Getting the nitty-gritty details into big data

Michael Cysouw
Phillips-University Marburg

How big is big ?

- "Big" does not necessarily mean "googol-size" (=10^{100}), or "google-ngram-size" ($\approx 10^{10}$)
- **"Big" often is much simpler, just: "too big to grasp in one human brain"**
- Humanity-scholars often spend decades to be able to grasp an awful amount of data in one human brain!

The importance of the nitty-gritty stuff

- By using large datasets, many problems can be automatically solved
- Often, 70~80% precision is easily reached, with each further percentage being difficult
- **Yet, many insights in the humanities depend upon unusual, strange, rare cases**
- It is difficult to get beyond the "we-knew-that-already" state when using big data

Ideal world …

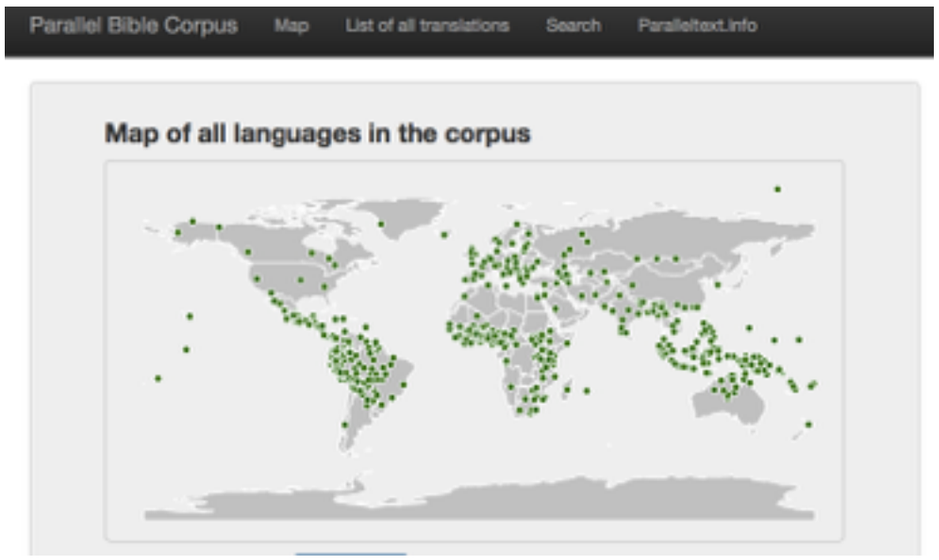
- **Let the big data do the easy part automatically**
- **So humans can focus on the interesting stuff!**

- Observation:
  - papers on big-data automatic analyses often report results only in the aggregate
  - Yet, it is normally the concrete decisions in individual cases that are more revealing!

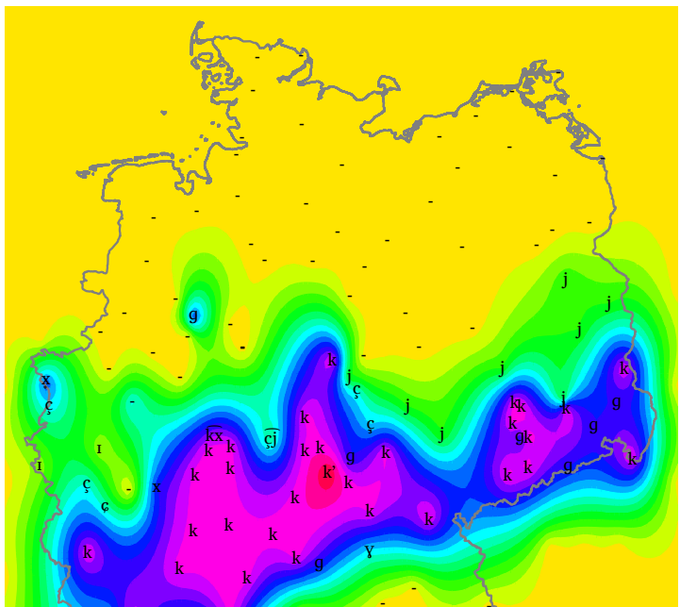
Three examples from ongoing research

Etymological Dictionaries

Massively Parallel Texts



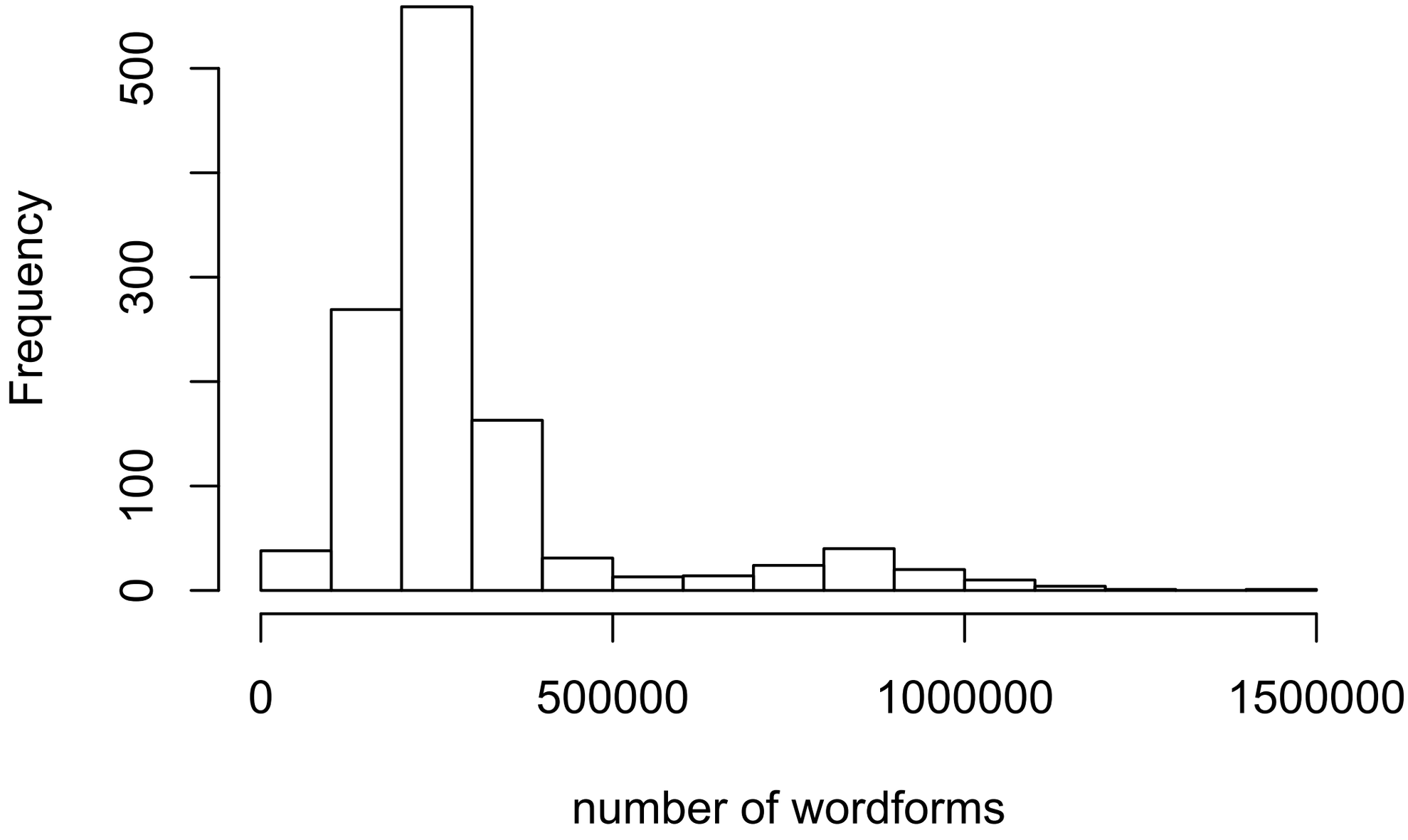
Dialekt Comparison



1. Bible corpus

paralleltext.info

- Bible translations in unified and cleaned format, tokenised for punctuation
- **Currently 1172 translation in 907 different ISO 639-3 codes (“languages”)**
- On average “only” **300 K** wordforms per translation (including punctuation…)
- Total corpus **350 M wordforms**!


Frequency
0
100
300
500
0
500000
1000000
1500000
number of wordforms

Word Alignment

- GIZA, berkeley aligner, fast align …
- One problem: with 1000 translations, there are **500.000 pairs**, which even optimistically calculated will take about a year to align everything
- What really is interesting for comparative linguistics is a **multiple-alignment**, and getting from pairs to multiple is not trivial either
- From a manual perspective, the quality is low

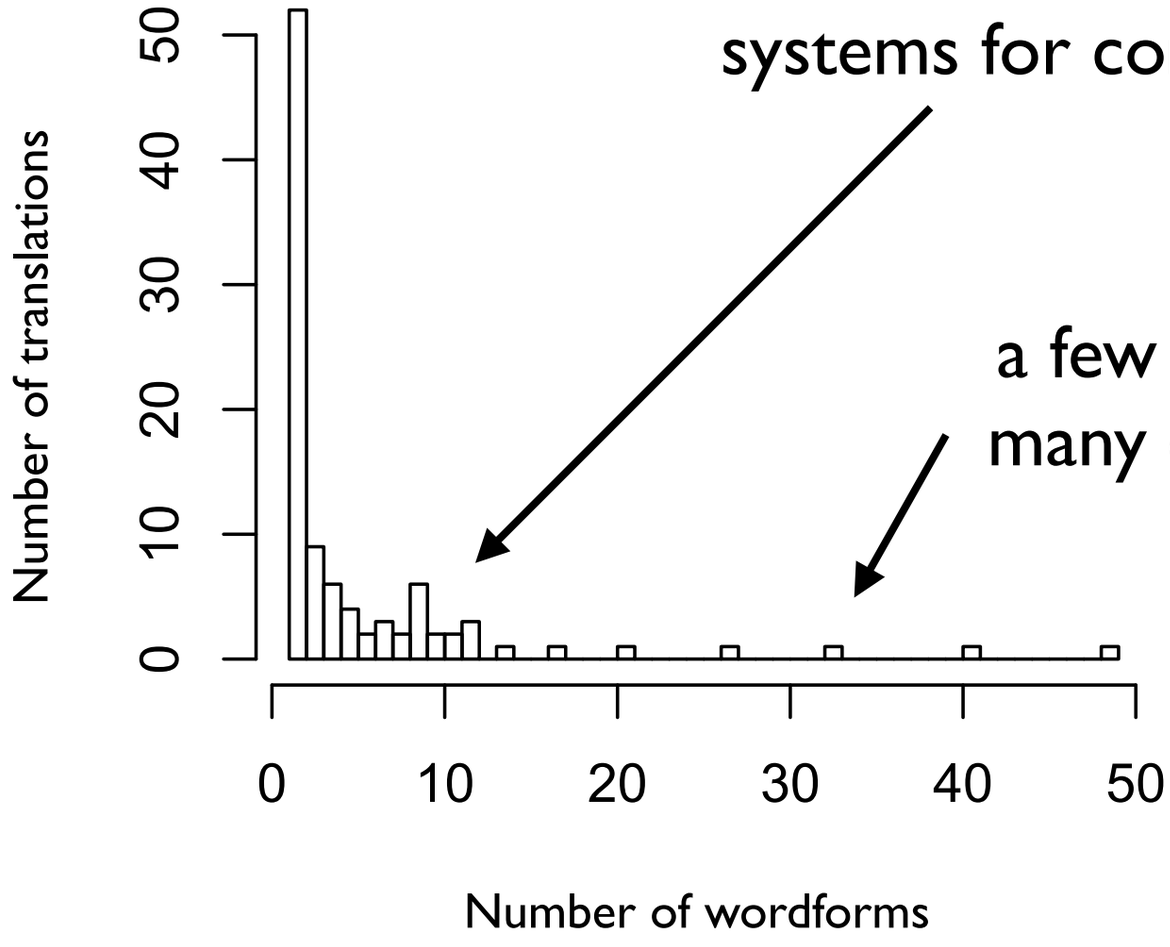
quick test: translation of "*Jerusalem*"

