De appel valt niet ver van de boom: afstanden berekenen tussen taalvarieteiten.

Tom Ruette

Manuscript van 10 december 2012

1 Afstanden en categorieën

De appel valt niet ver van de boom, of er is geen speld tussen te krijgen. En tussen de cultuur van België en Nederland ligt een wereld van verschil.

Om de complexe realiteit van alledag onder controle te houden schatten mensen de hele tijd afstanden, zoals de zegswijzen hierboven illustreren. Op basis van zulke afstanden kunnen we alle indrukken die op ons afkomen in groepjes of categorieën indelen: als twee indrukken zeer "dicht bij elkaar liggen", dan behoren ze tot dezelfde categorie. In zekere zin doen we iets gelijkaardigs met taal. Als iemand iets tegen ons zegt, beoordelen we bliksemsnel of die persoon een taal spreekt die dicht bij de onze ligt, of misschien een taal gebruikt die dicht bij die van een nieuwlezer aansluit. Aan de basis van zulke afstanden ligt het feit dat mensen taal gebruiken om te communiceren, maar ook om zichzelf een persoonlijkheid te verschaffen.

Met een technische term noemt men een type van taalgebruik dat typisch is voor bepaalde personen, karakters en/of situaties een *taalvariëteit*. Zo is een dialectspreker meestal oud en wekt hij doorgaans vertrouwen en gevoelens van authenticiteit op. Of een jongere die liever een "slikke jacka" draagt dan een mooie jas kan soms bevreemdend werken en onbegrip uitlokken. Deze voorbeelden tonen aan dateen taalvariëteit uit twee aspecten bestaat. Aan de ene kant is er het taalgebruik, zoals dialectische klanken of hippe woordkeuzes. Elke taalvariëteit bedient zich van een typisch groepje taalkenmerken dat we vanaf nu een *taalgebaseerde categorie* zullen noemen. Aan de andere kant is er ook het gevoel dat het gebruik van een taalvariëteit met zich mee brengt, zoals authenticiteit, vertrouwen, of bevreemding en onbegrip. Aan deze kant past ook de vaststelling dat een bepaalde taalvariëteit enkel gebruikt wordt door een specifieke groep van mensen of in bepaalde situaties. Zo is bijvoorbeeld de taalvariëteit *jongerentaal* uitsluitend geschikt voor jongeren, en is dialect niet gepast in een formele situatie. Groepen van mensen of situaties waarin bepaalde taalvariëteiten voorkomen, en de gevoelens die ze oproepen, noemen we vanaf nu *socio-situationele categorieën*.

Mijn onderzoek richt zich precies op het verband tussen wie we zijn — of wie we willen zijn —, de situatie waarin we ons bevinden, en ons taalgebruik. We vragen ons af de verschillen tussen socio-situationele categorieën gerelateerd zijn aan de verschillen in het taalgebruik. En zo komen we terug bij de afstanden waarmee dit

artikel begonnen is: *zijn de afstanden tussen socio-situationele categorieën gerelateerd aan afstanden tussen categorieën in het taalgebruik?* Of met behulp van een concreet voorbeeld: is de gevoelde afstand tussen de Wetstraat en de dorpstraat ook terug te vinden in de afstand tussen Wetstratees en Dorpstratees?

2 Fruitsap of suderans?

Vaak spelen persoonlijke gevoelens een grote rol bij het bepalen van afstanden tussen taalgebaseerde of socio-situationele categorieën. De puberende tiener schat de afstand tot volwassenheid bijzonder klein in, terwijl zijn ouders nog steeds een kind zien. En het Antwerpse dialect staat het dichtst bij de Standaardtaal volgens een rasechte sinjoor, maar niet volgens de Limburgse Maaslander. Hoewel deze gevoelsbaseerd afstanden bijzonder interessant zijn, beperkt mijn onderzoek zich tot objectief meetbare afstanden tussen taalgebaseerde categorieën. Bovendien spitst deze studie zich toe op de woordenschat, hoewel het taalgebruik verschillende aspecten kan beslaan, zoals uitspraak, woordvorming of zinsbouw. We onderzoeken in de woordenschat de variatie die er ontstaat doordat er verschillende mogelijkheden zijn om hetzelfde te zeggen. Zo kan men bijvoorbeeld een *broek van stevige stof, doorgaans in een blauwe kleur, geschikt voor niet al te formele situaties* benoemen met "jeans" of met "spijkerbroek". Een ander voorbeeld is het gebruik van "fruitsap" in Vlaanderen en "suderans" in Nederland om *drinkbaar sap van fruit* te benoemen. Zo een groep van woorden die hetzelfde uitdrukken noemen we een *lexicale variabele*.

