TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Liina Heinluht 182875IABM

VÕTMESÕNADE EKSTRAHEERIMINE EESTIKEELSEST TRANSKRIBEERITUD TEKSTIST

Magistritöö

Juhendaja: Ahti Lohk
PhD

Tallinn 2021

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Liina Heinluht

10.05.2021

Annotatsioon

Käesoleva magistritöö teemaks on võtmesõnade ekstraheerimine. Töö peamiseks eesmärgiks on selgitada välja parim võtmesõnade ekstraheerimise juhendamata õppe meetod eestikeelsetele transkribeeritud tekstidele. Töös kasutatakse „Kõnesalvestuse brauseri“ veebirakenduses asuvate helisalvestiste transkribeeritud tekstide kogumikke.

Eksperimendi raames rakendatakse valitud eestikeelsete raadiosalvestiste transkribeeritud tekstidele nelja erinevat võtmesõnade ekstraheerimise juhendamata õppe meetodit: TF-IDF, TextRank, RAKE ja YAKE!. Meetodite TF-IDF, TextRank ja RAKE algoritmide analüüs teostatakse programmeerimiskeele R jaoks loodud paketi UDPipe abil arenduskeskkonnas RStudio. YAKE! eesti keelele kohandatud Pythoni algoritmi rakendatakse aga Google Colaboratory’s.

Meetodeid võrreldakse täpsuse, saagise ja F_1-skoori abil. Töö tulemusena valmib analüüs, mille põhjal valitakse välja sobivaim meetod eestikeelsetele transkribeeritud tekstidele automaatselt kõige relevantsemate märksõnade määramiseks.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 38 leheküljel, 5 peatükki, 13 joonist, 8 tabelit.

Abstract

Keyword extraction from Estonian transcribed text

The topic of this master's thesis is keyword extraction. The main goal of the thesis is to find out the best unsupervised keyword extraction method for Estonian transcribed texts. The thesis uses collections of transcribed texts of sound recordings from "Kõnesalvestuse brauser" web application.

In an experiment, four different unsupervised keyword extraction methods TF-IDF, TextRank, RAKE and YAKE! are applied to the transcribed texts of selected Estonian radio recordings. The analysis of TF-IDF, TextRank and RAKE algorithms is performed using the UDPipe package created for the programming language R in the development environment RStudio. YAKE!'s Python algorithm adapted to the Estonian language is used in Google Colaboratory.

Methods are compared with precision, recall and F_1-score. As a result an analysis is completed, on the basis of which the most suitable method for automatically determining the most relevant keywords for Estonian transcribed texts is selected.

The thesis is written in Estonian and contains 38 pages of text, 5 chapters, 13 figures, 8 tables.

Lühendite ja mõistete sõnastik

ASR	*Automatic Speech Recognition* ehk automaatne kõnetuvastus on tehnoloogia, mille abil leitakse automaatselt sõnad ja laused, mis kõige paremini vastavad sisendiks olevale inimkõnele.
FN	*False Negative*, valenegatiivne tulemus, kus mudel ennustab negatiivse klassi valesti.
FP	*False Positive*, valepositiivne tulemus, kus mudel ennustab positiivse klassi valesti.
Klasterdamine	Objektide omavaheline grupeerimine sarnasuse alusel kasutades ainult objektide omadusi.
NLTK	*Natural Language ToolKit* on platvorm Pythoni programmide loomiseks inimkeele andmetega töötlemiseks.
Podcast	Taskuhääling on standard, mis võimaldab taskuhäälingusaadete tellija seadmes jooksval programmil märgata lisandunud saateid ning neid automaatselt isiklikku seadmesse alla laadida.
POS	*Part-of-speech*, sõnaliigid nagu nimisõna, omadussõna, tegusõna jne.
Parsimine	Protsess, milles arvutikeeles või andmestruktuurides esinevaid sõnesid analüüsitakse vastavalt loomuliku keele formaalse grammatika reeglitele.
Partikkel	Kaas- ja sidesõnade (ning hüüd- ja määrsõnade) ühisnimetus, abisõna.
Pipeline	Konveier ehk järjestikku ühendatud andmetöötluselementide komplekt, kusjuures ühe elemendi väljund on järgmise sisendiks.
Rekursiivne	Millegi kordamine viitega iseendale või enesesarnaselt.
Supervised method	Juhendatud õppega lähenemisviis, kus teisendatakse sisendandmed soovitud väljundandmeteks. Ette on antud märgistatud treeninguandmete kogu, mille põhjal tehakse uusi järeldusi.
Stopword	Stoppsõna on sõna, mis ei anna teksti sisu edasi (nt sidesõnad).
Teek	Funktsioonide, makrode, klasside, moodulite vms komponentide kogu, mida saab programmis vajadust mööda kasutada.

TN	*True Negative*, õige negatiivne tulemus, kus mudel ennustab negatiivse klassi õigesti.
TP	*True Positive*, õige positiivne tulemus, kus mudel ennustab positiivse klassi õigesti.
Transkribeerimine	Heli- või videosalvestise esitamine kirjalikus vormis.
TalTech	Tallinna Tehnikaülikool
Treebank	Puudepank on keele märkide vaheliste seostega märgendatud tekstikorpus.
Unsupervised method	Juhendamata õppega lähenemisviis, kus proovitakse leida märgistamata andmete kogust uusi struktuure.
Võtmesõna	Oluline märksõna ehk sisule viitav tekstisõna, millega saab teha päringu.
Võtmesõnade ekstraheerimine	Selliste sõnade ja fraaside eraldamine tekstist, mis vaadeldava teksti sisu kõige paremini edasi annavad.
Üneem	Pausi täitev häälitsus nt *mmm*, *äää*.
WER	*Word Error Rate*, sõnade veamäär, mis näitab valesti tuvastatud sõnade osakaalu tekstis.

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Digitaalsete andmete hulk meie ümber on viimaste kümnendite jooksul eksponentsiaalselt kasvanud. Üha suurenevas trendis on ka transkribeeritud ehk heli- või videosalvestise kirjalikus vormis esitatavate andmete hulk. Kliendikõnede teksti kujule viimine ning koosolekute protokollid on kaks kõige levinumat viisi transkribeerimise kasutamisest äriettevõtetes, kuid kasutusvaldkond on veelgi laiem.

Kuigi transkribeeritud tekstist on vajaminevat teavet küll oluliselt lihtsam leida kui heli- või videosalvestisest, on nende andmete struktureerimine ning korrastamine suureks väljakutseks. Sageli välja toodud statistika: „80% andmetest on struktureerimata“, viitab kasutamata tekstandmete ressursside suurele hulgale [1].

Andmete üheks korrastamise viisiks on võtmesõnade märkimine. Võtmesõnade, sealhulgas ka võtmefraaside ekstraheerimine aitab andmeid struktureerida ning teabe hankimise kiirust märkimisväärselt parandada, aidates inimestel saada tekstist kiiret ning täpset esmast teavet. Oskuslikul lähenemisel saab neid omakorda kasutada meelestatuse hindamisel, soovitussüsteemides ning tekstidest kokkuvõtete tegemisel. Seda kõike on võimalik ärihuvides ära kasutada.

Võtmesõnade ekstraheerimine on maailmas laialdaselt levinud ning nende tuvastamiseks on kirjutatud hulgaliselt algoritme. Autorile teadaolevalt ei ole võtmesõnade ekstraheerimist eestikeelsetele transkribeeritud tekstidele varasemates teadustöödes veel käsitletud. Käesolev magistritöö viib läbi uurimuse ja analüüsi, mille tulemusena valitakse sobivaim meetod, mille abil eestikeelsetest transkribeeritud tekstidest automaatselt kõige relevantsemaid märksõnu määrata.

1.1 Teema relevantsus

Mis tahes andmete kokkuvõtetel on tähtis roll olulise informatsiooni väljatoomisel ning ebaolulise teabe ignoreerimisel. Pikast tekstist välja tõmmatud võtmesõnad aitavad teksti

lugemisele aega kulutamata selle põhiideed olulisel määral lahti mõtestada ning seeläbi infootsija vaeva vähendada. See säästab palju aega nii äris, hariduses kui igapäevaelus.

Üheks keerulisemaks valdkonnaks traditsiooniliste kirjalike tekstide kõrval on automaatse kõnetuvastustehnoloogia ehk ASR (ingl *Automatic Speech Recognition*) abil transkribeeritud tekstidest võtmesõnade leidmine. Esiteks seetõttu, et spontaanne kõne erineb traditsioonilisest kirjalikust tekstist. Hästi vormistatud ning loogiliselt kulgevate lausete asemel koosneb transkribeeritud tekst sageli vormimata, grammatiliselt ebakorrektsetest ning mitteametlikest sõnadest ja lausetest. Lisaks lahjendavad kõnelejad olulist teavet sageli pause tehes, üksteist katkestades ja niisama vesteldes. Samuti süstivad ASR-ist tulenevad vead transkriptsioonidesse lisamüra [2].

Huvi ASR vastu on maailmas olnud juba aastakümneid. Laialdasemalt hakkas see levima 1980. aastatel [3]. Eestis on kõnetuvastusele pööratud suuremat tähelepanu alates 2004. aastast, kui E. Meister [4] algatas teadus- ja arendusprojekti „Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine“. Sellest ajast alates on eestikeelset kõnetuvastust teadusprojektide raames jõudsalt arendatud.

Tänasel päeval leiab ASR Eestis järjest laialdasemat kasutamist ka väljaspool teadusprojekte. ASR on kasutusel dokumentide dikteerimisel, kõne- ja videosalvestuste transkribeerimisel ning kõne abil arvutite ja seadmetega suhtlemisel [5].

Rahvusvahelise turundusuuringute firma MarketsandMarkets’i poolt 2019. aastal läbi viidud uuringu kohaselt on kõnetuvastuse tööstus 2022. aasta lõpuks enam kui kolmekordistunud [6]. Ennustatakse, et peagi võib iga ettevõte ja asutus vajada kõnetuvastustehnoloogiat. Kuigi ekspertide sõnul nõuab heli- ja videosisuga töötamine endiselt inimese sekkumist, kiirendab transkribeerimine vajalike ülesannete täitmist ja suurendab tööviljakust [7]. Lisades siia ka automaatselt ekstraheeritud märksõnad, saab personal keskenduda andmete analüüsimisele ning arendustele. Käsitsi märksõnade lisamine ei oleks siinjuures kuigi tõhus, sest nõuaks liialt inimressurssi, kes süveneks tekstidesse ning määraks loetu põhjal olulisemad sisu kajastavad võtmesõnad. Samuti lisanduksid ka subjektiivsuse probleemid.

Käesolev magistritöö tugineb eelnevalt viidatud probleemidele ning püstitab järgnevad uurimisküsimused:

- Kui lihtne on maailmas tuntud ning ingliskeelsetes teaduslikes artiklites viidatud võtmesõnade ekstraheerimise meetodeid rakendada eestikeelsetele transkribeeritud tekstidele?

- Kas transkribeeritud teksti veamäär ehk WER (ingl *Word Error Rate*) mõjutab võtmesõnade ekstraheerimisel tulemusi?

- Kui häid tulemusi näitavad viimaste aastate teaduslikes artiklites viidatud märksõnade ekstraheerimise meetodid eestikeelsetel transkribeeritud tekstidel? Milline on vaadeldud meetoditest parim?

Läbi nende uurimusküsimuste soovitakse saada vastus põhiküsimusele – kuidas eestikeelsetest transkribeeritud tekstidest automaatselt kõige relevantsemaid võtmesõnu ekstraheerida.

1.2 Eesmärgid

Käesoleva magistritöö peamiseks eesmärgiks on välja selgitada parim võtmesõnade ekstraheerimise meetod eestikeelsetele transkribeeritud tekstidele.

Selle eesmärgi saavutamiseks:

- Kaardistatakse kitsaskohad võtmesõnade ekstraheerimisel.

- Kaardistatakse transkribeeritud tekstide eripärad ning kitsaskohad, mis raskendavad võtmefraaside ekstraheerimist.

- Realiseeritakse ning rakendatakse transkribeeritud tekstidel erinevaid võtmesõnade ekstraheerimise juhendamata õppe meetodeid.

- Analüüsitakse ning hinnatakse realiseeritud meetodite efektiivsust ning kasutamise lihtsust eestikeelsete transkribeeritud tekstide peal. Analüüsi tulemusena selgitatakse valitud meetodite seast välja kõige efektiivsem meetod.

1.3 Ülevaade tööst

Eelnevalt püstitatud eesmärkideni jõudmiseks töötatakse magistritöö käigus läbi mitmed varasemalt publitseeritud artiklid antud valdkonnas. Vaadeldakse varasemaid uurimusi

(lõpuööd, magistritööd, doktoritööd) ja muid uuringuid ning praktikat, mida antud valdkonnas Eestis ning rahvusvahelisel tasandil läbi on viidud. Viimase paari aasta jooksul ilmunud teaduslikele artiklitele tuginedes selgitatakse välja teema relevantsus ning kaasnevad kitsaskohad, valitakse välja parimad rahvusvaheliselt tuntud võtmefraaside ekstraheerimise juhendamata õppe meetodid, rakendatakse neid eestikeelsete transkribeeritud tekstide peal ning valideeritakse saadud tulemused.

Magistritöö teoreetilises osas tutvustatakse võtmesõnade ekstraheerimise olemust ning selgitatakse praktilise osa jaoks väljavalitud meetodite tööpõhimõtteid. Antakse lühiülevaade transkribeerimisest ning selle eripäradest ja kitsaskohtadest.

Töö praktiline osa hõlmab töös kasutatavate tekstitöötlusvahendite tutvustust ja lähteandmete ettevalmistamist. Selgitatakse manuaalse ning automaatse võtmesõnade määramise protsesse ja valideerimise metoodikaid, viiakse läbi eksperimendid ning valideeritakse saavutatud tulemused.

Parim võtmesõnade ekstraheerimise meetod selgub kasutajate poolt manuaalselt määratud võtmesõnade ning automaatselt erinevate meetodite poolt pakutud märksõnade võrdlemisel. Lisaks valmib töö lõpus meetodeid võrdlev analüüs.

2 Teoreetiline raamistik ja seotud meetodid

Rahvusvahelisel tasandil on transkribeerimist ja võtmesõnade ekstraheerimist käsitletud nii eraldiseisvalt kui ka ühiselt mitmetes teaduslikes artiklites. Kõige sarnasemalt antud magistritööle on käsitlevad mõlemat teemat korraga A. Désilets, B. de Bruijn, ja J. Martin [8] ning F. Liu, D. Pennell, F. Liu ja Y. Liu [9].

Désilets, de Bruijn ja Martin uurivad oma töös [8] võtmefraaside (tavaliselt 1–3 järjestikuse sõna jada) ekstraheerimist suulistest helidokumentidest. Nad toovad välja seisukoha, et võtmefraaside ekstraheerimine on lihtsam ülesanne kui täisteksti transkriptsioon ning võtmefraase saab mõistliku täpsusega ära määrata ka transkribeeritud dokumentidest, mille sõnade WER on kuni 62%. Arvatakse, et võtmesõnade WER on tavaliselt madalam, kui mitte-võtmesõnade WER, ning neid esineb helidokumendis tihti üleliigselt [8].

F. Liu, Pennell, F. Liu ja Y. Liu uurivad oma töös [9] koosolekute transkribeeritud tekstidele võtmesõnade märkimist. Nad kasutavad oma uuringus ühte kõige levinumat juhendamata õppe baasmeetodit, mis näitab uurimuses teiste meetodite kõrval konkurentsivõimelisi tulemusi.

Võtmesõnade ekstraheerimist eestikeelsetest kirjalikest vabatekstidest on uuritud [10], kuid töös ei ole kasutatud varasemalt valmis loodud meetodeid. Võtmesõnade leidmist transkribeeritud tekstidest ei ole magistritöö autorile teadaolevalt hetkel veel käsitletud. T. Ennomäe on oma magistritöö [11] käigus uurinud telefonivestluste transkribeeritud tekste vaid meelestatuse hindamisel.

Eesti keele kõnetehnoloogiliste uuringute ja arendustöödega tegeleb valdavalt Tallinna Tehnikaülikooli (TalTech) Keeletehnoloogia laboratoorium. Laboratooriumi veebilehelt[1] on leitavad mitmed viimaste aastate uurimistööd, mis näitavad, et transkribeerimine on aktuaalne ning pidevas arengus.

[1] https://www.taltech.ee/en/laboratory-language-technology

2.1 Võtmesõnade ekstraheerimine

Võtmesõnade ekstraheerimine on tekstiline infotöötlusülesanne, mis hõlmab esinduslike ja iseloomulike sõnade automaatset eraldamist dokumendist, mis väljendavad selle sisu kõiki põhiaspekte. Võtmesõnad moodustavad tekstist lühikese kontseptuaalse kokkuvõtte, mis on kasulikud digitaalsetes infohaldussüsteemides semantilise indekseerimise, infootsingu, dokumentide klasterdamise ja klassifitseerimise jaoks [12].

Võtmesõnadena vaadeldakse antud töö raames nii üksikuid märksõnu kui ka kuni kolmest sõnast koosnevaid märksõnade fraase. Fraaside eelis võtmesõnade ees on see, et need pakuvad dokumendis kajastatud teemade täpsemaid kirjeldusi. Näiteks on fraas „geneetiline algoritm“ täpsem, kui sõnad „geneetiline“ või „algoritm“ või isegi märksõnade komplekt „geneetiline, algoritm“ (viimane sõna neist võiks võrdselt viidata nii geneetilistele algoritmidele kui ka genoomsete andmete kaevandamise algoritmile) [8].

Märksõnade ekstraheerimise meetodid võib jagada kahte liiki lähenemiseks:

- juhendatud (ingl *supervised method*) õppe meetodid;
- juhendamata (ingl *unsupervised method*) õppe meetodid.

Juhendatud õppe meetodite puhul teisendatakse etteantud sisendile õige väljund, toetudes seejuures eelnevalt käsitsi sisestatud suurtele andmekogudele. Seega vajab juhendatud õpe treenimisprotsessi ja nõuab keele olemuse kajastamiseks käsitsi märgendatud dokumendikogusid.

Juhendamata õpe tähendab aga seda, et süsteem õpib ise leidma etteantud sisendi andmetest õiget väljundit. Juhendamata õppel baseeruvate meetodite puhul ei sea keelte eripäradest tulenevad erisused meetodite rakendamisel suuri erinevusi [13].

Mitmetele, magistritöö autori poolt läbi töötatud, teaduslikele töödele ning artiklitele [8], [9], [12]– [14] tuginedes võetakse antud magistritöö raames võrdlusesse vaid juhendamata õppel baseeruvad meetodid. Seda eelkõige seetõttu, et uurimuste põhjal võib järeldada, et need meetodid ei sea eestikeelsetele tekstidele rakendamisel endas suuri piiranguid ning annavad seejuures ka relevantseid tulemusi.

