Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-88

Language Use in a Multilingual Tweet Corpus

Dmitrijs MILAJEVS [1]
Guest Researcher at National Institute of Standards and Technology, Maryland, USA

Abstract. A trilingual Latvian-Russian-English corpus of tweets is presented with an analysis of users, language and topics. The corpus consists of 1.4 million tweets that cover a period from April 2017 to July 2018. The language analysis reveals that the majority of users mostly use one language. Across topics, there is more Latvian content than in the whole collection. Among many potential use cases, the corpus can be used, for example, to study the public engagement of major Latvian media outlets and public figures, or the factors that determine language choice and content of a tweet.

Keywords. Corpus Linguistics, Latvian, Russian, English

1. Introduction

This paper presents a multilingual, location-anchored corpus of tweets from Latvia. The main challenge of collecting a socially representative corpus from this country is that several languages are used there: Latvian, the main communication language and the only official language, Russian, the language of the largest minority, and English.

Building a monolingual Latvian collection could be done by harvesting tweets that contain indicative Latvian words, which are not present in other languages, similarly to how it is done for Dutch [1]. However, such an approach is not suitable for tweets in Russian and English, as these languages are widely used outside of Latvia. For the same reason, a TREC-like collection building approach [2] of filtering the publicly available stream of tweets by language would not work. A tweet collection based on a curated list of users [3, 4] is effective, but extra care must be taken in building the list of users. An alternative approach would be to retrieve only geo-located tweets [5, 6]. Such a collection would neither be biased linguistically because it is not based on a list of keywords, nor it would be biased thematically because it is not based on a list of users. The downside is that a large number of tweets are not geo-located, which makes retrieval incomplete.

To keep a balance between objectivity and completeness, this work applies a hybrid approach by combining a geo-location based collection procedure with following a curated list of users, which is based on the accounts of Latvian media outlets, politicians, government institutions and public figures.

Our base assumption is that geo-located tweets are a representative and unbiased sample of tweets from Latvia. Thus, an objective collection should exhibit similar prop-

[1]Corresponding Author: Dmitrijs Milajevs E-mail: dmitrijs.milajevs@nist.gov.

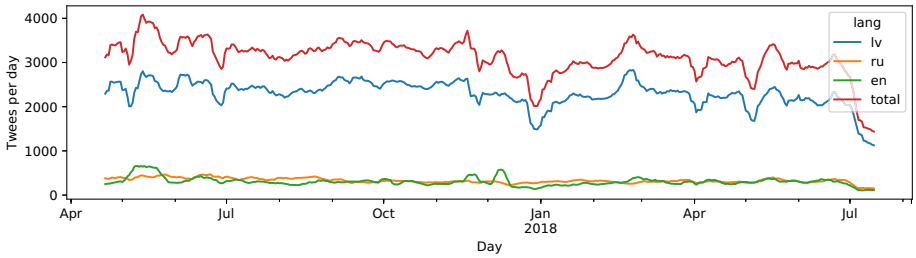

Figure 1. Tweet volume by day per language averaged over a rolling window of 7 days.

erties to the geo-located one. The main property we study in this work is the language proportion. We hypothesize, that an objectively extended collection should have similar language distribution as the initial collection.

To study the collection in more detail, we manually searched the collection for 16 topics of interest.

The analysis shows that—despite our attempt—the extended collection is biased toward content in Latvian: proportionally there are more tweets in Latvian in the extended collection than in the geo-located one. We observe this phenomena not only at the "global level" (the whole collection), but also at the "local level" (across the topics). In all topics, Latvian is more dominant than expected.

Building an unbiased collection is difficult. In our case the main reason of the introduced bias, might be that Russian and English content is less news-oriented and much more informal. In other words, while it is common to discuss current news in Latvian by engaging with the media, tweets in Russian and English tend to be personal and friend-oriented. Future work should verify whether this is actually the case.

2. Data collection

Over the period from 15 April 2017 to 16 July 2018 the initial set of 1 959 695 tweets was collected from the `POST status/filter` endpoint of the Twitter Streaming API. Following [6], the `locations` parameter was set to the bounding box of Riga, the capital of Latvia. It is small enough to fit into Twitter API restrictions and covers about 40% of the population of Latvia.[2]

In addition to the location, 420 accounts were followed.[3] The accounts are mostly Latvian (that is, coming from Latvia, but not necessary produce content in Latvian) news outlets, politicians, businesses, artists and sport clubs. We avoided following personal accounts to respect their privacy. We also avoided bootstrapping the list, as it might have introduced accounts that originate outside of Latvia, for example, `@BBC`. The whole list of screen names together with user IDs is available in the supplement file lv.cfg.

