Garland McNew[1], Curdin Derungs[2], Steven Moran[3*]
[1]Department of Geography, San Diego State University
[2]URPP Language and Space, University of Zurich
[3]Department of Comparative Linguistics, University of Zurich
*Corresponding author: steven.moran@uzh.ch

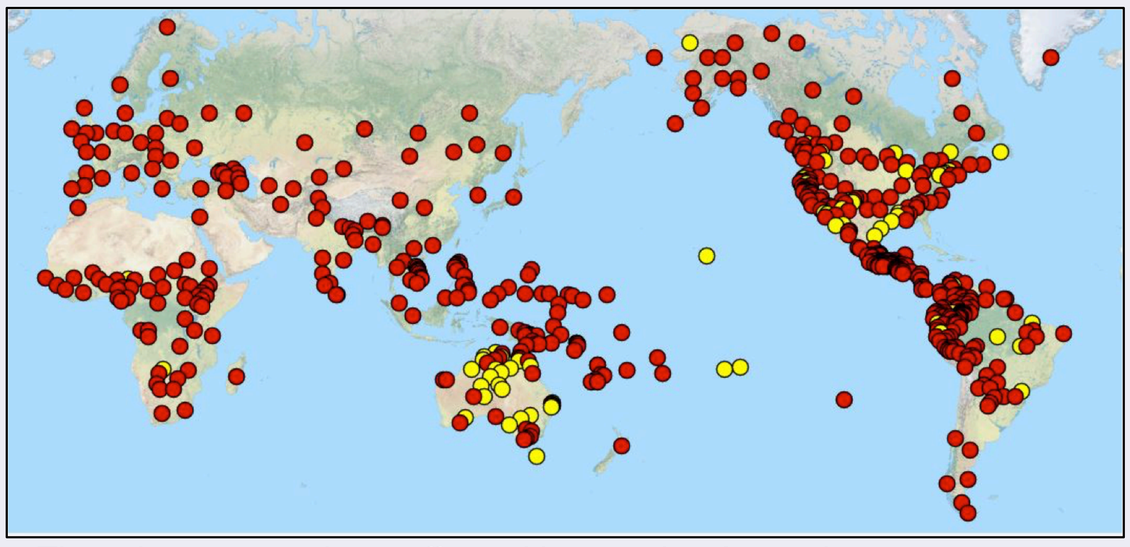
Towards faithfully visualizing global linguistic diversity

Overview

- There are approximately 7000 languages spoken in the world today and they are remarkably diverse in their sound systems, word formation strategies, and syntactic structures [1]
- Samples of linguistic diversity are encoded in many different typological databases, e.g. Cross-Linguistic Linked Data (http://clld.org/datasets.html)
- A popular strategy for visualizing worldwide linguistic diversity is to utilize point symbology to represent features of interest and to plot them as colored points on a Mercator map projection

The problem: illusions

- The World Atlas of Linguistic Structures (WALS) [2], Chapter 130 [3]
  - **Differentiation** (the language has separate words for 'finger' and 'hand', as in English) marked in red – 593 languages
  - **Identity** (a single word denotes both hand and finger) marked in yellow – 72 languages



- There is a strong visual effect with regard to Australia
- An explanation is proposed in [3], namely: "Farmers tend to lexically distinguish finger from hand more often than hunter-gatherers."