2.2 Juhendamata õppel baseeruvad võtmesõnade ekstraheerimise meetodid

E. Papagiannopoulou ja G. Tsoumakas toovad oma uurimuses [12] välja, et juhendamata õppel baseeruvad võtmesõnade ekstraheerimise meetodid on populaarsed, kuna on domeenist ehk valdkonnast (nt veebiuudised, teaduspublikatsioonid, foorumid, loengumaterjalid) sõltumatud ning ei vaja märgendatud treeningandmeid ehk käsitsi ekstraheeritud märksõnu, millega kaasnevad subjektiivsuse probleemid ning märkimisväärsed investeeringud ajas ja rahas.

Juhendamata õppe meetodid jagunevad võtmesõnade ekstraheerimisel [15]:

- statistilised;
- graafipõhised;
- kontekstuaalse sarnasuse põhised;
- keelemudelite põhised.

E. Papagiannopoulou ja G. Tsoumakas [12] on teinud ülevaate erinevatest juhendamata õppe meetoditest. Tööst on võimalik lugeda, et kõige levinumad on statistilised ning graafipõhised meetodid.

Juhendamata meetodite põhisammud võtmesõnade ekstraheerimisel [15]:

1. kandidaatvõtmesõnade valimine mõne heuristika põhjal;
2. valitud kandidaatvõtmesõnadele kaalude/skooride leidmine;
3. mitmesõnaliste võtmesõnade moodustamine;
4. võtmesõnade järjestamine.

Järgnevalt on välja toodud juhendamata õppe kaks kõige levinumat baasmeetodit ja kaks viimase paari aasta teaduslike artiklite [12], [13], [16], [17], [18], [19] võrdlustes olnud ning häid tulemusi näidanud meetodit. Lisaks uuriti neid meetodeid süvitsi, kas need võiksid anda häid tulemusi ka eestikeelsetele transkribeeritud tekstidele rakendades.

2.2.1 TF-IDF

TF-IDF (ingl *Term Frequency - Inverse Document Frequency*) on juhendamata õppe lähenemisviiside puhul üks kõige levinumaid baasmeetodeid. Statistiline meetod leiab igale sõnale kaalu selle lokaalset (dokumendi) ja globaalset (dokumentide kogum) mõju arvestades. Kuigi meetodi rakendamine on üsna lihtne, nõuab see juurdepääsu suurele tekstikorpusele, mis ei pruugi alati saadaval olla [13].

TF-IDF meetodi abil arvutatakse ja järjestatakse märksõnad ning fraasid valemiga

$$TfIdf = Tf \times Idf,$$

kus *Tf* on termini esinemissagedus antud dokumendis. *Idf* aga määrab termini globaalse esinemissageduse, mis arvutatakse valemiga

$$Idf = log_2 \frac{N}{1 + |d \in D: phrase \in d|},$$

kus *N* on dokumentide arv tekstikorpuses *D* ja $|d \in D: phrase \in d|$ on dokumentide arv, milles antud termin esineb [12].

Sellest valemist võib järeldada, et kui termin esineb konkreetses dokumendis palju kordi, kuid kogu tekstikorpuse dokumentides seda terminit ei esine või esineb harva, siis on sellel sõnal antud dokumendis suur kaal ehk oluline märksõna. Kui termin esineb kõikides dokumentides, siis tema väärtus konkreetses dokumendis on väga madal [15].

TF-IDF meetodit on kasutatud võrdlusmeetodina pea kõikides magistritöö autori poolt läbi töötatud teaduslikes artiklites. Lisaks on TF-IDF meetodit edukalt kasutatud uurimistöös [9], kus selle abil määrati võtmesõnu koosolekute transkribeeritud tekstidele. Eelnevalt mainitud uurimistöö lubab käesoleva töö autoril eeldada, et TF-IDF meetodit on sobilik kasutada ka antud magistritöö meetodite võrdluses.

2.2.2 TextRank

TextRank on esimene graafipõhine võtmesõnade ekstraheerimise meetod, mis pakuti välja Mihalcea ja Tarau (2004) poolt [12]. TextRanki võib samuti pidada juhendamata õppe lähenemisviiside üheks baasmeetodiks.

Graafipõhine järjestusalgoritm on sisuliselt viis graafi tipu olulisuse otsustamiseks, tuginedes kogu graafilt rekursiivselt saadud globaalsele teabele. Graafipõhise

järjestusmudeli põhiidee on „hääletamine“ või „soovitus“. Kui üks graafi tipp lingib teisele, annab see põhimõtteliselt hääle selle teise tipu poolt. Mida suurem on tipule antud häälte arv, seda suurem on tipu tähtsus. Veelgi enam, hääletava tipu tähtsus määrab, kui oluline on hääl ise ning seda teavet võtab arvesse ka järjestusmudel. Seega määratakse tipuga seotud skoor sellele antud häälte ja neid hääli andnud tippude skoori põhjal [20].

Formaalselt, olgu $G = (V, E)$ suunatud graaf koos tippudega V ja servade E kogumiga, kus E on $V \times V$ alamhulk. Konkreetse tipu V_i puhul olgu $In(V_i)$ tippude hulk, mis sellele osutavad (eelkäijad) ja olgu $Out(V_i)$ tippude hulk, millele tipp V_i osutab (järeltulijad). Tipu V_i skoor on määratletud valemiga:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j),$$

kus d on summutustegur, mille saab seada vahemikku 0 kuni 1 ning mille roll on mudelisse integreerida graafis antud tipust teise juhusliku tipuni hüppamise tõenäosus [20].

TextRank koostab graafi, kus sõlmed on terminid, mis koosnevad kindlaksmääratud sõnaliikidest ja servad ehk ühendused on sõlmede vahel, mis esinevad kindlaksmääratud aknas [13]. Akna suurus näitab, kui kaugel koosesinevad sõnad üksteisest võivad asetseda. Akna suurus võib olla 2 kuni 10 sõna. Arvatakse, et mida suurem on aken, seda väiksem on ekstraheeritavate sõnade täpsus, mis on seletatav sellega, et kaugemal asetsevate sõnade suhe pole piisavalt tugev, et neist tekstis seos määratleda [20].

Kuna käesoleva töö autorile teadaolevalt ei ole võtmesõnade ekstraheerimist eestikeelsetele transkribeeritud tekstidele varasemalt teadustöödes käsitletud, siis seetõttu on sooviks kasutada vähemalt kahte levinumat erineva lähenemisviisiga baasmeetodit. TextRank sai lisaks TF-IDF meetodile valitud võrdlusesse, sest seda peetakse graafipõhiste lähenemisviiside üheks alusmeetodiks.

2.2.3 RAKE

S. Rose, D. Engel, N. Cramer ja W. Cowley pakkusid 2010. aastal välja graafipõhise meetodi nimega RAKE (ingl *Rapid Automatic Keyword Extraction*). RAKE on juhendamata õppe lähenemisviisiga, domeenist ja keelest sõltumatu meetod märksõnade eraldamiseks üksikutest dokumentidest [21].

RAKE peab kandidaatsõnadeks terminite jadasid (piiritletud stoppsõnade (ingl *stopword*) või fraaside eraldajatega) ja ehitab sellest lähtudes terminite koosesinemise maatriksi. Seejärel arvutatakse iga kandidaatsõna jaoks terminiskoor selle liikmesõna skooride summana, lähtudes maatriksis olevate terminite astmest ja sagedusest:

- termini sagedus;
- termini aste (erinevate terminite arv, millega see samaaegselt esineb);
- astme ja sageduse suhe [13].

RAKE piiritleb kandidaatsõnad stoppsõnadega ja seetõttu ei sisalda eraldatud märksõnad sisemisi stoppsõnu. Kuna meetod on tekitanud tugevat huvi tänu oma võimele valida välja väga spetsiifiline terminoloogia, väljendati soovi ka selliste märksõnade tuvastamise osas, mis sisaldavad sisemisi stoppsõnu (näiteks ingl „*axis of evil*“). Selliste märksõnade leidmiseks otsib RAKE märksõnapaare, mis külgnevad üksteisega vähemalt kaks korda samas dokumendis ja samas järjekorras. Seejärel luuakse nende märksõnade ja nende sisemiste stoppsõnade kombinatsioonina uus kandidaatsõna. Uue märksõna skoor on selle liikmemärksõnade skooride summa.

Selliseid märksõnu eraldatakse aga üsna vähe, seetõttu suureneb nende olulisus. Kuna külgnevad märksõnad peavad dokumendis esinema kaks korda samas järjekorras, siis eraldatakse neid rohkem pigem pikematest tekstidest kui lühematest [21].

Kuna RAKE väljatöötamisel oli autorite üheks eesmärgiks, et meetod toimiks hästi just nende dokumentide puhul, kus ei järgita tavapäraseid grammatikareegleid [21], siis peaks olema see meetod väga heaks võrdlusmeetodiks just transkribeeritud tekstide juures.

2.2.4 YAKE!

YAKE! (ingl *Yet Another Keyword Extractor!*) loodi 2018. aastal ning seda kirjeldatakse teaduslikes artiklites meetodina, mida saab lihtsasti rakendada kõikidele keeltele vajamata treenimist soovitud liiki dokumentide kogumi osas. Kirjelduste kohaselt ei sõltu YAKE! keelespetsiifilistest sõnastikest, välistest korpustest, teksti pikkusest, keelest ega domeenist [13].

Oluliste märksõnade tuvastamiseks struktureerimata üksikutest dokumentidest, tuginetakse kohalikele tekstifunktsioonidele ja statistilisele teabele nagu näiteks terminite samaaegne esinemine ja sagedus [13].

YAKE! loojad toovad oma töös välja [13], et süsteem on töökindel ning seda saab hõlpsasti laiendada ulatuslike dokumentide ning kontekstideni. Asjaolu, et see tugineb üksikutele dokumentidele, võimaldab tegutseda suurte andmekogude olemasolust sõltumatult [13].

YAKE! algoritmil on viis peamist sammu [13]:

1. teksti eeltöötlus ja kandidaatsõnade tuvastamine;
2. tunnuste ekstraheerimine;
3. termini skoori arvutamine;
4. *n*-grammi (st *n*-sõnast koosneva fraasi) genereerimine ja kandidaatsõnade skooride arvutamine;
5. korduvate andmete eemaldamine ja paremusjärjestus.

Paremusjärjestus koostatakse märksõna skoori abil. Mida väiksem on märksõna skoor, seda tähtsam ning tähenduslikum on see sõna antud tekstis.

Järgnevalt on kujutatud YAKE! algoritmi voog (Joonis 1) [13].

Algorithm 1 Keyword extraction procedure.

Input: *text*, *w*, *n*, θ, *language*
1: # (**Step 1**) Text pre-processing and candidate term identification
2: *sentences* = split *text* into sentences
3: **for** each *sentence* in *sentences* **do**
4: Pre-process (*sentence*, *language*)
5: **end for**
6: # (**Step 2**) Feature extraction & (**Step 3**) Term score
7: **for** each *term* in *sentences* **do**
8: Feature extraction
9: Compute *term* weight
10: **end for**
11: # (**Step 4**) n-gram generation
12: **for** each *sentence* in *sentences* **do**
13: *chunks* = split *sentence* into chunks
14: **for** each *chunk* in *chunks* **do**
15: Build *n*-gram *candidateKeywords* list
16: **end for**
17: **end for**
18: # (**Step 4**) Candidate keyword score
19: **for** each *candidate* in *candidateKeywords* **do**
20: Compute *candidateKeywords* weight
21: **end for**
22: # (**Step 5**) Data deduplication
23: **for** each *candidate1*, *candidate2* in *candidateKeywords* **do**
24: **if** DistanceSimilarity(*candidate1*, *candidate2*) $> \theta$:
25: Remove *candidate* from *candidateKeywords*
26: **end for**
27: # (**Step 5**) Ranking
28: *Keywords* = sort (*candidateKeywords*) by ascending score
Output: (*Keywords*, *score*)

Joonis 1. YAKE! algoritmi mudeli voog.

YAKE! osutus magistritöö autori poolt valituks, kuna meetod esineb mitmete viimaste aastate teaduslike artiklite [13], [17], [18], [19] meetodite võrdlustes ning on näidanud neis ka arvestatavaid tulemusi. Lisaks on palju kiidetud meetodi lihtsat kasutatavust erinevates keeltes, vajamata muid keeletehnoloogilisi lisateadmisi.

2.3 Kitsaskohad võtmesõnade ekstraheerimisel

K. S. Hasan ja V. Ng toovad oma töös [22] välja, et võtmefraaside ekstraheerimist mõjutavad vähemalt neli tekstiga seotud tegurit:

1. **Pikkus**. Võtmefraaside märkimise raskus suureneb koos dokumendi pikkusega, kuna pikemad dokumendid annavad rohkem kandidaatsõnu (st fraase, mis sobivad võtmefraasideks) [22]. Hasan ja Ng järeldavad [22], et keerulisem on võtmefraase märkida teaduslikele artiklitele, tehnilistele aruannetele ja koosolekute transkriptsioonidele, kui kokkuvõtetele, e-kirjadele ja uudiste artiklitele.

E. Papagiannopoulou jõuab oma doktoritöös [14] järeldusele, et graafipõhised meetodid töötavad paremini lühikestel ning statistilised meetodid pikkadel dokumentidel.

2. **Struktuuriline ülesehitus**. Struktureeritud dokumendis on kindlad asukohad, kus võtmefraas kõige tõenäolisemalt ilmub. Näiteks peaks enamik teadusartiklite võtmefraase ilmuma lühitutvustuses ja sissejuhatuses. Kuigi struktuurset teavet on kasutatud teadusartiklitest (nt pealkiri, jaotisteave), veebilehtedest (nt metaandmed) ja vestlustest (nt dialoogiaktid) märksõnade väljavõtmiseks, on see kõige kasulikum, kui dokumentidel on struktuurne ülesehitus. Seetõttu on kõige lihtsam võtmefraase välja võtta teadusartiklitest ja tehnilistest aruannetest nende standardvormingu tõttu (st standardsed jaotised nagu lühitutvustus, sissejuhatus, järeldus jne). Seevastu struktuuri puudumine muud tüüpi struktureeritud dokumentides (nt veebilehtedel, mis võivad olla ajaveebid, foorumid või ülevaated) võib muuta struktuurilise teabe vähem kasulikuks [22].

3. **Teema muutus.** Teadusartiklite ja uudiste artiklite võtmefraaside väljavõtmisel kasutatakse tavaliselt tähelepanekut, et märksõnad ilmuvad mitte ainult dokumendi alguses, vaid ka lõpus. See tähelepanek ei kehti aga vestlustekstide (nt koosolekud, vestlused) puhul. Põhjus on lihtne: vestluses muutuvad teemad vestluse kulgemise jooksul ja nii muutuvad ka teemaga seotud märksõnad. Üks viis selle komplikatsiooni kõrvaldamiseks on vestluse tekstis teemavahetuse tuvastamine. Muutuste tuvastamine pole aga alati lihtne: kui ametlike koosolekute protokollide alguses on ära loetletud päevakorda tulevad teemad, siis juhuslikes vestlustes selliseid vihjeid ei ole [22].

4. **Teema korrelatsioon**. Üheks tähelepanekuks fraaside väljavõtmisel teadusartiklitest ja uudiste artiklitest on see, et dokumendi märksõnad on tavaliselt omavahel seotud. Kuid see tähelepanek ei kehti tingimata mitteametlike tekstide (nt e-kirjad, vestlused, mitteametlikud koosolekud, isiklikud ajaveebid) puhul, kus inimesed saavad rääkida mis tahes hulgast potentsiaalselt mitteseotud teemadel. Seostamata teemade olemasolu viitab sellele, et seoses olevaid teemasid ei ole võimalik kasutada ja seetõttu suureneb võtmefraaside väljavõtmise keerukus [22].

2.4 Transkribeerimine

Transkribeerimine tähendab kõnetuvastustehnoloogia abil heli- või videosalvestise esitamist kirjalikus vormis [23]. Kõnetuvastustehnoloogia leiab audiomaterjalist automaatselt sõnad ja laused, mis vastavad kõige paremini sisendiks olevale inimkõnele [24] ning esitab need kirjalikus vormis.

Üha enam ettevõtteid kasutab transkribeerimist kliendikõnede talletamiseks ning analüüsimiseks. Igapäevasemalt transkribeeritakse veel intervjuusid, koosolekute, seminaride, konverentside või loengute salvestisi, internetti üles laaditud videomaterjale, audiovestluseid, internetipõhiseid kursuseid, aga ka kohtuistungeid või arstide diktofonimärkmeid [23].

Heli- ning videosalvestiste transkribeerimine muudab nende sisu lihtsasti kättesaadavaks uutele sihtgruppidele ning paremini töödeldavaks erinevatele programmidele. Näiteks muutub audiomaterjalide sisu kättesaadavaks otsingumootoritele, mis üldjuhul ei suuda tuvastada audiofailide sisu [23].

Ühe uue sihtgrupina võib välja tuua kuulmisvaegusega inimesed, kes võivad loodud audiosisu mõista valesti või jääb see neile tugeva kuulmislanguse tõttu sootuks kättesaamatuks. Samuti luuakse transkribeerimisega lihtsasti ligipääsetav informatsioon inimestele, kes soovivad neist heli- ning videosisudest kiirelt leida endale vajalikku informatsiooni, kuid ei soovi selleks läbi kuulata kogu salvestist.

2.5 Transkribeeritud tekstide eripärad ja kitsaskohad

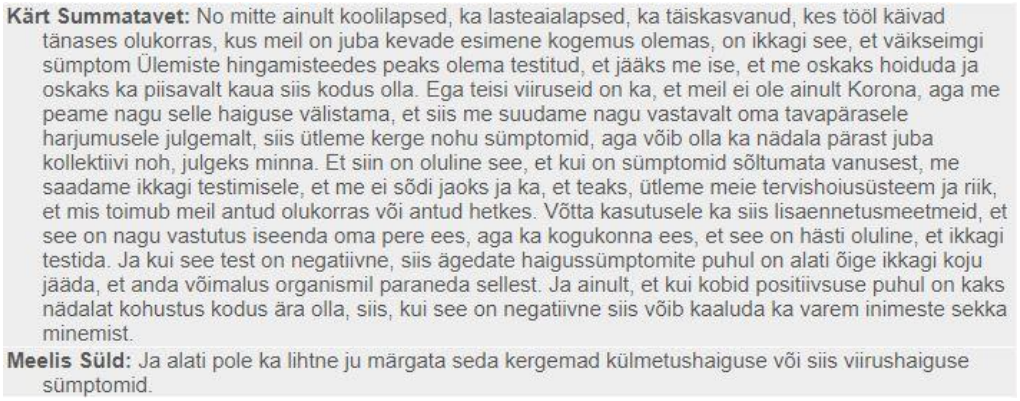
Transkribeeritud teksti sisendiks on suuline kõne, mis on oma olemuselt oluliselt keerukam, kui kirjalik kõne.

