TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Janar Saks

Eestikeelsete tekstide sisukokkuvõtja EstSum edasiarendamine

Bakalaureusetöö (9 EAP)

Juhendaja: Kaili Müürisep, PhD

Tartu 2018

Eestikeelsete tekstide sisukokkuvõtja EstSum edasiarendamine

Lühikokkuvõte:
Tänapäevaste informatsioonihulkade juures on sageli vaja saada kiiresti ülevaade olulisest informatsioonist. Seepärast võiks rakenduse poolt automaatselt genereeritud sisukokkuvõte kui lühendatud kiiresti kättesaadav ülevaade algallikast olla oluline informatsiooni kogumise vahend. Kuid nagu iga keeletehnoloogiline rakendus, sõltub see sihtkeele omapäradest, mille jaoks see on disainitud. Inglise keele tarbeks loodud sisukokkuvõtja ei sobi eesti keele jaoks, sest eesti keelele omane sõnavormide rohkus vajab hoopis teistsugust lähenemist.
Kuigi eestikeelsete tekstide kokkuvõtja EstSum kasutab lause kaalu arvutamisel võtmesõnade põhist skoori, siis skoori arvutatakse sõnavormide, mitte sõnade algvormide pealt. EstSumi ühendamine lingvistilise mooduliga, mis suudab analüüsida sõnade algvorme, tõstis kokkuvõtja tulemuslikkust võrreldes EstSumiga, millel vastav moodul puudub.
Töö tulemusena valminud automaatse kokkuvõtja uus versioon suudab eraldada rohkem olulist informatsiooni algallikast, kui seda tegi EstSumi vana versioon.

Võtmesõnad:
Lemma, sisukokkuvõtja, EstSum, EstNLTK, lemmatiseerija, kokkuvõtja hindamine

CERCS: P175 Informaatika, süsteemiteooria

The Development of Estonian Texts' Summarizer EstSum

Abstract:
In today's vast information quantity, there is often a need for a quick overview of important information. Therefore, a summary as a shortened overview of the source material, could be an important source of information collection. But like any other language technology application, they also depend on the peculiarities of the language they are designed for. A summarizer created for the English language is not implementable for the Estonian language, because of the multitude of word forms that is typical for the Estonian language and therefore requires a completely different approach.
Although Estonian texts' summarizer EstSum uses a keyword-based score to calculate a sentence´s weight, the score is calculated solely from word forms not the word´s stem. The addition of a linguistic module, that can analyze word stems, did increase the evaluation score compared to the regular EstSum.
Furthermore, the new developed version of the Estonian texts' summarizer is capable of separating more important information from the source than the old version of EstSum.

Keywords:
Word stem, summarizer, EstSum, EstNLTK, lemmatizer, summarization evaluation

CERCS: P175 Informatics, systems theory

Sisukord

1 Sissejuhatus

Tänapäevaste informatsioonihulkade juures on sageli vaja saada kiiresti ülevaade olulisest informatsioonist. Seepärast võiks rakenduse poolt automaatselt genereeritud sisukokkuvõte kui lühendatud kiiresti kättesaadav ülevaade algallikast olla oluline informatsiooni kogumise vahend. Kokkuvõte artiklitest, teadustöödest või dokumentidest võimaldab meil otsustada, kas soovime algtekstiga lähemalt tutvuda või saame selle kõrvale lükata. Samuti on väikestelt ekraanidelt ebamugav lugeda pikki tekste. Olukord muutuks oluliselt mugavamaks, kui oleks võimalik lühikese teksti põhjal otsustada, kas kogu teksti läbi töötamine on vajalik.
Automaatseid sisukokkuvõtjaid on loodud aastakümneid, neid on nii vabavaralisi kui ka kommertssüsteemide koosseisu kuuluvaid. Kuid nagu iga keeletehnoloogiline rakendus, sõltub see sihtkeele, millele see on disainitud, omapäradest. Inglise keele tarbeks loodud sisukokkuvõtja ei sobi eesti keele jaoks, sest eesti keelele omane sõnavormide rohkus vajab hoopis teistsugust lähenemist.
Eestikeelset sisukokkuvõtjat EstSum on arendatud mitu aastat. Pilleriin Mutso [P.05a] kirjeldas, kuidas EstSum määrab lausetele kaalu, mille põhjal koostatakse kokkuvõte. Mutso leidis, et kaalu määramisel on tähtis lause positsioon, formaat ja lauses paiknevad võtmesõnad. Mutso kasutas võtmesõnade analüüsis sõnade erivorme. Kaili Müürisep ja Pilleriin Mutso [P.05b] kirjeldasid, kuidas määrati parameetrid, mis näitavad lause kaalus positsiooni, formaadi ja võtmesõnade tähtsuse osakaalu. Parameetrid määrati väikse korpuse põhjal, kuhu kuulus kakskümmend teksti, milles keskmiselt kaheksateist lauset. Käesoleva töö eesmärk on lisada eestikeelsele sisukokkuvõtjale EstSum lingvistiline moodul, mis võimaldaks analüüsida sõnade algvorme. Samuti on töö eesmärk uuele Estsumile leida sobivad parameetrite väärtused autori koostatud korpuse põhjal ja hinnata selle tulemuslikkust.
Käesolev töö koosneb neljast osast. Töö esimeses osas defineeritakse sisukokkuvõtja mõiste ja tuuakse välja, kuidas saab erinevaid kokkuvõtjaid liigitada. Samuti antakse ülevaade EstSumist ja võrreldakse EstSumi rootsikeelse sisukokkuvõtja SweSumiga, mis oli eeskujuks EstSumi loomisel. Töö teises osas kirjeldatakse, kuidas hinnatakse sisukokkuvõtja tulemuslikkust ja antakse ülevaade eelnevatest töödest, kus on hinnatud EstSumi funktsionaalsust. Töö kolmandas osas räägitakse, kuidas ühendati EstSumiga lingvistiline moodul ja kuidas koostati uus sõnade algvormide sõnasagedustabel ja stoppsõnade loend. Töö neljandas osas kirjeldatakse, kuidas koostati arendus- ja testkorpus. Samas töö osas antakse ülevaade ka tulemustest: kuidas leiti EstSumile uued parameetrite väärtused ja hinnatakse tulemuslikkust. Samuti tuuakse välja autori nägemus EstSumi arendussuunast.

2 Eestikeelsete tekstide sisukokkuvõtja EstSum

Selles peatükis antakse ülevaade, kuidas liigitatakse sisukokkuvõtjaid. Lisaks kirjeldatakse EstSumi arhitektuuri ning tuuakse täpsemalt välja, kuidas EstSum valib kokkuvõtetesse lauseid. Seejärel võrreldakse EstSumi lausete valimise protsessi teiste sisukokkuvõtjate lausete valimiste protsessidega.

2.1 Sisukokkuvõtja

Inderjeet Mani [I.01] annab sisukokkuvõtte laia definitsiooni: „Sisukokkuvõtja on süsteem, mille eesmärk on sisendist toota tihendatud kujutus, mis on mõeldud inimestele tarbimiseks.“ Definitsioonist saab järeldada, et sisukokkuvõtja eesmärk on algallikast eraldada oluline teave, nii et lugejal oleks sellest kiiresti hoomatav ülevaade.
Sisukokkuvõtjaid saab eristada selle järgi, kuidas need genereerivad väljundi [I.01]:

1. väljavõtte (ingl k *extract*) meetodit kasutavad sisukokkuvõtjad, mille väljund koosneb ainult sisendist kopeeritud osadest nt sisendteksti laused;
2. ülevaate (ingl k *abstract*) meetodit kasutavad sisukokkuvõtjad, mille väljundis mingi osa puudub sisendist nt sisendtekstist nime asendamine asesõnaga.

Samas, traditsiooniliselt on sisukokkuvõtjad jaotatud kaheks [I.01]:

1. osundav (ingl k *indicative*), kus kokkuvõte annab ülevaate tervest sisendfailist;
2. informatiivne (ingl k *informative*), kus kokkuvõte annab ülevaate kogu olulisest informatsioonist sisendfailis.

Käesolevas bakalaureusetöös käsitletav sisukokkuvõtja EstSum on osundav ja kasutab kokkuvõtete tegemiseks väljavõtte meetodit.

2.2 Eesti keele sisukokkuvõtja

EstSum on veebiuudistele ja elektroonlistele ajaleheartiklitele orienteeritud eestikeelne sisukokkuvõtja [K.06]. Kuigi EstSumi on arendatud mitu aastat, siis arendus on enamasti seotud ainult diplomi- ja bakalaureusetöödega [K.06].
Kuna EstSum kasutab kokkuvõtete genereerimiseks väljavõtte meetodit, siis genereeritud kokkuvõte ei ole sidus [K.06]. EstSum on kirjutatud programmeerimiskeeles Perl ja koosneb kolmest moodulist [K.06].
Käesolevas töös valminud lähtekood uuest EstSumist on kättesaadav GitHubist[1]. Samuti on võimalik uut EstSumi katsetada veebis[2].

[1] Koodi repositoorium: https://github.com/janarsaks/EstSum_development
[2] EstSumi katsetamine: http://kodu.ut.ee/ janar/EstSum2.0/

2.3 EstSumi ülesehitus

Kaili Müürisep [K.06] järgi jaguneb EstSumi arhitektuur kolmeks osaks:

1. HTML-konverter, mille eesmärk on sisend viia SGML-formaati, kus sisendist on eemaldatud ebavajalikud elemendid ja lisatud vajalikud märgendid;
2. lausestaja, mis jagab regulaaravaldistega sisendi lauseteks;
3. lausete väljavalija, mis annab lausetele kaalu ja koostab kaalude järgi kokkuvõtte.

Käesolev töö keskendub EstSumi kolmandale osale, sest töö eesmärgiks on lisada lingvistiline moodul, mis võimaldaks lause kaalu hindamise protsessis arvestada sõnade algvormide esinemissagedust.

2.4 Lausete kaalu arvutamine

EstSum genereerib algtekstist kokkuvõtte, arvutades lausete kaalud ning nende tulemuste järgi valib laused, mille kaal on kõrgem. EstSumi lause kaalu arvutamise valem põhineb Edmundsoni paradigmal.
Inderjeet Mani [I.01] kirjeldas, kuidas Edmundson töötas teadusartiklite korpuse peal välja tingimused, mille abil saab määrata lause kaalu. See informatsioon aga võimaldas väljavalimismeetodi abil koostada algtekstist kokkuvõtte. Edmundson leidis, et lause kaalu mõjutavad märksõnad (ingl k *cue words*), võtmesõnad, lause positsioon ja pealkiri. Nende tingimuste alusel töötas Edmundson välja lause kaalu arvutamise valemi.

$$W(s) = \alpha C(s) + \beta K(s) + \gamma L(s) + \delta T(s) \tag{1}$$

Lause s kaalu leidmiseks tuleb liita kokku lause s märksõnade skoor C, võtmesõnade skoor K, positsiooni skoor L ja pealkirjaskoor T. α, β, γ ja δ on parameetrid, mille abil saab muuta tingimuse osakaalu lause kaalu arvutamises. Edmundson leidis, et lause positsioon oli parim tingimus ja võtmesõnad kõige nõrgem tingimus [I.01].
EstSum kasutab lause kaalu määramiseks kolme tingimust. Lause s kaal leitakse, kui liidetakse kokku lause s positsiooniskoor (P), formaadiskoor (F) ja sõnasageduste skoor (K).

$$W(s) = \alpha P(s) + \beta F(s) + \gamma K(s) \tag{2}$$

α, β, γ on parameetrid, millega saab reguleerida, kui suur on kindla skoori osakaal. Lause positsioonipõhise skoori jaoks on tähtsad lõikude esimesed laused. Samas, kõige tähtsam on esimene lause, mis järgneb teksti pealkirjale.
Lause formaadipõhist skoori suurendab see, kui lause on kirjutatud paksus kirjas või kaldkirjas. Skoor langeb, kui lauses on jutumärgid või lause lõppeb hüüu- või küsimärgiga. Samuti väheneb skoor, kui on tegemist piltide all olevate tekstidega.

