Social Media in SoNaR

Tot voor kort waren er weinig tot geen corpora van sociale media zoals chats, tweets en sms. Er was dus nauwelijks corpusonderzoek mogelijk naar deze nieuwe vormen van communicatie. Het SoNaR corpus opende nieuwe onderzoekswegen naar sociale media, doordat sociale media werden verzameld als onderdeel van het corpus. De auteurs waren verantwoordelijk voor de collectie van chats, tweets en sms.

Eric Sanders en Maaske Treurniet
CLST, Radboud Universiteit Nijmegen

Het SoNaR Nederlandstalig referentiecorpus is ontwikkeld als onderdeel van STEVIN. Eén van de doelen van het STEVIN programma was het realiseren van een adequate digitale taalinfrastructuur voor het Nederlands. Het ontwerpen van een referentiecorpus werd beschouwd als één van de vereisten voor het ontwikkelen van andere bronnen, tools en applicaties. Het corpus is van groot belang voor verder onderzoek in natural language processing. Toepassingen en onderzoek op het gebied van informatie-extractie, question-answering, documentclassificatie en automatisch samenvatten, die gebaseerd zijn op corpus gebaseerde technieken, kunnen profiteren van de grootschalige analyse van het corpus.

SoNaR bevat 500 miljoen woorden, uit Nederland en Vlaanderen. Het corpus is opgebouwd uit teksten in het hedendaagse Nederlands (vanaf 1954), verdeeld over uiteenlopende domeinen en genres. Bij het verzamelen van teksten is ook aandacht uitgegaan naar teksten waar gebruikers mee in aanraking komen via nieuwe, digitale media.

Nieuw taalgebruik

Met de komst van internet en mobiele communicatieapparatuur zijn er nieuwe manieren van communicatie ontstaan en daarmee ook nieuwe manieren van taalgebruik. De eerste methode die ontstond om met elkaar te communiceren via computernetwerken is e-mail. Deze vorm van communicatie heeft nog een grote gelijkenis met een conventionele brief. Snel daarna kwam echter het chatten, dat een compleet nieuwe manier van communiceren is. Chat is geschreven tekst, waarbij de deelnemers elkaar normaliter niet zien, maar die wel verloopt met de snelheid en beurtwisselingen die voorheen alleen bij spreken normaal waren. Men zou chats kunnen beschrijven als een soort getypte telefoongesprekken.
Sms is ontstaan toen de mobiele telefoon gemeengoed werd. Sms'en zijn losse berichten, met een maximale lengte van 160 karakters. Omdat invoer via een telefoon een lastig proces is, is het een relatief moeizame manier van communicatie.

Twitter is pas recent, sinds 2006, in gebruik, maar heeft sindsdien een duizelingwekkende vlucht genomen. Op moment van schrijven worden er dagelijks wereldwijd 300 miljoen tweets verstuurd, maar dit aantal verandert met de dag. Tweets hebben een grote gelijkenis met sms'en. Tweets bevatten maximaal 140 karakters en zijn normaliter losse berichten. Het grootste verschil met sms betreft het aantal ontvangers. Daar waar een sms meestal bij één persoon belandt, zijn tweets doorgaans openbaar en te lezen voor een soms grote schare volgers.

Deze nieuwe vormen van communicatie hebben elk geleid tot nieuwe specifieke vormen van taalgebruik. Het taalgebruik in chats is informeler dan eerder in geschreven taal gewoon was. Daarbij zijn chats zeer belangrijk geweest voor de ontwikkeling en het gebruik van emoticons (de zogenaamde 'smileys'). Vanwege de snelheid van chatconversaties en de beperkte lengte van sms- en twitterberichten is er een compact taalgebruik ontstaan met afkortingen en het weglaten van woorden en letters. In het Nederlands taalgebied is nog weinig onderzoek gedaan naar het taalgebruik in sociale media, onder

andere vanwege de afwezigheid van geschikte corpora.

Daar is nu verandering in gekomen. Als onderdeel van het SoNaR corpus, zijn chats, tweets en sms verzameld. Een collectie van chats en tweets aanleggen is, in tegenstelling tot sms, relatief eenvoudig. De meerwaarde van het corpus ligt echter in het verkrijgen van de rechten op gebruik en verspreiding, het verkrijgen van de persoonsgegevens van schrijvers en de anonimisering van de teksten. Hieronder staat kort beschreven hoe we de data, metadata en gebruikerstoestemming hebben verkregen en hoe de anonimisering al dan niet werd uitgevoerd.

