




Downloaded from <https://academic.oup.com/dsh/advance-article-abstract/doi/10.1093/dsh/fqy049/5113152> by Huilib user on 01 October 2018

Anna Hausmann , Henrikki Tenkanen , and Tuuli Toivonen 
Digital Geography Lab, University of Helsinki, Finland,
Department of Geography and Geosciences, University of Helsinki,
Finland and Helsinki Institute of Sustainability Science, University
of Helsinki, Finland

This article explores the linguistic landscape of social media posts associated with specific geographic locations using computational methods. Because physical and virtual spaces have become increasingly intertwined due to location-aware mobile devices, we propose extending the concept of linguistic landscape to cover both physical and virtual environments. To cope with the high volume of social media data, we adopt computational methods for studying the richness and diversity of the virtual linguistic landscape, namely, automatic language identification and topic modelling, together with diversity indices commonly used in ecology and information sciences. We illustrate the proposed approach in a case study covering nearly 120,000 posts uploaded on Instagram over 4.5 years at the Senate Square in Helsinki, Finland. Our analysis reveals the richness and diversity of the virtual linguistic landscape, which is also shown to be susceptible to continuous change.

tuomo.hiippala@helsinki.fi

Geotagged social media content also holds potential for sociolinguistic inquiry. In this article,

we adopt the term virtual linguistic landscape, which Ivkovic and Lotherington (2009) coined for discussing multilingualism on the web, to describe the languages present in geotagged social media content posted from a specific geographic location. We propose that the virtual linguistic landscape may be considered an extension of the physical linguistic landscape in the built environment. To explore the characteristics of virtual linguistic landscapes, we analyse nearly 120,000 posts uploaded on Instagram from the Senate Square in Helsinki, Finland, over a period of 4.5 years. We seek to answer the following research questions:

- (1) How to characterize virtual linguistic landscapes in terms of their linguistic richness and diversity?
- (2) How do virtual linguistic landscapes change over time?

Given the high volume of data, we adopt methods from the field of natural language processing, namely, automatic language identification and topic modeling. To measure linguistic richness and diversity, we use established indices from the fields of ecology and biology, which have been previously applied to the study of linguistic landscapes (Peukert, 2013; Manjavacas, 2016). We also perform temporal analyses at various timescales to examine changes in the virtual linguistic landscape. We do not, however, seek to compare or make claims about the respective characteristics of virtual and physical linguistic landscapes (cf. Deumert, 2014a, pp. 117–18). Instead, we aim to develop methods for studying high volumes of geotagged social media content, setting the stage for approaches involving mixed methods, which are ultimately necessary for achieving a comprehensive view of virtual linguistic landscapes.

2 Physical Places and Virtual Spaces

Androutsopoulos (2014) has observed that new sources of data for sociolinguistic inquiry are currently emerging at the intersection of research on computer-mediated communication (CMC) and linguistic landscapes. Whereas CMC covers private

and public communication in digital media, such as social media platforms, discussion forums, and email, the research on linguistic landscapes focuses on “signs and other artifacts in public space” (Androutsopoulos, 2014, p. 75, our emphasis). These definitions may reflect an emerging division of work between the aforementioned domains of sociolinguistic research, as the study of linguistic landscapes has traditionally focused on built environments, covering various locations ranging from tourist attractions (Bruyèl-Olmedo and Juan-Garau, 2015) to transportation hubs (Soler-Carbonell, 2016) and various media from billboards to shop signs (Gorter, 2013).

At the same time, the broader notion of public space, which Androutsopoulos (2014) assigns to the domain of linguistic landscapes, has been and continues to be transformed by digital technology in the form of both hardware and software (Dodge and Kitchin, 2005). In the field of human geography, one of the leading theorists of this transformation is Aharon Kellerman (see Kellerman, 2010, 2016), who has argued that mobile devices have enabled the emergence of a “double space” of intertwined physical and virtual spaces (see also Zook and Graham, 2007). This double space now increasingly envelopes its subjects, as access to the virtual space is no longer restricted by limitations arising from static hardware in the physical space, such as desktop computers.

Due to the increased potential for spatial mobility, this double space can now fill or support many basic human needs, including those originally defined by Abraham Maslow (Kellerman, 2014). For example, needs pertaining to esteem, such as status and reputation, are increasingly formed in virtual spaces (Kellerman, 2014, p. 542). Kellerman (2010, p. 2993) identifies multiple connections between the physical and virtual spaces, which are grouped along several dimensions: organization, or how such spaces are structured; movement, or the connections between spaces; and users, who populate these spaces. Two specific connections warrant further attention, namely, the convergence of physical and virtual places, and the languages encountered in virtual spaces, as both shape the virtual linguistic landscape.

created at the time of upload, as exemplified by the practice of posting content related to previous events under hashtags such as #throwback. Similarly, the content associated with a specific virtual location must not be necessarily created at the actual physical location.

Second, in terms of their linguistic characteristics, [Kellerman \(2010\)](#) suggests that physical spaces are characterized by domestic languages, whereas virtual spaces are dominated by English due to their international orientation. [Lee \(2017, p. 16\)](#) has observed that assumptions about the dominance of English in virtual spaces have been common among both academic and popular audiences ever since Internet became widely used. Yet measuring the actual linguistic diversity of virtual spaces remains a challenge ([Paolillo, 2007](#)), which is also affected by how such virtual spaces are defined and delimited ([Leppänen and Peuronen, 2012](#)). However, the current consensus seems to be that languages other than English are becoming increasingly prominent on the Internet ([Lee, 2016, p. 118](#)).

In virtual spaces, the linguacultural make-up of users has the potential to be extremely diverse, because online interactions do not require physical presence, but allow participation from distance, as illustrated in Fig. 1. Moreover, users may choose to use different languages for different audiences (Androutsopoulos, 2015). It is also important to acknowledge that online interactions can be asynchronous and unfold over longer periods of time. Moreover, not all social media content is necessarily