Lause sõnasageduste põhist skoori suurendavad sõnad, mis olid teksti pealkirjas. Skoor langeb, kui lauses on sõnad, mis paiknevad sõnavormide sagedustabelis või stoppsõnade loendis. Sagedustabelis oli 1057 sõnavormi ja stoppsõnade loendis 105 sõna. Pilleriin Mutso [P.05a] märkis, et sagedustabeli koostamiseks töödeldi läbi ajalehe „Postimees“ artiklitest koosnev korpus. Korpuses oli ligikaudu 390 000 sõna ning sagedustabelisse sisestati sõnavormide esinemissagedused 10 000 sõna kohta.

2.5 Lausete valimine võrdluses SweSumiga

EstSum on koostatud võttes eeskujuks sisukokkuvõtjat SweSum [K.06]. SweSum on rootsikeelne sisukokkuvõtja, mis on orienteeritud artiklite kokkuvõtmisele.
SweSum ja EstSum valivad kokkuvõtete jaoks lauseid erinevalt, kuigi mõlema kokkuvõtja jaoks on lause positsioon tähtis. SweSum annab lausetele skoori olenevalt sellest, millisel real nad paiknevad. Lause saab skoori vastavalt oma reanumbri pöördväärtusele [H.00]. Seevastu annab EstSum kõrgemaid skoore lõikude esimestele lausetele. Mõlemad kokkuvõtjad peavad tähtsaks lauset, mis järgneb algteksti pealkirjale [M.03].
Mõlema kokkuvõtja jaoks on oluline lause formaat. Kui lause on paksus kirjas, suurendab see lause kaalu [H.00]. Erinevalt SweSumist suurendab EstSum lause kaalu, kui see on kaldkirjas ja vähendab lause kaalu, kui selles on jutumärgid või see lõppeb hüüu- või küsimärgiga.
Lause kaalu hindamisel uurivad mõlemad kokkuvõtjad ka sõnu. Erinevalt EstSumist suurendab SweSum lause kaalu, kui lauses leidub numbriline väärtus [H.00]. Ent mõlemad kokkuvõtjad peavad tähtsateks pealkirjas paiknevaid sõnu [M.03], mis tähendab, et lause kaal suureneb, kui selles paikneb mõni pealkirja sõna. Suureks erinevuseks on asjaolu, et SweSum analüüsib lause kaalu arvutamisel ka sõnade algvorme [H.00]. SweSum kasutab sõnade algvormide leidmiseks sõnastikku, kus võtmeteks on sõnade erivormid ning väärtusteks sõna algvorm [M.03]. Lisaks, erinevalt EstSumist peab SweSum tähtsateks lauseid, mis sisaldavad samu numbrilisi väärtusi ja arvestab kaalu määramisel lause pikkusega [M.03].

3 Sisukokkuvõtja tulemuslikkus

Selles peatükis kirjeldatakse, kuidas hinnata sisukokkuvõtja tulemuslikkust. Lisaks tuuakse ülevaade hinnangutest EstSumi tulemuslikkusele, milleni on jõudnud eelnevad tööd.

3.1 Kokkuvõtja hindamine

Kokkuvõtete hindamine ei ole lihtne, sest pole võimalik objektiivselt väita, millised osad algtekstist on olulised ja millised mitte. Samuti pole olemas juhendit, mida järgides saaks anda hinnang kokkuvõtte ja seeläbi ka kokkuvõtja kohta. Üldiselt mõõdetakse kokkuvõtte hindamisel kahte tingimust [H.07]:

1. tihendatus (ingl k *compression rate*) (CR), mis näitab, kui palju lühem on kokkuvõte algtekstist;
2. andmepeetuse (ingl k *retention*) suhe (RR), mis näitab, kui palju informatsiooni on säilitatud.

Tihendatus saadakse, kui kokkuvõtte pikkus jagatakse algteksti pikkusega. Käesolevas töös sellele suhtele ei keskenduta, sest EstSumi arendamiseks ja testimiseks automaatselt genereeritavad ja käsitsi tehtud kokkuvõtted moodustavad umbes 30% algteksti pikkusest.

$$CR = \frac{kokkuvottepikkus}{algtekstipikkus} \tag{3}$$

Andmepeetuse suhe saadakse, kui kokkuvõttes paiknev informatsioon jagatakse algtekstis paikneva informatsiooniga. Seejuures tekib probleem, kuidas hinnata informatsiooni hulka kokkuvõttes või algtekstis.

$$RR = \frac{informatsioonkokkuvottes}{informatsioonalgtekstis} \tag{4}$$

Kõige üldisemad meetodid, millega üritatakse anda hinnang andmepeetuse suhtele, saab jagada sisemisteks (ingl k *intrinsic*) ja välimisteks (ingl k *extrinsic*) [I.01]. Sisemised meetodid uurivad kokkuvõtte sidusust ja informatiivsust, mis tähendab, et hinnang antakse kokkuvõtte sisulisele kvaliteedile [H.07]. Selleks võrreldakse kokkuvõtet ideaaliga, mis üldjuhul koostatakse inimese poolt [I.01][H.07]. Välised meetodid hindavad kokkuvõtte asjakohasust ja loetavust, mis tähendab, et hinnang antakse kokkuvõtte vastuvõetavuse ja praktilisuse kohta [H.07].
EstSumi tulemuslikkust on ka varem hinnatud ning iga eelnev hinnang on antud kasutades sisemist hindamismeetodit, kus võrreldakse kokkuvõtja genereeritud kokkuvõtet käsitsi koostatud kokkuvõttega [P.05a][K.06][Sel08]. Seepärast kasutati käesolevas töös lingvistilise mooduliga EstSumi tulemuslikkuse hindamisel samasugust lähenemist.

3.2 EstSumi tulemuslikkuse hinnangud

EstSumi tulemuslikkusele on hinnangu andnud kolm autorit. Pilleriin Mutso [P.05a] leidis, et EstSum suutis ajaleheartiklist eraldada keskmiselt umbes 60% informatsiooni, mis oli Mutso jaoks oluline. Hinnangu andmiseks valis ta 20 ajalehe „Postimees“ artiklit ning koostas käsitsi nende põhjal kokkuvõtted. Artiklite keskmine pikkus oli 25 lauset. Järgmiseks häälestas ta EstSumi lause kaalu valemi parameetreid ning siis võrdles EstSumi genereeritud kokkuvõtteid käsitsi koostatud kokkuvõtetega.
Sama kõrge tulemuslikkuse hinnangu said ka Kaili Müürisep ja Pilleriin Mutso koostöös [P.05b]. Nad leidsid, et EstSum valis kokkuvõtteid genereerides keskmiselt 60% ulatuses samu lauseid, mis olid ka käsitsi koostatud kokkuvõtetes. Nende testkorpus koosnes 11-st artiklist, mille keskmine pikkus oli kakskümmend kolm lauset.
Kolmandana on EstSumi tulemuslikkust hinnanud Keili Sellik [Sel08]. Ta hindas EstSumi tulemuslikkust kahel viisil. Esimese viisi jaoks koostas ta viiekümnest artiklist, mille keskmine pikkus oli 10,16 lauset, käsitsi kokkuvõtte ja võrdles neid EstSumi genereeritud kokkuvõtetega. Sellik sai tulemuseks, et keskmiselt 65,29% ulatuses kattusid genereeritud ja käsitsi koostatud kokkuvõtted. Teise meetodina kasutas ta ROUGE programmi, et hinnata EstSumi tulemuslikkust. Kasutades programmi moodulit ROUGE-L, mis võrdleb lauseid, saadi tulemuseks, et EstuSumi kattuvus on 68,96%.
Eelnevad hinnangud EstSumi tulemuslikkusele saadi lühikeste artiklite põhjal. Käesolevas töös valiti korpustesse pikemaid artikleid. Peatükis viis kirjeldatakse, kuidas loodi arendus- ja testkorpus, mille uue versiooniga EstSumi tulemuslikkust hinnatakse. Arenduskorpus koosneb kahekümnest artiklist, mille keskmine pikkus oli 40,4 lauset. Testkorpus koosneb kümnest artiklist, mille keskmine pikkus oli 40,3 lauset.

4 Lingvistilise mooduli ühendamine EstSumiga

Selles peatükis kirjeldatakse, kuidas EstSumile lisati lingvistiline moodul, mis kokkuvõtete genereerimisel arvestab sõnade algvormidega. Samuti kirjeldatakse, kuidas koostati võtmesõnade analüüsiks vajalik sõnade sagedustabel, mis on lingvistilise mooduli edukaks kasutamiseks hädavajalik.