- Identify wordforms that correspond to the English name *Jerusalem* in 100 translations
  - mostly automatic, minor manual work still needed
- **Many languages will just have one wordform, but some will have more than one**
- These different wordforms might give us information about case!

more than half of the languages have just one wordform

about 40% of the languages have 2-10 forms, which are informative systems for comparison

a few languages have very many different wordforms



Angaataha

(ISO 639-5 agm, spoken in Papua New Guinea)

- jerusarem+handa
- jerusarem+handaahap+
- jerusarem+hand+
- jerusarem+handaahiyai
- jerusarem+handaah+
- jerusarem+handaahap+hiyaun+
- jerusarem+handaahiyaisangi
- jerusarem+handaahap+hiya
- jerusarem+handaah+raap+
- jerusarem+handaah+h+
- jerusarem+handaah+he
- jerusarem+handam+
- jerusarem+manda
- jerusarem+handap+
- jerusarem+nd+
- jerusarem+handaahap+to
- jerusarem+handaahapaah+h+
- jerusarem+handi
- jerusarem+mandaahap+
- jerusarem+handaahun+
- jerusarem+handaahapun+
- jerusarem+handaahiya
- jerusarem+handam+hin+
- jerusarem+handaahap+hiyaatih+
- jerusarem+handaahap+hiyaate
- jerusarem+handaahiyaun+
- jerusaremos+hiyaate

Amharic

(ISO 639-3 amh, spoken in Ethiopia)

- **ኢየሩሳሌም**
- **በኢየሩሳሌም**
- **ከኢየሩሳሌም**
- **ኢየሩሳሌምም**
- **በኢየሩሳሌምም**
- **ኢየሩሳሌምን**
- **ከኢየሩሳሌምም**
- **የኢየሩሳሌም**
- **ለኢየሩሳሌም**
- **ለኢየሩሳሌምም**
- **የኢየሩሳሌምንም**
- **የኢየሩሳሌምም**

Amharic

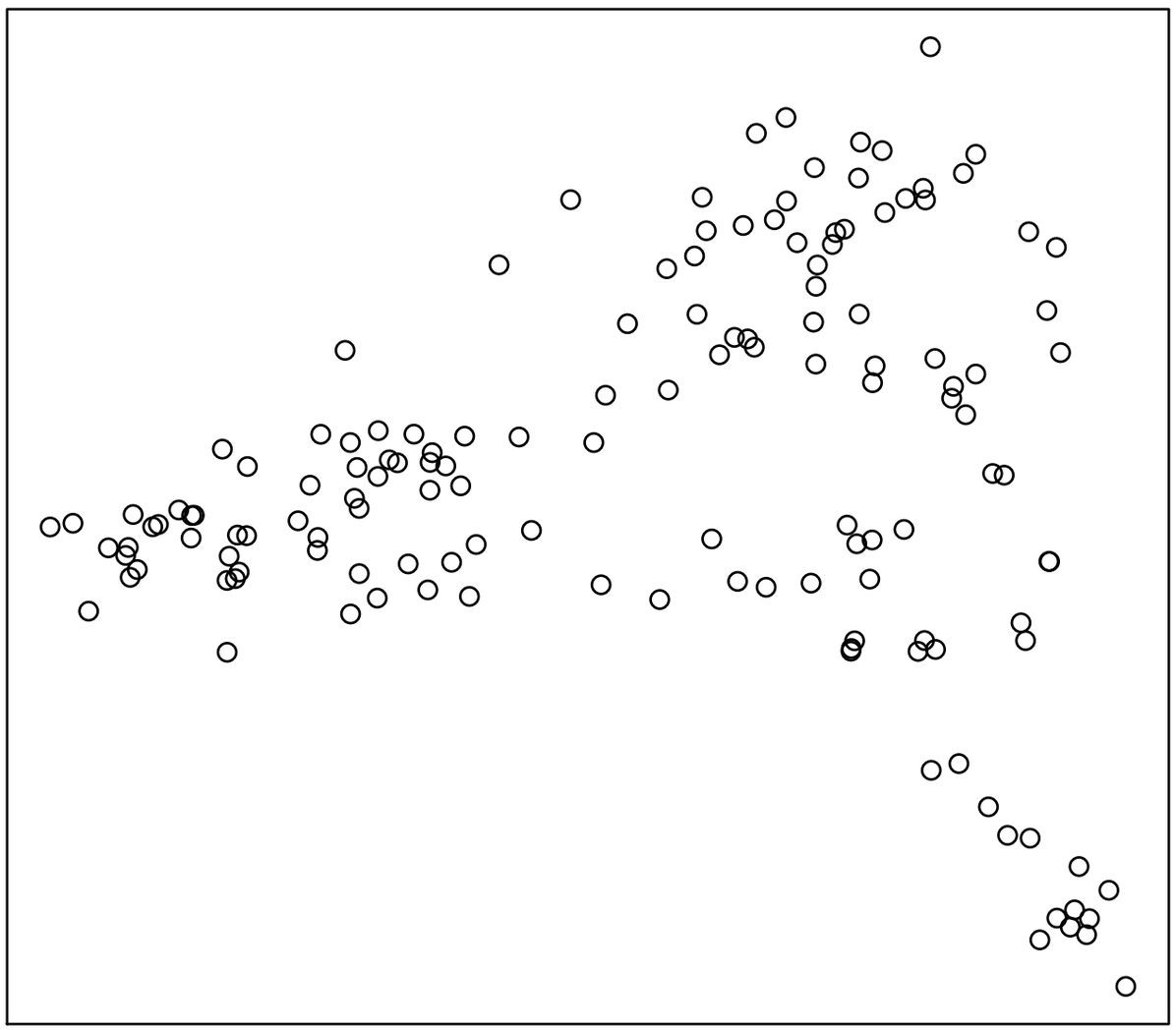
(ISO 639-3 amh, spoken in Ethiopia)

- ኢየሩሳሌም
- በኢየሩሳሌም
- ከኢየሩሳሌም
- ኢየሩሳሌምም
- በኢየሩሳሌምም
- ኢየሩሳሌምን
- ከኢየሩሳሌምምም
- የኢየሩሳሌም
- ለኢየሩሳሌም
- ለኢየሩሳሌምም
- የኢየሩሳሌምንም
- የኢየሩሳሌምም

98 languages, in total 520 different wordforms

We selected 167 verses including *Jerusalem*

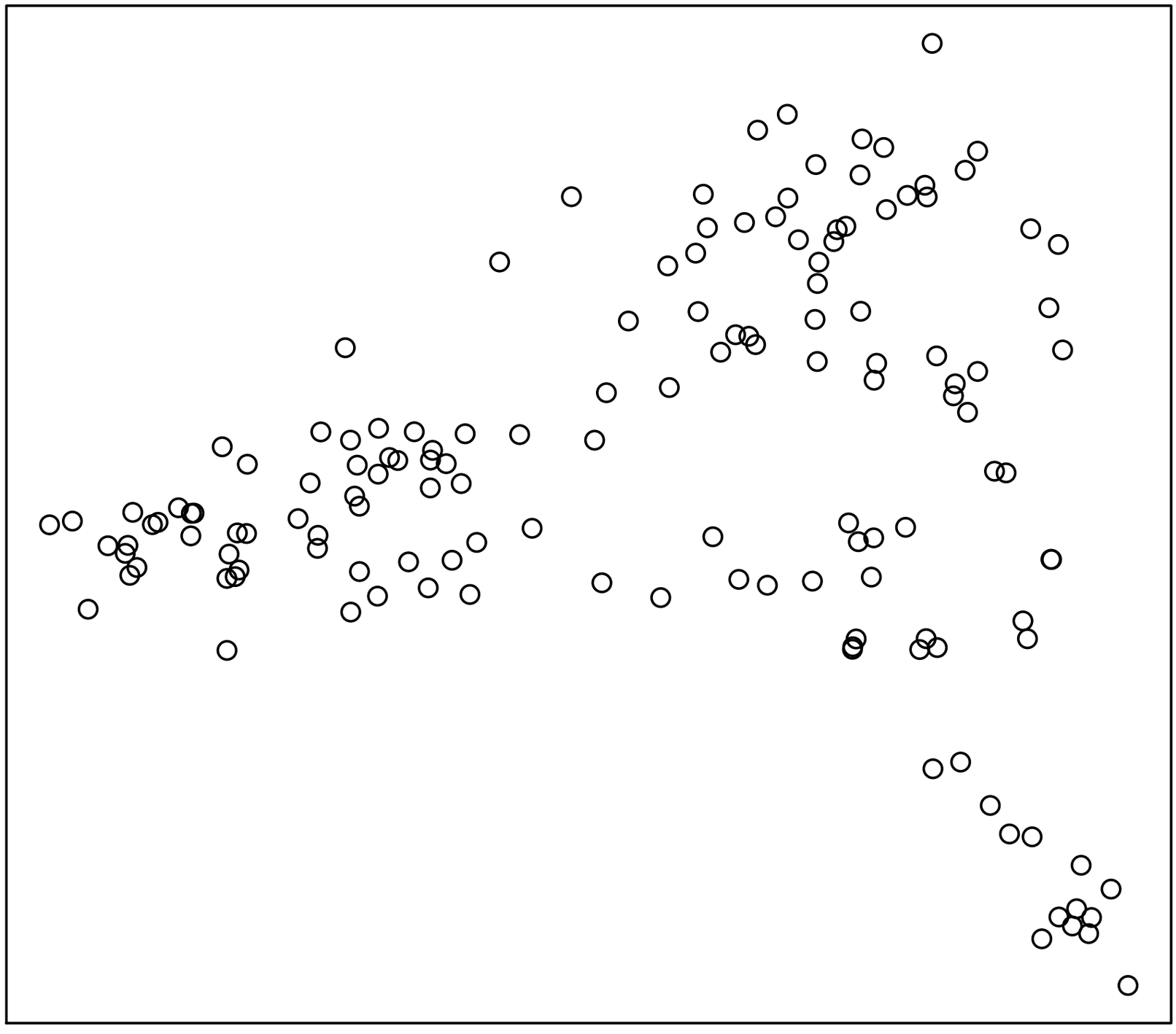
Matrix of size 520 x 167 coding the distribution of wordforms over verses



