

Eesti keele sõnavaratest LexEst

Kaidi Lõo¹, Katrin Leppik¹, Anton Malmi^{1,2}, Agu Bleive^{1,3}, Raymond Bertram⁴

¹Tartu Ülikool, ²Lancasteri Ülikool, ³Pekingi Välisõpingute Ülikool, ⁴Turu Ülikool

Eesti Rakenduslingvistika Ühingu kevadkonverents

25.04.2025

Kava

- Taust
 - Motivatsioon
 - Sõnavara ja keeleoskuse seos
 - Sõnavaratestid eesti ja teistes keeltes
- LexEsti koostamisest ja valideerimisest
 - Testsõnavara valik
 - Andmete kogumine ja testi valideerimine
 - Seosed LexEsti testi tulemuste ja küsimustiku vahel
- Kokkuvõte

Taust: motivatsioon eesti keele sõnavara testi loomiseks

- Keeleõppijate keeleoskuse hindamiseks kasutatakse tihti küsimustikke
- Enda keeleoskuse hindamine on subjektiivne ja keeruline
- Tundsime puudust objektiivsest ja kiirest keeleoskuse hindamine vahendist L2 teadusuurimuste tingimustes

Taust: olemasolevad eesti keele L2 testid

- Eesti keele tasemeeksamid (Haridus- ja Noorteamet)
- Eesti keele tasemetestid (sõeltestid, diagnostikatestid, Integratsiooni Sihtasutus)
- Eesti Keele Instituut on arendanud mitmeid materjale ja õppevahendeid (nt sõnavaraloendid keeletasemete kaupa)
- KaLaKe - sõnavaratest kakskeelsetele keelepuuetega lastele (Vaks et al., retsenseerimisel)

Taust: keeleoskus ja sõnavara

- Mitmed uurimused on leidnud, et L2 sõnavara suurus peegeldab hästi üldist keeleoskust (Read, 2000; Zareva *et al.*, 2005)
- Sõnavara peegeldab üldist keeleoskust kõigil keeleoskustasemetel (A1-C1) (Milton, 2010)
- Neljast osaoskusest korreleerub sõnavara oskus kõige enam lugemis- ja kirjutamisoskusega (Staehr, 2008; Milton, 2010; Miralpeix & Muñoz, 2018)
- Sõnavara ja keeleoskuse seoseid on uuritud peamiselt inglise keele näitel

Levinud sõnavaratestid inglise keeles

- **Vocabulary Levels Test** (VLT, Nation, 1983)
 - Sõnavara 4 tasemega, iga tase koosneb erineva esinemissagedusega sõnadest
 - Valikvastustega test
 - Sõnale tuleb valida õige definitsioon, nt *pencil - something used for writing*
- **Eurocentres Vocabulary Size Test** (EVST, Meara & Jones, 1988)
 - Test on jaotatud 10 plokki, igas plokkis erineva esinemissagedusega sõnu
 - Tuleb otsustada, kas tegemist on ingliskeelse sõnaga (nt **weekend**) või väljamõeldud sõnaga (nt **klimp**) (ingl *visual lexical decision task*)

LexTALE (vt täpsemalt <https://www.lextale.com/>)

- LexTALE ehk *Lexical Test for Advanced Learners of English* (Lemhöfer & Broersma, 2012)
- Mõeldud edasijõudnud õppijatele
- Kasutatakse laialdaselt püshholingvistika uurimustes
- Tuleb eristada päris sõnu (nt **hurricane**) väljamõeldud sõnadest (nt **kermshaw**) (40 sõna ja 20 väljamõeldud sõna)
- Koosneb 60 testsõnast ja võtab aega u 5 minutit

