De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak

Wilbert Heeringa en John Nerbonne[1]

Abstract

'Dialectometry', literally 'measurement of dialect', provides several methods for measuring dialect distances, each resulting in different outcomes. To sift among these, we formulate some requirements: first, representative data samples should be used as a basis; second, the comparison method should be sensitive enough to yield gradual distances, just as human perception; third, various linguistic levels should be involved. In this work we considered the levels of lexicon and pronunciation, using as data 360 Dutch sites from the *Reeks Nederlandse Dialectatlassen*. Lexical distances were measured using Goebl's 'gewichteter Identitätswert', a method in which the coincidence of rarely used words counts more heavily than those of more frequent ones. Pronunciation differences are measured using Levenshtein distance, a string edit distance measure. We also combine the measurements by taking the average of the normalized lexical and pronunciation distances. We compare these results to De Schutter's map, which is a simplification of Daan's map. Daan's map is based on the perception of dialect speakers, and De Schutter's map is considered by the author to reflect the 'communis opinio' of traditional dialectologists at the end of the 20th century. When comparing our results with De Schutter's map, we found the combined map closest. We found that both lexical and pronunciation differences have played a role in the perception of the dialect speakers who informed Daan's map.

1 Inleiding

In 1892 schreef Jellinghaus in het begin van *Die niederländischen Volksmundarten*: "Da in keinem Lande die Erforschung der Mundarten mit so viel Liebe und Ausführlichkeit betrieben ist, wie in den Niederlanden, so schien mir die Niederländische Volkssprache besonders geeignet für solche Betrachtung". In hetzelfde boek vinden we een kaart waarop het Nederlandse dialectgebied in gebieden is ingedeeld. Tot zover we weten is dit de oudste Nederlandse dialectkaart. Sindsdien zijn tal van andere kaarten verschenen, gebaseerd op verschillende methoden. Iedere kaart geeft steeds weer een ander beeld. Geerts (1975) schrijft dan ook: "De opvattingen over de indeling van het Nederlands taalgebied op basis van de onderscheiding van streektalen lopen nogal uiteen." En even verderop: "Omdat ze het niet helemaal eens zijn over de te volgen methode verschillen de dialektologen ook van mening over de indeling." (blz. 164).

Geïnspireerd door Jellinghaus' woorden in het begin van zijn boek willen we in dit artikel het onderzoek naar Nederlandse dialecten graag voortzetten. We willen ermee beginnen om temidden van de wirwar van verschillende indelingsmethoden te komen tot de formulering van een aantal eisen waaraan een goede indelingsmethode moet voldoen. Vervolgens presenteren we twee indelingsmethoden die in combinatie met elkaar ten dele aan die eisen voldoen en passen die toe op het Nederlandse dialectgebied. We focussen daarbij op lexicon en uitspraak. Het belang van lexicale variatie zien we bijvoorbeeld in de onderscheiding tussen het Germaanse taalgebied en het Romaanse taalgebied. Uitspraakvariatie lijkt vooral belangrijk te zijn op het gedetailleerdere niveau. Het Gronings kan bijvoorbeeld verdeeld worden in variëteiten waar *uit en thuis* uitgesproken wordt als *oet en thoes* en variëteiten waar men *uut en thuus* zegt.

In hoofdstuk 2 beginnen we met een kort overzicht van de belangrijkste dialectindelingskaarten van het Nederlandse dialectgebied. Aan de hand daarvan proberen we de eisen te vinden waaraan een goede indelingsmethode moet voldoen. In de rest van het artikel gebruiken we zelf een aantal indelingsmethoden. De gegevens waarop die methoden worden toegepast worden besproken in hoofdstuk 3. In hoofdstuk 4 bespreken we hoe we afstanden meten op basis van lexicale verschillen. Hoofdstuk 5 bespreekt een methode voor de meting van uitspraakverschillen. Gecombineerde metingen op basis van lexicon *en* uitspraak worden gepresenteerd in hoofdstuk 6. We eindigen in hoofdstuk 7 met het trekken van enkele conclusies.

2 Bestaande kaarten en methoden

In deze paragraaf geven we een overzicht van dialectindelingskaarten die voor het Nederlandse dialectgebied gemaakt zijn. Ons overzicht is onder andere gebaseerd op het overzicht van Goossens (1977), blz. 161-170. We noemen in ons overzicht alleen indelingsmethoden waarvan door middel van resultaten bewezen is dat ze toegepast kunnen worden op *heel* het Nederlands taalgebied. We eindigen het hoofdstuk met de formulering van eisen waaraan betrouwbare indelingsmethoden moeten voldoen.

2.1 Kaarten zonder vast criterium

In de loop van de tijd zijn verschillende kaarten gemaakt die een indeling van het Nederlandse dialectgebied weergeven. Wellicht de oudste kaart verscheen in 1892. De kaart was gemaakt door Jellinghaus. In het boek waarin de kaart is opgenomen, beschrijft Jellinghaus een groot aantal taalkundige eigenschappen van de verschillende dialecten, en deze zullen wellicht een rol gespeeld hebben bij de totstandkoming van de kaart.

In 1898 verscheen een kaart die werd gemaakt door Te Winkel. Deze kaart is gebaseerd op gegevens van 383 streken en plaatsen. De kaart verscheen voor het eerst in 1898 in een Duitse publicatie. In 1901 verscheen de kaart ook in een Nederlandse publicatie, waarbij de weergave van de kleuren nog meer in overeenstemming met de bedoeling van de auteur is gebracht (Te Winkel & Wieder, 1901). In 1913 verscheen een kaart van de hand van Van Ginneken, en in 1921 verscheen een kaart van Lecoutere. De kaarten van Te Winkel, Van Ginneken en Lecoutere lijken veel op elkaar. De kaart van Van Ginneken verschilt iets meer ten opzichte van de kaart van Te Winkel.

Het maken van dergelijke kaarten was in die tijd een hele prestatie. Wel is het jammer dat we niet precies weten waarop de dialectologen hun indeling baseerden. Voor de kaarten ontbreekt een verantwoording (Daan & Blok, 1969, blz. 17-19). Maar Goossens (1977) benadrukt wel dat met name Jellinghaus en Te Winkel op "de hoogte waren van een vrij groot aantal dialectgeografische tegenstellingen…" (blz. 163).

2.2 Isoglossenkaarten

In 1941 verscheen een isoglossenkaart van Weijnen.[2] Op deze kaart zijn 45 isoglossen getekend. In tegenstelling tot eerdere kaarten laat een isoglossenkaart beter zien hoe belangrijk grenzen zijn: een brede streng van isoglossen representeert een belangrijke grens, maar een grens van maar één enkele isoglosse is veel minder belangrijk. In 1958 verscheen opnieuw een kaart van Weijnen. De indeling op deze kaart is gebaseerd op 18 isofonen en isomorfen. Een andere isoglossenkaart verscheen in 1970 van de hand van Goossens. Goossens bekeek de Nederlandse dialecten vanuit Duits perspectief. Geerts (1975) heeft deze kaart ook opgenomen (blz. 165).

Het mooie van isoglossenkaarten is dat de resultaten verifieerbaar zijn. Maar het blijft wel onduidelijk hoe de ontwerper de keuze van de verschijnselen gemotiveerd heeft. Voor elk van de drie kaarten is de keuze van de isoglossen weer anders, met als gevolg dat elke kaart een andere indeling laat zien.

2.3 Pijltjeskaarten
2.3.1 Kaarten

In 1939 verstuurde het Dialectenbureau van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam een enquête naar ongeveer 1500 correspondenten. In die enquête werd onder andere gevraagd: "In welke plaats(en) in Uw omgeving spreekt men geheel of nagenoeg geheel hetzelfde dialecten als in de Uwe?" Op basis van de antwoorden op deze vraag construeerde Weijnen een dialectkaart voor Noord-Brabant. Met pijltjes verbond hij de plaatsen waarvan de dialecten volgens de sprekers "geheel of nagenoeg geheel" hetzelfde waren volgens de sprekers. "Zo

openbaren zich de stroken waar geen verbindingspijltjes doorlopen. En dit zijn dan duidelijk de in het dialectsprekersbewustzijn levende taalgrenzen" (Weijnen, 1958, blz. 138). Weijnens indelingskaart van Noord-Brabant verscheen in 1946 (zie ook Weijnen, 1958, blz. 140). Weijnen publiceerde ook een pijltjesmethodekaart voor Limburg (blz. 364). In 1955 publiceerde W.G. Rensink een indelingskaart voor heel Nederland. Deze kaart was eveneens gebaseerd op de antwoorden op de bovenvermelde vraag. De auteur benadrukte dat het ging om een voorlopig resultaat. De definitieve kaart werd gemaakt door J. Daan en verscheen in 1969 (Daan & Blok, 1969).

In Daan & Blok (1969, blz. 27-29) wordt erop gewezen dat grenzen die in het taalbewustzijn van dialectsprekers leven, niet altijd (zuiver) taalkundige grenzen zijn. Het kunnen geheel of gedeeltelijk ook geloofsgrenzen (protestant versus katholiek), sociale grenzen (industriëel versus agrarisch) of voormalige politieke grenzen zijn.

De enquête van het Dialectenbureau werd alleen verstuurd naar inwoners van Nederland. Daardoor bestrijkt de kaart van Rensink alleen Nederland. De kaart van Daan beslaat echter ook Noord-België en Frans Vlaanderen. De indeling van deze zuidelijke gebieden berust "op gegevens van taalkundigen die hier, in tegenstelling met Nederland, bijna allen dialektsprekers zijn. Dit geeft voldoende zekerheid dat ook in dit gebied de ervaring van de dialectspreker tot zijn recht is gekomen" (Daan & Blok, 1969, blz. 10).

In Figuur 1 is de kaart van Daan weergegeven zoals die geldt voor de 360 dialecten in de gegevensverzameling die beschreven wordt in hoofdstuk 3. Randgevallen, dat wil zeggen plaatsen die precies op of heel dicht bij een dialectgrens lagen, hebben we daarbij niet meegenomen. Verder ontbreken ten opzichte van de oorspronkelijke kaart van Daan op onze kaart twee gebieden, namelijk gebied 8 en gebied 16. Gebied 8 betreft een gebiedje in gebied 14 dat tegen de grens ligt tussen gebied 14 en gebied 1. Gebied 16 is een smalle overgangsstrook tussen gebied 15 en 17. De reden dat deze gebieden op onze kaart ontbreken, is het feit dat ze door geen enkele plaats in onze dialectenverzameling vertegenwoordigd zijn (zie paragraaf 3.2).

2.3.2 Kritiek

Hoewel Weijnen, Rensink en Daan alle drie de pijltjesmethode toepasten op hetzelfde materiaal, zijn de resultaten verschillend. Wanneer de pijltjes naar opgave van de sprekers op de kaart getekend zijn, kunnen grenzen blijkbaar toch nog op verschillende manieren getekend worden. Men kan hiervan een idee krijgen door de beide kaarten van Weijnen te bestuderen, omdat hierin zowel de pijltjes als de grenzen zijn weergegeven. Dat het pijltjesbeeld op verschillende manieren geïnterpreteerd kan worden, wordt ook bevestigd door het feit dat Noord-Brabant en Limburg op de beide kaarten van Weijnen in veel meer kleine gebieden verdeeld zijn dan op de kaarten van Rensink en Daan.

De pijltjesmethode is scherp bekritiseerd door Goossens (1977). De methode zou nooit consequent zijn toegepast. De dialectologen corrigeerden de gegevens van de correspondenten wanneer die gegevens niet helemaal voldeden of elkaar tegenspraken (blz. 167). Verderop zegt Goossens zelfs dat de dialectoloog de gegevens van de pijltjeskaarten verwerpt wanneer die iets nieuws opleveren. En dat betekent dat de ontwerpers van de pijltjeskaarten hun eigen methode niet serieus nemen. In Weijnen (1958, blz. 139) lezen we inderdaad dat fouten kunnen voorkomen, en dat het daarom belangrijk is om isoglossenkaarten te blijven tekenen.

Het resultaat van de pijltjesmethode is volgens Weijnen ook onvoldoende "om de grote gehelen te tonen". En dus kunnen we ook om die reden niet zonder isoglossenkaarten.

Om te toetsen in hoeverre Goossens' kritiek geldt voor de meest recente pijltjeskaart, de kaart van Daan, is het van belang de bladzijden 29 t/m 32 van Daan & Blok (1969) te lezen. We vinden dan slechts twee echte ingrepen in de resultaten van de pijltjesmethode. De eerste ingreep betreft Drenthe. De respons in dit gebied was zo laag, dat het vinden van grenzen niet mogelijk was. Er zijn dan twee mogelijkheden: herhaling van het experiment voor dat gebied òf terugvallen op één of meer isoglossen. De dialectologe koos voor de laatste optie. De tweede ingreep betreft een vermeende grens tussen Zuid-Holland en Utrecht. Uit de resultaten van de pijltjesmethode kwam deze grens niet echt duidelijk uit de verf. Toch was de dialectologe ervan overtuigd dat die grens er moest zijn. Ook toen vond ze de oplossing door terug te vallen op een isoglosse.

De eerste ingreep lijkt ons verantwoord. Bij de tweede ingreep moeten we toestemmen dat de oordelen van de sprekers niet volledig werden gerespecteerd. Meer van dergelijke ingrepen hebben we echter niet kunnen vinden, en we zijn daarom van oordeel dat de scherpe toon waarop Goossens zijn kritiek uit, niet terecht is. Bovendien is zijn kritiek dat de pijltjeskaarten niets nieuws opleverden niet juist. Wie bijvoorbeeld de kaart van Daan vergelijkt met oudere kaarten, komt daar snel genoeg achter.

We willen wel wijzen op een beperking in toepassingsmogelijkheden. De pijltjesmethode kan alleen uitsluitsel geven over de vraag of *aan elkaar grenzende* gebieden wel of niet gescheiden zijn door een grens. Daarbij is het belangrijk dat sprekers *elkaars dialecten kennen*, en dat zijn meestal alleen de dialecten van plaatsen die vlak in de buurt liggen. Daardoor is een vergelijking tussen bijvoorbeeld Nederlandse dialecten en Afrikaanse dialecten een stuk moeilijker. En vergelijking tussen Nederlandse dialecten en een onbekend Nedersaksisch dialect in Siberië is onmogelijk (voor een beschrijving van zo'n dialect zie Nieuweboer (1998)).

2.3.3 Communis opinio

Een vereenvoudigde versie van de kaart van Daan is de kaart van De Schutter (1994, blz. 440). De kaart is gegeven in Figuur 2. Op deze kaart vinden we zes groepen: Het Fries (dialecten 27 en 28 op de kaart van Daan), de noordoostelijke dialecten (dialecten 18 t/m 26), de centraal westelijke dialecten (dialecten 1 t/m 6), de zuidwestelijke dialecten (dialecten 7 en 9), de centraal zuidelijke dialecten (dialecten 8 en 10 t/m 15) en de zuidoostelijke dialecten (dialecten 16 en 17). De auteur kwam tot deze hoofdindeling door onder andere de kaart van Weijnen (1958) en een aantal kaarten van Goossens te raadplegen. Het doel van de kaart was een samenvatting te geven van de voornaamste indelingskaarten. De auteur beschouwt de kaart dan ook als een reflex van de 'communis opinio' onder traditionele dialectologen aan het einde van de twintigste eeuw.

