Teksti keerukusest

22.11.2019

Pille Eslon

Teksti keerukuse määramine formaalsete tunnuste alusel: sõnavorm, lause, tekst

- Sõnavormide arv tekstis (keskmine, absoluutarv)
  - pikim sõnavorm (tm)
  - lühim sõnavorm (tm)
  - keskmise pikkusega sõnavorm (tm)
- Lausete arv tekstis (keskmine, absoluutarv)
  - pikim lause
  - lühim lause
  - keskmise pikkusega lause

- Erineva pikkusega sõnavormide osakaal tekstis (arvutatud %, vahemikus 2-20 tm)
  - 2 tähemärki pikk
  - …..
  - 20 tähemärki pikk
- Erineva pikkusega lausete osakaal tekstis (arvutatud %, vahemikus 2-20 sõnet)
  - 2 sõnavormi pikk
  - ……
  - 20 sõnavormi pikk

Näide 1. A2-taseme eesti keele õppija: sõnavormi, lause ja teksti pikkus

	Üld-andmed	Keskmine sõna-vormide arv tekstis	Keskmine lausete arv tekstis	Lühim sõna-vorm (tm)	Pikim sõna-vorm (tm)	Keskmine sõna-vormi pikkus (tm)
Päring 1 (vene emakeel)	28 teksti 3660 sõna	130,71	13,75	1,71	14,0	5,17
Päring 2 (erinevad ema-keeled)	35 teksti 4864 sõnet	138,97	15,26	1,69	14,23	5,16

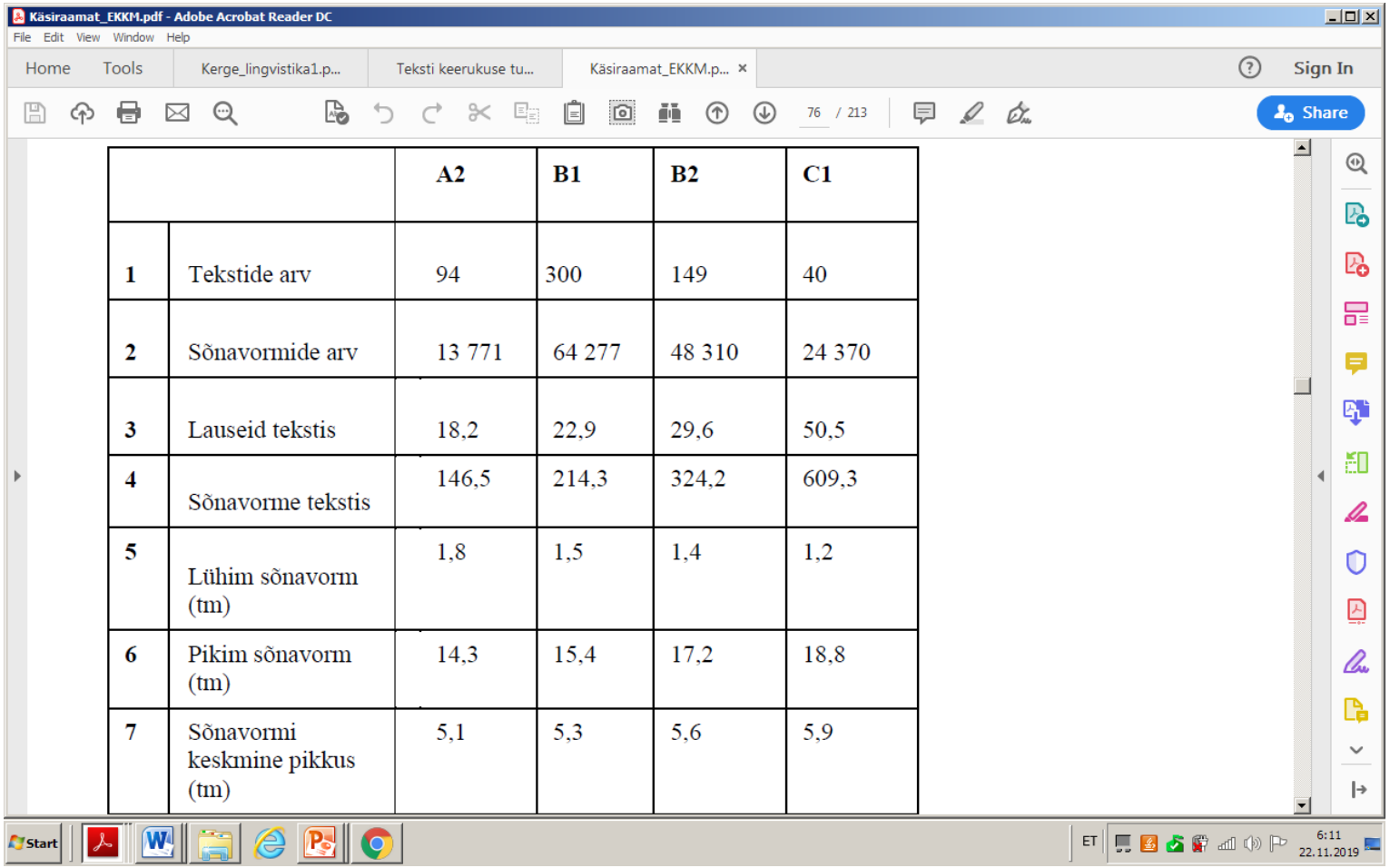
Erineva pikkusega sõnavormide osakaal A2-taseme tekstides