T. Hennoste kirjutab oma töös [25] järgmist: „Suulise ja kirjaliku kõne sõnavara ja lause erinevad teineteisest üsna palju, kusjuures kõne erijooned ei ole mitte kõneleja lohakuse või oskamatuse vili. Nad on tingitud kõnelemise protsessi ja kõneolukorra põhimõttelistest omadustest, mis ei kao ära ka siis, kui inimene kõneleb avalikus kohas ja räägib juba varem läbimõeldud teksti.“

K. Uibu [26] toob välja suulise ja kirjaliku kõne peamised erinevused: „Suuline kõne on spontaanne, sõltub suhtlussituatsioonist ja vestluspartneri reageeringutest. Ta on

šabloonsem kui kirjalik kõne, sisaldades suhtlussituatsioone, kus tuleb järgida etteantud käitumist. Lausestuse osas kasutatakse suulises kõnes enamasti lihtlauseid, mis võivad ühenduda liitlauseteks, või seotakse neid omavahel asesõnade ja kordustega. Suulise kõne kiire tempo tekib erinevate mõtete edastamisest, mis algselt on enamasti lühemad ega sarnane korrektsetele kirjakeele lausetele. Rohkesti esineb mõtte- ja kõhkluspause, mis täidetakse sõnakorduste, parasiitsõnade, partiklite, venituste ja/või üneemidega (st häälitsustega pausi täitmiseks ja kõnejärje hoidmiseks, nt *mmm, äää*). Suuline kõne on muutuv, reaalajas liikuv ja teisenev.“

Kirjaliku teksti puhul on lugejad harjunud, et tekst on üldjuhul korrektselt struktureeritud ning läbimõeldud ülesehitusega. Transkribeeritud tekst võib aga lugejale olla üsnagi raskesti jälgitav, sest seal esineb palju suulisele kõnele omaseid jooni. Keerukust lisavad ASR-ist tulenevad kõnetuvastusvead ning valesti märgistatud kirjavahemärgid (Joonis 2). Sellest tulenevalt võib transkribeeritud tekst osutuda võtmesõnade ekstraheerimise algoritmidele keerukamaks ülesandeks, kui tavapärane kirjalik tekst.


Kärt Summatavet: No mitte ainult koolilapsed, ka lasteaialapsed, ka täiskasvanud, kes tööl käivad tänases olukorras, kus meil on juba kevade esimene kogemus olemas, on ikkagi see, et väikseimgi sümptom Ülemiste hingamisteedes peaks olema testitud, et jääks me ise, et me oskaks hoiduda ja oskaks ka piisavalt kaua siis kodus olla. Ega teisi viiruseid on ka, et meil ei ole ainult Korona, aga me peame nagu selle haiguse välistama, et siis me suudame nagu vastavalt oma tavapärasele harjumusele julgemalt, siis ütleme kerge nohu sümptomid, aga võib olla ka nädala pärast juba kollektiivi noh, julgeks minna. Et siin on oluline see, et kui on sümptomid sõltumata vanusest, me saadame ikkagi testimisele, et me ei sõdi jaoks ja ka, et teaks, ütleme meie tervishoiusüsteem ja riik, et mis toimub meil antud olukorras või antud hetkes. Võtta kasutusele ka siis lisaennetusmeetmeid, et see on nagu vastutus iseenda oma pere ees, aga ka kogukonna ees, et see on hästi oluline, et ikkagi testida. Ja kui see test on negatiivne, siis ägedate haigussümptomite puhul on alati õige ikkagi koju jääda, et anda võimalus organismil paraneda sellest. Ja ainult, et kui kobid positiivsuse puhul on kaks nädalat kohustus kodus ära olla, siis, kui see on negatiivne siis võib kaaluda ka varem inimeste sekka minemist.
Meelis Süld: Ja alati pole ka lihtne ju märgata seda kergemad külmetushaiguse või siis viirushaiguse sümptomid.

Joonis 2. Väljavõte „Kõnesalvestuse brauseri“ lehe transkribeeritud tekstist [27].

2.5.1 Kõnetuvastus eestikeelsetele tekstidele

Kui suuremates riikides on kõnetuvastustehnoloogia arendajateks kommertsettevõtted, kuna sealne kasutajaskond on lai, siis Eestis on tehnoloogia arendamist siiani toetanud riik [28]. Siiski on Eesti Keeleressursside Keskuse[1] lehelt leitavad mitmed erinevad kõnetöötlusvahendeid.

[1] https://keeleressursid.ee/et/keeleressursid/konetootlusvahendid

Eestikeelset kõnetuvastustehnoloogiat on Eesti keeletehnoloogia riikliku programmi toel ning TalTech Keeletehnoloogia laboratooriumi poolt jõudsalt läbi aastate arendatud. Projekti tulemustest [24] on võimalik lugeda, et meeskond on kõnetuvastuse vigade osakaalu (WER) läbi aastate suutnud üha rohkem parandada (Tabel 1). Tabelis on toodud vigade osakaal protsentides ning väiksem arv tähendab paremat tulemust.

Tabel 1. Kõnetuvastuse vigade osakaal protsentides aastate lõikes [24].

Kõne tüüp	**2014**	**2015**	**2016**	**2017**
Raadio vestlussaated	16,9	15,7	12,4	9,9
Konverentsikõned	23,5	22,5	17,9	13,9
Aktuaalne Kaamera	19,6	17,1	15,5	9,6
Spontaanne kõne	39,9	31,6	22,4	17,6

Viimaste tulemuste põhjal [29] on raadio vestlussaadete WER 8.1%, konverentskõnede WER 12.9% ning erinevate kasutajate poolt tehtud salvestuste WER 22,7%.

T. Alumäe, TalTech Keeletehnoloogia laboratooriumi vanemteaduri, sõnul mingeid kindlaid andmeid valesti transkribeeritud sõnaliikide (näit. nimisõna, tegusõna) kohta pole tuvastatud (isiklik kommunikatsioon). Üldiselt on teada, et probleeme valmistavad võõrnimed ja ka tavalised nimi- ja muud sõnad, mis on hiljuti sõnavarasse lisandunud (näit. COVID-19).

Lisaks on Alumäe välja toonud [30] need kõne- ja suhtlussituatsioonid, kus praegune tuvastussüsteem palju eksib:

- mürane kõne (nt taustal on linnamüra vms);
- korraga kõneleb mitu inimest (nt mitme kõnelejaga koosolekud);
- ebaselge häädlus (nt eakate inimeste kõne);
- palju koodivahetust (nt ingliskeelsed terminid ja väljendid).

Kõige veavaesemad on raadiouudised ja vestlussaated, sest seal on üldjuhul professionaalsed kõnelejad ning kasutusel hea tehnika. Samuti puudub seal üksteisele

peale rääkimine. Kõne akustilise kvaliteedi langedes langeb ka kõnetuvastuse kvaliteet ning tekivad vead transkriptsiooni [28].

3 Eksperimendid

Käesolevas peatükis antakse ülevaade eksperimendis kasutatavatest andmetest ning sellest, millised osad otsustab magistritöö autor andmetest välja jätta. Kirjeldatakse töös kasutatavaid eeltöötluse erinevaid etappe ja kasutatavaid tekstitöötlusvahendeid. Viiakse läbi manuaalse võtmesõnade määramise eksperiment ning tutvustatakse automaatse võtmesõnade ekstraheerimise jaoks kasutatavaid algoritme.

3.1 Alusandmed

Eksperimendis toetutakse eestikeelsete transkribeeritud tekstide täpsuse hindamisel ning automaatsete märksõnade määramisel „Kõnesalvestuse brauseri“[1] veebirakenduses asuvate helisalvestiste transkribeeritud tekstide kogumikele.

„Kõnesalvestuse brauser“ on veebirakendus, mis on loodud TalTech Küberneetika Instituudi poolt "Eesti keele keeletehnoloogiline tugi (2006–2010)" raames ning võimaldab mugavalt sirvida automaatse kõnetuvastuse abil transkribeeritud kõnesalvestisi [27]. Veebirakenduse aluseks oleva programmi edasiarenduste lõppedes oli kõnetuvastuse vigade osakaal raadio vestlussaadetel 9,9% [24]. Kõige uuema (2018) viidatud informatsiooni kohaselt [29] on kõnetuvastuse süsteemi veelgi parandatud ning vestlustsaadete WER on 8,1%.

Veebirakenduses saab kuulata nelja Eesti raadio (Vikerraadio, Raadio 2, Raadio Kuku, Klassikaraadio) ja Geenius.ee taskuhäälingu (ingl *Podcast*) Algorütm vestlussaadete salvestusi ning lugeda saadetest automaatselt transkribeeritud teksti. Kõik raadiosaadete tekstid on automaatse nimetuvastusega jagatud kõnelejate kaupa lõikudeks.

Raadiosaadete juurde ei ole märgitud märksõnu ega lühikokkuvõtet. Samuti omavad saated kohati liiga üldist pealkirja (nt „Huvitaja. Tallinna Botaanikaaed“), mis ei anna täpsemat ülevaadet saates käsitletud teemade kohta. Kasutajamugavuse seisukohalt oleks

[1] http://bark.phon.ioc.ee/tsab/p/index

kindlasti vajalik nende lisamine, et lugejal oleks võimalik näha, kas salvestatud raadiosaade omab tema jaoks huvipakkuvat teemat või mitte.

Antud magistritöö raames kasutatakse vaid raadiosaadete vestlustekste ning välja jäetakse tekstilõikudele automaatselt lisatud kõnelejate nimed. Seda eelkõige seetõttu, et õigesti on tuvastatud vaid saatejuhid ning avalikkusele rohkem tuntud nimed. Valesti tuvastatud nimed toovad võtmesõnade märkimisel teksti juurde vaid müra ning ei anna ühtegi lisaväärtust.

Lisaks sisaldavad saated muusikaliste vahepalade transkriptsioone, mida kõnetuvastus on pidanud kõnelejate tekstiks. Analüüsis püüab töö autor selgeks teha, kas mainitud lõigud segavad automaatset märksõnade märkimist või mitte.

Praktilises osas võetakse TF-IDF meetodi jaoks võrdlusesse kümne juhuslikkuse alusel valitud raadiosaadete transkribeeritud tekstid. Seda põhjusel, et antud meetod vajab väljavalitud meetoditest ainsana märksõnade määramiseks suuremaid andmekogusid.

3.2 Teksti manuaalne parandamine

Ühe eeltööna enne võtmesõnade ekstraheerimise juurde asumist on magistritöö autoril vajalik teostada analüüsimiseks valitud raadiosaate transkribeeritud teksti manuaalne parandamine.

Manuaalse parandamise käigus parandatakse valesti tuvastatud sõnad ning eemaldatakse raadiosaate muusikaliste vahepalade transkriptsioonid. Muid Eesti kirjakeelele omaseid parandamisi (kirjavahemärkide lisamine ning eemaldamine) tekstis läbi ei viida. Vaid äärmisel vajadusel lisatakse lausete lõpumärgid kohtadesse, kus need saate kuulamisel tunduvad olevat kohased lisada.

Raadiosaate kuulamise ning transkribeeritud teksti vigade parandamisega elimineeritakse võimalus, et saate jaoks oluline välja öeldud sõna on kõnetuvastussüsteemi poolt kirja panemata või märgitud valesti. Samuti saab selle abil hiljem määrata, kas WER mõjutab kuidagi märksõnade automaatset märkimist. Manuaalselt parandatud transkribeeritud tekst on leitav magistritöö Lisas 4.

Vigade parandamise juures oli märgata olulist kõnetuvastuse eksimuste erinevust erinevate inimeste tekstide transkribeerimisel. Kõneleja puhul, kelle häädlus, sõnavara

ning lausutud mõte muutusid poole lause pealt, eksis kõnetuvastus oluliselt rohkem kui kõneleja puhul, kelle hääldus, traditsiooniline sõnavara ning lausutud jutt oli läbimõeldum. Viimase kõneleja teksti oli kasutajana kergem lugeda ning mõista ka ilma helisalvestist kuulamata, sest laused ei olnud veninud ebatavaliselt pikaks ning lohisevaks.

Lisaks oli märgata, nii nagu ka T. Alumäe mainis, et kõnetuvastus eksib palju hiljuti sõnavarasse lisandunud sõnade märkimisel. Antud raadiosaate puhul mängis olulist rolli ka asjaolu, et ühte konkreetset viirushaigust oli kõnelejate poolt nimetatud mitme eri nimetusega – COVID-19, SARS-CoV-2, koroonaviirus, koroona, Covid jne. Need kõik olid kõnetuvastuse jaoks uued sõnad. Ka kõnelejate rõhuasetus sõnade hääldamisel oli traditsioonilisest häälduses kohati üsna erinev.

Kuna viirushaigust oli nimetatud ja häälдatud erineval moel, oli ka kõnetuvastus neid tuvastanud erinevalt. Järgnevalt on loetletud mõned neist valesti tuvastatud sõnadest:

- Sars kov kahe;
- kobid;
- kobitestile;
- Coruna;
- Korona;
- Corona Virus;
- Kovid;
- Gorono.

Manuaalse märksõnade määramise juures saab inimene neid sõnu koondada ühise nimetuse alla. Automaatne süsteem neist sõnadest aga selliselt aru ei saa. Seetõttu võivad antud raadiosaate manuaalsed ja automaatselt määratud märksõnad suuresti erineda.

Tulemuste paremaks tõlgendamiseks ning võrdlemiseks kasutatakse manuaalselt parandatud teksti hiljem analüüsis, kus sellele rakendatakse samuti nelja eelnevalt välja

valitud võtmesõnade ekstraheerimise meetodit. Saadud tulemustega võrreldakse, kas ja kui palju erinevad märksõnad kahe teksti, transkribeeritud ja selle parandatud versiooni, vahel ning milline ekstraheerimise meetod annab seejuures kõige relevantsemad ehk manuaalselt määratud märksõnadele kõige sarnasemad tulemused.

3.3 Võtmesõnade ekstraheerimine

Järgnevalt kirjeldatakse manuaalselt ning automaatselt läbi viidud võtmesõnade ekstraheerimist. Eksperimendi jaoks valiti juhuslikkuse alusel välja 25.08.2020 salvestatud ning transkribeeritud Vikerraadio jutusaade Huvitaja. Saate teemaks oli „Nohused lapsed. Ruumide ventilatsioon ja viirused“[1].

3.3.1 Manuaalne võtmesõnade ekstraheerimine

Subjektiivse hinnangu vähendamiseks kaasati magistritöö raames manuaalsete võtmesõnade määramiseks kolm erapooletut inimest.

Osalistel paluti kuulata 54 minutit kestvat raadiosaadet ning seejärel määrata kuni 5 märksõna või fraasi (1 kuni 3 järjestikuse sõna jada), mis iseloomustaks kõige paremini kuulatud saadet. Märksõnade määramisel paluti neil kasutada sõnu, mis kuulatud tekstis vähemalt korra ilmnevad. Ühtegi teist piirangut kasutajatele ei seatud.

Erinevate kasutajate poolt raadiosaatele määratud märksõnad on välja toodud järgnevas tabelis (Tabel 2). Säilitatud on ka sõnade edastamise järjekord.

Tabel 2. Kasutajate poolt raadiosaatele määratud märksõnad

	Kasutaja 1	**Kasutaja 2**	**Kasutaja 3**
1	koroonaviirus	koroona	koroonaviirus
2	viirushaigused	sümptomid	sümptomid
3	nohused lapsed	gripp	viirushaigus
4	ventilatsioonisüsteem	ventilatsioon	testimine
5	juuste väljalangemine	juuste väljalangemine	ventilatsioon

[1] http://bark.phon.ioc.ee/tsab/p/play?trans=12498

Analüüsimisel viiakse kasutajate poolt määratud märksõnad nende algvormi ehk sõnad lemmatiseeritakse. Manuaalselt määratud märksõnad on kokku koondatult ning algvormi viidult järgnevad:

- koroonaviirus;
- viirushaigus;
- nohune laps;
- sümptom;
- testimine;
- gripp;
- ventilatsioon;
- juus väljalangemine.

3.3.2 Automaatne võtmesõnade ekstraheerimine

Võtmesõnade automaatseks ekstraheerimiseks on internetist võimalik leida nii vabavaralisi kui ka tasulisi rakendusi. Lisaks võrdluses olevate meetodite rakenduste vabale kättesaadavusele oli töö autorile testimise juures oluline programmide ja meetodite lihtne kasutatavus ning erinõuete puudumine arvutile.

Mitmetes võtmesõnu käsitlevates teaduslikes artiklites on viidatud meetodite Python rakendustele, mis on kasutamiseks saadaval GitHubis[1]. Kuigi Python rakenduste jaoks saab lihtsasti kasutada Google Colaboratory't[2], mis võimaldab kirjutada ja käivitada Pythoni keeles kirjutatud programme otse kasutaja brauseris, vajavad GitHubis jagatud algoritmid lisaks spetsiifilisemaid teadmisi mainitud programmeerimiskeelest, et rakendada neid eestikeelsetele tekstidele. Magistritöö autori eesmärgiks oli aga töö

[1] https://github.com/

[2] https://colab.research.google.com/notebooks/intro.ipynb

raames leida võimalus, kuidas ekstraheerida võtmesõnu eestikeelsetest tekstidest, omamata spetsiifilisemaid teadmisi ühestki programmeerimiskeelest.

Uurimustöö ning erinevate katsetuste tulemusel leidis töö autor, et R programmeerimiskeelele on loodud UDPipe[1] pakett, mida on võimalik rakendada eestikeelsete tekstide eeltöötluseks, omamata varasemat kogemust mainitud programmeerimiskeelest. Lisaks on R keelele loodud programmikoodid ning õpetused TF-IDF, TextRank ja RAKE meetodite kasutamiseks. Eelnevalt mainitud meetodite testimiseks võetakse töös kasutusele vabavaraline arenduskeskkond RStudio[2]. RStudiot on lihtne installeerida ning kasutusele võtta ning ettevõtetel on võimalik seda olemasolevate süsteemidega integreerida.

YAKE! meetodi jaoks R keeles paketti ega funktsiooni ei leidu. Lisaks ei anna YAKE! veebipõhine demorakendus[3] eestikeelse teksti peal rakendades relevantseid tulemusi. Kuigi demorakendus tunneb ära eestikeelse teksti, jääb väljatõmmatud sõnade loend tugevalt alla keskmise, sisaldades lihtsalt tekstis enim levinud sõnu nagu „aga, või, ole, väga, nagu“ jne. Autor on otsustanud jätta demorakenduse antud töö raames võrdlusest välja ning kasutada YAKE! rakendamiseks eesti keelele kohandatud Pyhtoni algoritmi[4], mille on loonud A. Lohk. Viidatud koodi testimiseks kasutatakse Google Colaboratoryt.

Andmete eeltöötlus

Manuaalne märksõnade märkimine ei vaja andmete eeltöötlusprotsessi, kuid enne automaatse ekstraheerimise juurde asumist tuleb teksti töödelda. Eeltöötlusetapp on esimene ning kõige suurem samm enne teksti esitamise ja analüüsi toimumist. Siin etapis tekst puhastatakse ning muudetakse masinloetavasse vormi: tuvastatakse olulised tükid ja eemaldatakse osad, mis on ebaolulised ning mürarikkad [13].

