HET GEBRUIK VAN TWITTER VOOR TAALKUNDIG ONDERZOEK

Erik Tjong Kim Sang
Rijksuniversiteit Groningen
erikt(at)xs4all.nl

Samenvatting

Twitter is een sociaal netwerk waarin gebruikers berichten van maximaal 140 tekens versturen die meestal door iedereen kunnen worden gelezen. Twitterberichten hebben een aantal eigenschappen die ze interessant maken voor taalkundig onderzoek. Ten eerste zijn er heel veel van beschikbaar, alleen al voor het Nederlands meer dan twee miljoen per dag. Ten tweede bevatten zij allerlei interessante metadata, zoals dag en tijdstip van versturen, informatie over de persoon die het bericht verstuurde en onderwerptags (de zogenoemde hashtags) die door de verzender worden toegevoegd.

In dit paper zullen we uitleggen hoe Twitterberichten kunnen worden gebruikt voor taalkundig onderzoek. We beschrijven automatische processen waarmee Twitterberichten kunnen worden verzameld, op taal kunnen worden geselecteerd, kunnen worden getokeniseerd en gevisualiseerd. Daarna presenteren we toepassingen gebaseerd op een analyse van de Twitterberichten, onder andere een vergelijking van de woorden gebruikt door mannen en vrouwen, en het volgen van het gebruik van woorden door de tijd.

1. Inleiding

Twitter is een snelgroeiend online sociaal netwerk dat dagelijks door miljoenen mensen wordt gebruikt. De kerndienst die door de website wordt aangeboden is uitwisseling van korte tekstberichten, van maximaal 140 tekens. Berichten kunnen de status publiek of privé hebben. Publieke berichten zijn leesbaar voor iedere internetgebruiker.

Het aantal profielen op Twitter werd begin 2011 op 175 miljoen geschat. Ook in Nederland wordt de dienst gebruikt. In maart 2011 werden meer dagelijks meer dan 2 miljoen publieke Twitterberichten of tweets in het Nederlands verstuurd. Dit enorme aantal maakt Twitter een interessante bron voor corpustaalonderzoek.

Naast de tekst van maximaal 140 tekens bevatten Twitterberichten ook meta-informatie die interessant kan zijn voor onderzoek. Zo zijn de tijd en datum van het versturen van het bericht beschikbaar net zoals summiere informatie over de verzender. Daarnaast kan de tekst van een bericht ook tags bevatten die het onderwerp van het bericht aangeven. Deze tags worden door verzenders zelf

toegevoegd om het voor lezers mogelijk te maken berichten over een bepaald onderwerp terug te vinden.

Dit alles maakt Twitter een interessante bron voor taalkundig onderzoek. De webservice speelt hier op in door zelf software aan te bieden waarmee nieuwe berichten door onderzoekers kunnen worden verzameld. Twitterteksten zijn in diverse talen geschreven. Als uit de berichtenstroom de teksten uit een bepaalde taal kunnen worden geselecteerd dan levert dit een mooie bron van hedendaags taalgebruik.

In dit paper zullen we uitleggen hoe Twitterberichten in een bepaalde taal automatisch kunnen worden verzameld. Daarnaast geven we twee voorbeelden van gebruik van de aldus verzamelde teksten.

2. Verzamelen van recente Twitterberichten

Twitter biedt zijn berichten op een aantal manieren aan. Ten eerste is er de zoekfunctie op de website, waarmee kan worden gezocht naar recente berichten die een bepaald woord bevatten. Daarnaast kunnen willekeurige selecties van berichten worden opgevraagd. De website biedt drie volumes aan: sprinkler (kleine selectie), gardenhose (grotere selectie) en firehose (alles). Voor het gebruik van de laatste twee moet een contract worden afgesloten met het bedrijf. Wij hebben alleen de Nederlandstalige berichten nodig en gebruiken daarom de zoekfunctie in combinatie met een lijst met veelvoorkomende Nederlandse woorden.

