A Computational Approach to Identifying Macro-Areas in Africa

Harald Hammarstrom
Max Planck Institute for the Science of Human History
Jena

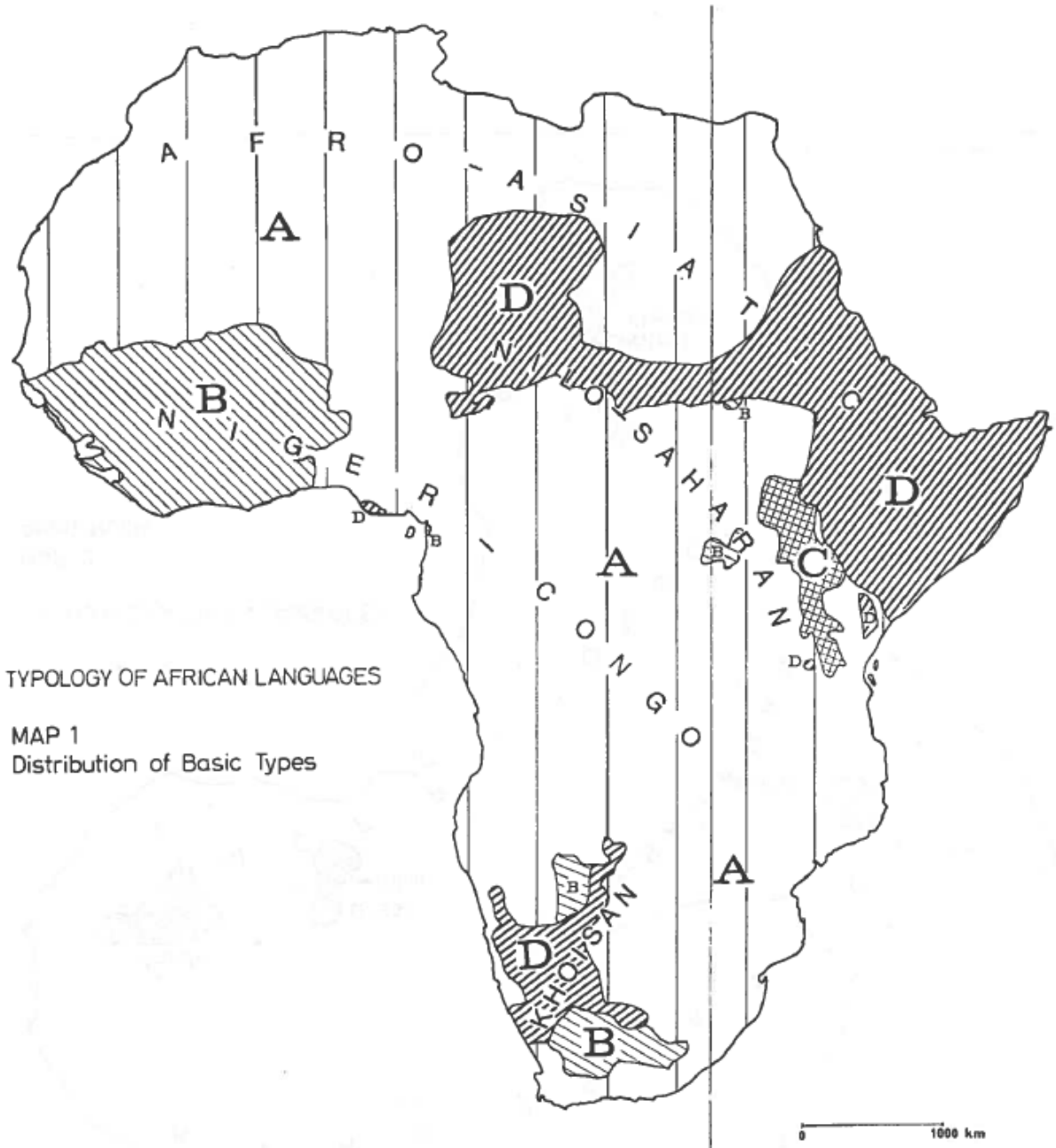
17 June 2016, Berlin

Areas of Typological Similarity in Africa

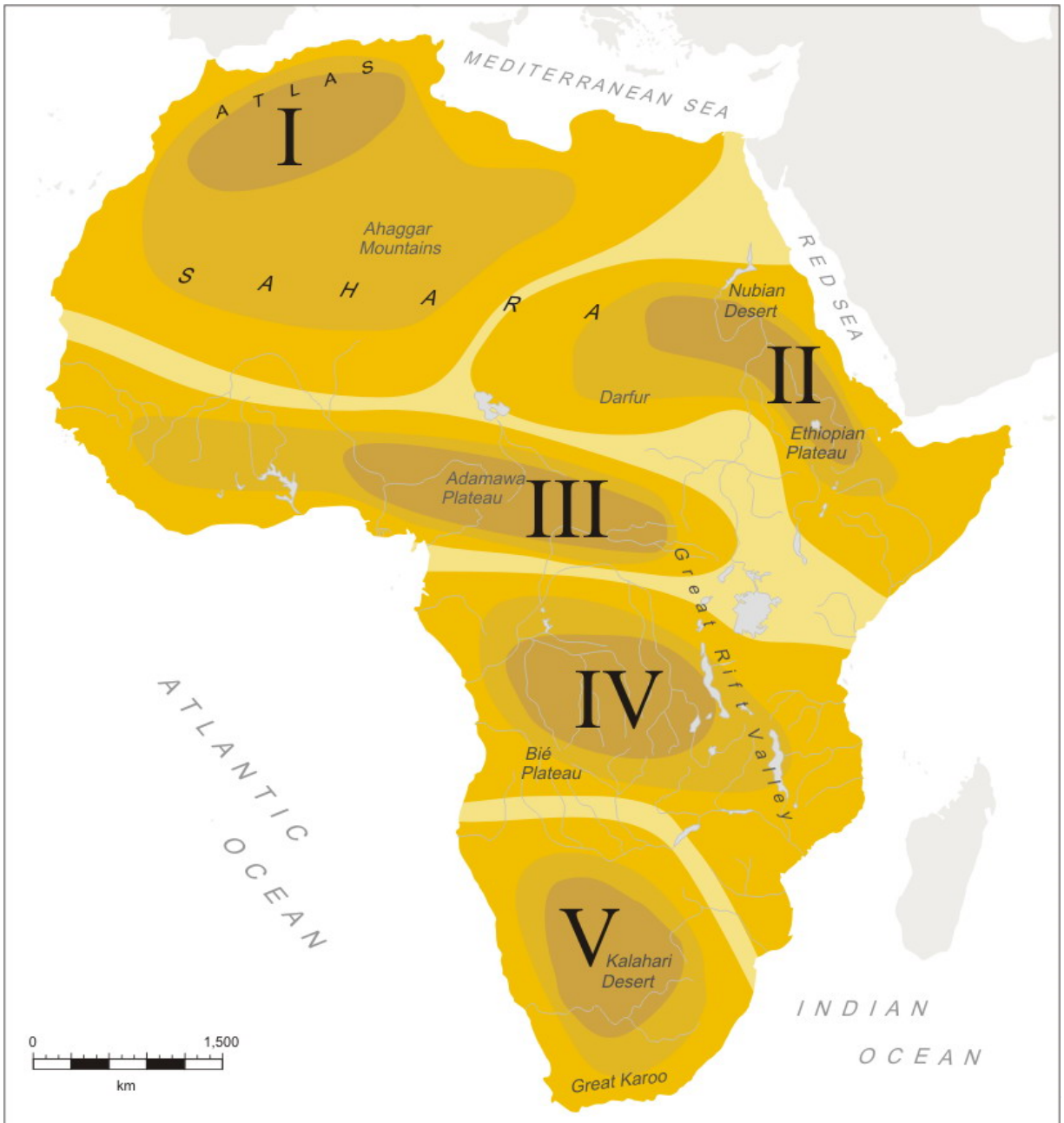
It has been observed since long that the languages of the African continent exhibit **typological similarities** *that are* **geographically conditioned**

- Some researchers have distilled large-scale convergence areas (e.g., Güldemann 2008, Heine 1976, 2011, Segerer 2015)
- These areas may reflect patterns of social interaction, ethnographic similarities, geographical conduits or barriers (mountains, rivers), language family expansions, . . .

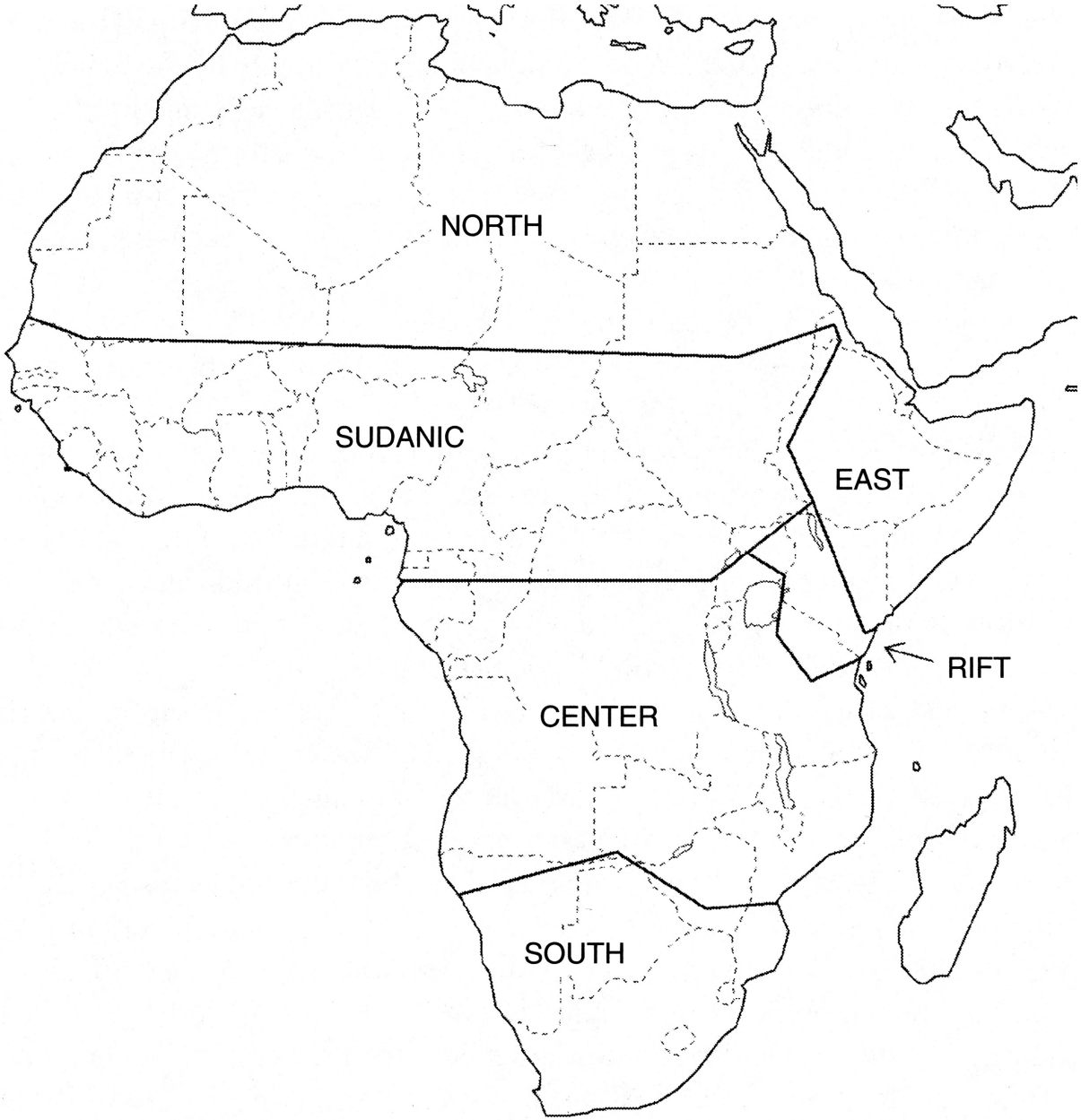
Heine 1976:90's Distribution of Basic Types



