TARTU ÜLIKOOL

Arvutiteaduse instituut

Informaatika õppekava

Jaan Teppo

Ajaväljendid Vikipeedia biograafilistes artiklites

Bakalaureusetöö (9 EAP)

Juhendaja: Siim Orasmaa

Tartu 2017

Ajaväljendid Vikipeedia biograafilistes artiklites

Lühikokkuvõte:

Käesoleva bakalaureusetöö eesmärk on uurida ajaväljendite kasutust Vikipeedia biograafilistes artiklites ja välja selgitada sünniaastate jagunemine, ajaväljendite liikide jagunemine, aastaarvuliste ajaväljendite jagunemine, ajaväljendite granulaarsus, ajaväljendite rikkalikkus ja sellele toetudes selgitada välja ajaväljendite sobivus Vikipeedia biograafiliste artiklite ajasemantiliseks võrdlemiseks. Töös antakse ülevaade autori poolt rakendatud Vikipeedia biograafiliste artiklite andmekaeve ja töötlemise protsessidest, ajaväljendite märgendamise ja statistika koostamise protsessidest, tuuakse välja ajaväljendite statistika Vikipeedia biograafilistes artiklites ja pakutakse välja meetod artiklite võrdlemiseks ajaväljendite alusel ja antakse meetodile esialgne hinnang.

Võtmesõnad:

Ajaväljend, Vikipeedia, biograafia, andmekaeve

CERCS: P175 Informaatika, süsteemiteooria

Temporal Expressions in Wikipedia Biographical Articles

Abstract:

The purpose of this Bachelor's thesis is to research the usage of temporal expressions in Wikipedia biographical articles and find out the distribution of birth years, distribution of temporal expression types, distribution of years in temporal expressions, granularity of temporal expressions, temporal richness and analyse the compability for comparison of the temporal expressions in Wikipedia biographical articles. This work gives an overview of data mining and processing of Wikipedia biographical articles, temporal expression tagging and statistics creation processes, statistics of temporal expressions in Wikipedia biographical articles is given and a method for comparing articles on the basis of temporal expressions and initial valuation for the method is proposed.

Keywords:

Temporal expression, Wikipedia, biography, data mining

CERCS: P175 Informatics, systems theory

Sisukord

Sissejuhatus

Aja ja ajapidamise kontseptsioone on inimesed kasutanud sündmuste paigutamiseks ja kirjeldamiseks juba aastatuhandeid. Samamoodi ulatub ka kalendrite ajalugu aastatuhandete tagusesse aega. Tänapäeva kalendri mõistele lähedasemad ehk fikseeritud kalendrid hakkasid selgemini kujunema esimesel aastatuhandel eKr, mis kulmineerus Juliuse kalendri loomisega 46 eKr, mille väikese modifitseeringuga varianti ehk Gregoriuse kalendrit kasutatakse kõige laialdasemalt juba aastasadu [1]. Ajaväljendid ja just kalendrilised ajaväljendid võimaldavad sündmusi ajateljele paigutada ja läbi selle sündmusi ja ka nendes osalenud isikuid omavahel seostada.

Internet ja tekstide digitaliseerimine on loonud olukorra, kus inimestel on ligipääs massilisele hulgale allikatele, artiklitele ja dokumentidele. Sellega kaasneb uudne probleem, kuidas suure hulga andmete või tekstide seest endale huvipakkuv üles leida. Seoste leidmiseks on loodud erinevaid otsingumeetodeid. Enamik meetodeid põhineb tekstide või võtmesõnade võrdlemisel ja kokkusobitamisel. Üheks vähem uurimist leidnud suunaks on aga aeg ja ajaväljendid. Viimased sisaldavad olulist infot, mis aitavad paigutada sündmused üheselt mõistetavale ajateljele. See omakorda loob võimaluse leida seoseid tekstide vahel, mis sõnastuselt võivad oluliselt erineda, aga on seotud ajaliselt. Ajaväljendite kasutamisest omavahel seotud tekstide leidmiseks võib kasu olla näiteks ajalootekstide, ajaloosündmuste ja biograafiliste tekstide uurimisel ning uudistekstide seostamisel.

Bakalaureusetöö esimene eesmärk on uurida ajaväljendite kasutust Vikipeedia biograafilistes artiklites ja välja selgitada sünniaastate, ajaväljendite liikide, aastaarvuliste ajaväljendite jagunemised, ajaväljendite granulaarsus, ajaväljendite rikkalikkus ja nende põhjal vastavalt välja selgitada, millised ajaväljendid ja millisel määral sobivad võrdlemiseks Vikipeedia biograafilistes artiklites.

Bakalaureusetöö teine eesmärk on vastavalt leiduvatele ajaväljenditele, kasutades ja vajadusel modifitseerides Omar Alonso jt [2] välja pakutud raamistikku, testida ja hinnata meetodit, leidmaks sarnaseid või ka seotud artikleid Vikipeedia biograafiliste artiklite hulgast. Nende raamistik pakub küllaltki lihtsa funktsiooni leidmaks kahe artikli vahelist kaugust ajaväljendite alusel. See on heaks alguspunktiks, et määrata artiklite vahelist seotust ajaväljendite alusel.

Töös on kasutatud eestikeelse Vikipeedia artikleid. Tööriistadeks on Python versioon 3.4.3 ja põhiliselt EstNLTK teek versioon 1.4.

Töö jaguneb neljaks suuremaks osaks. Esmalt on toodud taustainfo, kus kirjeldatakse lähemalt Vikipeediat, EstNLTK-d ja ajaväljendeid kasutavaid võrdlusmeetodeid. Teise osana on esitatud eeltöötlus, kus antakse ülevaade Vikipeedia artiklite töötlusest, biograafiate eraldamisest, ajaväljendite märgendamisest ja statistika loomisest. Seejärel tuuakse välja ja analüüsitakse ajaväljendite statistikat Vikipeedia biograafilistes artiklites. Viimases osas antakse ülevaade artiklite vahel seoste leidmiseks kasutatud meetodist, selle testimisest ja hinnangust meetodile seotud artiklite leidmiseks.

1. Taustainfo

Antud peatükis annab autor ülevaate Vikipeediast, mille artikleid kasutati töö jaoks loodud korpuse moodustamiseks, EstNLTK-st, mis oli põhiliseks tööriistaks korpuse ja andmete töötlemisel, ja võrdlemismeetodist, mida järgiti töö jaoks võrdlusmeetodi loomisel artiklite võrdlemiseks ja uurimiseks. Lisaks tuuakse välja paar alternatiivset artiklite võrdlusega seotud meetodit või uurimust.

1.1 Vikipeedia

Vikipeedia on 2001. aasta jaanuaris Jimmy Walesi ja Larry Sangeri poolt loodud internetis leiduv entsüklopeedia. Vikipeediat eristab traditsioonilistest entsüklopeediatest tema olemus, et igaüks võib luua ja muuta Vikipeedias olevaid artikleid. Seega võivad lugejad olla ka ise toimetajateks. Vikipeedia algusest, ajaloost ja loojatest on hea ülevaate andnud Andrew Lih [3]. Kui 2002. aasta alguseks sisaldas ingliskeelne Vikipeedia umbes 18 000 artiklit[1], siis 2017. aasta kevadeks on artiklite arv juba üle 5 350 000. Lisaks on 23. märtsi 2017. aasta seisuga Vikipeediaid 295 erineva keele jaoks. Eestikeelne Vikipeedia sisaldab sama kuupäeva seisuga natukene üle 155 000 artikli, mis asetab ta artiklite arvu poolest 44. kohale[2]. Vikipeedia olulisusest mitte ainult üldinfoallikana, vaid ka eesti keele ja kultuuri arendaja ja säilitajana rõhutab ka Mart Noorma algatatud projekt Miljon+ [4], mis näeb ette aastaks 2020 eestikeelse Vikipeedia artiklite arvu viimise miljonini.

Biograafia on kellegi elukäigu ja tegevuse kirjeldus. Vikipeedia esmane ülesanne on ikkagi olla entsüklopeedia, mis on võimalikult täpne ja tõetruu, aga samas on artiklid vabalt kõigi poolt lisatavad ja muudetavad. Seega on lihtne näha, kuidas antud eesmärk ja olemus eelkõige just biograafiate puhul võivad omavahel konflikte tekitada. Seda teemat on käsitlenud Pamela Graham [5]. Oma töös tõi ta välja, kuidas Vikipeedia ja internet on biograafiate olemust kujundamas ja kujundanud. Lisaks on ta oma töös käsitlenud Vikipeedia biograafiate kirjutamist ja sellega seosnevaid probleeme kirjutajate seisukohast.

Alternatiivse nurga alt uurisid Nir Ofek ja Lior Rokach [6], milliseid biograafiaid Vikipeedia aktsepteerib. Uurides mitte artiklite tekstide sisu, vaid keskendudes üheksale tegurile, nagu näiteks viidete arv või kas tegemist on esmakordse autori või anonüümse autoriga, pakkusid nad

[1] https://stats.wikimedia.org/EN/TablesWikipediaEN.htm (Vaadatud 23.03.2017)
[2] https://en.wikipedia.org/wiki/List_of_Wikipedias (Vaadatud 23.03.2017)

meetodi hindamaks, kas biograafiat aktsepteeritakse Vikipeedia poolt või mitte. Oma uurimuses tõid nad välja, kuidas olenemata artikli tekstist, keskendudes antud üheksale tegurile on väga kõrge täpsusega võimalik määrata, kas artikkel aktsepteeritakse biograafiana Vikipeedia poolt või mitte.

Vikipeedia oma suure artiklite ja andmete hulgaga pakub võimaluse uurimusteks andmekaevandamise seisukohalt. Vikipeedia artiklitel andmekaevandamisest on kirjutanud David Milne ja Ian Witten [7]. Oma töös tõid nad välja tarkvaralahenduse, mis pakub võimalust Vikipeedia artiklite andmekaevandamist korraldada. Sarnaselt on ka EstNLTK-le välja töötatud eraldi Vikipeedia liides, mis pakub samalaadset võimalust Vikipeedia andmete kaevandamiseks ja töötluseks. Sarnaselt antud tööle, kus on seotud andmekaevandamine ja Vikipeedia biograafiad, on uurimuse koostanud Ilia Reznik ja Vladimir Shatalov [8]. Oma töös kasutasid nad peamiselt biograafiates toodud sünni- ja surma-aastaid. Lisaks kasutasid nad kategooriaid, mille alla artiklid olid märgitud, et artikleid grupeerida ja sellest järeldusi teha. Järeldustena tõid nad näiteks välja, kuidas Vikipeedia biograafiate arvu kasv peegeldub umbes 1700. aasta paiku toimunud info leviku kasvus või kuidas Vikipeedia biograafiaid annab kasutada ajaloolise sotsioloogia õpinguteks.

1.2 EstNLTK

Üheks põhiliseks vahendiks antud uurimuses andmete töötlemisel oli EstNLTK. EstNLTK on Tartu Ülikooli juhtimisel loodud Pythoni teek, mis ühendab varasemalt keeletöötluseks loodud tarkvara ja muudab nad lihtsasti ligipääsetavaks ühise programmeerimisliidese all. EstNLTK-st on täpsema ülevaate kirjutanud teegi loojad Siim Orasmaa jt [9]. EstNLTK-st kasutati antud töös peamiselt Vikipeedia liidest ja ajaväljendite tuvastajat.

EstNLTK Vikipeedia liidese autoriks on Andres Matsin. Vikipeedia liides võimaldab kogu Vikipeedia artiklite hulga, mis on kokku pakitud ühte XML faili, lahti töödelda eraldi JSON failideks iga artikli kohta ja seejärel ka EstNLTK JSON failideks, mis võimaldab neid kasutada ka EstNLTK *Text*[3] klassiga. Lisaks märgendab ja grupeerib liides töötlusel igas artiklis üldandmed, välisviited, siseviited, artikli sektsioonid ja lõpuks ka artikli kogu teksti [10]. See

[3] *Text* klass on liides, mis koondab endas EstNLTK teksti töötluse baasoperatsioonid.

võimaldab uurijatel lihtsamini endale vajalikku infot eraldada ja vajadusel ka uurimiskorpuse moodustada.

EstNLTK ajaväljendite tuvastaja autoriks on Siim Orasmaa. Ajaväljendite tuvastaja on loodud EstNTLK-st varem iseseisva moodulina, millest on lähemalt kirjutanud Siim Orasmaa [11]. Ajaväljendite tuvastaja on sisse ehitatud EstNLTK *Text*'i klassi. Tuvastaja märgendab tekstis ajaväljendid ja moodustab neist loendi. Loend koosneb omakorda sõnastikest vastavalt iga märgendatud ajaväljendi kohta. Sõnastikud sisaldavad ajaväljendite semantikat. Kohustuslike atribuutidena on sõnastikest olemas ajaväljendi liik ja ajaväljendi kalendripõhine semantika, mis on toodud vastavalt atribuutides *type* ja *value*. Liigid jagunevad - kalendriliste toimumisaegade (ingl. k. *date*), nagu näiteks *22. aprill 1966* või *24. veebruar*; kellaajaliste toimumisaegade (ingl. k. *time*), nagu näiteks *pühapäeval kell 22:22* või *01. detsember 2016 kell 12:34;* ajavahemike või ajalise kestuse (ingl. k. *duration*), nagu näiteks *aastatel 1992-1999* või *seitse päeva*; ja ajaliste korduvuste (ingl. k. *set of times*), nagu näiteks *igal aastal* või *hommikuti*, vahel. Ajaväljendite kalendripõhine semantika sisaldab märgendatud ajaväljendit rahvusvahelisel kalendriaegade esitamise standardil ISO-8601. Esituse kuju jaguneb veel kuupõhiseks toimumisajaks, näiteks *22. aprill 1966* esitatakse kujul 1966-04-22, nädalapõhiseks toimumisajaks, näiteks kui käesolev aasta on 2017, siis ajaväljend *märtsi esimesel nädalal* esitatakse kujul 2017-W09, ja kestuseks, näiteks *seitseteist aastat* esitatakse kujul P17Y. Antud töö kontekstis on oluline ka märkida, et kuu-, nädala- ja kuupäevapõhised esitused võivad vastavalt ajaväljendi granulaarsusele paremalt pikeneda. Granulaarsus antud töö kontekstis näitab, millise täpsustusega on kalendriteljele paigutatav ajaväljend. Antud töös vaadatakse nelja erinevat granulaarsust. Alustades jämedaimast, milleks on aasta granulaarsus, näiteks ajaväljend *aasta 1969*, järgneb aasta ja kuu granulaarsus, näiteks *1996 jaanuar*, seejärel aasta ja nädala granulaarsus, näiteks *2016. a märtsi esimesel nädalal,* ja kõige kitsam aasta, kuu ja kuupäeva granulaarsus, näiteks *13. jaanuar 2015*. Näitena ajaväljendi esituse paremalt pikenemisest näiteks ajaväljend *aasta 1992* puhul on välja toodud lihtsalt aasta: 1992, aga *30. mai 1992* puhul on välja toodud aasta, kuu ja kuupäev: 1992-05-30. Kuna antud töö eesmärgiks on uurida ajaväljendeid seoses ajasemantilise võrdlemisega, siis on antud kaks kohustuslikku atribuuti peamiseks uurimisaluseks. Lisaks on samal põhjusel olulisel kohal ka atribuut *TemporalFunction*, mis määrab, kas tegemist on absoluutse või relatiivse ajaväljendiga. Absoluutsete ajaväljendite puhul on võimalik need paigutada ajateljele ilma, et oleks vaja lisainformatsiooni või teha lisaarvutusi, nagu näiteks kalendriliste

toimumisaegadega ajaväljendid *aasta 1936* või *24. veebruar 1926*. Relatiivsete ajaväljendite puhul ei piisa ajaväljendis toodud informatsioonist nende ajateljele paigutamiseks ja on tarvis lisada konteksti puudutavat informatsiooni, mis alati aga ei pruugi olla olemas või ei ole täpne. Näiteks kui ajaväljendiks on lihtsalt *30. mai* või *jõuludel*, siis on tegemist relatiivse ajaväljendiga ja kui ankrupunkti pole selgelt märgendatud, siis märgendatakse aastaks käesolev aasta. See aga võib artiklite vahel tekitada hulga ajaväljendeid, mis justkui langevad kokku, aga tegelikkuses ei tohiks. See loob eelduse võrdlemisel üldse relatiivsed ajaväljendid välja jätta [12].

