Induction to the max

Michael Cysouw
Philipps-University Marburg

Introducing the Parallel Text Corpus

Michael Cysouw
Philipps-University Marburg

Basic Problem of Language Comparison

- How to compare like with like
- Domain, Tertium Comparationis, Comparative Concept, Function, etc.
- Even better: **Contextually Situated Exemplars**
- Stimuli-based elicitation, translational equivalence

Parallel Bible Corpus

- 1169 translations
- 906 different ISO-639/3 codes
- In total more than 350 Million wordforms
- More than 17 Million different wordforms
- http://paralleltext.info/data

Demo

Software

- Contact me personally for access to the data
- R-package “qlcMatrix”
  http://cran.r-project.org/web/packages/qlcMatrix/index.html
  https://github.com/cysouw/qlcMatrix
- Python library
  https://github.com/tmayer/paralleltextprocessing

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

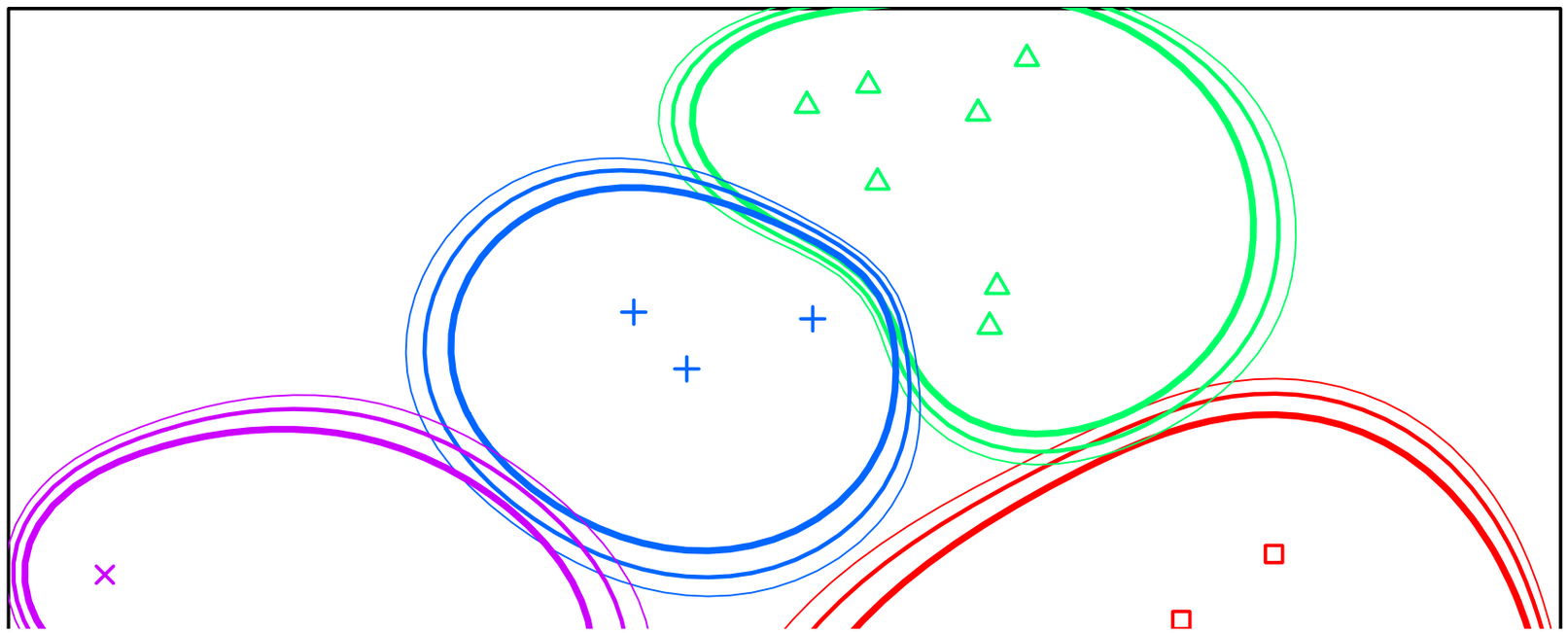
- Small-scale experiment: use fastalign for bitext-alignment on all pairs, build multi-text-alignment from there

- Only for 77 Germanic translations

- New Testament produced almost 100.000 Germanic alignments, which are directly comparable ‘words’

