Toward a Comparable Corpus of Latvian, Russian and English Tweets

Dmitrijs Milajevs
NIST
Gaithersburg, MD, USA
dmitrijs.milajevs@nist.gov

Abstract

Twitter has become a rich source for linguistic data. Here, a possibility of building a trilingual Latvian-Russian-English corpus of tweets from Riga, Latvia is investigated. Such a corpus, once constructed, might be of great use for multiple purposes including training machine translation models, examining cross-lingual phenomena and studying the population of Riga. This pilot study shows that it is feasible to build such a resource by collecting and analysing a pilot corpus, which is made publicly available and can be used to construct a large comparable corpus.

1 Introduction

Comparable corpora are widely used by the natural language processing community to build machine translation or information retrieval models. The goal of this work is to investigate in a pilot study whether it is possible to build a comparable linguistic resource of tweets that originates from one specific location–Riga, Latvia. Riga is a great location for this because it is a multilingual city in which Latvian and Russian are both widely used in everyday life, and English is a lingua franca in tourism and commerce.

Despite the fact that Latvian and Russian are widely used, there is little interaction between the two ethnic communities. The local media consists of two subsystems (Latvian and Russian) which use different sources and present different views on current affairs (Muižnieks, 2010). Even though large media portals tend to have separate Latvian and Russian web-sites, the same opinions are found in comments to controversial content on both versions of web-sites, making the Internet a public space for a dialogue between the ethnic communities (Šulmane, 2010). A corpus of user generated content from Riga is a fruitful resource for studying the integration of the two communities, by identifying what is being discussed; how, and most importantly why it is being discussed.

The pilot corpus[1] consists of tweets over the period of 5 months (November 2016 to March 2017). The main goal of the analysis is to investigate whether a creation of a comparable tweet corpus is feasible and what the corpus construction strategy should be. To see whether the pilot corpus is comparable, the peaks of Twitter usage were analysed. These peaks correspond to real world events (national celebrations, international political affairs and weather). The events are actively discussed in all languages, but in different proportions (Section 4).

All three languages are represented: 45.5% tweets are in Latvian, 33.9% in Russian and 20.7% in English.[2] By studying users' tweeting habits, we see that the majority of users (83.3%) mostly tweets in one language (Section 5), making the tweet collection strategy that considers only multilingual users incomplete.

The properties of the corpus correspond to the expectation that it will reflect the real world events and language use proportion, but its size is too small to draw solid conclusions. However, the construction of a reliable comparable corpus is a matter of the data collection procedure and corpus' application, because, as this study shows, not all topics are discussed equally.

2 Related work

Twitter provides an easy way to build a large text corpus for research. Numerous tweet collections are built for a variety of purposes. For example,

[1] https://doi.org/10.5281/zenodo.582300

[2] The ratio of ethnic Latvians to Russians in Riga is 46.2% to 37.7%.

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 26–30,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

Tjong Kim Sang and van den Bosch (2013) discuss the process of building a large collection of Dutch tweets and challenges of accessing the data. Their retrieval method is based on a list of frequent Dutch words.

Vicente et al. (2016) build a parallel multilingual corpus of tweets. Their process consists of two parts: retrieval and alignment. Retrieval is based on a list of multilingual users. The collected tweets are aligned using crowdsourcing. Ling et al. (2013) automatically extract parallel segments from Sina Weibo (a Chinese counterpart of Twitter). Gotti et al. (2013) use the parallel web pages mentioned in tweets of the agencies and organisations of Canada to train a statistical machine translation model.

There is a small but growing body of research of the Latvian Twittersphere, for example, work on sentiment analysis (Peisenieks and Skadiņš, 2014) and opinion mining (Špats and Birzniece, 2016). Both studies focus on Latvian.

3 Dataset construction

The initial set of tweets was retrieved by subscribing to the `POST status/filter` endpoint of the Twitter Streaming API.[3] The collected tweets had to be geo-located and had to originate from the area of Riga, the capital of Latvia.[4]

251 083 tweets were collected within the period from the 1st of November 2016 to the 31st of March 2017. On April 14th 2017, the collection was rehydrated[5] by querying the Twitter API with the collected tweet IDs to get rid of the deleted tweets. In addition, the tweets that originated from retweets were added to the collection: the JSON[6] representation of a retweet includes the original tweet, which was extracted and added to the collection. The rehydrated and expanded collection resulted in a total of 220 883 tweets.

