Using Twitter for Linguistic Research

Erik Tjong Kim Sang
erikt(at)xs4all.nl

18 June 2011

Twitter

Twitter is a website on which users broadcast text messages of up to 140 characters

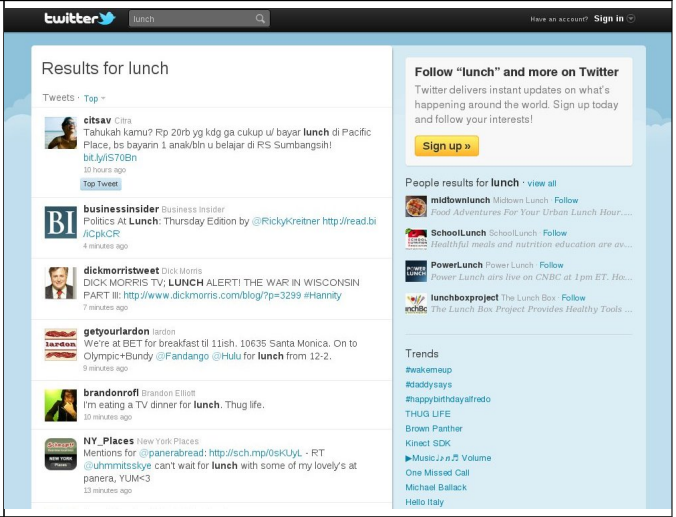
Users can subscribe to messages of users they like (follow users)

There are more than 200 million users world-wide

The users write more than 2 million Dutch messages each day

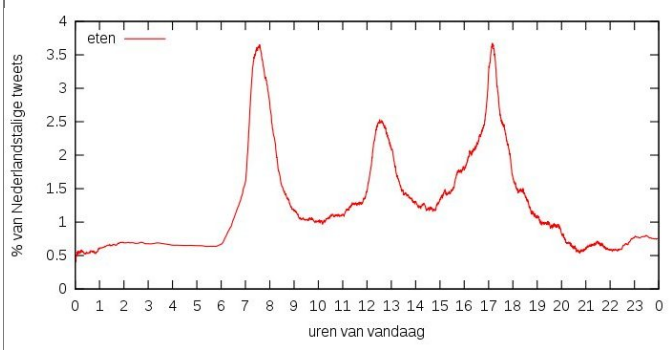
What do Twitter messages look like?

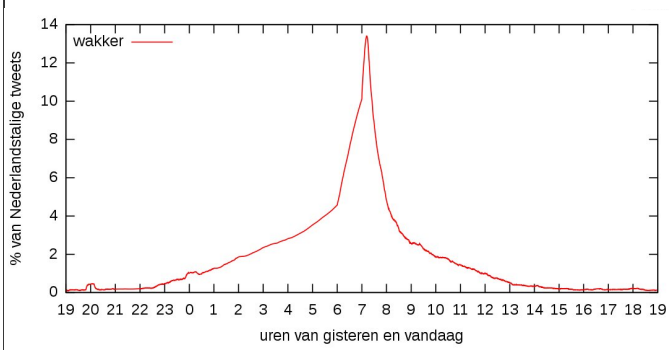
- a maximum of 140 characters of text
- the user name of the sender
- the date and time of the broadcast
- text may refer to other users like: @univgroningen
- text may include content tags like: #lunch

