Visualizing Linguistic Diversity in Vancouver

Roger Yu-Hsiang Lo, Namratha Rao, and Anika Sayara
roger.lo@ubc.ca, rao.namratha@gmail.com, sayanika@cs.ubc.ca

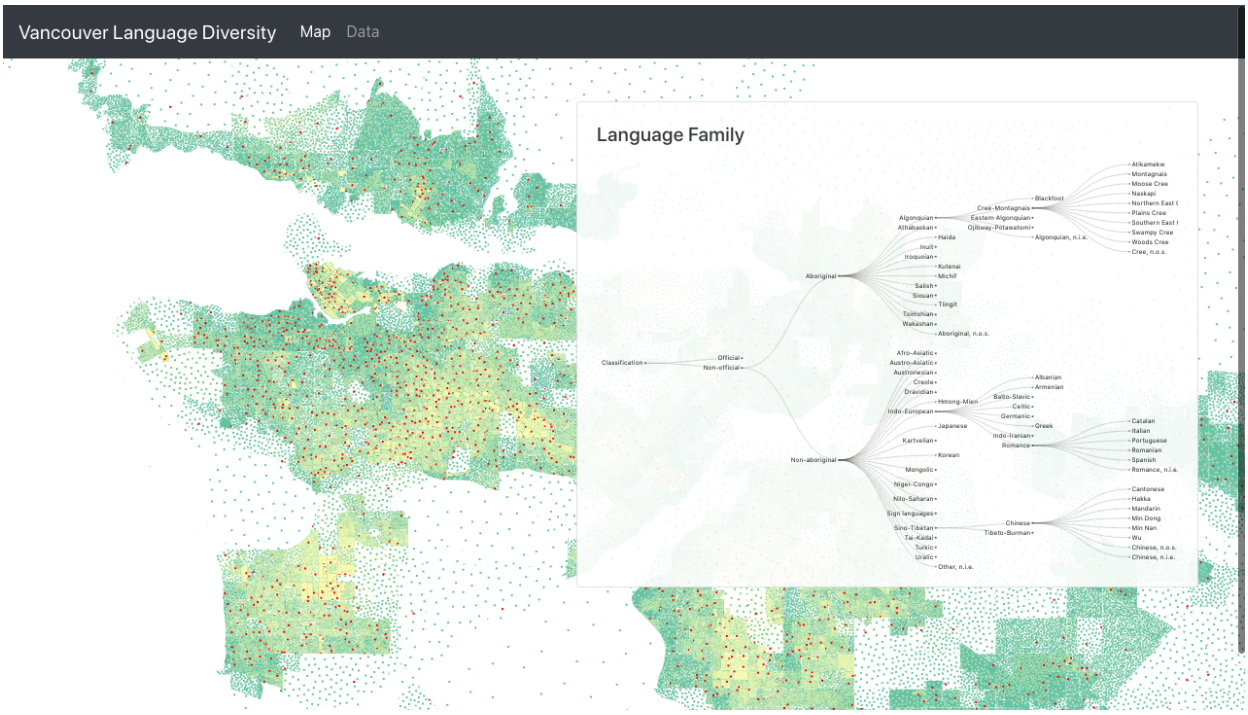

Fig. 1. Overall interface implemented so far

1 Introduction

With two official languages, diverse First Nation languages, and a long and rich history of immigration from a wide range of countries and cultures, Canada has now been home to a large number of languages. In fact, according to a ranking by UNESCO in 2009 [9], Canada ranked 79th in linguistic diversity, only after Belgium, Israel, Norway, Latvia, and Italy among the Organisation for Economic Co-operation and Development (OECD) countries.

This diversity in languages is especially prominent in a handful of Canada's largest cities, with Metro Vancouver being one of the hubs that attract a sizable number of international migrants. With immigrants comes new languages—it has been a common experience for Vancouverites to constantly brush shoulders with fellow residents who can speak a different language that they do not understand. In fact, it is also impossible to find a person in Metro Vancouver who does not have any experience with languages other than English or French.

