



TARTU ÜLIKOOL

Empiirilised andmed semantikas: kui juhuslik peaks olema polüseemia uurimisel kasutatav korpusvalim?

AIMI PIKKAAR

19. rakenduslingvistika konverents
"Keele mõju: andmetest tõendatud teadmiseni"
16.–17.juuni 2022, Tallinn

Millest räägin?

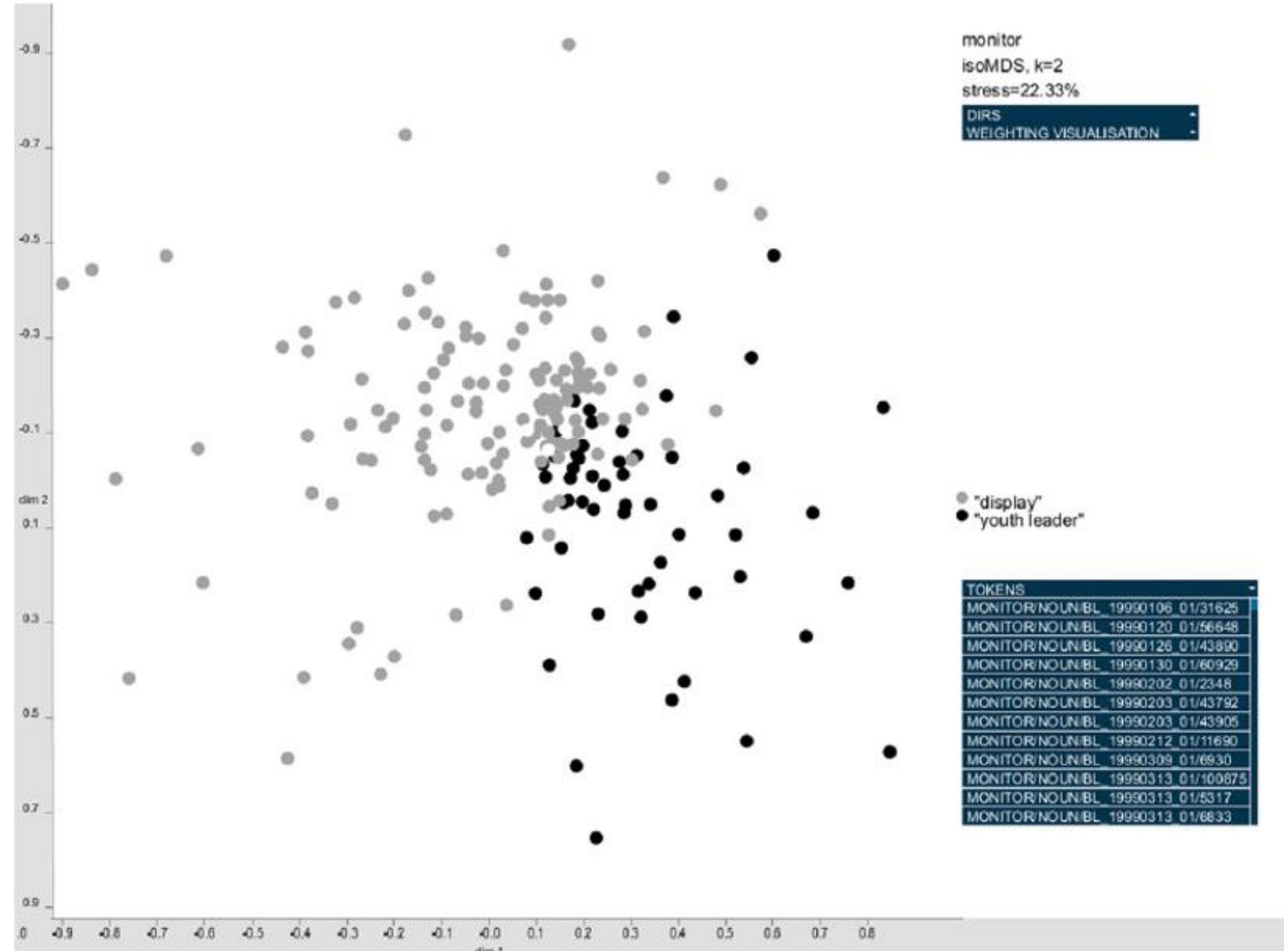
- **Distributiivsemantilised mudelid** – üks võimalus korpusandmete kasutamiseks polüseemia uurimisel
- **Valimi** moodustamise olulisusest
- Mida on võimalik teha distributiivsemantiliste mudelitega selle **uurimismeetodi praeguses arenguetapis?**

Distributiivsemantilised mudelid polüseemia uurimisel

- Tähenduste kirjeldamise vahendid, mis toetuvad arusaamale, et sõna tähendust on võimalik esitada kontekstide kaudu, milles see sõna loomulikus keelekasutuses esineb.
- **Distributiivse semantika hüpotees:**
sõnad, mis esinevad sarnastes kontekstides, kalduvad väljendama sarnast tähendust.
(Harris 1954)
 - Nt **halb** ja **kehv** (*ilm, olukord, tervis, enesetunne* jpm ühised sõnad)
- Polüseemia uurimisel on sarnasuste asemel fookuses pigem erinevused:
 - Kui üks sõna esineb mitmes erinevas kontekstis, siis sellest võib järeldada, et tal on mitu erinevat tähendust.
 - Mitme eri sõna tähenduste võrdlemise asemel võrreldakse polüseemia uurimisel omavahel **ühe sõna eri kasutusjuhtude tähendusi**, nt
 - *istusime **kuuma** päikese käes*
 - *andis **kuuma** suudluse*
 - *vaktsineerimine on **kuum** teema*

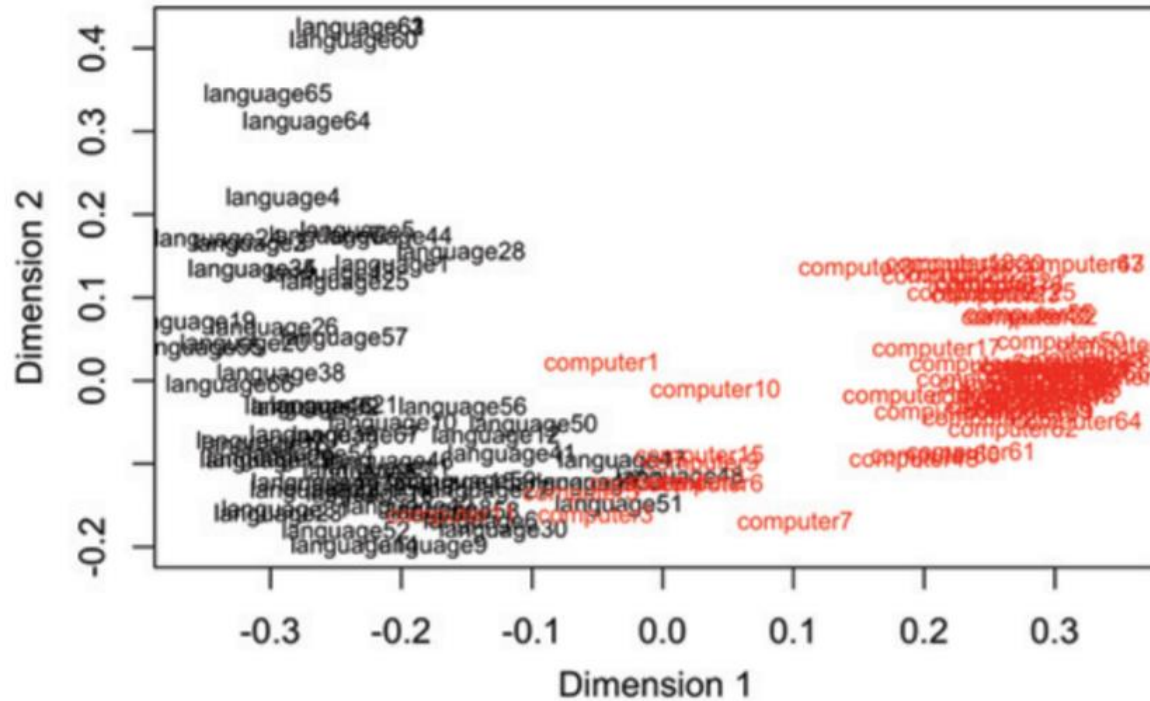
Distributiivsed mudelid: hollandi sõna *monitor* (Heylen jt 2015)

- Tähendused:
 - 'kuvar, ekraan'
 - 'noortejuht'
- Erineva tähendusega kasutusjuhtusid on võimalik eristada üksteisest leksikaalse konteksti alusel.
- Visualiseerimistehnikaks MDS
- Juhuvalim: korpusest 199 kasutusjuhtu sõnaga *monitor*



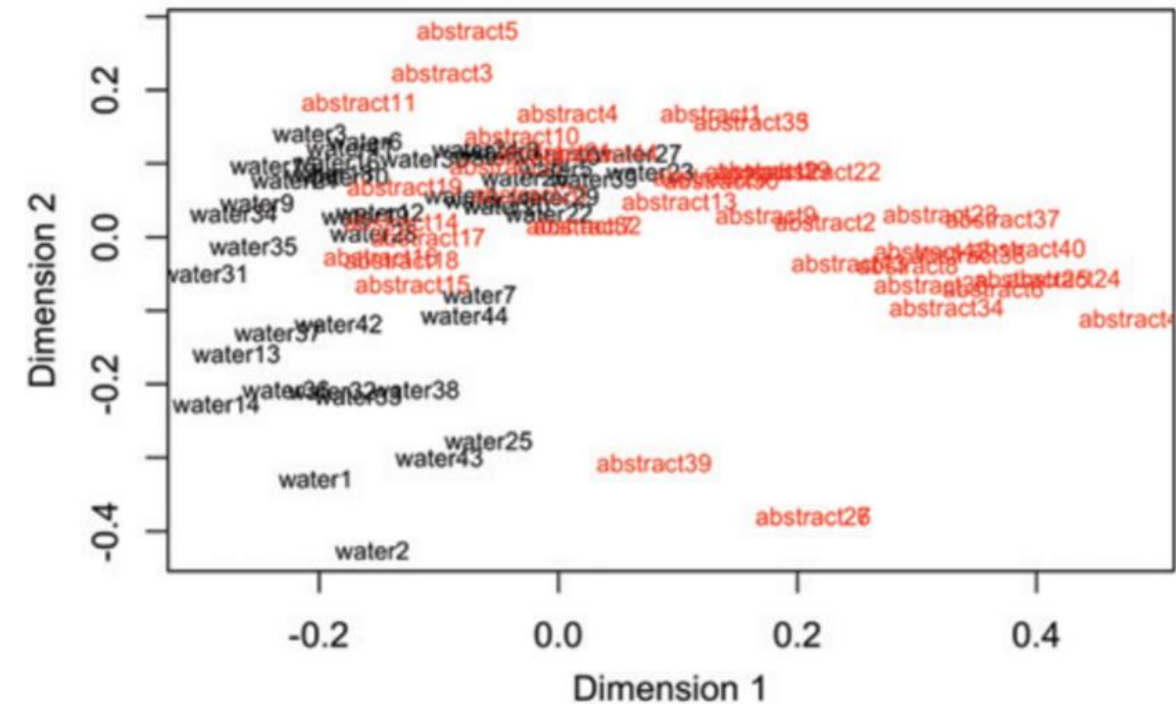
Distributiivsed mudelid: inglise sõnad *syntax* ja *wave* (Hilpert, Saavedra 2017)

syntax: language vs. computer



- Sõna *syntax* tähendused:
 1. loomuliku keelega seotud
 2. arvutiga seotud
- Tasakaalustatud proportsioonidega valim:
67 + 67 kasutusjuhtu

wave: water vs. abstract



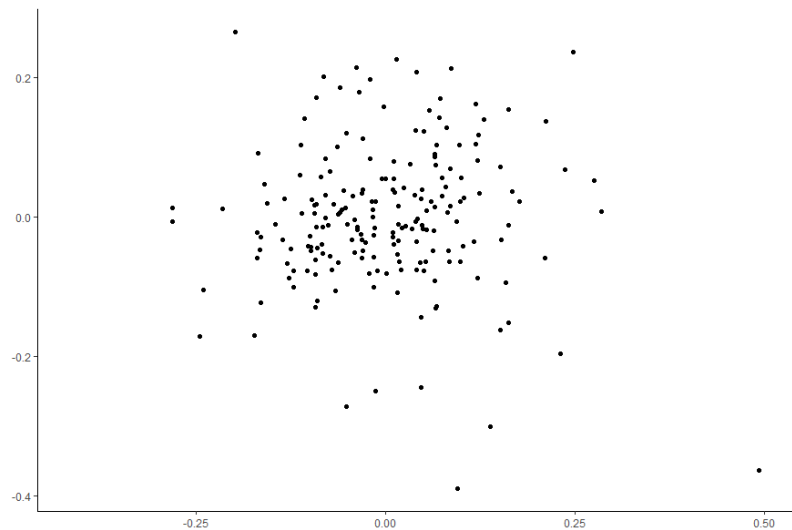
- Sõna *wave* tähendused:
 1. veelaine
 2. abstraktne, mitte veega seotud
- Tasakaalustatud proportsioonidega valim:
44 + 44 kasutusjuhtu

Minu uurimus: distributiivsed mudelid polüseemsete adjektiivide uurimisel

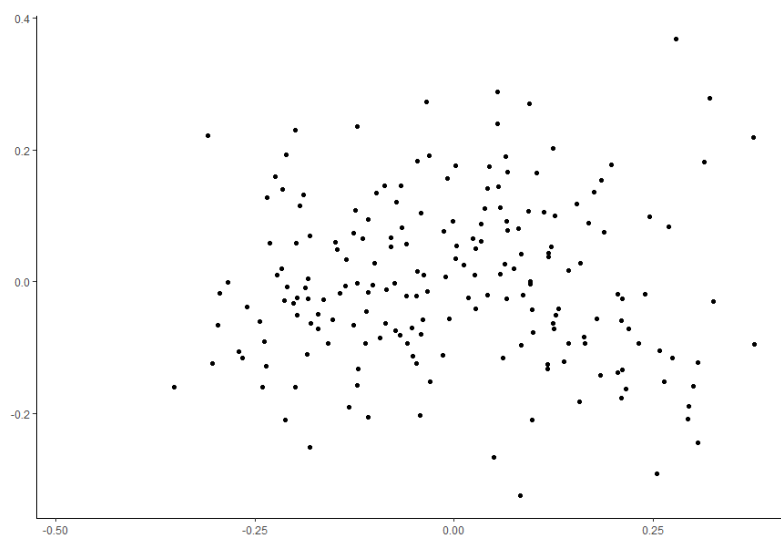
- Kahe ja kolme tähendusega adjektiivid
 - **juhuvalimid**
 - 200 kasutusjuhuga valimid Koondkorpusest
 - tähendusliigendused Sõnaveebist
- Omadussõna *kuum*
 - **erinevate põhimõtete alusel moodustatud valimid**
 - umbes 300 kasutusjuhuga valimid 2021. a ühendkorpusest
 - tähendusliigendus Sõnaveebist
- Taustaks: distributiivsete mudelite loomisel kasutasin
 - koosinussarnasuse alusel võrreldud sageduspõhiseid (PPMI) kasutusjuhuvektoreid
 - visualiseerimine: multidimensionaalne skaleerimine (MDS)
 - automaatne klasterdus: tiheduspõhine klasteranalüüs (HDBSCAN)

Kahe tähendusega (**konkreetne** ja **abstraktne**) adjektiivid: MDS

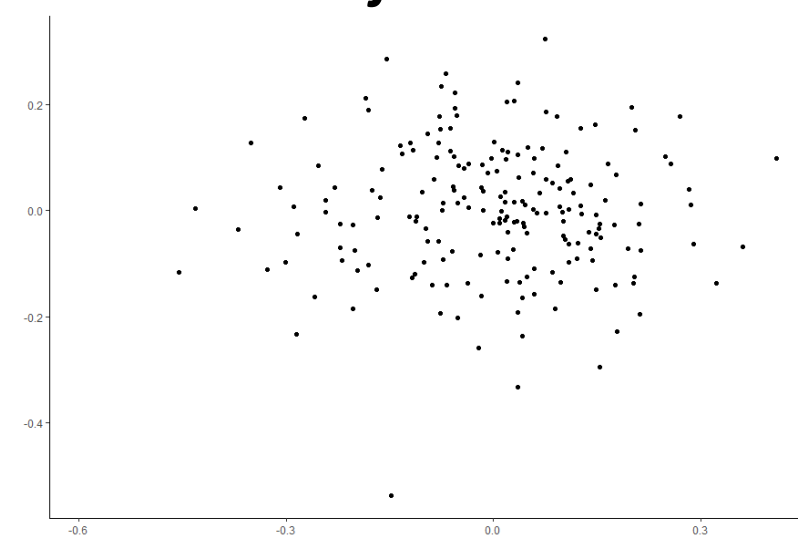
ere



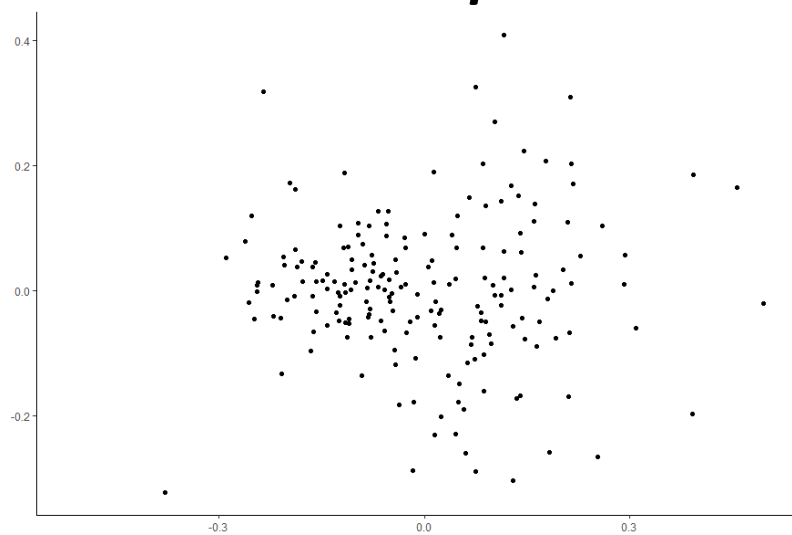
kahvatu



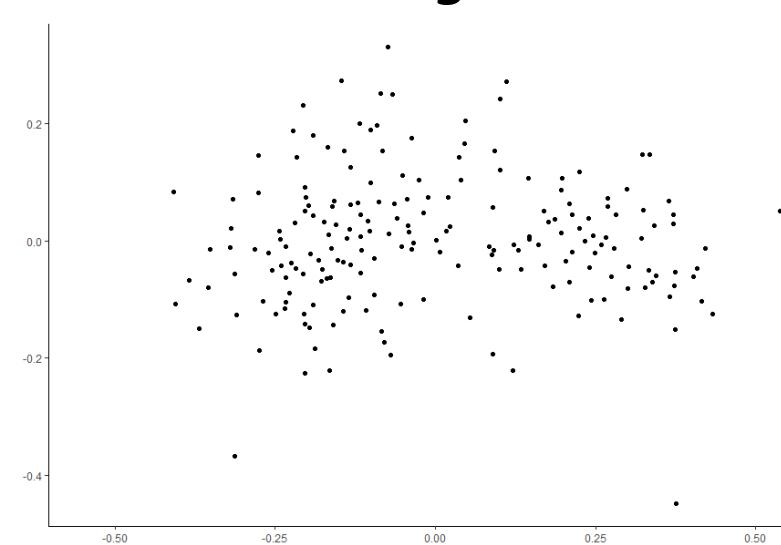
jahe



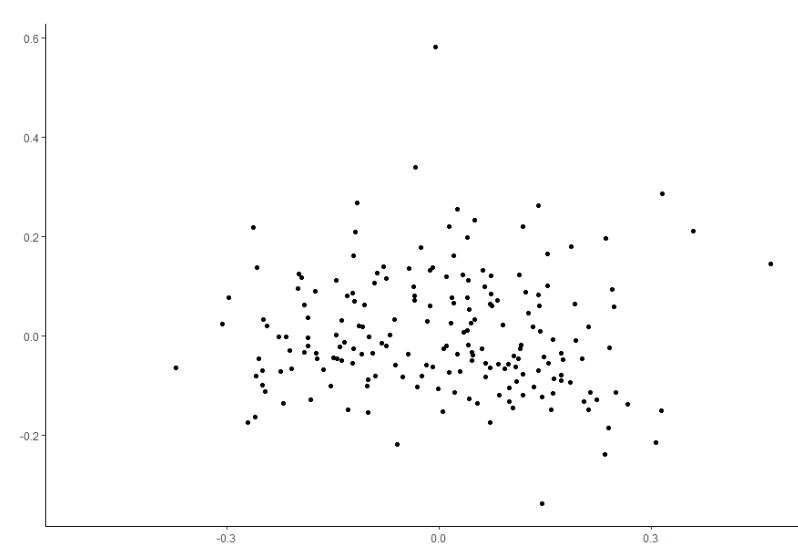
hapu



leige

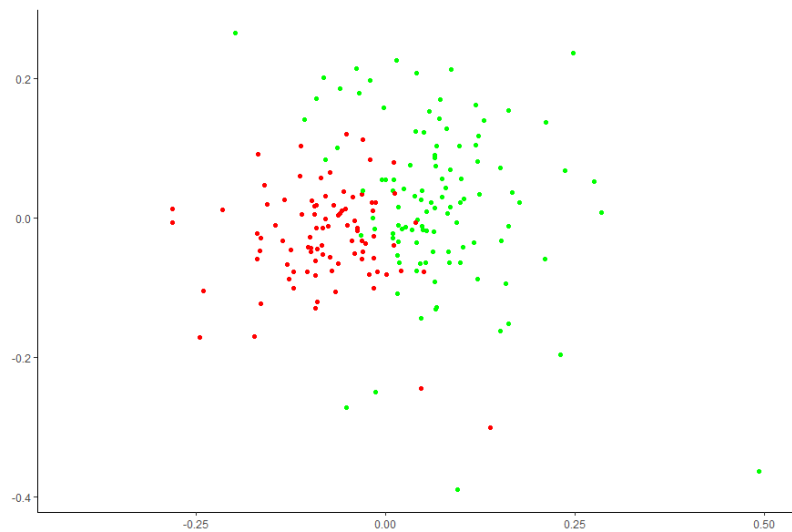


soolane

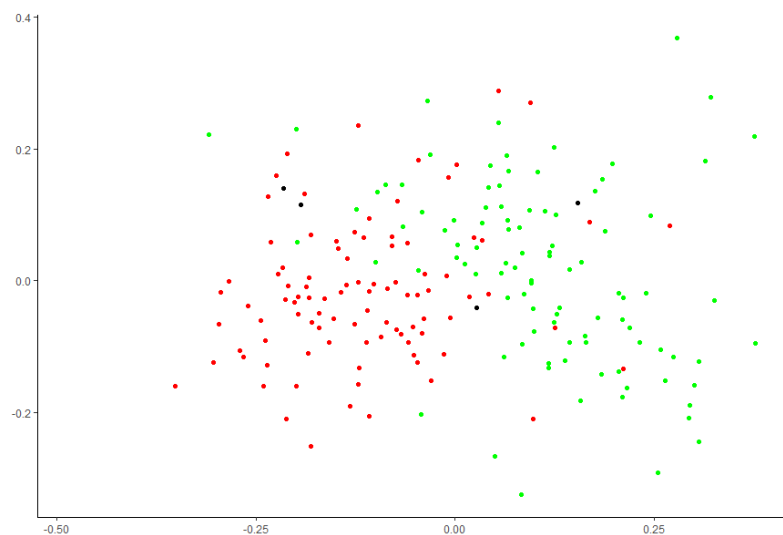


Kahe tähendusega (**konkreetne** ja **abstraktne**) adjektiivid: käsitsi märgendus