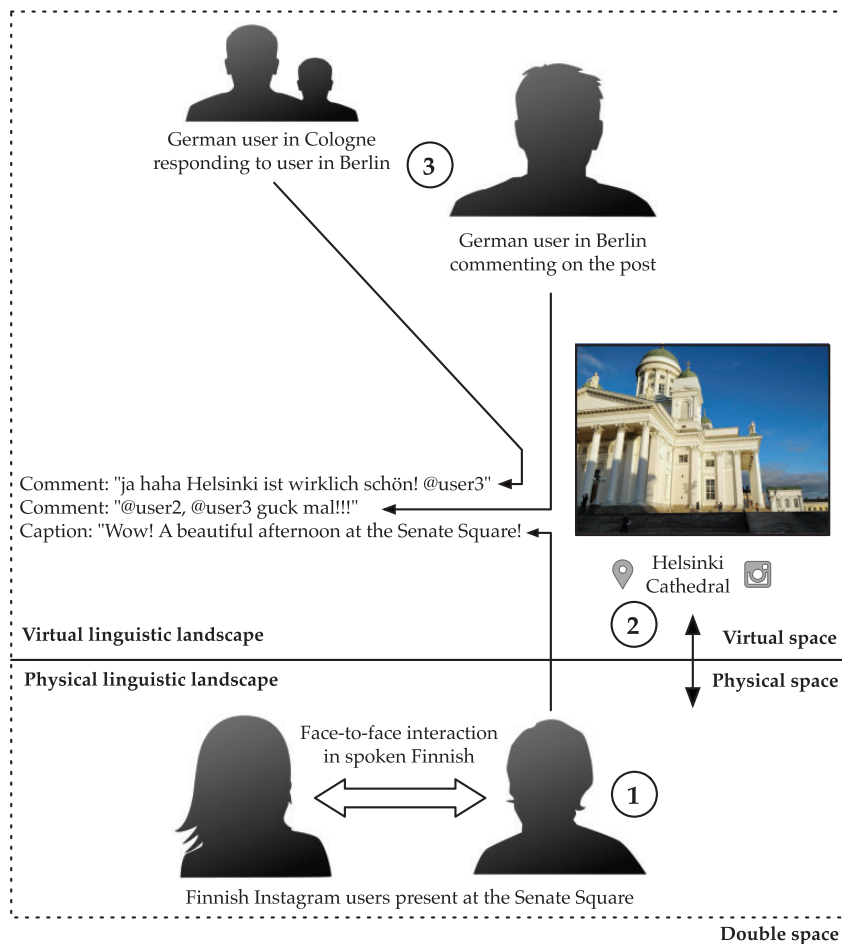


Fig. 1 A fictional example showing how (1) two Finnish users at the Senate Square speak Finnish with each other, but the other posts a photograph with an English caption on Instagram, having a number of international users in her social network. (2) Associating the photograph with the location named Helsinki Cathedral allows a German user who searches for content from Helsinki to discover the photograph. (3) Despite physical distance, German users can interact with the content and each other, contributing to the virtual linguistic landscape of the Senate Square. Each step in this chain of events involves language choices, which all contribute to the virtual linguistic landscape

3 Social Media Data and Computational Methods

3.1 Data and location

We collected data from Instagram,¹ a social media platform for sharing photographs and short videos, using the platform's application programming interface (API). In total, we collected 117,418 posts uploaded by 74,051 unique users between 4 July 2013 and 11 February 2018, that is, over a

period of roughly 4.5 years. As illustrated in Fig. 2, each geotagged post on Instagram is associated with a specific location pre-defined on the platform, which means the geographic coordinates of an individual data point do not provide GPS-level accuracy, unlike some other platforms, such as Twitter and Flickr.

Instead, the geographic coordinates associated with an Instagram post refer to what is commonly termed a point-of-interest (POI) in the field of



Fig. 2 Social media platforms such as Instagram (1), Twitter (2), and Flickr (3) all allow users to embed geographic metadata into their content at various degrees of accuracy from GPS coordinates to POI locations defined by the platforms

geoinformatics (Hochmair *et al.*, 2018). Instagram POIs are provided by the parent company, that is, Facebook. The response to any spatial query is therefore restricted to content associated with a POI on the platform. In our case, each post retrieved for the study was geotagged to a POI located within a 150-m radius from the point 60.169444 latitude and 24.9525 longitude (WGS-84), which lies at the centre of the Senate Square in downtown Helsinki, Finland.

We chose the location due to its status as a cultural landmark and a touristic attraction, which are likely to be reflected in its virtual linguistic landscape. Overlooked by the Lutheran Cathedral and surrounded by the main building of the University of Helsinki and the Government Palace, the Senate Square and its neoclassical architecture are widely recognized as one of the most important landmarks in Helsinki and in entire Finland. The Lutheran Cathedral, in particular, which is shown in Fig. 2, is often used as a symbol for the city of Helsinki (Jokela, 2014). In addition to its role as a touristic

attraction, the Senate Square serves as a venue for different events, ranging from concerts and festivals to protests and demonstrations.

3.2 Identifying the language of social media content

Like many other forms of digital data, geotagged social media content may be characterized as ‘big’ due to its high volume, velocity, and variety (Kitchin, 2013). Together, these characteristics present several challenges for the collection, processing, and analysis of social media data. Challenges related to volume and velocity may be met by adopting a programmatic approach, that is, collecting data systematically via an API and processing the data accordingly (see Tenkanen, 2017, p. 22). For mapping the languages that make up the virtual linguistic landscape, further processing involves automatic language identification, which is an active area of research within the broader field of natural language processing (Zubiaga *et al.*, 2016).

Automatic language identification, however, is not a straightforward task due to the variety of the data, which in this case takes the form of linguistic variation. Much has been written about the language of social media in recent years, revealing variation across different linguistic structures (see Zappavigna, 2013; Seargeant and Tagg, 2014; Hoffman and Bublitz, 2017). On a more practical level, the length of social media posts is typically limited, which encourages the use of abbreviations, non-standard spellings, and other forms of creative language use (Carter *et al.*, 2013, p. 196). Another challenge emerges from the use of hashtags, which are used to affiliate around shared values or topics (Zappavigna, 2011). Hashtags are often written in multiple languages (Barton, 2018; Lee and Chau, 2018), which injects multilingual material into otherwise monolingual texts. The same holds true for usernames on social media platforms.

Each of the aforementioned issues introduces additional challenges to performing automatic language identification. Yet it should be noted that identifying the language of a sentence is not a straightforward task for humans either due to ambiguous language use or orthographically similar words in multiple languages. For example, a caption consisting of a single proper noun, such as ‘Helsinki’, may represent Finnish, English, German, or some other language whose vocabulary includes this word, essentially preventing the identification of language.

We evaluated several state-of-the-art frameworks that provide pre-trained models for performing automatic language identification. The libraries considered for the current study are listed in Table 1 and introduced briefly below. The first framework, fastText, relies on word embeddings, which is a technique for learning numerical representations of words in a vocabulary by observing their distribution in their context of occurrence (Bojanowski *et al.*, 2017). The second framework, langid.py, is designed to provide reliable language identification across multiple domains, such as official documents, newspaper articles, and social media messages (Lui and Baldwin, 2012). Finally, the third framework, CLD2 or the Compact Language Detector 2, was originally developed for

Table 1 Language identification frameworks used in the study

Name	Reference	Number of languages supported
fastText	Bojanowski <i>et al.</i> (2017)	176
langid.py	Lui and Baldwin (2012)	97
CLD2	–	83

Google’s Chromium open-source project but has not been documented in a peer-reviewed publication. For this study, we used CLD2 via the polyglot natural language processing library.