Güldemann 2010:576's African Macro-Areas



Clements and Rialland 2008:37's African Macro-Areas



Delimiting Macro-Areas

- Even the even the largest previous studies ultimately rely on **eyeball inspection** of features plotted on a map to define the areas
- Today we will compare with a computerized approach that

  Given geolocated language data as input delineates the area(s) with the greatest homogeneity

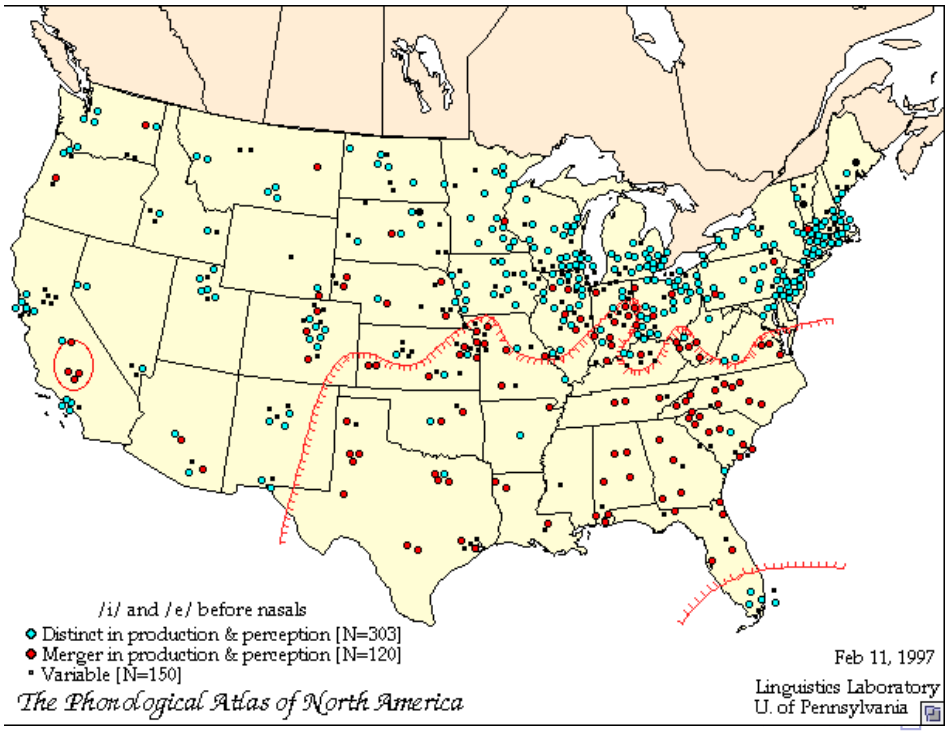
- Computational approaches have the advantage of containing no subjectivity, but, on the other hand, make some simplifying assumptions.
- Previous computational work typically searches for areal with some regularity in shape (circles, rectangles, size) and tests for geographical coherence (Daumé 2009, Michael et al. 2014, Muysken et al. 2015)
- Probably, a closer approximation of what humans are doing are captured by series of isogloss lines

Drawing Isogloss Lines

An isogloss is the geographical boundary of a certain linguistic feature, . . . such as the pronunciation of a vowel, the meaning of a word, or use of some syntactic feature (Wikipedia 8 June 2010)

- Widely used in dialectology
- Example, pin/pen merger as of Labov (1997):

`http://www.ling.upenn.edu/phono_atlas/maps/Map3.html`



Approaches to Isogloss Lines

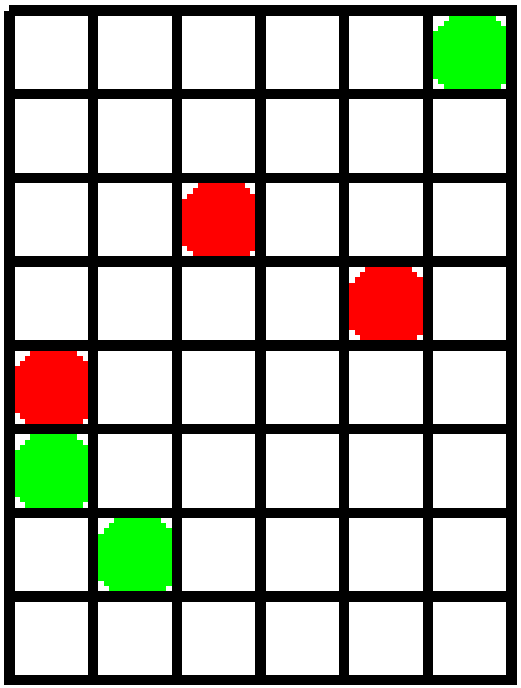
There appears to be no objective definition of an isogloss line, let alone an automated procedure for drawing one

- Dialectologists today draw isogloss lines by hand, based on intuition (p.c. Bert Vaux 2010)
- Today we will use a automated procedure which tries to maximize the homogeneity on either side of the line (Hammarström forthc.)

Problem Setting #1: Input

Given:

- 2D grid map with
- rings ("red") and crosses ("green") and empty positions

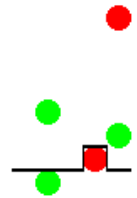


Problem Setting #2: "Line" Assumptions

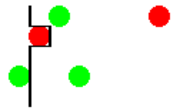
Assumptions about a "line":

- A line is not necessarily a straight line
- But, either
  - Runs from the west end to the east end on the map, crossing each column at exactly once OR
  - Runs from the north end to the south end on the map, crossing each row at exactly once

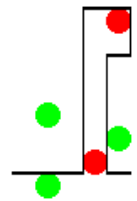
- Legal



- Legal



- NOT Legal



Definition of the Optimal Isogloss Line

Some straightforward alternatives

Absolute-Optimal The line that maximizes the total number of correctly classified points

Proportion-Optimal The line that maximizes the *proportion* of correctly classified points to the total number of points, on both sides

Homogeneity-Optimal The line that minimizes the weighted average *entropy* of the point distribution on either side (this is a generalization of proportion-optimality to non-binary maps)

Optimality: Example

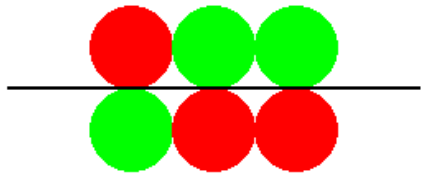
Absolute-Optimal: The max total number of correctly classified points

i) $2 + 2 = 4$ ii) $3 + 2 = 5$

i

Proportion-Optimal: The max *proportion* of correctly classified points to the total number of points, both sides

i) $2/3 + 2/3$ ii) $3/3 + 2/3$

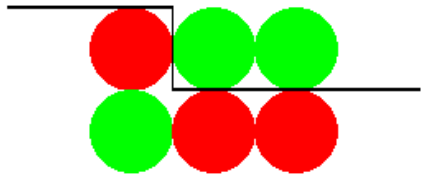


Homogeneity-Optimal: The minimal weighted average *entropy* of the point distribution on either side

ii

i $3 \cdot H(\frac{2}{3}, \frac{1}{3}) + 3 \cdot H(\frac{2}{3}, \frac{1}{3}) = 2.754 + 2.754 = 5.51$

ii $4 \cdot H(\frac{3}{4}, \frac{1}{4}) + 2 \cdot H(\frac{2}{2}) = 3.243 + 0.0 = 3.25$

