Speaking Swiss: Languages and Venues in Foursquare

Darshan Santani
Idiap Research Institute
EPFL, Lausanne, Switzerland
dsantani@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPFL, Lausanne, Switzerland
gatica@idiap.ch

ABSTRACT

Due to increasing globalization, urban societies are becoming more multicultural. The availability of large-scale digital mobility traces e.g. from tweets or checkins provides an opportunity to explore multiculturalism that until recently could only be addressed using survey-based methods. In this paper we examine a basic facet of multiculturalism through the lens of language use across multiple cities in Switzerland. Using data obtained from Foursquare over 330 days, we present a descriptive analysis of linguistic differences and similarities across five urban agglomerations in a multicultural, western European country.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Multilingualism, Foursquare, Linguistic Geography

1. INTRODUCTION

Urbanization is increasing at a faster pace than ever. For the first time in human civilization, a majority of the world's population lives in urban areas. Rural and transnational populations migrate towards cities in search for better opportunities and livelihood. As a result of these trends, it becomes relevant to understand its impact on urban societies, and especially on multiculturalism. Human mobility has been studied extensively in various disciplines. The availability of large-scale digital mobility traces e.g. from *tweets* or *checkins* is providing a new alternative to examine questions that until recently could only be addressed using survey-based research.

It is well established that multiculturalism and diversity promote innovation and creativity [8]. Previous studies have found correlations between the diversity of individuals' social connections and the economic prosperity of their respective communities. In this paper, we are interested in studying a basic facet of multiculturalism using data obtained from Foursquare. Our long-term goal is to investigate the feasibility of LBSN data to facilitate understanding of multicultural trends in cities. One such thread of inquiry could examine communication patterns amongst different ethnic communities and investigate heterogeneity of public places.

In many countries, multiculturalism is expressed directly through language use, both by the local population and migrant communities. In this paper, we take a first step towards characterizing cities through the lens of language used by Foursquare users. Our objective is to characterize basic patterns of multilingualism in cities at a country level. For our analysis, we have chosen Switzerland, that has a well-known geographic diversity of languages and where Foursquare is reasonably popular. Multilingual countries like Switzerland are facing cultural shifts. Swiss media have recently reported that "the language debate is reaching new heights in Switzerland with French and Italian speakers uneasy about the progressive abandonment of their languages by the German-speaking community in favor of English" [3].

This paper addresses three research goals. First, we characterize Swiss cities by their language distribution using place comments. Users, in addition to checking into given places or venues, also leave comments about the place. Comments usually serve the purpose of providing relevant feedback to the Foursquare community. One could potentially extend this research using reviews obtained from local directory services like Yelp or TripAdvisor. Our second objective is to compare these basic patterns of geographically grounded language usage with traditional demographic data sources, so as to assess the adequacy of Foursquare data. And last but not least, to extract basic content trends via topic models. Topic modeling represents a family of statistical methods to discover main themes (or topics) in an unstructured collection of documents [6].

Recent work has proposed a multi-level generative model to infer regional variations of latent topics using geotagged tweets [9]. In [5] the authors have presented a spatio-temporal topic model to discover topics in urban neighborhoods. Using checkin tweets obtained from Foursquare, the paper characterized the topics' spatial coverage and temporal evolution. A study similar to ours was conducted in parallel, to explore linguistic trends across geography in multilingual regions using Twitter [11]. Our research differs from these related work on two specific grounds. First, to capture cultural and linguistic idiosyncrasies better, we have focused our analysis on place metadata as opposed to checkin tweets. Second, the textual content of our dataset span 53 different

MM'13, October 21–25, 2013, Barcelona, Spain.

ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502133.

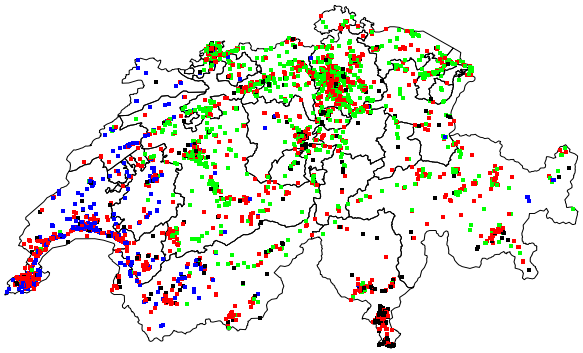


Figure 1: Map showing the spatial distribution of languages used for comments in our data. Each dot denotes a comment, with color depicting its language with red being EN, green is DE, blue is FR and black is IT. Language border is visible across different language-speaking cantons.

languages as opposed to merely English, which is mostly the case in prior literature.