2.4 Dialectometrische kaarten

De term 'dialectometrie' betekent letterlijk: de meting van het dialect. Deze term werd geïntroduceerd door Jean Séguy (Chambers & Trudgill, 1998). Jean Séguy was

directeur van de *Atlas linguistique de la Gascogne*. Samen met nog een aantal medewerkers publiceerde hij een zesdelige atlassenreeks. Séguy wilde de kaarten in deze atlassen op een objectievere manier analyseren dan mogelijk was met de traditionele analytische methoden. Séguy en zijn onderzoeksteam deden dit door voor elk tweetal naburige dialectplaatsen het aantal items te tellen waarvoor de naast elkaar gelegen dialectplaatsen verschillend waren. Dat aantal verschillen werd uitgedrukt in een percentage, en dat percentage representeerde vervolgens de taalkundige afstand tussen beide dialectplaatsen (Chambers & Trudgill, 1998, blz. 137-138).

Voor het Nederlandse dialectgebied werden dialectometrische methoden voor het eerst toegepast door de gebroeders Hoppenbrouwers in 1988. Spraakklanken kunnen worden beschreven als een reeks van onderscheidende kenmerken. Bijvoorbeeld klinkers kunnen voor, midden of achter in de mond worden uitgesproken (bijvoorbeeld gedefinieerd door het kenmerk *voor*). De tong kan daarbij een hoge, centrale of lage positie hebben (bijvoorbeeld gedefinieerd door het kenmerk *hoog*). De lippen kunnen gespreid of gerond zijn (bijvoorbeeld gedefinieerd door het kenmerk *rond*). Als we beschikken over een transcriptie van een bepaald dialect, dan kunnen we tellen hoeveel klanken in die transcriptie voorkomen die voor in de mond uitgesproken worden, of die met een hoge tongpositie uitgesproken worden, of die met geronde lippen uitgesproken worden. De featurefrequentiemethode telt voor elk van een reeks kenmerken (*features*) hoeveel klanken in de transcriptie aan dat kenmerk voldoen. Zijn er bijvoorbeeld veel hoge klanken, dan wordt de frequentie voor het kenmerk *hoog* heel hoog. Omgekeerd betekent een lage frequentie voor het kenmerk *hoog* dat er veel lage klanken zijn. De aantallen worden uitgedrukt in percentages. Het resultaat is een histogram, dat voor ieder kenmerk het aantal klanken representeert dat aan dat kenmerk voldoet. Maken we nu histogrammen voor meerdere dialecten, dan kunnen we op basis van die histogrammen taalkundige afstanden tussen de dialecten berekenen. De afstand tussen twee dialecten wordt in het eenvoudigste geval berekend als de som van de frequentieverschillen. Op basis van clusteranalyse kunnen we vervolgens dialectgroepen vinden.

In 2001 publiceerden de gebroeders Hoppenbrouwers hun boek *De indeling van de Nederlandse streektalen*. In dit boek laten zij de toepassing van hun methode zien op gegevens van 156 dialectplaatsen in het Nederlands taalgebied. Daarbij gebruikten zij transcripties uit de *Reeks Nederlandse dialectatlassen* (RND, Blancquaert & Pée, 1925-1982, zie ook hoofdstuk 3). Op blz. 60-61 vinden we een kaart met de hoofdindeling van de Nederlandse dialecten, en op blz. 66-67 vinden we een kaart met een nadere indeling. In vergelijking met de traditionele kaarten zijn er zowel verschillen als overeenkomsten. Opvallend is het onderscheid tussen kerndialecten en randdialecten. Randdialecten hebben het karakter van overgangszones.

Het mooie van dialectometrische methoden is dat tegenstrijdige informatie geen probleem is. Verschillende taalkundige verschijnselen in de transcripties kunnen verschillende indelingen suggereren. Met een dialectometrische methode wordt een soort gemiddelde indeling gevonden door alle verschijnselen in ogenschouw te nemen. Bij Séguys aanpak missen we wel een zekere gradualiteit. Twee items zijn gelijk of ongelijk. Bij de aanpak van de gebroeders Hoppenbrouwers missen we een bepaalde gevoeligheid, namelijk voor wat betreft de volgorde van klanken in een woord. Daardoor worden bijvoorbeeld [kəniˑn] (*konijn)* en [kniˑnə] niet onderscheiden.

2.5 Eisen waaraan een methode moet voldoen

In de vorige paragrafen hebben we een reeks methoden besproken waarmee dialectologen indelingskaarten van het Nederlandse dialectgebied hebben gemaakt. De verschillende methoden leiden tot verschillende indelingen, een probleem waar Geerts (1975, blz. 164) terecht op wijst. Het is daarom belangrijk om een verantwoorde keuze te kunnen maken uit de veelheid van indelingsmethoden. Voor dat doel is het belangrijk om scherp te stellen aan welke eisen een goede indelingsmethode moet voldoen. In paragraaf 2.5.1 geven we daarvan een overzicht. In paragraaf 2.5.2 geven we een korte beschrijving van de aanpak in dit artikel en houden die tegen het licht van de eisen uit paragraaf 2.5.1.

2.5.1 Te stellen eisen

Uit de bespreking van verschillende kaarten en methoden in de paragrafen 2.1 t/m 2.4 komen een achttal eisen naar voren.

a) Representativiteit
Het materiaal waarop een indelingsmethode wordt toegepast, moet representatief zijn voor de dialecten die onderzocht worden. Bij de pijltjesmethode is representativiteit gegarandeerd omdat de sprekers *zelf* beoordelen of dialecten verwant zijn of niet. Zij beschikken over alle kennis die voor zo'n beoordeling nodig is. Bij de isoglossenmethode zou een aselecte steekproef getrokken moeten worden uit de verzameling van alle mogelijke isoglossen. In de praktijk gebeurt dat niet, dialectologen hebben een voorkeur voor isoglossen die samenvallen. Bij dialectometrische methoden hangt representativiteit af van de keuze van items of zinnen. Bolognesi & Heeringa (2002) losten dit op door voor hun onderzoek naar Sardische dialecten de items (woorden) met de computer willekeurig te kiezen uit een groot corpus.

b) Talige perceptie
Wanneer we een indeling van dialecten willen bepalen, moeten we vooraf vaststellen wat we nu eigenlijk willen weergeven. *Ons* indelingscriterium is *talige perceptie*. De pijltjesmethode voldoet niet helemaal aan dit criterium. We zagen dat de gevonden grenzen soms ook geloofs-, sociale en politieke verschillen weerspiegelen. Voor de isoglossenmethode en de dialectometrische methoden weten we zeker dat taalkundige verschillen de basis voor de gevonden indeling vormen, maar inhoeverre zo'n indeling ook in overeenstemming is met de perceptie van de sprekers weten we niet meteen. Gooskens & Heeringa (2004) vonden een manier om daar achter te komen. Zij vergeleken de resultaten van een dialectometrische methode met de afstanden die door Noorse dialectsprekers zelf in een luisterexperiment gegeven werden. De dialectometrische methode die Gooskens & Heeringa in ogenschouw namen is sterk vergelijkbaar met de methode die we in paragraaf 5 beschrijven. De resultaten van deze methode bleken sterk te correleren met de resultaten van het luisterexperiment (r=0.67, p<0.001). Dat betekent dat de dialectometrische methode dialectafstanden berekent die een redelijke benadering vormen van de afstanden zoals die worden ervaren door de dialectsprekers zelf. Het is niet uitgesloten dat sjibbolets een rol hebben gespeeld in het luisterexperiment, maar die rol is hoogstens slechts beperkt geweest (Gooskens, 2005, blz. 43-44, 46).

c) Alle taalkundige niveaus

Het is belangrijk dat alle taalkundige niveaus in ogenschouw worden genomen bij de bepaling van de dialectindeling. Deze eis hangt uiteraard nauw samen met de twee hierbovengenoemde eisen: representativiteit en perceptie. De belangrijkste taalkundige niveaus zijn: lexicon, fonetiek, morfologie, syntax, prosodie. In het onderzoek van Séguy vinden we variabelen die variatie in diachrone fonetiek, fonologie, morfo-syntax, werkwoordsverbuigingen en lexicon representeren (Séguy, 1973). Elk van de variabelen wordt precies even zwaar gewogen. De vraag is of dat juist is. Door vergelijking van deze variabelen met de resultaten van een perceptieëxperiment zouden we daar meer zicht op kunnen krijgen.

d) Gradualiteit

Isoglosselijnen worden allemaal even dik getekend, ongeacht om welk taalkundig verschijnsel het gaat. Het komt niet uit de verf of bijvoorbeeld het onderscheid tussen [u] en [y] (bijvoorbeeld als klinker in *huis*: [hus] versus [hys]) taalkundig even belangrijk is als het verschil tussen [lopə] en [lopm̩]. Hetzelfde probleem vinden we bij de aanpak van Séguy: twee items zijn òf gelijk òf ongelijk. Bijvoorbeeld [hus] versus [hys] is even verschillend als [hus] versus [hœs]. Bij de featurefrequentiemethode van de gebroeders Hoppenbrouwers wordt juist heel goed rekening gehouden met deze gradualiteit. Omdat bij hun methode de fonetische kenmerken van klanken in de beschouwing betrokken worden, komt goed uit de verf dat het verschil tussen [hus] en [hys] kleiner is dan het verschil tussen [hus] en [hœs].

e) Sensitiviteit

Gradualiteit waarborgt nog niet altijd de gevoeligheid van een methode. We zagen dit met name bij de featurefrequentiemethode. De methode meet wel graduele verschillen (zie voorbeeld onder *d)*), maar is niet gevoelig voor de volgorde van segmenten in de uitspraak van een woord. Als voorbeeld noemden we in paragraaf 2.4 dat [kəniˑn] (*konijn)* en [kniˑnə] niet onderscheiden worden.

f) Vergelijkbaarheid

Zowel bij de isoglossenmethode als de pijltjesmethode kunnen alleen aan elkaar grenzende dialectgebieden met elkaar vergeleken worden. De abstractere dialectometrische methoden kennen deze beperking niet. Het is geen enkel probleem om bijvoorbeeld te bepalen of het Afrikaans sterker verwant is met het Zeeuws dan met het West-Vlaams. Daardoor is het bovendien mogelijk om binnen het Nederlandse dialectgebied elke dialectplaats met elke andere dialectplaats te vergelijken. Dit biedt ons de mogelijkheid om een eventuele verwantschap tussen dialecten in het noordoosten van Nederland en het noordwesten van België op het spoor te komen. De gebroeders Hoppenbrouwers vergeleken dan ook elk van de 156 dialecten met alle overige 155 dialecten, maar Séguy berekende alleen afstanden tussen dialecten die aan elkaar grenzen.

g) Eenduidige interpretatie

De procedure die leidt tot een indeling bestaat vaak uit een aantal stappen. Bij de isoglossenmethode worden eerst de isoglossen getekend, daarna volgt de interpretatie: waar vallen isoglossen samen, waar zijn grenzen, wat zijn de gebieden? Niet iedereen zal dezelfde grenzen en gebieden zien in het beeld dat de wirwar aan isoglossen te

zien geeft. Iets vergelijkbaars geldt voor de pijltjesmethode. Nadat de pijltjes naar opgave van de correspondenten op de kaart getekend zijn, vormen de witte stroken de grenzen. Dit lijkt eenvoudig, maar hierboven zagen we al dat het pijltjesbeeld soms verschillend geïnterpreteerd wordt. Ten slotte zien we iets vergelijkbaars ook voor de dialectometrische methoden. Na het meten van afstanden tussen dialecten voerden de gebroeders Hoppenbrouwers een clusteranalyse uit om dialectgebieden te vinden. Er bestaan echter minstens zeven verschillende clustermethoden, en die leiden tot verschillende indelingen. Toch is een verantwoorde keuze hier wel heel goed mogelijk (zie Heeringa, 2004, blz. 150-153 en Nerbonne & Siedle, 2005).

h) Verificatie
We noemden al dat Goossens (1977) benadrukt dat met name Jellinghaus en Te Winkel op "de hoogte waren van een vrij groot aantal dialectgeografische tegenstellingen…" (blz. 163). Toch ontbreekt een volledige verantwoording. Ook de indelingen op de kaarten van Van Ginneken en Lecoutere kunnen niet geverifieerd worden. Dit veranderde met de komst van de isoglossenmethode. De resultaten van deze methode zijn wel verifieerbaar. Hetzelfde geldt voor de pijltjeskaarten en de dialectometrische kaarten.

2.5.2 Onze aanpak

We hebben niet de pretentie in dit artikel een methodologie aan te bieden die aan alle eisen volledig voldoet. Wel willen we proberen met de huidige mogelijkheden zo dicht mogelijk in de buurt te komen. In onze aanpak worden dialecten met dialectometrische methoden met elkaar vergeleken. Daarbij kijken we naar twee taalkundige niveaus: het lexicale niveau en het uitspraakniveau. Voor de bepaling van lexicale afstanden gebruiken we de methode van Goebl (1984, blz. 85) (zie paragraaf 4.2). In navolging van Kessler (1995) berekenen we uitspraakafstanden tussen woordvarianten met de Levenshtein-afstand (zie paragraaf 5.2). In de laatste stap combineren we de indeling die we vonden op basis van de lexicale afstanden met de indeling die we vonden op basis van de uitspraakafstanden.

Om te bepalen in welke mate onze resultaten de talige perceptie weerspiegelen, willen we de resultaten in hoofdstuk 4, 5 en 6 vergelijken met de kaart van De Schutter.

3 Gegevensbron

De gegevensbron van ons onderzoek is de *Reeks Nederlandse Dialectatlassen* (RND). De 16 delen waaruit deze atlassenreeks bestaat, verschenen tussen 1925 en 1982 onder redactie van E. Blancquaert en W. Pée. In de atlassen vinden we transcripties van dialecten in Nederland, Noord-België, het uiterste noordwesten van Frankrijk en de Duitse graafschap Bentheim. Het atlasproject werd opgestart door Blancquaert. Toen Blancquaert overleed, nam Pée de leiding over. Onder zijn leiding is het project ook voltooid.

3.1 Woorden

In de RND wordt voor ieder dialect een vertaling gegeven van een reeks van 139 zinnen. Deze vertalingen zijn gegeven in fonetisch schrift. Blancquaert (1948) meldde dat de vragenlijst bedoeld was als een reeks van zinnen bestaande uit woorden die variatie in bepaalde klanken illustreren. Er was bijvoorbeeld voor gezorgd dat mogelijke ontwikkelingen in oud-Germaanse klinkers, diftongen en medeklinkers in de transcripties zouden kunnen worden teruggevonden. Morfologische en syntactische variatie zou eveneens door de zinnen worden gerepresenteerd (blz. 13).

