Ilia Moshnikov and Eugenia Rykova

Collecting minority language data from Twitter (X): A case study of Karelian

Abstract: The visibility of an endangered language online plays a crucial role in language revitalisation. The internet offers a new domain for using minority languages, especially for speakers living outside the language communities. This article investigates Karelian language visibility on X, formerly known as Twitter, and describes the first corresponding data collection using language-related keywords and hashtags. In total, 2,625 entries written fully or partially in Livvi, South and Viena Karelian were scraped with Postman API. The visibility of Karelian on Twitter (X) has been increasing considerably in the past few years, with Livvi-Karelian being the most prominent dialect. Automatic language detection was tested on such data for Karelian for the first time, and allows the identification of Livvi-Karelian (or a mix of dialects that include Livvi-Karelian) with 99.7% sensitivity, and South Karelian and Viena Karelian as Livvi-Karelian with 90% and 73.8% sensitivity, respectively. The entries were also analysed thematically, and 10 major topics were identified. Since the data was collected using keywords and hashtags related to the Karelian language itself, most of the entries are related to the language and vocabulary in sense of translation or language learning. Language status and policy is another important topic identified in the data. Although language-related topics are the most popular, there are a substantial number of entries on eight further topics. Excluding citations from religious texts and media headlines, 751 Twitter (X) entries could be used for linguistic and sociological research. Further data collection considerations are also discussed.

Keywords: automatic language recognition, data scraping, Karelian, language policy, language revitalisation, language status, minority languages, X (Twitter)

Ilia Moshnikov*, Karelian Institute, University of Eastern Finland, Joensuu, Finland,
e-mail: ilia.moshnikov@uef.fi
Eugenia Rykova*, University of Eastern Finland, Joensuu, Finland; Technical University of Applied Sciences TH Wildau, Wildau, Germany; and Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany, e-mail: eugenryk@uef.fi

* The two authors have contributed equally to the present paper.

https://doi.org/10.1515/9783111434018-007

1 Introduction

The use of the internet and social media has developed rapidly in recent years. Access to the internet offers enormous opportunities not only for majority languages, and minority languages have also found their place online. The internet has become a new domain for the use of minority languages, which supports language revitalisation initiatives. In addition to traditional websites, minority languages have begun to be used in social media and private communication, providing a rich ground for research.

The present article aims to investigate the Karelian language use on X, formerly known as Twitter. X is a leading microblog platform which can serve for data collection on various research questions (Grillenberger 2021), including the use and visibility of minority languages. Languages other than English have been receiving more attention in the last decade. However, studies focusing on minority languages are still scarce (Cunliffe 2019; Valijärvi and Khan 2023). Thus, the Karelian language is not even separately discussed in the X (Twitter) linguistic repertoire of Finland (Hiippala et al. 2020). A lack of corresponding automatic language processing tools further hinders the process.

This article describes an approach to investigate Karelian language visibility on X (Twitter) and collect the corresponding data, and considerations for further data collection are discussed. The following questions are addressed:

- How to collect data in Karelian from X (Twitter)?
- How present has Karelian been on X (Twitter) throughout the years?
- What dialects of Karelian are the most visible on X (Twitter)?
- What are the main topics of tweets published in Karelian?

2 Research background

2.1 The Karelian language and its usage online

Karelian is a minority, critically endangered Finnic language mainly spoken in Russia and Finland (see Figure 1). Currently, the total number of Karelian speakers ranges from 5,000 to 10,000 speakers in Finland and about 15,000 in Russia (Sarhimaa 2017: 115; Federal State Statistics Service 2021).

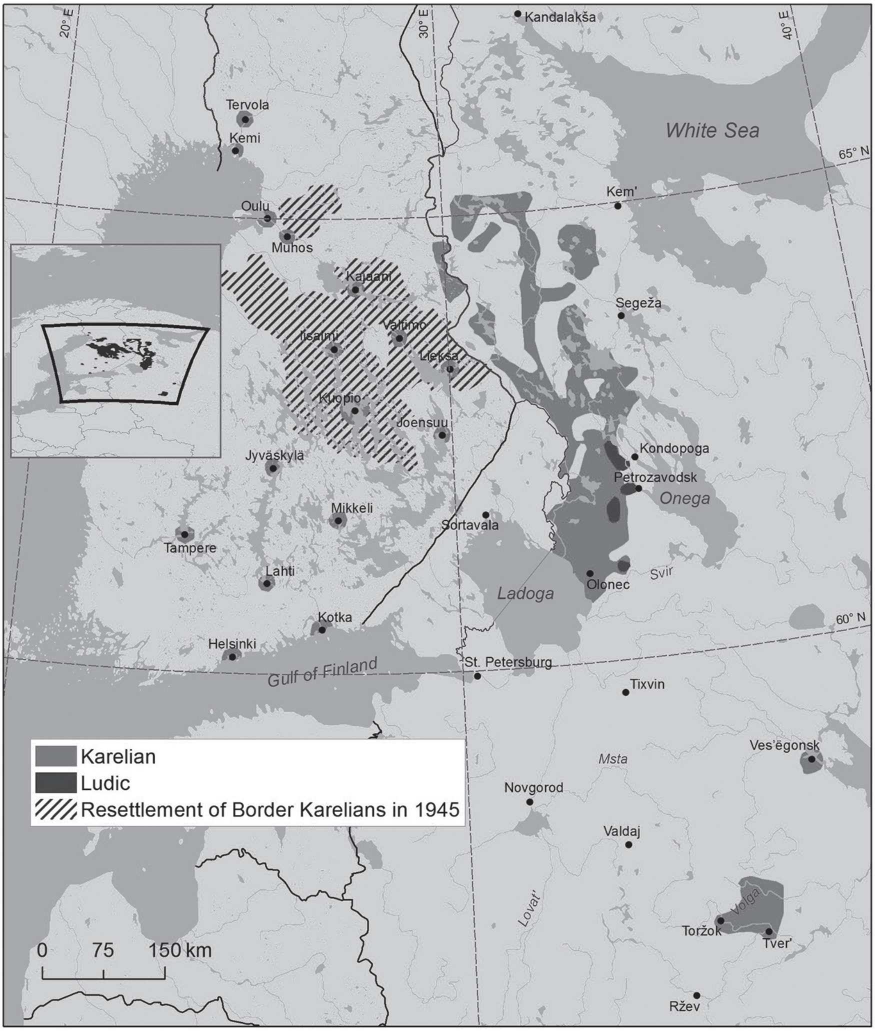
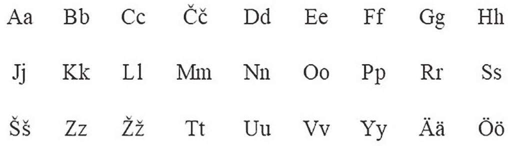


Figure 1: Karelian-speaking territories in the 2020s (Uralic Language Atlas 2024, see Roose et al. 2021).

