Introducing the *Parallel Text Corpus*

Michael Cysouw
Philipps-Universität Marburg

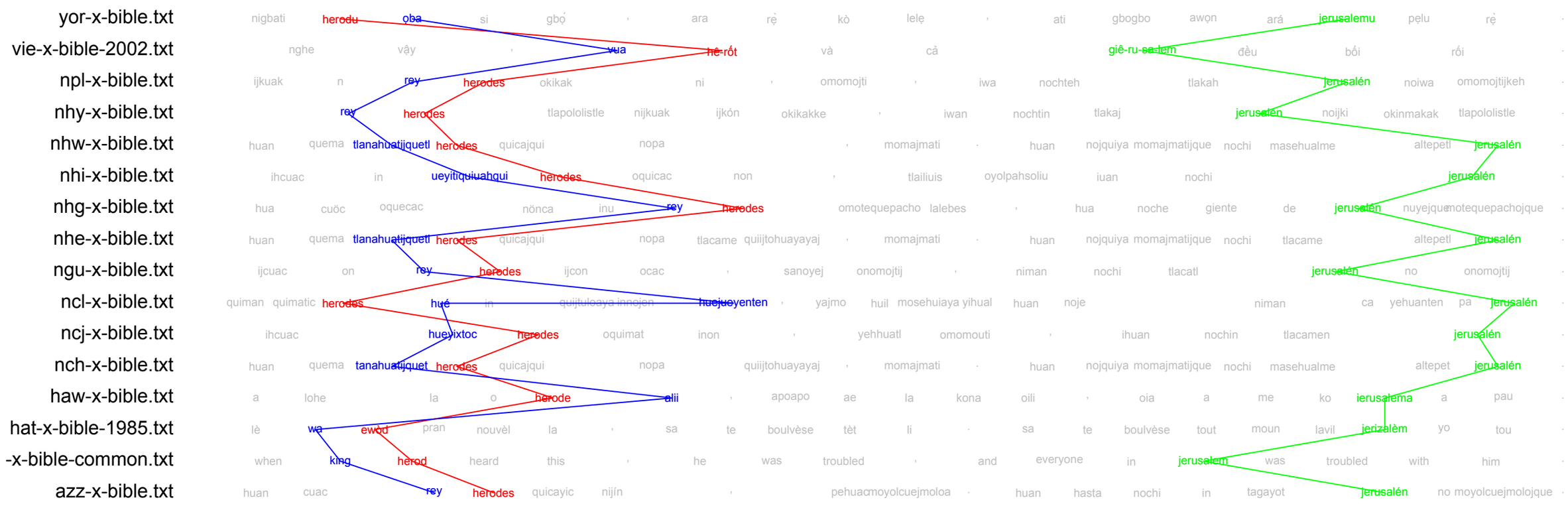
Parallel Text Corpus

- Universal Declaration of Human Rights
- Pamphlets of the Jehovas Witnesses
- Bible
- currently a private github repo: but I invite everybody to become a member

Parallel Bible Corpus

- 1169 translations online (soon 1600+)
- 906 different ISO-639/3 codes (soon 1300+)
- In total more than 350 Million tokens
- More than 17 Million different wordforms
- http://paralleltext.info/data

Demo

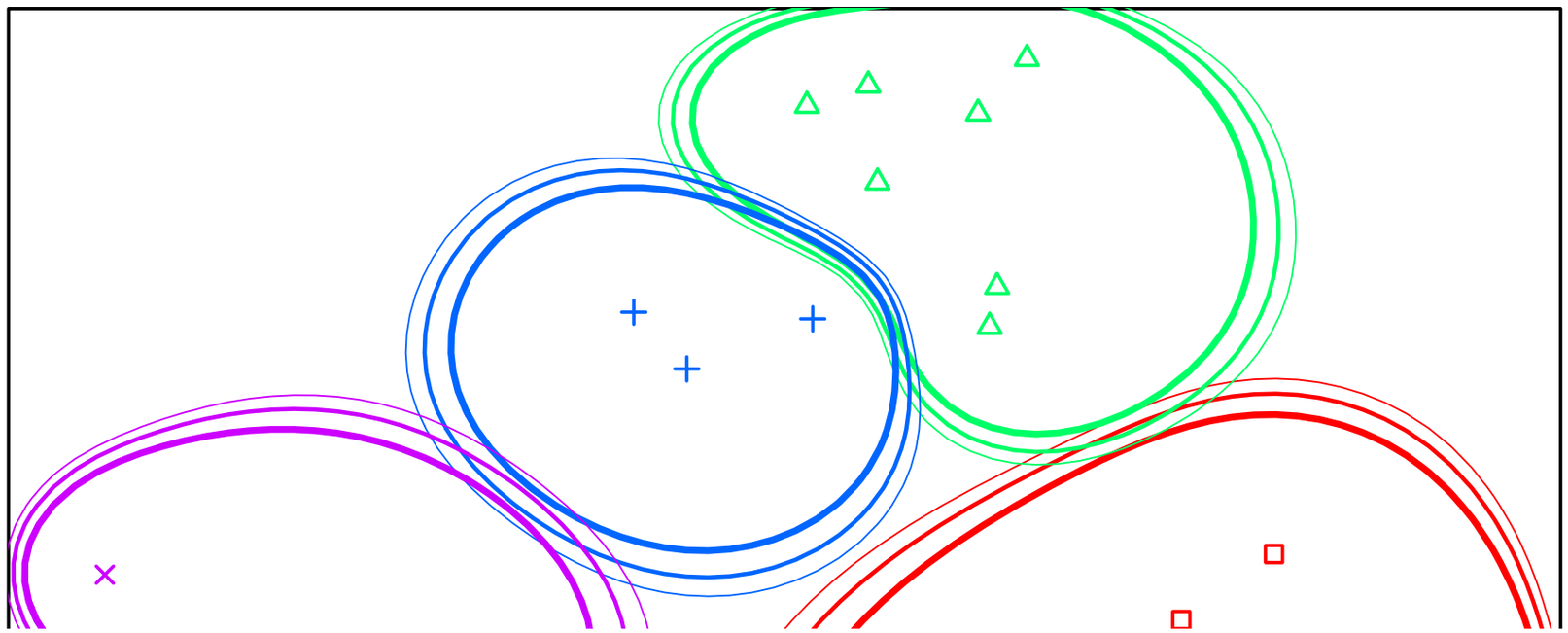

yor-x-bible.txt
nigbati herodu oba si gbọ́ , ara rẹ̀ kò lelẹ , ati gbogbo awọn ará jerusalemu pẹlu rẹ̀ .
vie-x-bible-2002.txt
nghe vậy , vua hê-rốt và cả giê-ru-sa-lem đều bối rối .
npl-x-bible.txt
ijkuak n rey herodes okikak ni , omomojti , iwa nochteh tlakah jerusalén noiwa omomojtijkeh .
nhy-x-bible.txt
rey herodes tlapololistle nijkuak ijkón okikakke , iwan nochtin tlakaj jerusalén noijki okinmakak tlapololistle .
nhw-x-bible.txt
huan quema tlanahuatijquetl herodes quicajqui nopa , momajmati . huan nojquiya momajmatijque nochi masehualme altepetl jerusalén .
nhi-x-bible.txt
ihcuac in ueyitiquiuahqui herodes oquicac non , tlailiuis oyolpahsoliu iuan nochi jerusalén .
nhg-x-bible.txt
hua cuöc oquecac nönca inu rey herodes omotequepacho lalebes , hua noche giente de jerusalén nuyejque motequepachojque .
nhe-x-bible.txt
huan quema tlanahuatijquetl herodes quicajqui nopa tlacame quiijtohuayayaj , momajmati . huan nojquiya momajmatijque nochi tlacame altepetl jerusalén .
ngu-x-bible.txt
ijcuac on rey herodes ijcon ocac , sanoyej onomojtij , niman nochi tlacatl jerusalén no onomojtij .
ncl-x-bible.txt
quiman quimatic herodes hué in quijtuloaya innojen huejueyenten , yajmo huil mosehuiaya yihual huan noje niman ca yehuanten pa jerusalén .
ncj-x-bible.txt
ihcuac hueyixtoc herodes oquimat inon , yehhuatl omomouti , ihuan nochin tlacamen jerusalén .
nch-x-bible.txt
huan quema tanahuatijquet herodes quicajqui nopa quiijtohuayayaj , momajmati . huan nojquiya momajmatijque nochi masehualme altepet jerusalén .
haw-x-bible.txt
a lohe la o herode alii , apoapo ae la kona oili , oia a me ko ierusalema a pau .
hat-x-bible-1985.txt
lè wa ewòd pran nouvèl la , sa te boulvèse tèt li . sa te boulvèse tout moun lavil jerizalèm yo tou .
-x-bible-common.txt
when king herod heard this , he was troubled , and everyone in jerusalem was troubled with him .
azz-x-bible.txt
huan cuac rey herodes quicayic nijín , pehuac moyolcuejmoloa . huan hasta nochi in tagayot jerusalén no moyolcuejmolojque .