trees and wood

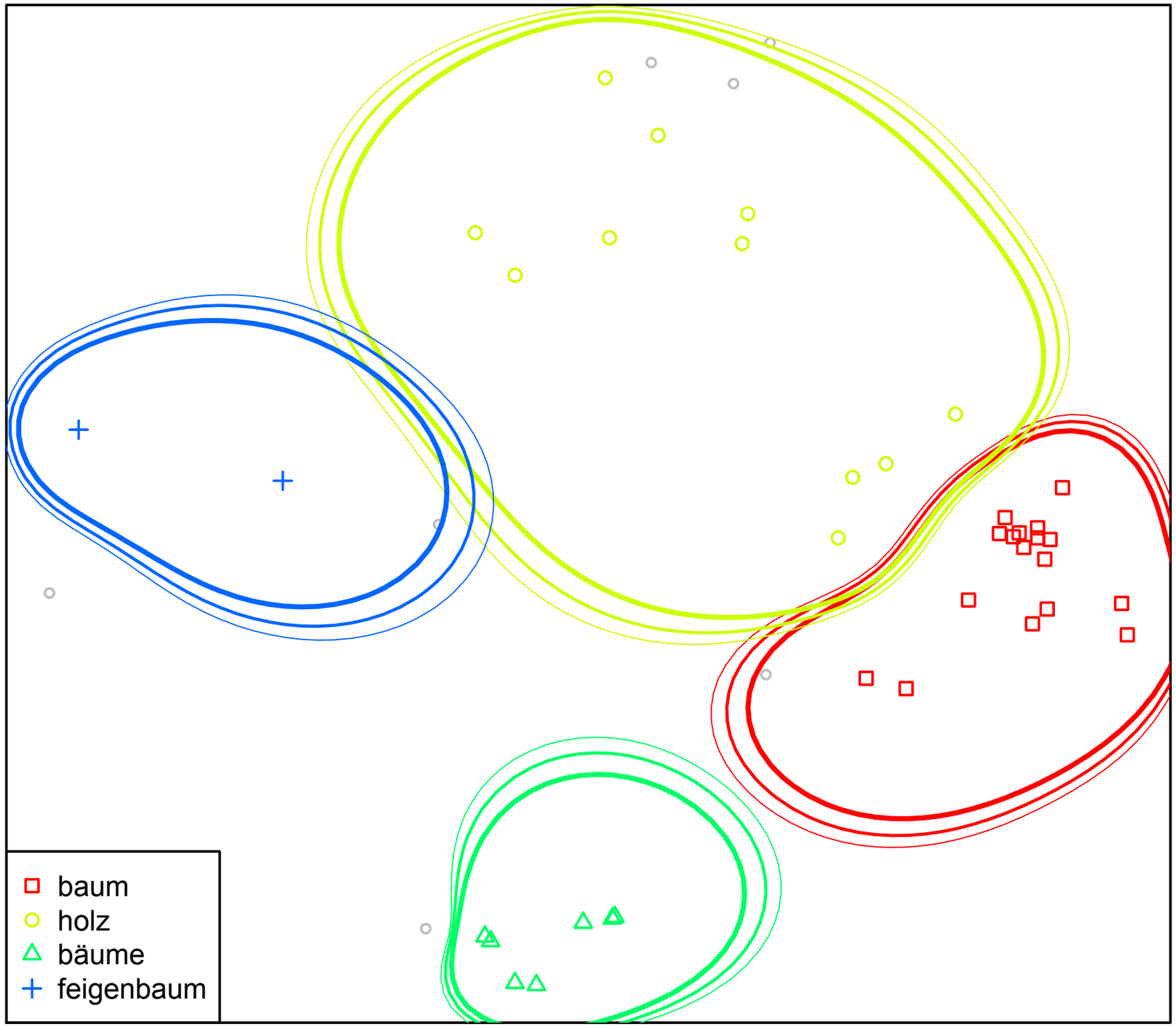
afr-x-bible-1953.txt



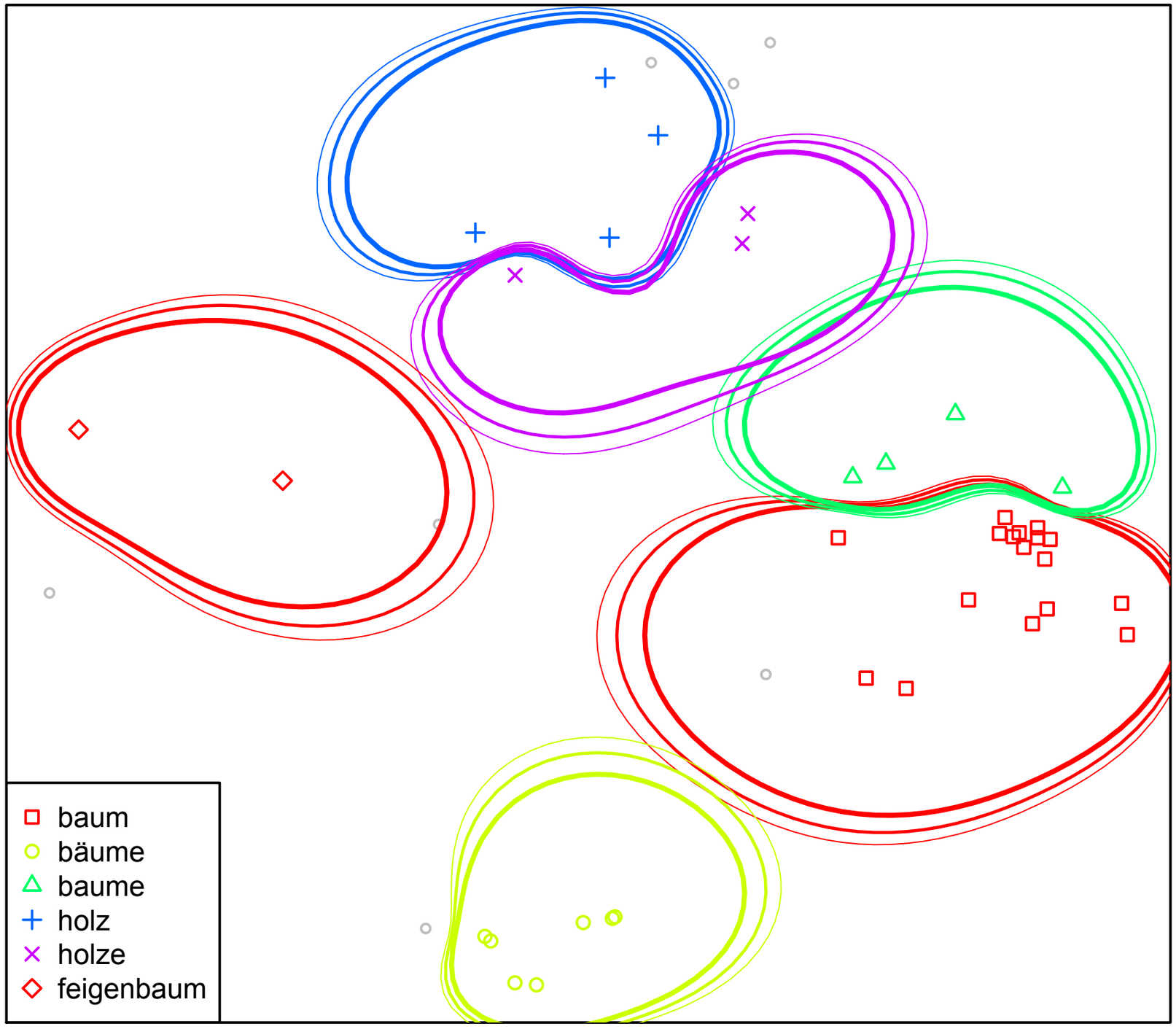
	tree	wood (stuff)	firewood	small forest	large forest
German	*Baum*	*Holz*		*Wald*	
Danish	*træ*			*skov*	
French	*arbre*	*bois*			*forêt*
Spanish	*árbol*	*madera*	*leña*	*bosque*	*selva*

