Proceedings of the Ninth International AAAI Conference on Web and Social Media

Audience and the Use of Minority Languages on Twitter

Dong Nguyen and **Dolf Trieschnigg**
University of Twente
Enschede, The Netherlands
{d.nguyen,d.trieschnigg}@utwente.nl

Leonie Cornips
Meertens Institute
Amsterdam, The Netherlands
leonie.cornips@meertens.knaw.nl

Abstract

On Twitter, many users tweet in more than one language. In this study, we examine the use of two Dutch minority languages. Users can engage with different audiences and by analyzing different types of tweets, we find that characteristics of the audience influence whether a minority language is used. Furthermore, while most tweets are written in Dutch, in conversations users often switch to the minority language.

Over 10% of the Twitter users tweet in more than one language (Hale 2014). Also within a single language, there is much geographical variation (Eisenstein et al. 2010). Every user has his or her own linguistic repertoire which the individual user can draw linguistic elements or codes (language varieties) from. We illustrate this with two tweets from an international fashion model from Friesland (a Dutch province) who can draw on English, Dutch and Frisian:

> We just touched down in London town #vsfashionshow

> @USER SKATSJE!!! Lekker genietsje fan heit en mem en Fryslan!! ik mis jim
> *Translation: @USER CUTIE!!! Enjoy with mom and dad in Friesland!! i miss you*

She mostly tweets in English, possibly to maximize her audience (Androutsopoulos 2014) and to create an international image. For example, the first tweet is written in English. Using *#vsfashionshow* the tweet becomes part of a public stream about a fashion show, increasing her audience even more. The second tweet is a response to a tweet from her sister and is written in Frisian, a minority language spoken in Friesland. Through tweeting in Frisian, she is constructing their shared localness or 'Frisianess'.

Which language and linguistic elements users select from their linguistic repertoire depends on various factors, including the audience, the topic or perspective, to mark something as humorous/serious, etc. (Androutsopoulos 2013). We focus on the influence of *audiences* on whether a minority language is used on Twitter. A speaker's style is influenced by the audience (Bell 1984), and in that sense, social media, and especially Twitter is interesting: multiple audiences (e.g. friends, colleagues) are collapsed into a single context (Marwick and boyd 2011). Users with public profiles on Twitter have potentially a limitless audience, but they often imagine an audience when writing tweets and may target tweets to different audiences (Marwick and boyd 2011).

We study Twitter users in two provinces in the Netherlands, where besides Dutch a minority language is spoken. Frisian (spoken in Friesland) is recognized as an official language in the Netherlands. Also Limburgish (spoken in Limburg) – a group of what people call dialects – have received minor recognition by the Netherlands, a signatory of the 1992 European Charter for Regional Languages or Languages of Minorities. There is a positive attitude towards both minority languages, but their use has declined (Riemersma, Gorter, and Ytsma 2001; Cornips 2013).

In this paper, we analyze the *language choices* of users on a tweet level, focusing on when users tweet in a minority language. An automatic language identification tool is used to classify tweets according to their language. We distinguish between two types of tweets: tweets that are a response to another tweet, and 'independent' tweets. We first focus on independent tweets, analyzing tweets with direct addressees (where the targeted audience may be reduced) and tweets with hashtags (where the audience may be expanded).

We then study language choices for tweets that are responses to other tweets by extracting conversations on Twitter. Speakers may often *code-switch* i.e. use multiple languages in a single speech exchange, for example within a speaker's 'turn', within a syntactic unit, or even hybrid ways of speaking in which the linguistic elements used cannot be attributed any longer to a specific language. In this study, we focus on code-switching on the tweet level. Following Androutsopoulos (2013), we take a restrictive view on what is considered a speech exchange and confine our attention to code-switching within Twitter conversations.

Our contributions can be summarized as follows:

- We show that Twitter users accommodate to their audiences by studying the influence of direct addressees and hashtag streams on language choice.
- We study code-switching patterns within Twitter conversations and find that characteristics of the conversation partner as well as previous language choices in the conversation influence language choice.

Related Work

Our study builds on two different lines of work. First, we draw from the frameworks of audience design (Bell 1984) and communication accommodation theory (Giles, Coupland, and Coupland 1991), and in particular recent studies that have applied these frameworks to social media settings. On Facebook, users maximize or partition their audience (when starting posts) or align or disalign (when responding) using their language choices (Androutsopoulos 2014). A small-scale study on Twitter revealed that bilingual Welsh/English users more often tweet in Welsh to a user who is also bilingual, and in English when posting a tweet that is not directed to particular users (Johnson 2013).

