University of Groningen

Using social media to measure language use

Zhuravleva, Anastasia; de Bot, Kees; Hilton, Nanna

Published in:
Journal of Multilingual and Multicultural Development

DOI:
[10.1080/01434632.2015.1111894](https://doi.org/10.1080/01434632.2015.1111894)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Zhuravleva, A., de Bot, K., & Hilton, N. (2016). Using social media to measure language use. *Journal of Multilingual and Multicultural Development*, 37(6), 601-614.
<https://doi.org/10.1080/01434632.2015.1111894>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Using social media to measure language use

Anastasia Zhuravleva, Kees de Bot & Nanna Haug Hilton

To cite this article: Anastasia Zhuravleva, Kees de Bot & Nanna Haug Hilton (2016) Using social media to measure language use, Journal of Multilingual and Multicultural Development, 37:6, 601-614, DOI: [10.1080/01434632.2015.1111894](https://doi.org/10.1080/01434632.2015.1111894)

To link to this article: <https://doi.org/10.1080/01434632.2015.1111894>

Published online: 18 Dec 2015.

Submit your article to this journal [↗](#)

Article views: 809

View related articles [↗](#)

View Crossmark data [↗](#)

Citing articles: 1 View citing articles [↗](#)

Using social media to measure language use

Anastasia Zhuravleva^{a*} , Kees de Bot^a and Nanna Haug Hilton^b

^aDepartment of Applied Linguistics, University of Groningen, Groningen, The Netherlands; ^bDepartment of Frisian Language and Literature, University of Groningen, Groningen, The Netherlands

ABSTRACT

Investigations of language use in multilingual regions are traditionally done through the usage of retrospective questionnaires, either distributed on paper, or digitally. The aim of the present study was to test a new method of measuring language choice and use: by using social media to contact individuals several times a day and ask them what language(s) they are using, how, and with whom. It was investigated whether this method could provide real-time data on an individual's language use. One hundred participants living in the Netherlands were given a language background questionnaire and were asked to provide information through Facebook®. It was found that the social media data generally reflect the questionnaire data. More importantly, the social media data include more exact information on cases of codemixing and allow to compile a typology of language contact situations. It was concluded that social media inquiries could prove useful in future language use studies. Methodological issues, in particular concerning privacy and problems related to crowd sourcing of language data through social media, are discussed.

ARTICLE HISTORY

Received 3 February 2015

Accepted 13 October 2015

KEYWORDS

Language use; language contact; social media; methodology; questionnaire

Measuring language contact

Research concerning the language behaviour of people in multilingual environments is of interest to researchers from fields as diverse as neuro- and psycholinguistics to sociolinguistics and fields concerned with language planning and policy. Naturally, researchers want to know what happens in real life and obtain data that reflects the objective reality of multilingual language competence. However, most existing methodologies for investigating multilingual language use employ retrospective self-report questionnaires, which provide the researchers with a subjective and time-delayed report made by the participants themselves of their language behaviour.

Traditional questionnaires used in language use and choice studies elicit information on socioeconomic background, age, gender, country of birth combined with information about the amount of contact informants have with different languages in different modalities and via different media. This is often combined with questions about conscious attitudes that informants hold towards languages, or language learning (for examples of language surveys, see Hoffman 1934; Rees 1954; Seliger and Gingras 1976; Seliger 1977; Bonnet 2002; Extra and Yağmur 2004; Berns, de Bot, and Hasebrink 2007; Lee, Browne, and Kusumoto 2011). We mention Gullberg and Indefrey's (2003) Language Background Questionnaire (henceforth LBQ) separately here, as it serves as the model for the questionnaire employed in the current investigation. The LBQ concerns the informants'

CONTACT Anastasia Zhuravleva amzhuravleva@gmail.com

^{*}Currently at Lomonosov Moscow State University, Moscow, Russian Federation.

background and history of language learning, as well as the amount of contact they have with various languages in different modalities/media.

Most methodologies for investigating language choice and use in multilingual communities, including the ones mentioned above, have in common the fact that they focus on reported data in a retrospective fashion, instead of monitoring language usage in real time. The usage of retrospective questionnaires in language surveying has received a fair amount of criticism in literature. Holmes (1997) offers an extensive review of the limitations of questionnaire surveys, among them the lack of depth that surveys offer, but also the inaccuracies and unreliable answers that the approach can bring with it. She points to the discrepancies that exist between different cultures' views of what constitutes a real language as potentially problematic. Another complication lies in the role that the social status of a language can play for its speakers to willingly report proficiency and usage (Holmes 1997). A similar point is made by Martin-Jones (1991, 51), who remarks that informants of immigrant backgrounds in western societies may idealise their usage of their home languages and over-report their usage. Milroy, Wei and Moffatt (1991) suggest that surveys of language use in multilingual communities should rely more on ethnographic methods, and combine structured approaches with more pragmatic ones. Baker (1997, 35) also emphasises that the surveying of language use in multilingual communities can be done through systematic or structured observations. It is the combination of a questionnaire with a systematic self-observation technique that we put to the test here in this article.

Use of social media in research

With the rise of social media, and more widespread use of the Internet in general, new possibilities present themselves to researchers interested in language use in multilingual communities. In the present study, Facebook® was chosen over other social media because it was thought that people today spend more time on it than on any other social networking service. As a social networking platform, Facebook® provides instant access to a large 'database' of people from all over the world. In 2013, the website boasted 1.11 billion users in total and 665 million daily active users, according to Facebook.com. Much research has been done on how and why people use Facebook® (Di Capua 2012). Mostly, people report using the platform for communicative purposes as well as for maintaining relationships.

