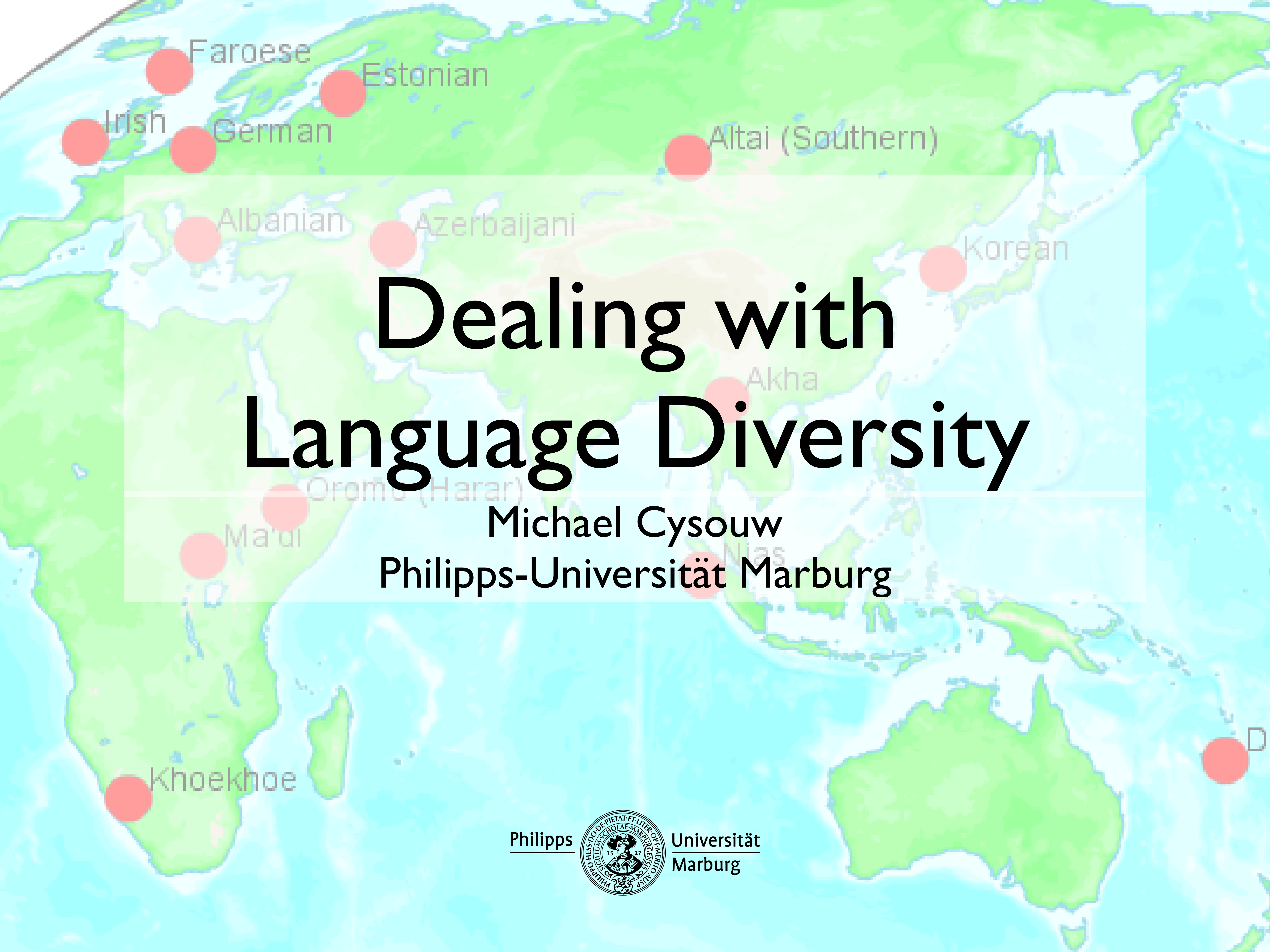
Dealing with Language Diversity

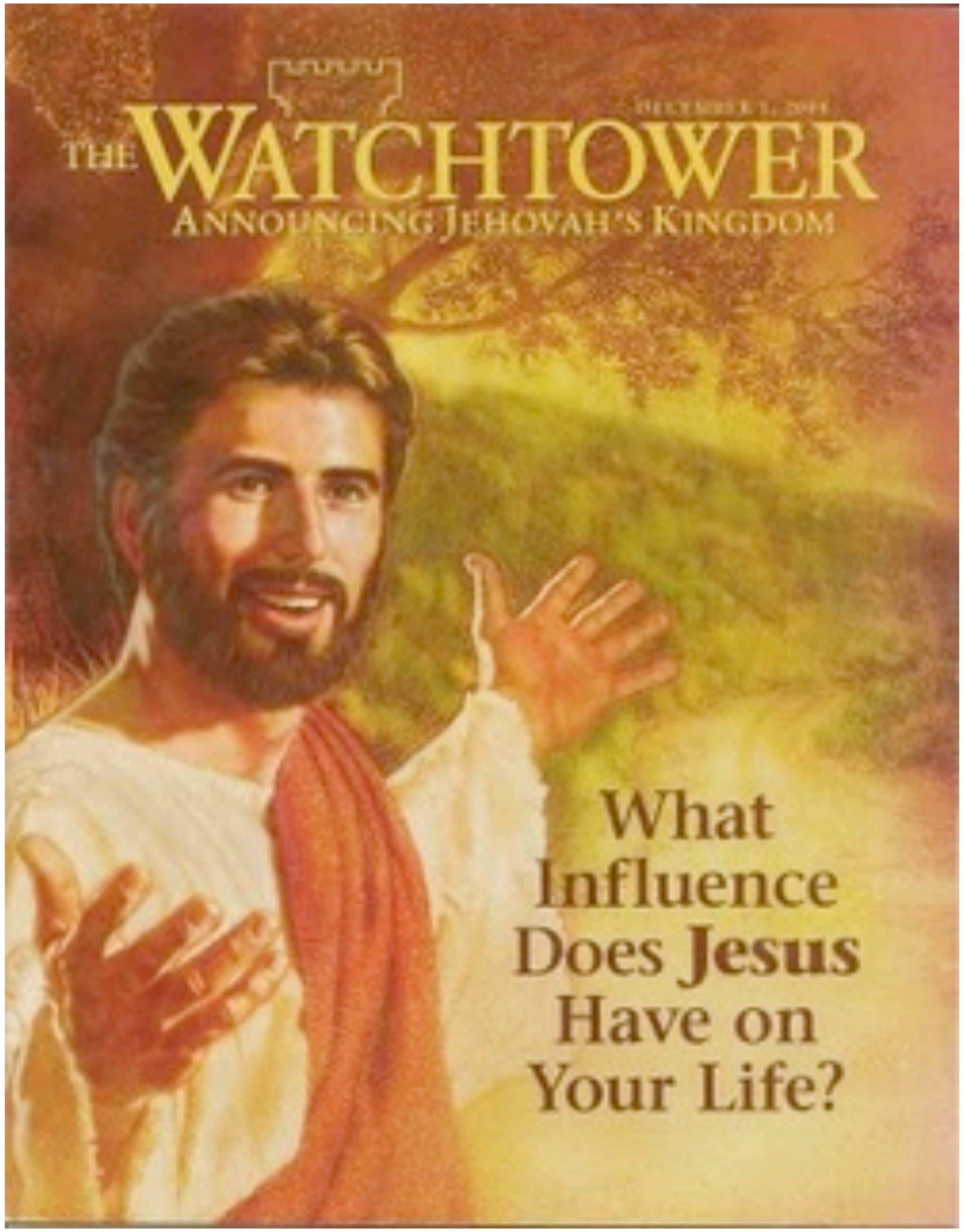
Michael Cysouw
Philipps-Universität Marburg

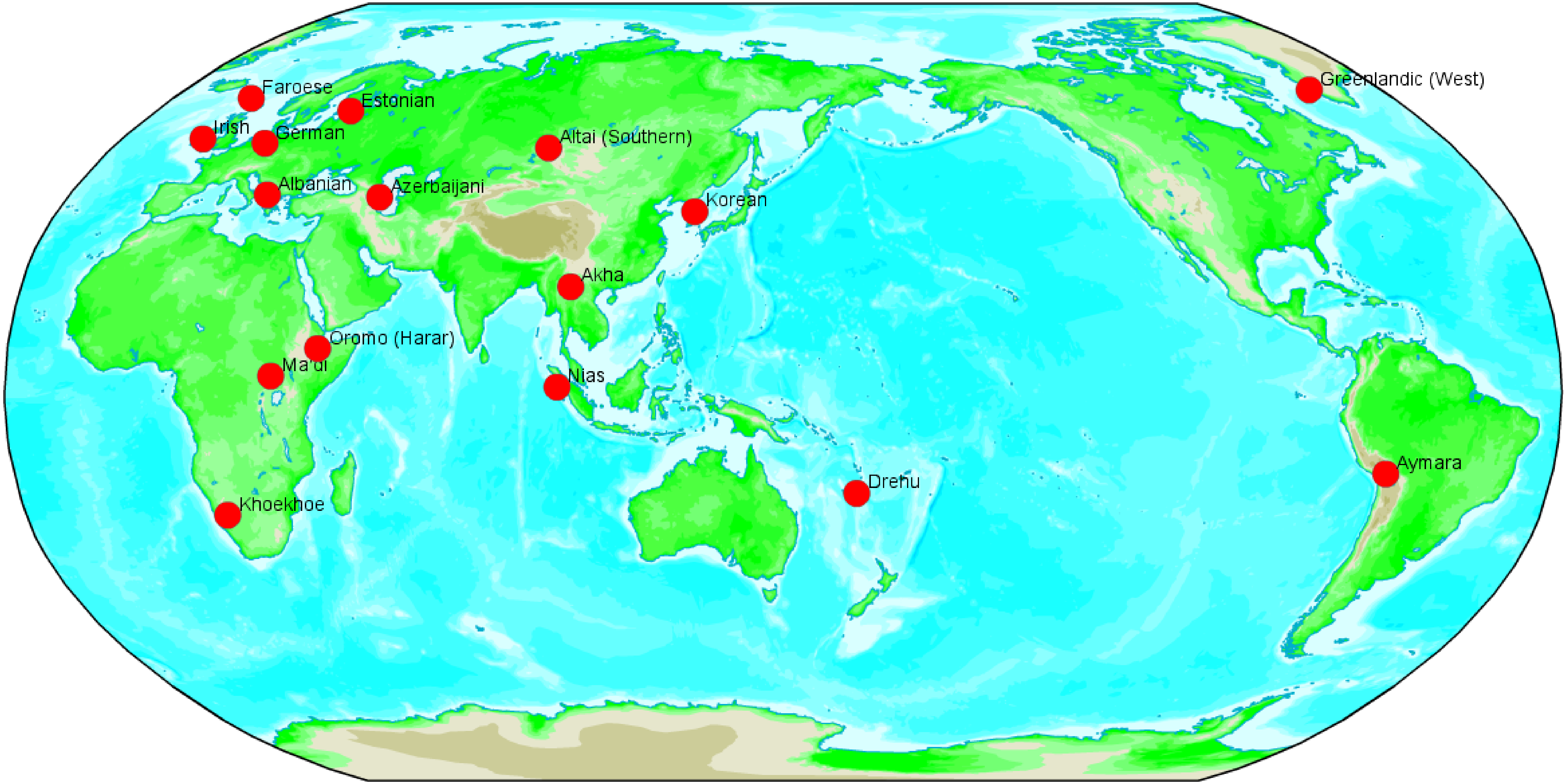


Philipps Universität Marburg

Functional Language Comparison

- Comparing morphosyntactic patterns in languages all over the world
- “Language Typology”
- Traditional problem how to compare disparate languages.
- Solution: Multi-alignment of parallel texts


DECEMBER 1,
THE WATCHTOWER
ANNOUNCING JEHOVAH'S KINGDOM
What Influence Does Jesus Have on Your Life?

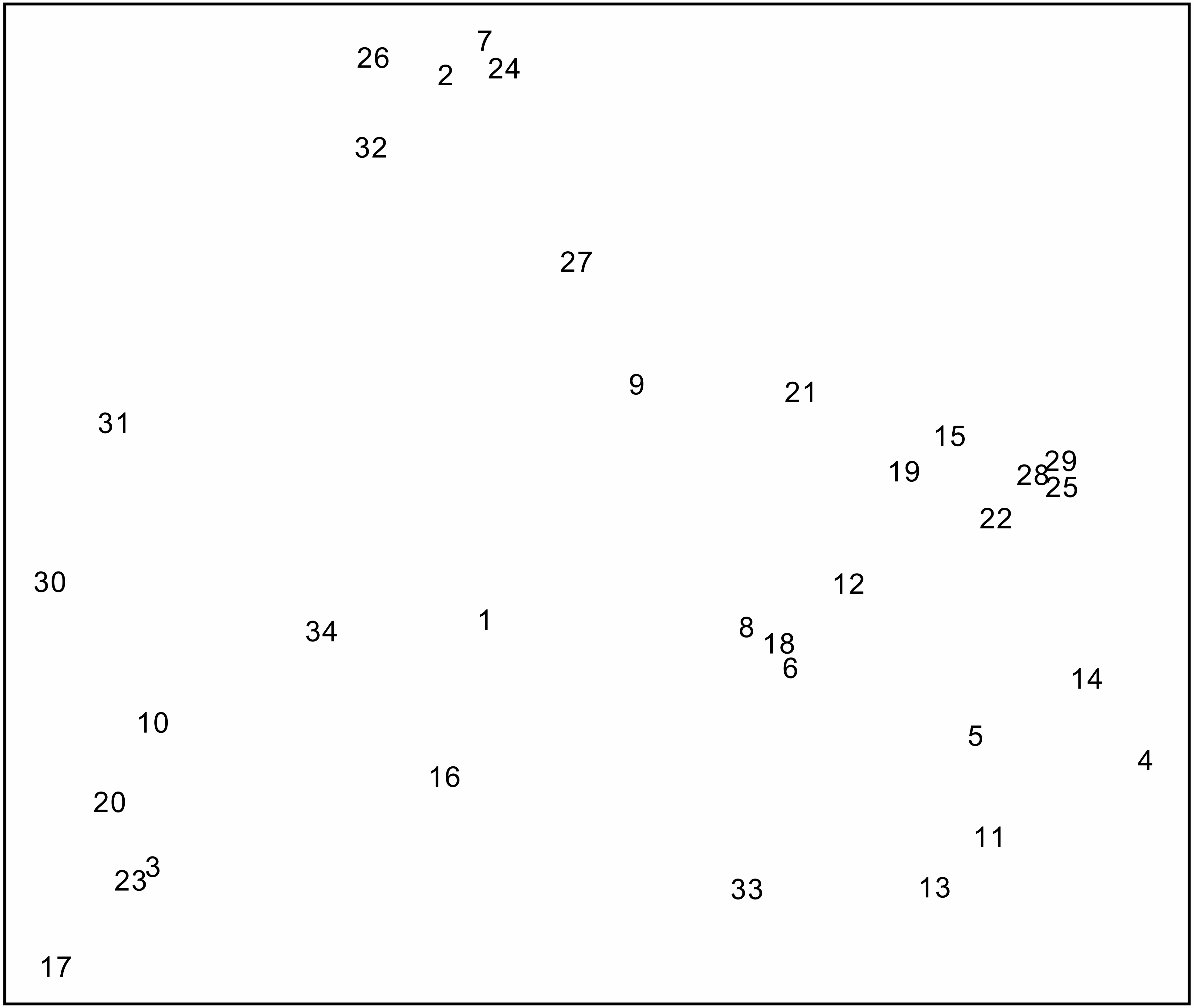

Faroese
Estonian
Irish
German
Altai (Southern)
Greenlandic (West)
Albanian
Azerbaijani
Korean
Akha
Oromo (Harar)
Ma'di
Nias
Drehu
Aymara
Khoekhoe

