Mapping language varieties

Koen Plevoets
Faculty of Applied Language Studies
University College Ghent

Mapping language varieties

- The use of correspondence analysis to represent linguistic distances

Mapping language varieties

- The use of correspondence analysis to represent linguistic distances
- The varieties of Belgian Dutch

Mapping language varieties

- Belgian Dutch: a lot of linguistic variation
- Formal-informal

Mapping language varieties

- OBJECTIVE: Measure difference between ‘varieties’/‘registers’ of Belgian Dutch based on their degree of (in)formality

	Adm	Ext	Fic	Ins	Jour	Non_f
verkrijgen	187	144	1	33	28	17
bekomen	91	69	0	23	3	6
zodra	47	48	10	17	85	26
van.zodra	7	4	0	3	5	0
…						

Mapping language varieties

2 datasets:

- TSS: Standard vs. vernacular
  - Spoken Dutch Corpus (CGN)
  - 14 linguistic variables
  - Is ‘tussentaal’ the new omni-situational standard?
- COMURE: Translated vs. non-translated
  - Dutch Parallel Corpus
  - 13 linguistic variables
  - Is translated language more formal than non-translated language?

Overview

- Correspondence analysis: Chi-square as a distance metric
- Within-chi-square
- ‘Multifactorial’ correspondence analysis
- Statistical inference
- Conclusions

Correspondence analysis

- Frequency Table: Chi-square
- X^2 = measure of heterogeneity

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

$$E_{ij} = \frac{N_{i+} * N_{+j}}{N_{++}}$$

Correspondence analysis

$$E_{ij} = \frac{N_{i+} * N_{+j}}{N_{++}}$$

	Adm	Ext	Fic	Ins	Jour	Non_f	N_{i+}
verkrijgen	187	144	1	33	28	17	410
bekomen	91	69	0	23	3	6	192
zodra	47	48	10	17	85	26	233
van.zodra	7	4	0	3	5	0	19
N_{+j}	332	265	11	76	121	49	854

Correspondence analysis

- Chi-square as a distance metric
- Anderson (2003: 80): X^2 of 2 varieties = Mahalanobis distance between the 2 varieties

$$D^2\left(Adm, Ext\right) = \left(Adm - Ext\right)^T * S^{-1} * \left(Adm - Ext\right)$$

- *S* = Covariance matrix of the linguistic items

Correspondence analysis

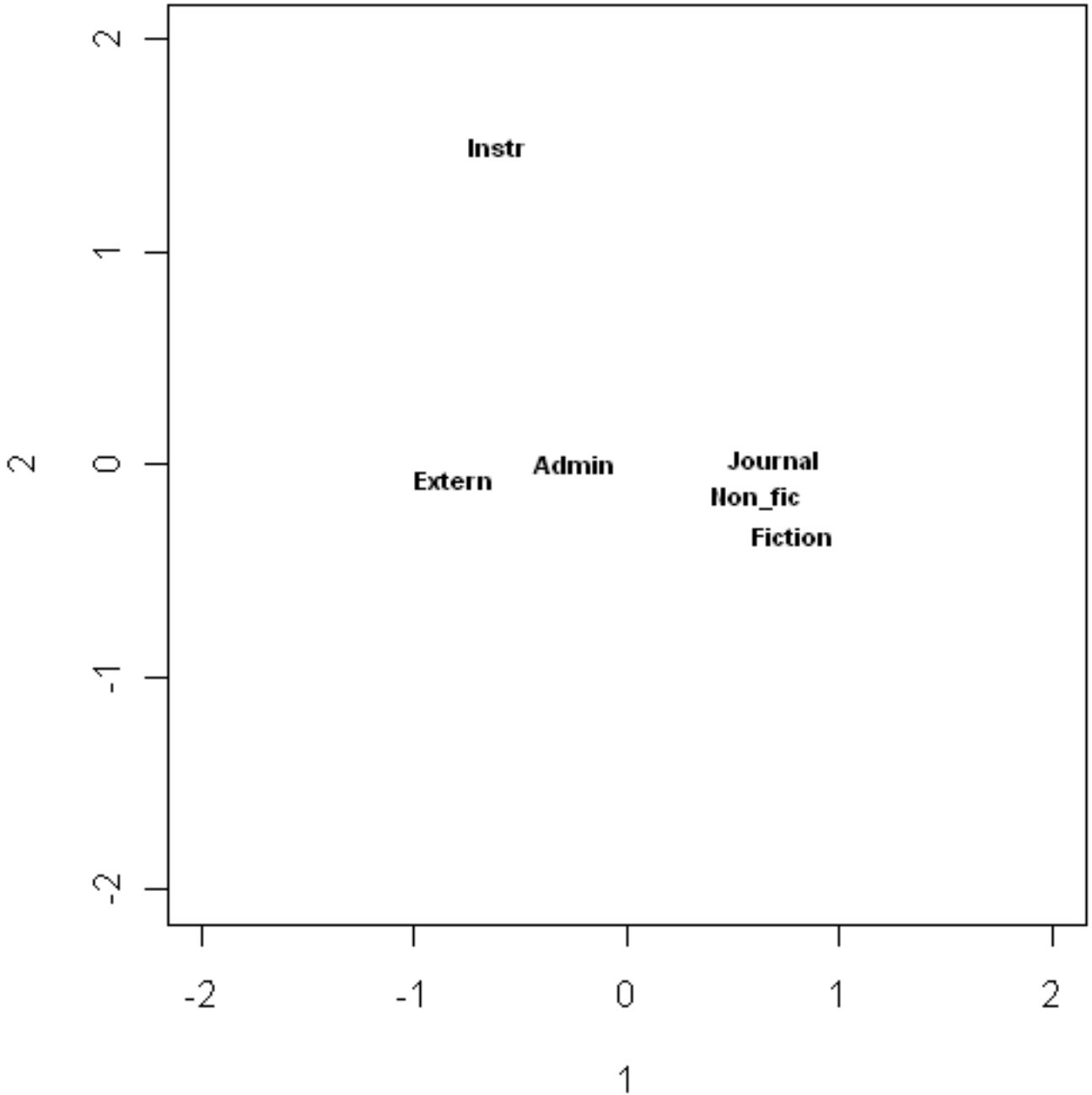
- OUTPUT: representation in low-dimensional space
- 'Singular Value Decomposition': $X = U*\Sigma*V^T$

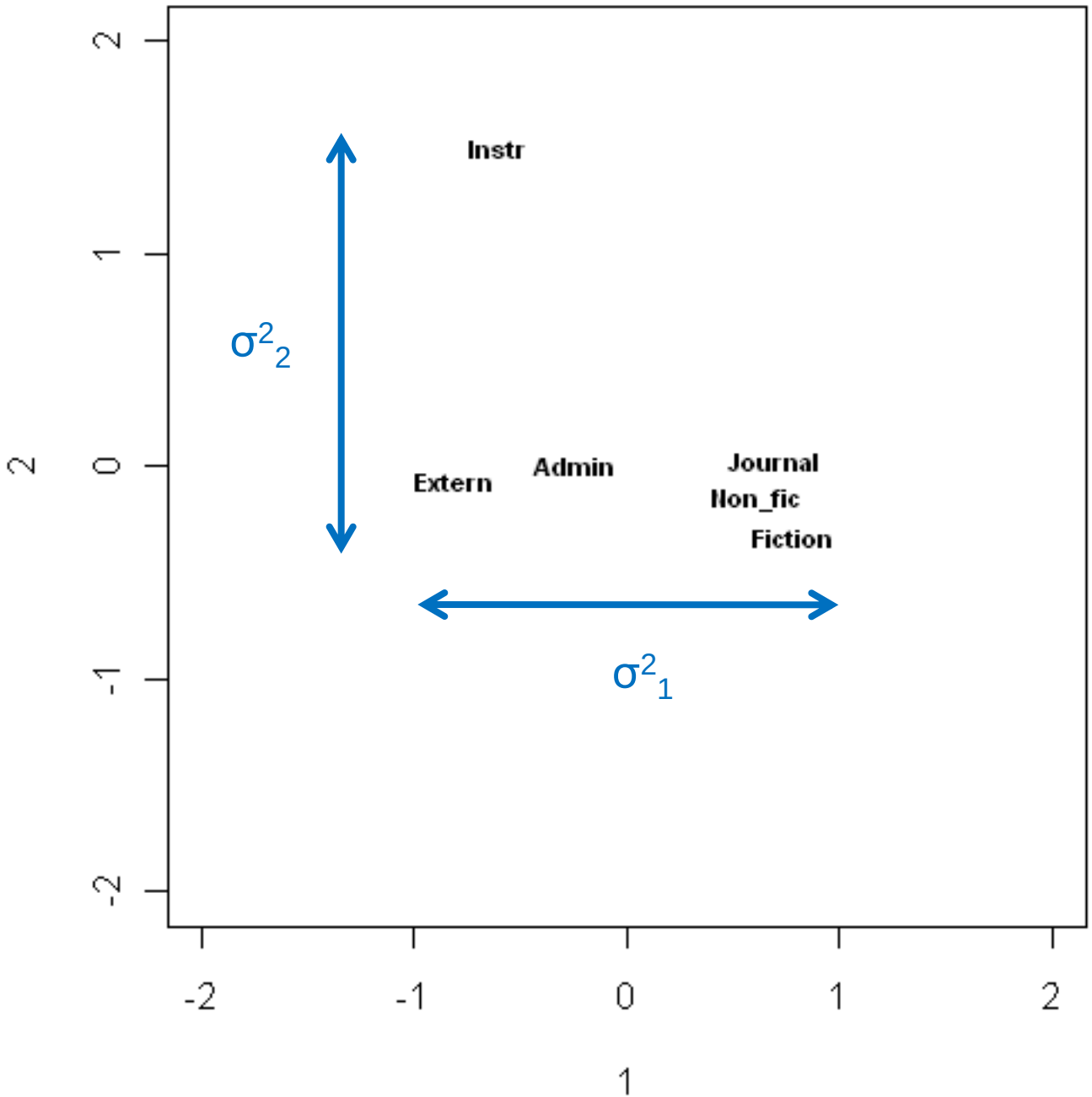
Correspondence analysis

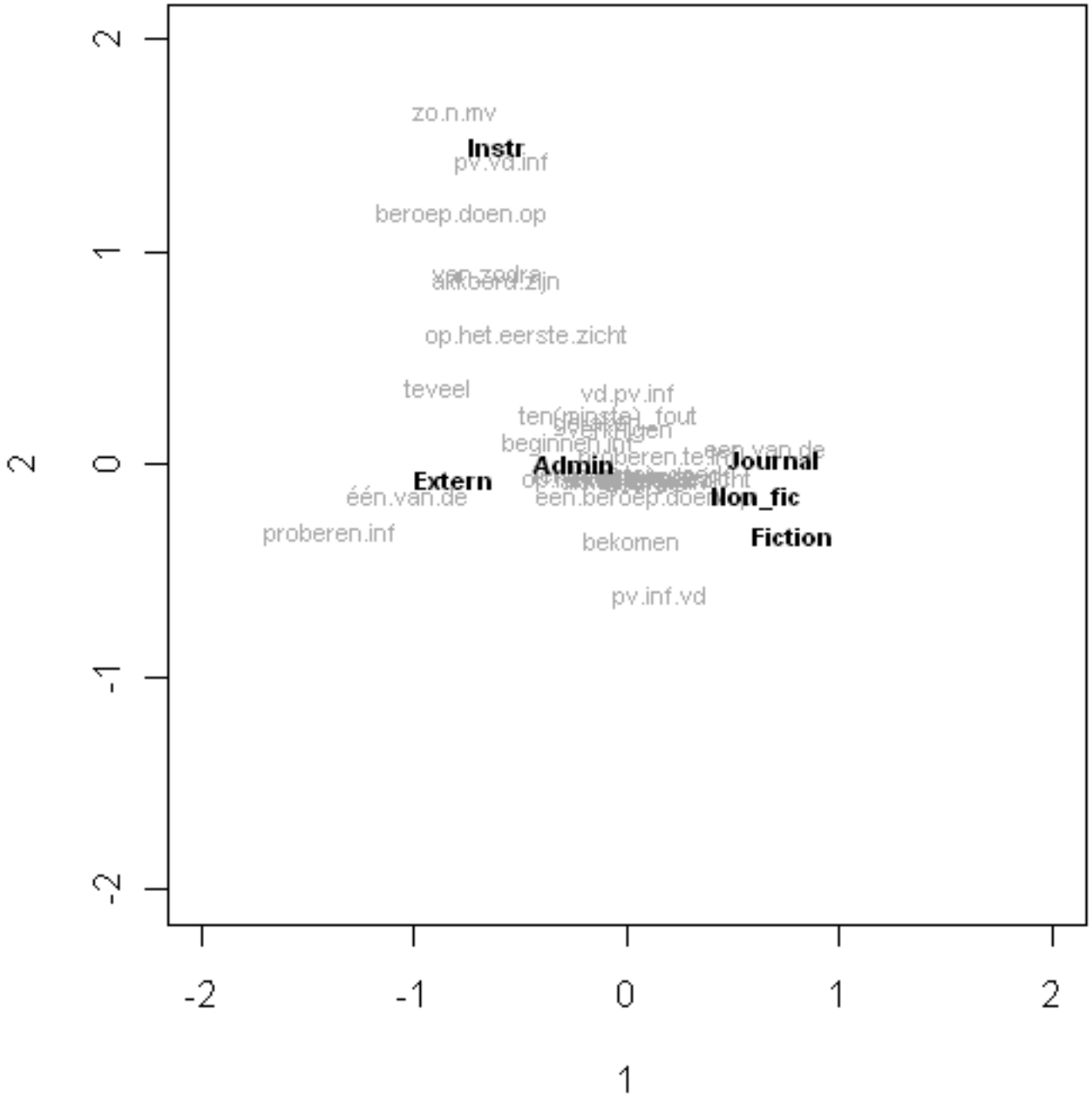
- OUTPUT: representation in low-dimensional space
- 'Singular Value Decomposition': $X = U*\Sigma*V^T$

