TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Edgar Miadzieles 131038IAPB

GUNNINGI FOG-I INDEKSI KOHALDATAVUS EESTI KEELELE

bakalaureusetöö

Juhendaja: Ermo Täks
filosoofiadoktor
Kaasjuhendaja: Ahti Lohk
filosoofiadoktor

Tallinn 2019

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: Edgar Miadzieles

[09.01.2019]

Annotatsioon

Antud töö eesmärk on uurida, kas Gunningi FOG-i indeks on sobilik meetod, et ennustada eestikeelse teksti loetavust. Töö käigus luuakse seoseid inglise- ning eesti keele vahel, et valemit saaks rakendada eestikeelsel tekstil. Loodud seosed realiseeritakse programmina, mis ootab sisendiks eestikeelset teksti ning arvutab selle põhjal loetavuse indeksi, vastavalt Gunningi FOG-i valemile. Tehtud valemi interpretatsiooni testimiseks luuakse eesti keele õpikute põhjal korpus, mis on jaotatud klassidesse alatest ühest kaheteistkümnendani. Samuti tehakse katseid keeleoskus tasemete A2–C2 järgi jaotatud tekstil ning uuritakse, kas Gunningi FOG-i indeks tõuseb vastavalt tasemele. Kolmas korpus koosneb seaduste tekstidest, millest on nii inglise- kui eestikeelne eksemplaar.

Õpikute korpuse tulemusena leiti, et antud valemi interpretatsioon ei vasta üks ühele, kuid kehtib väga tugev positiivne korrelatsioon. Keeleoskus tasemete korpuste keskmised Gunningi FOG-i indeksid tõusid vastavalt keeleoskus tasemele. Riigi Teataja seaduste korpuse põhjal tehtud testide tulemus näitas, et inglise- ja eestikeelsete seaduste tekstide loetavuse vahel kehtib tugev positiivne korrelatsioon ning mõlemas keeles jääb loetavus sarnasele tasemele. Rakendust saab kasutada eesti keele loetavuse ennustamiseks ühe tööriistana rohkemate tööriistade hulgas, et saada objektiivne ligikaudne loetavuse hinnang.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 29 leheküljel, 6 peatükki, 10 joonist, 5 tabelit.

Abstract

Applicability of Gunning FOG index to Estonian language

The goal of this thesis is to analyze if Gunning FOG index is applicable to Estonian language. Gunning FOG index demands certain constraints, specific to English language. These constraints were loosely defined so an interpretation of the formula was necessary. To meet the constraints of the chosen interpretation, relations between English and Estonian language had to be made. These relations were written as a computer program which takes Estonian text as input and displays calculated Gunning FOG index as output. The program also displays information about analyzed text such as average number of words, sentences and more.

To test the validity of this interpretation, a corpora of Estonian texts was necessary. A specific corpora had to be created that reflects the readability of the text. The chosen source for this corpora is the texts from Estonian textbooks sorted into formal education classes ranging from 1 to 12. Another corpora consisting of texts categorized by *CEFR* (*Common European Framework of Reference for Languages*) levels was used to determine if there is a relation between a higher level of language skill and Gunning FOG index. The third corpora consisted of Estonian law texts that had an English translation and Estonian text, to find out if there is any relation between language readabilities.

The results show that the expected result of Estonian textbooks corpora does not match the test results, but a very strong positive correlation exists. Applying the trend line equation to the result gives an approximate expected result. A similar finding came up using the latter corpora: the mean readability index rose when *CEFR* level increased. Third test also showed that a relation between English and Estonian text readability exists. The created application can be used as a tool in a set of tools, not to provide a definitive measure of readability of the text, but rather an approximation.

The thesis is in Estonian and contains 29 pages of text, 6 chapters, 10 figures, 5 tables.

Lühendite ja mõistete sõnastik

API	*Application programming interface*
CEFR	*Common European Framework of Reference for Languages* ehk Euroopa keeleoskus tasemete süsteem
CSS	*Cascading Style Sheets.*
HTML	*Hypertext Markup Language*.
HTTP	*Hypertext Transfer Protocol.* (hüperteksti edastuse protokoll)
JS	*JavaScript*
JSON	*JavaScript Object Notation*
Mallimootor	*Template engine.* Tehnoloogia, mis võimaldab staatilist ning dünaamilist sisu mugavalt kirjeldada ning kompileerida vastavasse formaati.
OS	*Operating System.* Operatsioonisüsteem (näiteks Windows)
Python	Interpreteeritud, kõrgtasemeline programmeerimiskeel.
Raamistik	*Framework*. Tarkvara raamistik pakub standardset viisi, et struktureerida koodi. Lisaks pakub raamistik lisafunktsionaalsust vastavalt tarkvara eesmärgile.
URL	*Uniform Resource Locator*. (Universaalne ressursilokaator ehk veebiaadress)

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Kirjutamisel kehtib põhimõte: mida läbimõeldum ja aeganõudvam on olnud teksti koostamine, seda lihtsam on selle lugemine ja tekstist arusaamine. Viimase saja aasta jooksul on loetavuse määramiseks ja teksti raskuse mõõtmiseks üritatud leida lahendusi, mis ennustavad teksti loetavust võimalikult objektiivselt. Antud töö eesmärgiks on uurida Gunningi FOG-i loetavuse valemi sobivust eesti keelele. Gunningi FOG-i indeks on üks esimestest laialdaselt kasutatavatest loetavuse valemitest.

Valem on algselt loodud inglise keele jaoks. Seetõttu tuleb valemis olevate parameetrite nõuete täitmiseks luua seoseid eesti- ning inglise keele vahel. Saadud seosed, tingimused ning reeglid realiseeritakse programmina, mis rakendab antud valemit. Eestikeelse teksti näol sisendi andmisel arvutab programm välja teksti loetavuse tulemuse, mida kutsutakse Gunningi FOG-i indeksiks ehk Gunningi uduindeksiks. Tarkvara näitab ka teisi tekstiga seotud parameetreid, näiteks sõnade arv, keerulisteks peetavate sõnade arv ning muid sõnaga seotud andmeid, mida on valemi jaoks kasutatud.

Võimalikult objektiivse tulemuse saamiseks valitakse korpused, mis esindavad seoseid hästi katsete tulemuste ja algandmete vahel. Koostatakse korpus eesti keele õpikutest, mis on jaotatud ametliku haridusklassi järgi. Teiseks korpuseks on A2–C2 keeleoskus tasemete järgi jaotatud materjalid. Kolmanda korpuse abil, mis koosneb Riigi Teataja seaduste eestikeelsest tekstist ja selle inglisekeelsest tõlkest, proovitakse leida keeltevahelisi Gunningi FOG-i indeksi seoseid.

Tehtud töö vastab küsimusele, kas Gunningi FOG-i valemit saab antud töös väljatoodud seadetes kasutada ka eestikeelsete tekstide jaoks ning kas tarkvara oleks sobilik lisatööriist ennustamaks eestikeelsete tekstide loetavust. Loetavuse ning teksti keerukuse omavahelise seose olemasolul saaks töö käigus loodud rakendust kasutada

näiteks korpuste loomiseks vastavale lugejaskonnale, ekstraheerides sobilikku teksti suurematest teksti kogumikest.

Töö struktuur on jaotatud järgnevatesse osadesse: teises peatükis tutvustatakse loetavuse üldist tausta ning loetavuse mõiste tausta seoses eesti keelega. Seejärel kirjeldatakse Gunningi FOG-i indeksit ning Gunningi FOG-i valemi interpretatsiooni seoses eesti keelega. Kolmandas peatükis on räägitud töö käigus valminud rakenduse ülesehitusest, mis Gunningi FOG-i valemi interpretatsiooni rakendab. Neljandas peatükis on kirjeldatud korpuseid ning katsete töö käiku, kus neid korpuseid kasutatakse. Viiendas peatükis on esitatud katsete tulemuste andmeid, tulemustega seotud arvutusi ja tulemuste analüüsi.

2 Teksti loetavuse ja ennustamise taust

Erinevates allikates, mis uurivad teksti keerukust, võib leida kontekste, kus loetavusest on räägitud kui teksti keerukusest, raskusest või jälgitavusest. Need mõisted on omavahel tugevalt seotud. Piir nende mõistete vahel hägustub veelgi, kui püüda tõlkida neid mõisteid inglise keelest eesti keelde. *Readability* on tõlgitud eesti keelde nii loetavuseks kui jälgitavuseks. Loetavuseks on aga võimalik ka lugeda inglisekeelset sõna *legibility* [21] . Samuti erineb loetavuse mõiste oma definitsioonilt ning funktsioonilt olenevalt allikast. Ühelt poolt peetakse oluliseks loetavuse mõõtmeks teksti ülesehitust arvestamata teksti sisulist mõtet [4] ning teiselt poolt lisatakse parameetrite sekka teksti sisulist tähendust omavaid aspekte [17] . Üldiselt peetakse loetavuseks teksti hoomamise raskuse näitajat [4] . Selliste erisuste tõttu on tarvis defineerida, mida loetavuse all mõeldakse antud kontekstis.

Teksti loetavuse täpseks mõõtmiseks on vaja korraldada katseid lugejatega. Mida loetavuse valemid üritavad teha, on ennustada teksti loetavust [13] . Antud töös on edaspidi loetavuse mõõtmise all mõeldud loetavuse ennustamist v.a juhul, kui loetavuse katseline (lugejate abiga) mõõtmine on välja toodud. Selles töös võtame loetavuse mõiste alla selle teksti osa, mis välistab teksti sisulise mõtte. See tähendab, et loetavuse mõiste ning sellega seotud näitaja ei sõltu tähenduslikust kontekstist. Välja võib arvata taolised tunnused, mida peetakse sisuliseks. Näiteks ei üritata hinnata erialapõhiste mõistete selgust või arusaadavust. Tegemist on näitajaga, mis peab olulisteks tunnusteks sõna- ja lauseehituse eripärasid. Küll aga on sellele reeglile antud töös seatud erandid, mida käsitletakse Gunningi FOG-i indeksi alampeatüki all.

Kergelt rakendatavaid loetavuse meetmeid on kasutusel palju, 1980. aastaks oli loetavuse valemeid ligikaudu 200 [4] ning tänapäeval võib see arv veelgi suurem olla. Enamus nendest on tehtud inglise keele jaoks. Tuntumatest võib välja tuua Flesch-Kincaid põhitaseme testi [3] , Edward Fry loetavuse meetodi [12] ning McLaughlini SMOG [19] valemi. Nende kõigi seos Gunningi FOG-i valemiga on tulemuse sõltuvus lause ning sõnade pikkusest. Näiteks toob McLaughlini SMOG sarnaselt Gunningi

FOG-i valemile välja „keerulised” sõnad, mis defineeritakse sarnaselt kolme või enama silbi olemasolul.

