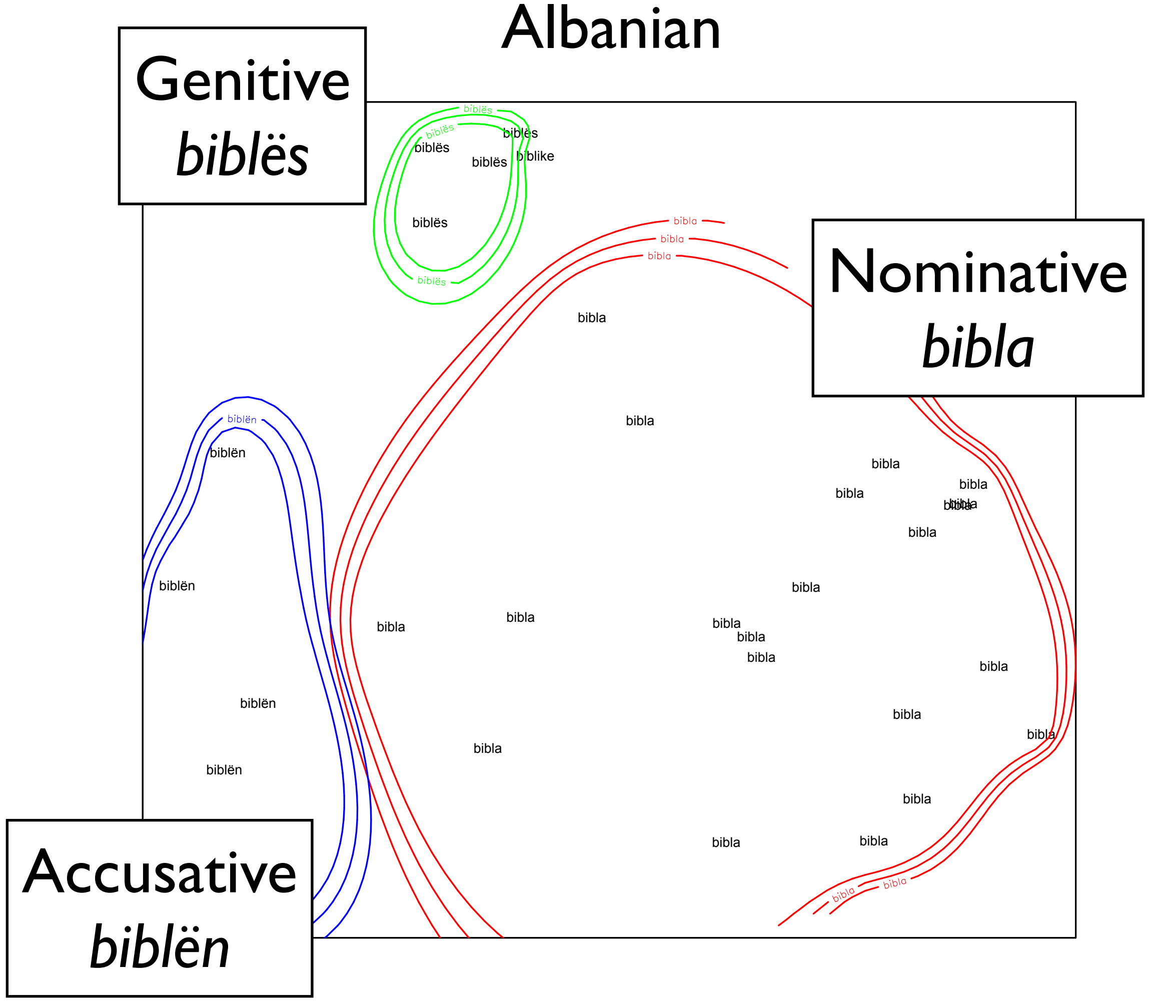
Using
language-internal similarities
to compare languages

Michael Cysouw

Research Unit "Quantitative Language Comparison"
Ludwig Maximilians University Munich

