Audience and the Use of Minority Languages on Twitter

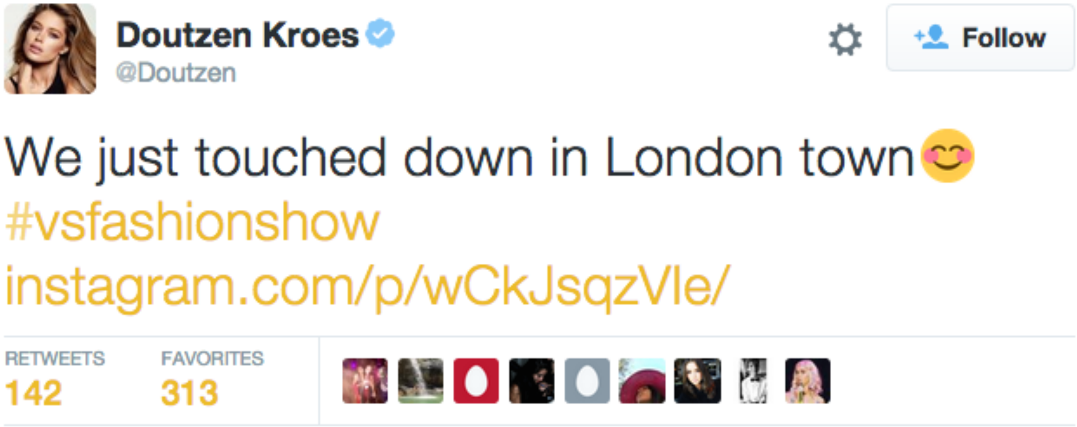
Dong Nguyen,

D. Trieschnigg, and L. Cornips

Meertens instituut

UNIVERSITY OF TWENTE.

