TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Martin Peedosk

Eesti keele digitaalsete ressursside ja tehnoloogiate rakendamine teksti lihtsustamise programmis

Bakalaureusetöö (9 EAP)

Juhendajad: Sven Aller
Kadri Vare

Tartu 2017

Eesti keele digitaalsete ressursside ja tehnoloogiate rakendamine teksti lihtsustamise programmis

Lühikokkuvõte:

Käesoleva bakalaureusetöö eesmärk oli uurida teksti lihtsustamise meetodeid ning luua veebipõhine rakendus, mis lihtsustaks eestikeelset teksti. Rakenduse loomiseks kasutati keeleressursse, nagu Eesti Wordnet, *word2vec'*i mudel, sagedusloend, võõrsõnade leksikon ja põhisõnavara sõnastik ning nendega leitakse sõnade keerukus ning sobivus teksti.

Võtmesõnad: lihtsustamine, eesti keel, semantilised suhted, *wordnet*, loomuliku keele töötlus

CERCS: P175, Informaatika, süsteemiteooria

Applying Estonian Digital Resources and Technologies in a Text Simplification Program

Abstract:

The purpose of this Bachelor's thesis was to research text simplification methods and to create a web-based application to simplify Estonian texts. The web application uses language resources such as the Estonian Wordnet, word2vec model, frequency dictionary, foreign word dictionary and basic vocabulary dictionary, which are used to identify word complexity and suitability to the text.

Keywords: simplification, Estonian, semantic relations, wordnet, natural language processing

CERCS: P175, Informatics, systems theory

Sisukord

Sissejuhatus

Keel on üks olulisemaid tunnusjooni rahvuse ja kultuuri seas. Statistikaameti andmetel rääkis 2016. aastal Eestis eesti keelt emakeelena 883 707 inimest, mis moodustas 68% kõigist Eesti elanikest [1]. Teistest rahvustest pärit elanike seas suhtlevad eesti keeles vabalt ainult 37% eesti keeles (s.t saavad aru, räägivad ja kirjutavad), 48% on passiivse eesti keele oskusega (s.t saavad aru ja veidi räägivad) ja 15% ei oska keelt üldse [2].

Eestikeelne kirjaoskus ja funktsionaalne lugemine on Eestis elava inimese jaoks oluline nii igapäevaelus kui ka tööturul osalemisel. Sageli võivad tekstid olla lugejale liiga rasked oma sõnavara poolest. Põhjuseks on kas vähene keeleoskus, vanus või puue. Eesti keele rääkijate arvu väiksuse tõttu on oluline, et leiduks ressursse, mille abil keelt paremini omandada. Siinse bakalaureusetöö eesmärk on uurida olemasolevaid lahendusi inglise keeles ning luua veebirakendus, mis asendab keerulisemad sõnad kas oma ülemmõiste või sünonüümidega. Kasutatavate keeleressursside hulgas on morfoloogilised tööriistad, Eesti Wordnet, *word2vec'*i mudel ja sõnaloendid. Rakendust saab kasutada, et lihtsustada eestikeelseid tekste.

Töö on jagatud kolmeks osaks. Esimene osa annab ülevaate erinevatest lihtsustamise meetoditest ning olemasolevatest rakendustest. Teises osas tuuakse välja töös kasutatavad keeleressursid. Kolmandas osas kirjeldatakse rakenduse ülesehitust ning analüüsitakse tulemusi.

1. Ülevaade teksti lihtsustamisest

Teksti lihtsustamine on tegevus, mille tulemusena saadakse algsest tekstist kergemini loetav ning arusaadavam tekst, jättes ajal samal tähenduse samaks. Horacio Saggion jt [4] on kirjutanud, et lihtsustatud tekste iseloomustavad selge ja otsene stiil, väiksem sõnavara ning kergem lauseehitus (nt vähem kõrvallauseid). Selle tõttu on teksti lihtsustamisel palju rakendusi:

- arusaadavuse parandamine madala kirjaoskustasemega inimestele;
- uudiste ja artiklite arvukuse parandamine inimestele, kellel on lugemispuuded, nagu afaasia, düsleksia jt;
- tekstide sisu selgemaks tegemine inimestele, kes omandavad keelt võõrkeelena;
- tehniliste tekstide (nt kasutusjuhendid, patendid, meditsiinilised tekstid) teisendamine lugejatele, kes ei ole tuttavad kasutatavate terminitega.

Selles peatükis kirjeldatakse teksti lihtsustamise teoreetilist poolt ja tuuakse välja mõned konkreetsed rakendused, mille abil on võimalik ingliskeelset teksti lihtsustada.

1.1 Lihtsustatud teksti vajadus

Eesti elanikest räägib eesti keelt emakeelena 68%, teistest rahvustest Eesti elanike seas suudab eesti keeles lugeda 58% [1, 2]. Seega on võõrkeele kõnelejate osakaal riigis suur ning paljud neist ei oska eesti keeles lugeda. On leitud, et võõrkeelsetest tekstidest arusaamine paraneb õppijate seas, kui teksti on lihtsustatud [3]. Lisaks võõrkeele õppijatele tekitavad keerulised tekstid muret ka lugemisraskustega inimestele. Rahvusvahelise Düsleksia Assotsiatsiooni andmetel võib üle maailma erinevate düsleksia astmete all kannatavate inimeste osakaal olla 15–20%, mis tähendab, et neile on harvaesinevate sõnade mõistmine raskendatud [4]. Lisaks düsleksiale on tekstidest arusaamine raskendatud nii kurtidele, kelle jaoks on probleemne süntaktiliselt keerukate lausete mõistmine, kui ka afaasia põdejatele, kellel on raskusi pikkade lausete ning harvaesinevate sõnade arusaamisega [3]. Keerukas sõnavara on probleemiks isegi tudengite seas [5].

1.2 Erinevad lihtsustamise meetodid

Teksti lihtsustamist on võimalik teostada mitmel erineval viisil, samuti saab erinevaid meetodeid omavahel koos kasutada. Olenevalt meetodist on võimalik seda rakendada käsitsi, automaatselt või mõlemal viisil. Käsitsi lihtsustamine on inimese poolt tehtav teksti lihtsustamine, kuid see on aeganõudev ja kallis tegevus, mistõttu leidub selliseid tekste

vähe [6]. Automaatne lihtsustamine on loomuliku keele (s.t keele, mida kasutatakse emakeelena) töötluse osa, kus soovitud tulemuse saavutamiseks kasutatakse arvuteid. Arvutite abiga saab meetodite rakendamist automatiseerida, mis suurendab lihtsustatud tekstide arvukust [7]. Siinne peatükk annab ülevaate enim levinud teksti lihtsustamise meetoditest.

1.2.1 Leksikaalne lihtsustamine

Leksikaalse ehk sõnavaralise lihtsustamise eesmärk on asendada keerulised sõnad sünonüümidega. H. Saggion jt [8] järgi jaguneb leksikaalne lihtsustamine üldjoontes kaheks etapiks. Esmalt leitakse asendatava sõna jaoks sünonüümide hulk. Selle hulga leidmiseks kasutatakse mitmeid lahendusi, mille hulgas on erinevad sõnastikud, sh ka *wordnet*, millega saab leida erinevate sõnatähendusvaheliste suhetega teisi sõnu. Seejärel asendatakse sõna kontekstist lähtuvalt kergemini mõistetava sünonüümiga. Õige tähenduse määramiseks kasutatakse erinevaid strateegiaid, nagu sõnatähenduse ühestamine ning sobivuse selgitamiseks arvutatakse sõnakeerukus [8]. Nii asendatava sõna leidmiseks kui ka uue sõna sobivuse kontrolliks võib kasutada erinevaid strateegiaid: sõna sagedus, sõna pikkus, sõna mitmetähenduslikkus või silpide arv sõnas [9].

Näide, kus on sõna hindamiseks kasutatud sõna sagedust: *Poiss lippas koolist koju.* → *Poiss lippama koolist koju.* → *Poiss {liduma, jooksma, lippama, pühkima} koolist koju.* → *Poiss jooksis koolist koju.*

1.2.2 Süntaktiline lihtsustamine

Süntaktilise ehk lausestruktuuri lihtsustamise eesmärk on moodustada keerulise ehitusega lausetest lihtsamad laused. Alljärgnev materjal on refereeritud J. D. Belderi ja M.-F. Moensi artiklist [10]. Eesmärgi saavutamiseks kasutatakse reeglipõhiseid süsteeme, et tuvastada ja asendada lauseüksusi (s.o lause või selle osa). Tüüpilised lihtsustatavad lauseüksused on lisand, passiiv- ja põimlaused, kuid leidub ka teisi. Täpsemalt on viidatud artiklis välja toodud mõned lihtsustamise reeglid:

- Lisand: kui lisand on tuvastatud, siis saab lause jagada kaheks, kus osalausest moodustatakse alus ning luuakse eraldiseisev lause.
  Näide*: Poiss kui sportlane on heas vormis.* → *Poiss on sportlane. Poiss on heas vormis.*

- Põimlause: sarnaselt lisandi lihtsustamisega jagatakse lause mitmeks osaks. Asesõna asendatakse sõnaga, millele see osutab ning moodustatakse uus lause.
  Näide*: Poiss, kes istus toolil, oli väsinud.* → *Poiss istus toolil. Poiss oli väsinud.*

Kui lause jaguneb lihtsustamise tulemusena mitmeks osaks, siis rakendatakse lihtsustamist ka tekkinud osadele, mis võimaldab lihtsustada lauseid, kus esineb rohkem kui kaks lauseüksust [6,10].

Näide: *Poiss, kes istus toolil, mis asus puu kõrval, oli väsinud.* → *Poiss istus toolil, mis asus puu kõrval. Poiss oli väsinud.* → *Poiss istus toolil. Tool asus puu kõrval. Poiss oli väsinud.*

1.2.3 Sisukokkuvõtte tegemine kui teksti lihtsustamine

Sisukokkuvõtte eesmärk on luua algtekstist lühendatud versioon, kuhu on alles jäetud vaid oluline informatsioon. Kaili Müürisep on oma artiklis [11] kirjutanud, et üldiselt on sisukokkuvõtte tegemiseks kaks lähenemist: väljavõte (ingl *extract*) ja ülevaade (ingl *abstract*). Väljavõtte juures valitakse sisendtekstist välja olulised laused ning kopeeritakse muutmata kujul väljundisse. Ülevaate puhul tekitatakse sisendteksti põhjal uued laused, mis ei pruugi algses tekstis leiduda. K. S. Jones [12] on kirjeldanud, et tüüpiline sisukokkuvõtmine toimub kolmes etapis. Esimesena analüüsitakse sisendit, mille käigus leitakse lausete ning sõnade piirid. Teises etapis rakendatakse algoritmi, mis teisendab teksti lihtsustatud kujule. Olenevalt lähenemisest, leitakse selles etapis lausete kaal, mis põhineb lause asukohal tekstis, võtmesõnade olemasolul, sõnade ja fraaside sagedusel, või leitakse masinõppe meetoditel sõnadevahelised seosed, mille põhjal otsustakse, kas sõna või lause peaks kuuluma tulemusse [11]. Kolmandas etapis teisendatakse tekst tagasi loomulikku keelde [12].

1.2.4 Masintõlkega teksti lihtsustamine

Masintõlge on protsess, mille käigus arvuti tõlgib treeningkorpuse põhjal teksti ühest loomulikust keelest teise (nt inglise keelest saksa keelde). Masintõlkemeetoditeks on näiteks statistiline masintõlge, kus leitakse statistiliselt kõige sobivam vaste, ja tehisnärvivõrkudel põhinev masintõlge, kus leitakse vaste komponentide omavaheliste seoste põhjal [13, 14]. Muude kasutusvaldkondade hulgas on masintõlget võimalik rakendada ka teksti lihtsustamiseks. Võimalikuks viisiks on kujutada algset teksti ja lihtsustatud teksti kui erinevaid keeli, millele rakendada masintõlget [13]. Sisuliselt tähendab see seda, et masintõlke programmile õpetatakse, millised on algsed tekstid ning millised nendele

vastavad lihtsustatud tekstid. Selle teadmise põhjal üritab masintõlge suvalises tekstis teha vajalikke asendusi. Selline lihtsustamine on võimalik, kui leidub tekstikorpus ning sellele põhjal muudetud lihtsustatud korpus. Selline lähenemine on rakendatav näiteks inglise keele jaoks, kus Wikipedia artiklitel on sageli olemas ka lihtsustatud keelega artikkel ning mille põhjal on võimalik masintõlget treenida. Masintõlke puuduseks ongi asjaolu, et paljudes keeltes ei leidu vajaminevate andmetega korpuseid.