Two verses with *Jerusalem* are similar when they often share the same wordform in language after language

here showing two main dimensions of variation of 167 contexts

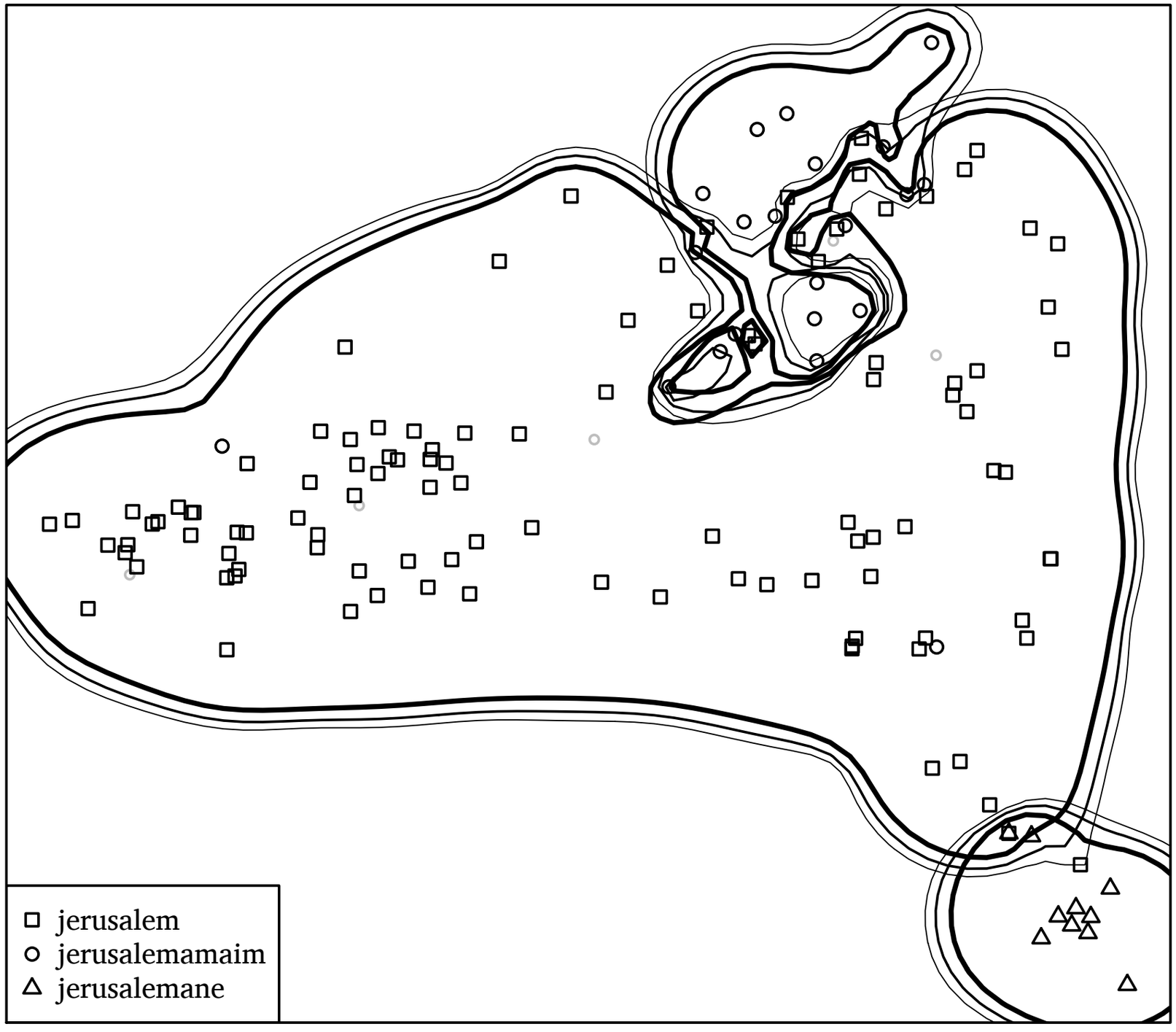
- the importance of dimensions depend strongly on the content of the corpus, which we cannot control
- only the first two dimensions are discussed here because of easy visualisation



aey

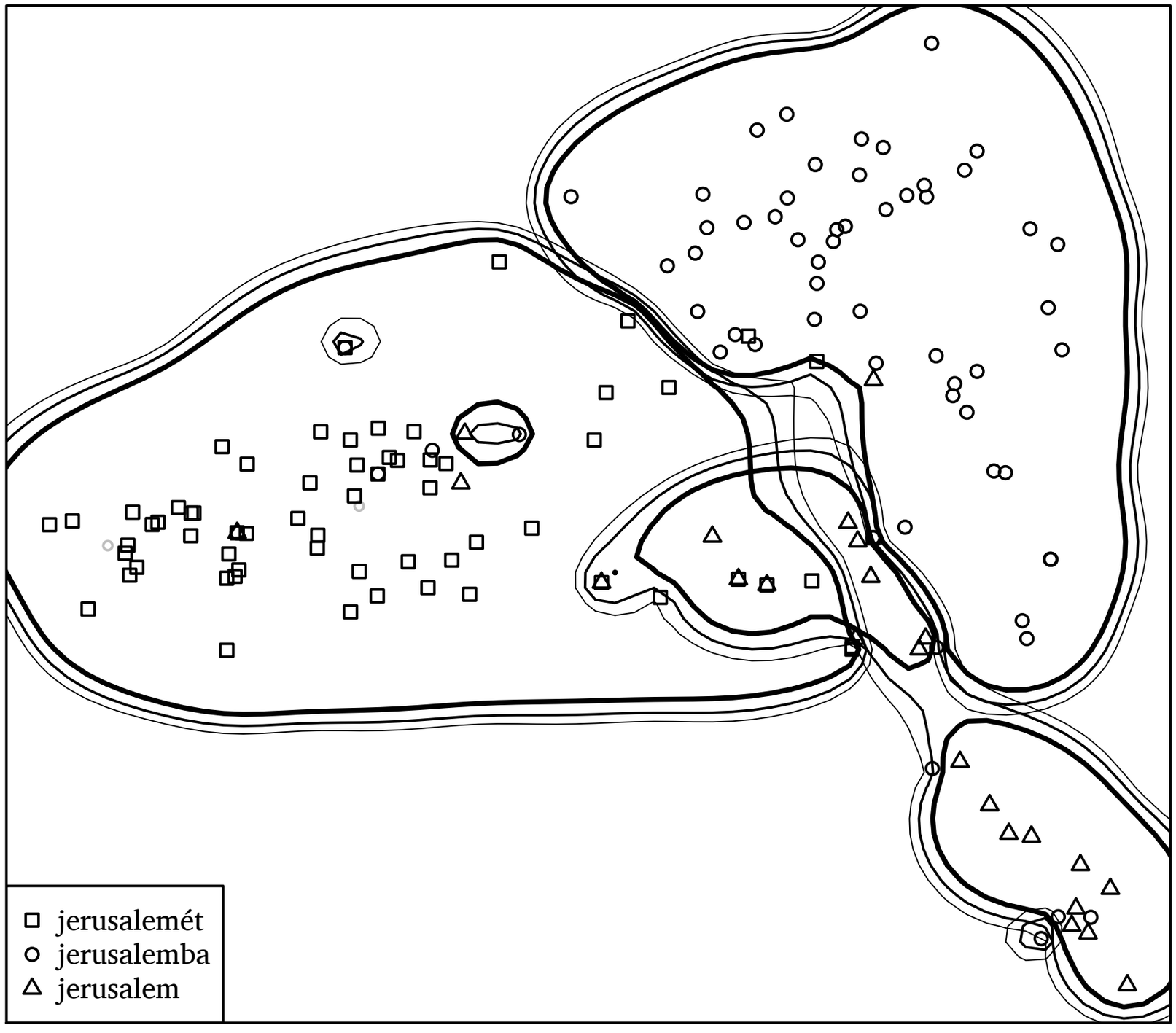
Amele (A language of Papua New Guinea)

