Mapping the languages of Twitter in Finland: Richness and diversity in space and time

TUOMO HIIPPALA, TUOMAS VÄISÄNEN,

TUULI TOIVONEN & OLLE JÄRV

Abstract Twitter is a popular social media platform for scholarly research, because the user-generated content on the platform can also include geographic and temporal information. We collect a corpus of 38 million Twitter messages with two million geographical coordinates to map the languages used across Finland at the level of regions and municipalities. To cope with the high volume of social media data, we use automatic language identification and place of residence detection. We estimate the linguistic richness and diversity of users and locations using measures developed within ecology and information sciences. The analyses reveal a rich, multilingual environment that varies geographically and temporally, particularly between coastal, rural and urban areas. The results, which underline the mutual benefits of collaboration between linguists and geographers, provide a more fine-grained, accurate and comprehensive view of the languages used on Twitter in Finland than previously available.

1. Introduction

Producing and consuming content on social media platforms has become an inseparable part of everyday life for many. The resulting user-generated content on social media platforms bears the hallmarks of big data, which is characterised by volume, velocity and variety, as massive amounts of data about a wide range of topics are created in real time (Kitchin 2014). These characteristics naturally apply to the linguistic content on these platforms as well. For this reason, social media is often considered a rich source of information on real-life language use, which is complemented by metadata about the users, their demographics, social networks and geographical locations (Herdağdelen 2013). In particular, geographical information available on social media is increasingly used in research on dialectology, linguistic landscapes and language choice, which has led to an increased exchange between linguistics and geography (see e.g. Grieve *et al.* 2018; Coats, 2019*a*; Hiippala *et al.* 2019; Derungs *et al.* 2020).

Against this backdrop, this article reports on an exploratory, data-driven study of the languages used on Twitter in Finland. Twitter is a popular social media platform used for a wealth of different communicative purposes,

which allows the users to optionally add geographical information to the content posted on the platform (Hu & Wang 2020). By leveraging the linguistic and geographical information available on Twitter, we seek to understand the linguistic richness and diversity of both users and locations across Finland. We collect a corpus of 38 million Twitter messages with two million geographical coordinates for this purpose. To cope with the high volume of data, we apply various computational methods and measures developed in the fields of language technology, geoinformatics, ecology and information sciences. We contrast our findings with previous research on multilingualism in Finland, while also discussing the challenges and opportunities of working with social media data at the intersection of linguistics and geography.

2. Social media data: challenges and opportunities

Access to large volumes of social media data has opened up new avenues for linguistic research (see e.g. Zappavigna 2013; Seargeant & Tagg, 2014; Bouvier & Machin, 2018). In a recent overview of the emerging field of computational sociolinguistics, Nguyen et al. (2016) observe that "the availability of social media and other online language data in computer-mediated formats is one of the primary driving factors for the emergence of computational sociolinguistics." New sources of data have rekindled an interest in the social aspects of language use among computational linguists. Purschke and Hovy (2019: 114), for instance, consider social media data particularly valuable, because they "represent unsupervised everyday practice rather than language use from carefully-designed experiments." The crucial link between everyday practice and language use is often established via metadata, that is, information about what kind of content was uploaded on the platform, in which language, when, where and by whom.

Many social media platforms allow users to associate their content with geographical locations via a practice known as geotagging (Humphreys & Liao 2011). This practice is enabled by the use of positioning technology in mobile phones and other consumer electronics (Heikinheimo et al. 2020). The widespread use of mobile devices equipped with positioning technology has led various fields of research to consider how this combination shapes both spaces and places, and affects human interactions at various spatial scales (see

e.g. Dodge & Kitchin 2005; Zook & Graham 2007; Humphreys 2010; Auer *et al.* 2014). Geographers, in particular, have paid attention to connection between physical and virtual experiences of space. Graham et al. (2013) argue that this connection establishes a form of augmented reality, which is characterised by a "material/virtual nexus mediated through technology, information and code, and enacted in specific and individualised space/time configurations" (Graham et al. 2013: 465). Aharon Kellerman has characterised this setting as a 'double space', which covers both physical places and their virtual extensions on digital platforms (Kellerman 2010, 2014, 2016). The notion of a double space is particularly interesting to linguists, because it connects language use on social media platforms to concrete places and social situations in the physical world (Hiippala et al. 2019).

2.1. Collecting geotagged social media content programmatically

Not surprisingly, the interest in geotagged social media content has grown among linguists in recent years, as exemplified by numerous data-driven studies and data collection efforts (Williams et al. 2013). One such example is the Nordic Tweet Stream (NTS), a real-time monitor corpus of Twitter data from Denmark, Finland, Iceland, Norway and Sweden (Laitinen et al. 2018). NTS collects user-generated content ('tweets') and their associated metadata (e.g. time, location and language) that have been geotagged to Nordic countries via Twitter's Application Programming Interface (API) (Laitinen et al. 2018: 351). An API allows computer programs to make requests and retrieve content from the platform without accessing the service via a user interface on the web or in a mobile application.

Like many other research projects involving Twitter, NTS uses the so-called Twitter Streaming API, which is freely available and provides a random sample of approximately 1% of tweets posted on the platform in real time. Operated by Twitter, the Streaming API is a 'black box', which is now increasingly scrutinised by researchers who wish to understand its operation and the characteristics of the sample returned (Pfeffer et al. 2018). Laitinen et al. (2018) report on several experiments from the Nordic countries: comparing the geotagged tweets captured by NTS to another data collection system that attempts to retrieve all tweets written in Swedish showed that

approximately 2.8% of the tweets contain geotags. This finding is on par with observations made elsewhere in the world using a premium Twitter API that provides access to a 10% sample of tweets in real time (Leetaru et al. 2013). Laitinen et al. (2018: 353) also found that the vast majority of geotagged tweets are included in the 1% sample available through the Streaming API, but this finding is based on a few users only. To summarise, Twitter can hardly be characterised as a location-driven social media platform (Martí et al. 2019).

2.2. Characteristics of social media data

Despite providing access to high volumes of data, using social media data for building linguistic corpora involves several challenges. Information about the age, gender and occupation of language users, their social networks and the context of language use allow structuring linguistic corpora and relating this information back to observations concerning language use (Biber 1993). For social media data, much of this information is unreliable, because it must be derived through proxies (Sloan 2017). Coats (2019*a*), for instance, matches Twitter users' self-reported names and home locations to lists of first names and places in the Nordic countries, to derive the information needed to study gender and language choice (see also Herdağdelen 2013). Laitinen et al. (2017), in turn, use the number of 'friends' and 'followers' on Twitter to approximate social network size and its impact on the preference for the English language, while Hiippala et al. (2019) estimate Instagram users' home location by examining the geographical history of their posts.

Compared to users' self-reported location information, GPS coordinates may appear as a more accurate source of location information, which the users often omit or use to convey information not related to location (Hecht et al. 2011). However, assuming that geotags allow establishing a clear link between everyday linguistic practices and specific physical locations warrants caution, particularly when working with high-volume data that cannot be verified manually. Based on a survey of 400 users, Tasse et al. (2017: 256) conclude that "geotags are postcards, not ticket stubs". They emphasise that geotagging is a conscious, performative act and the locations tagged are more likely to be those visited rarely, not routinely. Tasse et al. (2017: 257–258) also point out that the spatial accuracy of geotags is degrading, because

social media platforms are switching from coordinate to point of interest (POI) geotags (see also Hochmair et al. 2018; Hu & Wang 2020). Unlike coordinate geotags, which provide accurate longitude and latitude, point-of-interest locations are defined by the social media platform and function as 'magnets' to which the user-generated content may be attached, as exemplified by countries such as 'Finland' or cities such as 'Helsinki'.

Shifting the attention from locations to users, Artamonova & Androutsopoulos (2019) call for attention to 'mediational repertoires', which emphasises that social media platforms may serve different communicative purposes for users, which also influence their language choice. In other words, a Finnish user may draw on English to participate in public discussions on Twitter, while communicating mainly in Finnish to closer friends on Instagram (see also Lee 2016). Mediational repertoires across platforms are further complicated by spatial and temporal repertoires, that is, the influence of space and time on language use (Pennycook & Otsuji 2014). From a linguistic perspective, these inherently dynamic linguistic repertoires, which have been theorised under the umbrella of metrolingualism (Pennycook & Otsuji 2015), represent the kind of micro-level language use that is difficult to study at scale using social media data. Although computational techniques needed to process high volumes of data, such as automatic language identification, can now detect code-switching (Rijhwani et al. 2017), these tools are rarely available off-the-shelf.

To summarise, the high volume of linguistic data available on social media is offset by unreliable information about the language users and the inherent socio-demographic biases of social media platforms. The same applies to spatial location information, which may be spatially inaccurate or fuzzy, depending on whether location information is provided as coordinate or point-of-interest geotags (Toivonen et al. 2019).

2.3. Combining linguistic and geographic insights

> *"There are good reasons for arguing that linguists can learn from geographers, not only how to improve their descriptions … but also how to improve their explanations."* (Trudgill 1974: 232)

Our daily lives and social interactions take place in and are constantly shaped by geographical space (Giddens 1984), which also undoubtedly affects language use due to its central role in all things social (Halliday 1978). One subfield of linguistics that has long acknowledged the value of geographic information is dialectology (Trudgill 1974; Szmrecsanyi 2012), and particularly its branch of dialectometry, which develops computational and quantitative methods for dialectology (Nerbonne & Kretzschmar 2013; Wieling & Nerbonne 2015; Grieve 2017). Whereas dialectology has argued extensively for the need to account for factors such as geographic distance, accessibility and socio-demographic features in modelling the emergence and diffusion of linguistic phenomena, dialectometry has developed computational and quantitative methods for this purpose. The introduction of methods from modern geoinformatics to dialectometry, however, is a relatively recent development (Grieve et al. 2011).