1.3 Dokumentide võrdlus ajaväljendite alusel

Artiklite, tekstide, dokumentide võrdlusmeetodeid nende sarnasuse hindamiseks on küll loodud ja uuritud rohkesti, aga enamik neist põhineb otseselt tekstide või võtmesõnade võrdlemisel. Otseselt ajaväljendite põhjal loodud võrdlusmeetod seoste leidmiseks tekstide vahel on aga vähem uurimist leidnud suund.

Käesolevas töös on on aluseks võetud ja edasi arendatud Omar Alonso jt [2] töös välja pakutud otseselt ajaväljenditel põhinev võrdlemismeetod. Oma töös tõid nad välja dokumentidele omistatavate tunnustena ajaväljendite rikkalikkuse, mis näitab dokumendi ajaväljendite arvu määrana võrreldes kogu korpuse dokumentidega; dokumendi ajaväljendite spetsiifilisuse, mis näitab, millist tüüpi granulaarsusega ajaväljendeid leidub kõige enam; dokumendi ajalised piirid, määrates kõige varasema ja kõige hilisema dokumendis leiduva ajaväljendi. Granulaarsus oli neil jagatud viie tüübi vahel: aasta-, kuu-, nädala-, päeva- ja kellaaja täpsusega ajaväljendite vahel. Esiteks nägi nende meetod ette kõigi dokumendis leiduvate absoluutsete ajaväljendite laiendamist vastavalt kõige jämedama granulaarsusega ajaväljendile ja kokkulangevate grupeerimist. Näiteks tekst sisaldab ajaväljendeid, mille esitused on kujul 1936,1936,1945,1948-03,1949-04,1936-05-22,1948-03-12,1951-11-12, siis peale kõige jämedamale granulaarsusele viimist ja grupeerimist on esitused kujul 1936, 1937, 1938, …, 1945, 1946, 1947, 1948, 1949, 1950, 1951. Selle esituse põhjal pakkusid nad välja valemi kahe dokumendi vahelise kauguse hindamiseks, mis on toodud valemina (1).

$$dist(d, d') = \sum_{i=0}^{n-1} \left| \sum_{j=0}^{i} \left(t\text{-}freq_n(b_j) - t\text{-}freq_n(a_j) \right) \right| \qquad (1)$$

Valem võtab argumentidena kaks dokumenti, mis on märgitud kui *d* ja *d'*. Laiendades vastavalt antud dokumentide piire, et mõlemal dokumendil oleks sama alumine ja ülemine piir, ja lisades ajaväljendite gruppide hulka ühes dokumendis leiduvad ja teises puuduvad grupid ja vahepealsed grupid tühjade gruppidena, saame dokumendi *d,* mis koosneb ajaväljendite gruppidest $b_0,b_1,...,b_{n-1}$ ja dokumendi *d'*, mis koosneb ajaväljendite gruppidest $a_0,a_1,...,a_{n-1,}$ mis on oma pikkuselt ehk gruppide arvult võrdsed. Mõlemate dokumentide grupid on viidud samale, jämedaimale granulaarsusele. Näiteks on kaks dokumenti: üks, mis koosneb peale ajaväljendite grupeerimist ja samale granulaarsusele viimist gruppidest 1933,1937,1941,1950,1953,1954, ja teine, mis koosneb gruppidest 1932-02,1932-05,1932-08,1932-11,1933-01,1934-05, siis alandatakse esimese puhul alampiir 1932-ni ja lisatakse puuduvad ja teise puhul viiakse ajaväljendid aasta granulaarsusele ning ülemine piir 1954-ni ja lisatakse puuduvad vahepealsed grupid. Selle tulemusena saadakse esimese artikkel esitusena 1932,1933,1934,…,1952,1953,1954 ja teine artikkel 1932,1933,1934,…,1952,1953,1954. Valemis olev "$t\text{-}freq_n$" näitab vastavas grupis olevate ajaväljendite arvu. Ehk siis mõlema artikli puhul pikkus ehk *n* on 23. Seega valemi põhjal arvutatakse summa, kus *i* on nullist kuni 22-ni ja iga *i* korral leitakse vahesumma, mille absoluutväärtus liidetakse kogusummale. Vahesummaks on summa, kus *j* on nullist kuni *i*-ni ja *j* tähistab vastavalt gruppide asukohti jadas ehk siis antud näite põhjal, kui *i=0*, siis vahesummaks on summa, mis arvutatakse järgmiselt: aasta 1932 esinemiste arvust esimeses artiklis ehk 0-st lahutatakse aasta 1932 esinemiste arv teises artiklis ehk 4. Seega vahesummaks tuleb -4, mille absoluutväärtus ehk 4 liidetakse kogusummale. Seejärel *i=1* korral tuleb vahesummaks juba eelnevalt leitud -4-le juurde liita aasta 1933 esinemiste arv esimeses artiklis ehk 1, lahutada aasta 1933 esinemiste arv teises artiklis ehk 1. Seega vahesummaks tuleb taaskord -4, mille absoluutväärtus ehk 4 liidetakse kogusummale jne kuni *i=22*. Seega antud näite põhjal tuleb kahe dokumendi vahel valemi põhjal arvutatud kauguseks 73. Valem võrdleb põhimõtteliselt dokumentide histogramme ja leiab selle põhjal kauguse. Mida lähemal nullile, seda sarnasemad ajaväljendite põhjal antud dokumendid on. Meetod põhineb Sung-Hyuk Cha ja Sargur N. Srihari [13] poolt koos tõestustega esitatud valemil, mis näeb ette, et funktsioon leiab kauguse vastavalt sellele, kui mitmeid ümberpaigutamisi gruppide vahel peab kahe histogrammi vahel tegema.

Veel ühe sarnase alternatiivse hindamismeetodina tekstidevaheliste seoste leidmisel on tekstisarnasuse ja ajasemantika kombineerimine. Seda suunda on, kasutades muuseas ka Vikipeedia artikleid, uurinud ja meetodit ka hinnanud Daan Odijk jt [14]. Eesmärgiga leida

seosed ja sarnased tekstid just ajaloolist infot sisaldavatest artiklitest õnnestus neil oma töös näidata, kuidas tekstisarnasusel põhinevat võrdlemismeetodit on võimalik täpsemaks luua, lisades ajasemantilise võrdlemise.

2. Eeltöötlus

Antud peatükk annab ülevaate töös kasutatavate artiklite allalaadimisest, nende töötlusest ja statistika moodustamise protseduuridest. Kõik töö jaoks loodud skriptid on kättesaadavad lisast 1.

2.1. Eestikeelse Vikipeedia artiklid

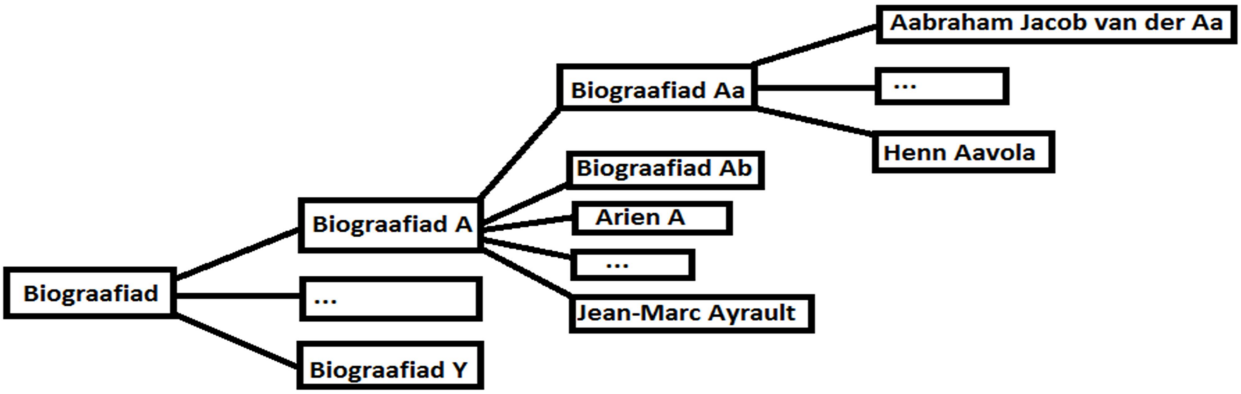
Esimesese sammu jaoks oli tarvilik kogu eestikeelsete Vikipeedia artiklite allalaadimine ja nende töötlemine korpuseks. Esmalt sai alla laetud kogu eestikeelse Vikipeedia sisutõmmis[4], kus on kõik artiklid koondatud ühte XML faili. EstNLTK-s on sisse ehitatud Vikipeedia liides [10], mis võimaldab antud sisutõmmist eraldi artikliteks lahti harutada ja need omakorda töödeldava korpuse kujule viia. Seejärel kasutati EstNLTK Vikipeedia liidese skripti *parser*, et sisutõmmise XML failist iga artikli ja ka suunamislehe kohta moodustada eraldi JSON fail. Tulemusena eraldati 200 001 JSON faili, mis lisaks artiklitele sisaldab kategooriate lehekülgi ja suunamislehekülgi. Seejärel kasutati EstNLTK Vikipeedia liidese skripti *convert*, et luua korpus, mis koosneb tekstifailidest vastavalt üks fail iga artikli, kategooria lehe või ka suunamislehe kohta.

2.2. Biograafiliste artiklite korpus

Kuna antud töö keskendub eestikeelse Vikipeedia biograafilistele artiklitele, siis järgmiseks sammuks oli kogu artiklite hulgast biograafiate ekstraheerimine.

Vikipeedia aritiklite hulgas on lehekülg Biograafiad[5]. Artikkel sisaldab sisemisi viiteid edasistele biograafiate loenditele mingi tähe täpsustusega või sisemisi viitasid biograafilistele artiklitele. Biograafiate loendid tähe täpsustusega sisaldavad omakorda selliste biograafiate loendeid, mille perekonnanimed algavad antud tähega või sisemisi viitasid biograafiate loenditele, kus on täpsustatud ka teinegi täht. Antud biograafiate loendi struktuuri on kujutatud joonisel 1.

[4] http://dumps.wikimedia.org/etwiki/latest/ (Vaadatud 26.11.2016)
[5] https://et.wikipedia.org/wiki/Biograafiad (Vaadatud 26.11.2016)



Joonis 1. Biograafiate loendite struktuur.

Joonisel on vasakpoolseimas kastis kujutatud lehekülge *Biograafiad* ja jooned kujutavad omakorda sisemisi viitasid, kas siis edasistele loenditele või biograafilistele artiklitele.

Biograafiate ekstraheerimisel on lähtutud biograafiate loendite struktuurist. Seega sai eelduseks võetud, et kui tegemist on biograafilise artikliga, siis esineb artikkel mõnes biograafiate loendis. Biograafiate eraldamiseks said loodud skript nimega *biograafiad*, milles on funktsioon, millele antakse ette esialgne biograafiate loendi fail ja mis eraldab failist regulaaravaldiste abil viitade hulga. Antud hulgast omakorda eraldatakse kõik pealkirjad ehk artiklite nimed ja käivitatakse funktsiooni rekursiivselt. Kui artikli puhul on tegemist biograafiate loendiga, siis kutsutakse funktsioon uuesti rekursiivselt välja, andes argumendiks loendi faili. Kui tegemist ei ole biograafiate loendiga, siis on tegemist biograafiaga ja kui antud nimega biograafia leidub kõigi failide hulgas, siis liigutatakse see eraldi biograafiate kausta. Töö käigus selgus, et biograafiate loendites on viitadena märgitud ka üksikud mittebiograafiad. Seega käis töö autor loendid käsitsi läbi ja loodavast korpusest sai käsitsi eemaldatud 38 mittebiograafiat. Mõni üksik mittebiograafia võis ka veel sisse jääda, aga see ei mõjuta oluliselt tulemusi. Kuna pealkirjade eraldamisel on osades pealkirjades sees unicode'i koodid, siis on veel loodud skript *korrastaUnicode*. Skript koosneb funktsioonist, millele teksti ette andes asendab unicode koodi vastava tähega, et oleks võimalik iga pealkirja puhul vastav failinimi leida. Kokku eraldas skript 200 001'st failist 33 453 faili ehk biograafilist artiklit, mis moodustavad antud töös uuritava korpuse.

2.3. Tekstide eraldamine

Järgmine samm oli biograafiate korpusesse kuuluvate artiklite failidest tekstiosade eraldamine. Sai loodud skript *tekstJaAjaväljendid*, mis sisaldab funktsiooni *koguTekstFailist*, millele antakse ette fail ja mis tagastab antud artikli kogu tekstiosa. Töö käigus selgus, et artiklite lõpus olevates sektsioonides, nagu näiteks pealkirjadega "Publikatsioonid", „Viited", „Kirjandus", „Välislingid", esineb mitmeid ajaväljendeid, mis ei pruugi biograafilise isikuga seotud olla, nagu näiteks Ramses II kohta käivas artiklis[6] sektsioonis Viited esineb ajaväljend *01. detsember 2016 kell 12:22*. Lisaks tuli aastaarvuliste ajaväljendite sajanditesse jagamisel välja, et esineb ebatõenäoliselt palju ajaväljendeid 23-ndast ja hilisematest sajanditest. Lähemal uurimisel selgus, et suurema enamuse puhul oli tegu valemärgendustega, mis leidsid aset antud lõpusektsioonides. Näiteks viidetes või kirjanduses välja toodud teoste puhul tuvastas märgendaja mitmetel juhtudel välja toodud lehekülgede numbrid ajaväljenditena. See omakorda võib mõjutada artiklite võrdlemise tulemusi. Seega on loodud veel teinegi funktsioon *filtreeritudTekstFailist*. Antud funktsioon jätab kogutekstist välja ette antud lõpusektsioonid.

2.4 Sünni- ja surma-aastate määramine

Vikipeedias on biograafilistel artiklitel sünni- ja surma-aastad, kui need on määratud, välja toodud kategooriatena iga artikli juures. Sellest tulenevalt tuli lisaks teksti eraldamisele esialgsetest korpusefailidest eraldada ka sünni- ja surma-aastad, kui need olid määratud. Seega sai skripti *tekstJaAjaväljendid* loodud funktsioon *sünniJaSurmaaasta*, mis kasutades regulaaravaldisi eraldab ette antud artikli failist kõik kategooriad. Seejärel vaadatakse, kas kategooriate hulgas on sünni- või surma-aasta kategooria ja kui on, siis märgendatakse see ajaväljendite märgendajaga ja eraldatakse sellest aasta ja määratakse vastavalt siis sünni- või surma-aastaks.

