A typology of ‘doing typology’

Michael Cysouw
Max Plank Institute for Evolutionary Anthropology
Work-in-Progress, 1 February 2006

Statistics

- Confirmatory Statistics
  - ...-test
  - $p < \ldots$
- Exploratory Statistics
  - cluster analysis
  - multidimensional scaling
    - *etcetera* ...

Data

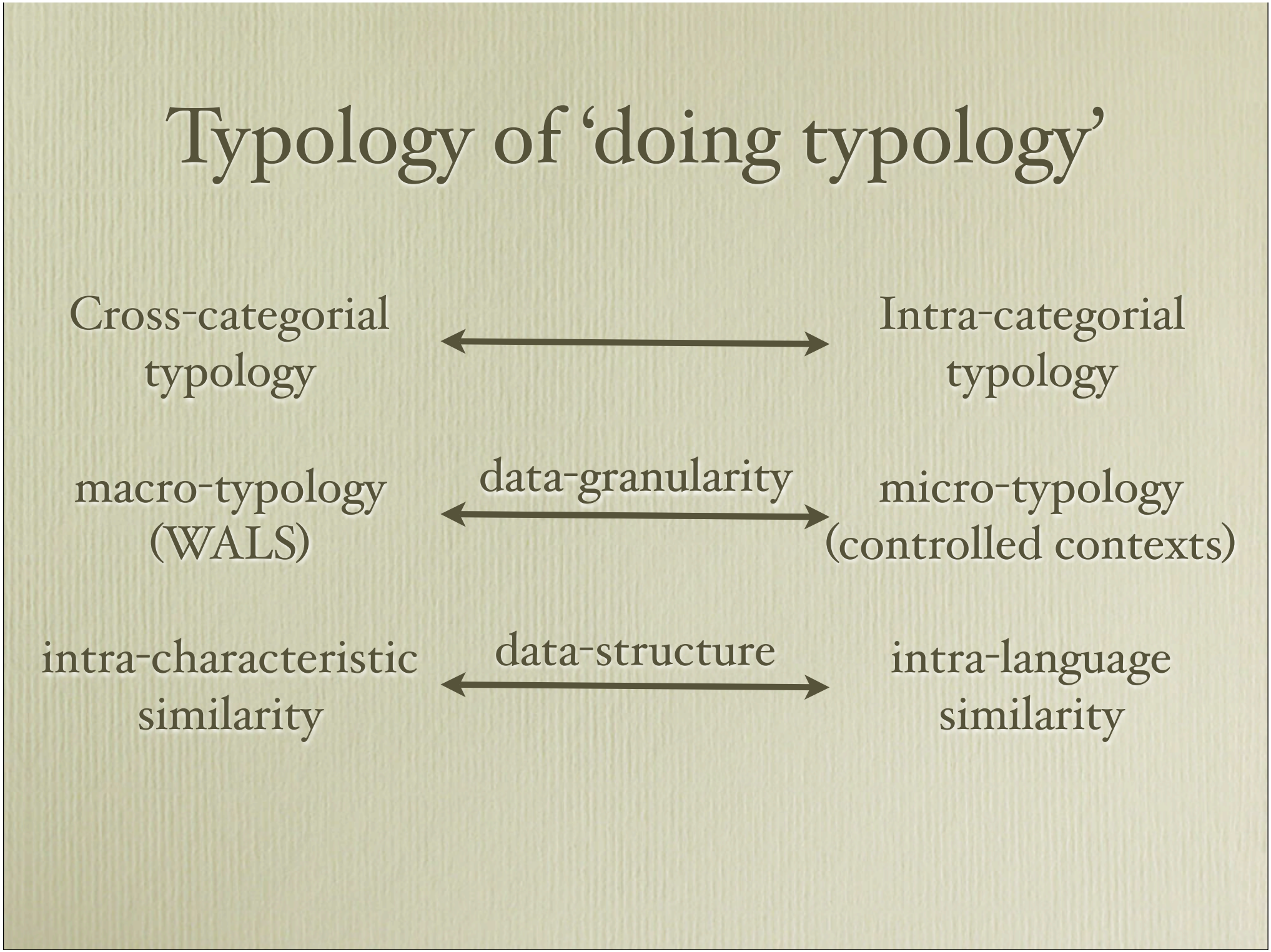
- Languages
  - languages
  - dialects/sociolect
  - idiolects
  - genre/register/text-specific ‘lects’
- Characteristics
  - macro-typologies (WALS)
  - fine-grained typologies (Autotyp)
  - questionnaires
  - controlled contexts (translations, stimulus reactions)

	L1	L2	L3	L4	English	L6	L7	Dutch	…
C1									
John 1:36								~*wandel*	
C3									
C4									
Velar Nasal					Not Initial				
C6									
C7									
C8									
…									

Data

- No independent/dependent variable distinction
- Discrete data (no linear similarity)

Typology of 'doing typology'