V =

	1	2
Admin	-0.24385	0.00865
Extern	-0.80581	-0.05939
Fiction	0.79083	-0.32532
Instr	-0.59552	1.496837
Journal	0.70657	0.026645
Non_fic	0.611468	-0.14888

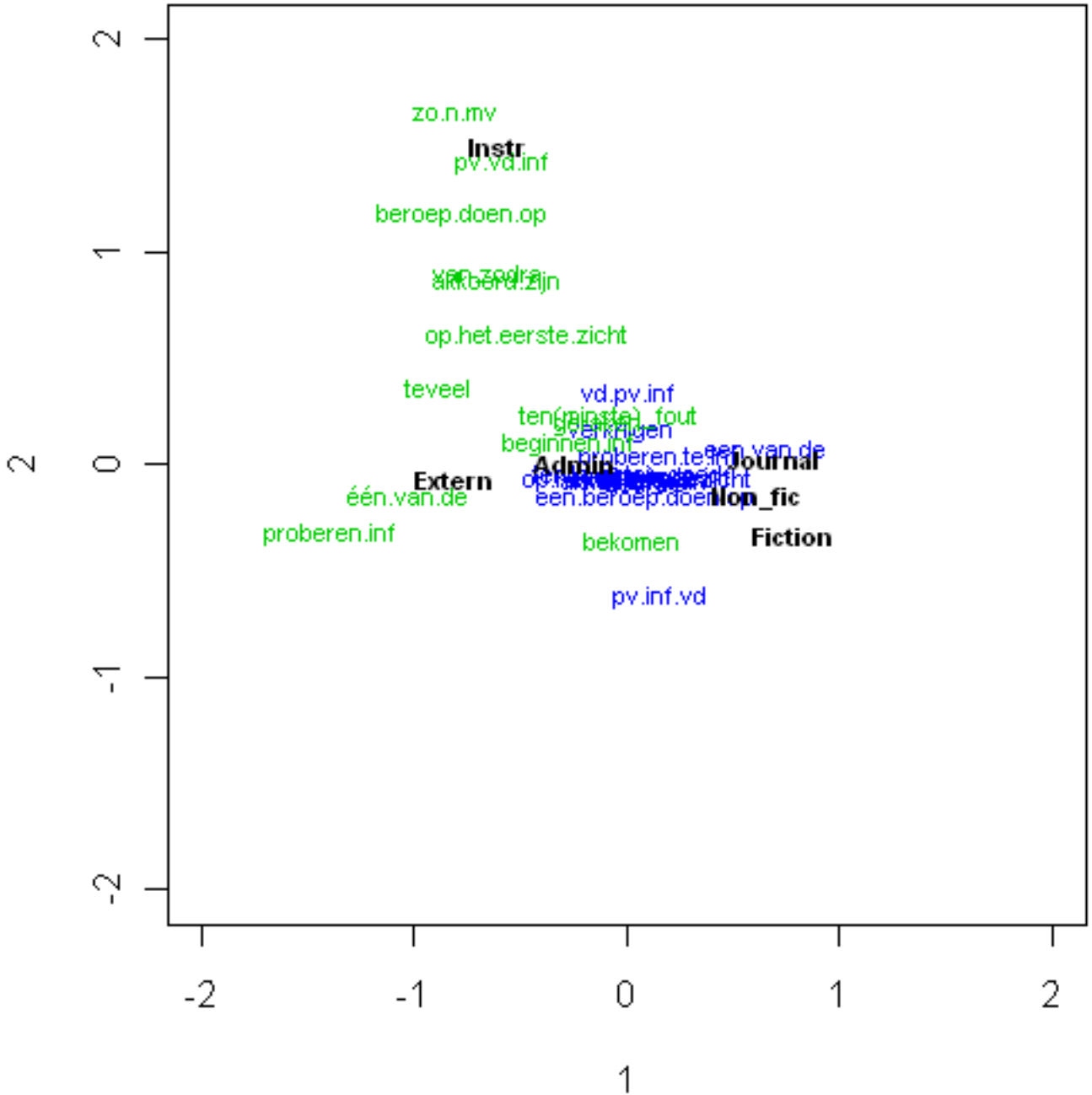

Instr
Extern
Admin
Journal
Non_fic
Fiction
2
1
0
-1
-2
-2
-1
0
1
2
1


Instr
Extern
Admin
Journal
Non_fic
Fiction
σ^2_2
σ^2_1
2
1
0
-1
-2
-2
-1
0
1
2
1


zo.n.mv
Instr
pv.vd.inf
beroep.doen.op
op.het.eerste.zicht
teveel
vd.pv.inf
Admin
Journal
Extern
Non_fic
één.van.de
een.beroep.doen
proberen.inf
bekomen
Fiction
pv.inf.vd
2
1
0
-1
-2
-2
-1
0
1
2
1

Biplot

- Gower, J., S. Lubbe & N. Le Roux (2011). *Understanding biplots*. Hoboken: Wiley.
- Greenacre, M. (2010). *Biplots in practice*. Bilbao: Fundación BBVA.


zo.n.mv
Instr
pv.vd.inf
beroep.doen.op
op.het.eerste.zicht
teveel
vd.pv.inf
beginnen.inf
Admin
Journal
Extern
één.van.de
een.beroep.doen
Non_fic
proberen.inf
bekomen
Fiction
pv.inf.vd
-2
-1
0
1
2
1
2

Within-chi-square

- ‘Linguistic variable’: alternation of linguistic variants
- HENCE: Linguistic items are partitioned

~ ‘Profile-based uniformity’ (Speelman, e.a. 2003)

Within-chi-square

	Adm	Ext	Fic	Ins	Jour	Non_f
verkrijgen	187	144	1	33	28	17
bekomen	91	69	0	23	3	6
zodra	47	48	10	17	85	26
van.zodra	7	4	0	3	5	0
pv.inf.vd	63	20	1	6	10	11
vd.pv.inf	20	14	1	9	19	1
pv.vd.inf	14	7	0	9	3	0
…						

Within-chi-square

	Adm	Ext	Fic	Ins	Jour	Non_f
verkrijgen	187	144	1	33	28	17
bekomen	91	69	0	23	3	6
zodra	47	48	10	17	85	26
van.zodra	7	4	0	3	5	0
pv.inf.vd	63	20	1	6	10	11
vd.pv.inf	20	14	1	9	19	1
pv.vd.inf	14	7	0	9	3	0
…						

Within-chi-square

- ‘Linguistic variable’: alternation of linguistic variants
- HENCE: Linguistic items are partitioned

- GOAL: Measure heterogeneity **per** variable (bracketing the heterogeneity across variables)

~ ‘Profile-based uniformity’ (Speelman, e.a. 2003)

Within-chi-square

- Huyghens' theorem (Greenacre 1984: 203-204)
- $X^2 = X^2_{Within} + X^2_{Between}$

$$X^2_{Within} = \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{j=1}^{J} \frac{\left(O_{ij} - N_{i+} * N_{+_k j} \Big/ N_{+_k +} \right)^2}{N_{i+} * N_{+j} \Big/ N_{++}}$$

$$X^2_{Between} = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(N_{+_k j} - N_{+_k +} * N_{+j} \Big/ N_{++} \right)^2}{N_{+_k +} * N_{+j} \Big/ N_{++}}$$

Within-chi-square

- Huyghens' theorem (Greenacre 1984: 203-204)
- $X^2 = X^2_{Within} + X^2_{Between}$

$$X^2_{Within} = \sum_{k=1}^{K} \sum_{i=1}^{I_k} \sum_{j=1}^{J} \frac{\left(O_{ij} - N_{i+} * N_{+_k j} \Big/ N_{+_k +} \right)^2}{N_{i+} * N_{+j} \Big/ N_{++}}$$

	Adm	Ext	Fic	Ins	Jour	Non_f	N_{i+}
verkrijgen	187	144	1	33	28	17	410
bekomen	91	69	0	23	3	6	192
zodra	47	48	10	17	85	26	233
van.zodra	7	4	0	3	5	0	19
…							

Within-chi-square

#	Variant 1	Variant 2	Variant 3	Translation
1	akkoord gaan	akkoord zijn		*to agree*
2	beginnen te +inf	beginnen + inf		*to begin to*
3	een beroep doen op	beroep doen op		*to appeal to*
4	een van de	één van de		*one of the*
5	op het eerste gezicht	op het eerste zicht		*at first sight*
6	proberen te + inf	proberen +inf		*to try to*
7	pv inf vd	vd pv inf	pv vd inf	*order of verbal end group*

Within-chi-square

#	Variant 1	Variant 2	Variant 3	Translation
8	raken	geraken		*to get*
9	te veel	teveel		*too much/many*
10	ten(minste)_goed	ten(minste)_fout		*at least*
11	verkrijgen	bekomen		*to obtain*
12	zodra	van zodra		*as soon as*
13	zulke + mv	zo'n + mv		*such + plural*

‘Multifactorial’ CA

- More than one factor coding the varieties
- Possibly with interactions

	DU_or	DU<EN	DU<FR
Adm	406	225	354
Ext	352	309	312
Fic	0	0	80
Ins	141	0	60
Jour	450	387	230
Non_f	363	0	93

‘Multifactorial’ CA

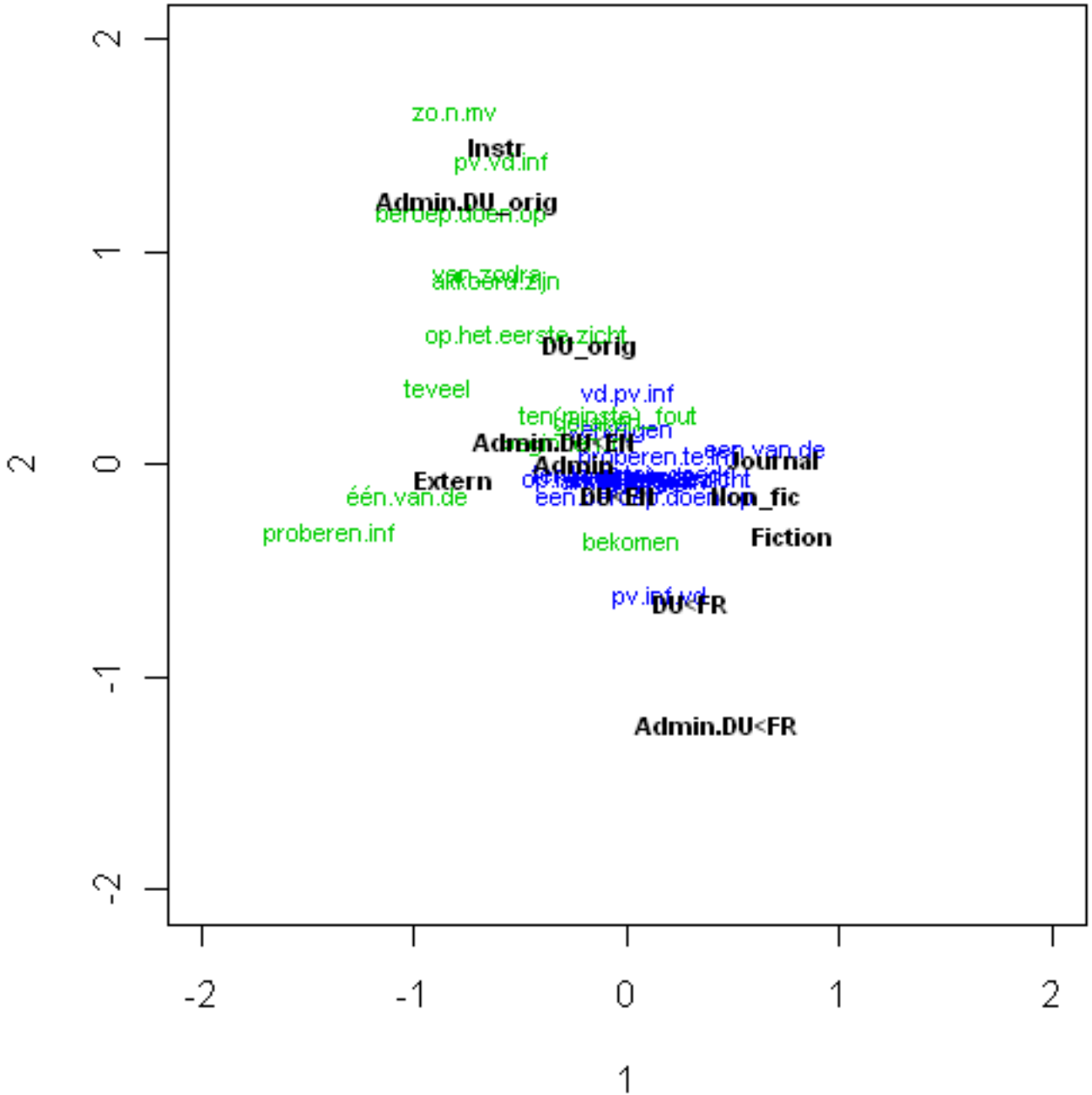
	Adm	Ext	Fic	Ins	Jour	Non_f	DU_or	DU<EN	DU<FR	…
verkrijgen	187	144	1	33	28	17	195	89	126	
bekomen	91	69	0	23	3	6	70	31	91	
zodra	47	48	10	17	85	26	93	37	103	
van.zodra	7	4	0	3	5	0	13	1	5	
pv.inf.vd	63	20	1	6	10	11	33	19	59	
vd.pv.inf	20	14	1	9	19	1	29	16	19	
pv.vd.inf	14	7	0	9	3	0	27	3	3	
…										

