GlottoVis: Visualizing Language Endangerment and Documentation

Thom Castermans* TU Eindhoven

Harald Hammarström† Max Planck Institute for the Science of Human History

Bettina Speckmann* TU Eindhoven

Kevin Verbeek* TU Eindhoven

Michel A. Westenberg* TU Eindhoven

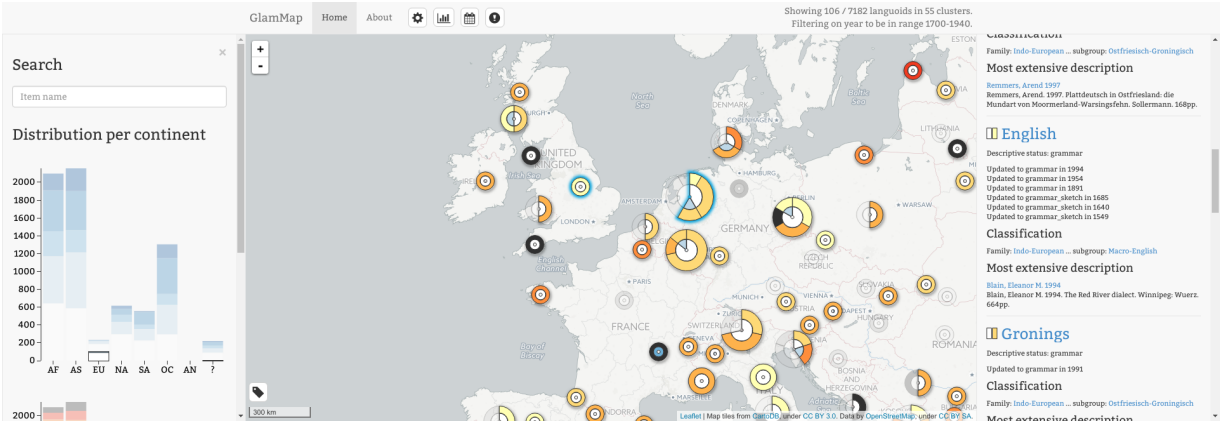

Figure 1: Status of European languages whose documentation changed between 1700 and 1940, two clusters (blue) are selected.

Abstract

We present GlottoVis, a system designed to visualize language endangerment as collected by UNESCO and descriptive status as collected by the Glottolog project. Glottolog records bibliographic data for the world's (lesser known) languages. Languages are documented with increasing detail, but the number of native speakers of minority languages dwindles as their population shifts to other more dominant languages. Hence one needs to visualize documentation level and endangerment status at the same time, to browse for the most urgent cases (little documentation and high endangerment) and to direct funding aimed at describing endangered languages. GlottoVis visualizes these two properties of languages and provides an interface to search and filter. Our tool is web-based and is comprised of glyphs on a zoomable geographic map. Clustering and (visual) data aggregation are performed at each zoom level to avoid overlap of glyphs. This reduces clutter and improves readability of the visualization. Preliminary tests with expert users confirm that our tool supports their desired workflow well.

Index Terms: I.3.3 [Computer Graphics]: Picture/Image Generation—Display algorithms; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Clustering; J.5 [Computer Applications]: Arts and Humanities—Linguistics

1 Introduction

There are approximately 7 000 languages spoken in the world at present. The diversity of these languages is an abundant resource for understanding the unique communication system of our species. All major languages are well-described in the sense that there are descriptive grammars, text collections, and large dictionaries available. For many smaller languages, however, very little information can be found. Consequently, documenting and describing all the lesser-known languages of the world is an on-going major objective for the field of linguistics. Given the size and breadth of the task, language description is an extremely decentralized activity, carried out by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, and linguists spanning several centuries. The Glottolog website [12] collects all relevant bibliographic data into one collection totalling over 275 000 references. Furthermore, each language is associated with a specific geographic location approximating where it is spoken. This extensive collection may be used to assess how much and what kind of descriptive materials exist for each language.

The task of describing languages is all the more urgent given the widespread tendency of speakers of minority languages to shift to another more dominant language. Such a shift starts with bilingualism in one generation, broken transmission to some later generation, and finally no transmission at all to the latest generation, leaving the language alive only as long as the oldest members of the early generation. A large number of languages are somewhere in this process and thus labeled *endangered languages* [10, 22]. This motivates the need to gain insight into documentation level and endangerment status at the same time, to browse for the most urgent cases (little documentation and high endangerment) and to direct funding aimed at describing endangered languages. An example of such efforts is the Endangered Languages Documentation Programme, `http://www.eldp.net`.