Voor het verzamelen van Twitterberichten, kan het programma cURL[1] worden gebruikt. Met dit programma kunnen bestanden automatisch worden opgehaald van websites. Voor de Twitterteksten heeft het programma drie soorten informatie nodig:

- het internetadres van de dienst die de Twitterberichten aanbiedt:
  http://stream.twitter.com/1/statuses/filter.json
- een geldige gebruikersnaam en wachtwoord voor de website van Twitter
- de locatie van een bestand met woorden waarnaar moet worden gezocht

Het commando met alle benodigde informatie ziet er als volgt uit:

```
curl -d '@BESTAND' URL -u 'GEBRUIKER:WACHTWOORD'
```

Hierbij is URL het bovenstaande internetadres van het programma dat Twitterberichten aanbiedt. GEBRUIKER en WACHTWOORD zijn een geldige Twittergebruikersnaam en bijbehorend wachtwoord die bij de website van Twitter http://twitter.com/ kunnen worden aangevraagd. BESTAND is de naam van een bestand met woorden waarnaar gezocht moet worden. Het bestand bestaat uit 1 regel: track= gevolgd door woorden gescheiden door komma's (zie Appendix).

3. Taalkeuze voor Twitterberichten

Voor ons onderzoek zijn we alleen geïnteresseerd in Nederlandstalige berichten. Echter Twitter is internationaal en de berichtenstroom bevat dan ook teksten in verschillende talen. We selecteren uit deze stroom berichten de Nederlandse met behulp van twee verschillende technieken:

1. We selecteren alleen berichten die een veelvoorkomend Nederlands woord bevatten, of een onderwerptag (hashtag) die vaak voorkomt in Nederlandstalige berichten
2. Vervolgens beoordelen we de overgebleven berichten met een automatische taalrader en bewaren we alleen de berichten die volgens het programma een grote kans hebben om in hun geheel Nederlandstalig te zijn

Voor de eerste selectie hebben we eerst gekeken naar wat de meestvoorkomende woorden waren in een verzameling van handmatig geselecteerde Nederlandstalige berichten.

1.	ik	de
2.	je	van
3.	de	het
4.	een	een
5.	en	in
6.	het	en
7.	niet	dat
8.	rt	te
9.	is	is
10.	dat	op

Tabel 1: Frequente woorden in Nederlandstalige Twitterberichten (linkerrijtje) en in Nederlandse krantentekst (rechts). Alle woorden zijn eerst omgezet in kleine letters

In Tabel 1 staan de tien meest voorkomende woorden in Nederlandstalige Twitterberichten. Opvallend is dat deze niet hetzelfde zijn als in krantentekst. Uit de tabel blijkt dat Twitter, wat woordfrequenties betreft, meer lijkt op gesproken taal dan geschreven taal. [2] Niet alle top-10-woorden zijn geschikt om Nederlandstalige berichten mee te selecteren. Zo zijn *je* en *de* ook frequente Franse woorden, is *en* regelmatig terug te vinden in Spaanse teksten en is *is* een veelgebruikt Engels woord. Tenslotte wordt *rt* (oorspronkelijk *RT* voor retweet) regelmatig gebruikt in de meeste talen die het Latijnse alfabet gebruiken. Deze woorden zijn daarom weggelaten bij de selectie van Nederlandse Twitterteksten. Een complete lijst van de gebruikte woorden is te vinden in de appendix.

Ondanks het inperken van de woordenlijst, bleken de geselecteerde berichten nog steeds teksten in diverse talen te bevatten. Om dit probleem op te lossen, hebben we een taalrader gebouwd die vermeende niet-Nederlandse

berichten kon herkennen. We hebben 500 willekeurige tweets handmatig geannoteerd als wel-Nederlandstalig of niet-Nederlandstalig. Vervolgens is met een leeralgoritme (Naive-Bayes) bepaald hoe vaak woorden optreden in Nederlandse en niet-Nederlandse Twitterberichten. Op basis hiervan kon een score over het Nederlands-zijn worden toegekend aan nieuwe berichten.

Na inspectie van de uitvoer van de taalrader hebben we ervoor gekozen om alleen Twitterberichten met een niet-negatieve score te gebruiken. Ook Twitterberichten met negatieve scores (tot ongeveer -10) konden Nederlandstalig zijn. We missen door deze aanpak een aantal berichten maar de dagelijkse verzameling Nederlandstalige berichten is al zo groot dat we niet verwachten de omissies te zullen missen.