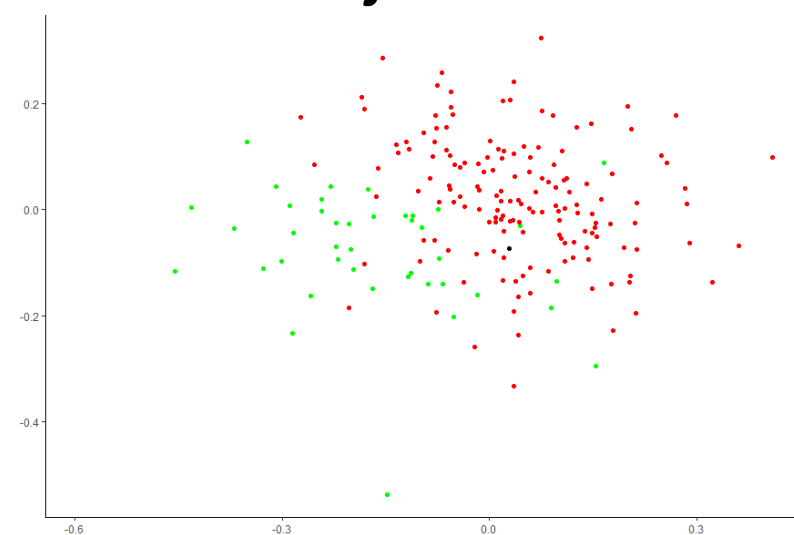
ere



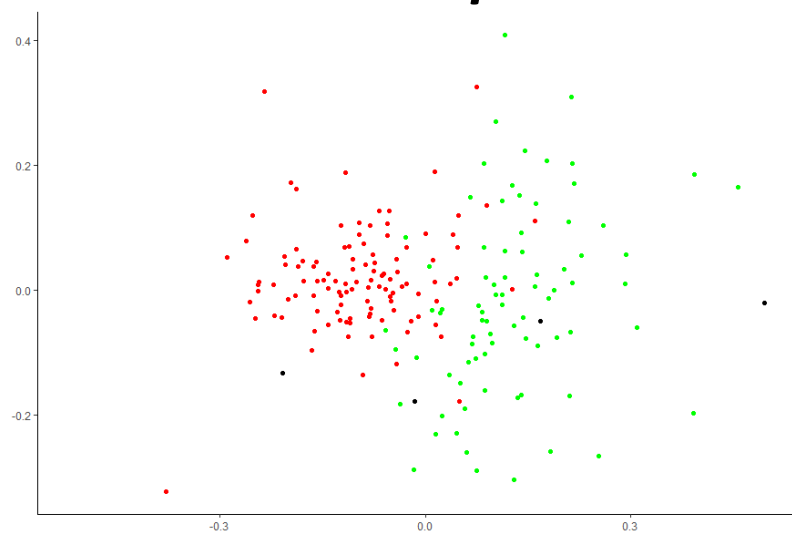
kahvatu



jahe



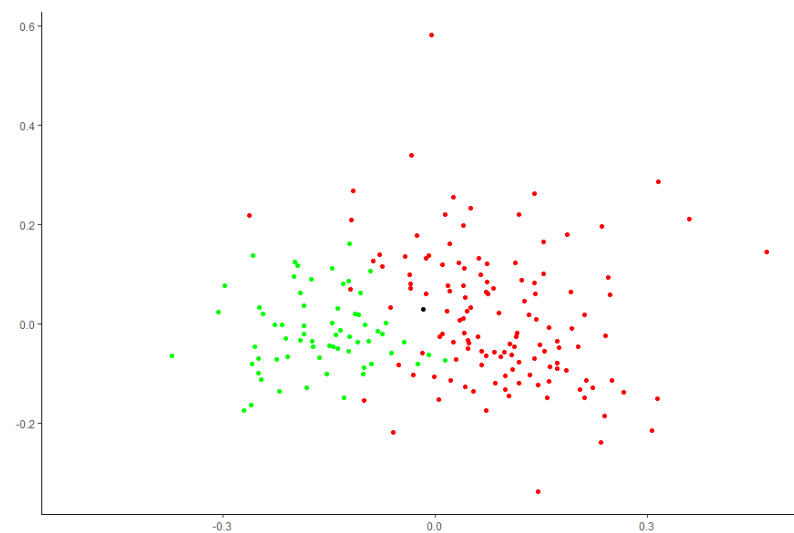
hapu



leige

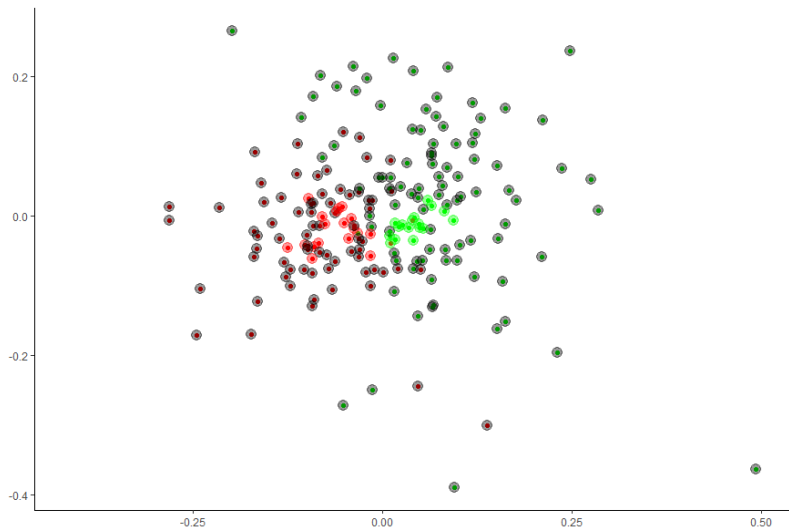


soolane

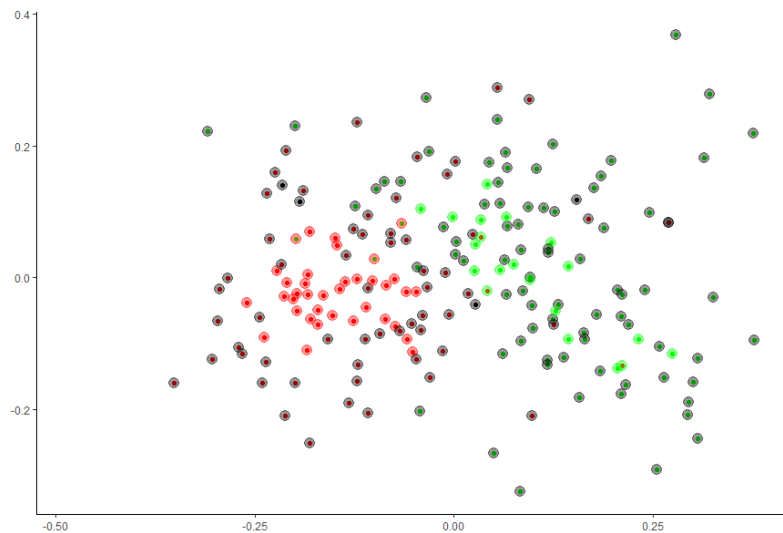


Kahe tähendusega (**konkreetne** ja **abstraktne**) adjektiivid: automaatne klasterdus

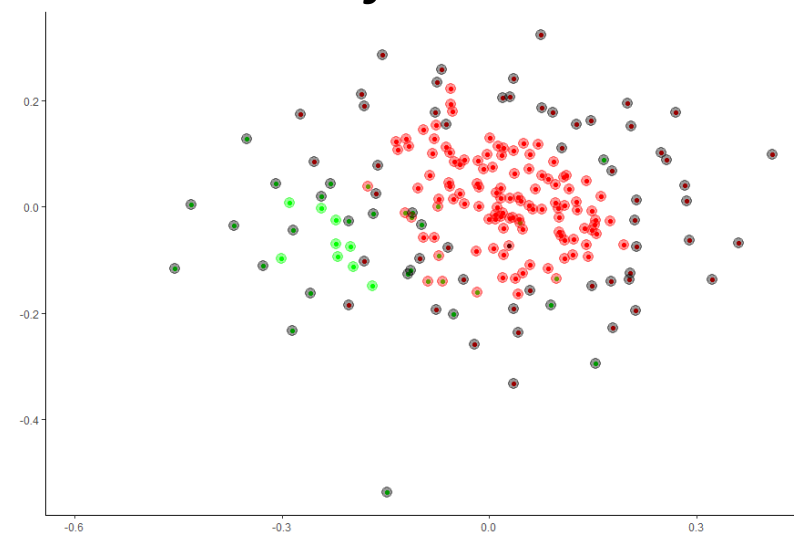
ere



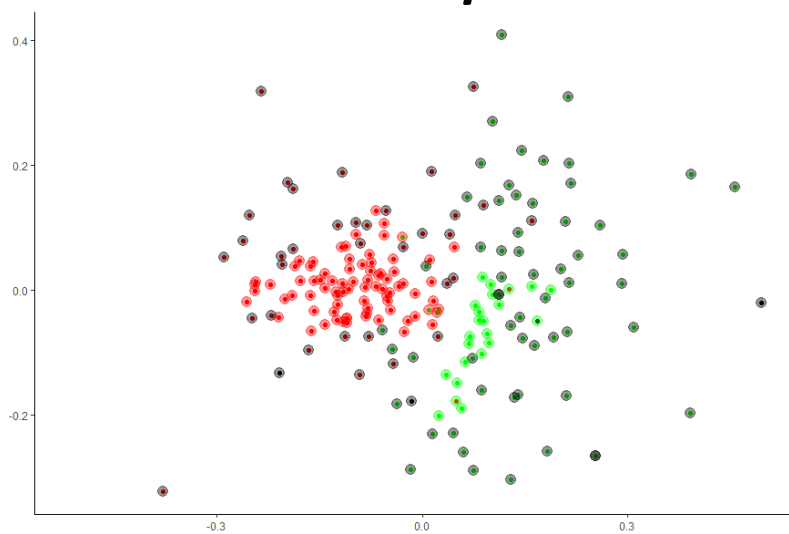
kahvatu



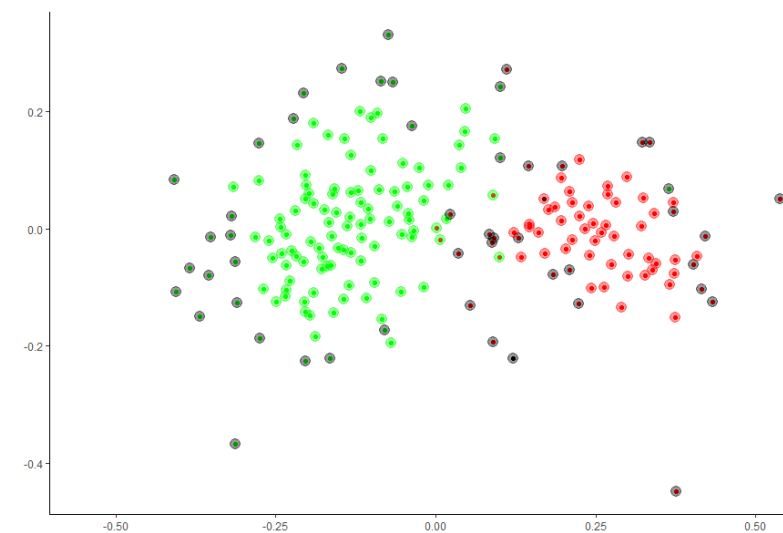
jahe



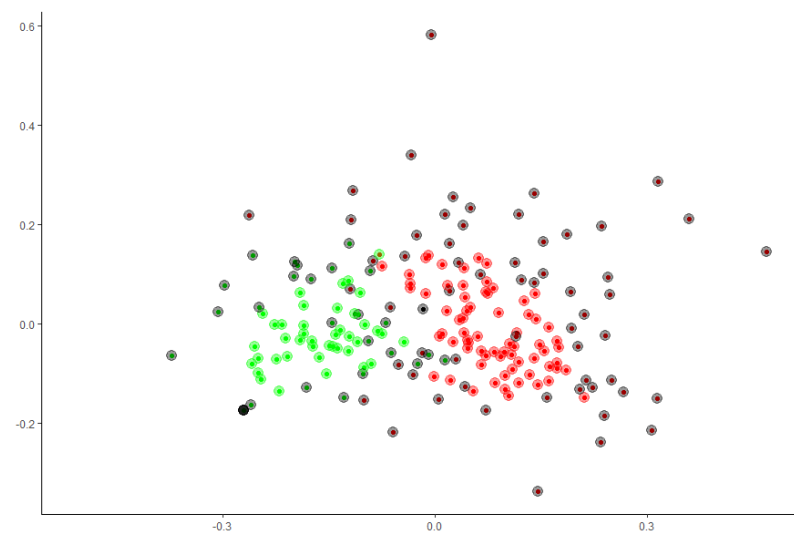
hapu



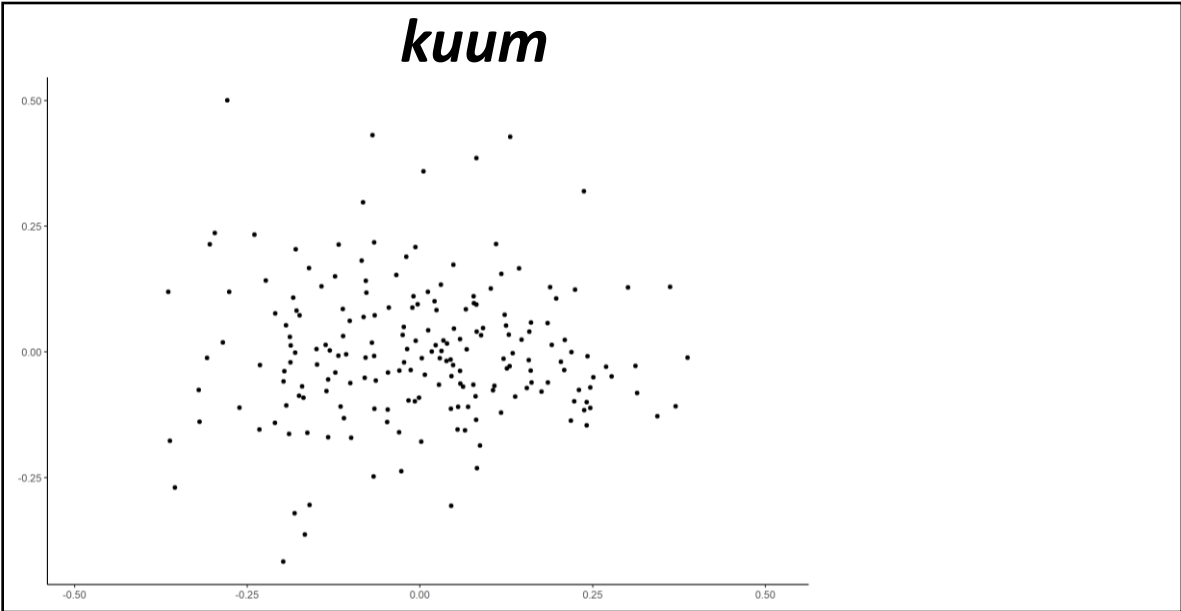
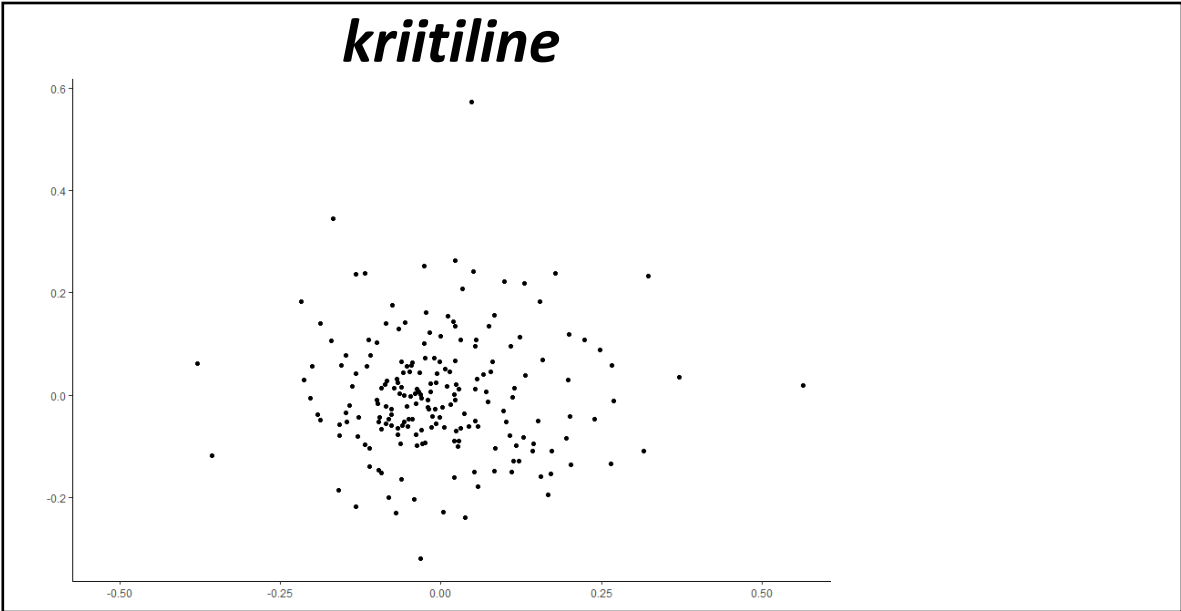
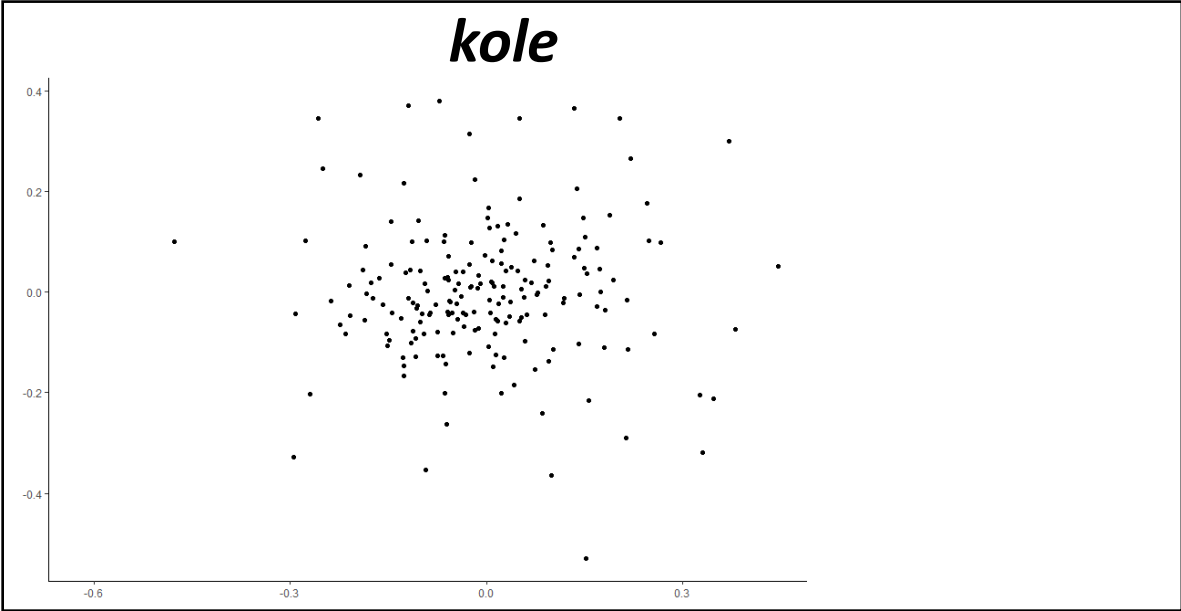
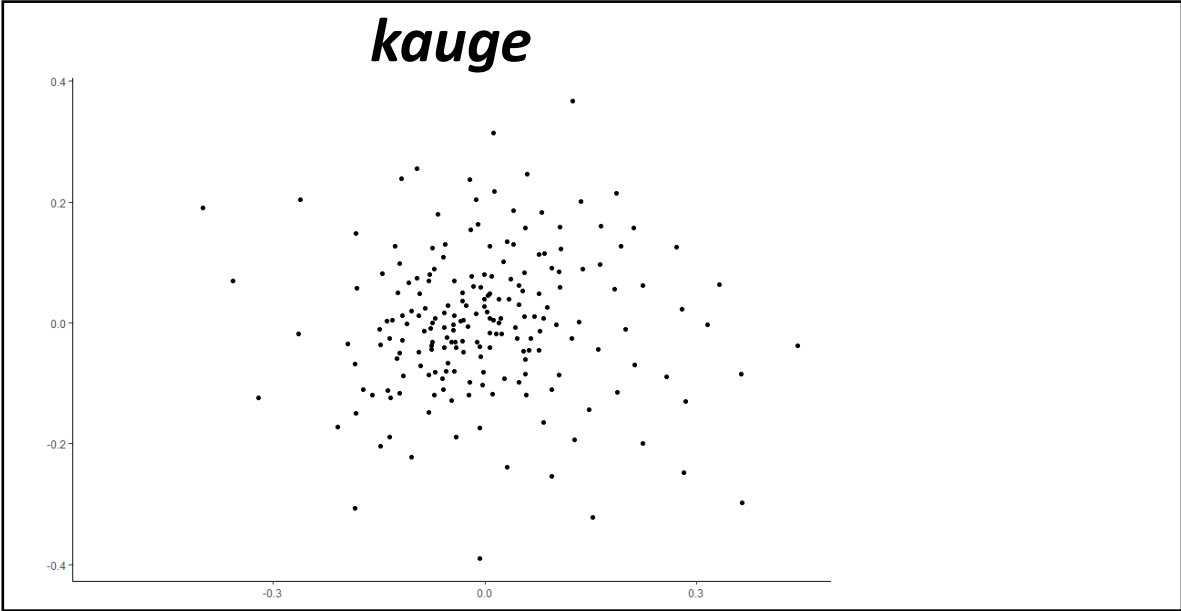
leige