2.5. Ajaväljendite märgendamine

Tekstide eraldamisele ja sünni- ja surma-aastate määramisele järgnes ajaväljendite märgendamine tekstides. See tegevus toimus samuti skriptis *tekstJaAjaväljendid*. Ajaväljendite märgendamiseks kasutati EstNLTK klassi Text ja sellest funktsiooni *timexes***.** Kuna ajaväljendite märgendaja on loodud ja testitud uudistekstide peal, siis võis eeldada, et antud tekstide

[6] https://et.wikipedia.org/wiki/Ramses_II (Vaadatud 18.02.2017)

märgendamisel võib esineda probleeme. Esialgsel märgendamisel selguski, et märgendaja ei märgenda kahe- ja kolmekohalisi aastaarve üldse. Lisaks ei märgendatud neljakohalisi aastaarve alates 1000-1899, mis esinevad üksikute numbritena, ilma kuupäevata või aasta tähiseta. Uudistekstides pole antud arvud tavaliselt mõeldud aastanumbritena, aga biograafilistes artiklites küll. Näiteks Galileo Galilei kohta käivas biograafilises artiklis[7] ainuüksi sektsioonis „Tuntumad saavutused“ on üheksa sellist ajaväljendit.

Seega viis EstNLTK ajaväljendite märgendaja looja Siim Orasmaa sisse uuenduse, mis lubas määrata ajaväljendite märgendajale märgendamiseks kasutatava reeglite faili. Käesoleva töö autor otsustas siduda ajaväljendite märgendamise sünniaastatega ja moodustas vastavalt sünniaastatele neli gruppi ning lõi esialgsele reeglite failile lisaks juurde veel kolm reeglite faili:

1. Esimese grupi moodustasid artiklid sünniaastatega vahemikust 99 e.m.a. kuni 99 m.a.j. ja lisaks ka artiklid surma-aastatega 99–1 e.m.a. Selle grupi jaoks sai loodud kõige laiemate reeglitega fail *reeglidesimene*, mis oli laiendatud nii kahe-, kolmekohaliste aastaarvude ja neljakohaliste aastaarvude jaoks vahemikust 1000–1899 ja kuupäevade puhul ka ühekohaliste aastaarvude tarvis.
2. Teise grupi moodustasid artiklid sünniaastatega 999–100 e.m.a. ja 100–999 m.a.j. ja lisaks ka artiklid surma-aastatega 999–100 e.m.a. Selle grupi jaoks said loodud reeglite fail *reeglidteine*, mis oli laiendatud nii kolmekohaliste aastaarvude kui ka neljakohaliste aastaarvude vahemikust 1000–1899 jaoks.
3. Kolmanda grupi moodustasid artiklid sünniaastatega ... –1000 e.m.a. ja 1000–1899 m.a.j. Selle grupi jaoks said loodud reeglite fail *reeglidkolmas*, mis oli laiendatud ka neljakohalistele aastaarvudele vahemikust 1000–1899.
4. Neljanda grupi moodustasid artiklid sünniaastatega 1900– ... m.a.j. ja lisaks veel artiklid, millel ei olnud sünniaastat määratud. Need artiklid said märgendatud esialgse, muutmata reeglite failiga.

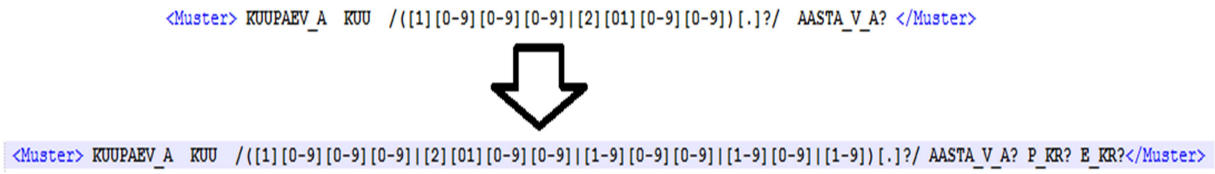
Kuna uued loodud reeglifailid olid kasvavas suunas laiendatud, aga aastad e.m.a. liiguvad ajateljel kahanevas järjekorras, siis oli nende puhul arvestatud ka surma-aastaid.

[7] https://et.wikipedia.org/wiki/Galileo_Galilei (Vaadatud 17.02.2017)

Näitena märgendamisest tooks näiteks Karl Suure kohta käiva artikli[8], milles on sünniaastaks märgendatud 740. aastate paiku, märgendades artiklit esialgse, muutmata reeglifailiga tuvastatakse 27 ajaväljendit, milles leidus vaid kahte korrektselt tuvastatud aastaarvu sisaldavat ajaväljendit, mis mõlemad olid sektsioonist "Kirjandus" vastavalt 1906 ja 1999. Märgendades sama artiklit reeglite failiga *reeglidteine*, tuvastatakse aga 63 ajaväljendit, milles oli 41 korrekselt tuvastatud aastaarvu sisaldavat ajaväljendit.

Ajaväljendite märgendamiseks on loodud funktsioon *ajaväljendid*, millele andes ette teksti ja sünni- ja surma-aastad märgendab ta vastavalt ajaväljendid ja tagastab need.

Kuna määramiseks koostatud reeglid koosnevad regulaaravaldistest, siis reeglites sisse viidud muudatused seisnesid enamasti aastaarve kirjeldavate regulaaravaldiste muutmisel selles, et tuvastataks arve väiksemaid kui 1900 ka ajaväljenditena. Reeglifailis oli juba olemas reegel, mis eraldi tuvastas sünniaasta vastavalt fraasile sündinud või sünd, ja kuna tegemist oli biograafiatega, siis sai lisatud ka samasugune reegel surma-aasta jaoks, mis tuvastas surma-aastat vastavalt fraasile surnud või surn. Uudistekstide puhul ei tähendaks antud fraasid koos arvuga enamasti kellegi spetsiifilist surma-aastat, aga biograafiate puhul küll. Näitena on toodud joonisel 2. reegel, mis tuvastab kuupäevalisi ajaväljendeid, näiteks nagu *30. detsember 2016* või *11. aprill 1993. aasta*. Joonisel on ülemine reegel esialgsest, muutmata reeglitefailist ja teine, kahe-, kolme- ja neljakohalistele arvudele laiendatud reeglite failist[9].



Joonis 2. Reeglite laienduste näide.

Joonisel on näha, et esimeseks tuvastab reegel kuupäeva osa, mille muster on juba eelnevalt regulaaravaldisega kirjeldatud. Seejärel kuu osa, mille muster on samuti juba eelnevalt kirjeldatud. Sellele järgneb aasta numbriline osa, millesse on sisse viidud muudatus, laiendades

[8] https://et.wikipedia.org/wiki/Karl_Suur (Vaadatud 24.11.2016)
[9] Ajaväljendite reeglite täpne kirjeldus - https://github.com/soras/Ajavt/blob/master/doc/writingRules.txt (vaadatud 24.04.2017)

seda regulaaravaldise osa ka kolme-, kahe- ja ka ühekohaliste aastaarvude jaoks. Sellele võib omakorda järgneda aasta tähis ja juurde on lisatud, et võib järgneda ka tähis, mis märgib, kas tegemist on kuupäevaga eKr või pKr.

Kuna ajaväljendite märgendamine on aeganõudev protsess, siis on skriptis kaks funktsiooni *koguTekstBiograafiatest* ja *filtreeritudTekstBiograafiatest*, mis eraldavad kogu etteantud biograafiate failide kaustast artiklitest nime, teksti (kogu või filtreeritud), ajaväljendid ja sünni- ja surma-aastad. Antud nelikud kirjutatakse funktsiooniga *andmedfailidesse*, mis kasutab andmete salvestamiseks Pythoni moodulit *pickle*, uude kausta. Uues kaustas on iga artikli jaoks eraldi fail, mis sisaldab artikli vastavaid andmeid.

2.6. Ajaväljendite statistika moodustamine

Ajaväljendite eraldamisele järgnes ajaväljendite põhjal statistika moodustamine. Selleks sai loodud skript *sagedusJaMuuStatistika*. Kuna ajaväljendid said iga artikli jaoks eraldi andmefaili kirjutatud, siis piisab antud skriptis lihtsalt Pythoni moodulit *pickle* kasutades andmed failist loendisse lugemisest. Selleks on loodud funktsioon *andmeteLoend*, mis etteantud kaustast kõik failid loendisse lisab.

Iga artikli kohta on loodud ajaväljendite loend, mis koosneb sõnastikest. Antud sõnastikud sisaldavad igaüks infot ühe tekstis esinenud ajaväljendi kohta. Näiteks näide ajaväljendi *22. juuni 1934* ajasemantilist infot sisaldavast sõnastikust on toodud joonisel 3.

```
{'end': 14, 'start': 0, 'value': '1934-06-22', 'id': 0, 'temporal_function': False,
'text': '22. juuni 1934', 'type': 'DATE', 'tid': 't1'}
```

Joonis 3. Näide ühe ajaväljendi semantilist infot sisaldavast sõnastikust.

Jooniselt on näha, et sõnastikus on toodud ajaväljendi lõpp- (ing. k. *end*) ja alguspunkt (ing. k. *start*) tekstis, ajaväljendi väärtus (ing. k. *value*) ehk kalendriline esituskuju, ajaväljendi identifikaator (ing. k. *id*), ajaväljendi absoluutsust (ing. k. *temporal function*) või relatiivsust näitav info, ajaväljendi tekst (ing. k. *text*), ajaväljendi tüüp (ing. k. *type*) ja veel ajaväljendi unikaalne identifikaator (ing. k. *tid*).

Ajaväljendite loendi töötlemiseks on loodud funktsioon *ajainfo,* mis võtab ette ühe artikli teksti, ajaväljendite loendi ja sünniaastate grupi, kuhu antud artikkel kuulus. Et ajaväljendite loendeid ei peaks mitmeid kordi läbima, on enamik ajaväljendite sõnastike seest uurimist puudutava info

eraldamine ja grupeerimine kõik korraga ühe funktsiooni sees ära tehtud. Funktsioon loeb näiteks kokku ühes artiklis esinevate kalendrilise toimumisajaga absoluutsete ja relatiivsete arvu, absoluutsete puhul ka granulaarsuste esinemised, koondab erinevatesse loenditesse ajaväljendite tüübid, ajaväljendite tekstid, ajaväljendite väärtused, absoluutsete kalendriliste toimumisaegadega ajaväljendite aastate väärtused jne. Lisaks on funktsioon *ajasagedus*, mis võtab ette kogu korpuse artiklite ajaväljendeid puudutava statistika, mis sai määratud eelnevalt kirjeldatud funktsiooniga. Läbides iga artikli kohta üksikud loodud loendid, moodustatakse funktsioonis kogu korpuse peale vastavalt statistika tüübile kas sagedusloendid või kokku liidetud loendid. Eraldi on käsitletud veel sünni- ja surma-aastaid puudutavat statistikat, mille jaoks on loodud funktsioon *sünniJaSurmaAastadSagedus.* Kuna sünni- ja surma-aastad olid juba metaandmetena artikliga kaasa pandud, antakse funktsioonile kohe ette andmete loend ja moodustatakse koguandmete põhjal vastavalt kas sagedusloendid või muu sünni- või surma-aastaid puudutav spetsiifilisem statistika. Lisaks on funktsioonid statistika failidesse kirjutamiseks. Statistika on kirjutatud topelt nii TXT-i failidesse kui ka CSV failidesse. Lisaks sisaldab skript erinevaid funktsioone pisemate spetsiifilist statistikat puudutavate detailide arvutamiseks.

2.7 Statistika visualiseerimine

Statistika visualiseerimiseks antud töö jaoks on kasutatud samuti Pythonit. Selleks on loodud skript *graafikud,* mis sisaldab funktsiooni nagu näiteks *sünnihistogramm*, *koguaastadhistogramm*, *surmaaastadgraafik* jt. Igas funktsioonis on loetud sisse vastavat statistikat sisaldav CSV fail ja loodud ning sätestatud graafik vastavalt vajadusele. Andmete lugemiseks on kasutatud Pythoni teeki Pandas ja visualiseerimiseks graafikute kujule on kasutatud Pythoni teeki Seaborn.

3. Ajaväljendid Vikipeedia biograafilistes artiklites

Antud peatükis tuuakse ülevaade ajaväljendite kasutusest ja jagunemisest Vikipeedia biograafilistes artiklites. Autor toob peatükis välja enne analüüsi tehtud eeldused ajaväljendite kasutuse kohta ning analüüsi põhjal toodud järeldused. Vastavalt sissejuhatuses sõnastatud esimesele eesmärgile leitakse peatükis vastused järgmistele uurimisküsimustele:

1. Kuidas jagunevad Vikipeedia biograafilised artiklid ajaliselt sünniaastate põhjal?
2. Kuidas jagunevad ajaväljendid Vikipeedia biograafilistes artiklites liigiti?
3. Millise granulaarsusega ajaväljendeid leidub kõige enam Vikipeedia biograafilistes artiklites?
4. Kuidas jagunevad ja mida näitavad aastaarvulised ajaväljendid Vikipeedia biograafilistes artiklites?
5. Kas ja kuidas mängib ajaväljendite rikkalikkuses rolli ajaperiood, kuhu kuulus Vikipeedia biograafilises artiklis kirjeldatud isik?

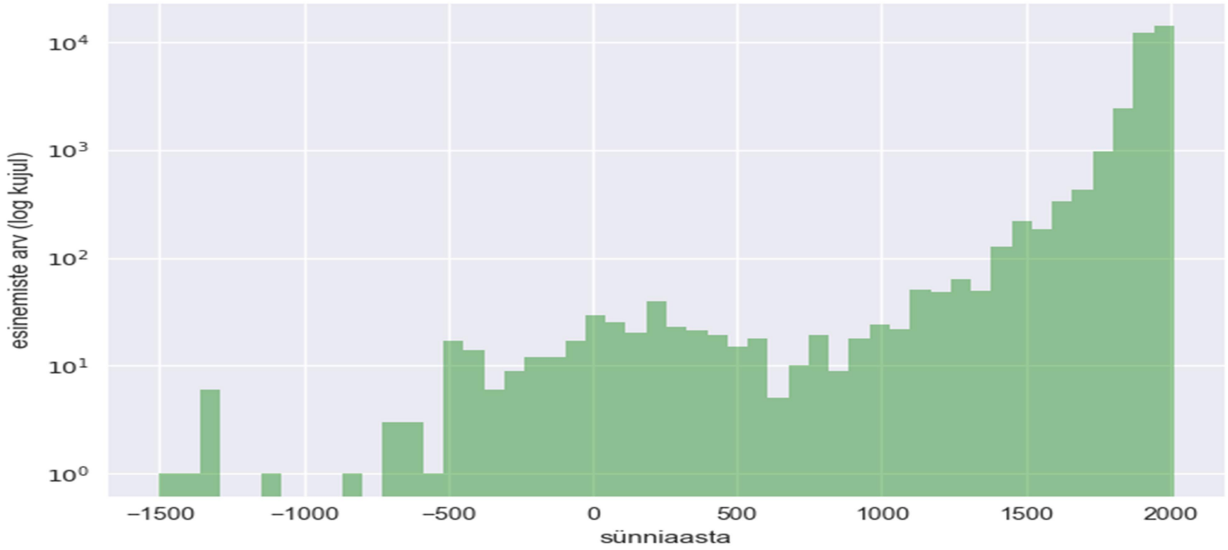
Antud peatükis toodud andmed ja statistika on kõik 24. novembri 2016 seisuga, sest siis viidi viimati töö jaoks algusest lõpuni läbi eeltöötlusprotsess. Uurimus on koostatud 33 453 biograafilisest artiklist koosneva korpuse peal. Statistika käib artiklite kogutekstidega korpuse kohta, kui just ei ole eraldi mainitud, et mingi osa põhineb välja jäetud lõpusektsioonidega tekstide korpusel. Antud statistika failid on ka toodud kõik lisas 1.