Further analysis of the extended rehydrated collection showed that there are 23 115 (10.5%) tweets that originated from check-ins on Foursquare.[7] This motivated additional filtering of the rehydrated collection, as "check-in tweets" follow a predefined template most of the time and thus do not reflect real language use.

Client	Tweet count	Share %
Twitter Web Client	93 705	42.4%
Twitter for iPhone	47 721	21.6%
Twitter for Android	34 277	15.5%
Foursquare*	23 115	10.5%
Instagram*	13 196	5.0%
Twitter for iPad	2 420	1.1%
Endomondo*	1 611	0.7%
Tweetbot of iOS	1 411	0.6%
World Cities*	1 361	0.6%
Linkis*	660	0.3%

Table 1: The top ten of Twitter clients in the rehydrated collection. *Clients that are not included in the final collection as they do not exhibit linguistic value.

Table 1 shows the top ten most popular clients in the rehydrated collection. Together with the tweets originating from Foursquare, tweets from Instagram, an image sharing service, and Endomondo, a workout tracking service, were removed. Tweets written using the World Cities client, which posts weather reports, and the Linkis client—a promotion website—were also removed.

The final collection resulted in 136 067 tweets which are in Latvian, Russian or English and created after the 1st of November 2016. The language of a tweet is provided by the corresponding field in the tweet JSON representation.

4 Tweet analysis

Out of 136 067 tweets that constitute the final collection, 45.5% are in Latvian, 33.9% are in Russian and 20.7% are in English, see Table 2 for tweet counts.

Language	Tweet count	Share %	Avg. token count
Latvian	61 869	45.4%	15
Russian	46 070	33.9%	11
English	28 128	20.7%	14

Table 2: Language distribution in the final collection.

Figure 1 shows the number of tweets per day over time for all three languages. There are several peaks in Twitter usage. Some of them affect all three languages, as in early January, some of them affect only one language, as in late January.

If the Twitter behaviour is affected by events in the real world, then these peaks should correspond

[3] https://dev.twitter.com/streaming/reference/post/statuses/filter

[4] The `locations` parameter was set to `23.9325829, 56.8570671, 24.3247299, 57.0859184`

[5] Since distribution of raw tweet data is not allowed, tweets IDs are shared instead. Hydration is the process of retrieval of raw tweet data by IDs.

[6] http://json.org

[7] https://foursquare.com

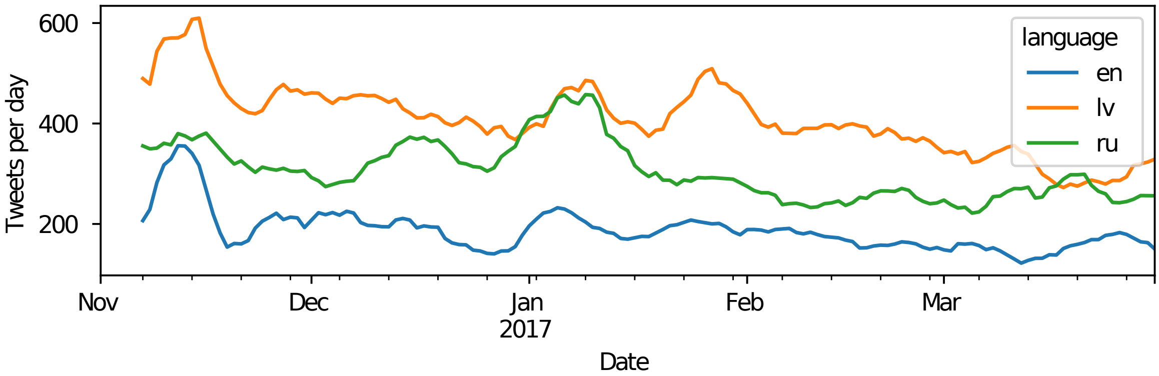

Figure 1: Tweet counts per day per language. The values are averaged over a week window at the right edge.

to events in the real world. The difference in peaks could then be explained as there are different real word events that trigger discussions on Twitter in Latvian, Russian and English. Table 3 suggests, that tweets in Latvian and English share similar behaviour. The Russian tweet timeline is distinct from both timelines, though its behaviour is more similar to the Latvian timeline than to the English.

