

Adjektiiv või mitte?

Prototüüpse adjektiivi korpusprofiil

Geda Paulsen, Maria Tuulik, Ene Vainik, Ahti Lohk
Eesti Keele Instituut | EKI

” Ettekande teemad:

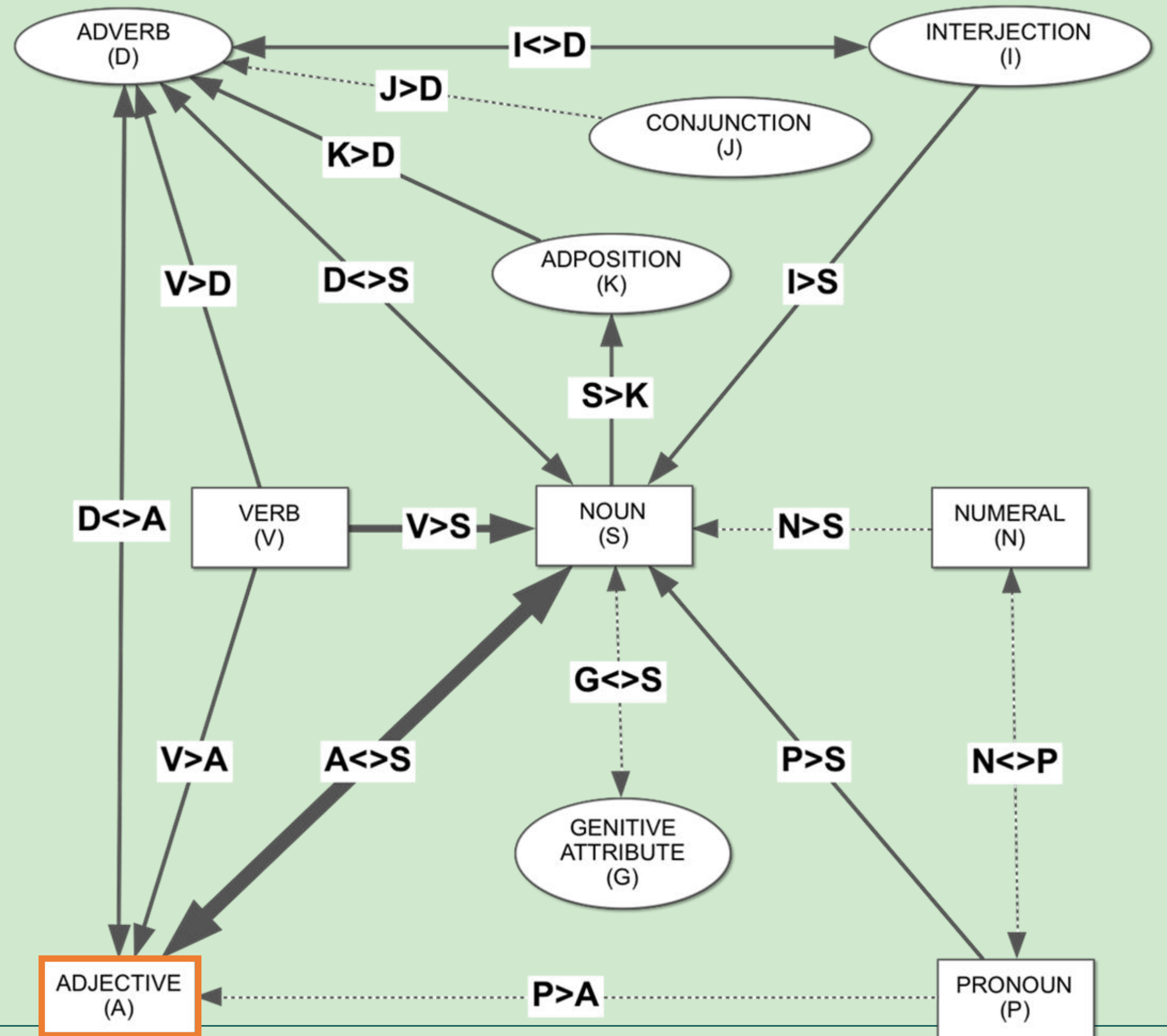
- Sõnaliigiline mitmesus sõnaraamatutöös
- Kuidas rakendada kvantitatiivseid meetodeid leksikograafi töös?
- Prototüüpse adjektiivi korpusprofiil
- Sarnasusmeetodid (kõrvalekaldeanalüüs, eukleidiline kaugus, koosinussarnasus)



Adjektiiv sõnaliigivõrgustikus

Paulsen, Vainik, Tuulik 2020: *Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias*

Vainik, Paulsen, Lohk 2020: *A typology of lexical ambiguities in Estonian*





Mis eristab adjektiivi teistest sõnaliikidest?

- Kuidas võiksid korpusandmed aidata hinnata partitsiibi, nimisõna, või adverbi nihkumist adjektiiviklassi?
- Milline on prototüüpne adjektiiv korpuses?
- Lähtekoht – morfosüntaktilised korpusmustrid:
 - atributiivne funktsioon
 - ühildumine
 - predikatiivne funktsioon
 - võrrete moodustamine
- Juhtum: **partitsiibid**



Test: parameetrid

1. atribuudiparameeter [**test_S**]
2. ühildumisparameeter [**test_S_ÜHILD**]
3. lausealguseparameeter [LA_**test_S**]
4. neljakohaline parameeter [V_?_**test_S**]
5. adverbiparameeter [**D_test**]
6. predikatiiviparameeter [*olema_test*; *olema_**D_test**]*
7. võrdeparameeter*

*võrdevormi esinemissagedus otsitakse etteantud keskvõrde vormi järgi



Testrühmad

1. **PSV protoadjektiivid (N = 100)**
avalik, inetu, ehtne, ohutu, sile, ükskõikne
2. **Ekilexi andmebaasi partitsiibid (Ekilex, N = 100)**
tuimestav, jälitatav, saastunud, planeeritud
3. **PSV verbidest moodustatud partitsiibid (verbilised, N = 100)**
loobuv, jätkatav, unustanud, armutud