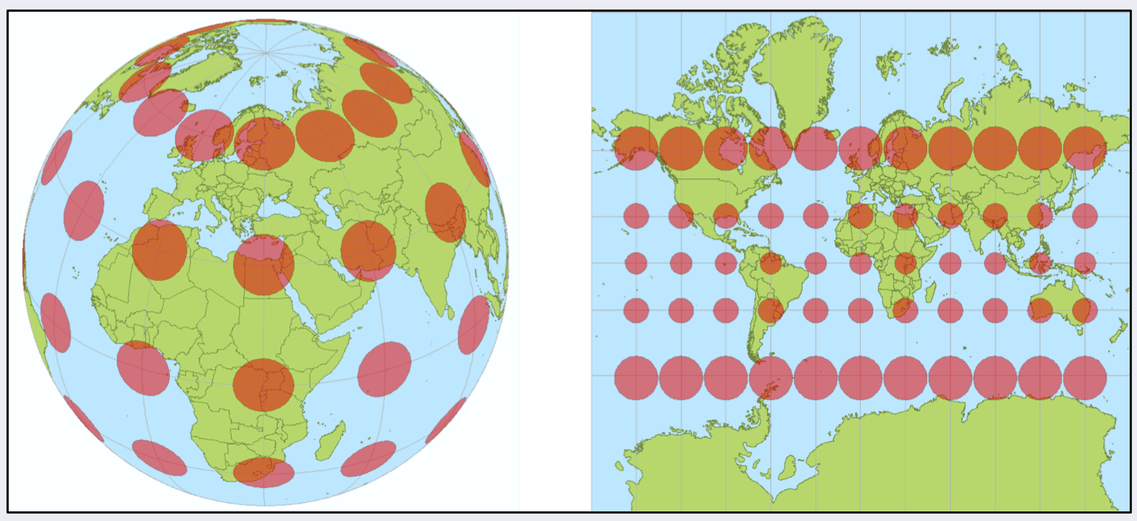
Identify the illusions

The bibliographic bias

- Typological datasets have poor cross-linguistic coverage because many languages have no documentation
- As such, data points typically come from convenience samples

The cartographic bias

- Mercator accurately represents lines of constant course (so-called rhumb lines that preserve directional accuracy and linear scale), but it distorts the area of objects as the latitude increases (north and south) from the equator due to the cylindrical nature of the projection
  - Consider that Greenland is roughly equal in land mass to Mexico, but due to Mercator projection it is has a much larger depiction



Languages as points

- Languages are not individual points, but are spoken by groups of people of varying population sizes and densities over different-sized, and often overlapping, geographic regions
- The lack of language points is also not indicated in typological atlases, so there is no visual cue for the absence of data

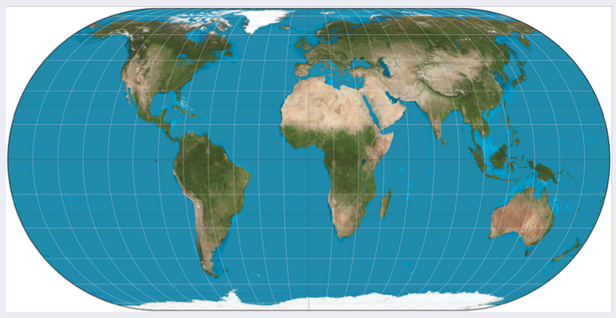
Colors dots or ranges of colored shapes

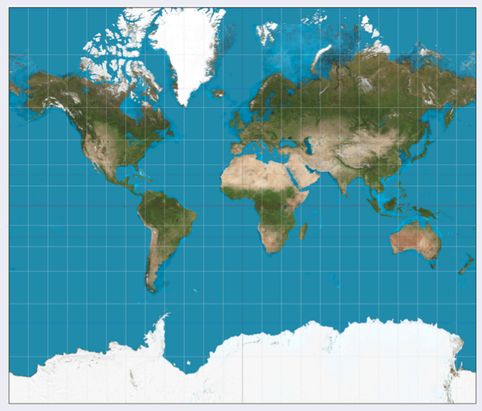
- How do you visualize for example consonant inventory sizes when they range from 6-to-140?
- A world map with 134 colored values would be uninterpretable (even if scaled by color) – one solution is to bin ranges
- But what statistical procedures went into determining these bins and how does their visualization change when the binning procedure is changed?

These biases must be taken into account when interpreting visual patterns of global linguistic diversity

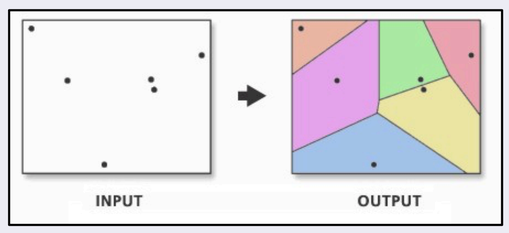
Towards a solution

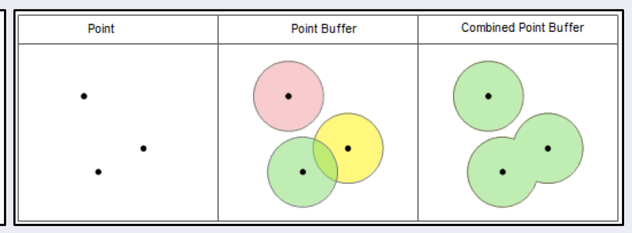
- One solution to the problem of projecting density visualizations is to not use Mercator (right), but instead use an equal area projection, such as Eckert IV (left)