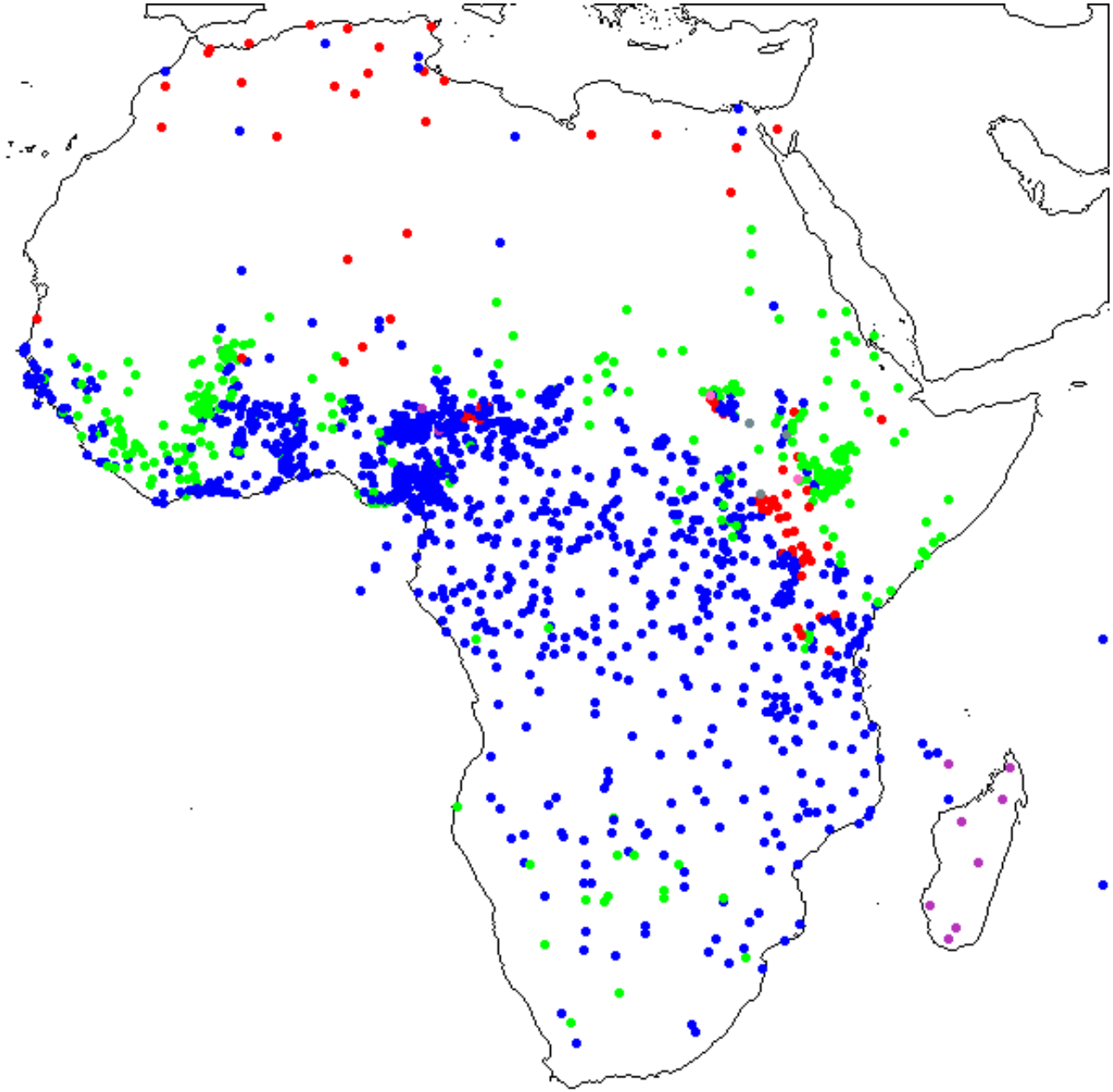


Line (ii) is better in all three cases of this example

Data: Three Databases

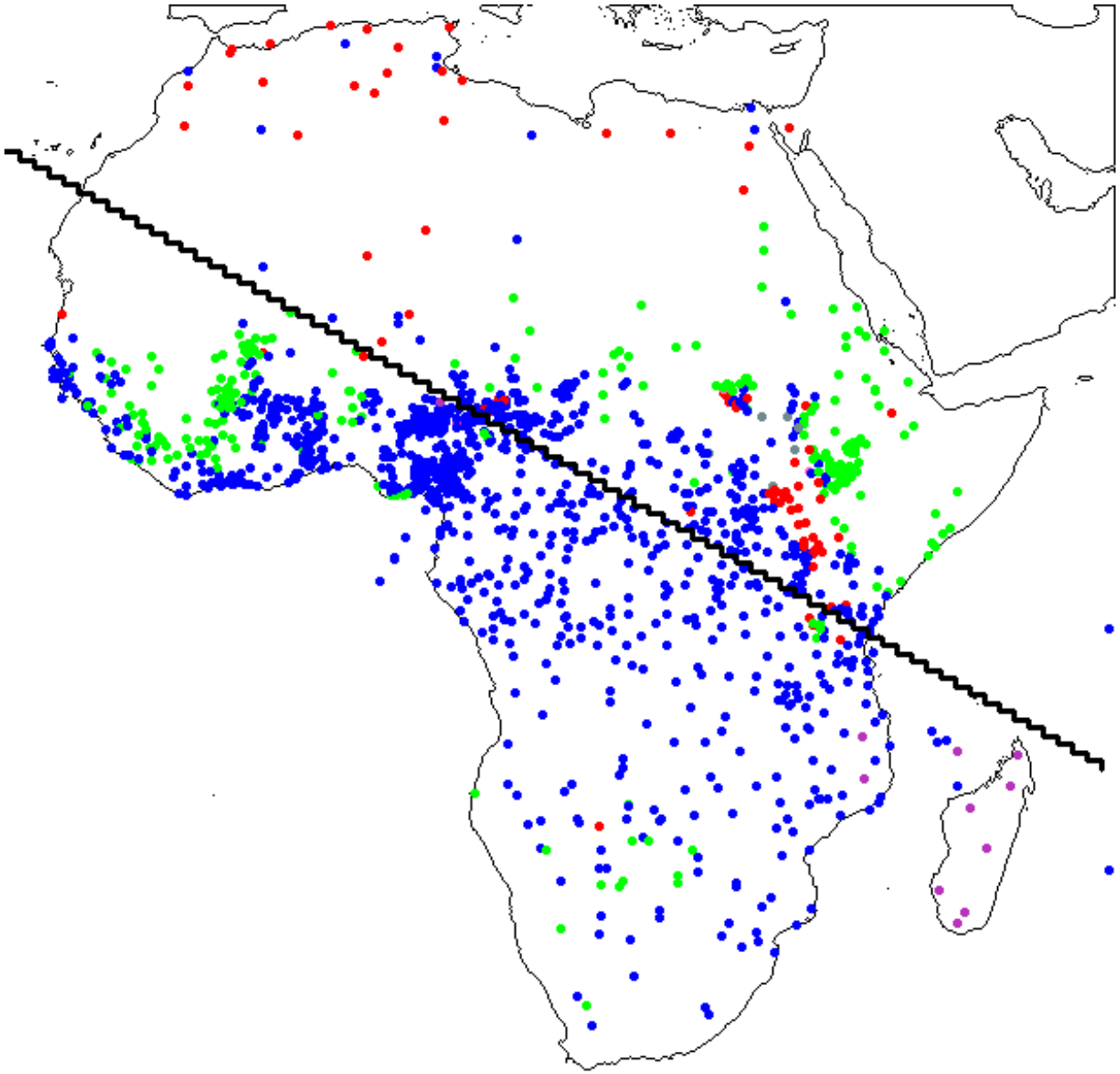
- **Constituent Order:** Basic constituent order in the transitive clause for 1431 spoken African languages (Own Database 2016)
- **Phonology:** Segmental inventories from 706 spoken African languages (Moran et al. 2015)
- **Morphosyntax:** 202 features from morphosyntax for 201 spoken African lanuages (Database developed at SHH Jena)

Constituent Order



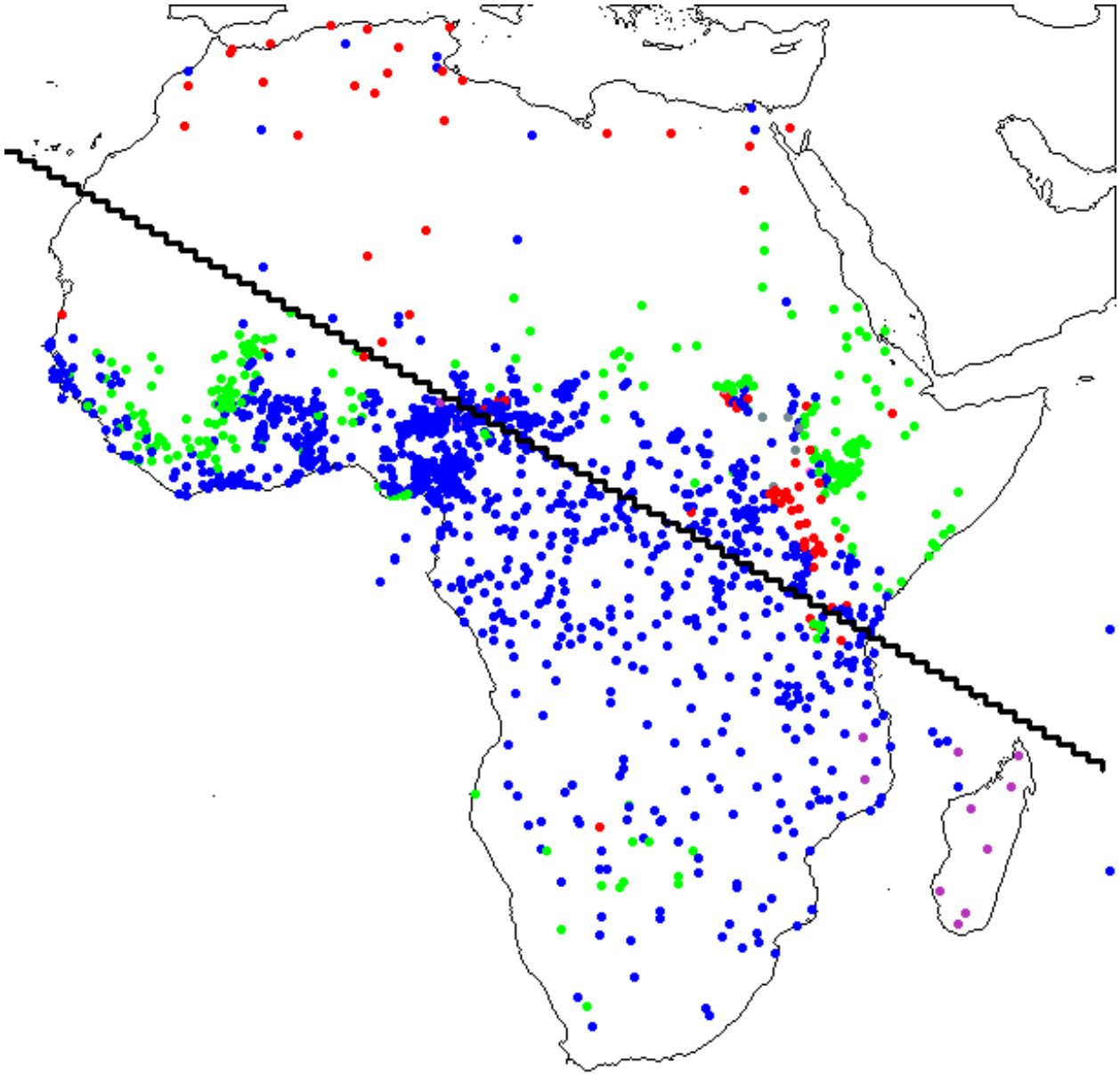
SVO	blue
VSO	red
OVS	slate gray
SOV	green
VOS	purple
NODOM	pink

