WILBERT HEERINGA, JOHN NERBONNE, RENÉE VAN BEZOOIJEN, MARCO RENÉ SPRUIT

Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied[1]

Abstract – It is a fundamental insight of dialectology that language variation is structured geographically (Nerbonne & Kleiweg 2006). Besides a low geographic distance, large population sizes may increase social contact between two locations and the chance that the respective dialects are influenced by each other. Analoguous to the gravity model in physics, Trudgill 1974 combined linguistic similarity, geography and population sizes in one model as an index of linguistic influence between dialect locations. Following Nerbonne & Heeringa 2006 we use a model which combines geography and population sizes only to explain variation in aggregate dialect distances. In contrast to Nerbonne & Heeringa we use data of a larger and less homogeneous area. The data set comprises 27 varieties in the Netherlands and north-Belgium. In accordance with Nerbonne & Heeringa 2006 we found geography to be an important predictor, but could not prove a significant additional value of population size in explaining linguistic variation.

1 Inleiding

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*. Séguy wilde de kaarten in deze atlassen op een objectievere manier analyseren dan mogelijk was met de traditionele methoden. Séguy en zijn onderzoeksteam deden dit door voor elk tweetal naburige dialectplaatsen eenvoudigweg 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).

Sterk verwant aan de methodologie van Séguy is het werk van Hans Goebl, ofschoon de basis van het werk van Goebl vrijwel onafhankelijk van Séguy is ontwikkeld. Met de hulp van Edgar Haimerl slaagde Goebl er ook in om de gemeten taalkundige afstanden geografisch weer te geven op kaarten. Een voorbeeld is een kaart die de afstanden van dialecten ten opzichte van een referentiepunt weergeeft. Zo'n referentiepunt kan een dialect zijn, of de standaardtaal. De mate van verwantschap wordt door middel van het regenboogschema weergegeven: rood betekent: is maximaal verwant, blauw betekent: is maximaal verschillend. In Goebl 2005 worden dergelijke kaarten gegeven voor het Frans, Engels en Italiaans. Aan iedere dialectplaats en het gebiedje rondom wordt een kleur toegekend, zodat patronen ontstaan en in één oogopslag te zien is welke dialectgebieden verwant zijn, en welke niet. Er zijn ook kaarten waarbij afstanden tussen naburige dialecten weergegeven worden. Als de taalkundige afstand groot is, worden ze gescheiden door een dikke donkere lijn. Bij een kleine afstand is de lijn juist dun en licht. Op die manier wordt het verloop in het dialectlandschap inzichtelijk weergegeven. Voorbeelden van deze kaarten vinden we onder andere in Goebl 2002 en Goebl 2003.

[1] Wij danken Peter Kleiweg voor het beschikbaar stellen van zijn programmatuur voor het tekenen van kaarten en voor het berekenen van afstanden tussen plaatsen. Graag bedanken we ook Frans Hinskens voor opmerkingen en suggesties bij de voorlaatste versie van deze bijdrage.

De idee van taalkundige afstandsmetingen tussen dialecten werd voor het Nederlandse dialectgebied voor het eerst toegepast door de gebroeders Hoppenbrouwers in 1988. Zij introduceerden de featurefrequentiemethode. Als we beschikken over fonetische transcripties van een reeks dialecten, dan kunnen we voor ieder dialect tellen hoeveel klanken in de bijbehorende transcriptie voorkomen die voor in de mond uitgesproken worden, of die met een hoge tongpositie uitgesproken worden, of die met geronde lippen uitgesproken worden. Behalve kenmerken (of features) van klinkers worden ook medeklinkerkenmerken in de beschouwing betrokken. De taalkundige afstand tussen twee dialecten is in het eenvoudigste geval gelijk aan de som van de verschillen in de frequentie van voorkomen van de diverse kenmerken. 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 vergelijkbare gegevens van 156 dialectplaatsen in het Nederlandse taalgebied.

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 (en ook Goebl's) 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. Beide problemen worden ondervangen door gebruik van de Levenshtein-afstand. 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 klanksegmenten worden toegevoegd, vervangen of verwijderd. Kesslers aanpak gaf plausibele resultaten 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 dit artikel willen we de vraag beantwoorden waarom sommige dialecten sterk verwant aan elkaar zijn, terwijl andere dialecten juist sterk van elkaar verschillen. Waarom lijken de dialecten van Amsterdam en Utrecht relatief veel op elkaar, maar zijn de dialecten van Amsterdam en Maastricht heel verschillend? We willen onderzoeken welke rol sociaal contact heeft in de verklaring van dialectvariatie in het Nederlandse taalgebied (Nederland en Noord-België). We maken daarbij gebruik van materiaal van 27 variëteiten dat in 2001 verzameld werd door Renée van Bezooijen. De mate van sociaal contact wordt vooral bepaald door de geografische afstand tussen twee plaatsen, en het aantal inwoners in beide plaatsen. Trudgill 1974 heeft deze beide factoren verenigd in één formule naar analogie van het zwaartekrachtmodel uit de natuurkunde. We bespreken dit model in paragraaf 2. Dialectvariatie meten we met de Levenshtein-afstand. Deze maat wordt uitgelegd en toegepast in paragraaf 3. In paragraaf 4 onderzoeken we de rol van

geografie en inwoneraantallen in de verklaring van dialectvariatie. We eindigen met het trekken van enkele conclusies in paragraaf 5.