Ekilex = Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteem (ekilex.ee)

PSV = Eesti keele põhisõnavara sõnastik

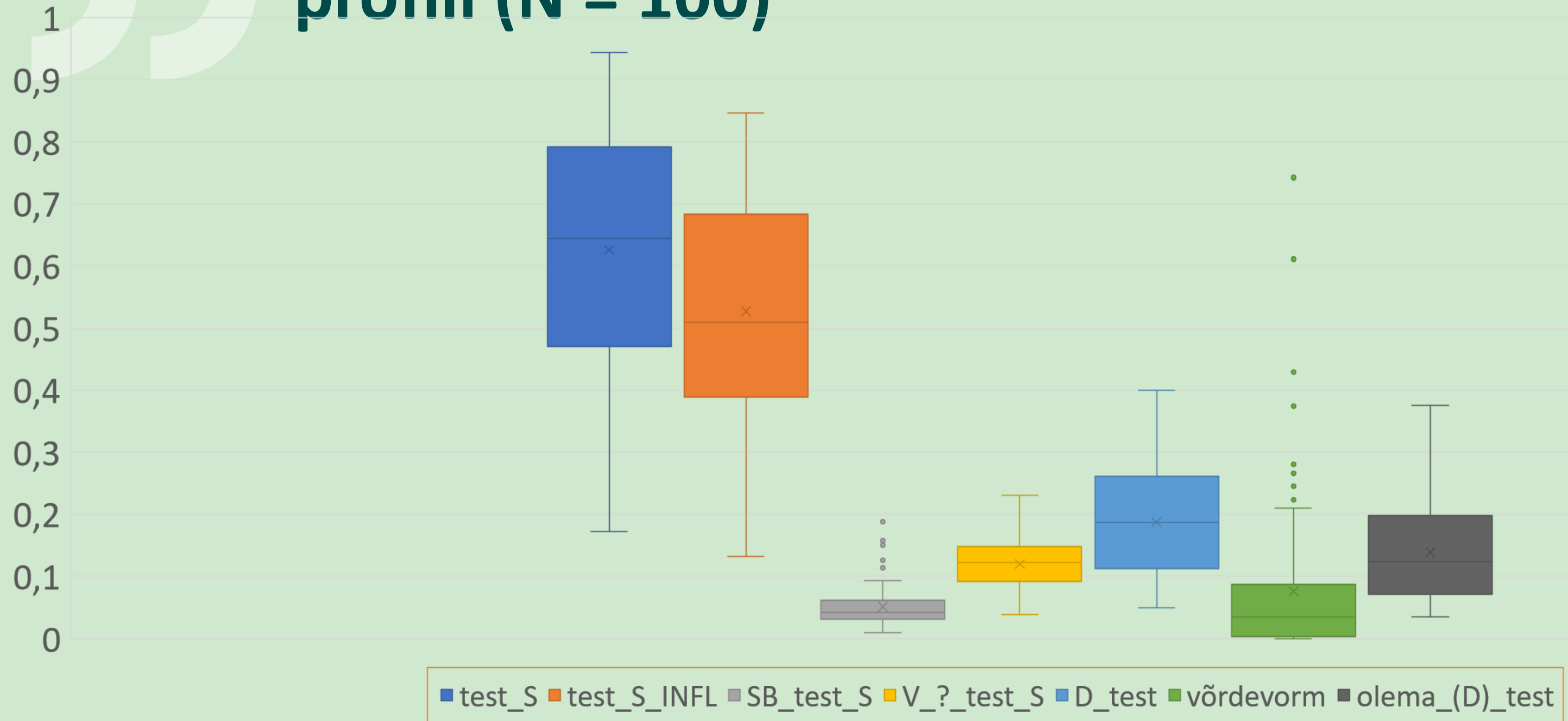


Ekstraheerimisprotsess

- Eesti keele ühendkorpus 2019
- EstNLTK
- Parameetrile vastamise **suhteline sagedus** (testsõna testmusteris esinemiste sagedus / lemma üldsagedus)
- Kood: [https://github.com/ahtilohk/PSG227/blob/main/Test-patterns occurrences in ENC2019 without estnltk corpus processing module.py](https://github.com/ahtilohk/PSG227/blob/main/Test-patterns%20occurrences%20in%20ENC2019%20without%20estnltk%20corpus%20processing%20module.py)

Tuulik, Vainik, Paulsen, Lohk 2022: Kuidas ära tunda adjektiivi? Korpuskäitumise mustrite analüüs.
– *ERÜ aastaraamat*.

Prototüüpse adjektiivi korpuskäitumise profiil (N = 100)