2. DATASET AND PREPROCESSING

2.1 Swiss Checkin Dataset

To guide our analysis, we have collected a place-level dataset (F) for Foursquare venues in Switzerland. We have gathered information from 8,846 users who have visited 26,034 places over a period of 330 days, starting from December 19, 2011. The place-level data includes information on the total number of checked in users, place comments, place category, etc.

We have focused our data collection to be within Switzerland. Geographically, Switzerland is located at the intersection of three major European countries: Germany, Italy and France. As a result, it has found itself at the confluence of linguistic and cultural influences by neighboring countries. It has four national languages: German, French, Italian and Romansh. As per 2000 population census, of all native residents, 63.7% speak German, 20.4% speak French, 6.5% speak Italian, 0.5% speak Romansh, while the remaining 9% speak other languages [4]. Furthermore, it is interesting to note that Switzerland has delineated regions (so called *language borders*, see Figure 1) where one language dominates over the rest. Being a federal republic of 26 cantons, 17 cantons are officially German-speaking, 4 French-speaking, 1 Italian-speaking, with 3 cantons officially declared as bilingual (German, French), while one canton is trilingual (German, Romansh, Italian).

2.2 Comments Dataset

In Foursquare, users are allowed to leave comments (or tips) to visited places. Comments usually serve the purpose of providing feedback to friends and other users. Given that our focus is to analyze user comments, we filter our original dataset F to keep all places where users have left at least one comment (dataset C). As a result of filtering, we are left with 9,010 places with 21,780 comments from 8,748 users in 53 automatically identified different languages, ranging from English to French to Slovak to Zulu. To identify the language used in comments, we have used an off-the-shelf language detection tool [10], which is pre-trained on 97 different languages, but excluding Romansh. (The list of identified languages likely contains false positives.) We perform the language detection at the comment level. Table 1 lists the proportion for the top 4 languages in our comments dataset. Figure 1 shows the comments by its language on the Switzerland map.

Language Used	% of Total Comments	% of Users
English (EN)	58.22	61.28
German (DE)	21.91	20.52
French (FR)	10.68	9.28
Italian (IT)	3.41	3.64

Table 1: Top Languages used for Comments Dataset C

2.3 Comments Translation

Instead of creating a polylingual model (due to the presence of multiple languages), we pursued our analysis in one language. For translation and topic modeling, we have ignored all comments which are not EN, DE, FR or IT. We have chosen English as our base language, and all comments in other languages were translated to English. We have used Systran Home Translator [1] for FR to EN translation, and Linguatec Personal Translator [2] for DE,IT to EN translations.

2.4 Venues as Documents

We treat all comments for each venue as a single text document. In the comment dataset C, more than 88% of places have less than 5 comments, while the average word count for a comment stands at 10.69, which implies that the majority of places have cumulative word count of roughly 50. Given such skewness in comment and word distribution, we define the analysis dataset (A) as a subset of dataset C, where all places having a cumulative word count of comments less than δ are filtered. We have set the value of $\delta = 100$. We also follow the standard text processing pipeline which includes tokenization, lower casing, including words and number only and (English) stopword removal.

3. ANALYSIS

3.1 Users and Languages

We begin our analysis by examining users' dominant language, where dominant language for a user is defined as the one used most often while expressing recommendations on Foursquare. Table 1 highlights the proportion of users who use EN, DE, FR, or IT as their dominant language. Interestingly, the majority of users employ English, while the rest of the language trends follow the official census statistics in terms of ranking (Section 2). In other words, the most common spoken language in Switzerland (German) correctly matches the second most popular language in Foursquare, and the same trend applies to the other top Swiss languages.

The dominance of English in a non-English native speaking country can be explained by multiple factors. First, a significant proportion of Foursquare users might be tourists, some of whom might use English to express themselves. Second, Switzerland is an attractive destination to a significant migrant population (both skilled and non-skilled) who do not speak a Swiss national language and use English as *lingua franca*, including international students and profession-

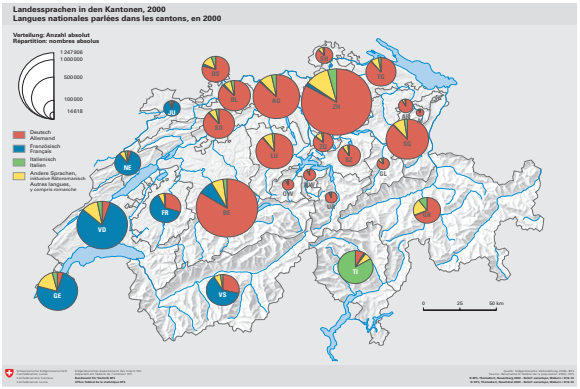


Figure 2: Relative popularity of national languages (DE,FR,IT) across all cantons in Switzerland as per 2000 census [4]. For example, in Ticino, IT is the most popular national language followed by DE, and then FR.

