Towards faithfully visualizing global linguistic diversity

Garland McNew[1], **Curdin Derungs**[2], **Steven Moran**[3]
[1]Department of Geography, San Diego State University
[2]University Research Priority Program Language and Space, University of Zurich
[3]Department of Comparative Linguistics, University of Zurich
garlandmcnew@gmail.com, curdin.derungs@geo.uzh.ch, steven.moran@uzh.ch

Abstract

The most popular strategy for visualizing worldwide linguistic diversity is to utilize point symbology by plotting linguistic features as colored dots or shapes on a Mercator map projection. This approach creates illusions due to the choice of cartographic projection and also from statistical biases inherent in samples of language data and their encoding in typological databases. Here we describe these challenges and offer an approach towards faithfully visualizing linguistic diversity. Instead of Mercator, we propose an Eckert IV projection to serve as a map base layer. Instead of languages-as-points, we use Voronoi/Thiessen tessellations to model linguistic areas, including polygons for languages for which there is missing data in the sample under investigation. Lastly we discuss future work in the intersection of cartography and comparative linguistics, which must be addressed to further advance visualizations of worldwide linguistic diversity.

Keywords: cartography, comparative linguistics, language diversity, data visualizations

1. The problem

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 (Evans and Levinson, 2009). Samples of linguistic diversity are encoded in many different typological databases. A popular way of visualizing these datasets is to map linguistic features to categorical values and then to plot them as colored points on a Mercator map projection. An example from The World Atlas of Linguistic Structures (WALS; Dryer and Haspelmath (2013)) is shown in Figure 1 (Brown, 2013).

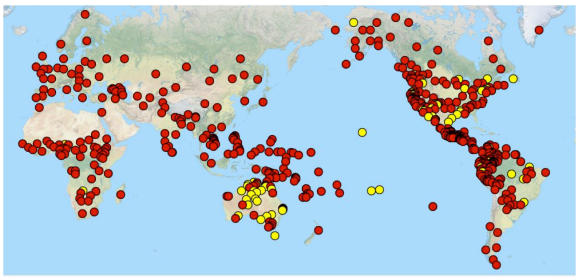

Figure 1: WALS Chapter 130: Finger and Hand

This and similar visualizations are unquestionably useful for doing exploratory analysis. Users can browse richly annotated digital maps that describe the known linguistic, cultural, and environmental diversity in the world (e.g. Kirby et al. (2016))[1] and researchers may use these visualizations to formulate hypotheses about linguistic and cultural phenomena, e.g. how they may have spread.

Moreover, many typological databases and digital atlases make their linguistic data and geographic coordinates openly available online. A fine example is Cross-Linguistic Linked Data (CLLD), which hosts over a dozen widely-used cross-linguistic comparative databases.[2] Hence the combination of increasingly easy-to-access tools for exploratory data analysis and access to the raw data has spurred hundreds of quantitative studies in comparative and historical linguistics. To highlight just two controversial studies that used WALS data, for example, consider whether linguistic tone has a genetic bias (Dediu and Ladd, 2007) and whether the worldwide distribution of phonemic diversity shows an out-of-Africa signal in ancient population movements (Atkinson, 2011).

Unfortunately there is a problem with the approach in visualizing global linguistic diversity described above: it creates potentially illusionary patterns that we – as humans and great pattern matchers – easily pick up on. These illusions are due to several factors that may misinform researchers if they are unaware of the model assumptions and statistical biases in the cartographic projection and linguistic data.

Let us consider for example Figure 1, Chapter 130 'Finger and Hand', which includes a global map displaying the distribution of the "two primary ways in which languages lexically treat the human finger and the hand of which it is a constituent" (Brown, 2013). The two values are either **identity** (a single word denotes both hand and finger) or **differentiation** (the language has separate words for 'finger' and 'hand', as in English). There are 72 languages of the identity type (marked with yellow) and 593 languages that mark differentiation (in red). The pattern of yellow versus red dots shows a strong visual signal that Australian languages are different than languages in the rest of the world