aai



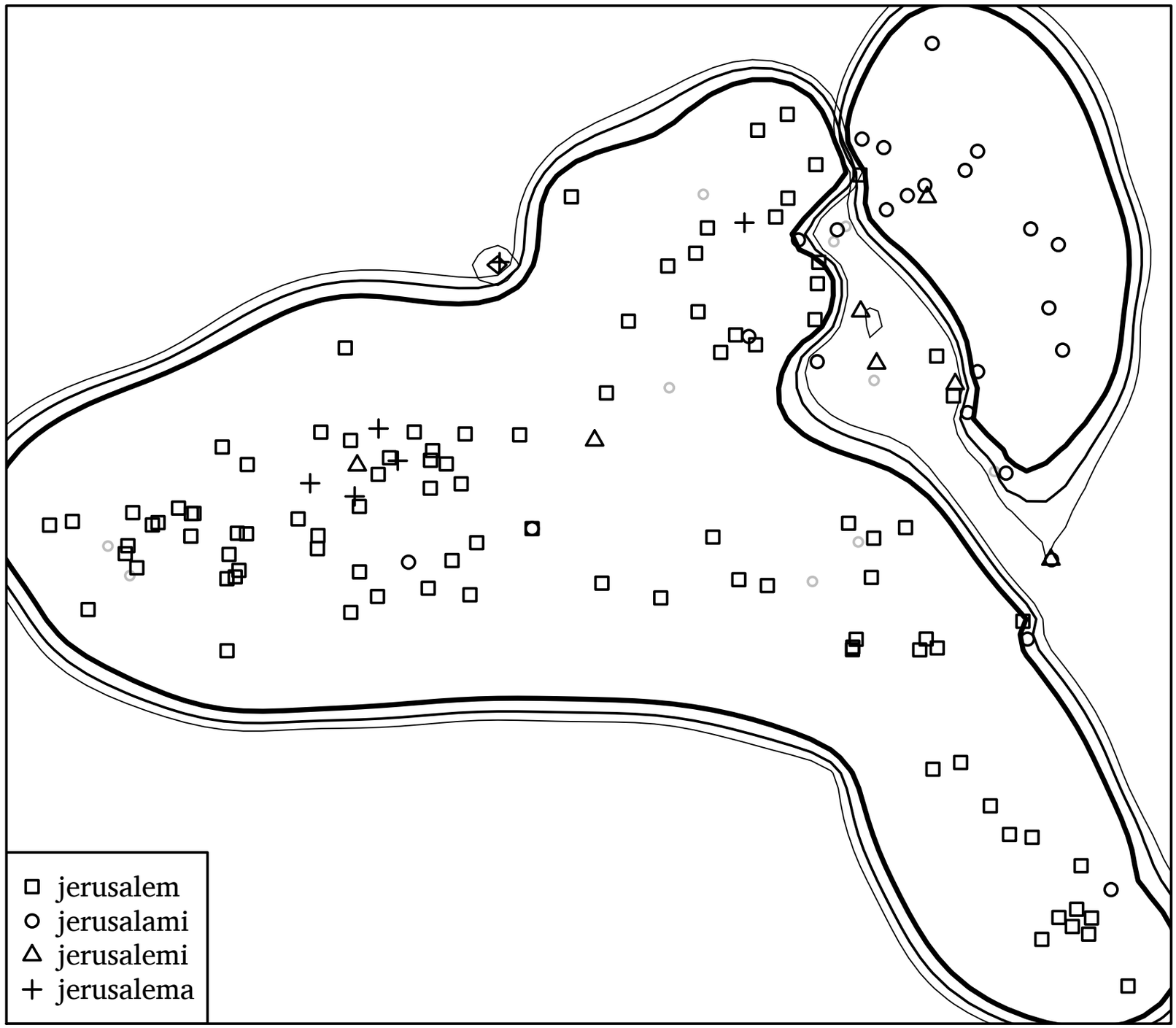
Arifama-Miniafia (a language of Papua New Guinea)

abt



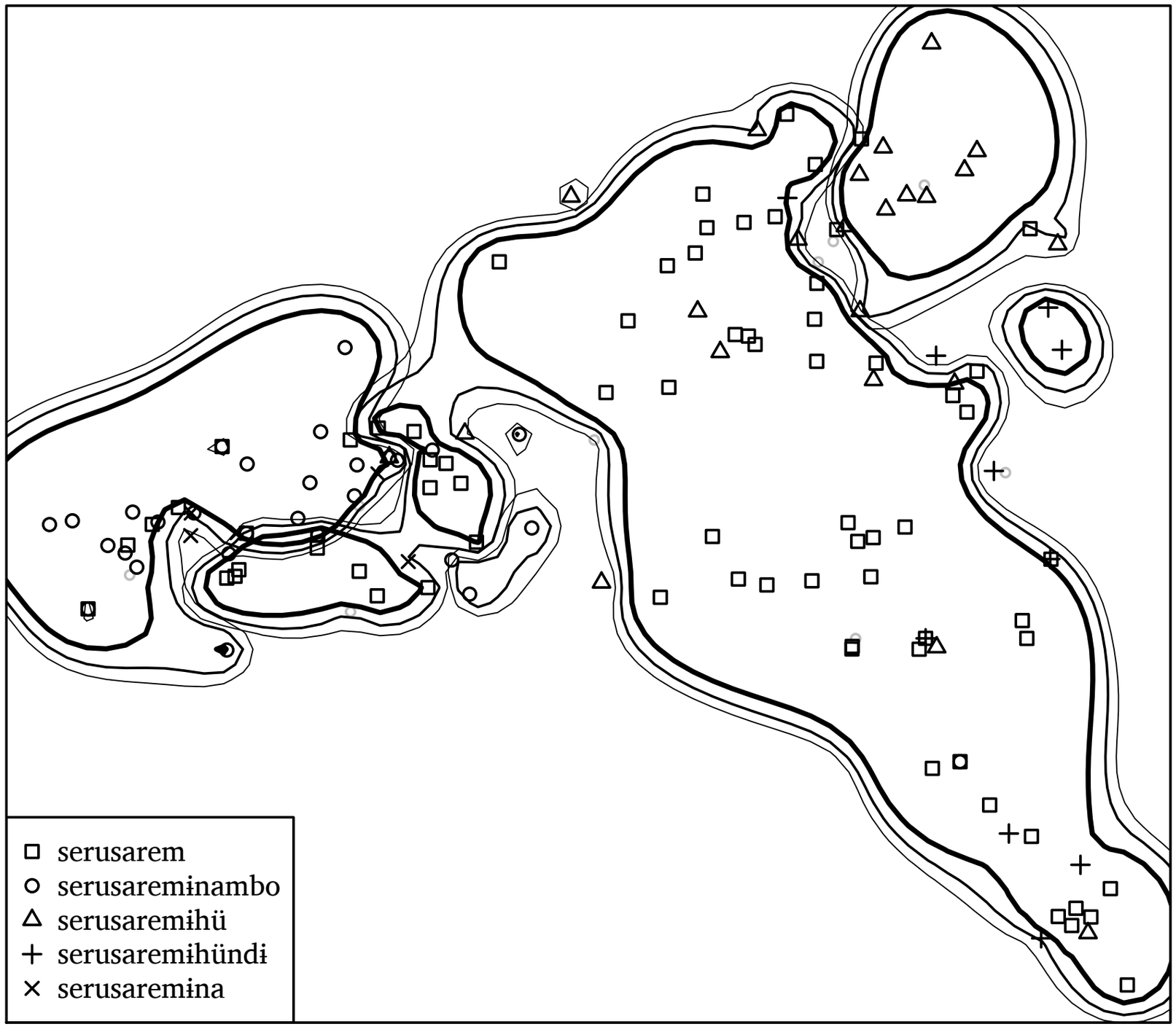
Ambulas (a language of Papua New Guinea)

aoj



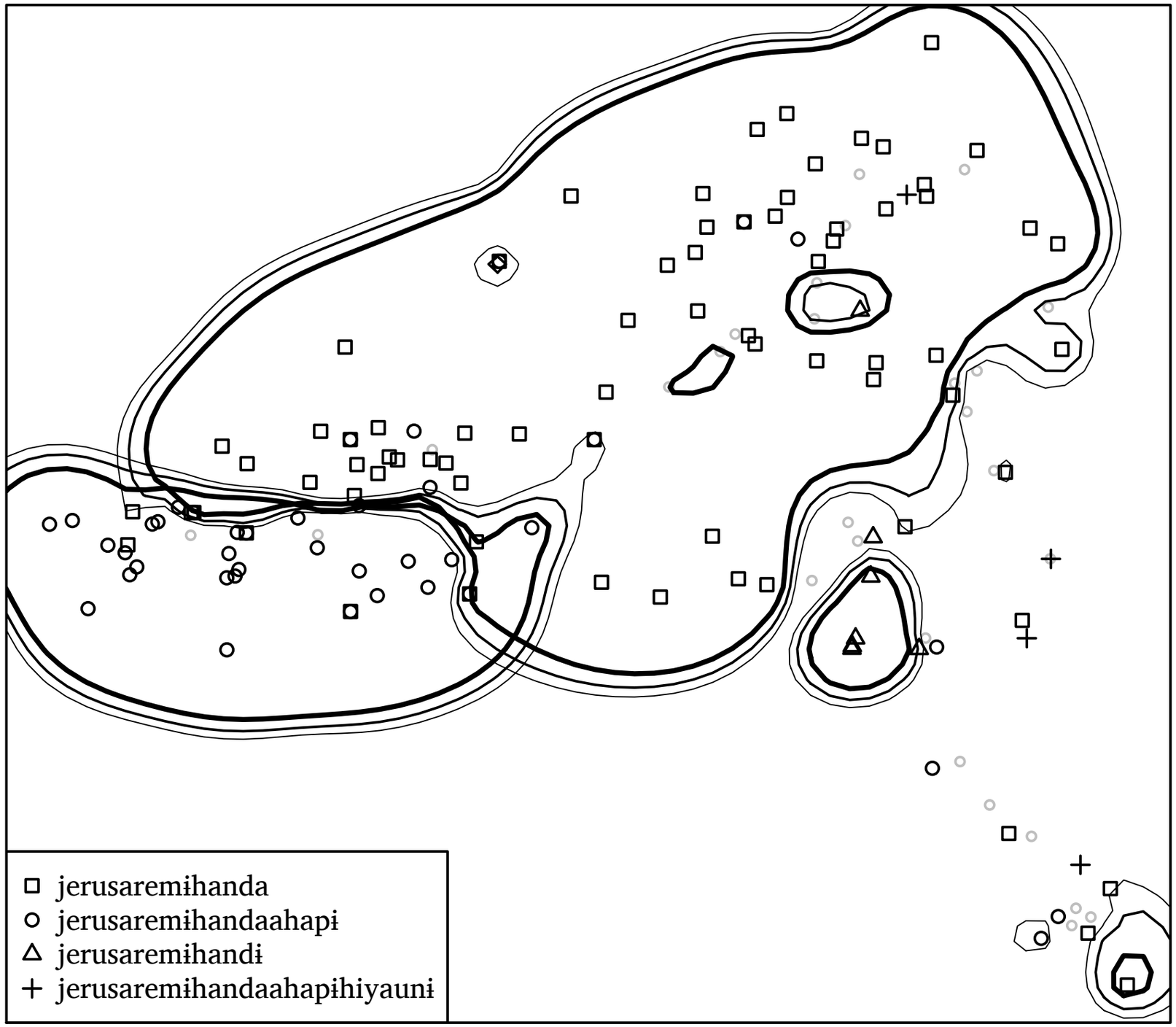
Muffian (a language of Papua New Guinea)