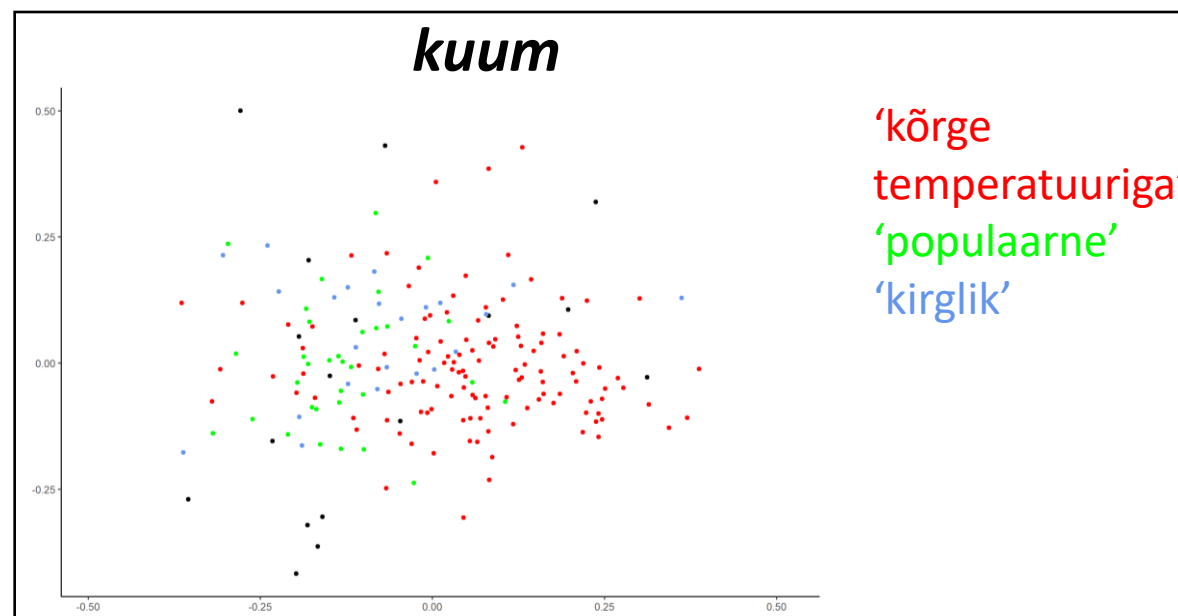
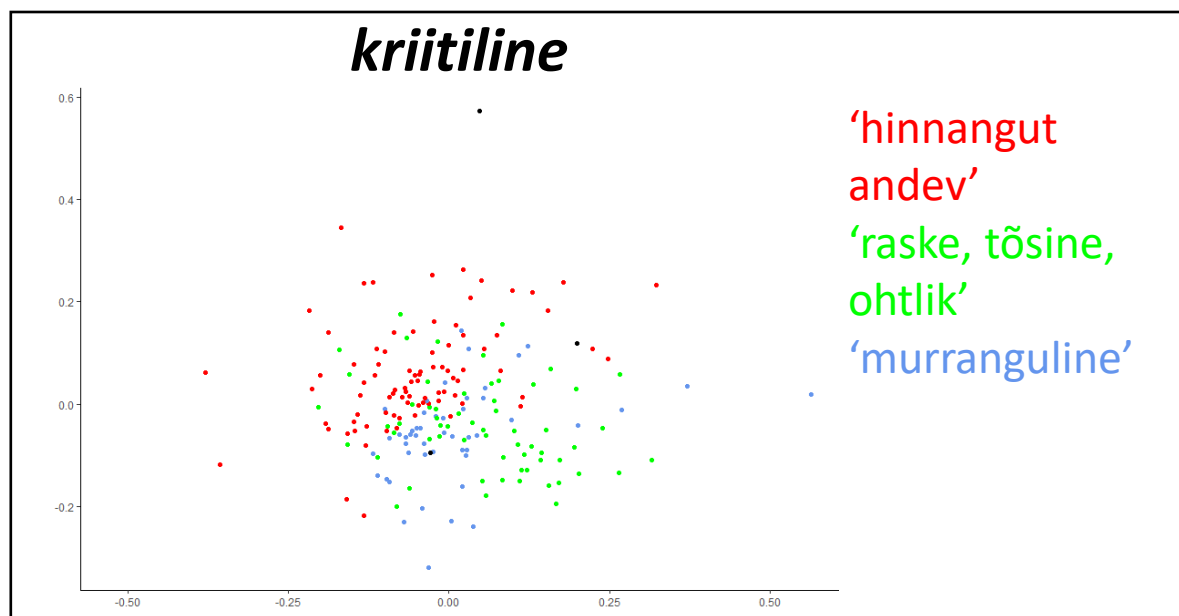
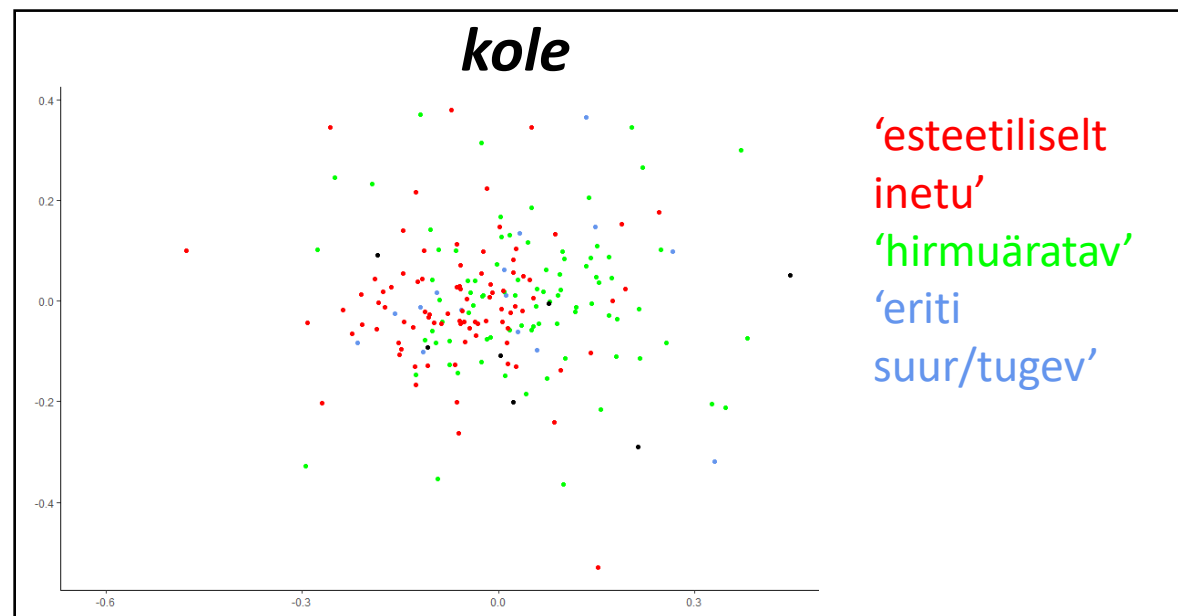
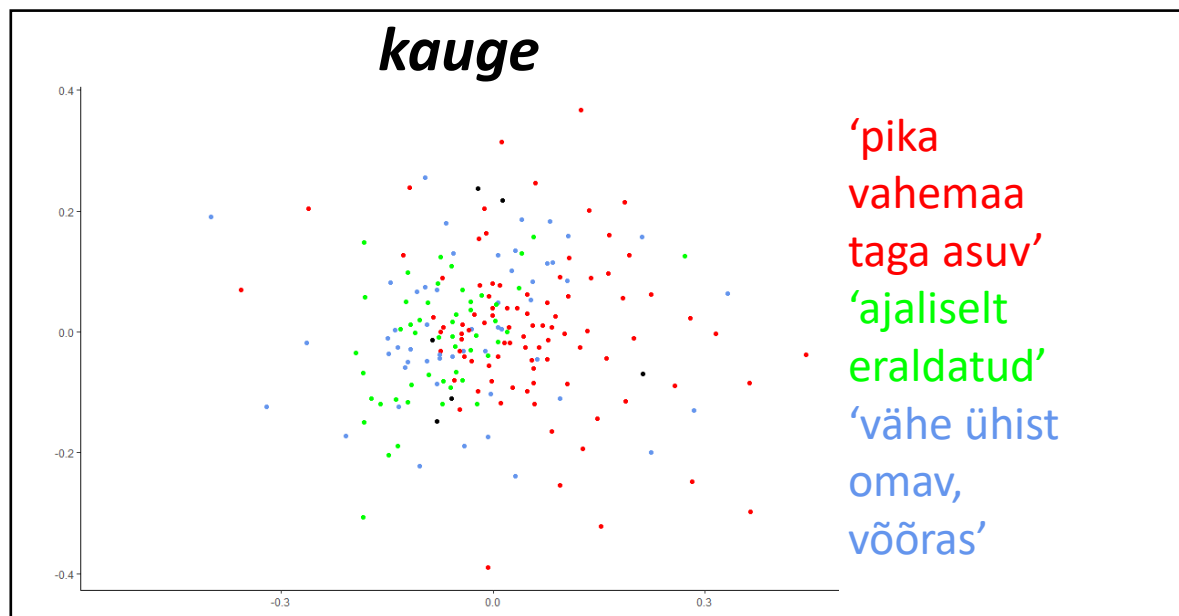
soolane



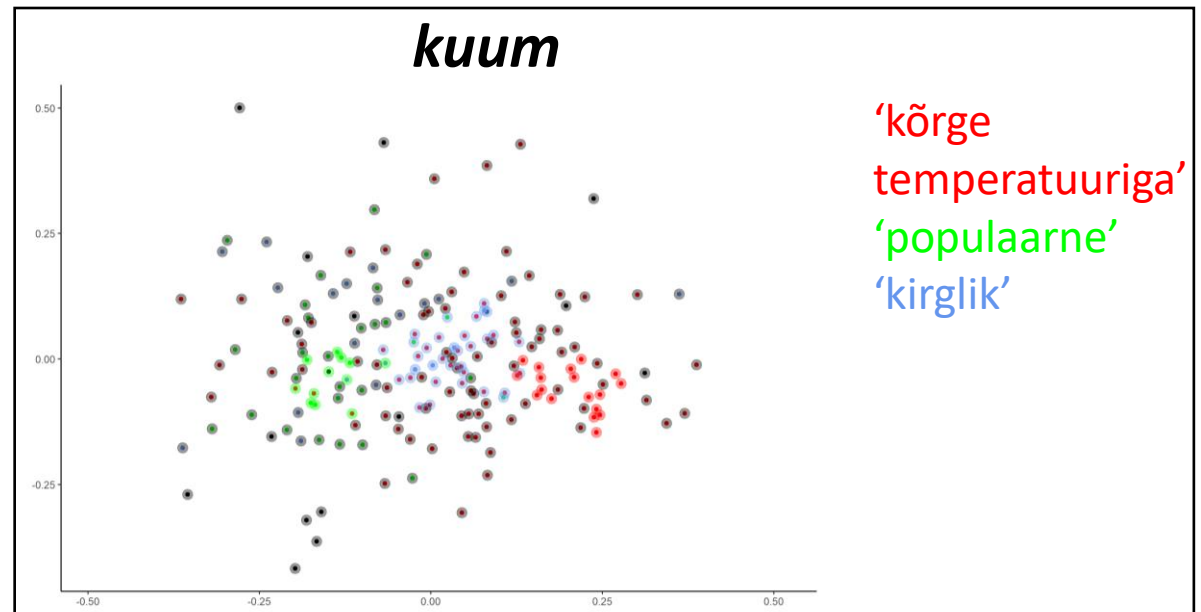
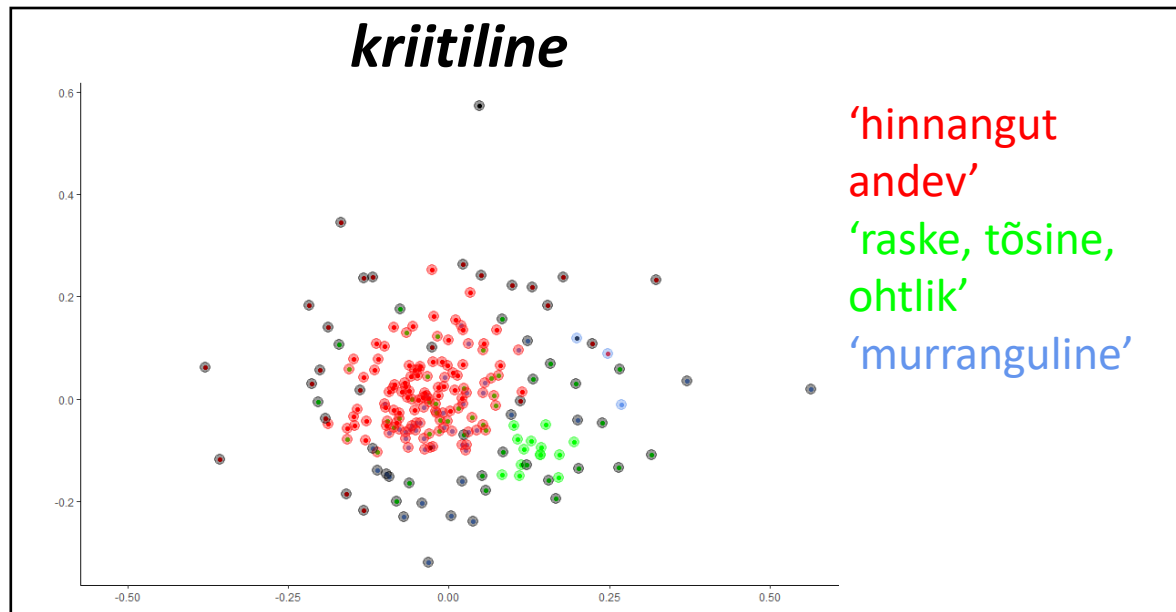
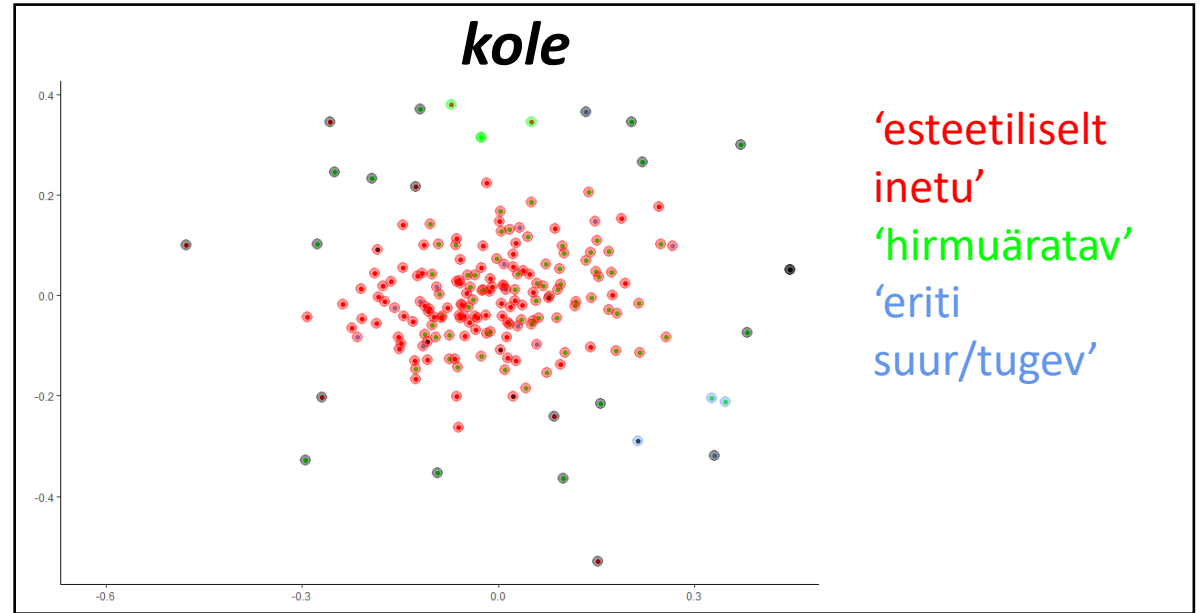
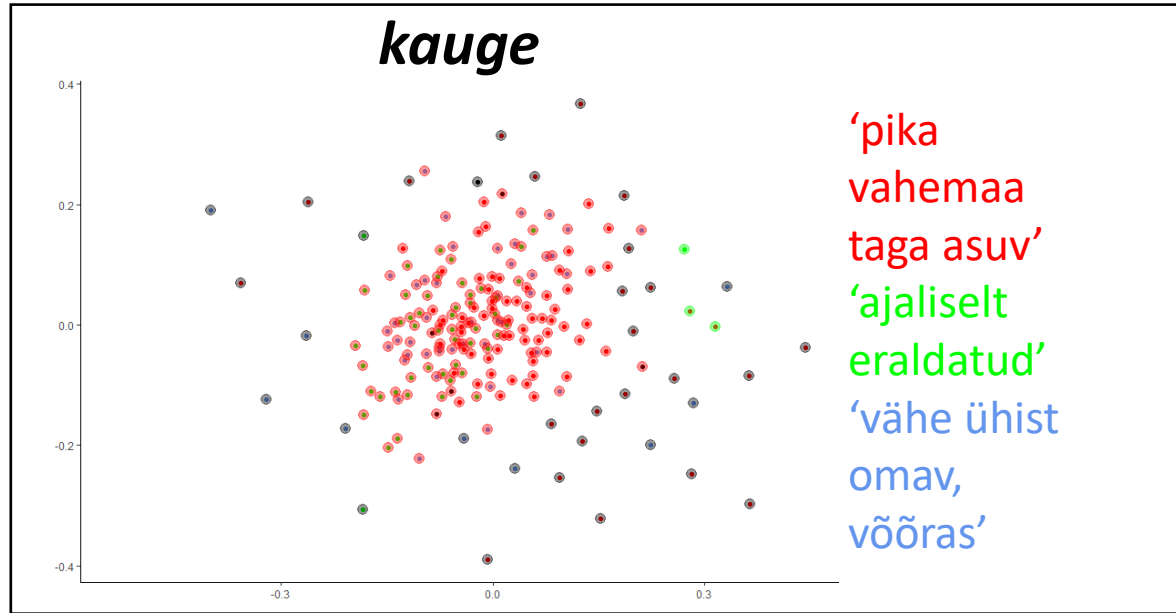
Kolme tähendusega adjektiivid: MDS



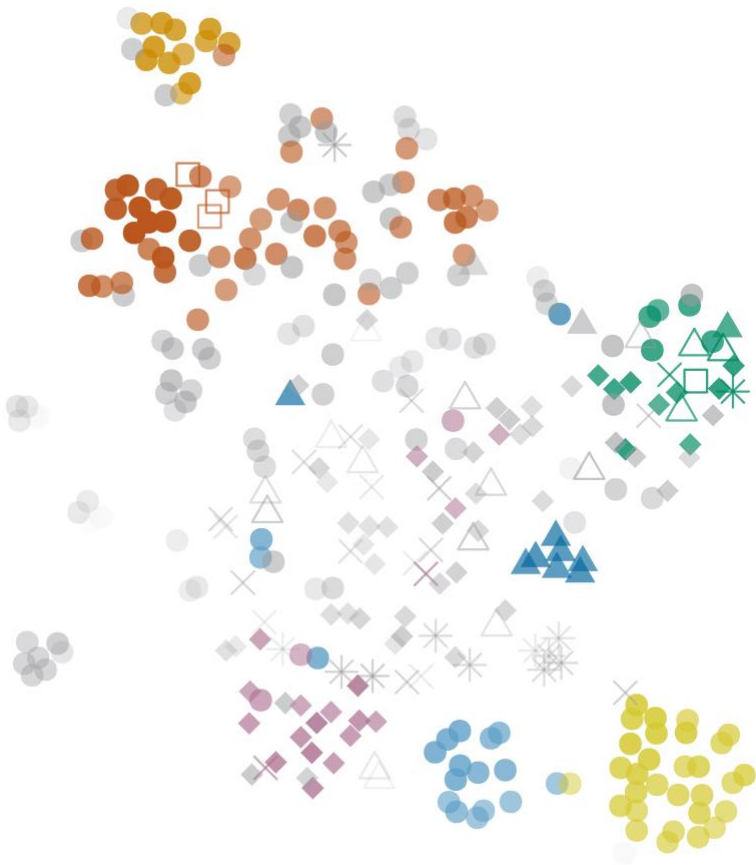
Kolme tähendusega adjektiivid: käsitsi märgendus (Sõnaveebi tähendusliigendused)



Kolme tähendusega adjektiivid: automaatne klasterdus



Hollandikeelne sõna *heep* 'kuum' (Montes 2021)



7 käsitsi märgendatud tähendust (andmepunkti kuju)

- (asjade kohta) 'väga soe'
- (ilma kohta) 'väga soe'
- (keha kohta) 'normaalsest kõrgema temperatuuriga'
- (toidu kohta) 'vürtsikas'
- (inimese kohta) 'atraktiivne'
- (sündmuste kohta) 'konfliktne'
- 'populaarne, huvitav või uus'

7 automaatanalüüsi klastrit (andmepunkti värv)



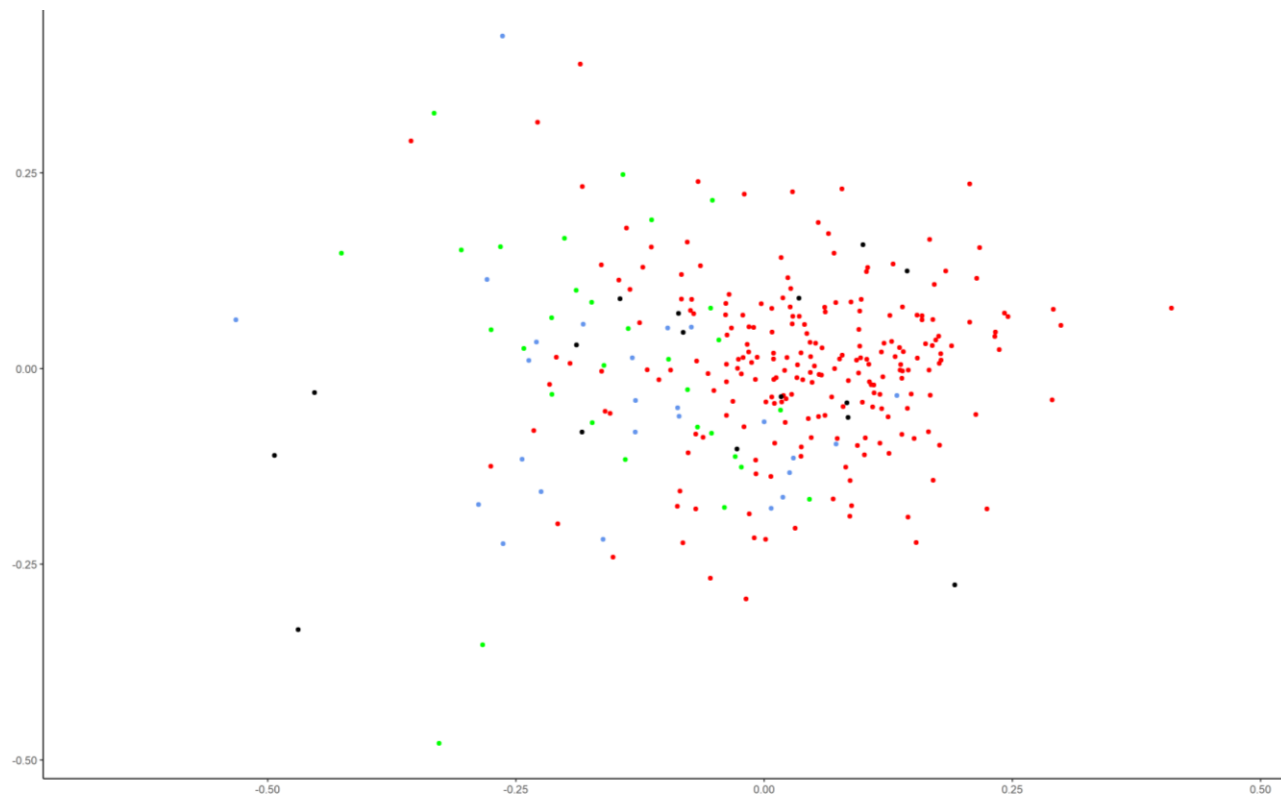
- Rohkem kui kahe tähendusega sõnade puhul ei ole **kontekstuaalsel sarnasusel** põhinevad automaatanalüüsiga saadavad klastrid üksüheses vastavuses **polüseemse sõna tähendustega**.
- Juhuvalimis on eri tähendusega kasutusjuhtusid erinevates proportsioonides.
- > Proovida tasakaalustada tähenduste proportsioone valimites (nagu Hilpert, Saavedra 2017)