[1] https://d-place.org/

[2] These databases include, among others, WALS (Dryer and Haspelmath, 2013), The World Loanword Database (Haspelmath and Tadmor, 2009), The Atlas of Pidgin and Creole Language Structures (Michaelis et al., 2013), The database of the Automated Similarity Judgement Program (Wichmann et al., 2013), PHOIBLE (Moran et al., 2014), and Glottolog (Hammarström and Nordhoff, 2011). See: http://clld.org/datasets.html

(although there are some yellow dots in the Americas).
An explanation is proposed by Brown (2013), namely that "Farmers tend to lexically distinguish finger from hand more often than hunter-gatherers." This explanation is inline with what we know about the Aboriginal Australians – they were mostly hunter-gatherers, living in relative isolation from their initial settlement some 65kya (Clarkson et al., 2017), until the late 18th century (Pugach et al., 2013).
Testing whether hunter-gatherer languages are more likely to make a lexical distinction than agriculturalists is interesting because it provides a data point towards understanding whether there are cross-linguistic differences in languages spoken by hunter-gatherers and agriculturalists. But before we spend much time investigating the purported correlation in detail (or investigating one of the hundreds of other visual correlations in linguistic variables plotted as points on a Mercator map projection), we should consider that it is easy, and arguably natural, to draw conclusions from visual patterns in data, especially on maps.
In this paper, we highlight some of the important considerations when formulating a hypothesis from a linguistic atlas or when deciding to test a hypothesis inspired by one. We do so by first identifying the major cartographic and linguistic data factors that lead to illusions from plotting language data as points on a Mercator map projection. Then we provide a solution for faithfully visualizing global linguistic diversity.

2. Identify the illusions

One illusion is because typological datasets have for the most part poor cross-linguistic coverage. For example, Chapter 130 contains 593 data points, but there are more than 7000 languages spoken in the world today. Hence the map in Figure 1 shows only roughly 12% of the complete picture. This is a bibliographic bias, i.e. linguists are restricted to the accessible data about languages (Bakker, 2011). Around half of all languages are poorly described or undescribed (Hammarström, 2010). If there is no existing data on a particular language or language family, then that data cannot be used in descriptive studies about the distribution, of say, a given linguistic feature. At best it can be inferred through other means, such as language genealogy or known areal contact.
The bibliographic bias is exasperated by the cartographic problems involved with the Mercator map projection. A projection is a planar representation of a spherical object. That is, a map projection is an attempt to portray the surface of the earth onto a flat surface. The most common type of map projection is the Mercator projection, which was originally created in 1569 for nautical navigation purposes. Mercator accurately represents lines of constant course (rhumb lines). Mercator preserves directional accuracy and linear scale, however it distorts the area of objects as the latitude increases (north and south) from the equator due to the cylindrical nature of the projection. An example is given in Figure 2.
Why is this a problem? Well consider that Greenland is roughly equal in land mass to Mexico, but due to Mercator projection it is has a much larger depiction. Mercator should not be used for density visualization purposes.