Languages Used	1	2	3	4
% of Users	81.0%	14.77%	3.05%	0.83%
% of Places	72.0%	21.71%	5.00%	0.98%
Language Pairs	EN-DE	EN-FR	EN-IT	DE-FR
% of Bilingual Users	42.57%	18.89%	8.36%	1.32%

Table 2: Users Multilingualism Characteristics

als (so called *expats*). The growing emergence of English is in fact one of the current sources of multicultural tensions in Switzerland [3]. Third, young Swiss natives often speak English and might use it (at least partially) to interact with non-Swiss friends and contacts.

To gain further insight into this matter, we analyze the breadth of language use.Table 2 lists the proportion of users who speak multiple languages on Foursquare. 81% of users adopt one primary language, but we also observe a compelling fraction of users (more than 18%) who use more than one language. These results clearly indicate multilingualism. We further examine the language pairs used by bilingual users. As shown in Table 2, out of 1,292 bi-lingual users, about half of them use a combination of EN and DE, while almost 19% use EN with FR, and 8% use EN-IT pairs. We also observe that more than 15% of bilingual users employ EN in combination with a language other than DE, FR or IT. This indicates that the third factor highlighted above occurs in practice. To validate first and second factors, one could develop a method to identify tourists vs. locals by studying the level of Foursquare activity across space and time, which is the subject of future work.

3.2 Venues and Languages

Now we turn our attention towards understanding the spatial distribution of language use across five cantons (or states) in Switzerland. Of 26 cantons we have chosen Zurich, Geneva, Bern, Vaud and Ticino for our current analysis. The key reasons for selecting them include the presence of cultural and linguistic diversity, in addition to high penetration of Foursquare in these cantons. It is important to note that each canton respectively contains a major urban agglomeration within it. Furthermore, Zurich and Geneva can be considered international cities with substantial expatriate population, while Bern and Lausanne (Vaud) are more local cities with a predominant local resident population.

Canton	EN	DE	FR	IT
Zurich (DE)	4215 (63.20%)	1912 (28.67%)	84 (1.26%)	114 (1.71%)
Geneva (FR)	2429 (71.74%)	37 (1.09%)	654 (19.31%)	61 (1.80%)
Bern (DE)	1091 (53.74%)	735 (36.21%)	58 (2.86%)	38 (1.87%)
Vaud (FR)	1117 (49.06%)	27 (1.19%)	947 (41.59%)	44 (1.93%)
Ticino (IT)	451 (50.79%)	54 (6.08%)	18 (2.02%)	319 (35.92%)

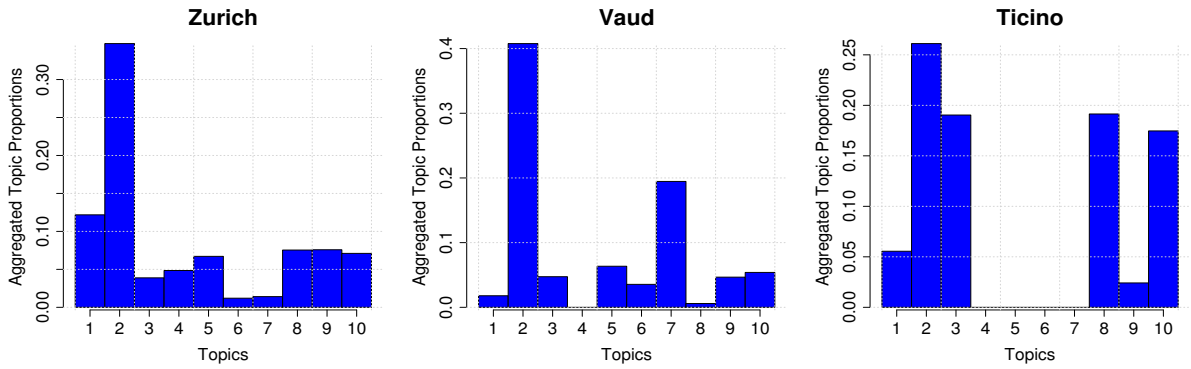
Table 3: Language use across cantons. For each canton, the total number of comments per language is shown along with their percentages (shown in brackets). The official language for each canton is shown in brackets.

Table 3 lists the languages used for commenting in venues across these 5 cantons. Following the analysis in Section 3.1, it is not surprising to find English topping the charts across all regions. More significantly, we observe that the official cantonal language (as per census, see the pie-charts in Figure 2) matches the second most popular language in that canton amongst Foursquare users. For example, Zurich is predominantly a German canton, and we observe German to be the second most popular language amongst Zürchers. This trend is equally observable amongst all other cantons. Furthermore, we highlight that the relative rankings of secondary spoken languages in different cantons (e.g, French or Italian in Zurich) as per the national census, matche with our results in Table 3. Another relevant pattern is that global cities (Zurich and Geneva) have a large proportion of English usage (63-71%) in comparison to more local cities (49-53%). This is specially interesting because it might respond to a different combination of the factors discussed earlier (tourists, migrants, and locals.)