Töös kasutatakse kahte tekstitöötlusvahendit, milleks on EstNLTK ja UDPipe. EstNLTK on kasutusel YAKE! meetodi juures ning UDPipe rakendatakse eeltöötluseks TF-IDF, TextRank ning RAKE meetodite peal. Mõlemad vahendid on töö autorile usaldusväärsed,

[1] https://cran.r-project.org/web/packages/udpipe/vignettes/udpipe-annotation.html

[2] https://rstudio.com/

[3] https://boiling-castle-88317.herokuapp.com/

[4] https://www.dropbox.com/s/irzllznob95oeyv/YAKE_original_estonian_developed_estnltk_1.6.pdf?dl=0

sest nende eestikeelsete mudelite väljatöötamise juures on suuremal või vähemal määral tegevad Tartu Ülikooli teadurid.

EstNLTK (NLTK ehk *Natural Language ToolKit*) on peamiselt Pythonis kirjutatud kogumik teeke eestikeelsete tekstide töötlemiseks, mida arendab Tartu Ülikooli arvutiteaduse instituut [31]. Teek sisaldab eesti keele sõna-, lause- ja osalausepiiride tuvastamist, lemmatiseerimist ehk sõnade algvormide määramist ning morfoloogilist analüüsi ja ühestamist, sõnaliikide määramist jne [31].

EstNLTK vajab spetsiifilisemaid teadmisi keeletehnoloogiast ning põhjalikku teadmist Python programmeerimiskeelest, et osata rakendada tekstitöötluseks vajalikke toiminguid. Lisaks on teegi erinevatel versioonidel erinev programmikood ning napp dokumentatsioon algajale kasutajale.

EstNLTK kõige hilisema versiooni installeerimine Python programmeerimiskeeles toimub Joonis 3 kuvatud käsureaga [15].

```
!pip install estnltk==1.6.7b0
```

Joonis 3. EstNLTK installeerimine Pythonis.

EstNLTK tekstitöötluse ilmestamiseks kasutab magistritöö autor järgnevat programmikoodi (Joonis 4) [15].

```
from estnltk import Text

txt = Text("Tekst, millele soovitakse eeltöötlust rakendada.")
txt.tag_layer("morph_analysis")
txt["morph_analysis"]
```

Joonis 4. EstNLTK programmikood teksti eeltöötluseks.

Koodi sisestatud tekst tagastatakse tabeli kujul (Joonis 5). Joonisel on näha, kuidas sisestatud lause tükeldatakse sõnadeks ja tuvastatakse lause osad. Suurtähed muudetakse väiketähtedeks, igale sõnale antakse tema algvorm ehk lemma ja tuvastatakse sõna tüvi ning tüvisõna lõpp. Tabeli viimases veerus on kirjas tuvastatud sõna liik ehk POS (ingl *Part-of-speech*).

text	normalized_text	lemma	root	root_tokens	ending	clitic	form	partofspeech
Tekst	Tekst	tekst	tekst	['tekst']	0		sg n	S
,	,	,	,	[',']				Z
millele	millele	mis	mis	['mis']	le		pl all	P
	millele	mis	mis	['mis']	le		sg all	P
soovitakse	soovitakse	soovima	soovi	['soovi']	takse		takse	V
eeltöötlust	eeltöötlust	eeltöötlus	eel_tööt=lus	['eel', 'töötlus']	t		sg p	S
rakendada	rakendada	rakendama	rakenda	['rakenda']	da		da	V
.	.	.	.	['.']				Z

Joonis 5. EstNLTK vaade tekstiosade märgendamisest.

UDPipe on Tšehhi Vabariigi Karli Ülikooli formaal- ja rakenduslingvistika instituudile (ÚFAL)[1] kuuluv konveier (ingl *pipeline*). UDPipe võimaldab sõnapiiride tuvastamist, kõneosade märgendamist, lemmatiseerimist ja sõltuvuse parsimist, mis on loomuliku keele töötlemisel oluline osa [32].

UDPipe Universal Dependencies (UD) 2.5 mudel [33] pakub 61 erinevale keele jaoks 94 eeltreenitud mudelit. Magistritöö autor kasutab oma töös eesti keele jaoks eeltreenitud mudelit Estonian-EDT. Estonian-EDT on teisendatud versioon EDT (ingl *Estonian Dependency Treebank*) mudelist, mis on loodud Tartu Ülikooli poolt ja sisaldab 30 972 puud ning 437 769 sõnet [34].

UDPipe'i UD 2.5 versiooni [33] kirjeldustest võib järeldada, et mudel on usaldusväärne ja sobilik magistritöös kasutamiseks, sest antud eeltreenitud mudeli sooritus sõnestamisel ehk teksti tükeldamisel sõnadeks on 100,0% ja lausestamisel ehk teksti tükeldamisel lauseteks on 91,6%.

UDPipe Estonian-EDT mudelit rakendatakse programmeerimiskeeles R. Joonis 6 kuvatud algoritmi [35] sisestamisel RStudios installeeritakse mudel ning viiakse läbi teksti eeltöötlus.

[1] https://ufal.mff.cuni.cz/

```
library(udpipe)

model <- udpipe_download_model(language = "estonian")
model <- udpipe_load_model(file = model$file_model)
x <- udpipe_annotate (model, x = "Tekst, millele soovitakse eeltöötlust
rakendada.")
x <- as.data.frame(x)
x
```

Joonis 6. UDPipe paketi algoritm teksti eeltöötluseks.

UDPipe mudeli tulemuste vaade tekstiosade märgendamisest on nähtav Lisas 2, kus kuvatakse sarnaseid tulemusi EstNLTK tulemuste tabeliga. Olulisemad veerud, mis tabelisse on lisandunud on „*upos*“ ja „*xpos*“. UPOS (ingl *Universal POS*) on UD universaalne märgis kõneosade märgendamiseks. XPOS (ingl *treebank-specific*) märgis on keelespetsiifiline ehk sõltub konkreetse keele märgendatud puudepangal (ingl *treebank*). Magistritöö käigus toetutakse veerus „*upos*“ toodud märgistele.

Meetodite rakendamine

TF-IDF meetodi rakendamiseks on R programmeerimiskeeles kasutusel funktsioon *document_term_frequencies_statistics* (Joonis 7) [36].

```
data(saated)
x <- document_term_frequencies(saated[, c("doc_id", "lemma")])
x <- document_term_frequencies_statistics(x)
head(x)
```

Joonis 7. TF-IDF meetodi algoritm.

Andmefail „saated“ sisaldab koodis analüüsitava raadiosaate ning lisaks TF-IDF meetodi jaoks juhuslikult valitud kümne raadiosaate transkriptsioone.

Algoritmi rakendamise järel tagastatakse tabel automaatselt ekstraheeritud märksõnadega (Joonis 8). Tabeli veeru „*doc_id*“ väärtusega on võimalik eristada erinevaid analüüsis olevaid raadiosaateid. Lisaks kuvatakse tabelis lemmatiseeritud märksõna, sõna esinemissagedus ning *tf, idf* ja *tf_idf* väärtused. Kõige suurema „*tf-idf*„ väärtusega sõna on valitud raadiosaate (*doc1*) kõige olulisem märksõna.

doc_id	term	freq	tf	idf	tf_idf
doc1	sümptom	17	0.0025207592	2.39789527	0.0060445166
doc1	ventilatsioon	18	0.0026690391	1.70474809	0.0045500394
doc1	gripp	11	0.0016310795	2.39789527	0.0039111578
doc1	kooli_maja	8	0.0011862396	2.39789527	0.0028444784
doc1	ventilatsioonisüsteem	8	0.0011862396	2.39789527	0.0028444784
doc1	juus	11	0.0016310795	1.70474809	0.0027805796
doc1	nakkus	10	0.0014827995	1.70474809	0.0025277997
doc1	temperatuur	7	0.0010379597	2.39789527	0.0024889186
doc1	ventilatsiooni_süsteem	6	0.0008896797	2.39789527	0.0021333588
doc1	haigus	14	0.0020759193	1.01160091	0.0021000019

Joonis 8. TF-IDF tulemuste kuvamine raadiosaate parandatud tekstile.

Joonis 8 võib välja lugeda, et kuigi sõna „ventilatsioon“ esineb analüüsitavas raadiosaates (*doc1*) kõige enam, mis tõstab tema olulisust antud raadiosaates, siis tema väärtus dokumendikogus on madal ehk teda esineb sõnast „sümptom“ teistes raadiosaadetes rohkem, mis kukutab sõna olulisust antud raadiosaates.

Tabelis on kuvatud kaks sarnast sõna „ventilatsioonisüsteem“ ning „ventilatsiooni_süsteem“. Uurides UDPipe mudeli poolt loodud märgendatud tekstiosade tabelit, siis on näha, et mudel ei suuda tuvastada sõna „ventilatsioonisüsteem“ erinevaid käändeid ning tuvastab need seetõttu erinevate sõnadena. Magistritöö autor antud sõnu ei koonda, sest sarnasel moel võib mudel eksida ka teiste sõnade suhtes mujal tekstides.

TextRank meetodi jaoks kasutatakse töö raames R keele paketti *textrank* ning võtmesõnade leidmiseks omakorda funktsiooni *textrank_keywords* (Joonis 9) [37].

```
library(textrank)
keywords <- textrank_keywords(x$token,
                              relevant = x$upos %in% c("NOUN", "ADJ"),
                              ngram_max = 2, p = 1/3, sep = " ")
keywords <- subset(keywords$keywords, ngram > 1)
head(keywords,10)
```

Joonis 9. TextRank meetodi algoritm.

Algoritm ehitab sõnade võrgu, mis on loodud viisil, kus vaadeldakse, millised sõnad üksteisele järgnevad ehk mis sõnad omavahel ettemääratud aknas koos esinevad. Mida sagedamini need sõnad omavahel järjestikku esinevad, seda suurem on nende vaheline side ning olulisus tekstis [37].

Töö autori poolt on algoritmi määratud, et fraasid peavad koosnema nimi- ja omadussõnadest ning nende koosesinemise akna suuruseks on 2. Algoritmi tulemused on kuvatud Joonis 10.

keyword	ngram	freq
kerge nohu	2	3
sümptomi ravi	2	2
värskes õhus	2	2
suletud ruumid	2	2
hea ventilatsioon	2	2
korda väiksem	2	2
tänases tervise	2	1
stuudios külas	2	1
külas perearst	2	1
sümptomitega patsientide	2	1

Joonis 10. TextRank tulemuste kuvamine raadiosaate parandatud tekstile.

Erinevad parameetrid said töö autori poolt algoritmis testitud, kuid kõige relevantsemad tulemused (Joonis 10) tagastasid just need sisendid. Ka TextRanki loojad [20] saavutasid eelnevalt mainitud parameetritega kõige paremad tulemused. Lisaks järjestab algoritm tulemused sageduse järjekorras.

RAKE meetodi rakendamiseks kasutatakse R keele funktsiooni *keywords_rake*, mille algoritm on kirjas Joonis 11 [38].

```
keywords <- keywords_rake(x = x, term = "lemma", group = "doc_id",
                          relevant = x$upos %in% c("NOUN", "ADJ"),
                          ngram_max = 3, n_min = 2, sep = " ")
head(keywords, 10)
```

Joonis 11. RAKE meetodi algoritm.

Sarnaselt TextRankiga soovib töö autor, et meetodiga ekstraheeritakse vaid nimi- ja omadussõnu. *Ngram_max* väärtus määrab maksimum sõnade arvu märksõnas ning *n_min* näitab sagedust, mitu korda peab termin tulemuse tagastamiseks tekstis minimaalselt esinema. *N_min* vaikimisi väärtuseks on 2. Algoritmi tulemus on kuvatud Joonis 12, kus

kõige suurema „*rake*“ veeru väärtusega sõna on meetodi jaoks raadiosaate kõige olulisem sõna. Ekstraheeritud võtmesõnad on järjestatud RAKE skoori kahanevas järjekorras.

	keyword	ngram	freq	rake
1	tervis	1	2	2.4000000
2	hea ventilatsioon	2	2	2.3492063
3	kerge _nohu	2	2	2.0500000
4	ohutu paik	2	2	2.0000000
5	sümptom ravi	2	2	2.0000000
6	väike ruum	2	2	1.8611111
7	esimene samm	2	2	1.8333333
8	viirushaigus	1	2	1.8000000
9	tõhus lahendus	2	2	1.7500000
10	inimene jagu	2	2	1.7000000

Joonis 12. RAKE tulemuste kuvamine raadiosaate parandatud tekstile.

YAKE! testimiseks kasutatakse eesti keelele kohandatud Pyhtoni algoritmi[1]. Antud algoritmi juures kasutatakse ainsana lisaks stoppsõnade loendit, mille sõnad on loetletud lõputöö Lisas 3. Joonis 13 kuvatakse Google Colaboratory’s algoritmi käivitamisel tagastatud tulemused.

```
find_candidate_keyword_scrore: 0.001020193099975586
[('laps', 0.003798772554769439),
 ('inimene', 0.0054947328349804585),
 ('ventilatsioon', 0.011360148297759458),
 ('aeg', 0.014260628668503188),
 ('haigus', 0.018464745055398815),
 ('viirus', 0.018745596758586953),
 ('näiteks', 0.02154228860832187),
 ('perearst', 0.021993337674029598),
 ('gripp', 0.02318367015450167),
 ('koroona', 0.028155512338720726),
```

Joonis 13. YAKE! meetodi tulemuste kuvamine raadiosaate parandatud tekstile.

YAKE! tulemused kuvatakse märksõnaskoori kasvamise järjekorras. Kõige väiksema skoori saanud sõna on meetodi jaoks raadiosaate kõige tähenduslikum sõna.

[1] https://www.dropbox.com/s/irzllznob95oeyv/YAKE_original_estonian_developed_estnltk_1.6.pdf?dl=0

4 Tulemuste valideerimine ja analüüs

Meetodite efektiivsuse hindamine põhineb traditsiooniliselt käsitsi märgendatud võtmesõnade ning automaatselt ekstraheeritud märksõnade võrdlemisel. Käsitsi märgendatud võtmesõnad on kuldstandardiks (ingl *gold standard*) ning meetod, mille märksõnad ühtivad kõige rohkem käsitsi märgendatud võtmesõnadega, saab kõige kõrgema hinde.

4.1 Hindamismõõdikud

Kõige tuntumad hindamismeetodid, mida võtmesõnade ekstraheerimisel kasutatakse on täpsus (ingl *precision*), saagis (ingl *recall*) ning F_1-skoor.

Täpsust hinnatakse valemiga:

$$täpsus = \frac{õigesti\ pakutud\ märksõnade\ arv}{ekstraheeritud\ võtmesõnade\ arv} = \frac{TP}{TP + FP},$$

kus TP (ingl *True Positive*) on õige positiivsete ja FP (ingl *False Positive*) on valepositiivsete sõnade arv [12]. Täpsus näitab, kui suur hulk märksõnadest ennustati õigesti [15].

Saagise arvutamise valem on järgmine:

$$saagis = \frac{õigesti\ pakutud\ märksõnade\ arv}{määratud\ sõnade\ koguarv} = \frac{TP}{TP + FN},$$

kus TP on õige positiivsete ja FN (ingl *False Negative*) valenegatiivsete sõnade arv [12]. Antud magistritöös näitab saagis, kui paljud automaatselt ekstraheeritud märksõnad kattusid kasutajate poolt käsitsi märgendatud sõnadega, st korrektselt ekstraheeritud sõnade osakaal käsitsi märgendatud sõnade seas.

F_1-skoori kasutatakse testi täpsuse mõõduna. F_1-skoor on täpsuse ja saagise harmooniline keskmine. Kui skoor saavutab oma kõrgeima väärtuse 1, siis tähendab see täiuslikku (100%-list) täpsust ja saagist. Halvimal juhul on F_1-skoori väärtus 0 [15].

F_1-skoori arvutatakse järgneva valemiga [12]:

$$F_{1-skoor} = 2 \times \frac{täpsus \times saagis}{täpsus + saagis}.$$

Hindamisel võrreldakse manuaalselt ja automaatselt märgitud võtmesõnade algvorme. Sealjuures eemaldatakse vajadusel ka sidekriipsud ning muud kirjavahemärgid.

Lisaks hinnatakse TF-IDF, TextRank, RAKE ja YAKE! meetodite täpsust, saagist ning F_1-skoori erinevate märksõnade hulga (5, 8 ja 10) ekstraheerimise juures. Sellega soovitakse teada, kui palju on mõistlik võtmesõnu tekstist ekstraheerida ning mis hetkel muutuvad väljatõmmatud sõnad ebavajalikuks.

Meetodite headust saab hinnata nii täpse kui osalise vaste teel. Töö autor toob mõlema lähenemise tutvustamisel näidiseks TextRank meetodi hindamise viie märksõna ekstraheerimisel.

4.1.1 Täpne vaste

Täpse vaste puhul peab manuaalselt määratud ning automaatselt ekstraheeritud sõna olema esitatud samal moel ehk kasutaja poolt määratud märksõna fraas „nohune laps“ peab ka automaatselt olema ekstraheeritud kujul „nohune laps“.

Kasutaja poolt määratud võtmesõnad ehk kuldstandardid raadiosaate originaaltekstile on järgnevad: {koroonaviirus, viirushaigused, nohused lapsed, sümptomid, testimine, gripp, ventilatsioon, juuste väljalangemine}. Peale sõnade algvormi viimist on tulemuseks:

{(koroonaviirus), (viirushaigus), (nohune, laps), (sümptom), (testimine), (gripp), (ventilatsioon), (juus, väljalangemine)}.

TextRank algoritmiga väljastati järgnevalt loetletud võtmesõnad: {kerge nohu, sümptomi ravi, värskes õhus, suletud ruumid, hea ventilatsioon}. Peale fraaside algvormi viimist saadi tulemuseks:

{(kerge nohu), (sümptom, ravi), (värske, õhk), (sulgema, ruum), (hea, ventilatsioon)}.

Kuigi algoritmi poolt pakutud sõnad on autori hinnangul raadiosaatest ülevaate andmiseks relevantsed, ei kattu ükski neist märksõnadest kasutajate poolt märgitud

võtmesõnadega. Seetõttu saab TextRank viie märksõna ekstraheerimisel täpse vaste hindamise teel tulemuseks 0 ehk algoritm ei ekstraheerinud ühtegi sõna korrektselt.

4.1.2 Osaline vaste

Osalise vaste puhul märksõnade fraasid tükeldatakse. See tähendab, et sõna „nohune laps“ võrreldakse üksikute sõnadena ehk „nohune“ ning „laps“.

Peale kasutajate poolt määratud märksõnade algvormi viimist ning fraaside tükeldamist on tulemus järgnev:

{(koroonaviirus), (viirushaigus), (nohune), (laps), (sümptom), (testimine), (gripp), (ventilatsioon), (juus), (väljalangemine)}.

Järgnevalt on TextRank algoritmi poolt väljastatud fraasid tükeldatud ning lisaks on rasvase kirjaga eristatud märksõnad, mis kattuvad manuaalselt märgitud võtmesõnadega:

{(kerge), (nohu), (**sümptom**), (ravi), (värske), (õhk), (sulgema), (ruum), (hea), (**ventilatsioon**)}.

