Abstract

The ongoing controversy regarding Ukraine’s language laws has highlighted the need for empirical research on language use in the country. Election and census results show that the country has two internal north-south borders: an electoral fault line that runs northeast to southwest along the eastern borders of Poltava and Kirovohrad oblasts (regions), and a linguistic border based on self-reported mother tongue that divides Luhansk, Donetsk, and Crimea from the rest of the country.

This article uses data from the microblogging service Twitter to analyze the geography of language in Ukraine. Based on a dataset of 2.4 million geotagged tweets collected over four months in 2015, we found that online language use largely reflects the country’s internal electoral border. In addition, we found widely different rates of bilingual communication: whereas more than half of those using Ukrainian also tweeted in Russian, fewer than one in ten of those using Russian also tweeted in Ukrainian. Use of both languages was higher in urban areas for both groups.
BORDERS OF IDENTITY AND ACTUAL LANGUAGE USE IN UKRAINE: AN ANALYSIS OF GEOTAGGED TWEETS

Introduction

Ukraine is a bilingual country. Most of the population can speak and understand both Ukrainian (the official language) and Russian. Generally, it is expected that everyone living in Ukraine should know both languages, at least passively. For Ukrainians, the two languages are mutually intelligible, which is proven by the common use of them in the media, education, and governmental institutions, including Ukraine’s Parliament. Ukrainians use the language that they consider to be most appropriate for the situation or more comfortable for personal use. Although language choice should not be considered the key marker of national identity in Ukraine (Kuzio, 2002; Bilaniuk, 2005), it is often associated with the political choice of an individual. Kiryukhin (2015) notes that

“The language in the case of Ukraine is one of those obvious and self-explanatory agents that allow, within the scope of identity politics, to draw the line between ‘us’ and ‘them’ … between ‘the Ukrainians’ and ‘the Russians.’ At the same time, a proportion of the ethnic Ukrainian population considers Russian to be their native language, and a number of Russian-speaking ethnic Ukrainians still count Ukrainian as their mother tongue.”

Kulyk (2011) concluded from his research on language identity and political cleavage in Ukraine that “language identity is a no less powerful predictor of Ukrainian citizens’ attitudes and policy preferences than language use.”

Language use is complex and flexible. As Fomina (2014) reported:

“It is not infrequent that while having a conversation, one person speaks Ukrainian and the other—Russian. Besides, especially in central Ukraine, many people speak so-called surzhyk, a combination of Russian and Ukrainian. Yet, when asked about their reliance on surzhyk, people may deny it and claim that they actually speak either Russian or Ukrainian.”

It is also quite common that people use both Ukrainian and Russian in the same conversation, switching from one language to another. Such code-switching takes place not only in big cities but also in small towns or
BORDERS OF IDENTITY AND ACTUAL LANGUAGE USE IN UKRAINE: AN ANALYSIS OF GEOTAGGED TWEETS

villages. After Ukraine gained its independence, it generally became culturally correct to treat language in public as transparent, reacting tolerantly and indifferently to language choice (Bernsand, 2001; Bilaniuk, 2005; Olszanski, 2012; Pavliy, 2014).

Set against this picture of pragmatic everyday bilingualism, Ukraine’s politics have been marked by a deep and consistent division between parties espousing closer ties with Russia, which enjoy strong support in the country’s eastern areas, and parties with a voter base in the west that tend to be suspicious of Moscow. One consequence of this divide has been seismic and controversial changes in language policy as different parties have alternated in power. Furthermore, issues of language rights and the legal status of languages have been, if not the cause, then at least one of the justifications used for political violence.

Ukraine’s unstable political situation and the danger that language issues will become politicized and used to justify violence make it more important than ever to understand how the country’s population is actually using language on a daily basis. The most recent census was carried out in 2001; since then, the country has experienced economic, political, and social upheaval. Additionally, the census only ascertained respondents’ mother tongue and not their actual language use. To what extent do patterns of language use reflect the country’s internal political and linguistic borders as expressed in election and census results?

The widespread adoption of social media by Ukrainians offers an affordable, non-invasive way of observing everyday communicative behavior and, hence, language use by that part of the country’s population that is active online. The results are available quickly, and it is possible to use location information (geotags) to map patterns of language use on a regional basis.

This article presents the first results of an ongoing project to examine language use by social media users in Ukraine. We used the location and language information of publically available geotagged Twitter posts to map the absolute and relative quantities of Ukrainian and Russian tweets sent in different parts of the country. We also used information about the users to calculate rates of bilingual use among those tweeting in the two languages, again, on an oblast-by-oblast basis.