Constituent Order: Example Line



- Suppose we draw an arbitrary line

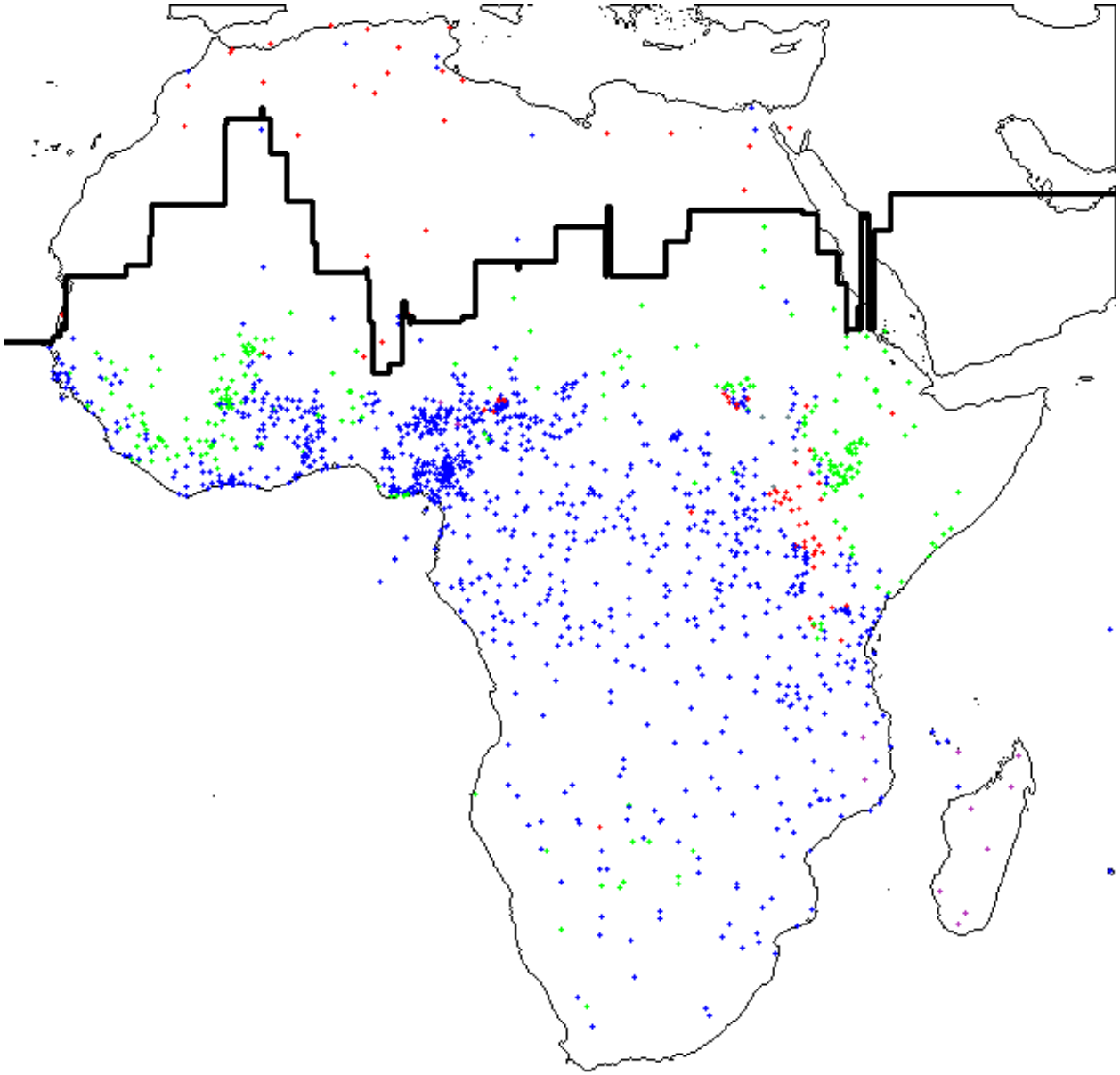
Constituent Order: Example Line



- Suppose we draw an arbitrary line
- Its homogeneity is 1721.3

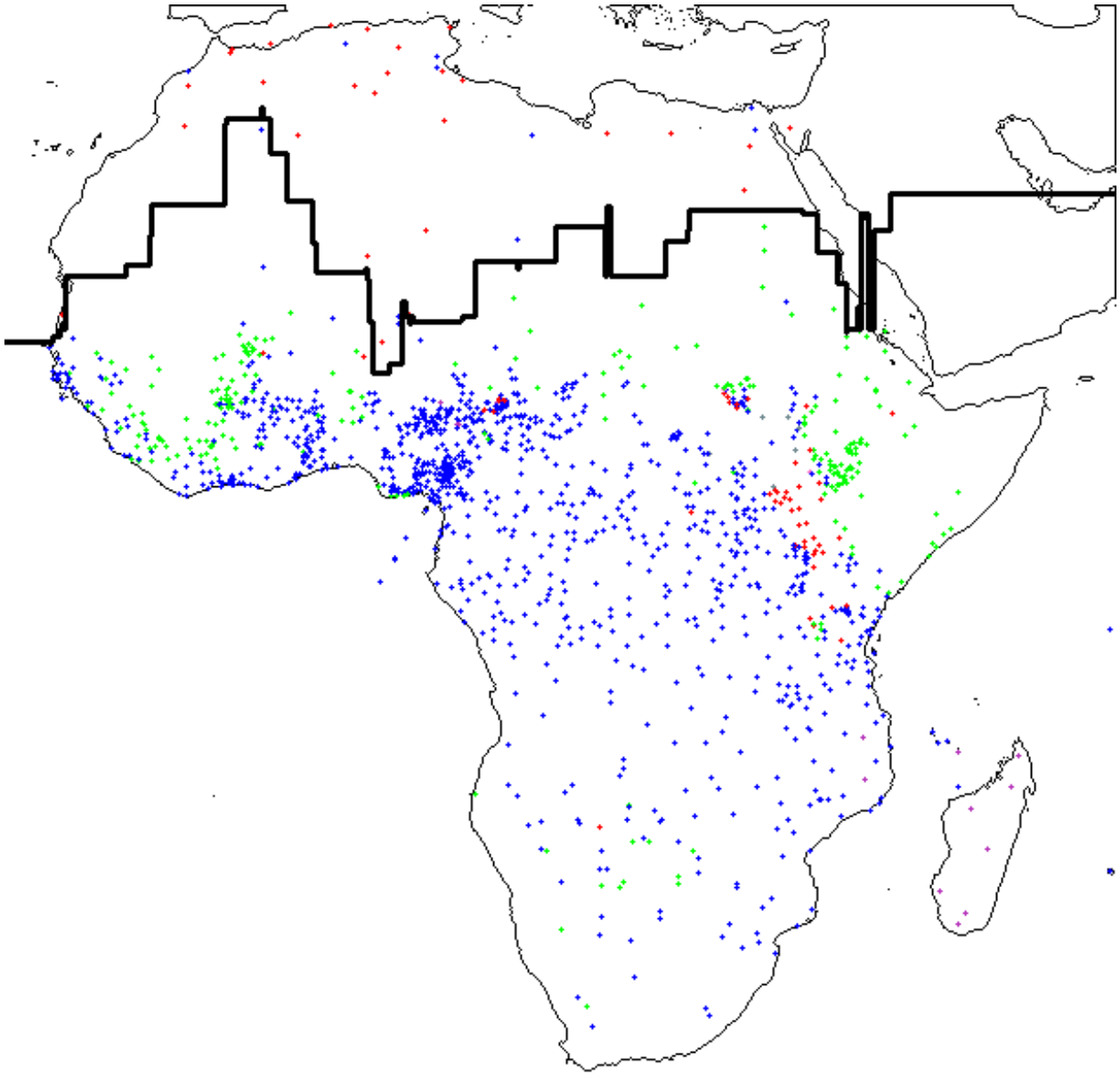
	Under	Over
SVO	663	286
SOV	177	244
VSO	7	74
VOS	13	1
OVS	1	6
NODOM	0	2
H	0.92	1.51
#	861	613

Constituent Order: East-West Line



- Suppose I let the computer find the *optimal* east-west line

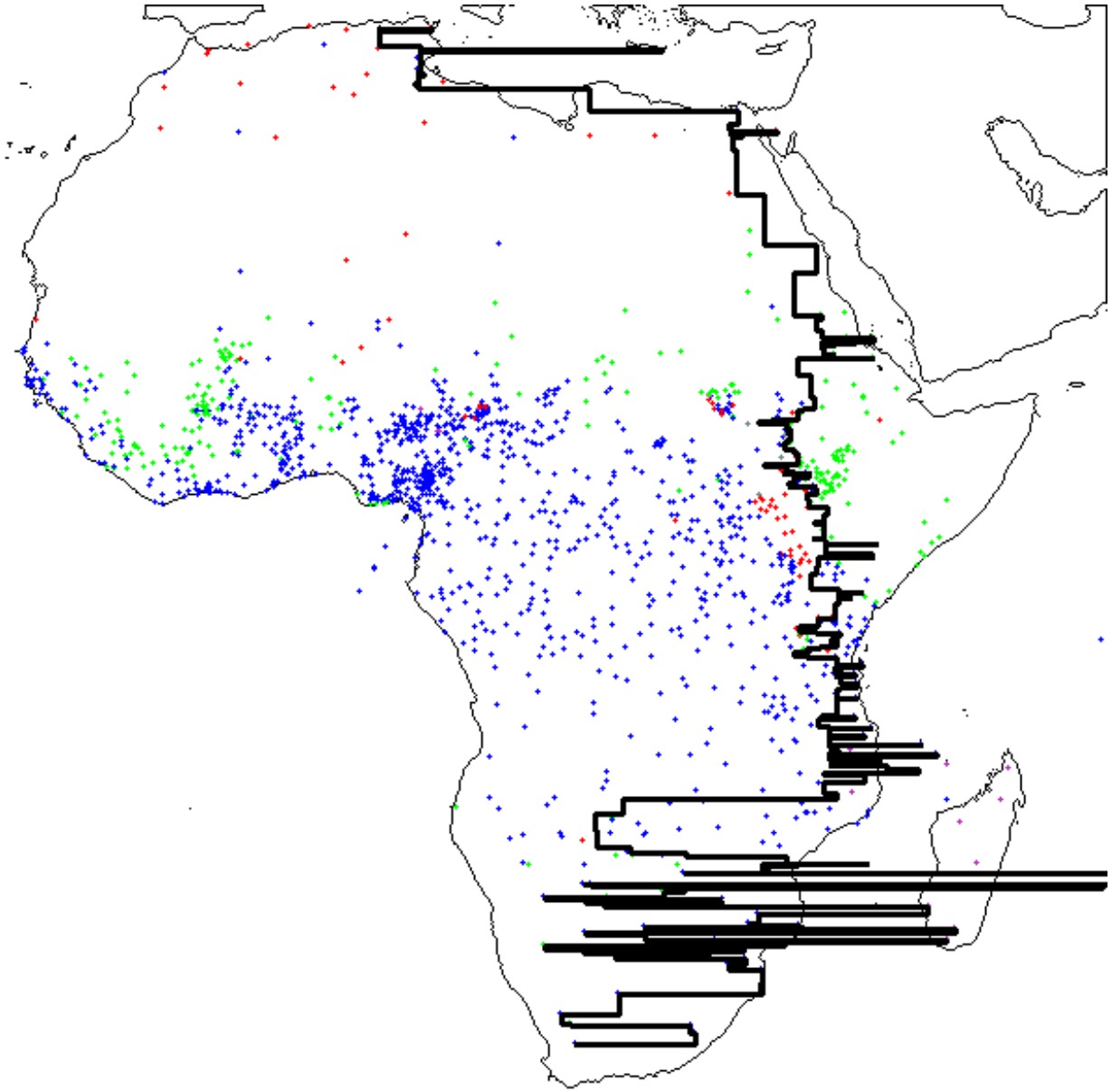
Constituent Order: East-West Line



- Suppose I let the computer find the *optimal* east-west line
- Its homogeneity is 1643.4

