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

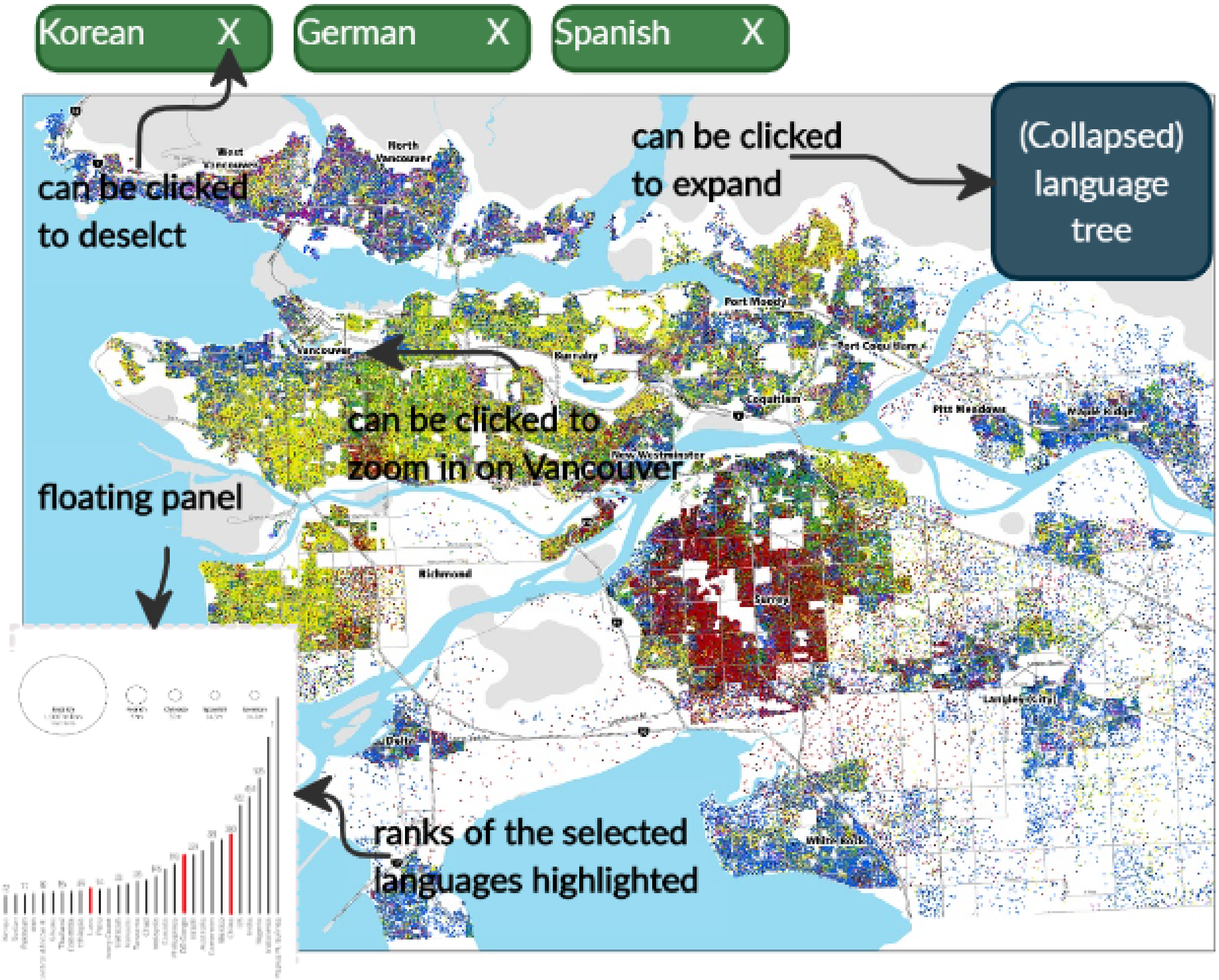

Fig. 1. Overall interface envisioned

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 [5], Canada ranked 79th in linguistic diversity, only after Belgium, Israel, Norway, Latvia, and Italy among 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 neighborhoods. 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 neighborhoods in Metro Vancouver. The aim of this project is to create a visualization that allows the user to find answers to these questions, and potentially further explore the linguistic landscape of Metro Vancouver.