3.1 Sünniaastad Vikipeedia biograafilistes artiklites

Esmane uurimispunkt töös oli välja selgitada, kuidas artiklid jagunevad sünniaastate põhjal. Arvesse võttes infohulga suurenemist läbi ajastute ja ka Vikipeedia olemust, kus artikleid luuakse ja uuendatakse vabalt kasutajate poolt, ning seda, et lisaks infohulga suurenemisele võib artiklite hulgas rolli mängida ka rahvaarvu kasv (rahvaarv 19. ja 20. sajandil mitmekordistus), siis võib eeldada artiklite laialdasemat jagunemist lähisajanditesse, kus enamus on kahekümnendast sajandist.

Sünniaasta oli korpuses määramata 1960 artiklil. See moodustab kogu korpusest ligikaudu 5,9%. See annab hea eelduse sünniaastaid kasutades leida artiklite ajapiirid, mida kasutada võrdluses. Varaseim artikkel, millel oli määratud sünniaasta, oli Egiptuse naisvaarao Hatšepsut'i kohta käiv

artikkel[10] sünniaastaga 16. saj eKr. Hiliseim artikkel, millel oli sünniaasta määratud, oli Rootsi printsess Estelle'i kohta käiv artikkel[11] sünniaastaga 2012. Määratud sünniaastatest moodustati histogramm, mis on kujutatud joonisel 4.



Joonis 4. Sünniaastate histogramm, sünniaastate arv kujutatud logaritmilisel kujul.

Joonisel on kujutatud sünniaastate esinemise arv ehk artiklite arv y-teljel ja aastate jagunemine x-teljel. Aastaarvud eKr tähistega on märgitud negatiivsetena. Esialgsel histogrammi loomisel selgus aastast 14. sajandi ja varasemate väga vähene esindatus võrreldes näiteks 20. sajandiga. Seega on loetavuse tõstmiseks viidud y-telg logaritmilisele kujule [15]. Kui sajandid 6. eKr ja varasemad on esindatud vaid mõne üksiku artikliga, siis sajandid 5. eKr kuni 11. pKr on igaüks juba kaetud mõnekümne artikliga. Mõningane langus on küll märgata sajanditest 5–9 sajanditeni 1–5, aga see langeb kokku Rooma riigi langusega ja sellele järgnenud vähem allikatega kaetud ajaga ajaloos. Histogrammil on näha hüppelist suurenemist biograafiate arvus alates 15. sajandist, mis tõuseb kuni 20. sajandini. Sünniaasta poolest 20. sajandisse kuuluvate artiklite arv on 22 221, mis moodustab kogu korpusest ligikaudu 66%. Lisaks kuulub sünniaasta poolest 19. sajandisse 6394 artiklit, mis moodustab kogu korpusest ligikaudu 19%. Artiklid sünniaastaga sajanditest 15–20 moodustavad kogu korpusest lausa 92,1%. Kui sinna juurde lisada veel artiklid, millel polnud sünniaastat määratud, jätab see 14. ja varasemate sajandite hulgaks ainult

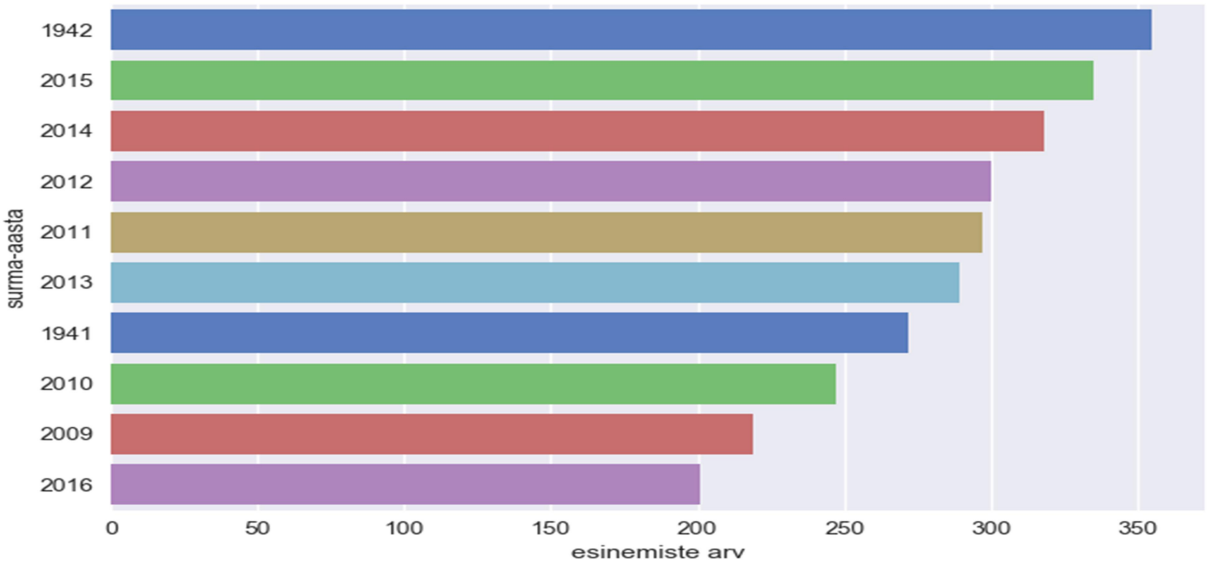
[10] https://et.wikipedia.org/wiki/Hat%C5%A1epsut (Vaadatud 12.03.2017)
[11] https://et.wikipedia.org/wiki/Estelle (Vaadatud 12.03.2017)

2%. Antud jaotus annab ettekujutuse kasutatava meetodi rakendamise võimalikest piiridest. Kuna sünniaastate abil piiritledes on lähisajandites ka enam võimalikke võrdlusobjekte, siis on tõenäolisem leida seoseid lähisajanditesse kuuluvate isikute vahel.

3.2 Surma-aastad Vikipeedia biograafilistes artiklites

Sünniaastate kõrval aitab biograafiliste artiklite ajapiire fikseerida ka surma-aasta, kui see on määratud. Võttes arvesse sünniaastaid, kus jaotus on ülekaalukalt lähisajanditesse koondunud ja moodustab väga suure enamuse 20. sajandist, siis võib ka eeldada, et päris arvestatav hulk isikutest, kelle kohta on Vikipeedias artikkel, on veel elus. Lisaks arvestades Vikipeedia olemust, kus artikleid lisatakse ja uuendatakse jooksvalt, võib eeldada, et enim leidub surma-aastatena just lähiaastaid.

Surma-aasta oli määramata 16 506 artiklil, mis moodustab uuritavast korpusest ligikaudu 49,3%. Artikleid, millel oli nii surma- kui ka sünniaasta määramata, oli 1258, mis moodustab kogu korpusest ligikaudu 3,8%. Artikleid, mille surma-aasta oli määramata, aga sünniaasta oli viimase 100 aasta sees, oli 15 000. See annab põhjust arvata, et päris suur hulk isikuid, kelle kohta on Vikipeedias biograafiline artikkel, on veel elus. Seda asjaolu tuleks arvesse võtta määratledes biograafiliste artiklite ajapiire, kui hakata omavahelisi seoseid leidma. Lisaks on surma-aastate sagedusloendis esimese 100 sagedamana esineva surma-aasta sees kõik aastad 20. või 21. sajandist. Joonisel 5 toodud 10 kõige sagedasemalt esinevat surma-aastat.



Joonis 5. 10 sagedamini esinevat surma-aastat.

Joonisel on y-teljel kujutatud surma-aastad ja x-teljel artiklite arv, milles esines antud surma-aasta. Vikipeedia ja selle artiklite olemusest annab aimu asjaolu, et 10 sagedasema aasta sees on kõik aastad vahemikust 2009–2016. Lisaks on huvitav asjaolu, et esikohal on aasta 1942 ja seitsmendal kohal on aasta 1941, mis on teise maailmasõja aastad. See omakorda läheb kokku Rezniki ja Shatalovi Vikipeedia biograafiateteemalise uurimusega [8], kus ka ingliskeelse Vikipeedia biograafiate surma-aastates peegeldus maailmasõdade mõju.

3.3 Ajaväljendite liigid Vikipeedia biograafilistes artiklites

Kuna EstNLTK ajaväljendite märgendaja jagab ajaväljendid nelja tüübi vahel, siis oli töö üheks uurimisküsimuseks, kuidas ajaväljendid Vikipeedia biograafilistes artiklites antud liikide vahel jagunevad. Eelnevalt oli töö autoril kasutada Siim Orasmaa koostatud ajaväljendite liikide jaotus aja-, ilu- ja teaduskirjandustekstides [16], mis on toodus tabelis 1.

Tabel 1. Ajaväljendite liigid aja,-ilu ja teaduskirjanduse tekstides automaatse analüüsi põhjal [16].

Ajaväljendite liigid aja,-ilu ja teaduskirjanduse tekstides		
Ajakirjandus	Ilukirjandus	Teaduskirjandus
DATE Absoluutseid: 9.9% Relatiivseid: 63.3%	DATE Absoluutseid: 3.7% Relatiivseid: 63.2%	DATE Absoluutseid: 30.1% Relatiivseid: 52%
TIME: 7.1%	TIME: 16.6%	TIME: 1.1%
DURATION: 16.8%	DURATION: 13%	DURATION: 15%
SET: 2.2%	SET: 3.4%	SET 1.5%

Tabelis on välja toodud aja-, ilu- ja teaduskirjanduse tekstides esinenud ajaväljendite jaotus vastavalt osakaalule kogu ajaväljendite arvust. Eraldi on kalendriliste toimumisaegadega ajaväljendite (ing. k. *date*) puhul toodud absoluutsete ja relatiivsete ajaväljendite osakaal. Tabelist on näha, et tekstiliikide vahelised erinevused ajaväljendite kasutuses tulevad kõige selgemalt esile, vaadates kalendrilise toimumisega (ing. k. *date*) ajaväljendeid ja täpsemalt just eristades nende vahel absoluutseid ja relatiivseid ajaväljendeid. Lisaks on märgata ka erinevused kellaajalise toimumisajaga (ing. k. *time*) ajaväljendite kasutuses. Eeldada võib, et Vikipeedia biograafiliste tekstide ajaväljendite jagunemine sarnaneb enam teaduskirjandustekstidele.

Kogu uuritud korpus sisaldas kokku 692 364 ajaväljendit. Ajaväljendite jagunemine eristatavate liikide vahel on toodud tabelis 2.

Tabel 2. Ajaväljendite liigid Vikipeedia biograafilistes artiklites.

Vikipeedia biograafilised artiklid	
Liik	Osakaal
DATE Absoluutseid Relatiivseid	 77.4% 8.2%
TIME	0.5%
DURATION	13.7%
SET	0.2%

Tabelis 2 on sarnaselt tabelile 1 toodud Vikipeedia biograafiliste artiklite ajaväljendite liikide osakaal kogu ajaväljendite arvust. Koheselt on tabelist märgata suurt erinevust kalendrilise toimumisajaga (ing. k. *date*) ajaväljendite osakaalus. Tabelist on näha, et enamik antud ajaväljenditest on absoluutsed. See läheb igati kokku asjaoluga, et tegemist on biograafiliste tekstidega. Lisaks tuleb antud asjaolu kasuks artiklitevaheliste seoste leidmisel. Relatiivsete ajaväljendite otsese võrdluse juures on seoste täpne määramine oluliselt raskendatud, sest see, kuidas relatiivseid ajaväljendeid Vikipeedia tekstides automaatselt normaliseerida, vajab veel eraldiseisvat uurimist. Ülejäänud tabeli osas ongi märgata kõige enam sarnasust teaduskirjandustekstidele. Ajalise kestuse (ing. k. *duration*) liiki ajaväljendeid on kõigis tekstiliikides enam-vähem sama palju. Lisaks on märgata teaduskirjandustekstidega võrreldes kellaajalise toimumise (ing. k. *time*) ja ajalise korduvuse (ing. k. *set*) ajaväljendite veelgi väiksem osakaal, mis on samuti seletatav asjaoluga, et tegemist on Vikipeedia biograafiliste artiklitega.

3.4 Ajaväljendite granulaarsus Vikipeedia biograafilistes artiklites

Eelmises alapeatükis selgus, et üle nelja viiendiku ajaväljenditest on aasta, kuu, nädala või päeva täpsustusega kalendrilised toimumisajad. Selle põhjal tekkis ka huvi, kuidas antud ajaväljendid granulaarsuse poolest jagunevad. Kuna antud ajaväljendid jagunesid veel absoluutsete ja relatiivsete vahel ja kuna relatiivsete puhul on tarvilik veel ajateljele paigutamiseks

lisainformatsioon, siis antud peatükis uuris autor vaid absoluutseid ajaväljendeid. Siis saab ka eelduseks võtta, et kitsama granulaarsusega ajaväljend, näiteks päeva täpsustusega, sisaldab ka jämedama granulaarsusega infot, näiteks kuu ja aasta täpsustust, aga nädala täpsustusega ajaväljend ei laiene kuu täpsustusele, vaid ainult aasta täpsustusele.

Kuna tegemist on biograafiliste tekstidega ja Vikipeedia artiklitega, siis eeldus on, et suuremas jaos on ajaväljendid just aasta granulaarsusega. Granulaarsuse poolest jagunemiste osakaal on toodud tabelis 3. Kokku määrati granulaarsus 536 197 ajaväljendil. Granulaarsus määrati vaid kalendrilise toimumisaegade absoluutsetel ajaväljenditel.

Tabel 3. Ajaväljendite granulaarsus Vikipeedia biograafilistes artiklites.

