Social Context and User Profiles of Linguistic Variation on a Micro Scale

Olga Kellert[1] and Nicholas H. Matlis[2]
[1]University of Göttingen, Germany
[2]Center for Free-Electron Laser Science CFEL,
Deutsches Elektronen-Synchrotron DESY, Germany
olga.kellert@phil.uni-goettingen.de and
nicholas.matlis@cfel.de

Abstract

This paper presents a new tweet-based approach in geolinguistic analysis which combines geolocation, user IDs and textual features in order to identify patterns of linguistic variation on a sub-city scale. Sub-city variations can be connected to social drivers and thus open new opportunities for understanding the mechanisms of language variation and change. However, measuring linguistic variation on these scales is challenging due to the lack of highly-spatially-resolved data as well as to the daily movement or users' "mobility" inside cities which can obscure the relation between the social context and linguistic variation. Here we demonstrate how combining geolocation with user IDs and textual analysis of tweets can yield information about the linguistic profiles of the users, the social context associated with specific locations and their connection to linguistic variation. We apply our methodology to analyze dialects in Buenos Aires and find evidence of socially-driven variation. Our methods will contribute to the identification of sociolinguistic patterns inside cities, which are valuable in social sciences and social services.

1 Introduction

Analysis of spatial patterns of linguistic variation is an important tool, not only for studying the dynamics of language change, but also as a probe of social dynamics which can be encoded in linguistic variation. The advent of social media and the growth of computational linguistic tools has created many opportunities for extending analysis of linguistic variation in new regimes. One in particular is the study of spatial or geographical patterns of linguistic varieties. Until now, most studies in this field have been limited to large scales of geographical analysis, from cities to countries (e.g., Eisenstein et al. 2010, Gonçalves & Sánchez 2016, Nguyen et al. 2016, Grieve et al. 2019, Hovy & Purschke 2020). However, urban dynamics, where social interaction and mixing occur, play an important role in language variation and also provide a window into linguistic variation on an urban scale (Abitbol-Levy et al. 2018, Kellert & Matlis 2022). Urban-scale analyses have been previously used to study the relation between urban location and language choice in multilingual cities (Mocanu et al. 2013, Kim et al. 2014). However, these studies focus on different languages and not on linguistic varieties of the same language.

Here we explore the use of Twitter data with precise geolocation information to map out patterns in the use of two linguistic variants (i.e., dialects) within the city of Buenos Aires, in order to get deeper insights into the relation between language use and urban structure. Our basic approach is to combine analysis of tweet metadata with analysis of tweet texts to first show that a large fraction of users in CABA is bi-dialectal (i.e., tweet in both variants) and then to determine how social context influences which dialect is used. Bi-dialectalism is established by exploiting the unique user ID metadata to perform linguistic profiling of the users, while the relevant geographical setting is extracted by using the precise GPS coordinates to pinpoint the location. The tweet texts then provide information on the associated topics of discussion which helps to complete the social-context picture.

Our work shows that combining social and geographical aspects of linguistic analysis, made possible by social-media data sources, opens new

Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 14–19
October 16, 2022. ©2022 Association for Computational Linguistics

opportunities to illuminate the mechanisms driving linguistic variation, especially on urban scales. Our analysis also helps to establish how well dialect-use patterns in social media conform to those in standard linguistic sources and provides complementary information about the users not easily accessible by other means.

2 Identification of the Dialects

We selected two Spanish varieties: Argentinian Spanish (ArgSp) and Standard Spanish (PanSp). These varieties are marked by the variation in the 2nd person singular pronoun (i.e., *vos* 'you' in informal address in ArgSp and *tú* 'you' in PanSp) together with the corresponding verbs that agree with the pronoun (e.g., *vos podés* vs. *tú puedes* 'you can') (Fontanella de Weinberg 1999). We use geolocated tweets in Spanish limited to Greater Buenos Aires (CABA), i.e., the city of Buenos Aires and its close surroundings, from October 2017 to March 2021 (Kellert & Matlis 2022). The selection of this city is motivated in §3.