Valim distributiivsemantiliste mudelite varieeritava parameetrina

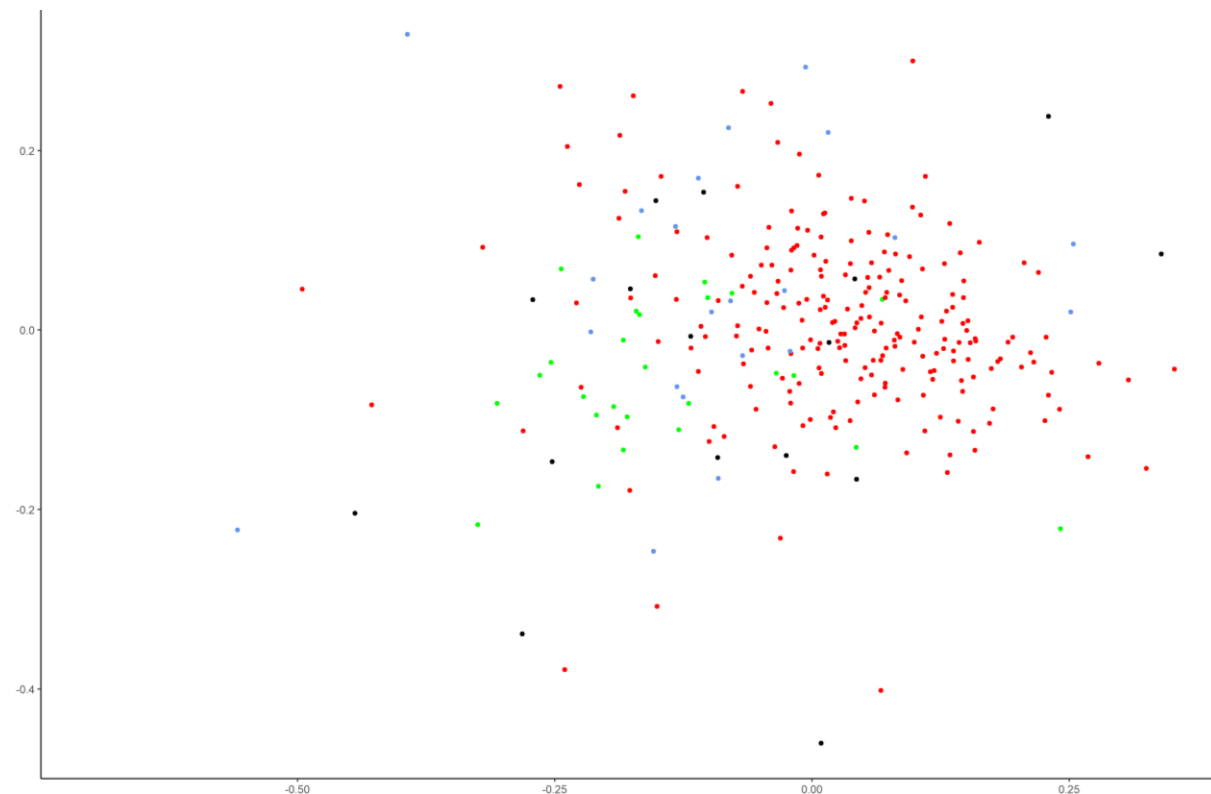
- Mida on võimalik teha distributiivsemantiliste mudelitega selle **meetodi praeguses arenguetapis?**
 - 1) ~~Kasutada polüseemse sõna tähenduste väljaselgitamise vahendina.~~
 - 2) Kasutada mingile introspeksioonipõhisele tähendusliigendusele empiirilise kinnituse saamise vahendina.
- **Valim** kui distributiivsemantiliste mudelite varieeritav parameeter
 - valimi moodustamise 6 erinevat põhimõtet
 - täiesti juhuslik >> tasakaalustatud proportsioonidega valim
- Näited omadussõna *kuum* analüüsides
 - valimid 2021. a ühendkorpusest (iga põhimõtte kohta kaks umbes 300 kasutusjuhuga)
 - tähendusliigendus Sõnaveebist
 - 'kõrge temperatuuriga'
 - 'populaarne'
 - 'kirglik'

Põhimõte (1): täiesti juhuslik valim (*kuum* omadussõnana)

valim1: *kuum*_(1)21nc300



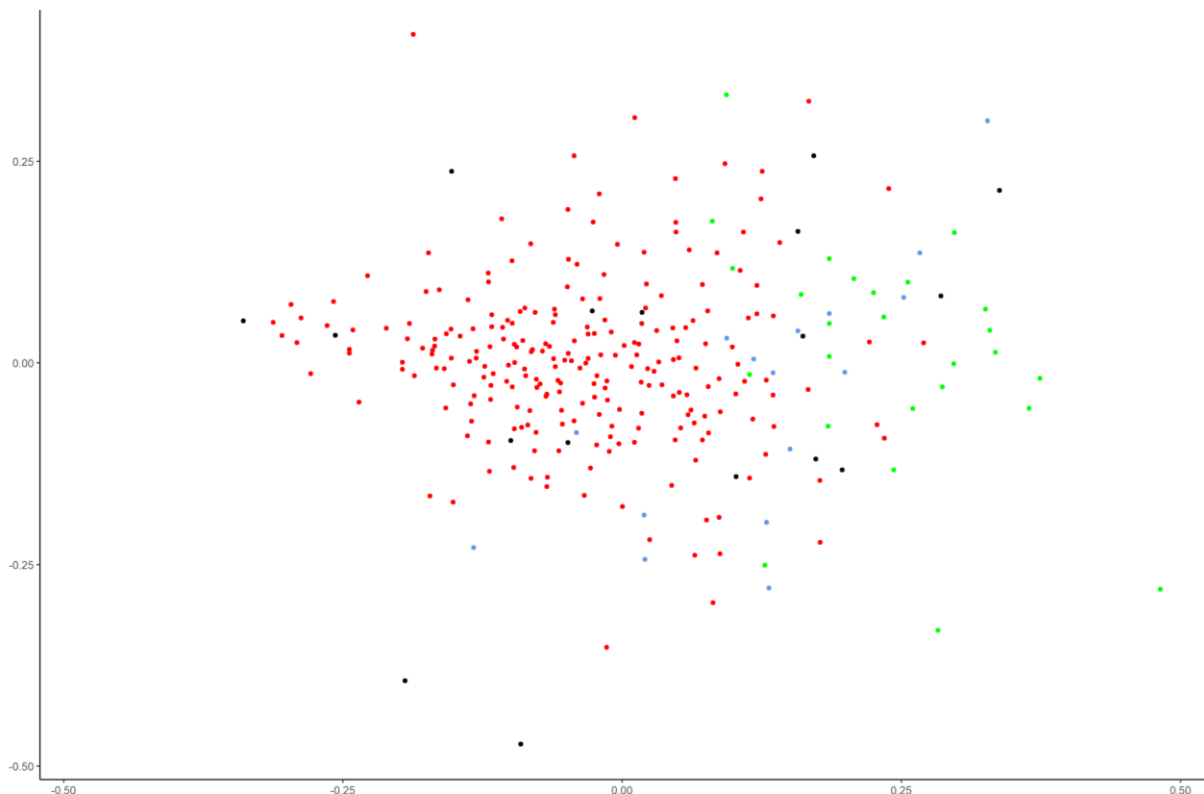
valim2: *kuum*_(1)21nc301



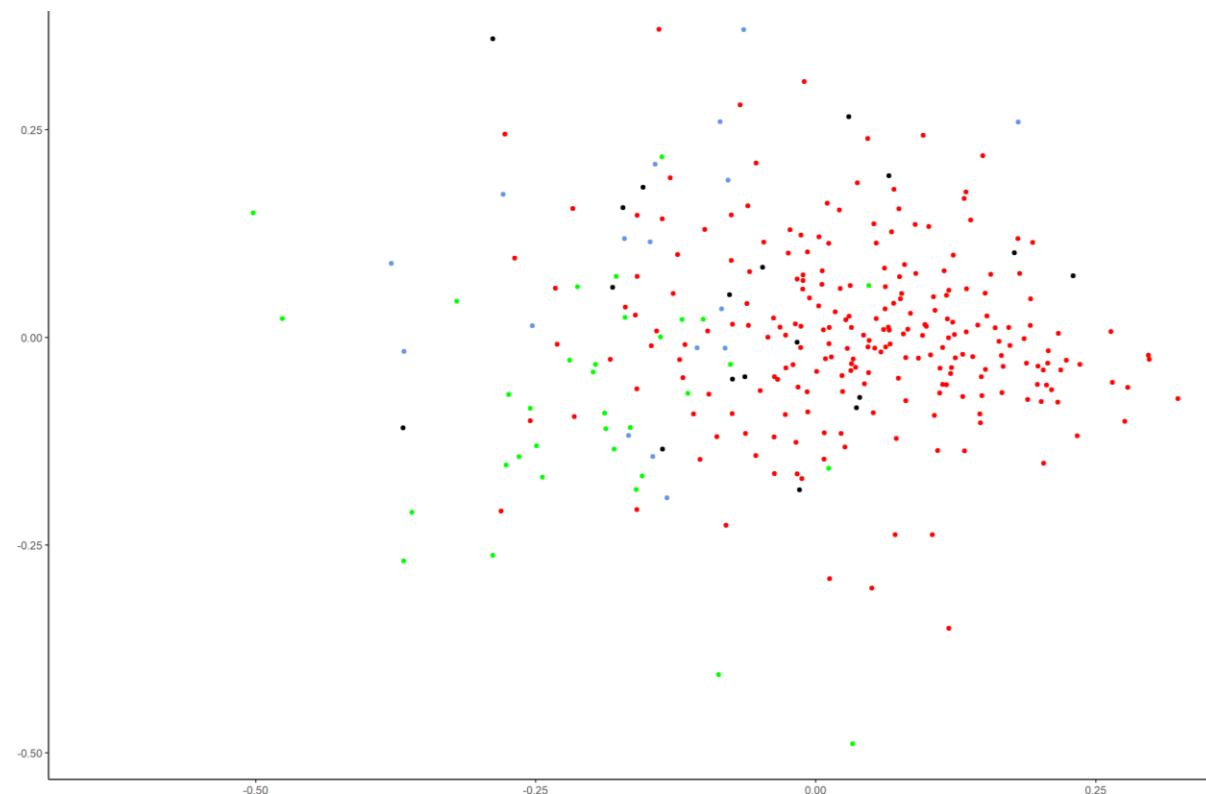
'kõrge
temperatuuriga'
'populaarne'
'kirglik'
määramata

Põhimõte (2): Adj + S ehk täiendi positsioonis omadussõnaga kasutusjuhtud

valim1: *kuum_(2)21nc300*



valim2: *kuum_(2)21nc301*



'kõrge
temperatuuriga'
'populaarne'
'kirglik'
märgendamata

Valimi piiramine täiendi positsioonis
omadussõna kasutusjuhtudega on
vajalik valimite moodustamise
järgmiste põhimõtete jaoks.

Omadussõna *kuum* tugevaimad nimisõnalised kollokaadid

(Eesti keele naabersõnade sõnastik)

kuum¹ omadussõna

1. kõrge temperatuuriga, väga soe

Nimisõnaga

kuum vesi
kuum ilm
kuum tee
kuum suvi
kuum jook
kuum suvepäev
kuum päike
kuum õhk
kuum päev
kuum kohv
kuum pann
kuum ahi
kuum õli
kuum aur
kuum saun
kuum koht

2. KÕNEK populaarne

Nimisõnaga

kuum teema
kuum kaup
kuum hitt
kuum kõneaine

3. kirklik, talitsematu; seksikas

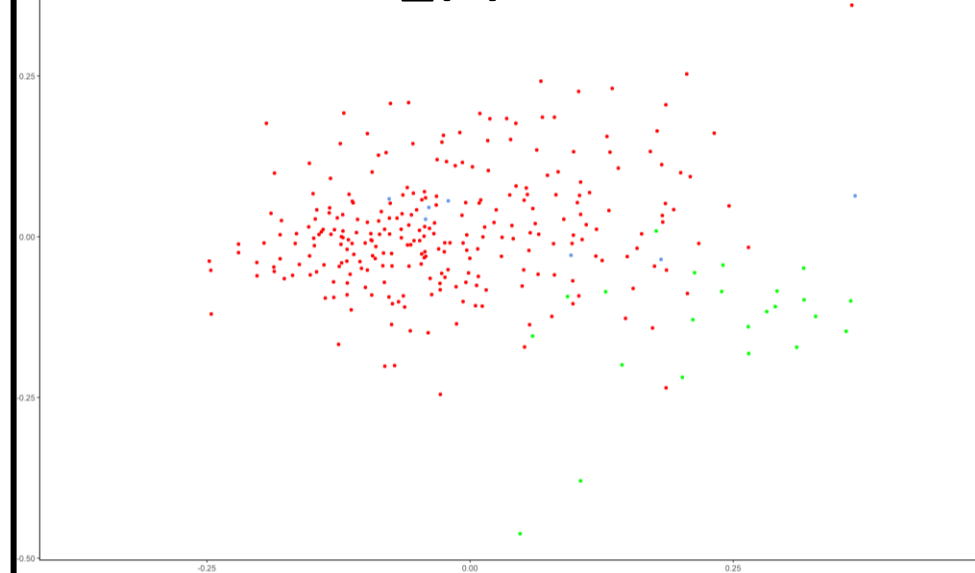
Nimisõnaga

kuum suhe
kuum mees
kuum suudlus
kuum seks
kuum keha
kuum naine
kuum temperament
kuumad tunded
kuumad rütmid

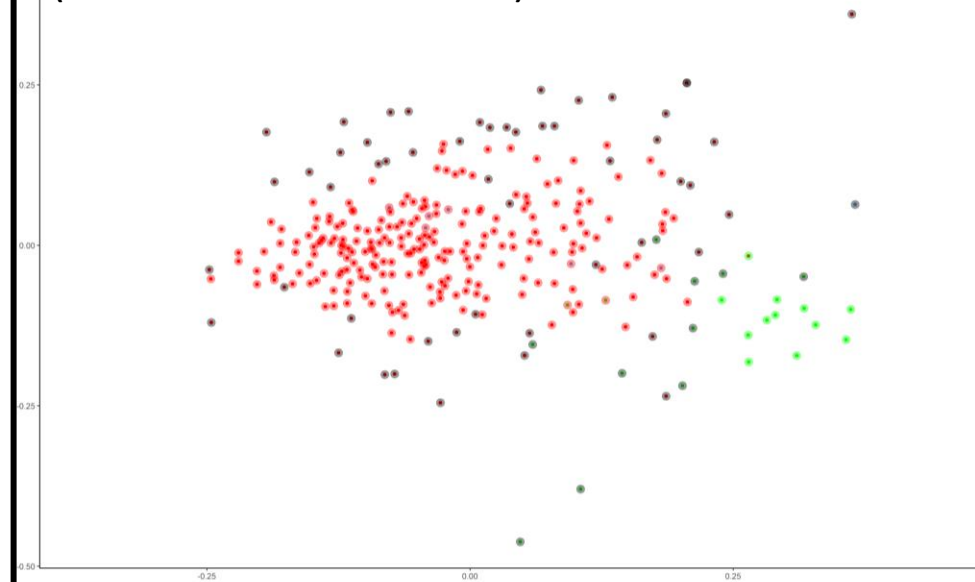
- Põhimõtte (3) jaoks veel üks piirang: nimisõna, mida omadussõna täiendab, peab kuuluma selle omadussõna tugevaimate kollokaatide hulka.
- Selle jaoks, et **tähenduste märgendust** saaks hakata tegema **automaatselt** nimisõnaliste kollokaatide põhjal (nt *kuum* koos sõnaga *vesi* > tähendus 'kõrge temperatuuriga').

Põhimõte (3): Adj + S, nimisõna tugevaimate kollokaatide hulgast

valim1: *kuum_(3)21nc300*

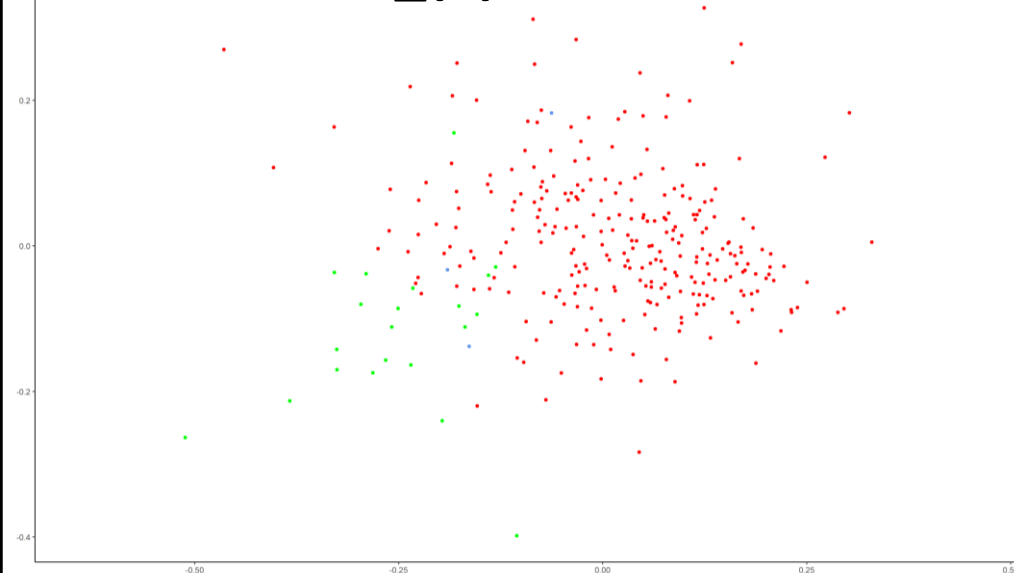


(automaatne klasterdus)

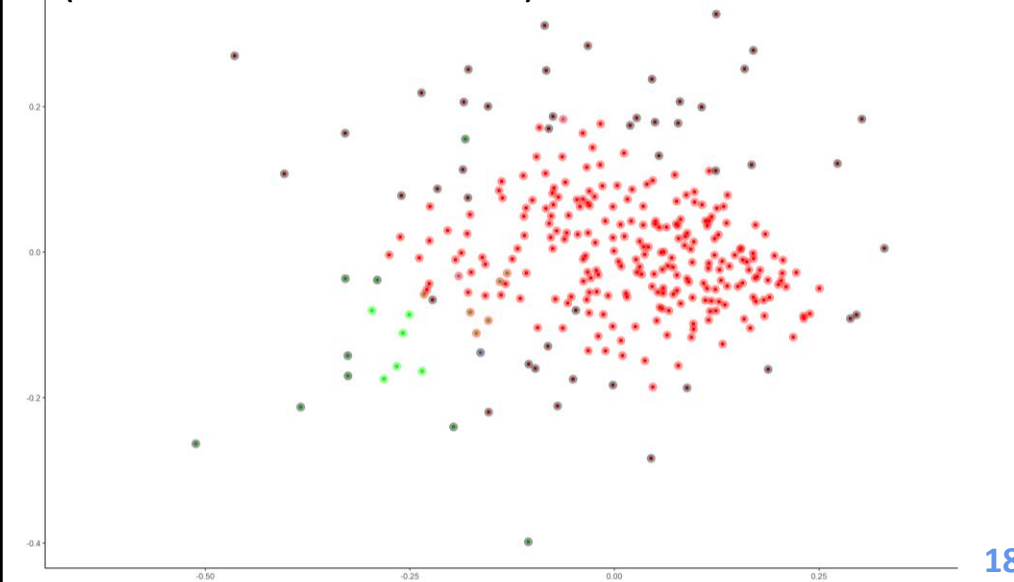


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

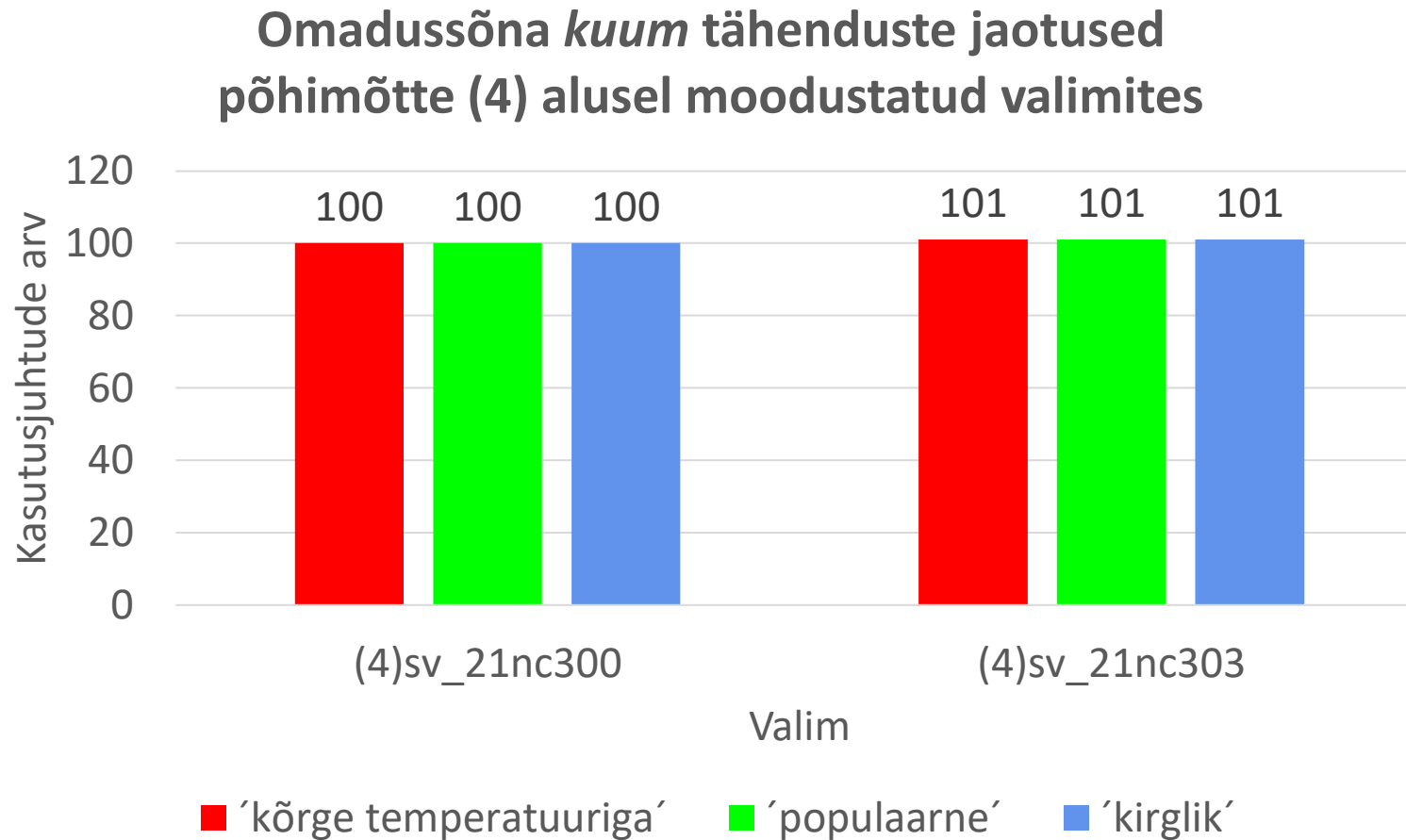
valim2: *kuum_(3)21nc301*



(automaatne klasterdus)