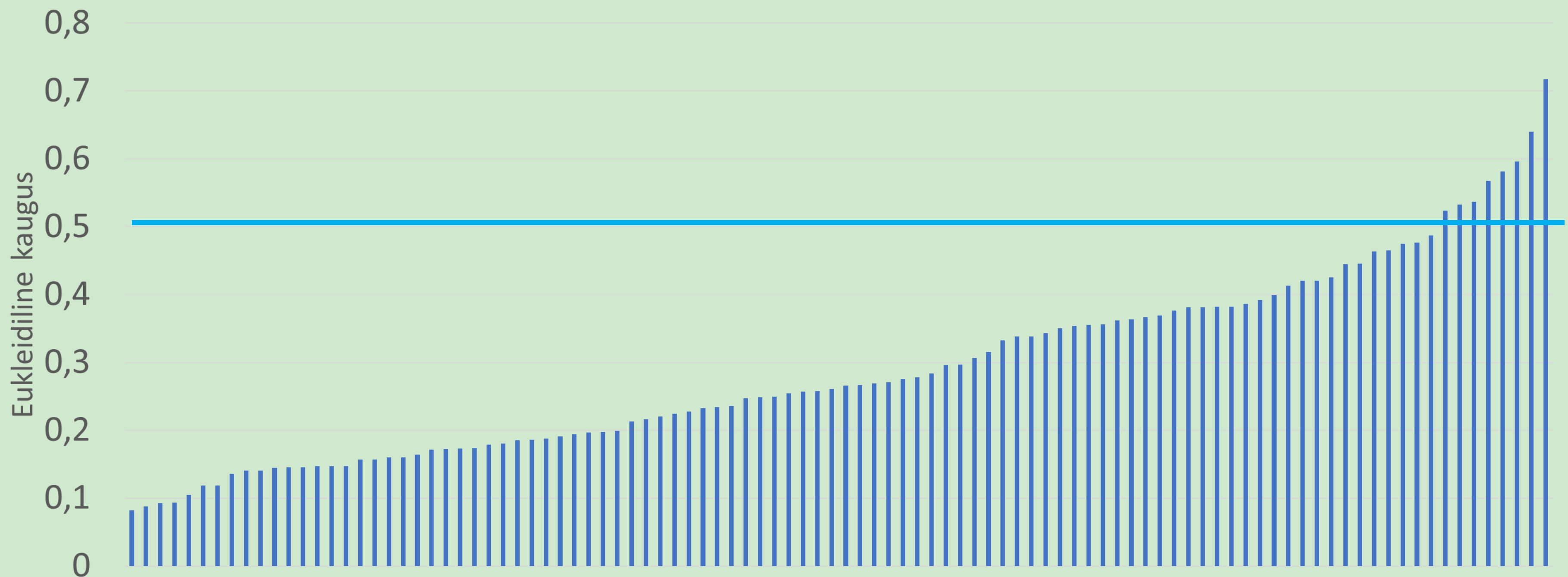
Sarnasusmõõdikud

Eukleidiline kaugus – ruutjuur uuritavate parameetrite ruutude summast. Mida väiksem number, seda suurem on lähedus ja seda sarnasemad on profiilid.

Koosinussarnasus – profiile saab käsitleda 7-dimensiooniliste vektoritena ja leida nurgafunktsiooni abil nende sarnasuse prototüüpset adjektiivi esindava vektoriga. Mida suurem näit, seda suurem sarnasus. Võib läheneda väärtusele 1.



Prototüüpse adjektiivi (N = 100) eukleidilised kaugused ja võimalik lävend (= 0,5)

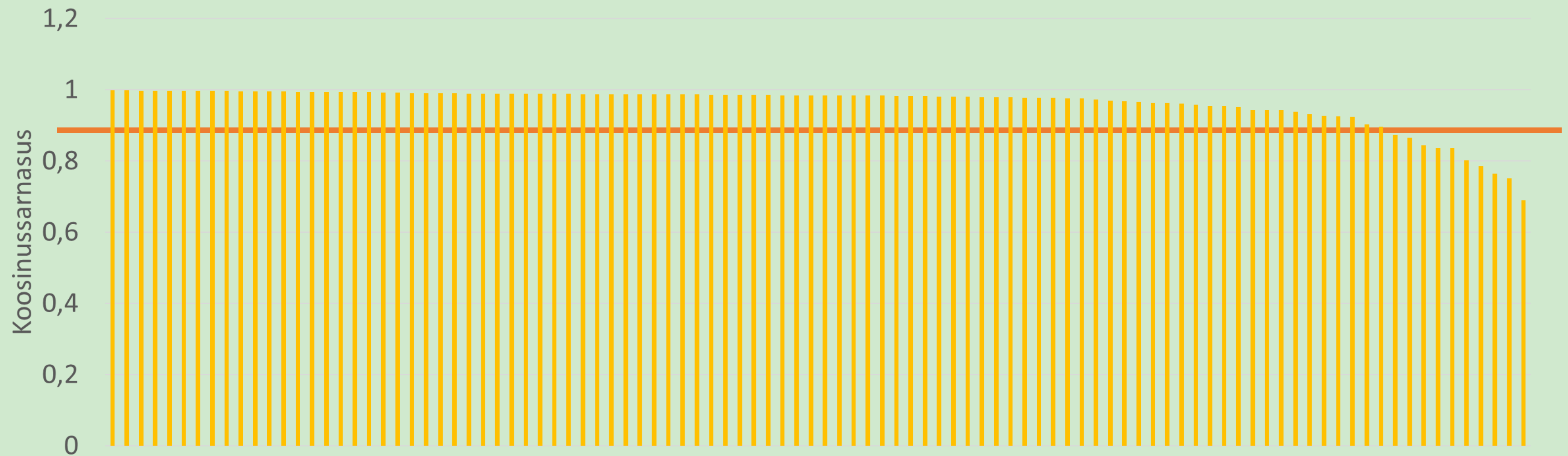


Prototüüpsete adjektiivide perifeeria eukleidilise kauguse põhjal

WORD	test_S	test_S_II	SB_test_	V_?_test	D_test	võrdevo	olema_()	Eukleide	koosinus
selge	0,256	0,202	0,011	0,064	0,205	0,071	0,275	0,523	0,844
ükskõikne	0,246	0,174	0,009	0,051	0,269	0,032	0,144	0,533	0,865
keeruline	0,344	0,270	0,016	0,086	0,399	0,281	0,306	0,537	0,785
kannatlik	0,247	0,207	0,018	0,039	0,209	0,070	0,376	0,568	0,764
lihtne	0,359	0,279	0,035	0,081	0,312	0,429	0,299	0,581	0,751
täpne	0,724	0,628	0,089	0,145	0,200	0,611	0,122	0,596	0,872
kade	0,173	0,132	0,013	0,038	0,266	0,004	0,295	0,640	0,689
kõrge	0,702	0,501	0,043	0,137	0,283	0,742	0,135	0,717	0,801