Linguistically, the Karelian language is divided into two main dialects: Olonets (or Livvi) Karelian, and Proper Karelian. The latter consists of Viena (North) Karelian and South Karelian (Koivisto 2018). Since 1989, several written standards of Karelian based on Latin script were created. In Finland it is common to use the three written standards based on the main dialects mentioned above, while in the Republic of Karelia in Russia only two are used – Livvi and Viena. In 2007, a unified alphabet of the Karelian language was approved (Figure 2).



Aa	Bb	Cc	Čč	Dd	Ee	Ff	Gg	Hh
Jj	Kk	Ll	Mm	Nn	Oo	Pp	Rr	Ss
Šš	Zz	Žž	Tt	Uu	Vv	Yy	Ää	Öö

Figure 2: Karelian alphabet.

Finnish and Karelian share between 2,700 and 2,950 relatively old words, and Karelian Proper shares more old vocabulary with the Finnish dialects than Livvi-Karelian, but the difference is not highly significant (Häkkinen 1990: 216; Söderholm 2012). The study of proximity level based on the 800 most frequent Finnish words shows an 83.34% degree of recognition in Proper Karelian and 66.38% in Livvi-Karelian on average, and 88% and 81.5%, respectively, as a maximum (Söderholm 2012).

The intentional revitalisation of Karelian started at the beginning of the 1990s. The Petrozavodsk State University in Russian Karelia and the University of Eastern Finland in Joensuu are offering university studies of the Karelian language. Since 2021, the University of Eastern Finland has been responsible for the revitalisation of Karelian in Finland, and Karelian can be studied as a minor subject there. The Finnish broadcasting company YLE has been producing internet and radio news in Karelian since 2015. In Russian Karelia there is a Karelian newspaper *Oma Mua* published in Petrozavodsk, and other media produced by the Karelia broadcasting company. Fiction books have also been published in Karelian. Karelian is taught at some schools in Russian Karelia, and there has been a language nest functioning in Vedlozero (Vieljärvi) from 2017 to 2022. Language courses for adult learners are arranged in both countries as well. But despite these revitalisation efforts, the number of Karelian speakers is rapidly declining (for further reading, see Karjalainen et al. 2013; Sarhimaa 2016; Riionheimo and Giloeva 2022; Karjalan kielen elvyttäminen 2024).

The first signs of Karelian being used online date from the late 1990s. The first websites in Karelian were launched in the early 2000s. From the 2010s, the use of Karelian on social media started to grow significantly. Salonen (2017) studied the use of the language in internet services and software, as well as the visibility of the language on social and digital media. According to the research, Karelian speakers show a higher passive use of the language online, which could be explained by the higher average age of the speakers, but also by some psychological factors such as a perceived lack of writing skills in the language, and fear of being teased or provoked (Moshnikov 2022a; Soria 2022). Moshnikov (2016, 2022b) studied the use of Karelian as a language of websites from the virtual linguistic landscape and the theory of language ideologies, as well as the use of the language online focusing on the social media platforms, and the motivation, benefits and challenges of using Karelian online by speakers (Moshnikov 2022a). While Facebook is the most popular social media platform for consuming and creating content in Karelian, the use of Karelian on other social media platforms, including X (Twitter) and Instagram, has increased. As a new domain, the language use on social media reveals ongoing trends and changes in the language itself, and also reflects certain sociocultural processes. Importantly, it has been noted that language use in different domains and its responsiveness to new domains and media are keystones in language survival and vitality (Drude and Intangible Cultural Heritage Unit's Ad Hoc Expert Group 2003).

2.2 X (Twitter)

X (formerly Twitter[1]) is a social media micro-blogging platform, where users can publish short messages (tweets), of a maximum of 280 characters (140 characters until November 2017) and receive feedback from other users (Fausto and Aventurier 2016). The platform was established in 2006 and sold to billionaire Elon Musk in late 2022, after which Twitter underwent significant changes, up to and including an official name change from Twitter to X in July 2023. As social media interaction in general, X is a multilingual source of data that corresponds to the Big Data definition: it has volume, velocity, and variety (Kitchin 2013). Unlike Facebook, Twitter had long allowed researchers to collect data via Twitter API free of charge. In February 2023, Twitter announced the elimination of free API access, which would make further data collection more difficult (Willingham 2023).

At the same time, the changes sparked a wave of protest that led users to leave X (Twitter) and look for alternatives. Several new platforms have recently emerged, such as Threads and Bluesky (Hurst 2023; Mehta 2023; Silberling, Stringer, and Corrall 2024). But as one door closes, another one opens: following ongoing changes on X allows us to see how Karelian speakers adapt to the new conditions, and whether they stay on X or move to another platform. However, building a new community from scratch might take some time, especially for smaller communities using an endangered minority language.

[1] In this article we use the names Twitter and X equally, as well as the forms 'tweet' and 'to tweet' as already established concepts that denote a Twitter post or a verb which refers to writing a Twitter post.

2.3 Minority languages and the internet

The use and visibility of a minority and/or endangered language online shows that it can be used in modern environments and new media, which increases the value of the language (Cunliffe 2019). The relationship between minority languages and technology can be described from three dimensions: availability, usability and how technology is developed for minority languages (Soria 2022). *Availability* includes the range of resources available in a given language, such as media, services, interfaces, and applications. While majority languages tend to have a wide range of resources available, minority languages have far fewer options, ranging from a lack of advanced technologies such as speech recognition to the unavailability of a keyboard. *Usability* of resources simply means which resources are used by speakers of a minority language. Minority language speakers easily switch to their dominant language when using language-based digital technologies, either because the technology is inherently better or because the range of services available is much wider. The third dimension describes how well the development of technology meets the *needs* of the language community. Companies often offer ready-made solutions without taking into account the real needs, desires, and expectations of minority language speakers. The context of each community is unique, and while some may value access to e.g., Wikipedia, others may simply value the opportunity to use the language.

Furthermore, minority languages are often overshadowed by major languages in the use and development of language technology tools, which makes minority languages vulnerable in digital environments as well. A minority language can be endangered not only linguistically, but also digitally (Soria 2016), and many minority language speakers are not well connected, or due to their age, do not have the ability or willingness to use the internet or social media, which can lead to a lack of digital competence (Cunliffe 2019).

Multilingual internet users are often faced with the choice of which language to use online, and how the online community will react to the use of a particular language or another, and whether other users will understand the post or comment (Mentrau Iaith Cymru 2014: 18). A minority language speaker constantly makes choices, which of the languages of their linguistic repertoire to use (Cunliffe 2019). In addition, speakers also evaluate their personal language skills, including the mistakes they are making, especially in public situations. Karelian speakers have expressed similar concerns, especially about their language skills and language purism (Moshnikov 2022a). Research further shows that the language choices of other users and the language use of public figures influence the language choices of minority language speakers (Mentrau Iaith Cymru 2014: 3–4).

It is very common for the community of speakers of a minority language to be dispersed. Karelian is no exception. Even if a minority lives compactly in one area, this does not guarantee that the members of the community have close contact with each other. A non-territorial minority language is in a weaker position in this respect, as its speakers are even more dispersed throughout the country or countries.