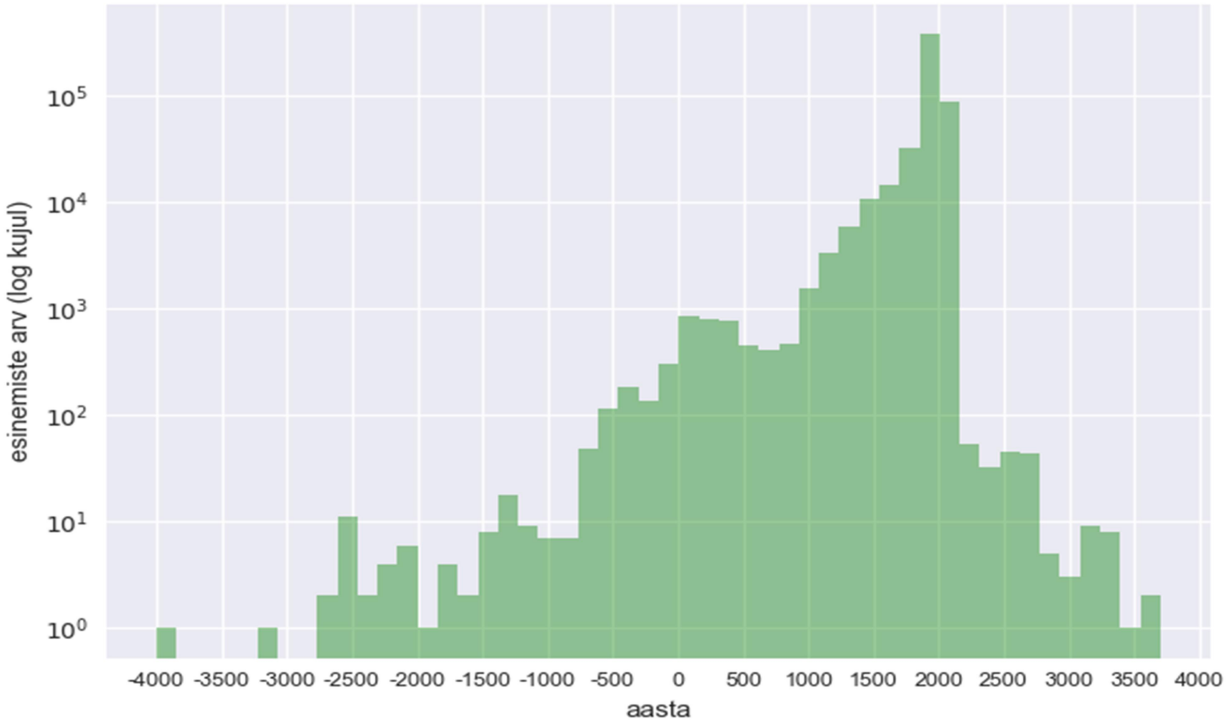
Vikipeedia biograafilised artiklid	
Granulaarsus	Osakaal
Aasta	77.2%
Kuu	3.3%
Nädal	0.0%
Päev	19.5%

Tabelis on toodud granulaarsuse jaotuste osakaal eristatud nelja tüübi – aasta, kuu, nädal, päev – vahel. Koheselt hakkab silma, et nädala granulaarsusega ajaväljendeid Vikipeedia biograafiates peaaegu ei kasutatagi. Neid esines kokku vaid 68 tükki. Lisaks selgus, et aasta granulaarsusega ajaväljendeid on ülekaalukalt kõige enam, mida võis ka eeldada. Peaaegu ühe viiendiku moodustavad ka päeva granulaarsusega ajaväljendid. Artiklite võrdlemise ja seoste otsimise näidetel annab jaotus peamiselt teada, et kõige enam peaks just aasta granulaarsusele keskenduma ja võib-olla kuu ja päeva granulaarsusega ajaväljendid laiendama samuti aasta granulaarsusele. Kuna päeva granulaarsusega ajaväljendeid oli päris arvestatav hulk, annab see eelduse katsetada ka päeva granulaarsust arvestavate ajaväljendite võrdlemist.

3.5 Aastate jagunemine Vikipeedia biograafiliste artiklite ajaväljendites

Siiani on vaadeldud artiklite jagunemist sünni- ja surma-aastate vahel, ajaväljendite jagunemist liikide vahel ja ajaväljendite granulaarsuste jagunemist. Eeltoodud tulemuste põhjal sobikski üheks keskseks uurimise punktiks võtta aastad Vikipeedia biograafilistes artiklites. Eeldada võib,

et aastate jagunemine näeb sarnane välja sünniaastate põhjal artiklite jagunemisega. Aastate uurimiseks sai nagu ka sünniaastate puhul loodud histogramm, mis on toodud joonisel 6.



Joonis 6. Vikipeedia biograafiliste artiklite aastate histogramm.

Joonisel on toodud x-teljel aastad 4000 eKr kuni 4000 pKr. Aastad eKr on märgitud negatiivsetena. Y-teljel on toodud aastate esinemise arv. Y-telg on nagu ka sünniaastate histogrammi puhul viidud logaritmilisele kujule. Koheselt on märgata võrreldes joonisega 4, kus oli kujutatud sünniaastate jagunemine, palju laiem jagunemine. Lähemal uurimisel selgus, et aastate 22. sajandist ja hilisemate puhul on peamiselt tegemist valetuvastustega, mida oli ka eeldada, vaadates nende hulka ja arvestades, et tegemist on biograafiliste artiklitega. Ajaväljendid 16. sajandist eKr ja varasemate puhul polnud aga enamasti tegu valetuvastustega, vaid leidus lihtsalt hulk biograafiad nendest sajanditest, millel polnud sünniaastat määratud, aga antud perioodide ajaväljendeid oli kasutatud. Järgmisena on näha võrreldes sünniaastate histogrammiga kordades suurem esinemiste arv, mida oli ka oodata, arvestades, et sünniaastaid esineb vaid üks iga artikli kohta, aga ajaväljendeid enamiku artiklite puhul rohkem. Kui vaadata kasvutrende, on võrdluses sünniaastate histogrammiga märgata selgeid sarnasusi. Näha on sajandite 5-1 eKr suurem esindatus võrreldes varasemate eKr sajanditega. Lisaks on märgata mõningane langus 6. sajandi paiku. Sarnaselt sünniaastatega on ka selged kasvutrendid alates 11.

sajandist, mis kulmineerib suurema enamusega 20. sajandist. Erinevustena tuleb välja, et kasvupunktid on isegi mõnevõrra tihedamad, millest võib järeldada, et lisaks artiklite arvule aastate lõikes suurenes ka ajaväljendite arv nendes artiklites. Lisaks on märgata suurt osa ajaväljendeid 21. sajandist, mida võis ka oodata, arvestades Vikipeedia olemust. Võrdluses sünniaastate histogrammiga saab välja tuua ka sajandite 1–5 pKr nähtavalt suurema esindatuse võrreldes sajanditega 1 –5 eKr, millest järeldub, et antud vahemikust artiklid sisaldavad enam ajaväljendeid ja on seega põhjalikumad.

Eeltöötluse peatükis oli välja toodud, et töö jaoks loodi kaks andmehulka. Üks, mis sisaldab artiklite kogutekste ja ajaväljendeid, ning teine, milles on osa artiklite lõpus olevaid sektsioone välja jäetud. Väljajätmise eelduseks oli, et väheneb valemärgenduste ja ajaväljendite, mis otseselt artiklit puudutava isikuga seotud pole, hulk. Tabelis 4 on seega toodud lisaks kogu tekstide hulgas esinenud aastate osakaalule ka välja jäetud sektsioonidega tekstide hulgas esinenud aastate osakaal.

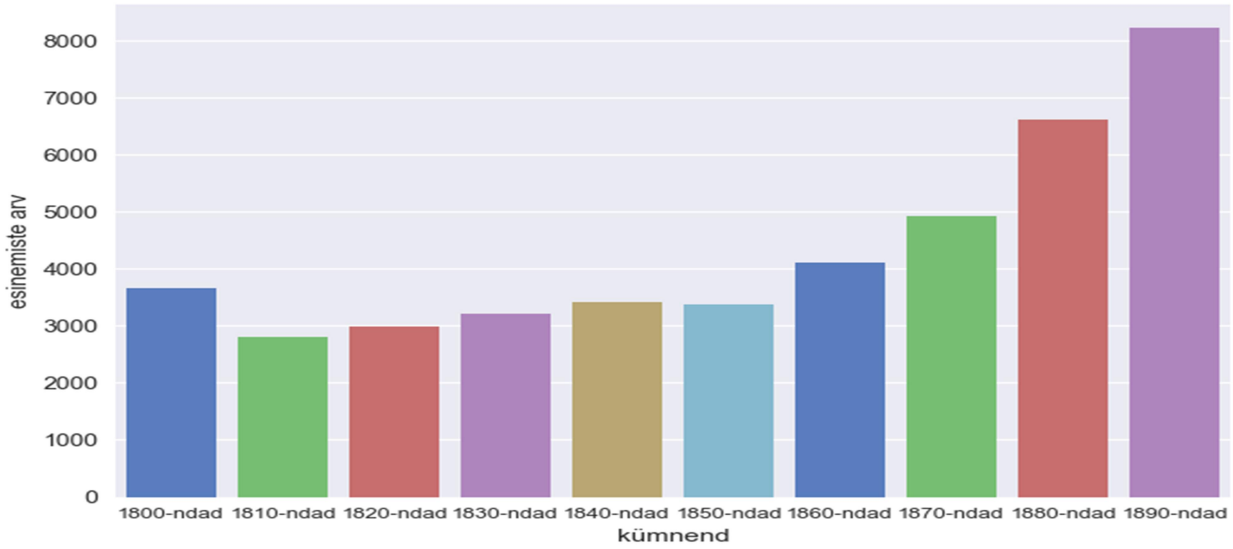
Tabel 4. Kogu tekstide ja välja jäetud lõpusektsioonidega tekstide aastate osakaal sajandite kaupa.

Kogu tekstide ja filtreeritud tekstide aastate osakaal sajandite kaupa		
Ajaperiood	Kogu tekstide aastaid – osakaal	Välja jäetud sektsioonidega tekstide aastaid – osakaal
21. sajand	132 817 – 24.8%	100 768 – 23.0%
20. sajand	302 671 – 56.4%	257 658 – 58.8%
19. sajand	43 787 – 8.2%	36 559 – 8.3%
11.–18. sajandid	51 702 – 9.6%	39 340 – 9.0%
1.–10. Ssajandid	4 055 – 0.8%	2 901 – 0.7%
eKr sajandid	834 – 0.2%	828 – 0.2%
22. sajand ja hilisemad	331 – 0.1%	135 – 0.0%

Tabelis on toodud aastaarvuliste ajaväljendite jagunemise osakaal sajandite kaupa, kus mõned sajandid on veel andmete esitamise selguse huvides grupeeritud. Tabelis on iga grupi kohta toodud nii esinemiste arv kui ka osakaal. Nagu ka joonisel 6 kujutatud histogrammi põhjal oli aru

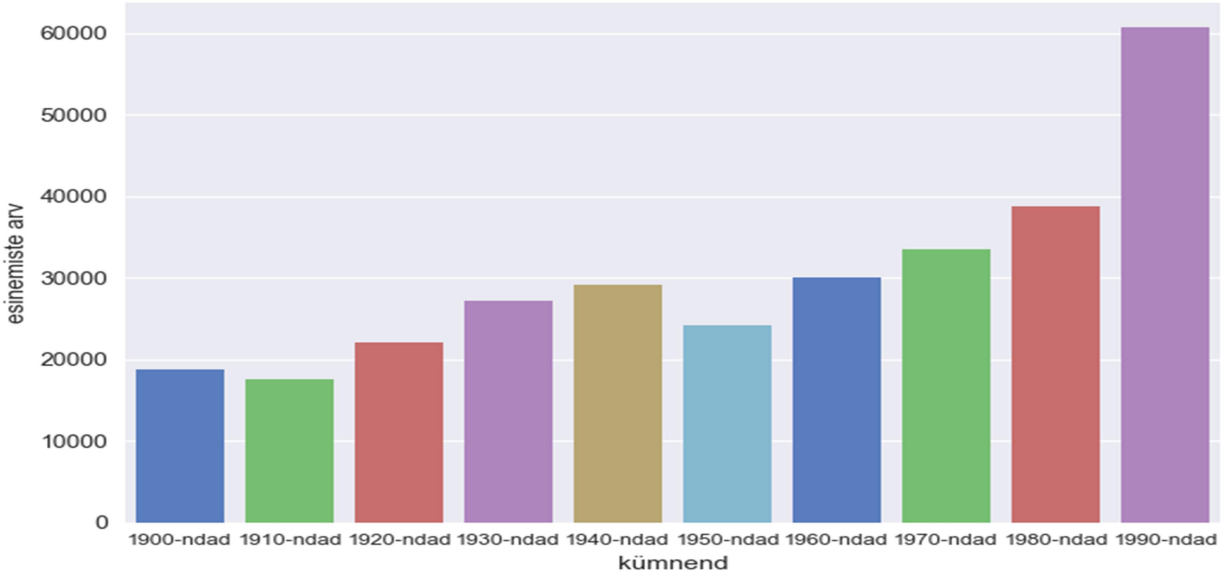
saada, moodustavad suurema osa aastad sajanditest 20 ja 21. Vikipeedia ja selle biograafiliste artiklite olemust kinnitab veelgi enam just 21. sajandi aastatele viitavate ajaväljendite vägagi suur osakaal, arvestades, et käimas on alles 17. aasta 21. sajandist. Kogu- ja filtreeritud tekstide võrdluses on märgata arvu poolest vähenemist kõigis gruppides. Kõige väiksem on vähenemine eKr aastate puhul, mida oli oodata. Kõige olulisemat vähenemist osakaalu mõttes on märgata 21. sajandist aastate puhul, mida võis ka eeldada, arvestades filtreeritud sektsioone. Huvitav on see, et on märgata ka selged vähenemised sajandite 1.–18. aastates. See omakorda annab aimu, et märgendusreeglite lõdvemaks muutmine tõi kaasa ka valemärgenduste suurenemise nende sajandite aastates, kuna antud aastaarvuliste ajaväljendite erilist esinemist välja jäetud lõpusektsioonides ei olnud eeldatav ja nende erilist esinemist välja jäetud lõpusektsioonides ei täheldanud ka töö autor.

Antud aastate jaotuse põhjal on näha, et huvipakkuvamad on just lähisajandid, sest need moodustavad niivõrd suure osa ajaväljenditest. Huvitav on just jälgida, kas sarnaseid kasvutrende on märgata ka läbi lähisajandite kümnendite. Joonisel 7 ja joonisel 8 on toodud vastavalt 19. sajandi ja 20. sajandi aastate jagunemine kümnenditesse.



Joonis 7. 19. sajandi Vikipeedia biograafilistes tekstides esineneud aastate jagunemine.

Joonisel on toodud x-teljel 19. sajandi kümnendid ja y-teljel antud kümnendist esinenud aastate arv. Jooniselt on näha, et aastate jagunemine sajandi esimesel poolel on jagunemine üsnagi võrdne. Kasvu on märgata alles 1860ndatest ja see jätkub stabiilselt kuni 1890ndateni.



Joonis 8. 20. sajandi Vikipeedia biograafilistes tekstides esineneud aastate jagunemine.

