Tallinna Ülikool
Digitehnoloogiate Instituut

Eesti vahekeele korpuse klasteranalüüsi vahendite kasutamine teksti keeletaseme prognoosimisel

Bakalaureusetöö

Autor: Virgo Hallik
Juhendaja: Jaagup Kippar

Autor:.....„„.....”2016
Juhendaja:.....„„.....”2016
Instituudi direktor:.....„„.....”2016

Tallinn 2016

Autorideklaratsioon

Deklareerin, et käesolev seminaritöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

....

(kuupäev) (autor)

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina _____(sünnikuupäev: _____)

(*autori nimi*)

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

   _____

   _____

   _____

   (*lõputöö pealkiri*)

   mille juhendaja on _____,

   (*juhendaja nimi*)

   säilitamiseks ja üldsusele kättesaadavaks tegemiseks Tallinna Ülikooli Akadeemilise Raamatukogu repositooriumis.
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.

Tallinnas/Haapsalus/Rakveres/Helsingis, _____

(digitaalne) allkiri ja kuupäev

Sisukord

Sissejuhatus

Bakalaureusetöö teema valikul lähtus autor seminaritöö raames tehtust, milles võrreldi Tallinna Ülikooli eesti vahekeele korpuse (edaspidi EVKK) eestikeelsete kirjalike tekstide sõna, lause ja teksti pikkust, täiendas eestikeelsete tekstide võrdlusleheküIge. Sellega seoses pakkus huvi automaatne tekstide keeletaseme prognoosimine.

Bakalaureusetöö eesmärk on EVKK klasteranalüüsi vahendeid kasutades luua veebirakendus kasutaja sisestatud eestikeelse tekstide automaatse keeletaseme prognoosimiseks. Eesmärki põhjendab vajadus keeletasemete automaatse hindamise järele eesti keele alusel.

Bakalaureusetöö ülesanneteks on 1. anda lühiülevaade automaatsest keeleoskuse hindamisest; 2. interpreteerida rakenduse loomiseks vajalikke klasteranalüüsi andmeid; 3. luua rakenduse prototüüp, kirjeldada selle arendust, katsetada töö tulemusi, et otsustada, missuguses suunas tööd jätkata.

Rakendust on võimalik testida leheküljel
http://greeny.cs.tlu.ee:18188/korpus/korpus/Search/klaster.html

1. Keelekorpused

Siin tutvustatakse korpuse mõistet keeleteaduses, Eesti vahekeele korpust ja antakse lühiülevaade teistest keelekorpustest Eestis.

Keeleteaduses mõeldakse sõna *korpus* all tavaliselt keeleainese kogu, mida kasutatakse uurimistöö
materjalina. Tänapäeval mõeldakse korpuse all peamiselt polüfunktsionaalseid elektroonilisel kujul
olevaid tekstikogusid, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev
tervik annaks tõepärase pildi keelest (vt Muischnek[1]).

1.1. Korpused Eestis

Eesti Keeleressursside Keskus on koondanud ühtsesse andmebaasi kõik korpused, mis eesti keele kohta koostatud. Esindatud on nii siinse uurimuse keeleainese allikas EVKK (eesti õppijakeel) kui ka eesti kirjakeele tasakaalus korpused, vana kirjakeele korpus, eesti-inglise paralleelkorpus, võru keele korpus, samuti piibli- ja murdekeele ning suulise kõne korpus. Spetsiaalsed ressursid on morfoloogiliselt ühestatud korpus, ühestatud sõnatähenduste korpus, pindsüntaktiliselt märgendatud korpus ja puudepank[2]. Enamus loetletud korpusressurssidest on kättesaadav Keeleveebi vahendusel[3]

Eesti vahekeele korpus (EVKK)

Eesti vahekeele korpus on eesti keele õppijate kirjalike tekstide kogu, mille praegune versioon on valminud Tallinna Ülikooli filoloogide, haridustehnoloogide ja informaatikute koostööna. Tegemist
on monitorkorpusega, kuna sellesse lisatakse pidevalt uusi tekste.

Korpust saab kasutada
1)empiirilist ja rakenduslikku laadi uurimistöös,

[1] Vt http://www.cl.ut.ee/kursused/korp_ling01 (23.04.2016)
[2] Vt https://keeleressursid.ee/et/keeleressursid/tekstikorpused (23.04.2016).
[3] Vt http://www.keeleveeb.ee (23.04.2016).

2)tulevaste õpetajate ning lingvistide koolitamisel,

3)tegevõpetajate täiendõppes,

4)eesti keele õpetamisel ja individuaalõppes.

Korpuse funktsionaalsusi ja kirjalike tekstide analüüsi vahendeid saavad kasutada kõik, eriõigused

on registreeritud kasutajatel, andmehalduril ja programmeerijal. (vt Eslon 2014: 438.)

EVKK praeguses versioonis on võimalik

1)teha päringuid dokumentide loendile, tekstidele ja statistikale erinevate parameetrite alusel

2)uurida korpuse statistikat

3)kasutada morfosüntaktilist analüüsi, sh eesti, soome ja vene keele morfoanalüüsi TreeTaggeriga

4)saada sisetatud teksti kohta statistilisi andmeid, nt sõna, lause ja teksti pikkuse alusel teksti keeleoskustaseme ennustamine

5)kasutada silbitajat

6)kasutada pöördsõnastikku

7)uurida korpuse tekstide sõna- ja vormisagedust

8)kasutada korpuse tekstide veataksonoomiat

Muud eesti keele korpused

Lisaks leidub Eesti Keeleressursside Keskuse all veel hulgaliselt muid keelekorpuseid. Siinkohal loetelu nendest.

•Eesti Kirjakeele Korpus 1890-1990 (10 alamkorpust)

http://www.cl.ut.ee/korpused/baaskorpus

•Eesti keele koondkorpus

http://www.cl.ut.ee/korpused/segakorpus/

•Tasakaalus korpus

http://www.cl.ut.ee/korpused/grammatikakorpus

•Morfoloogiliselt ühestatud korpus

http://www.cl.ut.ee/korpused/morfkorpus

•Ühestatud sõnatähenduste korpus

http://www.cl.ut.ee/korpused/semkorpus

•Vana kirjakeele korpus (VAKK)

http://www.murre.ut.ee/vakkur/Korpused/korpused.htm

•Pindsüntaktiliselt märgendatud eesti keele korpus

http://math.ut.ee/~kaili/Korpus/pindmine

•Inglise-eesti ja eesti-inglise paralleelkorpus

http://www.cl.ut.ee/korpused/paralleel

•Eesti keele puudepank

http://www.ut.ee/~kaili/Korpus/puud

•Eesti Keele Instituudi tekstikorpus

http://portaal.eki.ee/corpus

•Eesti piiblitõlke ajalooline konkordants

http://portaal.eki.ee/piibel

•Eesti keele spontaanse kõne foneetiline korpus

http://www.keel.ut.ee/et/foneetikakorpus

•Eesti murrete korpus

http://www.keel.ut.ee/et/keelekogud/murdekorpus

•Võru keele korpus

http://www.murre.ut.ee/voru/

2. Automaatne keeletaseme hindamine

Selles alapunktis antakse ülevaade mõningatest ingliskeelsete esseede automaatse hindamise tehnoloogiatest ja sammudest, mida on ette võetud teiste keelte tasemeoskuse automaatsel hindamisel, et luua vigade ennustamise automaatseid süsteeme.