We use a token-based analysis method to extract the two linguistic variants (Gonçalves & Sánchez 2016, Grieve et al. 2019, Kellert & Matlis 2022). This method is a classical method in social dialectology (Labov 2006). For clarity, we here refer to the Spanish varieties ArgSp and PanSp as Spanish *dialects*. However, since these varieties can be used by the same group of people under different social circumstances, as we will show in §4, one can also refer to them as *sociolects*.

A priority was placed on ensuring accuracy of the dialect definitions. The token sets were designed to be balanced, so that for each token of ArgSp, there was a corresponding token with the same meaning in the set representing PanSp (Kellert & Matlis 2022). Our grammatical token set consists of the most frequent tokens used in Argentina according to the corpus *Corpus del Español*, which is one of the biggest Spanish corpora.[1] We excluded all ambiguous tokens (e.g., ArgSp *seguí* 'follow!', which corresponds to PanSp *sigue* 'follow!', but also to 'he/she/it follows' in both dialects). Finally, we take special measures to account for differences in how people use accents in social media and standard language (Nguyen et al. 2016). In particular, accents in Tweets are frequently omitted (Eisenstein et al. 2010). We therefore included both accented and unaccented versions of all verbs but excluded those verbs where eliminating the accent results in an ambiguity in assigning the dialect (e.g., ArgSp *sabés* vs. PanSp *sabes*). The remaining verbs were distinguishable by the verb stem (e.g., ArgSp *tenés* vs. PanSp *tienes*). Our final token list contained 235 tokens for ArgSp and 198 tokens for PanSp dialect (Kellert & Matlis 2022).

3 Background on the Dialects

A little background on the dialects will help us to evaluate the results. The dialects ArgSp and PanSp have well-known and distinct historical and socio-linguistic roles in CABA (Fontanella de Weinberg 1999). ArgSp is the most prominent and is also the standard dialect in Argentina, Paraguay, Uruguay and Central America (ibid.). PanSp, on the other hand, is a variety that is very prominent elsewhere in the Spanish-speaking world. This distinction between the two varieties has previously been reported on the basis of geolocated tweets collected in 2016 (Bland & Morgan 2021), and here we confirm it using our tweet corpus by mapping out the differential distribution of the two tweet varieties on the world scale (see Figure 1).

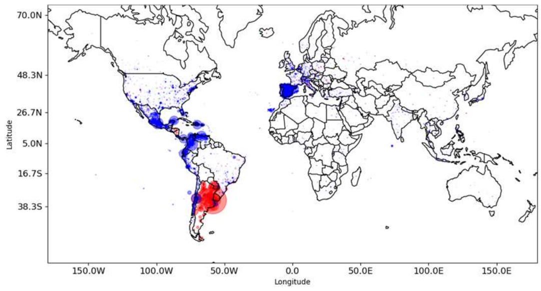


Figure 1: *ArgSp (red) and PanSp (blue) in the Twitter corpus collected from 2017-2021*. Map produced using *Cartopy*[2] on OpenStreetMap[3] data.

Despite the predominance of ArgSp in Argentina and CABA on the world scale, a closer inspection shows significant presence of both varieties within CABA (Kellert & Matlis 2022) which is to be expected for several reasons. First, PanSp is the Spanish variety that is used by Mass Media in Latin America and consequently also in CABA (Gonçalves & Sánchez 2016). Second, PanSp is used by tourists from PanSp speaking countries. And third, PanSp was long the standard variety in CABA and Argentina, before ArgSp took over this

[1] http:// www.corpusdelespanol.org/
[2] https://scitools.org.uk/cartopy/
[3] http://wiki.openstreetmap.org/wiki/Open_Database_License

role in the late 19th c. (Fontanella de Weinberg 1999). PanSp still exists in CABA as a substandard variety due to the region's history (ibid).

