

Dynamic universals in the linguistic marking of location

extracted from parallel texts

Michael Cysouw

Philipps



Universität
Marburg

A stylized map of Southeast Asia and Oceania. The landmasses are colored in shades of green and yellow, while the surrounding waters are light blue. Several colored dots are scattered across the map: yellow dots are located in the Philippines, Indonesia, and Papua New Guinea; red dots are in the Malay Peninsula and East Timor; and purple dots are concentrated in Australia. The text 'Introducing the massively Parallel Text Corpus' is overlaid in the center in a large, bold, black font.

Introducing the massively Parallel Text Corpus

Worldwide ‘survey’ data

- **Massively parallel texts**
 - ▶ Same text available in many languages (i.e. translations !)
 - ▶ Contextually situated comparable expressions
- **Including lesser-described languages**
 - ▶ Bible
 - ▶ Universal Declaration of Human Rights
 - ▶ Pamphlets of Jehova’s Witnesses

Bible corpus

- <http://paralleltxt.info>
 - ▶ 1169 translations online (soon 1850+)
 - ▶ 906 different ISO-639/3 codes (soon 1400+)
 - ▶ All data in a private GitHub repo (ask me!)
 - ▶ Old Testament 23K verses, New Testament 8K verses, Apocrypha 6K verses
 - ▶ All texts cleaned, aligned, normalised, punctuation separated
 - ▶ 4 GB raw text files
- Today data from
 - ▶ 1556 New Testament translations
 - ▶ 1163 different ISO-639/3 codes

Single-word comparative linguistics

The case of 'Jerusalem'

Angaataha

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

- jerusaremthanda
- jerusaremthandaahapt̩
- jerusaremthand̩t̩
- jerusaremthandaahiyai
- jerusaremthandaah̩t̩
- jerusaremthandaahapt̩hiyaunt̩
- jerusaremthandaahiyaisangi
- jerusaremthandaahapt̩hiya
- jerusaremthandaah̩traapt̩
- jerusaremthandaah̩t̩t̩
- jerusaremthandaah̩the
- jerusaremthandamt̩
- jerusaremthamanda
- jerusaremthandapt̩
- jerusarem̩nd̩t̩
- jerusaremthandaahapt̩to
- jerusaremthandaahapaah̩t̩t̩
- jerusaremthandi
- jerusarem̩mandaahapt̩
- jerusaremthandaahunt̩
- jerusaremthandaahapunt̩
- jerusaremthandaahiya
- jerusaremthandamt̩hint̩
- jerusaremthandaahapt̩hiyaat̩h̩t̩
- jerusaremthandaahapt̩hiyaate
- jerusaremthandaahiyaunt̩

Amharic

(ISO 639-3 amh, spoken in Ethiopia)

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

Amharic

(ISO 639-3 amh, spoken in Ethiopia)

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

Low hanging fruit ...

- Semantic structure of locative case
- Complexity of locative case
- Word order patterns
- Phylogenetic correlation between word order and complexity
- Reconstruction of language families based on functional differences

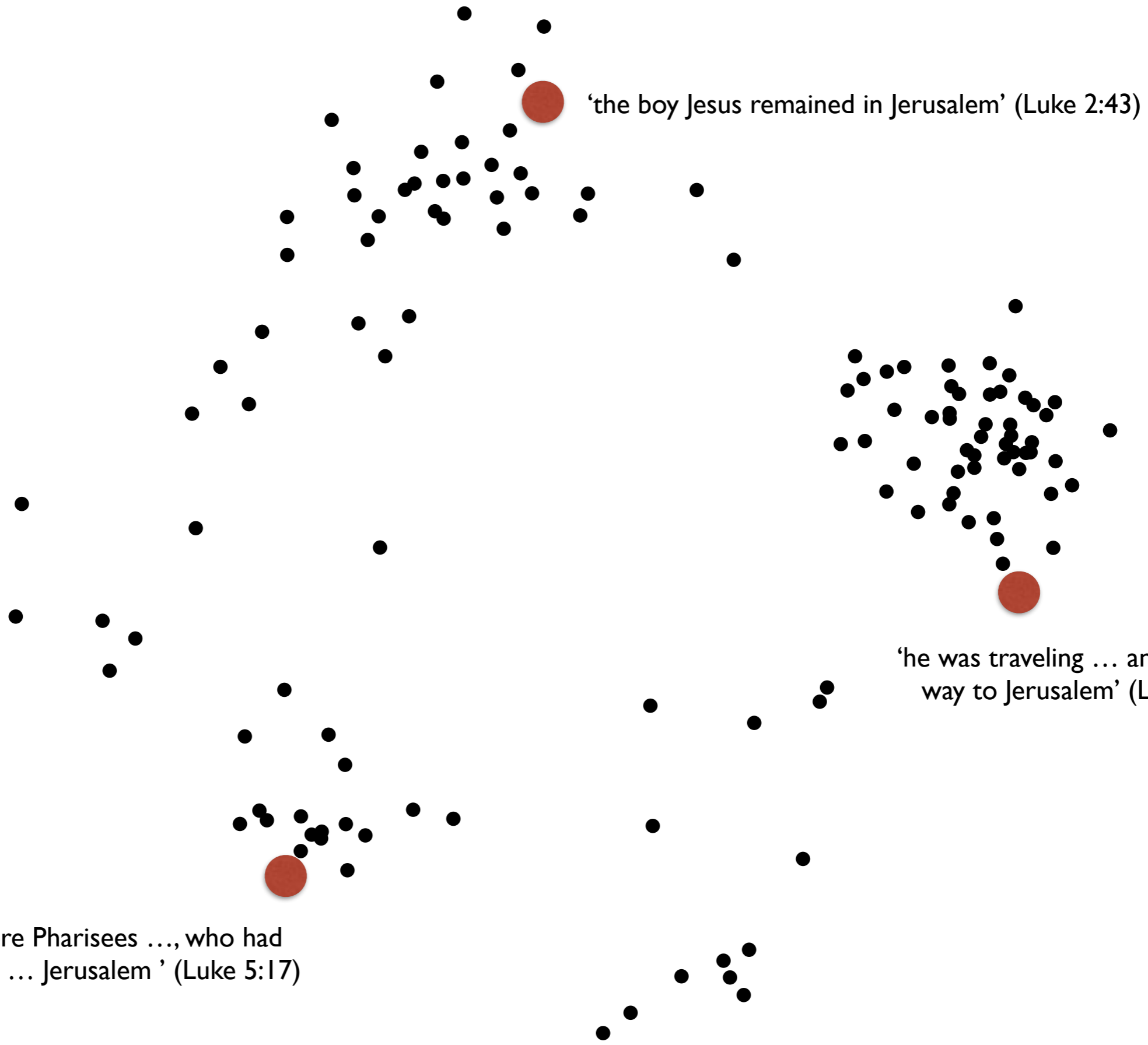


Low hanging fruit ...

- **Semantic structure of locative case**
- Complexity of locative case
- Word order patterns
- Phylogenetic correlation between word order and complexity
- Reconstruction of language families based on functional differences

Semantic map

- When two contexts recurrently show the same form in language after language, then these contexts have a similar meaning
- Use a low-dimensional reduction of the similarity space to compare languages

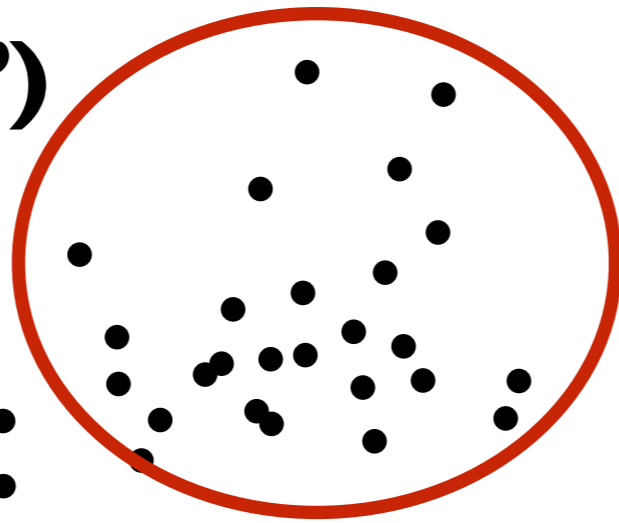


'the boy Jesus remained in Jerusalem' (Luke 2:43)

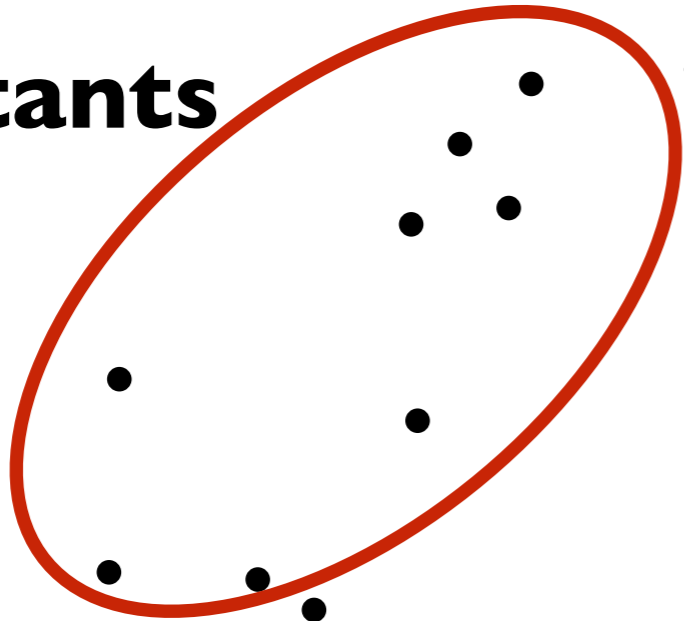
'he was traveling ... and making his way to Jerusalem' (Luke 13:22)

'there were Pharisees ..., who had come from ... Jerusalem' (Luke 5:17)

Essive ('in')



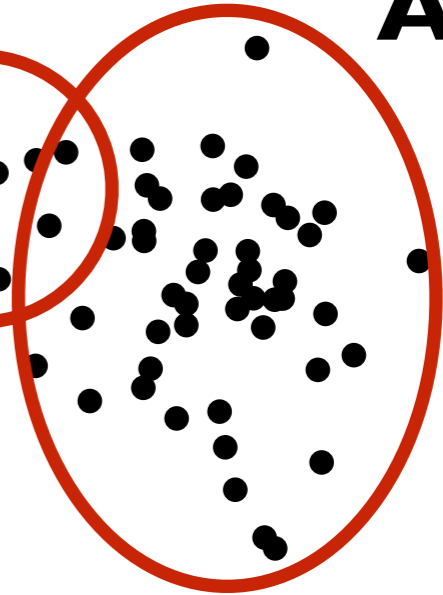
Inhabitants



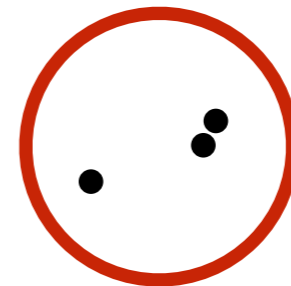
Inlative ('into')



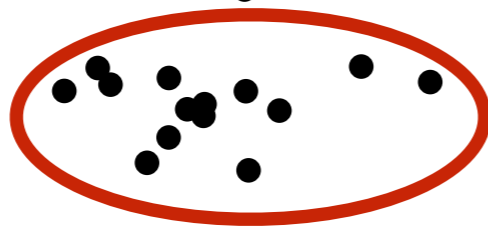
Allative ('to')



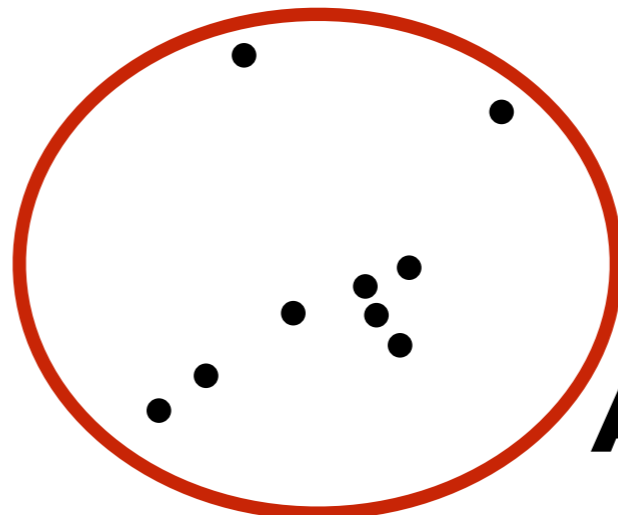
Apudlative ('near')



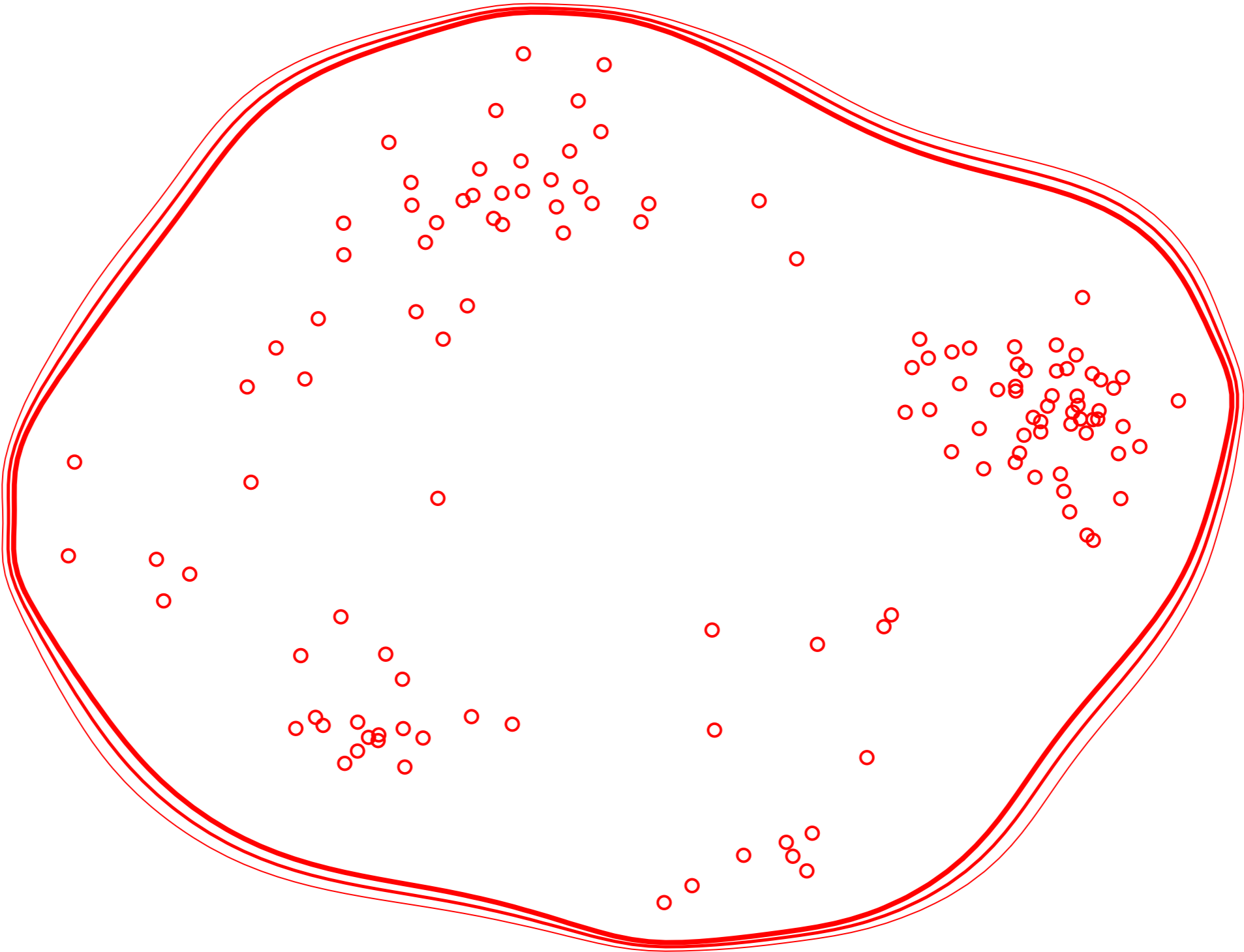
Ablative ('from')



Argument

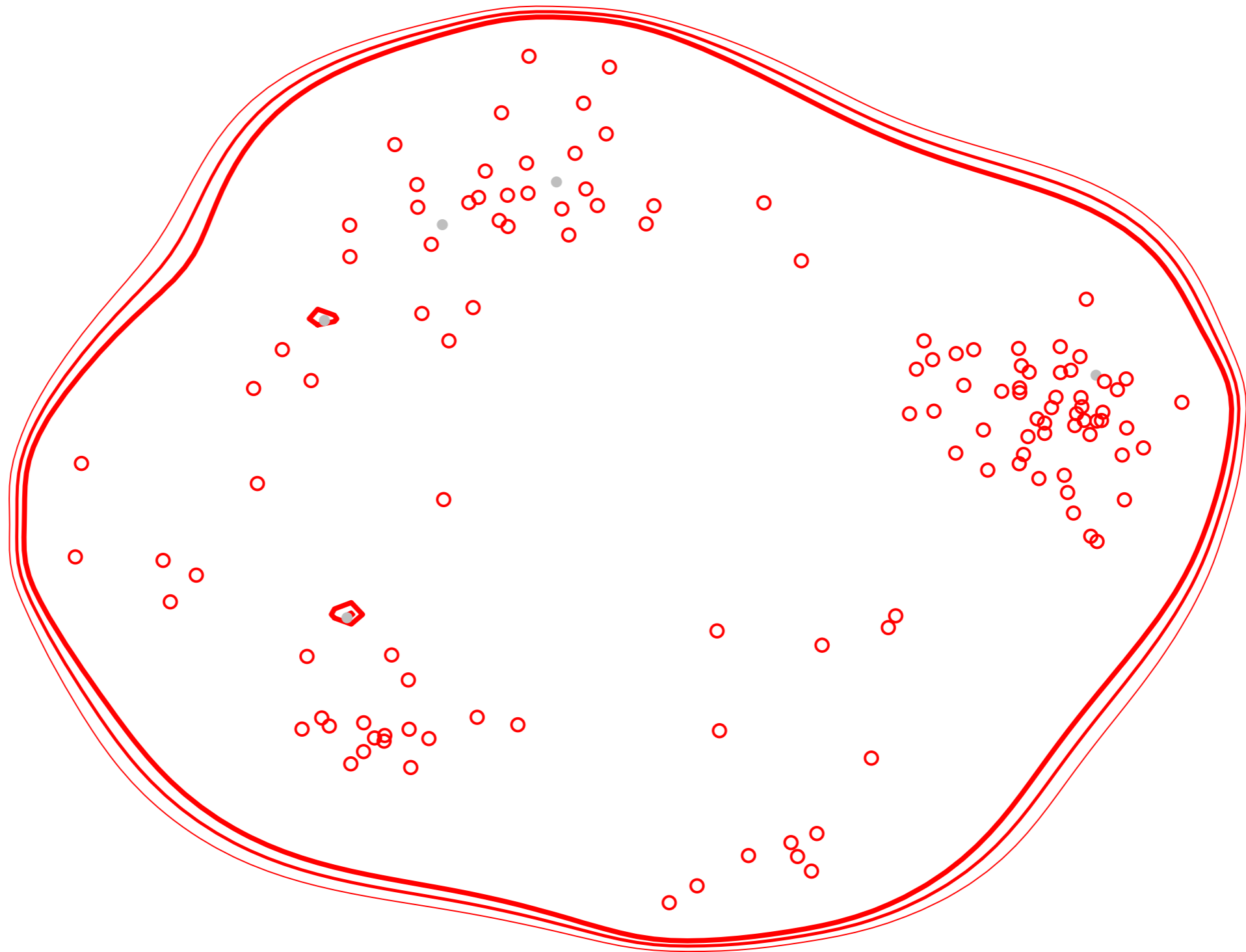


eng-x-bible-catholic.txt



○ Jerusalem

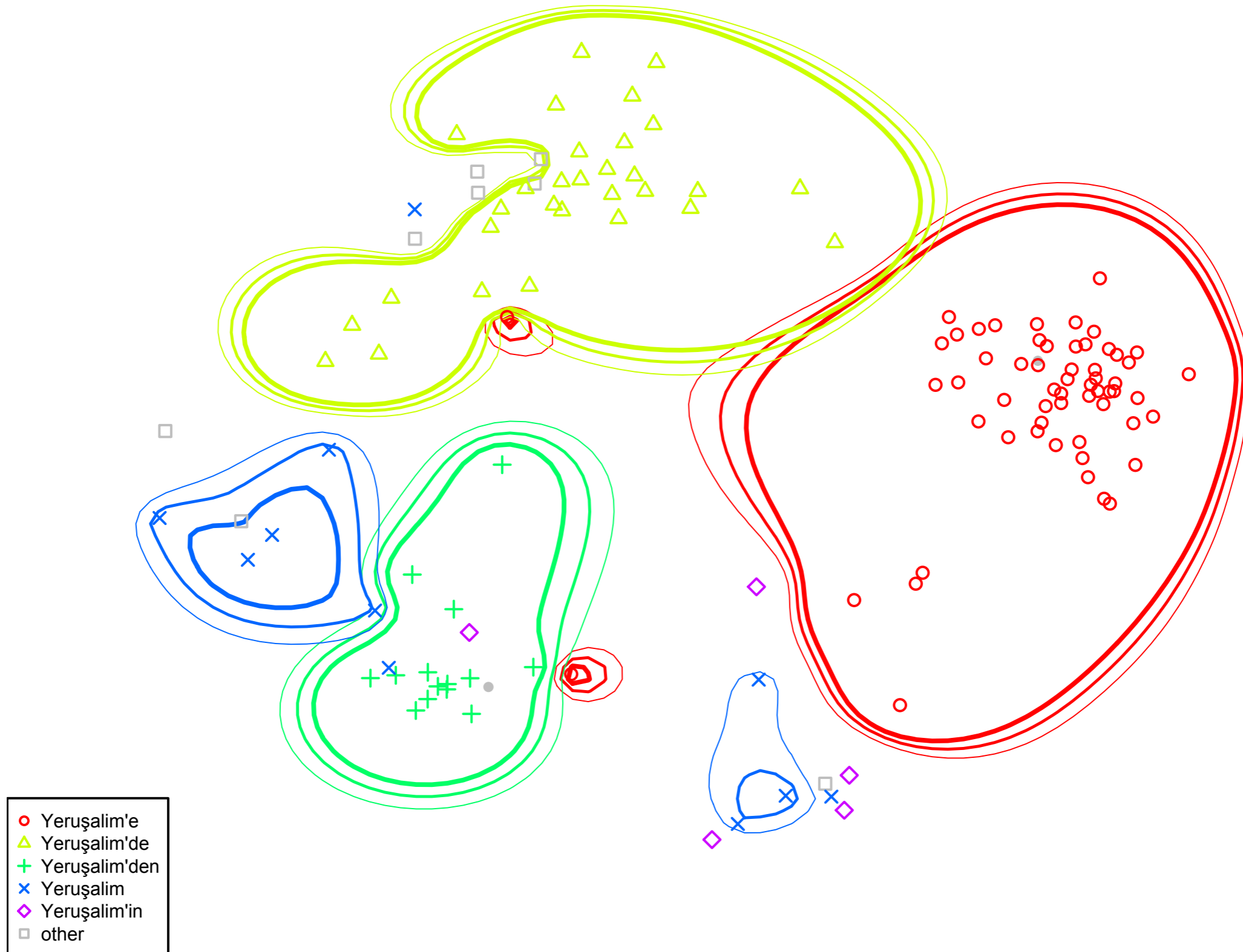
yle-x-bible.txt



○ Njedusalem

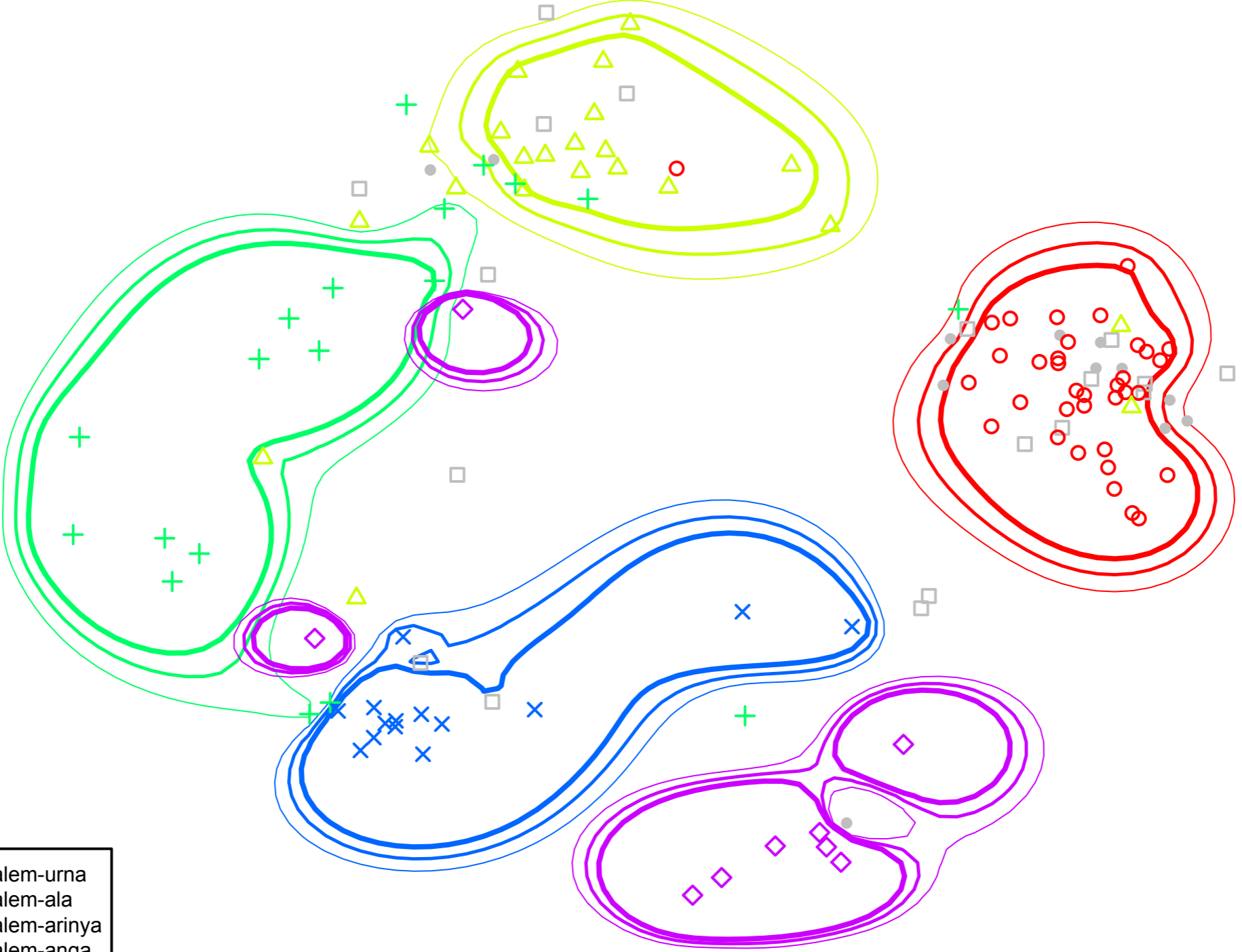
Levels drawn at 41-46-51%

tur-x-bible-2009.txt



Levels drawn at 41-46-51%

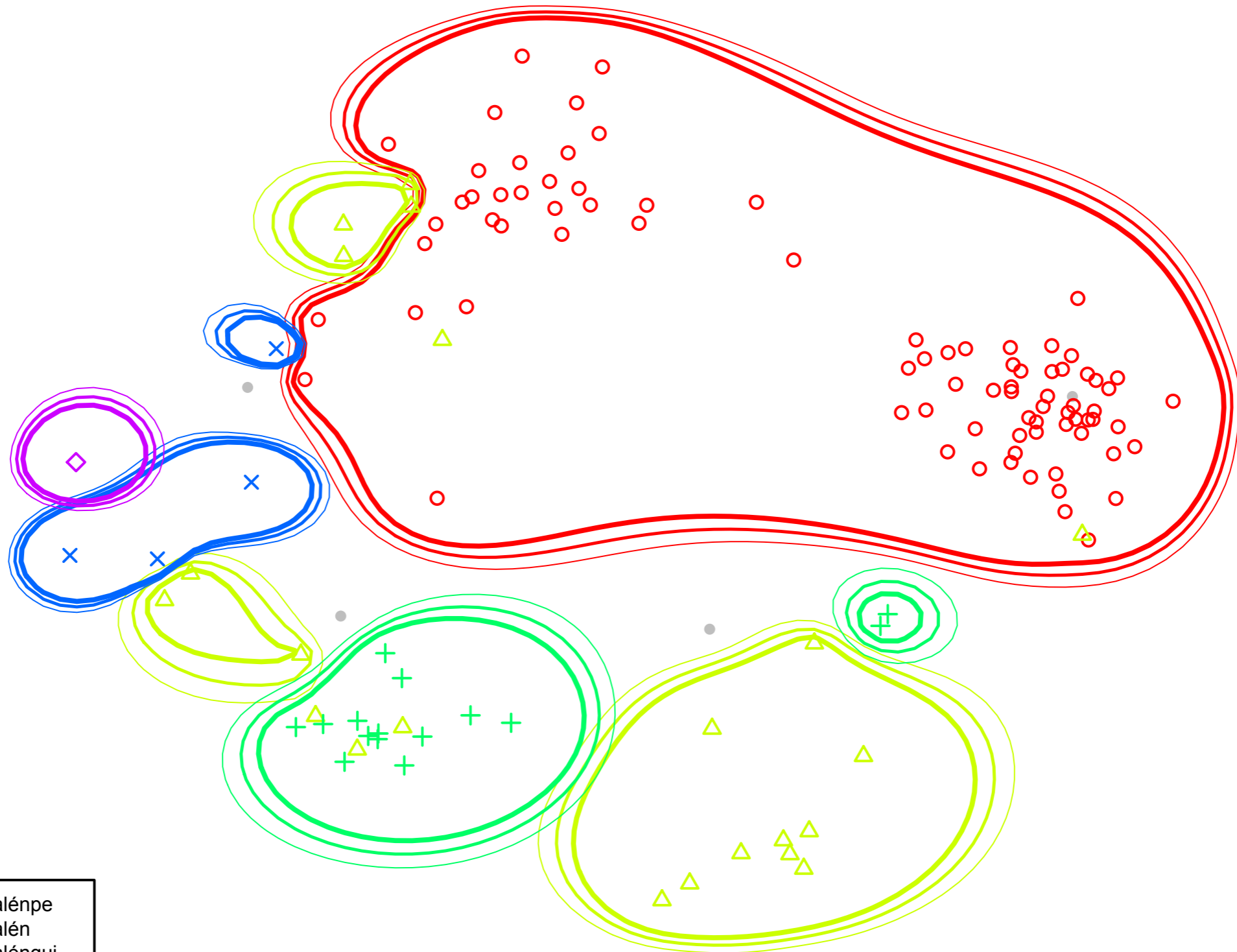
are-x-bible.txt



- Jerusalem-urna
- △ Jerusalem-ala
- + Jerusalem-arinya
- × Jerusalem-anga
- ◇ Jerusalem
- other