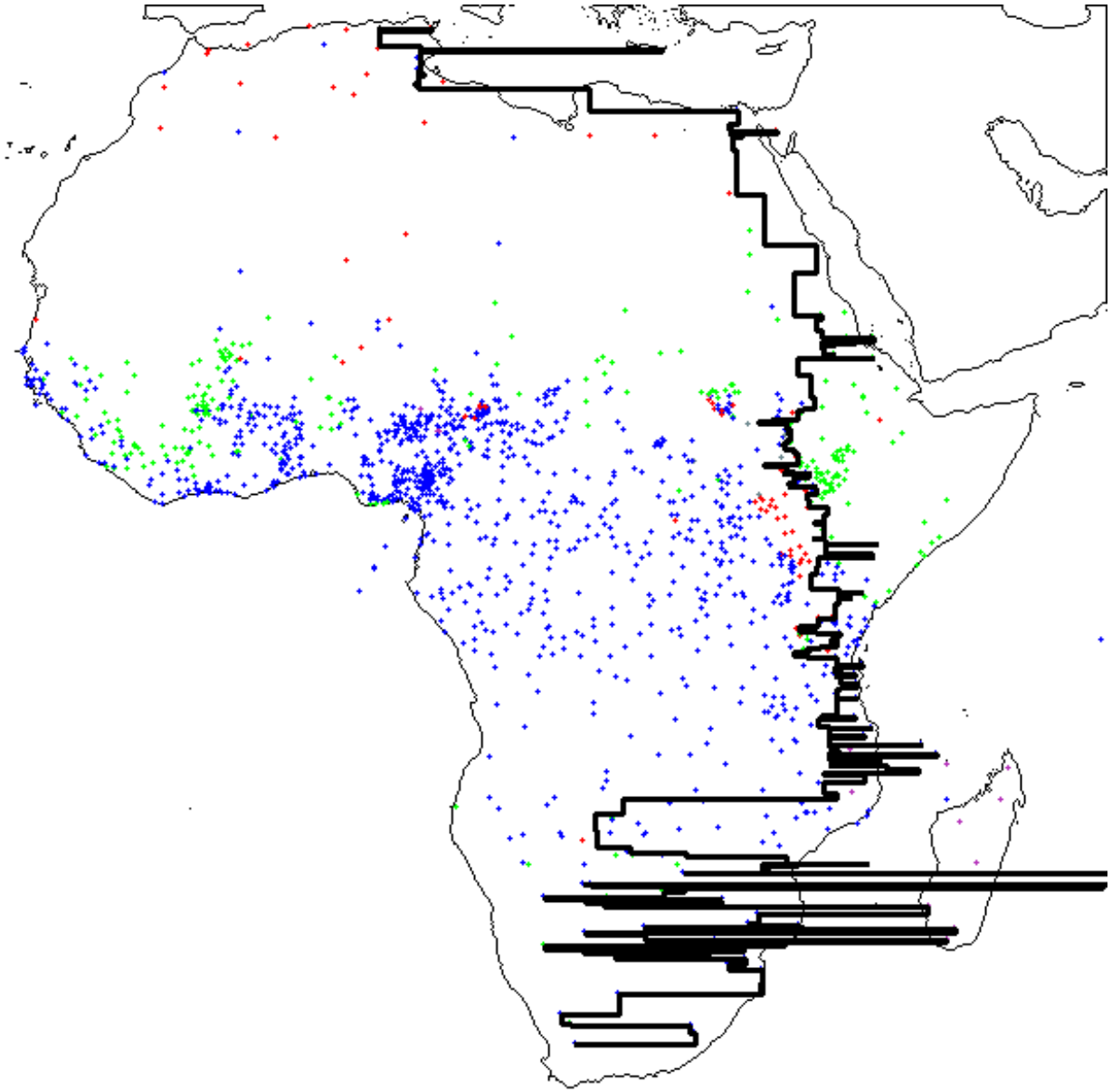
	Under	Over
SVO	1036	14
SOV	320	1
VSO	54	27
VOS	14	0
OVS	7	0
NODOM	2	0
H	1.12	1.07
#	1433	42

Constituent Order: North-South Line



- Suppose I let the computer find the *optimal* north-south line

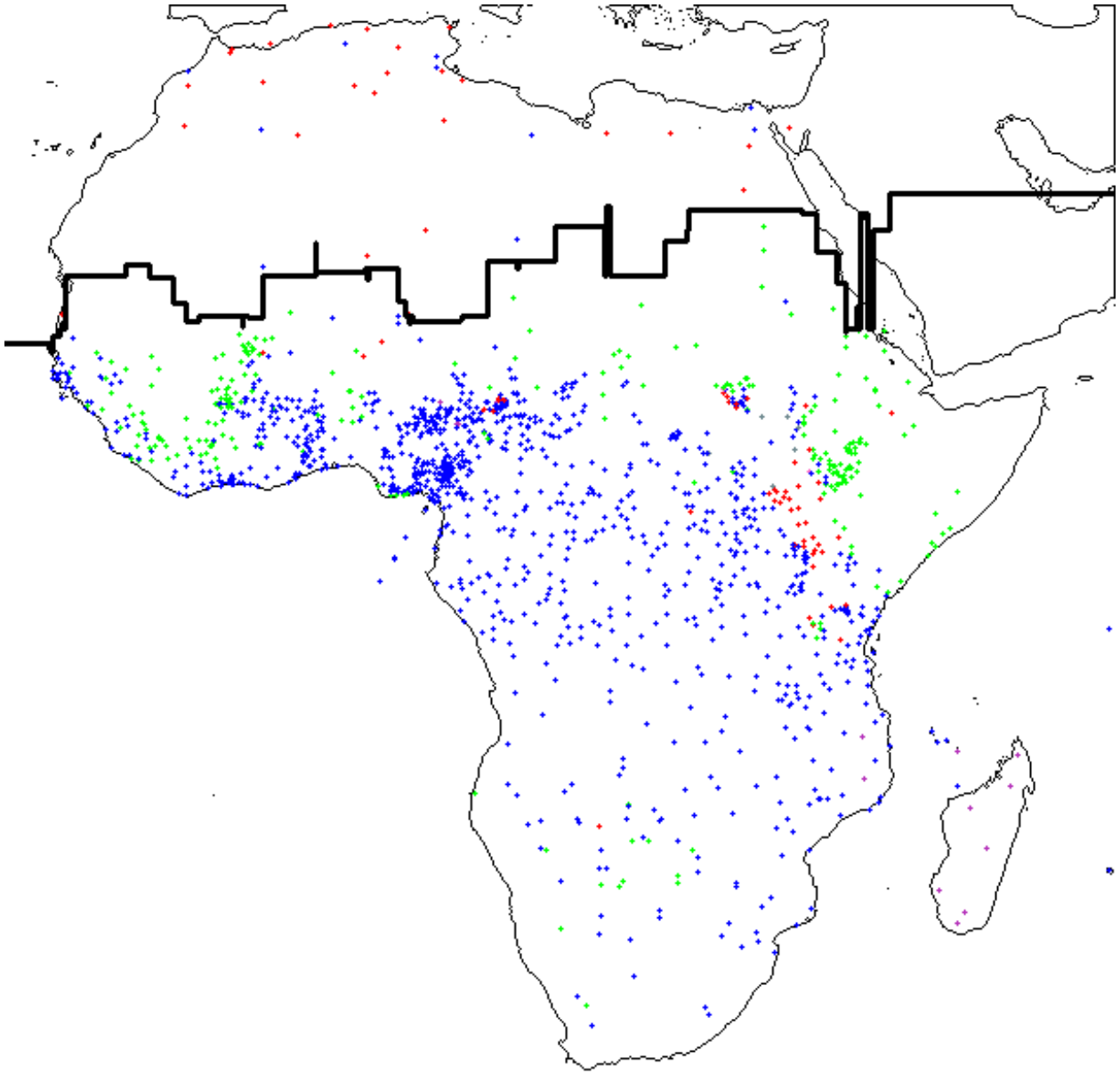
Constituent Order: North-South Line



- Suppose I let the computer find the *optimal* north-south line
- Its homogeneity is 1662.6

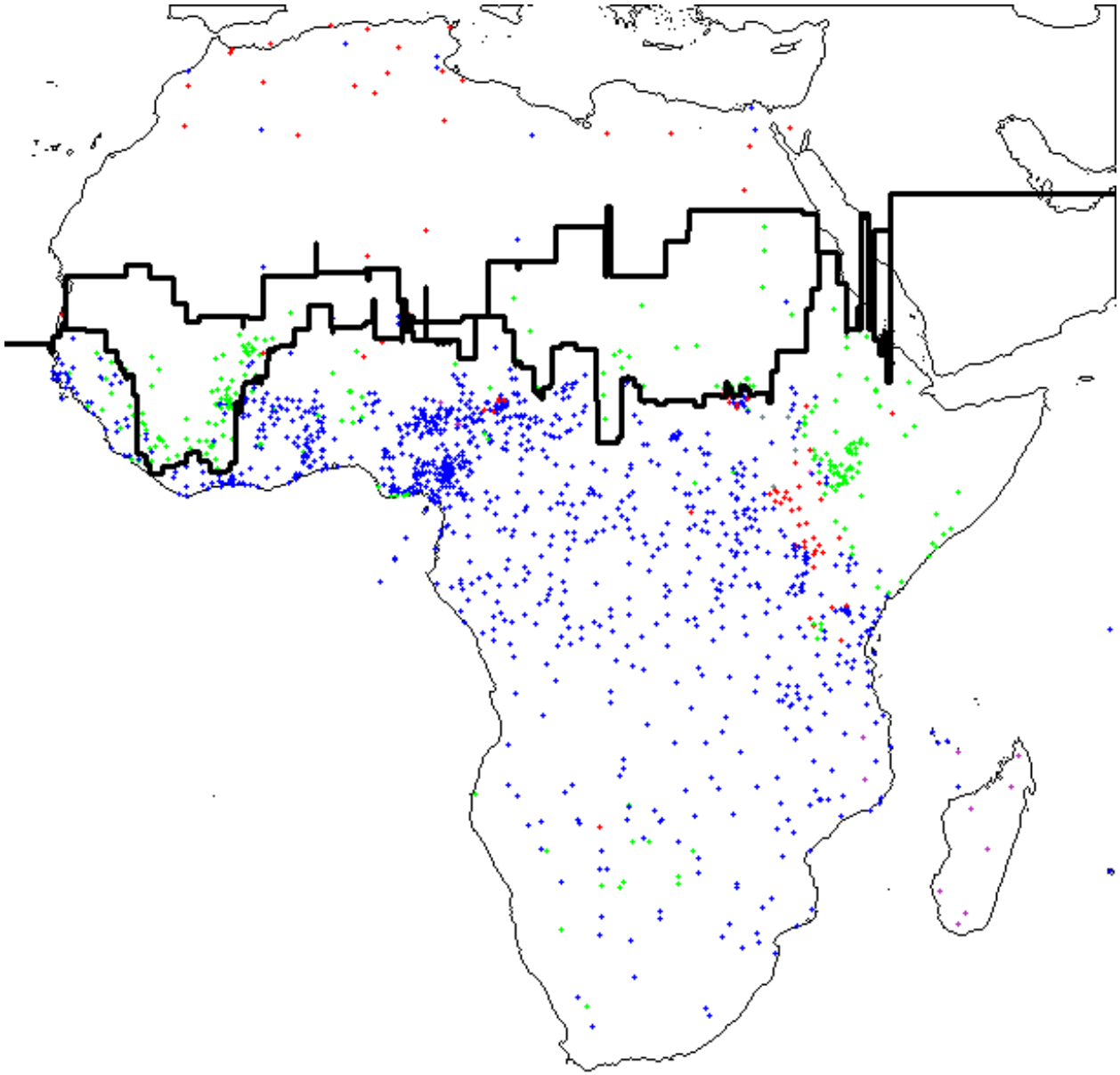
	Under	Over
SVO	957	94
SOV	234	87
VSO	72	9
VOS	4	10
OVS	7	0
NODOM	1	1
H	1.07	1.49
#	1275	201