Minority languages in social media



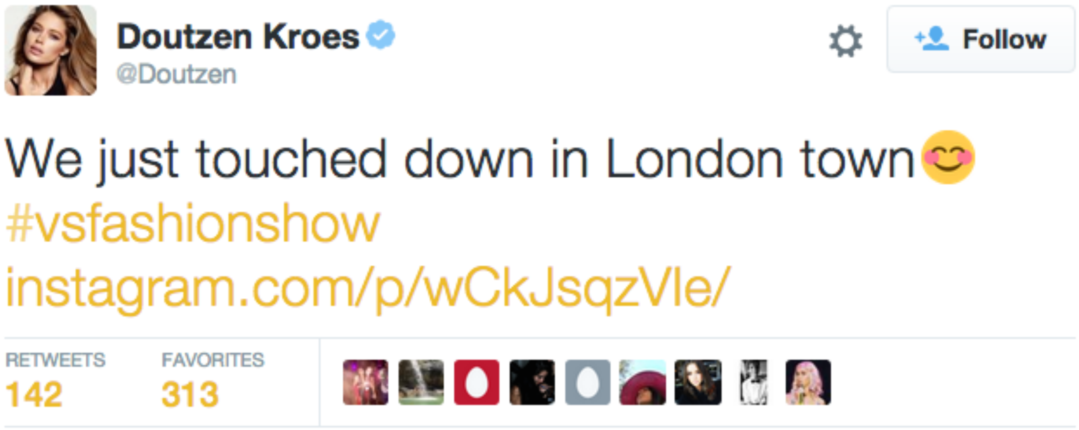
Doutzen Kroes
@Doutzen

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

RETWEETS 142

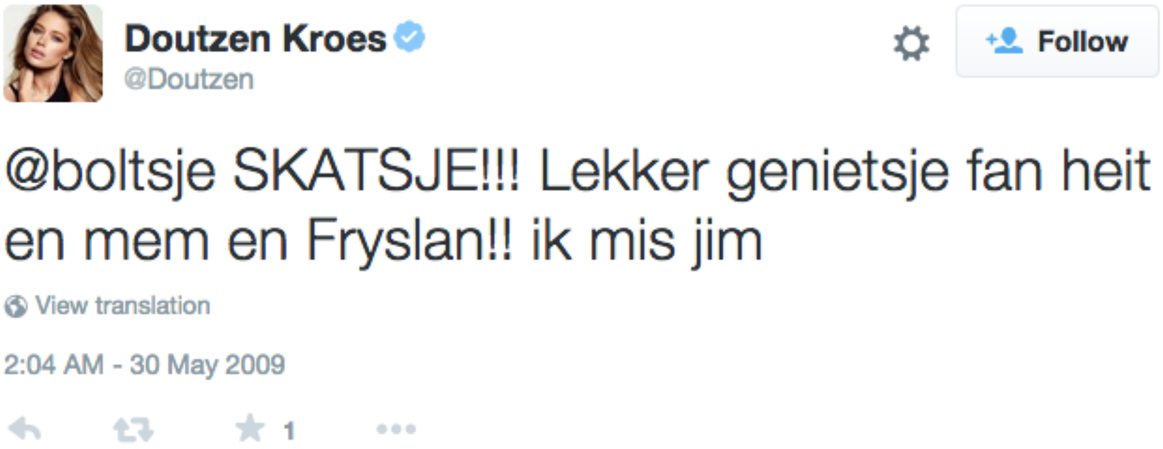
FAVORITES 313

Minority languages in social media



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

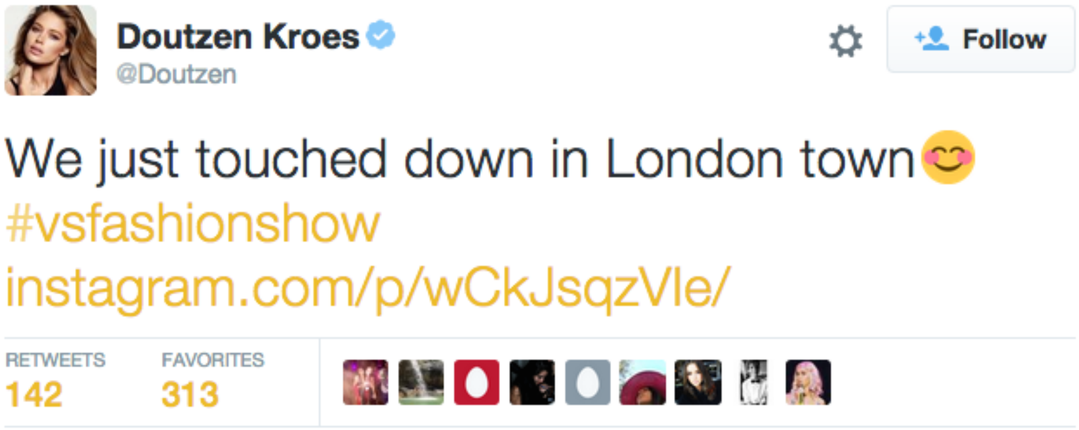
RETWEETS 142 FAVORITES 313



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

2:04 AM - 30 May 2009

Minority languages in social media

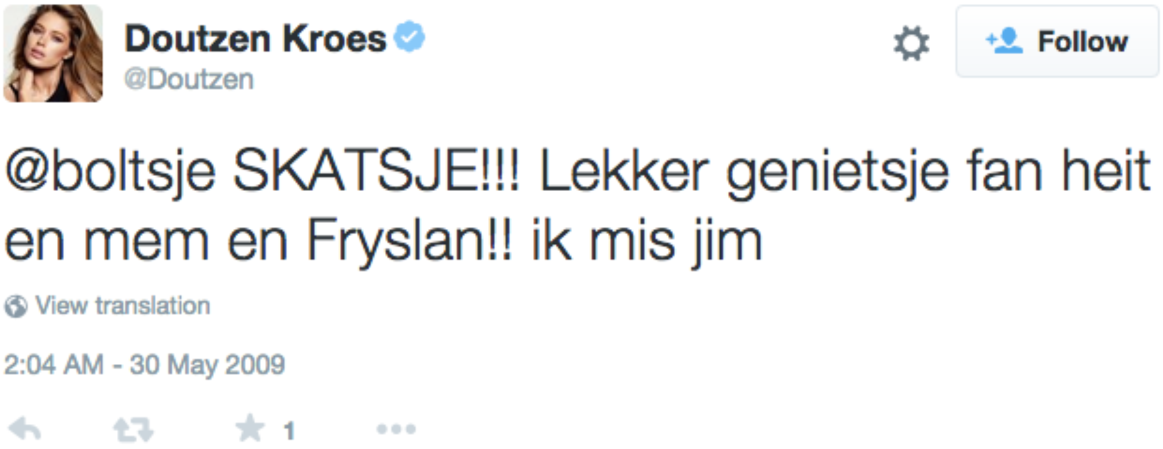


We just touched down in London town😊 #vsfashionshow instagram.com/p/wCkJsqzVle/

RETWEETS 142

FAVORITES 313

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



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

View translation

2:04 AM - 30 May 2009

1

Related work

- Audience design and Communciation 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