Levels drawn at 41-46-51%

gug-x-bible.txt



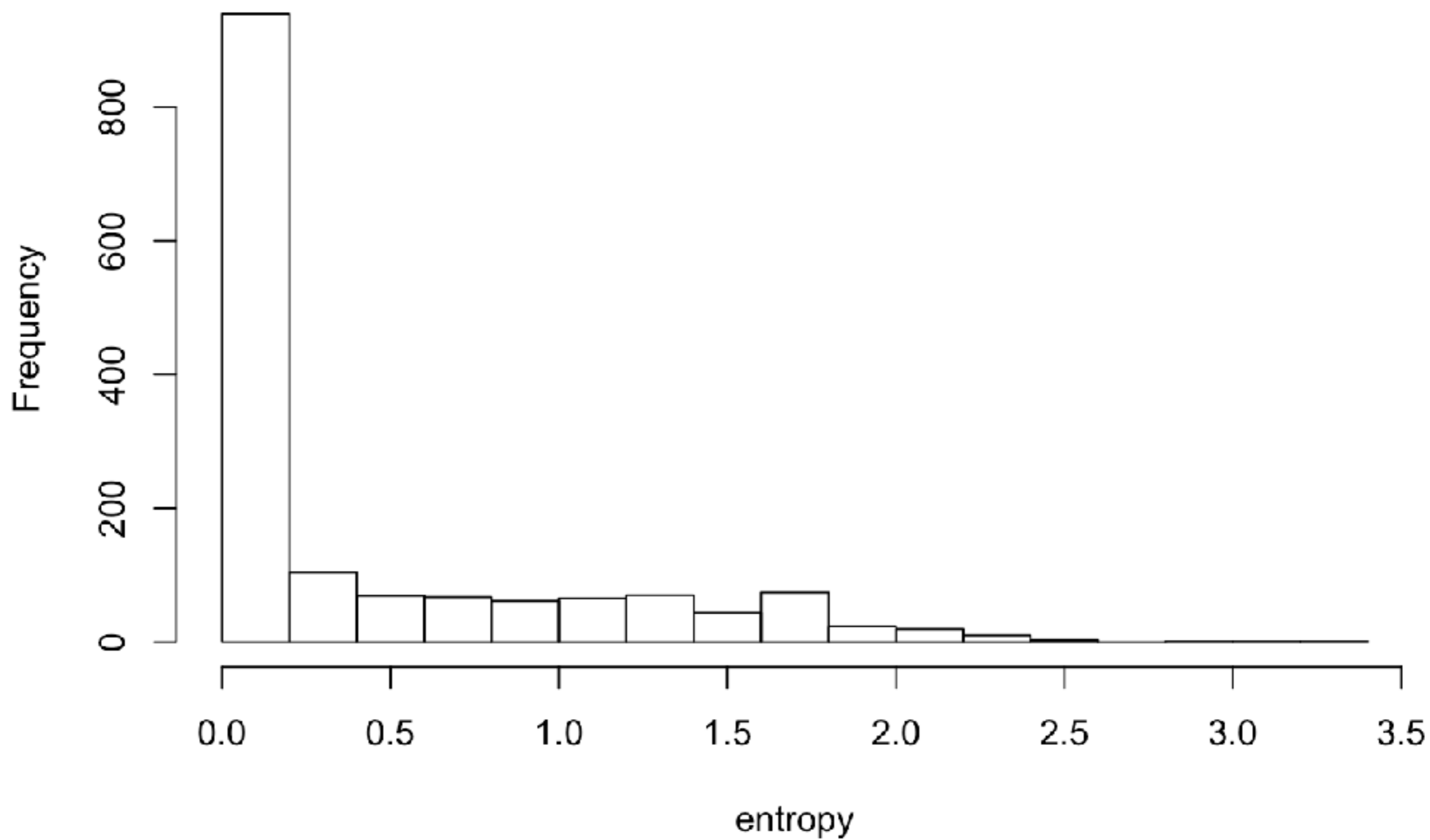
- Jerusalénpe
- △ Jerusalén
- + Jerusaléngui
- × Jerusalengua
- ◇ Jerusalenguáva

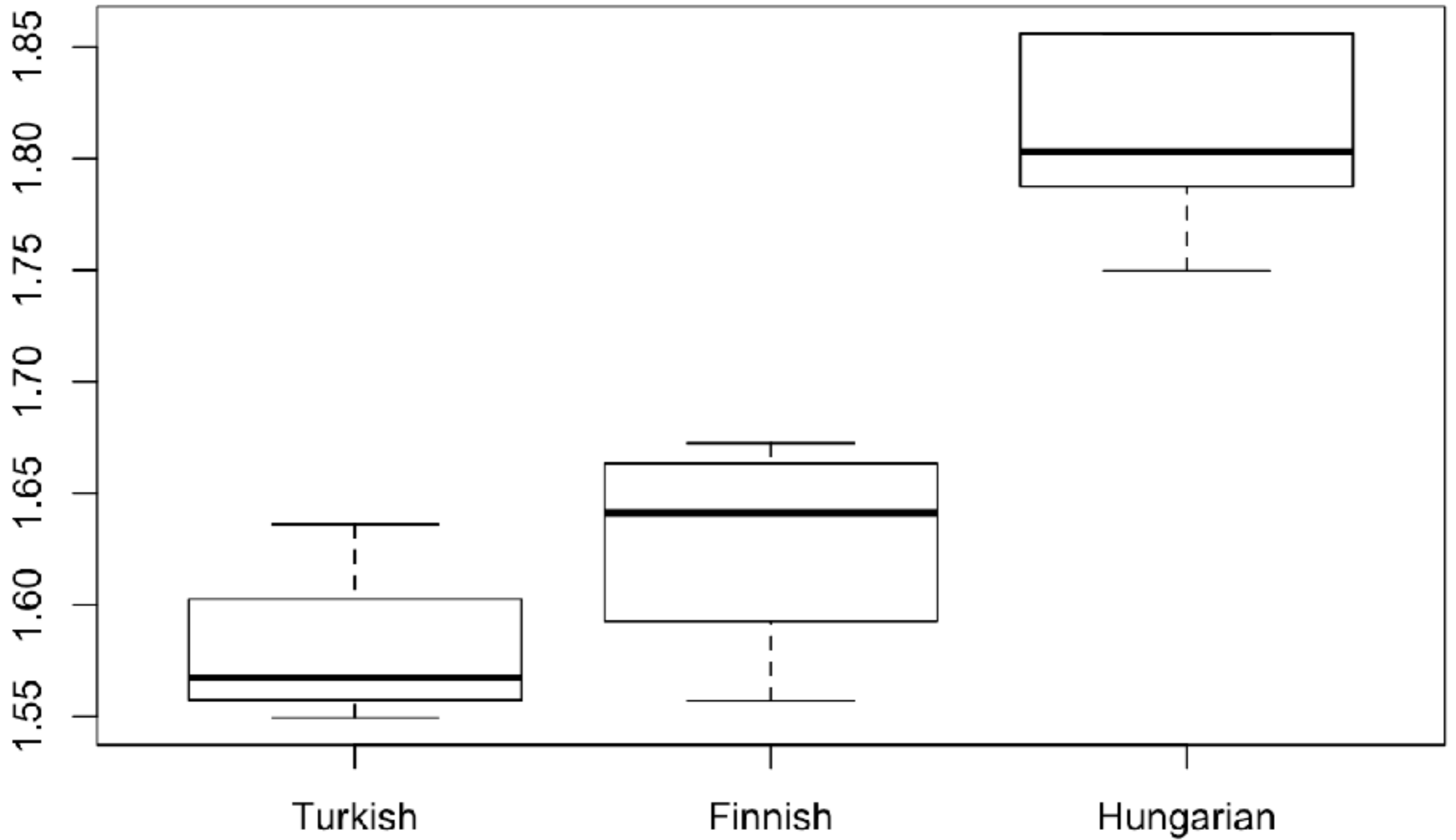


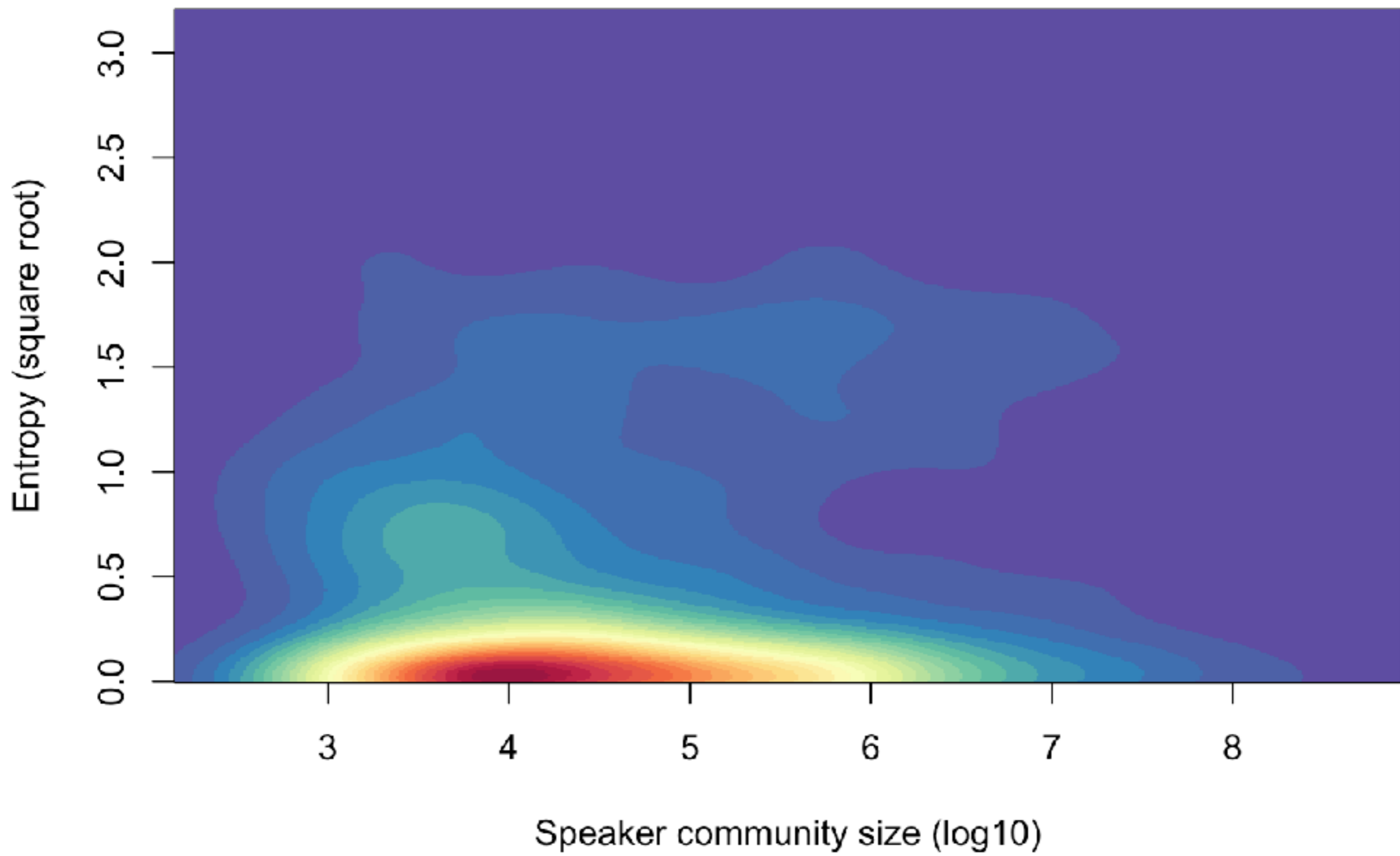
Low hanging fruit ...

- Semantic structure of locative case
- **Complexity of locative case**
- Word order patterns
- Phylogenetic correlation between word order and complexity
- Reconstruction of language families based on functional differences

Histogram of entropy

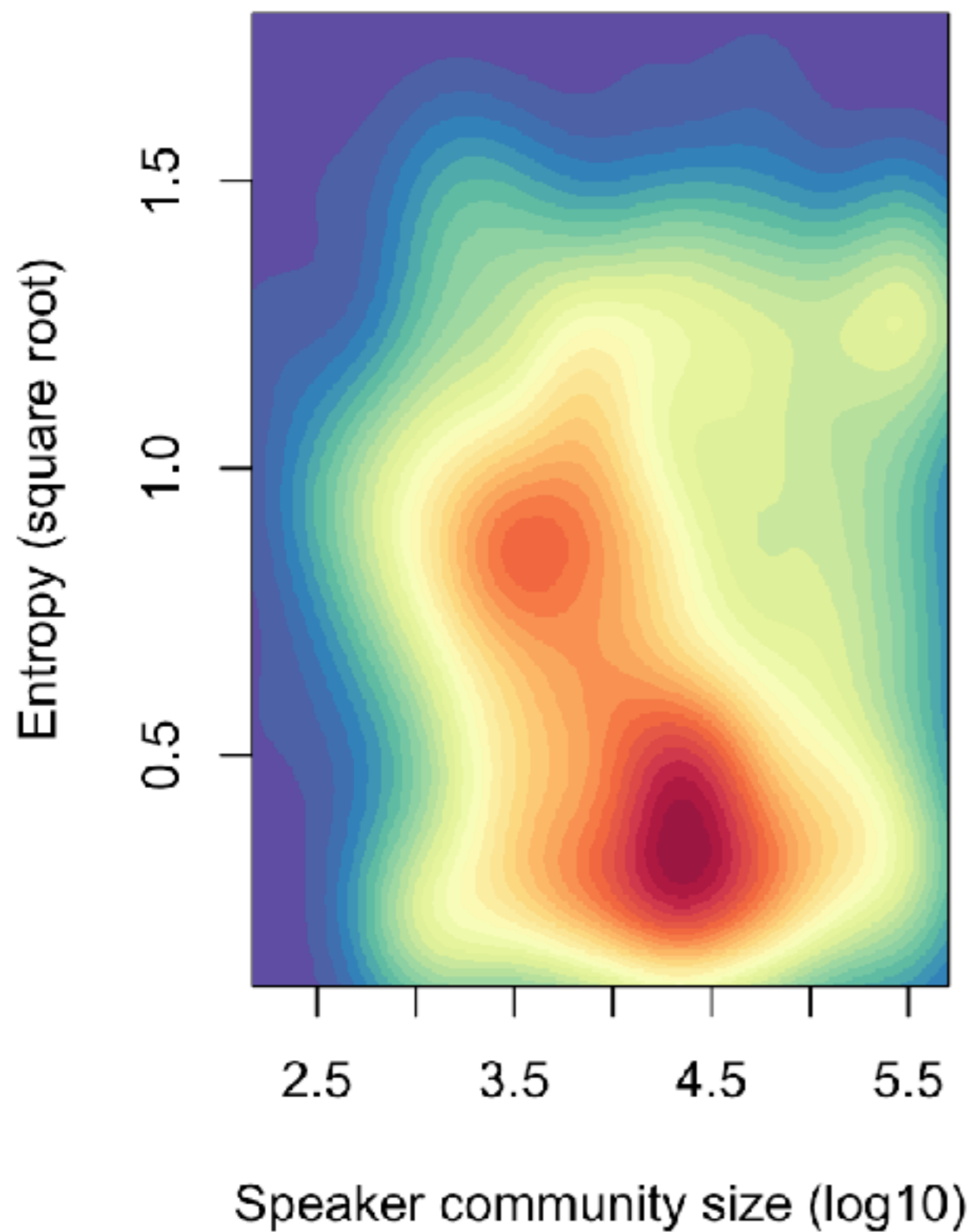






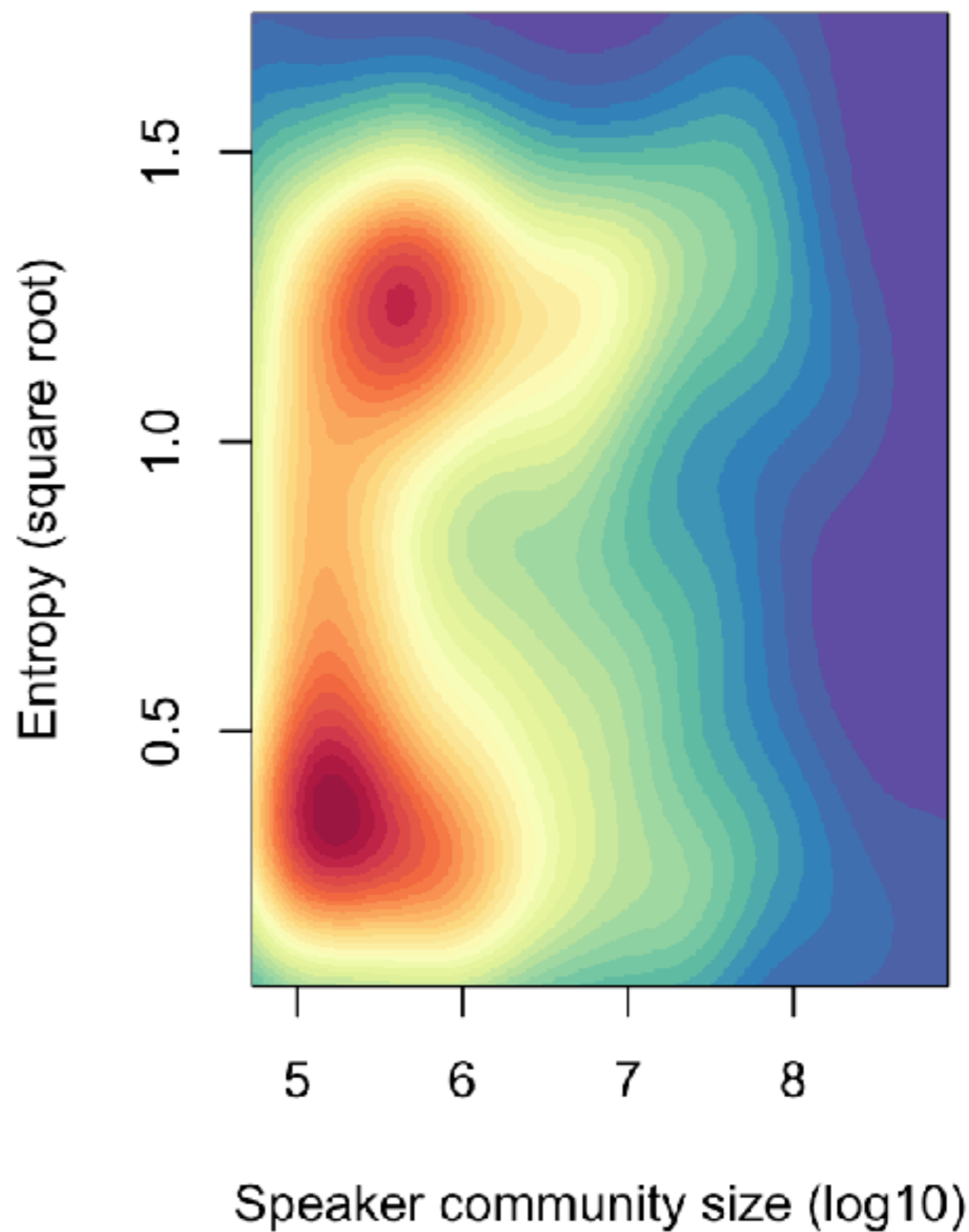
Small communities

ignoring languages with zero entro



Large communities

ignoring languages with zero entro

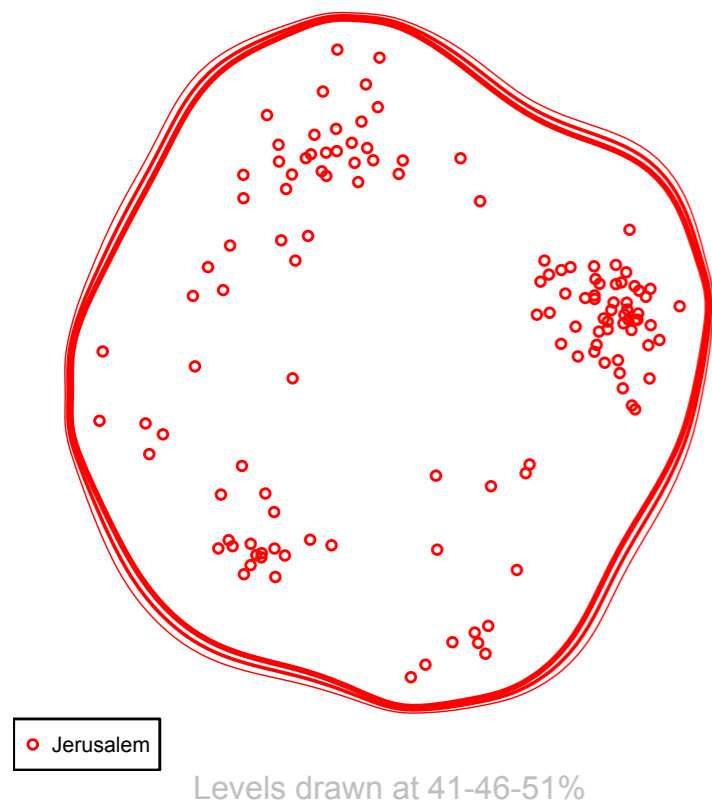




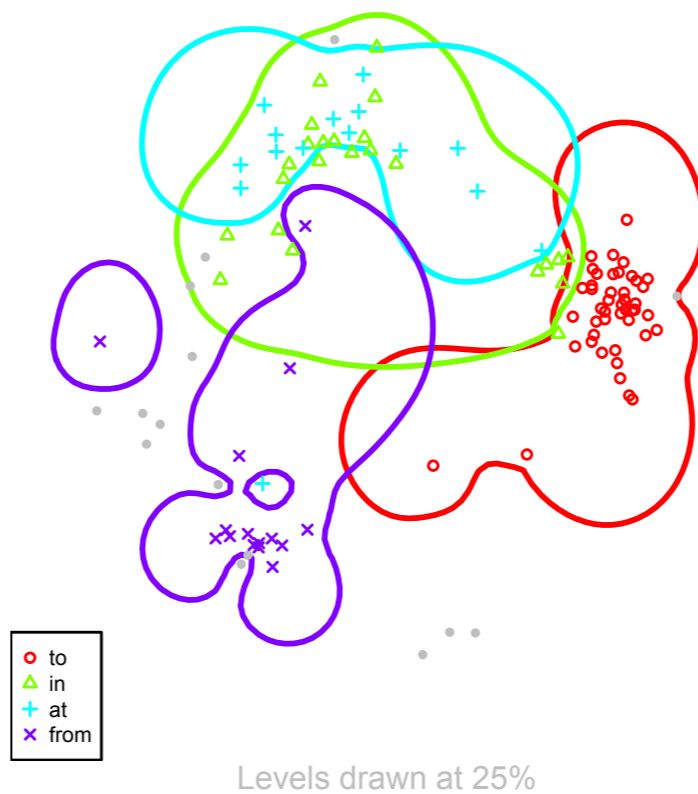
Low hanging fruit ...

- Semantic structure of locative case
- Complexity of locative case
- **Word order patterns**
- Phylogenetic correlation between word order and complexity
- Reconstruction of language families based on functional differences

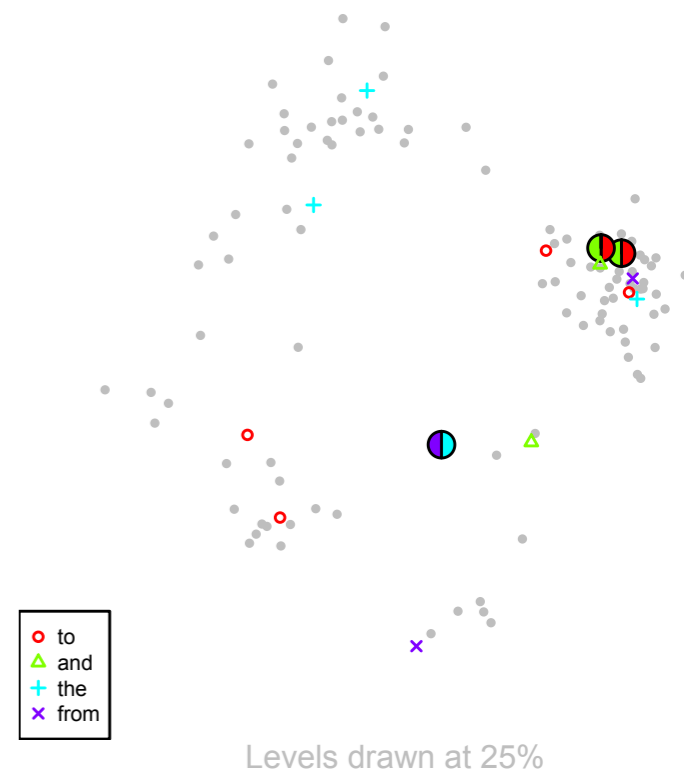
English (eng-x-bible-catholic.txt)
(wordforms)



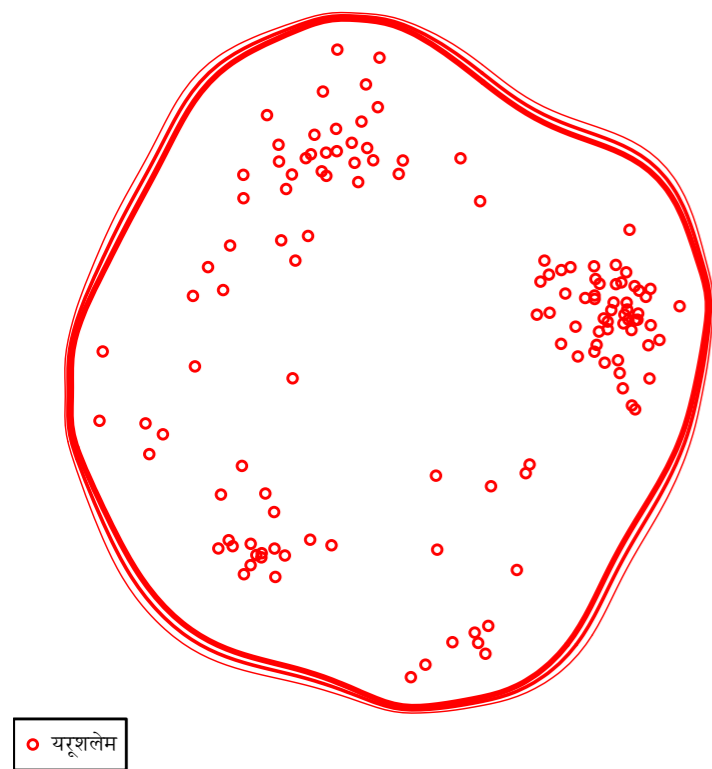
English (eng-x-bible-catholic.txt)
(before, binding: 18.59)



English (eng-x-bible-catholic.txt)
(after, binding: 2.33)

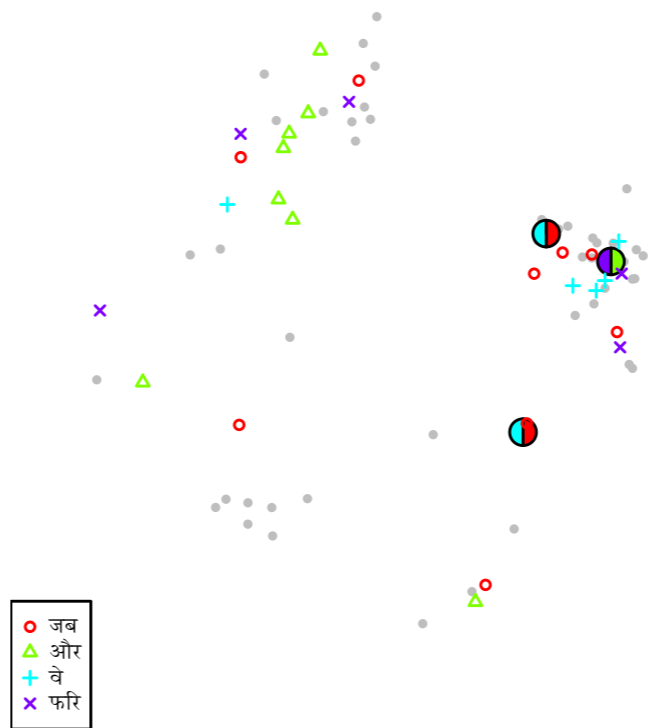


Hindi (hin-x-bible-easy.txt)
(wordforms)



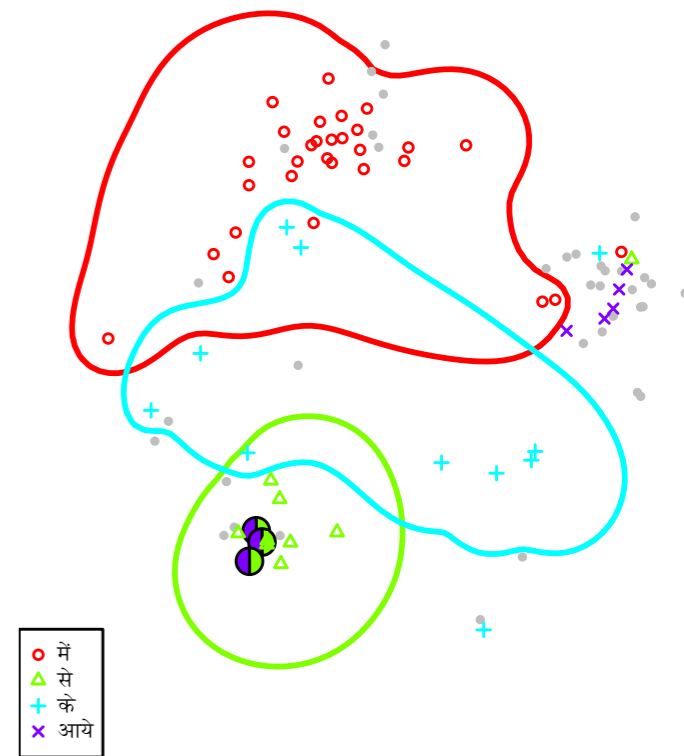
Levels drawn at 41-46-51%

Hindi (hin-x-bible-easy.txt)
(before, binding: 5.28)

