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

MAP 1
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 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.
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
- 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) $3 + 2 = 5$

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

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

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

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 languages (Database developed at SHH Jena)
Suppose we draw an arbitrary line.
Suppose we draw an arbitrary line. Its homogeneity is 1721.3.

<table>
<thead>
<tr>
<th></th>
<th>Under</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
<td>663</td>
<td>286</td>
</tr>
<tr>
<td>SOV</td>
<td>177</td>
<td>244</td>
</tr>
<tr>
<td>VSO</td>
<td>7</td>
<td>74</td>
</tr>
<tr>
<td>VOS</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>OVS</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>NODOM</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>0.92</td>
<td>1.51</td>
</tr>
<tr>
<td>#</td>
<td>861</td>
<td>613</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Under</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
<td>1036</td>
<td>14</td>
</tr>
<tr>
<td>SOV</td>
<td>320</td>
<td>1</td>
</tr>
<tr>
<td>VSO</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>VOS</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>OVS</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>NODOM</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>1.12</td>
<td>1.07</td>
</tr>
<tr>
<td>#</td>
<td>1433</td>
<td>42</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Under</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
<td>957</td>
<td>94</td>
</tr>
<tr>
<td>SOV</td>
<td>234</td>
<td>87</td>
</tr>
<tr>
<td>VSO</td>
<td>72</td>
<td>9</td>
</tr>
<tr>
<td>VOS</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>OVS</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>NODOM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1.07</td>
<td>1.49</td>
</tr>
<tr>
<td>#</td>
<td>1275</td>
<td>201</td>
</tr>
</tbody>
</table>
So the east-west line was the most homogeneous.

Now we draw the next optimal line, given the first one!
We now obtain areas

With arbitrary precision, as we draw further lines
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
Does the language have /s/?
Does the language have /ã/?
Does the language have /kp/?

PHOIBLE: 1373 (!) Segmental Features (706 Languages)
GB025: Does the language have /s/?

Hammarstrom
Macro-Areas Africa
17 June 2016, Berlin
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


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