Tartu Ülikool

Humanitaarteaduste ja kunstide valdkond

Eesti ja üldkeeleteaduse instituut

Üldkeeleteaduse osakond

Annely-Maria Liivas

Ajaväljendite korpusemärgenduse laiendamine sündmuste ja entiteetide ajalisteks omadusteks TimeML raamistikus

Bakalaureusetöö

Juhendaja Siim Orasmaa, PhD

Tartu 2022

Sisukord

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud lähtudes Tartu Ülikooli eesti ja üldkeeleteaduse instituudi lõputöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Sissejuhatus

Sündmused paigutuvad tekstis loomupäraselt aega, mis jääb teksti narratiivi sisse. Seetõttu on ajal baseeruvad sündmused aluseks maailmas toimuvate muutuste mõtestamisele. (Pustojevsky jt 2003) Tänapäeval eksisteerib väga suures mahus elektroonilisi tekste, mis on tekitanud huvi nende automaatse töötluse ja analüüsi vastu. Võimekus tekstist sündmuseid eraldada aitaks kaasa tekstide sisust automaatse analüüsi teel ülevaate saamisele.

Eesti keelele on loodud automaatne ajaväljendite tuvastaja (Orasmaa 2012), mis suudab piiritleda tekstis ajaväljendifraase ning esitada nende semantikat. Selle tööriista rakendusvõimalusi võiks olla aga võimalik laiendada uurides, mille kohta tekstis tuvastatud ajaväljendid käivad. Võimekus tekstist ajaväljenditega seotud sündmuseid tuvastada võimaldaks koostada tekstidest automaatselt või poolautomaatselt kronoloogiaid, millest oleks kasu nii ajakirjandustekstidest lühiülevaadete tegemisel kui ka digihumanitaarias. Samuti võimaldaks see näiteks vastata automaatselt ajafaktidega seotud küsimustele (Millal toimus mingi sündmus? Kas sellele sündmusele järgnes teine sündmus?). Rakenduste loomiseks tuleb aga eelnevalt treenida välja masinõppe mudelid, mille jaoks on vaja käsitsi märgendatud tekste. Semantiline märgendus on keeruline ning eksperimentaalne, seetõttu on oluline teostatud märgendust analüüsida. Selles bakalaureusetöös kasutatakse eksperimentaalset TimeML raamistikul (Pustojevsky jt 2003) põhinevat märgendusviisi ning uuritakse, milliseid ajalisi omadusi kirjeldavad ajaväljendid nendega seotud keelendite juures, milline on märgendite struktuur ning milline on märgendite žanriline jaotus. Ajaliste omaduste all mõeldakse siinses töös näiteks millegi toimumisaega, kestust, kellegi vanust vms.

See bakalaureusetöö koosneb nii praktilisest kui uurimuslikust osast. Töö praktilises osas lisatakse osale eesti keele ajaväljendite korpusest (EstTimexCorpora[1]) TimeML

[1] https://github.com/soras/EstTimexCorpora

märgendusskeemist lähtudes ajaväljenditega seotud sündmuste ning ajaväljendite ja sündmuste vaheliste ajaseoste märgendus. Varem on TimeML skeemi kasutanud eesti keeles Siim Orasmaa, kelle doktoritöö (Orasmaa 2016) raames loodi ajasemantilise märgendusega tekstikorpus (EstTimeMLCorpus[2]). Eesti TimeML korpuse märgendustööst aga ilmnes, et eesti keelele kohandatud TimeML märgendusskeemi saaks edasi arendada ning katsetada mõnevõrra erinevaid lähenemisi. Võrreldes sündmustevaheliste ajaseoste märgendusega oli TimeML korpuse märgendustöös ajaväljendite ja sündmuste vaheliste ajaseoste märgendamine kooskõlalisem ning esines vähem ebamääraseid seoseid. Seetõttu keskendutakse selles töös just ajaväljenditega seotud sündmustele.

Siin kasutatava lähenemise uudsus seisneb selles, et sündmustele lisatakse ka nende ajalise kestuse märgendus, mida pole töö autorile teadaolevalt varem eesti keeles tehtud, kuid mida on katsetatud näiteks inglise keeles (Pan jt 2006; Vashishtha jt 2019). Loomulikus keeles võivad ajasuhted olla väljendatud ebamääraselt ehk antud võib olla sündmuse toimumise laiem aeg, kuid välja jäetud täpsemad detailid, millest tulenevalt võib sündmuse toimumisaeg olla mitmeti tõlgendatav. Mitmesed tõlgendused suurendavad märgenduses loogiliste ebakõlade tekkimise ohtu ehk võivad takistada sündmuste korrektset ajalist järjestamist. Sündmuste kestuste lisamine märgendusse võimaldaks märgendusi valideerida (ehk tuvastada märgenduses korrastamist vajavad loogilised ebakõlad) ning aitaks sündmuseid tekstis täpsemalt ajaliselt järjestada (kus korrastatud märgenduse peal kasutatakse loogilist tuletamist, et täiendavalt ebaloogilisi järjestusi välistada).

Kuna ajaväljendid ei pruugi olla alati seotud sündmustega, vaid võivad kirjeldada hoopis mingi elusa või eluta objekti ajalisi omadusi, märgendatakse selle bakalaureusetöö raames lisaks sündmustele ka ajaväljenditega seotud entiteedid, mis muidu pole TimeML raamistikus märgendust leidnud. Samuti katsetatakse ajaväljendite ja sündmuste vaheliste ajaseoste märgendamist varasemaga võrreldes vähemate seosetüüpidega.

[2] https://github.com/soras/EstTimeMLCorpus

Bakalaureusetöö koosneb neljast peatükist. Esimeses peatükis kirjeldatakse lühidalt varasemaid sündmuste määratlemise viise, ajaväljendite märgendust TimeML raamistikus ja siinses töös rakendatavat sündmuste, nende kestuste, ajaseoste ning entiteetide märgendusviisi. Teises peatükis kirjeldatakse märgendatavaid tekste, tutvustatakse märgenduskeskkonda ning märgenduste analüüsis kasutatavat meetodit. Kolmandas peatükis kirjeldatakse märgendusprotsessi ja tuuakse välja märgendustöö käigus ilmnenud probleemkohad. Neljandas peatükis esitatakse märgenduse analüüsi tulemused.

1. Taust ja märgendusskeemi piiritlemine

Loomuliku keele töötluse (ingl *natural language processing*) kontekstis on püütud juba paarkümmend aastat erinevate lähenemiste teel tekstist sündmuste eraldamist teostada. Sündmuse mõiste osas ei ole aga jõutud ühise arusaamani: seda on aja jooksul korduvalt korrigeeritud ning eksisteerinud on erinevaid sündmuse definitsioone. (Sprugnoli, Tonelli 2017: 1)

Erinevad autorid on võtnud sõnade või fraaside sündmustena tõlgendamisel arvesse nii nende semantilisi, grammatilisi kui ajalisi omadusi. Ehkki valitseb üldine üksmeel, et sündmusel peab olema ajaline mõõde, ei ole jõutud ühisele meelele, millised võivad olla sündmused sõnaliigilt ning kas sündmusteks tuleks lugeda ainult millegi juhtumisele viitavaid keelendeid või võib laiendada sündmuse tõlgendust ka seisunditele.

Teoreetilise lingvistika poole pealt on Vendler (1957) eristanud nelja sündmuste ajalist klassi: tegevused (ingl *activities*), sooritused (ingl *accomplishments*), saavutused (ingl *achievements*) ja seisundid (ingl *states*). Tegevused on dünaamilised protsessid ajas, mille lõpp on lahtine (nt *jooksma*). Sooritused eeldavad, et dünaamilisel protsessil oleks ajaline lõpp (nt *ringi joonistama*). Saavutused toimuvad erinevalt sooritustest silmapilgu vältel ning ei eelda eelnevat protsessi (nt *(midagi) taipama*). Seisundid kestavad pikema või lühema ajaperioodi vältel ja on staatilised (nt *(midagi) teadma*). Vendleri lähenemises jagunevad sündmused loomupäraselt ajalistesse klassidesse ning terve ümbritseva lause kontekst ei mängi selles jagunemises tingimata rolli, pigem väljenduvad sündmuse ajalised omadused sõna või fraasi semantikas. Reichenbach (1947) on aga eristanud lause kontekstis sündmushetke, kõnehetke ja vaatlushetke. Sündmus- ja vaatlushetk võivad kõnehetkele eelneda, sellega samaaegselt kulgeda või sellele järgneda. Mõnikord võivad sündmus-, kõne- ja vaatlushetk langeda kokku ehk toimuda samaaegselt (nt *Täna kirjutan ma esseed*), teinekord jällegi mitte. Näiteks lauses *Eilseks õhtuks olid pooled tudengid oma esseed*

esitanud on esseede esitamise sündmus toimunud enne vaatlushetke eilseks õhtuks. Vaatlushetk paikneb kõnehetke suhtes minevikus ning esseede esitamise sündmus omakorda vaatlushetke suhtes minevikus. Monahan ja Brunson (2014) püüdsid lahendada sündmuse määratlemise küsimust hinnates keelendi "sündmuslikkust" (ingl *eventiveness*). Nende lähenemise järgi saab keelendi sündmuslikkust hinnata seitsme omaduse kaudu: toimumine ehk kas sündmuse mõjul peaks maailmas toimuma muutus, ruumilis-ajaline paigutus, leksikaalne aspekt ehk millised on sündmuse sisemised ajalised omadused (sündmuse ajaline piiritletus algus- ja lõpp-punkti näol ning kestus), agentsus ehk kui palju kontrolli ja tahet on sündmuse toimumisse panustatud, mõju ehk kui palju mõjutab sündmus selles osalejaid, spetsiifilisus ehk kui konkreetne sündmus on ja reaalsus ehk kas sündmus ka tegelikult toimub (Monahan, Brunson 2014: 61-64). Keelendid võivad olla selle lähenemise järgi vähem ja rohkem sündmuslikud ning sündmused võivad ka lauses puududa.

Osad autorid on tõlgendanud sündmustena ainult verbe, teised on laiendanud sündmuse tõlgendust ka teistele sõnaliikidele. Näiteks Katz ja Arosio (2001) pakkusid välja märgendusskeemi, kus märgendamisele kuulusid sündmusviited ning nende ajaline paigutus üksteise suhtes lausete tasandil. Selles lähenemises tõlgendati sündmustena vaid lauses esinevaid verbe. Schilder ja Habel (2001) tõlgendasid näiteks aga oma lähenemises sündmusi kas nimisõna- või verbifraasidena, millel on ajaline dimensioon, kusjuures nimisõnad leidsid sündmustena tõlgendamist vaid siis, kui need olid otseselt seotud ajaväljendiga, mistõttu oli nende lähenemises sündmustena tõlgendatavate nimisõnade hulk võrdlemisi piiratud.

Ilmneb, et sündmuste määratlemisel on aja jooksul olnud eri autoritel erinevaid seisukohti ning kasutatud on erinevaid sündmuste märgendusviise. Võib öelda, et sündmuste märgendamine on veel eksperimentaalne ja diskussioonid õige lähenemisviisi üle jätkuvad siiani.

1.1. TimeML märgendusskeem

Selle töö üks eesmärkidest on märgendada käsitsi eesti keele ajaväljendite korpuse alamosas TimeML märgendusskeemist lähtudes ajaväljenditega seotud sündmused, määrata ajaseosed sündmuste ja ajaväljendite vahel ning sündmuste tüüpilised kestused.

TimeML (Pustojevsky jt 2003) on raamistik loomulikus keeles esinevate sündmuste ja ajaväljendite märgendamiseks ning nendevaheliste ajaseoste esiletoomiseks. Märgendusskeemi iseloomustab see, et see seob sündmusviited laialt varieeruvate ajaväljenditüüpidega, võimaldab ajaliselt järjestada tekstis leiduvaid sündmusviiteid üksteise suhtes ning annab semantika vähem täpsustatud ajaväljenditele. Märgendamist leiavad ajaväljendid, sündmused ning sündmustevahelised või ajaväljendite ja sündmuste vahelised seosed. TimeML-i järgi on sündmus katusmõiste, mis hõlmab nii situatsioone kui seisundeid tähistavaid keelendeid. Seega on TimeML-i sündmuse mõiste üsna lai ning sinna alla kuuluvad keelendid võivad olla sõnaliigilt nii verbid, nimisõnad kui omadussõnad (näiteks inglise keele puhul ka predikatiivosalaused või eessõnafraasid). Sündmus võib olla punkt ajas või kesta mingi ajaperioodi vältel. Algselt loodi TimeML märgendusskeem inglise keelele, kuid selle hilisem edasiarendus ISO-TimeML (Pustojevsky jt 2008) on mõeldud rahvusvaheliseks kasutamiseks.

Eesti keeles on varem TimeML märgendusskeemi ajaväljendite ja sündmuste märgendamisel kasutanud Siim Orasmaa. Eesti keele TimeML märgendusprojekti (Orasmaa 2014a; Orasmaa 2014b) raames koostati TimeML raamistiku põhjal juhised eestikeelsetes tekstides sündmuste, ajaväljendite ja seoste märgendamiseks ning lasti kolmel märgendajal märgendada 80 eestikeelses ajaleheartiklis ajaväljendid, sündmused ning ajaseosed sündmuste ning ajaväljendite ja sündmuste vahel. Ilmnes, et märgendajatevaheline kooskõla oli suurim üksikust verbist koosnevate sündmuste märgendamisel, kuid probleemsemateks osutusid nimisõnalised sündmused ja mitmesõnalised üksused (Orasmaa 2014a: 1263). Tuli

ka välja, et selgete ajaliste vihjete (kas ilmutatud kujul ajaväljendi või verbi lihtminevikus vormi) olemasolu vähendas ebamääraste (*vague*) ajaseoste osakaalu ning suurendas märgendajatevahelist kooskõla ajaseoste määramisel. Samas toodi ka välja, et kasutatud seosetüüpide suur arv võis suurendada individuaalsete tõlgenduste esinemise riski ning soovitati edaspidi katsetada väiksemal hulgal seosetüüpide kasutamist (täpsemalt ainult eelnevus- ja järgnevusseoseid *before* ja *after* ning üldist ülekattuvusseost *overlap*). (Orasmaa 2014b: 218)

1.2. Ajaväljendite märgendus

Eesti keele ajaväljendite korpuses on ajaväljendid märgendatud ISO-TimeML märgendusskeemi järgi ja ümbritsetud TIMEX3 märgenditega. Ajaväljendite märgendusjuhendi (Orasmaa 2014c) järgi on ajaväljendite semantika toodud välja märgendite atribuutides. Ajaväljenditega seotud sündmuste märgendamise kontekstis on olulised atribuudid järgmised:

- Ajaväljendi liik (atribuudis *type*)
- Ajaväljendi normaliseeritud semantika vastavalt ISO aja märkimise standardile (atribuudis *value*)
- Ajaväljendi semantika täpsustus (atribuudis *mod*)
- Ajaliste korduvuste semantika (atribuutides *quant* ja *freq*)
- Ajaväljendi relatiivsus (atribuudis *temporalFunction*)
- Probleemse ajaväljendi korral on lisatud kommentaar (atribuudis *comment*)

Ajaväljendi liik peab olema määratud igal ajaväljendil. Võimalikud liigid on järgnevad:

- DATE – toimumisaeg, mis on määratud aasta (nt *2022. aastal*), kuu (nt *järgmise aasta aprillis*), nädala (nt *septembri teisel nädalal*) või päeva täpsusega (nt *eelmisel esmaspäeval*)

- TIME – toimumisaeg, mis on määratud kellaajalise täpsusega (nt *järgmisel kolmapäeval kell 10.00*), kuid võib olla ka päevaosa (nt *pühapäeva hommikul*)
- DURATION - ajavahemik, mille algus- ja lõpp-punkt võivad olla määramata (nt *kaks tundi*), varjatult määratud (nt *viimase kümne aasta jooksul*) või ilmutatult määratud (nt *aastatel 2005-2010*)
- SET - ajaline korduvus (nt *pühapäeviti*, *hommikuti*, *aastaringselt*)

Atribuudil *value* on 4 üldist ISO aja märkimise standardile[3] vastavat formaati:

- Kuupõhine formaat (*yyyy-mm-dd*T*hh:mm*)
- Nädalapõhine formaat (*yyyy*-W*nn-wd*T*hh:mm*)
- Ainult kellaaega sisaldav formaat (T*hh:mm*)
- Ajalise kestuse formaat (P*n1*Y*n2*M*n3*W*n4*DT*n5*H*n6*M)

Kuupõhises formaadis saaks näiteks ajaväljend *2022. aastal 1. jaanuaril kell 12* kuju *2022-01-01*T*12*. Nädalapõhises formaadis saaks näiteks ajaväljend *Selle aasta jaanuari esimesel nädalal* kuju *2022*-W*01*. Kellaaega sisaldavas formaadis saaks näiteks ajaväljend *kell 11.30* kuju *T11:30*. Ajalise kestuse formaat markeeritakse P-ga. Sellele järgnev *n*-väärtus tähistab arvu ning väärtused Y, M, W, D, H ja M vastavat ajaühikut ehk aastat, kuud, nädalat, päeva, tundi või minutit (nt *P2Y* (kaks aastat), *P5H* (viis tundi)).

