TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Allan Loo

Eestikeelsete ristsõnade generaator

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Aller, MSc

Tartu 2022

Eestikeelsete ristsõnade generaator

Lühikokkuvõte:

Bakalaureusetöö eesmärk oli luua eestikeelsete ristsõnade generaator, mis kasutab eesti keele keeleressursse ja kogub muu vajaliku info veebist. Tulemuseks saadud ristsõna on nii paberile trükitav kui ka lahendatav ja kontrollitav otse veebilehel. Genereerimisel kasutatavad andmed koguti eestikeelsest Vikipeediast, kasutades veebikoorimist (ingl *web scraping*), ja Eesti Wordnetist, kasutades selles olevad definitsioone ning EstNLTK morfoloogilist sünteesi. Ristsõna genereeritakse, kasutades kitsenduste rahuldamise meetodit. Töös kirjeldatakse andmete kogumise metoodikat, ristsõna genereerimist ja lõplikku veebirakendust ennast.

Võtmesõnad:

Keeleteadus, keeletehnoloogia, eesti keel, automaatne genereerimine, veebikoorimine, veebirakendus, React, EstNLTK, ristsõna.

CERCS:

P175 - Informaatika, süsteemiteooria.

Estonian crossword generator

Abstract:

The goal of this Bachelor's thesis was to create an Estonian crossword generator, that uses Estonian language resources and gathers the rest of the data from the web. The resulting crossword is printable to paper, solvable and checkable right on the webpage. The data used in the generation was gathered from the Estonian Wikipedia using web scraping and from the Estonian Wordnet using the definitions present in it and morphological synthesis. The crossword is generated using the constraint satisfaction method. In this thesis, data collection methods, crossword generation and the final web application are described.

Keywords:

Language science, language technology, Estonian language, automatic generation, web scraping, web application, React, EstNLTK, crossword puzzle.

CERCS:

P175 – Informatics, system theory

Sisukord

Sissejuhatus

Ristsõnad on pikalt omanud olulist rolli nii hariduses, meelelahutuses kui ka vaimse tervise valdkonnas. Ristsõnade lahendamine on üks aktiivse õppimise meetoditest. Aktiivne õppimine suurendab oluliselt õppija huvi, motivatsiooni ja autonoomiat, mis omakorda parandab õppija tulemusi võrreldes passiivse õppimisega [1]. Paljud inimesed harrastavad regulaarselt ristsõnade lahendamist. Näiteks juba üksnes mobiilirakendust The New York Times Crossword on alla laaditud üle miljoni korra[1]. Ristsõnade jätkuvat populaarsust näitab ka fakt, et paljud Eesti ajalehed ja ajakirjad (näiteks Postimees ja Kroonika) avaldavad endiselt regulaarselt ristsõnasid.

Ristsõnade lahendamise kasulikkus on eriti suur vanemaealistele, sest mitmed uuringud annavad põhjust uskuda, et ristsõnade lahendamisest tulenev aju stimulatsioon aitab aeglustada vaimset degradatsiooni [2,3].

Ristsõnade koostamine on aeganõudev protsess, mis ühtlasi nõuab koostajalt võrdlemisi suurt kompetentsi. Sellest tulenevalt on loogiline katsetada ristsõnade koostamise protsessi automatiseerimist. Leidub mitmeid veebirakendusi, mis genereerivad etteantud andmete põhjal ristsõna, aga nõuavad seetõttu ka kasutajalt koostamisandmete olemasolu. Näited sellistest veebirakendustest on The Teacher's Corner [4], Crossword labs [5], aga ka Eestis Tomozovi tehtud ristsõnade generaator [6]. Teistes keeleruumides leidub täisautomaatseid ristsõnade generaatoreid, näiteks üks esimestest oli Itaalias Rigutini jt poolt loodud ristsõnade generaator [7].

Täisautomaatseid ristsõnade generaatoreid eestikeelsel kujul ei eksisteeri, seega panustab loodud ristsõnade generaator eesti keeletehnoloogia arengusse.

Bakalaureusetöö eesmärk on luua eestikeelsete ristsõnade generaator, mis kasutab eesti keeleressursse ja kogub muu vajaliku info veebist. Samuti oli eesmärk teha programm selliselt, et lõppkasutaja ei peaks liiga kaua ootama ristsõna genereerimise taga. Tulemuseks saadud ristsõna oleks nii paberile trükitav kui ka lahendatav ja kontrollitav otse veebilehel. Teema valiti, sest töö autorile pakkus huvi veebirakenduse loomise protsess ja traditsiooniliste manuaalsete protsesside automatiseerimine. Genereerimisel kasutatavad andmed koguti eestikeelsest Vikipeediast, kasutades veebikoorimist (ingl *web scraping*), ja Eesti Wordnetist,

[1] The New York Times Crossword: https://play.google.com/store/apps/details?id=com.nytimes.crossword&hl=en&gl=US

kasutades selles olevaid definitsioone. Töös kirjeldatakse andmete kogumise metoodikat, ristsõna genereerimist ja lõplikku veebirakendust ennast.

Töö jaguneb kolmeks peatükiks. Esimeses kirjeldatakse ristsõnadega seonduvat tausta. Tuuakse esile ristsõnade olulisus hariduses ja eakate vaimses tervises. Samuti tuuakse välja näiteid olemasolevatest ristsõnade generaatoritest ja nende puudujääkidest ning erinevustest. Peatükk lõpetatakse ristsõnade genereerimise kui kitsenduste rahuldamise probleemi kirjeldusega. Teises peatükis tutvustatakse töös kasutatud olulisemaid tehnoloogiaid. Kolmandas peatükis antakse ülevaade valminud veebirakendusest, sealhulgas ülevaade arhitektuurist, andmete kogumisest, genereerimisalgoritmist, esinenud probleemidest, kogutud tagasisidest ja edasiarendusvõimalustest.

Mõisted ja terminid

API (ingl *application programming interface*) Reeglid ja vahendid rakendusprogrammi suhtluseks operatsioonisüsteemiga, andmebaasihalduse süsteemiga, muu juhtprogrammiga, sideprotokolliga. Rakendusliides määrab suhtlusvormingud ja sisaldab mooduli.[2]
Degradatsioon (ingl *degradation*) Organismi võimete vähenemine.[3]
Eesliides (ingl *frontend*) Veebilehele ilmuv kasutajaliides, mis võimaldab veebisaidi külastajal kahepoolselt suhelda saidi dünaamiliste osadega nagu andmebaasid, ostukorviprogrammid ja onlain-ostutöötlustarkvara.[4]
EstWN/Eesti Wordnet (ingl *Estonian Wordnet*) Omavahel viidetega ühendatud sünohulkade kogum.[5]
IEEE (ingl *Institute of Electrical and Electronics Engineers*) Elektri ja Elektroonika Inseneride Instituut.
Itereerima (ingl *iterate*) IT korduvate tsüklitena sooritama, sel viisil lahenduseni jõudma, kusjuures iga korduse puhul toetutakse eelneva seisu andmetele.[3]
Märgend (ingl *tag*) Hüpertekst-märgistuskeeltes nagu HTML, SGML, XML , XHTML jne, määravad teksti sisse kirjutatud märgendid ära selle, kuidas peavad brauseri poolt kuvatav tekst ja pildid arvutiekraanil välja nägema. Peale veebilehe väljanägemise ja selle elementide paigutuse kirjeldamise kasutatakse märgendeid ka linkimiseks, indekseerimiseks jms.[4]

[2] https://akit.cyber.ee/
[3] https://sonaveeb.ee
[4] http://www.vallaste.ee/

Raamistik (ingl *framework*)

Objektorienteeritud süsteemide puhul objektiklasside hulk, mis annab kasutajale või programmile omavahel seotud funktsioonide kollektsiooni.[4]

Rekursioon (ingl *recursion*)

Protsess, kus meetod kutsub iseennast ikka ja jälle, kuni mõni protsess on lõpule jõudnud või kuni saab täidetud mingi tingimus.[4]

REST (ingl *Representational state transfer*)

Hajustöötluse paradigma, eriti veebiteenuste programmeerimisel, mis[2]:

- taotleb veebi nüüdisnõuetele sobivat arhitektuuristiili ja ühtset liidest
- suurendab jõudlust, mastabeeritavust, muudetavust, nähtavust, porditavust, töökindlust
- lihtsustab liidest

Sünohulk (ingl *synonym set* või *synset)*

Wordneti tüüpi tesauruse põhiühik, mis koosneb kõigist ühte ja sama mõistet väljendavatest sõnadest (või sõnaühenditest).[5]

Tagaliides (ingl *backend*)

Teenus, mille ülesandeks on kaudselt toetada eesteenusi ning nad on tavaliselt lähemal vajalikule ressursile ja/või suudavad sellega suhelda.[4]

Tagurdus (ingl *backtracking*)

Lahendusmeetod alamülesannete lahendite katsetamise ja tagasivõtuga.[4]

Teek (ingl *library*)

Programmeerimises nimetatakse teegiks valmiskompileeritud alamprogramme, mida programm saab kasutada.[4]

[5] https://www.cl.ut.ee/ressursid/teksaurus/

Veebikoorimine (ingl *web scraping*)

Automaatprotsess andmete eraldamiseks veebilehtedelt, mis ei ole mõeldud olema masinloetavad, nagu pildid või formaaditud veebilehed.[2]

Veebirakendus (ingl *web application*)

Hajus rakendusprogramm, mis[2]:

- ei sõltu platvormist
- on klient-server-arhitektuuriga:
    - kliendiks on veebibrauser
    - serveriks on veebiserver
- andmeid hoitakse ja töödeldakse peamiselt serveril
- andmevahetus toimub võrgu kaudu

1 Ristsõnade kasutamine ja loomine

Käesolevas peatükis antakse ülevaade ristsõnade ühiskondlikust ja individuaalsest kasulikkusest. Peamiselt keskendutakse ristsõnadele kui õppevahendile ja vaimse tervise turgutajale. Samuti antakse peatükis ülevaade juba eksisteerivatest ristsõnade generaatoritest ja tuuakse välja nende puudusi. Viimasena kirjeldatakse ristsõnade genereerimist kui kitsenduste rahuldamise probleemi.

