# Bottom-up- und Top-down-Zugänge zu sprachvergleichenden Korpusanalysen

Michael Cysouw & Simon Kasper
Philipps-Universität Marburg

Philipps Universität Marburg
LOEWE Exzellente Forschung für Hessens Zukunft
DFG Deutsche Forschungsgemeinschaft
erc

# Language comparison today

- either **tertiary "grammar-based"**
  - tertiary interpretations of secondary analyses of primary sources
- or **secondary "questionnaire-based"**
  - questionnaires ideally filled out by specialists with direct access to primary sources
- or **primary "survey-based"**
  - collection of specific primary data, mainly possible in dialectology (and with massive monetary resources)

# Corpus-based comparison

- Alternative: reanalysis of existing primary data
  - pro: based on primary data
  - contra: no control about available data

- **LOEWE-database** (S. Kasper)
  - focussed manual annotation of monolingual corpora based on specific research topic (“top-down”)

- **Parallel-text project** (M. Cysouw & T. Mayer)
  - induce annotation from translational equivalents based on functional domain (“bottom-up”)

# Algorithmic corpus-based approaches to typological comparison

Michael Cysouw & Thomas Mayer (Marburg)
Uwe Quasthoff & Dirk Goldhahn (Leipzig)


# Induction of linguistic annotation

- Combination of two approaches
  - large-scale automatic monolingual corpus analysis (Leipzig)
  - preparation of massively parallel texts to link corpora across languages (Marburg)
- Underlying research question:
  - what kind of linguistic marking is “easy” to annotate, what is “difficult”
  - where is manual work by specialists needed

# Massively Parallel Texts

- Same text available in many languages (i.e. translations !)
- Including lesser-described languages
  - *Bible*
  - *Universal Declaration of Human Rights*
  - *Pamphlets of Jehova's Witnesses*

# Bible Corpus

- http://paralleltext.info
- Currently 1000 texts (90% only NT), up to 1500 reached by the end of next year
- Text cleaned, verse-aligned, text encoding harmonised, wordform tokenisation …
- Not everything completely freely available (ask us!)

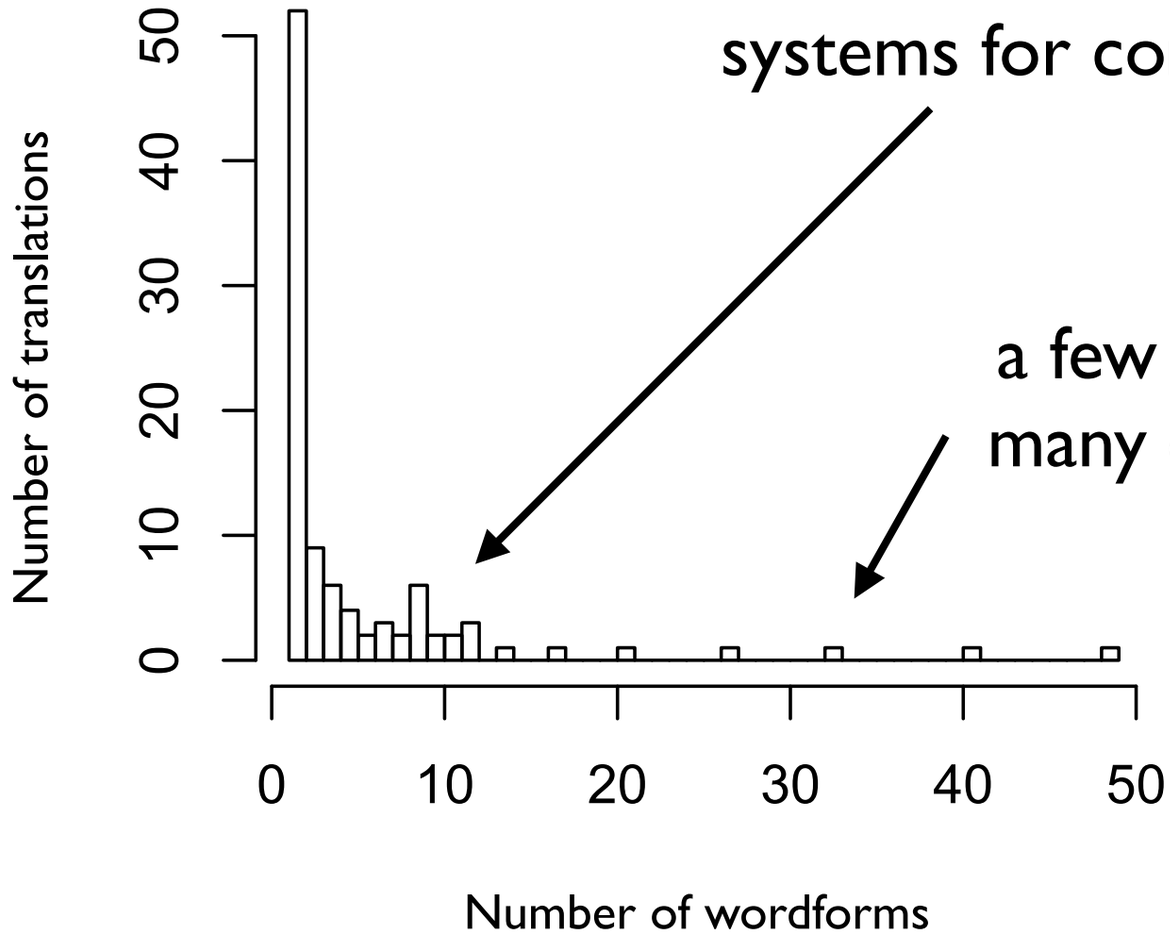
# quick test: translation of "*Jerusalem*"