Loetavuse valemite automatiseerimiseks on nende funktsionaalsust rakendatud programmina mitmeid. Üheks tuntumatest programmidest võib välja tuua Microsoft Wordi tarkvara, kus on rakendatud Flesch-Kincaidi põhitaseme test ning Fleschi kerglugemise test. Microsofti andmetel kuvab programm mõne Euroopa keele sõnaarvestuse ja keskmiste näitajate statistikat, jättes välja loetavuse andmed [2] . Täpsemat infot Microsoft selle kohta ei paku. Eestikeelse tekstiga testides loetavuse kohta informatsiooni programm ei näidanud.

Eelmainitud loetavuse mõõtmise meetodeid ei ole vastu võetud kriitikata. Suureks puudujäägiks peetakse selliste valemite puhul neid faktoreid, millega valemid ei arvesta. Valemid ootavad, et tekst on „ausalt kirjutatud” ehk teksti ei manipuleerita valemi jaoks. Sisulist mõtet arvesse võtmata ei arvestata ka lugeja motivatsiooniga ning teksti lugemise eesmärkidega [1] . Samuti erinevad tihtipeale erinevate valemite tulemused, kus sama teksti loetavuse (näiteks haridustaseme) näitajaks võib olenevalt valemist olla 8.9 või 12.3. Selline kriitika keskendub valemi tulemusele kindla teksti puhul, jättes hindamata loetavuste järjepidevust erinevate tekstide puhul [4] . Kriitikast olenemata on paljud organisatsioonid loetavuse ennustamise valemeid kasutusele võtnud [11] . Arvestades valemite nõrkusi on ilmselge, miks neid kasulikuks peetakse.

2.1 Eestikeelse teksti loetavuse ennustamise ja määramise taust

Eestikeelsete tekstide raskuse väärtuse leidmiseks on rohkem kasutatud katselist meetodit, kus palutakse katsegrupil vastata erinevatele küsimustele. Sellist lähenemist kasutas näiteks Jaan Mikk. Nõukogude kool 74-11 artiklis “Teksti raskuse mõõtmine” võtab ta käsile teksti raskuse mõõtmisega seotud probleemid. Ta tutvustab valemit, mis sobib objektiivseks kompassiks õpikute jõukohasuse mõõtmisel. Ligikaudu 1500st õpilasest koosnevast grupis viidi läbi erinevaid katseid, mis hindavad raskuse erinevaid aspekte, et leida õiged parameetrid, mis iseloomustavad teksti raskust. Nii saadakse matemaatiline seos teksti tunnuste ja selle raskuse vahel [17] . Valemi rakendamiseks peab kasutaja (või programm) olema teadlik põimlause erinevatest osadest ning vajab

ka sõnade sisulist mõistmist. Seetõttu on valemit arvutiprogrammina keeruline realiseerida.

2011. aastal Helin Puksand ja Krista Kerge on teksti loetavust nimetanud ka raskusastmeks ning jälgitavuseks. Kirjeldades mõistet selliselt: „jälgitavuse tagavad keelelised konstruktsioonid ja sõnavara, mille sobiv kooslus teeb teksti sihipärase kasutamise lihtsaks". Loetavuse ennustamise meetodiks ning raskusastme mõõtmiseks võeti kasutusele Carl Hugo Björnssoni Lix valem. Valemi kasutamiseks tuleb sõnade arv jagada lausete arvuga ning summeerida saadud tulemus keeruliste sõnade (enam kui 6 tähest koosnevad sõnad) osatähtsusega kõigi sõnade hulgas. Tulemusteks saadakse vahemikud (20–25 tähistab väga kerget, 30–35 kerget jne), mis väljendavad teksti loetavust. Nende kirjeldusel on valemi interpreteerimine küllaltki lihtne, sest sellist arvutust on võimalik teha ka tekstitöötlusprogrammis, tuues välja Microsoft Word eraldi näitena [21] . Eeldatavasti on siinkohal mõeldud arvutuste tegemist kasutades statistilisi andmeid, mida Microsoft Word (või mõni muu tekstitöötlusprogramm) pakub. Microsoft Word ise Lixi loetavuse arvutamise funktsionaalsust ei paku.

2.2 Gunningi FOG-i indeks

Gunningi FOG-i valem on välja töötatud Robert Gunning Associates firma poolt ja on ennekõike mõeldud teksti loetavuse mõõdupuuks, mille rakendamine ei ole liiga tülikas. See pidi olema piisavalt efektiivne, et aidata kirjastustel parandada enda väljaannete loetavust. See on "tööriist, mitte reegel" [14] . Gunningi FOG-i indeksiks nimetatakse valemi rakendamisel saadud tulemust, mis on võrdne ametliku haridussüsteemi aastate arvuga alates algkoolist [25] .

Tabel 1. Gunningi FOG-i indeksi vastavus ametliku haridusastmega

...	
13	Bakalaureuse 1. astme tudeng
12	Keskkooli 3. astme õpilane
11	Keskkooli 2. astme õpilane
10	Keskkooli 1. astme õpilane

9	Põhikooli lõpuklassi õpilane
...	

Gunningi FOG-i valemi rakendamiseks on tarvis ligikaudu saja sõnaline lõik. Lauseid jäetakse täielikuks ning ei poolitata. Saja sõnaline nõue on seatud vaid selleks, et andmete maht oleks piisav tulemuse saamiseks. See tähendab, et sõnade hulk võib ületada saja piiri [16] .

$$0.4 * \left[\left(\frac{sõnade\,arv}{lausete\,arv} \right) + 100 * \left(\frac{keeruliste\,sõnade\,arv}{sõnade\,arv} \right) \right]$$

Sõnade arv jagatakse lausete arvuga, et saada keskmine sõnade arv lauses. Keeruliste sõnade arv jagatakse sõnade arvuga ning tulemus korrutatakse sajaga. Keerulisteks sõnadeks loetakse sõnu, mis koosnevad kolmest või enamast silbist ning ei sisalda levinud järelliiteid (näiteks *-ing, -ed*). Keeruliste sõnade hulka ei arvestata ka liitsõnu, mille osad ei ole üle 2 silbi ning pärisnimesid, mis on kirjeldatud W. F. Kwoleki poolt kui sõnad, mis algavad suure tähega ning Charles H. Vervalin kirjeldab neid kui *proper names* ehk pärisnimed. Leitud tulemuse summa korrutatakse 0.4ga [15] , [16] .

Esimeseks oluliseks vastuoluks või pingepunktiks on oluline mainida, et Gunningi FOG-i valem hägustab piiri tähenduseta (sisulist mõtet mitte omava) ja tähendusliku vahel pärisnimede näol. Samuti ootab valem, et keeruliste sõnade lugemisel ei loetaks silpide hulka levinud järelliiteid (näiteks *-ing*, *-es*, *-ed*). See jätab ruumi valemi interpretatsiooniks, mis võib põhjustada erinevusi valemi kasutamise osas. Seda on näha erinevates allikates, mis on valemi kasutusele võtnud [15] , [16] . Mõned valemi kasutajad on pidanud oluliseks *-ing* (näiteks *working*) ja *-es* (*glasses*) [16] sufikseid ning teised lisavad sufiksite hulka ka *-ee* (näiteks *employee*), *-able* (näiteks *playable*) jne [24] . Mõned lihtsustavad valemit ning ei kasuta sufiksite piirangut. Nad loevad silpide hulka kõik sõnas esinevad silbid [4] . Lihtsamaks ei tee ka asjaolu, et inglise keeles saab valida järelliiteid väga paljudest eksemplaridest [7] .

2.2.1 Valemi interpreteerimine seoses eesti keelega

Nii inglise- kui eesti keeles muudavad sufiksid ning sõna asetus lauses sõna tähendust selliselt, et täiuslikku paralleeli keelte vahel ei ole tihtipeale võimalik leida. See tähendab, et leides omavahelise seose eesti ning inglise sufiksi vahel võib see seos

kehtetu olla teises kontekstis või lauseehituses. Näiteks võttes aluseks inglisekeelne sufiks *-ing*, siis selle sõna tähendus võib muutuda kontekstis [18] . Võttes aluseks ühe interpretatsiooni saame esialgu leida seose *-ing* ja -mine sufiksi vahel. *Runn-ing* ja jooks-mine tundub esialgu mõistliku seosena, kuid kasutades sõnu teistsugustes lausetes seos kaob: *I am runn-ing*; ma jooks-en.

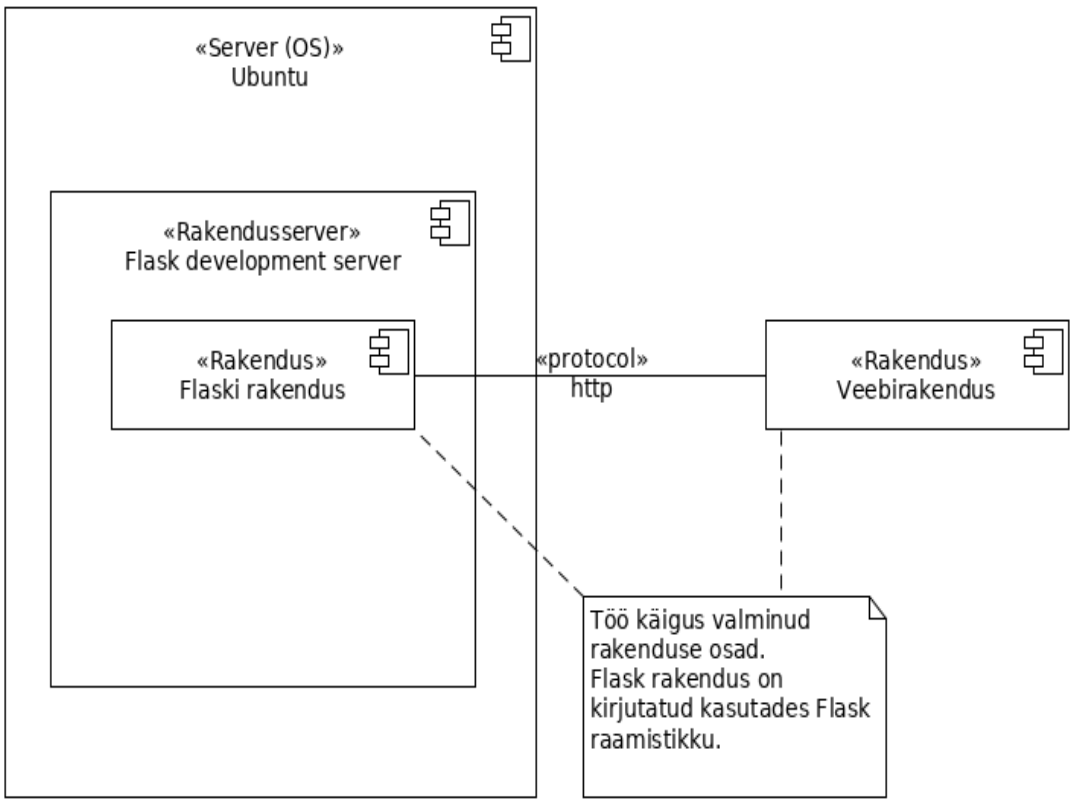
Antud töös interpreteeritakse reeglit selliselt, et järelliide ei tee sõna "keerulisemaks", lugedes ainult algvormi silpe, lähtudes Gunningi FOG-i valemist. Eesti keeles võib sõnatüvi vormimoodustuses muutuda, seetõttu tuleb vormimoodustuse kirjeldamisel aluseks võtta üks kindel tüvekuju, millest reeglite esitamisel lähtuda [8] . Samaselt Eesti keele käsiraamatuga saame kasutada reeglistikku, kus "käändsõnade algvormiks on ainsuse nimetav kääne (ema, auto, kirjanik, tantsijatar) ja pöördsõnade algvormiks tüvekuju, mis esineb ma-tegevusnime tunnuse ees (ela(ma), hakka(ma), kirjuta(ma), ilmne(ma)) [8] . Sellise reegli järgimine põhjendab ka tarkvaras kasutatav teek [20] . Sellest hiljemalt alampeatükis Estnltk teek. Tugevalt mõjutavad tulemust sellisel juhul sõnad, mille algvormid lõppevad näiteks -mine liitega või moodustatakse uus algvorm mitmest sõnast, näiteks seadus-lik-us-ta-mine. Kõik silbid loetakse sellisel juhul sõna hulka ning sõna peetakse keeruliseks.