Wolfram writes (in his blog, *Stephen Wolfram Blog*, Data Science of the Facebook World, on April 24, 2013) that clear clusters can be distinguished among a user's Friends. For example, users as young as 13 years old show clusters of Friends from school, family, and the neighbourhood. The number of these clusters increases with age, when university and the workplace, among others, are added. Ugander et al. (2011) find that users have mostly Friends from within their own country. Thus, it can be seen that Friends on Facebook® are grouped not only according to age, but also according to domains, using the language maintenance term, and geographically. Finally, Park, Kee and Valenzuela (2009) found that Facebook® plays a significant role in involving the younger population in social issues and politics by facilitating communication among people with common interests. In other words, users can also group themselves and their Friends according to common interests and beliefs.

Researchers have also looked at Facebook® not as a *subject* of research, but as a *tool* for research. By looking not only at the users' profiles on Facebook®, but also by tracing their activity (Likes, Timeline posts, and Status Updates), researchers have the opportunity to reach their desired demographic without ever leaving the office (van Gilder Cooke 2011). Facebook® has been found to be an effective tool for gathering participants from 'hard-to-reach' populations (for a summary of the main characteristics of these populations, see Marpsat and Razafindratsima 2010) and for distributing web links to online surveys (Baltar and Brunet 2012). However, it seems Facebook® is generally only used to contact the potential participants, while online surveys are distributed through other, specialised survey websites.

While Facebook® has an ‘encyclopedic stash of personal information’ (van Gilder Cooke 2011), it has also been – for much the same reason – a focus of constant controversy where the issues of privacy are concerned. The privacy policy of the social network is subject to frequent change, often leaving users confused about how to adjust their privacy settings. Quite often, Facebook® users may not even be aware that they are the subjects of research (van Gilder Cooke 2011). This means that researchers embarking on studies which employ Facebook®, or any other social media sites for that matter, face many ethical questions. Markham (2006) and boyd (2007) both address issues of privacy on social media sites in their work. Sometimes what a researcher perceives as a public space (e.g. a chatroom or a Facebook® group) may be perceived as a private space by the users of the site (Markham 2011). Consequently, users could see a researcher’s involvement as an intrusion into their private lives.

There is also the question of how to gain informed consent in a digital space. Furthermore, it is difficult to verify the identity (e.g. age, gender, and even interests) of potential participants when conducting Internet research (Markham 2006). In other words, before conducting a study, it is yet unclear how a researcher can be sure that their informant is of age, or whether the people they are interviewing in a chatroom are the same people who gave permission to be interviewed a month ago. The ethical boundaries of Internet research are under constant discussion and scrutiny. We refer to Markham and Buchanan (2012) for a further discussion and some guidelines concerning ethics for researchers wishing to conduct a study using the Internet.

Using social media to measure language contact

The present study aims to test a methodology that would facilitate the task of both the participant and the researcher by making use of social media – in this case, Facebook®. In contemporary society, people spend quite a lot of time online and especially on social media, usually accessing it several times a day. Since Internet has made the problem of distance obsolete, making it easier to communicate not only across distances, but also in real time without any delay in communication, a researcher can simply ask people several times per day over social media what language they are using at the moment and thus compile a picture of their language behaviour.

An experiment was conducted in the present study featuring two methods of data collection: the social media method (henceforth SMM) and the self-report questionnaire. All participants took part in both methods. The SMM data were first analysed independently, and then in comparison with the questionnaire. The aim of the study was to see whether SMM can be used in such studies, whether the data were informative, and whether there were any significant differences between the data elicited on social media and the data elicited using questionnaires.

Methodology

The current project had two phases. In the first, a pilot study was carried out to test the SMM methodology on a smaller scale and to acquaint the students of the University of Groningen with the experiment, since they would be collecting the data for the main study. On the basis of the findings in the pilot, the instruments and set up were adapted for the main study. The collection of data for the main study were conducted on 27 March –13 April 2013.

Participants¹

Each student who volunteered to contribute to the collection of data for the main study was asked to find from 6 to 10 participants among their Facebook® friends. The only criterion controlled for in the selection of the participants was place of living, which was chosen as the Netherlands. The students’ final grade for one of their courses depended for 5% on their participation in the present study and their reports on it.

The participants' background information was gathered by means of the first section of the (adapted) LBQ used in the study. There were 100 participants: 71 female and 29 male. They were between ages 14 and 62 ($M = 23$; $SD = 7.5$). Ninety-seven participants were born in the Netherlands, three – outside of the Netherlands. Eighty participants indicated Dutch as their native language, 16 indicated both Frisian and Dutch,² 1 indicated English, 2 – German, 1 – Bildts.³ Participants indicated knowledge of 1–4 foreign languages per individual ($M = 2.7$; $SD = 0.92$; Mode = 3). The following languages were mentioned: English (99 participants), German (77), French (54), Spanish (12), Frisian (11), Italian (4), Dutch (3), Portuguese (2), Swedish (2), Latin (2), Finnish (2), Greek (1), Limburgish (1), Hungarian (1), and Norwegian (1). Each student randomly divided their participants into two groups, A and B, by arranging their names in a list and alternatively assigning them to a group. As a result, Group A had 54 participants, Group B – 46 participants. The participants in the A-group did the SMM data collection first and then the self-report questionnaire, the participants in the B-group started with the questionnaire and then did the SMM part. In Chi-square tests the differences between the results obtained from the two groups (A and B) concerning language use in activities in the SMM data and the questionnaire data were not significant, $\chi^2(1) = 0.143$, $p = .71$.