The application of geoinformatics has already yielded valuable insights on language use, particularly in connection with social media data. In recent years, geotagged social media data has been used to study how the diffusion of lexical items between cities follows the hierarchy of urban settlement system – larger cities transmit linguistic features to smaller cities, but social media allows transmission to overcome large geographical distances (Eisenstein et al. 2014). Grieve et al. (2018) add to this work by incorporating temporal information into their spatial models of lexical innovation and use spatial statistics to uncover their spread in the United States. Ljubešić et al. (2018), in turn, reconstruct dialect regions by estimating the spatial distribution of distinctive linguistic features in the Balkans. These studies reflect many characteristics associated with the notion of moving 'beyond the geotag' in geographic information science (Crampton et al. 2013). These characteristics include acknowledging (1) the unreliability of location information, (2) the temporal aspect of any event, (3) the rapid spread of phenomena across large distances, (4) the content created by non-humans (e.g. automated 'bots') and (5) the need to verify insights from social media data against other sources of data.

Derungs et al. (2020: 278–279) argue that although dialectology has been content to describe the spatial distribution of linguistic phenomena, the

potential of spatial information and insights has not been fully embraced by linguists. Besides spatial dependencies and diffusion between geographical units, an alternative spatial perspective to linguistic phenomena involves turning towards the individual. The diversity of individual's linguistic repertoire, one's surrounding social environment, its spatiotemporal characteristics, socio-economic features and accessibility are tightly linked to common interests of linguists and geographers. In this complex, intersecting web, language can serve as a key marker of cultural difference, ethnicity and community membership (Alexander et al. 2007; Järv et al. 2015).

Several influential conceptual frameworks in human geography, such as activity spaces (Golledge & Stimson 1997) and time geography (Hägerstrand 1970), place individuals at the centre of attention. Activity spaces refer to physical locations where individuals engage in social activities (e.g. home, work, restaurant, gym) and trajectories of movement between these locations over time. These conceptual frameworks can help to uncover what kinds of social and linguistic environments individuals are exposed to and to trace their potential for social interaction (Järv et al. 2015). Tracing the activity spaces of individuals and their spatiotemporal constellations can provide additional perspectives to linguistic repertoires, urban multilingualism and exposure to linguistic diversity, not only at the locations visited, but also while moving between them. Although these perspectives are gradually incorporated into linguistic research, there is considerable potential in fully integrating geographical insights related to spatiotemporality and accessibility (Longley & Adnan 2016; Järv et al. 2018; Toivonen et al. 2019).

3. Data and methods

3.1. Twitter as a source of spatial data

Twitter allows users to optionally add geographic information to their tweets in three ways (Hu & Wang 2020). The first option is to use coordinate geotags, which indicate the exact location using coordinates for latitude and longitude (API field: coordinates). The second option is to select a location (e.g. Helsinki, Finland) from Twitter's point-of-interest database (API field: place). The third

option makes Twitter an ambiguous source of spatial data: the user may add a coordinate geotag to a tweet, while also choosing a point-of-interest location that does not match the geographic location of the coordinate geotag. In other words, the coordinate geotag may point to Rovaniemi, while the location refers to Helsinki. As explained in the documentation for Twitter API, this allows users to express that a tweet may be about a location rather than implying that the tweet was created at the location.[1] We focus on tweets with coordinate geotags, as the point-of-interest database introduces additional uncertainties (Hecht et al. 2011) and creates artificial hotspots, as large volumes of tweets are attached to POI locations that correspond to large geographical areas (e.g. 'Helsinki').

3.2. Data collection

We collected the data from Twitter via the platform's Application Programming Interface (API), focusing on tweets with coordinate geotags. We combined our collection of tweets with an existing dataset that had access to a full archive of geotagged tweets to increase our coverage (Poorthuis & Zook 2017). Collecting the data involved the following steps:

1. We first defined a bounding box that covered the geographical area of Finland to retrieve all tweets that had been geotagged within the bounding box.
2. We then filtered the tweets with a geographical mask following the land and maritime borders of Finland to retrieve all tweets, which had been geotagged in Finland to rule out tweets from the St. Petersburg area, which falls within the original bounding box.
3. We then retrieved the tweet history for each user who had geotagged a tweet within Finland, up to the limit determined by the API (3200 tweets).

1 Twitter API reference: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects

4. After collecting tweet history for each user, we used a home detection algorithm developed by Massinen (2019) to predict their probable country of residence. The algorithm approximates home location by counting the number of unique weeks spent in a country.
5. We included the users whose home location was predicted to be Finland into our primary dataset.

Table 1. An overview of the collected data.

Tweets	38 487 766
Geotagged tweets	2 030 499
Unique users	40 442
Users with at least two geotags	40 342
Finnish users with at least two geotags	33 932
Finnish users without languages that occur only once	33 322
Average posts per user	941
Average geotags per user	50

3.3. Methods

We first used the Punkt tokenizer (Kiss & Strunk 2006) via the Natural Language Toolkit (Bird et al. 2009) to split tweets into orthographic sentences. To identify the language of each sentence, we used a model trained using the fastText algorithm, which is capable of identifying 176 languages (Bojanowski et al. 2017). This model has been previously shown to perform well with social media content (Hiippala et al. 2019). We discarded predictions made with less than 70% confidence and removed all languages that occurred only once in the user's tweet history to reduce the impact of misclassifications by the automatic language identification algorithm. Finally, we applied various measures of richness and diversity developed within ecology and information sciences to observations across each user's tweet history and the spatial units defined (Peukert 2013). These measures are introduced in greater detail in connection with the results in Section 4.1.

We then aggregated our observations into two common administrative spatial units in Finland: regions (Fi *maakunta*) and municipalities (Fi *kunta*). This choice was motivated by the number of users and content, and the goal of making the maps informative: a municipality-level analysis would have left many rural areas with few Twitter users, whereas focusing on regions would have hidden potentially meaningful spatial patterns at a finer scale. We used the home detection algorithm developed by Massinen (2019) to estimate the users' home municipality in Finland, based on the number of weeks spent at each municipality. Massinen (2019, p. 41) reports that the algorithm detects the user's place of residence with an accuracy of 88.6%. In case of a tie, we added 0.5 to the user count for both municipalities to reflect multiple home locations. Finally, we chose 25 municipalities with highest user counts to complement the 19 regions in Finland. In case the user was not predicted to reside within these 25 municipalities, the user was assigned to the region where the predicted home municipality is located. These methods were implemented in the Python programming language and are openly available at https://doi.org/10.5281/zenodo.4279401".

4. Results

4.1. The linguistic diversity and richness of individual users

We begin by focusing on individual users and quantify their linguistic diversity and richness. Figure 1 shows several measures of diversity and richness calculated over the tweet histories of individual users. For a comprehensive introduction to these measures and their application in linguistics, the reader is referred to Peukert (2013). In Figure 1, the y-axis provides the frequency, whereas the x-axis provides the values for each measure.

The dominance index in Figure 1a reflects the balance of languages in a user's tweet history: 0 means that all languages are equally present, whereas 1 indicates the dominance of a single language. The dominance index reveals that although there is a large group of users with a single dominant language, the majority of users draw on more than one language, as indicated by users with a dominance value below 1.0. The absence of low values suggests that

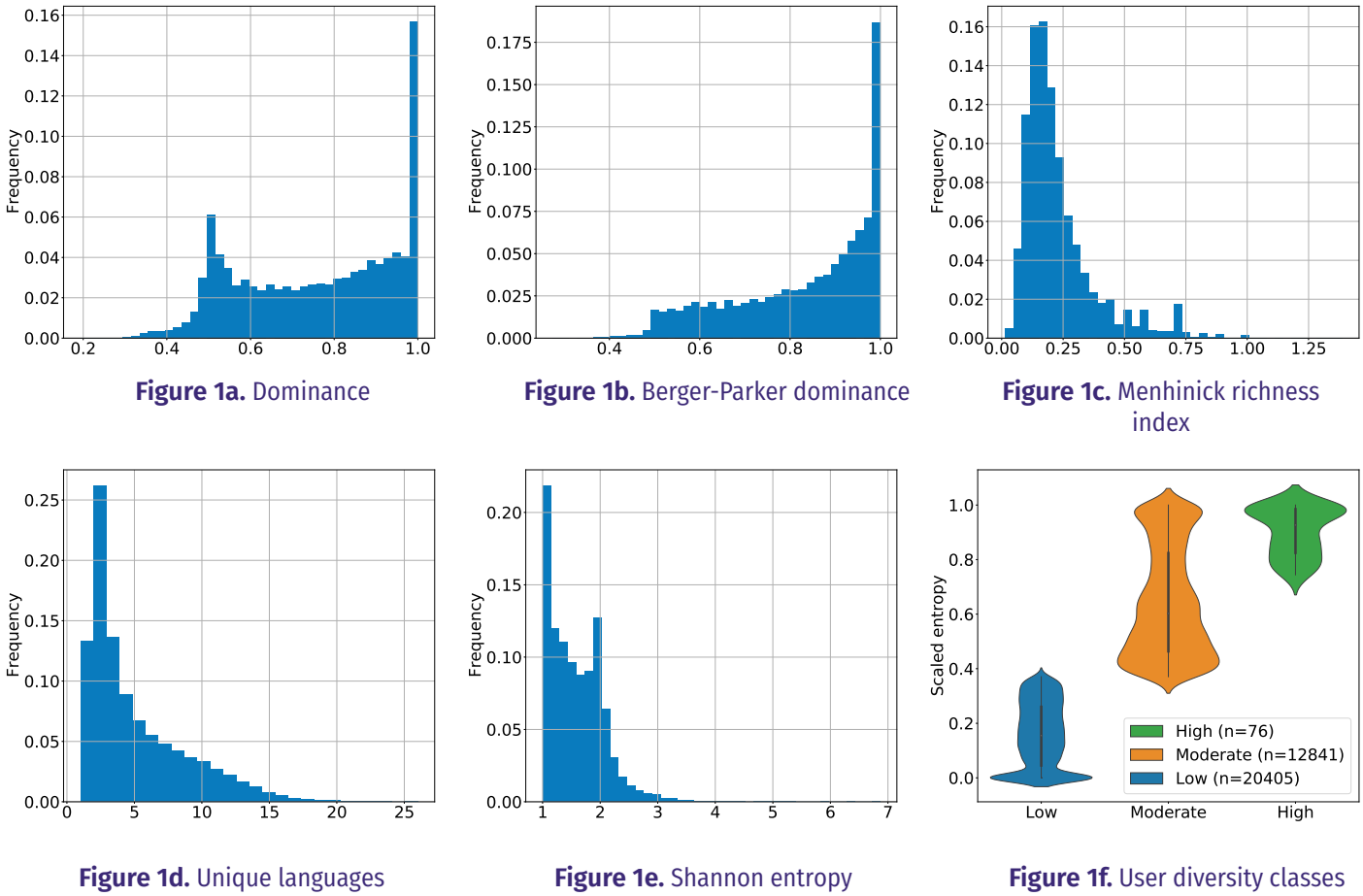

Figure 1a. Dominance

Figure 1b. Berger-Parker dominance

Figure 1c. Menhinick richness index

Figure 1d. Unique languages

Figure 1e. Shannon entropy

Figure 1f. User diversity classes

Figure 1. Diversity measures for languages in the tweet histories of individual users.

the users' linguistic repertoires are not equally balanced, but one language is always more prominent than others.