1.1 Ristsõnad hariduses

Ristsõnade lahendamine on paljudele meeldiv ajaveetmise viis, millega natuke testida oma teadmisi, aga kui kasulikud on ristsõnad õppevahendina? TALE (ingl *International Conference on Teaching, Assessment, and Learning for Engineering*) konverentsil esitatud uurimuses [8] analüüsisid Lottering jt ristsõnade efektiivsust õpilaste hindamisel. Nad õpetasid gruppi tudengeid kasutades traditsioonilisi õpetamise meetodeid, ning hindasid pooli neist kasutades traditsioonilist hindamismeetodit ja pooli kasutades ristsõnu. Lottering jt leidsid, et ristsõnadega hinnatud grupi tulemused olid oluliselt paremad teisest grupist. Journal of Food Science Education väljaandes on Mshayisa kirjutanud artikli sarnasel teemal kus on keskendutud ristsõnade kasulikkusele õppevahendina kogu õppeprotsessi vältel, mitte ainult lõpphinde moodustamisel. Uurimuses [9] avastati, et ristsõnad nõudsid õpilastelt kriitilist mõtlemist ja aitasid hinnata nende arusaama kursusel omandatud materjalist. Uuringu tulemused kinnitasid, et ristsõnad suurendasid õpilaste vahelist koostööd.

Mshayisa [9] uurimus tõi esile, kuidas mänguline lähenemine õppimisele ristsõnade näol aitas luua koostöörohke ja sõbraliku keskkonna õpilastele. Ristsõnade positiivset mõju õppekeskkonnale kinnitab ka Lotteringi jt [8] uurimus, kus tutvustati ristsõnu õppeprotsessis alles kõige lõpus, hindamisel, aga see ei osutunud õpilaste jaoks probleemiks, õpilased ei kogenud selle tõttu stressi ega ärevust.

Ristsõnasid on katsetatud hindamisvahendina meditsiini valdkonnas terminoloogia õpetamisel. Patrick jt [10] viisid läbi katse MBBS (ld *Medicinae Baccalaureus, Baccalaureus Chirurgiae*, arstikraadiga võrdväärne kraad) 2. kursuse 5. semestri tudengite seas, mille eesmärk oli uurida ristsõnade kasulikkust hindamisvahendina farmakoloogias. Õppimisel rakendati traditsioonilisi õppemeetodeid, ristsõnu rakendati ainult hindamisel. Tagasisidest selgus, et tudengid pidasid ristsõnu interaktiivseks ja meeldivaks hindamise viisiks. Samuti leidsid tudengid, et ristsõnad aitasid neil kinnistada erialast terminoloogiat ning motiveerisid neid varem ettevalmistumisega alustama. Kuna antud uuring ei võrrelnud käsitluse all olevat

hindamismeetodit traditsioonilise hindamismeetodiga, siis ei saanud tudengite tulemuste põhjal järeldusi teha. Seetõttu ei saa öelda, kas ristsõnade kasutamine hindamisvahendina on parem traditsioonistest hindamisvahenditest, aga tudengite positiivne tagasiside viitab selle võimalikkusele.

1.2 Ristsõnade kasulikkus eakate vaimsele tervisele

Ühiskonnas on levinud arusaam, et aju stimuleerimine aitab aeglustada vananemisega kaasnevaid vaimseid puudujääke. Seda hüpoteesi aitavad toetada mitmed uuringud. Näiteks Corki Ülikoolis läbiviidud uurimuses [2] analüüsisid Murphy jt kahe meetodi efektiivsust verbaalse tähesujuvuse (ingl PVF, *Phonemic verbal fluency*) võimekuse tugevdamisel vanemaealiste inimeste seas. Katses osalesid 37 vaimselt tervet 57–90 aasta vanust inimest. Katse sooritati läbi 4 nädala jooksul 2 grupis, kus üks grupp lahendas igapäevaselt ristsõna ja teine grupp täitis päevikut. Nad nägid oma tulemustest, et ristsõnade grupi PVF-i võimekus kasvas oluliselt rohkem võrreldes päeviku kirjutajate grupiga. Nende tulemused seega toetavad hüpoteesi, et ristsõnade lahendamise kaudu aju stimuleerimine toetab inimese vaimset tervist. Sarnaseid järeldusi tegi ka Pillai jt teadustöö [3], mis uuris mälu halvenemise ja ristsõnade lahendamise seost dementsuse all kannatavate inimeste seas. Töös jõuti seisukohale, et tulemused andsid küll põhjust uskuda ristsõnade lahendamise kasulikkusesse dementsuse ennetamisel, aga polnud piisavad, et järeldustes kindel olla.

1.3 Olemasolevad ristsõnade generaatorid ja nende puudused

Internetis leidub mitmeid ristsõnade genereerimist võimaldavaid veebirakendusi nagu Crossword Labs [5] ja The Teacher's Corner [4]. Sellised rakendused võtavad kasutajalt sisendiks küsimusi ja nende vastuseid ning loovad nende põhjal ristsõnu. Sellise meetodiga genereeritud ristsõnad on enamasti hõredalt kombineeritud ning nõuavad koostajalt võrdlemisi suurt kompetentsi ja aega. Teadaolevalt ei eksisteeri eestikeelset versiooni täiesti autonoomsest ristsõnade genereerimise süsteemist.

Heaks näiteks on itaalia keeles tehtud süsteem, mis on võimeline ilma inimese sisendita looma ristsõnu. See süsteem on Rigutini jt [7] tehtud ristsõnade generaator, mis kasutab veebikoorimist (ingl *web scraping*) ja loomuliku keele töötluse (ingl NLP, *natural language processing*) analüsaatorit. Generaatori definitsiooni eraldaja moodul on võimeline eeldefineeritud interneti lehekülgedelt eraldama toore teksti. See tekst läbib NLP analüsaatori, kus igale sõnale määratakse tema liik ja igale lausele selle liikmed: alus, öeldis, sihitis. Lausete osadest luuakse grupeeringud ja tehakse kindlaks, kas iga grupeering on subjekt, predikaat

või midagi muud. Lõpuks eraldatakse definitsioonid kategoriseeritud tekstist. Tulemuseks on suur hulk sõnu ja neile vastavaid definitsioone, mille seast valib ristsõna kompileerija moodul sobiva alamhulga, et moodustada ristsõna.

Rigutini jt töö testimisel kasutati itaaliakeelset Vikipeediat, kust suudeti eraldada 91 000 definitsiooni. Nendest 81% loeti korrektseks, 3% valeks ja 16% korrektseks, aga mitte kasutatavaks. Viimase grupi alla kuulusid definitsioonid, mis olid liiga pikad, sisaldasid iseendas vastust või olid liiga umbkaudsed [7].

Rigutini jt ristsõnade generaator on näide ühest kaugemale arendatud süsteemist. Enamik eksisteerivaid süsteeme nii võimekad ei ole. Ranaivo-Malancon jt [11] esitlesid 8. rahvusvahelisel Aasia infotehnoloogia konverentsil (ingl CITA, *International Conference on Information Technology in Asia)* enda esmast versiooni automaatsest ristsõnade generaatorist. Esitletud versiooni oli mitmeid piirangud. See ei olnud võimeline looma ristsõna ruudustikku ja oli mõeldud vaid Sarawaki ajaloo teemaliste ristsõnade genereerimiseks. Nende loodud süsteemi vastuste vihjed saavutasid 53% suuruse korrektsuse määra. Autorid mainisid, et tegu oli vaid projekti esmase versiooniga, mida oli plaanis veel oluliselt täiendada.

Ristsõnade genereerimise probleemiga on tegeletud ka Eestis. 2013. aastal kirjeldas Tomozov oma bakalaureusetöös ristsõnade genereerimist, kasutades kitsenduste rahuldamist ja libalõõmutamist (ingl *simulated annealing*). Töö keskendus suuresti teoreetilisele poolele ja varem mainitud kahe genereerimise meetodi võrdlemisele.

Tomozovi töö tulemusena valmis Java programm, mis on võimeline genereerima ruudustiku ja selle täitma sõnadega, kasutades ühte kahest meetodist: kitsenduste rahuldamist või libalõõmutamist. Realiseeritud generaatoris tuli sõnastikud kasutajal endal programmile ette anda .txt failina. Töös kogutud materjali kasutati Tehisintellekt I kursusel õppevahendina [6].

Selles töös loodi sarnane süsteem eelmainitutele, mis genereerib eestikeelseid ristsõnu, kogudes definitsioone eestikeelsetest allikatest. Peamine erinevus Tomozovi rakendusega [6] on töö programmiline andmete kogumine ja rakenduse kättesaadavus kasutajatele veebis.

1.4 Ristsõna genereerimine kui kitsenduste rahuldamise probleem

Ristsõnade genereerimine on oma olemuselt kitsenduste rahuldamise probleem (KR). KR on meetod keeruliste probleemide lahendamiseks, kus luuakse kitsendused, mille rahuldamise igal sammul jõutakse lõpuks aktsepteeritava lahenduseni [12]. Sedasi on võimalik oluliselt lihtsustada probleemide lahenduste implementatsioone.

Võime defineerida KR probleemi järgmiselt [13]:

- Muutujate hulk $\{x_1, x_2, .., x_n\}$.
- Igale muutujale x_i vastav piirkond D_i võimalikest väärtustest sellele muutujale.
- Kitsenduste hulk, mis kirjeldab seoseid muutujate ja väärtuste vahel.

Ristsõna genereerimisest võib mõelda kui probleemist, kus on vaja täita eeldefineeritud ruudustik mittekorduvate sõnadega, mis valitakse ühest kindlast sõnade hulgast [12].

KR probleeme liigitatakse muutujate järgi, mis saavad olla diskreetsed või pidevad ning nende väärtuste piirkonnad kas lõplikud või lõpmatud [14]. Ristsõnade genereerimine on näide diskreetsete muutujatega ja lõplike väärtuste piirkondadega KR probleemist. Seda sellepärast, et ristsõnasse on võimalik valida andmeid lõplikust hulgast andmetest.

Samuti liigitatakse eraldi kitsendusi [14]:

- Unaarsed kitsendused (ühe muutujaga): nt $X \neq 1$.
- Binaarsed kitsendused (kahe muutujaga): nt $X \neq Y$.
- Kõrgemat järku kitsendused.

Ristsõnade genereerimisel kõige tüüpilisemad on binaarsed kitsendused. Kui vaatleme iga ristsõna küsimuse vastust kui eraldi muutujat, siis ristsõna genereerimisel esineks binaarne kitsendus iga ristuva vastuse puhul. Ristuvatel vastustel peab ristumiskohal esinema sama sümbol. Näide unaarsest kitsendusest oleks otsitavate vastuste pikkus, kui ruudustik on enne paika pandud. Sellisel juhul peab ruudustikku valitav vastus olema kindla pikkusega.

KR probleemi lahendamiseks on kasutusel kolm meetodit [14]:

- Tagasipöördumisega otsing (ingl *backtracking search*).
- Edasivaatav otsing (ingl *forward checking*).
- Kitsenduste levitamine (ingl *constraint propagation*).

Tagasipöördumisega otsing pole ristsõnade loomisel kõige optimaalsem, sest muutujale seatud kitsendusi vaadeldakse alles siis, kui muutuja on juba väärtustatud. Sedasi tehakse suurel hulgal ebavajalike operatsioone.