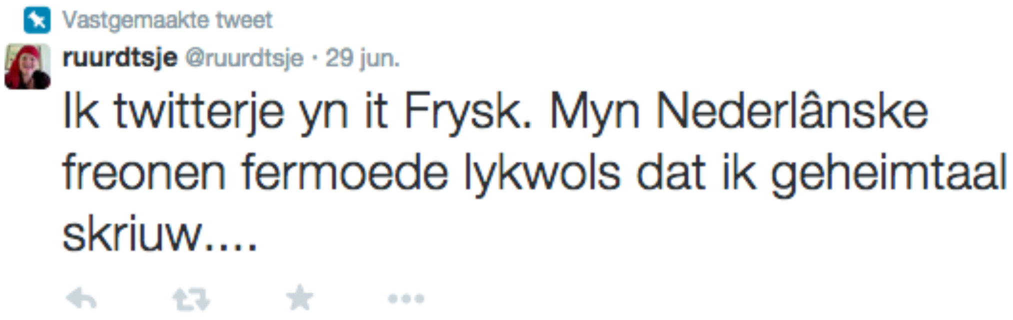
Oct 2013 [1]:

- 5 million accounts
- 1 million active users

They mostly tweet in Dutch, English and …

[1] PeerReach, 2013

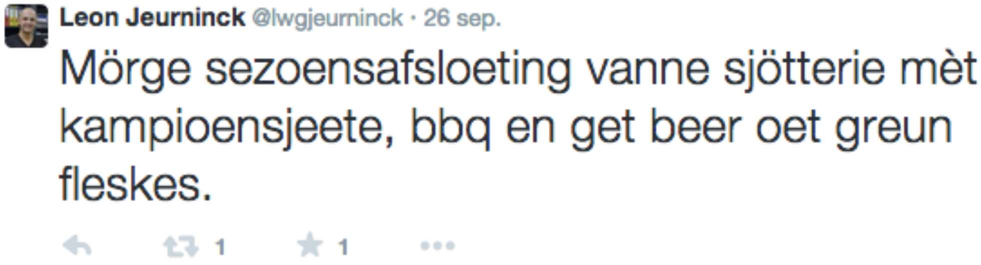
Dialects/minority languages/ regional languages



Vastgemaakte tweet

ruurdtsje @ruurdtsje · 29 jun.

Ik twitterje yn it Frysk. Myn Nederlânske freonen fermoede lykwols dat ik geheimtaal skriuw....



Leon Jeurninck @lwgjeurninck · 26 sep.

Mörge sezoensafsloeting vanne sjötterie mèt 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
N31
Goutum
N31
N32

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

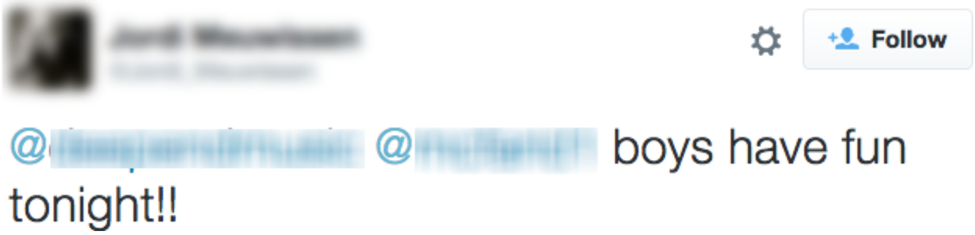
- Total number of users:
  - 2,069 from Friesland
  - 2,761 from Limburg