	2 tm	3 tm	4 tm	5 tm	6-9 tm	10-20 tm
Päring 1 (vene emakeel)	17,49%	11,29%	19,6%	13,08%	26,62%	7,21%
Päring 2 (erinevad ema-keeled)	17,78%	10,43%	19,15%	13,88%	27,27%	6,79%

Erineva pikkusega lausete osakaal A2-taseme tekstis

	2 sõna-vormi	3 sõna-vormi	4 sõna-vormi	5 sõna-vormi	6-9 sõna-vormi	10-20 sõna-vormi
Päring 1 (vene emakeel)	0,65%	6,19%	7,06%	11,34%	45,99%	33,42%
Päring 2 (erinevad ema-keeled)	0,64%	6,0%	8,23%	12,17%	45,99%	30,73%

Teksti keerukuse määramine formaalsete tunnuste alusel: sõnavorm, lause, tekst



		A2	B1	B2	C1
1	Tekstide arv	94	300	149	40
2	Sõnavormide arv	13 771	64 277	48 310	24 370
3	Lauseid tekstis	18,2	22,9	29,6	50,5
4	Sõnavorme tekstis	146,5	214,3	324,2	609,3
5	Lühim sõnavorm (tm)	1,8	1,5	1,4	1,2
6	Pikim sõnavorm (tm)	14,3	15,4	17,2	18,8
7	Sõnavormi keskmine pikkus (tm)	5,1	5,3	5,6	5,9

Keelekasutuse mustrid

- Sarnaste bigrammide, trigrammide ja tetragrammide esinemus
  - absoluutarvudes
  - protsentides
  - n-grammide optimaalsuskoefitsient $k \geq \sqrt{n} : 2$ (valimi optimaalne n-grammide hulk ehk n-grammide hulga olulisuse piirväärtus)

– erinevate valimite statistiliselt olulise sarnasuse/erinevuse sümmeetriline mõõdik, mis üles ehitatud log-tõepära funktsioonile
  - Nt Paul Rayson, vt http://ucrel.lancs.ac.uk/llwizard.html ja hii-ruut kalkulaator http://stattrek.com/online-calculator/chi-square.aspx
  - keeleandmete analüüsi puhul valitakse vea protsendiks tavaliselt 5%, st $LL \geq 3{,}84$ ($p < 0{,}05$); kasutatud ka 10% ja 1%.

