised probleemsed kohad. Ilmselgelt ei võimaldanud kirjeldatud materjali piiramine kõigi sõnastikust puuduvate ja samas korpuses olemas olevate korrektsete sõnade leidmist, kuid andis siiski läbilõike korpuses esinevast materjalist.

2745 läbi vaadatud verbist jäi järele 634-sõnane fail. Umbes 300 sellest moodustasid liitverbid, sagedasim EKSS'ist puuduv, kuid samas korpuses olemasolev liitverb oli *esilinastuma*, mis esines Koondkorpuses kokku 1463 korda. Lisaks liitverbidele esines korpuses slängisõnu (*chillima*, *tšekkama*, *plekkima*), toorlaene (*renderdama*, *driftima*, *linkima*) ja tuletisi, viimastest järgnevalt pikemalt. Korpuses leidus 170 *eeri*-liitega verbi, millega moodustatakse, lisaks saksa ja vene keele vahendusel kasutusele võetud tegusõnadele, ka uusi võõrpäritolu verbe (Mäearu 2011: 68), nt *konfigureerima, sponseerima*. Enesekohast tegevust väljendava sufiksi *-u* liitumisega saadakse *eeruma*-lõpulised verbid (Mäearu 2011:

68), neid leidus korpuses 89, nt *erekteeruma*, *kanaliseeruma*, *fokuseeruma*. Hausenberg (2009: 257) kirjutab, et kuigi paljudele *eeri*-/eeru-liitelistele sõnadele on pakutud ka otsetuletuse alla kuuluvat lühemat rööpvormi, eelistavad keelekasutajad siiski pikemat, näiteks sõna *arhiveerima* on sagedasem kui sõna *arhiivima*. 81 korral leidus korpuses *stama*-lõpulisi sõnu, nt *ühestama*, *sidustama*, *zombistama*. „Sulandmorfreem *-sta* on tekkinud *s*-lõpulise tuletustüvega *ta*-tuletiste analoogial tüve ja liite piiri ähmastumise tõttu“ (Kasik 2013: 68). 66 korral leidus ka *t/duma*-lõpuli verbe, nt *esilinastuma, tõstatuma, tähtsustuma*.

Substantiividest jäi pärast läbivaatamist järele 1007 erinevat sõna, üle poole kõigist leidunud sõnadest moodustasid verbidest moodustatud *mine*- ja *ja*-liitelised tuletised (648). Suurem osa neist (461) oli tuletatud *mine*-sufiksiga, nt *diagnoosimine*, *garanteerimine*, *kainenemine*. *mine*-sufiksi abil moodustatakse substantiive kõikidest verbidest, *mine*-sufiks muudab ainult sõnaliiki, ei lisa tuletusalusele verbile semantilisi tunnuseid. *mine*-tuletistest kantakse sõnaraamatusse tavaliselt vaid idiomatiseerunud tähendusega nimetused (Kasik 1996: 17). EKSS'i on lisatud nt tuletatud substantiivid *elamine*, *mõtlemine*, *otsimine*.

ja-tuletisega sõnu, mis väljendatavad tegijat iga tegijat võimaldava tegevuse korral (Kasik 1996: 102) leidus korpuses 185 korda, nende seas *arendaja*, *rahastaja*, *hääletaja*.

Eesti keeles levinud isikuliitega *-lane* (Kasik 1996: 105) tuletatud sõnu leidus korpuses 25 korral.

Leidus ka modifitseeriva liitega sõnu, mille puhul liide ei muuda sõna alustüve sõnaliiki ega semantilist välja (Kasik 1996: 128). Kasiku kohaselt (Kasik 1996: 128) jagunevad modifitseerivad liited kahte suuremasse gruppi neist esimesed kannavad naissoo tunnust (*-nna*, *-tar*) ning teise rühma moodustavad vähendava-meelitava varjundiga deminutiivsufiksid (*-ke(ne)*, *-u*). Korpuses leidunud sõnadest 12 oli tuletatud kasutades *nna*-tuletist kasutades, näiteks *islandlanna* ning viiel korral kasutati *tar*-tuletist , nt *pariisitar*. Teise modifitseerivate tuletiste gruppi kuulusid *ke*- ja *kene*-liitega „konkreetsetest substantiividest moodustatud deminutiivtuletised“ (Kasik 1996: 130). Neid esines korpuses 58 erinevat sõna, nt *harjake*, *ämbrike*, *kruusike*.