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

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 [4] 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 [3] 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.

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

3.1 Data Abstraction

Our dataset originates from the 2016 Canada Census data. The data are extracted from an R package, `Cansensus` [6]. This tool enables us to take advantage of the built-in vector search and selection tools to easily pull out the data of interest.

The dataset is stored in the GeoJson format and includes a total of 232 attributes, with 3450 data entries. Of these, 218 attributes are quantitative values, each corresponding to the number of speakers of a particular language. The detail of the remaining attributes are outlined in Table 1 and Table 2 according to the attribute type.

3.2 Task Abstraction

As touched upon in Introduction, we envision the user of our visualization comes with questions like 'What are the neighborhoods in Metro Vancouver that are most diverse linguistically?', 'How many languages are spoken in Surrey? And what are the proportions of speakers of these languages', or 'How are speakers of Korean distributed in Metro Vancouver?'. At a slightly abstract level, our visualization aims to provide search functions in two directions: (i) given one or more geographic locations, find out the language distributions in these locations, and (ii) Given one or more languages, find out the geographic distribution associated with these languages.

To abstract our task into an even higher level, our visualization aims to provide functionality for the user 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 user can perform, the user should be able to carry out tasks that involve lookup, locate, browse, and explore. At the lowest level, the user can retrieve information by identifying specific targets (e.g., a language in a particular neighborhood) or comparing across targets (e.g., several neighborhoods where a particular language is spoken).

4 PROPOSED SOLUTION

4.1 Geographic densities and distributions of languages/language families across the neighborhoods of Metro Vancouver

We plan to use a dot density map with color coding (Fig. 2) to visualize the distribution of languages/language families across Metro Vancouver. We chose dot density map for this purpose because it displays density differences and clusters intuitively.

Because there are many languages/language families that are spoken across Metro Vancouver and we are limited to the number of data classes we can represent using color coding, we plan to limit the number of language/language family distributions that can be visualized at once and allow users to make selections. We also plan to aid users to make exploratory selections via a radial language tree. At the moment, we are thinking of a radial tree, as shown in Fig. 3, where each node can be selected and deselected by clicking the small circles of the nodes and an entire branch of the tree can be selected/deselected at once by clicking on its root node. Since we expect this interactive radial tree to take up a lot of space on the screen, we plan to place it on a collapsible panel, which can be expanded when needed.

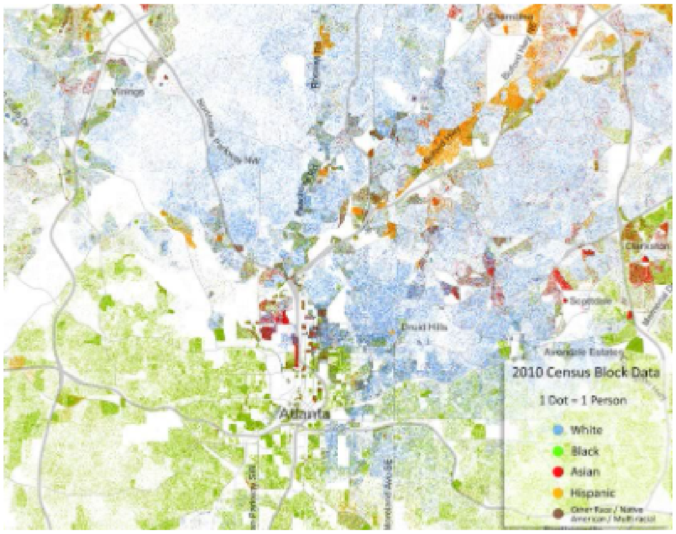

Fig. 2. Mockup of a dot density map for visualizing geographic densities and distribution of languages across Metro Vancouver

In order to support visualization of selected languages, we plan to use categorical color coding. For categorical color coding, the ColorBrewer 2.0 supports a maximum of 12 data classes. Based on this information we can assume that our solution would be able to display at least 12 different languages at once in a dot density map. The same approach is applicable for visualizing selected language families.

In order to visualize distribution of languages within two or more language families, we might want to use categorical color for coding language family and sequential color of the same hue for languages within that family.

In addition to allowing users to make selection on the languages/language families they want to visualize, we plan to also allow them to make region-specific filtering.