Louis Hjelmslev
Prolegomena to a Theory of Language (1963)

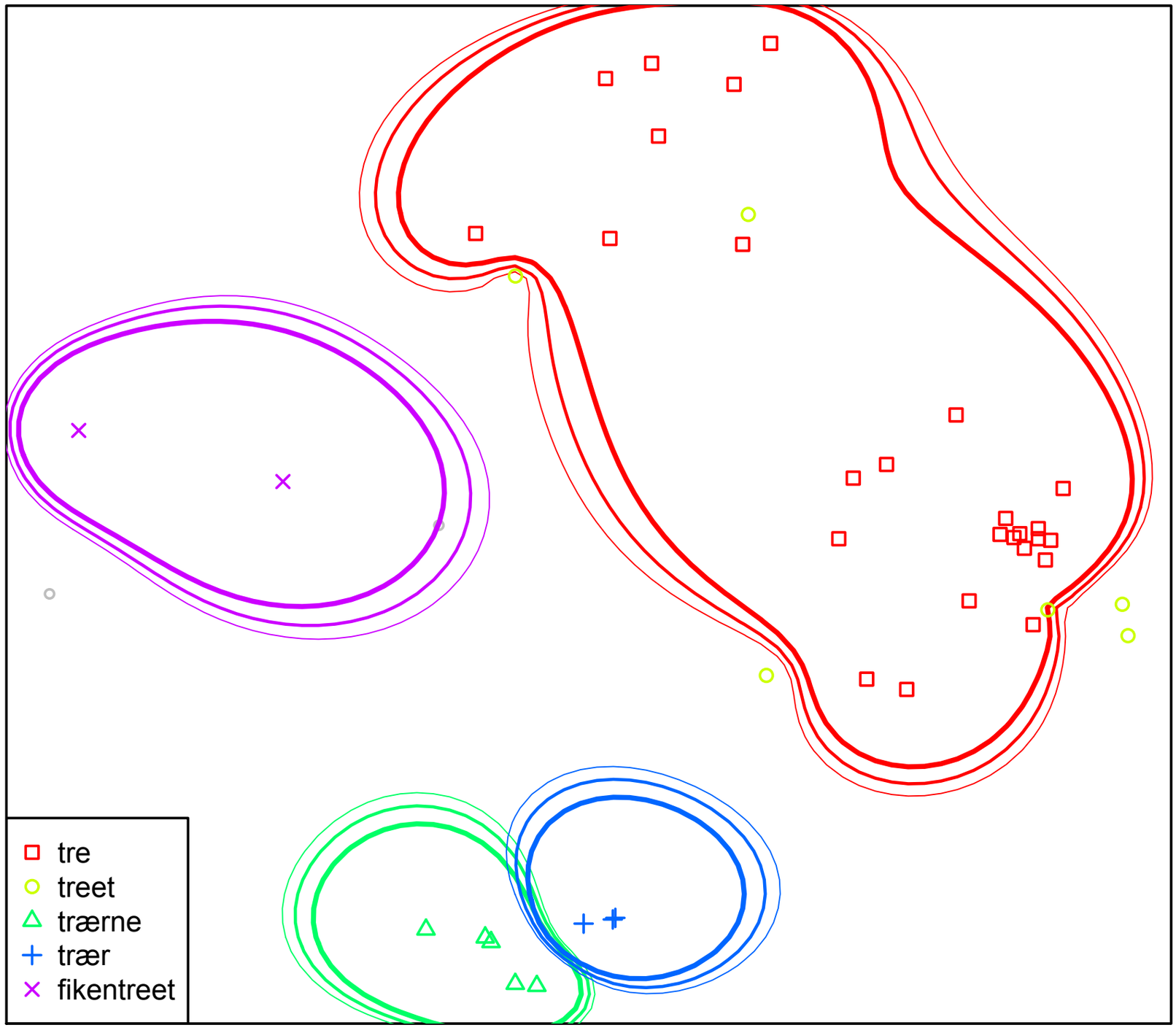
deu-x-bible-erben.txt