Edasivaatav otsing ja kitsenduste levitamine mõlemad vaatlevad kitsendusi enne muutuja väärtustamise ja on seetõttu oluliselt paremad valikud KR probleemi lahendamiseks selles olukorras.

Järgevas peatükis tuuakse välja, millised tehnoloogilised vahendid muutsid loodud veebirakenduse võimalikuks.

2 Kasutatud tehnoloogilised vahendid

Praktilises osas kasutati mitmeid vahendeid, mis võimaldasid andmeid koguda, töödelda ja kasutajale esitleda. Selles peatükis antakse ülevaade nende seast olulisematest, mis muutsid selle töö võimalikuks. Peamised loetletud vahendid on Pythoni programmeerimiskeel ja selle teegid, mis leidsid töös kasutust. Lisaks on kirjeldatud keeletöötlusvahendit EstNLTK, mälusisest andmehoidlat (ingl *data store*) Redis ja kasutajaliidese loomisel kasutatud võimalusi.

2.1 Python

Python on avatud lähtekoodiga programmeerimiskeel, mis keskendub lihtsusele ja kasutajasõbralikkusele [15]. Töös on kasutatud Pythonit tagaliidese (ingl *backend*) loomisel. Pythoni kasuks otsustati, sest selles oli saadaval mitmeid teeke, mis soodustasid töö valmimist. Olulisemad neist on välja toodud selles peatükis.

2.2 Scrapy

Scrapy on avatud lähtekoodiga Pythoni teek, mille peamine kasutusala on veebikoorimine (ingl *web scraping*) [16]. Antud töös kasutati Scrapy veebikoorimise võimalusi, et koguda andmeid eestikeelsest Vikipeediast[6], mida hiljem kasutatakse ristsõnade genereerimisel. Scrapy võimaldab mugavalt kasutada veebilehel leiduvaid viiteid, et rekursiivselt liikuda edasi oma tööga järgmisele veebilehele ja seeläbi koguda andmeid suurelt hulgalt lehtedelt.

2.3 Pandas

Pandas on avatud lähtekoodiga Pythoni teek, mida peamiselt kasutatakse andmete analüüsiks ja töötluseks [17]. Pandas oli antud töös peamine andmetöötlusvahend, mida kasutati nii andmete puhastamiseks kui ka genereerimisel andmete välja valimiseks. Andmeid hoitakse kasutamise ajal Pandase Dataframe objektina. Pandase kasuks otsustati, sest Pandas on populaarne teek andmeteaduses ja võimas tööriist andmete töötlemisel.

2.4 EstNLTK

EstNLTK on avatud lähtekoodiga keeletöötlusvahend, mis võimaldab muuhulgas märgistada sõnu, lauseid ja paragrahve, teha morfoloogilist analüüsi, nimetuvastust jpm [18]. Töös kasutati EstNLTK võimalusi andmete kogumiseks ja puhastamiseks. Puhastamise juures on võtmetähtsusega morfoloogilise analüsaatori lemmatiseerimise võimekus ja n-grammide loomine. Töös kasutati EstNLTK versiooni 1.6.9.1b0.

[6] Eestikeelne Vikipeedia: https://et.wikipedia.org

Andmete kogumise juures kasutati EstNLTK Eesti Wordneti (EstWN) liidest. EstWN on andmebaas, mis sisaldab sünohulki (sama mõistet väljendavad sünonüümsed sõnad) ja nendevahelisi semantilisi ühendusi [19]. Paljude sünohulkade juures on välja toodud ka sellele vastavad definitsioonid, mida kasutatigi siinses töös andmete genereerimisel.

Täpsem selgitus andmete tekitamisest tuuakse välja andmete kogumise peatükis 3.2.

2.5 Redis

Redis on avatud lähtekoodiga mälusisene andmehoidla (ingl *data store*), mis leiab peamiselt kasutust andmebaasina, vahemäluna ja sõnumi vahendajana [20]. Selles töös leidis Redis kasutust andmebaasina. Redis hoiab mälusiseses andmebaasis töö tulemusi ja aitab vahendada infot eesliidese ja tagaliidese vahel.

Võtmetähtsusega on antud töös Redis Queue, mis on Pythoni teek ülesannete järjekorda panemiseks, et neid tagaplaanil töödelda [21]. Redis Queue võimaldab veebirakendusel teenindada kasutajaid ja samaaegselt tegeleda ristsõnade genereerimisega.

2.6 Flask

Flask on mikroveebiraamistik, mis võimaldab luua Pythonit kasutades veebiaplikatsioone [22]. Siinses töös on kasutatud Flaski REST (ingl *Representational state transfer*) API (ingl *application programming interface)* loomisel, et eesliides ja tagaliides saaks omavahel suhelda. Flaski kasuks otsustati, tuginedes selle populaarsusele ja kasutusmugavusele.

2.7 React

React on avatud lähtekoodiga JavaScripti teek kasutajaliideste loomiseks [23]. Reacti üks tugevusi on selle komponentide süsteem, mis võimaldab kasutajaliidese jagada väiksemateks taaskasutavateks tükkideks. Komponentide süsteemi kasutavad ära kasutajaliidese raamistikud, mis sisaldavad endas laia valikut valmiskomponente.

Võimalik oli kasutada mitmeid erinevad teeke, raamistikke ja platvorme kasutajaliidese loomiseks, näiteks Vue.js[7] või Angular[8]. Otsustati Reacti kasuks, sest see on üks kõige laiemalt levinud raamistikke[9] ja autoril on sellega eelmainitutest kõige rohkem kogemust.

[7] Vue.js koduleht: https://vuejs.org/
[8] Angular koduleht: https://angular.io/
[9] Statista veebiraamistike populaarsuse küsitluse tulemused: https://www.statista.com/statistics/1124699/worldwide-developer-survey-most-used-frameworks-web/

Antud töös on kasutatud PrimeReacti[10] kasutajaliidese komponendi teeki. Teek valiti, sest sisaldab kvaliteetseid, kasutajasõbralikke ja moodsa disainiga komponente.

Järgmistes alapeatükkides kirjeldatakse, kuidas eelmainitud tehnoloogilisi lahendusi on veebirakenduses kasutatud.

[10] PrimeReact kasutajaliidese komponendi teegi koduleht: https://www.primefaces.org/primereact/

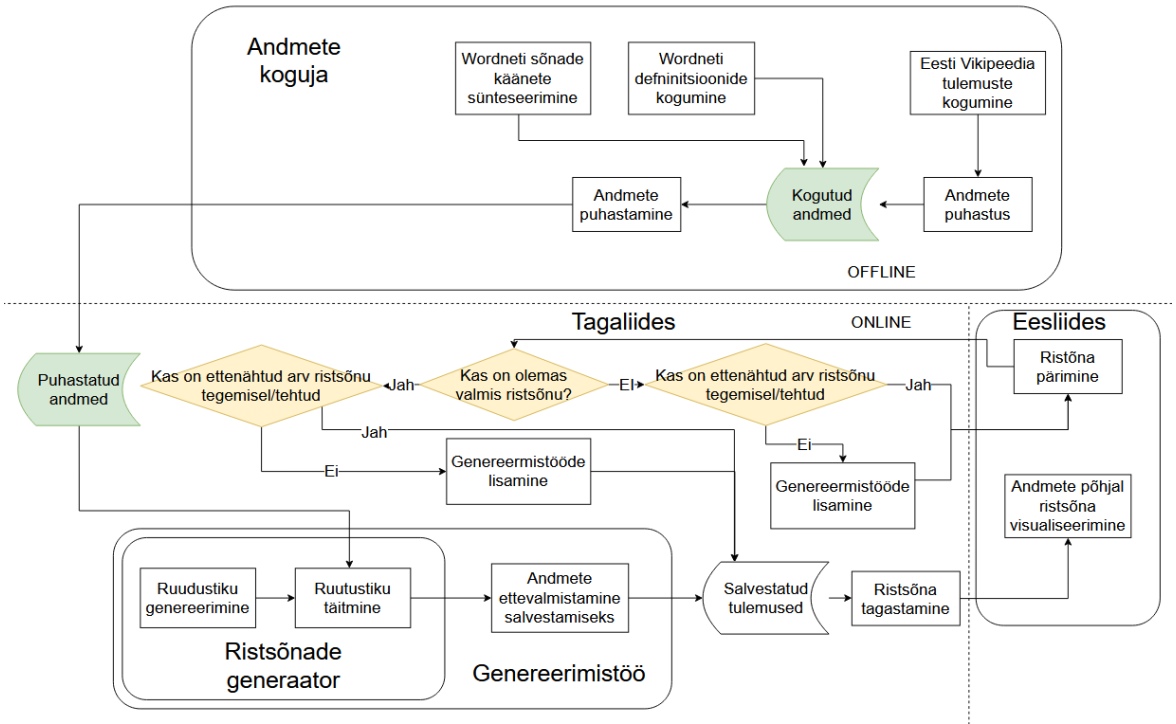
3 Veebirakendus

Antud peatükis antakse ülevaade töö käigus valminud veebirakendusest. Eraldi käsitletakse arhitektuuri, andmete kogumist, ristsõna genereerimist, andmete haldamist ja eesliidest. Veebirakendus on kättesaadav veebiaadressilt https://ristsonade-generaator.keeleressursid.ee/ ning lähtekood on leitav aadressilt https://github.com/VainLoo/Eestikeelsete-Ristsonade-Generaator.

3.1 Veebirakenduse arhitektuur

Selles peatükis antakse ülevaade veebirakenduse erinevatest komponentidest.

Loodud veebirakendust võib vaadelda kolme eraldi osana: andmete koguja, ristsõnade generaator ning eesliides (joonis 1). Andmete kogutakse varem kokku ja puhastatakse. Selle tulemusel valminud andmestik lisatakse tagaliidesesse, kus generaator seda kasutama hakkab.



Joonis 1. Vooskeem veebirakenduse loogikast, rohelisega on tähistatud lõplikke andmeid, nooltega on tähistatud andmete liikumine.

Kogumisele järgnev tegevus toimub tagaliideses. Generaator loob ristsõna ruudustiku ning seejärel võtab kogutud andmed ja kasutab neid, et täita see ruudustik andmetega. Sedasi genereeritakse määratud arv ristsõnu.

Kui eesliidesest tuleb päring ristsõna saatmiseks, saadetakse üks juhuslik genereeritud ristsõnadest päringu esitajale. Eesliideses võetakse ristsõna vastu ja esitatakse kasutajale. Iga kord, kui tagaliidese poole pöördutakse päringuga, kontrollitakse piisava arvu ristsõnade valmis ja tegemisel olemist. Vajadusel lisatakse juurde genereerimistöid tööde järjekorda, kust töölised (ingl *worker*) neid järjest töötlemisse võtavad.