2.1. Olemasolevad tehnoloogiad keeleoskuse hindamiseks inglise keele alusel

Esseede automaatset hindamist kasutatakse laialdaselt inglise keele peal, sealhulgas keeletestides nagu *Graduate Record Examination* (GRE[4]) ja *Graduate Management Admission Test* (GMAT[5]).

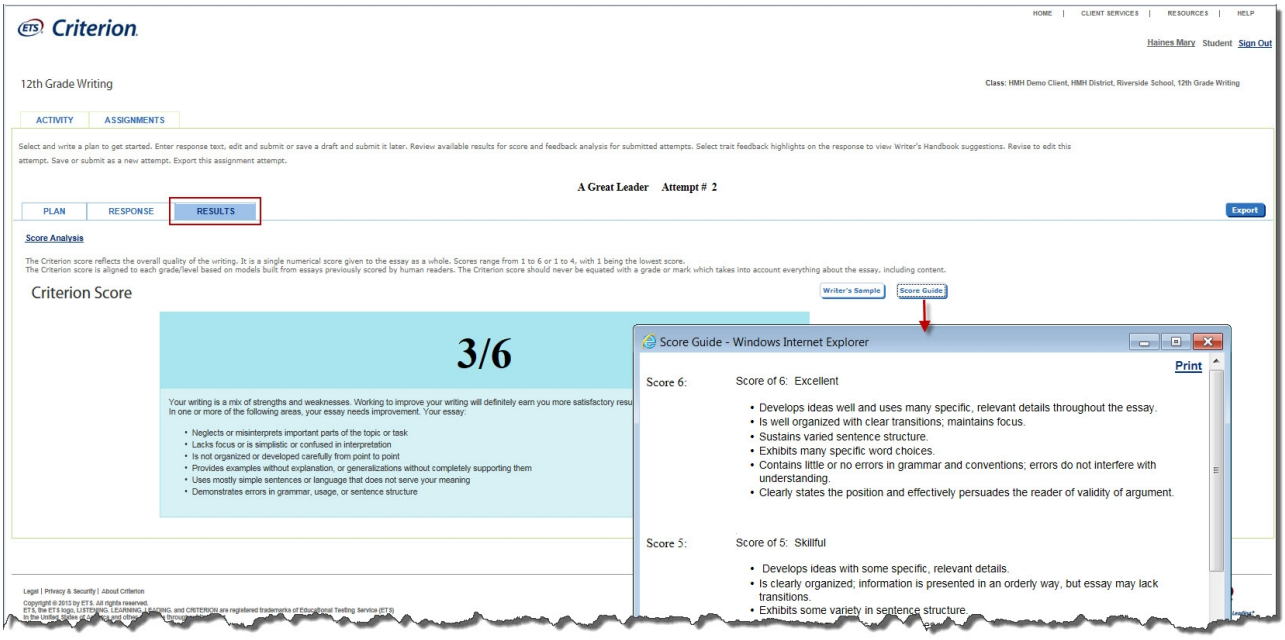
[4] Vt http://www.gre.org (20.04.2016).

[5] Vt http://www.gmac.com/gmat.aspx (20.04.1016).

Esseede automaatse hindamise tehnoloogiad on kasutajale enamasti tasulised. Siinkohal mõned näited.

ETS – e-rater[6]:

Hinnatakse grammatikat, kasutust (sõnad ja sõnavormid), sõnakombinatsioone, sõnavara keerukust, stiili, teksti organiseeritust ja selle ladusust, teemaarendust, märksõnade kasutatavust ja essee pikkust (Attali, Burstein 2004[7]). ETS – e-rater süsteemi on võimalik kasutada läbi teenuste Criterion[8] (vt joonis 1) ja ScoreItNow![9], viimasega hinnatakse ka GRE teste. Mõlema puhul on tegemist tasulise teenusega, mille hind teadmata.



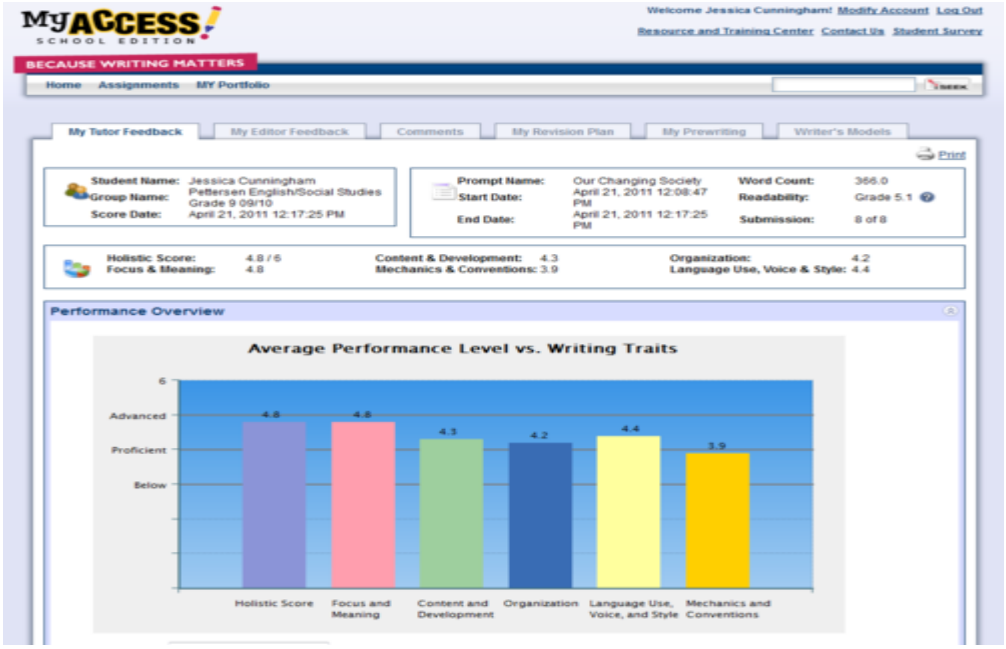
Joonis 1 Criterion'i tulemuselehe näide kodulehelt

Vantage Learning - Intellimetric[10]

Hinnatakse fookust ja mõtet, teksti organiseeritust, sisu ja teemaarendust, keelekausutust ja stiili, keelendite kombineerimist ja konventsiooni[11]. Kasutatakse GMAT testide hindamisel, mis võimalik läbi teenuse MyAccess, hind 3 õpilase jaoks 99.95$. Joonisel 2 on pilt MyAccess'i

[6] Vt https://www.ets.org/erater/about (20.04.2016).
[7] Vt https://www.ets.org/Media/Products/e-rater/erater_IAEA.pdf (20.04.2016)
[8] Vt https://www.ets.org/criterion/about/ (20.04.2016).
[9] Vt https://www.dxrgroup.com/cgi-bin/scoreitnow/index.pl (20.04.2016).
[10] Vt http://www.vantagelearning.com/products/intellimetric/ (20.04.2016).
[11] Vt http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/ (20.04.2016)

tulemuslehest.