4.1 Lemmatiseerija rakendamine

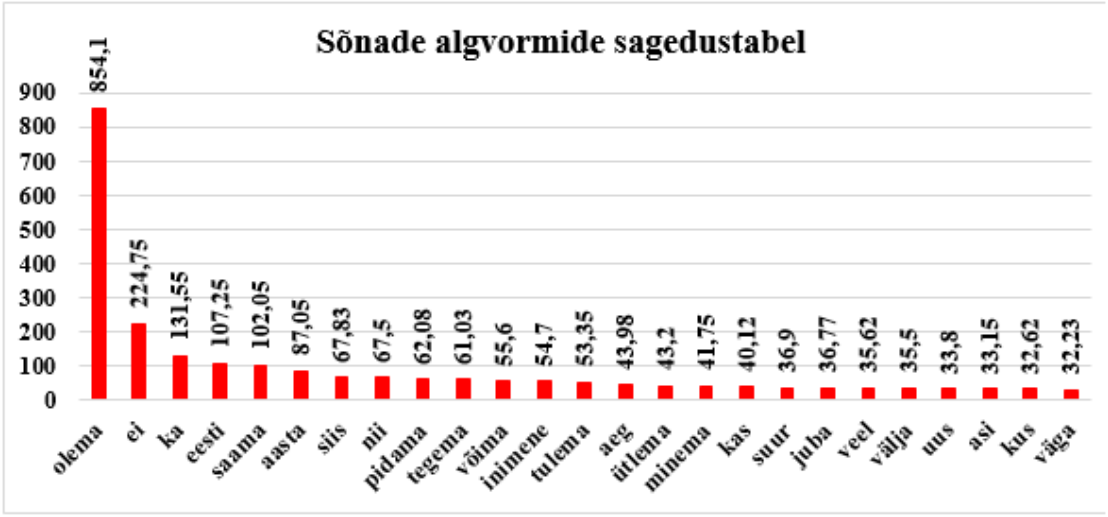
EstSum oli algselt kirjutatud programmeerimiskeeles Perl. Et mugavamalt siduda programmi eesti keele keeletehnoloogiliste moodulitega, otsustati EstSumi lähtekood kirjutada ümber programmeerimiskeelde Python. Pythonis kirjutatud programmides on mugav kasutada teeki EstNLTK [HJ16], mis võimaldab EstSumil leida võtmesõnade algvorme ehk lemmasid.
EstSumi esialgne versioon kasutas võtmesõnade analüüsimiseks sõnavorme, mitte lemmasid. See tähendab, et sõna kaal sõltus sõna kirjapildist, mitte tähendusest. Näiteks kui artikkel rääkis metsandusest, siis sõnavormid mets, metsa, metsas, metsast jne loeti kõik erinevateks võtmesõnadeks. Selle puudujäägi eemaldamiseks lisati EstSumile juurde moodul, mis muudab lauses olevad sõnad nende algvormiks (eespool toodud näites lemmaks 'mets'). Selleks kasutati EstNLTK Text klassi lemmatiseerijat. Lemmatiseerija lisati meetoditesse „analyze_title“, „analyze_line“ ja „word_based_score“, kus analüüsitakse lausetes või pealkirjades paiknevaid sõnu. Töö eesmärgi täitmiseks kasutati Pythoni versiooni 3.5 ning EstNLTK versiooni 1.4.1.1.
EstNLTK lemmatiseerija kasutab sõnade algvormide leidmiseks Vabamorfi sõnastikupõhist morfoloogilist analüsaatorit [HJ16]. Vabamorfi morfoloogiline analüüs võrdleb sõnu sõnastikus paiknevate lekseemide kombinatsioonidega ja eemaldab sõnadelt liiteid ja lõppe, et kontrollida seda leksikoniga, kus paiknevad sõnatüved [T.16a]. Kui morfoloogiline analüüs annab tulemuseks mitu varianti, siis kasutab Vabamorf ühestajat, et valida õige ning kui ühestaja ei suuda õiget valida, väljastab analüüs mitu varianti [HJ16].
EstSum kasutab võtmesõnade analüüsimiseks sõnade sagedustabelit. Esialgses versioonis oli selleks sõnavormide sagedustabel. Uues versioonis loodi uus sõnade algvorme sisaldav sagedustabel. Uue versiooniga EstSumi lähtekoodiga saab tutvuda GitHubis[3].

4.2 Sõnade algvormide sagedustabel ja stoppsõnad

EstSumi esialgne versioon kasutab võtmesõnade analüüsimiseks sõnade sagedustabelit ja stoppsõnade loendit. Kuid kuna nii sagedustabel kui ka loend koosnevad erinevates sõnavormidest, vajab uus, lingvistilise mooduliga ühendatud versioon võtmesõnade analüüsiks algvormidest koosnevat sagedustabelit ja stoppsõnade loendit. Seetõttu tuli

[3]Koodi repositoorium: https://github.com/janarsaks/EstSum_development

koostada uus sõnade sagedustabel ning stoppsõnade loend, kus on ainult sõnade algvormid.
Sõnade algvormide sagedustabeli loomiseks kasutati Tartu ülikooli „Keeletehnoloogia“ kursuse eestikeelset korpust, kus on poole aasta jagu ajaleheartikleid. Korpus koosnes 894 737 sõnast. Korpus oli TEI-formaadis ning loeti sisse kasutades teicorpus klassi EstNLTK teegis. Esmalt korpuse tekstid lausestati ja seejärel lemmatiseeriti. Sõna algvormi sagedusetabelis on toodud lemmade esinemissagedus 100 000 sõna kohta. Tabelisse lisati 1000 kõige kõrgema sagedusega lemmat. Samas välistati sõnade algvormid, mis paiknevad stoppsõnade loendis. Joonisel 1 on ära toodud 25 kõrgema sagedusega lemmat. Sagedustabeli sõnade loend ja sagedused on kogu suuruses toodud välja lisas I.



Joonis 1. 25 sagedamini esinevat sõna algvormi ajaleheartiklites.

Antud töös mõeldakse stoppsõnade all sõnu, millele antakse võtmesõnade analüüsi hetkel väärtuseks 0. See tähendab, et need sõnad ei mõjuta lause kaalu võtmesõnade skoori ehk need sõnad on tähenduse mõttes sisutühjad. Seepärast paiknevad uues stoppsõnade loendis ainult asesõnad ja sidesõnad. Nende loend saadi Eesti puudepangast ase- ja sidesõnade filtreerimise teel [T.16b]. Stoppsõnade loendis on 72 sõna ning need on välja toodud lisas II.

5 Tulemuslikkuse hindamine

Selles peatükis kirjeldatakse, kuidas koostati arendus- ja testkorpus, mis võimaldaks hinnata lingvistilise mooduliga ühendatud EstSumi tulemuslikkust. Samuti tuuakse välja, kuidas valiti uue mooduliga sobivad konstandid lause kaalu valemisse. Seejärel antakse hinnang uue mooduliga EstSumi tulemuslikkusele ja edastatakse autori nägemus, millises suunas peaks EstSumi arendamine tulevikus liikuma.

5.1 Arendus- ja testkorpus

Arenduskorpuse loomiseks valiti pisteliselt kakskümmend ajalehtede „Postimees“ ja „Õpetajateleht“ veebiartiklit. Arenduskorpusesse valiti ainult artiklid, mille pikkus oli keskmisest kõrgem ehk umbes viissada kuni tuhat sõna. Esimesed kümme valitud artiklit olid arvamused ning ülejäänud kümme olid uudised. Pooled arvamused ja pooled uudised valiti ühest ajalehest ja ülejäänud teisest.
Testkorpuse loomisel lähtuti samadest tingimustest nagu arenduskorpuse loomisel. Kuid testkorpusesse valiti kõigest kümme artiklit, millest viis olid arvamused ja viis uudised. Arvamustest valiti kolm ajalehest „Õpetajaleht“ ja kaks ajalehest „Postimees“. Uudisetest valiti kaks ajalehest „Õpetajaleht“ ja kolm ajalehest „Postimees“.
Valitud artiklitest koostas autor väljavalimismeetodit kasutades kokkuvõtted, mis moodustasid umbes 30% artikli kogupikkusest. Samas tuleb märkida, et kokkuvõtete tegemine on subjektiivne, sest ühe inimese tõlgendus artikli olulisest osast võib erineda teise inimese tõlgendusest. Seda näitas ka Martin Hassel [H.05], kes leidis, et kümne isiku koostatud artikli kokkuvõtetes oli igas kokkuvõttes ainult 33,9% sama informatsiooni. Tabelis 1 on ära toodud arendus- ja testkorpuse artiklite ja kokkuvõtete lausete ja sõnade arvud.

Tabel 1. Korpuses paiknevate artiklite ja kokkuvõtete pikkused

Arenduskorpus				
Artikkel	**Laused**	**Sõnad**	**Kokkuvõtte laused**	**Kokkuvõtte sõnad**
Arvamus1	52	666	15	198
Arvamus2	37	758	10	227
Arvamus3	51	897	16	270
Arvamus4	35	576	11	173
Arvamus5	43	816	12	239
Arvamus6	73	999	21	295
Arvamus7	46	711	15	214
Arvamus8	52	879	14	255
Arvamus9	34	578	11	176
Arvamus10	50	820	13	246
Uudis1	35	532	10	161

Uudis2	44	562	12	170
Uudis3	44	646	11	191
Uudis4	34	682	8	210
Uudis5	46	695	12	213
Uudis6	31	593	8	185
Uudis7	20	590	8	186
Uudis8	26	467	8	135
Uudis9	31	481	8	141
Uudis10	24	506	8	153
Keskmine	**40,4**	**672,7**	**11,55**	**201,9**
Testkorpus				
Arvamus1	61	958	20	289
Arvamus2	43	690	12	196
Arvamus3	44	535	12	164
Arvamus4	45	714	12	207
Arvamus5	57	960	19	278
Uudis1	36	530	9	156
Uudis2	31	535	9	153
Uudis3	22	498	8	141
Uudis4	32	514	10	148
Uudis5	32	741	9	217
Keskmine	**40,3**	**667,5**	**12,0**	**194,9**

Kuid korpuste koostamine ei piirdu ainult artiklitest kokkuvõtete tegemisega, sest kokkuvõtte genereerimiseks EstSumiga on vaja veel artiklid viia kindla märgendusega sisendfaili kujule. Lisas III on välja toodud näide artiklist, selle põhjal käsitsi koostatud kokkuvõttest ja EstSumi genereeritud kokkuvõttest. Korpuses paiknevad kokkuvõtted on kättesaadavad GitHubis[4].

5.2 Artiklite märgendamine

Selleks, et viia artikkel EstSumi jaoks vajalikule sisendfaili kujule, on tarvis artikleid eeltöödelda. Eeltöötlusprotsess jagunes kaheks faasiks.
Eeltöötluse esimeses faasis tuli tekstis käsitsi ära märgendada kindlad teksti osad, et järgmises faasis oleksid need osad äratuntavad. Artikli teksti tuli lisada järgmised märgendid:

1. pealkiri pandi „<h1>“ ja „</h1>“ märgendite vahele;
2. alampealkirjad pandi „<h2>“ ja „</h2>“ märgendite vahele;

[4]Koodi repositoorium: https://github.com/janarsaks/EstSum_development

3. paksus kirjas osa pandi „“ ja „“ märgendite vahele;

4. kaldkirjas osa pandi „“ ja „“ märgendite vahele.

Eeltöötluse teises faasis jagati artiklid lõikudeks ja lõigud lauseteks. Selleks kasutati EstNLTK Text klassi. See suudab teksti lõikudeks jagada, kuid eeldab, et lõigud on eraldatud kahekordse reavahetussümboliga[5]. Lõikude jagamisel lauseteks kasutati Text klassi lausestajat. Selline jagamine võimaldas lisada lõigud „ “ ja „ “ märgendite ning laused „<s>“ ja „</s>“ märgendite vahele. Samuti asendati eeltöötluse esimeses faasis lisatud märgendid sisendfaili jaoks sobilike märgenditega. Käsitsi lisatud märgendid võimaldasid algtekstis peal- ja alampealkirjad lisada „<head>“ ja „</head>“ märgendite vahele, kaldkiri lisada „<hi rend="italic»“ ja „</hi>“ märgendite vahele ja paks kiri lisada „<hi rend="bold»“ ja „</hi>“ märgendite vahele. Joonisel 2 on välja toodud kaks lõiku korrektselt märgendatud sisendfailist.

```
<div0><head><hi rend="bold">Riina Kabi: miks ei saa Eesti Punast Risti muuta projektipõhiseks?</hi></head>
<p>
<s><hi rend="bold">Eesti on alati pidanud tähtsaks elanike turvalisust.</hi></s>
<s><hi rend="bold">Enamik maailma riikidest on jõudnud äratundmisele, et turvalisuse tagamiseks on vaja veel midagi peale relvade ja kaitsejõudude.</hi></s>
</p>
<p>
<s>Tuleb hoolitseda haavatute eest, ühendada sõjas üksteist silmist kaotanud pereliikmeid, tegelda laste, orbude, vanurite ja invaliididega.</s>
<s>1864. aastal sõlmitud esimese Genfi konventsiooniga pandi nii alus humanitaarõigusele ja Punase Risti liikumisele.</s>
</p>
```

Joonis 2. Näide märgendatud tekstist.