Osalise vaste puhul väljastab algoritm 2 TP sõna (sümptom, ventilatsioon), 8 FP (kerge, nohu, ravi, värske, õhk, sulgema, ruum, hea) ning 8 FN (koroonaviirus, viirushaigus, nohune, laps, testimine, gripp, juus, väljalangemine) sõna. Osalise vaste hindamise teel on meetodi täpsus 0,2, saagis 0,2 ning F_1-skoor samuti 0,2. Seega on TextRanki tulemus viie märksõna ekstraheerimisel osalise vaste puhul 20% parem, kui täpse vaste puhul. Lõputöö autor viib kahe erineva lähenemisega hindamise läbi kõikide meetodite juures, mille tulemused on toodud järgmises peatükis.

4.2 Valideerimine ja analüüs

TF-IDF, TextRank, RAKE ja YAKE! meetodite abil ekstraheeritakse märksõnad ühe ja sama transkribeeritud raadiosaate kolmele erinevale versioonile:

- raadiosaate transkribeeritud originaaltekst (tabelites tähistatud sõnaga „originaal“);
- raadiosaate transkribeeritud originaaltekst, millest on eemaldatud muusikaliste vahepalade transkriptsioonid (tabelites tähistatud sõnaga „originaal-lauluta“);

- raadiosaate transkriptsiooni käsitsi parandatud tekst (tabelites tähistatud sõnaga „parandatud“).

Kolme versiooni tulemusi võrreldakse omavahel, et saada teada, kui palju mõjutab WER võtmesõnade ekstraheerimisel saavutatud tulemusi. WER hindab, mitu protsenti teksti sõnadest tuvastati valesti.

Raadiosaate transkribeeritud originaalteksti WER on 12%. Eemaldades sealt muusikalised vahepalad on WER 8,4%. WER % määramiseks kasutatakse Pyhtoni koodi, mille on kirjutanud A. Lohk (Lisa 5).

Järgnevas kahes tabelis on kuvatud meetodite täpsus ehk kui paljud kõikidest ekstraheeritud võtmesõnadest olid ka kasutajate poolt märgitud märksõnadeks. Tabel 3 on toodud täpsuse hindamise täpne vaste ning Tabel 4 osalise vaste tulemused. Veergude pealkirjad „originaal“, „originaal-lauluta“ ning „parandatud“ tähistavad ühe raadiosaate kolme erinevat versiooni. Number tähise „n“ järel tähistab tekstist ekstraheeritud võtmesõnade arvu. Parimad tulemused on märgitud rasvases kirjas.

Tabel 3. Meetodite täpsuse hindamine täpse vaste teel.

	originaal			**originaal-lauluta**			**parandatud**		
Meetod	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	**0.600**	0.375	0.333	**0.600**	0.375	0.333	**0.600**	0.375	0.333
TextRank	0	0	0	0	0	0	0	0	0
RAKE	0	0.125	0.100	0	0.125	0.100	0	0.125	0.100
YAKE!	0.200	0.250	0.200	0.200	0.250	0.200	0.200	0.125	0.200

Tabel 4. Meetodite täpsuse hindamine osalise vaste teel.

	originaal			**originaal-lauluta**			**parandatud**		
Meetod	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	**0.600**	0.500	0.444	**0.600**	0.500	0.444	**0.600**	0.500	0.444
TextRank	0.200	0.125	0.111	0.200	0.125	0.111	0.200	0.125	0.111
RAKE	0.200	0.200	0.158	0.200	0.200	0.158	0.222	0.214	0.167
YAKE!	0.200	0.250	0.200	0.200	0.250	0.200	0.200	0.250	0.300

Tabelitest Tabel 3 ja Tabel 4 järeldub, et TF-IDF meetod suudab õigesti tuvastada kõige rohkem märksõnu ning seda just viie märksõna ekstraheerimisel. Täpsemalt kattusid 60% meetodi sõnadest kasutajate poolt märgitutega ning seda nii osalise kui täpse vaste puhul. TF-IDF meetodi tulemustest nähtub veel, et enam kui viie märksõna väljatõmbamisel, hakkavad algoritmi ja kasutajate tulemused järjest rohkem erinema. Konkreetse analüüsi juures sõltub saadud tulemus sellest, et tekstitöötlusvahend ei suutnud tuvastada sõna „ventilatsioonisüsteem“ käändelõppe ning meetod ekstraheeris olulise märksõna kahel erineval moel.

YAKE! meetod ainsana tagastab parandatud teksti 10 võtmesõna ekstraheerimise juures paremaid tulemusi, siis võib siinkohal järeldada, et TF-IDF meetodi juures võiks piirduda viie ning ülejäänud meetodite juures kaheksa märksõna ekstraheerimisega, rohkemad sõnad muutuvad juba ebavajalikuks.

Kõige kehvemad tulemused on saavutanud TextRank, saades täpse vaste hindamisel tulemuseks 0. Seda eelkõige põhjusel, et meetod ekstraheeris kõige enam kahest sõnast koosnevaid fraase, kasutajad piirdusid aga märksõnade märkimisel üldjuhul ühesõnaliste terminitega. Sellegipoolest saavutab ka osalise vaste hindamise juures kõige kehvemad tulemused just see meetod.

Osalise vaste hindamise juurde aga tuleb vaadata ka saagise tulemusi, ehk korrektselt ekstraheeritud sõnade osakaalu käsitsi märgendatud sõnade seas. Seda põhjusel, et täpsust saab suurendada mudeliga, mis ennustab korrektseid tulemusi, aga tulemusi on seejuures arvuliselt vähem. Korrektsete sõnade osakaal on meetodil küll suur, aga väga palju olulisi märksõnu võis jääda ekstraheerimata.

Saagise arvutamisel on meetodite tulemused sarnased eelnevale ehk kõige paremad tulemused saavutab TF-IDF meetod (Tabel 5). 37,5% nimetatud meetodi poolt ekstraheeritud sõnadest esinevad kasutajate poolt märgitud märksõnades. TextRank meetod ei tagasta aga ühtegi kasutajate poolt märgitud fraasi.

Tabel 5. Meetodite saagise hindamine täpse vaste teel.

	originaal			**originaal-lauluta**			**parandatud**		
Meetod	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	**0.375**	**0.375**	**0.375**	**0.375**	**0.375**	**0.375**	**0.375**	**0.375**	**0.375**
TextRank	0	0	0	0	0	0	0	0	0
RAKE	0	0.125	0.125	0	0.125	0.125	0	0.125	0.125
YAKE!	0.125	0.250	0.250	0.125	0.250	0.250	0.125	0.125	0.250

Saagise hindamisel osalise vaste teel on näha (Tabel 6), et ka selle puhul tagastab parima tulemuse TF-IDF. Kuigi TextRanki tulemus on paranenud, ei sõltu see siiski erineva arvu võtmesõnade ekstraheerimisest. Teised meetodid saavutavad 10% võrra paremaid tulemusi enam kui viie märksõna väljatõmbamisel. Parandatud teksti juures paraneb YAKE! tulemus kümne võtmesõna ekstraheerimisel veel omakorda 10%.

Tabel 6. Meetodite saagise hindamine osalise vaste teel.

	originaal			**originaal-lauluta**			**parandatud**		
Meetod	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	0.300	**0.400**	**0.400**	0.300	**0.400**	**0.400**	0.300	**0.400**	**0.400**
TextRank	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
RAKE	0.200	0.300	0.300	0.200	0.300	0.300	0.200	0.300	0.300
YAKE!	0.100	0.200	0.200	0.100	0.200	0.200	0.100	0.200	0.300

Täpsust ja saagist tuleb vaadata kvaliteedinäitajatena siiski koos. Meetodite täpse vaste hindamise harmoonilised keskmised ehk F_1-skoorid on kuvatud Tabel 7 ning osalise vaste hindamise teel saadud tulemused Tabel 8.

Tabel 7. Meetodite F_1-skoor täpse vaste hindamise teel.

	originaal			**originaal-lauluta**			**parandatud**		
Meetod	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	**0.462**	0.375	0.353	**0.462**	0.375	0.353	**0.462**	0.375	0.353
TextRank	0	0	0	0	0	0	0	0	0
RAKE	0	0.125	0.111	0	0.125	0.111	0	0.125	0.111
YAKE!	0.154	0.250	0.222	0.154	0.250	0.222	0.154	0.125	0.222

Tabel 8. Meetodite F_1-skoor osalise vaste hindamise teel.

Meetod	originaal			originaal-lauluta			parandatud		
	n=5	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**	**n=5**	**n=8**	**n=10**
TF-IDF	0.400	**0.444**	0.421	0.400	**0.444**	0.421	0.400	**0.444**	0.421
TextRank	0.200	0.154	0.143	0.200	0.154	0.143	0.200	0.154	0.143
RAKE	0.200	0.240	0.207	0.200	0.240	0.207	0.211	0.250	0.214
YAKE!	0.133	0.222	0.200	0.133	0.222	0.200	0.133	0.222	0.300

Tabelitest Tabel 7 ja Tabel 8 nähtub, et valideerimise kokkuvõttes saavutas TF-IDF algoritm mõlema hindamisviisi juures kõige paremad tulemused. Meetodi täpsuseks võtmefraaside ekstraheerimisel raadiosaate transkribeeritud tekstist võib lugeda 46,2% ning selle tulemuse saavutab meetod viie võtmesõna väljatõmbamisel täpse vaste teel. Osalise vaste korral saavutab TF-IDF parima tulemuse (44,4%) kaheksa võtmesõna ekstraheerimisel.

RAKE algoritmi tulemused on paremuselt teisel kohal ning meetod on kõige efektiivsem kaheksa võtmesõna ekstraheerimisel. YAKE! saavutab vaid parandatud teksti kümne märksõna väljatõmbamise juures RAKE algoritmist parema tulemuse. Kõige kehvemad tulemused saab kirja TextRank.

4.3 Tulemuste analüüs ja järeldused

Eksperimendi tulemusi analüüsides toetub autor töö teoreetilises osas kirjeldatud võtmefraaside ekstraheerimist mõjutavatele teguritele.

Pikkus. Pikkuselt kaldus veidi alla seitsme tuhande märgisega raadiosaate originaaltekst pigem pika sisuga tekstide hulka. Võrdluseks võib siinkohal välja tuua, et väga lühikese sisuga dokumendiks loetakse 75 märgisega ning pikaks 8 tuhande märgisega tekste [13]. Kuna graafipõhised meetodid töötavad paremini lühikestel dokumentidel, siis võib sellest järeldada, et seetõttu andis graafipõhine TextRank kõige kehvemad tulemused. Samas suudab näiteks RAKE konkureerida statistiliste meetodi YAKE! tulemustega, jäädes täpse vaste puhul küll YAKE! algoritmile alla, kuid suutes osalise vaste puhul näidata sellest paremaid tulemusi. Seda ilmselt põhjusel, et RAKE tagastas vastupidiselt YAKE!

algoritmile märksõnade fraase, mis tükeldades ühtisid paljude kasutajate poolt määratud märksõnadega.

Struktuuriline ülesehitus. Eksperimendis kasutatud raadiosaade oli struktureeritud ülesehitusega. Raadiosaate alguses tutvustati kokkuvõtvalt arutusele tulevaid teemasid ning seejärel liiguti iga teema juurde süvitsi. Arutelus olnud teemasid oli saates kokku kolm. Kaks esimest teemat sisaldasid üsna võrdselt teksti mahtu (vastavalt ca 2900 ning 2040 märgist), kuid kolmas jäi ca 560 märgisega kõrvalisemaks. Kõik võrdluses olevad meetodid tuvastasid enim võtmesõnu just kahest mahukamast saateosast ning viimane teema jäi neil märksõnadena kajastamata. Oskuslikult jätsid kõik meetodid kõrvale saate lõpus järgmise saate tutvustava teksti.

Teema muutus ning korrelatsioon. Raadiosaated on oma ülesehituselt üldjuhul kindlas formaadis, kus kõik saate jooksul käsitletavad teemad on omavahel seotud. Eksperimendis kasutatud saate läbivaks teemaks oli koroonaviirus ning millest kõrvalekaldeid ei toimunud. Siiski ei suutnud ükski meetod raadiosaate originaaltekstist ekstraheerida sõna „koroona“ või sellega sarnast sõna. YAKE! algoritm tagastas sõna koroona 10 võtmesõna ekstraheerimisel viimase sõnana.

Käesoleva magistritöö peamised järeldused:

- Kõige relevantsemad märksõnad tagastab eestikeelsele transkribeeritud raadiosaate tekstile TF-IDF meetod.
- Võtmesõnade ekstraheerimisel võib piirduda viie kuni kaheksa võtmesõna väljatõmbamisega. Rohkemad märksõnad on pigem ebavajalikud.
- Täpse vaste hindamisel olid neljast meetodist kolme tulemused kesised ehk suudeti tuvastada vaid kuni 25% märksõnadest. Meetodite hindamistulemused võisid olla kehvad, sest manuaalsete märksõnade märkijateks olid amatöörid ning neile anti märksõnade määramisel suuresti vabad käed.
- Ühesõnaliste märksõnadega meetodid TF-IDF ning YAKE näitavad kokkuvõttes täpsuse hindamisel kõige paremaid tulemusi, kuid seda seetõttu, et ka kasutajate poolt määratud märksõnad olid enamjaolt ühesõnalised. Samas annab TF-IDF

algoritmi märksõna „sümptom“ autori hinnangul vähem kasulikku teavet teksti kohta, kui TextRanki „sümptomi ravi“ tulemus.

- Kõikide meetodite puhul võib järeldada, et originaalteksti 12% veamäär (WER) ei mõjuta veel kuidagi ühegi algoritmi efektiivsust.

- Kuigi mitmetes teaduslikes artiklites on kiidetud YAKE! meetodi häid tulemusi ning lihtsat rakendamist erinevatele keeltele, siis vajab algoritm spetsiifilisemaid teadmisi programmeerimiskeelest Python ning keeletehnoloogiast, et neid eestikeelsetele tekstidele rakendada.

- TF-IDF, TextRank ning RAKE meetodite rakendamine R keelel ei vaja lisateadmisi mainitud programmeerimiskeelest. R keelele kirjutatud algoritmid on lisaks lihtsasti rakendatavad.

Raadiosaated on kõnetuvastustehnoloogiale üks lihtsamaid domeene. Keerulisemaks osutuvad koosolekute transkriptsioonid, kus sõna võivad võtta mitmed erinevad inimesed läbisegi, esineda võib üksteisele peale rääkimist ning teemaväliseid vestluseid. Kõnelejate kaugus mikrofonist ning ruumi suurus lahjendab heli kvaliteeti ning langeb kõnetuvastuse kvaliteet. Seetõttu pakub töö autor välja teema edasiarendusena võtmesõnade automaatset ekstraheerimist koosolekute transkriptsioonidest. Koosolekute puhul on WER % suurem ning võimalus on paremini analüüsida sõnatuvastuse vigadest tulenevat võtmesõnade ekstraheerimise täpsust.

5 Kokkuvõte

Käesoleva töö eesmärgiks oli välja selgitada parim võtmesõnade ekstraheerimise meetod eestikeelsetele transkribeeritud tekstidele. Töös kasutati „Kõnesalvestuse brauseri“ lehel asuvaid raadiosaadete transkribeeritud tekstide kogumikke.

Eksperimendi käigus rakendati väljavalitud transkribeeritud raadiosaate kolmele erinevale versioonile nelja erinevat võtmesõnade ekstraheerimise juhendamata õppe meetodit. Parim võtmesõnade ekstraheerimise meetod eestikeelsete tekstide jaoks selgitati välja eksperimendi tulemuste analüüsi abil. Lisaks näitas teostatud analüüs, et eestikeelsete transkribeeritud raadiosaadete tekstide juures ei mõjuta väike veamäär võtmesõnade ekstraheerimisel tulemusi.

Töö peamisteks tulemusteks on:

- erinevate võtmesõnade ekstraheerimise algoritmide võrdlus;
- nii võtmesõnade ekstraheerimise kui ka transkribeeritud tekstide kaardistatud kitsaskohad ning eripärad, mis raskendavad võtmefraaside ekstraheerimist;
- valminud analüüs, milles hinnati realiseeritud meetodite efektiivsust ning kasutamise lihtsust eestikeelsete transkribeeritud tekstide peal;
- välja valitud kõige efektiivsem meetod eestikeelsete transkribeeritud tekstide võtmesõnade ekstraheerimiseks.

Kasutatud kirjandus

[1] M. Anandarajan, C. Hill, T. Nolan, Practical Text Analytics - Maximizing the Value of Text Data, Springer, Cham, 2019.

[2] P. Meladianos, A. Tixier, G. Nikolentzos, M. Vazirgiannis, „Real-Time Keyword Extraction from Conversations," %1 *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Spain, 2017.

[3] „Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition,* kd. 41, nr 10, pp. 2965-2979, 2008.

[4] E. Meister, „Eestikeelse kõnetuvastuse meetodite uurimine ja arendamine," Eesti Teadusinfosüsteem, [Võrgumaterjal] https://www.etis.ee/Portal/Projects/Display/3688dc6a-e592-49f4-8bd3-91a58c552472?tabId=tab_GeneralData. [Kasutatud 10 2020].

[5] S. Zupping, „2019. aasta keeleauhinnad," *Oma Keel,* kd. 40, 2020.

[6] M. Kumar, „The Future of Medical, Legal & General Transcription: Jobs Outlook," Transcription Certification Institute, [Võrgumaterjal]. https://www.transcriptioncertificationinstitute.org/blog/future-of-transcription-jobs-digital-age. [Kasutatud 10 2020].

[7] „Tehisintellekt teisendab tekstiks Leedu heli ja videosalvestise," Tilde.AI, [Võrgumaterjal]. https://www.tilde.ai/et/work/kantar-tns. [Kasutatud 10 2020].

[8] A. Désilets, B. de Bruijn, J. Martin, „Extracting Keyphrases from Spoken Audio Documents," %1 *Information Retrieval Techniques for Speech Applications*, Springer, 2002, pp. 36-50.

[9] F. Liu, D. Pennell, F. Liu, Y. Liu, „Unsupervised approaches for automatic keyword extraction using meeting transcripts," %1 *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, Colorado, 2009.

[10] R. Sirel, „Poolautomaatne teadmusbaaside konstrueerimine dialoogsüsteemidele," 2011. [Võrgumaterjal]. https://www.cl.ut.ee/yllitised/sirel2011.pdf. [Kasutatud 10 2020].

[11] T. Ennomäe, „Turunduspakkumiste edukus tulenevalt kliendi varasemast meelestatusest klienditeeninduses Euroopa telekomiettevõtte näitel," 2019. [Võrgumaterjal] http://dspace.ut.ee/bitstream/handle/10062/65310/ennomae_terje.pdf?sequence=1&isAllowed=y. [Kasutatud 10 2020].

[12] E. Papagiannopoulou, G. Tsoumakas, „A review of keyphrase extraction," *WIREs Data Mining and Knowledge Discovery,* kd. 10, nr 2, 2020.