Järgmistes peatükkides antakse põhjalikum ülevaade varem mainitud kolmest osast: andmete kogujast, ristsõnade generaatorist ning eesliidesest. Eraldi kirjeldatakse veel andmete haldamist.

3.2 Andmete kogumine

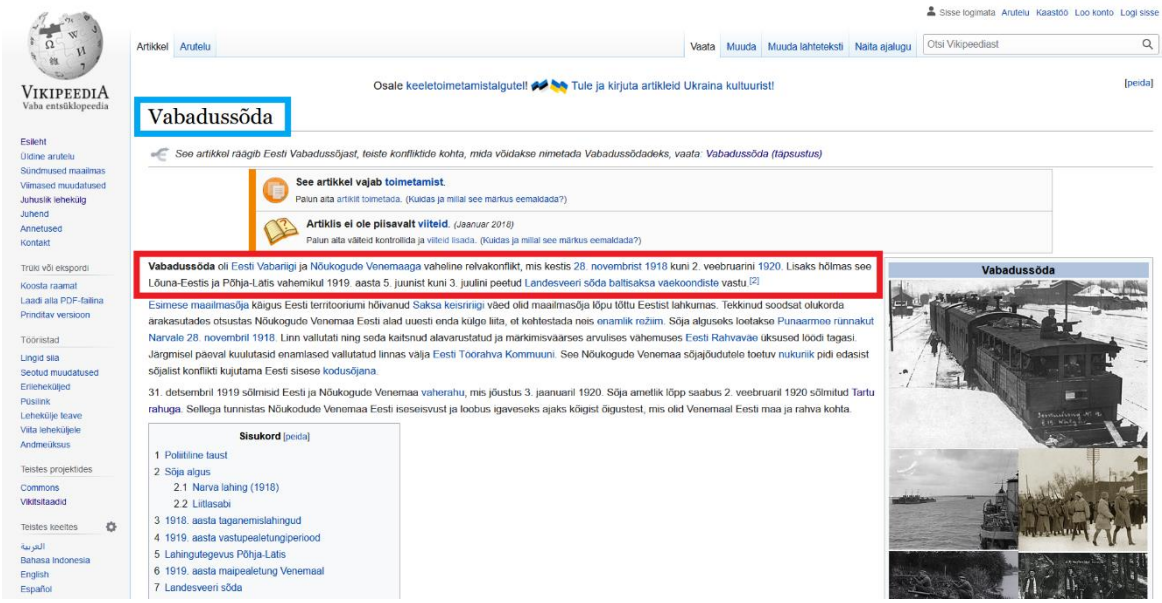
Ristsõnade genereerimiseks on vaja andmeid, millega neid täita. Vajalikeks andmeteks on sõnede paarid, kus üks on vastus ja teine on definitsioon. Antud töös koguti andmed programmiliselt kahest allikast: Vikipeediast ja EstWN-st.

3.2.1 Veebikoorimine Vikipeediast

Vikipeediast andmete kogumiseks kasutati veebikoorimist, mis on automaatprotsess andmete eraldamiseks veebilehtedelt, mis ei ole mõeldud olema masinloetavad, nagu pildid või formaaditud veebilehed. Selle saavutamiseks kasutati Scrapy Pythoni teeki. Lisaks kasutati Beautiful Soup[11] teeki, et eraldada leheküljelt vajalikud andmed.

Andmete kogumiseks tuvastati kindlad märgendid lehel, milles asusid otsitavad andmed (joonis 2). Andmeid loeti ainult lehtedelt, kus tuvastati definitsioon, mille algusest leiti regulaaravaldisega vastuse olemasolu. Seejärel eraldati definitsiooniks sõne ilma alguses oleva vastuseta. Leitud sõnede paare väljastati jooksvalt eraldi faili.

[11] Beautiful Soup dokumentatsioon: https://beautiful-soup-4.readthedocs.io/en/latest/



Joonis 2. Näide Vikipeedia lehelt kogutavastest andmetest, kus sinises ruudus on vastus ja punases ruudus on sellele vastav definitsioon[12].

Näiteks joonis 2 puhul tuvastati vikiartiklis, et sõna „Vabadussõda“ esineb leitud definitsioonis alguses ja eraldati vastus definitsioonist koos sellele järgneva sobiva tegusõnaga. Definitsioonina jäi alles „Eesti Vabariigi ja Nõukogude Venemaaga vaheline relvakonflikt, mis kestis 28. novembrist 1918 kuni 2. veebruarini 1920. Lisaks hõlmas see Lõuna-Eestis ja Põhja-Lätis vahemikul 1919. aasta 5. juunist kuni 3. juulini peetud Landesveeri sõda baltisaksa väekoondiste vastu“.

Järgnevalt valib Scrapy leheküljе sisu osast järgmise juhusliku hüperlingi, kus oma tegevust jätkata. Hüperlingi veebiaadress peab olema tingimustele vastav, mis garanteerib, et sihtlehekülg on Vikipeedia artikkel, kus on lootust leida sobivaid andmeid. Juhul, kui leheküljelt ei leita sobivaid andmeid, liigutakse edasi järgmisele lingile.

Veel üks lähenemine oleks olnud kasutada Vikipeedia andmetõmmist (ingl *data dump*) ja selle seest koguda andmeid. Veebikoorimise kasuks otsustati, sest sedasi oli võimalik saada kõige värskemad andmed.

3.2.2 Eesti Wordnet

Teeki EstNLTK on sisse ehitatud mugavad tööriistad, mis võimaldavad EstWNist kergesti andmeid koguda. Andmeid koguti kahe meetodiga: salvestatud definitsioone kogumine ja käänete sünteesimine.

[12] Näites toodud Vikipeedia leht: https://et.wikipedia.org/wiki/Vabaduss%C3%B5da

Definitsioone koguti itereerides üle kõikide sünohulkade. Iga sünohulga puhul kontrolliti definitsiooni olemasolu, mis on vajalik, sest definitsiooni olemasolu pole garanteeritud. Seejärel seati definitsiooni olemasolu korral vastav definitsioon paari iga selle sünohulga sõna algvormiga.

Näiteks kui on sünohulk, mis koosneb sõnadest „kass“, „kodukass“ ja „kiisu“ ning omab definitsiooni „kaslaste hulka kuuluv koduloom“, siis teeb programm nendest kolm paari:

- kass – kaslaste hulka kuuluv koduloom
- kodukass – kaslaste hulka kuuluv koduloom
- kiisu – kaslaste hulka kuuluv koduloom

Teine meetod Eesti Wordnetist andmete kogumiseks oli käändeliste vormide sünteesimine, kasutades morfoloogilist sünteesi. Morfoloogiline süntees on morfoloogilise analüüsi tagurpidine protsess, mis võimaldab sõna algvormist genereerida selle sõna käändeid [18]. Selles töös sünteesiti käändelised vormid ainult Eesti Wordneti nimisõnadest. Omadussõnade käänete ja tegusõnade pöörete lisamine oleks olnud samuti valik, aga see oleks tekitanud liigsel hulgal käänamise/pööramise tüüpi andmeid ja seeläbi aeglustanud sisse lugemise ja andmetest otsimise protsessi.

Eesti Wordnetis käidi tsükliliselt üle kõikide sünohulkade nimisõnade algvormide. Igale algvormile sünteesiti kõik eestikeelsed käändelised vormid (ainsuses ja mitmuses) välja arvatud nimetav, sest tegu on algvormiga ja kasutatakse küsimuses endas. Igale käänatud sõnale seatakse definitsiooniks „Sõna <käänatud sõna> käändes <ainsuse/mitmuse> <kääne>“.

Näiteks, sünohulk, mis koosneb sõnadest „kass“, „kodukass“ ja „kiisu“, ning omab definitsiooni „kaslaste hulka kuuluv koduloom“, siis teeb programm sõnast „kass“ järgmised ainsuse paarid:

- kassi – Sõna kass käändes ainsuse omastav
- kassi – Sõna kass käändes ainsuse osastav
- kassisse – Sõna kass käändes ainsuse sisseütlev
- kassis – Sõna kass käändes ainsuse seesütlev
- kassist – Sõna kass käändes ainsuse seestütlev
- kassile – Sõna kass käändes ainsuse alaleütlev
- kassil – Sõna kass käändes ainsuse alalütlev

- kassilt – Sõna kass käändes ainsuse alaltütlev
- kassiks – Sõna kass käändes ainsuse saav
- kassini – Sõna kass käändes ainsuse rajav
- kassina – Sõna kass käändes ainsuse olev
- kassita – Sõna kass käändes ainsuse ilmaütlev
- kassiga – Sõna kass käändes ainsuse kaasaütlev

Kui sõnale leitakse rohkem kui üks otsitava käände kuju, lisatakse andmestikku kõik. Näiteks, kui sõna kass mitmuse osastavas käändes võib olla kas „kasse“ või „kassisid“, siis andmestikku lisataks järgmised paarid:

- kasse – Sõna kass käändes mitmuse osastav
- kassisid – Sõna kass käändes mitmuse osastav

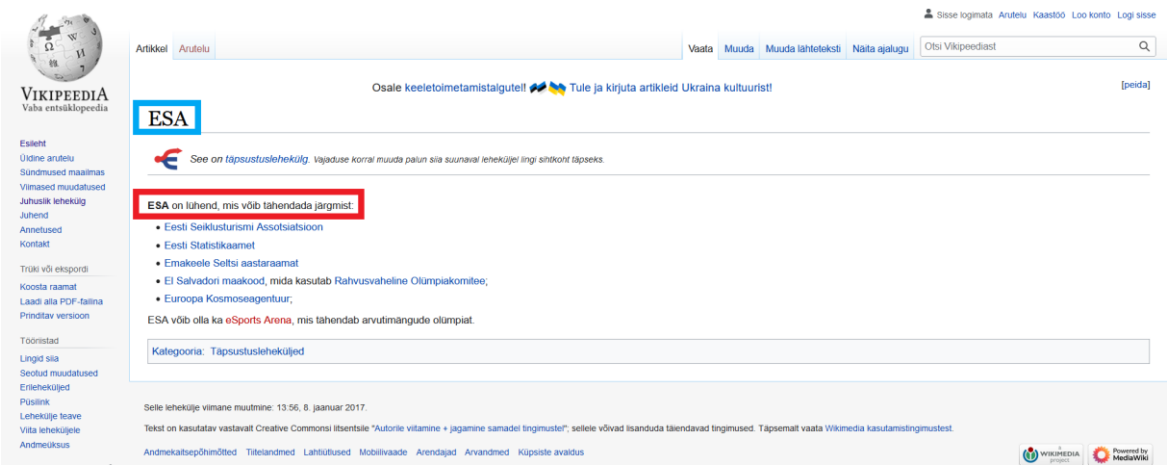
Kirjeldatud meetoditega sai kogutud piisavalt andmeid, et alustada ristsõnade genereerimisega, aga enne seda oli vajalik andmeid puhastada, et suurendada ristsõnades esinevate küsimuste ja vastuste kvaliteeti ning eemaldada potentsiaalseid veakohti süsteemist.