Berger-Parker dominance in Figure 1b, which corresponds to the proportion of dominant language, indicates that roughly 18% of Finnish users post in a single language on the platform. In contrast, the observations in the range between 0.3 and 0.99 suggests that the majority of users actively draw on multiple languages. Menhinick's richness index in Figure 1c, which captures the relation between the languages observed and the number of observations made, indicates that for most users, the linguistic repertoire is not as rich as the number of unique languages in Figure 1d suggests. In other words, for most users, the majority of observations fall within few languages. Shannon entropy in Figure 1e gives the amount of information needed to represent the distribution of languages observed in the user's tweet

history, which provides a measure of diversity (see also Coats 2019*b*). Overall, the diversity measures in Figures 1a–e suggest that most Finnish Twitter users draw on multiple languages to communicate on the platform, but these languages are not used in a balanced manner.

To explore the relative proportions of languages used by individual users, the violin plot in Figure 1f places the 33 322 users into high (n = 76), moderate (n = 12 841) and low (n = 20 405) classes based on their linguistic diversity. This method, which was originally developed for estimating the racial diversity of geographical areas, measures diversity using scaled entropy, whose values range from 0 to 1 (Holloway et al. 2012). Table 2 provides a closer look at this measure by showing the dominant languages and their proportion for each diversity class. To exemplify, 19.06% of users belonging to the moderate diversity class post most frequently in Finnish, whereas the corresponding number for the low diversity class is 33.65%.

Table 2 shows that most users in the low diversity class (33.65%) prefer to communicate in Finnish, whereas 21.46% prefer English. Figure 1f reveals that this class contains both monolingual users (scaled entropy = 0) and users who occasionally draw on other than the dominant language. Interestingly, users who prefer Spanish (1.38%) or Russian (1.29%) outnumber those that prefer Swedish (0.98%), the second official language of Finland. The moderate diversity class, whose members prefer to communicate using multiple languages, shows a smaller difference between Finnish (19.06%) and English (15.60%). Just as in the low diversity class, users in the moderate class exhibit a strong preference for Finnish and English, as the proportion of remaining languages is negligible compared to these two languages. Finally, the small size of the high diversity class is likely to result from the original purpose of the method in urban studies. Users in the high diversity class must have a scaled entropy value greater than or equal to 0.74 and no language can make up more than 45% of the sentences in the user's linguistic repertoire (Holloway et al. 2012: 69). Unlike the racial diversity of a population, this kind of 'balanced' linguistic diversity is extremely rare among Finnish Twitter users.

Figure 2 shows the spatial distribution of low and moderate diversity classes across Finland. Figure 2a reveals that low diversity users are more

Table 2. Proportion of users and number of sentences across the diversity classes and ten most common languages.

Diversity class	**High**		**Moderate**		**Low**	
Dominant language	Users	Sentences	Users	Sentences	Users	Sentences
Finnish	-	2 631	19.06 %	2 975 623	33.65 %	20 270 277
English	-	2 053	15.60 %	2 498 074	21.46 %	9 542 699
Swedish	-	1 814	1.07 %	226 015	0.98 %	551 298
Spanish	-	309	0.43 %	78 673	1.38 %	549 717
Russian	-	882	0.41 %	42 235	1.29 %	531 887
Japanese	-	0	0.09 %	39 582	0.34 %	226 594
French	-	33	0.16 %	27 658	0.20 %	116 143
Portuguese	-	0	0.08 %	7 234	0.41 %	143 075
German	-	81	0.07 %	11 542	0.07 %	35 489
Estonian	-	0	0.13 %	19 782	0.13 %	37 569
Other	-	586	1.44 %	154 166	1.34 %	367 666
Total	0.29 %	8 389	38.54 %	6 080 584	61.24 %	32 372 414

likely to be found in rural municipalities and regions. To some extent, these areas correspond to those where the use of Finnish is above the country-wide average (see Figure 5a). Contrasting the distribution of the low diversity class with the moderate diversity class in Figure 2b reveals that the observations mirror each other: users that belong to the moderate diversity class are less likely to be found in areas where the proportion of low diversity users is above the country-wide average. Although some rural areas appear balanced in terms of diversity classes, that is, they do not diverge from the average for low and moderate classes, there appears to be a contrast between coastal, urban and rural areas.

4.2. The linguistic diversity and richness of regions and municipalities

Having described the richness and diversity of individual users, we now turn towards regions (n = 19) and municipalities (n = 25), starting with the spatial distribution of unique languages, or linguistic richness in Figure 3a.

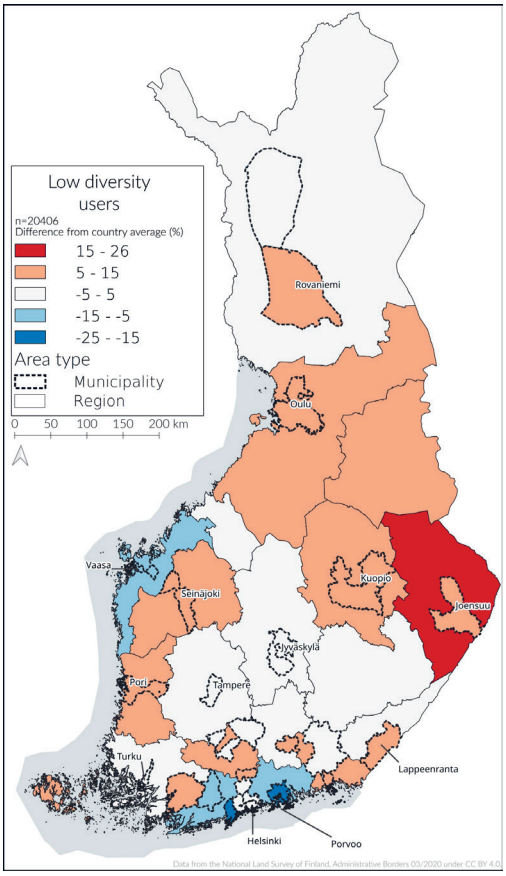

Figure 2a. Low diversity class

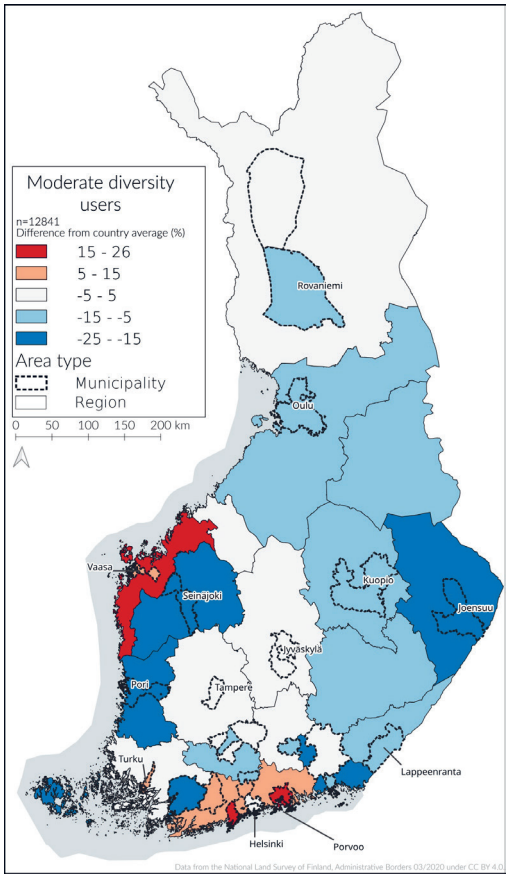

Figure 2b. Moderate diversity class

Figure 2. Predicted home locations for users in low and moderate diversity classes.

The cities of Helsinki and Espoo are the richest with 71 to 91 languages. Most regions and municipalities in the fastest growing area in Finland, roughly demarcated by a triangle formed by the cities of Helsinki, Tampere and Turku, feature from 53 to 70 languages. Coastal regions and municipalities are generally richer than northern and inland regions. Cities that function as regional centres (and host higher education institutions), such as Lappeenranta, Jyväskylä, Kuopio, Oulu, Rovaniemi and to lesser extent Joensuu, are an exception to this pattern. To summarise, in terms of linguistic richness, there is a contrast between coastal and inland regions, and cities and countryside. It should be noted, however, that even the linguistically 'poorest' regions feature 19 languages.

