Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Tim Aadam Laats

Divergentse assotsiatsiooni ülesande seosed vaimse võimekuse ja isiksusega olenevalt sõnavektori valikust

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Uku Vainik,
PhD Kristi Kuljus

Tartu 2025

DIVERGENTSE ASSOTSIATSIOONI ÜLESANDE SEOSED VAIMSE VÕIMEKUSE JA ISIKSUSEGA OLENEVALT SÕNAVEKTORI VALIKUST

Bakalaureusetöö

Tim Aadam Laats

Lühikokkuvõte

Divergentse assotsiatsiooni ülesande (DAT) eesmärgiks on mõõta inimese divergentse mõtlemise võimekust ja loovust. Test koosneb üksteisest võimalikult erinevate nimisõnade nimetamisest. Skoor arvutatakse läbi sõnadevahelise semantilise distantsi, mis leitakse sõnade vektoresituste kaudu. Antud töös vaadeldakse kahte erinevat eestikeelset sõnade vektoresituste komplekti ehk sõnavektorit. Kasutatav andmestik on saadud sotsiaalmeedias jagatud kolmest uuringust, mis sisaldasid isiksuse- ja kognitiivseid teste ning eneseraporteeritud loomingulisust. Uuritakse, millised on DAT-skooride seosed teiste tunnustega andmestikus ning kuidas mõjutab tulemusi sõnavektori valik.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: divergentne mõtlemine, loomingulisus, sõnavektor, lineaarne regressioon.

THE RELATIONSHIP OF THE DIVERGENT ASSOCIATION TASK WITH COGNITIVE ABILITY AND PERSONALITY DEPENDING ON THE CHOICE OF WORD VECTOR

Bachelor thesis

Tim Aadam Laats

Abstract

The goal of the Divergent Association Task (DAT) is to measure a person's divergent thinking ability and creativity. The test consists of naming nouns that are as different from each other as possible. The score depends on the semantic distance between words, which is calculated using word vector representations. In this thesis, two different sets of Estonian word vector representations, or word vectors, are examined. The dataset was obtained via three studies shared on social media, which included personality and cognitive tests and self-reported creativity. The relationships between DAT scores and other variables in the dataset are examined, as well as how the choice of word vector affects the results.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: divergent thinking, creativity, word vector, linear regression

Sisukord

Sissejuhatus

Divergentse assotsiatsiooni ülesanne (edaspidi DAT - *Divergent Association Task*) on 2021. aastal avaldatud psühholoogiline test, mille eesmärk on mõõta objektiivselt inimese divergentse ehk lahkneva mõtlemise võimet ning seeläbi ka tema loovust. Testi raames palutakse vastajatel nimetada 10 nimisõna, mis oleksid üksteisest võimalikult erinevad. Seejärel arvutatakse sõnadevahelised semantilised distantsid printsiibil, et sarnastes kontekstides kasutatavate sõnade vahel on väiksem distants ning lõpliku skoori saamiseks leitakse sõnapaaride skooride transformeeritud keskmine. (Olson *et al.*, 2021)

Skoori arvutamine sõltub ka sellest, millist digitaliseeritud tekstide kogumit ehk tekstikorpust kasutatakse ning millise algoritmi abil on sealt arvutatud sõnadevahelised kaugused. Taoliste algoritmide kaudu leitakse korpuses olevatele sõnadele vektoresitused, milles peegeldub siis sõnade omavaheline seos. Selliselt töödeldud korpust nimetame sõnavektoriks.

Käesolevas bakalaureusetöös võetakse vaatluse alla kaks eestikeelset sõnavektorit eesmärgiga välja selgitada, kuidas on nende põhjal arvutatud DAT-skoorid seotud muude inimest iseloomustavate tunnustega andmestikus. Samuti uuritakse, kuidas mõjutab seoseid sõnavektori valik.

1 Divergentse assotsiatsiooni ülesanne

1.1 Divergentne mõtlemine ja loomingulisus

Divergentne mõtlemine (*Divergent thinking* - DT) on termin, mille võttis kasutusele ameerika psühholoog J. P. Guilford 1950-ndatel aastatel[1] inimintelligentsuse tervikliku teooria loomise käigus, pidades seda üheks olulisemaks kategooriaks loomingulise mõtlemise kirjeldamisel ning seostas selle alla käivaid oskuseid „ideede genereerimisega olukordades, kus varieeruvus on olulisel kohal“ (Guilford, 1967a; Guilford, 1967b). EKI Haridussõnastikus on divergentne mõtlemine defineeritud kui „uute ideede tekkele ja ebatavalistele lahendustele viiv mõtlemine, probleemile rohkem kui ühe lahenduse pakkumine“ (Eesti Keele Instituut, i.a). Ka tänapäeval on loomingulisuse uurimisel erinevad divergentset mõtlemist mõõtvad testid väga laialdaselt kasutusel ning ehkki nende abil on võimalik mingil määral eelnimetatud oskusi mõõta, siis ei peeta kindlasti korrektseks divergentset ja loomingulist mõtlemist samastada ning DT-testid on pigem vaid üks viis hinnata loomingulist potentsiaali (Runco ja Acar, 2012).

1.2 DAT-testi valiidsus, seos divergentse mõtlemise ja loomingulisusega

DAT-testi valiidsuse hindamiseks on uuritud tulemuste korrelatsiooni juba kasutusel olevate divergentse mõtlemise testidega. Näiteks leiti 2024. aastal väljaantud uuringus, mis viidi läbi Hiina põhikooli õpilaste seas, DAT-testi positiivne seos kahe juba tuntud DT-testiga *Alternative Uses Task*[2] (AUT) ja *Bridge-the-Associative-Gap Task*[3] (BAG) (Ding *et al.*, 2024). Ka algsetes 2021. aasta uurin-

[1] Hiljem võttis ta küll *„thinking“* ebamäärasuse tõttu selle asemel kasutusele termini *„production“*.

[2] Vastajal tuleb etteantud igapäevasele objektile leida võimalikult palju kasutusviise.

[3] Vastajal tuleb kahe sõna jaoks leida kolmas, mis mõlemaga seotud oleks.

gutes leiti DAT-i positiivsed seosed mitme AUT dimensiooni ja BAG puhul. Lisaks olid DAT-skooride korrelatsioonid teiste DT-testide skooridega üldjuhul vähemalt sama tugevad kui teiste DT-testide skooride omavahelised korrelatsioonid (Olson *et al.*, 2021). Positiivsed seosed DAT-testi ja nimetatud DT-testide vahel ilmnesid ka 2024. aasta Jaapani uuringutes (Ishiguro *et al.*, 2024). Seega eeldades juba kasutusel olevate DT-testide kehtivust, võib divergentse mõtlemise testina ka DAT-testi tõenäoliselt valiidseks pidada.

Loomingulisuse hindamiseks kasutatakse tihti eneseraporteeritud käitumist mõõtvaid küsimustikke. Näiteks eelmainitud Jaapani uuringutes kasutati kolme sellist tüüpi skaalat:

- Loomingulise enesehinnangu skaala (*Creative Self Scale*) - 11 väidet loomingulise käitumise kohta skaalal 1 („ei nõustu üldse“) kuni 7 („nõustun täielikult“)
- Loomingulise käitumise skaala (*Creative Behavior Scale*) - 34 loomingulise tegevuse seast märkida ära sellised, millega viimase 12 kuu jooksul oldi tegeletud
- Loominguliste tegevuste ja saavutuste skaala (*Creative Activities and Achievement Scale*) - viimase 10 aasta jooksul etteantud loominguliste tegevuste tegemise sagedus ning saavutused kaheksas erinevas valdkonnas (kirjandus, muusika, kunst ja käsitöö, loominguline kokkamine, sport, visuaalkunst, lavakunst, teadus ja tehnika). Tegevuste sagedust ja saavutusi käsitleti eraldi tunnustena.

Kahes uuringus vaadeldi skaalade korrelatsioone 12 erineva sõnavektori DAT-skooriga. Esimeses uuringus loomingulise enesehinnangu seost DAT-skooriga ühegi sõnavektori puhul ei tuvastatud. Mitme sõnavektori korral leiti aga positiivsed seosed loomingulise käitumise ja loominguliste saavutustega, kus seoste tugevuseks hinnati olenevalt sõnavektorist $r = 0,18$ kuni $r = 0,27$. Teises uuringus tuvastati mingil

määral DAT-skoori positiivne seos kõigi loomingulisuse skaaladega. (Ishiguro *et al.*, 2024).

1.3 Seos IQ ja isiksusega

Ehkki DAT-testi ja IQ vahelist seost varem otseselt vaadeldud ei ole, siis on uuritud IQ seost DT-ga teiste kasutusel olevate DT-testide kaudu. Näiteks 2023. aasta metaanalüüsis leiti mõõdukas positiivne korrelatsioon ($r = 0,47$; 95% UI: [0,38; 0,54]) DT ja ulatusliku meenutamisvõime (*Broad Retrieval Ability*)[4] ehk Gr vahel, ning veidi vähemal määral ($r = 0,31$; 95% UI: [0,20; 0,41]) DT ja töötlemiskiiruse (*Processing speed*)[5] ehk Gs vahel (Miroshnik *et al.*, 2023).

Vaadeldud on ka Suure viisiku dimensioonide ning DAT-skoori omavahelisi korrelatsioone. Jaapani uuringutes (Ishiguro *et al.*, 2024) esimese uuringu puhul DAT-skoori seost ühegi sõnavektori korral Suure viisikuga ei tuvastatud. Teises uuringus leiti aga mitme sõnavektori puhul DAT-skoori positiivne seos avatusega kogemusele (*Openness*), kus seose tugevuseks hinnati olenevalt sõnavektorist $r = 0,13$ kuni $r = 0,23$. Kuigi seosed teiste Suure viisiku dimensioonidega ei olnud statistiliselt olulised, siis seos meelekindlusega oli mõlemas uuringus järjepidevalt negatiivne ning positiivne seos leiti vaid ühes uuringus ühe sõnavektori korral (Ishiguro *et al.*, 2024).

On ka uuritud Suure viisiku ja DT seost teiste DT-testide kaudu. Aastal 2023 tehtud metaanalüüsis leiti eelkõige seos avatusega ($r = 0,20$; 95% UI: [0,18; 0,22]) ning vähemal määral ka ekstravertsusega ($r = 0,09$; 95% UI: [0,06; 0,12]) (Grajzel, Acar ja Singer, 2023).

[4]Oskus ladusalt ja paindlikult verbaalset ja mitteverbaalset infot pikaajalisest mälust kätte saada (Schneider ja McGrew, 2018)

[5]Oskus kiiresti ja tõhusalt lihtsaid ja korduvaid kognitiivseid ülesandeid lahendada (Schneider ja McGrew, 2012)

1.4 Tekstikorpuse ja algoritmi valik

Testi tulemuste valiidsus võib olla oluliselt mõjutatud nii kasutatavast keelekorpusest kui ka semantiliste distantside leidmise algoritmist. Näiteks leiti Hiinas tehtud uuringus, et Word2vec algoritm oli eelistatav sellest uuematele BERT ja GPT algoritmidele, sest viimaste puhul oli DAT-i seos teiste DT testidega oluliselt nõrgem (Ding *et al.*, 2024). Kõige arvukamalt erinevate korpuste ja algoritmide korral DAT-i vaadeldud Jaapani uuringutes kasutati seitset erinevat jaapanikeelset korpust ning Word2vec algoritme *Continuous Bag of Words* (CBOW) ja Skip-gram, samuti ka Glove algoritmi. Ehkki kõigi sõnavektorite puhul tuvastati positiivne seos DAT-skoori ja tuntud DT-testi AUT dimensioonidega, siis seos loomingulisuse skaalade ja Suure viisikuga leiti vaid valitud sõnavektorite korral. Erines ka skooride jaotus sõltuvalt sõnavektori valikust (Ishiguro *et al.*, 2024).