However, the distribution of languages in Metro Vancouver is not random, as immigrants of the same ethnicity or original country tend to cluster around specific neighbourhoods. Many questions then can be asked about this *nonuniform* language distribution. For example, if one walks in different cities in Metro Vancouver, in which city is one most likely to meet someone who speaks a different language? Or a Bengali speaker might wonder to what extent is his/her language spoken in different neighbourhoods in Metro Vancouver. These questions can be easily answered with census data, but only if the person has substantial skills in extracting and wriggling raw data. The aim of this project is to create a visualization that lowers this bar and allows the user to find answers to these questions, and potentially further explore the linguistic landscape of Metro Vancouver. We hope that the visualization tool developed in this project will lower this bar and be of interest to linguists who want to understand the language distributions in Metro Vancouver and laypeople who perhaps wonder what language the sounds that people around them are constantly making belong to.

2 Related Work

To the best of our knowledge, there has been little work directly addressing the visualization of language diversity. Most of the previous works we identified are mainly from the field of cartography. For example, McNew, Derungs, and Moran [5] overcome the prominent biases in linguistic data and cartographic projection that tend to lead to visual illusions through the Eckert IV projections and Voronoi/Thiessen tessellations to model linguistic areas, instead of the more traditional languages-as-points approach. Luebbering, Kolivras, and Prisley [4] use the language diversity index [2] to study the extent of the presence of different languages in an area, based on the language categories of the U.S. Census.

Geo-spatial datasets are often unstructured. However, the locations, directions, distances, size and altitudes in the dataset can give them an inherent positional structure and shape. Shanbhag, Rheingans, and

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

deJardins [7] deal with the complexity of partitioning geo-spatial data by using a combination of three temporal visualization techniques – wedges, rings and time slices to visualize the changing patterns of population data while maintaining the spatial boundaries of the map. By contrast, Skupin and Hagelman [8] propose trajectory-based techniques that represent change in the attribute as movement of objects across a 2D Self-Organizing Map (SOM) surface to analyze population census data. The spatial location of the data is distorted by the computational changes and the developed patterns are visible. As the quantity and diversity of geo-spatial data increases, analysis becomes difficult due to their size and hidden patterns. Artificial Neural Networks are an evolving solution for data analysis and pattern recognition. Koua [3] also uses a SOM algorithm to uncover the structure and patterns from the geo-spatial dataset consisting of a collection of socio-economic indicators related to municipalities in a region of the Netherlands.He explores visualization techniques like mesh, multiple linked views of component planes, 2D and 3D surface plots of distance matrices that use spatial metaphors such as distances, regions, and scale. Dot maps visualize the geographic distribution of either univariate or multivariate count data. Walker [11] discusses pointillist interactive dot maps, which maps the colour of dots to a set of categorical values allowing the visualization of both the areas of categorical homogeneity and areas of diversity as the colours for different categories blend. He outlines techniques like dasymetric dot mapping, zoom-dependent data and styling, interactive filters, and linked charts to summarize data on the map to improve the user experience for viewers of the dot maps. This drives our design consideration for the interface envisioned.

Our project was inspired by Dmitry Shkolniks' work (https://www.dshkol.com/2017/language-diversity-in-canada/) in trying to identify most linguistically diverse regions in Canada, using the language diversity index. Even though he did provide maps that visualize the results, these maps are all static and only depict aggregated results. Our proposed visualizations expend upon these basic maps and add interactivity that allows the user to retrieve more detailed information.

3 Data and Task Abstraction

The data and task abstraction described below follows the framework detailed in Munzner [6].

