Esseede analüüs: teksti keerukus

20.11.2020

Kaisa Norak, Pille Eslon

I. Teksti üldise keerukuse määramisest

- Arvandmete põhjal
  - Sõnavormide ja lausete arv tekstis (keskmine, absoluutarv)
  - Erineva pikkusega sõnavormide osakaal tekstis (arvutatud %, vahemikus 2-20 tm)
  - Erineva pikkusega lausete osakaal tekstis (arvutatud %, vahemikus 2-20 sõnet)

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

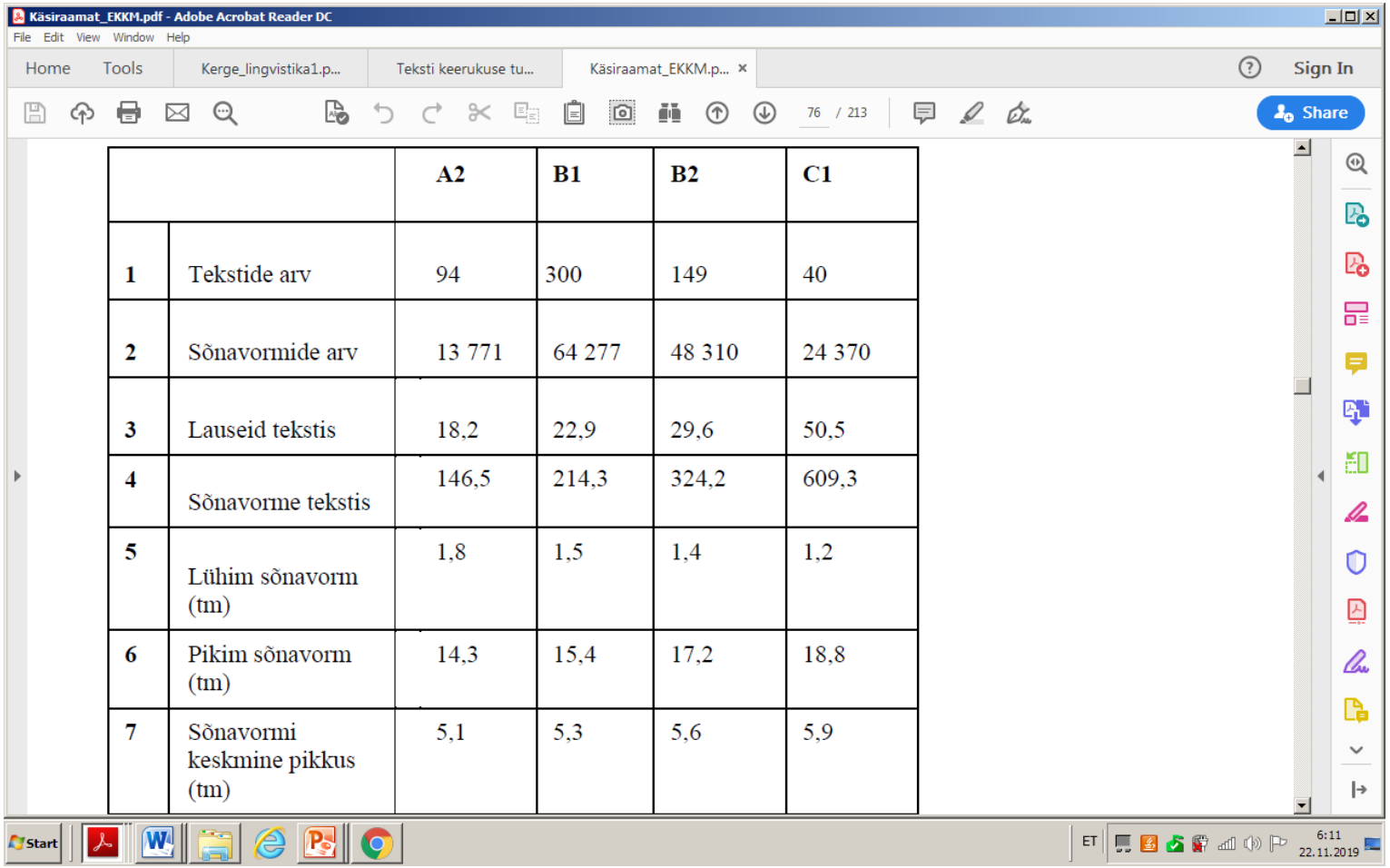
Näide 2. Erineva pikkusega sõnavormide osakaal A2-taseme tekstides

	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%

Näide 3. 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%

Näide 4. Teksti üldine keerukus: ülevaade keeleoskustasemetest



		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

II. Teksti keeleoskustaseme automaathindaja

- Rakenduse demoversioon: vt

https://github.com/centre-for-educational-technology/evkk/wiki/Demos

ja

http://minitorn.tlu.ee/~jaagup/oma/too/20/09/tasemed2.php)

- prognoosib eestikeelse teksti vastavust riiklikult hinnatavatele keeleoskustasemetele: A2 - esmane keeleoskus, B1 - suhtluslävi, B2 - edasijõudnu, C1 - vaba suhtlus
- praegu võtab rakendus hindamisel arvesse kolme mõõdet:
  - teksti üldine keerukus (teksti, sõnade ja lausete pikkus);
  - sõnavara mitmekesisus, ulatus, tihedus ja abstraktsus;
  - vormikasutus (sõnaliikide ja muutevormide osakaalud tekstis ning nende rohkus või esindatus absoluutarvudes)