Uit de 139 zinnen hebben we min of meer willekeurig 125 woorden gekozen. Het zou te veel tijd kosten om de complete teksten te digitaliseren. De woorden die we geselecteerd hebben representeren zo ongeveer alle klinkers (monoftongen en diftongen) en medeklinkers. De 125 woorden mogen beschouwd worden als een representatieve steekproef. Bij het digitaliseren werd erop gelet dat alle varianten van een woord dezelfde betekenis hadden. Lexicale variatie was wèl toegestaan.[3]

3.2 Variëteiten

De RND bevat transcripties van in totaal 1956 Nederlandse dialecten. Ook hier geldt dat het te veel tijd zou gaan kosten om de transcripties van alle dialecten te digitaliseren. Daarom maakten we een selectie van 360 dialecten. De 360 dialectplaatsen vormen over het algemeen een regelmatig net. Dit net is weergegeven in Figuur 3.

In Friesland kan men onderscheid maken tussen Friese dialecten en Friese mengdialecten. In het grootste deel van de provincie wordt Fries gesproken, maar op Ameland, in Het Bildt en in Stellingwerf wordt een mengdialect gesproken. In een groot aantal steden wordt eveneens een mengdialect gesproken. Deze steden vormen taaleilanden in het Friese dialectcontinuüm. We begonnen met het opzetten van een regelmatig net van dialecten die behoren tot het Friese dialectcontinuüm, en voegden daar vervolgens de taaleilanden aan toe. In Figuur 3 zijn de taaleilanden weergegeven als witte ruitjes.

In de RND worden voor Tjalleberd, Donkerbroek en Appelscha twee transcripties gegeven. Toen in Tjalleberd de RND-opnames gemaakt werden, spraken de meeste mensen daar Fries, maar een klein deel sprak het `Gietersk', een dialect dat geïntroduceerd werd door veenarbeiders uit Giethoorn en omstreken. Evenals Tjalleberd, ligt ook Donkerbroek in het Friese dialectcontinüum. Behalve Fries werd in deze plaats ook een Stellingwerfs dialect gesproken. Appelscha ligt in het Stellingwerfs dialectgebied. Net als in Donkerbroek werd in deze plaats zowel een Fries als een Stellingwerfs dialect gesproken. Het Friese dialect in deze plaats werd geïntroduceerd door veenarbeiders die afkomstig waren uit het Friese dialectgebied. Tjalleberd, Donkerbroek en Appelscha zijn in Figuur 3 aangegeven met grijze ruitjes.

3.3 Transcribenten

Hierboven noemden we al dat de RND uit 16 delen bestaat. Verder waren in totaal 16 veldwerkers betrokken bij het project. De verdeling van atlasdelen en veldwerkers is gegeven in Figuur 4.

De vragenlijsten die in de verschillende delen gebruikt zijn, verschillen soms een beetje. Ruwweg kunnen we onderscheid maken tussen een Vlaamse versie (deel 1 t/m 8) en een Nederlandse versie (deel 9 t/m 16). In de vragenlijst van deel 6 (West- en Frans-Vlaanderen) werden verschillende woorden vervangen door Franse equivalenten. Voor deel 15 (Friesland) werd voor de meeste plaatsen een Friese vragenlijst gebruikt. Onze lijst van 125 woorden bevat uitsluitend woorden die in alle vragenlijsten voorkomen. Een vragenlijst kan een enigszins sturende werking hebben op de informanten. Anderzijds sloot een gewijzigde vragenlijst beter aan bij het dialect van de informanten.[4]

Het was Blancquaerts doel dat de andere transcribenten op dezelfde manier zouden transcriberen als hij zelf deed, zodat de transcripties van de verschillende transcribenten consistent met elkaar zouden zijn. Goossens (1977) liet al zien dat dat niet helemaal gelukt is (blz. 71/72). Het is niet gemakkelijk om verschillen in notatiewijze uit de data weg te filteren. Het is moeilijk om vast te stellen of verschillen tussen twee transcripties dialectverschillen of notatieverschillen representeren. Toch lukte het voor een paar gevallen om verschillende notatievarianten terug te brengen tot één notatievariant. Het betrof de notatie van een klinker voor de /r/, de notatie van een syllabische nasaal na een andere nasaal, de notatie van een (syllabische) nasaal na een plofklank, en de notatie van een gepalataliseerde /t/.[5]

3.4 Lexemen

Omdat we lexicale afstanden willen meten en afstanden op basis van uitspraakvariatie, was een extra codering in onze data nodig. Voor elk van de 125 woorden moesten we bepalen welke lexemen voor dat woord onderscheiden kunnen worden. Bij de beoordeling of twee woorden lexicaal hetzelfde waren of niet zijn we steeds uitgegaan van het standpunt van een leek. Etymologie hebben we geen rol laten spelen. We illustreren dit aan de hand van enkele voorbeelden.

Het woord *tegen* wordt in het westen en zuiden van Nederland vaak uitgesproken als [teɣə], en in het noordwesten (Friesland) als [tsjĭ]. De vorm [tsjĭ] lijkt een variant te zijn op de eerste lettergreep van [teɣə]. De gemiddelde leek zal beide vormen echter beschouwen als verschillende woorden. Daarom verwerken we de twee vormen als lexicaal verschillende vormen.

Kijken we naar *zijn*, dan hebben we de vormen [bɪn̩] (Winschoten) en [bɛnə] (Schagen) tot een ander lexeem gerekend als de vormen [zɛɪn] (Haarlem) en [zɑn] (Geraardsbergen). Hoewel de vormen [bɪnt] (Ommen) en [zɪnt] (Kerkrade) veel op elkaar lijken, beschouwen we deze toch als verschillende lexemen.

Bij het coderen hebben we morfologisch verschillende vormen tot hetzelfde lexeem gerekend. Bijvoorbeeld [boˑmkə] (Roodeschool), [boˑmpin] (Roswinkel) en [bo̤ˑ^{u}mpjə] (Delft) zijn alle drie varianten van het lexeem *boompje*.

Het opdelen in lexemen was soms lastig. Bijvoorbeeld voor het woord *knuppel* vonden we de volgende vormen:

Haarlem:	[knʏpəl]
Gemert:	[klʏpəl]
Renesse:	[klɪpəl]
Geraardsbergen:	[klɪpər]
Lebbeke:	[klipər]

De vormen [knʏpəl] en [klipər] lijken op het eerste gezicht lexicaal verschillend te zijn. De tussenliggende vormen laten echter zien dat sprake is van een continuum. We hebben daarom al deze vormen gerekend tot hetzelfde lexeem.

4 Meting van lexicale verschillen

4.1 Verschillen tellen

Voor de bepaling van de lexicale afstand tussen twee dialecten is de aanpak van Jean Séguy goed bruikbaar (zie paragraaf 2.4). Séguy en zijn onderzoeksteam telden het aantal items waarvoor de dialectplaatsen verschillend waren. Dat aantal verschillen werd uitgedrukt als een percentage, en dat percentage representeerde vervolgens de taalkundige afstand tussen beide dialectplaatsen (Chambers & Trudgill, 1998, blz. 137-138). We illustreren dit aan de hand van een voorbeeld. Stel, we bepalen de lexicale afstand tussen het dialect van Middelstum en het dialect van Ommen op basis van zes woorden. De vergelijking ziet er dan als volgt uit:

	Middelstum	**Ommen**	
vriend	kɑmərʊˑt̬	kɑmərɔːt	0
schip	sxɪp	sxɪp	0
ver	v̥ɛːr	ʊit̬	1
zijn	bɪn̩	bɪnt	0
nog	nɔx	nɔx	0
duwen	støˑtn̩	drʏk̬ŋ	1

			2

Van de zes woordparen zijn er twee waarvan de beide woorden lexicaal verschillend zijn. De afstand wordt nu gelijk aan (2/6)*100=33%.

4.2 Weging van overeenkomsten

De aanpak die we beschreven in de vorige paragraaf werd verder verfijnd door Goebl (1984, blz. 85) door de frequenties van lexemen in rekening te brengen. We demonstreren zijn aanpak aan de hand van een voorbeeld. Voor alle 360 dialecten vinden we een vertaling voor *schip*. 353 dialecten hebben *schip*, 2 dialecten hebben *boot*, 1 dialect heeft *lager*, en 4 dialecten hebben *schuit*. Wanneer we dus twee willekeurige dialecten met elkaar vergelijken, is het niet uitzonderlijk als beide

dialecten het lexeem *schip* hebben. Het is echter wel bijzonder als beide dialecten het lexeem *schuit* hebben, en nòg uitzonderlijker is het als beide dialecten het lexeem *boot* hebben. In de aanpak van Goebl worden deze drie gevallen nu verschillend gewogen:

schip	vs.	schip	:	353/360	=	0.981
schuit	vs.	schuit	:	4/360	=	0.011
boot	vs.	boot	:	2/360	=	0.006

Wanneer twee lexemen verschillend zijn, is de afstand altijd 1. We zien dus dat schip vs. schip een afstand heeft die bijna gelijk is aan 1. Zou *schip* het enige lexeem zijn, dan zou schip vs. schip gelijk worden aan 360/360=1. Goebl duidt deze aanpak aan met *gewichteter Identitätswert*. Nerbonne & Kleiweg (2003) pasten in principe dezelfde methodologie toe bij de vergelijking van dialecten in het oosten van de Verenigde Staten.

Keren we terug naar ons voorbeeld waarin we Middelstum en Ommen met elkaar vergelijken op basis van zes woorden, dan wordt de berekening nu als volgt:

	Middelstum	**Ommen**			
vriend	kɑmərʊˑt̬	kɑmərɔːt	140/354	=	0.40
schip	sxɪp	sxɪp	353/360	=	0.98
ver	ɣ̊ɛːr	ʊit̬			1
zijn	bɪn̩	bɪnt	176/360	=	0.49
nog	nɔx	nɔx	354/355	=	1.00
duwen	støˑᵗn̩	drʏk̬ŋ			1

					4.87

De afstand wordt nu gelijk aan (4.87/6)*100=81.2%.

4.3 Ontbrekende woorden

We meldden al dat twee dialecten gewoonlijk op basis van 125 woordparen worden vergeleken. Het kan echter voorkomen dat voor één van beide dialecten, of zelfs voor beide dialecten geen vertaling beschikbaar is voor één of meer van de 125 woorden. Omdat het dan onmogelijk is om voor die woordparen een afstand te berekenen, worden ze genegeerd. Wanneer we voor *n* woordparen de woordafstanden (kunnen) berekenen, delen we de som van de woordafstanden door *n*.

Het bleek dat dialectparen gemiddeld op basis van 121 woorden konden worden vergeleken. Het minimale aantal woordparen was 103 en het maximale aantal was 125. De standaarddeviatie was 3.07.

De vraag die hier gesteld kan worden is of het aantal woorden dat we gebruiken om dialecten te vergelijken, voldoende is om betrouwbare resultaten te geven. Als maat voor betrouwbaarheid of consistentie gebruikten we Cronbachs α (zie voor meer details Heeringa (2004, blz. 170-173)). Binnen de sociale wetenschappen wordt in het algemeen een drempelwaarde van 0.70 geaccepteerd. Wij

kregen een waarde die net iets hoger is: 0.75. Ons aantal woorden is dus voldoende om betrouwbare resultaten te krijgen.

4.4 Meerdere varianten

Soms kan een woord in een dialect op meerdere manieren vertaald worden. Bijvoorbeeld in het dialect van Vreren vinden we in de RND voor *dochterje* drie mogelijkheden: [dʊxtər], [dʏxtərkə] en [mɛtskə]. In het dialect van Aalst vinden we twee mogelijkheden: [dʊxtərkən] en [mɑskən]. Voor Vreren hebben we dus drie vertalingen, en voor Aalst twee. We verwerken één en ander in een aantal stappen. In de eerste stap zorgen we ervoor dat we voor beide dialecten evenveel vertalingen krijgen. Daartoe dupliceren we de vertalingen van Vreren twee keer, en de vertalingen van Aalst drie keer. We krijgen dan zes vertalingen voor beide dialecten:

Vreren: [dʊxtər], [dʊxtər], [dʏxtərkə], [dʏxtərkə], [mɛtskə], [mɛtskə]
Aalst: [dʊxtərkən], [dʊxtərkən], [dʊxtərkən], [mɑskən], [mɑskən], [mɑskən]

In de volgende stap berekenen we de afstand van elk van de zes vertalingen van het dialect van Vreren ten opzichte van elk van de zes vertalingen van het dialect van Aalst. Op basis van deze afstanden vormen we woordparen. Omdat we zes vertalingen per dialect hebben, gebeurt dit in zes stappen. In de eerste stap kiezen we dàt paar dat de kleinste afstand heeft. In de tweede stap houden we dan nog vijf vertalingen per dialect over. Opnieuw kiezen we het woordpaar met de kleinste afstand. Dit herhalen we totdat we zes woordparen hebben. Ten slotte berekenen we het gemiddelde over de zes woordpaarafstanden. In ons voorbeeld krijgen we de volgende woordparen:

Vreren	**Aalst**	
[mɛtskə]	[mɑskən]	0.21
[mɛtskə]	[mɑskən]	0.21
[dʊxtər]	[dʊxtərkən]	0.45
[dʊxtər]	[dʊxtərkən]	0.45
[dʏxtərkə]	[dʊxtərkən]	0.45
[dʏxtərkə]	[mɑskən]	1

		2.77

De afstand wordt nu gelijk aan (2.77/6)*100=46.2%.

Een vergelijkbare maar enigszins verschillende aanpak voor de verwerking van meerdere varianten wordt beschreven door Nerbonne & Kleiweg (2003).

4.5 Resultaten
4.5.1 Clusteranalyse

In de vorige paragrafen bespraken we hoe we afstanden kunnen meten tussen dialecten op basis van lexicale verschillen. Omdat we 360 dialecten hebben (zie

paragraaf 3.2), meten we in totaal (360 × 359) / 2 = 64620 afstanden! Op basis van deze afstanden kunnen we tot een indeling in groepen komen door gebruik te maken van clusteranalyse. De groepen heten clusters. Clusters kunnen bestaan uit subclusters, subclusters uit subsubclusters, enz. Het resultaat is een hiërarchisch gestructureerde boom, waarbij de bladeren de dialecten zijn (Jain & Dubes, 1988). De takken in de bomen die door middel van clusteranalyse gemaakt zijn (de dendrogrammen), representeren de afstanden tussen dialecten en clusters. Er bestaan verschillende clustermethoden. Wij gebruikten UPGMA (Unweighted Pair Group Method using Arithmetic Averages). Het bleek dat de boom die met déze methode gemaakt werd, de originele afstanden – de afstanden tussen de 360 dialecten op basis waarvan de clusteranalyse werd uitgevoerd – het meest nauwkeurig weerspiegelt (zie Heeringa 2004, blz. 150-153). De afstanden tussen dialecten zoals die door onze UPGMA-boom gesuggereerd worden, verklaren voor 71% de variantie in de oorspronkelijke afstanden.