GlottoVis. In this paper we present *GlottoVis*, a system designed to visualize language endangerment and documentation status simultaneously. The core of our web-based system is a zoomable geographic map overlayed with bivariate glyphs (see `http://glammap.net/glottovis/`). We use clustering and (visual) data aggregation at each zoom level to avoid overlap of glyphs and provide an interface to search and filter languages. GlottoVis builds upon the core system of GlamMap [6], but its interface supports the workflow of our linguistic collaborators. Furthermore, GlamMap's glyphs are designed to display univariate data, whereas GlottoVis must handle bivariate data, which is a major design challenge. In this paper we focus in particular on a detailed problem analysis and the corresponding design decisions, and also report on preliminary user feedback. Details on the user interface can be found in the extensive "About" pages of GlottoVis.

*{t.h.a.castermans | b.speckmann | k.a.b.verbeek | m.a.westenberg}@tue.nl
†hammarstroem@shh.mpg.de

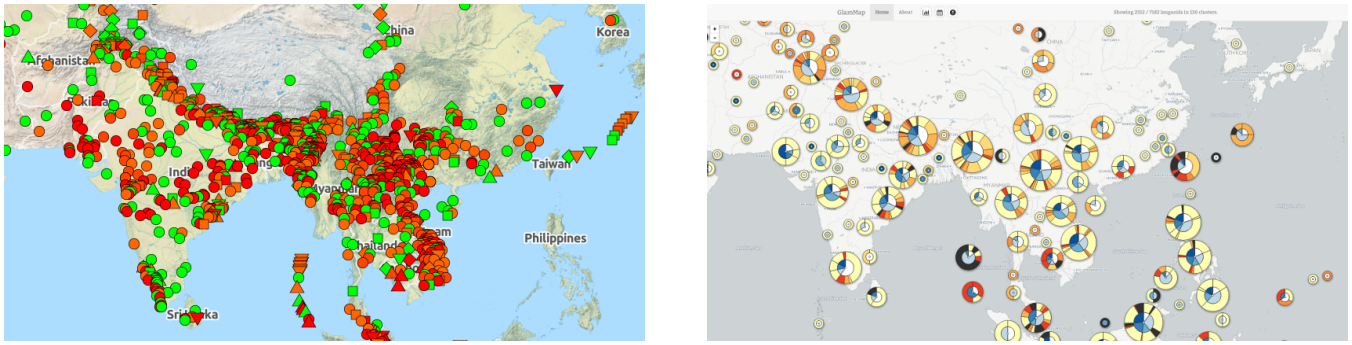

Figure 2: *Left:* Plotting languages as simple glyphs on a map. *Right:* Approximately the same view in GlottoVis.

Related work. The Glottolog team made an initial effort to visualize their data, see Fig. 2 (left). This "pins-on-Google-maps"-style visualization shows each language individually, which causes clutter and overplotting. Furthermore, the endangerment status of languages is indicated using shapes, which are not ordinal [4, 32]. Another language database is the Ethnologue [18, 29], a yearly publication that contains statistics for languages. The Ethnologue also provides a visualization of the available data, but this visualization offers very limited interaction. Users need to mouse over areas of the world to be able to view the status of languages in that area, making it hard to get an overview.

Just as GlamMap, GlottoVis shows items at their respective locations. The resulting layout problem hence resembles dynamic map labeling [2], with the crucial difference, that we aggregate overlapping items into disjoint glyphs instead of resolving overlaps by omitting items [34]. Related techniques have been used in a variety of settings, including network exploration (see, for example, Vehlow *et al.* [33]). A good overview of techniques is given by Scheepens *et al.* [26].

There are many techniques to display univariate geo-located data on maps, such as quadtree aggregation [3], choropleth and symbol maps [7], and binning [5, 25]. Several of these techniques employ circles [8, 16, 25] and hence have a visual appearance similar to GlottoVis. However, for our users it is of paramount importance to see both endangerment (do people still speak a language) and descriptive status (how well documented is a language) at the same time. We hence need a visual design which can accommodate bivariate geolocated data.