3.1 Data Abstraction

Our dataset originates from the 2016 Canada Census, provided in the R package `cansensus` [10]. The extracted dataset is stored in the GeoJSON format and includes a total of 3450 data entries and 230 attributes. Each data entry represent a single dissemination area (DA). In the original dataset, each DA item consists of eight categorical and 222 quantitative attributes. Among the categorical attributes are identifiers, such as CENSUS TRACT IDENTIFIER and CENSUS DIVISION IDENTIFIER, that are irrelevant to the current project and are therefore ignored. Only three attributes—TYPE, REGION NAME, and GEOMETRY—are pertinent, the descriptions and sample values of which are given in Table 1. Four out of the 222 quantitative attributes encode information about the number of households, dwellings, residents, and area in square kilometer in each DA respectively. The other quantitative attributes encode the numbers of speakers of different languages, with each attribute corresponding to one language in principle. More detailed information concerning the quantitative attributes is given in Table 2.

3.2 Task Abstraction

The main users we have in mind are the general public. We envision the users of our visualization come with questions like 'What are the neighbourhoods in Metro Vancouver that are most diverse linguistically?', 'How many languages are spoken in Surrey? And what are the numbers/proportions of speakers of these languages?', or 'How are Korean speakers distributed in Metro Vancouver?'. More abstractly, our visualization aims to support queries in two directions: (i) given a geographic location—a neighbourhood in our case—find out the distribution of languages spoken in this location, and (ii) given one or more languages, find out the geographic distributions associated with these languages.

To abstract our task into a higher level, the goal of our visualization is to provide functionality for the users to *discover features* from a geometry dataset and hopefully to *derive some enjoyment* along the process. As we do not constrain the type of search actions the users can perform, the users should be able to carry out tasks that involve *lookup*, *locate*, *browse*, and *explore*. At the lowest level, the users can retrieve information by *identifying specific targets*, such as finding the number of speakers of a language in a particular neighbourhood, or comparing across targets, for example, by comparing the numbers of speaker of a particular language among several neighbourhoods.

4 Solution

This section is organized as follows: we first describe the overall interface before diving into the two major components of the interface—a dot map and a collapsible tree. We provide justifications for the choice of the two chart types.

4.1 The Overall Interface

The layout of the interface is shown in Figure 1, with a dot map as the centre piece and a movable panel on top that contains a collapsible tree. As addressed in the task abstraction, this interface supports interactivity going in two directions: (i) the users can select a region on the dot map, and the linguistic information of the selected region will be automatically updated in the collapsible tree, or (ii) the users can pick some languages using the collapsible tree, and the geographical distributions of these languages will be depicted on the dot map. Given the task requirements, the two charts not only serve to display the data but also function as *control panels* for data filtering.

4.2 The Dot Density Map

Dot density maps are a technique for visualizing the geographic distribution of either univariate or multivariate count data, where each dot can represent either a single (i.e., one-to-one dot-to-data ratio) or multiple data points (i.e., one-to-many dot-to-data ratio). Typically, dot maps are realized as *pointillist* maps, in which the colour of dots is mapped to a set of categorical values. Pointillist maps allow for not only the visualization of the distribution of a geographic phenomenon but also the consideration of how this distribution varies among subcategories. Pointillist maps are therefore suited to showing the internal heterogeneity of areal units and to illustrating smoother demographic transitions between neighbourhoods.

As mentioned above, the chart has dual functions of displaying the geographical distributions of various languages and providing the interface for the users to select neighbourhoods, of which the distributions of different languages the users are interested in.

When the dot map is used as a selection interface, the users can select a specific neighbourhood by clicking on one of the predefined neighbourhood polygons, and the linguistic information of the selected neighbourhood will be displayed in the collapsible tree, as described in the next section.

When used to display the geographic distribution of (selected) languages, we provide two view modes, termed *overview* and *detail*. In the overview mode, we plot the distributions of *all* languages in Metro Vancouver. Because of the large number of languages spoken in Metro Vancouver, colour coding at the level of individual languages is not practical; some sort of aggregation is needed to map the languages to a smaller number of colours. The language family tree provides an intuitive way to achieve this aggregation. Specifically, we choose six language families with most speakers and only do colour coding at this level. For example, even though there are many First Nation languages spoken in Metro Vancouver, these languages tend to have a very small number of speakers, so we group all First Nation languages together and encode all of them with a single colour. In the detail mode, we allow the users to select specific languages for comparing their geographic distributions. The selection of languages is achieved via the collapsible tree, as will be explained in the following section. We allows the users to select up to 12 languages, as ColorBrewer 2.0

Table 1. Categorical attributes in the dataset.

