TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Kärt-Kristiin Jaagu

VERBI GRAMMATILISTE KATEGOORIATE ESINEMINE ERI TEKSTILIIKIDES

Bakalaureusetöö

Juhendaja PhD Kadri Muischnek

Tartu 2021

Sisukord

Sissejuhatus

Bakalaureusetöö teema on verbi grammatiliste kategooriate esinemine eri tekstiliikides. Töös võrreldakse verbi grammatiliste kategooriate sagedusi ilukirjanduse, tõlgitud ilukirjanduse, ajakirjanduse, populaarteadusliku kirjanduse, jututubade, seaduste ja suulise keelekasutuse tekstides.

Teemavalik lähtub teema aktuaalsusest ja isiklikust huvist. Käändsõna grammatiliste kategooriate ehk arvu ja käände sagedusloendid on tasakaalus korpuse põhjal koostatud (Kirt 2013), kuid verbi kategooriaid (isik, aeg, tegumood, kõneviis, kõneliik) on rohkem ning nende sagedusi tekstiliikides on vähe uuritud. Verbivormide sagedusest eesti keeles on varem kirjutanud Mare Kitsnik, kes vaatles B1- ja B2-taseme kirjalikku õppijakeelt (Kitsnik 2014). Samuti on teemat käsitlenud Juhan Tuldava ja Astrid Villup, kes on võrrelnud ilukirjandusproosa autorikõne sõnaliikide, sealhulgas ka verbide sagedusi erinevate allkeelte sõnaliikide sagedustega (Tuldava, Villup 1976).

Bakalaureusetöö eesmärk on välja selgitada, kas verbi grammatiliste kategooriate sageduse põhjal oleks võimalik eristada tekstiliike. Rakenduslik siht on leida verbi grammatilised kategooriad, mis võiksid toimida tunnustena tekstide klassifitseerimisel tekstiliikidesse. Teoreetiline siht on võrrelda tekstiliike nendes esinevate verbi grammatiliste kategooriate sageduste seisukohalt.

Eelnevast tulenevalt on bakalaureusetöös kolm peamist uurimisküsimust:

- kui sagedased on erinevad verbi grammatilised kategooriad eri tekstiliikides,
- kas ja kui palju erinevate tekstiliikide verbikategooriate sagedused erinevad,
- kas ainult verbikategooriate sageduse põhjal võiks teada saada, mis liiki tekstiga on tegu?

Uurimistöö autori püstitatud hüpoteesid on järgmised: tekstiliigiti on verbide morfoloogiliste kategooriate esinemissagedused oluliselt erinevad ning nende põhjal on võimalik määrata, millist liiki on tekst.

Bakalaureusetöös kasutatakse seitset korpust, millest igaühes on ligikaudu 100 000 sõna. Esindatud on järgmised tekstiliigid: ilukirjandus, tõlgitud ilukirjandus, ajakirjandus, populaarteaduslik kirjandus, seadused, jututubade keel ja suuline keel. Tegemist on kvantitatiivse uurimusega, mille analüüsiosas võrreldakse omavahel verbide morfoloogiliste kategooriate esinemissagedusi, mis on parema võrreldavuse huvides normaliseeritud.

Uurimus koosneb neljast suuremast peatükist. Esimeses peatükis tutvustatakse eesti keele verbi ja selle grammatilisi kategooriaid, kirjutatakse tekstiliigist ning antakse ülevaade varem uurimisalal tehtud tööst. Teises peatükis kirjeldatakse uuritavat materjali ja töö etappe, mis tulemusteni jõudmiseks läbiti. Kolmandas peatükis on analüüs, milles näidatakse verbi morfoloogiliste kategooriate esinemissagedusi eri tekstiliikides. Neljandas peatükis tuuakse esile tähtsamad tulemused ja järeldused.

Bakalaureusetöö autor avaldab tänu oma juhendaja Kadri Muischnekile kannatlikkuse, abivalmiduse vastutulelikkuse ja väärtuslike nõuannete eest.

1. Eesti keele verbist ja verbikategooriatest

Bakalaureusetöö esimene peatükk jaguneb kolmeks alapeatükiks. Esimeses osas tuuakse välja, mis on verb ja millised on verbide liigid. Teises alapeatükis esitatakse eesti keele verbi morfoloogilised kategooriad. Kolmandas osas kirjutatakse tekstiliikidest ja verbikategooriate esinemissageduse uurimisest.

1.1. Eesti keele verb ja verbide liigid

„Eesti keele käsiraamatu " (edaspidi ka EKK) järgi on verbid ehk pöörd- ehk tegusõnad pöördes, ajas, kõneviisis, tegumoes ja kõneliigis muutuvad sõnad. Need väljendavad tegevust ja on lauseliikmena öeldised. (EKK 2007: 172) Eesti keeles on liht-, liit-, ühend- ja väljendverbid. Lihtverbid on moodustatud ühest tüvest, näiteks *kõndima* ja *magama*. Liitverbid on liitsõnad ja koosnevad kahest tüvest, näiteks *kuritarvitama, ebaõnnestuma*. Ühendverbid on kokku pandud kahest sõnast, kusjuures üks on verb ning teine afiksaaladverb ehk abimäärsõna, mille abil saab muuta verbi tähendust. Ühendverbid on näiteks *ära jooksma, üles tõstma*. Väljendverbid moodustatakse verbist ja noomeni- ehk käändsõnavormist või pronoomeni- ehk asesõnavormist. Väljendverbid on näiteks *vanaks saama* ja *silma paistma*. (Samas 2007: 172–173) Verbivormid on liht- või liitvormid. Liitvormi puhul tuleb silmas pidada, et selle koostisesse kuulub alati mõni *olema*-verbi vorm või eitussõna, näiteks *ei*. (Samas 2007: 174)

Verbidel on viis morfoloogilist kategooriat, nendeks on aeg, kõneviis, kõneliik, pööre ja tegumood. Pöördsõna vormistikus kombineerub suurem osa kategooriatest teiste kategooriatega vaid osaliselt, seepärast on verbivormistiku koosseisu määratlemine nimisõnavormistikust raskem. (Viht, Habicht 2019: 104)

Morfoloogiliste kategooriate avaldumise põhjal jagunevad verbivormid finiitseteks ehk pöördelisteks ja infiniitseteks ehk käändelisteks (EKK 2007: 174). Finiitsed vormid

saavad olla lauses iseseisvaks öeldiseks, neis avalduvad morfoloogilised kategooriad. Infiniitsed vormid üldiselt lauses öeldiseks ei ole ning neis avalduvad morfoloogilised kategooriad puudulikult. Infiniitsetel vormidel ei ole kõiki tegusõnadele omaseid tunnuseid ning tähenduse ja süntaktilise käitumise poolest võivad infiniitsed verbivormid sarnaneda nimi-, omadus- ja määrsõnadega. (Samas: 261)

Pöördsõna finiitsetes lihtvormides saavad avalduda kõik verbi morfoloogiliste kategooriate liikmed, kuid pöörde ja tegumoe tunnused ei saa esineda samaaegselt. Finiitsetes liitvormides saavad esineda mõned ajakategooria liikmed ning eitava kõneliigi kujud. Infiniitsete lihtvormide alla võib lugeda kõik verbide infiniitsed vormid, töö autor toob need eraldi välja. Infiniitseid liitvorme ei ole eraldi liigitatud, kuid tarindeid, milles sisalduvad *olema*-verbi infiniitne vorm ja põhiverbi mineviku partitsiip, võib nendeks lugeda. (EKK 2007: 262–263)

Eesti keele verbide infiniitsed vormid on:

a) nimisõnalaadsed *da*-infinitiiv ehk *da*-tegevusnimi (*tantsida*), *vat*-infinitiiv ehk *vat*-tegevusnimi (*tantsivat*) ja supiin ehk *ma*-tegevusnimi (*tantsima*). *Da*-infinitiiv ei edasta grammatilisi tähendusi, kuid väljendab siiski tegevust ning võib olla ükskõik milline lauseliige, selle tunnusel on kolm varianti: *ta, da* ja *a*. *Vat*-infinitiiv on tekkinud verbi oleviku partitsiibi osastava käände vormist, võib lauses asendada *et*-sihitist ja kattub vormilt kaudse kõneviisi olevikuga. *Ma*-infinitiivi variandid on *ma-*, *mas-*, *mast-*, *maks-* ja *mata*-vorm, mis väljendavad vastavalt muule tegevusele järgnevat tegevust, samaaegset tegevust, eelnenud tegevust, tegevust kui otstarvet ja sooritamata tegevust;
b) omadussõnalaadsed partitsiibid ehk kesksõnad (*tantsiv, tantsitav, tantsinud, tantsitud*). Need aitavad väljendada tegevust kui seisundit või omadust. Eesti keeles on olemas oleviku kesksõna ja mineviku kesksõna, mõlemad esinevad isikulises ja umbisikulises tegumoes. Esimest kasutatakse, kui soovitakse kirjeldada tegevust, mis iseloomustab tegijat või tegevusobjekti olevikus, selle tunnused on *v* ja *tav*. Teist kasutatakse siis, kui omadus või seisund, mille kohta sõna käib, on tekkinud või tekitatud minevikus, tunnused on *nud, tud,* kusjuures *tud* võib esineda ka kujul *dud*;

c) määrsõnalaadne *des*-vorm ehk gerundiiv (*tantsides*). See avaldab mingi tegevusega samaaegselt toimuva ja seda iseloomustava tegevuse ning tunnusel on kolm varianti: *des*, *tes* ja *es*. (EKK 2007: 263–270)

1.2. Verbi grammatilised kategooriad

Pööre on isiku- ja arvukategooria, mis kordab osaliselt isiku ja arvu kohta käivat informatsiooni lauses. Pöördeid on eesti keeles kokku kuus, neist kolm on ainsuslikud ja kolm mitmuslikud. Pöörded saavad väljendada ka seda, kas kõnealune isik lauses on kindel, üldisikuline või umbisikuline. (Erelt 2017: 196–199) **Ainsuse 1. pöörde** tunnus on *n*. **Ainsuse 2. pöörde** tunnused on *d* ja *0*, kusjuures *0* esineb vaid käskivas kõneviisis. **Ainsuse 3. pöörde** tunnused on *b* ja *0*, tunnus *b* esineb ainult kindla kõneviisi olevikus. **Mitmuse 1. pöörde** tunnuse kujud on *me* ja *m*. **Mitmuse 2. pöörde** tunnused on *te* ja *0*, tunnust *0* kasutatakse vaid käskiva kõneviisi vormis, kus tunnusele eelneb kõneviisi tunnus. **Mitmuse 3. pöörde** tunnusel on kolm kuju: *vad*, mis esineb kindla kõneviisi olevikus; *d*, mis esineb kindla kõneviisi lihtminevikus ja tingiva kõneviisi vormides; ja *0*, mis esineb käskivas kõneviisis. (EKK 2007: 270–272)

Kõneliik näitab, kas lause on eitav või jaatav. Selle kategooria liikmed on eitus (negatiiv) ja jaatus (afirmatiiv). **Jaatust** märgitakse tavaliselt vaid küsimustele vastates, kuid mitte koos öeldisega. (Erelt 2017: 181) Jaataval kõneliigil ei ole tunnust **Eitava** kõneliigi abil on võimalik esitada lauses kirjeldatava tegevuse eitust, selleks kasutatakse eitusvormi. Eitusvormi juurde kuuluvad eitussõnad *ei* ja *ära*. (Viht, Habicht 2019: 116–117)

Tegumood kirjeldab nominaalsete lauseliikmete ja nende rollide vahekorda (Erelt 2017: 209). Eesti keeles on kaks tegumoodi: isikuline ehk personaal ja umbisikuline ehk impersonaal. **Isikuline tegumood** on kategooria liige, mis ei ole markeeritud, seega on isikuline tegumood tunnuseta. Isikulises tegumoes olevas lauses eksisteerib alus ehk tegija. **Umbisikuline tegumood** on markeeritud ja näitab, et lauses on tegijaks isik või isikud, keda ei väljendata, st näiteks ei saa teada seda, kas tegijaid on üks või mitu. Umbisikulise tegumoe tunnusel on seitse varianti, need on: *takse, dakse,*

akse, t, d, ta, da. (EKK 2007: 273) Umbisikulise tegumoe näitlikustamiseks võib tuua järgmised näited, milles tunnus on paksus kirjas: *joos**takse**, laul**dakse**, süü**akse**, elat**i**, mõel**d**i, joos**ta**vat, laul**da**vat.*

Aeg esitab verbiga märgitud lause situatsiooni ajalist suhet kõnehetke või teise situatsiooniga. Aeg võib väljendada minevikku, olevikku või tulevikku ning ajakategooriasse kuuluvad olevik (preesens), lihtminevik (imperfekt), täisminevik (perfekt) ja enneminevik (plu(skvam)perfekt), tingivas ja kaudses kõneviisis ka üldminevik (preteeritum). (Erelt 2017: 129, 136)

„Eesti keele käsiraamatu" järgi pole **olevikul** tänapäevases eesti keeles tunnust, teoses on kirjutatud järgnevalt: „Keeles on säilinud siiski mõningad morfeemid, mida võib pidada ajalooliseks oleviku tunnuseks: oleviku kesksõna tunnus *v*, ainsuse 3. pöörde tunnus *b*, mitmuse 3. pöörde tunnuses *vad* sisalduv *va* ning samuti umbisikulise tegumoe tunnusevariantides *takse, dakse* ja *akse* sisalduv *kse*" (EKK 2007: 276). Eesti keeles ei ole olemas morfoloogilist tulevikuvormi, selle asemel on tulevikku võimalik väljendada olevikuvormis verbiga, millele on vajadusel lisatud ajamäärused või tulevikutähendusega verbid (samas: 461). **Lihtminevik** väljendub jaatava ja eitava kõneliigi puhul erinevalt. Jaatavas kõneliigis on lihtmineviku tunnused *s, i*, ja minevikutüvi. Tunnus eitavas kõneliigis on vastavalt sellele, kas tegu on isikulise või umbisikulise tegumoega, *nud* või *tud, dud*. (Samas: 277) **Täisminevikus** lisandub põhiverbi mineviku kesksõnale sõna *olema* pöördeline olevikuvorm, näiteks *olen teinud*. **Enneminevikus** lisandub põhiverbi mineviku kesksõnale verbi *olema* pöördeline minevikuvorm, näiteks *olin teinud*. (Samas: 279) **Üldmineviku** väljendamiseks on lauses lisaks põhitegusõna mineviku kesksõnale verb *olema,* mis on tingivas, kaudses või möönvas kõneviisis, olevikuvormis ja isikulises tegumoes – näiteks *oleksin teinud*. Üldminevik esineb vaid tingivas, kaudses ja möönvas kõneviisis. (Samas: 279–280)

Kõneviis väljendab seda, kas lauses kirjeldatud sündmus on reaalne või irreaalne, deontilise või episteemilise modaalsusega, mis on lause suhtluseesmärk ja kas teade on lauses vahendatud või mitte. Eesti keeles on viis kõneviisi: kindel ehk indikatiiv, tingiv

ehk konditsionaal, käskiv ehk imperatiiv, möönev ehk jussiiv ja kaudne ehk kvotatiiv. (Erelt 2017: 160) **Kindlal kõneviisil** tunnust ei ole. **Tingiva kõneviisi** tunnuse kujud on *ksi* ja *ks*, olenevalt sellest, kas tunnusele järgneb pöördelõpp või mitte. **Käskiva kõneviisi** tunnuseid on viis ja need on *ge*, *ke*, *gu*, *ku* ja *0*, neist viimane esineb vaid ainsuse 2. pöördes. **Möönva kõneviisi** tunnused on *gu* ja *ku*, umbisikulises tegumoes esineb alati *gu*. **Kaudse kõneviisi** tunnuseks on *vat*. (EKK 2007: 281–283)

1.3. Allkeelest, tekstiliigist ja verbikategooriate sageduse uurimisest keelevariantides

Bakalaureusetöö keskendub verbi grammatiliste kategooriate esinemisele erinevates tekstiliikides. Sellest tulenevalt jaguneb töö kahte ossa: esitatakse sagedusandmed verbikategooriate kaupa ning võrreldakse töös kasutatud korpusi kui erinevate allkeelte esindajaid verbikategooriate sageduste alusel. Selleks, et teada, kas verbi grammatiliste kategooriate abil on võimalik tekstiliike eristada, on vaja välja selgitada, kas on verbikategooriaid, mis eristuvad piisavalt, et toimida tunnusena.