This article makes three main contributions to scholarship. First, it
Bogdan Pavliy, Jonathan Lewis

establishes the feasibility of using social media (specifically, Twitter) to undertake research on the geography of language use in Ukraine. Second, it offers the (to our knowledge) first geographical analysis of language use on social media in Ukraine—results that, in view of the ongoing unrest, will provide useful empirical material for researchers in political science and peace and conflict studies as well as sociologists and sociolinguists. Third, our analysis of bilingualism is (again, to our knowledge) a methodological innovation—one that is made possible by our large dataset.

We realize that the article might be seen as a continuation of research exploring the involvement of people in the issues of national identity, politics, and language policies that, in the case of Ukraine, has been conducted by a great number of scholars, including Arel (1995, 2002, 2014), Wilson (1995), Kuzio (1997, 2000, 2001, 2002, 2006), Laitin (1998), Janmaat (1999), Shulman (1999), Wolczuk (2000), Berndt (2001), Fournier (2002), Bilanski (2005), Fimyan (2008), Polese (2010, 2011), Zhurzhenko (2010, 2011), Kulyk (2011), and Olszanski (2012). However, this current research does not tackle any identity issues of the Ukrainian population on either a local or national level. Our approach is focused mainly on establishing methods of gathering and classifying data and analyzing the linguistic preferences of particular Twitter users based on their locations.

The article is structured as follows. First, we briefly outline the geographical aspects of the country’s electoral politics and language use and note the controversial language politics that provide the context and motivation for this study. In the second section, we discuss the advantages, limitations, and main elements of using social media data to investigate communicative behavior and note previous research that has mapped language use in other parts of the world. The third section describes in more detail our data collection and cleaning methods. Section four presents our results and map quantities of Ukrainian and Russian tweets and bilingualism. The final section offers some conclusions and discusses topics for further investigation.
Electoral Borders, Linguistic Borders, and Language Politics
Ukraine’s Political and Linguistic Divides

The results of national elections held in the country since 2004 show a consistent geographical divide between southern and eastern Ukraine on one hand and northern and western Ukraine on the other. The Presidential elections of 2004 and 2010 and the Parliamentary elections of 2007 and 2012 all showed a political fault line running northeast to southwest along the eastern borders of Poltava and Kirovohrad oblasts. Figure 1 illustrates this, using the results of the 2012 Parliamentary election; however, the border is identical for the other three elections.

According to the 2001 census, Russian was the majority native language in three oblasts: Luhansk, Donetsk, and Crimea. Hence, as Figure 2 shows, we can delineate two borders within the country: a linguistic divide, in terms of levels of the reported mother tongue in the far east, and an electoral divide farther west.

These political and linguistic divisions have been reflected in Ukraine’s
language politics since its independence. We will briefly outline the major developments in Ukrainian language politics in order to establish the context of and need for our empirical research on language use in the country.

The 2012 Language Law and the Status of Russian

The legal status of the Russian language has been a divisive issue in Ukraine since its independence. Until 2012, Ukrainian language policy was set under the 1989 law, “On the languages in the Ukrainian SSR.” The law recognized Ukrainian as the only official language in all of Ukraine and did not grant Russian the status of a second official or regional language in any administrative district. When Viktor Yanukovych’s Party of Regions took power in 2012, it enacted a new law under which, if the percentage of representatives of a national minority in any administrative district of
Ukraine exceeds 10% of the total population of the district, the status of regional language should be granted to the language of the minority. The language can then be used in all governmental institutions in the district (including schools, courts, and governmental offices), along with the official language, Ukrainian.

It was clear that the new law would enable Russian to become an official regional language in some oblasts and cities of Ukraine. There was considerable opposition to the law; protests were held in Kyiv and other Ukrainian cities, fistfights broke out in the Ukrainian Parliament, and the Chairman of the Parliament tried to resign in order to prevent the law from being ratified. His resignation was refused, and he had to sign the law, which was ratified by President Yanukovych and came into force on August 10, 2012.

Following the implementation of the law, Russian became a regional language in several oblasts (Odessa, Mykolayiv, Kherson, Dnipropetrovs’k, Zaporizhzhya, and Donetsk) and cities (Odessa, Mykolayiv, Kherson, Dnipropetrovs’k, Zaporizhzhya, Kharkiv, Luhansk, Krasnyi Luch, and Sevastopol’) of southern and eastern Ukraine.