Tuleb märkida, et Text klassi lausestaja teeb lausete eraldamisel vigu. Seetõttu tuli eeltöötluse tulemus käsitsi üle kontrollida. Lisas III on välja toodud näide märgendatud tekstist. Korpuses paiknevad märgendatud tekstid on kättesaadavad GitHubis[6].

5.3 Parameetrite uuendamine

EstSumile, millega on ühendatud lingvistiline moodul, tuli leida uued parameetrid lause kaalu arvutamise valemis, sest eeldati, et uues versioonis töötab võtmesõnade arvestamise moodul efektiivsemalt. Valemis on kolm parameetrit ning sobivate kolmikute leidmiseks prooviti läbi kõik variandid, kus kolmiku summa andis tulemuseks ühe, sest nii oli võimalik analüüsida erineva parameetri osakaalu tähtsust.
Parameetrite väärtuste vahemikuks võeti 0,1 kuni 0,8. Kolmiku sobivuse hindamiseks genereeriti vastava kolmikuga kakskümmend kokkuvõtet artiklite põhjal, mis paiknesid arenduskorpuses. Iga genereeritud kokkuvõtet võrreldi käsitsi koostatud kokkuvõttega. Võrdluse aluseks oli kattuvus ning tulemuseks oli kattuvusprotsent. Kattuvuse hindamiseks vaadati läbi nii genereeritud kui ka käsitsi koostatud kokkuvõtete laused ning leiti

[5]EstNLTK Text klassi lähtekood. Kättesaadav: https://estnltk.github.io/estnltk/1.4/_modules/estnltk/text.html
[6]Koodi repositoorium: https://github.com/janarsaks/EstSum_development

laused, mis esinesid mõlemas kokkuvõttes. Kattuvusprotsendiks oli kattuvate lausete sõnade osakaal genereeritud kokkuvõtte kogupikkusest. Seega saadi iga kolmiku sobivuse hindamiseks kakskümmend kattuvusprotsenti ning nende abil arvutati keskmine kattuvus. Mida kõrgem oli keskmine kattuvus, seda sobilikum oli parameetrite kolmik. Tabelis 2 on toodud viis parameetrite kolmikut, mis saavutasid kõrgeimad kattuvusprotsendid. Tabel kõikide parameetrite kolmikutega on ära toodud lisas IV.

Tabel 2. Viis kõrgeimate keskmiste kattuvustega parameetrite kolmikut.

α	0,4	0,5	0,4	0,4	0,7
β	0,4	0,4	0,3	0,5	0,2
γ	0,2	0,1	0,3	0,1	0,1
Keskmine kattuvus	62,24%	61,68%	61,65%	61,46%	61,39%

Tabelis 2 on välja toodud, et kõige kõrgema keskmise kattuvuse sai parameetrite kolmik 0,4, 0,4 ja 0,2. Kui analüüsiti tabelit 2, siis selgus, et lause kaalu arvutamise valemis on kõige olulisem parameeter α, mis näitab lause positsiooni skoori osakaalu, ning kõige vähem olulisem on parameeter γ, mis näitab lause sõnasageduste põhise skoori osakaalu. Et saada parem ülevaade erinevate parameetrite olulisusest, vaadati iga parameetri puhul kõiki parameetrite kolmikuid, kus vaadeldava parameetri väärtus on suurem kui 0,5. See tähendab, et vaadeldi kolmikuid, kus üks parameeter moodustas üle poole lause valemi skoorist. Vastavaid kolmikuid oli igal parameetril kuus.

5.3.1 Parameeter α

Parameeter α näitab lause kaalu valemis, milline on lause positsiooni skoori osakaal. Kui Kaili Müürisep [K.06] otsis parameetritele väärtusi, siis leidis ta, et kõige olulisem neist kolmest oli parameeter α. Tabelis 3 on välja toodud viis kõrgeimate keskmiste kattuvustega parameetrite kolmikut, kus α parameetri väärtus on kõrgem kui 0,5.

Tabel 3. Parameetrite kolmikud, kus α on suurem kui 0,5.

α	0,7	0,8	0,6	0,7	0,6	0,6
β	0,2	0,1	0,3	0,1	0,2	0,1
γ	0,1	0,1	0,1	0,2	0,2	0,3
Keskmine kattuvus	61,39%	61,34%	61,32%	61,28%	61,27%	60,26%

Kui vaadeldi tabelis 2 keskmiseid kattuvusi ja võrreldi tabeliga 3, siis täheldati, et tulemused on väga sarnased. Samas on tabelist näha, et keskmise kattuvuse langusega on seotud parameetri γ suurenemine ja parameetri α vähenemine.

5.3.2 Parameeter β

Parameeter β näitab lause kaalu valemis lause formaadi skoori osakaalu. Tabelis 4 on välja toodud 5 kõrgema keskmise kattuvusega parameetrite kolmikut, kus β parameetri väärtus on kõrgem kui 0,5.

Tabel 4. Parameetrite kolmikud, kus β on suurem kui 0,5.

α	0,3	0,2	0,2	0,1	0,1	0,1
β	0,6	0,7	0,6	0,8	0,7	0,6
γ	0,1	0,1	0,2	0,1	0,2	0,3
Keskmine kattuvus	61,25%	60,74%	59,94%	58,55%	55,83%	51,63%

Kui vaadata tabeli 4 keskmiseid kattuvusi, siis on näha, et keskmise kattuvuse langus on seotud parameetri α vähenemisega ja parameetri γ suurenemisega.

5.3.3 Parameeter γ

Parameeter γ näitab lause kaalu valemis, milline on lause sõnasageduste skoori osakaal. Tabelis 5 on välja toodud viis kõrgeimate keskmiste kattuvustega parameetrite kolmikut, kus γ parameetri väärtus on kõrgem kui 0,5.

Tabel 5. Parameetrite kolmikud, kus γ on suurem kui 0,5.

α	0,3	0,2	0,2	0,1	0,1	0,1
β	0,1	0,2	0,1	0,2	0,3	0,1
γ	0,6	0,6	0,7	0,7	0,6	0,8
Keskmine kattuvus	55,02%	51,05%	50,4%	47,64%	47,46%	45,18%

Kui vaadeldi tabeli 5 keskmisi kattuvusi, siis on täheldati, et need on tunduvalt madalamad kui keskmised kattuvused tabelites 3 ja 4. Samuti tundus huvitav, et kolmiku puhul, kus γ väärtus on 0,8, on keskmine kattuvus kõige madalam. Nagu tabelite 3 ja 4 puhul, siis ka tabelil 5 kehtib eripära, et keskmine kattuvus langeb, kui parameeter α väheneb ja parameeter γ suureneb.

5.3.4 Hinnang parameetritele

Lisas IV välja toodud tabel näitab, et kaheteistkümne parameetri kolmiku keskmine kattuvus oli suurem kui 60%. Nendes kolmikutes oli α keskmine väärtus 0,5 ning γ keskmine väärtus 0,29. See tähendab, et positsioonipõhine skoor on kõige olulisem ning

sõnasageduste põhine skoor on kõige vähemolulisem. Väidet toetab ka asjaolu, et kui vaadati tabeleid 3, 4 ja 5, siis täheldati asjaolu, et kui positsioonipõhine skoor moodustab lause kaalust üle poole, siis saadud keskmised kattuvused olid kõige kõrgemad. Seevastu kui sõnasageduste põhine skoor moodustab lause kaalust üle poole, siis keskmised kattuvused olid kõige madalamad. Samuti kehtis tabelites 3, 4 ja 5 eripära, mis viitas, et keskmised kattuvused langesid, kui parameeter α vähenes ja parameeter γ suurenes.
Seejärel leiti keskmine kattuvusprotsent arenduskorpuse artiklite kokkuvõtete ja EstSumi vanema versiooni genereeritud kokkuvõtete vahel. EstSumi vanem versioon kasutab lause kaalu valemis parameetreid 0,4, 0,4 ja 0,2. Keskmine kattuvusprotsent oli 58,08%. See viitas, et EstSumi vanema versiooni genereeritud kokkuvõtted sisaldavad vähem olulist informatsiooni kui tabelis 2 välja toodud kolmikutega genereeritud kokkuvõtted. Lisaks on EstSumi vanema versiooni keskmine kattuvus 4,16% madalam kui EstSumi uuemal versioonil, mis kasutab lause kaalu valemis samu parameetreid. Sellest järeldub, et lingvistilise mooduli ühendamine EstSumiga tõstab tulemuslikkust. Kindlasti tuleb aga möönda, et arenduskorpuses on kakskümmend artiklit ning käsitsi kokkuvõtete tegemine on subjektiivne. Samuti tuli saadud tulemusi kontrollida.

5.4 Tulemuste kontrollimine

Saadud tulemusi kontrolliti testkorpuse abil. Selleks kasutati testkorpuse artikleid, et leida käsitsi koostatud kokkuvõtete ja genereeritud kokkuvõtete kattuvusprotsendid ja arvutada nende põhjal keskmine kattuvusprotsent. Testkorpuse abil leiti keskmine kattuvus tabelis 2 väljatoodud kahe kõrgema keskmise kattuvusega parameetrite kolmikuga. Sama tehti ka tabelitega 3, 4 ja 5. Tabelis 6 on välja toodud saadud tulemused.

Tabel 6. Testkorpuse abil saadud keskmised kattuvused.

α	β	γ	**Keskmine kattuvus**
0,4	0,4	0,2	53,91%
0,5	0,4	0,1	53,17%
0,7	0,2	0,1	50,35%
0,8	0,1	0,1	50,35%
0,3	0,6	0,1	53,81%
0,2	0,7	0,1	53,9%
0,3	0,1	0,6	51,26%
0,2	0,2	0,6	52,84%

Tabelist 6 on saab järeldada, et keskmine kattuvus oli kõrge parameetrite kolmikutel, kus parameeter β oli kõrge. Samuti seda, et parameetri α tähtsus on langenud. Kui lähtuda ainult arenduskorpuse tulemustest, saanuks järeldada, et parameetri α väärtus tagab

kõrge keskmise kattuvuse. Kuid testkorpuse tulemused seda väidet ei toeta, vaid näitavad, et ka parameeter β on tähtis. Kui lähtuda mõlema korpuse tulemustest, siis kõrgemad keskmised kattuvused saadi, kui parameetrid α ja β olid osakaalult samatähtsad.
Seejärel leiti EstSumi eelmise versiooni kattuvusprotsendid ja keskmine kattuvusprotsent testkorpuse artiklitega, et saadud tulemusi võrrelda. Kui vaadata tulemusi parameetritega 0,4, 0,4 ja 0,2, siis EstSumi uuema versiooni keskmine kattuvus oli 53.91% ja EstSumi eelmise versiooni keskmine kattuvus oli 51,74%. Kuigi keskmiste kattuvuste erinevus vähenes 2,17%-ni, siis sellegipoolest olid EstSumi uuema versiooni genereeritud kokkuvõtted kõrgema keskmise kattuvusega. Arenduskorpuse abil saadud järeldus, et lingvistlise mooduli lisamine EstSumile tõstab tulemuslikkust, pidas paika ka testkorpuse puhul. Samuti tuleb märkida, et parameetrite kolmik 0,4, 0,4 ja 0,2 saavutas kõrgema keskmise kattuvuse arenduskorpuse artiklite seas. Lisaks saavutati kõrgeim keskmine kattuvus testkorpuse artiklite seas. Sellest järeldub, et keskmiselt kõige sobivam parameetrite kolmik EstSumi uuemale versioonile on 0,4, 0,4 ja 0,2. Ent siiski tuleb möönda, et käsitsi kokkuvõtete tegemine on subjektiivne ning üldiste järelduste tegemine oleks sobilikum suurema korpuse abil.