Märkimisväärne osa (88) sõnu oli moodustatud *us*-liitega, mis ei lisa alussõnale semantilisi tunnuseid, vaid muudab üksnes sõnaliiki. (Kasik 1996: 125) *Us*-liitega tuletatud substantiive moodustatakse võrdselt nii verbidest kui ka adjektiividest, nt *tõestamatus*, *hälbelisus*, *föderatiivsus*, *eleegilisus*.

Leidus ka produktiivset, kuid vähe kasutatavat *ng*-liidet (Kasik 1996: 97), (*sämpling*, *tuuning*, *krüpteering*), päritolult vana *ur*-sufiksit (*metsur*, *hoidur*, *hakkur*), *ik*-sufiksit (*kilbik*, *karbik*), *la*-sufiksit (*pöörla*), *kas*-sufiksit (*põlvakas)*, *kond*-sufiksit (*autorkond*) ning *is*-sufiksit (*traageldis*). Võõrsõnadest võib eraldi välja tuua erinevad *iin*-, *ii*-*l*, *iid*-, *iiv*- jm lõpulisi sõnu, nt *indapamiid*, *leptiin*, *gastroskiis*. Veel esines *loog*-lõpulisi võõrsõnu (*allergoloog*, *gastroenteroloog*, *mütoloog*) ja *tsioon*-lõpulisi sõnu (*iteratsioon*, *inhibitsioon*, *identifikatsioon*). Ülejäänud sõnade hulgas leidus põhiliselt võõrsõnu ja toorlaene, näiteks *indie*, *draiver*, *vagotoomia*.

6.5. Korpuse ja EKSS'i sõnavara võrdluse kokkuvõte

Kuuendas peatükis võrreldi Koondkorpuse ja selle alamosa Tasakaalus korpuse sagedusloendit valikuliselt EKSS'i märksõnaloendiga. Valikulisuse all mõeldakse seda, et võrdlemiseks kasutati korpuse ja sõnaraamatu märksõnaloendi verbe ja lihtnimisõnu. Võrdluses koostatud failid (*võrdluse_loendid_verbid*, *võrdluse_loendid_lihtsubstantiivid*).

Materjali kitsendamine oli vajalik peamiselt kahel põhjusel:

1) magistritöö maht on piiratud, seega otsustati lähemalt uurida kahe sõnaklassi kattuvust,

2) võrreldavad andmekogud on erinevad – liitnimisõnade moodustamine on eesti keeles väga produktiivne, kuid sõnaraamatute maht on piiratud ja neid sinna ei lisata - seega otsustati ka töös piirduda vaid lihtnimisõnadega.

Mõlema võrreldava sõnaklassi jaoks koostati kolm loendit (sulgudes esitatud loendite lühendid on ühtlasi ka töö lisades esitatud töölehtede pealkirjad):

1) nii Koondkorpuses kui ka EKSS'is esinevad verbid/substantiivid (Kjah_Ejah),

2) verbid/substantiivid, mis esinevad EKSS'is, kuid puuduvad Koondkorpusest (Kei_Ejah),

3) verbid/substantiivid, mis esinevad Koondkorpuses, kuid puuduvad EKSS'ist (Kjah_Eei).

Tabel 24. **Võrdluse kokkuvõte**

Sõnaliik	**Vaatluse all**	**Kjah_Ejah**	**Kei_Ejah**	**Kjah_Eei (va 1x)**
Verb	kõik sõnad	6829	3342	634
Substantiiv	lihtsõnad	20038	5012	1007 (iga kümnes sõna)

Järgnevalt esitatakse kokkuvõte peatükis koostatud korpuse ja sõnaraamatu märksõnaloendi võrdlusest. Tabel 24 võtab kokku võrdluses koostatud failides sisalduvad arvandmed.

Esmalt ülevaade sõnadest, mis esinesid nii Koondkorpuses kui ka EKSS'is, koostatud loendis leidus 6829 verbi ja 20038 lihtnimisõna. Saadud sõnade loendid jaotati omakorda Koondkorpuse alamosa Tasakaalus korpuse kolme allosa (aja-, ilu- ja teaduskirjanduse) vahel. Seejärel moodustati sõnadest, mis esinesid vaid ühes allkorpuse osas, eraldi tekstiklasse iseloomustavad sõnade loendid. Näiteks ajakirjandust iseloomustavate sõnade loendisse kuulusid need sõnad, mida ilukirjanduse ja teaduskirjanduse tekstides ei leidunud.