Multiple Alignment

- Small-scale experiment
  - use *fastalign* for bitext-alignment on all pairs
  - build multi-text-alignment using graph clustering
- Only for 77 Germanic translations
- New Testament produced almost 100.000 Germanic alignments, which are directly comparable ‘words’

trees and wood

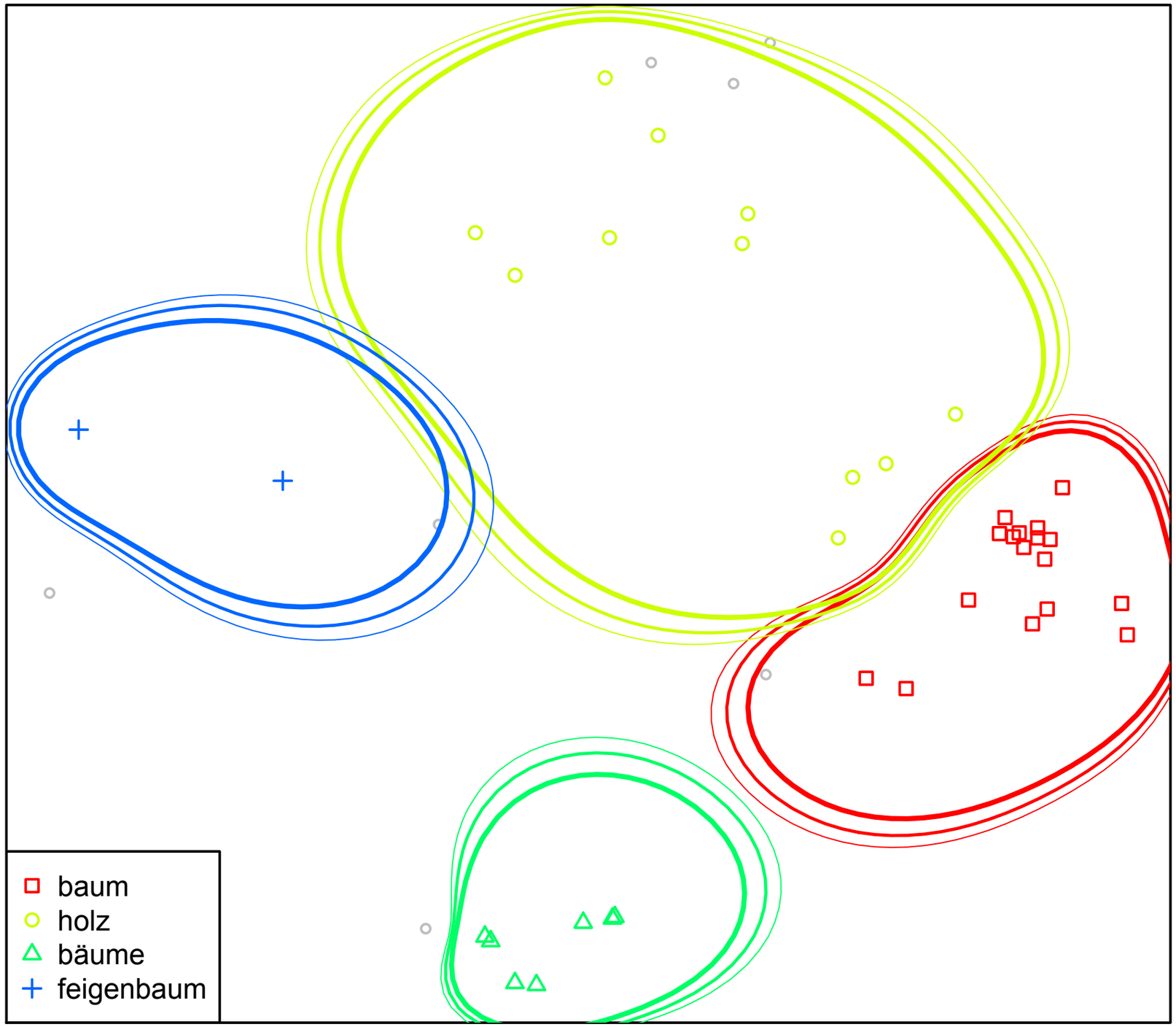
afr-x-bible-1953.txt



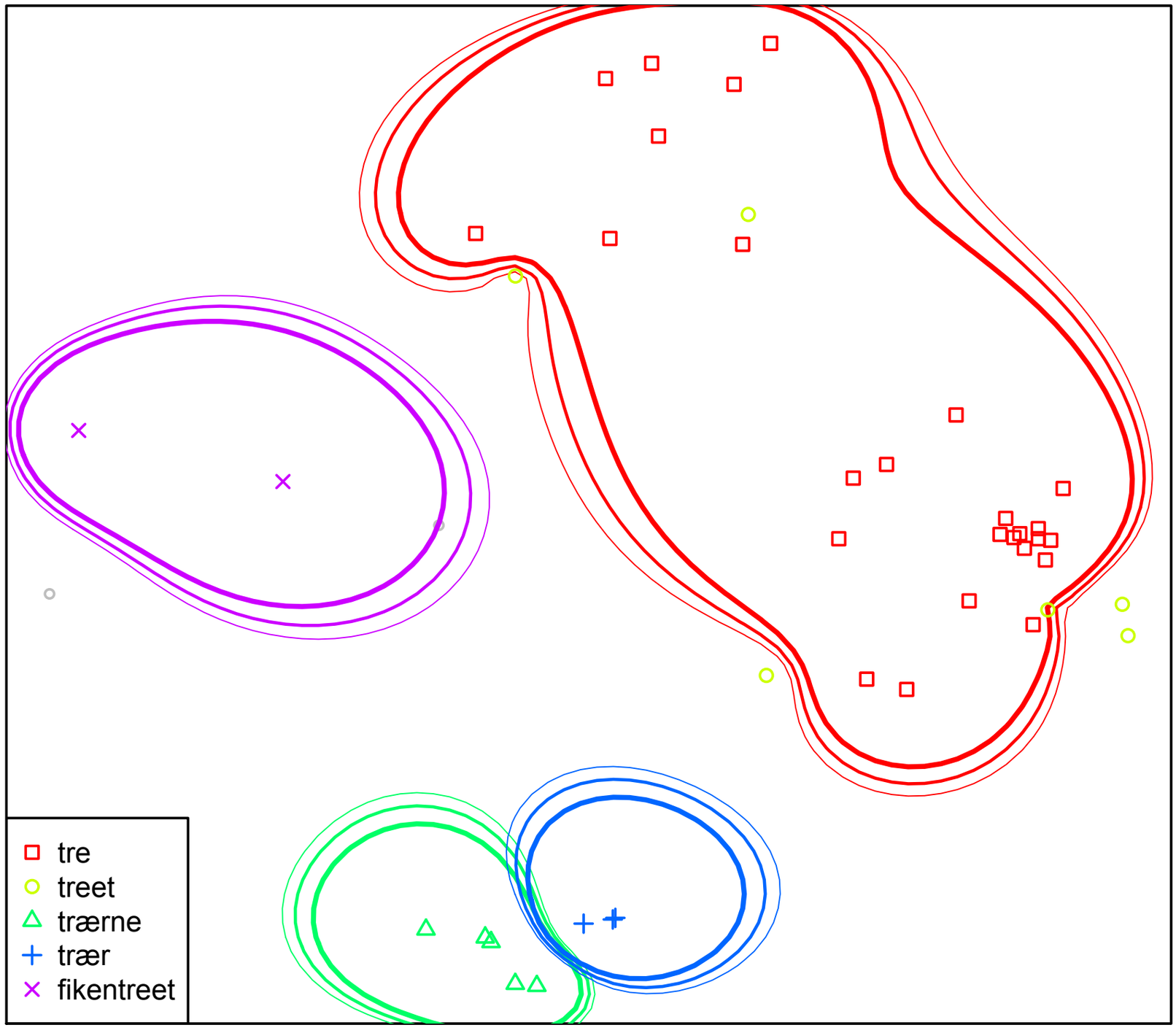
	tree	wood (stuff)	firewood	small forest	large forest
German	*Baum*	*Holz*		*Wald*	
Danish	*træ*			*skov*	
French	*arbre*	*bois*		*forêt*	
Spanish	*árbol*	*madera*	*leña*	*bosque*	*selva*

Louis Hjelmslev
Prolegomena to a Theory of Language (1963)