deu-x-bible-freebible.txt

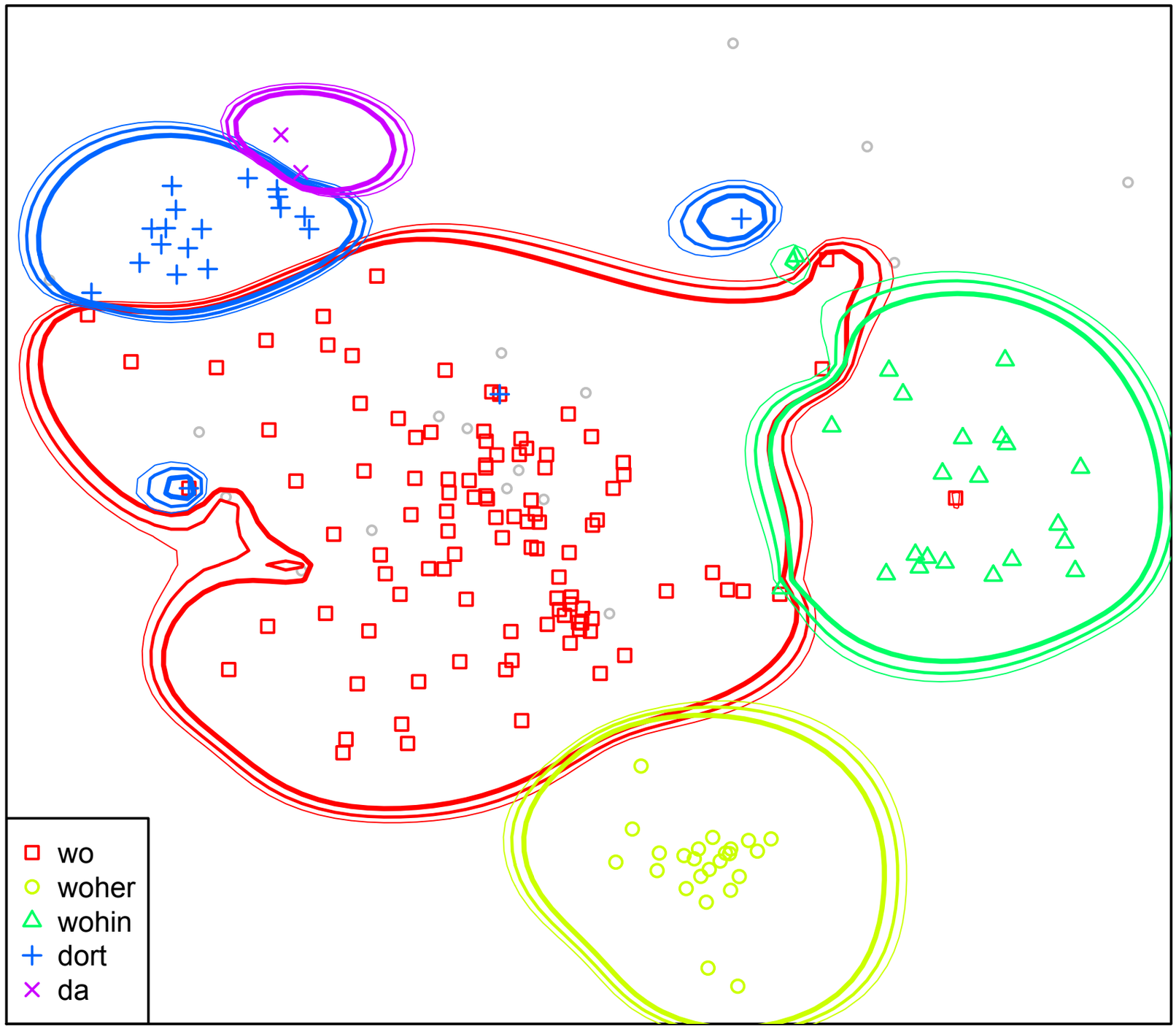


nob-x-bible-2007.txt

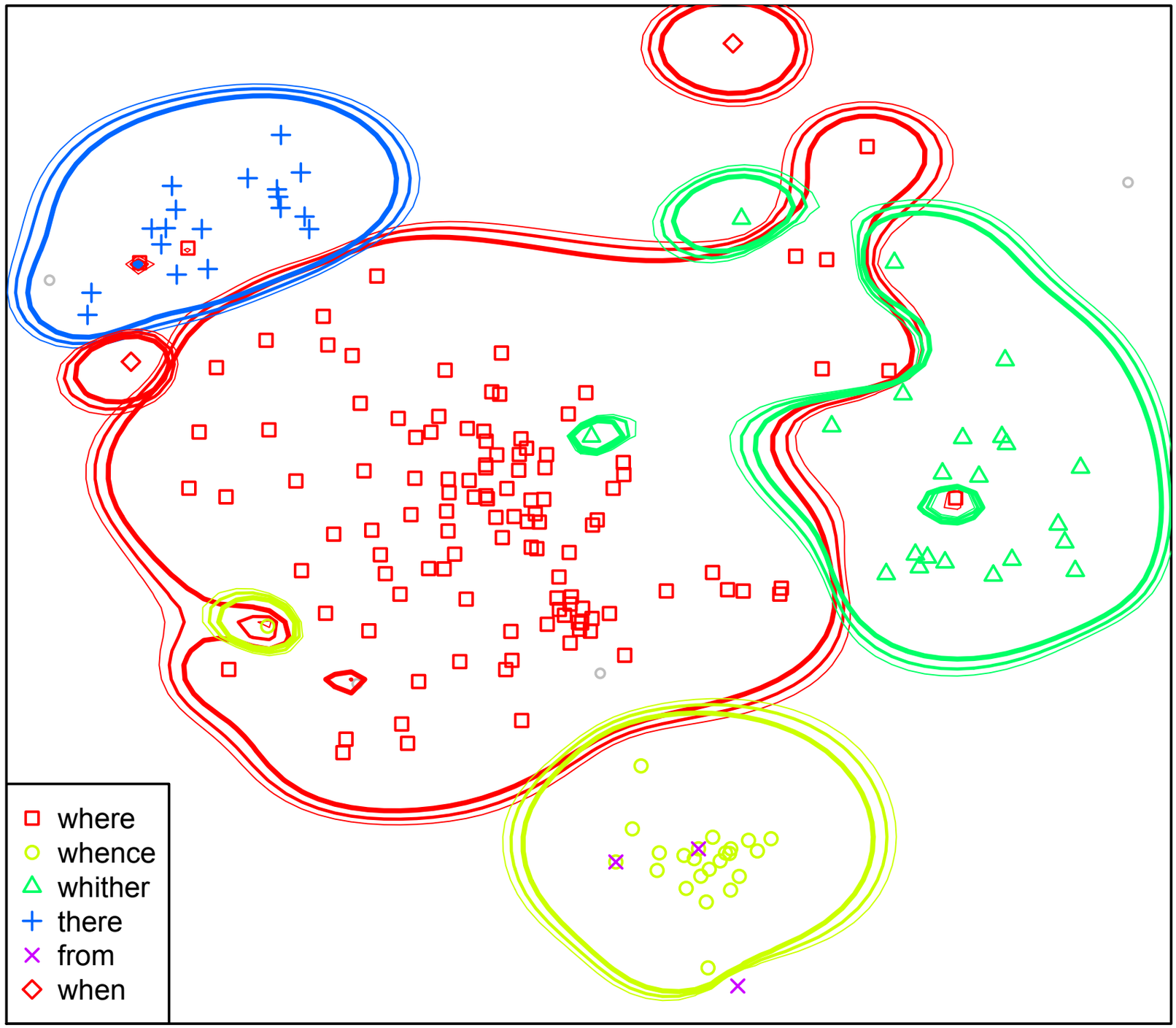