- 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

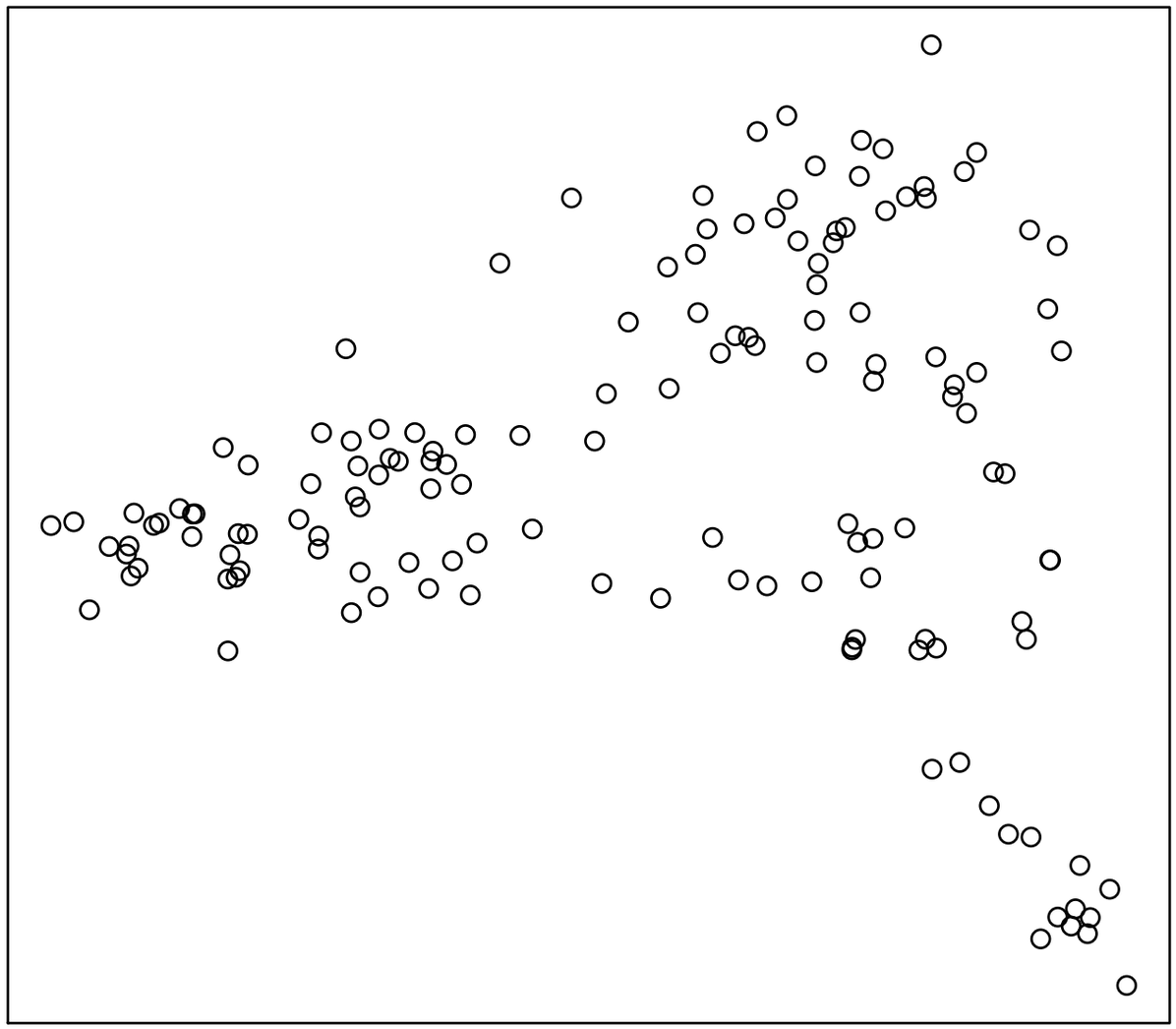
(ISO 639-3 amh, spoken in Ethiopia)