In Figuur 5 zien we het resultaat op basis van 360 dialecten. Een compleet dendrogram waarin alle bladeren, namelijk labels van de 360 dialecten, zijn weergeven, is niet erg overzichtelijk. Bovendien willen we de details die in een dergelijke boomstructuur te zien zijn, in dit artikel niet bespreken. Figuur 5 bevat daarom een vereenvoudigd dendrogram waarin alleen de 11 meest significante groepen te zien zijn. De labels in dit dendrogram zijn (met uitzondering van Zoutkamp) in feite dus niet de bladeren (de dialecten), maar representeren subgroepen met daarbinnen een verdere verdeling in dialecten. De 11 groepen worden geografisch weergegeven in Figuur 6. Het ruitje in het noordwesten representeert het dialect van Leeuwarden dat één groep vormt met het dialect van het Westerkwartier. We hadden ook kunnen kiezen voor een verdeling in 12 of meer groepen. We krijgen dan echter een meer versnipperd beeld. Het zou interessant zijn om over de taalkundige achtergronden daarvan te schrijven, maar we willen dat, om het artikel overzichtelijk te houden, nu buiten beschouwing laten.

4.5.2 Vergelijking

Wanneer we de kaart in Figuur 6 vergelijken met de dialectkaart van De Schutter, dan zien we zowel overeenkomsten als verschillen. Gebied 1 op onze kaart correspondeert ongeveer met de Friese groep. Gebied 2 vinden we niet op de kaart van De Schutter, maar wel op de gedetailleerde kaart van Daan (Kollumerlands, groep 25 in Figuur 1). Op de kaart van Daan vinden we rond de zuidoostgrens van Friesland een Nedersaksisch overgangsgebied: het Stellingwerfs (gebied 22 in Figuur 1). Volgens onze kaart hoort dit overgangsgebied bij het Fries, volgens de kaart van De Schutter hoort het bij de noordoostelijke dialecten. De gebieden 4 t/m 8 op onze kaart komen overeen met de noordoostelijke groep. De grens tussen de gebieden 4 en 6 op onze kaart vinden we op de kaart van Daan als de grens tussen de gebieden 24 en 26 (noordelijk) en gebied 23 (zuidelijk). Gebied 5 komt overeen met het noordelijke deel van gebied 20 op de kaart van Daan (Twents-Graafschaps), en groep 8 met gebied 18 op de kaart van Daan (Veluws).

Ons gebied 9 komt ongeveer overeen met de centraal westelijke groep, en ons gebied 10 met de zuidwestelijke groep. We zien dat Zeeland en het Zuid-Hollandse eiland Goeree-Overflakkee op de kaart van De Schutter behoren tot de zuidwestelijke dialecten, terwijl dit gebied op onze kaart behoort bij de centraal westelijke dialecten. Het grote gebied 11 op onze kaart omvat de centraal zuidelijke en de zuidoostelijke

groep. Het Limburgs wordt op onze lexicale kaart niet apart onderscheiden. Op de kaart van de Schutter is dat wel het geval: de zuidoostelijke dialectgroep.

Waar bovengenoemde verschillen niet te maken hebben met verschil in gedetailleerheid, kunnen ze eenvoudig verklaard worden uit het feit dat onze kaart zuiver lexicaal bepaald is, terwijl aan de indeling van de kaart van De Schutter ook andere taalkundige factoren ten grondslag kunnen liggen. We willen daarom in paragraaf 5.7.2 de kaart van De Schutter ook vergelijken met een kaart die we construeerden op basis van uitspraakverschillen.

We merken op dat we verschillen tussen onze kaart en de kaart van De Schutter niet of nauwelijks konden verklaren uit het feit dat de dialecttranscripties door verschillende veldwerkers gemaakt zijn (zie Figuur 4). Verschillen tussen transcribenten wreken zich eerder op het uitspraakniveau (zie paragraaf 5.7). Er zijn soms wel kleine verschillen tussen de gebruikte vragenlijsten (zie paragraaf 3.3), maar omdat deze daardoor juist vrij goed aansluiten bij de dialectgebieden waarvoor ze gebruikt werden, is de sturende werking daarvan blijkbaar heel beperkt gebleven.

5 Meting van uitspraakverschillen

5.1 Inleiding

In 1995 gebruikte Kessler de Levenshtein-afstand als instrument voor het meten van taalkundige afstanden tussen Ierse dialecten. De Levenshtein-afstand is gelijk aan de minimale kosten die nodig zijn om de ene reeks te veranderen in de andere. In het eenvoudigste geval zijn drie operaties mogelijk: een element toevoegen, een element vervangen door een ander element, of een element verwijderen. In het geval van Kessler worden woorduitspraken vergeleken. Woorduitspraken worden gerepresenteerd in de vorm van fonetische transcripties. Bij de bepaling van de Levenshtein-afstand tussen twee fonetische transcripties kunnen fonetische segmenten worden toegevoegd, vervangen of verwijderd.

Kesslers aanpak bleek succesvol en werd ook toegepast op Nederlandse dialecten (Nerbonne et al., 1996, Heeringa, 2004, pp. 213-278), Sardische dialecten (Bolognesi & Heeringa, 2002), Noorse dialecten (Gooskens & Heeringa, 2004) en Duitse dialecten (Nerbonne & Siedle, 2005). In ons onderzoek gebruiken we eveneens de Levenshtein-afstand. We meten daarbij uitsluitend uitspraakverschillen. Uitspraakverschillen omvatten zowel fonetische als morfologische verschillen.

Er zijn verschillende varianten van de Levenshtein-afstand mogelijk (zie Heeringa (2004), blz. 165-167). In paragraaf 2.5.1 meldden we echter dat we *talige perceptie* als ons indelingscriterium kozen. In Heeringa (2004, blz. 178-195) wordt voor een reeks dialectvergelijkingsmethoden onderzocht in hoeverre ze de perceptie van de dialectsprekers benaderen. We gebruiken in dit artikel daarom dìe variant die volgens dit onderzoek de talige perceptie het beste benadert. We geven hieronder een korte beschrijving van de methodologie. Voor een uitgebreidere uitleg verwijzen we naar Heeringa (2004, met name blz. 121-135). In paragraaf 5.2. introduceren we eerst een simpele versie van de Levenshtein-afstand waarbij klanken òf gelijk òf ongelijk aan elkaar zijn. In paragraaf 5.3 verfijnen we de aanpak door gebruik te maken van graduele klankafstanden.

5.2 De Levenshtein-afstand

Zoals we hierboven al meldden, introduceren we in deze paragraaf een een simpele versie van de Levenshtein-afstand om het principe beter te kunnen uitleggen. We gaan er daarbij gemakshalve van uit dat klanken òf hetzelfde òf verschillend zijn. We schreven hierboven al dat de Levenshtein-afstand gelijk is aan het minimale aantal operaties dat nodig is om de ene reeks (van fonetische segmenten) te veranderen in de andere reeks. We illustreren dit aan de hand van een voorbeeld. In het dialect van Amsterdam wordt *konijn* uitgesproken als [kənɛːn]. In het dialect van Zwollekerspel wordt hetzelfde woord uitgesproken als [kniːnə]. De ene uitspraak zou je kunnen veranderen in de andere op de volgende manier:

kənɛːn	vervang ɛː door iː	1
kəniːn	verwijder ə	1
kniːn	voegtoe ə	1
kniːnə		
---	---	---
		3

In feite kan men op heel veel verschillende manieren de ene uitspraak veranderen in de andere. De kracht van het Levenshtein-algoritme is echter dat deze de operaties zodanig kiest dat de totale kosten zo klein mogelijk blijven.

Omdat woorden taalkundige eenheden zijn, willen we normaliseren over de woordlengtes. Twee woorden die met elkaar vergeleken worden hoeven niet altijd dezelfde lengte te hebben. Zouden we bijvoorbeeld [kənɛːn] met [kniːn] vergelijken, dan bestaat het eerste woord uit vijf segmenten, en het tweede uit vier segmenten. Het is daarom het handigste om de gevonden Levenshtein-afstand te delen door de lengte van de oplijning (zie Heeringa 2004, blz 130-133 voor een gedetailleerde uitleg). Voor het voorbeeld uit het begin van deze paragraaf ([kənɛːn] versus [kniːnə]) ziet de oplijning er als volgt uit:

k	ə	n	ɛː	n	
k		n	iː	n	ə
---	---	---	---	---	---
0	1	0	1	0	1

Klankparen met een 0 zijn *matches*, dat wil zeggen klankparen waarvoor geldt dat beide klanken gelijk zijn. De overige klankparen betreffen toevoegingen, vervangingen of verwijderingen. De Levenshtein-afstand (1+1+1=3) delen we nu door de lengte van de oplijning (6). Dit geeft een woordafstand van 3/6 = 0.5, oftewel 50%.

5.3 Klankafstanden
5.3.1 Gradueel

In het voorbeeld dat we gebruikten in de vorige paragraaf, gingen we uit van een aanpak waarbij de operaties ‘toevoegen’, ‘vervangen’ en ‘verwijderen’ altijd hetzelfde

gewicht hebben, namelijk 1. Wanneer de klanken van een klankpaar gelijk zijn, hebben we een *match* met als gewicht de waarde 0. Deze aanpak kan verfijnd worden door rekening te houden met de mate van verwantschap tussen klanken. Bijvoorbeeld de [ɪ] en de [e] lijken meer op elkaar dan de [ɪ] en de [ɔ]. In dit artikel gebruiken we daarom graduele klankafstanden.

Voor het bepalen van graduele klankafstanden hebben we gebruik gemaakt van de cassette *The Sounds of the International Phonetic Alphabet* die uitgegeven werd in 1995. Op deze cassette worden alle klanken van het *International Phonetic Alphabet* (IPA) uitgesproken door twee sprekers: John Wells en Jill House. Medeklinkers worden steeds gevolgd door een [a] (bijv. [pa]) en soms bovendien voorafgegaan door een [a] (bijv. [apa]). We knipten de medeklinkers dan steeds uit hun context ([p]). Op de cassette worden niet alle klanken op dezelfde toonhoogte uitgesproken. Om de klanken zo vergelijkbaar mogelijk te maken, hebben we ze daarom gemonotoniseerd. Om afstanden tussen klanken te kunnen berekenen is het noodzakelijk dat de klanken een akoestische representatie hebben. Wij maakten gebruik van de spectrogramrepresentatie. Een spectrogram is een tweedimensionale representatie, waarbij de x-as de tijd, de y-as de frequentie en de z-as de geluidsintensiteit representeert. Voor een reeks tijdstippen wordt voor een reeks frequenties de intensiteit gegeven. Er zijn verschillende spectrogramrepresentaties mogelijk. Omdat we vooral de perceptie van de sprekers willen benaderen, kozen we voor een meer perceptief geöriënteerd model: het Barkfilter. Een gedetailleerde beschrijving van het Barkfilter is te vinden in Heeringa (2004, blz. 87–93).

Voor ‘vervangingen’ gebruikten we de graduele klankafstanden. Voor ‘toevoegingen’ en ‘verwijderingen’ gebruikten we echter ook graduele afstanden. Wanneer een klank wordt toegevoegd of verwijderd, gebruikten we als gewicht de afstand van die klank ten opzichte van ‘stilte’. De spectrogramrepresentatie van ‘stilte’ is eenvoudig: voor ieder tijdstip zijn de intensiteiten van alle frequenties gelijk aan 0. We vonden dat de [a] de grootste afstand heeft ten opzichte van stilte en de [ʔ] (glottisslag) de kleinste afstand.

5.3.2 Logaritmisch

Perceptief gezien spelen kleine klankverschillen een relatief grote rol ten opzichte van grote klankverschillen. Bijvoorbeeld *gaan* wordt in het grootste deel van de provincie Groningen uitgesproken als [ɣʊˑn], maar in de rest van het Nedersaksisch taalgebied (Drenthe, Overijssel, Noord-Gelderland) als [ɣɔˑn]. Hoewel dit eigenlijk niet meer dan een nuanceverschil is, is het wel bepalend. We gebruiken daarom logaritmische klankafstanden. Daardoor worden kleine klankafstanden relatief zwaarder gewogen dan grote afstanden. Omdat de logaritme van 0 niet gedefinieerd is, en de logaritme van 1 gelijk is aan 0, worden de klankafstanden, zoals voorgesteld in paragraaf 5.3.1, eerst verhoogd met 1. Daarna wordt de logaritme berekend. Omdat we klankafstanden bovendien willen uitdrukken als percentages, wordt elke logaritmische afstand gedeeld door de logaritme van de grootste klankafstand plus 1, en vermenigvuldigd met 100. De maximale afstand is de afstand tussen de [a] en ‘stilte’. In formulevorm wordt dit:

$$(ln(\text{afstand}+1) \,/\, ln(\text{maximale afstand}+1)) \times 100$$

5.3.3 Toegestane correspondenties

Wanneer we rekening willen houden met de syllabificatie in woorden, is het van belang om niet alle correspondenties in een oplijning toe te staan. Onze versie van het Levenshtein-algoritme houdt rekening met syllabificatie door de volgende regels in acht te nemen:

1. de [j] en de [w] mogen met een klinker (of andere medeklinker) corresponderen, bijvoorbeeld *hooi*: [hoː**j**] versus [hoː**i**];
2. De [i] en de [u] mogen met een medeklinker (of andere klinker) corresponderen, bijvoorbeeld *sneeuw*: [snəː**u**] versus [snəː**w**];
3. De [ə] mag met een sonorante medeklinker (of andere klinker) corresponderen, bijvoorbeeld *vuur*: [vyː**ə**] versus [vyː**r**];
4. alle andere klinkers mogen alleen corresponderen met klinkers, bijvoorbeeld *kaas* [kaːs] (standaard Nederlands) versus [keːs] (Gronings);
5. alle andere medeklinkers mogen alleen corresponderen met medeklinkers, bijvoorbeeld *schip* [s**x**ɪp] (standaard Nederlands) versus [s**k**ɪp] (Fries).

5.4 Aggregatie

In de vorige paragraaf bespraken we hoe de afstand tussen twee woorduitspraken wordt berekend als de Levenshtein-afstand. De afstand tussen twee dialecten wordt echter niet berekend op basis van één enkel woordpaar, maar op basis van een reeks woordparen, in ons geval 125 woordparen (zie paragraaf 3.1). We illusteren dit aan de hand van een voorbeeld. In dat voorbeeld berekenen we de afstand tussen Middelstum en Ommen op basis van zes woorden. Om het voorbeeld eenvoudig te houden gebruiken we hier geen graduele klankafstanden, maar de ruwere aanpak waarbij de drie gewichten (toevoegen, vervangen, verwijderen) altijd de waarde 1 hebben. Ook laten we diacritsche tekens buiten beschouwing. De berekening ziet er dan als volgt uit:

	Middelstum	**Ommen**			
schip	sxɪp	sxɪp	0/4	=	0
pet	pɛt	pɛtə	1/4	=	0.25
roepen	rɔupm	ərupm	2/6	=	0.33
springen	sprɪŋ	sprɪŋkt	2/7	=	0.29
kelder	kɛlər	kɛldər	1/6	=	0.17
huis	hus	hys	1/3	=	0.33

					1.37

In de vierde kolom geeft de teller steeds de Levenshtein-afstand, en de noemer de lengte van de langste oplijning. De laatste kolom geeft de genormaliseerde

Levenshtein-afstanden. Deze genormaliseerde Levenshtein-afstanden tellen we op. Anders gezegd: we aggregeren ze. De afstand tussen Middelstum en Ommen wordt nu gelijk aan (1.37/6)*100=22.8%.