where

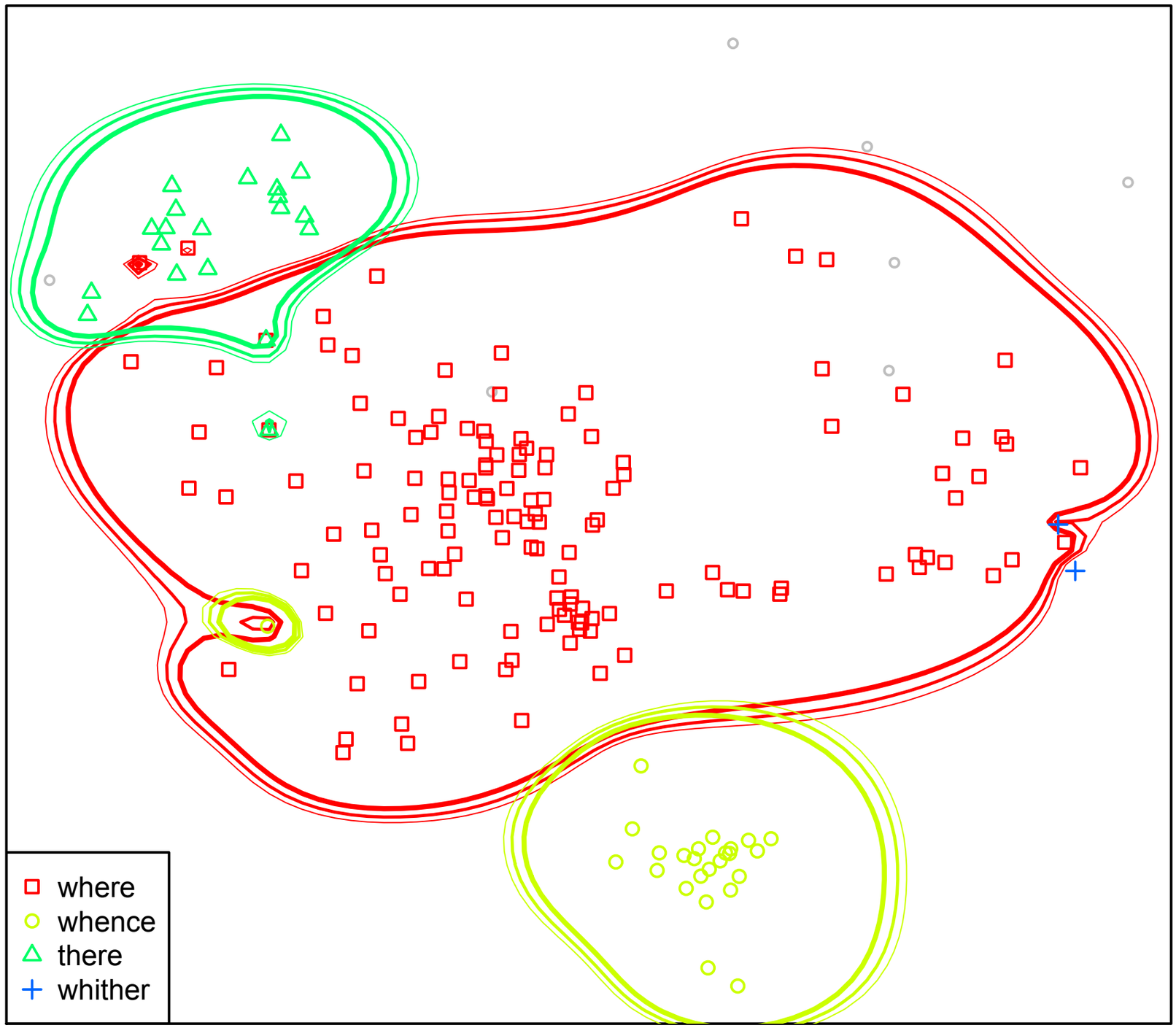
deu-x-bible-pattloch.txt



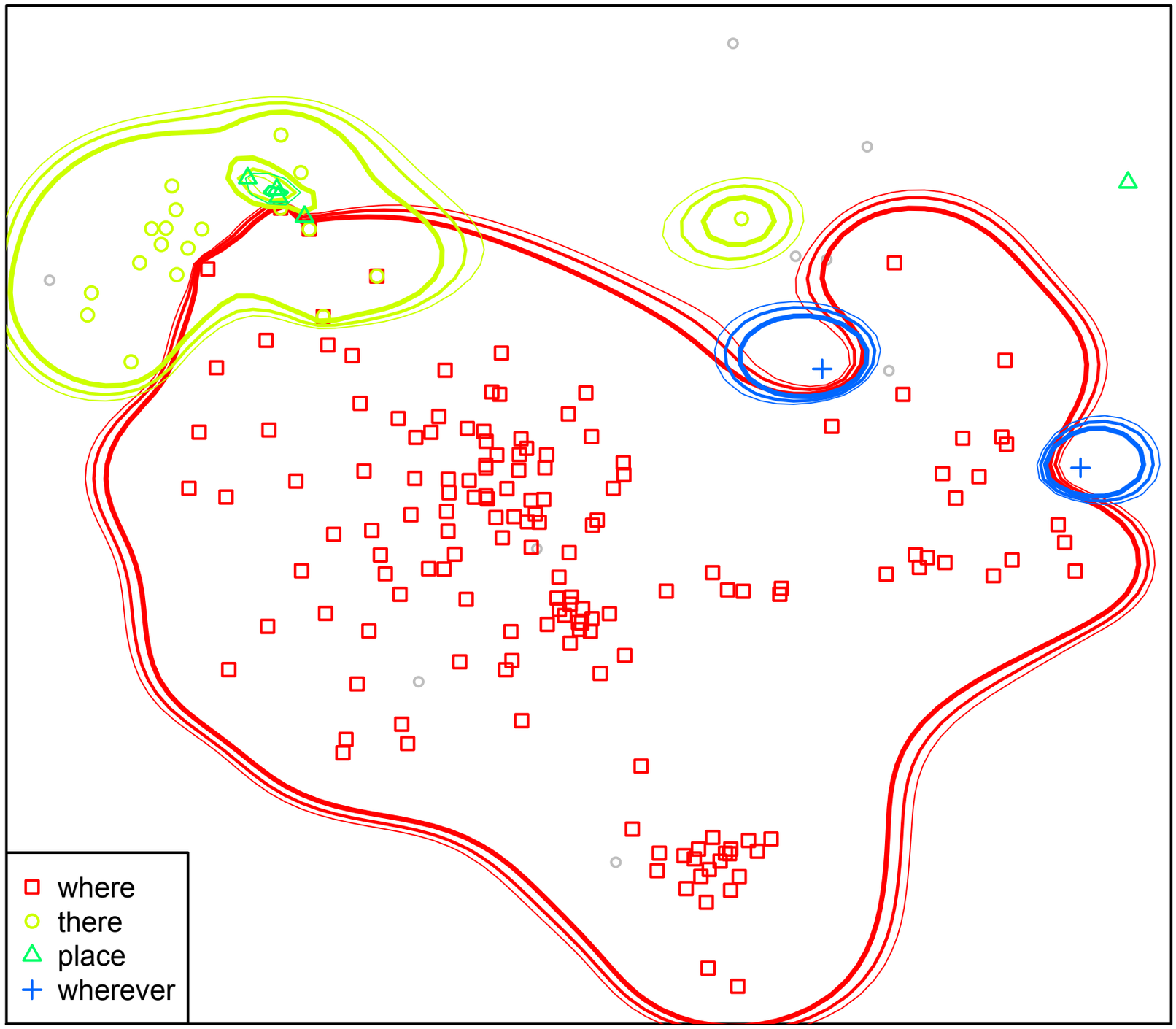