Prototüüpse adjektiivi (N =100) koosinussarnasused ja võimalik lävend (= 0,9)



weather	0.7549	0.8274	0.6436	0.7688	0.7839	0.6254	0.7693	0.774	0.7682
fragrance	0.7788	0.8721	0.8578	0.7386	0.841	0.8938	0.8256	0.7999	0.7978
positivity	0.6822	0.7733	0.6811	0.7479	0.7668	0.6279	0.7528	0.6996	0.7571
interest	0.7714	0.6628	0.6436	0.7833	0.7878	0.6738	0.7496	0.7688	0.7681
supper	0.484	0.4754	0.8284	0.8863	0.7984	0.8	0.7861	0.7688	0.7667
hatchbacks	0.6857	0.477	0.8687	0.7184	0.7441	0.8	0.7827	0.8294	0.788
remote-area	0.7574	0.6788	0.847	0.7488	0.788	0.8971	0.7814	0.7947	0.788
nanopore	0.6438	0.7073	0.8793	0.781	0.7341	0.767	0.783	0.8268	0.7667
cellulose	0.6668	0.8928	0.8726	0.7384	0.778	0.8188	0.7188	0.8881	0.7668
elbow	0.6754	0.6789	0.8884	0.7374	0.7988	0.889	0.8793	0.8884	0.7668
telephony	0.8738	0.6611	0.8844	0.8833	0.7387	0.8734	0.7811	0.7488	0.8881
hazmat	0.6681	0.8679	0.8881	0.8888	0.7883	0.8888	0.8888	0.8888	0.8887
handcuffs	0.7938	0.8277	0.8489	0.8878	0.8888	0.8881	0.8681	0.7181	0.781
sports	0.7413	0.827	0.8883	0.78	0.8891	0.8888	0.8833	0.7487	0.7668
black-chalk	0.7838	0.8881	0.8888	0.7487	0.7321	0.8683	0.7884	0.7888	0.8887
pillbox	0.4884	0.47	0.8883	0.8881	0.8883	0.8687	0.7498	0.7888	0.8881
silver-oval	0.6788	0.4787	0.8887	0.7379	0.7381	0.8887	0.8887	0.7888	0.7888
benches	0.8833	0.41	0.888	0.8833	0.7718	0.8884	0.7888	0.7887	0.8881
intermittent	0.6711	0.4888	0.8687	0.7788	0.8888	0.887	0.7418	0.7191	0.8881
lawn-tilt	0.8188	0.8788	0.8879	0.7888	0.7421	0.8887	0.8833	0.7817	0.8881
radiation	0.7774	0.8888	0.8778	0.7888	0.78	0.8881	0.8888	0.7887	0.8881
crack-leads	0.8881	0.8787	0.8881	0.8888	0.7891	0.8881	0.7881	0.7888	0.8881
over-the-ear	0.8879	0.7881	0.8881	0.749	0.7374	0.8887	0.8888	0.7888	0.8884
rhinoceros	0.6778	0.8888	0.8884	0.8881	0.8888	0.8881	0.7771	0.7488	0.8881
skirt	0.8888	0.8888	0.8881	0.8888	0.7888	0.8887	0.7888	0.8888	0.8888
napkins	0.8788	0.48	0.8884	0.778	0.7381	0.8881	0.7388	0.7887	0.8888
trucks	0.7888	0.8888	0.8884	0.7888	0.8888	0.8881	0.741	0.7888	0.8881
handbags	0.8788	0.8888	0.8887	0.7188	0.8881	0.8881	0.7887	0.788	0.8881
leaves	0.8887	0.8888	0.8888	0.8888	0.8884	0.788	0.7784	0.8887	0.8887
tennis-balls	0.7884	0.8881	0.8881	0.7888	0.7888	0.7887	0.8888	0.7887	0.8888
lidless	0.8888	0.7884	0.8881	0.7888	0.8887	0.8888	0.8881	0.7887	0.8884
on-dribble	0.8888	0.8888	0.8887	0.7888	0.788	0.8887	0.7888	0.7888	0.8888
film-film	0.8888	0.8888	0.8881	0.8888	0.7881	0.8881	0.8888	0.7887	0.8884
gloves	0.7888	0.7888	0.7887	0.7487	0.7884	0.8888	0.8888	0.7881	0.8888
hazmat	0.8888	0.8887	0.8887	0.8887	0.7887	0.8887	0.7887	0.7888	0.8887
handbags	0.8881	0.8881	0.8881	0	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881	0.8881
cellulose	0.8881	0.8881							

Prototüüpse adjektiivi parameetrite väärtused reastatuna koosinussarnasuse järgi

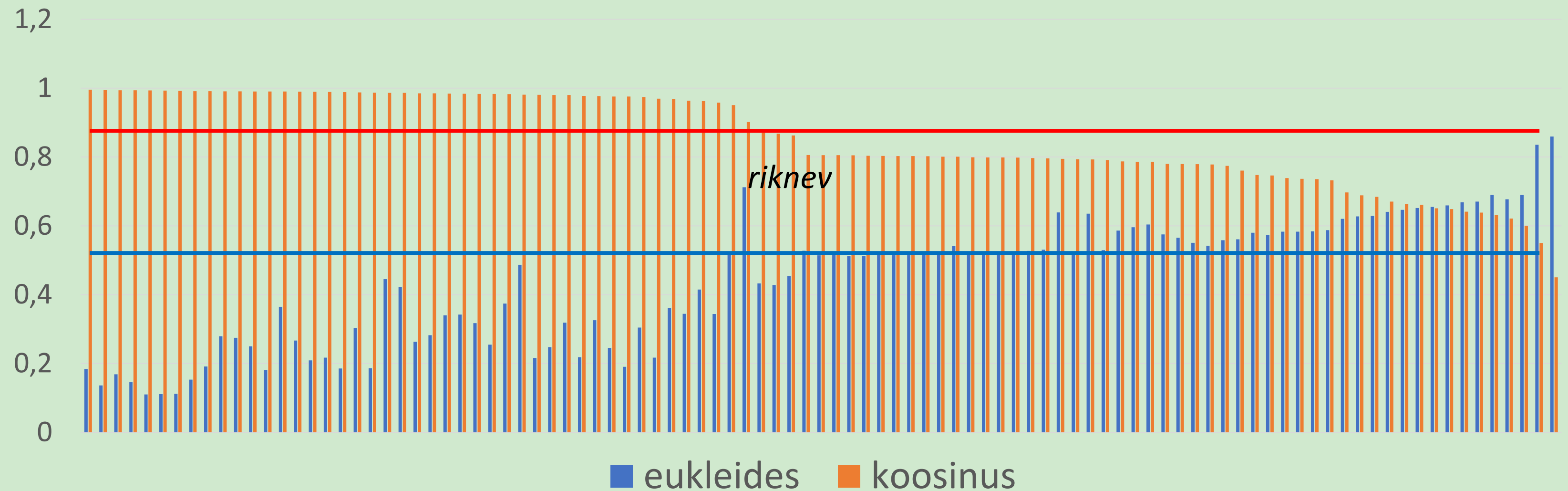
Kõige sarnasemad referentsprofiilile, st igas mõttes keskmised : *reaalne, traagiline, positiivne, inimlik*
Keskmissi ületavad, kuid tasakaalus: *tihe, täpne, kõrge*.