5.5 Ontbrekende woorden

Twee dialecten worden op basis van maximaal 125 woordparen vergeleken. Voor veel dialectparen zal het aantal woordparen waarvoor de afstand kan worden berekend, iets lager zijn. De oorzaak kan zijn dat sommige woordparen niet compleet zijn. Voor één van beide dialecten, of zelfs voor beide dialecten is geen vertaling beschikbaar. Een andere oorzaak kan zijn dat de beide woorden van het woordpaar lexicaal verschillend zijn. Omdat we alleen uitspraakverschillen willen meten, willen we geen afstanden meten tussen woorden die lexicaal verschillend zijn. Zo'n woordpaar valt dus uit.

Incomplete woordparen of woordparen met lexicaal verschillende woorden worden genegeerd. Wanneer we voor *n* woordparen de woordafstanden (kunnen) berekenen, delen we de som van de woordafstanden door *n*.

Het bleek dat dialectparen gemiddeld op basis van 110 woorden konden worden vergeleken. Het minimale aantal woordparen was 81 en het maximale aantal was 125. De standaarddeviatie was 6.51.

Net als in paragraaf 4.3 moeten we ook hier de vraag stellen of het aantal woorden dat we gebruiken om dialecten te vergelijken, voldoende is om betrouwbare resultaten te geven. We berekenen opnieuw Cronbachs α. Deze is gelijk aan 0.97 en ligt dus ruim boven de drempelwaarde van 0.70. Het aantal woorden is dus ruim voldoende om betrouwbare resultaten te krijgen.

5.6 Meerdere varianten

In paragraaf 4.4 bespraken we dat voor een woord in een dialect soms meerdere vertalingen gegeven worden. We illustreerden dit aan de hand van het woord *dochterje*. In de RND zijn voor het dialect van Vreren drie vertalingen gegeven ([dʊxtər], [dʏxtərkə] en [mɛtskə]), en voor het dialect van Aalst twee vertalingen ([dʊxtərkən] en [mɑskən]). We zorgden er dan voor dat we voor beide dialecten evenveel vertalingen kregen. Voor beide dialecten kregen we zes vertalingen en we berekenden de afstand van elk van de zes vertalingen van het dialect van Vreren ten opzichte van elk van de zes vertalingen van het dialect van Aalst. Daarna gingen we woordparen vormen. In de eerste stap kozen we dàt paar dat de kleinste afstand heeft. In de tweede stap hielden we nog vijf vertalingen per dialect over. Opnieuw kozen we het woordpaar met de kleinste afstand. Dit herhaalden we totdat we zes woordparen kregen. Nu we uitspraakafstanden berekenen in plaats van lexicale afstanden, gebruiken we vrijwel dezelfde procedure. Op basis van uitspraakafstanden krijgen we dan de volgende woordparen:

Vreren	**Aalst**	
[dʏxtərkə]	[dʊxtərkən]	0.22
[dʏxtərkə]	[dʊxtərkən]	0.22
[dʊxtər]	[dʊxtərkən]	0.33
[mɛtskə]	[mɑskən]	0.43
[mɛtskə]	[mɑskən]	0.43
[dʊxtər]	[mɑskən]	x

		1.63

Om het voorbeeld eenvoudig te houden, hebben we ook hier de ruwe aanpak gebruikt waarbij alle Levenshtein-operaties de waarde 1 krijgen. En ook nu zijn de diacritische tekens buiten beschouwing gelaten. Bij het vormen van paren worden paren met lexicale verschillen zo veel mogelijk voorkomen. In ons voorbeeld is het echter onvermijdelijk dat er een paar overblijft waarvan de woorden lexicaal verschillend zijn, namelijk het laatste paar. Woordparen met lexicale verschillen worden buiten beschouwing gelaten. De afstand in ons voorbeeld wordt nu gelijk aan (1.63/5)*100=32.6%.

5.7 Resultaten

5.7.1 Clusteranalyse

In de vorige paragrafen bespraken we hoe we afstanden kunnen meten tussen dialecten op basis van uitspraakverschillen. Net als in paragraaf 4.5.1 passen we ook hier clusteranalyse toe om te komen tot een indeling in groepen. We gebruiken weer UPGMA (Unweighted Pair Group Method using Arithmetic Averages). Evenals bij toepassing van clusteranalyse op lexicale afstanden, blijkt ook nu, bij toepassing op uitspraakafstanden, UPGMA een boom te geven die de originele afstanden het meest nauwkeurig weerspiegelt. De afstanden tussen dialecten zoals die door onze UPGMA-boom gesuggereerd worden, verklaren voor 70% de variantie in de oorspronkelijke afstanden.

In Figuur 7 zien we het resultaat op basis van 360 dialecten. Net als in Figuur 5 toont het dendrogram in Figuur 7 alleen de 11 meest significante groepen. De labels representeren dus geen bladeren (dialecten), maar subgroepen. Een uitzondering wordt hier gevormd door Urk. De 11 groepen worden geografisch weergegeven in Figuur 8. De ruitjes in het noordwesten representeren taaleilanden.

5.7.2 Vergelijking

We willen nu de kaart in Figuur 8 vergelijken met de dialectkaart van De Schutter (Figuur 2). We zien dat de groepen 1 en 2 op onze kaart overeenkomen met het Friese gebied op de kaart van De Schutter. Groep 2 omvat de Friese mengdialecten. Deze worden op de kaart van De Schutter niet apart onderscheiden, echter wel op de kaart van Daan (groep 22 en groep 28). Onze groep 3 komt overeen met de noordoostelijke groep op de kaart van De Schutter. Dat de zuidgrens van de noordoostelijke dialecten

op onze kaart noordelijker ligt dan op de kaart van De Schutter zou goed verklaard kunnen worden uit de nototatieverschillen tusen transcribenten (vergelijk Figuur 4).

De grens tussen de centraal westelijke groep en de centraal zuidelijke groep is op onze uitspraakkaart niet te vinden. Blijkbaar is dit een zuiver lexicale grens, want deze grens is wel te vinden in op de lexicale kaart in Figuur 6. De grens tussen onze groep 5 en groep 7 daarentegen lijkt vooral een grens te zijn die uitspraakverschillen representeert. Deze grens is te vinden op de kaart van Daan (Figuur 1), maar niet op de kaart van De Schutter, en ook niet op de lexicale kaart in Figuur 6.

Een overeenkomst met de kaart van De Schutter is dat Zeeland nu één groep vormt met de zuidwestelijke dialecten (vergelijk Figuur 6 en paragraaf 4.5.2). Het Oostvlaams (onze groep 8) is op de kaart van de Schutter niet als aparte groep gemarkeerd, maar op de kaart van Daan vinden we deze groep als gebied 11.

Op de kaart van De Schutter vormen de zuidoostelijke dialecten één groot gebied, evenals op de kaart van Daan (gebied 17). Op onze kaart is dit gebied verdeeld in drie gebieden (gebied 4, 10 en 11). Van dit zuidoostelijk gebied, het Limburgs, is bekend dat het dialectologisch gezien heel heterogeen is (Hoppenbrouwers & Hoppenbrouwers, 2001, blz. 187). In een gebied met veel variatie leven dialectgrenzen in het bewustzijn van de sprekers wellicht minder sterk, zodat de grenzen op onze kaart niet terug te vinden zijn op de kaarten van De Schutter en van Daan.

Tenslotte willen we de aandacht vestigen op twee kleine gebiedjes, die beide behoren tot groep 10 op onze kaart. Het gebiedje in het westen is Steenbergen, en dat in het oosten is Helmond. Zowel in de dialecttranscriptie van Steenbergen als van Helmond wordt de /r/ voornamelijk getranscribeerd als de uvulaire [ʀ]. In de omgeving rondom deze plaatsen wordt steeds de alveolaire [r] gebruikt. Om meer zicht te krijgen op de rol die de uitspraak van de /r/ hier speelt, hebben we een experiment gedaan waarbij we alle [ʀ]'s in de transcripties van deze plaatsen vervangen hebben door [r]'s. Het blijkt dat beide dialecten dan niet langer taaleilanden vormen, maar het meest verwant zijn aan geografische dichtbijgelegen dialecten. Het gebruik van de [ʀ] in Steenbergen kan mogelijk verklaard worden door migratie. Het kan ook zijn dat het gebruik van de [ʀ] niet algemeen is in het dialect van Steenbergen, maar een toevallige eigenschap is van één van de informanten. Voor Helmond is het gebruik van de [ʀ] eenvoudiger. Deze plaats ligt niet ver van het Limburgse dialectgebied, waar veelvuldig de [ʀ] wordt gebruikt.

6 Een indeling op basis van lexicon èn uitspraak

6.1 Het combineren van beide niveaus

In hoofdstuk 4 bespraken we de meting van lexicale afstanden tussen 360 dialecten in het Nederlandse dialectgebied, en in hoofdstuk 5 lieten we zien hoe uitspraakafstanden tussen dezelfde dialecten gemeten kunnen worden. Zowel op basis van lexicale afstanden als op basis van uitspraakafstanden voerden we een clusteranalyse uit. De respectievelijke resultaten projecteerden we op kaarten. Hoewel de lexicale kaart ten dele een ander beeld laat zien dan de uitspraakkaart, blijkt er wel een significante correlatie te bestaan tussen de lexicale afstanden en de uitspraakafstanden: r=0.63 met p<0.001.

In paragraaf 2.5.1 punt c zeiden we dat het belangrijk is dat alle taalkundige niveaus in ogenschouw genomen worden bij de bepaling van een dialectindeling. Dat betekent dat we een kaart willen hebben die niet *alleen* gebaseerd is op lexicale variatie of *alleen* op uitspraakvariatie, maar op de *combinatie* van lexicale variatie *en* uitspraakvariatie. Séguy betrok in zijn onderzoek vijf variabelen: diachrone fonetische afstanden, fonologische afstanden, morfo-syntactische afstanden, afstanden op basis van variatie in werkwoordsverbuigingen en lexicale afstanden (Séguy, 1973).[6] Voor alle variabelen werden de afstanden uitgedrukt in percentages. De afstand tussen twee dialecten was gelijk aan het percentage itemparen waarvan de items niet aan elkaar gelijk waren. Alle variabelen hadden dus dezelfde schaal, en Séguy nam dan ook simpelweg het gemiddelde over alle vijf variabelen om tot een gecombineerd resultaat te komen (Chambers & Trudgill, 1998). De aanpak van Séguy kunnen we niet zomaar overnemen. Bij Séguy waren de schalen van de verschillende variabelen vergelijkbaar doordat voor elke variabele hetzelfde meetprincipe werd toegepast: gewoon tellen hoeveel itemparen verschillend zijn, en dat uitdrukken als een percentage. Hoewel ook in ons onderzoek zowel de lexicale afstanden als de uitspraakafstanden uitgedrukt worden als percentages (zie de paragrafen 4.2 en 5.4), zijn de onderliggende meetprincipes nogal verschillend. Lexicale afstanden worden gemeten met de *gewichteter Identitätswert*-methode en de uitspraakafstanden met de *Levenshtein*-afstand. Als we tussen de lexicale afstanden en de uitspraakafstanden gewoon een gemiddelde berekenen, dan correleert het resultaat sterker met de uitspraakafstanden dan met de lexicale afstanden (r= 0.992 vs. r=0.723), vermoedelijk omdat verschillende schalen gebruikt worden. Als we daarentegen de lexicale afstanden c.q. de uitspraakafstanden eerst normaliseren en dan middelen, krijgen we een resultaat dat even sterk correleert met de lexicale afstanden als met de uitspraakafstanden (r=0.902). We hebben de lexicale afstanden en de uitspraakafstanden genormaliseerd door ze om te zetten naar z-waarden. Omzetten naar z-waarden wil zeggen dat de variabelen omgezet worden naar reeksen die *beide* een gemiddelde hebben van 0 en een standaarddeviatie van 1.

Het is ons duidelijk dat deze manier om lexicale afstanden en uitspraakafstanden te middelen niet optimaal hoeft te zijn. Gooskens & Heeringa (2006) vonden voor het Noors dat uitspraakvariatie een belangrijkere rol speelt in de perceptie van de dialectsprekers dan lexicale variatie. Dit zou voor Nederlandse dialecten ook kunnen gelden. Een onderzoek waarin de oordelen van de dialectsprekers worden gevraagd, zou daarover meer duidelijk kunnen geven. In paragraaf 7 punt c komen we hier nog op terug.

6.2 Resultaten

6.2.1 Clusteranalyse

In paragraaf 4.5.1 bespraken we de toepassing van clusteranalyse op basis van lexicale afstanden, en in paragraaf 5.7.1 op basis van uitspraakafstanden. In deze paragraaf bespreken we het gebruik van clusteranalyse om te komen tot de classificatie van onze 360 dialecten op basis van de gecombineerde afstanden uit paragraaf 6.1. Evenals bij de lexicale afstanden en de uitspraakafstanden, hebben we ook nu verschillende clustermethoden getest en vonden opnieuw dat UPGMA een boom geeft die de originele afstanden het meest nauwkeurig weerspiegelt. De afstanden tussen dialecten zoals die door de UPGMA-boom gesuggereerd worden, verklaren voor 73% de variantie in de oorspronkelijke afstanden.

In Figuur 9 vinden we het resultaat op basis van 360 dialecten. Zoals in de Figuren 5 en 7 toont ook dit dendrogram uitsluitend de 11 meest significante groepen. De labels representeren dus subgroepen. Uitzonderingen zijn hier Tienen en Urk. De 11 groepen worden geografisch weergegeven in Figuur 10. De drie ruitjes in het noordwesten representeren Tjalleberd (links), Donkerbroek (midden) en Appelscha (rechts). Tjalleberd en Donkerbroek liggen in groep 1, maar in beide plaatsen wordt ook het dialect van groep 2 gesproken. Appelscha ligt in groep 2, maar in deze plaats wordt ook het dialect van groep 1 gesproken.

6.2.2 Vergelijking

In paragraaf 4.5.2 vergeleken we de kaart gebaseerd op lexicale verschillen met de kaart van De Schutter, en in paragraaf 5.7.2 deden we hetzelfde met de kaart die een indeling weergeeft op basis van uitspraakverschillen. Vergelijken we onze combinatiekaart met de kaart van De Schutter, dan valt meteen op dat beide kaarten een heel vergelijkbaar beeld vertonen. Onze groep 1 (Friesland) komt overeen met het Fries op de kaart van De Schutter. Groep 2 (Westerkwartier, Stellingwerf en de plaatsen Appelscha, Donkerbroek en Tjalleberd) is op de kaart van De Schutter niet als aparte groep onderscheiden, maar op de kaart van Daan vinden we het Westerkwartier als groep 25, en de Stellingwerf als groep 22.