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

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



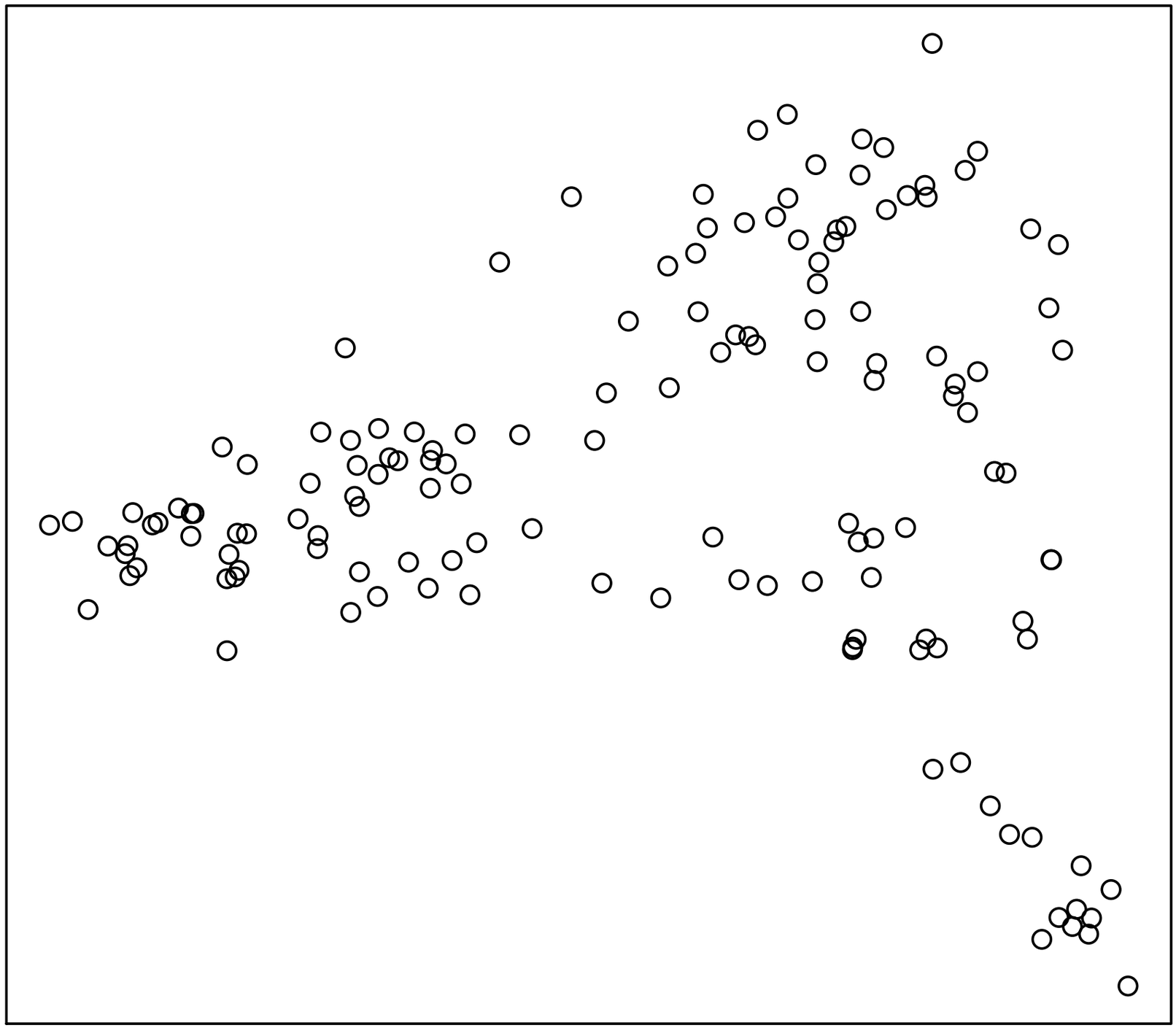
Two contexts of *Jerusalem*
are similar when they often
share the same wordform in
language after language

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

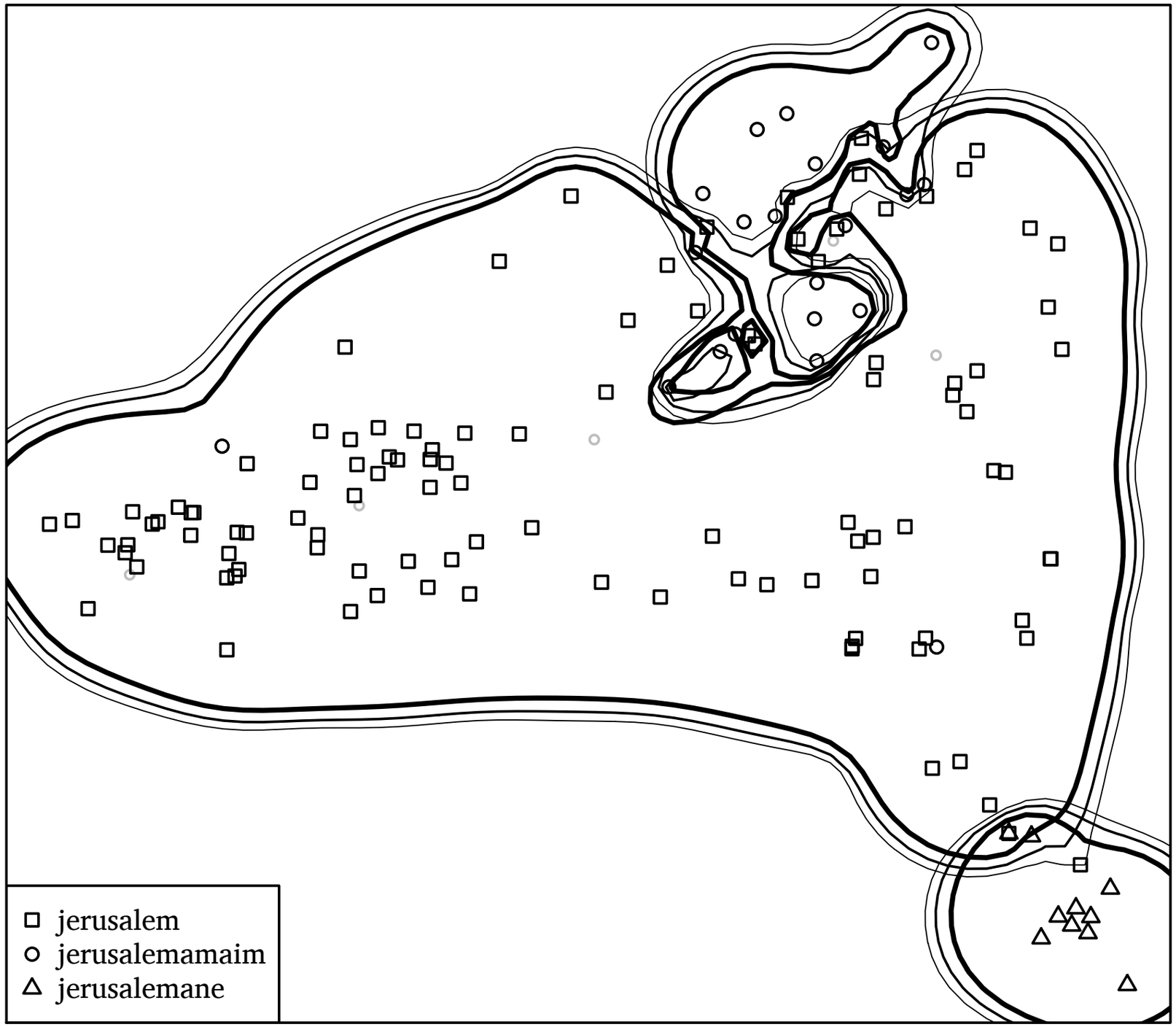
- Taking the different wordforms in a language as a system of marking
- How similar are the systems functionally?
- We can compare functions through distribution in the text