Chats

Het Nederlandse gedeelte van de chats bestaat uit vier delen. Voor drie delen was een speciale chat-server ingericht om doelgroepen deel te laten nemen aan chat-sessies. In twee gevallen betrof dit middelbareschoolleerlingen en in het derde geval collega's van de onderzoeksgroep Taal&Spraak aan de Radboud Universiteit in Nijmegen. Het vierde deel bestaat uit msn-conversaties van vrienden en bekenden, die waren opgeroepen hun chats te verzamelen. Eén corpus met chatdata van middelbareschoolleerlingen was al verzameld voor het SoNaR-project was begonnen. Dit betreft het ChatIG corpus, in 2005 en 2006 verzameld door Wilbert Spooren en Tessa van Charldorp van de Vrije Universiteit in Amsterdam.
Alle deelnemers, of hun ouders, van de Nederlandse chatdata hebben toestemming gegeven voor distributie van de data en hebben hun metadata (leeftijd, geslacht, woonplaats) gegeven. De (nick)namen van de chatters zijn onherkenbaar veranderd. Verdere anonimisering bleek te tijdrovend om handmatig uit te voeren en te complex om te automatiseren.

Voor het Vlaams is toestemming bemachtigd om data van de grote chatservice chat.be te gebruiken. Er is collectieve toestemming gegeven door de eigenaar van de chatsite. Gedurende een aantal maanden in 2011 is data van het chatkanaal verzameld. Voor deze data is geen individuele toestemming en er is ook geen metadata van de gebruikers beschikbaar.

Al met al bevinden zich 2.194.592 chatregels in SONaR.

Tweets

Voor Twitter bestaat een API waarmee tweets verzameld kunnen worden. Er is enige onduidelijkheid over toestemming voor hergebruik, maar wij hebben de gebruiksvoorwaarden zo geïnterpreteerd dat de tweets in SoNaR opgenomen mogen worden, mits de tekst en de gebruikersnaam onveranderd blijven. Daarmee zijn zowel het toestemmings- als het anonimiseringsprobleem opgelost.

Er zijn twee manieren gebruikt om de metadata te verzamelen. Bij de eerste methode is een tweet verstuurd met daarin de oproep aan twitteraars hun persoonsgegevens te sturen vanwege de aanleg van het SoNaR-corpus. Deze tweet werd geretweet en zo ontstond een sneeuwbaleffect dat uiteindelijk resulteerde in berichten op de hoofdpagina van de Radboud Universiteit (ru.nl) en de bekendste nieuwswebsite van Nederland (nu.nl), waarin mensen werd gevraagd

hun gegevens te mailen. Daarnaast hebben we van bekende en semi-bekende Nederlanders en Vlamingen geslacht, leeftijd en woon- of geboorteplaats opgezocht op hun persoonlijke homepage of op Wikipedia.

In SONaR zijn 1.532.251 tweets opgenomen.

Sms

Het verzamelen van sms-berichten vraagt een specifieke aanpak vanuit het oogpunt van privacy en anonimisering. Er is samengewerkt met de National University of Singapore, School of Computing, die al eerder een verzameling sms'en had aangelegd. Er werd een website opgezet, waar instructies te vinden waren over drie verschillende manieren waarop sms'en bijgedragen konden worden. Allereerst was er de mogelijkheid om sms'en handmatig te kopiëren in een online formulier. Daarnaast konden eigenaars van een Android smartphone een zogenaamde App downloaden, waarmee de sms'en automatisch geëxtraheerd werden uit de telefoon en vervolgens via Gmail naar

SoNaR verstuurd werden. Tenslotte bevatte de website instructies om de sms'en via de PC te downloaden vanuit een iPhone of Nokiatoestel, waarna dit bestand via de SoNaR dropbox ingestuurd kon worden. Om toestemming te verkrijgen van de auteurs van iedere sms, werden de voorwaarden gemeld bij het uploaden van de sms, respectievelijk in de te verzenden mail of door een melding bij de dropbox. Anonimisering van de berichten werd grotendeels overgelaten aan de eigenaars van de sms'en. De originele telefoonnummers werden vervangen door een unieke code en bij sms die via de App werden ingezonden, werden daarnaast automatisch bankrekeningnummers en IP-adressen onherkenbaar gemaakt.

De website werd gepromoot via communicatieafdelingen van de verschillende partneruniversiteiten van SoNaR, via flyers, persberichten, sociale media en onder de eigen kennissenkring van medewerkers.

Al met al bevat SoNaR 42.358 sms-berichten.