5.5 Tulevikusuund

Tehtud töö põhjal võib öelda, et lingvistilise mooduli lisamine EstSumile suurendab kokkuvõtja tulemuslikkust, kuid need tulemused on saadud väikeste korpuste pealt. Et saadud tulemusi kinnitada, tuleks tulevikus luua suurem korpus. Samuti tuleks korpuse loomisesse kaasata erinevaid inimesi, sest kui mitu inimest koostab ühele artiklile kokkuvõtte, on võimalik objektiivselt hinnata, mis on selles artiklis oluline enamiku inimeste jaoks. Sellise korpuse abil oleks võimalik täpsemalt leida lause kaalu valemile parameetrite väärtused, mis keskmiselt tagaksid kõrge kattuvusprotsendi.
Samuti tuleks korpusesse valida erinevat liiki artikleid, sest nii saaks kokkuvõtja parameetreid häälestada erinevat liiki artiklite jaoks. See võimaldaks aga EstSumile lisada artikliliigi seadistuse, millega saaks valida eelhäälestatud parameetrite kolmiku, mis keskmiselt tagaksid kõrge kattuvusprotsendi kindlale artikliliigile.

6 Kokkuvõte

Käesoleva töö eesmärgiks oli hinnata EstSumi tulemuslikkust, kui sellega on ühendatud lingvistiline moodul, mis võimaldaks kokkuvõtjal analüüsida sõnade algvorme.
Töös kirjeldatakse, kuidas kasutatakse EstNLTK lemmatiseerijat, et EstSumile lisada lingvistiline moodul. Lisaks on näidatud, kuidas koostati korpused, eesmärgiga leida EstSumi uuele versioonile uued parameetrid ja hinnata selle tulemuslikkust.
Töö tulemusena valmis uus versioon EstSumist, mis võtab arvesse sõnade algvorme. Samuti leiti, et kui võrrelda vana EstSumi genereeritud kokkuvõtet uue EstSumi genereeritud kokkuvõttega, siis uus EstSum eraldab rohkem autori jaoks olulist informatsiooni. EstSumi uus versioon eraldas arenduskorpuse peal 4,16% ja testkorpuse peal 2,17% rohkem olulist informatsiooni kui EstSumi vana versioon. Samuti leiti, et EstSumi uuele versioonile on kõige sobivam parameetrite kolmik 0,4, 0,4 ja 0,2. Samas tuleb mainida, et korpuste tegemine oli subjektiivne ja kolmekümne kokkuvõtte põhjal ei saa teha liiga üldiseid järeldusi.
EstSumi edasipidine arendussuund võiks olla seotud suurema korpuse loomisega, mille põhjal oleks võimalik anda üldisem järeldus EstSumi tulemuslikkuse kohta ja peenhäälestada parameetrite väärtuseid nii, et kasutajal oleks võimalik valida parameetrite väärtused vastavalt artikli liigile.

Viidatud kirjandus

[H.00] Dalianis H., 2000. SweSum - a text summarizer for Swedish. KTH-Stockholm. ftp://ftp.nada.kth.se/IPLab/TechReports/IPLab-174.pdf (08.05.2018).

[H.05] Hassel M., Dalianis H., 2005. Generation of Reference Summaries. KTH-Stockholm. https://pdfs.semanticscholar.org/a374/e9a8dbd0c395f28795625c8258e73bdbbb1a.pdf (14.05.2018).

[H.07] Hassel H. *Resource Lean and Portable Automatic Text Summerization*. PhD thesis, KTH School of Computer Science and Communication, 2007.

[HJ16] Orasmaa S., Petmanson T., Tkachenko A., Laur S., Kaalep H.-J., 2016. EstNLTK-NLP Toolkit for Estonian. Tartu Ülikool, Arvutiteaduse Instituut. https://pdfs.semanticscholar.org/d834/13f7f785aae067d49332239c3a36c346ba99.pdf (13.05.2018).

[I.01] Mani I. *Automatic Summarization*. Amsterdam: John Benjamins Publishing Company, 2001.

[K.06] Müürisep K. Eestikeelsete tekstide sisukokkuvõtjast estsum. *Keel ja Arvuti*, Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 6, 2006.

[M.03] Dalianis H., Hassel M., 2003. From SweSum to ScandSum - Automatic text summarization for the Scandinavian languages. KTH-Stockholm. https://people.dsv.su.se/~hercules/scandsum/ScandSumArsbog2002.pdf (10.05.2018).

[P.05a] Mutso P. Automaatne sisukokkuvõtete tegemine eestikeelsetest ajalehetekstidest: parameetrid ja hindamine, 2005. Diplomitöö. Tartu Ülikool, Arvutiteaduse Instituut.

[P.05b] Müürisep K., Mutso P. Estsum - estonian newspaper texts summarizer. Proceedings of The Second Baltic Conference on Human Language Technologies, Tallinn 2005.

[Sel08] Keili Sellik, 2008. Automaatse sisukokkuvõtja töö hindamine. Tartu Ülikool, Arvutiteaduse Instituut. http://lepo.it.da.ut.ee/~kaili/juhendamised/Baka_Sellik.pdf (12.05.2108).

[T.16a] Kaalep H.-J., Vaino T. Complete morphological analysis in the linguist's toolbox. In *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pages 9–16, 2016.

[T.16b] Muischnek K., Müürisep K., Puolakainen T. Estonian dependency treebank: from constraint grammar tagset to universal dependencies. Proceedings of LREC, 2016.
http://www.lrec-conf.org/proceedings/lrec2016/pdf/411_Paper.pdf (10.05.2018).

Lisad

I. Sõnade algvormide sagedustabel

olema 854.10
ei 224.75
ka 131.55
eesti 107.25
saama 102.05
aasta 87.05
siis 67.83
nii 67.50
pidama 62.08
tegema 61.03
võima 55.60
inimene 54.70
tulema 53.35
aeg 43.98
ütlema 43.20
minema 41.75
kas 40.12
suur 36.90
juba 36.77
veel 35.62
välja 35.50
uus 33.80
asi 33.15
kus 32.62
väga 32.23
mitte 32.15
mees 32.15
tahtma 32.00
pärast 30.15
teadma 30.10
rääkima 30.07
võtma 30.05
hakkama 29.55
jääma 29.48
hea 29.33
üle 28.80
kaks 28.30
ära 28.10
andma 27.62
esimene 26.93
tallinn 26.45
eest 26.02
töö 25.18
elu 25.18
kord 24.47
nägema 23.97
küll 23.63
kroon 23.30
kõige 22.70
ainult 22.63
kuidas 22.55
nüüd 22.05
tagasi 21.53
maailm 21.15
käima 21.02
raha 20.98
naine 20.90
viimane 20.77
sõna 20.73
panema 20.60
arvama 20.27
enam 20.18
vastu 20.15
päev 20.13
kogu 19.82
rohkem 19.70
seal 19.15
laps 19.07
riik 18.97
just 18.90
näide 18.88
ju 18.87
leidma 18.27
isegi 18.18
peale 18.12
läbi 17.75
siin 17.72
firma 17.65
miljon 17.65
vaatama 16.95
osa 16.90
ikka 16.75
kokku 16.52
keel 16.50
kolm 16.27
juht 16.20
käsi 15.80
pool 15.75
kirjutama 15.75
i 15.70
tegelikult 15.67
elama 15.40
auto 15.10
nimi 15.05
maja 14.73
kasutama 14.63
koos 14.55
enne 14.50
lihtsalt 14.45
miks 14.43
vene 14.30
noor 14.28
praegu 14.25
ette 13.68
jõudma 13.62
raamat 13.60
siiski 13.53
euroopa 13.40
vana 13.40
kohta 13.40
mõte 13.10
vahel 12.97
silm 12.93
muidugi 12.40
maksma 12.40
mõtlema 12.38
jaoks 12.32
küsimus 12.25
pilt 12.22
võimalus 12.20
tartu 12.20
tundma 11.92
lõpp 11.85
lugema 11.78
linn 11.35
kunagi 11.30
tooma 11.22
küsima 11.22
suurem 11.20
hästi 11.18
alla 11.10
eriti 11.08
pea 11.05
kohe 11.05
võimalik 11.05
eestlane 11.00
nädal 10.98
soome 10.97
all 10.97
ostma 10.88

hiljem 10.70
rahvas 10.70
näitama 10.70
töötama 10.67
maa 10.58
sisse 10.55
president 10.45
kunst 10.33
maha 10.32
parem 10.30
paar 10.30
viis 10.30
pakkuma 10.22
film 10.22
hoopis 10.20
alati 10.15
suutma 10.10
edasi 10.07
poolt 9.98
tunduma 9.95
viima 9.93
liiga 9.93
kodu 9.90
müüma 9.88
sõber 9.78
juurde 9.75
näitus 9.75
sõitma 9.70
kunstnik 9.57
eelmine 9.55
algus 9.47
tegemine 9.40
umbes 9.30
tee 9.30
üles 9.25
aru 9.22
ameerika 9.20
venemaa 9.20
lugu 9.18
puhul 9.17
kuu 9.15
toimuma 9.00
uskuma 8.90
mängima 8.88
järgmine 8.85
protsent 8.82
laskma 8.80
tähendama 8.75
järgi 8.65
lisa 8.65
kümme 8.62
kiri 8.62
vaja 8.57
samuti 8.55
ühe 8.53
probleem 8.52
ees 8.50
üldse 8.47
saatma 8.40
jutt 8.35
muusika 8.30
hoidma 8.28
kuuluma 8.22
ootama 8.12
põhjus 7.97
asuma 7.93
alles 7.92
autor 7.92
jätma 7.92
ilmselt 7.90
istuma 7.90
ajama 7.87
hetk 7.85
parim 7.85
usa 7.83
muutuma 7.82
pikk 7.80
ülikool 7.80
raske 7.77
tund 7.75
lubama 7.72
vanem 7.70
ilma 7.67
lõpuks 7.67
endine 7.67
tulemus 7.65
sajand 7.60
isa 7.58
kindlasti 7.55
kinni 7.47
meenutama 7.45
varem 7.40
kool 7.40
meeldima 7.38
valitsus 7.37
oskama 7.30
samas 7.28
plaat 7.28
õige 7.27
idee 7.25
neli 7.23
ekspress 7.20
erinev 7.17
the 7.12
kinnitama 7.12
püüdma 7.12
nõukogu 7.12
enamik 7.10
üsna 7.08
hind 7.03
vähemalt 7.03
sinna 7.00
juures 7.00
ümber 7.00
pank 6.98
poiss 6.97
mart 6.93
otsima 6.92
tekkima 6.90
raudtee 6.90
täiesti 6.88
kaasa 6.88
jooksul 6.83
täis 6.82
kindel 6.80
ühiskond 6.78
huvi 6.78
teema 6.77
päris 6.75
selge 6.72
ajalugu 6.72
pigem 6.72
kohalik 6.72
kuulma 6.72
kandma 6.70
tegelema 6.60
kutsuma 6.60
ilmuma 6.60
vähem 6.58
jälle 6.58
algama 6.57
poole 6.55
meel 6.55
teater 6.52
vastama 6.50
väike 6.50
valge 6.47
koht 6.35
tüdruk 6.35
nimetama 6.33
ainus 6.32
saksamaa 6.30
ajal 6.25
läti 6.23
rootsi 6.22
taga 6.20
andres 6.18
ema 6.18
kell 6.15
tänav 6.12