Erisõnade ning -väljendite väljatoomisel on vaja kontekstipõhist arusaamist, mistõttu erilist tähendust omavad sõnad ja väljendused, mida ei saa lugeda pärisnimedeks kõigis kontekstides, jäetakse keeruliste sõnade piirangust välja. Pärisnime piirangu austamiseks saab kasutatud teegi [20] raames arvestada sõnu, mida teek tunneb ära pärisnimede- ning lühenditena. Teistsugune lähenemine nõuaks üldnimedena kasutusel olevate sõnade analüüsimist sisulise konteksti põhiselt.

Sõnade hulka ei loeta numbreid, mis on tähistatud sümbolitega 1–9. Nende silpe ei saa lugeda, mistõttu ei saa nende "keerulisust" antud valemi kontekstis ka arvestada. Samuti ei loeta sõnadeks kirjavahemärke. Kirjavahemärgid on koma, koolon, semikoolon jne. Mõned erandid on veel esitatud tarkvara piirangute all.

3 Rakendus

Gunningi FOG-i indeksi kiireks mõõtmiseks koostati rakendus, mis ootab kasutajalt sisendina eestikeelset teksti. Valemi rakendamiseks loetakse andmete hulka kõik programmile sisestatud sõnad. Sõnad analüüsitakse ning määratakse, millised loetakse sõnade hulka ning millised loetakse keerulisteks. Seejärel rakendab programm Gunningi FOG-i valemit ning näitab kasutajale tulemust. Tulemuse hulgas on sõnade hulka loetud sõnade arv, lausete arv, keeruliste sõnade arv ning Gunningi FOG-i indeks. Kasutaja saab vaadata ka sõnade informatsiooni üksikult. Näiteks saab kasutaja näha, miks on antud sõna loetud keeruliseks või mitte.



Joonis 1. Rakenduse komponentide loogiline vaade

Sisuliselt on tegemist veebirakendusega, mis on jaotatud kaheks osaks (Joonis 1). Üks on serveripoolne rakenduse osa ning teine kliendipoolne rakenduse osa, mida edaspidi nimetatakse kliendiks. Kliendi võib jaotada veel omakorda kaheks: kasutajaliideseks ning rakenduse funktsionaalsusi täitvaks osaks. Kuna kliendi mõlemad osad on omavahel tihedalt seotud ning põimitud, on mõistlik nende funktsionaalsust kirjeldada koos.

3.1 Server

Serveripoolne rakendus on kirjutatud kasutades Flask raamistikku [10] ning veebiserverina kasutatakse Flaski poolt pakutud arendusserverit. Flaski abil arendatud rakendust kutsutakse edaspidi Flaski rakenduseks. Flaski rakendust on arendatud Ubuntu keskkonnas.

3.1.1 Estnltk teek

Teksti analüüsimiseks kasutatakse estnltk teeki versioon 1.4.1 [20] , mis on kirjutatud Pythoni programmeerimiskeeles. Teek võimaldab sõnade morfoloogilist analüüsi, mille abil saab näha sõnade algvorme, käändeid ning sõnaliiki. Teeki kasutades tuleb ette, et mõne sisendiga jookseb programm kokku pakkumata head põhjendust: „*Segmentation fault (core dumped)*”. Tavaliselt juhtub see siis, kui sisendis esineb erilisi sümboleid või tihedalt kasutatavaid sümboleid teatud järjestuse puhul. Näiteks võivad tarkvara kokkujooksmist põhjustada sümbolid (eraldatud komadega): -, /, \ jne. Kuna need sümbolid valemi tulemust olulisel moel ei mõjuta, on parem, kui need asendada tühikuga, et tagada programmi toimimine.

Antud töös ei otsi estnltk ega ülejäänud programm eritähenduslikku sisu omavat sümbolit või sümbolite kogu. Sellise näitena võib välja tuua e-posti aadressi. Estnltk loeb sümbolite kogu nimetus@email.ee kui nimetus @ email.ee. Samuti võib näitena tuua URLi (universaalne ressursilokaator ehk veebiaadress), mis võib ebaproportsionaalselt mõjutada saadud tulemusi. Sellised erandid eemaldatakse käsitsi sõnade loetelust. Üldjuhul ei tohiks antud töös kasutatud korpuses sellise sisuga tekste ette tulla.

Gunningi FOG-i indeksi kirjelduse peatükis on mainitud sõnade tüve leidmise tarkvaralisi piiranguid. Eesti keeles saab sõnatüvi vormimoodustises muutuda. Näiteks küla+line, siga+la jne. Estnltk loeb moodustatud sõna uueks sõnatüveks. Selliseid sõnamoodustisi lahti harutada on keeruline. Näiteks sõna seaduslikustama on moodustatud mitmest järelliitest, kus algvormiks võib lugeda seadus ning sõna järelliideteks -lik-us-ta-ma. Moodustatud sõna on uue tähendusega ning estnltk loeb sõna algvormiks „seaduslikusta". Ka eraldi automaatikat on selliste moodustiste juurde keeruline ehitada. Näiteks järelliite -lik eemaldamisel sõnalt sisalik ei saavutata tähenduslikku sõna. Seetõttu on antud töös sõna algvormiks loetud teegi poolt tuvastatud variant. Neid variante saab sõnast olenevalt olla mitu, seega antakse kasutajale võimalus ise õige variandi pakututest valida läbi kasutajaliidese (3.2 Rakenduse klient).

3.1.2 Flask raamistik

Antud tarkvara loomiseks on Flask [10] raamistik sobilik, sest see on sarnaselt estnltk teegile kirjutatud Pythonis. Teegi ja Flask rakenduse integreerimine on seetõttu lihtsam. Flask tuleb kaasa oma arendusserveriga, mida kasutatakse selle töö raames ainukese serverirakendusena. Seda seetõttu, et selle ülesseadmine on kiire ja mugav. Päringute saatmine ja vastuvõtmine toimub HTTP-d (*Hypertext Transfer Protocol*) kasutades ning raamistik pakub päringute marsruutimise tuge.

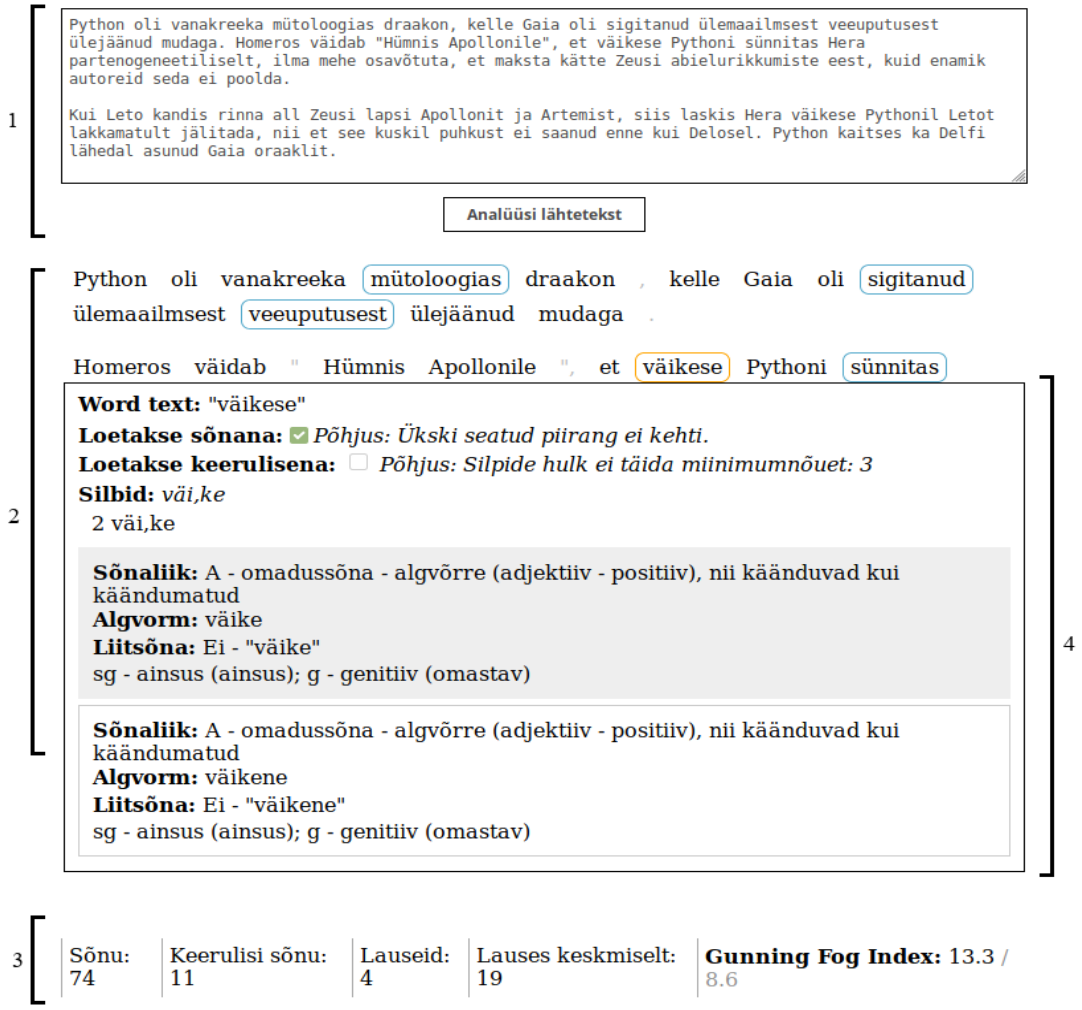
Kliendi poolt esialgse päringu saamisel URLile <url>/ (<url> tähistab domeeni nime näiteks „localhost", ning kokku „localhost/") saadab Flask rakendus kliendile vajalikud andmed, mille hulgas vajalikud HTML (*Hypertext Markup Language*) , JS (*JavaScript*), CSS (*Cascading Style Sheets*) failid, et kuvada kasutajaliides. Kasutajaliides on kirjutatud *ninja2* mallimootorit kasutades. See on eelseadistatud Flask raamistiku poolt ning annab võimaluse kasutada muutujaid HTML koodi vahel. Nii marsruutimist kui mallimootori kasutamine toimub Flaski abimeetodeid kasutades.

