TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Linda Freienthal

PRONOMINAALSETE VIITESUHETE ANALÜÜS ASENDUSSÕNADE SUHTES KÄSITSI MÄRGENDATUD KORPUSES

Bakalaureusetöö

Juhendajad dotsent Kadri Muischnek ja Kristiina Vaik

Tartu 2018

SISUKORD

SISSEJUHATUS

Paljudele keeletehnoloogilistele vahenditele on lisaks morfoloogilisele ja süntaktilisele infole vaja ka infot sõnadevaheliste seoste kohta. Näiteks sorteerib automaatne sisukokkuvõtja (ingl *text summarization*) tekstist välja informatiivsemad laused. Aga mis saab siis, kui kokkuvõttesse valitud lausetes on kõik alused ja sihitised pronoomenid ning lugeja ei tea, kellest või millest on jutt? Või kuidas teab masintõlge (ingl *machine translation*), kas eestikeelne *tema* viitab nais- või meessoost isikule, tõlkides teksti taolist infot vajavatesse keeltesse nagu inglise ja vene? Samuti, kuidas automaatsel infoeraldamisel (ingl *information extraction*) teada, missugused sõnad viitavad ühele ja samale asjale pärismaailmas? (Mitkov 2004: 275–276)

Teadmust pronoomenite tegelike tähenduste kohta saab arvutile anda asendussõnade automaatne lahendaja, mida keeletehnoloogias nimetatakse tüüpiliselt anafooride lahendajaks. Asendussõnade automaatne lahendaja loob allpoolses näites sõnade *ta* ja *Linda* vahele automaatselt viitesuhte. Selles viitesuhte näites on *ta* asendussõna ning *Linda* tema viitealus ehk see, millele asendussõna viitab või mida asendussõna asendab.

Linda on üliõpilane. *Ta* elab Tartus.

Selliste viitesuhete osas märgendatud sisendtekstide abil on masintõlkeprogrammil olemas teave viitealuse soo kohta ning võimalus asendada isikupronoomen pärisnimega, sisukokkuvõtja võib asendada pronoomenid nende viitealustega, ning infoeraldamisel saab kontrollida, millistel sõnadel on tegelikult sama tähendus.

Anafooride automaatse lahendamise suure nõudluse tõttu on sellega järjepidevalt 70ndatest-80ndatest alates tegeletud (Mitkov 2004: 277). Suurt huvi ja nõudlust illustreerivad näiteks vähemalt kaheksa korda toimunud kollokviumid „*Discourse Anaphora and Anaphor Resolution Colloquium“* (DAARC 2011), anafoori lahendamise teemal ilmunud artiklitekogumikud (Anaphora Processing… 2005; Anapora Resolution… 2016) ja suur hulk artikleid erinevate keelte anafooride lahendajate, nende meetodite või keerulisemate viitesuhete analüüsimise kohta.

Eesti keeles on asendussõnade automaatse lahendamisega tegelenud Pilleriin Mutso (2008) ja Tiina Puolakainen (2015), kuid nende tööd on jäänud katsetusteks ega ole eesti

keeletehnoloogias laialt kasutatavad või edasi arendatud. Rohkem ei ole töö autorile teadaolevalt eesti keelele loodud asendussõnade lahendajaid katsetatud. Kuna tegu on nii olulise keeletehnoloogilise abivahendiga, on vaja taoline lahendaja luua ka eesti keelele. See töö loobki aluse uue asendussõnade automaatse lahendaja valmimiseks.

Antud töö eesmärk on analüüsida selle korpuse põhjal pronoomenite *mina-meie, sina-teie, tema-nemad, kes, mis* ja *see-need* viitesuhteid ja leida tugevaid asendussõnade ja tema viitealus(t)e vahelisi seoseid, mida saaks kasutada tulevikus eesti keele pronominaalsete viitesuhete automaatse lahendaja loomisel. Selleks analüüsiti projekti „Sihipärane süntaks korpuse jaoks“ raames asendussõnade suhtes käsitsi märgendatud *ca* 107 000 tekstisõna suurust ajalehetekstide korpust[1]. Analüüsimise hõlbustamiseks lõi töö autor Pythonis kirjutatud programmi, mis kannab asendussõna ja tema viitealus(t)e morfoloogilised ja süntaktilised andmed paremini hallatavasse Exceli faili.

Töö koosneb kolmest peatükist. Esimeses peatükis süvenetakse keeleteaduse ja arvutilingvistika olulisemate viitesuhetega seotud mõistete definitsioonidesse. Samuti kirjeldatakse eesti keele pronoomeneid ja tutvustatakse põgusalt nende viitamisega seotud karakteristikuid. Esimese peatüki teises pooles kirjeldatakse, kuidas tavaliselt asendussõnu automaatselt lahendatakse ning tuuakse välja reeglid, millel põhines Mutso (2008) magistritööna valminud lahendaja.

Teises peatükis kirjeldatakse lähemalt töö aluseks olnud korpust, süüvitakse selle märgendamisreeglistikku ja märgendamise keerulisematesse tüüpjuhtudesse. Samuti kirjeldatakse analüüsimiseks loodud programmi algoritmi.

Kolmandas peatükis on toodud eesti pronominaalsete asendussõnade ja nende viitesuhete analüüs antud korpuse põhjal. Peatüki alguses on kirjeldatud parameetreid, mida töös iga pronoomeni puhul käsitleti. Sellele järgneb pronominaalsete viitesuhete analüüs ja parameetrite kirjeldus. Analüüsis võrreldakse tulemusi ka Mutso algoritmi reeglitega. Peatüki lõpus tuuakse välja analüüsi käigus leitud tähtsamad reeglid, mida saab tulevikus eesti keele pronoomenite lahendaja tegemisel ära kasutada.

[1] https://github.com/Lindafr/Asenduss6nadeAnalysaator

1. ASESÕNADE VIITEALUSTE TUVASTAMISE TEOREETILINE KÄSITLUS

1.1. Viitamise käsitlusi keeleteaduses

Viitamine on „keeleüksuse vastavusse viimine mingi reaalse või diskursusemaailma referendiga“ (Pajusalu 2017: 566). Referent on keeleväline olend, ese, omadus, tegevus, sündmus, tekst või tähistatav, millele viidatakse tervikuna. Midagi esimest korda mainides on tegu esmaviitamisega. Sidusat teksti kõneldes või kirjutades tuleb üldjuhul mainida ühte ja sama referenti aga mitu korda. Sel juhul on tegemist kordusviitamisega, mida võib jagada anafoorseteks ja katafooreteks. (Pajusalu 2017: 566–567)

Sõna *anafoor* tuleb kreeka keele sõnast *anaphora*, mis tähendab tagasitoomist või kordust. Keeleteaduses märgitakse sellega tagasiviidet või pronoomenit, „mis on samaviiteline mingi eespool esineva väljendiga” (VSL 2012). Näites 1 viitab anafoorne pronoomen *ta* sõnale *kass*. Sõna *kass* on selle pronoomeni viitealus (ingl *antecedent*) ehk see, millele pronoomen viitab. Anafoorile vastandub edasiviide ehk katafoor (Pajusalu 2017: 567). Näites 2 on katafoorne sõna *selleks*, millele järgneb tema viitealuseks olev kõrvallause *et õigeks ajaks kohale jõuda*. Kui viitealus on viitajaga erinevas lauses, on tegu lauseülese viitamisega (ingl *intersentential*), samas lauses lausesisese viitamisega (ingl *intrasentential*) (Mitkov 2002: 14–15). Lauseülest viitamist illustreerib näide 1, lausesisesest viitamist näide 2.

(1) *Kass* lamab päikeselaigus. *Ta* nurrub.

(2) *Selleks, et õigeks ajaks kohale jõuda,* asusime varakult teele.

Anafoorseid ja katafoorseid viitajaid on nende sõnaliigi alusel erinevalt liigitatud. „Võõrsõnade leksikon“ nimetab anafooriks tagasiviidet või tagasiviitavat pronoomenit (VSL 2012). „Eesti keele süntaksis” defineerib Renate Pajusalu (2017: 567) anafoori ja katafoori kui pronominaalset (nt *mina, tema*) või proadverbiaalset (nt *siis, millal*) viidet eelnevale/järgnevale samaviitelisele nimisõnafraasile. „Eesti keele grammatika II“ (EKG II 1993: 198, 207) annab anafoori ja katafoori viitajatele üldnimetuse *asendussõnad* ning jaotab neid pronoomeniteks, proadverbideks ja proadverbidena toimivateks verbideks (nt *tegema, juhtuma*). Siin ja edaspidi kasutataksegi sõna

asendussõna viitaja ehk viitava keelendi tähenduses. Seega on see sõna peatükis 3 sõnaga *pronoomen* võrdne, kui viimase all mõeldakse anafoorset või katafoorset pronoomenit.

Anafoor tähendab eesti keeles enamasti nähtust (tagasiviide), VSL (2012) järgi ka tagasiviitavat sõna (pronoomen). Tagasiviite ja tagasiviitaja mõlema anafooriks nimetamine võib tuleneda sellest, et inglise keeles kasutusel olevad kaks sarnast sõna on eesti keeles sulandunud üheks sõnaks. *Anaphora* on tagasiviitamine, *anaphor* tagasiviitaja ehk tagasiviitav sõna või üksus. (Mitkov 2004: 266–267) Käesolevas uurimistöös kasutatakse sõna *anafoor* üldjuhul tagasiviite, st nähtuse tähenduses.

Keeletehnoloogias kuuluvad anafoorsete viitajate alla kõikvõimalikud sõnad või fraasid, mis viitavad tagasi millelegi tekstis esinevale või omavad viidatavaga sama referenti (ehk keelevälist tähistatavat) (Mitkov 2004: 266–267). Ruslan Mitkov (2002: 8–14; 2004: 268–269), kes uurib viitesuhteid selleks, et neid automaatselt tuvastada, jaotab tagasiviited[2] anafoorivormi järgi pronominaalseks viiteks (ingl *pronominal anaphora*), nimisõnafraasiviiteks (ingl *lexical noun phrase anaphora*), verbianafooriks (ingl *verb anaphora*) ja nullanafooriks (ingl *zero anaphora*).

Pronominaalne viide on näites 3, kus asendussõnaks on pronoomen *teda*. Nimisõnafraasiviite näide on sama lause fraas *kadunud poiste juht*. Mõlemad viitavad *Peeter Paanile* ehk viitealusele. Verbianafoori illustreerib näide 4, kus verb *tegi* viitab verbifraasile *puges külje alla*. Nullanafoori illustreerib lause 5. Seal on anafoor ära jäetud (*Ø* märgend näitab, kus anafoor olema peaks). Kuna seda saab öeldise pöördevormi abil tuvastada, on tegu grammatilise ellipsiga (väljajätt, mis on kergesti leitav grammatika abil (Erelt 2017a: 591)).

(3) *Peeter Paani* nähti jälle. *Kadunud poiste juht* lendas eile öösel Eikunagimaa poole, *teda* jälitas kapten Konkskäsi.

(4) Kati *puges* talle *külje alla*, nagu *tegi* ka Maali.

(5) *Anni* jooksis eile kaua ja *Ø [= tema]* on nüüd väsinud.

Lisaks viitajavormile ja viitealuse asukohale võib viitesuhteid jagada ka asendussõna ja viitealuse samasuse järgi. EKG II (1993: 198) jagab viitesuhted kolmeks:

[2] Mitkov katafooridega ei tegele.

(a) viitealus tähistab asendussõnaga „sama objekti või nähtust“ (vt näiteid 1, 3 ja 5),
(b) viitealus tähistab „samaliigilist, ühenimelist objekti või nähtust“ (vt näide 6),
(c) viitealus tähistab „samalaadset tunnust, seisundit, tegevust, protsessi“ (vt näiteid 4 ja 7).

(6) Ostsin oma õele *kontserdipiletid,* aga tal olid *need* juba olemas.
(7) Tahtsin täna perele *suppi keeta,* aga Sina juba *tegid seda*.

Mitkov (2004: 266–269) jagab aga viitesuhted samasuse alusel kaheks. Esimesel juhul (kui asendussõnal ja viitealusel on sama referent ehk keeleväline tähistatav) nimetatakse anafoori ja tema viitealuse viitesuhet samaviiteliseks (ingl *coreferential)* (EKG II liik a). Muudel juhtudel (EKG liigid b ja c), kui asendussõnal on viitealus, aga neil puudub ühine referent, on tegu samatähendusliku viitega (ingl *identity-of-sense anaphora).*

Kõik ülaltoodud näited on otsese tagasiviitamise (ingl *direct anaphora*) näited, kus asendussõna ja viitealuse vaheline seos on üks järgnevatest: samasus (ingl *identity*), sünonüümsus (ingl *synonymy*), üldistus (ingl *generalization*) või spetsialiseerumine (ingl *specialization*). Kaudse tagasiviitamise (ingl *indirect anaphora*) alla kuuluvad sellised seosed, kus asendussõna referent on osa mingist tervikust või kuulub mingisse määratud hulka ja viitealus ongi see tervik või määratud hulk. (Mitkov 2004: 269) Näites 8 kuulub Merlyn Uusküla kui bändiliige bändi Nexus hulka ja on seega osa bändist ning viitab kaudselt bändile kui määratud hulgale.

(8) *Nexus* tuleb taas kokku, *Merlyn Uusküla* rääkis juba esimestest bändiproovidest.

Käesolev töö keskendub eesti keele pronoomenitele automaatselt otseste viitealuste leidmisele ega käsitle muid asendussõnade liike. Seepärast keskendub järgmine peatükk eesti keele pronoomenitele.

1.2. Eesti keele pronoomenid viitamise kontekstis

Pronoomen ehk asesõna on käändsõna, mis käitub nagu nimi-, omadus- või arvsõna, aga on täistähenduslike sõnadega võrreldes leksikaalselt sisuvaene ja abstraktse tähendusega

(EKG I 1995: 26; Erelt 2017b: 59). See tähendab, et need sõnad saavad endale tähenduse kontekstist. Pronoomeneid võib jagada käändsõnade liigituse alusel (EKG I 1995: 27; Erelt 2017b: 59):

- asenimisõnadeks ehk prosubstantiivideks (nt *mina, sina, tema, see, kes, mis*),
- aseomadussõnadeks ehk proadjektiivideks (nt *missugune, selline*),
- asearvsõnadeks ehk pronumeraalideks (nt *mitmes*).

Pronoomeneid võib liigitada ka nende funktsiooni alusel. Üks pronoomen võib esineda mitmes erinevas rollis (st funktsioonis). EKG I (1995: 27–31) jagab pronoomenid nende funktsiooni alusel kaheksaks. Järgnevalt EKG I jaotus koos näidetega:

- demonstratiivpronoomenid ehk näitavad asesõnad (pronoomen viitab asjale või tunnusele, mis on mõistetav vaid situatsioonis ja kontekstis, näiteks *sama, muu, toosama* või pronoomen *selle* lauses *võtsin laualt raamatu ja panin selle kotti*);
- determinatiivpronoomenid ehk määratlevad asesõnad (pronoomen tõstab nimisõna esile või rõhutab seda, näiteks *oma, igaüks, kõik, kumbki, terve* või pronoomen *ise* lauses *tahtsin sellega ise tegeleda*);
- indefiniitpronoomenid ehk umbmäärased asesõnad (pronoomeni viidatav pole konkreetne, näiteks *mingi, mingisugune, ükski, mitmes, mõni, paljud* või pronoomen *ühtki* lauses *ma ei leidnud endale ühtki sobivat lõhnaõli*);
- interrogatiiv-relatiivpronoomenid ehk küsivad-siduvad asesõnad (pronoomen on kõrvallause sidend või küsimustab väitlauset, näiteks *kes, mis, missugune, mitmes* või *milline* lauses *see, milline on Sinu reaktsioon, pole minu teha*);
- personaalpronoomenid ehk isikulised asesõnad (pronoomen viitab kõnelejale, kuulajale või mõnele muule isikule, näiteks *mina, tema* või pronoomen *sind* lauses *otsisin sind taga*);
- possessiivpronoomenid ehk omastavad asesõnad (pronoomen väljendab millegi omamist tegevussubjekti poolt, näiteks *iseenda, omaenda, enese* või pronoomen *oma* lauses *ta ei rääkinud oma saladusest kellelegi*);
- refleksiivpronoomenid ehk enesekohased asesõnad (pronoomen viitab tegijale, kellele tegevus on suunatud, näiteks *enese, iseenda* või pronoomen *omale* lauses *nad võtsid omale lemmiklooma*);

- retsiprookpronoomenid ehk vastastikused asesõnad (pronoomen viitab omavahel vastastikku toimivale kahele või enamale tegevussubjektile, näiteks *üksteise* või pronoomen *teineteist* lauses *nad armastasid teineteist*).

Kõik ülaltoodud pronoomenid ei ole viitavad (nt indefiniitpronoomen *mingisugune*) ja seega antud töös olulised. Renate Pajusalu (2009: 122) jagab eesti keele pronominaalse viite asendussõnad neljaks: personaalpronoomenid, demonstratiivpronoomenid, possessiivpronoomenid ja pronumeraalid.

Personaalpronoomeneid ehk isikulisi asesõnu on eesti keeles kokku kaksteist: *mina, sina, tema* ainsuses ja *meie, teie, nemad* mitmuses ning nende lühikesed vormid. Kui lauseõhk satub pronoomenile või pronoomenit kasutatakse vastandusena mõnele muule isikule, kasutatakse üldjuhul pikka varianti. (Pajusalu 2009: 123)

Eesti keele kõneleja intuitsioon võib pidada personaalpronoomenit *tema/ta* ilma kontekstita automaatselt ainult elusale referendile ja demonstratiivpronoomenit *see* ainult elutule referendile viitajaks. Tegelikkuses ei ole nende pronoomenite kasutus nii rangelt määratud. *Tema* viitab üldjuhul elusale referendile, s*ee* elutule, väga abstraktsele referendile. *Ta* on nende kahe pronoomeni vahel: võib olla nii elus kui ka elutu, viimasel juhul konkreetsem kui *see*. Kui kõnes viidatakse kahele eluta või kahele elus referendile, peab nendele viitavaid asendussõnu kuidagi eristama. Sel juhul viidatakse tavaliselt põhireferendile (fookuses, esile tõstetud, põhiteemana olev või esimesena mainitud referent) pronoomeniga *ta* ning teisele referendile pronoomeniga *tema* või *see*. (Pajusalu 2005: 112–117)

Viitavad demonstratiivpronoomenid on *see* ja *too*. Üle-eestiliselt on *too* harv nähtus ning seda kasutatakse rohkem Lõuna-Eestis. Sõna *see* võib edasi anda kõhklust, määratleda definiitsust või viidata millelegi. (Pajusalu 2009: 123) Käesoleva töö materjalis sõna *too* ei käsitletud, kuna korpuses polnud seda märgendatud.