We gebruiken de variatie bij het benoemen van begrippen om een *lexicale afstand* te meten tussen het taalgebruik in verschillende socio-situationele categorieën. Stel dat we het gebruik van "jeans" en "spijkerbroek" kunnen vaststellen in Belgische en Nederlandse modebladen. Mochten de modebladen uit beide landen altijd hetzelfde woord kiezen om een *broek van stevige stof, doorgaans in een blauwe kleur, geschikt voor niet al te formele situaties* te benoemen, dan is de afstand tussen het taalgebruik (in modebladen) van de landen niet zo groot. Dus, als er amper verschil is in de woordkeuze van beide landen (in modebladen), dan meten we een kleine afstand. Is het echter zo dat de Belgische modebladen voornamelijk "jeans" gebruiken, en de Nederlandse modebladen voornamelijk "spijkerbroek" gebruiken, dan is het verschil — en daarmee ook de afstand — tussen het taalgebruik in België en Nederland groot.

Mochten we die lexicale afstand tussen taalvariëteiten in het Nederlands enkel baseren op de bekende lexicale variatie met een België versus Nederland patroon (zoals "jeans" en "spijkerbroek" of "fruitsap" en "suderans") dan zouden we slechts een verschil kunnen vaststellen tussen België en Nederland. Maar natuurlijk varieert de woordenschat in het Nederlands lange meerdere dimensies dan enkel het nationale verschil. Daarom is het noodzakelijk om de lexicale afstandsmeting te baseren op (idealiter) alle mogelijke lexicale variabelen van het Nederlands. Om zo veel mogelijk lexicale variabelen te verzamelen maken we gebruik van een automatische methode die woorden met dezelfde betekenis kan vinden in een grote verzameling van teksten. Die methode wordt ook gebruikt in zoekmachines op het internet. Als je bijvoorbeeld met Google zoekt naar *delete all files*, dan wordt er impliciet ook gezocht naar *remove all files* en *clear all files*. Aan de basis van deze automatische methode ligt de vaststelling dat gerelateerde woorden ook in gelijkaardige contexten voorkomen. Zo zouden we de gelijkaardigheid van *ongeluk* en *ongeval* kunnen afleiden uit het feit dat deze woorden allebei vaak voorkomen

in de buurt van bijvoorbeeld *auto*, *slachtoffer* of *blikschade*. Met behulp van deze automatische methode zijn we erin geslaagd om een grote hoeveelheid groepjes van gelijkaardige Nederlandse woorden te vinden. En ook voor het Engels vonden we meerdere honderden groepjes.

3 Alle registers opengetrokken

Om taalgebaseerde afstanden te meten op basis van het woordgebruik in een aantal socio-situationele categorieën is aan de ene kant een grote verzameling van groepjes van gelijkaardige woorden nodig, zoals we hierboven hebben besproken. Aan de andere kant kiezen we ervoor om de socio-situationele categorieën te onderzoeken met behulp van teksten die representatief zijn voor deze categorieën, zodat taalgebaseerde verschillen geobserveerd kunnen worden in het woordgebruik in de teksten. De socio-situationele categorieën zijn zo gekozen dat ze zowel een nationale dimensie als een registerdimensie representeren. In het geval van de Engelstalige teksten is het ook mogelijk om een temporele dimensie te bestuderen. Heel concreet onderzoeken we voor het Nederlands het nationale verschil tussen Vlaanderen en Nederland, en het registerverschil tussen spontane conversaties, online discussies, populaire kranten, kwaliteitskranten en het Staatsblad. Voor het Engels onderzoeken we het nationale verschil tussen de Verenigde Staten en Groot-Brittanië, het registerverschil tussen verzonnen en informatieve teksten, en het temporele verschil tussen de jaren zestig en de jaren negentig. De socio-situationele categorieën waarvan we het taalgebruik willen vatten worden gevormd door de doorsnedes van deze dimensies. Zo hebben we voor het Nederlands bijvoorbeeld de categorieën *Vlaamse spontante conversatie* of *Nederlandse kwaliteitskranten*. Voor het Engels bestaan de categorieën uit de doorsnedes van de drie dimensies, bijvoorbeeld *Amerikaanse informatieve teksten uit de jaren negentig*.

