LANGUAGE CHOICE AND SOCIAL MEDIA IN UKRAINE

PAVLIY, Bogdan

Doctoral Dissertation
Graduate School of Social Sciences
Hitotsubashi University
SD152006

ウクライナにおけるソーシャルメディアと言語選択

パブリー・ボグダン

一橋大学審査学位論文
博士論文
一橋大学大学院社会学研究科博士後期課程

私は、博士学位請求論文を作成するにあたり、「一橋大学における研究活動に係る行動規範」*および、本研究科の「大学院生研究倫理規範」**を遵守したことを、ここに宣誓します。

＊「一橋大学における研究活動に係る行動規範」（2007年7月4日）

＊＊「一橋大学大学院社会学研究科　大学院生研究倫理規範」（2015年11月11日）

2017年 10月 31日

学位申請者(自署)：Bogdan Pavliy

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Jonathan Lewis, for his great support during the work on this dissertation. I was continually amazed by his patient guidance, encouragement, friendship, empathy and advice he has provided throughout my time as his student. He responded to my queries so promptly and cared so much about my work. It was a great privilege and honor for me to work on my thesis under his guidance.

I would also like to thank my beautiful wife, Lyudmyla, for her continued support and encouragement, and her willingness to check and doublecheck the content of hundreds of tweets. I am very much thankful to my son, Andriy and my daughter, Mariya for their love and understanding during my completion of this research work.

Completing this work would have been all the more difficult without the support of my parents and my brother, Ivan. I am really indebted to them for their constant love, prayers and encouragement.

I also express my gratitude to my colleagues, all those members and staff of Toyama University of International Studies, who were extremely patient and supportive during my work on this dissertation.

And finally, I would like to extend my deepest gratitude to God, for His profound and unconditional love for me in giving me this opportunity to start the research and enabling me to complete it successfully.

Publications

Parts of this thesis (ideas, tables, figures, results and discussions) have appeared previously in the following publications:

Pavliy, B., Lewis, J. (2017) Analyzing Actual Language Use in Ukraine Using Social Media Data: Gender, Location and Language, *Eurasia Cultura*, Vol.3, forthcoming.

Pavliy, B., Lewis, J. (2017).　Issues in Identifying the Age of Twitter Users in Ukraine. 富山国際大学現代社会学部紀要 第9巻, pp.179-190.

Pavliy, B., Lewis, J. (2016). The Use of Ukrainian and Russian on Facebook Pages of Governmental Organizations in Ukraine. 言語文化学会論集第46号, pp.47-61.

Pavily, B. (2016). Language choice and political preference in Ukraine: Can language unite the nation? In T. Ottman, Z. Ritchie, H. Palmer, & D. Warchulski (Eds.), *Peace as a global language: Peace and welfare in the global community*, Bloomington, IN: Iuniverse, pp. 165-179.

Pavliy, B., Lewis, J.(2016). The Performance of Twitter’s Language Detection Algorithm and Google’s Compact Language Detector on Language Detection in Ukrainian and Russian Tweets 富山国際大学現代社会学部紀要 第8巻, pp.99-106.

Pavliy, B., Lewis, J. (2015). Borders of Identity and Actual Language Use in Ukraine: An Analysis of Geotagged Tweets *Japanese Slavic and East European Studies*, Vol.36, pp.77-97.

Pavliy, B. (2015). Language and the cultural border in contemporary Ukraine 富山国際大学現代社会学部紀要 第7巻, pp.115-128.

Pavliy, B. (2014). The abolition of the 2012 language law in Ukraine: was it that urgent? 富山国際大学現代社会学部紀要 第6巻, pp.207-216.

Table of Contents

Chapter 1. Introduction

1.1. Background

Historically and culturally Ukraine has been divided into three regions: Galychyna and other parts of Western Ukraine, Central Ukraine and South-Eastern Ukraine. These regions have traditionally supported different political parties and had different language preferences. Ukrainian was institutionalized as the sole official national language following the country's independence in 1991 but Russian remains the most used language in many regions of the country, especially the South-East. A large proportion of the population is able to speak and write both languages competently; *surzhyk,* a Ukrainian-Russian mixed language which oversteps the Ukrainian-Russian language boundary (Bernsand, 2001:40), is also widely used (Del Gaudio and Tarasenko, 2009:327).

Generally, it is expected that everyone living in Ukraine should know both Ukrainian and Russian, at least passively. As Petro (2015:31) describes it: "Ukraine is, at its heart, bilingual and bicultural". For Ukrainians, the two languages are mutually intelligible, which is shown by the common use of them in the media, education, and governmental institutions, including Ukraine's Parliament.

In Ukraine language use is an important issue on both the state and the personal levels. On the state level, controversial shifts in language policy have taken place since the country's independence in 1991, and Russia's tendency to equate use of Russian with Russian nationality has been one factor undermining its relations with Ukraine as well as with other former Soviet states (Laitin 1998). Meanwhile on the individual level, language use is an important element of the complex, shifting and often ambivalent politics of national identity that have prevailed in Ukraine since independence. In general, as Fox and Miller-Idriss (2008) note, "language and other audible and visual cues trigger an awareness of category membership through everyday interaction" (Fox and Miller-Idriss, 2008:541). And in regard to Ukraine, Pirie (1996:1088) notes that "on the whole, the language divide proved to be a much more reliable predictor of political attitudes than self-declared nationality", while at the same time reminding us that "language usage is an important factor which informs national self-identification, and political attitudes, but it should not be regarded as the Alpha and Omega of national identity in Ukraine - other factors play a significant role in shaping identity as well" (Pirie, 1996:1081). Research by Kuzio (2002) and Bilaniuk (2005) suggests that language choice in Ukraine should be associated with the political choice of an individual rather than being considered a key marker of national identity.

Since independence, 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. "Until the crisis of 2014, national politicians were on the whole reluctant to address the issue beyond electoral pledges, and when they did, as in the case of legislation on language rights at the regional level, they provoked a strong social reaction" (Chaisty and Whitefield, 2017:8-9).

The legal status of the Russian language has been a divisive political issue in Ukraine since 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 pro-Russian 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.

After the Revolution of Dignity in winter 2013-2014, in which more than a hundred protesters and police were killed, and which finally led to Yanukovych's ouster on February 22, 2014, the Ukrainian Parliament made an attempt to abolish the 2012 law "On State Language Policy" and approve a new law with no status for Russian as a regional language. Since politicians on the nationalist side considered themselves winners, they sought to capitalize on their political momentum by changing the law in favor of the Ukrainian language. This attempt to change the language law has had multiple consequences for Ukrainian society. Russian authorities raised their voices against the changing of the law. Although the new bill contained no threat either to the Russian language, nor to Russian-speaking Ukrainians (Blacker, 2014), the situation with the possible implementation of a new bill was used by Russian propaganda as a reason for the annexation of Crimea, followed by support for a separatist movement that resulted in military conflict in the Donetsk and Luhansk regions. In their attempt to split Ukraine, pro-Russian political powers promoted the idea of "two Ukraines", a concept based primarily on the existence of severe language conflict in Ukraine. Fomina (2014) describes the situation as follows:

> *The narrative about two Ukraines is often employed to justify the proposals for the political division of Ukraine, either federalization or a split into two separate political entities, or uniting parts of Ukraine with another state (Russia). However, public opinion is*

predominantly hostile to any such changes, both in the west and in the east. More than half of the population in all the regions - with 53% in the east being the lowest score - are critical of the idea of the federalisation of Ukraine. This goes against the grain of popular perceptions about the widespread desire of eastern Ukrainians to see their region as part of a federation rather than the unitary state of Ukraine.

(Fomina, 2014:13)

In March 2014 Russia broke its obligations to protect Ukrainian territorial integrity, and occupied and annexed the Autonomous Republic of Crimea. After that Russia ignited and supported a separatist movement in the Donbas (Donetsk and Luhansk oblasts) region of Ukraine. To fight the separatists in April 2014 the Ukrainian government started the Anti-Terrorist Operation (ATO), which continues at the time of writing. By the end of August 2014, it had become clear~~,~~ that unless Russia undertook direct military aggression into Ukraine, the separatist states, which formed in Donbas region by that time, would be completely defeated by Ukrainian forces. To intensify its pressure on Ukraine, therefore, Russia started direct military aggression in the Donbas region. Russian plans to annex parts of Ukraine then spread to all regions with a high percentage of Russian native speakers, which Russia considered to be a part of “Russkiy mir” - the Russian World (Jilge, 2014; Laruelle, 2015). However, the reality was different from how it was portrayed by the Russian political elite. For example, during conflicts with the separatists and Russian military forces, Russians realized that many Russian-speaking Ukrainians and even ethnic Russians were fighting on the Ukrainian side against the separatists[1]. Language mattered much less then ideology.

On the other hand, language is an important element of political and social identity – one that "influences perception, thought, and, at least potentially, behavior" (Holmes, 2008:336). That makes language issues a very convenient tool for manipulating people and communities in their political choices. As Kiryukhin (2015) notes:

> *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.*
>
> (Kiryukhin, 2015:63-64)

Since Ukraine gained its independence the language issue was used to provoke confrontation related to national identity and ignite separatism, and we cannot exclude the

[1] See, for example, http://ru.telekritika.ua/kontent/2014-12-01/101031

possibility that language conflict will become once again a cause of political destabilization in Ukraine.

Ukraine's unstable political situation and the danger that language issues may become politicized, and used to justify anti-governmental protests or promote separatism, 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 serious economic, political, and social upheavals. Additionally, the census only ascertained respondents' reported mother tongue and not their actual language use. Given the 2012 change in language policy regarding regional administrations, political crisis (followed by Russia's annexation of Crimea, separatism and military conflict with Russia on the East of Ukraine) and the considerable variation in language use across Ukraine, it is important to monitor how individuals are using language in their everyday communications, and particularly in their interactions with government institutions. Such analysis, particularly if it is carried out over a longer period, can be useful in helping Ukrainian government institutions at all levels become more responsive to their citizens. It can also increase researchers' understanding of bilingualism.

To conduct ~~a~~ research on actual language use in Ukraine, I decided to use data from social media. According to Gemius[2], in 2016 Facebook.com was used by about 45 percent of Internet users in Ukraine, second only to the Russian site Vkontacte.ru. Twitter.com ranked fifth place in the ranking, used by nearly 14 percent of internet users. It should be said, though, that to my understanding of this report, it relates only to the websites, not the apps, so actual use of Twitter and Facebook would probably have been much higher.

In regard to weekly Internet use overall, according to BBG Gallup[3] in 2014 close to half of Ukrainians (46.0%) reported having used social networking services in the past seven days; the figure rose to 89.9% among those age 15 to 24. 74.5% and 66.0% respectively reported having used Vkontakte.ru and Odnoklassniki.ru in the past seven days, the figures for Facebook and Twitter were 42.9% and 21.6% respectively.

This widespread adoption of online social networks 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

[2] https://www.gemius.com/agencies-news/ukraine-through-the-prism-of-social-media-what-are-the-leading-portals-and-who-uses-them.html

[3] https://www.bbg.gov/wp-content/media/2014/06/Ukraine-research-brief.pdf

it is possible to use location information (geotags) to map patterns of language use on a regional basis.

1.2. Research aims

The main aims of the research are:

1. To identify the bilingual users sending geotagged tweets in Ukraine.
2. To identify the bilingual users on governmental sites of Facebook.
3. Consider the relation between the language preferences of the users of social media in Ukraine and their location, gender, and age.
4. To find out what connections between users (clusters of users) exist in Ukrainian online networks, and how can they influence users' language choice, or, vice versa, can be influenced by it.

1.3. Research questions

In this research I will try 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. What are the demographic characteristics of Twitter users in Ukraine?

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? (This question also relates to geography of the users, what region they are probably live in, etc.)

Concerning the first two questions, it is worth mentioning, that I will deal only with those users, whose geographical position (oblast or city) can be identified. Depending on the previous research, these users constitute only the small percent of all online communication network users, and the numbers are small compared to the whole population (Sloan et al.,2013; Sloan and Morgan, 2015), however, I can suggest that even if the numbers are small in my research I will be observing the behavior of more than a tiny group of enthusiasts.

According to the last census, which was held in 2001, the percentage of those whose mother tongue is Ukrainian totals 67.5% of the population of Ukraine (2.8 percentage points more than in 1989) and the percentage of those whose mother tongue is Russian totals 29.6%

of the population. The results of an exit poll by R&B group in 2010 show that the linguistic preferences in three big regions of Ukraine – West, Center and South-East — are quite different: while Russian is highly prioritized in the South-East and Ukrainian is prioritized in the West, the Center tends to be more bilingual. At the same time, the respondents' individual assessment 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%) (R&B poll, 2010).

While providing the comparative analysis of my findings with the above data, I do not intend to either support or refute this data. My thesis presents an attempt to describe and consider the language preferences of Ukrainian population in the context of the actual language use in Ukrainian online social media, which cannot be applied for describing the situation in the offline social networks in Ukraine.

1.4. Contributions of this thesis

This thesis aims to make the following contributions to scholarship:

Research on the geography of social media users and actual language use in Ukraine

Ongoing controversy regarding Ukraine's language laws has highlighted the need for empirical research on language use in the country. Although there is a lot of research on the use of social networks in political processes in Ukraine during the Revolution of Dignity in 2013-2014 (Bohdanova, 2014; Galushko and Zorba, 2014; Kuksenok, 2014; Onuch, 2014, 2015, 2015; Ronzhyn, 2014, 2016; Yasna, 2015; Chalupa, 2015; Gorchinskaya, 2015; Luxmoore, 2015), little attention has been paid to language; the only exception (to my knowledge) is Kuksenok (2015)'s research on multilingualism on social media in the Euromaidan movement. However, Kuksenok's dataset consists only of tweets associated with #EuroMaidan hashtags, and because she did not use geotagged tweets we do not know if the users tweeted from Ukraine, Russia or some other country. In contrast my research uses a dataset of geotagged tweets sent from the whole territory of Ukraine irrespective of content, offering a fuller picture of language use by Ukrainian social media users, whether or not they are politically engaged.

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. My findings show that there is a discrepancy between the language choices of Twitter users and the census data regarding mother tongue. In general Russian is used far more often in tweets than Ukrainian (more than six Russian tweets are sent for every Ukrainian one). In regard to the electoral border that has obtained in national elections since 2004, based on my analysis of tweets I identify a language border veering slightly to the west of the electoral border. My research also shows that rates of bilingual communication are highest in the oblasts containing the country's four largest cities (Kyiv City, Dnipro, Kharkiv and Odesa). The results of my Facebook analysis support my Twitter findings, showing the same tendency to prioritize Russian in the East and South of Ukraine as well as in the large cities.

A new approach to "everyday nationalism"

As a study of the complex role language plays in identity and of actual users' language choice in online communities, this thesis represents a new, quantitative approach to the hitherto almost exclusively qualitative, ethnographic research on "everyday nationalism", which is critical of the simplistic, mutually exclusive ethnic and other categories imposed from the top down in e.g. censuses. This research can thus provide a new perspective from which to understand the findings of Pirie (1996) regarding national identity in Southern and Eastern Ukraine, of Kulyk (2011) on language identity, linguistic diversity and political cleavages in different regions of Ukraine, and of Knott (2015) on the complexities of identity in kin-majorities (Russians in Crimea and Romanians in Moldova) and the role of ethnicity in post-Communist societies.

Research on the demographics of Twitter users in Ukraine

Although demographic attributes such as gender, age or nationality may form the basis of clusters of online communication, and are basic factors that must be considered in any sociological research project, with many social media platforms including Twitter and to some extent Facebook it is not possible to obtain such demographic information about users directly. I therefore explore ways of estimating users' age and gender, and investigate the relationship between these variables and users' linguistic choices. I develop an algorithm that is able to estimate Twitter users' gender with acceptable accuracy from the vocabulary and grammar of the text of their tweets in Ukrainian and Russian. My results suggest that female Twitter users outnumber male users by two to one throughout Ukraine, although the ratio of female to male users is lower in Kyiv than in other regions. I also observe a nationwide tendency for women to tweet more in Russian than men.

The main objectives here are to find some political, social, religious, or some other content which can be related to the demographics and language of use, and discuss the possible application of such findings in frame of the language policy of Ukrainian government and local authorities.

Government-citizen communication in a bilingual society

My thesis also covers language use on Facebook, and code-switching as a factor of a language choice in the daily communication of Ukrainians. Dealing with the data based on the Facebook updates, comments on updates, and comments on comments I found that local governments adapt their language use to that of their citizens. My findings also show that language use in comments on the pages tends to reflect regional statistics on language use. In general, page visitors tend to comment in the same language as the update, and to reply to comments in the same language as the comment. My data also show a clear trend over the last two years for local governments to post more updates and for users to post more comments, which may encourage other local governments to start Facebook pages.

Research on the accuracy of Twitter's and Google's language detection systems in identifying Ukrainian and Russian languages

In this research I also investigate the language preferences of Ukrainian users of Facebook to support my findings on the language behavior on Twitter. My investigation focuses on the language of updates and comments of residents on city and regional (oblast) Facebook pages, Twitter's language detection algorithm cannot be used for that purpose, and thus, I decided to use another detection tool – Google's Compact Language Detector. As my research deals with automatic language detection of a large amount of tweets, and a large amount of updates and comments on Facebook pages, which cannot be done or checked manually, it is necessary to make sure that the language detection systems, used for this research, identify the language correctly enough. As no research on the peculiarities of Ukrainian and Russian language detection (including performance of accessible language detection algorithms) has been presented yet, in this study I will describe the two of most accessible and popular language recognition systems - Twitter's language detection algorithm and Google's Compact Language Detector, provide their comparative analysis and give the suggestions for improving their performances for the next generation of sociolinguists, who would study the language of online communications between Ukrainians. Along with the comparison of Twitter's and Google's performance on the initial stage, I also consider the main obstacles for the correct recognition by Google and offer the procedure of cleaning as one methods for improving the identification of Ukrainian and Russian content of tweets.

1.5. Significance of the study

In this thesis I use the data from the microblogging service Twitter and the most popular social networking service Facebook to analyze the language preferences of online social media users in Ukraine. I will describe the current situation with the linguistic choices of Twitter and Facebook users and discuss characteristics of the actual language use in Ukrainian social media from the perspective of gender, age and geography of users. My research will be conducted on rather national than local level. I also discuss the advantages and limitations of using social media data to investigate communicative behavior and geography of language use in Ukraine.

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. However, nowadays online interaction comprises a significant part of actual daily interaction in the life of Ukrainians (especially those of young generation) and, as I can prognosticate, it will expand in near future and will encompass even more significant part than offline.

I believe that my research will contribute in various fields of social studies such as sociolinguistics (in particular the research on bilingualism and national identity), social media and online communications, language education and language policy, Ukrainian and Russian studies. This research also deals with the development and improvement of language recognition systems, such as language detection algorithms, and methods on collecting data on gender and age of social network users. I believe that my findings can contribute to the improvement of national language policy in Ukraine, and help Ukrainian scholars, lawmakers, politicians and social activists in developing language laws and language education on both regional and national levels.

1.6. Organisation

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. I will describe the current situation with the linguistic choices of Twitter and Facebook users and discuss characteristics of the actual language use in Ukrainian social media from the perspective of gender, age and geography of users. My research will be conducted on rather national than local level. I also discuss the advantages and limitations of using social media data to investigate communicative behavior and geography of language use in Ukraine.

After the introductory Chapter 1, I will give a review on the research on bilingualism and language use in social online communication networks in Chapter 2. In Chapter 3 I will describe methodology of my research; provide comparative analysis of Twitter's and Google's language detection systems and discuss limitations of research. In Chapter 4 the initial stage of the research on language use on Ukrainian Twitter will be discussed. Chapter 5 describes the language use in comments and updates on governmental sites on Ukrainian Facebook. In Chapter 6 the demographics (age and gender) of Ukrainian Twitter users and their relation to users' language behavior will be discussed. Chapter 7 deals with the network analysis and provides the information on the main clusters in Ukrainian Twitter network. My primary goal there is to explore the linguistic choices and priorities of main clusters in relation to the gender, age, topics and geographical location (region) of their members. Conclusions, implications and topics for future research are given in the Chapter 8.

Chapter 2. Previous research

In my research I am dealing with different fields of study, such as linguistics, social studies, and online communication studies. This chapter discusses previous research on bilingualism and social media in general and also in the specific case of Ukraine.

2.1. Bilingualism

Most researchers on the language situation in Ukraine assume that Ukrainians are bilingual (Arel, 1995, 2002; Kuzio, 1997, 2000, 2001, 2002; Janmaat, 1999; Bilaniuk, 2005; Kulyk, 2006, 2011; Søvik, 2007; Bilaniuk and Melnyk, 2008; Pavlenko, 2008; Zhurzhenko, 2010; Polese, 2011; Fomina, 2014; and many others). The typological closeness of Ukrainian and Russian languages means that even individuals lacking linguistic ability will have few difficulties in acquiring each of the languages in spoken or written form (Pavlenko, 2008:61). Therefore, the language choice of the environment often becomes crucial for the language choice of the individual. Such impact of the environment on the language choice makes it difficult for individuals to identify their native language, which leads to the phenomenon of people claiming to be Ukrainian speakers in their replies to census questions while using mostly Russian in their daily life (Arel, 2002). Ukrainian people practice bilingualism "with no clear boundaries between the respective domains of the Ukrainian and Russian languages. Therefore, Ukrainian may be ambivalently perceived as *both* the language of the country and one of the two languages, hardly the most significant for social interaction" (Kulyk, 2006:309).

2.1.1 Research on bilingualism

As mentioned above, researchers on the language situation in Ukraine assume that Ukrainians are bilingual. However, concerning the terms "bilingualism" and "bilingual", there is no consensus among scholars on the degree to which an individual should be able to operate in the language in order to qualify as a bilingual individual.

Although bilingualism is a fairly common phenomenon in modern society and current estimates are that more than 50 percent of the world population is bilingual (Malmkjær, 2010:51), it is difficult to describe bilingualism (or multilingualism) as the definition of it varies depending on the researcher. One of the best definitions of bilingualism is given by the famous American linguist, Leonard Bloomfield in his book *Language* (1933). He defines bilingualism as follows: "In the extreme case of foreign language learning, the speaker becomes so proficient as to be indistinguishable from the native speakers… In the cases where this perfect foreign language learning is not accompanied by loss of the native language, it results in bilingualism, (the) native-like control of two languages" (Bloomfield, 1933:55-56). This definition of bilingualism as the "native-like control of two languages" was challenged by researchers who arguing that bilinguals can be defined as individuals or groups of people who obtain sufficient communicative skills, with various degrees of proficiency, in order to interact with speakers of one or more languages in a given society (Weinreich, 1953, Macnamara, 1967, Mohanty and Perregaux, 1997, Butler and Hakuta, 2004), or simply as "the ability to use more than one language" (Makey, 1962:52). Moreover, some researchers even assert that we can call a person "bilingual" if he or she is able to "produce complete meaningful utterances in the other language" (Haugen, 1953:7).

As seen from the above, a major difference in the approach to who we can call bilingual is the degree of language competence of the individual: should it be native like (the "maximalist" approach), or just at the level of being able to produce understandable utterances (the "minimalist" approach)? Both approaches have their deficiencies. The maximalist approach describes the ideal bilingual who can speak both languages indistinguishably from two native monolinguals. However, in reality no bilingual person can function like two monolingual individuals, because the person has been influenced by both languages (Cook, 1999:191). In fact, the degree of competence in each language differs from individual to individual and it is impossible to decide who should serve as the model of the ideal representative of a native-speaker in each language (Cook,1991; Malmkjær, 2010). On the other hand, the "minimalist" approach fails to make a clear distinction between those bilinguals who are just able to produce meaningful utterances and those who actively use the language in their daily life (Chin and Wigglesworth, 2007:5-6).

"The simplest definition of a bilingual is a person who has some functional ability in a second language. This may vary from a limited ability in one or more domains, to very strong command of both languages (*which is sometimes called balanced bilingualism*)" (Spolsky 1998:45). Depending on the degree of fluency, we can distinguish between *balanced* bilinguals, who have similar degrees of proficiency in both languages, and *dominant* individuals, whose proficiency in one language is higher than in another (Peal and Lambert, 1962).

But fluency is not the only one dimension in classification of bilinguals. In regard to the social status of language, bilinguals can be either *folk* or *elite* (Fishman, 1977) depending on the social status they gain from their linguistic ability. Depending on the bilingual's interpretation of the linguistic codes and meaning units bilingualism can be discerned as *compound, coordinate* or *subordinate* (Weinreich, 1953). As Weinreich describes it, for coordinate bilingual of English and Russian the meaning of the English word "book" and the Russian word "kniga" (book) is slightly different. However, in the mind of compound bilingual these two words are absolutely equal: "book" = "kniga". In the case of the subordinate bilingualism, one language is dominant, and another language was learned with the help of dominant language. So the words in the subordinate language are interpreted through the words in the dominant language.

Bilingualism also differs depending on age of acquisition: it can be *early,* acquired in childhood or *late*, acquired after puberty (Genesee et al., 1978). Depending on the level of maintenance of languages researchers discern between *additive* bilingualism, when both languages a bilingual knows are maintained, and *subtractive*, where bilingualism is on a transitional stage from dominant use of one language to dominant use of another (Lambert, 1977; Landry and Allard, 1993).

Interaction between bilinguals is marked by *code-switching* and *code-mixing. Code-switching* refers to the "use of various linguistic units (words, phrases, clauses, and sentences) primarily from two participating grammatical systems across sentence within a speech event" (Ritchie and Bhatia, 2004:337). *Code-mixing* refers to the mixing of various linguistic units (morphemes, words, modifiers, phrases, clauses, and sentences) primarily from two participating grammatical systems within a sentence" (ibid). Both code-switching and code-mixing are motivated by socio-psychological factors and are widely spread in the bilingual language communities of Ukraine (Lakhtikova, 2017).