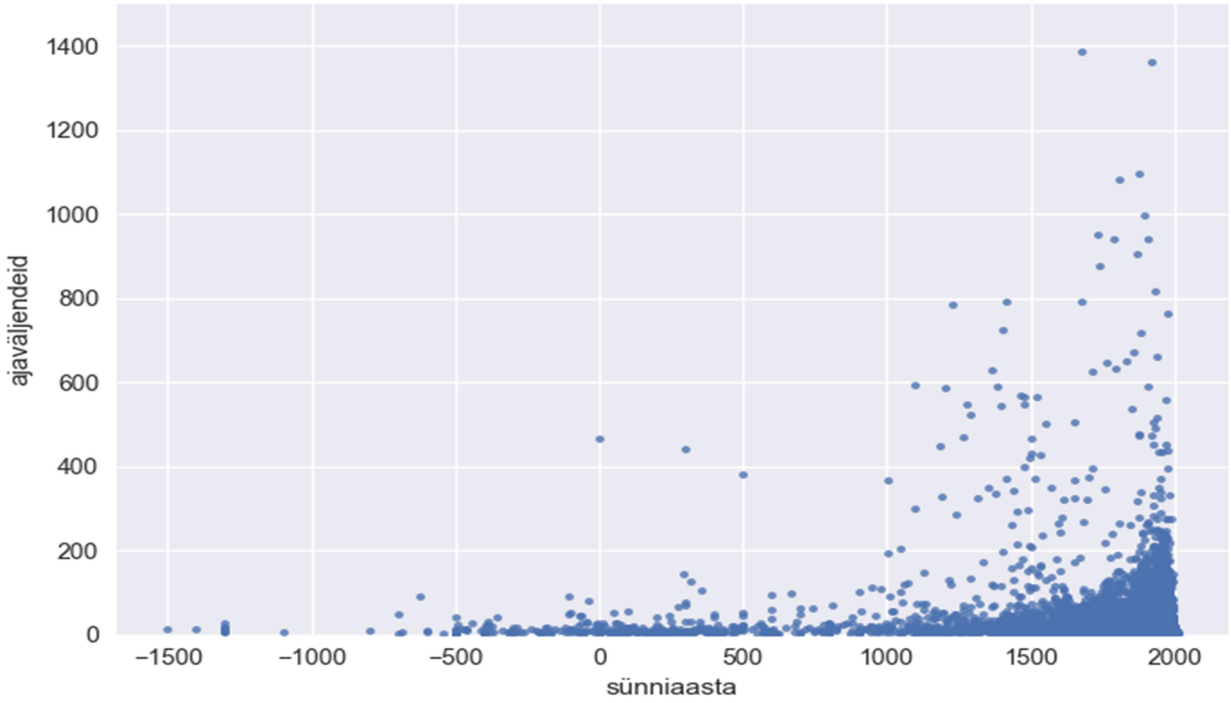
Joonis 8 on samasuguse ülesehitusega, aga käib 20. sajandi kümnendite kohta. Võrreldes 19. sajandi kümnendi jaotusega on arvud kordades suuremad. Kõige selgemalt jääb silma just 1990ndate osakaal võrreldes teiste kümnenditega. Seda oli mõneti ka oodata, arvestades jällegi Vikipeedia olemust ja infohulga kasvu. Stabiilset kasvu on ka antud kümnendite puhul märgata alates 1910ndatest kuni 1940ndateni. Sellele järgneb 1950ndates selge langus, mida annab arvatavasti taaskord seletada just teise maailmasõja mõjuga. Sellele järgneb stabiilne kasv kuni 1990ndateni, kus toimub juba oluline kasv.

Kolmandaks olulisemalt esindatud sajandiks oli käesolev ehk 21. sajand. Oodatult moodustavad 21. sajandi jaotusest 99,6% esimesed kaks kümnendit. Tõusu võrreldes 1990ndatega on märgata ka 2000ndate kasutuses. 2000ndaid aastaid oli kokku 88 871. Käesolevast kümnendist ehk 2010ndatest oli 43 357. Antud langus on täiesti loogiline, arvestades, et käesolev kümnend on alles käimas.

3.6 Vikipeedia biograafiliste artiklite ajaväljendite rikkalikkus

Omar Alonso jt [2] pakkusid oma töös välja ühe dokumendi mõõduna ajaväljendite alusel nn dokumendi ajaväljendite rikkuse. Nende töös oli antud mõõt defineeritud kui dokumendis esinenud ajaväljendite arv jagatud kogu korpuse ajaväljendite arvuga. Põhimõtteliselt väljendab mõõt dokumendi ajaväljendite osakaalu võrreldes kogu korpuse dokumentidega.

Seega võib antud töös uurimispunktina kasutada ajaväljendite rikkalikkust. Eelnevalt on alapeatükkides välja toodud artiklite jagunemine sünniaastate järgi, mis näitas artiklite hulga selget kasvu alates sünniaastatega 11. sajandist. Lisaks on välja toodud aastaid sisaldavate ajaväljendite jagunemine, mis näitab veel selgemini sarnast kasvu läbi sajandite. Antud tulemuste põhjal tekkis küsimus, kas sünniaastate poolest hilisemad artiklid on ka kõrgema ajaväljendite rikkalikkusega. See tähendab, et kas nad sisaldavad enam ajaväljendeid. Joonisel 9 on toodud artiklite sünniaastate põhisel jagunemisel ajaväljendite arvude graafik.



Joonis 9. Ajaväljendite arv artiklites sünniaastalise jagunemise põhjal.

Joonisel on x-teljel toodud sünniaastad alates 1500 eKr kuni 2000. Aastad eKr on märgitud negatiivsetena. Y-teljel on toodud ajaväljendite arv 0 kuni 1400. Jooniselt on välja jäänud kolm artiklit vastavalt 1900, 2500 ja 3100 ajaväljendiga, seda põhjusel, et ülejäänud jagunemine oleks selgemalt nähtav. Iga punkt tähistab ühte artiklit. Joonisel on loetavuse huvides ajaväljendite rikkalikkuse asemel toodud ajaväljendite arv, kuna töödeldud korpus oli niivõrd suur, et rikkalikkuse määrad tulid jagamisel liiga väikesed ja logaritmilisele skaalale oleks olnud liialt suur oht kaotada tõlgendatavus. Joonisena näevad rikkalikkuse ja ajaväljendite arvude jagunemine niikuinii identsed välja, kuna rikkalikkus määrati nii, et ajaväljendite arv artiklis

jagati kogu korpuse ajaväljendite arvuga, ja kuna kogu korpuse ajaväljendite arv on iga artikli puhul sama, siis on põhimõtteliselt lihtsalt see ajaväljendite arvu jagamise osa ära jäetud.

Jooniselt paistab selgemini välja artiklite, mis sisaldavad üle 200 ajaväljendi, jaotus. Selge on, et antud ajaväljendid esinevad peamiselt just artiklites, mis algavad sünniaastaga 1000 ja hiljem. Aga samas on nende jagunemine niivõrd hajunud, et mingeid väga selgeid järeldusi ajalise jaotuse ja nende artiklite rikkalikkuse põhjal teha ei saa. Küll on aga artiklite, mis sisaldavad alla 200 ajaväljendi, jaotusest märgata, et umbes aasta 1000 paiku on lisaks paigutuse tiheduse kasvule näha ka ajaväljendite arvu kasvu nendes tihedamalt paiknevate artiklite hulgas. Antud jaotus ei näita küll, et sünniaastate põhjal lähisajandite artiklid on kindlasti kõrgema ajaväljendite arvuga, sest artikleid lähisajanditest on niivõrd rohkem ja suur hulk neist võib olla ka madala ajaväljendite arvuga. Küll aga on jaotusest näha, et alates aastast 1000 edasi liikudes kasvab artiklite arv, mis sisaldavad rohkem ajaväljendeid ehk millel on kõrgem ajaväljendite rikkalikkus. Jooniselt paistev lähisajandite artiklite, mis sisaldavad ka näiteks 20 või enam ajaväljendit, küllaltki tihe paigutus näitab, et moodustades korpust, mille peal testida võrdlemismeetodit võib sätestada ka kõrgema ajaväljendite piiri ehk lävendi, millest rohkem ajaväljendeid võiks dokument võrdlemiseks sisaldada. Seda eesmärgil, et iga artikli puhul oleks võrdlemisalust materjali piisavalt.

4. Seotud või sarnaste artiklite leidmine ajaväljendite alusel

Antud peatükis analüüsitakse võrdlusmeetodit, mida kasutati seotud artiklite leidmiseks. Kirjutatakse ka antud meetodi testimisest ja antakse esialgne hinnang meetodile arvestades ajaväljendite põhjal leitud sarnastes artiklites kujutatud inimeste vahel seotust.

Eelnevas peatükis toodud andmed ja analüüs tõid esile, millised ja millisel määral ajaväljendid Vikipeedia biograafilistes artiklites võrdluseks sobivad. Kuna valdav enamus on just kalendrilise toimumisajaga ajaväljendid, siis see tähendab, et need saab paigutada ajateljele ja nende kokkulangemine ütleb palju enam, kui näiteks ajaliste korduvuste või ajaliste kestuste puhul. Lisaks on valdav enamus nendest absoluutsed ajaväljendid, mis ei vaja ajateljele paigutamiseks lisainfot ja seega sobivad võrdlemisaluseks paremini kui relatiivsed. Seega seoses alapeatükis 1.2 kirjeldatud probleemidega, mis võivad relatiivsete ajaväljendite puhul võrdlemisel ilmneda, on antud töös artiklite võrdlemisel relatiivsed ajaväljendid välja jäetud. Lähisajandite ajaperioode kirjeldab palju suurem hulk artikleid ja seega on lähisajanditest artikleid valides rohkem materjali, millega võrrelda. Lisaks just lähisajandite puhul on suurem hulk artikleid, mis on küllaltki suure ajaväljendite rikkalikkusega ehk sisaldavad vähemalt näiteks 10 või 15 ajaväljendit, mis tagab just lähisajandite artiklite puhul selle, et leidub piisavalt võrdlemisalust materjali ehk antud juhul ajaväljendeid.

4.1 Võrdlusmeetod leidmaks sarnaseid artikleid ajaväljendite alusel

Võrdlusmeetodi baasiks võeti alapeatükis 1.3 kirjeldatud Omar Alonso jt [2] esitatud raamistik. Kuna tegemist on kõigi artiklite puhul biograafiliste artiklitega, siis alampiiriks ja ülempiiriks on võetud vastavalt sünni- ja surma-aasta, erinevalt nende raamistikust, kus alampiiriks ja ülempiiriks olid vastavalt ajateljel varaseim ja hiliseim dokumendis esinenud ajaväljend. Kui surma-aasta ei olnud määratud ja kui sünniaasta oli viimase 100 aasta sees, siis ülempiiriks võeti aasta 2016. Kuna võrdlemiseks tuleb nende raamistiku järgi viia ajaväljendid samale granulaarsusele, siis antud töö kontekstis tähendaks see aastalise granulaarsusega ajaväljendite võtmist aluseks ja kuu- ja kuupäeva granulaarsusega ajaväljendite laiendamist aasta granulaarsusele. Töö käigus katsetas autor esialgu otseselt Omar Alonso jt [2] raamistikus välja pakutud valemit kahe artikli vahelise sarnasuse määramiseks. Seejärel, et vähendada juhuslikkust, mis tuleneb vaid aasta granulaarsusega ajaväljendite kokkulangemistel, otsustas autor sisse viia mõned muudatused, jättes küll kauguse arvutamise valemi olemuse ja

tööpõhimõtte samaks. Artiklite vahelise sarnasuse arvutamiseks, testimiseks ja hindamiseks on loodud skript *sarnased*.

Esimese erinevusena on see, et võrdlemisaluste gruppidena ei moodustata grupid alampiirist ülempiirini, vaid vastava granulaarsusega ajaväljendite põhjal, mis artiklis esinevad. Näiteks võttes granulaarsuseks aasta granulaarsuse, siis on esimene artikkel järgmine: see sisaldab aastaarvulisi ajaväljendeid, mille loend on toodud *artikkel1Aastad*, mis koosneb elementidest $b_0,b_1,...,b_{m-1}$ ehk näiteks 1898, 1917, 1917, 1918, 1920, 1924, 1925, ja teine artikkel on järgmine: see sisaldab aastaarvulisi ajaväljendeid, mille loend *artikkel2Aastad*, mis koosneb elementidest $c_0,c_1,...,c_{p-1}$ ehk näiteks 1901, 1909, 1917, 1917, 1918, 1919, 1924, 1924, 1928, 1931. Nende kahe artikli alusel moodustatakse võrdlemisalused grupid kahe loendi ühendit kujutavast hulgast ehk näiteks moodustatakse hulk *esinevadAastad*, mis sisaldab elemente $a_0,a_1,...,a_{n-1}$, mis antud näite puhul on järgmine: 1898, 1901, 1909, 1917, 1918, 1919, 1920, 1924, 1925, 1928, 1931. Elemendid on alati järjestatud kasvamise järjekorras. Loendis võib sama element mitu korda esineda, hulgas ei või. Seejärel leitakse kahe artikli vahel aastaarvuliste ajaväljendite kaugus sarnaselt alapeatükis 1.3 toodud valemis (1) esitatud loogikaga. Joonisel 10 toodud lõik koodist, kus on kujutatud kahe artikli vahel aastaarvuliste ajaväljendite põhjal kauguse leidmine.

```
for i in range(0,len(esinevadAastad)):
    vahesum = 0
    for j in range(0,i+1):
        a = esinevadAastad[j]
        vahesum += artikkel1Aastad.count(a) - artikkel2Aastad.count(a)
    kaugus += abs(vahesum)
```

Joonis 10. Näide aastaarvuliste ajaväljendite põhjal kahe artikli vahel kauguse arvutamisest

Jooniselt on näha, et kood koosneb kahest tsüklist. Esimene tsükkel kujutab valemis (1) kujutatud esimest summat nullist kuni hulga *esinevadAastad* elementide arvuni ehk *n*-ini. Esimese summa liidetavad on, sarnaselt valemi (1) teisele summale, omakorda absoluutväärtused summadena *j*-ist *i*-ni, mis on kujutatud teise tsüklina. Selle sees kujunevad vahesumma elemendid vastavalt esialgu grupi a_o, ehk toodud näites aasta 1898, esinemiste arv esimeses artiklis lahutada esinemiste arv teises artiklis. Ja niimoodi edasi, vastavalt *j* suurenedes võrreldes grupi a_j esinemiste arv esimeses artiklis lahutada esinemiste arv teises artiklis, kuni viimase grupini ehk a_{n-1} ehk antud näites aasta 1931. Vastav summa, mis on kujutatud kui muutuja

kaugus viitabki kahe artikli vahelisele kaugusele aastaarvuliste ajaväljendite alusel. Mida rohkem on kokku langevaid või lähestikuseid ajaväljendeid, seda väiksem tuleb kaugus.

Teise erinevusena on see, et kaugus ei leita mitte ainult kõige laiema granulaarsusega ajaväljendite alusel, vaid vaadetakse nelja erinevat hulka. Lisaks aastaarvulistele ajaväljenditele leitakse kaugus veel kuu granulaarsusega ajaväljenditel, kuupäeva granulaarsusega ajaväljenditel ja ka ajavahemikel. Seega leitakse meetodi sees neli kaugust täpselt sama struktuuri alusel, nagu sai kirjeldatud aastaarvuliste ajaväljendite näite puhul. Erinevuseks on vaid, et vaadeldakse hulkasid - *esinevadAastadKuud*, mis koosnevad võrreldavates artiklites esinevatest kuu granulaarsusega ajaväljenditest; *esinevadAastadKuudKuupäevad,* mis koosnevad võrreldavates artiklites esinevatest kuupäeva granulaarsusega ajaväljenditest; ja *esinevadAjavahemikud*, mis koosnevad võrreldavates artiklites esinevatest ajavahemikest. Lisaks vaadeldakse ainult absoluutseid ajaväljendeid, mis tähendab, et kitsama granulaarsusega ajaväljendid on esindatud ka laiema granulaarsusega hulkades ehk näiteks kuupäeva granulaarsusega ajaväljend *1936-05-22* esineb nii kuu granulaarsusega ajaväljendite hulgas kujul *1936-05* ja aastalise granulaarsusega ajaväljendite hulgas kujul *1936*. Samamoodi esinevad ajavahemike otspunktid teistes hulkades. Antud laiendused on tehtud põhjusel, et suurendada kuupäeva ja kuu granulaarsusega ajaväljendite kokkulangemiste mõju artiklite sarnasemaks tegemisel. Kui kuupäeva granulaarsusega ajaväljend langeb kokku nii aastaarvuliste ajaväljendite põhjal kauguse arvutamisel, kuu granulaarsusega ajaväljendite põhjal kauguse arvutamisel kui ka kuupäeva granulaarsusega ajaväljendite põhjal kauguse arvutamisel, siis tähendab, et kaugus tuleb madalam kolmes arvutatavas kohas, mis tõstab kitsama granulaarsusega ajaväljendite väärtust.