4 Van de Wetstraat naar de dorpsstraat

We berekenen de taalgebaseerde en lexicale afstanden tussen de socio-situationele categorieën met behulp van de methode die hierboven al uitgelegd werd. De methode, kort herhaald, gaat na hoevaak hetzelfde woord gebruikt wordt om een bepaald begrip uit te drukken. Wordt in twee socio-situationele categorieën voor vele begrippen doorgaans hetzelfde woord gebruikt, dan is de taalgebaseerde afstand tussen die categorieën klein. Wordt voor vele begrippen echter een ander woord gekozen, dan is de taalgebaseerde afstand groot.

Vergelijken we de taalgebaseerde afstanden tussen alle mogelijk paren van socio-situationele categorieën, dan valt op dat zowel voor het Nederlands als voor het Engels de verschillen tussen de registers meer uitgesproken zijn dan de verschillen tussen de naties. Met andere woorden, de dorpsstraat ligt verder van de Wetstraat dan Antwerpen van Amsterdam, en New York ligt dichter bij Londen, dan een (Engelse) gebruiksaanwijzing bij het laatste Harry Potter boek. Bovendien stellen we vast dat het verschil in woordkeuze in het Engels tussen de jaren zestig en de jaren negentig het minst uitgesproken is, in vergelijking met het registerverschil en het nationale verschil. Als smakelijk detail konden we wel vaststellen dat het vooral scheldwoorden en uitroepen die veranderd zijn tussen de jaren zestig en de jaren negentie. Voor het Nederlands vertoont het krantenmateriaal uit Vlaanderen

een duidelijk verschil in het taalgebruik (of meer precies: de woorkeuze) van populaire kranten, zoals *Het Laatste Nieuws* en kwaliteitskranten, zoals *De Standaard*. Dit uitgesproken verschil tussen populaire en kwaliteitskranten kunnen we niet terugvinden in het Nederlandse krantenmateriaal.

Deze bevinding lijkt aan te leunen bij de mening van Geert Mak in *De Groene Amsterdammer* van 24 oktober 2012, waarin hij de teloorgang van het NRC Handelsblad gepassioneerd beschrijft. Uit mijn onderzoek blijkt echter enkel dat de kwaliteitskranten van Nederland doorgaans dezelfde woorden kiezen als hun populaire tegenhangers om begrippen uit te drukken. Of dit te wijten is aan een laksere en commerciëlere aanpak van de redactie, zoals Geert Mak suggereert, kan daaruit uiteraard niet afgeleid worden. Dit toont tegelijkertijd wel aan dat in mijn onderzoek de focus op het woordgebruik slechts het objectieve aspect van de menselijke vaardigheid om afstanden tussen taalvariëteiten te schatten belicht. De subjectieve gevoelens die Geert Mak tot het schrijven van zijn open brief bewogen — en mogelijkerwijze zijn objectieve afstandsinschatting beïnvloed hebben — vallen buiten het bereik van mijn onderzoek. Het is echter een boeiende uitdaging om ook de subjectieve kant van die vaardigheid te onderzoeken op een doorgedreven wetenschappelijke manier.

Tom Ruette is wetenschappelijke medewerker aan de Humboldt Universiteit van Berlijn na het behalen van een doctoraat in de taalkunde onder de supervisie van Prof. Dr. Dirk Speelman en Prof. Dr. Dirk Geeraerts (Universiteit van Leuven). Tijdens zijn doctoraat werkte hij aan de computationele modelering van woordbetekenis en kwantitatieve methodologie voor grootschalig taalkundig onderzoek. Aan de Humboldt Universiteit van Berlijn combineert hij zijn taalkundige en technische achtergrond bij de opbouw, ontsluiting en studie van het eerste digitale corpus van Oud-Duitse teksten.