Muidugi on küsitav mingi tekstiliigi iseloomuliku sõnavara määratlemise aluseks võtta vaid tingimus, et teistes sõnaklassides seda sõna ei esinenud. Võrdlusest tulid küll välja sõnad, mida tõesti võis pidada ühele tekstiklassile iseloomulikumaks, kuid samas leidus ka sõnu, mis võisid lihtsalt Tasakaalus korpuse tekstivalikust lähtuvalt jääda ühte või teise tekstižanrisse – teise tekstivaliku puhul oleksid võinud nad olemas olla ka praegu puuduvast tekstiklassist. Kuid samas pakkus taoline sõnavara tekstiklassidesse jaotamine kiiret ülevaadet korpuse kindlas allosas esinenud sõnavarast.

Tekstiklassideks jaotatud sõnavara analüüsimisel selgus, et puhtalt sellist sõnavara, mis iseloomustaks ainult ajakirjanduse keelekasutust, esines vähem kui ainult teadustekstides või ainult ilukirjanduses kasutatavat sõnavara. Samas huvitaval kombel kerkisid ajakirjanduse substantiividest esile spordivaldkonda kuuluvad sõnad.

Ilukirjanduse loendit lugedes tundusid seal esinevad verbid kohe „ilukirjanduslikud". Ilukirjandussõnavara on omanäoline, poeetilise varjundiga ja sisaldab rikkalikult kirjeldavaid ning ilustavaid sõnu.

Teaduskirjanduse puhul tuli esile, et tekstid sisaldasid nii teaduskeelt üldisemalt iseloomustavat sõnavara kui ka spetsiifilist oskussõnavara. Kuna teadustekstid sisaldavad põhiliselt ainult kindla valdkonna eriala termineid, siis sellest lähtuvalt on teaduskirjanduse sõnavara kõige enam sellest, millise teadusala tekste korpus sisaldas.

Võrreldes sõltuvust korpuse tekstide valikust teadustekstide, aja- ja ilukirjanduse sõnavara vahel, võib öelda, et ajakirjanduse sõnavara on sellest kõige vähem mõjutatud. Näiteks Tasakaalus korpuse ajakirjanduse osa sisaldab mitme ajalehe erinevaid numbreid ning ühes ajalehenumbris on käsitletud erinevatesse valdkondadesse kuuluvaid teemasid. Samas

muidugi mitte nii põhjalikult, kui teaduskirjanduses seda tehtud oleks või nii värvikalt ja sõnarohkelt kui ilukirjanduses. Ajalehetekstides ei esine autorikesksuse probleem nii teravalt kui teistes tekstiliikides – ühe ajalehenumbri kallal töötab rohkem kui üks ajakirjanik. Mingi sõna võib aga siiski „kunstlikult“ populaarseks saada siis, kui korpusesse satuvad korraga ühest järsult aktuaalseks muutunud juhtumist kirjutavad ajalehed. Samas on see probleem väiksem, võrreldes teadus- ja ilukirjanduse tekstidega, neist esimene on enim mõjutatud autorikesksusest, ühele inimesele iseloomulikust sõnavarast ja nagu juba mainitud, on korpuse teaduskirjanduse sõnavara seotud suuresti korpusesse valitud tekstide temaatikaga.

Teise võrdluses koostatud sõnade grupi moodustasid EKSS'is olevad ja Koondkorpusest puuduvad verbid ja substantiivid. EKSS'is esines korpusest puuduvaid verbe 3342 ja substantiive 5012. Selle sõnade grupi puhul analüüsiti, mis põhjusel neid sõnu korpuses ei esinenud. Analüüsimisel kasutati EKSS'is esitatud sõnade stiiliregistri ja kasutusala märgendeid. Nii näiteks ei esinenud korpuses sõna *ganoid*, mis on sõnaraamatus saanud kasutusalamärgendi *ZOOL* (zooloogia) ning tähistab *'ganoidsoomustega kala, vaapkala'* (EKSS). Sõnade esinemine on sõltuvuses korpuse tekstide valikuga, see puudutab enim just erialatermineid. Üldkeelde kuuluvatel sõnadel puudub range tekstiklassiline/teemaline piiritletus, eelnevas näites nimetatud spetsiifilisemad *ganoidi*-sarnased sõnad tavaliselt eesti ajakirjanduse või ilukirjanduse tekstides ei esine, kui just nendega mingi aktuaalne uudis pole seotud või kui nad mõnele ilukirjanduse autorile eriliselt silma pole jäänud.