All programs developed for this study were written using the Python 3.6.3 programming language, to take advantage of the wide range of libraries available within the Python ecosystem. The libraries used include the Natural Language Toolkit (NLTK; Bird *et al.*, 2009), polyglot, spaCy, and gensim (Rehurek and Sojka, 2010) for natural language processing; scikit-bio for diversity measures; and pandas (McKinney, 2010) and scikit-learn (Pedregosa *et al.*, 2011) for storing and manipulating the data. All code written for this study is made publicly available with an open licence at: <https://doi.org/10.5281/zenodo.1404729>.

3.3 Evaluating language identification frameworks

To evaluate how the language identification frameworks introduced above perform on our data, we created a ground truth by randomly sampling the data without replacement for 1,476 captions. We then applied the preprocessing steps described in [Table 2](#) to these captions, extracting a total of 2,011 sentences. Two annotators, namely, the first and the second author, subsequently identified the language of each preprocessed sentence manually. We annotated each language using its ISO-639 code, such as ‘en’ for English, or using multiple codes joined by a + if the sentence featured more than one language, such as ‘en+fi’ for English and Finnish.

To assess the level of agreement between the two annotators, we used the common metrics for measuring inter-rater agreement surveyed in [Artstein and Poesio \(2008\)](#), such as Fleiss’ κ (0.929), Scott’s

Table 2 The individual steps of the preprocessing strategy were designed to counter common challenges in automatic language identification, such as emojis and smileys, excessive punctuation, multilingual hashtags and usernames, and sentence-level code-switching

1	The original caption includes hashtags, user mentions, and smileys and emojis Great weather in Helsinki!!! On holiday with @username.:-) #helsinki #visitfinland 🌞🌈
2	We begin by replacing any line breaks with whitespace and convert the emojis into their corresponding emoji shortcodes, which are wrapped in colons Great weather in Helsinki!!! On holiday with @username.:-) #helsinki #visitfinland:nerd_face_&_sunny_&_passenger_ship:
3	The colons make finding the emojis easy using a regular expression, which we then apply to remove them Great weather in Helsinki!!! On holiday with @username.:-) #helsinki #visitfinland
4	We then remove any words that begin with an @ symbol, which indicates a username Great weather in Helsinki!!! On holiday with:-) #helsinki #visitfinland
5	Next, we remove any hashtags, that is, any words beginning with a # Great weather in Helsinki!!! On holiday with:-)
6	Any remaining non-alphanumeric words in the caption, such as the smiley:-) are then removed using a regular expression Great weather in Helsinki!!! On holiday with
7	Longer sequences of exclamation or question marks (e.g. !!!), full stops, and other kinds of punctuation are shortened to just one of each character (e.g. !) Great weather in Helsinki! On holiday with
8	These sequences can confuse the Punkt sentence tokenizer (Kiss and Strunk, 2006), which outputs a Python list containing sentence tokens. These tokens are then fed to the language identification frameworks one at a time ["Great weather in Helsinki!", "On holiday with"]

π (0.929), and Krippendorff's α (0.929) as implemented in NLTK (Bird *et al.*, 2009). The average observed agreement between the two annotators was 0.948. Overall, these metrics suggest that the ground truth can be reliably used for evaluating the performance of language evaluation frameworks, particularly as the manual classification also accounted for code-switching within sentences. For the final ground truth, we dropped captions whose language we disagreed on, retaining a total of 1,374 captions with 1,863 sentences, which was further reduced to 1,688 by leaving out sentences whose language could not be manually identified or which contained sentence-internal code-switching.

We then evaluated the language identification frameworks against the ground truth and examined whether their performance would improve by excluding sentences with a low character count. fastText and langid.py had a slight advantage over CLD2, as they supported all manually identified languages present in the ground truth, whereas CLD2 did not support Latin. However, the ground truth contained only three sentences in Latin, so this disadvantage should not have a big impact on the

performance of CLD2. Table 3 reports the reliability of predictions for each framework at different character thresholds, using Krippendorff's α to correct for chance agreement. Average observed agreement—or accuracy—is given in parentheses.

As Table 3 shows, the fastText library and its pre-trained model provide superior performance compared to langid.py and CLD2 regardless of the character threshold. langid.py and CLD2 begin to match fastText's baseline performance only at the threshold of thirty characters or above, which simultaneously involves losing nearly 60% of the data. This trade-off is obviously unacceptable, which is why we chose fastText for automatic language identification.

3.4 Measuring richness and diversity

To measure the richness and diversity of the languages that make up the virtual linguistic landscape, we adopt common indices used in the fields of ecology and information sciences, such as richness, Menhinick's richness, Berger–Parker dominance, and Shannon entropy. Peukert (2013) provides a thorough introduction to using these indices to measure linguistic diversity, illustrating their application in a comparison of physical linguistic

Table 3 Krippendorff's α scores for language identification frameworks at different character thresholds for preprocessed sentence length

Framework	No threshold	>10 characters	>20 characters	>30 characters
CLD2	0.845 (0.895)	0.850 (0.899)	0.895 (0.928)	0.961 (0.974)
fastText	0.909 (0.939)	0.919 (0.946)	0.961 (0.974)	0.978 (0.985)
langid.py	0.787 (0.851)	0.799 (0.861)	0.868 (0.908)	0.917 (0.943)
Data loss	0% (0)	17.07% (318)	41.28% (796)	59.85% (1,115)

Note: Best result is marked in bold. For data loss, the value in parentheses reports the number of sentences lost.

landscapes in two neighbourhoods in Hamburg, Germany and showing how these indices may be used to measure and compare linguistic diversity across locations. Manjavacas (2016), in turn, applies similar indices to geotagged Twitter posts from Berlin, Germany. Because these indices are relatively new to the study of linguistic landscapes, we introduce them in greater detail in connection with the analyses of linguistic richness and diversity in Section 4.4.

4 Exploring the Virtual Linguistic Landscape

4.1 Temporal patterns in social media activity

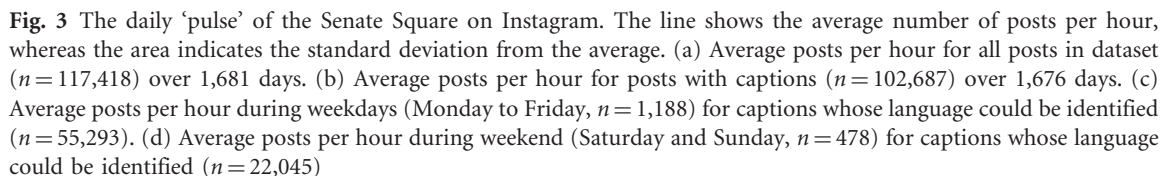
Fig. 3 presents Instagram activity around the Senate Square over 24 h. The figures show the average number of posts and their standard deviation for each hour of the day for four different samples: Fig. 3a shows the hourly frequency of all posts in the data set over 1,681 days, which also includes posts without any linguistic content ($n = 117,418$). Not surprisingly, this frequency reflects common hours of activity in the city, with approximately four to six posts per hour for daytime and evening hours. During the night, the number falls down to roughly two posts per hour. A similar pattern may be observed in Fig. 3b, which only includes posts with captions ($n = 102,687$).