Kliendi poolt tehtud päringud /gunningfog URLile vastatakse JSON (*JavaScript Object Notation*) formaadis. Päringu vastuste hulgas on analüüsitud tekst jaotatud sõnadeks ning lauseteks. API (*Application programming interface*) täieliku kirjelduse leiab Lisa

1 – API kirjeldus juures. See on rakenduse ainuke API URL tee. Kõik ülejäänud vajalikud arvutused seotud andmetega tehakse kliendivaates.

3.2 Rakenduse klient

Kasutajaliides koosneb ühest vaatest, mille alla kuulub neli osa (Joonis 2). Esimeses osas on lahter teksti sisestamiseks ning lahtri all on nupp sisestatud andmete saatmiseks Flask rakendusele. Pärast nupu vajutust saadab klient sisestatud andmed Flask rakendusele /gunningfog URLile ning Flask rakendus vastab päringule JSON formaadis vastusega. Seejärel analüüsib klient andmeid, et määrata millised sõnad loetakse sõnadeks ja millised keerulisteks sõnadeks.



Joonis 2. Kasutajaliidese vaade jagatud osadeks

Kui klient on analüüsiga valmis, kuvatakse kasutajale kõik sõnad teises osas, mis asub eelkirjeldatud nupu all. Sõnad, mida ei loeta valemisse kaasa, kuvatakse halli värviga. Sõnad, mis on määratud keerulisteks, ääristatakse sinise joonega. Sõnad, mida ei loeta keeruliseks, kuid võivad olla keerulised teistsuguse sõna seadistusega (täpsem kirjeldus antud funktsionaalsusest ülejärgmises lõigus) ääristatakse oranžilt.

Gunningi FOG-i indeksi kuvamine kuulub kolmanda osa alla, mis asub kasutajaliidese alumises servas. Kasutaja näeb tulemustega seotud andmeid, milleks on: sõnade arv, keeruliste sõnade arv, lausete arv, sõnade keskmine arv lauses ning Gunningi FOG-i indeks. Antud töö katse tulemuste põhjal on lisatud kolmandale osale ka eesti keele jaoks kohandatud Gunningi FOG-i indeksi tulemus. See indeks on leitud esialgse indeksi tulemuse kasutamisel trendijoone funktsioonis (Joonis 3) ning on näidatud pärast kaldkriipsu hallikana.

Tarkvara ei ole täiesti täpne, et määrata igas olukorras, millise sõnaliigi või -vormiga on tegu. Seetõttu, klikkides valitud sõnale, tuleb esile vaate neljas osa ning kasutajale näidatakse sõnaga seotud andmeid. Kui mõni analüüsi osa tundub vale, saab kasutaja käsitsi määrata õige seade. Käsitsi on võimalik määrata, kas sõna tuleks lugeda keeruliste sõnade hulka või kas sõna ise lugeda arvestavate sõnade hulka. Pärast seadete muutmist uuendab klient osas 2 olevaid sõnu ning osas 3 kirjeldatud tulemusi. Kui serveripoolne rakendus tuvastas mitu võimalikku varianti sõnaliigi või -vormiga seoses, esitatakse variandid samuti selles osas ning kasutaja saab määrata nende hulgast õige variandi. Valiku tegemisel uuendab klient kasutajaliideses olevad andmed eelnevalt kirjeldatud protseduurile. Selliseks sõnaks võib osutuda näiteks „väikese", kus rakendus tuvastab kaks erinevat algvormi: väikene ning väike. Antud töös eelistatakse väiksema arvu silpidega algvormi.

Eelmises lõigus kirjeldatud paranduse tegemine loetakse sisulise konteksti väliseks, kuigi sõnade sõltuvuslik kontekst jääb. See on tähtis seetõttu, et sõnade omavahelise sõltuvusliku struktuuri tõttu võib sõna algvorm erineda ning tarkvara poolt esimeseks jäänud variandi silpide arv ei pruugi olla õige. Estnltk teegi dokumentatsioonis on välja toodud näiteks sõna "mõeldud" [9] . Olenevalt lause struktuurist võib olla sõna omadussõna või tegusõna. Kasutaja saab teha vajaliku paranduse.

4 Korpused ning katsed nende põhjal

Gunningi FOG-i valemi testimiseks on vaja leida sobivad sisendandmed ehk tekst. Loetavuse või teksti keerukuse põhjal jaotatud korpuseid on keeruline leida. Enamik eestikeelseid korpuseid on jaotatud teistsuguste kategooriate järgi [5] . Näiteks on kategooriaks ilukirjandus, ajalehed, seadused jne. Nendest on antud töös vähe kasu, kui tuleb leida seos loetavuse ning Gunningi FOG-i indeksi vahel. Eelnevalt kirjeldatud probleemide tõttu valiti ning valmistati ette korpused vastavalt testi iseloomule.

4.1 Korpus 1: tekstid eesti keele õpikutest

Kuna Gunningi FOG-i indeks on väärtus, mis vastab üheselt ametliku haridustaseme astmega, on mõistlik valida korpus, mis on loodud vastavale lugejaskonnale. Sellist korpust on keeruline leida väljaspool ametlikku haridussüsteemi, seetõttu on esimese korpuse andmeteks valitud klasside järgi jaotatud õpikud. Korpuseks valiti eesti keele õpikud, mis algavad esimesest klassist ning lõppevad 12. klassiga. Oletatakse, et nendes õpikutes on kõige vähem erisõnu ning õpiku loetavused vastavad teatud haridusastme tasemele. Kirjandust, mis eelneb esimesele klassile ei lisatud, sest isegi üks sõna lauses annab Gunningi FOG-i indeksi minimaalseks tulemuseks 0.4 ja see ei kajasta haridusastet. Pärast 12. klassi on kirjandust haridusastme järgi keerulisem jaotada, seetõttu on taolised andmed korpusest välja jäetud.

Kuna kirjandus ei ole digitaliseeritud, oleks pidanud kõik õpikud digitaliseerima. Arvestades digitaliseerimise ajamahtu, piirduti antud korpuse tegemisel ligikaudu tuhande sõnaga igast õpikust. Tuhandesõnalise teksti koostamisel valiti osad lõigud õpiku algusest, õpiku keskelt ning õpiku lõpust. Eelistati tekste, mis olid pikemad ning jaotatud suurematesse osadesse. Väiksemad ülesande kirjeldused said lisatud korpusesse vaid siis, kui sõnade kogus oli alla tuhande.

Digitaliseerimiseks tehti raamatu lehekülgedest pildid, mis hiljem Tesseract tarkvara abiga tekstiks muudeti [23] , kasutades eesti keelel treenitud andmeid [22] . Tesseract ei

olnud iga kord täpne, mistõttu pidi veenduma tulemuse korrektsuses ning vajadusel seda parandama[1]. Korpuse failid on jaotatud klasside kaupa kaustadesse. Kasutatud korpuse faili nimetus on corpus.txt. Failis on märgitud tekstilõikude juurde leheküljenumbrid.

Antud korpuse puhul on oluline välja tuua, et teksti loetavust võib mõjutada see, et teksti on koostanud erinevad autorid. Mõne õpiku puhul on autoriteks mitu inimest ning mõne õpiku puhul on autoriks üksainus isik. Samuti võib loetavust mõjutada asjaolu, et tekstid valiti pisteliselt ning eelistati pikemaid lõike. Sellegipoolest eeldatakse, et antud korpus on objektiivne ning sobiv antud töö eesmärgi saavutamiseks.

4.1.1 Töö käik

Loetavuse indeksi leidmiseks kasutatakse töö autori koostatud tarkvara. Sisendandmete ettevalmistamiseks puhastatakse korpuse fail lisainfost: eemaldatakse leheküljenumbrid ning andmetevaheliseks eralduseks kasutatud sidekriipsud. Saadud tekst kopeeritakse tarkvara esimese vaate esimese osa lahtrisse ning tarkvara analüüsib teksti. Vaadatakse üle, kas mõni sõna on ekslikult peetud mittekeerukaks. Vajadusel tehakse parandus.

Eelnevalt kirjeldatule sarnaselt, analüüsitakse testitavad andmed ning saadud tulemused salvestatakse tabelisse. Salvestatud tulemuste põhjal moodustatakse hajuvusdiagramm ning pealtnäha lineaarse seose olemasolul kantakse graafikule trendijoon ja arvutatakse korrelatsioon. Kõikide andmete esitamiste ning saadud tulemuste põhjal arvutuste tegemisel kasutatakse Libreoffice Calc tarkvara. Hajuvusdiagramm koostatakse kasutades XY (*scatter*) graafiku funktsionaalsust. Korrelatsioon arvutatakse kasutades CORREL funktsiooni.

4.2 Korpus 2: EVKK tekstid jaotatud keeleoskus tasemete järgi

Teiseks katse korpuseks on valitud EVKK (Eesti vahekeele korpus), mis on kirjeldatud kui „eesti keele kui riigikeele ja võõrkeele õppijate kirjalike tekstide kogu” [6] . Korpuse andmete leidmiseks kasutatakse EVKK kodulehe päringusüsteemi, mis võimaldab filtreerida tulemusi keele valdamise taseme (*CEFR*'i) A1 kuni C2 järgi.

1 Korpuse failid leiab siit: https://bios.ee/loputoo/eestiopikutekorpus/

Antud töös tehakse katseid andmetega, mis algavad alates A2 tasemest, sest A1 taseme materjale pole piisavalt. Korpuse koostamise ajal oli A1 taseme tekstilõike vaid üks.

Kuna korpus esindab mitmeid autoreid ning nende loomingut erinevatel teemadel, on ootuspärane, et see kajastub ka tulemustes vahemikena. Samuti *CEFR* tasemed ei jaotu täpselt haridustasemete järgi ning neid ei saa võrrelda otseselt Gunningi FOG-i indeksiga, kuid tulemused võiksid kajastada loetavuse keerukuse tõusuna kui keeleoskuse tase on kõrgem. Mõned andmed on EVKK kogus koos ülesande kirjeldusega, seetõttu kopeeritakse testimise jaoks vaid kirjatüki sisu/keha[1].