agg



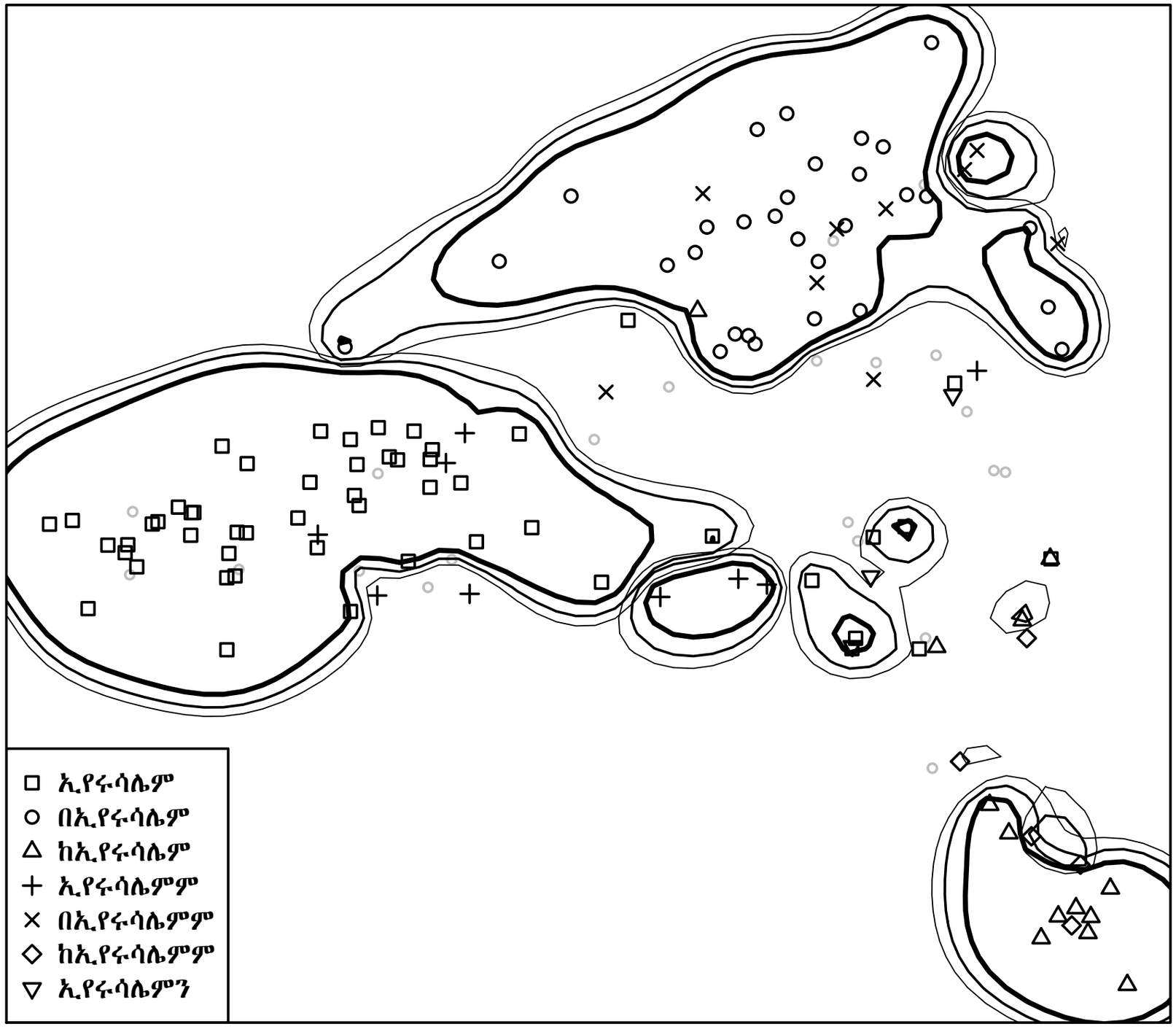
Angor (a language of Papua New Guinea)

agm



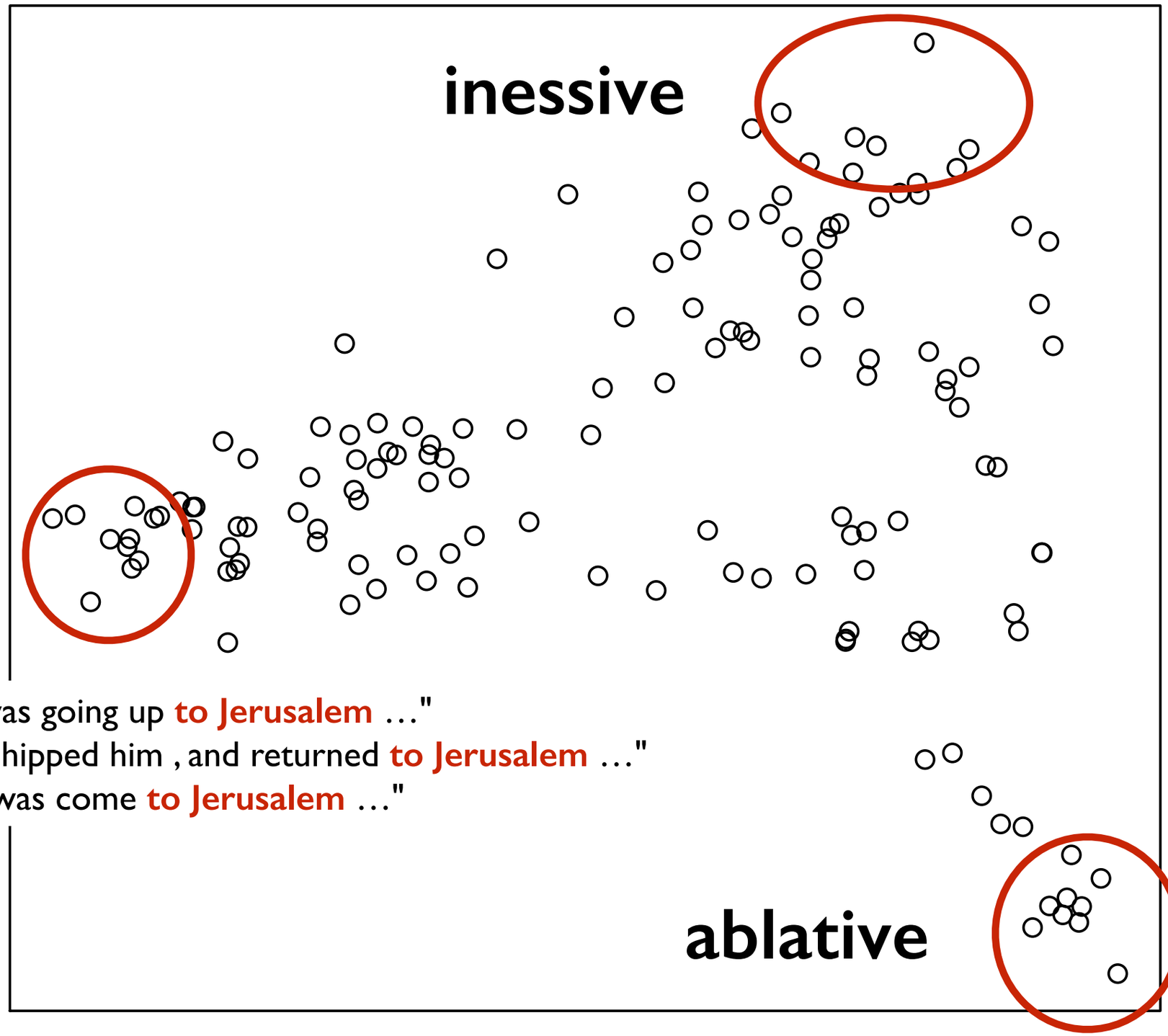
Angaatiha (a language of Papua New Guinea)

amh



Amharic (a language of Ethiopia)

42002025: "… there was a man **in Jerusalem** whose name was Simeon …"
44004005: "… their rulers and elders and scribes were gathered together **in Jerusalem** …"
44021011: "… So shall the Jews **at Jerusalem** bind the man that owneth this girdle …"



40020017: "… as Jesus was going up **to Jerusalem** …"
42024052: "… they worshipped him , and returned **to Jerusalem** …"
44009026: "… when he was come **to Jerusalem** …"

41003022: "And the scribes that came down **from Jerusalem** said …"
42005017: "… there were Pharisees and doctors of the law sitting by , who were come **out of** every village of Galilee and Judaea and **Jerusalem** …"
42010030: "… A certain man was going down **from Jerusalem** to Jericho …"

Prospects

- Parallel texts offer the possibility for detailed functional comparison across languages
- **The comparison is based on actual examples, so each typological generalisation can be scrutinised by specialists**
- Algorithmic assistance is possible, but manual decisions are needed

2. Dictionaries for historical comparison

- Comparative-historical method
  - use complete knowledge of languages
- Concept-list method ("Swadesh")
  - use small dataset per language (typically ~100 words) and process automatically
- Challenge: use large datasets automatically
  - http://quanthistling.info/data/

Jivaroan

github.com/cysouw/jivaro

- Experiment in collaboratively making an etymological dictionary
- 4 languages, 7 dictionaries, 46 K wordforms

Glottolog

Family: Jivaroan

Glottocode: jiva1245

Classification

open Jivaroan | expand all | collapse all

- Jivaroan (4)
  - Aguaruna
  - Shuaric (3)
    - Achuar-Shiwiar
    - Huambisa
    - Shuar