Baker and Prys Jones (1998) in their *Encyclopedia of Bilingualism and Bilingual Education* consider some other difficulties in measuring individual bilingualism, which relate to the difference between degree of language ability or competence and function (or actual use) of those two languages. A person may have the ability to speak both languages~~,~~ but prefer to speak only one. On the contrary, a person may be highly proficient in one language and comparatively poor in another, but has to use both languages regularly (Baker and Prys Jones, 1998:3). The authors also argue that the language proficiency of an individual may vary across the four language skills. For example, in some communities people may use one language mostly for oral communication and another for writing. Or people may understand the spoken or written language well, but have difficulties in speaking or writing it by themselves. Such persons may be said to have passive competence in a second language (ibid.). Based on my personal observation, I can say that these situations are quite common in some communities and regions in Ukraine. The eastern regions of Ukraine along with big industrial cities are marked with passive bilingualism (Lakhtikova, 2017). For example, in Donbas region many people can understand spoken and written Ukrainian, but have difficulties with writing documents, reports, etc. in Ukrainian and no difficulties with writing them in Russian, etc. Romaine (1989) in her research on bilingual children renders six main types of bilingual acquisition in childhood: "One person – one language", "Non-dominant home language", "Non-dominant home language without community support", "Double non-dominant home language without community support", "Non-native parents" and "Mixed languages" (Romaine 1989:166-168). It is difficult to decide which category prevails in Ukraine in general, because of different linguistic preferences on the regional level and sometimes even the community level. However, it is possible to suggest that the prevalent types would be "Mixed languages" (both parents are bilingual; community may also be bilingual and parents code-switch and mix languages), "Non-dominant home language" (parents have different native languages; language of one of the parents is dominant in the community; both parents speak non-dominant language to the child, who is exposed to the dominant language outside home), or, in some regions, or, in some regions, "Non-dominant home language without community support" (parents have the same native language; language of the parents is non-dominant in the community; parents speak their own language to the child). Such types or patterns, again, are affected by the use of surzhyk (mix of Ukrainian and Russian) in many regions and communities, which makes it even more complicated to distinguish between the types of bilingual acquisition in childhood in case of Ukraine.

Baker and Prys Jones (1998) argue that categorizing people as bilinguals can "depend on the purpose of categorization" (Baker and Prys Jones,1998:91) They insist that "making arbitrary cut-off points about who is bilingual or not along too few proficiency dimensions" is precarious and where it is possible "the avoidance of simplistic classification and categorization" should be preferred (ibid).

Another issue is that a bilingual person's competence in a language tends to change over time, as the second language may become dominant if the person lives for a long time in an area where the majority second language undermines the minority first language. If the person moves away from the area where his native language is spoken, or loses contact with those who speak it, the person may lose fluency in it (Baker and Prys Jones, 1998:3). This phenomenon is seen in immigrant families in many parts of the world, where "the second language is prestigious and powerful, used exclusively in education and employment, while the minority language is perceived as low in status and value" (Malmkjær, 2010:74). In that case the learning of the second language can undermine the minority first language. Immigrants experience pressure to use their second (dominant) language, because it is the language of the majority and they may feel embarrassed to use their first language. In Ukraine such a situation can be seen not only with immigrants, but also with the rural population when a person from a Ukrainian-speaking rural area moves to a big city, where Russian is used as "the language of convenience" (Plokhy, 2015:314) in all spheres of society.

It is important to note that while speaking of individual bilingualism we cannot isolate a person from the linguistic ecology in which he or she exists. The same person may feel it appropriate or inappropriate to use a given language depending on the social situation and language environment. As Edwards (2009) describes it: "Speaking a particular language means belonging to a particular speech community and this implies that part of the *social* context in which one's *individual* personality is imbedded, the context which supplies the raw materials for that personality, will be linguistic… In general, an influence of language upon personality may be assumed, if not easily demonstrated, but it will tend to link personalities and operate upon their socially overlapping spheres rather than distinguishing between them or producing idiosyncratic dispositions” (Edwards, 2009:23). People’s speech reflects the types of networks they belong to, so “when the people we mix with regularly belong to a homogeneous group, we will generally speak the way the rest of the group does” (Holmes 2008:195).

2.1.2. Bilingualism in Ukraine

In the Soviet Union, as well as in Tsarist Russia before that, the Russian language was used predominantly throughout the whole territory of Ukraine, except some western regions, most of which were under the influence of Austria-Hungary or Poland. From the late 1960s onwards the Soviet Union adopted a policy of “national-Russian bilingualism” *(natsional’no –russkoye dvuyazychie)* under which populations were to use a non-Russian local language as their first language and Russian as their second language” (Haarman 1998:249). Despite this, in Soviet Ukraine the Russian language became ever more widely used, as for a long time Ukrainian language had “a weak position as a cultural medium” (Skvirskaja 2009:192). Similar to other Soviet republics, in Ukraine Russian was “the driving force par excellence for social advancement” (Haarman 1998:249), and acted as a *lingua franca*. However in the Soviet Union the social roles of Russian were not limited only to the role of *lingua franca*, they became much broader and more diverse. Language was considered an important tool of Soviet ideology. The national language policy of the Communist Party and the Soviet state aimed at providing the most favorable conditions for the spread of the Russian language and the formation on this basis of national and Russian bilingualism. In the last decade of Soviet era the scale of national-Russian bilingualism was constantly expanding. Such bilingualism based on the language of interethnic communication favored the culture of the united Soviet people; the main means to make all nationalities familiar with the interethnic language was school education (Guboglo, 1984). Such bilingualism was intended to bring different nationalities closer to each other and to further the merging of all nationalities in the USSR into one "Soviet socialist nation" (Tsameryan, 1979).

Besides its ideological role, Russian was used as a language for daily communication in all businesses and spheres of life in Ukraine. Among the factors that motivated Ukrainians to master Russian and use it in their daily life were access to better education, professional knowledge and technology, and publications in general. A vast amount of literature – technical and specialized literature in particular – was available only in Russian. Another factor was the desire to be promoted to and accepted in higher levels of society. Proficiency in Russian was valued "as a means of social advance" (Haarman 1998:249), providing a so-called "social lift".

By the end of the Soviet era most Ukrainians along with other non-Russian nationalities in the USSR had adopted Russian without abandoning their native language and were able to speak Russian at different levels of proficiency (Haarman 1998:249-249).

Following the collapse of the Soviet Union national-Russian bilingualism continued to be a characteristic of the communicative and language behavior of non-Russian citizens of the Russian Federation and CIS countries. In this new socio-political environment such bilingualism became an advantage, raising the status of the bilingual person and opening up domains of professional activity which were now beyond the reach of Russian-speaking monolinguals. While in individual linguacultural behaviour of such persons bilingualism or multilingualism became a significant marker of their status, in Russian society in general, it was not encouraged. The Russian language identity (or as Burykin (2006) calls it "Russian language mentality") provides that use of any other language besides Russian in existing sociocultural context developed by Russian is unwelcome because the functional possession of the status of first language by any other language, besides Russian, results in irreversible shifts in mentality and language behaviour of Russian nationals and those who belong to Russian cultural environment. (Burykin, 2006)

What can be said concerning the language use and identity of present-day Ukrainian individuals? Pavlenko and Blackledge (2004) argue that identity is something fluid and dynamic, which can be constructed and reconstructed through verbal interaction. They consider identity as a "fragmented, decentered, and shifting" narrative (Pavlenko and Blackledge, 2004:18), which should be observed as a narrative emerging through language and examined in its linguistic (multilingual) contexts. However, I would rather support Pennycook (2003) in his statement that “it is not that people use language varieties because of who they are, but rather that we perform who we are by (among other things) using varieties of language" (Pennycook, 2003: 528). And in regard to the language identity of Ukrainians, Kulyk (2011) in his research on language identity and political cleavage argues that in Ukraine language choice may be considered as one of the best indicators of a person’s attitude towards political and social issues, and hypothesizes that Ukrainians’ preferences in their actual language use in their daily life are shaped largely by their identities. Kulyk’s findings show that “people are more likely to experience and assess the state’s policy in accordance with their language competence and practice” (Kulyk, 2011:641), and that the language behavior of Ukrainian people often is a predictor of people’s “attitudes and policy preferences with regard to both language use and other socially divisive issues, such as foreign policy and historical memory” (Kulyk, 2011:627). Søvik (2007) claims that:

> *A related argument as to why Russian should not or could not become the second state language in Ukraine is connected to a question of “mentality”. The Russian language as a symbol of solidarity with Russia is claimed to play a role as a political or ideological marker of loyalty to Russia rather than to Ukraine, and hence, an orientation towards Russian values, world views and interests rather than towards the Ukrainian ditto. Therefore, those who speak*

Russian are not to be fully trusted because they are not showing respect to or solidarity with the Ukrainian nation or state.

(Søvik, 2007:117).

Language use in Ukraine is complex and flexible. Ukrainians use the language that they consider to be most appropriate for the situation or more comfortable for personal use. “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 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” (Fomina, 2014:5). 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 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).

The Ukrainian and Russian languages are typologically close, so most individuals have few difficulties speaking and writing both languages (Pavlenko, 2008:61) According to a 2006 survey reported by Trach (2009:320) 93% of respondents were bilingual: 37% of respondents reported using both languages equally, 31% used more Ukrainian and 25% used more Russian. As a result, the language or languages being used in a given context often influences an individual's language choice. This influence of context makes it difficult for individuals to identify their native language and leads to the phenomenon of people claiming to be primarily Ukrainian speakers in their census return, while in fact using mostly Russian in their daily lives (Arel, 2002). Ukrainian people practice bilingualism "with no clear boundaries between the respective domains of the Ukrainian and Russian languages. Therefore, Ukrainian may be ambivalently perceived as *both* the language of the country and one of the two languages, hardly the most significant for social interaction" (Kulyk, 2006:309).

Another relevant phenomenon is *nonereciprocal bilingualism* (Bilaniuk, 2005) or *cooperative nonaccomodation* (Pavlenko, 2008): in conversation, bilingual individuals use their preferred languages "with the expectation of being understood and respected by the other party" (Pavlenko, 2008:62). In the Ukrainian context, this means that most people accept bilingualism as a normal practice of social interaction, and consider language choice a political or social rather than a linguistic matter.

Given that various practices impose the use and consumption of both Ukrainian and Russian languages and cultures as the most normal pattern, the society itself is assumed to be bilingual and bicultural not in the sense that it consists of two relatively homogenous parts, but *rather that every member combines the two elements in his/her identity and behavior. However, this assumption coexists in mass consciousness with the belief that the Ukrainian culture and tradition constitute a core value of society, which is reproduced by many other practices.*

(Kulyk, 2006:109-110)

In regard to the language choice of the individual, according to Zaliznyak (2009), those who tend to speak Ukrainian, or call themselves "Ukrainian speakers" tend to choose pro-European positions (47% for and 17% against) while those who tend to use Russian, or call themselves "Russian speakers", do not welcome European integration (30% for and 59% against). The same attitude is seen towards Ukraine joining NATO (Ukrainian speakers: 61% for and 23% against, Russian speakers 15% for and 52% against). By contrast, Russian speakers see the future of Ukraine in union with Russia (Russian speakers: 55% for and 17% against, Ukrainian speakers: 20% for and 63% against) (Zaliznyak, 2009:149-151).

In Western and Central Ukraine the dominance of Ukrainian in interaction with public services is clear: 92% of respondents in the West and 66% of respondents in Central Ukraine reported that they usually address state officials in Ukrainian, while 70% of respondents in the East and 62% in the South reported using Russian. On the other hand, public servants often show some linguistic tolerance and "are inclined to switch to the language of the client depending on the situation" (Trach, 2009:321). In Western and Central Ukraine the majority of public servants (86% and 53% respectively) tended to reply to clients in Ukrainian, while only 10% of the officials in the East and 25% in South held to the state language when their clients used Russian (ibid).

While dealing with relatively recent data, we should not forget, though, that language choice is neither static nor rigid. Language priorities may quickly change depending on the situation in the living environment of individuality, a family unit, community, city or region. During the 25 years since independence the relative statuses of the Russian and Ukrainian languages have been gradually changing. More citizens consider that Ukrainian should be a requirement for access to power and careers. Masenko (2009:113) notes that: "…there is an almost two-fold reduction in the number of Russian speakers (projected 22.5 per cent against the current 40.5 per cent) who would like their children to speak Russian only, which signifies gradual changes in mass attitude toward Ukrainian language".

It is plausible to suggest that the above-mentioned results may not reflect the recent situation in actual language choice in Ukraine due to a probable decrease in the popularity of Russian after 2014, which might be related to such events as the annexation of Crimea, separatism and Russian military invasion in Donbas along with the propaganda war against Ukraine in almost all Russian media. However, even with the presumable lack of popularity of

Russian, the balance of use of Russian and Ukrainian is not expected to have changed radically, hence, I can suggest that those using social media in Ukrainian in the east of the country are more likely than those in the western parts to also interact online in Russian.

Knott (2015) in her study of kin-identifications of the citizens in kin-related states (Romania-Moldova, Russia-Crimea) adopts a qualitative ethnographic approach and uses grounded theory to derive categories of nationalism from her interview transcripts. Her respondents' descriptions of their language use and their attitudes to language form an important part of her data. While investigating "everyday nationalism", Knott argues that government surveys such as censuses tend to use mutually exclusive categories which do not capture the complex realities of people's lived experiences of identity. In my research this phenomenon may also take place. Thus, people living in urban areas can be expected to have more diverse social contacts than those in rural areas, and, consequently, may tend to use social media in both languages (so called "bilingual Twitter users").

And finally, as it was pointed out that the Internet and the new media it facilitates can be challenging for less used languages (see the following Chapter concerning English/Welsh and Dutch/Frisian) because they strengthen~~s~~ the supremacy of the "dominant" language, I can expect that, as the Internet contains much more information in Russian than Ukrainian, the use of Russian in Ukrainian online social networks will be even more pronounced than in offline social networks.

2.1.3. Bilingualism (language use) in social media

Social networking sites feature significantly in the lives of many young people. Where these young people are bilingual, social networking sites may have an important role to play in terms of language use and in shaping perceptions of languages. Meanwhile, many of the world's languages are declining in terms of use and number of speakers. The findings on the endangered languages or languages of minorities, such as the linguistic choices of bilingual Welsh speakers, or on the use of Frisian in online social media (described in this Chapter) may not seem relevant to research on Ukrainian, the official language of a country with a population of 44 million. However, given that almost all Ukrainians are bilingual and can access and understand without any hindrance Russian content on the Internet, they will be facing the same challenges as Welsh or Frisian speakers: the pervasiveness of a "dominant" language (in our case Russian). As soon as a bilingual individual in Ukraine starts looking for some information online, this person will be immediately thrown into the ocean of Russian Internet resources with which the resources in Ukrainian cannot compete. Hence, many Ukrainian social media users will use Russian in order to gain more followers or to influence more people. Russian is understood by more users and allows Ukrainians (even those who prioritize Ukrainian in their offline interactions) to access more information, or gain access to a wider social network.

One of the most interesting studies on minority language and bilingualism in social media has been done by the CaML Group - a group of Welsh researchers led by Daniel Cunliffe – which has been studying aspects of the relationship between minority languages (especially Welsh) and information technology since 2000. In their quantitative and qualitative study they investigated the use of language in social networking sites by young Welsh speakers, focusing on Facebook (Cunliffe, Morris and Prys, 2013), and Twitter (Jones, Cunliffe and Honeycutt, 2013). The report of the Facebook project concluded that:

> *... the choice of language for a particular message may be influenced by the sender, the intended audience and the message itself. While there were characteristics of Facebook which appeared to influence this choice (e.g. all Friends see status updates), it did not appear that Facebook per se was influencing this choice. It seems plausible to suggest that there might be strong similarities between the choice and use of language on Facebook and oral language choice and use offline. However, there may also be some language behaviours that are specific to written environments...*
>
> (Cunliff, Morris and Prys, 2013:351)

Cunliff, Morris and Prys (2013) considered language choice in social networking sites in the context of offline language behavior and found out that there is a relationship between the language used with friends outside school and the language used on Facebook. Moreover, their research shows that the use of English in offline social networks is more strongly related to the use of English on Facebook than the use of Welsh in offline social networks is related to the use of Welsh on Facebook. However, a small number of students used mainly Welsh on Facebook even though in their offline social networks they used English. This phenomenon, Cunliff et al suggest, may be reflecting differences in the membership of these pupils' online and offline social networks. Surprisingly, almost one third of those who prefer to use more Welsh in their offline social networks use mainly English on Facebook. On the other hand, Cunliff et al also argue that the Internet and the new media it facilitates can prove challenging for "vulnerable" minority languages, and that it is doubtful that an increase in the amount and range of digital content in a minority language will help the survival of that language. They claim that although online communication networks provide opportunities for a minority language (Welsh), they also increase exposure to the majority language (English), and consolidate the dominance of English in Welsh online communities (Cunliffe, Morris and Prys, 2013).

The situation with Twitter users is similar: Jones, Cunliffe and Honeycutt (2013) used an online questionnaire to examine Welsh speakers' use of Twitter and came to the following conclusion:

> *The Welsh-speaking community appears to have responded positively to the Internet and the new media it facilitates…. The results show that Twitter has become a new domain for the production and consumption of the Welsh language, as well as facilitating new connections between members of the Welsh-speaking community. However, while Twitter may provide a new domain for the Welsh language, it is also a new domain for the production and consumption of the English language by Welsh speakers. While the presence of the Welsh language on Twitter should be seen as encouraging, the overall effect of Twitter on the maintenance of the Welsh language remains difficult to determine.*
>
> (Jones, Cunliffe and Honeycutt, 2013:653)

Another relevant report on bilingualism and language use in social networks comes from the Netherlands. Jongbloed-Faber et al. (2016) in their work on bilingualism in Dutch online communities explore the use of Frisian (a minority language spoken in a province in the northwest of the Netherlands) on social media by Frisian teenagers. Although Frisian is mainly used as a spoken language and only 12% of the respondents said that they could use Frisian as a written language, in recent years Frisian contributions have frequently shown up on social media. In this study, more than 2,000 Frisian teenagers aged between 14 and 18 years filled in a questionnaire about their language use, language preferences, language attitudes and language proficiency. Results show that, on social media, Frisian has been mainly used by mother tongue speakers, 87% of whom use it to some extent. The study indicates that the teenagers' peer group, language attitudes and writing proficiency can be considered reliable explanatory factors for the use or non-use of Frisian on social media. Although teenagers do not always keep with the official spelling rules of the Frisian language, they prefer to use it in social media. "Social media thus seem to have introduced Frisian into the written domain for an extended group of people, which is a positive sign of the vitality of the Frisian language" (Jongbloed-Faber et al., 2016:27).

2.2. Research on social media

Social media have become an important place for private and public communication, in Ukraine as elsewhere. Both state and non-state organizations are using services such as Facebook and Twitter to provide information and influence opinion on every topic, and individuals are responding to these institutional communications as well as disseminating their own news and views. Analyzing online communication promises to offer important insights into many aspects of Ukrainian society including language use and language policy. It can offer up-to-date findings, particularly important in a country where the last census was carried out in 2001. However, research on social media comes with a number of limitations and cannot be regarded as a replacement for other information sources or research methods.

2.2.1. Who is using social media and for what? How to identify users?

More than 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 one of the most popular social networks or "popular platforms for interaction, communication and collaboration between friends" (Wilson et al, 2009). There are two main reasons for the popularity of Twitter in social science research: 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 meeting specified criteria.

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).

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.

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).

2.2.2.Research on gender and social media

As I deal in this study with geographical, linguistic and demographical variables, it is necessary to refer to the existing research on the demographic attributes of Twitter users. A growing number of researchers from different parts of the world are investigating the age and gender characteristics of social media users. Past research has shown that gender differences exist in a variety of IT contexts.

One pioneering work was by Venkatesh and Morris (2000), whose study of the behavior of 342 workers being introduced to a new software system showed that compared to women, men's technology usage decisions were more strongly influenced by their perceptions of usefulness, while women were more strongly influenced by perceptions of ease of use and subjective norms.

An important contribution was made by Argamon et al. (2007), who studied blogs written by male and female blogger of age variations from teens to forties. By applying factor analysis and machine learning techniques, the researchers demonstrated consistent patterns of gender–linked variation in writing topics. They found that male bloggers wrote more on such topics as *Religion, Politics, Business,* and *Internet,* while female bloggers' writing more often fell into the categories *Conversation, AtHome, Fun, Romance,* and *Swearing*.

Herdağdelen and Baroni (2011), as part of an effort to provide artificial intelligence systems with "common sense" understanding of humans' stereotypes and expectations regarding gender, extracted gender-specific actions (i.e. whether an action was performed by a man or woman) from text corpora and Twitter, and compared them with human coders' results. They concluded that it is feasible to use natural text and a Twitter-derived corpus in order to augment common sense repositories with stereotypical gender expectations of actions.

Mandel et al. (2012) examined responses to the Hurricane Irene disaster on Twitter by location and gender. Their quantitative and qualitative analysis showed that females were more likely to express concern than males.

Soedjono (2012) used a corpus linguistics approach to Twitter communication, and her results showed gender differences in the use of pronouns, similarities in the use of abbreviations, and differences and similarities between genders in the use of vulgar words.

Bamman, Eisenstein and Schnoebelen (2014) conducted a study of the relationship between gender, linguistic style, and social networks on Twitter. They found a range of styles and interests reflecting the multifaceted interaction between gender and language. They also investigated individuals whose language better matches the other gender and found that such individuals have social networks that include significantly more individuals from the other gender, and that in general, social network homophily is correlated with the use of same gender language markers.

Cunha et al. (2014) studied differences between female and male language use on Twitter, with a particular focus on the hashtag designation process during political debates. Considering political hashtags as strategies of persuasion in Twitter, imperative tags could be understood as more overt ways of persuading and declarative tags as more indirect ones. Analyzing tweets with political content posted during Brazilian presidential campaigns, the researchers found that male Twitter users, when expressing their attitude toward a given candidate, tended to use imperative verbal forms in hashtags, while female users preferred declarative forms. This difference could be interpreted as a sign of distinct approaches in relation to other network members.

Holberg and Hellsten (2014) investigated gender differences among participants in the climate change debate on Twitter. They identified hashtags and usernames that were proportionately more frequently mentioned by either male or female tweeters, and found that female users mentioned significantly more campaigns and organizations with a convinced attitude towards anthropogenic impact on climate change, while male users mentioned significantly more private persons and usernames with a skeptical stance.

Some studies revealed changes in traditional gender roles and behavior. For example, Huffaker and Calvert (2005) investigated online identity and language use among male and female teenage bloggers. They examined the disclosure of personal information, sexual identity, emotive features, and semantic themes. The researchers found that male and female teenagers presented themselves similarly in their blogs in terms of revealing personal information. However, compared to females males used more emoticons, employed an active and resolute style of language, and were more likely to present themselves as gay. The researchers claimed that teenagers stay closer to reality in their online expressions of self than has previously been suggested, and that these explorations involve issues such as learning about their sexuality that commonly occur during the adolescent years.

Harp and Tremayne (2006) examined gender inequity among the most-read political blogs on the Web. Sampling over one year of blog rankings, they found that only 10% of the top bloggers were women. Discourse analysis of bloggers' explanations for gender disparity revealed three dominant traits: women do not blog about politics, women's blogs lack "quality" (as defined by the authors), and top bloggers do not link to women's sites.

In 2013-2015 a team of researchers from Cardiff University developed “techniques for collecting or estimating demographics from Twitter data including analyzing gender, language and location” (Sloan et al., 2013). They developed an algorithm for detecting the age of Twitter users using a set of pattern matching rules and building on previous research. The reliability of the results was demonstrated by expert human validation. The researchers found that the age distribution of Twitter users is much younger than the general UK population as of the 2011 Census, with a peak around ages 16 to 22 accounting for 67.5% of all users. The researchers claimed that more than half of the users (59.4%) belonged to the age group 13 to 20 (Sloan et al.,2015).

Human relationships in online networks

Previous research on the online networks in other countries showed that participants often used the Internet, especially social networking sites, to connect and reconnect with friends and family members and that there was overlap between participants’ online and offline networks. The pattern also suggested that emerging adults may use different online contexts to strengthen different aspects of their offline connections (Subrahmanyam et al., 2008). However, as I do not deal with any offline activities and surveys in this research, the information from my research will be relevant only to the groups of online users, who communicate with each other online, based on shared attributes or interests. Shared attributes might include location, gender, age, language, nationality. Interests include similar topics, hobbies, things to be discussed in tweets, etc.

It is relevant to mention here the findings of Wilson et al. (2009) on interaction activity on Facebook. Their research revealed that the interactive activity of each user is significantly skewed towards a small portion of each user's social links, which casts doubt on the assumption that all social links imply equally meaningful friend relationships. Moreover, the analysis of interaction graphs derived from their Facebook data revealed different characteristics than the corresponding social graph.

2.2.3. Research on social media use in Ukraine

Research on the use of social media such as Twitter or Facebook in Ukraine (Bachmann and Lyubashenko, 2014; Goban-Klas, 2014; Lyubashenko, 2014; Ronzhyn, 2014, 2016) along with the Yandex report on the Twitter usage and trendiest Twitter hashtags in Ukraine in 2013-2014, is mostly focused on political or socioeconomic aspects of use rather than the gender, linguistic or other demographic attributes of users. The geographic and demographic attributes of social media users in Ukraine are considered in the research of Onuch on the characteristics of Euromaidan protestors and ways of mobilizing them through social online communities (Onuch, 2014, 2015, 2015). However, in this type of research, both linguistic and demographic aspects are limited, as those involved in online communications—whether organizers, participants or supporters—are on one side of the barricades in their political preferences, and cannot represent the nation as a whole.

Although I could not find any relevant demographic research on Twitter users in Ukraine on the national scale, previous research suggests that Ukrainians using social networks including Twitter tend to be relatively young. Pentina, Basmanova and Zhang (2014) in their cross-national study of Twitter users' motivations and continuance intentions compare motivations and preferences of Twitter users in Ukraine and the United States. They did not opt for geotagged tweets but provided online surveys for selected Twitter users in both countries. Based on the previous research of Kostenko (2011), Pentina, Basmanova and Zhang (2014) suggest that "while typical Ukrainian Twitter user demographics are not available, the sample characteristics are representative of the Ukrainian Internet user who is characterized by a younger age and higher socioeconomic status that provides access to wireless and mobile communications" (Pentina, Basmanova and Zhang, 2014:41).