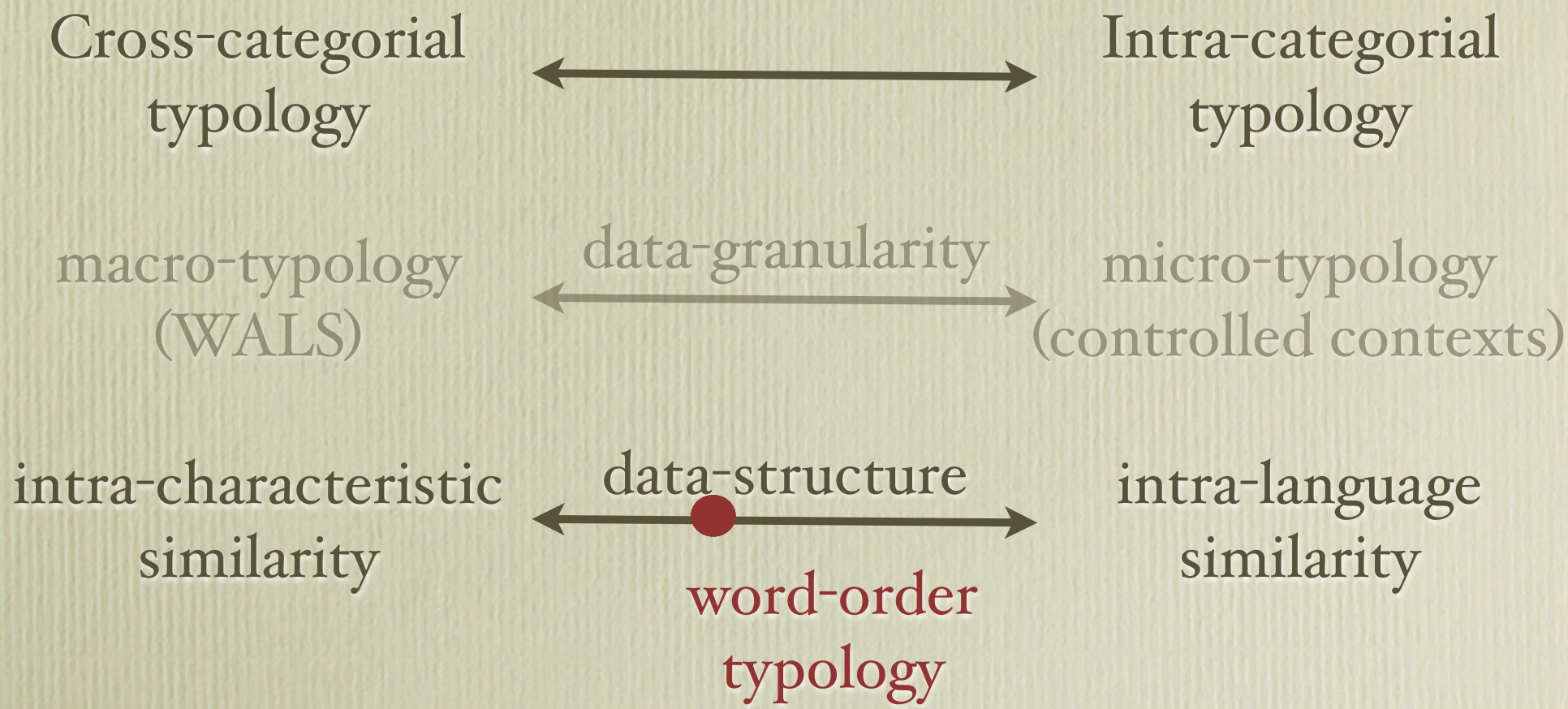


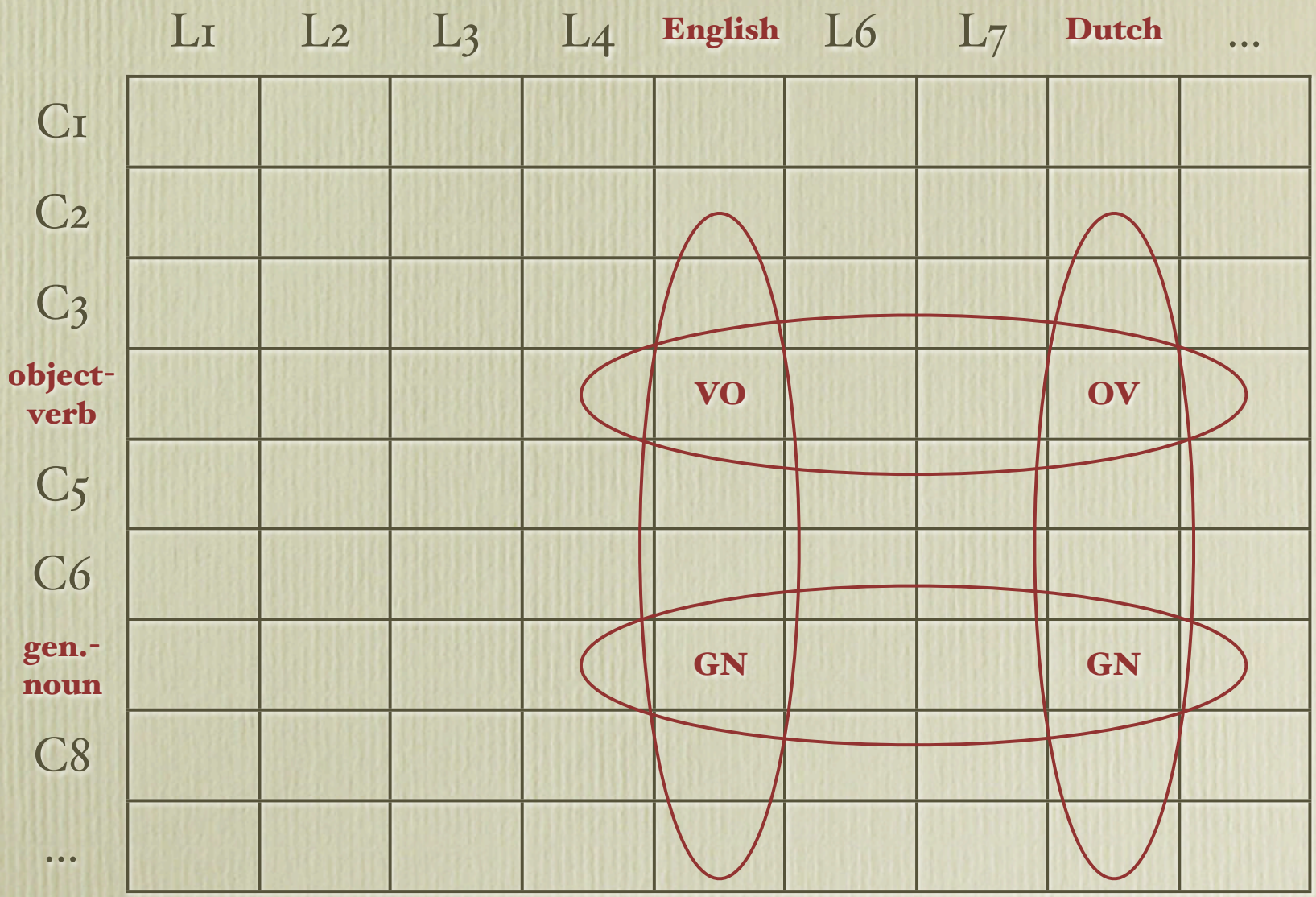
	L1	L2	L3	L4	English	L6	L7	Dutch	…
C1									
John 1:36					~walk			~wandel	
John 5:8					go			~wandel	
C4									
Velar Nasal					Not Initial			Not Initial	
polite-ness					No			Yes	
C7									
C8									
…									

	L1	L2	L3	L4	English	L6	L7	Dutch	…
C1									
John 1:36					*~walk*			*~wandel*	
John 5:8					*go*			*~wandel*	
C4									
Velar Nasal					Not Initial			Not Initial	
polite-ness					No			Yes	
C7									
C8									
…									

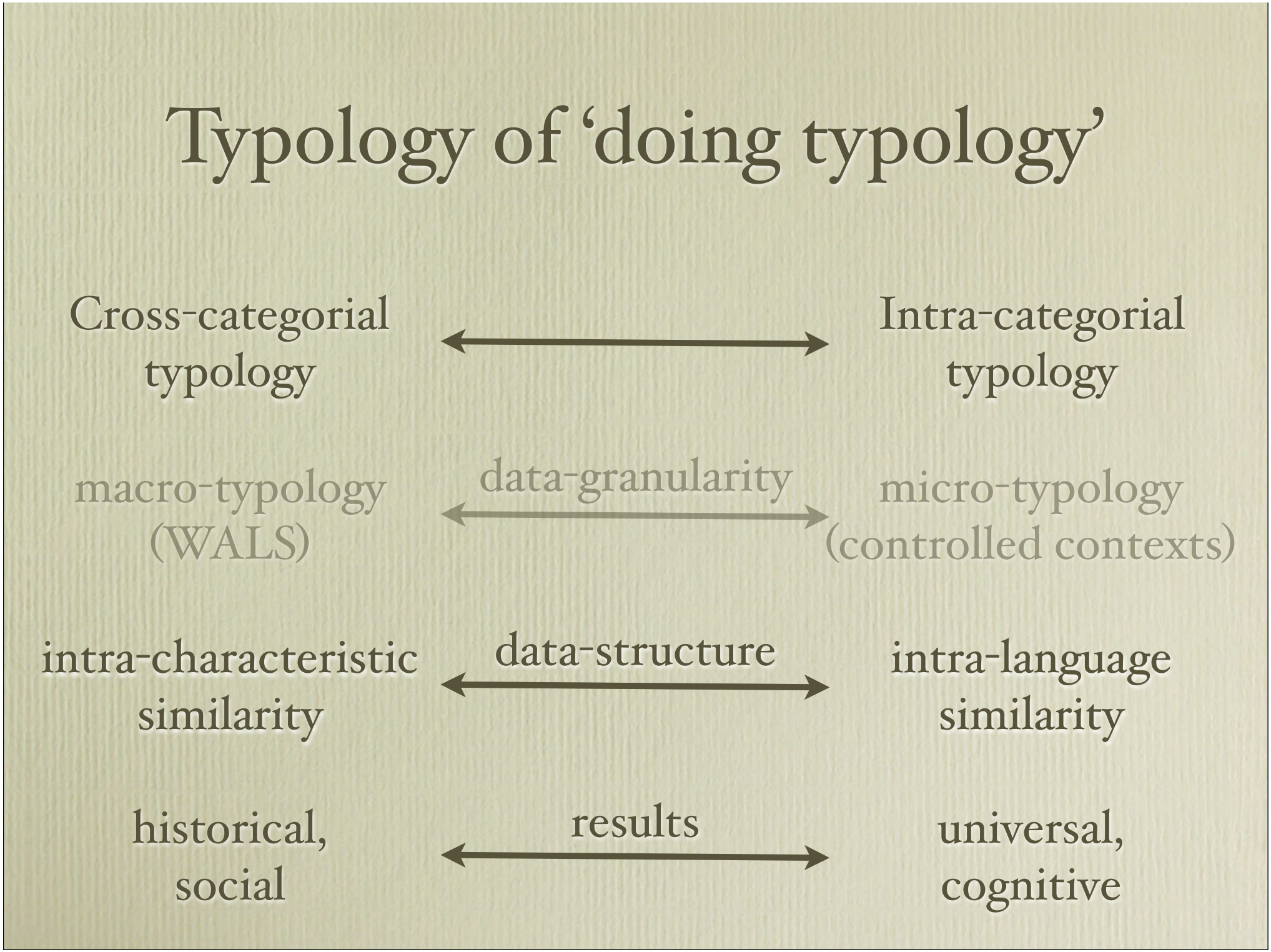
	L1	L2	L3	L4	English	L6	L7	Dutch	…
C1									
John 1:36					*walk*			*wandel*	
John 5:8					*go*			*wandel*	
C4									
Velar Nasal					Not Initial			Not Initial	
polite-ness					No			Yes	
C7									
C8									
…									

Typology of 'doing typology'




L1
L2
L3
L4
English
L6
L7
Dutch
...
C1
C2
C3
object-verb
C5
C6
gen.-noun
C8
...
VO
OV
GN
GN

Typology of 'doing typology'



Intra-Categorial Analysis

- Bernhard Wälchli's data
- 39 languages
- 335 clauses for each language
- clauses describing motion events in the bible
- only the lexical verb used

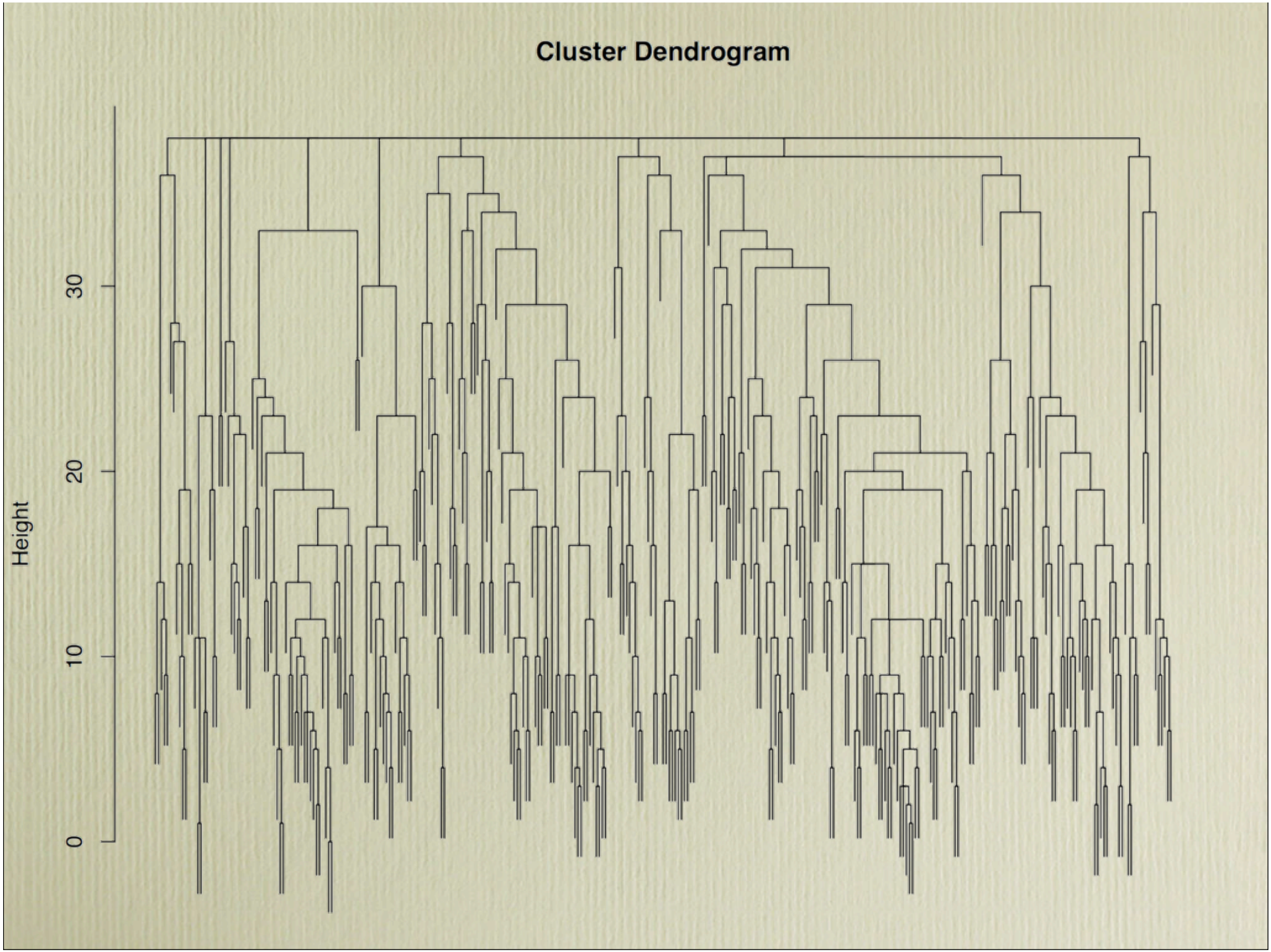
	MRD	LIT	ENG	FRE
1050	sams	eiti	go	aller
1070	sams	eiti	come	venir
1090	sams	eiti	come	venir
1104	lisems	kopti	come	sortir
1105	valgoms	zengti	descend	descendre
1114	NA	NA	come	se=faire=entendre
1120	vetjams	varyti	drive	pousser
1140	sams	eiti	come	se=rendre
1160	jutams	eiti	walk	marcher