Three pro-European Ukrainian political parties—Svoboda, UDAR, and Batkivshchyna—viewed this law as threatening to increase the status and use of the Russian language throughout Ukraine and, in turn, offering Russia more opportunities to bring the Ukrainian population into its political sphere of interest. The day after ousting President Yanukovych, on February 22, 2014, the three parties capitalized on their political momentum with the Ukrainian Parliament’s vote to abolish the 2012 law. They then started preparing a new bill on language policy that contained no provisions for granting regional language status to Russian (Pavliy, 2014; Moser, 2015).

This attempt to reduce the legal status of Russian in Ukraine attracted widespread criticism both domestically (Arel, 2014) and from neighboring governments, such as Russia and Poland. A few weeks later, it was cited by Russian media as one reason for the separatist movement and military conflict in Donetsk and Luhansk oblasts. Amidst these tensions, the abolition of the 2012 language law was vetoed by acting president Oleksandr Turchynov.

This, then, is the highly sensitive political context in which the discussion
of language rights and use in Ukraine takes place. Clearly, there is a need for empirical research on language use that will facilitate informed and constructive debate among academic researchers and policymakers, both in Ukraine and abroad.

**Using Microblogging Data to Examine Geographical Patterns of Language Use**

A decade after the mass adoption of social media around the world, social scientists have started to exploit its potential for offering insights into many aspects of human behavior. Of the popular social media, academic researchers have most commonly focused on the microblogging service Twitter. Twitter is the most researched social media for two main reasons: first, tweets are public, which makes them much easier to use from a research ethics perspective, as compared to services such as Facebook, where posts are generally only visible to a restricted group; second, Twitter provides application programming interfaces (APIs) that allow the automated collection of tweets that meet specified criteria.

Twitter also allows users to publish their location at the time of posting. For privacy reasons, the user is required to opt in to location publishing; as a result, only a small percentage of tweets are geotagged. Nevertheless, given sufficient total volume, even a small percentage of traffic can offer the researcher many insights. Twitter’s Streaming API lets researchers request all tweets geotagged within a given area, making it simple to set up a continuous collection of tweets sent from a certain territory. The convenience and accessibility of Twitter data have given rise to a large body of research that uses the microblogging service as a social sensor to examine aspects of human behavior as diverse as political debate (Conover et al., 2011), the spread of rumors following natural disasters (Takayasu et al., 2015), and positive and negative reactions during sporting events (Takeichi et al., 2014).

Investigation of the linguistic aspects of communication on Twitter is facilitated by the existence of language detection algorithms that permit the automatic identification of the language or languages used in the text of tweets. Twitter runs its own language detection algorithm on each tweet and provides the result—a single language tag for each tweet. In cases where Twitter’s own algorithm proves insufficiently accurate or where recognition
of multiple languages is required, other algorithms, such as Google’s Compact Language Detector, are available.

A number of researchers have made use of these data and tools to study the geographic aspects of linguistic behavior. Mocanu et al. (2013) produced a global language map of tweets and demonstrated the usefulness of geotagged tweets as a way of studying such phenomena as the linguistic homogeneity of different countries, seasonal movements such as tourism, and the geographical distribution of different languages in multilingual regions. They showed that it is possible to undertake fine-grained analysis of language behavior in urban areas with large volumes of Twitter traffic, illustrating this by showing where particular languages are used in Montreal and New York City. Other wired cosmopolitan cities, such as London, have also been the subject of this kind of analysis (for example, Cheshire and Uberti, 2014).

This article adopts the same descriptive, effectively atheoretical approach taken by the above-noted studies of language use in social media around the world. We believe this approach is justified because there is (to our knowledge) no existing research on the geography of language use in Ukrainian social media. The researcher's highest priority at this pioneering stage is, therefore, to establish methods of gathering and classifying data, analyzing its basic characteristics (in this case, location, language, and user ID), and relating those basic characteristics to other available data (in this case, the territories of administrative regions and election and census results). This can then serve as a foundation for research that adopts particular sociolinguistic and linguistic frameworks and methods.