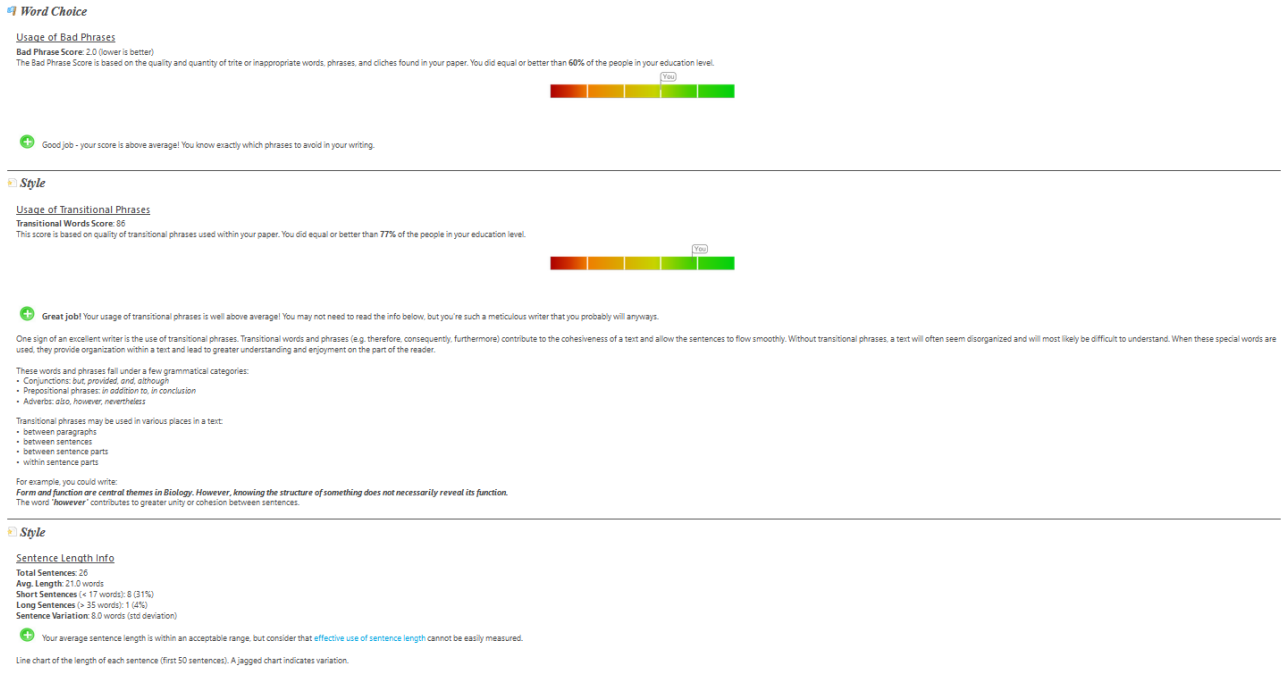
Joonis 2 Myaccess'i tulemuselehe näide kodulehelt

Measurment Inc. - Project Essay Grade[12]. Lehel ei ole välja toodud, mille alusel esseesid hinnatakse. Võimalik on kasutada teenust PEG Writing[13], hind teadmata (vt joonis 3).

Joonis 3 PEG Writing tulemuselehe näide kodulehelt

PaperRater[14]

Hindamisel lähtutakse mitmest tegurist, sealhulgas sõnade valik, õigekiri ja grammatika (vt joonis 4). Pakutakse nii tasuta kui ka tasulist teenust, kusjuures automaatset hindamist saab kasutada mõlemal juhul – seatud on mahulised piirangud.



[12] Vt http://www.measurementinc.com/products/peg ja http://www.pegwriting.com/about (20.04.2016).
[13] Vt https://www.pegwritingscholar.com/ (20.04.2016).
[14] Vt https://www.paperrater.com (20.04.2016).

Joonis 4 PaperRater'i tulemuselehe näide

Kriitika

Automaatsed hindamissüsteemid aitavad suuresti vähendada hindajate ja õpetajate koormust, individualiseerivad õppimisega seotud aspekte, soodustavad iseõppimist jm, ent sellegipoolest on need süsteemid saanud ka kriitika osaliseks.

Ühe miinusena on välja toodud, et need süsteemid ei ole võimelised inimlikult hindama.

Teisena on märgitud, et hindamissüsteemid on liialt pinnapealsed, hinnates peamiselt esseede mõõdetavaid tegureid nagu essee pikkus ja sõnade järjestus või jättes need kõrvale ja hinnates ainult sisu. (Dikli 2006: 24-26)[15]

Kolmandaks probleemiks või ohuks on peetud võimalust automaatset hindamissüsteemi petta, kui selle tööpõhimõtted ja hindamise alused on teada. Hindamissüsteemide arendajad lükkavad need väited ümber või lubavad süsteeme edasi arendada, et sarnast situatsiooni vältida.

Neljandaks ohuks peetakse seda, et õpilased hakkavad kirjutama „masinale" ja õpetajad õpetama „masinale" kirjutamist, sest kõik teavad, et töid hinnatakse automaatselt.

Viiendaks, kui töid hinnatakse automaatselt ja on teada, millistel teguritel on essee hindamisel (järelikult ka kirjutamisel) suurem kaal, siis arvatakse, et eksamineerija võib süsteemi osaliselt petta. (Blood 2011: 55-59)[16]

Eespool loetletud puudused on automaatsete hindamissüsteemide arendajad kas juba ümber lükanud või lubavad süsteeme edasi arendades neid probleeme vältida.

Töö autor on arvamusel, et inimfaktor ei kao esseede hindamisel kuhugi. Automaatse hindamissüsteemi ja vastavate tarkvaralahenduste eesmärk pole inimteguri kaotamine, vaid nimetatud kahe suuna sümbioos, koostoimimine, et anda keeleõppijatele ja hindajatele avaramaid võimalusi enesekontrolliks. Sellest on abi keele õppimisel, keelekorralduses, samuti keelekasutuse uurimisel.

[15] Vt http://files.eric.ed.gov/fulltext/EJ843855.pdf (25.04.2016).
[16] Vt http://tesol-dev.journals.cdrs.columbia.edu/wp-content/uploads/sites/12/2015/04/4.-Blood-2011.pdf (25.04.2016).

2.2. Arenguid keeleoskuse automaatseks hindamiseks mitte-inglise keelte alusel

Lisaks arvukatele uurimustele, mis tehtud inglise keele kui võõrkeele tasemeoskuste hindamiseks – üks viimasel ajal ilmunud uurimustest on korealaste inglise keele kirjutamisoskuse prognoosimine keelekasutuse lingvistilise keerukuse alusel (vt Kim, Ji-Young 2014) – on ka saksa (vt Hancke, Meurers 2013), rootsi (vt Östling, jt 2013) ja eesti keele põhjal (vt Vajjala, Lõo 2013) astutud samme neis keeltes kirjutatud esseede automaatseks hindamiseks. Keskne roll on siin Walt Detmar Meurersi töörühmal Thübingeni Ülikoolis[17]. Vajjala ja Lõo on seotud just selle uurimisrühma töödega. Eesti õppijakeele andmeallikana on nad kasutanud EVKK ressurssi. 2014. aastal jätkasid Vajjala ja Lõo eesti õppijakeele esseede automaatse hindamise võimaluste katsetamist. Loodi klassifikatsiooni ja regressiooni mudelid, selleks võrreldi tekste 72 tunnuse alusel. Mudelite võrdlemisel jõuti järeldusele, et arvestades keelevigu on klassifikatsioon efektiivsem kui regressioon. Klassifikatsiooni mudel võimaldas keeleoskustaset prognoosida 79% täpsusega, korrelatsioon 0.85. (Vajjala, Lõo 2014)

Siinne bakalaureusetöö on katse eristada eesti keele tasemeid mustrite alusel, mis eristuvad sõnaliigijärjenditena. Sellele eelnevalt on loodud EVKK tuumkorpuse põhjal statistiline mudel, mis määrab esseede taset sõna, lause ja teksti keskmise pikkuse alusel[18]. Iga soovija võib sisestada korpuse aknasse oma teksti ja saada vastuse, millisel tasemel tekstiga võiks nimetatud tunnuste põhjal olla tegu.