4.2 Rank of selected language/language families in a region

In order to visualize the rank, we can make use of bar chart (Fig. 4) and highlight the only the bars of selected language/language families.

4.3 Proportion of language families spoken across Vancouver and proportion of languages in a language family

In order to visualize the proportion of languages in a selected language family we plan to use treemaps (Fig. 5). The size of each rectangle will represent the number of speakers of a particular language in a language family. The treemap visualization is chosen as it allows quick perception of languages that are large contributors in each language family. If time allowed, We aim to implement spatial ordering in the treemap by coloring of absolute position as done in [7].

4.4 The five most popular languages in a selected region

To visualize the five most popular languages in a selected region, we can use size-coded circles like Fig. 6.

4.5 The overall interface

We envisioned the overall interface to be like Fig. 1. Since all our tasks require visualizing relevant data on the map of Metro Vancouver, we plan to place the map at the central focal position of the interface and

Table 1. Categorical data attributes

Name	Description	Cardinality/# of Categories	Sample value
Type	Constant value field indicating dissemination area	1	DA
GeoUID	Geographic identifier	3450	59150004
CSD_UID	Census subdivision identifier	39	5915001
CT_UID	Census tract identifier	461	9330002
CD_UID	Census division identifier	1	5915
CMA_UID	Census metropolitan area identifier	1	59933
Region Name	Region name of DA	25	Richmond
Geometry	Polygon geometry	3450	(POLYGON ((-123.28147116478 49.36803241352,))

Table 2. Quantitative data attributes

Name	Description	Min	Max	Median
Households	The number of families in each DA.	0	4923	213
Dwellings	The number of dwelling units in each DA.	0	5631	229
Population	The total number of people in each DA	0	8778	586
Shape area	Area of the respective DA in sq km	0.00228	846.8001	0.145185
v_CA16_1355	Total number of speakers in each DA	15	8540	580

;

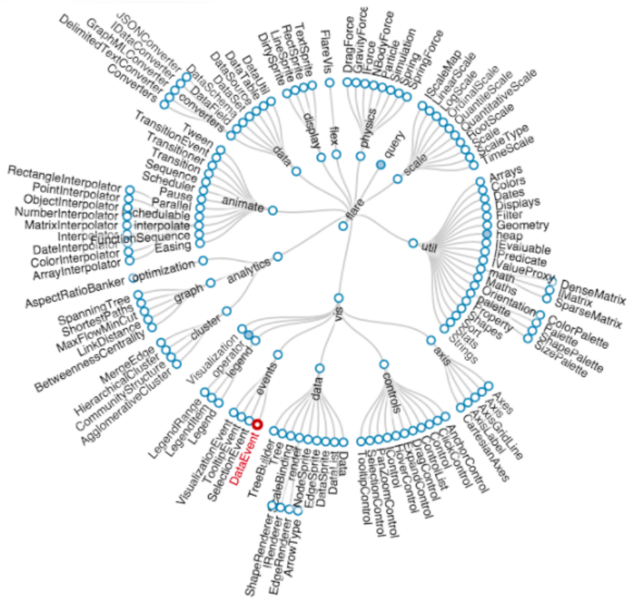

Fig. 3. Mockup of language family tree to aid users make exploratory selections by clicking on the small circles of the nodes

arrange necessary widgets and floating panes around it. The widgets will be used to manipulate "what" to visualize and floating panes to provide additional details about the data being visualized. Additionally, we plan to make the map interactive to support zooming in on cities or neighborhoods.

4.6 Implementation Approach

The solution will be developed as a web application. Therefore, we are going to use HTML, CSS, Javascript and their libraries. The visualization itself is going to be implemented using the D3 library [1]. We will also use the Leaflet library for mapping.

4.7 Results

4.7.1 Scenario 1

A user wants to see how speakers of the Korean language are distributed in Metro Vancouver, so he either searches for the Korean language using the search bar or selects it using the radial language family tree of the selection widget in Fig. 3. Now he can see how speakers of the Korean language are distributed in Metro Vancouver in a dot density map. This also allows him to locate clusters and identify regions which have more Korean language speakers than others. Additionally, he also sees in a floating panel the rank of the Korean language among all the languages spoken in Metro Vancouver (Fig. 4).