3.2.3 Andmete puhastamine

Erinevatest allikatest kogutud andmetes esinesid erinevad probleemid, mida pidi lahendama enne, kui andmeid sai kasutusele võtta. Siin osas kirjeldatakse andmetega tehtud toiminguid, millega suurendati nende kvaliteeti.

Vikipeediast kogutud andmed sisaldasid endas kõige rohkem vigasid. Andmetes esines vigu definitsiooni väljas, kus võis esineda puuduvaid väärtuseid, üksikuid tähemärke ja tühjasid sõnesid. Sellised vigased read eemaldati, kasutades Pandas teegi vahendeid.

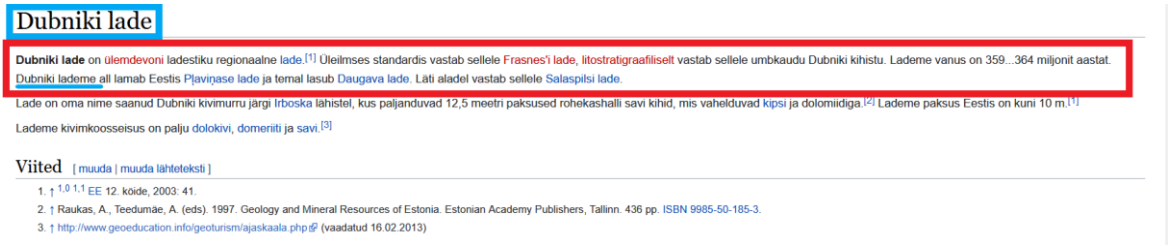
Vikipeediast kogutud andmetes esines vigaseid andmeid, mis olid seotud lühenditega. Lühendite puhul esineb definitsiooni sisaldavas märgendis loetelu kirjeldus, mis ei sobi definitsiooniks. Joonis 3 puhul kogutaks leheküljelt andmepaar „ESA – lühend, mis võib tähendada järgmist:“, mis kindlasti ei sobiks ristsõna andmeteks.



Joonis 3. Näide Vikipeedia lehelt[13] kogutavastest andmetest, kus sinises ruudus on vastus ja punases ruudus on sellele vastav definitsioon.

Kuna kogutud informatsioonis puudusid piisavad andmed, et parandada vigaseid ridu, siis need read eemaldati. Valdav enamus sellistest vigastest andmetest tuvastati ja eemaldati.

Korduv viga Vikipeediast kogutud andmetes on vastuse esinemine definitsioonis endas (joonis 4). Selle probleemi lahendamiseks on töös kasutatud EstNLTK morfoloogilise analüüsi vahendeid nagu *layer rolling* ja lemmatiseerimine.



Joonis 4. Näide Vikipeedia lehelt[14] kogutavatest andmetest, kus sinises ruudus on vastus ja punases ruudus on sellele vastav definitsioon. Sinisega on alla joonitud vastuse esinemine definitsioonis endas.

Layer rolling võimaldab üle definitsiooni itereerida n-grammide kaupa ja iga n-grammi puhul kontrollida, et selle algvorm ei ole sama mis vastus. Juhul, kui on, siis asendatakse iga täht alakriipsuga. Selline lähenemine võimaldab enamus juhtudel korrektselt asendada ka definitsioonis vastused, mis esinevad tüve muutusega käändes. Samuti võimaldab see korrektselt eemaldata definitsioonist vastused, mis koosnevad mitmest sõnast. Näites joonis 4 esineb sõnasid „Dubnik“ ja „lade“ definitsioonis eraldi seisvalt, aga puhastamisel neid ei

[13] Näites toodud Vikipeedia leht: https://et.wikipedia.org/wiki/ESA
[14] Näites toodud Vikipeedia leht: https://et.wikipedia.org/wiki/Dubniki_lade

eemaldata, eemaldatakse vaid „Dubniki lademe“. Seda sellepärast, et sealt leitakse n-gramm, mille sõnade algvormid on võrdsed lahenduse omadega. Kirjeldatud lähenemine ei tööta igal juhul, sest morfoloogiline analüüs ei tuvasta alati korrektset algvormi.

Eelnev andmete puhastus toimus ainult Vikipeediast kogutud andmete peal. Järgmised protseduurid viidi läbi kõikide kogutud andmete peal.

Andmetes võis esineda sümboleid, mis ei kuulu eesti tähestikku ja seega pole kasutaja jaoks klaviatuurilt sisestatavad. Seetõttu eemaldati kõik sellised andmeread. Kuna ristsõna maksimaalseks suuruseks on seatud 12x12, siis saab ristsõna maksimaalne sõna pikkus olla kuni 12 tähemärki. Seetõttu eemaldati kõik pikemad vastused andmestikust. Andmestikust eemaldati ka kõik read, mis sisaldasid puuduvaid väärtuseid või tühjasid sõnesid.

Lõpuks kasutusele võetud andmestik sisaldas endas 851 938 vastuse ja definitsiooni paari, millest enamuse moodustasid sünteesitud käänetest moodustatud tulemused. Vältimaks liigset käänamise küsimuste kasutust, on kasutatav andmestik jagatud kaheks: põhiandmestik ja lisaandmestik. Vikipeedia andmed ja EstWNi definitsiooni andmed moodustasid põhiandmestiku ja sünteesitud käändeküsimused lisaandmestiku. Põhiandmestikus esines 168 384 vastuse ja definitsiooni paari ning lisaandmestikus 683 554. Ristsõna loomisel on prioriteetsem kasutada põhiandmestikku, et tagada lõppkasutajale huvitavam ristsõna.

Järgmises peatükis on täpsemalt kirjeldatud ristsõna genereerimise protsessi.

3.3 Ristsõna genereerimine

Selles peatükis antakse ülevaade, kuidas ristsõnade genereerimise probleem on antud töös lahendatud.

Töös saab ristsõna genereerimise jagada üldjoontes kolmeks osaks: ruudustiku genereerimine, ruudustikus sõnaväljade tuvastamine ja ristsõna täitmine.

3.3.1 Ruudustiku genereerimine

Ruudustiku genereerimine on üks viis, kuidas luua peatükis 1.4 mainitud kitsendusi. Sedasi peab terve ristsõna koosnema kindlast arvust sõnedest, kus igal sõnel on oma kindel pikkus ja kindlad ristumiskohad teiste sõnedega.

Ruudustiku genereerimisel tuleb veenduda, et ruudustik vastaks ristsõna struktuuri nõuetele. Selles töös on genereerimisel ristsõnadel järgmised nõuded:

- Ruudustiku kõik lahenduste väljad peavad olema omavahel seotud.
- Ristsõnas kõik sõned on vähemalt pikkusega 3.

Esineb olukordi, kus genereeritakse ruudustik, mis ei vasta eelmainitud nõuetele. Sellise olukorra tuvastamisel taasalustatakse ruudustiku genereerimise protsessi.

3.3.2 Ruudustikus sõnaväljade tuvastamine

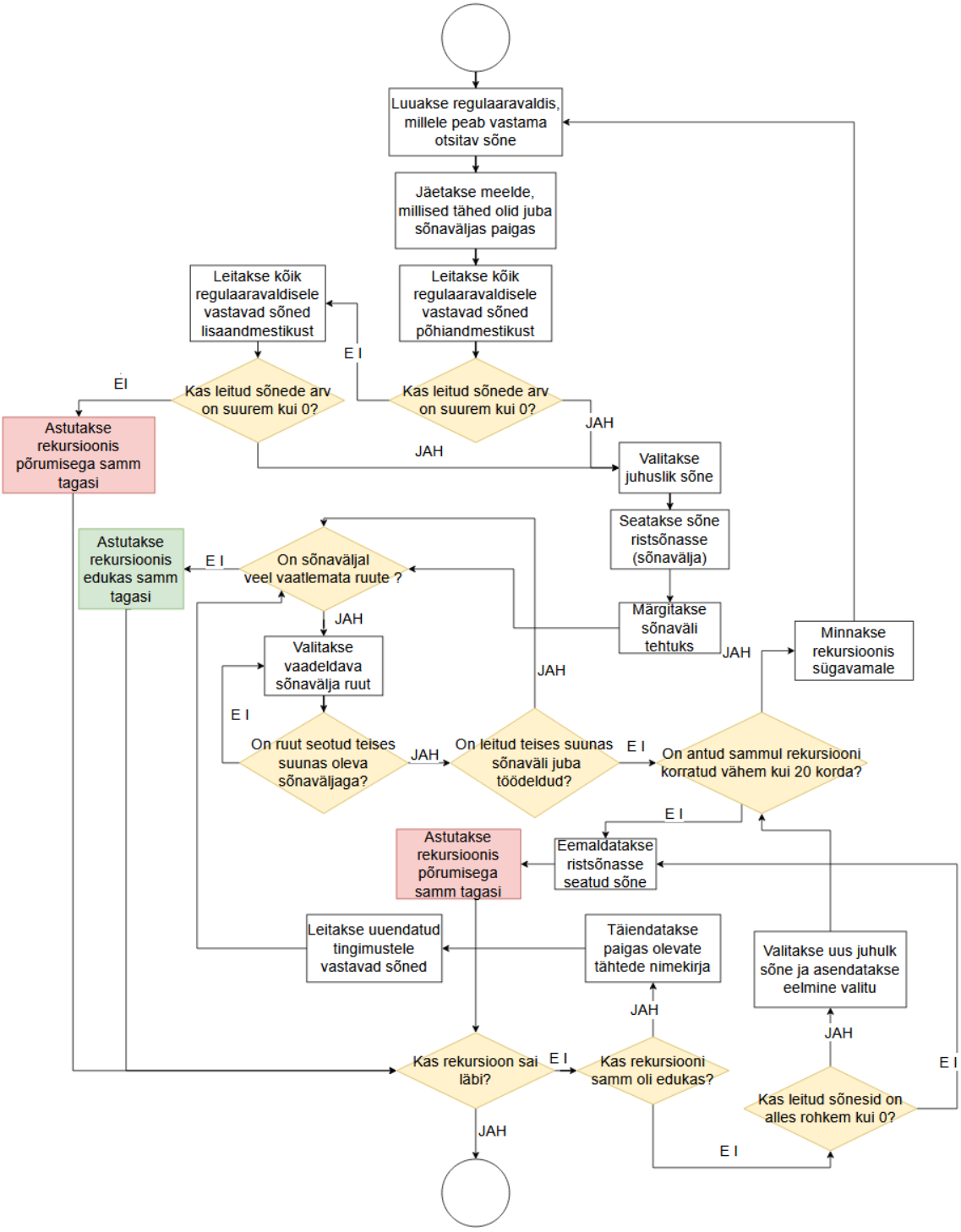
Selles töös luuakse muutuja iga täitmist vajava sõne jaoks ruudustikus, edaspidi nimetatakse neid muutujaid sõnaväljadeks. Kitsendused, mis peavad olema rahuldatud, on järgmised:

- Vastuste pikkus peab vastama sõnavälja pikkusele.
- Ristuvate sõnaväljade ristumiskohad peavad sisaldama sama sümbolit.
- Ükski lisatav vastuse ja definitsiooni paar ei tohi ristsõnas korduda.
- Kõik sõnaväljad peavad saama täidetud.