Materials

The edited questionnaire, based on the LBQ by Gullberg and Indefrey (2003), contains questions concerning the background of the participants (education, job, etc.), their proficiency in various skills of all the languages they know, details on what languages they use with whom (family members, colleagues, etc.), in what activities (watching TV, reading, etc.), and for how long. Questions concerning speakers' attitudes to their languages and learning these languages, which are present in the original LBQ, were not included in the version used for this study because they were deemed irrelevant to the present study. The edited version was also translated into Dutch for those participants who felt they were not proficient enough in the English language to fill in the English version.

In the SMM part, participants were sent three questions over the Facebook website's messaging service:

- (1) What language(s) are you using?
- (2) How? (Interaction, class, film, book, music, etc.)
- (3) If interaction, with whom?

Each data collection prompt⁴ included all three questions. It was decided to include only three questions so as to decrease the time a participant would need to answer these questions, in order not to demotivate them. Interaction was perceived as communication between the participant and another person. Detailed instructions were included in the first message sent to participants (see Appendix 1).

Procedure

Each student collected the data independently (the researcher was available through e-mail and mobile phone to answer any questions that arose). They were presented with the schedule, the materials, the time frames, and a coding template. In response to their concerns that they may not be able to send a data collection prompt strictly according to the schedule because of being in class or out of range of a computer, it was decided to allow them a range of 1 hour around the scheduled data collection to complete this. The students first asked their A-group participants to fill in the edited LBQ through SurveyMonkey.com. After that, the students sent the data collection prompts over the Facebook® messaging service according to the following schedule: Wednesday and Friday – 10:00, 14:00, 23:00; Saturday – 10:00, 14:00, 23:00, 01:00. This was done over the course of

three weeks (so nine days of data collection in total). Participants were asked to answer the questions in the data collection prompts according to what was happening at the time they *read* the prompts, not at the time the prompts were *sent* (so not retrospectively). Finally, the students' B-group participants were asked to fill in the questionnaire.

The students then coded the results into an Excel coding template (see Appendix 2). The participants usually provided full sentence answers (e.g. 'I'm using Dutch and talking with my classmates'), which were then coded according to the template. The students' individual coding templates were collected by the researcher and combined into one document for ease of analysis.

Not all of the data that was provided by the students could be used for analysis. The results provided by the students were filtered and participants were selected according to the following criteria: (1) all planned data collection prompts were sent by the student; (2) the questionnaire was filled in completely; (3) the questionnaire was filled in on time (Group A – 26–28 March; Group B – 13–16 April). Out of 281 initial participants, only 100 were retained for data analysis.

Methodological concern I: privacy on social media

One of the problems encountered during the study concerned privacy in the SMM. In designing the method, it was made sure that the information asked for would not infringe upon the participants' privacy. For example, it was decided early on that they would not be asked what topic (e.g. school, work, and romantic) they were interacting on. In the pilot study, the researcher sent out the questions used in the SMM to the students of the University of Groningen and the latter responded to these over the course of one week. For this pilot study, the students were asked to add the researcher as a Friend on Facebook® and many of them were reluctant to mix their social and educational environments, concerned that the researcher would have access to their private information and news feed. Only after the students were reassured that the researcher would not pry into these areas did they comply with the request. Such a problem was not encountered in the main study because the students gathered participants from their existing Facebook® friends.

The issue of privacy is relevant concerning the employment of the SMM over any social media site/application. In social media such as WhatsApp®, which are more like smartphone messaging services, there are no profiles containing information which participants may not want a researcher to see, only their phone number and status. One way the issue of getting access to the informants on Facebook® or Twitter® (or other social media with profile pages/news feeds) could be resolved is through a signed agreement that the researcher would not explore the participants' profiles and only use the messaging service to conduct the survey. It is also possible to employ the messaging function in Facebook® without becoming Friends with the research participants. There is an option to send direct (private) messages to participants on Twitter® as well.

Another way is through the adjustment of privacy settings on social media. Participants could be instructed to set their privacy settings so that the researcher would not be able to see any updates from them, nor anything on their timelines or profiles, and would only be able to send/receive messages. However, participants could be demotivated by the amount of 'work' they have to do to take part in this study: not only do they have to answer surveys and questionnaires, they also have to change their privacy settings in their social media profile.

Methodological concern II: crowd sourcing of data on social media

During the course of the study, it became apparent that the SMM of collecting data required a great deal of motivation not only from the participants, but also from the many students involved in the study. The students were not monitored closely by the researcher every day of data collection, so the diligence with which they collected the data depended on them. In the end, some of the students failed to send data collection prompts for reasons which were not always objective (e.g. 'I wanted to go to sleep'). As a consequence, the data collected by these students were not reliable and

could not be used. If this method were to be used in the future, it would have to be ensured that the people recruiting participants understand that all collection prompts have to be sent. As an alternative, the whole SMM study could be conducted by one researcher only or through an application.