eng-x-bible-kingjames.txt



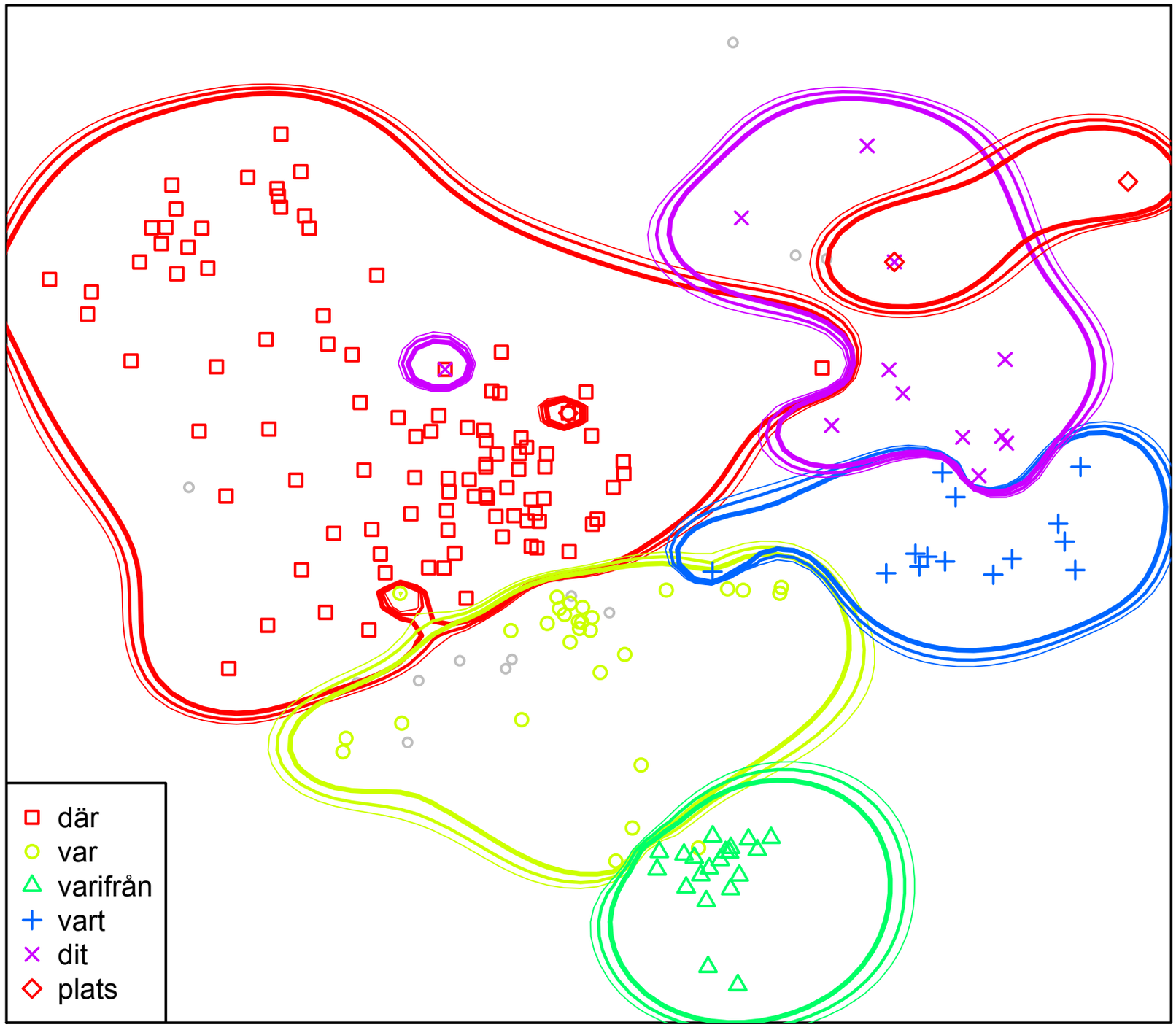
eng-x-bible-darby.txt



eng-x-bible-treeoflife.txt