1.5 Sõnavektorid

Antud töös on vaadeldud kahte sõnavektorit. Neist esimene pärineb NLPL (*Nordic Language Processing Laboratory*) repositooriumist ning on saadud Eesti CoNLL17 tekstikorpuse põhjal kasutades Word2vec pideva Skip-gram algoritmi (Fares *et al.*, 2017). Teine sõnavektor pärineb Fasttexti projektist, mille raames treeniti sõnavektorid 157 keele jaoks kasutades Wikipedia ja Common Crawl korpuseid kasutades CBOW algoritmi, mis on modifitseeritud versioon originaalsest Word2vec algoritmist (Grave *et al.*, 2018).

Sõnavektorite leidmise algoritmid lähtuvad distributiivse semantika kesksest ideest, et sõna tähendus on kirjeldatav selle kaudu, mis kontekstides see esineb (Grave *et al.*, 2018). Word2vec Skip-gram puhul on treenimise ülesandeks iga korpuses esineva sõna jaoks ennustada teda ümbritsevaid[6] sõnu. CBOW töötab aga põhimõttel,

[6]Ümbritsevate sõnade arv on määratud hüperparameetri valikuga. Näiteks antud töös NLPL sõnavektori treenimisel kasutati akna suurust $n = 10$, ehk võeti arvesse sõnale 10 eelnevat ja 10 järgnevat sõna.

et proovitakse ennustada ümbritsevate sõnade kaudu keskmist sõna (Mikolov *et al.*, 2013). Oluline ei ole aga otseselt nimetatud ülesannete lahendamine, vaid lahendamise optimeerimise tulemusena saadud sõnade vektoresitused, mis võimaldavad näiteks antud töös arvutada sõnadevahelist semantilist distantsi läbi vektoresituste-vahelise koosinuskauguse. Antud töös kasutatud sõnavektorite korral olid vektoresitused näiteks elementide arvuga 100 (NLPL) ja 300 (fasttext). Lisaks sellele, et vektoresituste-vaheline koosinuskaugus peegeldab vastavate sõnade semantilist distantsi, on teatud vektoresituste-vahelistel liitmis- ja lahutamistehetel ka intuitiivne tõlgendus. Tuntuimaks näiteks on (**kuningas** − **mees** + **naine** ≈ **kuninganna**), st liites sõnade „kuningas“ ja „naine“ vektoresitused ning lahutades neist sõna „mees“ vektoresituse, saadakse vektor, mis on väga lähedane „kuninganna“ vektoresitusele. Mõned teistsugused seosed on näiteks (**pariis** − **prantsusmaa** + **poola** ≈ **varssavi**) ja (**autod** − **auto** + **õun** ≈ **õunad**), mis ilmestavad vastavalt lingvistilisi mustreid nagu riigi pealinn ja mitmus (Vylomova *et al.*, 2015). Ka üks peamiseid sõnavektorite kvaliteedi hindamise viise põhineb taoliste suhete kehtimise täpsusel (Mikolov *et al.*, 2013; Grave *et al.*, 2018).

Lisaks sellele, et antud töös vaadeldud sõnavektorid on tuletatud erinevate korpuste pealt ning põhinevad erineval vektoresituste leidmise algoritmil (fasttext - CBOW ja NLPL - Skip-gram), on peamine erinevus n-grammide kasutus Fasttexti CBOW meetodis. Antud fasttext sõnavektori puhul kasutati n-gramme pikkusega 5 (Grave *et al.*, 2018). See tähendab, et sõnade vektoresitus arvutatakse sõna viie sümboli pikkuste alasõnede ja sõna enda vektoresituse (kui see korpuses leidub) summana, võttes arvesse ka sõna algust ja lõppu tähistavaid sümboleid (< ja >). Näiteks sõna „rohuroheline“ vektoresitus on leitud n-grammide `<rohu, rohur, ohuro, huroh, urohe, rohel, oheli, helin, eline, line>` ja sõna enda (kui see korpuses leidub) `<rohuroheline>` vektoresituste summana. Üks esile toodud eeliseid sellise lähenemise juures on, et korpuses harva (või mitte üldse) esinevate sõnade jaoks on võimalik leida kvaliteetsemaid vektoresitusi, mis võib olulist rolli mängida näiteks paljude käänete või pööretega keelte puhul (Bojanowski *et al.*, 2017). Kuna

DAT-testis küsitakse aga vaid nimisõnu, siis on arvatavasti antud töös n-grammide kasutamine tähtis eelkõige liitsõnade korral. Näiteks isegi kui „rohuroheline“ esineb korpuses väga harva, jagab see n-gramme oluliselt sagedamini kasutatava sõnaga „roheline“, mis võimaldab potentsiaalselt leida sõnale „rohuroheline“ kvaliteetsema vektoresituse.

2 Mitmene lineaarne regressioon

Järgnev peatükk toetub õpikule Heumann, Schomaker ja Shalabh (2016) ning annab ülevaate statistilisest metoodikast, mida antud töö raames kasutati.

Mitmest lineaarset regressiooni kasutatakse juhul, kui sõltuv tunnus Y võib olla mõjutatud rohkem kui ühe, näiteks p selgitava tunnuse $X_1, \ldots, X_p$ poolt. Siis avaldub mudel kujul

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e, \tag{1}$$

kus $\beta_0, \beta_1, \ldots, \beta_p$ on regressioonikoefitsiendid ning e on mudeli viga. Täpsemalt nimetatakse β_0 ka vabaliikmeks.

Olgu meil n vaatlusega andmestik, kus kõik vaatlused rahuldavad võrrandit (1). Siis saame mudeli esitada maatrikskujul

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kus

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Siin $\mathbf{X}$ koosneb n vaatlusest selgitavate tunnuste $X_1, \ldots, X_p$ kohta ning ühtede veerust, mis märgib vabaliikme olemasolu. Hinnang regressioonikoefitsientidele $\boldsymbol{\beta}$ leitakse vähimruutude meetodil, mis põhineb vigade ruutude kogusumma minimiseerimisel, st minimiseeritakse $\sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$. Vähimruutude hinnang avaldub kujul

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}.$$

Et $\boldsymbol{\beta}$ hinnangutele saaks konstrueerida usaldusintervallid ning testida nendega seonduvaid statistilisi hüpoteese, on vajalik eeldus vigade normaaljaotuse kohta. Täpsemalt eeldame, et $e_i \sim \mathcal{N}(0, \sigma^2)$, kus σ^2 on vigade dispersioon. Seega vigade keskväärtus $\mathrm{E}(e_i) = 0$ ja $\mathrm{Var}(e_i) = \sigma^2$ (homoskedastsus) iga e_i korral. Nihketa σ^2 hinnang avaldub kujul

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (p+1)} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n - (p+1)} = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \hat{e}_i^2,$$

kus mudeli vead on hinnatud kui $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Formaalselt testimaks, kas ühe kindla koefitsiendi hinnang $\hat{\beta}_j$ on erinev nullist, st $H_0 : \beta_j = 0$ ja $H_1 : \beta_j \neq 0$, kasutatakse T-statistikut

$$T = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}},$$

mille väärtused on t-jaotusega, t_{n-p-1}. Siin $\hat{\sigma}_{\hat{\beta}_j} = \sqrt{s_{jj}\hat{\sigma}^2}$, kus s_{jj} on maatriksi $(\mathbf{X}'\mathbf{X})^{-1}$ j-s diagonaalil asuv element. Nullhüpotees kummutatakse juhul, kui $|T| > t_{n-p-1;1-\alpha/2}$, kus α on valitud olulisuse nivoo. Regressioonikoefitsientide $\hat{\beta}_j$ hinnangute usaldusintervallid olulisuse nivool α on

$$\hat{\beta}_j \pm t_{n-p-1;1-\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_j}.$$

On võimalik testida ka mudeli üldist olulisust järgneva nullhüpoteesi kaudu:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

Ehk uurime, kas leidub vähemalt üks nullist erinev β_i. Selleks kasutame F-statistikut

$$F = \frac{n - p - 1}{p} \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i e_i^2},$$

kus $\hat{y}_i$ on mudelist saadud hinnangud ja mille väärtused on F-jaotusega, $F_{p,n-p-1}$. Nullhüpotees olulisuse nivool α kummutatakse juhul, kui $F > F_{1-\alpha;p,n-p-1}$.

Olukorras, kus potentsiaalselt informatiivseid selgitavaid tunnuseid on palju ja ei ole ilmne, millist kombinatsiooni nendest mudelis peaks kasutama, on üks võimalus süstemaatiliselt tunnuseid mudelisse lisada või eemaldada. Antud töös kasutati tagurpidi elimineerimist p-väärtuste alusel, mis töötab järgmiselt:

1. Alustatakse mudeliga, mis sisaldab kõiki vaadeldavaid tunnuseid $\mathcal{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}\}$.
2. Valitakse olulisuse nivoo.
3. Kui mudelis leidub tunnuseid, mille p-väärtus on valitud olulisuse nivoost kõrgem, siis eemaldatakse tunnus $\mathbf{x_i} \in \mathcal{X}$, millel on kõrgeim p-väärtus.
4. Sammu 3. korratakse, kuni olulisuse nivoost kõrgema p-väärtusega tunnuseid mudelis ei leidu.

Mudelite kirjeldavuse määra hindamiseks jaotatakse sõltuva tunnuse koguhajuvus (SQ_{Kogu}) kaheks: regressioonimudelist tulenev ruutude summa ($SQ_{\text{Regressioon}}$) ja jääkidest tulenev ruutude summa ($SQ_{\text{Jääk}}$):

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SQ_{\text{Kogu}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SQ_{\text{Regressioon}}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SQ_{\text{Jääk}}}.$$

On selge, et tahaksime jääkidest tulenevat ruutude summat $SQ_{\text{Jääk}}$, mis oleks võimalikult lähedal nullile. See tähendaks, et mudelist saadud hinnangud $\hat{y}_i$ on väga sarnased päriselt valimis esinenud väärtustele y_i, ehk mudel kirjeldab sõltuva tunnuse varieeruvust väga hästi. Mida suurem on aga $SQ_{\text{Jääk}}$, seda suuremad on erinevused hinnangute ja valimi väärtuste vahel, mistõttu on mudeli kirjeldavus kehvem. Mudeli kirjeldavuse määr defineeritakse kujul:

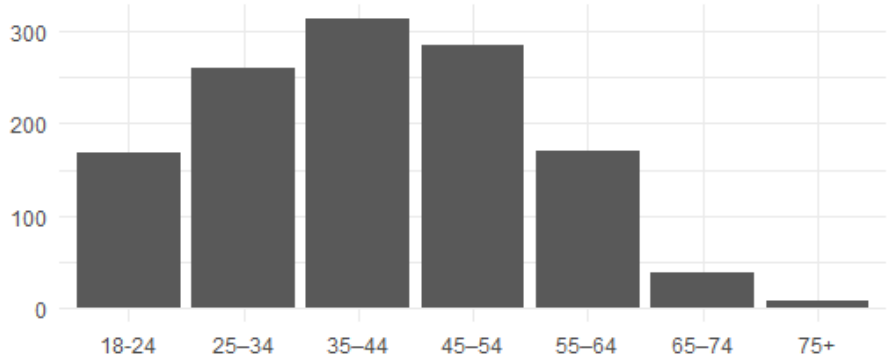
$$R^2 = \frac{SQ_{\text{Regressioon}}}{SQ_{\text{Kogu}}} = 1 - \frac{SQ_{\text{Jääk}}}{SQ_{\text{Kogu}}}, \tag{2}$$

kus $0 \leq R^2 \leq 1$. Üks piiranguid valemis (2) toodud definitsiooniga on, et R^2 väärtus kasvab igal juhul tunnuste mudelisse juurde lisamisel, isegi kui need on ebaolulised. Seega kasutatakse ka korrigeeritud kirjeldavuse määra R^2_{adj}, mis võtab arvesse mudelisse kaasatud tunnuste arvu ja mida saab kasutada erinevate tunnuste arvuga mudelite võrdlemisel:

$$R^2_{adj} = 1 - \frac{SQ_{\text{Jääk}}/(n-p-1)}{SQ_{\text{Kogu}}/(n-1)}.$$

3 Andmestiku ülevaade

Analüüsiks kasutati kolme veidi varieeruvate testide ja küsimustega uuringu andmestikku, mis on kõik kogutud Facebook-i kaudu. Vastajatele anti küsimustiku eduka läbimise korral automaatne tagasiside viie peamise isiksusejoone skoori kohta. Andmestikest valiti välja vaid tunnused, mille puhul sooviti uurida nende seost DAT-skooriga. Lõplik valim suurusega $n = 1241$ koosnes vastajatest, kes täitsid korrektselt ära DAT-testi ning vähemalt osaliselt ka ülejäänud küsimustiku ning keda ei jäetud analüüsist kõrvale madalapoolse eesti keele taseme või vastamisel kõrvalise abi kasutamise tõttu. Neist 1064 (86%) olid naised ja 177 (14%) mehed. Kõige noorem vastaja oli 18 ning vanim vastaja 86 aastat vana, vanuste jaotus on toodud joonisel 1.



Joonis 1: Vastajate arv vanusegruppide kaupa

3.1 DAT-skoorid ja sõnade komplektid

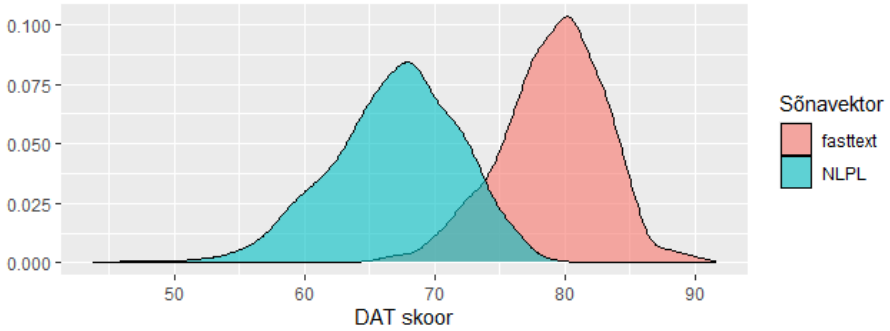
Sarnaselt DAT-testi loojatele (Olson *et al.*, 2021) kasutati DAT-testi skoori arvutamisel vaid seitset esimest valiidset sõna[7]. Seega jäi analüüsist välja 19 inimest, kes ei nimetanud vähemalt ühe sõnavektori vaates piisavalt valiidseid sõnu. Andmestike

[7]Valiidse sõna all on mõeldud sõna, millel on kasutatava sõnavektori puhul olemas vektoresitus.

jaoks arvutati DAT-testi raames kogutud nimisõnade põhjal DAT-skoor nii fasttexti (Grave *et al.*, 2018) kui ka NLPL (Fares *et al.*, 2017) eestikeelsete sõnavektorite jaoks, kasutades sõnavektorite sisselugemiseks ja töötlemiseks Gensimi (Řehůřek ja Sojka, 2010) fasttexti ning word2vec implementatsioone ja Pythoni näidiskoodi (Lisa 1. Kood sõnadevahelise sarnasuse arvutamiseks). DAT-skoori leidmiseks arvutati kõigepealt kasutatava 7 sõna vektoresituste vahelised paariviisilised koosinussarnasused (kokku 21 sõnapaari). Seejärel võeti leitud koosinussarnasustest aritmeetiline keskmine. Lõppskoor saadi kujul[8] (1− keskmine koosinussarnasus) ×100.

DAT-testile ei seatud vaadeldud uuringutes ajapiirangut, kuid suur osa vastajatest lahendasid testi siiski üsna lühikese aja jooksul. Lahendamisaja mediaan oli 141 sekundit ning 81% vastanutest lahendasid DAT-testi vähem kui 4 minuti jooksul, mis oli testi loojate (Olson *et al.*, 2021) poolt kasutatud ajapiirang.

Jooniselt 2 on näha skooride jaotus mõlema sõnavektori korral. Fasttext skoorid olid antud valimis nii kõrgema keskväärtusega (fasttext 79,0; NLPL 66,9) kui ka väiksema hajuvusega (standardhälve fasttext 4,1; NLPL 5,0). Samuti näitas sõnavektorite skooride vaheline Pearsoni korrelatsioonikordaja $r = 0,60$, et tegemist on küllaltki erinevate sõnavektoritega ja analüüsi käigus oleks oodatav näha ka erinevat käitumist olenevalt sõnavektori valikust. Neli kõige väiksemat DAT-skoori



Joonis 2: DAT-skooride hinnatud tihedus vastavalt sõnavektorile

saanud sõnade komplekti olid mõlema sõnavektori korral identsed. Neist minimaal-

[8]koosinuskaugus = 1− koosinussarnasus

ne (skooriga fasttext - 50,58 ja NLPL - 30,48) leiti sõnade komplekti[9] (**jalanõud**, **tuhvlid**, **sussid**, **plätud**, **sandaalid**, kossid, **pastlad**, **papud**, butsad, kingad) puhul, mis on selge näide testi ülesande valest tõlgendamisest ning jäeti seetõttu edasisest analüüsist välja. See kattub ka originaaluuringus tehtud järeldusega, et kõige madalamad skoorid leiti tihti olukordades, kus ülesandest oli valesti aru saadud (Olson *et al.*, 2021).

1. Teised madalaimate skooridega sõnade komplektid on näiteks:
   (**Empaatilisus**, **Sõbralikkus**, **Töö**, **Armastus**, **Hoolimine**, **Positiivsus**, **Konkreetsus**, Ausus, Kuulamine, Jalutamine);
   fasttext - 59,86; NLPL - 46,44
   (**pere**, **laps**, **õde**, **koer**, **kass**, **kodu**, **sõbrad**, päike, suvi, sügis);
   fasttext - 64,38; NLPL - 46,74

2. Keskmiseid skoore iseloomustavateks sõnakomplektideks võib tuua:
   (**Rekka**, **Õun**, **Teater**, **Sein**, **Rohi**, **Võilill**, **Prügikast**, Mäda, Lihased, Kruus);
   fasttext - 79,03; NLPL - 65,24
   (**Laud**, **Puu**, **Pilv**, **Kassa**, **Arvuti**, **Koer**, **Pall**, Tolm, Müra, Vesi);
   fasttext - 80,82; NLPL - 66,86

3. Kõige kõrgema skoori sai fasttexti korral komplekt:
   (**keskendumisvõime**, **öö**, **aeg**, **ussipesa**, **armastus**, must auk, **radioloogia**, **tunked**, venelane, lilliput);
   fasttext - 91,71; NLPL - 73,51
   Mõlema sõnavektori puhul ei leitud sõnu „must auk“ ja „lilliput“.

   NLPL korral sai kõrgeima skoori:
   (**aed**, **professor**, **mõte**, **beebi**, **sõiduk**, **turg**, **antiik**, kanal, torn, vanker)
   fasttext - 82,87; NLPL - 79,51

[9]Rasvase tekstiga on tähistatud need 7 sõna, mida skoori arvutamisel kasutati.

3.2 Keeleoskus ja haritus

Järgevalt on toodud keeleoskuse ja haritusega seotud tunnuste kirjeldused:

eestikTase - Küsimusele „Mis tasemel Te valdate eesti keelt?“ on vastatud skaalal:

1 - „A1 - algtase“

2 - „A2 - madalam kesktase“

3 - „B1 - kesktase“

4 - „B2 - kõrgem kesktase“

5 - „C1 - edasijõudnud“

6 - „C2 - professionaalne“

Tabelis 1 toodud sageduste põhjal jäeti regressioonanalüüsist välja vastanud, kes märkisid keeletasemeks B1 või madalam ($n = 7$). Tasub märkida, et küsimust sisaldas vaid Uuring 3, mille vastused moodustasid 23% koguvalimist. Teistest uuringutest võisid seega olla aga kaasatud ka madala keeleoskusega inimesed, tõenäoliselt sarnase proportsiooniga nagu uuringus 3. Tunnust käsitleti analüüsis faktortunnusena.

Tabel 1: Keeletaseme vastuste sagedustabel

Keeletase	A1	A2	B1	B2	C1	C2
Sagedus	1	0	6	27	68	198

haridusAastad - Haridustasemed on kodeeritud arvulisteks väärtusteks, mis peegeldavad haridustee pikkust aastates. Kodeerimine baseerub tabelil S1 artiklis Rietveld *et al.* (2014).

1 - „Alghariduseta“

7 - „Algharidus“

10 - „Põhiharidus“

13 - „Keskharidus“

15 - „Rakenduslik kõrgharidus“

19 - „Kõrgharidus (bakalaureuse- või magistrikraad)“

22 - „Teaduskraad (doktorikraad)“

Tunnust käsitleti analüüsis pidevana, tabelis 2 on toodud vastuste jaotus.

Tabel 2: tunnuse haridusAastad sagedustabel

haridusAastad	1	7	10	13	15	19	22
Sagedus	0	6	96	693	116	342	9

3.3 Testide läbimine

Järgnevalt on toodud testide läbimisega seotud tunnuste kirjeldused:

datMeeldis - DAT testi meeldivus skaalal 1 („Üldse ei meeldinud“) kuni 5 („Väga meeldis“)

datKeeruline - DAT testi keerulisus skaalal 1 („Väga lihtne“) kuni 5 („Väga raske“)

testMot - hinnang vastaja motivatsioonile kõigi uuringu käigus läbitud testide täitmisel. Kasutatud on *Student Opinion Scale* (Thelk, Horst ja Finney, 2009) eestikeelset küsimustikku, mis sisaldab näiteks väiteid nagu „Minu jaoks oli oluline, et mul läheks testides võimalikult hästi“ ja „Ma oleksin olnud võimeline teste paremini sooritama, kui oleksin rohkem pingutanud“. Skoorid on vahemikus 1-5.

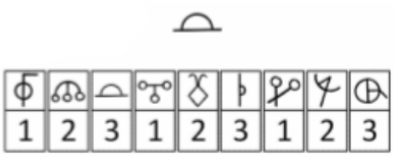
Abi - „Kas kasutasite ülesannete tegemisel kõrvalist abi?“:
Analüüsist jäeti kõrvale kõik mingil viisil abi otsimist tunnistanud vastajad ($n = 27$).