Eesti keelt peetakse traditsiooniliselt artiklita keeleks. Siiski võivad demonstratiivpronoomenid *see* ja *too* esineda definiitsete määratlejate (ingl *definite determiners*) ehk artiklite rollis. Kuna nende pronoomenite kasutamine artiklina ei ole kohustuslik, ei peeta eesti keelt artikliga keeleks. (Pajusalu 1997: 146, 172–173;

Pajusalu 2009: 130) Pronoomeni *see* artiklilaadsus oli korpuse märgendamisel keeruline väljakutse (vt lähemalt ptk 2.1.2).

Mitte iga kord ei viita pronoomen kindlale referendile, mõnikord võib tegu olla hägusa rühmareferendiga. Hägusat rühmareferenti saab vaadelda kui ebamääraste liikmetega tervikut. Teisisõnu, viidatakse rühmale, mille tegelikke liikmeid ja arvu ei tuvastata, kuid millele saab hiljem viidata näiteks sõnadega *nemad* või *meie*. Hägusat rühmareferenti annavad edasi impersonaalsed verbivormid (viitab, et kõneleja ei kuulu rühma) ja isikuta konditsionaalivormid (kõneleja viitab endale). Pronoomenitest viitab hägusale rühmaelemendile näiteks pronoomen *kõik*. (Pajusalu 2017: 586) Näites 9 ei ole võimalik luua täisloetelu õitsema hakkavatest referentidest, *kõik* viitab hägusale rühmareferendile.

(9) Kevadel hakkab *kõik* õitsema.

Eesti keele viitavate pronoomenite ja nende referentide suhet ning kasutust on uuritud palju keeleteaduslikust vaatenurgast. See info aitab ennustada, mida asendussõnade lahendamisel oodata tasub, kuid pole keeletehnoloogias nii oluline ja kasutatav, kui võiks eeldada. Asendussõnade lahendajate jaoks analüüsitavaid tunnuseid ning kasutatavaid meetodeid kirjeldab järgnev alapeatükk.

1.3. Asendussõnade automaatne lahendamine

Selle alapeatüki esimene pool kirjeldab erinevaid lähenemisviise asendussõnade automaatsele lahendamisele ja lühidalt Eestis tehtud tööd. Teine pool kirjeldab eesti keelele loodud pronominaalsete anafooride lahendaja algoritmi.

1.3.1. Erinevad lähenemisviisid asendussõnade automaatsele lahendamisele

Asendussõnade automaatne lahendamine ehk asendussõnade automaatne märgendamine tähendab tekstist automaatselt asendussõnade ja nende viitealuste tuvastamist ning nende vahele seose loomist (ehk seose üles märkimist). Teisisõnu, arvuti lahendab ehk märgib üles, mida tähendab ehk millele viitab tekstis iga konkreetne *tema* ja *meie* või mõni muu asendussõna (vt lähemalt ptk 1.1), ehk lahendab need.

Üldjuhul keskendutakse vaid anafooride lahendamisele (ingl *anaphora resolution*), jättes katafoorid tähelepanuta. Katafooride lahendamine oli vahepeal jäänud lausa sedavõrd tähelepanuta, et 2002. aastal toimunud neljandas „*Discourse Anaphora and Anaphor Resolution Colloquium"* nimelises kollokviumis tõstatasid Reuland ja Avrutin (2005) katafooridele kui „tagurpidiste anafooridele" (ingl *backward anaphora*) viidates üles teema, millega lootsid innustada teisigi katafoorseid viitesuhteid taas uurima. Ise uurisid nad inglise ja hollandi keele katafoore.

Anafooride automaatseid lahendamisviise on erinevaid ja neid võib nende strateegiate alusel jagada kaheks (Mitkov 1999: 1, 6):

- traditsioonilised lähenemisviisid (reeglipõhised, mis sorteerivad kandidaatsõnade hulgast välja viitealuse) ja
- alternatiivsed lähenemisviisid (viitealus leitakse näiteks statistika vms mitte reeglipõhise tehnika abil).

Reeglipõhised anafooride lahendajad annavad viitealuse leidmiseks kandidaatsõnadele punkte erinevate reeglite põhjal või just vähendavad punktisummat. Hinnatakse näiteks viitealuse ja anafoori ühildumist soos ja arvus, süntaktilist sarnasust ja kandidaatsõna korduvust lõigus. Need reeglid võivad olla välistavad (ingl *eliminating*) või eelistavad (ingl *preferential*). Välistavad reeglid välistavad kandidaatsõnu (viskavad kandidaatide nimekirjast välja ja edaspidi neid ei vaadatagi) ning eelistavad reeglid lisavad kandidaatsõnadele plusspunkte (tõstavad kandidaatsõna viitealuseks olemise tõenäosust, viitealuseks valitakse kõrgeima punktisummaga kandidaat, võrdsete punktide korral lähim). Vastavalt sellele, kas algoritm põhineb pigem välistavatel või eelistavatel reeglitel, võib anafooride traditsioonilisi lahendajaid jagada nii eelistavateks kui ka välistavateks lahendajateks. (Mitkov 1999: 3–4, 6)

Ülaltoodud reeglite näited kasutavad ainult süntaktilist, morfoloogilist jms infot, jättes kõrvale semantilise info reaalse maailma kohta. Selliseid lähenemisi kutsutakse teadmistevaesteks lähenemisteks (ingl *knowledge-poor approaches*). Süntaktilisest ja morfoloogilisest infost aga ei piisa näiteks näidete 10 ja 11 pronoomeni *ta* viitealuse tuvastamiseks. Nendes näidetes on kandidaatsõnad *Bill* ja *John* morfoloogiliselt võrdsed, süntaktiliselt saab kõrgemad punktid aluse funktsioonis olev *John*, mis näite 10 puhul on

semantikale tuginedes vale. Viitealuse saab nendes näidetes järeldada teadmiste põhjal reaalse maailma kohta. Sellist lähenemisviisi, mis kasutab semantilist infot ja teadmisi päris maailma kohta, kutsutakse teadmistel põhinevaks lähenemisviisiks (ingl *knowledge-based approach*). (Lappin 2005: 4–6; Mitkov 2002: 30–32) Näited 12 ja 13 illustreerivad samuti olukorda, kus õigeks lahendamiseks on tarvis semantikat, kuna ilma selleta ei pruugi arvuti vaid morfoloogia, kauguse ja süntaksi abil tuvastada, et esimese näite *selle* viitealuseks on *arvutist* ja teise näite *selle* viitealuseks on *disketi*.

(10) John peitis Billi võtmed ära. *Ta* oli purjus. (Lappin 2005: 7)

(11) John peitis Billi võtmed ära. *Ta* mängis talle vingerpussi. (Lappin 2005: 7)

(12) Vincent eemaldas disketi arvutist ja lülitas *selle* siis välja. (Mitkov 1999: 5)

(13) Vincent eemaldas disketi arvutis ja siis kopeeris *selle*. (Mitkov 1999: 5)

Loomuliku keele töötluse (ingl *natural language processing*) süsteemid eelistavad pigem ressursiliselt odavaid ja kiireid anafooride lahendajaid, mistõttu eelistatakse ka teadmistevaeseid lähenemisi. Teadmistevaesed lähenemised (sinna kuuluvad näiteks traditsioonilised reeglipõhised ja ka alternatiivsed masinõppe lähenemisviisid) on robustsemad ja kiiremad just seetõttu, et vajavad vähem lingvistilist taustinfot ja seetõttu ka vähem teksti eeltöötlemist. Mitkov väidab, et isegi ilma süntaktilise analüüsita on võimalik saada hästi töötav anafooride lahendaja. (Mitkov jt 2007: 180; Mitkov 2002: 145) Lappin (2005: 6) nõustub, et teadmistevaesed lähenemised on kiiremad, robustsemad ja tehnilises mõttes odavamad, kuid lisab, et teksti on siiski vaja süntaktiliselt ja morfoloogiliselt analüüsida ning masinõpe nõuab mahukat käsitsi märgendatud korpust, st palju töötunde. Samuti räägib teadmistevaeste lähenemiste vastu suutmatus lahendada olukordi, mida illustreerisid näiteid 10 ja 11 ning 12 ja 13.

Lisaks küsimusele, millist teadmust ja infot teksti kohta lahendajat luues kasutada (kas läheneda teadmistevaeselt või teadmisel põhinevalt; võtta arvesse ka süntaksit või vaadata ainult sõnaliiki, kaugust ja viitealuse kandidaadi morfoloogiliste tunnuste ühildumist asendusssõnaga), on vaja ka otsustada, millist meetodit kasutada. Nii traditsioonilised kui ka alternatiivsed lähenemisviisid nõuavad palju eeltööd. Ühed põhjalikke analüüse reeglite koostamisel (nagu selle töö puhul), teised suurte korpuste märgendamist. Küsimus on pigem lahendaja tulemi kvaliteedis ehk lahendaja efektiivsuses. Lee jt (2017: 2, 6–7) leiavad, et reeglipõhiseid meetodeid ei tasu suures masinõppe tuhinas

iganenuks ja ebaefektiivseks kuulutada: nad leidsid, et teatud juhtudel töötavad reeglipõhised lähenemised väga hästi. Seda eriti juhul, kui on vaja kasutada vähe parameetreid ning kontekst on ette määratav (näiteks kõnevoorude vahetused dialoogidega tekstides). Samas ei tule need eriti hästi toime suure hulga leksikaalse informatsiooniga. Siin tulevad nende sõnul appi masinõppel põhinevad süsteemid, mis saavad hakkama suure hulga parameetritega. Lee jt (2017) kirjeldavadki oma artiklis „A Scaffolding Approach to Coreference Resolution Integrating Statistical and Rule-based Model" reeglipõhise lähenemisviisi ja masinõppel põhineva lähenemisviisi hübriidi ja selle katsetamist inglise keeles.

Eestis on pronominaalsete anafooride automaatse lahendamisega tegelenud nii Pilleriin Mutso kui ka Tiina Puolakainen, mõlemad teadmistevaeselt ja reeglipõhiselt. Puolakainen (2015) kasutas oma töös kitsenduste grammatikat (ingl *constraint grammar*) ning suutis ajalehetekstides lahendada 70–79% pronoomenitest. Tema lahendaja otsis nii anafoorseid kui ka katafoorseid viitesuhteid: asendussõna viitealuse kandidaate otsiti samast lausest või 7 lauset eest- või tagantpoolt. Mutso töö (2008) käsitleb aga ainult anafoorseid viitesuhteid. Järgnev peatükk kirjeldab lähemalt tema tööd.

1.3.2. Eesti keelele kohandatud Mitkovi teadmistevaene anafooride lahendaja

Pilleriin Mutso lõi oma magistritöö „Knowledge-poor Anaphora Resolution System for Estonian" (2008) raames pronoomenite *tema* ja *nemad* automaatse reeglipõhise lahendaja Mitkovi teadmistevaese anafoori lahendaja ja selle edasiarenduse MARS (Mitkov 2002) põhjal. Mitkovi lahendajat on lisaks eesti keelele kohandatud ka poola, araabia, prantsuse ja bulgaaria keelele. (Mitkov 2002: 153–157, 172)

Originaalis katsetas Mitkov oma teadmistevaest reeglipõhist lahendajat ja selle edasiarendust MARS inglisekeelsetel tehnilistel manuaalidel (Mitkov 2002: 152–153, 169). Mutso kohandas Mitkovi töö eesti keelele eesti morfoloogiliselt ühestatud korpuse põhjal, millest leiab juriidilisi, teadus-, ilukirjandus- ja ajalehetekste, ning eesti keele süntaktiliselt ühestatud korpuse põhjal, millest leiab ilukirjandus-, teadus- ja ajalehetekste. Need korpused kattuvad osaliselt. (Mutso 2008: 22–23)

Esimese ja teise pöörde isikupronoomeneid Mutso ei käsitle, kuna leiab, et nende pronoomenite puhul võib vaja minna tekstivälist ehk semantilist infot, mida pole

korpustes, millel ta töö põhineb. Samuti leiab ta, et kolmanda pöörde isikupronoomeneid leidub korpuse žanrites (ajalehe- ja teadustekstides) rohkem kui esimese ja teise isiku omi. Lisaks lahendab ta pronoomeneid *tema* ja *nemad* Mitkoviga sarnaselt vaid anafoorselt (see tähendab, et ta ei vaata asendussõnale järgnevaid sõnu). (Mutso 2008: 22, 29).

Mitkovi lahendaja kohandamine mingile keelele tähendab Mitkovi lahendajate reeglite ümberhindamist, eemaldamist või neile uue reegli lisamist vastavalt sihtkeele eripäradele. Ka Mutso valis ja kohandas Mitkovilt saadud indikaatoreid ehk reegleid vastavalt eesti keele eripäradele ja tema töö aluseks olevate korpuste võimalustele ning žanritele. Enne kandidaatsõnade hindamist vastavalt reeglitele ehk indikaatoritest läbi laskmist järgib Mutso (2008: 32) Mitkovi algoritmi (2002: 146) ja filtreerib välja vaid need kandidaatsõnad, mis ühilduvad oma asendussõnaga arvus[3]. Järgmises loetelus on toodud Mutso lahendaja indikaatorid (2008: 33–37) ja neile vastavad punktijagamise süsteemid, mis on kas otse või väheste muudatustega võetud Mitkovi lahendajast ja selle edasiarendusest MARS.

- Sageduse indikaator ehk leksikaalse korduvuse indikaator (ingl *frequency indicator, lexical reiteration*). Kui kahe pealkirja vahel olevas tekstis on kandidaatsõna mainitud kaks korda, siis lisatakse sellele sõnale üks punkt. Kui on mainitud rohkem, kui kaks korda, siis antakse kaks punkti. Eesti keele käänderohkuse tõttu vaatab Mutso kandidaatsõnade korduvust nende tüvede esinemissageduse järgi.
- Mustri indikaator (ingl *pattern indicator*). Kui nii kandidaatsõna kui ka asendussõna mõlemad eelnevad või järgnevad verbile, siis saab kandidaatsõna ühe punkti. Niimoodi saab kätte ajalehetekstides sagedasti esinevad otsekõne alustavad või lõpetavad fraasid jutumärkide vahetus läheduses nagu *räägib Mari* või *Juhan hüüatas*.
- Viitekauguse indikaator (ingl *referential distance indicator*). Kandidaadid, mis asuvad asendussõnaga samas liitlauses, saavad kaks punkti. Asendussõna lausele eelnevas lauses olevad kandidaadid saavad ühe punkti.

[3] Mitkovi algoritmis läbivad kandidaatsõnad lisaks arvulise ühildumise filtrile ka soos ühildumise filtri (Mitkov 2002: 146). Seda ei ole ilmselgelt vaja eesti keeles teha, kuna eesti keeles pole sugusid.

- Indikatiivse verbi või keelendi indikaator (ingl *indicative verbs indicator*). Kandidaatsõna saab kaks punkti, kui see eelneb või järgneb mingile sõnale valitud sõnade listist. Selles listis on kõneaktiverbid ja muud otse- või kaudkõne saatelauses esinevad keelendid nagu *möönis*, *ütles* ja *hinnangul*, mille vahetus naabruses on üldjuhul esilduvad sõnad nagu näiteks nimed fraasides *Liis möönis* ja *Juhani hinnangul*.
- Pronoomeni indikaator (ingl *boost pronoun indicator*). Pronoomenitest kandidaatsõnad saavad ühe punkti. Mitkov loodab selle indikaatoriga kätte saada pronoomenite viiteahelad, mille kaudu saaks kätte muidu lahendaja tegevusraadiusest väljas oleva viitealuse (Mitkov 2002: 165-166). See tähendab, et kui pronoomeni ta_1 viitealus ei ole viitealuse otsimisvahemikus, kuid seal leidub kandidaadina mõni teine pronoomen, näiteks ta_2, siis viidatakse sellele. Ta_2 kaudu saab kätte tema viitealuse, mis võib ühtlasi olla ka ta_1 viitealus. (Mitkov 2004: 165–166 põhjal)
- Nime indikaator (ingl *name indicator*). Mutso kogus listi kõik korpuses esinenud pärisnimed, mis esinesid kõrvuti mingi teise pärisnimega (nii saab kätte näiteks inimese ees- ja perekonnanime või mitmesõnalise institutsiooni nime). Kui kandidaat esineb selles listis, antakse talle juurde kaks punkti.
- Käände indikaator (ingl *declination indicator*). Nimetavas käändes kandidaadid saavad kaks punkti, osastavas käändes kandidaadid ühe punkti. Oma asendussõnaga käändeliselt ühilduvad viitealuse kandidaadid saavad samuti ühe punkti.
- Otsekõne indikaator (ingl *quotation indicator*). Kui asendussõna on otsekõnes (jutumärkide vahel) ja tema viitealuse kandidaatsõna mitte, siis võetakse kandidaadilt kolm punkti ära, kuna suure tõenäosusega on nii asendussõna kui ka tema viitealus mõlemad otsekõnes või sellest väljas.

Mutso katsetas veel kolme indikaatorit (2008: 33–36), mis lõppversioonist siiski erinevatel põhjustel välja jäid:

- Antuse indikaator (ingl *givenness indicator*). Kui asendussõna asub liitlauses, siis lisatakse selle lause esimesele nimisõnafraasile üks punkt. See reegel osutus ebatõhusaks peamiselt eesti keele üsna vaba sõnajärje tõttu.

- Süntaktilise ühildumise indikaator (ingl *syntactic parallelism indicator*). Kui nii kandidaatsõna kui ka tema asendussõna on mõlemad kas aluse või sihitise funktsioonis, siis saab kandidaatsõna endale ühe punkti. Mutso lahendaja efektiivsus tõusis 0,5% võrra pärast indikaatori välja jätmist.
- Pealkirja eelistamise indikaator (ingl *section heading indicator*). Pealkirjades esinevad kandidaadid saavad ühe punkti. See indikaator jäi välja ebatõhususe tõttu.