	MRD	LIT	ENG	FRE
1050	**sams**	eiti	go	aller
1070	**sams**	eiti	**come**	venir
1090	**sams**	eiti	**come**	venir
1104	lisems	kopti	**come**	sortir
1105	valgoms	zengti	descend	descendre
1114	NA	NA	**come**	se=faire=entendre
1120	vetjams	varyti	drive	pousser
1140	**sams**	eiti	**come**	se=rendre
1160	jutams	eiti	walk	marcher

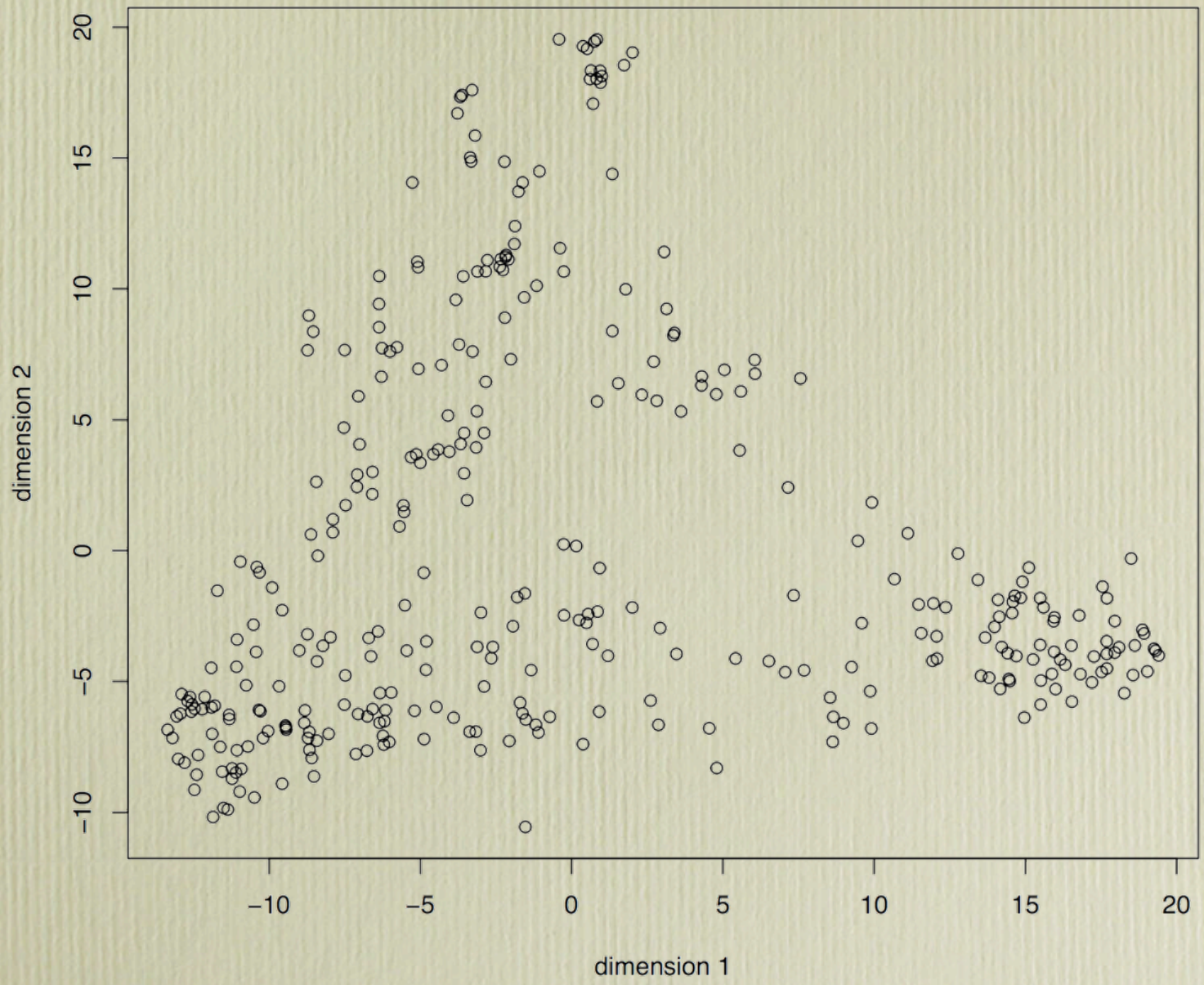
	MRD	LIT	ENG	FRE
1050	sams	eiti	go	aller
1070	sams	eiti	come	venir
1090	sams	eiti	**come**	venir
1104	lisems	kopti	**come**	sortir
1105	valgoms	zengti	descend	descendre
1114	NA	NA	come	se=faire=entendre
1120	vetjams	varyti	drive	pousser
1140	sams	eiti	come	se=rendre
1160	jutams	eiti	walk	marcher

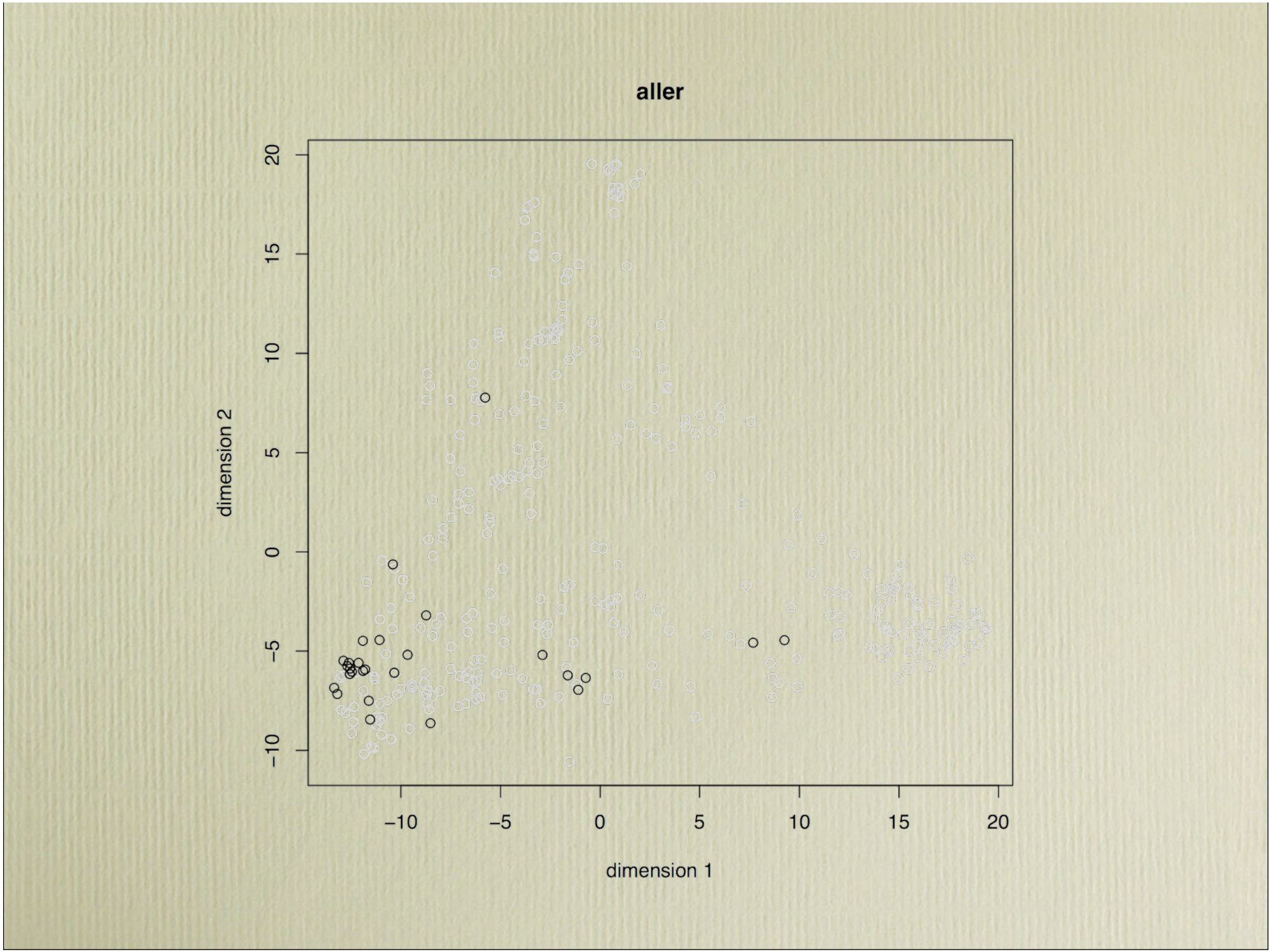
	MRD	LIT	ENG	FRE
1050	sams	eiti	go	aller
1070	sams	eiti	come	venir
1090	**sams**	**eiti**	**come**	venir
1104	lisems	kopti	come	sortir
1105	valgoms	zengti	descend	descendre
1114	NA	NA	come	se=faire=entendre
1120	vetjams	varyti	drive	pousser
1140	**sams**	**eiti**	**come**	se=rendre
1160	jutams	eiti	walk	marcher