[2]The coordinates of the bounding box are `23.9325829, 56.8570671, 24.3247299, 57.0859184`.

[3]In this paper, by "following users", we mean that their user IDs where passed to the Twitter API to collect tweets that they created, retweets of those tweets, or the tweets that are replies by other users to their tweets. Refer to the Twitter documentation for a complete description. Keep in mind that this is different from the case when a user follows another user.

Table 1. Global statistics. "By followed" is the percentage of tweets that are created by a followed account, this excludes retweets and replies by not followed accounts.

Client	Tweets			Latvian		Russian		English		Other	
	Number	Share	By followed	Number	Share	Number	Share	Number	Share	Number	Share
Web Client	484 547	34.2%	52.7%	400 533	82.7%	15 403	3.2%	40 316	8.3%	28 295	5.8%
Android	231 631	16.4%	8.5%	157 577	68.0%	22 911	9.9%	33 818	14.6%	17 325	7.5%
iPhone	212 021	15.0%	14.6%	127 317	60.0%	34 537	16.3%	32 445	15.3%	17 722	8.4%
TweetDeck	108 826	7.7%	92.0%	106 660	98.0%	76	0.1%	1 532	1.4%	558	0.5%
TVNET	61 116	4.3%	96.6%	27 634	45.2%	32 750	53.6%	26	<0.1%	706	1.2%
dlvr.it	47 781	3.4%	98.4%	47 209	98.8%	145	0.3%	135	0.3%	292	0.6%
Facebook	38 152	2.7%	95.1%	14 341	37.6%	22 013	57.7%	462	1.2%	1 336	3.5%
Foursquare	31 493	2.2%	<0.1%	24 853	78.9%	221	0.7%	1 902	6.0%	4 517	14.3%
Instagram	25 774	1.8%	1.8%	9 242	35.9%	2 504	9.7%	8 619	33.4%	5 409	21.0%
SKATIES	24 184	1.7%	97.9%	24 166	99.9%	-	-	-	-	18	0.1%

To comply with Twitter's terms of service, on 16 July 2018, the raw tweet data was re-downloaded to get rid of deleted tweets. The tweets that originated a retweet were added to the collection. Also, we noticed a large number of tweets that came from Sweden (probably because of the imprecise `locations` parameter value), so the tweets with the location country code `SE` were omitted. This resulted in 1 415 984 tweets that formed the final collection presented here.[4]

3. Twitter users

On average, about 3 098 tweets were collected per day, which is more than 1 500 by the geo-location based technique in [6]. Most of the tweets came from the official Twitter clients for Web, Android or iPhone. Together these clients contributed more than 60% of all collected tweets, refer to Table 1 for more details.

The top 5 of most active uses of the Twitter Web Client consists only of followed Latvian media accounts: `@DienaLV` (a newspaper), `@LA_lv` (another newspaper), `@TV3_Play` (a TV channel), `@JaunsLV` (a media portal), and `@dblv` (a newspaper).

The proportion of tweets coming from followed accounts is lower for Android (8.5%) than for iPhone (14.6%). All top 5 Android users are personal accounts, while for iPhone there are 2 personal accounts in the top 5. `@Lattelecom` (a telecommunication company) is the most active iPhone user. `@LRZinas` (a news account of Latvian Radio) is the second most active iPhone user. `@TV3zinas` (a news account of a TV channel) is the fifth most active iPhone user. All three accounts tweet exclusively in Latvian and were followed during corpus collection.

For the TweetDeck, TVNET, dlvr.it and SKATIES client applications the top 5 users are followed media accounts that write dominantly in Latvian. The exceptions are a business account `@Kompresori` that was not followed and tweets in three languages: Latvian, Russian and English, `@RusApollo` and `@TVNET_rus` who tweet identical content exclusively in Russian, and `@SejasLV`, an exclusively Latvian account writing about celebrities, which was not followed.

TweetDeck is used by `@DelfiLV` (a major media portal), `@LV_Portals`, (the official government gazette), `@lsmlv` (a publicly funded radio and television organization,

[4]The supplement files are available at Zenodo: https://zenodo.org/record/1317574. Collected tweet IDs are available as collected_tweets.csv and the final collection is available as rehydrated_tweets.csv. The supplement files are licensed under a Creative Commons CCZero 1.0 License/Waiver.