- Conversations:
  - 3,916 conversations, containing a total of 10,434 tweets

Language choice

Language choice

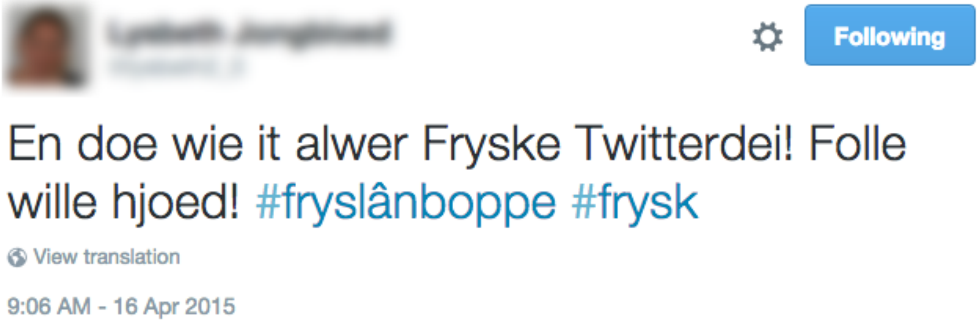
- Independent tweets (no replies/retweets)
- Addressee: the targeted audience is often shifted towards the addressed user (audience is reduced)



[redacted]

@[redacted] @[redacted] boys have fun tonight!!

- Hashtags: Tweets are included in public hashtag streams. Causes an expansion of the audience.



[redacted]

En doe wie it alwer Fryske Twitterdei! Folle wille hjoed! #fryslânboppe #frysk

9:06 AM - 16 Apr 2015

Language choice: Addressee

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

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 #doartefreegjen: all tweets in the minority language

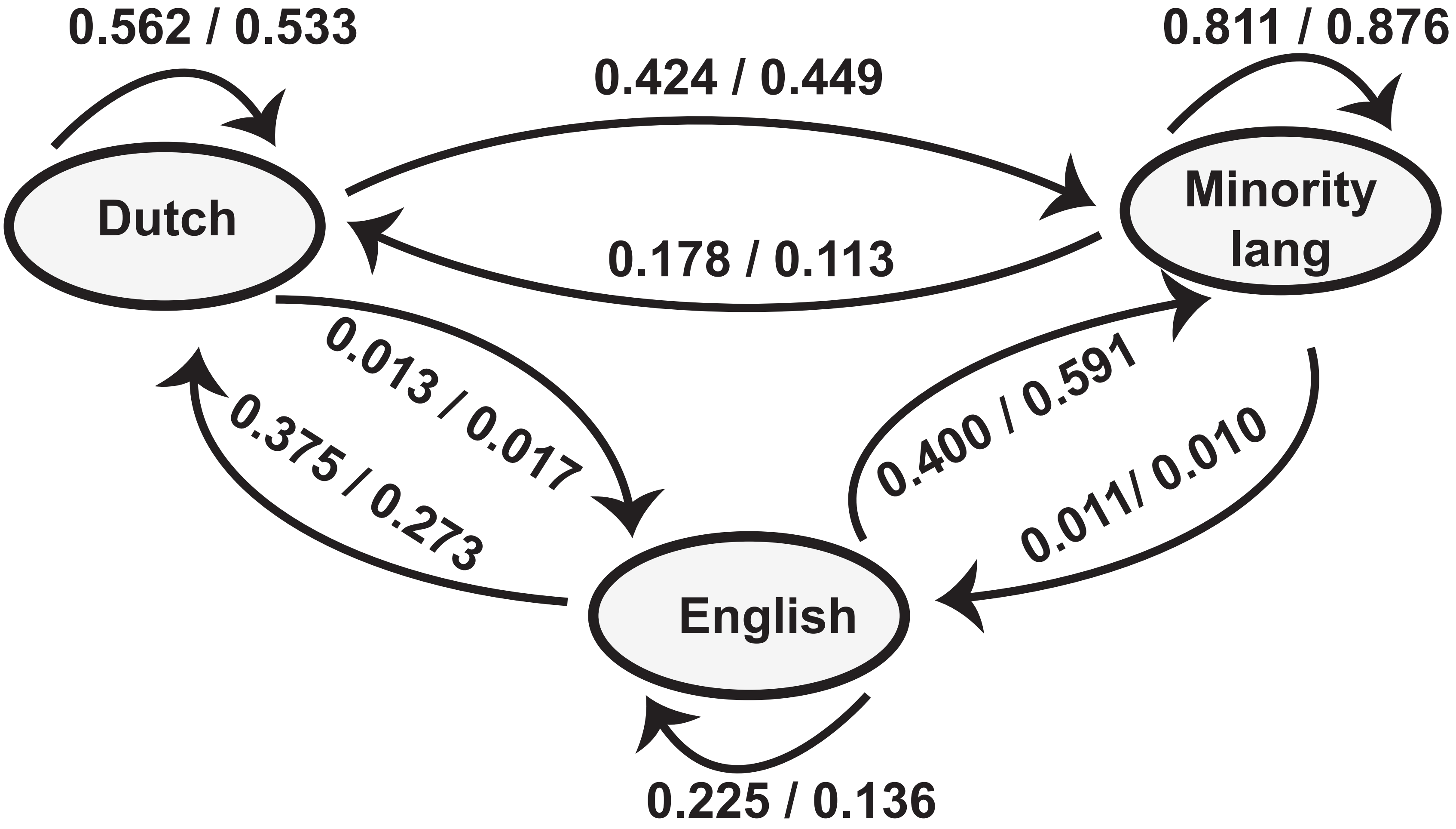
	Coefficient	Std. Error
Intercept	-3.718^{***}	0.453
Use of minority lang. by user	4.984^{***}	0.819
Use of minority lang. in stream	6.489^{***}	1.352
Hashtag about local entity	0.513	0.435