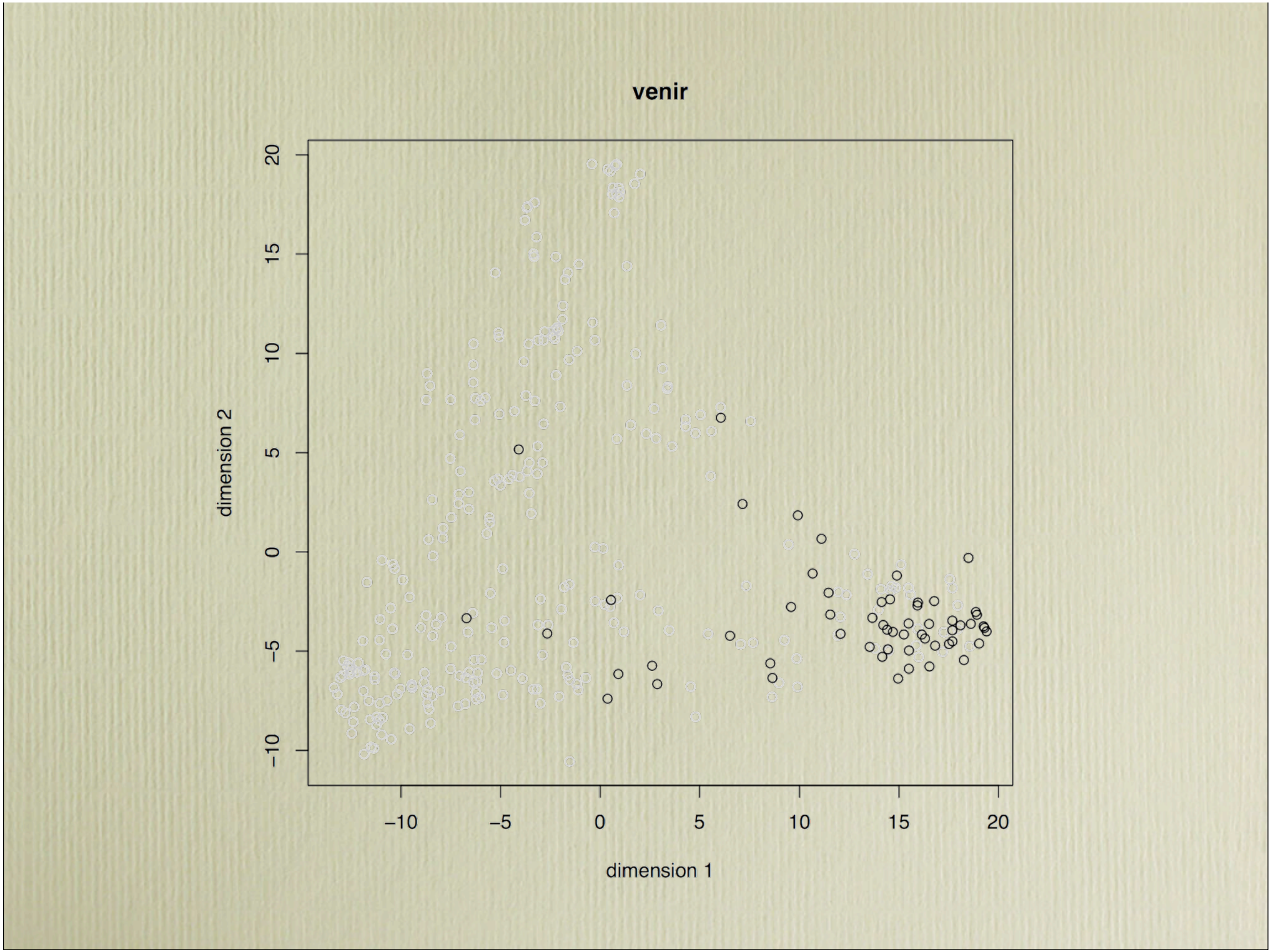
Hierarchical Clustering

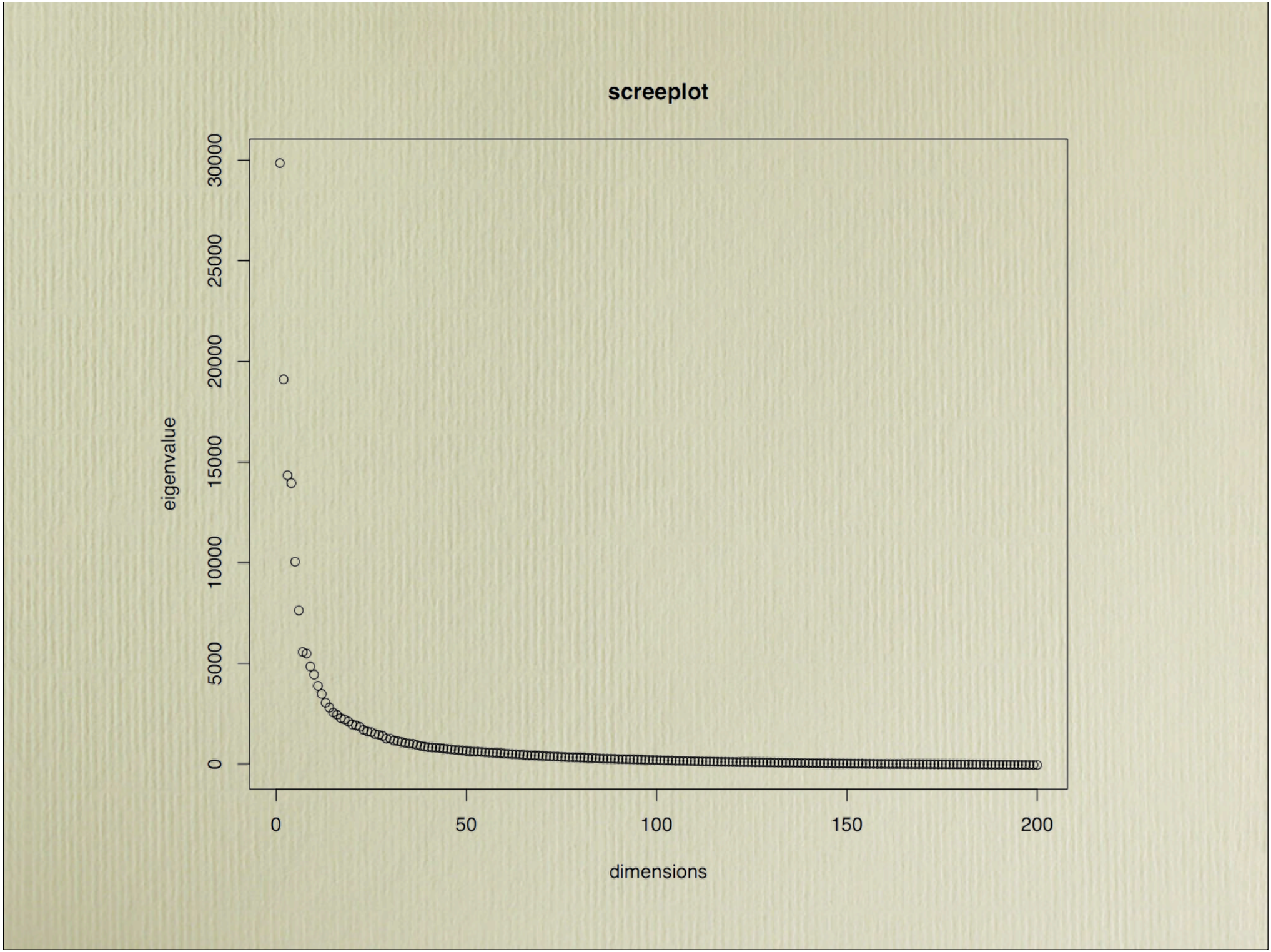

Cluster Dendrogram
Height
0
10
20
30

Multidimensional Scaling


dimension 2
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
dimension 1


aller
dimension 2
dimension 1
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20