A few things must be said not only about the students' motivation, but motivation on the part of the participants. When the method of the experiment was designed, it was kept in mind that participants might simply become annoyed at being sent multiple messages per day and stop replying or ask to be removed from the experiment. For this reason, it was decided to send three to four data collections per day and sufficiently spaced apart, so as to decrease this risk. While the majority of the participants answered regularly, there were two recorded instances of a participant explicitly asking to be removed from the project and six cases of when participants simply stopped replying mid-way through the second week of the experiment, although they filled in the questionnaire and did not ask to be removed from the project. This is interpreted as a sign of 'annoyance' because the unanswered collection prompts were not distributed evenly throughout the participants' data sets, but clustered in the second half. All these six participants were in Group B. They were not contacted post-experiment regarding the reason why they stopped answering.

All the participants of the current study were Friends of the students. Naturally, this made the gathering of participants less difficult: many of the participants may have agreed to take part in the study as a favour to the student. On the other hand, the resulting sample is not representative of the overall population in the Netherlands. This brings up the question of how effective – and even possible – it would be to gather participants through social media from outside one's Friends or contacts list, and whether this sample would be representative of the overall population.

The Netherlands is a country which has remarkable linguistic diversity. Thus, a study in language use would yield interesting results. If the SMM were employed in such a study on a larger scale than in the present one, then it would be necessary, in our opinion, to control more strictly for the geographical location of the participants. The presence of Frisian and its vitality in the present study owes much to the fact that the study was conducted in the north of the Netherlands. Perhaps if it were conducted in the south, Limburgish would be one of the languages often recorded. It may be feasible to conduct several independent studies using the SMM in different Dutch cities and then to combine the results for analysis to attain a complete and representative view of multilingual language use throughout the Netherlands.

Concerning the large immigrant population in the Netherlands and the study of the use of immigrant languages, one option would be to reach out to this population through schools. Teenagers usually have a large presence on social media and they could involve their families in the study. Naturally, to set up this experiment it would be only ethical to meet with the families and their children face to face to explain everything and to obtain permissions from schools and/or parents or guardians.

Crowdsourcing data and gathering a representative sample through social media without meeting face to face may be feasible. When investigating, for example, language shift in a society, a researcher can contact the target population (e.g. speakers of an endangered language) through, for example, a Facebook® group or a public page based on a common interest in the target language and ask them to also contact any Friends of theirs who speak the language, thus employing the snowball sampling technique. In other social media, where public groups/pages such as those mentioned above do not exist (e.g. WhatsApp® and Twitter®), a researcher would most likely have to find participants either in real life (for example, at a school/university), or by browsing their social media profiles/feeds and then be referred by them to further participants – thus again employing the snowball sampling technique. A researcher may also find their target population by browsing through Twitter® hashtags connected with languages.

The following sections contain the results of the main study and a discussion of these results in order to provide future researchers with a picture of the kind of data which can be obtained through the SMM.

Results of the main study

Amount of data obtained with the SMM

The effectiveness of the SMM was determined through a quantitative analysis of the number of data collection prompts in total and the percentages of answered/unanswered collection prompts. There were 3000 collection prompts in total, the majority of which (2371) were answered. This left 629 unanswered collection prompts. Each participant received 30 prompts. The number of unanswered prompts varied per individual, from 0 to 25; on average, 6.29 prompts per individual were not answered. Fifty-one percent of the participants missed five or fewer prompts, 81% missed nine or fewer prompts.

Descriptive analysis of the SMM data

The languages recorded in the SMM data and the frequencies are presented in Table 1. The number of times each language was recorded was calculated according to the ‘use’ of the language. For example, if the participants answered that they were using Dutch for reading and interaction, it was recorded that the language was used two times. The results of these calculations can be seen in Table 1, which also contains the number of participants who used the languages regularly according to the SMM data, as well as the number of participants who claim in the questionnaire to use the language regularly and to know the language.

Dutch is recorded most often. English is also strongly present in the everyday lives of the participants, accounting for 28% of language use. Frisian accounts for 4% of language use, and the rest of the languages even less. The category ‘Other’ includes 18 different languages. According to the number of speakers of each language, it can be seen that the use of English is evenly dispersed throughout the group of participants. Hungarian, on the other hand, is only used by one participant, though often. In addition, more than half of the participants who claim to know Frisian actually use the language in everyday life, whereas, in the case of German, only about 13% of the participants who know the language use it.

Data showing whether participants were using one or two or more languages simultaneously was also collected. The results were as follows: single language – 1914 times; multiple languages – 407; none – 50. During the 407 cases of the simultaneous use of multiple languages, different language combinations were singled out, some more frequent than others: Dutch/English (304 times), Dutch/Frisian (15), Dutch/Other (35), English/Other (19), three or more languages at a time (32), and other combinations (2). The Dutch/English combination dominates everyday language use, amounting to 75%, while the other combinations are more or less evenly distributed across the data, with Dutch/Other and three or more languages at a time taking 9% and 8% respectively, and Dutch/Frisian and English/Other – 4% and 5% respectively.

The contexts in which the simultaneous use of multiple languages occurred were also analysed. The data provided various context combinations (e.g. Interaction, Homework/Music, Reading, TV/Subtitles etc.). Fifty-one percent of these cases involved the use of media such as music, radio, or television. In contrast, just 7% of the simultaneous use of multiple languages was limited to only interaction (i.e. communication with people) in several languages. Most of it was carried out in Dutch/English (54%) or Dutch/Frisian (25%).