;

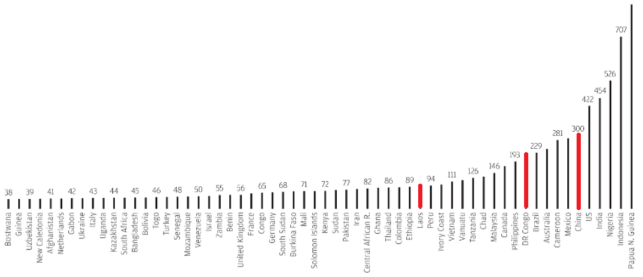

Fig. 4. Mockup to visualize rank of selected languages in a region

Now he is wondering how the distribution of Korean speakers compares to that of German and Spanish language speakers. So he expands the selection widget and selects German and Spanish by clicking on the language family tree. Now he sees the distribution of all three selected languages in the dot density map and can also see the ranks of all the languages in the bar chart (Fig. 4). The bars for Korean, Spanish, and German are highlighted and hence he can easily compare the numbers of these three languages using the bar chart. If he wishes, he can further zoom in on a city to view the same information.

4.7.2 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 a selected region. Hence he switches from the map view to the treemap view (Fig. 5) by clicking on a collapsed pane. Now he can see the relative number of speakers in various language families and also the proportion of each language within each language family.

5 Milestone

We plan to spend about 237 hours together towards the project. Table 3 provides a rough estimate of the project's tasks.

Table 3. Project timeline

Task	Est. hours	Deadline	Description
Proposal writeup	5	23 Oct.	—
Update writeup	10	17 Nov.	
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	13 Oct.	Meeting note preparation
Post-update meeting	1	24 Nov.	Meeting note preparation
Literature review	20	1 Nov.	Browse/read relevant papers
Tool familiarization	20	8 Nov.	Parallel learning during implementation
Dataset preprocessing	10	1 Nov.	Data cleaning and attribute derivation
Implementation			
- Main view 1 (density)	25	15 Nov.	Create map layout; add diversity data
- Control widget 1	10	18 Nov.	Region selection; information display
- Main view 2 (language family tree)	30	22 Nov.	Create the hierarchy tree for the language families
- Control widget 2	10	25 Nov.	Language selection; information display
- Main view 3 (treemap)	50	6 Dec.	Create geographic treemap

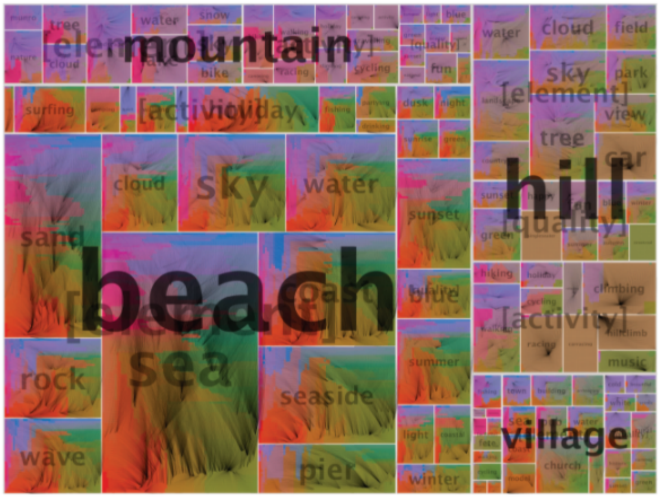

Fig. 5. Mockup for Proportion of language families spoken across Metro Vancouver and proportion of languages in a language family

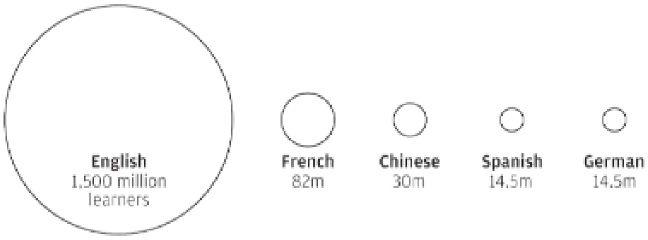

Fig. 6. Mockup to visualize the five most popular languages across regions

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] 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

[4] 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.

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

[6] 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.

[7] J. Wood and J. Dykes. Spatially ordered treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1348–1355, 2008.