‘Multifactorial’ CA

…	Adm_or	Adm<EN	Adm<FR	Ext_or	Ext<EN	Ext<FR	Fic_or	Fic<EN	Fic<FR	…
	91	48	48	58	38	48	0	0	1	
	25	11	55	19	20	30	0	0	0	
	17	3	27	17	11	20	0	0	10	
	7	0	0	0	1	3	0	0	0	
	14	9	40	2	5	13	0	0	1	
	8	6	6	5	6	3	0	0	1	
	12	1	1	5	2	0	0	0	0	

‘Multifactorial’ CA

…	Ins_or	Ins<EN	Ins<FR	Jour_or	Jour<EN	Jour<FR	Non_or	Non<EN	Non<FR
	19	0	14	11	3	14	16	0	1
	18	0	5	3	0	0	5	0	1
	11	0	6	28	23	34	20	0	6
	3	0	0	3	0	2	0	0	0
	5	0	1	2	5	3	10	0	1
	7	0	2	8	4	7	1	0	0
	7	0	2	3	0	0	0	0	0


zo.n.mv
Instr
pv.vd.inf
Admin.DU_orig
beroep.doen.op
op.het.eerste.zicht
DU_orig
teveel
vd.pv.inf
Admin.DU_Elf
Admin
Extern
één.van.de
proberen.inf
bekomen
Journal
Non_fic
Fiction
DU<FR
Admin.DU<FR
2
1
0
-1
-2
-2
-1
0
1
2

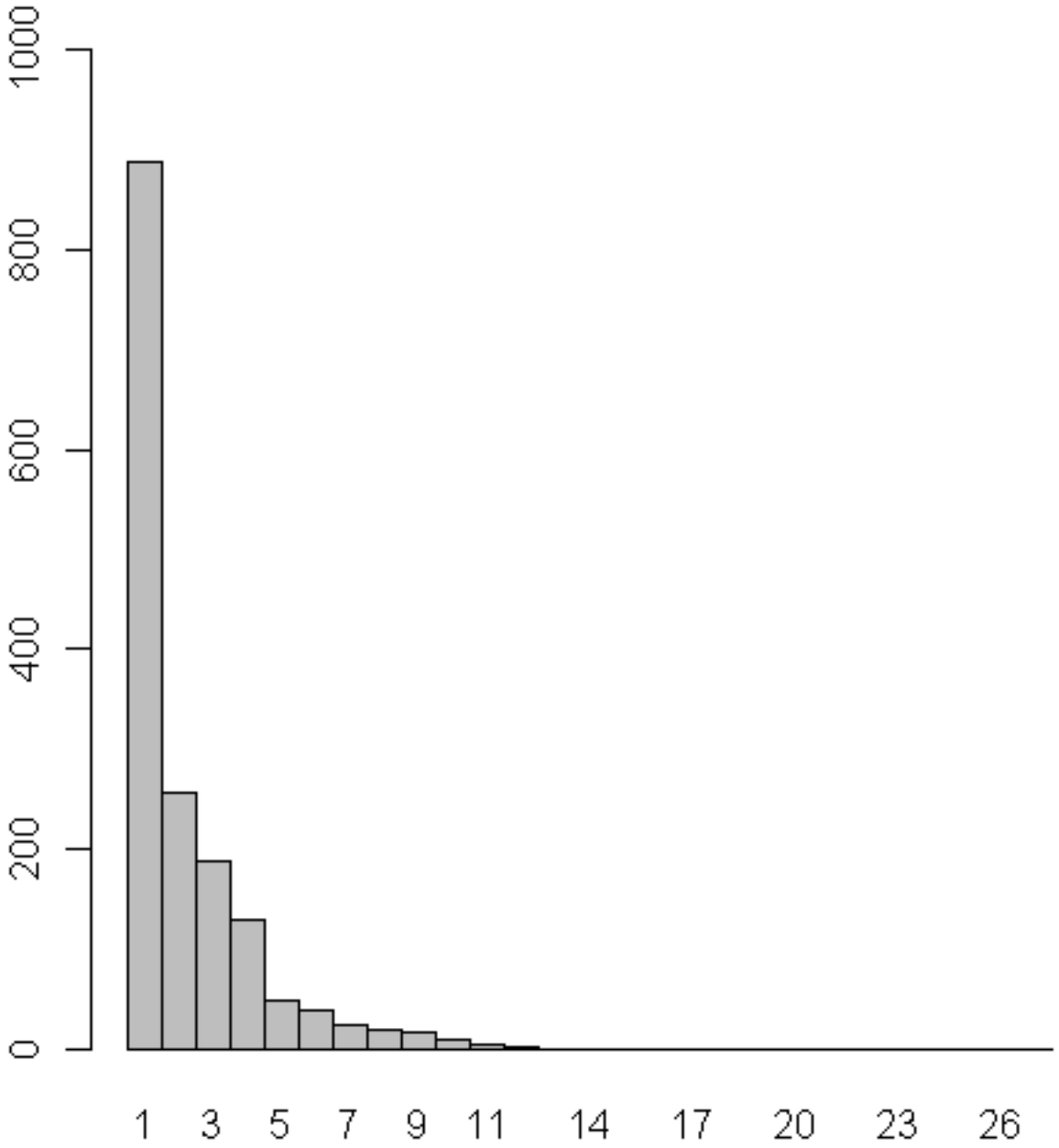
Statistical inference

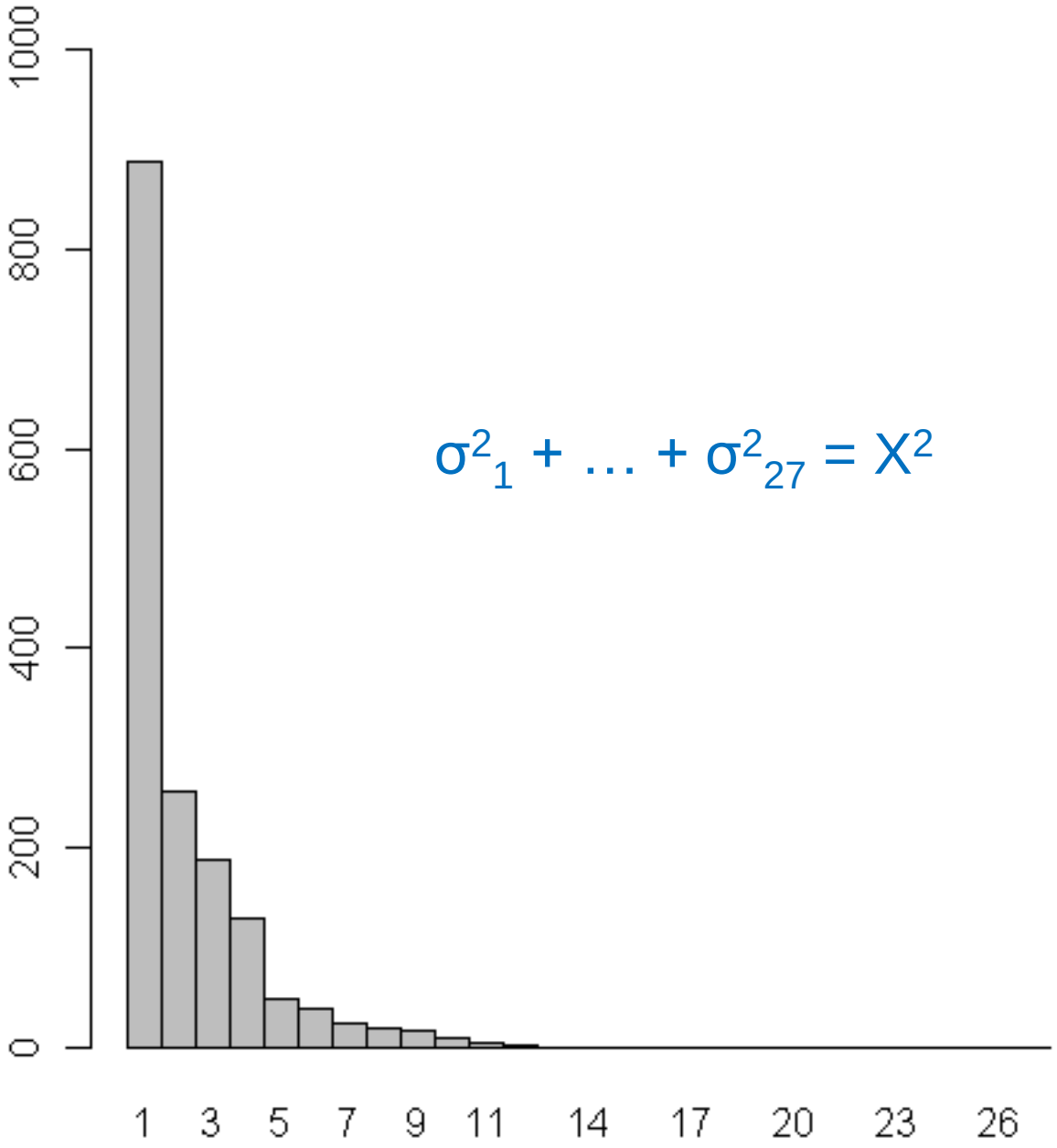
- How many dimensions?