The only research focused on language in Ukrainian online communities (to my knowledge) is Kuksenok (2015). In her research on multilingualism on social media in the Euromaidan movement she downloaded and analysed tweets associated with #EuroMaidan and the corresponding Ukrainian and Russian hashtags before and after the Revolution of Dignity and found that on February, 21 2014 there were nearly the same amount of tweets in Ukrainian as in Russian, while on March 4, 2014 there proportion of Ukrainian to Russian was 1 to 10, and Russian tweets continued to increase. On the other hand, only from 10 to 15 percent of tweets were either in Ukrainian or Russian. Kuksenok explains this disproportion in relation to the goals of twitter users, reasoning that one of the goals which emerged after the end of the Revolution of Dignity was to show Russians the real state of things in Euromaidan from the point of view of Ukrainians. Although the language of tweets was the target of Kuksenok's research, because she did not differentiate between geotagged and non-geotagged tweets we do not know if the users tweeted from Ukraine, Russia or some other country.

Chapter 3. Methodology and limitations

My thesis presents the results of a two and a half year exploratory study into linguistic preferences and actual language use in online social networks in Ukraine. The research focuses on the use of Ukrainian and Russian on Twitter and Facebook. As (to my knowledge) there was no existing research on the geography of language use, users' demographics or network segments (clusters) of users in Ukrainian social media, one aim of this study was to gather initial data that would provide both insight and direction for future research.

In this research, I have to deal with a massive amount of data which consists of tweets and updates or comments on Facebook pages of those living or staying on the territory of Ukraine. To identify the language of tweets, I used Twitter's own language detecting algorithm of Twittter. To identify the language of updates and comments on Facebook pages I used Google's Compact Language Detector. Using these automated detection systems I am unable to make a clear distinction between those bilinguals who are only able to produce meaningful utterances and those who are proficient enough to actively use the language in their daily life. Moreover, it is impossible to investigate if a user who sent a tweet or wrote a comment on Facebook is actually competent in the language, or using some translating device to write her message. For that reason, I am bound to cling to the "minimalist" approach in this research and use the term “bilingual” simply to meaning a user who is tweeting or commenting on Facebook in both Ukrainian and Russian.

3.1. Methodology

Due to the lack of any existing research of a similar nature, which could provide with some testable information that would merit my study, I believe that it is justified to apply the same descriptive approach taken by Venkatesh and Morris (2000), Argamon et al. (2007), Herdağdelen and Baroni (2011), Mandel et al. (2012), Soedjono (2012), Bamman, Eisenstein and Schnoebelen (2014) Cunha et al. (2014), Kuksenok (2015), and other researchers of language use in social media around the world. Therefore, my highest priority at this pioneering stage was to establish methods of gathering and classifying data, analyzing its basic characteristics (in this case, location, language, and user ID), and explore the relation of those basic characteristics to other available data (in this case, the territories of administrative regions, users’ gender and age (if available), and topics of their tweets). This can then serve as a foundation for research that adopts particular sociolinguistic and linguistic frameworks and methods.

Although my research uses both quantitative and qualitative methods at the analysis stage, the data collection was done using only quantitative techniques. The quantitative data used in this thesis constitute observations on the language behavior of social network users. As the users could not know that their language behavior was being studied, the danger of deliberate misrepresentation or misreporting is avoided, and overall reliability of the data is good. At the analysis stage it was necessary to use some qualitative methods, which mostly meant the reading and interpretation of the content of tweets and Twitter profiles, and updates and comments on Facebook. This was largely dictated by the lack of basic demographic information about the users.

In order to gain a more precise understanding of language use in Ukrainian online social media, additional qualitative research such as offline observation of user's language behavior would be useful. Such research would have to deal with sampling challenges, the practical difficulties and costs of carrying out such observation, and of course ethical issues such as privacy. It might be possible to work out a strategy to deal with such issues in the future.

3.2. Data collection and cleaning

3.2.1. Twitter

At the first stage of the research geotagged tweets sent from the territory of Ukraine (including Crimea) from the Twitter Streaming API between April 11 and September 15, 2015 were collected. 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, this program received it and stored it in a Postgres database. The program ran almost continually, with a small number of interruptions due to connectivity and server maintenance issues. Tweets sent from areas in the bounding box that were outside the territory of Ukraine were then excluded[5] along with the tweets generated by the location service Foursquare that merely included text information about the user's location.

3.2.2. Facebook

In order to be more confident that my findings from Twitter were not due to some peculiarities of Twitter and/or its users in Ukraine, I decided to also investigate the linguistic choices of Ukrainian users of Facebook. Unlike Twitter, most updates and comments written

[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 the database and imported a shapefile of the Ukrainian national territory and administrative areas downloaded from http://www.gadm.org The geo-coordinates provided by the Twitter API in the format (longitude, latitude) were converted 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 all tweets in the database, 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.

by general users on their Facebook pages are not public by default, and the data on those pages (whether public or not) cannot be retrieved through Facebook's API. However, a certain kind of Facebook page (akin to a public website created within Facebook) is public, and updates and comments to it can be collected through the API. I was able to identify 24 of these Facebook sites run by Ukrainian city and regional governments and succeeded in downloading updates and comments on them using a Python program to access the Facebook API and save the results to a Postgres database.

On these sites government institutions broadcast information to local residents, and residents respond in the form of comments. Some of these pages are updated frequently, and some attract a large number of comments. As the pages are local in character and the updates are written for people living in the city or region, it is highly probable that the comments are written either by residents or by former residents who still have a strong connection to the area. Thus, although I do not have precise information on users' locations in the same way that I do with the geotagged Tweets, I can be relatively confident of the location of the users. Furthermore, as some of the comments attract replies or comments on comments, it enables me to observe communication between citizens in a local context and explore if the language use in comments on updates shows a correlation to reported language use statistics for different

Ukrainian regions. Analyzing communication on these sites has the added benefit of offering insights into language behavior in government-citizen and citizen-citizen communications.

3.3. Comparative analysis of Twitter’s and Google’s language detection systems

The large number of tweets in my dataset makes it impossible to check the language of each manually. Twitter's language detection algorithm provides a language tag for each tweet, but for a study on language behaviour on Twitter I need to check the accuracy of Twitter’s language tags.

I also needed a way to identify the language of the Facebook updates and comments automatically. I used Google’s Compact Language Detector (CLD) for this purpose. Once again, it was necessary to check the accuracy of the CLD in recognizing the Ukrainian and Russian languages.

Moreover, although the language of tweets can be identified by Twitter’s language detection algorithm, if Google’s Compact Language Detector proves to be more accurate than Twitter’s algorithm in detecting and tagging the language of the tweets written in Ukrainian or Russian languages, it can be used for the research on Twitter as well. Thus, there was a need to provide a comparative analysis of Twitter’s and Google’s language detection systems.

To check the level of accuracy of Twitter’s and Google’s language detection systems, I asked native Ukrainian-Russian bilingual speakers to read and identify the language of a sample of 4000 tweets. I then compared their results with those from Twitter's and Google's language detection algorithms.

3.3.1. Method of analysis

I selected at random 2000 tweets recognized by Twitter's language detection algorithm as written in Ukrainian and 2000 tweets recognized as written in Russian. I tagged the tweets using the Google CLD. Then I had the tweets coded for language by bilingual native speakers of Ukrainian and Russian languages and compared the results, which are summarized in Table 1.

Tweets language detected as UK by Twitter	Tweets language detected as RU by Twitter	Both Twitter and native speaker detected as UK	Both Twitter and native speaker detected as RU	Both Google and native speaker detected as UK	Both Google and native speaker detected as RU
2000	0	1568	0	1272	0
0	2000	0	1846	0	1337

Table 1. Tweets detected as Ukrainian and Russian by detection algorithms and manually

As we can see, the performance of Twitter's language detection algorithm is considerably better than Google's. In case of Ukrainian language detection, the correctness of Twitter is 78%, while the initial level of Google's correctness is 64%. In case of Russian language detection, the performance of Twitter is even better: 92%, while the initial level of Google's correctness is 67%.

Clearly, Twitter's algorithm achieves a higher level of accuracy than Google's, especially in identifying tweets written in Russian. However, I investigated the variations in language tags and found that Google's Compact Language Detector often identifies the language of a tweet as NONE. Hence it would not be correct to assert that Twitter's performance on language recognition always surpasses Google's. In my sample of 4000 tweets Google detected as NONE 1510 tweets; of these 807 tweets were tagged by Twitter as Ukrainian and 703 tweets as Russian. As the number of such tweets exceeded one third of all tweets in the sample, I decided to analyze their content and find ways to improve the language detection in case of using Google's algorithm.

3.3.2. Tweets tagged by Google as NONE

Having analyzed the content of tweets tagged by Google as NONE, I came to the following conclusions:

1. In general, most of the tweets in the NONE category contained very short messages or utterances where even native speakers sometimes had difficulty understanding the meaning of the tweets.

2. Among the tweets in the NONE category, there were tweets written in *surzhyk* (a

mixture of both Ukrainian and Russian), and in some cases language identification was problematic for native speakers of both languages.

3. The similarity of Ukrainian and Russian expressions is a major problem in detecting the language for native speakers. As some expressions are identical in both languages the expressions alone cannot be detected as either Ukrainian or Russian, so even the native speakers decided to mark them as NONE. (e.g. “Христос Воскрес!” = Jesus Has Risen! (identified by native speakers as NONE))

4. Use of emoticons, hashtags, abbreviations (e.g. Шалено 🙆🙆🙆🕺🔥🔥✨🚀🚀🚀😻😻😻😻 #бумбокс = Crazy 🙆🙆🙆🕺🔥🔥✨🚀🚀🚀😻😻😻😻 #boombox (identified by native speakers as Ukrainian))

5. Expressions with no meaning, mimicking sounds (e.g. Бам Бам Бам Бам = Bum Bum Bum Bum (identified by native speakers as NONE))

6. Mixing languages (code-mixing) by using English words inside of Ukrainian or Russian phrase (e.g. “Де твій Online коли ти так потрібна” = Where is your Online when I need you so (identified by native speakers as Ukrainian))

7. Writing English expressions in Cyrillic alphabet (e.g. “май фейфоріт піца” = My favorite pizza (identified by native speakers as NONE)).

8. Use of slang, spelling mistakes, ungrammatical writing or compressed writing

(“мала,з др” = Happy Birthday, baby (identified by native speakers as Ukrainian))

9. Mixture of Latin and Cyrillic letters in one word (e.g. “Sportик” (identified by native speaker as NONE).

10. Repetition of some letter(s) in emotional expressions (e.g. “ТИ СЕРЙООЗНОО” = ARE YOU SEERIOOOUS (identified by native speakers as Ukrainian)).

It would be plausible to suggest that the above problems caused more than one third of the tweets from my batch of 4000 tweets to be recognized as NONE by Google’s algorithm. Consequently, I decided to perform cleaning of the tweets content and discuss how the accuracy of the algorithm, based on the results of language detection by native speakers, could be improved.

3.3.3. Process of cleaning and the results of cleaning

To perform the cleaning of the 4000 tweets, I used a Python script which removed URLs, @usernames and #hashtags from the text of tweets before running the Google language detection algorithm.

After careful selection and analyses of the results (both positive and negative) of cleaning, I found that running the cleaning script had five effects for tweets identified by Twitter as either Ukrainian or Russian and by Google as NONE (see Table 2 for details):

1. *Changed incorrectly*

The new language tag does not match native speakers' detection.

2. *Changed to Ukrainian correctly*

The new language tag matches native speakers' language detection as Ukrainian.

3. *Changed to Russian correctly*

The new language tag matches native speakers' language detection as Russian.

4. *Erroneous change from NONE to some language*

Google's initial identification of the tweet's language was correct and matched native speakers' detection as NONE, but after cleaning Google erroneously identified the language as UK, RU or some other language.

5. *An improvement of the recognition of tweets written in the Belarusian language.*

Three tweets initially tagged as NONE were correctly recognized as Belarusian after cleaning. However, as I target only Russian and Ukrainian languages in this study, this will not be discussed here.

Type of change Detected by Twitter	**1) Changed incorrectly** (do not match native speakers' detection)	**2) Changed to Ukrainian correctly** (match native speakers' language detection as Ukrainian)	**3) Changed to Russian correctly** (match native speakers' language detection as Russian)	**4) Erroneous change from NONE** (matched native speakers' NONE detection but was erroneously given some language tag after cleaning)	**TOTAL**
as Ukrainian	**59**	**204**	**42**	**14**	**319**
as Russian	**9**	**1**	**258**	**8**	**277**
Total	**68**	**205**	**300**	**22**	**596**

Table 2. Types of change caused by cleaning in tweets identified by Google as NONE

From this data I can conclude that, based on native speakers' recognition of the language of each tweet, I got better results in language identification by Google after cleaning. From my sample of 4000 tweets, of the 1510 tweets that were initially tagged as NONE, after cleaning recognition improved for 507 tweets (205 Ukrainian, 300 Russian, 3 Belarusian), while negative changes caused by recognition errors due to cleaning happened only for 22 tweets.

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. This is probably to be expected:

Twitter has its algorithm optimized for shorter texts and probably also ignores URLs, hashtags and usernames when identifying the language of each tweet. 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.

3.4. Limitations

At the end of the day, the opportunities offered by the Internet for developing social science research must be taken for what they are, namely technical possibilities which increase the researcher's control over some aspects and reduce it over others.

(Frippiat, Marquis and Wiles-Portier, 2010:307)

It is important to note that using social media such as Twitter or Facebook does 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, Kelly and Lotan (2015). The latter outlined a set of dangers which every researcher of Twitter would be exposed to when dealing with Twitter data:

(1) Analysis of message content without regard to network structure,

(2) Using social media as a perfect substitute for traditional public opinion

(3) Having a simple conception of the Internet as comprising a 'global conversation', without sufficient attention to global vs. local contexts, and the relationship between languages and cross-national information flow

(Zeitzoff, Kelly and Lotan, 2015:380).

Most critically, social media users are a self-selecting sample and not a random sample of the population. As Boyd and Crawford (2012) pointed out: "Twitter does not represent 'all people' and it is an error to assume 'people' and 'Twitter users' are synonymous: they are a very particular sub-set. Neither is the population using Twitter representative of the global population. Nor can we assume that accounts and users are equivalent" (Boyd and Crawford, 2012:669). Moreover, even manual reading of users' Twitter profiles cannot give us full assurance that people are who they claim to be or that their linguistic behavior online is similar to their offline behavior.

Furthermore, we lack reliable initial demographic information about Twitter users such as age or gender that could be used to correct sample bias. Thus, however many tweets are collected, one cannot draw conclusions regarding the population as a whole. For this reason I cannot idealize the findings of this research, especially in the sections dealing with the comparison of actual language use with the results of the national census. Zeitzoff, Kelly and Lotan (2015) remark that "caution should be exercised when attempting to draw a direct relationship to social media and public opinion. Social media, and the Internet more broadly, represent a field of communicative engagement among diverse sets of actors, only some of which are subsets of 'the public'" (Zeitzoff, Kelly and Lotan, 2015:380).

Self-selecting bias, which seriously impacts the effectiveness of online surveys was described and discussed by Couper (2000), Bosnjak, Tuten and Bandilla (2001), Börsch-Supan et al. (2004), Bethlehem and Stoop (2007), Bethlehem (2010), Frippiat, Marquis and Wiles-Portier (2010), Das, Ester and Kaczmirek (2011), Khazaal et al. (2014) and many other researchers. One of the main factors of misrepresentation in online surveys is differential access to the Internet which results in distortion in sample composition, so called "Internet bias", which misleads the researcher in interpretation of the data. Another factor, so-called "self-selection", means that individuals knowingly or unknowingly select themselves for the study, in other words that they are willing to participate in activity. The same applies to research on Twitter or Facebook: the internet researcher is not in control over the selection process, and depends on subjects' willingness to participate in the research.

Self-selection means that principles of probability sampling cannot be followed. Bethlehem (2010) stresses that "by selecting a random sample, probability theory can be applied, making it possible to construct unbiased estimates", while "many web surveys rely on self-selection of respondents instead of probability sampling…. The theory of probability sampling cannot be applied and estimates are often substantially biased" (Bethlehem, 2010:162). Even concerning Twitter users or Facebook users we cannot neglect the fact that "self-selection into an online sample is the joint effect of two factors: internet access and willingness to participate" (Börsch-Supan et al., 2004). In the case of Twitter, for the purposes of this research willingness to participate is the willingness of user to switch the geolocation of his or her device on and then willingness to actually tweet or respond to tweets. In the case of Facebook it is the desire of a user to visit a local government Facebook page, read some update or a comment on the update and comment on it.

However, 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, argue 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, 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.

In regard to research on Facebook, it has even more limitations compared to the research on Twitter. Along with above-mentioned self-selection bias, and lack of demographic details of those commenting on the Facebook governmental pages, another limitation exists: I can hardly assume if people's language use in commenting on public Facebook pages is the same as it is in face-to-face interactions with government officials (e.g. at the city office) or they prefer the certain language only for interaction on Facebook. Linguistically, interactions on social media have elements of both written communication, which tends to be more formal, and spoken communication, which tends to be informal. Furthermore, online communication may be between individuals who know each other in the "real world" and may or not have met offline, or under conditions of explicit or de facto anonymity where users have no clear idea of the gender, age, ethnicity or other attributes of their interlocutors.

Another issue is that the relatively limited number of Facebook pages and the low levels of replies to comments prevent me from drawing statistically significant conclusions regarding several topics of interest.

Chapter 4. Language use on Twitter in Ukraine

In this chapter, I consider the actual language use of individuals in Ukrainian Twitter compared to the results of the 2001 census and to electoral preferences in Ukraine during 2004-2012.

4.1. Results of census and polls

The results of the last census, which was held in 2001, show that the percentage of those whose mother tongue is Ukrainian totals 67.5% of the population of Ukraine (2.8 percentage points more than in 1989) and the percentage of those whose mother tongue is Russian totals 29.6% of the population. This division by mother tongue can be seen from Table 3.

Reported as mother tongue:	*Ukrainians*	*Russians*	*Other nationalities*	**Total**
Ukrainian	85.2%	3.9%	11.8%	**67.5%**
Russian	14.8%	95.9%	31.1%	**29.6%**
Other language	0.0%	0.2%	57.1%	**2.9%**

Table 3. Mother tongues of Ukrainian citizens by the results of 2001 year Census

Source: 2001 census, http://2001.ukrcensus.gov.ua/eng/results/general/language/

Consolidated data of four polls carried out by KIIS (Kyiv International Institute of Sociology) during two years before the census practically coincide with the results of the census concerning native language. However, KIIS polls gave respondents the option to name both Ukrainian and Russian as their mother tongues. Using such an approach, while the Russian language is the sole native language for 29.6% people (by data from census; by data from poll about 30.4 ± 1.2%), when we take into account those who call Russian their mother tongue simultaneously with Ukrainian, we found that Russian is native for 42,8 ± 1,4 % of the adult population of Ukraine.

Reported as mother tongue(s) :	*Ukrainians*	*Russians*	*Other nationalities*	**Total**
Ukrainian	71.3%	2.8%	7.8%	**54.4%**
Russian	14.9%	87.6%	36.6%	**30.4%**
Both Russian and Ukrainian	13.6%	9.5%	6.3%	**12.4%**
Other language	0.2%	0.1%	49.3%	**2.8%**

Table 4. Mother tongues of two main national groups by the results of KIIS poll

Source: Consolidated data of four polls (5226 respondents in total) held by KIIS for two years before the census by direct interviews to representatives of adult people of Ukraine (18 years and older)(in Khmelko, 2004):

http://www.kiis.com.ua/materials/articles_HVE/16_linguaethnical.pdf

An advantage of having no option to choose both Ukrainian and Russian simultaneously as mother tongues is that it forces the respondent to choose just the language which has higher value for her. The disadvantage is that the term “mother tongue” is ambiguous and its practical meaning differs depending on the respondent’s understanding (Khmelko, 2004).

4.2. Electoral sympathies of voters

The results of national elections held in the country since 2004 show a consistent geographical divide between southern and eastern Ukraine on the 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.[6] Figure 1 illustrates this, using the results of the 2012 Parliamentary election; however, the border is identical for the other three elections.

[6] Official election results can be found at http://www.cvk.gov.ua/pls/vp2004/wp0011e.

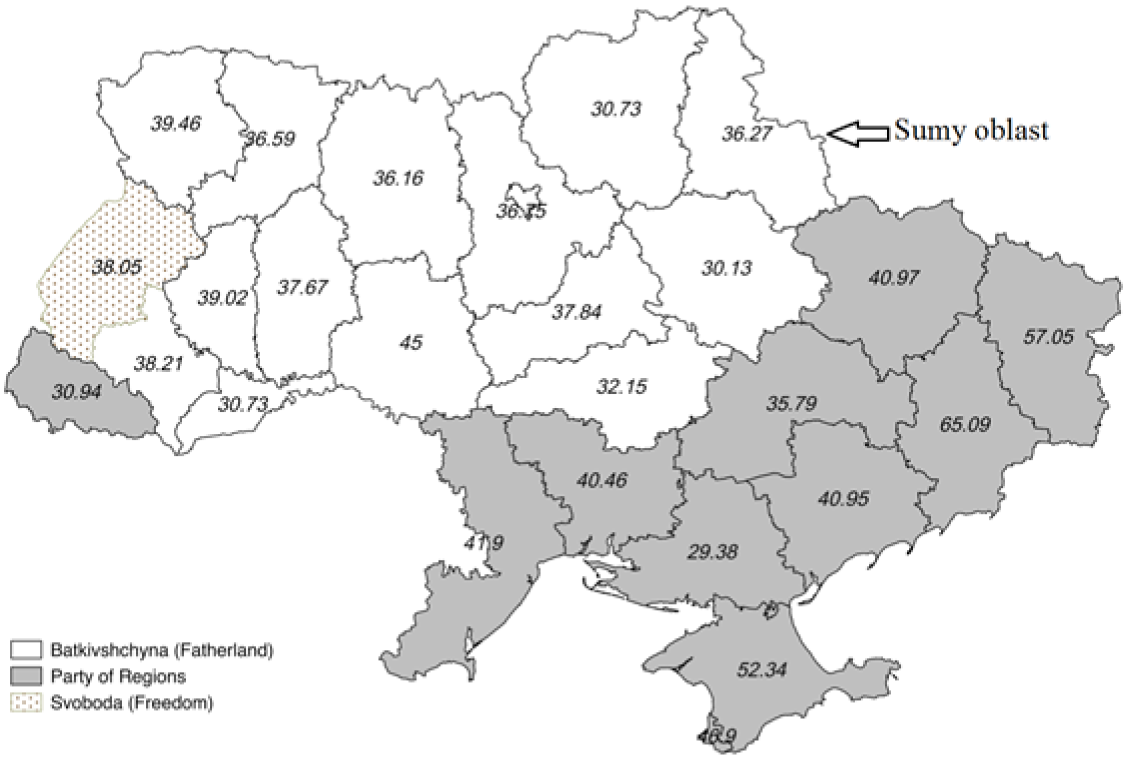

Figure 1. Results of 2012 Parliamentary Elections

Source: http://www.cvk.gov.ua/pls/vnd2012/wp001E

According to the 2001 census, Russian was the majority native language in three oblasts: Luhansk, Donetsk, and Crimea.[7] Hence, as Figure 2 shows, two borders can be delineated within Ukraine: a linguistic divide, in terms of levels of the reported mother tongue in the far east, and an electoral divide farther west.

[7] An exit poll conducted by the R&B 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%).

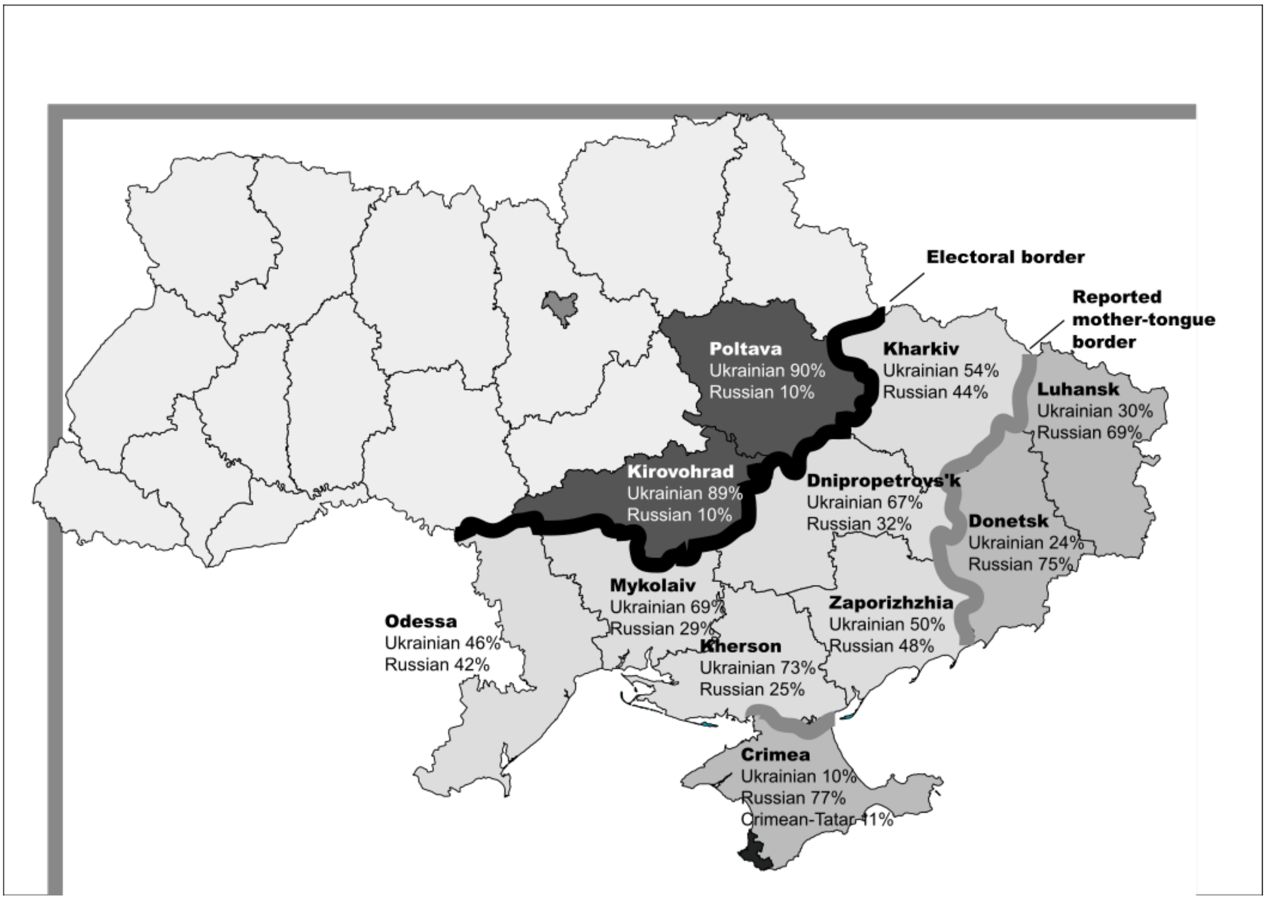

Figure 2. Border by electoral choice and border by native language.