Using social media such as Twitter does, however, have significant limitations and in no way replaces more traditional survey methods. These limitations have been pointed out by, among others, Boyd and Crawford (2012) and Zeitzoff et al. (2015). Most critically, social media users are a self-selecting sample and not a random sample of the population. Furthermore, we lack reliable basic demographic information about Twitter users, such as age or gender, that could be used to correct sample bias. Thus, however many tweets we collect, we cannot draw conclusions regarding the population as a whole. Some researchers have regarded this self-selecting bias as a strength rather than a shortcoming: for example, Pearce et al. (2014), in their study of communication regarding climate change on Twitter,
argued that those tweeting regularly about the issue are likely to be opinion leaders and, hence, particularly worth studying. Furthermore, to the extent that social media users tend to have a lower average age than the population as a whole, we can make the case that trends observed in online behavior could provide early indicators of changes among the general population. Finally, the pervasive nature of social media use makes it likely that language use online does correspond to language use offline to a considerable extent, although that supposition requires further investigation.

Data Collection and Cleaning

We collected geotagged tweets sent from the territory of Ukraine (including Crimea) from the Twitter Streaming API between April 11 and September 15, 2015. This was achieved by writing a Python script using the tweepy library, which established an open connection to the Twitter Streaming API and specified the geo-coordinates of a bounding box that contained the territory of Ukraine (including Crimea).4 Whenever a geotagged tweet was sent from within the bounding box, our program received it and stored it in a PostGreSQL database. Our program ran almost continually, with a small number of interruptions due to connectivity and server maintenance issues. We then excluded tweets sent from areas in our bounding box that were outside the territory of Ukraine.5 We also excluded tweets generated by the location service Foursquare that merely included text information about the user’s location. Table 1 shows the numbers of tweets in the resulting dataset, grouped by language according to Twitter’s language detection algorithm.6

Results

First, we counted the number of Ukrainian and Russian tweets sent in each of Ukraine’s 27 oblasts; Table 1 gives the results, together with the number of tweets sent in all languages and the ratio of Ukrainian to Russian tweets.

In order to test whether these results are consistent over time or merely the result of brief, intensive bursts of activity in particular areas, we calculated the Ukrainian-to-Russian ratios for each of the complete calendar months in
our dataset, i.e., May to August 2015, and then compared the average monthly ratios and standard deviations for each oblast. In no oblast was the standard deviation greater than 0.45 of the average monthly ratio, and, in all except three oblasts, the standard deviation was less than 0.2 of the average; hence, we concluded that these results are consistent over the period covered by our dataset.

The total numbers of tweets may seem impressive; however, how many users are sending them? Table 3 gives the breakdown by oblast for total number of users, those tweeting in Russian, in Ukrainian, and in both languages. While the numbers are small as compared to the whole population, it is clear that we are observing the behavior of more than a tiny group of enthusiasts. Furthermore, when we recall that geotagged tweets have been determined to be only a few percent of all tweets sent, we have reason to claim that our dataset has the potential to offer insights into the everyday communicative behavior of a sizeable, if not a representative, portion of those living in Ukraine.

Table 2 makes it clear that Russian tweets heavily outnumber Ukrainian tweets everywhere. In the country as a whole, more than six Russian tweets are sent for every Ukrainian tweet; in only two oblasts, Ivano-Frankivs'k and Transcarpathia, does the ratio of Ukrainian to Russian tweets creep above

### Table 1 Number of tweets in dataset by language

| Language   | Number of tweets |
|------------|------------------|
| Russian    | 1528181          |
| Ukrainian  | 240732           |
| English    | 205773           |
| Unknown    | 136214           |
| Slovene    | 70984            |
| Polish     | 64999            |
| Bulgarian  | 49952            |
| Spanish    | 21777            |
| Turkish    | 15723            |
| Bosnian    | 9039             |
| Others     | 66234            |
| Total      | 2409608          |

87
1:2. However, Figure 3, which plots the contents of Table 2 on a map, does show a clear trend of higher ratios of Ukrainian to Russian tweets in the west of the country. Of the five bands of values used in Figure 3, which are derived using the Jenks natural breaks classification method, the band with
the lowest values forms a contiguous block containing 11 eastern and southern oblasts. The border between that band and its western neighbors is almost the same as the electoral border in Figure 2, except that Sumy oblast is on the eastern side of the border.

Clearly, there is a discrepancy between this observed online behavior and the census data regarding mother tongue, as noted above. In the absence of other explanations, we hypothesize that many Twitter users in Ukraine who regard Ukrainian as their mother tongue are using Russian to communicate online.

The last two columns of Table 3 offer support for this hypothesis. They show that, of all users tweeting in Ukrainian, overall, more than half also tweet in Russian, whereas of all users tweeting in Russian, only one in ten also tweets in Ukrainian. Figures 4 and 5 map the proportions of users tweeting in Ukrainian and Russian, respectively, who tweet in both languages. While neither figure shows a distinct east-west divide similar to the maps derived from electoral and census (mother-tongue) data, Figure 4 does suggest that those tweeting in Ukrainian in the east of the country are even more likely than those in the western parts to also tweet in Russian.