Kasutusel on ka mõned kokkuleppelised eritähised. Kuupõhises, nädalapõhises ja ainult kellaaega sisaldavas formaadis võivad olla nendeks päevaosad, mis asendavad tunde ja minuteid: MO (hommik), AF (pärastlõuna), EV (õhtu), NI (öö) ja DT (päevane aeg). Näiteks *käesoleva aasta 5. mai hommikul* saaks kuju *2022-05-05*T*MO*.

[3] https://www.cl.cam.ac.uk/~mgk25/iso-time.html (vaadatud 29.05.2022)

Nädalapõhises formaadis võivad olla kasutusel eritähised WD (tööpäev) ja WE (nädalalõpp), mis asendavad nädalapäeva. Näiteks *käesoleva aasta jaanuari teisel nädalavahetusel* saaks kuju *2022*-W*02*-*WE*.

Kuupõhises formaadis võivad olla kasutusel aastaaegade või kvartalite eritähised, mis asendavad kuud: SP (kevad), SU (suvi), FA (sügis), WI (talv) või Q1 (esimene kvartal), Q2 (teine kvartal), Q3 (kolmas kvartal), Q4 (neljas kvartal), QX (teadmata kvartal). Näiteks *möödunud aasta kevadel* saaks kuju *2021-SP* ning *käesoleva aasta esimeses kvartalis* saaks kuju *2022-Q1*.

Ajaväljendid võivad olla ka ebamäärased viited minevikule, olevikule ja tulevikule, mille kohta kalendrilist informatsiooni ei ole. Sellistel juhtudel kasutatakse ajaväljendi *value* atribuudis kokkuleppelisi väärtuseid PAST_REF (viide minevikule, nt *hiljuti*), PRESENT_REF (viide olevikule, nt *praegu*) ja FUTURE_REF (viide tulevikule, nt *varsti*).

Kui ajaväljend sisaldab mingit kalendrilist informatsiooni, kuid täpset väärtust pole võimalik määrata, esitatakse ajaväljendi semantika X sümbolitega. Näiteks *ühel päeval* esitataks kujul *XXXX-XX-XX*,

Ajaväljendi semantika täpsustatakse atribuudis *mod*, kui ajaväljend viitab ajastatava kalendriaja algus-, kesk- või lõpuosale (nt *2022. aasta alguses*) või esimesele või teisele poolele (nt *1990. aastate esimeses pooles*).

Atribuute *quant* ja *freq* kasutatakse ajalisele korduvusele viitavate ajaväljendite puhul. Näiteks ajaväljendi *kord tunnis* puhul märgitaks *freq* väärtuseks *“EVERY”* ja *quant* väärtuseks *“1X”*.

Ajaväljendi relatiivsus tuuakse välja, kui ajaväljendit ei ole selles sisalduva informatsiooni põhjal võimalik ajateljele paigutada ning tarvis oleks täiendavat informatsiooni kontekstist ja kalendriarvutusi (nt *üheksakümnendatel*, millest ei ilmne, millise sajandi kohta see käib).

Sellisel juhul märgitakse atribuuti *temporalFunction "true"*. Kui ajaväljend paigutub ajateljele, on selle atribuudi väärtuseks *"false"*.

1.3. Sündmuste märgendus

Sündmuste märgendamisel lähtutakse eesti keelele kohandatud TimeML märgendusskeemi sündmuste märgendusjuhendist (Orasmaa 2014e). Järgnevalt tuuakse välja, mis tingimustel ja kuidas märgendatakse eri sõnaliikidest sündmuseid ning kuidas määratakse neile klass.

1.3.1. Verbid

Sündmuste märgendusjuhendi järgi loetakse grammatilisi predikaate ehk öeldiseid üldiselt sündmustena märgendatavaks (nt *Laps sööb saia*). Pikemate verbifraaside, osalausete või lausete korral märgendatakse enamasti vaid üksuse verbiline peasõna.

Teatavaid erinevusi märgendusviisis esineb keerukamate verbifraaside puhul. Ahelverbide puhul märgendatakse üldiselt mõlemad verbifraasi liikmed lahus ehk eraldi sündmustena (nt *Varsti saame viia läbi uuringu*, *Ta hakkas lugema*, *Ta laseb endale süüa teha*). Erandiks on koloratiivtarindid, kus võib jätta koloratiivverbi märgendamata ning märgendada vaid infiniitse verbivormi (nt *Ta rõkatas naerda*).

Väljendverbide ja ühendverbide puhul märgendatakse vaid verbiline fraasiliige (nt *pähe võtma*).

Sariverbide puhul võib märgendada eraldi sündmustena kõik konstruktsioonis osalevad verbid (nt *Tule too mulle mu mantel*).

Sisuvaesest finiitsest verbivormist (nt verbid *olema*, *saama*, *jääma*, *pidama*) ja infiniitsest verbivormist koosnevate ühendite ehk perifrastiliste verbivormide puhul jäetakse tavaliselt sisuvaene verb märgendamata (nt *Kodutööd on tehtud*). Sisuvaene finiitne verbivorm ja infiniitne verbivorm võidakse märgendada eraldi sündmustena juhul, kui sisuvaene verb kannab modaalset või aspektuaalset tähendust (nt *Tänane päev saab olema igav*), samuti juhul, kui ühendi kasutus pole eriti regulaarne või edastatav tähendus pole selgelt morfoloogiline (nt *Ta arvab, et tal on õigus teha seda, mida ise tahab*). Muudel juhtudel peaks sisuvaene verb jääma märgendamata.
Sekundaartarindite ehk moodustajate, kus puudub alus ja öeldis puhul märgendatakse infiniitsed verbivormid sündmustena juhul, kui need on fraasi peasõna positsioonis (nt *Ta plaanis [teatrisse minna]*).

Konstruktsioonide puhul, mis sisaldavad *olema*-verbi on märgendusreeglites toodud aga välja palju erijuhte. Enamasti märgendatakse nii *olema*-verb kui teine konstruktsiooni liige ühe sündmusena, kui tegemist on *olema*-verbi ja öeldistäite positsioonis nimisõna või omadussõna konstruktsiooniga (nt *Patsiendi seisund on stabiilne*). Kui *olema*-verb esineb koos alusega ning alust on võimalik tõlgendada eraldiseisva sündmusena, märgendatakse nii *olema*-verb kui alus lahus (nt *Südame- ja veresoonkonnahaigused on Eestis ühed levinumad surmapõhjused*). Kui aga alus ja *olema*-verb moodustavad terviku, võib märgendada need ühe sündmusena (nt *On kahtlus, et minister on osalenud korruptsioonis*). Ainult *olema*-verb märgendatakse siis, kui aluse positsioonis nimisõna on keeruline sündmusena tõlgendada või kui see on pigem tegevuse asjaolusid, mitte seisundit kirjeldav määrus (nt *Laua peal on raamatud*). Seisundit kirjeldav nimisõnaline määrus märgendatakse koos *olema*-verbiga ühe sündmusena (nt *Ta on nõus minuga kohtuma*). *Olema*-verbi ja kvantiteedi konstruktsioonides märgendatakse vaid *olema*-verb (nt *2011. aasta rahvaloenduse andmetel oli Eestis 1,3 miljonit elanikku*). Kui *olema*-verb esineb koos infiniitse verbivormiga, märgendatakse vaid infiniitne verbivorm (nt *kodutööd on tehtud*).

Eituse korral jäetakse *ei* märgendamata (nt *ei söönud*, *ei joo*). Kui eitusena on kasutatud sõna *pole*, lähtutakse märgendamisel *olema*-verbi konstruktsioonide märgendusreeglitest.

1.3.2. Nimisõnad

Nimisõnu loetakse sündmusteks, kui need on ajaliselt määratletavad. Kui sündmus avaldub nimisõnafraasina, märgendatakse üldiselt vaid nimisõnafraasi peasõna (nt *presidendi visiit*). Erandiks on juhud, kus täiendiga nimisõnafraasi mõlemat liiget saab tõlgendada eraldiseisvate sündmustena (nt *avarii pealtnägemine*).

Mõnevõrra eriline kategooria on *sortal states*. Need võivad väljenduda kas tegijanimede (nt *ehitaja*), rolle või ameteid tähistavate nimisõnade (nt *peaminister*) või täpsete viitajate kaudu (nt *Hiina viimane keiser*). Selliseid keelendeid märgendatakse seisunditena vaid siis, kui need on seotud kas siduva öeldisega (nt *Praegu on ta Eesti president*), oleku tekkimisel viitava öeldisega (nt *Ta sai presidendiks 2021. aastal*), aspektuaalse öeldisega (nt *Ta alustas saatejuhina 2010. aastal*), oleku muutust tähistava öeldisega (nt *Ta taandus ametikohalt peaministrina 2021. aastal*) või olekut hindava või kirjeldava öeldisega (nt *Teda hinnatakse parimaks Eesti peaministriks*).

1.3.3. Omadussõnad

Enamikel juhtudel omadussõnu sündmustena ei tõlgendata (nt *suur maja*, *ilus lill*). Sündmusena võivad need leida märgendamist vaid siis, kui tegemist on lauses seisundimääruse või öeldistäitena esinevate omadussõnadega (nt *Laps jäi kurvaks*, *Mees näis õnnelik*). Oluline on ka see, et selline omadussõna peab kirjeldama mittepüsivat olekut, olema ajaliselt piiritletud ja/või olema esitatud kui kellegi arvamus, teadmine või uskumus. Ebaselguse korral jäetakse omadussõna pigem märgendamata.

1.3.4. Sündmuste klassid

Sündmuste märgendusjuhendi järgi jagunevad sündmused kaheksasse klassi: *reporting*, *perception*, *aspectual*, *i_action*, *i_state*, *state*, *modal* ja *occurrence*. Klassi *reporting* kuuluvad millegist teavitamisele viitavad sündmused (kellegi sõnul, teatel või arvates toimus midagi). Klassi *perception* kuuluvad sündmused, mis viitavad millegi tajumisele (keegi nägi, kuulis või tundis midagi). Klassi *aspectual* kuuluvad mingi teise sündmuse algusele, jätkule või lõpule viitavad sündmused (nt *Pidu <u>algas</u> kell kaheksa*). Klassi *i_action* kuuluvad sündmused, mis väljendavad tahtlikku tegevust (nt *Ta <u>otsustas</u> loobuda suhkru tarbimisest*). Klassi *i_state* kuuluvad tahtlikku seisundit väljendavad sündmused (nt *Ta <u>kardab</u> lennukiga reisimist*). Klassi *state* kuuluvad seisundid, milles kellegi tahte osalemine pole nõutud (nt *Eestis <u>on</u> ligikaudu 1,3 miljonit elanikku*). Klassi *modal* kuuluvad modaalverbid *võima, tohtima, saama, pidama, näima, paistma* ja *tunduma*, mis viitavad üldiselt mingi teise sündmuse võimalikkusele või vajalikkusele antavatele hinnangutele. Klassi *occurrence* kuuluvad kõik ülejäänud märgendatavad sündmused (nt *kontsert*).

Klassid *reporting*, *perception*, *aspectual*, *i_action*, *i_state* ja *modal* nõuavad tekstis ilmutatud kujul argumentsündmuse olemasolu (nt *Pidu algas kell kaheksa* puhul on ülemsündmus *algas*, mille argumentsündmus on *pidu*). Selleks, et markeerida sündmuste argumendisuhe, võetakse märgendamisel kasutusele seosetüüp *has_argument*. Kui sündmuse argument avaldub osalauses, siis argumendisuhet ei markeerita. Ülesande lihtsustamiseks märgendatakse vaid vahetu argument ehk kui argumentsündmus on omakorda klassist, mis nõuab argumentsündmuse olemasolu, siis järgmist argumentsündmust enam ei märgendata. Ehkki süstemaatilisem oleks märgendada kõik lauses esinevad argumentsündmused, pole see otseselt vajalik, sest nagu tõdeti ka eesti keele TimeML märgendusprojektis, on võimalik tuvastada sõnadevahelised seosed lauses automaatsüntaksiga (Orasmaa 2014a), mille töö lihtsustamisel on kasu ka vähem põhjalikust märgendusest. Klassid *state* ja *occurrence*

argumentsündmuse olemasolu ei nõua. Seetõttu määratakse nendesse klassidesse ka sündmused, mis muidu võiksid kuuluda mõnda teise sündmuste klassi, kuid ei rahulda argumentsündmuse olemasolu nõuet.

1.4. Sündmuste kestuste määramine

Osad autorid on püüdnud laiendada sündmuste märgendust, määrates märgendatavatele sündmustele ka nende ajalisi kestuseid. Pan jt (2006) kasutasid näiteks oma uurimistöös süsteemi, kus märgendajad märgendasid sündmuste võimalike kestuste alam- ja ülempiirid. Tulemusena pakuti TimeML märgendusskeemi laiendamiseks välja meetod, kus sündmuse kestuse alam- ja ülempiir tuleks märkida märgendatava sündmuse kahes eri atribuudis. Näiteks kui sündmuse kestus võiks olla vahemikus 5 sekundit ja 5 minutit, siis märgitaks atribuuti *lowerBoundDuration* "PT5S" ja atribuuti *upperBoundDuration* "PT5M". Vashishtha jt (2019) võtsid sündmuste kestuste määramisel kasutusele aga mõnevõrra lihtsama skeemi ehk kestuste tüübid *instantaneous*, *seconds*, *minutes*, *hours*, *days*, *weeks*, *months*, *years*, *decades*, *centuries* ja *forever*.

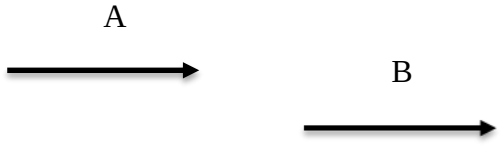
Eesti keeles pole autorile teadaolevalt sündmuste tüüpiliste kestuste märgendust siiani tehtud. Siinses töös võetakse eeskuju Vashishtha jt (2019) lähenemisest, mis on lihtsam ning kasutatakse sündmustele kestuste määramisel kestuste tüüpe. Kasutatavad tüübid on *instant* (silmapilk), *seconds* (sekundid), *minutes* (minutid), *hours* (tunnid), *days* (päevad), *weeks* (nädalad), *months* (kuud), *years* (aastad), *centuries* (sajandid) ja *forever* (igavik). Vashishtha jt lähenemise tulemustest selgus, et lühema kestusega sündmused esinevad palju sagedamini, kui aastase või veelgi pikema kestusega sündmused. Võib eeldada, et ka selle töö raames tehtavas märgenduses esineb rohkem lühema kestusega sündmuseid, sest suur osa märgendatavatest tekstidest pärineb ajakirjandusrubriikidest, kus tõenäoliselt kajastuvad rohkem päevakajalised sündmused. Seetõttu ei pruugi olla tarvis pikemate kestustüüpide eristamisega väga detailseks minna ning märgendusest jäetakse välja aastakümnete pikkune

kestustüüp. Sajanditepikkune kestustüüp jääb kasutusele ajalooalastele teadusartiklitele mõeldes, kus võib esineda suurema tõenäosusega väga pika kestusega sündmuseid.

1.5. Ajaseosed sündmuste ja ajaväljendite vahel

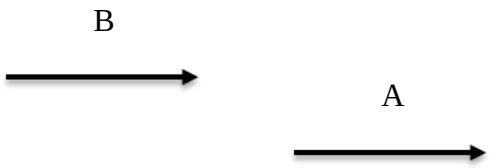
Selle töö raames märgendatakse ajaväljendite ja sündmuste vahelised ajaseosed ehk TLINK seosed ning katsetatakse nende seoste tõmbamist ka ajaväljendite ja entiteetide vahele. TimeML-i järgi vastab nii ajaväljend kui sündmus mingile intervallile ning TLINK näitab, milline on nende intervallide vaheline ajaseos ehk kuidas need üksteise suhtes paiknevad. Seoseid märgendatakse paarikaupa ehk korraga määratakse kahe liikme (ajaväljendi ja sündmuse või ajaväljendi ja entiteedi) seosed. Seoste märgendusjuhendi (Orasmaa 2014d) järgi on TLINK seoseid üheksa liiki. Kuna eesti keele TimeML märgendusprojekti käigus ilmnes, et nende kõigi kasutamine suurendab märgendajatevahelisi ebakõlasid (Orasmaa 2014b: 218), ei võeta siinses märgenduses kõiki märgendusjuhendis välja toodud seosetüüpe kasutusele. Märgenduses kasutatavad seosetüübid on *before*, *after*, *simultaneous*, *includes*, *is_included* ja *vague*. Kasutatavad seosetüübid erinevad Orasmaa (2014b: 218) poolt välja pakutuist selle poolest, et üldise ülekattuvusseose (*overlap*) asemel kasutatakse täpsemaid ülekattuvusseoseid (*simultaneous*, *includes*, *is_included*), sest need võimaldavad saada ajaväljendite ja sündmuste kattumiste kohta rohkem informatsiooni.