kermshaw

no

yes

hurricane

no

yes

LexTALE'i kohandamine teistele keeltele ja keeletasemetele

- Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English (Lemhöfer & Broersma, 2012)
- Assessment Tool for Italian L2 Receptive Vocabulary Size (Amenta et al., 2021)
- Lexize: A test to quickly assess vocabulary knowledge in Finnish (Salmela et al. 2021)
- Lextale FR a fast, free, and efficient test to measure language proficiency in French (Brysbaert, 2013)
- Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size (Izura, Cuetos & Brysbaert, 2014)
- LEXTALE CH: A quick, character-based proficiency test for Mandarin Chinese (Chan & Chang, 2018)
- LexCHI: A quick lexical test for estimating language proficiency in Chinese (Wen et al., 2024)
- LextPT: A reliable and efficient vocabulary size test for L2 Portuguese proficiency (Zhou & Li, 2022)
- LexMAL: A quick and reliable lexical test for Malay speakers (Lee et al., 2024)
- LexArabic: A receptive vocabulary size test to estimate Arabic proficiency (Alzahrani 2024)

LexEsti esialgne variant

- LexEst tehtud algse inglise, soome jt sõnavaratestide (Lemhöfer ja Broesma 2012, Salmela jt. 2021) eeskujul koostöös Turu ülikooliga
- Andmete kogumine Turu ülikooli õppeplatvormil ViLLE (<https://ville.utu.fi/>; mai 2023-okt 2023)
 - **Taustaküsimustik** (küsimused vanuse, soo, enesehinnangute keeletasemele, eesti keele motivatsiooni ja kasutuse ja läbitud eesti keele kursuste kohta jne)
 - **Sõnavaratest** (kas ekraanil olev sõna on päris eesti keele sõna vt. Löö jt 2018)

Algne testsõnavara valik 1

- Stiimulsõnad Tasakaaluskorpuse põhjal tehtud sõnasageduslistist
- **90 sõna** kuuest erinevast sagedusvahemikust
- 50 nimisõna (nt **suusk**), 24 omadussõna (nt **lauge**), 16 verbi (nt **tikkima**)
- 3-8 tähemärki
- Vältisime murde-, sõimu-, liitsõnu, läbipaistvate tuletusliidetega sõnu, mitmetähenduslikke sõnu

Algne testsõnavara valik 2

- **45 pseudosõna**

- Tasakaaluskorpusest 45 uut sarnase sageduse ja pikkusega sõna
- Muutsime 1-3 tähte, et teha väljamõeldud pseudosõnu (nt **keep** > **teep**)
- 3-7 tähemärki
- Kontrollisime, et pseudosõnad ei oleks päris sõnad teistes tuntumates keeltes (nt inglise, soome, saksa keeles) ega eesti keele murretes

Testiekraan 1

SOOV

yes

S

no

L

Testiekraan 2

umuk

yes

S

no

L

Osalejad

- 485 osalejat (vanus 18–77 aastat; 348 naist, 129 meest, 8 mittebinaarset).
 - 274 olid eesti keelt emakeelena kõnelejad (L1)
 - 211 eesti keelt teise keelena kõnelejad (L2)
 - Erineva keeletasemega, 24 erineva emakeelega

Table 1: Distribution of L2 speakers across proficiency levels.

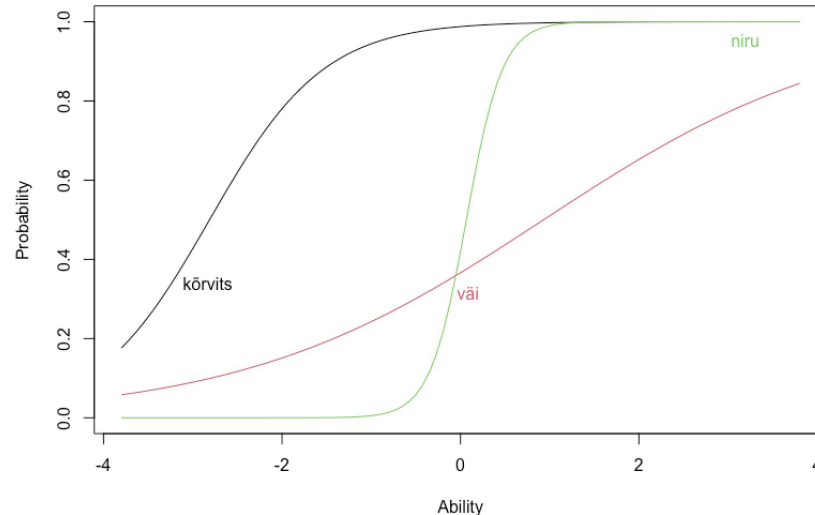
L2 proficiency level	Number of L2 speakers
A1	26
A2	22
B1	31
B2	38
C1	40
Total	157

Table 3: Distribution of native languages of L2 speakers.