Table 1. Total number of instances of language records.

	Dutch	Frisian	English	German	French	Spanish	Hungarian	Finnish	Other
Records	1883	106	842	54	15	12	8	12	37
Speakers	100	15	92	8	6	7	1	2	5
Quest. Speakers	100	17	92	10	8	6	1	2	
Quest. Know	100	26	99	77	54	12	1	2

Figure 1. The use of languages in isolation and simultaneously with other languages.

The proportion of how much a language was used in isolation and simultaneously with other languages was also analysed. Dutch has the largest difference in proportion between single and multiple situation use. English has a smaller difference between the situations. Frisian, like Dutch, is mostly used in isolation from other languages. In the case of German practically an equal proportion can be seen (single – 25; multiple – 29), whereas the use of other languages is largely relegated to situations of simultaneous use. Figure 1 presents the results.

The contexts of use for individual languages were also analysed. It was found that the use of English is largely relegated to media: music, television, and reading. Seventy percent of the use of English is in these contexts, regardless of whether it is used in isolation or in situations of simultaneous use. The other uses of English are as follows: homework – 7%; interaction – 8%; other – 15%. The latter category includes writing, gaming, Internet usage, class, etc.

Frisian appears as a language largely used in interaction: 99 times out of 106 (93%). Moreover, 66% of this interaction occurs within the family, 28% with friends, 0% with a partner, and 4% at work. There were two data collection prompts where the participant did not indicate with whom the interaction was occurring.

German also appears as a language primarily used in interaction (56%). French and all other languages are mostly used when doing homework, with 40% and 46% allocated to this use respectively.

The questionnaire data and its comparison to the SMM data

The data collected through the questionnaires largely reflect the SMM data. All of the non-Dutch language use in interaction indicated in the questionnaires that happened for a major period of time each day (60 minutes per day and longer) was also found in the SMM data.

However, the questionnaire data showed a lower number of languages encountered in participants' everyday lives than indicated in the SMM data. Overall, 20 participants were found to use more languages on SMM than they indicated in the questionnaire. The use of these languages varied from media and gaming to interaction and homework. It is thought that most of these language uses were not accidental, meaning that participants encounter and use these languages systematically throughout their everyday lives.

The number of speakers of the most often recorded languages in the SMM data (Dutch, Frisian, English, German, French, Spanish, Hungarian, and Finnish) practically coincided with the number of users according to the questionnaires (see Table 1). When the data were broken into the number of users per each separate activity, a Chi-Square analysis showed a significant difference between the SMM data and the questionnaire data in the case of the use of English, $\chi^2(3) = 19.83$, $p < .001$. Significantly fewer participants were found to use English in various activities when compared to the questionnaire. The numbers can be seen in Table 2.

Table 2. The number of participants using English in various activities according to the questionnaire and the Facebook® data.

	Reading	TV	Music/Radio	Internet
Questionnaire	66	78	72	75
Facebook®	40	60	52	15

The correlations of other languages could not be performed due to the small size of the sample.

Discussion

Effectiveness of the SMM

Since the SMM was proposed as an alternative or an addition to the questionnaire, it was first determined whether the SMM provides the researcher with a sufficient amount of data. Naturally, participants would not be able to answer all the data collection prompts sent to them over Facebook® due to the simple fact that they are not online 24 hours a day. They may be in class, sleeping, or driving, among other things, which would prevent them from answering the data collection prompts.

The quantitative analysis showed that participants replied quite regularly and, even though there were some who did not reply to a majority of the data collection prompts, there was a sufficient amount of data left for analysis. Even following a schedule of only three to four data collection prompts per day, it was possible to gather a great amount of data. There was enough space between collection prompts for the participant to have time to check their Facebook® accounts and reply. There do not seem to be reasons for increasing the frequency of data collection prompts, since it would probably lead to more of them being missed by the participant. On many mornings, participants even replied to data collection prompts which were sent to them the evening before. This could be seen as a sign that not only were the participants not adverse to replying on days which were not singled out as data collection days, but also that, if a participant was motivated to take part in the experiment, they did it quite conscientiously.

Language use patterns encountered

Overall, participants were found to use 20 different languages. The number of times the languages were used was recorded along with the number of speakers so as to acquire a picture of the probable dispersion of the use of a specific language over this sample. The extensive use of Dutch by every participant is an expected result and hardly requires analysis, since it is the native language of most of the participants.

English appears as a language that is used in everyday life by an overwhelming majority of the participants. English has a great presence in the Netherlands, not only due to its spread in media and the Internet, but also in education. Many of the courses at universities are taught in English, while even those that are taught in Dutch use academic literature in English alongside that in Dutch. A large part of the participants were at the age when they are still at university, so they would encounter English not only in these contexts, but also in interaction at the university with international students. Overall, this data confirmed previous knowledge concerning the spread of English in different walks of life also in countries where English is not an official language (Graddol 2006).

The differences in numbers of participants who claim to *know* a language in the questionnaire, who claimed to *speak* it and who, according to the SMM data, actually did, presents an interesting picture of language vitality in the society in question. The fact that more than half of the participants who know Frisian actually speak it – and quite often, as its use amounts to 4% of all language use, according to the SMM data – shows that the language holds a firm place in certain domains of

language use. In contrast, German is spoken regularly only by 8–10 speakers out of 77 participants, which leads to the assumption that the use of German may be governed more by other factors, such as practice in order to maintain the level of proficiency, rather than by necessity. These results are not unexpected in keeping with the design of the present research.