Seosetüübid *before* ja *after* on eelnevus- ja järgnevusseosed ehk need näitavad, kumb seoseliige esineb ajaliselt varem ja kumb hiljem. *Before* tüüpi seos näitab, et esimene seoseliige esineb teisest ajaliselt varem (ehk *A before B*).



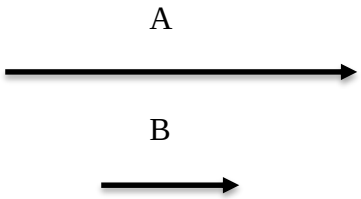
Joonis 1. **A before B.**

After tüüpi seos näitab vastupidiselt, et esimene seoseliige esineb teisest ajaliselt hiljem (ehk *A after B*).



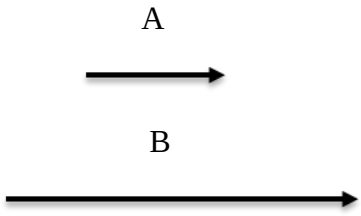
Joonis 2. **A after B.**

Seosetüübid *includes* ja *is_included* on ajalise sisaldumise ja kaasaarvamise seosed. *Includes* tüüpi seos näitab, et esimene seoseliige on teisest seoseliikmest ajaliselt pikem ning teine seoseliige sisaldub esimese sees (ehk *A includes B*).



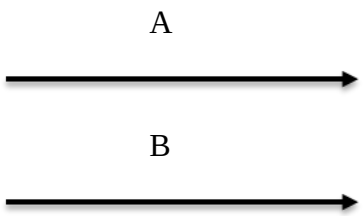
Joonis 3. **A includes B.**

Is_included tüüpi seos näitab, et esimene seoseliige on teisest seoseliikmest ajaliselt lühem ning sisaldub teise seoseliikme sees (ehk *A is_included B*).



Joonis 4. **A is_included B.**

Seosetüüp *simultaneous* näitab, et seoseliikmed toimuvad samaaegselt (ehk *A simultaneous B*).



Joonis 5. **A simultaneous B.**

Seosetüüpi *vague* kasutatakse juhtudel, mil liikmetevahelist ajaseost on teksti põhjal keeruline määrata.

1.6. Entiteedid

Kuigi selles töös on eesmärgiks märgendada TimeML skeemist lähtudes ajaväljenditega seotud sündmused, ei pruugi ajaväljend alati olla seotud sündmusega: see võib kirjeldada ka mingi elusa või eluta objekti ajalisi omadusi. Seetõttu võetakse märgendamisel lisaks sündmuse märgendile kasutusele ka entiteedi märgend *entity*. Kuna entiteetide märgendamine TimeML raamistikus on eksperimentaalne ning selle töö raames püütakse alles kaardistada, kui suur on entiteetide roll ajaväljenditega seotud keelendite seas, siis hoitakse siinne entiteetide märgendus võimalikult lihtsana ning neile ei määrata klasse ega muid atribuute. Entiteetidena tõlgendatakse lisaks elusatele (nt *tüdrukud*, *koer*, *puu*) ja eluta objektidele (nt *hülgeluu*, *müür*, *romaan*) ka organisatsioone (nt *litsentsikomisjon*) ja kohti (nt *Saaremaa*). Samuti tõlgendatakse siin töös entiteetidena viiteid teostele autori nime (nt *Faehlmann*) või pealkirja (nt *"Keisri Hull"*) kaudu.

2. Korpus ja meetod

Eesti keele ajaväljendite korpus koosneb 113 eestikeelsest tekstist. Tekstid on jaotatud nende liigi alusel üheksasse kategooriasse: arvamusuudised, kohalikud (eesti) uudised, kultuuriuudised, majandusuudised, spordiuudised, välisuudised, ajalooalased teadusartiklid, riigikogu stenogrammid ning seadustekstid. Korpuse koostas Siim Orasmaa ajaväljendite tuvastaja loomise ja testimise raames ja selles sisalduvad tekstid on valitud Eesti Keele Koondkorpusest juhuslikult (Orasmaa 2012). Korpusetekstides on kokku 1905 ISO-TimeML märgendusskeemi järgi märgendatud ajaväljendit.

Selle töö raames lisati sündmusmärgendus eesti keele ajaväljendite korpuse alamosale, milleks oli 85 teksti. Märgendamisel kasutati märgendustööriista BRAT[4]. Märgendatavate tekstide seas olid esindatud tekstid kõigist kategooriatest peale seadustekstide. Seadustekstid jäeti välja seetõttu, et nendes kipub olema suurel hulgal ajaväljendeid, millega seotud keelendeid on keeruline selle töö raames kasutatud märgendusskeemi järgi sündmustena tõlgendada. Lõik dokumendist *sea_eesti_x50009k:*

(1) *Sotsiaalmaksu makstakse töötajale või teenistujale kuu eest makstud tasult , kuid mitte vähem kui eelarveaastaks riigieelarvega kehtestatud kuumääralt proportsionaalselt sellel kuul töötatud ajaga , järgmistel juhtudel :*
1) töötaja puhul , kelle tööleping on peatunud Eesti Vabariigi töölepingu seaduse (RT 1992 , 15/16 , 241 ; 1993 , 10 , 150 ; RT I 1993 , 26 , 441 ; 1995 , 14 , 170 ; 16 , 228 ; 1996 , 3 , 57 ; 40 , 773 ; 45 , 850 ; 49 , 953 ; 1997 , 5/6 , 32 ; 1998 , 111 , 1829 ; 1999 , 16 , 276 ; 60 , 616 ; 2000 , 25 , 144 ; 51 , 327 ; 57 , 370 ; 102 , 669 ; 2001 , 17 , 78 ; 42 , 233 ; 53 , 311) § 55 punktide 3-11 alusel ,
2) teenistuja puhul , kelle teenistussuhe on peatunud avaliku teenistuse seaduse (RT I 1995 , 16 , 228 ; 1999 , 7 , 112 ; 10 , 155 ; 16 , 271 ja 276 ; 2000 , 25 , 144 ja 145

[4] https://korpused.keeleressursid.ee/brat/#/ajasemantika/ (vaadatud 31.05.2022)

; 28 , 167 ; 102 , 672 ; 2001 , 7 , 17 ja 18 ; 17 , 78 ; 42 , 233 ; 47 , 260) § 108 punktide 2-9 alusel või

3) töötaja või teenistuja puhul , kes on asunud tööle või lahkunud töölt sellel kuul.

Näites 1 on näha, et seadustes on sündmus sageli abstraktne: selle konkreetne toimumisaeg, koht ja osalised pole täpselt määratletud, nagu näiteks ajakirjanduses või ajalooalastes artiklites. Ajaväljendid märgivad sageli seaduste jõustumise kuupäevi või on viited Riigi Teataja avaldamisaegadele.

Märgendatavad tekstid, millest on välja jäetud seadustekstid kuuluvad rangelt võttes kolme suuremasse žanrisse (ajakirjandusrubriigid, ajalooalased teadusartiklid ja riigikogu stenogrammid). Tinglikult nimetatakse aga siin ja edaspidi ka ajakirjanduse eri rubriike žanriteks.

Märgenduse analüüsimiseks tuli märgendatud dokumendid teisendada Pythoni teegi EstNLTK (Laur jt 2020) jaoks sobivateks *Text* objektideks. BRAT-i märgenduste teisendamisel *Text* objektideks kasutatati skripti *convert_brat_to_estnltk_json.py*[5].

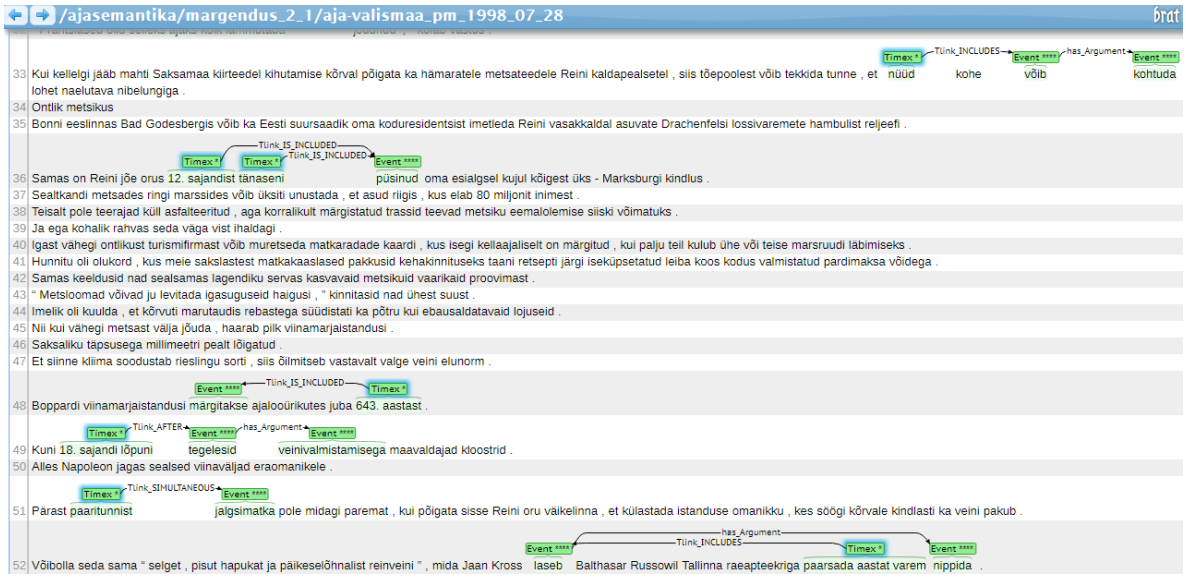
[5] https://github.com/soras/EstTimexCorpora/blob/master/scripts/convert_BRAT_to_estnltk_json.py

3. Märgendamine

Selles peatükis kirjeldatakse märgendusprotsessi ning tuuakse välja tekstide märgendamise käigus ilmnenud probleemkohad, mida töö autor märgendustöö jooksul dokumenteeris. Probleemkohtade kaardistamine on oluline, sest see annab informatsiooni selle kohta, millele tulevikus sündmuste märgendusviisi parandamisel tähelepanu osutada.

3.1. Üldine protsess

Töös märgendati eesti keelele kohandatud TimeML märgendusjuhendist lähtuvalt esiteks ajaväljenditega seotud sündmustele viitavad keelendid märgendiga *event*. Kui ajaväljend oli seotud entiteediga, kasutati märgendit *entity*. Seejärel määrati sündmuste puhul märgendi atribuutides sündmuse klass, kindlustunde tase sündmuse klassi määramisel (kõrge (*high*), keskmine (*neutral*) või madal (*low*)), sündmuse kestus ning kindlustunde tase kestuse määramisel (kõrge (*high*), keskmine (*neutral*) või madal (*low*)). Kui tegemist oli mitmesõnalise sündmusega, mille liikmed tuli märgendada ühe sündmusena, märgendati need kas ühes plokis (kui need paiknesid tekstis kõrvuti) või kasutati dialoogiaknas valikut *add frag* (juhul, kui sündmuse liikmed ei paiknenud tekstis kõrvuti), mille abil sai liikmeid ühendada. Kui sündmus kuulus argumenti nõudvasse klassi ning tekstis oli argumentsündmus ilmutatud kujul olemas, tõmmati ülemsündmusest argumentsündmuseni seos *has_argument*. Seejärel määrati sündmuse ja ajaväljendi vaheline ajaseos, mis tõmmati alati ajaväljendist sündmuseni. Argumendiga sündmuste puhul määrati ajaseos mõlemale sündmusele vaid juhtudel, kus näis, et nende sündmuste suhe ajaväljendiga on erinev. Kui ülemsündmuse ja argumentsündmuse suhe ajaväljendiga oli ühesugune, määrati seos vaid ajaväljendi ja ülemsündmuse vahel. Problemaatilistes kohtades lisati sündmusmärgendustele kommentaar.



Joonis 6. **BRAT-i märgendusaken**.

3.2. Tehnilised probleemid

Märgendustöö käigus ilmnes kaks tehnilist probleemkohta, mis puudutasid märgendustööriista BRAT. Esimeseks probleemkohaks olid juhud, kus ajaväljendiga seotud sündmus või entiteet kattus kas osaliselt või täielikult ajaväljendiga. Ajal, mil selle töö raames märgendust teostati ei olnud BRAT-is lubatud üksteise sisse jäävate märgenduste tegemine. Seetõttu jäid sellised sündmused või entiteedid esialgu märgendamata või said vaid osaliselt märgendatud. Näiteid märgendatud tekstidest:

(2) *Kui tulelõõm korraga liiga suur sai, jooksis viieaastane välja abi otsima-kutsuma.*

(3) *"Kirjavahetus Tallinna residentuuriga 1939-40"*

Näites 2 on *viieaastane* märgendatud ajaväljendina, kuid see viitab ka viieaastasele isikule ehk entiteedile. BRAT-i konfiguratsiooni tõttu jäi entiteet märgendamata. Näites 3 on tegemist toimiku pealkirjaga, mis selles märgenduses peaks leidma märgendamist

entiteedina. Kuna ajavahemik *1939-40* on märgendatud ajaväljendina, ei olnud BRAT-i konfiguratsiooni tõttu võimalik seda entiteeti täies mahus märgendada, mistõttu jäi märgendus poolikuks.

Teiseks probleemkohaks olid juhud, kus sündmus koosnes mitmest sõnast, mis tuli märgendada ühe sündmusena, kuid mis ei asunud lauses üksteise kõrval. Kui sündmuse fragmentide vahele jäi teine märgendus (siinses märgenduses valdavalt ajaväljend), polnud võimalik neid fragmente ühe sündmusena märgendada. Näide märgendatud tekstist:

(4) *Me ei ole ka praegu lõpuni kindlad, kas tegemist on autentse materjaliga.*

Märgendusreeglite järgi oleks tulnud näites 4 märgendada *ole* ja *kindlad* ühe sündmusena, kuid kuna nende vahele jäi ajaväljend *praegu* ei võimaldanud BRAT seda teha.

3.3. Mitmeti tõlgendatavus

Märgendustöös tuli kohati ette, et ajaväljendiga seotud keelendit oleks olnud võimalik tõlgendada nii sündmuse kui entiteedina. Sellisel juhul pidi autor otsustama, kummana mingit tüüpi keelendeid märgendada ja sellele süsteemile märgendustöös läbivalt kindlaks jääma. Ei saa aga öelda, et autori poolt valitud tõlgendus kindlalt õigem oleks ning tulevikus pole teistsugused tõlgendused välistatud. Näiteid märgendusest:

(5) *Siin on kaks suurkuju (Joyce ja Faulkner), kellest ei saa üle ega ümber ükski 20. sajandi maailmakirjanduse käsitlus.*
(6) *Ühtekokku on ta kogunud 12 tundi helisalvestisi.*
(7) *1987. aasta lepingus ei olnud loomulikult midagi selle kohta, mis sel juhul maalidest saab.*

Näites 5 saaks *maailmakirjandust* tõlgendada *occurrence* klassist sündmusena ehk maailmas toimuva teoste kirjutamise ja avaldamisena, kuid seda saab tõlgendada ka kui 20. sajandi jooksul maailmas ilmunud teoseid ehk entiteete. Näites 6 on samuti võimalik tõlgendada *helisalvestisi* sündmustena (ehk heli salvestamise tegevusena), kuid võimalik on tõlgendada neid ka kui heli salvestamise tegevuse tulemusi ehk entiteete. Näites 7 saab *lepingut* tõlgendada kui lepingu sõlmimise sündmust, kuid võimalik on ka tõlgendus, et leping on füüsilisel kujul (paberil) eksisteeriv objekt. Autor on otsustanud seda tüüpi juhtumite puhul entiteedi tõlgenduse kasuks, kuid nagu eelnevalt sai mainitud, ei ole entiteedi tõlgendus ainuõige.