Kui kõik kandidaatsõnad on kõigi indikaatorite poolt hinnatud ja oma punktid kätte saanud, siis korrutatakse asendussõnale eelnevas lauses olevate kandidaatsõnade tulemus 0,75-ga ja sellele eelnevas lauses olevate kandidaatsõnade tulemus 0,5-ga. Asendussõnaga samas lauses olevad kandidaatsõnad jäävad samaks (mõtteliselt võib need läbi korrutada ühega). Seejärel määratakse viitealuseks kõrgeima tulemusega kandidaatsõna. Kui esikohal on mitu sõna, siis valitakse asendussõnale lähim kandidaat. (Mutso 2008: 29–31)

Mutso lahendaja võtab kandidaatsõnadeks vaid substantiivid ja pronoomenid ega vaata muid sõnaliike. Ta leidis, et vaid 1,5% viitesuhetest jäid välja tema lahendaja otsinguulatusest, milleks oli kolm lauset: asendussõnaga sama lause ja kaks sellele eelnevat lauset. Üleüldiselt jäi tema lahendaja edukus alla 74%, mis on siiski võrdlemisi hea tulemus, võrreldes teiste keelte lahendajate tulemustega. (Mutso 2008: 29, 40, 50)

2. MATERJAL JA ANALÜÜSIMIST HÕLBUSTAV PROGRAMM

See peatükk koosneb kahest osast. Esimene pool kirjeldab asendussõnade suhtes käsitsi märgendatud korpust, millel põhineb töö autori analüüs järgmises peatükis. Selles osas kirjeldatakse lähemalt ka märgendamisreegleid, millele märgendajad korpuse loomisel tuginesid, ning märgendamisel tekkinud keerulisi kohti ja nende lahendusi. Peatüki teises pooles kirjeldatakse korpuse analüüsimise hõlbustamiseks loodud programmi, selle algoritmi ja tulemit (Exceli faili).

2.1. Asendussõnade suhtes käsitsi märgendatud korpus

Aastatel 2015–2017 viidi läbi projekt „Sihipärane süntaks korpuste jaoks" (Sihipärane süntaks…), mille eesmärk oli parendada eesti keele automaatset süntaktilist analüüsi ja edendada korpuste märgendamist. Selle projektiga rahastati ka anafooride suhtes märgendatud eesti keele sõltuvuspuude panga loomist, millega alustati 2016. aasta sügisel. Töö kestis 2017. aasta detsembrikuuni. Selle töö autor osales projektis ühe märgendajana. Kokku märgendati *ca* 107 000 tekstisõna suurune ajalehetekstide korpus. Iga artiklit märgendas kaks inimest, seejärel ühtlustati märgendamiserinevused vastavalt töö käigus loodud märgendamisreeglitele. Peamine ühtlustaja oli käesoleva töö autor.

Märgendajatele anti ette ajalehetekstide korpus, kus olid automaatselt esile tõstetud järgnevad pronoomenid kõigis käändevormides:

- isikulised asesõnad *mina, sina, tema, meie, teie* ja *nemad;*
- küsiv-siduvad asesõnad *kes* ja *mis*;
- näitav asesõna *see* ja *need*.

Korpuses on paar korda esile tõstetud ka pronoomenid *missugune, mitmes* ja *selline,* kuid need jäävad käsitlusest välja, sest esimesed kaks on esile tõstetud vaid paaris failis, millega katsetati korpuse märgendamissüsteemi, ning viimane on pronoomenina esile tõstetud tõenäoliselt märgendaja apsuna vaid ühel korral.

Märgendaja ülesanne oli leida ära märgitud pronoomenitele viitealus ning luua selle pronoomeni ja viitealuse vahele seos. Tähelepanu tuleb pöörata sellele, et märgendades polnud oluline, kas tegu on anafoorse või katafoorse pronoomeniga. See tähendab, et märgendaja võis vajadusel luua viiteseose ka pronoomenile järgnevale sõnale, kuigi korpust nimetatakse erialaslängist tulenevalt anafooride suhtes märgendatud eesti keele sõltuvuspuude pangaks. Seega on sõna *anafoor* nii korpuses kui ka programmis asendussõna, täpsemalt pronominaalse asendussõna tähenduses, või siis lihtsalt viitamise (mitte ainult tagasiviitamise) tähenduses. Viiteseoseid võis luua nii lausetesiseselt kui ka lauseteüleselt. Ajalehenumbrid tükeldati failideks automaatselt sõnade arvu järgi, seetõttu võis üks ajaleheartikkel jaguneda mitme faili vahel ning asendussõnade seosed viitealustega ei ole täielikud, kuna mõni viitealus võis jääda teise faili. Märgendajad märgendasid vabavaralise tarkvara *brat* (Brat rapid…) abil.

Korpuse formaat järgib „Eesti keele sõltuvuspuude panga (EDT)“ (Muischnek jt 2014) formaati, kuhu lisati pronoomenite ja nende viitealuste märgendid. Lisa 1 illustreerib korpuse formaati. Seal on toodud korpusest leitud kolme pronoomeniga (pronoomenid alla joonitud) lause *puutüved on miljoneid aastaid vastu pidanud tänu* sellele*, et* neid *ümbritses kiht liiva,* mis *võis olla tekkinud mõnest tugevast liivatormist*. Selle lause viitealused on vastavalt *ümbritses*, *puutüved* ja *kiht*. Lisa 1 esimene rida *"<s id="36">"* märgib lause algust ja lausenumbrit (antud näitelause on failis 36. lause). Viimane rida *"</s>"* märgib lauselõppu. Märgend {Pronoomen} viitab pronoomenile, mida korpuses märgendati. Märgend {Viitealus} märgib viitealust. Märgend {Coref:36.12} näitab viiteseost. Esimene number näitab lausenumbrit, teine sõnanumbrit antud lauses. Näiteks viitab märgend {Coref:36.12} selle artikli 36. lause 12. sõnale. Mitme viitealuse korral on numbrid eraldatud komadega. Kui antud pronoomenil viitealust ei ole, siis {Coref:36.12}-märgend puudub, aga {Pronoomen}-märgend jääb alles.

Iga lause tekstisõna ja kirjavahemärgi analüüs on toodud kahel real, mida illustreerivad nii näide 14 kui ka lisa 1. Esimesel real on sõna lauses esinenud kuju. Teine rida algab vastava sõna lemmaga, millele järgneb morfoloogiline ja süntaktiline info. Nendel ridadel leiduvate lühendite selgitused leiab töö lõpust. Sõnanumber, mis näitab mitmes sõna see

lauses on, antakse edasi pärast märki #. Noolega viidatakse selle sõna numbrile, millele antud sõna allub. Kõige viimasena tuleb viitesuhete info.

(14) "<kiht>"
"kiht" L0 S com sg nom @SUBJ #13->12 {Viitealus}

Lisas 1 on kolm viitesuhet. Pronoomen *neid* viitab sõnale *puutüved*, pronoomen *mis* viitab sõnale *kiht* ja pronoomen *sellele* viitab sõnale *ümbritses*. Täpsemad viitealuste märgendamisreeglid, mille põhjal nii viidati, leiab järgmisest peatükist. Anafooride suhtes märgendatud korpuse võib leida veebikeskkonnast github.com[4].

2.1.1. Viitealuste märgendamisreeglid asendussõnade suhtes käsitsi märgendatud korpuses

Märgendamisreeglid loodi märgendajatele juhiseks ja järjepidevuse hoidmiseks, pidades silmas korpuse kasutusvõimalusi tulevikus. Järgnevalt on välja toodud viitealuste märgendamise põhireeglid.

Viitealusena märgendatakse alati üksiksõna. See nõue tuleneb sellest, et korpuse süntaktiline märgendus põhineb sõltuvussüntaksi põhimõtetel, mille järgi süntaktiline sõltuvussuhe on alati kahe sõnavormi vahel. See sõltuvussuhe on ebavõrdne: üks sõnavorm on ülemus ja teine selle alluv (vt sõltuvussüntaksi kohta lähemalt ka Muischnek, Müürisep 2016). Samuti on üksiksõnalisi viitealuseid anafooride lahendajal lihtsam lahendada kui mitmesõnalisi viitealuseid.

Sisuliselt võib viitealus olla sõna, fraas, osa- või täislause. Ka pronoomen ise võib olla viitealuseks. Juhul, kui viitealuseks on rohkem kui üks sõna, tuleb märgendades valida fraasipõhi. Näites 15 on fraasi *Neeme küla* fraasipõhjaks *küla* ja näites 16 on fraasi *väike tüdruk* fraasipõhjaks *tüdruk*. Valides viitealuseks viitealuse fraasipõhja, on vajadusel võimalik automaatselt kätte saada terve viitealuse fraas.

(15) Neeme *küla*, mille...

(16) Väike *tüdruk*, kes...

Kvantorfraaside nagu *kiht liiva* ja *kast tomateid* puhul on viitealuseks kvantor ehk hulgasõna (antud juhul sõnad *kiht* ja *kast*). Kui viitealusteks on koordinatsiooniseoses

[4] Korpuse aadress on https://github.com/EstSyntax/EstAnaphora.

olevad sõnad, siis märgendatakse iga viitealus eraldi, st ühel pronoomenil võib olla mitu koordinatsiooniseoses olevat viitealust. Näiteks lauses 17 viitab pronoomen *need* kolmele viitealusele: *luik*, *haug* ja *vähk*.

(17) *Luik, haug* ja *vähk* – need vedasid...

Asendussõnal võib olla mitu koordineeritud viitealust (vt näidet 17), kuid tal on alati üks ühe tähendusega viitealus. See tähendab, et mitmest võimalikust samatähenduslikust viitealuse variandist valitakse alati üks. Kui viitealused on võrdsed (üks ei ole teisest täpsem), siis lähtutakse kauguseprintsiibist: eelistatakse neid viitealuseid, mis on asendussõnale lähemal. Näites 18a–18c on asendussõnale *ta* sobivad viitealused nii näite 18a *donžuan* kui ka näite 18b *südametemurdja*. Kuna *südametemurdja* on asendussõnale lähemal, siis valitakse viimane viitealuseks.

(18) a) Kohalik *donžuan* on suur kalastushuviline.
b) See suur *südametemurdja* käib lausa igal pühapäeva hommikul kalal.
c) Mõnikord võtab *ta* oma naise ka kaasa.

Kui viitealus on terve osalause, siis märgendatakse viitealusena selle osalause öeldisverb. Näites 19 on pronoomeni *mis* viitealuseks märgitud verb *liigub*. See tähendab, et tegelik viitealus pronoomenile *mis* on terve osalause *maapind liigub Liibanoni ranniku lähedal*.

(19) Maapind *liigub* Liibanoni ranniku lähedal, *mis* tähendab...

Kui öeldiseks on ahelverb või verbi liitvorm, siis tuleb märgendada infiniitne komponent kui selle vormi kõrgeim ülemus. Kui näites 20 tahta viidata tervele näitelausele, siis tuleb viitealuseks valida sõna *teinud*.

(20) Ta oli *teinud* palju paksu pahandust.

Osalause kõrgeima ülemuse määramise põhimõtetest saab täpsemalt lugeda Kadri Muischneki ja Kaili Müürisepa artiklist „Eesti keele sõltuvuspuude pank ja selle keeleteoreetilised lähted“ (2016).

2.1.2. Märgendamisel esinenud probleemid ja nende lahendamine

Eelnevas peatükis toodud reegleid silmas pidades ei ole siiski võimalik kahel märgendajal üheselt märgendada, kuna ühele asendussõnale võis leiduda mitu erinevat ja sobivat viitealust. Selles alapeatükis kirjeldan märgendamisel ilmnenud keerulisi kohti ja nende lahendusi.

Referenti märgitakse tekstis tihti mitme erineva sõnaga. Näiteks viidatakse korpuses olevas artiklis ühele ja samale isikule fraasidega *härra Tamm, vanahärra* ja *endine metsamees*. Kuna selle korpuse põhjal loodud anafooride lahendaja üheks kasutusalaks võib olla automaatselt koostatud sisukokkuvõttes pronoomenite asendamine isikunimedega ning ühe referendiga asendussõnad märgendati alati ühe viitealusega, siis otsustati märgendada kõigi inimesele viitavate pronoomenite viitealuseks võimalusel alati isiku pärisnimi (antud juhul *Tamm*). Seetõttu on viitealus mõnikord asendussõnast kaugemal, kui mõni teine viitealuse kandidaat. Kui pärisnimi puudus, siis valiti lähim sobivatest viitealuse kandidaatidest (nagu näites 18a–18c).

Püsiväljendite koostises olevatele asesõnadele ei leidu tihti viitealuseid. Näiteks ei märgendata konstruktsiooni „mida X, seda Y". Samuti ei leidunud viitealust mõne üksiku erandiga ajaga seotud püsiväljendites nagu *sel aastal* ja *sellel suvel*. Samuti ei saa endale viitealust küsisõnade rollis olevad *kes* ja *mis*, kuna need ei ole anafoorsed. Näites 21 ei ole pronoomenil *mis* viitealust.

(21) *Mis* teeb inimesest dirigendi?

Pronoomenile *meie* antakse viitealused vaid siis, kui kõik *meie* alla kuuluvad referendid on tekstist kättesaadavad täisloeteluna. Näites 22 leiduvad pronoomeni *meie* alla kuuluvad kõik referendid täisloeteluna ja seega saab igale referendile eraldi viidata. Kui *meie* esineb hägusa rühmaelemendina (vt ptk 1.2), siis sellele üldjuhul viitealust ei leia (vahel leiab hägusale rühmaelemendile väga üldise tähenduse nagu *eestlased, eesti ühiskond*). Pronoomenit *meie* võib seostada grupi nimega, kui pronoomeni ja viitealuse vahetamine tähendust ei muuda. Näites 23 võib sõna *meie* asendada fraasiga *Eesti Ekspress*, seega on *meie* viitealuseks *Ekspress*.

(22) *Jüri*, *Mari* ja *mina* väitsime: „*Meie* pole seda teinud!"

(23) Eesti *Ekspress* väidab: „*Meie [= Eesti Ekspress]* pole seda teinud.”

Demonstratiivpronoomenit *see*/*need* on problemaatiline märgendada. See pronoomen ühendati viitealusega vaid siis, kui sai viidata kindlale sõnale, tekkis loogiline seos ja viitealusega asendamine andis loogilise lause.

Kui tekstis on toodud täisloetelu koos kokkuvõtva sõnaga (nt. *Ta sai oma lemmikõppeainetes – matemaatikas ja geograafias – ainult viisi*.), siis eelistati täisloetelu (*matemaatikas* ja *geograafias*). Kui viitealuseid leidub asendussõnast nii ees- kui ka tagapool, siis eelistatakse eespoolset (loetud) teksti. Võõrkeelsetes pealkirjades on sõltuvussüntaksis ülemuseks viimane sõna. Pealkiri allub järellisandina lisandi põhjale. Näites 24 on kogu fraasi põhjaks (ja seega ka viitealuseks) *plaat*.

(24) uus *plaat* „head music”

Viitamata jäid pronoomenid, millele ei leidunud otsest täis- või samatähenduslikku vastet; mis viitasid tekstist osaliselt või täielikult välja (ei saanud kätte referentide täisloetelu või –viidet rühmasõna kujul) või mis olid üldsõnalised nagu näited 25 ja 26 ning mis töötasid mõnes muus funktsioonis nagu rõhutav sõna näites 27 ja küsisõna näites 21. Vaata ka kolmandast peatükist viitealuseta asendussõnade näiteid 28–34.

(25) Kui *me* kaotame füüsilise kontakti oma kehaga, kaotame *me* füüsilise kontakti ka maailmaga, paradiisiga.

(26) Muide, raadios on alati olnud tunduvalt kõrgema intellektuaalse tasemega toimetajad kui televisioonis. Raadio nõuab, et *sa* oskaksid kirjutada ja emakeelt hoolikalt pruukida.

(27) Jaanus Männiku arvates aga ilmselt siis, kui „õiged“ on kogu põllumaale ükskõik *mis* moel käpa peale pannud.

2.2. Analüüsimaterjali koostamine korpuse põhjal

Asendussõnade ja nende viitealuste analüüsimiseks ei sobi korpuse algne kuju (vt lisa 1 või näidet 14), mis eeldaks käsitsi failide läbi töötamist ja andmete Exceli tabelisse kandmist. Töö autor lõi ligi 475-realise Pythonis (versioon 3.7.0a2) kirjutatud

asendussõnade analüsaatori[5], mis sorteerib korpusefailidest välja pronoomenid ja viitealused ja tõstab analüüsimise hõlbustamiseks nende morfoloogilise ja süntaktilise info Exceli faili.

Asendussõnade analüsaator töötleb ja analüüsib iga faili eraldi ning kogub andmed sõnastikku, kus võtmeteks on analüüsitavad pronoomenid ja iga sõnastiku võtme väärtuseks on järjend, mis sisaldab kahte elementi. Selle sõnastiku info teisendatakse hiljem Exceli faili kujule. Sõnastiku väärtuseks on järjendite järjend, mille iga elemendi esimene element on analüüsitava asendussõna infojärjend. Teine element on järjend tema viitealuste infojärjendi(te)ga. Kui viitealus puudub, on elemendiks vastav sõne.

Esimesena loeb programm faili ridade kaupa sisse. Enne {Pronoomen}- ja {Viitealus}-märgenditega ridade välja sorteerimist lisab programm vastava faili iga lemma inforea lõppu selle lemma lausenumbri ja lauses oleva sõnade arvu. Samuti loob programm eraldi järjendi, kuhu on kogutud kõikide lausete pikkused sõnedes (lauses arvestatakse ka kirjavahemärke, sh lauselõpupunkte, eraldi sõnedeks). Seda on vaja hiljem asendussõna ja tema viitealuse vaheliste kauguste arvutamiseks. Kui lemma inforeal on {Pronoomen}- või {Viitealus}-märgend, siis lisab programm rea lõppu ka sõna algvormi, mille leiab inforeale eelnevalt realt.