It is also noteworthy that, while the percentages of Russian users tweeting
in both languages are much lower throughout the country, for both groups, rates of bilingual communication are highest in the oblasts containing the country’s three largest cities: Kyiv City, Dnipropetrovs'k, and Kharkiv. This

Table 3 Numbers of users by language and oblast

| Oblast       | Total unique users | Tweeting in Uk only | Tweeting in Ru only | Tweeting in both | % Uk users tweeting in both | % Ru users tweeting in both |
|--------------|--------------------|---------------------|---------------------|-----------------|----------------------------|----------------------------|
| Kyiv City    | 45758              | 2696                | 25858               | 5946            | 68.80                      | 18.70                      |
| Dnipropetrovs'k | 41662           | 1906                | 24909               | 4739            | 71.32                      | 15.98                      |
| Kharkiv      | 28240              | 1466                | 17681               | 2298            | 61.05                      | 11.50                      |
| Odessa       | 27718              | 1489                | 17430               | 2115            | 58.68                      | 10.82                      |
| Donetsk      | 26345              | 1380                | 16872               | 1800            | 56.60                      | 9.64                       |
| Kyiv         | 23659              | 1584                | 14743               | 1818            | 53.44                      | 10.98                      |
| L'viv        | 23532              | 1983                | 14098               | 1741            | 46.75                      | 10.99                      |
| Zaporizhzhya | 22040              | 1289                | 14135               | 1444            | 52.84                      | 9.27                       |
| Crimea       | 18107              | 1053                | 11973               | 874             | 45.36                      | 6.80                       |
| Vinnytsya    | 17177              | 1147                | 10773               | 1258            | 52.31                      | 10.46                      |
| Cherkasy     | 16737              | 1146                | 10533               | 1068            | 48.24                      | 9.21                       |
| Mykolayiv    | 15500              | 935                 | 9993                | 815             | 46.57                      | 7.54                       |
| Kherson      | 14373              | 882                 | 9300                | 656             | 42.65                      | 6.59                       |
| Poltava      | 14287              | 857                 | 9230                | 701             | 44.99                      | 7.06                       |
| Chernivtsi   | 14070              | 905                 | 8964                | 725             | 44.48                      | 7.48                       |
| Chernihiv    | 13083              | 803                 | 8391                | 684             | 46.00                      | 7.54                       |
| Kirovohrad   | 12861              | 796                 | 8286                | 544             | 40.60                      | 6.16                       |
| Volyn        | 11490              | 890                 | 7171                | 576             | 39.29                      | 7.44                       |
| Luhansk      | 11421              | 713                 | 7471                | 436             | 37.95                      | 5.51                       |
| Khme'nyst'kyv | 11276           | 914                 | 7095                | 532             | 36.79                      | 6.98                       |
| Zhytomyr     | 10524              | 760                 | 6688                | 485             | 38.96                      | 6.76                       |
| Rivne        | 9608               | 762                 | 5959                | 493             | 39.28                      | 7.64                       |
| Ivano-Frankivs'k | 9343            | 789                 | 5777                | 371             | 31.98                      | 6.03                       |
| Sumy         | 8091               | 556                 | 5158                | 302             | 35.20                      | 5.53                       |
| Sevastopol'  | 7605               | 490                 | 4983                | 210             | 30.00                      | 4.04                       |
| Transcarpathia | 7490             | 617                 | 4626                | 256             | 29.32                      | 5.24                       |
| Ternopil'    | 6217               | 536                 | 3835                | 219             | 29.01                      | 5.40                       |
| TOTAL        | 468214             | 29344               | 291932              | 33106           | 53.01                      | 10.19                      |
Figure 4: Proportions of users tweeting in Ukrainian who also tweet in Russian.

Figure 5: Proportions of users tweeting in Russian who also tweet in Ukrainian.
makes intuitive sense, in that people living in urban areas can be expected to have more diverse social contacts than those in rural areas, and offers further support for our use of Twitter data to measure the geography of language use in the country.