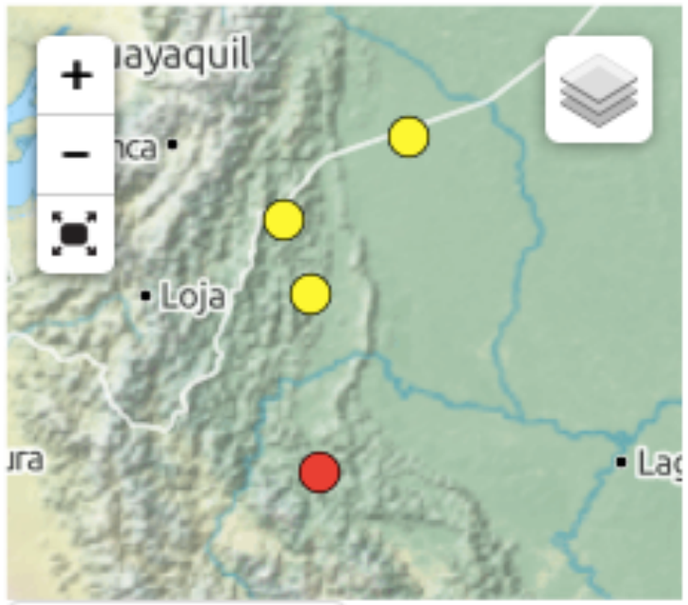
Family membership references

- Adelaar, Willem F. H. and Muysken, Pieter C. 2004

Subclassification references

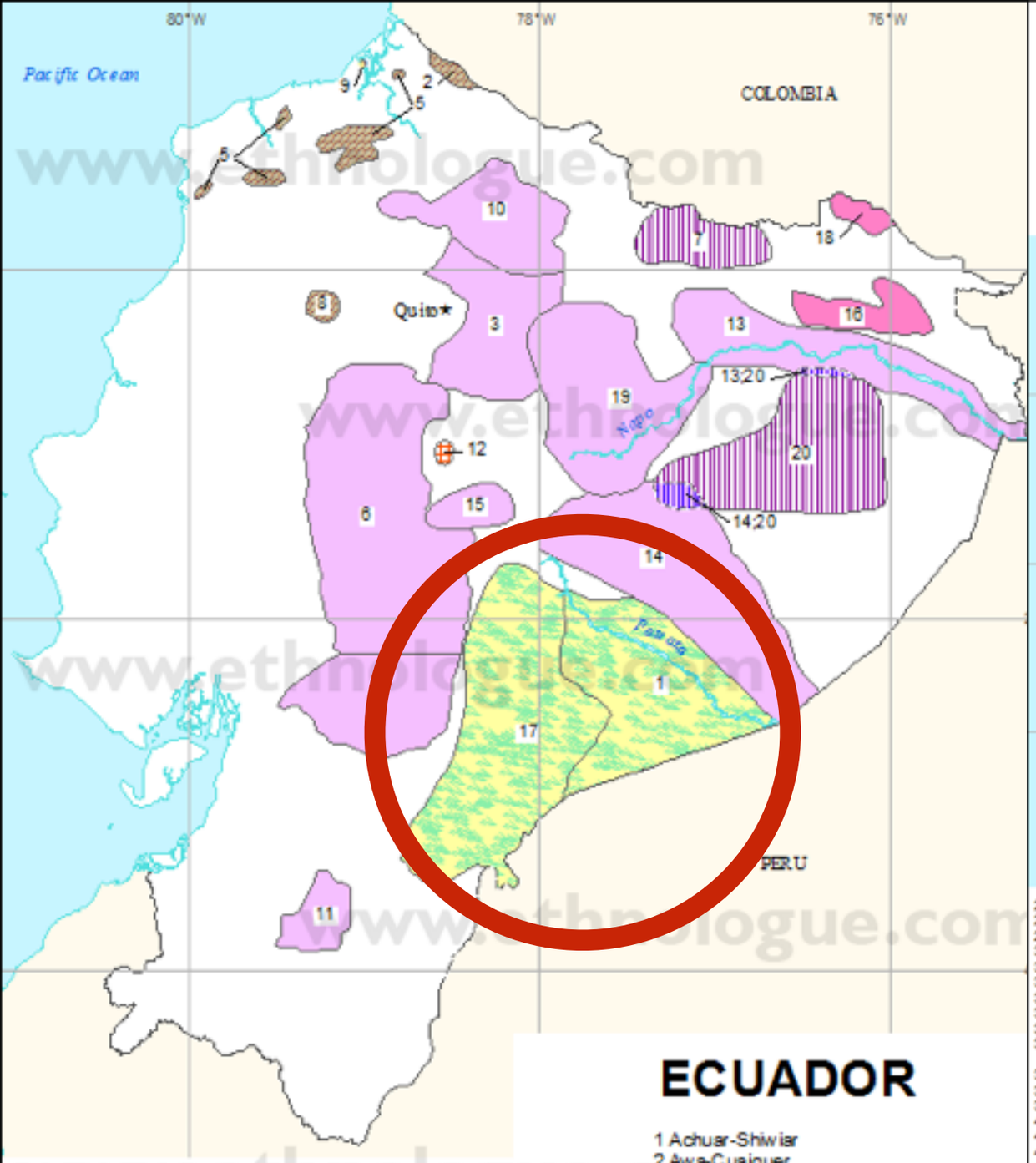
- Overall, Simon 2007

Map



show big map

Links



ECUADOR

1 Achuar-Shiwiar
2 Awa-Cuaiquer
3 Calderón Highland Quichua
4 Cañar Highland Quichua
5 Chachi (2)
6 Chimborazo Highland Quichua
7 Cofán
8 Colorado
9 Epena
10 Imbabura Highland Quichua
11 Loja Highland Quichua
12 Media Lengua
13 Napo Lowland Quichua (2)
14 Northern Pastaza Quichua (2)
15 Salasaca Highland Quichua
16 Secoya
17 Shuar
18 Siona
19 Tena Lowland Quichua
20 Waorani (3)

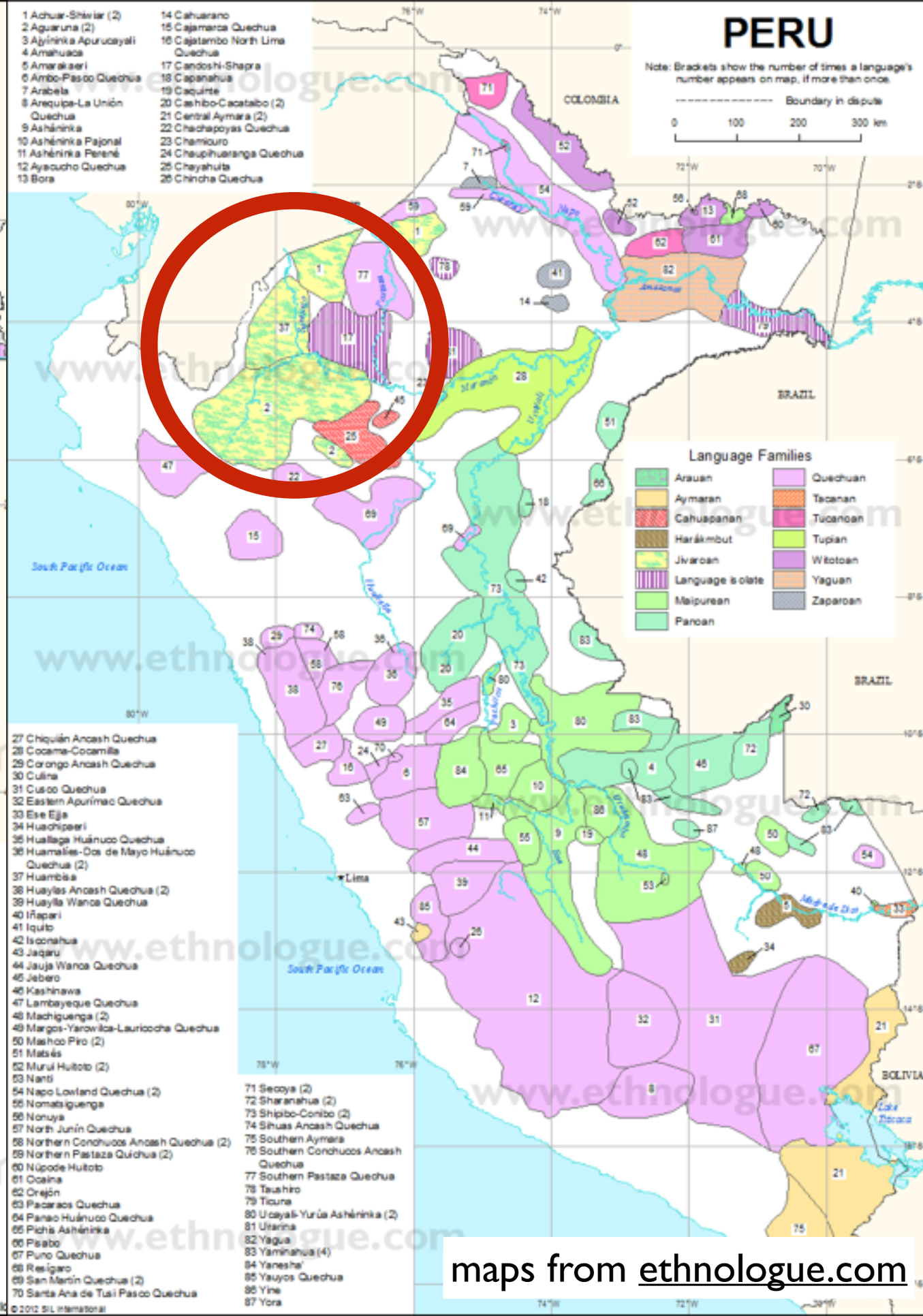
Language Families

- Barbacoan
- Chocoan
- Jivaroan
- Language isolate
- Mixed language
- Quechuan
- Tucanoan

Boundary in dispute

Notes:
1. Dashed lines show overlap of language areas.
2. Brackets show the number of times a language's number appears on map, if more than once.

0 50 100 150 200 km



maps from ethnologue.com

Using the github-approach

- Keep everything in one big file
  - github.com/cysouw/jivaro
- Automatic grouping in ‘cognate sets’
  - using quick-and-dirty sparse matrix algebra from github.com/cysouw/qlcMatrix
- Automatic alignment
  - using LingPy from github.com/lingpy
- Manual corrections using Alignment Editor
  - github.com/digitallinguists/msa-editor

EDICTOR ABOUT DISPLAY CREATE ANALYZE RELOAD

<data.tsv> Showing 1 - 16 of 16 entries START

ID	LANGUAGE	SOURCE	ETYMONID	ALIGNMENT	TRANSLATION
53	Achuar-Shiwiar	fastmowitz2008	21	a ch í - m k a - - t i n -	agarrarse
62	Achuar-Shiwiar	fastmowitz2008	21	a ch i - - - u - - - - - -	agarrador; uno que agarra
8460	Huambisa	jakway2008	21	a ch i - - - - - - - a m u	agarrar; capturar; sacar chonta
8464	Huambisa	jakway2008	21	a ch i k m - a - - - - - u	agarrado
8468	Huambisa	jakway2008	21	a ch i - n - a - - t - - -	agarrar; capturar; sacar chonta
14812	Aguaruna	larson1966	21	a ch í - t - - - - - - - -	agarrar; prender; clavar
17642	Aguaruna	mori1981	21	a ch í - t - - - - - - - -	agarrar
29172	Aguaruna	wipiodeicat1996	21	a ch i - m - á - - t - - -	agarrar; prender
35619	Aguaruna	yagkug1998	21	a ch í - - - - - - - a m u	clavar; hincar un objeto; sacar ...
35637	Aguaruna	yagkug1998	21	a ch i - m - á - - t - - -	agarrarse en algún objeto
35638	Aguaruna	yagkug1998	21	a ch í - m - a - u - - - -	agarrarse en algún objeto
35639	Aguaruna	yagkug1998	21	a ch i - m k á - - t a s a	agarrarse en algún objeto
35644	Aguaruna	yagkug1998	21	a ch i - n - í - - - a m u	agarrarse mutuamente; darse l...
35645	Aguaruna	yagkug1998	21	a ch i - n - í k - t a s a	agarrarse mutuamente; darse l...
35646	Aguaruna	yagkug1998	21	a ch i - n - í - - t - - -	agarrarse mutuamente; darse l...
35649	Aguaruna	yagkug1998	21	a ch í - t - - - - - - - -	clavar; hincar un objeto; sacar ...