4. Tokenisatie van Twitterberichten

De volgende stap in het voorbewerkingsproces was het onderverdelen van Twitterberichten in woorden: tokeniseren. Voor deze taak is standaardsoftware beschikbaar. We hebben hier gekozen voor een eenvoudig zelfgeschreven algoritme dat leestekens verwijderd van de voorkant en de achterkant van een woord. Een voordeel van het gebruik van een eigen programma is dat we rekening konden houden met de speciale vorm van Twitterberichten. We hebben bijvoorbeeld hashtags behouden: de hashtekens zijn niet verwijderd van de woorden. Hierdoor konden we de hashtags ook in de latere stadia van het verwerkingsproces herkennen.

5. Vergelijken van mannelijke spraak en vrouwelijke spraak

Twitter wordt door zowel mannen als vrouwen gebruikt en het zou interessant zijn om te kijken of het taalgebruik in de twee groepen verschillend is. Helaas bevatten de gebruikersprofielen op Twitter geen informatie over het geslacht van de gebruiker. Daarom hebben we een deel van de gebruikers handmatig verdeeld in drie groepen op basis van hun gebruikersnaam: mannelijk, vrouwelijk of onbekend. De derde categorie was nodig omdat niet altijd uit een naam te zien is wat het geslacht is van een gebruiker, bijvoorbeeld bij gebruikers met fantasienamen (ikeetkabouters) of namen van organisaties (nieuwsberichtnl) . De berichten van de gebruikers met dergelijke namen hebben we niet gebruikt bij deze studie.

Tweehonderd veelschrijvende Twittergebruikers hebben we op basis van hun gebruikersnaam ingedeeld in de groep mannen (100) en vrouwen (100). Voorbeelden namen van de eerste gebruikersgroep zijn joost_vanharen, Patriick_B, Kevinvk123, JobBeintema en jeremynac. Voorbeelden van de tweede groep zijn MirjaamMannesse, DeniesjeD, DoniiaB, anna_wetering en DikraEH. De 200 gebruikers schreven samen ongeveer 22.000 Twitterberichten. De

doorgestuurde berichten van andere gebruikers, de zogenaamde retweets, zijn hier bij niet meegenomen.

We zijn geïnteresseerd in woorden die vaker door mannen worden gebruikt dan door vrouwen en omgekeerd. Om deze woorden automatisch uit de berichten te halen, hebben we een statistische metriek gebruikt, voorgesteld in Church, Gale, Hanks en Hindle (1991). Deze metriek berekent het verschil van de relatieve frequenties van een woord in twee verzamelingen teksten en deelt die door de som van de geschatte variantie van de twee frequenties:

$P_m(w)$ = relatieve frequentie van woord w in mannenteksten
$P_v(w)$ = relatieve frequentie van woord w in vrouwenteksten
variantie = $sqrt(P_m(w)/N_m + P_v(w)/N_v)$
$t = (P_m(w) - P_v(w)) /$ variantie

Hier worden de relatieve frequenties berekend door het aantal keer dat een woord voorkomt te delen door het totaal aantal woorden in de bewuste verzameling. De variantie die hoort bij een woord wordt geschat door de twee relevante relatieve frequenties door de totale woordaantallen (N_m en N_v) te delen en van de som de wortel te berekenen. Tenslotte is de score t van een woord gelijk aan het verschil tussen de twee frequenties gedeeld door de variantie. Woorden met een positieve score worden vaker door mannen gebruikt en woorden met een negatieve score vaker door vrouwen.

Deze metriek kent aan laagfrequente woorden vaak extreme scores toe. Deze woorden zijn minder interessant en we hebben daarom voor dit onderzoek alleen maar gekeken naar woorden die in minstens één groep door 10 of meer personen werd gebruikt. Op basis daarvan hebben we de informatie in Tabel 2 afgeleid: de tien woorden met de meest extreme positieve en negatieve scores berekend op basis van één dag Nederlandstalige Twitterdata:

	Mannenwoorden	Vrouwenwoorden
1.	noemen	misselijk
2.	gingen	kleren
3.	makkelijk	omg
4.	één	schattig
5.	hè	gebeurd
6.	ipod	mams
7.	Nederland	diegene
8.	reactie	jaloers
9.	chill	vieze
10.	twee	mam

Tabel 2: Top-10 van woorden in de Twitterdata die meer door mannen worden gebruikt dat door vrouwen (linkerrijtje) en andersom (rechts)