Second, we follow recent large-scale quantitative studies of language choice and code-switching based on automatic language identification (Kim et al. 2014; Jurgens, Dimitrov, and Ruths 2014; Eleta and Golbeck 2014; Hale 2014). Traditional sociolinguistic studies rely on qualitative analyses (cf. Androutsopoulos (2014)) or quantitative analyses using questionnaires or manual coding (see Androutsopoulos (2013)). While revealing valuable insights, these studies have been limited to small sets of speakers. Larger datasets that are automatically tagged by language can complement such studies. So far, large-scale studies have mostly focused on the networks of multilingual users, finding that multilingual users connect users who only tweet in one language (Kim et al. 2014; Eleta and Golbeck 2014; Hale 2014). In these studies, users were represented by a language label or the language distribution of their tweets, thus focusing on language choice on a *user level*. In comparison, this study focuses on language choice on a *tweet level*.

Dataset

Twitter users from the Dutch provinces Friesland and Limburg were collected by starting with seed users and expanding using followers and followees. The seed users were manually identified users and users with a geotagged tweet from within these provinces (streaming API: January 2013 - July 2014). Users were mapped to locations (city, province, country) based on their provided profile location. For each user we collected the most recent 200 tweets.

An automatic language identifier was used to label the tweets. A training set of over 38k tweets was manually compiled with tweets labeled as English, Dutch, Limburgish or Frisian. Tweets containing multiple languages were labeled according to the predominant language. A logistic regression classifier obtained a cross-validation accuracy of 98%. Because performance was lower on very short tweets, tweets with less than 4 tokens were not labeled by the classifier. Manual rules were constructed to label a subset of the very short tweets. Similar to other studies on language choice (e.g. Kim et al. (2014)), we applied a threshold to determine whether a user uses a minority language on Twitter. We only retained users with at least 7.5%[1] of their tweets marked as containing Frisian or Limburgish, resulting in 2,069 users from Friesland and 2,761 users from Limburg.

[1]Threshold was based on data analysis, retaining approximately 23% of the users

We extract conversations based on information from the Twitter API, which provides the identifier of the original tweet in case of a reply. We excluded conversations with tweets from only one user (users can reply to themselves) and conversations for which the first tweet was a response to a missing tweet. We extracted 3,916 conversations, containing a total of 10,434 tweets. Most conversations were of length 2 (mean: 2.664, max 23).

Language Choice

In this section, we focus on tweets that are not a response to another tweet and are not a retweet. We analyze tweets with an explicit mention of another user. In such cases the targeted audience is often shifted towards the addressed user. We also study tweets with a hashtag, which causes a possible expansion of the audience as they are included in public hashtag streams. These differences in audiences are reflected in statistics normalized by user: When users mention a specific user, they are more likely to employ a minority language than when they use hashtags (e.g. users from Limburg use Limburgish in 33.8% of their tweets with a user mention vs. only in 28.6% of their tweets with a hashtag).

Addressee We first study the influence of addressees on language choice. We restrict our analysis to tweets that start with a user mention (@user). Such tweets are often directed towards the addressed user, in comparison to just tagging a user. For each user, we sampled up to two tweets.

We aim to analyze if addressees influence whether a minority language is used, while controlling for a user's tendency to use a minority language. We use logistic regression, which allows analyzing which factors explain the language choice. We fit a model with the dependent variable being the language choice, modeled as a binary variable (minority language or not). Independent variables are the use of minority language by the addressee, measured as the proportion of the last 100 tweets (before the tweet of interest) containing a minority language, and a binary variable indicating whether the addressee is from the same province. We collected additional data for addressees who were not in our dataset. However, for some we were not able to obtain data and these were excluded from the analysis. The results (Table 1) indicate that Twitter users are more likely to use a minority language, when addressing a user who often uses the minority language. From manual inspection we do observe that users not always accommodate to their addressee. For instance, sometimes even celebrities or international companies are addressed in a minority language.

	Coefficient	**Std. Error**
Intercept	-2.010***	0.149
Use of minority lang. by user u	2.685***	0.299
Use of minority lang. by user a	3.221***	0.293
Same province	0.160	0.149

Table 1: Logistic regression model of influence of addressee a on language choice of user u; n = 1272; $^{***}p < 0.001$

Hashtags Hashtags have become a common practice on Twitter and they are often included to join public discussions (Huang, Thornton, and Efthimiadis 2010). We study the influence of the audiences of these public discussions on the language choice for tweets with hashtags.