3. Keeletasemete klasteranalüüsi tulemustest

Selles alapunktis kirjeldatakse analüüsi taustainfot, vaadeldakse lähemalt iga keeletaseme 15-t suurema esinemusega klastrit, mille põhjal luuakse rakenduse prototüüp. Klastrite kirjeldus on tehtud nende osakaalu muutumise põhjal, liikudes alamalt keeleoskustasemelt kõrgemale. Klastrid tulevad esile sõnaliigijärjendite alusel, iga sõnaliiki tähistab vastav lühend (vt allpool toodud lühendite seletused).

[17] http://www.sfs.uni-tuebingen.de/~dm/index.html (30.04.2016).
[18] Vt http://evkk.tlu.ee/wordtree/usertext.html?GetInfo (30.04.2016).

Sõnaliigi lühendite seletused[19]

A - omadussõna - algvõrre (adjektiiv - positiiv), nii käänduvad kui käändumatud, nt *kallis* või *eht*

C - omadussõna - keskvõrre (adjektiiv - komparatiiv), nt *laiem*

D - määrsõna (adverb), nt *kõrvuti*

G - käändumatu omadussõna (genitiivatribuut), nt *balti*

H - pärisnimi, nt *Edgar*

I - hüüdsõna (interjektsioon), nt *tere*

J - sidesõna (konjunktsioon), nt *ja*

K - kaassõna (pre/postpositsioon), nt *kaudu*

N - põhiarvsõna (kardinaalnumeraal), nt *kaks*

O - järgarvsõna (ordinaalnumeraal), nt *teine*

P - asesõna (pronoomen), nt *see*

S - nimisõna (substantiiv), nt *asi*

U - omadussõna - ülivõrre (adjektiiv - superlatiiv), nt *pikim*

V - tegusõna (verb), nt *lugema*

X - tegusõna juurde kuuluv sõna, millel eraldi sõnaliigi tähistus puudub, nt *plehku*

Y - lühend, nt *USA*

Z - lausemärk, nt -, /, ...

3.1. Analüüsi taustainfo

Kõiki EVKK-s olevaid esseesid analüüsiti keeletasemete kaupa EVKK morfosüntaktilise analüüsi vahendiga Klastrileidja (vt 5.1). Päring võimaldab varieerida klastri pikkust (bi-, tri-, tetragrammid jne) ja analüüsi lingvistilist aspekti (morfoloogiline, süntaktiline, morfosüntaktiline). Käesoleva töö jaoks valiti klastri pikkuseks 3 üksust (trigrammid) ja lingvistilise parameetrina „Ainult sõnaliigid", st morfoloogiline aspekt. Välja on jäetud keeletasemed A1 ja C2, milles esseed puudusid. Üldvalim koosneb neljast alamvalimist:

1) A2-tase, 81 dokumenti, teksti keskmine pikkus – 143 sõnet ehk tekstisõna

2) B1-tase, 250 dokumenti, teksti keskmine pikkus – 210 sõnet

3) B2-tase, 124 dokumenti, teksti keskmine pikkus – 385 sõnet

4) C1-tase, 59 dokumenti, teksti keskmine pikkus – 868 sõnet.

[19] Vt http://www.filosoft.ee/html_morf_et/morfoutinfo.html (24.04.2016).

Iga keeletaseme analüüsimisel andis Klastrileidja tulemuseks 200 klastrit, väikseimaks klastrite hulgaks 5, kõikide tasemete peale on kokku 245 erinevat klastrit. Järgnevalt analüüsitakse kõikide keeletasemete 15 levinumat klastrit, tuues tasemete vahel esile erinevused ja muutused klastrite esinemuses ja varieerumisel. Valim on piiratud 15 klastriga, et testida rakenduse prototüüpi esmalt võimalikult madala arvu klastritega keeletaseme kohta, mis oleks samaaegselt piisav, et näha keeletasemete eristumist klastrite alusel.

3.2. Levinumad klastrid

A2 taseme 15 levinumat klastrit, järjestatuna nende osakaalu järgi, on asesõna-tegusõna-nimisõna ehk PVS, nimisõna-tegusõna-nimisõna ehk SVS, sidesõna-asesõna-tegusõna ehk JPV, nimisõna-sidesõna-nimisõna ehk SJS, asesõna-tegusõna-tegusõna ehk PVV, asesõna-nimisõna-tegusõna ehk PSV, tegusõna-nimisõna-nimisõna ehk VSS, asesõna-tegusõna-määrsõna ehk PVD, tegusõna-omadussõna-nimisõna ehk VAS, tegusõna-nimisõna-sidesõna ehk VSJ, nimisõna-tegusõna-omadussõna ehk SVA, nimisõna-tegusõna-määrsõna ehk SVD, tegusõna-määrsõna-omadussõna ehk VDA, tegusõna-asesõna-nimisõna ehk VPS, tegusõna-tegusõna-nimisõna ehk VVS ja tegusõna-määrsõna-nimisõna ehk VDS.

B1 taseme 15 levinumat klastrit on nimisõna-tegusõna-nimisõna ehk SVS, sidesõna-asesõna-tegusõna ehk JPV, asesõna-tegusõna-nimisõna ehk PVS, asesõna-tegusõna-tegusõna ehk PVV, tegusõna-asesõna-nimisõna ehk VPS, tegusõna-omadussõna-nimisõna ehk VAS, nimisõna-sidesõna-nimisõna ehk SJS, tegusõna-nimisõna-nimisõna ehk VSS, asesõna-nimisõna-tegusõna ehk PSV, nimisõna-tegusõna-määrsõna ehk SVD, nimisõna-nimisõna-tegusõna ehk SSV, nimisõna-tegusõna-omadussõna ehk SVA, tegusõna-määrsõna-omadussõna ehk VDA, sidesõna-nimisõna-tegusõna ehk JSV ja asesõna-tegusõna-määrsõna PVD.

B2 taseme 15 levinumat klastrit on tegusõna-asesõna-nimisõna ehk VPS, nimisõna-tegusõna-nimisõna ehk SVS, tegusõna-nimisõna-nimisõna ehk VSS, nimisõna-sidesõna-nimisõna ehk SJS, asesõna-nimisõna-tegusõna ehk PSV, sidesõna-asesõna-tegusõna ehk JPV, asesõna-tegusõna-nimisõna ehk PVS, tegusõna-omadussõna-nimisõna ehk VAS, nimisõna-tegusõna-määrsõna ehk SVD, nimisõna-nimisõna-tegusõna ehk SSV, nimisõna-tegusõna-tegusõna ehk SVV, asesõna-tegusõna-tegusõna ehk PVV, sidesõna-nimisõna-tegusõna ehk JSV, nimisõna-nimisõna-nimisõna

ehk SSS ja nimisõna-tegusõna-asesõna ehk SVP.