The pattern changes when choosing different timescales and preprocessing the data for language identification ($n = 77,338$), as illustrated in Fig. 3c and d, which show the average number of hourly of posts for weekdays ($n = 1,118$) and weekends ($n = 478$), respectively. Whereas the weekdays

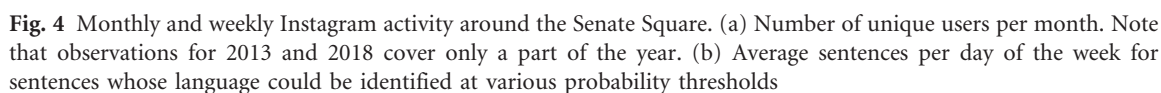
show a peak around lunch hours, the activity increases considerably towards the evening during weekends. A D'Agostino–Pearson test showed that none of the hourly observations in Fig. 3c and d follow a normal distribution, which means that the statistical differences between hourly activity may be evaluated using Levene's test and the Mann–Whitney U-test. For Levene's test, which compares the variance of samples, the differences were found to be statistically significant for Hours 2 ($W = 4.947$, $P = 0.027$), 4 ($W = 6.971$, $P = 0.009$), 5 ($W = 17.829$, $P = <0.001$), 7 ($W = 5.536$, $P = 0.019$), 9 ($W = 8.387$, $P = 0.004$), and 16 ($W = 7.111$, $P = 0.008$). The Mann–Whitney U-test, which examines the difference in averages, showed a statistically significant difference for Hour 2 ($U = 18,043.5$, $P = 0.025$).

This suggests that social media activity is subject to temporal variation, which can be revealed by examining the data on different timescales. In other words, studying the activity at lunch hour during the working week will reveal a different picture than an analysis focusing on the late hours on the weekend. This variation will undoubtedly affect the appearance of the virtual linguistic landscape on the daily scale and beyond. As a culturally valued landmark and a tourist attraction, the Senate Square also experiences seasonal variation, attracting a higher number of users during the summer months and Christmas holidays, as shown in Fig. 4a. The seasonal pattern becomes increasingly pronounced due to the rapidly growing popularity of Instagram as a social media platform.

Fig. 4b, in turn, shows the average number of sentences per day of the week, which reveals increased activity during the weekend. This trend, however, becomes less pronounced due to loss of



Including all predictions regardless of their level of confidence is likely to increase the number of errors, as very short sentences force fastText to make uninformed guesses based on limited data. To improve the quality of language identification while preserving the temporal features of Instagram activity at the Senate Square, we exclude predictions that fall into the first decile either in terms of their associated probability (<0.4231) or character length after preprocessing (<10), amounting to a loss of 17.31% of the data. This left us with



The graphs in Fig. 5 are presented in pairs. On the left-hand side, the Y-axes show the daily relative frequency, which calculated given by dividing the number of observations for each language by the total number of daily observations for all languages. This measurement is intended to capture the power relations and visibility of different languages in the

Generally, the ‘big three’—English, Finnish, and Russian—make up the vast majority of the virtual linguistic landscape. What is particularly worth noting in [Fig. 5a and b](#) is that Finnish overtook Russian as the second most common language only in 2015. Traditionally, Helsinki has been a popular destination among Russians due to its proximity and accessibility via road, rail, sea, and