At the grassroots level, local communities adapt social media platforms for their purposes and interests using specific hashtags. Speakers of a particular language create their own hashtag systems, which makes it easier to find tweets or other posts based on a concrete topic, place, or language (Cocq 2015; Outakoski, Cocq, and Steggo 2018; McMonagle et al. 2019). Communities of speakers also create networks to support and encourage language use and learning. The minority language can be the only connecting thing between users of social media platform, and in small communities, the role of an individual active user could be crucial.

The Indigenous Tweets portal (2024) provides the following data: the total number of tweets in certain minority languages and their distribution among the users. Apparently, the fewer tweets written in a particular language, the fewer unique users produce these tweets. Thus, from 19,936 tweets in Udmurt, 88.6% are written by top 15 users; and from 5,698,636 in Welsh, only 10% are written by top 15 users. The Māori language is in the middle: from 346,434 tweets, 35% are written by top 15 users. As noted by Keegan, Mato, and Ruru (2015), the statistics for the latter (and some other languages) is skewed by individual users: as for 2014, top three users, which might be the same person/organisation, were responsible for 68.4% of all the tweets in Māori. These tweets contained exclusively translated Bible passages.

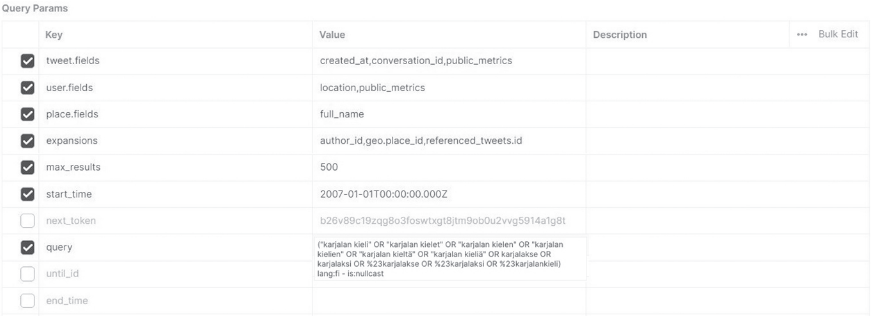
3 Research data and methods

3.1 Data scraping

Postman API software (2022) was selected to collect data from Twitter using Academic Research access, due to its convenient way of modifying the search parameters and stating the necessary information sections to be retrieved (see Rykova et al. 2023). Unlike other language-specific Twitter data collections (e.g., AbdelHamid 2022; Rykova et al. 2023), Karelian data cannot be collected via specifying the language in the query as Karelian is not among those languages whose identification is supported by Twitter API, nor is it built-in to other software libraries for

Twitter data collection. Post-hoc language identification of Karelian (cf. Ljubešić, Fišer, and Erjavec 2014; Nguyen, Trieschnigg, and Cornips 2015) is also difficult due to a scarcity of corresponding resources and dialect variability. Thus, the applicability of the HeLI-OTS 1.4 language identifier (Jauhiainen, Jauhiainen, and Lindén 2022) to Twitter entries is firstly examined in the current paper.

First, a full-archive search was performed with the help of keywords and hashtags (case and special characters can be ignored), which can be seen in Figure 3. It was assumed that the users would use these hashtags to highlight the use of the language as the speakers of other minority languages often do (Cocq 2015; McMonagle et al. 2019), and keep Karelian apart from the Finnish language. The hashtag #karjala ('Karelia' as a territory or 'Karelian' as a language) was not included in the search because it might be used in irrelevant posts about the Karjala beer produced by the Hartwall brewery or be part of discussions related to the loss of a significant part of Finnish Karelia after the Second World War. The forms of the nominative, genitive, and partitive have been chosen according to their frequency and usage in the closely related Finnish language (Hakulinen et al. 2004, §1228).



Query Params

	Key	Value	Description	••• Bulk Edit
☑	tweet.fields	created_at,conversation_id,public_metrics		
☑	user.fields	location,public_metrics		
☑	place.fields	full_name		
☑	expansions	author_id,geo.place_id,referenced_tweets.id		
☑	max_results	500		
☑	start_time	2007-01-01T00:00:00.000Z		
☐	next_token	b26v89c19zqg8o3foswtxgt8jtm9ob0u2vvg5914a1g8t		
☑	query	("karjalan kieli" OR "karjalan kielet" OR "karjalan kielen" OR "karjalan kielien" OR "karjalan kieltä" OR "karjalan kieliä" OR karjalakse OR karjalaksi OR %23karjalakse OR %23karjalaksi OR %23karjalankieli) lang:fi - is:nullcast		
☐	until_id			
☐	end_time			

Figure 3: Full-archive search query in Postman.

Since Karelian cannot be detected via Twitter API, but is recognised as Finnish, the search query contained the parameter of Finnish as the language of the entry. Additionally, tweets had to be organic, and not an advertisement. The data was retrieved starting from 2007.

After the initial search, additional searches for the parent tweets of the retrieved comments (multiple tweets query) and user information (user lookup query) were performed. Thus, the data included entries, information on public metrics, location (if given), the author, and their 'about' information. For the comments, entries that allowed tracing the conversation back (references) to the parent tweet were included.

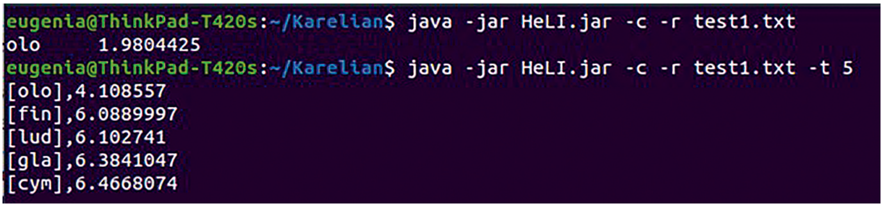
3.2 Data reduction

As of March 14, 2023, the collected data consisted of 15,428 entries. Removing retweets – sharing someone's tweets – reduced the number of entries to 8,463. Removing duplicates – tweets with the same content from the same or different users which were not marked as retweet and could be copying the same links or self-repetitions resulted in the final number of 8,224 multilingual entries, which were subject to manual language labelling.

3.3 Data labelling

The text of entries was subject to automatic language detection with the help of HeLI-OTS 1.4 (Jauhiainen, Jauhiainen, and Lindén 2022). This language identifier includes two dialects of Karelian: Livvi-Karelian (*olo*) and Ludic Karelian (*lud)*, although nowadays Ludic is generally considered as an independent language (Pahomov 2017: 286). HeLI-OTS is based on scoring the frequency of each word (or a shorter chunk of the word if necessary) in the analysed text against the language models, which is then measured with a negative logarithm of the relative frequency of the word in each language. The final decision corresponds to the average of the scores of the words (named relative frequency score below). The algorithm can output the detected language together with the confidence score or the indicated number of best suiting languages with respective relative frequency scores. The confidence score is the absolute difference between the relative frequency scores of the detected language and the second-best one. An example of the outputs can be seen in Figure 4: the language of the text is detected as Livvi-Karelian (*olo*) with the confidence score 1.98, while the best five suiting languages are Livvi-Karelian (*olo*) with the relative frequency score of 4.11; Finnish (*fin*) with the relative frequency score of 6.09; Ludic Karelian (*lud*) with the relative frequency score of 6.1; Scottish Gaelic (*gla*) with the relative frequency score of 6.38; and Welsh (*cym*) with the relative frequency score of 6.47.