Language	Number of L2 speakers
Russian	110
German	26
Finnish	17
English	11
Ukrainian	8
Latvian	4
Hungarian	4
Swedish	4
French	3
Polish	3
Dutch	3
Turkish	2
Spanish	2
Portuguese	2
Italian	2
Chinese	2
Tatar	1
Norwegian	1
Nepali	1
Komi	1
Hausa	1
Bulgarian	1
Azerbaijani	1
Arabic	1

Analüüs: algse testsõnavara valideerimine ja kitsendamine

- Algse testsõnavara raskusastme ja eristusvõime mõõtmine üksikvastuste teooria (ingl. *item response theory*) analüüs (Rizopoulos 2006):
- Lõplik LexEst testi sõnavara kitsendamine 135 stiimulilt 90 stiimulile (60 sõna ja 30 pseudosõna)

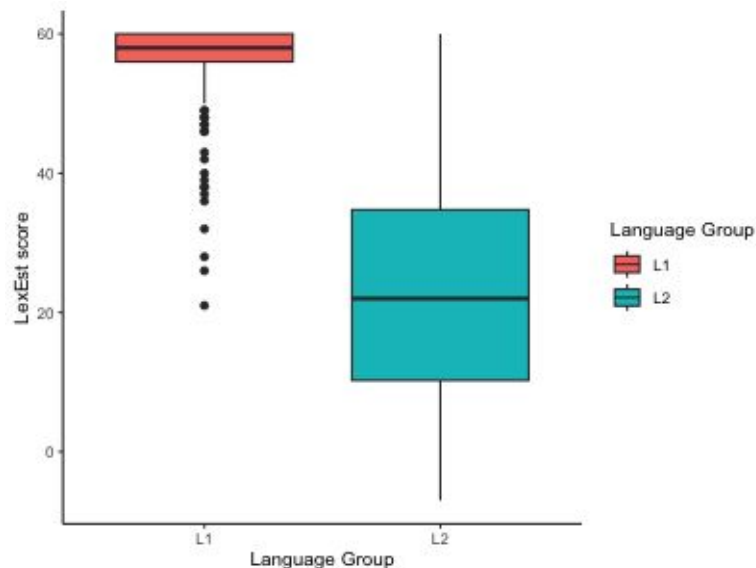


L1 ja L2 kõnelejate punktisummad

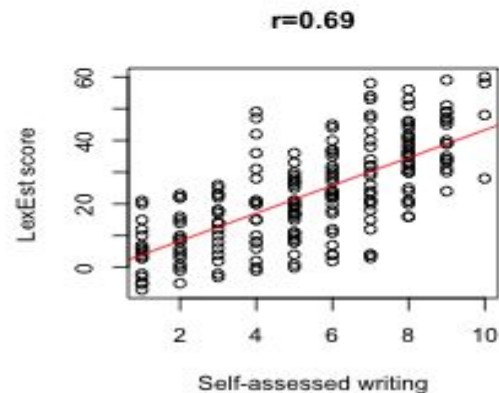
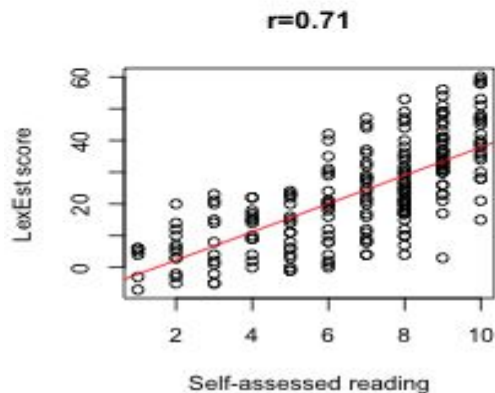
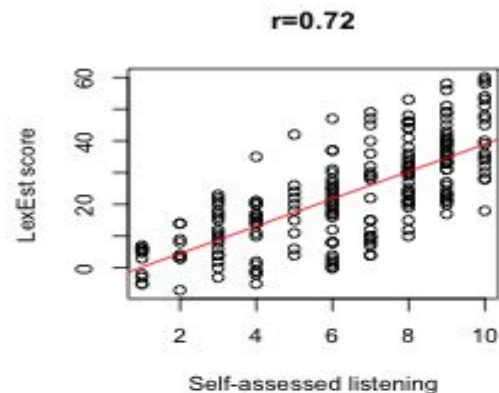
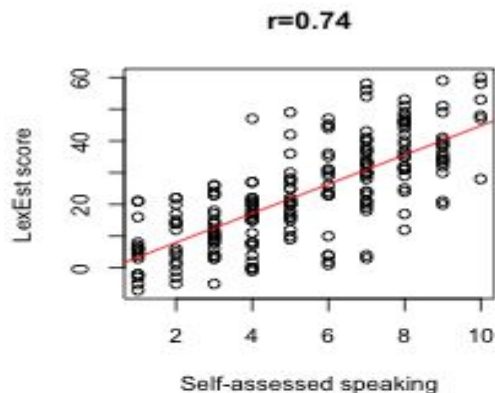
- Õigete ja valede vastuste põhjal punktisumma arvutamine (vt Lemhöfer ja Broesma 2012; Brysbaert 2013), max 60 punkti

LexEst Score = N Yes to words - 2 * N Yes to pseudowords

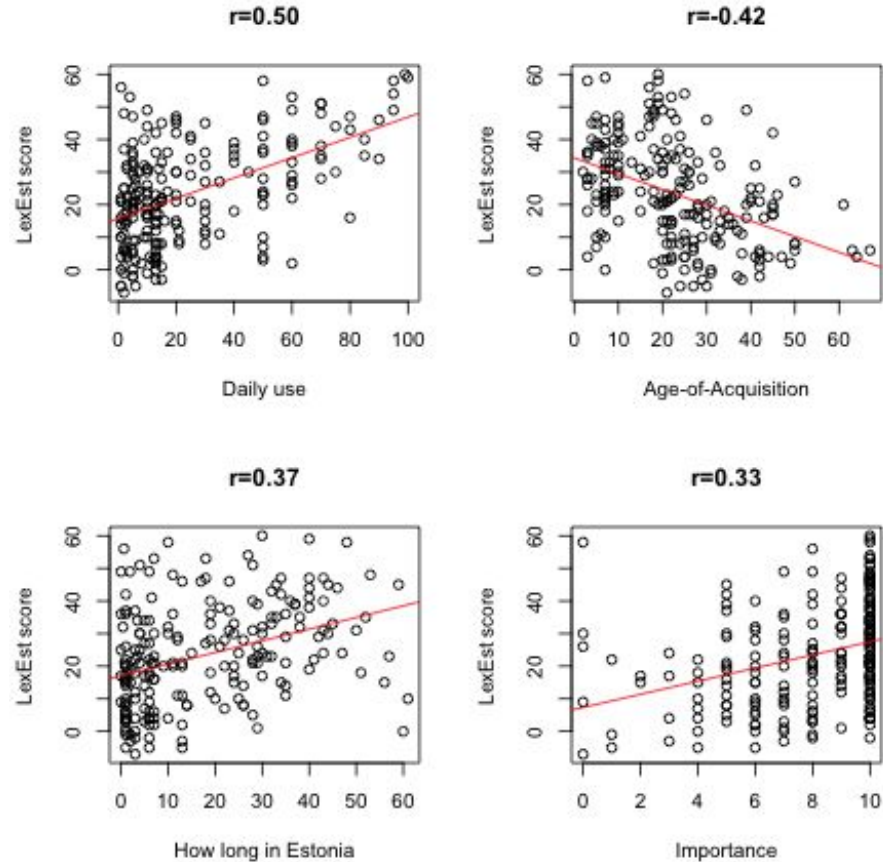
- Meie uurimuses
 - L1 vahemikus 20 kuni 60 punkti
 - L2 vahemikus -7 kuni 60 punkti