The Berger-Parker dominance index in Figure 3b, which corresponds to the proportion of dominant language out of all observations made, reveals that dominance is lowest around Vaasa and the surrounding region of Ostrobothnia, which are home to Finnish-Swedish communities that speak a regional variety of Swedish (Østern 2001). Finland Swedish communities are

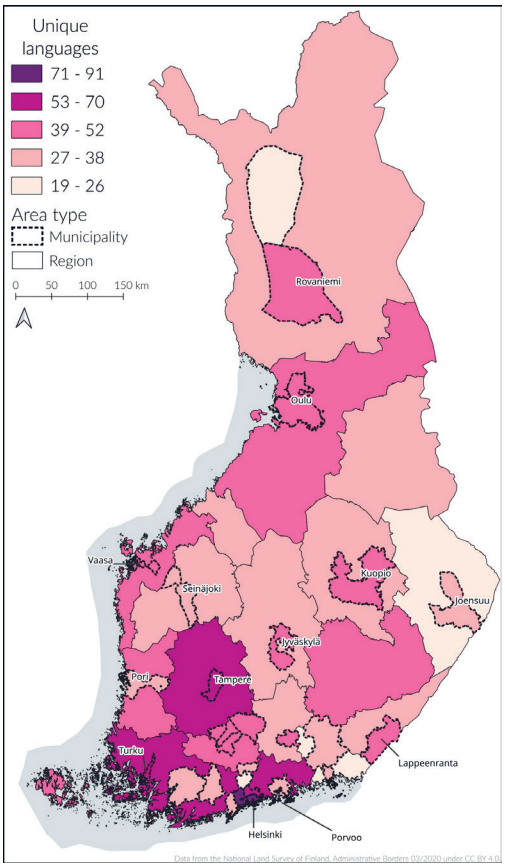

Figure 3a. Unique languages

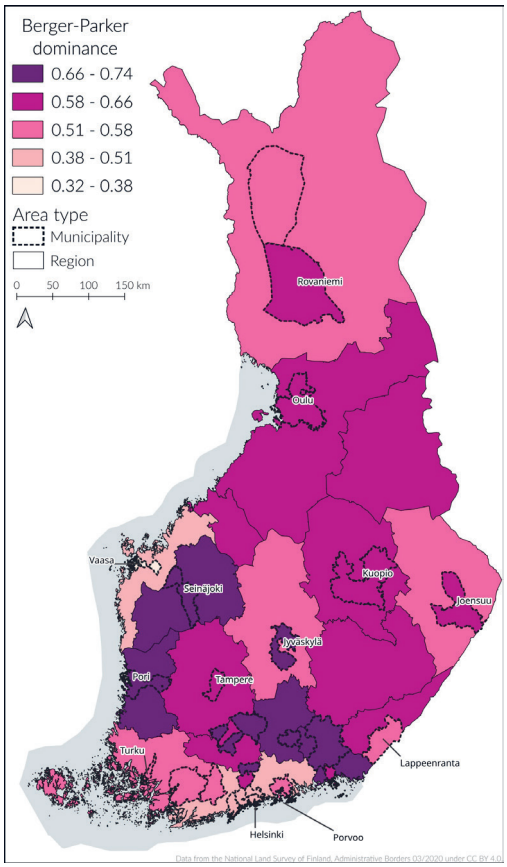

Figure 3b. Berger-Parker dominance

Figure 3. Unique languages and Berger-Parker dominance per spatial unit.

also located on the southern coast in the municipalities surrounding Helsinki and Porvoo, which feature a low dominance index (see Figure 6a). In Vaasa and Ostrobothnia, the low Berger-Parker dominance index is likely to reflect the active use of three languages, Swedish, Finnish and English, as the Finland Swedish communities also draw on English as a lingua franca (Sjöholm 2004). The same applies to the Swedish-speaking municipalities along the southern coast, although for the Helsinki Metropolitan Area, the low value for Berger-Parker dominance index is likely to be explained by urban multilingualism (see e.g. Lehtonen, 2016) and linguistic richness (see Figure 3a).

In contrast to the coastal areas with Finland Swedish populations, the regions around Seinäjoki, Pori and south of Jyväskylä show the highest values for the Berger-Parker dominance index. A similar view is provided by the dominance index and Shannon entropy in Figure 4. Taken together, these indices suggest a preference for the Finnish language, as revealed by Figure 5a, which shows that these regions feature more tweets in Finnish and fewer

in English compared to the country-wide average. This low-diversity belt, which extends across Finland, stands in strong contrast to the linguistically diverse coastal areas, as indicated by the high values for Shannon entropy. Interestingly, Oulu, which is a main regional centre with strong international connections due to higher education and technology industries, has a surprisingly low linguistic diversity.

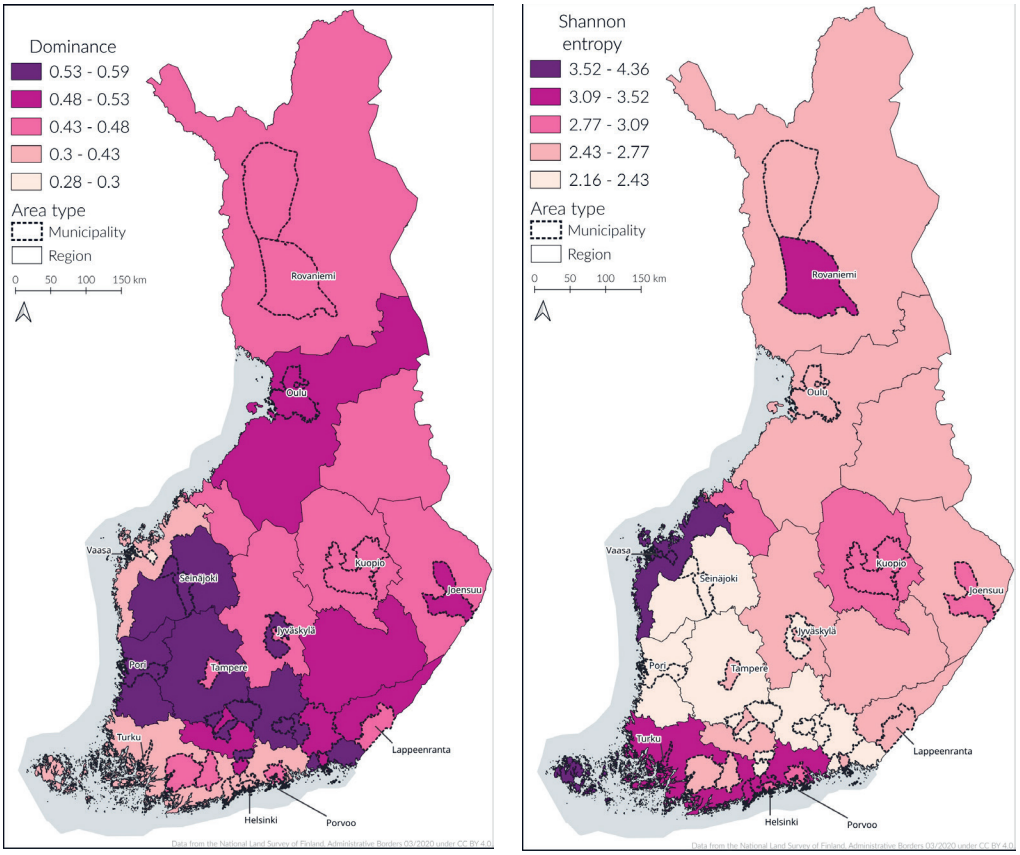

Figure 4a. Dominance

Figure 4b. Shannon entropy

Figure 4. Dominance and Shannon entropy per spatial unit.

4.3. The spatial distribution of individual languages

To explore how the languages used on Twitter differ across regions and municipalities in Finland, we first calculated an average for each language across the entire country and then determined how much each region or municipality diverges from the country-wide average. The calculations were based on tweets with coordinate geotags, which were split into orthographic sentences before language identification. Figure 5 shows that observations for the most common languages, Finnish and English, are evenly distributed

across Finland, apart from the areas where users tweet more in Finnish or Swedish. The Finland Swedish regions of Åland Islands, Vaasa and Ostrobothnia feature fewer tweets in Finnish compared to the rest of the country, whereas Finnish is used slightly more in the low-diversity belt that extends across Finland (see Figure 4b).

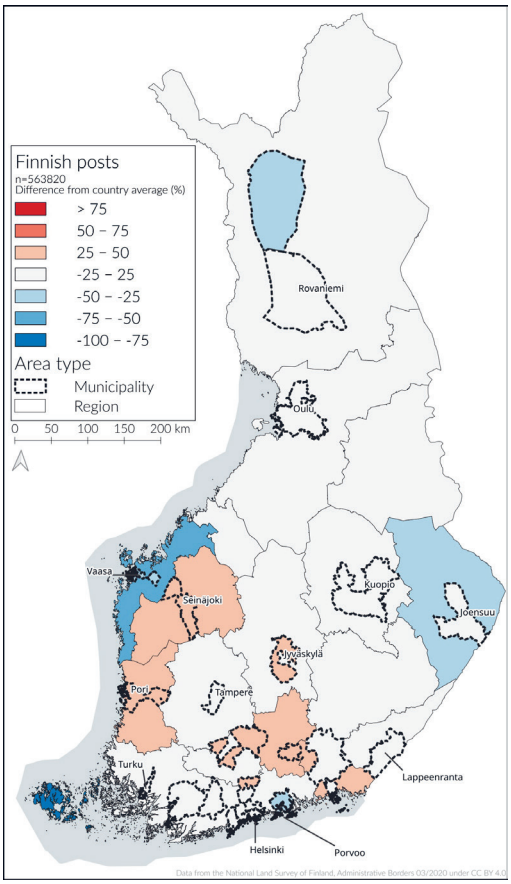

Figure 5a. Finnish

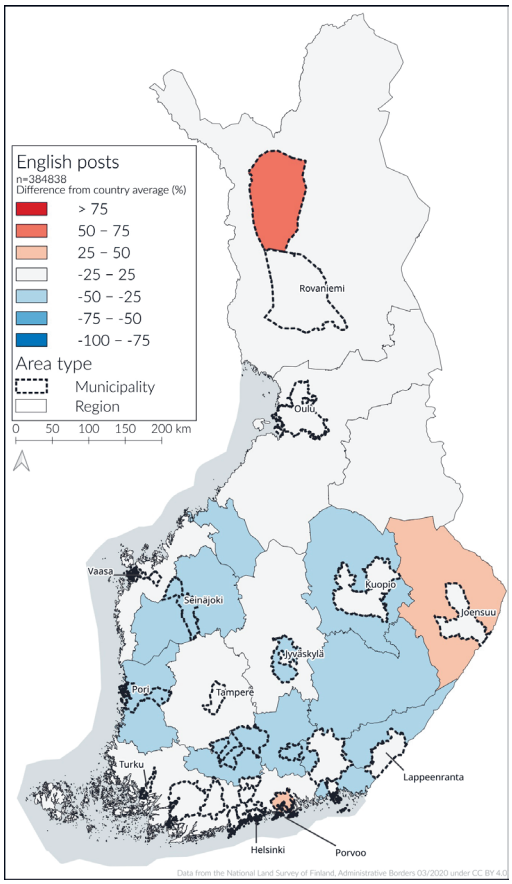

Figure 5b. English

Figure 5. The spatial distribution of tweets in Finnish and English among Twitter users in Finland.