1.3 Lihtsustamise rakendused

Tekstide lihtsustamist on uuritud paljudes keeltes: inglise, prantsuse, itaalia, jaapani, hispaania jt [3]. Lisaks on keeletehnoloogia arenguga tekkinud erinevaid lihtsustamis-teenuseid pakkuvaid veebirakendusi. Kuigi rakendusi leidub ka teiste keelte jaoks, on siinses töös piirdutud ainult ingliskeelsete näidetega. Järgnevates peatükkides on katsetatud kolme erinevat teksti lihtsustamise rakendust ning kõikides on ühtluse mõttes kasutatud sama sisendteksti, et tulemusi omavahel võrrelda. Rakendused põhinevad leksikaalse lihtsustamise meetodil ehk lihtsustusi tehakse ainult sõnavaraliselt. Sisendtekstiks on valitud esimene lõik Mary Shelley teosest „Frankenstein“. Lõik sai valitud põhjusel, et tekstis leidub autori hinnangul piisavalt keerulisi sõnu, mida lihtsustada. Näidistekst on järgmine [15]:

„You will rejoice to hear that no disaster has accompanied the commencement of an enterprise which you have regarded with such evil forebodings. I arrived here yesterday, and my first task is to assure my dear sister of my welfare and increasing confidence in the success of my undertaking.“

1.3.1 Tekstilihtsustamise rakendus Rewordify.com

Veebirakendus Rewordify.com [16] pakub võimalust lihtsustada kogu veebilehekülge või ainult sisestatud teksti. Rewordify.com kirjutab, et veebirakendus leiab sõnadele kontekstipõhiselt õiged asendused tipptasemel loomuliku keele töötluse abil. Asendamisel uurib rakendus nii erinevaid sõnu kui ka terveid fraase. Pärast lihtsustamist säilivad otsekõne ja sõnavormid. Rakendus on suuteline lihtsustama enam kui 58 000 sõna ja fraasi [16]. Näidisteksti lihtsustamise tulemusena on saadud tekst:

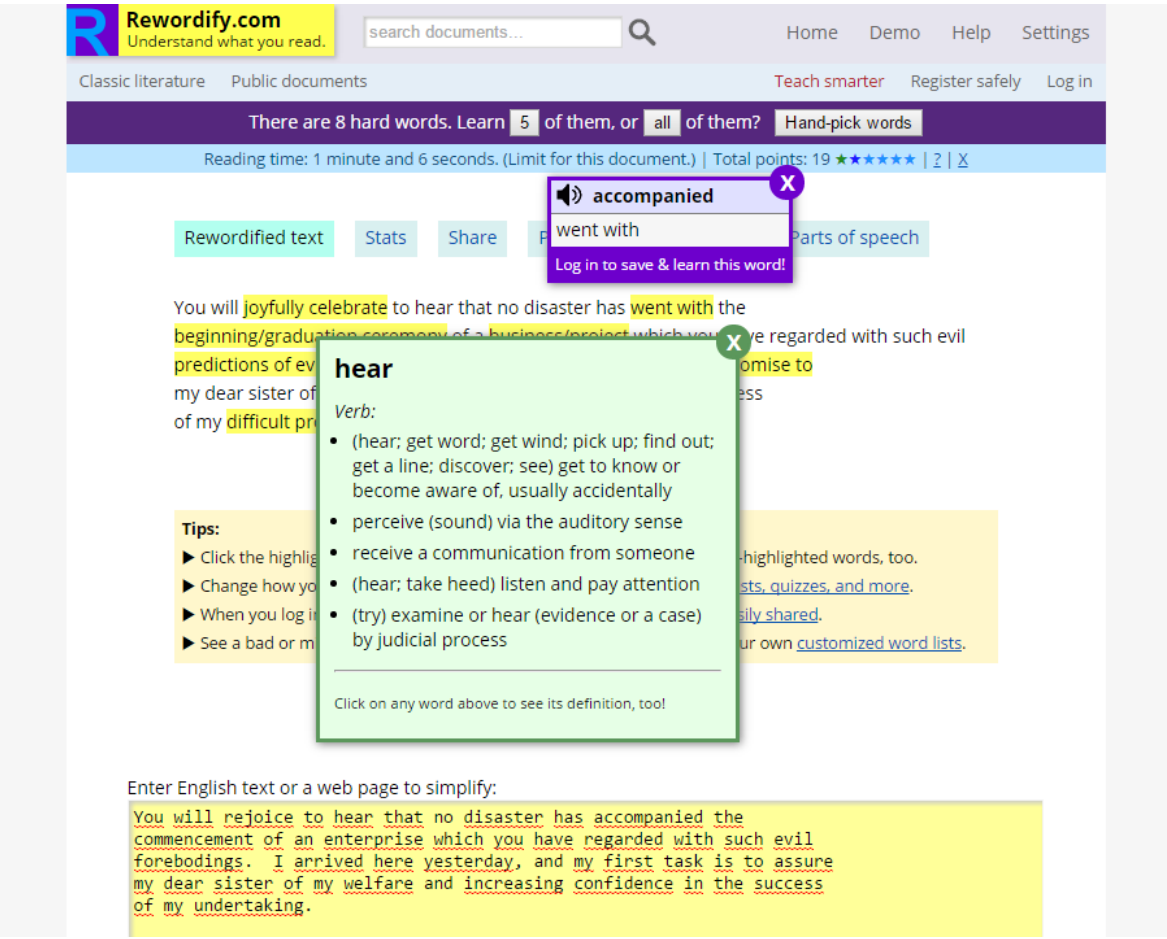
„You will joyfully celebrate to hear that no disaster has went with the beginning/graduation ceremony of a business/project which you have regarded with such evil predictions of evil.

I arrived here yesterday, and my first job is to promise to my dear sister of my welfare and increasing confidence in the success of my difficult project.“

Tekstis on asendatud järgmised sõnad:

- *rejoice → joyfully celebrate;*
- *accompanied → went with;*
- *commencement → beginning/graduation ceremony;*
- *enterprise → business/project;*
- *forebodings → predictions of evil;*
- *task → job;*
- *assure → promise to;*
- *undertaking → difficult project.*

Kuigi igale sõnale on kontekstis korrektne vaste leitud ning asendused on autori hinnangul lihtsamad, siis kohati ei ole tekst ladus, nt ’*such evil predictions of evil*’.



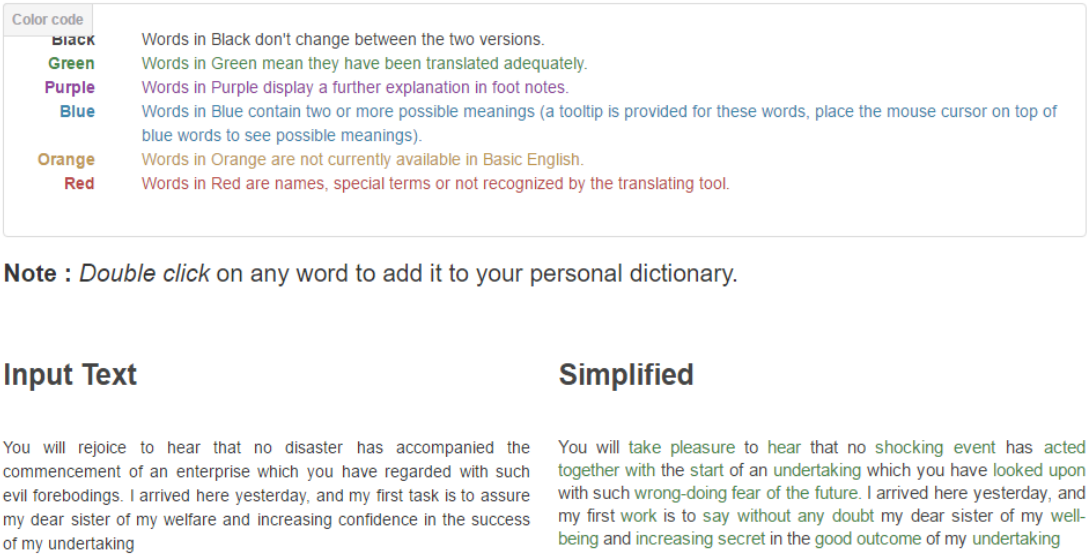
Joonis 1. Veebirakenduse Rewordify.com väljund näidistekstile [16].

Lisaks teistele funktsioonidele on jooniselt 1 näha, et asendamata sõna (*hear*) peale vajutades on kuvatud sõna definitsioon. Asenduse (*went with)* peale vajutades on näha nii

asendus kui ka algne versioon ning samuti on võimalik kuulata mõlema häälдust. Lisaks pakub veebirakendus teksti kohta statistikat, kus on kirjas teksti sõnade arv, silpide arv jm. Rakenduses on võimalik näidistekstideks valida erinevate klassikalise kirjanduse tekstikatkete hulgast [16].

1.3.2 Tekstilihtsustamise rakendus Simplish

Veebirakendus Simplish [17] pakub võimalust kontrollida õigekirja, teha sisukokkuvõtet ja lihtsustada teksti. Lihtsustamise juures on kirjeldatud, et rakenduse sõnavara hulka kuulub *ca* 120 000 sõna, mille abil üritatakse sisendtekst teisendada lihtsustatud inglise keelde (*Simple Basic English*), mis on leksikon, kuhu kuulub umbkaudu 1000 sõna [17].



Joonis 2. Veebirakenduse simplish väljund näidistekstile [17].

Jooniselt 2 on näha, et näidisteksti lihtsustamise tulemusena on saadud tekst:

„You will take pleasure to hear that no shocking event has acted together with the start of an undertaking which you have looked upon with such wrong-doing fear of the future. I arrived here yesterday, and my first work is to say without any doubt my dear sister of my well-being and increasing secret in the good outcome of my undertaking.”.

Tekstis on asendatud järgmised sõnad:

- *rejoice → take pleasure;*
- *disaster → shocking event;*
- *accompanied → acted together with;*
- *commencement → start;*
- *enterprise → undertaking;*

- *regarded → looked upon;*
- *evil → wrong-doing;*
- *forebodings → fear of the future;*
- *task → work;*
- *assure → say without any doubt;*
- *welfare → well-being;*
- *confidence → secret;*
- *success → good outcome.*

Võrreldes Rewordify.com väljundiga on Simplish teinud rohkem asendusi. Autori hinnangul on tekst paremini loetav, kui see oli Rewordify.com-i puhul, kuna igale sõnale on leitud täpselt üks asendus (*enterprise* asemel *undertaking*, mitte *business/project*), mistõttu on tekst tervikuna ladusam, kuid lihtsustused ei nii kvaliteetsed, kui eelmise rakenduse puhul. Tähenduse mõttes on tehtud ka mõned vead, nt *confidence* on asendatud *secret*'iga, mis ei ole antud kontekstis õige. Märkida tasub, et kõigi sõnade puhul on mõlemad rakendused teinud erinevaid asendusi (nt Rewordify.com *task → job*, Simplish *task → work*).

Lisaks on jooniselt 2 näha, et väljundis on sõnad esile toodud erinevate värvidega. Must värv väljendab, et asendusi pole tehtud; roheline, et asendus on leitud, ja punane, et sõna või terminit ei tuntud ära [17].

1.3.3 Tekstilihtsustamise rakendus Article Simplifier

Veebirakendus Article Simplifier [18] pakub võimalust lihtsustada teksti, kus keerulistele sõnadele leitakse sõnastiku abil lihtsam asendus.