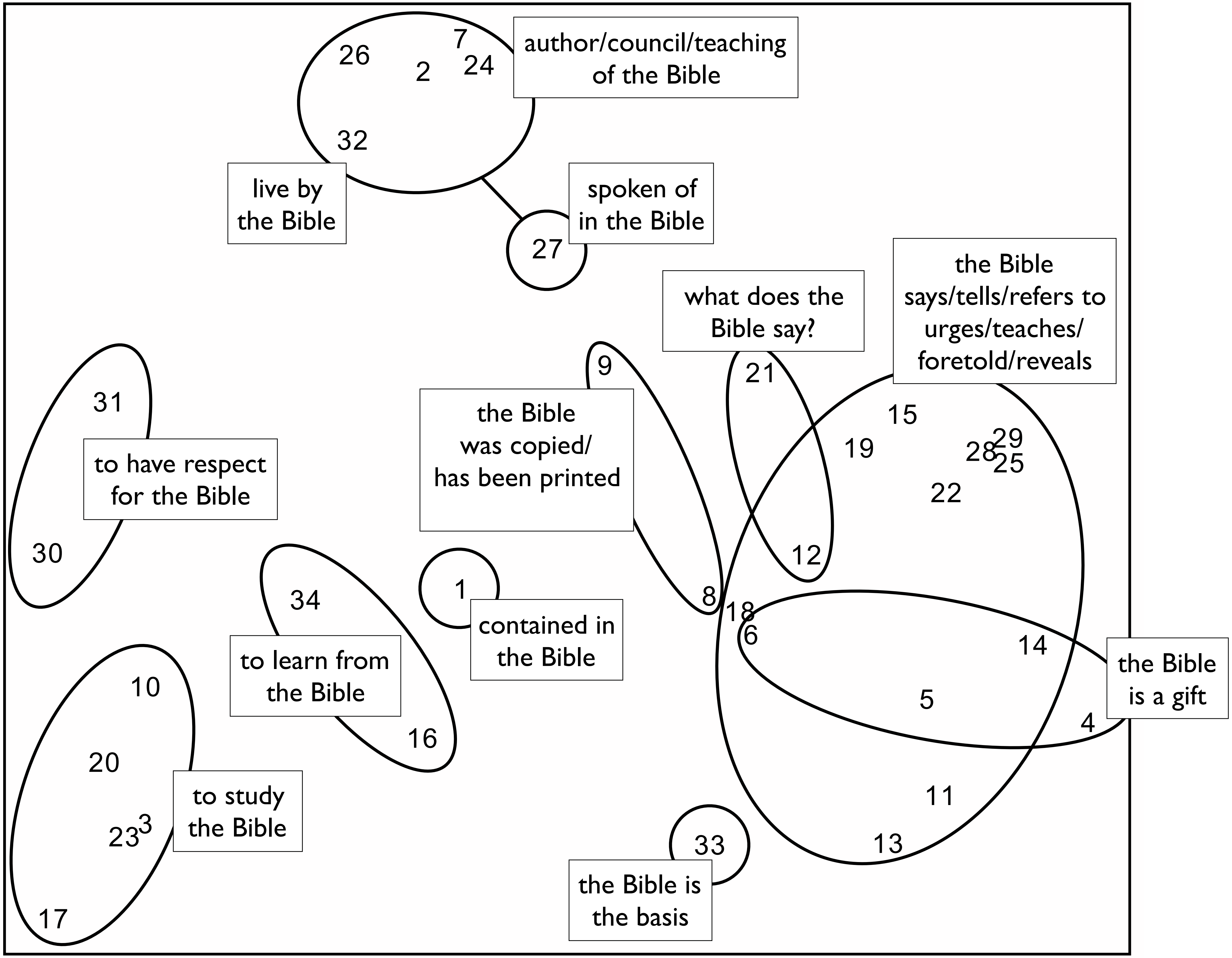
1 What important information is contained in the Bible?
2 Who is the Bible's author?
3 Why should you study the Bible?
4 The Bible is a precious gift from God.
5 The Bible alone tells us what we must do to please God.
6 The Bible was written by some 40 different men over a period of 1,600 years, beginning 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.
11 But the Bible tells us that there is only one TRUE God.
12 But when the Bible was written, the name Jehovah appeared in it some 7,000 times
13 God is a Spirit, says the Bible.
14 The Bible reveals Jehovah's personality to us.
15 The Bible tells us that he is also merciful, kind, forgiving, generous, and patient.
16 We learn about God from creation and from the Bible.
17 Another way we can learn about God is by studying the Bible.
18 By disobeying God's command, the first man, Adam, committed what the Bible calls sin.
19 This is what the Bible refers to as the ransom.
20 Some of your loved ones may become very angry because you are studying the Bible.
21 What is the Bible's view of separation and of divorce?
22 The Bible says that a husband is the head of his family.
23 Parents need to spend time with their children and study the Bible with them,
24 When marriage mates have problems getting along together, they should try to apply Bible counsel.
25 The Bible urges us to show love and to be forgiving.
26 But God does not approve of them if they come from false religion or are against Bible teachings.
27 The only two birthday celebrations spoken of in the Bible were held by persons who did not worship Jehovah.
28 The Bible teaches that only a few people are on the narrow road to life.
29 The Bible foretold that after the death of the apostles, ...
30 True Christians love one another, respect the Bible, and preach about God's Kingdom.
31 Another mark of true religion is that its members have a deep respect for the Bible.
32 They try to live by the Bible in their everyday life.
33 The Bible is the basis for what is taught.
34 By now you have learned many good things from the Bible.

Cysouw, Michael. 2014. Inducing semantic roles. In Silvia Luraghi & Heiko Narrog (eds.), *Perspectives on Semantic Roles,* Amsterdam: Benjamins.

Albanian	Faroese	Estonian	Greenlandic
bibla Nominative	*biblian* Nominative	*þiibel* Nominative	*biibiliþ* Ergative
biblën Accusative	*bibliuna* Accusative	*þiiblit* Partitive	*biibli* Absolutive
biblës Genitive/Dative	*bibliunnar* Genitive	*þiibli* Genitive	*biibilmik* Instrumental
…	*bibliuni* Dative	*þiiblis* Inessive	*biibilmi* Locative
	…	*þiiblist* Elative	…
		…	

Context	Albanian	Faroese	Estonian	Greenlandic
1	bibla	bíbliuni	piibel	biibili
2	biblës	bíbliunnar	piibli	biibilimik
3	biblën	bíbliuna	piiblit	biibili
4	bibla	bíblian	piibel	biibili
5	bibla	bíblian	piibel	biibilip
6	bibla	bíbliuna	piibli	biibili
7	biblës	bíbliunnar	piibli	biibilimut
8	bibla	bíblian	piiblit	biibilip
9	bibla	NA	piiblit	biibili
10	biblën	bíbliuna	piiblit	biibilimik
11	bibla	bíblian	piibel	biibilimili
12	bibla	bíblian	piibel	biibilili
13	bibla	bíblian	piibel	biibilimi
14	bibla	bíblian	piibel	biibilimi
15	bibla	bíblian	piibel	biibilimi
16	bibla	bíbliuni	piibli	biibililu
17	biblën	bíbliuna	piiblit	biibilimik
18	bibla	bíblian	piiblis	biibilip
19	bibla	bíblian	piiblis	biibilimi
20	biblën	bíbliuna	piiblit	biibilimik
21	NA	bíblian	piibel	biibilimi
22	bibla	bíbliuni	piibel	biibili
23	biblën	bíbliuna	piiblit	biibilimillu
24	biblike	bíblian	piibli	biibilimi
25	bibla	bíblian	piibel	biibilimi
26	biblës	bíbliunnar	piibli	biibilimi
27	bibla	bíblian	piiblis	biibilimi
28	bibla	bíblian	piibel	biibilimi
29	bibla	bíblian	piibel	biibilimi
30	biblën	bíbliuna	piiblist	biibilimik
31	biblën	bíbliuni	piibli	biibilimik
32	biblës	bíbliuni	piibli	biibili
33	bibla	bíbliuna	piibel	biibilimik
34	bibla	bíbliuni	piiblist	biibilimeersunik


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



author/council/teaching of the Bible

live by the Bible

spoken of in the Bible

what does the Bible say?