Much has been written about the status of English in Finland in recent decades, particularly in relation to the tension between Finnish and English (see e.g. Taavitsainen & Pahta 2003, 2008; Leppänen et al. 2011). Contrasting the observations for Finnish (Figure 5a) and English (Figure 5b) shows that in most areas, the use of both languages is within the country-wide average. There are, however, regions and municipalities where Finnish is used more and English is used less, which largely coincide with rural areas of low linguistic diversity (see Figure 4b). Although postulating a division into

'have-nots', 'haves' and 'have-it-alls' based on age, educational background and location is tempting (Leppänen et al. 2011: 165), such a division cannot be derived from spatial information alone, but would require a more detailed estimation of the users' geotemporal demographics (Longley & Adnan 2016).

Figure 6a shows the distribution of tweets in Swedish, which are closely aligned with the locations of the Finland Swedish communities (Sjöholm 2004: 640). The contrast between these coastal regions and municipalities and the rest of Finland is rather striking. The coastal communities have strong social and historical ties with Sweden, which have been argued to form a 'transnational social field' between Sweden and Finland Swedes (Hedberg 2007). The Swedish language naturally plays a major role in maintaining this transnational space, which is likely to be reflected in language choices among Finland Swedes on Twitter. What is also remarkable that the use of English is within the national average in the Swedish-preferring areas (see Figure 5b). Compared to the regions where Finnish is preferred over English, the users in the Swedish-speaking areas do not seem to favour Swedish over English.

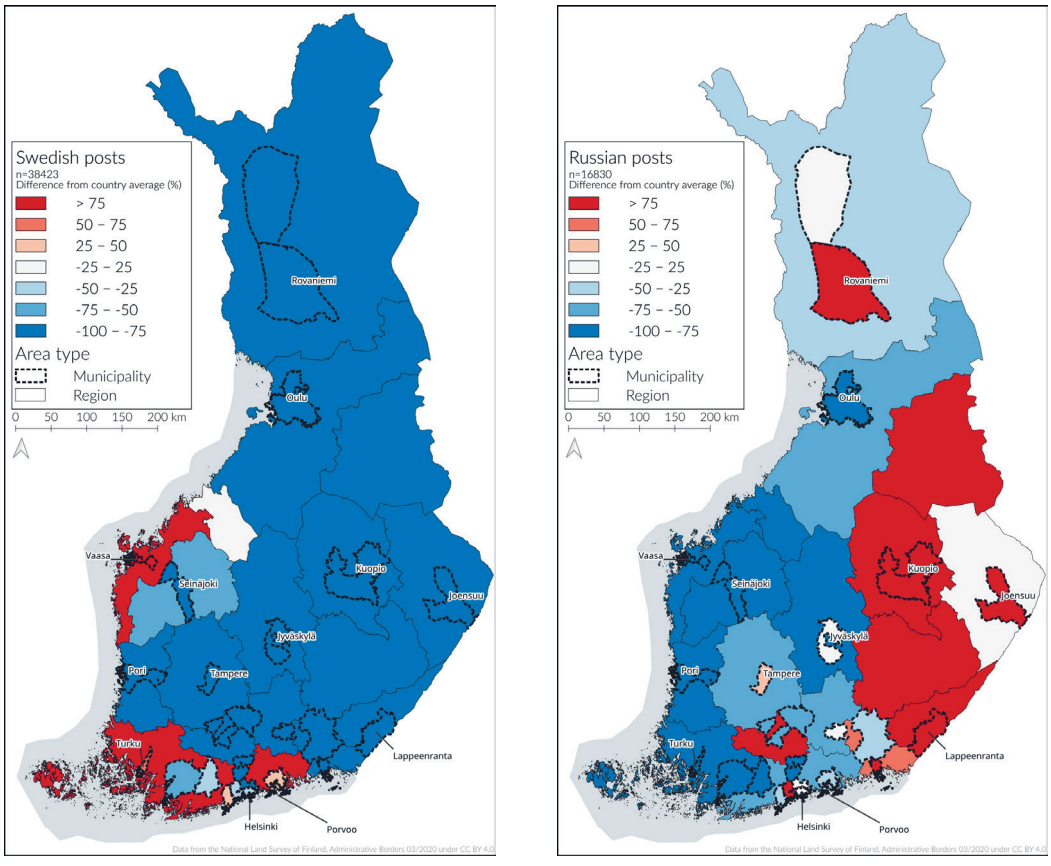

Figure 6a. Swedish

Figure 6b. Russian

Figure 6. The spatial distribution of tweets in Swedish and Russian among Twitter users in Finland.

The use of Russian, shown in Figure 6b is likely to reflect both historical and contemporary developments. Apart from Finns with Russian heritage, there have been several waves of migration from Russia to Finland in recent history. First, approximately 30 000 Ingrian Finns remigrated to Finland following the dissolution of the Soviet Union (Kyntäjä 1997), which may explain the use of Russian in the city of Espoo located west of Helsinki and the region of Tavastia Proper between Helsinki and Tampere. Second, the use of Russian in the regions and municipalities close to the Russian border, which spans from the city of Lappeenranta to the region of Kainuu, is likely to be explained by later migration of Russians to Finland, Russians who maintain a second home in Finland and their cross-border social networks. Moreover, Finnish businesses are likely to communicate with potential Russian customers in their native language, which may explain why Rovaniemi, a gateway for visitors to Lapland, stands out as well.

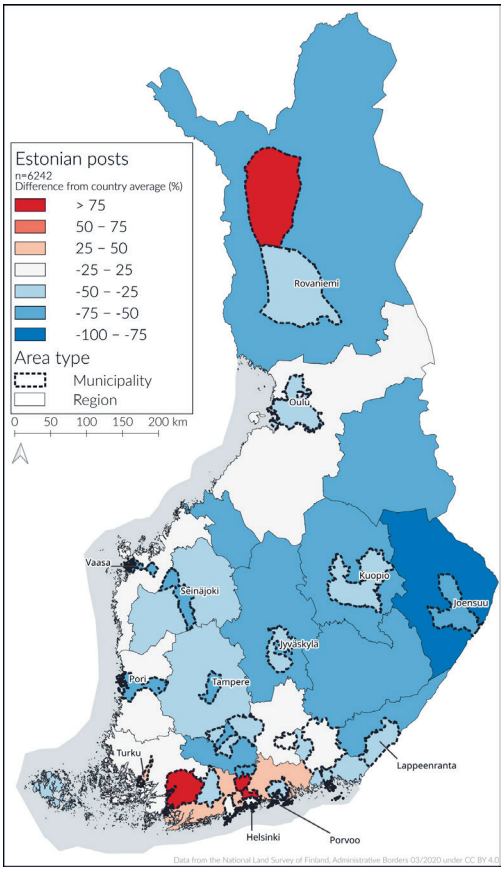

Figure 7a. Estonian

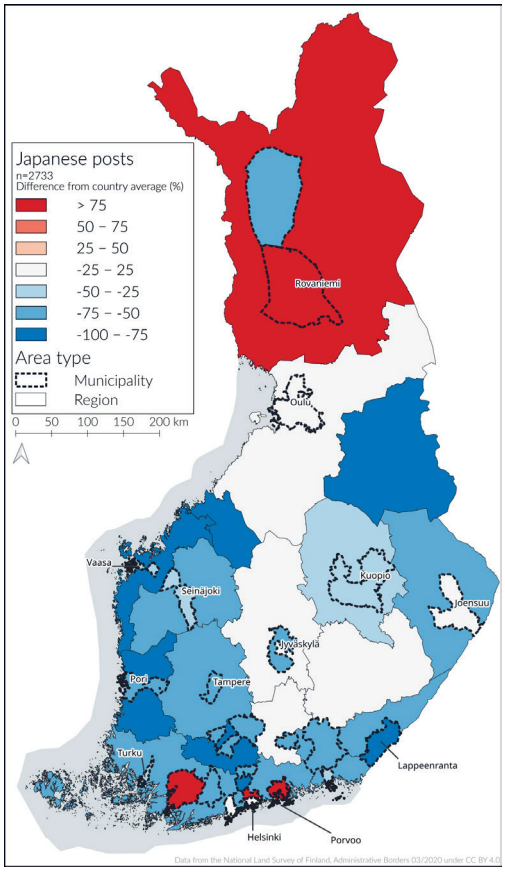

Figure 7b. Japanese

Figure 7. The spatial distribution of tweets in Estonian and Japanese among Twitter users in Finland.