We use an agglomerative hierarchical clustering algorithm. A similar approach has been used by GeoTemCo [16]. However, there are two important differences. First of all, GeoTemCo does not aggregate multivariate data into a single glyph but instead uses multiple circles for a single location, hence weakening the association of symbol and location. Secondly, the clustering algorithm proposed by GeoTemCo does not succeed completely in avoiding overlaps, since it checks for overlaps only among neighbors within the Delaunay triangulation. It is hence comparatively easy to construct instances where the glyphs will overlap. Our clustering algorithm guarantees that glyphs are disjoint. Two related clustering approaches are NanoCubes [19] and ElasticSearch [1, 9]. However, both of these methods are targeted towards agglomerating points into grid cells for a quick overview and to improve performance. Our goal is solely to improve the clarity of the map and hence our aggregation merges as few glyphs as possible while still ensuring that all glyphs are disjoint. It is also important to note that our clustering algorithm purposefully does not use demographic or political boundaries to cluster. The temporal extent of the data set is such that no particular set of such borders will ever be valid for all data. To not wrongly place the data into an invalid geopolitical context we guide our clustering simply by the geometric proximity of the language locations.

2 Problem Definition

In this section we first give additional background on our data set and then analyze the problem of effectively visualizing multivariate geolocated data. We do so by defining a series of tasks that should be supported by the visualization and also consider various visual aspects that influence task efficiency.

2.1 Data Description and Definitions

Language. A *language* is, for our purposes, an entity with properties, one of which is a unique identifier. The name of a language (Dutch or German, for example) will in practice serve this role, but our input has other identifiers too, like a Glottocode [12] and an ISO-639-3 code. The other three properties which we consider are *point location*, *endangerment* and *descriptive status*.

Endangerment status. The endangerment status of a language is an ordinal property describing the risk of a language to go extinct. There are six values, which are defined by UNESCO [22]: extinct; critically or severely or definitely endangered; vulnerable; safe.

Descriptive status. The descriptive status of a language is an ordinal property assigned by the Glottolog project, with nine possible values describing how well the language has been documented by capturing its most extensive description. Examples are *grammar*, *phonology* and *wordlist*. For the visual encoding we created five categories, in consultation with domain experts. In GlottoVis, the more fine-grained Glottolog classifications are visible when viewing the details of a language.

The descriptive status of a language might change over the years. For example, a language may have been described by a wordlist in 1823, by a grammar sketch in 1947, by a grammar in 1952 and a more extensive grammar in 1993.

Language status. Endangerment and descriptive status are independent from each other and together form the *language status*.

2.2 Problem Analysis

Our input consists of a set of languages with the properties described above (and possibly more). Together with our collaborators from the Glottolog project we defined the following essential tasks to analyse a set of such languages:

- ... see geographical distribution of **T1** endangerment status; **T2** descriptive status, and **T3** language status;
- ... see statistical distribution of **T4** endangerment status; **T5** descriptive status, and **T6** language status;
- **T7** determine individual language status;
- **T8** find languages with a specific language status;
- **T9** find languages in the geographic neighborhood of another one.

Visual support for these tasks allows a user to answer questions such as: Identify a region with many endangered languages (**T1**); In

terms of documentation, is it correct to say that South America is the "least known continent"? (**T2**); Find one region which has many little documented and endangered languages (**T3**); Which continent has the highest proportion of extinct languages? (**T4**); Which continent has the highest (absolute) number of languages with full grammars? (**T5**); Suppose you are managing a research fund set up to improve the language documentation levels in neglected regions. Where would you invest and why? (**T3** and **T6**); The Lafofa language is spoken in the south of the Nuba mountains in North Sudan, not far from the border with South Sudan, what is its endangerment and documentational status? (**T7**); Find a language in the Sepik region (North Papua New Guinea) which is definitely endangered (or worse) and has no more than a wordlist of documentation (**T8**); Which are the two critically endangered languages closest (as the crow flies) to Kathmandu? (**T9**).

There are many techniques for visualizing a set of ordinal variable values (**T4**, **T5**). Examples are pie charts, bar charts, box plots [21], dot plots [35] and spine plots [15]. Visualizing two related ordinal variables (**T6**) is less straightforward, but still various chart types are known, including mosaic plots [11, 14], contingency tables [23] and treemaps [17, 27]. However, all of these visualizations lack the ability to perform **T7** and **T8**. While it is in principle possible to label cells in a mosaic plot, labels will overlap or become tiny when many languages are visualized. These types of visualizations are essentially designed to give an overview and thus only work well for the first part of the visual information-seeking mantra [28] "overview first, zoom and filter, then details-on-demand". A notable exception are treemaps, which can be used in combination with tooltips in an interactive setting. This would partially enable **T7** and **T8**.