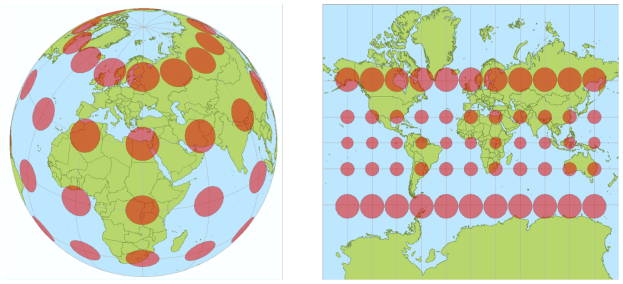

Figure 2: Mercator projection

Points are pulled or stretched away from one another visually when plotting them on a Cylindrical/Mercator type projection. Although positionally accurate, the areal distortion creates an erroneous perspective of density. Two points of equal distance will appear more clustered near the equator than they would near the poles. This is especially problematic in language diversity studies because the majority of the world's languages are spoken near the equator.
A second illusion comes from plotting languages as dots. 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. What impact does a single dot for 'the English language' have, if someone compares its geographic distance with all other languages? Crucially, the **lack** of language points is also not indicated in typological atlases. That is, there is no visual cue for the absence of data. This creates the illusion that all data points displayed represent the population under investigation, when in fact it is often a skewed sample, e.g. the data points come from a convenience sample derived from well documented and described languages.
A third illusion is due to the use of colored dots or ranges of colored shapes (in the case of non-binary feature values). For example, consider the number of consonants in phonological inventories cross-linguistically. How do you visualize consonant inventory sizes? A different colored dot for each value would likely mask any patterns in visualizing the global data because consonant inventories range greatly in size, from a low of 6 to a high of 140. A world map with 134 colored values would be uninterpretable even if scaled by color. To handle this variation, one approach is to bin continuous values into ranges, e.g. small, medium, and large (Maddieson, 2013). But what statistical procedures went into determining these bins and how does their visualization change when the binning procedure is changed? These factors must be taken into account when interpreting visual patterns of global linguistic diversity. Furthermore, there are linguistic-specific factors of genealogical descent and areal contact that affect the distribution of linguistic features, for which there is currently no good cartographic solution – we return to this the issue in the Discussion Section.

3. Towards a solution

One solution to the problem of projecting density visualizations is to not use Mercator, but instead use an equal area

projection, such as Eckert IV, shown in Figure 3.

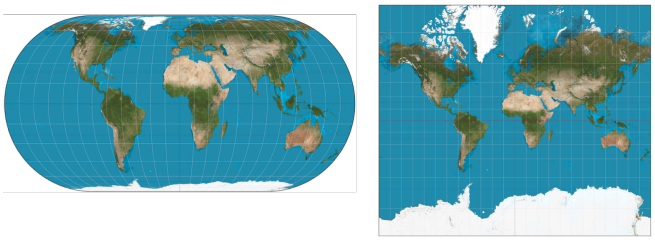

Figure 3: Eckert IV vs Mercator map projections

The Eckert IV projection preserves land mass and object size, but highly distorts lines of constant course. Since we are not concerned with maritime navigation, this projection is an acceptable compromise of distortion because it preserves land shape and area.

With a base map projection in place, we turn to the issue of languages-as-points. Linguistic visualizations often utilize point symbology to represent objects or features of interest, but languages are obviously spoken in areas and not in dots. Our solution is to create a Voronoi/Thiessen tessellation layer from the language locations (typically longitude and latitude coordinates). These polygons are generated from a set of sample points. Each Thiessen polygon defines an area of influence around its sample point, so that any location inside the polygon (Euclidian distance-wise) is closer to that point than any of the other sample points. Thus we no longer utilize the points to represent the language data, but instead the derived discrete polygons. An example is given in Figure 4.

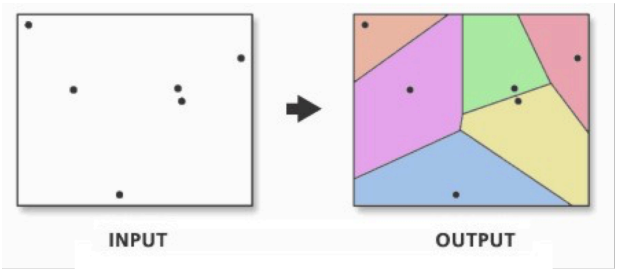

Figure 4: Conversion of point symbology to Thiessen polygons

However, one problem is that the entire geographic extent is accounted for when we create these polygons on a worldwide scale. That is, areas with low sampling density will have large polygons radiating far beyond a reasonable distance of influence. For example, consider Figure 5 and how the Thiessen tessellations grow disproportionately in areas of low linguistic diversity (data points in the tip of South America also radiates into Antarctica, where there are no native languages).

One approach to solving this issue is to create a dissolved buffer around each point that creates the polygons and then use this buffer layer to clip the Thiessen polygon layer, which we illustrate in Figure 6.