# aey

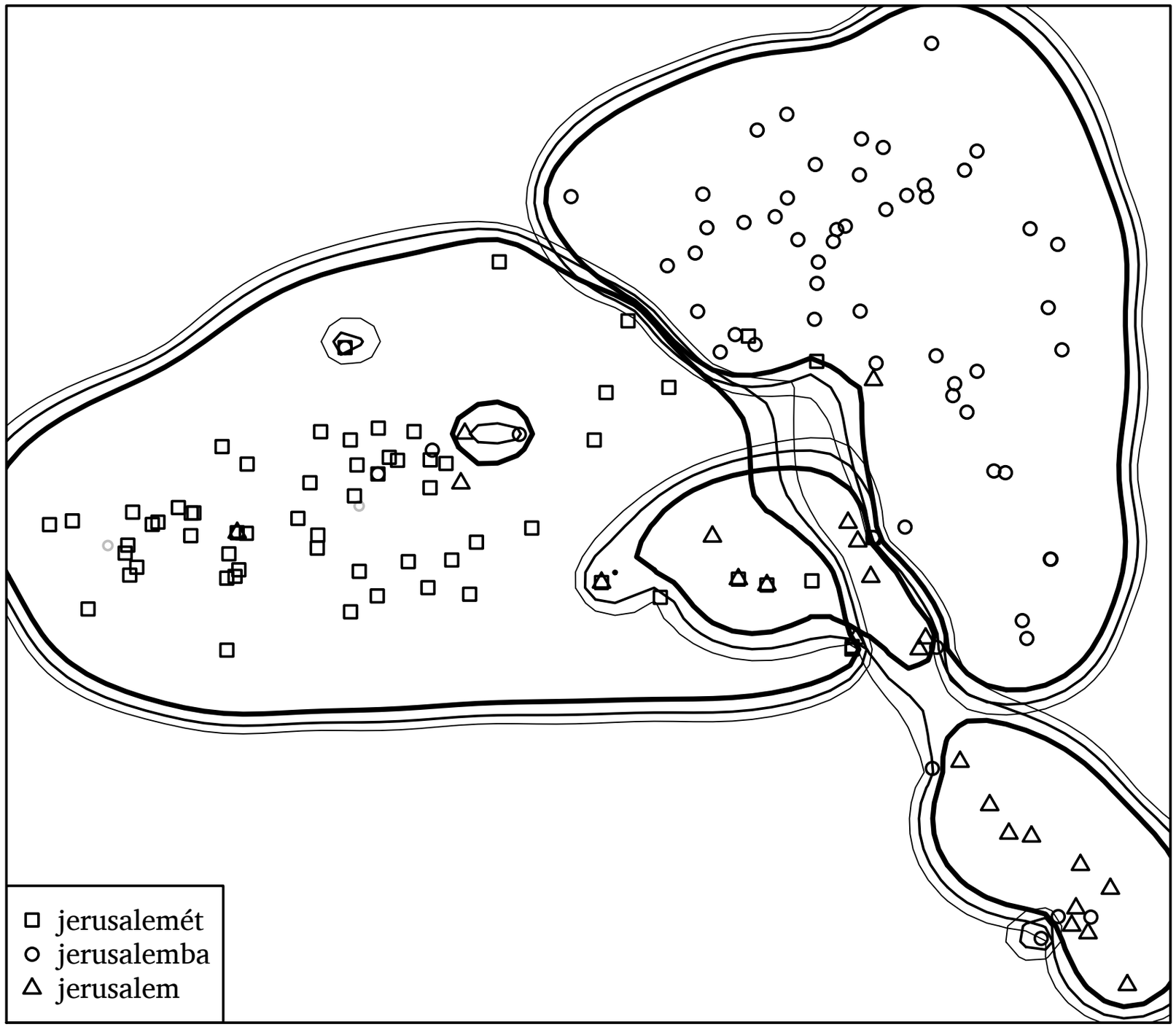

Amele (A language of Papua New Guinea)