Constituent Order: Optimal Line



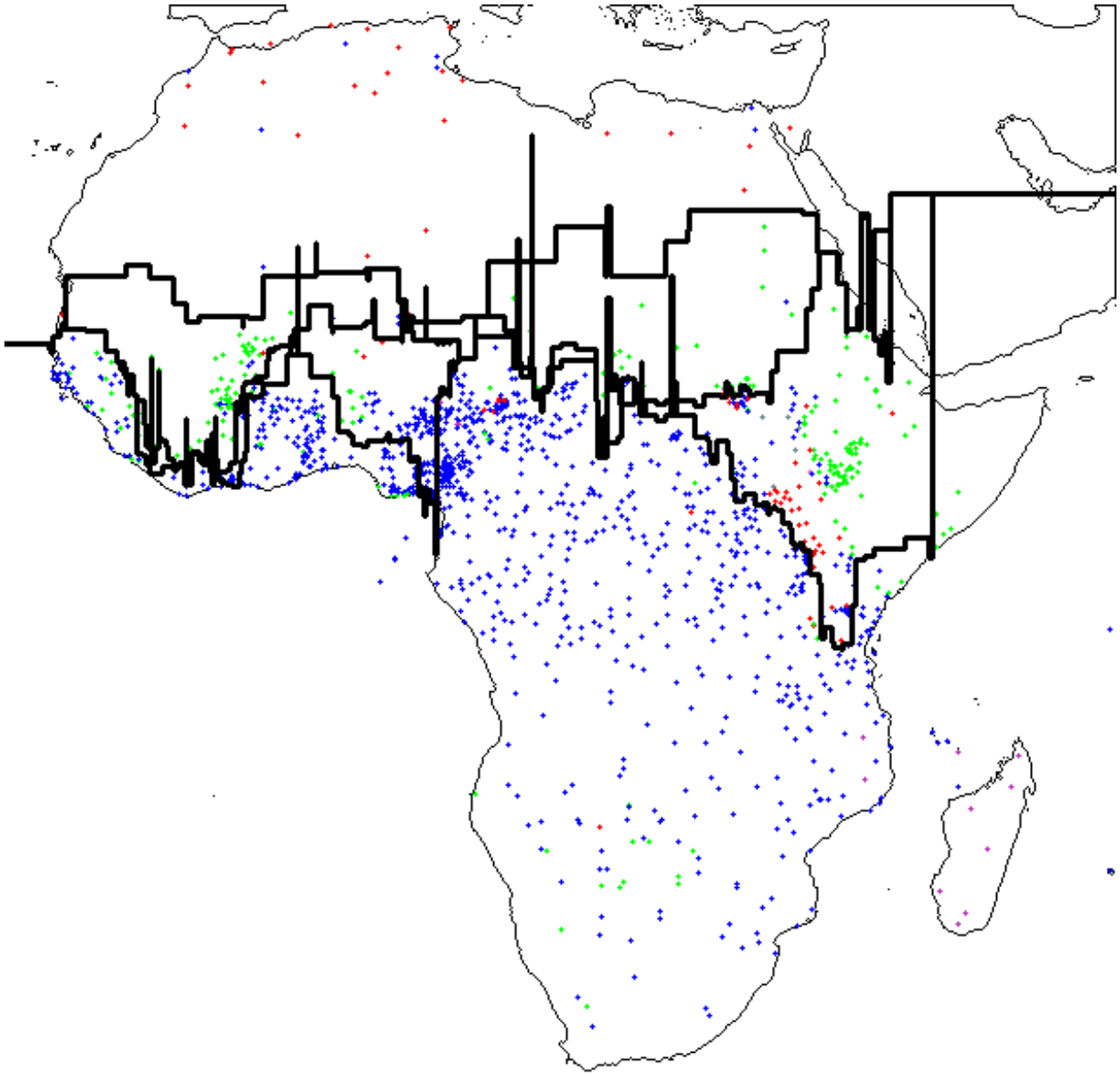
- So the east-west line was the most homogeneous
- Now we draw the *next* optimal line, given the first one!