De groepen 3 (Groningen) en 4 (Overijssel) op onze kaart vormen samen de noordoostelijke dialecten op de kaart van De Schutter. De grens tussen de noordoostelijke dialecten (op onze kaart gebied 4) en de centraal westelijke dialecten (op onze kaart gebied 9) ligt op onze kaart een stuk noordelijker dan op de kaart van De Schutter. Net als bij de uitspraakkaart (zie Figuur 8 en paragraaf 5.7.2.) merken we hier wellicht de invloed van transcribentverschillen (vergelijk de veldwerkerskaart in Figuur 4). Onze groep 9, de centraal westelijke dialecten, vinden we op de kaart van De Schutter ook als de centraal westelijke dialecten. Onze groep 5 (zuidwestelijke dialecten) komt ongeveer overeen met de zuidwestelijke dialecten op de kaart van De Schutter. Net als op de lexicale kaart (zie Figuur 6) vormt Zeeland op de combinatiekaart één groep met de centraal westelijke dialecten. De kaart van De Schutter sluit hier aan bij de uitspraakkaart (zie Figuur 8) waar Zeeland hoort bij de zuidwestelijke dialecten (zie paragrafen 4.5.2 en 5.7.2). Op de kaart van De Schutter valt de oostgrens van de zuidwestelijke dialecten voor een groot deel samen met de provinciale grens tussen West-Vlaanderen en Oost-Vlaanderen. Op onze kaart ligt deze grens verder naar het oosten, zodat het grootste deel van Oost-Vlaanderen ook valt onder de zuidwestelijke dialecten. Wanneer we kijken naar onze uitspraakkaart (Figuur 8), dan vinden we zowel de oostgrens van onze combinatiekaart als een grens die ongeveer overeenkomt met de oostgrens op de kaart van De Schutter. Beide ‘oostgrenzen’ representeren een contrast in uitspraak. Volgens de kaart van de Schutter representeert de ‘westelijke oostgrens’ het grootste contrast. Een overzicht van de taalkundige verschijnselen die aan beide oostgrenzen ten grondslag liggen wordt gegeven door Taeldeman (1978) (zie ook Taeldeman, 1979).

De centraal zuidelijke dialecten op de kaart van De Schutter zijn op onze kaart terug te vinden als groep 7 (centraal zuidelijke dialecten). De zuidoostelijke dialecten op de kaart van de Schutter bestaan op onze kaart ongeveer uit de groepen 8 (Tienen), 6 (Zuidwest-Limburg) en 11 (Oost-Limburg). Het zuidoostelijke dialectgebied op de kaart van De Schutter werd op de lexicale kaart niet apart onderscheiden, maar wel op de uitspraakkaart. Ook op de combinatiekaart wordt dit gebied onderscheiden, maar

net als op de uitspraakkaart is het verdeeld in meerdere kleinere gebieden (gebied 6 en 11). In paragraaf 5.7.2 wezen we er al op dat door sterke variatie in dit gebied het voor de sprekers misschien moeilijker is om dialectgrenzen te onderkennen, zodat het Limburgs op de kaart van Daan (en dus op de kaart van De Schutter) niet verder verdeeld is in kleinere gebieden. Verder zien we dat de begrenzing van het Limburgs op de kaart van De Schutter, de uitspraakkaart en de combinatiekaart steeds verschillend is. De noordwestelijke grens op de kaart van De Schutter is vermoedelijk vooral een uitspraakgrens (vergelijk Figuur 6), de noordelijke grens is wellicht ook lexicaal bepaald: zowel op de kaart van de Schutter als op de combinatiekaart ligt de grens zuidelijker dan op de uitspraakkaart.

Opmerkelijk is groep 8: de plaats Tienen. Door de combinatie van lexicale en uitspraakverschillen heeft deze plaats een zelfstandige positie gekregen.

7 Conclusies

We begonnen ons artikel met de constatering dat er een wirwar van verschillende methoden bestaan die allemaal weer tot een andere indeling in dialectgebieden leiden. Dit feit noopte ons tot de formulering van een reeks eisen waaraan een goede indelingsmethode moet voldoen. Die eisen hebben we op een rijtje gezet in paragraaf 2.5.1. In paragraaf 2.5.2 presenteerden we onze aanpak. In deze aanpak stelden we voor om afstandsmetingen tussen dialecten te verrichten op het lexicale niveau en op het uitspraakniveau, en om de resultaten van beide niveaus vervolgens te combineren. In dezelfde paragraaf gingen we na in hoeverre deze aanpak voldoet aan de eerder geformuleerde eisen. Ook onze aanpak blijkt niet volledig aan alle eisen te voldoen, maar we komen wel een eind in de goede richting. Nu we de resultaten van onze aanpak in de hoofdstukken 3, 4 en 5 gepresenteerd en besproken hebben, is het zinvol om alle eisen nog eens na te lopen.

a) Representativiteit
We voerden het onderzoek uit op basis van dialectteksten uit de *Reeks Nederlandse Dialectatlassen*. Iedere tekst bestaat uit 139 zinnen. Uit deze zinnen kozen we min of meer willekeurig 125 woorden. Voor zover de 139 zinnen representatief zijn, vormen de 125 woorden een representatieve steekproef (zie paragraaf 3.1). Daarmee is aan de eis van representativiteit voldaan.

b) Talige perceptie
In paragraaf 2.5.1 kozen we *talige perceptie* als onze indelingscriterium. Om een idee te krijgen in hoeverre we daarin slaagden, vergeleken we onze resultaten steeds met de kaart van De Schutter. De kaart van De Schutter is een vereenvoudigde versie van de kaart van Daan. In paragraaf 2.3.1 schreven we al dat de kaart van Daan ook niet zuiver perceptief is, maar het is dit moment wel de beste optie.

Bij vergelijking met de kaart van De Schutter vonden we overeenkomsten, maar ook verschillen. De verschillen laten zich soms verklaren door transcribentverschillen. Hoewel Blancquaert z'n best heeft gedaan om ervoor te zorgen dat de veldwerkers op dezelfde manier zouden transcriberen als hij, blijkt dit niet helemaal gelukt te zijn. De gevolgen hiervan doen zich voornamelijk gelden bij de meting van uitspraakafstanden. Echt dramatische gevolgen hebben we niet gevonden, meestal was slechts sprake van een kleine grensverschuiving.

Vergelijking van de lexicale kaart, de uitspraakkaart en de combinatiekaart met de kaart van De Schutter (en soms met de kaart van Daan) liet zien welke grenzen die in de perceptie van de dialectsprekers bestaan, vooral het gevolg zijn van lexicale verschillen, en welke het gevolg zijn van uitspraakverschillen. Verder zagen we dat grenzen in een heterogeen gebied door de dialectsprekers zelf niet altijd herkend worden.

c) Alle taalkundige niveaus

Uit het feit dat we afstanden willen meten die de talige perceptie weerspiegelen, vloeit ook voort dat we *alle taalkundige niveaus* in de beschouwing willen betrekken. Uit onze resultaten blijkt dan ook dat resultaten op basis van meerdere niveaus beter zijn dan resultaten op basis van één enkel niveau. Dit wordt het snelste duidelijk wanneer we de lexicale kaart (Figuur 6), de uitspraakkaart (Figuur 8) en de combinatiekaart (Figuur 10) vergelijken met de dialectkaart van De Schutter (Figuur 2). Op de lexicale kaart ontbreekt de zuidoostelijke dialectgroep (Limburg), op de uitspraakkaart ontbreekt de grens tussen de centraal westelijke dialectgroep (Noord-Holland, Zuid-Holland, Utrecht, Zeeland) en de centraal zuidelijke dialectgroep (Noord-Brabant, Antwerpen, Vlaams-Brabant). Op de combinatiekaart vinden we beide elementen wel terug, terwijl een aantal minder belangrijke grenzen die op de kaart van De Schutter niet weergegeven zijn, ook op onze combinatiekaart niet meer voorkomen. Aggregatie van meerdere niveaus leidt dus tot betere resultaten.

Bij het combineren van het lexicale niveau en het uitspraakniveau hebben we beide niveaus precies even zwaar laten wegen (zie paragraaf 6.1). 'Even zwaar' betekent hier dat de gecombineerde afstanden even sterk correleren met de lexicale afstanden als met de uitspraakafstanden. De vraag is echter of beide niveaus inderdaad even zwaar moeten wegen. In de toekomst willen we de weging van beide niveaus in de perceptie van de dialectsprekers verder onderzoeken.

We hebben alleen het lexicale niveau en het uitspraakniveau in ons onderzoek betrokken. Het uitspraakniveau omvat fonetische en morfologische variatie (zie paragraaf 3.4). Het syntactische niveau ontbreekt hier nog. Spruit (2005) classificeerde de dialecten in het Nederlands taalgebied op basis van een reeks syntactische verschijnselen. Opvallend in zijn resultaten is dat er geen scherp contrast is tussen de Friese en de Nedersaksische dialecten, terwijl in onze resultaten op basis van lexicon en uitspraak deze beide dialectgebieden wel sterk contrasteren. De resultaten van Spruit bevestigen Hoekstra (1998) die liet zien dat er op het syntactisch niveau tal van overeenkomsten bestaan tussen het Fries en het Gronings. Het is de bedoeling om in de toekomst de lexicale metingen en de uitspraakmetingen van dit artikel met de syntactische metingen van Spruit te vergelijken en te combineren.

d) Gradualiteit

Wanneer we letten op de methodologie die we gebruikten, dan zien we dat onze aanpak voldoet aan de eisen van *gradualiteit*. De lexicale metingen zijn gradueel door toepassing van Goebls *gewichteter Identitätswert*-methode (zie paragraaf. 4.2). De uitspraakafstanden zijn eveneens gradueel door toepassing van de Levenshtein-afstand (zie paragraaf 5.2). Dit leidt tot graduele woordafstanden, maar binnen het algoritme wordt bovendien gebruik gemaakt van graduele segmentafstanden (zie paragraaf 5.3).

e) Sensitiviteit
De door ons gebruikte methoden voldoen ook aan de eis van *sensitiviteit*. In het geval van lexicale afstandsmetingen hangt die natuurlijk ook af van de redacteur van de gegevensbron: wanneer waren volgens hem twee woorden lexicaal hetzelfde, en wanneer niet. Bij de metingen van afstanden op basis van uitspraak hangt dit af van de mate van detaillering: worden àlle diacritische tekens verwerkt, of is een grovere, en dus minder gevoelige aanpak beter? Wij hebben, in navolging van Hoppenbrouwers & Hoppenbrouwers (1988, 2001) slechts een beperkte verzameling van diacritische tekens in beschouwing genomen. Omdat niet alle transcribenten een even gevoelig gehoor hebben, is het, terwille van een consistente verwerking van heel het Nederlands taalgebied, verstandiger alleen dìe diacritische tekens mee te nemen die wellicht door alle transcribenten zouden zijn genoteerd.

f) Vergelijkbaarheid
Bij gebruik van onze methodologie kan elk dialect met elk ander dialect worden vergeleken, ongeacht of ze aan elkaar grenzen of juist ver uit elkaar liggen. Voor elk van de 360 dialecten hebben we daarom de afstand gemeten ten opzichte van alle andere 359 dialecten. Daardoor wordt een goed beeld van de onderlinge verhoudingen tussen de variëteiten verkregen, en onze classificaties zijn op dàt beeld gebaseerd (zie paragraaf 4.5.1).

g) Eenduidige interpretatie
Wanneer de taalkundige afstanden tussen dialecten eenmaal gemeten zijn, kunnen we op basis van die afstanden komen tot een indeling in dialectgroepen door toepassing van clusteranalyse. Er bestaan verschillende clustermethoden, maar we hebben onze keuze gemotiveerd door díe methode te kiezen waarmee een boom verkregen wordt die de originele afstanden het meest nauwkeurig weerspiegelt (zie paragraaf 4.5.1). Daarmee is aan de eis van eenduidige interpretatie voldaan.

h) Verificatie
Ten slotte noemen we dat onze methodologie ook voldoet aan de eis van verificatie. Op basis van het materiaal in de *Reeks Nederlandse Dialectatlassen* enerzijds en de beschrijving van de methodologie in dit artikel en in Heeringa (2004) anderzijds kan een ieder de resultaten in dit artikel reconstrueren. Graag noemen we in dit verband ook de webstek http://www.let.rug.nl/~kleiweg/L04/ . Op deze webstek kan het pakket R*u*G/L^{04} gratis geladen worden. Dit pakket is ontwikkeld door Peter Kleiweg en speciaal bedoeld voor het meten van dialectafstanden en het in kaart brengen van dialectvariatie.

We hebben in dit artikel bepaald dat lexicale variatie en uitspraakvariatie sterk met elkaar correleren (r=0.63). In toekomstig onderzoek kan gezocht worden naar een verklaring voor deze samenhang. Ook kan nader bepaald en verklaard worden waar grote lexicale verschillen bestaan terwijl de uitspraakverschillen klein zijn, of omgekeerd.

Bibliografie

Blancquaert, E. (1948). *Na meer dan 25 jaar dialect-onderzoek op het terrein.* Nr. 28; Reeks III. Gent, koninklijke Vlaamse academie voor taal- en letterkunde.

Blancquaert, E. & W. Pée, red. (1925-1982). *Reeks Nederlands(ch)e dialectatlassen*. Antwerpen, De Sikkel.

Bolognesi, R. & W. Heeringa (2002). De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT; tijdschrift voor taalwetenschap*, 9 (1), 45-84. Beschikbaar via: http://www.let.rug.nl/~heeringa/ dialectology/ papers/.

Chambers, J. & P. Trudgill (1998). *Dialectology*. Cambridge, Cambridge University Press.

Daan, J. & D. P. Blok (1969). *Van randstad tot landrand. Toelichting bij de kaart: dialecten en naamkunde.* Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam 37, Amsterdam, N.V. Noord-Hollandsche uitgevers maatschappij.

Geerts, G. (1975). Voorlopers en varianten van het Nederlands; een gedokumenteerd dia- en synchroon overzicht. Leuven, uitgeverij Acco.

Van Ginneken, J. (1913). *Handboek der Nederlandsche taal. Deel I. De sociologische structuur der Nederlandsche taal I.* Nijmegen, L.C.G. Malmberg.

Goebl, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Deel 3. Tübingen, Max Niemeyer.

Gooskens, Ch. (2005). How well can Norwegians identify their dialects? *Nordic Journal of Linguistics*, 28(1), 37-60.

Gooskens, Ch. & W. Heeringa (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language variation and change*, 16, 189-207.

Gooskens, Ch. & W. Heeringa (2006). The relative contribution of pronunciation, lexical and prosodic differences to the perceived distances between Norwegian dialects. Te verschijnen in: *Literary and Linguistic Computing*.

Goossens, J. (1970). Niederländischen Mundarten - vom Deutschen aus gesehen. *Niederdeutsches Wort*, 10, 61-80.

Goossens, J. (1977). *Inleiding tot de Nederlandse dialectologie*, Groningen, Wolters-Noordhoff.

Heeringa, W. (2001). De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen. *TABU; bulletin voor taalwetenschap,* 31, 61-103.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Proefschrift rijksuniversiteit Groningen, Groningen. Beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/thesis/.

Hoekstra, E. (1998). Oer de oerienkomst tusken de dialekten fan Grinslân en it Frysk. In: *Philologia Frisica Anno 1996, Lêzingen en neipetearen fan it fjirtjinde Frysk filologekongres, 23, 24 en 25 oktober 1996*. Ljouwert, Fryske akademy, 117-137.

