LANGUAGE CHOICE AND SOCIAL MEDIA IN UKRAINE

Bogdan Pavliy

Summary

In this thesis I use the data from the microblogging service Twitter and social networking service Facebook to analyze the language preferences of online social media users in Ukraine. This research describes the current situation with the linguistic choices of Twitter and Facebook users and discusses characteristics of the actual language use in Ukrainian social media from the perspective of gender, age and geography of users.

In Chapter 1, the introduction of the research topic is given. I describe the background of the study, the research aims and research questions. The main aims of the research are: to identify the bilingual users sending geotagged tweets in Ukraine, consider the relation between the language preferences of the users of social media in Ukraine and their location, gender, and age; and finally to find out what connections between users exist in Ukrainian online networks, and how can they influence users' language choice, or, vice versa, can be influenced by it. In my research I aim to answer the following questions: 1. What languages are preferred for online communication in Ukraine? 2. What is the geography of the users? What linguistic preferences can be found on the regional level, based on the data from Twitter and Facebook? 3. To what extent do patterns of language use reflect the country's internal political and linguistic borders as expressed in election and census results? 4. Which demographics are represented among the Twitter users excessively? 5. How to identify users, their gender and age? 6. If it is possible to identify age and/or gender of users, and there are some

bilingual users, who use both languages for online communication, which language is prioritized depending on age and/are gender and why? Then I describe what contributions my thesis aims to make to scholarship. I also argue on the significance of my study and provide the organisation of this thesis. Although my thesis describes and discusses the linguistic preferences of Ukrainians only in online social networks, and does not deal with their language behavior in offline networks, nowadays online interaction comprises a significant part of actual daily interaction in the life of Ukrainians and, as I can prognosticate, it will expand in near future and will encompass even more significant part than offline.

In Chapter 2 I give a review on the research on bilingualism and language use in social online communication networks. After the review on the research on bilingualism in general, which is given in the first part of this chapter, I describe the research on bilingualism in Ukraine and bilingualism in social media. In this part I also discuss how my thesis is related to the research on "everyday nationalism" and language identity of Ukrainian people. In the second part of Chapter 2 I discuss who is using social media and for what. The questions I deal with here are: how to identify users? What research on social media and gender and what research on social media use in Ukraine has been conducted up to date?

In Chapter 3 I describe methodology and limitations of my research. This chapter begins with the description of the data collection from Twitter and Facebook and possible ways of cleaning the data. The technical aspects of data collection and processing are described and suggestions for successful data cleaning are provided. Then I provide comparative analysis of Twitter's and Google's language detection systems, which will be used in my research for language identification of tweets and for language identification of updates and comments on Facebook webpages. It includes the description of method of analysis, the process of cleaning and the results of cleaning. The results at the first stage of language identification by Twitter and Google showed that Twitter's performance, especially in Russian language recognition, is generally better that Google's and without cleaning the difference is immense. At this stage of the research, it can be concluded that even after cleaning data for the Google algorithm Twitter still seems to be more accurate than Google in recognizing Ukrainian and Russian languages in tweets. However, the difference is not significant, and it is likely that after proper cleaning of the content of the tweets, Google's recognition could be improved more, possibly even to that extent when it surpasses Twitter's. The further improvement can be attained through the process of more accurately cleaning the tweets tagged by Google as NONE (or unidentified). The conclusion for this stage of my research is that both Twitter's and Google's language detection systems can be acceptably accurate in Ukrainian and Russian language recognition and may be used further in my research on use of Ukrainian and Russian in online social networks. In this chapter I also discuss deficiencies and limitations of my research among which most serious ones are: so called "Internet bias" (differential access to the Internet which results in distortion in sample composition), self-selection bias and lack of demographic details of those commenting on the Facebook governmental pages.

In Chapter 4 I provide an account of the initial stage of my research on language use on Twitter in Ukraine. In the first sections of the Chapter 4 I describe the results of census and polls in Ukraine and electoral sympathies of voters. Later I compare the results of offline data with my data taken online and come to conclusion that there is a discrepancy between online language behavior and the results of census concerning mother tongue. In this chapter I discuss my findings on language behavior of Twitter users in Ukraine, including data on users in each oblast and their language behaviour depending on region. Last sections of this chapter provide my findings on users' bilingual behaviour followed by my conclusions at this initial stage of the research. My findings suggest that in relation to actual language use, the borderline between stronger and more moderate use of Russian language 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. However, when I compare the results of my analysis of Twitter traffic to election and census results, it is vital 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 my counts of tweets and users include anyone who happens to be in a given location, rather than only local residents. I also found a stark difference in language behavior 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.