Viimase asjana enne märgenditega ridade järjendisse sorteerimist lisab programm pronoomenitele, millel on inforeal {Viitealus}-märgend, inforeale juurde {Pronoomen}-märgendi. Märgendamiskeskkonnast korpust teisele kujule teisendades jäi asendussõnadele, mis funktsioneerisid ka viitealustena mõnele teisele asendussõnale, inforeale alles vaid {Viitealus}-märgend. See aga segaks programmi tööd, mis otsib {Pronoomen}- ja {Viitealus}-märgenditega ridade listist {Pronoomen}-märgendeid. Seetõttu lisab programm puuduoleva märgendi juurde.

Igale järjendis olevale pronoomenile leiab programm samast järjendist selle viitealuste inforead ning kannab saadud info koos pronoomeniga sobival kujul sõnastikku. Selleks, et inforead infojärjenditeks teisendada, on loodud mahukad meetodid. Mahukust lisab asjaolu, et iga sõnaliiki tuli käsitleda eraldi nende morfoloogiliste erisuste tõttu.

[5] Asendussõnade analüsaatori ja selle tulemi (Exceli faili) leiab aadressilt https://github.com/Lindafr/Asenduss6nadeAnalysaator.

Kui igast failist on sobiv info sõnastikku kantud, teisendatakse andmed sõnastikus Exceli tabeliks. Iga asendussõna jaoks luuakse Exceli tabelis kaks lehekülge. Esimesel leheküljel on iga asendussõna info ühe korra ühel real. Kõigi tema viitealuste info järgneb asendussõnale samal real. Sellel leheküljel on hea analüüsida asendussõnade karakteristikuid, kuid kõigi viitealuste infot on raske kätte saada, kuna need ei ole üksteise all, vaid on oma asendussõnaga samal real erineval kaugusel sõltuvalt viitealuste arvust. Teisel leheküljel on viitealuste paremaks analüüsimiseks tõstetud iga viitealuse info koos tema pronoomeni infoga eraldi reale. Sellel leheküljel on hea analüüsida viitealuste karakteristikuid ning asendussõnade ja viitealuste vahelisi seoseid, kuna sealt saab iga viitealuse info mugavalt kätte (viitealuste inforead on üksteise all).

Mõlema lehekülje iga rea alguses on asendussõna info: asendussõnale määratud unikaalne indeks, viitealuste arv, sõna algne kuju, sõnade arv lauses, lause number, koht lauses, süntaktiline roll, arv, kääne, pronoomeni liik, pööre, ülemusverbi number. Sellele järgneb viitealuse info, mille esimesed üheksa lahtrit on alati samasugused: sõnaliik, sõna algne kuju, sõnade arv lauses, lause number, koht lauses, süntaktiline roll, arv, kaugus, ülemusverbi number. Vastavalt sõnaliigile jätkavad inforida järgnevad lahtrid:

- pronoomen viitealusena puhul: pronoomeni liik, kääne, pööre;
- numeraali puhul: numeraali liik (põhiarvsõna või järgarvsõna), kääne, esitusviis (number, rooma number, sõna);
- verbi puhul: verbitüüp (põhiverb, abiverb või modaalverb), kõneviis (indikatiiv e kindel kõneviis, imperatiiv e käskiv kõneviis, konditsionaal e tingiv kõneviis, kvotatiiv e kaudne kõneviis) või infiniitse verbi vorm (*da*-infinitiiv e infinitiiv, *des*-vorm e gerundiiv, *ma*-infinitiiv e supiin, partitsiibid), aeg, pööre, tegumood, jaatus, eitus, kääne (olemas osade infiniitvormide puhul);
- lühendi puhul: lühendi sõnaliik (nimisõna-, omadussõna-, määrsõna-, tegusõnalühend), kääne;
- adjektiivi puhul: võrre (alg-, kesk- või ülivõrre), kääne;
- substantiivi puhul: sõnaliik (üld- või pärisnimi), kääne.

Määrsõnal lisalahtreid ei ole. Kui vastavat infot inforeal ei leidu või tunnus puudub, jääb lahter tühjaks või täitub kolme X-iga.

3. EESTI ASENDUSSÕNADE ANALÜÜS ASENDUSSÕNADE SUHTES KÄSITSI MÄRGENDATUD KORPUSE PÕHJAL

Selles peatükis on esitatud korpuse põhjal tehtud pronoomenite ja viitealuste morfoloogilise ja süntaktilise info ning nendevaheliste seoste analüüs. Analüüsi eesmärk on leida asendussõna ja tema viitealuse vahelisi seoseid, mida saaks eesti keele asendussõnade lahendaja loomisel ära kasutada. Samuti peetakse analüüsis silmas Mutso lahendaja (vt ptk 1.3.2) algoritmi ja mõne tema indikaatori mõistlikkust asendussõnade suhtes märgendatud korpuse kontekstis.

Nagu peatükis 2.1 mainitud, on korpuses viitealuste suhtes märgendatud pronoomenid *mina/ma, sina/sa, tema/ta, meie/me, teie/te, nemad/nad, kes, mis* ja *see/need* (Mutso analüüsib vaid *tema-nemad* paari (vt ptk 1.3.2)). Korpuses märgitakse mitmuses isikupronoomeneid ainsuse pronoomeni lemmaga (st sõna *meie* lemmareal on märgitud lemmaks *mina* ning arvuks mitmus). Seetõttu on ka käesolevas analüüsis vaadeldud isikupronoomenite paare *mina-meie, sina-teie, tema-nemad* ja *see-need* üldjuhul korraga, kui ei ole just vastupidi märgitud nende vahel leitud olulisi erinevusi. Seega tuleb lugedes arvestada, et näiteks sõna *see* all mõeldakse nii ainsuse kui ka mitmuse vormi ning juhul, kui soovitakse tuua välja mingi oluline erinevus selle pronoomeni ainsuse ja mitmuse vahel, siis öeldaksegi „ainsuses *see*“ ja „mitmuses *see*“.

Analüüsis on vaadatud kõikvõimalikke seoseid ja võrreldud näitajaid, mis leiduvad korpuse igal lemma morfoloogilise ja süntaktilise info real. Lisades 2–7 leiab iga asendussõna viitealuste süntaktiliste rollide jaotuse asendussõnade kaupa ja viitealuste käändelise jaotuse asendussõna käänete kaupa. Pronominaalsete asendussõnade analüüsis tuuakse välja järgmised parameetrid:

- viitealus(t)ega ja viitealuseta asendussõnade osakaal, mis annab vastuse küsimusele, kui suure tõenäosusega peaks programm asendussõnale viitealust üldse otsima hakkama ning kui suure tõenäosusega võib programm eeldada, et viitealust ei ole;
- asendussõnade käändeline ja arvuline (ainsus ja mitmus) jaotus, mis iseloomustab valimit täpsemalt;

- viitealuste sõnaliigiline jaotus, mis näitab, millist liiki sõnu viitealuste otsimisel võiks eelistada;
- viitealuste morfoloogiline info, mis iseloomustab natukene täpsemalt viitealuste morfoloogilisi näitajaid viitealuste sõnaliikide kaupa;
- asendussõnade ja viitealuste arvuline ühildumine, mis on oluline Mitkovi ja Mutso lahendaja parameeter (vt ptk 1.3.2), kuid nii Mitkov, Mutso kui ka töö autor leiavad, et ära ei tohiks unustada ka ainsuses hulgasõnu (näiteks *rühm* ja *grupp*), mille asendussõna võib olla mitmuses (nt *nemad* ja *see*) (Mitkov 2002: 146; Mutso 2008: 32);
- viitealuste süntaktilised rollid koos nende asendussõnade süntaktiliste rollidega, mis näitab, kui suurel hulgal viitealustel on oma asendussõnaga sama süntaktiline roll ning kas nende rollide vahel leidub mingeid seoseid;
- käändsõnaliste viitealuste käänded asendussõnade käänete kaupa, mis illustreerib, kui suurel hulgal viitealustel on oma asendussõnaga sama kääne ja kas viitealuste ja asendussõna käänete vahel leidub mingeid tendentse või tugevaid seoseid;
- viitealuse kaugus tema asendussõnast sõnedes (töös kasutatakse kaugusühikuks termini *lemma* asemel terminit *sõne*, mis tähendab korpuse lemmat ehk sõnasid ja kirjavahemärke[6]);
- viitealuse kaugus tema asendussõnast lausetes, mis näitab, kui kaugelt peaks programm viitealuseid otsima;

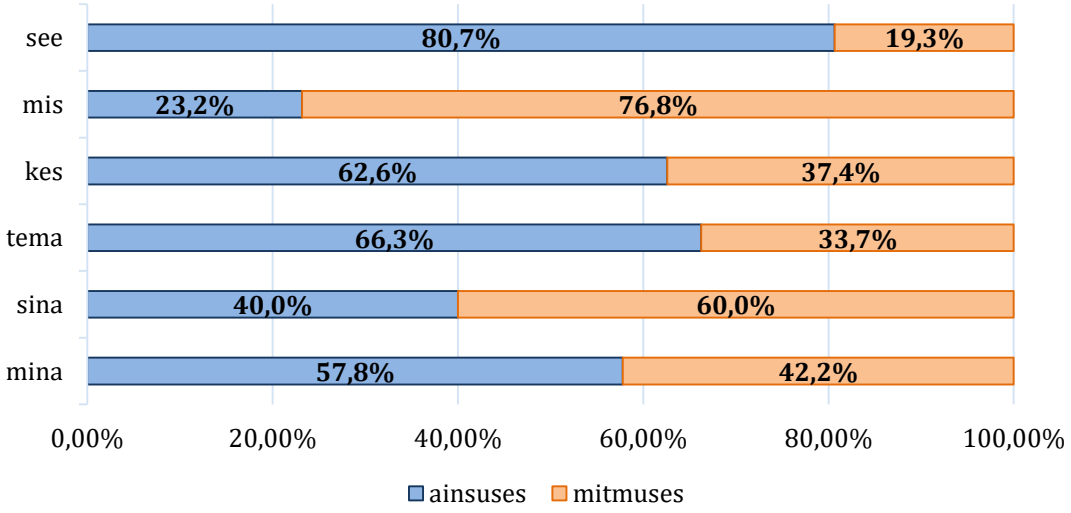
Mõnel üksikul pronoomenil puudus {Pronoomen}- või {Viitealus}-märgend ja jäi seetõttu analüüsist välja. Kõige rohkem esines asendussõnana korpuses pronoomen *see* (1489 korda), sellele järgnesid pronoomenid *tema* (922 korda), *mis* (716), *mina* (680) ja *kes* (337). Kõige vähem esines korpuses pronoomen *sina* (kokku 135 korda). Võib eeldada, et suurem valim tähendab varieeruvamat ja seetõttu ka täpsemat ning ulatuslikumat andmestikku, mille põhjal võib kindlamaid järeldusi teha. Seetõttu võib olla eriti julge pronoomenite *see* ja *tema* kohta üldistusi tehes.

[6] Korpuses loetakse ka kirjavahemärke (sh punkte) lemmadeks ja neile on antud oma kohanumber lauses.

On loogiline, et kõige rohkem viitealuseid leidub pronoomenil *see* (1220), kõige vähem viitealuseid (vaid 84) pronoomenil *sina*, kuna neid pronoomeneid esines korpuses vastavalt kõige rohkem ja kõige vähem. Pronoomenil *kes* on 299 viitealust, pronoomenil *mina* 378, pronoomenil *mis* 582 ja pronoomenil *tema* 958.

3.1. Asendussõnade käändeline ja arvuline jaotus

Joonisel 1 on toodud asendussõnade arvuline jaotus. Sealt võib näha, et asendussõnad, millest enamus esineb korpuses ainsuses, on *mina* (57,8% ainsuses), *kes* (62,6%), *tema* (66,3%) ja *see* (80,7%). Asendussõnal *mis* on mitmuses lausa 76,8% esinemisjuhtudest ning asendussõnal *sina* 60%. Üllatuslikul kombel ei sarnane omavahel asendussõnad *kes* ja *mis*, millest autor eeldas kasutussarnasuste tõttu ka morfoloogilisi sarnasusi. Korpuses kalduvad tugevalt mitmuse või ainsuse poole asendussõnad *see* ja *mis*. Äärmisel juhul võib nende hulka lugeda ka ainsuse poole kalduva asendussõna *tema*, kuid nii selle kui ka ülejäänu puhul on protsentides suurimad osad liiga väiksed taolise üldistuse tegemiseks.



Joonis 1. **Asendussõnade arvuline jaotus.**

Pronoomenite käändeline varieeruvus on võrdlemisi suur. Pronoomen *kes* esines viies erinevas käändes, pronoomen *mis* kuues, pronoomenid *mina* ja *sina* kaheksas, pronoomenid *tema* ja *see* kümnes erinevas käändes. Neid käändeid võib vaadata täpsemalt lisades 2–7.

Kõigi asesõnade levinuim kääne on nimetav. Asendussõnadel *kes* ja *see* on sageduselt teisel kohal seestütlev kääne (*kellest* ja *sellest/nendest*). Asendussõnadel *tema* ja *mina* on sageduselt teisel kohal see-eest omastav kääne (*tema/ta* ja *minu/mu*). Asendussõna *sina* puhul saab teise kohana välja tuua kaks käänet: omastav ja seesütlev kääne (*sinu/su* ja *sinus).* Omastav kääne on selle asendussõna puhul teisel kohal mitmuses asendussõnade seas (*nende*), seesütlev kääne ainsuses asendussõnade seas (*selles*). Asendussõna *mis* puhul jagavad teist kohta osastav ja sisseütlev kääne; vastavalt 20,9%-ga ja 21,4%-ga.

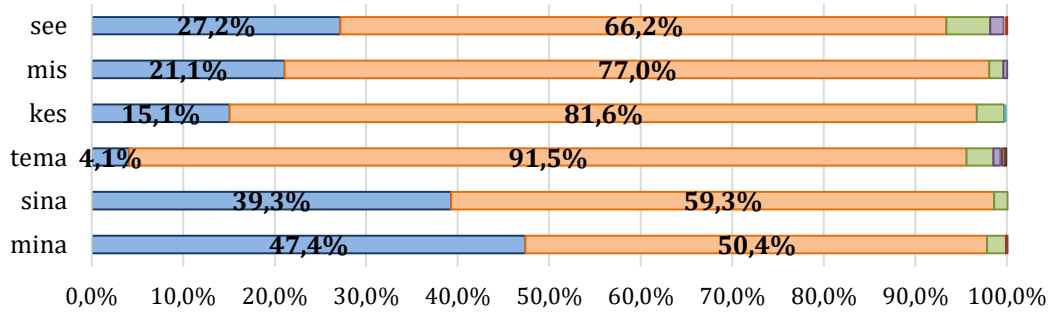
Veel võib välja tuua asendussõnade *see* ja *mina* sageduselt kolmandal kohal olevad käänded: asendussõnal *see* on kolmandal kohal osastav, asendussõnal *mina* seesütlev kääne. Teise asendussõnade puhul tõusid protsentides tugevalt esile vaid kaks sagedasimat käänet, ülejäänuid käändeid esines vähe.

3.2. Viitealus(t)ega ja viitealuseta asendussõnade osakaal

Joonisel 2 on toodud erineva viitealuste arvuga asendussõnade osakaalud pronoomenite kaupa. Iga pronoomeni puhul on üle poolte ühe viitealusega: pronoomenitest *mina* on 50,4% ühe viitealusega, pronoomenitest *sina* 59,3%, pronoomenitest *see* 66,2%, pronoomenitest *mis* 76,3%, pronoomenitest *kes* 81,6% ning pronoomenitest *tema* lausa 91,5%. Seega võib automaatse lahendaja reegleid koostades pidada silmas, et asendussõnal *tema* on üsna suure tõenäosusega üks viitealus. Järelikult toetab ka minu andmestik Mutso otsust leida asendussõnadele *tema* ja *nemad* üks viitealus (vt ptk 1.3.2). Samuti võib eelistada ühe viitealuse otsimist pronoomenitele *kes* ja *mis*, kuigi ühe viitealuse osakaal nende pronoomenite puhul ei lase välistada muid võimalusi nagu viitealuse puudumine või rohkem kui ühe viitealuse olemasolu. Ka pronoomeni *tema* puhul välistab kindlasti ühe viitealuse määramise reegel 8,5% ülejäänust korpuses esinenud juhtudest.

Pronoomenitel *mina, sina* ja *see* puudub suurel osal asendussõnadest viitealus (vastavalt 47,4%-l, 39,3%-l ja 27,2%-l) ja seetõttu tuleb asendussõnade lahendajat koostades arvestada, et nendel pronoomenitel ei pruugigi olla viitealust ja lahendajal peab jääma võimalus jätta pronoomenile viitealus määramata. Seda eriti pronoomeni *mina* puhul, mille tõenäosus, et asendussõnal on üks või enam viitealust, on vaid napilt üle 50% (vt

viitealuseta *meie* näidet 25 ja viitealuseta *mina* näidet 28). Ka küsiv-siduvate pronoomenite *kes* ja *mis* puhul ei saa eeldada viitealuse kindlat olemasolu: pronoomenil *kes* puudub 15,1%-l viitealus ning pronoomenil *mis* 21,1%-l. Küsivas funktsioonis pronoomenil viitealust ei ole (vt ptk 2.2) ja seetõttu võib viitealuse puudumisel nende pronoomenite puhul eeldada küsivat (või ka haruldasemat rõhutavat) funktsiooni. Viitealuseta pronoomeni *mis* näited on 21 ja 27, viitealuseta asendussõnu *sina* ja *see* illustreerivad vastavalt näited 26, 29 ja 30, 31. Kui pronoomen on tekstis määratlev (nagu pronoomen *see* näites 30), siis tal viitealus puudub (vt ptk 2.2).



	mina	sina	tema	kes	mis	see
0 viitealust	47,4%	39,3%	4,1%	15,1%	21,1%	27,2%
1 viitealus	50,4%	59,3%	91,5%	81,6%	77,0%	66,2%
2 viitealust	2,1%	1,5%	2,9%	3,0%	1,5%	4,8%
3 viitealust			0,8%		0,4%	1,5%
4 viitealust			0,1%	0,3%		0,1%
5 viitealust						0,2%
6 viitealust			0,1%			
7 viitealust	0,1%		0,3%			0,1%
8 viitealust			0,1%			

Joonis 2. **Erinevate viitealuste arvu osakaalud pronoomenite kaupa.**

(28) Hiljuti kutsuti *mind* suurele Rõuge rahvapeole rahuvalvajaks. /…/ Rõuge omadel oli plaanis rahulikult, ilma mürata pidutseda… Aga *ma* ajasin nende ülla kava nurja /…/. [Kirjutaja isikut pole kordagi terves tekstis välja toodud.]