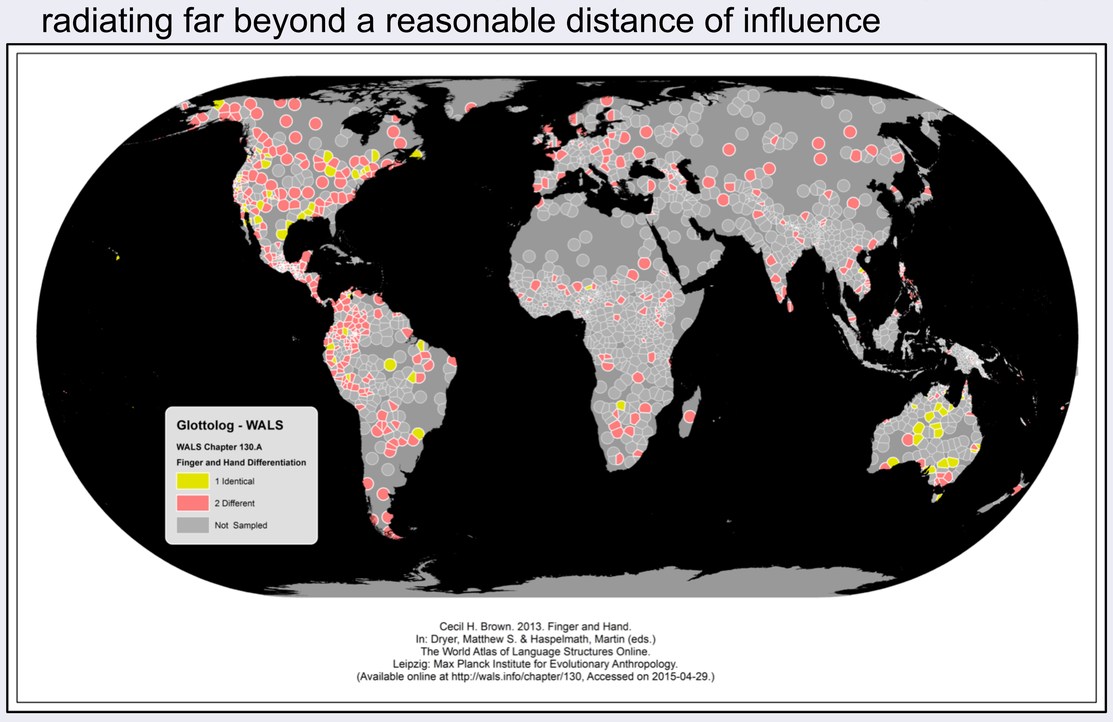


- Instead of language as points create a Voronoi/Thiessen tessellation layer from the language locations (longitude and latitude coordinates)





- Create a dissolved buffer around each point that creates the polygons and then use this buffer layer to clip the Thiessen polygon layer so that areas with low sampling density will not have large polygons radiating far beyond a reasonable distance of influence



- To address the illusion caused by missing data, the gray polygons show areas for which we know languages exist, but for which the input database is lacking information about the linguistic variable

Discussion

Preliminary solution and open source code

- Our approach presents a much less distorted picture of the reality of linguistic diversity and it can be applied to any set of typological data in CLLD databases:
  - https://github.com/bambooforest/visualizing-typology-data

Future research

- Language change involves to two factors: **relatedness** and **contact**
- Our approach here offers a preliminary stab at the problem of language contact by creating polygons that show which languages are likely to have been in contact
- But it lacks a straightforward way to encode language relatedness (beyond using point symbology or adding genealogical markers, such as coloring or border effects to the Thiessen polygon layer)
- Both **relatedness** and **contact** processes are implicit factors in visualizations of language data from typological databases
- They must be considered when evaluating visual patterns (or resulting illusions) in atlases of linguistic diversity
- The problems of visualizing genealogical relatedness and areal contact between languages on a global scale is an area that needs further research

References

1. Evans, N. and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. Behavioral and Brain Sciences, 32(05):429–448.
2. Dryer, M. S. and Haspelmath, M. (2013). WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.
3. Brown, C. H. (2013). Finger and hand. In Matthew S. Dryer et al., editors, The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.