4.2.1 Töö käik

EVKK veebilehel otsitakse A2–C2 keeleoskustasemete järgi sissekandeid järjest. Iga taseme materjalidest võetakse järjest 68 eksemplari, jättes teksti keha tervikuks, eemaldades vaid pealkirjad ning ülesande kirjeldused, kui need peaks tekstis leiduma. Kui tekst ei koosne vähemalt sajast sõnast võetakse otsingutulemustest järgmine kanne. 68 teksti piirang tekkis sellest, et A2 algandmete hulgast olid sobilikud 68 tekstilõiku, kõigist ülejäänud keeleoskus taseme tekstidest piirduti samuti 68-ga.

Iga tekst sisestatakse töö käigus valminud programmi ning tulemus salvestatakse tabelisse. Tulemused ümardatakse täisarvuni, sest nii on tulemusi lihtsam esitada graafikuna ning tulemuste täpsus on piisav. Arvutatakse oskusastme tulemuste keskmised ning mediaanid. Tulemustest koostatakse tabelid ja joonised.

4.3 Korpus 3: inglise- ja eestikeelsed tekstid Riigi Teataja seadustest

Kolmas korpus on moodustatud tekstist, millel on olemas nii inglise kui eestikeelne tõlge. Tõlge peaks olema suunatud samale lugejaskonnale, olema sama eesmärgiga ning üheti mõistetav. Seetõttu valiti kolmandat korpust esitama Riigi Teatajas olevad eestikeelsete seaduste tekstid ning vastavate seaduste inglisekeelsed tõlked. See tähendab, et samast seadusest peab olemas olema inglisekeelne ning eestikeelne variant. Eeldatakse, et seaduse tõlge ühest keelest teise säilitab oma sisulist kavatsust ning on

1 Korpuse failid leiab veebiaadressilt https://bios.ee/loputoo/keeleoskustasemetekorpus/

üheselt mõistetav teatavale lugejaskonnale. Uurimise alla läheb seos inglisekeelse teksti loetavuse ning eestikeelse teksti loetavuse vahel.

Seadused on valitud pisteliselt. Tekst ei tohi olla liiga lühike. Kui seaduse tekstis on sõnu üle tuhande mõlemas keeles, loetakse tekst sobivaks. Eelistatakse pikemaid tekstilõike ning jäetakse välja pealkirjad, muutmismärked jne. Iga paragrahvi, lõigu või muu teksti osa kohta on sellele vastav tõlge teises keeles korpuses olemas. Korpuse tekstid on valitud kahekümnest erinevast seadusest mõlemas keeles. Korpuse tekstid on jaotatud seaduste järgi[1].

4.3.1 Töö käik

Eestikeelse seaduse teksti Gunningi FOG-i indeksi leidmiseks kasutatakse töö käigus valminud rakendust. Inglisekeelse seaduse teksti Gunningi FOG-i indeksi leidmiseks kasutatakse veebirakendust, mis asub veebiaadressil http://gunning-fog-index.com/. Testides inglisekeelse teksti jaoks mõeldud veebirakendust leiti, et veebirakenduse interpretatsioon ei arvesta liitsõnade piiranguga - rakendus loeb kõik liitsõna silbid kokku. Antud töö käigus valminud rakenduse algseadetes on piiranguga arvestatud, mistõttu tehakse 2 katset.

Esimese katse jaoks jäetakse antud töö käigus valminud rakenduses liitsõnade piirang alles. Liitsõnal arvestatakse silpe iga liitsõna osa kohta eraldi. Teise katse jaoks võetakse piirang seadetest maha ning programm arvestab liitsõna osade silpide summat. Inglisekeelse teksti jaoks kasutatud rakenduses saadud tulemused jäetakse mõlemal korral samaks.

Kõik tulemused salvestatakse tabelisse, leitakse korrelatsioon ning inglisekeelse ja eestikeelse tulemuste vahe absoluutväärtus. Tulemused sorteeritakse vahe väärtuse järgi tõusvas järjekorras väiksemast suuremani. Moodustatakse joonised, kuhu kantakse inglise- ning eestikeelsed tulemused ettenähtud järjekorras, et illustreerida keelte tulemuste vahe väärtust.

1 Korpus asub veebiaadressil: https://bios.ee/loputoo/seadustekorpus/

5 Tulemuste analüüs

Iga korpuse kohta tehtud katsete tulemused on esitatud eraldi alampeatükis. Kõiki tulemusi on esitatud ka graafiliselt, et anda saadud tulemustest parem ülevaade. Kõikide tulemuste salvestamiseks, graafikute tegemiseks ning arvutuste sooritamiseks on kasutatud Libreoffice Calc funktsionaalsusi. Rohkete andmetega katse tulemusi on esitatud histogrammina.

5.1 Korpus 1 tulemused ja analüüs

Saadud teksti analüüsi tulemused on esitatud 2 veeruga ning 12 reaga tabelis Tabel 2. Tabeli esimene veerg iseloomustab Gunningi FOG-i indeksi eeldatavat tulemust, ehk haridusastet. Kui õpik oli mõeldud 1. klassi jaoks, on tabelis sümbol 1, kui 2. klassi jaoks siis sümbol 2 jne. Teine veerg on katsetulemuste jaoks.

Tabel 2. Korpus 1 põhjal tehtud katse tulemuste tabel

Eeldatav tulemus	**Tulemus**
1	5.8
2	6.1
3	9.1
4	8.5
5	10
6	9.5
7	14.7
8	14.3
9	14.5
10	13.3
11	14.4

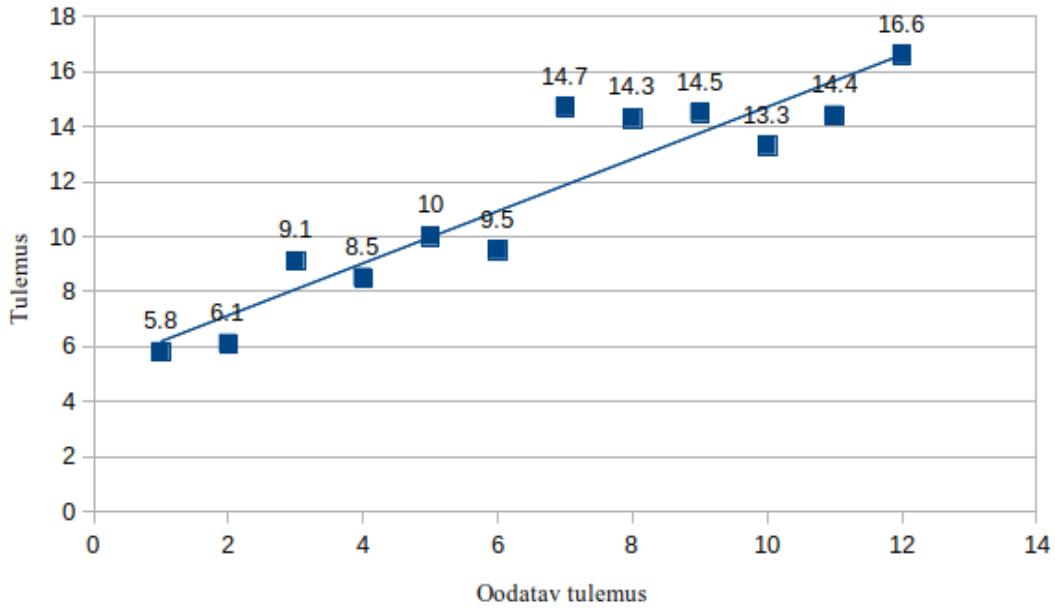
Eeldatav tulemus	Tulemus
12	16.6

Hajuvusdiagrammil Joonis 3 on esitatud andmed järgnevalt: horisontaaltelje väärtus on oodatav tulemus ning vertikaaltelje väärtuseks on saadud tulemus. Hajuvusdiagrammile on lisatud trendijoon, mis on genereeritud Libreoffice Calc abil.

Trendijoone funktsiooniks on:

$$f(x) = 0.948 * x + 5.24.$$

Kasutades CORREL funktsiooni saadi korrelatsioonikordajaks: **0.93**, mis vastab väga tugevale positiivsele suhteseosele.



Joonis 3. Koprus 1 põhjal tehtud katse tulemuste hajuvusdiagramm ning trendijoon

5.1.1 Analüüs

Tulemused on madalamate klassi õpikute puhul ootavast tulemusest kõrgemad ning kõrgema klassi poole liikudes vahe oodatava ning katse tulemuse vahel väheneb, kuid on siiski oodatust kõrgem. Küll aga on näha, et oodatava- ning saadud tulemuse vahel on tugev positiivne korrelatsioon. Trendijoont jälgides on näha, et oodatava tulemuse tõustes tõuseb üldiselt ka katse tulemus. See tähendab, et Gunningi FOG-i indeksi

tulemuse sidumine eesti keele haridustasemetega ei ole ühene, kuid trendijoone funktsiooni rakendades saab ligikaudse õige tulemuse.

Haridusastmete 1–6 ning 10–12 tulemused on trendijoonele väga lähedal ning lineaarset trendi kõigutavad mõnevõrra astmed 7–9, asudes trendijoonest kõrgemal. Selle erisuse üheks põhjustajaks võivad olla korpuse koostamisel valitud õpikud. Õpikud klassidele 7–9 on koostanud Priit Ratassepp ning teiste õpikute koostajate hulgas teda ei esinenud. Kahjuks peab selle põhjuse erisuse välja selgitamiseks kasutama erinevat korpust.

Puuduvad andmed ka kõrgemate loetavustasemete (üle 12. klassi materjali) testimiseks. Kuna ühene seos oodatud ning saadud tulemuse vahel puudub, on võimalik järeldusi teha vaid trendijoont mööda paremale liikudes. Ligikaudu 104. haridusastme juures ristuvad saadud- ning oodatava tulemuse väärtused, kus oodatav tulemus on edaspidi suurem eeldatavast saadud tulemusest. Valemi iseloomust tingituna on see loomulik, et kasutades keerulisemaid sõnu ning pikemaid lauseid tõuseb ka Gunningi FOG-i indeks ning seda teeb ka trendijoon. Kõrgemate haridusastmete ning Gunningi FOG-i indeksi tulemuse vahel seose tekitamiseks on vaja rohkem katseid vastava korpusega.