deu-x-bible-erben.txt

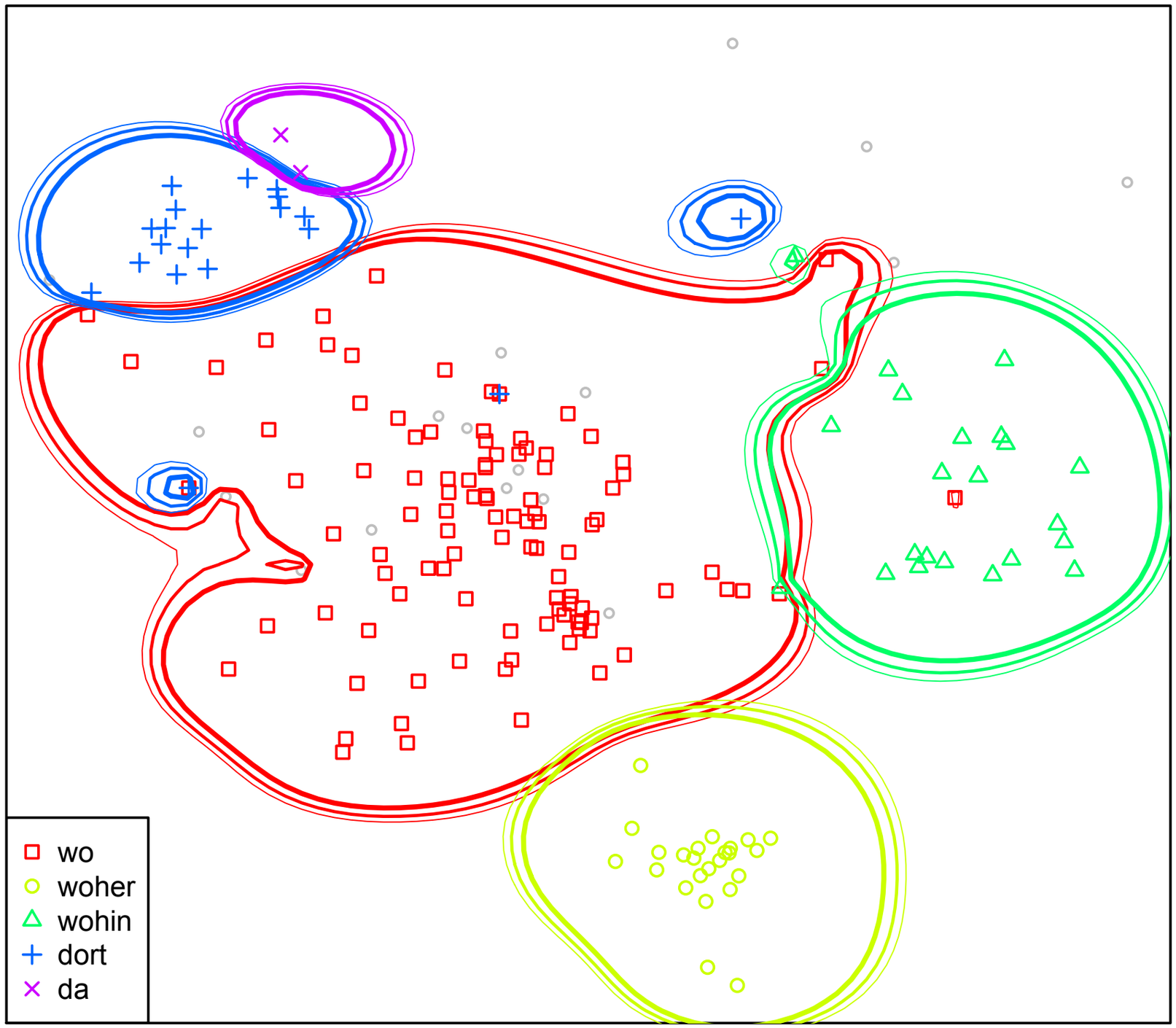


nob-x-bible-2007.txt

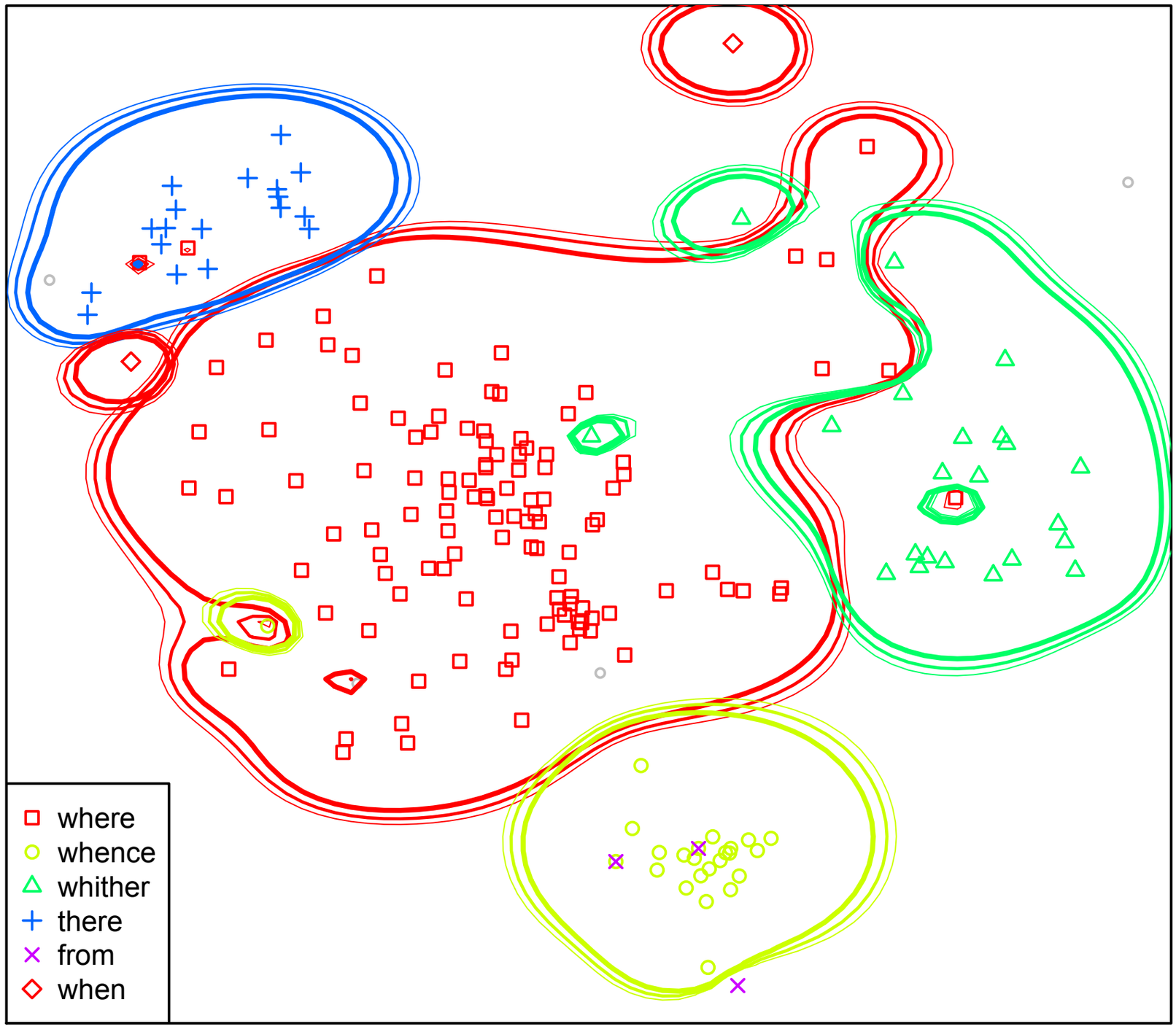


where

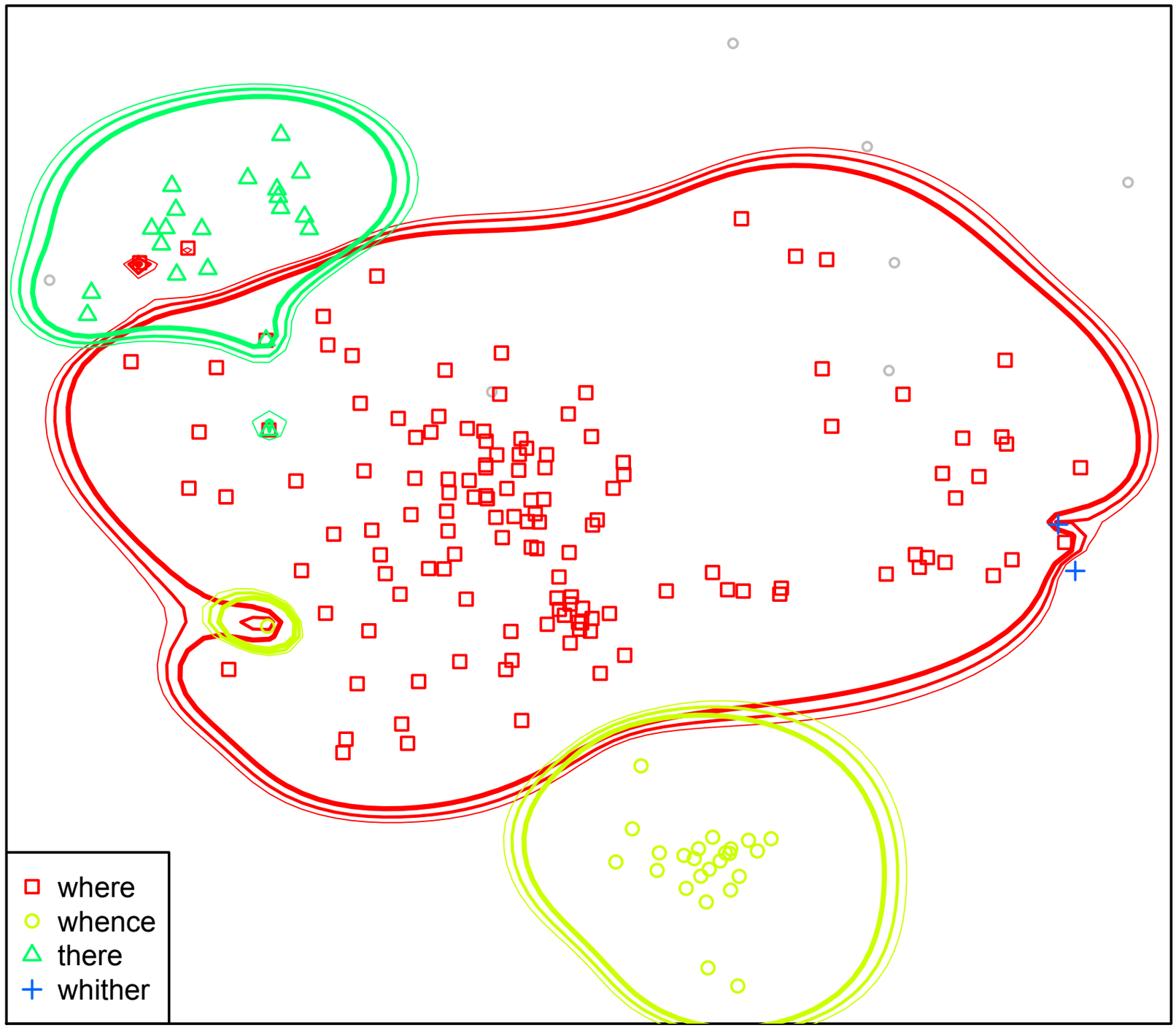
deu-x-bible-pattloch.txt



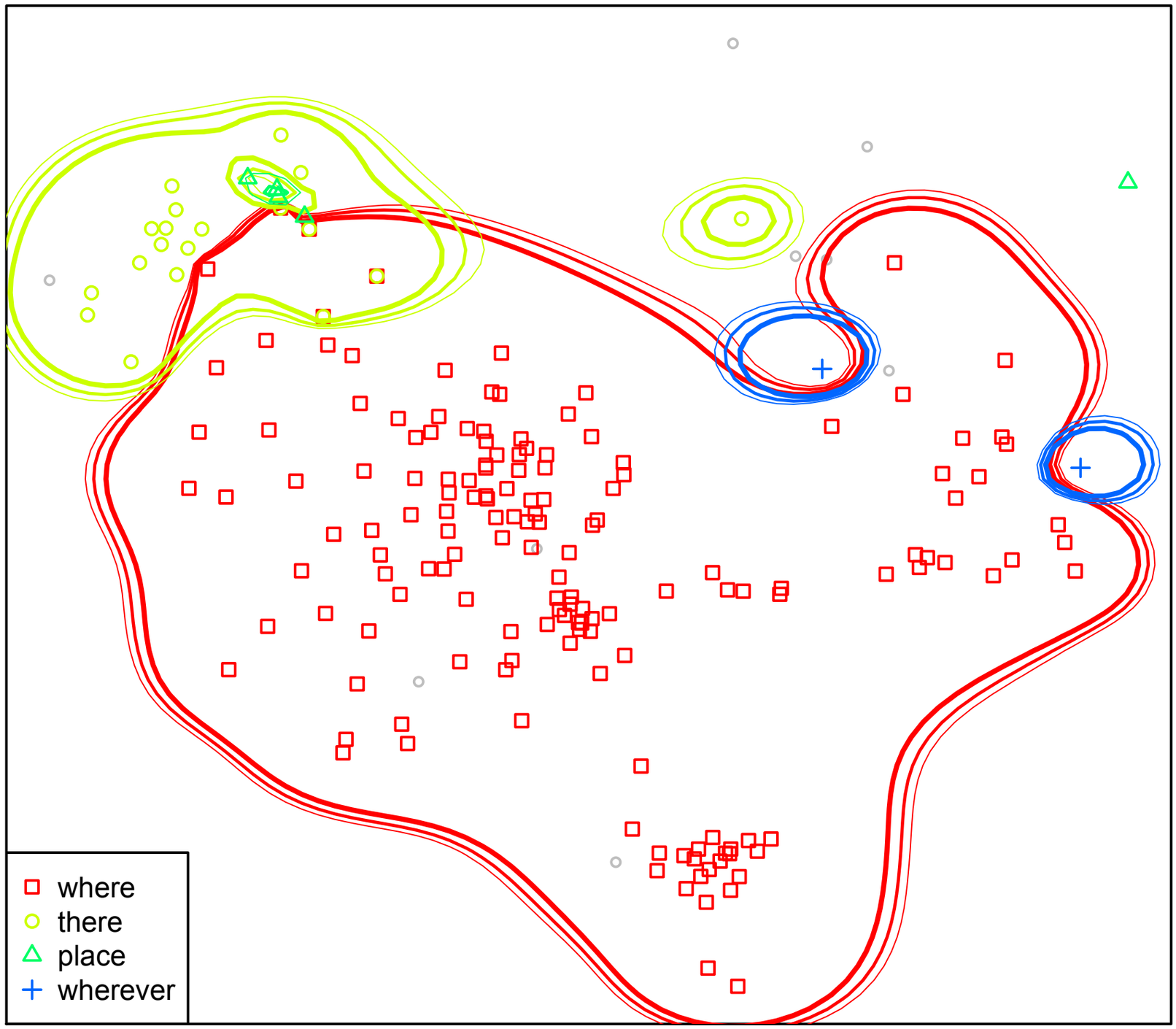
eng-x-bible-kingjames.txt



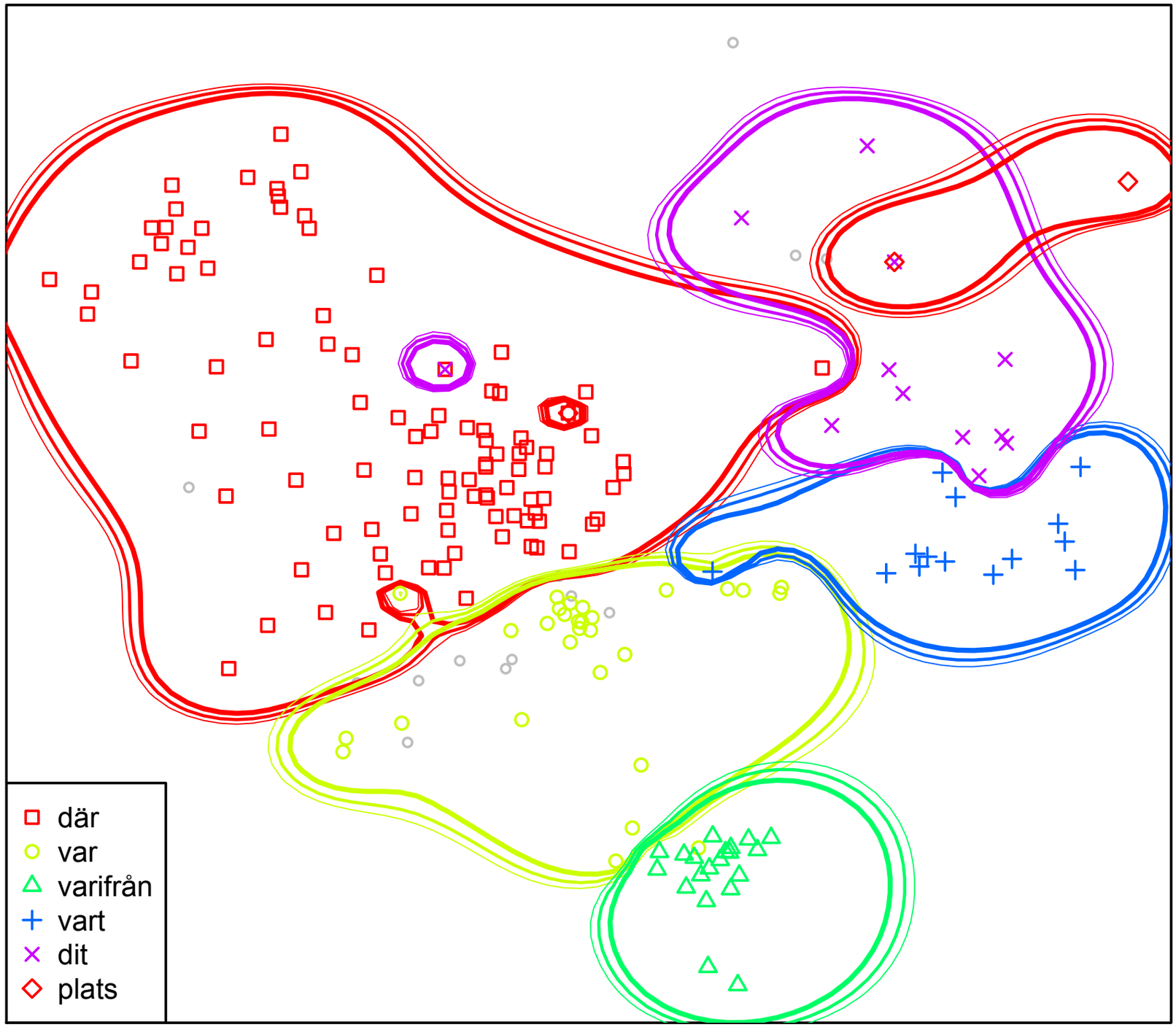
eng-x-bible-darby.txt



eng-x-bible-treeoflife.txt



swe-x-bible-folk1998.txt



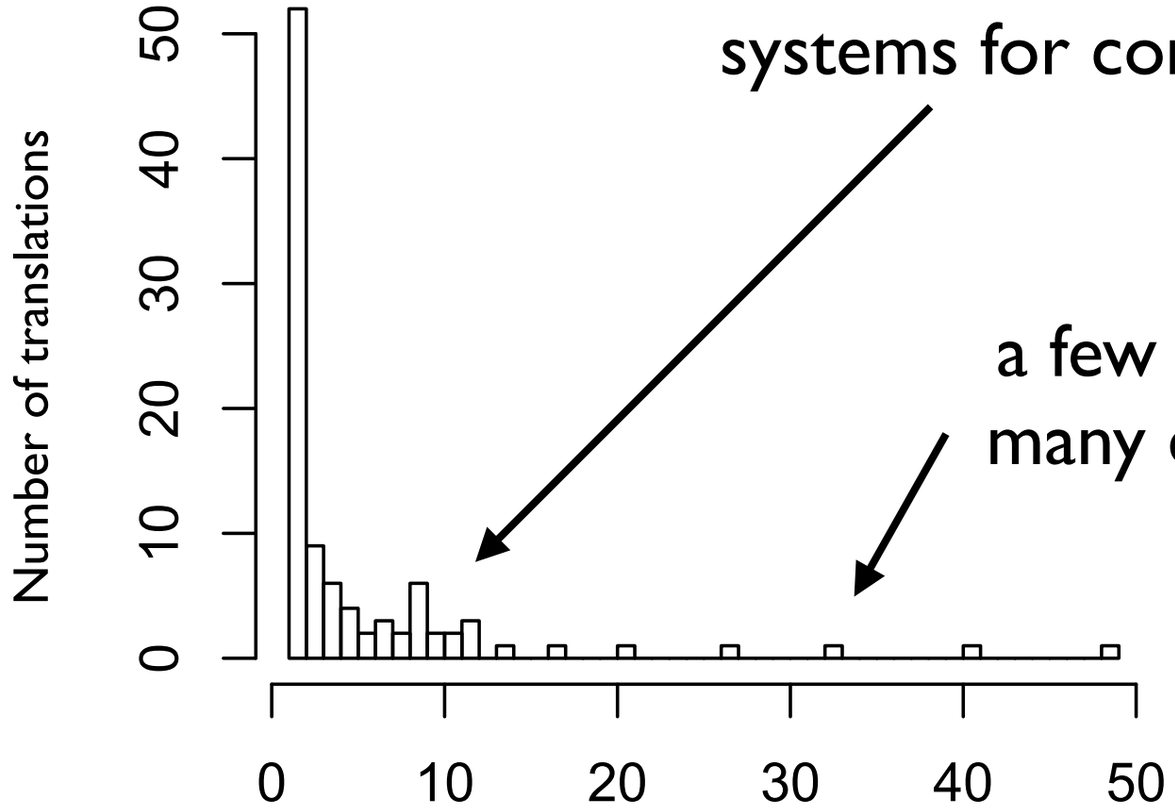
quick test: translation of "*Jerusalem*"

- 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 local case functions

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



Bamanankan

("Bambara", ISO 639-3 bam, spoken in Mali)

Word ID	Bamanankan	Best English back-translation	Best German back-translation	Relevance
1441	jerusalɛm	jerusalem	jerusalem	1
1443	jerusalɛmkaw	jerusalem	jerusalem	0,67
1442	jerusalɛmka	---	---	0,6

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
- ከኢየሩሳሌምም
- የኢየሩሳሌም
- ለኢየሩሳሌም
- ለኢየሩሳሌምም
- የኢየሩሳሌምንም
- የኢየሩሳሌምም

Highland Puebla Náhuatl

(ISO 639-3 azz)