the Bible says/tells/refers to urges/teaches/ foretold/reveals

to have respect for the Bible

the Bible was copied/ has been printed

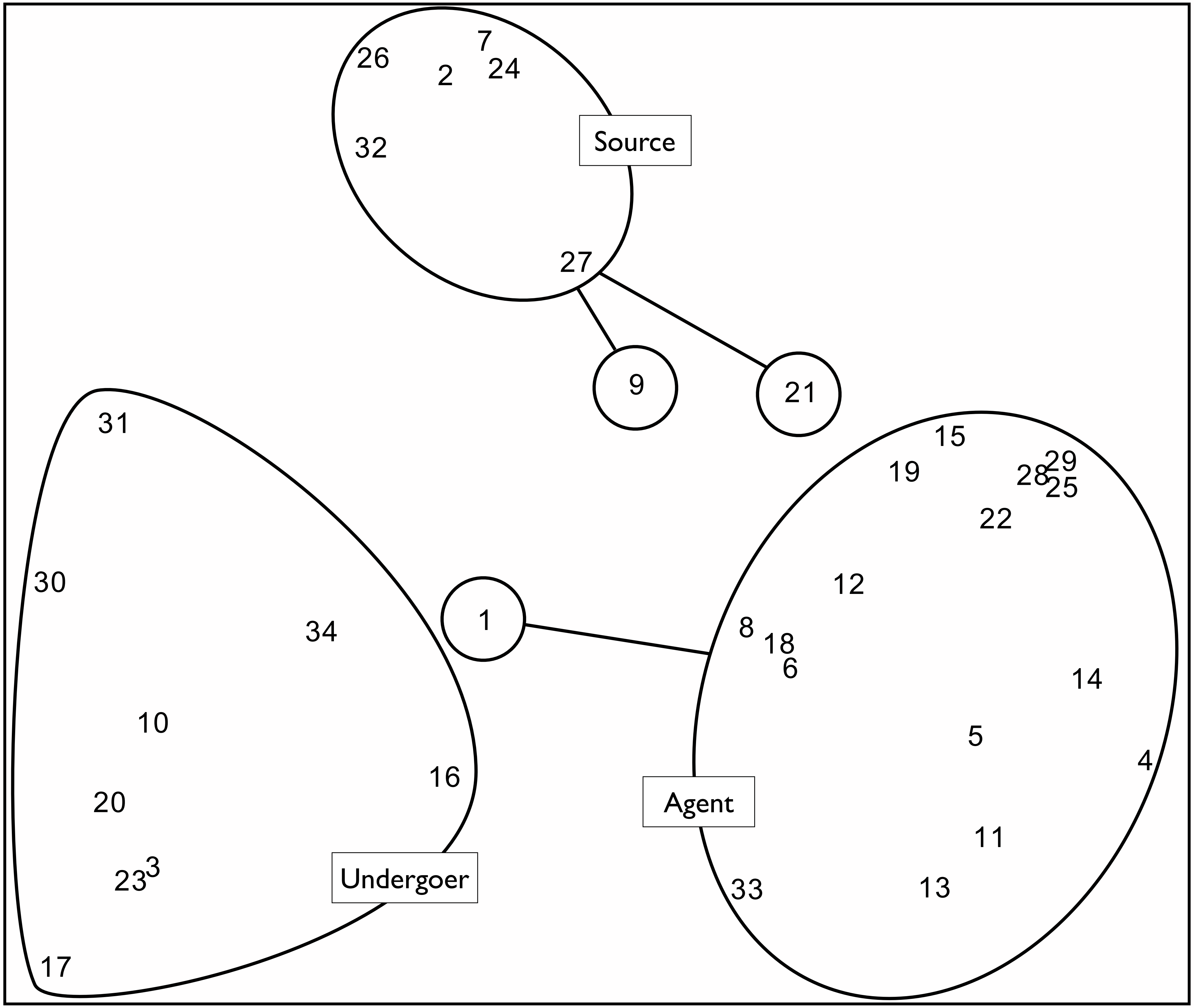
contained in the Bible

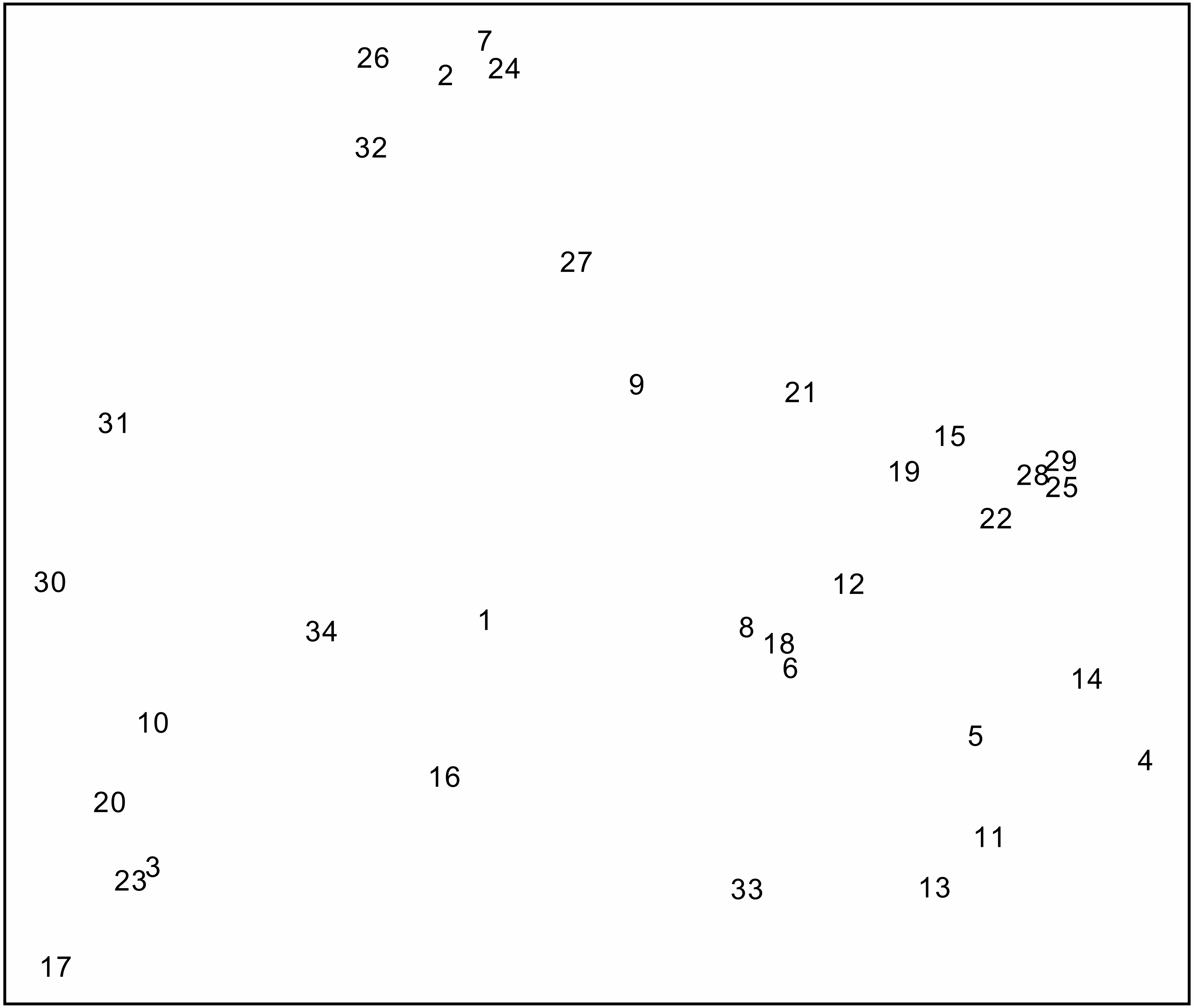
to learn from the Bible

the Bible is a gift

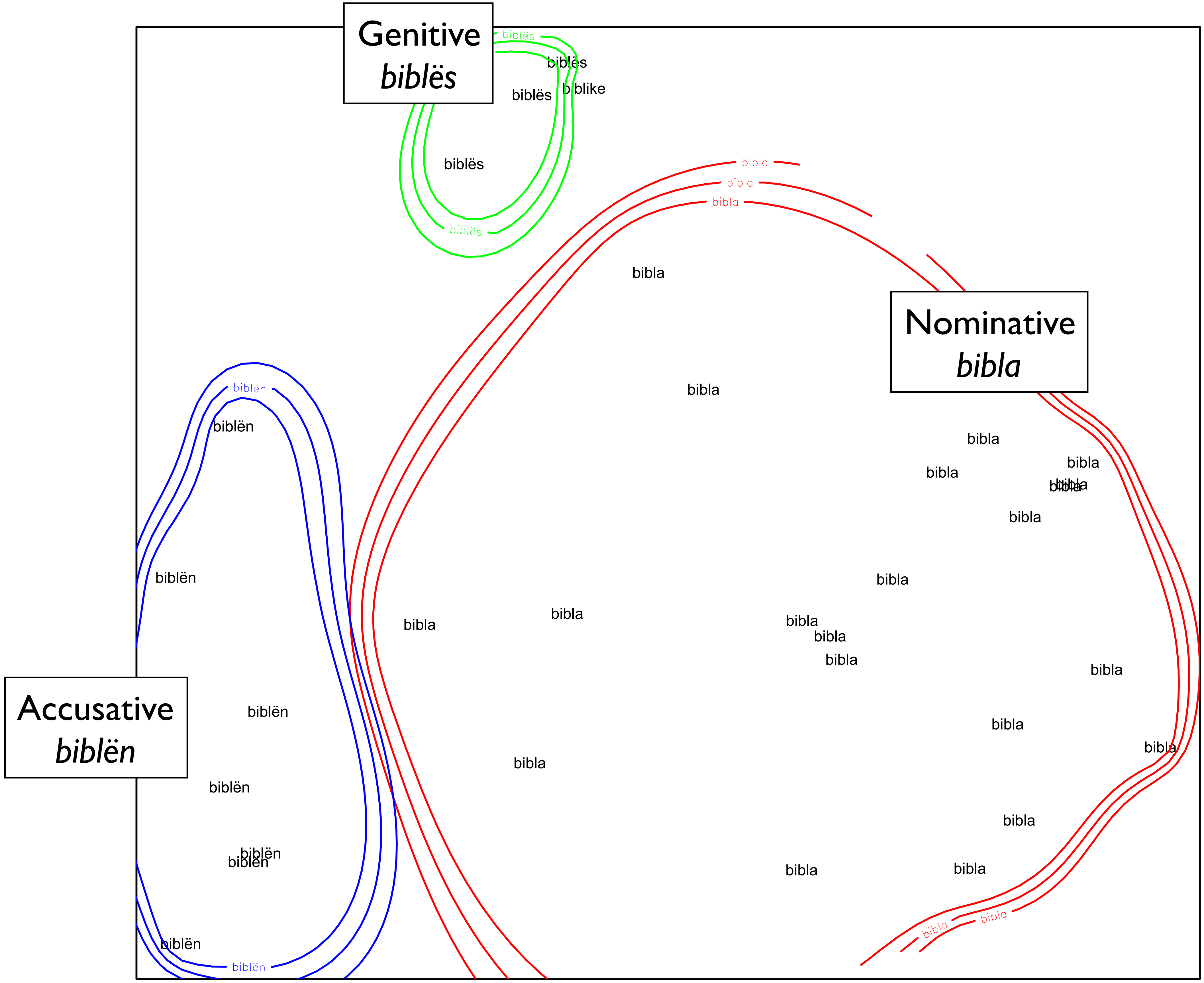
to study the Bible

the Bible is the basis


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


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

albanian

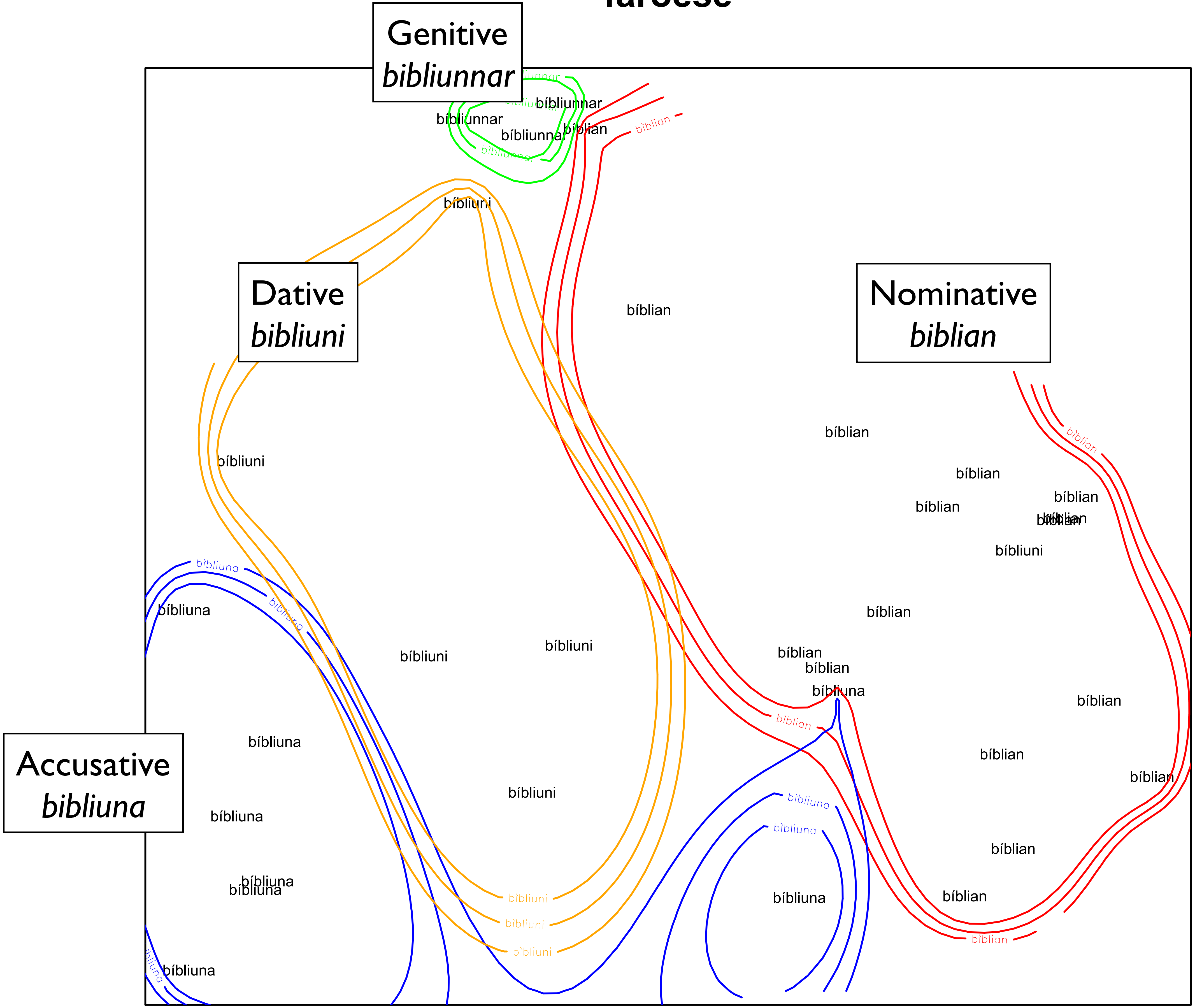


Genitive *biblës*

Nominative *bibla*

Accusative *biblën*

faroese



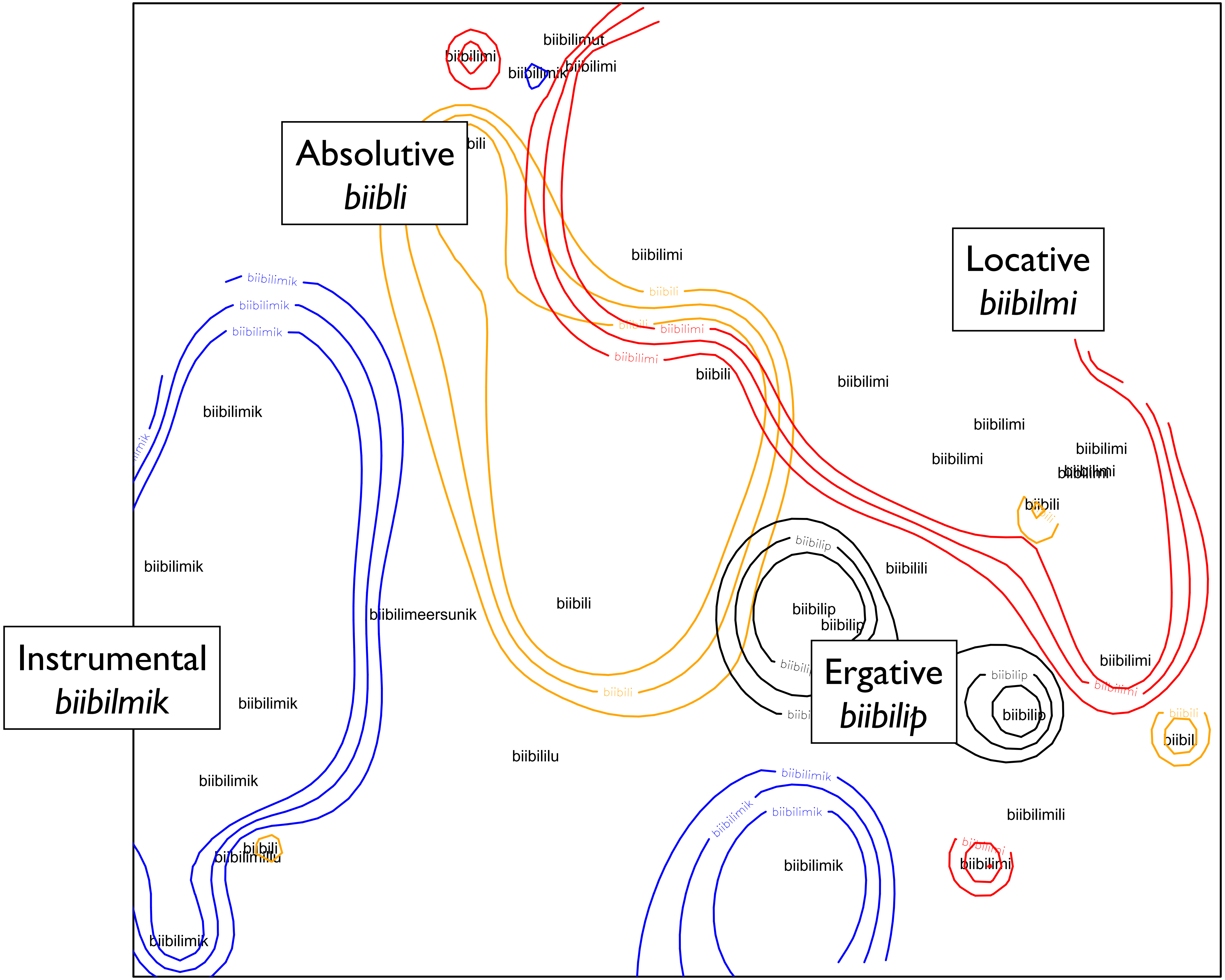
Genitive
bibliunnar

Dative
bibliuni

Nominative
biblian

Accusative
bibliuna

greenlandic



Absolutive
biibli

Locative
biibilmi

Instrumental
biibilmik

Ergative
biibilip

Parallel Bible Corpus

- more than 1600 translations
- (about 1200 ISO 639-3 codes ~ ‘languages’)
- http://paralleltext.info/data
- (closed GitHub source because of copyright)

Multiple Alignment

- Based on sentence-by-sentence alignment,
  induce word-by-word alignment
- Translations can be (and often are!) quite different
- Bi-text alignment is widely researched problem
- Mulit-text alignment not so much
  (but multi-string alignment in bio-informatics is!)