We use the dissolved buffer distances based on the data

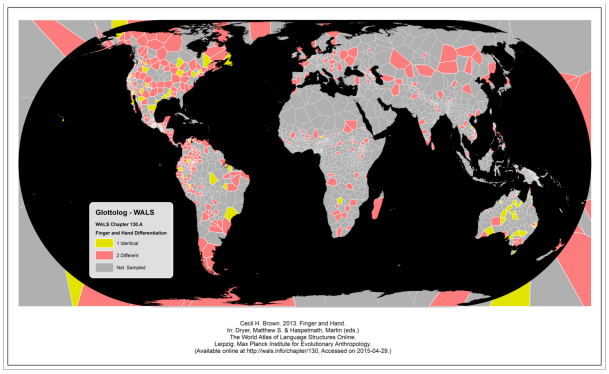

Figure 5: Worldwide Thiessen tessellations without a buffer

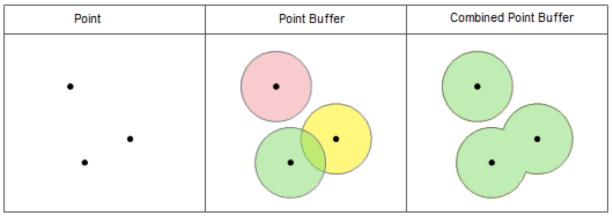

Figure 6: Dissolved point buffers

points provided by a particular typological database. For example in Figure 7, we apply a 200km buffer around each point and dissolve all of the buffers to account for the overlap of buffer zones.

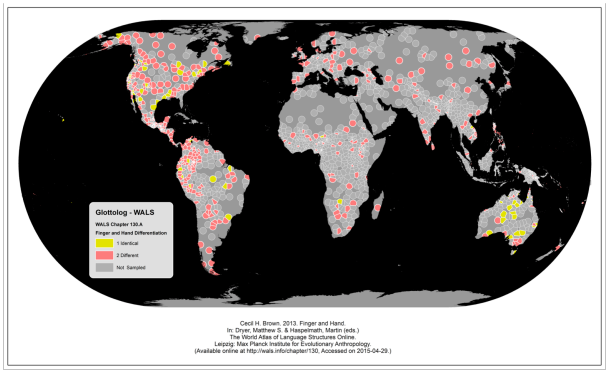

Figure 7: Worldwide Thiessen tessellations with buffer

In Figure 8, we provide an example in which we zoom in on the Americas. 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. Figures 7 and 8 show a global visualization of WALS Chapter 130 without its current cartographic illusions and linguistic biases. The yellow polygons show **identity**, the red polygons **differentiation**, and the gray polygons unknown values. Our approach presents a much less distorted picture of the reality of linguistic diversity and can be applied to any set of typological data.[3]

[3]In our initial research we used the geographic in-

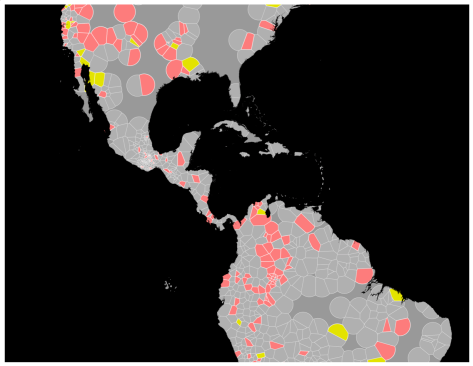

Figure 8: Zoomed-in Eckert IV projection with buffered Thiessen tessellations

4. Discussion

A popular strategy for visualizing worldwide linguistic diversity encoded in typological databases is to utilize point symbology to represent features of interest. We have shown here one way to overcome the most prominent biases in the linguistic data and cartographic projection that lead to visual illusions in global maps projecting linguistic diversity. Language data, however, also presents two additional challenges in visualizing worldwide linguistic diversity. When comparative or historical linguists investigate today's (synchronic) linguistic diversity, they are interested in the historic (diachronic) distribution of the world's languages due to how languages change over time. Language change involves to two factors: retention and contact.

Retention is the so-called **vertical** process of language change because it captures the idea that linguistic features are genealogically inherited from their parent language. Retention can be defined as the probabilistic likelihood of retaining a linguistic feature in descendant languages from the parent language. In other words, today's descendant languages have a high probability that they share features with their parent languages because they are genealogically related. An example is the Romance languages French, Italian, Spanish and Portuguese. They all share, to varying degrees, phonological and grammatical similarities with Latin.