MODIFY ENTRY 53/5

EDICTOR ABOUT DISPLAY CREATE ANALYZE RELOAD

<data.tsv> Showing 1 - 16 of 16 entries START

ID	LANGUAGE	SOURCE	ETYM	SLATION
53	Achuar-Shiwiar	fastmowitz2008	21	rse
62	Achuar-Shiwiar	fastmowitz2008	21	dor; uno que agarra
8460	Huambisa	jakway2008	21	; capturar; sacar chonta
8464	Huambisa	jakway2008	21	do
8468	Huambisa	jakway2008	21	; capturar; sacar chonta
14812	Aguaruna	larson1966	21	; prender; clavar
17642	Aguaruna	mori1981	21	
29172	Aguaruna	wipiodeicat1996	21	; prender
35619	Aguaruna	yagkug1998	21	hincar un objeto; sacar …
35637	Aguaruna	yagkug1998	21	se en algún objeto
35638	Aguaruna	yagkug1998	21	a ch í - m - a - u - - - - agarrarse en algún objeto
35639	Aguaruna	yagkug1998	21	a ch i - m k á - - t a s a agarrarse en algún objeto
35644	Aguaruna	yagkug1998	21	a ch i - n - í - - - a m u agarrarse mutuamente; darse l…
35645	Aguaruna	yagkug1998	21	a ch i - n - í k - t a s a agarrarse mutuamente; darse l…
35646	Aguaruna	yagkug1998	21	a ch i - n - í - - t - - - agarrarse mutuamente; darse l…
35649	Aguaruna	yagkug1998	21	a ch í - t - - - - - - - clavar; hincar un objeto; sacar …

ETYMONID "21" links the following 16 entries:

Achuar-Shiwiar	a	ch	í	-	m	k	a	-	-	t	i	n	-
Achuar-Shiwiar	a	ch	í	-	-	-	u	-	-	-	-	-	-
Huambisa	a	ch	i	-	-	-	-	-	-	-	a	m	u
Huambisa	a	ch	i	k	m	-	a	-	-	-	-	-	u
Huambisa	a	ch	i	-	n	-	a	-	-	t	-	-	-
Aguaruna	a	ch	í	-	t	-	-	-	-	-	-	-	-
Aguaruna	a	ch	í	-	t	-	-	-	-	-	-	-	-
Aguaruna	a	ch	i	-	m	-	á	-	-	t	-	-	-
Aguaruna	a	ch	í	-	-	-	-	-	-	-	a	m	u
Aguaruna	a	ch	i	-	m	-	á	-	-	t	-	-	-
Aguaruna	a	ch	í	-	m	-	a	-	u	-	-	-	-
Aguaruna	a	ch	i	-	m	k	á	-	-	t	a	s	a
Aguaruna	a	ch	i	-	n	-	í	-	-	-	a	m	u
Aguaruna	a	ch	i	-	n	-	í	k	-	t	a	s	a
Aguaruna	a	ch	i	-	n	-	í	-	-	t	-	-	-
Aguaruna	a	ch	í	-	t	-	-	-	-	-	-	-	-
IGNORE	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

EDIT SUBMIT EXPORT CLOSE

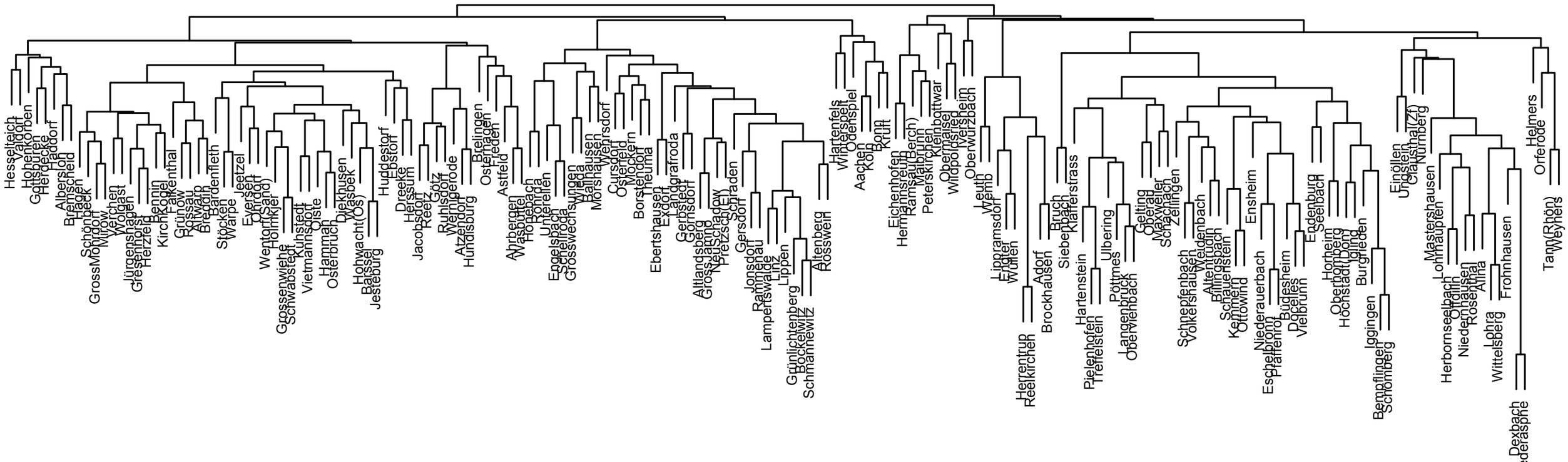
3. German Dialect Data

github.com/cysouw/pad

- 31 K phonetically transcribed words, distributed over 186 meanings and 183 villages

- How to make sense of the details?

- Agglomerative methods are trivial
  - e.g. using sparse matrix algebra from github.com/cysouw/qlcMatrix



“we knew that already …”

“… except that you go wrong here and there!”

Sound Alignment

LOCATION	WORD
Aachen	a̟ːph
Adorf	a̟ːbʱə
Ahrbergen	o̝→o̟phə̟
Albersloh	ɑːpᵝ̈ə̟
Allna	ɑɸh
Altenberg	ʌfɛ̟
Altentrüdin	a̟f
Altlandsberg	ɑʾfəʾ
Altwarp	oːph
Astfeld	pˤːpʰə
Atzendorf	afɛ̟
Ballhausen	ʌʾfə
Bardenfleth	o̟ːp͡ɸ
Barssel	o̟ːpʲə
Bempflingen	afː
Bennin	o̝pʰ
Billingsbach	a̟f
Bockelwitz	ʌ̟və̟
Bonn	ɑːpʼ
Borstendorf	ɤ̝fː
Breddin	ɒːph
Brelingen	ɑ̟f͡βə
Bremscheid	pˤːpʰ̟ə̟
…	…

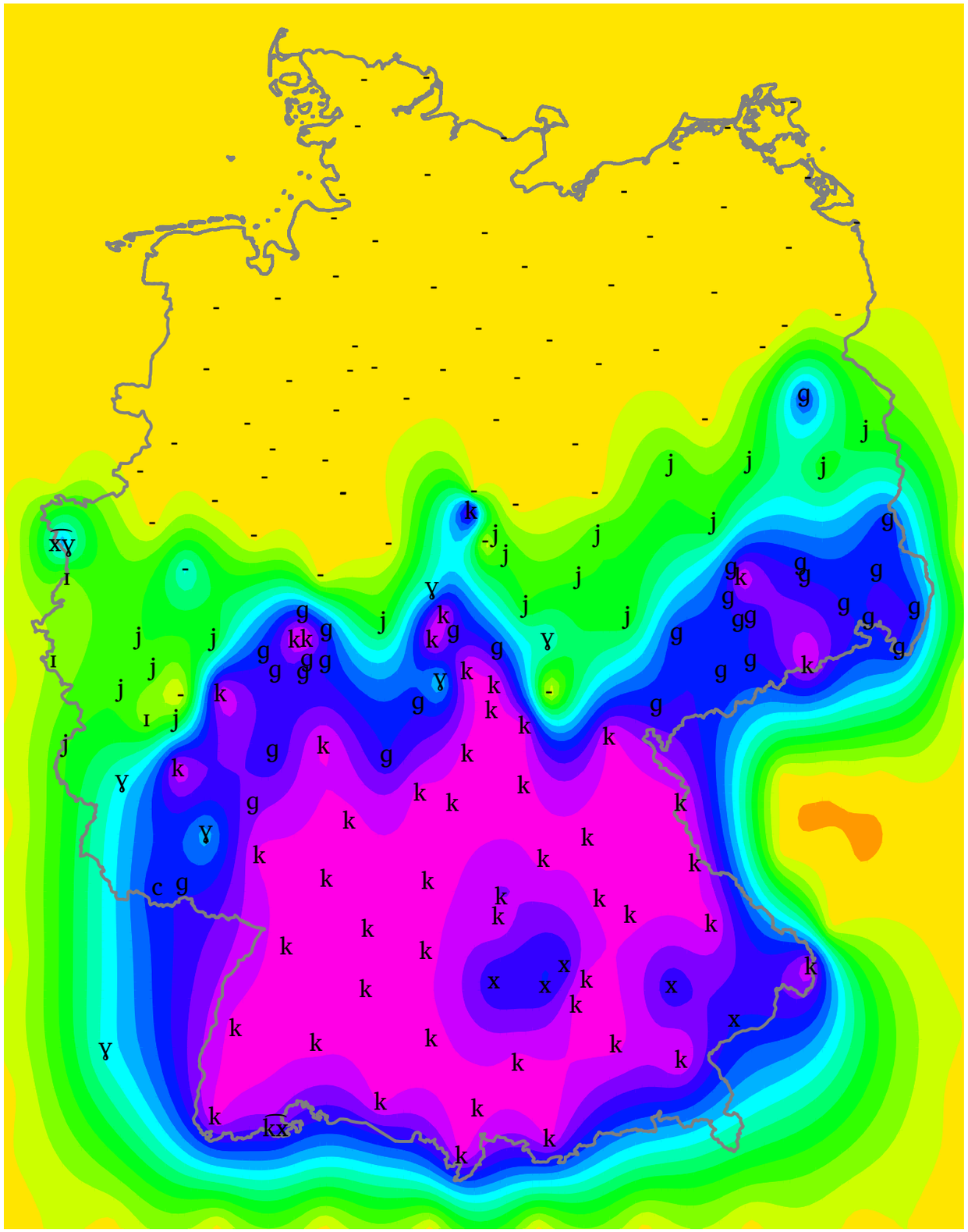
A	FF	E
a̟ː	ph	-
a̟ː	bʱ	ə
o̝→o̟	ph	ə̟
ɑː	pᵝ̈	ə̟
ɑ	ɸh	-
ʌ	f	ɛ̟
a̟	f	-
ɑʾ	f	əʾ
oː	ph	-
pˤː	pʰ	ə
a	f	ɛ̟
ʌʾ	f	ə
o̟ː	p͡ɸ	-
o̟ː	pʲ	ə
a	fː	-
o̝	pʰ	-
a̟	f	-
ʌ̟	v	ə̟
ɑː	pʼ	-
ɤ̝	fː	-
ɒː	ph	-
a̟	f͡β	ə
pˤː	pʰ̟	ə̟
…	…	…