Märgendamise käigus võis tulla ka ette, et keeruline oli määrata sündmuse klassi. Eelkõige võis seda juhtuda siis, kui keelendit oli võimalik tõlgendada nii sündmuse kui seisundina. Näide märgendusest:

(8) *Täna on Limesis renoveeritud esimene müüritorn sellest hiigelrajatisest, mis pidas idast peale tungivaid horde kinni peaaegu 200 aastat.*

Näites 8 kuulub märgendamisele *pidas*, mida saaks tõlgendada nii *occurrence* tüüpi sündmusena (kui tõlgendada kinni pidamist müüri või selle kaitsjate aktiivse tegevusena) kui seisundina (klass *state*). Kuna hiigelrajatis ilmselt siiski ise midagi ei teinud, vaid lihtsalt eksisteeris seal kohas selle ajaperioodi vältel, on antud juhul otsustatud seisundi tõlgenduse kasuks. Klassi määramise kindlustasemeks on aga märgitud keskmine kindlustase, mis indikeerib seda, et tõlgenduses ei saa täiesti kindel olla.

3.4. Retoorilised võtted

Märgenduses tuli ette kohti, kus oli kasutatud retoorilisi ehk loo rääkimise võtteid. Täpsemalt tekitas see probleeme siis, kui ajaväljendiga seotud sündmust oli kasutatud loo rääkimise

võttena, sest see mõjutas ajaseose määramist ajaväljendi ja sündmuse vahel. Näiteid märgendusest:

(9) *Hongkongi aktsiaturg on olnud jätkuvas vabalanguses juba teist nädalat , kui möödunud nädala reede välja arvata , teatas Baltic News Service.*
(10) *Vilniuse börsi A-grupi aktsiate käive oli 53,8 miljonit krooni , mis on ligi veerandi võrra väiksem kui eelmisel nädalal.*
(11) *Rocki kümnendeist suuresti sõltumatu vandersellivägi.*

Näites 9 kirjeldatakse aktsiaturu vabalanguses olemise sündmust, millest arvatakse loo jutustaja poolt välja üks lõik (*möödunud reede*). Kuna välja arvamise sündmus leiab aset hiljem, määratakse ajaväljendi ja sündmuse vahele ajaseos *before* ehk ajaväljend eelneb sündmusele. Näites 10 on retooriline võte võrdlus (*on väiksem*) eelneva nädalaga. Kuna võrdlemine toimub hiljem, on ajaseos nende vahel *before*. Näites 11 räägitakse ajaperioodist (*Rocki kümnendid*) ning kasutatakse sellest sõltumatu olemist retoorilise võttena. Seega tegelikult see ajaperiood justkui ei loe. Märgendati *sõltumatu* kui seisund, milles sisaldusid *Rocki kümnendid*, sest ümbritsevas kontekstis oli viidatud, et sõltumatu olemine kestis nii selle ajaperioodi ajal kui ka hiljem. See konkreetne fraas on aga keeruline ning märgenduse korrektsuses ei saa täiesti kindel olla.

3.5. Ilmutatud kujul sündmuste puudumine

Märgendustöö käigus ilmnes, et mõnes tekstis võivad ilmutatud kujul sündmused puududa. Selline oli dokument *aja-eesti_pm_1995_12_20*, mille puhul oli tegemist infoloendiga, mis ei sisaldanudki rangelt võttes lauseid. Mõned read tekstist:

(12) *Mustvee kirik*

24. detsembril kell 18 , 25. detsembril kell 10.30 , 26. ja 31. detsembril kell 17 ja 1. jaanuaril kell.10.30.

Alatskivi kirik

24. detsembril kell 17 , 25. , 26. , 31. detsembril ja 1. jaanuaril kell 11.

Vara kirik

24. ja 31. detsembril kell 16.

Selle teksti puhul oli sündmusviide olemas dokumendi metaandmetes (*jõulu- ja aastavahetuse jumalateenistused*), kuid kuna teiste tekstide puhul metaandmeid ei märgendatud, ei tehtud sellist märgendust ka selles tekstis.

3.6. Ebamäärased ajaväljendid

Üsna sageli tuli märgenduses ette ebamääraseid ajaväljendeid ehk ebamääraseid viiteid minevikule, olevikule või tulevikule. Selliste ajaväljendite puhul ei ole võimalik nende täpset kestust määrata ning seetõttu võivad sellised ajaväljendid osutuda problemaatilisteks sündmuste kestuste, samuti nende ja nendega seotud sündmuste vaheliste ajaseoste määramisel.

Osadel juhtudel väljendab ebamäärane ajaväljend suhteliselt selgelt vaatlushetke ehk lühikest ajaperioodi, mille vältel millegi kohta infot saadi. Kontekstist sõltuvalt ei pruugi aga sündmuse kestus taanduda vaid vaatlushetkele ning võib olla vaatlushetkest pikem. Näiteid märgendusest, kus ebamäärane ajaväljend on vaatlushetk ja sündmuse puhul võib eeldada vaatlushetkest pikemat kestust:

(13) *Praegu on Kaagveres 71 tüdrukut, lühikese aja jooksul on kooli juurde tulnud kümme uut õpilast.*

(14) *Nüüd on tuulik taas püsti.*

Näite 13 muudab keeruliseks see, et pole teada kui kiiresti tüdrukute arv muutub, kuigi on viidatud sellele, et nende arv on muutunud lühikese aja jooksul (ja sellest tulenevalt võib ka edaspidi jätkata lühikese aja jooksul muutumist). Samuti pole teada vaatlushetke täpne kestus ehkki eeldatavalt on see lühike. Selle näite puhul on võetud seisukohaks, et vaatlushetk on artikli kirjutamise aeg ning isegi kui tüdrukute arv muutub näiteks päevadega, on vaatlushetke kestus tõenäoliselt lühem kui koolis nimetatud arvul tüdrukute eksisteerimine. Täiesti kindel ei saa töö autor aga selle tõlgenduse korrektsuses olla, sest tüdrukute arv võib muutuda ka kiiremini. Näites 14 on vaatlushetkeks *nüüd* ja sündmuseks seisund *püsti*. Oletatavasti on tuulik olnud püsti kauem kui näiteks artikli kirjutamise aeg, mõned sekundid, minutid või tunnid, kuid täpset seisundi kestust pole võimalik määrata.

Teisalt võib ebamäärane ajaväljend viidata pikemale ajaperioodile. Lihtsamad juhtumid on näiteks *tulevik* ja *minevik*, mille puhul saab eeldada väga pikka ajaperioodi. Keerulisem on aga näiteks *praegu*, mida võib tõlgendada väga lühikese kestusena, kuid mis võib olla mingites kontekstides tõlgendatav ka väga pika ajaperioodina. Näide märgendusest:

(15) *Praegu moodustavad peamise tööjõu nõukogude ajal hariduse saanud inimesed.*

Näite 15 puhul võib eeldada, et tööjõu põhiline isikkoosseis ei muutu eriti kiiresti (uute põlvkondade piisavas mahus tööturule jõudmine, et muutuda seal domineerivaks grupiks võtab ilmselt aega aastaid, kui mitte aastakümneid). Seetõttu võib oletada, et *praegu* all mõeldakse aastaid kestvat tööjõu moodustamist. Samas ei ole ühtse tõlgenduse leidmine seda tüüpi juhtumitele autori jaoks lihtne olnud. Näiteks fraaside puhul nagu *praegune litsentsikomisjon*, *praegused riigikeeleõpetajad* ja *praegune olukord* on autor tõlgendanud jällegi ajaväljendit *praegu* sündmuse/entiteedi kestusest ajaliselt lühemana. Selline tõlgendus on tekkinud sellest, et näiteks litsentsikomisjoni saab tõlgendada pika aja jooksul eksisteeriva organisatsioonina, mille isikkoosseis vahetub. Seega autori arvates viitas *praegu* seal tolle

aja isikkoosseisule, mis ajaliselt on vaid väike lõik organisatsiooni eksisteerimisest. Sellist lähenemist on rakendatud paljude teiste sama tüüpi juhtumite puhul. Selliste juhtumite juures tekib valikukoht, kas lihtsustada märgendust ehk märgendada süstemaatiliselt ebamäärane ajaväljend ning sellega seotud entiteet või sündmus samaaegsena või püüda olla märgendamisel võimalikult täpne ja tuua välja, et ebamääraste ajaväljenditega seotud sündmuste või entiteetide kestused varieeruvad (nt litsentsikomisjoni isikkoosseis võib muutuda kiiremini kui tööealine elanikkond). Autor püüdis märgendamisel olla täpsem, kuid tema hilisem seisukoht on, et parem oleks olnud tõlgendada selliste juhtumite puhul ajaväljendi ja sündmuse/entiteedi kestust süstemaatiliselt samaaegsena. Masinõppe kontekstis on süstemaatiline märgendus lihtsamini õpitav. Täpsema märgenduse miinuseks ebamääraste ajaväljendite puhul on ka asjaolu, et kuna selliste ajaväljendite kestus on avatud erinevateks tõlgendusteks, siis ei saa täpsema märgenduse korrektsuses alati kindel olla.

3.7. Sisuvaesed verbid ja olema-verbi konstruktsioonid

Märgendustöös võis olla autoril kohati keeruline otsustada sisuvaest verbi sisaldavate konstruktsioonide puhul, kas konstruktsiooni sisuvaene verb kannab piisavalt tähendust, et seda eraldi sündmusena märgendada. Kuna individuaalsel tunnetusel näib olevat selles otsustusprotsessis suur kaal, võivad selliste konstruktsioonide märgendamisel tekkida vead. Näiteid märgendusest:

(16) *Järgmine kohtumine peetakse kolmapäeval Kehras.*
(17) *Varsti peame tõsiselt mõtlema hakkama, kas festivali saabki suurendada.*

Näite 16 puhul on autoril olnud tunnetus, et kohtumise pidamises pole pidamise osa olnud piisavalt tähenduslik, et seda eraldi sündmusena märgendada. Näite 17 puhul on aga autorile tundunud, et pidamine kannab modaalset tähendust (väljendab kohustust mõtlema hakata)

ning leiab seetõttu eraldi sündmusena märgendamist. Nende tõlgenduste korrektsuses ei saa autor aga täiesti kindel olla.

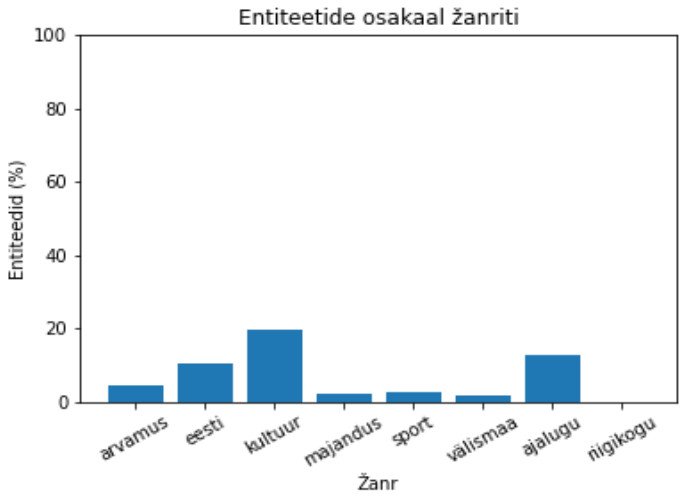
Olema-verbi konstruktsioonide märgendamise muutis keeruliseks see, et nende konstruktsioonide puhul eksisteerib märgendusreeglites palju erijuhte. Seetõttu on võimalik, et *olema*-verbi konstruktsioonide märgendamisel on tehtud kohati vigu. Võrreldes sisuvaest verbi sisaldavate konstruktsioonidega võiks olla nende märgendamine siiski lihtsam, sest juhised on üsna konkreetsed ja jäetud on vähem ruumi individuaalseteks tõlgendusteks.

4. Märgenduse analüüs

Lisaks eesti keele ajaväljendite korpuse alamosas sündmuste, entiteetide, sündmuste kestuste ning ajaseoste märgendamisele on selle töö eesmärgiks uurida, milliseid ajalisi omadusi ajaväljendid nendega seotud keelendite juures kirjeldavad, milline on märgendite struktuur ning milline on nende žanriline jaotus. Selleks teostatakse märgenduse sagedusanalüüs. Töö raames märgendati eesti keele ajaväljendite korpusest 85 teksti, mis sisaldasid 1282 märgendatud ajaväljendit. Töö autor märgendas kokku 1329 üksust, millest ajaväljendiga seotud sündmuseid oli 1233 ja ajaväljendiga seotud entiteete 96. Ajaseoseid märgendati kokku 1290.

4.1. Ajaväljendite seotus entiteetidega

Entiteetide märgendamine TimeML raamistikus oli eksperimentaalne. Selles töös soovitakse kaardistada, kui suurt rolli ajaväljenditega seotud entiteedid märgenduses üldse mängivad ning kas tekstižanrite lõikes joonistub välja erisusi entiteetide jaotuses. Märgendatud tekstides moodustavad entiteedid vaid 7,2% kõigist autori poolt märgendatud üksustest, seega on ajaväljenditega seotud entiteetide esinemissagedus võrreldes sündmustega madal.



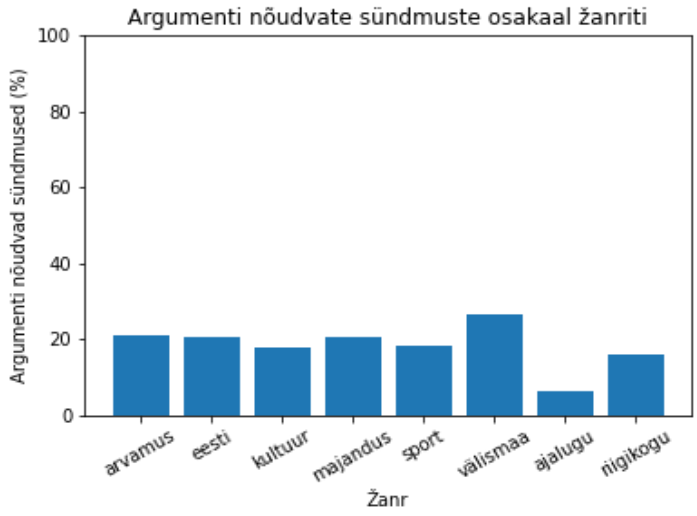
Joonis 7. **Ajaväljendiga seotud entiteetide osakaalud eri žanrites**.

Jooniselt 7 on näha, et kuigi entiteetide osakaal on kõigis žanrites võrdlemisi madal, varieerub nende osakaal žanrite lõikes. Majandusuudistes, spordiuudistes ning välisuudistes leidub ajaväljenditega seotud entiteete väga vähe ning riigikogu stenogrammides ei esine ühtegi ajaväljendiga seotud entiteeti. Mõnevõrra rohkem esineb ajaväljenditega seotud entiteete kultuuriuudistes, ajalooartiklites ja eesti uudistes. Entiteetide lähemal uurimisel ilmnes, et siinses märgenduses on entiteetidena märgendatud peamiselt isikuid (nt *poisike*, *Tõnu*, *peatunnistaja*), organisatsioone (nt *Palestiina valitsus*) ja esemeid (nt *veesõiduk*). Vähem esineb entiteetide seas kohti (nt *Ruhnu*). Entiteetide seas esineb ka arvestataval määral (kokku 17 korral) viiteid pealkirjadele (nt *"Kristluse ajalugu I"*), kuid enamus neist (15) esineb ühes eesti uudiste rubriiki kuuluvas tekstis, kus on toodud välja nimekiri toimikute pealkirjadest, mida polnud BRAT-i konfiguratsiooni tõttu võimalik lõpuni märgendada (nt *Eesti poliitiline politsei 1932-39*). Üks viide pealkirjale esineb kultuuriuudiste rubriiki kuuluvas artiklis ja üks ajalooalases teadusartiklis. Viide autorile esineb märgenduses vaid ühel korral (*Faehlmann, 1839*) ajalooalases teadusartiklis.

4.2. Sündmuste klassid ja struktureeritus

Sündmuse struktureeritus mõjutab seda, kui lihtne või keeruline on sündmuseid käsitsi märgendada ning automaatselt analüüsida. Töö autor uurib, kui palju esineb märgenduses struktureerituid ehk argumenti nõudvaid sündmuseid, kas žanrite lõikes esineb erinevusi struktureeritud sündmuste osakaaludes ning kuidas jagunevad sündmused klassidesse. Kuna sündmustele klassi määramisel märkis autor ka klassi määramise kindlustaseme, uuritakse, kas mingite klasside määramine oli autori jaoks keerulisem.