Kasutatakse seitset morfoloogiliselt märgendatud allkorpust: ilukirjandustekstide korpus, tõlgitud ilukirjanduse korpus, ajakirjandustekstide korpus, populaarteadusliku kirjanduse korpus, seadustekstide korpus, suulise keelekasutuse korpus ja jututubade tekstide korpus. Keelevariante võrreldakse sellistena, nagu nad korpustes esindatud on. Keelevariantide nimetamiseks on mitmeid võimalusi: allkeel, tekstiliik, žanr, tekstitüüp.

Tiit Hennoste on kirjutanud, et loomulik keel liigendub erinevateks variantideks ehk allkeelteks ning neid variante saab üksteisest eristada sõnavara või grammatika abil (Hennoste 2000a: 9). Allkeelte liigitamine on küll kohati problemaatiline ja allkeelte piirid võivad olla ebaselged, ka Hennoste ütleb, et autoritel on allkeelte rühmitamisel erinevaid arusaamu (samas: 38). Tiit Hennoste (samas: 10) jagab keelevariandid kahte rühma: kasutajakesksed variandid ehk murded ja kasutuskesksed ehk situatiivsed variandid. Situatiivse variandi all mõtleb Hennoste seda, et kindel kommunikatiivne situatsioon seostub kindlate keeleliste joontega. Samas tõdeb ta ka, et situatiivsete

variantide määratlemisel on palju terminoloogilist segadust, kasutatud on termineid stiil, žanr, register, funktsionaalstiil. (Samas: 15–18)

Reet Kasik kasutab Hennoste situatiivse variandi kohta terminit tekstiliik (ehk žanr) ning mõistab tekstiliikidena kasutusolukorrast sõltuvaid keelevariante, eristades neid tekstitüüpidest selle poolest, et tähelepanu pole niivõrd sisul, kuivõrd esitusviisil. Tekstiliikidena nimetab Kasik ilukirjanduskeelt, tarbekeelt ja argikeelt, kuid oluline on mainida, et tekstiliikidel on alaliigid, millel võivad samuti olla alaliigid, niisiis ei ole tekstiliike ainult kolm. (Kasik 2007: 34–35)

Bakalaureusetöö autor on otsustanud uuritavaid keelevariante nimetada tekstiliikideks, sest peab seda piisavalt täpseks definitsiooniks. Selle töö materjali moodustavad tekstid, mille saab siinkirjutaja meelest liigitada järgnevalt: ilukirjanduse ja G. Orwelli „1984“ tekstid kuuluvad ilukirjanduskeele alla; ajakirjanduse, Horisondi ja seaduste tekstid tarbekeele alla; suulise keele ja jututubade korpuse tekstid argikeele alla. Autor leiab, et tekstiliikideks nimetamine aitab säilitada lihtsust ning võimaldab hoida mahulist tasakaalu, sest kõigis kolmes kategoorias on üsna sarnane arv sõnu. See omakorda lubab uurimistöös saadud tulemusi pidada adekvaatseks.

Töös vaadeldakse verbikategooriate sagedusi eri tekstiliikides. Peatüki järgnevas osas antakse ülevaade sellest, mida on varem eesti keeles verbi ning selle grammatiliste kategooriate sageduste kohta uuritud.

Mare Kitsnik on tegelenud verbivormidega B1- ja B2-taseme kirjalikus õppijakeeles, tuginedes riiklike tasemeeksamite kirjutamisülesannetele, ja tema uurimuse tulemus on, et kõige sagedamini kasutati isikulise tegumoe, kindla kõneviisi oleviku või lihtmineviku, tingiva kõneviisi oleviku ning infinitiividena esinenud verbivorme. Samas nimetas Kitsnik oluliseks erinevuseks selle, et kindla kõneviisi kasutamine vähenes ning tingiva kõneviisi kasutamine kasvas, kui keeleoskus arenes. (Kitsnik 2014: 20–23)

Verbide sagedust eesti ilukirjandusproosa autorikõnes on käsitlenud Juhan Tuldava ja Astrid Villup. Nende uurimistöö tulemusel selgus, et verbid moodustavad mainitud tekstiliigis 22,5% sõnadest ning sõnaliikide sageduselt teisel kohal (lk 69). Samuti on autorid oma uurimuses esitanud tabeli, milles kajastatakse sõnaliikide sagedusi tekstis erinevate eesti keele allkeelte põhjal. Välja toodud allkeeled tabelis on järgmised: ilukirjandusproosa autorikõne (verbe 22,5%), „Tõde ja õigus I“ autorikõne (verbe 23,6%), spordiateated ajakirjanduses (22,8%), Nõukogude Sotsialistlike Vabariikide Liidu Telegraafiagentuuri ehk TASS-i sõnumid ajakirjanduses (19,6%) ning rahvalaul (24,9%). (Tuldava, Villup 1976: 69–70)

Sõnaliigi, sh verbi sageduse uurimisega tekstiliikides on tegeletud ka eesti keele sugulaskeeltes, võrdluseks võib näite tuua ungari keelest. Ilukirjanduses, seadusetekstides ja ajakirjandustekstides esinevaid sõnaliikide sagedusi selles on uurinud Ungari keeleteadlane Veronika Vincze. Ta toob artiklis „Domain Differences in the Distribution of Parts of Speech and Dependency Relations in Hungarian“ välja, et verbi sagedus 189 751–sõnalises ilukirjanduskorpuses oli 34 805, seadusetekstide 225 207-sõnalises korpuses 15 557 ja 190 406-sõnalises ajakirjandustekstide korpuses 20 751 (Vincze 2013: 317).

2. Materjal ja meetod

Peatükk „Materjal ja meetod“ koosneb kahest alapeatükist. Esimene alapeatükk annab ülevaate bakalaureusetöö aluseks olevatest materjalidest ning teine kirjeldab, milline oli töö käik lõpptulemuseni jõudmiseks.

2.1. Tekstikorpused

Bakalaureusetöö uurimismaterjaliks on tekstikorpuste tekstid. Nende töötlemiseks ning neist vajaliku informatsiooni saamiseks kasutati peamiselt Bitvise SSH Clienti terminali ja loodud *shell*'i skripte, mis on leitavad autori GitHubi kontolt (GitHub 2021).

Tekstikorpus on tekstide kogum, milles olevad tekstid iseloomustavad keele seisundit või varianti: tekstikorpusel on kindel eesmärk ja selle tekstid peaksid keelevarianti maksimaalselt esindama, seepärast on korpuste koostamisel oluline teha valik ühiskonnas levivate tekstide hulgast (Hennoste, Muischnek 2000: 185). Töös kasutatakse olemasolevaid korpusi, mis sisaldavad ja esindavad erinevaid tekstiliike. Valitud korpused on morfoloogiliselt ühestatud korpuse (TÜAU 2018) alamkorpused. Valitud korpustes on kokku umbes 697 000 sõna. Uurimiseks võeti tõlgitud ilukirjandust esindava G. Orwelli „1984“ tekstide (75 500 sõna), ilukirjanduse (104 000 sõna), ajakirjanduse (111 000 sõna), populaarteaduslikku kirjandust esindava Horisondi tekstide (98 000 sõna), seaduste (121 000 sõna), suulise keelekasutuse (100 000 sõna) ja jututubade (87 500 sõna) korpus, sest neis kõigis oli ligikaudu 100 000 sõna.

Morfoloogiliselt ühestatud korpus (edaspidi ka käsitsi märgendatud korpus) sisaldab tekste, mis on ühestatud käsitsi. Korpuse koostamine algas 1990. aastate teisel poolel, kuid suurem osa materjalist ühestati 2000. aastate alguses. Korpusesse kuuluvad tekstiliigid on järgmised: ilukirjandus, mh tõlgitud ilukirjandus, ajakirjandus, seadused, populaarteaduslik kirjandus, infotekstid, suulise keelekasutuse tekstid Hiljem on

korpusesse lisatud jututubade tekstid. (TÜAU 2018) Bakalaureusetöös on analüüsimiseks kasutatud ilukirjanduse, ajakirjanduse, seaduste, suulise keelekasutuse ja jututubade tekste. Tõlgitud ilukirjandust kui tekstiliiki esindab inglise keelest tõlgitud G. Orwelli „1984“ ja populaarteaduslikku kirjandust kui tekstiliiki esindab ajakiri Horisont. Infotekstid jäid valimist välja, sest selles osas olevaid sõnu oli väga vähe ning tulemused ei oleks olnud võrreldavad.

Siinkohal on vajalik selgitada suulise keelekasutuse korpuse olemust ja suulist keelt kui tekstiliiki, sest see on tõenäoliselt vähem tuntud kui sellised tekstiliigid nagu ilukirjandus, aimekirjandus, ajakirjandus ja seadusetekstid. Suulise keele korpusest on bakalaureusetöös kasutatud suhtlussituatsioonides salvestatud kõne transkribeeringuid, seega on korpuses spontaansed suhtlussituatsioonidest pärinevad suulised tekstid, mis esindavad erinevaid keelevariante ja on omadustelt varieeruvad (SEKK 2020). Tiit Hennoste on oma artiklis „Sissejuhatus suulisesse eesti keelde“ (2000b) toonud välja mõned suulisele keelele kui tekstiliigile omased nähtused. Suulises keeles võib esineda keelevääratusi, rohkelt partikleid, kirjalikust keelest enam lühenenud sõnu ja parasiitsõnu (Hennoste 2000b: 49–50). Kirjaliku keele lausete asemel on suulises kõnes lausungid, nende puhul on tähtis intonatsioon, mida pole võimalik kirjalikult edasi anda. Lausungitega kaasneb parandamine ja täpsustamine. (Samas: 53–54) Hennoste ütleb, et suulise keele sõnajärg on sageli erinev kirjaliku keele omast. Samuti lisab ta, et üks suulise kõne tekstivorme võib olla spontaanne suhtlus interneti jututubades. (Samas: 55–56) Seepärast loeti bakalaureusetöös jututubade korpuse tekstid suulist keelekasutust esindavateks.

Morfoloogiliselt ühestatud korpuse failide morfoloogilise analüüsi puhul on tegu käsitsi ühestatud märgendusega, edaspidi on see mainitud ka kui KÜM-kuju. See tähendab, et morfoloogiline analüüs on esitatud näiteks nii:

Ei ei+0 //_V_ aux neg //
vaidle vaidle+0 //_V_ main indic pres ps neg //
. . //_Z_ Fst //
(TÜAU 2018).

KÜM-kuju puhul on iga tekstisõna ja kirjavahemärk omaette real. Rea alguses on analüüsitav element ehk see sõnavorm, millena sõna tekstis esines. Sellele järgneb lemma ehk algvorm, millele on +-märgiga lisatud käände- või pöördelõpp, kui see eksisteerib, või 0, kui käände- või pöördelõppu ei ole võimalik tuvastada. Kahekordsete kaldkriipsude vahel on kirjas sõna morfoloogiline analüüs. Alakriipsudega eraldatud suurtähed näitavad sõnaliiki, tähistused on esitatud tabelis 1.

Tabel 1. **Sõnaliikide tähistused KÜM-kuju puhul** (TÜAU 2018).

A	adjektiiv ehk omadussõna
D	adverbiaal ehk määrus
G	genitiivatribuut ehk omastavaline täiend
I	interjektsioon ehk hüüdsõna
J	konjunktsioon ehk sidesõna
K	adpositsioon ehk kaassõna
N	numeraal ehk arvõna
P	pronoomen ehk asesõna
S	substantiiv ehk nimisõna
Z	Kirjavahemärk
T	tundmatu sõna
V	verb ehk tegusõna
X	adverb ehk määrsõna
Y	Lühend

Pärast sõnaliigi märgendit tulevad lühendid, mis näitavad, millistesse morfoloogilistesse kategooriatesse sõna kuulub. Verbid on märgendatud kolmel viisil*:* *_V_ main* on põhiverb, *_V_ aux* on abiverb ja *_V_ mod* on modaalverb.

Jututubade korpuses on internetijututubadest kogutud kirjalikud tekstid 2003. ja 2006. aastast. Korpuse materjal ei sisalda automaatselt tuvastatud võõrkeelseid osasid, samuti on eemaldatud hüperlingid ja meiliaadressid on muudetud tundmatuks. Korpuse koostajad on märkinud, et korpuse sisu on kui ülesmärgitud lavastus, milles igal osalejal

on laused. Kirjutatu on normeeritud kirjakeelest väga erinev, ent kasutajate kirjaviis on siiski säilitatud muutmatuna, lisatud on vaid tühikud kirjavahemärkide ümbrusesse. (TÜAU2 2019) Bakalaureusetöös on korpusesse kuuluvast 300 failist, mis sisaldavad kokku 7 miljonit sõna, kasutatud 14 faili. Korpuse 7 miljoni sõna hulgast jäi valimisse umbes 87 500 sõna – tulemuseni jõuti, kui failidest eemaldati lause alguse- ja lõpumärgendid <s> ja </s> ning kirjavahemärgid, selleks kasutati skripti *jututuba-sõnadearv.sh*.

Jututoa teksti tekstiliigina saab iseloomustada, tuginedes Sigrid Salla artiklile „Jututuba kui võrgusuhtlusvorm“ (2002). Jututubade vestlused on slängirohked ja lühendirikkad (Salla 2002: 133). Samuti võib tekstides esineda emotikone ehk graafilisi pildikesi ja onomatopoeetilisi sõnu, mis püüavad edasi anda emotsioone, helide ja muu auditiivse jäljendamist (samas: 138–140). Selleks, et mõne vestluskaaslase tähelepanu äratada, kasutatakse jututubades tihti kasutajanime või pseudonüümi mainimist (samas: 136–137). Bakalaureusetöö kirjutaja nõustub eespool nimetatud Tiit Hennoste seisukohaga, et jututubade tekstid sarnanevad suulise keelega. Sama meelt on ka Salla, kes kirjutab, et jututoas võivad teksti kirjutamise-saatmise ajal tekkida takerdused ning vestluses kasutatakse suulisele dialoogile sarnaselt lauseosa väljajätmist ja antakse tihti edasi vaid teksti tuumosa (2002: 142–143).

Jututubade korpuse failide algne märgendus oli Giellatekno (Kaalep jt 2018: 48) morfoloogiliste kategooriate märgendamise süsteemi järgi. Giellatekno kuju teisendati töö käigus KÜM-kujule. Giellatekno märgenduse näide on järgmine:

ma mina+0 //_P_ Sg Nom, //
seni seni+0 //_D_ //
sööma. söö+ma //_V_ Pers Sup Ill, //
) :) //_E_ //
(TÜAU2 2019).