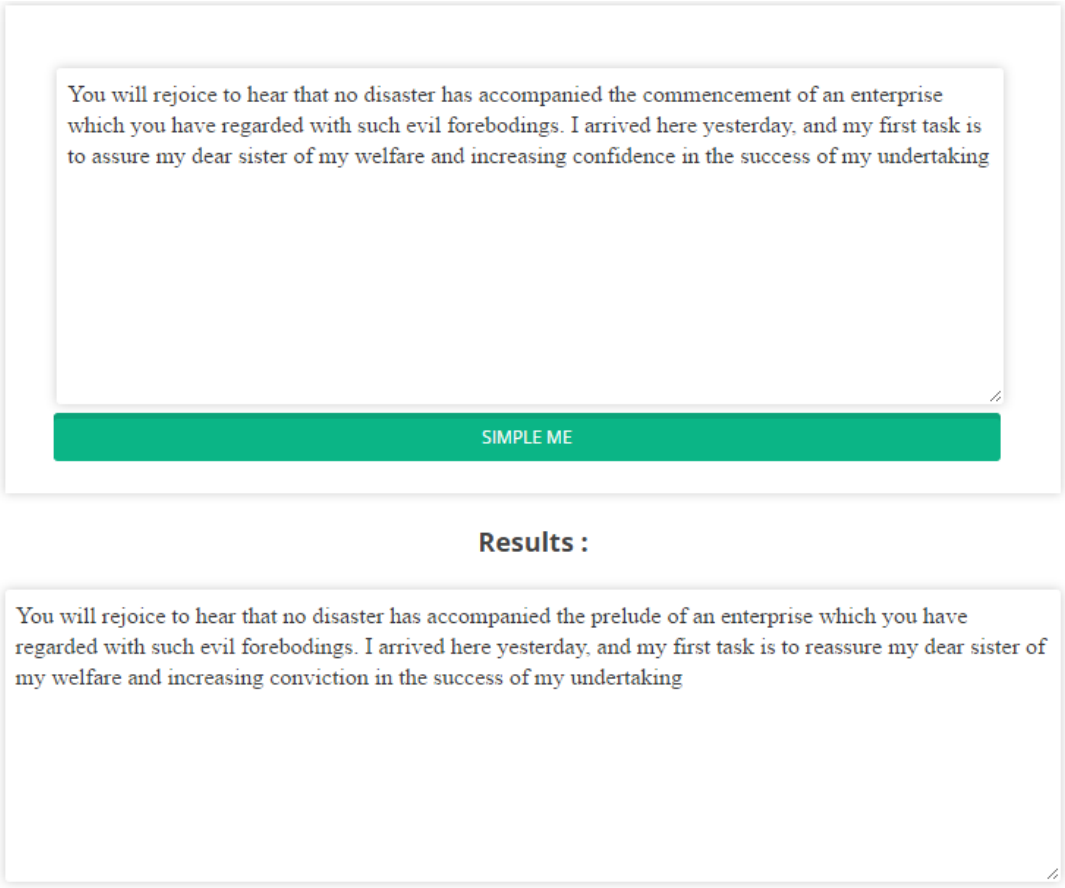
Jooniselt 3 on näha, et näidisteksti lihtsustamise tulemusena saadi tekst:

„*You will rejoice to hear that no disaster has accompanied the prelude of an enterprise which you have regarded with such evil forebodings. I arrived here yesterday, and my first task is to reassure my dear sister of my welfare and increasing conviction in the success of my undertaking.*"

Tekstis on asendatud järgmised sõnad:

- *commencement → prelude*;
- *assure → reassure;*
- *confidence → conviction.*

Article Simplifier on teinud eespool kirjeldatud programmidest kõige vähem asendusi ning üldine teksti keerukus ei ole autori hinnangul muutunud, kuna tehtud asendused ei ole oluliselt kergemad.



Joonis 3. Veebirakenduse Article Simplifier väljund näidistekstile [18].

Article Simplifier on lihtsa ülesehitusega ning ei paku lisaks lihtsustatud teksti kuvamisele teisi tegevusi.

2. Keeleressursid

Keeleressursid on masinloetaval kujul elektroonilised andmekogumid (sh ka tarkvara), mida kasutatakse loomuliku keele uurimiseks, keeletehnoloogia arendamiseks ning keeletarkvara väljatöötamiseks [19, 20]. Keeleressursid on näiteks tekstikorpused, kõneandmebaasid (nt helisalvestused jms), leksikaalsed ressursid (nt sõnastikud, sagedusloendid jms), tekstitöötlusvahendid (nt morfoloogiline analüüs, speller jms) ja kõnetöötlusvahendid (nt tekst-kõne-süntees) [18]. Eesti keele jaoks on võimalik keeleressursse leida Eesti Keeleressursside Keskuse veebilehelt[1]. Erinevate keeleressursside abil on võimalik koostada keerulisi süsteeme. Selles peatükis on kirjeldatud keeleressursse, mida on kasutatud teksti lihtsustamise rakenduse tegemiseks.

2.1 *Wordnet*

Wordnet on leksikosemantiline ehk sõnatähenduslik andmebaas, mis rühmitab nimi-, tegu-, määr- ja omadussõnad tähenduslikesse üksustesse ehk sünohulkadesse [21]. Iga sünohulk koosneb ühte ja sama mõistet väljendavatest ehk sünonüümsetest sõnadest, kus kõik sõnad on ühest sõnaliigist [22]. Sünohulgad on ühendatud mõistetevaheliste semantiliste või leksikaalsete suhete abil [23]. Ingliskeelse WordNeti arendamist alustati 1980. aastate keskpaiku Princetoni Ülikoolis ning see on olnud eeskujuks teistele keeltele [21, 23]. *Wordnet*-tüüpi andmebaase on loodud üle maailma enam kui 50 erineva keele jaoks ning need on loomuliku keele töötluses üks enim kasutatud ressursse [23]. Eesti Wordnet (edaspidi EstWN) on loodud Princeton WordNeti eeskujul, EstWN-i arendamist alustati 1996. aastal, kui Tartu Ülikool liitus EuroWordNet projektiga [22]. 2017. aasta jaanuari seisuga sisaldas EstWN (versioon 73) enam kui 77 800 mõistet (sh sõnu umbkaudu 106 200) ning üle 248 000 semantilise suhte [24]. EstWN-is on mõisted omavahel ühendatud 43 eri liiki suhtega, millest kõige sagedamini on määratletud ülem- ja alammõiste [23]. Lisaks mõistetevahelistele suhetele on EstWN-i üheks omaduseks mitmekeelsus ehk kõik mõisted on ühendatud keeltevahelise indeksi (ingl *InterLingual Index*) abil ingliskeelse Princeton WordNetiga. EstWN on koostatud peamiselt käsitsi eesti keele eripära arvestades ning aastas lisandub EstWN-i umbkaudu 7000 uut mõistet [23, 25].

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

2.2 Semantilised suhted

Semantilised ehk tähenduslikud suhted on seosed, mis esinevad sõnade tähenduste, erinevate fraaside tähenduste või lausete tähenduste vahel [26]. *Wordnet*-tüüpi andmebaasidel on suhted erinevate sünohulkade vahel. Sagedasemad suhted, mille abil ühendatakse kaks sünohulka, on sünonüümia, hüperonüümia ja osa-terviku suhe, mis koos määravad ära *wordnet'*i konstruktsiooni iseärasuse [23]. Viidatud artiklis on kirjutatud, et teist tüüpi suheteks on leksikaalse suhted, näiteks antonüümia suhe, mis seob sõnaga tema vastandtähendusliku sõna. Sünonüümia ja hüperonüümia suhteid on järgnevates peatükkides põhjalikumalt seletatud.

Näide antonüümia suhtest: *vana* ↔ *noor*.

Näide osa-terviku suhtest: *pea* ↔ *silmad, kõrvad, nina, juuksed jne.*

2.2.1 Sünonüümia

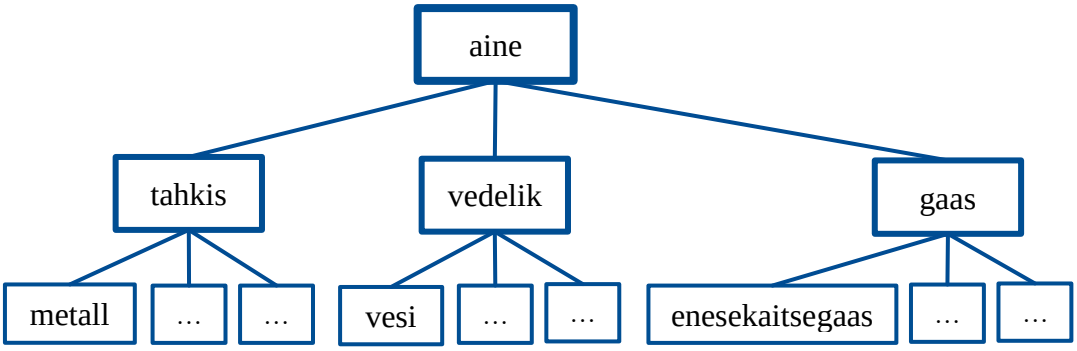
Sünonüümia ehk samatähenduslikkus on fundamentaalne suhe *wordnet'*i jaoks, mille alusel on sõnad jaotatud erinevatesse sünohulkadesse [27]. Sünonüümia võib olla lähedasem ja kaugem, lähedane sünonüümia on näiteks täissünonüümia, mille moodustavad sõnad, mille kõik tähendused on samad kõigis kontekstides [28]. Kuna täissünonüümiat leidub loomulikus keeles harva, sisaldab leksikon hulganisti piiratud asendatavusega sünonüüme, mistõttu on enamikes *wordnet'*ides kasutusele võetud osa- või lähisünonüümia suhe [23]. Osa- või lähisünonüümid on peaaegu sarnase tähendusega sõnad, kuid tähendus on kontekstipõhine [28]. EstWN-is on osa- või lähisünonüümia suhte nimetuseks *near_synonym* [23].

Näide täissünonüümist: *inimene* ↔ *isik.*

Näide osa- või lähisünonüümist: *jooksma* ↔ *liduma.*

2.2.2 Hüponüüm ja hüperonüüm

Hüponüümia on leksikaalne suhe kahe või enama mõiste vahel, kus kõrgema astme tähendus on alumisest laiem. Kõrgemal astmel olevat tähendust nimetatakse hüperonüümiks (ülemmõiste) ja madalamal olevat tähendust hüponüümiks (alammõiste). Kõrgem tähendus sisaldab endas temast allpool olevad mõisted [28]. EstWN-is on sõna ülem- ja alammõiste suhete nimetuseks vastavalt *has_hyperonym* ja *has_hyponym* [29].



Joonis 4. Ülemmõiste ja alammõiste hierarhia [30].

Joonisel 4 on kujutatud ülemmõiste ja alamõiste vahelisi seoseid. Sõna ’tahkis’ puhul on ülemmõisteks sõna ’aine’, alammõisteks aga sõna ’metall’. Sama sõna saab olla nii hüponüüm kui ka hüperonüüm, näiteks nagu ’vesi’ kui vedeliku suhtes ning ’vesi’ kui allikavee suhtes [30].

2.3 Eesti keele morfoloogiline analüsaator, süntesaator ja ühestaja

Eestikeelne morfoloogiline analüsaator ja süntesaator on tekstitöötlemise vahendid, mille abil on võimalik eesti keele tekst muuta kergemini töödeldavaks ning vastupidi, et muuta töödeldud tekst tagasi loomulikku keelde [31]. Morfoloogilise ühestamise eesmärk on mitmetähenduslike sõnade puhul kontekstipõhiselt tuvastada kõige sobivam variant [31]. Eesti keele jaoks on nimetatud vahendid kasutatavad Pythoni teegis EstNLTK [32].

Teksti morfoloogiliseks analüüsimiseks kasutatakse morfoloogilist analüsaatorit, mis leiab antud sõna vormi abil sõnale vastava algvormi (**lemma**), struktuuri (nt tüvi, lõpp), sõnaliigi ja kategooriad, kui sõna on mitmeti analüüsitav, siis väljastatakse kõik võimalused [31].

Morfoloogiline süntesaator on vastand analüsaatorile ning sünteesib etteantud lemma ja sõnaliigi, käände või pöörde põhjal selle sõna muutevormi või muutevormid [31].