One obvious approach to enable users to locate languages is by using the language location. Hence we need to support tasks **T1**, **T2**, **T3** and **T9**. Univariate georeferenced data is commonly visualized by plotting colored or scaled circles on a map [8, 16, 25]. Such approaches can be extended to multivariate data by plotting symbols that provide more information. Assuming that either there are legible labels or tooltips (in an interactive setting), it should be easy to perform **T7**, **T8** and **T9**. However, performing **T1**, **T2**, **T3**, **T4**, **T5** and **T6** will be harder, if not impossible, whenever we face overplotting problems. In the univariate case, overplotting can be avoided by, for example, using binning [5], a technique which unfortunately does not extend to multivariate data. Motivated by this discussion, we consider several visual aspects that may influence the task efficiency:

- **V1** the visualization accurately represents the input data;
- **V2** it has clean, easy to read symbols;
- **V3** these symbols stand out from the background, and
- **V4** these symbols are clearly separated from one another.

First and foremost, we need to accurately visualize the data (**V1**). Our output needs to have a firm grounding in the input data. Hence languages should be displayed (approximately) at their associated point location. Absolute precision is however not needed, since a point location is just a very rough approximation of the region where a language is spoken. Furthermore we need to ensure that the geographical distribution of language status (**T1**, **T2** and **T3**) is displayed in such a manner that it can be read easily by users. An aspect of the visualization that can play an important role here is the choice of colors [20].

Once our visualization is grounded in the input, we need to facilitate that users can read it quickly and easily. Mostly to be able to perform tasks **T1**, **T2** and **T3**, but to a lesser degree also **T4**, **T5** and **T6**. To do so, we need to use glyphs that are easy to read. There are several considerations. Firstly, the glyphs should stand out from the background (**V3**). Because glyphs are shown on a map, care should be taken to ensure that the map does not interfere with the glyphs. Secondly, every glyph should have as low a visual complexity as possible (**V2**) since many glyphs will potentially be shown close to each other. On the one hand, users need to be able to distill as much information as possible from a single glyph, but on the other hand scanning many glyphs should give an impression of the overall distribution of language status. Ideally glyphs are not too close to each other but this might be hard to avoid without distorting the input data ($\neg$**V1**). Glyphs should be outlined by a border or other means, so that they are clearly separated (**V4**).

3 Design Decisions

Our system GlottoVis was developed according to the problem analysis in Section 2. In this section we discuss the design decisions that shaped GlottoVis and the rationale behind those decisions.

3.1 Glyphs on a map

To support tasks **T1**, **T2**, **T3** and **T9** we decided to build our visualization around the geographical location of all languages. This decision is further supported by the fact that users have a strong association between a language and the approximate location where this language is spoken.

Simply plotting glyphs on a map at the correct location results in a very cluttered and potentially hard to read visualization. Fig. 2 (left) shows such a design based on initial ideas of the Glottolog team. There are easy improvements possible, such as always drawing endangered languages on top of less endangered ones, and choosing different colors and symbols. Still, given the number of languages and corresponding locations in Glottolog, any visualization in this style will always suffer from clutter and overplotting.

We resolve these issues by merging overlapping glyphs into a bigger glyph that displays the accumulated data of all merged glyphs. A single clean glyph is easier to read than an arbitrary number of overlapping ones. We increase the size of each glyph proportional to the number of languages it represents. Increasing its size can cause the newly constructed glyph to overlap other glyphs, so the merging process needs to be repeated until no more overlap remains. To allow users to see more details as they zoom in, we do not scale the glyphs at the same rate as the map when zooming in / out. Thus, the overlap needs to be evaluated at every zoom level.

The clustering of glyphs must be both consistent and efficient. If the user pans, the clustering should not change. Furthermore, if two glyphs are in the same cluster at a particular zoom level, then they should remain grouped at lower zoom levels. Hence we compute a hierarchical clustering for all glyphs on all zoom levels. From this hierarchy we can efficiently extract the clustering at any zoom level.