Mall 1: ametlik, harilik, pidulik, avalik, tehniline, tänane, üksikasjalik, põhiline, haruldane, alaline, sõjaline, eelmine

Mall 2: loogiline, õudne, kurb, vastupidav, ükskõikne, selge, kõhn, lahja, keeruline, kannatlik, lihtne, kade

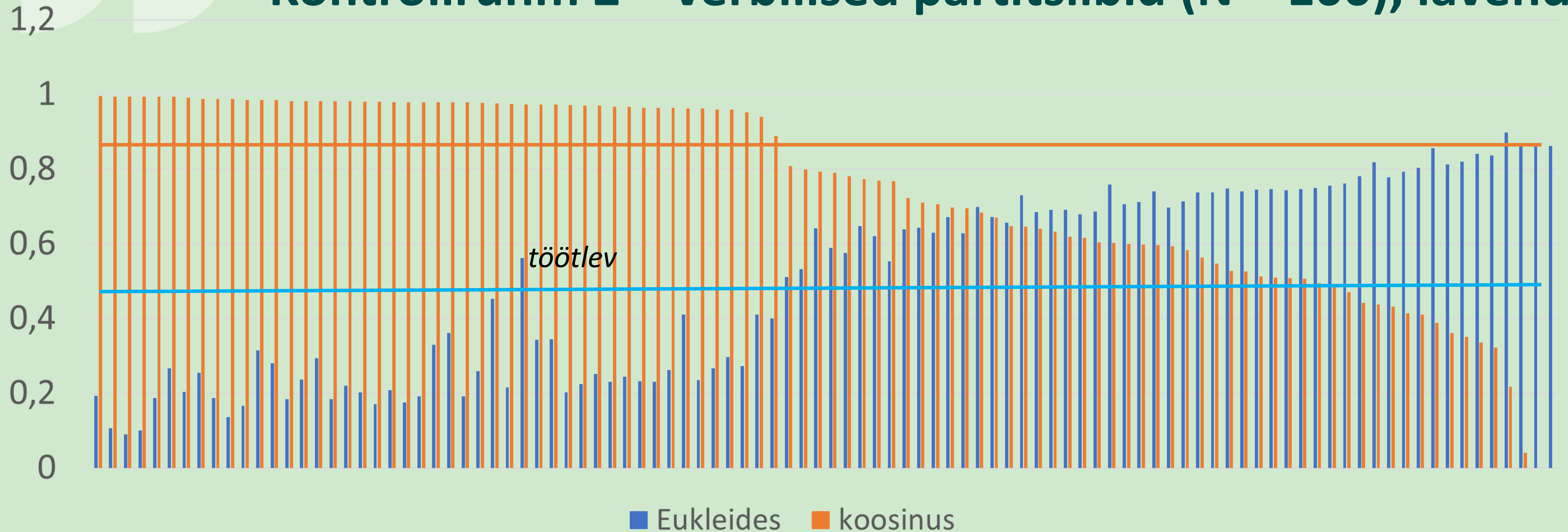


Kontrollrühm 1 – Ekilexi partitsiibid (N =100), lävendid





Kontrollrühm 2 – verbalised partitsiibid (N = 100), lävendid





Kõrvalekaldeanalüüs

- näitab, kas ja mil määral hälbivad teised testrühmad (Ekilexi partitsiibid ja verbilised partitsiibid) prototüüpsete adjektiivide järgi sätitud piirmääradest
- igale parameetrile on prototüüpsete adjektiivide alusel määratud miinimum- ja maksimumväärtus ja uurime, kuidas teiste testrühmade sõnad määratud vahemikku mahuvad
- võimaldab välja selgitada parima eristusvõimega parameetrid
- **EELDUS:** Ekilexi partitsiipide seas rohkem adjektiviseerunud partitsiipe, kui verbiliste partitsiipide rühmas



Piirvahemiku seadistamine

- võrdevormiparameeter seadistatud sõna keskvõrde korpuse üldsageduse põhjal:

$$5 < \dots$$

- teised parameetrid prototüüpsete adjektiivide vastavate osakaalude alusel (parameetrite test_S ja LA_test_S juures jäetud arvestamata üksikud rühmast tugevalt hälbivad sõnad)



Adjektiivsuse skaala

Kattuvate parameetrite arv	0	1	2	3	4	5	6	7
ekilex			3	9	16	39	27	6
verbilised	1	11	21	16	20	14	16	1

5, 6, 7 parameetri põhjal prototüüpsete adjektiividega sarnased:

72% Ekilexi partitsiipidest; **31%** verbilistest partitsiipidest (rohkem kui poole vähem!)

Prototüüpsetest adjektiividest kaugemad = 0, 1, 2 parameetriga kattuvus:

ainult **3%** Ekilexi partitsiipidest; **33%** verbilistest partitsiipidest



Adjektiivsuse skaala näited

	0	1	2	3	4	5	6	7
ekilex			soostunud			hõõguv	kergendav	kleepuv
ekilex						muteerunud	tolmav	selgitav
verbilised		lootnud	hakkav			müüv	korratav	
verbilised	kihlutud	langetud	jätnud			kiusav	vaidlev	hinnatav