Morphosyntactic comparison: Case Marking

Albanian

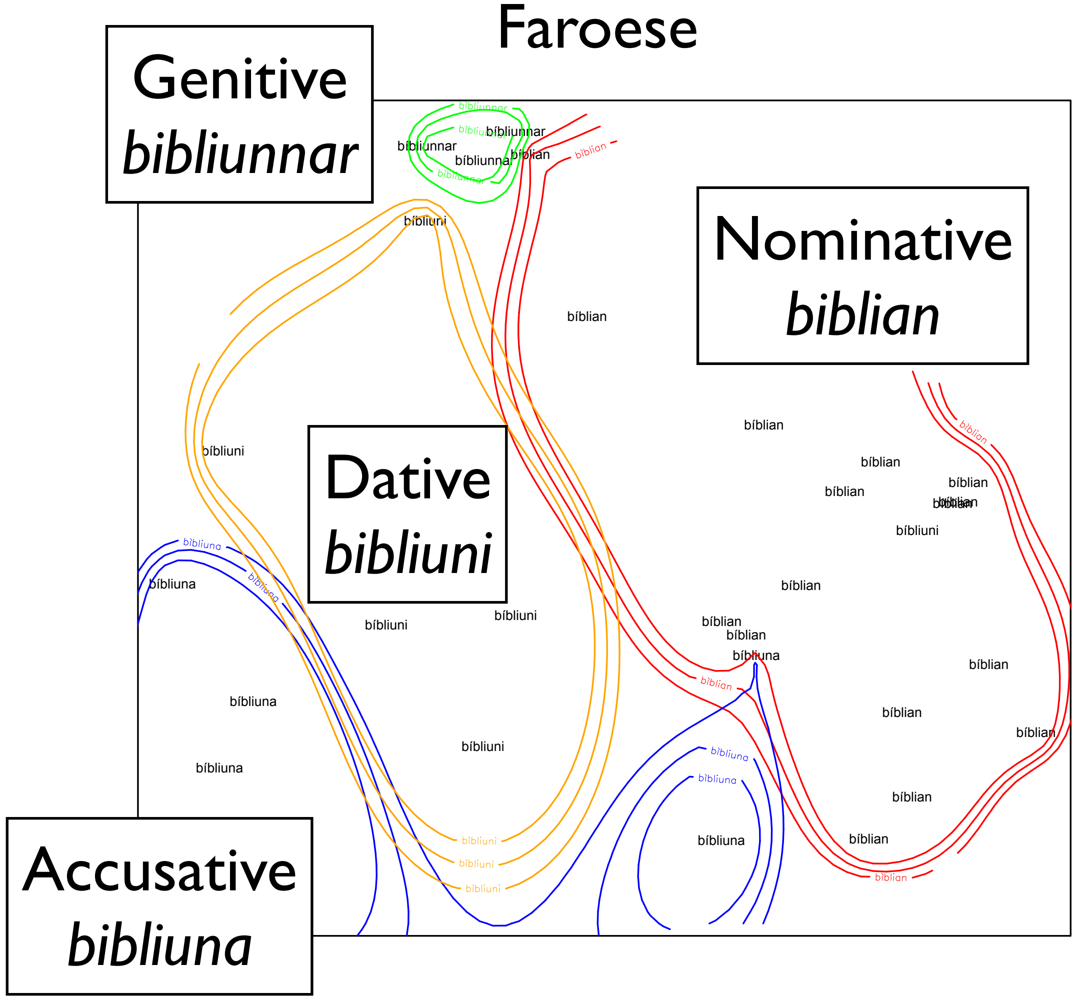


Genitive
biblës

Nominative
bibla

Accusative
biblën

Faroese



Genitive
bibliunnar

Nominative
biblian

Dative
bibliuni

Accusative
bibliuna

Assumptions (apparently slightly radical)

- Linguistic structures are language-specific
  - there is no universal ‘nominative’
    there are many different ‘Nominatives’
  - same holds also for phonemes!
- Nature of a language-specific structure is determined by the set of all occurrences
  - e.g. the meaning of the Albanian Nominative is defined by the collection of contexts in which it is used
  - a phoneme is defined by the set of
    all concrete occurrences in context

Case marking case-study

- Texts from watchtower.org
  - (Pamphlets of Jehovah's Witnesses)
- Translated into hundreds of languages
- Selection of 34 contexts in 15 languages that contain the word ‘Bible’
- Only consider bound ‘case’ marking in orthographic representation of word ‘Bible’

First 10 selected contexts

1. What important information is **contained in the Bible**?
2. Who is **the Bible's author**?
3. Why should **you study the Bible**?
4. **The Bible is** a precious gift from God.
5. **The Bible alone tells us** what we must do to please God.
6. **The Bible was written** by some 40 different men over a period of 1,600 years, beginning in 1513 B.C.E.
7. So God in heaven, not any human on earth, is **the Author of the Bible**.
8. God made sure that **the Bible was accurately copied** and preserved.
9. **More Bibles have been printed** than any other book.
10. Not everyone will be happy to see **you studying the Bible**, but do not let that stop you.