Giellatekno märgenduse põhikonstruktsioon on sarnane eelnevalt kirjeldatud KÜM-kujul märgendusele. Sõnaliikide tähistele on jututubade korpuse puhul lisandunud emotikoni

märkiv E, lausepartiklit märkiv B ja sõnast eraldi trükitud osa märkiv Q (MÜJK 2020). KÜM-kujust erinevad on lühendid, mida kasutatakse morfoloogiliste kategooriate märkimiseks. KÜM-formaadis on märgendatud abi- ja modaalverbid, vastavalt on nende märgenditeks *aux* ja *mod*. Giellatekno märgenduses neid ei kasutata, mistõttu tuli autoril need käsitsi lisada.

2.2. Meetod

Alapeatükis 2.2. kirjeldatakse, millised olid tööülesanded ja mis etapid tuli lõpptulemuse saavutamiseks läbida.

2.2.1. Kasutatavate korpuste valimine

Esmalt valis autor koostöös bakalaureusetöö juhendajaga välja tekstikorpused. Määravaks sai nii korpuse suurus kui ka see, kas korpus on morfoloogiliselt märgendatud või mitte, sest märgendamine oleks olnud mahukas lisatöö. Eelistatud oli KÜM-kujul ehk käsitsi ühestatud märgendus, ent jututubade korpuse morfoloogiline märgendus oli algselt Giellatekno formaadis. Selleks, et kogu materjal oleks ühesuguse märgendusega ning edasine töö sujuks kiiremini, oli tähtis, et märgenduste formaadid oleksid ühtlustatud. Juhendaja soovitusel otsustati pärast materjaliga tutvumist, et edasi minnakse KÜM-kujuga, sest suurem osa ainesest oli märgendatud selle variandi kaudu ning lisaks oli Giellatekno märgenduse teisendamine KÜM-kujule tänu olemasolevale morf2morf.py skriptile mõistlikum. Kuna Giellatekno märgenduses puudusid abiverbide märgendid *olema*-abiverbide juurest ja modaalverbide märgendid sõnade *pidama, saama* ja *võima* juurest, siis oleks Giellatekno kujuga jätkamine toonud töö märgendite lisamise näol või ebatäpsema analüüsi, kui märgendeid poleks lisatud.

2.2.2. Morfoloogilise analüüsi teisendamine

Märgenduse teisendamiseks kasutati skripti *morf2morf.py*. Skript on mõeldud morfoloogilise märgenduse teisendamiseks Giellatekno kujult KÜM-kujule. Töö autor sai skripti juhendajalt ning skriptiga samas kaustas oli dokumentatsioon koos kasutatavate teisendusreeglitega ning sõnastik, mis võrdles Giellatekno ja KÜM-kuju märgendust.

Töötamiseks kasutati Bitvise SSH Clienti terminaliakent. Skriptiga samasse kausta lisati failid, mis vajasid märgenduse teisendamist. Terminalis olles sisenes autor jututubade korpuse faile sisaldavasse kataloogi ning sisestas käsureale käsu *python morf2morf.py AdiChat151103.yhenegt*. Käsurea esimene sõna tähistab, et tegu on programmeerimiskeeles Python kirjutatud skriptiga, *morf2morf.py* on skripti nimi ja *AdiChat151103.yhenegt* jututoa korpuse ühe faili nimi. Viimast muudeti vastavalt töödeldavale failile nii, et kõik korpuse failid saaksid läbi käidud ja õigele kujule teisendatud. Töödeldud failide nimedele lisandus pärast protsessi lõppu tähis *.mrf*, mis näitas, et tehtu on salvestatud uue failina.

Kõikide jututubade korpuse failide morfoloogilise märgenduse kuju teisendamine õnnestus. Teiste morfoloogilise ühestatud korpuste maht jäi umbes 100 000 sõna juurde ja jututubade korpusest sai KÜM-kujul märgenduse 141 350 elementi, pärast lausemärgendite ja kirjavahemärkide eemaldamist jäi järele 87 500 sõna. Bakalaureusetöö kirjutaja leidis, et sobiv kogus sõnu on olemas ning teisendatud märgendusega jututubade korpus on edaspidises töös ülejäänud korpustega võrreldav.

Teisendatud märgendusega failidest olid puudu modaalverbide *saama*, *võima* ja *pidama* märgendid ning nimetatud verbid olid analüüsitud põhiverbideks. Lisaks olid abiverbide asemel põhiverbidena esitatud ka verbi liitvormides esinevad *olema*-vormid. Selleks, et töö lõpptulemus saaks võimalikult täpne, ja võttes arvesse, et teiste kasutatud korpuste märgenduses olid olemas modaalverbid ja *olema*-vormis abiverbid, tuli kõik jututubade korpuse failid üle vaadata ja vajadusel käsitsi parandused sisse viia. Autor kasutas viimistlemiseks tekstiredaktorit Notepad ++. Korpusest otsiti kõiki *olema*, *saama*, *võima* ja *pidama* verbe ning vaadati neid ümbritsevat konteksti, et teha kindlaks, kas verb on lauses põhi-, abi- või modaalverbina. Etapi lõpus olid jututubade korpuse failides tehtud järgnevad muudatused:

- *olema* sai verbi liitvormis esinenuna abiverbi märgenduse, seega määrati *main* asemel märgenduseks *aux*. Näide sellest on järgmine:
  "<on>"
  "ole" L0 V *aux* indic pres ps3 sg ps af

"<uuendatud>"
"uuenda" Ltud V main partic past imps;

- *saama*, *võima* ja *pidama* said vajadusel – ehk siis, kui esinesid koos *da*-infinitiiviga (*saama* ja *võima*) või *ma*-infinitiiviga (*pidama*) – märgenduse *main* asemel märgenduseks *mod*. Näide sellest on järgmine:
  "<saavad>"
  "saa" Lvad V *mod* indic pres ps3 pl ps af
  "<teada>"
  "tead" La V main inf.

2.2.3. Verbianalüüsi kogumine

Pärast morfoloogilise märgenduse ühtlustamist tuli failid verbikategooriate kohta käiva informatsiooni kogumiseks ette valmistada. Selleks kirjutati kaks skripti: *kym-verbid.sh* ja *jututuba-verbid.sh*. Esimene oli mõeldud ilukirjanduse, tõlgitud ilukirjanduse, populaarteadusliku kirjanduse, ajakirjanduse, seadusetekstide ja suulise keele tekstide jaoks ning teine jututubade korpuse failide jaoks, kuid nende tööpõhimõte on sama.

Autor otsustas kahe skripti kasuks, sest suurem osa failidest oli laiendiga *.kym*, jututubade korpuse failid aga faililaiendiga *.mrf*. Morfoloogiliste analüüside esitamise kuju oli erinev, näiteks puudusid *.mrf*-failidest alakriipsud analüüsi osade vahel. Jõuti otsusele, et analüüsikujude ühesuguseks teisendamine oleks liialt ajamahukas. Skriptid olid mõeldud selleks, et sorteerida välja kõik elemendid, mis olid saanud verbianalüüsi; eemaldada kogu üleliigne informatsioon, sealhulgas tekstisõna ja lemma; jätta alles vaid morfoloogiline analüüs, mis annaks teadmisi verbi morfoloogiliste kategooriate kohta. Saadud tulemused salvestati eraldi failidesse.

Skriptide kasutamisel oli oluline, et skript oleks samas kaustas kui failid, mida see töötleb. Skript haaras kaustast kõik sobiva laiendiga failid ning tekitas uue faili, millele andis nimeks *verbid.1txt*. Number 1 lisati laiendisse selleks, et hiljem verbikategooriate sagedusi kogudes võtaks selleks etapiks koostatud skript kaasa vaid verbianalüüsi sisaldava faili ja väljastaks õiged tulemused. Hiljem muutis autor igas kaustas *verbid.1txt*

nime selliseks, et see kajastaks, millise alakorpuse verbide morfoloogilisi analüüse see sisaldab. Lõpuks tekkis seitse faili nimedega *jututoad-verbid.1txt, ilukirjandus-verbid.1txt, 1984-verbid.1txt, horisont-verbid.1txt, ajakirjandus-verbid.1txt, suuline-verbid.1txt, seadused-verbid.1txt.*

2.2.4. Verbi morfoloogiliste kategooriate sageduste leidmine

Järgmine ülesanne oli kokku lugeda kõikide verbi morfoloogiliste kategooriate liikmete esinemise arv igas tekstiliigis. Koostöös juhendajaga otsustati, et seda on kõige mõistlikum teha põhiverbide, abiverbide ja modaalverbide kaudu. Selleks uuriti, millist infot annavad *V main, V aux* ja *V mod* märgendid. Selgus, et käsitsi ühestatud märgenduse ehk KÜM-kuju puhul saab verbi lihtvormide kõiki kategooriaid puudutava kätte *V main* märgendiga sõnadest; liitvorme, niisiis eituse ja liitaegade teavet on näha *V main* ja *V aux* märgendiga sõnadest; kategooriad, mis modaalverbidega esinevad, kajastuvad *V mod* märgendiga sõnades.

Kirjutati skript *verbid.sh*, mis luges iga korpuse kaustas failist, mille laiend on *.1txt*, kokku verbimärgendi saanud sõnade arvu; põhiverbide, abiverbide ja modaalverbide arvud; kindlas, tingivas, käskivas ja kaudses kõneviisis verbid; olevikus, lihtminevikus, täisminevikus, enneminevikus, üldminevikus verbid, ainsuses ja mitmuses ning nii ainsuse kui mitmuse 1., 2. ja 3. pöördes olevad verbid; isikulises ja umbisikulises tegumoes verbid; jaatavas ja eitavas kõneliigis verbid, *da*-infinitiivid, supiinid, partitsiibid ja *des*-vormid. Skripti kasutamise tulemusel loodi *verbiinfo.txt* failid, mis sisaldasid kategooriate nimetusi ja esinemiste arve korpuses.

Eitavas kõneliigis tegusõnade arv koguti kahes osas ning kasutati lisaskripti. Esmalt loeti *verbid.sh* skriptiga kokku kõik *V aux neg* analüüsi saanud verbid. Tulemuses ei kajastunud ühe sõnaga väljendatud liitaegade eitus ehk sõna *pole* erinevad vormid. Selleks, et ka kõik *pole* vormid kätte saada, kirjutati kaks skripti, mida kasutati korpuse algfailide peal (laiendiga *.kym*) – jututubade korpuse puhul KÜM-kujule teisendatud failidel (laiendiga *.mrf*) –, sest need skriptid ei lugenud sõnu kokku mitte analüüsi, vaid tekstisõna põhjal. Kahe skripti kasutamise vajaduse tingis algfailide erinev

märgendamissüsteemi kuju. Skripti *pole-jututoad.sh* kasutati jututubade korpuse failidel, skripti *pole.sh* ülejäänud korpuste failidel. Skriptide tööpõhimõte oli sama: loeti kokku kõik sõnad, mis algasid ühendiga *pol* ja olid saanud põhiverbi analüüsi *V main*. Skripti väljastatud tulemus, *pole* vormide arv lisati käsitsi nendele *verbiinfo.txt* failidele, mis tekkisid eelnevas etapis, kui kasutati skripti *verbid.sh* ning seejärel oli võimalik kokku arvutada, mitu korda eitust esines.

Mineviku partitsiipide loendamiseks kirjutati lisaskriptid *minevikupartitsiip.sh* ja *minevikupartitsiip-jututoad.sh*, mis kogusid kokku kõik adjektiiv positiivi analüüsi saanud ja ühenditega *nud, tud* või *dud* lõppevad sõnad, väljastasid nende koguarvu ja kirjutasid tulemuse *mp.txt* failidesse. Hiljem kirjutas autor kõikide failide nimedesse tähise, mis näitab, millise korpuse andmed failis sisalduvad. Saadu lisati samuti käsitsi eelnevalt *verbid.sh* skriptiga loodud failidele. Jututubade korpusele mõeldud mineviku partitsiipide loendamiseks mõeldud skript töötab failidel, mis on laiendiga *mrf.inforem*. Töö käigus selgus, et *.mrf* laiendiga failides ei ole mineviku partitsiipe ka omadussõnadena märgendatud, vaid need on samasuguse märgendiga nagu verbi liitaja osad. Näiteks oli sõna *kadunud* samasuguse morfoloogilise märgendusega kui sõna *näinud* ühendis *pole näinud* – mõlema morfoloogiline märgendus oli *V main partic past ps*. Teistes, *.inforem* laiendiga failides olid mineviku partitsiibid märgendatud kui adjektiiv positiivid. Samuti tuleb siinkohal öelda, et *.inforem* failidest puudusid kaks jututubade korpuse faili, seega on skriptiga läbi vaadatud korpuse suurus veidi väiksem kui ülejäänud kategooriates.

Oluline on märkida, et isiku kategooria statistika on koostatud verbi jaatavate vormide põhjal, sest eitavates vormides oli kategooria märgendatud vaid käskivas kõneviisis. Lisaks puudub tööst oleviku partitsiipide analüüs, sest oleviku partitsiibid olid märgitud omadussõnadena ning nende eraldamine ei olnud võimalik. Näiteks olid samasuguse analüüsiga sõnad *erilist* ja *kompromiteerivat*, mis olid märgendatud kui adjektiiv positiiv singular partitiiv: *//_A_ pos sg part //*. Korpustes ei olnud märgendatud möönvat kõneviisi ehk jussiivi ja *vat*-infinitiive, niisiis ei ole uurimuses nende sagedusi kajastatud.

2.2.5. Sageduste normaliseerimine

Eelnevalt kirjeldatud töö tulemusel saadud informatsioon normaliseeriti, et oleks võimalik võrrelda, milline on verbi grammatiliste kategooriate esinemissagedus tekstiliigiti. Normaliseerimine tähendab skripti abil kogutud sageduste ümberarvutamist korpuse mahtu arvestades. Kogutud sagedused on nähtuse absoluutsagedused korpuses, need tuleb jagada korpuse sõnede koguarvuga ja korrutada tulemus valitud baasiga („Kuidas mõista andmestunud maailma?“ 2020: terminisõnastik). Normaliseerimine oli vajalik, sest kuigi korpused olid peaaegu sarnases mahus, umbes 100 000 sõna, peeti mõne korpuse sõnade arvu siiski liialt erinevaks, et saada täpsed andmed: 75 500-sõnalise korpuse võrdlemine 121 000-sõnalise korpusega ei oleks olnud mõistlik.

Esmalt normaliseeriti verbide osakaal korpuse sõnade koguarvust, et võrrelda korpusi omavahel, keskendudes verbide rohkusele. Selleks võeti igast korpusest verbianalüüsi saanud sõnade koguarv, mis jagati korpuse sõnade koguarvuga ning korrutati 100 000-ga.

Järgnevalt normaliseeriti põhiverbide, abiverbide ja modaalverbide osakaal, et ka nende esinemissagedust oleks võimalik uurida. Põhiverbid, abiverbid ja modaalverbid otsustati normaliseerida verbide koguarvu kasutades ning baasiks võeti 20 000, sest ligikaudu nii palju verbe oli igas korpuses. Seega tuli iga korpuse põhi-, abi- ja modaalverbide arvud jagada verbide koguarvuga korpuses ja korrutada 20 000-ga.

Viimase sammuna normaliseeriti kõikide korpuste verbi grammatiliste kategooriate esinemise arvud, et saada kätte kategooria verbide osakaal korpuse verbide arvust. Kui seda poleks tehtud, ei oleks korpused olnud võrreldavad: nende korpuste puhul, milles oli rohkem verbe, oleks võinud tulla tulemuseks see, et on ka rohkem näiteks olevikus või umbisikulises tegumoes verbe. Ka kolmandal normaliseerimisel otsustati verbide baasarvuks võtta 20 000. Niisiis jagati iga kategooria kohta esinenud verbide arv põhiverbide arvuga korpuses ja korrutati 20 000-ga. Normaliseeriti põhiverbide arvu järgi, sest verbi liitvormide puhul loeti kategooriate sagedust ainult ühest liitvormi osast.