The numbers are the percentages of Ukrainian and Russian native speakers reported in the 2001 census.

Source: http://2001.ukrcensus.gov.ua/eng/results/general/language

In the following section I will address the following two research questions: What regional linguistic preferences can be found, based on the data from Twitter? And to what extent do patterns of language use reflect the country's internal political and linguistic borders as expressed in election and census results?

4.3. Language behavior of Twitter users in Ukraine

As described in the previous Chapter, I collected a total of 2,409,608 geotagged tweets sent from the territory of Ukraine (including Crimea) from the Twitter Streaming API between April 11 and September 15, 2015. Table 5 gives the numbers of tweets sent in Ukrainian and Russian, according to the language tags supplied by Twitter.

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

Table 5. Number of tweets in dataset by language

In order to test whether these results are consistent over time or merely the result of brief, intensive bursts of activity in particular areas, I calculated the Ukrainian-to-Russian ratios for each of the complete calendar months in my dataset (May to August 2015), and then compared the average monthly ratios and standard deviations for each oblast.[8] In no oblast was the

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

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, it can be concluded that these results are consistent over the period covered by the dataset.

4.2.1. Data on users in each oblast and their language behaviour

I mapped the coordinates of each tweet in my dataset to Ukraine’s 27 oblasts (prefectures). Table 6 gives the breakdown by oblast for total number of users, those tweeting in Russian, in Ukrainian, and in both languages.

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
Dnipro	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
Khmel'nyts'kyy	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
Zakarpattya	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

Table 6. Numbers of users by language and oblast

468,214 unique users were identified; this constitutes 1.03% of the 45.15 million population of Ukraine. While this is of course a self-selecting sample and by no means to be taken as representative of the whole population, it is clear that here I am observing the behavior of more than a tiny group of enthusiasts. Furthermore, given that geotagged tweets have been found to be only a few percent of all tweets sent, I have reason to claim that the 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 7 provides counts of tweets by language and oblast:

Oblast	All languages	Ukrainian	Russian	ratio Uk:Ru
Kyiv City	417298	44945	257569	0.174
Dnipro	403656	26805	272761	0.098
Kharkiv	171623	11031	112206	0.098
Donetsk	158066	8812	109428	0.081
Odesa	150060	9416	98576	0.096
Lviv	115301	26107	58015	0.450
Kyiv	108915	11237	69737	0.161
Zaporizhzhya	107575	6524	73735	0.088
Vinnytsya	73917	12660	41142	0.308
Mykolayiv	66394	4420	44938	0.098
Cherkasy	66091	10817	38943	0.278
Crimea	60339	3733	40343	0.093
Chernivtsi	52739	6167	32839	0.188
Kherson	49923	2910	34347	0.085
Poltava	48675	4319	31698	0.136
Chernihiv	43911	4127	28697	0.144
Kirovohrad	43815	3815	28357	0.135
Volyn	37899	9530	18897	0.504
Luhansk	37398	2024	26225	0.077
Khmelnytskyi	34673	6349	19666	0.323
Zhytomyr	30189	3944	18657	0.211
Ivano-Frankivsk	28367	5100	12864	0.396
Rivne	28202	6950	14196	0.490
Sumy	22988	1730	15411	0.112
Zakarpattya	20053	2365	10839	0.218
Sevastopol	16419	1059	10811	0.098
Ternopil	15122	3836	7284	0.527
TOTAL	2409608	240732	1528181	0.158

Table 7. Counts of tweets by language and oblast

As we can see from Table 7, 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, Ternopil and Volyn, does the ratio of Ukrainian to Russian tweets creep above 1:2. (0.527 and 0.504 respectively).

However, Figure 3, which plots the contents of Table 7 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,[9] the band with the lowest values forms a contiguous block containing 11 eastern and southern oblasts.

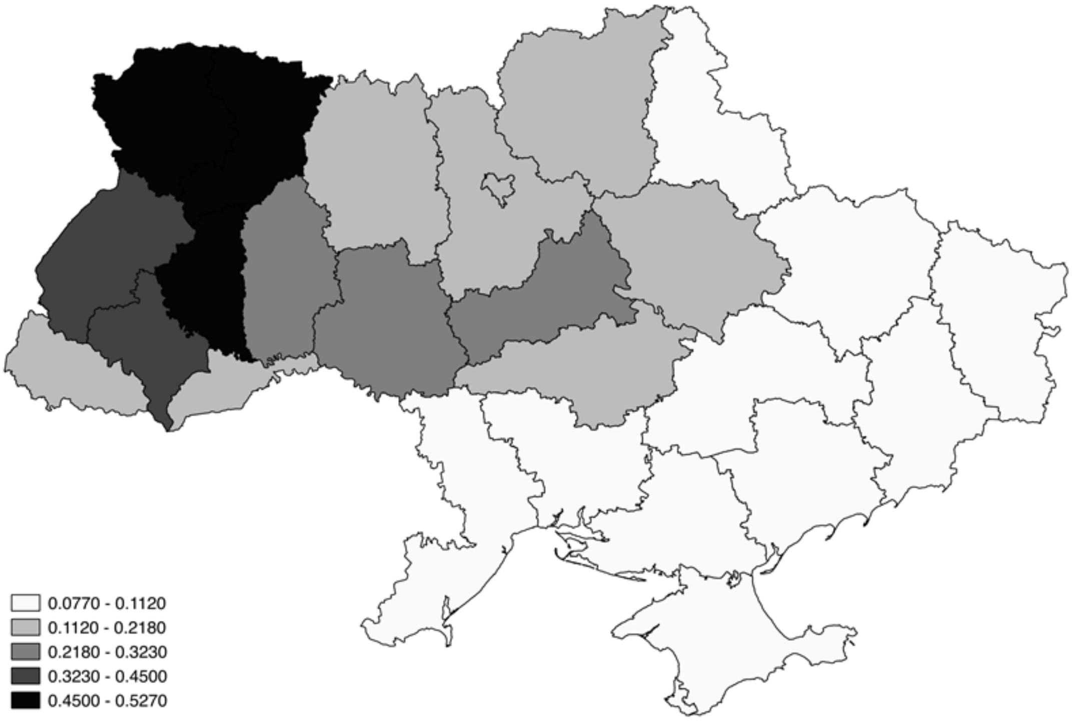

Figure 3. Ratio of Ukrainian to Russian tweets by oblast

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

The border between the 0.0770-0.1120 band (colored white in the map) and its western neighbors is almost the same as the electoral border of 2012 elections in Figure 1, 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. My data suggest that many Twitter users in Ukraine who regard Ukrainian as their mother tongue prefer to use Russian in online communication for some reason. It might be the case that many of those users who prefer to interact in Ukrainian in their offline networks tend to use the "dominant" language (Russian) in their online communications. On the other hand, Søvik (2010) in her research on language behavior in Kharkiv showed that census data do not reflect actual language use even in offline communities:

"According to data from the 2001 census, among the population of Kharkiv oblast, 53.8%, declared Ukrainian as their native language and 44.3% stated Russian. Thus, the Ukrainian language is considered the native language of a majority, but the primary language (L1) of most of the Kharkiv population is, by all accounts, Russian. Ukrainian may thus be designated as a second language (L2) for those for whom it is not L1 or those who are functionally bilingual. When the two demographic measures of ethnicity and native language are put together, they display some incongruence." (Søvik, 2010:11).

Hence, it is also possible that what we are seeing here is just Ukrainians' offline linguistic behaviour moving online.

The last two columns of Table 6 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. I return to this in the following section.

Another possible explanation of this discrepancy between the observed online behavior and the census data regarding mother tongue is that there is a gap between the "language identity" and the "actual language use" of Ukrainians. Such an explanation would be supported by the findings of Kulyk (2011), who designates language identity and language use separately and claims that "language identity is a no less powerful predictor of Ukrainian citizens' attitudes and policy preferences than language use" (Kulyk, 2011:644).

4.2.2. Users' bilingual behaviour

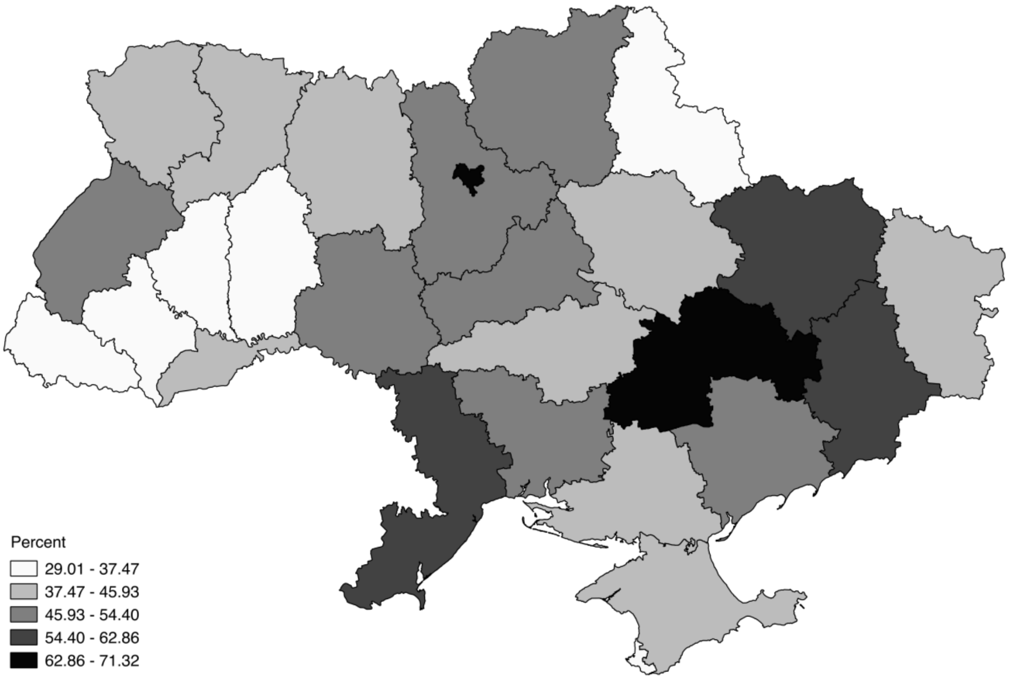

Figure 4. Proportions of users tweeting in Ukrainian who also tweet in Russian

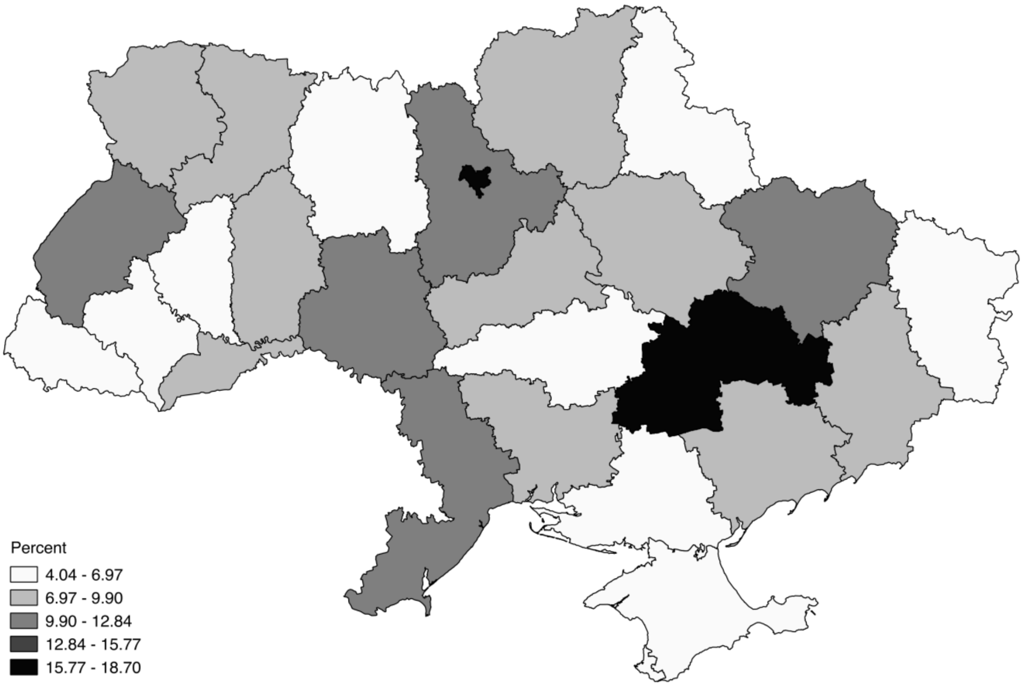

Figure 5. Proportions of users tweeting in Russian who also tweet in Ukrainian

As noted above, of all users tweeting in Ukrainian, more than half also tweet in Russian, whereas of all users tweeting in Russian, only one in ten also tweets in Ukrainian. 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 four largest cities: Kyiv City, Dnipro, Kharkiv and Odesa.

This 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 my use of Twitter data to measure the geography of language use in the country.

Clearly, the active Twitter users in my 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 my 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 I cannot claim that my sample is representative of the country's population as a whole. I can merely argue that this 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 the question of mobility was also ignored. Most users, who sent geotagged Tweets, did so from their mobile devices. As this study counts tweets sent in different regions of Ukraine, it is almost certain that my dataset includes tweets sent from different regions by the same user. However, it can be seen as a feature not a bug, in the sense that my 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.

4.3. Conclusions

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.[10]

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.

In this research I demonstrated the feasibility of using data from social media (specifically, Twitter) to capture and process information about the language use of a large number of people across the country. While I cannot make claims about the whole Ukrainian population on the basis of my self-selecting sample, these 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, I am 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. Further work is needed to investigate my result, and all simplistic explanations and extrapolations should be avoided.

[10] 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.

This chapter showed strong regional variations in the ratio of Ukrainian to Russian geotagged tweets. I used information about the users to calculate rates of bilingual use among those tweeting in the two languages, again on a regional (oblast-by-oblast) basis and I found a clear trend of higher ratios of Ukrainian to Russian tweets in the west of the country, where some oblasts (Ternopil, Volyn, Lviv, Rivne) have about 50 percent of the tweets written in Ukrainian, while in some southern and eastern regions the share of tweets in Ukrainian is less than 10 percent. Moreover, of all users tweeting in Ukrainian, overall, more than half also tweeted in Russian, whereas of all users tweeting in Russian, only one in ten also tweeted in Ukrainian. I also found that those tweeting in Ukrainian in the east of the country were more likely than those in the western parts to also tweet in Russian.

Chapter 5. Language use on Ukrainian governmental pages on Facebook

Although my findings in the previous chapter showed strong regional variations in the ratio of Ukrainian to Russian geotagged tweets, given that users who send geotagged tweets in Ukraine constitute only a small group of social media users, and Facebook is much more popular social network than Twitter, there was a need to investigate the language behaviour of the Ukrainian Facebook users to either support or disprove my findings in the research on Twitter. In this chapter I describe the findings from an exploratory study of language use on Facebook pages maintained by city and regional (*oblast*) councils in Ukraine. City and regional pages were chosen because they show government institutions broadcasting information to citizens (in the form of updates to the Facebook pages), and citizens responding both to the government communications and to other citizens (in the form of comments). Furthermore, they are local in character: the updates are written for people living in the city or region, and it is likely that the comments are written either by residents or by former residents who still have a strong connection to the area. This allowed me to study language use in government-citizen and citizen-citizen communications on a region-by-region basis.

Due to the lack of previous studies on this topic, my first concern was to establish what it is possible to achieve: what kind of data is available, from which sources, in what quantities, and over what period. The answers to these questions will clearly influence the other research questions I will be able to formulate and answer.

I want to observe three different instances of language choice: by government officials in their updates to the pages; by users (citizens) in commenting on those updates; and by users (citizens) in reply to other users' comments. I will analyze the data both on a page-by-page basis in order to identify regional trends in language use; and on a comment-by-comment basis in order to understand exchanges on a micro level.

5.1. Tasks and questions

As noted above, my first task is to establish the availability and extent of publicly available data. This then, is my first question in research on Facebook:

RQ1: Is enough data available from city and regional councils' Facebook pages to carry out meaningful research?

Assuming I am able to collect sufficient data, I wish to investigate how language use by city and regional governments varies across Ukraine. Therefore my second research question is:

RQ 2: Does language use by page maintainers in updates show a correlation to reported language use statistics for different Ukrainian regions?

I also want to investigate language use by citizens in their interactions with government, at least in the public realm of social media. In terms of this research, I will observe region-by-region trends in language use among Facebook users posting comments on updates, thus:

RQ 3: Does language use in comments on updates show a correlation to reported language use statistics for different Ukrainian regions?

The above two research questions investigate region-by-region trends and therefore have the Facebook page as their unit of analysis. However, I am also interested in language choice in interactions between citizens and government. It can be operationalized here as follows:

RQ4: Is there a relationship between the language of updates and the language used in comments on those updates?

Because users can also reply to other users' comments, language choice in exchanges between citizens can be observed. Thus:

RQ5: Is there a relationship between the language of comments and the language used in replies to those comments?

5.2. Data collection and language detection

Facebook provides an Application Programming Interface (API) that allows researchers to download updates and comments from specified pages. However, the pages publicly available through this API are strictly limited; pages created by individuals as well as many organizations are not available unless the user is a friend of the page owner. After extensive searching I identified and accessed 24 Facebook pages, run by city and regional governments that were available through the API. A total of 31,370 updates and 11,044 comments posted on these pages have been downloaded. The oldest update was from July 2010 and the most recent from June 2016. Table 8 provides details of the pages along with numbers and dates of updates, comments and replies; see also the map in Figure 6. Figures 7 and 8 show the numbers of updates and comments over time; we can see the number of updates starting to increase in 2014, and the number of comments rising from 2015. This suggests that city and regional governments are making increasingly active use of Facebook as a means of communicating with citizens, and that increasing numbers of citizens are using Facebook to provide feedback to their local administration.

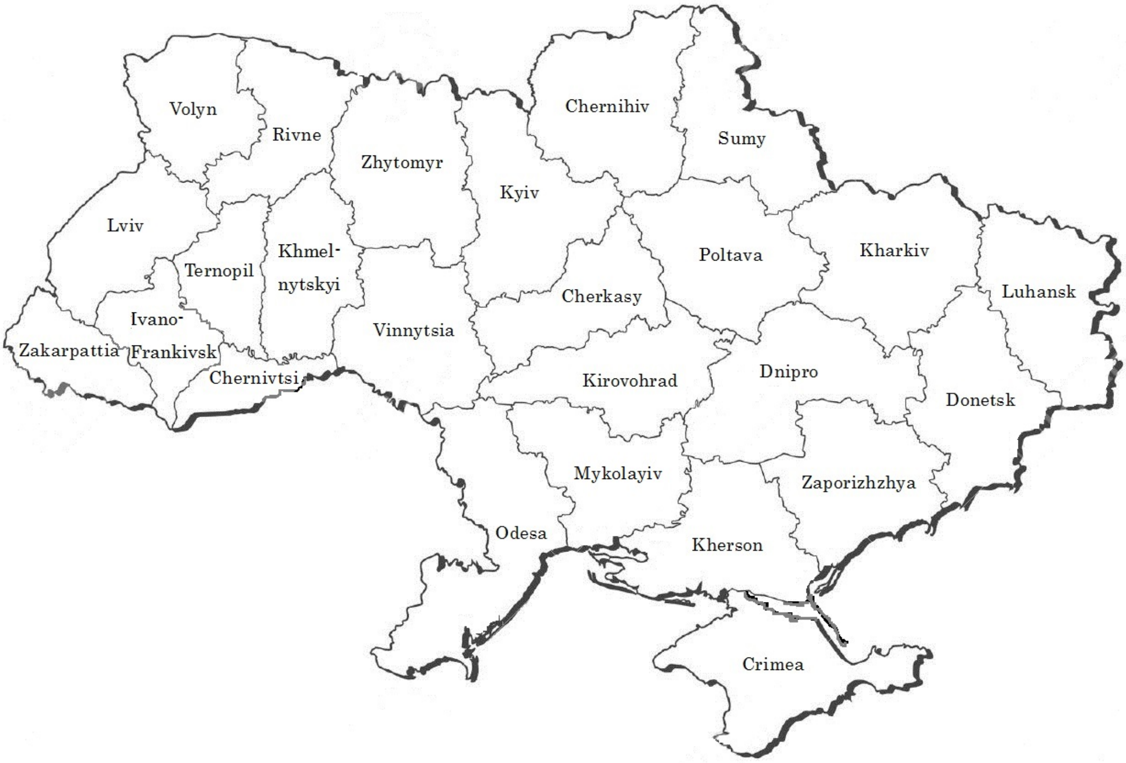

Figure 6. Map of Ukraine's regions (*oblasts*).

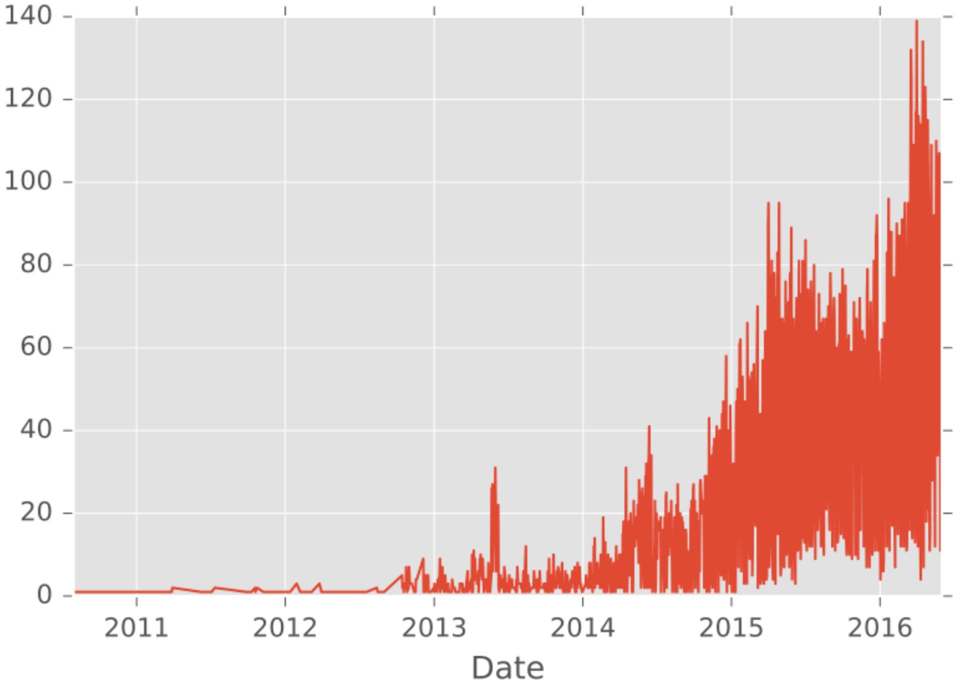

Figure 7. Numbers of updates posted each day.

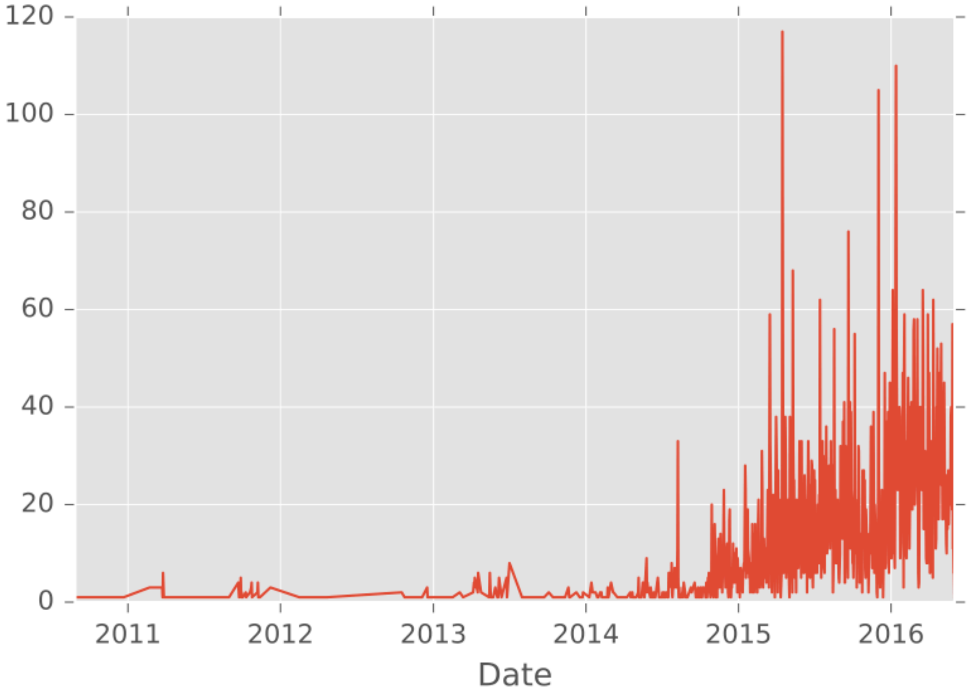

Figure 8. Numbers of comments and replies to comments posted each day.