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
...	...	...	...	...	...

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

The dilemma of language comparison:

There is no objective way to decide which structure from language A is to be compared with which structure from language B

(without any a-priori definition,

which would typically be a Eurocentric one)

Solution
to the dilemma:

Language-specific metrics

(similarity measures)

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian
1	bibla
2	biblës
3	biblën
4	bibla
5	bibla
6	bibla
7	biblës
8	bibla
9	bibla
10	biblën
...	...

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla										
2	biblës										
3	biblën										
4	bibla										
5	bibla										
6	bibla										
7	biblës										
8	bibla										
9	bibla										
10	biblën										
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0									
2	biblës		0								
3	biblën			0							
4	bibla				0						
5	bibla					0					
6	bibla						0				
7	biblës							0			
8	bibla								0		
9	bibla									0	
10	biblën										0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0									
2	biblës		0								
3	biblën			0							
4	bibla				0						
5	bibla					0					
6	bibla						0				
7	biblës							0			
8	bibla								0		
9	bibla									0	
10	biblën										0
...	...										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0			0	0	0		0	0	
2	biblës		0								
3	biblën			0							
4	bibla	0			0	0	0		0	0	
5	bibla	0			0	0	0		0	0	
6	bibla	0			0	0	0		0	0	
7	biblës							0			
8	bibla	0			0	0	0		0	0	
9	bibla	0			0	0	0		0	0	
10	biblën										0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0			0	0	0		0	0	
2	biblës		0								
3	biblën			0							
4	bibla	0			0	0	0		0	0	
5	bibla	0			0	0	0		0	0	
6	bibla	0			0	0	0		0	0	
7	biblës							0			
8	bibla	0			0	0	0		0	0	
9	bibla	0			0	0	0		0	0	
10	biblën										0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0			0	0	0		0	0	
2	biblës		0					0			
3	biblën			0							
4	bibla	0			0	0	0		0	0	
5	bibla	0			0	0	0		0	0	
6	bibla	0			0	0	0		0	0	
7	biblës		0					0			
8	bibla	0			0	0	0		0	0	
9	bibla	0			0	0	0		0	0	
10	biblën										0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0			0	0	0		0	0	
2	biblës		0					0			
3	biblën			0							
4	bibla	0			0	0	0		0	0	
5	bibla	0			0	0	0		0	0	
6	bibla	0			0	0	0		0	0	
7	biblës		0					0			
8	bibla	0			0	0	0		0	0	
9	bibla	0			0	0	0		0	0	
10	biblën										0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0			0	0	0		0	0	
2	biblës		0					0			
3	biblën			0							0
4	bibla	0			0	0	0		0	0	
5	bibla	0			0	0	0		0	0	
6	bibla	0			0	0	0		0	0	
7	biblës		0					0			
8	bibla	0			0	0	0		0	0	
9	bibla	0			0	0	0		0	0	
10	biblën			0							0
…	…										

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0	1	1	0	0	0	1	0	0	1
2	biblës	1	0	1	1	1	1	0	1	1	1
3	biblën	1	1	0	1	1	1	1	1	1	0
4	bibla	0	1	1	0	0	0	1	0	0	1
5	bibla	0	1	1	0	0	0	1	0	0	1
6	bibla	0	1	1	0	0	0	1	0	0	1
7	biblës	1	0	1	1	1	1	0	1	1	1
8	bibla	0	1	1	0	0	0	1	0	0	1
9	bibla	0	1	1	0	0	0	1	0	0	1
10	biblën	1	1	0	1	1	1	1	1	1	0
…	…										