- Jerusalén
- Jerusaléncopaca

Huasteca Náhuatl

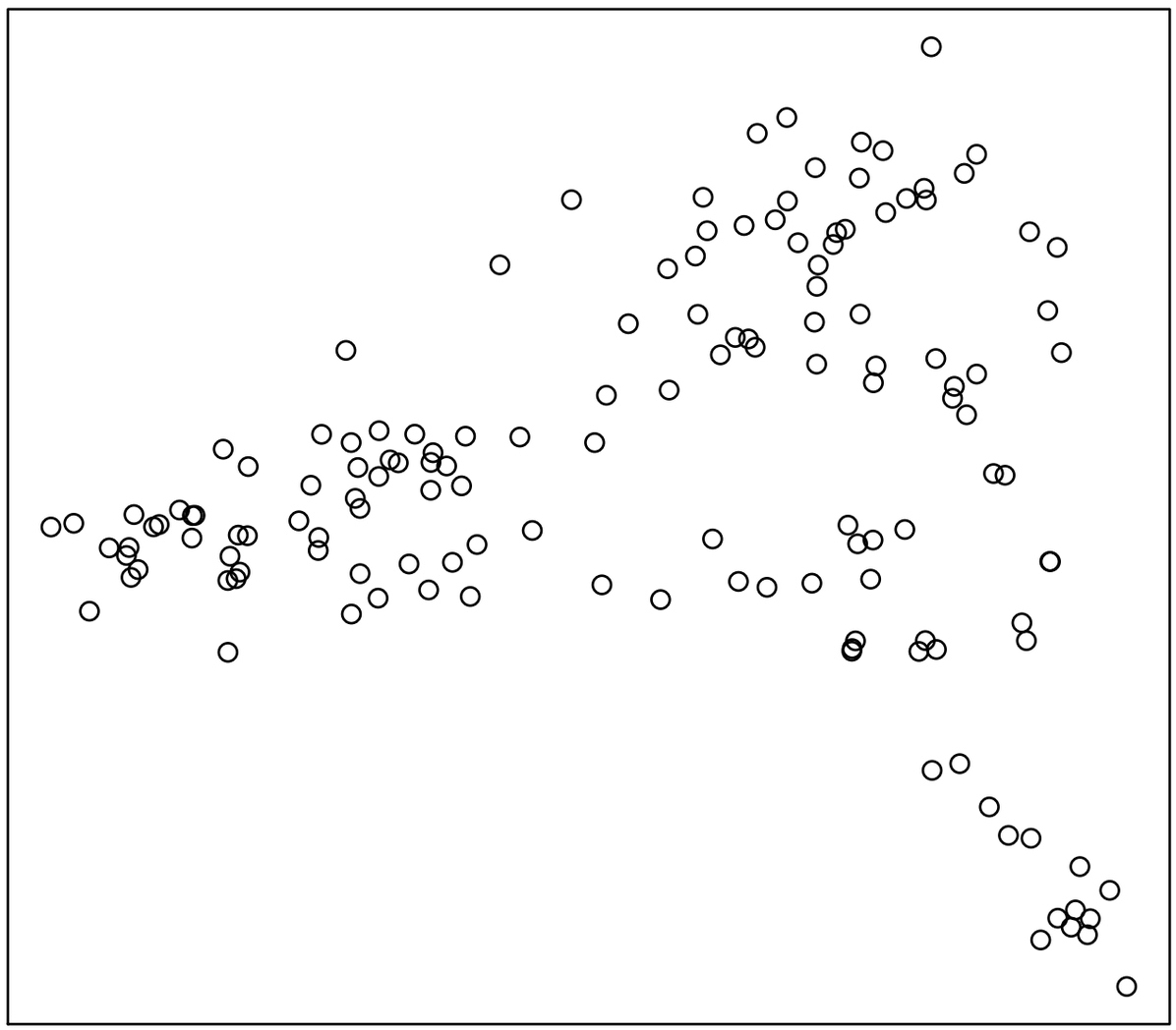
(ISO 639-3 nch)

- Jerusalén
- anJerusalén

98 languages, in total 520 different wordforms

I selected 167 verses including *Jerusalem*
only once in more than 40 languages

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



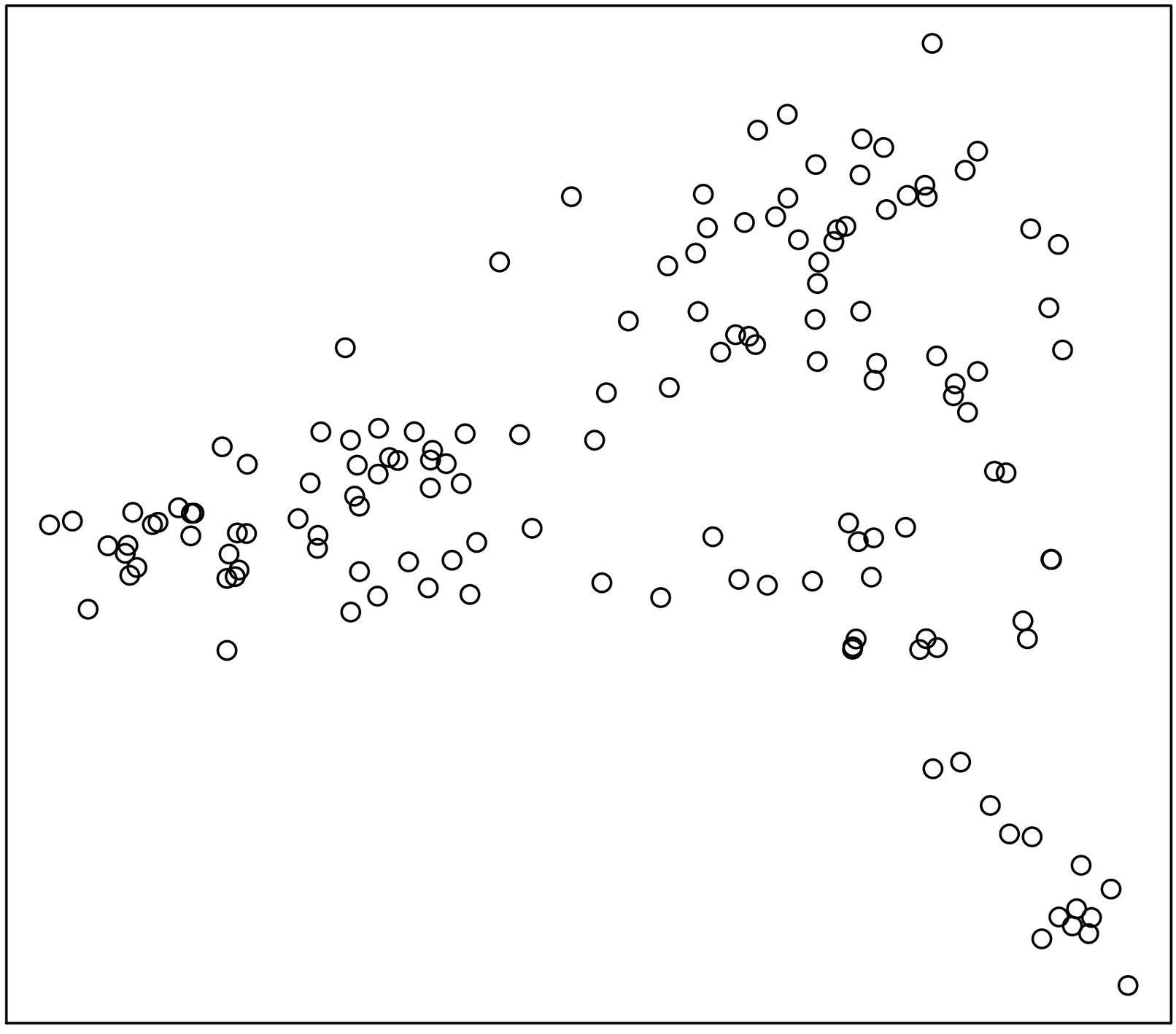
Two contexts of *Jerusalem*
are similar when they often
share the same wordform in
language after language

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

System Typology

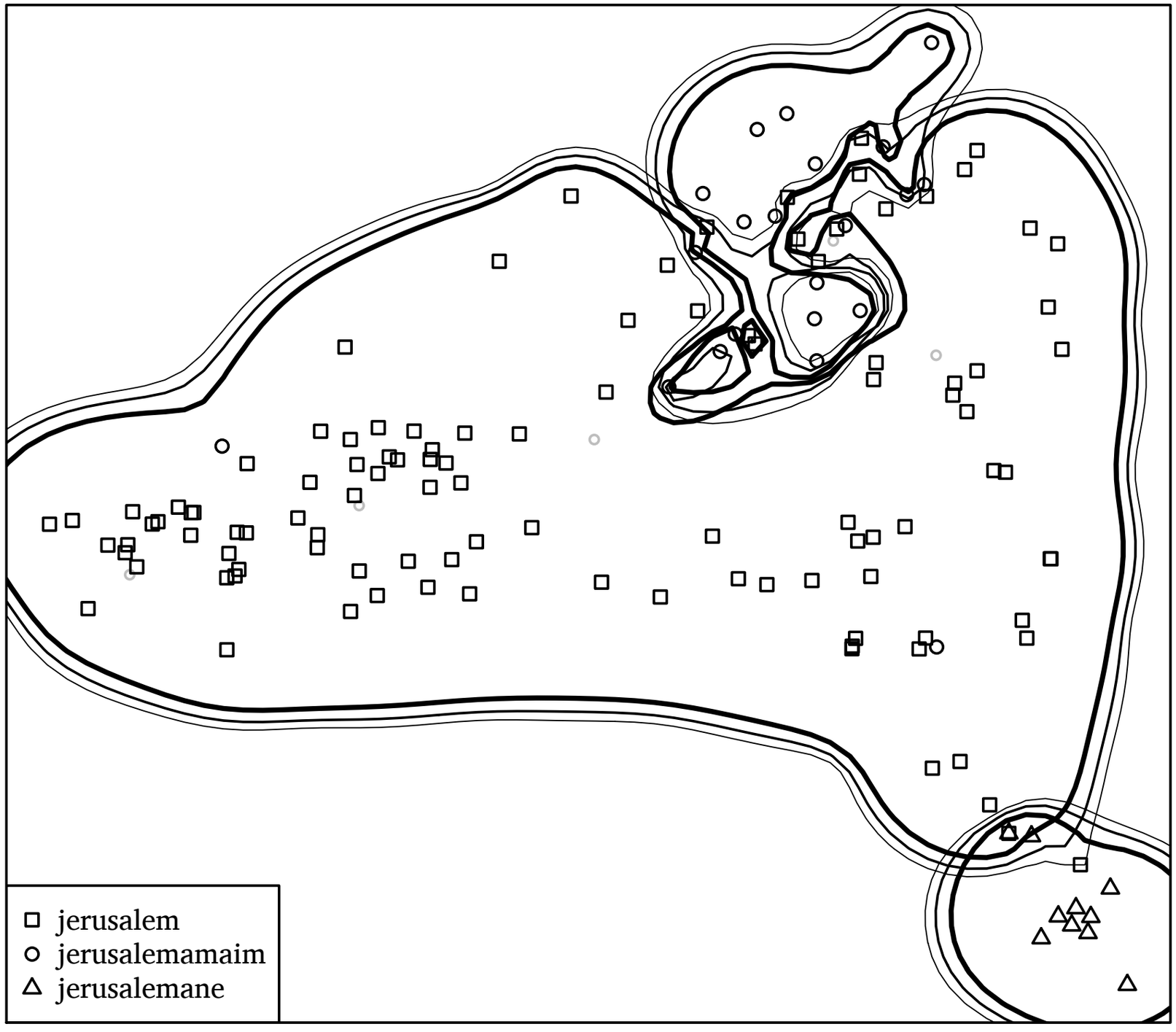
- Taking the different wordforms in a language as a system of marking
- How similar are the systems functionally?
- We can compare functions through distribution in the text