For our dataset, we stored the detected language, information on the algorithm confidence score, the second probable language, and the relative frequency scores for both. As the data was not normally distributed, statistical comparisons were performed with a Mann-Whitney U test with the help of Python scipy.stats library (Virtanen et al. 2020). Splitting the multilingual entries into sentences was performed with sent_tokenize tokenizer from Python nltk library (Bird, Klein, and Loper 2009).



```
eugenia@ThinkPad-T420s:~/Karelian$ java -jar HeLI.jar -c -r test1.txt
olo	1.9804425
eugenia@ThinkPad-T420s:~/Karelian$ java -jar HeLI.jar -c -r test1.txt -t 5
[olo],4.108557
[fin],6.0889997
[lud],6.102741
[gla],6.3841047
[cym],6.4668074
```

Figure 4: Example outputs of HeLI-OTS 1.4.

The language was also labelled manually by the first author of the present study, who is a native Livvi-Karelian speaker. Manual labels included more specific information on Karelian dialects: in column ‘language’ generally marked as *olo* or *krl*, and the latter further specified in a separate column ‘dialect’ (South or Viena Karelian). If an entry contained several sentences written in up to five different languages, the languages were listed in order of appearance. Non-text entries, ones with languages mixed within a sentence, or separate sentences written in more than five languages were labelled as “other”.

Manual labelling also included assigning topics to entries written fully or partially in one of the Karelian dialects. The selection of topic was data-driven, and relevant groups were identified and refined during the labelling process.

4 Results

4.1 General results

There are 2,625 entries (2,201 tweets and 424 comments) written either fully or partially in one of the Karelian dialects in the final dataset, which constitutes 32% of the cleaned data. The distribution of these entries by year and dialect is shown in Figure 5. Year 2023 is not included in the graph because the data were not collected for the entire year due to the changes in Twitter policies. If an entry contains more than one dialect, it is counted for each of them. Thus, the total number of entries in the graph is higher than the actual number of entries in the database. In the whole dataset, there are 2,394 entries that include Livvi-Karelian, from which 2,038 are written in this dialect only; 231 entries that include South Karelian, from which 149 are written in this dialect only; and 112 entries that include Viena Karelian, from which 41 are written in this dialect only.

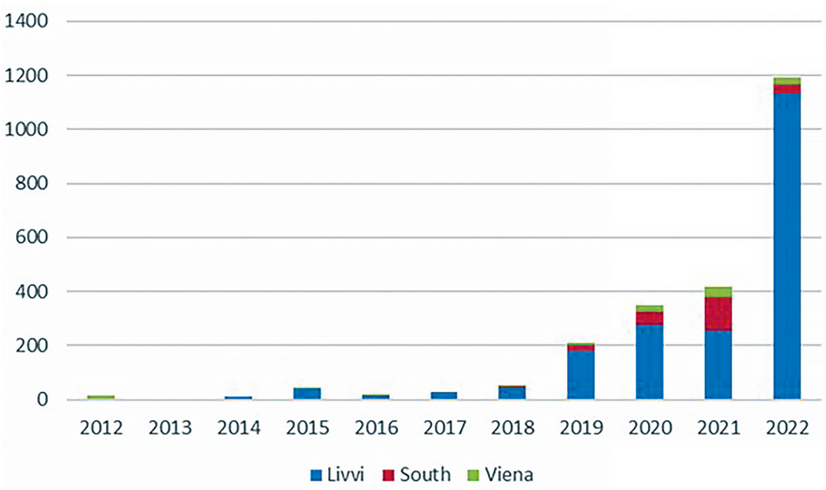


Figure 5: Entries in Karelian per year.

From the hashtags used in the search, the hashtag #karjalakse ('in Karelian', a translative case form for the word 'Karelian' in Livvi-Karelian) is present in 1,529 entries, #karjalankieli ('the Karelian language') in 155 entries, and #karjalaksi ('in Karelian', translative case form for the word 'Karelian' in Finnish, and South and Viena Karelian) in 3 entries only. However, with the help of #karjalankieli in the search, 1,180 entries were found with the hashtag #KarjalanKieliEläy (with orthographic variations meaning 'The Karelian language lives'). One more variation of #karjalankieli is present in 90 entries, and contains an underscore: #karjalan_kieli.

4.2 Language detection

The confusion matrix for automatic and manual language labelling can be seen in Figure 6. It must be noted that this matrix is not a classic confusion matrix as its true and predicted labels are asymmetric. Manual labelling allows more than one language to be included: for example, *krl* + *eng/fin* means South or Viena Karelian followed by either English or Finnish, *3+ (olo)* means three and more languages, including Livvi-Karelian. The automatic detection algorithm outputs only one language and has (erroneously) output languages that are not present in the original data. The language marked as Indonesian (*ind*) has such a label based on the location of the entries author. However, this variety of Malay is not included separately in the languages detected by HeLI-OTS 1.4, which makes the Malay macrolanguage (*msa*) the closest possible label for the absent *ind*. Sami languages are manually marked as a group (*sami*), without further distinction.

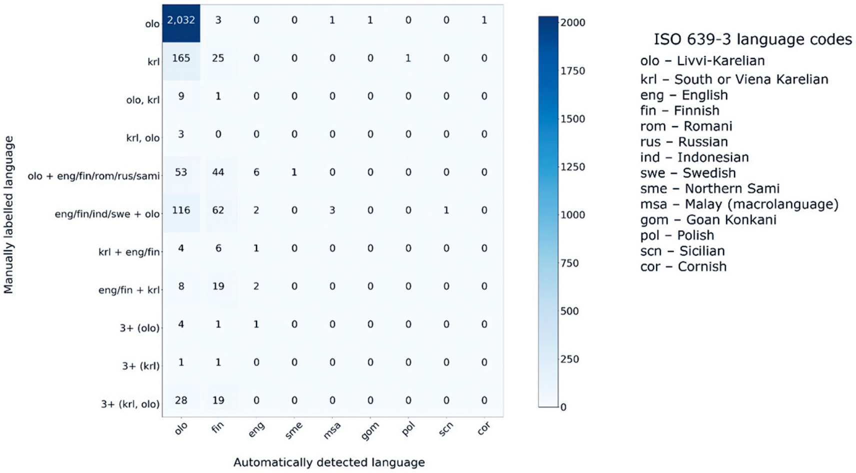


Figure 6: Confusion matrix of manually and automatically identified languages.

Livvi-Karelian is recognised as such in 99.7% of the cases, and as Finnish in 0.15% of the cases. The mean confidence score of the algorithm in cases when Livvi-Karelian is recognised as such, is 1.58±0.43 (standard deviation). The mean relative frequency score in such cases is 4.22±0.5. If Livvi-Karelian is recognised as Finnish, the mean confidence score is 0.7±0.91, and the mean relative frequency score is 5.06±0.9 (there are only 3 data points, which causes such a dispersity of data). When Karelian dialects are not recognised as Livvi-Karelian or Finnish, the mean confidence score is 0.31±0.24, and the mean relative frequency score is 5.15±0.78.