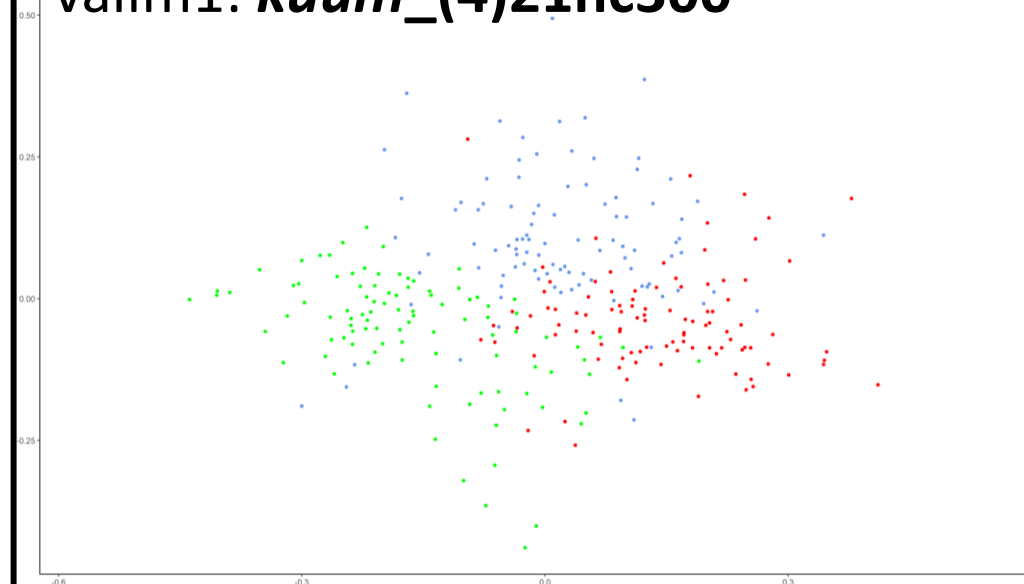


Põhimõtte (4): Adj + S, tähenduste proportsioonid tasakaalustatud

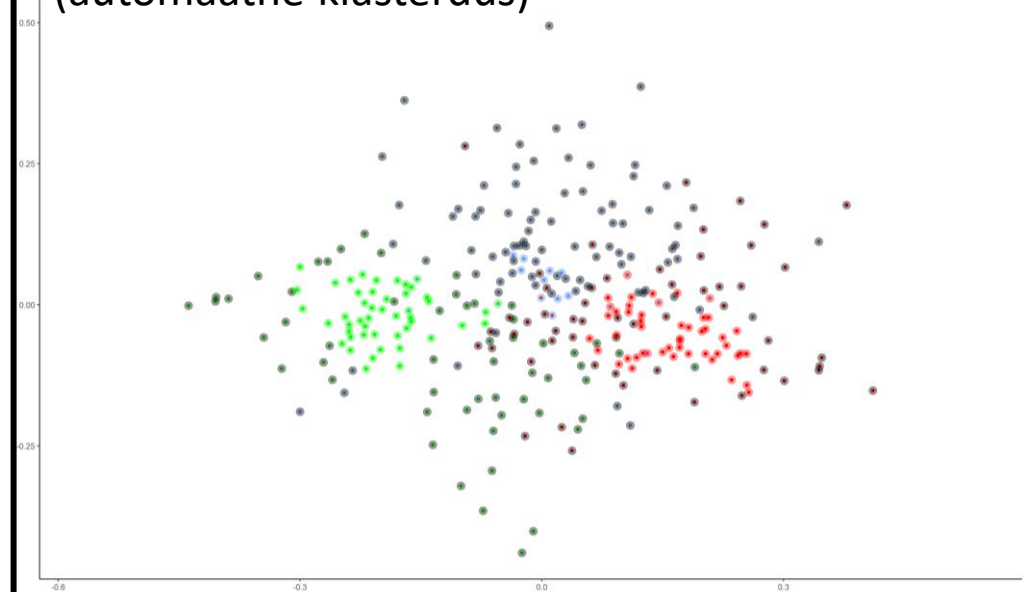


Põhimõte (4): Adj + S, tähenduste proportsioonid tasakaalustatud

valim1: *kuum_(4)21nc300*

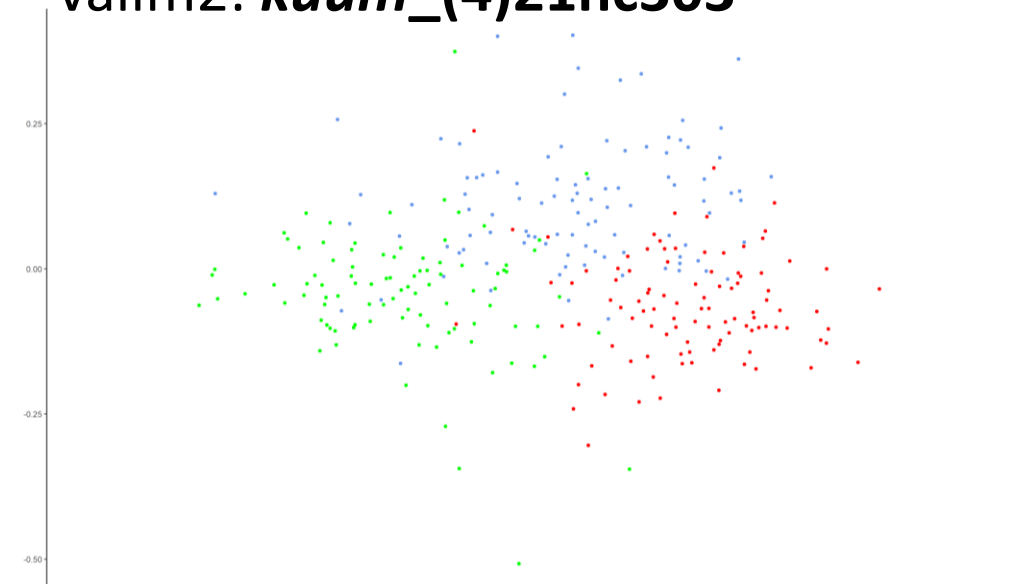


(automaatne klasterdus)

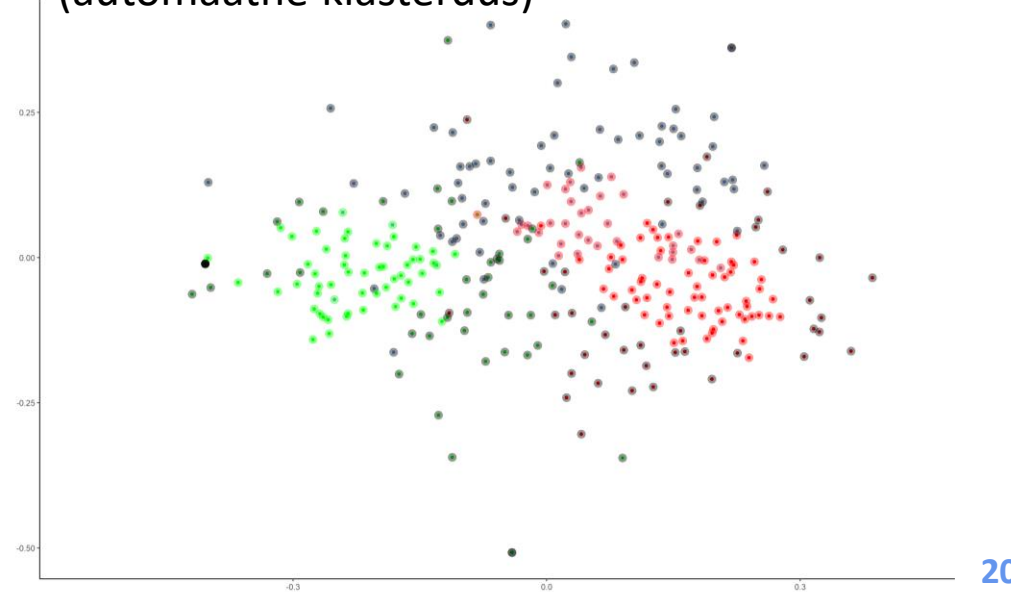


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

valim2: *kuum_(4)21nc303*

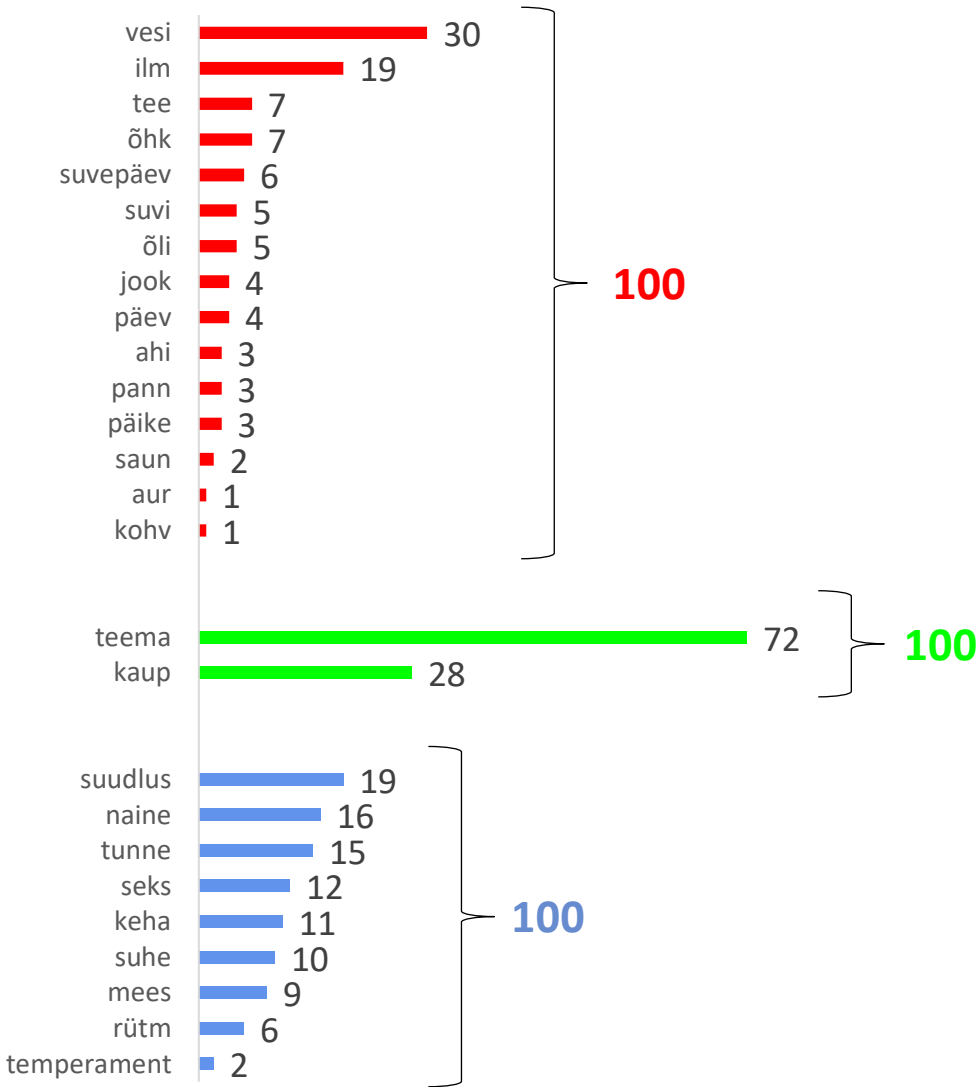


(automaatne klasterdus)



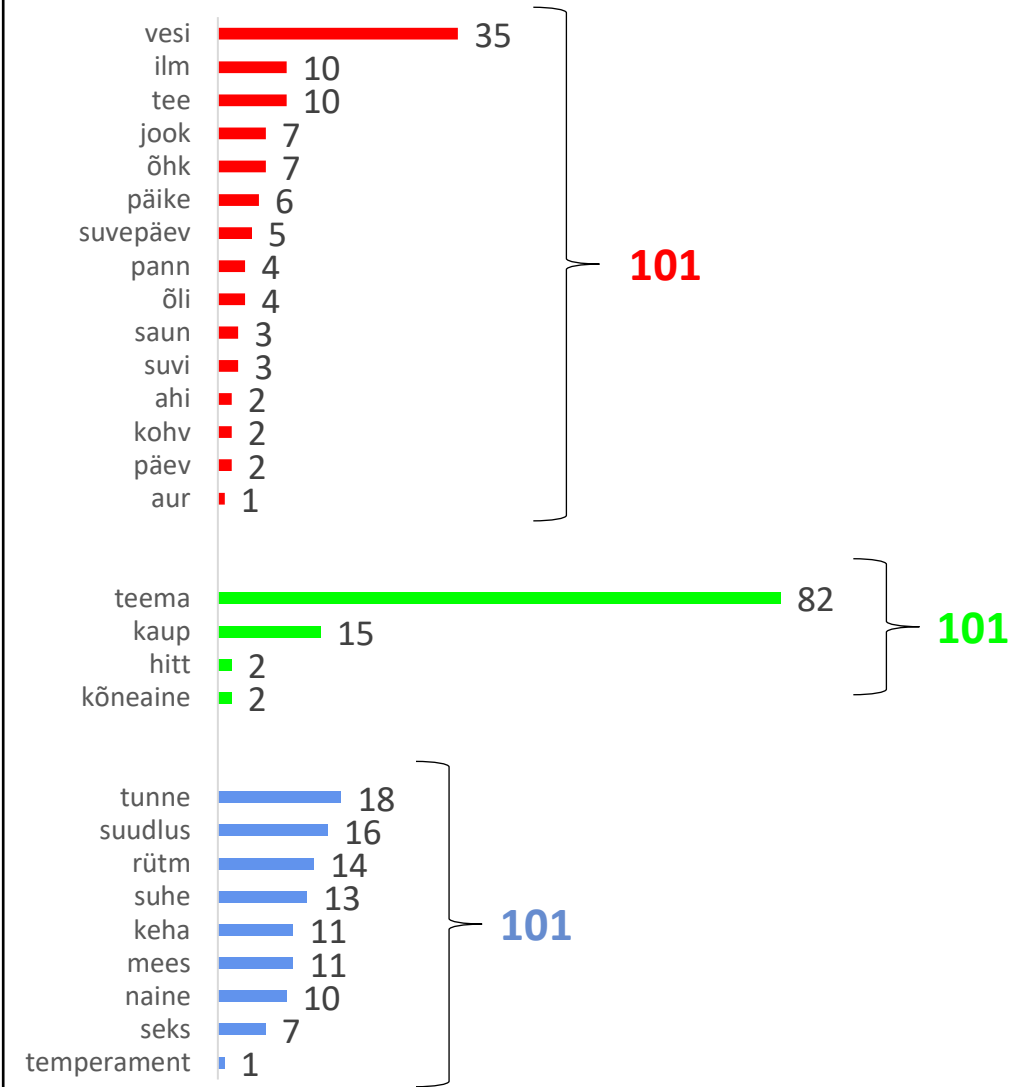
Põhimõte (4): Adj + S, tähenduste proportsioonid tasakaalustatud

kuum_(4)sv_21nc300

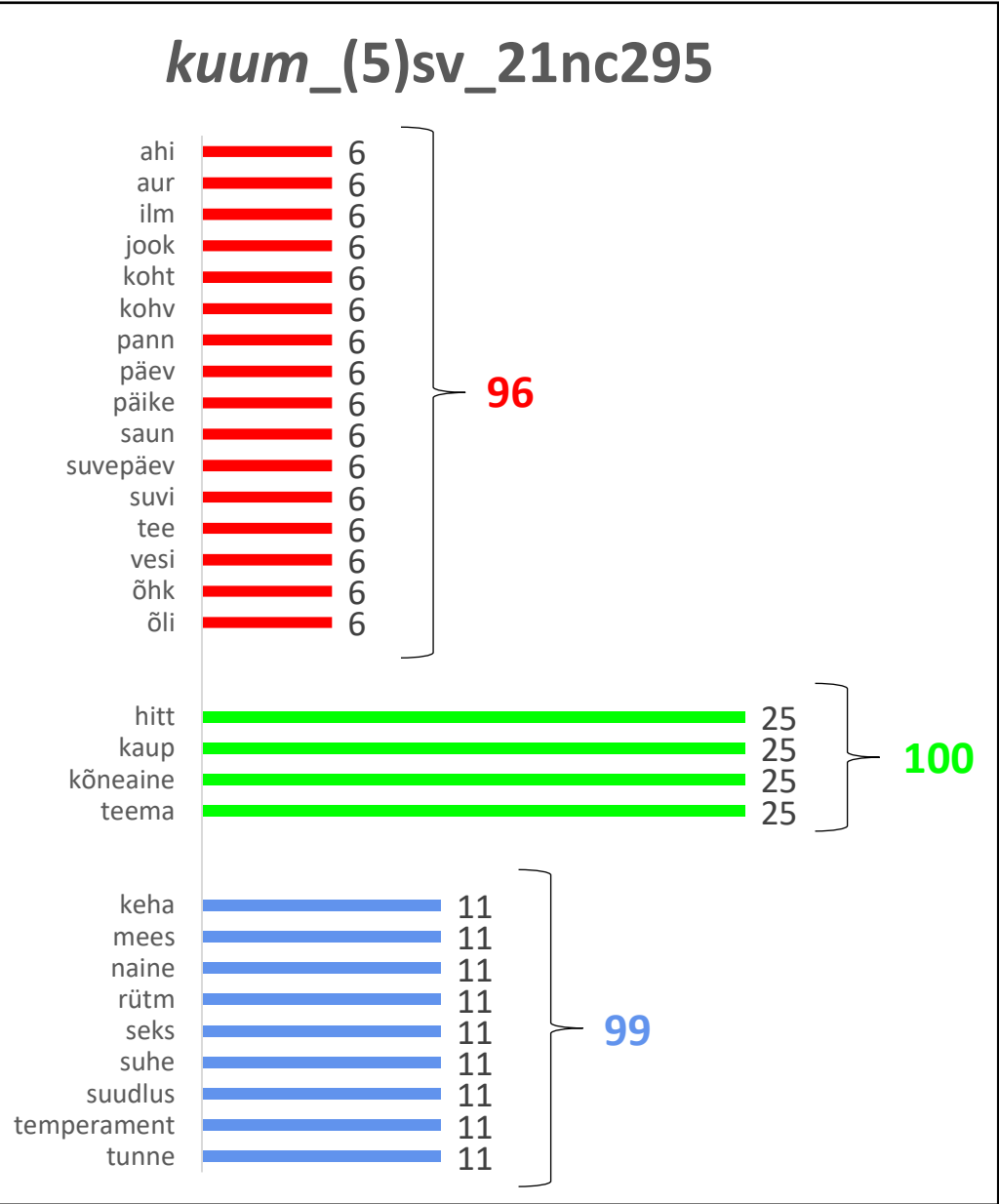


‘kõrge
temperatuuriga’
‘populaarne’
‘kirglik’

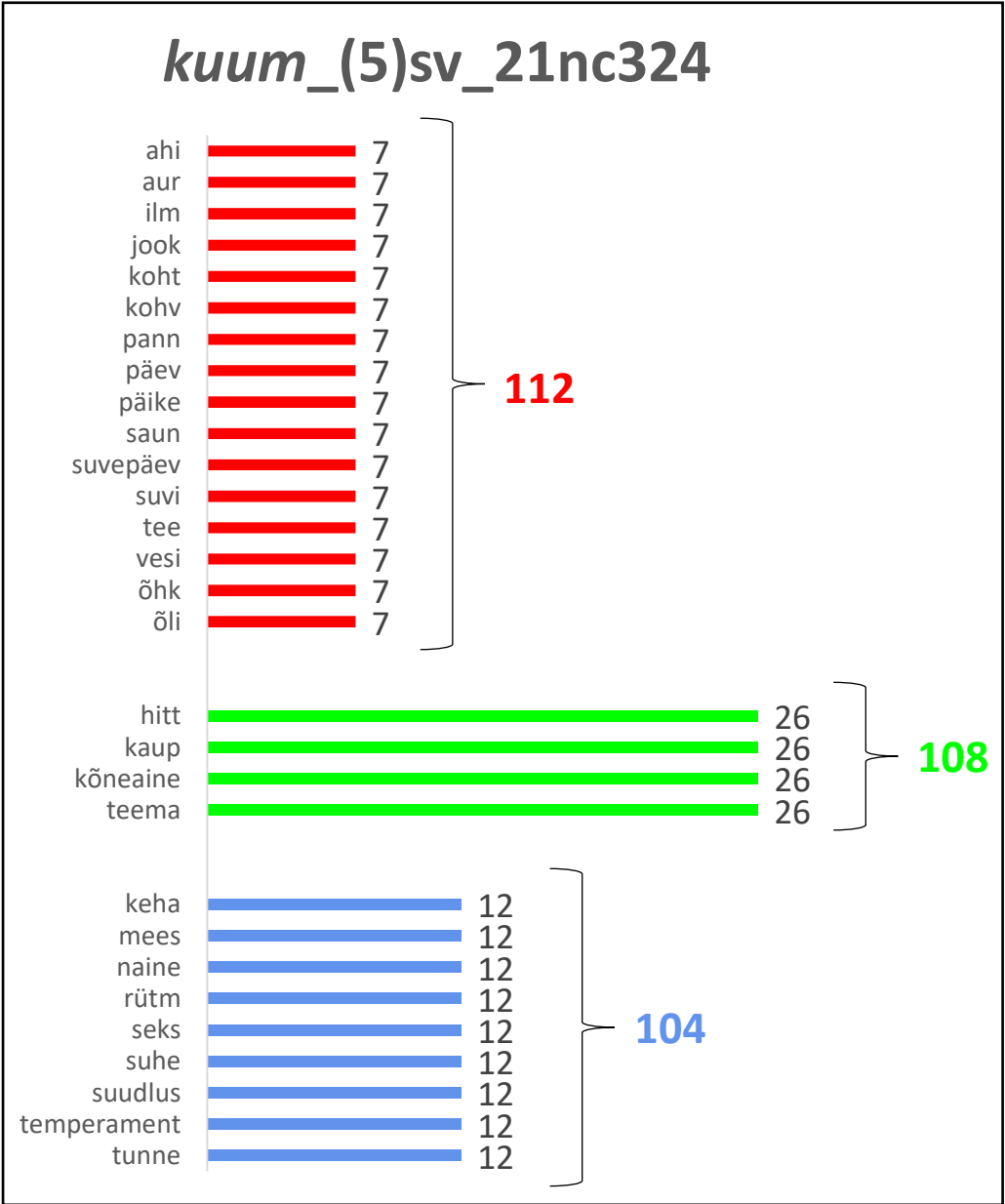
kuum_(4)sv_21nc303



Põhimõte (5): Adj + S, kollokaatide proportsioonid tähenduste sees tasakaalustatud