→ Bootstrap confidence intervals for eigenvalues

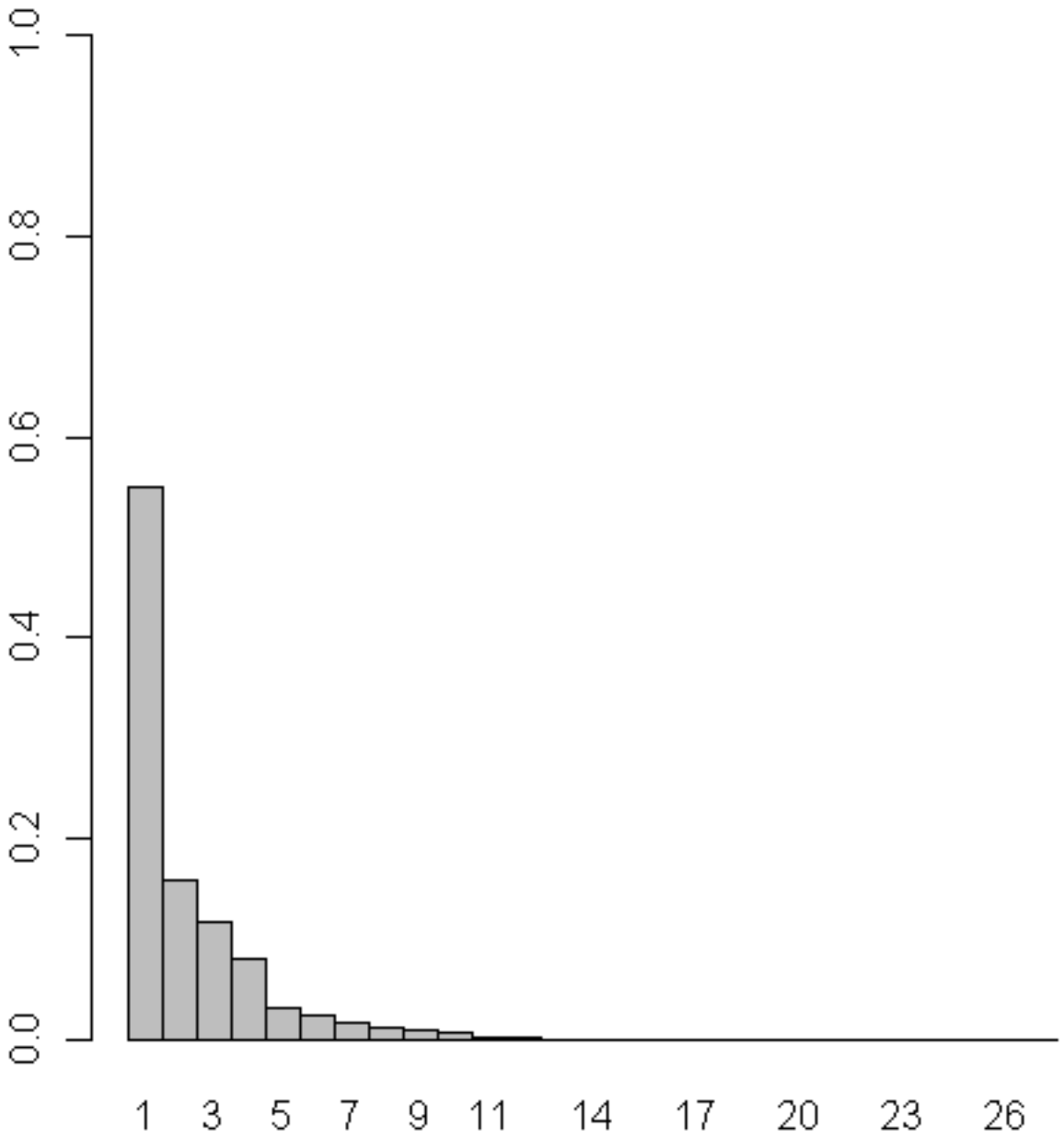
- Distances between varieties significant?

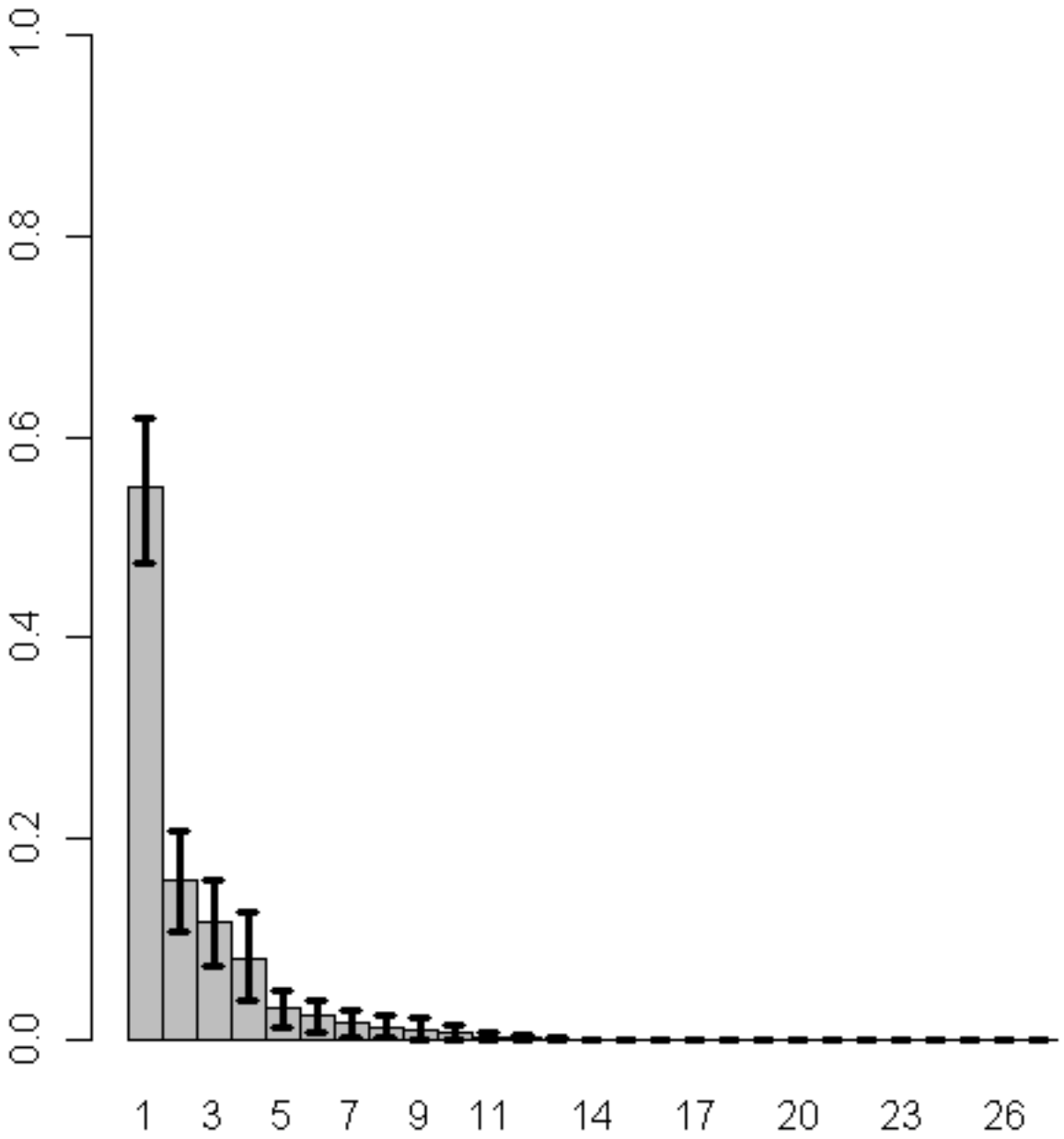
→ Bootstrap confidence regions for coordinates