A mix of Karelian dialects is recognised as Livvi-Karelian with a mean confidence score of 1.31±0.52. The mean relative frequency score is 4.17±0.85. These results do not statistically significantly differ from the results for Livvi-Karelian alone according to the Mann-Whitney U rank test ($p > 0.05$). When a mix of dialects was recognised as Finnish, the confidence score is 0.25, and the relative frequency score is 5.31.

If an entry is written in South Karelian, its language is detected as Livvi-Karelian in 90% of the cases and as Finnish – in 9.3% of the cases. The mean confidence score of the algorithm in cases when South Karelian is recognised as Livvi-Karelian is 0.98±0.6, and the mean relative frequency score in such cases is 4.67±0.68. If

South Karelian is recognised as Finnish, the mean confidence score is 0.28±0.21, and the mean relative frequency score is 5.37±0.54. The differences in confidence scores and relative frequency scores between two recognition patterns are statistically significant according to the Mann-Whitney U rank test ($p < 0.01$).

From entries written in Viena Karelian, the detected language is Livvi-Karelian in 73.8% of the cases, and Finnish otherwise (26.2%). The mean confidence score of the algorithm in cases when Viena Karelian is recognised as Livvi-Karelian is 0.79±0.56. The mean relative frequency score in such cases is 4.5±0.67. If Viena Karelian is recognised as Finnish, the mean confidence score is 0.33±0.28, and the mean relative frequency score is 4.68±0.55. According to the Mann-Whitney U rank test, the difference in confidence scores between two recognition patterns is statistically significant ($p = 0.008$), but the difference in relative frequency is not ($p = 0.85$).

For multilingual entries (two or more languages): when the recognised language is one of the entry languages (*olo* is considered as language for *krl*), the mean confidence score is 0.54±0.44 and the mean relative frequency score is 4.87±0.58. From 383 multilingual entries, in 214 (56%), the language was detected as *olo*. From the remaining 169 entries, after automatic sentence split, the language was detected as *olo* for at least one sentence of the entry in 108 cases. It must be noted, however, that automatic sentence splitting does not always perform as ideally desired due to its internal flaws or the nature of data (especially where phrases in different languages do not constitute separate sentences).

In 73% of all cases, when any of the Karelian dialects or their mix is not recognised as Livvi-Karelian (for cases when the language is recognised as Finnish, this is 79%), the latter is still the second language in the language probabilities list. Incorporating thresholds based on discovered mean confidence and relative frequency scores should help to detect entries in Karelian even if their language is erroneously recognised. For example, taking the second option *olo* instead of the detected language, when the confidence score of the algorithm is below 0.3 or the relative frequency score for the detected language is higher than 4.8, this allows a retrieval of 17 entries – 59% for which the detected language is Finnish.

4.3 Users

In total, there were 161 usernames in the final dataset. Information on the 15 users with the greatest amounts of tweets is presented in Table 1. It must be noted that “Tweets in Karelian” means only those that were scraped with the proposed method. As we can see from the table, 89.4% of the collected entries (n=2,347) in Karelian are produced by the top 15 users. Also, it is more common to use one dialect of Karelian among individual users, while organisations or media are posting multidialectal content (e.g., Karelian Youth Organisation, Karelian Cultural Society).

Table 1: Top-15 users posting tweets in Karelian (data as of March 14, 2023).

Description of the user	N of entries in Karelian	Author location (if defined)	N of followers	following	N of tweets	Dialects used
Music teacher, religion activist	1,476	Oulu, Finland	666	1,184	14,622	olo
Professor of linguistics	220	Joensuu, Finland	1,452	1,170	7,487	south, viena, olo
Language activist, account not active anymore	158		722	74	9,809	olo, south, viena
Linguistic researcher, language activist	93	Joensuu, Finland	465	503	332	olo
Researcher, language activist	69	Tampere, Finland	9,044	1,709	32,990	olo
Musician, teacher, language activist	66	Jyväskylä, Finland	82	121	126	south
Karelian Youth Organisation	56	Suomi, Finland	1,167	104	870	olo, south, viena
Language activist, account not active anymore	50	Helsinki, Finland	218	971	1,233	olo, south, viena
Language activist	39		441	952	26,060	viena, south olo
Language activist, account not active anymore	29		888	656	7,769	south, olo
Karelian Cultural Society	24	Helsinki, Finland	249	134	221	viena, olo
Language activist	23		441	494	134,177	viena, south, olo
Language activist	16		344	196	1,880	south
Music band	16	Suomi, Finland	312	282	261	olo, viena, south = olo
Karelian News Portal	12		52	377	29	viena, olo

4.4 Topics

Ten topics identified during the manual labelling procedure are presented in Table 2, including the corresponding number of entries and an example. It must be noted that 98% of the largest topic Religion (n=1,483) comprises extracts from the Bible or other (Christian) religious texts, posted by the same user, starting in February 2021 (see Table 1). Only 28 tweets are posted by other users, and these are mainly related to the greetings on church holidays. The second largest group of tweets labelled Personal (n=327) represent personal opinions and experiences. This material is particularly interesting for studying how Karelian speakers themselves use Karelian online. Topics Vocabulary (n=313), Research (n=54), and Language learning (n=44) are respectively related to the vocabulary, research and learning of the Karelian language in a broad context. Users can use Vocabulary tweets to learn new words, Language learning – to get information about the courses available, and Research – to review or even participate in scientific studies. Tweets on these topics are published not only by educational institutions and organisations, but also by researchers and native speakers themselves.

Most of the entries in the topic Media (n=106) are links to news sources, accompanied by the title (and sometimes subtitle) of the corresponding news article. Usually, these tweets do not contain any personal information or opinions. The purpose is to share the latest news in the Karelian community. From a research perspective, the message of the posts or links is notable as well as the reactions (likes, re-tweets, quotes) and possible discussion based on these tweets. News headlines are also modifying the visibility of the Karelian language online.

There are 217 tweets related to language status and language policy. Despite the growing visibility of the Karelian language on the internet and in the media in general, there are still regular discussions about the status of the Karelian language, the status of Karelian as a language (pro dialect), language policy, and the revitalisation process in general. The mixing of the Karelian language with the Karelian dialects of Finnish, as well as the internal naming of Karelian dialects and their status, are also discussed in these tweets.

Culture (n=41) and Politics (n=12) have a significantly smaller number of tweets in the data. Tweets related to this topic include posts about Karelian music and literature, as well as events related to Karelian culture. Often, such tweets can be interpreted in a variety of ways, from personal to language policy. Political tweets are clearly related to politics, including political parties and elections. Usually, such tweets appear in the run-up to an election to attract voters. Tweets related to language status and policy are included in their own group.

The last group of Other tweets (n=29) includes some randomly written tweets related to different topics often overlapping with other topics. In the most cases, it is simply difficult to label them as belonging to just one topic.