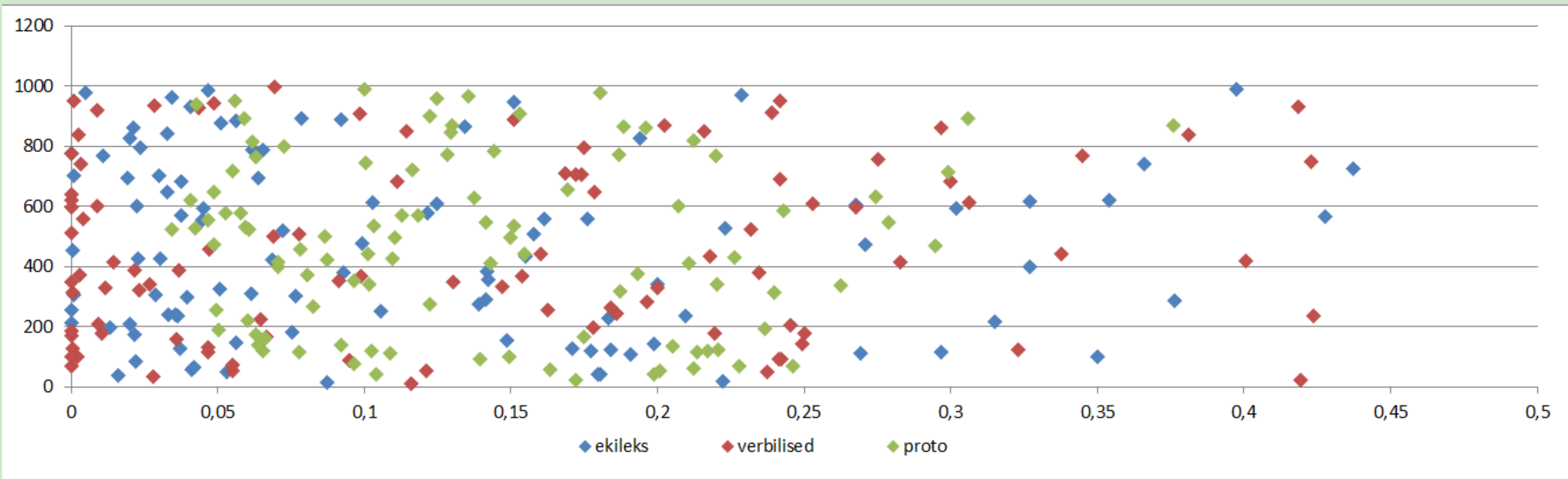
Parameetrite eristusvõime

	test_S	test_S_ÜHILD	LA_test_S	V_?_test_S	D_test	võrre	olema_(D)_test
ekilex	93	51	74	91	89	26	72
verbilised	62	49	25	63	75	14	66

- võrdeparameeter eristab hästi verbilisi partitsiipe prototüüpsetest adjektiividest, ent välistab ka hulga Ekilexi partitsiipe
- atribuudi- ja neljakohalise parameetri tulemused kattuvad, piisaks ühest (vähemalt partitsiipide analüüsil)
- parim eristaja: lausealguseparameeter
- nõrgemad eristajad: ühildumisparameeter, predikatiiviparameeter

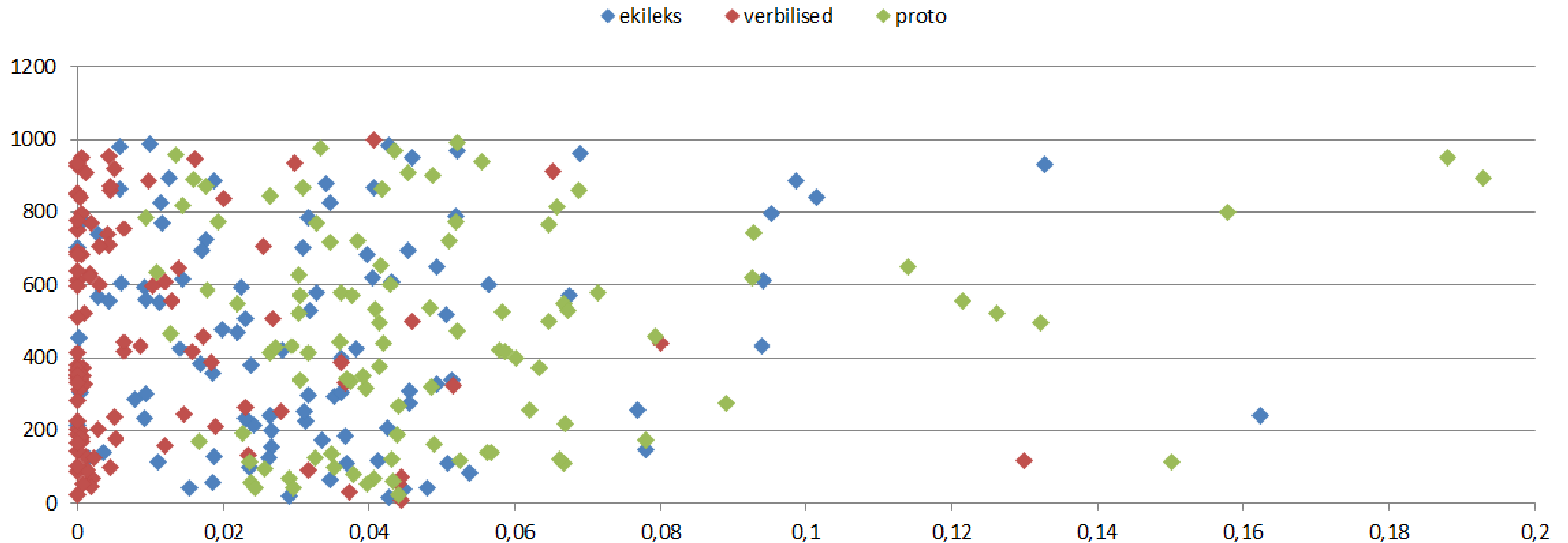


Predikatiiviparameeter testrühmade lõikes (nõrk eristaja)





Lausealguseparameeter testrühmade lõikes (tugev eristaja)





Kokkuvõtteks I

- Adjektiivide rühma sees varieeruvust, paigutuvad skaalale, perifeeria.
- Prototüüpset adjektiivi meie valimis ei iseloomusta võrdevormide kasutamine.
- Sarnasusmõõdikud lasevad mõõta profiili sarnasust prototüübiga
- Käänduvate verbipartitsiipide puhul (-v, -tav) on nende atributiivne ja ühilduv kasutus domineeriv, mistõttu sarnasus adjektiividele suur.
- Kontrollrühmade adjektiivsust määrab täpsemalt kõrvalekaldeanalüüs, mis arvestab adjektiividega sarnanemist parameetrite kaupa ja vajadusel kitsendab tingimusi.