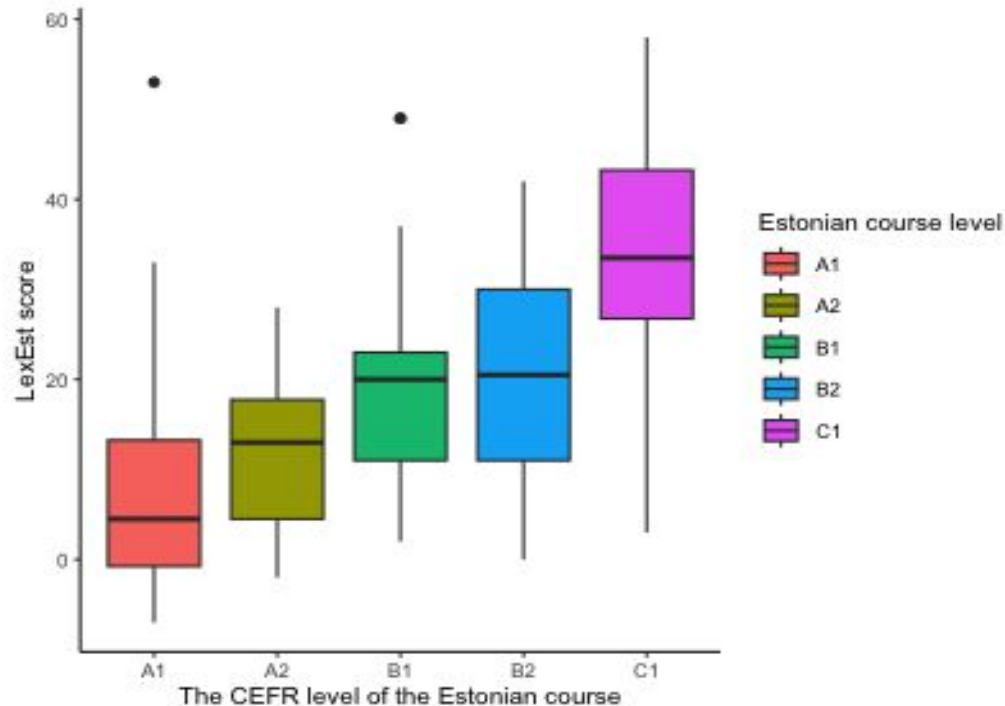
LexEsti L2 testitulemuste korrelatsioonid enesehinnangutega



LexEsti L2 tulemuste korrelatsioonid keelekasutusega



LexEsti L2 tulemuste korrelatsioonid eesti keele kursuse tasemega



LexEst kokkuvõte ja tulevikusuunad

- . Artikkel LexEsti koostamisest ja valideerimisest hetkel retsenseerimisel ajakirjas Applied Psycholinguistics
- . Erinevad eesti L2 uurimustes LexEst juba kasutavad (vt. Lutter 2025, Mackiewicz 2025)
- . Võrukeelse testi koostamine?
- . Testi adapteerimine eesti (kooli)laste sõnavara mõõtmiseks?
- . Eesti-soome kakskeelsete sõnavaratest?

LexEst veebis ja paberil: <https://lexest.ut.ee/>

The LexEst logo consists of a blue rectangular background with a white diagonal stripe running from the top-left corner to the bottom-right corner. The word "LexEst" is written in white, sans-serif font, centered on the blue background.

megel

Tere tulemast!