Through the updates of Facebook pages, officials of city councils and oblast inform the residents of current events in their administrative unit. For example, consideration of the petition against the gay-parade in the capital of Ukraine (Kyiv), opening of a fan-zone for Euro 2016 (Kharkiv), news about an ecological catastrophe on the site of a solid waste landfill in the village of Hrybovychi (Lviv), news about the opening of the Second Festival of Classical Music (Odesa), and a photo-report on the Agro-2016 International Exhibition (Cherkasy).

To identify the language of the Facebook updates and comments I used Google's Compact Language Detector (CLD). I showed in Chapter 3 that CLD is acceptably accurate in recognizing both Russian and Ukrainian. To check the performance of the CLD in recognizing the language of Facebook comments and updates, I manually tagged the language of 775 Facebook comments from the dataset and compared the results to those of the CLD. I found that, after removing URLs, the Cohen's Kappa for inter-coder agreement between the native speaker and the CLD was 0.827, which falls in the "nearly perfect agreement" range (Landis and Koch 1977).

5.3. Results

My dataset contained a total of 12,598 updates in Ukrainian and 10,322 updates in Russian. 8,394 updates were tagged as being in no language because they comprised only a picture and no text, and 56 updates were tagged as being in other languages. Turning to comments, I found that 5,582 were in Russian, 3,230 were in Ukrainian, 1,928 were tagged as none, and 304 as being in other languages.

As Table 8 shows, there is great variation in the number of updates and comments posted on the various sites. The total number of updates in Russian is greatly boosted by the Kharkiv Regional council's page, which has more than twice as many updates and comments as any other page.

Page	Region	Updates					Comments		
		Total	Oldest	Newest	% Uk	% Ru	Total	% Uk	% Ru
Chernihiv. Region.Rada	Chernihiv	144	2015/4/21	2016/5/30	95.83	0	15	6.67	66.67
CvCouncil	Chernivtsi	1448	2015/3/19	2016/5/27	37.91	0.14	139	51.8	20.14
Dnepropetrovsk	Dnipro	24	2010/7/30	2012/8/31	8.33	66.67	83	4.82	71.08
Dneprorada	Dnipro	568	2016/3/2	2016/5/27	98.59	0.35	195	20	67.69
Hmmiskrada	Khmelnytskyi	563	2015/12/25	2016/5/29	90.59	0	1406	77.67	8.89
Kharkov government	Kharkiv	3401	2013/5/22	2016/5/10	2.21	88.27	2884	2.05	69.59
Kontcenter	Kirovohrad	95	2016/2/11	2016/6/1	100	0	0	0	0
kyiv.city.council	Kiev City	1419	2014/10/17	2016/5/27	96.69	0.99	1095	39.18	49.22
lviv.adm	Lviv Oblast	1424	2016/3/16	2016/5/28	99.44	0	187	67.91	6.42
Miskaradazt	Zhytomyr	301	2015/1/24	2016/4/5	99.34	0	41	34.15	46.34
mrada.if.ua	Ivano-Frankivsk	1399	2013/4/4	2016/5/27	42.89	0	140	81.43	0.71
Mykoblrada	Mykolaiv	670	2014/5/5	2016/5/31	93.73	1.19	190	30.53	55.79
Oblradaks	Kherson	461	2015/12/14	2016/5/31	88.07	10.63	127	28.35	51.18
oda.odesa	Odesa	268	2015/7/23	2016/5/31	33.21	29.85	373	22.79	56.57
Odalug	Luhansk	2681	2014/7/2	2016/6/1	95.56	3.99	1794	27.98	57.13
poltava.council	Poltava	521	2014/5/17	2016/5/31	98.27	0	70	52.86	27.14
rivne.rada	Rivne	679	2013/4/19	2015/8/26	0.15	0	3	33.33	0
Slavrada	Donetsk	710	2015/1/26	2016/5/31	87.04	2.39	59	35.59	49.15
sorada.gov.ua	Sumy	3346	2014/6/3	2016/6/1	0.24	0	76	80.26	13.16
Ternopil.rada	Ternopil	907	2011/10/19	2016/5/23	79.27	0.11	132	75	1.52
UAPRCKODA	Cherkasy	112	2015/8/4	2016/5/31	57.14	0	12	66.67	0
Vinnytsia	Vinnytsya	1100	2014/6/27	2016/5/27	1.82	0.09	64	65.62	31.25
zakrada.gov.ua	Zakarpattya	1349	2013/12/6	2016/5/31	93.92	0	284	72.18	7.04
Zaporozhye	Zaporizhzhya	7780	2015/3/31	2016/5/31	1.13	90.27	1318	4.7	70.71

Table 8. Details of updates and comments collected from city and regional council Facebook pages. Uk = Ukrainian, Ru = Russian.

My first research question (RQ1) asks whether there is enough data available from city and regional councils' Facebook pages to carry out meaningful research. As the figures in Table 8 imply, my answer is a very qualified affirmative. Some but by no means all city and regional councils around Ukraine are maintaining public Facebook pages whose content researchers can easily obtain and analyze. Some of these pages are updated very frequently, and some attract a large number of comments from visitors to the page. Furthermore, some of the comments attract replies, and that allows researchers to observe communication between citizens in a local context. The number of pages, updates and comments makes it possible to identify patterns in language use in both government-citizen and citizen-citizen interactions. The relatively small number of pages and the low levels of replies to comments overall limit the number of statistically significant conclusions that can be drawn from this data. Nevertheless, the data show a clear trend over the last two years for local governments to post more updates and for users to post more comments. This may prompt other local governments to start Facebook pages and the less active ones to post more frequently; if that happens then researchers may be able to draw more detailed and statistically significant conclusions about language use in Ukrainian social media.

In order to investigate language use in updates (RQ2) and comments (RQ3), I compared the percentage of Ukrainian updates with the percentage of 2001 census respondents reporting Ukrainian as their mother tongue in the 2001 census. I excluded pages with fewer than 50 updates (for my analysis of updates; leaving 19 pages) or 50 comments (for my analysis of comments; leaving 20 pages).

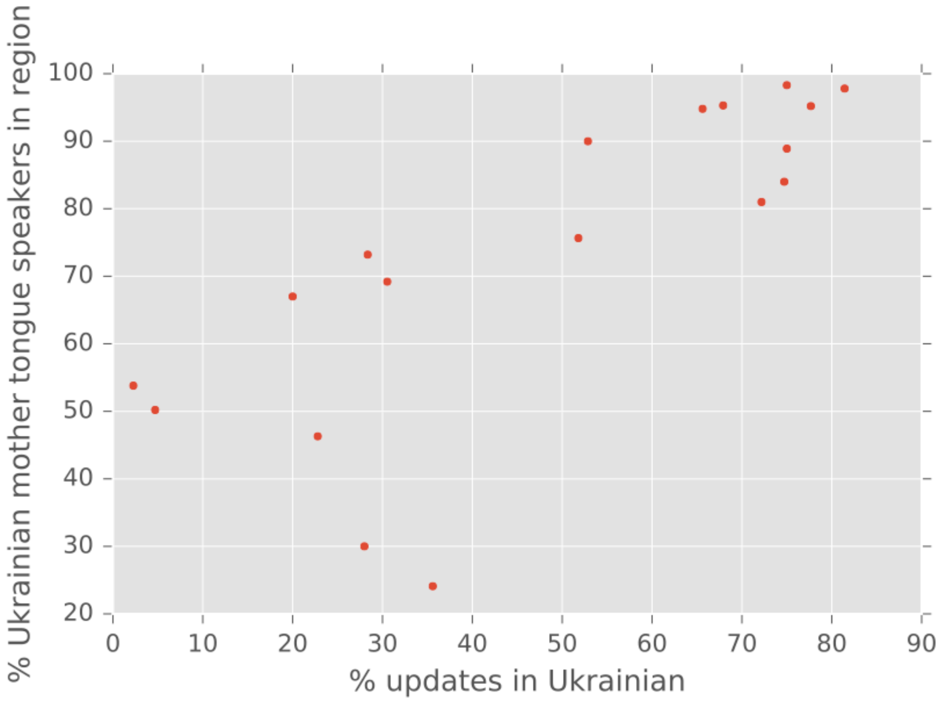

Figure 9. Relationship between language of updates on city and regional council Facebook pages and 2001 census data for the corresponding region (*oblast*).

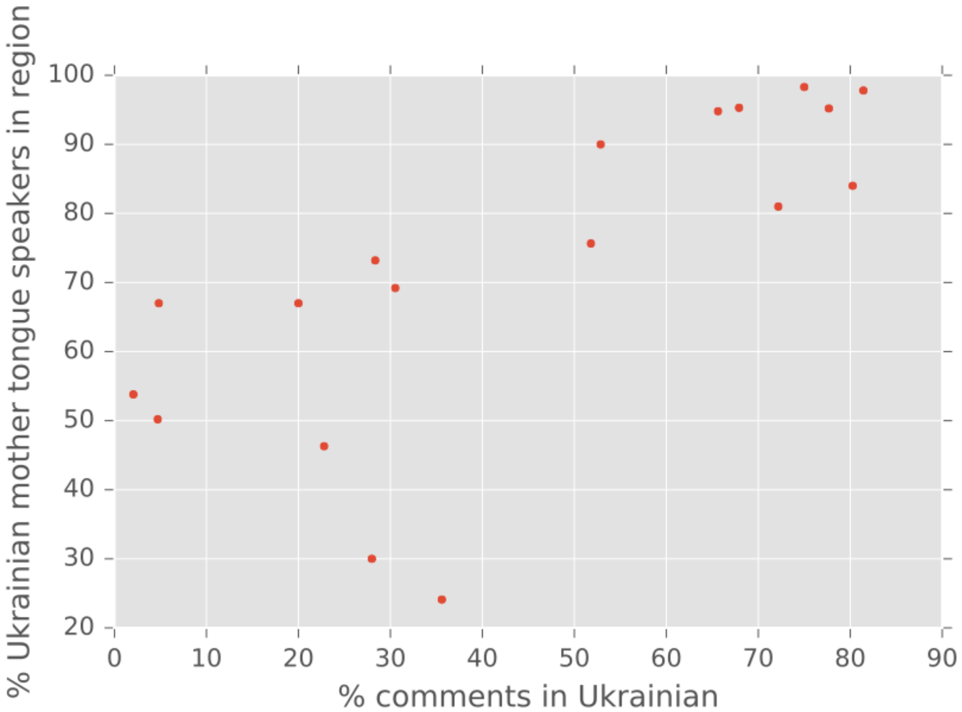

Figure 10. Relationship between language of comments on city and regional council Facebook pages and 2001 census data for the corresponding region (*oblast*).

As the plots in Figures 9 and 10 suggest, pages for cities and regions in areas with more Ukrainian native speakers do tend to get both more updates and comments in Ukrainian. For updates, the Pearson's correlation coefficient between the two variables is 0.77, and for comments the coefficient is 0.73. These numbers provide evidence that there is a positive linear relation between the language of updates and comments on Facebook pages and the mother tongue numbers on a regional basis. However, the number of observations (19 and 20 Facebook pages for updates and comments respectively) is too small to carry out tests for statistical significance, so more data is needed to reach firmer conclusions.

The next step is to ascertain whether users tend to comment on updates in the same language as the update (RQ4), and on other comments in the same language as the comment to which they are replying (RQ5). To find the answers, I first removed updates and comments not tagged as being in Ukrainian or Russian, and also removed "orphan" comments written in Ukrainian or Russian on updates and comments that were not tagged as either Ukrainian or Russian. Then, to analyze comments on updates (RQ4), I removed replies to other comments from my dataset. Table 9 shows the numbers of comments in Ukrainian and Russian on updates in the two languages.

		Update	
		Ukrainian	Russian
Comment	Ukrainian	2148	142
	Russian	1672	2690

Table 9. Languages of Facebook page updates and the comments on those updates.

A chi-squared test of the relationship between the two variables (language of update and language of comment) showed a p-value of 2.19e-269, which allows me to reject our null hypothesis of the two variables being independent of each other. Hence, there is a statistically significant probability that a comment will be in the same language as the update it addresses.

Turning to interactions between citizens (RQ5), I reduced my dataset to replies to other comments. There were a total of 1,437 replies: 467 – Ukrainian to Ukrainian, 82 – Russian to Ukrainian, 227 – Ukrainian to Russian, and 661 – Russian to Russian (Table 10).

		Comment	
		Ukrainian	Russian
Reply	Ukrainian	467	227
	Russian	82	661

Table 10. Languages of comments on Facebook page updates and the replies to those comments.

Once again, I analyzed the relationship between the language of a reply and the language of the comment it was attached to. A chi-squared test of the relationship between the two variables (language of comment and language of reply) showed a p-value of 2.62e-06, which allows me to reject my null hypothesis of the two variables being independent of each other. Hence, there is a statistically significant probability that a reply will be in the same language as the comment it addresses.

Does my dataset permit me to investigate language use, and in particular bilingualism, by individual users? In order to do that I would need to find users switching their language to fit the language used by another user in her comment. Clearly I need users posting more than one reply in order to see whether they switch languages. Unfortunately only 29 individuals posting replies in both Ukrainian and Russian were identified. This number is insufficient to carry out meaningful quantitative analysis.

5.4. Conclusions and topics for future research on Facebook

This exploratory study has shown that local and regional council Facebook pages in Ukraine can offer some useful and up-to-date insights into the use of language by local administrations, the use of language by citizens in their interactions with the state, the use of language in online interactions between citizens in the context of a government-controlled online space.

Previous research (Azhnyuk, 2008:345; Olszanski, 2012:22-23) showed 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. My analysis suggested that the findings are holding true in the increasingly important arena of online communication. 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 order to pursue that question it might be necessary to choose Facebook pages with larger amounts of comments and replies; however such pages may have the disadvantage of lacking a regional focus.

The relatively small number of pages and the low levels of replies to comments limit the number of statistically significant conclusions that can be drawn from this data. Nevertheless, the data show a clear trend over the last two years for local governments to post more updates and for users to post more comments. This may prompt other local governments to start Facebook pages and the less active ones to post more frequently; if that happens then researchers may be able to draw more detailed and robust conclusions about language use in Ukrainian social media.

Chapter 6. Age, gender and language use on Ukrainian Twitter

The lack of reliable demographic information about users is perhaps the greatest problem confronting researchers using social media data to investigate social phenomena

(Boyd and Crawford 2012).

6.1. Age and language use

In any sociological research project the demographic attributes of users such as gender, age or nationality are basic factors that must be considered as they form the basis of clusters of online communication and provide vital information for the researcher. I therefore make an attempt to explore ways of estimating users' age and gender, and investigate the relationship between these variables and users' linguistic choices. Establishing age and other demographic characteristics of Twitter users is not easy. A group of researchers from the UK developed "techniques for collecting or estimating demographics from Twitter data including analyzing gender, language and location" (Sloan et al. 2013). Building on this work and adapting techniques used by other researchers, Sloan et al. (2015) developed a Twitter user age detection algorithm based on a set of pattern matching rules. They validated the reliability of their algorithm using expert human testers. Applying their technique to British Twitter users, they found that the age distribution of Twitter users is much younger than the UK population as of the 2011 Census, with a peak around ages 16 to 22 accounting for 67.5% of all users. They claimed that more than half of the users (59.4%) belonged to the age group 13 to 20. Their results also pointed to the existence of over half a million Twitter users over the age of 40 in the UK. However they conceded that due to the limitations of their detecting technique, the desirable accuracy can be attained only by analyzing the content of these tweets cross referencing with social survey data.

Because of time constraints it was not feasible to reproduce Sloan et al's algorithm for Ukrainian tweets. However, the necessity to know age of our audience prompts me to explore ways-of estimating users' ages from the content of their tweets.

In this research, I do not intend to make any algorithm for investigating users' age, however I try to identify a group of users in the last years of high-school. For that reason I discuss the relationship between the examination periods to Ukrainian universities and peaks of tweeting activity, establishing the method of detecting the age of Twitter users through the use of words "exam" or External independent evaluation or External independent testing (in short "ЗНО" in both Ukrainian and Russian) in the content of geotagged tweets.

For this stage of the research, I used Twitter's Streaming API to collect geotagged tweets sent from within the territory of Ukraine (including Crimea) between 11 April and 30 September 2015. My dataset has undergone cleaning in process of which some obviously "robot" tweets such as those sent by the FourSquare application were deleted. The resulting dataset comprised 2,458,953 tweets. Twitter tags each tweet with a language, the result of its own language detection algorithm; I have previously shown Twitter's language tags to be acceptably accurate with regard to Russian and Ukrainian tweets. 1,553,787 or 63.2% of the collected tweets were in Russian and 242,829 or 9.9% were in Ukrainian; English tweets accounted for 8.7%, followed by a long tail of other languages such as Slovenian, Polish and Bulgarian with less than 3% each. 5.7% of the tweets were tagged as "language unidentified". The Ukrainian and Russian tweets were sent by a total 70,429 distinct users. As I mentioned in previous chapters, although geotagged tweets account for only a few percent of all tweets sent, it can be said that I am dealing with a sizable sample of Ukrainian online community.

The number of tweets retrieved each day shows great variation over the six-month period, as we can see on Figure 10:

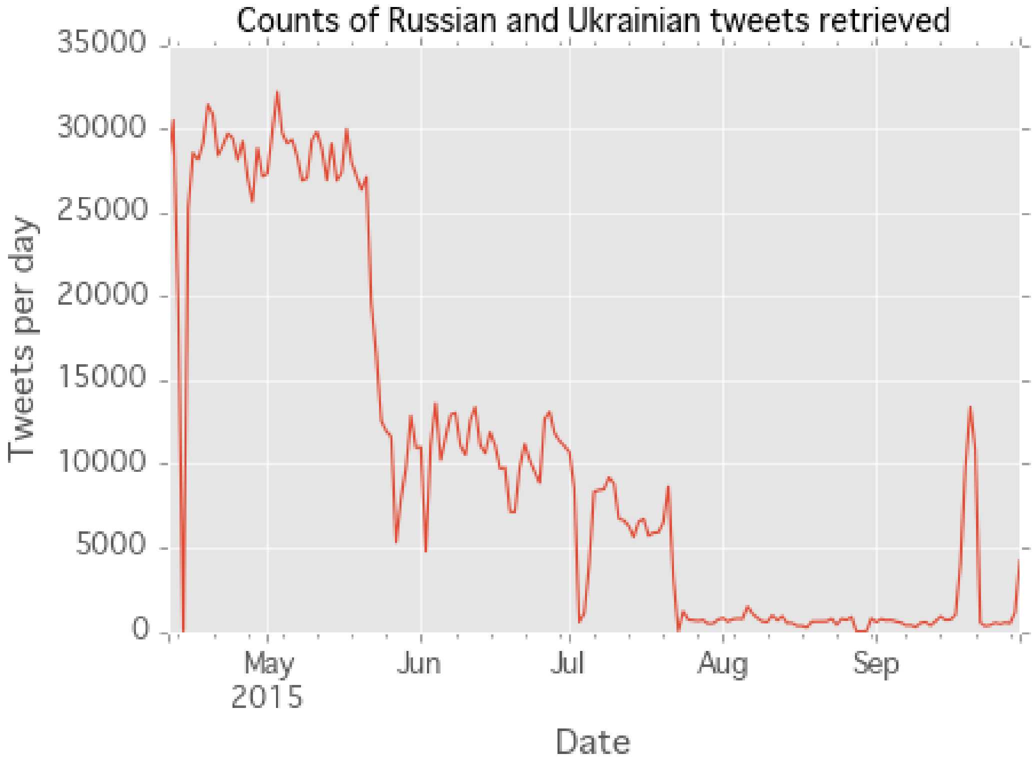

Figure 10. Counts of Russian and Ukrainian tweets

Much higher activity is observed between April and June compared to the summer months. While my data collection was not running during 100% of the period – as shown by the occasional dips of the graph to 0 – it was not the case that data collection was interrupted more frequently in the summer months than in the spring.

Although there are some regional differences in tweeting activity, the April-June peak is a nationwide phenomenon.

6.1.1. Hypothesis

As we can see from Table 11, more than a third of Twitter users worldwide are of young age.

Twitter Age Demographic	Number of Users	Percentage of User Base
18-29	95 Million	35%
30-49	54 Million	20%
50-64	30 Million	11%
65+	13.5 Million	5%

Table 11. Age variations of Twitter users in 2014 worldwide.

Source : Jetscram.com, http://jetscram.com/blog/industry-news/social-media-user-statistics-and-age-demographics-2014/

In regard to Ukraine, as mentioned in Chapter 6, according to demographic data on weekly Internet use 46% of Ukrainians overall say they have used social networking services in the past seven days, with that figure rising to 89.9% among those age 15 to 24. However, I cannot depend solely on these data as it includes all online social networks, while Twitter is used weekly only by 21.6% of all Internet users in Ukraine.

Can I get some information on Twitter users age, then, by using some other methods or techniques?

I hypothesized that if the main reason for this seasonal variation in the number of tweets is high school students' online hyperactivity in the periods of high school and university examinations in May – June, and start of the new academic year in September, then we can find relatively exact percent of a very narrow (15-17 y.o.) age group among the Twitter users in

Ukraine. High school and university students in Ukraine have their examination period in May and June, so the reason for tweeting may be the queries about the content of the exams, requests for assistance, or, most probable, expressing to friends and classmates personal feelings and emotions related to the exams. The period of entrance examinations to the universities in 2015 was from July 10 to September 1, after which a decrease in tweets is observed. Hypothetically, the high school examinations and especially EIT (external independent testing) examination rush could make an impact on this drastic increase of tweets in April-May. Subsequently, if my suggestion proves to be correct, I can state that a reasonable amount of geotagged tweets has been sent by high school students, graduates or university students during or before their examination period.

6.1.2. Examinations in Ukraine

Examinations in Ukraine can be divided into six groups: high school summer final examination, high school graduation examinations, external independent evaluation or external independent testing = EIT (in Ukrainian 3HO), university entrance examinations, university examinations, and university graduation examinations. Among them the most intense ones are the EIT examinations, because they are unified for all educational facilities, they are also independent, which means that they are not biased or influenced by the personal relationship with the examiner. Moreover their results impact the future career of the examinees as they cannot apply to the universities in case they have lower than threshold score. Table 12 shows the dates of the EIT “ЗНО” high school graduates examinations in Ukraine in 2015.

Discipline	Number of participants	Date of exam	Exam result announcement
Ukrainian language & literature (basic) (advanced)	267,394 21,583	24-04-2015	13-05-2015
French	840	03-06-2015	25-06-2015
German	3,172	05-06-2015	25-06-2015
Spanish	179	08-06-2015	25-06-2015
English	81,318	10-06-2015	25-06-2015
Math (basic) (advanced)	129,142 17,650	12-06-2015	25-06-2015
Russian	3,645	15-06-2015	03-07-2015
Biology	98,372	17-06-2015	03-07-2015
History of Ukraine	158,556	19-06-2015	03-07-2015
Physics	51,463	22-06-2015	03-07-2015
Geography	65,541	24-06-2015	08-07-2015
Chemistry	39,730	26-06-2015	09-07-2015

Table 12. ЗНО (EIT) Schedule in 2015

To check the correctness of my suggestion, I calculated how often the words “exam” or “ЗНО” appear in the content of the tweets. Depending on the results, it might be possible to suggest what percent of geotagged twitter users are of high school/university age and to what extent the examination period impacted the tweeting activity in Ukrainian communities in 2015.

I found that the word “exam” “экзамен” (or its derivatives, in lower and upper case) in Russian appears in 6,375 tweets and “екзамен” (or its derivatives) in Ukrainian appears in 610 tweets. Moreover, in 90 tweets which were identified as Russian the word “екзамен” (or its derivatives) is spelled in the Ukrainian way. In those cases the senders presumably either did not know the correct spelling of the word exam in Russian, or wrote it deliberately in the Ukrainian manner.

I also counted occurrences of the usage of “ЗНО” – the Ukrainian/Russian abbreviation for external independent evaluation or external independent testing. My count was case-insensitive and excluded instances where “зно” was used as an adverbial suffix. I found 2,931 occurrences in Russian and 1,328 in Ukrainian, which means that almost one third of all tweets with the word “ЗНО” were sent in Ukrainian.

Third, I calculated the number of tweets including words EXAM or its derivatives

("ЭКЗАМЕН" in Russian and "ЕКЗАМЕН" in Ukrainian) and "ЗНО" on weekly basis. I also checked if the increase in tweets is related to the most intense "ЗНО" exam periods.