# aai



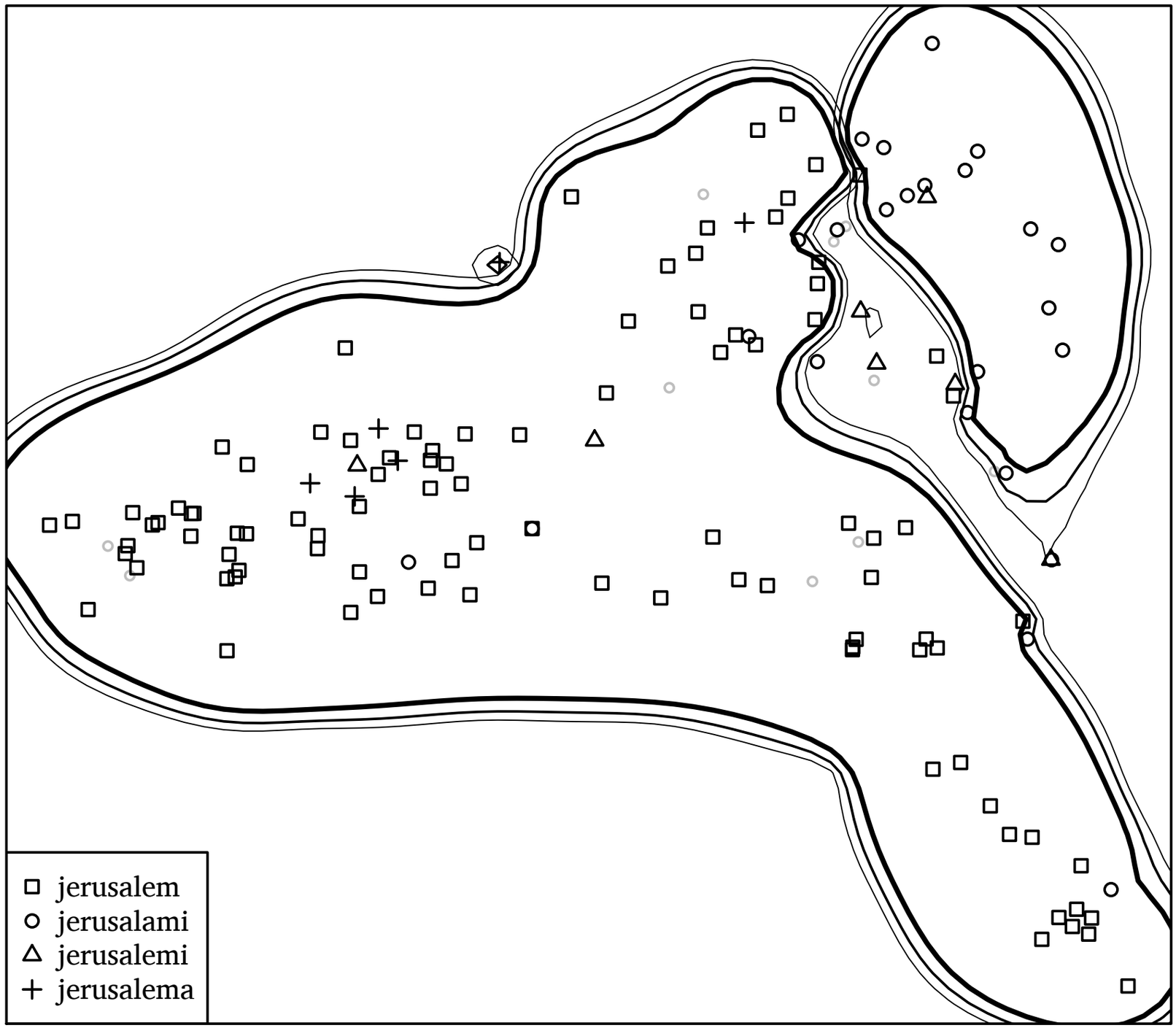

Arifama-Miniafia (a language of Papua New Guinea)