Limitations and Assumptions

Clearly, the active Twitter users in our sample are but a small minority of the Ukrainian population, and their language use on Twitter might not necessarily correspond to their language use in other areas of activity. It is also possible that our sample contains a higher proportion of Russian native speakers than the population of the country, to the extent that urban areas in Ukraine have a higher proportion of Russian native speakers than do rural areas, and that city dwellers tend to be younger and more Internet-connected than those living in the country. Even without these limitations specific to the case of Ukraine, the lack of demographic information available about Twitter users makes it important to emphasize that we are not claiming that our sample is representative of the country’s population as a whole. We merely argue that our sample is representative of the large Twitter-using population in the country, assuming that no particular part of the Twitter-using population is more likely to geotag its tweets than another.

For this study, we also ignored the question of mobility. Most users who send geotagged Tweets do so from mobile devices. As this study counts tweets sent in different regions of Ukraine, it is almost certain that our dataset includes tweets sent from different regions by the same user. We see this as a feature not a bug, in the sense that our dataset provides insights into the communicative behavior of people where they actually are, rather than the single locations where they may be officially registered for purposes of voting or residing but where they may not, in fact, spend much of their time. However, when we compare the results of our analysis of Twitter traffic to election and census results, we need to remember the qualitative differences between the three sources. Unlike election and census data, counting tweets will give greater weight to those who tweet more often; and our counts of tweets and users include anyone who happens to be in a given location, rather than only local residents.
Conclusions

The ongoing controversy regarding Ukraine’s language laws has fueled domestic divisions and international tensions. In order to offer a more solid foundation for future language policies, more empirical research is needed to understand how people living in different parts of the country use language on an everyday basis.

We have helped to address this need by demonstrating the feasibility of using data from social media (specifically, Twitter) to capture information about the language use of a large number of people across the country. While we cannot make claims about the whole Ukrainian population on the basis of our self-selecting sample, our results are sufficiently in tune with other data, such as electoral maps and higher rates of bilingualism in urban areas, to justify treating them as a valid source of sociological data. However, we are certainly not claiming to provide a complete account of the complex realities of language use in Ukraine, which could only be achieved through a more comprehensive multi-disciplinary study.

Our analysis suggests that Ukraine’s language border lies not in the country’s far east, where the majority of those surveyed by the 2001 census declared Russian to be their mother tongue, but rather more centrally, either following or even veering to the west of the electoral border that has been drawn in national elections since 2004.

We also found a stark difference in language use between those who tweet in Ukrainian and those who tweet in Russian. Whereas more than half of those using Ukrainian also tweeted in Russian, fewer than one in ten of those using Russian also tweeted in Ukrainian. Use of both languages was higher in urban areas for both groups. More work is needed to investigate this result, and we must avoid simplistic explanations and extrapolations; however, it does suggest that we may need to reassess the picture of all-round bilingualism painted by previous research cited in the introduction.

This article suggests a number of avenues for future research. One is a more micro-level analysis of language use, including bilingualism in particular oblasts, e.g., the largest cities or areas experiencing political violence. Another is a chronological analysis, both of shorter-term seasonal
trends in movements within the country and of longer-term developments over a number of years. We could capitalize on the mobile nature of most Twitter use to examine the movements of people around the country and investigate how this relates to bilingualism and other aspects of language use. Finally, analysis of the content of tweets could offer insights into the geographical and linguistic aspects of bilingual communication as well as the use of surzhyk.

Acknowledgments

We thank NIFTY Corporation for the use of the C4SA cloud hosting service for data collection. We also thank the developers of the following open source software packages used in this research: PostGreSQL, Python, and QGIS.

Notes

1. See for example the online records of parliament http://rada.gov.ua/meeting/stenogr/.
2. Official election results can be found at http://www.cvk.gov.ua/pls/vp2004/wp0011e.
3. More recently, an exit poll conducted by the R&I group in 2010 showed that the linguistic preferences in three major regions of Ukraine—the west, the center, and the southeast—are quite different, with Russian being highly prioritized in the southeast. At the same time, respondents’ individual assessments of their knowledge of both languages showed that their level of knowledge of Russian (speaking, writing, and reading) is higher (76%) than the level of their Ukrainian (69%).
4. Corners of the bounding box (west, south, east, north): 21.64, 44.10, 40.26, 52.64
5. We installed the PostGis extension to our database and imported a shapefile of the Ukrainian national territory and administrative areas downloaded from http://www.gadm.org. We converted the geo-coordinates provided by the Twitter API in the format (longitude, latitude) to a geospatial Point object using a query similar to the following:
   UPDATE statuses SET geom = ST_GeomFromText('POINT(36.209359 49.985544)')
   WHERE id = 599514912814190592.
   (Here, statuses is the table containing our tweets, the geom column contains the Point data, and the ID column contains the unique ID of the tweet.) Then, we can delete tweets sent from outside Ukrainian territory using the following query:
   DELETE FROM statuses WHERE NOT ST_CONTAINS(ukr_adm0.geom,statuses.geom).
   Where ukr_adm0 is the table containing details of the national territory, stored as a geospatial MultiPolygon object in the geom column.
6. Further research is required to assess the accuracy of Twitter’s algorithm in recognizing Ukrainian and Russian and other languages used in the region. It would also be useful to
investigate use of surzhyk creole in Tweets. Such research is beyond the scope of this paper; therefore, here we assume that Twitter’s algorithm achieves acceptable levels of accuracy.