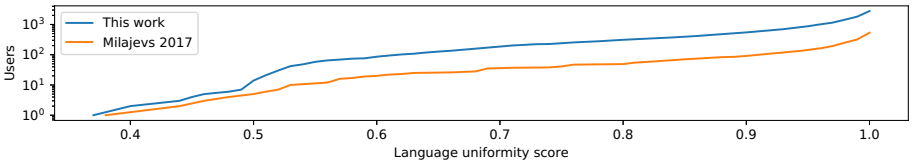

Figure 2. Accumulative number of users by the uniformity score. Note the logarithmic Y-scale. Milajevs 2017 is [6].

LSM), `@ltvzinas` (the news account of Latvian Television) `@RietumuRadio` (a radio station).

TVNET is a media company that controls several media portals. The most active accounts that use their application are `@TVnet_portals`, `@RusApollo`, `@TVnet_rus`, `@SportsTVNET`, and `@SejasLV`. As mentioned above, `@RusApollo` and `@TVnet_rus` post identical messages.

The tweet delivery service `dlvr.it` is used by the newspaper `@nralv` and their sport portal `@Sportacentrs`, the photo account of the national news agency (LETA) `@letafoto`, `@Kompresori` and the sport account of Latvian Television `@ltvsports`.

SKATIES delivers video content from several sources. The most active accounts are `@skatieslatvija`, `@lnt_lv`, `@tv3lv`, `@BezTabuTV3` and `@TV3Zinas`.

Some tweets are crossposts from other social media networks. Facebook content mostly comes from the followed media accounts. The top 5 most active users are: `@mixnews_lv` (the Russian edition of the Mixnews.lv media portal), `@Otkrito` (the Russian edition of `@JaunsLV`), `@labdienlv`, `@ekonimikalv` and `@nozare` (all three belong to the same media company and tweet in Latvian).

4. Language

Latvian is the dominant language: 1 046 412 (73.9%) tweets are written in it.[5] There are 149 366 (10.5%) tweets in Russian, 138 492 (9.8%) in English and 81 714 (5.8%) in other languages. This distribution is very different from a geo-located method used in [6], where the distribution is 44.5% Latvian, 33.9% Russian and 20.7% English. Figure 1 shows the tweet volume over time.

The tweets were written by 41 443 users. We refer to a user who produced at least 50 tweets as an *active user*. There are 2 786 (6.7%) active users who produced 1 229 010 (86.8%) tweets in total.

Among active users, 797 are monoligual. Exclusively in Latvian write 643 (80.7%), in Russian 40 (5.0%) and in English 114 (14.3%).

The *language uniformity score* is defined in [6]. It is the number of tweets in the most frequent language of a user normalized by the total number of tweets of that user. Figure 2 compares the language uniformity scores of active users in this work and in [6]. Apart of the higher number of active users in this work (which is expected, as the corpus covers a longer period), the currently presented corpus has more active users with the language uniformity score about 0.5. Even though the majority is mostly monoligual, this work captured more (at least) bilingual users.

[5]Twitter labels tweets with the language they are written in.

5. Topics

To investigate the collection further, 16 topics of interests were defined. The collection was manually searched using keywords to get relevant tweets. To compensate for rich morphology in Latvian and Russian, keywords were manually stemmed. For example, `hokej` was used instead of Latvian *hokejs* to search for the tweets about ice hockey.

The topics are:[6]

- `LV001`: **Latvia 100** The Centennial Celebration of the Republic of Latvia.
- `LV002`: **Ice Hockey** Ice hockey.
- `LV003`: **Refugees** Refugee crisis in Europe, attitude to immigration and immigrants.
- `LV004`: **Brexit** Withdrawal of the United Kingdom from the European Union known as Brexit.
- `LV005`: **Olympics** The Olympic games and the Latvian team.
- `LV006`: **Winter** Winter, snow and cold weather.
- `LV007`: **May 4** The Restoration of Independence Day.
- `LV008`: **May 8-9** Remembrance of the end of Word War II (May 8). It is also honored as the Victory Day on May 9. The Europe Day is observed on May 9. Tweets about any of the three events are relevant.
- `LV009`: **Midsummer** Midsummer celebration on June 23-24.
- `LV010`: **November 11** A memorial day for soldiers who fought for the independence of Latvia.
- `LV011`: **November 18** Proclamation Day of the Republic of Latvia.
- `LV012`: **Christmas** Christmas.
- `LV013`: **New Year** New Year.
- `LV014`: **March 16** Remembrance Day of the Latvian legionnaires.[7]
- `LV015`: **The Chronicles of Melanie** Melānijas Hronika (The Chronicles of Melanie) is a Latvian movie that was selected for the Foreign-Language Category for the Oscars.
- `LV016`: **Blizzard of Souls** A Latvian movie Dvēseļu Putenis (Blizzard of Souls) that is in production by the time of writing.