Table 2: Topics identified in the collected data.

Topic	Description	N of entries	Example
Religion	Tweets related to religious holidays or the Bible.	1,455+28	*Hyviä äijänpäivän pruazniekkua! Kristoz voskres! Hristos nouzi kuollielois! #äijänpäivy #äijypäivy #karjalakse* 'Happy Easter holidays! Christ has resurrected! Christ has risen from the dead!'
Personal	Tweets identified as an opinion or experience.	326	*Tänäpiänä lähen otpuskah, ga loma vuottau dačalla! Hyviä heinäkuudu #karjalakse #tiedäjättiijetäh* 'I'm going on holiday today, but the iron scrap is waiting for me at the cottage! Have a good July'
Vocabulary	Tweets related to the learning of the language from the perspective of the vocabulary (e.g., translations and presenting variants from the different dialects of Karelian).	313	*Tänäpäi aijankohtaine sanaine karjalakse on huračču = vasenkätinen.* *Huraččuloin päiviä pietäh 13. elokuudu jo vuvves 1976. #sanainekarjalakse* 'A relevant word in Karelian today is 'huračču' – left-handed. Left Handers Day has been celebrated on 13 August since 1976.'
Language status and policy	Tweets related to the language status, policy, and revitalisation process in a broad understanding.	217	*Karjalan kieli on oma kieli, ei suomen kielen murreh.* 'Karelian is a proper language, not a dialect of Finnish.'
Media	Tweets of news or other mass-media sources.	106	*Yle Uudizet karjalakse: Päivännouzu-Suomen yliopisto tahtou jatkua karjalan kielen elvytändiä da kehitändiä* 'Yle News in Karelian: The University of Eastern Finland would like to continue its work on the revitalisation and development of the Karelian language.'
Research	Research related topics.	54	*Hyvä karjalan kielen maltai! Vastua kyzelyh karjalan kielen käyttöh näh! #karjalakse #karjalankieli* 'Dear Karelian speaker! Answer the questionnaire on the use of the Karelian language!'

Topic	Description	N of entries	Example
Language learning	Education related topics including university studies and other language courses.	44	*Zavodimma egläin Karjalan Liiton karjalan kursan. Mie opastan varzinkarjalua / suvikarjalua. Opastujat ollah kaikin puolin Suomie, 20 rištikanzuo. Keski-igä ozapuilleh 27,5 vuotta.* ‘Yesterday, together with the Karelian Union, we started a Karelian language course. I am teaching Karelian Proper/South Karelian. The students are from all over Finland, 20 people. Average age of the participants is 27.5 years old.’
Culture	Culture related topics.	41	*Elbyygö karjalan kieli? Tulgua terveh Lieksan 11. kul’ttuuraseminuarah piätinččänä 4. muarienkuuda. Väl’l’ä piäzy!* ‘Will the Karelian language be revived? Welcome to the 11th Lieksa Cultural Seminar on Friday, 4 March. Free entry!’
Politics	Tweets related to elections or political parties.	12	*Minule mugon! Iänestä minuu Jovensuun kunduvalličuksis 2021!* ‘For me it’s like this. Vote for me in the Joensuu Municipal Election 2021!’
Other	Other topics not related to other groups mentioned here.	29	*Hyviä puolistusvoimien flagupruazniekkua! #karjalan #kieli #puolistusvoimat #flagu #pruazniekku #Suomi* ‘Happy Flag Day of the Finnish Defence Forces!’

5 Discussion

5.1 Karelian presence on Twitter (X)

The method of using language-related keywords and hashtags has proven to be successful to collect X (Twitter) entries in Karelian. From the data available from Twitter until March 2023, 2,625 original entries in three Karelian dialects were collected. The predominant dialect is Livvi-Karelian, which is in line with other research (Moshnikov 2022a). There are no official statistics about the number of speakers of each Karelian dialect, but there might be slightly more speakers of Livvi Karelian than speakers of other dialects. Some other factors, such a Wikipedia and Yle news in Livvi Karelian also increase a visibility of Livvi Karelian dialect on X (Twitter).

Despite the use of such specific hashtags, language-related topics were not the only ones identified in the data. The research data shows that certain events and holidays increase the activity of users. For example, the launch of Yle News in Karelian (Yle Uudizet karjalakse) in 2015 or the establishment of the Association of Young Karelians in Finland (Karjalazet Nuoret Suomes, KNŠ) in November 2019 clearly increased activity in Karelian on Twitter. Some spikes in the use of Karelian on Twitter can also be observed on specific dates, for example, Karelian Language Day, 27th of November (cf. Keegan, Mato, and Ruru 2015). The use of Karelian on Twitter (X) seems to have grown in the last five years. However, this tendency can be affected by "technical" reasons: some user accounts get deleted or become private, so that their entries are no longer available for data scraping.

5.2 Automatic language detection

Automatic detection of language with the help of HeLI-OTS 1.4 (Jauhiainen, Jauhiainen, and Lindén 2022) allows the identification of Livvi-Karelian (or a mix of dialects that include Livvi-Karelian) with 99.7% sensitivity. Two other Karelian dialects, namely South Karelian and Viena Karelian, are identified as Livvi-Karelian with 90% and 73.8% sensitivity, respectively, while the mean confidence of the algorithm becomes lower, and the score based on a negative logarithm of the relative word frequency becomes higher. Each of the three dialects can be confused with Finnish: Livvi-Karelian the least (0.15%), followed by South Karelian (9.3%), and Viena Karelian the most (26.2%). South Karelian is recognised as Finnish with a lower mean confidence score and higher (negative) relative frequency score than Viena Karelian. In fact, for the latter, the (negative) relative frequency score is not different between cases when the text language is recognised as Livvi-Karelian or Finnish. Such results are in line with the knowledge on Karelian dialects' lexicon proximity to Finnish (Söderholm 2012), and suggest that the automatic language detection tool could be of use for language and dialects proximity research. The information on confidence and (negative) relative frequency scores could also provide information on the possible dialect.

Furthermore, when Karelian dialects are erroneously recognised as Finnish, Livvi-Karelian is the second language in the language probabilities list in most of the cases, which with a lower percentage also holds true for other erroneous outputs. In this case, incorporating thresholds or more elaborated statistical models based on confidence and (negative) relative frequency scores seems to be promising for not missing relevant texts in Karelian. Corresponding modelling and testing with data in Finnish itself are left for further research.

Entries with separate phrases or sentences in different languages which include any of the Karelian dialects are usually identified with one of the languages present in the entry. Livvi-Karelian is detected as the language of 56% of such multilingual entries. When automatic sentence splitting is used, the percentage of entries for which Karelian is detected as the language for at least one sentence (or the whole entry) rises to 84%, despite the imperfections of the splitting algorithm. Since the Karelian alphabet contains specific characters (see Figure 2) and the HeLI-OTS 1.4 scoring is based on words, it might be useful to include automatic splitting into words and consider the entry to be at least partially written in Karelian if it contains a certain number of words identified as belonging to Livvi-Karelian.