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

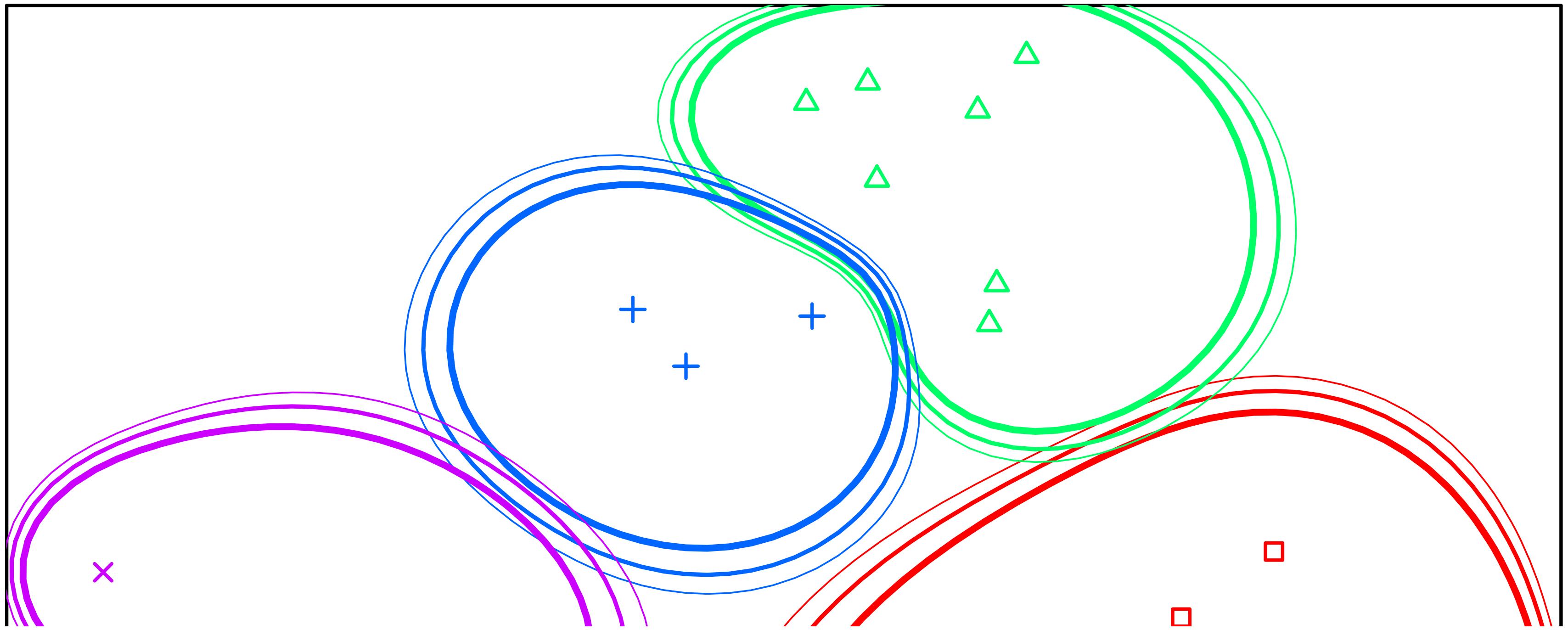
Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Multiple Alignment

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

trees and wood

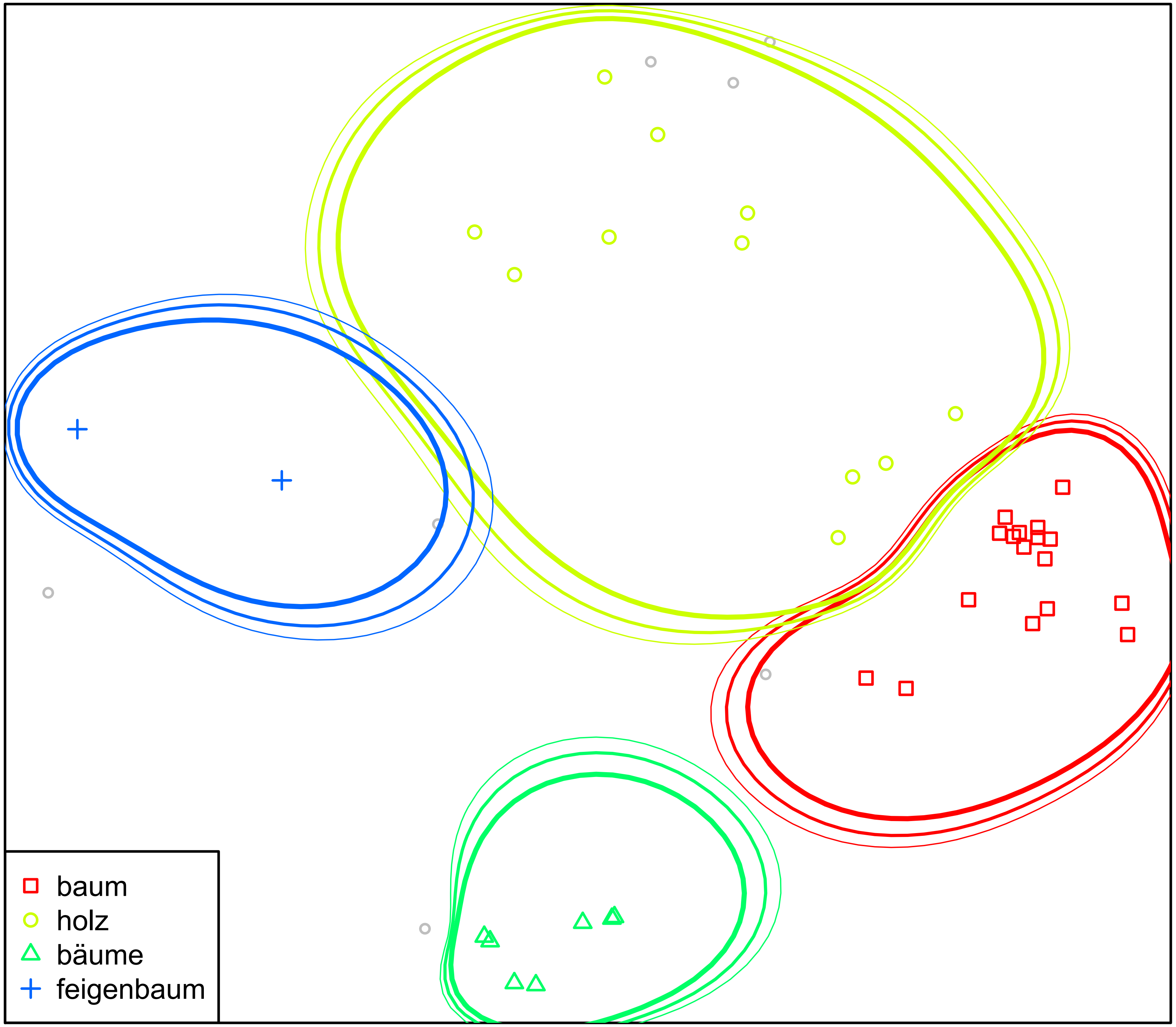
afr-x-bible-1953.txt



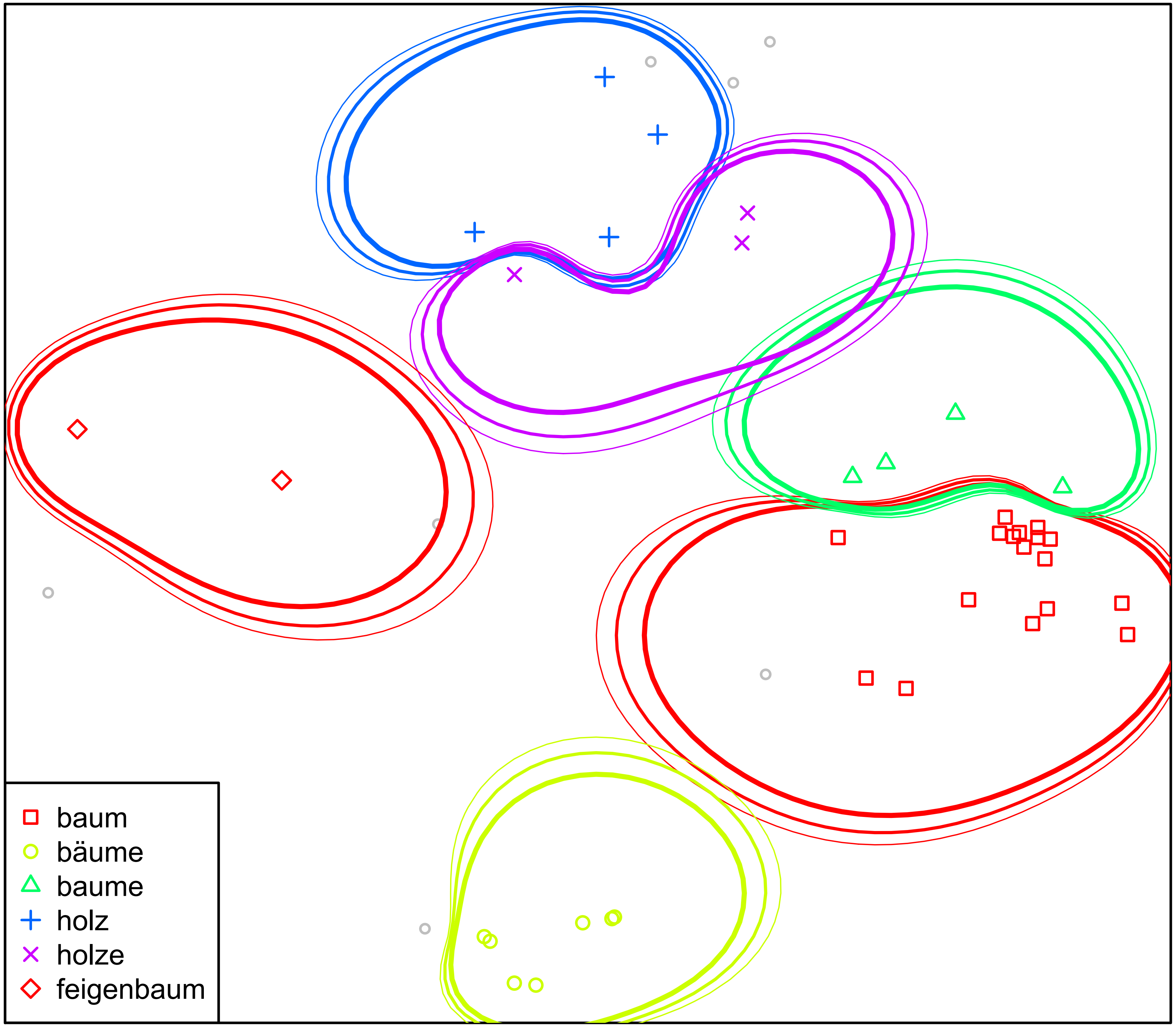
	tree	wood (stuff)	firewood	small forest	large forest
German	*Baum*	*Holz*		*Wald*	
Danish	*trae*			*skov*	
French	*arbre*	*bois*			*forêt*
Spanish	*árbol*	*madera*	*leña*	*bosque*	*selva*

Louis Hjelmslev
Prolegomena to a Theory of Language (1963)