Kui ruudustik on genereeritud, tuvastatakse kõik sõnaväljad, millest ruudustik koosneb. See tähendab, et itereeritakse üle ruudustiku ja tuvastatakse iga seal asetseva sõnavälja pikkus, suund ning täpne asetus ruudustikus. Iga ruudustiku ruudu puhul märgitakse ära, kas see asetseb ülalt-alla sõnaväljas, paremalt-vasakule sõnaväljas või mõlemas. Selline sidusus sõnavälja ja sellega seotud ruutude vahel muudab ristsõna täitmise protsessi lihtsamaks.

3.3.3 Ristsõna täitmine

Ristsõna täitmine sõnedega toimub rekursiivselt, kasutades kitsenduste rahuldamise edasivaatava otsingu meetodit. Algoritmi aitab kirjeldada joonis 5.



Joonis 5. Vooskeem ruudustiku täitmise algoritmist, kus otsus on tähistatud kollasega, edukas rekursiooni sammu tagastus rohelisega ja edutu punasega.

Algoritm külastab rekursiivselt sõnaväljasid üritades kitsendusi rahuldades täita neid sõnedega. Iga sõnavälja puhul külastatakse rekursiivselt iga selle sõnaväljaga seotud teist sõnavälja. See loogika seletab ka, miks on üks ruudustiku genereerimise nõue kõikide sõnaväljade seotus üksteisega. Vastasel juhul ei oleks võimalik rekursioonil jõuda iga sõnaväljani.

On siiski olukordi, kus rekursioon lõpetab töö enne, kui on täitnud kogu ristsõna. Selline olukord tuvastatakse tagastusväärtuse põhjal ja seejärel genereerimise protsess käivitatakse uuesti.

Algoritm kasutab tagurdust (ingl *backtracking*), et optimeerida genereerimise protsessi. Jõudes olukorda, kus sõnaväljale ei leidu ühtegi tingimustele vastavat vastust andmestikust, astutakse rekursioonis samm tagasi ja üritatakse asendada eelnevat täidetud sõnavälja teistsuguse vastusega. Kuna iga järgnev valitud sõnaväli on vaadeldava sõnaväljaga ristuv, saab kindel olla, et eelneva sõnavälja asendamine mõjutab tupikusse jooksnud sõnavälja.

Eelnev annab ülevaate, kuidas ristsõnad lõpuks valmivad, aga selleks, et kasutaja saaks ristsõna ka lahendada, on tarvilik esmalt see temani toimetada. Seda protsessi kirjeldatakse järgmises peatükis.

3.4 Andmete haldamine

Järgneval antakse kõrgema taseme ülevaade veebirakenduse osade vahelisest kommunikatsioonist.

Ristsõnade genereerimine on potentsiaalselt palju aega nõudev protsess. Selleks, et lõppkasutaja ei peaks liiga kaua ootama ristsõna genereerimise taga, on veebirakenduses kasutusel Redis ja Redis Queue. Veebirakendus pidevalt töötab selle nimel, et määratud arv ristsõnu oleks Redise andmebaasi salvestatud. Kasutatakse Redise töölisi (ingl *worker*), et samaaegselt genereerida mitut ristsõna. Samuti kasutatakse Redis Queue võimalusi, et vajalik arv ristsõnade genereerimise töid seada järjekorda, kust töölised neid protsessimiseks võtta saavad.

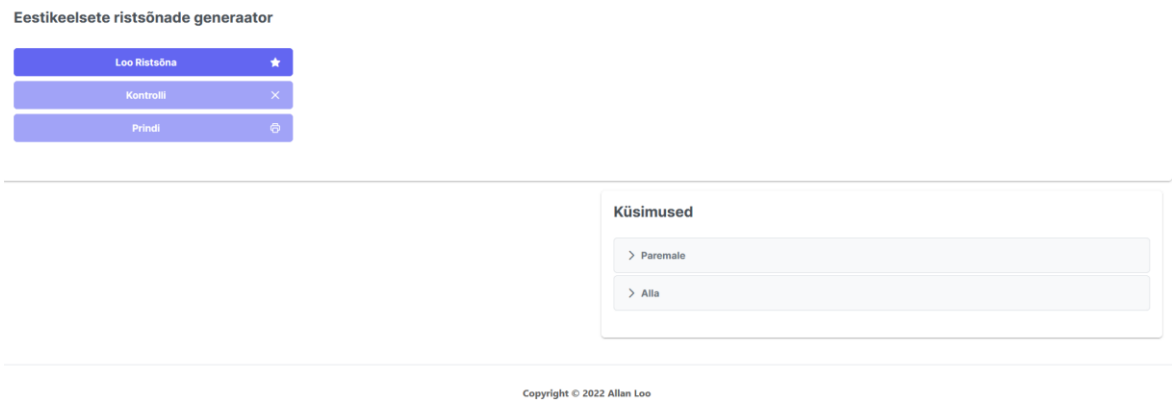
Eesliides suhtleb tagaliidesega kasutades Flask veebiraamistikku. Eesliideselt saabub päring, kui kasutaja soovib endale ristsõna. Päring võetakse vastu ja üritatakse andmebaasist tagastada juhuslik valmis genereeritud ristsõna. Kui leitakse sobiv ristsõna, siis see tagastatakse eesliidesele ja kustutatakse andmebaasist. Sedasi saab kasutaja igal päringul alati unikaalse ristsõna.

Kui andmebaasis ei leidu valmis ristsõna, siis tagastatakse eesliidesele vastus, mille põhjal eesliides teab, et valmis ristsõnu ei leidunud. Sellises olukorras jätkab eesliides automaatselt regulaarselt päringute saatmist, kuni valmis ristsõna tagastatakse. Sellist olukorda, kus ühtegi valmis ristsõna pole andmebaasis, ei tohiks kasutaja kogeda, sest süsteem üritab alati hoida andmebaasis seadistatud hulga ristsõnu.

3.5 Eesliides

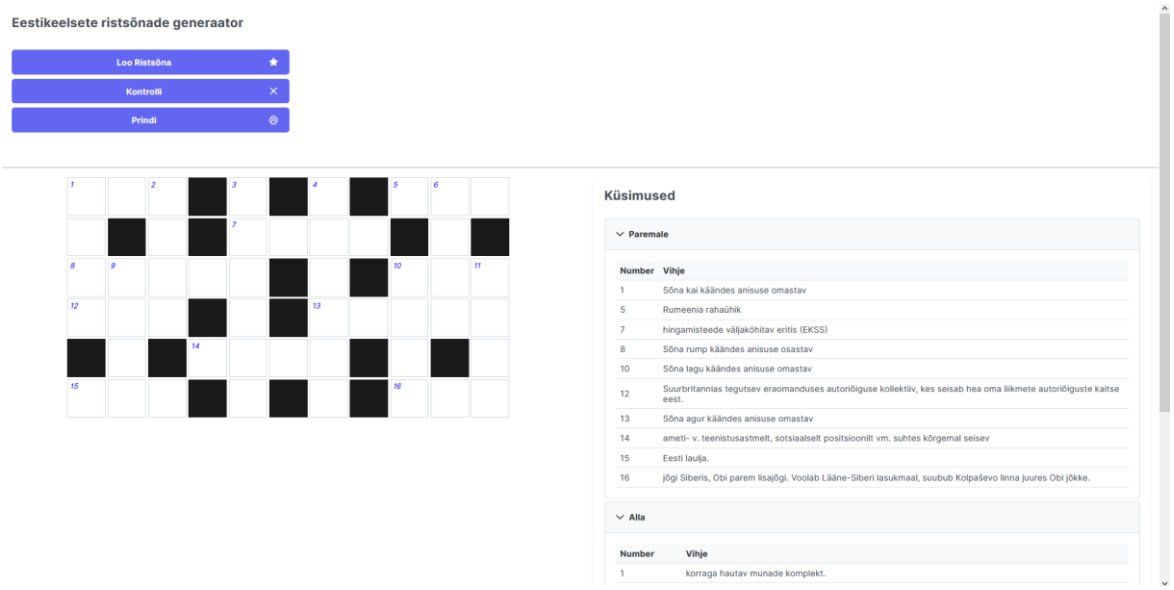
Selles peatükis antakse ülevaade veebirakenduse kasutajaliidesest ja selle funktsionaalsustest.

Eesliidese loomisel kasutati React teeki koos PrimeReact kasutajaliidese raamistikuga. Kasutades neid vahendeid, loodi heledates toonides minimalistlik kasutajaliides (joonis 6). Avalehel on kasutajal võimalik endale koostada ristsõna, klikkides nupule „Loo Ristsõna“.



Joonis 6. Veebirakenduse avaleht.

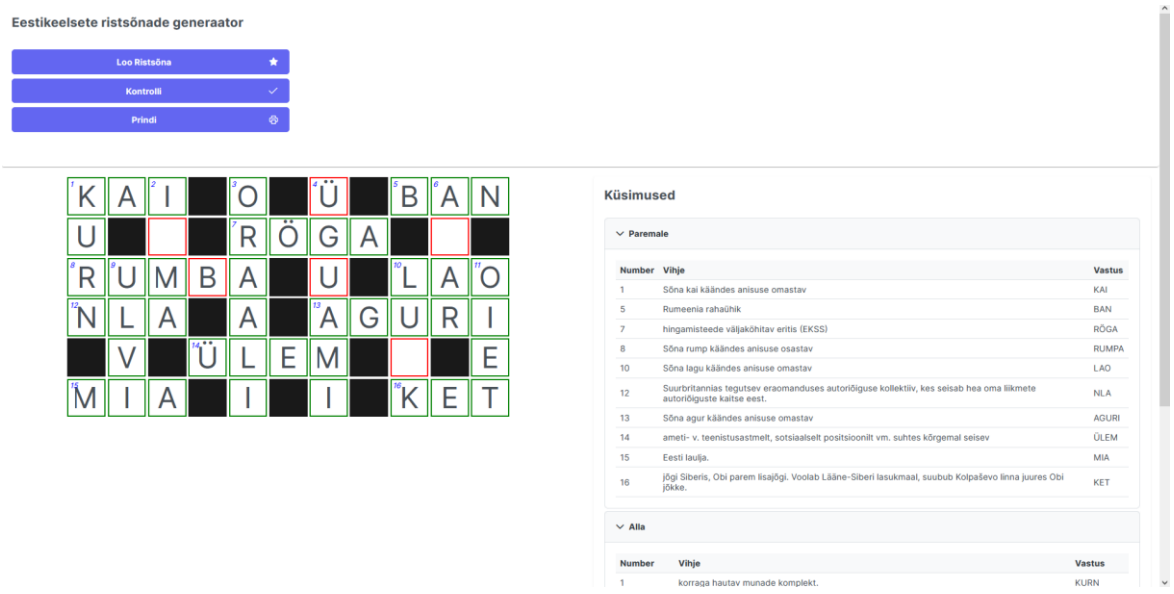
Peale ristsõna pärimist visualiseeritakse kasutajale ristsõna ja täidetakse küsimuste tabelid küsimustega (joonis 7). Küsimused on jagatud kaheks vastuste suuna järgi: paremale ja alla. Mõne välja ülemises vasakus nurgas on näha number, mis tähistab küsimuse numbrit. Kui samalt väljalt algab vastus nii paremale kui alla, on sama numbriga küsimus olemas nii paremale kui ka alla küsimuste seas.