4 Detailed Methodology and Results

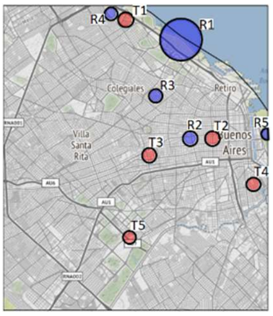
Our analysis, which is based on calculation of the spatial variations in the use of ArgSp and PanSp across the city, is done in three steps. First, the prevalence of the two dialects and the degree of bi-dialectalism are quantified. The presence of bi-dialectalism in CABA offers the opportunity to directly evaluate how social circumstances influence linguistic variation, since bi-dialectal users can choose when to use each variant. Second, regions where each dialect is most prominent are evaluated to look for correlations between social context and tweet content. And third, locations expected to have specific social functions (e.g., soccer stadiums) are evaluated to look for prominence of one or the other dialect.

4.1 ArgSp & PanSp vs Bi-Dialectalism

The observation in §3 that ArgSp is the predominant dialect of CABA is confirmed by the fact that ArgSp tweets outnumber PanSp tweets three to one (i.e., 18,731 vs 5,607 respectively). However, by using the unique user-ID metadata to associate multiple tweets to individual users, we found that a considerable number of CABA users (11%) are "bi-dialectal" in that they tweet using both dialects. Some users even mix the two dialects in a single tweet (e.g., *vos puedes* or *tú podés* 'you can'). The existence of bi-dialectal users and the existence of mixing dialects in a single tweet suggests that PanSp plays an important role in communication of CABA citizens and that it is not exclusively used by people of foreign background such as tourists or immigrants. However, ArgSp is a more important variety than PanSp because bi-dialectal users tweet twice as often in ArgSp as in PanSp (6,299 vs 3,040 respectively) and because there are more tweets posted by mono-dialectal users than by bi-dialectal users (14,999 vs. 9,339, respectively). The latter observation indicates that tweeting in both linguistic varieties is not the standard tweeting behavior of CABA citizens.

4.2 Analysis 2: Dialects in Geocontext

In this analysis, we focus on regions with the greatest prominence of each of the dialects to look for correspondences between dialect use and social context. The regions of prominence are determined by generating spatial distributions of each variant calculated by partitioning the city into small areas or "bins", corresponding roughly to the size of a city block, which define the spatial resolution of the maps (Schlosser et al. 2021, Kellert & Matlis 2022). We then selected five bins with the greatest prominence of each variant, based on the normalized difference in tweet counts (Kellert & Matlis 2022), and examined the associated geographical setting and tweet content (Figure 2).



Bin	GPS Coordinates
T1:	-58.4130, -34.5570
T2:	-58.4068, -34.6113
T3:	-58.4298, -34.5879
T4:	-58.4598, -34.5428
T5:	-58.3556, -34.6088
R1:	-58.4501, -34.5461
R2:	-58.3918, -34.6113
R3:	-58.4342, -34.6205
R4:	-58.3644, -34.6364
R5:	-58.4474, -34.6657

Figure 2: *Left: Bins with the greatest representation of PanSp (blue) and ArgSp (red). Right: GPS coordinates of the bins.*

The geographical setting ("geocontext") was determined by using the GPS coordinates to identify buildings located within each bin, and the tweet-content was evaluated by performing uni-gram and bi-gram analyses of the tweet texts from each bin, using the software package NLTK, to determine the most frequent words and word pairs. (Bird et al. 2009). The results of the analysis for the most prominent bins of each type (bins T1 and R1 in Figure 2) are shown in Table 1.

Geocontext	Tweet-content features
PanSP Jorge Newbery International Airport	• location names mentioning CABA • Picture postings in CABA • Weekly horoscope • travel club postings • happy birthday wishes • good/happy day/night wishes
ArgSP Soccer stadium: Estadio "Monumental" Antonio V. Liberti	• location names mentioning the soccer stadium • reports about soccer matches • sentiments about soccer matches and players

Table 1: *Geographical context and tweet content features in most prominent bins for each dialect.*

The most prominent bin for PanSp covers the international airport in the northern part of the city (Figure 2, R1), and the associated tweets mention CABA in various forms (e.g., “Buenos Aires, BsAs”) as well as picture postings and other references to travel. By contrast, the most prominent bin for ArgSp covers the famous soccer stadium River Plate (Figure 2, T1), and the associated tweets refer to this stadium and discuss the matches and players.