Frisian is a language that is native to the north Netherlands and has an official status along with Dutch. The participants who spoke the language are the ones that are bilingual in both languages. German, on the other hand, is a foreign language in the Netherlands; therefore, it is natural that it is not used so often. However, it would be interesting to look at this proportion between the times of language use and the number of speakers in a society where the language situation is more ambiguous. It seems that the SMM provides an interesting picture of language vitality and the ‘intensity’ of language use, which may be used in research that, according to Verdoodt (2000), always should precede the drawing up of language policies.

The SMM also provided an opportunity to look at how much participants used one language and how much they used multiple languages simultaneously. These answers were anticipated before the conducting of the experiment, but what was surprising was the popularity of replies indicating that the participant was not using any language at all. Strictly speaking, they were using Dutch or English to reply to the data collection prompt, but it was decided at the beginning of the study to cut this factor out of the data analysis. So it appeared that there were at least 50 situations when the participants decided that they were not using any language. The contexts of use for these situations were not always mentioned and varied from ‘Sleeping’ to ‘Listening to classical music’, making it impossible to trace a pattern. It must be asked, though, whether there are any situations in an adult person’s life when they are indeed not using any language at all? There were enough records of the use of a language when ‘Thinking’ (29 records) to venture to say that participants usually indicated this instance in their replies and did not overlook it. These answers may indicate that the participants were unwilling to reply in detail about their language use at that time and typed ‘None’ to brush aside the researcher, or that the participants were not aware of their language use at the time.

The cases of the simultaneous use of multiple languages were looked at in detail. Among the language combinations singled out, Dutch/English was firmly in the lead. Since Dutch is the native language of 96 participants, it is also often used in combination with all foreign languages. English was recorded most often because it is the most widespread foreign language. The Dutch/Frisian combination is due to the presence of bilinguals in these languages among the participants. English is also used in combination with other languages. However, an overwhelming majority of these instances involve English music or English television and film. The prevalence of English media will be discussed in more detail further; however, it is safe to say that English is used in the same domains both in Dutch/English and English/Other combinations. It is highly unlikely that the languages involved in cases of simultaneous use would be used to an equal degree. The results from the SMM showed that 51% of these instances involved the use of media (e.g. music, television, and radio) in at least one language. Thus, this language was not being used actively, but as ‘background noise’. In contrast, only 7% of the ‘multiple languages’ records involved two languages used solely in interaction. A question arises: while these 7% could be classified as codeswitching without doubt, do the remaining records fall into this category?

Clearly, this sort of data – the combinations of different languages and the contact situations – is not available through questionnaires. The SMM provides unique data on cases of codemixing, mainly because participants report on their language behaviour shortly after it occurred. Moreover, the SMM allows for compilation of a typology of contact situations.

The SMM also provided the opportunity to compile a picture of the uses of each individual language. It is interesting to see that native languages (Dutch and Frisian) are mostly used in isolation, while foreign languages have either an almost equal proportion between the cases of isolated and simultaneous use or the proportion is skewed in favour of simultaneous use with other languages. English, looked at in isolation, appears to be used primarily through media, such as music and television. Although it is firmly present in the participants’ everyday lives, it is rarely

used productively (e.g. writing and interaction) and is mostly used receptively (e.g. media and reading). This large presence of English media confirms previous research on the subject (Berns, de Bot, and Hasebrink 2007).

Differently from English, Frisian is shown as a language of interaction, primarily in the family. There were bilingual participants who were raised in both languages, and it is evident that the use of Frisian is maintained in the family domain. It may overlap a little with the friends domain, but not significantly. Fishman (1972) wrote that, in the case of minority languages and language shift, multilingualism either starts in the family, or withdraws to it. The SMM data on the Frisian language confirm this, showing that the language is very much 'alive', yet mostly in the family domain.

Comparison of the SMM data with the questionnaire data

In short, the numbers of speakers of the most often used languages coincided in both sets of data. On the other hand, the questionnaire data showed a higher usage of smaller languages than the SMM data. This is likely to be an effect of the design of the experiment: the data collection prompts were sent over Facebook® only three/four times per day. The probability of these prompts coinciding with the use of a language which is used 60 or more minutes per day is much greater than with a language which is used 5 minutes per day. A researcher in the field of multilingualism would typically be more interested in the more often used language; for example, the extended and habitual use of a minority language in such cases may have some effects on language maintenance and shift. Of course, some languages may be used less because they are used with people with whom the participant does not maintain regular and constant contact. A young person who moved away from home may be using his home language with his parents only five minutes a day. This would be harder to record through the SMM.

The discrepancies concerning the languages which participants mentioned can most likely be explained by factors such as the lack of space in the questionnaire to include all the languages the participants encounter, or the lack of motivation to fill in all these languages. Finally, the participants may simply not have been aware that they use/encounter these languages at all.

It must be mentioned that due to the different types of data collected through the SMM and the questionnaire it was impossible to fully compare the data. In other words, the number of minutes/hours a language is spoken according to the questionnaire cannot be compared to the number of times the language is recorded in the SMM data. However, the number of participants claiming to use a language, as well as the domains in which they use a language, can be compared. Future studies should take this into account.