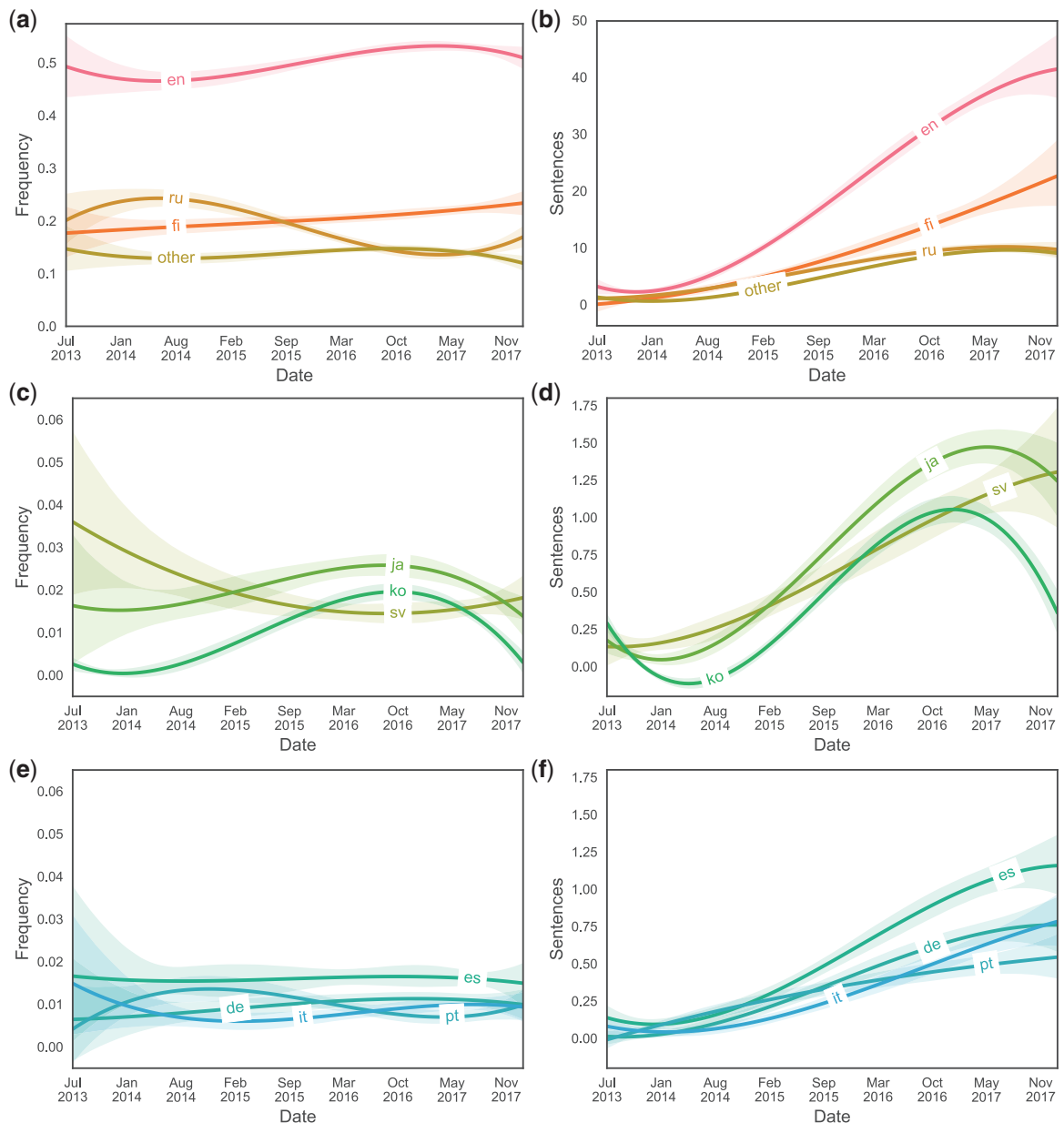


Fig. 5 Daily relative frequencies for languages identified using fastText, with 99.9% confidence intervals estimated using 10,000 bootstrapped samples from the underlying data, which are marked by the shaded areas. The lines show a third-order polynomial regression fitted using ordinary least squares. (a) Daily relative frequencies for the top-3 languages: English (en), Finnish (fi), Russian (ru) and other languages ($n = 77$). (b) Daily sentence counts for the top-3 languages. (c) Daily relative frequencies for the top 4–6 languages: Japanese (ja), Korean (ko) and Swedish (sv). (d) Daily sentence counts for the top 4–6 languages. (e) Daily relative frequencies for the top 7–10 languages: Spanish (es), German (de), Italian (it) and Portuguese (pt). (f) Daily sentence counts for the top 7–10 languages

air. Interestingly, the decline of the Russian language coincides with the economic sanctions imposed on Russia due to the invasion of Ukraine, which caused the number of Russian tourists visiting Helsinki to dip in 2015 and 2016 (Official Statistics of Finland, 2018). Comparing the difference between the daily relative frequencies for Russian in 2014 and 2015–16 using the Kruskal–Wallis H -test was found to be statistically significant at $H = 31.503$, $P = <0.001$.

Figure 5c–f zooms into the languages outside the top three, which were grouped together under the label ‘other’ in Fig. 5a and b. Note that this move is accompanied by a changes of scale, as the relative frequencies and sentence counts for these languages are considerably lower than those in Fig. 5a and b. The observations are split into different figures for a clearer view, but if Fig. 5c–f were presented in a single graph, the confidence intervals would overlap for many languages, indicating that the differences in their frequencies and counts are not statistically significant. The way the relative frequencies of these languages fluctuate suggests that they contribute sporadically in the virtual linguistic landscape, which is also supported by their low sentence counts.

Nevertheless, Fig. 5c and d shows how geographically remote languages such as Japanese (ja) and Korean (ko) contribute to the virtual linguistic landscape, even temporarily surpassing Swedish, the second official language of Finland. The relatively low proportion of Swedish in the virtual linguistic landscape stands in stark contrast with the physical linguistic landscape, in which Swedish remains very prominent, as public signs are required to be bilingual if the number of minority speakers in the municipality exceeds 8% or 3,000 individuals (Syrjälä, 2017, p. 118). This is naturally the case with Helsinki as well, which is historically a bi- and multilingual city. However, fastText cannot distinguish between standard Swedish and Finland-Swedish, which means these observations should not be associated exclusively with the Swedish-speaking minority in Finland, but include visitors from Sweden as well.

Coming back to Japanese and Korean, it should be noted that although tourism statistics for Helsinki show that visitors from European countries

outnumber Asians three to one (Official Statistics of Finland, 2018), the widespread adoption of mobile technology among Japanese and Korean users may explain their prominence in the virtual linguistic landscape. These languages, however, decline towards the present, although tourism statistics show that arrivals from Japan and Korea continue to increase, which may suggest that these users are abandoning Instagram. European visitors, in turn, are likely to include a sizeable number of business travellers, who may be less likely to contribute to the virtual linguistic landscape at the Senate Square, which may explain the relatively low proportion of major languages spoken in Europe such as Spanish, German, Italian, and Portuguese.

4.3 Language choices among users

The most striking feature of the virtual linguistic landscape at the Senate Square is the dominance of the English language, as it is unlikely that half of the users active at the location would speak English as their first language. To investigate language choices among users, we retrieved the time and location of posts for up to thirty-three previous posts for each user, who were naturally limited to those users who had posted captions whose language we could identify. To determine the likely country of origin for each user, we first retrieved the administrative region of each coordinate/timestamp pair in the location history using a point-in-polygon query. Next, we used the timestamps to determine the overall duration of user’s activity within each region by calculating the time between the oldest and newest posts. In addition to storing the region with the longest period of activity, we also recorded the region with the most activity. Finally, we calculated the average duration of activity for each user by dividing the time spent at each region by the total number of regions visited.

The initial data for estimating the users’ country of origin contained 75,685 posts by 49,842 unique users. On the average, the location history of a user contained 18.02 coordinate/timestamp pairs ($SD = 8.46$), whereas the average period of activity amounted to 152 days ($SD = 161$). To make our estimation more reliable, we discarded the first quartile for both coordinate/timestamp pairs and

Table 4 The distribution of the six most common languages among the users originating in ten most common countries

Country	Finnish	English	Russian	Swedish	Japanese	Korean	All
Finland	10,691	10,629	673	468	57	17	23,127
Russia	73	903	8,157	2	–	1	9,261
The USA	100	2,687	97	8	1	4	2,987
The UK	82	1,813	31	7	5	5	1,998
Germany	78	836	53	2	7	3	1,281
Sweden	88	528	59	308	4	3	1,061
Spain	72	478	112	4	1	10	1,048
Italy	55	554	133	6	5	7	1,019
France	37	474	110	3	9	12	817
Japan	14	247	1	1	364	14	674

Note: The countries are ranked by their popularity in the leftmost column. The rightmost column gives the total number of sentences written by users from the particular country in all languages.

the longest period of activity. In practice, this meant excluding users with eleven or fewer coordinate/timestamp pairs and whose longest period of activity was 44 days or less. For the final estimation, we retained a total of 45,685 posts by 31,442 unique users. For these users, we assumed that the administrative region where the users had been active for the longest period of time could be used to approximate their country of origin.

Table 4 presents the distribution of sentences in the six most frequent languages shown in Fig. 5 among users from the ten most frequent countries of origin. As may be expected, the majority of users active in the vicinity of the Senate Square come from Finland, but what is surprising is that Finnish users post nearly as much in English as in Finnish. Previous surveys on the role of the English language in Finland have emphasized the popularity and importance of English, particularly among the youth (Leppänen *et al.*, 2011). This may be a source of bias, as youth are also more likely to use social media (Longley *et al.*, 2015; Hausmann *et al.*, 2018). Nevertheless, the high proportion of sentences (45.9%) written in English warrants closer attention, as similar findings have been reported for other social media platforms, namely Twitter, by Laitinen *et al.* (2018).