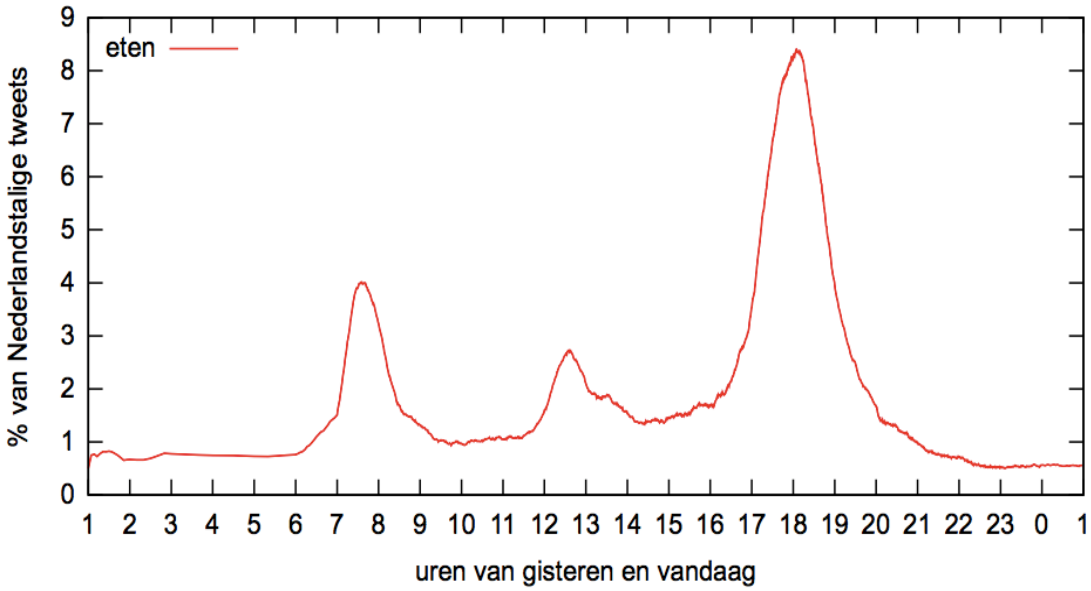
Bij de mannen vallen de woorden met accenten op op plaatsen 4 en 5 en het technische woord op plaats 6. Bij de vrouwen vallen op de twee varianten voor moeder op plaatsen 6 en 10, en de gevoelswoorden op plaatsen 4 en 8. De vrouwengroep bevat kennelijk diverse jongeren en communiceert vaker over gevoelens dan de mannengroep. Dit zijn geen verrassende observaties. Zij laten wel zien dat Twitter kan worden gebruikt voor het vergelijken van het taalgebruik van groepen personen. Het zou interessant zijn om meer van dit soort Twitterdata in te zetten voor vergelijkbaar onderzoek.

6. Het volgen van woorden door de tijd

Aan elk bericht op Twitter is een datum en een tijd gekoppeld van het moment van verzenden. Dit maakt het mogelijk om dit medium te gebruiken voor het volgen van het gebruik van taaluitingen over de tijd. We moeten daarbij wel nadenken hoe dit het best in beeld kan worden gebracht. Als we zonder meer het gebruik van een woord per seconde zouden weergeven dan krijgen we een grafiek met een paar pieken van één gebruik per seconde en vele malen geen gebruik van het woord. Dit is niet erg interessant.

We kunnen een interessantere grafiek tekenen als we niet elk gebruik van een woord apart aangeven in de grafiek maar als we voor een langere tijdsduur aangeven hoe vaak woorden zijn gebruikt. We hebben er zelf voor gekozen om niet per vaste tijdseenheid berichten te tellen maar per vast aantal berichten. Per blok van 50.000 berichten tellen we hoe vaak elk woord is gebruikt. Daarna halen we de 1000 oudste berichten uit het blok, plaatsen er 1000 nieuwe bij en tellen we opnieuw, enzovoorts. Door op deze manier te tellen met een zogenaamd *sliding window*, kunnen we een goed beeld krijgen over hoe vaak woorden worden gebruikt op Twitter.