# abt



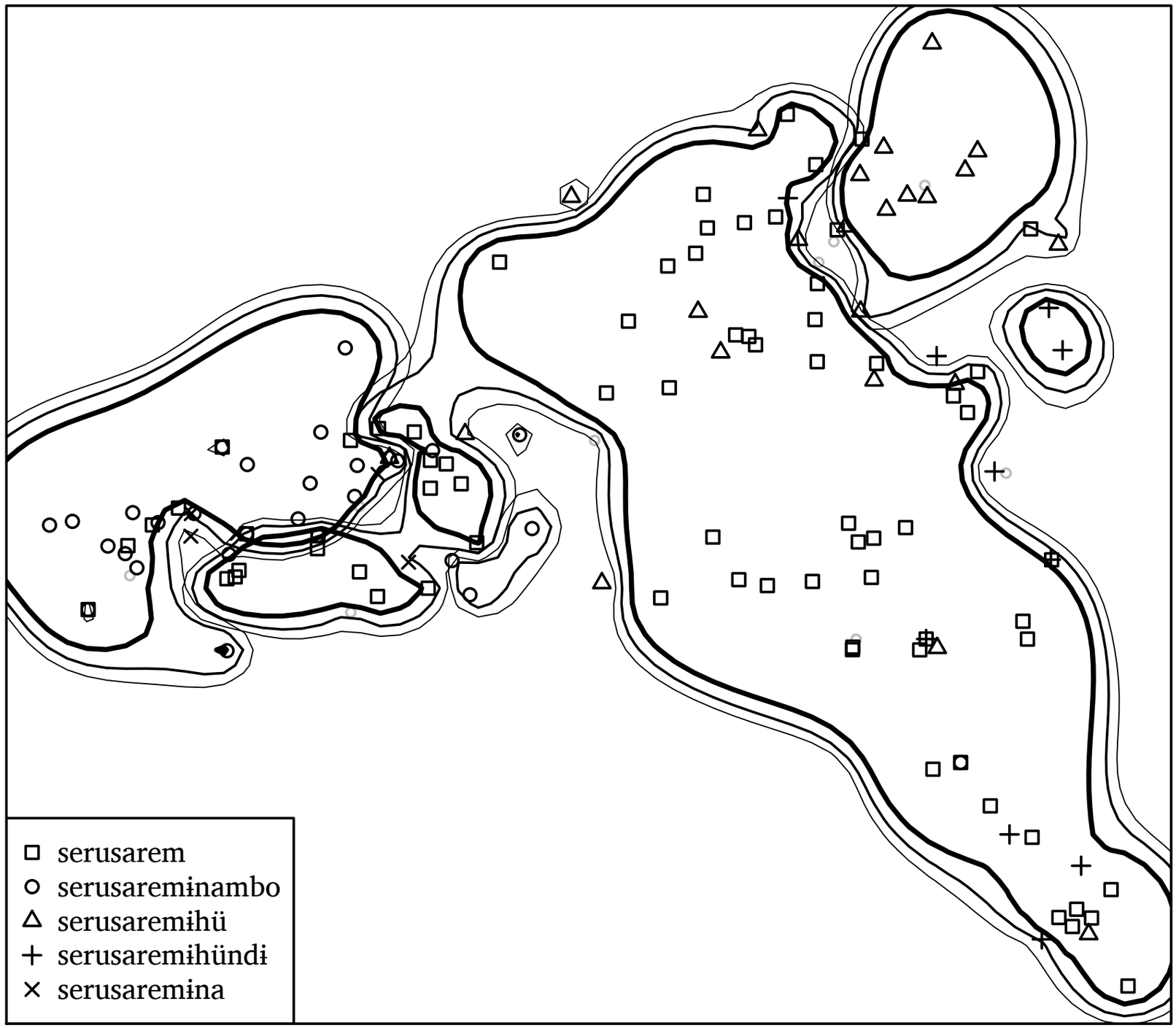

Ambulas (a language of Papua New Guinea)

# aoj



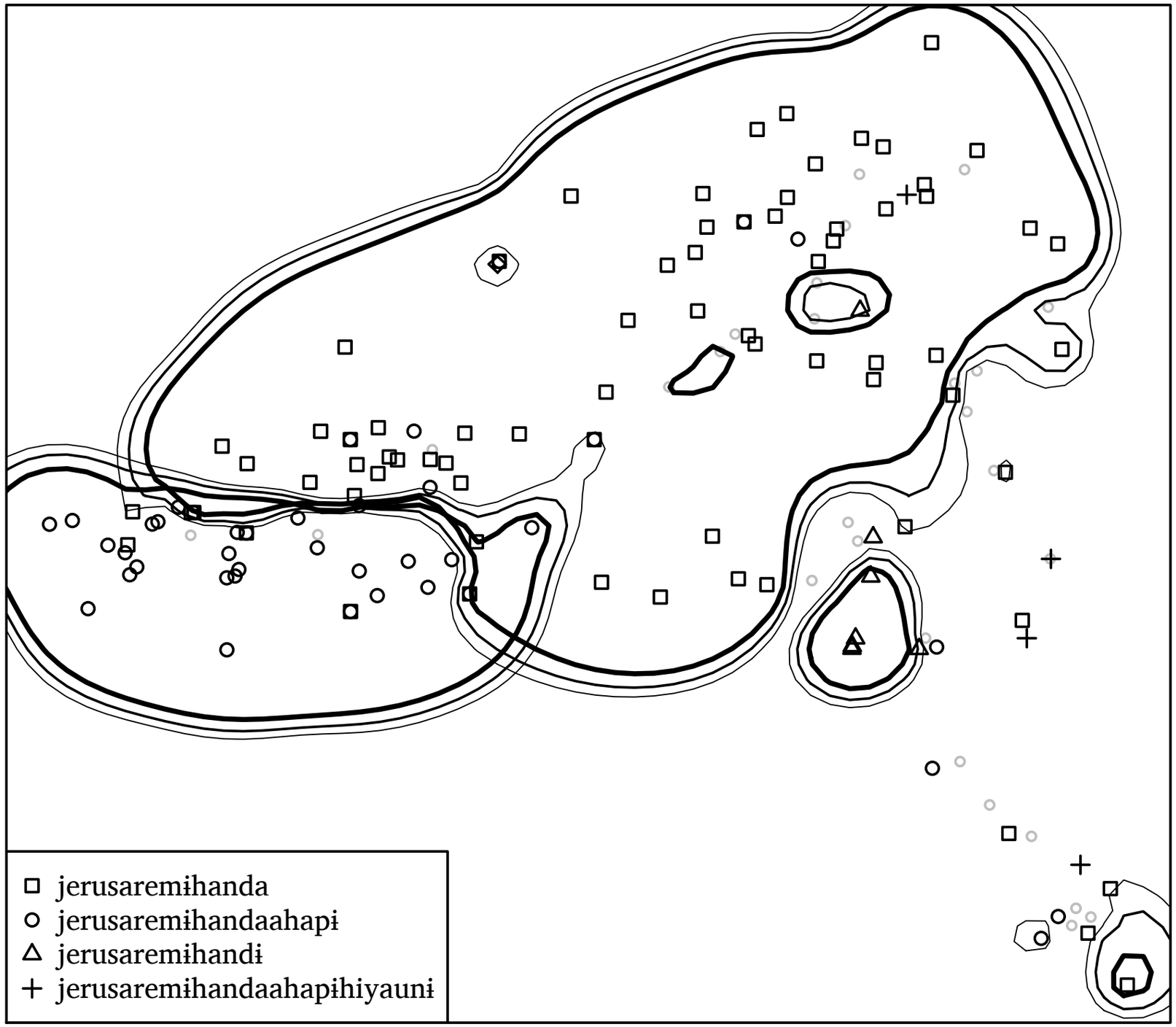

Muffian (a language of Papua New Guinea)