To do so, we trained a topic model over monolingual English captions posted by users whose country of origin was estimated to be Finland. These data consisted of 8,636 captions with 5,552 unique words after removing rare and frequent

words that appeared in a single sentence or in more than 25% of the sentences. The model was trained using the Latent Dirichlet Allocation algorithm for 150 iterations with ten passes through the corpus, using the implementation provided in the gensim library (Rehurek and Sojka, 2010). To preprocess the data, we adopted the procedure set out in Table 2. We also removed stopwords defined in NLTK (Bird *et al.*, 2009) and lemmatized the words using the lookup table for English in spaCy. Finally, we calculated a coherence score, C_v , for each topic, which has been suggested to correlate strongly with human evaluations of topic coherence (Röder *et al.*, 2015).

Table 5 gives the ten most prominent topics with their ten most frequent words. Some of the coherence scores are fairly low, which is not surprising given the noisy social media data and the small size of the corpus. Nevertheless, the topics can provide insights into the nature of the content posted in English by Finnish users. To begin with, several topics seem to be strongly associated with the location, weather, leisure, and celebrations such as Christmas and New Year's Eve (1 and 3) and the Lux light carnival (6). Many topics also feature words associated with a positive sentiment (3, 5–7, 9, and 10). This suggests that Finns use English to connect with international audiences, appraising the physical location and the activities associated with it in the virtual space.

Finnish users appear to participate in maintaining the identity of the location as a culturally valued

Table 5 A topic model trained over 8,636 captions written in English by Finnish users, with one topic per column

1	2	3	4	5	6	7	8	9	10
Helsinki	Get	Year	Make	Love	Helsinki	Good	Town	Look	Day
Christmas	Cold	Happy	Start	Night	Lux	Pizza	Run	Go	One
Cathedral	Menu	New	Open	Great	Light	Morning	Conjurer	Lot	Independence
Light	Thing	Well	Art	Enjoy	Finland	Beautiful	Afternoon	Let	Church
Market	Finally	Time	Night	People	Sunday	Walk	Friday	Know	Back
Senate	Ready	Take	Welcome	See	Home	Lovely	Finnish	Special	Nice
Square	New	Week	Way	Last	Festival	City	Colour	Right	Finland
Time	May	Picture	Wine	Come	Snow	Sun	Well	Exhibition	Sunny
Lunch	Always	Thank	Drink	December	Amaze	Blue	Look	Pretty	Big
Winter	Taste	Get	Spring	Weekend	Wait	Today	Know	Like	Last
0.342	0.263	0.492	0.292	0.3	0.37	0.345	0.254	0.356	0.289

Note: The words (rows) associated with each topic are sorted by their weight in a descending order. The final row gives the coherence score C_v for the topic (Röder *et al.*, 2015).

landmark, at the same time construing the location as a tourist attraction. The role of English as the lingua franca of tourism (Francesconi, 2014), which may also explain the choice of language, is also supported by a positive view of the language and a high level of proficiency in Finland (Leppänen *et al.*, 2011). However, the preference for English holds for most, but not all linguistic groups contributing to the virtual linguistic landscape: Table 4 shows that Russians clearly prefer their native language over English.

4.4 The diversity of the virtual linguistic landscape

Finally, we turn towards the richness and diversity of the virtual linguistic landscape, applying the indices introduced in Section 3.4. The following discussion focuses on Fig. 6, which shows several indices applied to the results of automatic language identification. We introduce these indices and explain their implications below.

Fig. 6a shows the linguistic richness, or simply the number of unique languages per day, and the number of singletons, that is, how many languages appear only once a day. In Fig. 6a, the parallel increase in unique languages and singletons suggests that smaller languages are driving the increase in linguistic richness. This observation was supported by a strong positive correlation for Pearson's r between 30-day rolling averages for unique languages and singletons ($r=0.975$, $n=1,633$, $P = <0.001$). Increasing linguistic richness also correlated with

the increase in unique users ($r=0.899$, $n=1,633$, $P<0.001$), as shown in Fig. 6b. To summarize, Fig. 6a and b suggests that the growing popularity of Instagram has resulted in an increasingly rich virtual linguistic landscape at the Senate Square, as smaller linguistic groups have adopted the platform.

Simple richness index, however, does not account for the growing volume of data due to the increasing popularity of the platform. This perspective can be provided by Menhinick's richness index, which emphasizes the relationship between data volume and richness. Menhinick's richness index, shown in [Fig. 6c](#), reveals a decreasing trend over the 4.5 years. This trend suggests that despite increasing linguistic richness, driven by the increase in smaller languages, the virtual linguistic landscape is increasingly dominated by languages such as English, Finnish, and Russian (cf. [Fig. 5a and b](#)). In other words, the growing volume of data has made the dominant languages increasingly prominent in the virtual linguistic landscape, which is reflected in a decreasing value for Menhinick's richness index.

Measuring the diversity of the virtual linguistic landscape requires indices that account for both the number of languages observed and their relative proportions. One such index is the Berger–Parker dominance index, shown in Fig. 6d, which gives the fraction of observations for the language with the most posts per day. Given the observations in Fig. 5a, approximately half of the time the dominant language is English. The decreasing