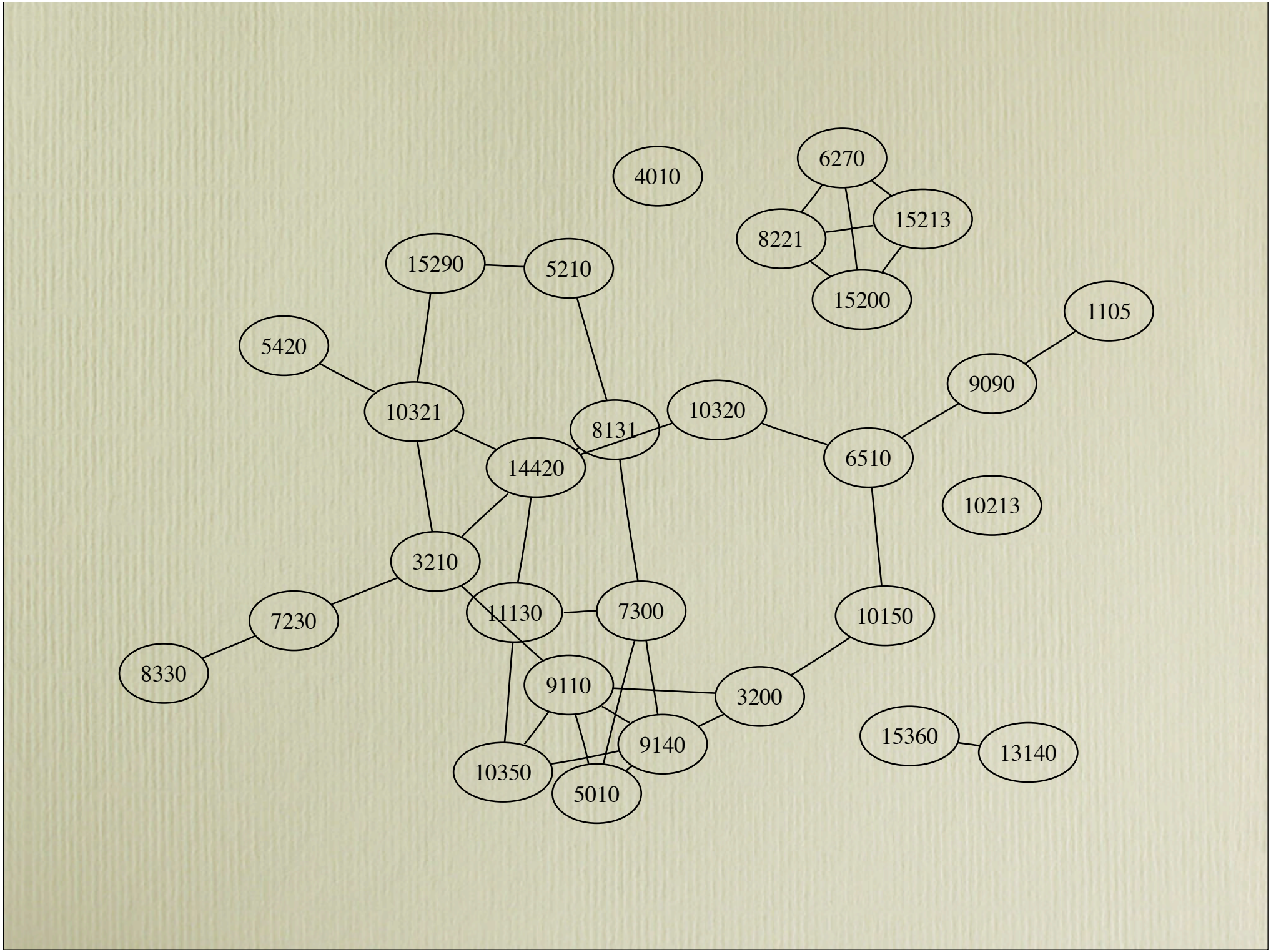

venir
dimension 2
dimension 1
−10
−5
0
5
10
15
20


screeplot
eigenvalue
0
5000
10000
15000
20000
25000
30000
0
50
100
150
200
dimensions

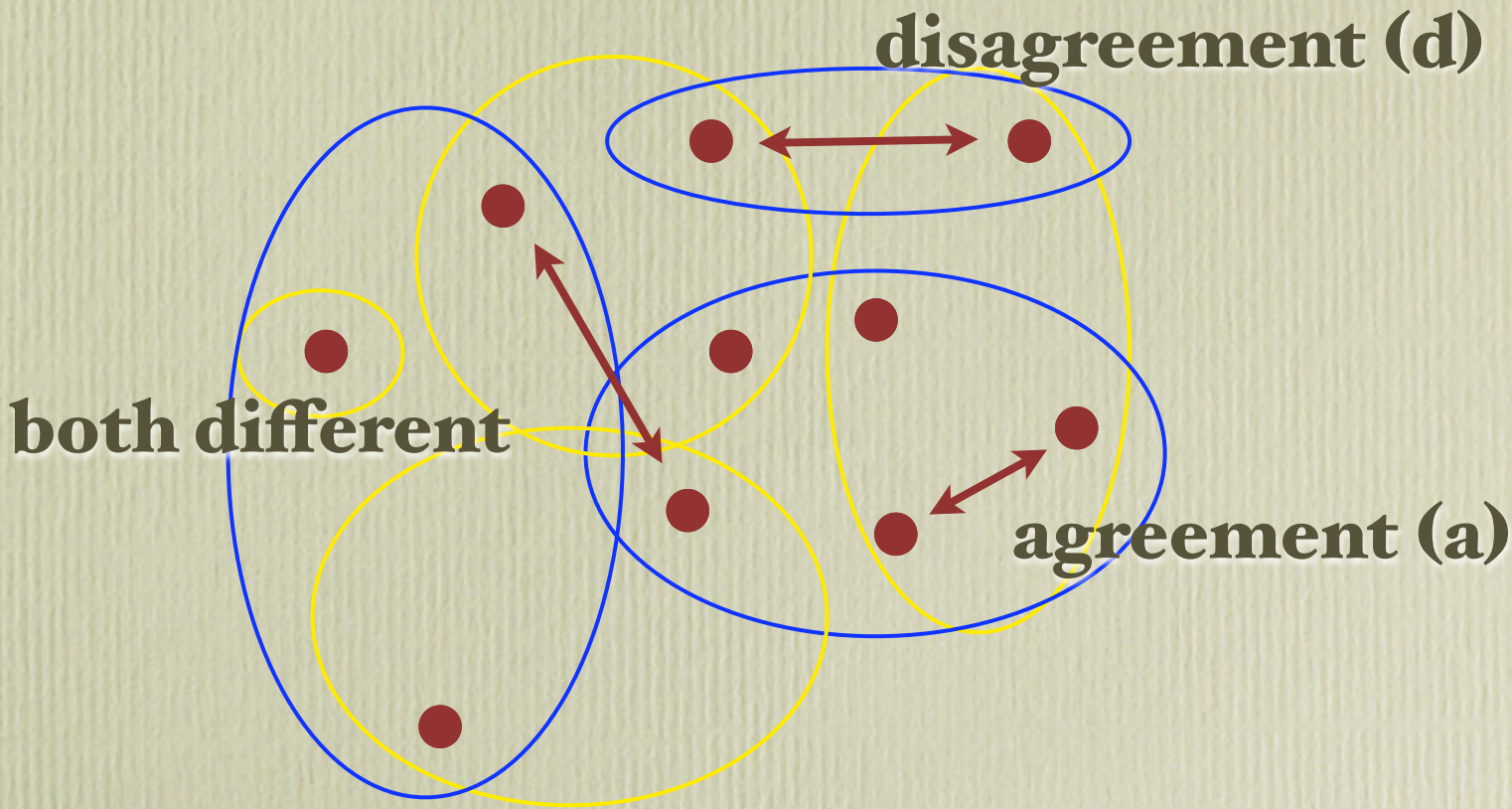
Semantic Map

Graph Condensation

- How to select which characteristics to display ?
- How to deal with equivalent graph structures ?
- Should the least amount of lines be preferred ?
- Should frequency play a role ?
- Are crossing lines allowed ?

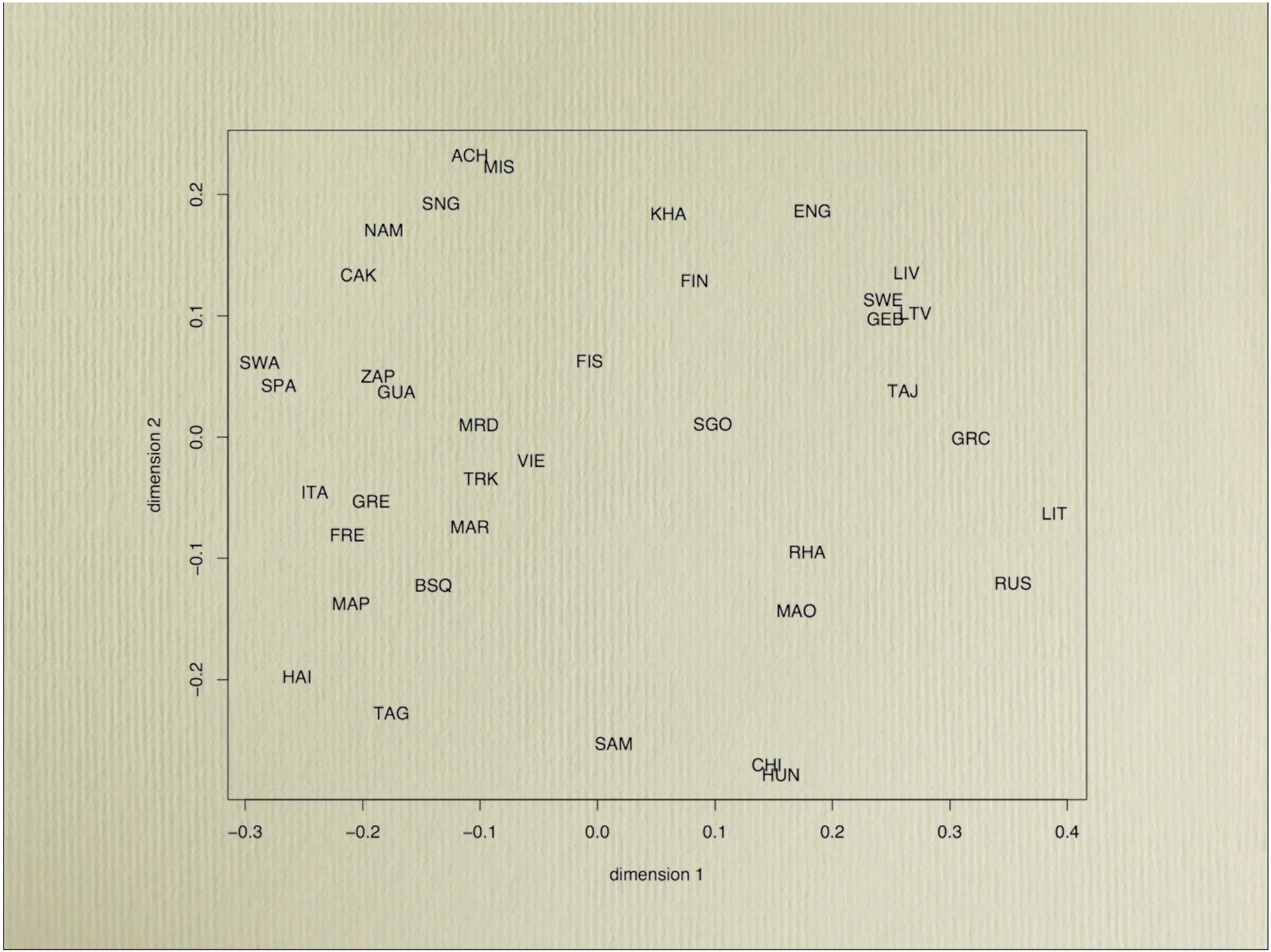

4010
6270
15213
8221
15200
15290
5210
1105
5420
9090
10321
10320
8131
6510
14420
10213
3210
11130
7300
10150
7230
8330
9110
3200
15360
13140
9140
10350
5010