Joonis 7. Näide veebirakendusest genereeritud ristsõnaga.

Igale väljale on võimalik sisestada vaid üks eesti tähestikku kuuluv täht. Küsimuste lahendused võivad sisaldada ka tühikuid. Kuna kasutajad tihti ei oota ristsõnades tühikute esinemist, siis tühikute lahtrid on eeltäidetud ja märgistatud. Kasutaja ei pea ise ära arvama tühikute olemasolu.

Kui ristsõna on päritud, on kasutajal igal ajahetkel võimalik vajutada „Kontrolli“ nuppu, et näha, millises väljades on korrektne täht ja millised on õiged vastused (joonis 8). Õigetele väljadele tekib roheline äär ja valedele punane. Õiged vastused küsimustele tekivad küsimuste tabelitesse uude tulpa nimega „Vastus“.



Joonis 8. Veebirakendus ristsõna kontrollimise olekus.

Veebirakendusel on ka printimise funktsionaalsus. Vajutades nuppu „Prindi“ saab printida välja ristsõna olekus, milles see parasjagu on. Printimisel väljastatakse vaid ristsõna ise ja küsimuste tabelid.

3.6 Ristsõnade generaatorile antud tagasiside

Eestikeelsete ristsõnade generaatori testimiseks kasutati küsimustikku (lisa 1). Katseisikutel paluti üritada lahendada veebirakenduses vähemalt 3 ristsõna ning seejärel täita küsimustik. Küsimustikus küsiti inimeste ristsõnade lahendamise harjumuste ja veebirakenduse kohta. Küsimustiku tagasisidet kasutati ilmnenud puudujääkide likvideerimiseks.

3.6.1 Testijate taust

Küsimustikule vastas 8 inimest vanusevahemikus 21-25. Ühelgi vastajal tehnilisi tõrkeid ei esinenud. Peaaegu kõik vastajad lahendasid ristsõnu erineva sagedusega. Ei leidunud ühtegi vastajat, kes lahendaks ristsõnu rohkem kui kord nädalas. Vastajate kogemuse erinevus aitab koguda andmeid erinevatest vaatenurkadest.

3.6.2 Eesliidese tagasiside

Kasutusmugavust hinnati halvast väga heani, kus 1 vastas „halb“, 1 „keskmine“, 3 „hea“ ja 3 „väga hea“. Toodi välja, et ristsõna täitmine on ebamugav, kui rakendus ei liigu lahtri täitmisel automaatselt järgmise ruudu juurde ja soovitati lisada nooltega liikumine ruutude vahel. Kiideti veebirakenduse minimalistlikku disaini.

5 vastajat 8-st pidas kasutajaliidese selgust väga heaks ja 3 heaks. Üldjoontes oldi arvamusel, et kasutajaliidese juures on kõik selge ja arusaadav. Märgiti, et mittefunktsionaalses olekus nupud võiks kasutaja eest ära peita ning et eelistataks ristsõna ruutude vahede puudumist. Üks vastaja tõi esile, et iga kasutaja ei pruugi märgata, et „Kontrolli“ nupu vajutamisel ilmuvad küsimuste juurde ka õiged vastused.

Tagasisides soovitati lisada rakendusele värvi, et see silmale meeldivam oleks. Soovitati ka jagada kontrollimine ja vastuste vaatamine kaheks eraldi funktsionaalsuseks.

3.6.3 Ristsõnade tagasiside

Keskmiselt suudeti ristsõnu täita umbes 50% ulatuses. Vastajate täidetud ruutude hulk oli vahemikus 21%-80%. Ristsõnade küsimuste raskustaset määrati keskmiseks ja raskeks võrdse jaotusega. 4 vastajat koges testimise käigus küsimusi, mis olid vigased või arusaamatud. Sellised küsimused täpsustati tagasisides.

Küsimuste kvaliteeti hinnati erinevalt, kõik arvamused peale „väga hea“ olid esindatud. Enamus vastajad hindasid negatiivselt liiga suurt sõna käänamise küsimuste hulka. Oldi ka vastupidisel arvamusel: leiti, et lihtsamad käänamise küsimused muutsid keerulisemate küsimuste vastamist lihtsamaks. Toodi esile paljude sisukamate küsimuste keerulisust.

Tuleb arvestada, et testimisel kasutatud versioon veebirakendusest ei kasutanud veel kaheks jaotatud andmestikku, mis oluliselt vähendas käänamisküsimuste osakaalu. See implementeeriti hiljem tagasisidele toetudes.

Hariduslikku väärtust peeti 3 vastaja poolt heaks, 1 poolt keskmiseks, 2 poolt halvaks ning 2 puudus arvamus. Keerulisi küsimusi nähti nii positiivsetena kui negatiivsetena. Mainiti, et oli huvitav vastuseid internetist otsida ja seeläbi silmaringi avardada, aga üks vastaja leidis, et veebirakendus õpetab vaid käänamist. Toodi soovitusena esile, et rakenduse hariduslik väärtus oleks oluliselt suurem, kui ristsõnadel oleks kategooria.

Tagasisides avaldati soovitust keskenduda küsimuste kvaliteedi tõstmisele, mis keskenduks peamiselt vigaste küsimuste, liigselt raskete ja umbkaudsete küsimuste eemaldamisele.

6 vastajat 8-st soovitaks veebirakendust ka mõnele tuttavale.

3.6.4 Tuvastatud vead

Kasutajad kirjeldasid veebirakenduse testimise käigus kogetud vigu. Toodi esile kirjaviga, mis sai järgmises versioonis parandatud. Toodi esile veel küsimusi, mis olid kas liiga üldised või vigased. Samuti toodi esile küsimusi, kus vastus esines küsimuses endas. Sellised olukorrad võimalikud, kui morfoloogiline analüsaator pole võimeline korrektselt tundma ära sõna algvormi. Sellises olukorras ei oska programm definitsioonist korrektselt vastust tuvastada.

Tuvastati olukord, kus ristsõnas korrektseks märgitud täht polnud sama, mis vastuses märgitud. Selline situatsioon oli tingitud veast sõnaväljadesse vastuste valimisel, kui vastuse tingimused olid rekursiooni jooksul muutunud. Viga parandati hilisemas versioonis.

Leidus vastuseid, mis vastajate arvates ei olnud piisavas korrelatsioonis küsimusega. Kontrollimisel selgus, et küsimused ei olnud vigased. Vastuste sõnad olid sellistel juhtudel vähe kasutusel olevad sõnad, tekitades vastajates segadust.

Tagasisidest saadi palju mõtteid ja ideid veebirakenduse parandamiseks, mida ei jõutud realiseerida. Järgmises peatükis antakse ülevaade neist.

3.7 Edasised arenguvõimalused

Töö põhieesmärk oli luua kõigile kättesaadav automaatne eestikeelsete ristsõnade generaator, mis kogub andmed programmiliselt. Põhieesmärk sai täidetud, aga leidub veel mitmeid viise, kuidas loodud veebirakendust kaugemale arendada.

Kõige rohkem võidaks kasutaja küsimuste kvaliteedi täiendamisest. Vikipeediast saaks koguda veel täiendavat informatsiooni, et formuleerida küsimusi, mis siinses töös välja jäid.

Näiteks jäid välja enamus Vikipeedias kirjeldatud lühenditest. Üks edasiarenduse võimalusi oleks veebikoorimist täiendada, et koguda ka informatsiooni lühendite tähenduste kohta. Sedasi saaks luua küsimuste vastuste paarid, kus vastus on lühend ja vastus selle üks võimalikest tähendustest.

Ruudusiku genereerimisel ei tohiks kunagi tekkida nõuetele mittevastavaid ruudusikke. Edasiarendus oleks tagada alati korrektsete ruudustike genereerimise. See vähendaks vajalike operatsioonide arvu ja seeläbi optimeeriks genereerimise protsessi.

Ristsõna täitmise ajal jääb algoritm vahel pikalt ühe tsükli peale kinni, kui peaksid kõrvuti sattuma tähed, mis ei võimalda leida sobivat sõna, näiteks D ja T. Algoritmi oleks võimalik edasi arendada tuvastama selliste olukordade teket, et neid vältida ja seeläbi kiirendada genereerimise protsessi.

Kasutajaliidese puhul märgiti ära ebamugavusi, mida samuti saaks parandada. Tähe sisestamise järel peab kasutaja ise valima uue ruudu kas hiirega või vajutades tabeldusklahvi, mis aga võimaldab ainult paremale järgmisele ruudule liikuda. Oluliselt mugavam oleks ristsõna täita, kui tähe sisestamisel muutuks järgmine ruut automaatselt aktiivseks. Samuti parandaks kasutusmugavust nooltega liikumise võimalus.

Küsimused võisid olla sageli äärmiselt rasked vastata. Selle parandamiseks oleks võimalik küsimustele määrata raskusastmeid nende vastuste esinemissageduste järgi eesti keeles. See muudaks võimalikuks ka erinevate raskusastmetega ristsõnade loomise. Samuti oleks üks võimalik edasiarendus küsimusi kategoriseerida, et ristsõnad keskenduks kindlale teemale ja omaksid seeläbi suuremat hariduslikku väärtust.

Kokkuvõte

Bakalaureusetöö eesmärk oli luua veebirakendus, mis koostaks programmiliselt kogutud andmete põhjal automaatselt ristsõnu ning oleks nii veebilehel lahendatav kui ka paberile trükitav. Valminud veebirakendus on hariv ja meelelahutuslik tööriist kõikidele vanustele. Osade küsimuste keerukusest tulenevalt on soovitatav kasutada väliseid allikaid, et otsida vastuseid ja seeläbi enda silmaringi laiendada.