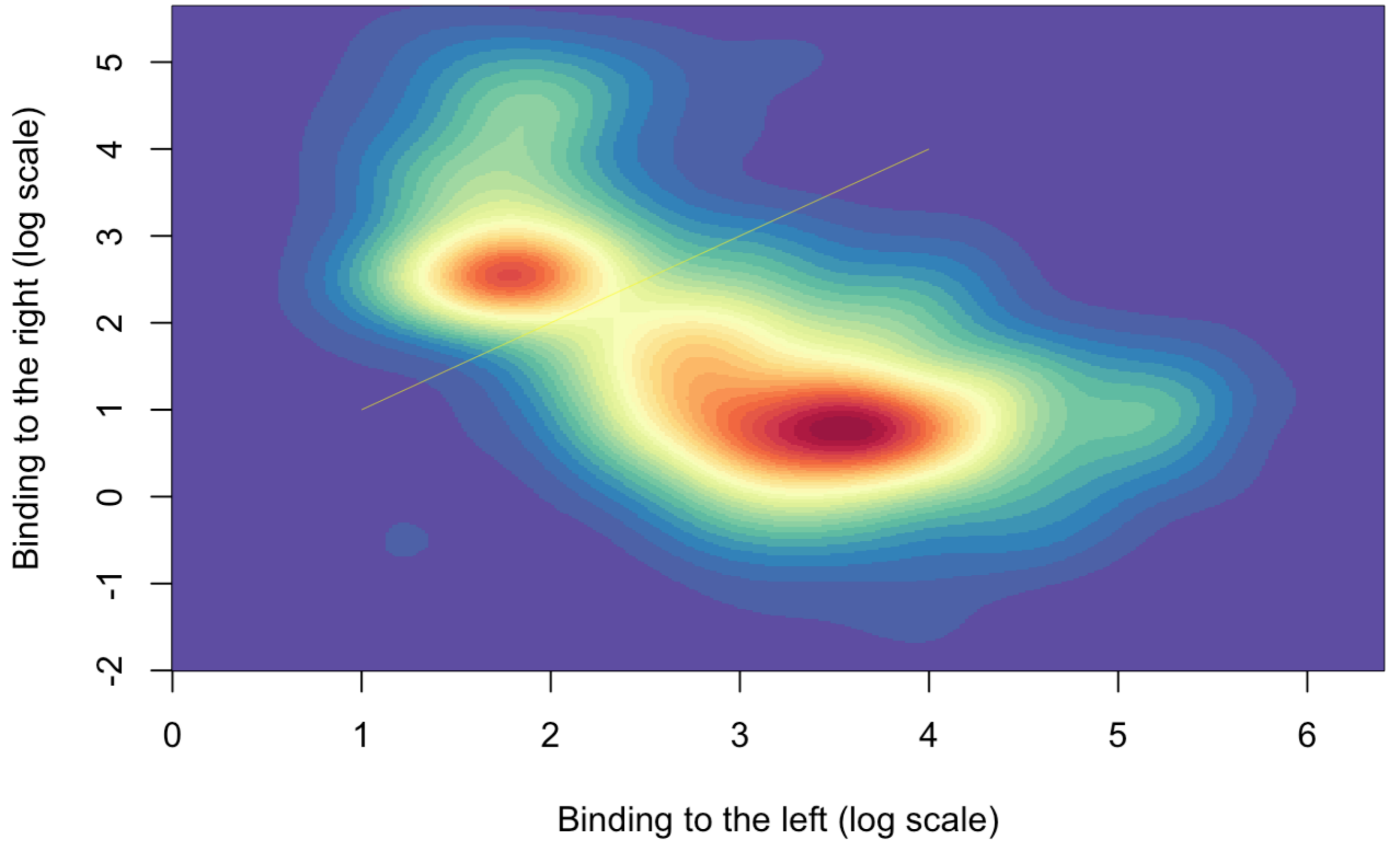


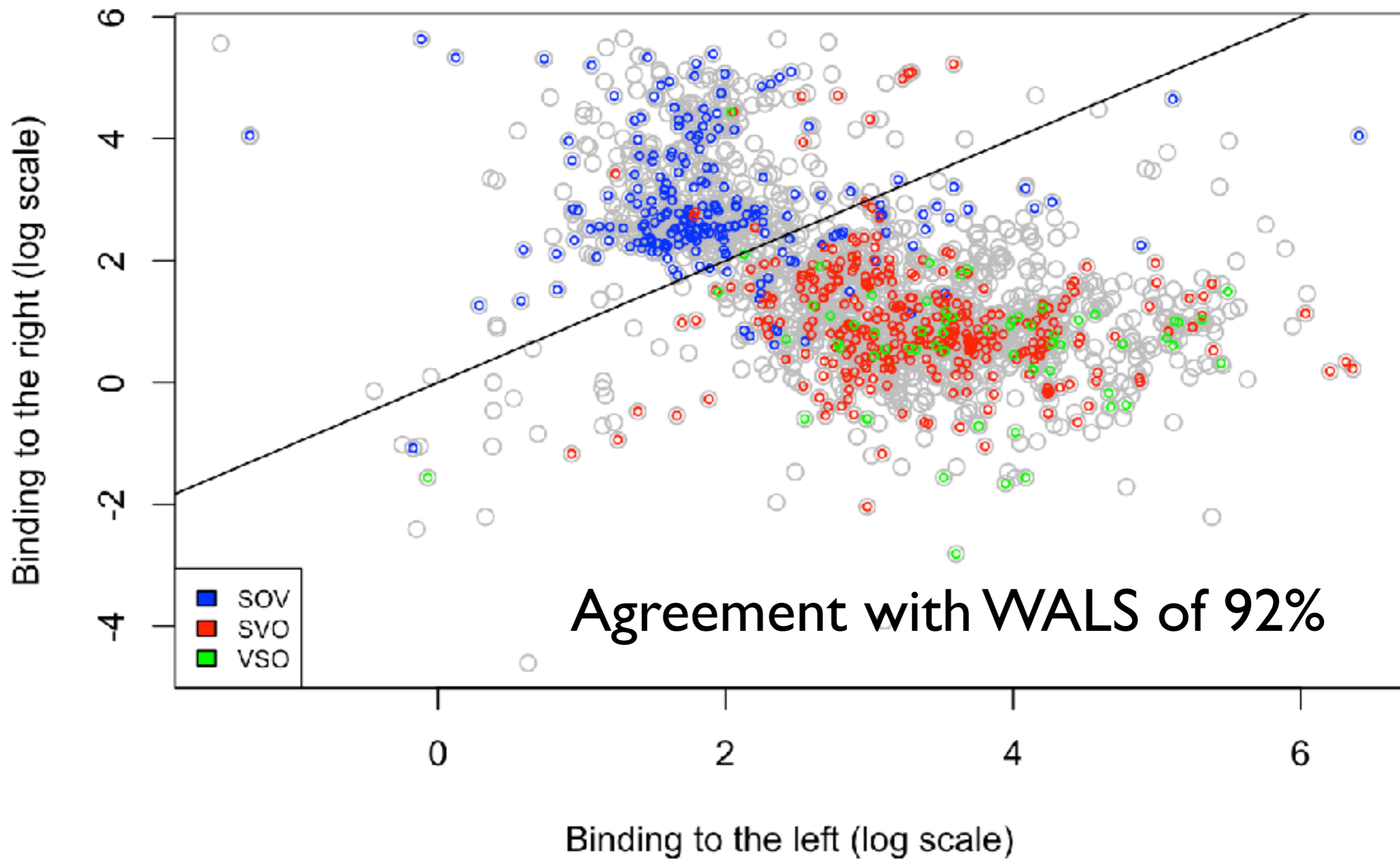
Levels drawn at 25%

Hindi (hin-x-bible-easy.txt)
(after, binding: 30.75)



Levels drawn at 25%





“OV”

Difference in binding (right - left)

6

4

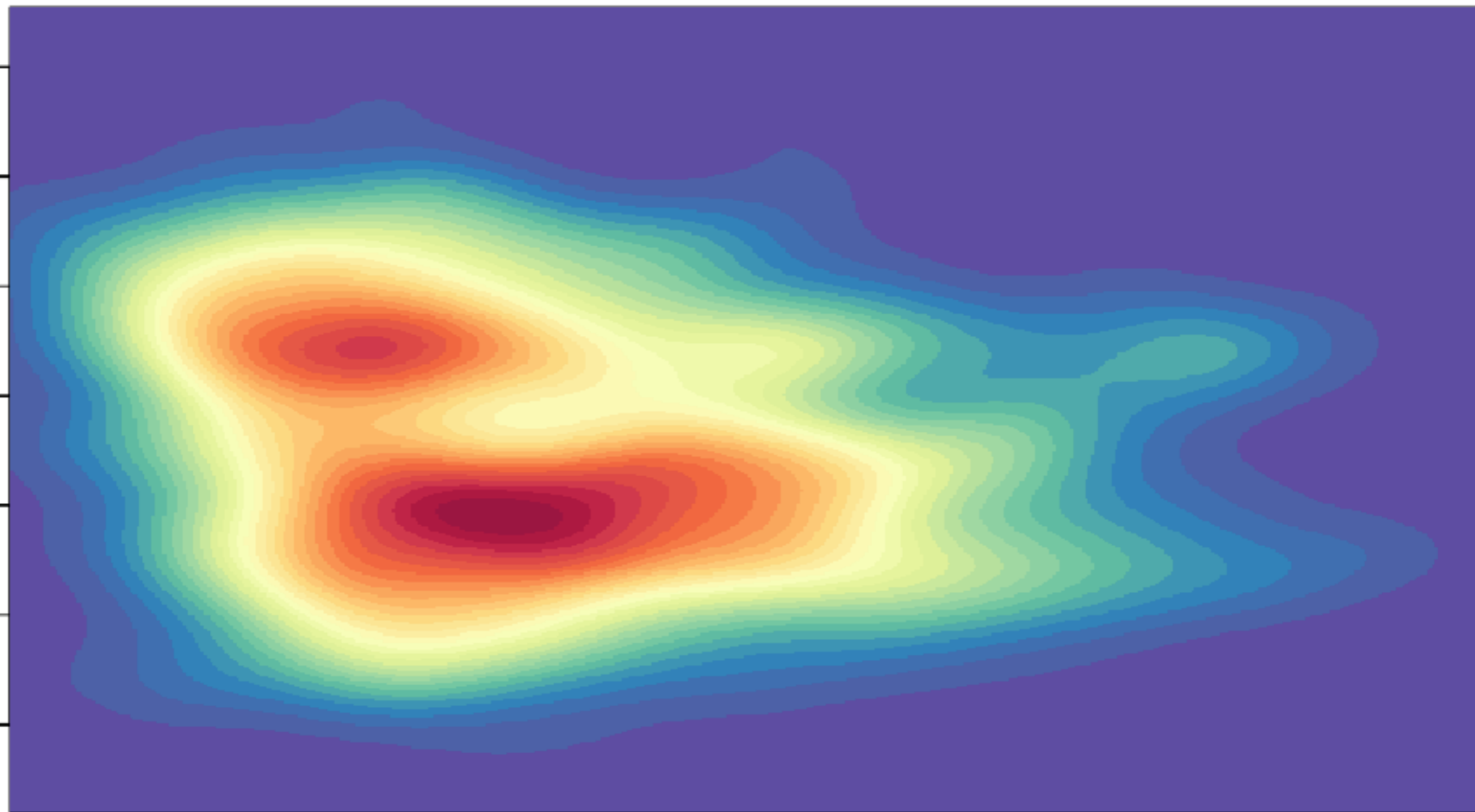
2

0

-2

-4

-6



3

4

5

6

7

8

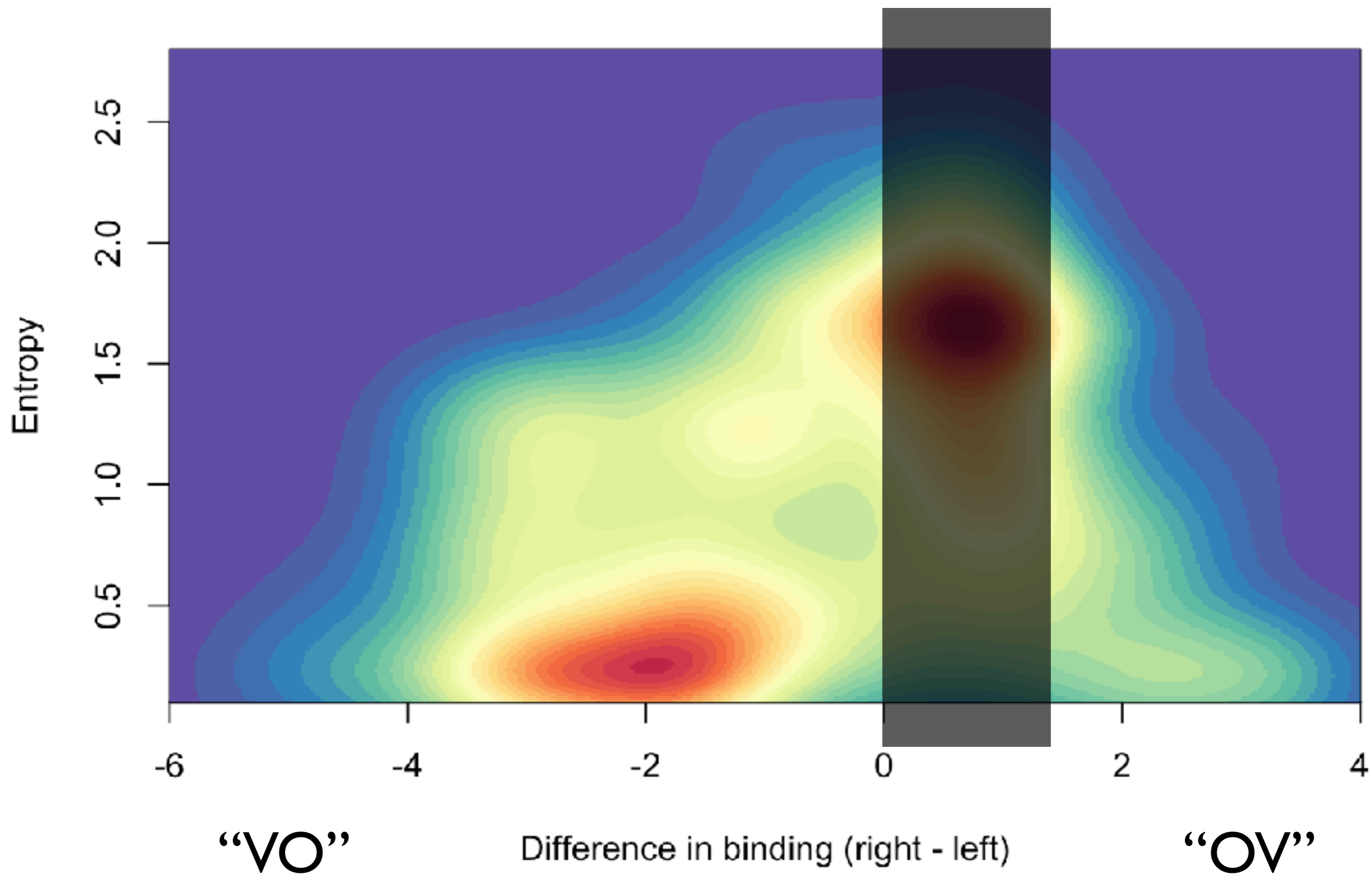
Speaker community size (log10)

“VO”

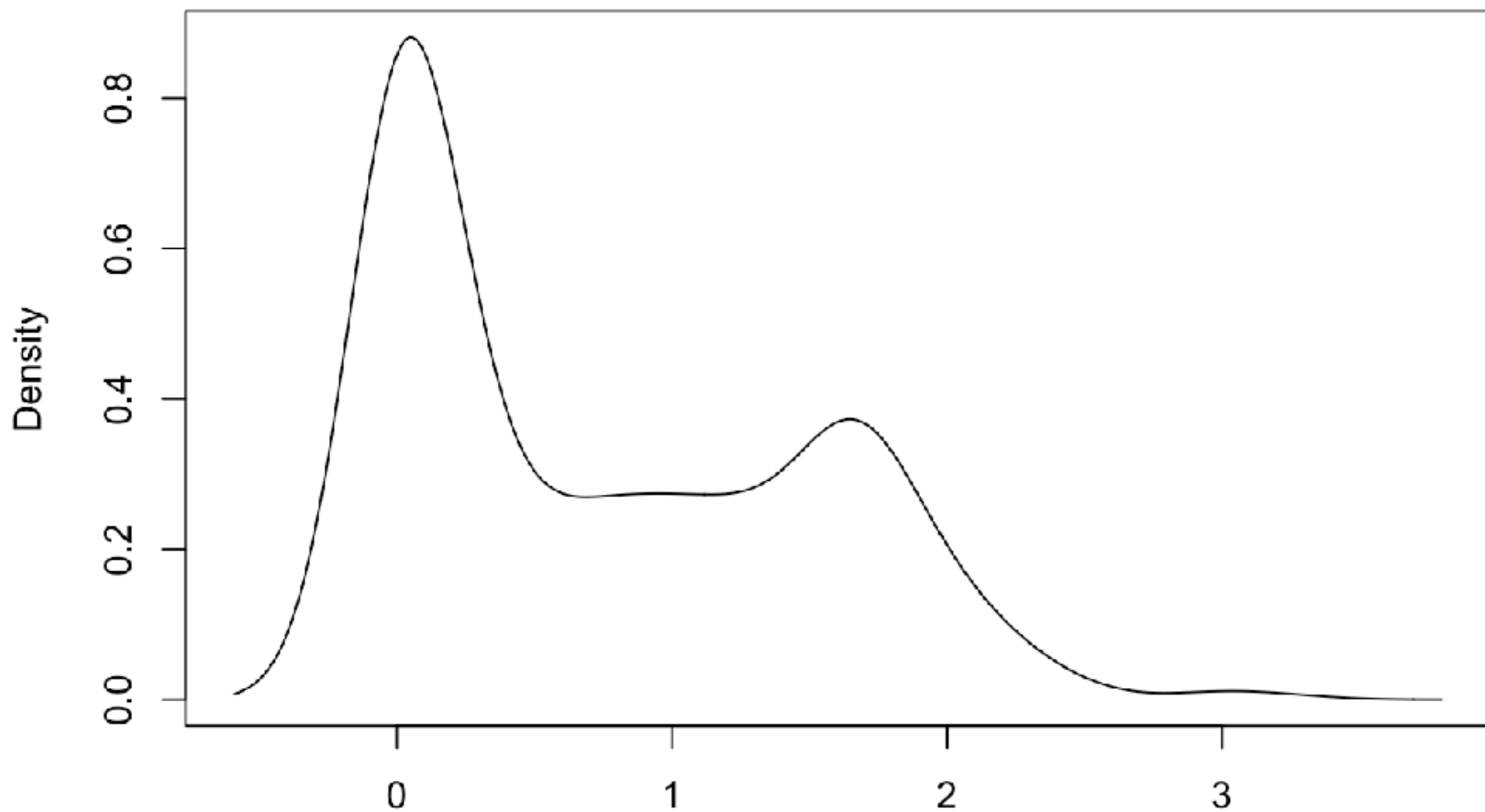


Low hanging fruit ...

- Semantic structure of locative case
- Complexity of locative case
- Word order patterns
- **Phylogenetic correlation between word order and complexity**
- Reconstruction of language families based on functional differences



Density of entropy (only OV languages)



N = 477 Bandwidth = 0.1973

Phylogenetic modeling



High entropy (1.5+)



Mid entropy (0.1 - 1.5)

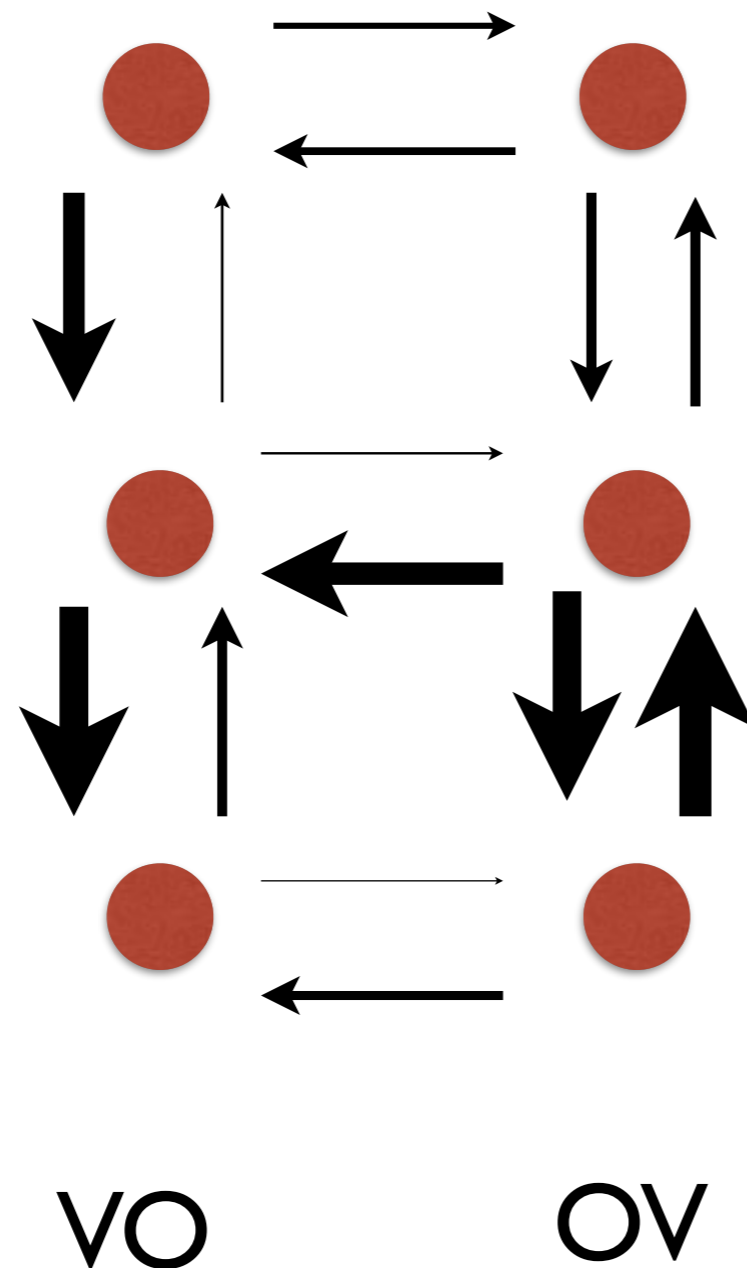


Low entropy (0 - 0.1)

VO

OV

Phylogenetic modeling



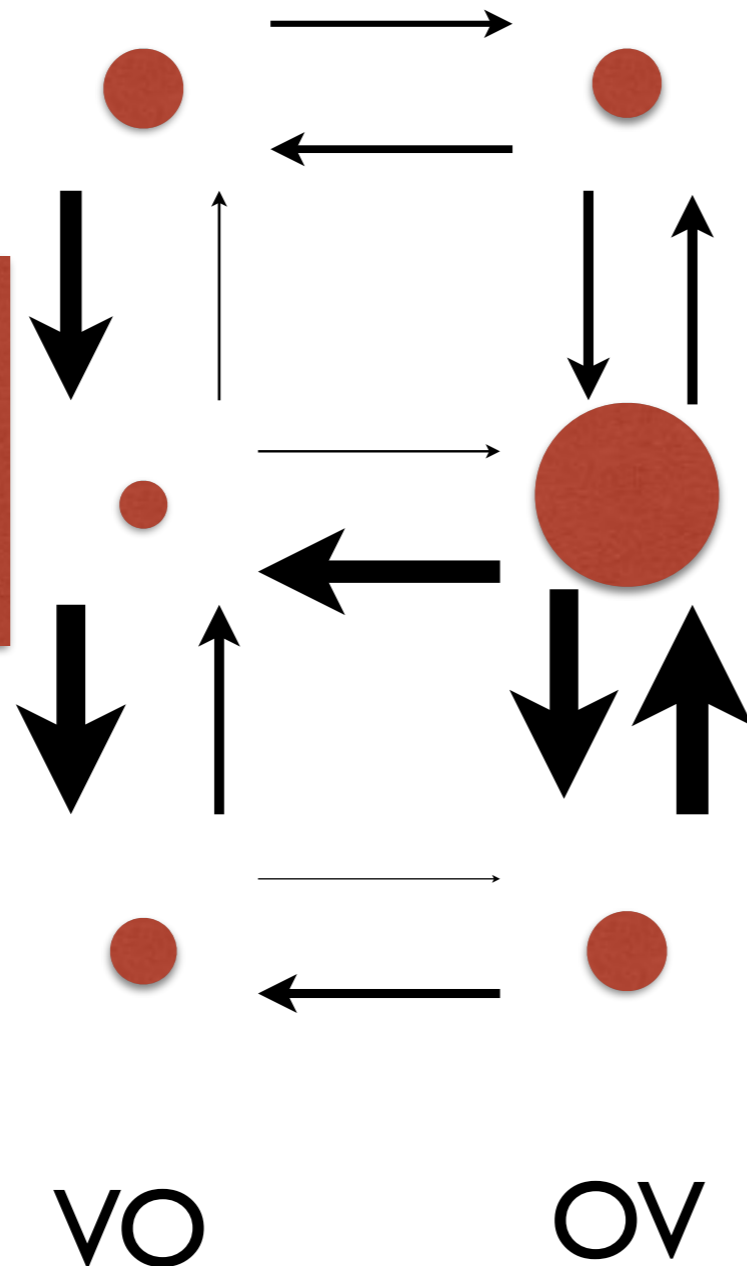
High entropy (1.5+)

Mid entropy (0.1 - 1.5)

Low entropy (0 - 0.1)

Phylogenetic modeling

Average of reconstruction per family



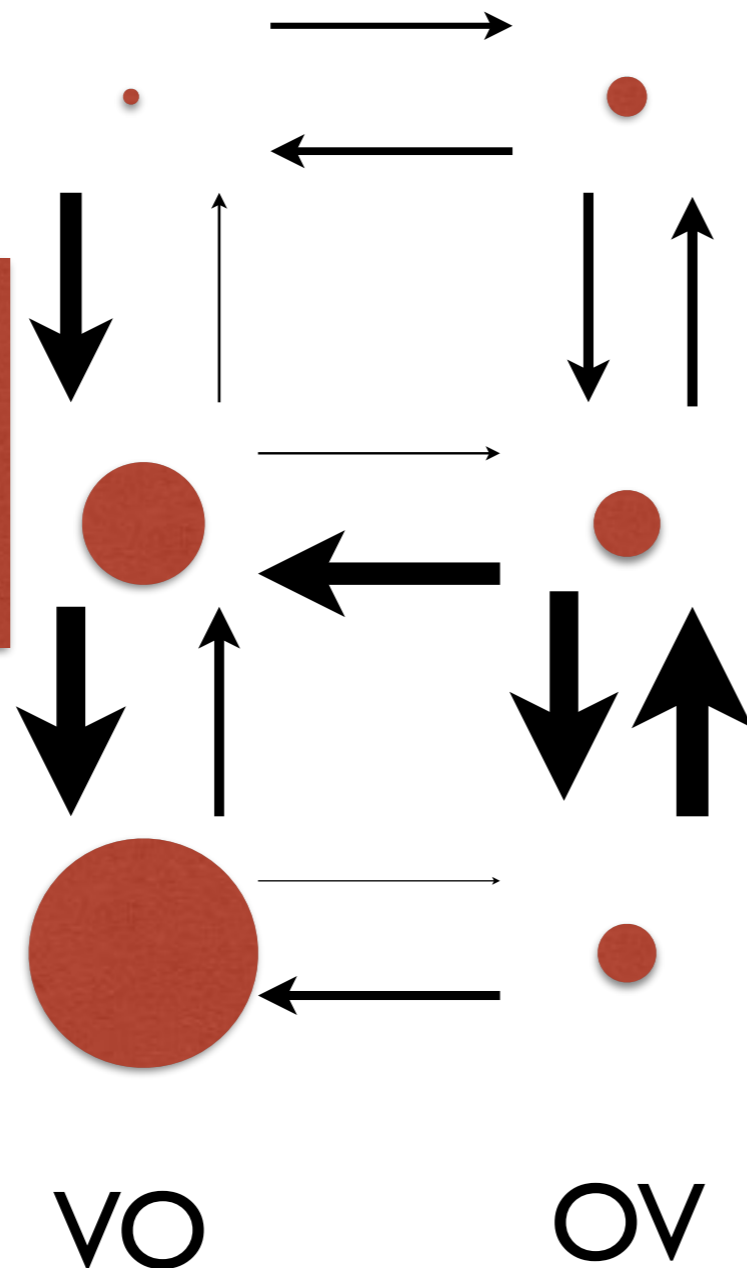
High entropy (1.5+)

Mid entropy (0.1 - 1.5)

Low entropy (0 - 0.1)

Phylogenetic modeling

Actual frequencies of languages



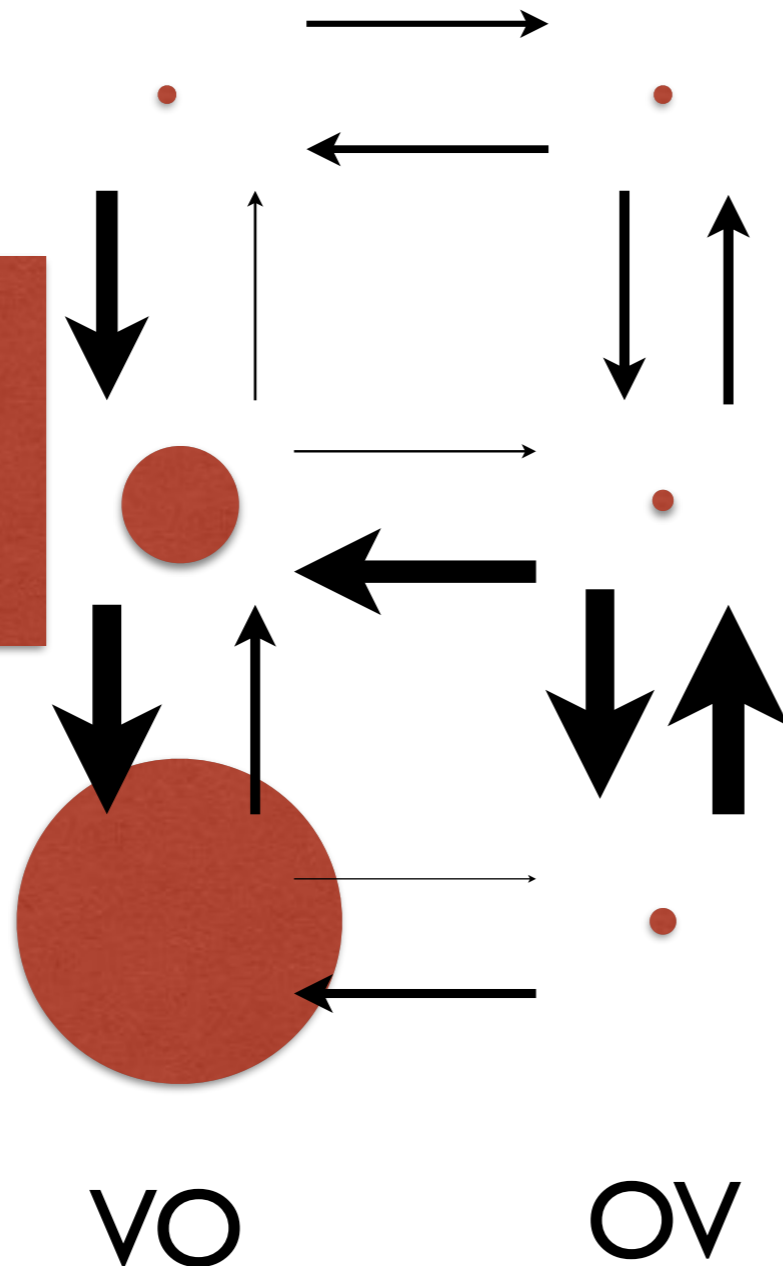
High entropy (1.5+)

Mid entropy (0.1 - 1.5)

Low entropy (0 - 0.1)

Phylogenetic modeling

Steady state



High entropy (1.5+)

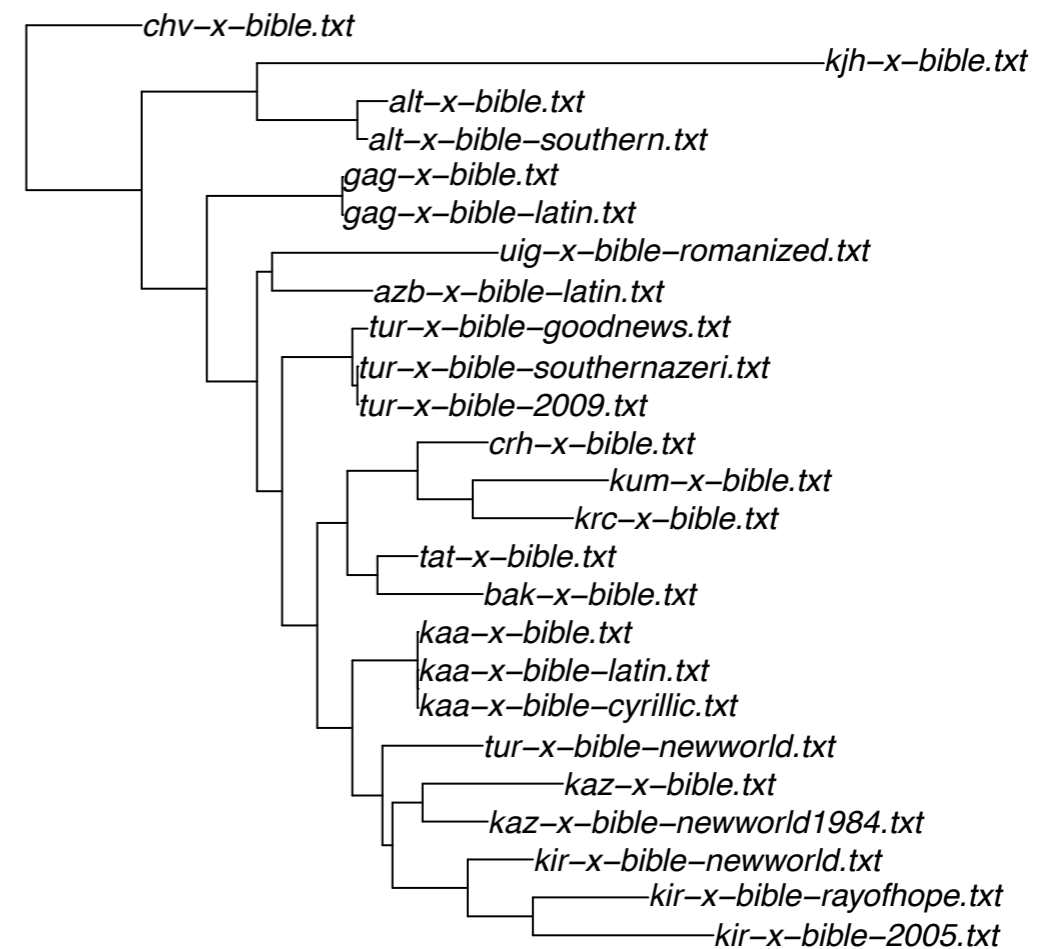
Mid entropy (0.1 - 1.5)