WEEK	Total Number of tweets	Total Number of tweets with EXAM + ЗНО	% of Total	Tweets with Exam in RU	Tweets with Exam in UK	Tweets with ЗНО	Tweets with ЗНО % of Total
2015/4/12	59947	69	0.12	36	1	32	0.05
2015/4/19	166799	634	0.38	230	31	373	0.22
2015/4/26	211406	3036	1.44	538	56	2442	1.16
2015/5/03	205698	554	0.27	347	51	156	0.08
2015/5/10	208653	519	0.25	384	49	86	0.04
2015/5/17	206371	1092	0.53	677	61	354	0.17
2015/5/24	162951	967	0.59	789	75	103	0.06
2015/5/31	72790	592	0.81	513	53	26	0.04
2015/6/07	78298	1317	1.68	1126	113	78	0.10
2015/6/14	85198	1047	1.23	681	85	281	0.33
2015/6/21	68848	676	0.98	422	76	178	0.26

2015/6/28	80548	535	0.66	368	39	128	0.16
2015/7/05	49218	146	0.30	125	3	18	0.04
2015/7/12	59131	70	0.12	45	2	23	0.04
2015/7/19	44165	52	0.12	39	2	11	0.02
2015/7/26	21596	30	0.14	25	1	4	0.04
2015/8/02	4476	4	0.09	2	0	2	0.04
2015/8/09	6377	2	0.03	1	0	1	0.02
2015/8/16	4574	3	0.07	3	0	0	0.00
2015/8/23	3937	2	0.05	2	0	0	0.00
2015/8/30	2831	2	0.07	0	0	2	0.07
2015/9/06	4711	6	0.13	2	1	3	0.06
2015/9/13	3182	4	0.13	0	1	3	0.09
2015/9/20	31908	22	0.72	6	0	16	0.05
2015/9/27	14162	11	0.08	4	0	7	0.05
TOTAL	**1857775**	**11392**	**0.61**	**6365**	**700**	**4327**	**0.23**

Table 13. The weekly results on tweets with the words EXAM and/or 3HO

From 3HO (EIT) Schedule table 13 I can conclude that the most important weeks for

“ЗНО” participants are:

1) April 20 –April 26 (as the exam on Ukrainian language and literature is on April 24) in which the number of tweets including the word ЗНО was the highest = 2,442 tweets;

2) May 11 – May 17 (as the announcement of the results of the exam on Ukrainian language and literature is on May,13) = 354 tweets;

3) June 8 – June 14 (as the exam on Math is on June 12) = 281 tweets;

4) June 15 – June 21 (as there are two major exams on this week: Biology – June 17, and History of Ukraine – June 19) = 178 tweets;

5) June 22 – June 28 (as the announcement of the results of the Math, English, Spanish, French, German exams on is on June 25) = 128 tweets;

6) June 29 – July 5 (as the announcement of the results of the Biology, History of Ukraine, Russian, and Physics is on July 3). I found an unexpectedly low number of tweets here = 18 tweets.

Although the intensification of Twitter activities is seen on the above weeks, the numbers of tweets using the word ЗНО are lower than I expected, except the April 20 –April 26 week. The total amount of tweets with either word “exam” (7,065) or word “ЗНО” (4,332) reaches 11,397. However, it is seen that this portion constitutes around 0.6% of all 1,857,775 geotagged tweets sent in April – September 2015 in Ukraine. As “ЗНО” is significantly important only for those high-school students, who aim to enter prestigious universities in Ukraine, these data do not prove or contradict the results of the previous research on Twitter activity, implying that the typical representative of Twitter users in Ukraine is a user characterized by young age and relatively high social and economic status (Kostenko, 2011; Pentina, Basmanova and Zhang, 2014).

6.1.3. Findings

In Chapter 4 I found that in the period of April-August 2015 in the country as a whole, more than six Russian tweets were sent for every Ukrainian tweet. Now I can state that the same tendency is seen with the tweets related to exams: Russian tweets heavily outnumber Ukrainian tweets even during the periods of “ЗНО” exams, which are conducted in Ukrainian.

The ratio of Ukrainian to Russian tweets concerning exams is around 1:10, which means that in 2015 the tendency among high-school students and graduates in Ukraine in tweets related to examinations is even less than in general tweets (1:6).

WEEK	Exam RU	Exam UK	% of exam UK to exam RU
2015/4/12	36	1	2.78
2015/4/19	230	31	13.48
2015/4/26	538	56	10.41
2015/5/3	347	51	14.70
2015/5/10	384	49	12.76
2015/5/17	677	61	9.01
2015/5/24	789	75	9.51
2015/5/31	513	53	10.33
2015/6/7	1126	113	10.04
2015/6/14	681	85	12.48
2015/6/21	422	76	18.01
2015/6/28	368	39	10.60
2015/7/5	125	3	2.40
2015/7/12	45	2	4.44
2015/7/19	39	2	5.13
2015/7/26	25	1	4.00
2015/8/2	2	0	0.00

2015/8/9	1	0	0.00
2015/8/16	3	0	0.00
2015/8/23	2	0	0.00
2015/8/30	0	0	0.00
2015/9/6	2	1	50.00
2015/9/13	0	1	n/a
2015/9/20	6	0	0.00
2015/9/27	4	0	0.00
Total:	6365	700	11.00

Table 14. Proportion of tweets with word “EXAM” in Russian to those in Ukrainian

My research on users’ age has shortcomings such as an inability to check the real age of a tweet sender, access only to geotagged tweets, but still I can tentatively conclude that there is no visible trend among high-school students or graduates to prefer Ukrainian language over Russian in their daily life. Depending on my data I can state that the peak of tweeting activity is related to the examination period in ЗНО (EIT). However, the percentage of tweets including the words ЗНО is less than 1 percent in all periods except the first week before the obligatory ЗНО exam of Ukrainian and literature (April 24). As previously mentioned, “ЗНО” is important only for high-school students aiming to enter prestigious universities, so relatively seldom usage of it does constitute evidence of a lower percentage of high-school students among Ukrainian Internet users. This research showed that portion of tweets related to school or university examinations constitutes around 0.6% of all 1,857,775 geotagged tweets sent in April – September 2015 in Ukraine. With such an insignificant percent of these tweets in my data it is impossible to prove or refute the statement that the typical representative of Twitter users in Ukraine is a user characterized by young age. Moreover, I found it extremely difficult to identify the age of tweet senders with high accuracy, solely depending on the contents of their tweets.

As for the language in examination related tweets, I found that there is no visible tendency among young Ukrainians to prioritize Ukrainian language in their tweets concerning school or university examinations. Russian tweets concerning examinations outnumber Ukrainian ones by ten to one, higher than the six to one ratio in all geotagged tweets sent in the same period.

6.2. Gender and language use

Although it is difficult to estimate the ages of those who sent geotagged tweets from the territory of Ukraine, gender variations can be distinguished more easily. Some grammar attributes of the Ukrainian and Russian languages can provide hints on the gender of the author of a tweet. While previous research has, as introduced in Chapter 2 of this thesis, discussed gender differences in Twitter use in terms of discourse, usage of certain words, level of emotional response, users' motivations or aspirations, content of the tweets in relation to political issues, attitudes toward disclosure of personal information, sexual identity, emotive features, and semantic themes, to my knowledge no previous research has investigated whether language choice on Twitter varies with gender.

In this section my dependent variable is language choice (i.e. Ukrainian versus Russian) in tweets, which I operationalize as a binary variable which is True if a user sends more than the average proportion of their tweets in Ukrainian and False otherwise. My independent variables are the user's location and the user's gender.

Given the many differences in social media use by women and men shown by previous research, it would come as no surprise to find gender differences in language choice, but neither previous research nor theory leads me to expect language choice and gender among Ukrainian Twitter users to be related in a particular direction. My first hypothesis is therefore:

H1 there is no relationship between user gender and language choice.

Previous research on gender and Twitter use also makes no mention of a correlation between gender and location, e.g. of women in rural locations being more likely than similarly located men to use social media. Hence, my second hypothesis is:

H2 there is no relationship between user gender and location.

The following section outlines how data for this study has been obtained.

6.2.1. Data collection

I increased the number of tweets in my dataset and in this section I work with larger data than in Chapter 4. Between 11 April 2015 and 26 June 2016 3,807,456 tweets from Twitter's Streaming API geotagged for Ukraine (including Crimea) were collected. After excluding just over one million tweets in languages other than Ukrainian and Russian (see the following section on "identifying language") I had a dataset of 2,738,022 tweets sent by 103,307 users.

6.2.2. Identifying language, gender and location

Identifying language

As explained in Chapter 3, I investigated the level of accuracy of language detecting systems and came to conclusion that that Twitter's language tagging of Ukrainian and Russian tweets is by no means perfect but is nevertheless sufficiently accurate for its results to be used in research on language use. My dataset contained 2,370,496 tweets in Russian and 367,526 tweets in Ukrainian.

Identifying gender

Concerning the gender identification, it is possible to estimate users' gender using linguistic clues in the text of tweets. For Russian tweets, I was able to make use of Helmut Schmid's part-of-speech tagger[11] together with the Russian parameter file provided by Serge Sharoff[12]. This identifies verbs in the text and labels verbs in the past tense as masculine, feminine, neutral or plural.

My gender detection first used this part-of-speech data to look for first person pronouns in each tweet; when it found one, it took the gender of the first verb following the first person pronoun as indicating the gender of the writer. In tweets without a first person pronoun, if the first word of the tweet was a verb then the gender of that verb was taken to indicate the gender of the writer.

When this part-of-speech information did not yield a result, which was of course the case for all our Ukrainian tweets, I used a dictionary of female and male words and word endings in Ukrainian and Russian. For each tweet, I counted female and male words based on the word being included in the dictionary of words or having a male or female ending. Tweets with a greater count of female to male words were then tagged as being written by females, and vice versa; tweets without any gender words or with equal numbers of male and female users were tagged as gender unknown.

I then excluded users who had ten or fewer tweets in Russian or Ukrainian in my dataset.

[11] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[12] http://corpus.leeds.ac.uk/mocky/

The remaining 32,243 users were tagged as “female”, “male” or “unknown” depending on whether they sent a greater number of female or male tweets or equal numbers of both. The right-hand column of Table 5 shows the results; female users outnumber male users by 1.95 to 1. I excluded the 3,795 users of unknown gender from the rest of my analysis.

To provide a sufficient analysis of the accuracy of gender identification, I selected at random 200 profiles of users, among whom 140 were recognized by our gender detection algorithm as female and 60 as female. Then I checked their profiles manually and compared my results. My findings show that out of those 140 female users’ profiles 114 were female, 3 male and 23 impossible to identify; out of 60 male profiles - 38 were male, 17 female and 5 impossible to identify. These results show the relative correctness of the algorithm for gender identification, and that it is appropriate to use this algorithm for the research on gender. If anything, the manual check suggests that the algorithm may overestimate the number of male users.

My dataset is comprised of geotagged tweets, so I know the user’s geographical coordinates at the time of sending for all the tweets in our dataset. I mapped these coordinates to Ukraine’s oblasts (prefectures). For each user I then counted which oblast they sent most of their tweets from, and set that oblast as the user’s location. The oblasts were grouped into four regional categories: Kyiv City, Dnipro, the West and the South-East. The South-East category included the nine oblasts Crimea, Donetsk, Kharkiv, Kherson, Luhansk, Mykolayiv, Odesa, Sevastopol’ and Zaporizhzhya; the remaining 16 oblasts were assigned to the West category. I decided on this four-way regional categorization because of the traditional division of Ukraine into West and South-East regions, having Dnipro oblast as a representative of Central Ukraine and one the most powerful industrial oblasts. As for Kyiv City, with its massive number of users it represents urban Ukraine, and as a capital city invites people from the different regions of Ukraine. It was also necessary to have sufficient numbers of users in each regional category to allow statistical analysis. Table 15 shows the number of male and female users in each regional category.

6.2.3. Results on location and gender

First, I examined the relationship between two independent variables, location and gender. As Table 15 shows, while female users outnumbered male users everywhere, they did so by only 1.76 to 1 in Kyiv, but by as many as 2.22 to 1 in the South and East. When I did a chi-square test, the resulting p-value (1.13 e-11) was much lower than 0.05, so I could reject my second hypothesis and conclude that there is a statistically significant correlation between users’ genders and locations. This presents me with a puzzle for which I do not currently have an answer. I will discuss possible reasons for the puzzle in the end of this chapter.

	Dnipro	Kyiv	South and East	West	Total
Female	5389	7655	3288	2458	18790
Male	2653	4352	1479	1174	9658
Total	8042	12007	4767	3632	28448
Female/male	2.03	1.76	2.22	2.09	1.95

Table 15. Location and Gender

6.2.4. Results on location and language

Based on the previous findings in my research I would expect to find a very high proportion of Russian tweets in the South and East, and a fairly high proportion of Ukrainian tweets in the West. For each user in my dataset, which includes only users sending more than ten tweets identified as male or female, I calculated the proportion of their tweets written in Ukrainian and a binary value for whether this proportion was higher than the national average or not. The results, shown in Table 16, confirm the strong relationship between region and language. 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 (11.2%) of their tweets in Ukrainian.

	Users tweeting on or less than national average in Ukrainian	Users tweeting more than national average in Ukrainian	Less than / more than
Dnipro	7309	733	9.97
Kyiv	10436	1571	6.64
South and East	4428	339	13.06
West	1782	1850	0.96

Table 16. Language and Location

6.2.5. Results on gender and language

I examined the numbers of female and male users tweeting above or below the national average in Ukrainian. I found that while female users were 5.96 more likely to be tweeting below the national average (11.2%) in Ukrainian, male users were only 4.38 times more likely to do so (see Table 17). The results are statistically significant (p-value of 3.396e-20 in the chisquare-test).

This finding forces me to reject my first hypothesis, and represents another puzzle: why do men seem (somewhat) more likely to tweet in Ukrainian than women? This puzzle will be discussed later in this chapter.

	Users tweeting on or less than national average in Ukrainian	Users tweeting more than national average in Ukrainian	Less than / more than
Female	16091	2699	5.96
Male	7864	1794	4.38

Table 17. Gender and Language Use

Figure 11 summarizes my results, and shows that the tendency for women to tweet more in Russian than men is a nationwide one: in Kyiv, Dnipro and the South and East the gaps between the percentages of women and men tweeting below average in Ukrainian are 5.7, 4.0 and 4.6 percentage points respectively. Only in the West is the gap much smaller, at 0.5 percentage points.

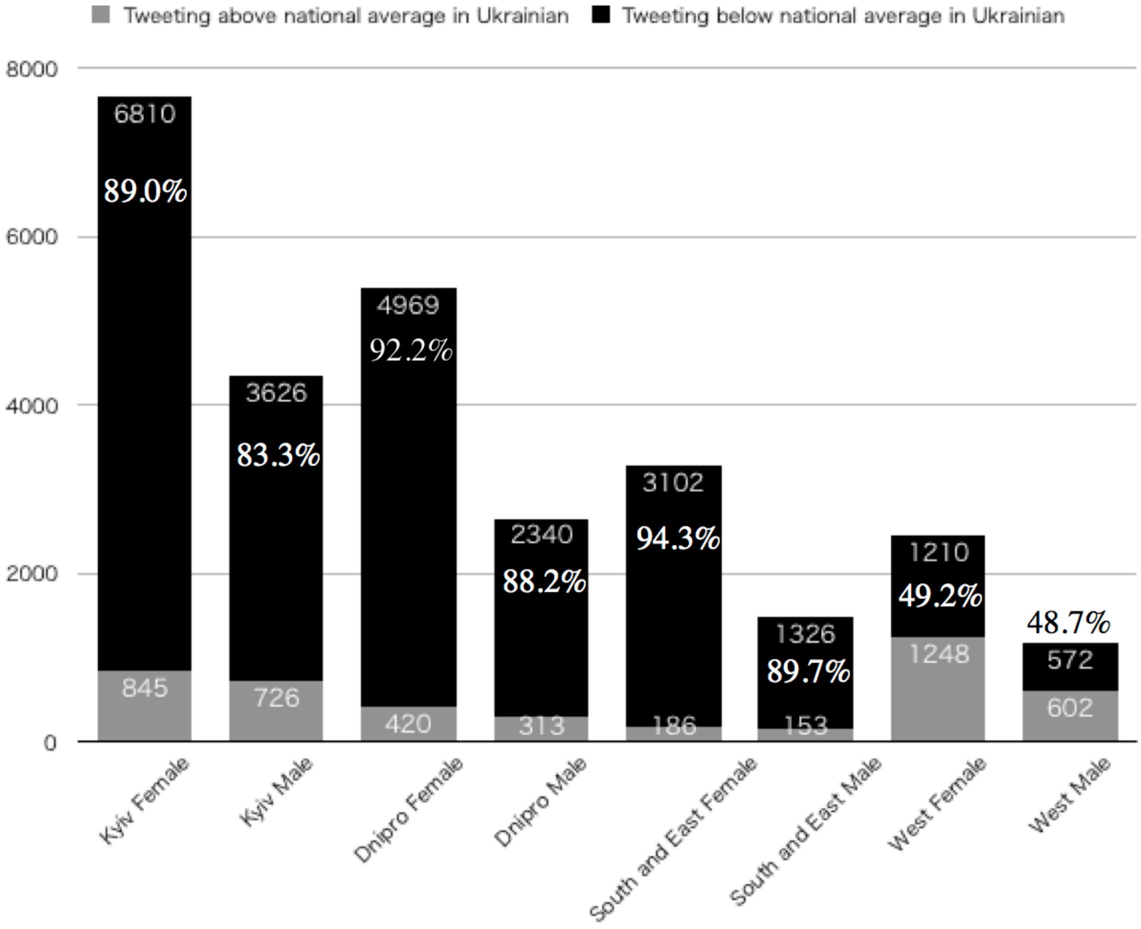

Figure 11. Language Use by Region and Gender

My findings have shown considerable differences in male and female social media use in Ukraine. It is clear that female users outnumber male users by just under two to one throughout the country; this in itself is not necessarily surprising, although it suggests that further investigation should be carried out to understand the reasons for this imbalance. My results have, however, thrown up two puzzles regarding the gender of social media users in Ukraine. The first puzzle is the apparent existence of a regional gender imbalance, with the ratio of female to male users clearly less in Kyiv than in the three other regions. The second puzzle is what appears to be a stronger preference for Russian among female users. Neither previous research nor my current analysis allows me to offer explanations for either of these puzzles.

While it is possible that further rechecking of this data along with the improvement of gender detection algorithms might reveal some methodological bias, assuming that my findings are accurate, how they might be explained? One possible explanation could be a behavioral difference between men and women using social media in bilingual societies: it is possible to speculate that female users are more likely than men to adjust their language to that of those with whom they interact on Twitter. Further analysis of my existing dataset may provide some empirical evidence for or against that possibility.

Another explanation might be that certain topics or kinds of conversation are more likely to be carried out in a single language and others are more likely to be carried out in two languages in a bilingual society; and if Argamon et al. (2007)'s findings that men and women blogging in English tend to write about different topics are applicable to the Ukrainian Twittersphere, this may help to explain what I am observing on this stage of my research. The first step to investigate this explanation would be to implement the same unsupervised text categorization method employed by Argamon et al on my dataset. My findings thus open up some intriguing questions for deeper and comparative research on gender and social media use, particularly in bilingual countries.

Chapter 7. Language homophily in Ukrainian Twitter networks

In previous chapters of this thesis, I examined the language use, location and gender of individual social media (Twitter and Facebook) users in Ukraine. However, given that language is always used in a particular social context, is important to address the "social" aspects of social media, and in particular to investigate the connections between users and the characteristics of the groups of users that form in social media.

7.1. Language homophily.

The flows of messages and status updates serve to create and maintain ties between social media users, each of whom every day makes many decisions about connecting to, reacting to, or disconnecting from other users and about what, how and when to communicate. This is as true for Ukraine as for other countries, and online communications offer the social scientist many insights into the behavior and attitudes of the country's residents. As I mentioned in the beginning of my thesis, language use on the individual level is an important element of the complex, shifting and often ambivalent politics of national identity that have prevailed in Ukraine since independence. In general, as Fox and Miller-Idriss (2008:541) note, "language and other audible and visual cues trigger an awareness of category membership through everyday interaction".

A large body of sociological research has shown the pervasive influence of *homophily*, or assortative mixing, in social networks (McPherson, Smith-Lovin and Cook, 2001). Given the choice, people generally opt to connect to others of similar age, sex and race to themselves, and even to those in similar psychological states. More recent work has shown homophily at work in online settings too (Bollen et al. 2011).

In the case of Ukraine, almost all Internet users are able to read both Ukrainian and Russian and in many cases to write both languages without difficulty. On the other hand, research from a number of countries has shown that people are sensitive to subtle linguistic signals of identity (Giles, 1979: 255–9; Gumperz, 1982: 32–3; Woolard 1989). In this last part of my study, therefore, I want to investigate to what extent Ukrainian social media users tend to engage with others who have similar language use? In other words, do those who use mostly Ukrainian communicate more with others who also use mostly Ukrainian, and similarly for those who use mostly Russian?

Of course, homophily operates on several dimensions, so I would not expect to find Ukrainian social media users selecting their contacts solely on the basis of language use; gender, location and other shared attributes such as age and profession can also be expected to form the basis of many online communities. My aim therefore is to establish the relative importance of language use as a shared attribute forming the basis of online communities. While I once again note the complex nature of identity in contemporary Ukraine, which means that any temptation to draw simplistic conclusions about political or ethnic divides on the basis of language use data should be resisted, nevertheless analyzing the extent to which language use is forming the basis of Ukrainians' online ties is an important first step to understanding the structure and dynamics of identity formation online, as well as to understanding how political ideas, information and misinformation might spread through the country.

7.2. Analyzing homophily among Twitter users in Ukraine

Twitter was used as my data source due to the public availability of the data (unlike e.g. Facebook where many posts are only available to friends) and because Twitter users can opt to publish their location at the time of posting. This time I worked with the same dataset as in the section on gender in Chapter 6, which consisted of tweets collected between 11 April 2015 and 26 June 2016. My dataset contained 3,807,456 tweets from Twitter's Streaming API geotagged for Ukraine (including Crimea). After excluding just over one million tweets in languages other than Ukrainian and Russian (see the section below on "identifying language"). This time I worked with. I had in my database 2,738,022 tweets sent by 103,307 users.

My strategy in analyzing homophily among Ukrainian Twitter users was as follows:

1. Create a network of the connections between users.

2. Detect clusters where users have particularly dense connections to one another.

3. Examine the distribution of user attributes across these clusters in order to see which

shared user attributes seem to be the basis for the clusters.

As noted above, online users connect with each other on the basis of a large number of shared characteristics, interests and states. The limited demographic data available from Twitter restrict researchers in the attributes they can measure in their study. For this research I could only use information from Twitter regarding users' location and the language of their tweets, and estimated users' gender using an algorithm (see below). I also carried out a qualitative content analysis by reading a sample of tweets sent by users in each cluster. In future, I believe, topic modelling could be used to detect thematic differences between clusters, and algorithms might be developed to estimate other key user attributes such as age.

In this study I do not venture into aspects of homophily such as whether the shared attributes observed in clusters are the result of users choosing to connect with people with whom they share common traits (*homophilic attachment*) or of users influencing the behavior of others with whom they are already connected (*contagion*). Neither do I aim to unpack the connections within clusters in order to distinguish between each user's connections with other individual users (*pairwise assortativity*) as opposed to the user's overall pattern of connections (*neighborhood assortativity*) (Bollen et al. 2011). I intend to conduct such investigations in the future.

7.2.1. Identifying location

As I described in Chapter 4 number of tweets differed greatly depending on oblast, so dividing the tweets and dealing with the data by oblast, would make my analysis unwieldy. Because relatively few tweets were sent from some rural oblasts, I decided to group the oblasts into four regions: Kyiv City, Dnipro, the West and the South-East. The South-East category contained Crimea, Donetsk, Kharkiv, Kherson, Luhansk, Mykolayiv, Odesa, Sevastopol and Zaporizhzhya; the remaining 16 oblasts were assigned to the West category. This categorization follows the traditional division of Ukraine into West and South-East regions. Kyiv City and Dnipro were treated as separate regions, both due to their large share of tweets and their role as urban centers, attracting people from around the country.

7.2.2. Identifying gender

Gender information, along with other basic demographic data such as age or occupation, is not available for Twitter users. Therefore an algorithm for estimating users' gender based on the grammar and vocabulary used in Russian and Ukrainian tweets has been developed. This technique is described in detail elsewhere, but uses part-of-speech tagging for Russia and dictionaries for Ukrainian and Russian, to estimate the gender of the author of each tweet. Similarly to location, I then defined female users as those who sent more female than male tweets, male users as those who sent more male than female tweets, and gender-unknown users as those who sent either equal numbers of male and female tweets or no gender-specific tweets. In order to have sufficient numbers of gender-estimated tweets to decide each user's gender, I limited my dataset to the 32,243 users who sent more than 10 tweets. Of these, the algorithm tagged 18,790 as female, 9,658 as male and 3,795 as gender unknown. The gender unknown users were then removed from my dataset.

7.2.3. Identifying language

For language identification Twitter language detection algorithm has been used. Twitter language detection algorithm tags each tweet with a language when it is sent, and this language tag is returned by the API. In Chapter 3 of this thesis I showed that Twitter's tagging of Russian and Ukrainian is sufficiently accurate for use in research on language use. Based on its language identification, my whole dataset contained 2,370,496 Russian and 367,526 Ukrainian tweets.

7.3. Connections between users

Twitter users connect to each other in three main ways: through following one another, through retweeting one another's posts, and through mentioning each other by user name (e.g. user A sends a post "Hello @userB"). All three kinds of connections have been used in studies of Twitter networks (e.g. Conover et al. 2011, Bollen et al. 2011). My dataset does not currently contain follower information, and although it would be possible to obtain it for those users in my dataset whose accounts are still active, following another user's tweets does not represent an active attempt to communicate with that user. Similarly, retweeting is widely used to forward information a user finds interesting or amusing; and while this behavior can be very useful in estimating users' awareness of or opinions on certain issues, retweeting does not entail a direct communication from one user to another. Hence, of the three behaviors, mentioning denotes the most direct and intentional connection between two users; of course, the content of the communication might be either negative or positive.

In this study, therefore, I examined the ties between users who mentioned each other in their geotagged tweets. I limited the population of users for this analysis to the 32,243 users with at least ten tweets in my database because our gender detection algorithm only produced estimates for those users' gender.

Within our population of 32,243 Twitter users, 21,494 mentioned at least one other user, and I found a total of 60,799 connections between our users. Of course, the users in my dataset also mentioned many other Twitter users, but those mentioned users had not themselves sent geotagged tweets that were contained in our database so we did not include them in our network. We then reduced our network to its main component, which contained 18,085 users linked by 57,778 mention connections. Table 18. **Users by region, language use and gender.**

2 shows the central part of the resulting network visualization[13].

As Figure 12 shows, the network contains a number of clusters i.e. groups of Twitter users who mention each other particularly frequently. We used a community detection algorithm to identify these clusters. At a standard resolution of 1.0 the algorithm identified 111 clusters with a modularity score of 0.609. A modularity score of 0.609 suggests that the network is divided into quite distinct clusters. Many of the clusters contained very few nodes, so we removed all except the largest seven, each of which contained at least 294 users. After removing users who did not belong to the largest seven clusters I had 2,833 users in my network. I labelled the seven clusters A thru G.