As noted in the introduction, our algorithm merges glyphs strictly based on their geometric proximity, not based on country or continent borders or even seas. The reader might wonder why we do not political or linguistic borders to improve the quality of the clustering. The reason is that borders are ill defined and change over time. War causes countries to split or merge and languages change over time due to internal and external influences. Trying to set consistent, non-disputed borders and then use them in the clustering is hence essentially impossible.

3.2 Glyph design

Recall that we have two variables per language that should be visualized with the glyphs: the endangerment status (six different values) and the descriptive status (five different values). While designing the glyphs, we take the following concerns into account:

- **C1** users need to see not only endangerment status and descriptive status separately, but also their correlation to determine the language status of every language that is represented (**T1**, **T2**, **T3**, **T7**, **T8**, **V1**);
- **C2** glyphs should be of as low a visual complexity as possible (**V2**);
- **C3** glyphs should stand out from the map and each other (**V3**, **V4**).

Reducing the number of variables. The first concern (C1) implies that a number of standard, easy to read glyph designs are not viable. Our collaborators from the Glottolog team were initially very interested in a single scale. We therefore considered various approaches to merge the endangerment and descriptive status into one scale. We explored four possibilities to combine the scales, but eventually concluded that any resulting scale would have many categories, be hard to read and not intuitive for users. Hence we decided to show descriptive status and endangerment status separately, using a hierarchical glyph design.

Sunburst charts. To address C3, it is helpful if glyphs have a basic geometric shape like a square or circle. Since we needed a hierarchical glyph, we decided to use sunburst charts [31], which are well suited for visualizing hierarchies (C1). Parent-child relation is indicated by a neighbor relation of areas instead of using additional shapes such as arrows, which keeps visual complexity low (C2).

There is an ongoing debate about the usage of pie charts and donut charts (see, for example, `https://eagereyes.org/blog/2015/ye-olde-pie-chart-debate`). This debate naturally extends to sunburst charts and raises the question how well users can read them. Skau and Kosara [30] recently conducted a user study on the effectiveness of pie charts and donut charts. While they did not explicitly investigate sunburst charts, their work likely extends to such charts. Skau and Kosara concluded that users perform unexpectedly well in situations where they have to estimate a percentage by area only. The main intent of our glyphs is to let users estimate percentages relative to the languages represented by that glyph. Hence estimating percentage by area is sufficient. Furthermore, nesting is an effective method to indicate a hierarchical relation between variables.

We leave a small hole in the center of our glyphs since otherwise many lines (separating wedges) might meet in a single point, giving a cluttered impression. Skau and Kosara [30] saw no adverse effect from leaving out the center of the chart.

Color palette. To make the glyphs pop out from the map (C3) we chose to use a map layer in grayscale. This contrasts nicely with our color palette for the endangerment status of languages, which is the outer ring of all glyphs. We use the 6-class `YlOrRd` palette from ColorBrewer2.org [13], a scale ranging from yellow to red. Since the color red is associated with danger [24] it is intuitive to use darker red for more endangered languages. After feedback from our expert users we replaced the darkest red with almost black: extinct languages have a special status and need to be very visible.

The descriptive status of languages is displayed in the inner ring, for which we chose the 5-class `Blues` palette, which ranges from gray to blue. We replaced the lightest color of that palette with white, to improve the contrast with our gray map. This indicates the highest level descriptive status, a grammar.

Darker colors in both scales can be associated with a worse status: dark red or black means very endangered while dark blue means not, or barely, described. Thus, color lightness is a meaningful indication of language status. Scanning visually for either descriptive status or endangerment status can then be done by focusing on a specific hue. This should improve performance for tasks T1, T2, T3 and T9.

Drop shadows. We added black drop shadows to all glyphs, again addressing concern C3. We do not suffer from z-ordering issues, since overlap is eliminated by our clustering algorithm. Thus, a shadow is always cast on the map and not on another glyph. The drop shadows are also utilized to indicate selection; selected glyphs have a blue shadow.

4 GlottoVis

We now highlight web functionality of GlottoVis and refer the reader to the GlottoVis about page[1] to learn more.

[1] http://glammap.net/glottovis/about/

The main view of GlottoVis consists of an interactive map with overlaid glyphs, that can be panned and zoomed. Mousing over a glyph opens a tooltip, which shows all languages (truncated to at most five) represented by the glyph, and their language status (see Fig. 3). The tooltips are meant to serve as quick reference, mostly in situations when not too many languages are aggregated in a glyph.