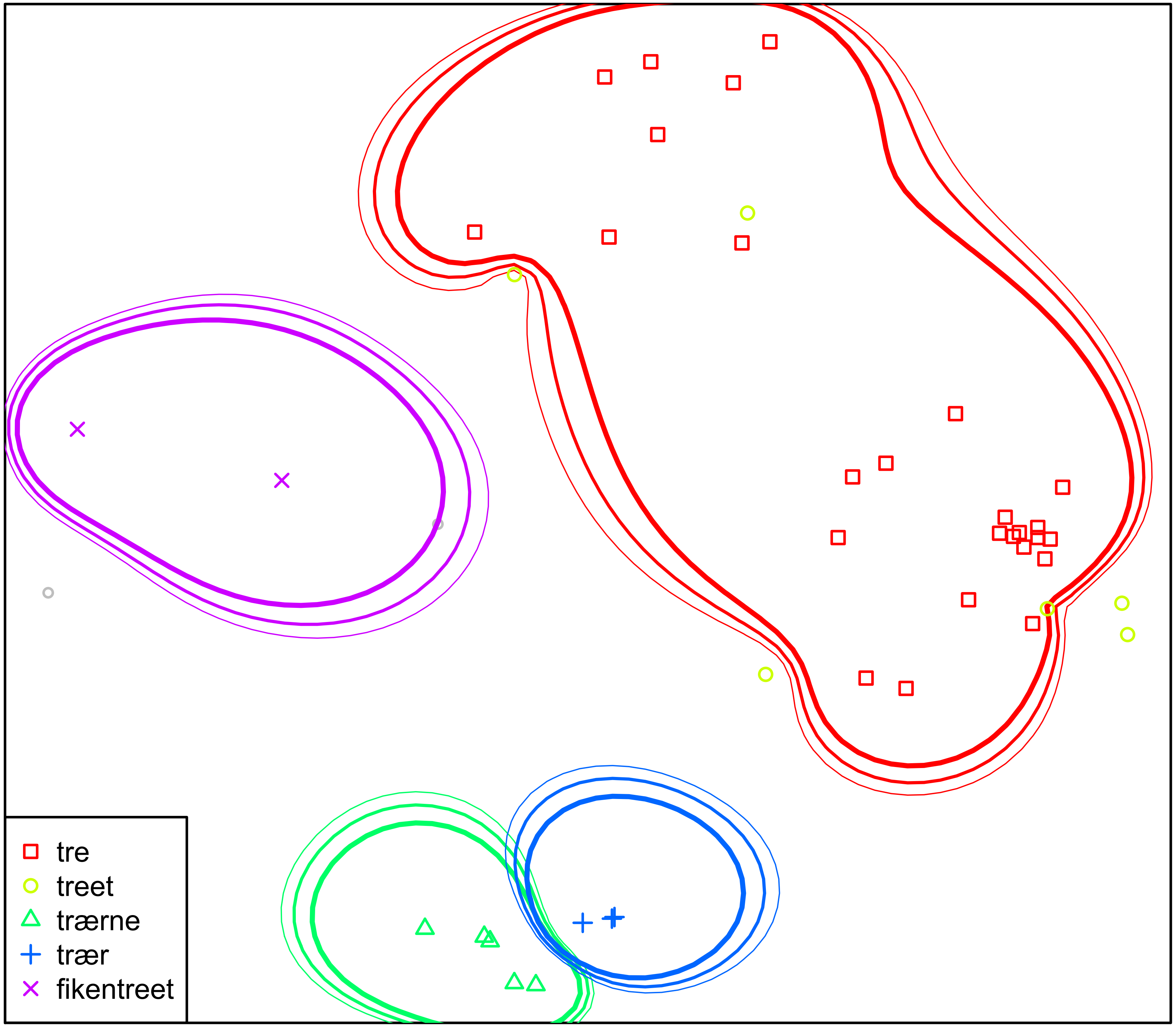
deu-x-bible-erben.txt



deu-x-bible-freebible.txt

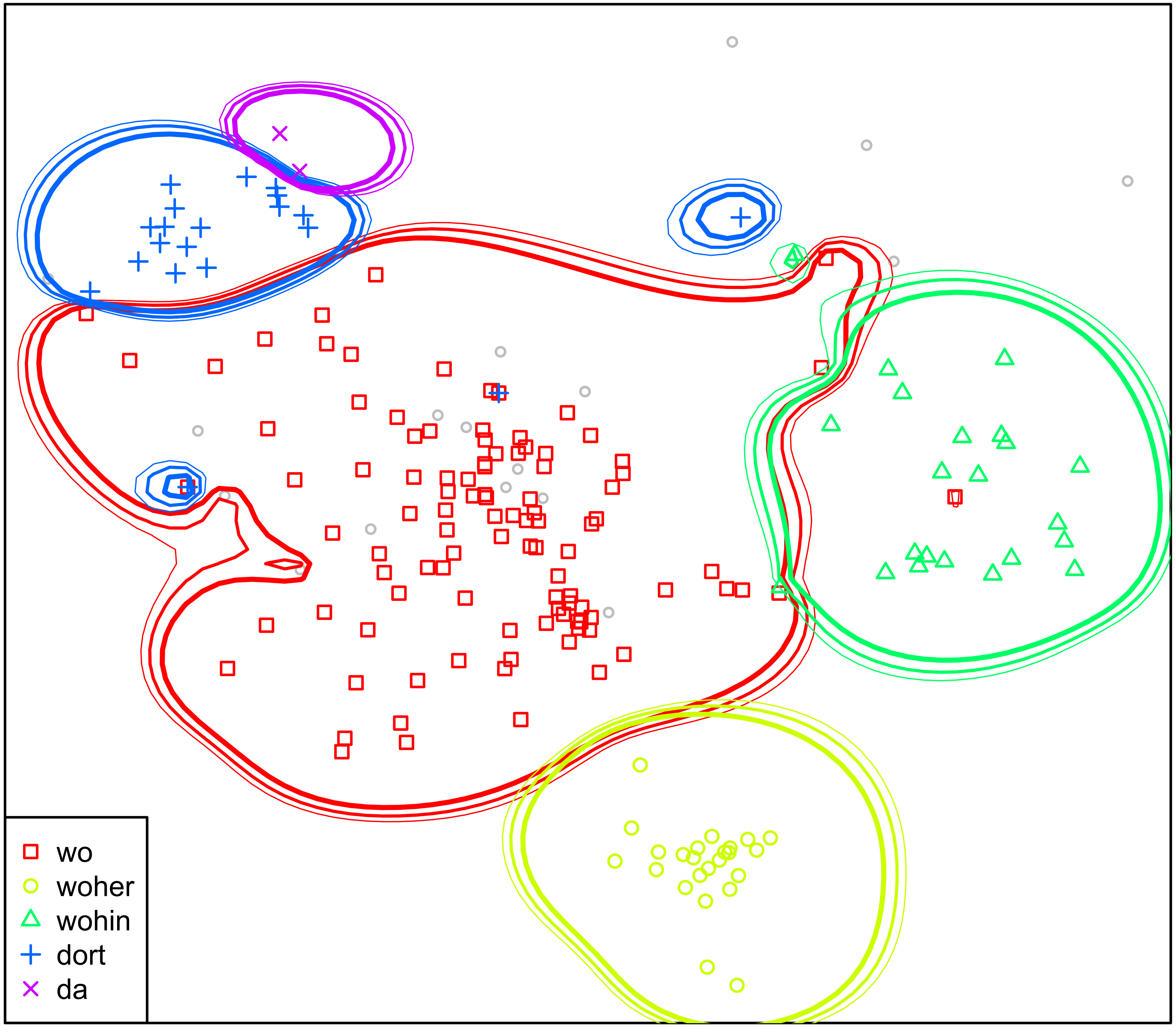


nob-x-bible-2007.txt

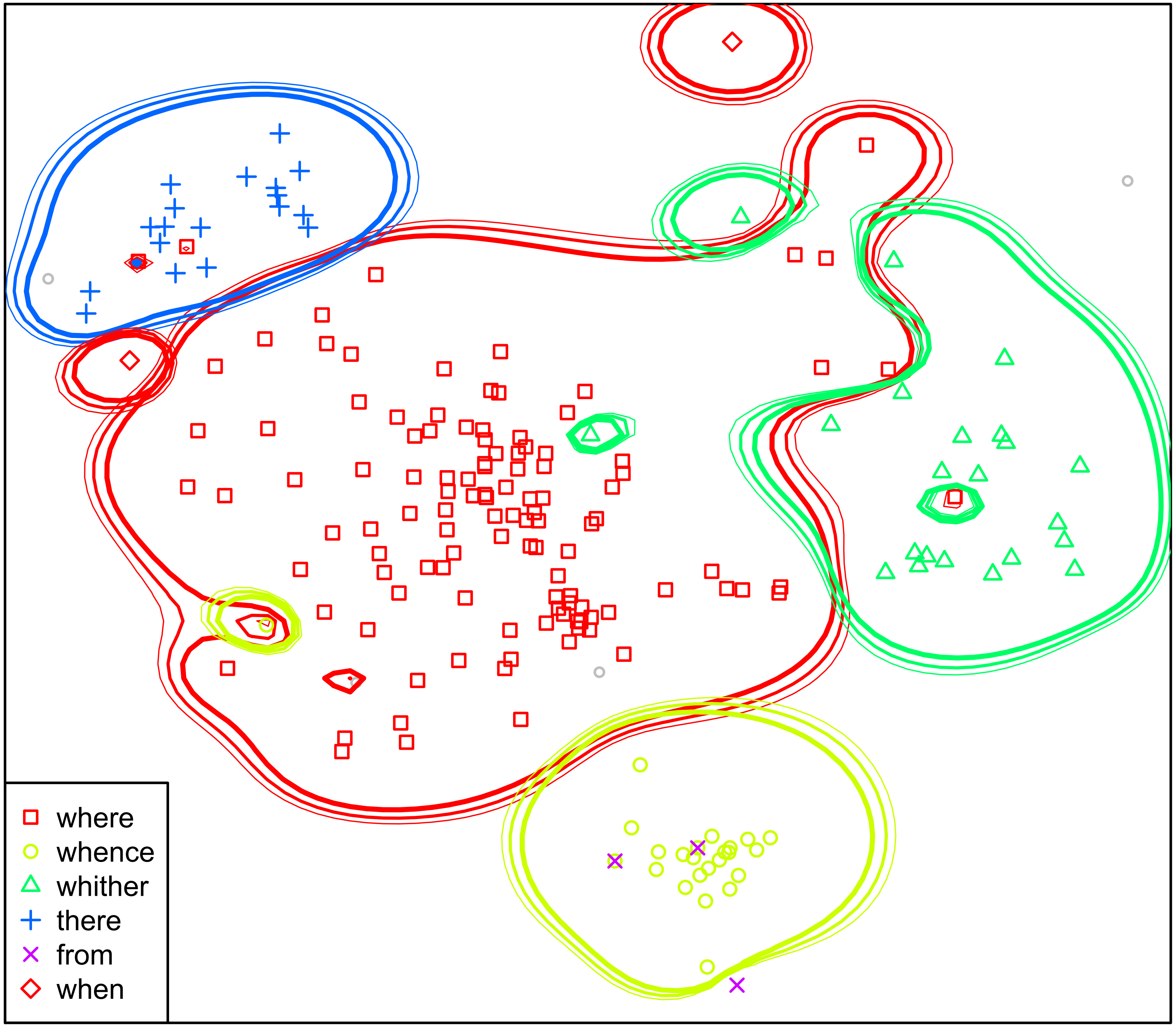


where

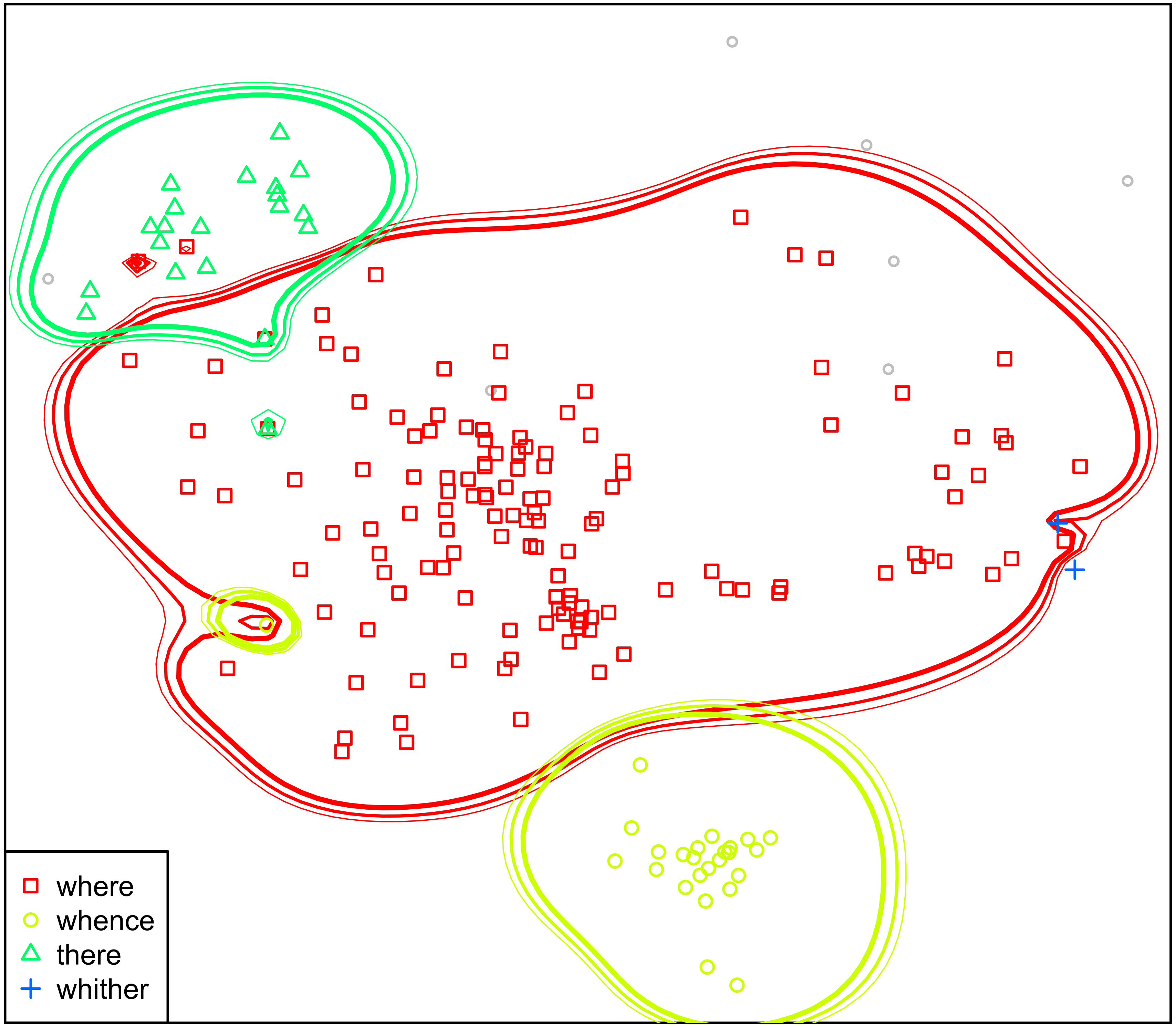
deu-x-bible-pattloch.txt



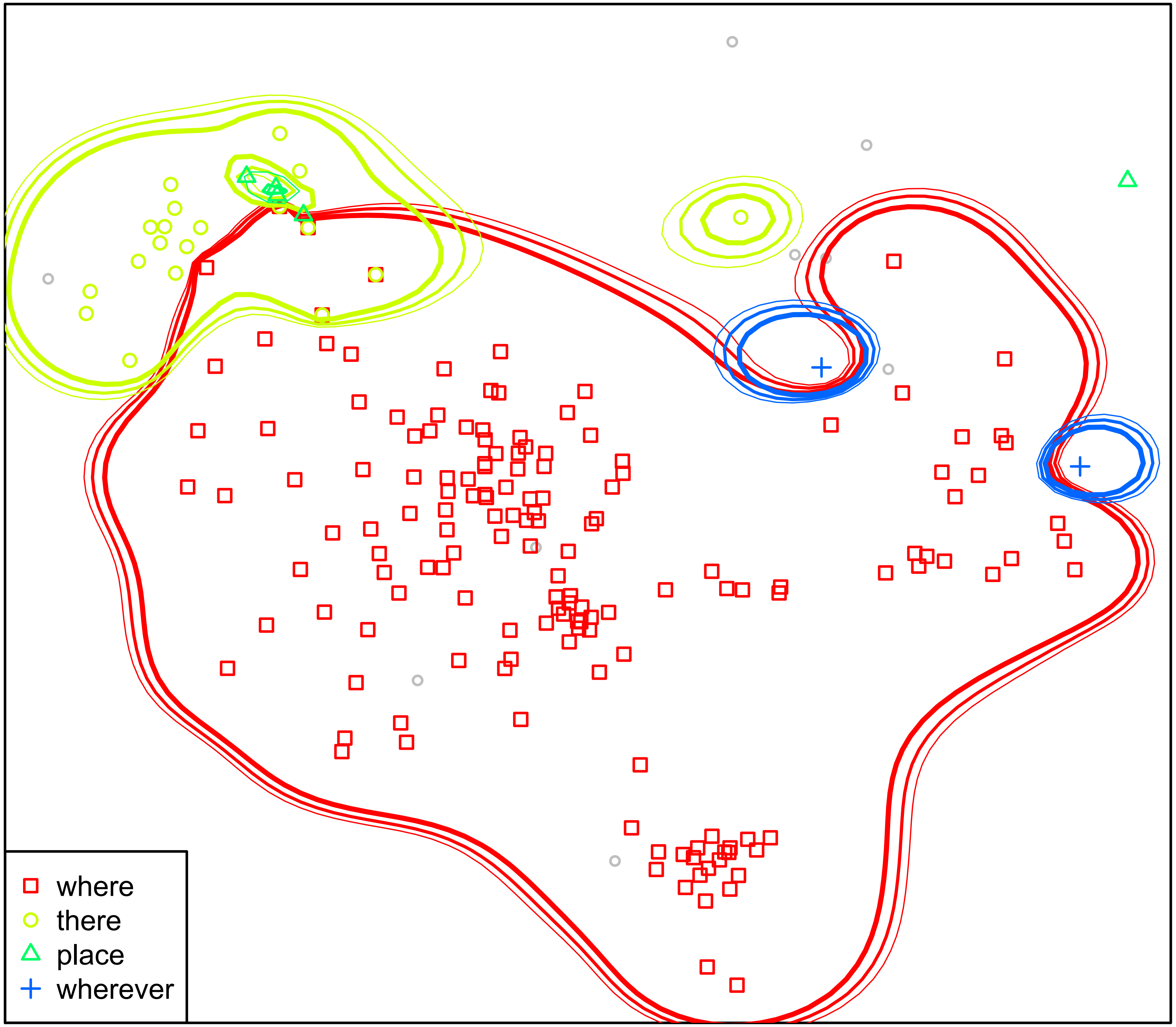
eng-x-bible-kingjames.txt



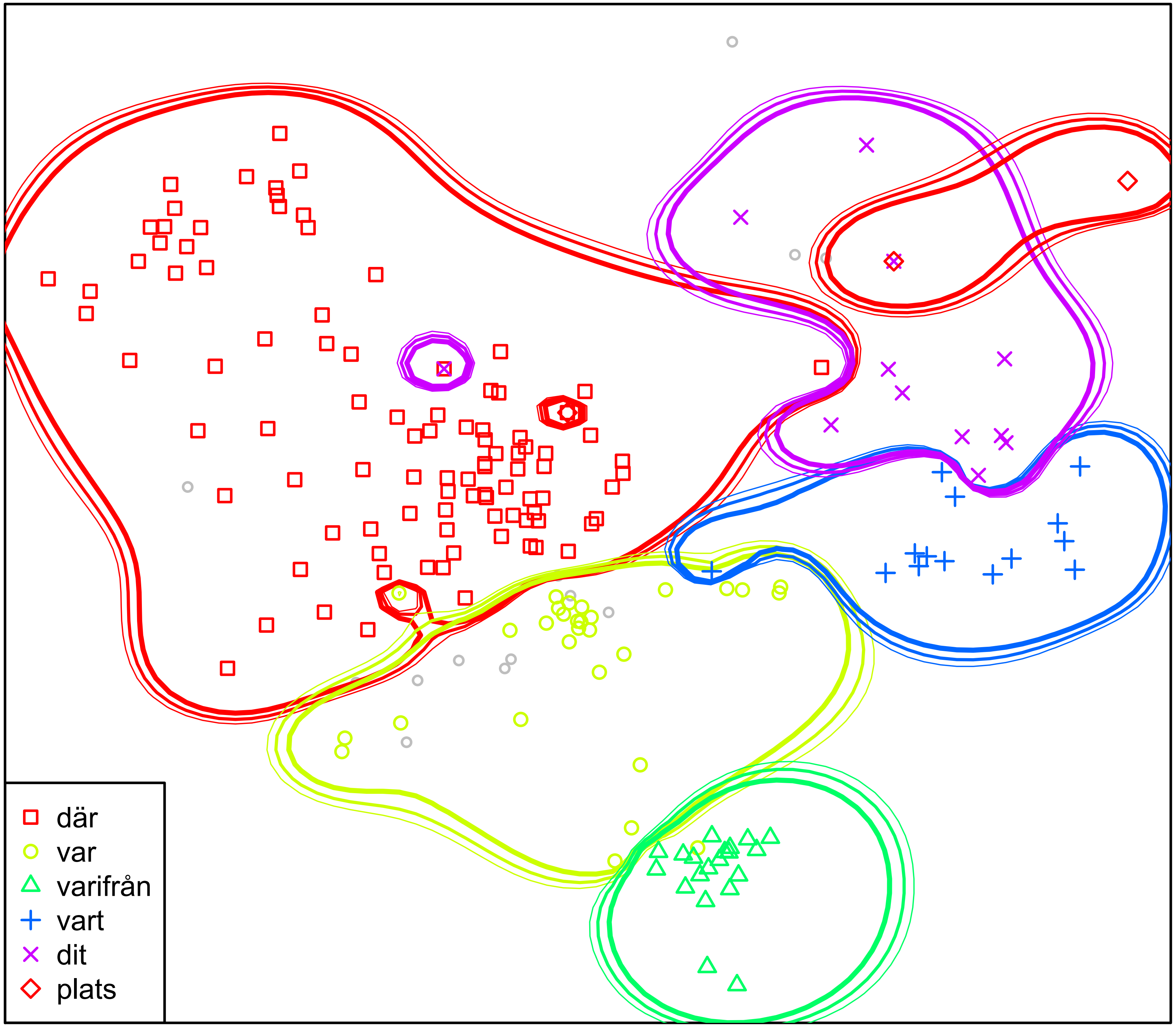
eng-x-bible-darby.txt



eng-x-bible-treeoflife.txt



swe-x-bible-folk1998.txt



Translation of “*Jerusalem*”

- Identify wordforms that correspond to the English name *Jerusalem* in 100 translations
- 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

Angaataha

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

- jerusaremɨhanda
- jerusaremɨhandaahapɨ
- jerusaremɨhandɨ
- jerusaremɨhandaahiyai
- jerusaremɨhandaahɨ
- jerusaremɨhandaahapɨhiyaunɨ