[13] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, „YAKE! Keyword extraction from single documents using multiple local features," *ScienceDirect,* kd. 509, pp. 257-289, 2020.

[14] E. Papagiannopoulou, „Keyphrase Extraction Techniques,“ 08 02 2021. [Võrgumaterjal] http://ikee.lib.auth.gr/record/327188/files/GRI-2021-29716.pdf. [Kasutatud 04 2021].

[15] A. Lohk, „Tekstikaeve - loeng ITB8828,“ A. Lohk, Tallinn, 2020.

[16] I. Gagliardi, M. T. Artese, „Unsupervised Automatic Keyphrases Extraction Algorithms,“ %1 *On the Move to Meaningful Internet Systems: OTM 2018 Workshops*, 2019.

[17] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, A. Jatowt, „A Text Feature Based Automatic Keyword Extraction Method for Single Documents,“ %1 *European Conference on Information Retrieval*, 2018.

[18] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, M. M. Rahman, „TeKET: a Tree-Based Unsupervised Keyphrase Extraction Technique,“ *Cognitive Computation*, kd. 12, p. 811–833, 2020.

[19] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, C. Zhang, „SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model,“ *IEEE Access*, kd. 8, pp. 10896-10906, 2020.

[20] R. Mihalcea, P. Tarau, „TextRank: Bringing Order into Texts,“ %1 *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Spain, 2004.

[21] S. Rose, D. Engel, N. Cramer, W. Cowley, „Automatic Keyword Extraction from Individual Documents,“ %1 *Text Mining: Applications and Theory*, John Wiley & Sons, 2010, pp. 1-20.

[22] K. S. Hasan, V. Ng, „Automatic Keyphrase Extraction: A Survey of the State of the Art,“ %1 *The 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 2014.

[23] „Transkriptsioon - kõnest kirja,“ Keeleait, 2017. [Võrgumaterjal] https://www.keeleait.ee/transkriptsioon-konest-kirja/. [Kasutatud 08 2020].

[24] „Kõnetuvastus 2,“ Eesti Keeletehnoloogia, [Võrgumaterjal] https://www.keeletehnoloogia.ee/et/ekt-projektid/konetuvastus-2. [Kasutatud 10 2020].

[25] T. Hennoste, „Sissejuhatus suulisse eesti keelde,“ *Oma Keel*, 09 2000.

[26] K. Uibu, „Internetikeele mõjud kirjakeeles,“ Tartu Ülikooli Dspace, [Võrgumaterjal] https://dspace.ut.ee/bitstream/handle/10062/15383/peamised_erinevused.html. [Kasutatud 11 2020].

[27] „Kõnesalvestuse brauser,“ Tallinna Tehnikaülikool, [Võrgumaterjal] http://bark.phon.ioc.ee/tsab/p/about.

[28] T. Alumäe, Interviewee, *TalTechi teadur: keeleandmete kasutamisel automaatse kõnetuvastuse arendamiseks on palju halli ala.* [Intervjuu]. [Kasutatud 08 2020].

[29] T. Alumäe, O. Tilk, Asadullah, „Advanced Rich Transcription System for Estonian Speech,“ *Human Language Technologies – The Baltic Perspective*, kd. 307, pp. 1-8, 2018.

[30] „"Eesti keeletehnoloogia: Baastehnoloogiad ja -ressursid" projekt EKTB24,“ Eesti Teadusinfosüsteem, 2018. [Võrgumaterjal] https://www.etis.ee/Portal/Projects/Display/51cb784d-b515-4aae-a1fd-92a4c5cee516. [Kasutatud 11 2020].

[31] S. Laur, „EstNLTK,“ Tartu Ülikool, [Võrgumaterjal] https://estnltk.github.io/.

[32] „udpipe - R package for Tokenization, Tagging, Lemmatization and Dependency Parsing Based on UDPipe,“ DataCamp, [Võrgumaterjal] https://www.rdocumentation.org/packages/udpipe/versions/0.8.5. [Kasutatud 04 2021].

[33] „UDPipe Models | ÚFAL,“ Institute of Formal and Applied Linguistics, [Võrgumaterjal]. Available: https://ufal.mff.cuni.cz/udpipe/1/models#universal_dependencies_25_models_performance. [Kasutatud 04 2021].

[34] „UD Estonian EDT,“ Universal Dependencies, [Võrgumaterjal]. Available: https://universaldependencies.org/treebanks/et_edt/index.html. [Kasutatud 04 2021].

[35] „Natural Language Processing for non-English languages with udpipe,“ R-bloggers, [Võrgumaterjal] https://www.r-bloggers.com/2018/01/natural-language-processing-for-non-english-languages-with-udpipe/. [Kasutatud 03 2021].

[36] „document_term_frequencies_statistics function - RDocumentation,“ DataCamp, [Võrgumaterjal] https://www.rdocumentation.org/packages/udpipe/versions/0.8.5/topics/document_term_frequencies_statistics. [Kasutatud 03 2021].

[37] „textrank_keywords function - RDocumentation,“ DataCamp, [Võrgumaterjal] https://www.rdocumentation.org/packages/textrank/versions/0.3.1/topics/textrank_keywords. [Kasutatud 03 2021].

[38] „keywords_rake function - RDocumentation,“ DataCamp, [Võrgumaterjal] https://www.rdocumentation.org/packages/udpipe/versions/0.8.5/topics/keywords_rake. [Kasutatud 03 2021].

[39] Z. A. Merrouni, B. Frikh, B. Ouhbi, „Automatic keyphrase extraction: a survey and trends,“ *Journal of Intelligent Information Systems,* kd. 54, pp. 391-424, 2020.

[40] P. Ludwig, M. Thiel, A. Nürnberger, „Unsupervised Extraction of Conceptual Keyphrases from Abstracts,“ %1 *Lecture Notes in Computer Science*, Springer, Cham, 2016, pp. 37-48.

[41] E. Papagiannopoulou, G. Tsoumakas, „Local word vectors guiding keyphrase extraction,“ *Information Processing & Management,* kd. 54, nr 6, pp. 888-902, 2018.

[42] Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, L. Panpan, „A Graph-based Approach of Automatic Keyphrase Extraction,“ *Procedia Computer Science,* kd. 107, pp. 248-255, 2017.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Liina Heinluht

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Võtmesõnade ekstraheerimine eestikeelsest transkribeeritud tekstist“, mille juhendaja on Ahti Lohk
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

10.05.2021

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

Lisa 2 – UDPipe mudeli tulemuste vaade RStudios

	doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	upos	xpos	feats	head_token_id	dep_rel	deps	misc
1	doc1	1	1	Tekst, millele so...	1	Tekst	tekst	NOUN	S	Case=Nom|Number=Sing	0	root	*NA*	SpaceAfter=No
2	doc1	1	1	Tekst, millele so...	2	,	,	PUNCT	Z	*NA*	4	punct	*NA*	*NA*
3	doc1	1	1	Tekst, millele so...	3	millele	mis	PRON	P	Case=All|Number=Sing|PronType=Int,Rel	4	obl	*NA*	*NA*
4	doc1	1	1	Tekst, millele so...	4	soovitakse	soovima	VERB	V	Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Pass	1	acl:relcl	*NA*	*NA*
5	doc1	1	1	Tekst, millele so...	5	eeltöötlust	eeltöötlus	NOUN	S	Case=Par|Number=Sing	6	obj	*NA*	*NA*
6	doc1	1	1	Tekst, millele so...	6	rakendada	rakendama	VERB	V	VerbForm=Inf	4	xcomp	*NA*	SpaceAfter=No
7	doc1	1	1	Tekst, millele so...	7	.	.	PUNCT	Z	*NA*	1	punct	*NA*	SpacesAfter=\n

Lisa 3 – Stoppsõnade loend

aga
ei
et
ja
jah
kas
kui
kõik
ma
me
mida
midagi
mind
minu
mis
mu
mul
mulle
nad
nii
oled
olen
oli
oma
on
pole
sa
seda
see
selle
siin
siis
ta
te
ära

Lisa 4 – Raadiosaate parandatud transkriptsioon

Raadiosaate transkriptsiooni parandatud ning lisatud sõnad on märgitud rasvases kirjas.

Vestleja nimi	**Vestlustekst**
Meelis Süld	Tervist, head kuulajad tänases tervise teemalises Huvitajas ja õppeaasta algul uurime, mida teha **nohuste** lastega, kes tavaolukorras saadetaks kooli või lasteaeda. Aga koroonaviiruse levikuga seoses tuleb olla ettevaatlikum. Lähemalt selgitab olukorda perearst Vanda Kristjan. Teine teema, kui palju me teame **SARS-CoV-2** levimise kohta siseruumide õhus ja millised on soovitused seoses ruumide **ventilatsiooniga** nüüd? Sellest teemast räägib meile professor Jarek Kurnitski Tallinna Tehnikaülikoolist. Saate lõpuosas tuleb stuudiosse Novaatori toimetaja Maarja **Merivoo-Parro** ja räägib, kuidas võib olla seotud läbipõetud **koroonahaigus** ja juuste väljalangemine. Mina olen saatejuht Meelis Süld, helipuldis on abiks Maarika **Leetmäe**. Head kuulamist ja kaasamõtlemist.
Meelis Süld	**Tänases** teisipäevases tervise teemalises Huvitaja saates räägime veidi koroonaviiruse olukorrast, aga selles võtmes, et nüüd, kui õppeaasta algab, siis lapsed lähevad kooli. Me teame, et varasemalt, kui lapsel oli nohu või või kerge köha, et siis ega teda koju ei jäetud, viidi ikkagi **seltskonda**, ta läks kas lasteaeda või või kooli, aga mida nüüd siis sellel õppeaastal teha? Meil on stuudios külas perearst, Vanda Kristjan, tervist.
Vanda Kristjan	Tervist.
Meelis Süld	Alustuseks, milline see olukord ühes perearstipraksises hetkel välja näeb, et vastu **esmaspäeva** oli, et kolm uut haigusjuhtu nüüd viimased andmed peaksid ka siin lähema tunni jooksul eelmise ööpäeva kohta saabuma. Aga milline see reaalne olukord perearstikeskuses on?
Vanda Kristjan	Perearstikeskuses hetkel on suhteliselt rahulik muidugi võib-olla võrreldes juulikuuga või augusti algusega on ikkagi viirushaiguste sümptomitega **patsientide** hulk suurenenud ja me iga nädalaga aina noh, rohkem ja rohkem küll tasapisi. Suuname siis nii nagu tänasel päeval kehtiv reegel ette näeb, et ülemiste hingamisteede infektsioonitunnustega kõik patsiendid peaksid minema testimisele ehk siis **Covid testile** suunamine, et see ei ole mingi keeruline protsess, tuleks pöörduda ikkagi oma perearsti keskusesse ja, ja esimene samm on ikkagi testimine.
Meelis Süld	Aga nüüd seesama kooliaeg kui varem oli selge, et mingisuguse lihtsama probleemiga me ei hakka ju last koju jätma, palavikku tal ei ole näiteks. Aga mida nüüd siis teha selles uues olukorras? Ja, ja vanemate jaoks, kes käivad tööl, on see ju tõsine küsimus, et mida nüüd siis teha?

Vanda Kristjan	No mitte ainult koolilapsed, ka lasteaialapsed, ka täiskasvanud, kes tööl käivad tänases olukorras, kus meil on juba kevade esimene kogemus olemas, on ikkagi see, et väikseimgi sümptom **ülemistes** hingamisteedes peaks olema testitud, et jääks me ise, et me oskaks hoiduda ja oskaks ka piisavalt kaua siis kodus olla. **Et ega** teisi viiruseid on ka, et meil ei ole ainult **koroona**, aga me peame nagu selle haiguse välistama, et siis me suudame nagu vastavalt oma tavapärasele harjumusele julgemalt, siis ütleme kerge nohu **sümptomiga**, võib olla ka nädala pärast juba kollektiivi noh, julgeks minna. Et siin on oluline see, et kui on sümptomid sõltumata vanusest, me saadame ikkagi testimisele, et me **ise teaks** ja ka, et teaks, ütleme meie **ka** tervishoiusüsteem ja riik, et mis toimub meil antud olukorras või antud hetkes. **Et** võtta kasutusele ka siis lisaennetusmeetmeid, et see on nagu vastutus iseenda oma pere ees, aga ka kogukonna ees, et see on hästi oluline, et ikkagi testida. Ja kui see test on negatiivne, siis ägedate haigussümptomite puhul on alati õige ikkagi koju jääda, et anda võimalus organismil paraneda sellest. Ja ainult, et kui **Covid** positiivsuse puhul on kaks nädalat kohustus kodus ära olla, siis, kui see on negatiivne siis võib kaaluda ka varem inimeste sekka minemist.
Meelis Süld	Ja alati pole ka lihtne ju märgata seda **kergemat** külmetushaiguse või siis viirushaiguse **sümptomit**.
Vanda Kristjan	Ja no ega absoluutset garantiid me ei saagi anda, sellepärast ongi see, et me ikkagi potentsiaalselt käituks ise iga sammuna, nii et me ikka väga lähestikku ei ole, väga suurtes rahvamassides ei ole, sest et me ju täna teame seda, et väga sageli on haigus, eriti lastel, noorematel lastel haigus positiivsed nad on, aga nendel ei ole mitte mingeid sümptomeid, ehk siis sümptomi vaba haiguse kulg on täiesti võimalik ja ja sellepärast ka, et, et ikkagi me peaksime võtma kui potentsiaalset ohtu, ehk siis kaaluma igat oma sammu. Kas see on meie jaoks oluline, kas me peame minema kusagile kogunemisele või siis kus me noh, vahetult oleme lähestikku teiste inimestega võõraste inimestega pere on pere, et seal on loomulikult see, see suhtlus ikkagi perekeskkonnas, nii nagu ta on aga just, et sellised võib-olla meelelahutustegevused massiüritused, kus on väga emotsionaalne see asi, et siis võib-olla praegu ikkagi veel ei ole mõistlik selles osaleda.
Meelis Süld	Võtame praktilise näite on nelja-aastane laps, näiteks hommikul on nina tatine. Mida nüüd see ema või isa peaks tegelikult tegema?
Vanda Kristjan	No võib-olla veel enne, kui see nelja-aastane laps kodus tatine on, mõelda täna läbi, siis kui kõik veel terved on, et mida see pere siis tegelikult teeb? Kindlasti see laps peab jääma koju kindlasti see pere peaks võtma ühendust oma perearstiga ja kindlasti **perearst** andes selle suunamise kutsutakse nad testimisele sinna ka minema seda, kes nüüd koju lapsega jääb, kuidas see on, see on ikkagi läbimõtlemise koht, sellepärast et täna me ei saa või ei ole õige moraalselt kaasata oma vanavanemaid sellesse protsessi. Sest kui me teame tänaseks teadaoleval informatsioonil, mida on siis uuritud ja analüüsitud, et ikkagi lapsed-noored põevad neid haigusi kergemini ja see riskigrupp, kellel on siis ohtlik see eluohtlik see haigus võib-olla et need on ikkagi küpsemas eas ja **vanemas eas** inimesed või need, kellel on mingid

	kaasuvad haigused. Et sellepärast see täna ei ole see vanaema, see kõige õigem inimene või vanatädi, kes siis võiks selle lapsega jääda, kas on mingeid teisi võimalusi, kas on, ma ei tea, nooremaid õdesid-vendi on või on kellelgi veel sugulastel **lapsi samas** olukorras, et siis võib-olla seal on seal noh, **kokkulepete koht** või, aga noh, iga pere peaks läbi mõtlema, et mis siis teeb, kes jääb koju ja kojujäämine on kindlasti esimene samm. Igal juhul.
Meelis Süld	See lapsevanem võtab haiguslehe.
Vanda Kristjan	Jah, kui tal on, no see on ka nii või ju varemgi olnud, et kui ikkagi haigussümptomitega laps on, et siis ja laps on selles vanuses, kus teda ta vajab hooldust siis ta kindlasti lapsevanem saab selle hooldusleh. Ja kaks nädalat on ta ka varem olnud, sest et laste vanuses **noh** lasteaia ja koolilastel küll vähem on see kaks nädalat, ongi tõenäoliselt selle pärast kaks nädalat, et see on tavaliselt sellise, ka sellise keskmise raskusega viirushaigusest paranemiseks piisav aeg.
Meelis Süld	Ja testitakse ära siis terve perekond.
Vanda Kristjan	Ei, testitakse ära see, kellel on **haigussümptom**.
Meelis Süld	Ehk siis nelja aastane laps läheb testimisele, võetakse nina kaudu see proov.
Vanda Kristjan	Ninaneelus kaabe ei ole meeldiv. Meditsiinis ei olegi meeldivaid asju, eriti et ka vaktsineerimine või, või mingi analüüsi võtmine. Ka vereanalüüsi võtmine ei ole meeldiv protseduur. Aga ta on ebameeldiv, üleelatav, ma olen seda kogenud.
Meelis Süld	Nii et kui see laps on testitud ja kui on negatiivne, läheb suhteliselt lihtsasti.
Vanda Kristjan	Siis on vastavalt sümptomitele, kuidas **laps** on valmis kollektiivi minema, et ta nagu suudaks osaleda selles kollektiivi tegevustes **kas** siis õppimises või, või lasteaiatöös siis nagu on ikka nii, nagu tavaviirushaigus on, ja varem on, et perearstide selts on **ju** välja töötanud siin eelmine üle-eelmine aasta soovitused, et ikkagi lapse üldseisund on see, mis annab võimaluse kollektiivi minna või mitte minna. Ja kindlasti palavikus **laps** ei ole kollektiivi mõeldud.
Meelis Süld	Ja kui nüüd see laps on koroonaviiruse osas positiivne, siis on kaks nädalat kindel, et tuleb olla kodus.
Vanda Kristjan	Jah nii see on ja seal ei ole midagi päev sinna-tänna ja seal ei sõltu sellest, kas sümptomid on või ei ole. See on uuringutel põhinev ja see on selline ta, noh, meditsiinis üldse **sajaprotsendilisi** garantiisid ei ole, aga see on see tõenäosuse asi, et kui see kaks nädalat on ikkagi kodus ära põetud, siis on üpris tõenäoline, et enam ta ei ole. Risk teistele.
Meelis Süld	Kas last siis uuesti testitakse või mitte?
Vanda Kristjan	Ei ei pea testima.
Meelis Süld	Kuidas mõjutab see olukord nüüd perearstikeskuste tööd et kui neid **nohuseid köhaseid** lapsi on palju siis noh, neid haiguslehti tuleb ka **ilmselt** vormistada päris palju. Et kas te saate **hakkama**?