Low entropy (0 - 0.1)



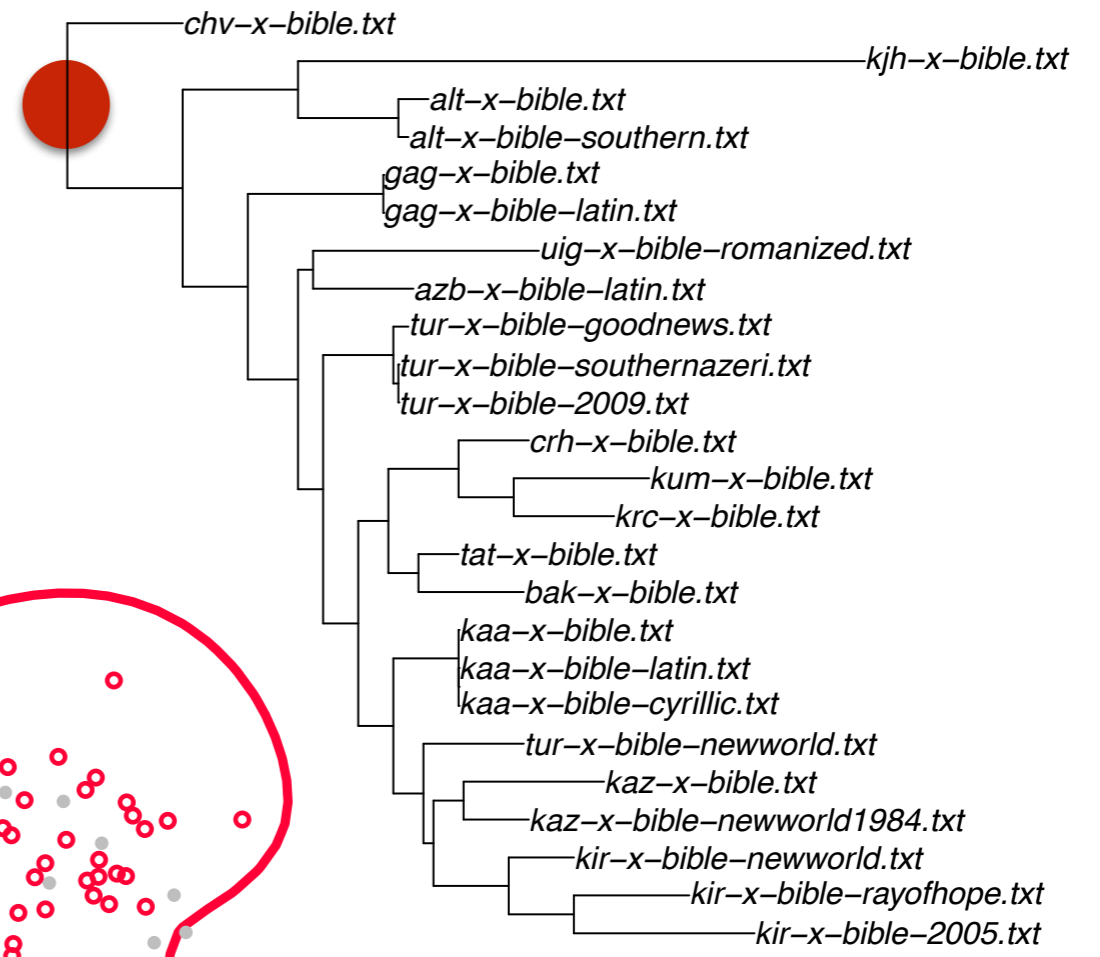
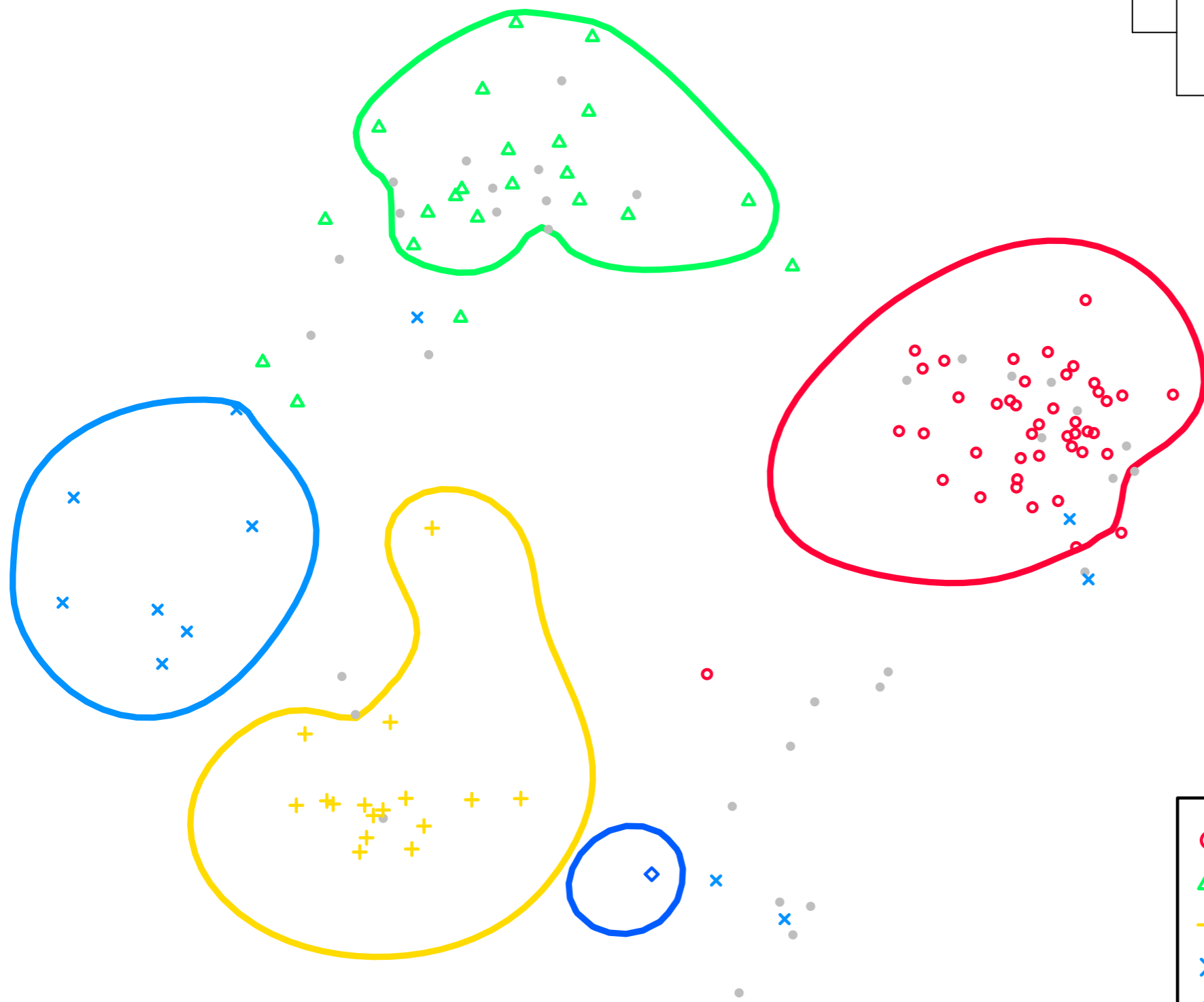
Low hanging fruit ...

- Semantic structure of locative case
- Complexity of locative case
- Word order patterns
- Phylogenetic correlation between word order and complexity
- **Reconstruction of language families based on functional differences**



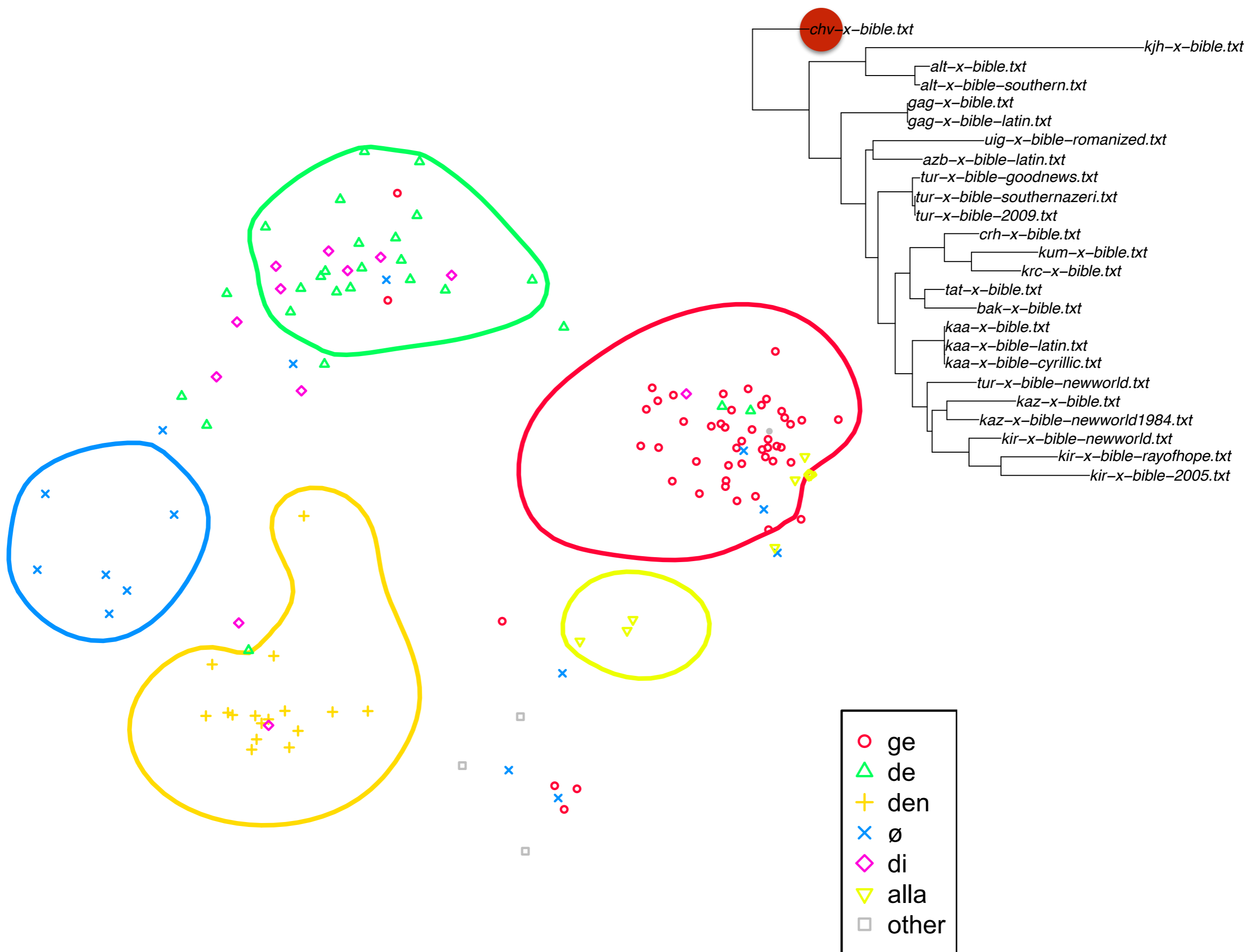
● Phylogenetic reconstruction

- ▶ Using distribution of cognate affixes over contexts in parallel text (Cysouw & Forker 2009)
- ▶ Simple parsimony ratchet from R-library 'phangorn'
- ▶ Turkic family as an example

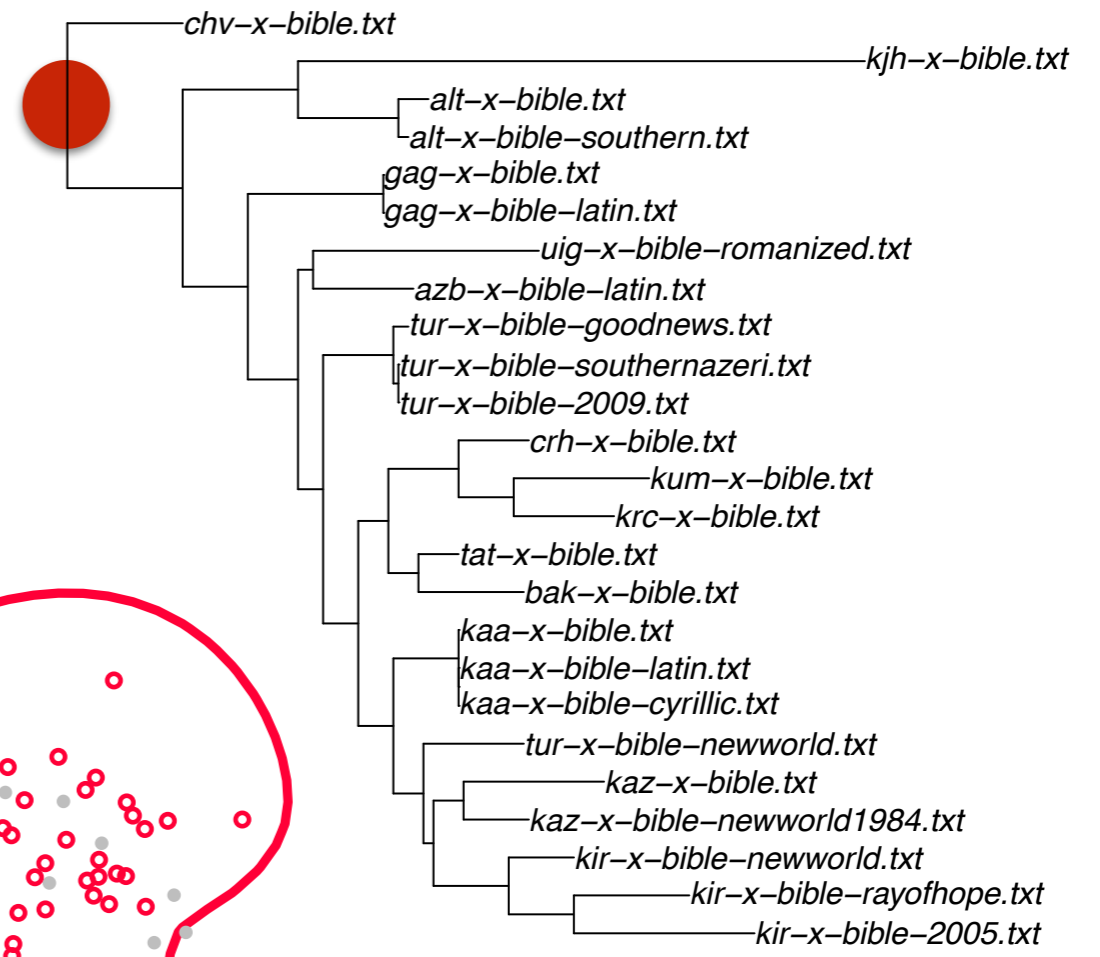
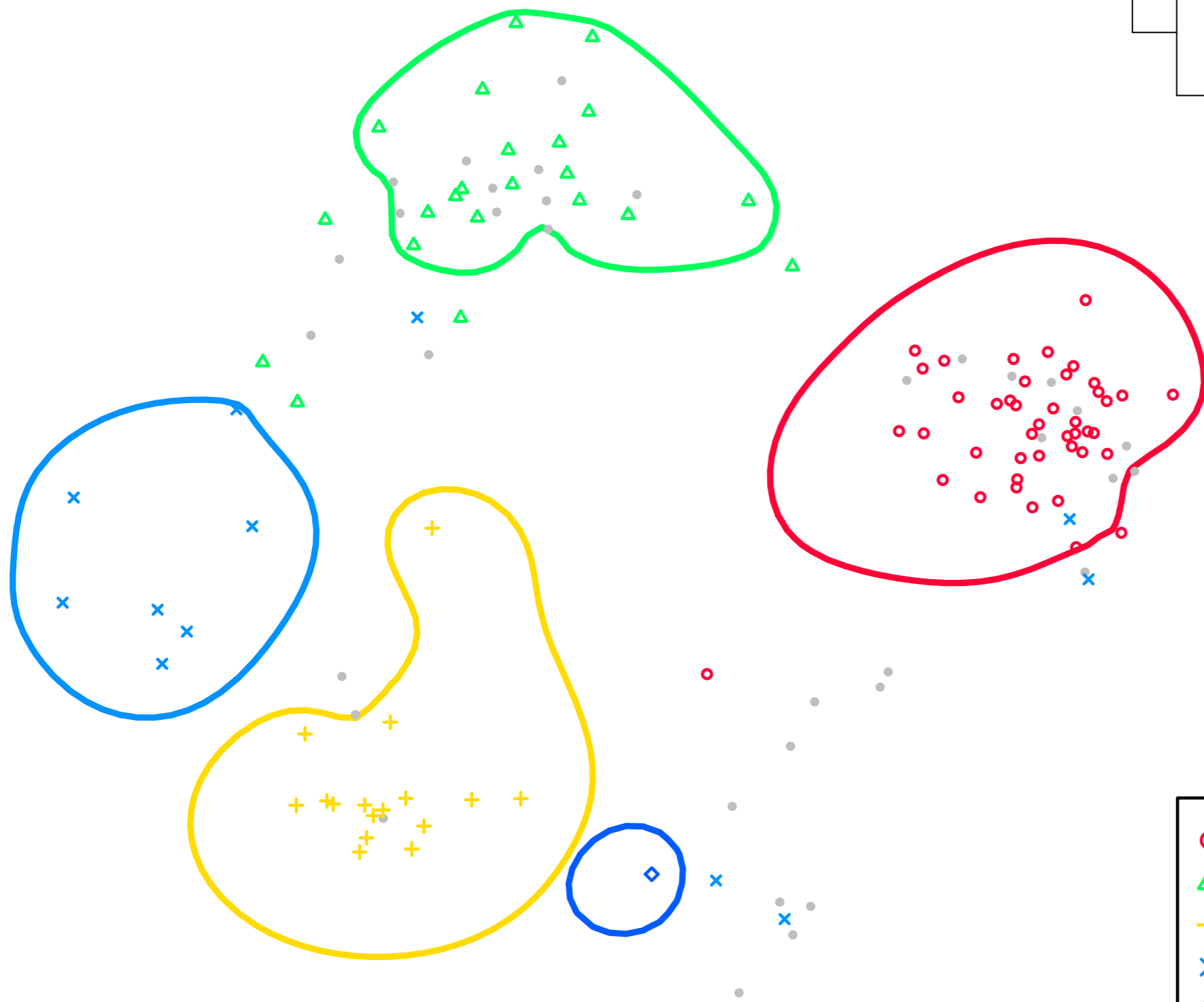


- ge
- △ de
- + den
- × ø
- ◇ in
- ▽ alla
- other

Levels drawn at 50%

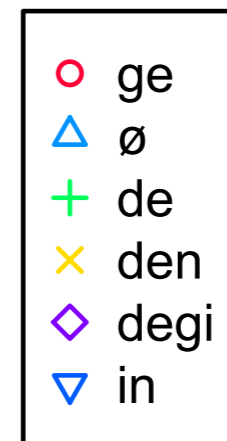
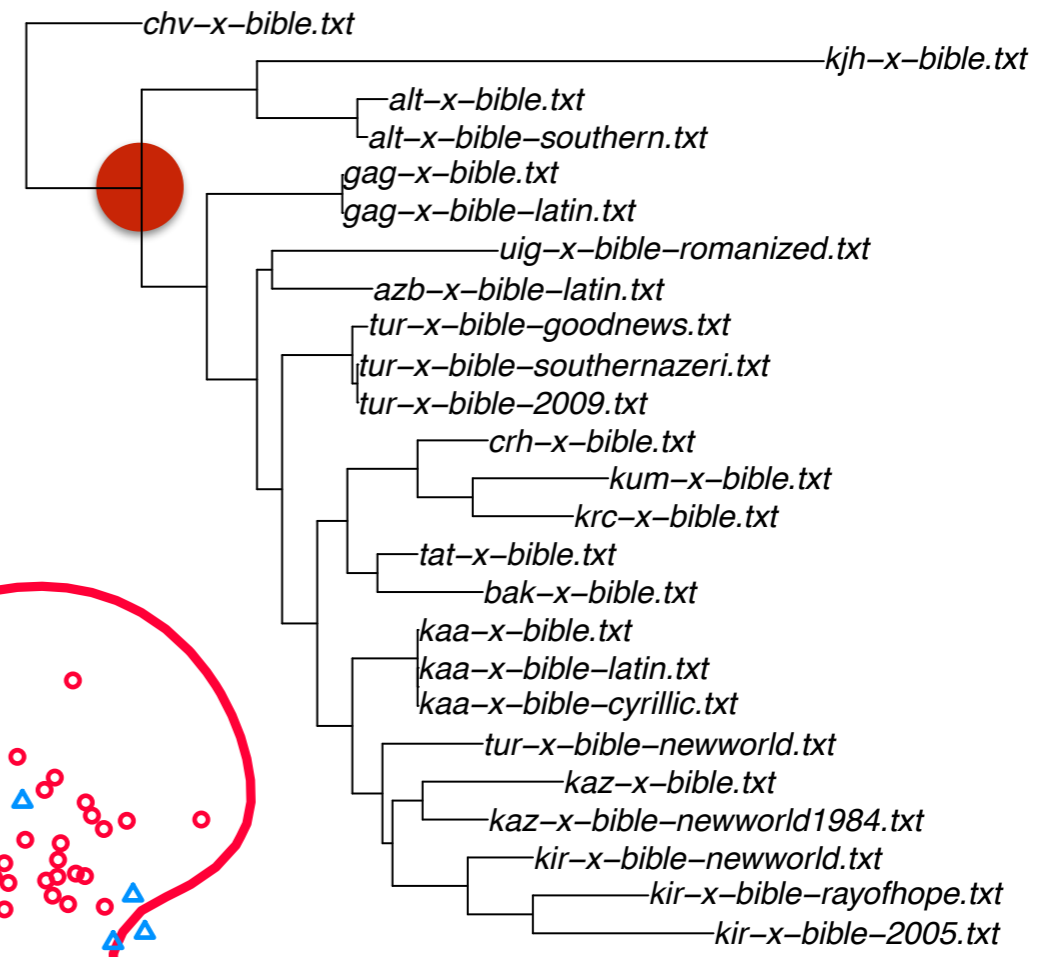
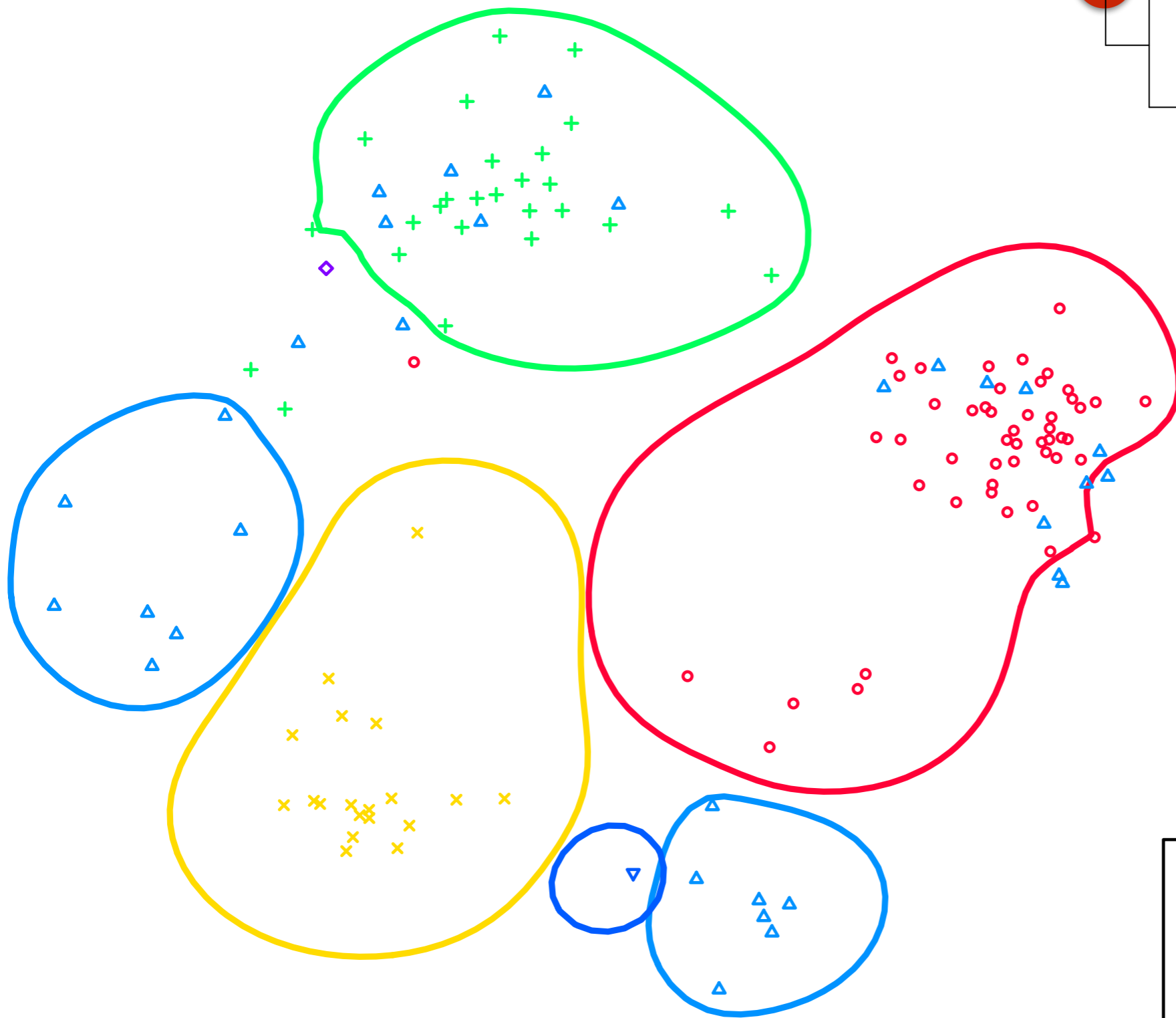