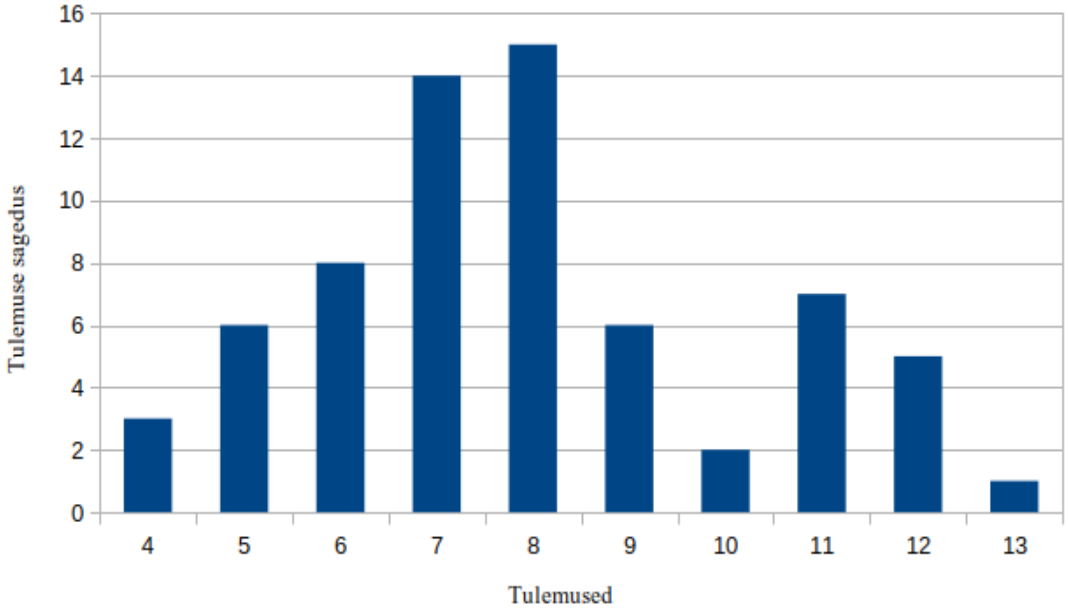
5.2 Korpus 2 tulemused ja analüüs

Keeleoskus tasemete korpuse katsete tulemuste keskmised ning mediaanid on esitatud tabelis Tabel 3 ning sissekanded on reastatud keeleoskus A2 - C2 tasemete järgi alustades kõige kergemast.

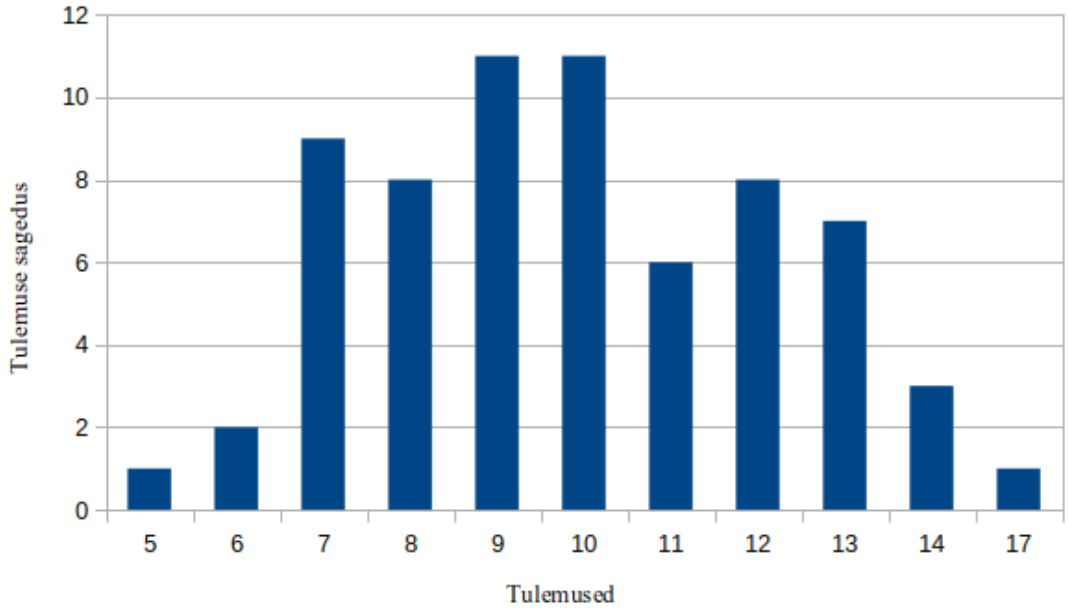
Tabel 3. Korpus 2 põhjal tehtud katse tulemuste keskmiste tabel

	Keskmine	**Mediaan**
A2	8	8
B1	10	10
B2	11	11
C1	13	13
C2	15	15

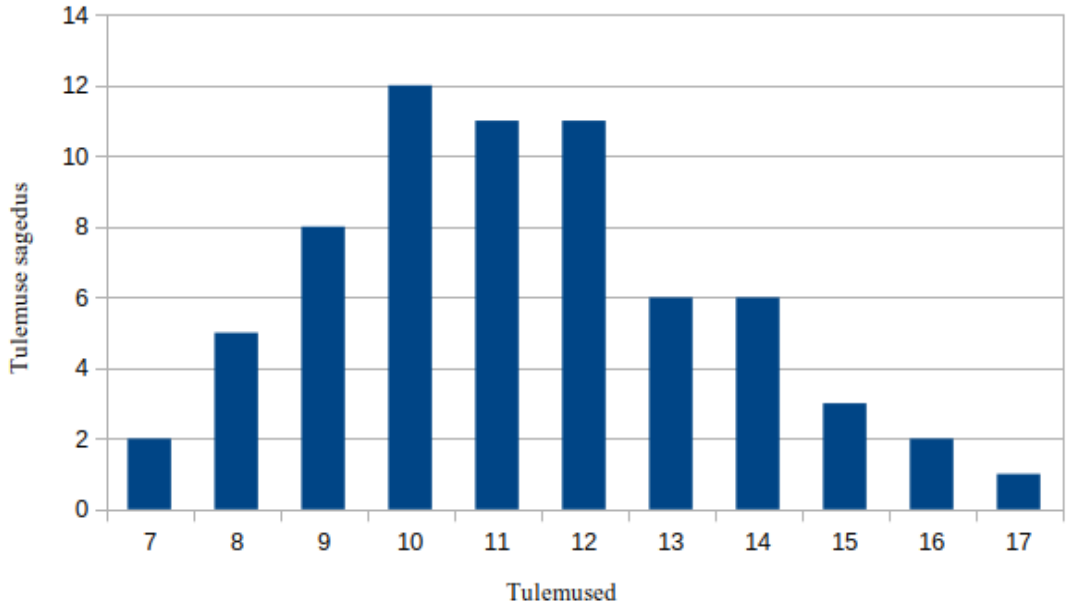
Joonistel Joonis 4, Joonis 5, Joonis 6, Joonis 7 ja Joonis 8 on esitatud keeleoskus tasemete korpuse põhjal tehtud katsete tulemuste sagedused, kus x teljel on tulemuste väärtused ning y teljel, mitu korda antud tulemus esines.



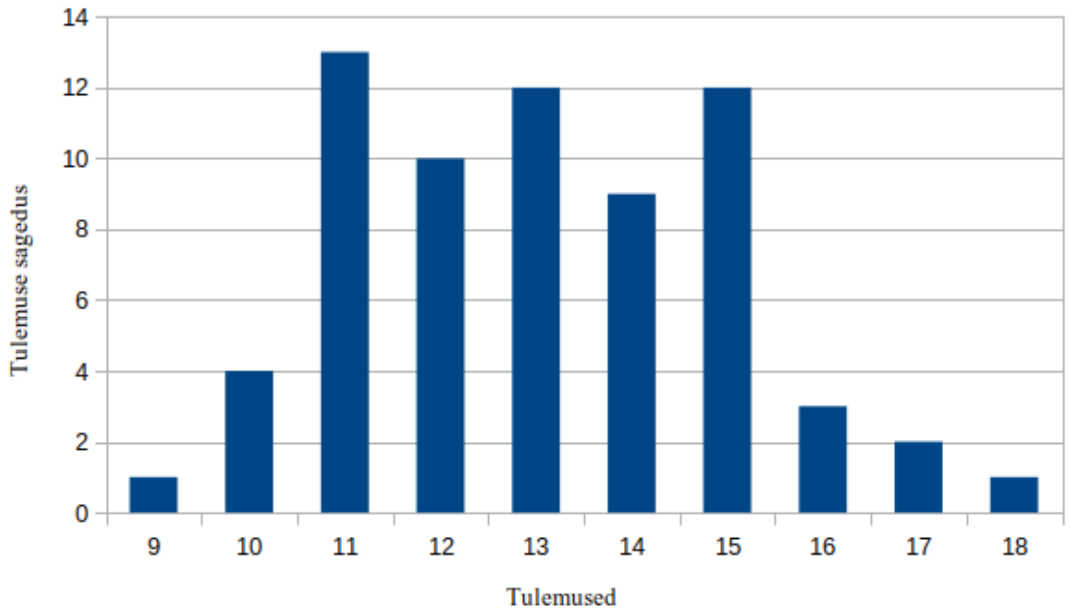
Joonis 4. Keeleoskus taseme A2 korpuse põhjal tehtud katse tulemuste sagedused



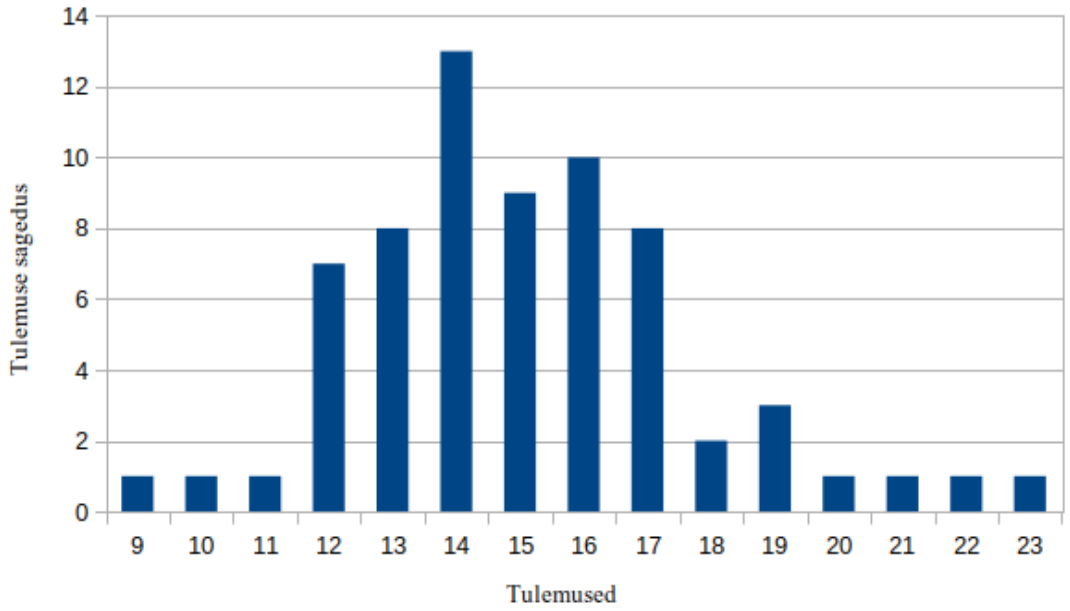
Joonis 5. Keeleoskus taseme B1 korpuse põhjal tehtud katse tulemuste sagedused



Joonis 6. Keeleoskus taseme B2 korpuse põhjal tehtud katse tulemuste sagedused



Joonis 7. Keeleoskus taseme C1 korpuse põhjal tehtud katse tulemuste sagedused



Joonis 8. Keeleoskus taseme C2 korpuse põhjal tehtud katse tulemuste sagedused

5.2.1 Analüüs

Keeleoskus tasemete korpuse keskmiste tulemuste tabeli Tabel 3 põhjal on näha, et keeleoskus taseme tõustes tõuseb ka vastava taseme Gunningi FOG-i indeksi keskväärtus ning mediaan. Ka selle katse korral, sarnaselt eesti keele õpikute korpuse põhjal tehtud katsetele, algavad Gunningi FOG-i indeksi väärtused madalamate keeleoskustasemete juures kõrgemast oodatavast väärtusest. A1 keeleoskus taseme kohta puuduvad andmed ning antud järeldust tuleks eraldi kinnitada lisakatsetega vastavat korpust kasutades.