Saadud tulemuste põhjal oli võimalik omavahel võrrelda kõikide korpuste verbikategooriaid.

2.2.6. Erinevate verbide hulk korpustes

Töö käigus otsustati vaadelda ka seda, kui palju on igas korpuses erinevaid põhiverbe. Selle tarbeks kirjutati taas kaks skripti, *tegusonad-jututoad.sh* ja *tegusonad.sh*, esimene jututubade korpuse jaoks ning teine ülejäänud korpustele. Jututubade korpusele (laiendiga *.mrf*) mõeldud skript võttis korpusest välja põhiverbid, eemaldas üleliigse informatsiooni, jättes järele vaid tegusõna lemma, lisas lemmale *ma*-tunnuse, sorteeris sõnad ning kirjutas eraldi faili kõik unikaalsed tulemused. Faili lõppu lisati arv, mis näitas, mitu sõna kokku tuli. Teistele korpustele (laiendiga *.kym*) mõeldud skripti tööpõhimõte on samasugune – erandiks on vaid see, et kuna ilukirjanduse, tõlgitud ilukirjanduse, ajakirjanduse, populaarteadusliku kirjanduse, seadusetekstide ja suulise keelekasutuse korpuste failides olid täpitähtede asemel HTML-koodi olemid, oli oluline need asendada. Kõikidest korpustest saadud erinevate verbide arvud normaliseeriti 100 000 sõna peale, et saavutada võrreldavad tulemused. Selle tööetapi tulemuste analüüs on alapeatükis 3.1.2.

2.2.7. Verbikategooriate sageduse ühtlus failiti

Pärast verbikategooriate kohta saadud informatsiooni normaliseerimist ja analüüsimist otsustati kontrollida, kas verbi grammatiliste kategooriate esinemine võib tekstiliigi failides erineda. Selleks valiti välja kolm verbikategooriate liiget, mis esinesid sageli, kuid mille sagedus kõikus tekstiliike võrreldes piisavalt, et oleks võimalik vaadelda, kas tegu on tekstiliigile omase tunnusega. Tulemusi on kirjeldatud alapeatükis 3.4.

Koostati skriptid *failiti-jututuba.sh* ja *failiti.sh*, esimene oli mõeldud jututubade korpusele ja teine ülejäänud korpustele. Kahe skripti kasuks otsustati taas seepärast, et kasutama pidi algfaile (laiendiga *.kym*), jututubade korpuse puhul KÜM-kujule teisendatud faile (laiendiga *.mrf*), ning morfoloogilise märgenduse esitamise formaadid erinesid. Lisaks oli jututubade korpuse sõnade arv eelnevalt *jututuba-sonadearv.sh* skripti abil kokku loetud, seega sai mainitud skripti kasutada põhjana, et lugeda kokku failides olev sõnade arv. Aja kokkuhoiu tarbeks ei muudetud jututubade korpuse esitusformaati, vaid kasutati tükke

skriptidest *failiti.sh* ja *jututuba-sõnadearv.sh*. Valminud skriptide tööpõhimõte kattus. Skriptid tõsteti samasse kausta iga korpuse failidega ning käivitati, mille tulemusel koguti igast failist andmed faili suuruse, põhiverbide arvu, umbisikulise tegumoe, *da*-infinitiivi ja oleviku esinemise kohta ning kirjutati need faili nimega *failiti.txt*. Hiljem lisas autor iga *failiti.txt* faili nimesse, millise korpuse andmeid see sisaldab. Lõpuks tekkis seitse faili: *1984-failiti.txt, aja-failiti.txt, hor-failiti.txt, ilu-failiti.txt, jututoad-failiti.txt, sea-failiti.txt* ja *suu-failiti.txt.*

Järgmisena vaadati, kas korpustes olevad failid on korpuse lõikes võrreldavate suurustega. Selgus, et inglise keelest tõlgitud ilukirjanduse korpuse faili suurus oli umbes 94 000 sõna, populaarteadusliku kirjanduse ja ilukirjanduse korpuse failide suurusjärk oli 2000 sõna, seadusetekstide ja suulise keelekasutuse korpuse failid koosnesid ligikaudu 10 000 sõnast. Failide mahud varieerusid suuremal määral jututubade ja ajakirjandustekstide korpuses: jututubade korpuse suurim fail koosnes 15 900 ning väikseim 800 elemendist, ajakirjanduskorpuse suurimas failis oli 26 000 ja väikseimas 2500 elementi.

Selleks, et mõista, kas verbikategooriate esinemissagedus on omane tekstiliigile või vaid konkreetsele failile, vaadeldi, kas valitud kolme kategooria – umbisikulise tegumoe, *da*-infinitiivi ja oleviku – sagedused on failides sarnases suurusjärgus. Tõlgitud ilukirjanduse, populaarteadusliku kirjanduse, ilukirjanduse, seadusetekstide ja suulise kõne tekstide korpuse puhul ei olnud normaliseerimine vajalik, kuid jututubade ja ajakirjandustekstide puhul otsustati valida korpuses olevate failide suuruste aritmeetiline keskmine ning seda kasutades verbikategooriate sagedused normaliseerida.

Ajakirjanduskorpuses oli 11 faili ning keskmiseks failisuuruseks tuli 12 000 elementi. Jututubade korpuse 14 faili keskmine suurus oli 6250 elementi. Normaliseerimiseks valiti ajakirjanduskorpuse puhul baasiks 12 000 ja jututubade puhul 6250. Iga faili verbikategooria esinemissagedus jagati failis olnud põhiverbide arvuga ja korrutati valitud baasiga, et võrrelda, kuivõrd sarnane on verbikategooria esinemissagedus igas failis.

3. Analüüs

Analüüsi esimeses osas kirjeldatakse korpuste verbide üldhulkade, põhi-, abi- ja modaalverbide ning erinevate verbide hulga normaliseerimisel saadud tulemusi. Teises osas kirjutatakse kategooriate esinemissagedusest tekstiliikides. Kolmandas osas on välja toodud verbi infiniitsete vormide esinemine korpustes. Neljandas osas esitatakse umbisikulise tegumoe, oleviku ja *da*-infinitiivi sageduste esinemiste arvud korpuste kõikides failides.

3.1. Verbide üldhulk

Autor normaliseeris kõikide korpuste verbide üldhulgad 100 000-sõnalisele korpusele, et välja selgitada, millises korpuses on verbide osakaal tekstis kõige suurem. Verbide üldhulkade normaliseerimisel saavutatu on tabelis 2. Tabelid 2–13 on koostatud samal põhimõttel. Tabeli päises on korpuste nimetused: tõlgitud ilukirjandus esindab inglise keelest tõlgitud ilukirjanduse ehk G. Orwelli „1984“ tekstide korpust, ajakirjandus ajakirjandustekstide korpust, populaarteaduslik kirjandus ajakirjast Horisont pärinevate populaarteaduslike tekstide korpust, ilukirjandus ilukirjandustekstide korpust, jututoad jututubade tekstide korpust, seadused seadusetekstide korpust ja suuline esindab suulise keelekasutuse tekstide korpust. Viimase tulba pealkiri „keskmine“ näitab, et viimane tulp sisaldab andmeid iga nähtuse keskmise esinemissageduse kohta kõikides korpustes. Tabeli vasakus servas on kirjas vaadeldav nähtus.

Tabel 2. **Verbide üldhulkade normaliseeritud sagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
verbide sagedus	24095	18441	17738	24028	10016	12354	18531	17687

Normaliseerimise põhjal võib öelda, et suurim verbide esinemissagedus oli inglise keelest tõlgitud ilukirjanduse korpuses (24 095), sellele järgnesid kahanevas järjekorras ilukirjanduse (24 028), suulise keele (18 531), ajakirjanduse (18 441), populaarteadusliku kirjanduse (17 738) ja seaduste tekstid (12 354) ning kõige väiksem oli sagedus jututubade tekstides (10 016).

Ilukirjandustekstide ja tõlgitud ilukirjanduse teksti andmetest tuleb töö autori meelest hästi välja see, et ilukirjandust sisaldavates korpustes on verbide osakaal väga sarnane. Ka ajakirjanduses ja populaarteadusliku kirjanduse tekstides, mis pärinevad ajakirjast Horisont, on verbide esinemissagedus peaaegu võrdne.

Suulise keele tekstides on pöördsõnade osakaal sarnane ajakirjandustekstidele, mis on üsna huvitav. Selle järgi võiks öelda, et suuline keel ja ajakirjanduskeel on sarnased. Tulemust võib mõjutada see, et ajakirjanduses tehakse suulisi intervjuusid, millest artiklitesse lauseid kirjutatakse.

Seaduste tekstides on verbe eelnevalt nimetatud korpustega võrreldes pigem vähe. Vahe on tõenäoliselt tingitud sellest, et seadused on sõnastatud nominaalstiili kasutades, mistõttu on tekstides palju nimisõnu ja vähe tegusõnu, seaduskeel võib olla kantseliitlik (Narits 2002).

Teiste korpustega võrreldes võib jututubade tekstide väikese verbide sageduse põhjustada soov informatsiooni kiirelt ja konkreetselt edastada ning kui verb ei ole lauses

ilmtingimata vajalik, jäetakse see kirjutamata, et hoida kokku aega ja moodustada lühem tekstijupp. Seega on jututubades tihti kasutusel konstruktsioon, mis sarnaneb lausele *„ma korra eemale“* – puudub tegusõna, kuid põhiidee on vestluskaaslastele siiski mõistetav ning lause täidab info edastamise eesmärki. Verbide üldhulga sagedusandmed ei toeta väidet, et suulise keele ja jututubade tekstid on sarnase keelekasutusega.

3.1.1. Põhiverbid, abiverbid, modaalverbid

Põhiverbide, abiverbide ja modaalverbide hulgad normaliseeriti 20 000-sõnalisele korpusele. Selle abil on võimalik teada saada, millises korpuses on nimetatud nähtuste hulk suurim. Normaliseerimisel saavutatu on tabelis 3.

Tabel 3. **Põhi-, abi- ja modaalverbide esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
põhiverbid	16005	16664	16700	16632	16483	15070	16292	16303
abiverbid	3140	2397	2338	2636	2862	3249	2763	2738
modaalverbid	855	939	962	732	655	1682	945	959

Põhiverbide esinemissagedus oli suurim populaarteadusliku kirjanduse tekstides (16 700) ja väikseim seadusetekstides (15 070). Nähtuse esinemissageduse põhjal koostatud kahanev korpuste järjestus on järgnev: populaarteadusliku kirjanduse korpus (16 700), ajakirjanduskorpus (16 664), ilukirjanduskorpus (16 632), jututubade tekstide korpus (16 483), suulise kõne tekstide korpus (16 292), tõlgitud ilukirjanduse korpus (16 005) ja seadusetekstide korpus (15 070). Põhiverbide keskmine esinemissagedus kõikide korpuste peale oli 16 303, mis näitab, et esinemissageduselt kõige lähedasem on suulise kõne tekstikorpus.

Abiverbe esines kõige sagedamini seadusetekstide korpuses (3249) ning kõige harvemini populaarteadusliku kirjanduse tekstides (2338). Kahanev sagedusjärjestus on järgmine:

seadusetekstid (3249), tõlgitud ilukirjandus (3140), jututubade tekstid (2862), suulise kõne tekstid (2763), ilukirjandustekstid (2636), ajakirjandus (2397), populaarteadusliku kirjanduse korpus (2338). Keskmine esinemissagedus oli 2738, kõige rohkem sarnaneb näitajaga abiverbide esinemissagedus suulise kõne tekstide korpuses.

Modaalverbe oli enim seadusetekstides (1682) ja kõige vähem jututubade tekstides (655). Modaalverbide kahaneva esinemissageduse järgi loodud korpuste järjestus on järgmine: seadusetekstide korpus (1682), populaarteaduslik kirjandus (962), suulise kõne tekstid (945), ajakirjandustekstid (939), tõlgitud ilukirjandus (855), ilukirjandus (732) ja jututubade korpus (655). Nähtuse esinemise keskmine sagedus oli 959, niisiis on modaalverbide esinemissageduselt lähim populaarteadusliku kirjanduse tekstide korpus.

3.1.2 Erinevad verbid

Lisaks eri sõnaliikide osakaalule saab tekste ja tekstiliike iseloomustada ka nende sõnavara rikkuse kaudu. Tabel 4 sisaldab andmeid selle kohta, kui palju erinevaid verbe vaadeldud tekstiliikides esineb ja kui suur oli erinevate verbide sagedus igas korpuses.

Tabel 4. **Erinevate verbide sagedus korpuses.**

	tõlgitud ilukirjandus	ajakirjandus	populaarteaduslik kirjandus	ilukirjandus	jututoad	seadused	suuline
erinevaid verbe	1657	1247	1236	1702	787	422	665

Erinevate verbide sagedus oli kõige suurem ilukirjandustekstides (1702), kõige väiksem seadusetekstides (422). Sagedusnäitaja kahanemise järgi loodud järjestus korpustest on järgmine: ilukirjandustekstide korpuse (1702), tõlgitud ilukirjandus (1657), ajakirjandus (1247), populaarteaduslik kirjandus (1236), jututubade tekstide korpus (787), suulise keele tekstid (665) ja seadusetekstid (422).

3.2. Kategooriad

Verbide hulgad kõikides kategooriates normaliseeriti samuti 20 000-sõnalisele korpusele ja selleks kasutati põhiverbide arvu. Tulemusi vaadeldes saab välja selgitada, milline oli iga kategooria liikme esinemissagedus korpustes.

3.2.1. Pöörde kategooria

Pöörde ehk isiku ja arvu kategoorias olevad liikmed on ainsus, mitmus, ainsuse 1., 2. ja 3. pööre ning mitmuse 1., 2. ja 3. pööre. Kõikide liikmete esinemissagedus igas korpuses on tabelis 5.

Tabel 5. **Pöörde kategooria liikmete esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
ainsus	11393	10770	10037	11843	12222	9534	11956	11120
mitmus	2785	2890	3678	2227	1714	1229	3267	2637
ainsuse 1. pööre	482	446	332	1996	2827	2	1889	1093
ainsuse 2. pööre	353	254	101	690	3708	4	1257	707
ainsuse 3. pööre	10558	10070	9604	9156	5621	9529	8810	9315
mitmuse 1. pööre	249	286	551	426	321	0	665	375
mitmuse 2. pööre	626	121	81	206	473	0	992	349
mitmuse 3. pööre	1910	2484	3045	1595	919	1229	1610	1913

Ainsust oli kõige sagedamini kasutatud jututubade korpuse tekstides (12 222). Kõige vähem esines ainsuse kasutust seadusetekstides (9534). Kahaneva esinemissageduse järgi loodud järjestus korpustest on järgmine: jututubade tekstid (12 222), suulise kõne tekstid (11 956), ilukirjanduse korpus (11 843), tõlgitud ilukirjandus (11 393), ajakirjandustekstid (10 770), populaarteaduslik kirjandus (10 037),

seadusetekstid (9534). Nähtuse esinemissagedus korpustes oli keskmiselt 11 120, see on kõige lähedasem tõlgitud ilukirjanduse sagedusnäitajale.

Mitmuse esinemissagedus oli suurim populaarteadusliku kirjanduse korpuses (3678) ja kõige väiksem seadusetekstide korpuses (1229). Vastavalt kahanevale mitmuse esinemissagedusele on korpuste järjestus selline: populaarteaduslik kirjandus (3678), suulise kõne tekstid (3267), ajakirjandus (2890), tõlgitud ilukirjandus (2785), ilukirjandus (2227), jututubade korpus (1714), seadusetekstid (1229). Mitmuse keskmine esinemissagedus oli 2637, sarnases suurusjärgus on nähtuse sagedusnäitaja tõlgitud ilukirjanduse tekstides.