LexEst on **kire** ja **lühike sõnavaratest**, mille abil on võimalik hinnata eesti keele õppijate üldist keeleoskust. Õppijate enda hinnangud keeleoskusele pole tihti täpsed, kuna enda oskusi kiputakse üle- või alahindama. Tesmetestid on väga mahukad, nende hindamine on keeruline ja võtab kaua aega. Varasemad uurimused (Milton, 2010; Laufer ja Ravenhorst-Kalovski, 2010) on aga näidanud, et õppijate **sõnvara teadmised annavad üpris hea ülevaate nende üldisest keeleoskusest**. Seega oleme LexTALE1 (Lemhöfer ja Broersma, 2012) eeskujul koostanud ka eesti keele jaoks sõnvara tundmisel põhineva keeleoskustesti.

LexEst seisneb eesti keeles olevate **sõnade eristamisest** väljendatud sõnadest ehk **pseudosõnadest**. Pseudosõnad on sõnad, mis on oma struktuurilt sarnased eestikeelsetele sõnadega, kuid millele pole omistatud tähendust. Testi tegija näeb ekraanil sõna ja peab otsustama, kas see on tema arust päris eesti keele sõna või väljendatud sõna. Kokku esitatakse testi sooritajale 90 sõna, millest 60 on eesti keele sõnad ja 30 pseudosõnad. **Testi tegemine võtab aega umbes 5 minutit**. Tulemuste salvestamiseks järgi testis antavaid juhiseid, tulemusi ei salvestata automaatselt!

Pärast testi sooritamist kuvatakse õigete ja valede vastuste põhjal arvutatud **punktsumma**. Test on koostatud eelkõige tööristaks uurijatele, et saada aimu uurimustes osalevate eesti keele õppijate üldisest sõnvara tundmisest. Soovitame lisaks LexEstile uurimuses kasutada ka teisi keeleoskuse mõõdikuid.

Test on üles laetud meie GitHubi.

Projekti toetab Kadi, Nikolai ja Gerda Rõugu fond.



Viited

- Alzahrani, A. (2024). LexArabic: A receptive vocabulary size test to estimate Arabic proficiency. *Behavior Research Methods*, 56(6), 5529–5556.
- Amenta, S., Badan, L., & Brysbaert, M. (2021). LexITA: A quick and reliable assessment tool for Italian L2 receptive vocabulary size. *Applied Linguistics*, 42(2), 292–314.
- Brysbaert, M. (2013). Lextale_FR a fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, 53(1), 23–37.
- Chan, I. L. & Chang, C. B. (2018). LEXTALE CH: A quick, character-based proficiency test for mandarin chinese. In *Proceedings of the Annual Boston University Conference on Language Development*: Cascadilla Press.
- Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, 35(1), 49–66.
- Kallas, J., Koppel, K., Pool, R., Tsepelina, K., Üksik, T., Alp, P., & Epner, A. (2021). Eesti keele kui teise keele õpetaja tööriistad Eesti Keele Instituudi keeleportalis Sõnaveeb. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 17, 61–80.
- Lee, S. T., van Heuven, W. J., Price, J. M., & Leong, C. X. R. (2024). LexMAL: A quick and reliable lexical test for Malay speakers. *Behavior Research Methods*, 56(5), 4563–4581.
- Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, 44, 325–343.
- Lõo, K., Järvikivi, J., & Baayen, R. H. (2018). Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing. *Cognition*, 175, 20–25.
- Salmela, R., Lehtonen, M., Garusi, S., & Bertram, R. (2021). Lexize: A test to quickly assess vocabulary knowledge in Finnish. *Scandinavian journal of psychology*, 62(6), 806–819.
- Wen, Y., Qiu, Y., Leong, C. X. R., & van Heuven, W. J. (2024). LexCHI: A quick lexical test for estimating language proficiency in Chinese. *Behavior Research Methods*, 56(3), 2333–2352.
- Zhou, C. & Li, X. (2022). LextPT: A reliable and efficient vocabulary size test for L2 Portuguese proficiency. *Behavior Research Methods*, 54(6), 2625–2639.