Lingvistilised tunnused

- sõnaliikide jaotumine tekstis (nimi-, omadus-, ase-, arvsõnade vs. tegusõnade ja määrsõnade osakaal %)
  - konkreetsete/abstraktsete nimisõnade jagunemine (%)
  - terminoloogilise ja üldkeelse sõnavara suhe (%)
  - põhisõnavara eristamine üldsõnavarast (Sirts & Võhandu 2009. Korpuiste tükeldamine: rakendusi silpide ning allkeeltega. – Eesti Rakenduslingvistika Ühingu aastaraamat 5, lk 251–266)
  - aktiivse (sageli) kasutatava sõnavara ja grammatika eristamine passiivsest (harva) jne

Jaan Mikk & teksti keerukuse mõõtmine

- Tunnused:
  - Fraasi keskmine pikkus tm
  - Lause keskmine pikkus tm
  - Iseseisva lause keskmine pikkus sõnades
  - Iseseisva lause keskmine pikkus tm
  - Sõna keskmine pikkus tm
  - 10 ja enam tm sisaldavate sõnade %
  - 20 ja enam tm sisaldavate sõnade %
  - VL-s esinevate sõnade %
  - Erinevate tundmatute sõnade %
  - Korduvate nimisõnade keskmine abstraktsus

Õpetajate Leht 1. august 2003

Kas tekst sobib õpikusse?

Teksti keerukuse arvutamise valem (J. Mikk):

$K = 0{,}131\ X_9 + 9{,}84\ X_{22} - 4{,}59$

K - teksti keerukuse näitaja

X_9 - iseseisva lause keskmine pikkus tm

X_{22} korduvate nimisõnade keskmine abstraktsus

Liiga keeruline tekst on korrelatsioonis õppeedukuse langusega:

- raske aru saada, ei tekita huvi, puudub motivatsioon lugeda jne

Teksti keerukuse määramise lingvistiliste tunnuste põhjal

- Keerukas lausestruktuur, pikad laused
  - K. Kerge (2002) näide – nominalisatsioon, *mine*-tuletised, kiilud jm lisainfot edastavad lõigud ajakirjandus- ja ilukirjandustekstide lausestuses
    - Formaalsed tunnused – kirjavähemärgid (nende olemasolu või puudumine)
- Tundmatu sõnavara (arhaismid, historismid, teadusterminid, slängisõnad, eriala metakeel jm)
- Sõnatähenduse abstraktsus

Jaan Miku näiteid (1)

Eespool juba mainiti, et Rutherford tegi radioaktiivse lagunemise uurimisel katseliselt kindlaks radioaktiivsete ainete aktiivsuse ajast sõltuvuse iseloomu – radioaktiivse lagunemise põhiseaduse.

Rutherford tegi katseliselt kindlaks, et radioaktiivsete ainet aktiivsus väheneb aja jooksul. Nii näiteks väheneb radooni aktiivsus iga 10 minutiga kaks korda.

Jaan Miku näiteid (2)

Jõudu, millega maa tõmbab keha enda poole antud kohas, nimetatakse raskusjõuks.

Raskusjõuks nimetatakse jõudu, millega maa tõmbab keha enda poole.

Müüa import meeste pükse.

Müüa meeste importpükse.

Jaan Miku näiteid (3)

Kõrvuti asetsevate koodonitega seostunud tRNA molekulide otste küljes asuvate aminohapete vahele sünteesitakse ribosoomis oleva ensüümi kaasabil peptiidside.

???

Mõtlemine areneb kaasasündinud reflekside baasil assimilatsiooni ja akommodatsiooni abil.

???

Jaan Miku näiteid (4)
Pikad laused lihtsamaks!