Levels drawn at 50%

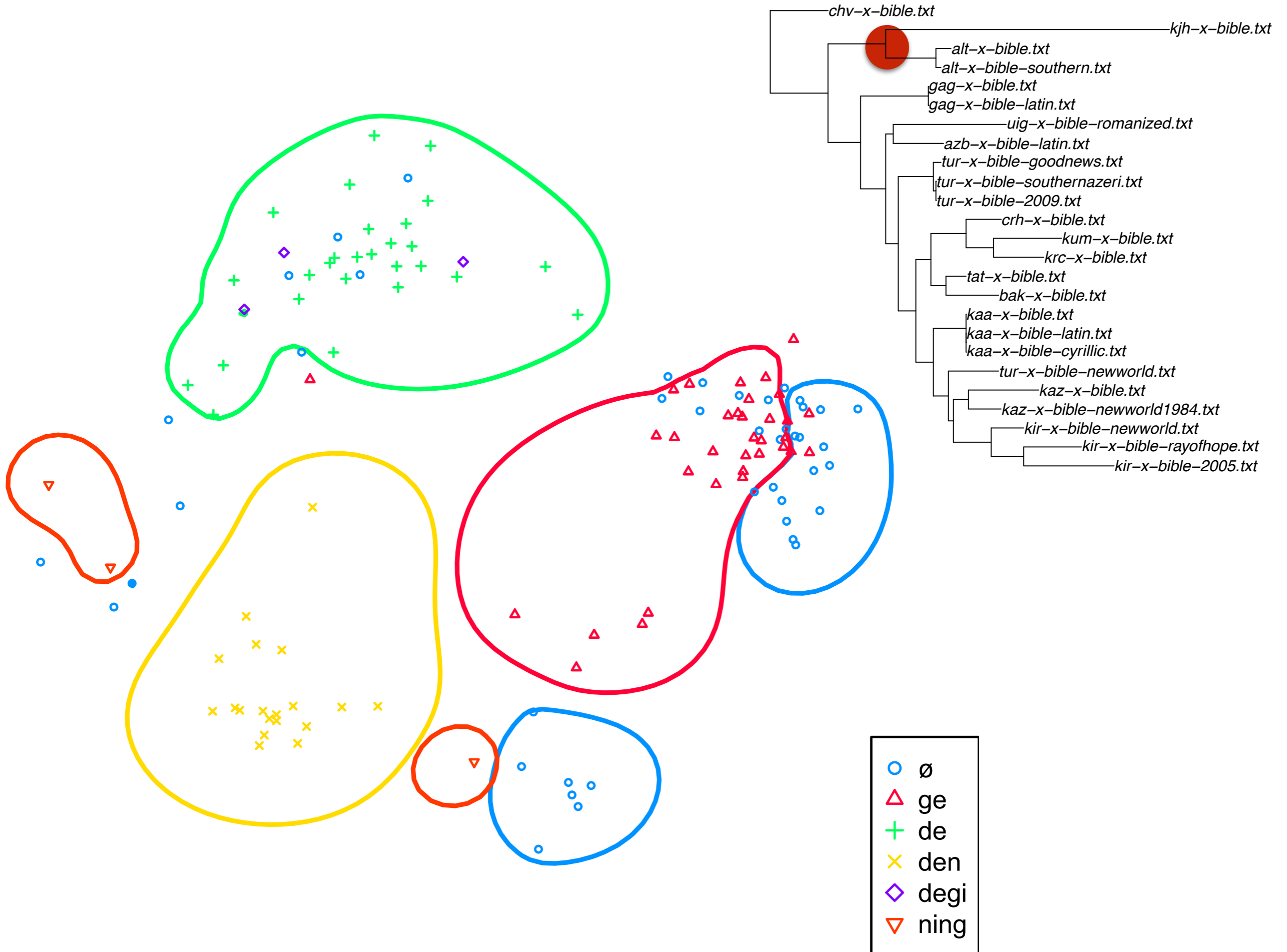


- ge
- △ de
- + den
- × ø
- ◇ in
- ▽ alla
- other

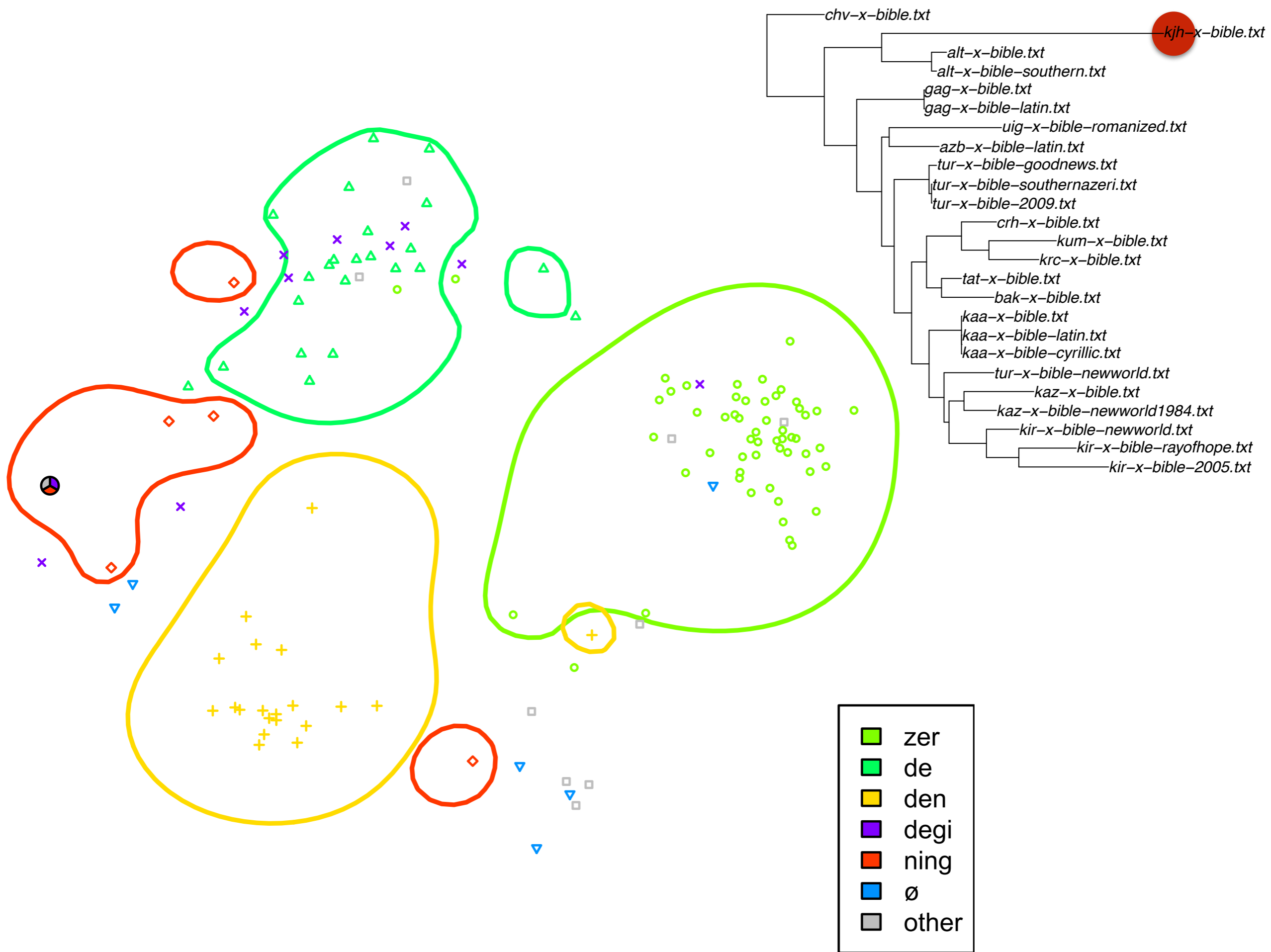
Levels drawn at 50%



Levels drawn at 50%

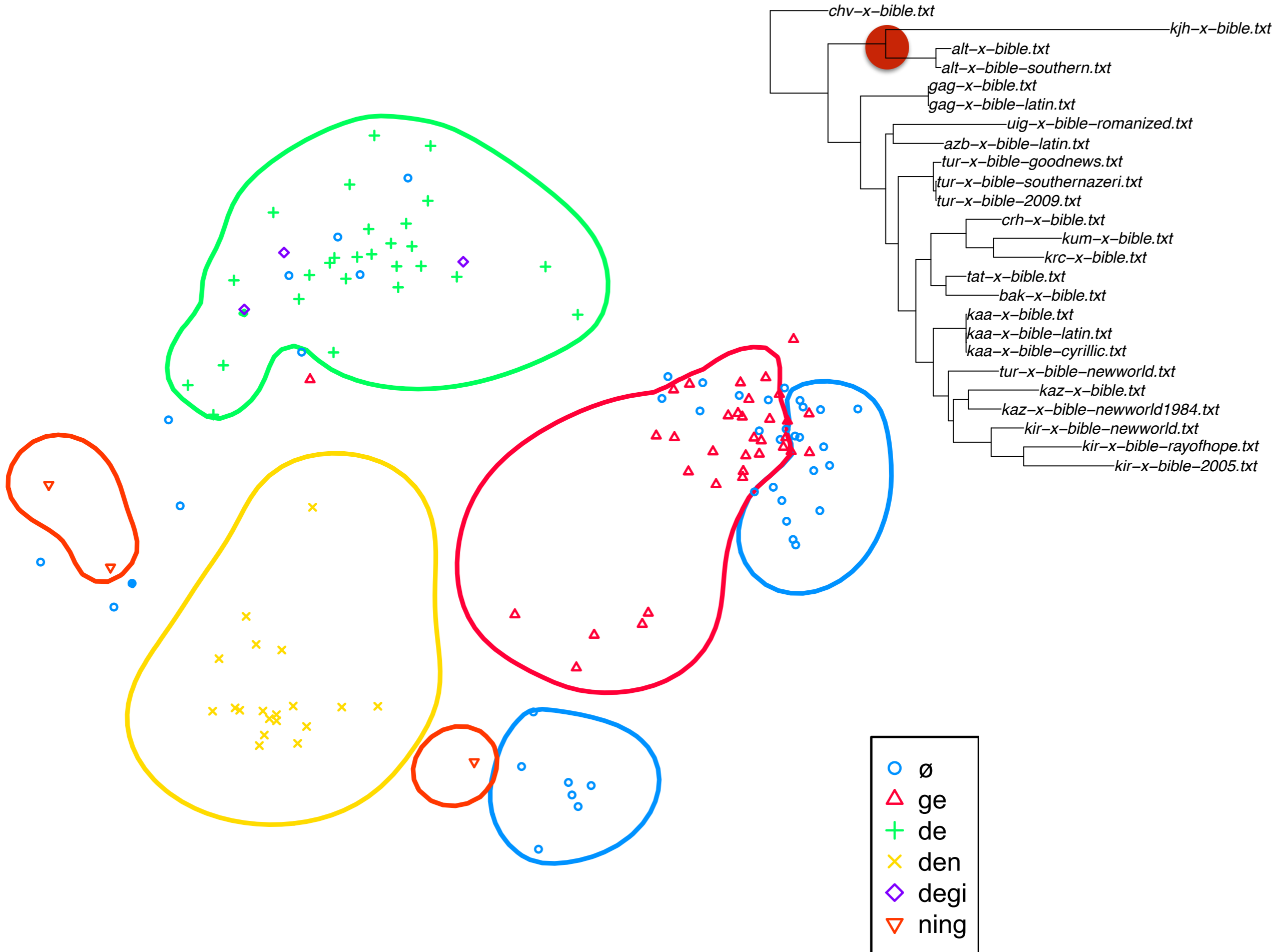


Levels drawn at 50%

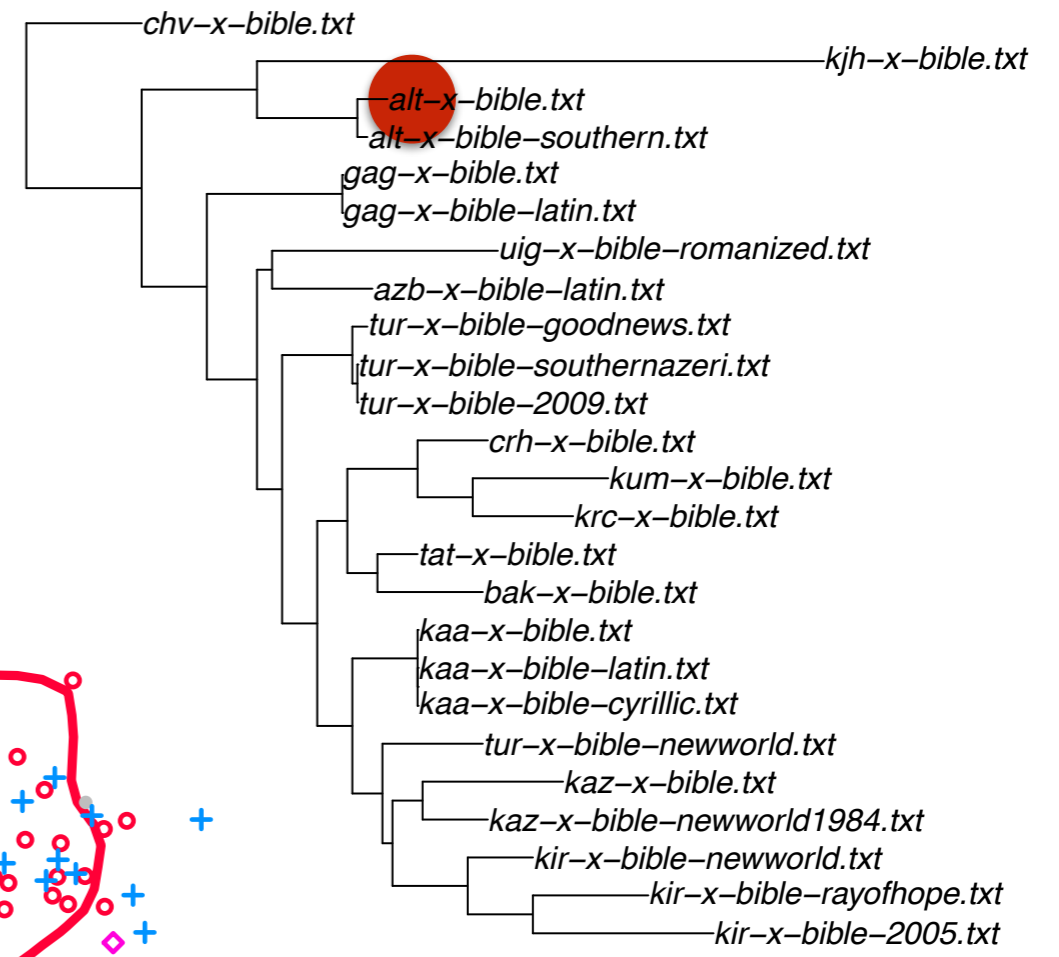
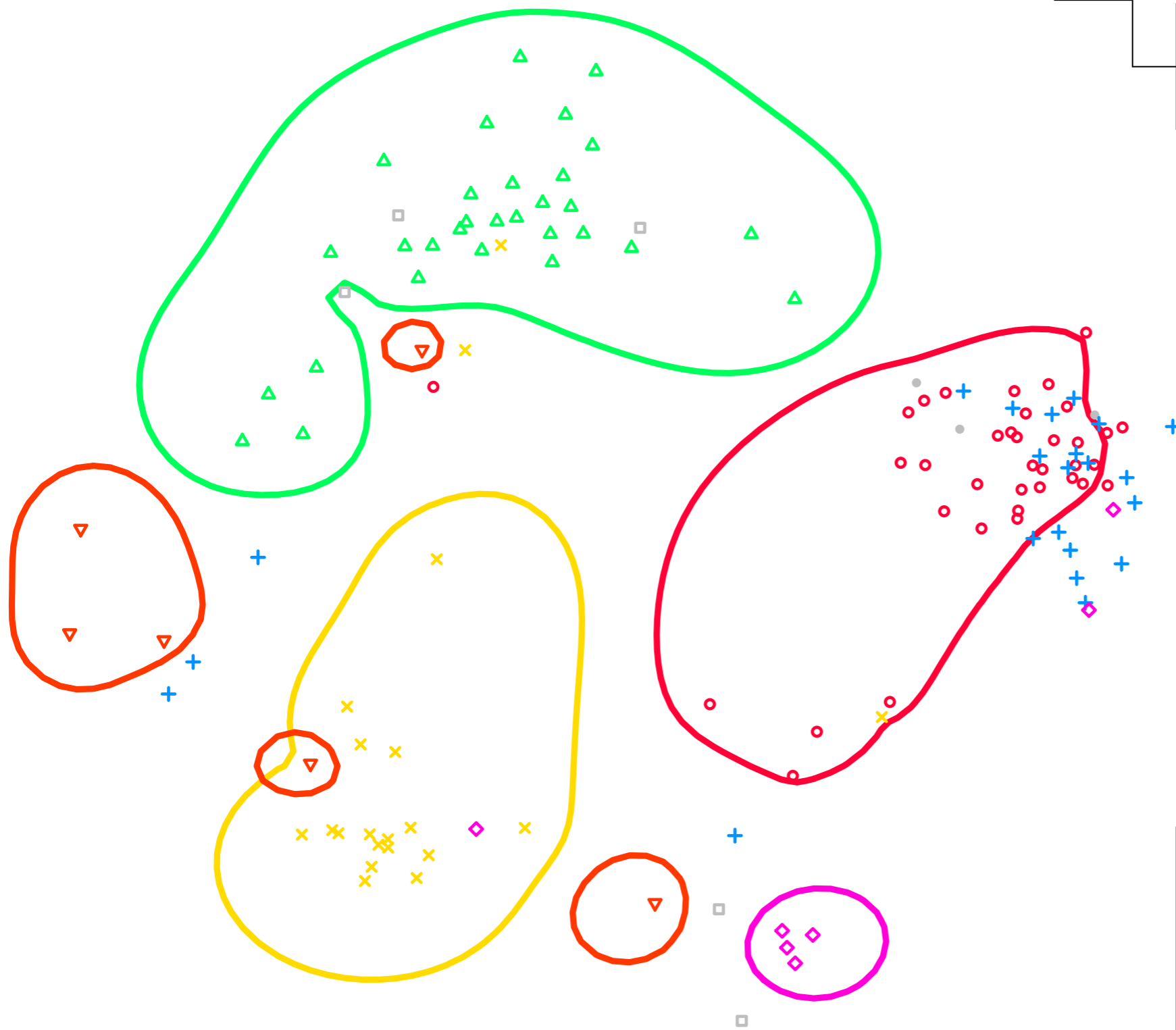


- zer
- de
- den
- degi
- ning
- ø
- other

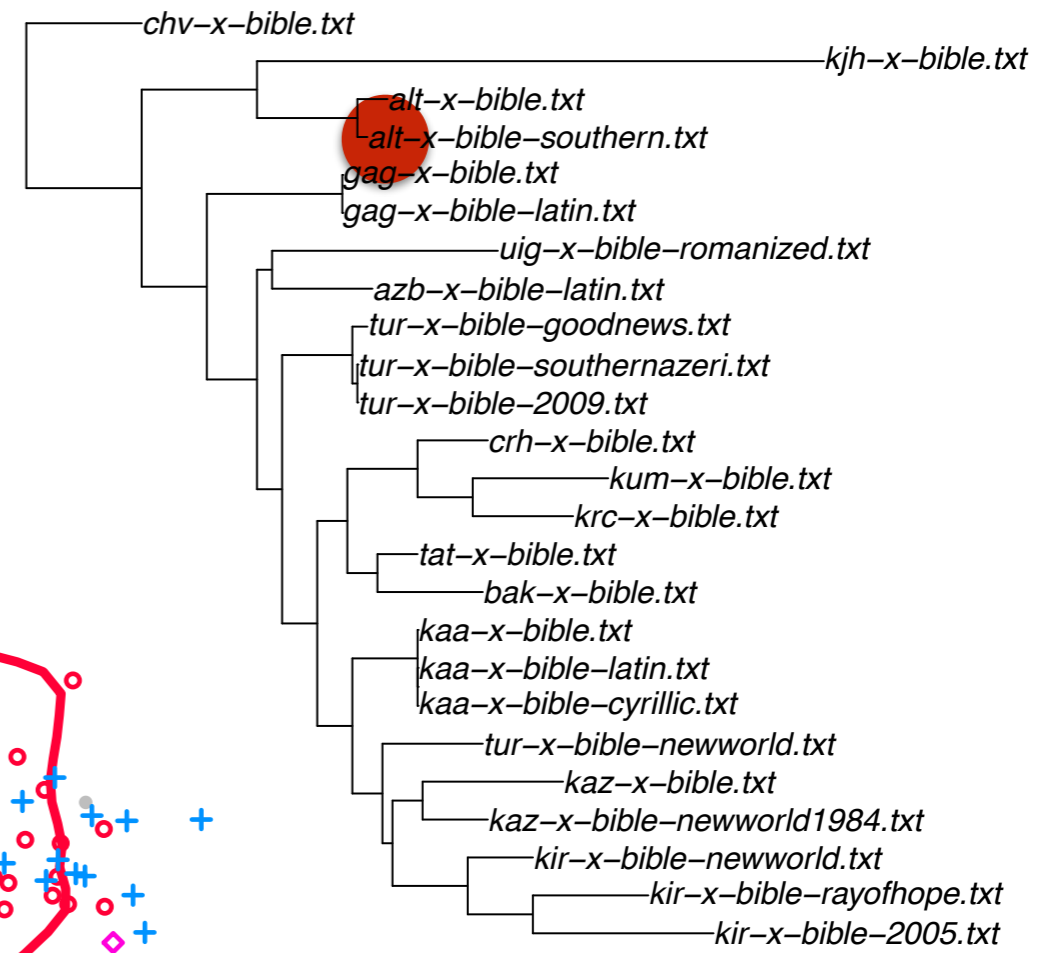
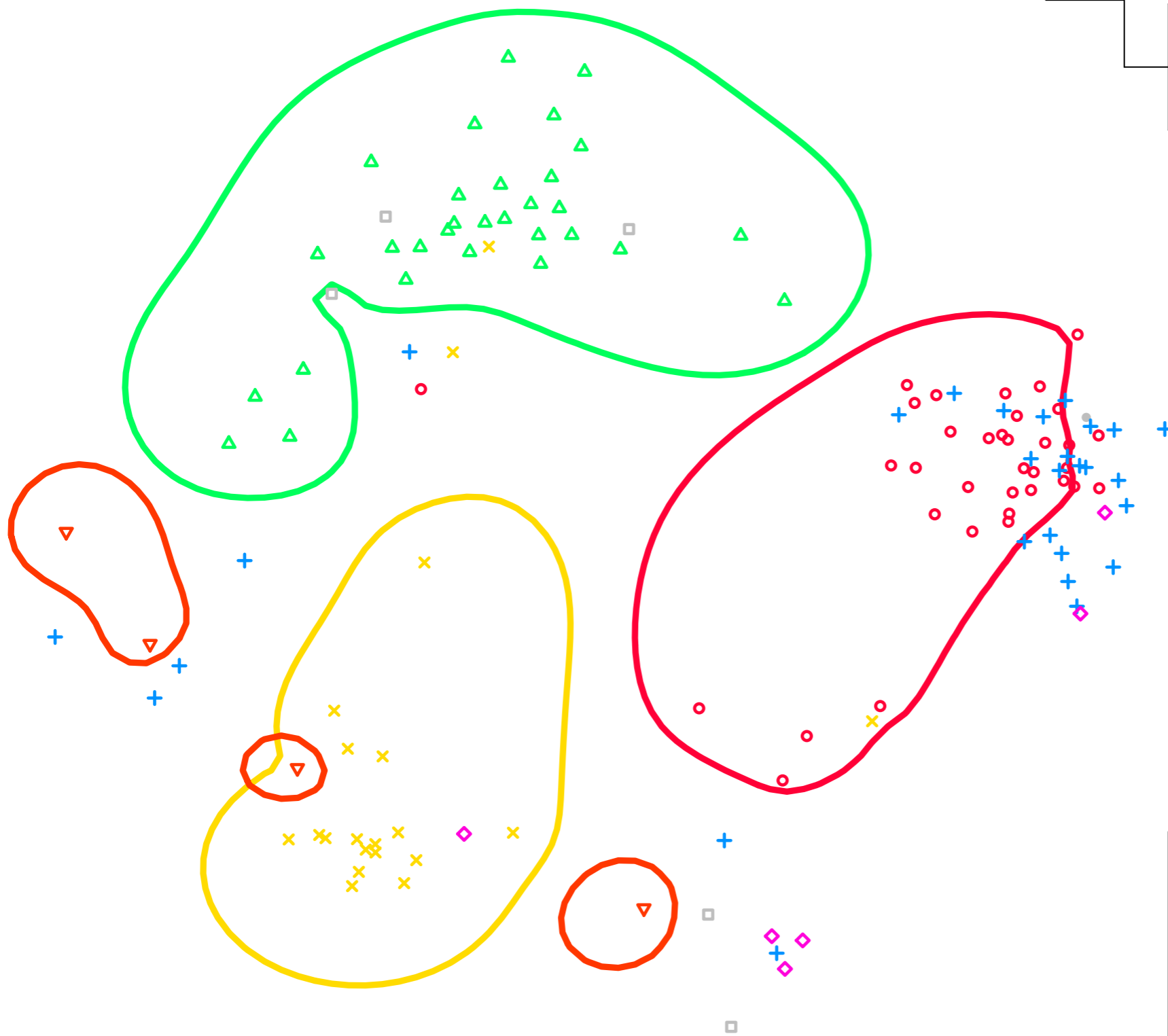
Levels drawn at 50%



Levels drawn at 50%

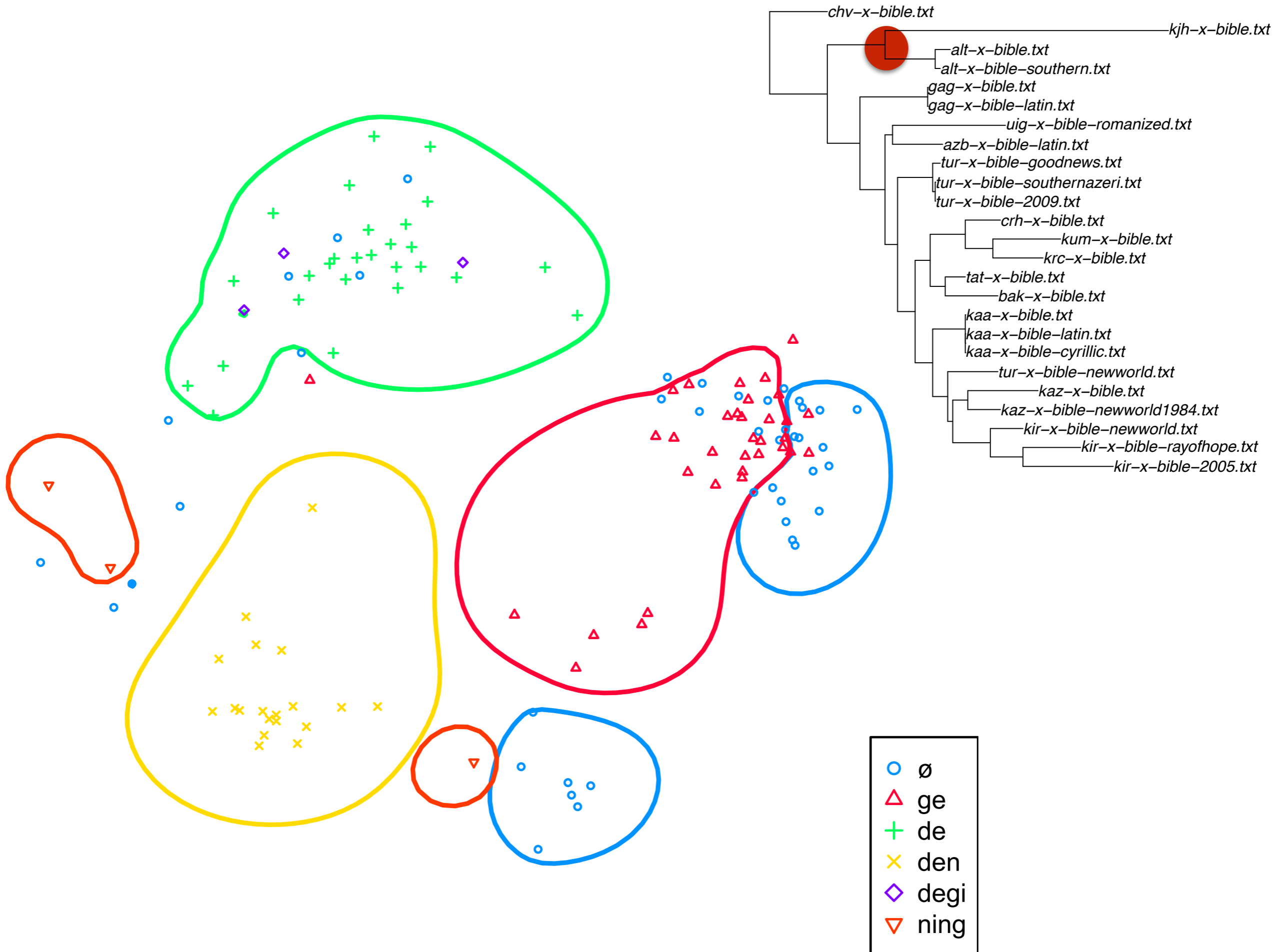


Levels drawn at 50%

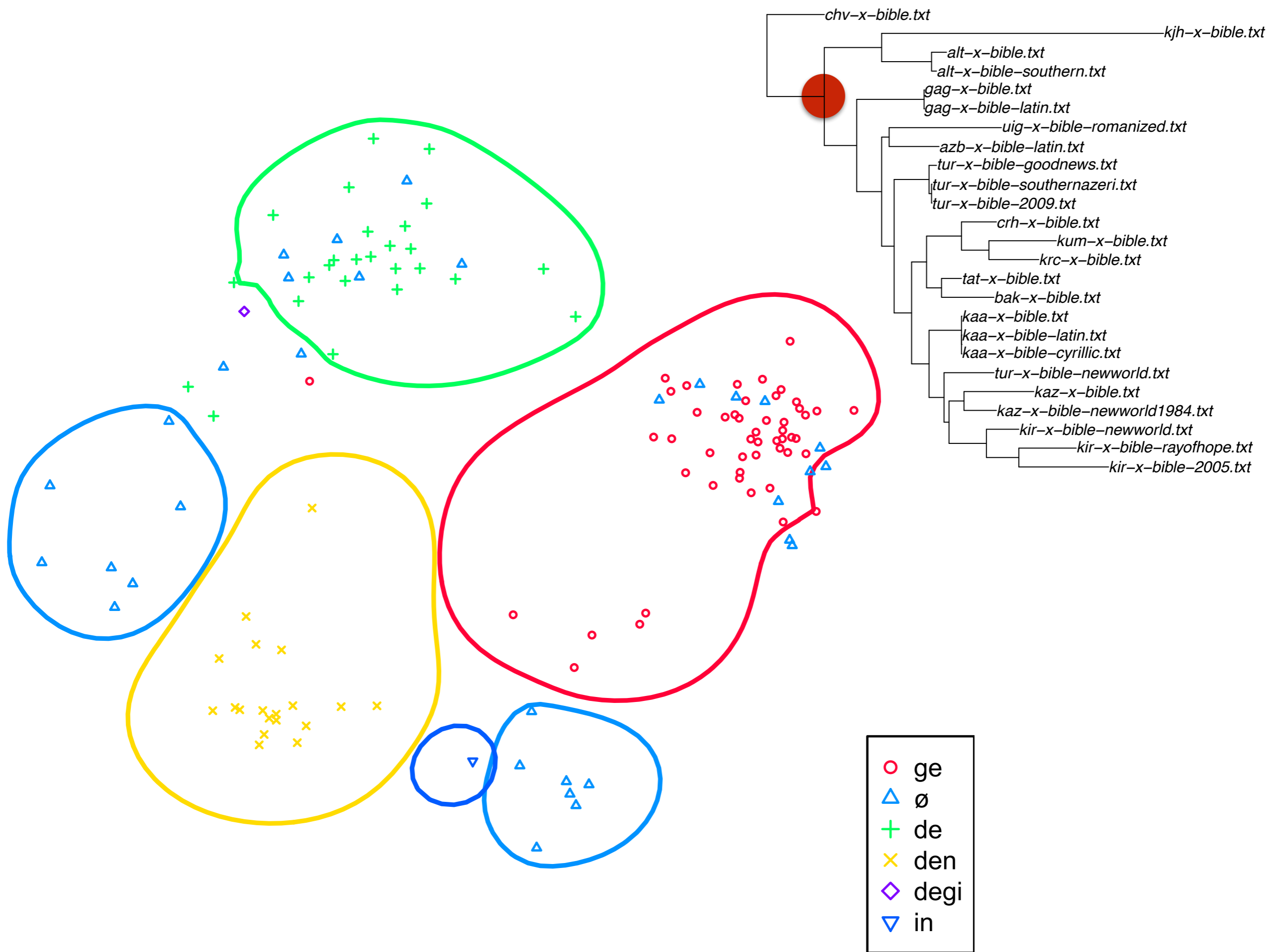


- ge
- △ de
- + ø
- × den
- ◇ di
- ▽ ning
- other

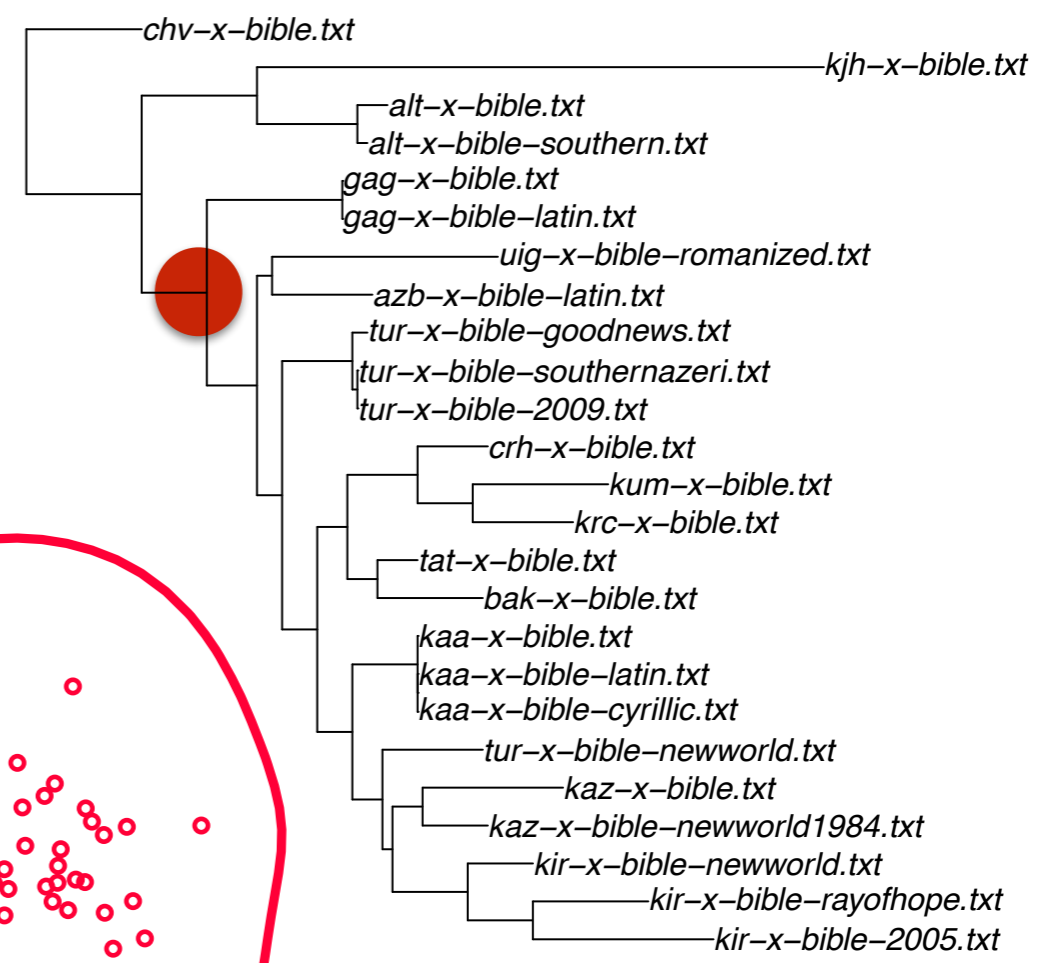
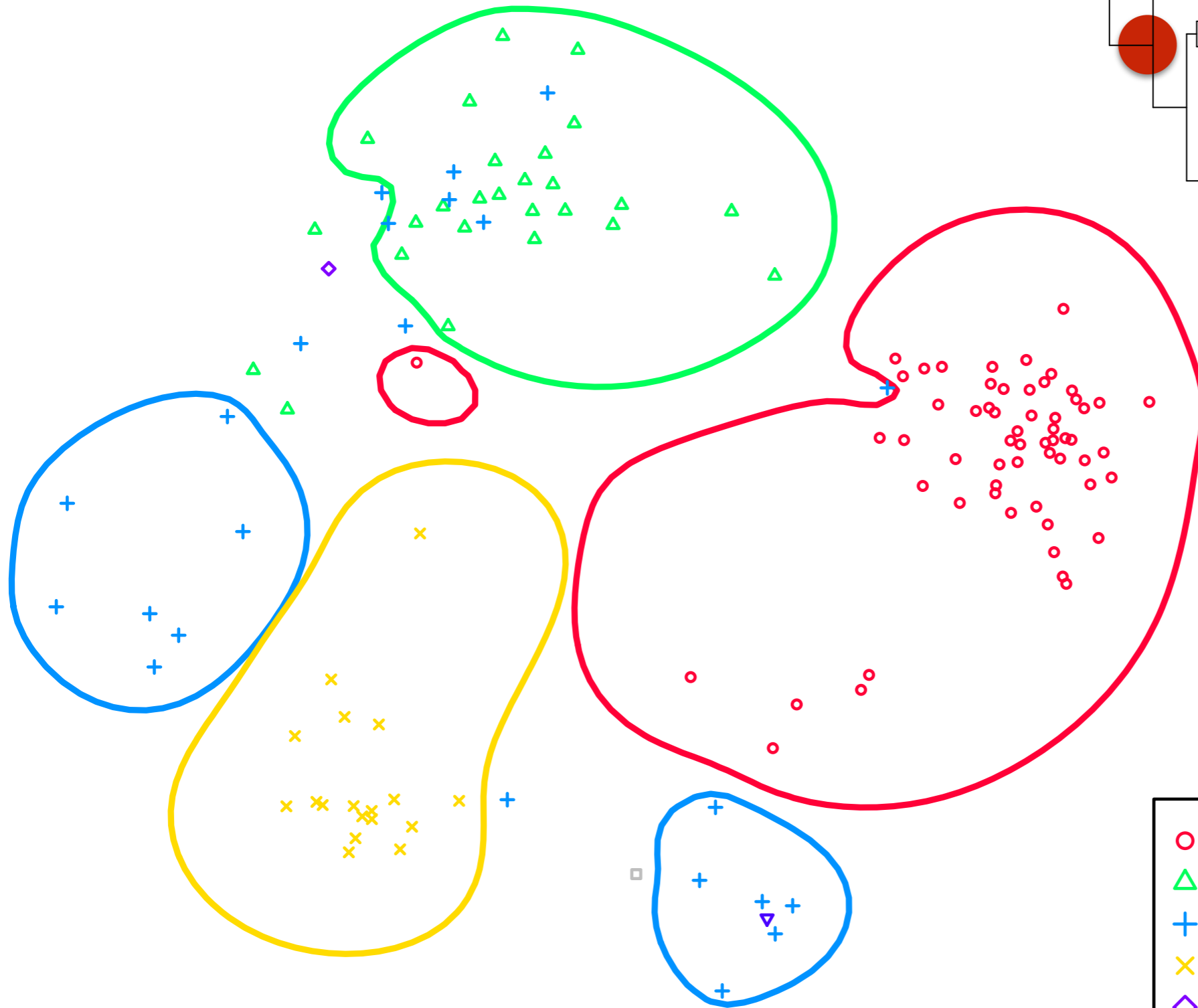
Levels drawn at 50%