Constituent Order: Optimal Line #2

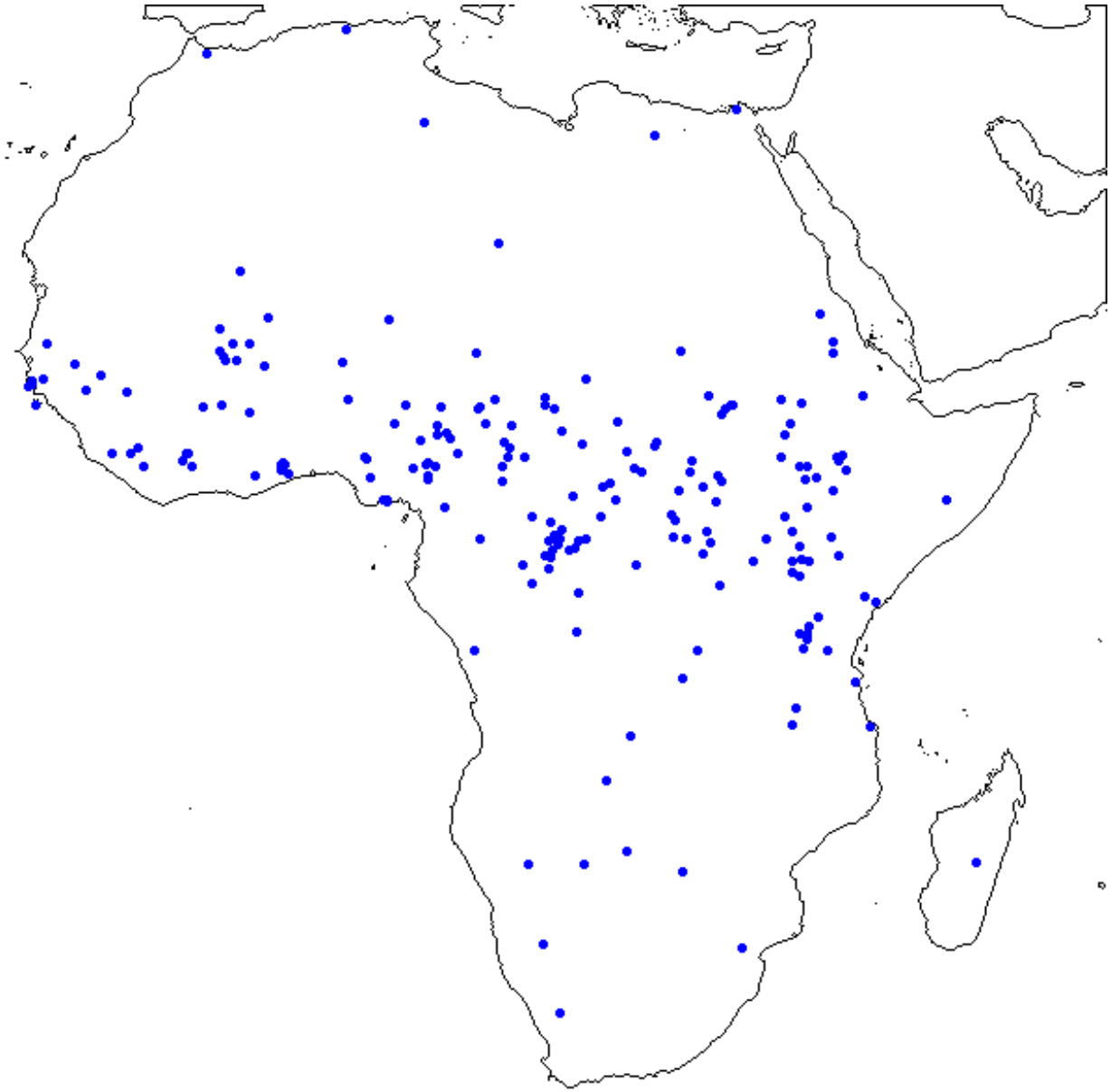


- We now obtain areas
- With arbitrary precision, as we draw further lines

Constituent Order: Optimal Line #3

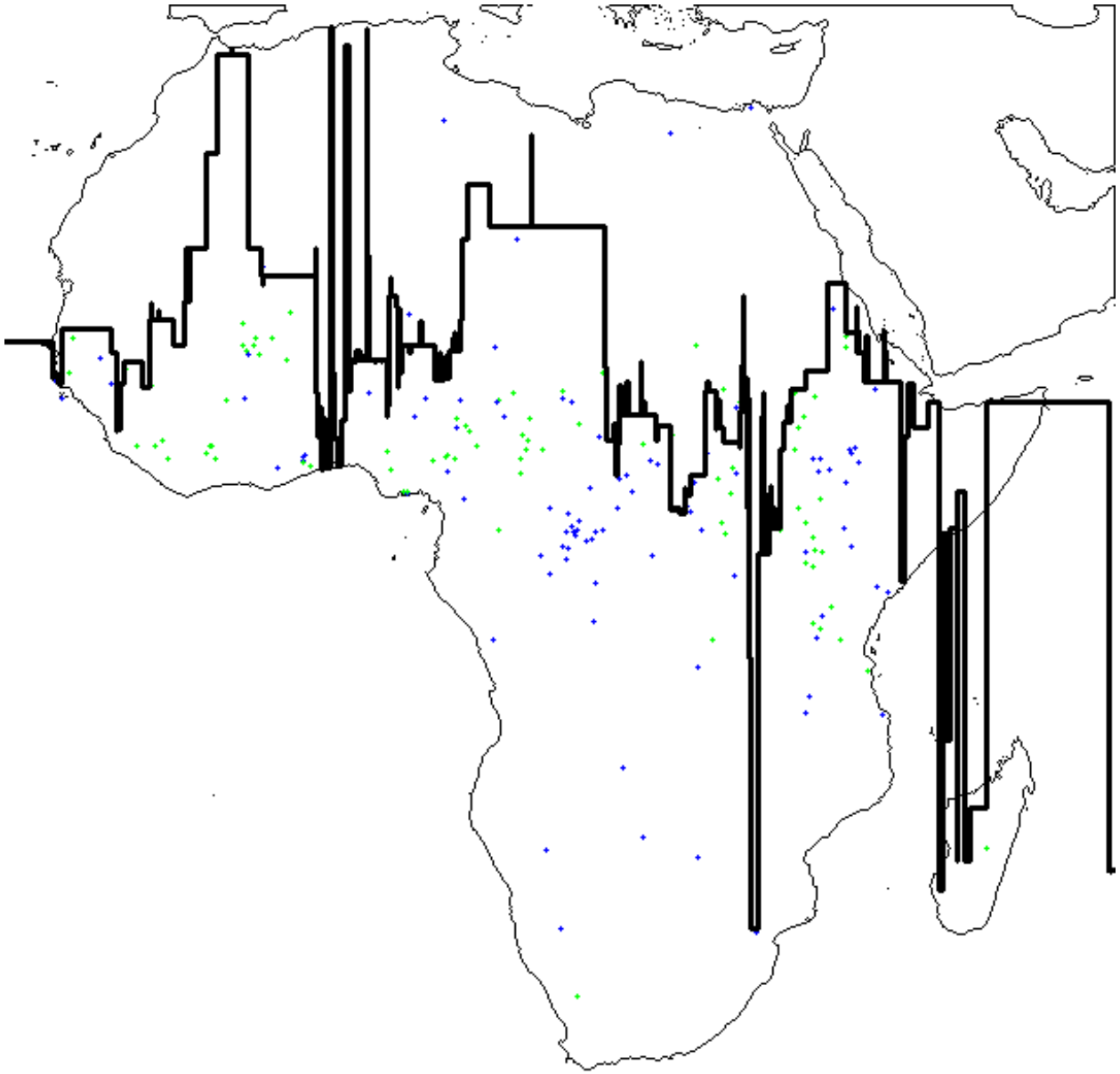


Grambank: 202 Morphosyntactic Features (201 Languages)

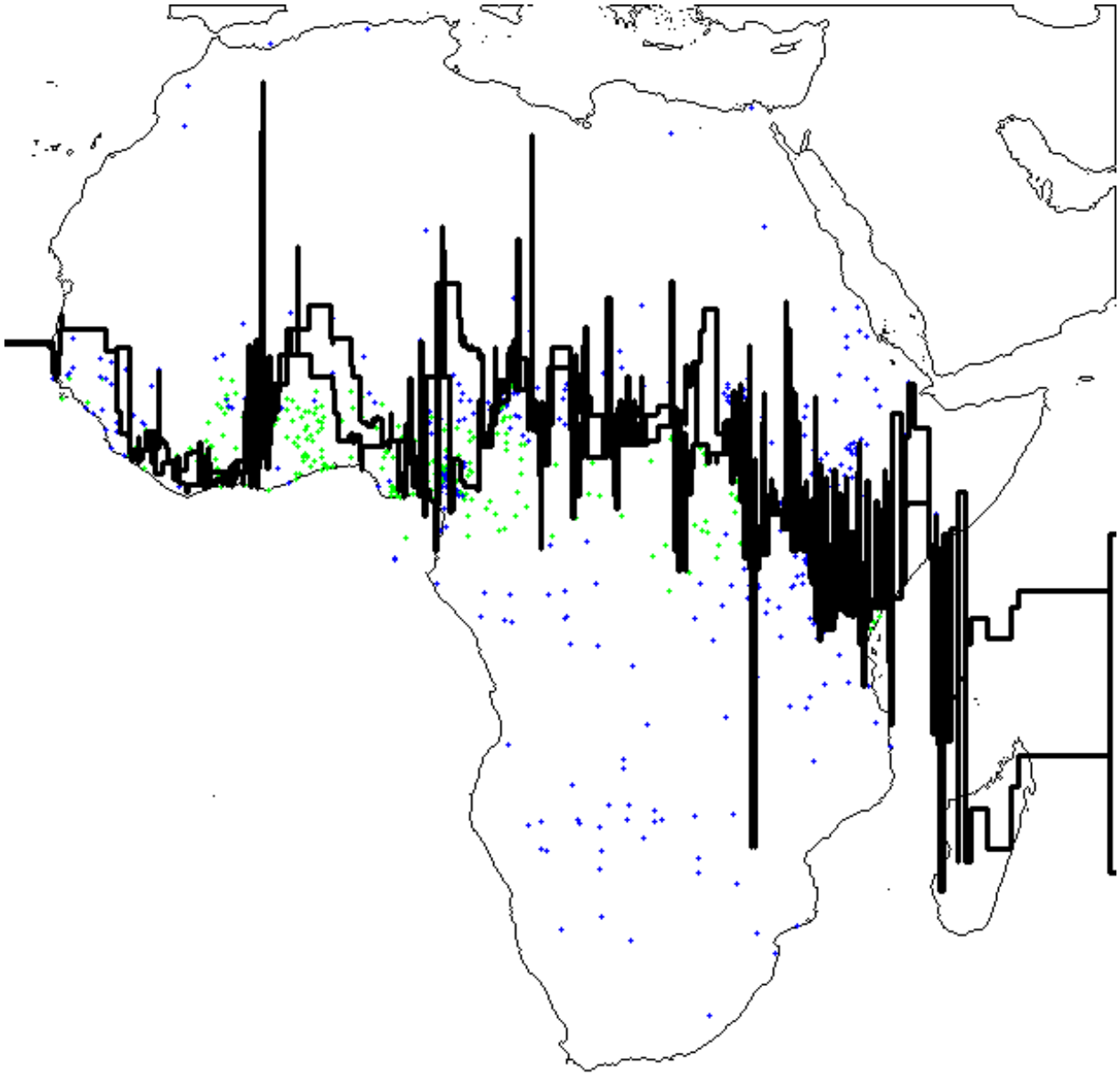


GB044: Can plural number be marked on the noun itself?
GB031: Is there a dual (or unit augmented) in addition to a plural (or augmented) number category in pronouns?
GB030: Is there a gender distinction in 3rd person pronouns (or demonstratives, if no 3rd person pronouns)?
GB025: What is the order of demonstrative and noun in the NP?
...

GB030: Is there a gender distinction in 3rd person pronouns?