	Irish	1	2	3	4	5	6	7	8	9	10
1	bhíobla										
2	bhíobla										
3	mbíobla										
4	bíobla										
5	bíobla										
6	bhíobla										
7	bhíobla										
8	bíobla										
9	bhíobla										
10	bhíobla										
…	…										

	Irish	1	2	3	4	5	6	7	8	9	10
1	bhíobla	0	0	1	1	1	0	0	1	0	0
2	bhíobla	0	0	1	1	1	0	0	1	0	0
3	mbíobla	1	1	0	1	1	1	1	1	1	1
4	bíobla	1	1	1	0	0	1	1	0	1	1
5	bíobla	1	1	1	0	0	1	1	0	1	1
6	bhíobla	0	0	1	1	1	0	0	1	0	0
7	bhíobla	0	0	1	1	1	0	0	1	0	0
8	bíobla	1	1	1	0	0	1	1	0	1	1
9	bhíobla	0	0	1	1	1	0	0	1	0	0
10	bhíobla	0	0	1	1	1	0	0	1	0	0
…	…										

Language-specific metric
on sampled contexs

- Many more intricate possibilities to define such a similarity measure
- Important: the similarity only uses structure of one single language
- There is no language comparison (yet)!

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

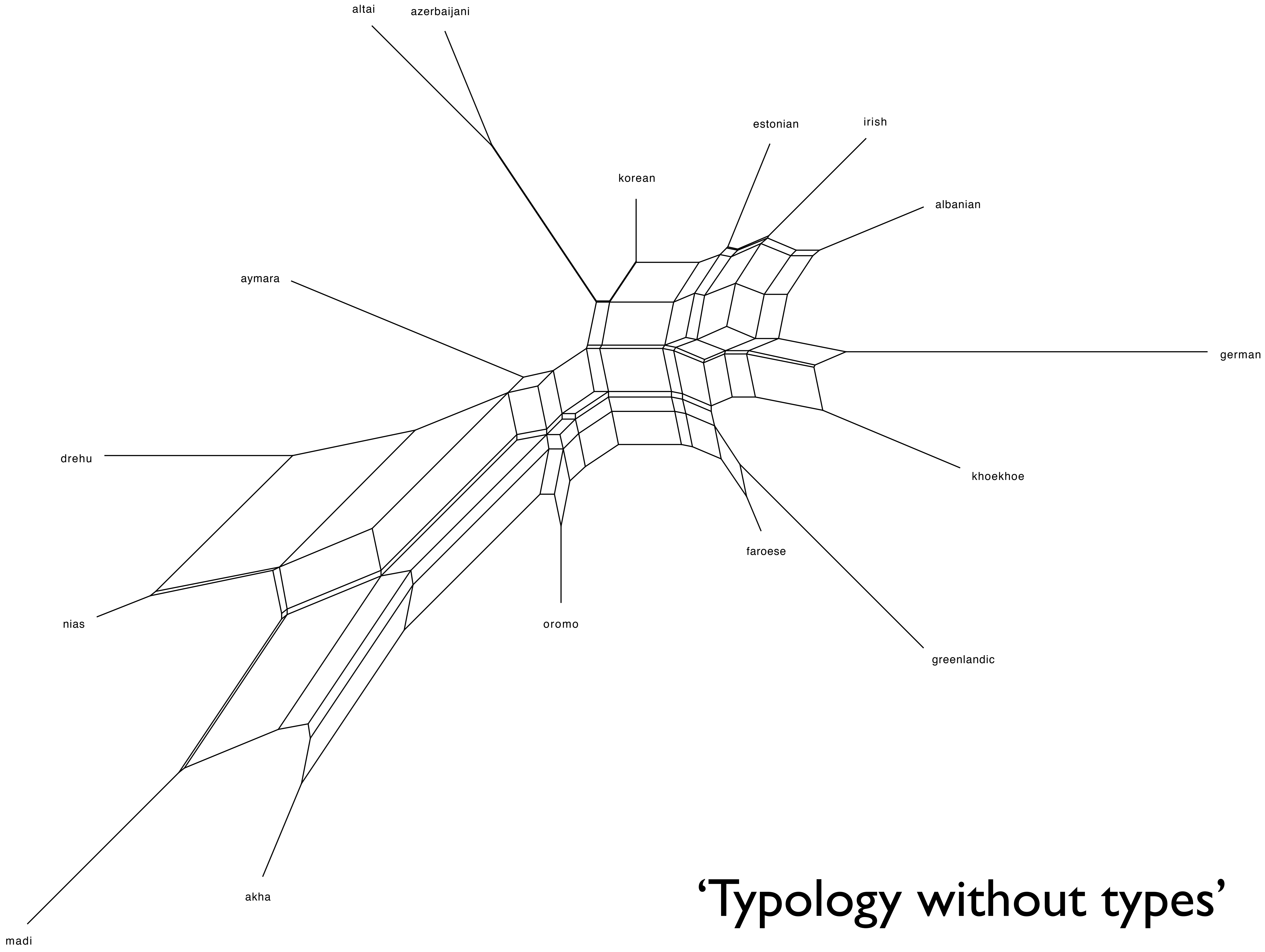
Similarity between languages ...