Logistic regression model (*** $p < 0.001$).

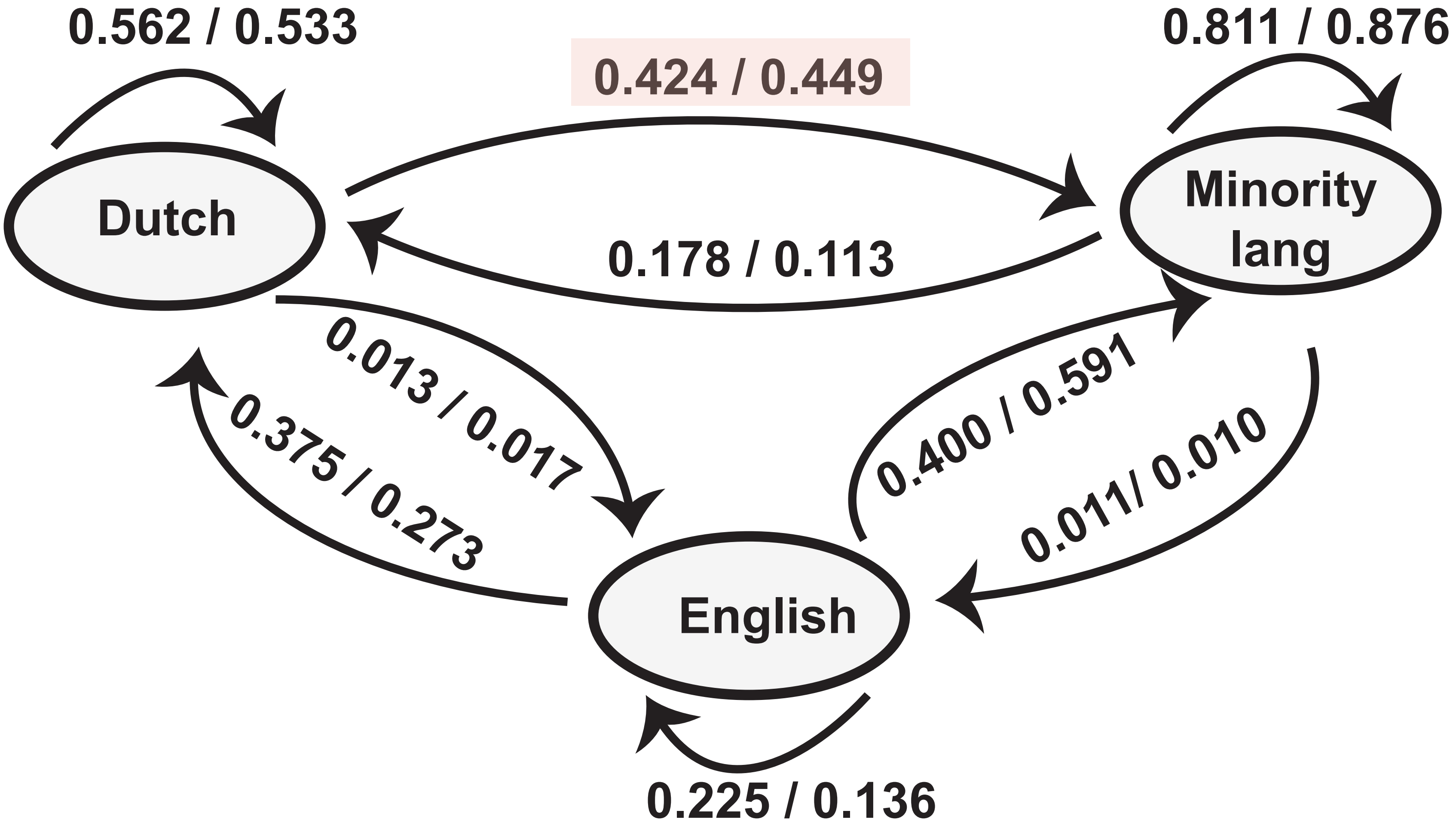
Dependent variable = Tweet in minority language?