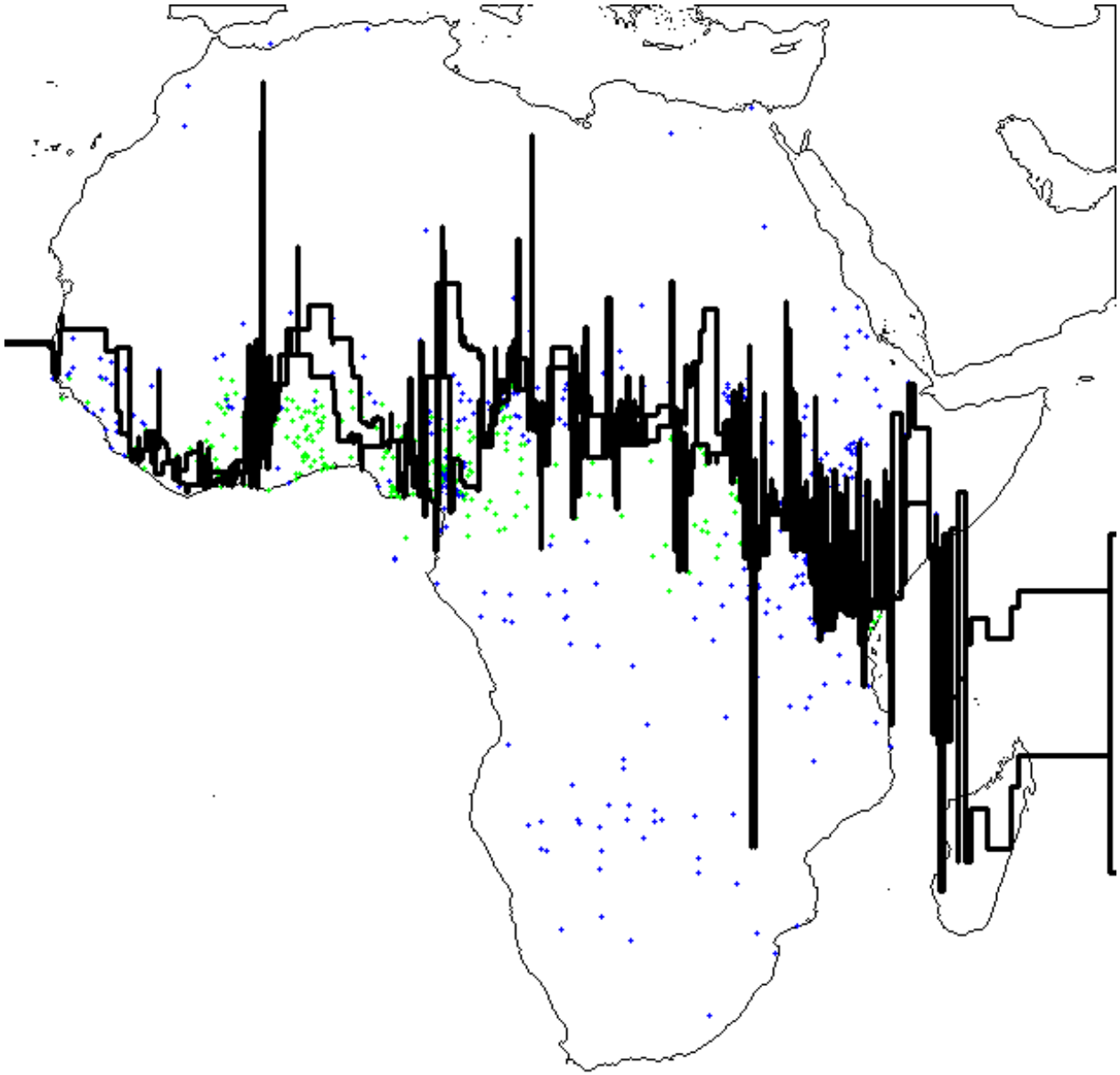


Grambank: **All** 202 features at the same time line #1



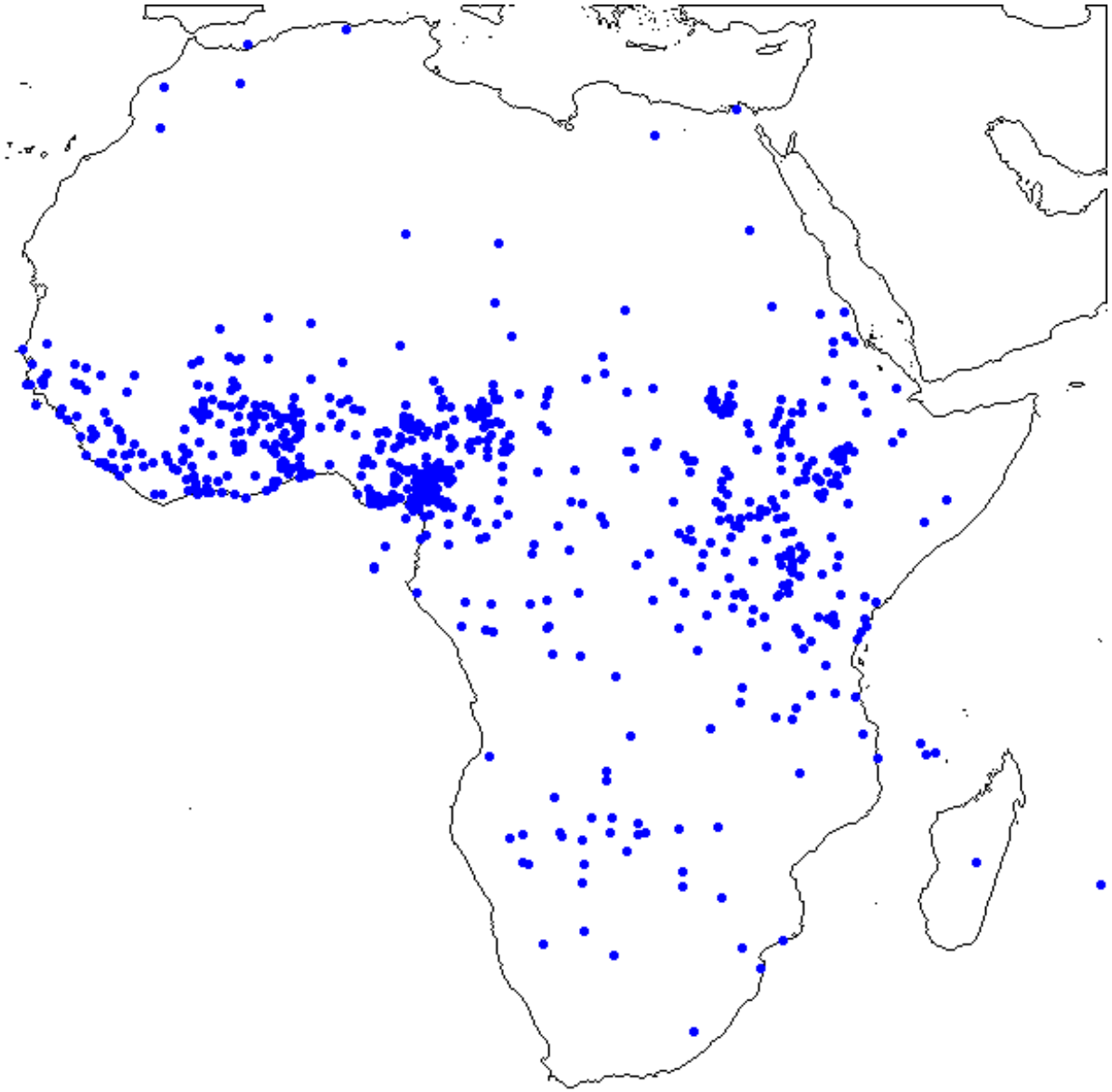
- We start to recognize this contour

Grambank: **All** 202 features at the same time line #2



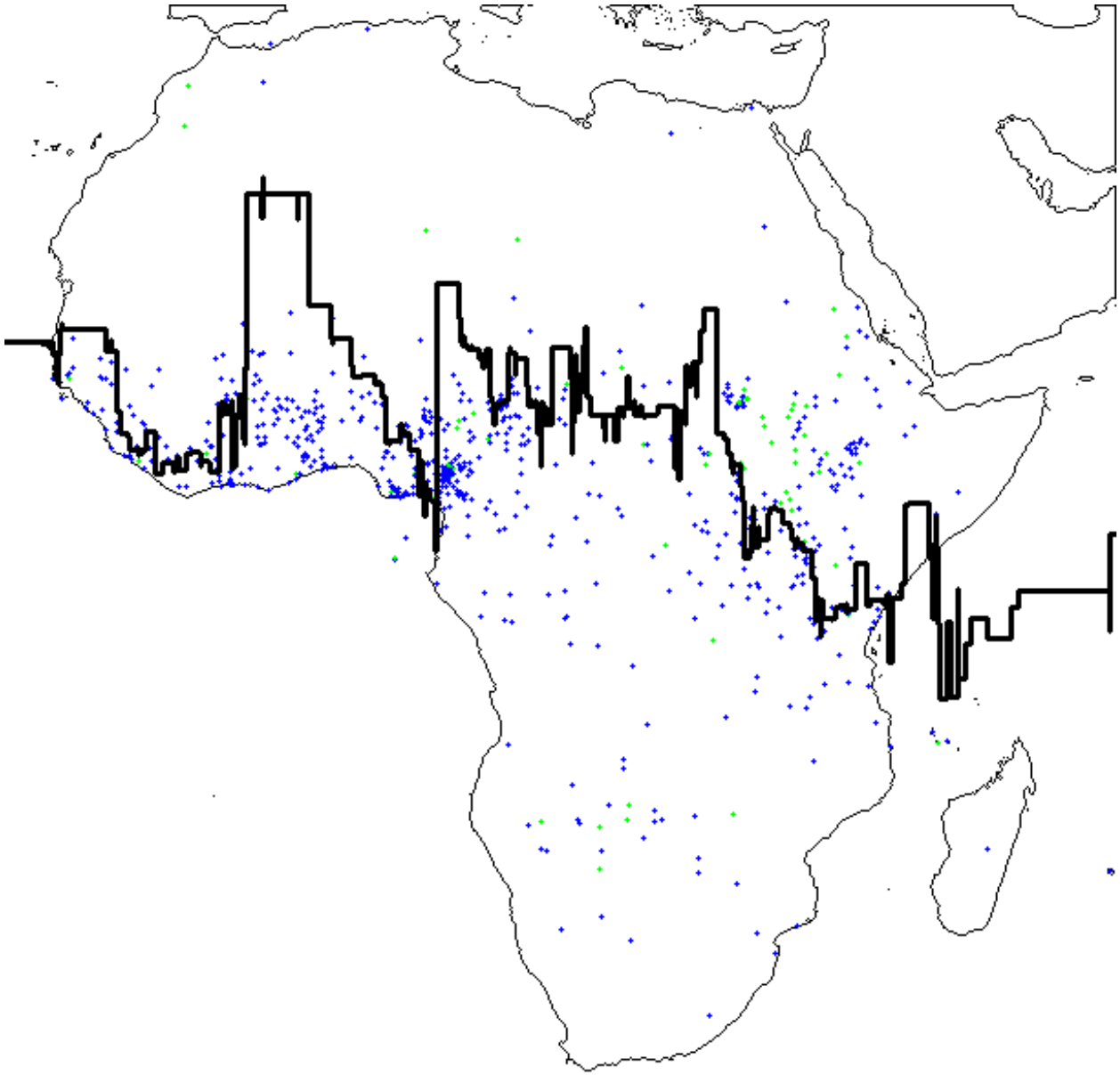
- Difficult to make sense of

PHOIBLE: 1373 (!) Segmental Features (706 Languages)



Does the language have /s/?
Does the language have /ã/?
Does the language have /kp/?
...

GB025: Does the language have /s/?



Conclusions

- Presented one automated technique for dividing geolocated data into areas with resemblance to what humans (aim to) do
- Unfortunately, difficult to make sense of isogloss lines which combine more than one or a few features
- Ideas on how to weigh/combine features greatly appreciated
- More work is needed before a serious comparison with human area-dividing can be done

Thank you

Clements, N. and Rialland, A. (2008). Africa as a phonological area. In Heine, B. and Nurse, D., editors, *A linguistic geography of Africa*, pages 36–87. Cambridge: Cambridge University Press.

Daumé, III, H. (2009). Non-parametric bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 593–601, Morristown, NJ, USA. Association for Computational Linguistics.

Güldemann, T. (2008). The macro-sudan belt: towards identifying a linguistic area in northern sub-saharan africa. In Heine, B. and Nurse, D., editors, *A linguistic geography of Africa*, pages 151–185. Cambridge: Cambridge University Press.

Güldemann, T. (2010). "sprachraum" and geography: Linguistic macro-areas in africa. In Lameli, A., Kehrein, R., and Rabanus, S., editors, *Language and Space: An International Handbook of Linguistic Variation Volume 2: Language Mapping*, volume 30/2 of *Handbooks of Linguistics and Communication Science*, pages 561–585. Berlin: Mouton de Gruyter.

Heine, B. (1976). *A Typology of African Languages*, volume 4 of *Kölner Beiträge zur Afrikanistik*. Berlin: Dietrich Reimer.

Heine, B. (2011). Areas of grammaticalization and geographical typology. In Hieda, O., König, C., and Nakagawa, H., editors, *Geographical Typology and Linguistic Areas: With Special Reference to Africa*, pages 41–66. Amsterdam: John Benjamins, Amsterdam.

Michael, L., Chang, W., and Stark, T. (2014). Exploring phonological areality in the circum-andean region using a naive bayes classifier. *Language Dynamics and Change*, 4(1):27–86.

Moran, S., McCloy, D., and Wright, R. (2015). Phoible online. Leipzig: Max Planck Institute for Evolutionary Anthropology (Available online at http://phoible.org, Accessed on 2015-10-01.).

Muysken, P., Hammarström, H., Birchall, J., van Gijn, R., Krasnoukhova, O., and Müller, N. (2015). Linguistic areas, bottom up or top down? the case of the guaporé-mamoré region. In Comrie, B. and Golluscio, L., editors, *Language Contact and Documentation*, pages 205–238. Berlin: DeGruyter Mouton.

Segerer, G. (2015). How databases shape research: labial-velars distribution in africa. Paper presented at the Workshop Areal

Phenomena in Northern Sub-Saharan Africa (8th World Congress of African Linguistics), August 20-24, 2015, Kyoto, Japan.