Levels drawn at 50%

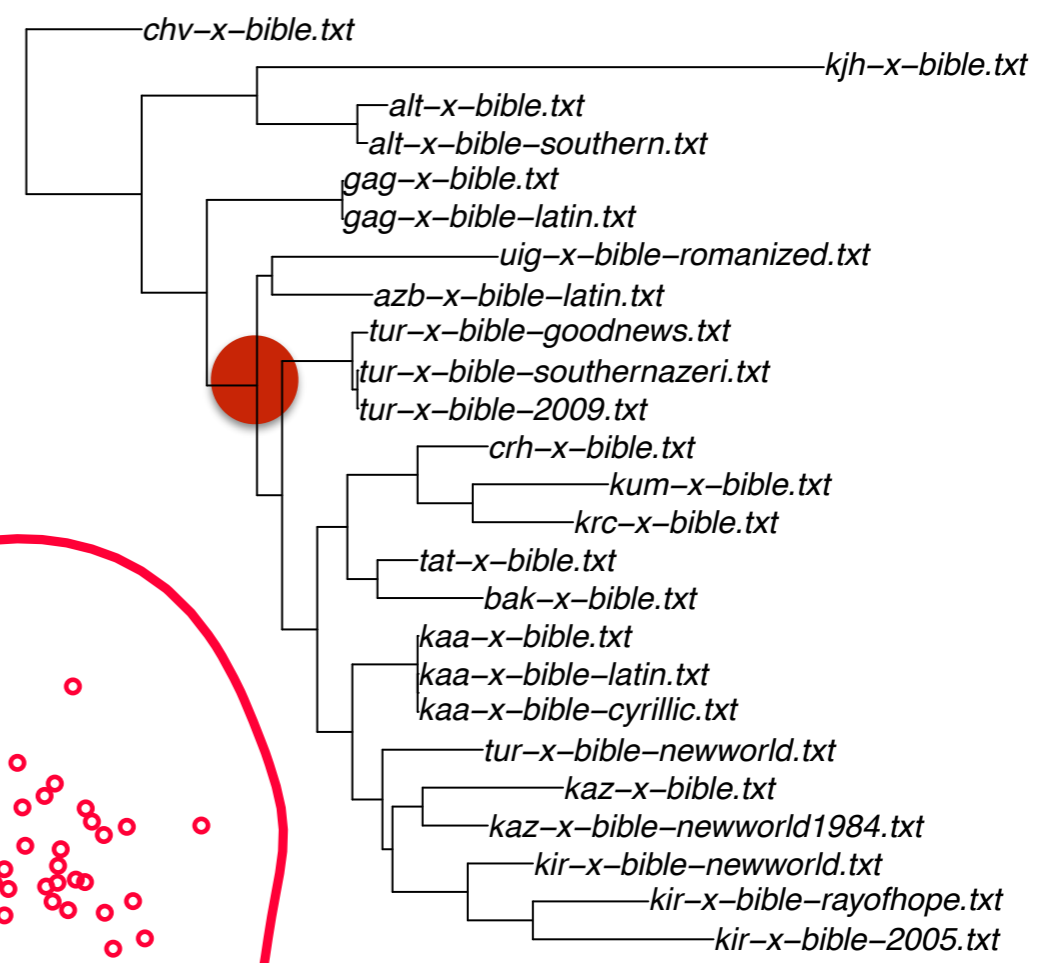
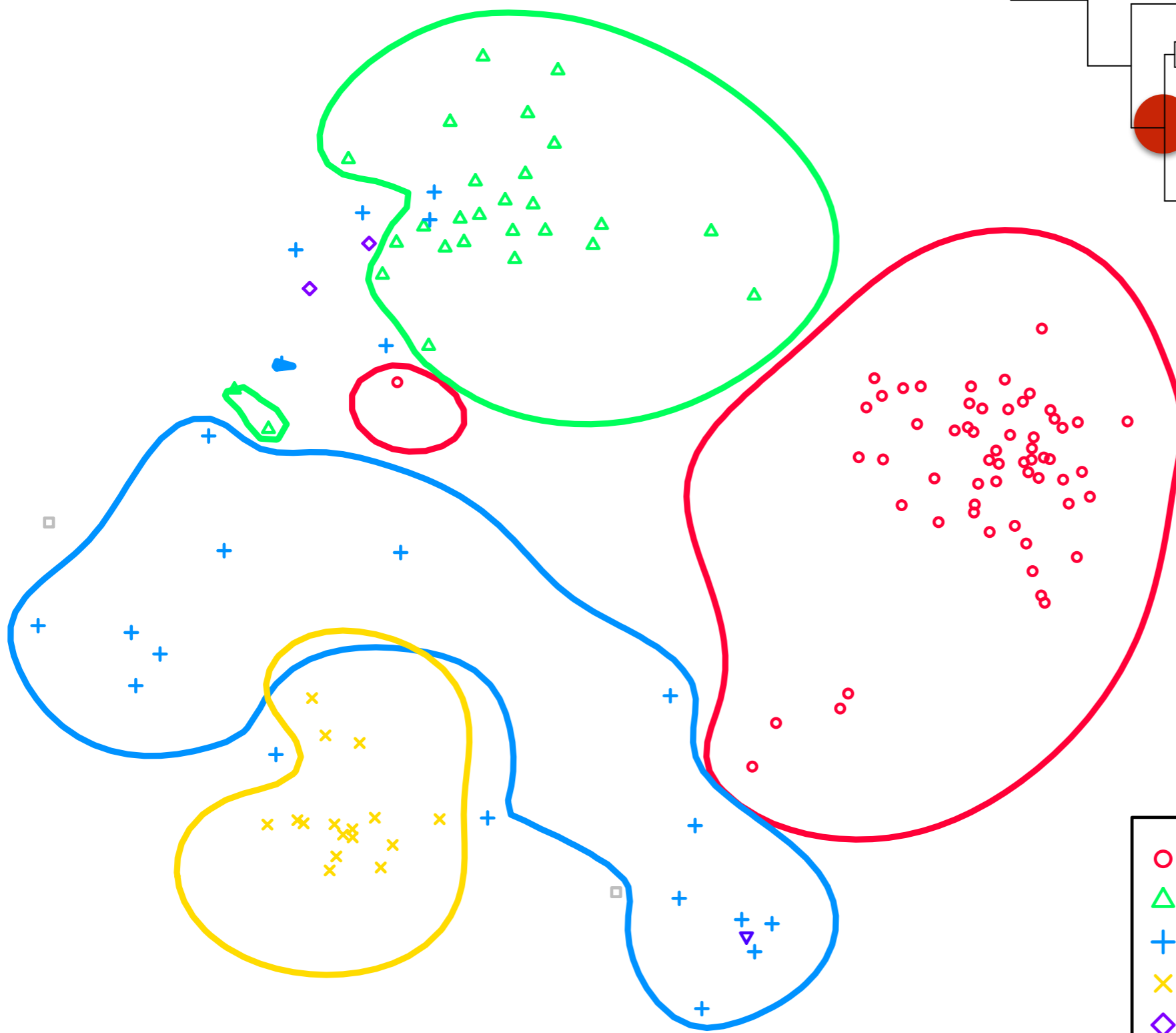


Levels drawn at 50%



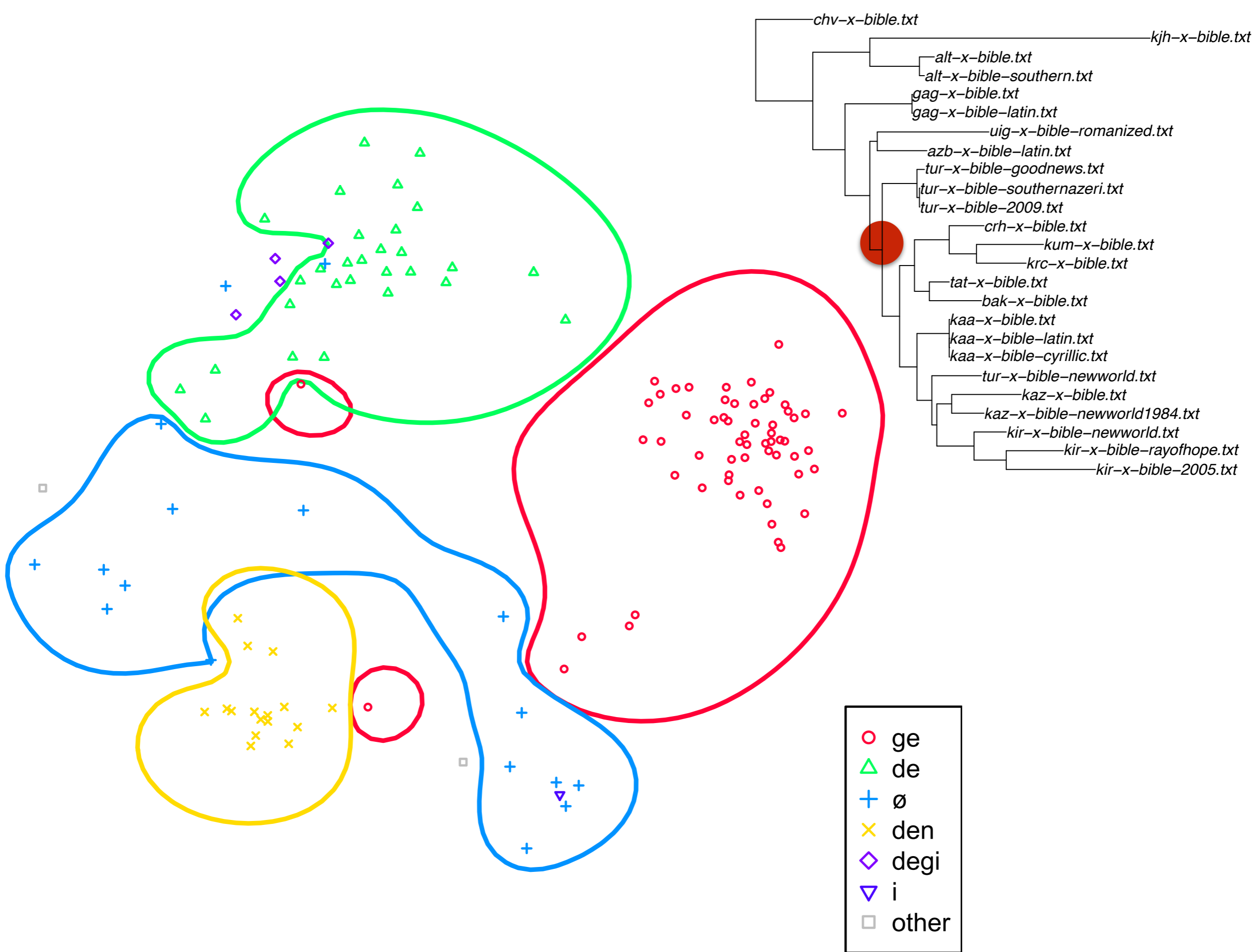
- ge
- △ de
- + ∅
- × den
- ◇ degi
- ▽ i
- other

Levels drawn at 50%

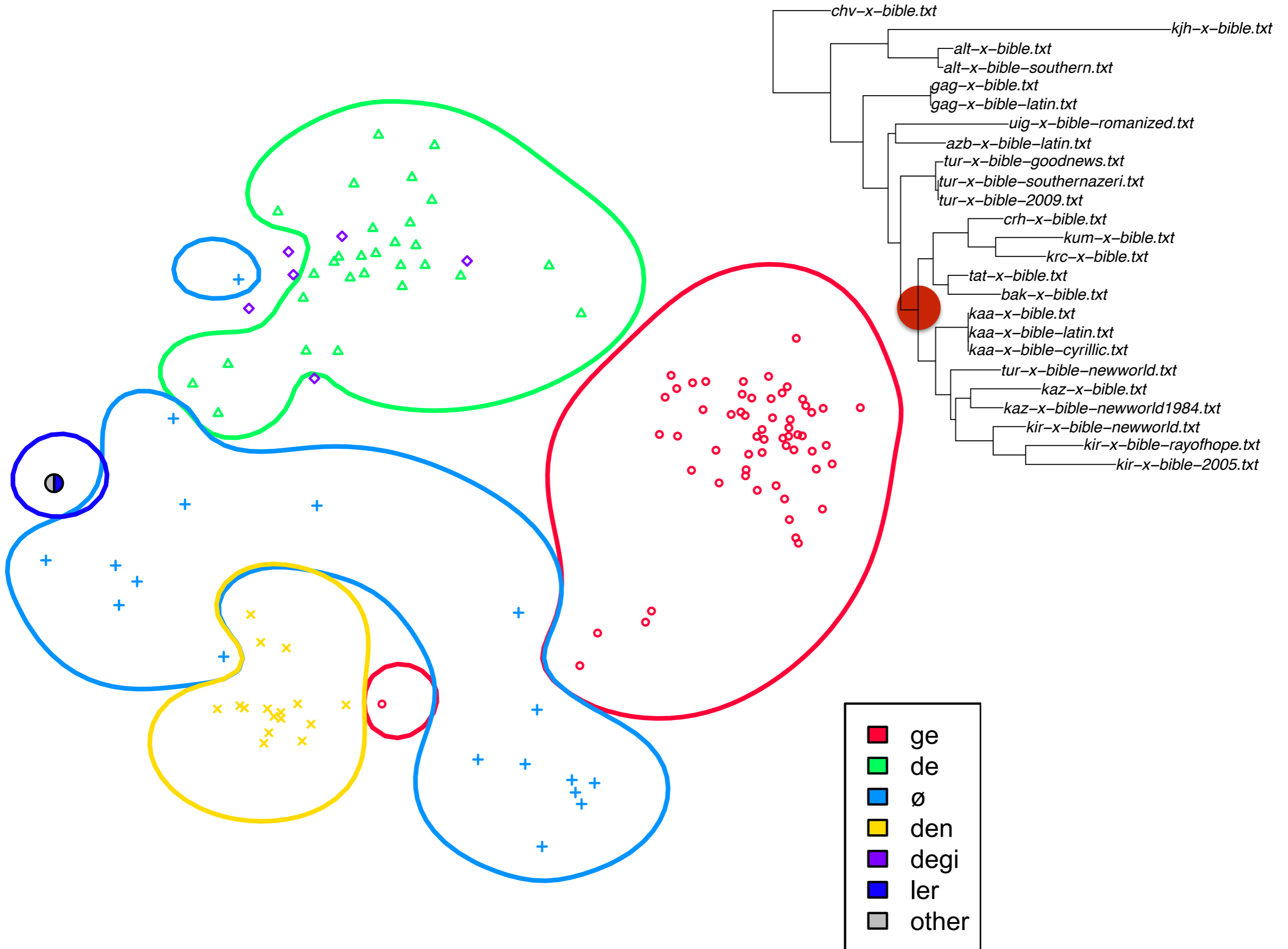


- ge
- △ de
- + ø
- × den
- ◇ degi
- ▽ i
- other

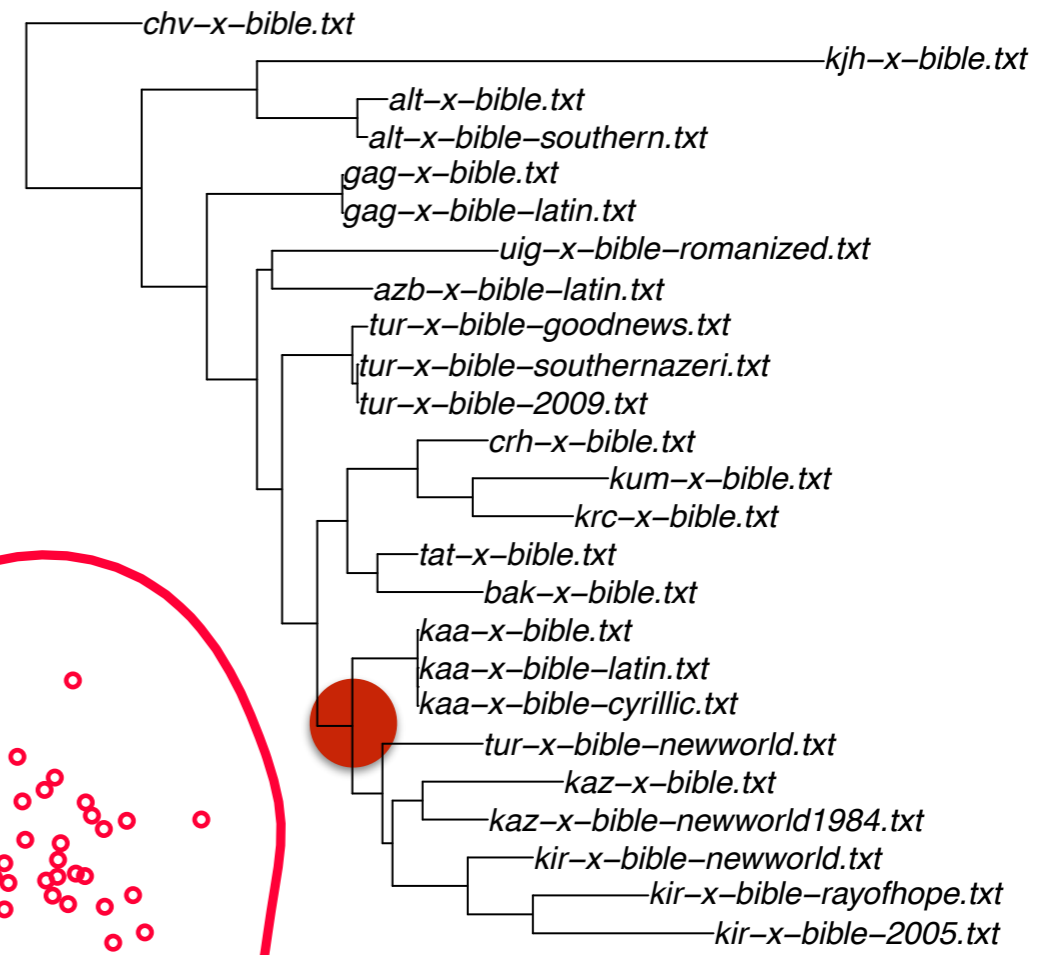
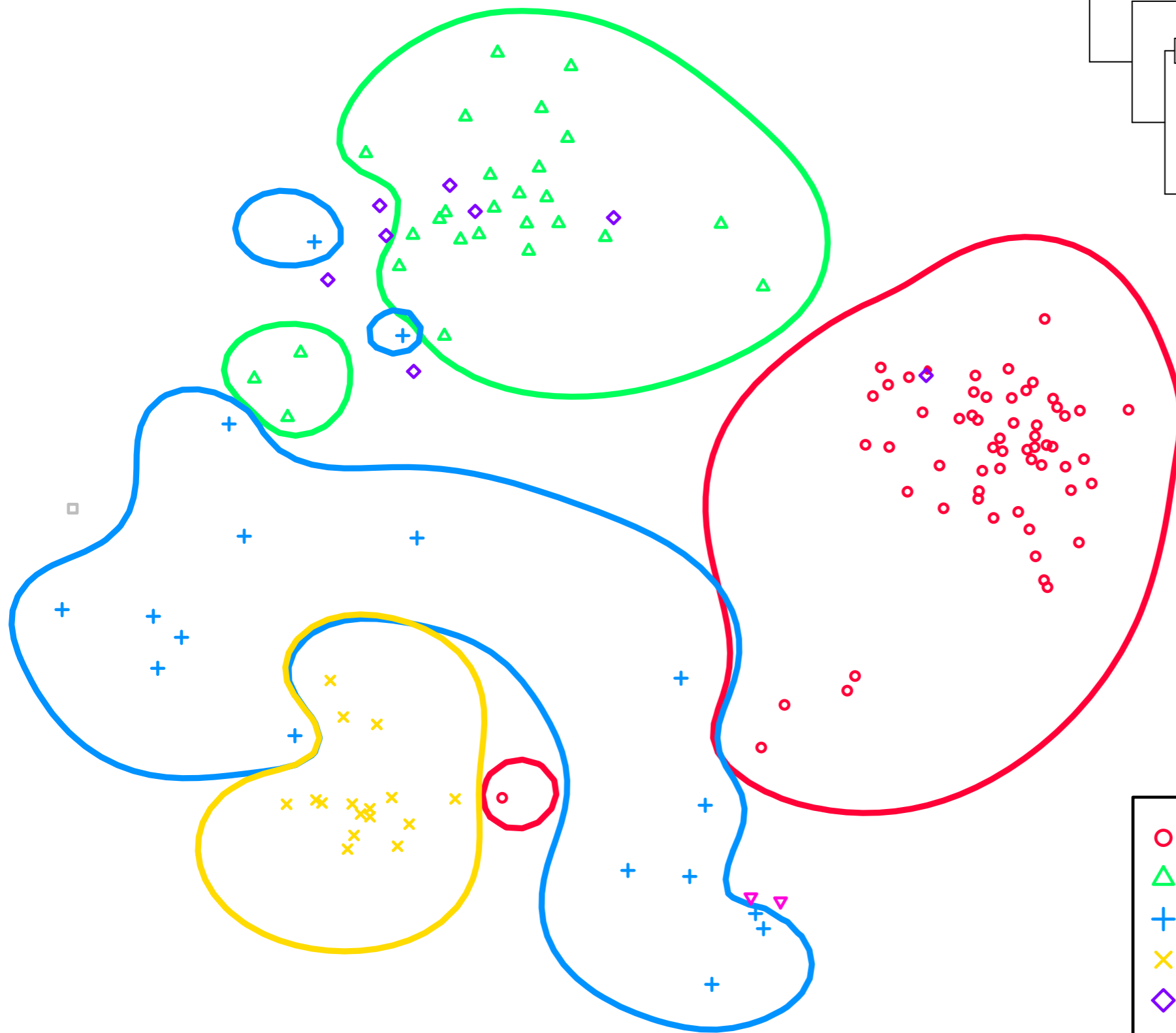
Levels drawn at 50%