Ainsuse 1. pööret esines kõige rohkem jututubade korpuses (2827) ja kõige vähem seadusetekstides (2). Kahaneva sagedusnäitaja põhjal on korpuste järjestus järgmine: jututubade tekstid (2827), ilukirjandus (1996), suulise kõne tekstid (1889), tõlgitud ilukirjandus (482), ajakirjandus (446), populaarteadusliku kirjanduse korpus (332), seadusetekstid (2). Nähtuse keskmine sagedus oli 1093, kõige lähedasem sellele on tõlgitud ilukirjanduse tekstide sagedusnäitaja.

Ainsuse 2. pööret oli enim jututubade korpuses (3708) ja kõige vähem esines seda seadusetekstides (4). Nähtuse kahaneva esinemissageduse järgi loodud korpuste järjestus on järgmine: jututubade tekstid (3708), suulise kõne tekstid (1257), ilukirjandus (690), tõlgitud ilukirjandus (353), ajakirjandus (254), populaarteaduslik kirjandus (101), seadusetekstid (4). Ainsuse 2. pöörde keskmine esinemissagedus oli 707, mis on kõige lähemal nähtuse esinemissagedusele ilukirjanduse korpuses.

Ainsuse 3. pöörde sagedus oli suurim tõlgitud ilukirjanduse tekstides (10 558) ja väikseim jututubade tekstides (5621). Vastavalt nähtuse kahanevale esinemissagedusele on korpuste järjestus selline: tõlgitud ilukirjandus (10 558), ajakirjandustekstid (10 070), populaarteaduslik kirjandus (9604), seadusetekstid (9529), ilukirjandus (9156), suulise kõne tekstide korpus (8810), jututubade tekstide korpus (5621). Keskmiselt oli nähtuse esinemissagedus kõikide korpuste peale 9315, mis on lähim sagedusnäitajale

ilukirjandustekstides. Väga lähedane keskmisele esinemissagedusele on ka seadusetekstide sagedusnäitaja.

Mitmuse 1. pööret leidus kõige rohkem suulise kõne tekstides (665) ning see puudus seadusetekstidest (0). Kahaneva mitmuse 1. pöörde esinemissageduse põhjal saab korpused järjestada järgnevalt: suulise kõne tekstide korpus (665), populaarteaduslik kirjandus (551), ilukirjandus (426), jututubade korpus (321), ajakirjandus (286), tõlgitud ilukirjandus (249), seadusetekstid (0). Mitmuse 1. pöörde puhul oli keskmine sagedus korpustes 375, mis sarnaneb enim sagedusnäitajaga jututubade korpuses.

Mitmuse 2. pöörde sagedus oli kõige suurem suulise kõne tekstide puhul (992) ning nähtus puudus seadusetekstidest (0). Nähtuse kahaneva esinemissageduse järjestus korpustest on järgmine: suulise kõne korpus (992), tõlgitud ilukirjanduse tekstid (626), jututubade tekstid (473), ilukirjandus (206), ajakirjandus (121), populaarteadusliku kirjanduse korpus (81), seadusetekstid (0). Mitmuse 2. pöörde keskmine esinemissagedus oli 349, kõige sarnasemad sellele on sagedusnäitajad jututubade ja ilukirjanduse tekstide puhul.

Mitmuse 3. pöörde sagedus oli suurim populaarteadusliku kirjanduse korpuses (3045) ja kõige väiksem jututubade korpuse tekstide puhul (919). Kahaneva esinemissageduse järgi saab korpused järjestada selliselt: populaarteaduslik kirjandus (3045), ajakirjandus (2484), tõlgitud ilukirjandus (1910), suulise kõne korpus (1610), ilukirjandustekstid (1595), seadusetekstid (1229), jututubade korpus (919). Keskmiselt oli mitmuse 3. pöörde esinemissagedus korpuste peale 1913, see sagedusnäitaja on kõige lähemal tõlgitud ilukirjanduse tekstide sagedusnäitajale.

Kui vaadelda pöörde kategooria esinemist tekstiliigiti, võib märgata, et seadusetekstides on pöörde kategooria sagedus väike, kõige sagedasem on 3. pööre. Jututubade tekstides on sageli kasutatud ainsust, seejuures enim ainsuse 1. ja 2. pööret. Kõikides tekstiliikides on ainsuse sagedus mitmuse sagedusest suurem ning kõige sagedasem on 3. pööre.

3.2.2. Kõneliigi kategooria

Kõneliigi kategoorias on kaks liiget: jaatav ja eitav. Mõlema liikme esinemissagedus igas korpuses on tabelis 6.

Tabel 6. **Kõneliigi kategooria liikmete esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
jaatav	14729	14670	15343	14482	14800	14101	15557	14816
eitav	2271	1793	1498	2215	3481	2465	2858	2263

Jaatava kõneliigi sagedust vaadeldes on näha, et suurim on see suulise kõne tekstides (15 557) ning kõige vähem on jaatavat kõneliiki kasutatud seadusetekstides (14 101). Jaatava kõneliigi kahaneva esinemissageduse põhjal saab korpused järjestada järgnevalt: suulise kõne korpus (15 557), populaarteadusliku kirjanduse tekstid (15 343), jututubade korpus (14 800), tõlgitud ilukirjandus (14 729), ajakirjandustekstid (14 670), ilukirjandustekstid (14 482), seadusetekstid (14 101). Keskmine nähtuse esinemissagedus oli 14 816, seega on see kõige lähemal jututubade tekstide korpuse näitajale.

Eitavat kõneliiki esines enim jututubade korpuses, kus esinemissagedus oli 3481. Kõige vähem esines seda populaarteadusliku kirjanduse tekstides, milles sagedus oli 1498. Vastavalt nähtuse kahanevale esinemissagedusele on korpuste järjestus selline: jututubade tekstid (3481), suulise kõne tekstide korpus (2858), seadusetekstid (2465), tõlgitud ilukirjandus (2271), ilukirjandus (2215), ajakirjandustekstid (1793), populaarteaduslik kirjandus (1498). Keskmine sagedusnäitaja oli 2263, sarnaseks võib pidada eitava kõneliigi esinemissagedust tõlgitud ilukirjanduse korpuses.

Kui vaadelda kõneliigi kategooria liikmete esinemissagedust tekstiliigiti, võib märgata, et see on üsna ühtlane. Suuri erinevusi üheski tekstiliigis teistega võrreldes ei esine.

3.2.3. Tegumoe kategooria

Tegumoe kategoorias on kaks liiget: isikuline ja umbisikuline. Mõlema liikme esinemissagedus igas korpuses on tabelis 7.

Tabel 7. **Tegumoe kategooria liikmete esinemissagedus.**

	tõlgitud ilukirjandus	ajakirjandus	populaarteaduslik kirjandus	ilukirjandus	jututoad	seadused	suuline	keskmine
isikuline	17277	16122	15587	17491	18220	12638	18569	16623
umbisikuline	1067	1562	2415	785	172	5624	751	1687

Isikulise tegumoe esinemissagedus oli suurim suulise kõne korpuses (18 569) ja väikseim seadusetekstide korpuses (12 638). Vastavalt kahanevale isikulise tegumoe esinemissagedusele on korpuste järjestus selline: suulise kõne tekstide korpus (18 569), jututubade tekstid (18 220), ilukirjandustekstid (17 491), tõlgitud ilukirjanduse korpus (17 277), ajakirjandustekstid (16 122), populaarteaduslik kirjandus (15 587), seadusetekstid (12 638). Nähtuse keskmine esinemissagedus korpustes oli 16 623, kõige sarnasem on see arv sagedusnäitajaga ajakirjandustekstides.

Umbisikulist tegumoodi esines kõige sagedamini seadusetekstides, kus sagedusnäitaja oli 5624. Kõige harvem oli nähtust jututubade tekstides (172). Kahaneva esinemissageduse põhjal on korpuste järjestus järgnev: seadusetekstid (5624), populaarteadusliku kirjanduse tekstid (2415), ajakirjandustekstide korpus (1562), tõlgitud ilukirjandus (1067), ilukirjandus (785), suulise kõne korpus (751), jututubade tekstide korpus (172). Keskmine umbisikulise tegumoe sagedus korpustes oli 1687, mis tähendab, et keskmise sagedusega kõige sarnasem on umbisikulise tegumoe sagedusnäitaja ajakirjandustekstide puhul.

Kui vaadelda tegumoe kategooria liikmete esinemist tekstiliigiti, võib näha, et jututubade tekstidele on iseloomulik väike umbisikulise tegumoe sagedus. Umbisikulise tegumoe rohke kasutuse poolest paistavad välja seadusetekstid.

3.2.4. Aja kategooria

Aja kategoorias on viis liiget: olevik, lihtminevik, täisminevik, enneminevik ja üldminevik. Kõikide liikmete esinemissagedus igas korpuses on tabelis 8.

Tabel 8. **Aja kategooria liikmete esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
olevik	5692	9954	9054	6669	14661	13488	13399	9779
lihtminevik	9126	4940	6116	8465	2581	614	4029	5654
täisminevik	551	1241	1200	457	850	2083	603	943
enneminevik	1232	157	335	829	22	53	125	451
üldminevik	334	114	113	274	58	21	76	159

Olevikku oli kõige rohkem kasutatud jututubade tekstides, kus oleviku esinemissagedus oli 14 661, ning kõige vähem tõlgitud ilukirjanduse korpuses, kus sagedus oli 5692. Esinemissageduse alusel moodustatud kahanev järjestus on järgnev: jututubade korpus (14 661), seadusetekstid (13 488), suulise kõne tekstid (13 399), ajakirjandustekstid (9954), populaarteaduslik kirjandus (9054), ilukirjandustekstid (6669) ja tõlgitud ilukirjanduse tekstid (5692). Kõikide korpuste keskmine oleviku esinemissagedus oli 9779, seega võib näha, et kõige sarnasem on see ajakirjanduse korpuse tekstidega.

Lihtminevikku leidus enim tõlgitud ilukirjanduse tekstikorpuses, esinemissageduseks 9126. Kõige vähem oli seda seadusetekstides, kus esinemissagedus oli 614. Kahaneva esinemissageduse põhjal on korpuste järjestus järgnev: tõlgitud

ilukirjanduse korpus (9126), ilukirjandus (8465), populaarteaduslik kirjandus (6116), ajakirjandustekstid (4940), suulise kõne tekstid (4029), jututubade korpus (2581) ja seadusetekstide korpus (614). Keskmine lihtmineviku esinemissagedus korpustes oli 5654, niisiis esindab keskmist lihtmineviku kasutust kõige paremini populaarteaduslik kirjandus.

Täismineviku esinemissagedus oli suurim seadusetekstides (2083) ja kõige väiksem ilukirjandustekstides (457). Täismineviku kahaneva esinemissageduse järgi loodud korpuste järjestus on järgmine: seadusetekstid (2083), ajakirjandustekstid (1241), populaarteaduslik kirjandus (1200), jututubade tekstid (850), suulise kõne korpuse tekstid (603), tõlgitud ilukirjandus (551) ja ilukirjandustekstid (457). Nähtuse esinemissagedus korpustes oli keskmiselt 943, mis tähendab, et keskmise sagedusega kõige sarnasem on lihtmineviku sagedusnäitaja jututubade korpuses.

Ennemineviku sagedust vaadeldes võib märgata, et suurim on see tõlgitud ilukirjanduse (1232) puhul ning kõige vähem on enneminevikku kasutatud jututubades (22). Ennemineviku kahaneva esinemissageduse põhjal saab korpused järjestada järgnevalt: tõlgitud ilukirjandus (1232), ilukirjandus (829), populaarteadusliku kirjanduse tekstid (335), ajakirjandustekstid (157), suulise kõne tekstid (125), seadusetekstid (53) ja jututubade tekstid (22). Keskmine ennemineviku esinemissagedus oli 451, seega on see kõige lähemal populaarteadusliku kirjanduse näitajale.

Üldminevikku oli enim tõlgitud ilukirjanduse korpuses (334), kõige vähem aga seadusetekstides (21). Kahaneva esinemissageduse järgi loodud järjestus korpustest on järgmine: tõlgitud ilukirjanduse tekstid (334), ilukirjanduse korpus (274), ajakirjandustekstid (114), populaarteaduslik kirjandus (113), suulise kõne korpus (76), jututubade tekstid (58) ja seadusetekstid (21). Keskmine nähtuse sagedus kõikide korpuste peale oli 159, sarnaseks võib pidada nii ajakirjanduse kui populaarteadusliku kirjanduse sagedusnäitajat.

Kui vaadelda aja kategooria liikmete sagedust tekstiliigiti, võib märgata, et seadusetekstides on minevikuaegadest kõige sagedasem täisminevik. Ilukirjanduse ja tõlgitud ilukirjanduse tekstides esineb minevikuaegu olevikust rohkem.

3.2.5. Kõneviisi kategooria

Kõneviisi kategoorias on esindatud neli liiget ehk neli uuritavat nähtust: kindel, tingiv, käskiv ja kaudne kõneviis. Kõikide liikmete esinemissagedus igas korpuses on tabelis 9.

Tabel 9. **Kõneviisi kategooria liikmete esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
kindel	15822	15347	16028	15316	15473	16159	16722	15814
tingiv	740	820	617	876	806	85	817	707
käskiv	262	193	116	482	1911	0	722	433
kaudne	144	100	83	107	8	2	21	76

Kindla kõneviisi esinemissagedus kõigis seitsmes vaadeldud korpuses oli kõige ühtlasem, teiste kõneviiside puhul olid erinevused märgatavamad. Kindlat kõneviisi oli enim suulise kõne korpuses, esinemissagedus oli 16 722. Kõige vähem leidis kindel kõneviis kasutust ilukirjanduse korpuses, kus esinemissagedus oli 15 316. Vastavalt kahanevale esinemissagedusele, täpne sagedus on sulgudes, sai korpustest luua järgneva järjestuse: suulise kõne korpus (16 722), seadusetekstide korpus (16 159), populaarteadusliku kirjanduse korpus (16 028), tõlgitud ilukirjanduse korpus (15 822), jututubade korpus (15 473), ajakirjanduse korpus (15 347), ilukirjandustekstide korpus (15 316). Keskmine kindla kõneviisi esinemissagedus kõigi seitsme korpuse kohta oli 15 814, seega sarnaneb keskmisele sagedusele enim tõlgitud ilukirjanduse korpuse sagedusnäitaja.

Tingiva kõneviisi puhul jäi silma vähene esinemissagedus seadusetekstide korpuses (85). Kõige enam oli tingivat kõneviisi kasutatud ilukirjanduses (876), võrreldes seadusetekstidega on erinevus umbes kümnekordne. Nähtuse kahaneva esinemissageduse järjestus korpustest on järgmine: ilukirjandustekstid (876), ajakirjandustekstid (820), suulise kõne tekstid (817), jututubade tekstid (806), tõlgitud ilukirjandus (740), populaarteaduslikud tekstid (617), seadusetekstid (85). Keskmine esinemissagedus tingival kõneviisil oli 707, mis on lähim tõlgitud ilukirjanduse näitajale.

Käskiv kõneviis esines seitsmest vaadeldud korpuses kuues. Mitte kordagi ei kasutatud käskivat kõneviisi seadusetekstides. Kõige rohkem oli seda jututubade tekstides, kus esinemissagedus oli 1911. Võrreldes teiste korpustega on arv üsna suur. Kahanev sagedusjärjestus on selline: jututubade tekstid (1911), suulise kõne korpus (722), ilukirjandustekstide korpus (482), tõlgitud ilukirjandus (262), ajakirjandustekstid (193), populaarteaduslik kirjandus (116), seadusetekstid (0). Keskmine esinemissagedus oli 433, sarnane on ilukirjandustekstide sagedusnäitaja.