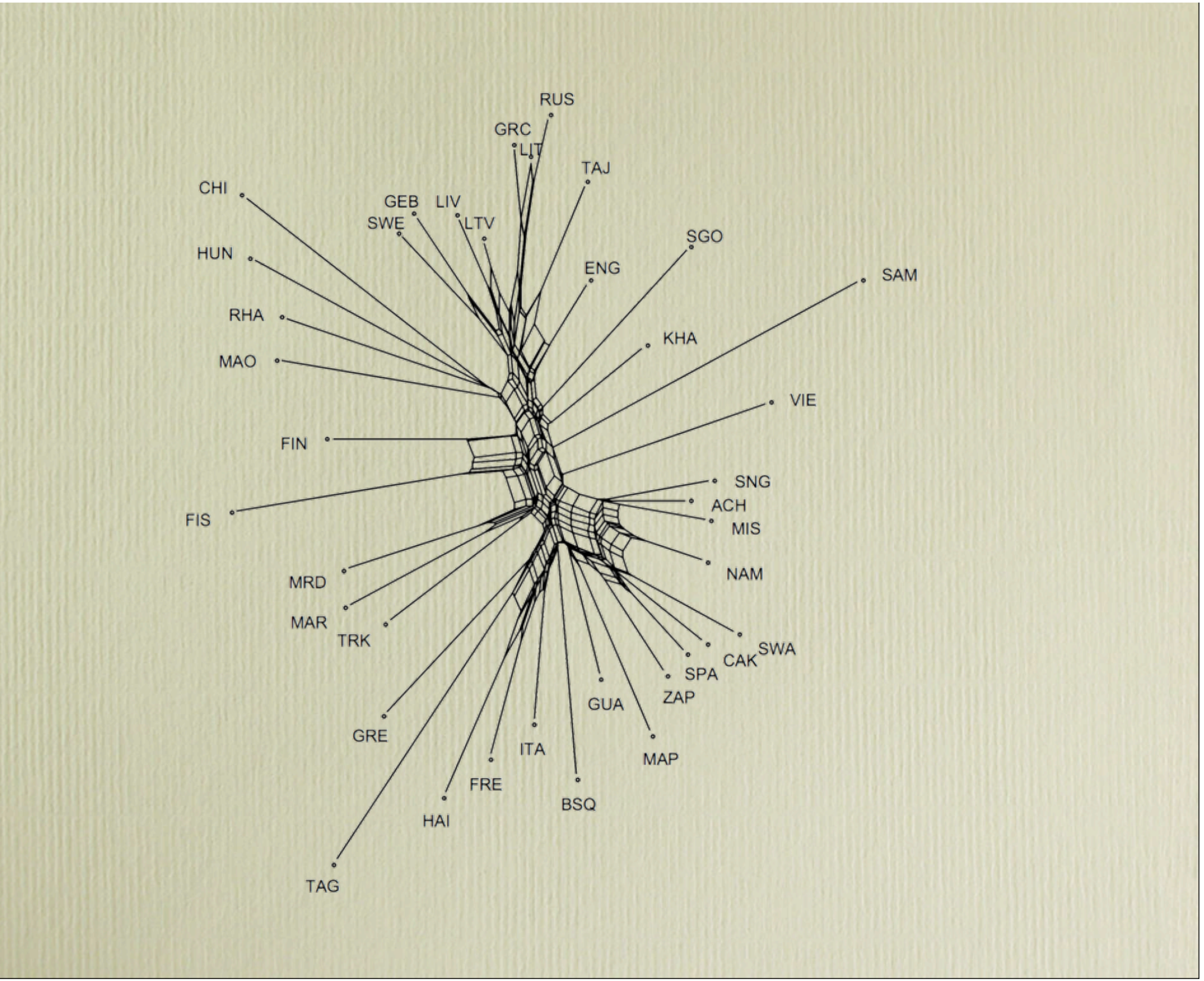
Language Similarity



Jaccard distance:

$d/_{a+d}$


ACH
MIS
SNG
KHA
ENG
NAM
CAK
FIN
LIV
SWE
LTV
GEB
SWA
FIS
SPA
ZAP
GUA
TAJ
MRD
SGO
GRC
VIE
TRK
ITA
GRE
LIT
MAR
FRE
RHA
BSQ
RUS
MAP
MAO
HAI
TAG
SAM
CHI
HUN
dimension 2
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
dimension 1


RUS
GRC
LIT
TAJ
CHI
GEB
LIV
SWE
LTV
SGO
HUN
ENG
SAM
RHA
KHA
MAO
VIE
FIN
SNG
ACH
FIS
MIS
NAM
MRD
MAR
TRK
SWA
CAK
SPA
ZAP
GUA
GRE
ITA
MAP
FRE
BSQ
HAI
TAG

Inter-Categorial Analysis

- Data from WALS
- hundreds of languages
- 140 characteristics
- notion of similarity between languages given

Similarity between Characteristics

Pairwise interaction

	No Politeness	Yes Politeness
No Velar Nasal	57	21
Yes Velar Nasal	51	12

Comparing WALS-maps

- Correlation tests: is there an interaction ? (chi-square, Fisher's exact, Dryer's test)
- Not much … (velar nasal/politeness: p = 0.088)
- Maybe the type-distinctions do not match ?
- Explore interactions, not test them !

Highly significant interaction

	no we	I=we	we	I=exclusive	inclusiv+ exclusive
no we	1	5	36	1	27
I=we	1	1	9	0	1
we	0	2	75	0	2
I=exclusive	0	0	0	4	5
inclusiv+ exclusive	0	2	0	0	28

Most prominent combinations

	no we	I=we	we	I=exclusive	inclusiv+ exclusive
no we	1	5	36	1	27
I=we	1	1	9	0	1
we	0	2	75	0	2
I=exclusive	0	0	0	4	5
inclusiv+ exclusive	0	2	0	0	28

Higher than expected

	no we	I=we	we	I=exclusive	inclusiv+ exclusive
no we	1	5	36	1	27
I=we	1	1	9	0	1
we	0	2	75	0	2
I=exclusive	0	0	0	4	5
inclusiv+ exclusive	0	2	0	0	28

This is not the same !

Absolute

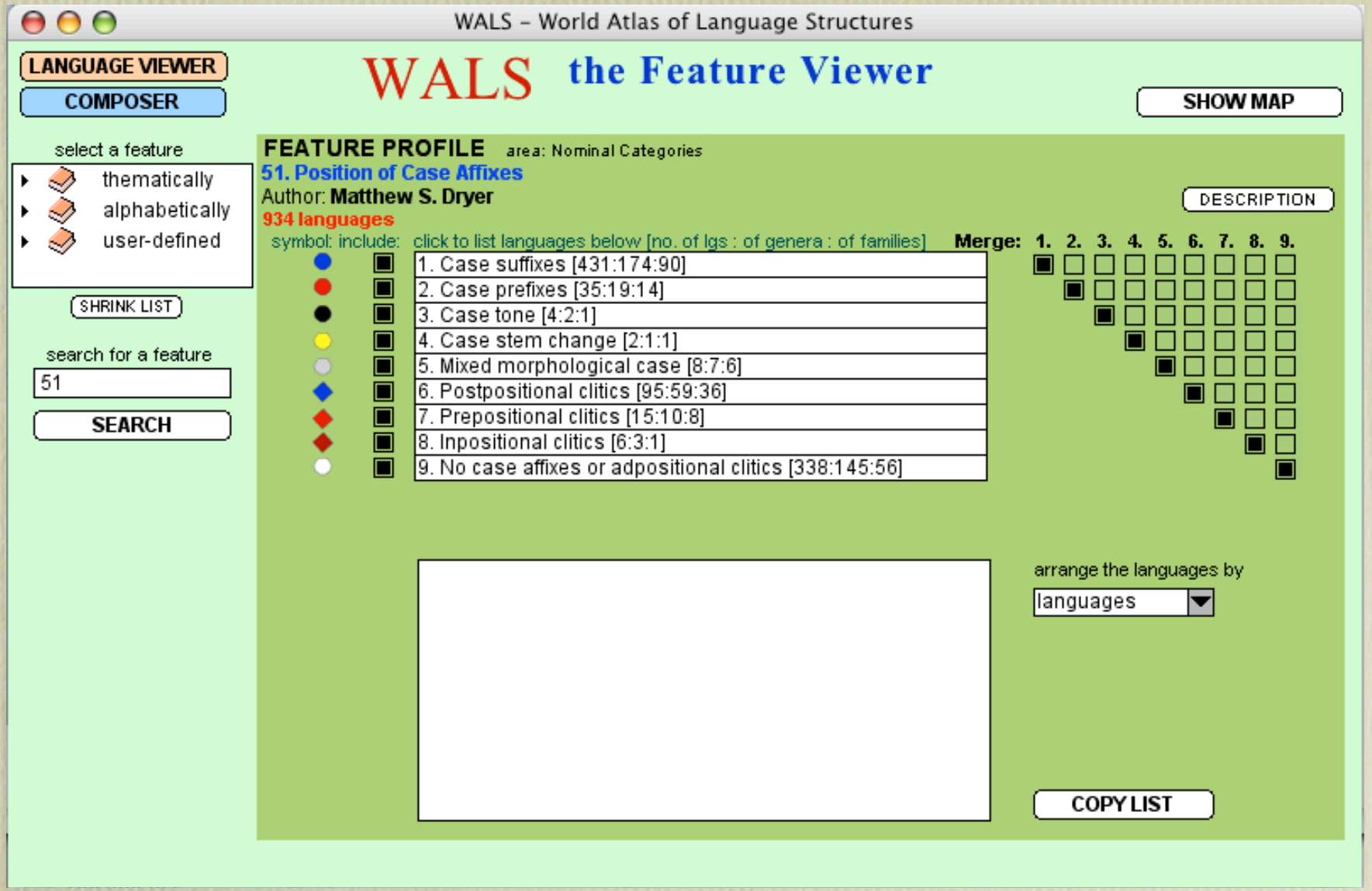
	no we	I=we	we	I=exclusive	inclusiv+
no we	1	5	36	1	27
I=we	1	1	9	0	1
we	0	2	75	0	2
I=exclusive	0	0	0	4	5
inclusiv+	0	2	0	0	28