7. In comparative statistics, the lower the standard deviation as compared to the average, the more reliable the average is as a guide to the size of the individual values.

8. The Jenks natural breaks classification method is one of five classification methods available in the QGIS software used to prepare maps for this article.

9. Of course, election results and census results are also qualitatively different. For example, election results do not reflect the opinions of those who do not vote, whereas census results should, in theory, offer a complete snapshot of the country’s population.

References
Arel, D. (1995) “Language Politics in Independent Ukraine: Towards One or Two State Languages,” Nationalities Papers, 23 (3), pp. 597–621.
Arel, D. (2002) “Interpreting ‘Nationality’ and ‘Language’ in the 2001 Ukrainian Census,” Post-Soviet Affairs, 18 (3), pp. 213–249.
Arel, D. (2014) “Double-Talk: Why Ukrainians Fight Over Language,” Foreign Affairs, March 18, 2014. Available: http://www.foreignaffairs.com/articles/141042/dominique-arel/double-talk [Accessed on September 29, 2015]
Bernsand, N. (2001) “Surzhyk and National Identity in Ukrainian Nationalist Language Ideology,” Berliner Osteuropa-Info, 17. Available: http://www.oei.fu-berlin.de/media/publikationen/boi/boi17/11_bernsand.pdf [Accessed on September 2, 2015]
Bilaniuk, L. (2005) Contested Tongues: Language Politics and Cultural Correction in Ukraine, Ithaca: Cornell University Press.
Boyd, D. and Crawford, K. (2012) “Critical Questions for Big Data,” Information, Communication & Society 15:5, pp. 662–679.
Cheshire, J. and Uberti, O. (2014) London: The Information Capital, London: Particular Books.
Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011) “Political Polarization on Twitter.” Fifth International AAAI Conference on Weblogs and Social Media. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/index.
Fimyar, O. (2008) “Educational Policy-making in Post-communist Ukraine as an Example of Emerging Governmentality: Discourse Analysis of Curriculum Choice and Assessment Policy Documents” (1999–2003), Journal of Education Policy, 23 (6), pp. 571–594.
Fomina, J. (2014) “Language, Identity, Politics: The Myth of Two Ukraines,” Policy Brief, Institute of Public Affairs/Bertelsmann Foundation. Available: http://www.lse.ac.uk/IDEAS/publications/reports/pdf/SR019/SR019-Stewart.pdf [Accessed on September 28, 2015.]
Fournier, A. (2002) “Mapping Identities: Russian Resistance to Linguistic Ukrainization in Central and Eastern Ukraine,” *Europe-Asia Studies*, 54 (3), pp. 415–433.

Janmaat, J. (1999) “Language Politics in Education and the Response of the Russians in Ukraine,” *Nationalities Papers*, 27 (3), pp. 475–501.

Kiryukhin, D. (2015) “Roots and Features of Modern Ukrainian National Identity and Nationalism,” in *Ukraine and Russia: People, Politics, Propaganda and Perspectives*, Pikulicka-Wilczewska, A. and Sakwa, R. (eds.), Bristol: E-International Relations, pp. 59–69.

Kulyk, V. (2011) “Language Identity, Linguistic Diversity, and Political Cleavages: Evidence from Ukraine,” *Nations and Nationalism*, 17 (3), pp. 627–648.

Kymlicka, W. and Patten, A. (2003) *Language Rights and Political Theory*. Oxford: Oxford University Press.

“Kuzio, T. (1997) *Ukraine under Kuchma: Political Reform, Economic Transformation and Security Policy in Independent Ukraine*; London and New York: McMillan and St. Martin’s Press.

Kuzio, T. (2000) “Nationalism in Ukraine: Towards a New Framework,” *Politics*, 20 (2), pp. 133–162.

Kuzio, T. (2001) “Identity and Nation-building in Ukraine: Defining the ‘Other,’” *Ethnicities*, 1 (3), pp. 343–366.