Summarising the above points, the proposed language detection algorithm can be used for scraping data in Karelian from social media. For better recognition results, in particular avoiding false negatives, automatic splitting into sentences or words and certain mechanisms based on confidence scores, (negative) relative frequency scores, and a language probabilities list should be applied.

5.3 Twitter (X) users posting in Karelian

Certain authors (both individuals and organisations) who regularly write in Karelian on Twitter have been identified. The top 15 users published 89.4% of tweets in Karelian, which is similar to the rate for Udmurt language (Indigenous Tweets portal 2024) and supports the idea of higher concentration of tweets per user for less represented minority languages on Twitter (X). Individuals are posting mostly by using one of the dialects and written standards of Karelian, but organisations are usually posting multidialectal entries. Further data collections could focus on these particular authors and their interactions with other Twitter users (cf. Ljubešić, Fišer, and Erjavec 2014; Nguyen, Trieschnigg, and Cornips 2015).

5.4 Topics of Karelian Twitter (X)

During the manual labelling of the entries, 10 topics were identified in the data. However, the topic of Religion, comprising the largest number of entries, mainly consists of direct citations of (Christian) religious texts, comparable to the case of Māori (Keegan, Mato, and Ruru 2015). While it is relevant to the general visibility of Karelian online, these collected texts cannot be considered as representing personal voices on social media. The same holds true for the majority of entries in the topic Media, because they only copy the titles and subtitles of the news articles and provide corresponding links. The topic Vocabulary predominantly contains word or phrase lists translated into one or more of the Karelian dialects. While such posts are also relevant for visibility and could be used as a language learning resource, their applicability to other research fields is questionable. That reduces our dataset to 751 Twitter entries written fully or partially in one or more Karelian dialect, that could be used for deeper linguistic and sociological analysis. The data can be analysed in the context of language or dialect contacts, lexical and morphological variation, and from the perspective of translation studies and discourse analysis. Tweets and discussions related to the status of the Karelian language are interesting from the perspective of language revitalisation and policy. The modern use of Karelian online has also an important symbolic meaning for the Karelian-speaking community. The collected corpus becomes even more important in the current context of changes in Twitter API access.

6 Conclusion

To the best of authors' knowledge, this paper describes the first corpus of X (Twitter) entries in Karelian. Using language-related keywords and hashtags, 2,625 original entries corresponding to 10 different topics were collected. 29% of the material can be seen as useful for further linguistic and sociological analysis.

The recent changes on X (Twitter) made the new data collection challenging. Regarding the accessibility of the resources mentioned above (Soria 2022), it should be noted that access to Twitter in Russia has been restricted since 1 March 2022. Facebook and Instagram are also blocked. The use of X itself is constantly changing, and the future will show whether the Karelian language will retain its position on X, or whether speakers will move to another platform. Nevertheless, minority languages, including Karelian, have found their place in the online space. Accordingly, it is important to support the use of endangered languages online on a personal and institutional level, and this will support the vitality and revitalisation of the language.

References

AbdelHamid, Medyan, Assef Jafar & Yasser Rahal. 2022. Levantine hate speech detection in Twitter. *Social Network Analysis and Mining* 12. 121. https://doi.org/10.1007/s13278-022-00950-4.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit.* O'Reilly Media, Inc. https://www.nltk.org/book/ (last accessed 11 March 2024).

Cocq, Coppélie. 2015. Indigenous voices on the web: Folksonomies and endangered languages. *Journal of American Folklore* 128 (509). 273–285. https://www.jstor.org/stable/10.5406/jamerfolk.128.509.0273 (last accessed 11 March 2024).

Cunliffe, Daniel. 2019. Minority languages and social media. In Gabrielle Hogan-Brun & Bernadette O'Rourke (eds.), *The Palgrave handbook of minority languages and communities,* 451–480. London: Palgrave Macmillan.

Drude, Sebastian and Intangible Cultural Heritage Unit's Ad Hoc Expert Group. 2003. *Language vitality and endangerment*. https://unesdoc.unesco.org/ark:/48223/pf0000183699 (last accessed 11 March 2024).

Fausto, Sibele & Pascal Aventurier. 2016. Scientific literature on Twitter as a subject research: Findings based on bibliometric analysis. In Clement Levallois, Morgane Marchand, Tiago Mata, & André Panisson (eds.), *Twitter for Research Handbook* 2015–2016, 1–14. Lyon: EMLYON Press.

Federal State Statistics Service. 2021. *Vserossijskaja perepis' naselenija 2020 [Russian Census 2020].* https://rosstat.gov.ru/vpn/2020 (last accessed 11 March 2024).

Grillenberger, Andreas. 2021. Twitterdaten analysieren mithilfe der blockbasierten Programmiersprache SNAP! [Analyse Twitter data using the block-based programming language SNAP!]. *LOG IN* 41. 54–60. https://dl.gi.de/items/01f69d2c-a8a4-4934-b1ee-40ef56870f1a (last accessed 11 March 2024).

Hakulinen, Auli, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen & Irja Alho (eds.), 2004. *Iso suomen kielioppi [Descriptive grammar of Finnish]*. Online version. Helsinki: Finnish Literature Society. http://scripta.kotus.fi/visk (last accessed 11 March 2024).

Hiippala, Tuomo, Tuomas Väisänen, Tuuli Toivonen & Olle Järv. 2020. Mapping the languages of Twitter in Finland: Richness and diversity in space and time. *Neuphilologische Mitteilungen* 121 (1), 12–44. https://doi.org/10.51814/nm.99996.

Hurst, Luke. 2023. With Twitter gone and users unsure about X, is Bluesky the future? We try it out. *Euronews*. https://www.euronews.com/next/2023/08/15/with-twitter-gone-and-users-unsure-about-x-is-bluesky-the-future-we-try-it-out (last accessed 11 March 2024).

Häkkinen, Kaisa. 1990. *Mistä sanat tulevat. Suomalaista etymologiaa* [Where do words come from: Finnish etymology]. Tietolipas 117. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jauhiainen, Tommi, Heidi Jauhiainen & Krister Lindén. 2022. HeLI-OTS, Off-the-shelf language identifier for text. *Proceedings of the Thirteenth Language Resources and Evaluation Conference.* 3912–3922. Marseille, France. European Language Resources Association. https://aclanthology.org/2022.lrec-1.416/ (last accessed 11 March 2024).

Karjalainen, Heini, Ulriikka Puura, Riho Grünthal & Svetlana Kovaleva. 2013. Karelian in Russia: ELDIA casespecific report. *Studies in European Language Diversity* 26. Mainz: Research consortium ELDIA https://phaidra.univie.ac.at/detail/o:314612 (last accessed 11 March 2024).

Karjalan kielen elvyttäminen. 2024. *The revitalisation of the Karelian language.* University of Eastern Finland. https://blogs.uef.fi/karjalanelvytys/ (last accessed 11 March 2024).

Keegan, Te Taka, Paora Mato & Stacey Ruru. 2015. Using Twitter in an indigenous language: An analysis of te reo Māori tweets. *AlterNative: An International Journal of Indigenous Peoples* 11:1. 59–75. https://doi.org/10.1177/117718011501100105.