Kolmandas võrdluse osas koostatud sõnade grupi moodustasid Koondkorpuses esinevad, kuid EKSS'ist puuduvad sõnad. Koostatud loendid vajasid ülevaatamist, kuna neis esines ka trüki- ja õigekirjavigadega sõnavorme ning morfoloogilise ühestamise vigu. Faili kontrollimisel vaadati läbi korpuses vähemalt kaks korda esinenud lemmad. Pärast kontrollimist jäi verbide faili alles 634 EKSS'ist puudunud erinevat lemmat (enne läbivaatamist esines korpuses 9050 lemmat (koos üks kord esinenud lemmadega)). Substantiivide puhul jäi vaatluse alla jäi iga kümnes korpuses esinenud lemma, pärast faili läbivaatamist jäi alles 1007 substantiivi (algselt esines korpuses 277989 sõna (koos üks kord esinenud lemmadega).

Loendite ülevaatamine oli mahukas töö, samas pärast korpusetekstide eripära (kirjavead, trükivead vms) kõrvale jätmist tulid välja sõnaraamatust puuduvad sõnad. Nende seas leidus arvukalt *mine-* ja *ja-* liitelisi tuletisi, mida lisatakse sõnaraamatusse vaid siis, kui nende tähendus on leksikaliseerunud. Leksikaliseerumine ei ole alati korrelatsioonis korpuses

esinemise absoluutsagedusega, kuid üldjuhul leksikaliseerunud sõnade esinemissagedus on kõrgem leksikaliseerumata sõnade korpuses esinemise sagedusest. (Bauer 2004: 13) Ühe sõnaraamatust puudunud sagedase *mine*-liitelise sõnana kerkis esile sõna *kohtumine*, mida esines Koondkorpuses koguni 64629 korda (Tasakaalus korpuses 2338 korda). Sõnaraamatus esineb *kohtumine* ainult liitsõna viimase komponendina, teda on kasutatud ka teiste sõnaartiklite definitsioonides, nt *tippkohtumine – 'riigijuhtide kohtumine'*, *töökohtumine 'tööasjus aset leidev kohtumine'*, *ärikohtumine- 'äriasjus aset leidev kohtumine'*, kuid omaette märksõnana pole teda sõnaraamatus esitatud. Tuli välja ka teisi sõnaraamatusse sobivaid sõnu, nt *destillaator, allergoloog.*

Kokkuvõtteks võib lugeda katse võrrelda valikuliselt korpuse sagedusloendit sõnaraamatu märksõnaloendiga kordaläinuks ning koostatud failid on esitatud töö lisades.

Kokkuvõte

Käesoleva magistritöö eesmärgiks oli eesti keele sagedusloendite koostamine Tasakaalus korpuse põhjal ja saadud sagedusloendite valikuline võrdlus „Eesti keele seletava sõnaraamatu“ märksõnaloendiga. Töö võrdluse osas kasutati lisaks 15-miljoni sõne suurusele Tasakaalus korpusele ka „Eesti keele koondkorpusest“ eraldatud aja-, ilu- ja teaduskirjanduse tekste, mida töös nimetati kokkuvõtvalt Koondkorpuseks.

Töö alguses anti ülevaade sagedussõnastikest: Eestis koostatud sõnastikest, sõnastike kasutamisest ja koostamise võimalustest (ptk 1). Eraldi peatükid käsitlesid nii sõnavara jagunemist (ptk 2), töös kasutatud keeleteaduslikke mõisteid (ptk 3) ning kasutatud materjali tutvustust (ptk 4).

Viiendasse peatükki koondati sagedusloendite koostamise alla käivad ülesanded alustades materjali ettevalmistamisest kuni tehtud sagedusloendite kirjelduseni. Kuigi sagedusloendite alusmaterjalina kasutati t3mestaga morfoloogiliselt ühestatud korpust, osutus sagedusloendite koostamisel vajalikuks korpusematerjali järelühestamine. Järelühestamise kahe etapi läbimise tulemusel saadi sagedusloendite koostamiseks vajalik alusmaterjal.