aey

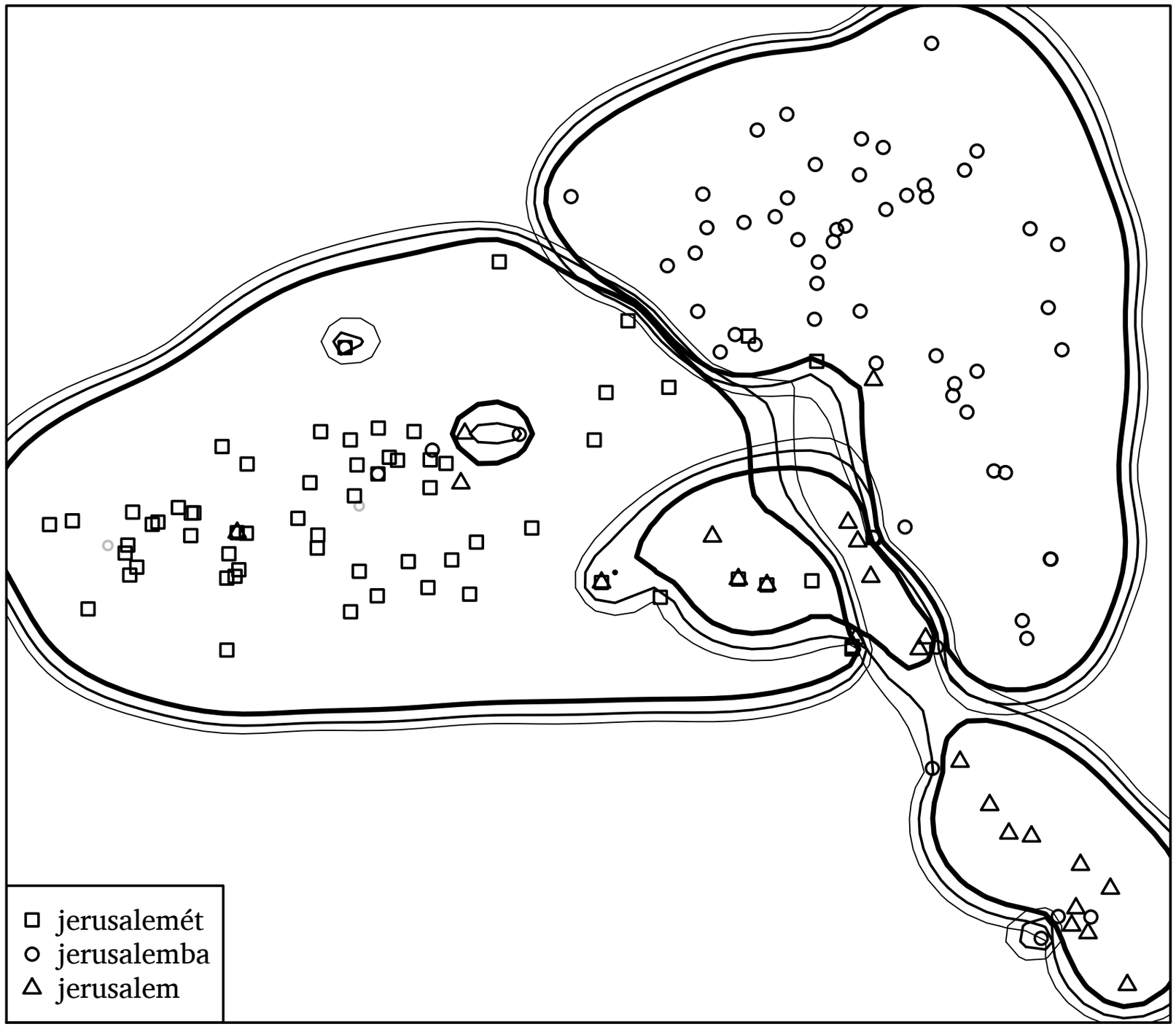
Amele (A language of Papua New Guinea)

aai



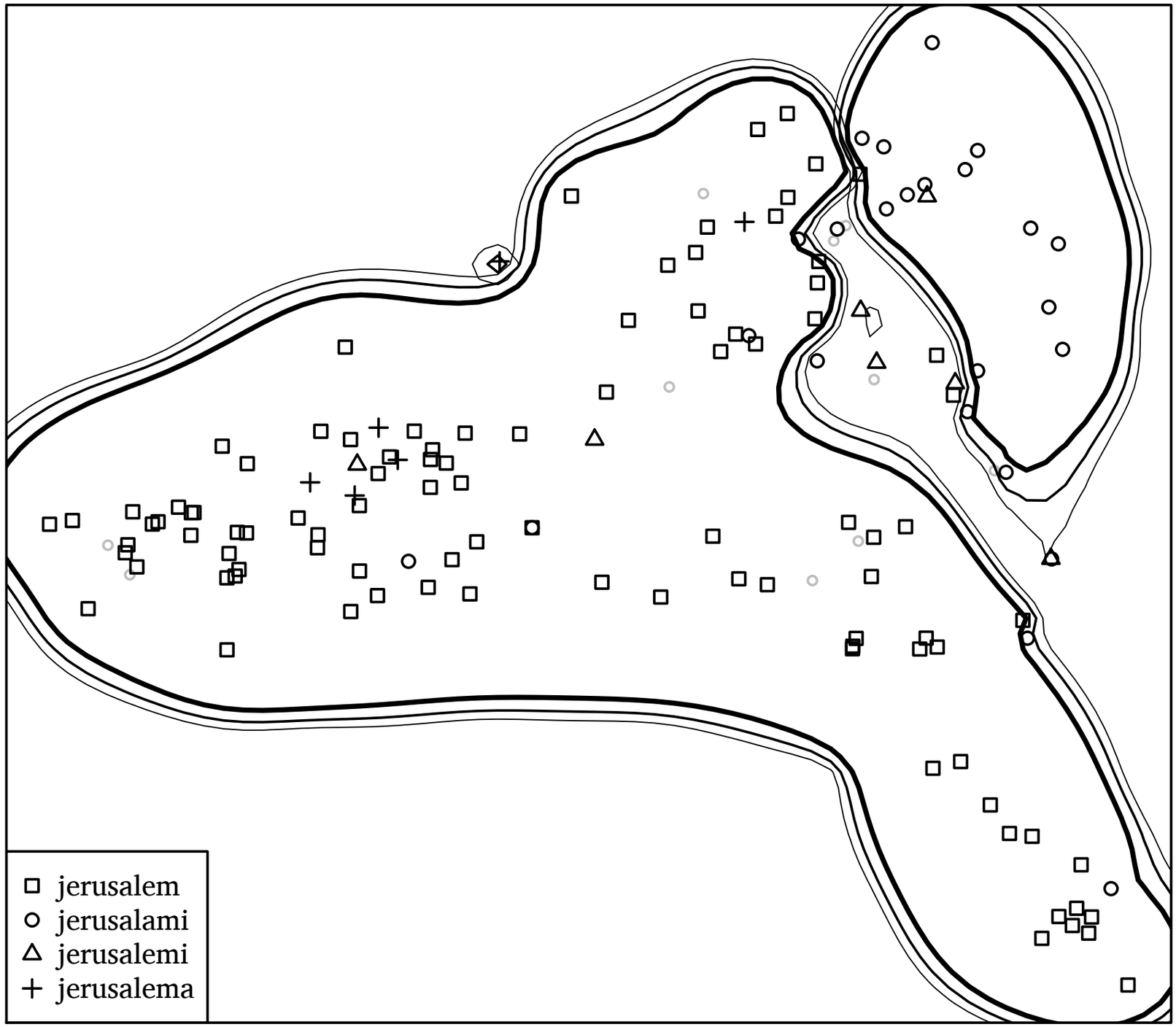
Arifama-Miniafia (a language of Papua New Guinea)

abt



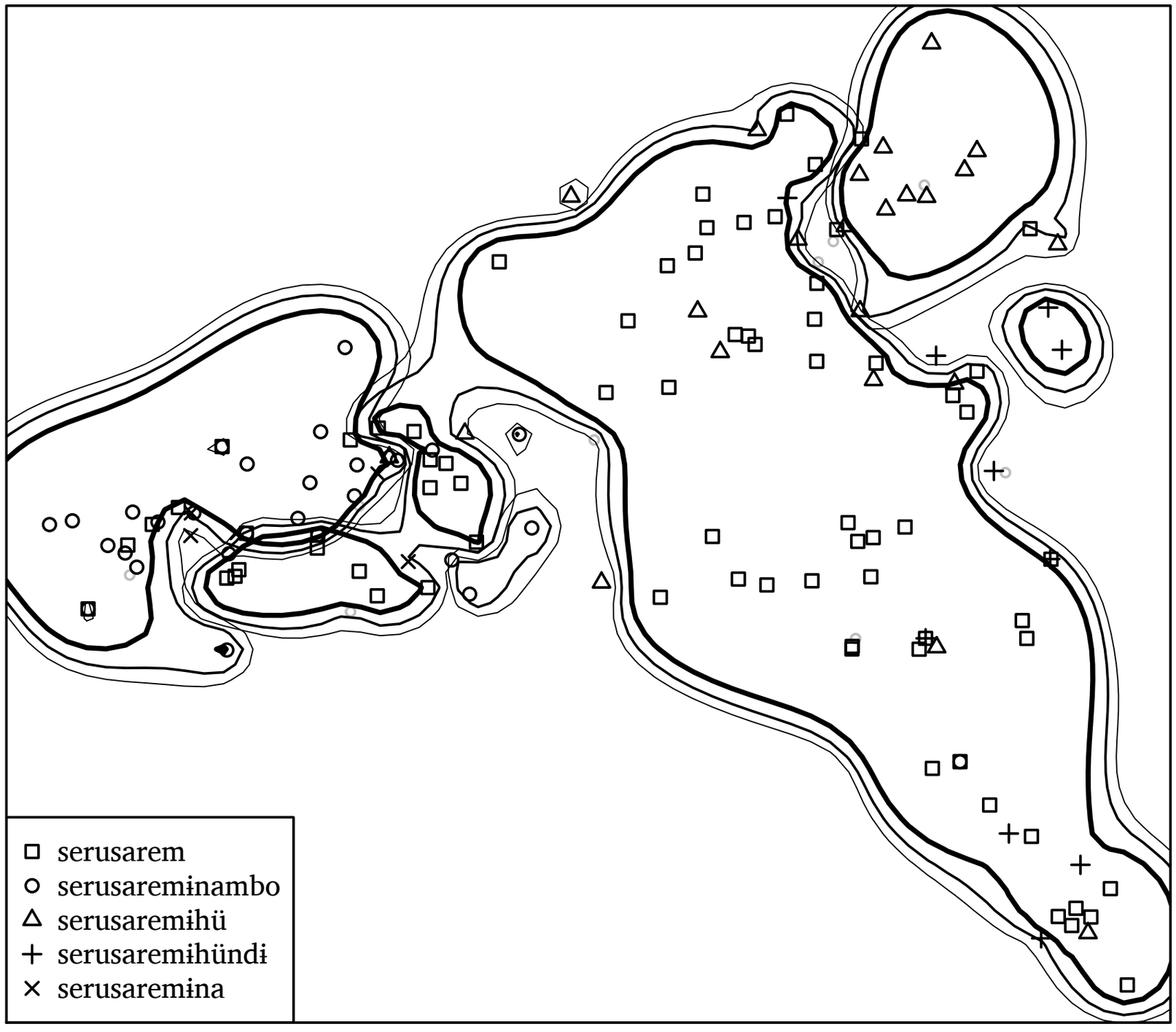
Ambulas (a language of Papua New Guinea)