Kõiki testide läbimisega seotud tunnuseid (v.a Abi, mida kasutati vaid vaatluste kõrvaldamisel) käsitleti analüüsis pidevatena.

3.4 Kognitiivsed testid

Alapeatükk on kirjutatud veebilehel *testmybrain.org* (TestMyBrain, 2025) toodud kirjelduste ning artiklite (Passell *et al.*, 2019; Singh *et al.*, 2021) põhjal. Järgnevalt on toodud kognitiivsete testide kirjeldused, mille skoore analüüsis kasutati:

DSC - numbri-sümboli vastavuse test: ette antakse erinevad numbri-sümboli kombinatsioonid ning sümboli ilmumisel tuleb kasutajal võimalikult kiiresti vajutada vastavale numbrile. Näiteks joonisel 3 toodud näite korral oleks korrektne vajutada numbrile 3. Test mõõdab töötlemiskiirust ning visuaalset lühiajalist mälu.



Joonis 3: number-sümboli sobitamise testi näide

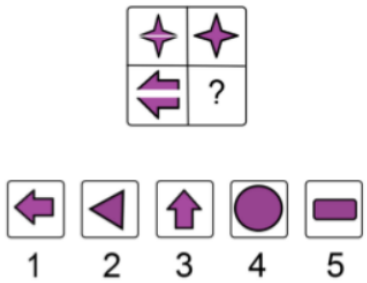
VPA - sõnapaaride ajendatud meenutamise test: esimeses faasis kuvatakse meeldejätmiseks ükshaaval 25 sõnapaari, teises faasis tuleb kasutajal kõikide nähtud sõnade jaoks leida valikuvariantide hulgast korrektne paariline. Test mõõdab verbaalset ning episoodilist mälu.

Vocab - sünonüümide sõnavara test: iga antud sõna jaoks tuleb kasutajal leida viie sõna hulgast sellele tähenduslikult sarnaseim. Test mõõdab pikaajalist verbaalset mälu, kristalliseeritud ja üldist intelligentsust ning verbaalset mõtlemist.

forwardSpan (backwardSpan) - numbrite mälu-ulatuse testid: esitatakse järjest pikemaid numbrijadasid, kus numbreid näidatakse ükshaaval lühikese viivitusega. Pärast iga numbrijada esitust tuleb kasutajal korrektselt sisestada jada kas samas (forwardSpan) või vastupidises järjekorras (backwardSpan). Test mõõdab lühiajalist mälu, tähelepanuvõimet ja töömälu.

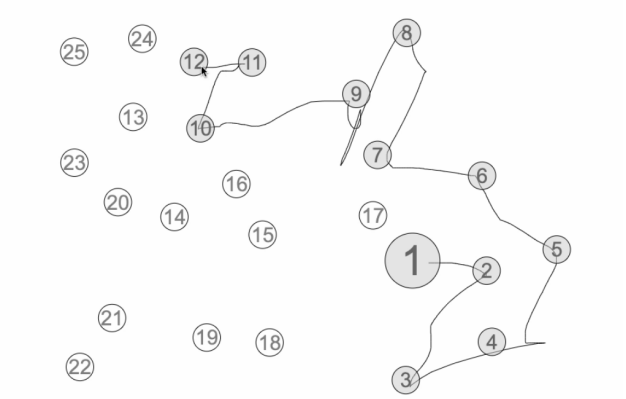
Matrix - maatriksid: esitatakse järjest keerulisemaid piltide maatrikseid, millest on üks pilt puudu. Kasutajal tuleb leida vastusevariantide hulgast puuduvale kohale sobivaim pilt teatud loogilise reegli järgi. Näiteks joonisel 4 toodud näite puhul oleks korrektne vastus 1. Test mõõdab fluiidset intelligentsust ning mitteverbaalset mõtlemist.

TrailsA - punktide ühendamise test A: kasutajal tuleb numbritega tähistatud ringid võimalikult kiiresti õiges järjekorras ühendada, vt näidet joonisel 5. Test



Joonis 4: maatriksi testi näide

mõõdab töötlemiskiirust ja ühelt ülesandelt teisele ümberlülitamise oskust (*task switching*).



Joonis 5: punktide ühendamise test A näide

TrailsB - punktide ühendamise test B: kasutajal tuleb numbrite ja tähtedega tähistatud ringid võimalikult kiiresti vahelduvalt õiges järjekorras ühendada (i.e., "1", "A", "2", "B", "3", "C"... "13") (Treviño *et al.*, 2021). Test mõõdab kognitiivset paindlikkust.

SimpleRT - lihtreaktsiooniaeg: kasutajal tuleb võimalikult kiiresti vajutada nupule (või ekraanile) niipea, kui kuvatav punane ruut muutub roheliseks. Test mõõdab psühhomotoorset reageerimisaega.

GradCPT - järk-järgulise algusega pideva sooritusse test: ekraanile kuvatakse üksteise järel kas mägede pilte (10-20% piltidest) või linnavaate pilte, vt näidet joonisel 6. Kasutajal tuleb võimalikult kiiresti linnavaate puhul nupule vajutada

ning mägede korral nupule vajutamisest hoiduda. Test mõõdab püsivat tähelepanuvõimet, pidurduskontrolli ning kognitiivset kontrollimisvõimet.

Joonis 6: järk-järgulise pideva soorituse testis esinevate piltide näide

Valimis esinenud skooride kirjeldus on toodud tabelis 3.

Tabel 3: Kognitiivsete testide skooride kirjeldus valimis

Tunnus	Keskväärtus	Standardhälve	Min	Max	n
DSC	43.72	8.04	12.00	66.00	1071
VPA	16.95	5.04	3.00	25.00	807
Vocab	19.63	7.31	4.00	45.00	1198
forwardSpan	6.24	1.60	2.00	11.00	277
backwardSpan	5.03	1.58	2.00	11.00	559
Matrix	25.83	4.78	5.00	34.00	557
TrailsA	42.05	18.14	12.52	107.06	283
TrailsB	28.43	11.91	12.14	77.06	289
SimpleRT	30.03	5.55	13.15	50.37	573
GradCPT	71.53	14.69	15.62	100.00	495

3.5 Suur viisik ja loomingulisus

Suure viisiku test (avatus kogemusele, meelekindlus, ekstravertsus, koostöövalmidus ja neurootilisus) koosnes vastavalt uuringule kas 10-st või 12-st väitest iga dimensiooni kohta. Nende alusel on arvutatud iga dimensiooni jaoks agregeeritud skoorid, võttes vastava dimensiooni väidete tulemustest aritmeetilise keskmise. Et varem on näidatud eelkõige DT ja DAT-i seost avatusega kogemusele (Grajzel, Acar

ja Singer, 2023; Ishiguro *et al.*, 2024), on vaatluse alla võetud selle dimensiooni väited ka ükshaaval. Kirjeldatud tunnuseid käsitleti analüüsis pidevatena. Vastatud on skaalal 1 („Täiesti vale“) kuni 6 („Täiesti õige“):

Avatus01 - „Mind huvitab teadus“

Avatus04R[10] - „Soovin, et mind nähakse korraliku ja tavapärase inimesena“

Avatus05 - „Mul on rikas sõnavara“

Avatus06 - „Mässan autoriteetide vastu“

Avatus07 - „Mind peetakse elutargaks inimeseks“

Avatus08 - „Mind peetakse isemoodi inimeseks“

Avatus09 - „Töötan enda parandamise nimel“

Avatus09R - „Väldin filosoofilisi arutelusid“

Avatus10 - „Oskan paljusid asju hästi“

Avatus11 - „Ma ei kõhkle, kui on vaja ebapopulaarne arvamus välja öelda“

Avatus12 - „Armastan uusi asju õppida“

Avatus15 - „Pean kunsti oluliseks“

Avatus17 - „Mulle meeldib nuputada üha uusi viise, kuidas asju teha“

Avatus18 - „Pean saama end loominguliselt väljendada“

Avatus20 - „Mulle meeldib keerulisi probleeme lahendada“

Avatus24 - „Mul on elav kujutlusvõime“

Avatus28 - „Mind huvitavad paljud asjad“

Loomingulisus1 - küsimusele „Kas Te peate ennast loominguliseks inimeseks?“ on vastatud skaalal 1 („Ei, ei ole üldse loominguline“) kuni 5 („Jah, olen väga loominguline“)

Loomingulisus2 - küsimusele „Kas Te olete enda arvates hea uute ja teistmoodi seoste loomisel?“ on vastatud skaalal 1 („Ei, see on väga raske minu jaoks“) kuni 5 („Jah, saan sellega väga hästi hakkama“)

[10]Tähega R lõppevate väidete puhul on kodeeritud skoorid ümberpööratult, ehk kõrgem tulemus vastab siiski kõrgemale avatuse skoorile.

4 Regressioonanalüüs

Antud töö eesmärk oli tuvastada potentsiaalselt olulisi DAT-skoori kirjeldavaid tunnuseid ning uurida, kuidas mõjutab tulemusi sõnavektori valik. Ülesande lahendamiseks kasutati mitmest lineaarset regressioonanalüüsi, kasutades mudelites sõltuva tunnusena DAT-testi skoore (kas fasttext või NLPL) ning selgitavate tunnustena eelmises peatükis kirjeldatud tunnuseid, mis võiksid DAT-skoori mõjutada. Et iga uuringu puhul kasutati mõneti erinevat testide ja küsimuste valikut, siis leiti regressioonimudelid iga uuringu jaoks eraldi ning ka valimite jaoks, mis kombineerivad uuringuid (1,2), (1,3), (2,3) ja (1,2,3). Seega näiteks valimi (1,2) puhul võeti algsesse regressioonimudelisse vaid sellised tunnused, mida sisaldasid nii Uuring 1 kui ka Uuring 2.

Tabelist 4 on näha vastuste arv iga uuringu ja tunnuse jaoks. Tuli näiteks välja, et leidus tunnuseid, mida vaid ühesse uuringusse kaasati (nt TrailsA, TrailsB, SimpleRT, GradCPT ja mitmed avatuse väited uuringus 2 ning eestikTase, datMeeldis, datKeeruline, forwardSpan uuringus 3). Uuringutes 1 ja 3 kattusid eelkõige avatuse väited ning uuringutes 2 ja 3 kattusid hästi mitmed kognitiivsed testid ja loomingulisuse küsimused. Seega iga valim, mille jaoks mudel leiti, oli erinevate tunnuste, tunnuste arvu ja valimi suurusega.

