Wie is van...

Woorden, woorden en nog eens woorden: via sociale media gooien we er petabytes tegelijk van op het internet. Martijn Wieling van de Rijksuniversiteit Groningen vertelt hoe de computationele taalkunde daarvan handig gebruikmaakt. ‘Het besef dat in dit vakgebied veel banen te vinden zijn, wordt steeds groter.’ *Door Enith Vlooswijk*

Snel een aardbeving detecteren met Twitter

‘Engelstaligen zetten het puntje van hun tong verder naar voren dan Nederlanders als ze de “th” uitspreken. Ik onderzoek zulke verschillen in de uitspraak van het Engels tussen Nederlanders en Engelstaligen, kijk hoe de beweging van de tong hun uitspraak beïnvloedt. Zulke kennis is bruikbaar voor mensen die Engels als tweede taal leren.

Ik heb Informatica gestudeerd, maar vond taal altijd al leuk. Daarom volgde ik ook enkele vakken bij Informatiekunde. Later promoveerde ik bij de Rijksuniversiteit Groningen op de kwantitatieve analyse van verbanden tussen dialecten en verschillende sociale en geografische factoren.
Mijn vakgebied, de computationele taalkunde, maakt veel gebruik van computertechnologie. We gebruiken grote hoeveelheden data – woorden, zinnen, transcripties van geluidsopnames – en leggen verbanden.

Dat lijkt iets actueels, maar al ruim voor de eeuwwisseling, in 1992, startten wetenschappers in de Verenigde Staten het Linguistic Data Consortium op. Er was toen een groot tekort aan digitaal beschikbare data die spraak- en taaltechnologen konden gebruiken bij het ontwikkelen van nieuwe systemen. Als je tot die tijd digitale data wilde gebruiken in je onderzoek, moest je die grotendeels zelf verzamelen. Internet was nog in opkomst. Voor kranten, televisie en radioprogramma’s verzorgde een kleine groep mensen de inhoud. Nu creëert iedereen zijn eigen data op Facebook, Twitter, Reddit, enzovoorts. Vaak is dat geschreven taal en daar kunnen wij heel interessante dingen mee doen.

Martijn Wieling (1981) combineert taalkunde met informatiekunde en statistiek. Hij promoveerde in 2012 aan de RUG op het proefschrift ‘A Quantitative Approach to Social and Geographical Dialect Variation’. Sinds 2013 werkt hij daar aan een Veni-project waarin hij tong- en lipbewegingen van mensen bestudeert die een tweede taal leren. Wieling is lid van De Jonge Akademie.

Op grond van Twitterdata kun je bijvoorbeeld vaak sneller een aardbeving detecteren dan seismografen dat kunnen. En bij de verkiezingen van de Provinciale Staten in 2011 hebben collega’s van me uitslagen voorspeld op basis van uitspraken op Twitter. Uiteindelijk zaten ze niet veel verder van de verkiezingsuitslag af dan de grote opiniepeilers.

We moeten wel oppassen dat we de ontdekte verbanden niet gaan zien als wetmatigheden. Met technieken uit de computationele taalkunde zijn we bijvoorbeeld behoorlijk goed in staat om te bepalen of twee teksten door dezelfde persoon zijn geschreven. Dat betekent echter niet automatisch dat die technieken ook goed genoeg zijn om te gebruiken bij het identificeren van misdadigers op grond van geschreven tekst.
Het besef dat in dit vakgebied veel banen te vinden zijn, wordt steeds groter. Dat merken we hier aan de groeiende populariteit van de studie Informatiekunde: eerst hadden we relatief weinig studenten, nu schieten de cijfers omhoog naar vijftig of zestig nieuwe studenten per jaar. Mensen zien dat grote bedrijven als Google heel veel gebruik maken van methoden uit de computationele taalkunde. Daar komen steeds meer bedrijven bij. We willen dat onze telefoon gesproken opdrachten steeds beter begrijpt; en het zou natuurlijk fantastisch zijn als we ooit door middel van automatische vertaling een gesprek konden voeren met een anderstalige. Om dat voor elkaar te krijgen, kunnen we niet zonder taaltechnologie.

Ons vakgebied leent zich erg goed voor het openbaar delen van onderzoeksmethoden en -resultaten. Ik vind het belangrijk dat anderen mijn onderzoek kunnen repliceren en controleren. Als ik een programma heb geschreven om data te analyseren, is het bovendien fijn dat anderen het wiel niet opnieuw hoeven uit te vinden.

Toch zijn veel wetenschappers terughoudend met het delen van data en methoden. Soms zijn ze bang dat anderen ermee aan de haal gaan en er eerder over publiceren dan zij. En als hun methode achteraf een fout blijkt te bevatten, moet hun artikel misschien worden teruggetrokken. Ik denk dat wetenschappers minder hun eigen belang voorop moeten stellen en meer de vooruitgang van de wetenschap als geheel.’ I/O

Meer informatie: *www.martijnwieling.nl*