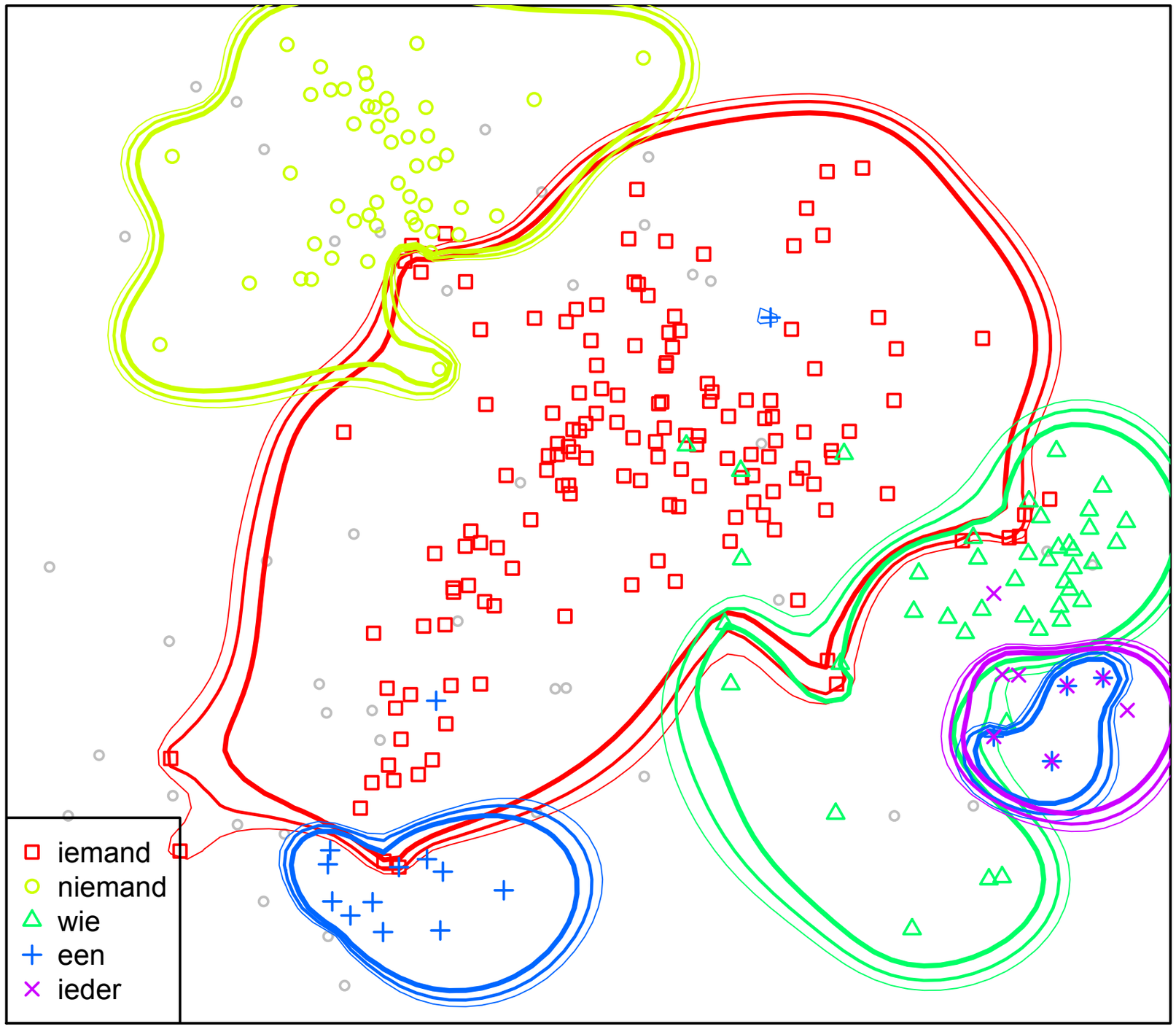


swe-x-bible-folk1998.txt

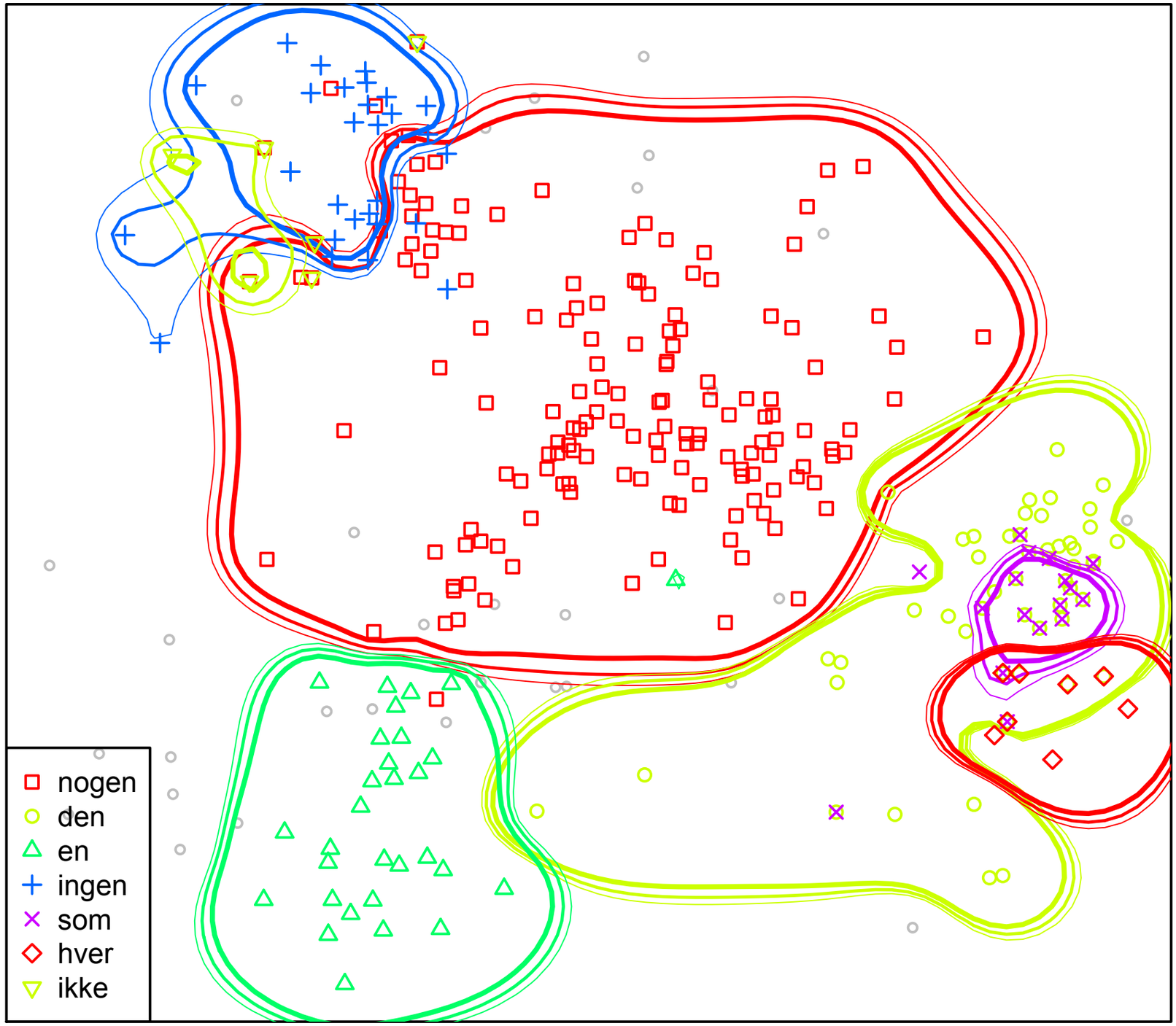


Indefinite person (someone, anyone)