Topics exhibit various temporal patterns, as seen in Figure 3. The volume of tweets is constant for long-lasting news stories such as the centennial celebration, the refugee crisis, Brexit and ice hockey. During the Ice Hockey World Championship the volume of hockey related tweets increases. Similarly, event-based topics—such as Christmas, New Year and Olympics—are the most active during corresponding events, the volume of tweets builds up as an event approaches.

In case of Christmas, we see unexpected activity in March which is due to the discussion of making the Orthodox Christmas a public holiday in Parliament. This topic exhibits some cultural differences and language use. Tweets in Latvian peak during Advent and Christmas in December, tweets in Russian reach maximum both in December and early January when Orthodox Christmas are celebrated. It is worth noting that during the discussion about whether Orthodox Christmas should be a public holiday, it sometimes

[6]The topic file is available as topics.json.txt. Relevance judgments are available as relevance_judgments.csv.

[7]For more information about the event, refer to http://www.mfa.gov.lv/en/policy/information-on-the-history-of-latvia/the-latvian-government-s-position-on-16-march-events.

Table 2. Number of relevant tweets per topic and language distribution.

TopicID	Title	Latvian	Russian	English	Tweets
LV001	Latvia 100	87.2%	1.8%	11.0%	7 536
LV002	Ice Hockey	94.7%	2.5%	2.8%	25 699
LV003	Refugees	89.2%	6.9%	4.0%	3 136
LV004	Brexit	85.2%	5.0%	9.7%	2 119
LV005	Olympics	93.5%	4.1%	2.3%	7 371
LV006	Winter	86.1%	10.0%	3.9%	9 334
LV007	May 4	79.7%	1.6%	18.7%	1 459
LV008	May 8-9	84.8%	12.9%	2.3%	1 193
LV009	Midsummer	86.9%	8.1%	5.0%	2 176
LV010	November 11	90.3%	1.9%	7.8%	959
LV011	November 18	74.3%	5.4%	20.3%	936
LV012	Christmas	87.3%	4.8%	7.9%	3 421
LV013	New Year	77.6%	14.5%	7.9%	1 383
LV014	March 16	94.4%	3.5%	2.1%	479
LV015	The Chronicles of Melanie	89.8%	1.0%	9.1%	197
LV016	Blizzard of Souls	100.0%	-	-	219

was referred as “Russian Christmas.” In spring 2018, there was a discussion to make November 11 a public holiday, thus a spike in activity. Other event-related topics (May 4, May 8-9, etc.) behave similarly: their activity peaks during the event.

Topics about movies are the smallest volume-wise. For a movie that was shown at several international festivals, we see multilingual content, while tweets about a movie that is still in production are solely monolingual.

The topic about Winter is an example of a seasonal topic, which is mostly about the weather, the pictures of snow and driving conditions.

Table 2 shows the total number of tweets per topic and language share for Latvian, Russian and English. For all topics, the share of Latvian tweets is higher than on average in the collection (Section 4).

While for Russian only few topics are more active than expected, it is different for English, where some topics are twice more active than expected. The topics where share of Russian tweets is greater than expected 9.8% are New Year, May 8-9 and Winter. English most active topics are Latvia 100, May 4 and November 18, all of which are public celebrations.

6. Conclusion and future work

This paper presented a multilingual tweet collection with some analysis of users, language use and content. The analysis revealed differences in language use between the geo-located collection, global collection and topical sub-collections.

However, the question of whether the collection is any good remains open. It would be easy to test its extrinsic properties, for example, whether it leads to improvements in a language identification system when used as training data. But does not reveal its intrinsic properties. Nevertheless, we believe the presented dataset is useful across different studies.