aoj



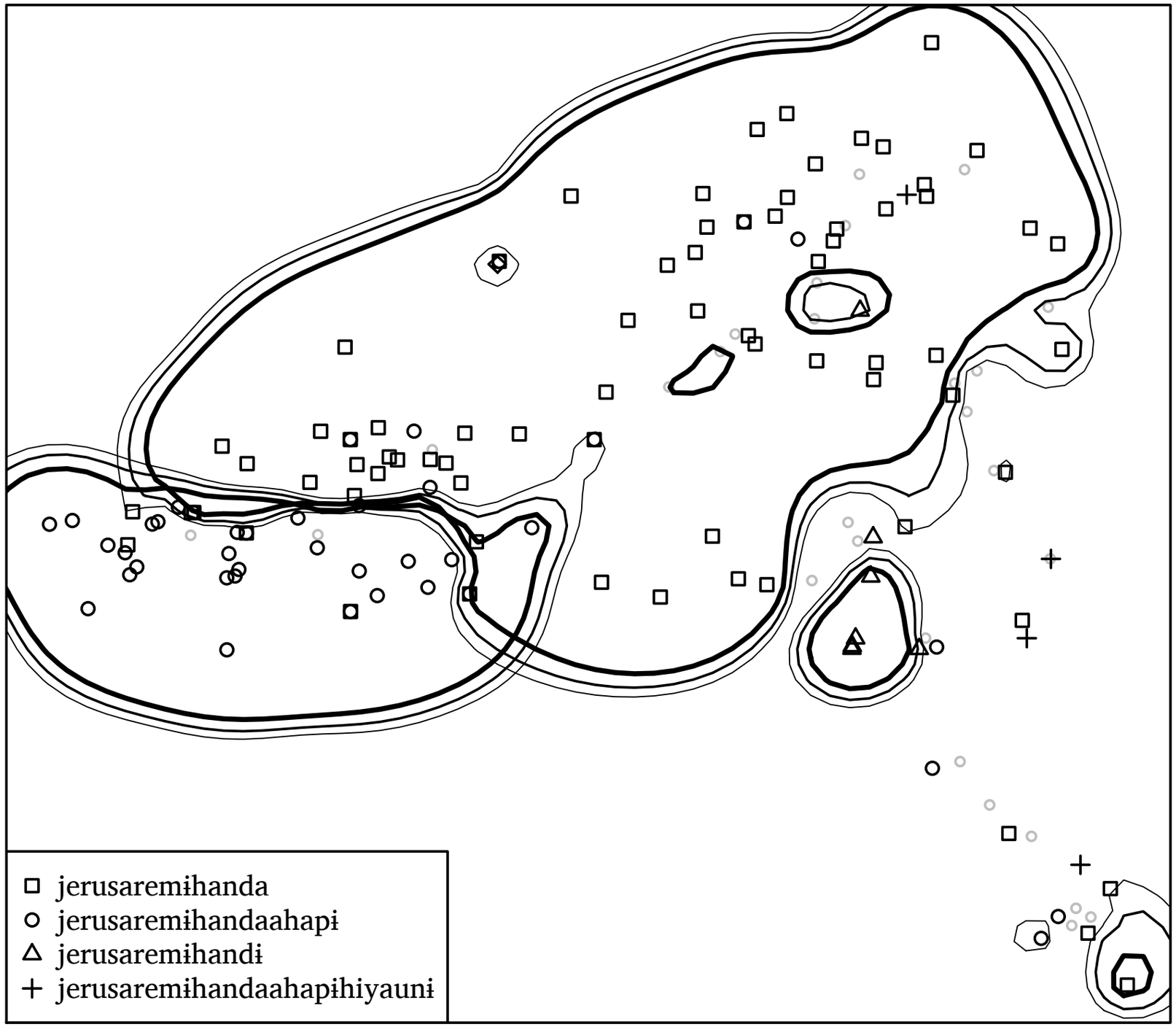
Muffian (a language of Papua New Guinea)

agg



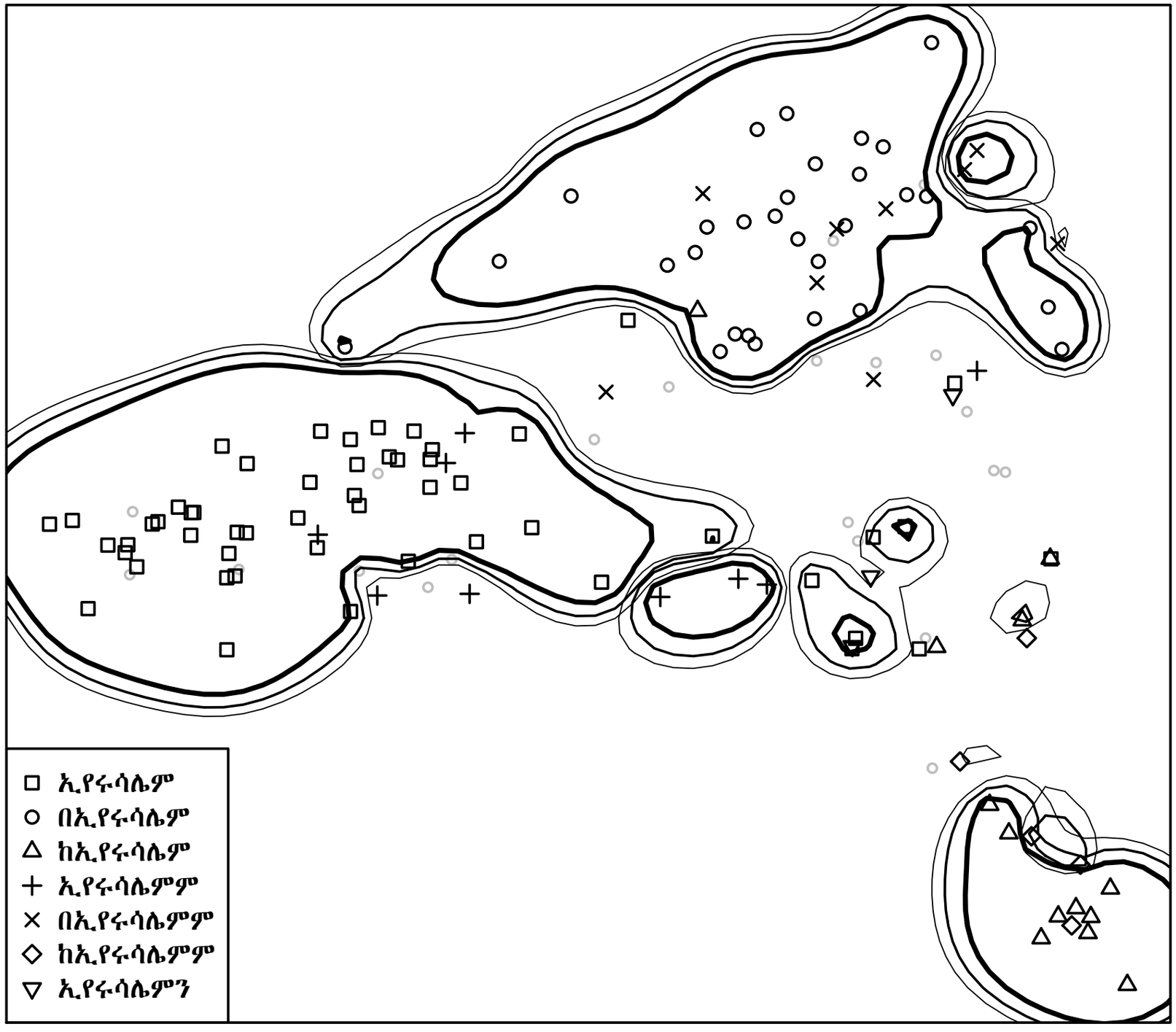
Angor (a language of Papua New Guinea)