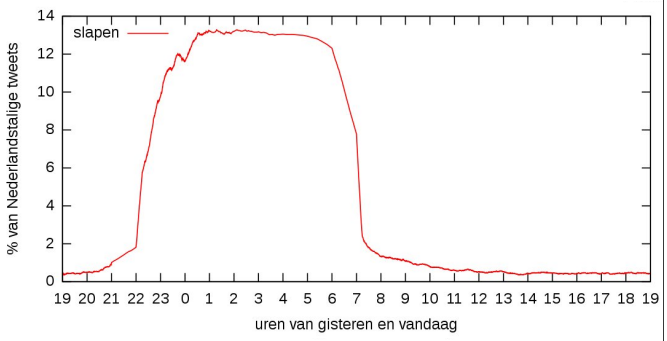


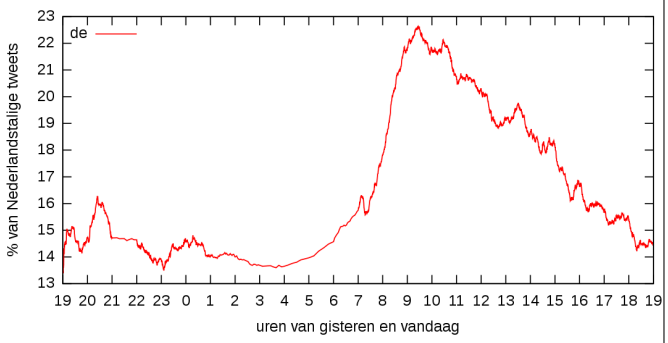
WHAT CAN WE DO WITH THIS DATA?

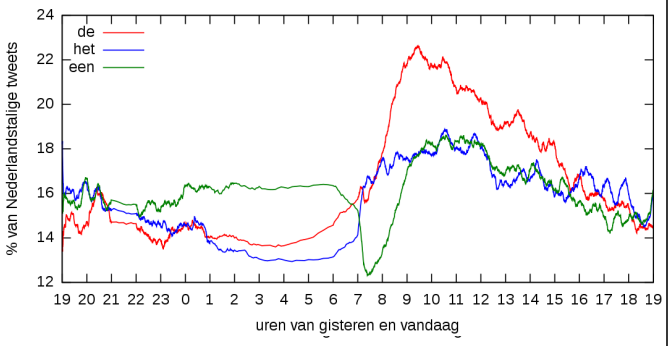
TRACKING WORDS











MALE VS FEMALE SPEECH

WHAT WORDS DO THEY USE?

Can we determine the gender of a Twitter user?

Short answer: no

But we could try to guess the gender from the user name:

- Male: Riccardo24ZD, MerlijntjeNL, dannykarimi, ...
- Female: L1nda_M, DeniesjeD, _Aliceeeh, ...
- Unknown: TaalNazi, nieuwsmedia, ikeetkabouters

Experiment setup

- Names of the top 1250 frequent posters were inspected
- 200 presumed male and female users were selected
- All their messages were collected: 265,000 male and 408,000 female
- Words in the two text collections were counted and compared with
  $\frac{f_{male}(w) - f_{female}(w)}{\sqrt{f_{male}(w) + f_{female}(w)}}$ (t-score: Church, Gale, Hanks, Hindle, 1991)
- Words used by fewer than 10 members of any group, were ignored

Typical words used per gender on Twitter

1.	ik	11.	omg	1.	goedemorgen	11.	natuurlijk
2.	me	12.	he	2.	de	12.	tweeps
3.	echt	13.	oke	3.	een	13.	dag
4.	mn	14.	gwn	4.	fijne	14.	vd
5.	hihi	15.	papa	5.	ha	15.	dan
6.	mama	16.	heb	6.	dank	16.	jouw
7.	ga	17.	even	7.	voor	17.	ghehe
8.	dr	18.	ma	8.	mooie	18.	vrijheid
9.	eh	19.	ben	9.	gewenst	19.	radio
10.	wil	20.	niet	10.	koffie	20.	goede

Typical word pairs used per gender on Twitter

1.	ik wil	1.	fijne dag
2.	ik ga	2.	voor jou
3.	ik heb	3.	je klaar
4.	ik moet	4.	mooie dag
5.	ik ben	5.	dank je
6.	als ik	6.	aan de
7.	dat ik	7.	voor je
8.	en ik	8.	een mooie
9.	ik haat	9.	van de
10.	oke ik	10.	slaap lekker

Concluding remarks

- Twitter can provide us with a lot of text data
- Data is available in many languages, for example in Dutch
- Basic analysis of this data can provide interesting results
- It is an excellent source for corpus linguists

THE END

Create your own Twitter graphs at http://www.let.rug.nl/erikt/twitter/