Regressioonimudelite leidmisel võeti aluseks kõik tunnused, mis vastavas uuringus sisaldusid (uuringute kombinatsioonide korral olemasolevad tunnused kõigis vastavates uuringutes). Et aidata tuvastada tunnuseid, mis DAT-skooride kirjeldamisel oluliseks võivad osutuda, kasutati tagurpidi eliminatsiooni (*backward elimination*) p-väärtuste alusel ($p = 0,1$) rakendades R paketti `olsrr` (Hebbali, 2024). Kõikide valimite jaoks leiti alguses ka eraldi mudelid, kaasates kas ainult agregeeritud skoori Avatus või vaid kõiki vastava valimi korral leiduvaid avatuse väiteid. Kuna selgus, et valimites avatuse väited eraldi tunnustena kirjeldavad DAT-skoore paremini kui koondtunnus, siis on esitatud vaid tulemused, mis kaasavad avatuse väiteid eraldi. See kattub ka koguvalimil korrelatsioonide uurimisel saadava järeldusega, vt

Tabel 4: Vastuste arv uuringute ja tunnuste kaupa, 0 – antud uuringus vastav küsimus ei esinenud

Tunnus	Uuring 1	Uuring 2	Uuring 3	Kokku
Sugu	262	689	290	1241
Vanus	262	689	290	1241
eestikTase	0	0	290	290
haridusAastad	262	689	290	1241
datMeeldis	0	0	240	240
datKeeruline	0	0	240	240
testMot	239	315	240	794
DSC	256	563	252	1071
VPA	0	555	252	807
Vocab	248	685	265	1198
forwardSpan	0	0	277	277
backwardSpan	0	309	250	559
Matrix	241	316	0	557
TrailsA	0	283	0	283
TrailsB	0	289	0	289
SimpleRT	0	573	0	573
GradCPT	0	495	0	495
Avatus	229	304	228	761
Meelekindlus	229	304	228	761
Ekstravertsus	229	304	228	761
Koostöövalmidus	229	304	228	761
Neurootilisus	229	304	228	761
Avatus01	0	304	0	304
Avatus04R	229	0	228	457
Avatus05	0	304	0	304
Avatus06	229	0	228	457
Avatus07	0	304	0	304
Avatus08	0	304	0	304
Avatus09	0	304	0	304
Avatus09R	229	304	228	761
Avatus10	0	304	0	304
Avatus11	0	304	0	304
Avatus12	229	304	228	761
Avatus15	229	304	228	761
Avatus17	229	0	228	457
Avatus18	229	0	228	457
Avatus20	229	304	228	761
Avatus24	229	304	228	761
Avatus28	229	0	228	457
Loomingulisus1	0	685	290	975
Loomingulisus2	0	685	290	975

Lisa 3. Paariviisilised korrelatsioonid – koguvalim. Kõikides uuringuid kombineerivates valimites jäeti mudelisse ka uuringu numbrit tähistav faktortunnus, mida koondtabelites välja ei tooda, kuna see osutus iga valimi korral ebaoluliseks.

4.1 Valimi 2,3 mudelite näide

Järgnevad mudelid on leitud valimi põhjal, mis kombineeris uuringuid 2 ja 3. Puuduvate väärtuste eemaldamise järel jäi valimisse $n = 502$ vaatlust. Algsest mudelist, mis sisaldas 19 selgitavat tunnust, saadi eliminatsiooni järel tabelites 5 ja 6 toodud mudelid, kus sõltuvaks tunnuseks oli vastavalt fasttexti ja NLPL sõnavektorite põhjal arvutatud DAT-skoor.

Tabel 5: fasttext regressioonimudel peale eliminatsiooni – Uuring 2,3 valim

	Koefitsient	SE	t(494)	p
(Vabaliige)	77.30	1.57	49.10	$< .0001$
Vanus	-0.09	0.02	-4.27	$< .0001$
VPA	0.10	0.04	2.73	0.0066
Vocab	0.16	0.04	4.04	$< .0001$
backwardSpan	0.24	0.12	2.06	0.0396
Meelekindlus	-0.44	0.22	-1.99	0.0472
Avatus12	0.32	0.16	1.94	0.0528
Uuring (3)	0.27	0.39	0.69	0.4928
$R^2_{adj} = 0.100$, $\hat{\sigma} = 3.912$				

Tabel 6: NLPL regressioonimudel peale eliminatsiooni – Uuring 2,3 valim

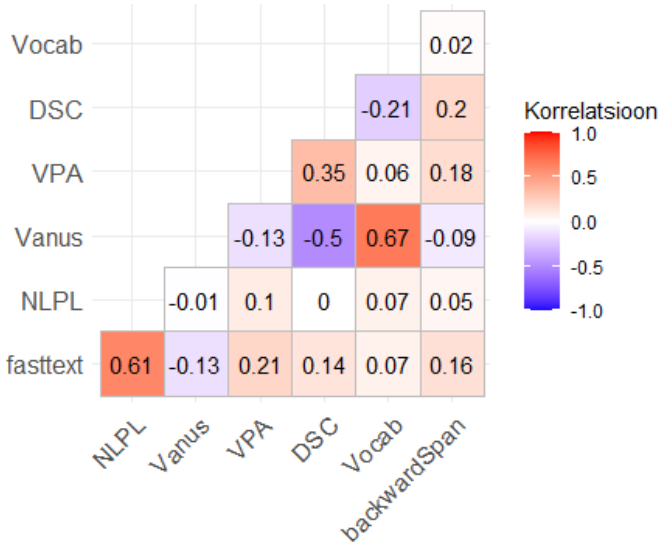
	Koefitsient	SE	t(496)	p
(Vabaliige)	67.05	1.24	53.99	$< .0001$
Vanus	-0.05	0.02	-1.88	0.0603
Vocab	0.12	0.05	2.38	0.0179
Avatus09R	0.32	0.15	2.04	0.0414
Loomingulisus2	-0.52	0.24	-2.14	0.0328
Uuring (3)	0.42	0.45	0.95	0.3413
$R^2_{adj} = 0.018$, $\hat{\sigma} = 4.829$				

Tabelitest 5 ja 6 on näha, et fasttexti mudeli kirjeldavuse määr (R^2_{adj}) on kõrgem. Mõlema mudeli kirjeldavuse määrad on aga küllaltki madalad ning ka fasttexti

mudel kirjeldab ära vaid umbes 10% DAT-skoori varieeruvusest. Lisaks on kõrged mudelite jääkide standardhälbed $\hat{\sigma}$, mille väärtused on pea sama suured, kui DAT-skooride standardhälbed koguvalimis (fasttext 4,1; NLPL 5,0).

Mõlema sõnavektori korral tuvastati negatiivne seos vanusega ja positiivne seos sünonüümide sõnavara testiga Vocab. Seos Vocab-testiga on huvitav, sest tegemist on intuitiivselt DAT-testile küllaltki sarnase ülesandega (Vocab testib tähenduslikult sarnaste ning DAT tähenduslikult erinevate sõnade leidmist). Fasttext mudelis leiti ka positiivne seos VPA-testiga (sõnapaaride ajendatud meenutamise test) ning mõnevõrra väiksema statistilise olulisusega positiivne seos tunnustega backwardSpan (numbrite mälu-ulatuse test) ja Avatus12 („Armastan uusi asju õppida“) ning negatiivne seos meelekindlusega.

Vaadates DAT-skooride, vanuse ja kognitiivsete testide korrelatsioonimaatriksit antud valimis (Joonis 7) on näha, et vanus on negatiivselt korreleeritud kõigi kognitiivsete testidega peale Vocab-testi.



Joonis 7: Korrelatsioonid valimis 2,3 – vanus ja kognitiivsed testid

Ehkki Vocab-testi ja vanuse vahel on küllaltki tugev positiivne korrelatsioon, siis mudelites on DAT-skooriga vanusel negatiivne ning Vocab-testil positiivne seos,

mis kehtib ka korrelatsioonide puhul.

Joonisel 7 toodud korrelatsioonide põhjal jääb ka mulje, et fasttext DAT-skooriga on kognitiivsete testide hulgast Vocab-testi tulemustel kõige nõrgem seos. Pärast vanuse mõju arvesse võtmist on aga kognitiivsete testide seast Vocab-testi osakorrelatsioon fasttext DAT-skooriga üks tugevamaid (Tabel 7). Võrdluseks on vastavad osakorrelatsioonid toodud ka NLPL DAT-skooriga, kus seosed on oluliselt nõrgemad (Tabel 8).

Tabel 7: Osakorrelatsioonid fasttext DAT-skooriga arvestades vanuse mõju – Uuring 2,3 valim

	Hinnang	p
Vocab	0.21	$< .0001$
VPA	0.20	$< .0001$
DSC	0.09	0.0363
backwardSpan	0.15	0.0006

Tabel 8: Osakorrelatsioonid NLPL DAT-skooriga arvestades vanuse mõju – Uuring 2,3 valim

	Hinnang	p
Vocab	0.10	0.0196
VPA	0.10	0.0231
DSC	0.00	0.9724
backwardSpan	0.05	0.2802

4.2 Valimi 1,3 mudelite näide

Järgnevad mudelid on leitud valimi põhjal, mis kombineerisid uuringuid 1 ja 3 ($n = 446$). Tegemist oli ainukese kahte või enamat uuringut kombineeriva valimiga, milles kattusid täielikult avatuse väited. Tabelites 9 ja 10 on toodud pärast eliminatsiooni saadud regressioonimudelid.

Sarnaselt eelmisele näitele leiti fasttexti mudeli korral negatiivne seos vanuse ja meelekindlusega ning positiivne seos Vocab-testiga. Ükski nimetatud tunnus NLPL

Tabel 9: fasttext regressioonimudel peale eliminatsiooni – Uuring 1,3 valim

	Koefitsient	SE	t(437)	p
(Vabaliige)	81.03	1.64	49.55	$< .0001$
Sugu (Naine)	1.06	0.59	1.79	0.0734
Vanus	-0.09	0.02	-4.25	$< .0001$
Vocab	0.11	0.03	3.23	0.0013
Meelekindlus	-0.72	0.23	-3.11	0.0020
Avatus09R	0.34	0.15	2.34	0.0197
Avatus20	0.40	0.16	2.48	0.0135
Avatus24	-0.31	0.16	-1.97	0.0498
Uuring (3)	0.77	0.49	1.57	0.1178
$R^2_{adj} = 0.086$, $\hat{\sigma} = 3.842$				

Tabel 10: NLPL regressioonimudel peale eliminatsiooni – Uuring 1,3 valim

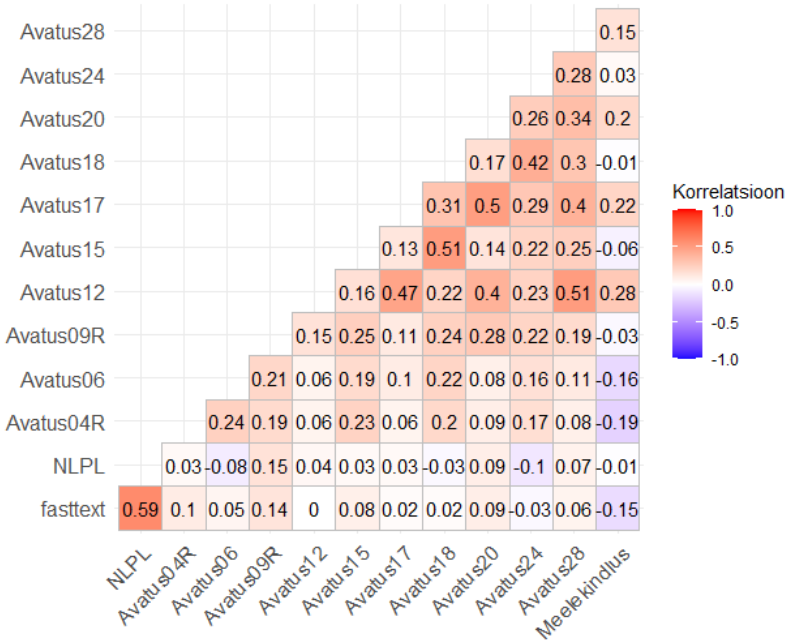
	Koefitsient	SE	t(439)	p
(Vabaliige)	65.02	1.59	40.88	$< .0001$
Sugu (Naine)	1.29	0.76	1.70	0.0900
Avatus06	-0.35	0.17	-2.03	0.0430
Avatus09R	0.75	0.19	4.06	$< .0001$
Avatus24	-0.61	0.20	-3.07	0.0023
Avatus28	0.40	0.22	1.84	0.0664
Uuring (3)	-0.41	0.47	-0.87	0.3838
$R^2_{adj} = 0.053$, $\hat{\sigma} = 4.893$				

mudelis ei esinenud. Mõlema sõnavektori puhul leiti positiivne seos Avatus09R („Väldin filosoofilisi arutelusid“)[11] ja negatiivne seos Avatus24 („Mul on elav kujutlusvõime“) tunnustega. Mõnevõrra väiksema statistilise olulisusega jäi mõlemasse mudelisse ka soo tunnus.