- jerusaremɨhandaahiyaisangi
- jerusaremɨhandaahapɨhiya
- jerusaremɨhandaahɨraapɨ
- jerusaremɨhandaahɨhɨ
- jerusaremɨhandaahɨhe
- jerusaremɨhandamɨ
- jerusaremɨmanda
- jerusaremɨhandapɨ
- jerusaremɨndɨ
- jerusaremɨhandaahapɨto
- jerusaremɨhandaahapaahɨhɨ
- jerusaremɨhandi
- jerusaremɨmandaahapɨ
- jerusaremɨhandaahunɨ
- jerusaremɨhandaahapunɨ
- jerusaremɨhandaahiya
- jerusaremɨhandamɨhinɨ
- jerusaremɨhandaahapɨhiyaatihɨ
- jerusaremɨhandaahapɨhiyaate
- jerusaremɨhandaahiyaunɨ

Amharic

(ISO 639-3 amh, spoken in Ethiopia)

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

Amharic

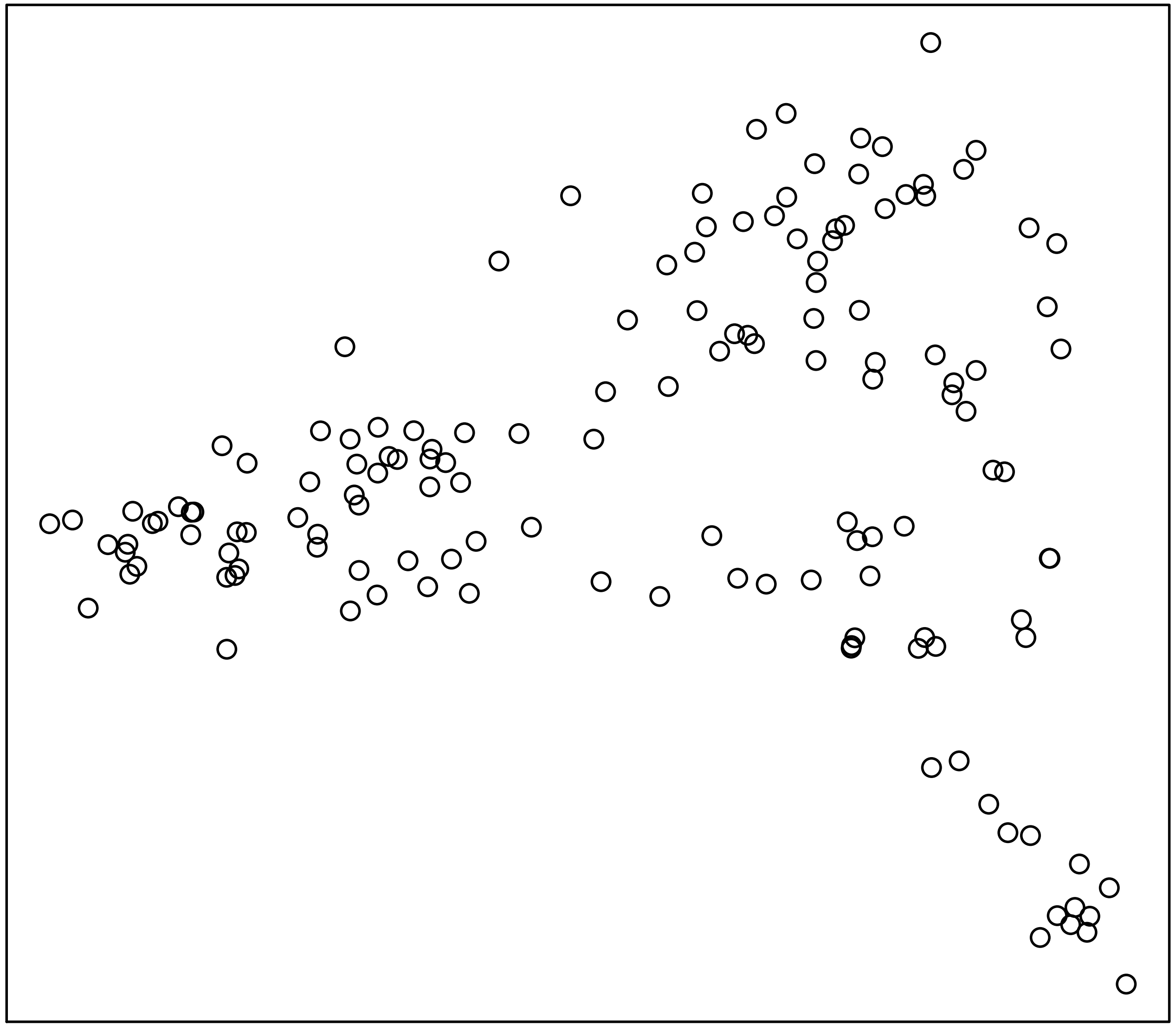
(ISO 639-3 amh, spoken in Ethiopia)

- 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

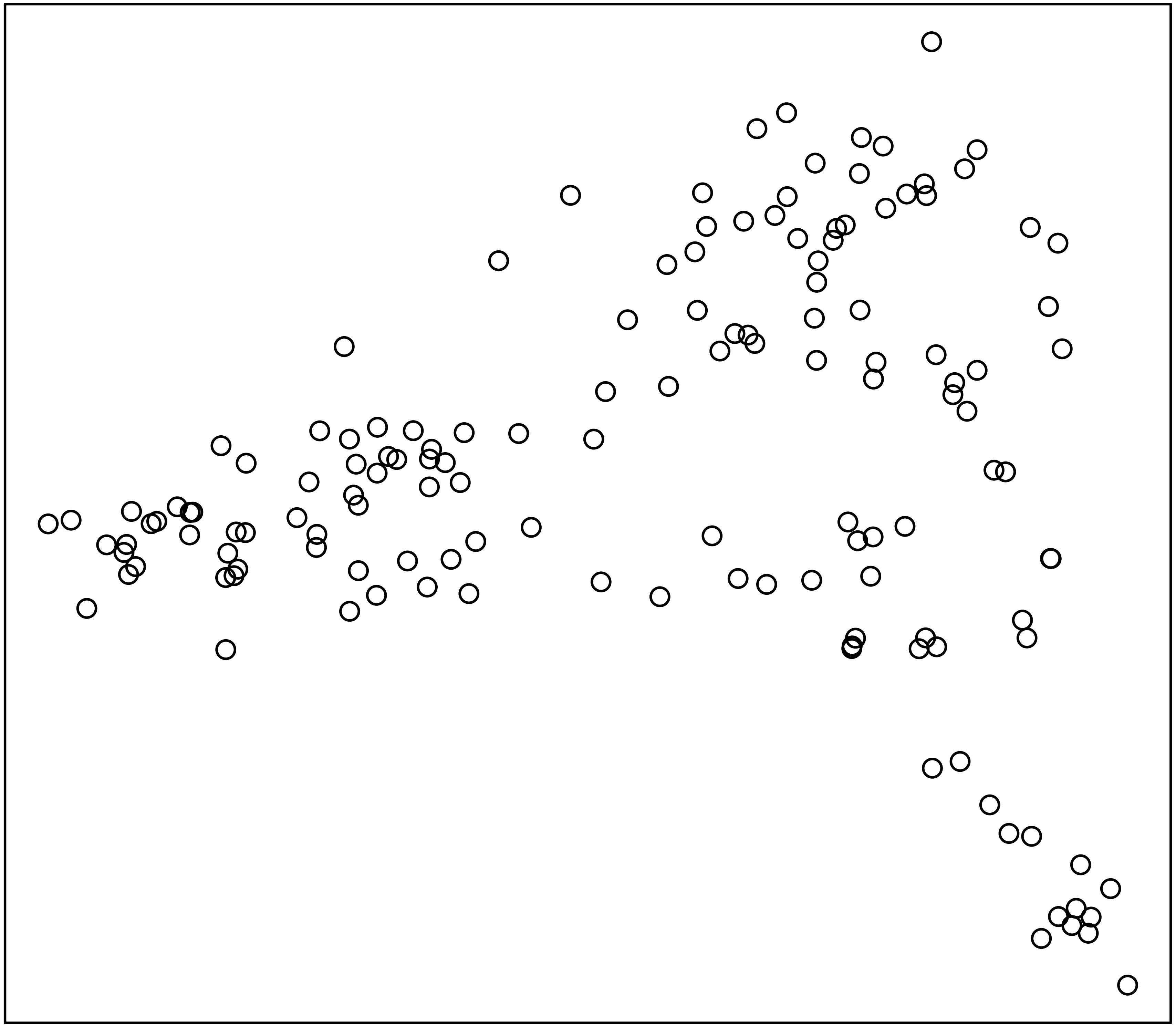
Matrix of size 520 x 167 coding the distribution of wordforms over verses