Clicking a glyph reveals a sidebar on the right (see Fig. 1), showing languages sorted alphabetically by name. The sidebar displays for every language its name, language status, brief overview of changes to its descriptive status over time, classification into a language family and the reference to its most extensive description. These all link back to Glottolog, where more details can be found.

Another sidebar, on the left, can be toggled by a button. It contains histograms detailing the statistical distribution of endangerment status and descriptive status of the languages currently in view (tasks T4, T5, and T6). See Fig. 1, left. To provide context, the distribution of all languages in Glottolog is shown faded out in the background. Clicking a bar will apply a filter to the glyphs, so that only languages in the clicked continent are visible. When some of the languages represented by a glyph must be filtered out, they are grouped together and shown in grayscale, in a semitransparent manner and without a drop shadow. This way they blend into the map, but still give the user the appropriate context. Filtering can be stopped by clicking the same bar again. Clicking a different bar will change the filter.

We have associated point locations with continents using data from `GeoNames.org`. This data is not fully accurate, and as a result not all languages are associated with a continent. These languages are shown in the histograms as a bar labeled with a question mark.

We briefly mentioned in Section 2 that there is a temporal aspect to the data. Changes to the descriptive status of languages are tracked in Glottolog. We provide a stream graph view of this data in a sidebar at the bottom, that can be used to filter glyphs. Due to space constraints, we will not further discuss this view.

5 Evaluation & Conclusion

One of the co-authors (a linguist) demoed GlottoVis to various experts during the final symposium of BAULT (Building and Using Language Technology) in Helsinki, Finland. Two independent experts filled out a questionnaire after executing the tasks and answering the questions described in Section 2.2. Additionally, two other independent experts, Matti Miestamo and Tapani Salminen, performed the same tasks and provided more in-depth feedback in separate sessions, which were recorded (screen and audio). The linguists did not receive an extensive training, but were only given a brief instruction.

The glyphs were intuitive to understand for the linguists. It was clear to them how the rings relate to each other and how the size of the glyphs indicates the number of languages represented. Tapani discussed the density of languages, which can be read from the size of glyphs in a region. One linguist remarked that "it is a lot of information to fit on a map" during the demo session, and he was consequently impressed how much he could read at once. The experts used the tooltips frequently to search for languages in a region that they know. They also made extensive use of selection and the links to the Glottolog website. Concerning zooming, they

Figure 3: Languages (and their status) represented by a glyph. Shown when mousing over any glyph.

mostly made use of the extremes and less of intermediate zoom levels. One extreme is an overview of the whole world, which gave some experts surprising insights (South American language status is not as bad as they thought). The other extreme zoom level is used to look at regions they know very well.

The linguists also managed to work with the filtering on continent using the histograms, but here we encountered minor issues such as the lack of an explicit 'undo' or 'go back' option.

Finally, Matti mentioned that he would like to be able to change the glyph size. Especially when zooming in considerably, the size of glyphs could be reduced to see more details on the map. Another extension that was suggested during the demo sessions was the possibility to search for geographical landmarks, such as rivers, regions, and cities.

Conclusion. We have presented GlottoVis, a web-based tool to visualize the data collected in the Glottolog project. The particular challenge of this data set is the fact that two variables (language endangerment and descriptive status) need to be visible simultaneously and in close correlation. Our design uses aggregated hierarchical glyphs on a zoomable geographic map and has various options to search and filter. Visual clutter is reduced by clustering overlapping glyphs; our agglomerative clustering algorithm guarantees that glyphs are truly disjoint.

Future work. We plan to integrate GlottoVis fully into the Glottolog web-page, as the main visual interface to their data. This integration will be coupled with more extensive user testing.

Glottolog also contains relational data. Specifically, for each language it is not only known when and to which degree it was documented, but also which language was used for the documentation (for example, modern Greek was documented into English in 1987). This poses additional visualization challenges that we plan to address in the future.

Acknowledgments

The Netherlands Organisation for Scientific Research (NWO) is supporting B.S. under project no. 639.023.208, K.V. under project no. 639.021.541, and T.C. under project no. 314.99.117.

References

[1] S. Banon. Elasticsearch. `https://www.elastic.co/`, 2013.

[2] K. Been, E. Daiches, and C. Yap. Dynamic map labeling. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):773–780, 2006.

[3] M. Behnisch, G. Meinel, S. Tramsen, and M. Diesselmann. Using quadtree representations in building stock visualization and analysis. *Erdkunde*, 67(2):151–166, 2013.