	Albanian	1	2	3	4	5	6	7	8	9	10
1	bibla	0	1	1	0	0	0	1	0	0	1
2	biblës	1	0	1	1	1	1	0	1	1	1
3	biblën	1	1	0	1	1	1	1	1	1	0
4	bibla	0	1	1	0	0	0	1	0	0	1
5	bibla	0	1	1	0	0	0	1	0	0	1
6	bibla	0	1	1	0	0	0	1	0	0	1
7	biblës	1	0	1	1	1	1	0	1	1	1
8	bibla	0	1	1	0	0	0	1	0	0	1
9	bibla	0	1	1	0	0	0	1	0	0	1
10	biblën	1	1	0	1	1	1	1	1	1	0
…	…										

	Irish	1	2	3	4	5	6	7	8	9	10
1	bhíobla	0	0	1	1	1	0	0	1	0	0
2	bhíobla	0	0	1	1	1	0	0	1	0	0
3	mbíobla	1	1	0	1	1	1	1	1	1	1
4	bíobla	1	1	1	0	0	1	1	0	1	1
5	bíobla	1	1	1	0	0	1	1	0	1	1
6	bhíobla	0	0	1	1	1	0	0	1	0	0
7	bhíobla	0	0	1	1	1	0	0	1	0	0
8	bíobla	1	1	1	0	0	1	1	0	1	1
9	bhíobla	0	0	1	1	1	0	0	1	0	0
10	bhíobla	0	0	1	1	1	0	0	1	0	0
…	…										

... by establishing a similarity between
language-specific similarities

	Language	1	2	3	4	5
1	Albanian	0.00	0.28	0.21	0.30	0.76
2	Irish	0.28	0.00	0.35	0.27	0.74
3	Faroese	0.21	0.35	0.00	0.33	0.42
4	Estonian	0.30	0.27	0.33	0.00	0.59
5	Aymara	0.76	0.74	0.42	0.59	0.00
…	…					



‘Typology without types’

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

	Albanian	Irish	Faroese	Estonian	Aymara
1	bibla	bhíobla	bíbliuni	piibel	bibliaxa
2	biblës	bhíobla	bíbliunnar	piibli	bibliax
3	biblën	mbíobla	bíbliuna	piiblit	bibliat
4	bibla	bíobla	bíblian	piibel	bibliax
5	bibla	bíobla	bíblian	piibel	bibliakiw
6	bibla	bhíobla	bíbliuna	piibli	bibliax
7	biblës	bhíobla	bíbliunnar	piibli	bibliaxa
8	bibla	bíobla	bíblian	piiblit	bibliaxa
9	bibla	bhíobla	(n.a.)	piiblit	bibliawa
10	biblën	bhíobla	bíbliuna	piiblit	bibliat
…	…	…	…	…	…