Kitchin, Rob. 2013. Big data and human geography: Opportunities, challenges, and risks. *Dialogues in Human Geography* 3 (3). 262–267. https://doi.org/10.1177/2043820613513388.

Koivisto, Vesa. 2018. Border Karelian dialects – a diffuse variety of Karelian. In Marjatta Palander, Helka Riionheimo & Vesa Koivisto (eds.), *On the border of language and dialect,* 56–84. Studia Fennica Linguistica 21. Helsinki: Finnish Literature Society.

Ljubešić, Nikola, Darja Fišer & Tomaž Erjavec. 2014. TweetCaT: A tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2279–2283. Reykjavik, Iceland. European Language Resources Association (ELRA). https://aclanthology.org/L14-1642/ (last accessed 11 March 2024).

McMonagle, Sarah, Daniel Cunliffe, Lysbeth Jongbloed-Faber & Paul Jarvis. 2019. What can hashtags tell us about minority languages on Twitter? A comparison of #cymraeg, #frysk, and #gaeilge. *Journal of Multilingual and Multicultural Development* 40 (1). 32–49. https://doi.org/10.1080/01434632.2018.1465429.

Mehta, Ivan. 2023. What is Instagram's Threads app? All your questions answered. *TechCrunch online magazine*. https://tcrn.ch/44d3hB3 (last accessed 11 March 2024).

Mentrau Iaith Cymru. 2014. *The Welsh language and social networks*. Llanrwst: Mentrau Iaith Cymru.

Moshnikov, Ilia. 2016. Karjalankieliset verkkosivut virtuaalisena kielimaisemana [Developing websites in the Karelian language as part of virtual linguistic landscape]. *Lähivõrdlusi. Lähivertailuja* 26. 282–310. http://dx.doi.org/10.5128/LV26.09.

Moshnikov, Ilia. 2022a. The use of the Karelian language online: Current trends and challenges. Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. *Journal of Estonian and Finno-Ugric Linguistics* 13 (2). 275–305. https://doi.org/10.12697/jeful.2022.13.2.09.

Moshnikov, Ilia. 2022b. The use of the Karelian language online: websites in Karelian. In Tanja Seppälä, Sirkku Lesonen, Päivi Iikkanen & Sigurd D'hondt (eds.), *Kieli, muutos, yhteiskunta - Language, change, society*. AFinLA Yearbook 2022, 192–216. https://doi.org/10.30661/afinlavk.113920.

Nguyen, Dong, Dolf Trieschnigg & Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. *Proceedings of the ninth international AAAI conference on web and social media*, 9 (1). 666–669. https://ojs.aaai.org/index.php/ICWSM/article/view/14648/14497 (last accessed 11 March 2024).

Outakoski, Hanna, Coppélie Cocq & Peter Steggo. 2018. Strengthening indigenous languages in the digital age: Social media–supported learning in Sápmi. *Media International Australia* 169 (1). 21–31. https://doi.org/10.1177/1329878X18803700.

Pahomov, Miikul. 2017. *Lyydiläiskysymys: Kansa vai heimo, kieli vai murre?* [The Ludian Question: Nation or tribe, language or dialect?]. Helsinki: University of Helsinki & Lyydiläinen Seura.

Postman. 2023. *Postman API Tool*. https://www.postman.com/ (last accessed 11 March 2024).

Riionheimo, Helka & Natalia Giloeva. 2022. Karjalankielinen yliopisto-opetus – vastavirtaan soutamista? [University education in Karelian - rowing against the stream?]. *Kieli, koulutus ja yhteiskunta* 13/5. https://www.kieliverkosto.fi/fi/journals/kieli-koulutus-ja-yhteiskunta-lokakuu-2022/karjalankielinen-yliopisto-opetus-vastavirtaan-soutamista (last accessed 11 March 2024).

Roose, Meeli, Tua Nylén, Harri Tolvanen & Outi Vesakoski. 2021. User-centered design of multidisciplinary spatial data platforms for human-history research. *SPRS International Journal of Geo -Information* 10/7, 467. https://doi.org/10.3390/ijgi10070467.

Rykova, Eugenia, Christine Stieben, Olga Dostovalova & Horst Wieker. 2023. Connected driving in German-speaking social media. *Social Sciences* 12 (1): 46. https://doi.org/10.3390/socsci12010046.

Salonen, Tuomo. 2017. Karelian – a digital language? In Claudia Soria, Irene Russo & Valeria Quochi (eds.), *Reports on digital language diversity in Europe*. https://www.dldp.eu/sites/default/files/documents/DLDP_Karelian-Report.pdf (last accessed 11 March 2024).

Sarhimaa, Anneli. 2016. Karelian in Finland. ELDIA case-specific report. *Studies in European Language Diversity* 27. Mainz: Research consortium ELDIA. https://fedora.phaidra.univie.ac.at/fedora/get/o:471733/bdef:Content/get (last accessed 11 March 2024).

Sarhimaa, Anneli. 2017. *Vaietut ja vaiennetut. Karjalankieliset karjalaiset Suomessa* [Silent and being forced to be silent: Karelian-speaking Karelians in Finland]. *Tietolipas* 256. Helsinki: Finnish Literature Society.

Silberling, Amanda, Alyssa Stringer & Cody Corrall. 2024. What is Bluesky? Everything to know about the app trying to replace Twitter. *TechCrunch online magazine*. https://tcrn.ch/3HDTvi7 (last accessed 11 March 2024).

Soria, Claudia. 2016. What is digital language diversity and why should we care? In Josep Cru (ed.), *Digital media and language revitalisation. Els mitjans digitals i la revitalitzacio lingüística. Linguapax Review* 13–28. https://www.linguapax.org/wp-content/uploads/2015/03/LinguapaxReview2016web.pdf (last accessed 11 March 2024).

Soria, Claudia. 2022. *Decolonizing Minority Language Technology.* https://internetlanguages.org/en/stories/decolonizing-minority-language/ (last accessed 11 March 2024).

The Indigenous Tweets portal. 2024. http://indigenoustweets.com/ (last accessed 18 July 2024).

Twitter Developers. 2023. *Announcing new access tiers for the Twitter API. Twitter.* https://twittercommunity.com/t/announcing-new-access-tiers-for-the-twitter-api/188728 (last accessed 11 March 2024).

Valijärvi, Riitta-Liisa & Lily Kahn. 2023. The role of new media in minority- and endangered-language communities. In Eda Derhemi and Christopher Moseley (eds.), *Endangered languages in the 21st century,* 139–157. Abingdon, Oxfordshire, UK: Routledge. https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003260288-12/role-new-media-minority-endangered-language-communities-riitta-liisa-valij%C3%A4rvi-lily-kahn (last accessed 11 March 2024).

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt & SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17 (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Willingham, AJ. 2023. *Why Twitter users are upset about the platform's latest change.* CNN. https://edition.cnn.com/2023/02/03/tech/twitter-api-what-is-pricing-change-cec/index.html (last accessed 11 March 2024).