For example, one of the most popular Dutch hashtag streams on Twitter is *#dtv* or *#durftevragen* ('dare to ask'). In these streams, Twitter users post questions on various topics, ranging from questions about software, opinions about news, to looking for a certain service. These streams have local variants (albeit less popular), such as for Limburgish *#durftevraoge* and *#durftevroage* and for Frisian *#doartefreechjen* and *#doartefreegjen*. Reaching the right audience is key here, since users are looking for an answer to their question. In our dataset, tweets using the local Limburgish and Frisian hashtag variants are all written in a minority language, whereas 84.6% of the tweets using the Dutch hashtag variants are written in Dutch.

Not all hashtags are added to join public discussions, for example some indicate a feeling (e.g. *#sad*) (Jurgens, Dimitrov, and Ruths 2014). We therefore confine our analysis to hashtags referring to named entities. These are interesting, because they tend to be used to link to a public discussion and most of them do not imply a language choice on their own. We excluded hashtags that had local variants, such as names referring to cities. We manually annotated a random subset of the tweets (at most 1 tweet per user). In addition, we annotated whether the hashtag was referring to a local (e.g. local music festival) or (inter)national entity (e.g. show on national television). Of the annotated hashtags, 44.2% were marked as named entities and 51.7% of these referred to a local entity.

Similar to our previous analysis, we fit a logistic regression model with the language choice being the dependent variable. On Twitter the exact audience is unknown, but users use cues from the environment to imagine their audience (Marwick and boyd 2011). In a similar way, we use the last 100 tweets with the same hashtag written *before* the tweet of interest as cues. The tweets were collected using the search on the Twitter website, which allows searching in historical tweets. For each hashtag instance, we calculated the proportion of tweets in the stream containing a minority language. The audience may also consist of lurkers, but their language choices cannot be analyzed.

Table 2 shows the results. When many tweets in the hashtag stream contain a minority language, it is more likely that a user will use a minority language as well (even after controlling for the use of minority language by the user and whether the hashtag refers to a local or national entity).

	Coefficient	Std. Error
Intercept	-3.718***	0.453
Use of minority lang. by user	4.984***	0.819
Use of minority lang. in stream	6.489***	1.352
Hashtag about local entity	0.513	0.435

Table 2: Logistic regression model of influence of hashtag on language choice; n = 236; $^{***}p < 0.001$

Code-Switching in Twitter Conversations

In this section, we study the language choices of Twitter users when participating in conversations. Multilingual users often switch language during a conversation (i.e. *code-switching*). Initial tweets are frequently targeted towards a broader audience, but during a conversation the audience often shifts towards the direct conversation partner(s). Speakers accommodate to each other during conversations (Giles, Coupland, and Coupland 1991), which has also been observed on Twitter (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011). While the previous section focused on independent tweets, this section focuses on conversations, and thus an additional factor that influences language choice are the previous language choices made within the conversation.

Influence of previous tweet We calculate the probability of a language choice for a tweet ($lang_i$) given the language of the tweet the user is responding to ($lang_{i-1}$), i.e. $P(lang_i|lang_{i-1})$ shown in Fig. 1. Most of the time users align their language choice with the language of the tweet they are responding to (i.e. the self loop probabilities for Dutch and the minority languages are all above 0.5), and this trend is particularly strong when responding to tweets written in a minority language. However, this is not the case for English, which may be explained by the fact that English is most often used emphatically, for example by only inserting '*nice*' or '*thanks*', and thus it is less expected that the conversations continue in English. We also find that users from the Limburg province more often tweet in their minority language than users from Friesland.

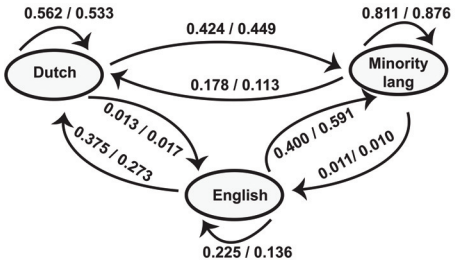

Figure 1: Switching behavior. The probabilities are reported for both provinces using [Friesland]/[Limburg]

In our next analysis, we also take into account the previous use of minority language by the users. Only tweets at the second position in a conversation were included in the analysis, to eliminate effects of other language choices. For each user, we sampled at most two tweets. Similar to the previous analyses, we fit a logistic regression model (Table 3) with as the dependent variable the language choice ($lang_i$). As independent variables, we include the use of minority language by both users as well as the language of the previous tweet. Location information was not included, since both users are from the same province. The results indicate that while the use of the minority language by the conversation partner is significant, the language of the previous tweet has a larger influence on the language choice.