Language	Latvian	Russian	English
Latvian	1.0	0.4	0.6
Russian	0.4	1.0	0.3
English	0.6	0.3	1.0

Table 3: Pairwise Pearson's-ρ correlation coefficients between Latvian, Russian and English timelines.

What are the distinctive and similar properties of the timelines? To answer the question, we first identify the real world events that happened at the time of the highest peaks.

Mid November 11th of November is Lāčplēsis Day, a memorial day for soldiers who fought for the independence of Latvia. 18 November is the Proclamation Day of the Republic of Latvia. Also, on the 8th of November, the United States presidential election took place.

The number of tweets significantly increased for Latvian and English, and not so much for Russian. Manual inspection of the tweets in that period reveals that the US elections are discussed in all three languages, while the national celebrations of the 11th and the 18th of November are mostly discussed in Latvian. The discussion includes such topics as the news related to celebrations, historical notes, reminders of working hours of businesses, greeting and advertisement.

Manual inspection also shows that events are language sensitive. For example, the election results were discussed by Latvians in English. Also, businesses reported their working hours during the national celebrations in Latvian and do not duplicate this information in Russian.

Early January In early January a snowstorm hit Riga. In Latvian and Russian the discussed topics were the same, namely, appreciation of snow, the transportation difficulties and outdoor photos. Tweets in English mostly contained photos showing how beautiful Latvia is in Winter.

Late January The inauguration of the 45th President of the United States was held on 20th of January 2017. The number of Latvian tweets increases, while for other languages it stays roughly the same. The reason why there are relatively little politics-oriented Russian tweets might be that 60% of citizens and 47% of non-citizens are interested in politics (Aldermane et al., 2000). Out of citizens, 60% are ethnic Latvians, 27% are ethnic Russians. Out of non-citizens, 66% are Russians, and less than 1% are Latvians.[8]

5 User analysis

We have seen an evidence that topics are languge dependant. How many Twitter users switch between languages?

[8] https://lv.wikipedia.org/wiki/Nepilsoņi_(Latvija)

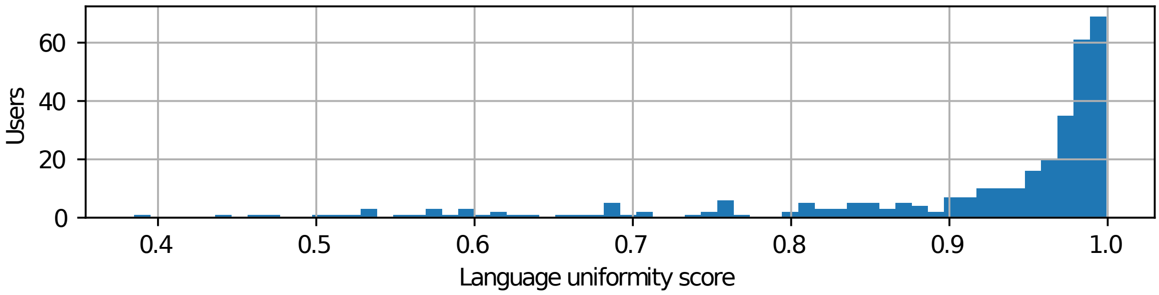

Figure 2: Histogram of language use uniformity scores. Low values mean that distinct languages are used, while high values mean that a single language is preferred.

We consider 507 users for whom at least 50 tweets were collected. 180 or 35.5% of them tweet exclusively in one language (75 users tweet only in Latvian, 43 in Russian and 62 in English). Others tweet in several languages.

To get more insight on how languages are used, we compute the language uniformity score defined as:

$$\frac{\max(n_{lv}, n_{ru}, n_{en})}{n_{lv} + n_{ru} + n_{en}} \quad (1)$$

where n_{lv} corresponds to the number of tweets in Latvian for a given user, n_{ru} to the number of tweets in Russian, and n_{en} to the number of tweets in English.

The higher the score, the more dominant a language. The lowest possible value of 0.33 means that all three languages are used equally. The value of 0.5 means that 50% of tweets are written in a dominant language. The value of 1 means that the user tweets exclusively in one language.