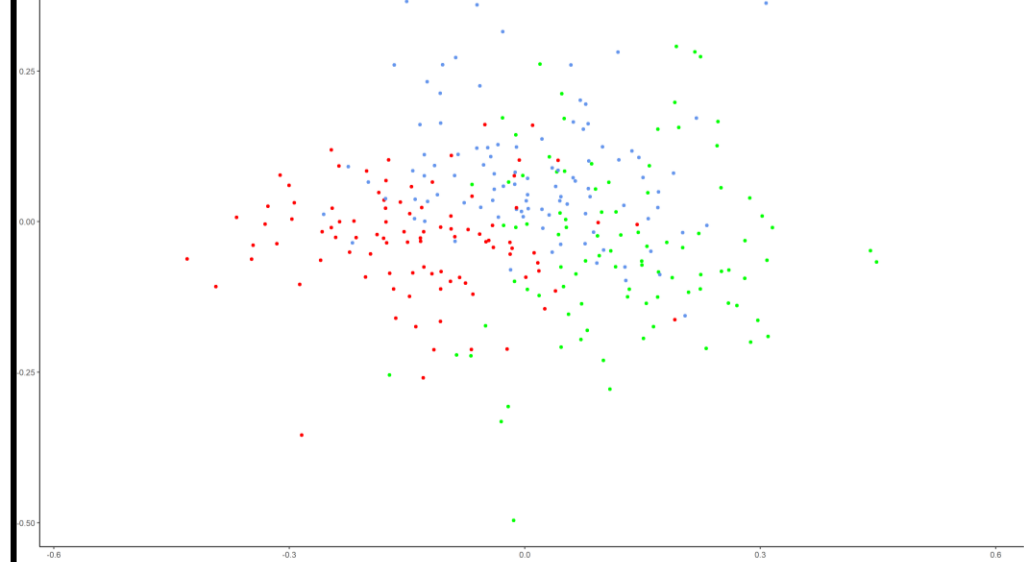


‘kõrge
temperatuuriga’
‘populaarne’
‘kirglik’

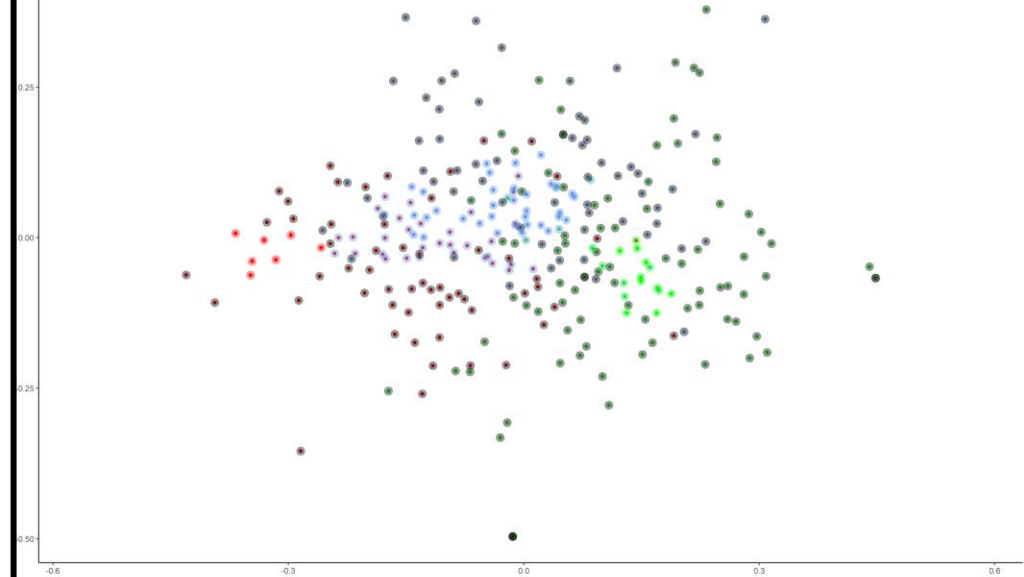


Põhimõte (5): Adj + S, kollokaatide proportsioonid tähenduste sees tasakaalustatud

valim1: *kuum_(5)21nc295*

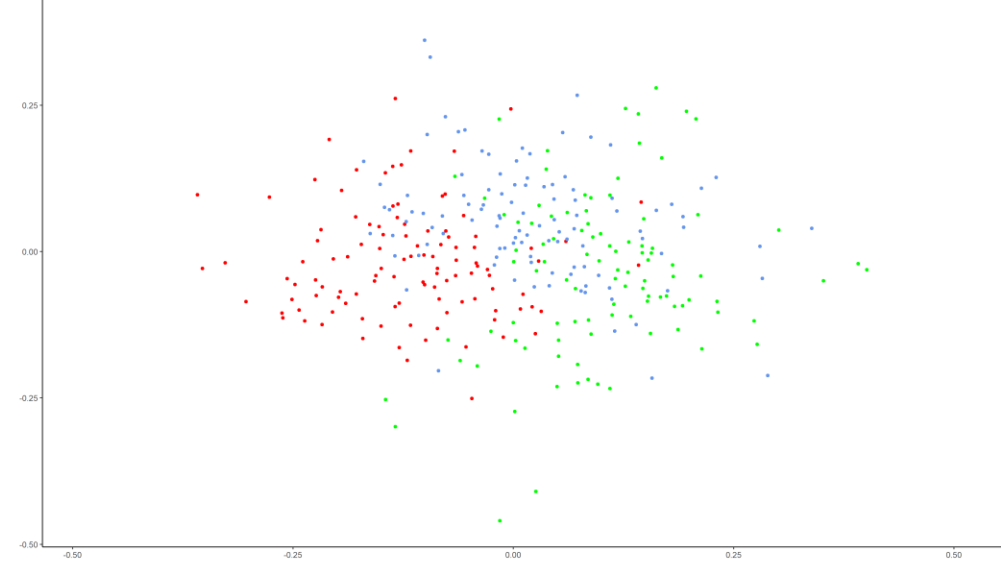


(automaatne klasterdus)

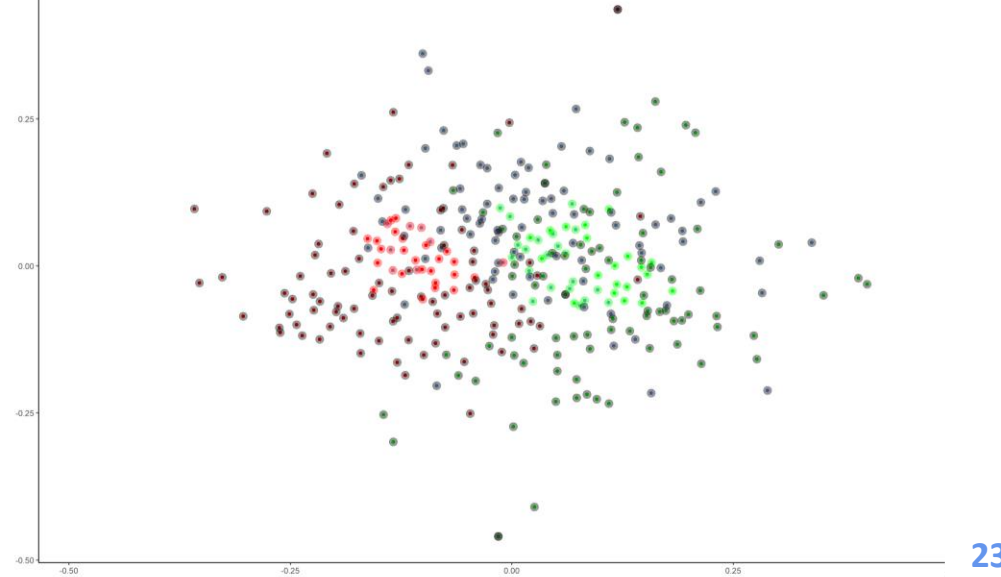


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

valim2: *kuum_(5)21nc324*



(automaatne klasterdus)



Omadussõna *kuum* tugevaimad nimisõnalised kollokaadid

(Eesti keele naabersõnade sõnastik)

kuum¹ omadussõna

1. kõrge temperatuuriga, väga soe

Nimisõnaga

kuum vesi
kuum ilm
kuum tee
kuum suvi

kuum jook
kuum suvepäev
kuum päike
kuum õhk
kuum päev
kuum kohv
kuum pann
kuum ahi
kuum õli
kuum aur
kuum saun
kuum koht

2. KÕNEK populaarne

Nimisõnaga

kuum teema
kuum kaup
kuum hitt
kuum kõneaine

3. kirelik, talitsematu; seksikas

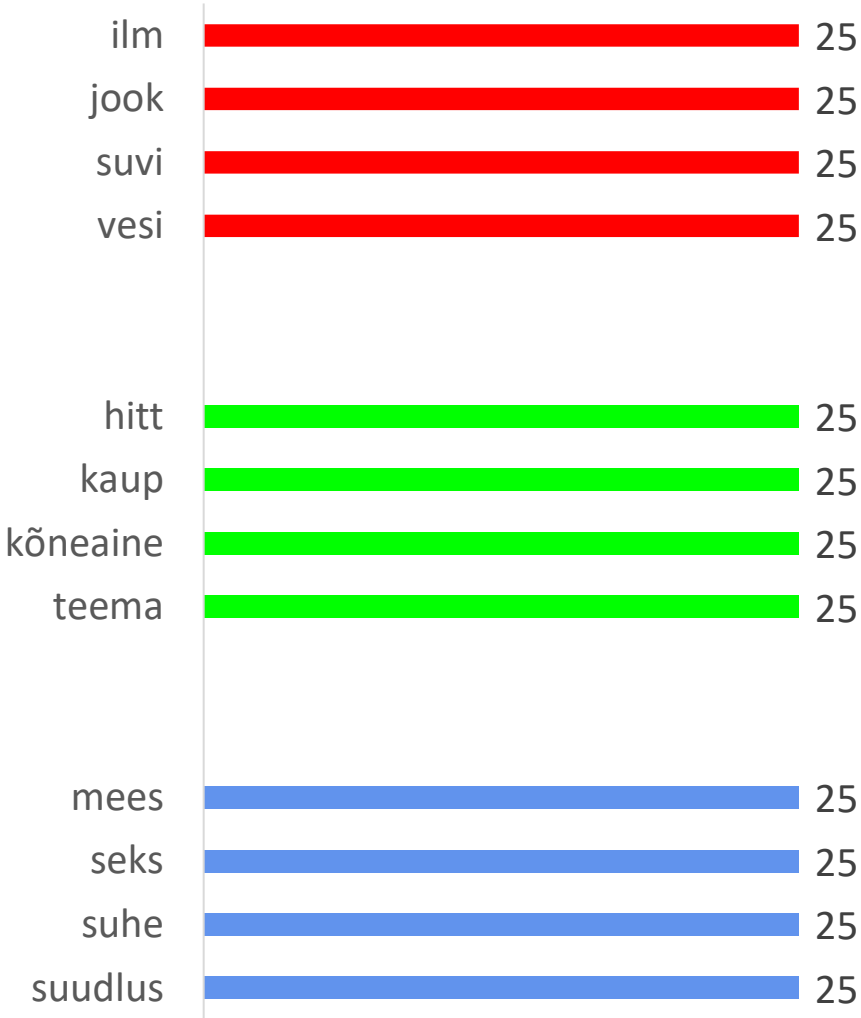
Nimisõnaga

kuum suhe
kuum mees
kuum suudlus
kuum seks

kuum keha
kuum naine
kuum temperament
kuumad tunded
kuumad rütmid

Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate

kuum_(6)sv_21nc300



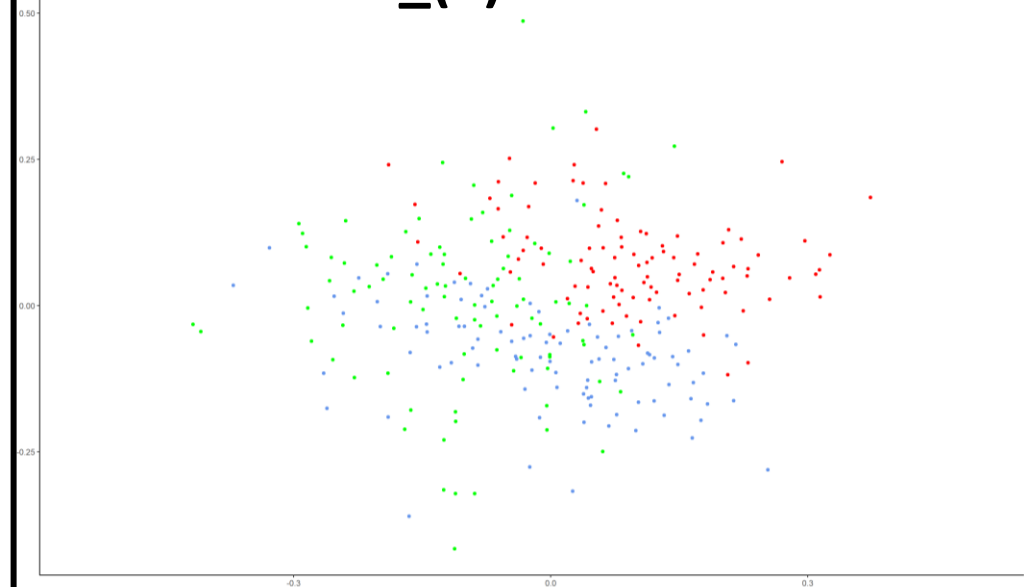
'kõrge
temperatuuriga'
'populaarne'
'kirglik'

kuum_(6)sv_21nc312

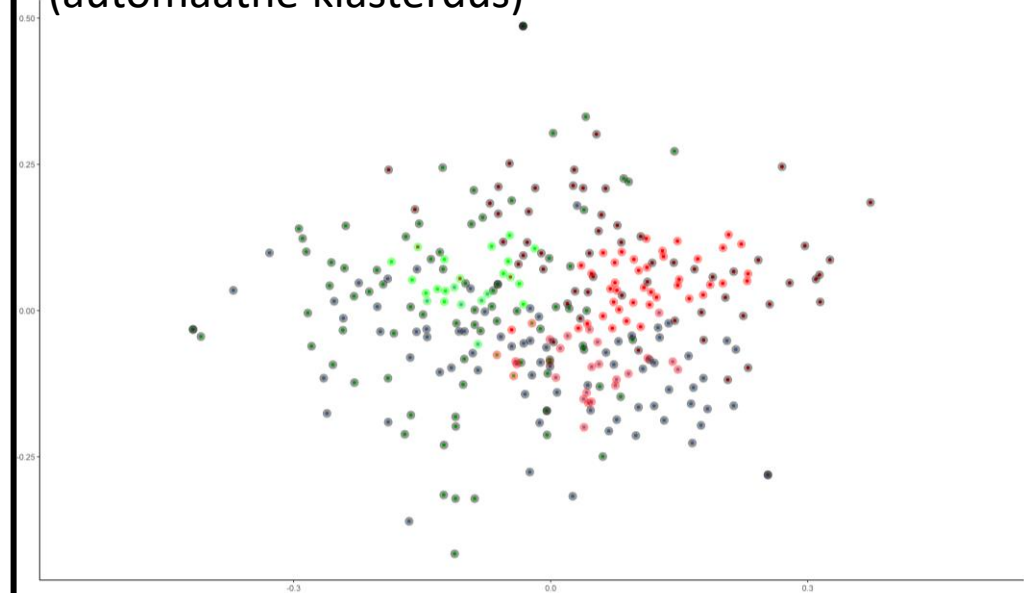


Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate

valim1: *kuum_(6)21nc300*

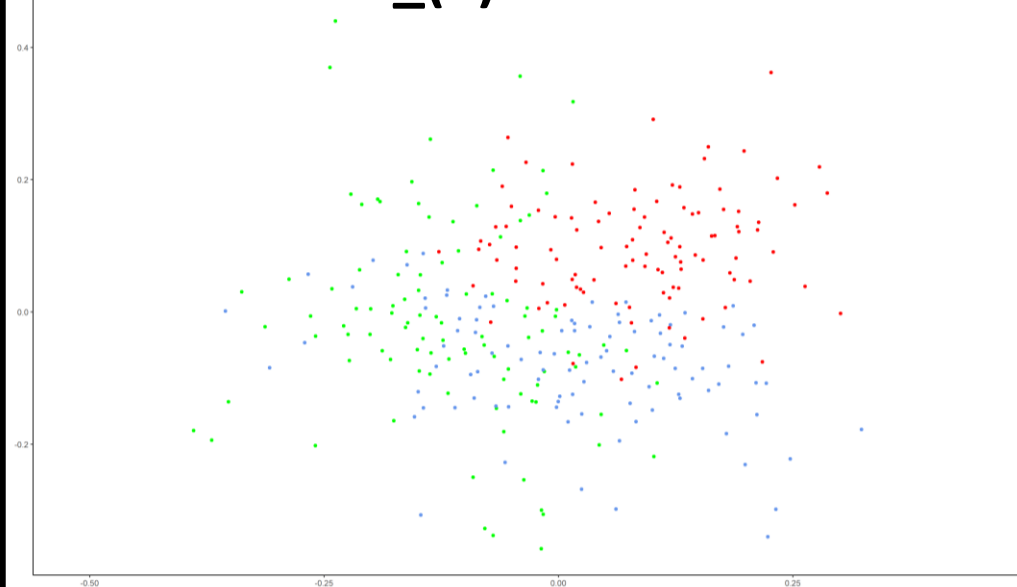


(automaatne klasterdus)

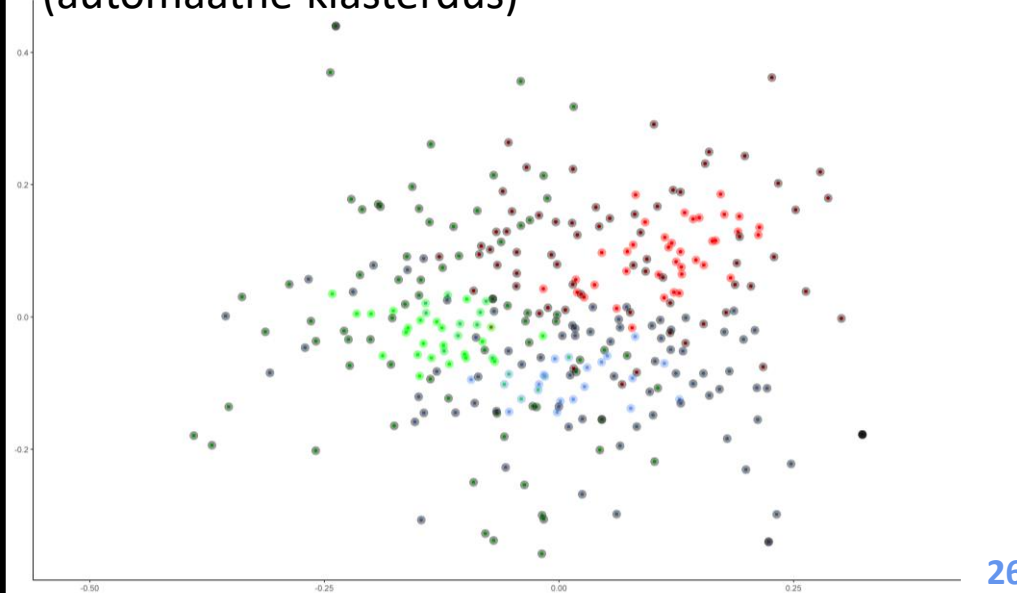


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

valim2: *kuum_(6)21nc312*

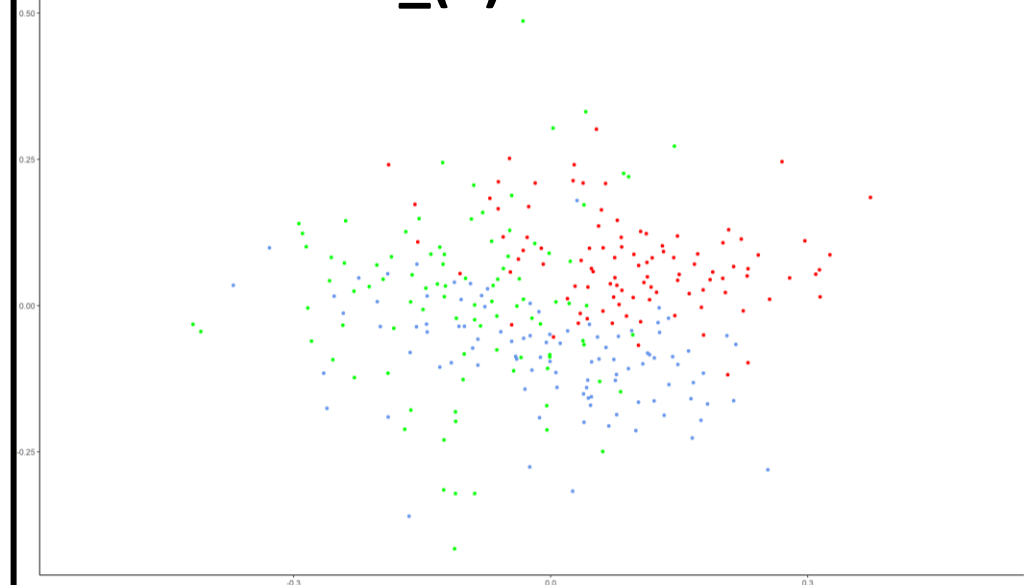


(automaatne klasterdus)