- **Workflow:**
  - Tokenisation of segments using **qlcTokenize** (github.com/cysouw/qlcTokenize)
  - Automatic alignment using **LingPy** (github.com/lingpy)
  - Manual correction using **Alignment Editor** (github.com/digitallinguists/msa-editor)
  - Separation of cognates (e.g. *Samstag* vs. *Sonnabend*)
  - Annotation of columns (e.g. many-to-one alignments, metathesis)

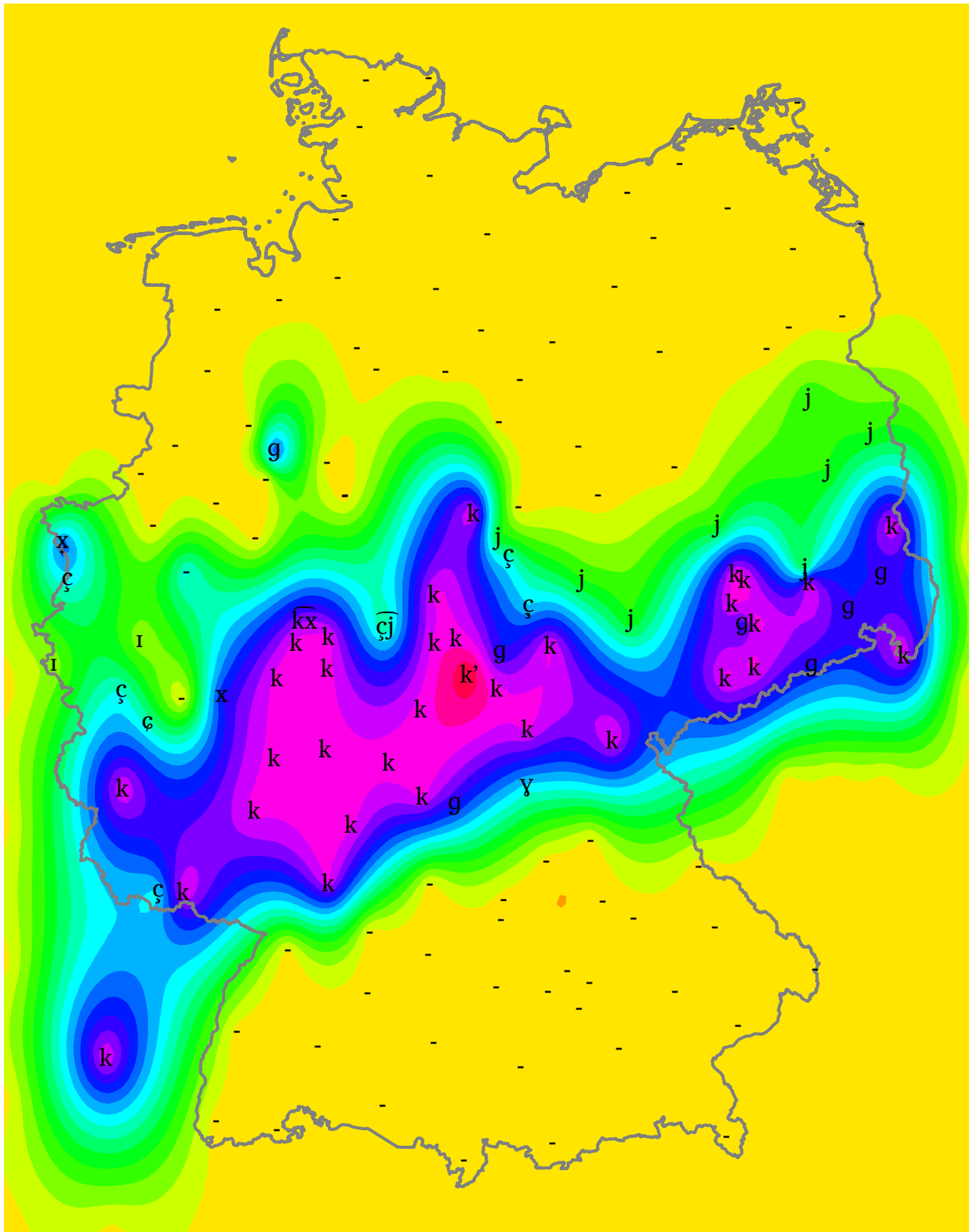
Investigation of Detail

- **Consonant** [g] in German Perfect-prefix <ge>
- Ordered to ‘strength’, visualised as ‘height’
- (NULL) ɣ̣ ɪ j̩ j j̲ ʔ ç͡j̩ z̩ ʑ̩ ʒ ɕ̩ ɣ̩ ç̩ c͡j ɣ ɟ̩ ç̲ ɣ̲ t͡ç g͡ɣ x͡ɣ g̩ x̩ c g k͡ç k͡j x ç̲ x̲ kʰ k͡x k̩ ʀ̥ χ k k̲ q k’

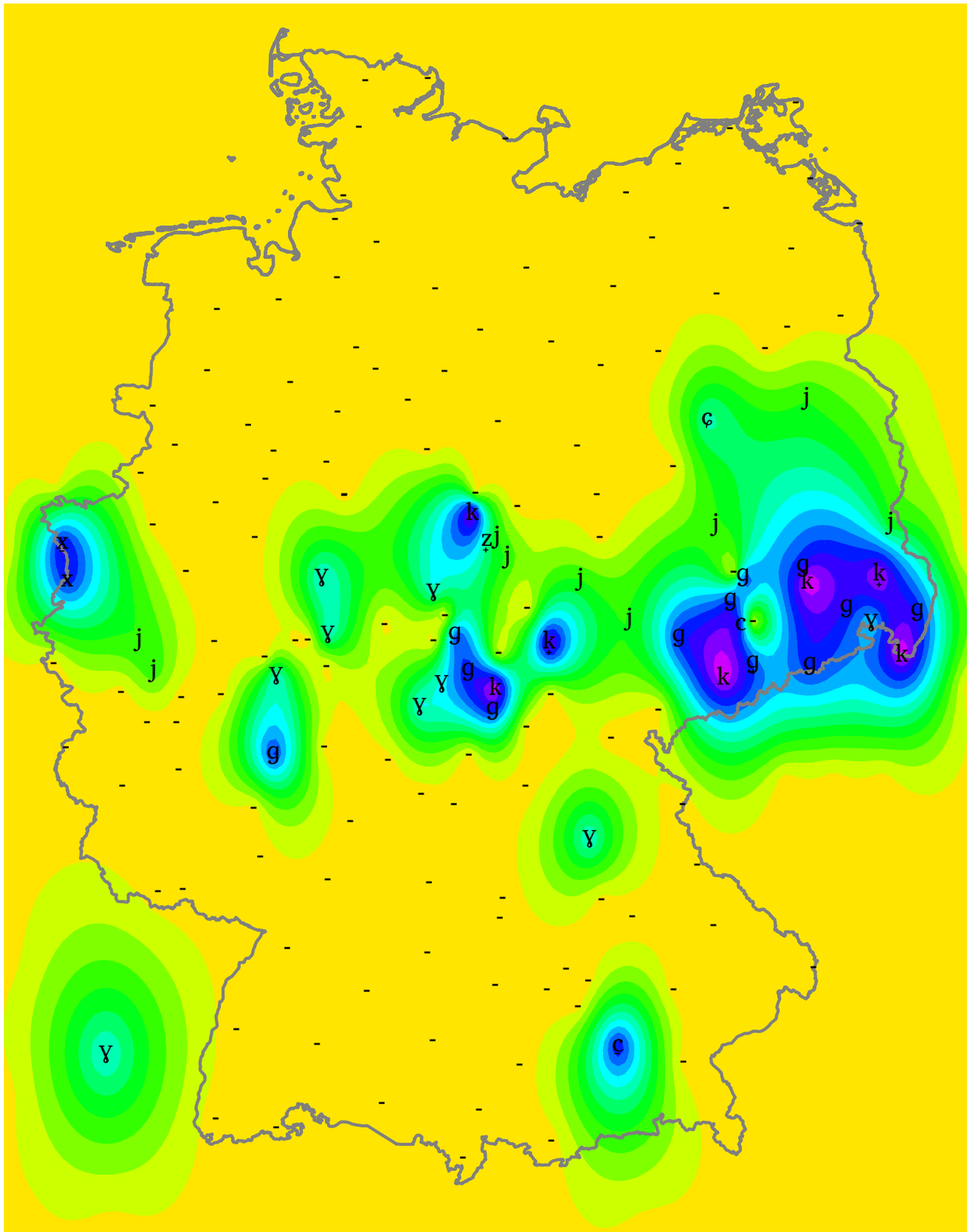
gefahren



gebrannt



gekommen



Big Data and Small Stories

- Big data are a curse and blessing
- Big (agglomerative and approximative) stories and not a step forward
- Large numbers of cases allow for much detail, but the details have to be identified first
- The detail and minutiae that can be told with big data are the real improvement!