Kolmandaks kuna uurimus näitas, et aasta granulaarsusega ajaväljendeid on ülekaalukalt kõige rohkem ja kuna nad pole nii spetsiifilised, siis langevad nad ka tihedamini kokku. Seega on kuu granulaarsusega, kuupäeva granulaarsusega ja ajavahemike põhjal arvutatavad kaugused läbi korrutatud teguriga, mis sõltub kahe artikli peale esinevate aastaarvuliste ajaväljendite arvust, mis on jagatud siis vastavalt kahe artikli peale esinevate kuu granulaarsusega ajaväljendite või kuupäeva granulaarsusega ajaväljendite või ajavahemike arvuga. Ehk siis kuupäeva granulaarsuste põhjal arvutatud kaugus korrutatakse läbi täisarvulise teguriga, mis saadakse hulga *esinevadAastad* elementide arvu jagamisel hulga *esinevadAastadKuudKuupäevad* arvuga. Kuna kuupäeva granulaarsusega ajaväljend on alati laiendatuna aasta granulaarsusele toodud ka

hulgas *esinevadAastad,* on hulk *esinevadAastad* alati suurem või võrdne hulgaga *esinevadAastadKuudKuupäevad*. See korrutamine normaliseerib mõnevõrra kuupäeva granulaarsusega ajaväljendite põhjal leitud kaugust aastalise granulaarsusega ajaväljendite põhjal leitud kaugusega vastavalt hulkade suuruse erinevusele. See kordajaga korrutamine aitab arvesse võtta kauguse arvutamisel natukene ka artiklite ülesehitust just vastavate ajaväljendite sisalduse mõttes. Näiteks vaadates kolme artiklit ja mõeldes ainult nende vaheliselt arvutatavale ajavahemike kaugusele, kui esimesed kaks artiklit sisaldavad mõlemad ühte ajavahemikku, mis omavahel ei kattu ja kolmas ei sisalda ühtegi ajavahemikku, siis vastavalt valemile ajavahemike põhjal arvutatav kaugus nende kõigi vahel tuleb sama ehk 1. Aga kuna esimesed kaks artiklid sisaldavad mõlemad ühte ajavahemikku, siis nad on ülesehituselt sarnasemad ja kuna nende vahel on hulga *esinevadAjavahemikud* elementide arv suurem, siis sellest tulenevalt korrutatav tegur väiksem, mis lõpptulemuses tagab, et nende vaheline kaugus on väiksem kui näiteks esimese ja kolmanda artikli kaugus.

Kahe artikli vaheline kogukaugus on leitud nende nelja mainitud arvutatava kauguse summana. Skript *sarnased* sisaldab funktsiooni *kaugusfunktsioon*, mis saab argumentideks kaks artiklit ja seejärel tagastab nendevahelise kauguse vastavalt kirjeldatud meetodile. Kirjeldatud võrdlusmeetodi olemus ja loodud *kaugusfunktsiooni* esitus on toodud kui pseudokood 1.

Pseudokood 1. Kaugusfunktsioon ajaväljendite põhjal kahe artikli vahelise kauguse leidmiseks.

```
1  kaugusfunktsioon(a1,a2):
2        a_o,a_1,...,a_n-1 = {a1.aastad}U{a2.aastad}
3        b_o,b_1,...,b_m-1 = {a1.kuud}U{a2.kuud}
4        c_o,c_1,...,c_o-1 = {a1.päevad}U{a2.päevad}
5        d_o,d_1,...,d_p-1 = {a1.ajavahemik}U{a2.ajavahemik}
6        kaugusAasta = ∑_{i=0}^{n-1} |∑_{j=0}^{i} (a1.aastad.count(a_j) − a2.aastad.count(a_j))|
7        kaugusKuu = ∑_{i=0}^{m-1} |∑_{j=0}^{i} (a1.kuud.count(b_j) − a2.kuud.count(b_j))|
8        kaugusKuupäev = ∑_{i=0}^{o-1} |∑_{j=0}^{i} (a1.päevad.count(c_j) − a2.päevad.count(c_j))|
9        kaugusAjavahemik = ∑_{i=0}^{p-1} |∑_{j=0}^{i} (a1.ajavahemik.count(d_j) − a2.ajavahemik.count(d_j))|
10       kaugusKuu = kaugusKuu *  int(n/m)
11       kaugusKuupäev = kaugusKuupäev * int(n/o)
12       kaugusAjavhemik = kaugusKuu * int(n/p)
13       kaugus = kaugusAasta + kaugusKuu + kaugusKuupäev + kaugusAjavahemik
14       return kaugus
```

Funktsioon saab sisendiks kaks artiklit *a1* ja *a2*, millele on mõlemal määratud loendid – *aastad*, aastaarvuliste ajaväljendite jaoks; *kuud*, kuu granulaarsusega ehk aasta ja kuu osast koosnevate ajaväljendite jaoks; *päevad*, kuupäeva granulaarsusega ehk aasta ja kuu ja kuupäeva osast koosnevate ajaväljendite jaoks; *ajavahemik*, ajavahemikkude jaoks. Loogeliste sulgudega ümbritsemine tähistab hulga võtmist ehk ära jäetakse elementide kordused. Funktsioon *count* tähistab vastavast loendist elemendi esinemiste arvu võtmist, kus võetav element on toodud sulgude sees. Funktsioon *int* tähistab vastavalt sulgude sees toodud osast täisarvu osa võtmist. Pseudokoodis toodud jadad – $a_o,a_1,\ldots,a_{n-1}$; $b_o,b_1,\ldots,b_{m-1}$; $c_o,c_1,\ldots,c_{o-1}$; $d_o,d_1,\ldots,d_{p-1}$ tähistavad vastavalt hulkasid, mis meetodi kirjelduses ja implementatsioonis olid kirjeldatud kui *esinevadAastad*, *esinevadAastadKuud*, *esinevadAastadKuudKuupäevad* ja *esinevadAjavahemikud*. Antud hulkades on kõikides elemendid ehk ajaväljendid järjestatud ajateljel varaseimast hiliseimani. Pseudokoodis toodud kaugused, mis on toodud summana, on kõik realiseeritud vastavalt joonisel 10 toodud näitele. Ridadel 10 kuni 12 on näha vastavalt kordajaga läbikorrutamised, mis koodis realiseeruvad vaid juhul, kui vastavad hulkade pikkused *m*, *o*, *p* on nullist suuremad.

4.2 Meetodi esimene testimine

Seoses valetuvastustega ja biograafias kujutatud isikuga mitte seotud ajaväljenditega (kirjeldatud alapeatükis 2.3), kasutati testimiseks ja hindamiseks korpust, milles olid lõpusektsioonid „kirjandus“, „välislingid“, „viited“ ja „publikatsioon“ välja jäetud. Esialgse testimise käigus selgus, et eriti just kirjanike kohta käivate biograafiliste artiklite puhul, kus on pikad teoste loetelud, on küll väga tihti just aastaarvuliste ajaväljendite puhul kokkulangevusi, aga kuna tegemist on just teoste ilmumisaastatega, mida on artiklite vahel väga keeruline seostada ja mis tihtipeale võib negatiivselt mõjutada tulemusi, seega korrati välja jäetud sektsioonidega korpuse moodustamise protsessi, millesse oli juurde lisadtud välja jäetavate märksõnade hulka ka „väljaanded“ ja „teos“.

Kuna statistika moodustamiseks loodud korpused sisaldavad kõiki ajaväljendeid ühes loendis, aga võrdluseks kasutati ainult absoluutseid kalendrilise toimumisajaga ajaväljendeid ja ajavahemikke, millede otspunktid olid absoluutsed ajaväljendid, siis skriptis *sarnased* on loodud funktsioon *artiklist*, mis võtab argumendiks artikli faili statistika jaoks moodustatud korpusest. Funktsioonis koostatakse neli loendit – *aastad*, mis sisaldab kõiki artiklis esinevaid aastaarvulisi ajaväljendeid; *aastadKuud*, mis sisaldab kõiki aastast ja kuust koosnevaid ajaväljendeid;

aastadKuudKuupäevad, mis sisaldab kõiki aastast, kuust ja kuupäevast koosnevaid ajaväljendeid; *ajavahemikud*, mis sisaldab kõiki ajavahemikke, kus otspunktid on absoluutsed ajaväljendid. Lisaks kõik kitsama granulaarsusega, näiteks kuupäeva granulaarsusega ehk aastast, kuust ja kuupäevast koosnevad ajaväljendid esinevad ka laiema granulaarsusega loendis ehk aasta osa sisaldub aastate loendis. Samuti määratakse funktsioonis alampiir ja ülempiir vastavalt sünni- ja surma-aastatele. Kuna alam- ja ülempiire pole kasutatud mitte võrreldavate gruppide moodustamiseks, vaid lihtsalt määramiseks, kas artiklite eluperioodid langesid kokku ja läbi selle kas on üldse mõtet artiklite kaugust mõõta, siis eeldusel, et esimese 15 eluaasta jooksul suuremal osal isikutel puuduvad saavutused või sündmused, mida selgelt seostada teiste artiklitega, on alampiiris sünniaastale veel liidetud 15. Sellega üritab autor kitsendada võrdlemispiire, mis on määratud alam- ja ülempiiride põhjal, et vältida ebavajalike kauguste leidmist. Funktsioon tagastab seitsmest kauguse ja sarnaste artiklite leidmiseks vajalikust osast koosneva muutuja. Tagastatakse artikli nimi, alumine piir, ülemine piir, aastate loend, aastate ja kuude ajaväljendite loend, aasta ja kuu ja kuupäeva ajaväljendite loend, ajavahemike loend.

Järgmisena, et kogu aeg ei peaks kordama kogu ajaväljendite hulgast vajalike eraldamist ja grupeerimist, on loodud funktsioon *moodustakorpus*. Funktsioonil on viis argumenti. Kõigepealt kaust, mille põhjal võrdluskorpus moodustatakse ehk välja jäetud sektsioonidega korpuse kaust. Teisena kaust, kuhu võrdluskorpus kirjutatakse. Kolmandana aastaarvuliste ajaväljendite arv ehk minimaalne aastaarvuliste ajaväljendite arv, mida artikkel peab sisaldama. Seda põhjusel, et garanteerida, et iga võrdluskorpuses artikkel sisaldaks piisavalt võrdlemismaterjali. Neljandaks ja viiendaks vastavalt sünnivahemike alumine ja ülemine piir ehk vahemik, millesse peab artikli sünniaasta jääma, et saaks uurida kindlaid ajaperioode. Vastavalt uurimuse tulemusele keskenduti testimisel ja hindamisel just 19. ja 20. sajandist sünniaastatega artiklitele. Esialgu testiti ka varasemaid perioode, aga nagu eeldada võis, ei leidunud nende puhul piisvalt võrdlusmaterjali.

Võrdluskorpusest sarnaste artiklite leidmiseks on loodud funktsioon *leiaSarnased*, mis võtab argumentideks artikli, millele otsitakse sarnaseid, võrdluskorpuse, mille seast sarnaseid artikleid otsitakse ja ka ajaväljendite miinimumarvu, mis peavad kahe artikli vahel kattuma, et kaks artiklit omavahel sarnaseks lugeda. Seda põhjusel, et kuna väga palju on tegemist juhuslike kokkulangevustega ja eesmärgil, et suurendada tõenäosust, et tegemist ei ole juhusliku

kokkulangevusega. Funktsioon määrab antud artikli ja võrdluskorpuse artiklite vahel kaugused kasutades *kaugusfunktsioon*'i. Funktsioon määrab kauguse vaid juhul, kui võrdluskorpuse artikli alumine või ülemine piir jääb argumendina antud artikli alumise ja ülemise piiri vahele. Funktsioon reastab viis sarnasemat vastavalt viie väikseima kauguse järgi ja väljastab ekraanile nende artiklite nimed, kaugused ja ka hulga ajaväljenditega, mis kahe artikli vahel kattusid.

Esialgne meetodi testimine toimus juhuslikult valitud artiklite abil. Koheselt selgus, et kui artiklid on kauguse poolest sarnased, ei tähenda, et artiklites kujutatud isikud tingimata seotud oleks. Väga suur osa sellest tulenes just aastaarvuliste ajaväljendite kokkulangemiste juhuslikkusest, mis oli ka üheks peamiseks põhjuseks, miks kauguse arvutamisel said juurde lisatud ka kuu, kuupäeva granulaarsusega ajaväljendid ja ajavahemikud. Samas, kuna aasta granulaarsusega ajaväljendeid leidub lihtsalt nii palju enam ja lisaks on ka nende kokkulangemine kordades tõenäolisem, siis alati pole nendest lisatäpsustustest kasu. Lisaks pole mingeid garantiisid, et artiklile leidub üldse viis sarnast artiklit, mille vahel isikud otseselt seotud oleks või seotus on mingis nii kitsas ajalises piiris ja ülejäänud artikli osad on niivõrd erinevad, et kauguse põhjal ei tule need artiklid esile. Küll aga märkas töö autor ka mitmeid positiivseid tulemusi, kus valitud artikli puhul sai väita, et kõik viis ajaväljendite kauguse põhjal leitud artiklit sai ühel või teisel moel valitud artikliga siduda.

Näiteks Julius Ellandi, kes oli 20. sajandi esimese poole Eesti ja Saksamaa sõjaväelane, kohta käiva artikli[12] puhul olid kauguse poolest sarnasemad viis artiklit, kus isikud olid kõik samuti sõjaväelased samast ajaperioodist. Näiteks üks nendest viiest oli Hans Hirvelaan[13], kes oli samuti sama ajaperioodi Eesti sõjaväelane. Artiklite puhul on näha seoseid just läbi suuremate sündmuste, nagu näiteks Eesti vabadussõda või esimene maailmasõda, kus sündmused on läbi paljude ajaväljendite hästi artiklis välja toodud ja läbi selle ka ajaväljendite põhjal arvutatavad kaugused on väiksemad. Lisaks võib näha selgeid sarnasusi artiklite ülesehituses välja toodud sektsioonides, artikli pikkustes ja ajaväljendite kasutuses ja olemuses.

Teise huvitava näitena tooks välja Nikolai Ogarkovi, kes oli 20. sajandi Nõukogude Liidu sõjaväelane, kohta käiva artikli[14]. Ja samuti kõigi viie sarnasema artikli puhul olid isikud

[12] https://et.wikipedia.org/wiki/Julius_Ellandi (Vaadatud 30.04.2017)
[13] https://et.wikipedia.org/wiki/Hans_Hirvelaan (Vaadatud 30.04.2017)
[14] https://et.wikipedia.org/wiki/Nikolai_Ogarkov (Vaadatud 30.04.2017)

Nõukogude liidu sõjaväelased ja mis võib-olla veelgi huvitavam, on asjaolu, et nii Ogarkov kui ka kõik viis sarnasemat isikut olid auastmelt Nõukogu Liidu marssalid. Näiteks üks neist viiest oli Pjotr Koševoi[15]. Samuti on lisaks sõjaväelisele seosele artiklite vahel näha ka selgeid sarnasusi artiklite ülesehituses nii sektsioonides kui ka ajaväljendite kasutuses ja tüüpides.