(29) Kui *teil* on vähe aega, kuulake plaadi nimilugu ning fantaseerige oma rikutuse piirides sinna- ja siiapoole seda ideaalse popmuusika telgjoont. [Artikli autor soovitab lugejaile uut plaati, sõna *teile* viitab lugejale.]

(30) Vererõhul on omadus tõusta ka sel ajal, kui teda mõõdetakse. Eriti ilmne on *see* tõus meditsiiniasutuses või meditsiinitöötaja poolt vererõhku mõõtes. [On aru saada, mida tähendab *see tõus*, kuid otseselt pole võimalik kuhugi viidata.]

(31) Janar ütleb, et tema on *need* kolm ja pool aastat õppimise vaeva ette võtnud selle nimel, et Tallinnas tööd saada. [Esmakordne õpingu pikkuse mainimine.]

Kuna vaid 4,1%-l pronoomenitest *tema* puudub viitealus (ja seega 95,9%-l on vähemalt üks viitealus), võib lahendajat koostades kirjutada reegli, mis määrab igale pronoomenile *tema* viitealuse (viitealuseta *tema* ja *nemad* on näidetes 33 ja 34). Samuti puudub vaid väiksel osal pronoomenist *kes* (15,1%-l) viitealus, mistõttu võib ka selle puhul arvestada vähemalt ühe viitealuse olemasolu suurema tõenäosusega (viitealuseta *kes* on näites 32). Ülejäänu puhul tuleb lahendajat kirjutades kindlasti arvestada võimalusega, et viitealus puudub. Seda eriti pronoomenite *mina* ja *sina* puhul.

(32) Küsisin, *kes* on tema kauba sihtgrupp. [Siin on tegu küsisõnaga.]

(33) Igaüks peaks endale selgeks tegema, mida *ta* elult üldse tahab. Kui *ta* kavatseb Venemaale elama minna, siis *tal* muidugi pole mõtet eesti keelt õppida. [Puudub loogiline viitealus, ei sobi : *igaühel muidugi pole mõtet*.]

(34) „Eksamid olid meil karmid, ei veetud kummiga hindeid pikaks, kolmekümnest alustanust lõpetas kolmteist," mäletab ta. Ütleb, et on töötanud Jüri Uluotsa juhtnööride põhjal, kes *neile* kursustel agraarõigust luges. [Kuskil pole mainitud õppijate kollektiivi ega nimesid.]

Kahe viitealuse esinemissagedus on iga pronoomeni puhul üsna väike ja jääb pronoomenitel *kes, mis, mina, sina* ja *tema* lausa alla kolme protsendi. Pronoomenil *sina* ei leidu korpuses ühtegi üle kahe viitealusega asendussõna. Pronoomenil *kes* esineb üks (0,3%) nelja-viitealuseline asendussõna, pronoomenil *mina* üks (0,1%) seitsme-viitealuseline asendussõna ning pronoomenil *mis* kolm (0,4%) kolme-viitealuselist asendussõna. Pronoomenil *tema* on 1,4%-l asendussõnadest kolme kuni kaheksa viitealusega. Pronoomenitest *see* 4,8% on kahe viitealusega ning kolm kuni seitse viitealust on 1,9% *see* pronoomenitest. Näide 35 illustreerib korpuses olnud viie-viitealuselist mitmuses pronoomenit *see* ja tema viitealuseid. Võib väita, et üle ühe

viitealuse on väga väiksel osal asendussõnadest ning seitsme või kaheksa viitealusega asendussõnad on üksikud erandjuhud.

(35) Kliinilised uuringud on näidanud, et kord juba aktiveeritud eosinofiilid degranuleeruvad ja vabastavad mitmeid tsütotoksilisi proteiine nagu eosinofiilne katioonne *proteiin* (<ECP>, eosinophil cationic protein), suur *põhiproteiin* (<MBP>, major basic protein), eosinofiil-vahendatud *neurotoksiin* (eosinophil-derived neurotoxin), eosinofiili *peroksüdaas* (eosinophil peroxidase) ja Charcot'-Leydeni *kristallproteiin*. Tsütotoksiliste toime kaudu suurendavad *need* proteiinid limaskesta epiteelirakkude rakumembraani läbitavust ja /…/.

3.3. Viitealuste sõnaliigiline jaotus

Joonisel 3 on toodud viitealuste sõnaliigiline jaotus asendussõnade kaupa. Jooniselt võib näha, et kõigi asendussõnade viitealuste levinuim sõnaliik on substantiiv. Asendussõnal *sina* on kõik viitealused substantiivid ja seega võib asendussõna lahendajas vaadelda sellele asendussõnale viitealuseid otsides vaid substantiive. Asendussõnal *mina* on 97,9% viitealustest substantiivid, asendussõnal *tema* 97,6%, asendussõnal *mis* 85,9%, asendussõnal *kes* 81,6% ning asendussõnal *see* 56,3%. Vaid asendussõnal *see* on tõenäosus, et tema viitealus on substantiiv, alla 60%. Ülejäänute puhul on tõenäosus, et viitealus on substantiiv, üle 81% ja seega võib eelistada substantiividest viitealuse kandidaate. Seda eriti asendussõnade *sina* ja *tema* puhul.

Jooniselt 3 võib veel näha, et asendussõnade *tema*, *sina* ja *mina* puhul pole peale substantiiv-viitealuste ühtegi teist liiki viitealust või on neid vähe. 98,3% asendussõna *tema* viitealustest on kas substantiivid või pronoomenid. Seega on Mutso valik võtta kandidaatsõnadeks ainult need sõnaliigid (vt ptk 1.3.2) õigustatud. Märkimisväärse osa viitealustest lisaks substantiiv-viitealustele moodustab asendussõna *see* puhul verbid (41%), asendussõna *kes* puhul pronoomenid (17,1%) ning asendussõna *mis* puhul pronoomenid (8,4%) ja verbid (3,8%). Kuna verbidele viitamine tähendab osalausele viitamist (vt ptk 2.1.1), siis saab öelda, et 41% pronoomeni *see* ja 3,8% pronoomeni *mis* viitealustest viitavad osalausele.

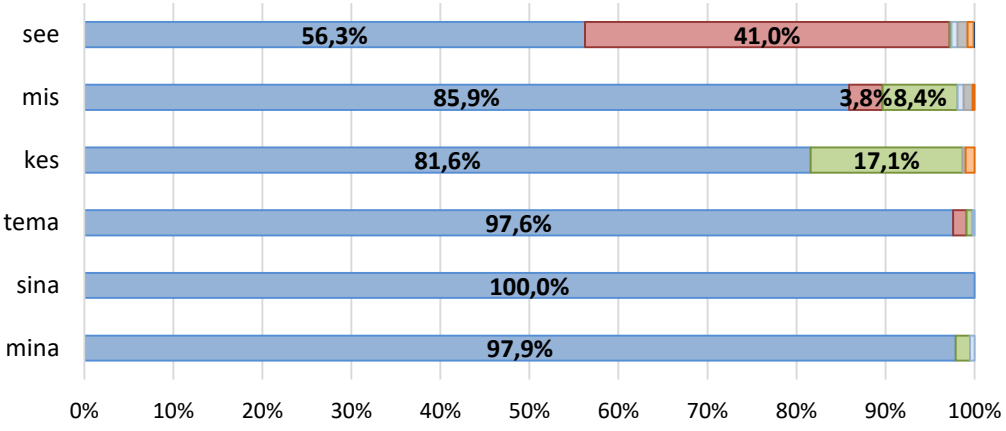
Korpusest võetud näide 36 illustreerib, kuidas viitab pronoomen *see* osalausele, ja näide 37 illustreerib sama asendussõna *mis* kohta. Näited 38 ja 39 illustreerivad asendussõnade *kes* ja *mis* viitamist pronoomenile (pronoomen *kellel* viitab pronoomenile *neile* ning pronoomen *mida* viitab pronoomenile *see*, lisaks viitavad mõlemad *kes* ja *mis* viitealused omakorda osalausetele (*neile* viitealuseks on *minna* ja *see* viitealuseks on *nägid*)).

(36) Võin rahulikult jälgida, kuna ei vastuta *selle* eest, mis sealt kastist *tuleb*.

(37) Maapind *liigub* Liibanoni ranniku ääres umbes iga 1500 aasta järel, *mis* tähendab, et 551. aastale ligilähedast maavärinat ja tsunamit võib oodata lähiajal.

(38) Soovitus *neile*, *kellel* on kalduvus koduteel kaduma minna.

(39) *See*, *mida* sa nägid, oli rutiinne protseduur.



	mina	sina	tema	kes	mis	see
substantiiv	97,9%	100,0%	97,6%	81,6%	85,9%	56,3%
verb			1,5%		3,8%	41,0%
pronoomen	1,6%		0,7%	17,1%	8,4%	0,2%
lühend	0,5%		0,2%		0,7%	0,7%
numeraal				0,3%	1,0%	1,1%
omadussõna				1,0%	0,2%	0,7%
adverb						0,1%

Joonis 3. **Viitealuste sõnaliigiline jaotus asendussõnade kaupa.**

Ülejäänud sõnaliikide esinemissagedused jäävad iga asendussõna viitealuste puhul vahemikku 0,1% kuni 1,6%. Vähemalt üks numeraal-viitealus ja üks adjektiiv-viitealus esineb asendussõnadel *kes*, *mis* ja *see*. Vähemalt kaks lühend-viitealust esineb asendussõnadel *mis*, *mina*, *tema* ja *see*. Vähemalt kolm pronoomen-viitealust esineb

asendussõnadel *mina, tema* ja *see*. Asendussõnal *tema* on 14 verb-viitealust (1,5% viitealustest) ning asendussõnal *see* on üks adverb *joostes* (0,1% viitealustest). Viimane on aga analüüsi käigus avastatud korpuseviga, kuna õige viitealus oleks vastavalt süntaksipuu reeglitele olnud *zhongleerimine* (see koht on toodud näites 40).

(40) Levinuim on *zhongleerimine* joostes (joggling). *Selles* osas on legendaarseim oma ala meister /…/.

Igal asendussõnal esineb pronoomen-viitealuseid 3–7 erinevas sõnaliigis. Lahendajat koostades tuleb arvestada, et sageduselt esikohal (ja seega eelistatav) on alati substantiiv-viitealus. Asendussõnal *see* tuleb eelistada ka verb-viitealuseid ning asendussõnal *kes* pronoomen-viitealuseid. Ülejäänud sõnaliigid esinevad viitealuste seas nii vähe, et nendega ei pea lahendajat koostades arvestama.

3.4. Viitealuste iseloomustus sõnaliikide kaupa

Viitealuste morfoloogilist infot analüüsides on näha, et isikuasesõnalistelt asendussõnadelt, millelt võiks eeldada pärisnimedele viitamist, moodustavad nende substantiiv-viitealuste seas kõrgeima osakaalu just pärisnimed. Asendussõna *mina* substantiiv-viitealustest 78,1% on pärisnimed, asendussõna *sina* substantiiv-viitealustest 81%. Asendussõna *tema* substantiiv-viitealuste pärisnimede osakaal on 51,3% ehk napilt üle poole. See on väikseim osakaal isikupronoomenite seas ja kinnitab asendussõna *tema* omadust viidata nii elusale kui ka elutule referendile (vt ptk 1.2).

Kui asendussõnade *mis* ja *see* substantiiv-viitealuste pärisnimede osakaal jääb alla viie protsendi (vastavalt 1,9% ja 4,8%), siis asendussõna *kes* substantiiv-viitealustest moodustavad pärisnimed üllatuslikult lausa veerandi.

Asendussõnade *mis* ja *see* puhul tuleb anafooride lahendajat tehes eelistada üldnimelisi kandidaate. Asendussõnal *tema* on võrdsed võimalused mõlemal substantiiviliigil. Asendussõnade *mina* ja *sina* viitealused kalduvad olema pärisnimed, asendussõnal *kes* üldnimed. Viimase kolme asendussõna suurema osakaaluga substantiiviliigid jäävad protsentides vahemikku 75%–81%. Asendussõnade lahendajat luues võib pronoomenitel *mina* ja *sina* eelistada pärisnimesid, kuid ei tohiks välistada ka üldnimesid. Pronoomenitel

mis ja *see* peaks kindlasti eelistama üldnimesid. Pronoomenil *kes* võib pigem eelistada üldnimesid ning pronoomenil *tema* peab eelistama mõlemaid substantiiviliike.

Nendest vähestest adjektiiv-viitealustest, mis korpuses esinesid, on vaid üks asendussõnale *see* kuuluv viitealus keskvõrdes. Kõik ülejäänud adjektiiv-viitealused on algvõrdes. Numeraal-viitealuseid esines korpuses samuti vähe. Nendest vaid üks (pronoomeni *see* viitealus) on järgarvsõna. Ülejäänud numeraal-viitealused on põhiarvsõnad. Kõik korpuses leiduvad lühendid on nimisõnalühendid.

Lausa 85,7% asendussõna *mis* pronoomen-viitealustest on näitavad asesõnad. Korpusest toodud näide 41 illustreerib, kuidas osastavas käändes asendussõna *mis* viitab näitavale pronoomenile *see* (sama olukord on ka näites 39). Ka asendussõna *kes* pronoomen-viitealustest moodustavad suurima osa näitavad asesõnad (54,9%). Seda illustreerib näide 42, kus asendussõna *kes* viitab näitavale asesõnale *see*. Asendussõnal *tema* on 42,9% pronoomen-viitealustest näitavad asesõnad, asendussõnal *see* 33,3%. Seega võib pronoomen *mis* puhul eelistada viitealuse kandidaate, mis on näitavad asesõnad, ülejäänu puhul on see protsent liiga madal.

(41) *See*, *mida* kunstiteraapias kujutatakse, on metafoor.

(42) Eurokõnelustel kaotab *see*, *kes* enne lõppu end lõdvaks laseb.

Isikulistest asesõnadest viitealuseid esineb asendussõnadel *kes* (29,4% viitealustest) ja *mina* (100% viitealustest). Asendussõna *mina* kõik kuus pronoomen-viitealust on ainsuses esimese pöörde personaalpronoomenid (*mina*) ja neil on kaheviitealuseline asendussõna. See tähendab, et pronoomen-viitealus *mina* on üks osa asendussõna *meie* referentide täisloetelust (vt ptk 3.7). Näide 43 illustreerib üht sellist suhet, kus sõna *meil* alla kuulub nii minategelane kui ka viidatav prints.

(43) *Ma* usun, et lordkantsleri isiku kaudu saab Suurbritannia oluliselt mitmeski küsimuses Eestit toetada. /…/ *Meil* oli neil teemadel *printsiga* meeldiv mõttevahetus.

Umbmäärastest asesõnadest viitealuseid leidub asendussõnadel *kes* (13,7% asendussõna *kes* pronoomen-viitealustest), *mis* (10,2%) ja *tema* (42,9%). Määratlevaid asesõnu ja küsiv-siduvaid asesõnu leidub viitealusena väga vähe. Asendussõna *kes* pronoomen-

viitealustest 2% on määratlevad asesõnad, asendussõna *mis* pronoomen-viitealustest 4,1%. Asendussõnal *see* moodustavad küsiv-siduvad asesõnad pronoomen-viitealustest 66,7% (tegelikult kaks viitealust kolmest asesõna-viitealusest) ja asendussõna *tema* pronoomen-viitealustest 14,3%. Näide 44 on üks kahest asendussõna *see* küsiv-siduvast pronoomen-viitealusest.

(44) Oma ülesannet ajakirjanikuna näeb ta selliselt: esineda *nende* inimeste nimel, *keda* tahetakse unustada.

Verb-viitealused esindavad tervet (osa)lauset (vt ptk 2.1). Verb-viitealuseid leidub asendussõnadel *mis, tema* ja *see*. Asendussõnadel *mis* ja *tema* on verb-viitealuseid alla nelja protsendi viitealustest. Asendussõnal *see* on verb-viitealuseid 41%. Peaaegu kõik korpuses esinenud verb-viitealused on põhiverbid. Asendussõnal *see* on lisaks paar modaalverbi. Suur osa verbi-viitealustest on kas kindlas kõneviisis või *da*-infinitiivid. Asendussõna *mis* verb-viitealustest on 59,1% kindlas kõneviisis ja 27,3% *da*-infinitiivid. Asendussõna *tema* verb-viitealustest on 42,9% kindlas kõneviisis ja 57,1% *da*-infinitiivid. Asendussõna *see* verb-viitealustest on 62,2% kindlas kõneviisis ja 18,6% *da*-infinitiivid. Korpuses esineb ka teisi verb-viitealuste liike ja tunnuseid, kuid nende osakaalud on sedavõrd väiksed, et neid pole põhjust asendussõnade lahendajat tehes arvestada. Näide 45 illustreerib asendussõna *see* verb-viitealust kindlas kõneviisis (pronoomen *nende* viitab verbile *langevad*) ning näide 46 *da*-infinitiivis (pronoomen *selle* viitab verbile *pärandada*).

(45) Süstoolne arteriaalne vererõhk ja keskmine arteriaalne vererõhk *langevad*, aga *nende* muutustega ei kaasne reflektoorset tahhükardiat.

(46) Tundkem parem rõõmu *selle* üle, mida ta suutis meile *pärandada*.

3.5. Asendussõnade ja viitealuste arvuline ühildumine

Asendussõnadel *kes, mis, mina* ja *tema* ühildub märkimisväärne osa viitealustest oma viitajaga arvus. Asendussõnal *tema* on 81,7% viitealustest oma asendussõnaga samas arvus, asendussõnal *mina* 82% viitealustest, asendussõnal *kes* 90,6% viitealustest ning asendussõnal *mis* 91,2% viitealustest. 60,7% asendussõna *see* viitealustest ning 42,9% asendussõna *sina* viitealustest on oma viitajaga samas arvus.

Viitealuste asendussõnaga arvulise ühildumise suure osakaalu tõttu on ka eesti keele puhul loogiline kaaluda seda näitajat asendussõnade lahendaja ühe parameetrina, nagu teevad seda Mitkovi ja Mutso teadmistevaesed anafooride lahendajad (vt ptk 1.3.2). Mõlema lahendajad filtreerivad asendussõnaga arvuliselt ühilduvad kandidaadid välja ega kaalugi mitteühilduvaid sõnu viitealuse kandidaatideks. See pole aga mõistlik pronoomenite *see* ja *sina* puhul. Samuti võiks 80protsendise ühildumistõenäosusega pronoomenite *tema* ja *mina* puhul kaaluda pigem punkti andmist kui välja filtreerimist.