The remaining eight most-prominent bins show a similar pattern. For the PanSp bins, the geocontext includes Irish bars, tourist attractions and wealthy neighborhoods in the northern part of the city which are attractive to tourists, while all of the ArgSp bins contain geocontext features relevant to locals of CABA such as soccer clubs, soccer stadiums such as La Bombonera (Figure 2, T4), dance schools, small commercial businesses and residential buildings in neighborhoods such as *Villa Soldati*, located in the South-West of the city. Similarly, all of the PanSp bins have associated tweets with location mentions and photo postings as top-ranked topics whereas none of the ArgSp bins do. Mentioning the name of the city and posting pictures are typical activities of tourists or of users addressing tourists (e.g., in advertisements of touristic attractions) (Kim et al. 2014). The analysis therefore suggests that tourism plays an important role in the use of PanSp dialect in CABA and that national sports clubs, local commerce and residential buildings tend to prioritize ArgSp.

4.3 Analysis 3: Dialects in Social Contexts

We have chosen several social contexts defined as bins containing buildings of a selected, well-defined social function. The building types chosen were: 1) tourist attractions, 2) Starbucks cafes, 3) soccer stadiums, and 4) hospitals. We then counted how many bins of each type demonstrated a relative prominence of ArgSp vs PanSp.

In Figure 3, maps for each category are presented showing bins with a relative prominence of ArgSp and PanSp marked by red and blue circles, respectively. Empty bins containing no tweets of either type are marked in grey. For the tourist attractions, 52% of the non-empty bins showed a relative prominence of PanSp, while for Starbucks Cafes, 59% of non-empty bins are PanSp oriented. While these numbers are far from conclusive, due to the sparse statistics, the pattern is consistent with the connection, observed in §4.2, between PanSp and tourism, if one accepts that tourists are likely to visit Starbucks cafés. For the soccer-stadium case, 100% of non-empty bins demonstrate a relative preponderance of ArgSp, which also reinforces the connection between ArgSp and soccer found in in §4.2. Finally, hospital bins showed no preference for either dialect.

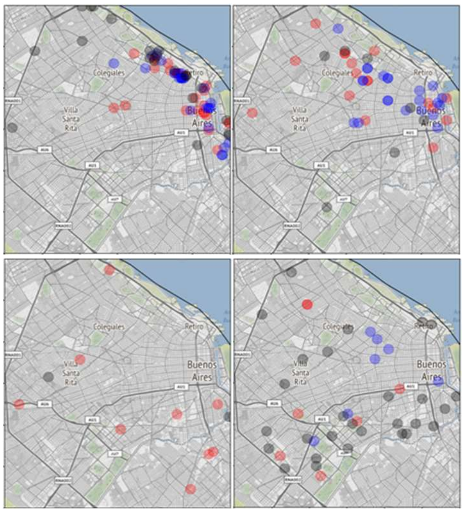


Figure 3: *Distribution of bins with a relative prominence of PanSp (blue) and ArgSp (red) tweets for specific social contexts*. Top left: touristic hotspots, Top right: Starbucks cafés, Bottom left: soccer stadiums, Bottom right: Hospitals. Black= no data

5 Discussion

The three analyses presented show ways in which different elements of the tweets can be combined to extract valuable information about the user’s linguistic behavior and about relevant geographical and social contexts that can influence this behavior. The ability to connect multiple tweets to individual users via the user ID is a powerful tool enabling determination of user attributes such as bi-dialectalism which are unavailable via textual analysis alone. Similarly, the presence of precise GPS coordinates allows connections to be made between text and local geographical features that can be used to characterize the associated social contexts. In the work presented here, although evidence of a pattern is present, some analysis was done manually, leading to small data sets and low statistical significance. Development of algorithms to detect geocontext features and characterize tweet content would allow automation of bin analysis, greatly increasing the statistics and hence the strength of the approach.