# agg



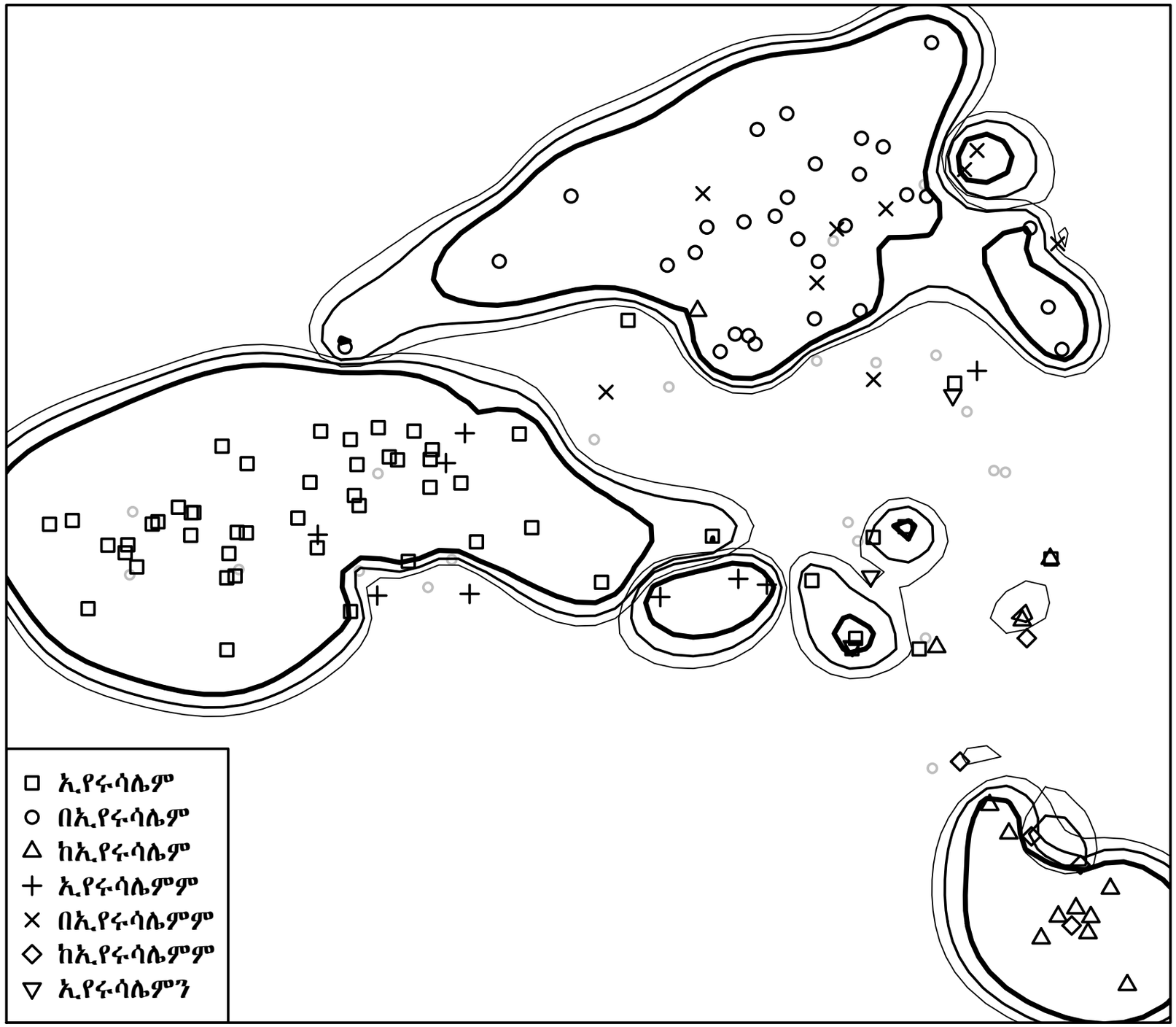

Angor (a language of Papua New Guinea)

