Audience and the Use of Minority Languages on Twitter

Dong Nguyen, D. Trieschnigg, and L. Cornips
Minority languages in social media

Doutzen Kroes
@Doutzen

We just touched down in London town😊
#vsfashionshow
instagram.com/p/wCkJsqzVle/

RETWEETS:
142

FAVORITES:
313
Minority languages in social media

Doutzen Kroes
@Doutzen

We just touched down in London town😊
#vsfashionshow
instagram.com/p/wCkJsqzVle/

142 RETWEETS 313 FAVORITES

Doutzen Kroes
@Doutzen

@boltsje SKATSJE!!! Lekker genietsje fan heit en mem en Fryslan!! ik mis jim

View translation
2:04 AM - 30 May 2009
Minority languages in social media

the influence of audiences on the use of minority languages on Twitter
Related work

• Audience design and Communication Accommodation Theory applied to social media (Androutsopoulos 2014; Johnson 2013)

• Large-scale studies on language choice and codeswitching using automatic language identification (Kim et al. 2014; Jurgens, Dimitrov, and Ruths 2014; Eleta and Golbeck 2014; Hale 2014)
Dataset
The Dutch Twitter landscape

Oct 2013 [1]:
- 5 million accounts
- 1 million active users

They mostly tweet in Dutch, English and ... 

[1] PeerReach, 2013
Dialects/minority languages/regional languages

ruurdtsej @ruurdtsej · 29 jun.
Ik twitterje yn it Frysk. Myn Nederlânske freonen fermoede lykwols dat ik geheimtaal skriuw....

Leon Jeurninck @lwjeurninck · 26 sep.
Mörge sezoensafsloeting vanne sjötterie mêt kampioensjeete, bbq en get beer oet greun fleskes.
Data Collection: user selection I

- Twitter users from the Dutch provinces Limburg and Friesland
- Seed users: Manually selected and based on geotagged tweets
- Expanded using social network (followers/followees)
Automatic Location Identification

Leeuwarden 1307 69.1%
leeuwarden 145 7.7%
Leeuwarden, The Netherlands 49 2.6%
Ljouwert 33 1.7%
Leeuwarden, Netherlands 25 1.3%
Leeuwarden, Friesland 14 0.7%
Leeuwarden, the Netherlands 13 0.7%
Leeuwarden, Nederland 13 0.7%
Leeuwarden, NL 8 0.4%
Leeuwarden, Holland 8 0.4%

Leeuwarden - Fryslân - Holland 1 0.1%
Stenden Leeuwarden 1 0.1%
°Leeuwarden° 1 0.1%
Leeuwarden, Techum 1 0.1%
Prinsentuingracht, Leeuwarden 1 0.1%
de blokhuispoort leeuwarden 1 0.1%
Binnenstad Leeuwarden 1 0.1%
#leeuwarden 1 0.1%
leeuwarden # freeceland 1 0.1%
Crystalic, Leeuwarden 1 0.1%
Leeuwarden - Bussum - Holland 1 0.1%
Kollum..Leeuwarden..Hoogezand 1 0.1%
Ureterp en Leeuwarden 1 0.1%
Stiens e.o. en Leeuwarden 1 0.1%
Emmakade, Leeuwarden 1 0.1%
... ... ...
Total 1891
Automatic Language Identification

- Languages labeled on a tweet level: English, Dutch, Limburgish or Frisian
- Features based on character n-grams
- Short tweets (less than 4 tokens) were skipped. Some were labeled using manual rules.
- Automatic classifier: accuracy of 98%
Data Collection: user selection II

• Only users with at least 7.5% of their tweets marked as Frisian or Limburgisch

• Total number of users:
  – 2,069 from Friesland
  – 2,761 from Limburg

• Conversations:
  – 3,916 conversations, containing a total of 10,434 tweets
Language choice
Language choice

• Independent tweets (no replies/retweets)

• Addressee: the targeted audience is often shifted towards the addressed user (audience is reduced)

• Hashtags: Tweets are included in public hashtag streams. Causes an expansion of the audience.
## Language choice: Addressee