Näide keeleoskustaseme automaathindaja tööst

- **Sisend: kirjalik tekst**
- **Väljund:**
  - **Tekst vastab tasemele:**
    B2
    Tõenäosus: 62%
  - **Teiste tasemete tõenäosus:**
    - **B1**: 27%
    - **A2**: 11%
    - **C1**: 0%

- **Teksti üldine keerukus:**
  B1 (tõenäosus 84%)
  Arvesse on võetud üldise keerukuse tunnused: teksti, sõnade ja lausete pikkus.
- **Vormikasutus:**
  B2 (tõenäosus 97%)
  Arvesse on võetud sõnaliikide ja muutevormide osakaalud ning sõnade vormirohkus.
- **Sõnavara:**
  A2 (tõenäosus 100%)
  Arvesse on võetud sõnavaliku mitmekesisus ja ulatus (unikaalsete sõnade hulk, harvem esineva sõnavara osakaal), sõnavara tihedus (sisusõnade osakaal) ja nimisõnade abstraktsus.
- **Koondhinnang (B2)** põhineb nii sõnavara-, vormikasutuse kui ka üldise keerukuse tunnustel.

III. Teksti keelekasutusmustrite leidmine ja nende keerukuse analüüs

- Keelekasutusmustrid leitakse automaatselt
  - tekstitöötlus (tekstikaeve) Klastrileidja rakenduse abil, tulemuste kvalitatiivne tõlgendamine
- Sisend: eesti keele süntaksianalüsaatoriga automaatselt märgendatud tekst
- Väljund: keelekasutusmustrid
  - nt kahest, kolmest, neljast komponendist koosnevad regulaarsed sõnaliikide järgnevused tekstis (n-grammid)

Mustrite esinemus

- Kahest, kolmest, neljast komponendist koosnevate mustrite 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 ≥ 3,84 ($p < 0,05$); kasutatud ka 10% ja 1%.

B1-taseme keelekasutusmustri näide

Soomekeelsed õppijad	Venekeelsed õppijad	Emakeelekõnelejad Trainis, Allkivi (2014: 286) andmetel
Rõhusõnad *nii* ja *väga*	Tüüpiliselt rõhusõna *väga*, mõnel korral *suhteliselt*, nt *väga tore et*, *väga hea ja*, *suhteliselt huvitav ja*	Muster ei kuulu statistiliselt oluliste hulka
Rikkalik valik omadussõnu		
Tavaliselt rinnastav sidesõna *ja*, mõnel juhul rinnastav-alistav *kui*, harva alistavad sidesõnad *et*, *sest*	Tavaliselt rinnastav sidesõna *ja* (harva *ning*), harvem alistavad sidesõnad *et*, *sest*	

sm – *nii rahulik* (*huvitav*, *pime*, *tore* jne) *ja*, *väga lahja* (*tark*, *ilus*, *vaikne* jne) *ja*, *eriti huvitav ja*, *ka tore kui*, *niisama soe ja*; vn – *väga tore et*, *väga hea ja*, *Muidugi õnnelik ning*, *ka tark ning*, *nii väärtuslik kui*. See muster on tulnud mõlema rühma õppijatel esile B1-tasemel, esinemus soomekeelsete õppijate tekstides on keskmine (11%) ja venekeelsetel pigem väike (6%).

Näide esseest

Muster nimisõna-sidesõna-nimisõna: *õppimist ja roboteid, juhtimist ja programmeerimist, õppimist ja õpetamist; saamist ja laps; loogika ja loovus; mõtlemine ja eeldused, teadus ja tehnika, süle- ja tahvelarvutid, süle- või tahvelarvuti, tehnika ja infotehnoloogia, meeskonnatööoskus ning loovus; mängimine ja robootikat; informatikaga sest roboteid; tehnoloogiaga ja informaatikaga* jne

Mustrite sõnavara rikkus

Sõnavormid	Sõnaliigid	Näited
Essee 1: 1022 sõnavormi	464 nimisõna, 196 tegusõna, 98 omadussõna, 85 määrsõna, 83 sidesõna, 65 asesõna, 12 kaassõna, 6 arvsõna	Lapsed, lasteaiad, (mini)robotid; olema, kasutama, arenema, arendama, õppima,saama, võima, programmeerima; erinev, mänguline, uus, loogiline; palju, väga, kõige; ja (ning, et); see, need; abil, läbi

IV. Kvalitatiivne lingvistiline analüüs: teksti keerukuse tunnused

- sõnaliikide jaotumine tekstis (nimi-, omadus-, ase-, arvsõnade vs. tegusõnade ja määrsõnade osakaal %)
  - konkreetsete/abstraktsete nimisõnade jagunemine (%)
  - terminoloogilise ja üldkeelse sõnavara suhe (%)
  - aktiivse (sageli) kasutatava sõnavara ja grammatika eristamine passiivsest (harva) 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, uued sõnad, terminid, võõrsõnad, slängisõnad, eriala metakeel jm)
- Sõnatähenduse abstraktsus

Esseede keerukuse hindamine

- Küsimused:
  - Kas tegemist on loogiliselt üles ehitatud arusaadava tekstiga?
  - Kas käsitletav küsimus on põhjendatult fookuses? Kas autori eesmärk on saavutatud?
  - Kas tekst sisaldab asjatut infomüra? Millest see on tingitud?
    - Sisu tundmine ja piiritlemine
    - Sõnastamisoskus, keeleline korrektsus, sõnakasutuse täpsus, võõrpäritolu sõnad ja terminid

Kirjandust

- Kerge, Krista 2002. Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX sajandil. Tallinn.