C1 taseme 15 levinumat klastrit on tegusõna-asesõna-nimisõna ehk VPS, asesõna-nimisõna-tegusõna ehk PSV, sidesõna-asesõna-tegusõna ehk JPV, asesõna-tegusõna-tegusõna ehk PVV, nimisõna-tegusõna-nimisõna ehk SVS, nimisõna-sidesõna-nimisõna ehk SJS, asesõna-tegusõna-asesõna ehk PVP, tegusõna-omadussõna-nimisõna ehk VAS, nimisõna-tegusõna-tegusõna ehk SVV, nimisõna-tegusõna-asesõna ehk SVP, sidesõna-nimisõna-tegusõna ehk JSV, nimisõna-tegusõna-määrsõna ehk SVD, nimisõna-nimisõna-tegusõna ehk SSV, tegusõna-nimisõna-nimisõna ehk VSS, nimisõna-nimisõna-nimisõna ehk SSS ja tegusõna-tegusõna-asesõna ehk VVP.

3.3. Klastrite dünaamika

Järgnevalt kirjeldatakse klastrite dünaamikat, st seost klastri ja selle esinemissageduse vahel ning selle põhjal võrreldakse tasemeid omavahel (vt joonis 5). Ühesuguseid klastreid, mis tulid läbivalt esile kõigil neljal keeleoskustasemel, on kokku 23 - SVS, JPV, VPS, SJS, PSV, PVV, VAS, PVS, VSS, SVD, SSV, JSV, PVD, SVV, VDA, SVA, SVP, VVS, SSS, PVP, VSJ, VDS, VVP. Ülejäänud klastrite esinemus varieerub tasemelt teisele liikudes, mõned klastrid kaovad, mõned tulevad juurde.

15 sagedamat klastrit igal tasemel

SVS (nimisõna-tegusõna-nimisõna), nt *suvilas käisid siilid*. A2-tasemel **teisele** kohal (3,54%), B1-tasemel tõuseb **esimesele** kohale (3,03%), B2-tasemel lanegeb **teisele** kohale (2,53%), C1-tasemel langeb **viiendale** kohale (1,98%).

JPV (sidesõna-asesõna-tegusõna), nt *sest see võimaldab*. A2-tasemel **kolmandal** kohal (3,15%), B1-tasemel tõuseb **teisele** kohale (2,94%), B2-tasemel langeb **kuuendale** kohale (2,14%) ja C1-tasemel tõuseb uuesti **kolmandale** kohale (2,33%).

VPS (tegusõna-asesõna-nimisõna), nt *valisin selle eriala*. A2-tasemell **kolmeteistkümnendal** kohal (1,70%), B1-tasemel tõuseb **viiendale** kohale (2,4%), B2-tasemel tõuseb **esimesele** kohale

(2,57%), C1-tasemel on **esimesel** kohal (3,09%).

SJS (nimisõna-sidesõna-nimisõna), nt *vanaema ja vanaisa*. A2-tasemel **neljandal** kohal (2,82%), B1-tasemel langeb **seitsmendale** kohale (2,26%), B2-tasemel tõuseb **neljandale** kohale (2,34%), C1-tasemel- langeb **kuuendale** kohale (1,97%).

PSV (asesõna-nimisõna-tegusõna), nt *selliseid juhtumeid kohtab*. A2-tasemel **kuuendal** kohal (2,38%), B1-tasemel langeb neljateistkümnendale kohale (1,69%), B2-tasemel tõuseb **viiendale** kohale (2,21%), C1-tasemel tõuseb **teisele** kohale (2,43%).

PVV (asesõna-tegusõna-tegusõna), nt *ma läksin õppima*. A2-tasemel **viiendal** kohal (2,39%), B1-tasemel tõuseb **neljandale** kohale (2,45%), B2-tasemel langeb **kaheteistkümnendale** kohale (1,81%), C1-tasemel tõuseb **neljandale** kohale (2,02%).

VAS (tegusõna omadussõna algvõrre nimisõna), nt *seisab valge kapp*. A2-tasemel **üheksandal** kohal (1,97%), B1-tasemel tõuseb **kuuendale** kohale (2,33%), B2-tasemel langeb **kaheksandale** kohale (2%), C1-tasemel jääb **kaheksandale** kohale (1,71%).

PVS (asesõna-tegusõna-nimisõna), nt *ma teen pannkooke*. A2-tasemel **esimesel** kohal (4,1%), B1-tasemel langeb **kolmandale** kohale (2,73%), B2-tasemel langeb **seitsmendale** kohale (2,08%), C1tasemel langeb kahekümne esimesele kohale (1,25%).

VSS (tegusõna-nimisõna-nimisõna), nt *on meenutus noortele*. A2-tasemel **seitsmendal** kohale (2,31%), B1-tasemel langeb **kaheksandale** kohale (2,08%), B2-tasemel tõuseb **kolmandale** kohale (2,38%), C1-tasemel langeb **neljateistkümnendale** kohale (1,43%).

SVD (nimisõna-tegusõna-määrsõna), nt *draamakirjandus on võrdlemisi*. A2-tasemel **üheteistkümnendal** kohal (1,75%), B1-tasemel langeb **kümnendale** kohale (1,78%), B2-tasemel tõuseb **üheksandale** kohale (2%), C1-tasemel langeb **kaheteistkümnendale** kohale (1,57%).

SSV (nimisõna-nimisõna-tegusõna), nt *maailmas sensatsiooni tekitanud*. A2-tasemel

üheksateistkümnendal kohal (1,33%), B1-tasemel tõuseb **üheteistkümnendale** kohale (1,77%), B2-tasemel tõuseb **kümnendale** kohale (1,95%), C1-tasemel langeb **kolmeteistkümnendale** kohale (1,53%).

JSV (sidesõna-nimisõna-tegusõna), nt *ning õhtul külastasime*. A2-tasemel seitsmeteistkümnendal kohal (1,47%), B1-tasemel tõuseb **teisele** kohale (2,94%), B2-tasemel langeb **kolmeteistkümnendale** kohale (1,79%), C1-tasemel tõuseb **üheteistkümnendale** kohale (1,61%).

PVD (asesõna-tegusõna-määrsõna), nt *nad on väga*. A2-tasemel **kaheksandal** kohal (2,38%), B1-tasemel langeb **üheksandale** kohale (2,02%), B2-tasemel langeb üheksateistkümnendale kohale (1,28%), C1-tasemel langeb kuueteistkümnendale kohale (1,4%).
SVV (nimisõna–tegusõna-tegusõna), nt *noored tahavad olla*. A2-tasemel kahekümne neljandal kohal (1,08%), B1-tasemel tõuseb kahekümnendal kohale (1,36%), B2-tasemel tõuseb **üheteistkümnendale** kohale (1,9%), C1-tasemel tõuseb **üheksandale** kohale (1,7%).

VDA (tegusõna-määrsõna-omadussõna algvõrre), nt *loen palju huvitavaid*. A2-tasemel **kaheteistkümnendal** kohal (1,72%), B1-tasemel langeb **kolmeteistkümnendale** kohale (1,71%), B2-tasemel langeb kuueteistkümnendale kohale (1,5%), C1-tasemel langeb üheksateistkümnendale kohale (1,34%).

SVA (nimisõna-tegusõna-omadussõna algvõrre), nt *laulukultuur on ainulaadne*. A2-tasemel **kümnendal** kohal (1,79%), B1-tasemel langeb **kaheteistkümnendale** kohale (1,72%), B2-tasemel langeb seitsmeteistkümnendale kohale (1,36%), C1-tasemel langeb kahekümne kuuendale kohale (1,13%).

SVP (nimisõna-tegusõna-asesõna), nt *järeltulijad mäletavad neid*. A2-tasemel kolmekümne teisel kohal (0,74%), B1-tasemel tõuseb kahekümne esimesele kohale (1,23%), B2-tasemel tõuseb **viieteistkümnendale** kohale (1,52%), C1-tasemel tõuseb **kümnendale** kohale (1,68%).