Töös uuriti ristsõnade kasulikkust hariduses, meelelahutuses ja meditsiinis. Samuti uuriti juba eksisteerivaid realisatsioone ristsõnade generaatoristest ja nende puudujääke.

Valminud veebirakendus Eestikeelsete ristsõnade generaator on saadaval eraldi veebilehel[15]. Andmeid koguti eestikeelsest Vikipeediast, kasutades veebikoorimist (ingl *web scraping*), ja Eesti Wordnetist. Viimasest saadi andmed, kasutades seal leiduvaid definitsioone ja EstNLTK morfoloogilist sünteesi. Tulemuseks loodi kaks andmestikku: põhiandmestik, kus 168 384 vastuse ja definitsiooni paari, ja lisaandmestik, kus 683 554. Ristsõna genereeriti, kasutades kitsenduste rahuldamise meetodit rekursiivse algoritmiga. Veebirakendusele loodi kasutajaliides, kasutades React teeki. Valminud veebirakendus on eesti keeleruumis ainulaadne oma programmiliselt kogutud andmete tõttu. Teadaolevalt ei eksisteeri teist täisautomaatset lahendust eestikeelsest ristsõnade generaatorist.

Veebirakendust testis ja andis tagasisidet 8 inimest. Kogutud tagasiside põhjal tehti muudatusi, et parandada ristsõna küsimusi, genereerimist ja kasutajaliidest. Üldiselt suhtuti veebirakendusse positiivselt ning 6 inimest 8-st soovitaks veebirakendust ka mõnele tuttavale. Arvestades üldjoontes positiivset tagasisidet saab öelda, et töö eesmärk sai täidetud. Veebirakendus töötas ilma tehniliste probleemideta ja tagas kasutajatele kiirelt mitmekülgseid ristsõnu. Nõrgem külg, mida ka kasutajad esile tõid, oli ristsõna küsimuste kvaliteet ja kohati ebamugav tähtede sisestamise meetod. Leidub mitmeid arenguvõimalusi, et tõsta küsimuste kvaliteeti ja kasutajamugavust, mida tulevikus kaaluda.

Töö tulemusena sai loodud huvitav ja silmaringi laiendav veebirakendus, mis pakkus väljakutseid terve arenduse jooksul.

[15] Eestikeelsete ristsõnade generaator: https://ristsonade-generaator.keeleressursid.ee/

Viidatud kirjandus

[1] Mahmood Muhammad.Asim, Tariq Maria., Javed Saira., STRATEGIES FOR ACTIVE LEARNING: AN ALTERNATIVE TO PASSIVE LEARNING, *Academic Research International*, vol. I, no. 3, pp. 194-198, 2011.

[2] Murphy Mike., O'Sullivan Katie., Kelleher Kieran.G, Daily crosswords improve verbal fluency: A brief intervention study., *International Journal of Geriatric Psychiatry*, vol. 29, no. 9, pp. 915-919, 2014.

[3] Pillai J.A et al., Association of Crossword Puzzle Participation with Memory Decline in Persons Who Develop Dementia, *Journal of the International Neuropsychological Society*, vol. 17, no. 6, pp. 1006-1013, November 2011.

[4] The Teacher's Corner. https://worksheets.theteacherscorner.net/make-your-own/crossword/ (4.Aprill.2022)

[5] Crossword Labs. https://crosswordlabs.com/ (4.Aprill.2022)

[6] Tomozov Pavel., Crossword Construction using Constraint Satisfaction and Simulated Annealing, Tartu Ülikool, Tartu, TÜ arvutiteaduse instituudi bakalaureusetöö 2013.

[7] Rigutini Leonardo., Diligenti Michelangelo., Maggini Marco., Gori Marco., A Fully Automatic Crossword Generator, in *Seventh International Conference on Machine Learning and Applications*, Siena, 2008.

[8] Lottering R., Hans R., Lall M., The impact of Crossword Puzzles on Students' Performance: Does Pre-exposure to Puzzles Matter?, in *IEEE International Conference on Teaching, Assessment, and Learning for Engineering*, Wollongong, 2018, pp. 545-550.

[9] Mshayisa V.V, Students' perceptions of Plickers and crossword puzzles in undergraduate studies, *Journal of Food Science Education*, vol. 19, no. 2, pp. 49-58, April April 2020.

[10] Patrick Shilpa. et al., The usefulness of crossword puzzle as a self-learning tool in pharmacology, *Journal of Advances in Medical Education and Professionalism*, vol. 6, no. 4, pp. 181-185, 2018.

[11] Ranaivo-Malancon B., Lim T., Minoi J-L., Jupit A.J.R, Automatic generation of fill-in clues and answers from raw texts for crosswords, in *International Conference on Information Technology in Asia (CITA)*, Kota Samarahan, 2013.

[12] Beacham Adam., Chen Xinguang., Sillito Jonathan., Beek Peter.van, Constraint Programming Lessons Learned from Crossword Puzzles, in *14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, Ottawa, 2001.

[13] Department of Computer & Information Sciences. https://cis.temple.edu/~giorgio/cis587/readings/constraints.html (20.Märts.2022)

[14] Koit Mare., Roosmaa Tiit., *Tehisintellekt*. Tartu: Tartu Ülikooli Kirjastus, 2011.

[15] Python. https://www.python.org/about/ (6.Märts.2022)

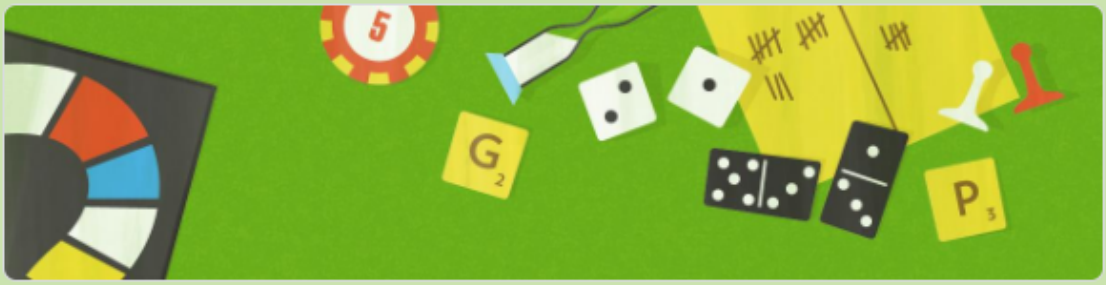
[16] Scrapy dokumentatsioon. https://docs.scrapy.org/en/latest/ (6.Märts.2022)

[17] Pandas. https://pandas.pydata.org/about/ (6.Märts.2022)

[18] EstNLTK Github.
https://github.com/estnltk/estnltk (6.Märts.2022)
[19] Eesti Wordnet.
https://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=et (6.Märts.2022)
[20] Redis.
https://redis.io/ (6.Märts.2022)
[21] Redis Queue.
https://python-rq.org/ (13.Märts.2022)
[22] Flask.
https://flask.palletsprojects.com/en/2.0.x/foreword/ (6.Märts.2022)
[23] React.
https://reactjs.org/ (6.Märts.2022)

Lisad

I Küsimustik



Veebirakenduse Eestikeelsete ristsõnade generaator testimisküsimustik

Käesolev küsimustik on tagasiside kogumiseks Allan Loo bakalaureusetöö käigus loodud veebirakendusele Eestikeelsete ristsõnade generaator.
Veebirakendus kasutab programmiliselt kogutud andmeid, et genereerida eestikeelseid ristsõnu. Võimalik on sisestada vaid eesti tähestiku tähti.
Küsitlus on anonüümne ning vabatahtlik. Saates ära vastused annab osaleja nõusoleku kasutada vastuseid eelmainitud töös. Küsimustiku tulemusi saab näha 2022 Tartu Ülikoolis kaitstavas bakalaureusetöös.
Kontaktisik: Allan Loo

Edenemise salvestamiseks logige Google'isse sisse. Lisateave

* Kohustuslik

Vanus *

Teie vastus

Kui tihti lahentate ristsõnu? *

- ◯ Peaaegu mitte kunagi
- ◯ Kord aastas
- ◯ Kord kuus
- ◯ Kord nädalas
- ◯ Rohkem kui kord nädalas

Juhised

Veebirakendus on leitav aadressilt http://35.228.28.13/. Üürita lahendada, kasvõi osaliselt, vähemalt kolme ristsõna ja peale seda täida ülejäänud küsimustik. Kuna andmed on kogutud programmiliselt, võib seal leiduda problemaatilisi küsimusi ja nõrkasid seoseid küsimuste ja vastuste vahel. Kontrolli nupp võimaldab näha õigeid vastuseid. Kontrollimise funktsionaalsust saab sama ristsõna peal vabalt sisse-välja lülitada. Jätke meelde tekkivad probleemid, et neid küsimustikus välja tuua.

Kui suures mahus keskmiselt suutsite ristsõna ära täita? *

- ◯ 0%
- ◯ 1% - 20%
- ◯ 21% - 40%
- ◯ 41% - 60%
- ◯ 61% - 80%
- ◯ 81% - 99%
- ◯ 100%

Kuidas hindade küsimuste raskustaset? *

- ◯ Väga raske
- ◯ Raske
- ◯ Keskmine
- ◯ Kerge
- ◯ Väga kerge
- ◯ Puudub arvamus

Kas kogesite küsimusi, mis olid vigased või arusaamatud? *

- ◯ Jah
- ◯ Ei

Kui jah, siis mis oli probleemiks? (Võimalikult täpselt)

Your answer

Anna hinnang järgnevatele omadustele *

	Väga halb	Halb	Keskmine	Hea	Väga hea	Puudub arvamus
Kasutusmugavus	◯	◯	◯	◯	◯	◯
Kasutajaliidese selgus	◯	◯	◯	◯	◯	◯
Küsimuste kvaliteet	◯	◯	◯	◯	◯	◯
Hariduslik väärtus	◯	◯	◯	◯	◯	◯

Kasutusmugavuse kommentaarid:

Your answer

Kasutajaliidese selguse kommentaarid:

Your answer

Küsimuste kvaliteedi kommentaarid:

Your answer

Haridusliku väärtuse kommentaarid:

Your answer

Kas soovitaksite ristsõnade generaatorit ka mõnele tuttavale? *

- Jah
- Ei

Kui esines tehnilisi tõrkeid, siis kirjeldage neid:

Your answer

Kommentaare veebirakenduse kohta:

Your answer

Submit Clear form

II Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina,

Allan Loo,

(autori nimi)

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

Eestikeelsete ristsõnade generaator,

(lõputöö pealkiri)

mille juhendaja on

Sven Aller,

(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Allan Loo

10.05.2022