Vanda Kristjan	No ma usun, et me oleme karastunud küll ja küll, et mis on võib-olla tavavõõrast või varasemast erinev, mida ma tahaks väga südamele panna, on see, et et kallid patsiendid ärge tulge **ette** leppimata perearstikeskustesse. See tõstab riski. Et kui inimene arvab, et tal ei ole midagi, aga ta homme on ja ta võib-olla juba potentsiaalselt ohtlik. Et lepime kokku nii, et helistamine või e-meili teel või mingi muu võimalus, mida on võib-olla perearstikeskus võimaldanud, noh, see on selline, iga perearst on sihuke või perearstikeskus oma näoga ja oma tavatöörütmis, aga täna on oluline, et me lepime kokku, millal ta tuleb ja ta tuleb sellele õigel ajal ehk siis, et ta ei ohustaks teisi ja **teised ei** ohustaks teda. Kindlasti perearstid jätkavad oma tööd jällegi sõltuvalt sellest, kuidas on ruumi võimalused, kui palju personali on, aga hoitakse lahus viirusinfektsioonitunnustega patsiendid ja muude haigusmuredega patsiendid. **Et kuidas** keegi keskuses seda organiseerib, ongi noh, iga keskuse otsustada, aga kindlasti ärme ohusta personali, ärme ohusta ise ennast, et, et lepime kokku telefoni teel ja lähme sellel õigel ajal kohale.
Meelis Süld	Ja testimine on praegu korraldatud, siis mismoodi, kus kohas seda testi koroonaviiruse suhtes saab teha?
Vanda Kristjan	Et see test käib siis ikkagi endiselt nii nagu ka kevadel käis, ta käis tegelikult suvi läbi siin, et ainult see vajadus oli väiksem. Perearst saadab infosüsteemi kaudu saatekirja, et testkeskus võtab ühendust patsiendiga. Nüüd vahepeal neid oli vähem neid testkeskuseid, sest et ei olnud nii suurt nõudlust, eks vastavalt sellele vajadusele **neid testkeskuseid** võib-olla jälle lisatakse või, või nende läbilaskvus võime, suurendatakse, et see kõik näitab aeg, kuidas meil seda vajadust on, et, et sellele vastavalt siis korrigeeritakse seda tööd.
Meelis Süld	Tuleb kohale minna, tullakse koju, sõltub olukorrast.
Vanda Kristjan	Jah, need inimesed, kes ikkagi liikuda ei saa, ei ole võimalik, mis iganes põhjusel, see on kokkuleppeline, aga ma usun, et enamus inimesed siiski ka teistkeskustesse **jõuavad** ja nii nagu ta oli ka varem, et kas on nüüd noh, seal on kokkulepe, et praegu suvel näiteks ei olnud seda Drive In süsteemi, aga seda lihtsalt nii vähe oli neid inimesi, et kui nad ju ka üksinda seal tänaval seisid teistest eemal, et siis ta **nüüd** ei ohustanud kedagi. Aga ma usun, et seal ka võivad need jällegi see disain muutuda sõltuvalt sellest töökoormusest.
Meelis Süld	Koroona viirushaiguse sümptomite kohta veel, et kas lastel ja noortel on ka mingisugused eriomased sümptomid, kas on võimalik noh, hinnata, et ja vat lapsel võib nüüd olla tõepoolest **koroona**.
Vanda Kristjan	Ei, mina ei jääks nende asjade juurde, sellepärast et elu on näidanud, et see on niivõrd **variaabel**, et seal võib olla ükskõik mis muu viirus ja võib-olla ka **koroona** seal alates ütleme, Kellel kurk kõditab, kellel nina sügeleb, kellel kõht valutab ja seal on hiljem selgunud, et on **koroonaviirus**, kas see on nüüd olnud selle sümptomi põhjus või mitte. **See on väga** keeruline, sest et ega ei ole sellist reeglit, et üks viirust saab inimesel korraga olla, ega nad võivad ka seal kombineeruda, et mida saab ikkagi kindlalt öelda, on see, kui

	inimene on testitud. Nii võimalik kui see on loomulikult seal on omad nüansid veel. Et kui hästi õnnestub see analüüs võtta kui õigesti, et võib ju ka tulla väikseid vigu, aga ma usun, et see proportsionaalselt ikkagi valdav enamus seda informatsiooni on tõene, mida me saame, kui me tõesti ära teeme.
Meelis Süld	Ja see kodune ravi lapse puhul näiteks on siis missugune.
Vanda Kristjan	Nii nagu ikka, vot mis on viirushaigus, on viirushaigus ja see on teada-tuntud tõde, et viirushaigusele väga spetsiifilist välja arvatud gripp ravi ei ole, et sümptomi ravi, et sümptomi ravi tähendab seda, et kui mul kurk valutab, siis ma üritan kurguvalu leevendada. Kui mul on nina kinni, **siis** üritan nina limaskesta turset vähendada ja **kui mul** on köha, siis ma üritan siis köha leevendada, ehk siis mis annaks mulle mu organismile tuge üle elada, see aeg, kui organism ise selle viirushaigusega ikkagi hakkama saab ja minu immuunsüsteem selle ära pärsib. Nüüd gripp on ainus erand ja miks ta on erand, on see, **et gripp** on vana tuntud haigustega on ju nii kaua meiega olnud ja kuna gripp on ka ikkagi keskmise või raske kuluga haigus, mis ka võib lõppeda surmaga siis seetõttu on tal vaktsiini võimalus ja on ka ravimivõimalused, kui ikkagi ei ole vaktsineeritud ja patsient jääb haigeks ja tal on gripp, ehk siis testidega on see kindlaks tehtud. Et siis on võimalik ka tablettraviga seda ravida. Aga see on ka ainult ma ütlen, et ta on meie jaoks tuntud haigus ja ta on pikka aega olnud ja kuna ta on võrreldes teiste viirustega, mis noh, kerge nohu või köha mis ei ole **ohtlik**, ei muutu kunagi, siis see gripp on selles suhtes erand. Et see **Covid** on ka ikkagi noh, **ohtlikumgi veel** kui gripiviirus, nii et ma loodan, et aeg töötab meie kasuks. Et saavad olema ka võimalused, et saada nii ravi kui ka ennetust.
Meelis Süld	Sellesama gripivaktsiiniga seoses, et nüüd on just ju algamas gripivaktsiini tegemise hooaeg või, või need vaktsiinid peaksid siis ka ju jõudma Eestisse. Et kas on mõistlik see kaitsesüst ära teha, sest noh, mõnes mõttes jällegi meil on nagu siis üks mure vähem või võime loota, et, et me vähemasti siis gripi ei jää.
Vanda Kristjan	Just et ma olen väga nõus sellega, et meil on nii palju haigusi, mida me ennetada ei saa. Et miks me jätame ennast ilma ennetusest, mis meil on võimalik teha. Et ma olen kakskümmend aastat ühes kohas perearstina ametit pidanud, kõik kakskümmend aastat, olen ennast vaktsineerinud ja oma pereliikmeid ja ma ütlen, et kui mul noh, viimane talv mul isegi nohu ei olnud, et, et kui mul kerge nohu on, et siis see on ka kõik, ma ei karda **grippi**, ma teenindan oma patsiente ja mul on üks konkreetne kogemus isegi näiteks tuua, et oli juba kevadekuu märtsikuu ja tuli väike laps vastuvõtule, gripp oli juba taandumas, noh juba ei olnud nädalaid olnud patsiente, kes gripidiagnoosi sai ja lapse läbivaatusel ta aevastas või **köhatas** mulle silma oma lima. Ja kui ma pärast saatsin ta testile, selgus, et tal oli nii A kui B-gripp. Ma mõtlesin seda, et nii, et no nüüd ma ikkagi jään haigeks, et see oli nii vahetu ja ka massiivne, sest lapsel olid väga ägedad sümptomid. Ja ütlesin, et tuleb, mis tuleb, et mina olen kõik teinud, et terveks jääda. Ja õhtul veel ma tundsin, et midagi vist hakkab tulema, aga hommikuks ma olin terve, mul ei olnud mingeid sümptomeid, kõik taandus

	ära, et ja see oli üks **reaalne**, minu kogemus selles, et see õigustas ennast. Ja, ja ma, see ei ole isegi koht, mida ma mõtlen, kas teha kindlasti teha, et kui üks haigus vähem on, see on suur asi, siis kui sellest praegu nagu ma ütlesin enne ka, et gripp on ka surmaga, noh, võib lõppeda surmaga ja kui me mõtleme nagu laiemalt, et püüame olla inimlikud, siis mõtleme sellele, kui palju on meie ümber ikkagi haigeid inimesi, kes ei saa ennast vaktsineerida. Siis võiks gripp olla nagu see väike mure, et **vaktsineerime** siis ennast, et mitte levitada seda.
Meelis Süld	Ja siis muud sellised tervisesoovitused, et kas me peaksime püüdma olla nii-öelda terved, **no** nii palju, kui see meie võimuses on sööma siis nüüd ma ei tea, mis vitamiine sisse ja, ja kõiksuguseid preparaate, jätma ujumistunnid äkki ära, sellepärast et tihti pere laps on nohune pärast ujumistundi, kui on selline **kogemus juba** olnud, mida soovitada?
Vanda Kristjan	No see, mida teie loetlesite, see kindlasti ei hoia meid tervena selles mõttes, et **vitamiine** tuleb süüa siis, kui meil on vitamiinipuudus, kui me oleme toitunud nii nagu on õige ja mitmekülgselt, siis tänapäeval me saame oma vitamiini toidust kätte. Et arvates, et me nüüd võtame mingi lisaasja, et see nüüd tõstab mingi lisaväärtuse annab või tugevuse ei ole tõendust leidnud, et ikkagi, kui me toitume õigesti, kui me puhkame, kui me oskame oma stressiga hakkama saada, kui me liigume värskes õhus piisavalt, et selle ujumise koha pealt mina ütleks seda, et kindlasti mina ei soovitaks seda ära jätta. Sest et **külmetus**, noh, kui inimene saab külma, ei tee inimest haigeks, haigeks teeb inimeste ikkagi viirused, mis tal limaskesta peal on, aga külmetus soodustab selle viiruse nagu paljunemist, et noh kaitsevõimet natuke vähendab jah, tuleb jälgida, et oleks vastav **riietus**, et ei oleks liiga palav, et ei oleks külm, et võib-olla peale ujumist ikkagi kuivatada juukseid ära. Vesi on selline asi, mis pigem uhub, need liigsed viirused **ninast** välja, et, et kõige muu jamaga, et mina nagu ei kindlasti ei soovita loobuda ujumisest värskes õhus sportimisest, trennides käimisest, et see on pigem, on ju see, mis meie immuunsüsteemi karastamine on oluline, et praegu on veel selline soojad ilmad, et pigem on karastamine õige, et me sellise **iga** jahedama õhu peale kohe haigeks ei jääks. Seda, et meil viiruseid on kogu aeg limaskesta peal olemas nad ongi seal aga lihtsalt see, et kas see limaskesta peal olev viirus suudab hakata nagu paljunema rakku sisenema, sõltubki selles, et kui me oleme nõrgad, kui me oleme nõrgestatud, meie immuunsüsteem ei suuda neid ära pärssida, siis me jäämegi haigeks, selleks võib olla üleväsimus ja stressreaktsioon mingi **kontrolltöö näiteks** või on mingi magamata öö või mis iganes seal noh, et mis nagu võib nõrgestada, aga see, et, et meil viiruseid seal limaskesta peal noh, et me oleme puhtalt sellest **seda me** ei saa **loota**. Need on alati keskkonnas, on suvel ka, ainult et me oleme, me liigume väljas, me oleme positiivsed, me puhkame ja **me ei jää** haigeks ja me oleme õues rohkem, eks ole, et, et miks on sügis ka selline, et see on nagu soodne selle viiruse jaoks keskkond soodsam. Ja ka me oleme ruumis rohkem sees, me oleme rohkem üksteisele lähedal. Et lapsed ninapidi koos, et, et see tõenäoliselt määrab selle sessoonsuse ka.
Meelis Süld	Nii et vähem muretsemine teeb meid küll tervemaks.

Vanda Kristjan	Kindlasti see on õige otsus.
Meelis Süld	Ja aitäh tulemast täna Vikerraadios Huvitaja **saatesse** perearst Vanda Kristjan.
Vanda Kristjan	Aitäh, olgem terved ja olgem negatiivsed, siis **koroonanegatiivsed**.
Meelis Süld	Koroonaviiruse teemal veidi ka jätkame ja räägime sellest, kui palju me teame viiruse levimise kohta õhu kaudu ja meil on külas Tallinna Tehnikaülikoolist professor Jarek Kurnitski, kes on just selle valdkonna ventilatsiooni sisekliimavaldkonna ekspert. No kui palju me tõesti teame, et pindadest on palju räägitud, neid me hoolega **desinfitseerime**, puhastame käsi, aga, aga viiruse levimine just õhu kaudu ja, ja siseruumides.
Jarek Kurnitski	Enamus ruumid on siseruumid ja võib-olla mind ongi natukene häirinud, et kõnepruuki on tulnud **siuke** termin nagu suletud ruumid. Me siin raadios istume ka suletud ruumis ja hooned ongi tehtud selleks, et inimesi ilmastiku eest kaitsta, järelikult nad peavad suletud olema, aga noh, loomulikult need suletud ruumid on siis kas on sisekliima tagamisega või ilma sisekliima tagamisega. Ja kui ventilatsioon ei ole, siis on, asjad on halvasti. Aga kui nüüd vaadata, et kuidas **see** viirus on levinud ja kuidas need nakkused on saanud saadud, siis on ikkagi kaks põhijuhtumit, mis saab välja tuua kus inimesed on väga lähestikku koos ehk siis niinimetatud lähikontaktis. Kindlasti on saadud väga-väga palju ja seda võib saada ka välisõhus mõnel kontserdil, kui inimesed trügivad külg külje kõrval, et selleks ei pea kindlasti olema siseruumis, et kindlasti see füüsilise distantsi nõue seda, seda tuleks hoida ja see on väga oluline. Ja teine koht, kus siis neid nakkusi on saadud, on ülerahvastatud siseruumid, kus on puudulik ventilatsioon. Et meil joonistub nagu kaks juhtumit välja, et kui me oleme kellegagi ninapidi koos, siis on oht väga suur ja teine väga suur oht on siis, kui me oleme ka kaugemal, näiteks viie või kümne meetri kaugusel nakkuskandjast, aga ruum ei ole ventileeritud, et siis me võime samamoodi selle viiruse nakkuse saada.
Meelis Süld	Nüüd, kui me toome selle uuringute poolega siia juurde, et väga palju tehakse praegu selles valdkonnas uurimistööd, et teada saada, kuidas nakkus levib ja, ja mida teha, et, et vältida mis on võib-olla seoses just sisekliimaga olnud sellised kõnekamad näited uuringute poole pealt.
Jarek Kurnitski	No kindlasti kõik teavad seda **Diamond Printsess** kruiisilaeva või siis mõningaid koorilaulujuhtumeid, mis on üle maailma toimunud ja tegelikult need kirjeldavad selliseid inimkatseid, mida tavaliselt uuringutest teha ei saa, sest see on lihtsalt ebaeetiline. Aga kuna need asjad on niimoodi juhtunud, siis nad on teadlastele väga oluline materjal või siis ka siin näiteks Saksamaal üks lihatööstuse juhtum, kus **sammuti** ilma **ventilatsioonita** tehases saadi väga palju nakkusi. **Et** sellised sellised juhtumid on, on kindlasti palju avalikkuses olnud ja teadlased neid põhjalikult uurivad ja sealt saadakse ka väga olulist informatsiooni.

Meelis Süld	Ohumärk on siis see, kui ventilatsioon on puudulik, et ööklubid, baarid, noh, muidugi, võib-olla ööklubis ka väga hea ventilatsioon, aga, aga no mitte ka alati, et siin tulevadki siis ka mitmete hoonete puudused välja.
Jarek Kurnitski	Tulevad puudused välja, aga no nüüd üks oluline faktor on veel, mis on välja tulnud ja mida võib-olla teiste viiruste puhul ei ole tähele pandud või ei ole nii palju uuritud, on see, et kui inimesed karjuvad või laulavad või räägivad siis kõrgendatud hääletoonil, et siis me paratamatult puhume rohkem õhku välja, mida rohkem õhku välja puhume, seda rohkem on seal neid väikeseid **piiskasid** ja aerosoole ja see nakkusallika intensiivsus võib-olla tuhat korda suurem, kui siis rahulikult istuval inimesel. Ja selline tuhande kordne vahe on siis ka midagi sellist, et millega ventilatsioon muutub võimetuks, et siis meil peaks olema nagu tõmbekapp pea kohal, et siis oleks seda võimalik eemaldada, aga no ilmselt ööklubides juhtub tihti, et need halvad asjad satuvad kokku, inimesed on väga lähestikku, siis seal on nii vali muusika, et nad peavad karjuma, et suhelda, mis tähendab, kui kellegil tõesti, sest nakkus on, siis ta levitab seda väga-väga korralikult ja võib ka juhtuda see, et mõnes ööklubis ei ole ventilatsioon töökorras. Et seal võib tegelikult kõik need kolm halba asja võivad nagu kokku sattuda, kui, kui niimoodi läheb. Aga nagu näha, sellel on nüüd leitud väga tõhus lahendus, et kui alkoholi müüki piiratakse ja ilmselt **peaks veel** neid detsibelle ka piirama, et inimesed saaksid vaiksemalt rääkida, et siis, siis võib-olla on see probleem kuidagi nagu lahendatav, aga aga see, see on nagu üks ilmekas näide sellest, et kui, kui kolm halba asja satuvad kokku, siis see tõenäosus on väga suur ja mitmes kohas Eestis seda on juhtunud ja loomulikult igal pool mujal maailmas samamoodi.
Meelis Süld	Kui ma võtan nüüd koolikeskkonna peale, siis noh, küllap on ka koolimaju kenasti renoveeritud ja ja on tehtud seal ventilatsioonisüsteeme paremaks, aga on ka vanu koolimaju, kus ei ole nii hästi ventilatsioon lahendatud. Koridorides võib-olla ei ole jällegi nii hea ventilatsioon kui klassiruumis lapsed jooksevad, karjuvad. Tundub üsna ohtlik selle eelneva näite valguses.
Jarek Kurnitski	No siiski, koolimajad on üks, võib-olla kõige ohutumates kohtadest sest koolimajade puhul enamikel koolimajadel siiski on ventilatsioonisüsteemid ja ma loodan, et suve jooksul need on nüüd ka üle kontrollitud, näiteks Tallinna Tehnikaülikool on ka kõik oma **ventilatsioonisüsteemid** üle kontrollinud ja veendunud, et nad on töökorras ja leidnud üles ka mõne audika, kus siis ei ole korraliku ventilatsiooni ja püüab neid vastavalt vähem kasutada. Loodan, et kõik hoonete omanikud ja, ja eriti koolimajades on sellesse panustatud, sest noh, selleks on ka riik tegelikult kohustanud on olemas isegi selline eriolukorra ventileerimise määrus, mis siiamaani kehtib ja see on väga oluline, et me me teame, missugune ventilatsioon meil on ja mis seisukorras need süsteemid on ja koolimajade puhul on. Mis teeb selle olukorra paremaks, on see, et klassiruum on üks hea suur ruum, see on selline kuuskümmend ruutmeetrit, seitsekümmend ruutmeetrit, seal võib olla kolmkümmend inimest ja seal on kolmekümne inimese jagu seda õhuvahetust ja kui siis sinna satub üks nakkuskandja siis **lahjeneb** seal