astuma 6.10
seisma 6.10
hääl 6.08
eesmärk 6.05
täpselt 6.05
oluline 6.05
toomas 6.05
terve 6.00
õppima 6.00
vist 5.98
meri 5.95
vein 5.95
rahvusvaheline 5.92
lahti 5.90
olukord 5.90
nõudma 5.90
aitama 5.88
tunnistama 5.82
kultuur 5.80
tänapäev 5.78
tekst 5.77
valima 5.75
ikkagi 5.73
liige 5.73
muutma 5.70
turg 5.70
täna 5.70
valmis 5.67
tegevus 5.65
ilus 5.65
väitma 5.65
projekt 5.62
tõesti 5.53
mõistma 5.53
ajakirjanik 5.53
ligi 5.53
venelane 5.53
töötaja 5.52
omanik 5.52
suurim 5.50
must 5.50
asemel 5.50
otsus 5.47
õhtu 5.47
otsustama 5.47
lihtne 5.45
selguma 5.45
ruum 5.45
vesi 5.42
uks 5.40
nägu 5.38
kolmas 5.35
teos 5.35
praegune 5.32
moskva 5.32
roll 5.25
kohus 5.20
tõttu 5.20
hulk 5.20
juhtuma 5.15
teatama 5.15
mööda 5.12
uurima 5.12
pisut 5.12
peaaegu 5.10
teenima 5.10
ometi 5.08
plaan 5.08
tunne 5.07
leping 5.05
kaudu 5.05
jalg 5.00
kõrval 5.00
peeter 4.98
tegelane 4.97
huvitav 4.97
ettevõte 4.95
tavaline 4.95
liit 4.95
jah 4.93
esitama 4.92
kirjanik 4.90
müük 4.90
politsei 4.90
kasvama 4.90
tulevik 4.88
jüri 4.88
kuhu 4.85
äri 4.85
sattuma 4.85
võrdlema 4.83
lõpetama 4.83
palk 4.82
üritama 4.80
artikkel 4.80
kuulama 4.80
leht 4.78
seega 4.78
võim 4.77
avaldama 4.75
dollar 4.75
näiteks 4.75
looma 4.75
tähtis 4.75
surm 4.72
õigus 4.70
tuttav 4.70
arvuti 4.67
vastus 4.67
poliitiline 4.65
vaba 4.65
mood 4.60
inglise 4.60
alustama 4.58
vähe 4.58
direktor 4.58
esi 4.57
suhe 4.55
pärit 4.55
ärimees 4.53
eriline 4.53
näitleja 4.53
puuduma 4.53
ajaleht 4.50
paistma 4.50
kuidagi 4.50
loom 4.50
ajakiri 4.48
jõud 4.47
itaalia 4.45
soovima 4.45
a 4.45
lootma 4.45
tuhat 4.45
avalik 4.42
veidi 4.40
mäng 4.40
märts 4.40
hommik 4.38
new 4.38
internet 4.38
hoone 4.38
tohtima 4.35
järele 4.35
hulgas 4.33
poliitik 4.33
kust 4.33
saksa 4.33
seni 4.33
esinema 4.33
tõeline 4.30
paluma 4.30
natuke 4.28
märk 4.28
piir 4.27
kartma 4.25
sööma 4.25
s 4.23
peal 4.22
klient 4.20

selg 4.18
stiil 4.17
sada 4.17
partei 4.17
kujutama 4.17
tekitama 4.17
loomulikult 4.15
andmed 4.15
publik 4.15
mäletama 4.15
perekond 4.10
korter 4.10
pere 4.10
number 4.08
sobima 4.08
lava 4.08
võibolla 4.05
miljard 4.05
abi 4.05
pääsema 4.03
soov 4.03
aprill 4.00
paremini 4.00
liikuma 4.00
arvamus 4.00
nimelt 3.98
halb 3.98
ehitama 3.98
jaanuar 3.98
veebruar 3.95
jaan 3.95
lööma 3.95
preemia 3.95
näima 3.95
kuulus 3.95
tavaliselt 3.93
sündima 3.93
lausa 3.93
mõjuma 3.90
ajakirjandus 3.88
materjal 3.88
seletama 3.88
korraldama 3.88
kuus 3.85
seetõttu 3.85
foto 3.85
muidu 3.83
olev 3.83
tähelepanu 3.83
üha 3.83
looming 3.83
eks 3.82
muuseum 3.82
põlvkond 3.80
seas 3.80
sealt 3.80
jooma 3.80
valik 3.80
suvi 3.78
kõlama 3.78
vahele 3.77
no 3.77
taas 3.77
ala 3.77
esimees 3.75
keskmine 3.75
tõstma 3.75
lisama 3.75
puhas 3.73
enamasti 3.73
jagama 3.73
tõusma 3.73
london 3.73
kaduma 3.72
juhtima 3.72
kirjandus 3.72
vajama 3.72
mark 3.70
võitma 3.70
tegija 3.70
juut 3.70
kiiresti 3.70
ringi 3.70
kolleeg 3.68
eri 3.68
isiklik 3.68
album 3.67
maal 3.67
keha 3.67
tegu 3.65
peaminister 3.65
tõmbama 3.65
koer 3.63
tulviste 3.62
selgitama 3.62
minut 3.62
punane 3.60
jne 3.60
sein 3.60
avama 3.57
lugeja 3.57
sees 3.55
hiljuti 3.55
sageli 3.55
saabuma 3.55
kuju 3.55
poeg 3.55
mulje 3.55
york 3.55
kogemus 3.55
tase 3.53
arhitekt 3.53
kõrge 3.52
estonia 3.50
tugev 3.50
meelest 3.48
pidevalt 3.45
otse 3.45
riie 3.43
külg 3.43
õnn 3.43
piisavalt 3.42
holland 3.42
seisukoht 3.40
isik 3.40
laul 3.40
suhtes 3.40
seadus 3.40
väiksem 3.38
pärnu 3.38
riigikogu 3.38
sõda 3.38
sügis 3.38
rida 3.38
kohale 3.38
uudis 3.38
välismaa 3.37
surema 3.35
vorm 3.35
arst 3.35
kavatsema 3.35
meeter 3.35
tühi 3.35
prantsuse 3.35
etendus 3.35
kasu 3.33
hotell 3.33
tütar 3.32
mil 3.30
poliitika 3.30
tõsine 3.27
sakslane 3.27
pealt 3.25
areng 3.25
soovitama 3.25
aktsia 3.25
korralik 3.25
klass 3.25
esindaja 3.25
leedu 3.25

lahkuma 3.25
tükk 3.23
tõenäoliselt 3.22
vahe 3.20
juhtum 3.20
alates 3.20
elav 3.20
sõnum 3.20
reklaam 3.20
sadam 3.20
seltskond 3.18
meedia 3.18
järel 3.18
veelgi 3.18
vend 3.18
seitse 3.18
amet 3.17
sarnane 3.15
intervjuu 3.15
jumal 3.15
süsteem 3.15
siia 3.15
osalema 3.15
kallis 3.15
jätkuma 3.15
vajalik 3.15
hansapank 3.15
moodne 3.15
kirik 3.15
haigla 3.13
väärtus 3.13
kaua 3.13
nr 3.13
aken 3.12
nn 3.10
ameeriklane 3.10
helistama 3.10
eelkõige 3.10
täitma 3.10
vabariik 3.10
kirjastus 3.08
värske 3.08
ooper 3.08
minister 3.08
tihti 3.07
viga 3.05
uuesti 3.05
toit 3.05
paik 3.05
saade 3.02
kelam 3.02
lavastus 3.02
kaotama 3.02
suu 3.02
abil 3.00
kiire 3.00
hoolimata 3.00
süda 3.00
kontsert 3.00
ülesanne 3.00
armastama 3.00
pilk 2.98
mägi 2.98
jälgima 2.98
seepärast 2.97
telefon 2.95
erastamine 2.95
viskama 2.93
sellepärast 2.92
mõis 2.92
tegutsema 2.92
taha 2.92
paraku 2.92
laud 2.90
sisu 2.90
sündmus 2.90
asuv 2.90
tore 2.90
arvestama 2.90
kirjeldama 2.90
meeskond 2.88
protsess 2.88
kohtuma 2.88
pudel 2.88
järjest 2.87
kõvasti 2.85
ajalooline 2.85
tänu 2.85
kõne 2.85
saal 2.85
leiduma 2.83
hans 2.83
masin 2.83
mets 2.83
tähendus 2.83
hindama 2.82
juhataja 2.82
luule 2.82
avastama 2.82
laulma 2.82
algul 2.82
kaheksa 2.80
tuba 2.80
valimine 2.80
üritus 2.80
koostöö 2.80
lääs 2.80
ammu 2.80
vajadus 2.80
nali 2.78
abikaasa 2.78
kerge 2.78
tüüp 2.78
poolest 2.77
võõras 2.77
väide 2.77
august 2.77
ühte 2.75
suhteliselt 2.75
märkama 2.75
otsa 2.75
nimel 2.75
piirkond 2.75
riia 2.75
tehas 2.75
lahendus 2.75
elanik 2.73
klubi 2.73
pind 2.73
haigus 2.73
pealkiri 2.73
kaup 2.73
päritolu 2.73
siinne 2.72
info 2.72
vedama 2.72
mullu 2.70
korral 2.70
laar 2.70
savisaar 2.70
pood 2.68
õnnestuma 2.68
sotsiaalne 2.67
uuring 2.67
anne 2.67
paber 2.67
tiina 2.67
moodustama 2.67
proovima 2.65
kõrvale 2.65
vahetama 2.65
käigus 2.65
märkima 2.65
kalev 2.65
tasuma 2.65
ametnik 2.62
odav 2.62
etv 2.62
jooksma 2.62
avalikkus 2.62