Üldjuhul kehtib reegel, et ainsuses asendussõnadega viitealuste seas on enamus viitealustest samuti ainsuses ning mitmuses asendussõnadega viitealuste seas on enamus viitealustest samuti mitmuses. Sellele reeglile leiab pronoomenite *mina* ja *sina* näol erandi. 71,6% mitmuses pronoomeni *mina* (ehk *meie*) viitealustest on ainsuses. Seda saab põhjendada kahe kasutusvariandiga, millest mõlemad esinesid korpuses. Esiteks võib pronoomenil *meie* olla mitu ainsuses viitealust ehk siis täisloetelu osalejatest nagu illustreerib näide 47. Näites 47 on viimases otsekõnes oleva asendussõna *me* viitealuseks esimeses lauses olevad nimed *Lestrade* ja *Watson*.

(47) Vaene inspektor *Lestrade* lamas kamina ees ja doktor *Watson* määris tema kannikaid minu odekolonniga. „Oi, andke andeks, ma vist segasin teid!“ ütlesin ma viisakalt. „Oh, pole midagi, kulla Sherlock. *Me* just lõpetasime,“ vastas Watson.

Teiseks võib viitealuseks olla hägusale rühmareferendile viitava ainsuses sõnaga (vt ptk 1.2) nagu seda on näiteks sõna *Eesti* näites 48. Selles näites on viitealuse *Eesti* asendussõna pronoomen *meie*.

(48) Ta rääkis *Eesti* pürgimisest ELi ja NATOsse ning *meie* rahva väärtushinnangutest, mis toetuvad ka traditsioonidele.

Pronoomenil *sina* on 60% ainsuses viitealustest mitmuses asendussõnaga (ehk *teie* mingis vormis asendussõnal on ainsuses viitealus). Vaid kahel juhul viitab pronoomeni *teie* mingi vorm kahele ainsuses sõnale. Üks nendest kahest viitesuhtest on leitav näites 47 (*teid* viitab inspektor Lestradele ja doktor Watsonile). Neljal juhul viitab mitmuses asendussõna *teie* mitmuses viitealusele. Ülejäänud juhtudel viitab asendussõna *sina/teie*

vaid ühele ainsuses viitealusele. Seega saab selle pronoomeni puhul eeldada pronoomeni *teie* viisaka pöördumisvormi funktsiooni sagedast kasutamist ajalehetekstides ja lahendajat koostades võib eelistada ainsuses viitealuseid (kuid ei tohi välistada mitmuses viitealuseid). *Teie* esineb väga harva hägusa rühmareferendile viitajana või kahe inimese poole viitajana.

3.6. Asendussõnade ja viitealuste süntaktiliste funktsioonide ühildumine

Tabel 1 esitab andmed asendussõna ja tema viitealuse süntaktiliste funktsioonide kohta: esimeses tulbas leiab alus-asendussõnaga viitealuste aluse funktsiooni osakaalu, teises tulbas kõigi viitealuste asendussõna süntaktilise funktsiooniga ühildumise osakaalu ja kolmandas tulbas sihitis-asendussõnaga viitealuste sihitise funktsiooni osakaalu. Tabelist võib näha, et vaid 18%–43% viitealustest oma asendussõnaga samas süntaktilises funktsioonis. Kõige vähem ühilduvad oma asendussõnaga süntaktiliselt pronoomeni *see* viitealused (18,5%). Siin mängib kindlasti rolli verb-viitealuste rohkus, mistõttu on kõige levinum süntaktiline funktsioon pronoomeni *see* viitealuste hulgas infiniitse öeldise roll. Kõige rohkem ühilduvad oma asendussõnaga süntaktiliselt pronoomeni *mina* viitealused (42,6%).

Tabel 1. **Viitealuste erinevate süntaktilise funktsioonide asendussõnaga ühildumiste osakaalud asendussõnade kaupa.**

Asendussõna	Alus-asendussõnaga viitealuste aluse funktsiooni osakaal	Kõigi viitealuste asendussõna süntaktilise funktsiooniga ühildumise osakaal	Sihitis-asendussõnaga viitealuste sihitise funktsiooni osakaal
tema	57,72%	38,8%	25,9%
see	21,4%	18,5%	13,58%
sina	68,29%	35,7%	0
mina	74,5%	42,6%.	4%
mis	34,6%	26,6%	34,85%
kes	44,44%	36,8%	10%[7]

Pronoomeni *mina* ja tema viitealuse süntaktilise ühildumise võib leida näites 49 (sõnad *mina* ja *suursaadik* on samas funktsioonis). Näide 50 illustreerib pronoomeni

[7] Kokku oli vaid 10 sihitise rollis asendussõnaga viitealust, seega esines selline suhe korpuses vaid ühe korra.

mitteühildumist oma viitealusega (viitealus *Tali* on korpuses määratud aluse funktsiooni ning asendussõna *minu* määruse funktsiooni). Viimane näide võib olla üsna sage juhtum, kus minategelase saab intervjuudes kätte tema poole pöörduvatest küsimustest. Viitealus ja asendussõna ei pruugi siiski olla nii lähestikku kui näites 50, kuna sageli mainitakse intervjueeritava nime artikli alguses, kuid vastaja räägib endast minavormis terve intervjuu vältel.

(49) „No *ma* ei tea," keeldus proua *suursaadik* kommenteerimast, kui just sellekohase ettepaneku tegi St Johni kolledži president Michael Scholar Eesti presidendile.

(50) Anu *Tali*, millist maailma tipporkestrit tahaksite dirigeerida ja miks? Viini Filharmoonikuid, sest *minu* arvates on sellel orkestril kõige ilusam ja mitmekesisem kõla.

Ühegi pronoomeni puhul ei ole süntaktilise ühildumise osakaal piisavalt suur, et asendussõnade lahendajat luues viitealuste ja asendussõnade süntaktiliste funktsioonide samasust parameetriks võtta nagu seda teeb Mitkov. See seletab ka, miks Mutso lahendaja efektiivsus tõusis poole protsendi võrra pärast süntaksi indikaatorist loobumist (vt ptk 2.3.2). Pronoomeni *tema* ühildumise protsent on analüüsitavas korpuses vaid 38,8%.

Viitealuste süntaktiliste funktsioonide hulk on pronoomeniti varieeruv. Asendussõna *sina* viitealustel on 7 erinevat funktsiooni, asendussõna *mina* viitealustel 8, asendussõna *kes* viitealustel 9, asendussõna *mis* viitealustel 12, asendussõna *tema* viitealustel 10 ja asendussõna *see* viitealustel 15 erinevat funktsiooni. Viitealuste suure varieerumise tõttu ei paista ükski süntaktiline funktsioon viitealuste seas välja oma osakaalu suuruse poolest. Viitealuste funktsioone ja nende jaotust saab lähemalt vaadelda lisades 2–7.

Andmetega tutvudes võib näha, et mõnda süntaktilist funktsiooni esindab vaid paar üksikut viitealust ning kõige suurem osakaal on iga asendussõna, v.a asendussõna *see*, viitealuste seas aluse funktsioonil. Asendussõnal *see* on aluse funktsioon viitealuste seas sageduselt teisel kohal (esimesel kohal on infiniitse öeldise funktsioon, see tähendab mitmesõnalise öeldise infiniitset komponenti). Seega võib Mutso otsus anda kaks punkti aluse funktsioonis olevatele kandidaatidele olla õigustatud. Samas annab Mutso lahendaja need punktid vaid siis, kui ka asendussõna on aluse funktsioonis (vt ptk 1.3.2). Ühegi

aluse funktsioonis oleva asendussõna viitealuste jaotuses ei ületa alus-viitealuste osakaal 75% (vastavad protsendid on toodud tabeli 1 teises tulbas), mistõttu ei ole ei ole mõistlik hinnata iga alus-asendussõna puhul ainult aluse funktsioonis olevaid viitealuse kandidaate, kuigi seda võib eelistada pronoomenite *mina* ja *sina* puhul. Selle indikaatori ebatõhusust näitab lisaks antud analüüsi tulemustele ka Mutso otsus jätta antud indikaator lahendajast välja.

Tabelis 1 on veel näha, et sihitis-asendussõnadega viitealuste seas pole kas ühtegi sihitis-viitealust või jääb nende osakaal alla 35%. Seega ei toeta antud korpuse andmestik sihitise funktsioonis asendussõnade viitealuste kandidaatidele sihitise funktsiooni eest lisapunkti andmist, kuigi pronoomenitel *kes, mis* ja *see* puhul on sihitise funktsioon viitealuste funktsioonide seas sageduselt esikolmikus.

3.7. Asendussõnade ja viitealuste käändeline ühildumine

Mutso annab viitealuse kandidaatidele punkte, kui viitealus on nimetavas käändes, osastavas käändes või oma asendussõnaga samas käändes (vt rohkem ptk 1.3.2). Tabelis 2 on toodud vastavate näitajate osakaalud iga pronoomeni käändsõnaliste viitealuste[8] seas analüüsitavas korpuses.

Tabel 2. **Erinevad käändsõnaliste viitealuste asendussõnaga käändelise ühildumise osakaalud asendussõnade kaupa.**

Asendussõna	Käändsõnaliste viitealuste asendussõnaga käändelise ühildumise osakaal	Käändsõnaliste viitealuste nimetava käände osakaal	Käändsõnaliste viitealuste osastava käände osakaal
kes	35,5%	47,8%	18,1%
mis	28,6%	44,1%	22,3%
mina	44,7%	78,8%	1,1%
sina	35,7%	83,3%	0
tema	43,7%	63,2%	9,1%
see	25,3%	42,2%	22,4%

[8] Verbid jäid välja, kuna üldjuhul verbidel käändelist vormi ei eeldata ning korpuses esines sisse- ja seesütlevat käänet verbidel harva.

Tabeli 2 teises tulbas on näha, et asendussõnade käändsõnalistest viitealustest jääb oma asendussõnaga samas käändes olevate viitealuste osakaal vahemikku 25%–45%. Need osakaalud jäävad iga asendussõna puhul alla poole ja seetõttu ei ole põhjendatud viitealuse kandidaadi eelistamine ainult samakäändelisuse põhjal, kuid tuleb arvestada selle suure (üle veerandi ja alla poole) tõenäosusega. Kindlasti tuleb silmas pidada, et Mitkovi töö põhineb keeltel, millel on eesti keelest vähem käändeid ja seetõttu võib tõenäosus viitealuste ja asendussõnade käändelisele kokkulangevusele olla suurem. Korpuses esines aga iga pronoomeni puhul viitealuseid kaheksas või rohkem erinevas käändes. Suur arv käändeid vähendab ühe käände osakaalu viitealuse käänete seas.

Nimetav kääne on kõigi asendussõnade viitealuste seas enimlevinud kääne. Tabeli 2 kolmandas tulbas on toodud nimetava käände osakaalud. Selle põhjal võib järeldada, et nimetavat käänet võib lahendajat luues eelistada asendussõnade *mina* ja *sina* puhul, kus nimetavas käändes viitealuseid on vastavalt 78,8% ja 83,3% viitealustest. Need protsendid on siiski liiga väiksed, et jätta välja viitealuse teises käändes olemise võimalus. Vahemikku 42%–48% jääb asendussõnade *kes*, *mis* ja *see* nimetavas käändes viitealuste osakaal. Asendussõna *tema* viitealustest on 63,2% viitealuseid nimetavas käändes, kuid see protsent on reegli loomiseks liiga madal. Osastava käände eelistamisele ei leia töö autor erinevalt Mutso käsitlusele (vt ptk 1.3.2) põhjust. Vahemikku 18%–23% jääb asendussõnade *kes, mis* ja *see* osastavas käändes viitealuste osakaal. Asendussõna *tema* viitealustest on 9,1% osastavas. Asendussõna *mina* puhul jääb see osakaal alla kolme protsendi ja asendussõnal *sina* polegi ühtegi osastavas käändes viitealust.

Töö materjali põhjal ei ole Mutso käänete indikaatoris kasutatud näitajate eelistamine iga asendussõna puhul õigustatud. Ainukesena võib eelistada asendussõnade *mina* ja *sina* nimetavas käändes viitealuse kandidaate. Analüüsist ei tulnud välja muid tugevaid seoseid, mida lahendajat tehes silmas pidada. Ülejäänud käänded esinesid üksikult ja harva, loomata mingeid tugevamaid seoseid. Iga asendussõna viitealuste käändelist jaotust asendussõna käänete kaupa saab vaadata lisades 2–7.

3.8. Viitealuse kaugus asendussõnast sõnedes

Kõik asendussõnade *kes* ja *mis* viitealused on oma asendussõnast kuni 19 sõnet eespool. See on võrreldes teiste asendussõnadega üsna kitsas viitealuste leidumusvahemik. Ülejäänud asendussõnadest on kõige kitsam vahemik, kus viitealus leidub, pronoomenil *see* (172 sõnet asendussõnast eespool on eespoolseim ja 21 sõnet tagapoolseim viitealus). Veidi laiem vahemik on asendussõnal *tema*, millel eespoolseim viitealus on 350 sõnet eespool ja tagapoolseim 15 sõnet tagapool.

Kõige silmatorkavamad on asendussõnade *mina* ja *sina* kaugusvahemikud. See võib tulla intervjuu eripäradest: intervjuudes mainitakse intervjueeritava nime harva ning loo autori nimi on vaid korra artikli lõpus nimetatud. Asendussõna *mina* eespoolseim viitealus on 1351 sõnet eespool ja tagapoolseim viitealus on 700 sõnet tagapool. Asendussõna *sina* eespoolseim viitealus on 1086 sõnet eespool ja tagapoolseim on 627 sõnet tagapool.

Vahemikud venitavad tihti nii laiaks vaid mõned üksikud viitealused ning suurem osa viitealustest koondub kitsamatesse vahemikesse, kust asendussõnade lahendajal tasub viitealuste kandidaate otsida. 84% asendussõna *kes* viitealustest ja 79% asendussõna *mis* viitealustest on oma asendussõnast kaks sõnet eespool. Kuna sõnedeks loetakse ka kirjavahemärke (sh komasid), võib julgelt eeldada, et need viitealused kuuluvad „….viitealus, kes/mis…“ konstruktsioonidesse ehk relatiivlausesse, mida illustreerivad korpusest võetud näited 51a ja 51b, kus asendussõna *mida* viitealuseks on *argument* ja asendussõna *kes* viitealuseks on *tudengiga*. Seda kaugust võib nende asendussõnade puhul eelistada (kuid teisi kauguseid ei tohi välistada).

(51) a) Peamine *argument*, *mida* Saksa valitsus Euroopa Liidu laienemise puhul kasutab, on praegu see, et avaneb suur turg.
b) Taanis tutvus ta kahe *tudengiga*, *kes* kutsusid ta väikesele peole ja pakkusid öömaja.

96% asendussõna *kes* viitealustest on oma asendussõnast kaks kuni kuus sõnet eespool. Vahemikku kaks kuni seitse sõnet asendussõnast *mis* ettepoole jääb 94,7% viitealustest. 83% asendussõna *tema* viitealustest on vahemikus 1–28 sõnet asendussõnast eespool. Asendussõnade *sina* ja *mina* puhul mingit tugevat viitealusevahemikku välja ei joonistu.

89,2%-l asendussõna *see* viitealuste puhul on viitealuse ja tema asendussõna vahel kuni 19 eespoolset sõnet või kuni 8 tagapoolset sõnet.

3.9. Viitealuse kaugus asendussõnast lausetes

Asendussõnadele *kes* ja *mis* tasub otsida viitealust asendussõnaga samast lausest, sest nende kõik viitealused on oma asendussõnaga samas lauses, v.a üks viitealus asendussõnal *mis*.

Asendussõnale *see* tasub viitealust otsida viitajaga samas või sellele eelnevas lauses, sest sellesse vahemikku jääb 93,6% asendussõna *see* viitealustest. Vaid üks asendussõna *see* viitaja on oma asendussõnast kaks lauset tagapool, ülejäänu on samas lauses või kuni 12 lauset eespool.

Asendussõna *tema* viitealuste leidumusvahemik on võrreldes asendussõnaga *see* ühe lause võrra laiem. 90,1% asendussõna *tema* viitealustest on oma viitajaga samas lauses või kuni kaks lauset eespool. Ülejäänu viitealustest on kuni 25 lauset eespool. Mitte ükski asendussõna *tema* viitealus ei ole asendussõnale järgnevas lauses. Mutso otsus mitte otsida kandidaate kaugemalt, kui kuni kaks lauset asendussõna lausest eespool (vt ptk 1.3.2), välistaks selle korpuse puhul 9,9% viitesuhetest.

Asendussõnade *mina* ja *sina* puhul ei leidu kitsast lauselist vahemikku, kuhu suurem osa viitealustest on koondunud. Asendussõna *mina* viitealused jäävad lausevahemikku 122 lauset asendussõnast eespool kuni 50 lauset asendussõnast tagapool. Veidi üle poole (56,1%) asendussõna *mina* viitealustest on oma asendussõnaga samas lauses või kuni neli lauset eespool. Asendussõna *sina* viitealused jäävad vahemikku 95 lauset asendussõnast eespool kuni 44 lauset asendussõnast tagapool. Asendussõnaga samas lauses või sellele eelnevas lauses on 35,7% kõigist viitealustest. Samas lauses või kuni 13 lauset eespool on 63,1% viitealustest. Mõlema asendussõna puhul ei ole väljatoodud lausetevahemike osakaalud piisavalt tugevad, et neid asendussõna lahendaja reeglistikus rakendada.

Kui asendussõnade *kes, mis, see* ja *tema* puhul pole ühtegi viitealust või on vaid üks viitealus asendussõnale järgnevates lausetes, siis asendussõnade *mina* ja *sina* puhul on

vastavalt 14,8% ja 13,1% viitealustest asendussõnale järgnevates lausetes. Nende kahe asendussõna puhul tuleb lahendajat tehes arvestada ka lauseüleste katafooridega.