	kenasti ära ja on väga suur tõenäosus, et sellises ruumis keegi ei saagi seda nakkust endale külge. Ohtlikumad on, palju ohtlikumad on väikesed ruumid, näiteks me siin raadio stuudios istume praegu kolmekesi ruumis ja siin ilmselt on ka kolme või nelja inimese jagu seda ventilatsioonimäära sest on tegemist väikese ruumiga ja kui sellises väikeses ruumis siis on üks nakkuse kandja, siis, siis need kaaslased kindlasti ka selle nakkuse saavad väga kergelt. Sest ventilatsiooni summaarselt on lihtsalt vähem võrreldes klassiruumiga, seal on kümme korda rohkem kui näiteks siin ruumis ja kui on kümme korda rohkem, see on kümme korda rohkem lahjendatud kümme korda väiksem kontsentratsioon kümme korda väiksem doos ja suure tõenäosusega seda nakkust lihtsalt ei saa. Nii et koolid on tegelikult üks kõige ohutumaid kohtasid ja veel siis lisaks selle viiruse eripära, millest viroloogid ja meditsiiniteadlased palju on rääkinud, et mida väiksemad lapsed, seda vähem nagu külge jääb, et noh, see on veel teine aspekt, aga puht nagu ventilatsiooni-tehniliselt on klassiruum väga ohutu koht.
Meelis Süld	Seoses talve tulekuga, külmemate ilmade saabumisega ventilatsioon on ju seotud ka mõnes mõttes meie küttearvetega ja, ja võib-olla siis seda ventilatsiooni ka ei, ei kasutata nii tõhusalt selleks et küttearveid vähendada, et millised on praegu siin need suunised, arvestades viiruse leviku ohtu ja, ja, ja soove siiski ka kokkuhoidlikult majandada.
Jarek Kurnitski	Ja ka minule **meenuvad** aastatetagused, mingisuguseid juhtumeid, kus mõni koolimaja ei lülitanud oma ventilatsioonisüsteeme sisse selleks et energiat kokku hoida, aga siis kas nad kokku hoidsid või mitte, seda keegi ei tea, sest siis ikkagi avati aknaid ja, ja see välisõhu kütmine on ka väga kallis tegevus, et noh, võib-olla nad midagi kokku ei **hoidnudki aga nad** arvasid, et kui nad ventilatsiooni sisse ei lülita, et siis on võimalik küttekulu kokku hoida. Et, et selle tõttu on väga oluline, et need ventilatsioonisüsteemid lülitatakse **sisse**. Ja praegu on rõhutatud seda, et süsteemid pead, peaksid töötama natukene rohkem kui tavapärases olukorras ehk kaks tundi enne hoone kasutusaega juba sisse lülitada, tavaliselt on üks tund ja samuti siis kaks tundi käitada veel peale hoone kasutusaega, kui inimesed on ära läinud, et kindlasti oleks kõik välja ventileeritud, et mis, mis neid saasteaineid hoones võib-olla, nii et selline ventilatsioonisüsteemide käidu aja pikendamine on kõige lihtsam ja kindlasti ka kõige tõhusam ennetusmeede. Loomulikult need süsteemid peavad olema töökorras ja kui nüüd hoones ei ole korralikku ventilatsioonisüsteemi siis see on väga nagu juhtumipõhine **tegevus** ja siis tuleks, kindlasti tuleks hajutada, sest selge, et paari kuuga uut **ventilatsioonisüsteemi** ei suuda ehitada hoonetesse, aga ma loodan, et ka riik selle peale mõtleb ja, ja, või kui meil selliseid lasteaedasid või koolimaju veel leidub, et kus ei ole ventilatsioonisüsteemi, siis kiirendatud korras kindlasti need tuleks ette võtta.
Meelis Süld	Aga lahendus on siis akende lahti hoidmine või, või tuulutamine, eks ole, et muud muud varianti vähe väga näha ei ole.
Jarek Kurnitski	Jah, praegusel hetkel akna kaudu tuulutamine töötab veel väga hästi, väljas on soe, aga kindlasti külmal talvel, siis sellega hakkab õpilasi ära külmetama ja siis on jällegi teistpidi kaasnevad riskid, et siis tuleb ikkagi

	tuulutada vahetundide ajal ja tuleb seda teha intensiivselt. Aga kindlasti need hooned, kus ei ole sisekliima tagamist, ei ole töökorras, ventilatsioonisüsteeme seal see akende avamine. Noh, see, see tuleb teha õpetajate kohustuseks, kuigi õpetajatele selle eest palka ei maksta, et nad nüüd aknaid peaksid lahti kiskuma. Aga et selle viirusega kuidagi nüüd selline periood üle elada ja kannatada ära, **siis** sinnamaani, kuni need ventilatsioonisüsteemid välja ehitatakse, need remondid suudetakse ära teha. See kindlasti on raske periood ja on, on, tuleb mõelda siis juhtumipõhiselt, et kuidas sellega hakkama saadakse.
Meelis Süld	Kas ventilatsioonisüsteem peaks ka suuremal võimsuses **või** võimsusel töötama lisaks sellele, et ta varem sisse lülitatakse ja, ja siis kauem töötab pärast inimeste lahkumist?
Jarek Kurnitski	No see on tavaliselt olemasolevates ventilatsioonisüsteemides, et nad lülitatakse niinimetatud **nominaalsele** kiirusele, mis on siis projekteeritud võimsus ja kui sealt näiteks veel kümme protsenti tahaks seda õhuvooluhulka kasvatada, siis selle mõju on väga väike. Aga see võib kaasa tuua müraprobleeme ja paljudes süsteemides ei ole lihtsalt nagu võimalik. Küll on see võimalik sellistes süsteemides, mis on nõudluspõhised ja mida kasutatakse näiteks paljudes büroohoonetes ja, ja muudes mitteelamutes, kus ventilatsiooni juhitakse näiteks süsihappegaasitaseme järgi või temperatuuri järgi. Ja sellistes hoonetes on tõesti see soovitatud, et see Nõudlus **nõudluspõhisus** tuleb välja lülitada ja need süsteemid peaksid kogu aeg töötama nominaalselt kiirusel, nii et me olemasolevalt süsteemilt ei saa nõuda seda, et ta nüüd meile kaks korda rohkem õhku annaks. Aga me saame teda hoida nii-öelda täiskiirusel lihtsalt töös.
Meelis Süld	Kas on ka mingisugused, ma ei tea, kemikaalid õhu puhastamise, tehnilised muud vahendid. **Osoneerimine**, ultraviolettlampide paigaldamine, kas, kas need on kuidagi muud mõistlikud, vajalikud?
Jarek Kurnitski	Need muud vahendid on, on kindlasti. Igalühel on oma kasutusotstarve. Nüüd üks president ka naljakalt ütles, et neid inimesi võiks proovida ka desinfitseerida nagu seestpoolt, et saaks viirusest lahti, aga kahjuks inimene sureb ära, kui sellist asja teha, nii et see väga-väga tõhus lahendus ei ole. Aga kui nüüd tõsiselt rääkida, siis on, on sellised ventilatsioonisüsteemid, kus on **tagastusõhu** kasutamine ehk niinimetatud **retsirkulatsioon**. Ja näiteks mõningates kauplustes selliseid süsteeme kasutatakse, kus siis õhuga köetakse ja samamoodi õhuga jahutatakse ja seal õhk ringleb. Ja siis seda **ringlusõhku** ehk see **väljatõmbeõhk**, mis tagasi suunatakse sissepuhkeõhuks. Selle filtreerimist on oluline parandada ja seal on kindlasti võimalik ultraviolett C-kiirguse kasutamine mida **küll** Eestis on äärmiselt vähe tänase päevani tehtud ja võib-olla ettevõtetel ei ole ka väga sellist võimekust seda pakkuda. Aga kuna me räägime siin ikkagi selliste peenosakeste filtreerimisest, sest need viiruse osakesed ei ole **paljad**, et nad on ikkagi nende peenosakeste sees, mida inimene suust välja hingab ja me räägime siin ühe mikromeetri üks kuni kümme mikromeetrit kuni viiskümmend mikromeetrit suurusjärgust. Nii et tavaliste välisõhu peenfiltritega on need osakesed välja filtreeritavad väga suures koguses. Nii

	et see on ka üks selline soovitus, mis on antud tagastusõhuga ventilatsioonisüsteemidele, et seal tuleb seda väljatõmbeõhu **filtreerimist** parandada, et paremad, lihtsalt tavalised paremad filtrit panna sellised filtrid, mis on tavaliselt siis välisõhupool ja see võib olla olemasolevates süsteemides on natukene lihtsam lahendus. Ja pikemas perspektiivis muidugi sellist **tagastusõhu** kasutamisest tuleks üleüldse loobuda, **sest see** ei ole hügieeniline ventilatsioonilahendus.
Meelis Süld	Saunad, kas need on ohutud paigad?
Jarek Kurnitski	See viirusele selline kõrge temperatuur ei meeldi tõesti, et kui kui, kui ma nüüd oma mälus **sobran** ja kui temperatuur on seitsekümmend kraadi, kas ta kümme minutit peab vastu või ei pea. Nii et kui ma vist viisteist minutit kaheksakümmend kraadi leili võtan, siis kindlasti seal ühtegi aktiivset viirust ei ole, nii et see, see juhtumisi on ohutu paik, siis selle viiruse jaoks.
Meelis Süld	Aurusaunaga vist sama lugu ei ole, kahjuks.
Jarek Kurnitski	Seal on ka temperatuur, on siiski ilmselt on piisavalt kõrge, et ta lihtsalt ei pea sellele kõrgele termotöötlusele vastu.
Meelis Süld	Vähemalt on mõned ruumid, kus saame rahulikult nautida ja ja, ja end turgutada, aga aitäh tulemast täna saatesse Jarek Kurnitski Tallinna Tehnikaülikoolist ja meie vaatame siis, et need ventilatsiooniseadmed ja sisekliima oleksid kaasaegsed ja, ja süsteemitu töötaksid hästi. Huvitaja saates teisipäeviti vahendame ka Novaatori tervise ja teadusuudiseid. Maarja Merivoo-Parro on meil stuudios. Millega sa meid täna rõõmustad?
Maarja Merivoo-Parro	Ma ei tea, kas ma rõõmustan, aga oma infot annan ikka **koroona** kohta. **Koroona** on ju veel väga uus haigus ja tema kulg ja ka järelmõjud on alles selgitamisel ja üks viimasel ajal enam tähelepanu saanud uutest võimalikest nähtudest on juuste väljalangemine.
Meelis Süld	Ära hirmuta.
Maarja Merivoo-Parro	Teadete järgi see tabab **eelkõige** inimesi, kes on koroona juba ära põdenud ja sellest näiliselt justkui terveks saanud, aga keda mõned sümptomid on siiski kummitama jäänud, selliseid inimesi olla päris palju. Kes nagu ei suuda seda haigust maha raputada. Põhjalikke teadusuuringuid veel sel teemal tehtud ei ole, pigem on meediasse jõudnud arutelud saanud alguse sotsiaalmeediast, kus selle konkreetse probleemi ehk juuste väljalangemisega kimpus endised ja praegused koroona patsiendid on leidnud kaaskannatajaid, kellega oma muret jagada ja muljeid vahetada ja **arstid** arvavadki, et tegelikult ilmselt see viirus ise siiski seda juuste väljalangemist otseselt ei põhjusta. Küll aga võib selleni viia stress, mis pandeemia kontekstis on nii füüsiline kui emotsionaalne ja mõjutab nii haigeks jäänud kui ka tegelikult neid, kes elavad nakatumise hirmus. Sest nagu on ju ammu teada, juuksed võivad tõesti stressi või šoki tulemusel välja langeda ka ilma **koroonata** näiteks lähedase inimese kaotuse, õnnetuse üleelamise, töökohalt koondamise või mingite majanduslike murede tõttu või kasvõi lõputöö kirjutamise tõttu, kui inimene ise seda

	tõesti väga hingega valuliselt üle elab ja keha ei vii seda protsessi läbi mitte mingi lisakaristusena vaid seetõttu, et ta soovib üliväga endal hinge sees hoida. Nimelt kui juhtub midagi väga suurt, midagi väga negatiivset, mis inimest põhjalikult raputab, siis keha saabki signaali, et käes on eriolukord, tuleb ressursse strateegiliselt kasutada ja keskenduda kõige tähtsamale, karvakasv ju ei ole kõige tähtsam. Ja nii enamasti väikse viivitusega hakkabki inimene karva ajama. Jah, see juuste kaotus iseenesest on üsna ebameeldiv ja võib samuti tunduda hirmutav ja juba põetavale stressile veel **vindi** peale keerata. Aga jah, tegelikult seda võivad põhjustada ka teised tervisehädad ja kui teid selline mure tabab, siis arst võib suunata teid ka hoopis **kilpnäärmeuuringule või** vaadata. Kuidas teil toitainete omastamisega on?
Meelis Süld	Ei tasu siis liigselt muretseda jällegi teist korda, ütleme tänases saates.
Maarja Merivoo-Parro	Jah, põhiline sõnum on jah, et kui, kui tahad võita pead karvasena, siis püüdke igal võimalusel stressi vältida, peaasi et see stressi vältida püüdmisest omakorda suur stressi **ei saa ja** üldse. Võib-olla tasuks mõelda ka, et noh, juuste peale üldse, et meie kultuuriruumis nad on nagu kuidagi hirmus olulised. Aga kui selle üle juurdlema hakata, siis äkki nad on liiga olulised see aeg ja energia, see raha, mis nende peale pannakse. Ja nüüd siis see aeg ja energia, mis nende üle põdemisele pannakse ka siis, kui nad on täiesti peas, aga võib-olla ei ole piisavalt lokkis või piisavalt sirged või piisavalt siledad või piisavalt paksud siis see kõik kokku lugedes on ikkagi väga suur. Samal ajal on terve seltskond väga väärikaid inimesi, kes saab hakkama ilma igasuguste juusteta. Ehk et me oleme võib-olla jah, need soengud ja sära ja läike ja **lopsakuse** ja värvitoonid asetanud kuskile ideaalide maailma. Ja, ja arstid tõesti on ka tõdenud, et juuste kaotamine mõjub patsientidele väga **demoraliseerivalt**. Ja muidugi noh, see juuksekaotus nüüd nende **Covid-19 põdejate** puhul on pigem ajutine aga selle mõju on pikk, sest juuksed teadupärast kasvavad aeglaselt ja seda endist paksust ja pikkust võivad kannatanud oodata aastaid.
Meelis Süld	Aitäh tulemast täna saatesse Maarja **Merivoo-Parro** portaalist Novaator. **Küllap** on neid, kellel veel puhkus kestab ja kellel ka ei kesta, siis pärast tööpäeva lõppu on võimalik minna nii mõneski kohas veel suplema **või** ennast vähemalt karastada. Üheksateist kraadi on mõõdetud veetemperatuuriks siin hommikul kell kümme Peipsi Kauksi rannas. Pikakari ja Pärnu rand näitavad ka üheksateist kraadist temperatuuri, aga Pärnus on miskipärast punane lipp väljas. Kaheksateist kraadi on merevee temperatuur Haabneeme Stroomi ja Võsu ja Pirita ranna puhul. Pirital on kollane lipp väljas. Harku järves on samuti veetemperatuur kaheksateist kraadi ja seitseteist kraadi on Paralepas kuusteist kraadi Kakumäel, nii et need on siis hommikused mõõtmised. Küllap teiste temperatuuride kohta jõuab infoga siis päeva jooksul. Aga kui meil Huvitaja saade läbi saab, on eetris virgutusvõimlemine, on ka järjejutt ja pärast keskpäevaseid uudiseid on Uudis pluss. Ja saatejuht Mirko Ojakivi Tartu stuudiost. Tervist.
Mirko Ojakivi	Tere enne lõunat.

Meelis Süld	Missugused on sinu teemakäsitlused täna?
Mirko Ojakivi	No avaloos me arutleme Trinity advokaadibüroo vandeadvokaadi abi Maarja **Pildiga** selle üle, kuidas oleks võimalik teatritel ja kinodel **inimeste** isikuandmeid rahva tervise huvides koguda, et me oleme siin eelnevatel päevadel kuulnud, miks seda teha ei saa, samas kui vaadata Saksamaal Belgias tuntud sellistes andmekaitse andmekaitset kummardavates riikides selliseid andmeid kogutakse, aga Eestis juba kõik teavad, et meil ei saa, aga kas kuidagi ikka saab siis päevakommentaaris, arutleb Külli Taro teemal, kas inimese ja riigisuhet on taandatav kliendisuhtele, nagu siin vahel ju ka riik saadab teavitusi inimestele, et hea klient, et kas see on nagu taandatav lihtsalt kliendisuhtele siis peaminister Jüri Ratas viie minuti pärast alustab otsesaates Otse uudistemajast Anvar **Samostiga** välispoliitika teemadel rääkimist, me teeme sellest ka Uudis plussis. Väikese kokkuvõtte. Olulisematest küsimustest olulisematest vastustest ja täna saates veel Tartu on saanud lodjakoja võrra rikkamaks. Kellele ja milleks lodjakoda on mõeldud, sellest räägime lodjakoja peremehe Priit **Jagomägiga**. Ja valitsust nõustanud majandusinimesed on valmis saanud majanduse elavdamise ettepanekutega. Me räägime neist ekspertkogusse kuuluva majandusprofessori Raul Eametsaga.
Meelis Süld	Need teemad siis keskpäevaste uudiste järele Uudisplussi saates aitäh Mirko praegu selle lühikese ülevaate eest ja, ja head päeva ja, ja ettevalmistust veel saate eel. Mina, Meelis Süld, tänan kõiki kuulamast ja kaasa mõtlemast. Aitäh Marika **Leetmäele** helipuldis kaunist teisipäeva jätku.

Lisa 5 – WER arvutamine

```
# -*- coding: utf-8 -*-
"""
Created on Thu Mar 18 13:01:10 2021
@author: Ahti Lohk
"""

import re

def read_file(file_name)-> list:
    with open(file_name, mode = "r", encoding = "utf-8") as
input_file:
        file_content = input_file.read()

    return re.findall(r'[a-zA-ZäöüõÄÖÜÕšžŠŽ]+', file_content.lower())

def make_word_freq_dict(words):
    freq_dict = dict()
    for word in words:
        freq_dict[word] = freq_dict.get(word, 0) + 1
    return freq_dict

def overlap_coefficent(word_set1, word_set2):
    intersection = word_set1 & word_set2
    union = word_set1 | word_set2
    return len(intersection) / len(union)

original_words = read_file("fail-originaal.txt")
repair_words = read_file("fail-parandatud.txt")

orig_freq_dict = make_word_freq_dict(original_words)
repair_freq_dict = make_word_freq_dict(repair_words)

print(overlap_coefficent(set(original_words), set(repair_words)))

# Unique words in repair file not in original:
print("\REPAIRED")
reprair_unique = set(repair_words) - set(original_words)
print(reprair_unique)

# Unique words in original file not in repair:
print("\nORIGINAL")
original_unique = set(original_words) - set(repair_words)
print(original_unique)
```