Eesti keel on rikka morfoloogiaga keel, seega sagedusloendid koostati eraldi nii lemmade kui sõnavormide põhjal. Kokku koostati nelja erinevat tüüpi sagedusloendid. Esiteks koostati sagedusloendid nii kogu Tasakaalus korpuse kui ka selle iga viie miljoni sõna suuruse allkorpuse põhjal, loendid sorteeriti nii alfabeetilise kui ka sageduse kahanemise alusel. Teiseks koostati loend, milles Tasakaalus korpuse sagedusele järgnesid sagedused tema kolmes allosas. Kolmandaks koostati kõigile kolmele tekstiklassile ühiste sõnade loend. Kõik eelnevalt nimetatud loendid on koostatud eraldi korpuses vähemalt kümme korda esinenud lemmade ja sõnavormide põhjal. Neljandas loendis esitati korpuses vähemalt kümme korda esinenud lemma järel kõik tema korpuses esinenud sõnavormid. Erinevad koostatud loendid võimaldavad kasutajal endal valida oma ülesande täitmiseks kõige sobilikuma loendi. Koostatud sagedusloendid leiavad kindlasti praktilist kasutust nii keeleteaduses, aga loodetavasti ka väljaspool seda, nt psühholoogias, pedagoogikas ja informaatikas.

Viimase osa tööst moodustas peatükk 6, milles võrreldi korpuses leiduvat materjali valikuliselt „Eesti keele seletava sõnaraamatu“ märksõnaloendiga. Võrdluse alla võeti verbid ja lihtsubstantiivid ning nendega koostati kolm suuremat loendit. Esimeses loendis esitati sõnad, mis esinesid nii Koondkorpuses kui ka EKSS’i märksõnaloendis, seega tegemist oli

reaalselt kasutatavate sõnadega, mis on oma koha leidnud ka sõnaraamatu märksõna loendis. Teise koostatud loendisse kuulusid EKSS'is esinevad sõnad, mida ei leidunud korpuses. Töös selgus, et üle poole korpusest puudunud sõnadest olid EKSS'i märksõnaloendis saanud kas kasutusala või stiiliregistri märgendi. Viimane koostatud loend koosnes korpuses esinenud sõnadest, mis puudusid sõnaraamatu märksõnaloendist. Kuna korpuse tekstid sisaldasid võõrkeelseid sõnu õigekirja- ja trükivigu ning korpuse töötlemisel tekkinud morfoloogilise ühestamise probleeme, siis osutus vajalikuks korpuseloendi käsitsi ülevaatamine. Puhastatud sõnade loendid sisaldasid nii produktiivseid tuletisi kui ka sõnaraamatusse sobivaid sõnu.

Sagedusloendid on vajalik keeleressurss, mida tuleks koostada vastavalt vajadusele. Üldkeele sagedusloend, mille eesmärgiks on pakkuda kõigile midagi, võib konkreetsema ülesande puhul jääda liialt üldiseks. Samas võimaldab korpuse põhjal koostatud sagedussõnastik kiiret pilguheitu just alusmaterjaliks olnud korpuse sõnavarale ning osutub seetõttu vajalikuks uurimismaterjaliks. Kokkuvõtteks võib öelda, et töös püstitatud eesmärgid said täidetud ning saadud tulemuste failid on esitatud töö lisades.

Kirjandus

Bauer, L. 2004. Adjectives, compounds and words. *Nordic Journal of English Studies 3(1)*, 7-22.

Carroll, J. B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index (SFI). *Computer Studies in the Humanities and Verbal Behavior, 3*, 61-65.

Čermák, F., & Křen, M. 2005. Large Corpora, Lexical Frequencies and Coverage of Texts. *Corpus Linguistics Conference 2005.* Birmingham .

Ehala jt = Ehala, M., Kerge, K., Lepajõe, K., & Sõrmus, K. 2010. *Kõrgkoolide üliõpilaste eesti keele oskuse tase: Uuringukokkuvõte.* Tartu : Tartu Ülikool.

EKK = Erelt, M., Erelt, T., & Ross, K. 2000. *Eesti keele käsiraamat.* Tallinn: Eesti Keele Sihtasutus.

EKG = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., et al. 1995. *Eesti keele grammatika I. Morfoloogia, sõnamoodustus.* Tallinn: Eesti TA Keele ja Kirjanduse Instituut .

Eslon, P., & Matsak, E. 2009. Eesti keele kasutusvariandid: korpusest tulenev käändevormide võrdlev analüüs. *Eesti Rakenduslingvistika Ühingu aastaraamat 5*, lk 79 - 110.