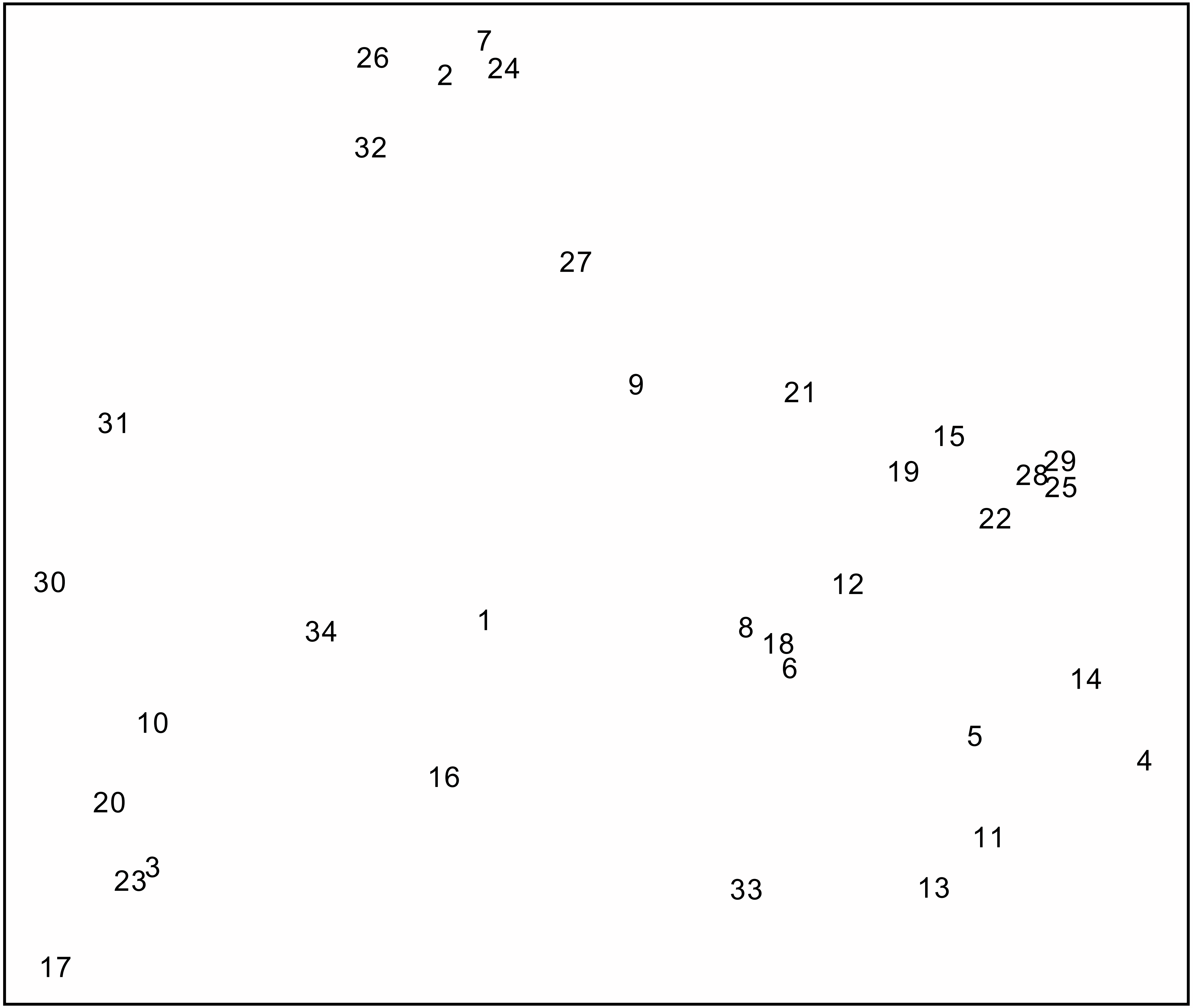
identically coded in 2 out of 5 languages