# agm



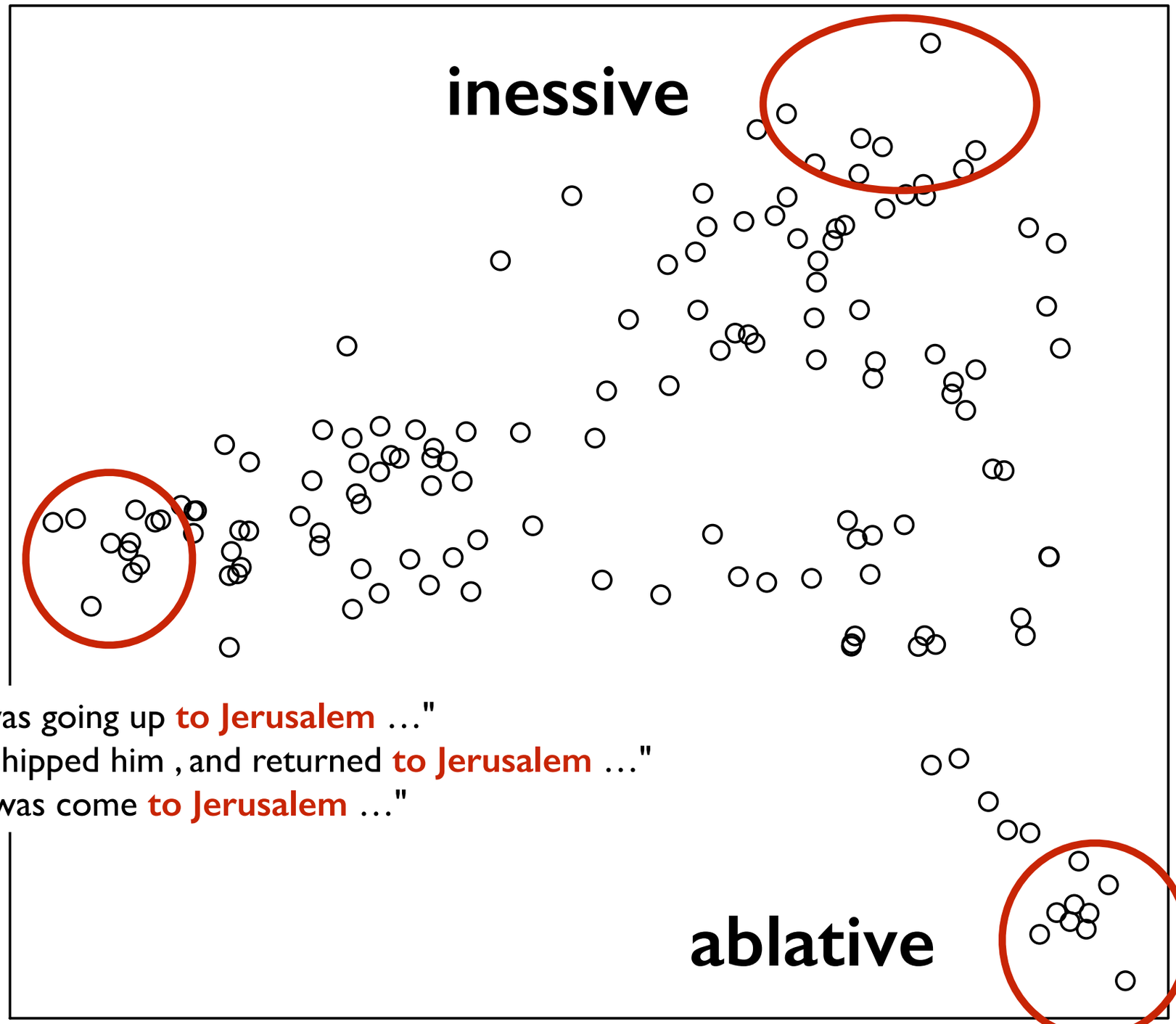

Angaatiha (a language of Papua New Guinea)

# amh




Amharic (a language of Ethiopia)

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 …"

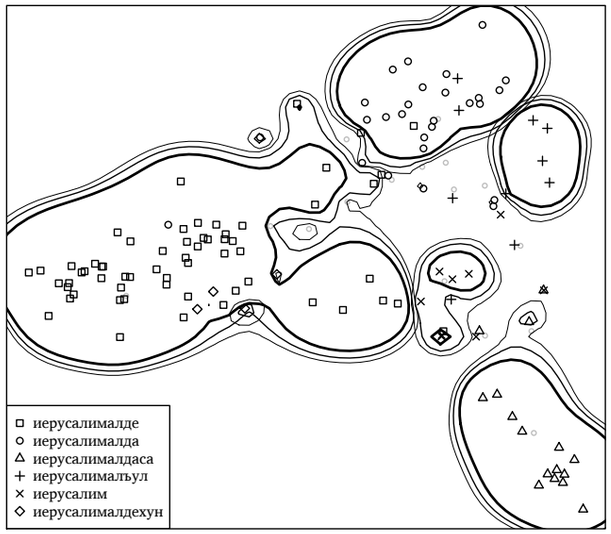



40020017: "… as Jesus was going up **to Jerusalem** …"
42024052: "… they worshipped him , and returned **to Jerusalem** …"
44009026: "… when he was come **to Jerusalem** …"

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 …"

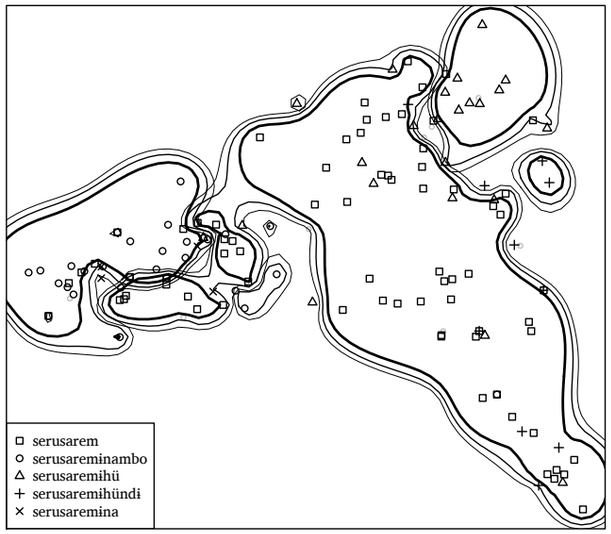
# allative dominated

**ava**



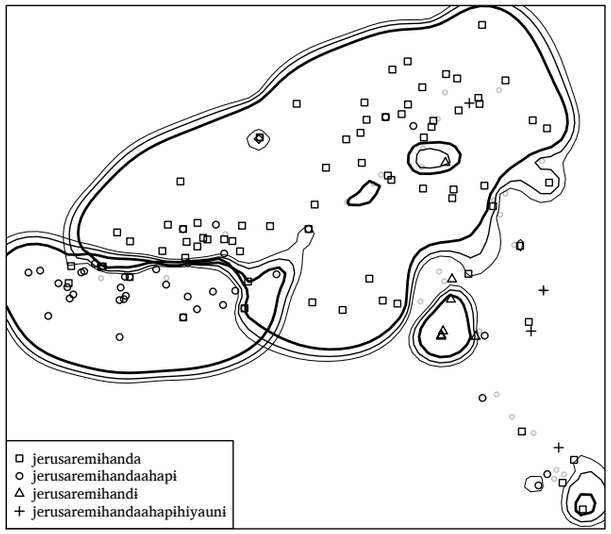

# ablative dominated

**agg**




# inessive dominated

**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