For each of the 2,833 users remaining in our network, I counted the numbers of Ukrainian and Russian tweets they sent. The 568 users writing half or more of their tweets in Ukrainian were then classified as tweeting "mostly in Ukrainian", and the 2,265 users writing more than half of their tweets in Russian were classified as tweeting "mostly in Russian."

This two-way categorization of language use might seem at odds with the "everyday nationalism" approach which tends to place individuals into a larger number of more nuanced categories. However, it is justified by the distribution of language use: 77.8% of these users

[13] Gephi's ForceAtlas2 layout algorithm was used to visualize the network; the logic of force-directed layout algorithms such as ForceAtlas2 is that nodes that have more direct and indirect connections are positioned closer to each other. Our visualization does not show links between nodes in order to make the diagram clearer.

wrote at least 70% of their tweets in Russian, 16.6% of them wrote at least 70% of their tweets in Ukrainian, and only 5.59% used a more balanced mix of Russian and Ukrainian.

Table 18 provides descriptive statistics of the users in our network by region, language use and gender.

Region	Total	Language use	Total	Gender	Total
Kyiv City	1,352	>= 50% Uk	253	Female	127
				Male	126
		> 50% Ru	1,099	Female	532
				Male	567
Dnipro	854	>= 50% Uk	75	Female	41
				Male	34
		> 50% Ru	779	Female	449
				Male	330
West	359	>= 50% Uk	208	Female	127
				Male	81
		> 50% Ru	151	Female	66
				Male	85
South and East	268	>= 50% Uk	32	Female	10
				Male	22
		> 50% Ru	236	Female	115
				Male	121

Table 18. Users by region, language use and gender.

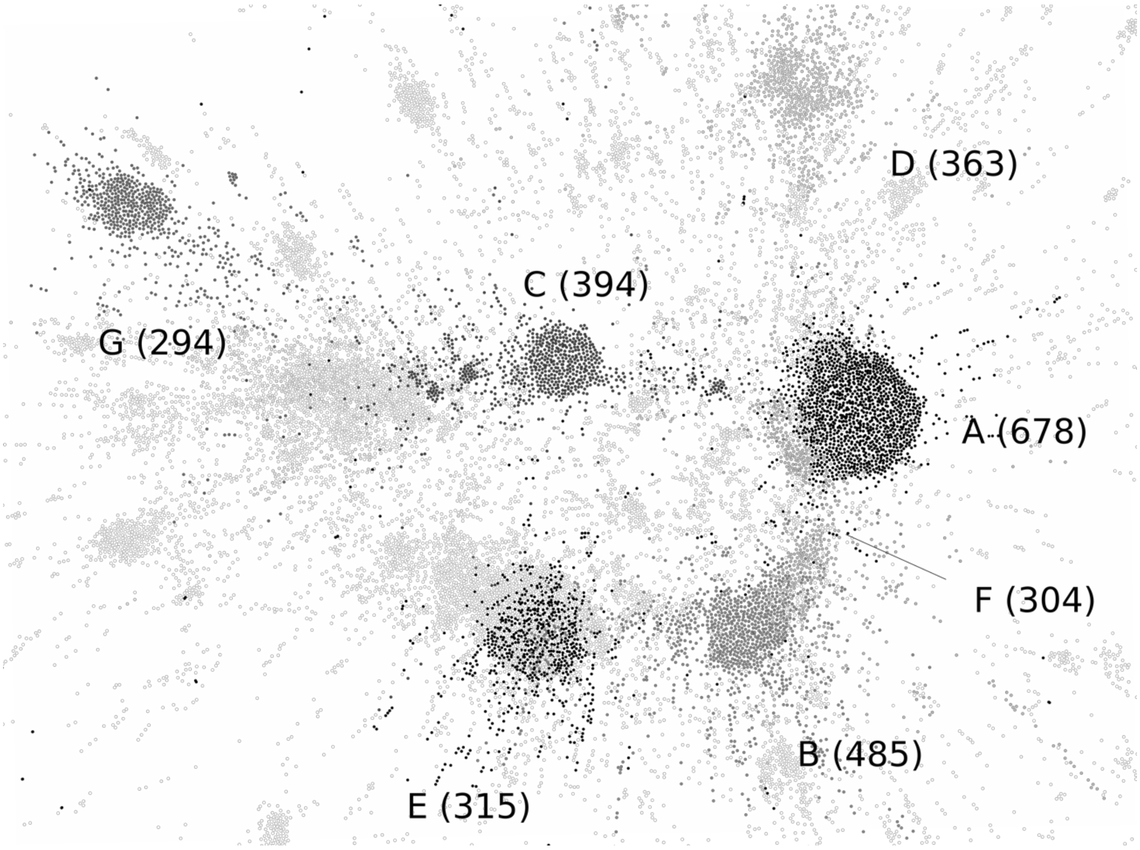

Figure 12. Structure of the mention network

Then the attributes of the users assigned to each of these seven clusters have been investigated. Figure 13, Figure 14 and Figure 15 show the breakdown by language use, location and gender for each of the seven clusters. In addition, to gain some qualitative insights into the content of tweets sent by users in the clusters, I read a random sample of 1000 tweets sent by users in each cluster. The next section provides quantitative and qualitative descriptions of each cluster.

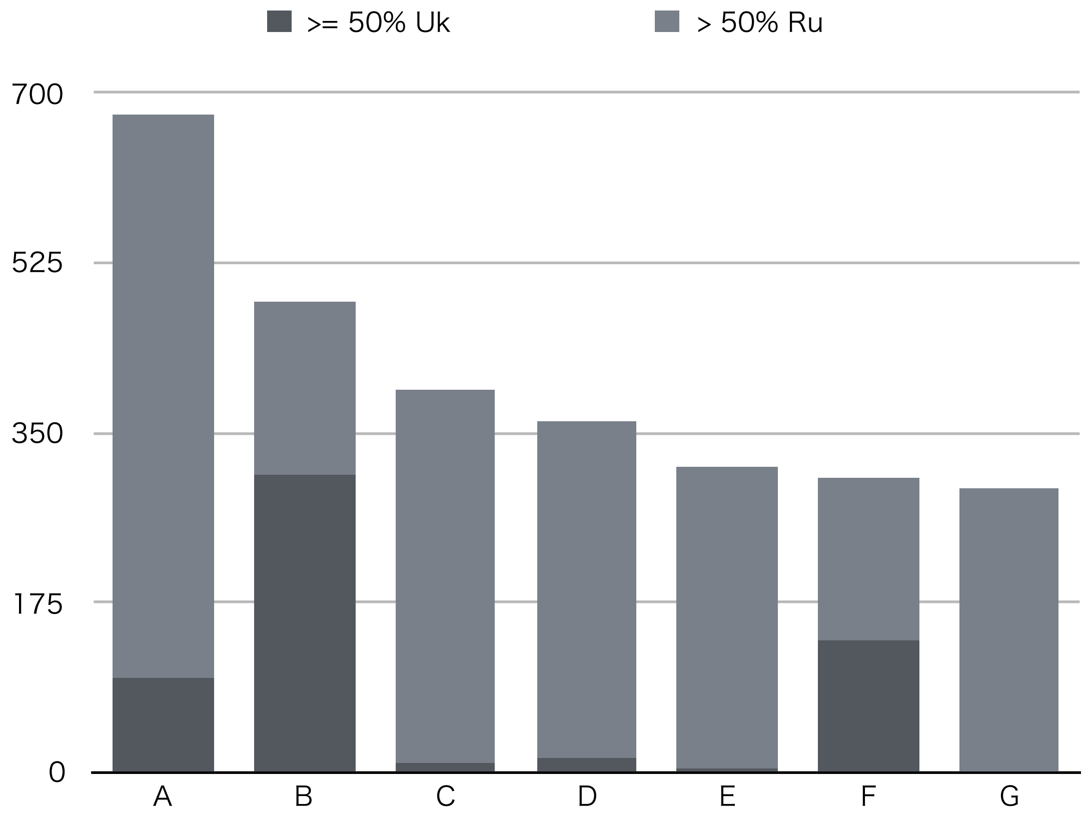

Figure 13. Language use and clusters in the mention network

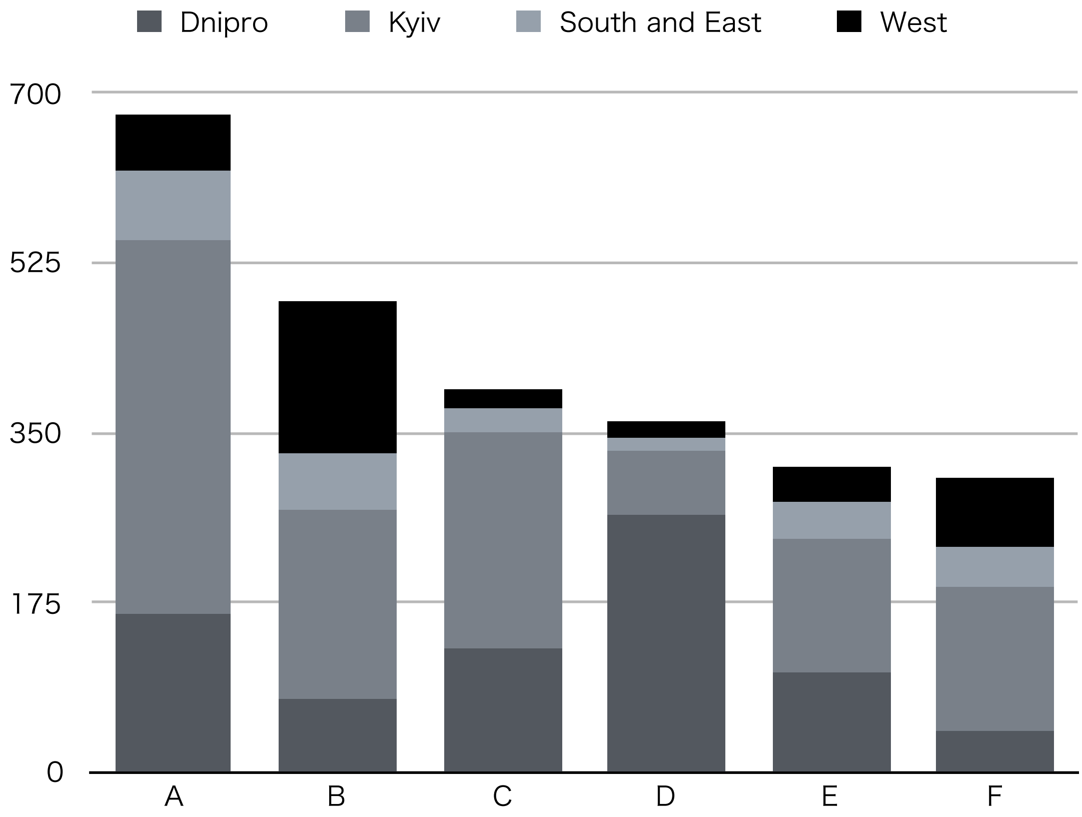

Figure 14. Location and clusters in the mention network

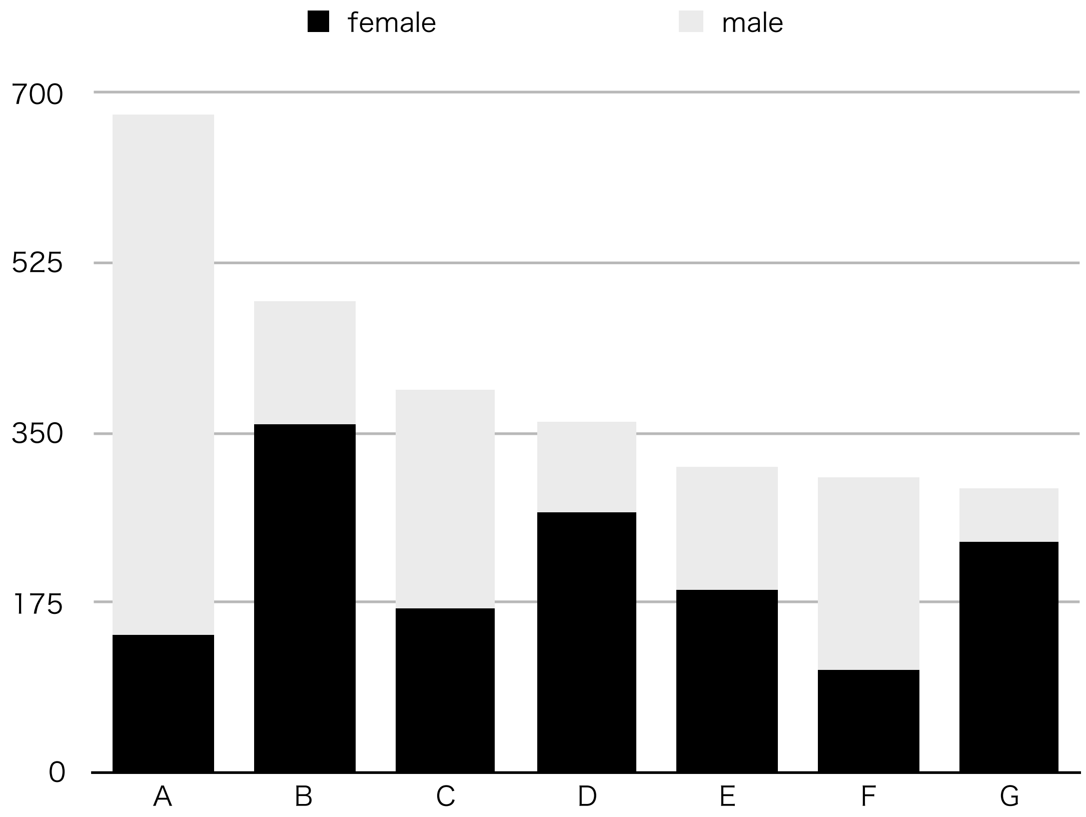

Figure 15. Gender and clusters in the mention network

7.4. Description of clusters

Cluster A: Nationwide / male / Russian / politics

The language of the users is predominantly Russian (85.5%). While as noted above female users outnumber male users in our dataset by about 2 to 1, cluster A consists mainly of male users (79.2%). Most of those who belong to the cluster tweet from Kyiv (56.6%) and Dnipro (24.3%), but some users are tweeting mostly from the South-East (10.6%) and the West (8.4%). So this cluster is nationwide and linked by language use and gender rather than location.

This cluster is characterized by a large number of tweets related to political and social news and issues in Ukraine. However, the political preferences and sympathies of the users are often opposite, e.g.

"Саакашвили готов занять премьерское кресло Яценюка (Хвиля) Как же блаженно в дурдоме" (Saakashvili is ready to occupy Prime Minister Yatseniuk's chair (Khvilya) How blissful it is in a madhouse);

"область приграничная вот в чем опасность. Надеюсь все таки что Русскую весну не попытаются там повторить." (The danger is that this oblast (region) is in a border area. I hope they are not going to repeat that "Russian Spring" here.)

"Путин готовится к майским праздникам. Россияне, готовьте себе бронежилеты" (Putin is getting ready for the May holidays. Russians, prepare for yourselves body armor.);

"либеральный идиот гавкающий из подворотни." (liberal idiot barking out of the gateway).

Cluster B: Western Ukraine / female / Ukrainian / various topics

The language of the users in this cluster is mainly Ukrainian (63.0%). Users in this cluster are mostly female (74.0%), and a relatively large proportion (32.3%) are from the Western part of Ukraine.

Tweets in our random sample were on a variety of topics such as school life, travelling, pets, food and drinks, fashion, movies and TV dramas and music.

Cluster C: Nationwide / male / mixed languages / Belarusian community in Ukraine

The tweets sent by users of this cluster often contain information related to Belarus, e.g.

“В Минске какой-то радиоактивный дождь. В Уручье пахнет какой-то гарью…” (In Minsk there is some radioactive rain. In Uruchya it smells like fumes...)

“Ура! Беларусь и Европейский Союз договорились взаимно упростить визовый режим” (Hooray! Belarus and the European Union have agreed to simplify their mutual visa regime).

Moreover, some tweets sent by users in this cluster were actually written in the Belarussian language but have been mistakenly tagged as Ukrainian or Russian by Twitter’s language detector.

The content of the tweets is diverse: daily life, politics, music, school life, friends, drinks and parties, movies, travelling, pets and so on. As for gender, we can see almost the same percentage of male and female users. Most of those who belong to the cluster tweet from Kyiv (56.7%) and Dnipro (31.9%), with smaller proportions of users tweeting mostly from the South-East (7.0%) and West (4.4%) regions of Ukraine. The high concentration of Belarusian content, including frequent mentions of Belarusian cities and towns (Minsk, Vitebsk, Gomel, Mogilev, Brest, Rogachev, etc.), suggests that many users in this cluster are Belarusians either living in Ukraine or visiting it frequently. Hence, we can describe this cluster as a Belarusian

Twitter community in Ukraine.

The prominence of this Belarusian community among Ukrainian Twitter users is perhaps surprising. According to data from Gemius[14], in 2014 the largest group of Twitter users in Belarus was between 18 and 24 years of age (34%), whereas in Ukraine the largest group of Twitter users was between 25 and 34 years of age (29%). It is possible that younger people might be more willing to let their tweets be geotagged, leading to a larger than expected number of Belarusian users in our dataset.

Cluster D: Dnipro / female / Russian / various topics

The language of the tweets is predominantly Russian (95.6%), and most of the users are female (73.9%) and based is the Dnipro region (73.0%). Reading 1000 random tweets did not reveal any particular common topics; most of the tweets were about daily matters related to school life, food, music and so on. However, the content of the tweets underscored that many users in this cluster were living in the Dnipro region, e.g.:

"Во втором туре за кресло мэра Кривого Рога сразятся Вилкул и Милобог" (Mr.Vilkul and Mr.Milobog will compete in the second round for the post of mayor of Kriviy Rig.)

"Господи, моя тетя летит из Киева в Днепр, такая погода" (Oh my God, my aunt is

[14] https://www.gemius.com/agencies-news/age-groups-of-social-media-users.html

flying from Kiev to Dnipro, and in such weather.).

Cluster E: Nationwide / female / Russian / movies and music

Movies, TV shows and music are a frequent topic of tweets sent by users in cluster E e.g.

“я только сегодня успела посмотреть Королевы крика и Готэм” Today I finally managed to see *Scream Queens* and *Gotham*

“песни ОЕ оргазм для ушей также как и песни Ширана” Songs by OE [Okean Elzy – a Ukrainian rock group] are orgasms for the ears as well as the songs of Sheeran.

Other topics are daily matters, school life, literature, etc.

99.0% of the users in this cluster tweet mainly in Russian. There is a fairly even gender balance (60% female, 40% male), and no noticeable regional concentration.

Cluster F: Nationwide / male / mixed language / various topics

This cluster is marked by a somewhat high proportion of male users (65.5%). It is the most balanced in terms of the language use of its members, with a 55-45 split between Russian and Ukrainian. This cluster is nationwide and its members tweet on diverse topics.

Cluster G: Central Ukraine / female / Russian / various topics

The members of this cluster tweet on a wide range of topics such as their interests, feelings, priorities, daily matters, school life, movies, music, pets, fashion and so on. We noticed that the tweets of users in this cluster contain more vulgar words compared to other clusters. All the users in this cluster tweet mostly in Russian, 80.6 % of them are female, and most are based in the central part of Ukraine: 66.3 % in Kyiv and 26.2% in Dnipro.

7.5. Which user attributes affect membership of clusters?

The above descriptions of the seven clusters in our network suggest that they vary along several dimensions of homophily: some have a pronounced gender bias, others a regional bias, and others are marked by greater use of either Russian or Ukrainian. In order to get a sense of how important each of these user attributes is as a basis for cluster formation, we carried out a series of multinomial logistic regressions. My dependent variable was the cluster, and my three independent variables were gender, region and language use.

Before carrying out the analysis, I made one small modification to my data: because cluster G had no users using mostly Ukrainian, in order to avoid errors when running the regression analysis I chose a single user at random from group G and changed their language use from mostly Russian to mostly Ukrainian.

Before analyzing the independent variables in combination, I first carried out multinomial logistic regressions using single independent variables in order to assess the strength of the correlation between each user attribute and cluster membership. For gender, overall pseudo-R-squared was 0.057; for region, pseudo-R-squared was 0.060; and for language use, pseudo-R-squared was 0.088. All three results were statistically significant with LLR p-values of less than <0.001. This suggests that knowing a user’s language use is more useful than knowing their gender or location when predicting which cluster they belong to.

Using both gender and language in my model raised the pseudo-R-squared to 0.12, and using all three independent variables raised the pseudo-R-squared to 0.18 (both results were statistically significant). This might seem rather low, but as Frost (2013) notes “any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%”; it also reflects the fact that different clusters are distinguished by different attributes.

7.6. Analysis and discussion

In a country where many people are functionally bilingual, especially younger ones who are also more likely to be social media users, I would expect to find communities of users who use both languages to exchange information about e.g. topics of common interest or the region they live in. In this study, I did indeed find such bilingual communities among Ukrainian Twitter users who mention each other in their geotagged tweets. However, my statistical analysis suggests that language use, which I operationalize as whether a user tweets mostly in Ukrainian or Russian, is more important than both gender and location as a basis for the formation of online communities in Ukraine.

The implication of this is that information and ideas in the digital space are likely to flow more easily among those mostly Ukrainian on the one hand, and among those using mostly Russian on the other. While it certainly not the case that the Ukrainian internet is divided into those using Russian and those using Ukrainian, any more than it is divided into male and female users, nevertheless my findings offer empirical evidence that language use is playing a major role in structuring online information exchange in the country.

My research concerning network analysis in Ukraine has many limitations. I used a single data source, geotagged Tweets. The strength of this data is that I can be confident of where each tweet was sent from, but it comes at the price of both quantity—the research ended up with a network of only 2,833 users—and quality: this being Twitter, I lacked basic demographic information about users such as their age or profession. I used a simple two-way classification of language use which obscured the (albeit relatively few) users who made frequent use of both languages. Also, cluster detection in network analysis means assigning each node (user) to only one cluster. In reality, however, people can be simultaneously members of numerous overlapping groups, possibly using a different language depending on the group. Further research using expanded data sources, and more refined methods is therefore required to see to what extent my findings represent a wider phenomenon in Ukrainian online society.

Chapter 8. Conclusions, implications and topics for future research

In this thesis I set out to answer the following research 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. Can Twitter users be considered a representative sample of society? If not, which demographics are represented among the Twitter users excessively?

5. How to identify users' general demographic characteristics such as 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/or gender?

My survey of existing research showed that the Ukrainian Internet user is characterized by a younger age and higher socioeconomic status that provides access to wireless communications. Previous research on gender and Twitter use makes no mention of a correlation between gender and location, and no relationship between user gender and language choice is expected. There should be no relationship between user gender and location: we would expect to find approximately the same number of male and female users in each region and nationwide. Twitter and other social network users may use different online contexts to strengthen different aspects of their offline connections, so their online and offline networks will overlap. As for online language use, from the findings of previous research it can be expected that Russian will be used predominantly for online communication in Ukraine especially in large cities and in the South and East of Ukraine.

However, the changing socio-political situation in Ukraine makes it important to assess the actual language use across the country. In addition, the adoption of social media is itself an important arena of language use that needs to be researched, and offers scholars the potential to study the daily language use of a large number of people across Ukraine.

In order to answer the research questions I used an original dataset of 3,807,456 tweets from Twitter's Streaming API geotagged for Ukraine (including Crimea) collected between 11 April 2015 and 26 June 2016 and analyzed the data using algorithms for language detection and gender detection. In order to support my findings on the language use on Twitter, I also undertook a limited investigation of language use on Facebook. I identified and accessed 24 Facebook pages of city and regional governments, and downloaded total of 31,370 updates and 11,044 comments posted on these pages from July 2010 to June 2016.

8.1.Conclusions

My research questions were answered as follows:

1. *What languages are preferred for online communication in Ukraine?*

The answer is clear: most of the Ukrainian population prefer to use Russian in their online interaction. My findings described in Chapter 4 show that 1,553,787 or 63.2% of the collected tweets were in Russian and only 242,829 or 9.9% were in Ukrainian. As for the 26.9% of the tweets left, English tweets accounted for 8.7%, tweets tagged as "language unidentified" - 5.7%, other languages such as Slovenian, Polish and Bulgarian - less than 3% each. My research on Facebook, as shown in Chapter 5, confirms this outcome: from my Facebook dataset, which contained a total of 12,598 updates in Ukrainian and 10,322 updates in Russian, I found that Russian was prioritized in users' responses: 5,582 comments on updates were in Russian, 3,230 were in Ukrainian.

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?*

As described in Chapter 4, the data from Twitter shows that most of the geotagged tweets were sent from Kyiv city and Dnipro oblast (45,758 and 41,662, respectively), Kharkiv (28,240) and Odesa (27,718), followed by Donetsk (26,345) constitute the top five locations in number of tweets. This means that most tweets were sent from oblasts containing cities with populations of one million or more. In general, my findings show that the proportion of Ukrainian to Russian is higher in the West. And the numbers of bilingual users are higher in the cities.

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?

I found and described the discrepancy between census data regarding reported mother tongue and the actual online behavior of Twitter users. My findings in Chapter 4 show that the language border in online networks can be drawn more centrally than both the political border that has obtained in national elections since 2004 to 2014 and the linguistic border based on reported mother tongue data from the 2001 census.

However, I realize that my findings should be treated with caution because of the danger of binary divisions, as we all know language use and national identity are much more complex than that. We should not forget that maps are political so they should be treated with extreme caution.

4. What are the demographic characteristics of Twitter users in Ukraine?

My findings show that female users outnumber male users by just under two to one and the reasons for such a great imbalance are not clear. As it was mentioned in Chapter 6, in itself this is not necessarily surprising, although it suggests that further investigation should be carried out to understand the reasons for this imbalance.

5. How to identify users' general demographic characteristics such as gender and age?

As I described in Chapter 6, I successfully developed an algorithm for identifying users' gender; however, I failed to propose any effective method of identifying users' age. In future, I will try to explore the possibilities of creating an algorithm for identifying users' age, depending on the content of their tweets.