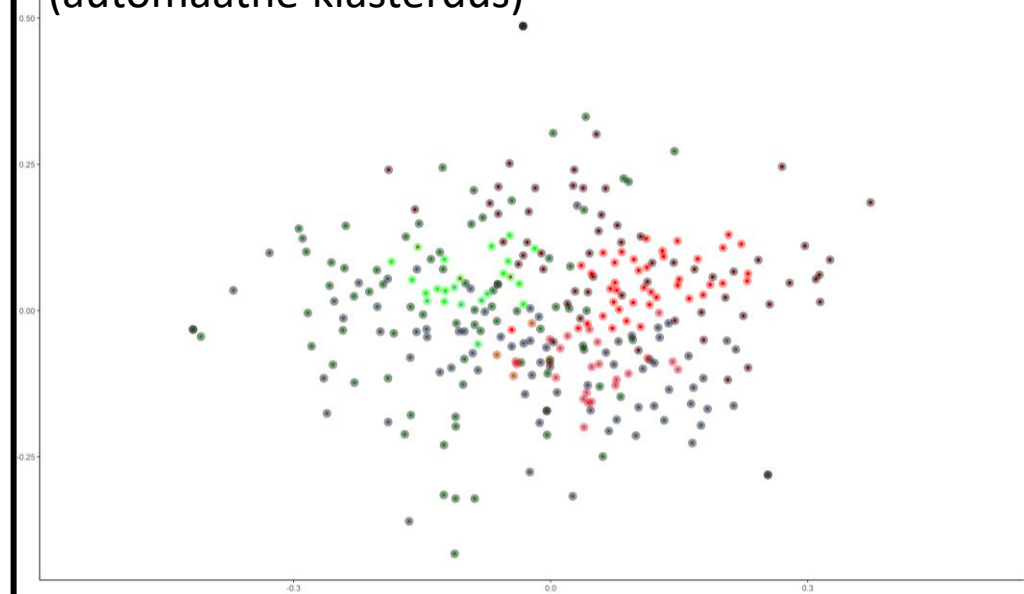


Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate

valim1: *kuum_(6)21nc300*

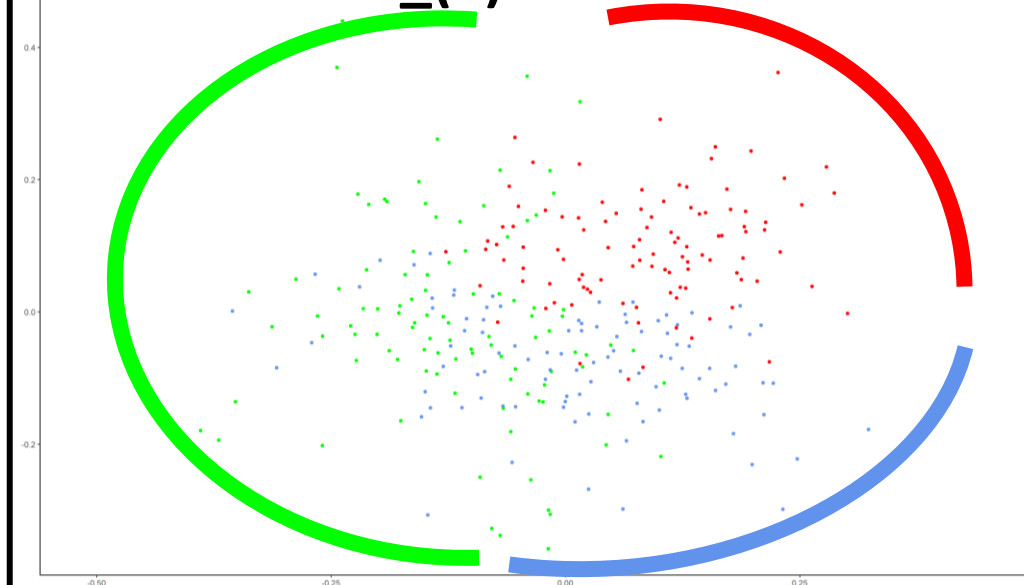


(automaatne klasterdus)

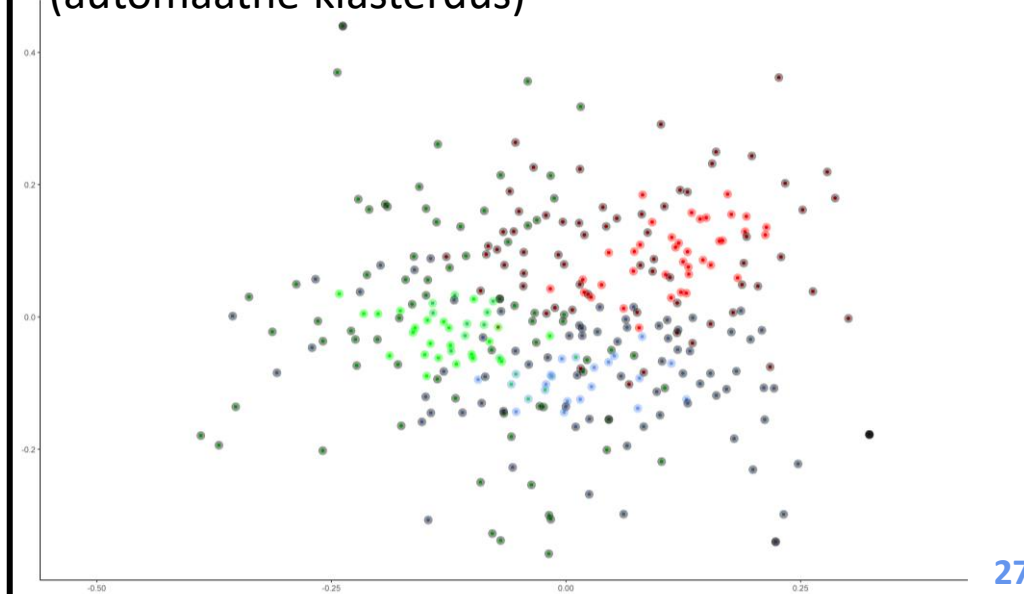


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

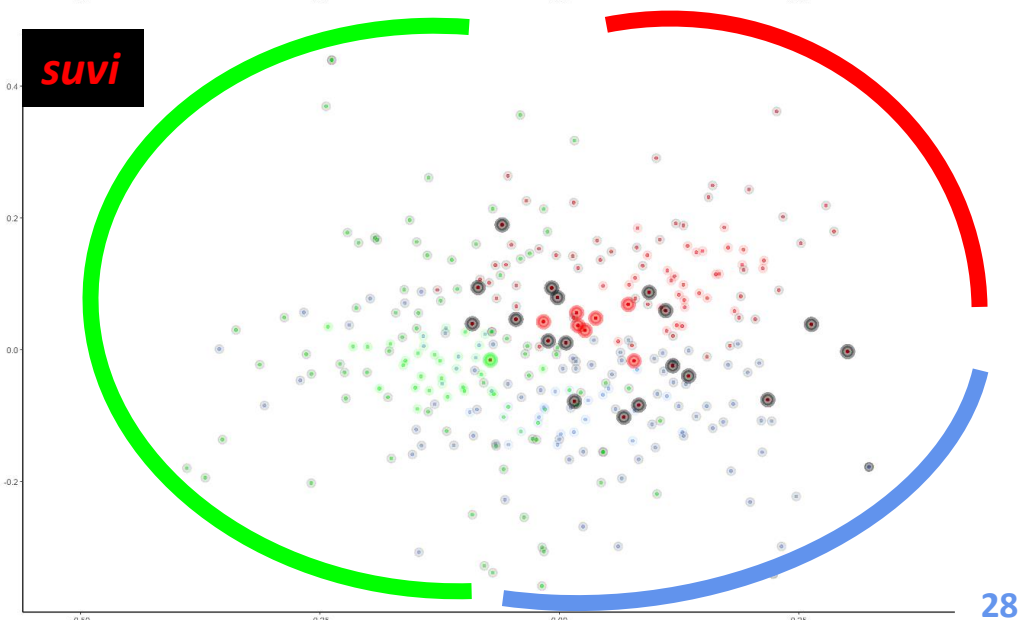
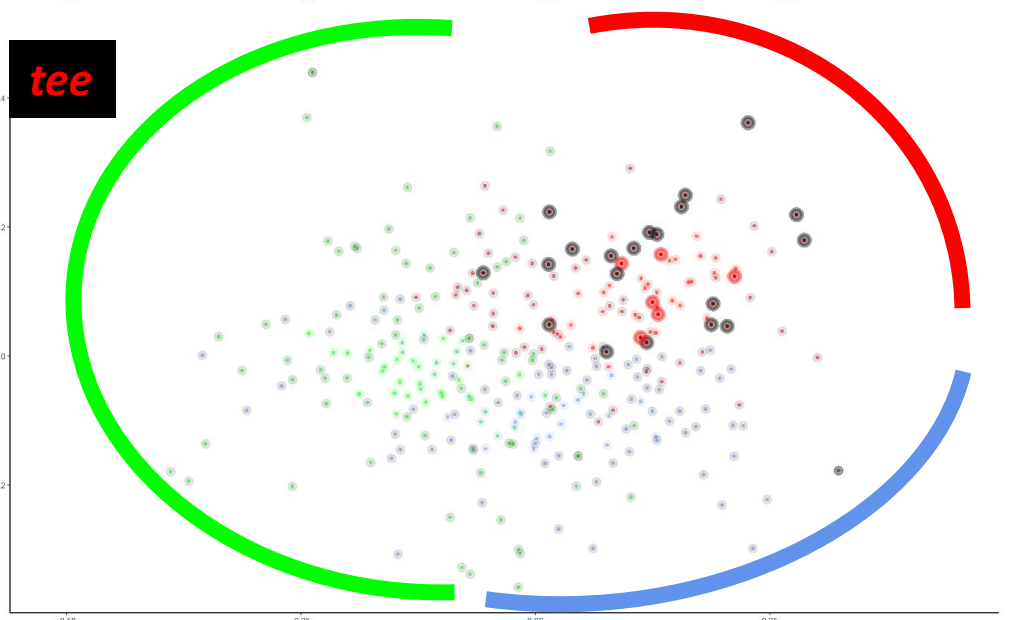
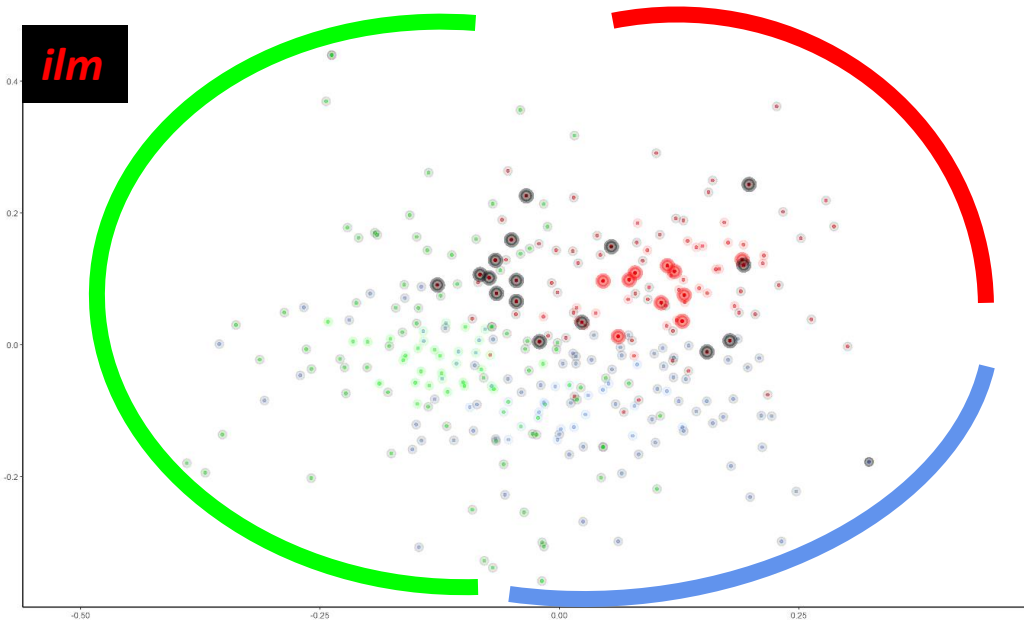
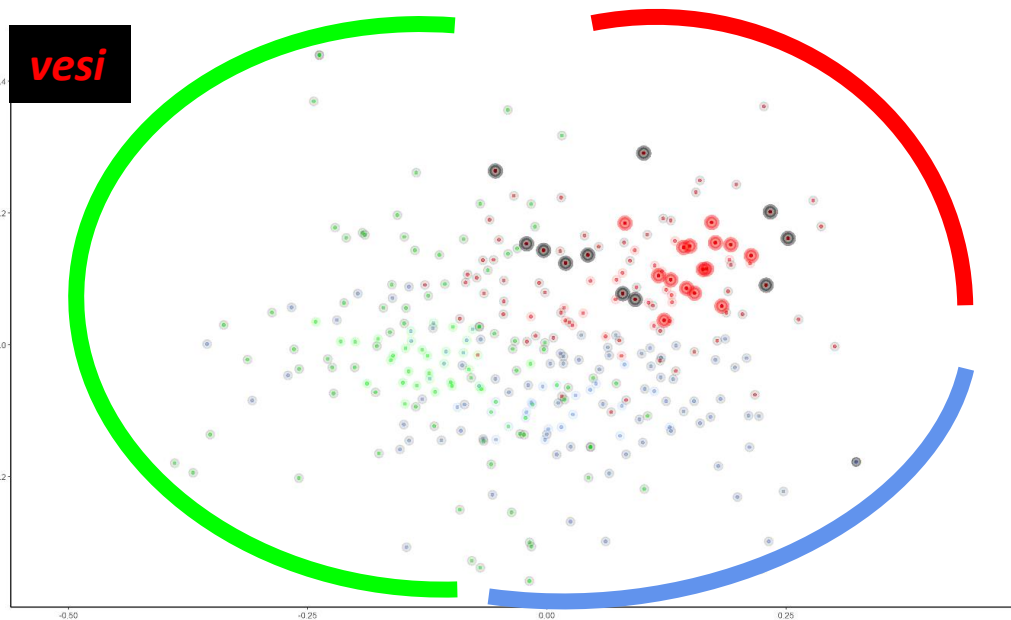
valim2: *kuum_(6)21nc312*



(automaatne klasterdus)

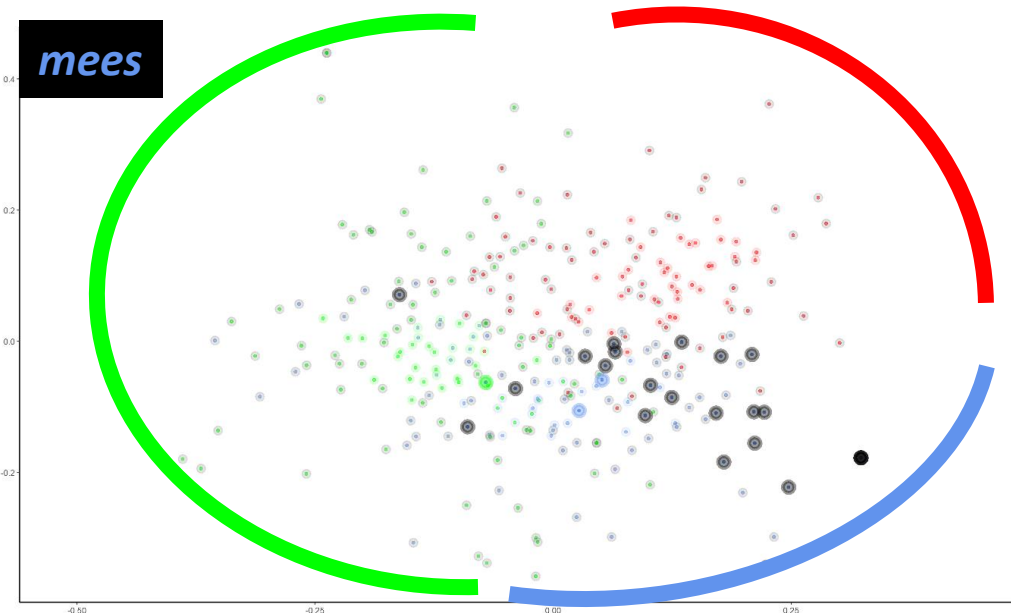
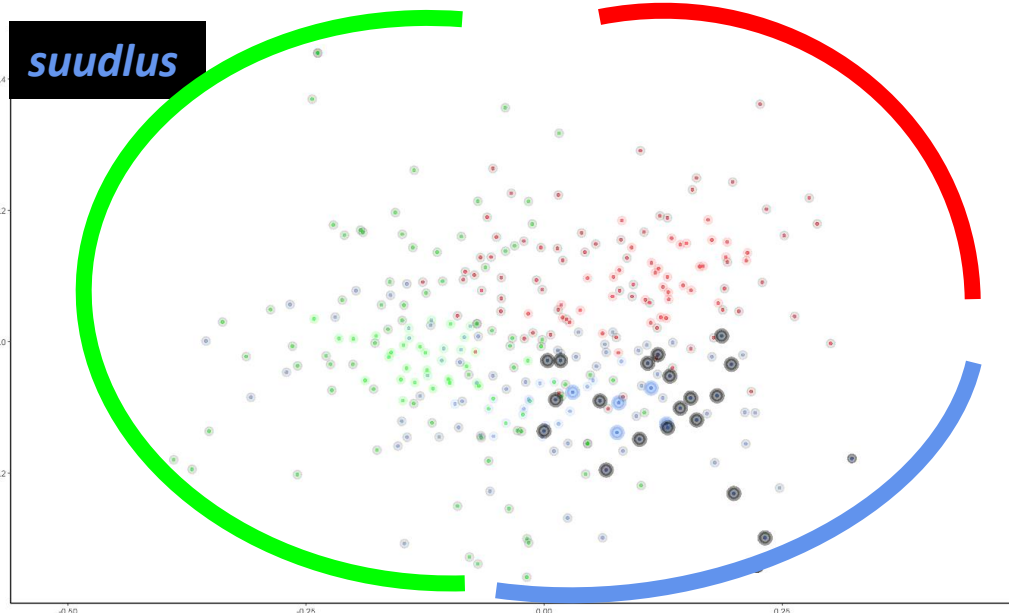


Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate

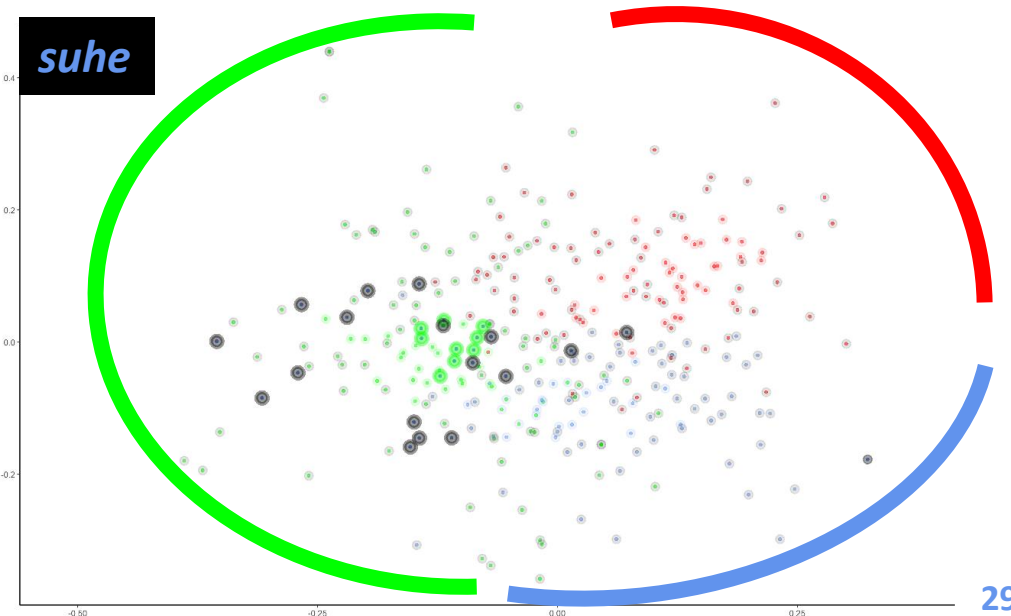
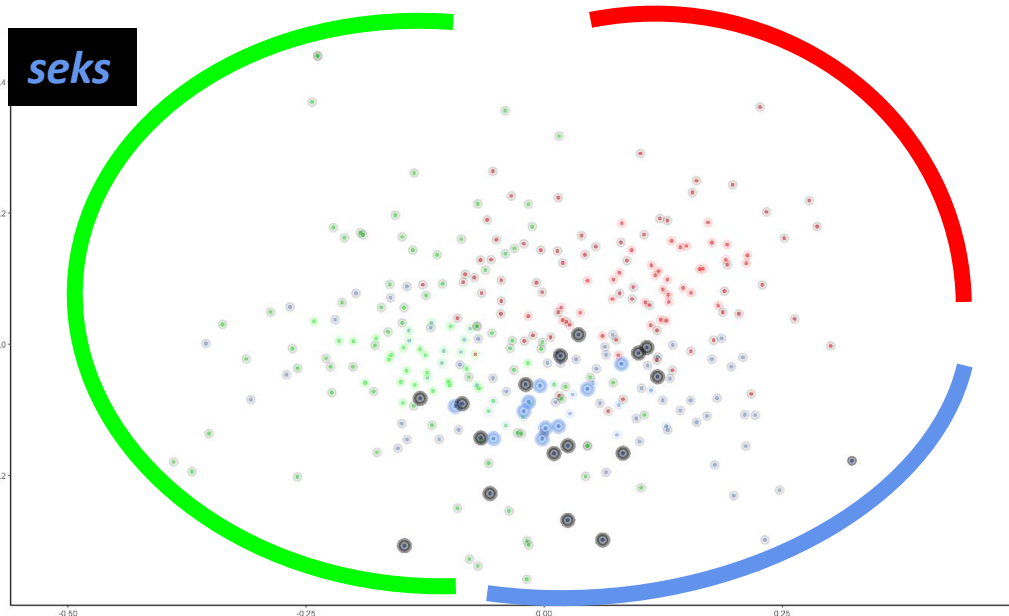


'kõrge
temperatuuriga'
'populaarne'
'kirglik'