Estonian, shown in Figure 7a, is used frequently in the Helsinki Metropolitan Area, the surrounding regions and municipalities, and the municipality of Kittilä in the Finnish Lapland. The prominence of the Estonian language in the capital region is likely to reflect the migration pattern of both permanent movers and transnational people, as some Estonians share their life between Finland and Estonia, and the Helsinki harbour is the main gateway between the two countries (see Silm et al. 2020). Apart from Kittilä, the spatial pattern for Estonian illustrates how transnational spaces and mobilities are intertwined: despite the proximity of the two countries, access to the main mode of transportation between Estonia and Finland appears to constrain the spatial distribution of Estonian. This becomes evident when contrasted with the spatial distribution of Russian in Figure 6b, which benefits from proximity and better accessibility.

Finally, Figure 7b shows the distribution of Japanese tweets in Finland, which illustrates the difficulty of interpreting spatial patterns for languages without a geographic (Swedish, Russian and Estonian) or linguistic frame of reference (Finnish, English). Although previous research can fill in some of these gaps, namely that the pattern for Japanese corresponds to domestic tourism preferences of Japanese residing in Finland (Matilainen & Saanalahti 2018), the small size of Japanese population raises the question whether the spatial pattern for Japanese reflects places of residence or visits. Alternatively, the data may be generated by businesses communicating with Japanese tourists, or Finns who study Japanese. To address these issues, one would have to conduct a detailed analysis of the linguistic content, in order to evaluate whether the geotags represent 'postcards' or 'ticket stubs' (Tasse et al. 2017).

4.4. Temporal changes across regions and municipalities

The previous analyses of richness, diversity and geographical distribution of languages revealed distinctive spatial patterns across regions and municipalities in Finland. This kind of spatial view, however, does not capture the inherently temporal nature of society, which is manifested in the daily, weekly and seasonal rhythms that govern much of our daily lives (Järv et al. 2015; Silm et al. 2020). Research on human geography and urban

multilingualism have often emphasised the temporal dimension of individual (linguistic) behaviour and activity, which also shapes the surrounding environment (Hägerstrand 1970; Golledge & Stimson 1997; Pennycook & Otsuji 2015).

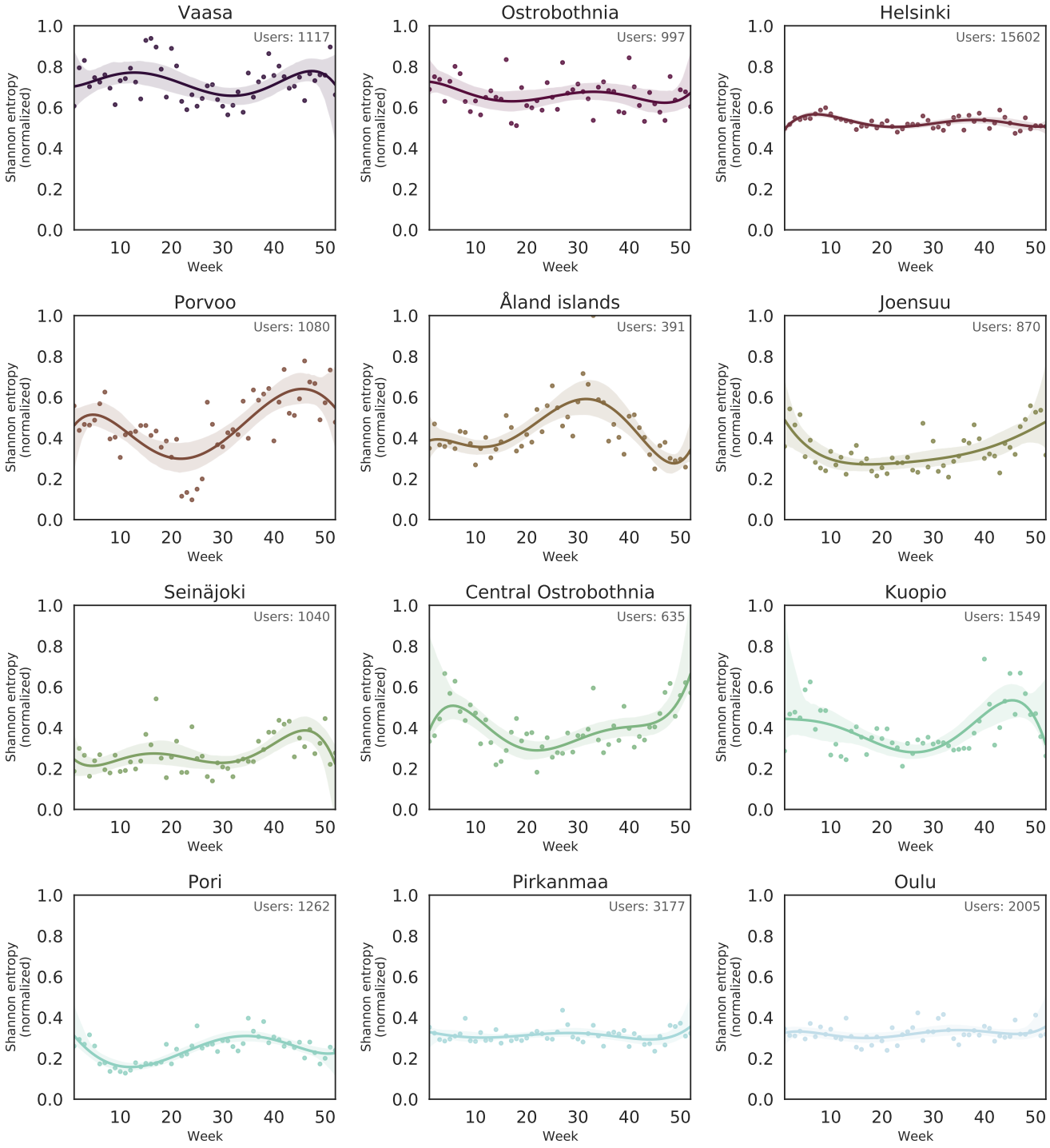

Figure 8. Normalised Shannon entropy over 52 weeks across 12 different spatial units.

To analyse temporal changes in linguistic diversity, we fitted a fifth-order polynomial regression to the values for Shannon entropy over 52 weeks at selected locations, which are visualised in Figure 8. The entropy values have been normalised across locations into the range from 0 to 1. The shaded areas represent 99.9% confidence intervals that have been estimated using 1000 bootstrapped samples. Put differently, the mean value lies within the shaded area with 99.9% probability.

Figure 8 reveals that linguistic diversity fluctuates over time. Regions and municipalities exhibit different temporal patterns, which are likely to emerge from social and environmental differences, such as population size, diversity and geographical location. Moreover, some areas may also be affected by seasonal holidays and domestic tourism, semesters in educational institutions and annual events that increase the number of short-term visitors.

Beginning from top left, the city of Vaasa and the surrounding rural region of Ostrobothnia are the most diverse in Finland (see also Figure 4b). Whereas Vaasa shows a moderate seasonal rhythm, the surrounding region of Ostrobothnia does not have a distinctive seasonal pattern. Although both Vaasa and Ostrobothnia are predominantly Swedish-speaking areas (see Figure 6a), the difference in temporal patterns are likely to be caused by the contrast between urban and rural environments. Vaasa hosts a sizeable student population, which is large enough to generate a strong seasonal pattern.

The capital city of Helsinki remains moderately diverse throughout the year. Helsinki's status as the largest city in Finland is reflected in the high number of observations compared to other sites (n = 15 602). The high number of observations allows estimating the mean value with greater accuracy, as reflected by the smaller confidence intervals. The absence of a clear temporal pattern is not surprising due to the number of observations, the weekly temporal scale and the extent of the geographical area covered: any temporal patterns are lost in the 'noise' from the surrounding activity in the city. For this reason, identifying temporal patterns would require zooming in at specific locations within the city (cf. e.g. Hiippala et al. 2019).

The remaining areas in Figure 8 feature various temporal patterns. The municipality of Porvoo and region of Åland reflect high seasonality that are

likely to result from tourism seasons in winter and summer, respectively. Regional centres with higher education institutions, such as Joensuu, Seinäjoki and Kuopio, and the region of Central Ostrobothnia feature seasonal peaks towards the end of the year. Curiously, the city of Oulu, a major regional centre and a host to several higher education institutions, and the region of Pirkanmaa are an exception to this pattern. The city of Pori, in turn, hosts political and musical festivals each summer, which may explain the increase in diversity during summertime.

5. Discussion

In this study, we sought to combine perspectives from human geography with methods from geoinformatics in order to better understand the richness and diversity of languages used on Twitter in Finland and their spatiotemporal characteristics. Our analysis revealed that what is often colloquially called the 'Finnish Twitter' is a diverse linguistic community that varies from one geographical location to another. These findings represent a significant advance over the level of detail provided by previous country-level analyses (see e.g. Coats 2019*a*, 2019*b*). Many Finnish Twitter users draw on multiple languages to communicate on the platform, but the combination of languages and their proportions vary between users (see Table 2). In addition to producing new knowledge about the linguistic repertoires of individual users, our method enabled making observations about language use on Twitter that correspond to the everyday linguistic realities in Finland, such as the prominent role of English and the geographical distribution of Swedish, Russian and Estonian, and their proximity to their respective transnational 'zones of influence' (Hedberg 2007; Silm et al. 2020).

Although detailed linguistic analyses fell outside the scope of this study, future studies can build on our work to create stratified corpora tailored for answering particular research questions. To exemplify, our results could be used to build corpora for studying the use of English as a foreign language among users with low linguistic diversity that reside in areas where the use of Finnish is above the country-wide average. Geography- and diversity-based stratification allows imposing structure on large volumes of Twitter data, which offers ample opportunities for studying the features of English on

Twitter in Finland (Coats 2016) or the extending the study of Finnish dialects to Twitter (Hyvönen et al. 2007), to name just a few examples.