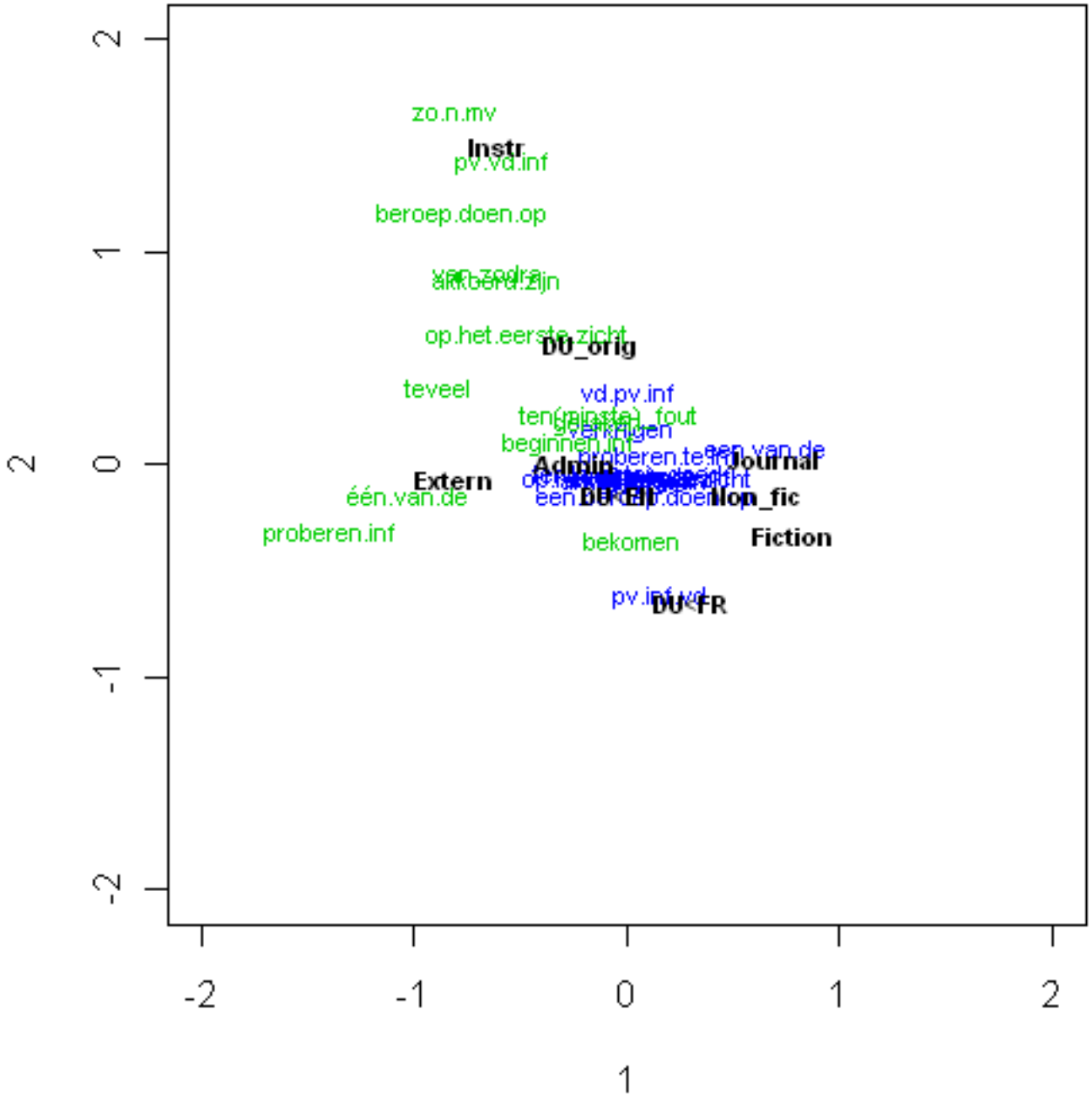

0
200
400
600
800
1000
1
3
5
7
9
11
14
17
20
23
26

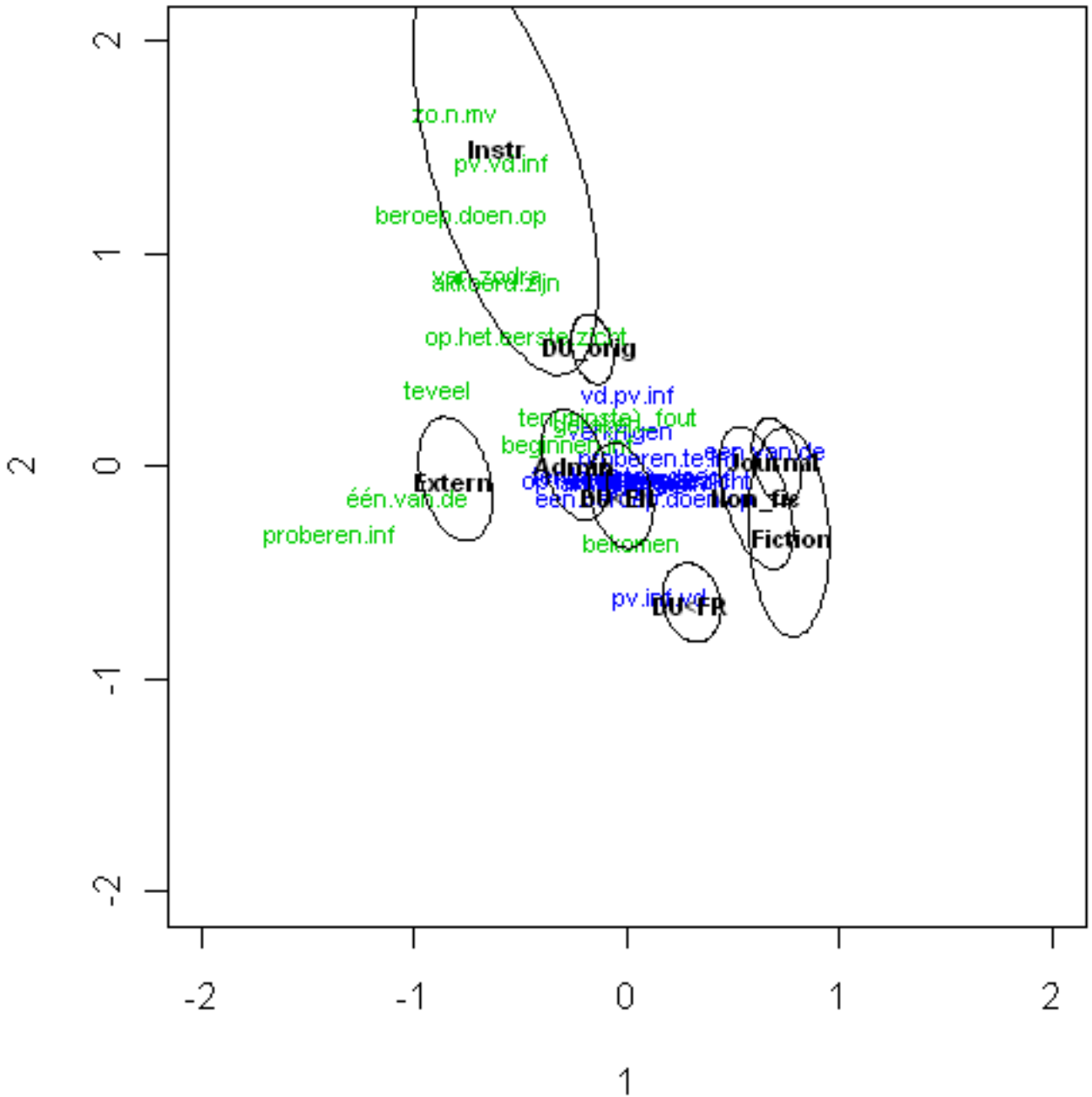


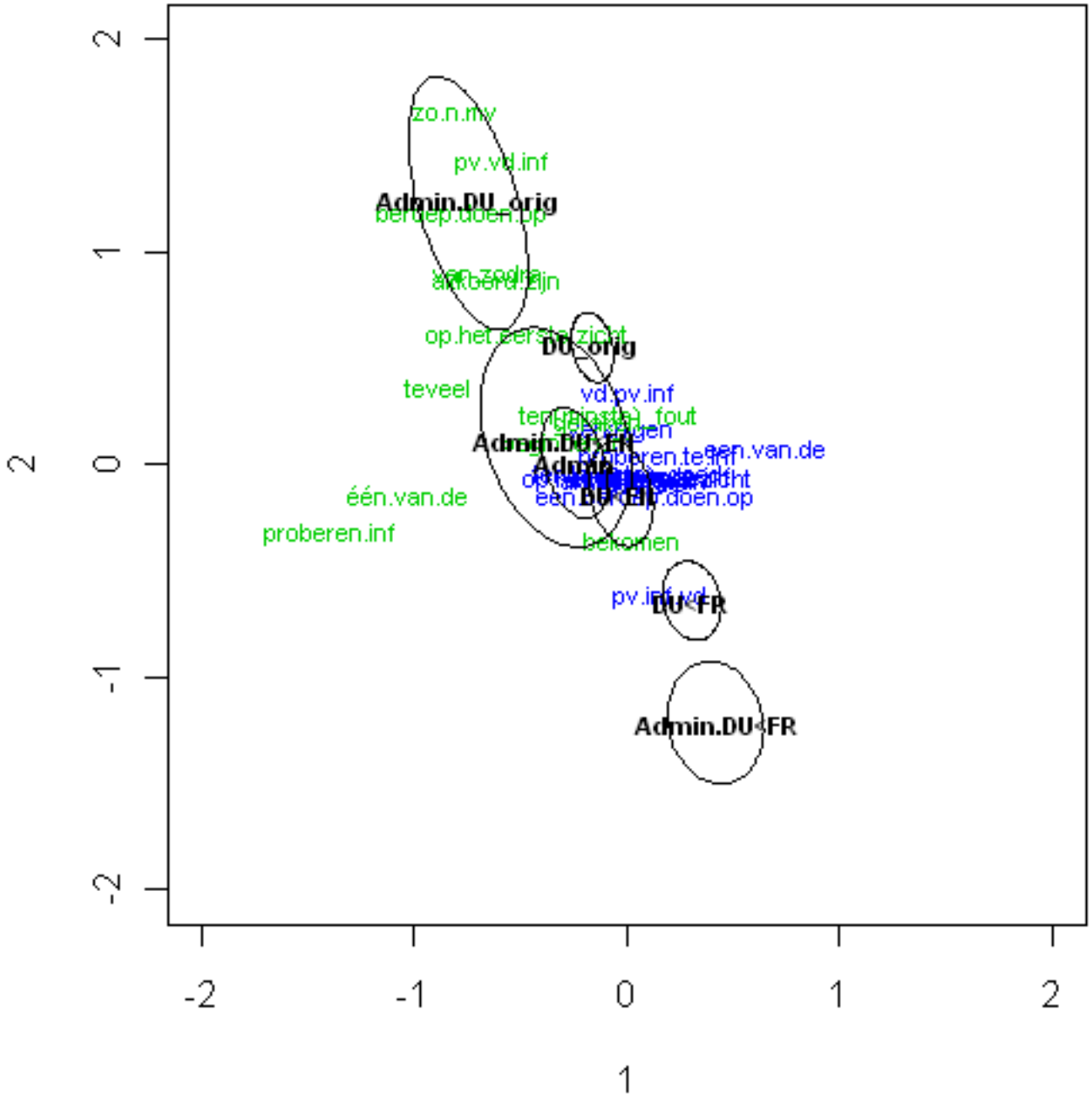
$\sigma^2_1 + \ldots + \sigma^2_{27} = X^2$

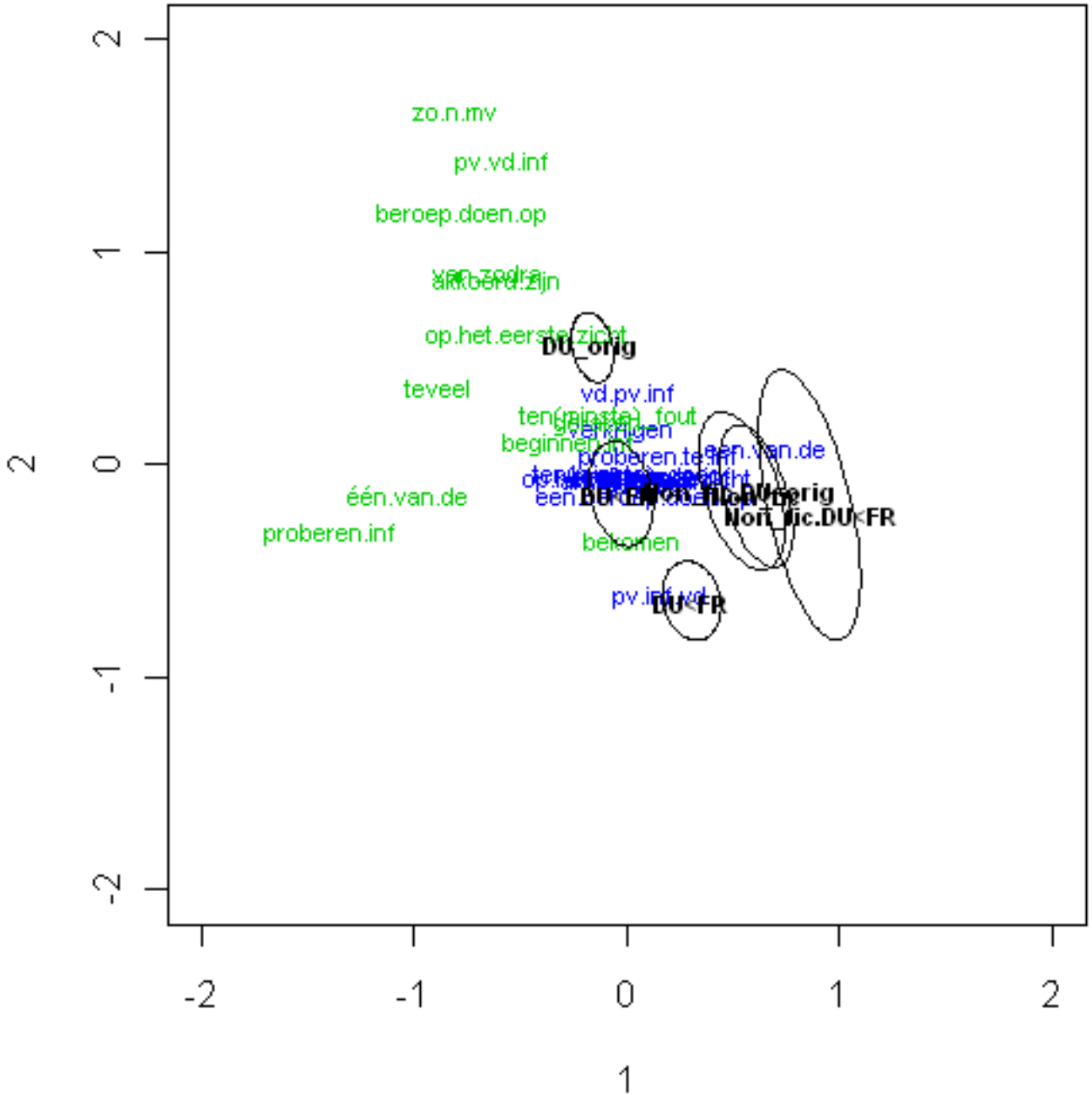

0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
14
17
20
23
26


0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
14
17
20
23
26


zo.n.mv
Instr
pv.vd.inf
beroep.doen.op
op.het.eerste.zicht
DU_orig
teveel
vd.pv.inf
beginnen.inf
Admin
Journal
Extern
één.van.de
Non_fic
proberen.inf
bekomen
Fiction
DU_FR
2
1
0
-1
-2
-2
-1
0
1
2
1


zo.n.mv
Instr
pv.vd.inf
beroep.doen.op
op.het.eerste.zicht
DU_orig
teveel
vd.pv.inf
beginnen.inf
Admin
Extern
één.van.de
proberen.inf
bekomen
pv.inf.vd
DU_FR
Fiction
2
0
-1
-2
-2
-1
0
1
2
1


zo.n.mv
pv.vd.inf
Admin.DU_orig
op.het.eerste.zicht
DU_orig
teveel
vd.pv.inf
Admin.DU<FR
een.van.de
één.van.de
proberen.inf
DU<FR
Admin.DU<FR
2
1
0
-1
-2
-2
-1
0
1
2

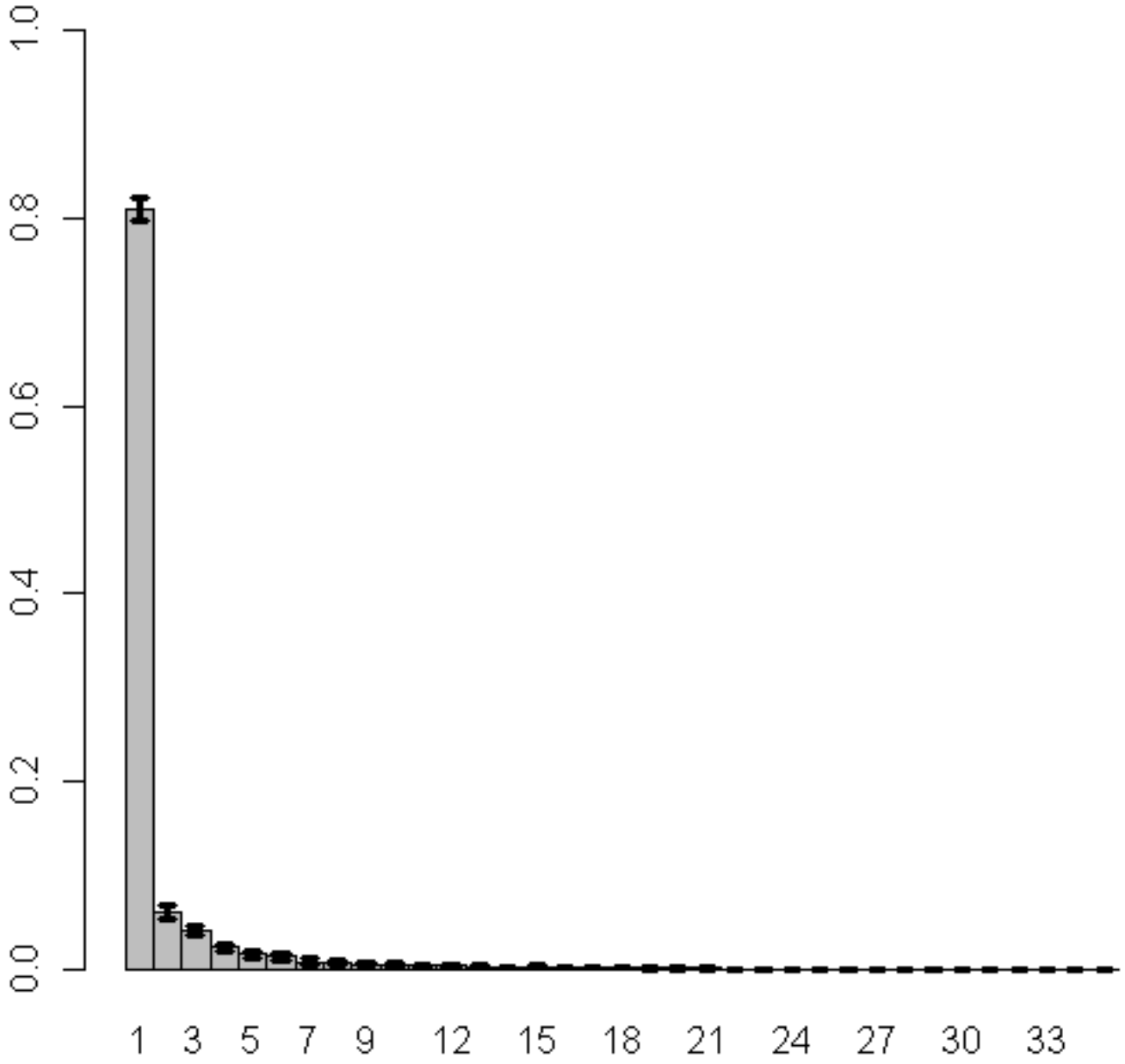

zo.n.mv
pv.vd.inf
beroep.doen.op
op.het.eerste.zicht
DU_orig
teveel
vd.pv.inf
fout
beginnen.inf
één.van.de
proberen.inf
Nort_dic.DU-FR
bekomen
DU-FR
2
1
0
-1
-2
-2
-1
0
1
2
1