Anatoomia õpik, 33 sõna	Toimetatud tekst
(1) Hingetoru valendik on alati lahti ning **(2)** ükskõik millises asendis meie keha ka ei viibiks, **(3)** läheb õhk hingetorust vabalt läbi, **(4)** kuna selle seintes asetsevad kõhrest poolrõngad (hingetorukõhred), **(5)** mis on omavahel ühendatud sidemete ja lihastega.	**(4)** Hingetoru seintes on poolrõngakujulised kõhred. **(5)** Need on omavahel ühendatud sidemete ja lihastega. (1) Kõhred hoiavad hingetoru avatuna. **(3)** Seetõttu pääseb õhk hingetorust alati vabalt läbi.

Reinsalu 2012 näiteid

Originaal	Toimetatud
Leping kuulub täitmisele Poolte õigusjärglaste poolt samadel tingimustel.	Poolte õigusjärglased täidavad lepingut samadel tingimustel.
Tähtaegadeks mittetasumisel kuuluvad töövõtja omandis olevad materjalid töövõtja poolt demonteerimisele.	Kui summat ei tasuta ettenähtud ajaks, on töövõtjal õigus demonteerida tema omandis olevad materjalid.
Käesoleva lepingu alusel tasumisele kuuluvad summad võib toetuse saaja eest välja maksta Sihtasutus Keskkonnainvesteeringute Keskus (KIK).	Lepingu alusel tasutavad summad võib toetuse saaja eest välja maksta sihtasutus Keskkonnainvesteeringute Keskus (KIK).

BA-töö näide: Vajak 2019: 12-13

Antud peatükis keskendutakse andmeanalüüsi meetoditele, mida varasemates töödes kasutati Milliseid masinõppe algoritme ja klassifikatsiooni tüüpe on kasutatud.

Üks levinud masinõppe algoritm oli SVM (*support vector machine*) masinat, mis arvutab välja iga mustri jõudluse. SVM masinat kasutati *radial basis kernel*-iga, kuna teatakse, et see toimib kõige kindlamalt (Giot & Rosenberger, 2012). Kasutati kahte SVM-i rakendust, esimeseks treenitakse SVM-i treenimise mustritega, milleks on klahvivajutuse peiteaeg.

Teiseks kasutatakse SVM-i koos Gaussian *radial basis* funktsiooni *kernel mapping*-uks. Väärtuste uuendamiseks kasutati kuut erinevat meetodit: *gradient descent with adaptive learning rate, conjugate gradient backpropagation with Fletcher-Reeves updates, BFGS quasi-Newton method, one-step secant backpropagation, scaled conjugate gradient backpropagations ja Levenberg-Marquardt backpropagation* (Uzun, Bicakci, & Uzunay, 2016). SVM kõrval kasutati veel nelja populaarset algoritmi binaarse klassifikatsiooni jaoks: logistiline regressioon, lähim naaber, C4.5 ja Random Forest (Pentel, 2017).

Probleemid

- Eestikeelne terminoloogia
  - *support vector machine*
  - *radial basis kernel*
  - *radial basis*
  - *kernel mapping*
  - Meetodid: *gradient descent with adaptive learning rate, conjugate gradient backpropagation with Fletcher-Reeves updates, BFGS quasi-Newton method, one-step secant backpropagation, scaled conjugate gradient backpropagations ja Levenberg-Marquardt backpropagation*

- Eriala terminoloogia, nt
  - binaarne klassifikatsioon
  - logistiline regressioon
  - masinõppe algoritmid
  - klahvivajutuse peiteaeg
  - C4.5 ja Random Forest

Kirjandust

- Kerge, Krista 2002. Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX sajandil. Tallinn.
- Mikk, Jaan 2002. Kuidas hinnata õppetekstide keerukust. Õpetajate Leht, 1.08.2002; Mikk 2012. Õppeteksti keerukus.
- Reinsalu, Riina 2012. Lepingukeele keerukus. – Õiguskeel 2.
- Vajak, Henri 2019. Põlvkondade interaktsioonide erinevused arvutite sisendseadmete kasutamisel. Bakalaureusetöö. Tallinn: TLÜ digitehnoloogiate instituut.