Hausenberg, A.-R. 2009. Kuhu lähed eesti keel? Sõnavara muutumine jätkub. *Keel ja Kirjandus 4*, 249-259.

Hlavácová, J. 2006. New Approach to Frequency Dictionaries - Czech Example. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, (lk 373-378). GENOA.

Hütt, M. 2012. *Kakskeelsete eelkooliealiste laste grammatilised oskused: kolme juhtumi kirjeldus.* Tartu: Tartu Ülikool.

Kaalep, H.-J., & Muischnek, K. 2002. *Eesti kirjakeele sagedussõnastik.* Tartu: Tartu Ülikooli Kirjastuse trükikoda.

Kaalep jt = Kaalep, H.-J., Muischnek, K., & Kirt, R. 2012. A trivial method for choosing the right lemma. *Human Language Technologies – The Baltic Perspective*, 82 - 89.

Kaalep, H.-J. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. *Computers and the Humanities*, lk 115 - 133.

Kaalep, H.-J. 1999. *Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös.* Tartu: Tartu Ülikooli Kirjastus.

Kaalep, H.-J.; Muischnek, K. 2004. Frequency Dictionary of Written Estonian of the 1990ies. *The First Baltic Conference Human Language Technologies: The Baltic Perspective. Commission of the Official Language at the Chancellery of the President of Latvia.*, (lk 57 - 60). Riga.

Kaalep, H.-J.; Vaino, T. 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. *Keel ja Kirjandus 1*, lk 30-38.

Kaasik jt = Kaasik, Ü., Soontak, J., Viilup, A., & Ääremaa, K. 1977. *Töid keelestatistika alalt. II, Keelestatistika.* Tartu: TRÜ trükikoda.

Kallas, J., & Tuulik, M. 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. *Eesti Rakenduslingvistika Ühingu aastaraamat 7*, lk 59-75.

Karlsson, F. 2002. *Üldkeeleteadus.* Tallinn: Eesti Keele Sihtasutus.

Kasik, R. 1996. *Eesti keele sõnatuletus.* Tartu: Tartu Ülikooli Kirjastuse trükikoda.

Kasik, R. 2003. Ajakirjanduskeel. rmt: *Eesti kirjakeele kasutusvaldkondade seisundi uuringud* (lk 118 - 148). Tallinn: TPÜ Kirjastus.

Kasik, R. 2011. *Stahli mantlipärijad.* Tartu: Tartu Ülikooli Kirjastus.

Kasik, R. 2013. *Komplekssete sõnade struktuur.* Tartu.

Kerge, K. 2008. *Vilunud keelekasutaja. C1-taseme eesti keele oskus.* Tallinn: Eesti Keele Sihtasutus.

Kilgarriff, A. 1997. Putting freguencies in the dictionary. *International Journal of Lexicography 10*, 135-155.

Krikman, A. 2004. *"Sai hea obaduse vastu obadust" : löömist ja peksmist märkivad väljendid eesti keeles* . Tartu: Eesti Kirjandusmuuseum.

Krikman, A. 1997. *Sissevaateid folkloori lühivormidesse.* Tartu: Tartu Ülikooli Kirjastuse trükikoda.

Landau, S. I. 2006. *Dictionaries. The Art and Craft of Lexicography. 2nd ed.* Cambridge: Cambridge University Press.

Langemets, M. 2010. *Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras.* Tallinn: Eesti Keele Sihtasutus.

Leech jt = Leech, G.; Rayson, P.; Wilson, A. 2001. *Word Frequencies in Written and Spoken English.* Longman, Pearson Education.

Li, H. 2010. Word frequency distribution for electronic learner's dictionaries. *ELexicography in the 21st century: new challenges, new applications* (lk 217-228). Louvain-la-Neuve : Presses univ. de Louvain.

Mäearu, S. 2011. eerima-tegusõna: üksi ja teistega. *Õiguskeel 2011*, 68 - 72.

McEnery, T., & Andrew, W. 1997. *Corpus Linguistics.* Manchester: Edinburgh University Press.

Metsmägi jt = Metsmägi, I., Sedrik, M., & Soosaar, S.-E. 2012. *Eesti etümoloogiasõnaraamat.* Tallinn: Eesti Keele Sihtasutus.

Nation, I. S. 2001. *Learning Vocabulary in Another Language.* Cambridge University Press.