The second type of language change is due to areal contact between speakers of different languages. These languages may or may not be genealogically related. For example, the lexicon of Brazilian Portuguese has long been influenced through contact with native South American languages. Beyond just borrowing words for new semantic concepts, language contact can lead to the borrowing of sounds and grammatical features (Matras, 2009).

Both retention and contact are implicit factors in the visualizations of language data from typological databases. However, they must be considered when evaluating patterns in maps of linguistic diversity because these factors are typically not overtly encoded. One procedure, employed by the CLLD, uses different colored shapes to indicate language family affiliation. However, this is only used when displaying the entire language sample as points globally or when a specific scalar variable is plotted on a map, e.g. the cross-linguistic distribution of the voiceless labiodental fricative [f].[4]

An alternative approach to visualize language genealogy is the application of the sunburst visualization (Stasko and Zhang, 2000) to the WALS and PHOIBLE databases (Mayer et al., 2014).[5] This application is successful in visualizing the hierarchical relatedness in language genealogy, but it does not address issues of geographic proximity and it offers nothing in particular for visualizing both known and unknown data points.

Our approach here offers a preliminary stab at the problem of contact by creating polygons that show which languages are likely to have been in contact. But it fails in discerning which languages are or were in contact (for this we need tertiary social information). Our approach also 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 proposed here. Thus the problem of visualizing genealogical relatedness and areal contact between languages on a global scale is an area that needs more research.

5. Acknowledgements

Many thanks to Robert Forkel, Martin Haspelmath, André Skupin, and to participants at the workshops *Visualizing Linguistic Data* organized by the URPP Language and Space at the University of Zurich, and *Language Comparison with Linguistic Databases* at the Max Planck Institute of Evolutionary Anthropology in Leipzig, Germany. We also thank three anonymous reviewers for their feedback.

6. Author contributions

SM, GM designed the research. GM implemented the cartographic methods. CD created the open source code. SM, GM wrote the paper.

7. Bibliographical References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332:346–359.

Bakker, D. (2011). Language sampling. In J. J. Song, editor, *Handbook of Linguistic Typology*. Oxford University Press, Oxford, UK.

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.

Clarkson, C., Jacobs, Z., Marwick, B., Fullagar, R., Wallis, L., Smith, M., Roberts, R. G., Hayes, E., Lowe, K.,

formation system software ArcGIS. We are now working on an open source version using R. The code is available on Github: `https://github.com/bambooforest/visualizing-typology-data`.

[4] `http://phoible.org/parameters/86E80A7DA80F7F58CA76F85BACA48F86#1/14/144`

[5] See: `http://www.th-mayer.de/wals/` and `http://tmayer.github.io/PhoibleVis/`.

Carah, X., et al. (2017). Human occupation of northern Australia by 65,000 years ago. *Nature*, 547(7663):306–310.

Dediu, D. and Ladd, D. R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences*, 104(26):10944–10949.

Dryer, M. S. and Haspelmath, M. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

Hammarström, H. and Nordhoff, S. (2011). Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language*, 3(2):31–43.

Hammarström, H. (2010). The status of the least documented language families in the world. *Language Documentation and Conservation*, 4:177–212.

Martin Haspelmath et al., editors. (2009). *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., et al. (2016). D-PLACE: A global database of cultural, linguistic and environmental diversity. *PloS one*, 11(7):e0158391.

Maddieson, I. (2013). Consonant inventories. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matras, Y. (2009). *Language Contact*. Cambridge University Press.

Mayer, T., Wälchli, B., Rohrdantz, C., and Hund, M. (2014). From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. *Language Processing and Grammars. The role of functionally oriented computational models*, pages 13–38.

Michaelis, S. M., Maurer, P., Haspelmath, M., and Huber, M. (2013). *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Moran, S., McCloy, D., and Wright, R. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M., and Stoneking, M. (2013). Genome-wide data substantiate holocene gene flow from India to Australia. *Proceedings of the National Academy of Sciences*, 110(5):1803–1808.

Stasko, J. and Zhang, E. (2000). Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization*, pages 57–65. IEEE.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., , and Valenzuela, P. (2013). *The ASJP Database (version 16)*. Max Planck Institute for Evolutionary Anthropology.