	Coefficient	Std. Error
Intercept	-1.005***	0.112
Use of min. lang. by user of tweet$_i$	2.053***	0.241
Use of min. lang. by user of tweet$_{i-1}$	0.773**	0.248
Tweet$_{i-1}$ in minority language	1.478 ***	0.132

Table 3: Logistic regression model for language choice in conversations; n = 1863; $^{***}p < 0.001$, $^{**}p < 0.01$

Language choice over time Figure 2 shows the language distribution by position within a conversation. The analysis is based on all conversations, but we note that the same trends are observed when only including longer conversations. Most of the initial tweets are written in Dutch, possibly to maximize the audience (Androutsopoulos 2014). However, as conversations progress, it becomes more unlikely for a tweet to be written in Dutch. Once a switch has been made to a minority language, users tend to continue in that minority language (see also Figure 1).

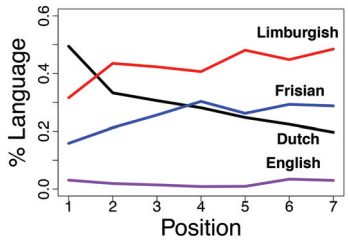

Figure 2: Language distribution by position in conversation

By replying in a different language, a user may be trying to negotiate the language choice (Androutsopoulos 2014). We expect that once a base language has been established, the probability of switching language decreases. For each $tweet_i$, we find the longest consecutive sequence ending at the previous tweet ($tweet_{i-1}$) written in $lang_{i-1}$ as an indication of the extent of negotiation going on. As expected, there is a significant, negative correlation between the lengths of these sequences and whether a switch occurs (Pearson's r=-.150, p $<$ 0.0001). The position in a conversation and whether a switch occurs correlate only slightly (Pearson's r=-.058, p $<$ 0.001) and controlling for this did not lead to notable changes in the trend.

Conclusion

In this paper, we studied the use of minority languages on Twitter across various settings. Our findings indicate that users tend to adapt their language choice to their audiences. When users address other Twitter users, the minority language is more likely to be used when the addressed users often make use of the minority language as well. In Twitter conversations, the language choices of users are also influenced by the language of the tweet they are responding to. Furthermore, while many tweets are written in Dutch to reach a broader audience, users often switch to the minority language during a conversation.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO), grants 314-98-008 (Twidentity) and 640.005.002 (FACT). The authors would like to thank Theo Meder, Lysbeth Jongbloed, Jolie van Loo, Anna Jørgensen and Jannis Androutsopoulos.

References

Androutsopoulos, J. 2013. Code-switching in computer-mediated communication. In *Pragmatics of Computer-Mediated Communication*. De Gruyter Mouton.

Androutsopoulos, J. 2014. Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media* 4–5(0):62 – 73.

Bell, A. 1984. Language style as audience design. *Language in Society* 13(02):145–204.

Cornips, L. 2013. Recent developments in the Limburg dialect region. In Hinskens, F., and Taeldeman, J., eds., *Language and Place. An International Handbook of Linguistic Variation*. De Gruyter Mouton.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of WWW*.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*.

Eleta, I., and Golbeck, J. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior* 41(0):424 – 432.

Giles, H.; Coupland, N.; and Coupland, J. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation*. Cambridge University Press.

Hale, S. A. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of CHI*.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational tagging in Twitter. In *Proceedings of Hypertext*.

Johnson, I. 2013. Audience design and communication accommodation theory: Use of Twitter by Welsh-English biliterates. In *Social Media and Minority Languages: Convergence and the Creative Industries*. Multilingual Matters.

Jurgens, D.; Dimitrov, S.; and Ruths, D. 2014. Twitter users# codeswitch hashtags!# moltoimportante# wow. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of Hypertext*.

Marwick, A. E., and boyd, d. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1):114–133.

Riemersma, A.; Gorter, D.; and Ytsma, J. 2001. *Frisian in the Netherlands*. Multilingual Matters. 103–118.