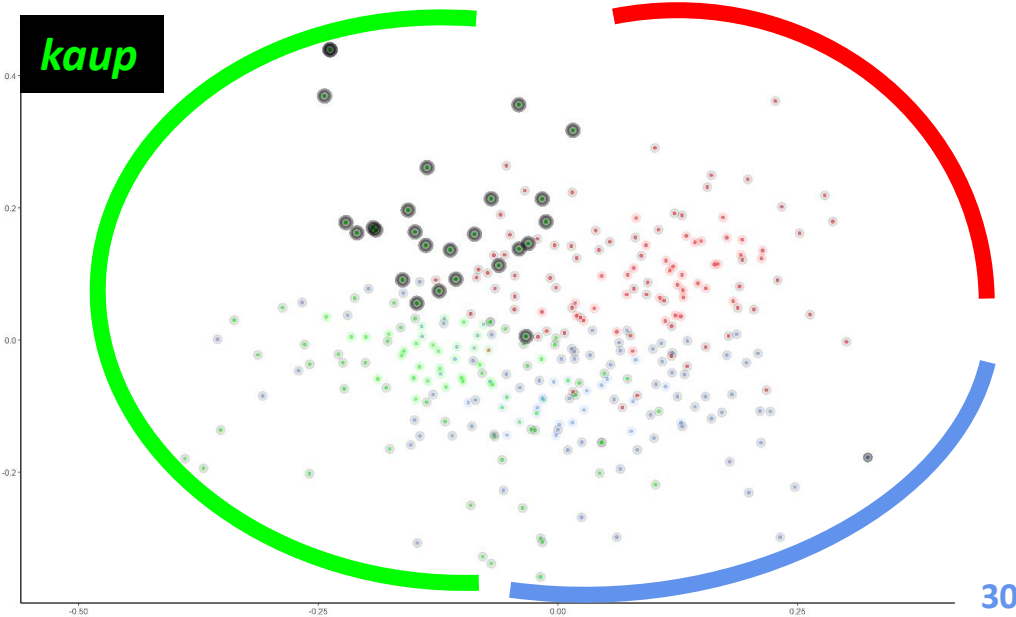
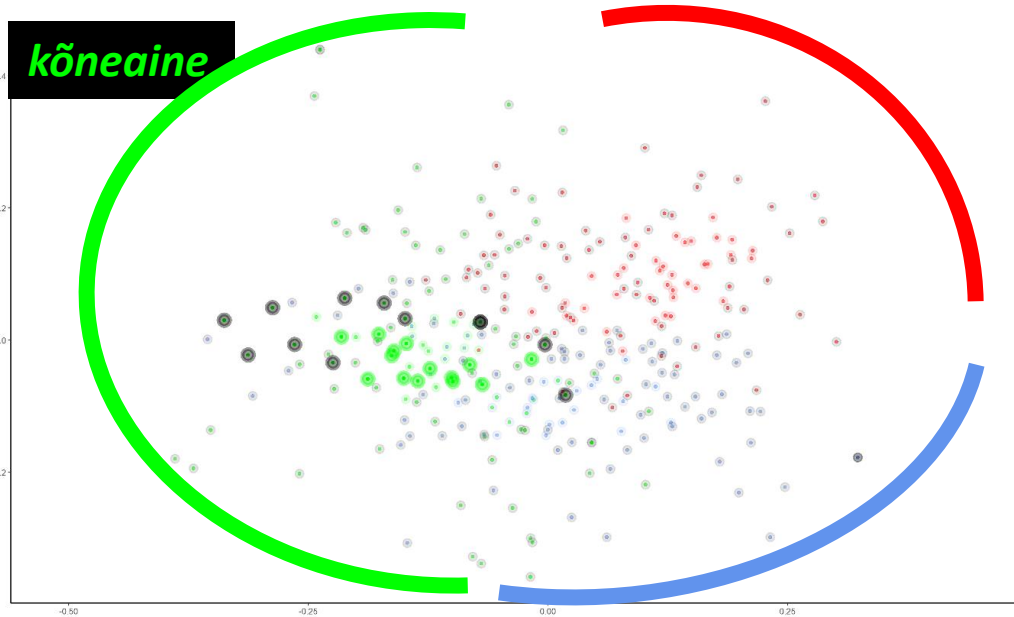
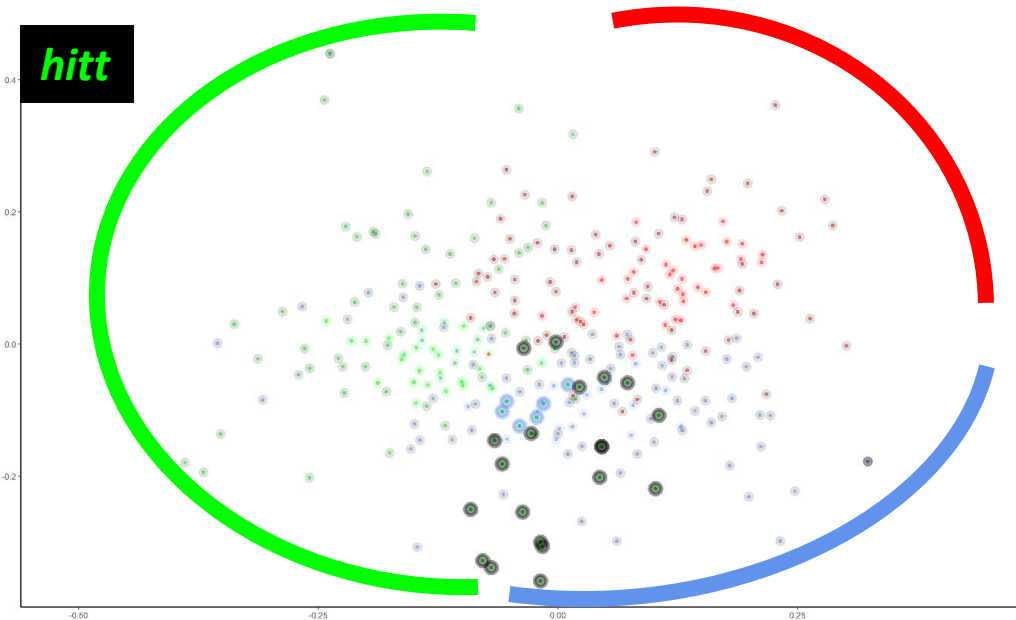
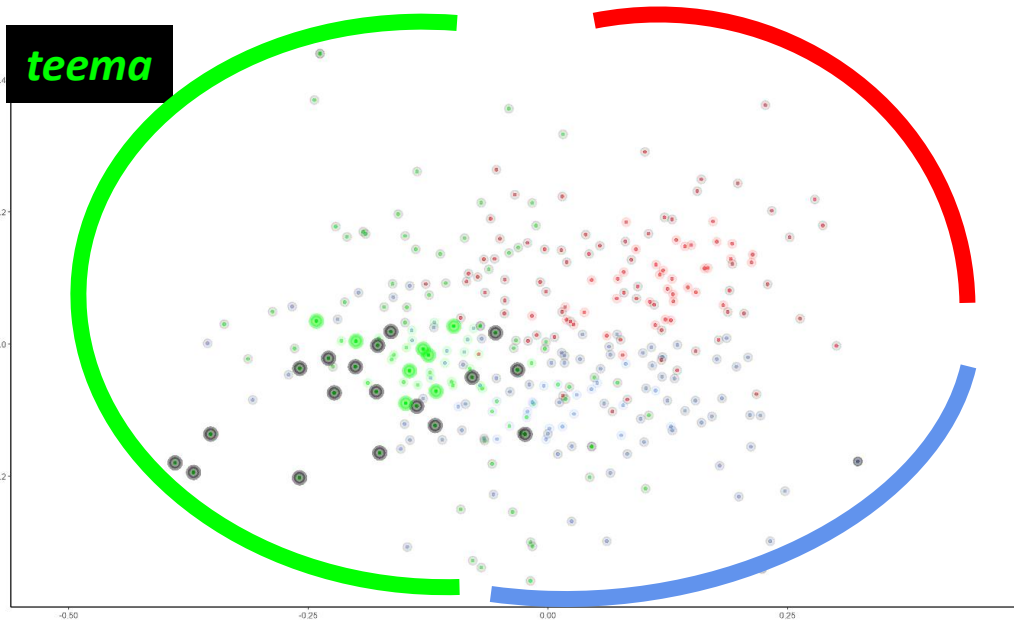
Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate



'kõrge
temperatuuriga'
'populaarne'
'kirglik'



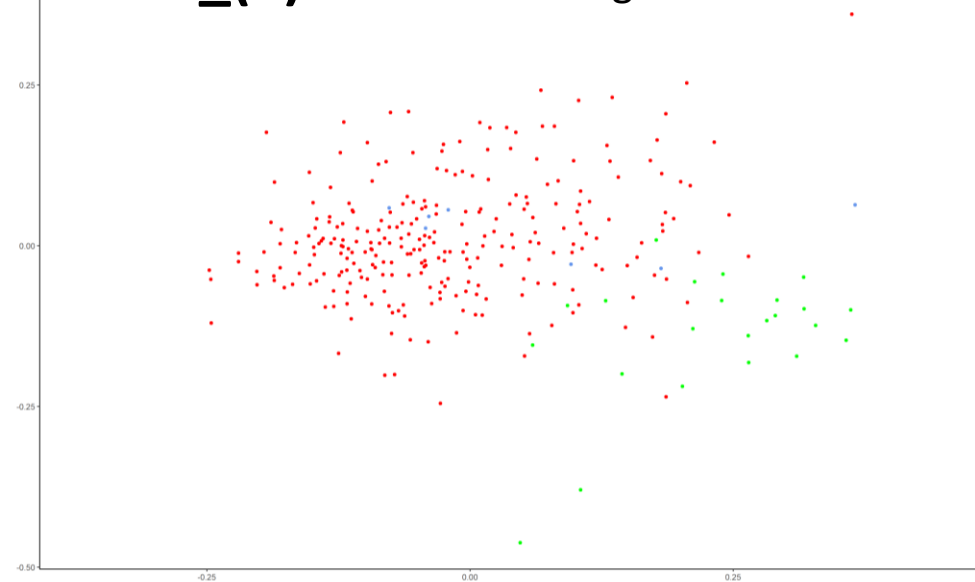
Põhimõte (6): Adj + S, iga tähenduse kohta võrdne arv nimisõnalisi kollokaate



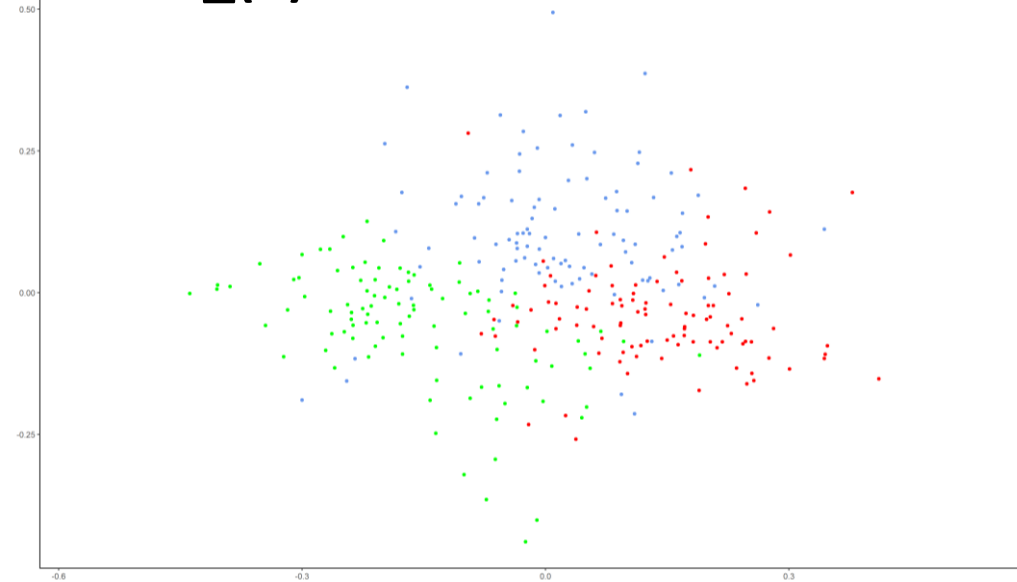
'kõrge
temperatuuriga'
'populaarne'
'kirglik'

Põhimõtete (3) kuni (6) alusel moodustatud valimite analüüsitulemuste võrdlus

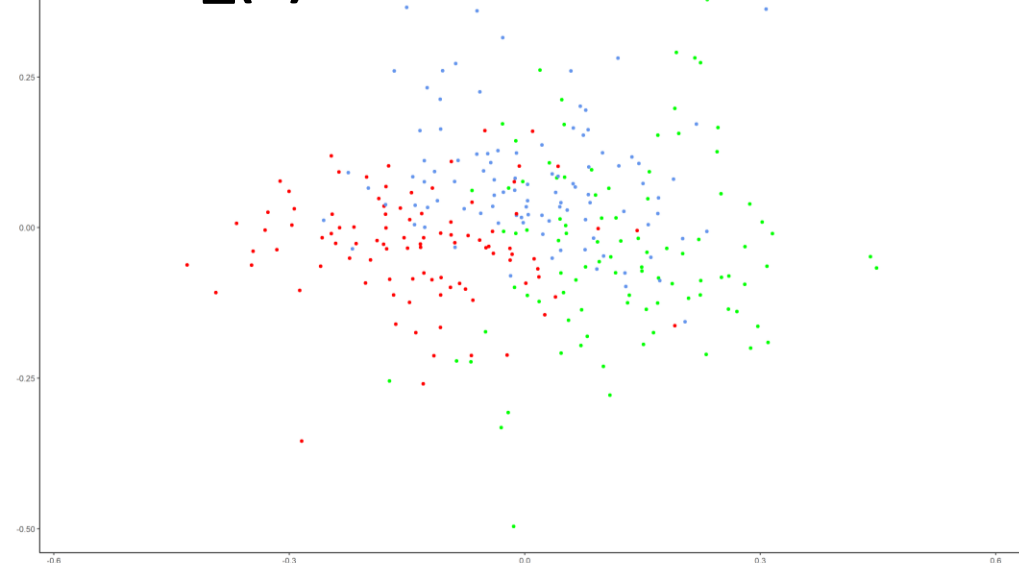
kuum_(3)21nc300 – tugevaimad kollokaadid



kuum_(4)21nc300 – tähendused tasakaalus

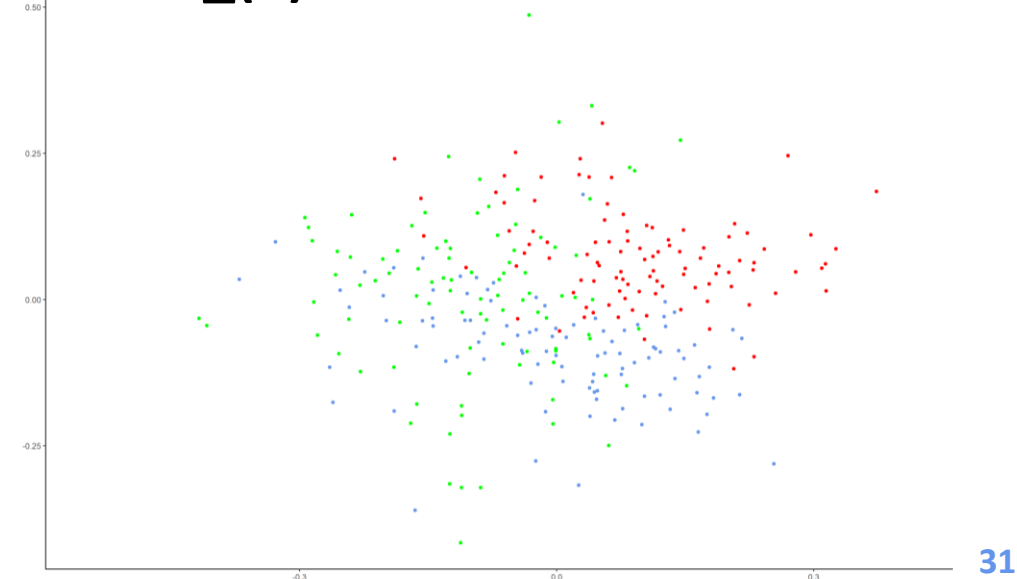


kuum_(5)21nc295 – kollokaadid tasakaalus



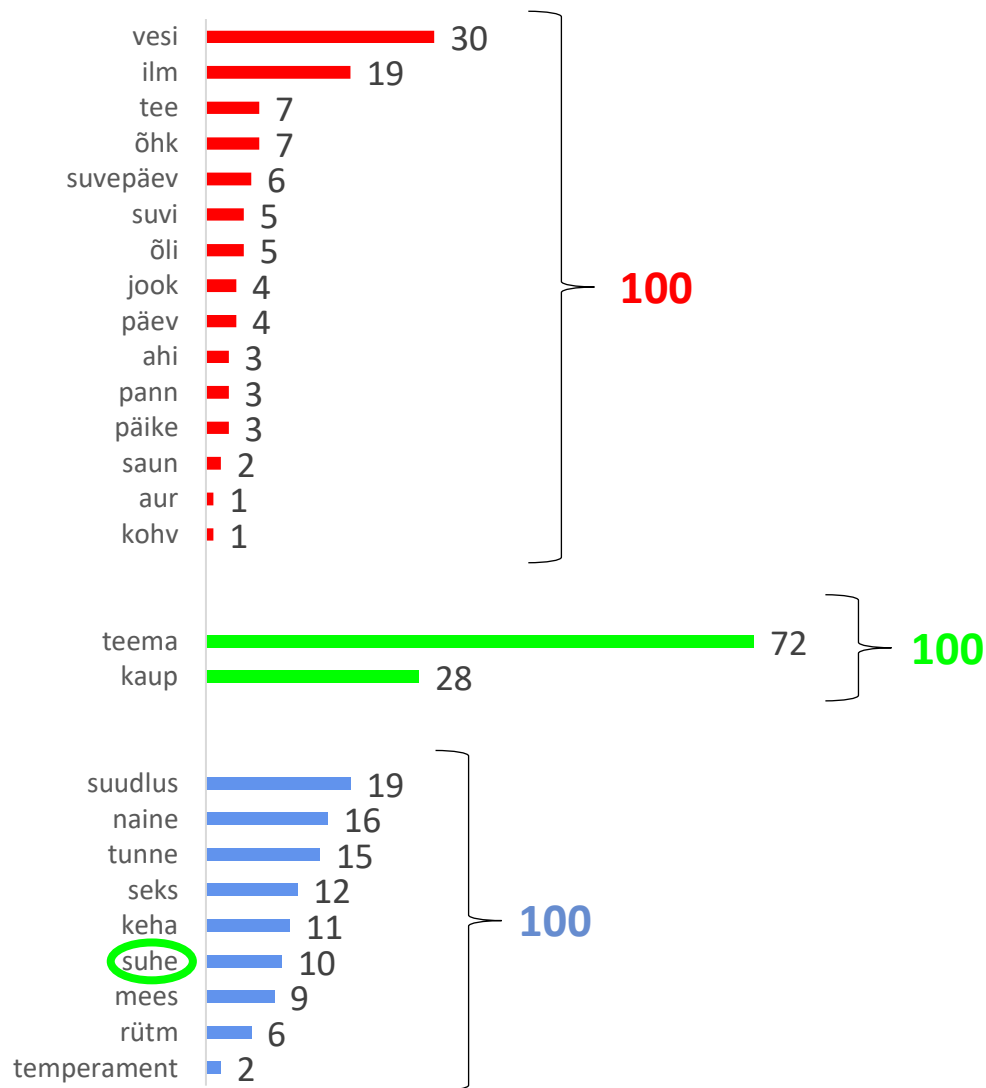
'kõrge
temperatuuriga'
'populaarne'
'kirglik'

kuum_(6)21nc300 – kollokaate võrdselt



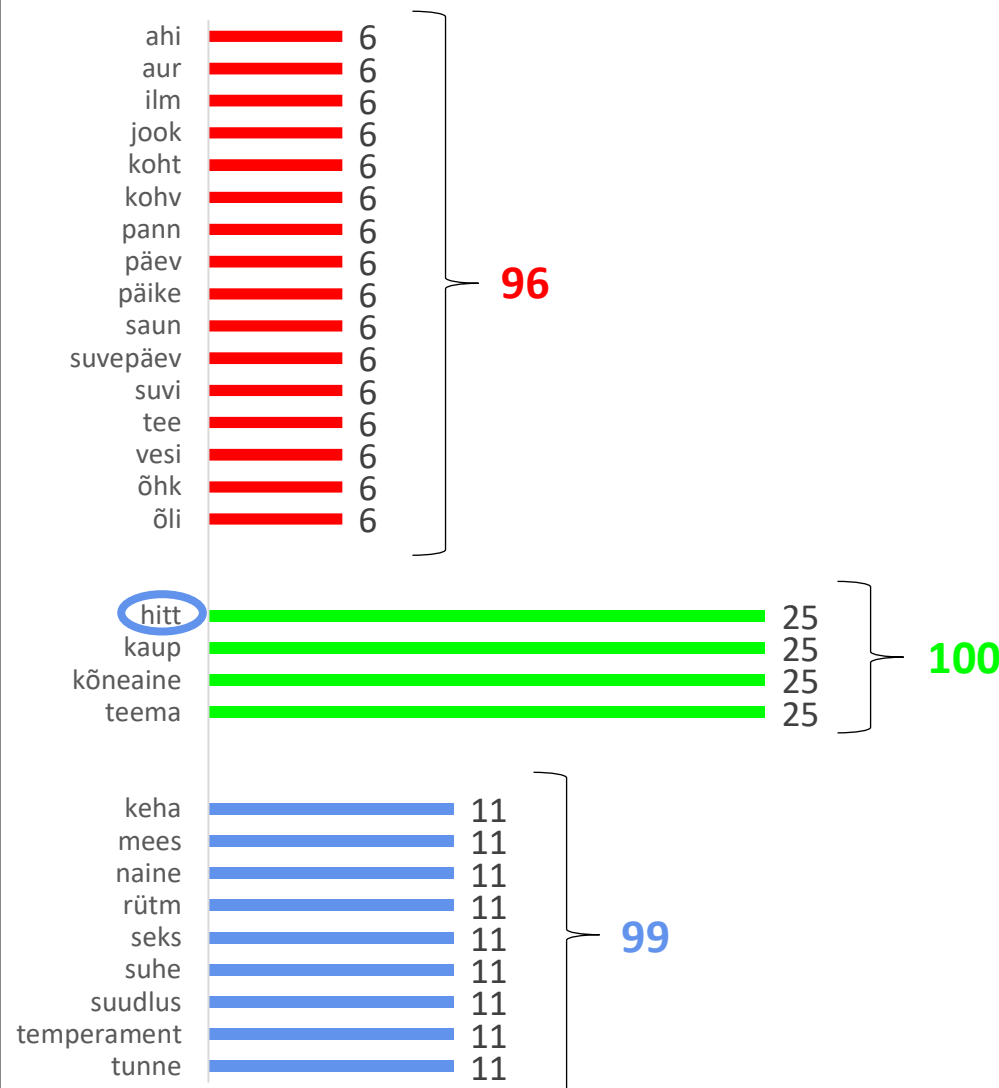
Põhimõtete (4) ja (5) alusel moodustatud valimite kollokaatide jaotuste võrdlus

kuum_(4)sv_21nc300



'kõrge
temperatuuriga'
'populaarne'
'kirglik'

kuum_(5)sv_21nc295



Kokkuvõte

- Distributiivsemantilised mudelid suudavad praegu eristada üsna hästi polüseemsete sõnade **konkreetse ja abstraktse** tähendusega kasutusjuhtusid.
- Enam kui kahe tähenduse eristamisel on üheks oluliseks probleemiks see, et täiesti juhuslikus valimis on eri tähendusi esindavate kasutusjuhtude **proportsioonid väga erinevad**.
- Valimite varieerimine näitas, et eri tähendusi esindavate kasutusjuhtude **proportsioonide tasakaalustamine valimis** parandab oluliselt mudelite võimet eristada kontekstuaalse info põhjal tähendustele vastavaid kasutusjuhtude rühmi.
- Mõne kollokaadi puhul ei lange introspektsioonipõhiselt määratud tähendus kokku kontekstipõhise analüüsi tulemustega (nt *kuum **suhe*** ja *kuum **hitt***, aga ka *kuum **kaup***).
- Järgmise sammuna oleks vaja tegeleda eri tähenduste sees **kollokaatide proportsioonide tasakaalustamise** või tasakaalustamata jätmise küsimusega.
- Kaugem siht – saada mudelid tööle **ilma teoreetilise tähendusliigenduse etteandmiseta**.

Harris, Zellig 1954. Distributional structure. – Word, 10(2–3), 146–162.

Heylen, K.; Wielfaert, T.; Speelman, D.; Geeraerts, D. 2015. Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. – Lingua, 157, 153–172.

Hilpert, M.; Saavedra, D. C. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. – Corpus Linguistics and Linguistic Theory, 1–32.

Montes, M. 2021. Cloudspotting – Visual analytics for distributional semantics. Dissertation. Leuven.