3.10. Analüüsi tähtsamad tulemused

Järgnevas loetelus on analüüsist välja tulnud tähtsamad ja tugevamad seosed, mida saab kasutada eesti keele pronominaalsete asendussõnade automaatse lahendaja koostamisel. Verb ’võima’ märgib siin natukene vähemkindlamaid (alla 90%-seid) seoseid, verbid ’tulema’ ja ’pidama’ tugevamaid seoseid.

- Asendussõna *tema* puhul võib või lausa peab leidma ning asendussõnade *kes* ja *mis* puhul võib leida vähemalt ühe viitealuse. Teiste asendussõnade puhul tuleb kindlasti arvestada võimalusega, et viitealust asendussõnal ei olegi.
- Asendussõnade *kes* ja *mis* puhul võib eelistada substantiiv-kandidaate, asendussõnade *mina*, *sina* ja *tema* puhul lausa peab. Pronoomeni *see* puhul ei tohi unustada ka verb-viitealuse olemasolu võimalust.
- Asendussõnade *mina* ja *sina* korral võib eelistada substantiiv-kandidaatide puhul pärisnimesid, asendussõna *kes* korral üldnimesid.
- Asendussõnade *mis* ja *see* korral tuleb eelistada substantiiv-kandidaatide puhul üldnimesid.
- Verb-viitealuste puhul (asendussõnal *see*) tuleb eelistada põhiverbe kindlas kõneviisis või *da*-infinitiivis.
- Asendussõnadel *kes* ja *mis* peab eelistama ning asendussõnadel *tema* ja *mina* võib eelistada asendussõnaga samas arvus olevaid kandidaatsõnu.
- Mitmuses asendussõna *mina* (ehk *meie*) puhul võib eelistada ainsuses kandidaate.
- Asendussõnade *mina* ja *sina* korral võib eelistada aluse funktsiooni ühildumist.
- Asendussõnade *mina* ja *sina* korral võib eelistada nimetavas käändes kandidaate.
- Asendussõnade *kes* ja *mis* korral võib eelistada kandidaate, mis on asendussõnast kaks sõnet eespool.
- Asendussõna *kes* puhul tuleb eelistada kandidaate, mis on asendussõnast kaks kuni kuus sõnet eespool.

- Asendussõna *mis* puhul tuleb eelistada kandidaate, mis on asendussõnast kaks kuni seitse sõnet eespool.
- Asendussõna *tema* puhul võib eelistada 1–28 sõnet eespool olevaid kandidaate.
- Asendussõna *see* puhul võib eelistada kandidaate, mis jäävad vahemikku kuni 19 eespoolset sõnet ja kuni 8 tagapoolset sõnet.
- Asendussõnade *kes* ja *mis* viitealuseid tuleb otsida samast lausest.
- Asendussõna *see* puhul võib või lausa tuleb eelistada kandidaate, mis jäävad asendussõnaga samasse või sellele eelnevasse lausesse.
- Asendussõna *tema* puhul võib või lausa tuleb eelistada kandidaate, mis jäävad asendussõnaga samasse või sellele eelnevasse või üle-eelnevasse lausesse.

KOKKUVÕTE

Bakalaureusetöö eesmärk oli leida asendussõnade suhtes käsitsi märgendatud korpusest, mille üks märgendaja oli ka töö autor ise, asendussõnade ja nende viitealuste vahelisi seaduspärasusi, mida saaks eesti keelele pronominaalsete asendussõnade automaatset lahendajat luues ära kasutada. Autori hinnangul see eesmärk täideti.

Eesmärgist lähtuvalt annab bakalaureusetöö ülevaate viitesuhetest ja sellega seotud tähtsamatest mõistetest, eesti keele pronoomenitest ja erinevatest asendussõnade lahendamise meetoditest, keskendudes ühe eesti keele pronoomenitele *tema* ja *nemad* loodud reeglipõhise anafoorsete asendussõnade lahendaja reeglitele. Samuti kirjeldab antud töö *ca* 107 000 sõnalist korpust, millel analüüs põhineb, selle märgendamisreegleid ja -otsuseid ning selle viitesuhete analüüsimise hõlbustamise jaoks loodud programmi[9].

Korpuses on märgendatud pronoomenid *mina-meie, sina-teie, tema-nemad, see-need, kes* ja *mis*. Tuleb tähele panna, et analüüsis vaadeldi neid paare ühe lemmana, kuna nii olid need ka korpuses märgitud (sõne *nemad* lemmaks on korpuses *tema*). Siiski toodi analüüsis välja ka nende paaride vahelised tähtsamad erinevused, kui neid leidus.

Iga pronoomeni ehk asendussõna puhul vaadati nende viitealuste morfoloogilisi ja süntaktilisi tunnuseid ning nende tunnuste ühilduvust asendussõnaga. Samuti uuriti viitealuse ja tema asendussõna vahelisi kauguseid ja asendussõnade viitealuste arvu. Analüüsis võrreldi saadud näitajaid ka eesti keele pronoomenitele *tema* ja *nemad* loodud anafoorsete asendussõnade lahendaja reeglitega ning vaadati, kas need reeglid on vastavuses uuritava korpuse andmetega.

Analüüsist selgus, et asendussõnadel *mina-meie, sina-teie* ja *see-need* puudub suurel osal asendussõnadest viitealus (vastavalt 47,4%-l, 39,3%-l ja 27,2%-l) ja seetõttu peab lahendajal jääma võimalus jätta pronoomenile viitealus määramata. Asendussõnadel *kes* ja *mis* on suure ning asendussõnal *tema-nemad* väga suure tõenäosusega üks viitealus, mistõttu neile võib või lausa peab leidma kindlasti ühe viitealuse.

[9] Pronoomenite uurimiseks lõi autor asendussõnade analüsaatori, mis on koos oma tulemiga leitav aadressil https://github.com/Lindafr/Asenduss6nadeAnalysaator ja mille algoritmi kohta saab lugeda peatükis 2.2.

Asendussõnadel *mina-meie*, *sina-teie* ja *tema-nemad* on kõik või suur osa viitealustest substantiivid ja seetõttu tuleb lahendajat tehes nende pronoomenite puhul eelistada substantiivseid viitealuste kandidaate. Asendussõnadel *mis* ja *see-need* on väga suur osa substantiivsete viitealuste kandidaatidest üldnimed, mistõttu tuleb nende asendussõnade puhul eelistada üldnimelisi kandidaate. Silma paistis ka asendussõna *see-need* verbidest viitealuste rohkus (41% kõigist viitealustest), mille puhul tuleb eelistada põhiverbe kindlas kõneviisis või *da*-infinitiivis.

Asendussõnade *kes* ja *mis* viitealused on kaks kuni kuus sõnet asendussõnast eespool ja kindlasti samas lauses. Asendussõna *see-need* puhul on mõistlik otsida kandidaate, mis on asendussõnaga samas või sellele eelnevas lauses. Asendussõna *tema-nemad* puhul leiab suure osa viitealustest samast, sellele eelnevas või üle-eelnevast lausest. Analüüsi tulemuste põhjal võib lahendajat tehes teha muidki otsuseid, mida saaks kasutada reeglitena pronoomenitest asendussõnade automaatses lahendajas.

Seni eesti keelega seotud lahendajad ei ole veel eesti keeletehnoloogias laialt kasutusel, vaid on jäänud katsetusteks. Samas on aga taoline lahendaja oluline eeldus või abivahend paljudele keeletehnoloogilistele vahenditele, mistõttu on oluline, et muidu väga silmapaistval eesti keeletehnoloogial oleks samuti asendussõnade automaatne lahendaja tagataskust võtta. Selle analüüsi tulemusi saab edaspidi reeglipõhise pronominaalsete asendussõnade automaatse lahendaja loomisel arvesse võtta.

KIRJANDUS

Anafooride suhtes märgendatud Eesti sõltuvuspuude pank. https://github.com/EstSyntax/EstAnaphora. Vaadatud 03.05.2018

Anaphora Processing… = Anaphora Processing: Linguistic, cognitive and computational modelling. 2005. Toim Branco, A., McEnery, T., & Mitkov, R. Amsterdam: John Benjamins Publishing.

Anaphora Resolution… = Anaphora Resolution. Algorithms, Resources, and Applications. 2016. Toim Poesio, M., Stuckardt, R., Versley, Y. Berlin: Springer.

Asendussõnade analüsaator. https://github.com/Lindafr/Asenduss6nadeAnalysaator. Vaadatud 05.05.2018

Brat rapid… = Brat rapid annotation tool. Online environment for collaborative text annotation. brat.nlplab.org/. Vaadatud 03.05.2018

DAARC 2011 = The 8th Discourse Anaphora and Anaphor Resolution Colloquium kodulehekülg. daarc2011.clul.ul.pt/. Vaadatud 21.04.2018

EKG I = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. 1995. Eesti keele grammatika I. Peatoim. Mati Erelt. Tallinn: Eesti Keele ja Kirjanduse Instituut.

EKG II = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. 1993. Eesti keele grammatika II. Peatoim. Mati Erelt. Tallinn: Eesti TA Keele ja Kirjanduse Instituut.

Erelt, M. 2017a. Ellips. *Eesti keele süntaks.* Toim. Mati Erelt ja Helle Metslang. Tartu: Tartu Ülikooli Kirjastus, lk 590–601.

Erelt, M. 2017b. Sissejuhatus süntaksisse. *Eesti keele süntaks.* Toim. Mati Erelt ja Helle Metslang. Tartu: Tartu Ülikooli Kirjastus, lk 53–88.

Lappin, S. 2005. A Sequenced Model of Anaphora and Ellipsis Resolution. – Branco, A., McEnery, T., & Mitkov, R. *Anaphora Processing: Linguistic, cognitive and computational modelling*. Amsterdam: John Benjamins Publishing, pp 3–16.

Lee jt = Lee, H., Surdeanu, M., & Jurafsky, D. 2017. A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, *23*(05), pp 733–762. https://doi.org/10.1017/S1351324917000109

Mitkov jt = Mitkov, R., Evans, R., Orăsan, C., Ha, L. A., Pekar, V. 2007. Anaphora Resolution: To What Extent Does It Help NLP Applications? *Anaphora: Analysis, Algorithms and Applications.* Berlin: Springer, pp 179–190.

Mitkov, R. 1999. *Anaphora Resolution: The State Of The Art*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf. Vaadatud 20.04.2018

Mitkov, R. 2002. Anaphora Resolution. Eastbourne: Pearson Education.

Mitkov, R. 2004. The Oxford Handbook of Computitional Linguistics. Anaphora Resolution. Ed. by Ruslan Mitkov. New York: Oxford University Press.

Muischnek jt = Muischnek, K., Müürisep, K., Puolakainen, T., Aedmaa, E., Kirt, R., Särg, D., 2014. Estonian Dependency Treebank and its annotation scheme. In: Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, Adam Przepiórkowski (Ed.). *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories* (TLT13), pp 285−291.

Muischnek, K., Müürisep, K. 2016. Eesti keele sõltuvuspuude pank ja selle keeleteoreetilised lähted. – *Emakeele Seltsi aastaraamat 62*, lk 122–145. http://doi.org/10.3176/esa62.04

Mutso, P. 2008. Knowledge-poor Anaphora Resolution System for Estonian. University of Tartu, Institute of Computer Science.

Pajusalu, R. 1997. Is there an article in (spoken) Estonian? – Estonian: typological studies II. Ed. Mati Erelt. Tartu: Tartu University Press, pp 146–177.

Pajusalu, R. 2005. Anaphoric pronouns in spoken Estonian. – Minimal reference. The use of pronouns in Finnish and Estonian discourse. Ed. Riva Laury. Helsinki: Finnish Literature Society, pp 107–135.

Pajusalu, R. 2009. Pronouns and reference in Estonian. - STUF, Akademie Verlag, 62, 1/2, 122–139.

Pajusalu, R. 2017. Viiteseosed. *Eesti keele süntaks.* Toim. Mati Erelt ja Helle Metslang. Tartu: Tartu Ülikooli Kirjastus, lk 566–589.

Puolakainen, T. 2015. Anaphora resolution experiment with CG rules. *Proceedings of the Workshops on „Constraint Grammar – methods, tools and applications" at NODALIDA*. Lithuania: Vilnius, pp 35–38; www.ep.liu.se/ecp/113/006/ecp15113006.pdf. Vaadatud 04.05.2018

Reuland E., Avrutin, S. 2005. Binding and Beyond: Issues in Backward Anaphora. – Branco, A., McEnery, T., & Mitkov, R. *Anaphora Processing: Linguistic, cognitive and computational modelling*. Amsterdam: John Benjamins Publishing, pp 139–161.

Sihipärane süntaks… = Sihipärane süntaks korpuste jaoks. Eesti keeletehnoloogia lehekülg. https://www.keeletehnoloogia.ee/et/ekt-projektid/sihiparane-suntaks-korpuste-jaoks. Vaadatud 03.05.2018

VSL = Väärі, E., Kleis, R., Silvet, J., Paet, T., Rehemaa, T. 2012. Võõrsõnade leksikon. 8., põhjalikult ümber töötatud trükk. Toim. T. Paet. Tallinn: Valgus; http://www.eki.ee/dict/vsl/index.cgi?Q=anafoor&F=M&C06=et. Vaadatud 30.01.2018

ANALYSIS OF PRONOMINAL COREFERENCES IN CORPUS WITH MANUALLY ANNOTATED COREFERENCE RELATIONS

Various types of language technology devices need coreference resolution for better performance. There is currently no actively used coreference resolution programme for Estonian and the aim of this thesis is to find out and formulate rules based on which a coreference resolution programme for Estonian pronouns can be created. In order to gain necessary knowledge, this thesis analyses the pronominal coreferences in a 107 000-word corpus, which is manually annotated with pronominal coreference relations. A programme which transfers the data from the corpus into an Excel file was made to simplify the process of analysing.

The thesis also gives an overview of the main terms and concepts related to coreference, Estonian pronouns, methods of resolving coreferences automatically and of one attempt to make an automatical resolution programme for Estonian pronouns *tema* and *nemad*. It describes the corpus, its annotation and the programme created for gathering and structuring information from the corpus.

The analyse investigates Estonian pronouns *mina-meie 'I-we', sina-teie 'you(sg)-you(pl)', tema-nemad 'he/she-they', see-need 'it-these', kes 'who'* and *mis 'what'* and their antecedents' morphological and syntactical features and the distance between the pronoun and its antecedent(s). Among other important findings it is clear that the resolution has to have an option to leave pronouns *mina-meie 'I-we', sina-teie 'you(sg)-you(pl)'* and *see-need 'it-these'* without an antecedent. Pronoun *tema-nemad 'he/she-they'* on the other hand most certainly requires an antecedent. Substantives should be preferred when looking for antecedent(s) for pronouns *mina-meie 'I-we', sina-teie 'you(sg)-you(pl)'* and *tema-nemad 'he/she-they'*. If the antecedent candidate is substantive, then pronouns *mis* and *see* prefer appellatives. 41% of pronoun *see*'s *'it'* antecedents are verbs and among them those with finite indicative or infinite reading should be preferred. Pronouns *kes 'who'* and *mis 'what'* has antecedents definitely in the same sentence and 2–7 strings before the pronoun. Pronoun *see-need 'it-these'* mostly has antecedents in the same sentence or a sentence before it. Most of the antecedents of *tema-nemad 'he/she-they'* are in the same sentence with it or 1–2 sentences before.

LÜHENDID

<DN	*määrsõnaline järeltäiend*
<INFN	*verbi infinitiitne vorm järeltäiendina*
<NN	*nimi-, ase- ja põhiarvsõna järeltäiendina*
<P	*eessõna laiend*
<Q	*kvantori järellaiend*
abes	*abessiiv, ilmaütlev kääne*
abl	*ablatiiv, alaltütlev kääne*
ad	*adessiiv, alalütlev kääne*
adit	*aditiiv, lühike sisseütlev kääne*
ADVL	*adverbiaal, määrus*
all	*allatiiv, alaleütlev kääne*
AN>	*omadus- ja järgarvsõna eestäiendina*
asenduss.	*asendussõna*
el	*elatiiv, seestütlev kääne*
es	*essiiv, olev kääne*
FCV	*olema liitaegades ning modaalverbid ahelverbides*
FMV	*öeldis*
gen	*genitiiv, omastav kääne*
ill	*illatiiv, pikk sisseütlev kääne*
IMV	*infiniitne verbivorm ahelverbi kooseisus (nt* küsida *ahelverbis* võib küsida*)*
in	inessiiv, seesütlev kääne
kom	*komitatiiv, kaasaütlev kääne*
NN>	*nimi-, ase- ja põhiarvsõna eestäiendina*
nom	*nominatiiv, nimetav kääne*
OBJ	*objekt, sihitis*
P>	*tagasõna laiend*
part	*partitiiv, osastav kääne*
PRD	*predikatiiv, öeldistäide*
SUBJ	*subjekt, alus*
Sum	*summa, kokku*
term	*terminatiiv, rajav kääne*
tr	*translatiiv, saav kääne*
viiteal.	*viitealus*
VOC	*üte*

LISA 1. KORPUSE NÄIDE

```
"<s id="36">"

"<Puutüved>"
	"puu_tüvi" Ld S com pl nom @SUBJ #1->6 {Viitealus}
"<on>"
	"ole" L0 V aux indic pres ps3 pl ps af @FCV #2->6
"<miljoneid>"
	"miljon" Lid N card pl part l @NN> #3->4
"<aastaid>"
	"aasta" Lid S com pl part @ADVL #4->6
"<vastu>"
	"vastu" L0 D @Vpart #5->6
"<pidanud>"
	"pida" Lnud V main partic past ps @IMV #6->0
"<tänu>"
	"tänu" L0 K pre @ADVL #7->6
"<sellele>"
	"see" Lle P dem sg all @<P #8->7 {Pronoomen} {Coref:36.12}
"<,>"
	"," Z Com #9->9
"<et>"
	"et" L0 J sub @J #10->12
"<neid>"
	"tema" Ld P pers ps3 pl part @OBJ #11->12 {Pronoomen} {Coref:36.1}
"<ümbritses>"
	"ümbritse" Ls V main indic impf ps3 sg ps af @FMV #12->6 {Viitealus}
"<kiht>"
	"kiht" L0 S com sg nom @SUBJ #13->12 {Viitealus}
"<liiva>"
	"liiv" L0 S com sg part @<Q #14->13
"<,>"
	"," Z Com #15->15
"<mis>"
	"mis" L0 P inter rel sg nom @SUBJ #16->19 {Pronoomen} {Coref:36.13}
"<võis>"
	"või" Ls V mod indic impf ps3 sg ps af @FCV #17->18
"<olla>"
	"ole" La V aux inf @ICV #18->19
"<tekkinud>"
	"tekki" Lnud V main partic past ps @IMV #19->13
"<mõnest>"
	"mõni" Lst P sg el @NN> #20->22
"<tugevast>"
	"tugev" Lst A pos sg el @AN> #21->22
"<liivatormist>"
	"liiva_torm" Lst S com sg el @ADVL #22->19
"<.>"
	"." Z Fst #23->23
"</s>
```