Name	Description	Cardinality	Sample value
TYPE	Constant value field indicating dissemination area (DA)	1	'DA'
REGION NAME	The name of the neighbourhood the DA is in	25	'Richmond'
GEOMETRY	Polygon geometry of the DA	3450	POLYGON ((-123.28147116478 49.36803241352), ...)

Table 2. Quantitative attributes in the dataset. Note that there are 222 attributes with the name v_CA16_XXXX. We only list one such attribute for illustration.

Name	Description	Min	Max	Median
HOUSEHOLDS	The number of families in the dissemination area (DA)	0	4923	213
DWELLINGS	The number of dwelling units in the DA	0	5631	229
POPULATION	The total number of people in the DA	0	8778	586
SHAPE AREA	Area of the DA in sq. km.	0.002	846.8	0.1
v_CA16_1367	The number of speakers of language 1367 (French)	0	85	0

supports a maximum of 12 data classes for categorical data. After the users selected the languages they are interested in, the dot map will be updated to reflect the distributions of these languages.

4.3 The Collapsible Tree

Given that language families assume a tree structure, a natural way to visualize this hierarchical structure is through a tree diagram, with leaves representing individual languages and non-terminal nodes different language families or branches within a language family. However, due to the depth of the language family trees and to avoid clutter the view, we opt for a collapsible tree, with only a few branches expanded by default based on the number of speakers in the branches. The users can manually expand other branches if they so wish. We implement the collapsible tree in the *tidy tree* layout, as opposed to the cluster dendrogram or the radial layout, because tidy trees are visually more compact.

Again the tree diagram serves two purposes—to indicate the languages as well as the number of speakers of these languages in a selected neighbourhood, and to function as an input interface for the users to specify the languages of which the distributions will be plotted on the dot map. In the former case, when a neighbourhood is selected or when the users first enter the application, the presence/absence of a specific language in the neighbourhood is indicated by the saturation of the coloured nodes and texts, with languages absent being desaturated. The number of speakers of languages present in the selected neighbourhood is encoded by the size of the nodes, with the size of leaf nodes mapping to the number of speakers of individual language and the size of non-terminal nodes the sum of numbers of speakers of languages under the node. To support the function of specifying languages for visualization on the dot map, a checkbox is attached to the name of each language or language branch, and by checking these checkboxes, the users can select a combination of languages and language branches, up to 12 of them, for visualization on the dot map.

5 IMPLEMENTATION

The solution is implemented as a web application, built with HTML, CSS, Javascript, and associated libraries. We use Javascript library D3 [1] for visualization. R is used for data preprocessing.

While it is possible to use D3 to generate dot map using SVG elements, it turns out that generating dot maps this way is computationally expensive and therefore extremely slow. To solve this problem, we turn to Mapbox (https://www.mapbox.com/) for create dot maps. The maps from Mapbox also support zooming and panning.

6 RESULTS

6.1 Scenario 1

A user wants to see how speakers of the Korean language are distributed in Metro Vancouver, so she selects it using the collapsible language family tree. Now she can see how speakers of the Korean language are distributed in Metro Vancouver in the dot map. This also allows her to locate clusters and identify regions which have more Korean language speakers than others.

Now she is wondering how the distribution of Korean speakers compares to that of German and Spanish language speakers, so she expands the tree and selects German and Spanish by clicking on the language text. Now she sees the distribution of all three selected languages in the dot map and can also get a sense the numbers of speakers of the languages by comparing the size of the nodes of these three languages.