Kokkuvõtteks II

- Täpsemalt eristas verbilisi partitsiipe prototüüpsetest adjektiividest lausealgusparameeter ja võrdeparameeter.
- Predikatiiviparameeter ja adverbiparameeter vajavad edasist häälestamist, praegusel kujul nad verbilisi partitsiipe (nt *on räägitud*, *hästi tehtud*) adjektiviseerunud verbivormidest ega prototüüpsetest adjektiividest (nt *on ohutu*, *väga kõrge*) eristada ei suuda.
- Parameetrite komplekt tervikuna paigutas Ekilexi partitsiipide rühma adjektiividele lähemale kui verbiliste partitsiipide rühma.

“

AITÄH

”



Kirjandus

Kerge, Krista 1998. *Vormimoodustus, sõnamoodustus ja leksikon: oleviku kesksõna võrdluse all*. Tallinn: TPÜ Kirjastus.

Paulsen, Geda; Vainik, Ene; Tuulik, Maria 2020. Sõnaliik leksikograafi töölaual: sõnaliikide roll tänapäeva leksikograafias. – *ERÜ aastaraamat* 16, 177–202.

Tuulik, Vainik, Paulsen, Lohk 2022: Kuidas ära tunda adjektiivi? Korpuskäitumise mustrite analüüs. – *ERÜ aastaraamat* 18.

Vainik, Ene; Paulsen, Geda; Lohk, Ahti 2020. A typology of lexical ambiforms in Estonian. – Z. Gavriilidou, M. Mitsiaki, A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion*, Vol. 1. Alexandroupolis: Democritus University of Thrace, 119-130.