VVS (tegusõna tegusõna nimisõna), nt *tahan minna vanalinna*. A2-tasemel **neljateistkümnendal**

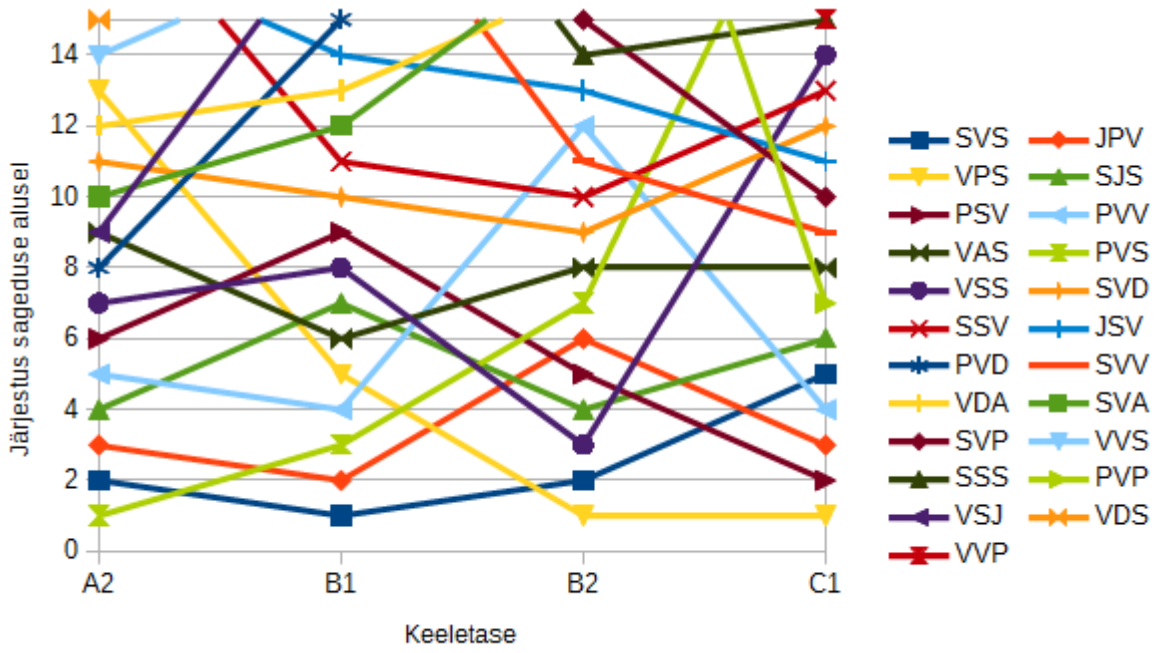
kohale (1,54%), B1-tasemel langeb seitsmeteistkümnendal kohale (1,46%), B2-tasemel langeb kaheksateistkümnendal kohale (1,3%), C1-tasemel langeb kahekümnendal kohale (1,32%).
SSS (nimisõna-nimisõna-nimisõna), nt *Bentoni võit Eurovisioonil*. A2-tasemel kaheksateistkümnendal kohal (1,46%), B1-tasemel langeb kahekümne kolmandale kohale (1,14%), B2-tasemel tõuseb **neljateistkümnendale** kohale (1,65%), C1-tasemel langeb **viieteistkümnendale** kohale (1,42%).

PVP (asesõna-tegusõna-asesõna), nt *mul pole midagi*. A2-tasemel kahekümnendal kohal (1,24%), B1-tasemel langeb kahekümne teisele kohale (1,21%), B2-tasemel langeb kahekümne seitsmendale kohale (1,03%), C1-tasemel tõuseb **seitsmendale** kohale (1,88%).

VSJ (tegusõna-nimisõna-sidesõna), nt *lugesin ajalehest, et*. A2-tasemel **üheksandal** kohal (1,97%), B1-tasemel langeb kaheksateistkümnendale kohale (1,39%), B2-tasemel langeb kahekümne neljandale kohale (1,09%), C1-tasemel langeb neljakümne teisele kohale (0,71%).

VDS (tegusõna-määrsõna-nimisõna), nt *käib juba koolis*. A2-tasemel **viieteistkümnendal** kohal (1,51%), B1-tasemel langeb üheksateistkümnendale kohale (1,37%), B2-tasemel langeb kahekümne esimesele kohale (1,6%), C1-tasemel langeb kahekümne seitsmendale kohale (1,01%).

VVP (tegusõna–tegusõna-asesõna), nt *minna kaitsma oma*. A2-tasemel neljakümne kolmandale kohale (0,51%), B1-tasemel tõuseb kolmekümne viiendale kohale (0,79%), B2-tasemel tõuseb kolmekümnendale kohale (0,99%), C1-tasemel tõuseb **viieteistkümnendale** kohale (1,42%).



Joonis 5. Keeletasemete 15 levinud klastrit

4. Teksti keeletaseme prognoosimise rakendus

Selles osas kirjeldatakse rakenduse arendamiseks kasutatud tehnoloogiaid, rakendust, selle testimise tulemusi ja tehakse järeldus, millises suunas võiks rakendust edasi arendada.

Rakendus on arendatud testkeskkonnas, et mitte häirida korpuse tavapärast kasutamist. Bakalaureusetöö käigus tehtud arendustöö testkeskkonnaks on Tallinna Ülikooli Digitehnoloogiate instituudi testserver greeny.cs.tlu.ee, mille Zope serveris on EVKK testkeskkond.

4.1. Kasutatud tehnoloogiad

Python[20] – laialtlevinud kõrgema taseme programmeerimiskeel

Zope[21] – tasuta ning avatud lähtekoodiga objektorienteeritud veebirakenduse server, mis on kirjutatud Pythonis ning millele on ehitatud EVKK.

Klastrileidja – Sander Otsa loodud rakendus, tekstide klasteranalüüsimiseks

[20] Vt https://www.python.org/
[21] Vt http://www.zope.org/

EVKK morfosüntaktiliselt analüüsitud tekstid – klasteranalüüsi sisend, Klastrileidja otsib süntaksi- ja morfomärgendite, samuti morfosüntaktiliste märgendite lineaarseid järjendeid. Antud töös kasutatakse klasteranalüüsi, mis jaguneb kaheks etapiks:
1) eelmärgenduseta teksti(de) märgendamine,
2) eelmärgendatud tekstide klasteranalüüs Klastrileidjaga.

Esimesel etapil saadetakse tekstid automaatseks märgendamiseks eesti keele morfoanalüsaatoriga ESTMORF[22], seejärel VISL kitsenduste grammatika parseriga[23], milles tekst ühtsustatakse morfoloogiliselt, teostatakse pindsüntaktiline ja sõltuvusanalüüs, selle väljundit kasutatakse Klastrileidjas.

Teisel etapil sisestatakse eelmärgendatud tekst(id) Klastrileidjasse, mis võimaldab otsida nii morfoloogilisi, süntaktilisi kui ka morfosüntaktilisi järjendeid, mille komponentide puhul kattuvad nii vorm kui ka funktsioon lauses (vt Ots 2012). Hiljem on lisatud ka lihtne võimalus sõnaliigijärjendite otsimiseks. Klastrileidja koondab valitud parameetri alusel sarnaste sõnavormide n-gramme läbi järkjärgulise liitmise. N-gramm on n pikkuseline järjestikuliste elementide jada, arvutilingvistikas on jadas keelelised elemendid.. Klastrileidja tuvastab n-gramme libiseva tükeldamise teel, liikudes iga n-järjendi jaoks sõnahaaval edasi. Klastrileidjas saab n-grammi pikkust valida. Klasteranalüüsi tulemus salvestatakse ja eksporditakse CSV (komadega eraldatud väärtuste) failina. Selle sisu kuvatakse ka veebilehel, kus kasutaja saab samuti faili töödelda vastavalt soovile.