Joonistel Joonis 4, Joonis 5, Joonis 6, Joonis 7 ja Joonis 8 on näha, et tulemuste vahemikud on olnud suhteliselt suured iga keeleoskus taseme kohta. Seda võis põhjustada korpuses kasutatud tekstide autorite keeleoskuse erinevus. Tulemuste suure vahemiku tõttu on loetavuse taset keeleoskustasemetega keeruline siduda. Sellegipoolest on näha, et tulemused moodustavad või meenutavad normaaljaotust, kus suurem osa sagedamatest tulemustest koonduvad keskväärtuse juurde, ning tulemuste keskmist väärtust võib pidada loetavuse indeksi piirkonnaks.

5.3 Korpus 3 tulemused ja analüüs

Riigi Teataja seaduste korpuse põhjal tehtud testi tulemuste tabelid Tabel 4 ja Tabel 5 koosnevad neljast tulbast, millest esimene on järjekorra number, et tulemused oleksid joonistel selgemini eristatavamad. Tabeli teises tulbas on inglisekeelse seaduse teksti tulemused ning neile vastavad eestikeelsed tulemused on kolmandas tulbas. Tulemused on reastatud vahe absoluutväärtuse järgi alustades väikseimast.

Tabel 4 esindab esimese katse tulemusi, kus eestikeelsed tulemused on leitud liitsõnade piirangut arvestades.

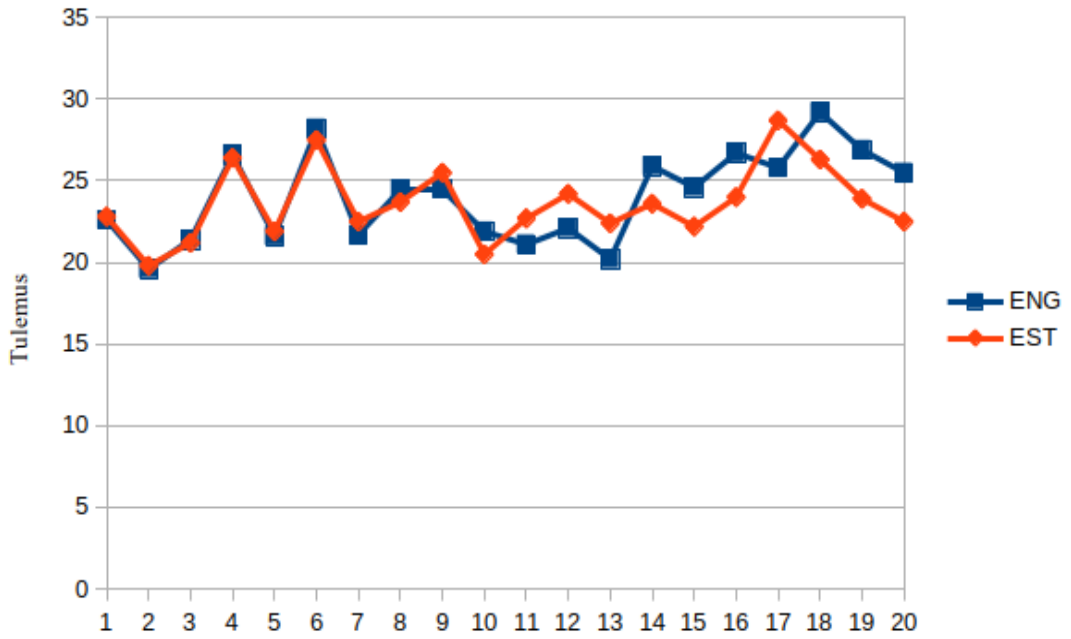
Tabel 4. Korpus 3 põhjal tehtud esimese katse tulemuste tabel

Järjekorra number	**Inglisekeelse teksti tulemus**	**Eestikeelse teksti tulemus**	**Tulemuste vahe**
1	22.6	22.8	0.2
2	19.6	19.8	0.2
3	21.4	21.2	0.2
4	26.6	26.4	0.2
5	21.6	21.9	0.3
6	28.2	27.5	0.7
7	21.7	22.5	0.8
8	24.5	23.7	0.8
9	24.5	25.5	1.0
10	21.9	20.5	1.4
11	21.1	22.7	1.6
12	22.1	24.2	2.1
13	20.2	22.4	2.2
14	25.9	23.6	2.3
15	24.6	22.2	2.4
16	26.7	24.0	2.7
17	25.8	28.7	2.9

Järjekorra number	Inglisekeelse teksti tulemus	Eestikeelse teksti tulemus	Tulemuste vahe
18	29.2	26.3	2.9
19	26.9	23.9	3.0
20	25.5	22.5	3.0

Inglisekeelsete ja eestikeelsete tekstide põhjal tehtud katsete tulemuste korrelatsioon on leitud kasutades Libreoffice Calc funktsiooni CORREL. Esimese katse tulemuste korrelatsioonikordajaks on **0.74**, mis vastab tugevale suhteseosele.

Joonis 9 on tabeli Tabel 4 andmete põhjal tehtud, kus y-teljel olevad väärtused esindavad katse tulemusi. Väärtused x-teljel on tabelis vastavad järjekorranumbrid. Eestikeelsete tekstide tulemuste andmed on tähistatud punase värviga ning inglisekeelsete tekstide tulemused sinisega.



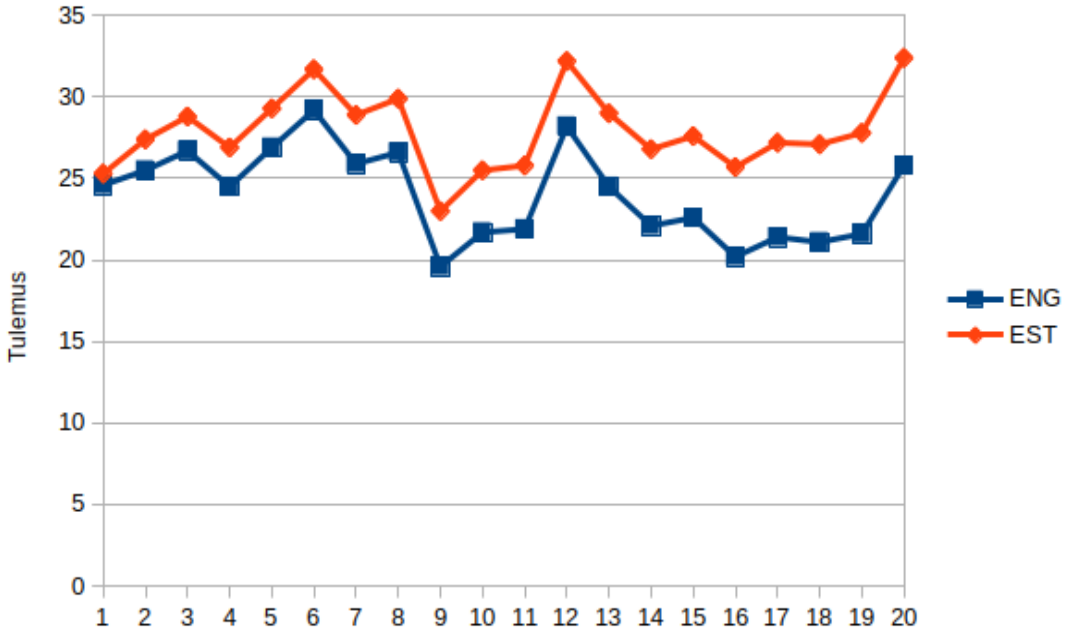
Joonis 9. Korpus 3 põhjal tehtud esimese katse tulemuste joonis

Tabelisse Tabel 5 on sisse kantud Riigi Teataja seaduste korpuse põhjal tehtud teise katse tulemused. Teise katse eestikeelsete tekstide Gunningi FOG-i indeksi leidmiseks ei ole kasutatud liitsõnade piirangut. Teise katse tulemuste korrelatsioonikordajaks on **0.81**, mis vastab tugevale suhteseosele.

Tabel 5. Korpus 3 põhjal tehtud teise katse tulemuste tabel

Järjekorranumber	**Inglisekeelse teksti tulemus**	**Eestikeelse teksti tulemus**	**Tulemuste vahe**
1	24.6	25.3	0.7
2	25.5	27.4	1.9
3	26.7	28.8	2.1
4	24.5	26.9	2.4
5	26.9	29.3	2.4
6	29.2	31.7	2.5
7	25.9	28.9	3.0
8	26.6	29.9	3.3
9	19.6	23	3.4
10	21.7	25.5	3.8
11	21.9	25.8	3.9
12	28.2	32.2	4
13	24.5	29	4.5
14	22.1	26.8	4.7
15	22.6	27.6	5
16	20.2	25.7	5.5
17	21.4	27.2	5.8
18	21.1	27.1	6
19	21.6	27.8	6.2
20	25.8	32.4	6.6

Joonis 10 on tabeli Tabel 5 andmete põhjal tehtud, kus y-teljel olevad väärtused esindavad katse tulemusi ning x-telje väärtused on tabelis vastavad järjekorranumbrid. Eestikeelsete tekstide tulemuste andmed on tähistatud punase värviga ning inglisekeelsete tekstide tulemused sinisega.



Joonis 10. Riigi Teataja seaduste korpuse põhjal tehtud teise katse tulemuste joonis

5.3.1 Analüüs

Jättes liitsõnade piirang alles, on tabelist Tabel 4 näha, et inglise- ning eestikeelsete seaduste tekstide põhjal tehtud tulemuste vahe jäänud väiksemaks kui seda on teise katse tulemuste Tabel 5 vahed, kus liitsõnade piirang on eemaldatud. Sellegipoolest on teise katse puhul näha, et tulemused inglise- ja eestikeelsete tekstide loetavuse osas on järjepidevamad, kus eestikeelse teksti loetavuse tulemus jääb alati inglisekeelse teksti loetavusest kõrgemaks. Seda toetab ka korrelatsioonikordaja suurem väärtus.