Kaudset kõneviisi oli enim tõlgitud ilukirjanduse tekstide korpuses, kus esinemissagedus oli 144. Kõige vähem esines seda seadusetekstides, siinkohal sagedusega 2. Vastavalt kahanevale esinemissagedusele on korpuste järjestus selline: tõlgitud ilukirjanduse tekstide korpus (144), ilukirjandustekstid (107), ajakirjandustekstid (100), populaarteaduslik kirjandus (83), suulise keelekasutuse korpus (21), jututubade korpus (8), seadusetekstid (2). Keskmine sagedusnäitaja oli 76, sarnaseks võib pidada kaudse kõneviisi esinemissagedust populaarteadusliku kirjanduse korpuses.

Kui vaadelda kõneviisi kategooria liikmete esinemist tekstiliigiti, võib näha, et kõikides tekstiliikides kasutati enim kindlat kõneviisi. Seadusetekstides puudub käskiv kõneviis, seevastu jututubades on seda kasutatud märgatavalt rohkem kui teistes tekstiliikides.

3.3. Infiniitsed vormid

Verbi infiniitsete vormide alla kuuluvad *da*-infinitiiv, *ma*-infinitiiv, mineviku partitsiip ja *des*-vorm. Selles alapeatükis on esitatud nimetatud vormide sagedus korpustes.

3.3.1. *Da*-infinitiiv

Da-infinitiivi esinemissagedus igas korpuses on tabelis 10.

Tabel 10. ***Da*-infinitiivi esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
infinitiiv ehk *da* -infinitiiv	2250	2992	2676	2174	1595	3528	1789	2449

***Da*-infinitiivi** oli enim seadusetekstide korpuses (3528), kõige vähem aga jututubade tekstides (1595). Kahaneva esinemissageduse põhjal loodud järjestus korpustest on järgmine: seadusetekstid (3528), ajakirjandus (2992), populaarteaduslik kirjandus (2676), tõlgitud ilukirjandus (2250), ilukirjandustekstid (2174), suulise kõne tekstid (1789), jututubade korpus (1595). Keskmine nähtuse sagedus kõikide korpuste peale oli 2449, sarnaseks võib pidada tõlgitud ilukirjanduse sagedusnäitajat.

3.3.2. *Ma*-infinitiiv

Ma-infinitiivi ja selle kõikide vormide esinemissagedus igas korpuses on tabelis 11. Tabelis on iga liikme juures sulgudes esitatud ka lühendid, mida kasutati skriptis, et failidest vormide esinemise arv leida.

Tabel 11. ***Ma*-infinitiivi ja selle vormide esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
ma -infinitiive kokku	1252	1103	971	1292	1160	1499	991	1176
ma -vorm (sup ps ill)	808	901	795	1032	1052	1190	844	934
mas -vorm (sup ps in)	257	102	92	139	75	5	127	122
mast -vorm (sup ps el)	43	14	8	21	0	39	0	19
maks -vorm (sup ps tr)	1	14	17	4	0	11	0	7
mata -vorm (sup ps abes)	139	69	58	95	33	254	19	94
ma -infinitiiv (sup imps)	4	4	1	1	0	0	1	2

***Ma*-infinitiivide** sagedus oli kõige suurem seadusetekstide puhul (1499) ja kõige väiksem populaarteadusliku kirjanduse puhul (971). Nähtuse kahaneva esinemissageduse põhjal on korpustest võimalik koostada järgnev järjestus: seadusetekstid (1499), ilukirjandus (1292), tõlgitud ilukirjandus (1252), jututubade tekstid (1160), ajakirjandus (1103), suulise kõne tekstid (991), populaarteaduslik kirjandus (971). Keskmine supiinide sagedus kõikide korpuste kohta oli 1176, mis on lähim supiinide esinemissagedusele jututubade korpuses.

***Ma*-vormi** esines kõige sagedamini seadusetekstides (1190) ja kõige harvem populaarteadusliku kirjanduse tekstides (795). Vastavalt *ma*-vormi kahanevale esinemissagedusele on korpuste järjestus järgmine: seadusetekstid (1190), jututubade korpus (1052), ilukirjandustekstid (1032), ajakirjandus (901), suulise kõne tekstid (844), tõlgitud ilukirjandus (808), populaarteaduslik kirjandus (795). Keskmine nähtuse esinemissagedus oli 934, mis on kõige lähemal ajakirjandustekstide sagedusnäitajale.

***Mas*-vormi** sagedus oli suurim tõlgitud ilukirjanduses (257) ja kõige väiksem seadusetekstides (5). Sagedusnäitaja kahanemisel põhinev korpuste järjekord on selline: tõlgitud ilukirjandus (257), ilukirjandus (139), suulise kõne korpus (127), ajakirjandustekstid (102), populaarteaduslik kirjandus (92), jututubade tekstid (75), seadusetekstid (5). Keskmine sagedusnäitaja korpuste kohta on 122, see sarnaneb enim nähtuse sagedusnäitajaga suulise kõne tekstide puhul.

***Mast*-vormi** leidus kõige rohkem tõlgitud ilukirjanduses (43), vorm puudus jututubade tekstidest (0) ja suulise kõne tekstidest (0). Sagedusnäitaja kahanemise järgi loodud järjestus korpustest on järgmine: tõlgitud ilukirjandus (43), seadusetekstid (39), ilukirjandustekstid (21), ajakirjandustekstide korpus (14), populaarteadusliku kirjanduse tekstid (8), jututubade ja suulise kõne korpus (0). Nähtuse keskmine sagedus korpustes oli 19, see on lähim ajakirjandustekstide sagedusnäitajale.

***Maks*-vorm** oli kõige sagedasem populaarteaduslikus kirjanduses (17), seda ei esinenud jututubade (0) ja suulise keele tekstides (0). Kahaneva esinemissageduse põhjal on korpuste järjestus järgnev: populaarteaduslik kirjandus (17), ajakirjandus (14), seadusetekstid (11), ilukirjandus (4), tõlgitud ilukirjandus (1), jututubade ja suulise keele korpused (0). Keskmine sagedusnäitaja oli 7, sellele on kõige lähemal ilukirjandustekstide sagedusnäitaja.

***Mata*-vorm**i esines enim seadusetekstides (254) ning kõige vähem suulise kõne tekstide korpuses (19). Vastavalt nähtuse kahanevale esinemissagedusele on korpuste järjestus selline: seadusetekstide korpus (254), tõlgitud ilukirjandus (139), ilukirjandus (95), ajakirjandus (69), populaarteadusliku kirjanduse tekstid (58), jututubade korpus (33), suulise keele tekstid (19). Keskmine *mata*-vormi sagedus korpustes oli 94, mis tähendab, et keskmise sagedusega kõige sarnasem on nähtuse sagedusnäitaja ilukirjanduse puhul.

Umbisikulist *ma*-infinitiivi esines kõige sagedamini tõlgitud ilukirjanduse (4) ja ajakirjanduse (4) tekstide puhul. Seda vormi ei esinenud jututubade (0) ja seadusetekstide (0) korpustes. Ülejäänud korpustes – populaarteadusliku kirjanduse,

ilukirjanduse ja suulise kõne korpustes – oli nähtuse esinemissagedus 1. Seega oleks kahaneva esinemissageduse põhjal loodud järjestus korpustest selline: tõlgitud ilukirjandus ja ajakirjandus (4), populaarteadusliku kirjanduse, ilukirjanduse ja suulise kõne tekstid (1), jututubade ja seadusetekstide korpused (0). Nähtuse keskmine esinemissagedus oli 2, mis on lähim populaarteadusliku kirjanduse, ilukirjanduse ja suulise keele tekstidele.

3.3.3. Mineviku partitsiibid

Mineviku partitsiibi, mille tunnused on *nud*, *tud* ja *dud*, esinemissagedus igas korpuses on tabelis 12.

Tabel 12. **Mineviku partitsiipide esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
mineviku partitsiip	632	1311	1494	671	19	6156	174	1386

Mineviku partitsiipe oli enim kasutatud seadusetekstides (6156), kõige väiksem oli nende sagedus jututubade tekstides (19). Vastavalt esinemissageduse kahanemisele on korpuste järjekord järgmine: seadusetekstide korpus (6156), populaarteadusliku kirjanduse tekstid (1494), ajakirjandustekstid (1311), ilukirjandus (671), tõlgitud ilukirjandus (632), suulise kõne tekstid (174), jututubade tekstid (19). Keskmine mineviku partitsiipide esinemissagedus korpustes oli 1386, mis on kõige lähemal nähtuse esinemissagedusele ajakirjandustekstides.

3.3.4. Gerundiivid

Gerundiivi ehk *des*-vormi esinemissagedus igas korpuses on tabelis 13.

Tabel 13. ***Des*-vormi esinemissagedus.**

	tõlgitud ilukirjandus	**ajakirjandus**	**populaarteaduslik kirjandus**	**ilukirjandus**	**jututoad**	**seadused**	**suuline**	**keskmine**
des-vorm	490	462	486	433	44	453	53	371

***Des*-vormi** esines kõige sagedamini tõlgitud ilukirjanduses (490), kõige harvem aga jututubade tekstides (44). *Des*-vormi sageduse kahanemise järgi loodud korpuste järjestus on selline: tõlgitud ilukirjanduse tekstid (490), populaarteadusliku kirjanduse korpus (486), ajakirjandustekstid (462), seadusetekstid (453), ilukirjandus (433), suulise keele tekstid (53), jututubade korpus (44). Keskmiselt *des*-vormi esinemissagedus korpuste peale 371, see sagedusnäitaja on kõige lähemal ilukirjandustekstide sagedusnäitajale.

3.4. Verbikategooriate sagedused failiti

Alapeatükis kirjeldatakse, kas umbisikulise tegumoe, oleviku ja *da*-infinitiivi sagedus iga tekstiliigi failides oli ühtlane või erines märgatavalt. Selle analüüsiosa põhjal on võimalik välja selgitada, kas eelnevates alapeatükkides esitatud verbikategooriate sagedused on omased tekstiliigile tervikuna või on need üldistatud ning varieeruvad korpuse failides sellisel määral, et nende sagedusest ei saa rääkida tekstiliigi kui terviku omadusena. Oluline on märkida, et iga korpuse failid olid peamiselt samas suurusjärgus, kuid oli üksikuid faile, mis olid teistest suuremad või väiksemad – see võib tulemusi mõjutada. Tabelis 14 on esitatud valitud kategooriate esinemiskordade arv tõlgitud ilukirjanduse failis. Tabelid 15–32 on koostatud selliselt, et vasakus servas on kirjas vaadeldud nähtus, ülemisel real esinemiskordade vahemikud ning alumisel real see, mitmes failis vahemik

kajastus. Viimase tulba pealkiri „keskmiselt failides“ tähendab, et viimases tulbas on välja toodud, mitu korda nähtust failides keskmiselt esines.

3.4.1. Tõlgitud ilukirjanduse korpus

Inglise keelest tõlgitud ilukirjanduse korpuses oli üks fail. Niisiis puudub siinkohal võrdlusmaterjal mitme erineva faili näitel. Tabelis 14 on umbisikulise tegumoe, oleviku ja *da*-infinitiivi esinemine 75 500-sõnalises tõlgitud ilukirjanduse failis.

Tabel 14. **Valitud verbikategooriate esinemine tõlgitud ilukirjanduse failis.**

umbisikuline tegumood	777
olevik	4143
da -infinitiiv	1642

Tõlgitud ilukirjanduse korpuse faile oli üks, seega ei olnud võimalik võrrelda, kui palju erines verbikategooriate esinemine failiti. Olemasolevas failis esines umbisikulist tegumoodi 777, olevikku 4143 ja *da*-infinitiivi 1642 korda.

3.4.2. Ilukirjandustekstide korpus

Ilukirjanduskorpuse failide suurus oli umbes 2000 sõna. Tabelis 15 on umbisikulise tegumoe esinemine failides.

Tabel 15. **Umbisikulise tegumoe esinemine ilukirjandustekstide failides.**

	1–9	**10–19**	**20–29**	**30–39**	**40–49**	**keskmiselt failides**
umbisikuline	11 faili	30 faili	7 faili	4 faili	1 fail	15 korda

Tabelist 15 võib näha, et 30 failis esines umbisikulist tegumoodi 10–19 korda, 11 failis 1–9 korda, 7 failis 20–29 korda, 4 failis 30–39 korda ja ühes failis 40–49 korda. Keskmine nähtuse esinemine faili kohta oli 15. Umbisikulist tegumoodi esines failides pigem ühtlaselt, sest suurem osa failidest kuulus vahemikku, millesse jäi ka aritmeetiline keskmine.

Tabel 16 sisaldab andmeid oleviku esinemise kohta ilukirjanduskorpuse failides.

Tabel 16. **Oleviku esinemine ilukirjandustekstide failides.**

	10–99	**100–179**	**200–299**	**330–349**	**keskmiselt failides**
olevik	18 faili	24 faili	9 faili	2 faili	131 korda

Tabelist 16 selgub, et 24 failis esines olevikku 100–179 korda, 18 failis 10–99 korda, 9 failis 200–299 korda ja 2 failis 330–349 korda. Keskmiselt esines olevikku failis 131 korda. Suurem osa failidest kuulus vahemikku, millesse jäi aritmeetiline keskmine. Samas oli palju ka neid faile, milles olevikku esines vähem kui 100 korda. Töö autori meelest võib tulemuse tingida isikustiili esilekerkimine – mõni autor kirjutab nii, et teose tegevus toimub olevikus, mõni otsustab minevikuaegade kasuks.

Tabelis 17 on välja toodud *da*-infinitiivi esinemine ilukirjanduskorpuse failides.

Tabel 17. ***Da*-infinitiivi esinemine ilukirjandustekstide failides.**

	1–29	**30–49**	**50–69**	**keskmiselt failides**
da -infinitiiv	10 faili	30 faili	13 faili	43 korda

Tabelist 17 tuleb välja, et 30 failis esines *da*-infinitiivi 30–49 korda, 13 failis 50–69 korda ja 10 failis 1–29 korda. Keskmiselt oli infinitiivi failis 43 korda. Nähtuse esinemist failiti võib pidada ühtlaseks, sest suurem osa failidest kuulus vahemikku, millesse jäi aritmeetiline keskmine.

Umbisikulist tegumoodi ja *da*-infinitiivi esines ilukirjanduskorpuse failides pigem ühtlaselt. Oleviku puhul märgati kõikumist, mille põhjuseks võib olla kirjandusteose autori isikustiil. Ilukirjandusteksti verbikategooriate sagedusi saab pidada tekstiliigi omaduseks.

3.4.3. Populaarteadusliku kirjanduse korpus

Populaarteadusliku kirjanduse korpuse failid koosnesid umbes 2000 sõnast. Tabelis 18 on umbisikulise tegumoe esinemine failides.

Tabel 18. **Umbisikulise tegumoe esinemine populaarteadusliku kirjanduse korpuse failides.**

	10–29	**30–59**	**60-79**	**100-109**	**keskmiselt failides**
umbisikuline	13 faili	27 faili	4 faili	1 fail	40 korda

Tabeli 18 põhjal on näha, et 27 failis esines umbisikulist tegumoodi 30–59 korda, 13 failis 10–29 korda, 4 failis 60–79 korda ja 1 failis 100–109 korda. Keskmiselt oli nähtust failis 40 korda. Suurem osa failidest kuulus vahemikku, millesse jäi aritmeetiline keskmine ja seega saab umbisikulise tegumoe esinemist failiti pidada ühtlaseks.