Joonisel 8 on toodud DAT-skooride, avatuse väidete ja meelekindluse korrelatsioonimaatriks. Ehkki kõik avatuse väited on omavahel positiivselt korreleeritud, siis eriti NLPL DAT-skooriga on mitmel avatuse väitel hoopis nõrk negatiivne seos, mis tuli välja ka mudelites. Üldiselt on avatuse väidete korrelatsioonid DAT-skooridega aga väga madalad. Korrelatsioonide põhjal paistab DAT-skooridega tugevaim seos olevat Avatus09R tunnusel ning seose tugevus on sarnane mõlema sõnavektori kor-

[11] Positiivne seos Avatus09R tunnusega tähendab sisuliselt seda, et filosoofilisi arutelusid vältivatel inimestel on madalamad DAT-skoorid.

ral. Fasttext mudelis leitud negatiivne seos DAT-skoori ja meelekindluse vahel kehtib ka korrelatsiooni puhul.



Joonis 8: Korrelatsioonid valimis 1,3 – avatuse väited ja meelekindlus

4.3 Mudelite kokkuvõte

Tabelites 11 (fasttext) ja 12 (NLPL) on kõikide mudelite koefitsientide ja p-väärtuste kokkuvõte iga valimi jaoks, kuhu on jäetud alles vaid tunnused, mis ühe või teise sõnavektori korral mõnesse mudelisse pärast elimineerimist jäid. Samuti on toodud mudelite kirjeldavuse määrad, jääkide standardhälbed ning valimite suurused. Kõikide kontrollitud tunnustega tabelid on toodud lisades (Lisa 2. Täielik mudelite kokkuvõte).

Tabel 11: fasttext mudelid, eliminatsiooni tulemused valimite kaupa – kolme punktiga (...) on tähistatud need tunnused, mis elimineerimise käigus vastavast mudelist kõrvale jäeti. Tühikutega tähistatud tunnused vastavas valimis ei esinenud.

Tunnus	1	2	3	1,2	1,3	2,3	1,2,3
Sugu (Naine)	...	...	1.75	...	1.06	...	...
Vanus	...	...	-0.12^{***}	-0.05^{**}	-0.09^{***}	-0.09^{***}	-0.07^{***}
haridusAastad	...	...	...	0.19^{*}	...	...	0.14^{*}
testMot	...	...	...	...	...	...	...
DSC	0.07	...	...	...	...	...	0.04
VPA		0.18^{*}	...			0.10^{*}	
Vocab	...	...	0.26^{***}	...	0.11^{*}	0.16^{***}	0.09^{*}
forwardSpan			0.35				
backwardSpan		...	...			0.24	
Matrix	...	0.18^{*}		0.10			
TrailsB		0.05					
GradCPT		−0.05					
Meelekindlus	−0.73	...	−0.80	−0.42	-0.72^{*}	−0.44	-0.53^{*}
Avatus04R	...		...		...		
Avatus05		0.47					
Avatus06	...		...		...		
Avatus09R	0.43	...	...	0.22	0.34	...	0.24
Avatus10		−0.54					
Avatus12	...	...	0.44	...	...	0.32	...
Avatus15	...	...	...	0.27	...	...	0.20
Avatus20	...	...	...	...	0.40	...	...
Avatus24	...	...	...	...	−0.31	...	...
Avatus28	...		...		...		
Loomingulisus1		...	...			...	
Loomingulisus2		...	...			...	
R^2_{adj}	0.051	0.164	0.146	0.093	0.086	0.100	0.084
$\hat{\sigma}$	3.782	3.672	3.830	3.882	3.842	3.912	3.918
n	222	196	217	519	446	502	744

*** $p < 0.0001$; ** $p < 0.001$; * $p < 0.01$

Tabel 12: NLPL mudelid, eliminatsiooni tulemused valimite kaupa – kolme punktiga (...) on tähistatud need tunnused, mis elimineerimise käigus vastavast mudelist kõrvale jäeti. Tühikutega tähistatud tunnused vastavas valimis ei esinenud.

Tunnus	1	2	3	1,2	1,3	2,3	1,2,3
Sugu (Naine)	...	...	3.08*	...	1.29	...	0.92
Vanus	...	...	−0.08	...	...	−0.05	...
haridusAastad	...	...	...	...	...	...	...
testMot	...	...	...	...	...	...	0.52
DSC	...	...	...	...	...	...	...
VPA		0.23*	...			...	
Vocab	...	...	0.25**	...	...	0.12	...
forwardSpan			...				
backwardSpan		...	...			...	
Matrix	...	...		...			
TrailsB		...					
GradCPT		−0.07*					
Meelekindlus	...	...	...	...	...	...	...
Avatus04R	...		0.49		...		
Avatus05		...					
Avatus06	−0.60		−0.47		−0.35		
Avatus09R	1.10**	...	0.50	0.53**	0.75***	0.32	0.55***
Avatus10		...					
Avatus12	...	...	...	...	...	...	...
Avatus15	...	...	...	...	...	...	...
Avatus20	...	0.58	...	...	...	...	...
Avatus24	−0.71*	...	...	−0.50*	−0.61*	...	−0.45*
Avatus28	...		...		0.40		
Loomingulisus1		...	−0.92*			...	
Loomingulisus2		−0.64	...			−0.52	
R^2_{adj}	0.089	0.095	0.110	0.032	0.053	0.018	0.030
$\hat{\sigma}$	4.810	4.509	4.653	4.943	4.893	4.829	4.902
n	222	196	217	519	446	502	744

*** $p < 0.0001$; ** $p < 0.001$; * $p < 0.01$

Ehkki peale Uuring 1 valimi olid kõik fasttext mudelid võrreldes NLPL mudelitega kõrgema kirjeldavuse määraga, siis olid kirjeldavuse määrad üldiselt väga madalad. Samuti ei leidunud ühtegi tunnust, mis oleks esinenud ühe või teise sõnavektori mudelites iga valimi puhul. Kõik seosed olid siiski järjepidevad selles mõttes, et iga tunnuse koefitsient esines mudelites sama märgiga olenemata valimist või sõnavektorist.

Fasttext mudelites oli küllaltki selge negatiivne seos vanusega, mis tuvastati kõrge statistilise olulisusega uuringus 3 ja kõigis valimeid kombineerivates mudelites. NLPL puhul jäi seos vanusega üldiselt ebaoluliseks. Mõlema sõnavektori puhul esines mõne valimi mudelites soo tunnus, kus naiste DAT-skoor oli võrreldes meestega kõrgem. Positiivne seos haridusaastatega tuvastati vaid fasttexti mudelites kahe valimi korral.

Fasttext mudelites tuli eelkõige välja positiivne seos mitmete kognitiivsete testidega. Kuna varem on näidatud DT seost kognitiivse võimekusega, siis oli ka DAT-testi puhul loogiline näha siin positiivseid seoseid. Nende hulgast kõige selgemalt tuvastati mõlema sõnavektori vaates seos Vocab (sünonüümide sõnavara test) ja vähemal määral ka VPA-testiga (sõnapaaride ajendatud meenutamise test). Tegemist on samuti loogilise tulemusega, sest Vocab ja VPA olid vaadeldud kognitiivsetest testidest ainukesed, mis kaasavad verbaalseid oskuseid. Lisaks on Vocab intuitiivselt DAT-testile sisu poolest mõnevõrra sarnane, sest kaasab otseselt sõnade tähendusliku sarnasuse tundmist. Ainukene negatiivse seosega kognitiivne test oli GradCPT (järk-järgulise algusega pideva soorituse test), mida vaadeldi vaid uuringu 2 valimis. NLPL mudelites olid seosed kognitiivsete testidega nõrgemad, vaid üksikutes valimites leiti seos VPA, Vocab ja GradCPT-ga.

Suure viisiku dimensioonidest tuvastati kõikides fasttext mudelites peale Uuring 2 valimi negatiivne seos meelekindlusega[12], mis üheski NLPL mudelis ei esinenud. Kuigi DT või DAT-i seost meelekindlusega varasemalt otseselt näidatud ei ole,

[12]Meelekindluse Uuring 2 fasttexti mudelisse kaasamisel oleks tunnuse koefitsient olnud positiivne (0,37).

siis ka Ishiguro *et al.* (2024) uuringutes olid DAT-skoori ja meelekindluse vahel järjepidevalt negatiivsed seosed praktiliselt kõigi vaadeldud sõnavektorite korral. Seal leitud seosed ei olnud aga statistiliselt olulised, mis on sarnane antud töö fasttext mudeli tulemustele, kus mudelitesiseselt oli tegu statistiliselt pigem väheolulise tunnusega. NLPL mudelites leiti aga eelkõige seosed mõne avatuse väitega, kus rohkem kui ühe valimi vaates tuvastati positiivne seos tunnusega Avatus09R („Väldin filosoofilisi arutelusid“)[13] ning negatiivne seos tunnustega Avatus06 („Mässan autoriteetide vastu“) ja Avatus24 („Mul on elav kujutlusvõime“). Neist kõige selgemalt tuli esile Avatus09R positiivne seos, mis oli ka fasttexti puhul avatuse väidetest kõige suurema arvu valimite puhul mudelitesse jäänud tunnus. Ülejäänud avatuse väited paistsid esinema fasttexti mudelites küllaltki juhuslikult. Nimetatud avatuse väidete osas jääb siiski ebaselgeks, kuidas saadud tulemusi tõlgendada. Avatus09R puhul ei ole näiteks ilmne, mida tunnus täpselt mõõdab ja kas tegu võiks olla informatiivse tunnusega divergentse mõtlemise või loomingulisuse uurimisel. Tabeli 12 põhjal tuleb näiteks esile muster, et Avatus09R on kõrge olulisuse (ja kõrgema koefitsendiga) just nende valimite mudelites, kust kognitiivsed testid on välja jäetud. Seda võib näha ka fasttexti puhul (Tabel 11), sest Avatus09R on jäänud mudelitesse, kus kognitiivsete testide olulisus on madalapoolne. Niisiis on üks võimalik selgitus, et Avatus09R (filosoofiliste arutelude mittevältimine) mõõdab mingil määral sarnaseid oskuseid, nagu teatud kognitiivsed testid. Sellele viitavad osaliselt ka tunnustevahelised korrelatsioonid koguvalimis (Tabel 13).