Orlov jt = Orlov, J. K., Boroda, M. G., & Nadarejšvili, I. Š. 1982. *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Pajupuu jt = Pajupuu, H., Kerge, K., & Alp, P. 2009. *Eesti Rakenduslingvistika Ühingu aastaraamat 5*, lk 187-196.

Popescu, I.-I. 2009. *Word Frequency Studies.* Göttingen: Walter de Gruyter.

Rammo, S. 2010. *Eesti keele õpik täiskasvanud keeleõppijale.* Tartu: Tartu Ülikooli Raamatukogu.

Summers, D. 1996. Corpus Lexicography - The importance of representativeness in relation to frequency. *Longman Language Review 3*, 3-9.

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

Voll, P. 2009. Seletussõnaraamat valmis, varsti ka parandatud ja täiendatud. *Keel ja Kirjandus1*, 1 - 10.

KK = *Eesti keele koondkorpus*. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas http://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et

EKSS = *Eesti keele seletav sõnaraamat*. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas Eesti Keele Instituut: http://www.eki.ee/dict/ekss/ekss.html

TK = *Tasakaalus korpus*. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas: http://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=et

Viks, Ü., & Hein, I. 2001. *Seadusetekstide grammatiline sagedussõnastik.* Kasutamise kuupäev: 20. jaanuar 2013. a., allikas Eesti Keele Instituut: http://www.eki.ee/teemad/seadused_dic/sonastikud.pdf

wired = *wired.com*. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas http://www.wired.com/wiredscience/2012/01/hapax-legomena-and-zipfs-law

wfd = *Word frequency data*. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas http://www.wordfrequency.info/uses.asp

Sagedused. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas http:://www.cl.ut.ee/ressursid/sagedused1/

ESTMORF. http://www.filosoft.ee/html_morf_et/morfoutinfo.html#1. Kasutamise kuupäev: 20. jaanuar 2013. a., allikas http:://www.cl.ut.ee/ressursid/sagedused1/

Summary

Word frequency lists based on the "Balanced Corpus of Estonian" and selective comparison of corpora frequency lists with keywords from the „Explanatory Dictionary of Estonian“

The purpose of the current thesis is twofold: first, the creation of word frequency lists from the "Balanced Corpus of Estonian" and second, the selective comparison of corpora frequency lists with keywords from the „Explanatory Dictionary of Estonian“. In the comparison part, the "Balanced Corpus of Estonian" (15 million word tokens) along with a part of the "Reference Corpus of Estonian", containing newspaper, fiction and scientific texts was used.

First, an outline of frequency lists was provided, an outline of respective dictionaries composed in Estonian, the usage of these dictionaries and the possibilities to compose them (Chapter 1). The next chapters considered vocabulary distribution (Chapter 2), relevant linguistic terms (Chapter 3) and a summary of the material used in the current work.(Chapter 4).

The fifth chapter described the process of frequency lists creation: from the preparation of the material till to the description of complete frequency lists. As Estonian is a language with varied morphology, two types of frequency lists were made: the frequency list of lemmata and the frequency list of word forms.

The last part of the the current work, Chapter 6, selectively compared the material with key-words from the „Explanatory Dictionary of Estonian“. Verbs and monomorphemic nominals were investigated and the concurrency between the two lists was analysed.

To sum up, the goals of the current work were met and the results are in the form of files in Appendix. Frequency lists are necessary language resources, which should be composed according to the needs. Frequency lists of general language, which have the purpose to offer something to everyone, might be not sufficient for a concrete task. Nevertheless, they offer a quick insight into the underlying vocabulary of the corpus and therefore are a useful material to explore.

Lisad

CD sagedusloendid ja võrdluse failid

Tööle on lisatud CD, mis sisaldab magistritöö viiendas peatükis koostatud Tasakaalus korpusel põhinevaid sagedusloendeid ja kuuendas peatükis koostatud korpusesõnavara ja „Eesti keele seletava sõnaraamatu“ märksõnaloendi võrdlemisel saadud faile. CD sisaldab ka failides sisalduvat materjali kirjeldavat dokumenti.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Riin Kirt
(*autori nimi*)
(sünnikuupäev: 20.01.1988)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning „Eesti keele seletava sõnaraamatu“ märksõnaloendi võrdlus“,
(*lõputöö pealkiri*)

mille juhendaja on Kadri Muischnek,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus/Tallinnas/Narvas/Pärnus/Viljandis, Kirt (*20.05.2013*)