2.4 Sagedusloend

Sagedusloend on sõnaloend, kus on välja toodud kõige sagedamini esinevad sõnad. Töös kasutatud sagedusloend on loodud 15 miljoni sõna suuruse korpuse baasil, kus sõnad on võetud aja-, ilu- ja teaduskirjandusest [33]. Sagedusloendis on paaridena esitatud sõna lemma ja sagedus, loendis esineb üle 43 000 sõna, kus sagedused varieeruvad alates 10-st kuni 639 802-ni[2]. Siinses töös kasutatakse terminit lemma esinemissagedus, mis väljendab konkreetse lemma sageduse arvu sagedusloendis.

[2] Tasakaalus korpuse lemmad sageduse järjekorras. https://keeleressursid.ee/images/cl-ut-ee/sagedusloendid/lemma_kahanevas.txt

2.5 Eesti keele põhisõnavara sõnastik

Eesti keele põhisõnavara sõnastik koosneb 5000 eesti keele olulisemast sõnast, mis aitavad eesti keelt õppida, sõnastikus on lisaks tähenduste ja käänete vaatamisele võimalik kuulata sõna häädlust ja näha illustreerivaid pilte [34]. Käesolevas töös kasutatakse põhisõnavara märksõnaloendit[3].

2.6 Eesti keele võõrsõnade leksikon

Eesti keele võõrsõnade leksikon koosneb üle 33 000 levinuimast mõistest, mille hulka kuuluvad võõr-, tsitaat- ja laensõnad kui ka lühendid, sententsid ja väljendid [35]. Siinses töös on kasutatud võõrsõnade leksikoni märksõnaloendit, mis koosneb üksikutest lemmadest, ning sõnatähenduste loendit, kus iga mõiste jaoks on välja toodud ka mõiste seletus[4].

2.7 *Word2vec*

Word2vec on meetod, millega leitakse sõnade jaoks vektoresitus [36]. Viidatud artiklis on kirjeldatud, et tegemist on kahekihilise tehisnärvivõrguga: sisendkorpuse sõnadega leitakse korpuses esinevatele sõnadele vektorid, kus saadud vektorid rühmitavad sarnaseid sõnu. Kahe sõna vaheline sarnasus leitakse neile vastavate vektoritevahelise koosinuskaugusega, mistõttu jääb sarnasust määrav arv vahemikku -1 kuni 1.

Tabel 1. *Word2vec*'i sarnasused sõnaga *'Sweden'* [37].

Sõna	Koosinuskaugus sõnaga '*Sweden'*
Norway	0,760124
Denmark	0,715460
Finland	0,620022
Switzerland	0,588132
Belgium	0,585835
Netherlands	0,574631
Iceland	0,562368
Estonia	0,547621

[3] Eesti keele põhisõnavara sõnastik 2014 märksõnaloend. https://www.eki.ee/litsents/
[4] Võõrsõnade leksikon XML, märksõnaloend. https://www.eki.ee/litsents/

Slovenia	0,531408

Tabelis 1 on välja toodud *word2vec'*i sarnasused sõnale *'Sweden'*, millest on näha, et Rootsiga kõige sarnasemad sõnad on tema naaberriikide nimed Norra, Taani ja Soome. Lisaks lähimate sarnaste sõnade leidmisele on *word2vec'*i abil võimalik leida kahe etteantud sõna sarnasus. Näiteks sõnade *'Finland'* ja *'Sweden'* vaheline sarnasus on 0,620022.

Tabel 2. *Word2vec'*i sarnasused sõnaga 'lõputöö'.

Sõna	Koosinuskaugus sõnaga 'lõputöö'
diplomitöö	0,846841
kursusetöö	0,798264
bakalaureusetöö	0,790251
magistritöö	0,784765
doktoritöö	0,770832
seminaritöö	0,730663
väitekiri	0,659725
doktoriväitekiri	0,656953
semestritöö	0,639264
kandidaaditöö	0,612238

Tabelis 2 on *word2vec'*i EstNLTK mudeli [38] põhjal leitud 10 lähedasemat sõna sõnale 'lõputöö', kus sarnasemad sõnad on erinevad lõputööde liigid.

3. Eestikeelse teksti lihtsustamise veebirakendus

Bakalaureusetöö käigus on loodud eestikeelset teksti lihtsustav veebirakendus, mis põhineb leksikaalse ehk sõnade lihtsustamise meetodil. Rakendus paikneb Eesti Keeleressursside Keskuse veebiserveris[5] ning lähtekood on kättesaadav GitHubis[6].

3.1 Veebirakenduse tutvustus

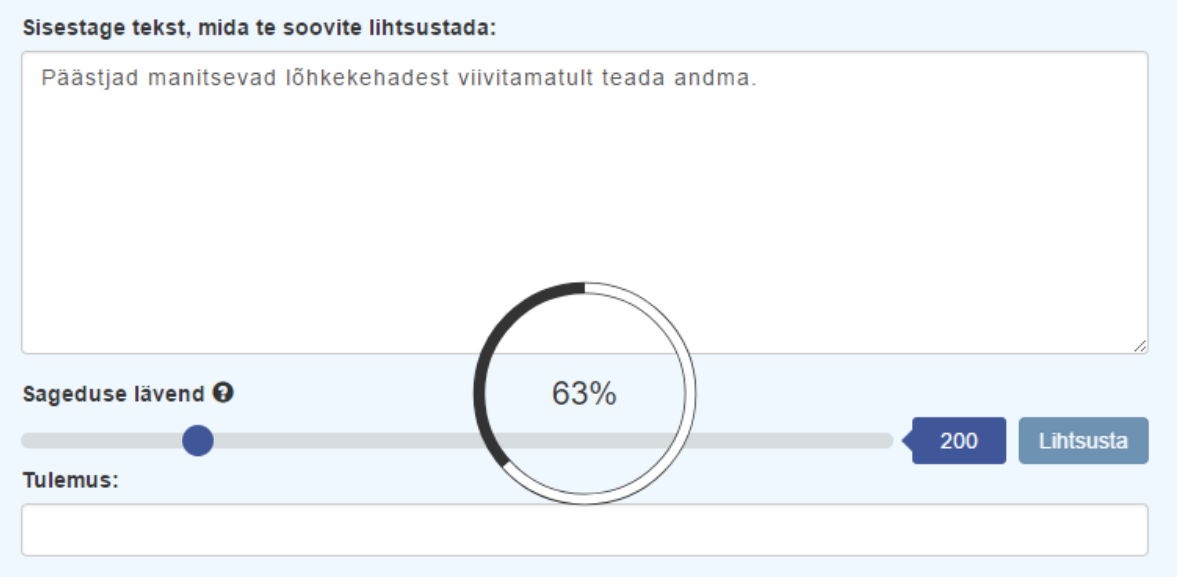
Loodud veebirakenduse kasutajaliides koosneb ühest lehest, mis on kujutatud joonisel 6.

Joonis 5. Veebirakenduse esileht.

Rakenduse üldine disain on sarnane Article Simplifier ülesehitusega. Põhikohal on teksti sisestamise väli, lihtsustamise nupp ning tulemusekast. Kasutajal on võimalik tekstikasti ise kirjutada või kopeerida tekst kusagilt mujalt. Samuti on rakenduses muudetav sageduse lävend, mis määrab, milliseid sõnu tuleb lihtsustada. Sagedusloendis on iga lemmaga seotud temale vastav esinemissagedus ning lihtsustada üritatakse sõnu, mille esinemissagedus jääb alla lävendi. Seega lävendi kasvades suureneb lihtsustavate sõnade hulk. Näiteks sõna ’manitseb’ puhul on sagedusloendis sõna ’manitsema’ sagedusega 167, mistõttu lävendi 150 juures seda sõna ei lihtsusta. Vajutades lävendi juures asetsevale küsimärgi ikoonile, on

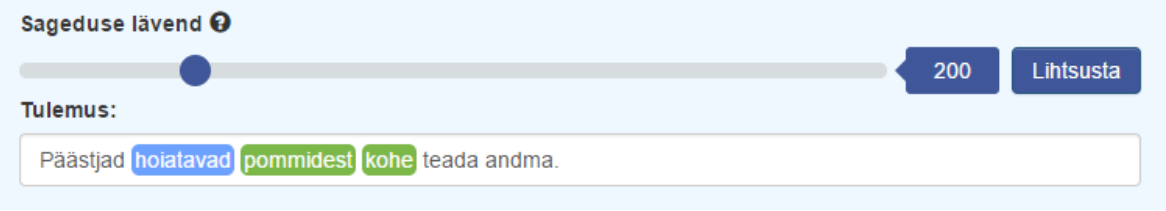
[5] http://prog.keeleressursid.ee:4567/
[6] https://github.com/mpeedosk/teksti-lihtsustamine

võimalik lugeda lävendi kohta lisainfot. Lävendi vaikeväärtus on 150, millest jääb sagedusloendist allapoole 77% sõnadest. Teksti lihtsustamiseks tuleb vajutada nuppu „Lihtsusta“, mispeale saadetakse päring serverile ning kasutajale kuvatakse progressiriba (kujutatud joonisel 6).



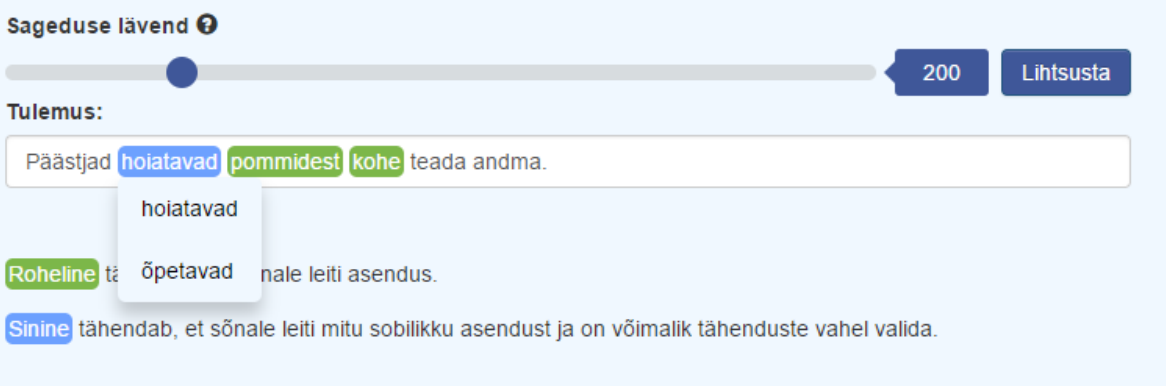
Joonis 6. Hinnanguline progressiriba.

Progressiriba eesmärk on anda kasutajale tagasisidet, et rakendus on tema tegevusele reageerinud ja tulemuse saamiseks tuleb oodata. Progressiriba on hinnanguline ja ei pruugi vastata tegelikkuses kuluvale ajale. Hinnangu määramiseks kasutatakse sisestatud teksti pikkust ning lävendi suurust. Mida pikem on tekst, seda kauem võib töötlemine kesta, samamoodi on lävendiga: suurema lävendi puhul on võimalike asenduste arv suurem. Kui serverilt saadakse vastus, siis progressiriba peidetakse. Joonisel 7 on kujutatud veebirakendust, kuhu kasutajale on tulemus juba kuvatud. Sarnaselt Simplish ja Rewordify.com rakendustele on asendatud sõnad välja toodud värviliselt, et neid oleks kergem eristada. Samuti on erinevatel värvidel erinevad tähendused rakenduse seisukohalt.



Joonis 7. Veebirakenduse poolt lihtsustatud tekst.

Roheline tähendab, et rakendus leidis algse sõna jaoks täpselt ühe asenduse. Sinine aga väljendab olukorda, kus ühele sõnale leiti mitu sobivat asendussõna. Jooniselt 8 on näha, et sõnale ’manitsevad’ on leitud asenduseks nii ’hoiatavad’ kui ka ’õpetavad’.