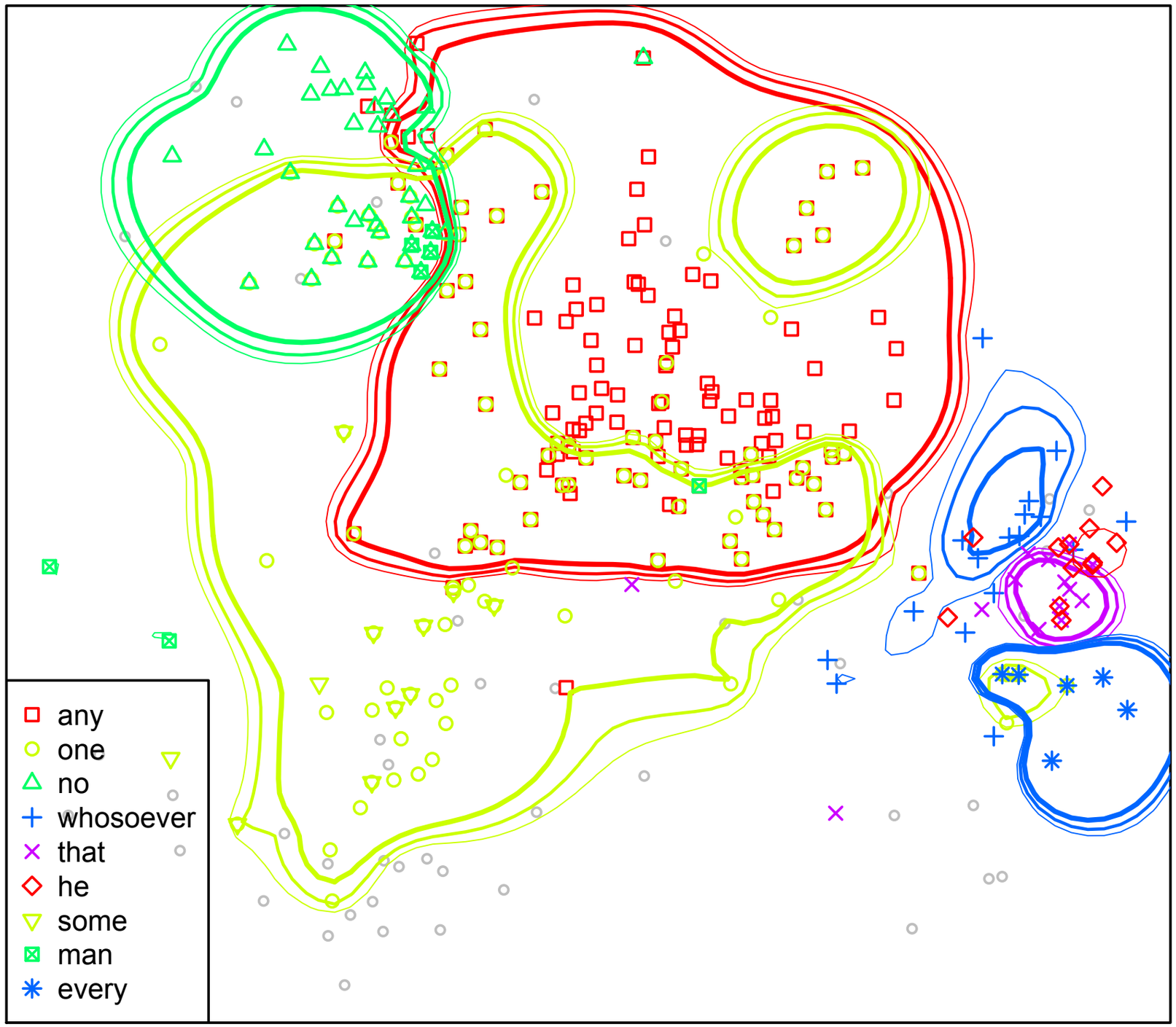
nld-x-bible-1951.txt



dan-x-bible-1931.txt



eng-x-bible-darby.txt



Lexical comparison

Verse 43019036 (John 19:36):
These things happened so that the scripture would be fulfilled : " Not one of his bones will be broken , "

been, ben, been, bein, bein, bein, knochen, bein, knochen, bein, bein, bein, knochen, knochen, knochen, bein, knochen, bein, bein, knochen, bones, bones, bones, bones, bone, bones, bones, bones, bones, bone, bone, bones, bones, bones, bones, bones, bones, bones, bone, bones, bones, been, beenderen, botten, been, bein, ben, ben, ben, bein, ben, bein, ben, ben, ben

Different contexts give different cognates !

	Matthew 1:16	Matthew 1:19	Matthew 19:10	Mark 10:12
afr-x-bible-boodskap.txt	*man*	*man*	*mens*	*man*
afr-x-bible-1953.txt	*man*	*man*	*man*	*man*
deu-x-bible-volxbibel.txt	*mann*	*mann*		*ehemann*
eng-x-bible-common.txt	*husband*	*husband*	*man*	*man*
eng-x-bible-literal.txt	*husband*	*husband*	*husband*	*husband*

Conclusion

- Massively parallel texts are a goldmine for language comparison
- Much experimentation is needed to find suitable methods to enrich the data
- Collaboration welcome (using git-approach)