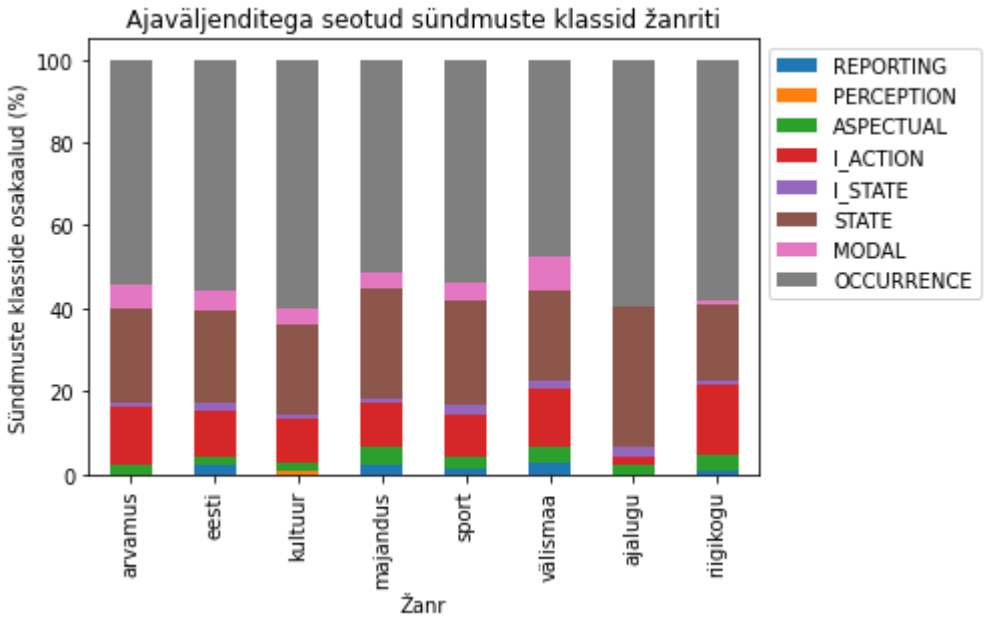
Märgenduses kasutatud eesti keelele kohandatud TimeML skeemi järgi on argumenti nõudvaid sündmuste klasse kuus: *reporting*, *perception*, *aspectual*, *i_action*, *i_state* ja *modal*. Märgendatud tekstides on kokku 266 argumenti nõudvast klassist sündmust, mis moodustavad 21,6% kõigist märgendatud sündmustest. Seega on argumenti nõudvaid ehk rohkem struktureeritud sündmusi märgendatud materjalis suhteliselt vähe.



Joonis 8. **Argumenti nõudvate sündmuste osakaal eri žanrites**.

Jooniselt 8 on näha, et enamike žanrite lõikes on argumenti nõudvate sündmuste osakaalud suhteliselt võrdsed. Erinevus tuleb välja ajalooartiklite puhul, kus argumenti nõudvate

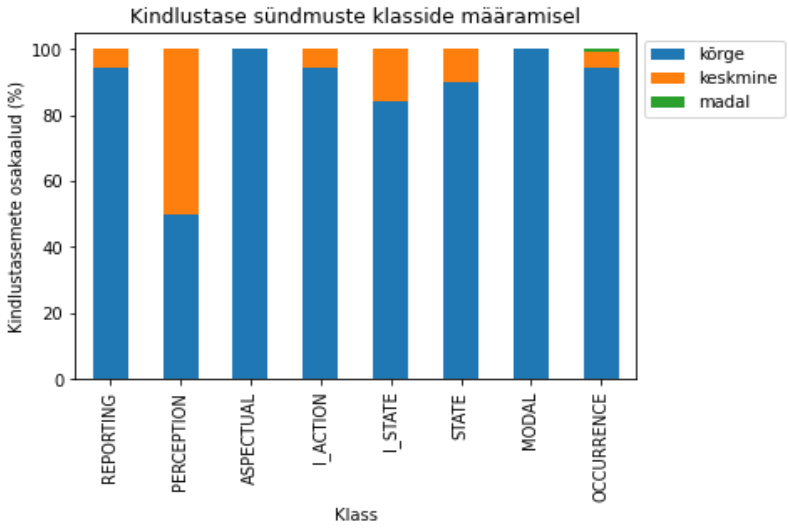
sündmuste osakaal on võrreldes teiste žanritega märgatavalt väiksem. Seega võiks olla ajalooartikleid võrreldes teiste žanritega mõnevõrra lihtsam automaatselt analüüsida, sest struktuuri on vähem, kuid seda saab lõplikult kinnitada alles masinõppe eksperimentidega. Kuna aga kõigis žanrites on suurema osakaaluga argumenti mittenõudvad sündmused ei saa öelda, et mõnda neist võiks olla teistega võrreldes oluliselt keerulisem automaatselt analüüsida.



Joonis 9. **Ajaväljendiga seotud sündmuste klasside osakaalud eri žanrites**.

Jooniselt 9 on näha, et kõikides žanrites on suurima osakaaluga klassi *occurrence* kuuluvad sündmused ning suuruselt teise osakaaluga klassi *state* kuuluvad sündmused. See on ootuspärane, sest klassid *state* ja *occurrence* on kõige üldisemad sündmusklassid. Samuti on ilmnenud varasemates uurimustes, et need kaks klassi domineerivad teiste üle (Bittar jt 2011). Kui muidu on sündmuste klasside jagunemine žanrite lõikes sarnane, on näha märgatavat erinevust ajalooartiklite puhul. Ajalooartiklites esineb võrreldes teiste žanritega vähem sündmuseid klassist *i_action* ning sündmused klassidest *reporting*, *perception* ja

modal puuduvad täielikult. See viitab sellele, et ajalooartiklites puudub sagedamini sündmuses tahtlik osaleja ning puudub täielikult raporteeriv stiil (keegi nägi või teatas midagi). Modaalverbide puudumine viitab sellele, et vähemalt siinses märgenduses puuduvad ajalooartiklites hinnangud teiste sündmuste võimalikkusele või vajalikkusele. Klassi *perception*, mis hõlmab millegi füüsilisele tajumisele viitavaid sündmuseid esines vaid eesti uudistes ja kultuuriuudistes. Mõlemas žanris esines sellest klassist sündmusi aga vaid ühel korral, mistõttu ei saa selle sündmusklassi esinemise kohta neis žanrites eriti järeldusi teha.



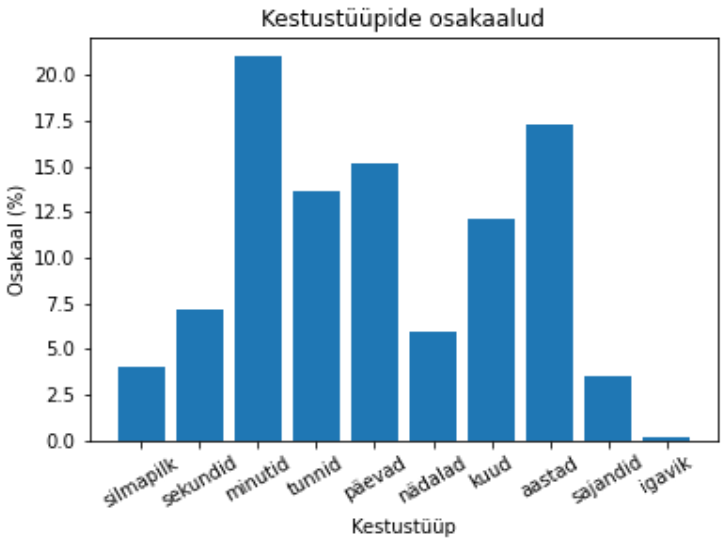
Joonis 10. **Klassi määramise kindlustasemete osakaalud sündmuste klasside lõikes**.

Jooniselt 10 on näha, et valdavalt on töö autori kindlustase sündmuse klassi määramisel olnud kõrge. Kuna klassi *perception* esines märgenduses vaid kahel korral, ei saa selle klassi määramise kindlustasemete kohta kindlaid järeldusi teha. Nähtavasti on keskmist kindlustaset esinenud mõnevõrra rohkem klasside *i_state* ja *state* määramisel, mis võib olla seletatav sellega, et seisundite puhul võis esineda kohati mitmeti tõlgendatavust (ehk kas tegemist oli seisundi või *occurrence* klassist sündmusega). Tähelepanuväärne on, et kuigi madalat kindlustaset klassi määramisel esines kokku vaid viiel korral, esinesid kõik need juhud klassi *occurrence* määramisel. See võib olla samuti seletatav mitmeti tõlgendatavuse probleemiga (kas kahtlus klasside *state* ja *occurrence* vahel või kahtlus, kas tõlgendada

keelendit entiteedi või sündmusena). Ainult klassi *modal* määramisel oli kindlustase alati kõrge ilmselt seetõttu, et sellesse klassi said kuuluda ainult modaalverbid.

4.3. Sündmuste ajalised kestused

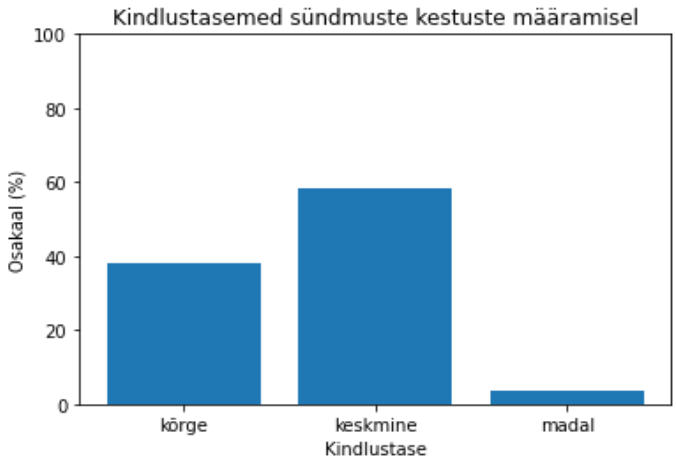
Selles töös märgendati ajaväljenditega seotud sündmuste ajalised kestused, mida pole autorile teadaolevalt eesti keele kontekstis varem tehtud. Töö autor uurib, kuidas jagunevad märgenduses sündmuste kestused ja kindlustasemed kestuste määramisel. Kuna töö raames märgendatud tekstid pärinevad erinevatest žanritest (ajakirjanduse eri rubriigid, ajalooartiklid ning riigikogu stenogrammid) uuritakse ka sündmuste kestuste jagunemist žanrite lõikes. Samuti uurib autor, kas sündmuse klassil on seos sündmuse kestusega.



Joonis 11. **Sündmuste kestustüüpide osakaalud märgendatud tekstides**.

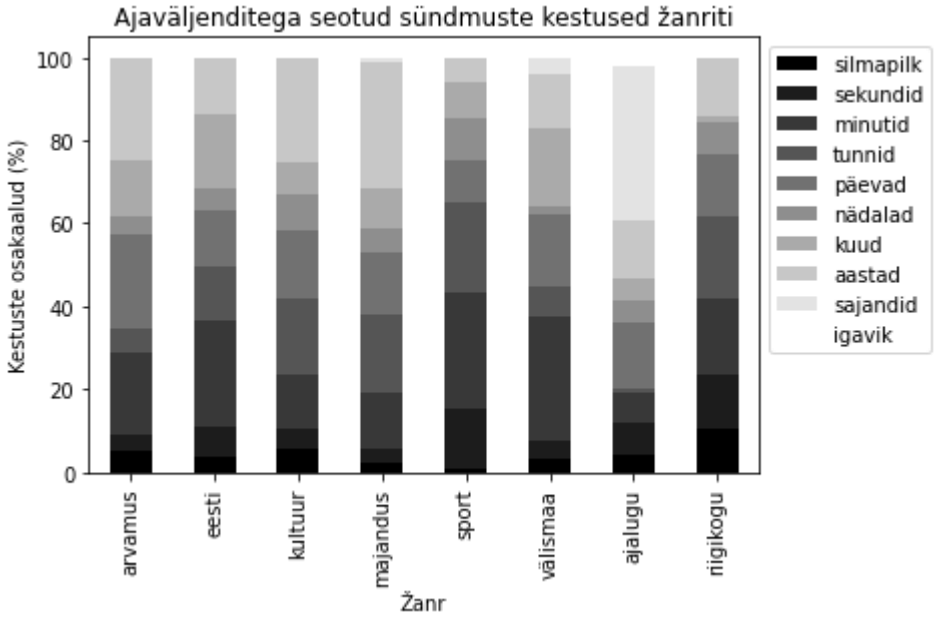
Jooniselt 11 on näha, et märgendatud tekstides on suurima osakaaluga minuteid kestnud sündmused. Samuti esineb palju tunde, päevi, kuid ja aastaid kestnud sündmuseid. Märgatavalt vähem esineb silmapilkseid või sekundeid kestnud sündmuseid, samuti

sajandeid või igavesti kestnud sündmuseid. Sellest järeldub, et eriti lühikese ja eriti pika kestusega sündmuseid esineb tekstides vähe ning domineerima kalduvad pigem keskmise pikkusega sündmused. Vähe esineb aga ka nädalaid kestnud sündmuseid, mis on huvitav, sest tegemist on pikkuselt üsna keskmise kestustüübiga. Autor ei oska sellele seletust pakkuda.



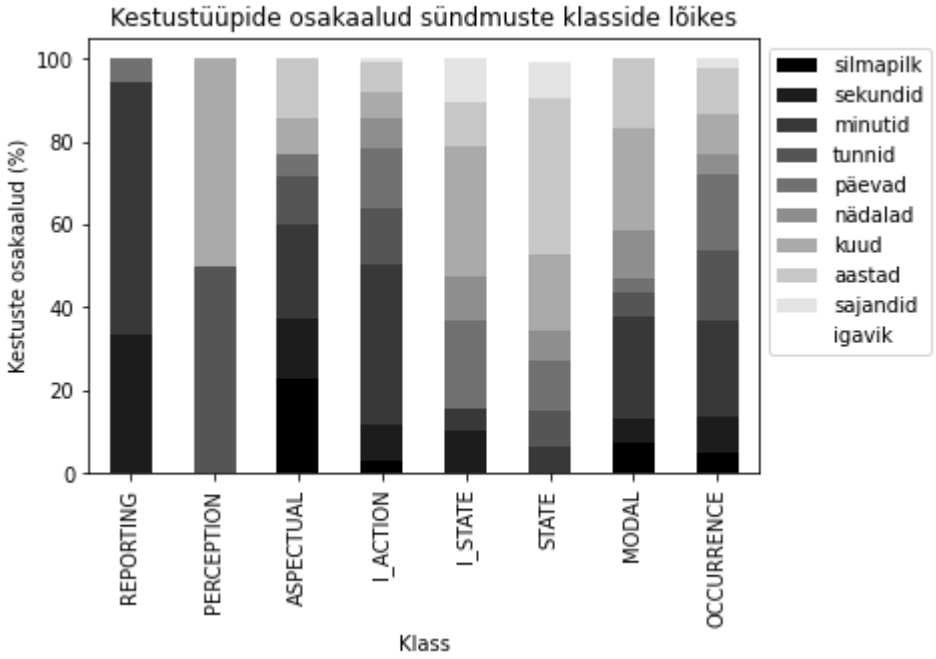
Joonis 12. **Sündmuste kestuste määramise kindlustasemete osakaalud**.

Jooniselt 12 on näha, et kõige sagedamini oli kindlustase sündmuse kestuse määramisel keskmine, millele järgnes kõrge kindlustase. Madalat kindlustaset esines harva. See viitab sellele, et valdavalt oli autoril kas kontekstist või isiklikest teadmistest tulenevalt mingi ettekujutus sündmuse võimalikust kestusest.



Joonis 13. **Ajaväljenditega seotud sündmuste ajaliste kestuste osakaalud žanrite lõikes**.

Jooniselt 13 on näha, et žanrite lõikes on sündmuste ajaliste kestuste jagunemises erinevusi. Ootuspärane on, et ajalooartiklites on väga pika kestusega sündmuste osakaal suurim ning palju on kajastatud sajandeid kestnud sündmuseid. Teiste žanrite puhul on näha, et valdavalt küündivad neis kajastatud sündmuste kestused aastateni, vaid üksikutel juhtudel on majandus- ja välisuudistes kajastatud ka sajandeid kestnud sündmuseid. Lühema kestusega sündmused domineerivad enim spordiuudistes ja riigikogu stenogrammides. See on üsna ootuspärane, sest spordiuudistes kajastatakse sageli spordisündmuseid, mille kestus ei ole enamasti pikk ning raporteeritakse nende spordisündmuste jooksul toimunut. Riigikogu stenogrammid aga kajastavad Riigikogu istungeid, mille kestus on samuti suhteliselt lühike ning kus räägitakse ilmselt peamiselt päevakajalistel teemadel.