Several other important issues must also be considered in going forward. First, despite the large size of our Twitter corpus (1.9M geolocated tweets), the quantity of data was a limiting factor for analyzing the small scales. Considering a binning of 100 x 100, one can expect only 190 tweets per bin, on average. Of course, due to spatial variations in population density, the number of tweets in the city center are far higher than those on the outskirts. Data scarcity is aggravated by the small fraction (~1%) of tweets for which the variants could be identified. Methods to improve variant identification are thus highly relevant. For instance, the number of tokens (and hence the number of collected tweets) could likely be increased by using a morphological tagger such as FreeLin[4]. However, since precise identification of the variants was our top priority, we opted for a manually-crafted set of tokens for which the lack of ambiguity could be verified.

Second, the social context within the bins may not be uniform. For instance, in Analysis 3, the Starbucks cafes represented only one of several buildings within the bins and therefore may not have been representative of the overall social context. By contrast, soccer stadiums, which are much larger, are more likely to fill an entire bin, thus providing a uniform social context. This increased context uniformity may partially account for the strong correlation observed between soccer stadiums and ArgSp dialect in Figure 3, bottom left.

Larger bins tend to encompass a greater diversity of social contexts lessening the degree of correlation between the chosen context and the prevalence of specific linguistic features. On the other hand, smaller bins tend to suffer from insufficient numbers of tweets, requiring optimization of the bin size. A similar problem arises, due to the lack of altitude information in most social-media GPS coordinates, when the bins contain multi-story buildings with different businesses on each floor.

The work presented here represents only a first step in the application of this methodology. Many opportunities exist for future work, including use of tweet-selection methods that do not rely on specific tokens (Nguyen et al. 2016) and expansion of the range of social contexts considered. These tools hold great promise to provide insights into the relation between language use and social dynamics, especially on small spatial scales.

References

Jacob Abitbol-Levy, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, and Eric Fleury. 2018. Socioeconomic dependencies of linguistic patterns in twitter: A multivariate analysis. In *WWW '18: Proceedings of the 2018 World Wide Web Conference*. Association for Computing Machinery Inc, pages 1125–34.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly, Sebastopol.

Justin Bland and Terrel A. Morgan. 2021. Geographic variation of *voseo* on Spanish Twitter. In Diego Pascual y Cabo and Idoia Elola (eds.), *Current Theoretical and Applied Perspectives on Hispanic and Lusophone Linguistics*, pages 7–38. John Benjamins, Amsterdam.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1277–87.

María Beatriz Fontanella de Weinberg. 1999. Sistemas pronominales de tratamiento usados en el mundo hispánico. In Violeta Demonte and Ignacio Bosque (eds.), *Gramática descriptiva de la lengua española,* vol 1., pages 1399–1426. Espasa Calpe, Madrid.

Bruno Gonçalves and David Sánchez. 2016. Learning about Spanish dialects through Twitter. *Revista Internacional Lingüística Iberoamericana*, 14(2):65–75.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence,* 2(11). https://doi.org/10.3389/frai.2019.00011.

Dirk Hovy and Christoph Purschke. 2020. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. Association for Computational Linguistics, pages 4383–94. https://doi.org/10.18653/v1/D18-1469.

Olga Kellert and Nicholas Matlis. 2022. Geolocation of multiple sociolinguistic markers in Buenos Aires,

[4] https://nlp.lsi.upc.edu/freeling/

PLoS One 17(9):e0274114. https://doi.org/10.1371/journal.pone.0274114.

Suin Kim, Alice Oh, Ingmar Weber, and Li Wei. 2014. Sociolinguistic Analysis of Twitter in Multilingual Societies. In *Proceedings of the 25th ACM conference on Hypertext and social media.*

William Labov. 2006. *The social stratification of English in New York city*. 2nd ed. Cambridge University Press, Cambridge, UK.

Guy Lansley and Paul A. Longley. 2016. The geography of Twitter topics in London. *Computers Environment Urban Systems*, 58:85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002.

Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping World Languages through Microblogging Platforms, *PLoS ONE* 8(4):e61981. https://doi.org/10.1371/journal.pone.0061981.

Dong-Phuong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska M.G. de Jong. 2016. Computational sociolinguistics. A survey, *Computational Linguistics,* 42(3):537–593.

Stephan Schlosser, Daniele Toninelli, and Michela Cameletti. 2021. Comparing methods to collect and geolocate tweets in Great Britain. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(1):1–20.