agm



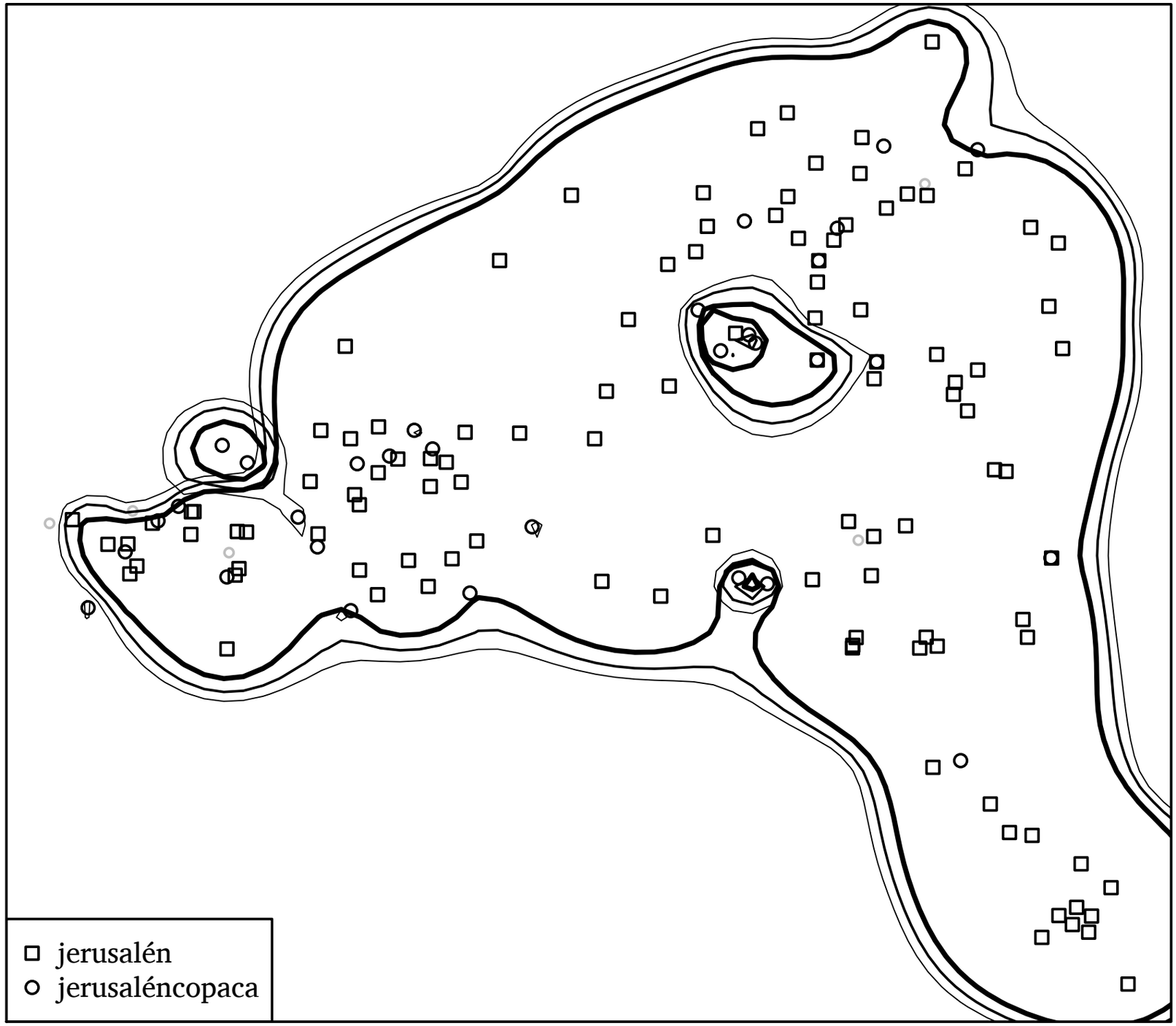
Angaatiha (a language of Papua New Guinea)

amh



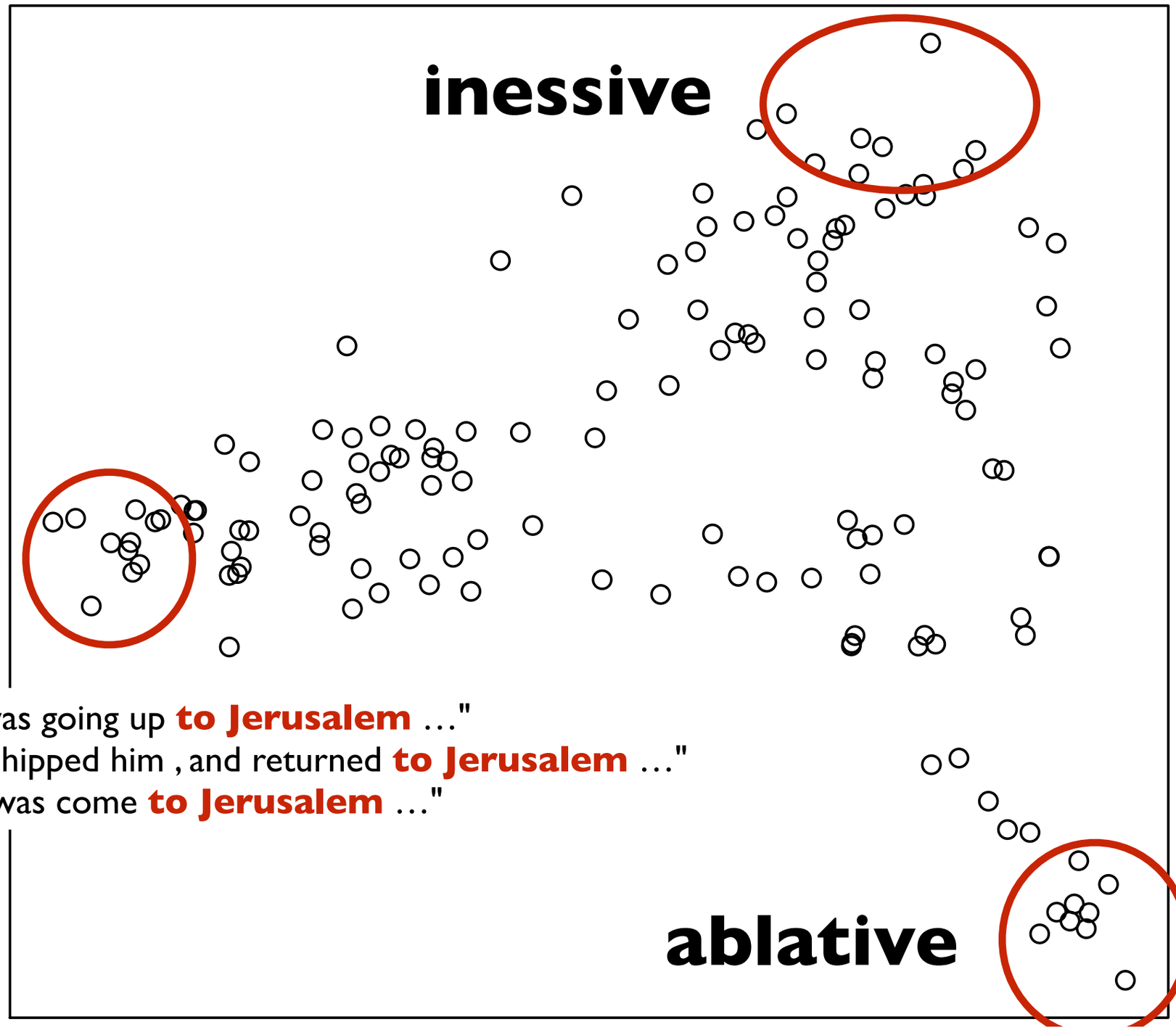
Amharic (a language of Ethiopia)

azz



Highland Puebla Náhuatl

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** …"

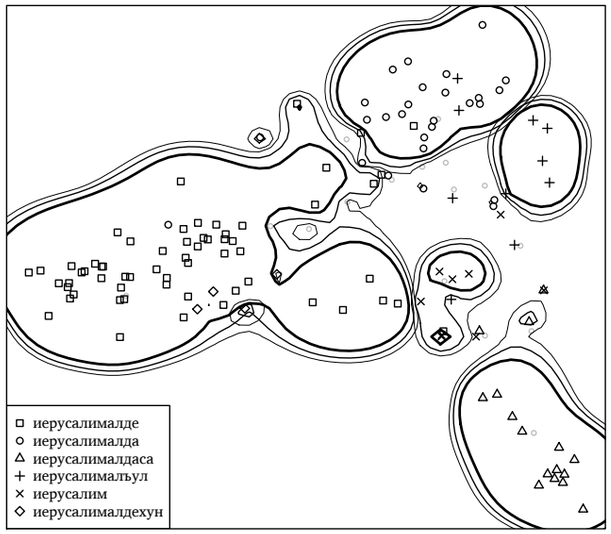
41003022: "And the scribes that came down **from Jerusalem** said …"
42005017: "… there were Pharisees and doctors of the law sitting by , who were come **out of** every village of Galilee and Judaea and **Jerusalem** …"
42010030: "… A certain man was going down **from Jerusalem** to Jericho …"
…"

allative
dominated

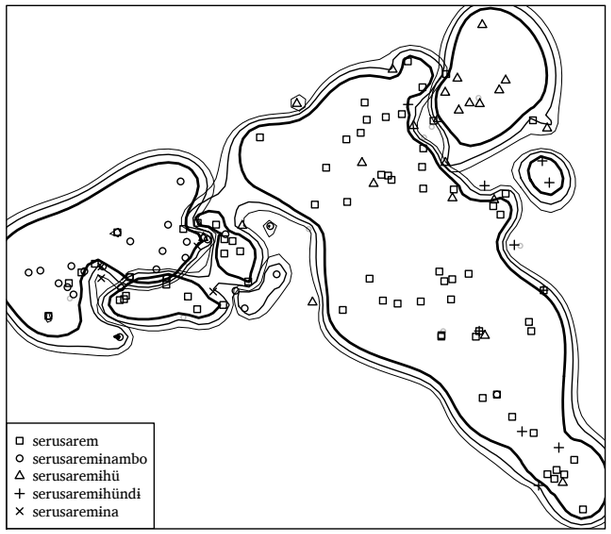
ablative
dominated

inessive
dominated

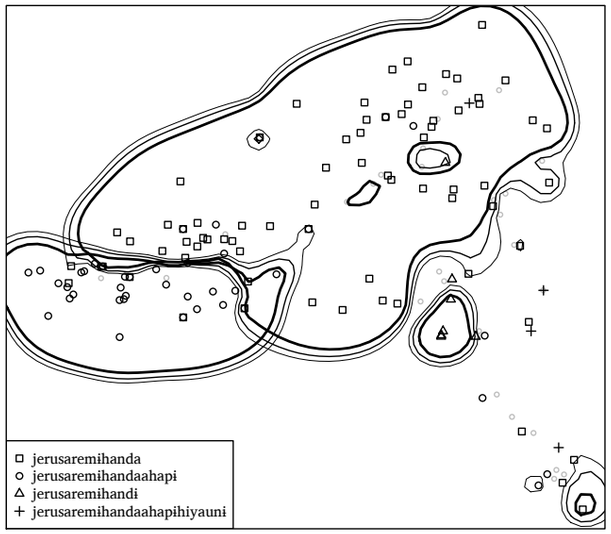
ava



agg



agm



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, so manual decisions