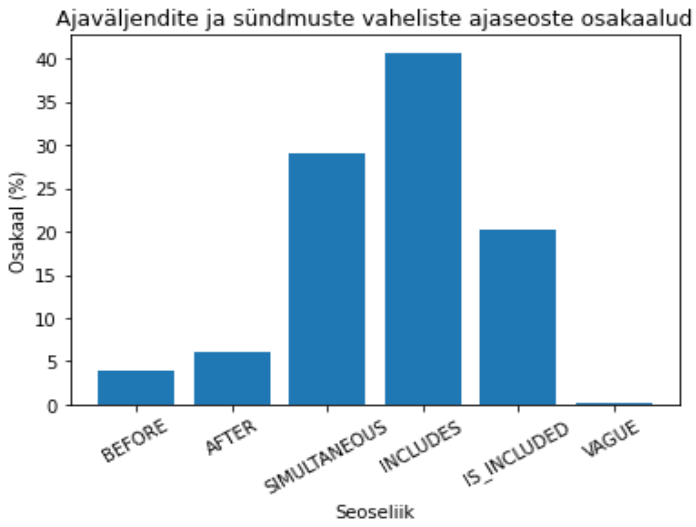
Joonis 14. **Sündmuste kestustüüpide osakaalud sündmuste klasside lõikes**.

Jooniselt 14 nähtub, et argumenti nõudvatest sündmusklassidest domineerivad lühikesed kestustüübid klassides *aspectual*, *reporting* ja *i_action*. See viitab sellele, et millegist teavitamisele ja mingi teise sündmuse algusele, keskosale või lõpule viitavad ning ka tahtlikku tegevust väljendavad sündmused kestavad üldiselt lühikest aega. Klassides *i_state* ja *state* on lühikeste kestustüüpide osakaal kõige väiksem. See viitab sellele, et nii tahtlikud kui tahtest sõltumatud seisundid kestavad tavaliselt pikema ajaperioodi vältel. Võrdlemisi väike on lühikeste kestustüüpide osakaal ka klassis *modal*, mis viitab sellele, et mingi sündmuse toimumise võimalikkusele või vajalikkusele antavad hinnangud kalduvad kestma mõnevõrra pikema ajaperioodi vältel. Klassi *perception* esines siinses märgenduses vaid kahel korral, seetõttu ei saa selle klassi kohta erilisi järeldusi teha. Klassis *occurrence* on aga jällegi lühemate kestustüüpide osakaal üsna suur, mis võib olla seletatav sellega, et sellesse klassi määratakse palju eri tüüpi sündmuseid, kaasa arvatud sündmuseid, mis muidu sobituksid klassidesse *aspectual*, *reporting* ja *i_action* (mille seas domineerivad lühikesed

kestustüübid). Samuti võib seda seletada see, et seisundeid kirjeldavaid sündmuseid, kus domineerivad pikemad kestustüübid, *occurrence* klassi ei määrata.

4.4. Ajaväljendite ja sündmuste või entiteetide vahelised ajaseosed

Ajaseosed määravad ajaväljendite ja sündmuste vahelised suhted. Siinses töös kasutati varasemaga võrreldes väiksemal arvu eri ajaseosetüüpe ning katsetati üldise ülekattuvusseose asemel täpsemate ülekattuvusseoste kasutuselevõttu. Töö autor uurib, kui suurel määral tähistab märgenduses ajaväljend sündmuse toimumisaega (ehk sündmus toimub ajaväljendi sees või sellega samaaegselt) ning mil määral märgib ajaväljend sündmuse juures midagi muud. Selleks vaadeldakse eri ajaseosetüüpide jagunemist ajaväljenditega seotud sündmuste lõikes.



Joonis 15. **Ajaväljendite ja sündmuste vaheliste ajaseoste (A seos B, kus A on ajaväljend ja B sündmus) osakaalud märgendatud tekstides**.

Jooniselt 15 on näha, et märgendatud tekstides on ajaväljendite ja sündmuste vahel enim *includes* tüüpi ajaseoseid ehk sündmuse sisaldumist ajaväljendis. Samuti esineb märgatavalt

simultaneous tüüpi ajaseoseid ja mõnevõrra vähem *is_included* tüüpi ajaseoseid. See viitab sellele, et valdavalt tähistab ajaväljend sündmuse toimumisaega. Näiteid märgendatud tekstidest:

(18) *11. mai kontserdil esineb ka Arbo Valdma klaveril.*
(19) *Juba mitu päeva mälub Eesti press seda teemat nagu lehm võilille ning hiilgab luterlikust eetikast kantud avaldustega.*

Näites 18 on ajaväljend *11. mai* ja sündmus *kontsert* toimub selle kuupäeva sees (seos *includes*), sest on eeldatav, et kontsert ei kesta terve kuupäeva vältel. Näites 19 on ajaväljend *mitu päeva* ning kontekstist tulenevalt toimub sündmus *mälub* ajaväljendiga samaaegselt (seos *simultaneous*).

Mõnevõrra vähem esineb *is_included* tüüpi seoseid ehk ajaväljendi sisaldumist sündmuse toimumisaja sees. Näide märgendusest:

(20) *Autojuhist isa oli hommikust peale tööl.*

Näites 20 on ajaväljend *hommikust* ning sündmusena on märgendatud seisund *oli*. Kontekstist selgub, et kuigi tööl olemine sai alguse hommikul, on see kestnud hommikust kauem. Seega selles näites piiritleb ajaväljend sündmuse algust ning sisaldub sündmuse sees (seos *is_included*).

Samas ei pruugi seosetüüp *is_included* alati piiritleda sündmuse toimumisaja algust või lõppu: see võib ka markeerida, et ajaväljend on lühem vaatlushetk sündmuse pikema toimumisaja sees. Näide märgendusest:

(21) *Veel mõni aasta tagasi valdavalt venekeelne linn räägib nüüd hoolega ukraina keelt.*

Näites 21 on seotud ajaväljend *mõni aasta tagasi* ja seisund *venekeelne* ning ajaväljend *nüüd* ja sündmus *räägib*. Mõlemad ajaväljendid näitelauses on ebamäärased, kuid võib oletada, et vene keel on olnud linnas laialdaselt kasutusel pika aja vältel ning mõne aasta tagune aeg kirjeldab lühemat vaatlushetke pikemast ajaperioodist. Samuti võib oletada, et *nüüd* kirjeldab pigem lühikest oleviku vaatlushetke ning ukraina keele rääkimine linnas kestab ilmselt edasi ka tulevikus. Seega sisaldub mõlemal juhul vaatlushetk sündmuse pikema toimumisaja sees (seos *is_included*).

Seosetüübid *before* ja *after* ehk eelnevus- ja järgnevusseosed on esinenud harvem. See tähendab, et enamasti on ajaväljendid kirjeldanud sündmuse ajalisi omadusi ning harvemad on olnud juhud, kus tekstis on kirjeldatud ajaväljendi ja sellega seotud sündmuse paigutust ajateljel üksteise suhtes või on ajaväljendiga seotud sündmust kasutatatud retoorilise võttena. Näide märgendatud tekstist:

(22) *Kehra oli selleks ajaks tabanud neli korda.*

Näide 22 pärineb käsipallivõistlust kirjeldanud artiklist. Ajaväljend *selleks ajaks* kirjeldab mingit ajalist punkti (vaatlushetke) mängu jooksul, tabamise sündmused on toimunud selle ajalise punkti suhtes varem. Võib öelda, et selle näite puhul on seos *after* tähistanud vaatluspunkti nihkumist.

Seosetüüpi *vague* ehk teksti põhjal raskesti määratavat ajaseost ajaväljendi ja sündmuse vahel esines märgenduses vaid ühel korral:

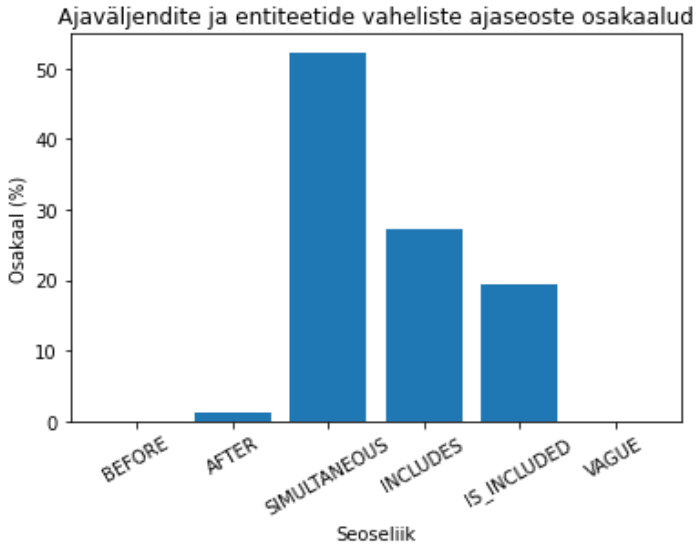
(23) *Iisraeli väed laiendasid teisipäeval oma haaret Läänekaldal, tappes neli palestiinlasest politseinikku, seda mõni tund pärast seda, kui USA president George Walker Bush kutsus üles vahetama Palestiina juhtkonda.*

Näites 23 määrati seostüüp *vague* ajaväljendi *mõni tund* ja sündmuse *vahetama* vahele. Põhjuseks oli see, et juhtkonna vahetamise sündmuse (mis oli üleskutse sündmuse argument)

kohta polnud teada, kas see ka tegelikult hiljem toimus (seega tegemist oli irreaalse sündmusega). Teadaolevalt toimus vaid üleskutse juhtkonna vahetamiseks.

Võib järeldada, et enamasti on ajaväljend märkinud sellega seotud sündmuse toimumisaega ning piiritlenud sündmuse kestust (piirid seab ajaväljendis sisaldumine (seos *includes*) ning sellega samaaegselt toimumine (seos *simultaneous*)). Sellest tulenevalt peaks saama enamasti ajaväljendi põhjal ka sündmuse kestust oletada.

Töö autor uurib ka seda, milliseid ajaseoseid ajaväljendite ja entiteetide vahel esineb ja kuidas need jagunevad, sest see võimaldab teha järeldusi selle kohta, mida entiteediga seotud ajaväljend entiteedi juures tavaliselt kirjeldab. Kõiki tekstides esinenud entiteete analüüsi ei võeta, sest osad neist jäid BRAT-i konfiguratsiooni tõttu lõpuni märgendamata ning seetõttu ei märgendatud ka nende seost ajaväljendiga.

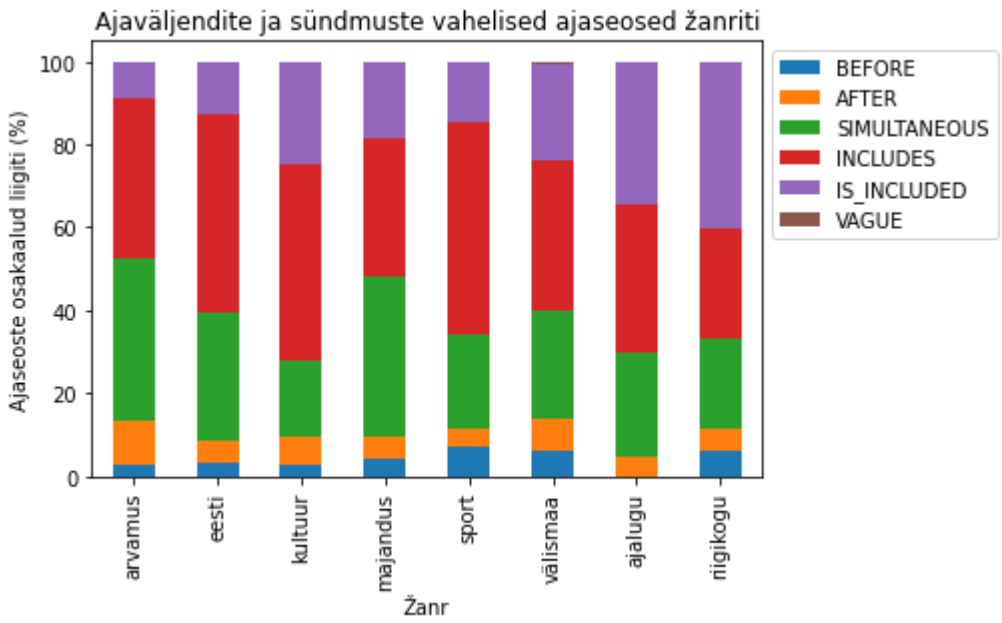


Joonis 16. **Ajaväljendite ja entiteetide vaheliste ajaseoseoste (A seos B, kus A on ajaväljend ja B entiteet) osakaalud märgendatud tekstides**.

Jooniselt 16 on näha, et kõige sagedamini on entiteedi ja ajaväljendi vaheline ajaseos olnud *simultaneous* ehk samaaegne. Märgendatud entiteetide lähemal uurimisel ilmneb, et selle seosetüübi puhul on ajaväljend enamasti kirjeldanud mingi eluta või elusa objekti vanust (nt *17-aastased tüdrukud*) või kestust (nt *48-minutiline film*).

Ajaseosed *includes* ja *is_included* on samuti märgatavalt esindatud. Näiteks fraasis *lähimineviku taiesed* on taiesed kui kunstiteosed tekkinud lähimineviku kui pikema ajaperioodi sees. Fraas *"Keisri hull" (1978)* kirjeldab aga näiteks teost ehk entiteeti, mis on ilmunud 1978. aasta sees. Fraasis *praegune Saaremaa* viitab ajaväljend vaatlushetkele, mille kestus on Saaremaa kui geograafilise nähtuse eksisteerimise kestusest tõenäoliselt lühem. Entiteetide, ajaväljendite ja nende vaheliste seoste lähema vaatluse tulemusel võib öelda, et sellised sisaldumis- ja kaasaarvamisseosed ajaväljendite ja entiteetide vahel viitavad valdavalt kas mingi objekti ilmumis- või tekkimisajale (entiteet ajaväljendi sees) või on tekstis kirjeldatud lühemat vaatlushetketke entiteedi pikemast eksisteerimisest (ajaväljend entiteedi sees).
Eelnevus- ega järgnevusseoseid (*before* ja *after*) ei esinenud ajaväljendite ja entiteetide vahel peaaegu üldse (*after* tüüpi seost esines vaid ühel korral) ning seosetüüpi *vague* ehk ebamäärast seost entiteetide ja ajaväljendite vahel ei esinenudki.

Viimaks uurib töö autor, kas ja kuidas varieeruvad ajaväljendite ja sündmuste vaheliste eri tüüpi ajaseoste osakaalud tekstižanrite lõikes. See võib anda aimu, kas mingites žanrites võib olla lihtsam määrata sündmuste kestuseid.



Joonis 17. **Ajaväljendite ja sündmuste vaheliste ajaseoste (A seos B, kus A on ajaväljend ja B sündmus) osakaalud žanrite lõikes**.