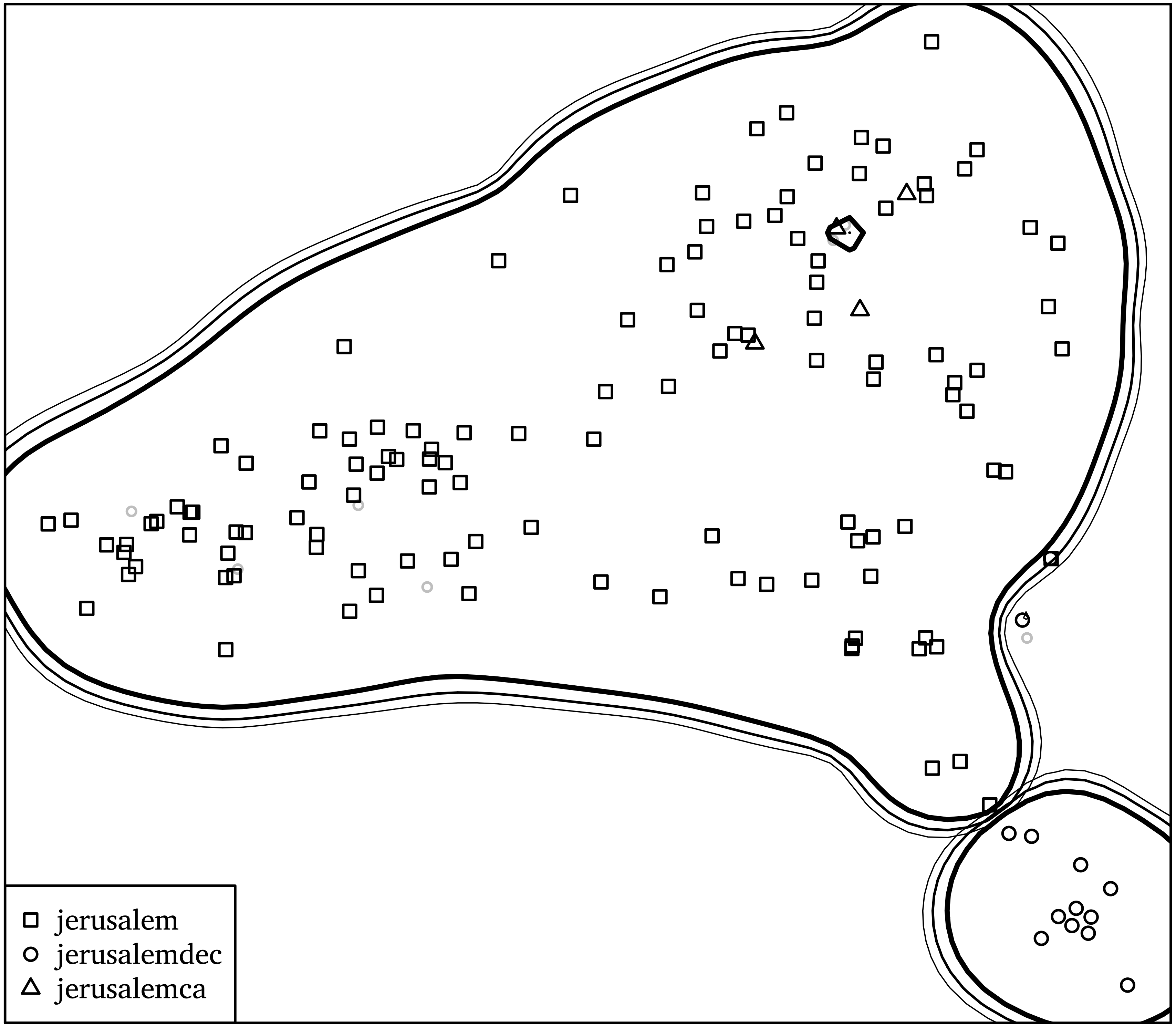
Two contexts of *Jerusalem* are similar when they often share the same wordform in language after language

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

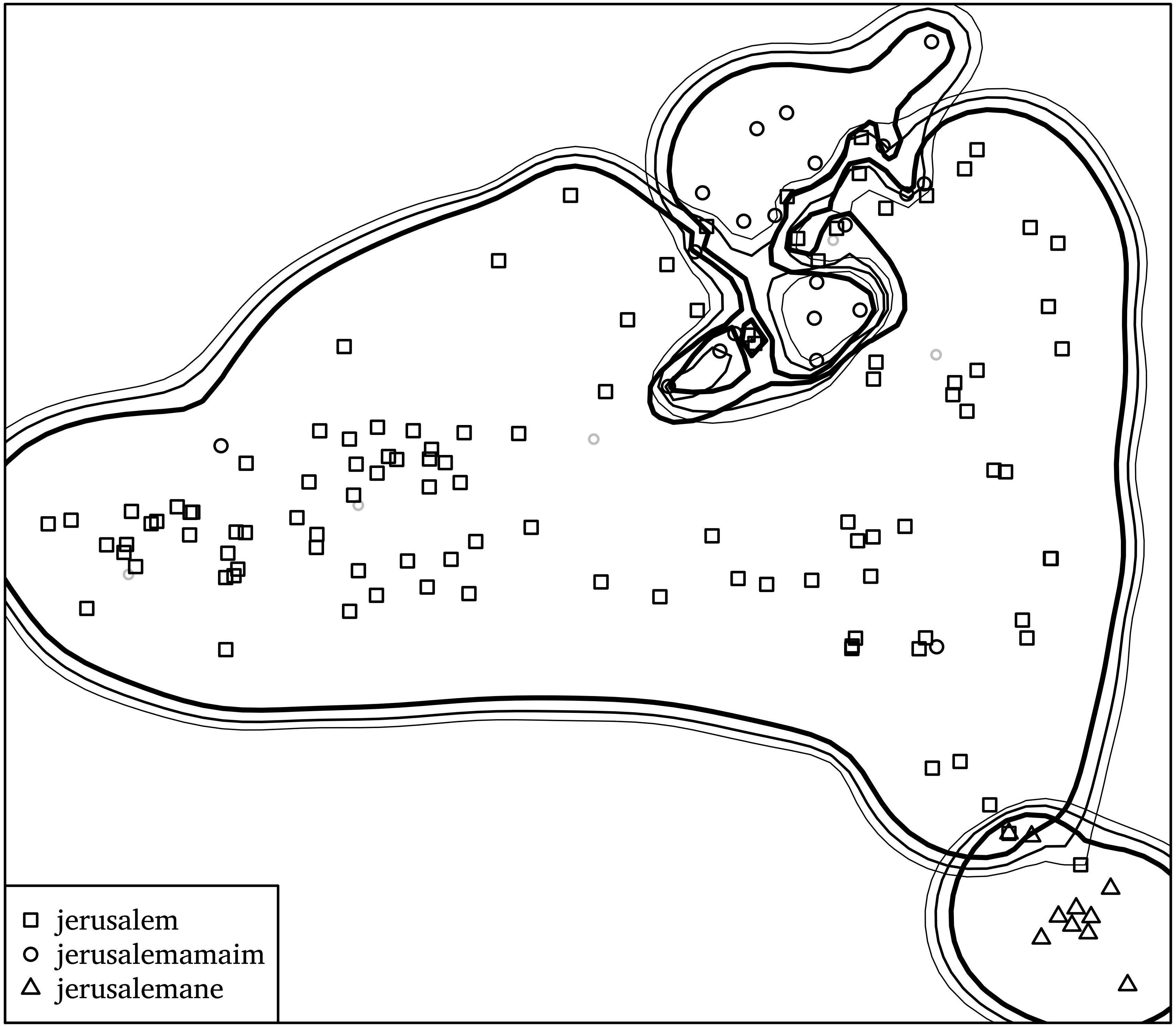


aey



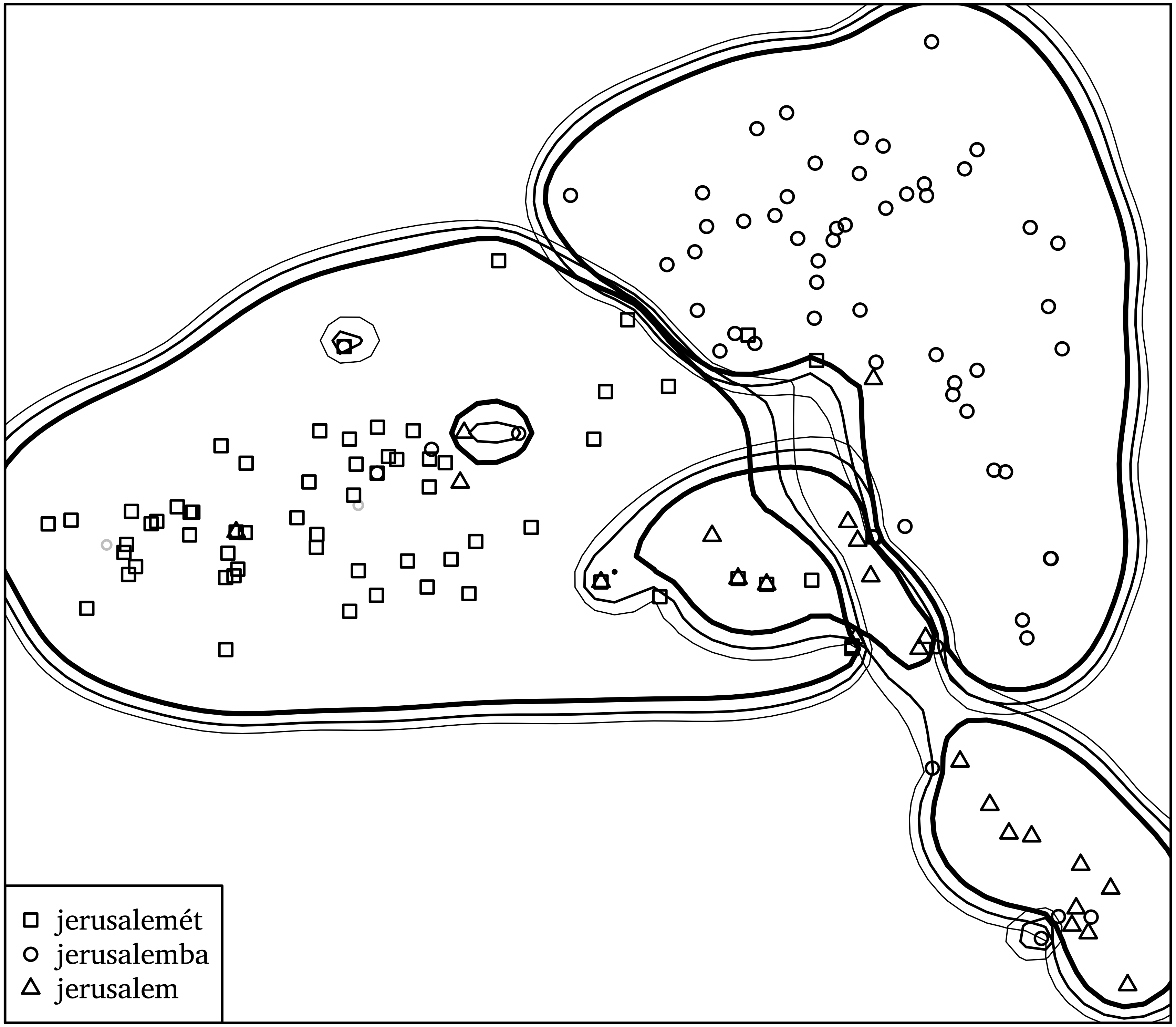
Amele (A language of Papua New Guinea)

aai



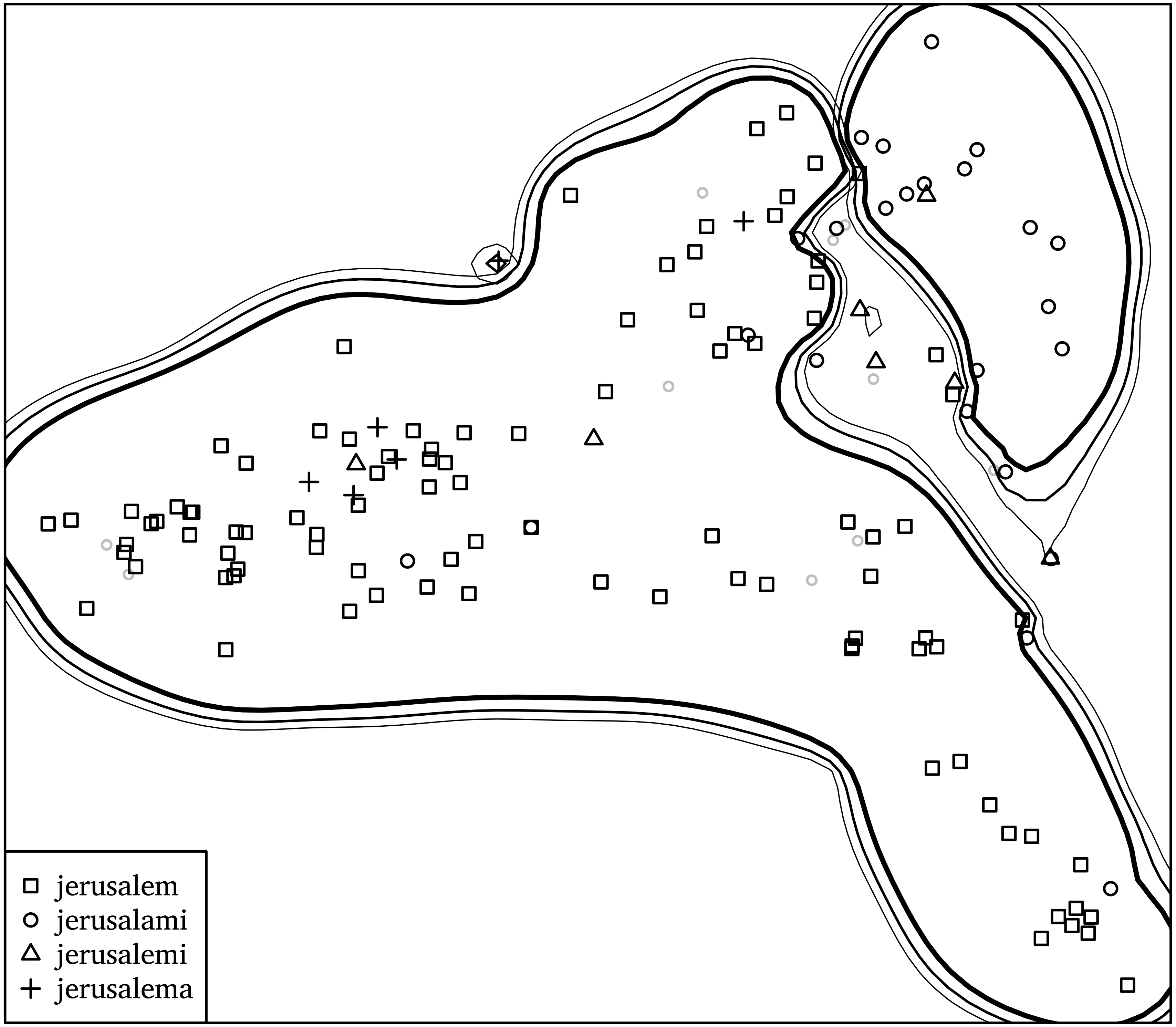
Arifama-Miniafia (a language of Papua New Guinea)