Given that our findings are based on the application of automatic language identification (Bojanowski et al. 2017) and place of residence detection (Massinen 2019), which are prerequisites for working with high volumes of social media data, they must be considered in light of the methods used. What comes to the multilingual nature of Twitter in Finland, Pennycook and Otsuji (2015: 47) have argued that the multilingual reality of everyday life resists quantification, because it does not respect definitions of abstract definitions of languages such as 'Finnish' or 'English'. Much of language technology, such as automatic language identification, relies on such abstractions, which obviously limits its capability to provide insights on multilingualism at the level of an individual. Furthermore, languages are not equally resourced in terms of language technology: because the language identification model we used did not support Sámi, we could not analyse its use (Cocq 2015).

Going forward, there is considerable potential in recent techniques developed within computational sociolinguistics, which move away from treating languages as abstract, categorical entities by learning formal distinctions between languages directly from the data (see e.g. Purschke & Hovy 2019). This also warrants considering what should be used as a yardstick for evaluating how well computational methods are able to account for multilingualism. Embodied interactions taking place at physical locations in a city, as described by Pennycook and Otsuji (2015: 47), should not be equated with mediated interaction on social media platforms such as Twitter, as these two communicative situations differ considerably in their characteristics (Bateman et al. 2017: 107–110). For this reason, Twitter should be considered just one platform within the double space that encompasses both physical and virtual environments (Crampton et al. 2013; Kellerman 2016; Hiippala et al. 2019). Any conclusions drawn about language use must account for the platform and the demographics of its users.

Regardless of how social interaction takes place, the spatiotemporal nature of human activity and the dynamic that underlies our societies remains the same. This study has clearly shown that space and time matter in relation to language use on social media platforms and beyond, which is why we advocate bringing conceptual frameworks, such as activity spaces and time

geography, into closer dialogue with linguistics. We thus strongly agree with Derungs et al. (2020) that methods for space-time analyses developed in the field of geoinformatics can find productive applications in linguistic research. These contributions are not limited to quantitative studies, as Larsson et al. (2020) have recently shown by integrating geographical information into qualitative analysis of social interaction. Geovisualization techniques, in particular, can offer interactive tools for exploring the spatiotemporal dynamics of linguistic phenomena. To summarise, geographical insights can provide linguists with new perspectives on language choice, multilingualism and factors affecting linguistic richness and diversity. For geographers, who often consider languages as markers of cultural identity and community membership, linguistic perspectives can offer a deeper understanding of the relationship between language use, individuals and societies (Alexander et al. 2007; Järv et al. 2015).

It should also be acknowledged that the future of scholarly social media research is uncertain. Bruns (2019) provides a succinct overview of the developments following the Cambridge Analytica scandal, which caused most social media platforms to shut down the APIs that allowed programmatic access to user-generated content. Public APIs that allowed critical research have been replaced by vague promises of data access and 'corporate data philantropy', which allows the platforms to influence the kind of research conducted using their content (Bruns, 2019: 1551–1552). This has led researchers to consider alternatives that may violate the platforms' terms of service, such as 'scraping' the content via their web interfaces (Freelon 2018). In short, as one of the remaining platforms with a public API, Twitter is currently an invaluable source of data.

Nevertheless, social media is not the only source of big data for conducting research at the intersection of linguistics and geography. Largely untapped sources of data include location information from mobile phones for mapping user flows and activity spaces, combined with application usage information to understand the platforms that make up the virtual environment in the double space. These observations can be then mapped to register data with information about languages spoken in given geographical areas and a rich array of demographic factors. Like social media, working with such sources

of data requires researchers to adhere to highest ethical standards when conducting research (Zook et al. 2017).

6. Conclusion

In this article, we reported on an empirical, data-driven study of the languages used on Twitter by Finnish users and their richness, diversity and geographical distribution. By combining automatic language identification and place of residence detection, we showed that language use on the platform may be characterised as rich, diverse and inherently multilingual, and it exhibits geographical and temporal variation at the level of municipalities and regions. The observations based on social media data largely reflect everyday multilingual realities in Finland. The coastal areas where Swedish is spoken as a minority language are more diverse than the rural areas located inland. Urban and rural areas also differ in terms of temporal patterns, which reflect the rhythm of human activity over daily, weekly and seasonal timescales. Given the spatiotemporal underpinnings of all human interaction, we suggest that there is much potential in integrating perspectives from human geography, geoinformatics and linguistics, in order to better understand the relationships between language, people and places over time in Finland and abroad. N

TUOMO HIIPPALA
TUOMAS VÄISÄNEN
TUULI TOIVONEN
OLLE JÄRV

UNIVERSITY OF HELSINKI

* The authors thank the Emil Aaltonen Foundation for funding this research.

References

ALEXANDER, Claire, Rosalind Edwards & Bogusia Temple 2007. Contesting cultural communities: language, ethnicity and citizenship in Britain. *Journal of Ethnic and Migration Studies* 33 (5): 783–800. doi: 10.1080/13691830701359223.

ARTAMONOVA, Olga & Jannis Androutsopoulos 2019. Smartphone-based language practices among refugees: mediational repertoires in two families. *Journal für Medienlinguistik* 2(2): 60–89. doi: 10.21248/*jfml*.2019.14.

AUER, Peter, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.) 2014. *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*. Berlin and Boston: De Gruyter.

BATEMAN, John A., Janina Wildfeuer & Tuomo Hiippala 2017. *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

BIBER, Douglas 1993 Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.

BIRD, Steven, Ewan Klein & Edward Loper 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.

BOJANOWSKI, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135–146.

BOUVIER, Gwen and David Machin 2018. Critical discourse analysis and the challenges and opportunities of social media. *Review of Communication* 18(3): 178–192. doi: 10.1080/15358593.2018.1479881.

BRUNS, Axel 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society* 22(11): 1544–1566. doi: 10.1080/1369118X.2019.1637447.

COATS, Steven 2016. Grammatical feature frequencies of English on Twitter in Finland. *English in Computer-Mediated Communication: Variation, Representation and Change*, ed. Lauren Squires. 179–210. Boston and Berlin: De Gruyter.

COATS, Steven 2019*a*. Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics* 42(1): 31–55.

COATS, Steven 2019*b*. Online language ecology: Twitter in Europe. *Building Computer-Mediated Communication Corpora for Sociolinguistic Analysis*, eds. Egon Stemle & Ciara Wigham. 73–96. Clermont-Ferrand: Presses universitaires Blaise Pascal.

COCQ, Coppélie 2015 Indigenous voices on the web: folksonomies and endangered languages. *Journal of American Folklore* (128)509: 273–285.

CRAMPTON, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson & Matthew Zook 2013. Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science* 40(2): 130–139. doi: 10.1080/15230406.2013.777137.

DERUNGS, Curdin, Christian Sieber, Elvira Glaser & Robert Weibel (2020) Dialect borders – political regions are better predictors than economy or religion.

Digital Scholarship in the Humanities 35(2): 276–295. doi: 10.1093/llc/fqz037.

DODGE, Martin & Rob Kitchin 2005. Code and the transduction of space. *Annals of the Association of American Geographers* 95(1): 162–180.

EISENSTEIN, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing 2014. Diffusion of lexical change in social media". *PLOS ONE* 9(11): 1–13. doi: 10.1371/journal.pone.0113114.

FREELON, Deen 2018. Computational research in the post-API age. *Political Communication* 35(4): 665–668. doi: 10.1080/10584609.2018.1477506.

GIDDENS, Anthony 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley, CA: University of California Press.

GOLLEDGE, Reginald G. & R. J. Stimson 1997. *Spatial Behavior: A Geographic Perspective*. Guilford Press: New York, NY.

GRAHAM, Mark, Matthew Zook & Andrew Boulton 2013. Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Institute of British Geographers* 38(3): 464–479. doi: 10.1111/j.1475-5661.2012.00539.x.

GRIEVE, Jack 2017. Spatial statistics for dialectology. *The Handbook of Dialectology*, eds. Charles Boberg, John Nerbonne & Dominic Watt. 415–433. Oxford: Wiley.

GRIEVE, Jack, Dirk Speelman & Dirk Geeraerts 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2): 193–221. doi: 10.1017/S095439451100007X.

GRIEVE, Jack, Andrea Nini & Diansheng Guo 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46(4): 293–319. doi: 10.1177/0075424218793191.

HÄGERSTRAND, Torsten 1970. What about people in regional science? *Papers of the Regional Science Association* 24(1): 6–21. doi: 10.1007/BF01936872.

HALLIDAY, Michael A. K. 1978. *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. London: Arnold.

HECHT, Brent, Lichan Hong, Bongwon Suh & Ed H. Chi 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*. 237–246. Vancouver, BC, Canada: Association for Computing Machinery. doi: 10.1145/1978942.1978976.

HEDBERG, Charlotta 2007. Direction Sweden: migration fields and cognitive distances of Finland Swedes. *Population, Space and Place* 13(6): 455–470. doi: 10.1002/psp.462.

HEIKINHEIMO, Vuokko, Henrikki Tenkanen, Claudia Bergroth, Olle Järv, Tuomo Hiippala & Tuuli Toivonen 2020. Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning* 201(103845). doi: 10.1016/j.landurbplan.2020.103845.

HERDAĞDELEN, Amaç 2013. Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation* 47(4): 1127–1147. doi: 10.1007/s10579-013-9227-2.

HIIPPALA, Tuomo, Anna Hausmann, Henrikki Tenkanen & Tuuli Toivonen 2019. Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*

34(2): 290–309. doi: 10.1093/llc/fqy049.

HOCHMAIR, Hartwig H., Levente Juhász & Sreten Cvetojevic 2018. Data quality of points of interest in selected mapping and social media platforms. *Progress in Location Based Services* 2018, eds. Peter Kiefer, Haosheng Huang, Nico Van de Weghe & Martin Raubal. 293–313. Cham: Springer.

HOLLOWAY, Steven R., Richard Wright & Mark Ellis 2012. The racially fragmented city? Neighborhood racial segregation and diversity jointly considered. *The Professional Geographer* 64(1): 63–82. doi: 10.1080/00330124.2011.585080.