LISA 2. ASENDUSSÕNA *KES* ANDMETABELID

Tabel 2.1. **Asendussõna *kes* viitealuste süntaktiliste funktsioonide jaotus asendussõna süntaktiliste funktsioonide kaupa.**

Asenduss. *kes* süntaks	**Viitealuste süntaks**									
	<NN	**<P**	**<Q**	**ADVL**	**NN>**	**OBJ**	**P>**	**PRD**	**SUBJ**	**Sum**
ADVL	1 1,54%		5 7,69%	21 32,31%		5 7,69%	3 4,62%	3 4,62%	27 41,54%	65 100%
NN>	1 4,55%		2 9,09%	6 27,27%		1 4,55%		4 18,18%	8 36,36%	22 100%
OBJ	1 10%			1 10%	1 10%	1 10%	1 10%	1 10%	4 40%	10 100%
P>						1 33,33%		1 33,33%	1 33,33%	3 100%
PRD				1 100%						1 100%
SUBJ	16 8,08%	5 2,53%	8 4,04%	32 16,16%	8 4,04%	22 11,11%	8 4,04%	11 5,56%	88 44,44%	198 100%
Sum	**19 6,35%**	**5 1,67%**	**15 5,02%**	**61 20,40%**	**9 3,01%**	**30 10,03%**	**12 4,01%**	**20 6,69%**	**128 42,81%**	**299 100%**

Tabel 2.2. **Asendussõna *kes* viitealuste käändeline jaotus asendussõna käänete lõikes.**

Asenduss. *kes* käänded	**Viitealuste käänded**											
	abl	**ad**	**all**	**el**	**es**	**gen**	**in**	**kom**	**nom**	**part**	**tr**	**Sum**
ad		5 35,71%	1 7,14%	1 7,14%					6 42,86%	1 7,14%		14 100%
el	1 1,35%	8 10,81%	8 10,81%	2 2,70%		3 4,05%		3 4,05%	33 44,59%	15 20,27%	1 1,35 %	74 100%
es						2 66,67%			1 33,33%			3 100%
nom	1 0,5%	2 1.01%	23 11,56%	13 6,53%	1 0,5%	17 8,54%	1 0,5%	6 3.02%	98 49,25%	37 18,59%		199 100%
part				2 22,22%		1 11,11%			5 55,56%	1 11,11%		9 100%
Sum	**2 0,67%**	**15 5,02%**	**32 10,7%**	**18 6,02%**	**1 0,33%**	**23 7,69%**	**1 0,33%**	**9 3,01%**	**143 47,83%**	**54 18,06%**	**1 0,33%**	**299 100%**

LISA 3. ASENDUSSÕNA *MIS* ANDMETABELID

Tabel 3.1. **Asendussõna *mis* viitealuste süntaktiliste funktsioonide jaotus asendussõna *mis* süntaktiliste funktsioonide kaupa.**

Viiteal. süntaks	**Asendussõna *mis* süntaks**					
	ADVL	**NN>**	**OBJ**	**P>**	**SUBJ**	**Kokku**
<INFN	1 2%			1 2,17%		2 0,34%
<NN	3 6%	4 6,15%	5 3,79%	2 4,35%	16 5,54%	30 5,15%
<P	2 4%		4 3,03%		1 0,35%	7 1,2%
<Q	6 12%	1 1,54%	1 0,76%	2 4,35%	9 3,11%	19 3,26%
ADVL	9 18%	13 20%	29 21,97%	6 13,04%	48 16,61%	105 18,04%
FMV		1 1,54%	2 1,52%		11 3,81%	14 2,41%
IMV	2 4%				2 0,69%	4 0,69%
NN>			2 1,52%		1 0,35%	3 0,52%
OBJ	8 16%	17 26,15%	46 34,85%	17 36,96%	71 24,57%	159 27,32%
P>			7 5,30%		8 2,77%	15 2,58%
PRD	6 12%	8 12,31%	10 7,58%	2 4,35%	22 34,6%	48 8,25%
SUBJ	13 26%	21 32,31%	26 19,7%	16 34,78%	100 34,6%	176 30,24%
Kokku	**50 100%**	**65 100%**	**132 100%**	**46 100%**	**289 100%**	**582 100%**

Tabel 3.2. **Asendussõna *mis* käändsõnaliste viitealuste käändeline jaotus asendussõna käänete kaupa.**

Käändsõnaliste viiteal. käänded	**Asendussõnade käänded**						
	ad	**el**	**es**	**ill**	**nom**	**part**	**Sum**
abl					1 0,33%		1 0,18%
ad			1 8,33%	2 1,45%	1 0,33%	3 3,26%	7 1,25%
adit				1 0,72%			1 0,18%
all			1 8,33%	5 3,62%	10 3,27%	2 2,17%	18 3,21%
el		2 18,18%	1 8,33%	5 3,62%	23 7,52%	12 13,04%	43 7,68%
es				2 1,45%	9 2,94%		11 1,96%
gen		1 9,09%		10 7,25%	29 9,48%	10 10,87%	50 8,93%
ill						1 1,09%	1 0,18%
in		1 9,09%		1 0,72%	5 1,63%	5 5,43%	12 2,14%
kom		1 9,09%		9 6,52%	12 3,92%	5 5,43%	27 4,82%
nom		4 36,36%	5 41,67%	66 47,83%	144 47,06%	28 30,43%	247 44,11%
part	1 100%	2 18,18%	2 16,67%	30 21,74%	65 21,24%	25 27,17%	125 22,31%
term			1 8,33%	1 0,72%	1 0,33%		3 0,54%
tr				5 3,62%	3 0,98%	1 1,09%	9 1,61%
XXX			1 8,33%	1 0,72%	3 0,98%		5 0,89%
Sum	**1 100%**	**11 100%**	**12 100%**	**138 100%**	**306 100%**	**92 100%**	**560 100%**

LISA 4. ASENDUSSÕNA *MINA* ANDMETABELID

Tabel 4.1. **Asendussõna *meie* viitealuste süntaktiliste funktsioonide jaotused asendussõna süntaktiliste funktsioonide kaupa.**

Asenduss. *mina* süntaks	**Viitealuste süntaks**								
	<NN	**<Q**	**ADVL**	**AN>**	**NN>**	**OBJ**	**SUBJ**	**VOC**	**Sum**
ADVL			7 8,54%		3 3,66%	2 2,44%	65 79,87%	5 6,1%	82 100%
NN>	5 7,04%		7 9,86%	2 2,82%	16 22,54%	2 7,04%	39 54,93%		71 100%
OBJ	3 12%				1 4%	1 4%	20 80%		25 100%
P>					2 13,33%		13 86,67%		15 100%
SUBJ	11 5,95%	2 1,08%	11 5,95%	1 0,54%	20 10,81%	2 1,08%	137 74,05%	1 0,54%	185 100%
Sum	**19 5,03%**	**2 0,53%**	**25 6,61%**	**3 0,79%**	**42 11,11%**	**7 1,85%**	**274 72,49%**	**6 1,59%**	**378 100%**

Tabel 4.2. **Asendussõna *mina* viitealuste käändeline jaotus asendussõna käänete kaupa.**

Asenduss. *mina* käänded	**Viitealuste käänded**								
	ad	**all**	**el**	**gen**	**in**	**kom**	**nom**	**part**	**Sum**
ad				2 3,57%	2 3,57%	2 3,57%	50 83,29%		56 100%
all		2 10%		2 10%			16 80%		20 100%
gen	1 2,08%	2 4,17%	1 2,08%	20 41,67%	2 4,17%	1 2,08%	20 41,67%	1 2,08%	48 100%
in	1 1,3%	2 2,6%		9 11,69%			65 84,42%		77 100%
kom							1 100%		1 100%
nom	4 2,33%	2 1,16%	1 0,58%	14 8,14%	1 0,58%	4 2,33%	145 84,3%	1 0,58%	172 100%
part				1 25%			1 25%	2 50%	4 100%
Sum	**6 1,59%**	**8 2,12%**	**2 0,53%**	**48 12,7%**	**5 1,32%**	**7 1,85%**	**298 78,84%**	**4 1,06%**	**378 100%**

LISA 5. ASENDUSSÕNA *SINA* ANDMETABELID

Tabel 5.1. **Asendussõna *sina* viitealuste süntaktiliste funktsioonide jaotus asendussõna süntaktiliste funktsioonide kaupa.**

Asenduss. *sina* süntaks	**Viitealuste süntaks**							
	<NN	**<Q**	**ADVL**	**NN>**	**OBJ**	**SUBJ**	**VOC**	**Sum**
ADVL			2 18,18%			8 72,73%	1 9,09%	11 100%
NN>	1 5,88%		2 11,76%			13 76,47%	1 5,88%	17 100%
OBJ	1 7,69%			1 7,69%		8 61,54%	3 23,08%	13 100%
P>						2 100%		2 100%
SUBJ	3 7,32%	1 2,44%	5 12,2%	2 4,88%	1 2,44%	28 68,29%	1 2,44%	41 100%
Sum	**5 5,95%**	**1 1,19%**	**9 10,71%**	**3 3,57%**	**1 1,19%**	**59 70,24%**	**6 7,14%**	**84 100%**

Tabel 5.2. **Asendussõna *sina* viitealuste käändeline jaotus asendussõnade käänete kaupa.**

Asenduss. *sina* käänded	**Viitealuste käänded**								
	abl	**all**	**el**	**gen**	**in**	**kom**	**nom**	**part**	**Sum**
ad					2 28,57%		5 71,43%		7 100%
all							3 100%		3 100%
gen		2 14,29%					12 85,71%		14 100%
in							14 100%		14 100%
nom	2 5,13%	1 2,56%	1 2,56%	2 5,13%		1 2,56%	30 76,92%	2 5,13%	39 100%
part				1 14,29%			6 85,71%		7 100%
Sum	**2 2,38%**	**3 3,57%**	**1 1,19%**	**3 3,57%**	**2 2,38%**	**1 1,19%**	**70 83,33%**	**2 2,38%**	**84 100%**

LISA 6. ASENDUSSÕNA *TEMA* ANDMETABELID

Tabel 6.1. **Asendussõna *tema* viitealuste süntaktilised funktsioonid asendussõna süntaktiliste funktsioonide kaupa.**

Asenduss. *tema* süntaks	**Viitealuste süntaks**										
	<NN	**<P**	**<Q**	**ADVL**	**FMV**	**NN>**	**OBJ**	**P>**	**PRD**	**SUBJ**	**Sum**
<NN	11 64,71%		4 23,53%				1 5,88%			1 5,88%	17 100%
ADVL	5 3,57%		4 2,86%	16 11,43%	1 0,71%	16 11,43%	19 13,57%	1 0,71%		78 55,71%	140 100%
NN>	6 3,47%	1 0,58%	4 2,31%	21 12,14%		35 20,23%	17 9,83%	3 1,73%	1 0,58%	85 46,13%	173 100%
OBJ	2 2,47%		4 4,94%	8 9,88%		6 7,41%	11 13,58%	3 3,7%	1 1,23%	46 56,79%	81 100%
P>	1 3,45%		2 6,9%	2 6,9%	1 3,45%	3 10,34%	2 6,9%			18 62,07%	29 100%
SUBJ	40 7,72%		11 2,12%	62 11,97%	4 0,77%	49 9,46%	46 8,88%	4 0,77%	3 0,58%	299 57,72%	518 100%
Sum	**65 6,78%**	**1 0,1%**	**29 3,03%**	**109 11,38%**	**6 0,63%**	**109 11,38%**	**96 10,02%**	**11 1,15%**	**5 0,52%**	**527 55,01%**	**958 100%**

Tabel 6.2. **Asendussõna *tema* käändsõnaliste viitealuste käändeline jaotus asendussõna käänete kaupa.**

Asendus. *tema* käänded	**Käändsõnaliste viitealuste käänded**										
	abl	**ad**	**all**	**el**	**es**	**gen**	**in**	**kom**	**nom**	**part**	**Sum**
abl						1 20%		1 20%	3 60%		5 100%
ad		2 3,51%	6 10,53%			10 17,54%		1 1,75%	34 59,65%	4 7,02%	57 100%
all		1 1.96%	3 5,88%			5 9,8%	1 1,96%	1 1,96%	35 68,63%	5 9,8%	51 100%
el		1 3,57%	3 10,71%	3 10,71%					16 57,14%	5 17,86%	28 100%
es										1 100%	1 100%
gen	1 0,47%	6 2,8%	7 3,27%	4 1,87%	1 0,47%	48 22,43%		6 2,8%	122 57%	19 8,88%	214 100%
in		1 33,3%				1 33,33%				1 33,33%	3 100%
kom			1 14,29%			2 28,57%			3 42,86%	1 14,29%	7 100%
nom	4 0,78%	22 4,29%	16 3,12%	9 1,75%	1 0,19%	67 13,06%	3 0,58%	9 1,75%	344 67,06%	38 7,41%	513 100%
part		2 3,17%		2 3,17%		6 9,52%		3 4,76%	38 60,32%	12 19,05%	63 100%
Sum	**5 0,53%**	**35 3,72%**	**36 3,82%**	**18 1,91%**	**2 0,21%**	**140 14,86%**	**4 0,42%**	**21 2,23%**	**595 63,16%**	**86 9,13%**	**942 100%**

LISA 7. ASENDUSSÕNA *SEE* ANDMETABELID

Tabel 7.1. **Asendussõna *see* viitealuste süntaktiliste funktsioonide jaotus asendussõna süntaktiliste funktsioonide kaupa.**

Viiteal. süntaks	**Asendussõnade süntaktilised rollid**								
	<NN	**<P**	**ADVL**	**NN>**	**OBJ**	**P>**	**PRD**	**SUBJ**	**Kokku**
<DN				1 0,3%					1 0,1%
<INFN				1 0,3%	2 1%			1 0,3%	4 0,3%
<NN	8 26,7%		7 3,8%	32 9,4%	1 0,5%			3 0,8%	51 4,2%
<P			1 0,5%	2 0,6%	2 1%	1 1,2%		1 0,3%	7 0,6%
<Q	5 16,7%		4 2,2%	11 3,2%	8 3,9%	2 4,3%		10 2,6%	40 3,3%
ADVL		3 15,8%	22 12%	30 8,8%	16 7,8%	1 2,1%		33 8,6%	105 8,6%
AN>				1 0,3%				2 0,5%	3 0,2%
FCV				2 0,6%	1 0,5%		1 10%		4 0,3%
FMV	6 20%	9 47,4%	54 29,3%	56 16,4%	41 20%	21 44,7%	9 90%	125 32,6%	321 26,3%
IMV	1 3,3%	4 21,1%	24 13%	20 5,8%	15 7,3%	7 14,9%		48 12,5%	119 9,8%
NN>	1 3,3%	1 5,3%	3 1,6%	61 17,8%	10 4,9%	1 2,1%		13 3,4%	90 7,4%
OBJ	2 6,7%		34 18,5%	42 12,3%	53 25,9%	6 12,8%		54 14,1%	191 15,7%
P>			1 0,5%	9 2,6%	3 1,5%			7 1,8%	20 1,6%
PRD	1 3,3%		2 1,1%	5 1,5%	3 1,5%	2 4,3%		4 1%	17 1,4%
SUBJ	6 20%	2 10,5%	32 17,4%	69 20,2%	50	6 12,8%		82 21,4%	247 20,2%
Kokku	**30 100%**	**19 100%**	**184 100%**	**342 100%**	**205 100%**	**47 100%**	**10 100%**	**383 100%**	**1220 100%**

Tabel 7.2. **Asendussõna *see* käändsõnaliste viitealuste käändeline jaotus asendussõna käänete kaupa.**

Käändsõnaliste viiteal. käänded	**Asendussõna *see* käänded**										
	abl	**ad**	**all**	**el**	**es**	**gen**	**in**	**kom**	**nom**	**part**	**Sum**
abes									1 0,4%		1 0,14%
abl									1 0,4%		1 0,14%
ad		7 33,33%		2 1,09%		2 3,08%			3 1,19%		14 1,94%
adit										1 0,68%	1 0,14%
all				8 4,35%	2 6,06%	2 3,08%			5 1,98%	1 0,68%	18 2,5%
el		2 9,52%	1 9,09%	3 1,63%	2 6,06%	1 1,54%			11 4,37%	4 2,74%	24 3,33%
es				1 0,54%					1 0,4%	1 0,68%	3 0,42%
gen	1 100%	2 9,52%		36 19,57%	4 12,12%	24 36,92%			41 16,27%	27 18,49%	135 18,75%
ill									3 1,18%		3 0,42%
in				2 1,09%		2 3,08%			1 0,4%	1 0,68%	6 0,83%
kom				9 4,89%	1 3,03%				10 3,97%	5 3,42%	25 3,47%
nom		7 3,33%	8 72,73%	78 42,39%	15 45,45%	20 30,77%	3 60%	1 50%	109 43,25%	63 43,15%	304 42,22%
part		3 14,29%	1 9,09%	36 19,57%	8 24,24%	12 18,46%	2 40%	1 50%	59 23,41%	39 26,71%	161 22,36%
term				1 0,54%							1 0,14%
tr				3 1,63%					1 0,4%	4 2,74%	8 1,11%
XXX			1 9,09%	5 2,72%	1 3,03%	2 3,08%			6 2,38%		15 2,08%
Sum	**1 100%**	**21 100%**	**11 100%**	**184 100%**	**33 100%**	**65 100%**	**5 100%**	**2 100%**	**252 100%**	**146 100%**	**720 100%**

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Linda Freienthal (sünnikuupäev: 27.02.1996),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Pronominaalsete viitesuhete analüüs asendussõnade suhtes käsitsi märgendatud korpuses“, mille juhendajad on Kadri Muischnek ja Kristiina Vaik,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 25.05.2018