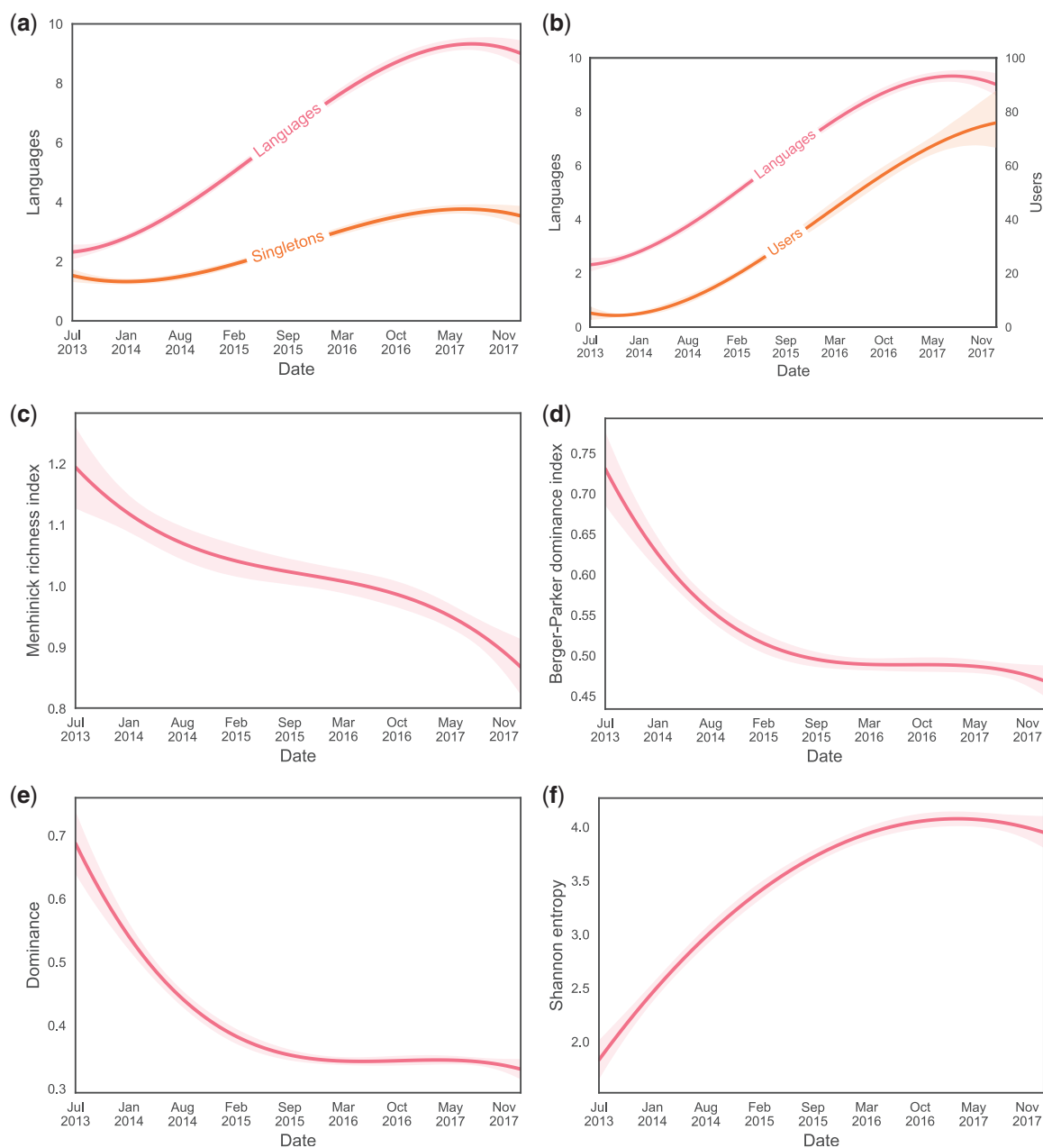


Fig. 6 Various diversity measures applied to the data set, with 99.9% confidence intervals estimated using 10,000 bootstrapped samples from the underlying data. The line shows a third-order polynomial regression fitted using ordinary least squares. (a) Richness and singletons. (b) Richness and daily unique users. (c) Menhinick richness. (d) Berger—Parker dominance. (e) Dominance. (f) Shannon entropy

Berger–Parker index suggests that the dominant languages are losing ground to smaller languages, showing a drop of thirty points during the 4.5 years, which suggests that the virtual linguistic landscape of the Senate Square is becoming increasingly diverse. This observation is also supported by the decreasing dominance index in Fig. 6e, which measures the respective proportions of languages: a dominance index of 0 would indicate that all languages are equally present, whereas an index of 1 would mean the total dominance of a single language.

Finally, the observed increase in diversity is also supported by Shannon entropy, shown in Fig. 6f, which captures the amount of information required to describe the degree of order/disorder in a system. The higher the degree of disorder—in this case, the variety of languages and their respective probabilities of occurrence—the more information is required to describe the state of the system, that is, the virtual linguistic landscape. Interestingly, the index for Shannon entropy peaks in 2017. This may suggest that the virtual linguistic landscape of the Senate Square has reached its maximal degree of diversity (with slightly over eight languages on the average day, as shown in Fig. 6a possible within the current userbase of Instagram).

To summarize, several conclusions may be drawn from the indices in Fig. 6. The richness of the virtual linguistic landscape increases as the number of users grows. Although the number of languages found in the virtual linguistic landscape grows, dominant languages such as English, Finnish, and Russian gain the most from the growth, enabling them to consolidate their position. Yet the proportion of dominant languages is decreasing, which indicates increasing diversity. Put differently, smaller languages are gaining on the share of the dominant languages. At the same time, the virtual linguistic landscape at the Senate Square seems to have reached a point where the linguistic diversity no longer increases. In other words, the number of languages in the virtual linguistic landscape remains the same, but the smaller languages change.

5 Discussion and Conclusion

Our results suggest that virtual linguistic landscapes can be effectively characterized using computational

methods, which are necessary for handling high volumes of social media data. With carefully planned preprocessing, automatic language identification and other natural language processing techniques can do most of the analytical work in a sufficiently reliable manner. However, insights provided by automatic language identification are limited without the means to evaluate the respective proportions of the observed languages. Our analysis revealed a rich and diverse virtual linguistic landscape at the Senate Square, which is dominated by English, as the language is used extensively by both locals and tourists.

The results also emphasize the role of Senate Square as a highly valued cultural landmark and a tourist attraction (Jokela, 2014). The cultural importance is manifested in the high number of posts by locals, whereas the impact of tourism is reflected by the high number of foreign visitors. In this respect, our findings support Kellerman's (2010) view that qualities associated with the physical place may be carried over to the corresponding virtual space. Although we did not explicitly touch upon the issue in the analysis, it should be noted that global mobility and tourism are a privilege of a select few rather than the many, which is likely to be reflected in the linguistic landscape. Choosing an alternative location for the study, such as a local transportation hub, would have likely yielded very different results (cf. Soler-Carbonell, 2016).

The richness and diversity of the virtual linguistic landscape also resonate with Lee's (2016, p. 119) proposal that user-generated social media content increases the potential for exposure to foreign languages. Geotagged social media content may be particularly effective for this purpose, as content associated with a location can be accessed through map interfaces instead of using hashtags or search terms in some specific language. This effect is further reinforced by Instagram, which allows locations defined on the platform to have multilingual names. All the content associated with the locations named in different languages is then aggregated under a single point of interest. This is also likely to drive the formation and maintain the double space, as conceptualized by Kellerman (2010).

In addition, the nature of Instagram as a platform must be taken into account when interpreting

the results. Unlike Twitter, which acts as a forum for public discussion, Instagram may be preferred for sharing personal experiences (Zappavigna, 2011, 2016; Tenkanen *et al.*, 2017). Together with the intended audience, the platform may affect language choices among users (Androutsopoulos, 2015). Tracing these linguistic repertoires would, however, require a much closer analysis of longitudinal data for individual users, which was beyond the scope of this article. However, our proposed method could be easily adopted for a large-scale study of what Pennycook and Otsuji (2014, p. 166) have called “a geography of linguistic happenings”. Such analyses, however, would still be limited by the spatial accuracy of Instagram, as observed in Section 3.1. Users may, for instance, associate content with locations higher in the POI hierarchy (such as ‘Helsinki’ instead of ‘Senate Square’) or choose the wrong location altogether.