korrus 2.62
lavastaja 2.62
valmistama 2.62
huvitama 2.60
kõva 2.60
siit 2.60
klaas 2.60
rahvuslik 2.60
tarmo 2.60
värv 2.60
mihkel 2.60
õpetaja 2.60
lai 2.58
tellima 2.58
saar 2.58
kohaselt 2.58
maitse 2.58
alt 2.58
reegel 2.58
paul 2.58
romaan 2.58
õde 2.58
restoran 2.58
viin 2.57
korraga 2.57
punkt 2.57
tõepoolest 2.55
hiina 2.55
kasutamine 2.55
loodus 2.53
häda 2.52
seejärel 2.52
langema 2.52
kuulutama 2.52
lennuk 2.52
veri 2.52
taust 2.52
arv 2.52
lennart 2.50
soomlane 2.50
vabadus 2.50
hulka 2.50
nokia 2.50
toetama 2.50
keeruline 2.50
of 2.50
tingimus 2.50
de 2.50
võistlus 2.48
sposato 2.48
variant 2.48
inglismaa 2.48
sundima 2.47
suurepärane 2.47
narva 2.47
edu 2.45
öö 2.45
koguma 2.45
külm 2.45
von 2.43
noormees 2.43
eelistama 2.43
teadlane 2.43
mootor 2.42
mõju 2.42
kukkuma 2.42
komme 2.42
valgus 2.42
haridus 2.40
tõsiselt 2.40
koguni 2.40
kestma 2.40
kilomeeter 2.40
juhatus 2.40
sõltuma 2.40
tootmine 2.40
samm 2.40
soe 2.40
väljas 2.40
tasu 2.38
majandus 2.38
linnapea 2.38
loobuma 2.38
vale 2.38
advokaat 2.38
kõigepealt 2.38
mari 2.38
õnnelik 2.38
unt 2.38
sisaldama 2.38
konkreetne 2.37
võimaldama 2.35
riiklik 2.35
tänane 2.35
endiselt 2.35
lähedal 2.35
õpetama 2.35
detsember 2.35
vaataja 2.35
vähene 2.35
park 2.35
julgema 2.33
äkki 2.33
keskel 2.33
katse 2.33
põhjal 2.33
tähtsam 2.33
nõunik 2.33
keerama 2.32
edukas 2.32
vaene 2.32
kriitik 2.32
ots 2.32
pealegi 2.32
tänavu 2.30
kummaline 2.30
noorem 2.30
unustama 2.30
võitlus 2.30
tiit 2.30
grupp 2.30
vahepeal 2.30
omavahel 2.30
suund 2.30
seadma 2.30
indrek 2.27
osutuma 2.27
varsti 2.27
is 2.27
toode 2.25
barbie 2.25
investor 2.25
käik 2.25
kaunis 2.25
lühike 2.25
õhk 2.25
keskus 2.25
kodanik 2.25
normaalne 2.23
professor 2.23
lk 2.23
mujal 2.23
lause 2.23
rikas 2.23
armastus 2.22
peamiselt 2.22
teenus 2.20
pealinn 2.20
arhitektuur 2.20
tarvis 2.20
tunduvalt 2.20
määrama 2.20
valitsema 2.20
staar 2.20
hinnang 2.20
n 2.20
karu 2.20
fotograaf 2.18
uurimine 2.17
oht 2.17

luuletus 2.17
t 2.17
kätte 2.17
erakond 2.17
linnavalitsus 2.17
rein 2.17
st 2.17
paavo 2.17
kohtumine 2.15
nina 2.15
vaim 2.15
lootus 2.15
ohver 2.15
kommenteerima 2.15
ühine 2.15
prantsusmaa 2.15
tootma 2.15
jürgenson 2.15
kuluma 2.13
lükkama 2.13
raadio 2.12
juuni 2.12
suhtlema 2.12
õpilane 2.12
itaallane 2.12
ühtlasi 2.12
kahjuks 2.12
urmas 2.12
hull 2.10
kvaliteet 2.10
tegelik 2.10
jätkama 2.10
saladus 2.10
seekord 2.10
laev 2.10
küla 2.10
kõrgem 2.10
vaatamata 2.10
pikem 2.10
helsingi 2.10
magama 2.10
esindama 2.10
kuum 2.10
naerma 2.10
teadmine 2.08
lõppema 2.08
kohal 2.08
niivõrd 2.08
ilves 2.08
paks 2.08
klassikaline 2.08
mure 2.08
tõnu 2.08
kangelane 2.08
siiani 2.08
saavutama 2.07
kujunema 2.07
objekt 2.07
süüdistama 2.07
osakond 2.07

II. Stoppsõnade loend

aga
ega
ehk
elik
ent
ja
kui
kuid
kuni
nagu
ning
vaid
või
ehkki
et
justkui
kuigi
kuna
nagu
olgugi
otsekui
selmet
sest
iga
igasugune
igaüks
ise
iseenese
keegi
kes
kõik
kumbki
milline
mina
mingi
mingisugune
mis
miski
missugune
mitmesugune
mitu
mõlema
mõni
muu
nemad
niipaljuke
niisugune
oma
omaenese
omasugune
palju
säärane
sama
samasugune
see
seesama
seesamune
selline
sihuke
sina
sinusugune
teiesugune
teine
teineteise
teistsugune
tema
temasugune
too
toosama
üks
ükski
üksteise

III. Näide algtekstis, märgendatud tekstist, käsitsi koostatud kokkuvõttest ja EstSumi genereeritud kokkuvõttest

Algtekst

Brüssel tuleb täna välja uue eelarvekavaga

Täna Euroopa Komisjoni avaldatav järgmine pikaajaline eelarvekava on seotud vähemalt kahe suure probleemiga: Suurbritannia ehk ühe suurema netomaksja lahkumine Euroopa Liidust ning Brüsseli soov kasutada suuremat tükki eelarvest tsentraalselt.

Brüsseli soov peegeldab tahet järgida Prantsusmaa presidendi Emmanuel Macroni üleskutset muuta Euroopa Liit tulevaste kriiside puhuks rahanduslikult võimekamaks, kirjutas Financial Times, kellel on õnnestunud eelarveperspektiivi mustand kätte saada.

Tegemist on 2021.–2027. aastani hõlmava eelarvekava esimese ettepaneku ehk perspektiiviga, kus arvatavasti konkreetseid arve veel ei ole. Kindlad summad võivad tulla juuni keskpaigas, aga riikidevahelised läbirääkimised nii omavahel kui ka Euroopa Komisjoniga kestavad ilmselt kaks aastat.

Vahepeal toimuvad aga Euroopa Parlamendi valimised, mille tulemusena vahetub Euroopa Komisjoni president ja ilmselt ka paljud volinikud. Vahepeal toimuvad mitmes riigis, näiteks Eestis, ka parlamendivalimised, mis samuti võib jõuvahekordi muuta. Eelarvekava vastuvõtmiseks peavad sellega nõustuma kõik Euroopa Liidu liikmesriigid. Riikidevahelised läbirääkimised eelarvekava üle on proovikivi riigipeadele. Valitsusjuhid teavad väga hästi, et nende läbirääkimisvõimekust mõõdetakse eurodes, õigemini miljonites ja miljardites eurodes. Seitsmeaastase eelarve kogumaht on triljon eurot.

«See on alati komplitseeritud. See võtab alati eeldatust rohkem aega. Ja seal on alati suur hulk dramaatilisust. Alati,» ütles eelarveläbirääkimistega tihedalt seotud Euroopa Komisjoni ametnik. «On hämmastav, et me üldse sellega toime tuleme. Aga me tuleme,» lisas ta.

Juba praegu väljendub kahe suurriigi Prantsusmaa ja Saksamaa suurim lahkheli suhtumises vajadusse Euroopa kriisifondide puhvreid suurendada.

«Vajame suuremat manööverdamisruumi,» ütles Euroopa Komisjoni ametnik Financial Timesile, nimetades põhjusena ohtu, et Euroopa fondid, nagu Euroopa Finantsstabiilsusmehhanism (EFSM), võivad kaotada kõrgeima (AAA) krediidireitingu.

Brüsseli ametnikud püüavad Euroopa laenusüsteemi muuta nii, et see ei põrkuks mitme riigi, eelkõige Saksamaa ja Hollandi vastuseisule. Saksamaa kantsleri Angela Merkeli sõnum teistele valitsusjuhtidele märtsikuisel tippkohtumisel oli, et fondide fiskaalvõimekus tuleb hoida nii madalal kui võimalik ning seda tuleks kasutada pigem investeeringuteks kui lihtsalt majandustoetusteks.

Nagu öeldud, lööb Suurbritannia lahkumine Euroopa Liidust eelarvesse märgatava eelarveaugu – 10–15 miljardit eurot aastas. Selle katmiseks on plaanis suurendada sissemakset praeguselt ühelt protsendilt 1,2–1,3 protsendile sisemajanduse kogutoodangust. Kaalu-

takse maksubaasi laiendamist, näiteks ettevõtte tulumaksu ühtlustamist.
Samas on esimest korda oodata ühtekuuluvusfondide kahanemist. Seni on eurotoetuste jagamise aluseks peamiselt üks kriteerium. Kuuldavasti kaalutakse uue perspektiivi väljatöötamisel veel lisakriteeriume: näiteks pagulaste vastuvõtmisest tingitud koormus ja noorte tööpuudus. Peale selle plaanitakse euroraha jagamisel arvesse võtta ka majandusreforme ja seaduste vastamist õigusriigi kriteeriumitele.
Sellega saavad ilmselt suurima löögi Poola ja Ungari, kes on eelmistel aastatel olnud suurimad Euroopa Liidu toetuste saajad. Bloombergi andmetel sai Poola aastatel 2014–2016 Euroopa Liidult netotoetusi keskmiselt 10,1 miljardit eurot aastas.
Samas on diplomaadid hoiatanud, et nende karistamine võib avada Pandora laeka ja päädida veel mõne riigi lahkumisega Euroopa Liidust. Igatahes Poola on väljendanud sellele kavatsusele juba vastuseisu.
«Näeme selles soovi avaldada mõnedele riikidele enne läbirääkimisi poliitilist survet,» lausus uudisteagentuurile AP Poola välisminister Jacek Czaputowicz. «Seepärast suhtume taolistesse ideedesse väga negatiivselt.»
Bloombergi andmetel plaanitakse vähendada põllumajandustoetusi, seni nullilähedasi kaitsekulutusi aga märkimisväärselt suurendada. Märkimisväärselt on kavas tugevdada ka piirivalvet, suurendades töötajate arvu praegusega võrreldes enam kui viis korda.
«See on hädavajalik Euroopa Liidu piiri mitme lõigu, näiteks Kreeka saarte tõttu,» rääkis Euroopa Komisjoni allikas Reutersile. Uudisteagentuuri teatel kulutatakse käimasoleval seitseaastakul Euroopa Liidu piiride kaitseks neli miljardit eurot, piirivalvurite arvu suurendamiseks vähemalt 3000 võrra tuleb eelarvet kasvatada 25 miljardi euroni.