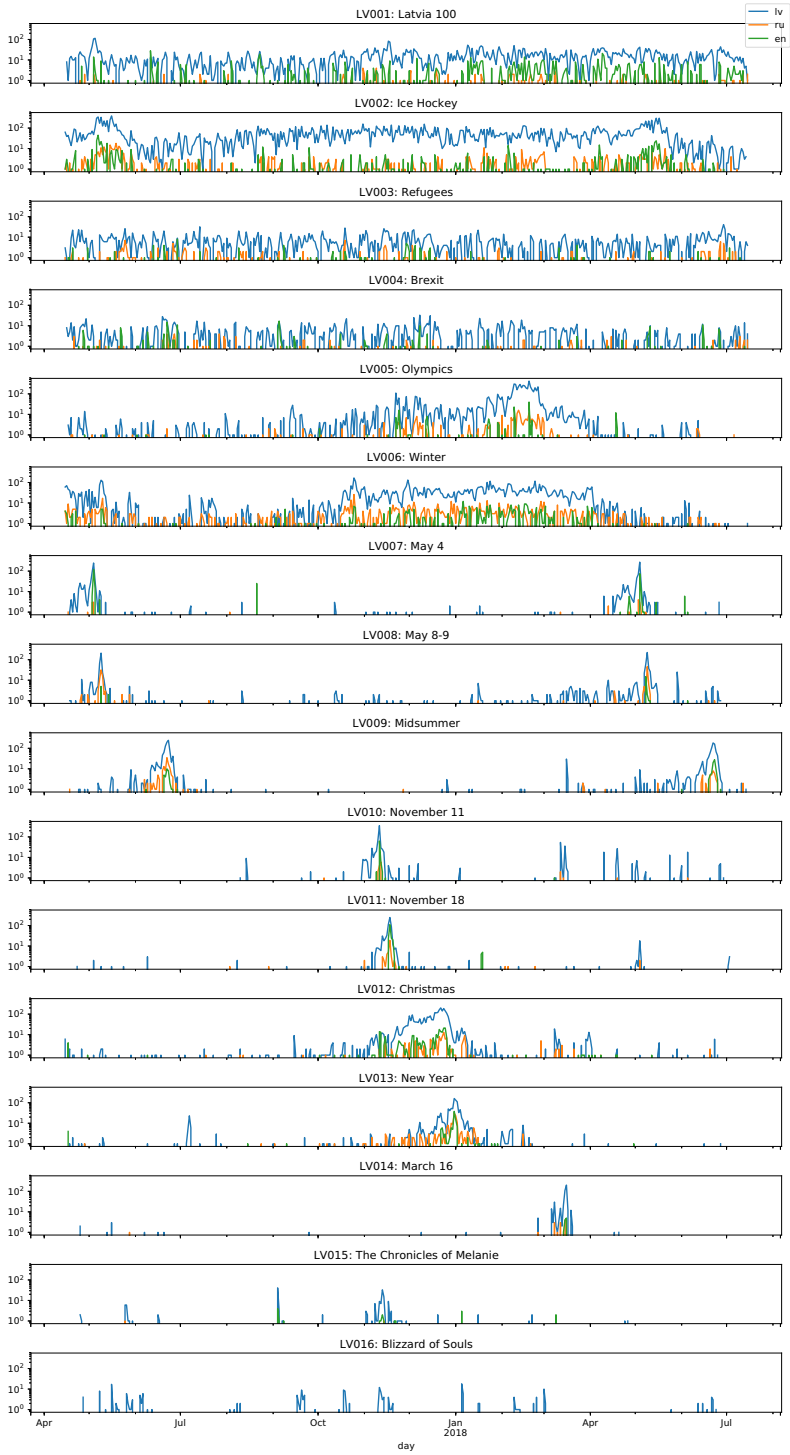

Figure 3. Topic timeline. Note the logarithmic Y-scale, which is the number of tweets.

As the corpus includes tweets of major media outlets and public figures, together with retweets and replies, it might be used to analyze public engagement.

From the sociolinguistic perspective, instead of avoiding collection artifacts such as topical bias, one could study what linguistic environment the tweets represent. What are the factors determining the language, contents and stance of a tweet? How do these factors interact? Who are the people writing those tweets?

Collection exploration using keyword search showed that there is a lot of content on various topics, and it should be feasible to run a shared retrieval task based on a multilingual collection from the Baltic region to evaluate retrieval systems and study what factors determine relevance.

References

[1] Erik Tjong Kim Sang and Antal van den Bosch. Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134, 12/2013 2013. ISSN 2211-4009.

[2] Jimmy Lin, Salman Mohammed, Royal Sequiera, Luchen Tan, Nimesh Ghelani, Mustafa Abualsaud, Richard McCreadie, Dmitrijs Milajevs, and Ellen Voorhees. Overview of the TREC 2017 Real-Time Summarization Track. In *Proceedings of the 26th text retrieval conference, TREC*, 2017.

[3] Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. TweetMT: A Parallel Microblog Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

[4] Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. Tweetcat: a tool for building twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA), 2014.

[5] Steven Coats. *European Language Ecology and Bilingualism with English on Twitter*. CMC-Corpora conference series, Sep 2017.

[6] Dmitrijs Milajevs. Toward a Comparable Corpus of Latvian, Russian and English Tweets. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 26–30, Vancouver, Canada, August 2017. Association for Computational Linguistics.