The histogram in Figure 2 shows the score distribution. 420 (82.8%) users tweet mostly in one language (their scores are greater than 0.9). For 83 (16.4%) users the score is between 0.5 and 0.9. There are only four (0.8%) users whose dominant language share is less than 50%.

Among the four Twitter users whose score is less than 0.5—meaning that they use all three languages extensively—three are personal accounts and one is a company account. Other interesting accounts that tweet equally in Latvian and Russian, but do not tweet in English are the accounts of a library and a football club.

To illustrate the language usage pattern between multilingual users, their first most frequently used language, their second most frequently used language and their third most frequently language were identified. If a user tweeted equally in two (three) languages, then the two (three) languages were given the maximal preference. A user who tweeted equally in Latvian and Russian, but less in English, is counted as Latvian and Russian being their first preference, English as the third.

Latvian is not only the most used language among the monoligual users, but also is the first and third most common choice between the multilingual users. The preference for Russian is similar to Latvian, despite the numbers being slightly lower, suggesting its significant role in everyday life. English is almost the ultimate second choice, proving its role as a lingua franca, as Table 4 shows.

	Latvian	Russian	English
Monoligual	**75**	43	62
Multi, first	**150**	135	42
Multi, second	56	19	**266**
Multi, third	**29**	26	9

Table 4: Language preference between users.

6 Conclusion

We have seen that location-based tweet collection produces adequate results. Tweets in all three target languages were collected, and the resulting collection reflects real world events.

How comparable are the language samples within the corpus? Topics are language dependent, so it is not the case that all topics are discussed in every language. There are "monolingual topics" such as the independence day in Latvia. Even "multilingual topics" vary in content, as with the snowstorm tweets, where Latvian and Russian tweets shared common topics, but tweets in English were distinct.

The final answer is that it depends on the application. For machine translation, it is important to have similar content, so the parallel segments can be extracted, for example from Latvian and Russian snowstorm tweets. For a social study, the corpus has to be representative, so that the topics can lead to the analysis of the communities as in the case of why the president inauguration was discussed much less in Russian than in Latvian.

Acknowledgements

The author thanks Dr. Sascha Griffiths for valuable discussion and Tatiana Chepurko for literature suggestions.

Disclaimer

Certain commercial products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the identified products are necessarily the best available for the purpose.

References

Eiženija Aldermane, Reinis Āboltiņš, Heidi Bottolfs, Boriss Cilēvičs, Jānis Jaudzems, Anita Jākobsone, Ābrams Kleckins, Falks Lange, Jānis Mažeiks, Ilmārs Mežs, Nils Muižnieks, Artis Pabriks, Aija Priedīte, Ilona Stalidzāne, Inese Šūpule, Ramona Umblija, Elmārs Vēbers, and Brigita Zepa. 2000. "Towards a Civic Society-2000" Survey of Latvian Inhabitants. Baltic Institute of Social Sciences.

Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. 2013. Translating Government Agencies' Tweet Feeds: Specificities, Problems and (a few) Solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*. Association for Computational Linguistics, Atlanta, Georgia, pages 80–89. http://www.aclweb.org/anthology/W13-1109.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 176–186. http://www.aclweb.org/anthology/P13-1018.

Nils Muižnieks, editor. 2010. *How integrated is Latvian society: An Audit of Achievements, Failures and Challenges*. University of Latvia Press.

Jānis Peisenieks and Raivis Skadiņš. 2014. Uses of Machine Translation in the Sentiment Analysis of Tweets. *Frontiers in Artificial Intelligence and Applications* 268(Human Language Technologies-The Baltic Perspective):126131. https://doi.org/10.3233/978-1-61499-442-8-126.

Gatis Špats and Ilze Birzniece. 2016. Opinion Mining in Latvian Text Using Semantic Polarity Analysis and Machine Learning Approach. *Complex Systems Informatics and Modeling Quarterly* (7):51–59.

Ilze Šulmane. 2010. The Media and Integration. In Nils Muižnieks, editor, *How integrated is Latvian society: An Audit of Achievements, Failures and Challenges*, University of Latvia Press.

Erik Tjong Kim Sang and Antal van den Bosch. 2013. Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal* 3:121–134.

Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. 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)*. European Language Resources Association (ELRA), Paris, France.