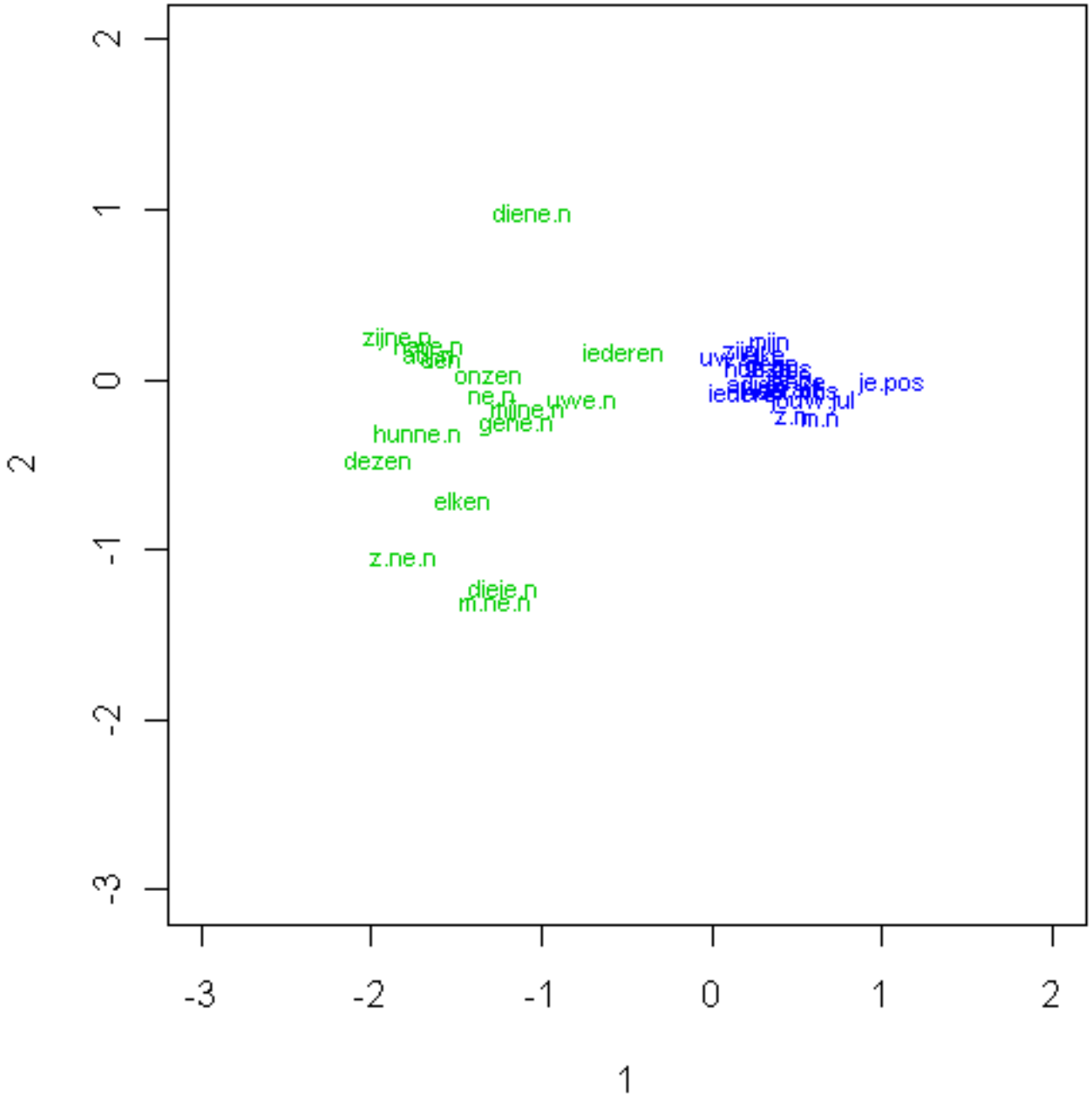
TSS

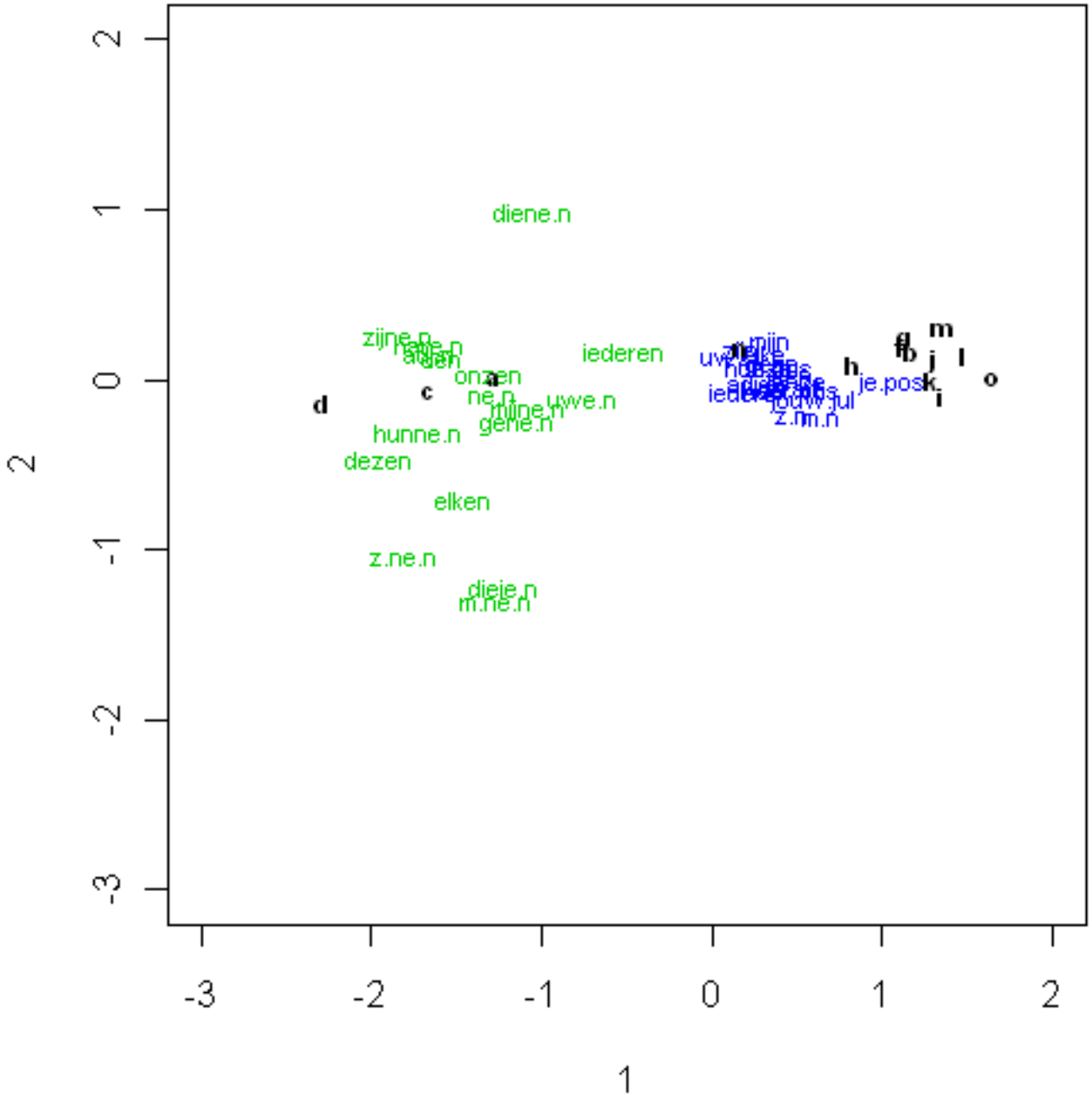
#	Variant 1	Variant 2	Variant 3	Variant 4	Translation
1	de	den			*the*
2	een	ne(n)			*a*
3	geen	gene(n)			*no*
4	elke	elken			*each*
5	iedere	iederen			*every*
6	deze	dezen			*this*
7	die	dieje(n)	diene(n)		*that*

TSS

#	Variant 1	Variant 2	Variant 3	Variant 4	Translation
8	mijn	m'n	mijne(n)	m'ne(n)	*my*
9	je	jouw/jullie	uw	uwe(n)	*your*
10	zijn	z'n	zijne(n)	z'ne(n)	*his*
11	haar	hare(n)			*her*
12	onze	onzen			*our*
13	hun	hunne(n)			*their*
14	ADJ-e	ADJ-en			*adjectival inflection*


0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
12
15
18
21
24
27
30
33


diene.n
zijne.n
iederen
onzen
ne.n
uwe.n
mijne.n
gene.n
hunne.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
je.pos
z.m.n
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
1

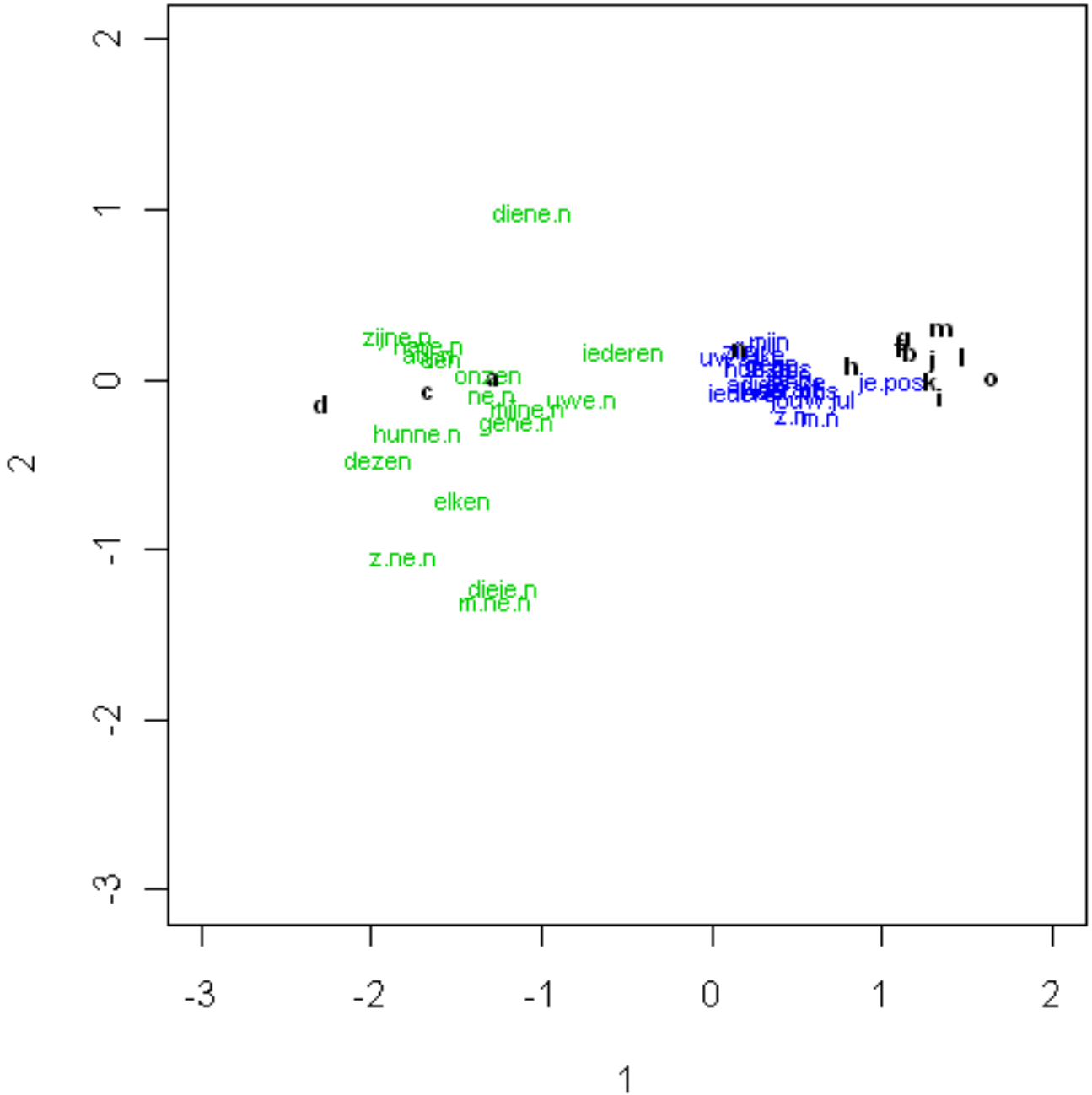

diene.n
zijne.n
iederen
onzen
ne.n
uwe.n
mijne.n
gene.n
hunne.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
je.posk
z.m.n
a
b
c
d
e
f
g
h
i
j
k
l
m
o
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
1
2