6. Is there any relationship between bilingual use of social media and gender?

The identification of bilingual users of both genders was described in Chapter 7. While dealing with clusters of Twitter users who geotagged tweets from Ukraine, I identified a network of 2,833 users who mentioned each other, and counted the numbers of Ukrainian and Russian tweets each of them sent. The 568 (305 bilingual female and 263 bilingual male) users writing half or more of their tweets in Ukrainian were then classified as tweeting "mostly in Ukrainian", and the 2,265 (1,162 bilingual female and 1,103 bilingual male) users writing more than half of their tweets in Russian were classified as tweeting "mostly in Russian." This two-way categorization was justified by the distribution of language use: 77.8% of these users wrote at least 70% of their tweets in Russian, 16.6% of them wrote at least 70% of their tweets in Ukrainian, and only 5.59% used a more balanced mix of Russian and Ukrainian.

8.2. Discussions

This thesis has shown the viability of using Twitter and Facebook data to obtain information about the language behavior of a large number of people in their online activities. Because I am dealing with a self-selecting sample, I cannot claim that the whole Ukrainian population have higher rates of bilingualism in urban areas, however, my finding should be considered a valid source of sociological data. I also succeeded in identifying the bilingual users on Ukrainian social networks (Twitter and Facebook). The relation between the language preferences of the users of social media in Ukraine, their demographics and geographic location has been investigated. I discovered the main groups (clusters) of users existing in Ukrainian Twitter (based on the geotagged tweets), and discussed their language behavior.

In Chapter 6, I came across some puzzles, as 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 (2012), where the proportion of female and male users is almost one to one (according to Statista.com [15]).

[15] https://www.statista.com/statistics/272993/gender-distibution-of-the-global-internet-population/

Moreover, in Europe males (51%) outnumbered females (49%). It is difficult to clarify the reasons for this imbalance. I cannot even tell if it shows that there are much more female than male Twitter users Twitter in Ukraine, or that Ukrainian women more than men tend not to switch their geolocation off. 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 plausible that behavioral differences between men and women exist and female users are more likely than men to adjust their language to that of those with whom they interact on Twitter. Further analysis of my existing dataset may provide some empirical evidence for or against that possibility.

8.3. Implications and topics for future research

The discrepancy between the census data regarding mother tongue and the actual online behavior of Twitter users, which was found in my research, raises the following question: is the Ukrainian language identity of social network users in danger? As discussed in Chapter 2, online networks effectively increase the prevalence of the "dominant" language over less used languages (Cunliffe, Morris and Prys, 2013; Jones, Cunliffe and Honeycutt, 2013). My data shows that of all users tweeting in Ukrainian, more than half also tweet in Russian, whereas of all users tweeting in Russian, only one in ten used Ukrainian in their tweets. If many of those who prefer to interact in Ukrainian in their offline networks, choose the "dominant" language

(Russian) to communicate online, it will have a negative impact on the future of the Ukrainian language and the language identity of the members of Ukrainian society.

To protect and popularize Ukrainian language on TV and radio, the Ukrainian supreme legislative unit Verkhovna Rada, has made amendments to the Law of Ukraine “On Television and Radio Broadcasting”[16] ratified by the president Petro Poroshenko on June 16, 2016, concerning the quota of Ukrainian language songs on TV and radio. It states that from the date of entry into force of this Law the share of Ukrainian language songs should be at least 25% during the first year, rising to 35% in the third year; and the share of programs in Ukrainian should be at least 50% during the first year, rising to 60% during the third year. Although I understand that any similar Law concerning Internet cannot be implemented, 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 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

[16] http://zakon4.rada.gov.ua/laws/show/1421-19?test=4/UMfPEGznhh5xr.Zi2q6njoHI4Wgs80msh8Ie6

number of years. Researchers 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.

Bibliography

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*, Vol. 18 (3), pp. 213–249.

Argamon, S., Koppel, M., Pennebaker, J., Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression, *First Monday, 12.* Available: http://firstmonday.org/ojs/index.php/fm/article/view/2003/1878 [Accessed on Aug 28, 2017].

Bachmann, K., Lyubashenko, I. (2014). The role of digital communication tools in mass mobilisation, information and propaganda. In K. Bachmann. and I. Lyubashenko (Eds.) *The Maidan uprising, separatism and foreign intervention : Ukraine's complex transition*, Frankfurt: Peter Lang., pp.349-378.

Baker, C., Prys Jones, S. (1998). *Encyclopedia of Bilingualism and Bilingual Education*. Clevedon, UK: Multilingual Matters.

Baker, C (2006). *Foundations of bilingual education and bilingualism (4th ed.)*. Clevedon, UK: Multilingual Matters.

Bamman, D., Eisenstein, J., Schnoebelen, T. (2014). Gender in Twitter: Styles, Stances, and Social Networks. *Journal of Sociolinguistics,* Vol. 18, pp. 135-160.

Bernsand, N. (2001). Surzhyk and National Identity in Ukrainian Nationalist Language Ideology. *Berliner Osteuropa-Info,* Vol.17, pp. 38-47. Available:

http://www.oei.fu-berlin.de/media/publikationen/boi/boi_17/11_bernsand.pdf [Accessed on Aug 22, 2017].

Bethlehem, J., Stoop, I. (2007). Online panels. A paradigm theft? In Trotman et al. (Eds), *The Challenges of a Changing World*, Southampton, UK: Association for Survey Computing, pp. 113-131.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, Vol.78(2), pp.161-188.

Bilaniuk, L. (2005). *Contested Tongues: Language Politics and Cultural Correction in Ukraine*, Ithaca: Cornell University Press.

Blacker, U. (2014). No real threat to Ukraine's Russian speakers.*Opendemocracy.net* Available:

https://www.opendemocracy.net/od-russia/uilleam-blacker/no-real-threat-to-ukraine%E2%80%99s-russian-speakers-language-law-ban [Accessed on Sep 3, 2017].

Bloomfield, L. (1933) *Language.* London: Allen and Unwin.

Bohdanova, T. (2014) Unexpected revolution: the role of social media in Ukraine's Euromaidan uprising. *European View,* Vol. 13 (1), pp. 133–142.

https://doi.org/10.1007/s12290-014-0296-4

Bosnjak, M., Tuten, T., Bandilla, W. (2001). Participation in web surveys: A typology. *ZUMA - Nachrichten,* Vol. 48, pp. 7-17.

Bollen, J., Gonçalves, B., Ruan, G., Mao, H. (2011). Happiness is Assortative in Online Social Networks. *Artificial Life*, Vol. 17 (3), pp. 237–251.

Boyd, D., Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society,* Vol. 15 (5), pp. 662–679.

Butler, Y., Hakuta, K. (2004). Bilingualism and second language acquisition. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism,* Malden, MA: Blackwell, pp. 114-144.

Chalupa, A (2015) #digitalmaidan United Ukrainians Everywhere over Social Media. In W.Schreiber, and M. Kosienkowski, (Eds.) *Digital Eastern Europe.* Wroclaw: Kew publishing.

Cheshire, J., Uberti, O. (2014) *London: The Information Capital*, London: Particular Books.

Chin, N., Wigglesworth, G. (Eds.) (2007). *Bilingualism. An Advanced Resource Book.* Andover; New York: Routledge.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., Menczer, F. (2011). *Political Polarization on Twitter.* Fifth International AAAI Conference on Weblogs and Social Media. Available:

http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/index. [Accessed on Sep 3, 2017].

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly,* Vol. 33, pp.185-209.

Chaisty, P., Whitefield, S. (2017). Citizens' Attitudes towards Institutional Change in Contexts of Political Turbulence: Support for Regional Decentralisation in Ukraine, *Political Studies.* Available:

http://journals.sagepub.com/doi/10.1177/0032321716684845 [Accessed on Aug 29, 2017].

Cunha, E., Magno, G., Gonçalves, M., Cambraia, C., Almeida, V. (2014). He Votes or She Votes? Female and Male Discursive Strategies in Twitter Political Hashtags. *PloS ONE, 9(1)* Available:

https://doi.org/10.1371/journal.pone.0087041 [Accessed on Aug 30, 2017].

Cunliffe, D., Morris, D., Prys, C. (2013). Young Bilinguals' Language Behaviour in Social Networking Sites: The Use of Welsh on Facebook, *Journal of Computer-Mediated Communication,* Vol. 18 (3), pp. 339-361.

Couper, M. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly,* Vol. 64, pp. 464-494.

Das, M., Ester, P., Kaczmirek, L. (ed.) (2011). *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies,* New York: Routledge

Del Gaudio, S., Tarasenko, B. (2009). "Surzhyk: Topical Questions and Analysis of a Concrete Case", in J. Besters-Dilger (Ed.), *Language Policy and Language Situation in Ukraine. Analysis and Recommendations*, Frankfurt am Main: Peter Lang, pp. 327-354.

Edwards, J. (2004). Foundations of Bilingualism. In T. K. Bhatia and W. C. Ritchie (Eds.) *The Handbook of Bilingualism,* Malden, MA: Blackwell, pp. 7-31.

Fishman, J. (1977) The social science perspective. In Fishman, J. (ed.) *Bilingual education: Current perspectives, Vol. 1: Social science* Arlington: Center for Applied Linguistics, pp.14-62.

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 Aug 28, 2017].

Fox, J., Miller-Idriss, C. (2008). Everyday Nationhood. *Ethnicities*, Vol. 8(4), pp. 536-563.

Frippiat, D., Marquis, N., Wiles-Portier, E. (2010) Web Surveys in the Social Sciences: An Overview, *Population (English Edition, 2002-)*, Vol. 65 (2), pp. 285-311.

Frost, J. (2013). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? *The Minitab Blog*. Available:

http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [Accessed on Aug 31, 2017]

Galushko, K., Zorba, N. (2014). Ukrainian Facebook-Revolution? Social Networks Against a Background of Euromaidan Social research (November – December 2013) *Forum for Ukrainian Studies. A Project of the Canadian Institute of Ukrainian Studies.* Available:

http://ukrainian-studies.ca/2014/10/03/ukrainian-facebook-revolution [Accessed on Sep 31, 2017].

Genesee, F., Hamers, J., Lambert, W., Mononen, L., Seitz, M., Starck, P. (1978). Language processing in bilinguals. *Brain and Language,* Vol. 5, pp. 1-12.

Giles, H. (1979). Ethnicity Markers in Speech. In K. Scherer and H. Giles (Eds.) *Social Markers in Speech*. Cambridge, UK.: Cambridge University Press, pp. 251–89.

Gumperz, J. (Ed). (1982). *Language and Social Identity*. Cambridge, UK: Cambridge University Press.

Gunter, B. (1999). *Media Research Methods: Measuring Audiences, Reactions and Impact*. SAGE Publications Ltd.

Gorchinskaya, K. (2015). YanukovychLeaks. In W. Schreiber and M. Kosienkowski (Eds.) *Digital Eastern Europe.* Wroclaw: Kew publishing.

Haarmann, H. (1998). Multilingual Russia and its Soviet heritage. In C. Paulston and D. Peckham (Eds.) *Linguistic minorities in Central and Eastern Europe*. Clevedon, UK: Multilingual Matters, pp. 224-254.

Harp, D., Tremayne, M. (2006). The Gendered Blogosphere: Examining Inequality Using Network and Feminist Theory. *Journalism & Mass Communication Quarterly,* Vol. 83(2), pp. 247-264.

Haugen, E. (1953) *The Norwegian Language in America*. Philadelphia: University of Pennsylvania Press.

Holmberg, K., Hellsten, I. (2014). Analyzing the climate change debate on twitter -content and differences between genders. *Proceedings of the 2014 ACM conference on Web science - WebSci '14*, New York:ACM, pp. 287-288.

Holmes, J. (2008) *An Introduction to Sociolinguistics.* Harlow: Pearson Education Limited.

Huffaker, D., Calvert, S. (2005) Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication 10(2).*Available:

http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00238.x/full [Accessed on Sep 30, 2017].

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

Jongbloed-Faber, L., Van de Velde, H., Van der Meer, C., Klinkenberg, E. (2016). Language use of Frisian bilingual teenagers on social media. *Treballs de Sociolingüística Catalana,* núm. 26, pp. 27-54.

Jones, R., Cunliffe, D., Honeycutt, Z. (2013). Twitter and the Welsh language. *Journal of Multilingual and Multicultural Development,* Vol. 34 (7), pp. 653-671.

Khazaal, Y., Van Singer, M., Chatton, A., Achab, S., Zullino, D., Rothen, S., Khan, R., Billieux, J., Thorens, G. (2014). Does Self-Selection Affect Samples' Representativeness in Online Surveys? An Investigation in Online Video Game Research. *Journal of Medical Internet Research*, Vol.16 (7):e164.

Kiryukhin, D. (2015). Roots and Features of Modern Ukrainian National Identity and Nationalism. In Pikulicka-Wilczewska, A., and Sakwa, R. (eds) *Ukraine and Russia: People, Politics, Propaganda and Perspectives*, Bristol: E-International Relations, pp.57-65.

Knott, E. (2015). What does it mean to be a kin majority? Analyzing Romanian identity in Moldova and Russian identity in Crimea from below. *Social Science Quarterly*, Vol. 96 (3), pp. 830-859.

Kulyk, V. (2006). Constructing common sense: Language and ethnicity in Ukrainian public discourse. *Ethnic and Racial Studies,* Vol. 29 (2), pp. 281-314.

Kuksenok, K. (2014). Hope, Lies and the internet: Social Media in Ukraine's Maidan movement. *CMDS Working Paper 2014.2. Center for Media, Data and Society. School of Public Policy. Central European University.* Available:

https://cmds.ceu.edu/sites/cmcs.ceu.hu/files/attachment/article/689/hopeliesandtheinternet.pdf [Accessed on Oct 25, 2017].

Kuksenok, K. (2015). Multilingualism on social media in the Maidan. In W. Schreiber and M. Kosienkowski (Eds.) *Digital Eastern Europe.* Wroclaw: Kew publishing.

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

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*, Vol. 20 (2), pp. 133-162.

Kuzio, T. (2001). Identity and nation-building in Ukraine: Defining the ‘Other’. *Ethnicities,* Vol. 1, pp. 343-365.

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

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

Lakhtikova, A. (2017). Understanding Passive Bilingualism in Eastern Ukraine. *Critical Multilingualism Studies,* Vol.5 (1), pp.144-173.

Lambert, W. (1977). The effects of bilingualism on the individual: Cognitive and sociocultural consequences. In P.A. Hornby (Ed.), *Bilingualism: Psychological, social, and educational implications,* New York: Academic Press, pp. 15-27.

Landis, J., Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics,* Vol. 33(1), pp. 159-174.

Landry, R., Allard, R. (1993). Beyond socially naive bilingual education: the effects of schooling and ethnolinguistic vitality of the community on additive and subtractive bilingualism. *Annual Conference Journal (NABE’90–’91)*, pp. 1-30.

Laruelle, M. (2015). The “Russian World”: Russia’s Soft Power and Geopolitical Imagination, *Washington, D.C.: Center on Global Interests*, Available:

http://globalinterests.org/wp-content/uploads/2015/05/FINAL-CGI_Russian-World_Marlene-Laruelle.pdf [Accessed on Sep 21, 2017]

Mackey, W. (1962). The Description of Bilingualism. *The Canadian Journal of Linguistics,* Vol. 7, pp. 51-85.

Macnamara, J. (1967). The bilinguals linguistic performance: a psychological overview. *Journal of Social Issues*, Vol. 23, pp. 59-77.

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., Rodrigue, J. (2012). A Demographic Analysis of Online Sentiment During Hurricane Irene. *Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics*, pp. 27-36.

Malmkjær, K. (Ed.) (2010). *The Routledge Linguistics Encyclopedia* (3rd ed.) London and New York: Routledge.

Masenko, L. (2009). Language situation in Ukraine: Sociolinguistic analysis. In J. Besters-Dilger (Ed.) *Language Policy and Language Situation in Ukraine: Analysis and Recommendations*. Frankfurt am Main: Peter Lang, pp. 101-137.

McPherson, M., Smith-Lovin, L., Cook, J. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, Vol. 27 (1), pp. 415-444.

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

Mohanty, A., Perregaux, C. (1997). Language acquisition and bilingualism. In J. Berry, P. Dasen and T. Saraswathi (Eds.), *Handbook of Cross-Cultural Psychology.* Vol.2, *Basic Processes and Human Development*, Boston: Allyn and Bacon, pp. 217-253.

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

Onuch, O. (2014). The Middle Class Median Protester: EuroMaidan and Democratization in Ukraine. *Journal of Democracy,* Vol.25 (3), pp. 44-51.

Onuch, O. (2015). EuroMaidan Protests in Ukraine: Social Media Versus Social Networks. *Problems of Post -Communism*, Vol. 62, pp.1-19.

Onuch, O. (2015). "Facebook Helped Me Do It" Understanding The EuroMaidan Protester 'Tool-Kit'. *Studies in Ethnicity and Nationalism,* Vol. 15 (1), pp. 170-184.

Pavlenko, A., Blackledge, A. (Eds.) (2004). *Negotiation of identities in multilingual contexts*, Clevedon, UK: Multilingual Matters.

Pavlenko, A. (2008). Russian in Post-Soviet Countries. *Russian Linguistics*, Vol. 32 (1), pp. 59-80.

Peal, E., Lambert, W. (1962). The relation of bilingualism to intelligence. *Psychological Monographs,* Vol. 76, pp. 1-23.

Pearce, W., Holmberg, K., Hellsten, I., Nerlich, B. (2014). Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report. *PLoS ONE* Vol. 9 (4): e94785. doi:10.1371/journal.pone.0094785.

Pennycook, A. (2003). Global Englishes, rip slyme, and performativity. *Journal of Sociolinguistics,* Vol. 7 (4), pp. 513-533.

Pentina, I., Basmanova, O., Zhang, L. (2014). A cross-national study of Twitter users' motivations and continuance intentions. *Journal of Marketing Communications* (doi:10.1080/13527266.2013.841273).

Petro, N. (2015). Understanding the other Ukraine: Identity and allegiance in Russophone Ukraine. In A. Pikulicka-Wilczewska and R. Sakwa (Eds.) *Ukraine and Russia: People, Politics, Propaganda and Perspectives*, Bristol: E-International Relations, pp.18-34.

Pirie, P. (1996). National identity and politics in southern and eastern Ukraine. *Europe-Asia Studies,* Vol. 48 (7), pp. 1079-1104.

Plohy, S. (2015). *The Gates of Europe: A History of Ukraine*. New York: Basic Books. Available:

http://shron.chtyvo.org.ua/Plokhii_Serhii/The_Gates_of_Europe__A_History_of_Ukraine_anhl.pdf [Accessed on Sep 24, 2017].

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

R&B group (2010). 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». Available:

http://rb.com.ua/rus/projects/omnibus/5078/ [Accessed on Oct 12, 2017].

Ritchie, W., Bhatia, T. (2004). Social and Psychological Factors in Language Mixing. In T. K. Bhatia and W. C. Ritchie (Eds.) *The Handbook of Bilingualism,* Malden, MA: Blackwell, pp. 336-352.

Ronzhyn, A. (2014). *The use of Facebook and Twitter during the 2013–2014 protests in Ukraine.* Proceedings of the European Conference on Social Media: ECSM, 2014, Reading, UK: Academic Conferences Ltd. Available:

https://www.researchgate.net/publication/268979057_The_Use_of_Facebook_and_Twitter_During_the_2013-2014_Protests_in_Ukraine [Accessed on Oct 1, 2017].

Ronzhyn, A. (2016). *Social Media Activism in Post-Euromaidan Ukrainian Politics and Civil Society.* International Conference for E-Democracy and Open Government 2016, Conference Paper.

Romaine, S. (1989). *Bilingualism*. Oxford: Basil Blackwell.

Russkiy Mir: "Russian World" On the genesis of a geopolitical concept and its effects on Ukraine DGAP Berlin Expert Round Table. A conversation with the Ukraine expert DGAP’s Wilfried Jilge (2016). Available:

https://dgap.org/en/node/28188 [Accessed on Sep 21, 2017]

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., Rodrigue, J. (2012). A Demographic Analysis of Online Sentiment During Hurricane Irene. *Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics*, pp. 27-36.

Skvirskaja, V. (2009). “Language is a Political Weapon” or on Language Troubles in post-Soviet Odessa. In J. Besters-Dilger (Ed.) *Language Policy and Language Situation in Ukraine: Analysis and Recommendations*. Frankfurt am Main: Peter Lang, pp. 175-200.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., Rana, O. (2013). “Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter.” *Sociological Research Online,* Vol. 18(3), article number: 7 (doi: 10.5153/sro.3001).

Sloan, L., Morgan, J., Burnap, P., Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *Plos One,* Vol. 10 (3), article number: e0115545.

Sloan, L., Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *Plos One,* Vol. 10 (11), article number: e0142209.

Soedjono, A. (2012). The Comparisons Between the Language Used by Male and Female Peers in Twitter. *Anglicist ,1,* pp. 1-6.

Søvik, M. (2007). *Support, resistance and pragmatism: An examination of motivation in language policy in Kharkiv, Ukraine*. Stockholm: Acta Universitatis Stockholmiensis.

Søvik, M. (2010). Language Practices and the Language Situation in Kharkiv: Examining the Concept of Legitimate Language in Relation to Identification and Utility. *International Journal of the Sociology of Language*, pp. 5-28.

State Statistics Committee of Ukraine. All-Ukrainian population census 2001. Всеукраїнський перепис населення 2001/English version/Results/General results of the census/ Language. Available:

http://2001.ukrcensus.gov.ua/eng/results/general/language/ [Accessed on Aug 20, 2017].

Subrahmanyam, K., Reich, S., Waechter, N., Espinoza, G. (2008). Online and offline social networks: Use of social networking sites by emerging adults, *Journal of Applied Developmental Psychology,* Vol. 29 (6), pp.420-433.

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., Takayasu, H. (2015). "Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study. *PLoS ONE, Vol*. 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.

Trach, N. (2009). Language Policy and Language Situation in the Sphere of Legal Proceedings and Office Administration in Ukraine. In J. Besters-Dilger (Ed.), *Language Policy and Language Situation in Ukraine. Analysis and Recommendations*, Frankfurt am Main: Peter Lang, pp. 287-326.

Valenzuela, S., Valdimarsson, V., Egbunike, N., Fraser, M., Sey, A., Pallaev, T., Chachavalpongpun, P., Saka, E., Lyubashenko, I. (2014). The Big Question: Have social media and/or smartphones disrupted life in your part of the world? *World Policy Journal,* Vol. 31(3), pp. 3-8.

Venkatesh, V., Morris, M. (2000). Why Don't Men Ever Stop to Ask for Directions? Gender, Social Influence, and Their Role in Technology Acceptance and Usage Behavior. *MIS Quarterly,* Vol. 24(1), pp. 115-139.

Weinreich, U. (1953). *Languages in Contact.* The Hague: Mouton.

Wilson, C., Sala, A., Puttaswamy, K., Zhao, B. (2009). *User interactions in online social networks and their implications*. Proceedings of the 4th ACM European conference on Computer systems, pp. 205-218.

Woolard, K. (1989). *Double Talk: Bilingualism and the Politics of Ethnicity in Catalonia*. Stanford, CA: Stanford University Press.

Yasna, I. (2015). Ukraine – 2014: Which Way Will the Digitalization Pendulum Swing? In Schreiber, W., and Kosienkowski, M. (Eds.) *Digital Eastern Europe.* Wroclaw: Kew publishing.

Zeitzoff, T., Kelly, J., 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*, Vol. 52 (3), pp. 368–383.

Zalizniak, H. (2009). Language Orientations and the Civilisation Choice for Ukrainians. In J. Besters-Dilger (Ed.), *Language Policy and Language Situation in Ukraine. Analysis and Recommendations*, Frankfurt am Main: Peter Lang, pp. 139-174.

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

Sources in Ukrainian:

Ажнюк, Б. (2008) “Шляхи і методи розширення сфери застосування української мови: концептуальні й практичні аспекти” с.343-366, *в О.Майборода (ред.) Мовна ситуація в Україні: між конфліктом і консенсусом* ІПіЕНД імені І.Ф.Кураса НАН України. Available:

http://www.ipiend.gov.ua/img/monograph/file/movna_sit_49.pdf [Accessed on Oct 21, 2017].

ЗАКОН УКРАЇНИ «Про внесення змін до деяких законів України щодо частки музичних творів державною мовою у програмах телерадіоорганізацій» (Відомості Верховної Ради (ВВР), 2016, № 31, ст.547). Available: http://zakon4.rada.gov.ua/laws/show/1421-19?test=4/UMfPEGznhh5xr.Zi2q6njoHI4Wgs80msh8Ie6 [Accessed on Sep 29, 2017].

ЗАКОН УКРАЇНИ «Про засади державної мовної політики» (Відомості Верховної Ради (ВВР), 2013, № 23, ст.218). Available: http://zakon4.rada.gov.ua/laws/show/5029-17 [Accessed on Sep 29, 2017].

Хмелько, В. (2004) "Лінгвоетнічна структура україни: регіональні особливості й тенденції змін за роки незалежності." *Наукові записки НаУКМА 32*. Соціологічні науки, c. 3–15. (Khmelko, V. (2004) Linhvo-etnichna struktura Ukrainy: rehional'ni osoblivosti ta tendentsii zmin za roky nezalezhnosti).

Sources in Russian:

Бурыкин, А.А. (2006) Ментальность, языковое поведение и национально-русское двуязычие. Санкт-Петербург. Available:

http://abvgd.russian-russisch.info/articles/10.html [Accessed on Aug 28, 2017].

Губогло, М.Н. (1984) автореферат диссертации по истории, диссертация на тему: Этносоциальный аспект развития национально-русского двуязычия в СССР, доктора исторических наук. Available:

http://cheloveknauka.com/etnosotsialnyy-aspekt-razvitiya-natsionalno-russkogo-dvuyazychiya-v-sssr#ixzz4ata4aNPf [Accessed on Aug 29, 2017].

Цамерян, И.П. (1979) *Нации и национальные отношения в развитом социалистическом обществе.* Москва:Наука.