6.2 Scenario 2

A user is interested in seeing what languages are the biggest contributors to the total number of speakers of each language family in Richmond. Hence he selects the neighbourhood labelled as 'Richmond' on the dot map by clicking on it. Now he can see the relative number of speakers in various language families in terms of the size of the nodes in the collapsible language family tree.

7 MILESTONE

We plan to spend about 222 hours together towards the project. Table 3 provides a rough estimate of the project's tasks and the actual numbers of hours spent on each task.

REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[2] J. H. Greenberg. The measurement of linguistic diversity. *Language*, 32(1):109–115, 1956.

[3] E. L. Koua. Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets. In *Proceedings of the 21st International Cartographic Conference (ICC)*, pp. 1694–1702, 2003.

[4] C. R. Luebbering, K. N. Kolivras, and S. P. Prisley. Visualizing linguistic diversity through cartography and GIS. *The Professional Geographer*, 65(4):580–593, 2013. doi: 10.1080/00330124.2013.825517

[5] G. McNew, C. Derungs, and S. Moran. Towards faithfully visualizing global linguistic diversity. In *Proc. LREC 2018*, pp. 805–809. European Language Resources Association, Miyazaki, Japan, 2018.

[6] T. Munzner. *Visualization analysis & Design*. CRC Press, 2014.

[7] P. Shanbhag, P. Rheingans, and M. desJardins. Temporal visualization of planning polygons for efficient partitioning of geo-spatial data. In *IEEE Symposium on Information Visualization*, pp. 211–218, 2005.

[8] A. Skupin and R. Hagelman. Attribute space visualization of demographic change. In *Proceedings of the 11th ACM international symposium on Advances in geographic information systems (GIS)*, pp. 56–62, 2003.

[9] UNESCO. *Investing in cultural diversity and intercultural dialogue: UNESCO world report*. UNESCO, 2009.

[10] J. von Bergmann, D. Shkolnik, and A. Jacobs. *cancensus: R package to access, retrieve, and work with Canadian Census data and geography*, 2020. R package version 0.3.2.

Table 3. Project timeline.

Task	Est. hours	Act. hours	Deadline	Description
Proposal writeup	5	5	23 Oct.	—
Update writeup	10	8	17 Nov.	Incorporate feedback from the proposal writeup
Final writeup	25	-	14 Dec.	Finalize paper
Peer project review	5	-	19 Nov.	Slide preparation; presentation time
Final presentation	5	-	10 Dec.	Slide preparation; presentation time
Pre-proposal meeting	1	1	13 Oct.	Meeting note preparation
Post-update meeting	1	-	24 Nov.	Meeting note preparation
Literature review	20	20	1 Nov.	Browse/read relevant papers
Tool familiarization	20	20	8 Nov.	Parallel learning during implementation
Dataset preprocessing	10	15	1 Nov.	[Completed] Data cleaning and attribute derivation
Implementation				
- Interface	10	5	18 Nov.	[In progress] - Region selection - information display - set up overall layout
- Dot map	30	41	22 Nov.	[Completed] - Create map layout - Implement dot density by plotting one dot for every 10 person - Applied Poisson Disc Sampling to evenly space the dots - Reduced loading time to some extent by converting geojson data to topojson with visvalingam algorithm using mapshaper.org [In Progress] - Current implementation takes about 1.5 minutes to load. So we are trying to improve performance by using node stream to support parallel processing while loading the data - Add Mapbox gl.js to support interactions and prevent pixilation when zoomed - If possible plot one dot per person. With current implementation approach memory runs out if we try to plot one dot per person.
- Collapsible tree	30	1	22 Nov.	[In progress] - Create the hierarchy tree for the language families
- Tree maps	20	5	6 Dec.	[In Progress] - Create treemap to visualize proportion of spoken languages within a family The view is occluding, currently experimenting to see if zoomable treemap would be a viable choice.
- Interaction	30	-	25 Nov.	[Not started] - Language selection - Region selection - information display
Total	222	121		

[11] K. E. Walker. Scaling the interactive dot map. *Cartographica*, 53(3):171–184, 2018.