Kuzio, T. (2002) *Ukraine: State and Nation Building*. Routledge Studies of Societies in Transition, 2nd edition. New York: Routledge.

Kuzio, T. (2006) “National Identity and History Writing in Ukraine,” *Nationalities Papers*, 34 (4), pp. 407–427.

Laitin, D. (1998) *Identity in Formation: Russian-Speaking Populations in the Near Abroad*, Ithaca: Cornell University Press.

“Language law comes into force in Ukraine” *KyivPost*, (August 10, 2012), http://www.kyivpost.com/content/ukraine/language-law-comes-into-force-in-ukraine-311340.html [Accessed on June 14, 2015]

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., and Vespignani, A. (2013) “The Twitter of Babel: Mapping World Languages through Microblogging Platforms,” *PLoS ONE* 8 (4): e61981. doi:10.1371/journal.pone.0061981.

Moser, M. (2015) “The Languages of Ukraine and the War for Russian ‘Compatriots,’” *The Institute for Human Sciences, Tr@nsit online*. Available: http://www.iwm.at/read-listen-watch/transit-online/the-languages-of-ukraine [Accessed on September 28, 2015]

Olszanski, T. (2012) *The Language Issue in Ukraine: An Attempt at a New Perspective*, Warsaw: Centre for Eastern Studies.

Pavliy, B. (2014) “The Abolition of the 2012 Language Law in Ukraine: Was It That Urgent?” *Bulletin of Toyama University of International Studies Faculty of Contemporary Society*, 6, pp. 207–216.

Pearce, W., Holmberg, K., Hellsten, I., and Nerlich, B. (2014) “Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report,”
BORDERS OF IDENTITY AND ACTUAL LANGUAGE USE IN UKRAINE: AN ANALYSIS OF GEOTAGGED TWEETS

PLoS ONE 9 (4): e94785. doi:10.1371/journal.pone.0094785.

Polese, A. (2010) “The Formal and the Informal: Exploring Ukrainian Education in Ukraine: Scenes from Odessa,” Comparative Education, 46 (1), pp. 47–62.

Polese, A. (2011) “Language and Identity in Ukraine: Was it Really Nation Building?” Studies of Transition States and Societies, Issue 3.3, pp. 36–50.

R&B group. The results of the national research on the use of Ukrainian and Russian languages in Ukraine. Результаты национального исследования "Практика использования украинского и русского языков в Украине." http://rb.com.ua/rus/projects/omnibus/5078/ [Accessed on June 30, 2015]

Shulman, S. (1999) “The Cultural Foundations of Ukrainian National Identity,” Ethnic and Racial Studies, 22 (6), pp. 1011–1036.

Statistics Committee of Ukraine (2001) All-Ukrainian population census 2001. Всеукраїнський перепис населення 2001|English version|Results|General results of the census: Language, http://2001.ukrcensus.gov.ua/eng/results/general/language/ [Accessed on June 30, 2015].

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., and Takayasu, H. (2015) “Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study,” PLoS ONE 10 (4): e0121443. doi:10.1371/journal.pone.0121443.

Takeichi, Y., Sasahara, K., Suzuki, R., and Arita, T. (2014) “Twitter as Social Sensor: Dynamics and Structure in Major Sporting Events,” ALIFE 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems. doi:10.7551/978-0-262-32621-6-ch126.

Wilson, A. (1995) “The Donbas between Ukraine and Russia: The Use of History in Political Disputes,” Journal of Contemporary History, 30 (2), pp. 265–289.

Wolczuk, K. (2000) “History, Europe and the ‘National Idea’: the ‘Official’ Narrative of National Identity in Ukraine,” Nationalities Papers, 28 (4), pp. 672–694.

Zeitzoff, T., Kelly, J., and Lotan, G. (2015) “Using Social Media to Measure Foreign Policy Dynamics: An Empirical Analysis of the Iranian–Israeli Confrontation (2012–13),” Journal of Peace Research, 52 (3), pp. 368–383.

Zhurzhenko, T. (2002) “‘Language Politics’ in Contemporary Ukraine: Nationalism and identity formation,” in Questionable Returns, Andrew Bove (ed.), IWM Junior Fellows Conferences, Vol. XII (2) [Online]. Available: http://www.iwm.at/wp-content/uploads/jc-12-02.pdf [Accessed on January 26, 2016]

Zhurzhenko, T. (2010) Borderlands into Bordered Lands: Geopolitics of Identity in Post-Soviet Ukraine, Stuttgart: Ibidem Verlag.