HU, Yingjie & Ruo-Qian Wang 2020. Understanding the removal of precise geotagging in tweets". *Nature Human Behaviour*. doi: 10.1038/s41562-020-00949-x.

HUMPHREYS, Lee 2010. Mobile social networks and urban public space. *New Media & Society* 12(5): 763–778.

HUMPHREYS, Lee & Tony Liao 2011. Mobile geotagging: reexamining our interactions with urban space. *Journal of Computer-Mediated Communication* 16(3): 407–423.

HYVÖNEN, Saara, Antti Leino & Marko Salmenkivi 2007. Multivariate analysis of Finnish dialect data – an overview of lexical variation. *Literary and Linguistic Computing* 22(3): 271–290.

JÄRV, Olle, Kerli Müürisepp, Rein Ahas, Ben Derudder & Frank Witlox 2015. Ethnic differences in activity spaces as a characteristic of segregation: a study based on mobile phone usage in Tallinn, Estonia. *Urban Studies* 52(14): 2680–2698.

JÄRV, Olle, Henrikki Tenkanen, Maria Salonen, Rein Ahas & Tuuli Toivonen 2018. Dynamic cities: location-based accessibility modelling as a function of time. *Applied Geography* 95: 101–110.

KELLERMAN, Aharon 2010. Mobile broadband services and the availability of instant access to cyberspace. *Environment and Planning A* 42: 2990–3005.

KELLERMAN, Aharon 2014. The satisfaction of human needs in physical and virtual spaces. *The Professional Geographer* 66(4): 538–546.

KELLERMAN, Aharon 2016. *Daily Spatial Mobilities: Physical and Virtual*. New York and London: Routledge.

KISS, Tibor & Jan Strunk 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics 32(4): 485–525. doi: 10.1162/coli.2006.32.4.485.

KITCHIN, Rob 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society* (1)1: 1–12. doi: 10.1177/2053951714528481.

KYNTÄJÄ, Eve 1997. Ethnic remigration from the former Soviet Union to Finland – patterns of ethnic identity and acculturation among the Ingrian Finns. *Finnish Yearbook of Population Research* 34: 102–113. doi: 10.23979/fypr.44924.

LAITINEN, Mikko, Jonas Lundberg, Magnus Levin & Alexander Lakaw 2017. Revisiting weak ties: using present-day social media data in variationist studies. *Exploring Future Paths for Historical Sociolinguistics*, eds. Tanja Säily, Arja Nurmi, Minna Palander-Collin, and Anita Auer. 303–325. Amsterdam: Benjamins.

LAITINEN, Mikko, Jonas Lundberg, Magnus Levin & Rafael Martins 2018. The Nordic Tweet Stream: a dynamic real-time monitor corpus of big and rich language data. *Proceedings of the*

Digital Humanities in the Nordic Countries 3rd Conference, eds. Eetu Mäkelä, Mikko Tolonen & Jouni Tuominen. 349–362. Helsinki, Finland.

LARSSON, Jens, Niclas Burenhult, Nicole Kruspe, Ross S. Purves, Mikael Rothstein & Peter Sercombe 2020. Integrating behavioral and geospatial data on the timeline: towards new dimensions of analysis. *International Journal of Social Research Methodology*. doi: 10.1080/13645579.2020.1763705.

LEE, Carmen 2016. Multilingual resources and practices in digital communication. *The Routledge Handbook of Language and Digital Communication*, eds. Alexandra Georgakopoulou & Tereza Spilioti. 118–132. New York and London: Routledge.

LEETARU, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan & Eric Shook 2013. Mapping the global Twitter heartbeat: the geography of Twitter. *First Monday* 18(5). doi: 10.5210/fm.v18i5.4366.

LEHTONEN, Heini 2016. What's up Helsinki? Linguistic diversity among suburban adolescents. *Linguistic Genocide or Superdiversity? New and Old Language Diversities*, eds. Reetta Toivanen & Janne Saarikivi. 65–90. Bristol: Multilingual Matters.

LEPPÄNEN, Sirpa, Anne Pitkänen-Huhta, Tarja Nikula, Samu Kytölä, Timo Törmäkangas, Kari Nissinen, Leila Kääntä, Tiina Räisänen, Mikko Laitinen, Päivi Pahta, Heidi Koskela, Salla Lähdesmäki & Henna Jousmäki 2011. *National Survey on the English Language in Finland: Uses, Meanings and Attitudes*. Helsinki: University of Helsinki.

LJUBEŠIĆ, Nikola, Maja Miličević Petrović & Tanja Samardžić 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography* 6(2): 100–124. doi: 10.1017/jlg.2018.9.

LONGLEY, Paul A. & Muhammad Adnan 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science* 30(2): 369–389. doi: 10.1080/13658816.2015.1089441.

MARTÍ, Pablo, Leticia Serrano-Estrada & Almudena Nolasco-Cirugeda 2019. Social media data: challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems* 74: 161–174. doi: 10.1016/j.compenvurbsys.2018.11.001.

MASSINEN, Samuli 2019. *Modeling Cross-Border Mobility Using Geotagged Twitter in the Greater Region of Luxembourg*. MA thesis. Faculty of Science, University of Helsinki.

MATILAINEN, Anne & Sanna Saanalahti 2018. Finland as a tourist destination through the eyes of the Japanese: an interview study of Japanese people living in Finland. Ruralia Institute Reports 180. Mikkeli and Seinäjoki: Ruralia Institute, University of Helsinki.

NERBONNE, John & William A. Kretzschmar, Jr. 2013. Dialectometry++. *Literary and Linguistic Computing 28*(1): 2–12.

NGUYEN, Dong, A. Seza Doğruöz, Carolyn P. Rosé & Franciska de Jong 2016. Computational sociolinguistics: a

survey". *Computational Linguistics* 42(3): 537–593.

ØSTERN, Anna 2001. Swedish in Finland. *The Other Languages of Europe*, eds. Guus Extra & Durk Gorter. 159–173. Clevedon: Multilingual Matters.

PENNYCOOK, Alastair & Emi Otsuji 2014. Metrolingual multitasking and spatial repertoires: "pizza mo two minutes coming". *Journal of Sociolinguistics* 18(2): 161–184. doi: 10.1111/josl.12079.

PENNYCOOK, Alastair & Emi Otsuji 2015. *Metrolingualism: Language in the City*. New York and London: Routledge.

PEUKERT, Hagen 2013. Measuring linguistic diversity in urban ecosystems. *Linguistic Superdiversity in Urban Areas: Research Approaches*, eds. Joana Duarte & Ingrid Gogolin. 75–93. Amsterdam: Benjamins.

PFEFFER, Jürgen, Katja Mayer & Fred Morstatter 2018. Tampering with Twitter's sample API. *EPJ Data Science* 7(50). doi: 10.1140/epjds/s13688-018-0178-0.

POORTHUIS, Ate & Matthew Zook 2017. Making big data small: strategies to expand urban and geographical research using social media. *Journal of Urban Technology* 24(4): 115–135. doi: 10.1080/10630732.2017.1335153.

PURSCHKE, Christoph & Dirk Hovy 2019. Lörres, Möppes, and the Swiss: (re) discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography* 7(2): 113–134.

RIJHWANI, Shruti, Royal Sequiera, Monojit Choudhury, Kalika Bali & Chandra Shekhar Maddila 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (ACL 2017). 1971–1982. Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/P17-1180.

SEARGEANT, Philip and Caroline Tagg (eds.) 2014. *The Language of Social Media*. Basingstoke: Palgrave.

SILM, Siiri, Jussi S. Jauhiainen, Janika Raun & Margus Tiru 2020. Temporary population mobilities between Estonia and Finland based on mobile phone data and the emergence of a cross-border region. *European Planning Studies*. doi: 10.1080/09654313.2020.1774514.

SJÖHOLM, Kaj 2004. Swedish, Finnish, English? Finland's Swedes in a changing world. *Journal of Curriculum Studies* 36(6): 637–644.

SLOAN, Luke 2017. Social science 'lite'? Deriving demographic proxies from Twitter. The SAGE Handbook of Social Media Research Methods, eds. Anabel Quan-Haase and Luke Sloan. 90–104. Los Angeles, CA: SAGE.

SZMRECSANYI, Benedikt 2012. *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.

TAAVITSAINEN, Irma & Päivi Pahta 2003. English in Finland: globalisation, language awareness and questions of identity. *English Today* 19(4): 3–15.

TAAVITSAINEN, Irma & Päivi Pahta 2008. From global language use to local meanings: English in Finnish public discourse. English Today 24(3): 25–38.

TASSE, Dan, Zichen Liu, Alex Sciuto & Jason I. Hong 2017. State of the geotags: motivations and recent changes. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017)*. 250–259.

TOIVONEN, Tuuli, Vuokko Heikinheimo, Christoph Fink, Anna Hausmann, Tuomo Hiippala, Olle Järv, Henrikki Tenkanen & Enrico Di Minin 2019. Social media data for conservation science: a methodological overview. *Biological Conservation* 233: 298–315. doi: 10.1016/j.biocon.2019.01.023.

TRUDGILL, Peter 1974. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 3(2): 215–246.

WIELING, Martijn & John Nerbonne 2015. Advances in dialectometry. *Annual Review of Linguistics* 1(1): 243–264.

WILLIAMS, Shirley A., Melissa M. Terras & Claire Warwick 2013. What do people study when they study Twitter? Classifying Twitter related academic papers. Journal of Documentation 69(3): 384–410.

ZAPPAVIGNA, Michele 2013. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Continuum.

ZOOK, Matthew, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson & Frank Pasquale 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology* 13(3). doi: 10.1371/journal.pcbi.1005399.

ZOOK, Matthew & Mark Graham 2007. Mapping DigiPlace: geocoded internet data and the representation of place. *Environment and Planning B* 34(3): 466–482.