abt



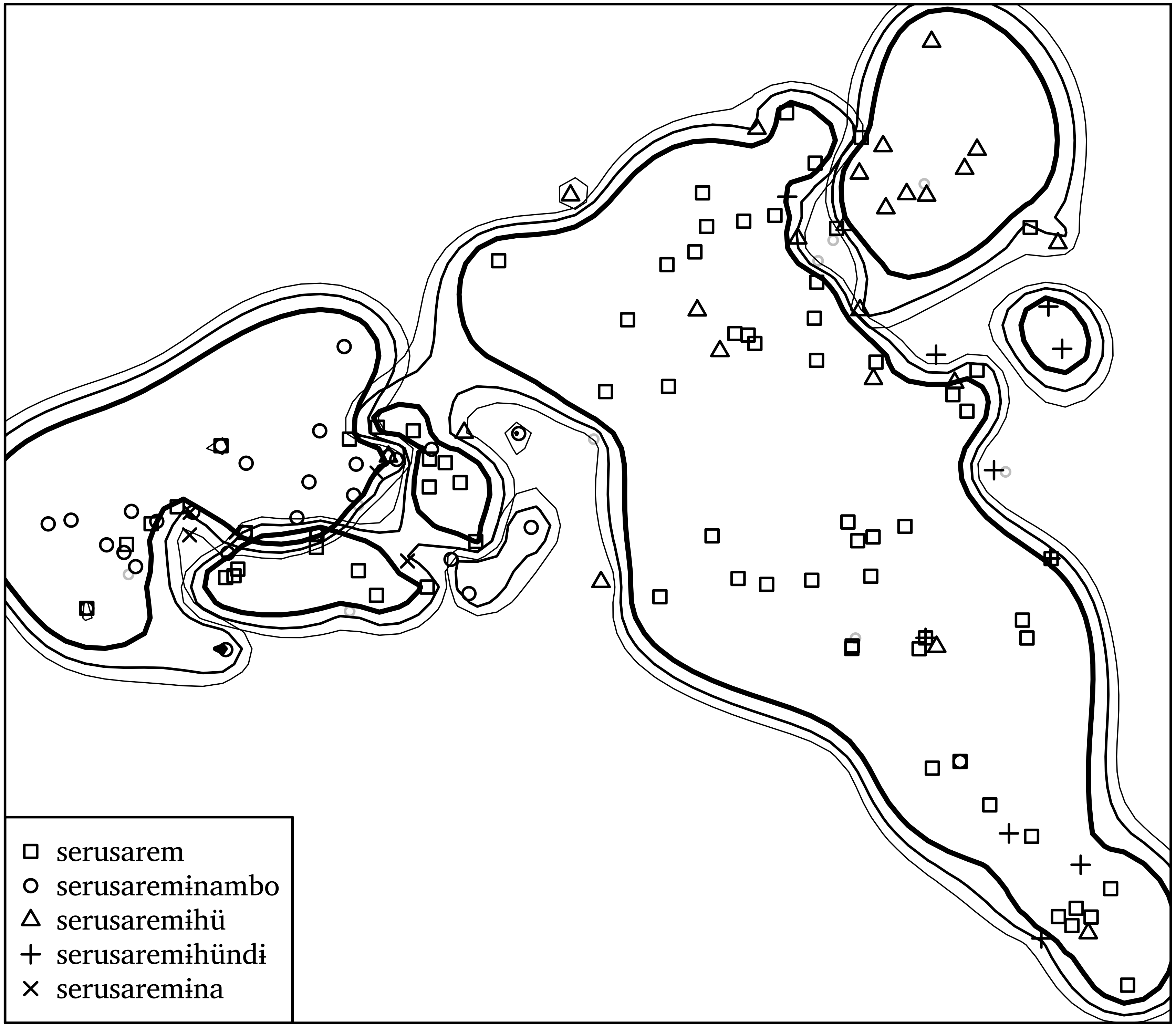
Ambulas (a language of Papua New Guinea)

aoj



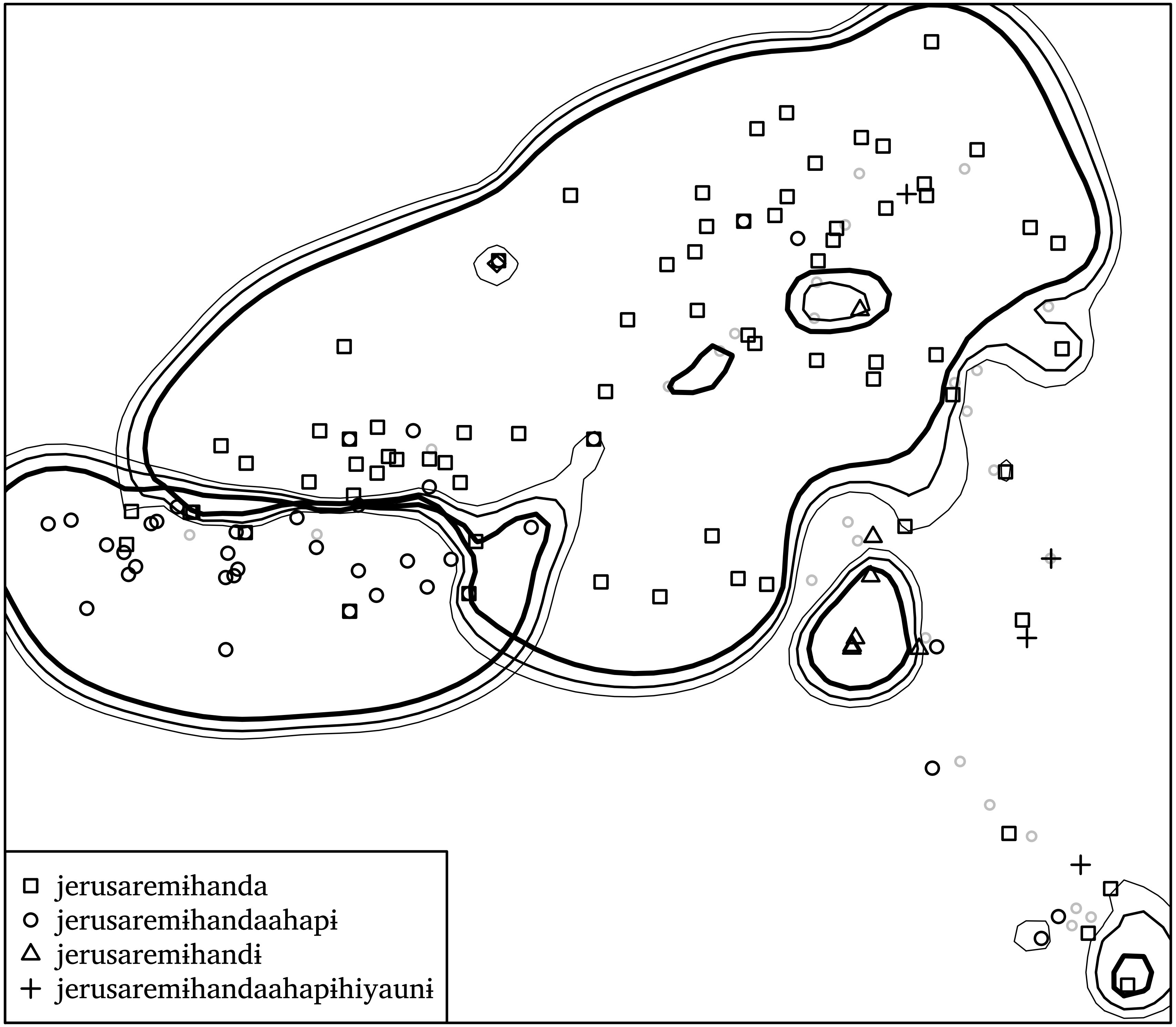
Muffian (a language of Papua New Guinea)

agg



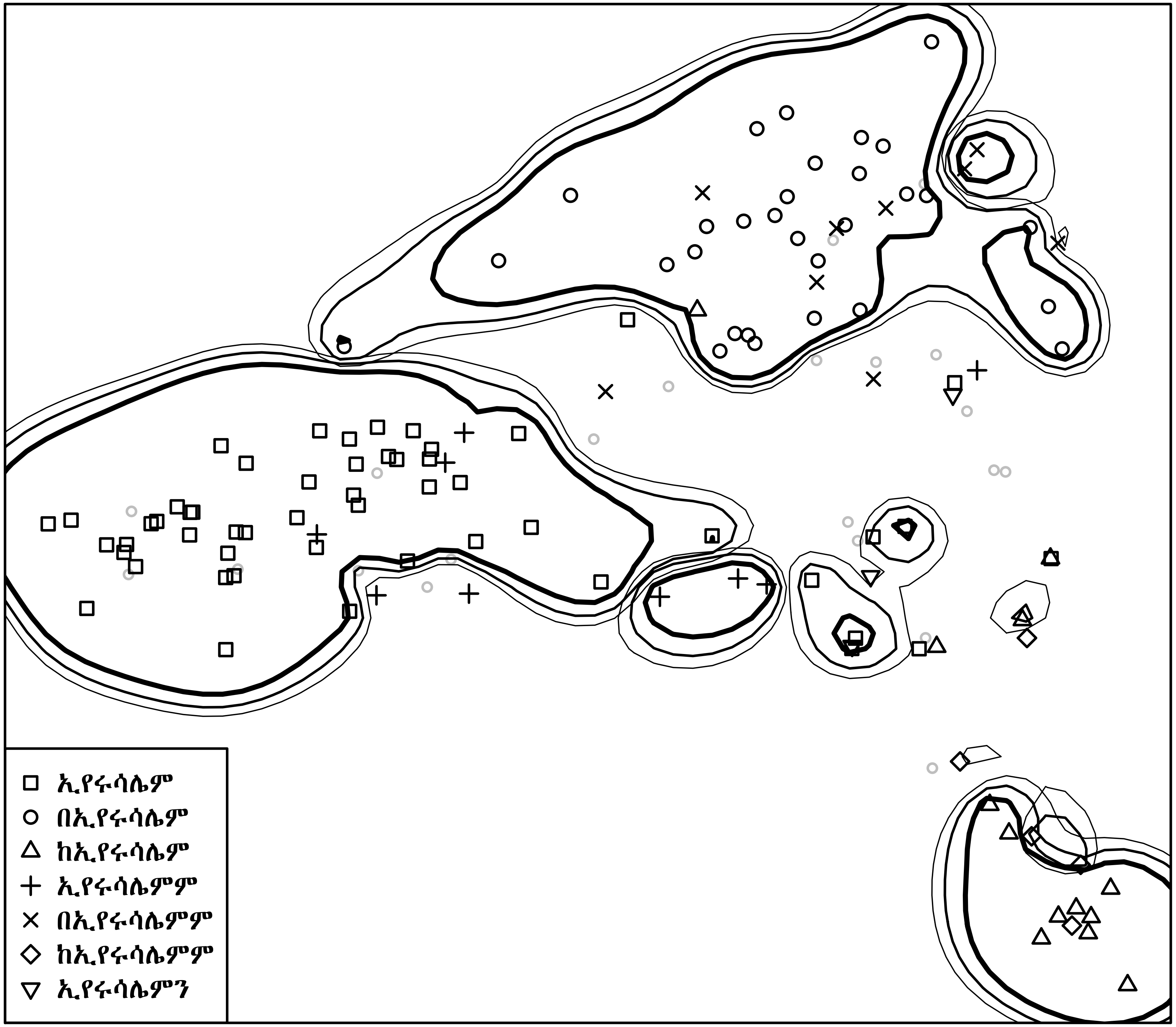
Angor (a language of Papua New Guinea)

agm



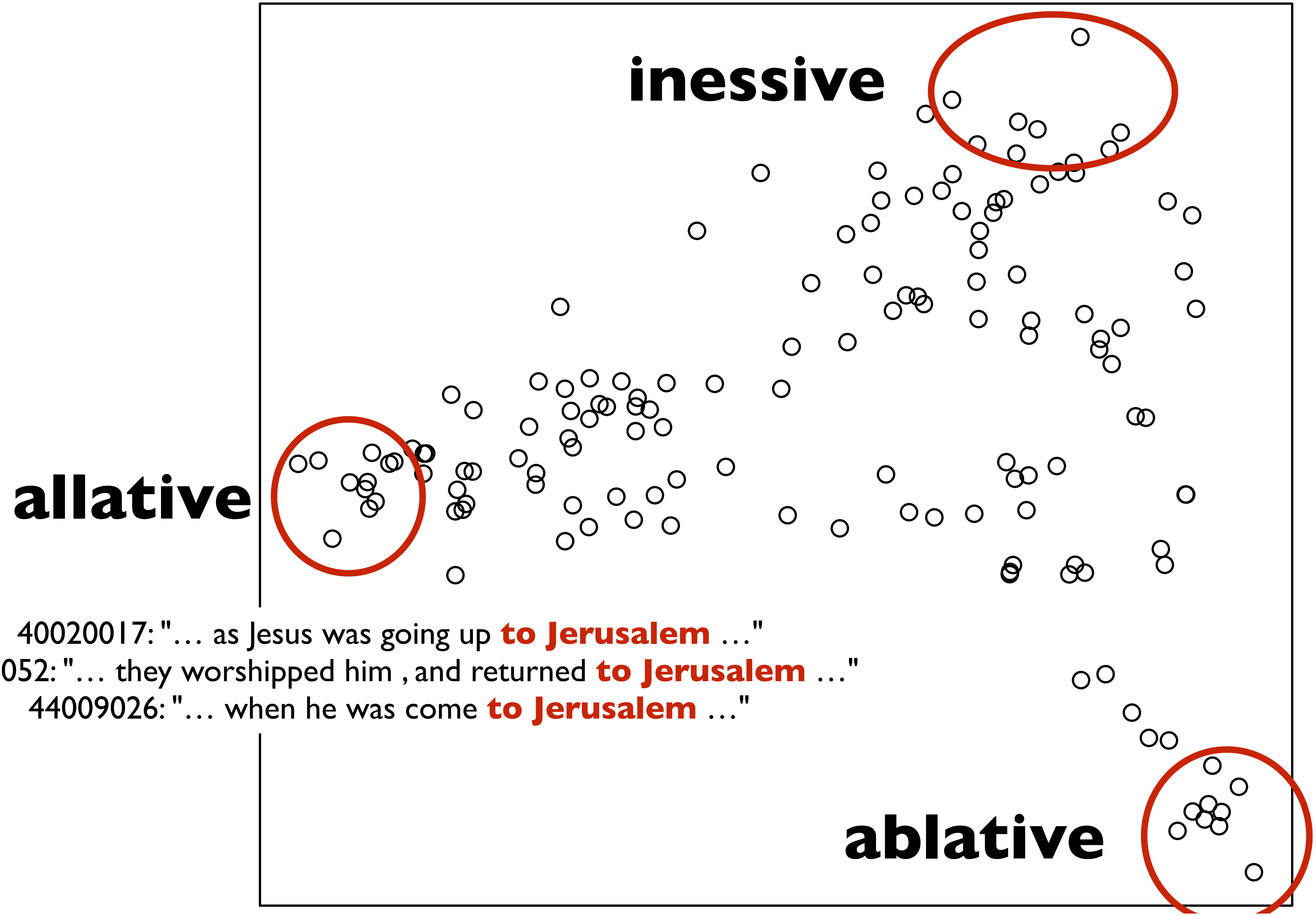
Angaatiha (a language of Papua New Guinea)

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 …"

Moving Forward

- Language diversity is much greater than often assumed
- Diversity is both a curse and a blessing
- Automatic analysis of each language individually is still unsatisfactory
- How can we better treat ‘exotic’ languages without Eurocentric bias ?