Conclusion

This study has examined the usability of social media for examining language use in multilingual communities. We compared reports of language use obtained through a language background questionnaire with those obtained by regularly contacting participants several times a day through Facebook® and asking them what language(s) they were using at that time, how and with whom. The data were compared manually and tested statistically when possible.

It was found that the data from the SMM largely reflect the data from the questionnaires where interaction with people is concerned. Concerning language use in various activities, it has been concluded that the SMM may provide more exact information regarding this than the questionnaire. When viewed independently, the SMM was found to provide a wealth of real-time data that could be analysed from many different points of view. Most importantly, it provided the opportunity to look at each foreign language recorded from different points of view, not only at its popularity and spread, but also the domains and activities in which it is used, the frequency with which it is used in isolation or in codemixing situations, and with which other languages it is used. The SMM provided further insight into codemixing situations

through the compilation of a typology of these situations. Information of this sort is indispensable in studies of multilingual environments.

One of the disadvantages of the SMM is that it is much more difficult to set up than a questionnaire. The researcher must keep in mind not only such factors as motivation and ‘annoyance’, but also issues of privacy. The questionnaire is less invasive and has to be filled out only once. However, social media will not stop being of interest as a tool for conducting research, especially large-scale surveys on a vast geographical scale. Despite its ‘trickiness’, if executed well, it could provide valuable information. The data would be of interest in studies focused on immigrant/minority communities, where the use of a majority language would be relegated to certain domains only, or hard-to-reach populations. Moreover, information concerning the domains and uses of a language (e.g. the spread of English media) could have an impact on education policies and teaching methods. It could also be adjusted to fit a specific research question, not only such general ones as in the present study.

We conclude that researchers should make use of social media in research on the descriptive make up of language contact situations. Facebook® has a clear advantage over other social media because of its popularity and the degree of people’s involvement in it. It would be interesting, however, to employ this method on other social media sites. Finally, for a continuation of this research the focus could be on the design of a specific social media application, under the auspices of and authorised by a university or another official institution, which would automatically collect language use data from voluntary participants several times a day. This would eliminate the human error possible on the part of a researcher and protect participants’ privacy.

Notes

1. Here and further the term ‘student’ refers to a student of the University of Groningen who gathered the data which was later analyzed by the researcher; the term ‘participant’ refers to the person the student contacted to take part in the experiment and who provided the data.
2. These participants indicated Frisian as their native language, and Dutch as the first language among the other languages they know. This may be due to the format of the questionnaire. However, due to their age (early twenties) and the current language situation in the Netherlands, it is highly unlikely that Frisian is indeed their native language in the strictest sense of the word, and Dutch a second language in the strictest sense. Most likely, they are bilingual in Frisian and Dutch. Therefore, it has been decided to code them this way.
3. This participant indicated Bildts, a variety spoken in Friesland, as their native language, and Frisian, Dutch, and English as other languages they know. However, a conclusion such as the one detailed in the footnote above cannot be drawn about this participant’s native languages. They may be trilingual, or bilingual, but that is not known with any degree of certainty.
4. Here and further ‘data collection prompt’ means a time when the three questions were sent as one message to a participant through Facebook®.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Anastasia Zhuravleva <http://orcid.org/0000-0002-2362-5842>

Kees de Bot <http://orcid.org/0000-0001-7442-2693>

Nanna Haug Hilton <http://orcid.org/0000-0003-0141-3539>

References

- Baker, Colin. 1997. “Survey Methods in Researching Language and Education.” In *Encyclopedia of Language and Education*, edited by N. H. Hornberger and David Corson, 35–46. Dordrecht: Springer.