Code-switching

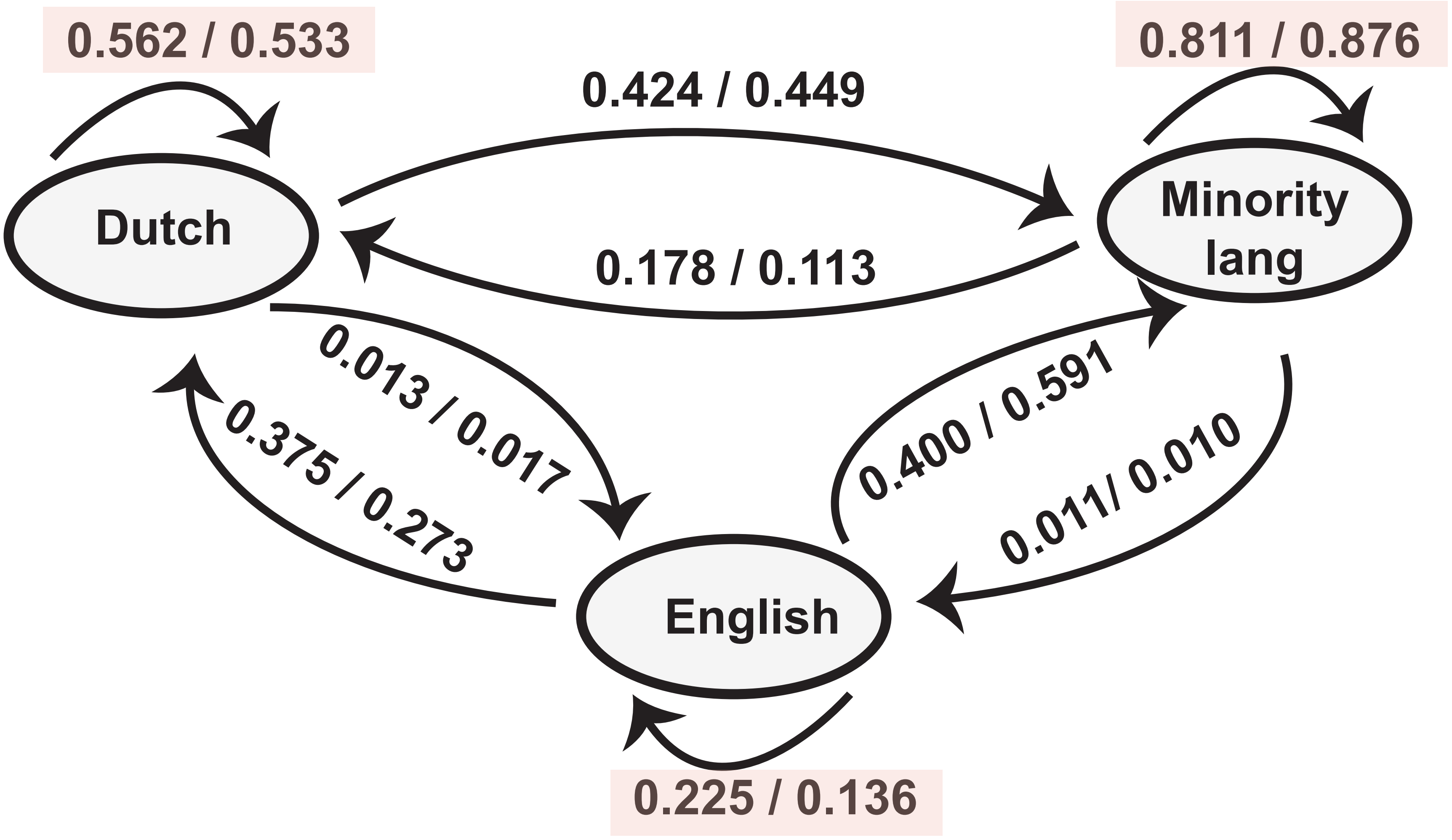
Influence of previous tweet I



Influence of previous tweet I



Influence of previous tweet I

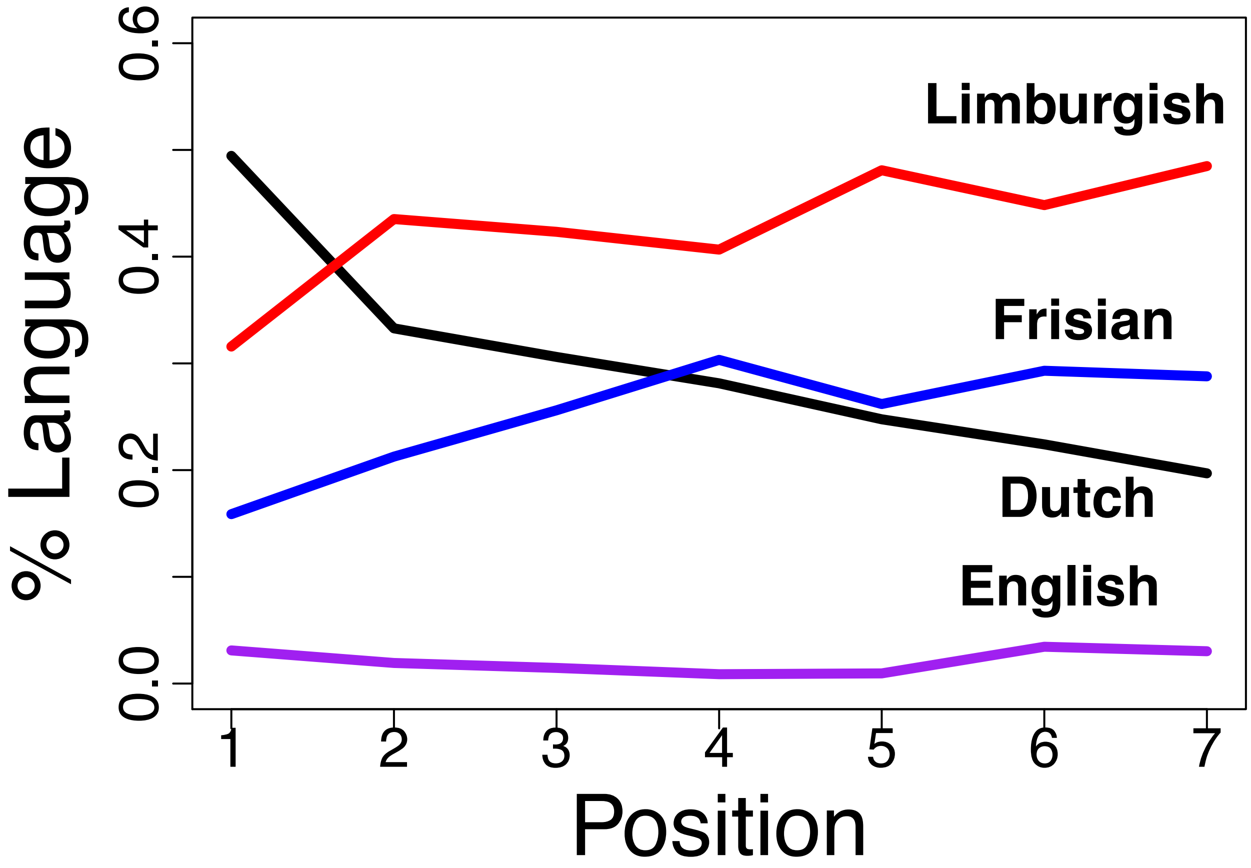


Influence of previous tweet II

	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

Logistic regression model (*** $p < 0.001$, ** $p < 0.01$)
Dependent variable = Tweet in minority language?

Language choice over time



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 entitities

Automatic Language Identification

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

… languages are not bounded, countable entitities

- 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