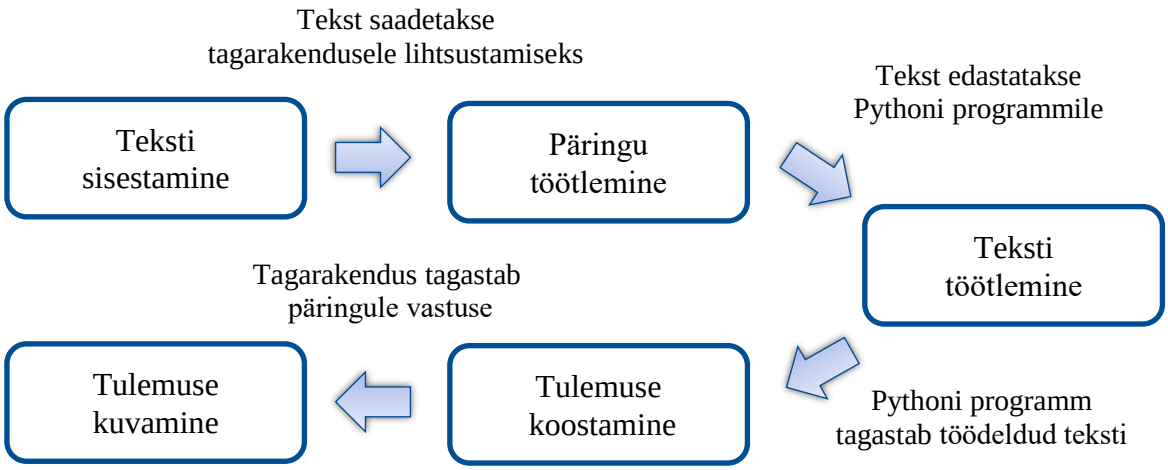


Joonis 8. Menüü, kui sõnale leiti mitu asendussõna.

Menüü sees on sõnad järjestatud sobivuse alusel. Näidatud lause puhul sobib algoritmi põhjal sõna ’hoiatavad’ paremini kui ’õpetavad’.

3.2 Kasutatud tehnilised vahendid

Loodud veebirakendus koosneb kolmest osast: eesrakendus (ingl *front end)*, tagarakendus (ingl *back end)* ja teksti lihtsustamise programm. Eesrakenduse kaudu saab kasutaja sisestada teksti, mida soovitakse lihtsustada, ja seejärel lugeda tulemust. Tagarakendus tegeleb kasutajaliidese poolt tulevate päringute töötlemisega ning edastab vajalikud parameetrid koos tekstiga edasi teksti lihtsustavale programmile. Lihtsustamise programmis rakendatakse tekstile algoritmi, mis on täpsemalt lahti seletatud peatükis 3.3. Programmi töö lõppedes tagastatakse lihtsustatud tekst tagarakendusele, mis tagastab tulemuse kasutajale. Veebirakenduse töövoog on kujutatud joonisel 9.



Joonis 9. Teksti lihtsustamise töövoog.

Erinevate osade tehnilised lahendused on järgnevalt alampeatükkide kaupa kirjeldatud.

3.2.1 Eesrakendus

Eesrakenduse osa on kirjutatud HTMLi, CSSi ja JavaScripti abil, mis on toetatud enamus kaasaegsete veebilehitsejate poolt [39]. HTML (ingl *Hypertext Markup Language)* on märgenduskeel, mille abil kirjeldatakse veebidokumentide struktuuri. CSS (ingl *Cascading Style Sheets)* on keel, millega määratakse dokumendi kujundus. JavaScript on programmeerimiskeel, mida kasutatakse interaktiivsete veebilehtede loomiseks. Lisaks on töös kasutatud JavaScripti teeke jQuery [40], mis lihtsustab serveriga suhtlemist ja ProgressBar.js [41], mis aitab visualiseerida ooteaega. Veebirakenduses on kasutatud ka raamistikku Bootstrap [42], mis pakub moodsat veebilehe kujundusstiili.

3.2.2 Tagarakendus

Tagarakenduse osa on kirjutatud programmeerimiskeeles Java, kasutades versiooni 1.8. Veebirakenduse ülesseadmiseks on kasutatud raamistikku Sparkjava [43], mis lihtsustab Java-põhiste veebirakenduste loomist. Vaadete töötlemiseks on kasutatud Thymeleafi [44], mis aitab staatilist HTML koodi muuta dünaamilisemaks. Lisaks on kasutatud teeki Deeplearning4j [45], mis pakub erinevate masinõppe meetodite teostusi, sh *word2vec*. *Word2vec'*i juures on kasutatud EstNLTK mudelit, mis on treenitud 16 miljoni lause suuruse korpuse põhjal [38].

```
File gModel = new File("lemmas.cbow.s200.w2v.bin");
Word2Vec vec = WordVectorSerializer.readWord2VecModel(gModel);
double cosSim = vec.similarity("tartu", "tallinn");
System.out.println(cosSim);

> 0.6300944089889526
```

Koodiplokk 1. Deeplearning4j *word2vec'*i sarnasus sõnade Tallinn ja Tartu vahel.

Koodiplokis 1 on välja toodud näide Deeplearning4j teegi *word2vec*'i kasutamisest, kus on leitud kahe sõna vaheline sarnasus.

3.2.3 Teksti lihtsustamine Pythonis

Teksti lihtsustamise osa on kirjutatud programmeerimiskeeles Python (versiooni 3.5). Eesti keele töötlemiseks kasutatakse peatükis 2.3 mainitud teeki EstNLTK, mis lisaks muudele funktsioonidele võimaldab teha morfoloogilist analüüsi ja sünteesi, ühestamist ning päringuid EstWN-ile [32].

```
import estnltk
text = estnltk.Text('Poiss jooksis koolist koju.')
print(text.analysis)
> [
[{'root_tokens': ['poiss'], 'clitic': '', 'root': 'poiss', 'ending':
'0', 'form': 'sg n', 'partofspeech': 'S', 'lemma': 'poiss'}],
[{'partofspeech': 'V', 'ending': 'is', 'root': 'jooks', 'clitic': '',
'root_tokens': ['jooks'], 'lemma': 'jooksma', 'form': 's'}],
[{'root_tokens': ['kool'], 'clitic': '', 'root': 'kool', 'ending':
'st', 'form': 'sg el', 'partofspeech': 'S', 'lemma': 'kool'}],
[{'root_tokens': ['kodu'], 'clitic': '', 'root': 'kodu', 'ending': '0',
'form': 'adt', 'partofspeech': 'S', 'lemma': 'kodu'}],
[{'root': '.', 'form': '', 'lemma': '.', 'partofspeech': 'Z', 'clitic':
'', 'ending': '', 'root_tokens': ['.']}]
]
```

Koodiplokk 2. EstNLTK morfoloogiline analüüs.

Koodiplokis 2 on näidatud, et pärast sõnade analüüsi on leitud sõnade algvorm (*lemma*), sõnaliik (*partofspeech*), sõnavorm (*form*), kliitik (*clitic*), sõna lõpp (*ending*), tüvi (*root*) ja liitsõnade puhul ka kõigi osasõnade tüved (*root_tokens*).

```
import estnltk
print(estnltk.synthesize('jooksma', 's'))
> ['jooksis']
print(estnltk.synthesize('niit', 'pl n'))
> ['niidid', 'niidud']
```

Koodiplokk 3. EstNLTK morfoloogiline süntees.

Koodiplokis 3 on toodud näide morfoloogilisest sünteesist, kus on etteantud lemma ning soovitud sõnavorm. Kui lemmal leidub antud sõnavormis mitu kuju, siis väljastatakse need kõik.

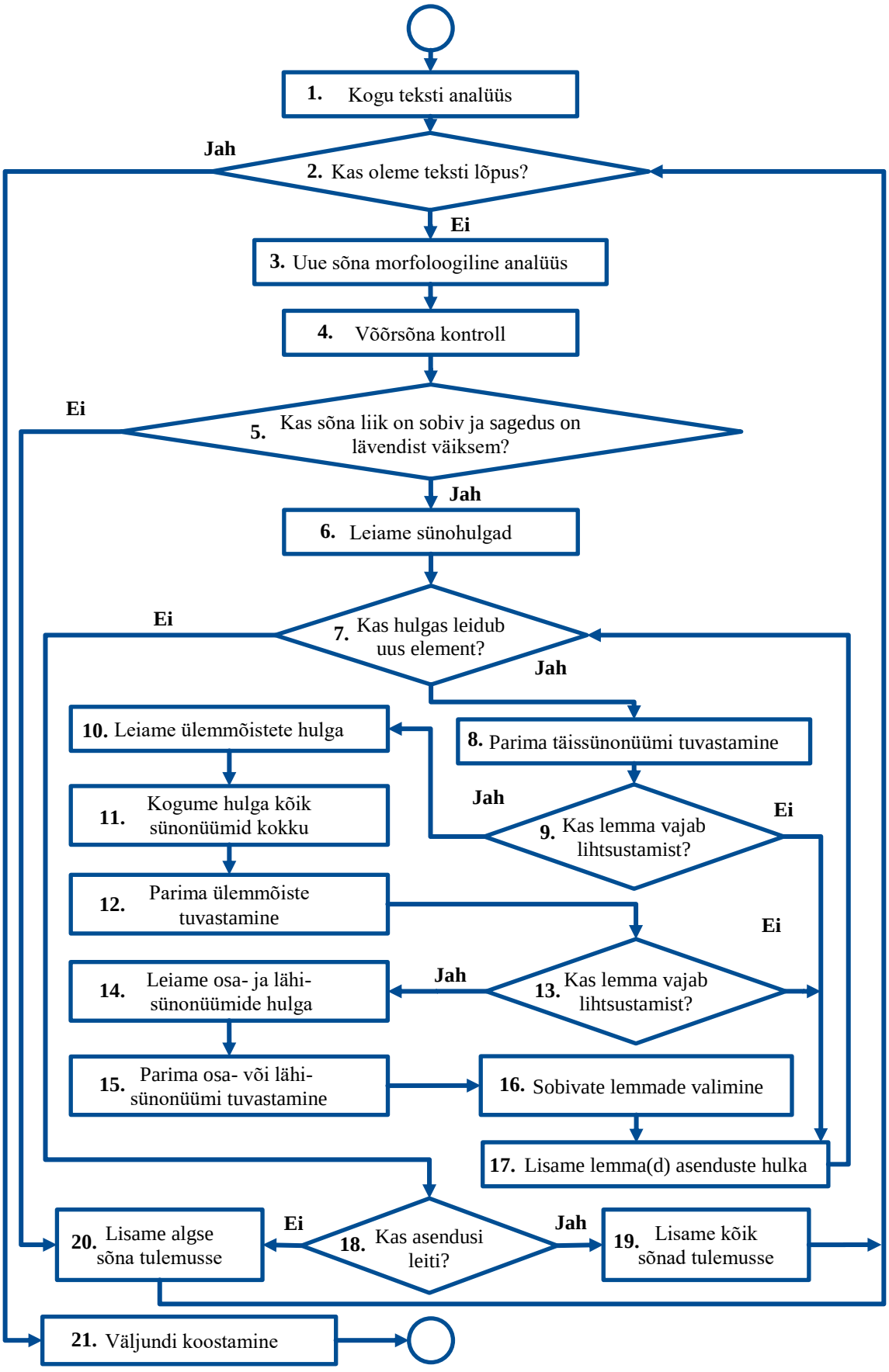
```
from estnltk.wordnet import wn
synset = wn.synsets("liduma", pos=wn.VERB)
print(synset)
> ["Synset('punuma.v.02')", "Synset('lippama.v.01')"]
hypernym = synset[0].hypernyms()
print(hypernym)
> ["Synset('jooksma.v.02')"]
```

Koodiplokk 4. EstNLTK EstWN-i päring.

Koodiplokis 4 on toodud näide EstWN-i kasutamisest EstNLTK-i kaudu. Etteantud lemma ja sõnaliigi põhjal leitakse kõik sobivad sünohulgad. Sünohulkadevahelised päringud tagastavad vastava seosega sünohulgad.

3.3 Algoritm

Teksti lihtsustamise algoritm on kujutatud joonisel 10. Algoritm on põhjalikumalt kirjeldatud sammude kaupa joonise järel.



Joonis 10. Teksti lihtsustamise algoritm.

Iga samm on väljendatud loendis järjekorra numbriga.