3.3 Topical Analysis

In previous sections, we have explored users' multilingualism and places' spatial distribution of language use across Switzerland. In this section, we examine what users are talking about specifically in these places. We are interested in extracting basic place themes and explore similarities and differences (if any) between places across different regions. Please be reminded that topic analysis is performed on analysis dataset A, which does not take into account the canton differentiation made in the previous section.

To discover topics from user comments, we treat each place as a single document as described in Section 2.4 and apply Latent Dirichlet Allocation (LDA) model [7]. After experimenting with different values for the number of topics and given the document set size (348) and vocabulary count (7,361 tokens after preprocessing), we set the number of topics to be 10. Table 4 lists five key discovered topics along with their most frequent words. Based on manual in-



(a) (b) (c)

Figure 3: Aggregate Topic Proportions for all places in a) Zurich b) Vaud and c) Ticino

T1: airport train swiss ticket hour don't zurich security class chf wi-fi area long buy shop switzerland shopping expensive open trains
T2: excellent beer delicious restaurant geneva eat don't wine super terrace brunch day price people drink small zurich drinks order perfect
T3: pizza italian pizzas pasta pie tiramisu shopping pretty terrible restaurant chocolate vis al store bernerhof made special gnocchi manor luxemburgerli
T7: burger burgers de drink forget fries chips kind generous dont cheese owners backpacker hostel pub guinness romandie big live hamburgers
T8: valid actions rosti salad st price season compared sausage bratwurst veal unzumutbar goods backyard sauber delivery o'clock noise mixed frying

Table 4: Topics discovered for Analysis (A) dataset, along with top 20 frequent words. Out of 10 topics, only topics 1, 2, 3, 7, 8 are shown.

spection, discovered topics conforms to our understanding of Foursquare, in particular in its ability to uncover keywords associated with daily rhythms of city life.

We inspect the spatial distribution of topics to understand topical peculiarities of cantons. For all places belonging to a canton, we aggregate place-topic probabilities for a given topic. Figure 3 shows normalized topic proportions for all topics across 3 cantons. (Other cantons are omitted due to space constraints.) We observe that Topic 2, being a generic topic, is predominant across all cantons. Topic 1 (*transport*) plays a bigger role in Zurich, which is the major transport hub in the country, while topic 7 (*fast food*) in canton Vaud is prominent possibly due to a large student community. For Ticino, being an Italian canton, topic 3 (*italian cuisine*) and topic 8 (*swiss-german food*) are dominant. Ticino is known to be a destination for Swiss Germans. These analyses suggest that the spatial distribution of topics can capture multicultural idiosyncrasies of a city.

4. CONCLUSIONS

In this paper we studied linguistic patterns in Switzerland, using place metadata collected from Foursquare. We found clear evidence of multilingualism. English is the dominant language amongst the majority of Foursquare users, and global cities show a larger proportion of English usage in comparison to more local cities. Another finding is that the relative rankings of secondary languages in different cantons (e.g, French in German-speaking Bern) as per national census data match Foursquare's trends. While obviously Foursquare users do not represent a fair sample of the Swiss population, our findings suggest the potential of Foursquare to complement census data to monitor cultural trends in urban settings. We believe that these results can be equally appealing to other contemporary multicultural nations.

5. ACKNOWLEDGMENTS

We thank Andrei Popescu-Beli (Idiap) for discussions and Thomas Meyer (Idiap) for his assistance in translation. This work was funded by the SNSF through the HAI project.

6. REFERENCES

[1] http://www.systran.co.uk/.
[2] http://www.linguatec.net/.
[3] English challenges multilingual Switzerland. http://bit.ly/LlzaJt. [Online; accessed May, 2013].
[4] Swiss Federal Statistical Office (SFSO), Swiss Statistics Web site. http://bit.ly/VRH1rA. [Online; accessed May, 2013].
[5] S. Bauer et al. Talking Places: Modelling and Analysing Linguistic Content in Foursquare. In *IEEE International Conference on Social Computing*, 2012.
[6] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
[7] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
[8] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 2010.
[9] J. Eisenstein et al. A latent variable model for geographic lexical variation. In *EMNLP*, 2010.
[10] M. Lui and T. Baldwin. langid. py: An off-the-shelf language identification tool. In *ACL*, 2012.
[11] D. Mocanu et al. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PloS one*, 2013.