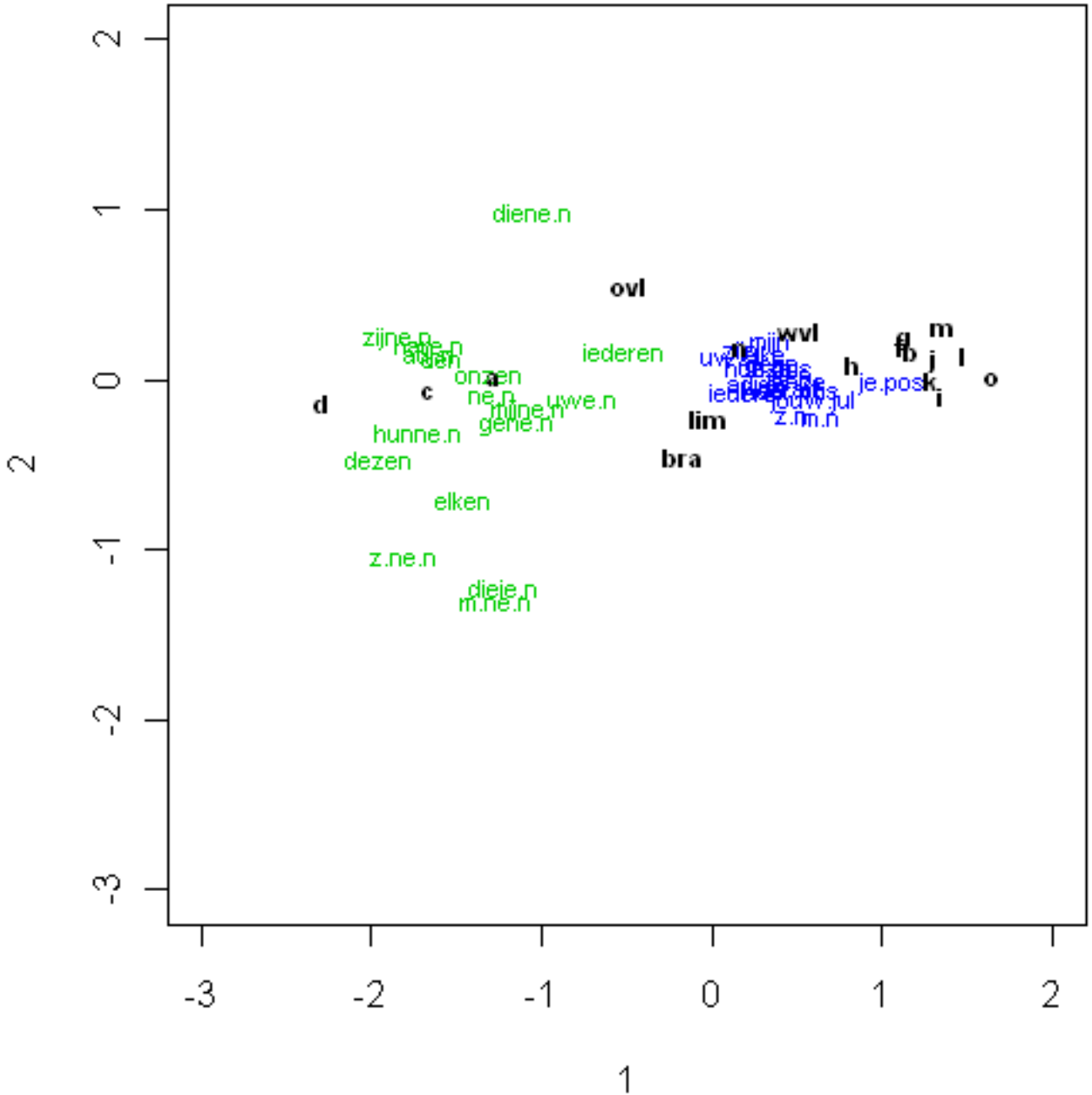
‘Components’

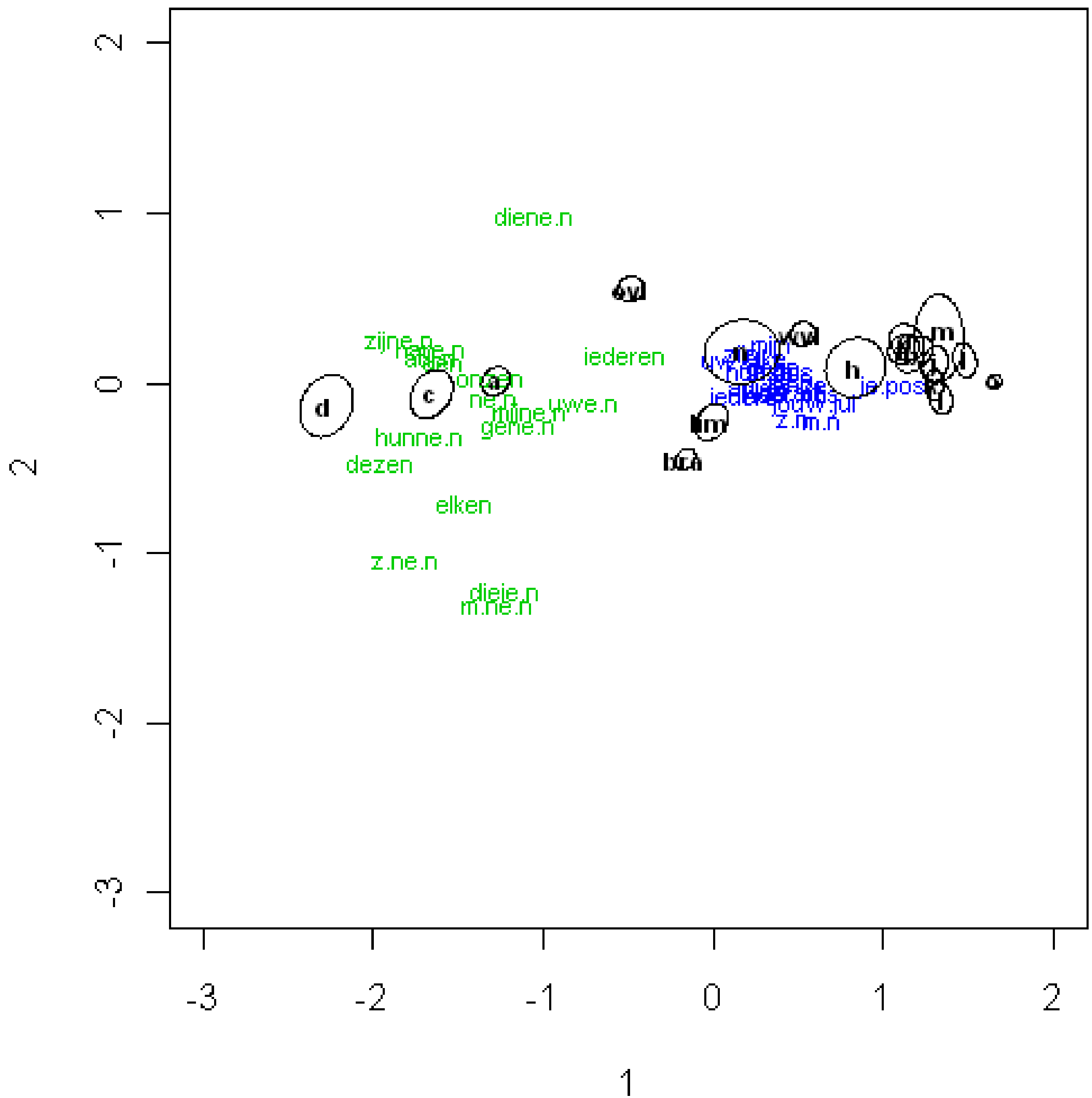
- a: Face-to-face conversations
- b: Interviews with teachers of Dutch
- c: Telephone dialogues (switchboard)
- d: Telephone dialogues (mini disc)
- f: Broadcast discussions/debates
- g: Non-broadcast discussions/debates
- h: Lessons recorded in classroom

‘Components’

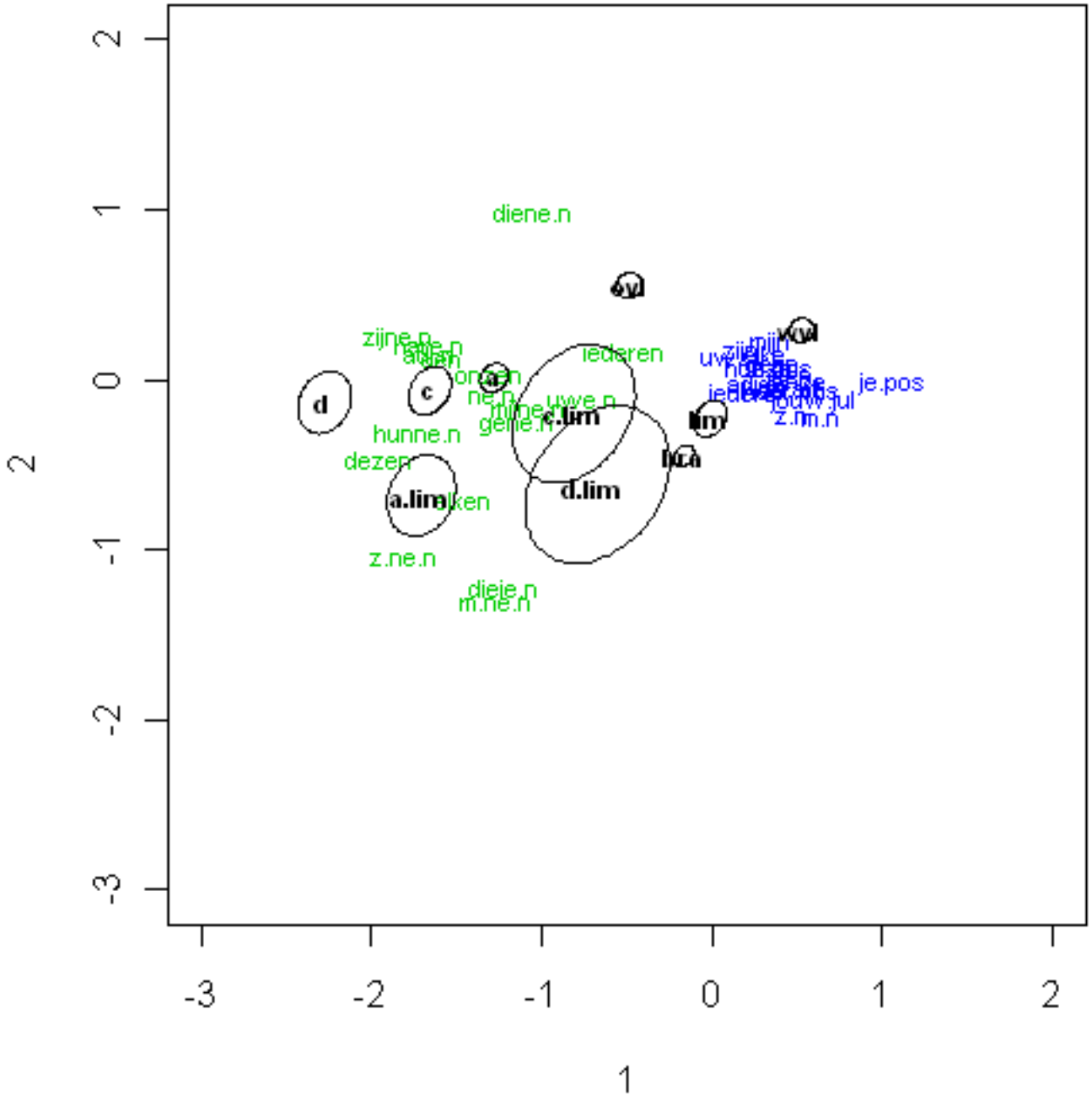
- i: Live commentaries (sports)
- j: News reports/reportages
- k: News bulletins
- l: Commentaries/columns/reviews
- m: Ceremonious speeches/sermons
- n: Lectures/seminars
- o: Read texts