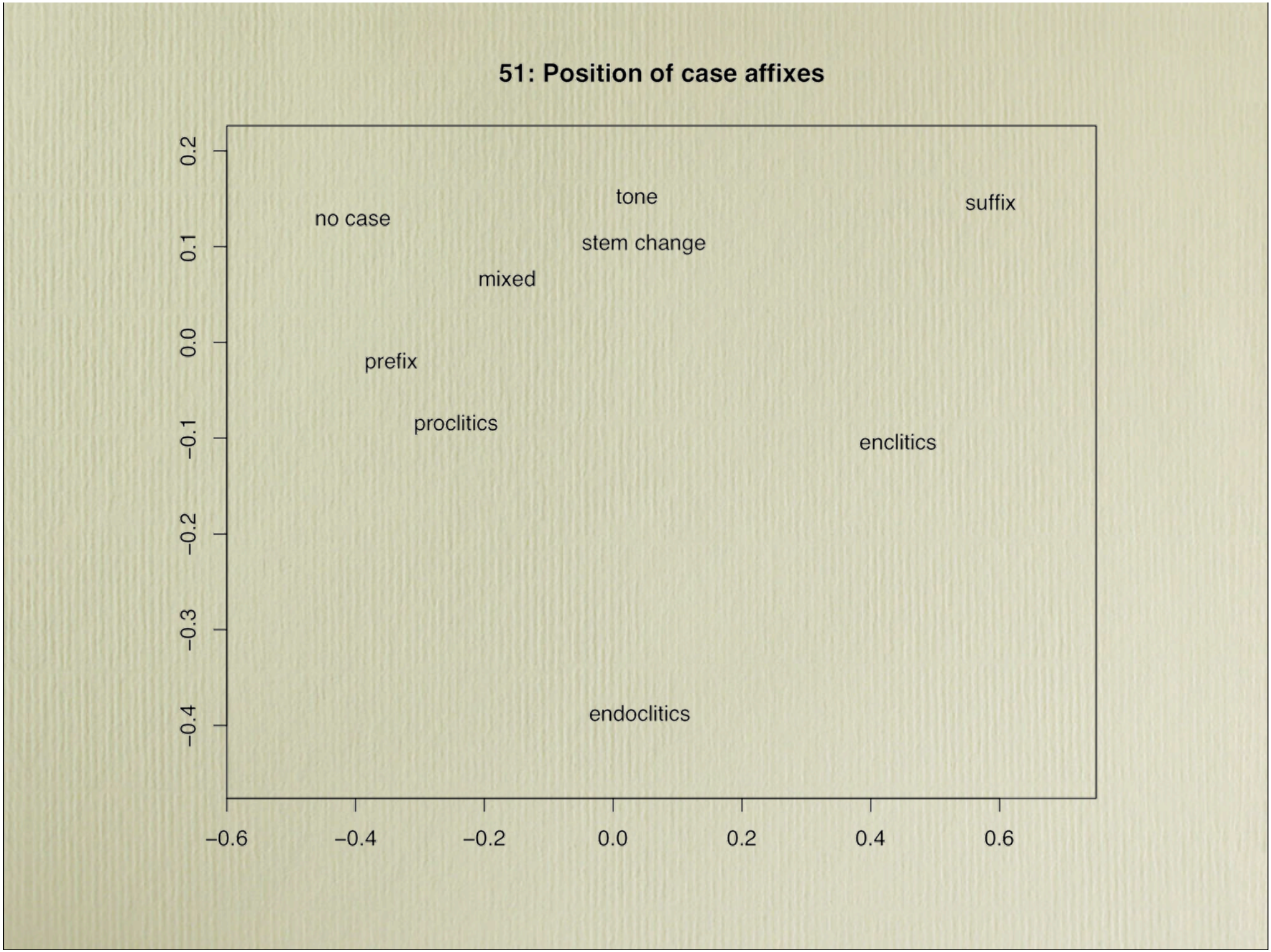
Relative

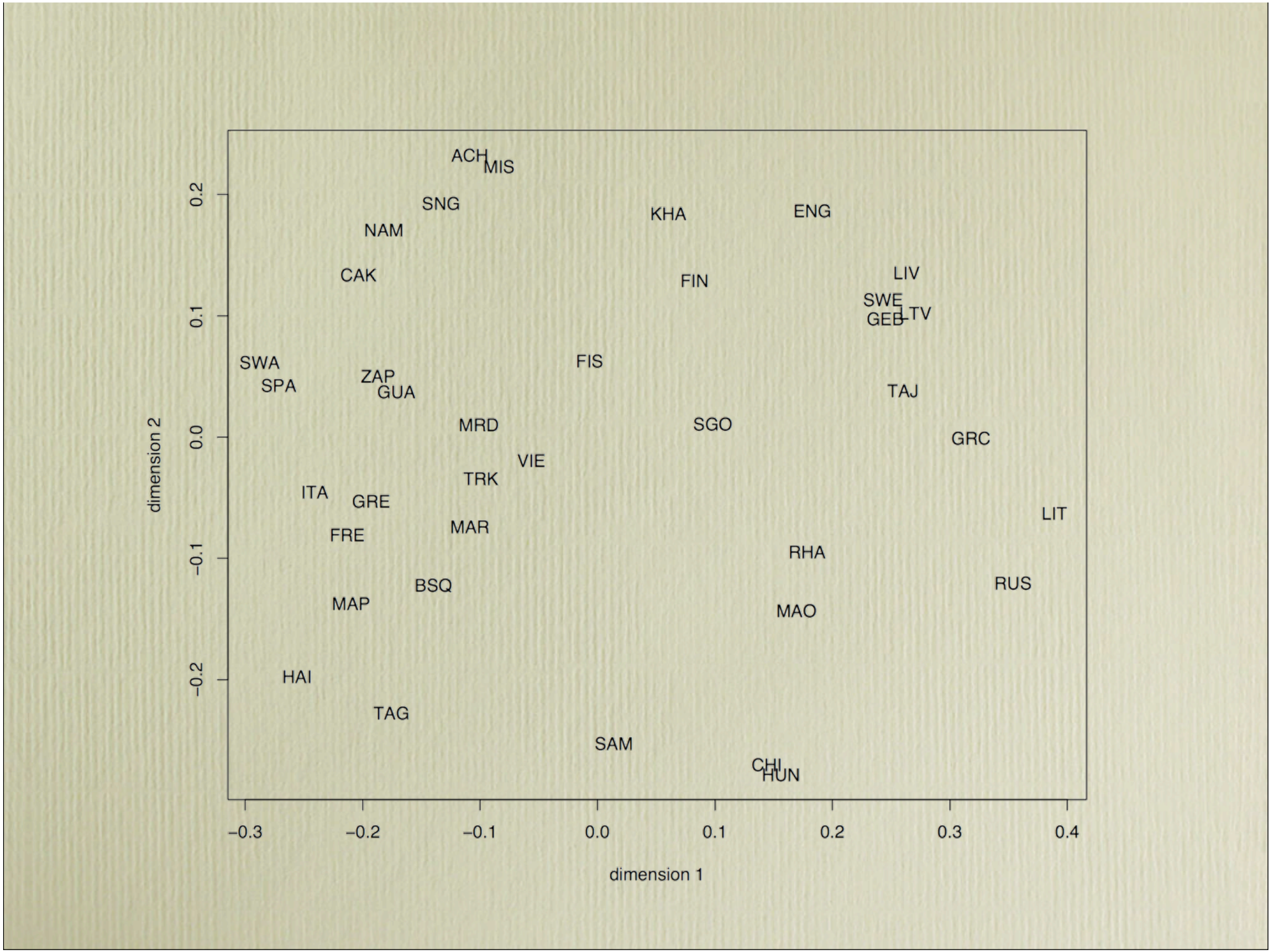
	no we	I=we	we	I=exclusive	inclusiv+
no we	1	5	36	1	27
I=we	1	1	9	0	1
we	0	2	75	0	2
I=exclusive	0	0	0	4	5
inclusiv+	0	2	0	0	28

‘Grain’ of Typology

- How many types to distinguish ?
- Make more fine-grained typologies !
- How to deal with typologies that distinguish very many types ?