[4] J. Bertin. *Semiology of graphics: diagrams, networks and maps*. University of Wisconsin Press. Originally in French: *Sémiologie graphique*, 1967.

[5] D. B. Carr, A. R. Olsen, and D. White. Hexagon mosaic maps for display of univariate and bivariate geographical data. *Cartography and Geographic Information Society*, 19(4):228–236, 1992.

[6] T. H. A. Castermans, B. Speckmann, K. A. B. Verbeek, M. A. Westenberg, A. Betti, and H. van den Berg. GlamMap: Geovisualization for e-Humanities. In *Workshop on Visualization for the Digital Humanities (Vis4DH)*, 2016.

[7] B. D. Dent. *Cartography: thematic map design*. McGraw-Hill, 5th ed., 1999.

[8] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008.

[9] Elasticsearch. Geo aggregations. Available online at `https://www.elastic.co/guide/en/elasticsearch/guide/current/geo-aggs.html`.

[10] N. Evans. *Dying words: endangered languages and what they have to tell us*. John Wiley & Sons, 1st ed., 2009.

[11] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

[12] H. Hammarström, R. Forkel, and M. Haspelmath. Glottolog 3.0. Jena: Max Planck Institute for the Science of Human History. Available online at `http://glottolog.org`, 2017.

[13] M. Harrower and C. A. Brewer. ColorBrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. Available online at `http://colorbrewer2.org/`.

[14] J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In *Proceedings of the 13th Symposium on the Interface of Computing Science and Statistics*, pp. 268–273. 1981.

[15] J. Hummel. Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11(1):23–33, 1996.

[16] S. Jänicke, C. Heine, and G. Scheuermann. GeoTemCo: Comparative visualization of geospatial-temporal data with clutter removal based on dynamic delaunay triangulations. In *Computer Vision, Imaging and Computer Graphics – Theory and Applications*, number 359, pp. 160–175. 2013.

[17] B. Johnson and B. Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the IEEE Conference on Visualization*, pp. 284–291, 1991.

[18] M. Lewis. *Ethnologue: languages of the world*. SIL International Publications, 19th ed., 2016.

[19] L. Lins, J. T. Klosowski, and C. Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.

[20] A. M. MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 1st ed., 1995.

[21] R. Mcgill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[22] C. Moseley, ed. *Atlas of the world's languages in danger*. UNESCO Publishing, 3rd ed., 2010.

[23] K. Pearson. *Karl Pearson's early statistical papers*. University Press, 1956. Original article: On the theory of contingency and its relation to association and normal correlation, 1904.

[24] K. Pravossoudovitch, F. Cury, S. G. Young, and A. J. Elliot. Is red the colour of danger? Testing an implicit red-danger association. *Ergonomics*, 57(4):503–510, 2014.

[25] R. E. Roth, K. S. Ross, B. G. Finch, W. Luo, and A. M. MacEachren. A user-centered approach for designing and developing spatiotemporal crime analysis tools. In *Extended abstracts of the 8th International Conference on Geographic Information Science*, 2010. `http://giscience2010.org/indexfea5.html`.

[26] R. Scheepens, H. van de Wetering, and J. J. van Wijk. Non-overlapping aggregated multivariate glyphs for moving objects. In *Proceedings of the 7th IEEE Pacific Visualization Symposium*, pp. 17–24, 2014.

[27] B. Shneiderman. Tree visualization with tree-maps: 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, 1992.

[28] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pp. 336–343, 1996.

[29] G. F. Simons, M. Lewis, and C. D. Fennig. Ethnologue: languages of the world. `https://www.ethnologue.com/`, 2016.

[30] D. Skau and R. Kosara. Arcs, angles, or areas: individual data encodings in pie and donut charts. *Computer Graphics Forum*, 35(3):121–130, 2016.

[31] J. Stasko and E. Zhang. Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the 6th IEEE Symposium on Information Visualization*, pp. 57–65, 2000.

[32] E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, vol. 2. Graphics Press, 1983.

[33] C. Vehlow, T. Reinhardt, and D. Weiskopf. Visualizing fuzzy overlapping communities in networks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2486–2495, 2013.

[34] M. O. Ward. Multivariate data glyphs: principles and practice. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pp. 179–198. Springer Berlin Heidelberg, 2008.

[35] L. Wilkinson. Dot plots. *The American Statistician*, 53(3):276, 1999.