In terms of other limitations, the results are naturally affected by how widely Instagram has been adopted by potential users of social media, and should be evaluated in the light of the inherent bias towards younger population found in social media data (Longley *et al.*, 2015; Hausmann *et al.*, 2018). Furthermore, the proposed method cannot provide a fine-grained view of the linguistic landscape, because automatic language identification cannot detect code-switching within sentences, or distinguish between varieties of a single language, such as American and British English or Finland-Swedish and Standard Swedish, unless explicitly trained to do so.

Despite these limitations, our results suggest that Instagram and other social media platforms with geolocated content do nevertheless hold much potential for sociolinguistic inquiry, as suggested by Androutsopoulos (2014). Tapping further into this potential, however, would benefit from collaborating with geographers, to leverage more advanced methods for spatiotemporal analysis. Such analyses could be used, for instance, to reveal where and when particular linguistic groups are active, to evaluate the potential for interaction between these groups. Longitudinal analyses for individual users, in turn, could be used to investigate their linguistic repertoires. Finally, because computational methods

develop rapidly, analytical tools should be shared openly to enable the replication and reproduction of research, which would benefit the entire field of study.

A natural extension to the current work would be to take on what Jaworski and Thurlow (2010) have conceptualized as semiotic landscapes, whose analysis would include other modes of expression besides language in the virtual linguistic landscape. Although research on artificial intelligence is making rapid progress in processing multimodal data (Bateman *et al.*, 2017, pp. 163–4), identifying fine-grained patterns of multimodal communication in high volumes of geotagged social media data is likely to remain a long-term endeavour. Nevertheless, sufficiently mature computational techniques can already support the study of both virtual and physical linguistic landscapes, and their potential applications should be explored further.

Funding

This work was supported by the Finnish Cultural Foundation and the Kone Foundation.

References

- Allen, P. T., Fatah, A., and Robison, D. (2018). Urban encounters reloaded: Towards a descriptive account of augmented space. In Jung, T. and tom Dieck, M. C. (eds), *Augmented Reality and Virtual Reality: Empowering Human, Place and Business*. Cham: Springer, pp. 259–73.
- Androutsopoulos, J. (2014). Computer-mediated communication and linguistic landscapes. In Holmes, J. and Hazen, K. (eds), *Research Methods in Sociolinguistics: A Practical Guide*. Oxford: Wiley, pp. 74–90.
- Androutsopoulos, J. (2015). Networked multilingualism: some language practices on Facebook and their implications. *International Journal of Bilingualism*, 19(2): 185–205.
- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4): 555–96.

- Lee, C. and Chau, D. (2018). Language as pride, love, and hate: archiving emotions through multilingual Instagram hashtags. *Discourse, Context and Media* 22, 21–9.
- Leppänen, S. and Peuronen, S. (2012). Multilingualism and the internet. In Chapelle, C. A. (ed.), *The Encyclopedia of Applied Linguistics*. Oxford: Wiley-Blackwell.
- Leppänen, S., Pitkänen-Huhta, A., Nikula, T., Kytölä, S., Törmäkangas, T., Nissinen, K., Kääntä, L., Räisänen, T., Laitinen, M., Pahta, P., Koskela, H., Lähdesmäki, S., and Jousmäki, H. (2011). *National survey on the English language in Finland: Uses, meanings and attitudes, Vol. 5 of Studies in Variation, Contacts and Change in English*. Helsinki: University of Helsinki.
- Longley, P. A., Adnan, M., and Lansley, G. (2015). The geotemporal demographics of Twitter usage. *Environment and Planning A: Economy and Space*, 47(2): 465–84.
- Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju Island, Korea, July 10. Association for Computational Linguistics, pp. 25–30.
- Manjavacas, E. (2016). *Mapping urban multilingualism through Twitter*. Master’s thesis, The Free University of Berlin.
- McKinney, W. (2010). Data structures for statistical computing in Python. In van der Walt, S. and Millman, J. (eds), *Proceedings of the 9th Python in Science Conference*, Austin, Texas, United States, June 28–July 3, pp. 51–6.
- Official Statistics of Finland (2018). Accommodation statistics. <http://www.stat.fi/til/matk/index.html> (accessed 6 July 2018).
- Paolillo, J. C. (2007). How much multilingualism? Language diversity on the internet. In Danet, B. and Herring, S. C. (eds), *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford: Oxford University Press, pp. 408–30.
- Papen, U. (2012). Commercial discourses, gentrification and citizens’ protest: The linguistic landscape of Prenzlauer Berg, Berlin. *Journal of Sociolinguistics* 16(1): 56–80.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–30.
- Pennycook, A. and Otsuji, E. (2014). Metrolingual multitasking and spatial repertoires: ‘pizza mo two minutes coming. *Journal of Sociolinguistics*, 18(2): 161–84.
- Peukert, H. (2013). Measuring linguistic diversity in urban ecosystems. In Duarte, J. and Gogolin, I. (eds), *Linguistic Superdiversity in Urban Areas: Research Approaches*. Amsterdam: Benjamins, pp. 75–93.
- Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of 7th Language Resources and Evaluation Conference: Workshop on New Challenges for NLP Frameworks*, ELRA, pp. 45–50.
- Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM’15)*, ACM, pp. 399–408.
- Sergeant, P. and Tagg, C. (eds) (2014). *The Language of Social Media*. Basingstoke: Palgrave.
- Soler-Carbonell, J. (2016). Complexity perspectives on linguistic landscapes: a scalar analysis. *Linguistic Landscapes*, 2(1): 1–25.
- Syrjälä, V. (2017). Naming businesses – in the context of bilingual Finnish cityscapes. In Ainiala, T. and Östman, J.-O. (eds), *Socio-onomastics: The Pragmatics of Names*. Amsterdam: Benjamins, pp. 183–202.
- Tenkanen, H. (2017). *Capturing Time in Space: Dynamic Analysis of Accessibility and Mobility to Support Spatial Planning with Open Data and Tools*. PhD thesis, Department of Geosciences and Geography, University of Helsinki. <http://urn.fi/URN:ISBN:978-951-51-2935-9>.
- Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., and Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports* 7(17615).
- Villi, M. (2015). “Hey, I’m here right now”: Camera phone photographs and mediated presence. *Photographies* 8(1): 3–22.
- Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media and Society* 13(5): 788–806.

- Zappavigna, M.** (2013). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Continuum.
- Zappavigna, M.** (2016). Social media photography: construing subjectivity in Instagram images. *Visual Communication*, 15(3): 271–92.
- Zook, M. A. and Graham, M.** (2007). Mapping digiplace: Geocoded internet data and the representation of place. *Environment and Planning B*, 34(3): 466–82.
- Zubiaga, A., Vicente, I. S., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., and Fresno, V.** (2016). TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4): 729–766.

Note

1 <http://www.instagram.com>