1. Teksti lihtsustamise algoritm alustab kogu teksti analüüsiga: tekst ühestatakse, jagatakse lauseteks ning lausetest eraldatakse sõnad (sh kirjavahemärgid).
2. Sõnad analüüsitakse ükshaaval läbi kuni teksti lõpuni. Kui kõik sõnad on analüüsitud, asub algoritm 21. sammu juurde.
3. Igale sõnale tehakse morfoloogiline analüüs, kus algoritmi jaoks on olulisel kohal sõnavorm, sõnaliik ja sõna lemma.
4. Esimesena kontrollitakse, kas sõna lemma leidub võõrsõnade leksikonis. Kui leksikonis selline lemma leidub ning sellele vastab eesti omasõna, siis lisatakse omasõna asenduste hulka.
5. 3. sammul analüüsitud sõna puhul kontrollitakse, kas sõnaliik on sobiv, st et sõnaliik on kas nimi-, tegu-, määr- või omadussõna. Kui sõnaliik ei ole sobiv, siis ei ole seda võimalik EstWN-i abil lihtsustada ja algoritm asub 20. sammu juurde. Sobiva sõnaliigi puhul analüüsitakse, kas lemma vajab lihtsustamist. Selleks kontrollitakse, kas lemma leidub sagedusloendis ja kas selle esinemissagedus on väiksem kui kasutaja määratud lävend või on tegemist võõrsõnaga.
6. Kui lemma vajab lihtsustamist, leitakse sellele vastav EstWN-i kirje ehk ühtlasi ka sünohulgad. Kui lemmale vastab mitu EstWN-i kirjet, siis järjestatakse need kontekstipõhiselt, kus iga sünohulga jaoks leitakse seal leiduvate lemmade keskmine sarnasus algsele sõnale eelneva ja järgneva lemmaga.
7. Leitud sünohulkade kõik elemendid analüüsitakse ükshaaval läbi. Kui kõik elemendid on analüüsi läbinud, asub algoritm 18. sammu juurde.
8. Ühest sünohulgast parima lemma tuvastamiseks leitakse iga sünonüümi jaoks selle esinemissagedus ja *word2vec'*i abil sarnasus algse sõna lemmaga. Sobivuse hindamiseks kasutatakse valemit (2). Kõikide sünonüümide seast valitakse sünonüüm, millel on suurim hinnang.
9. Kui parima sünonüümi esinemissagedus ei vasta lävendile, analüüsitakse algse sõna lemma ülemmõisteid. Kui aga sünonüüm sobib asenduseks, asub algoritm 17. sammu juurde.
10. Algse sõna lemmale leitakse vastava seosega ülemmõistete hulk ning ülemmõisted järjestatakse kontekstipõhiselt.
11. Iga ülemmõiste jaoks lisatakse selle sünohulga sünonüümid ühisesse järjendisse.

12. Saadud sünonüümide järjendi igale elemendile leitakse esinemissagedus ja sarnasus algse sõna lemmaga, mille põhjal arvutatakse valemit (5) kasutades hinnang. Sünonüümide seast valitakse lemma, mille hinnang on suurim.
13. Leitud ülemmõiste puhul kontrollitakse, kas see on piisavalt sarnane algse sõna lemmaga. Kui see sobib, siis asub algoritm 17. sammu juurde, vastasel juhul analüüsitakse algse sõna lemma osa- ja lähisünonüüme.
14. Algse sõna lemmale leitakse vastava seosega osa- ja lähisünonüümide sünohulgad.
15. Osa- ja lähisünonüümide sünohulkade seast leitakse esinemissageduse ja sarnasuse põhjal valemiga (5) lemma, mille hinnang on suurim.
16. Leitud sünonüümi, ülemmõiste ja osa- või lähisünonüümi lemmad üritatakse kõik asenduste hulka lisada.
17. Leitud lemmale või lemmadele sünteesitakse vastav sõnavorm, mis lisatakse asenduste hulka, kui lemma esinemissagedus on suurem algse sõna lemma esinemissagedusest. Seejärel asub algoritm tagasi 7. sammu juurde ning üritab järgmist sünohulka analüüsida.
18. Kui kõik algse sõna lemma sünohulga elemendid on läbi vaadatud, kontrollitakse, kas analüüsi käigus on asendusi leitud.
19. Kui algoritm leidis sõnale vähemalt ühe asenduse, siis lisatakse kõik asendused lõpptulemusse. Asendused järjestatakse algse sõna lemma sarnasuse alusel ning kasutajal on võimalik nende vahel valida. Algoritm asub tagasi 2. sammu juurde, kus üritatakse sisendi järgmist sõna analüüsida.
20. Kui ühtegi asendust ei leitud, lisatakse lõpptulemusse algne sõna. Algoritm asub tagasi 2. sammu juurde, kus üritatakse sisendi järgmist sõna analüüsida.
21. Kui teksti kõik sõnad on analüüsitud, koostatakse lõpptulemuse järjendi põhjal uus tekst, kus on algoritmi poolt tuvastatud lihtsustused sooritatud.

3.3.1 Loenditest ebasobivate sõnade eemaldamine

Jõudluse parandamiseks tehti eeltöötlust nii sagedusloendile, põhisõnavara sõnastiku ja võõrsõnade leksikoni märksõnaloenditele kui ka võõrsõnade leksikoni sõnatähenduste loenditele. Eeltöötluse käigus eemaldati loenditest sõnad, millele ei leidunud EstWN-is vastet, lisaks eemaldati sõnad, mille puhul EstWN-is täissünonüümid puudusid või millel ei leidunud ühtegi vajalikku seost (ülemmõistet, alammõistet ega osa-või lähisünonüümi). Võõrsõnade puhul koostati töödeldud märksõnade loendi ja tähendussõnastiku põhjal uus loend, kuhu lisati kõik tähendussõnastiku sõnad koos vastavate seletustega ning need sõnad

märksõnadest, mis puudusid põhisõnavara sõnastikust. Eeltöötluse tulemusena vähenes sagedusloendi sõnade arv 40%, põhisõnavara loendi sõnade arv 9,2% ning kahe võõrsõna loendi asemel kasutatakse ühte loendit, kus on 61% vähem sõnu kui varasemalt kahe loendi peale kokku.

3.4 Probleemid ning lahendused

Programmi katsetamise käigus ilmnesid probleemid nii sõna valiku, morfoloogilise sünteesi kui ka sõnaühendite juures.

3.4.1 Esinemissageduse põhjal asendamine

Algoritmi koostamisel katsetati asenduste tegemist ainult esinemissageduse põhjal. Sellisel juhul valiti alati sünonüümide ja ülemmõistete hulgast sagedusloendis kõige suurema esinemissagedusega sõna. Kõige kriitilisem viga sellise lähenemise puhul oli see, et pärast asendust kaotas lause tihti algse tähenduse.

Näide 1: *Õpetaja valdas mitut võõrkeelt → Õpetaja oli mitut võõrkeelt.*

Vea põhjuseks oli asjaolu, et ’valdama’ (esinemissagedusega 547) ülemmõiste sünohulgas leidus sõna ’olema’ esinemissagedusega 639 802. Seega sageduse mõttes tehti õige asendus, kuid selles kontekstis selline asendus ei sobi. Teiseks suuremaks veaks oli olukord, kui asendus oluliselt lause tähendust ei muutnud, kuid siiski ei sobinud autori arvates selles kontekstis.

Näide 2: *Põrandal tatsab mardikas → Põrandal käib putukas.*

Sõna ’tatsab’ ülemmõistete hulgas on sõnad ’käima’, ’kõndima’ ja ’sammuma’, mille hulgast sobib autori hinnangul kõige paremini sõna ’kõndima’.

Probleemi lahendamiseks võeti lisaks esinemissagedusele kasutusele sõnade sarnasus, mis leitakse *word2vec*’i abil. Sõnade ’tatsama’ ja ’käima’ vaheline seos on arvuliselt 0,517, kuid sõnade ’tatsama’ ja ’kõndima’ vaheline seos on 0,780, seega on sõnad ’tatsama’ ja ’kõndima’ tugevamalt seotud. Seda teadmist on võimalik ära kasutada sobivuse hindamiseks.

3.4.2 Asenduse sobivuse hindamine

Õige asenduse valimisele lisas keerukust esinemissageduse suur varieeruvus. Probleemi lahendamiseks kasutatakse asenduse sobivuse hindamiseks (tähistatud *h*) valemeid, mis

arvestaksid nii *word2vec*'i abil leitud sõnade sarnasuse väikeseid väärtusi kui ka esinemissageduse suuri väärtusi.

Erinevaid valemeid katsetati jooksvalt testimise käigus, et välja selgitada, milline neist annab praktikas paremaid tulemusi. Valemi (1) puhul osutus sõna sagedus liiga domineerivaks, mistõttu peaaegu alati valiti asendamiseks suurima sagedusega sõna. Siin kaaluti ka sageduse teisendamist vahemikku [0,1], kuid see ei lahendaks probleemi, sest sõna 'olema' oleks ikka 400 korda sagedasem kui sõna 'viibima'.

$$\boldsymbol{h = sagedus * sarnasus} \tag{1}$$

Järgmisena suurendati sarnasuse osakaalu hinnangule arvutamisel. Valemi (2) puhul mõjutas sõnade sarnasus hinnangu väärtust juba olulisel määral. Suurte sageduste, kuid väikeste sarnasuste asemel eelistati sageli asenduseks sõnu, millel oli keskmine sarnasus ja sagedus. See valem osutuks heaks täissünonüümide valimisel, sest nende puhul võis eeldada, et need on omavahel väga sarnased.

$$\boldsymbol{h = sagedus * sarnasus^2} \tag{2}$$

Kuid siiski oli probleemiks sõnad, mille sagedusete erinevus oli üle paarisaja. Valemi (3) puhul katsetati logaritmi kasutamist. Asendusi tehti võrdlemisi tasakaalukalt, suured sagedused ei kallutanud enam hinnangut nii oluliselt, kuid probleemseks osutusid suured sagedused, kus hinnang oluliselt ei muutunud sageduse kasvades.

$$\boldsymbol{h = ln(sagedus) * sarnasus} \tag{3}$$

Valemis (4) otsustati kuupjuure abil sõnade sageduse suhet parandada, mille tulemusena oli 'olema' vaid 2,2 korda sagedasem kui sõna 'viibima'. Selle valemi puhul tehti asendusi samuti võrdlemisi tasakaalukalt, kuid sarnasused mõjutasid hinnangut veidi palju.

$$\boldsymbol{h = \sqrt[3]{sagedus} * sarnasus^2} \tag{4}$$

Valemi (5) juures otsustati sarnasuse ruutuvõtmise asemel kasutada teist lähenemist. See valem osutus katsetatud valemite seas kõige efektiivsemaks, suured varieeruvused sageduse ja sarnasuse seas mõjutasid hinnangut soovitud määral.

$$\boldsymbol{h = \sqrt[3]{sagedus} * \frac{sarnasus}{\sqrt{1 - sarnasus}}} \tag{5}$$

Tabel 3. Autori katsetused hinnangu arvutamisel.

Sõna	kõndima	käima	sõna1	sõna2	sõna3
Esinemissagedus (sagedus)	1298	13 680	25 000	2500	250
Sarnasus sõnaga ’tatsama’	0,78	0,52	0,25	0,50	0,75
Hinnang valemiga (1)	1 012,44	7 113,60	6250	1250	187,50
Hinnang valemiga (2)	789,70	3 699,01	1562,50	625	140,63
Hinnang valemiga (3)	5,59	4,95	2,53	3,91	4,14
Hinnang valemiga (4)	6,64	6,47	1,83	3,39	3,54
Hinnang valemiga (5)	18,14	17,95	8,44	9,60	9,45

Tabelis 3 on välja toodud autori poolt proovitud valemid, mille abil katsetati erinevaid lähenemisi asenduse sobivuse hindamiseks. Omavahel on võrreldud sõnu ’kõndima’ ja ’käima’ ning lisaks on välja toodud 3 sõna, mille eesmärk on illustreerida valemite käitumist erinevates olukordades. ’Sõna1’ väljendab olukorda, kus sõna on suure esinemissageduse, kuid väikese sarnasusega; ’sõna2’ olukorda, kus sõna on keskmise esinemissagedusega ning keskmise sarnasusega ja ’sõna3’ olukorda, kus sõna on väikese esinemissagedusega, kuid suure sarnasusega. Valemi (1) puhul on näha, et üldiselt on suurem hinnang sõnadel, millel on sagedus suurem. Valemi (2) juures on samamoodi, kuid erinevus hinnangute vahel ei ole enam nii suur. Valemi (3) ja (4) puhul on suurem hinnang sõnadel, mille sagedus on suurem. Valemi (5) puhul on hinnang autori arvates heas tasakaalus: keskmise sageduse ja sarnasusega sõnal on suurem hinnang, kui suure sageduse, kuid väikese sarnasusega sõnal ja väikese sageduse kuid suure sarnasusega sõnal.