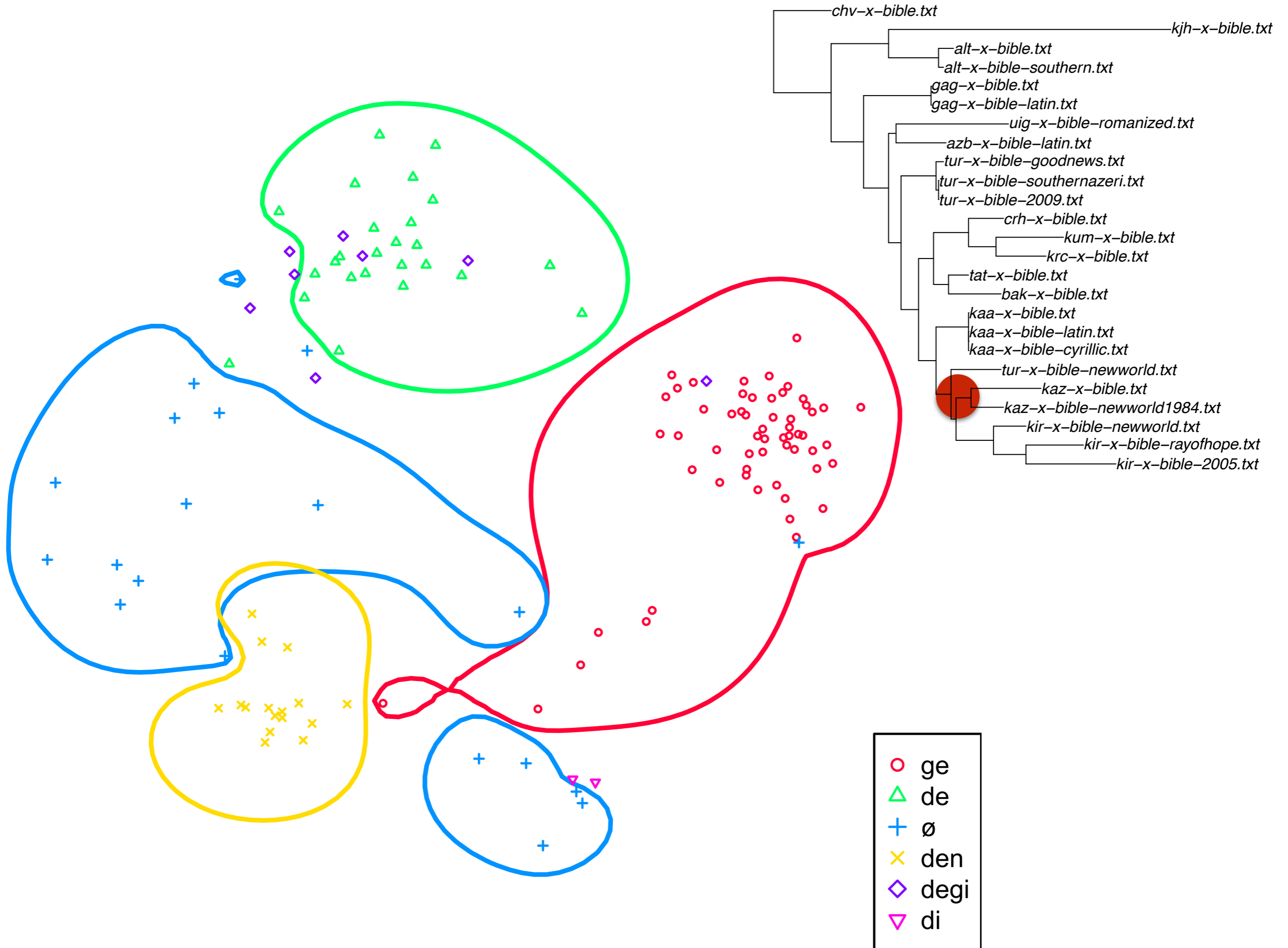
Levels drawn at 50%



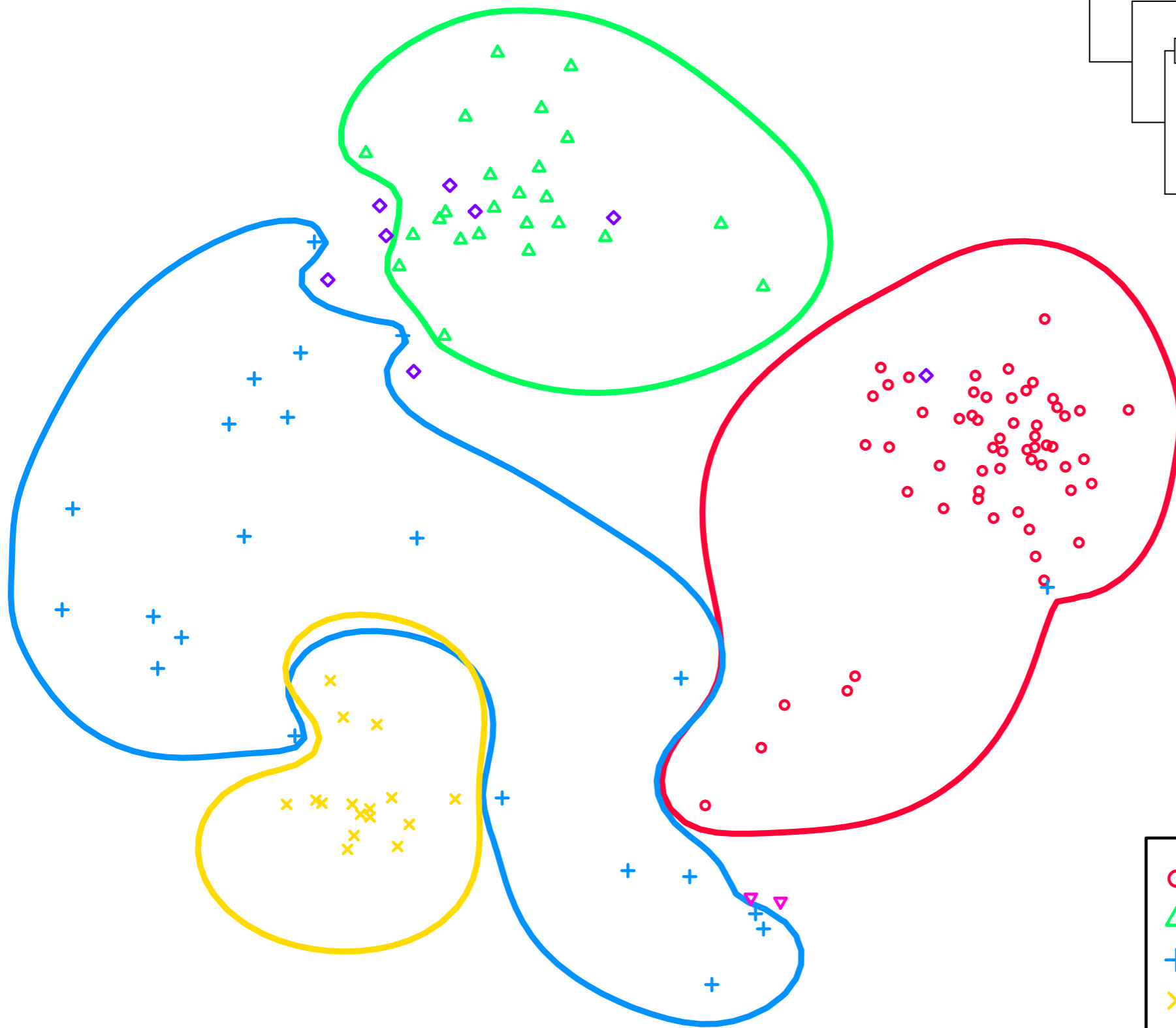
Levels drawn at 50%



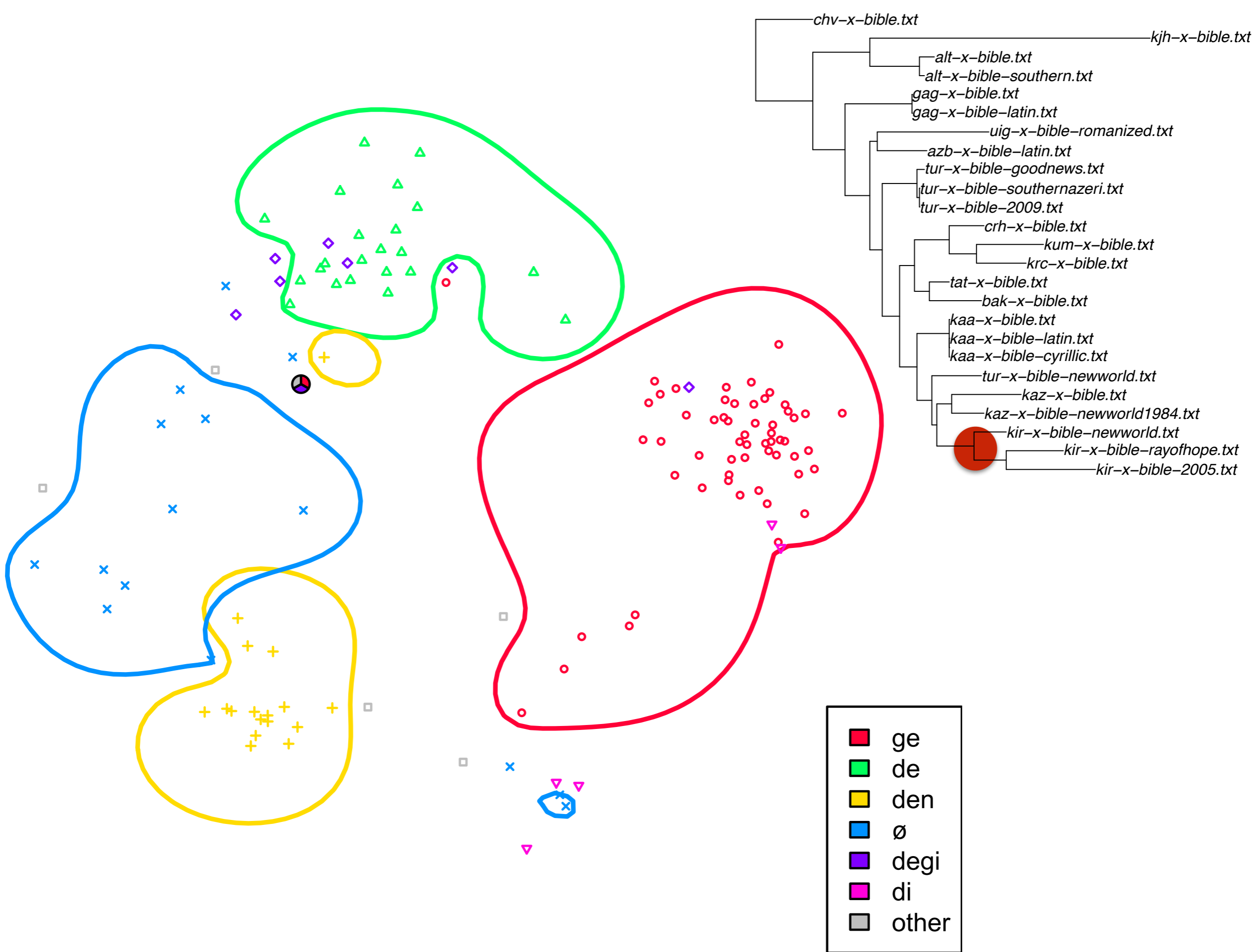
Levels drawn at 50%



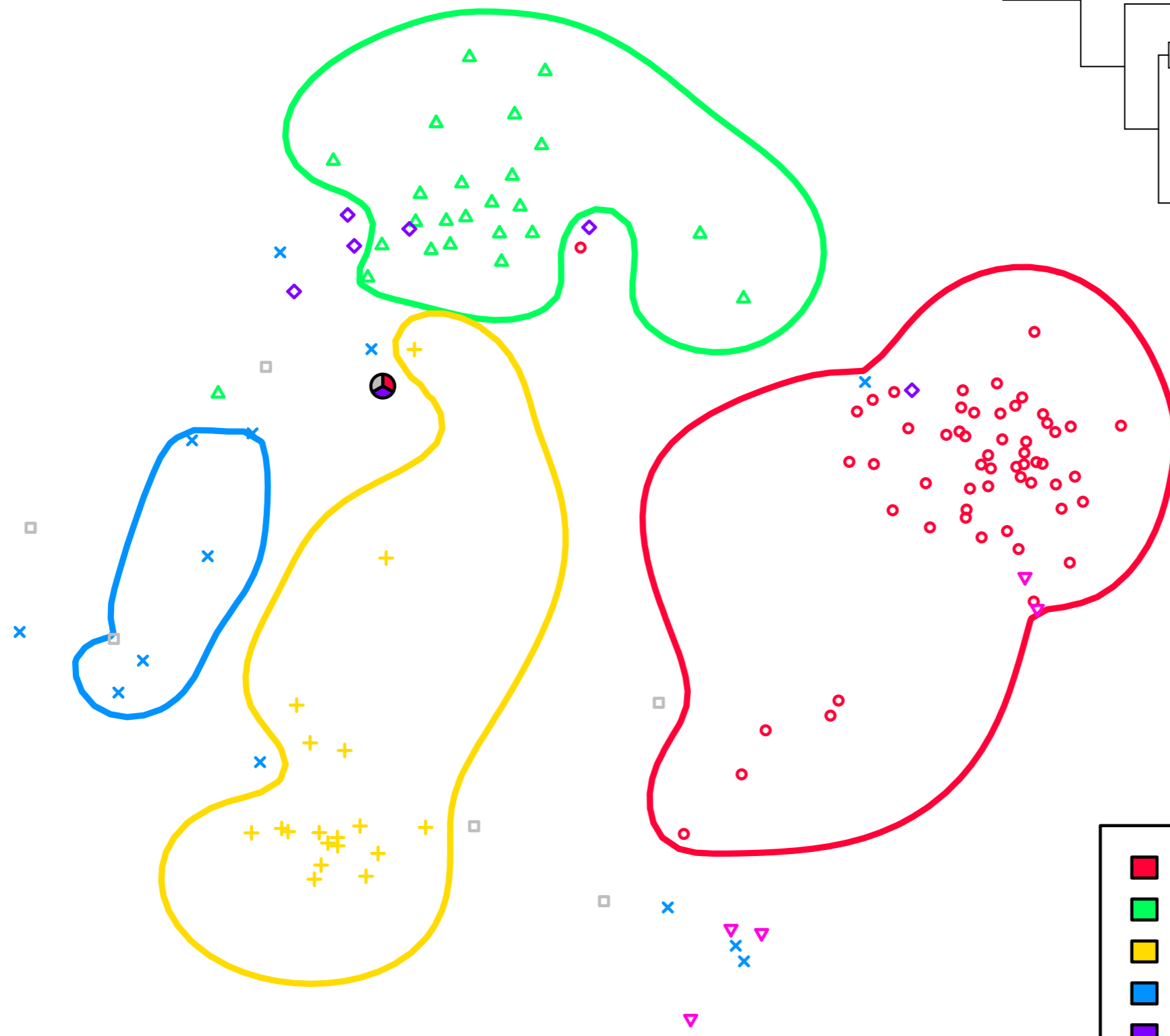
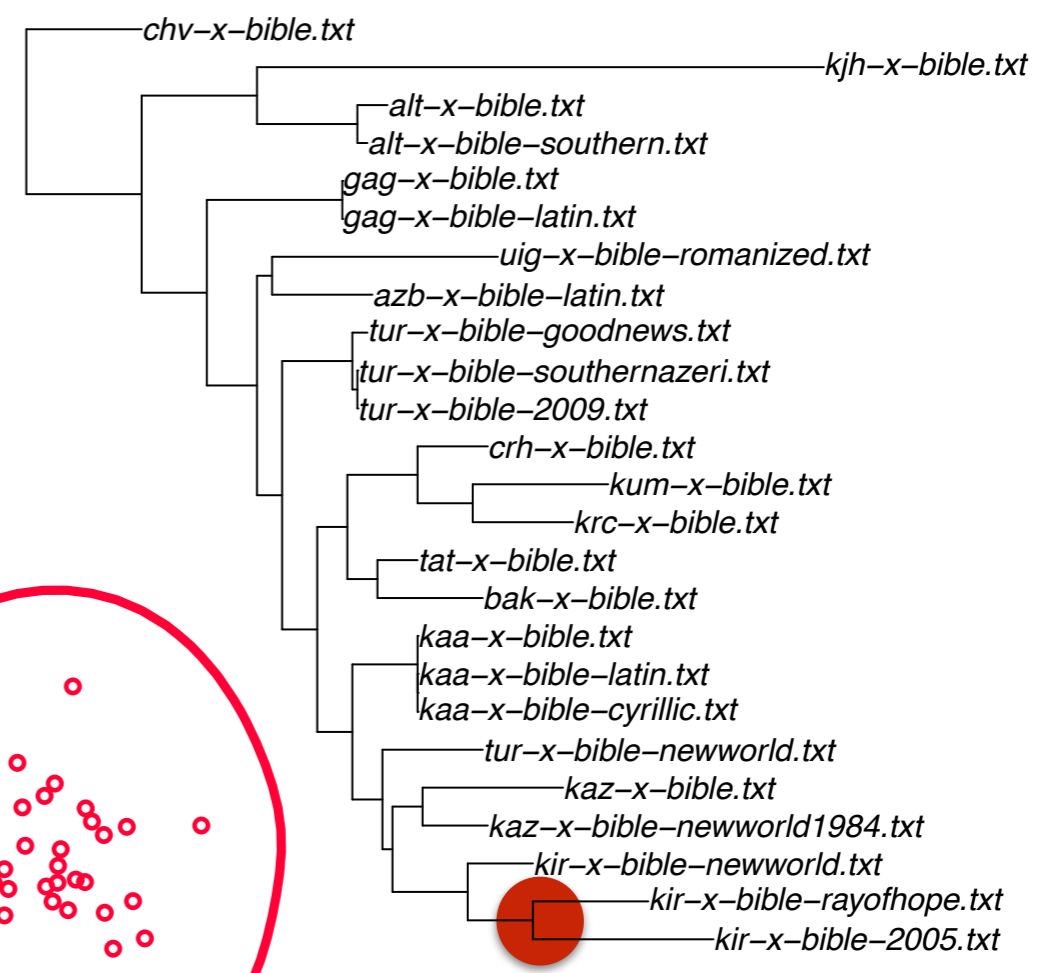
Levels drawn at 50%



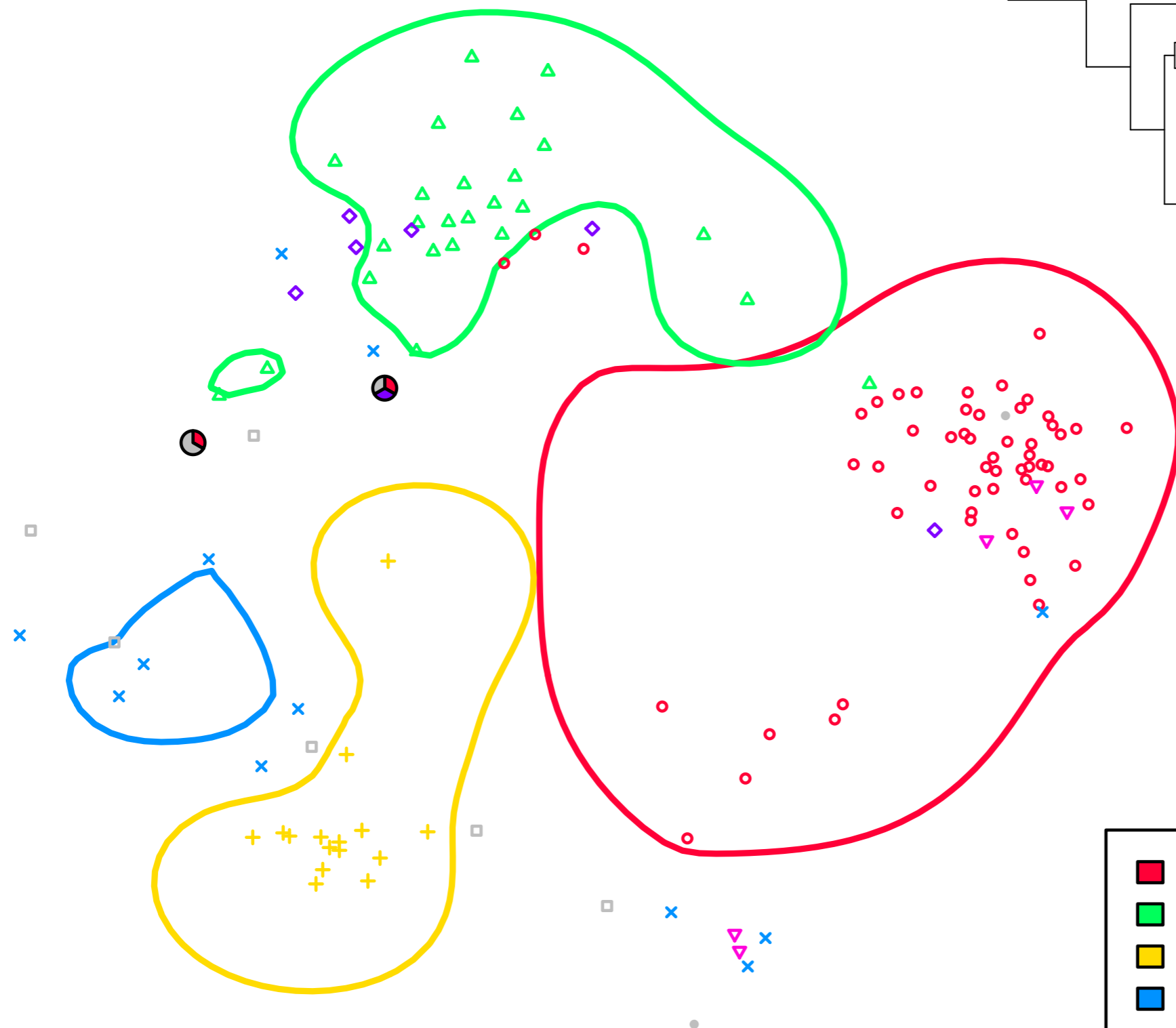
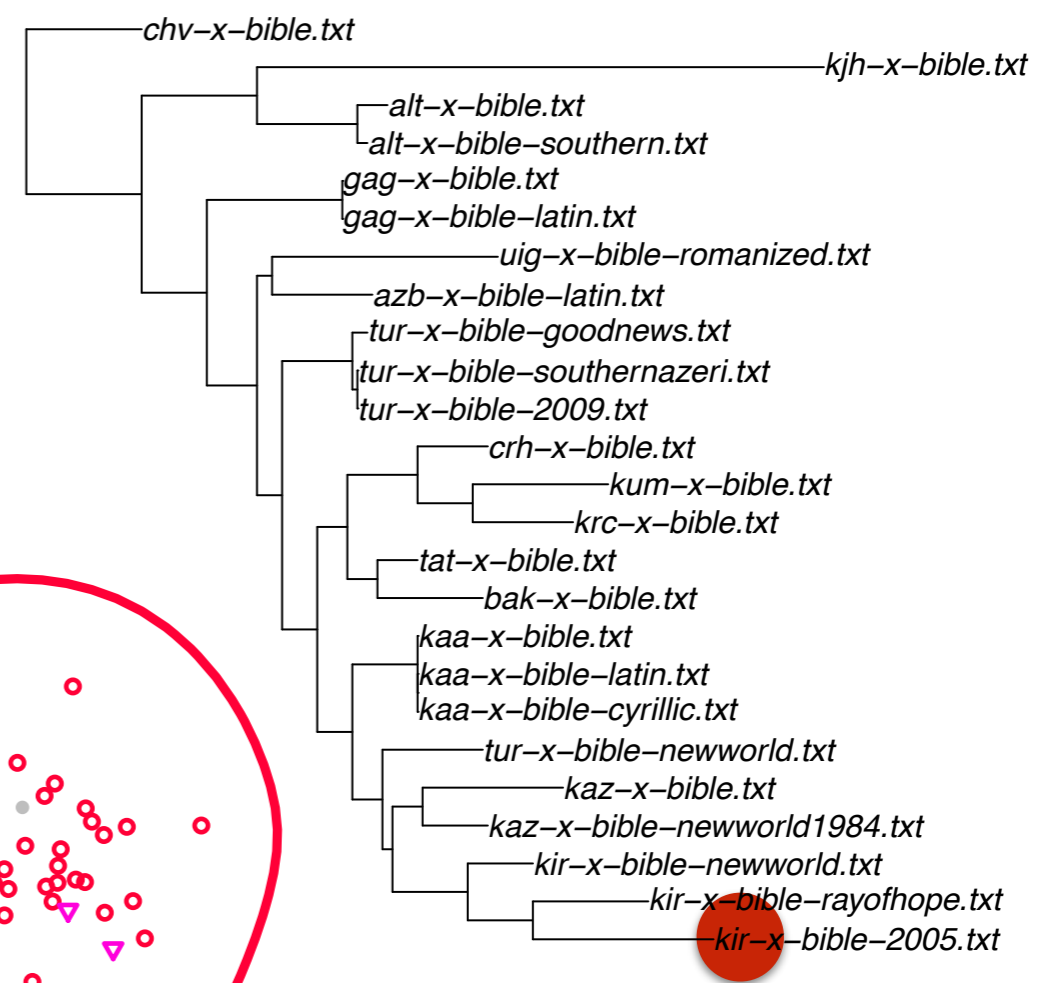
Levels drawn at 50%