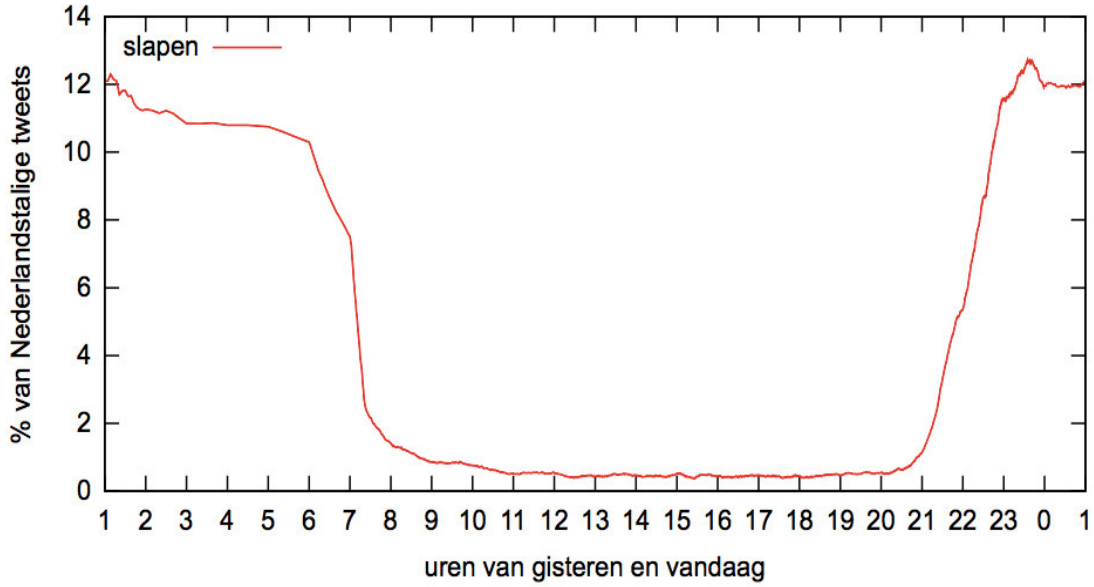
De frequentie van de Twitterberichten is niet constant over de gehele dag. Als we absolute aantallen woorden zouden bijhouden per tijdseenheid dan kan een stijging betekenen dat het woord populairder is geworden. Maar het zou ook kunnen dat het woord even populair is gebleven maar dat het volume van berichten is toegenomen. Om deze laatste verklaring uit te sluiten, meten we de relatieve frequentie van een woord: het aantal keer dat een woord is gebruikt gedeeld door het totale aantal woorden in de meetperiode.



Figuur 1: Relatieve frequentie van het woord *eten* op dinsdag 5 april 2011. De grafiek bevat drie pieken die corresponderen met de tijdstippen van ontbijt, lunch en diner.

Figuur 1 laat het gebruik van het woord *eten* zien op een doordeweekse dag. De grafiek bevat drie pieken zien die lijken te corresponderen met ontbijt (kwart voor acht), lunch (half een) en diner (zes uur). Deze grafiek biedt weinig nieuwe informatie. Maar wat hem bijzonder maakt is dat hij geheel automatisch is samengesteld. Mensen gebruiken Twitter om berichten rond te sturen over hun activiteiten. Deze grafiek laat zien dat als mensen gaan eten zij daar ook over twitteren. Het is net alsof je via Twitter een rechtstreekse verbinding hebt met het brein van de gemiddelde Twitteraar.

Een grafiek over het dagelijkse woordgebruik op Twitter kan ook vragen oproepen. Kijk maar eens naar de grafiek bij Figuur 2:



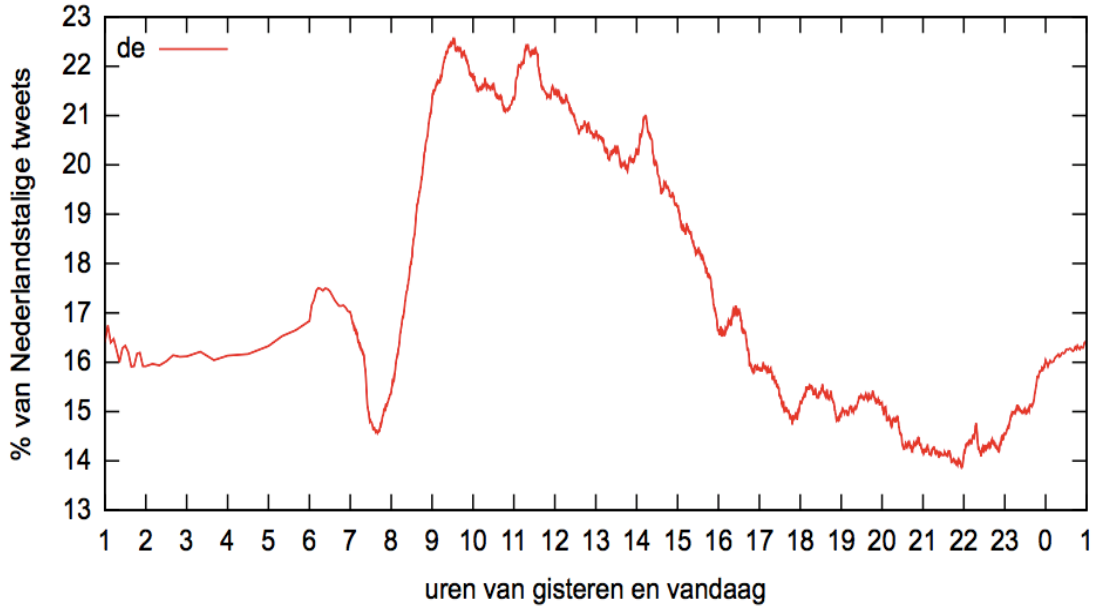
Figuur 2: Relatieve frequentie van het woord *slapen* op dinsdag 5 april 2011. De grafiek bevat een langdurende piek tussen ongeveer tien uur 's avonds en zeven uur 's morgens.

Figuur 2 laat zien hoe vaak het woord *slapen* op een doordeweekse dag wordt gebruikt. De grafiek bevat een piek tussen tien uur 's avonds en zeven uur 's morgens en dat zijn de tijdstippen waartussen de meeste mensen slapen. Verrassend is dat het woord dezelfde populariteit behoudt gedurende de hele nacht. Hoe is dit mogelijk? Dat mensen twitteren tijdens het eten is nog voorstelbaar maar niet dat zij twitteren terwijl ze slapen.

De vorm van de grafiek kan worden verklaard door het feit dat het gaat om relatieve frequenties. Het woord slapen wordt minder gebruikt rond drie uur 's nachts dan om elf uur 's avonds maar op het eerste tijdstip worden in totaal ook minder berichten verstuurd. De mensen die slapen, twitteren niet maar onder de mensen die niet slapen blijft het woord populair en dat is wat te zien is in de grafiek.

Tenslotte nog een grafiek die lastiger is om uit te leggen. Figuur 3 laat het gebruik van het woord *de* op een doordeweekse dag zien. Dit is een veelvoorkomend functiewoord en men zou kunnen verwachten dat het relatieve gebruik van het woord constant is. Dit is echter niet het geval. Het woord *de* wordt 's morgens het meest gebruikt waarna het relatieve gebruik sterk daalt tot zes uur 's middags. Hoe kan dit worden verklaard?

Een mogelijke aanwijzing zou kunnen liggen in het gebruik van het woord *ik* waarvan de grafiek precies andersom verloopt: laag in de ochtend en dan 's middags sterk stijgend tot acht uur 's avond. Misschien wordt Twitter in de middag en in de avond meer gebruikt voor persoonlijke berichten en wordt daarbij meer gebruik gemaakt van de eerste persoon enkelvoud en minder van lidwoorden.



Figuur 3: Relatieve frequentie van het woord *de* op dinsdag 5 april 2011.

Geheel onverwacht is dit frequente lidwoord populairder in de ochtend dan in de avond. Deze drie grafieken laten zien dat met eenvoudige middelen interessante informatie over het dagelijks gebruik van woorden uit Twitterdata. Deze informatie kan op veel manieren worden toegepast, bijvoorbeeld voor het schatten van kijkcijfers voor tv-programma's tot zelfs het meten van politieke voorkeuren van de Twitterpopulatie, zie bijvoorbeeld Tumasjan, Sprenger, Sandner & Welpe (2010).