3.4.3 Sõnade vähene sarnasus

Mõne asenduse puhul oli leitud sõna sarnasus algse sõnaga liialt väike, mistõttu muutusid laused ebaloogiliseks. Selliseid olukorrad esinesid, kui sõnadel puudusid täissünonüümid ja ülemmõiste oli liialt üldine.

Näide: *Tal on palju saladusi* → *Tal on palju seisundeid*.

Probleemi lahendamiseks lisati algoritmi sarnasuse alampiir, millest väiksema sarnasusega sõnu tulemusse ei lisata.

3.4.4 Sünteesivead

Probleemiks osutus ka EstNLTK morfoloogiline süntesaator, kus mõne lemma puhul leidus sünteesitud sõnade hulgas sõnu, mis ei ole eestikeelsed.

Näide: *Kõik on kriminaalkoodeksi ees võrdsed → Kõik on seadsa ees võrdsed.*

Probleemi lahendamiseks lisatakse kõik sünteesi käigus saadud sõnad tulemusse, mille hulgast on kasutajal ise võimalik õige valida.

3.4.5 Keeruliste sõnade puudumine EstWN-is

Kuigi EstWN-is leidub hulgaliselt mõisteid, leidub samuti sõnu, mille puhul seal kirje puudub. Selliste sõnade puhul jätab rakendus sõna lihtsustamata ning liigub järgmise sõna juurde.

3.5 Tulemused

Eesti keele jaoks puudub lihtsustatud teksti nn kuldstandard, mille alusel saaks rakenduse väljundit hinnata. Seega tulemuste hindamiseks tehti küsitlus, kus oli välja toodud 7 lihtsustatud näidisteksti erinevatest liikidest. Lisaks lihtsustatud tekstide hindamisele paluti vastajatel katsetada veebirakendust mõne enda leitud tekstiga ning anda selle lihtsustamisele tagasisidet. Lõpetuseks küsiti rakenduse kohta üldist tagasisidet. Küsimustiku küsimused on näha lisas 1. Küsimustiku eesmärk ei olnud teha kvantitatiivset analüüsi, vaid pigem saada ettepanekuid ning tagasisidet tulemustele.

Küsimustikule vastas 16 inimest, kellest 10–19 aastased moodustasid 19%, 20–29 aastased 44%, 30–39 aastased 12% ja 40–59 aastased 25%. Saadud tagasiside oli positiivne, lihtsustatud näidistekstid tundusid 87% vastanute jaoks lihtsamad. Lisaks leidis 11% vastanutest, et teksti oleks saanud paremini lihtsustada, tuues välja sõnad, mida rakendus ei asendanud. Märgitud sõnade puhul aga ei leidunud ühelgi juhul EstWN-i kirjet, seega EstWN-i täienedes peaks selliste probleemide esinemine vähenema. Märgitud vigade seas olid põhilised sellised, kus ülemmõistega asendades kaob ära oluline informatsioon, näide sellest on ’leetrid’, kus asenduseks on ’nakkushaigus’. Ise proovis tekste lihtsustada kolmveerand vastanutest, kellest 67% pidas lihtsustatud teksti lihtsamaks. Vead osutusid üldjoones samaks mis näidistekstide puhul, kuid leidus ka olukordi, kus asendused oli lause

kontekstis vale – ’intervallide’ oli asendatud ’ajavahemikkudega’, mis oli muusikaga seotud lause kontekstis vale. Üldisest tagasisidest järeldus, et rakenduse lävendi selgitus on ebaselge, sest paljudele jäi lävendi mõte arusaadamatuks. Sellele tuginedes kirjutati lävendi seletus ümber. Lisaks pakuti välja, et tekstikastid võiksid olla paralleelselt, mis aitaks paremini erinevusi tuvastada.

3.6 Edasiarendusvõimalused

Veebirakenduses on võimalik Java asemel kasutada täielikult Pythonit, kui asendada Sparkjava mingi muu raamistikuga, nagu Django või Flask. Selline asendus parandaks rakenduse jõudlust, sest ei ole vajadust kahe erineva keele vahelist suhtlust kasutada.

Mitmetähenduslike sõnade puhul saaks parandada esinemissageduse kaalu. Sagedusloendis on kõige suurema sagedusega tüüpiliselt sellised, millel on palju erinevaid tähendusi. Näiteks sõna ’olema’ jaoks leidub EstWN-is 9 tähendust, ’saama’ puhul 12 tähendust ning ’pidama’ jaoks 14 tähendust. Seega on sageduse põhjal eksitav valida asenduseks selline sõna, millel on palju tähendusi, sest ei saa olla kindel, millises tähenduses see sõna kõige rohkem esines. Probleemi lahendamisele võib kaasa aidata, kui eeltöötluse käigus lisada sagedusloendisse juurde uus veerg, kus on märgitud selle lemma erinevate tähenduste arv, mida hiljem algoritmis esinemissageduse leidmisel kasutada.

Näide: *’olema 638802’ asemel oleks vastav rida siis ’olema 638802 9’*.

Lisaks saaks lisada toe sõnaühenditele. Sagedusloend sisaldab ainult üksikuid lemmasid ja esinemissagedusi, seega ei ole selle abil võimalik hinnata sõnaühendite keerukust. Praegu lihtsustab veebirakendus sõnaühendite asemel üksikuid sõnu, mistõttu võib tulemus olla vigane. Võimalik lahendus on kasutusele võtta fraseoloogiasõnaraamat, kus on paljude eestikeelsete fraasidele toodud ühesõnalised vasted.

Näide: *aia taha minema → nurjuma*.

Lisaks sõnavaralisele lihtsustamisele võiks programm eelnevalt teha süntaktilist lihtsustamist, mille abil lihtsustatakse lause struktuuri. Tulemuseks oleks siis nii struktuurselt kui ka sõnavaraliselt lihtsam tekst. Osalausete tuvastamist on võimalik teha EstNLTK abil.

Lihtsa teksti üks omadusi on väiksem sõnavara. Selle saavutamiseks võiks asenduse valimisel eelistada sõna, mis antud tekstis juba esineb.

Teksti arusaamise seisukohalt oleks võimalik lisada keeruliste sõnade juurde võimalus kasutajal vaadata lisaks asendustele ka sõnade definitsiooni või seletust, mis leidub EstWN-is.

Muudetava lävendi asemel saaks kasutada kindlaks määratud konstanti, mille juures on lihtsustamine kõige tulemuslikum. Sellisel juhul ei peaks kasutaja enam katsetama erinevaid lävendi numbreid, et leida hea lihtsustamise tasakaal. Samuti ei oleks siis lävendist mittearusaamine kasutajale probleemiks.

Teiste optimeerimiste seas on algoritmi juures on võimalik parima sünohulga leidmiseks laiendada vaadeldavat konteksti.

4. Kokkuvõte

Käesolevas bakalaureusetöös uuriti teksti lihtsustamise meetodeid ning kirjeldati kolme inglise keele lihtsustamise veebirakendust. Välja toodud meetodid annavad ülevaate, millised on levinuimad teksti lihtsustamise viisid ning kuidas ja mida need lihtsustavad. Vaadeldud rakendused olid eeskujuks eestikeelse teksti lihtsustamise veebirakenduse loomisele.

Praktilise osa tulemusena loodi eestikeelset teksti lihtsustav veebirakendus, mis põhineb leksikaalse ehk sõnavaralise lihtsustamise meetodil. Veebirakenduses on omavahel koos kasutatud mitmeid keeleressursse. Eesti Wordnetiga leitakse sõnadele ülemmõisted ja sünonüümid, *word2vec'*i abil leitakse sõnadevahelised sarnasused. Sagedusloendi, võõrsõnade leksikoni ja põhisõnavara sõnastiku abil hinnatakse sõna keerukust. Asenduse sobivuseks kontrolliks koostati valem, mis leiab sõnade sarnasuse ja keerukuse põhjal hinnangu, mis arvestab nii suuri sagedusi kui ka väikeseid sarnasusi.

Tulemuste hindamiseks läbiviidud küsitluse tagasiside oli positiivne. Näidistekstide põhjal pidas lihtsustatud tekste tegelikult lihtsamaks 87% vastanutest. Võimalikeks edasiarendusteks on sõnaühendite tuvastamine ja lihtsustamine ning tekstide üldise sõnavara vähendamine.

5. Viidatud kirjandus

[1] Tammur, A. Kui palju räägitakse Eestis eesti keelt? Statistikaamet, 2017, https://statistikaamet.wordpress.com/2017/03/13/kui-palju-raagitakse-eestis-eesti-keelt/ (11.05.2017).

[2] Kruusvall, J. Keelteoskus ja keelte praktiline kasutamine. Eesti ühiskonna lõimumismonitooring 2015, 2015, http://www.kul.ee/sites/kulminn/files/6peatykk.pdf (11.05.2017).

[3] Siddharthan, A. A survey of research on text simplification. *The International Journal of Applied Linguistics*, 2014, pp 259–298.

[4] Dyslexia Basics. International Dyslexia Association. https://dyslexiaida.org/dyslexia-basics (11.05.2017).

[5] Ehala, M., Kerge, K., Lepajõe, K., Sõrmus, K. „Kõrgkoolide üliõpilaste eesti keele oskuse tase“ kordusuuring. Uuringu kokkuvõte. Tartu Ülikool. 2015.

[6] Saggion, H., Gómez Martínez, E., Etayo Gil, E., Anula Rebollo, A., Bourg, L. Text simplification in Simplext: making texts more accessible. *Procesamiento del Lenguaje Natural*, 2011, nr 47, pp. 341-342.

[7] Drndarević B., Štajner S., Bott S., Bautista S., Saggion H. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. *Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science*, 2013, nr 7817, pp. 488-500.

[8] Saggion, H., Bott, S., Rello, L. Comparing Resources for Spanish Lexical Simplification. *Statistical Language and Speech Processing. SLSP 2013. Lecture Notes in Computer Science*, 2013. nr 7978, pp. 236-247.

[9] Sanjay, S. P., Anand, K, Soman, K. P. AmritaCEN at SemEval-2016 Task 11: Complex Word Identification using Word Embedding. *Proceedings of SemEval*, 2016, pp. 1022–1027.

[10] Belder, J. D., and Moens, M.-F. Text simplification for children. *Proceedings of the SIGIR workshop on accessible search systems,* 2010, pp. 19–26.

[11] Müürisep, K. Eestikeelsete tekstide sisukokkuvõtjast EstSum. *Keel ja arvuti,* 2006, nr 6, lk 115-125.

[12] Jones, K. S. Automatic Summarising: Factors and Directions. *Advances in Automatic Text Summarization*, 1998, pp. 1-12.

[13] Specia, L. Translating from complex to simplified sentences. *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, 2010, nr 6001, pp. 30–39.

[14] Wang, T., Chen, P., Rochford J., Qiang J. Text simplification using neural machine translation. *In Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 4270-4271.

[15] Project Gutenberg's Frankenstein, by Mary Wollstonecraft Shelley. http://www.gutenberg.org/cache/epub/84/pg84.txt (11.05.2017).

[16] Rewordify.com. https://rewordify.com/ (11.05.2017).

[17] simplish. http://www.simplish.org/ (11.05.2017).

[18] Article Simplifier. http://seotoolzz.com/article-simplifier.php (11.05.2017).

[19] Eesti Keeleressursside Keskus. https://keeleressursid.ee/et/keeleressursid (11.05.2017).

[20] Riiklik programm. Eesti keeletehnoloogia 2011-2017. http://vana.keeletehnoloogia.ee/EKT2011-2017-programm-uuendet.pdf (11.05.2017).