Levels drawn at 50%

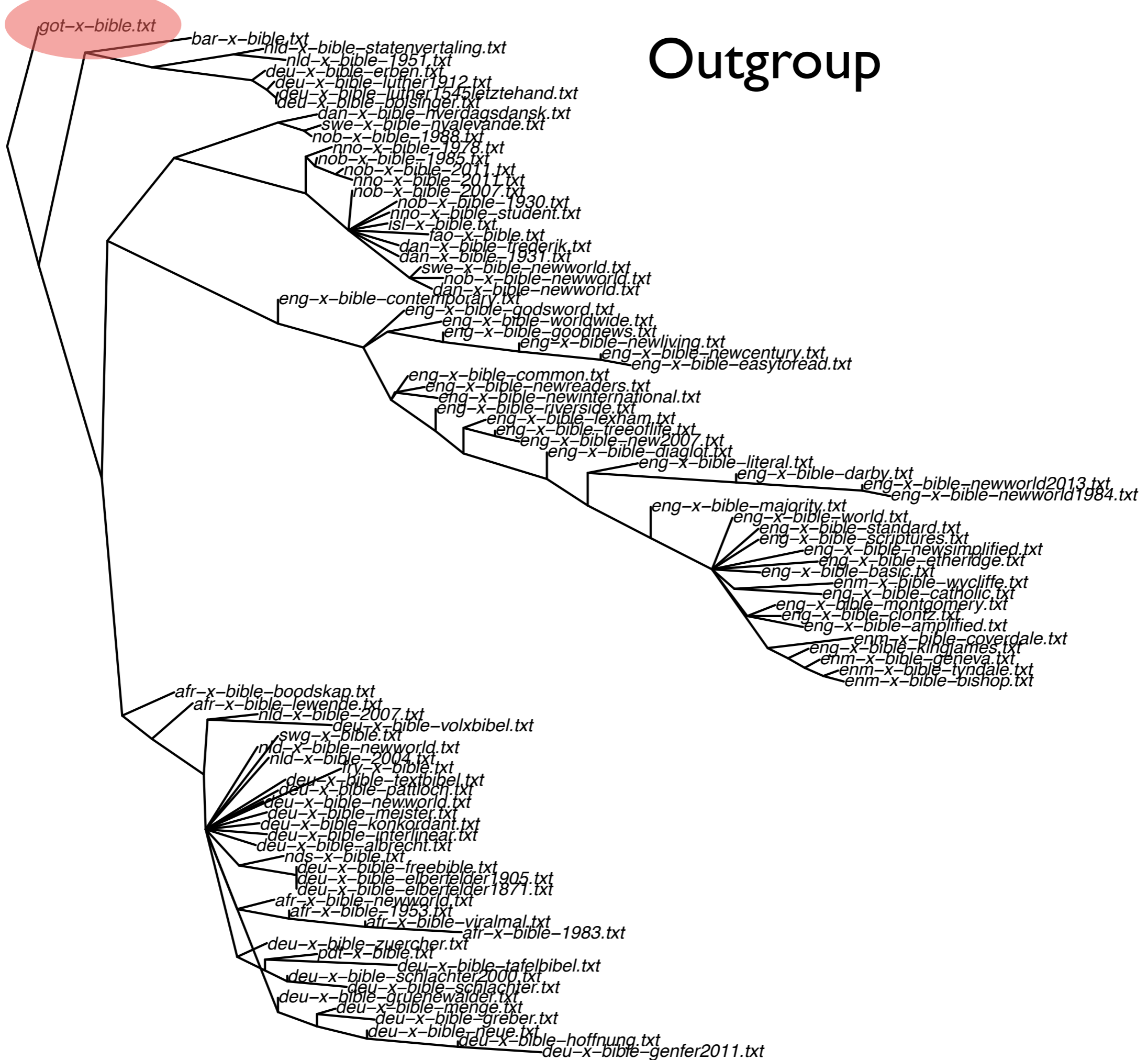


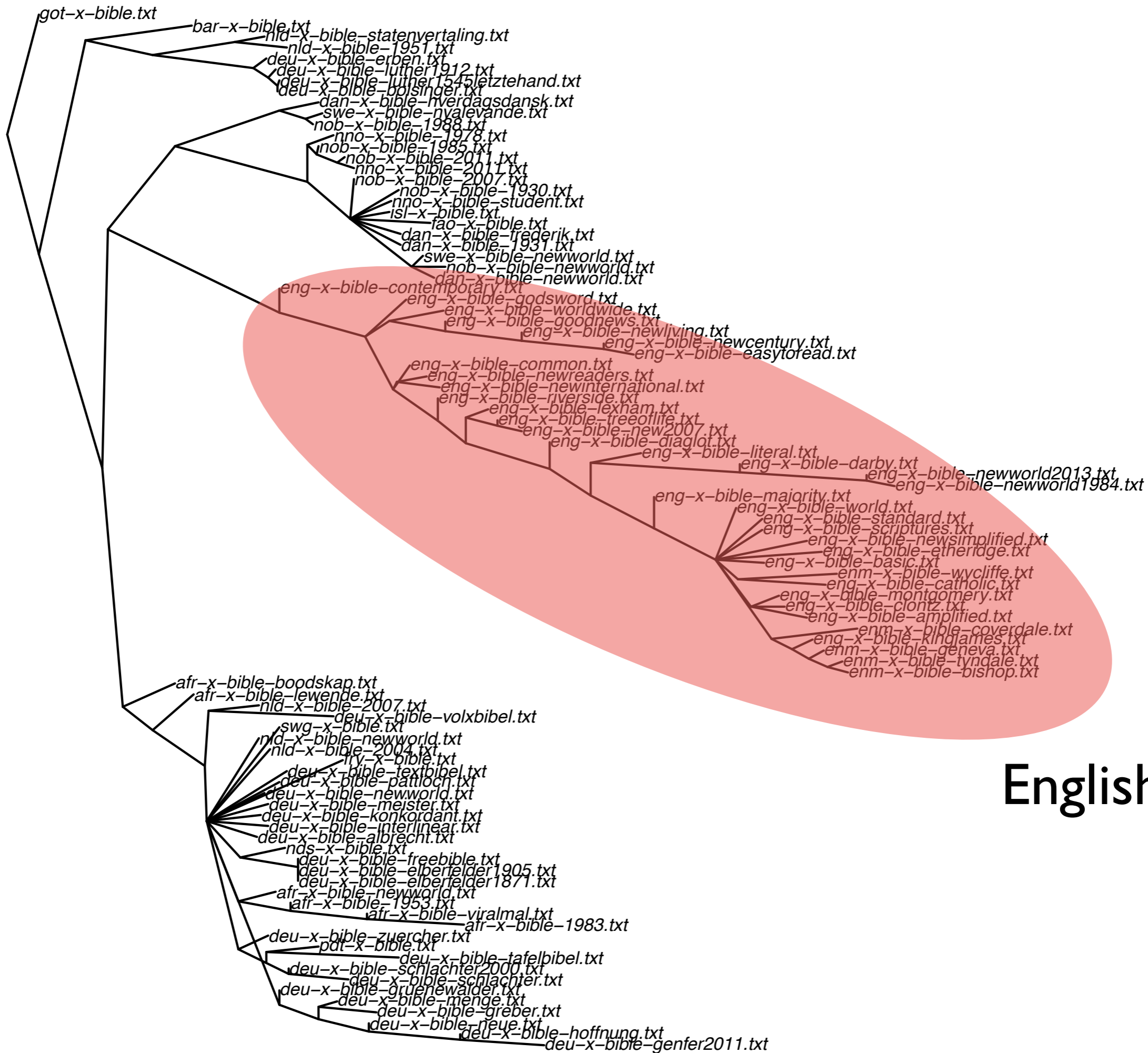
Levels drawn at 50%



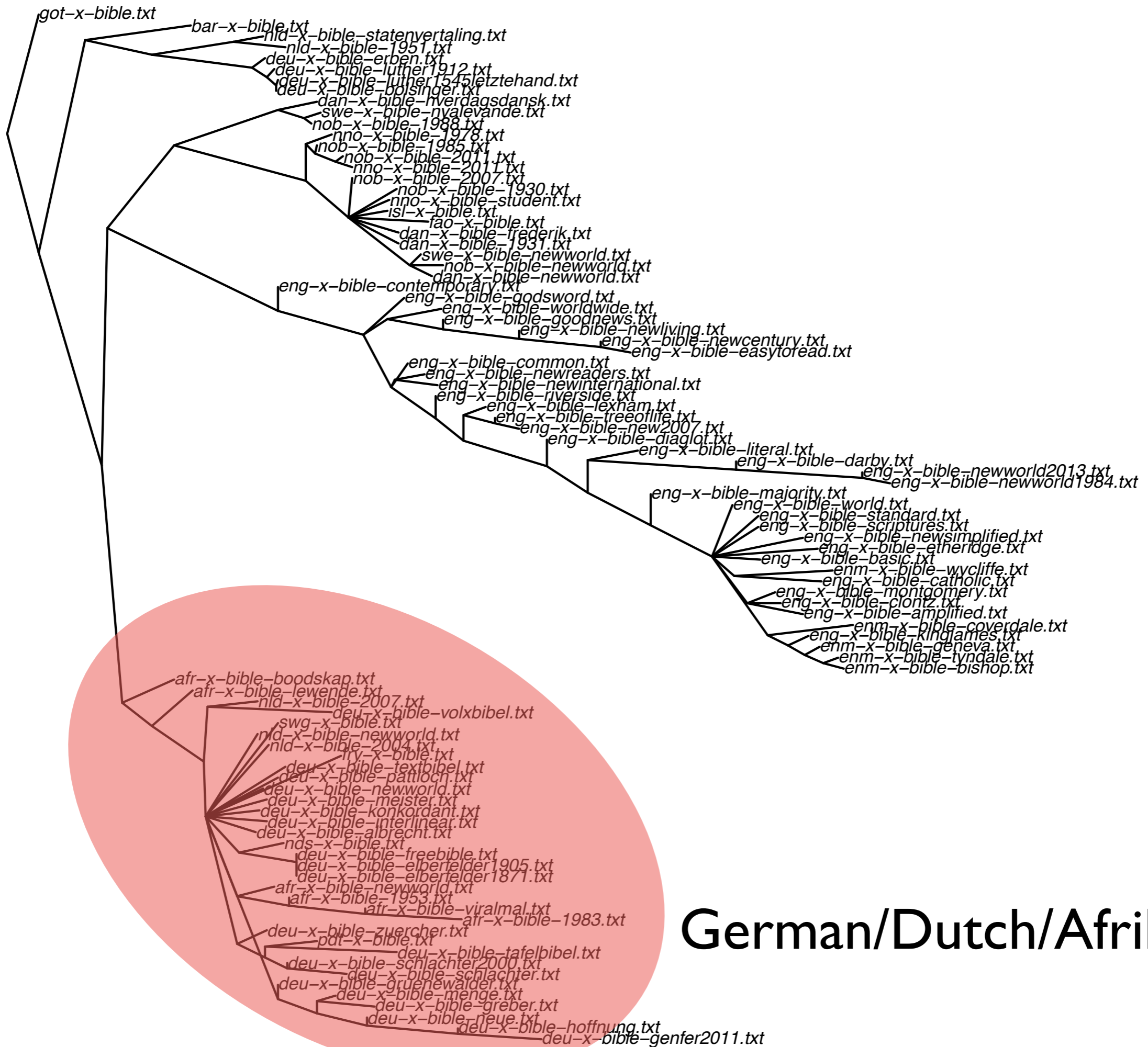
Levels drawn at 50%

Outgroup



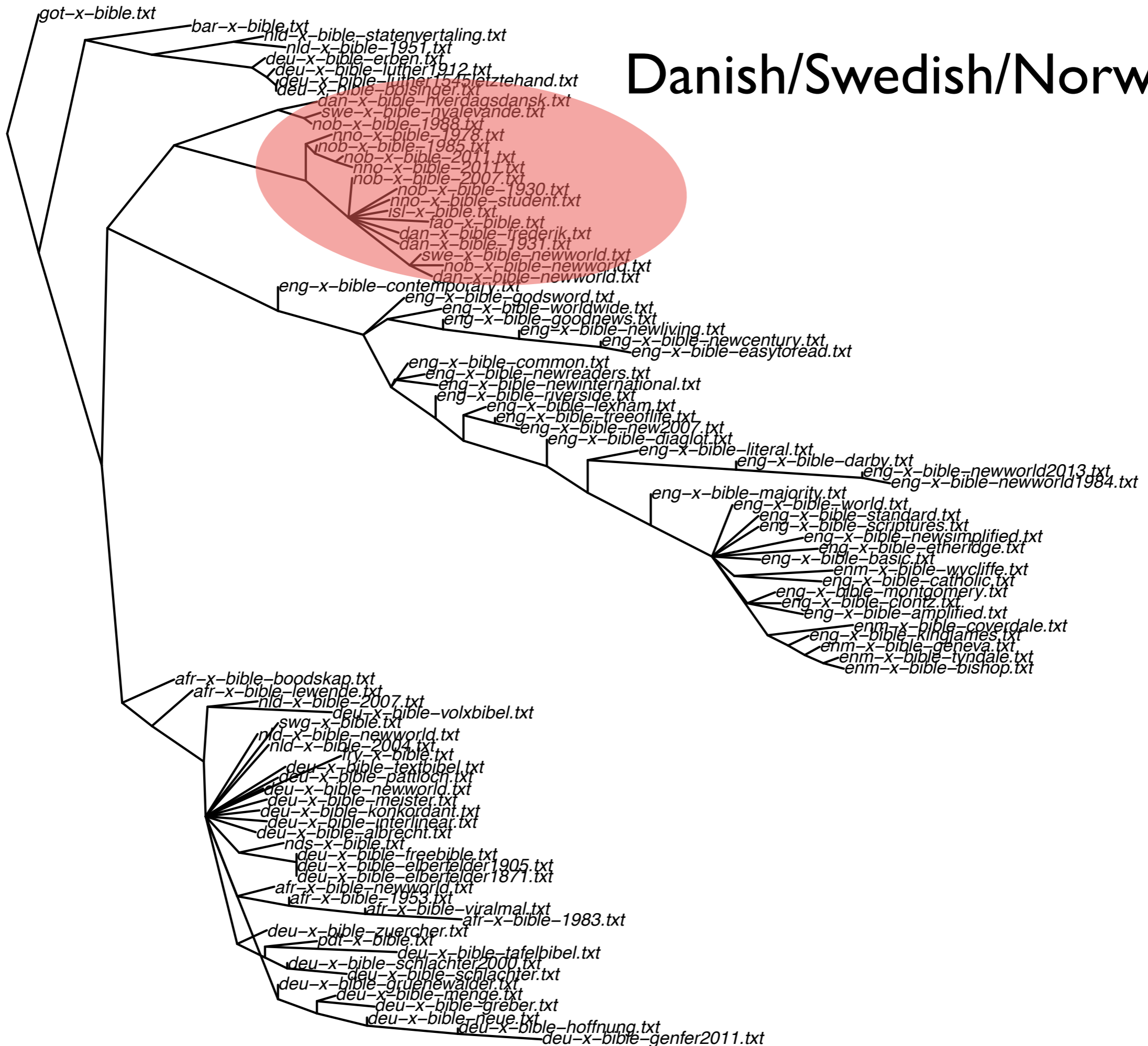


English



German/Dutch/Afrikaans

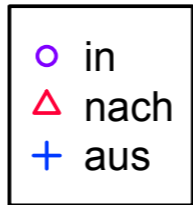
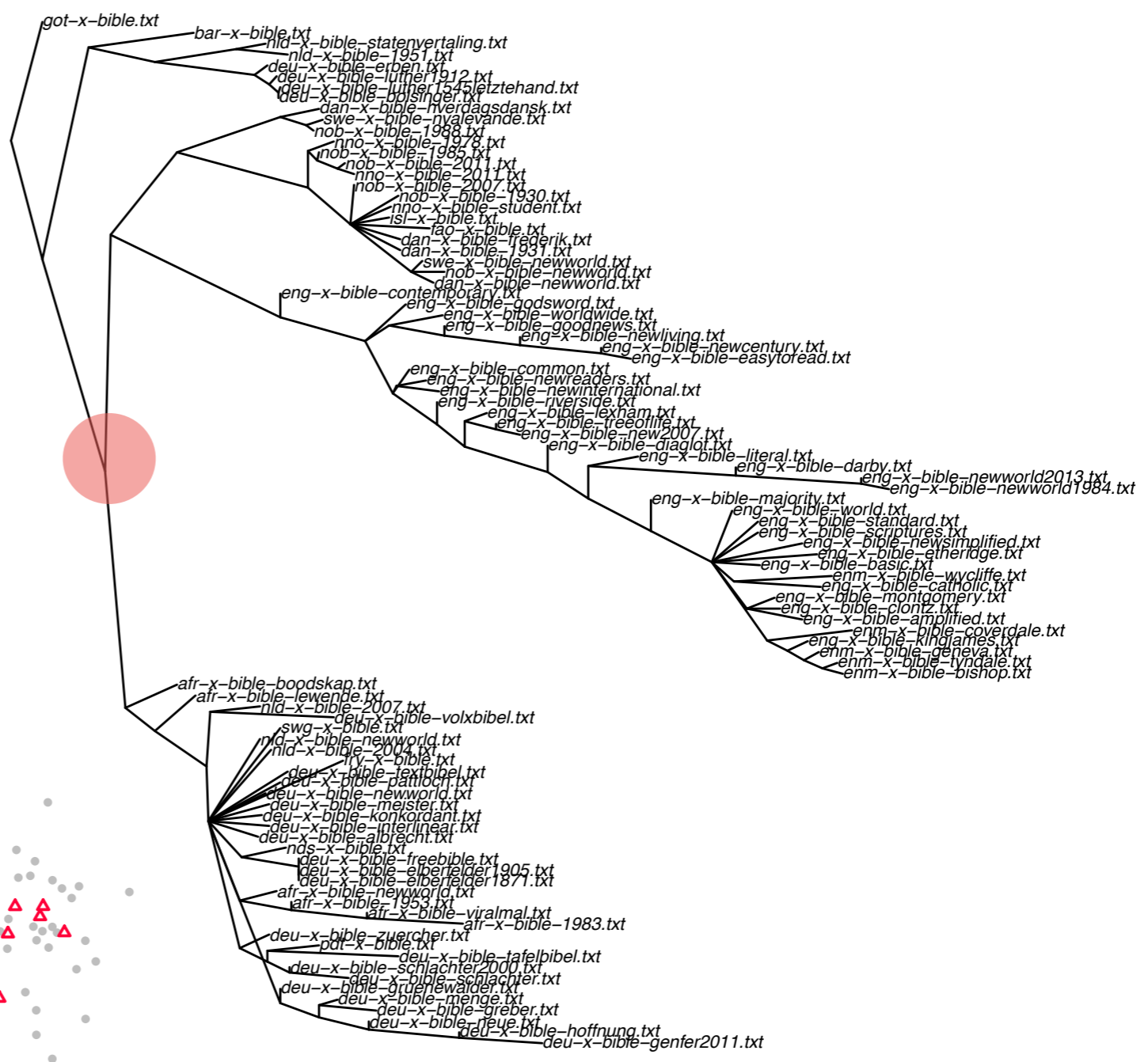
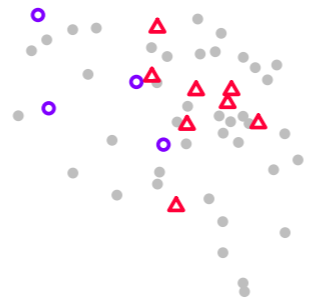
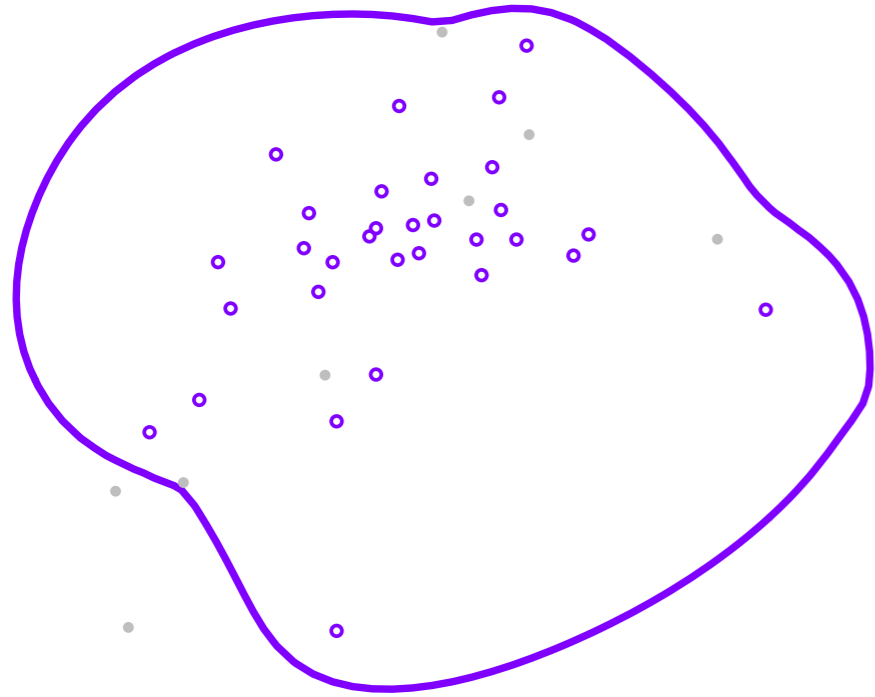
Danish/Swedish/Norwegian



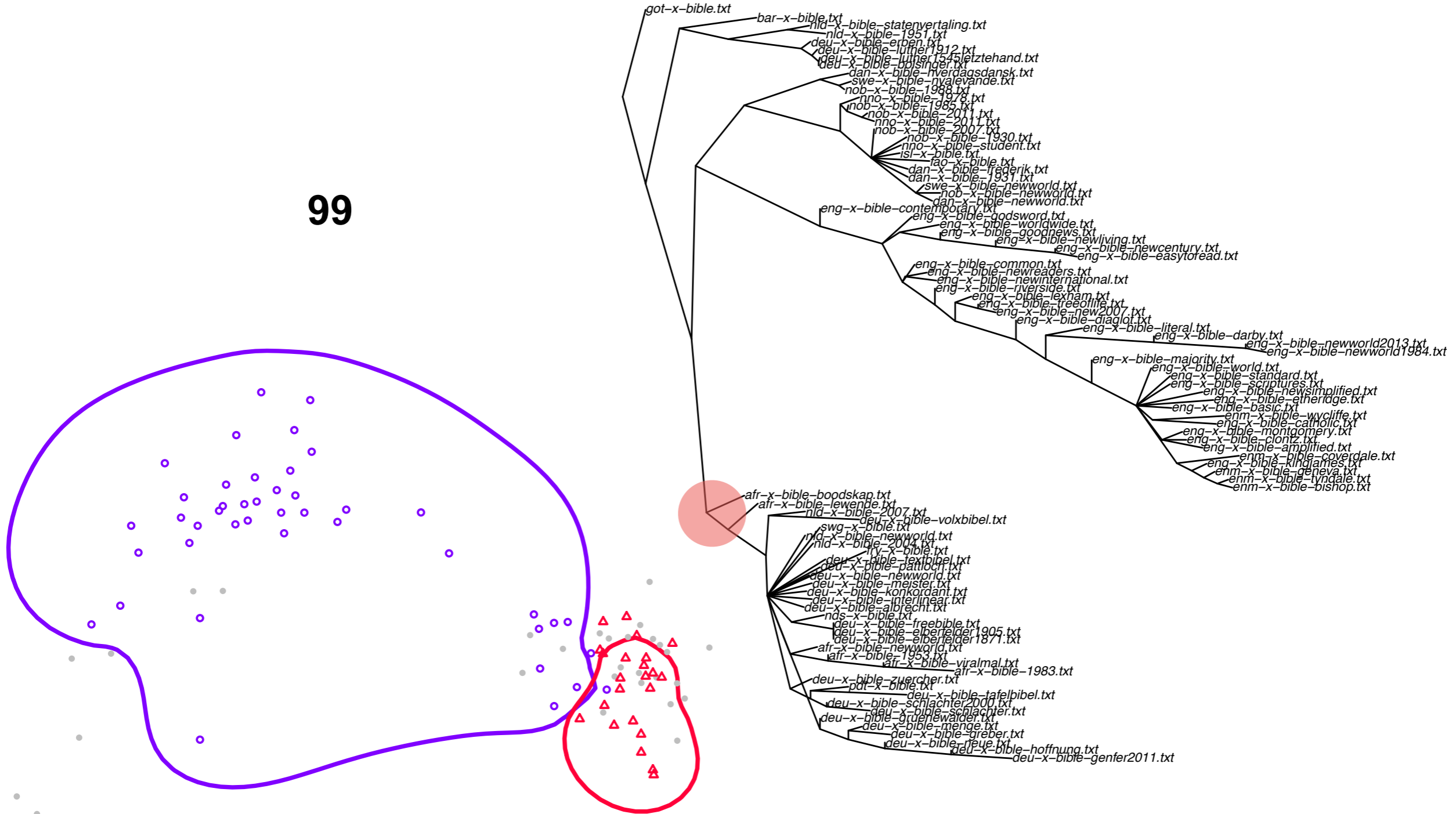


???

101

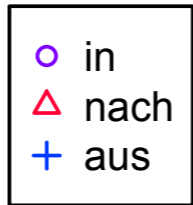
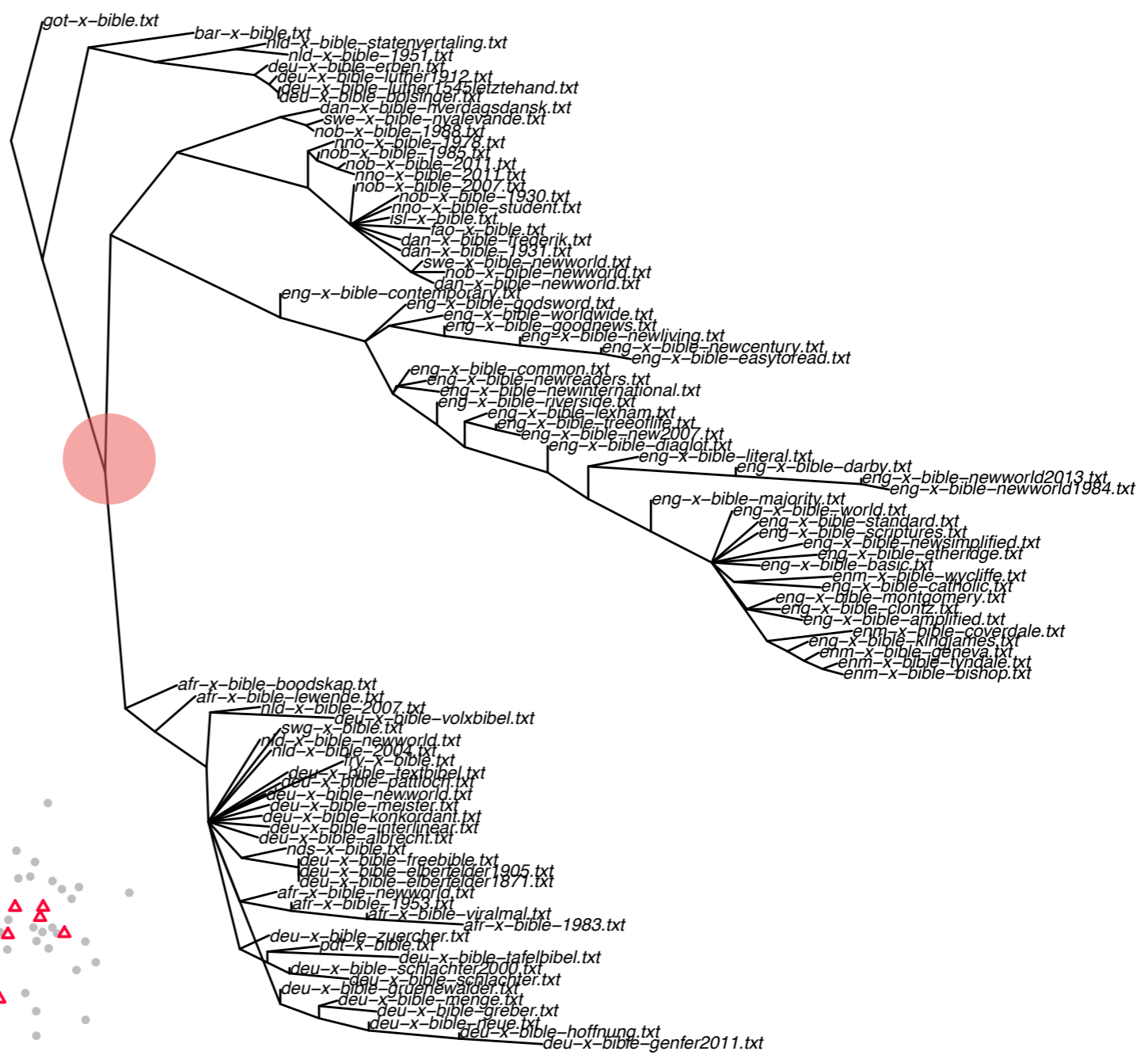
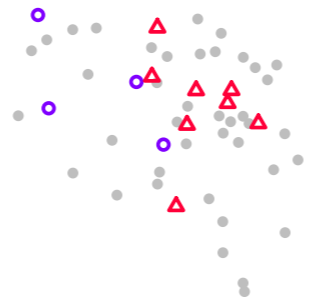
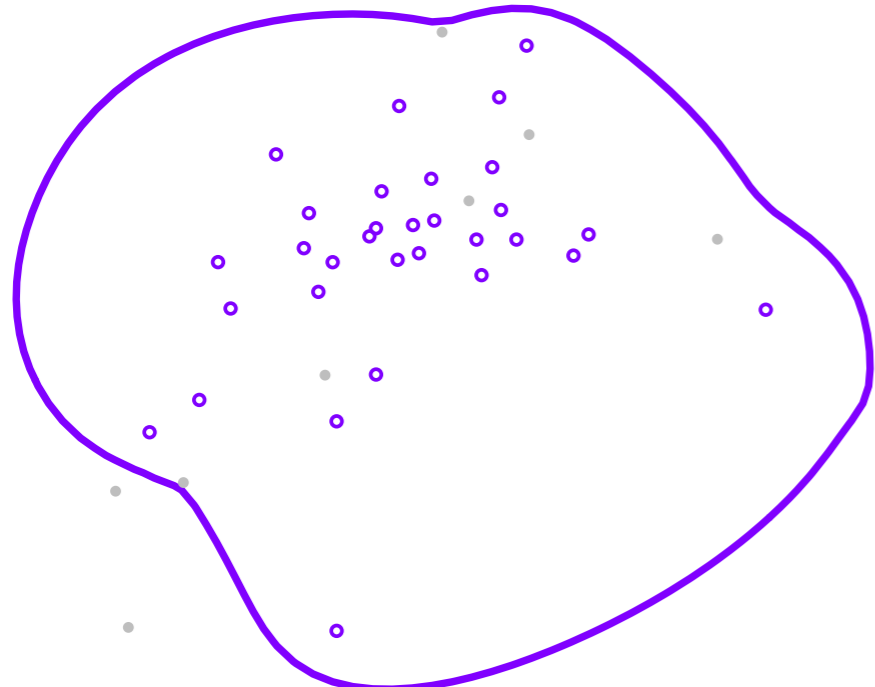


99

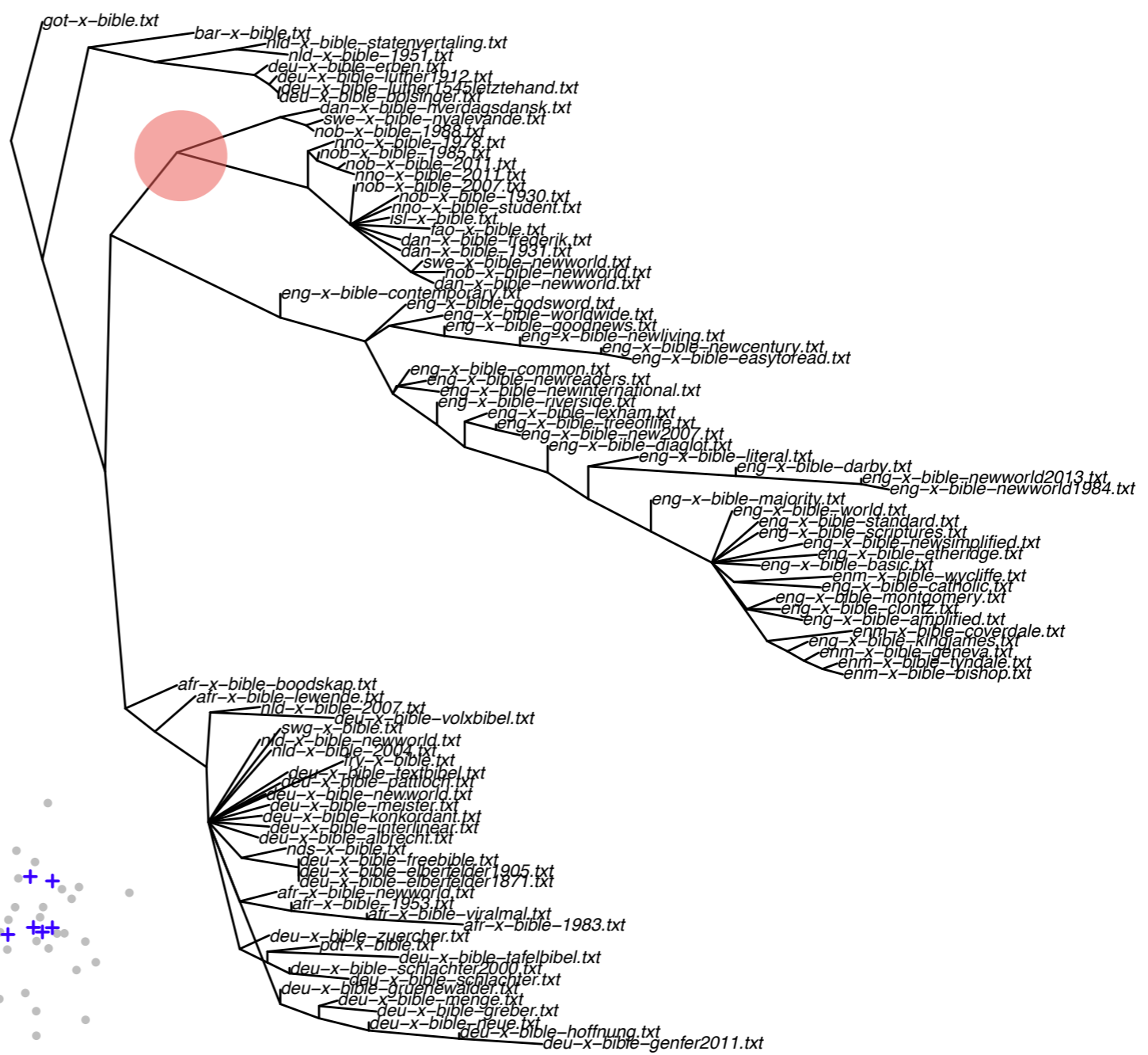
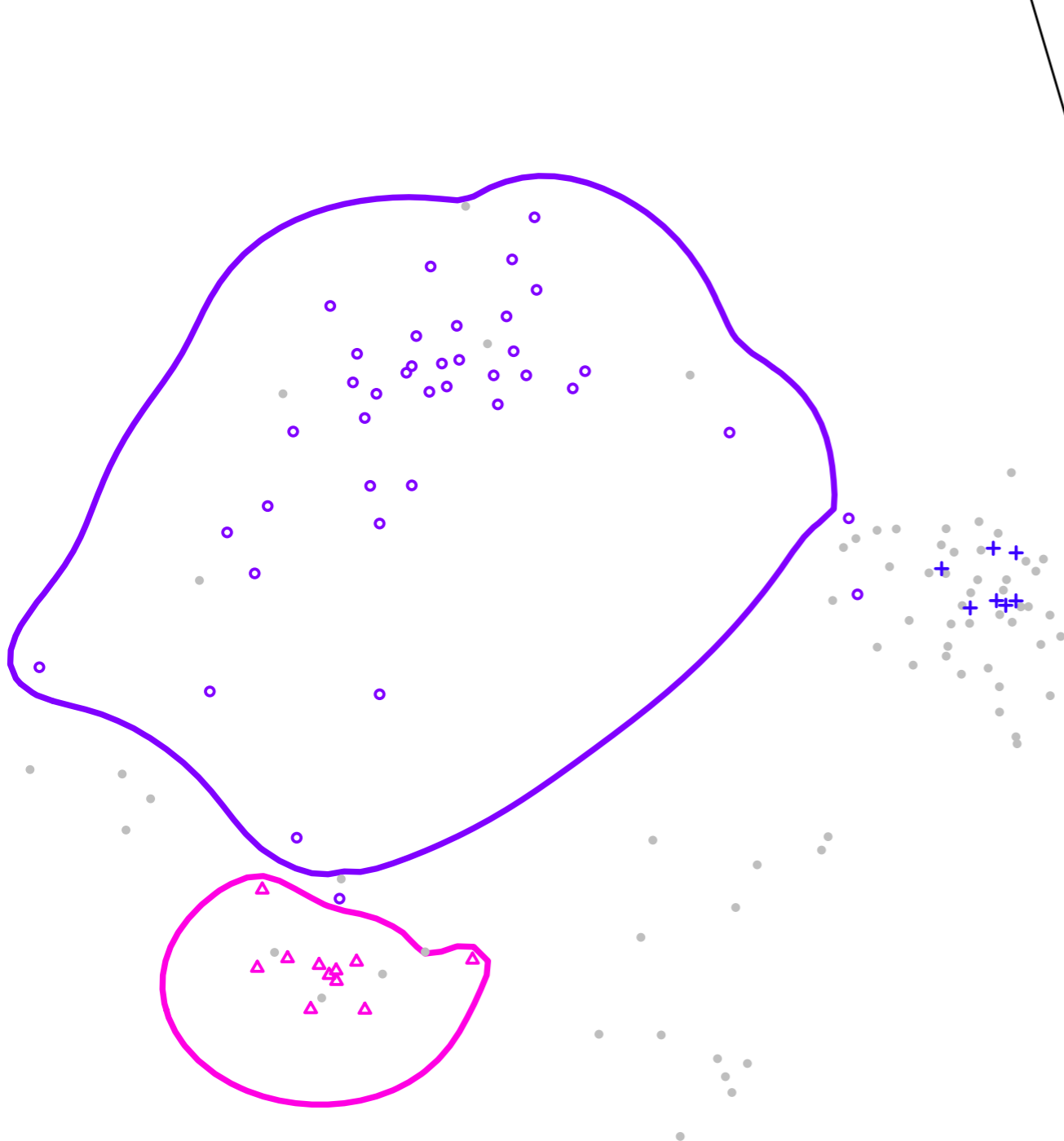


- in
- △ nach
- + aus
- × von

101

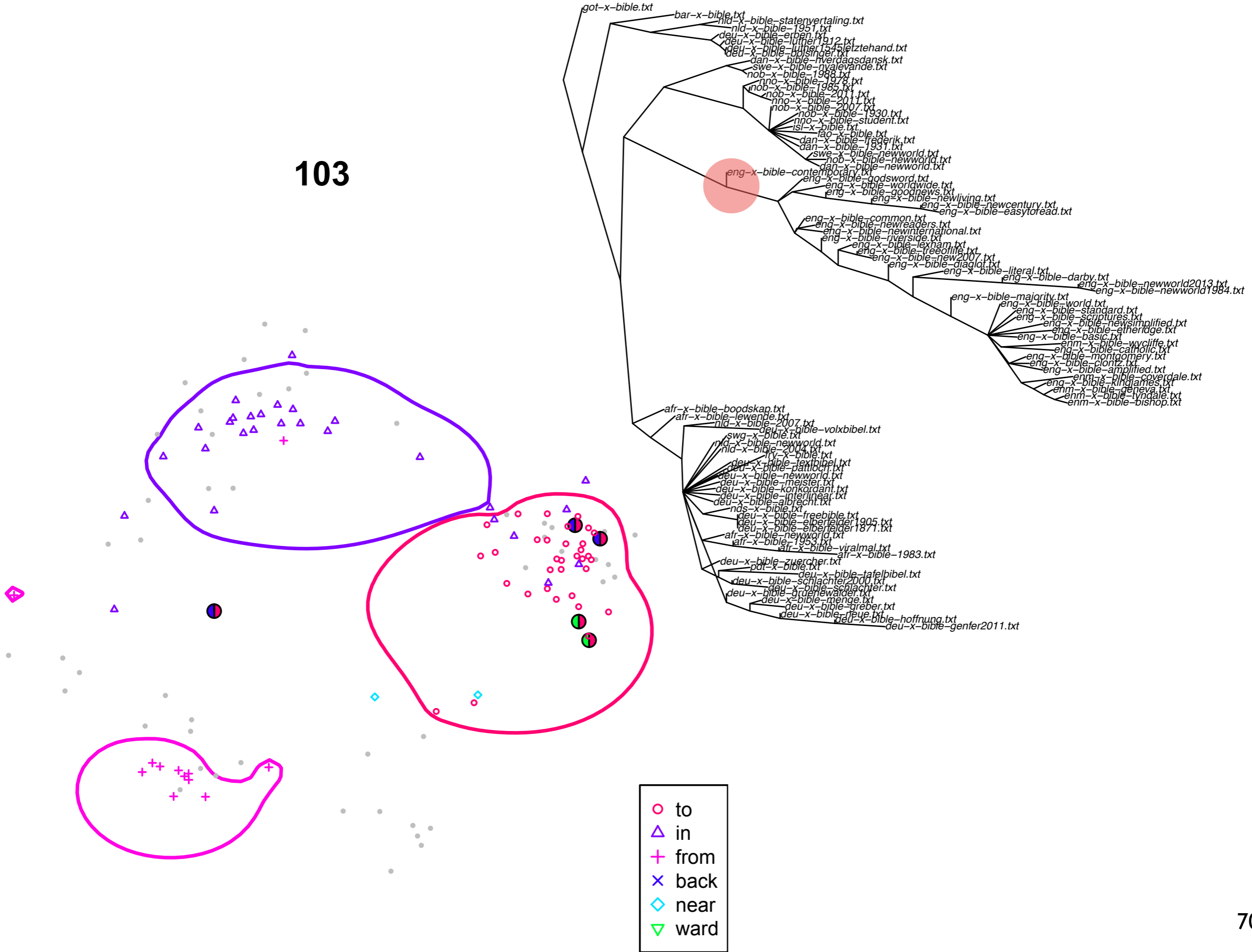


128



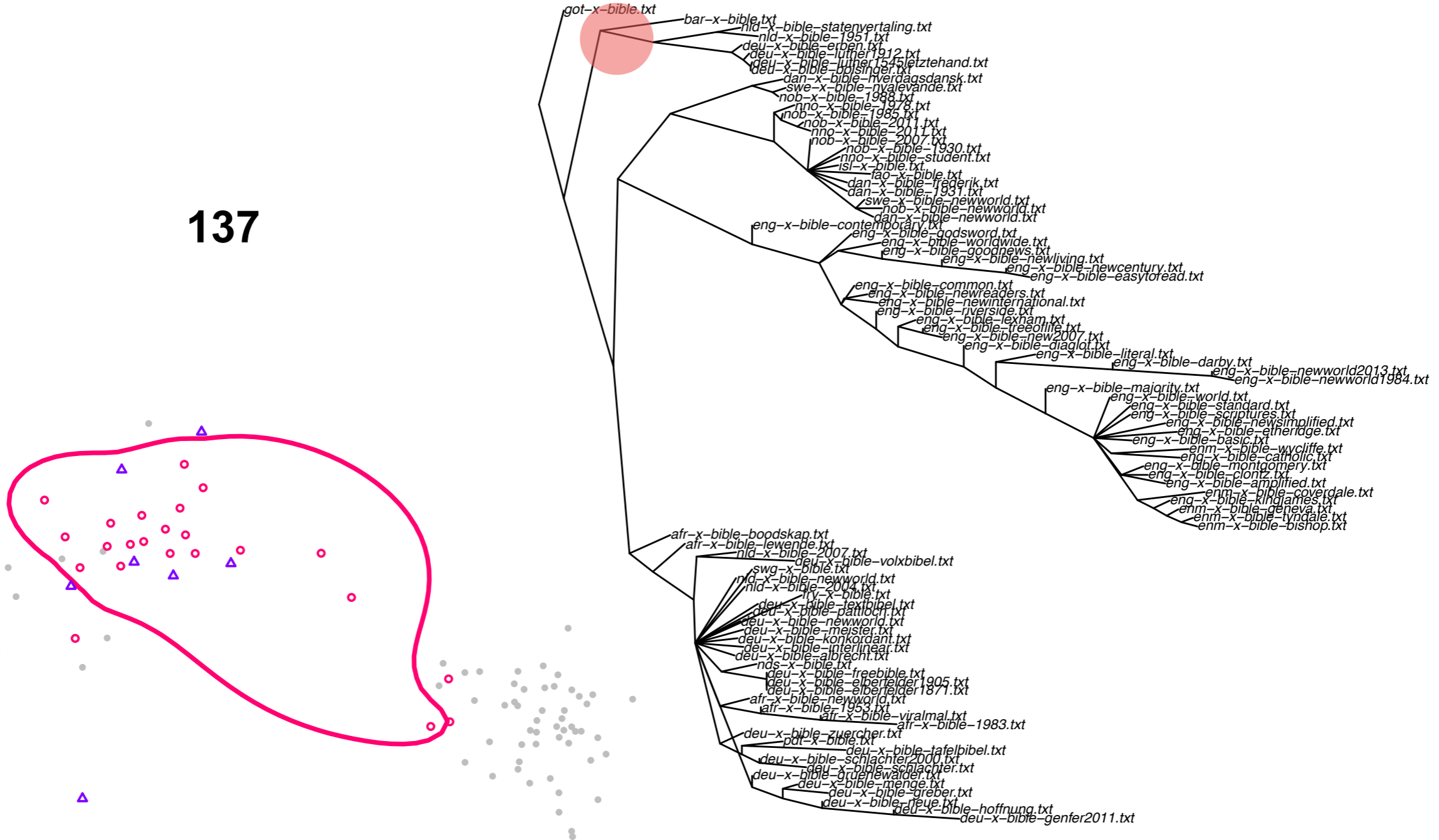
- in
- △ from
- + back

103



- to
- △ in
- + from
- × back
- ◇ near
- ▽ ward

137



- to
- △ in
- + aus

138