Tabel 13: Korrelatsioonid Avatus09R ja kognitiivsete testide vahel – Koguvalim

	Hinnang	p	n
VPA	0.17	6.868e-05	524
Vocab	0.08	0.0191	761
Matrix	0.14	0.0015	531

Negatiivsete märkidega avatuse väited ei viita aga kuidagi loogiliselt tõlgendatava-

[13] Positiivne seos Avatus09R tunnusega tähendab sisuliselt seda, et filosoofilisi arutelusid vältivatel inimestel on madalamad DAT-skoorid.

tele seostele, sest need võeti vaatluse alla lähtudes avatuse positiivsest seosest DT ja DAT-iga varasemates uuringutes. Lisaks sellele, et enamuse avatuse väidetega olid DAT-skooridel väga nõrgad seosed, selgitavad ka negatiivsete seostega väited, miks DAT-skoori ja avatuse koondtunnuse vaheline seos nii nõrgaks osutus. Niisiis võib DAT-i ja avatuse seose uurimisel väidete eraldi käsitlemine olla informatiivne, sest seos paistab antud töö tulemuste põhjal olema erinevate väidete korral kohati väga erinev. See aitaks ka aru saada, millistest avatuse väidetest võib positiivne seos avatuse koondtunnusega eelkõige tingitud olla. Tuleks aga arvatavasti mõelda selle peale, mida taolised positiivsete seostega väited mõõdavad. Näiteks antud töös tulid lisaks tunnusele Avatus09R mõnes mudelis positiivse märgiga sisse väited Avatus05 („Mul on rikas sõnavara“) ja Avatus20 („Mulle meeldib keerulisi probleeme lahendada“). On intuitiivselt loomulik, et rikas sõnavara seostub edukusega DAT-testis, kuid seda oskust saaks tõenäoliselt paremini mõõta objektiivsete testide kaudu. Samuti on keeruliste probleemide lahendamise meeldimine intuitiivselt seotud kõrgema motivatsiooni ja oskusega ka DAT-testi lahendamisel. Kahtluse alla võib aga seada taoliste seoste informatiivsuse DT ja loomingulisuse uurimise kontekstis.

Loomingulisuse küsimustega leiti NLPL mudelites kohati negatiivsed seosed, fasttext mudelites üheski valimis seoseid ei tuvastatud. Üks potentsiaalne põhjus, miks nende küsimuste puhul puudusid varasemate uuringute põhjal oodatavad loogilised positiivsed seosed, on antud töös kasutatud küsimuste lühidus (skaalad põhinevad vaid ühel küsimusel) ning vähene objektiivsus (vastamine sõltub suuresti enda subjektiivsest hinnangust). Seetõttu oleks arvatavasti soovitatav kasutada eneseraporteeritud loomingulisuse hindamiseks skaalasid, mis koosneksid suuremast arvust küsimustest ja/või küsimustest, mis võimaldaksid loomingulist käitumist hinnata objektiivsemalt, nagu näiteks Ishiguro *et al.* (2024) uuringutes kasutatud skaalad.

Kokkuvõte

Antud bakalaureusetöö eesmärk oli uurida, kuidas on DAT-skoorid seotud inimest iseloomustavate tunnustega ning kuidas mõjutab tulemusi sõnavektori valik. Vaatluse alla võeti kahe erineva eestikeelse sõnavektoriga arvutatud DAT-skoorid ning uuriti lineaarse regressioonanalüüsi kaudu skooride seost erinevate inimest kirjeldavate tunnustega andmestikus. Kuna andmestik sisaldas kolme erineva uuringu andmeid, kus tunnused mõnevõrra erinesid, siis vaadeldi lisaks eraldi uuringutele ka uuringuid kombineerivaid valimeid.

Analüüsi tulemusena leiti, et tulemused võivad olla olulisel määral mõjutatud sõnavektori valikust. Fasttext sõnavektori mudelites tulid kõige selgemalt välja negatiivne seos vanusega ja positiivsed seosed kognitiivsete testidega, seda eriti verbaalseid oskuseid kaasavate testide puhul. Vähemal määral tuli esile ka negatiivne seos meelekindlusega. NLPL mudelites tuvastati peamiselt positiivne seos tunnusega Avatus09R ning vähemal määral negatiivsed seosed tunnustega Avatus06 ja Avatus24. Jääb aga ebaselgeks, kuidas need väited on relevantsed DT või loomingulisuse uurimisel. Kummagi sõnavektori mudelites ei tuvastatud loogilist seost loomingulisuse küsimustega. Kuigi fasttexti mudelid olid üldiselt kõrgema kirjeldavuse määraga ning mudelites tuvastatud seoseid võib lugeda mõnevõrra loogilistemaks, siis on keeruline nende tulemuste alusel väita, et üks sõnavektor on usaldusväärsem kui teine. Puudusid tunnused, mis oleksid ühe või teise sõnavektori korral osutunud mudelites oluliseks iga valimi korral ning mudelite kirjeldavuse määrad olid üldiselt madalad.

Kasutatud andmestike üheks puudujäägiks oli, et need saadi sotsiaalmeedias levitatud testide kaudu, mis võib vähendada tulemuste usaldusväärsust. Teiseks ei leidunud andmestikus piisavalt kvaliteetseid tunnuseid, mis oleksid konkreetsemalt võimaldanud uurida seost DT ja loomingulisusega. Üks võimalus oleks kaasata DAT-testi uurimisel juba kasutusel olevaid DT-teste, mille abil saaks hinnata eestikeelse DAT-testi valiidsust DT-testina erinevate sõnavektorite korral. Samuti tuleks

arvatavasti kasutada loomingulisuse hindamisel detailsemaid ja objektiivsemaid küsimustikke.

Kasutatud allikad

Bojanowski, P., E. Grave, A. Joulin ja T. Mikolov (2017). *Enriching Word Vectors with Subword Information*. DOI: https://doi.org/10.48550/arXiv.1607.04606.

Ding, G., Y. He, K. Yi ja S. Li (2024). "Using the divergent association task to measure divergent thinking in Chinese elementary school students". *Thinking Skills and Creativity* 52. DOI: 10.1016/j.tsc.2024.101503.

Eesti Keele Instituut (i.a). *Haridussõnastik*. URL: https://arhiiv.eki.ee/dict/haridus/ (vaadatud 19.04.2025).

Fares, M., A. Kutuzov, S. Oepen ja E. Velldal (2017). "Word vectors, reuse, and replicability: Towards a community repository of large-text resources". Teoses: *Proceedings of the 21st Nordic Conference on Computational Linguistics*. Toim. J. Tiedemann ja N. Tahmasebi. Gothenburg, Sweden: Association for Computational Linguistics, lk. 271–276. URL: https://aclanthology.org/W17-0237/.

Grajzel, K., S. Acar ja G. Singer (2023). "The Big Five and divergent thinking: A meta-analysis". *Personality and Individual Differences* 214. DOI: 10.1016/J.PAID.2023.112338.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin ja T. Mikolov (2018). "Learning Word Vectors for 157 Languages". Teoses: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. DOI: https://doi.org/10.48550/arXiv.1802.06893.

Guilford, J. P. (1967a). "Creativity: Yesterday, today and tomorrow". en. *J. Creat. Behav.* 1.1, lk. 3–14.

– (1967b). *The Nature of Human Intelligence*. New York: McGraw-Hill.

Hebbali, A. (2024). *olsrr: Tools for Building OLS Regression Models*. R package version 0.6.1.9000, https://github.com/rsquaredacademy/olsrr. URL: https://olsrr.rsquaredacademy.com/ (vaadatud 21.03.2025).

Heumann, C., M. Schomaker ja S. Shalabh (2016). *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*. Springer Cham, lk. 249–289.

Ishiguro, C., S. Suzuki, M. Hattori, L. Abe ja K. Yang (2024). "Development and validation of Japanese version of divergent association task". Kasutatud Google Translate masintõlget. URL: https://doi.org/10.11225/cs.2024.038.

Mikolov, T., K. Chen, G. Corrado ja J. Dean (2013). *Efficient Estimation of Word Representations in Vector Space*. DOI: https://doi.org/10.48550/arXiv.1301.3781.

Miroshnik, K. G., B. Forthmann, M. Karwowski ja M. Benedek (2023). "The relationship of divergent thinking with broad retrieval ability and processing speed: A meta-analysis". *Intelligence* 98. DOI: https://doi.org/10.1016/j.intell.2023.101739.

Olson, J. A., J. Nahas, D. Chmoulevitch, S. J. Cropper ja M. E. Webb (2021). "Naming unrelated words predicts creativity". *Proceedings of the National Academy of Sciences* 118 (25), e2022340118. DOI: 10.1073/pnas.2022340118.

Passell, E., D. G. Dillon, J. T. Baker, S. C. Vogel, L. S. Scheuer, N. L. Mirin, L. A. Rutter, D. A. Pizzagalli ja L. Germine (2019). "Digital cognitive assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) field test battery report". DOI: https://doi.org/10.31234/osf.io/dcszr.

Řehůřek, R. ja P. Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". English. Teoses: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, lk. 45–50.

Rietveld, C. A., D. Conley, N. Eriksson, T. Esko, S. E. Medland, ... ja Social Science Genetics Association Consortium (2014). "Replicability and robustness of genome-wide-association studies for behavioral traits". en. *Psychol. Sci.* 25.11, lk. 1975–1986. DOI: 10.1177/0956797614545132.

Runco, M. A. ja S. Acar (2012). *Divergent Thinking as an Indicator of Creative Potential*. DOI: 10.1080/10400419.2012.652929.

Schneider, W. J. ja K. S. McGrew (2012). "The Cattell-Horn-Carroll model of intelligence". Teoses: *Contemporary intellectual assessment: Theories, tests, and issues*, lk. 99–144.

– (2018). "The Cattell-Horn-Carroll theory of cognitive abilities". Teoses: *Contemporary intellectual assessment: Theories, tests, and issues*. Toim. D P M Flanagan E. The Guilford Press, lk. 73–163.

Singh, S., Roger W Strong, Laneé Jung, Frances Haofei Li, Liz Grinspoon, Luke S Scheuer, Eliza J Passell, Paolo Martini, Naomi Chaytor, Jason R Soble ja Laura Germine (2021). "The TestMyBrain Digital Neuropsychology Toolkit: Development and psychometric characteristics". en. *J. Clin. Exp. Neuropsychol.* 43.8, lk. 786–795. DOI: 10.1080/13803395.2021.2002269.

TestMyBrain (2025). *TMB Cognitive Tests*. https://www.testmybrain.org/research-tools/tmb-tests.html. (Vaadatud 05.03.2025).

Thelk, A., S. Horst ja S. Finney (2009). "Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance". *Journal of General Education* 58, lk. 129–151. DOI: 10.1353/jge.0.0047.

Treviño, M., X. Zhu, Y. Y. Lu, L. S. Scheuer, E. Passell, G. C. Huang, L. T. Germine ja T. S. Horowitz (2021). “How do we measure attention? Using factor analysis to establish construct validity of neuropsychological tests”. en. *Cogn. Res. Princ. Implic.* 6.1. DOI: 10.1186/s41235-021-00313-1.

Vylomova, E., L. Rimell, T. Cohn ja T. Baldwin (2015). “Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning”. DOI: https://doi.org/10.48550/arXiv.1509.01692.