Selle korpuse põhjal tehtud teise katse korral, sarnaselt eelmiste korpuste põhjal tehtud katsetele, jääb eestikeelse teksti loetavus kõrgemaks võrreldavast. Keelte erinevusest olenemata näitavad testide tulemused tugevat positiivset korrelatsiooni loetavuste osas ning vahemikud tulemuste vahel ei jää suureks. Väga kõrge inglisekeelse teksti loetavuse indeksi puhul on ka eestikeelse teksti Gunningi FOG-i indeks kõrgem. Täpsemate tulemuste saavutamiseks oleks tarvis valmis teha siiski tööriist, mis mõõdaks inglisekeelse teksti Gunningi FOG-i indeksi loetavust sarnaselt töös valminud rakendusele.

5.4 Kokkuvõtlik analüüs

Kõigil kolmel korpusel tehtud katsete puhul (v.a Riigi Teataja seaduste korpuse esimese katse puhul) on näha seaduspära, et eestikeelsete tekstide põhjal tehtud katsete tulemused jäävad võrreldavatest ligikaudu 5 indeksit kõrgemale. Kõigi katsete puhul, eestikeelse teksti keerukuse tõustes tõuseb ka Gunningi FOG-i indeksi väärtus. Sellest järeldatakse, et Gunningi FOG-i indeksi põhjal on võimalik ennustada eestikeelse teksti umbkaudne loetavuse väärtus.

Sellegipoolest ei saa väita, et Gunningi FOG-i indeks on piisav ning väga täpne meetod kuidas hinnata eestikeelsete tekstide loetavust. Seda takistavad paljud erinevused inglise- ning eesti keele vahel. Kuid antud korpustega tehtud testi tulemused näitavad, et tulemuste vahel esineb tugev positiivne korrelatsioon ning eesti keele õpikute korpuse katse tulemustest ilmunud trendijoone funktsiooni rakendamine rakendusse aitab jõuda lähemale eestikeelse teksti objektiivsele loetavuse hindamisele. Mistõttu on valem küll ebasobilik andmaks lõplikku hinnangut eesti keele loetavuse osas, kuid piisavalt sobilik, et kasutada valemit ning rakendust kui tööriista teiste meetodite hulgas, et anda teksti loetavusele objektiivne hinnang.

6 Kokkuvõte

Antud töö eesmärk oli uurida Gunningi FOG-i indeksi loetavuse valemi sobivust eestikeelse teksti loetavuse hindamiseks. Esmalt loodi seosed inglise- ning eesti keele vahel, lähtudes valemi iseloomust. Katsete läbiviimiseks valmistati rakendus, mis implementeerib Gunningi FOG-i valemit ning eelnevalt loodud eesti- ning inglise keele vahelisi seoseid. Töö käigus valminud rakendust kasutati, et sooritada testid kolme korpuse põhjal. Eesti keele õpikute põhjal valminud korpuse põhjal uuriti seost haridusastme järgi jaotatud eestikeelse teksti loetavuse ning oodatava Gunningi FOG-i indeksi vahel. Teiseks sooritati katsed keeleoskus tasemete järgi jaotatud korpuse tekstidel, et saada ülevaade sellest, kas Gunningi FOG-i indeksit mõjutab keeleoskuse tase. Kolmanda katse korpuse, Riigi Teataja seaduste inglise- ning eestikeelsete tekstide põhjal uuriti inglise- ning eestikeelse teksti Gunningi FOG-i indeksi vahelisi seoseid.

Võttes aluseks eesti keele õpikute korpuse põhjal tehtud katsete tulemused saab öelda, et õpikutekstide loetavuse ja oodatava Gunningi FOG-i indeksi vahel esineb väga tugev positiivne korrelatsioon. Eesti keele õpikute tekstide loetavus jääb oodatavast tulemusest järjepidevalt ligikaudu viie indeksi võrra suuremaks.

Keeleoskus tasemete korpuse katsete tulemused näitasid sarnaseid tulemusi, kus keeleoskus taseme tõustes, tõusis ka vastava teksti Gunningi FOG-i indeksi tulemuste keskväärtus. Keskväärtused jäid oodatud tulemustest mõnevõrra kõrgemaks. Kuigi tulemuste vahemikud olid küllaltki suured, koondus suurem osa tulemustest keskväärtuse ümber meenutades normaaljaotust. Tulemuste suurimaks mõjutajaks võis olla see, et tekstid olid kirjutatud erinevate autorite poolt.

Eesti-inglise seadustekstide tõlkekorpusi kasutades jõuti tulemusteni, mis näitasid, et eesti- ning inglisekeelsete tekstide Gunningi FOG-i indeksi vahel kehtib tugev positiivne korrelatsioon. Inglisekeelse teksti kõrge tulemuse puhul jäi eestikeelse teksti loetavuse tulemus samuti kõrgeks.

Loetavuse ning keeleoskuse vahelise seose olemasolul, saab töö käigus valminud rakendust kasutada näiteks vajaliku keerukusega tekstide leidmiseks teatava lugejaskonna jaoks. Selliseid tekste saaks ekstrahheerida juba olemasolevatest korpustest, mis ei pruugi olla jaotatud keerukuse järgi. Loomulikult ei saa rakendust kasutada eestikeelse teksti loetavuse määramisel ainukese näitajana. Ainuüksi antud korpuste ning tulemuste põhjal ei saa teha põhjapanevaid järeldusi ning üldistusi. Sellegipoolest võib pidada töö käigus loodud rakendust kasulikuks ühe tööriistana teiste hulgas, et ennustada eestikeelsete tekstide loetavust.

Kasutatud kirjandus

[1] Bruce, B., Rubin, A., Starr, K. Why readability formulas fail. – *IEEE transactions on professional communication*, 1981, PC-24, 1, 53. [Online] https://ieeexplore.ieee.org/abstract/document/6447826 (10.12.2018)

[2] Dokumendi loetavuse testimine. [WWW] https://support.office.com/et-ee/article/dokumendi-loetavuse-testimine-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2#__toc342546558 (09.10.2018)

[3] DuBay, W.H. The Classic Readability Studies. – *ERIC Institute of Education Sciences*, 2009, 96-98 [Online] https://eric.ed.gov/?id=ED506404 (10.12.2018)

[4] DuBay, William H. The Principles of Readability. – *ERIC Institute of Education Sciences,* 2004, 2, 3, 24. 56 [Online] https://eric.ed.gov/?id=ED490073 (15.10.2018)

[5] Eesti keele koondkorpus. [WWW] https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (15.11.2018)

[6] Eesti vahekeele korpus. – *Tallinna Ülikool.* [Online] http://evkk.tlu.ee/ (12.12.2018)

[7] English words by suffix. [WWW] https://en.wiktionary.org/wiki/Category:English_words_by_suffix (22.10.2018)

[8] Erelt, M., Erelt, T., Ross, K. Eesti keele käsiraamat. – Tallinn: Eesti Keele Sihtasutus, 2007 [Online] https://www.eki.ee/books/ekk09/index.php?p=3&p1=2 (15.10.2018)

[9] Estnltk – Word analysis [WWW] https://estnltk.github.io/estnltk/1.4.1/tutorials/text.html#word-analysis (20.09.2018)

[10] Flask. [WWW] http://flask.pocoo.org/ (20.09.2018)

[11] Florio, T., Ley, P. The use of readability formulas in health care. – *Psychology, Health & Medicine*, 2007, 7. [Online] https://www.tandfonline.com/doi/abs/10.1080/13548509608400003 (27.12.2018)

[12] Fry, E. A Readability Formula That Saves Time. – *Journal of Reading*, 1968, 11, 7, 577. [Online] https://www.jstor.org/stable/pdf/40013635 (10.12.2018)

[13] George R. Klare. Assessing Readability. – *Reading Research Quarterly,* 1974-1975, 10, 1, 62. [Online] https://www.jstor.org/stable/747086 (09.11.2018)

[14] Gunning, R. The Fog After Twenty Years. – *Journal of Business Communication*, 1969, 4-5, 10 [Online] https://journals.sagepub.com/doi/abs/10.1177/002194366900600202 (15.10.2018)

[15] Hervalin, Charles H. Checked your fog index lately?. – *IEEE Transactions on Professional Communication*, PC-23, 2, 87. [Online] https://ieeexplore.ieee.org/document/6501857

[16] Kwolek, William F. A Readability Survey of Technical and Popular Literature – Journalism & Mass Communication Quaterly, 1973, 255-256. [Online] https://journals.sagepub.com/doi/pdf/10.1177/107769907305000206 (09.11.2018)

[17] Mikk, J. Teksti raskuse mõõtmine. – *Nõukogude kool*, 1974, 11, 934-938. [Online] https://www.digar.ee/arhiiv/nlib-digar:235682 (12.10.2018)

[18] Mirriam Webster -*ing* [WWW] https://www.merriam-webster.com/dictionary/-ing (12.11.2018)

[19] Moncada, Fatima M., Pabico, Jaderick P. On Gobbledygook and Mood of the Philippine Senate: An Exploratory Study on the Readability and Sentiment of Selected Philippine Senators' Microposts – *Asia Pacific Journal on Education, Arts, and Sciences*, 2015, 2. [Online] https://arxiv.org/abs/1508.01321 (10.12.2018)

[20] Open source tools for Estonian natural language processing. [WWW] https://github.com/estnltk/estnltk (20.09.2018)

[21] Puksand, H., Kerge, K. Õpiteksti analüüs kirjaoskuse omandamise kontekstis. – *Emakeele Seltsi aastaraamat 57*, 2011, 175. [Online] http://www.kirj.ee/20807/?c_tpl=1064 (11.12.2018)

[22] Tesseract est traineddata. [WWW] https://github.com/tesseract-ocr/tessdata/raw/4.00/est.traineddata (22.11.2018)

[23] Tesseract Open Source OCR Engine (main repository). [WWW] https://github.com/tesseract-ocr/tesseract (22.11.2018)

[24] Topic 8: Indexes and Ratings [WWW] http://jcsites.juniata.edu/faculty/roth/QM/topic08a.htm (20.11.2018)

[25] Yasseri, T., Kornai, A., Kertész, J. A practical approach to language complexity: a Wikipedia case study. – *Cornell University*, 2012, 3-4. [Online] https://arxiv.org/abs/1204.2765 (30.11.2018)

Lisa 1 – API kirjeldus

Teksti analüüsi info saamiseks tuleb teha JSON formaadis API päring URLile: <url>/gunningfog

Parameetrid: {fog_index_input: „<tekst>”}

Päringu vastuse näide: [{"tokens": {"text": "maaparandushoiu", "paragraphs": [{"end": 15, "start": 0}], "words": [{"text": "maaparandushoiu", "syllables": ["maa", "pa", "ran", "dus", "hoi", "u"], "analysis": [{"partofspeech": "S", "root": "maa_parandus_hoid", "root_tokens": ["maa", "parandus", "hoid"], "syllables": [["maa"], ["pa", "ran", "dus"], ["hoid"]], "form": "sg g", "lemma": "maaparandushoid", "clitic": "", "ending": "0"}], "end": 15, "start": 0}], "sentences": [{"end": 15, "start": 0}]}, "sentence": "maaparandushoiu"}]