https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p119-130.pdf



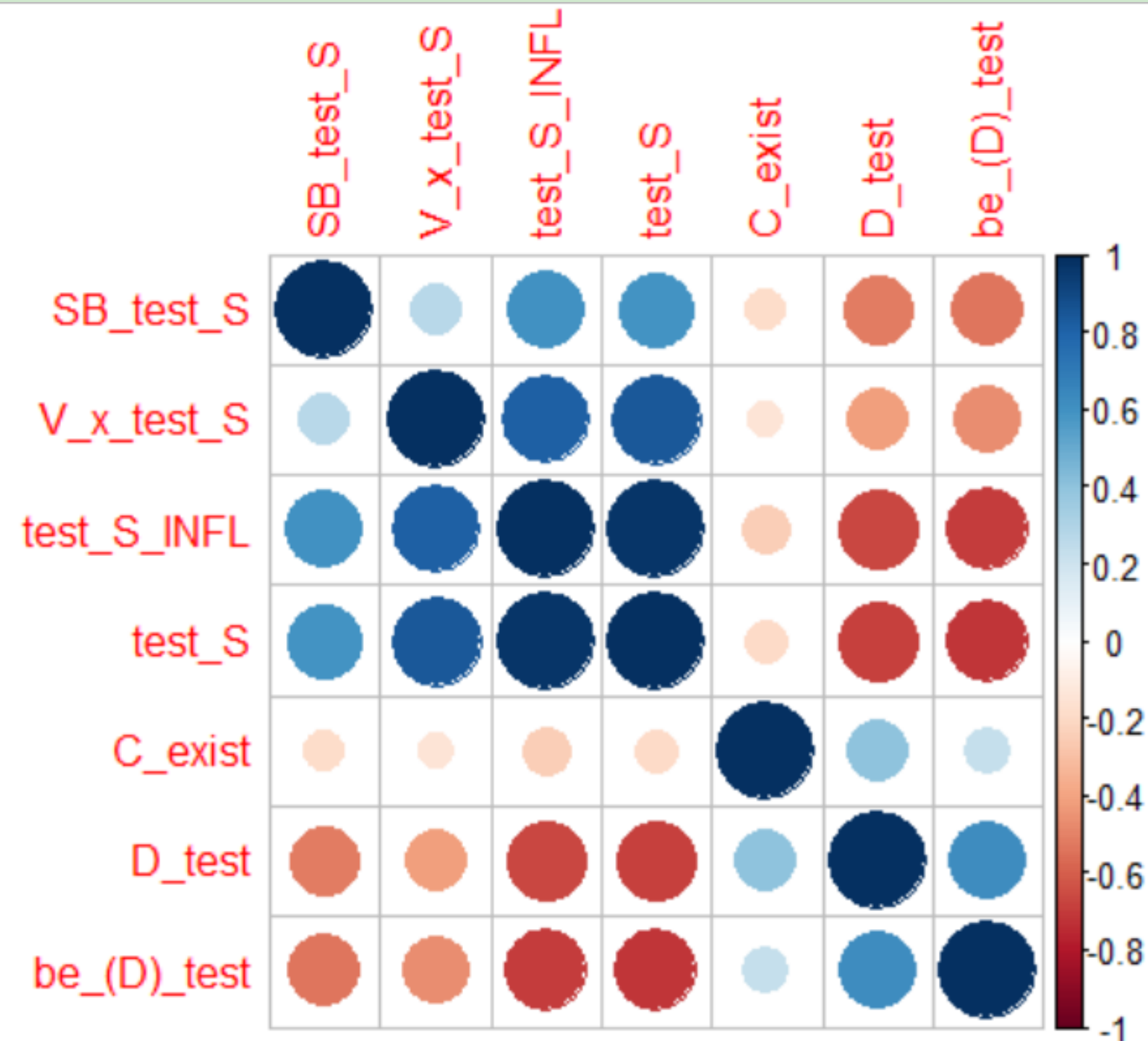
Arutelu

- Adjektiiv võib esineda väga mitmesugustes fraasitüüpides, mida meie parameetrid ei kata – oleme valinud (oletuslikult) kõige kesksamad lausemusterid.
- Predikatiiviparameeter toob kaasa mingil määral müra (*kohale on kandideeritud*).
Õeldistäide nimetavasse? *nud-/tud*-partitsiibid ei käändu, *tav*-vorm käitub silmapaistvalt adjektiivselt (*katse on korratav*).
- Adjektiivi "duaalne profiil", mis tuleneb vähemalt osaliselt parameetrite keskendumisest atributiivsele ja muule käitumisele (predikatiivsus, adverbifraas, keskvõrre).
--> kaks erinevat prototüübiprofiili, atribuudi- ja predikatiivikalduvustega adjede jaoks?
- **Rakenduse** teostatavus – ühendamine SkE-ga (testsõna sõnaliik määramata!), multiparameetriline rakendus. Parameetrite arv ja täiustamine. Partitsiipidele lisaks ka atüüpilised adjektiivid (käändumatud, adverbilised...)

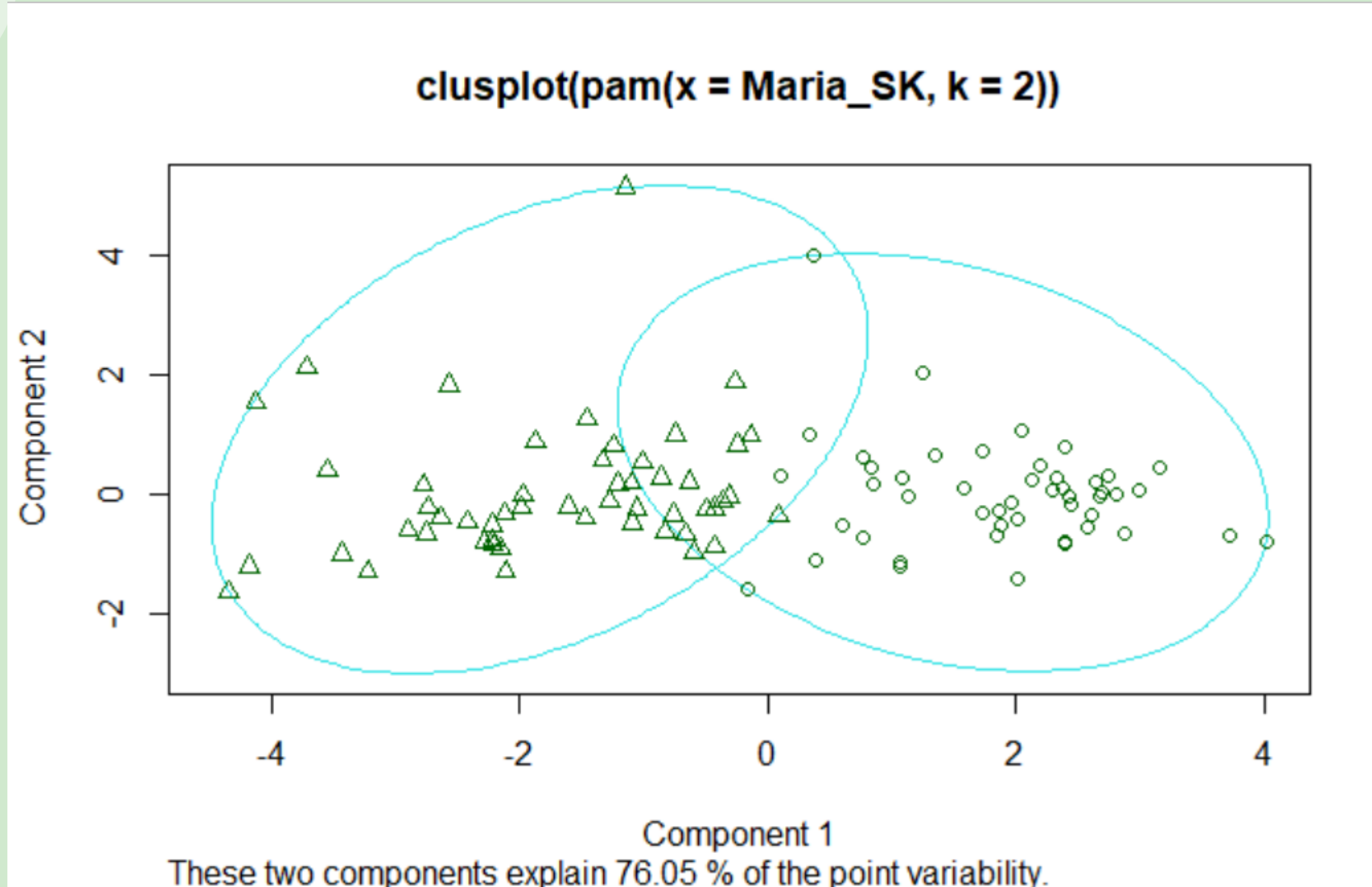


Lisamaterjal

Parameetrite korrelatsioonid (prot adje)



Klasteranalüüs (prot adje)



	Nimetus	Taanis	Loogiline avaldis
1.	atribuudiparameeter	test_S	lemmas[i].lower() in test_words and postags[i+1] == 'S'
2.	ühildumisparameeter	test_S_ÜHILD	lemmas[i].lower() in test_words and postags[i+1] == 'S' and forms[i] == forms[i+1]
3.	lausealguseparameeter	LA_ test_S	i == 0 and lemmas[i].lower() in test_words and postags[i+1] == 'S'
4.	neljakohaline parameeter	V_? test_S	i < sent_len - 3 and postags[i] == "V" and lemmas[i+2] in test_words and postags[i+3] == "S"
5.	adverbiparameeter	D_test	postags[i] == 'D' and lemmas[i+1].lower() in test_words
6.	predikatiiviparameeter	<i>olema_test</i> <i>olema_D_test</i>	i > 0 and lemmas[i] == 'olema' and lemmas[i].lower() in test_words "olema_D_test": i > 1 and lemmas[i-2] == 'olema' and postags[i-1] == 'D' and lemmas[i].lower() in test_words
7.	võrdeparameeter		lemmas[i] in test_comparatives