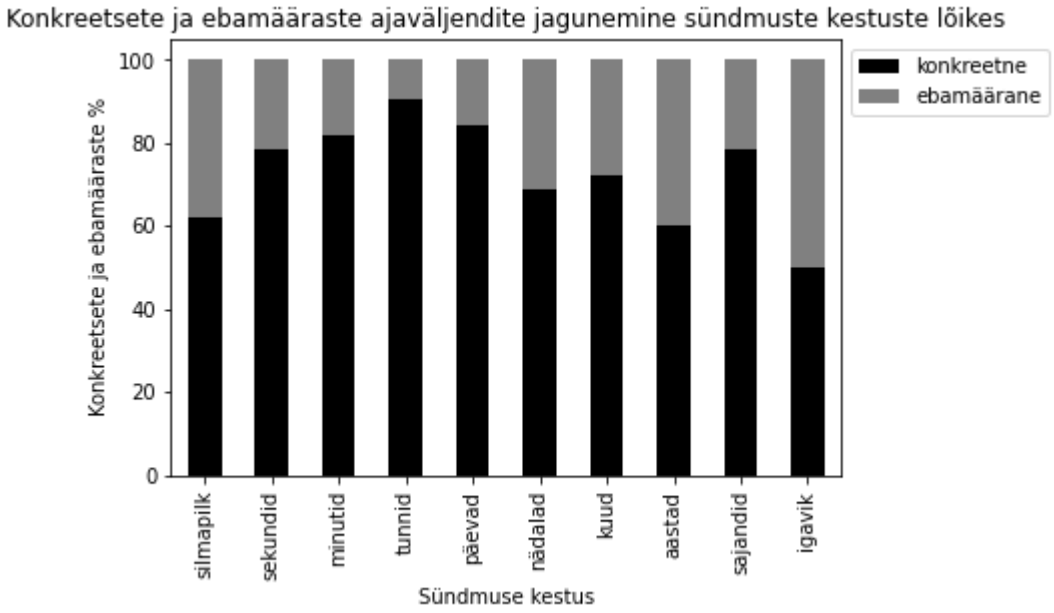
Jooniselt 17 on näha, et ajaväljendite ja sündmuste vaheliste ajaseoste osakaalud mõnevõrra varieeruvad žanrite lõikes. Üldiselt on žanrite lõikes enim esindatud *includes* ja *simultaneous* tüüpi seosed, mis näitab, et ajaväljend on tavaliselt tähistanud sündmuse toimumisaega. Joonistub välja, et ajaväljendiga samaaegselt toimunud sündmuste osakaal on suurim arvamus- ja majandusuudistes. Seosetüüpi *is_included* on esinenud enim riigikogu stenogrammides (kus see on domineeriv seosetüüp) ja ajalooartiklites, aga arvestataval määral ka kultuuri- ja välisuudistes (umbes võrdselt seosetüübiga *includes*). See viitab sellele, et iseäranis riigikogu stenogrammides ja ajalooartiklites on rohkem esinenud juhtumeid, kus ajaväljend on kas piiritlenud sündmuse algust või lõppu või on ajaväljend olnud lühem vaatlushetk sündmuse pikema toimumisaja sees. Eelnevus- ja järgnevusseosed (*before* ja *after*) on kõigis žanrites vähe esindatud. Sündmustele usaldusväärselt kestuse määramise kontekstis on see informatsioon väärtuslik: *includes* ja *simultaneous* tüüpi seoste suuremad osakaalud mingites žanrites näitavad, et neist žanritest tekstides võiks saada sündmuste kestuseid kindlamalt määrata, sest need on rohkematel juhtudel ajaväljenditega

piiritletud. Siinsete tulemuste põhjal näib, et arvamus- ja majandusuudistes võib olla mõnevõrra lihtsam sündmuste kestuseid määrata sealse mõnevõrra suurema *simultaneous* tüüpi ajaseose esinemise tõttu, mis viitab kõige täpsemale sündmuse kestuse piiritlemisele ajaväljendi poolt. Kuna aga ka *includes* tüüpi ajaseos lihtsustab sündmuse kestuse määramist, sest see määrab sündmuse võimaliku kestuse ülempiiri, võib öelda, et enamikes ajakirjandusrubriikides võiks olla sündmuste kestuste määramine lihtsam. Oluline on aga märkida, et *includes* ja *simultaneous* tüüpi ajaseosed lihtsustavad sündmuse kestuse määramist vaid siis, kui sündmusega seotud ajaväljendil on konkreetne väärtus. Ebamäärased ajväjendid sündmuse kestuse kohta täpset informatsiooni ei anna. Mõnevõrra keerulisem võib sündmuste kestuste määramine olla kultuuriuudistes, välisuudistes, ajalooartiklites ning riigikogu stenogrammides sealse mõnevõrra suurema *is_included* tüüpi ajaseose esinemise tõttu, sest see seosetüüp ei märgi alati sündmuse algust või lõppu, vaid võib märkida ka lühemat vaatlushetke pikemast sündmusest, mis sündmuse kestuse kohta informatsiooni ei anna.

4.5. Ajaväljendi konkreetsuse seos sündmuse kestuse määramisega

Ajaväljendeid on eri liiki ning need jagunevad konkreetseteks ja ebamäärasteks aegadeks. Konkreetsed ajaväljendid võivad olla kalendrilised ajad (nt *02.02.2022*), ajavahemikud (nt *kuue kuu jooksul*) või ajalised korduvused (nt *pühapäeviti*). Ebamäärased ajaväljendid on ebamäärased viited minevikule (nt *hiljuti*), olevikule (nt *praegu*) või tulevikule (nt *varsti*), milles puudub kalendriline informatsioon. Kuna selle töö raames tehtud märgenduses on iga sündmus seotud ajaväljendiga, uurib töö autor, kas ajaväljendi konkreetsusel on seos sündmuse kestusega ning kindlustasemega sündmuse kestuse määramisel. Ajaväljendi konkreetsuse seost sündmuse klassiga autor ei uuri, sest juba mingi keelendi sündmuseks lugemisel pidi ajaväljendiga seotuse tingimus olema täidetud ning sündmuse klassi määramisel lähtuti keelendi grammatilistest ja semantilistest omadustest, millel puudub seos ajaväljendi väärtusega. Ebamäärased ajaväljendid moodustavad siinses märgenduses 21,8%

kõigist ajaväljenditest, seega esineb neid konkreetsete ajaväljenditega võrreldes vähem, kuid siiski arvestataval hulgal. Analüüsist jääb välja 6 ajaväljendi ja sündmuse paari, sest nende puhul puudub ajaväljendile määratud väärtus.

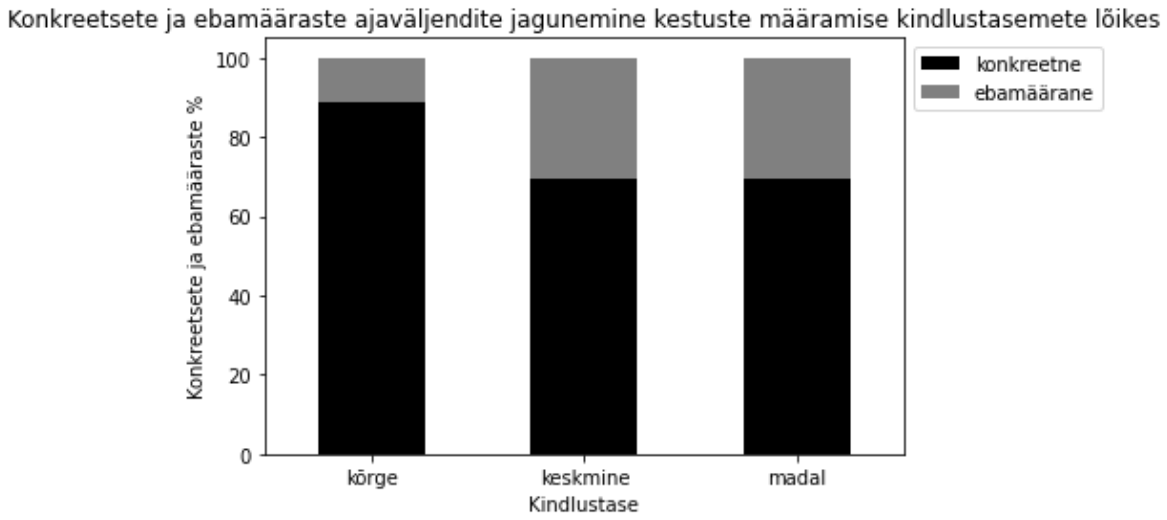


Joonis 18. **Konkreetsete ja ebamääraste ajaväljendite jagunemine sündmuste kestustüüpide lõikes**.

Jooniselt 18 on näha, et konkreetsete ja ebamääraste ajaväljendite osakaalude suhe kestustüüpide lõikes varieerub. See viitab sellele, et ajaväljendi konkreetsusel võib olla mõju sündmuse kestuse määramisele. Tuleb märkida, et sündmuseid, mille kestuseks määrati igavik esineb märgendatud tekstides kokku vaid kahel korral. Seetõttu ei saa selle kestustüübi kohta kindlaid järeldusi teha.

Jooniselt 18 nähtub, et üldiselt esinevad lühema kestusega sündmused (mille kestus ulatub sekunditesse, minutitesse, tundidesse või päevadesse) võrreldes pikema kestusega sündmustega (mille kestus ulatub nädalatesse, kuudesse või aastatesse) rohkem koos konkreetsete ajaväljenditega. Lühemad kestustüübid on täpsemad, sest nende vahelised ajalised erinevused on väiksemad (nt ajaline erinevus 30 sekundi ja kahe minuti vahel on

väiksem kui erinevus kuue kuu ja kahe aasta vahel). Sellest tulenevalt on lühema kestusega sündmuste puhul keerulisem otsustada, milline lühem kestustüüp neile täpselt sobitub ning on tõenäolisem, et kestustüübi määramisel võetakse arvesse sündmusega seotud ajaväljendi väärtust. Erandiks on lühemate kestustüüpide seas silmapilk. Silmapilgu vältel kestvate sündmuste puhul esineb võrdlemisi rohkem seotust ebamääraste ajaväljenditega. See on seletatav sellega, et mingile sündmusele silmapilkse kestuse määramisel lähtutakse sageli sündmuse semantikast või ümbritsevast kontekstist ning vähem ajaväljendi väärtusest. Pikemate kestustüüpide seas eristub teistest sajanditepikkune kestus, mille puhul esineb võrdlemisi vähem seotust ebamääraste ajaväljenditega. See on ootuspärane, sest tegemist on eriti pika ajalise kestusega ning otsustamisel, kas määrata pika kestusega sündmuse kestustüübiks aastaid või sajandeid, on võetud tõenäoliselt arvesse ajaväljendi väärtust.



Joonis 19. **Konkreetsete ja ebamääraste ajaväljendite jagunemine sündmuste kestuste määramise kindlustasemete lõikes**.

Jooniselt 19 on näha, et ajaväljendi konkreetsusel on tõenäoliselt olnud mõju sellele, kui enesekindel on töö autor olnud sündmuste ajaliste kestuste määramisel. Kindlus sündmuse kestuse määramisel on olnud kõrge valdavalt siis, kui sündmusega seotud ajaväljend on olnud konkreetne. Keskmise ja madala kindlustaseme vahel aga erinevusi ei ole ning mõlema

puhul on ebamäärased ajaväljendid võrreldes kõrge kindlustaseme juhtudega rohkem esindatud.

Kuigi statistikast ilmneb, et ajaväljendi konkreetsusel ilmselt on seos nii sündmuse kestuse määramisega kui autori kindlustasemega kestuse määramisel, tuleb võtta arvesse, et kuigi sündmuste kestuste määramisel lähtuti märgendustöös sageli tõepoolest ajaväljendite väärtustest, pidi autor kohati lähtuma isiklikest teadmistest sündmuste tüüpiliste kestuste kohta või võimalikku kestust oletama.

Selle, kas ajaväljendi väärtusest sõltub sündmuse kestus, määrab kontekst. Järgnevalt tuuakse märgendatud tekstidest välja mõned näitelaused, kus eri tüüpi ajaväljendid on määranud sündmuste kestust:

(24) *Istusin seal keldris umbes tund aega, kuid mehed olid tegelikult väga viisakad*

(25) *Üleeilne päev oli ilmselt üks rõõmsamaid lehelugejatele ja masendavamaid ajakirjanikele.*

(26) *Muudatusettepanekute tähtajaks pakuti 20. september kell 12.*

Näites 24 on ajaväljend *umbes tund aega duration*-tüüpi ehk ajaline kestus ning sündmus *istusin* toimub selle ajaperioodiga samaaegselt ehk aastaid. Näites 25 on ajaväljend *üleeilne päev date*-tüüpi ehk päeva täpsusega määratud kalendriaeg ning seisund *oli* kestab kogu selle kuupäeva vältel ehk tunde. Näites 26 on ajaväljend *22. september kell 12 time*-tüüpi ehk kellaaja täpsusega kalendriaeg ning sündmus *tähtaeg* on sellest tulenevalt ajaliselt silmapilkne. Selliste juhtumite puhul märkis autor kestuse määramise kindlustasemeks *high* ehk kõrge.

Kui ajaväljendi kestusel puudus konteksti põhjal seos sündmuse kestusega, pidi autor sündmuse võimalikku kestust oletama. Enesekindluse tase kestuse määramisel oli kõrge peamiselt juhtudel, kus sündmuse kestus oli autorile teada (nt *olümpiamängud*).

Paljudel juhtudel ei olnud sündmuse kestus aga autorile täpselt teada, kuid autoril oli ettekujutus sündmuse ligikaudsest kestusest.

(27) *Helistasin hiljem mitu korda mobiilile ning lõpuks ütles teine mees, et Ennul polevat sigu enam vaja olnud.*

Näites 27 toimub helistamise sündmus ebamäärase ajaväljendi *hiljem* sees, kuid kontekstist ei tule välja, kui kaua helistamine kestab. Võib aga oletada, et helistamine on suhteliselt lühikese kestusega sündmus ning ei saa kesta näiteks aastaid, kuid või nädalaid. Tõenäoliselt kestab see sekundeid või minuteid, kuid kummagi valikus ei saa täiesti kindel olla. Selliste juhtumite puhul märgiti kestuse määramise enesekindluse tasemeks tavaliselt *neutral* ehk keskmine kindlustase.

Selliste sündmuste puhul, mille kestus ei tulnud kontekstist välja ning mille kestust oli autoril keeruline oletada, märgiti kestuse määramise kindlustasemeks *low* ehk madal. Madalat kindlustaset esines aga märgenduses vaid 45 sündmuse puhul, mis tähendab, et enamasti leidus siiski vihjeid sündmuse kestusele kas sündmusega seotud ajaväljendis, ümbritsevas kontekstis või oli autoril ettekujutus sündmuse võimalikust kestusest.

4.6. Järeldused ja ettepanekud

Selles töös märgendati ajaväljenditega seotud sündmused eesti keelele kohandatud TimeML skeemi järgi, millele lisati juurde entiteetide ja sündmuste ajaliste kestuste märgendus ning võeti kasutusele varasema eesti keeles tehtud TimeML märgendusega võrreldes vähem ajaseosetüüpe. Märgendustöö käigus tulid esile mõned probleemkohad. Esiteks piiras BRAT-i konfiguratsioon kattuvate märgenduste tegemist, mille tulemusel võisid jääda mõned märgendused tegemata või poolikuks. Seda probleemi saaks tulevikus soovi korral

lahendada konfiguratsioonis kattuvate märgenduste lubamise teel, kuid kattuvad märgendused võivad samas automaatanalüüsi keerulisemaks muuta. BRAT-i konfiguratsioon piiras ka mitmesõnaliste sündmuste üheks sidumist juhul, kui nende vahele jäi mingi muu märgend (ajaväljend). Selle probleemi võiks tulevikus kindlasti konfiguratsiooni muutmise teel lahendada. Kuigi selliseid juhtumeid siinses märgenduses eriti palju ette ei tulnud, pole välistatud nende suurem esinemine tulevastes märgendustes.

Teiseks suureks probleemkohaks osutusid ebamäärased ajaväljendid, iseäranis ebamäärased viited olevikule (kõige sagedamini ajaväljend *praegu*), mis olid avatud erinevateks tõlgendusteks. Eesti keele Wordnetis[6] leidub näiteks sõnale *praegu* kaks erinevat tõlgendust:

1. *(AEG) kõnehetke ajal; seda hetke hõlmaval lühemal ajavahemikul*
2. *(AEG) kõnehetkega samal ajavahemikul (pikem periood olevikus)*

Kuna siinses märgenduses tuli ebamääraseid ajaväljendeid sageli ette, võiks tulevikus uurida, kas konteksti/tekstižanri põhjal oleks võimalik jaotada ebamääraseid olevikuviiteid kahte kategooriasse: lühemad vaatlushetked ja pikemad olevikuperioodid. Nii võiks olla võimalik vähendada mitmeti tõlgendatavuse probleemi ulatust. Teine ja lihtsam variant oleks jätta ebamäärased ajaväljendid vahele ning keskenduda vaid konkreetsetele ajaväljenditele.

Kolmandaks võiks tulevikus uurida ka seda, kas sündmuste märgendusviisi oleks võimalik veelgi lihtsustada. Hetkel on osade konstruktsioonide (*olema*-verbi konstruktsioonid ja konstruktsioonid sisuvaeste verbidega) märgendamine üsna keeruline või avatud erinevateks tõlgendusteks, millest tulenevalt võib selliste konstruktsioonide märgenduses tekkida suurema tõenäosusega vigu või ebakõlasid.