Hoppenbrouwers, C. & G. Hoppenbrouwers (1988). De featurefrequentie-methode en de classificatie van Nederlandse dialecten. *TABU; bulletin voor taalwetenschap*, 18, 51-92.

Hoppenbrouwers, C. & G. Hoppenbrouwers (2001). *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen, Koninklijke Van Gorcum B.V.

Jain, A. & R. C. Dubes (1988). *Algorithms for clustering data*, Prentice Hall, Englewoord Cliffs, N.J.

Jellinghaus, H. (1892). *Die niederländischen Volksmundarten*. Norden en Leipzig, Diedr. Soltau's Verlag.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In: *Proceedings of the 7th conference of the European chapter of the association for computational linguistics*. Dublin, EACL, 60-67.

Lecoutere, C. P. F. (1921). *Inleiding tot de taalkunde en tot de geschiedenis van het Nederlandsch*. Brussel.

Nerbonne, J., W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten & W. van de Vis (1996). Phonetic distance between Dutch dialects. In: G. Durieux, W. Daelemans & S. Gillis (red.), *CLIN VI, Papers from the sixth CLIN meeting*. Antwerpen, university of Antwerp, center for Dutch language and speech, 185-202. Beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/ papers/.

Nerbonne, J. & P. Kleiweg (2003). Lexical variation in LAMSAS. In: J. Nerbonne & W. Kretzschmar (red.), *Computers and the humanities, special issue on computational methods in dialectometry*, 37, 339-357. Beschikbaar via: http://www.let.rug.nl/ ~nerbonne/paper.html.

Nerbonne, J. & C. Siedle (2005). Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. Ingediend bij: *Zeitschrift für Dialektologie und Linguistik*. Beschikbaar via: http://www.let.rug.nl/~nerbonne/paper.html.

Nieuweboer, R. (1998). *The Altai dialect of Plautdiitsch West-Siberian Mennonite Low German*. Proefschrift rijksuniversiteit Groningen, Groningen.

De Schutter, G. (1994). Dutch. In: E. König & J. van der Auwera (red.), *The Germanic languages*, Routledge Language Family Descriptions, London en New York, Routledge.

Rensink, W. G. (1955). *Dialectindeling naar opgaven van medewerkers. Mededelingen der centrale commissie voor onderzoek van het Nederlandse volkseigen 7*, 20-23.

Séguy, J. (1973). *Atlas linguistique de la France par régions, atlas linguistique de la Gascogne, complément du volume VI*. Paris, centre national de la recherche scientifique.

Spruit, M. R. (2005). Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited. In: J. Doetjes & J. van de Weijer (red.), *Linguistics in the Netherlands 2005*, 179-190. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Taeldeman, J. (1978). *De vokaalstructuur van de "Oostvlaamse" dialekten. Een poging tot historische en geografische situering in het Zuidnederlandse dialektlandschap.* Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam 51. Amsterdam, N.V. Noord-Hollandsche uitgevers maatschappij.

Taeldeman, J. (1979). Het klankpatroon van de Vlaamse dialekten. Een inventariserend overzicht. In: *Inleiding tot het Woordenboek van de Vlaamse dialekten*, 48-120. Tongeren, G. Michiels.

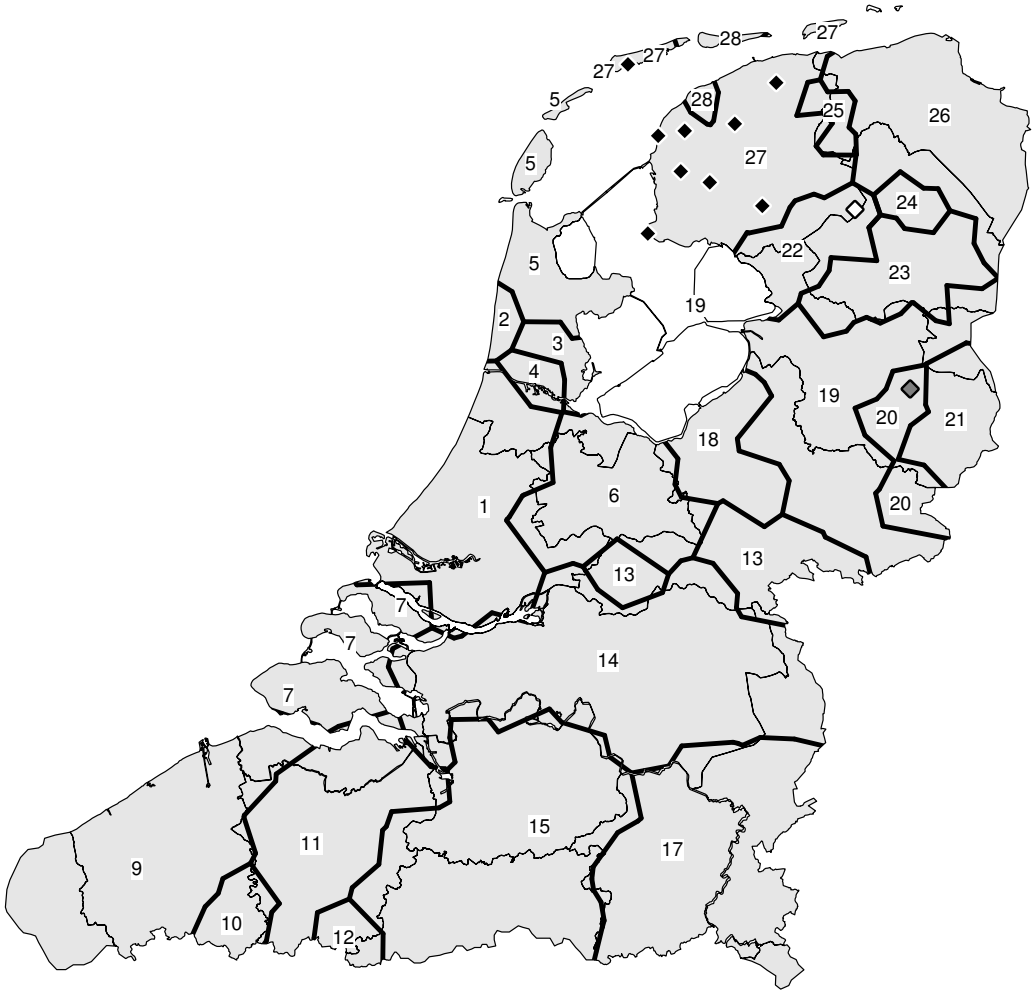
Weijnen, A. (1941). *De Nederlandse dialecten*. Groningen, Noordhoff.

Weijnen, A. (1946). De grenzen tussen de oost-noord-Brabantse dialecten onderling. In: *Oost-Noordbrabantse dialectproblemen: lezingen gehouden voor de dialectencommissie der koninklijke Nederlandsche akademie van wetenschappen op 12 april 1944*. Amsterdam, N.V. Noord-Hollandsche uitgeversmaatschappij.

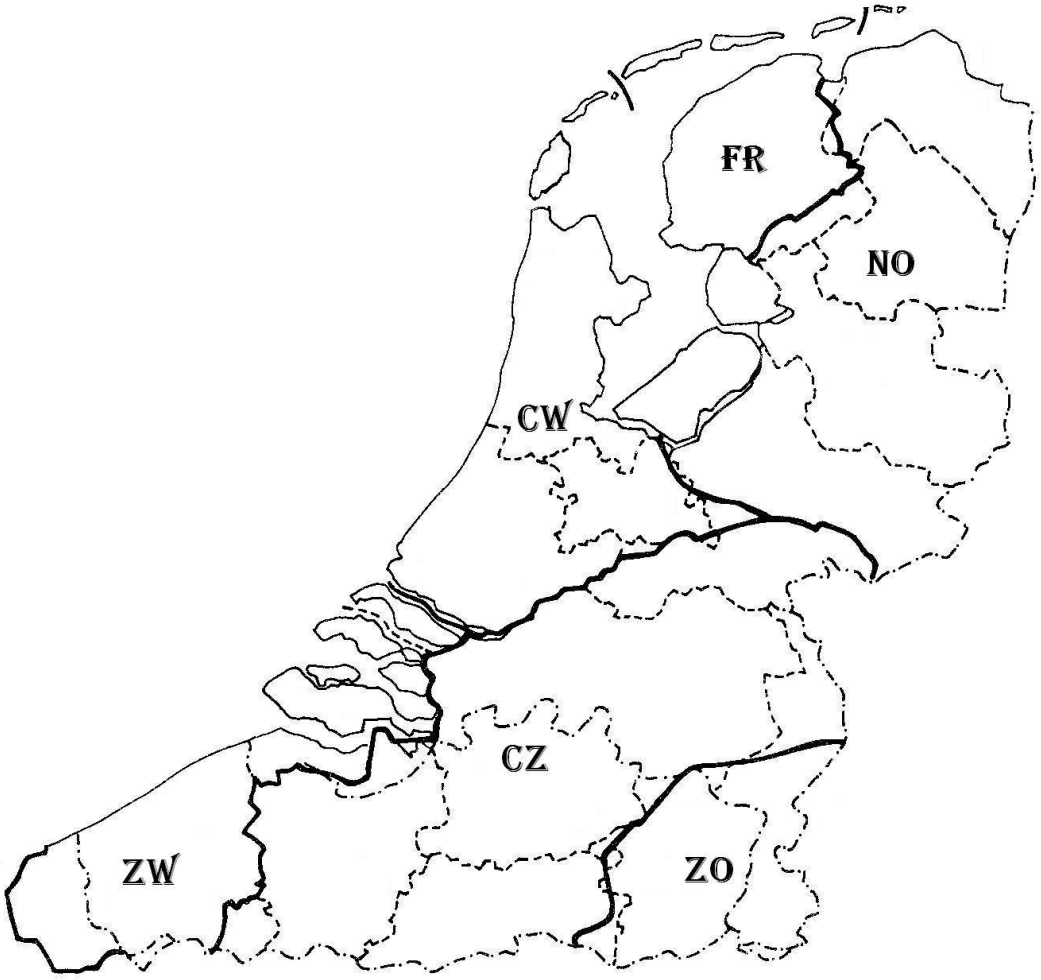
Weijnen, A. (1958). *Nederlandse dialectkunde*. Assen, Van Gorcum & Comp. N.V. - G.A. Hak & Dr. J. Prakke.

Te Winkel, J. & F. C. Wieder (1901). *Geschiedenis der nederlandsche taal, naar de tweede Hoogduitsche uitgave (1898) met toestemming van den schrijver vertaald door Dr. F. C. Wieder.* Culemborg, Blom & Olivierse.

Appendix



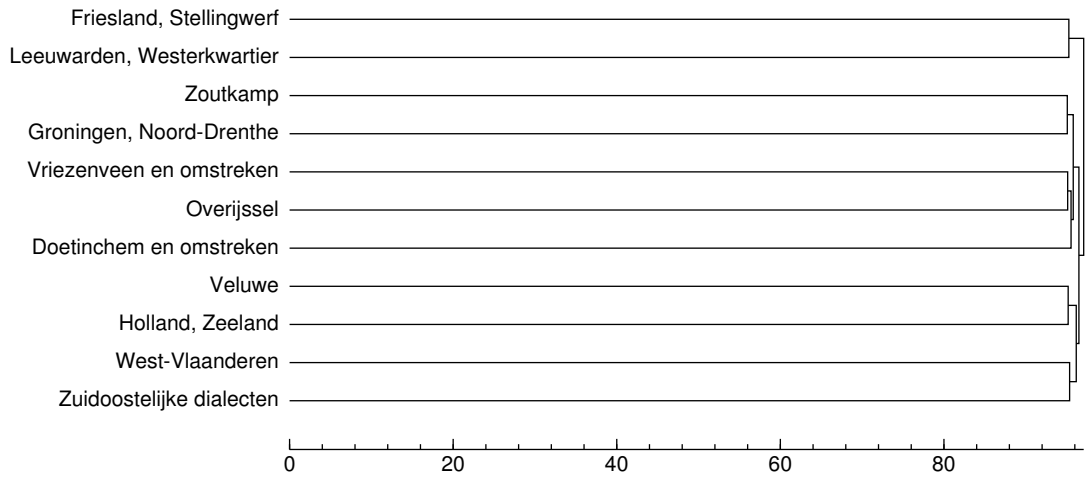
Figuur 1. De dialectkaart van Daan zoals deze van toepassing is op de 360 dialecten die in dit artikel gebruikt worden. De dunne lijnen geven grenzen tussen provincies en/of landen weer, de dikke lijnen geven grenzen tussen dialectgebieden weer. Ruitjes representeren dialecteilanden. De zwarte ruitjes representeren de Friese steden die behoren tot groep 28 (stadsfries). Het witte ruitje representeert Appelscha. In Appelscha wordt zowel het dialect van groep 22 (Stellingwerfs) als van groep 27 (Fries) gesproken. Het grijze ruitje representeert het dialect van Vriezenveen dat sterk constrasteert met het dialect dat in de omringende plaatsen gesproken wordt. De verdeling van de 360 variëteiten wordt weergegeven in Figuur 3.



Figuur 2. De dialectkaart van De Schutter. De kaart is een vereenvoudigde versie van de kaart van Daan en laat de zes belangrijkste gebieden zien. De afkortingen hebben de volgende betekenissen: FR=Fries, NO=noordoostelijke dialecten, CW=centraal westelijke dialecten, ZW=zuidwestelijke dialecten, CZ=centraal zuidelijke dialecten, ZO=zuidoostelijke dialecten. De kaart is overgenomen uit De Schutter (1994), waarbij de namen van de dialectgebieden zijn vertaald uit het Engels in het Nederlands.

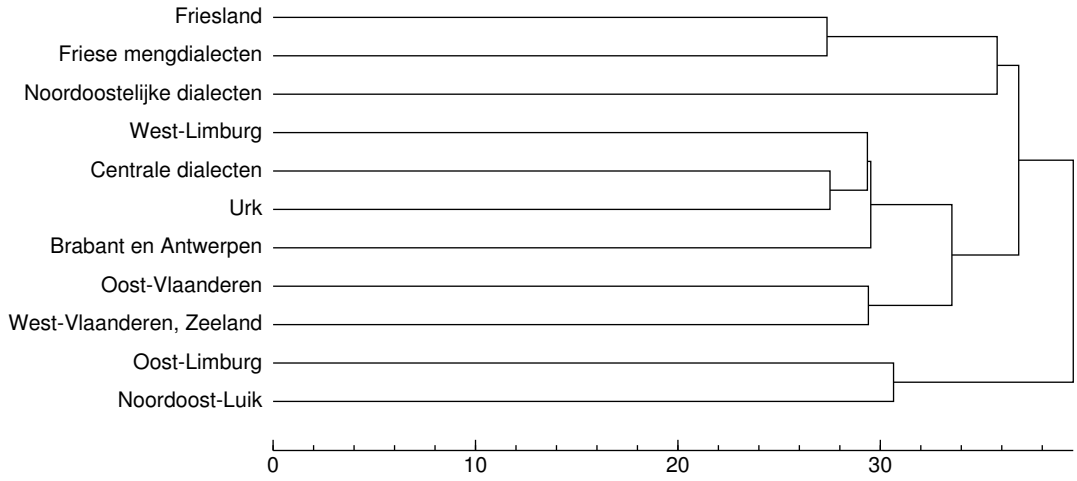
Figuur 3. Verdeling van de 357 plaatsen, corresponderend met 360 verschillende variëteiten. Witte ruitjes representeren taaleilanden (Friese steden), en grijze ruitjes plaatsen waar twee dialecten worden gesproken. Daarbij geldt één van de twee dialecten als een taaleiland. Het betreft Tjalleberd (meest linkse ruitje), Donkerbroek (middelste ruitje) en Appelscha (meest rechtse ruitje). Kleine cirkeltjes representeren kleine geografische eilanden.