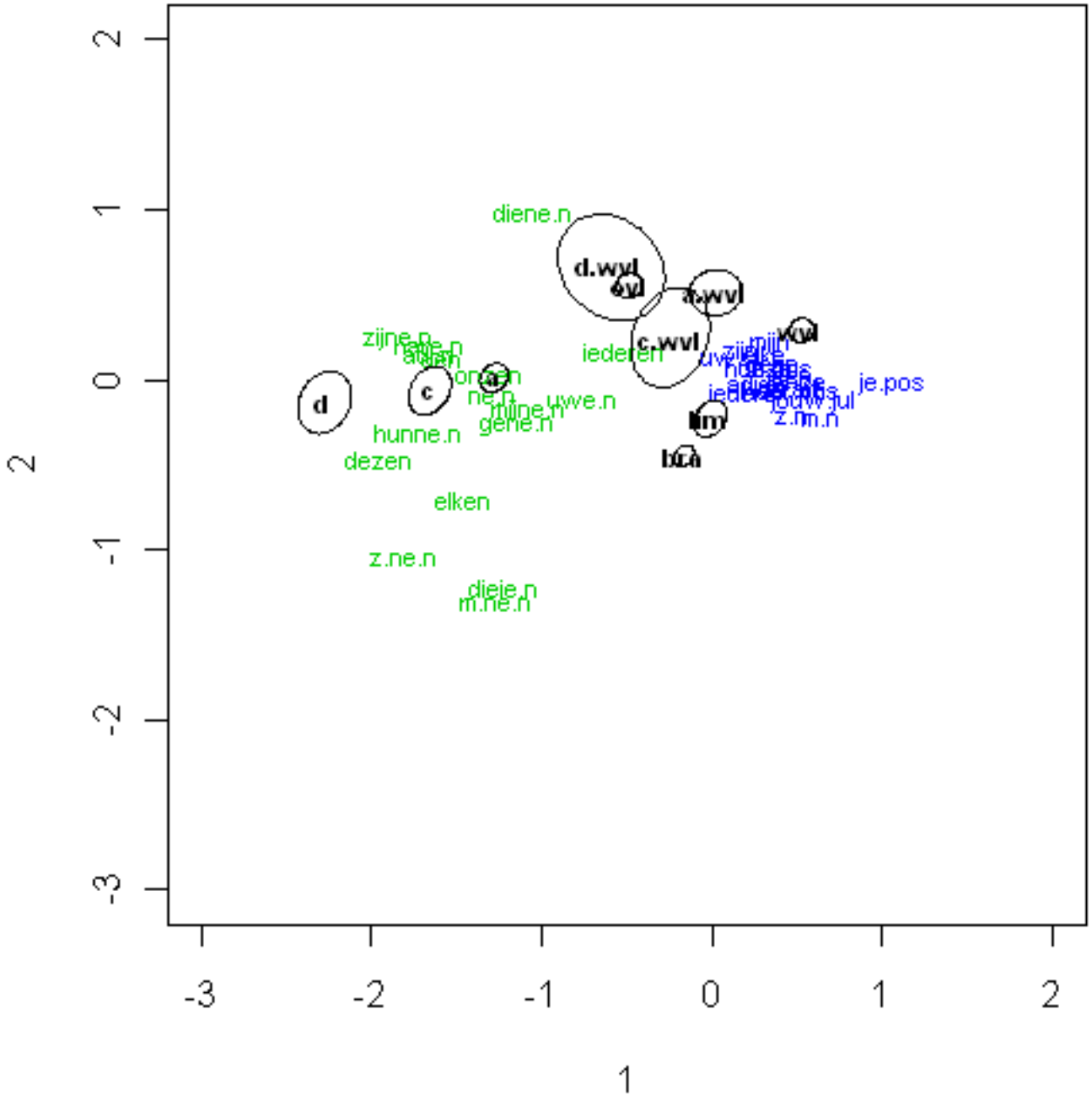

diene.n
zijne.n
iederen
onzen
ne.n
uwe.n
mijne.n
gene.n
hunne.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
je.posk
z.m.n
a
b
c
d
e
f
g
h
i
j
k
l
m
o
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
1
2

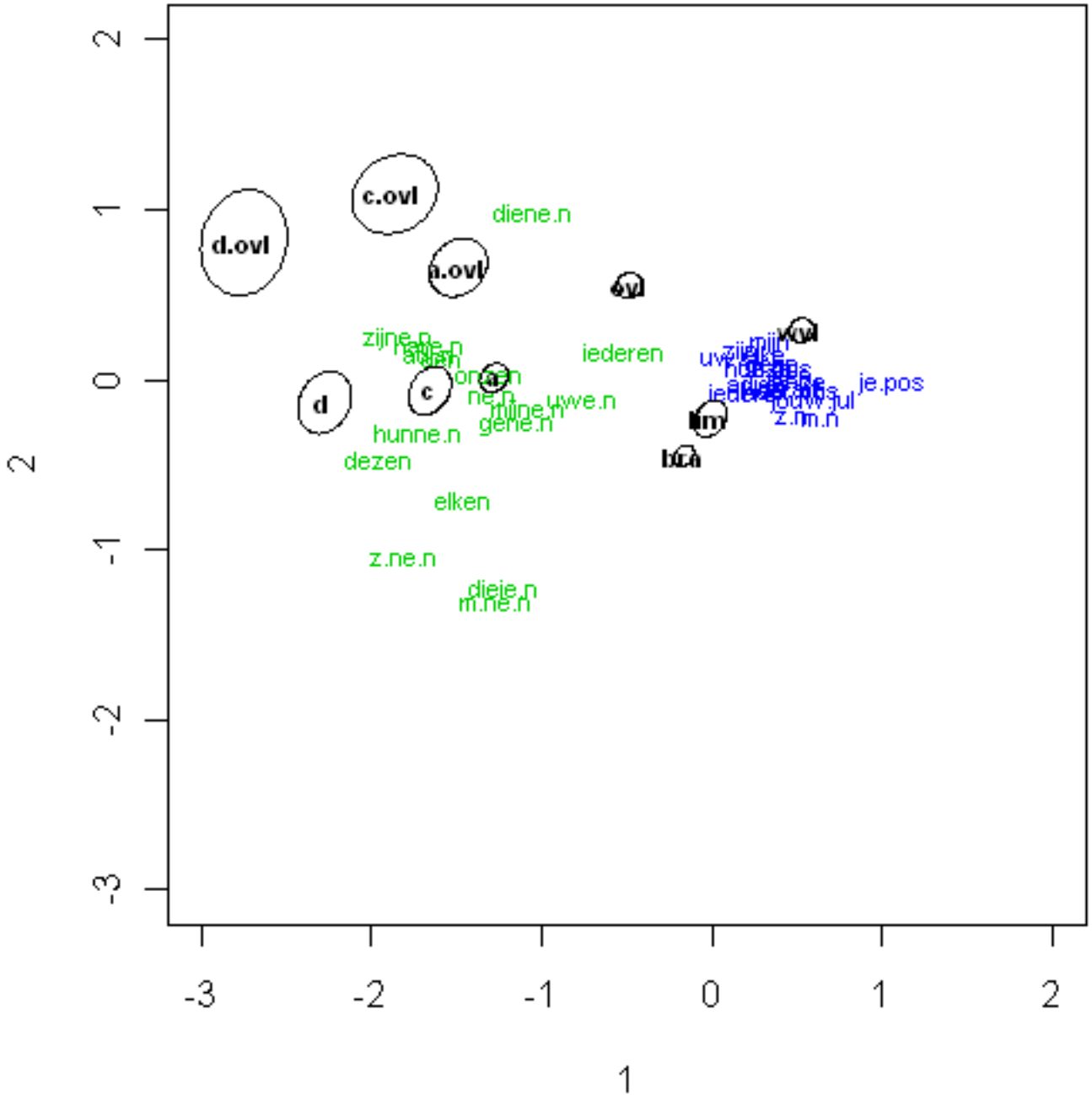

diene.n
ovl
zijne.n
iederen
wvl
m
l
o
onzen
c
d
uwe.n
lim
hunne.n
gene.n
dezen
bra
elken
z.ne.n
dieje.n
m.ne.n
-3
-2
-1
0
1
2
1
2

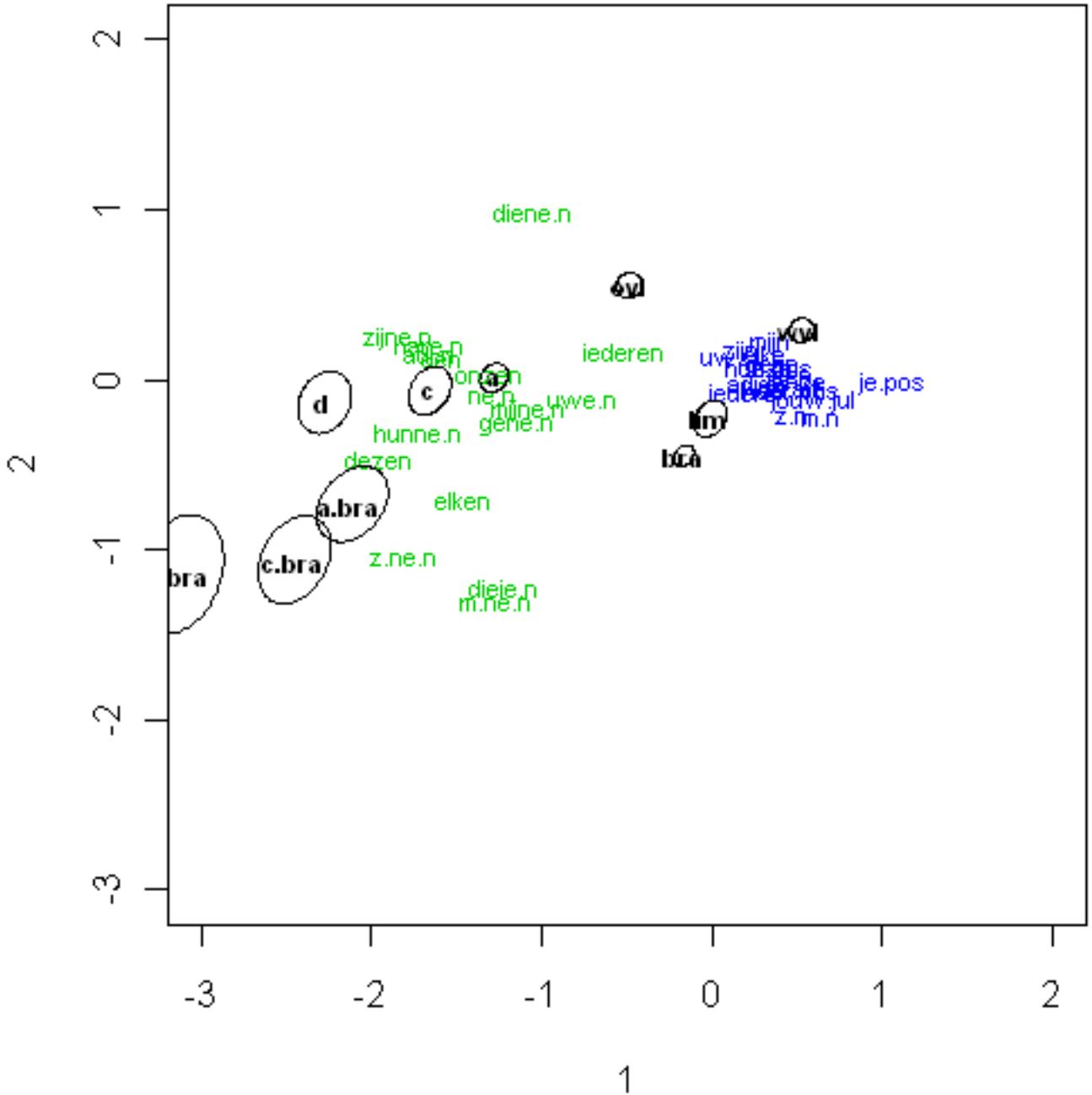


2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
1
2
diene.n
iederen
zijne.n
ne.n
uwe.n
gene.n
hunne.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
je.pos
d
c
a
h
m


diene.n
6vl
vvl
zijne.n
iedereen
uwe.n
je.pos
d
c
c.lim
lim
hunne.n
dezen
d.lim
a.lim
z.ne.n
dieje.n
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2


diene.n
d.wvl
a.wvl
c.wvl
wvl
zijne.n
iedereen
je.pos
d
c
uwe.n
hunne.n
gene.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
1


c.ovl
diene.n
d.ovl
a.ovl
6vl
vvl
zijne.n
iederen
je.pos
uwe.n
d
c
hunne.n
gene.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
z.m.n
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2


diene.n
6vl
vvl
zijne.n
iedereen
je.pos
uwe.n
hunne.n
gene.n
dezen
elken
z.ne.n
dieje.n
m.ne.n
d
c
a.bra
c.bra
bra
bca
z.m.n
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2

Conclusions

- (In)formality depends on context: different situations ask for different linguistic items
- (In)formality depends on power: dominant individuals can afford themselves more leniency

Thank you!

koen.plevoets@hogent.be