Ajaväljenditega seotud entiteetide märgendusse lisamise tulemustest saab öelda, et võrreldes sündmustega esineb neid tekstides vähe. Seega kirjeldavad ajaväljendid tekstides valdavalt

[6] https://teksaurus.keeleressursid.ee/ (vaadatud 31.05.2022)

ikkagi sündmuseid. Kui aga soovida ka edaspidi entiteetide märgendust TimeML raamistikus teostada, võib kaaluda viidete märgendamiseks eraldi märgendi (näiteks *reference*) kasutusele võtmist. Siinses märgenduses esines märgendatud entiteetide seas viiteid teoste pealkirjadele arvestataval hulgal, kuid tuleb nentida, et suur hulk neist esines ühes eesti uudiste rubriiki kuuluvas artiklis, mida võib pidada erandlikuks juhtumiks. Ülejäänud kaks viidet pealkirjadele ja üksik viide autorile esinesid kultuuriuudistes ja ajalooalastes teadusartiklites. Hetkel ei saa erandliku juhtumi ja viidete vähese esinemise tõttu mujal nende žanrilise jaotuse kohta veel järeldusi teha. Seda võiks uurida tulevikus suuremal korpusel, alles mille põhjal saaks otsustada ka viidete jaoks eraldi märgendi kasutusele võtmise vajalikkuse üle. Entiteetide ja ajaväljendite vahel esines kõige enam *simultaneous*-tüüpi ajaseost, mida kasutati juhtudel, kus ajaväljend kirjeldas entiteedi ajalist omadust (peamiselt kestust või vanust). Kuna entiteetide puhul näitas selline ajaseos võrreldes sündmustega mõnevõrra erinevaid omadusi, oleks variant võtta tulevikus kasutusele eraldi ajaseos (näiteks *is_property*), mis markeeriks seda, et ajaväljend kirjeldab entiteedi ajalist omadust (nt *[viieaastane] is_property [peatunnistaja]*, kus ajaväljend kirjeldab entiteedi vanust). Seosetüübid *includes* ja *is_included* olid entiteetide puhul lihtsamini tõlgendatavad (viitasid entiteedi tekkimisajale või (analoogselt sündmustele) lühemale vaatlushetkele entiteedi pikemast eksisteerimisest), mistõttu ei näe töö autor vajadust nende juures muudatuste tegemiseks.

Ajaväljendite ja sündmuste vahel esines palju ülekattumisseoseid (*includes, is_included* ja *simultaneous*), neist enim seosetüüpe *includes* ja *simultaneous*. See viitab sellele, et täpsemate ülekattumisseoste kasutuselevõtt üldise ülekattumisseose *overlap* asemel oli ilmselt õigustatud ning saab järeldada, et enamasti märgivad ajaväljendid sündmuste toimumisaega, mis tähendab ühtlasi, et enamasti võiks olla võimalik üsna usaldusväärselt sündmustele ka nende kestuseid määrata, kui lahendada ebamääraste ajaväljendite probleem. Kui sündmustele on märgenduses määratud kestused, saaks tulevikus kestuste ja ülekattumisseoste *includes* ja *simultaneous* abil märgendusi valideerida, sest need piiritlevad ajaliselt sündmuseid. Näiteks kui ajaväljend on määratud kuu täpsusega (*2022. aasta mai*)

ning selle ja sellega seotud sündmuse vahele on tõmmatud seos *includes*, siis peab sündmuse kestus olema kindlasti lühem kui üks kuu, sest muidu seos ei kehti. Loogiliste ebakõlade tuvastamine võimaldaks märgendusi korrastada. Korrastatud märgenduses saaks juba sündmuseid täpsemalt järjestada ning loogilise tuletamise teel välistada täiendavalt ebaloogilisi järjestusi. Näiteks kui üks ajaväljend on *2022. aasta jaanuaris* ja sellega seotud sündmus kestab päevi ning teine ajaväljend on *2022. aastal* ja sellega seotud sündmus kestab kuid, saab järeldada, et esimese sündmuse toimumine peab jääma kas teise sündmuse toimumisperioodi sisse või sellele eelnema ning välistatud on esimese sündmuse järgnemine teisele. Mõnevõrra vähem, kuid siiski märgatavalt esines märgenduses seosetüüpi *is_included*. See seosetüüp võib küll piiritleda sündmuse algus- või lõpuaega, kuid võib ka viidata lühemale vaatlushetkele sündmuse pikemast kestusest. Seetõttu ei saa võtta seisukohta, et tegemist on just sündmust ajaliselt piiritleva seosetüübiga ning žanrites, kus seda seosetüüpi rohkem esineb (siinses märgenduses ajalooalased teadusartiklid ja riigikogu stenogrammid) ei pruugi olla sündmuste kestuseid nii lihtne määrata.

Ajaväljendiga seotud sündmuste struktureerituse uurimisest ilmnes, et struktureeritumaid sündmuseid esineb märgenduses suhteliselt vähe ning ka eri tekstižanrite lõikes pole nende esinemises suuri erinevusi. Sündmuste klasside jaotuse uurimisel saadi ootuspärane tulemus, et kõigis žanrites valitsevad üldisemad sündmusklassid *occurrence* ja *state*. Teiste žanritega võrreldes esines ainult ajalooalastes teadusartiklites mõnevõrra vähem struktureeritumaid sündmuseid ja rohkem üldisematest klassidest sündmuseid. Seega ajalooalaseid teadusartikleid võib olla võrreldes teiste žanritega mõnevõrra lihtsam automaatselt analüüsida, kuid selle kinnitamiseks tuleks viia läbi masinõppe eksperimente. Ajalooalastes artiklites olid võrreldes teiste žanritega ka rohkem esindatud pikemate kestustega sündmused, mis näitab, et sajanditepikkuse kestustüübi kasutuselevõtt sellele žanrile mõeldes oli õigustatud. Huvitav tulemus oli see, et joonistusid välja erinevused eri klassidest sündmuste kestustes: seisundid kaldusid olema pikema kestusega ja raporteerivad või tahtlikule tegevusele viitavad sündmused kaldusid olema lühema kestusega. Modaalsust ehk teise sündmuse võimalikkusele või vajalikkusele antavaid hinnanguid väljendavad

sündmused olid kestuste pikkuselt vahepealsed. Töö autor ei oska seda trendi veel millegi laiemaga siduda, kuid see vääriks kindlasti edasist uurimist.

Kokkuvõte

Selle bakalaureusetöö eesmärk oli katsetada eesti keele ajaväljendite korpuse alamosal eksperimentaalset TimeML raamistikul põhinevat märgendusviisi ning uurida teostatud märgenduse põhjal, milliseid ajalisi omadusi kirjeldavad ajaväljendid nendega seotud keelendite juures, milline on märgendite struktuur ning milline on märgendite žanriline jaotus. Töö autor märgendas 85 tekstis 1233 ajaväljenditega seotud sündmust, 96 ajaväljenditega seotud entiteeti ja 1290 nendevahelist ajaseost. Märgendatud materjali saab tulevikus kasutada masinõppe otstarbel.

Töö olulisus seisneb selles, et võimekus tekstist ajaväljenditega seotud sündmuseid eraldada aitaks tekste automaatselt analüüsida. Tekstide automaatne analüüs eeldab käsitsi märgendatud korpuse põhjal masinõppe mudelite loomist. Õiget sündmuste märgendusviisi pole aga veel välja mõeldud ning semantilise märgenduse keerukuse tõttu tuleb eksperimentaalseid märgendusviise analüüsida.

Märgenduse analüüsi tulemustest ilmnes, et enamasti kirjeldavad ajaväljendid sündmuste juures nende toimumisaega, entiteetide juures aga enamasti entiteetide ajalisi omadusi (vanus, kestus). Sellest tulenevalt võiks tulevikus kaaluda entiteetide jaoks uue seosetüübi *is_property* kasutusele võtmist, mis markeeriks juhte, kus ajaväljend kirjeldab entiteedi ajalist omadust. Tulevikus võiks ka uurida entiteetide märgendust suuremal korpusel, et selgitada välja viiteid tähistavate entiteetide jaoks eraldi märgendi *reference* kasutusele võtmise mõttekus.

Märgenduses esineb ootuspäraselt enim üldisematesse sündmusklassidesse (*state* ja *occurrence*) kuuluvaid sündmuseid ning tunduvalt vähem struktureeritumatesse ehk argumenti nõudvatesse klassidesse kuuluvaid sündmuseid. Eri žanrite lõikes on struktureeritumate sündmuste osakaal üldiselt suhteliselt võrdne. Ainult ajalooalastes teadusartiklites on täheldatav teistega võrreldes mõnevõrra madalam struktureeritumate sündmuste esinemine. Seda, kas ajalooalaseid artikleid võiks olla tõepoolest natuke lihtsam käsitsi märgendada ja automaatselt analüüsida, tuleks aga kontrollida masinõppe eksperimentidega.

Sündmuste märgendusse toodi siinses töös sisse nende ajaliste kestuste märgendus. Joonistus välja, et eriti lühikese (silmapilgust sekunditeni) ja eriti pika (sajanditest igavikuni) kestusega sündmuseid esineb märgenduses vähe ning valdavalt küündivad ajaväljenditega seotud sündmuste kestused minutitest aastateni. Erandiks oli nädalatepikkune kestus, mille vähest esinemist autor seletada ei oska. Žanrite lõikes ilmnes, et ajakirjandusrubriikides ja riigikogu stenogrammides küündivad ajaväljenditega seotud sündmuste kestused enamasti aastateni ning väga pika kestusega sündmuseid esineb neis vähe. Ajalooalastes teadusartiklites esines aga rohkem pika ja eriti pika kestusega ajaväljenditega seotud sündmuseid. Huvitav avastus oli see, et osades sündmusklassides domineerisid lühema ja osades pikema kestusega sündmused. See trend vääriks edasist uurimist.

Lisaks dokumenteeriti märgendustöö jooksul ilmnenud probleemkohti, mille kaardistamine võiks aidata tulevikus sündmuste märgendusviisi parandada. Enim nõuaksid tulevikus tähelepanu ebamäärased ajaväljendid, sest neid tuli siinses märgenduses üsna sageli ette ning nendega seotud sündmustele on keeruline määrata usaldusväärselt kestust, samuti on keeruline määrata usaldusväärselt nende ja nendega seotud sündmuste vahelisi ajaseoseid. Tulevikus võiks ka uurida, kas sündmuste märgendusviisi oleks võimalik veel lihtsustada. *Olema*-verbi konstruktsioonide märgendusviis on hetkel üsna keeruline paljude märgendamise erijuhtude tõttu, mistõttu võib nende märgendamisel tekkida suurema tõenäosusega vigu. Samuti võib osutuda problemaatiliseks sisuvaest verbi sisaldavate

konstruktsioonide märgendus, sest nende märgendamisel on hetkel ruumi individuaalseteks tõlgendusteks, mis võib suurendada märgenduses ebakõlade tekkimise riski.

Kirjandus

Bittar jt = Bittar, André, Pascal Amsili, Pascal Denis, Laurence Danlos 2011. French TimeBank: an ISO-TimeML anntotated reference corpus. – Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 130–134.

Katz, Graham, Fabrizio Arosio 2001. The annotation of temporal information in natural language sentences. – Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing.

Laur jt = Laur, Sven, Siim Orasmaa, Dage Särg, Paul Timmo 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – In Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 7152–7160.

Monahan, Sean, Mary Brunson 2014. Qualities of eventiveness. – Proceedings of the Second Workshop on Events: Definition, Detection, Coreference and Representation, 59–67.

Orasmaa, Siim 2012. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. – Eesti Rakenduslingvistika Ühingu aastaraamat 8, 153–169.

Orasmaa, Siim 2014a. Towards an integration of syntactic and temporal annotations in Estonian. – Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Iceland: ELRA, 1259–1266.

Orasmaa, Siim 2014b. How availability of explicit temporal cues affects manual temporal relation annotation. – Human Language Technologies – The Baltic Perspective 268, 215–218.

Orasmaa, Siim 2014c. Ajaväljendite märgendamise juhised; https://github.com/soras/EstTimeMLCorpus/blob/master/docs-et/ajav2ljendite_m2rgendamine_06.pdf. Vaadatud 13.06.2022.

Orasmaa, Siim 2014d. Seoste märgendamine; https://github.com/soras/EstTimeMLCorpus/blob/master/docs-et/seoste_m2rgendamine_08.pdf. Vaadatud 13.06.2022.

Orasmaa, Siim 2014e. Sündmuste märgendamise juhised; https://github.com/soras/EstTimeMLCorpus/blob/master/docs-et/syndmuste_m2rgendamine_12.pdf. Vaadatud 13.06.2022.

Orasmaa, Siim 2016. Explorations of the problem of broad-coverage and general domain event analysis: The Estonian experience. – Dissertationes Mathematicae Universitatis Tartuensis 108. Tartu: Tartu Ülikooli kirjastus.

Pan jt = Pan, Feng, Rutu Mulkar-Mehta, Jerry R. Hobbs 2006. Extending TimeML with typical durations of events. – Proceedings of the Workshop on Annotating and Reasoning about Time and Events, 38–45.

Pustojevsky jt = Pustojevsky, James, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, Dragomir R. Radev 2003. TimeML: Robust specification of event and temporal expressions in text. – New Directions in Question Answering 3, 28–34.

Pustojevsky jt = Pustojevsky, James, Kiyong Lee, Harry Bunt, Branmir Boguraev, Nancy Ide 2008. Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events. – International Organization.

Reichenbach, Hans 1947. The tenses of verbs. – Time from concept to narrative construct: a reader. Toim. Jan Christoph Meister, Wilhelm Schernus. Berlin, 21–32.

Schilder, Frank, Christopher Habel 2001. From temporal expressions to temporal information: Semantic tagging of news messages. – Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing.

Sprugnoli, Rachele, Sara Tonelli 2017. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. – Natural Language Engineering 23(4), 485–506.

Vashishtha jt = Vashishtha, Siddharth, Benjamin Van Durme, Aaron Steven White 2019. Fine-grained temporal relation extraction. – arXiv preprint arXiv:1902.01390.

Vendler, Zeno 1957. Verbs and times. – The philosophical review 66(2), 143–160.

Extending temporal expression corpus annotation to temporal properties of events and entities in TimeML framework. Summary

The aim of this bachelor's thesis was to experiment with and examine TimeML (Pustojevsky *et al.* 2003) framework extended with event duration and entity annotation. To carry out the experiment, events, entities and temporal relations (TLINKs) between temporal expressions and events/entities were manually annotated in a subpart of EstTimexCorpora. Different genres were represented among annotated texts: news rubrics, historical articles and parliamentary transcripts.

Event, entity and temporal relation annotation was added to 85 texts containing 1282 temporal expressions. In total, 1233 events, 96 entities and 1290 temporal relations were annotated. Python library EstNLTK (Laur *et al.* 2020) was used to analyse the annotation.

The annotation analysis showed that temporal expressions are most often related to events. Compared to previous work on TimeML annotation in Estonian (Orasmaa 2014a; Orasmaa 2014b), fewer temporal relation types were used in order to lessen the prevalence of individual interpretations. However, precise overlap relations (*includes*, *is_included* and *simultaneous*) were still used, since those provide more information on the type of overlapping between temporal expression and an event/entity compared to general *overlap*. The results show that there is most often *includes* or *simultaneous* type of overlap between temporal expressions and events, which indicates that in events temporal expressions usually describe the time of event occurrence. Between temporal expressions and entities dominates *simultaneous* type of overlap, which in entities indicates that temporal expressions usually describe an entity's temporal properties (such as age or duration). Based on this, a new type of relation *is_property* is proposed for marking such cases in future. There were some instances where an entity represented a reference to a work of literature either through the surname of the author or title of the work. This raised an idea of using a separate tag for

references. However, since there were very few of such cases in this annotation, it should be studied on a larger corpus before deciding on the necessity of a separate reference tag.

It was found that the more general event classes *state* and *occurrence* dominated in this annotation and event classes that require a second event as an argument occurred less. There was not much difference in the distribution of more general and argument requiring events between different genres, only historical articles seemed to have a smaller percentage of events related to temporal expressions that belonged to an argument requiring class. This indicates that, compared to other genres it might be easier to manually annotate and automatically analyse historical articles; however machine learning experiments should be conducted before drawing this conclusion.

In this thesis, event durations were added to the event annotation. Annotation analysis showed that the duration of events related to temporal expressions is typically between minutes and years. Events with very short (instant, seconds) and very long (centuries, forever) durations were much rarer. An exeption was found in historical articles, where events related to temporal expressions lasted more often for centuries. An interesting trend was discovered in the distribution of duration types between event classes: reporting events and events indicating intentional activity tended to last for a shorter time period, intentional and unintentional states on the other hand tended to last for a longer time period. Modal verbs, which indicate an assessment on the possibility or necessity of another event were somewhere in between.

Some problem areas were documented during the annotation process. Most notable issue was related to temporal expressions that were vague references to past, present or future. The exact time period they cover cannot be determined, which means that it is not possible to confidently determine the durations of events related to them nor temporal relations between them and events. In future works it could be studied, whether it would be possible to assign such temporal expressions to different categories based on context or text genre. Another and

simpler possibility would be to omit such cases from future annotations. Another issue that calls for attention in the future is the complexity of event annotation guidelines, especially regarding the annotation of verb constructions involving the verb *be* that have many different rules for their tagging, which raises the possibility that mistakes might occur in annotation. Other difficult cases involve constructions involving a semantically weak verb. Whether to tag semantically weak verb or not is often open to individual interpretations, which can cause inconsistencies in annotation.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Annely-Maria Liivas,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Ajaväljendite korpusemärgenduse laiendamine sündmuste ja entiteetide ajalisteks omadusteks TimeML raamistikus“, mille juhendaja on Siim Orasmaa, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Annely-Maria Liivas

16.06.2022