Figuur 4. Verdeling van de 360 dialectplaatsen over 16 atlasdelen en 16 veldwerkers. De dunne lijnen geven grenzen tussen provincies en/of landen weer, de dikke lijnen geven grenzen tussen transcribentgebieden weer. De drie-letterige afkortingen hebben de volgende betekenis: bla=E. Blancquaert, boe=K. Boelens, cla=J.C. Cleassens, daa=Jo Daan, ent=H. Entjes, gof=W. Goffin, hol=A.R. Hol, mee=P. J. Meertens, oye=L. Oyen, pas=J. Passage, pee=Willem Pée, sas=A. Sassen, ste=A. Stevens, van=H. Vangassen, wei=A. Weijnen, wou=G. van der Woude. De cijfers zijn de nummers van de atlasdelen. De opnamen in deel 2 werden òf door Blancquaert òf door Vangassen gemaakt. De plaatsen in deel 5 werden gezamenlijk bezocht door Blancquaert en Meertens. Zwarte ruitjes representeren opnamen van Blancquaert, witte diamenten gezamenlijke opnamen van Blancquaert en Oyen, en grijze ruitjes gezamenlijke opnamen van Blancquaert, Boelens en Van der Woude.

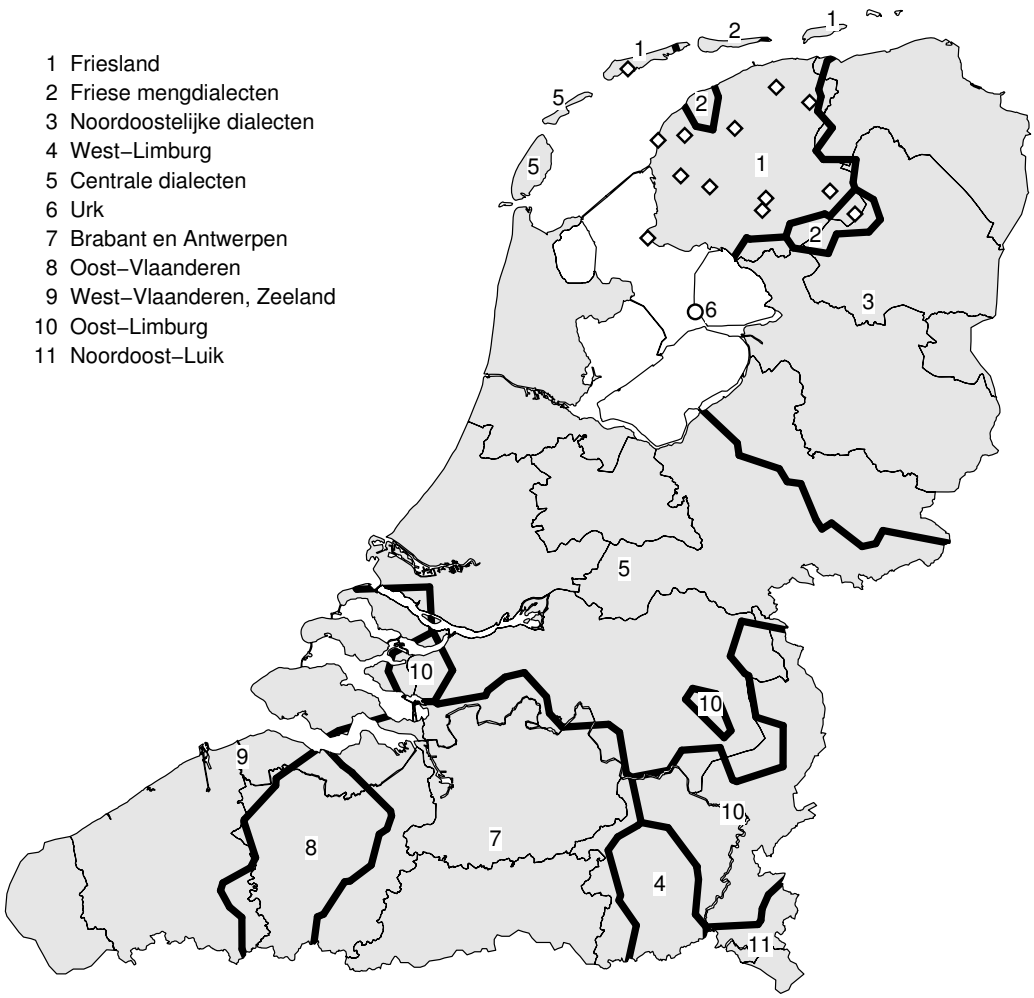


Figuur 5. Dendrogram op basis van de onderlinge lexicale afstanden tussen 360 dialecten. De schaal representeert percentages (zie paragraaf 4.2). De 11 labels zijn (met uitzondering van Zoutkamp) niet de bladeren van de bomen, maar representeren subgroepen met daarbinnen een verdere verdeling in dialecten. De 11 groepen worden geografisch weergegeven in Figuur 6. De boomstructuur verklaart 71% van de variantie.

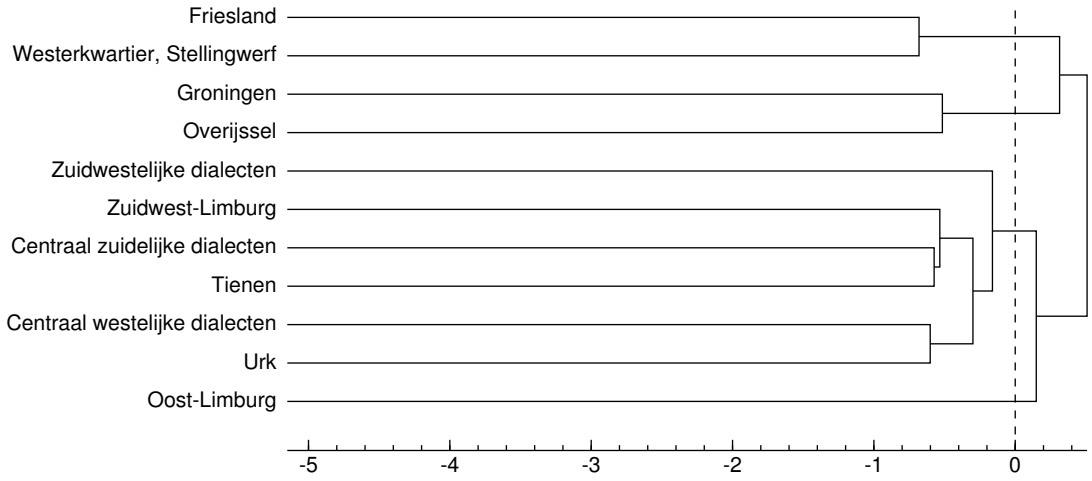
Figuur 6. Verdeling op basis van lexicale afstanden. De kaart toont de 11 meest significante groepen uit het dendrogram in Figuur 5. Het ruitje in het noordwesten representeert het dialect van Leeuwarden dat één groep vormt met het dialect van het Westerkwartier (groep 2).



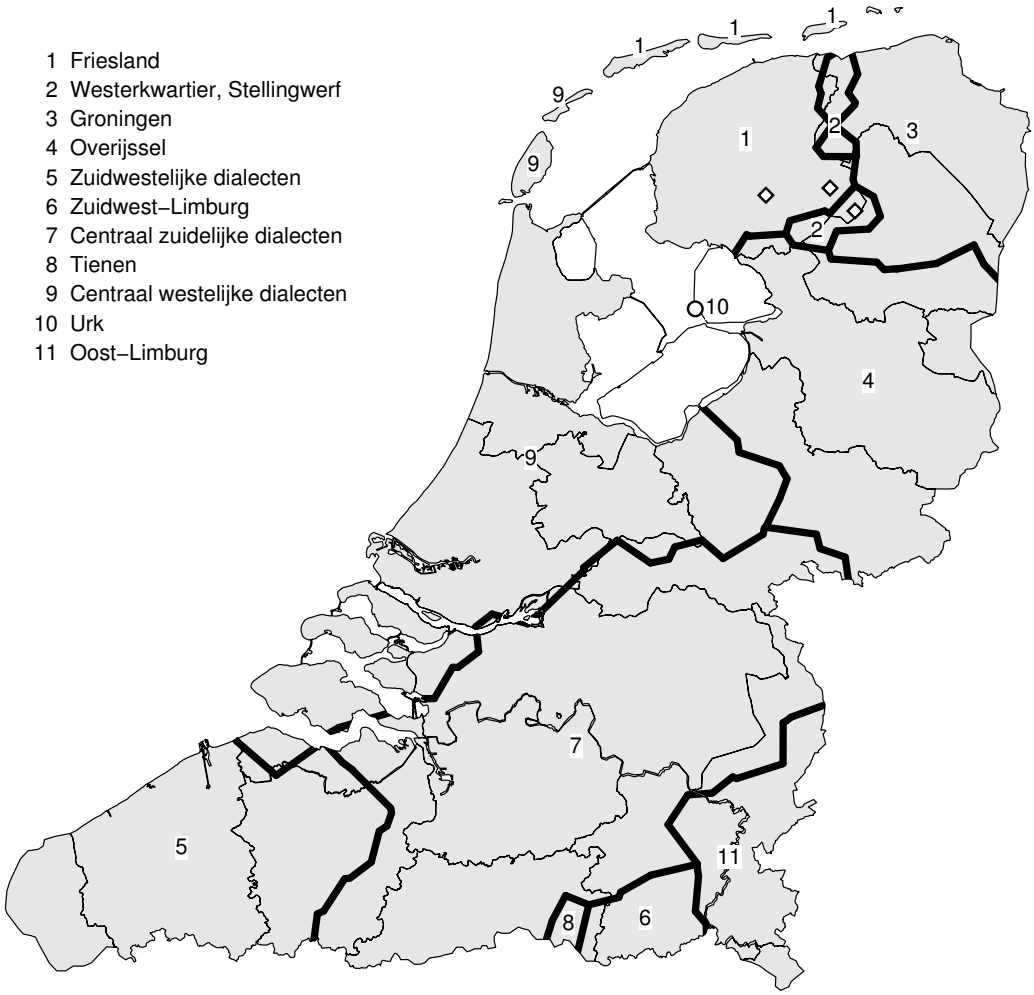
Figuur 7. Dendrogram op basis van de onderlinge uitspraakafstanden tussen 360 dialecten. De schaal representeert percentages (zie paragraaf 5.4). De 11 labels zijn (met uitzondering van Urk) niet de bladeren van de bomen, maar representeren subgroepen met daarbinnen een verdere verdeling in dialecten. De 11 groepen worden geografisch weergegeven in Figuur 8. De boomstructuur verklaart 70% van de variantie.



Figuur 8. Verdeling op basis van uitspraakafstanden. De kaart toont de 11 meest significante groepen uit het dendrogram in Figuur 7. De ruitjes in het noordwesten representeren dialecteilanden die behoren tot de Friese mengdialecten (groep 2).



Figuur 9. Dendrogram op basis van gecombineerde afstanden (lexicon en uitspraak) tussen 360 dialecten. De schaal representeert z-waarden (zie paragraaf 6.1). De 11 labels zijn (met uitzondering van Tienen en Urk) niet de bladeren van de bomen, maar representeren subgroepen met daarbinnen een verdere verdeling in dialecten. De 11 groepen worden geografisch weergegeven in Figuur 10. De boomstructuur verklaart 73% van de variantie.



Figuur 10. Verdeling op basis van gecombineerde afstanden (lexicon en uitspraak). De kaart toont de 11 meest significante groepen uit het dendrogram in Figuur 9. De drie ruitjes in het noordwesten representeren Tjalleberd (links), Donkerbroek (midden) en Appelscha (rechts). In deze plaatsen wordt zowel een Fries dialect (groep 1) als een Saksisch/Fries mengdialect (groep 2) gesproken.

Eindnoten

[1] Wilbert Heeringa, Rijksuniversiteit Groningen, Faculteit der Letteren, Vakgroep Informatiekunde, Postbus 716, 9700 AS Groningen, w.j.heeringa@rug.nl . John Nerbonne, Rijksuniversiteit Groningen, Faculteit der Letteren, Vakgroep Informatiekunde, Postbus 716, 9700 AS Groningen, j.nerbonne@rug.nl . Graag danken wij Peter Kleiweg voor het beschikbaar stellen van de programmatuur waarmee de dendrogrammen en de meeste kaarten in dit artikel gemaakt zijn.
Het onderzoek zoals beschreven in dit artikel werd uitgevoerd in het kader van het project *The Determinants of Dialectal Variation*. Dit project wordt gesubsidieerd door de *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (360-70-121). Graag bedanken we ook de anonieme beoordelaars van dit artikel voor hun waardevolle opmerkingen.

[2] Een isoglosse is een lijn op een kaart die variatie met betrekking tot een taalkundig verschijnsel representeert, zodanig dat een gebied met de ene vorm afgegrensd wordt van een gebied met een andere vorm. Een bekend voorbeeld is de slot-n-isoglosse die oostelijk van de provincie Utrecht en noordelijk van Arnhem loopt. Aan de noordoostelijke kant van deze isoglosse wordt een woord als *lopen* uitgesproken als [loːpm̩], en aan de zuidwestelijke kant als [loːpə].

[3] Een uitvoerige bespreking van de keuze van de woorden uit de RND kan worden gevonden in Heeringa (2001). De gedigitaliseerde data zijn met toestemming van uitgeverij De Sikkel (later opgegaan in De Boeck, Antwerpen) vrij beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/atlas/.

[4] In Heeringa (2001) wordt uitgebreider ingegaan op de verschillen tussen de vragenlijsten.

[5] Deze vier gevallen worden uitvoerig besproken in Heeringa (2004, blz. 224-226).

[6] Voor elk van de eerste vier variabelen geven we een voorbeeld uit het werk van Séguy. Diachroon fonetische variabele: hoe wordt een bepaalde consonant tussen twee vocalen in uitgesproken? 0=stemloos, 1=stemhebbend. Fonologische variabele: eindigt een woord op een /p/? 0=nee, 1=ja. Morfo-syntactische variabele: volgt na een prepositionele IF (infinitief) een postpositie? 0=nee, 1=facultatief, 2=ja, is verplicht. Werkwoordsverbuigingsvariabele: wat is de aard van het accent op de IP (indicatief presens)? 0=zwak, 1=hybride, 2=krachtig, 3=polymorf.