Kolmanda huvitavama näitena tooks välja August Alekõrsi, kes oli 20. sajandi Eesti kohtunik, kohta käiva artikli[16]. Samuti nagu eelnevate näidete puhul oli selgeid seoseid märgata kõigi viie kauguse poolest määratud sarnasema artikli puhul. Artiklite ülesehituses olid taaskord jällegi selged sarnasused, kuna kõik artiklid olid ühelõigulised, küllaltki lühikesed ja koosnesid ajaväljendite poolest peamiselt ajavahemikest. Kuigi artiklid ei olnud nii väga põhjalikud, kui sõjaväelaste näite puhul, sai siiski selgeid näited tuua nende kohta kus artiklite puhul kattuvad õpinguaastad Tartu Ülikooli õigusteaduskonnas või kuulumine Eesti Üliõpilaste Seltsi või jällegi osalemine Eesti vabadussõjas. Näiteks üks viiest artiklist oli Harald Oerti kohta käiv artikkel[17].

Toodud näited annavad aimu, et kuigi artikleid on palju ja nende vahel on just aastaarvuliste ajaväljendite puhul kokkulangemiste tõenäosused kõrged, siis olenevalt artikli tüübist ülesehituse mõttes ja artiklis kujutatava inimese elukutsest ja sündmustest, kus isik osales, või organisatsioonidest, kuhu isik kuulus, on võimalus ajaväljendite põhjal seoseid tõmmata küll.

4.3 Meetodi hindamine

Üks töö eesmärk oli puhtalt ajaväljenditel põhinevale võrdlusmeetodile ka esialgne hinnang anda. Hinnangu määramisel võeti eeskujuks infootsingu hindamisel tuntud meetod täpsus k-s (ing. k. *precision at k*), kus k on ühes otsingus kuvatavate tulemuste arv [17]. Kasutati metoodikat, kus juhusliku valikuga valiti artikleid võrdluskorpusest ja igale artiklile leiti võrdlusmeetodiga kauguse poolest viis sarnasemat. Hindamisel võeti eesmärgiks hinnata seotust just artiklites kujutatavate isikute vahel artiklite sisu põhjal. Kasutati kaalutud hindamist, kus hinne 1,0 omistati siis, kui artiklitevahelised isikud olid mõlemad osalised mõnes samas sündmuses, nagu näiteks Eesti vabadussõda, või kui artiklite- vahelised isikud kuulusid samadel või kattuvatel ajaperioodidel samasse organisatsiooni, näiteks Tartu Ülikooli õigusteaduskond. Hinne 0,5 omistati, kui artiklitevahelised isikud olid kattuvatel ajaperioodidel samade või

[15] https://et.wikipedia.org/wiki/Pjotr_Ko%C5%A1evoi (Vaadatud 30.04.2017)
[16] https://et.wikipedia.org/wiki/August_Alek%C3%B5rs (Vaadatud 30.04.2017)
[17] https://et.wikipedia.org/wiki/Harald_Oert (Vaadatud 30.04.2017)

sarnastel elukutsetel või ametikohtadel ja asukoha poolest piisavalt lähestikku, et võib loogiliselt eeldada mingisugust seost. Näiteks kaks Eesti vaimulikku, üks Võru praostkonna praost, teine Valga praostkonna praost. Ülejäänud juhtudel omistati hinne 0,0. Hinnangu tulemus saadi vastavalt jagades hinnete summa kogu määratud artiklite arvuga. Hindamiseks moodustati erinevaid võrdluskorpuseid, mille artiklitel jäid sünniaastad vahemikku 1800–2000 ja mis sisaldasid minimaalselt 8–20 aastaarvulist ajaväljendit. Valides võrdluskorpusest korraga 10 juhuslikku artiklit, määrati igaühele viis sarnasemat, millel oleks vähemalt 2–5 kokkulangevat võrreldavat ajaväljendit. Kokku valiti juhuslikult 70 artiklit, mille põhjal määrati kokku 350 sarnasemat artiklit. Üle kõigi juhuslikult valitud artiklite saadi antud meetodil hinnang, et ajaväljendite põhjal sarnaselt määratud artiklid on seotud keskmise täpsusega 0,18. Paremaid tulemusi oli märgata, kui isikud olid elukutselt sõjaväelased, vaimulikud või poliitikud või kui isikud olid osalised mingites suuremates ja paremini kaardistatud sündmustes, näiteks nagu Eesti vabadussõda. Paremaid tulemusi oli näha ka siis, kui tegemist oli rahvuselt eestlastega või tegutsetud oldi Eesti aladel. Väga palju oli siiski märgata juhuslikkust ja eriti just aastate kokkulangemisel. Tulemuste põhjal võib väita, et otseselt seotud artiklite määramiseks kõigi artiklite puhul antud meetodist, mis põhineb ainult ajaväljenditel, antud korpuse puhul ei piisa. Kuna artiklite hulk on küllaltki suur ja artikleid on väga erineva sisuga ja ajaväljendite poolest peamiselt aastaarvulistest ajaväljenditest koosnevad, siis juhuslike kokkulangevuste hulk on suurem kui selgetest seostest tingitud kokkulangevuste hulk.

Kokkuvõte

Bakalaureusetöö esimene eesmärk oli uurida ajaväljendite kasutust Vikipeedia biograafilistes artiklites ja välja selgitada sünniaastate, ajaväljendite liikide, aastaarvuliste ajaväljendite jagunemised, ajaväljendite granulaarsus, ajaväljendite rikkalikkus ja vastavalt välja selgitada, millised ajaväljendid ja millisel määral sobivad võrdlemiseks Vikipeedia biograafilistes artiklites. Bakalaureusetöö teine eesmärk oli vastavalt leiduvatele ajaväljenditele, kasutades ja vajadusel modifitseerides Omar Alonso jt [2] välja pakutud raamistikku, testida ja hinnata meetodit, leidmaks seotud või sarnaseid artikleid Vikipeedia biograafiate hulgast.

Bakalaureuse töö esimeses eesmärgis sõnastatud uurimuse jaoks moodustati 33 453 Vikipeedia biograafilisest artiklist koosnev korpus. Korpuse artiklite peal tuvastati sünniaastate põhjal artiklite jagunemine. Sünniaasta oli toodud ligikaudu 94% artiklitest. Sünniaastate poolest jagunesid artiklid peamiselt just lähisajanditesse. Ligikaudu 66% oli 20. sajandist ja ligikaudu 92% oli sajanditest 15–20. Kasutades EstNLTK ajaväljendite tuvastajat, millel töö autor pidi vastavalt tuvastusreeglites aastaarvuliste ajaväljendite tuvastuspiire natukene laiendama, märgendati korpuses 692 364 ajaväljendit, mille abil tuvastati ajaväljendite liikide jagunemine. Suurimat osakaalu ehk üle kolme neljandiku ajaväljendeid olid absoluutsed kalendrilise toimumisajaga ajaväljendid. Absoluutsetel kalendrilise toimumisajaga ajaväljenditel määrati ka granulaarsused. Ülekaalukalt ehk üle kolme neljandiku ajaväljendeid oli aasta granulaarsusega. Üpris suure osa, umbes ühe viiendiku, moodustasid ka kuupäeva granulaarsusega ajaväljendeid. Aastaarvuliste ajaväljendite jagunemine näitas veelgi suuremat jagunemist just lähisajanditesse. Ligikaudu ühe neljandiku puhul oli tegemist ka juba aastatega 21. sajandist, mis andis aimu Vikipeedia järjest uuenevast olemusest ja sellest, kuidas infot ja sündmusi, mida biograafiates kirjeldatakse, leidub järjest enam just lähiaastatest. Aastad 19–21. sajandist moodustasid kogu aastaarvulistest ajaväljenditest peaaegu 90%. Näha oli ka, et palju enam leidus kõrgema ajaväljendite rikkalikkusega ehk ajaväljendite arvuga artikleid just lähisajanditest sünniaastate poolest.

Töö raames pakkus autor välja meetodi kahe artikli põhjal sarnasuse leidmiseks ainult ajaväljendite alusel. Meetod kasutab baasina Omar Alonso jt [2] poolt pakutud raamistikku kauguse leidmisel. Sisse on viidud ka mitmed muudatused. Meetod kombineerib kauguse arvutamise aastaarvuliste ajaväljendite, aastast ja kuust koosnevate ajaväljendite, aastast ja kuust

ja kuupäevast koosnevate ajaväljendite, ajavahemike, mille otspunktid on absoluutsed ajaväljendid, vahel. Meetodit testiti erinevate võrdluskorpuste abil, leides mitmeid näited, kus ajaväljendite kauguse poolest lähedasemad artiklid on seotud ka sündmuste ja ka artiklite ülesehituse ja ajaväljendite olemuse poolest. Meetodile anti esialgne hinnang seotud artiklite leidmise nurga alt. 70 juhuslikult valitud artikli ja nende järgi määratud 350 sarnase artikli põhjal võis autori hinnangul väita, et artiklis kujutatud isikud võib seotuks lugeda täpsusega 0,18. Tulemuste põhjal võib väita, et otseselt seotud artiklite määramiseks kõigi artiklite puhul antud meetodist, mis põhineb ainult ajaväljenditel, vaadeldud korpuse puhul ei piisa. Kuna artiklite hulk on küllaltki suur ja artikleid on väga erineva sisuga ja ajaväljendite poolest peamiselt aastaarvulistest ajaväljenditest koosnevad, siis juhuslikke kokkulangevuste hulk on suurem kui selgetest seostest tingitud kokkulangevuste hulk.

Tööl on autori arvates mitmeid edasiarenduse võimalusi. Töö käigus selgus, et valdav enamus ajaväljendeid Vikipeedia biograafilistes artiklites on aastalise granulaarsusega. Võrdlemise seisukohast on see küll hea, sest tõenäosus, et aastaarvuliste ajaväljendite juures on kokkulangevusi palju rohkem, kui näiteks kuupäevalise granulaarsusega ajaväljendite puhul. Samas aga on palju kõrgem nende ajaväljendite juhuslik kokkulangevus ja mitte kokkulangevused sellest, et artiklites kujutatud isikud oleksid kuidagi seotud. Küll aga oli selgeid seoseid märgata, kui ajaväljendid olid seotud kindlate suuremate sündmustega nagu sõjad, näiteks Eesti vabadussõda või esimene maailmasõda. Seega oleks üheks edasiarenduse suunaks uurida ja kaardistada Vikipeedia biograafilistes artiklites enam esinevad suuremad sündmused ja võtta võrdlemisaluseks ajaväljendid, mis on seotud nende sündmustega. Teiseks edasiarenduse võimaluseks oleks uurida ja kaardistada Vikipeedia biograafilistes artiklites avalduvad elukutsed ja ametid. Seejärel, sidudes elukutsed ja ametid ajaväljenditega, oleks võimalik uurida, kuidas avalduvad seosed näiteks samal ajaperioodil tegutsenud muusikute või korvpallurite või poliitikute vahel. Kolmandaks võimaluseks oleks uurida ja kaardistada Vikipeedia biograafilistes artiklites esinevad asukohad. Sidudes esinevad ajaväljendid asukohtadega, kus isikud ühel või teasel ajahetkel viibisid, aitaks kõvasti vähendada näiteks aastaarvuliste ajaväljendite kokkulangemise juhuslikkust ja läbi selle parandada seotud artiklite leidmist. Lisaks antud töö edukas uurimus Vikipeedia biograafiliste artiklitest võib olla eeskujuks edasistele uurimustele, mis võtavad aluseks just eestikeelse Vikipeedia artiklid.

Kasutatud kirjandus

[1] Stern S. Calendars in Antiquity: Empires, States, and Societis. Oxford: Oxford University Press. 2012.

[2] Alonso O., Gertz M., Baeza-Yates R. Temporal Analysis of Document Collections: Framework and Applications. *International Symposium on String Processing and Information Retrieval*. Springer-Verlag, 2010, pp 290-296.

[3] Lih A. The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia. New York: Hyperion. 2009.

[4] Projekt Miljon+ veebileht. http://www.miljonpluss.ut.ee/ (Vaadatud 23.03.2017)

[5] Graham P. „An Encyclopedia, Not an Experiment in Democracy": Wikipedia Biographies, Authorship, and the Wikipedia Subject. *Biography*, 2015, nr 38 (2), pp 222-244.

[6] Ofek N., Rokach L. A Classifier to Determine Which Wikipedia Biographies Will Be Accepted. *Journal of the Association for Information Science and Technology*, 2015, nr 66 (1), pp 213-218.

[7] Milne D., Witten I. An Open-Source Toolkit for Mining Wikipedia. *Artificial Intelligence*, 2013, nr 194, pp 222-239.

[8] Reznik I., Shatalov V. Hidden Revolution of human priorities: An analysis of biographical data from Wikipedia. *Journal of Informetrics*, 2016, nr 10, pp 124-131.

[9] Orasmaa S., Petmanson T., Tkachenko A., Laur S., Kaalep H.-J. EstNLTK – NLP Toolkit for Estonian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, http://www.lrec-conf.org/proceedings/lrec2016/summaries/332.html (Vaadatud 04.04.2017)

[10] EstNLTK Vikipeedia liides. https://estnltk.github.io/estnltk/1.4/tutorials/wikipedia.html (Vaadatud 03.03.2017)

[11] Orasmaa S. Automaatne ajaväljendite tuvastamine eestikeelsetes tekstides. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 2012, nr 8, lk 153-169.

[12] Orasmaa S. Ajaväljendite tuvastaja märgendusformaat. https://github.com/soras/Ajavt/blob/master/doc/margendusformaat_et.pdf?raw=true (Vaadatud 06.04.2017)

[13] Cha S.-H., S. N. Srihari. On measuring distance between histograms. *Pattern Recognition,* 2002, nr 35 (6), pp 1355-1370.

[14] Odijk D., Garbacea C., Schoegje T., Hollink L., de Boer V., Ribbens K., van Ossenbruggen J. Supporting Exploration of Historical Perspectives Across Collections. *Research and Advanced Technology for Digital Libraries - 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Proceeding*. Springer-Verlag, 2015, pp 238-251.

[15] Hopper T. Graphing Highly Skewed Data. 2010, https://tomhopper.me/2010/08/30/graphing-highly-skewed-data/ (Vaadatud 12.02.2017)

[16] Orasmaa S. Loomuliku keele tekstide automaatne ajasemantiline analüüs (*stendiettekanne*). Eesti-uuringute interdistsiplinaarsed dialoogid, 2016.

[17] Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press. 2008. pp 158-163, https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html (Vaadatud 08.05.2017)

Lisad

I. Töö käigus moodustatud skriptid, statistika ja graafikud ja korpused.

Töö jaoks moodustatud skriptid koos kasutusjuhenditega ja töö käigus moodustatud ja esitatud statistika ja graafikud, mis on loodud Vikipeedia biograafiliste artiklitest 24.11.2016 seisuga, on kättesaadavad aadressilt: https://github.com/jteppo/ajavVikBioArt (Vaadatud 11.05.2017).

Et oleks võimalik täpselt töö tulemusi korrata, on ka töö jaoks moodustatud skriptid koos kasutusjuhentidega ja töö käigus moodustatud ja esitatud statistika, graafikud ning ka töö käigus loodud korpused Vikipeedia biograafiliste artiklitega ja märgendatud ajaväljenditega, mille põhjal loodi statistika ja võrdluskorpused ning ka kokkupakitud kogu 24.11.2016 seisuga Vikipeedia artiklite andmefail kokkupakitult kättesaadavad aadressilt: https://www.dropbox.com/s/sfi10vjcro2ed1f/ajavVikBioArt2.zip?dl=0 (Vaadatud 10.05.2017).

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Jaan Teppo**,
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Ajaväljendid Vikipeedia biograafilistes artiklites,
(*lõputöö pealkiri*)

mille juhendaja on Siim Orasmaa,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **11.05.2017**