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.010***</td>
<td>0.149</td>
</tr>
<tr>
<td>Use of minority lang. by user</td>
<td>2.685***</td>
<td>0.299</td>
</tr>
<tr>
<td>Use of minority lang. by addressee</td>
<td>3.221***</td>
<td>0.293</td>
</tr>
<tr>
<td>Same province</td>
<td>0.160</td>
<td>0.149</td>
</tr>
</tbody>
</table>

Logistic regression model (**p < 0.001).  
Dependent variable = Tweet in minority language?
Language choice: Hashtags

Example:
• #dtv or #durftevragen (‘dare to ask’): 84.6% tweets are in Dutch
• Local variants: Limburgish #durftevraoge and #durftevroage; Frisian #doartefreechjen and #doartefreeegjen: all tweets in the minority language

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.718***</td>
<td>0.453</td>
</tr>
<tr>
<td>Use of minority lang. by user</td>
<td>4.984***</td>
<td>0.819</td>
</tr>
<tr>
<td>Use of minority lang. in stream</td>
<td>6.489***</td>
<td>1.352</td>
</tr>
<tr>
<td>Hashtag about local entity</td>
<td>0.513</td>
<td>0.435</td>
</tr>
</tbody>
</table>

Logistic regression model (**p < 0.001).
Dependent variable = Tweet in minority language?
Code-switching
Influence of previous tweet I

- Dutch: 0.562 / 0.533
- English: 0.424 / 0.449
- Minority lang: 0.811 / 0.876

Dutch to English: 0.375 / 0.273
English to Minority lang: 0.400 / 0.591
Minority lang to Dutch: 0.011 / 0.010
Dutch to Minority lang: 0.013 / 0.017
Minority lang to English: 0.225 / 0.136

Influence of previous tweet I

0.562 / 0.533

Dutch

0.424 / 0.449

English

0.375 / 0.273

Minority lang

0.400 / 0.591

0.225 / 0.136

0.013 / 0.017

0.011 / 0.010

0.811 / 0.876
Influence of previous tweet I

Dutch → Minority lang

0.562 / 0.533

English

0.400 / 0.591

Minority lang → Dutch

0.811 / 0.876

0.424 / 0.449

0.178 / 0.113

English → Minority lang

0.375 / 0.273

0.013 / 0.017

Dutch

0.225 / 0.136

0.011 / 0.010
Influence of previous tweet II

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.005***</td>
<td>0.112</td>
</tr>
<tr>
<td>Use of min. lang. by user of tweet i</td>
<td>2.053***</td>
<td>0.241</td>
</tr>
<tr>
<td>Use of min. lang. by user of tweet i - 1</td>
<td>0.773**</td>
<td>0.248</td>
</tr>
<tr>
<td>Tweet i–1 in minority language</td>
<td>1.478***</td>
<td>0.132</td>
</tr>
</tbody>
</table>

Logistic regression model  (**p < 0.001, **p < 0.01)
Dependent variable = Tweet in minority language?
Language choice over time

% Language

Position

Limburgish
Frisian
Dutch
English
Discussion & Conclusion
Automatic Language Identification

• Difficult cases:
  – Treintje naar A'foort, dagke stage tot 4
  – Nice!
Automatic Language Identification

- Difficult cases:
  - Treintje naar A'foort, dagke stage tot 4
  - Nice!

... languages are not bounded, countable entities
Automatic Language Identification

• Difficult cases:
  – *Treintje naar A'foort, dagke stage tot 4*
  – *Nice!*

... languages are not bounded, countable entities

• But... these problems occur in any quantitative study! Quantitative studies require a simplification of the phenomenon.

• Next step: Automatic language identification at the word level (Nguyen & Dogruoz, EMNLP 2013), or maybe even morpheme level?
On computational methods & social media data

• Social media offers massive amounts of interesting data
• We need computational methods to fully leverage this data!
• Computational studies can complement existing sociolinguistic studies
Conclusion

• Users adapt their language choice towards their audiences
• Most tweets are written in Dutch, but users often switch to the minority language during a conversation

• See also: D. Nguyen, D. Trieschnigg and L. Cornips: Audience and the Use of Minority Languages on Twitter at ICWSM 2015
Thanks!

Questions/comments?

✉️ d.nguyen@utwente.nl
🐦 @dongng