WALS – World Atlas of Language Structures
LANGUAGE VIEWER
COMPOSER
WALS the Feature Viewer
SHOW MAP
select a feature
thematically
alphabetically
user-defined
SHRINK LIST
search for a feature
51
SEARCH
FEATURE PROFILE area: Nominal Categories
51. Position of Case Affixes
Author: Matthew S. Dryer
934 languages
DESCRIPTION
symbol: include: click to list languages below [no. of lgs : of genera : of families]
Merge: 1. 2. 3. 4. 5. 6. 7. 8. 9.
1. Case suffixes [431:174:90]
2. Case prefixes [35:19:14]
3. Case tone [4:2:1]
4. Case stem change [2:1:1]
5. Mixed morphological case [8:7:6]
6. Postpositional clitics [95:59:36]
7. Prepositional clitics [15:10:8]
8. Inpositional clitics [6:3:1]
9. No case affixes or adpositional clitics [338:145:56]
arrange the languages by
languages
COPY LIST

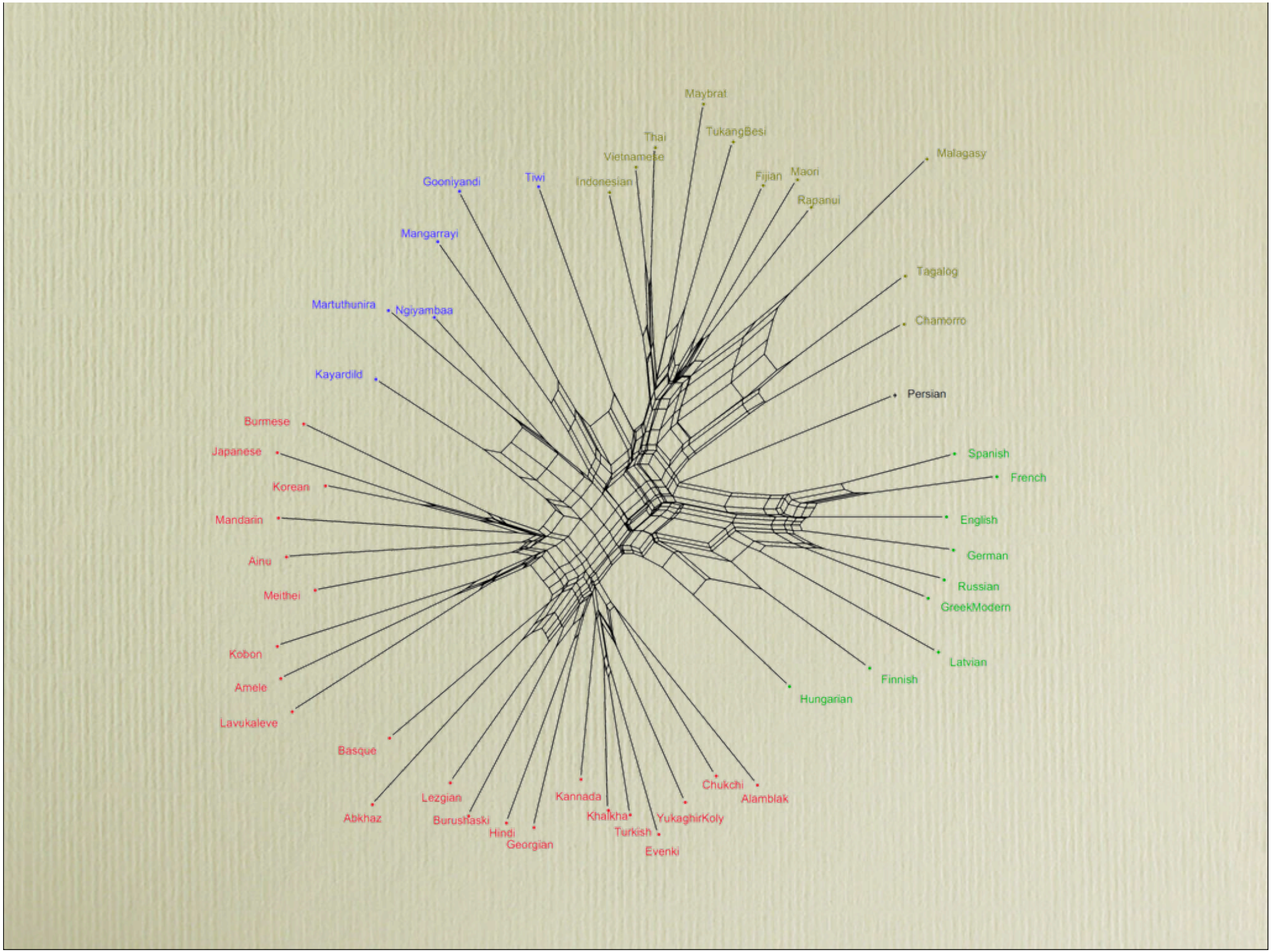

51: Position of case affixes
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
tone
suffix
no case
stem change
mixed
prefix
proclitics
enclitics
endoclitics
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

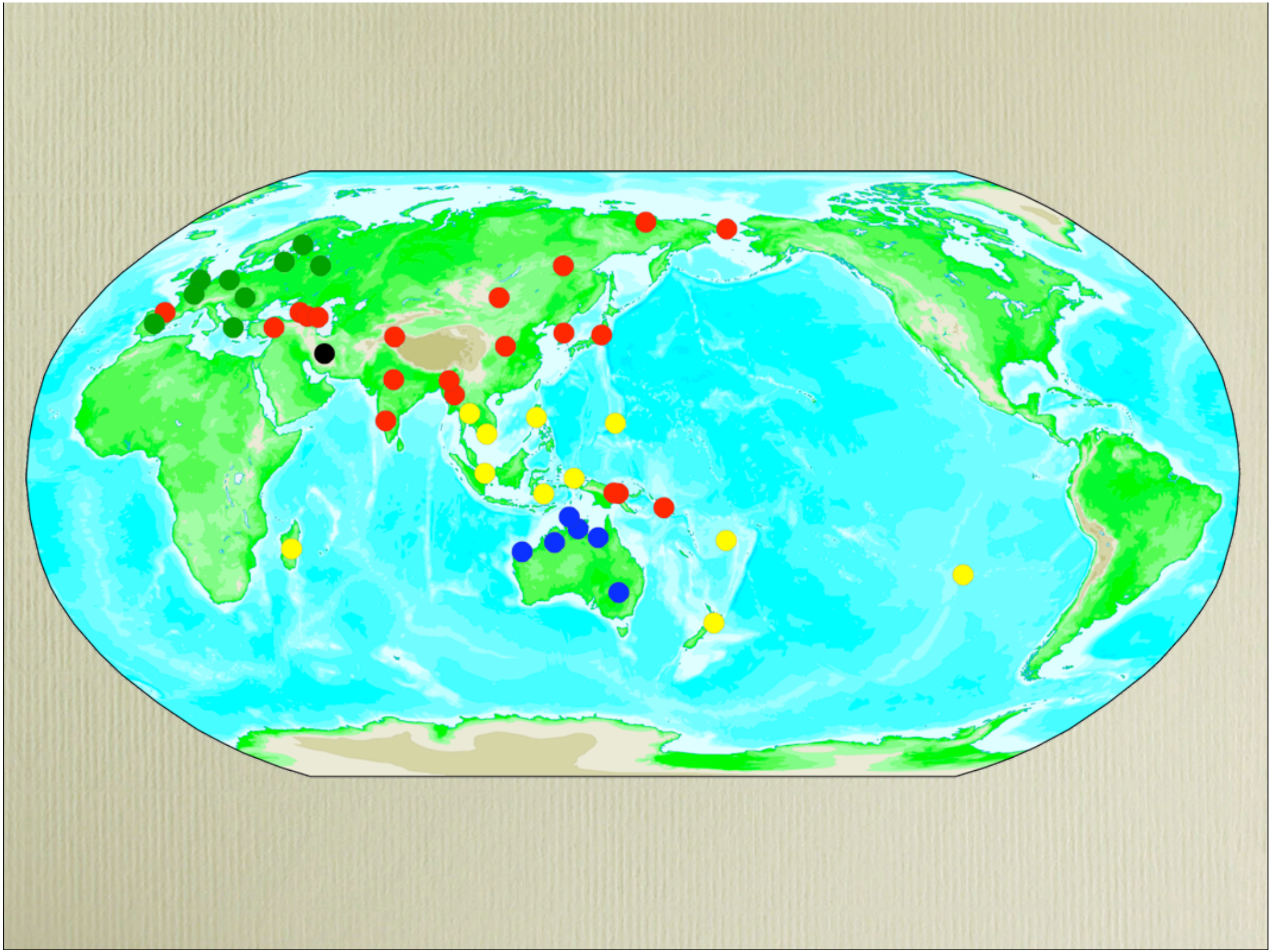

ACH
MIS
SNG
KHA
ENG
NAM
CAK
FIN
LIV
SWE
LTV
GEB
SWA
FIS
SPA
ZAP
GUA
TAJ
MRD
SGO
GRC
VIE
TRK
ITA
GRE
LIT
MAR
FRE
RHA
BSQ
RUS
MAP
MAO
HAI
TAG
SAM
CHI
HUN
dimension 2
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
dimension 1

Goals

- Fine-grained typologies
- Ideal: every language is its own type
- Establish relative similarity between values
- Establish transition probabilities by using internal structure of a genealogical tree
- We need *real* genealogical stratification of typological samples !

Similarity between Languages


Maybrat
TukangBesi
Thai
Vietnamese
Malagasy
Fijian
Maori
Indonesian
Tiwi
Gooniyandi
Rapanui
Mangarrayi
Tagalog
Martuthunira
Ngiyambaa
Chamorro
Kayardild
Persian
Burmese
Spanish
Japanese
French
Korean
English
Mandarin
German
Ainu
Russian
Meithei
GreekModern
Kobon
Latvian
Amele
Finnish
Lavukaleve
Hungarian
Basque
Chukchi
Lezgian
Kannada
Alamblak
Abkhaz
Khalkha
YukaghirKoly
Burushaski
Turkish
Hindi
Georgian
Evenki



Goals

- Interaction between genealogy and typological characteristics
- Interaction between geography and typological characteristics