[21] WordNet A lexical database for English. https://wordnet.princeton.edu/ (11.05.2017).

[22] Orav, H., Kerner, K., Parm, S. Eesti Wordnet'i hetkeseisust. *Keel ja Kirjandus*, 2011, nr 54, lk 96-106.

[23] Orav, H., Zupping, S., Vare, K. Leksikosemantiliste suhete hägusus Eesti Wordnetis. *Emakeele Seltsi Aastaraamat*, 2015, nr 60, lk 171–193.

[24] Estonian Wordnet (kb73-LAST). https://metashare.ut.ee/repository/browse/estonian-wordnet-kb73-last/31b5a794e6aa11e5a6e4005056b400244539990e22ec4bb9b3f585115ff1cb6b/ (11.05.2017).

[25] Eesti Wordnet'i täiendamine. Eesti keeletehnoloogia. https://www.keeletehnoloogia.ee/et/ekt-projektid/eesti-wordneti-taiendamine (11.05.2017).

[26] Zapata, A. A. Unit 1: Semantic Relationships. 2008. http://webdelprofesor.ula.ve/humanidades/azapata/materias/english_4/unit_1_semantic_relationships.pdf (11.05.2017).

[27] Orav, H. Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. Magistritöö. Tartu. 1998.

[28] Pajusalu, R. Sõna ja tähendus. Eesti Keele Sihtasutus. 2009.

[29] Eesti Wordnet. http://www.cl.ut.ee/ressursid/teksaurus/index.php (11.05.2017).

[30] Eesti keele käsiraamat. https://www.eki.ee/books/ekk09/index.php?p=6&p1=4 (11.05.2017).

[31] Vabavaraline Morfoloogiatarkvara. https://www.keeletehnoloogia.ee/et/ekt-projektid/vabavaraline-morfoloogiatarkvara (11.05.2017).

[32] EstNLTK. https://estnltk.github.io/ (11.05.2017).

[33] Sagedusloendid. https://keeleressursid.ee/et/keeleressursid-cl-ut/ressursid/83-article/clutee-lehed/256-sagedusloendid (11.05.2017).

[34] Eesti keele põhisõnavara sõnastik. http://www.eki.ee/dict/psv/ (11.05.2017).

[35] Võõrsõnade leksikon. http://www.eki.ee/dict/vsl/ (11.05.2017).

[36] Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv*, 2013, nr 1301.3781.

[37] Introduction to Word2Vec. https://deeplearning4j.org/word2vec (11.05.2017).

[38] EstNLTK word2vec mudel. https://github.com/estnltk/word2vec-models/blob/master/README.md (11.05.2017).

[39] Usage of JavaScript for websites. http://w3techs.com/technologies/details/cp-javascript/all/all (11.05.2017).

[40] jQuery. https://jquery.com/ (11.05.2017).

[41] ProgressBar.js. https://kimmobrunfeldt.github.io/progressbar.js/ (11.05.2017).

[42] Bootstrap. http://getbootstrap.com/ (11.05.2017).

[43] Sparkjava. http://sparkjava.com/ (11.05.2017).

[44] Thymeleaf. http://www.thymeleaf.org/ (11.05.2017).

[45] Deeplearning4j. https://deeplearning4j.org/ (11.05.2017).

Lisad

I. Küsimustik

Eestikeelse teksti lihtsustamine

Küsimustik on koostatud teksti lihtsustamise veebirakenduse tulemuste hindamiseks ja testimiseks. Veebirakendus asub aadressil http://prog.keeleressursid.ee:4567/ ning seal on võimalik lihtsustamist ise katsetada eestikeelsete tekstidega.

* Kohustuslik

Juhised

Üksteise all on välja toodud teksti lihtsustamise sisend ja tulemus. Palun kirjutage iga teksti juurde, kas lihtsustatud tekst tundub Teie jaoks lihtsam. Lisage juurde oma arvamus, mida oleks võinud teisiti lihtsustada või mis on Teie arust valesti lihtsustatud. Asendatud sõnad on välja toodud sümbolite "<" ja ">" vahel.

Sisestage oma vanus *

Teie vastus

Kas eesti keel on Teie emakeel? *

- ◯ Jah
- ◯ Ei

Tekst 1: Pealkiri 1

Päästjad manitsevad lõhkekehadest viivitamatult teada andma [1].

Tekst 1 lihtsustatult:

Päästjad <hoiatavad> <pommidest> <kohe> teada andma.

Kas lihtsustatud tekst 1 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 2: Pealkiri 2

Triennaali peapreemia pälvinud töö ühendab ilu ja funktsionaalsuse mõiste [2].

Tekst 2 lihtsustatult:

<Festivali> <peaauhinna> pälvinud töö ühendab ilu ja <otstarbekuse> mõiste.

Kas lihtsustatud tekst 2 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 3: Pealkiri 3

Rohkemat kui lihvitud liigutuste hukatuslik rutiin [3].

Tekst 3 lihtsustatult:

Rohkemat kui <kohendatud> liigutuste <hävitav> <harjumus>.

Kas lihtsustatud tekst 3 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 4: Ilukirjanduslik tekst

Kusagilt doodzhi pargaselt saadud ilmatu suures veneetsia laternas, mis rippus suure, tammepuuga vooderdatud eesruumi laes, loitis alles kolm gaasituld: nad paistsid peente siniste õilmelehtedena, mida raamis valge leek. Ta kustutas need, ja visanud kübara ning mantli lauale, läks läbi raamatukogu oma magamistoa poole, mis asus esimesel korral ja oli suur kaheksanurgeline ruum; oma vasttekkinud toredustundmuses oli ta ise lasknud selle dekoreerida ja oli ta katnud haruldaste renessansiaegsete gobeläänidega, mis olid leitud kusagilt Selby Royali katusekambrist [4:120].

Tekst 4 lihtsustault:

Kusagilt doodzhi <tööpaadilt> saadud <hirmus> suures veneetsia <lambis>, mis rippus suure, tammepuuga vooderdatud <esiku> laes, <säras> alles kolm gaasituld: nad paistsid peente siniste õilmelehtedena , mida <ümbritses> valge leek. Ta kustutas need, ja visanud kübara ning mantli lauale, läks läbi raamatukogu oma magamistoa poole, mis asus esimesel korral ja oli suur kaheksanurgeline ruum; oma vasttekkinud toredustundmuses oli ta ise lasknud selle <kaunistada> ja oli ta katnud haruldaste renessansiaegsete <piltvaipadega>, mis olid leitud kusagilt Selby Royali <kambrist>.

Kas lihtsustatud tekst 4 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 5: Arheoloogiline tekst

Liigi, mille teadlased ristisid baidatõriks (ladina keeles Baidabatyr clivosus), jäänused avastati Venemaalt Kemtšugi jõe äärsest liivakivipaljandist Krasnojarski oblastis. Leiukoht asub vaid viie kilomeetri kaugusel müstilisest dinosauruste nekropolist, mis on varem paleontoloogidele suurt huvi pakkunud [5].

Tekst 5 lihtsustatult:

Liigi, mille teadlased <nimetasid> baidatõriks (ladina keeles Baidabatyr clivosus), jäänused avastati Venemaalt Kemtšugi jõe äärsest liivakivipaljandist Krasnojarski <haldusüksuses>. <Kasvukoht> asub vaid viie kilomeetri kaugusel <salapärasest> <ürgsisalike> <surnuaiast>, mis on varem <teadlastele> suurt huvi pakkunud.

Kas lihtsustatud tekst 5 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 6: Meditsiiniline tekst

Viirusi esineb peamiselt Aasia ja Aafrika arengumaades. Turistidega levib haigus ka Euroopasse, kus tekib aeg-ajalt nakkuspuhanguid. Leetrite vastu vaktsineeritakse lapsi alates 1967. aastast. Haigus on kergesti nakkav, sageli kaasnevad tüsistustena hingamisteede bakteriaalsed infektsioonid, nagu pneumoonia, harvemini entsefaliit. Leetrid on laste surma sage põhjus arengumaades [5].

Tekst 6 lihtsustatult:

Viirusi esineb peamiselt Aasia ja Aafrika arengumaades. <Reisijatega> levib haigus ka Euroopasse, kus tekib aeg-ajalt nakkuspuhanguid. <Nakkushaiguste> vastu vaktsineeritakse lapsi alates 1967. aastast. Haigus on kergesti nakkav, sageli kaasnevad tüsistustena hingamisteede <pisikulised> <nakkused>, nagu <kopsupõletik>, harvmenini <ajupõletik>. <Nakkushaigused> on laste surma sage põhjus arengumaades.

Kas lihtsustatud tekst 6 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Tekst 7: Teaduslik tekst

Eesti Wordnet on leksikosemantiline andmebaas, mida koostatakse üldjoontes inglise Princetoni WordNeti põhimõtteid järgides. Eesti Wordneti loomist alustati aastal 1995 Tartu Ülikoolis ning praeguseks sisaldab see üle 72 000 mõiste (sh sõnu u 98 700) ja üle 230 000 semantilise suhte. Sõnaliikidelt koosneb Eesti Wordnet adjektiividest, substantiividest, verbidest ja adverbidest, mis iga sõnaliigi sees on koondatud paljudesse tähenduslikesse üksustesse ehk sünohulkadesse [7:171].

Tekst 7 lihtsusatult:

Eesti Wordnet on leksikosemantiline andmebaas, mida koostatakse üldjoontes inglise Princetoni WordNeti põhimõtteid järgides. Eesti Wordneti loomist alustati aastal 1995 Tartu Ülikoolis ning praeguseks sisaldab see üle 72 000 mõiste (sh sõnu u 98 700) ja üle 230 000 <tähendusliku> suhte. Sõnaliikidelt koosneb Eesti Wordnet <omadussõnadest>, <nimisõnadest>, <tegusõnadest> ja <määrsõnadest>, mis iga sõnaliigi sees on koondatud paljudesse tähenduslikesse üksustesse ehk sünohulkadesse.

Kas lihtsustatud tekst 7 on lihtsam? Mida oleks võinud teha teisiti? *

Teie vastus

Katsetused

Katsetage veebirakenduses (http://prog.keeleressursid.ee:4567/) lihtsustamist paari lausega, proovige sama teksti ka erinevate lävenditega. Suurema lävendi abil on võimalik suurendada sõnade hulka, mida programm üritab lihtsustada.

Kopeerige katsetatud laused siia (algtekst ja lävend)

Teie vastus

Andke tagasisidet saadud tulemustele (Kas asendused on õiged, loogilised, vigased?)

Teie vastus

Kui Teil on veebirakenduse kohta ettepanekuid (disain, kasutusmugavus, funktsionaalsus jm), siis kirjutage siia

Teie vastus

Viited

[1] Päästjad manitsevad lõhkekehadest viivitamatult teada andma. Postimees. http://www.postimees.ee/4066651 (05.05.2017)
[2] Triennaali peapreemia pälvinud töö ühendab ilu ja funktsionaalsuse mõiste. Postimees. http://kultuur.postimees.ee/4090115 (05.05.2017)
[3] Rohkemat kui lihvitud liigutuste hukatuslik rutiin. Postimees. http://kultuur.postimees.ee/4081385 (05.05.2017)
[4] Wilde, O. Dorian Gray portree. Readme. 2014
[5] Siberist avastati dinosaurused hävitanud katastroofi üle elanud ürgne «hamster». Postimees. http://atlas.postimees.ee/4093293 (05.05.2017)
[6] Vaktsiin kaitseb nakkuse eest, Vaktsineeritult on ohutum reisida. Postimees. http://www.postimees.ee/2485925 (05.05.2017)
[7] Orav, H., Zupping, S., Vare, K. Leksikosemantiliste suhete hägusus Eesti Wordnetis. Emakeele Seltsi Aastaraamat. 2015. Nr 60. Lk 171–193. http://kirj.ee/public/ESA/2014/esa_60_2014_171-194.pdf (05.05.2017)

SAADA ÄRA

Ärge saatke paroole kunagi Google'i vormide kaudu.

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Martin Peedosk**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **Eesti keele digitaalsete ressursside ja tehnoloogiate rakendamine teksti lihtsustamise programmis**, mille juhendajateks on Sven Aller ja Kadri Vare,
   1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **11.05.2017**