Lisa 1. Kood sõnadevahelise sarnasuse arvutamiseks

Näidiskood on kirjutatud Yaroslav Opanasenko (Tartu Ülikooli haridusteaduste instituut) poolt varasema projekti raames. Autori poolt on lisatud sõnade komplektidest vaid esimese seitsme valiidse sõna kaasamine keskmise sõnadevahelise sarnasuse arvutamisel, samuti on fasttexti mudeli töötlemisel tehtud vajalikud muudatused koodis.

```
import zipfile
import gensim
import pandas as pd
import numpy as np

# Path to the pre-trained model
model_path = "41.zip"

# Unzipping the model archive and loading the model
with zipfile.ZipFile(model_path, 'r') as archive:
    model_path = archive.extract('model.bin')
    model = gensim.models.KeyedVectors.load_word2vec_format(
        model_path, binary=True)

# Fasttext (autori lisatud)
#from gensim.models.fasttext import load_facebook_vectors
#model = load_facebook_vectors('cc.et.300.bin')

# Path to the csv table
dataframe = pd.read_csv('raw_formr_ee.csv')
```

```
# Selecting the columns containing words for each session
words_columns = dataframe.iloc[:, 5:15]

# List of words for each session
words_sessions = [words_columns.iloc[i].tolist()
                  for i in range(len(dataframe))]

# Computing the similarity matrix for words in each session
matrix = []
not_found_words = []  # List of words that were not found in the model
words_ind = 0
for words in words_sessions:
    row = []
    for word1 in words:
        column = []
        for word2 in words:
            try:
                # Formatting the word before searching in the model
                word1_formatted = word1.strip().lower()
                word2_formatted = word2.strip().lower()
                similarity = model.similarity(word1_formatted,
                                              word2_formatted)
            except KeyError:
                similarity = np.nan
            column.append(similarity)
        row.append(column)
    matrix.append(row)
    # Checking each word in the current session for existence in the model
```

```
    not_found = [word for word in words if word.strip().lower()
                 not in model]
    # Fasttext (autori lisatud)
    #not_found = [word for word in words if word.strip().lower()
    #             not in model.key_to_index]
    not_found_words.append(not_found)

    # Autori lisatud
    # If there are less than 7 valid (more than 3 invalid) words,
    # exclude the session from the average similarity calculation
    if len(not_found) > 3:
        words_sessions[words_ind] = []
    # Otherwise take the first 7 valid words
    else:
        s = set(not_found)
        words_sessions[words_ind] = [x for x in words if x not in s][:7]
    words_ind += 1

# Creating a DataFrame with similarity matrices for each session
df = pd.DataFrame(matrix, index=dataframe['session'],
                  columns=words_columns.columns)

# Computing the average similarity for each session
average_similarities = []
for words in words_sessions:
    total_similarity = 0
    num_pairs = 0
    for i in range(len(words)):
        for j in range(i + 1, len(words)):
```

```
            try:
                similarity = model.similarity(words[i].strip().lower(),
                                              words[j].strip().lower())
            except KeyError:
                continue
            total_similarity += similarity
            num_pairs += 1
    if num_pairs > 0:
        average_similarity = total_similarity / num_pairs
    else:
        average_similarity = np.nan
    average_similarities.append(average_similarity)

# Adding a new column with average similarity to the DataFrame
dataframe['Average Similarity'] = average_similarities

# Creating a column for the not found words in the DataFrame
dataframe['Not Found Words'] = not_found_words

new_tablecsv = 'datscores_word2vec7.csv'
dataframe.to_csv(new_tablecsv, index=False)
```

Lisa 2. Täielik mudelite kokkuvõte

Tabel 14: fasttext mudelid, kõigi tunnustega eliminatsiooni tulemused valimite kaupa – kolme punktiga (...) on tähistatud need tunnused, mis elimineerimise käigus vastavast mudelist kõrvale jäeti. Tühikutega tähistatud tunnused vastavas valimis ei esinenud.

Tunnus	1	2	3	1,2	1,3	2,3	1,2,3
Sugu (Naine)	...	...	1.75	...	1.06	...	...
Vanus	...	...	-0.12^{***}	-0.05^{**}	-0.09^{***}	-0.09^{***}	-0.07^{***}
eestikTase			...				
haridusAastad	...	...	...	0.19^{*}	...	...	0.14^{*}
datMeeldis			...				
datKeeruline			...				
testMot	...	...	...	...	...	...	...
DSC	0.07	...	...	...	...	...	0.04
VPA		0.18^{*}	...			0.10^{*}	
Vocab	...	...	0.26^{***}	...	0.11^{*}	0.16^{***}	0.09^{*}
forwardSpan			0.35				
backwardSpan		...	...			0.24	
Matrix	...	0.18^{*}		0.10			
TrailsA		...					
TrailsB		0.05					
SimpleRT		...					
GradCPT		-0.05					
Meelekindlus	-0.73	...	-0.80	-0.42	-0.72^{*}	-0.44	-0.53^{*}
Ekstravertsus	...	...	...	...	...	...	...
Koostöövalmidus	...	...	...	...	...	...	...
Neurootilisus	...	...	...	...	...	...	...
Avatus01		...					
Avatus04R	...		...		...		
Avatus05		0.47					
Avatus06	...		...		...		
Avatus07		...					
Avatus08		...					
Avatus09		...					
Avatus09R	0.43	...	...	0.22	0.34	...	0.24
Avatus10		-0.54					
Avatus11		...					
Avatus12	...	...	0.44	...	...	0.32	...
Avatus15	...	...	...	0.27	...	...	0.20
Avatus17	...		...		...		
Avatus18	...		...		...		
Avatus20	...	...	...	...	0.40	...	...
Avatus24	...	...	...	...	-0.31	...	...
Avatus28	...		...		...		
Loomingulisus1		...	...			...	
Loomingulisus2		...	...			...	
R^2_{adj}	0.051	0.164	0.146	0.093	0.086	0.100	0.084
$\hat{\sigma}$	3.782	3.672	3.830	3.882	3.842	3.912	3.918
n	222	196	217	519	446	502	744

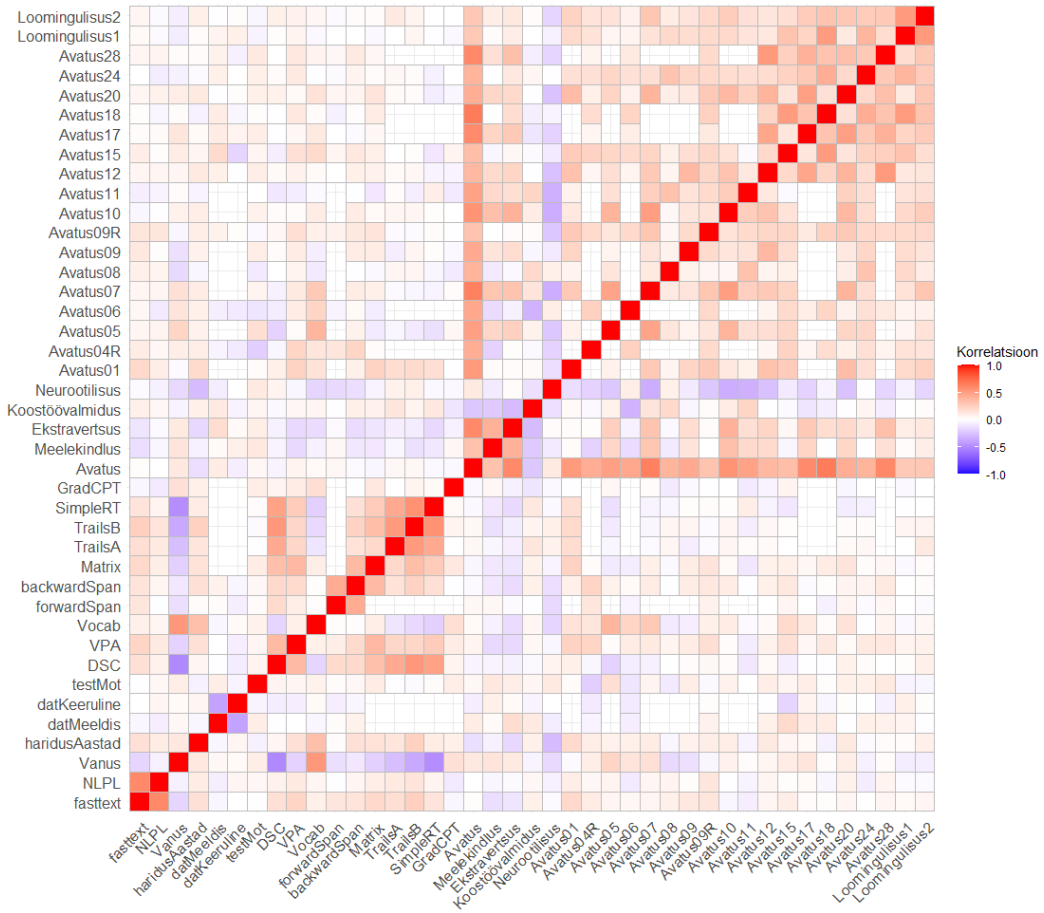
*** $p < 0.0001$; ** $p < 0.001$; * $p < 0.01$

Tabel 15: NLPL mudelid, kõigi tunnustega eliminatsiooni tulemused valimite kaupa – kolme punktiga (...) on tähistatud need tunnused, mis elimineerimise käigus vastavast mudelist kõrvale jäeti. Tühikutega tähistatud tunnused vastavas valimis ei esinenud.

Tunnus	1	2	3	1,2	1,3	2,3	1,2,3
Sugu (Naine)	...	...	3.08*	...	1.29	...	0.92
Vanus	...	...	−0.08	...	...	−0.05	...
eestikTase			...				
haridusAastad	...	...	...	...	...	...	...
datMeeldis			...				
datKeeruline			...				
testMot	...	...	...	...	...	...	0.52
DSC	...	...	...	...	...	...	...
VPA		0.23*	...			...	
Vocab	...	...	0.25**	...	...	0.12	...
forwardSpan			...				
backwardSpan		...	...			...	
Matrix	...	...		...			
TrailsA		...					
TrailsB		...					
SimpleRT		...					
GradCPT		−0.07*					
Meelekindlus	...	...	...	...	...	...	...
Ekstravertsus	...	...	...	...	...	...	...
Koostöövalmidus	...	...	...	...	...	...	...
Neurootilisus	...	...	...	...	...	...	...
Avatus01		...					
Avatus04R	...		0.49		...		
Avatus05		...					
Avatus06	−0.60		−0.47		−0.35		
Avatus07		...					
Avatus08		...					
Avatus09		...					
Avatus09R	1.10**	...	0.50	0.53**	0.75***	0.32	0.55***
Avatus10		...					
Avatus11		...					
Avatus12	...	...	...	...	...	...	...
Avatus15	...	...	...	...	...	...	...
Avatus17	...		...		...		
Avatus18	...		...		...		
Avatus20	...	0.58	...	...	...	...	...
Avatus24	−0.71*	...	...	−0.50*	−0.61*	...	−0.45*
Avatus28	...		...		0.40		
Loomingulisus1		...	−0.92*			...	
Loomingulisus2		−0.64	...			−0.52	
R^2_{adj}	0.089	0.095	0.110	0.032	0.053	0.018	0.030
$\hat{\sigma}$	4.810	4.509	4.653	4.943	4.893	4.829	4.902
n	222	196	217	519	446	502	744

*** $p < 0.0001$; ** $p < 0.001$; * $p < 0.01$

Lisa 3. Paariviisilised korrelatsioonid – koguvalim



Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Tim Aadam Laats,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Divergentse assotsiatsiooni ülesande seosed vaimse võimekuse ja isiksusega olenevalt sõnavektori valikust“, mille juhendajad on Uku Vainik ja Kristi Kuljus, 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.

Tim Aadam Laats
14.05.2025