4.2. Keeletaseme prognoosimise rakendus

Bakalaureusetöö raames on loodud rakendus, mis prognoosib kasutaja sisestatud teksti keeletaset. Rakendus ennustab morfosüntaktiliselt eelmärgendatud EVKK õppijatekstide keeletaset ja on kasutatav ka mis tahes tekstide analüüsimiseks, kui need eelnevalt märgendada automaatselt töötava eesti keele tarkvaraprogrammide abil.

[22] Vt http://www.filosoft.ee/html_morf_et/
[23] Vt http://beta.visl.sdu.dk/cg3.html

Teksti keeletaseme prognoosimiseks tuleb kasutajal sisestada tavaline tekst või eelnevalt märgendatud tekst ja valida klasterdamise tulemuse parameetrid: millise pikkusega klastreid saada tahetakse ja millisel kujul. Keeletaseme ennustamist parameetrite valik ei mõjuta. Testkeskkonnas töötab rakendus ainult eelnevalt märgendatud tekstiga. Teksti saab sobivale kujule, kasutades EVKK veebilehe vahendusel süntaksianalüüsi[24]. Seejärel läbib sisestatud tekst erinevad funktsioonid.

Kasutaja sisendi funktsioonid

Kõigepealt analüüsitakse tekste Klastrileidjaga. Selleks avatakse esmalt üheksa tekstifaili ja kasutaja teeb valikud analüüsi tulemuste eksplitseerimiseks. Seejärel avatakse üheksa Klastrileidja protsessi.

Esimene protsess: Klastrileidja analüüsib tekste vastavalt kasutaja valitud parameetritele, tulemuseks on kas morfoloogilised, süntaktilised või morfosüntaktilised klastrid. Analüüsi tulemused salvestatakse CSV-faili, analüüsi tulemus kuvatakse veebilehel. Nii failis kui ka veebilehel saab andmeid sortida.

Ülejäänud kaheksa protsessi võib jaotada esmalt kaheks: 1) bigrammide leidmine (klastrianalüüsi tulemuseks on kahest üksusest koosnevad klastrid) ja 2) trigrammide leidmine (kolmest üksusest koosnevad klastrid).

Seejärel võib protsessi jaotada neljaks: 1) morfoloogiliste klastrite analüüs, 2) süntaktiliste klastrite analüüs, 3) morfosüntaktiliste klastrite analüüs ja 4) sõnaliigi klastrite analüüs. Kõikide protsesside analüüsi tulemused salvestatakse esmalt tekstifailidena ja seejärel CSV-failidena, mida kasutatakse hiljem erinevate keeletasemete andmete võrdlemisel.

Teises sisendi funktsioonis on kasutaja sisestatud teksti analüüsimisel saadud tulemuse kuvamine. Esimese sisendi analüüsi tulemused eraldatakse vastavalt klastri pikkusele ridade kaupa alammassiivideks kujul [klastrite hulk, klastri n-is liige, ..., tekst]. Saadud tulemus kuvatakse tulemuste lehel.

[24] Vt http://evkk.tlu.ee/Search/search_reeglid.html

Teksti võrdlemise funktsioonid

Kasutaja sisestatud teksti võrdlemine koosneb kaheteistkümnest funktsioonist, mis jagunevad kolme kategooriasse: 1) failide sisselugemine, 2) sisseloetud andmete töötlemine ja 3) keeletaseme prognoosimine.

Esimeses kategoorias on kaks funktsiooni: esimene bigrammide ja teine trigrammide jaoks.
Teises kategoorias on samuti kaks funktsiooni: esimene bigrammide ja teine trigrammide jaoks.
Kolmandas kategoorias on kaheksa funktsiooni: neli bigrammide ja neli trigrammide jaoks.

Sisestatud teksti keeletaset prognoositakse morfoloogiliste, süntaktiliste, morfosüntaktiliste ja sõnaliigi klastrite alusel. Järgnevalt kirjeldatakse funktsioonide tööd lähemalt.

Failide sisselugemise funktsioonid

Loetakse sisse kasutaja sisestatud teksti bigrammide ja trigrammide alusel saadud klastrite andmetega CSV-failid ja EVKK eseede A2-, B1-, B2- ja C1-keeletaseme bigrammide ja trigrammide CSV-failid.

Sisseloetud andmete töötlemise funktsioonid

Sisseloetud failide sisu viiakse järgmise protsessi jaoks sobivale kujule. Igale reale määratakse järjekorranumber, klastriliikmed ühendatakse, eemaldatakse mittevajalikud tühikud ja alljooned ning moodustatakse nendest massiivi alammassiivid.

Andmete võrdlemise ja teksti keeletaseme prognoosimise funktsioon

Kolmandas funktsioonis viiakse läbi sisestatud teksti klastrite ja keeletasemeid iseloomustavate klastrite võrdlus, selle tulemusena prognoositakse kasutaja sisestatud teksti keeletase. Sisseloetud andmete lõikumise funktsioonist saadakse andmed kasutaja sisestatud teksti klastrite ja keeletasemeid iseloomustavate klastrite kohta.

Iga keeletaseme jaoks luuakse massiiv, millesse moodustatakse alammassiivid kõikide võimalike sisendi klastrite ja keeletaseme klastri paari kohta. Esmalt määratakse alammassiivile esimeseks elemendiks muutuja koht väärtusega -1. Seejärel kontrollitakse kas klastrite nimed on samad.

Nimede ühtimisel muudetakse parameetri väärtus („Koht”) samaks keeletasemele iseloomuliku klastri järjekorranumbriga. Alammassiivi teine element on sisendi klastri järjekorranumbri ja esimese elemendi vahe, kolmandaks elemendiks on klastri nimi.

Keeletaseme massiividest eemaldatakse alammassiivid, mille esimese elemendi väärtuseks jäi -1 ehk klastrid, millel ei olnud samanimelist paarilist. Seejärel mõõdetakse massiivide pikkust, et teada saada mitu alammassiivi massiivis on.

Iga keeletaseme massiivi kohta luuakse uus muutuja. Kui keeletaseme massiivi pikkuseks on 0, määratakse väärtuseks „Puudulikud andmed”, muul juhul määratakse väärtuseks massiivi alammassiivide teise koha elementide absoluutarvuliste väärtuste summa jagatud keeletaseme massiivi pikkusega.

Muutujad lisatakse massiivi „Prognoos” alammassiivide teise koha elementideks, esimeseks elemendiks on „Keeletase“. Massiiv „Prognoos” sorteeritakse teise elemendi järgi, alustades väikseimast ehk klastrite võrdlemise tulemusena kõige tõenäolisemast.

Seejärel antakse funktsiooni väljund. Kui varem mõõdetud keeletaseme klastrite pikkuste summa on 0, väljastatakse „Vigane sisend või puuduvad andmed keeletaseme määramiseks”, muul juhul väljastatakse massiiv „Prognoos”.

Kasutajale kuvatakse tulemuste lehel klastrite andmed, väljastatakse keeletaseme prognoosi tulemused morfoloogiliste, süntaktiliste ja morfosüntaktiliste bi- ja trigrammide alusel.