Tabelis 19 on kirjas oleviku esinemine korpuse failides.

Tabel 19. **Oleviku esinemine populaarteadusliku kirjanduse korpuse failides.**

	10–99	**100–199**	**200–279**	**keskmiselt failides**
olevik	14 faili	21 faili	7 faili	146 korda

Tabeli 19 põhjal saab öelda, et 21 failis oli olevikku 100–199 korda, 14 failis 10–99 korda ja 7 failis 200–279 korda. Keskmiselt oli olevikku ühes failis 146 korda. Suurem osa failidest kuulus vahemikku, millesse jäi aritmeetiline keskmine, niisiis esines olevikku failiti pigem ühtlaselt.

Tabel 20 kajastab *da*-infinitiivi esinemist korpuse failides.

Tabel 20. ***Da*-infinitiivi esinemine populaarteadusliku kirjanduse korpuse failides.**

	10–29	**30–79**	**90–109**	**keskmiselt failides**
da -infinitiiv	11 faili	32 faili	2 faili	45 korda

Tabelis 20 olevatest tulemustest selgub, et 32 failis esines *da*-infinitiivi 30–79 korda, 11 failis 10–29 korda ja 2 failis 90–109 korda. Keskmiselt oli olevikku failis 45 korda. Suurem osa failidest kuulus samasse vahemikku, millesse jäi aritmeetiline keskmine ja seega saab öelda, et *da*-infinitiivi esinemine populaarteadusliku kirjanduse korpuse failides oli ühtlane.

Populaarteadusliku kirjanduse korpuse puhul esines nii umbisikulist tegumoodi, olevikku kui *da*-infinitiivi failiti üsna võrdselt. Selle korpuse tekstide verbikategooriate sagedusi saab pidada tekstiliigi omaduseks. Populaarteadusliku kirjanduse korpus sellisena, nagu seda esindab ajakiri Horisont, on verbikategooriate esinemise poolest ühtlane.

3.4.4. Seadusetekstide korpus

Seadusetekstide korpuse failid koosnesid umbes 10 000 sõnast. Tabelis 21 on andmed umbisikulise tegumoe esinemisest failides.

Tabel 21. **Umbisikulise tegumoe esinemine seadusetekstide korpuse failides.**

	40–79	**100–189**	**200–279**	**360–369**	**keskmiselt failides**
umbisikuline	2 faili	6 faili	6 faili	1 fail	198 korda

Tabeli 21 põhjal on näha, et 6 failis esines umbisikulist tegumoodi 100–189 korda ja 6 failis esines seda 200–279 korda, 2 failis 40–79 korda ja ühes failis 360–369 korda. Keskmiselt oli umbisikulist tegumoodi failides 198 korda, niisiis võib väita, et nähtust esines failides ühtlaselt.

Tabelis 22 on oleviku esinemine korpuse failides.

Tabel 22. **Oleviku esinemine seadusetekstide korpuse failides.**

	100–399	**400–699**	**800–899**	**keskmiselt failides**
olevik	6 faili	8 faili	2 faili	475 korda

Tabelist 22 selgub, et 8 failis oli olevikku 400–699 korda, 6 failis 100–399 korda ja 2 failis 800–899 korda. Keskmine oleviku esinemine failides oli 475. Saab öelda, et oleviku esinemine failides varieerub veidi.

Tabelis 23 on välja toodud *da*-infinitiivi esinemine.

Tabel 23. ***Da*-infinitiivi esinemine seadusetekstide korpuse failides.**

	10–99	**100–189**	**200–209**	**keskmiselt failides**
da -infinitiiv	5 faili	7 faili	4 faili	127 korda

Tabeli 23 andmed näitavad, et 7 failis esines *da*-infinitiivi 100–189 korda, 5 failis 10–99 korda, 4 failis 200–209 korda. Keskmiselt oli *da*-infinitiivi failis 127 korda. Kuigi suurem osa faile kuulub samasse vahemikku, milles on aritmeetiline keskmine, võib siiski öelda, et *da*-infinitiivi esinemine failiti oli pisut varieeruv. Selle põhjus vajab edasist uurimist.

Seadusetekstide failides esines umbisikulist tegumoodi üsna ühtlaselt, oleviku ja *da*-infinitiivi puhul võis täheldada varieerumist. Autor arvab siiski, et verbikategooriate sagedusi võib pidada tekstiliigi omaduseks.

3.4.5. Suulise keelekasutuse tekstide korpus

Suulise keelekasutuse korpuse failid koosnesid umbes 10 000 sõnast. Tabelis 24 on andmed umbisikulise tegumoe esinemisest failides.

Tabel 24. **Umbisikulise tegumoe esinemine suulise keelekasutuse korpuse failides.**

	10–19	**30–69**	**70–89**	**keskmiselt failides**
umbisikuline	1 fail	5 faili	4 faili	57 korda

Tabelist 24 ilmneb, et 5 failis oli umbisikulist tegumoodi 30–69 korda, 4 failis 70–89 korda, ühes failis 10–19 korda. Keskmine umbisikulise tegumoe esinemise arv failis oli 57. Kuigi suurem osa failidest jääb samasse vahemikku, milles on aritmeetiline keskmine, saab siiski öelda, et umbisikulise tegumoe puhul on märgata varieerumist.

Tabel 25 esitab oleviku esinemise failides.

Tabel 25. **Oleviku esinemine suulise keelekasutuse korpuse failides.**

	600–899	**900–1199**	**1300–1499**	**keskmiselt failides**
olevik	5 faili	2 faili	3 faili	1011 korda

Tabeli 25 põhjal on näha, et 5 failis esines olevikku 600–899 korda, 3 failis 1300–1499 korda ja 2 failis 900–1199 korda. Keskmiselt oli olevikku failis 1011 korda. Selgub, et vahemikus, millesse jääb aritmeetiline keskmine, oli kõige vähem faile. Oleviku esinemine suulise keelekasutuse failides on seega varieeruv.

Tabelis 26 on *da*-infinitiivi esinemise andmed.

Tabel 26. ***Da*-infinitiivi esinemine suulise keelekasutuse korpuse failides.**

	60–99	**150–179**	**210–239**	**keskmiselt failides**
da -infinitiiv	5 faili	2 faili	3 faili	139 korda

Tabel 26 toob välja, et 5 failis oli *da*-infinitiivi 60–99 korda, 3 failis 210–239 korda ja 2 failis 150–179 korda. Keskmiselt oli nähtust failides 139 korda. Vahemikus, mis on

aritmeetilisele keskmisele kõige lähemal, on kõige vähem faile. *Da*-infinitiivi sagedus suulise keelekasutuse failides on kõikuv.

Suulise keelekasutuse korpuse failides varieerus nii umbisikulise tegumoe, oleviku kui ka *da*-infinitiivi esinemine. Oletatav põhjus saab olla näiteks erinevate keelekasutajate kõnelemise stiil ja keeleline taust. Võib olla, et suulise keelekasutuse puhul on esitatud verbikategooriate sagedused üldistatud ning neid ei saa käsitleda tekstiliigi kui terviku omadusena.

3.4.6. Ajakirjandustekstide korpus

Ajakirjandustekstide korpuse failid normaliseeriti 12 000 sõna peale. Tabelis 27 on umbisikulise tegumoe esinemine failides.

Tabel 27. **Umbisikulise tegumoe esinemine ajakirjandustekstide korpuse failides.**

	600–899	**900–1999**	**1300–1599**	**keskmiselt failides**
umbisikuline	4 faili	5 faili	2 faili	1017 korda

Tabelist 27 selgub, et 5 failis esines umbisikulist tegumoodi 900–1999 korda, 4 failis 600–899 korda, 2 failis 1300–1599 korda. Keskmiselt oli nähtust failis 1017 korda. Suurem osa failidest kuulus samasse vahemikku, milles on aritmeetiline keskmine. Umbisikulise tegumoe esinemise puhul võib märgata väikest varieerumist, kuid üldiselt võib faile pidada üsna ühtlaseks.

Tabel 28 näitab oleviku esinemist failides.

Tabel 28. **Oleviku esinemine ajakirjandustekstide korpuse failides.**

	5000–5299	**5500–5699**	**6500–6699**	**keskmiselt failides**
olevik	4 faili	4 faili	3 faili	5734 korda

Tabeli 28 põhjal on nähtav, et 4 failis esines olevikku 5000–5299 korda ja 4 failis 5500–5699 korda, 3 failis 6500–6699 korda. Keskmiselt oli olevikku failis 5734 korda. Aritmeetilisele keskmisele lähim vahemik on 5500–5699, kuid oleviku esinemine failiti on veidi varieeruv.

Tabel 29 toob välja *da*-infinitiivi esinemise failides.

Tabel 29. ***Da*-infinitiivi esinemine ajakirjandustekstide korpuse failides.**

	1500–1699	**1700–1999**	**2000–2199**	**keskmiselt failides**
da -infinitiiv	4 faili	5 faili	2 faili	1814 korda

Tabeli 29 andmetest selgub, et 5 failis oli *da*-infinitiivi 1700–1999 korda, 4 failis 1500–1699 korda ja 2 failis 2000–2199 korda. Keskmiselt oli *da*-infinitiivi failis 1814 korda. Suurem osa failidest kuulub samasse vahemikku, millesse jääb aritmeetiline keskmine, ent siiski võib nähtuse esinemise puhul failides märgata väikest varieerumist.

Ajakirjandustekstide korpuse failides on umbisikulise tegumoe, oleviku ja *da*-infinitiivi puhul näha mõningast kõikumist, mis vajab edasist uurimist. Oletada võib, et ajakirjanduse puhul on võimalik eristada alamliike, näiteks spordiuudist ja arvamuslugu, milles on verbivorme kasutatud erinevalt. Autori saab leitud verbikategooriate sagedusi pigem siiski kasutada tekstiliigi kui terviku omadustest rääkides.

3.4.7. Jututubade tekstide korpus

Jututubade tekstide korpuse failid normaliseeriti 6250 sõna peale. Tabelis 30 on esitatud umbisikulise tegumoe esinemise andmed.

Tabel 30. **Umbisikulise tegumoe esinemine jututubade tekstide korpuse failides.**

	0	**20–69**	**90–109**	**150–159**	**keskmiselt failides**
umbisikuline	1 fail	9 faili	3 faili	1 fail	57 korda

Tabelist 30 on näha, et 9 failis esines umbisikulist tegumoodi 20–69 korda, 3 failis 90–109 korda, ühes failis 150–159 korda ja ühes failis mitte kordagi. Keskmiselt oli nähtust failis 57 korda. Suurem osa failidest kuulus samasse vahemikku, millesse jäi aritmeetiline keskmine. Võib öelda, et umbisikulise tegumoe esinemine failides oli ühtlane.

Tabel 31 näitab, kui palju esines korpuse failides olevikku.

Tabel 31. **Oleviku esinemine jututubade tekstide korpuse failides.**

	3200–3499	**4100–4599**	**4800–5099**	**keskmiselt failides**
olevik	2 faili	6 faili	6 faili	4474 korda

Tabeli 31 põhjal on selge, et 6 failis esines olevikku 4100–4599 korda ja 6 failis 4800–5099 korda, 2 failis oli olevikku 3200–3499 korda. Keskmiselt oli olevikku failis 4474 korda. On näha, et oleviku esinemine on failiti veidi varieeruv.

Tabel 32 toob välja *da*-infinitiivi esinemise korpuse failides.

Tabel 32. ***Da*-infinitiivi esinemine jututubade tekstide korpuse failides.**

	300–399	**400–499**	**500–599**	**600–699**	**keskmiselt failides**
da -infinitiiv	4 faili	3 faili	6 faili	1 fail	487 korda

Tabeli 32 andmed näitavad, et 6 failis oli *da*-infinitiivi 500–599 korda, 4 failis 300–399 korda, 3 failis 400–499 korda ja ühes failis 600–699 korda. Keskmiselt esines nähtust failis 487 korda. *Da*-infinitiivi esinemine failides varieerus.

Jututubade tekstide korpuse failides oli kõige ühtlasem umbisikulise tegumoe esinemine, oleviku esinemine varieerus veidi ja *da*-infinitiivi esinemine kõikus failiti märgatavalt. Kuna suhtlus jututubades võib olla spontaanse suulise kõne tekstivorm, siis võib arvata, et siingi mõjutavad tulemusi vestlejate kõnelemise stiil ja keeleline taust. Võib olla, et jututubade tekstide puhul on esitatud verbikategooriate sagedused üldistatud ning neid ei saa käsitleda tekstiliigi kui terviku omadusena.

4. Järeldused

Selles peatükis iseloomustatakse tekstiliike neis esinenud verbi grammatiliste kategooriate kaudu.

Ilukirjanduse tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega esines selles kõige sagedamini tingivat kõneviisi ning kõige harvem leidus kindlat kõneviisi ja täisminevikku. Selles tekstiliigis esinenud kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: käskiv kõneviis, ainsuse 2. pööre, ainsuse 3. pööre, mitmuse 2. pööre, *maks*-vorm, *mata*-vorm, umbisikuline *ma*-infinitiiv, *des*-vorm. Ilukirjanduse omaduseks võib pidada suurt erinevate verbide sagedust, mis võib tuleneda ilukirjandusteose autori püüdest kirjutada isikupärast ja väljendusrikast teksti. Ilukirjandustekstide verbikategooriad esinesid failides pigem ühtlaselt ja neid saab arvestada tekstiliigi omadusena.

Tõlgitud ilukirjanduse tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega oli selles suurim verbide, kaudse kõneviisi, lihtmineviku, ennemineviku, üldmineviku, ainsuse 3. pöörde, *mas*-vormi, *mast*-vormi, umbisikulise *ma*-infinitiivi ja *des*-vormi esinemissagedus. Võrreldes teiste korpustega oli tõlgitud ilukirjanduses kasutatud kõige vähem oleviku aega. Tõlgitud ilukirjanduses esinenud kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: kindel kõneviis, tingiv kõneviis, ainsus ja selle 1. pööre, mitmus ja selle 3. pööre, eitav kõneviis, *da*-infinitiiv. See tekstiliik oli kõige sarnasem ilukirjanduse tekstiliigiga. Kuna tõlgitud ilukirjanduse korpus koosnes ühest failist, ei saanud selle puhul võrrelda verbikategooriate esinemist failiti.

Ajakirjanduse tekstiliiki iseloomustab enim see, et ühtegi vaadeldud verbi grammatilist kategooriat ei esinenud selles tekstiliigis kõige rohkem. Kõige väiksem oli selles teiste tekstiliikides võrreldes eitava kõneviisi sagedus. Ajakirjandustekstides esinenud kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: olevik,

üldminevik, isikuline tegumood, umbisikuline tegumood, *ma*-vorm, *mast*-vorm, mineviku partitsiip. Ajakirjandustekstide failides võis verbikategooriate esinemise puhul märgata varieeruvust, mille põhjustajaks võib olla see, et ajakirjanduses saab eristada teksti alamliike ja neis võib verbivormide kasutus erineda – see vajab edasist uurimist. Autor usub, et sellest hoolimata võib siinses töös kasutada verbikategooriate sagedusi tekstiliigi kui terviku omadusena.

Populaarteadusliku kirjanduse tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega oli selles suurim põhiverbide, mitmuse ja selle 3. pöörde ning *maks*-vormi esinemissagedus. Kõige vähem oli populaarteaduslikus kirjanduses abiverbe, eitavat kõneliiki, *ma*-infinitiive ja sealhulgas *ma*-vormi. Selles tekstiliigis esinenud kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: modaalverb, kaudne kõneviis, lihtminevik, enneminevik, üldminevik, umbisikuline *ma*-infinitiiv. Populaarteadusliku kirjanduse tekstides esinesid verbikategooriad failides ühtlaselt ning neid võib pidada tekstiliigi omaduseks.