- Baltar, Fabiola, and Ignasi Brunet. 2012. "Social Research 2.0: Virtual Snowball Sampling Method Using Facebook." *Internet Research* 22 (1): 57–74. doi:10.1108/10662241211199960.
- Berns, Margie, Kees de Bot, and Uwe Hasebrink, eds. 2007. *In the Presence of English: Media and European Youth*. New York, NY: Springer.
- Bonnet, Gérard, ed. 2002. *The Assessment of Pupils' Skills in English in Eight European Countries* 2002. European Network of Policy Makers for the Evaluation of Education Systems. Paris: Ministry of Education.
- boyd, danah. 2007. "Social Network Sites: Public, Private, or What?." *Knowledge Tree* 13, May. <http://www.danah.org/papers/KnowledgeTree.pdf>.
- Di Capua, Ivan. 2012. "A Literature Review on Research on Facebook* Use." *The Open Communication Journal* 6: 37–42. <http://benthamopen.com/tocommj/articles/V006/37TOCOMMJ.pdf>.
- Extra, Guus, and Kutlay Yağmur, eds. 2004. *Urban Multilingualism in Europe: Immigrant Minority Languages at Home and School*. Clevedon: Multilingual Matters.
- Fishman, J. A. 1972. "Language Maintenance and Language Shift as a Field of Inquiry: Revisited." In *Language in Sociocultural Change: Essays by Joshua A. Fishman*, edited by A. S. Dil, 76–115. Stanford, CA: Stanford University Press.
- van Gilder Cooke, Sonia. 2011. "Walls Have Eyes: How Researchers Are Studying You on Facebook*." *Time*, November 14. <http://www.time.com/time/business/article/0,8599,2099409,00.html>.
- Graddol, David. 2006. *English Next*. London: British Council.
- Gullberg, Marianne and Peter Indefrey. 2003. *Language Background Questionnaire. Developed in the Dynamics of Multilingual Processing*. Nijmegen: The Max Planck Institute for Psycholinguistics. <http://www.mpi.nl/departments/other-research/research-projects/the-dynamics-of-multilingual-processing/tools/Lang-Hist-Quest-Engl.pdf>.
- Hoffman, M. N. H. 1934. *The Measurement of Bilingual Background*. New York, NY: Teachers College, Columbia University. <http://catalog.hathitrust.org/Record/001642672>.
- Holmes, Janet. 1997. "Keeping Tabs on Language Shift in New Zealand: Some Methodological Considerations." *Journal of Multilingual and Multicultural Development* 18 (1): 17–39. doi:10.1080/01434639708666300.
- Lee, R. A., K. C. Browne and Yoko Kusumoto. 2011. "Measuring EFL Learners' Environment; English Contact and Use Outside the Classroom at a Japanese International University." *Polyglossia* 20: 15–25. <http://jairo.nii.ac.jp/0026/00002154/en>.
- Markham, Annette N. 2006. "Ethic as Method, Method as Ethic." *Journal of Information Ethics* 15 (2): 37–54. <http://markham.internetcinquiry.org/writing/markhamijie2006reprint.pdf>
- Markham, Annette N. 2011. "Internet Research." In *Qualitative Research: Theory, Method and Practices* (3rd ed.), edited by David Silverman, 111–127. London: Sage.
- Markham, Annette N., and Elizabeth Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (version 2.0). <http://aoir.org/reports/ethics2.pdf>
- Marpsat, Maryse, and Nicolas Razafindratsima. 2010. "Survey Methods for Hard-to-Reach Populations: Introduction to the Special Issue." *Methodological Innovations Online* 5 (2): 3–16. doi:10.4256/mio.2010.0014.
- Martin-Jones, Marilyn. 1991. "Sociolinguistic Surveys as a Source of Evidence in the Study of Bilingualism: A Critical Assessment of Survey Work Conducted among Linguistic Minorities in Three British Cities." *International Journal of the Sociology of Language* 90 (1): 37–56. doi:10.1515/ijsl.1991.90.37.
- Milroy, Lesley, Wei Li, and Suzanne Moffatt. 1991. "Discourse Patterns and Fieldwork Strategies in Urban Settings: Some Methodological Problems for Researchers in Bilingual Communities." *Journal of Multilingual and Multicultural Development* 12 (4): 287–300. doi:10.1080/01434632.1991.9994465.
- Park, Namsu, K. F. Kee, and Sebastián Valenzuela. 2009. "Being Immersed in Social Networking Environment: Facebook* Groups, Uses and Gratifications, and Social Outcomes." *CyberPsychology & Behavior* 12 (6): 729–733. doi:10.1089/cpb.2009.0003.
- Rees, M. E. G. 1954. *A Welsh Linguistic Background Scale: Pamphlet No.2*. Faculty of Education, University College of Wales.
- Seliger, H. W. 1977. "Does Practice Make Perfect? A Study of Interaction Patterns and L2 Competence." *Language Learning* 27 (2): 263–278. doi:10.1111/j.1467-1770.1977.tb00122.x.
- Seliger, H. W., and Rosario Gingras. 1976. Who Speaks How Much and to Whom?: A Study of Interaction Patterns in Second Language Classrooms. Unpublished paper presented at colloquium on verbal and non-verbal behavior in second language learning. National TESOL Convention, New York.
- Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The Anatomy of the Facebook* Social Graph. <http://arxiv.org/abs/1111.4503>.
- Verdoodt, A. F. 2000. "The Demography of Language." In *The Handbook of Sociolinguistics*, edited by Florian Coulmas, 33–45. Oxford: Blackwell Publishers.

Appendices

Appendix 1: SMM instructions and questions

Facebook® survey instructions (to be included in first message):

Please answer the questions immediately upon opening the message. Indicate which language you are using (it can be several languages, e.g. Dutch and English) and in which context (e.g. interaction with other people, listening/speaking in class, doing homework, watching a film, reading a book, listening to music, chatting with a friend on Facebook®/WhatsApp®/Skype®, writing an email, etc.). If you answer 'interaction', please indicate with whom (e.g. parent, sibling, teacher, friend, etc.). If you want to give a more detailed answer because it may be useful to the project, please feel free to do so. Thank you for your participation!

Questions:

What language(s) are you using?

How? (Interaction, class, film, book, music etc.)

If interaction, with whom?

Appendix 2: Coding template

The coding template with examples of coding can be seen below. Data collection prompt 5 is an example of coding an unanswered collection prompt. This is not the actual data used in the study (Table A1).

Table A1. Coding template with examples.

Group	Participant	Collection	Day	Time	Multiple/Single languages	Language	Context	Interaction with who
A	Participant	1	Wed	11:05	Single	Dutch	Reading	.
A	Participant	2	Wed	15:08	Single	Dutch	TV	.
						Dutch	Interaction	Partner
A	Participant	3	Thu	08:00	Multiple	English	TV	.
						Dutch	Interaction	Partner
A	Participant	4	Fri	12:35	None	.	.	.
A	Participant	5	.	.	.	.	.	.