4.3. Testimine

Rakendust testiti sõnaliigi trigrammide põhjal, saadud tulemust võrreldi eelnevalt hinnatud tekstide keeletasemete trigrammidega. Sisendtekstidena kasutati erinevate keeletasemetesse kuuluvaid EVKK esseesid.

Rakendust testiti sisestatud teksti sõnaliigi klastrite võrdlemisel A2, B1, B2 ja C1 tasemete 15 levinuma sõnaliigi klastriga. Prognoos ja tegelik keeletase ühtisid 32-st katsest 8-l juhul, 9-l juhul oli tegelik keeletase prognoosis teisel kohal.

Seejärel korrati teste kasutades 30 levinumat sõnaliigi klastrit ning seekord ühtisid prognoositav ja reaalne keeletase 9-l korral, 13-l juhul oli reaalne keeletase prognoosis teine valik.

Lõpuks viidi läbi test 200 klastriga, prognoos ja reaalne keeletase ühtisid 6-l juhul, 14-l juhul ühtis keeletaseme valikus teisena asetunud prognoosiga.

Testidest võib järeldada, et kuigi rakenduse prognoosid ei ole eriti täpsed, on tulemust võimalik parandada, kui suurendada veidi võrreldavate klastrite hulka.

4.4. Edasiarendused

Rakendus ei ole praeguseid lahendusi kasutades eriti täpne, ning vajab mõningast edasiarendust ja efektiivsemaks muutmist.

Hetkel kasutatav võrdlemine klastrite sagedust märkivate järjekorranumbrite alusel ei ole just kõige efektiivsem, seda põhjusel, et suur hulk sisendteksti klastritest on sama suurusega, eriti lühemate tekstide ja madalamate keeletasemete puhul. Tekib olukord, kus mahult võrdsed klastrid ei ole järjestuses seetõttu võrdse väärtusega ja eespool asetsevad klastrid mängivad keeletaseme prognoosimisel liiga suurt rolli, muutes prognoosi ebatäpseks. Parem lahendus oleks anda võrdse suurusega klastritele sama väärtus ehk järjekorranumber.

Teise võimalusena oleks mõttekas eraldi analüüsida testitavate tekstide morfoloogilisi, süntaktilisi, morfosüntaktilisi ja sõnaliigi klastreid ning kujundada tasemekirjeldus kõigi nelja analüüsi andmete alusel. Saadud tulemused tuleks siduda sõna, lause ja teksti pikkuse võrdlusandmetega, mis on kasutuses korpuse veebirakendusena.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli luua EVKK klasteranalüüsi vahendeid kasutades rakendus, mis aitab prognoosida kasutaja sisestatud teksti keeletaset. Eesmärgi täitmiseks uuriti, milliseid automaatseid eseede hindamise lahendusi on olemas inglise keele jaoks ja millised on arendamise suunad teiste keelte baasil. Teostati A2-, B1-,B2- ja C2-keeletasemete sõnaliigiliste trigrammide klasteranalüüs ja uuriti lähemalt iga keeletaseme 15 levinuma klastri dünaamikat nimetatud nelja taseme vahel. Bakalaureusetöö eesmärk sai rakenduse loomisega täidetud, kuid testimisel ilmnes, et rakendus ei ole siiski veel piisavalt täpne ning vajab avalikuks kasutamiseks edasiarendamist.

Bakalaureusetöö tulemusena valminud tarkvara oli autoril huvitav arendada ja autor jätkab selle edasiarendamist. Lisaks programmeerimiskogemusele sai autor uusi teadmisi lingvistikast, mis avardab silmaringi, muudab sisulist arusaama uurimisobjektist ja loob informaatikule soodsama aluse keeletehnoloogiliste vahendite arendamiseks.

Kasutatud kirjandus

Attali, Y., Burstein, J. (2004). Automated Essay Scoring with E-rater V.2.0. https://www.ets.org/Media/Products/e-rater/erater_IAEA.pdf (30.04.2016).

Blood, I. (2011). Automated Essay Scoring: A Literature Review. APPLE Award Winning Papers in TESOL & AL. Vol 11, No 2, pp. 40-64.
http://tesol-dev.journals.cdrs.columbia.edu/wp-content/uploads/sites/12/2015/04/4.-Blood-2011.pdf (24.04.2016)

Dikli, S. (2006). An Overview of Automated Scoring of Essays. The Journal of Technology, Learning, and Assessment. Vol. 5, No. 1.
http://files.eric.ed.gov/fulltext/EJ843855.pdf (24.04.2016)

Eslon, P. (2014). Eesti vahekeele korpus. Keel ja kirjandus. 6, lk 436-451. http://kjk.eki.ee/ee/issues/2014/6/507 (25.10.2015).

Hancke, J. (2013). Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language.[Master's thesis] International Studies in Computational Linguistics.
Seminar für Sprachwissenschaft, Universität Tübingen.
http://merlin-platform.eu/docs/MA-Thesis-Julia-Hancke.pdf (24.04.2016)

Ji-young, K. (2014). Predicting L2 Writing Proficiency Using Linguistic Complexity Measures: A Corpus-Based Study. 27 English Teaching, Vol 69, No 4, pp. 27-51.
http://journal.kate.or.kr/wp-content/uploads/2015/01/kate_69_4_2.pdf (24.04.2016)

Vajjala, S., Lõo, K. (2013). Role of morpho-syntactic features in Estonian proficiency classification. In Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8) pp. 63-72, Association for Computational Linguistics.
http://aclweb.org/anthology/W/W13/W13-1708.pdf (24.04.2016)

Vajjala, S., Lõo, K. (2014). Automatic CEFR Level Prediction for Estonian Learner Text.Proceedings of the third workshop on NLP for computer-assisted language learning. NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: pp. 113–127. http://www.ep.liu.se/ecp/107/009/ecp14107009.pdf (24.04.2016)

Östling, R., Smolentzov, A., Tyrefors Hinnerich, B., and Höglin, E. (2013). Automated essay scoring for swedish. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 42–47, Atlanta, Georgia. Association for Computational Linguistics.
http://aclweb.org/anthology/W/W13/W13-1705.pdf (24.04.2016)

Summary

Title: Using Estonian Interlanguage Corpus's Cluster Analysis Tools to Predict Language Level of Text

The purpose of this Bachelor Thesis was to develop a web application that can predict language level of the input text.

First chapter of the thesis provides an overview of linguistic corpuses and gives a closer look of Estonian Interlanguage Corpus.

In second chapter the author gives an overview of different automated essay grading systems, criticism they recieve and developments of automated language level assesment for non-english languages.

Third chapter gives an analysis of top fifteen wordtype clusters of each language level and explains their dynamics inbetween language levels.

Fourth and final chapter gives an overview of the application, it's test results and possible ideas of further developments.

In conclusion the goal of developing an application using cluster analysis tools was met, but the prediction accuracy of language level is too low, meaning the application needs further development.

Lisad

Lisa1

Bakalaureusetöö käigus loodud ja muudetud lähtekoodi failid asuvad tööga kaasas oleval CD-plaadil ja veebis aadressil www.tlu.ee/~virgoh/Bakalaureusetoo,

Failide loetelu:

Search.py – Päringu funktsioonide fail, selles failis on ka rakenduse funktsioonid.

klaster.pt – Tekstisisestamise ja parameetrite valiku leht.

klaster_vastus.pt – Analüüsi ja prognoosi tulemuste leht.