Jututubade tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega oli selles suurim käskiva kõneviisi, oleviku, ainsuse ning selle 1. ja 2. pöörde ning eitava kõneliigi sagedus. Kõige väiksem oli verbide, modaalverbide, ennemineviku, ainsuse 3. pöörde, umbisikulise tegumoe, *da*-infinitiivi, mineviku partitsiibi ja *des*-vormi esinemissagedus. Jututubade korpuses ei esinenud *ma*-infinitiivi *mast*-vormi, *maks*-vormi ja umbisikulist vormi. Korpuse tekstides esinenud kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: täisminevik, mitmuse 1. pööre, mitmuse 2. pööre, jaatav kõneliik, *ma*-infinitiiv. Jututubade tekstides oli ühtlane umbisikulise tegumoe esinemine, oleviku ja *da*-infinitiivi esinemine varieerus ning siinkohal on keeruline öelda, kas verbikategooriate sagedusi saab pidada tekstiliigi kui terviku omaduseks.

Suulise keelekasutuse tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega oli selles kõige suurem kindla kõneviisi, mitmuse 1. ja 2. pöörde, isikulise tegumoe, jaatava kõneliigi sagedus. Kõige vähem esines tekstiliigis *mata*-vormi. Suulise kõne korpusest puudusid *ma*-infinitiivi *mast*-vorm ja *maks*-vorm. Korpuse tekstides esinenud

kategooriatest olid keskmise esinemissagedusega kõige sarnasemad need: põhiverbid, abiverbid, *mas*-vorm, umbisikuline *ma*-infinitiiv. Suulist keelekasutust esindava korpuse failides varieerus umbisikulise tegumoe, oleviku ja *da*-infinitiivi esinemine, seega ei saa kinnitada, et verbikategooriate sagedusi saab pidada tekstiliigi kui terviku omaduseks.

Seaduste tekstiliiki iseloomustab see, et võrreldes teiste tekstiliikidega oli selles suurim abiverbide, modaalverbide, täismineviku, umbisikulise tegumoe sagedus. Korpuses oli väikseim põhiverbide, tingiva kõneviisi, kaudse kõneviisi, lihtmineviku, üldmineviku, ainsuse ning selle 1. ja 2. pöörde, mitmuse, isikulise tegumoe ja jaatava kõneliigi esinemissagedus. Seadusetekstide korpuses ei esinenud käskivat kõneviisi, mitmuse 1. ja 2. pööret. Seaduste tekstiliigi sagedusnäitajad ei sarnanenud ühegi nähtuse keskmise sagedusega. Seadusetekstide omaduseks võib pidada väikest erinevate verbide sagedust, mis võib tuleneda sellest, et seadusetekstide puhul püütakse kirjutada üheselt mõistetavat teksti ning esitada sama tegevust sama verbiga. Võrreldes ilukirjandustekstidega, kus erinevate verbide sagedus oli suur, on seaduseteksti kommunikatiivne eesmärk erinev. Seadusetekstide failide verbikategooriate sagedused olid pigem ühtlased ning autori hinnangul saab neid pidada tekstiliiki omaduseks. Samuti võib märkida, et seadusetekste eristavad verbikategooriate sagedused hästi ning selle tekstiliigi puhul tulevad erisused kõige paremini esile.

Teooriaosas esitati võimalus, et ilukirjandustekstide ja inglise keelest tõlgitud ilukirjanduse teksti korpused esindavad ilukirjanduskeelt; ajakirjandustekstid, populaarteaduslik kirjandus ja seadusetekstid tarbekeelt; suulise keelekasutuse tekstid ja jututubade tekstid argikeelt. Järgnevalt tuuakse välja tekstide erinevused ja sarnasused nende keelevariantide kaudu.

Ilukirjanduskeele tekstides sarnanesid järgmiste kategooriate sagedused: verbide üldhulk, modaalverb, erinevate verbide osakaal, ainsus ja selle 2. pööre, eitav kõneliik, isikuline tegumood, täisminevik, üldminevik, kaudne kõneviis, *da*-infinitiiv, *ma*-infinitiiv, *ma*-infinitiivi *maks*-vorm ja *mata*-vorm ning mineviku partitsiip. Ilukirjanduses oli tõlgitud ilukirjandusest rohkem ainsuse 1. pööret, mitmuse 1. pööret, käskivat

kõneviisi. Tõlgitud ilukirjanduses oli enam mitmuse 2. pööret ja umbisikulist *ma*-infinitiivi.

Tarbekeele tekstidest eristus enim seadusetekstide korpus. **Seadusetekstid** sarnanesid teistega järgmistes kategooriates: ainsus ja selle 3. pööre, *da*-infinitiiv, *ma*-infinitiivi *maks*-vorm, *des*-vorm. **Populaarteaduslik kirjandus ja seadused** sarnanesid kindla kõneviisi sageduse poolest. **Ajakirjandustekstides ja populaarteadusliku kirjanduse** tekstides olid sarnased verbide üldhulk, põhiverbide ja abiverbide sagedus, erinevate verbide osakaal, ainsuse 1. pöörde ning mitmuse 2. ja 3. pöörde sagedus ja eitava kõneliigi, isikulise ja umbisikulise tegumoe, täismineviku, üldmineviku, käskiva ja kaudse kõneviisi, *ma*-infinitiivi *mas*-, *mast*- ja *mata*-vormi ning mineviku partitsiibi sagedus.

Seadusetekste eristas suurem abi- ja modaalverbide, täismineviku, umbisikulise tegumoe, oleviku, *ma*-infinitiivi ja selle *mata*-vormi ning mineviku partitsiipide sagedus; väiksem erinevate verbide osakaal; väike mitmuse, ainsuse 1. pöörde, jaatava kõneliigi, lihtmineviku, ennemineviku, üldmineviku, tingiva ja kaudse kõneviisi, *ma*-infinitiivi *mas*-vormi sagedus. **Populaarteaduslikus kirjanduses** oli kõige rohkem mitmust ning selle 1. pööret, lihtminevikku, jaatavat kõneliiki; kõige vähem eitavat kõneliiki ja olevikku. **Ajakirjandustekstides** oli enim ainsuse 2. pööret, tingivat kõneviisi, umbisikulist *ma*-infinitiivi, kõige vähem leidus modaalverbe.

Argikeele ehk jututubade ja suulise keele tekstid sarnanesid järgmistes kategooriates: põhiverbide ja abiverbide sagedus, erinevate verbide osakaal, ainsus, isikuline tegumood, üldminevik, tingiv ja kaudne kõneviis, *da*-infinitiiv, *ma*-infinitiivi *mata*-vorm, *des*-vorm. Mõlemast korpusest puudusid *ma*-infinitiivi *mast*- ja *maks*-vorm. **Suulises keelekasutuses** oli uuritud failide põhjal jututubade tekstidest suurem verbide üldhulk, rohkem modaalverbe, ainsuse 3. pööret, mitmust ja selle kõiki pöördeid, jaatavat kõneliiki, umbisikulist tegumoodi, lihtminevikku, ennemineviku, umbisikulist *ma*-infinitiivi, mineviku partitsiipe**. Jututubades** oli enam ainsuse 1. ja 2. pööret ja käskivat kõneviisi.

On näha, et kõige selgemini tulevad verbi grammatiliste kategooriate abil esile seadusetekstide erijooned. Omavahel on kõige sarnasemad ilukirjanduse ja tõlgitud ilukirjanduse tekstid, samuti sarnanevad ajakirjanduse ja populaarteadusliku kirjanduse tekstid. Suulise keelekasutuse ja jututubade tekstide vahel märgati erinevusi.

Kokkuvõte

Bakalaureusetöö „Verbi grammatiliste kategooriate esinemine eri tekstiliikides“ keskendus verbikategooriate sagedusele ilukirjanduse, inglise keelest tõlgitud ilukirjanduse, ajakirjanduse, populaarteadusliku kirjanduse, seaduse, suulise keelekasutuse ja jututubade korpustes.

Töö neljast peatükist esimene andis teoreetilise ülevaate eesti keele verbist, grammatilistest kategooriatest ja tekstiliikidest, samuti varem sel teemal tehtud uurimustest. Teises peatükis esitati informatsioon uuritava materjali ja kasutatava meetodi kohta ning kirjeldati töö käiku. Kolmas peatükk sisaldas analüüsi verbikategooriate esinemissagedusest. Neljandas peatükis toodi välja tähtsamad tulemused ja järeldused.

Peamine eesmärk oli välja selgitada, kas verbi grammatiliste kategooriate sageduse põhjal on võimalik teha kindlaks, mis liiki on tekst. Töö rakenduslik siht oli leida verbikategooriad, mis võiksid toimida tunnustena tekstide klassifitseerimisel tekstiliikidesse ning teoreetiline siht oli võrrelda tekstiliikides esinevate verbikategooriate sagedusi. Töö autori hinnangul sai eesmärk täidetud. Eristusid verbi grammatilised kategooriad, mida võiks lugeda tekstiliigi tunnuseks, sest neid esines tekstis sagedasti või harva, kuid mõned neist võivad eristada pigem üksikuid autoreid või kõnelejaid kui tekstiliike tervikuna. Uurimisküsimused said vastused. Töös toodi välja verbi grammatiliste kategooriate sagedused käsitletud tekstiliikides, võrreldi sageduste erinevusi ning võib öelda, et verbikategooriate sageduse põhjal on võimalik teada saada, mis liiki on tekst. Ka autori püstitatud hüpoteesid said kinnitatud.

Bakalaureusetöö võib edaspidi saada võrdlusaluseks mõne keelevariandi iseloomustamisel.

Kirjandus

EKK = Erelt, Mati, Tiiu Erelt, Kristiina Ross 2007. Eesti keele käsiraamat. Kolmas, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.

Erelt, Mati 2017. Öeldis. – Eesti keele süntaks. (= Eesti keele varamu III) Toim. Mati Erelt, Helle Metslang. Tartu: Tartu Ülikooli Kirjastus, 94–239.

GitHub 2021. Verbikategooriad. Kärt-Kristiin Jaagu bakalaureusetöö skriptid. https://github.com/kkristiin/Verbikategooriad. Viimati muudetud 15.05.2021.

Hennoste, Tiit 2000a. Eesti keele allkeeled. Tartu Ülikooli eesti keele õppetooli toimetised 16. Toim. Tiit Hennoste. Tartu: Tartu Ülikooli Kirjastuse trükikoda.

Hennoste, Tiit 2000b. Sissejuhatus suulisesse eesti keelde. – Oma Keel, 1, 48–57.

Hennoste, Tiit, Kadri Muischnek 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Toim. Tiit Hennoste. Tartu: Tartu Ülikooli kirjastus, 183–217.

Kaalep, Heiki-Jaan, Sjur Nørstebø Moshagen, Trond Trosterud 2018. Estonian Morphology in the Giella Infrastructure. – Human Language Technologies – The Baltic Perspective, 307, 47–54. https://doi.org/10.3233/978-1-61499-912-6-47. Vaadatud 03.05.2021.

Kasik, Reet 2007. Sissejuhatus tekstiõpetusse. Tartu: Tartu Ülikooli Kirjastus.

Kirt, Riin 2013. Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning „Eesti keele seletava sõnaraamatu“ märksõnaloendi võrdlus. Magistritöö. Käsikiri Tartu Ülikooli eesti keele osakonnas.

Kitsnik, Mare 2014. Verbivormid B1- ja B2-taseme kirjalikus õppijakeeles. – Eesti ja soome-ugri keeleteaduse ajakiri = Journal of Estonian and Finno-Ugric Linguistics, 5 (3), 9−35. https://doi.org/10.12697/jeful.2014.5.3.01. Vaadatud 13.02.2021.

Kuidas mõista andmestunud maailma? Metodoloogiline teejuht 2020. Koost. Ja toim. Anu Masso, Katrin Tiidenberg, Andra Siibak. Tallinn: TLÜ Kirjastus.

MÜJK = Morfoloogiliselt ühestatud jututubade korpus 2020. https://www.cl.ut.ee/korpused/jutumorf/. Vaadatud 02.02.2021.

Narits, Raul 2002. Juriidiline semantika ehk õiguskeel Eesti õiguskorra kontekstis. Riigikogu toimetised 5. https://rito.riigikogu.ee/eelmised-numbrid/nr-5/juriidiline-semantika-eesti-oigusloomes/. Vaadatud: 27.03.2021.

Salla, Sigrid 2002. Jututuba kui võrgusuhtlusvorm. Tartu Ülikooli eesti keele õppetooli toimetised 23. Tekstid ja taustad. Artikleid tekstianalüüsist. Toim. Reet Kasik. Tartu: Tartu Ülikooli Kirjastuse trükikoda.

SEKK = Suulise eesti keele korpus 2020. https://keeleressursid.ee/et/220-suulise-eesti-keele-korpus. Vaadatud 02.02.2021.

Tuldava, Juhan, Astrid Villup 1976. Sõnaliikide sagedusest ilukirjandusproosa autorikõnes. Tartu Riikliku Ülikooli toimetised. Keelestatistika 1. Toim. Jaan Soontak. Tartu: Tartu Riiklik Ülikool.

TÜAU = Tartu Ülikooli arvutilingvistika uurimisrühm 2018. Morfoloogiliselt ühestatud korpus. https://cl.ut.ee/korpused/morfkorpus/. Vaadatud 12.01.2021.

TÜAU2 = Tartu Ülikooli arvutilingvistika uurimisrühm 2019. Uue meedia korpus: jututoad.
https://www.cl.ut.ee/korpused/segakorpus/uusmeedia/jututoad/index.php?lang=et.
Vaadatud 17.01.2021.

Viht, Annika, Külli Habicht 2019. Eesti keele sõnamuutmine. Toim. Helle Metslang. Eesti keele varamu IV. Tartu: Tartu Ülikooli Kirjastus.

Vincze, Veronika 2013. Domain Differences in the Distribution of Parts of Speech and Dependency Relations in Hungarian . – Journal of Quantitative Linguistics, 4, 314–338. https://doi.org/10.1080/09296174.2013.830553. Vaadatud 28.04.2021.

Verb Categories in Different Genres of Text. Summary

Bachelor's thesis „Verb Categories in Different Genres of Text" belongs to the field of computational linguistics. The purpose of this research was to study the frequency of verb categories in genres of text in order to find out if it is possible to ascertain what kind of a genre the text is from. In the author's opinion, the purpose was achieved.

Seven morphologically disambiguated corpuses were used, every one had about 100 000 words and about 20 000 of these were verbs. Corpuses were fiction, fiction translated from English, newspaper texts, legal texts, texts from a scientific magazine, oral speech and chat rooms' texts. It was a quantitative research. Practical aim was to find categories of verbs that could work as a distinctive feature for a genre, theoretical aim was to compare different genres according to the frequencies.

The research questions were:

- how frequent are verb categories in different genres of text;
- whether and how much the observed categories differed in genres;
- is it possible to recognize from which genre the text is, according to the frequency of verb categories?

The first hypothesis was that verb categories' frequencies are significantly different and the second hypothesis was that by frequencies it is possible to assay what kind of text it is. All the questions were answered and the hypotheses were true.

The conclusion of this research was that it is possible to use the categories of verbs as a distinctive feature, although sometimes the frequency might represent text's author or speakers rather than the whole genre.

There are not many studies about verb categories and their frequencies in the Estonian language. This bachlelor’s thesis could be used as a criterional when characterizing some version of language.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kärt-Kristiin Jaagu,

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

„Verbi grammatiliste kategooriate esinemine eri tekstiliikides“ ,

mille juhendaja on Kadri Muischnek,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kärt-Kristiin Jaagu
24.05.2021