Märgendatud tekst

<div0><head><hi rend="bold»Brüssel tuleb täna välja uue eelarvekavaga</hi></head>

<s><hi rend="bold»Täna Euroopa Komisjoni avaldatav järgmine pikaajaline eelarvekava on seotud vähemalt kahe suure probleemiga: Suurbritannia ehk ühe suurema netomaksja lahkumine Euroopa Liidust ning Brüsseli soov kasutada suuremat tükki eelarvest tsentraalselt.</hi></s>

<s>Brüsseli soov peegeldab tahet järgida Prantsusmaa presidendi Emmanuel Macroni üleskutset muuta Euroopa Liit tulevaste kriiside puhuks rahanduslikult võimekamaks, kirjutas Financial Times, kellel on õnnestunud eelarveperspektiivi mustand kätte saada.</s>

<s>Tegemist on 2021.–2027. aastani hõlmava eelarvekava esimese ettepaneku ehk perspektiiviga, kus arvatavasti konkreetseid arve veel ei ole.</s>

<s>Kindlad summad võivad tulla juuni keskpaigas, aga riikidevahelised läbirääkimised nii omavahel kui ka Euroopa Komisjoniga kestavad ilmselt kaks aastat.</s>

<s>Vahepeal toimuvad aga Euroopa Parlamendi valimised, mille tulemusena vahetub Euroopa Komisjoni president ja ilmselt ka paljud volinikud.</s>
<s>Vahepeal toimuvad mitmes riigis, näiteks Eestis, ka parlamendivalimised, mis samuti võib jõuvahekordi muuta.</s>
<s>Eelarvekava vastuvõtmiseks peavad sellega nõustuma kõik Euroopa Liidu liikmesriigid.</s>

<s>Riikidevahelised läbirääkimised eelarvekava üle on proovikivi riigipeadele.</s>
<s>Valitsusjuhid teavad väga hästi, et nende läbirääkimisvõimekust mõõdetakse eurodes, õigemini miljonites ja miljardites eurodes.</s>
<s>Seitsmeaastase eelarve kogumaht on triljon eurot.</s>

<s>«See on alati komplitseeritud. See võtab alati eeldatust rohkem aega. Ja seal on alati suur hulk dramaatilisust. Alati,» ütles eelarveläbirääkimistega tihedalt seotud Euroopa Komisjoni ametnik.</s>
<s>«On hämmastav, et me üldse sellega toime tuleme. Aga me tuleme,» lisas ta.</s>

<s>Juba praegu väljendub kahe suurriigi Prantsusmaa ja Saksamaa suurim lahkheli suhtumises vajadusse Euroopa kriisifondide puhvreid suurendada.</s>

<s>«Vajame suuremat manööverdamisruumi,» ütles Euroopa Komisjoni ametnik Financial Timesile, nimetades põhjusena ohtu, et Euroopa fondid, nagu Euroopa Finantsstabiilsusmehhanism (EFSM), võivad kaotada kõrgeima (AAA) krediidireitingu.</s>

<s>Brüsseli ametnikud püüavad Euroopa laenusüsteemi muuta nii, et see ei põrkuks mitme riigi, eelkõige Saksamaa ja Hollandi vastuseisule.</s>
<s>Saksamaa kantsleri Angela Merkeli sõnum teistele valitsusjuhtidele märtsikuisel tippkohtumisel oli, et fondide fiskaalvõimekus tuleb hoida nii madalal kui võimalik ning seda tuleks kasutada pigem investeeringuteks kui lihtsalt majandustoetusteks.</s>

<s>Nagu öeldud, lööb Suurbritannia lahkumine Euroopa Liidust eelarvesse märgatava eelarveaugu – 10–15 miljardit eurot aastas.</s>
<s>Selle katmiseks on plaanis suurendada sissemakset praeguselt ühelt protsendilt 1,2–1,3 protsendile sisemajanduse kogutoodangust.</s>
<s>Kaalutakse maksubaasi laiendamist, näiteks ettevõtte tulumaksu ühtlustamist.</s>

<s>Samas on esimest korda oodata ühtekuuluvusfondide kahanemist.</s>
<s>Seni on eurotoetuste jagamise aluseks peamiselt üks kriteerium.</s>
<s>Kuuldavasti kaalutakse uue perspektiivi väljatöötamisel veel lisakriteeriume: näiteks pagulaste vastuvõtmisest tingitud koormus ja noorte tööpuudus.</s>
<s>Peale selle plaanitakse euroraha jagamisel arvesse võtta ka majandusreforme ja seaduste vastamist õigusriigi kriteeriumitele.</s>

<s>Sellega saavad ilmselt suurima löögi Poola ja Ungari, kes on eelmistel aastatel olnud suurimad Euroopa Liidu toetuste saajad.</s>
<s>Bloombergi andmetel sai Poola aastatel 2014–2016 Euroopa Liidult netotoetusi keskmiselt 10,1 miljardit eurot aastas.</s>

<s>Samas on diplomaadid hoiatanud, et nende karistamine võib avada Pandora laeka ja päädida veel mõne riigi lahkumisega Euroopa Liidust.</s>
<s>Igatahes Poola on väljendanud sellele kavatsusele juba vastuseisu.</s>

<s>«Näeme selles soovi avaldada mõnedele riikidele enne läbirääkimisi poliitilist survet,» lausus uudisteagentuurile AP Poola välisminister Jacek Czaputowicz.</s>
<s>«Seepärast suhtume taolistesse ideedesse väga negatiivselt.»</s>

<s>Bloombergi andmetel plaanitakse vähendada põllumajandustoetusi, seni nullilähedasi kaitsekulutusi aga märkimisväärselt suurendada.</s>
<s>Märkimisväärselt on kavas tugevdada ka piirivalvet, suurendades töötajate arvu praegusega võrreldes enam kui viis korda.</s>

<s>«See on hädavajalik Euroopa Liidu piiri mitme lõigu, näiteks Kreeka saarte tõttu,» rääkis Euroopa Komisjoni allikas Reutersile.</s>

<s>Uudisteagentuuri teatel kulutatakse käimasoleval seitseaastakul Euroopa Liidu piiride kaitseks neli miljardit eurot, piirivalvurite arvu suurendamiseks vähemalt 3000 võrra tuleb eelarvet kasvatada 25 miljardi euroni.</s>
</p>
</div0>

Käsitsi koostatud kokkuvõtete

Brüssel tuleb täna välja uue eelarvekavaga
Täna Euroopa Komisjoni avaldatav järgmine pikaajaline eelarvekava on seotud vähemalt kahe suure probleemiga: Suurbritannia ehk ühe suurema netomaksja lahkumine Euroopa Liidust ning Brüsseli soov kasutada suuremat tükki eelarvest tsentraalselt.
Tegemist on 2021.–2027. aastani hõlmava eelarvekava esimese ettepaneku ehk perspektiiviga, kus arvatavasti konkreetseid arve veel ei ole.
Vahepeal toimuvad aga Euroopa Parlamendi valimised, mille tulemusena vahetub Euroopa Komisjoni president ja ilmselt ka paljud volinikud.
Riikidevahelised läbirääkimised eelarvekava üle on proovikivi riigipeadele.
Juba praegu väljendub kahe suurriigi Prantsusmaa ja Saksamaa suurim lahkheli suhtumises vajadusse Euroopa kriisifondide puhvreid suurendada.
Brüsseli ametnikud püüavad Euroopa laenusüsteemi muuta nii, et see ei põrkuks mitme riigi, eelkõige Saksamaa ja Hollandi vastuseisule.
Samas on esimest korda oodata ühtekuuluvusfondide kahanemist.
Sellega saavad ilmselt suurima löögi Poola ja Ungari, kes on eelmistel aastatel olnud suurimad Euroopa Liidu toetuste saajad.
Poola on väljendanud sellele kavatsusele juba vastuseisu.
Bloombergi andmetel plaanitakse vähendada põllumajandustoetusi, seni nullilähedasi kaitsekulutusi aga märkimisväärselt suurendada.

EstSumi genereeritud kokkuvõte

<s><hi rend="bold»Täna Euroopa Komisjoni avaldatav järgmine pikaajaline eelarvekava on seotud vähemalt kahe suure probleemiga: Suurbritannia ehk ühe suurema netomaksja lahkumine Euroopa Liidust ning Brüsseli soov kasutada suuremat tükki eelarvest tsentraalselt.</hi></s>
<s>Brüsseli soov peegeldab tahet järgida Prantsusmaa presidendi Emmanuel Macroni üleskutset muuta Euroopa Liit tulevaste kriiside puhuks rahanduslikult võimekamaks, kirjutas Financial Times, kellel on õnnestunud eelarveperspektiivi mustand kätte saada.</s>
<s>Vahepeal toimuvad aga Euroopa Parlamendi valimised, mille tulemusena vahetub Euroopa Komisjoni president ja ilmselt ka paljud volinikud.</s>
<s>Riikidevahelised läbirääkimised eelarvekava üle on proovikivi riigipeadele.</s>

<s>Juba praegu väljendub kahe suurriigi Prantsusmaa ja Saksamaa suurim lahkheli suhtumises vajadusse Euroopa kriisifondide puhvreid suurendada.</s>
<s>Brüsseli ametnikud püüavad Euroopa laenusüsteemi muuta nii, et see ei põrkuks mitme riigi, eelkõige Saksamaa ja Hollandi vastuseisule.</s>
<s>Nagu öeldud, lööb Suurbritannia lahkumine Euroopa Liidust eelarvesse märgatava eelarveaugu – 10–15 miljardit eurot aastas.</s>
<s>Sellega saavad ilmselt suurima löögi Poola ja Ungari, kes on eelmistel aastatel olnud suurimad Euroopa Liidu toetuste saajad.</s>

IV. Parameetrite kolmikute keskmine kattuvus arnedus-korpusel

α	β	γ	**Keskmine kattuvus**
0,4	0,4	0,2	62,24%
0,5	0,4	0,1	61,68%
0,4	0,3	0,3	61,65%
0,4	0,5	0,1	61,46%
0,7	0,2	0,1	61,39%
0,8	0,1	0,1	61,34%
0,6	0,3	0,1	61,32%
0,7	0,1	0,2	61,28%
0,6	0,2	0,2	61,27%
0,3	0,6	0,1	61,25%
0,5	0,2	0,3	61,08%
0,5	0,3	0,2	60,85%
0,2	0,7	0,1	60,74%
0,3	0,5	0,2	60,7%
0,3	0,4	0,3	60,27%
0,6	0,1	0,3	60,26%
0,2	0,6	0,2	59,94%
0,4	0,2	0,4	59,41%
0,1	0,8	0,1	58,55%
0,3	0,3	0,4	58,2%
0,2	0,4	0,4	57,94%
0,2	0,5	0,3	57,47%
0,5	0,1	0,4	57,21%
0,3	0,2	0,5	56,49%
0,1	0,7	0,2	55,83%
0,4	0,1	0,5	55,65%
0,2	0,3	0,5	55,3%
0,3	0,1	0,6	55,02%
0,1	0,6	0,3	51,63%
0,2	0,2	0,6	51,05%
0,1	0,5	0,4	50,44%
0,2	0,1	0,7	50,4%
0,1	0,2	0,7	47,64%
0,1	0,4	0,5	47,56%

0,1	0,3	0,6	47,46%
0,1	0,1	0,8	45,18%

V. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Janar Saks**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   PEALKIRI

   mille juhendaja on Kaili Müürisep

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

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

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

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

Tartus, 14.05.2018