7. Conclusies

Berichten op het online sociale netwerk Twitter bevatten maximaal 140 tekens maar door het grote aantal, meer dan twee miljoen Nederlandstalige berichten per dag, zijn zij een prachtige bron van data voor taalkundig onderzoek. In dit paper hebben we laten zien hoe Nederlandse Twitterberichten automatisch kunnen worden verzameld. Daarna hebben we twee toepassingen van de data besproken: het vergelijken van het taalgebruik van mannen en vrouwen, en het meten van het gebruik van woorden over de tijd.

De databron kan op diverse andere manieren worden ingezet bij onderzoek. Een voorbeeld is een studie naar het gebruik van grammaticale constructies op Twitter. Hiervoor zou het handig zijn als de berichten automatisch syntactisch kunnen worden ontleed. Wij zijn op dit moment aan het kijken hoe onze op krantentekst getrainde syntactische parser Alpino kan worden ingezet voor automatische grammaticale analyse van Twitterberichten.

Een andere toepassing voor Twitterdata is het bepalen en vergelijken van het taalgebruik van groepen gebruikers, bijvoorbeeld op basis van leeftijd of

geografische locatie. Dergelijke informatie is slechts beperkt gespecificeerd in de profielen van de gebruikers. Het is daarom goed dat de Universiteit van Nijmegen in januari 2011 in het kader van het SoNAR-project is begonnen met het verzamelen van dergelijke metadata voor Nederlandse Twittergebruikers.

Noten

[1] cURL is gratis beschikbaar vanaf http://curl.haxx.se

[2] Woorden die horen bij dialogen, zoals *ja*, zijn wat minder frequent dan bij spreektaalcorpora gevonden wordt. Hetzelfde geldt voor pauzewoordjes als *eh*. *Ja* staat op de plaats 2 in de lijst van meest frequente woorden in De Jong e.a. (1979), en *eh* op plaats 3. Woorden die een tekstverband aangeven, zoals *dat* (Bouma 2008), nemen ook een iets minder prominente positie in, vanwege de geringe lengte van de gemiddelde Tweet.

Bibliografie

Bouma, Gerlof

2008 *Starting a sentence in Dutch. A corpus study of subject and object fronting.* Dissertatie, Rijksuniversiteit Groningen.

Church, Kenneth, William Gale, Patrick Hanks & Donald Hindle

1991 'Using Statistics in Lexical Analysis.' In: Uri Zernik (red.), *Lexical Acquisition - Exploiting On-Line Resources to Build a Lexicon*, Lawrence Erlbaum Associates, 115-164.

Jong, Eveline D., red.

1979 *Spreektaal. Woordfrequenties in gesproken Nederlands.* Bohn, Scheltema & Holkema, Utrecht.

Tumasjan, Andranik, Timm Sprenger, Philipp Sandner & Isabell Welpe

2010 'Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.' In: *Proceedings of the Fourth AAAI conference on Weblogs and Social Media*, 178-185.

Appendix: praktische informatie

Het gebruikte Linuxcommando voor het verzamelen van berichten:

```
curl -d '@sleutelwoorden.txt'
http://stream.twitter.com/1/statuses/filter.json -u
"NAAM:WACHTWOORD"
```

Inhoud van het bestand `sleutelwoorden.txt`:

```
track=een,het,ik,niet,maar,voor,ook,als,heb,naar,nog,echt,m
oet,weer,mijn,zijn,bij,jij,toch,lekker,geen,gewoon,gaat,mee
r,slapen,weet,mensen,alleen,kijken,leren,heeft,vandaag,eens
,hoor,uur,jou,veel,denk,maken,leuk,heel,zou,daar,komt,eten,
iets,vind,hebben,altijd,vanavond,jullie,thuis,iemand,helema
al,waar,waarom,wakker,komen,beetje,nieuwe,worden,steeds,gez
ellig,straks,kunnen,zeggen,iedereen,ofzo,omdat,werken,allem
aal,moeten,andere,jaar,terug,staat,kut,ging,erg,zien,vroeg,
bijna,zelf,zegt,vrij,zeker,werk,#gtst,#tienerthings,#bzv,#w
iedoethet,#durftevragen,#nieuws,#tienerfeiten,#dutchteenage
rs,#dwdd,#penw,#widm,#slajezelf,#voetbalfans,#pownews,#geen
zin,#slapen,#lekker,#rtl7
```

Adres zoekdemo: `http://www.let.rug.nl/erikt/twitter`