Cross-linguistic metric

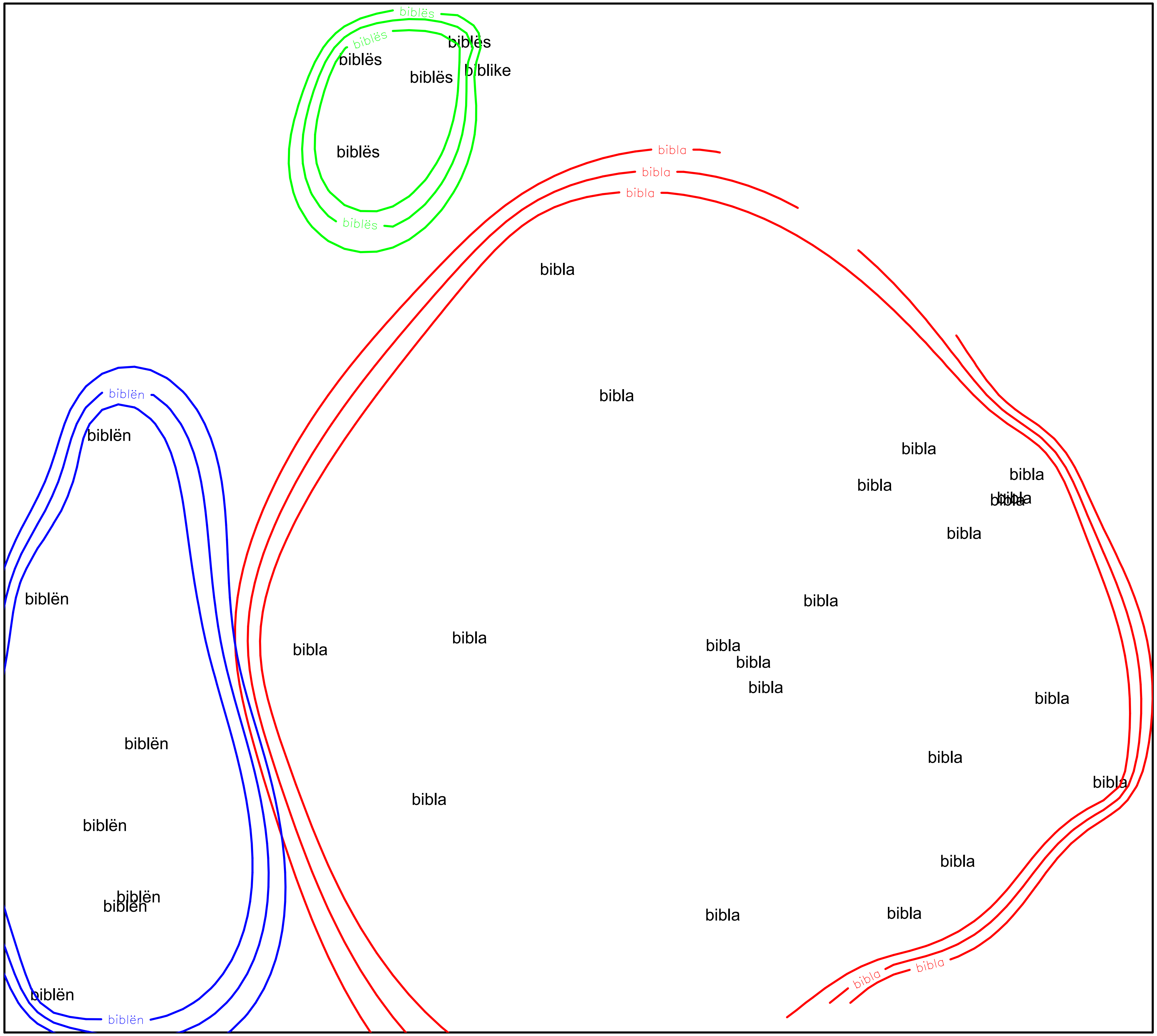
on sampled contexs

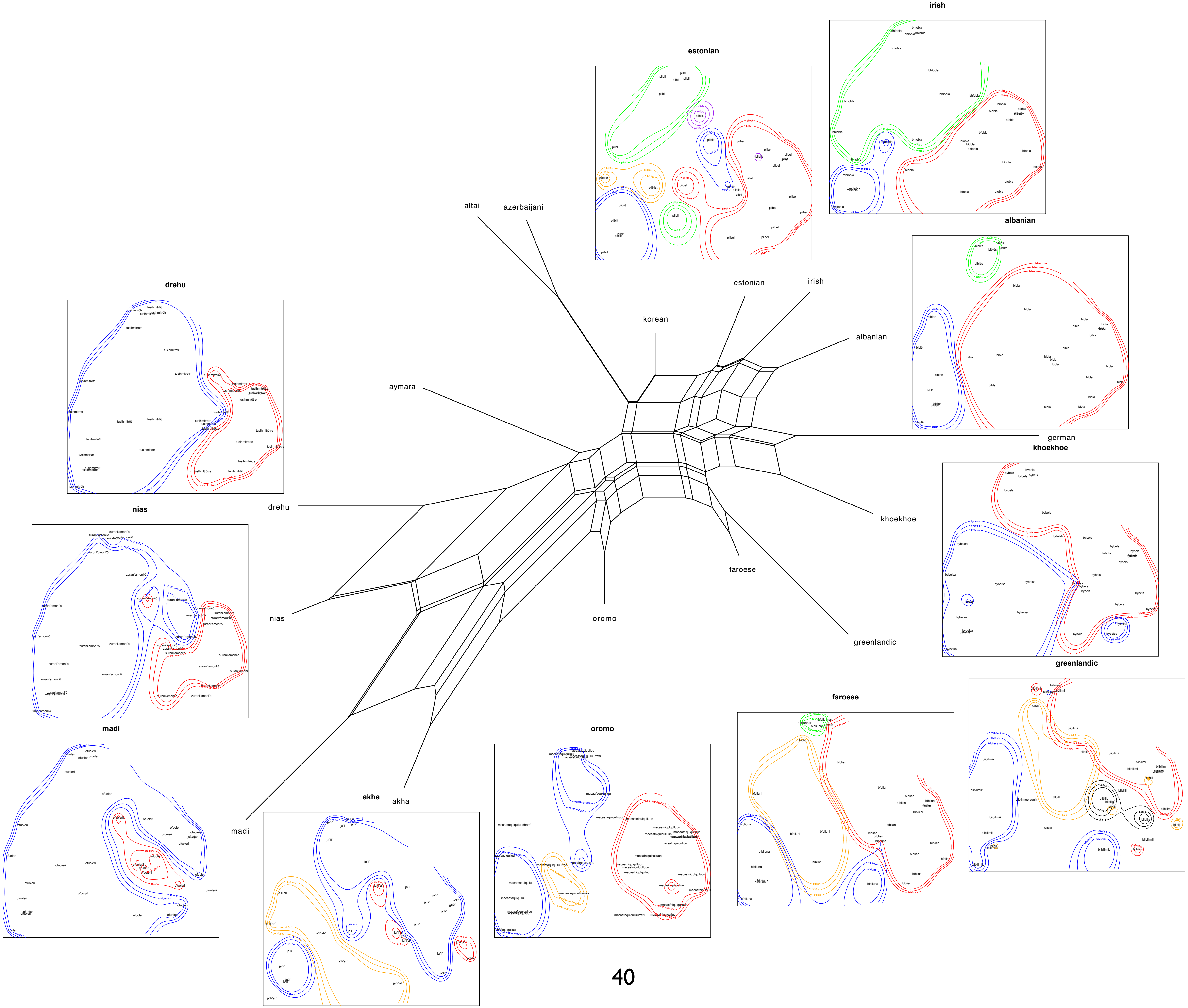
	1	2	3	4	5
1	0	0.28	0.21	0.30	0.76
2	0.28	0	0.35	0.27	0.74
3	0.21	0.35	0	0.33	0.42
4	0.30	0.27	0.33	0	0.59
5	0.76	0.74	0.42	0.59	0

(average of language-specific metrics)


7
26
24
2
32
27
9
21
31
15
29
19
28
25
22
30
12
1
8
34
18
6
14
10
5
4
16
20
11
3
23
13
33
17

albanian


biblës
biblës
biblës
biblës
biblës
biblës
biblës
biblike
biblës
bibla
bibla
bibla
bibla
bibla
biblën
biblën
biblën
biblën
biblën
biblën
biblën
biblën
biblën
biblën
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla
bibla


irish
estonian
albanian
drehu
nias
madi
akha
oromo
faroese
greenlandic
khoekhoe
altai
azerbaijani
korean
aymara
estonian
irish
albanian
german
khoekhoe
faroese
greenlandic
oromo
drehu
nias
madi
akha

Summary

- Language comparison has to be based on a suitable set of exemplars in context
  - Language-specific metrics on contexts
  - Only uses knowledge about the structure of one language
- Language comparison is reduced to a metric on language-specific metrics
  - ‘typology without types’
- Compare language-specific metric to an ‘average’ metric
  - Cross-linguistic metric on contexts
  - Visualize structure of individual languages on this basis