To support my findings at the first stage of the research I considered investigating the language use in comments and updates on governmental sites on Ukrainian Facebook. Chapter 5 describes this process. In the beginning I clarify tasks and questions of the research on Facebook. Then I discuss the details of data collection and language detection. Finally, I consider the results of the research and provide my conclusions. My analysis suggested that local governments in some regions ignore the status of Ukrainian as the only language for official use and adapt their language use to that of their citizens. The use of the two languages in page updates by local governments tends to reflect the language use of citizens in their areas as reported in the 2001 census. My findings also show that language use in comments on the pages tends to reflect regional statistics on language use. However the number of pages was insufficient to obtain statistically significant results. As far as bilingualism is concerned, I obtained statistically significant results showing that page visitors tended to comment in the same language as the update, and also they tended to reply to comments in the same language as the comment. In regard to the language change in multiple comments, unfortunately the number of users replying to multiple comments was too small to explore whether individual users switch languages to fit the language used by others. In general, the results of this research support the results of the research on Twitter.

In Chapter 6 I discuss the demographics (age and gender) of Ukrainian Twitter users and their relation to users' language behavior. This chapter consists of two parts: research on age and research on gender of users. The first part - age and language use - gives an account on the research on age of Twitter users in Ukraine which was conducted based on hypothesis that a significant part of Twitter users may be of the high school age. In these sections I describe the process of high school and university examinations in Ukraine, periods of examinations and number of tweets related to the examinations. Then I discuss my findings on age of the Twitter users based on the content of tweets. The second part of this research - gender and language use - gives an account on the research on gender of Twitter users in Ukraine. First, I examined the relationship between two independent variables, location and gender and found that female users outnumber male users by 1.95 to 1. Then I examined the relationship between region and language and found that this relationship is strong. Only in the West do the number of users tweeting more than the national average in Ukrainian outnumber those tweeting less than the national average, while in the South and East fewer than one user in 13 is writing more than the national average of their tweets in Ukrainian. Finally, I examined the relationship between gender and language. I examined the numbers of female and male users tweeting above or below the national average in Ukrainian and found that while female users were 5.96 more likely to be tweeting below the national average in Ukrainian, male users were only 4.38 times more likely to do so.

Chapter 7 deals with the network analysis and provides the information on the main clusters in Ukrainian Twitter network. My primary goal there was to explore the linguistic choices and priorities of main clusters in relation to the gender, age, topics and geographical location of their members. This chapter also provides the identification of bilingual users of both genders. While dealing with clusters of Twitter users who sent geotagged tweets from Ukraine, I found the network of 2,833 users, and counted the numbers of Ukrainian and Russian tweets each of them sent. The bilingual female and bilingual male users writing half or more of their tweets in Ukrainian were then classified as tweeting "mostly in Ukrainian", and the bilingual female and bilingual male users writing more than half of their tweets in Russian were classified as tweeting "mostly in Russian. It has been found that in the case of bilingual users, who tweet in both languages, both genders prioritize Russian in a similar manner.

Conclusions, discussions, implications and topics for future research are given in the Chapter 8. I provide the answers to research questions and discuss the puzzles that I came across in dealing with my data. My findings show considerable differences in male and female social media use in Ukraine. I found that in Ukrainian Twitter among those, who sent geotagged tweets throughout the country, females outnumber males by almost two to one. This tendency controverts the worldwide data on gender of Internet users, where the proportion of female and male users is almost one to one. The second puzzle is the gender imbalance by region. The third puzzle is that female users have a stronger preference for using Russian than male users. It is difficult to clarify the reasons for these puzzles: further research is needed to investigate them. As for the implications and my suggestions from this research, I suggest that the legislators of Ukraine should reconsider their strategy and find solutions on how to endorse the use of Ukrainian in online social media, and protect online networks from dominance of the Russian language. My thesis also suggests a number of avenues for future research among which a more micro-level analysis of language use, including bilingualism in particular oblasts; or a chronological analysis, both of shorter-term seasonal trends in movements within the country and of longer-term developments over a number of years.