2 Geografie en inwoneraantallen

2.1 Model

De *zwaartekracht* of *gravitatie* is een aantrekkende kracht die twee massa's op elkaar uitoefenen. De zwaartekracht zorgt ervoor dat een voorwerp altijd naar beneden valt wanneer je het loslaat. Isaac Newton heeft de rol van de zwaartekracht voor het eerst in een formule vastgelegd:

$$F = G\frac{m_1 \times m_2}{r \times r}$$

waarin F de zwaartekracht tussen twee objecten (in Newton) is, m_1 en m_2 de respectieve massa's van twee objecten (in kg), r de afstand tussen de objecten (in m), en G de gravitatieconstante, de kracht in Newton die twee objecten met elk een massa van 1 kg, op een afstand van 1 m op elkaar uitoefenen. (Wikipedia-bijdragers 2006).

Naar analogie van dit model formuleerde Zipf 1946 een model voor de beschrijving van migratiebewegingen tussen steden: de migratie tussen plaats *i* en plaats *j* is recht evenredig met het product van de inwoneraantallen van beide steden en omgekeerd evenredig met de geografische afstand tussen beide steden.

Het zwaartekrachtmodel werd voor het eerst geïntroduceerd in de dialectologie door Peter Trudgill in 1974. Trudgill 1974 gebruikte het model als index van de taalkundige invloed tussen plaatsen en – in aangepaste vorm – van de ene plaats op de andere plaats. Trudgill paste het model toe op dialecten in het zuiden van Noorwegen (Trudgill 1974 en Trudgill 1983) en het oosten van Engeland (Trudgill 1983). Het model werd ook toegepast door Hinskens (1992; 1993) op enkele Limburgse dialecten. Het idee is dat taalkundige invloed bepaald wordt door drie factoren: taalkundige verwantschap, geografische afstand en inwoneraantallen. Volgens Trudgill 1983 (p. 74) nemen dialectsprekers taalkundige verschijnselen gemakkelijker van elkaar over naarmate hun respectieve dialecten taalkundig verwanter zijn. In het model van Trudgill representeert de constante G de taalkundige verwantschap.

Wat betreft geografie: naarmate plaatsen dichter bij elkaar liggen, zal er ook meer contact zijn tussen de inwoners van beide plaatsen. In het zwaartekrachtmodel wordt deze afstand r gekwadrateerd. Vanuit een gegeven punt kan een inwoner immers in alle richtingen gaan (noord, noordoost, oost, enz.). De kans dat een inwoner gaat naar een punt op een denkbeeldige cirkel rond zijn of haar woonplaats is gelijk aan $1/r^2$.

De mate van invloed wordt ook verondersteld afhankelijk te zijn van het aantal inwoners in de plaatsen. Iedere inwoner uit de ene plaats kan in contact komen met iedere inwoner in de andere plaats. Stel dat de twee plaatsen respectievelijk m_1 en m_2 inwoners hebben, dan zal de kans op wederzijdse taalkundige invloed toenemen met het product $m_1 \times m_2$.

In dit artikel gebruiken we eveneens het zwaartekrachtmodel. Wij gebruiken het model echter niet als index van taalkundige *invloed*, maar – in navolging van Nerbonne

& Heeringa 2006 – als verklarend model van taalkundige *afstand*. Het idee is dat het model fungeert als index van sociaal contact. Onze hypothese is dat sociaal contact een goede voorspeller is voor taalkundige afstand, en dat sociaal contact – vereenvoudigd voorgesteld – bepaald wordt door de afstand tussen twee plaatsen en het aantal inwoners in beide plaatsen. In het model van Trudgill representeert de constante *G* de mate van taalkundige overeenkomst. Omdat wij het zwaartekrachtmodel willen gebruiken om taalkundige variatie in dialecten te verklaren, willen we die variatie juist niet in het model opnemen. De constante vervalt dus in ons geval.

Nerbonne & Heeringa 2006 gebruikten het model voor de verklaring van variatie tussen 52 dialecten in het noordoosten van Nederland. Wij passen het model toe op 27 dialecten die verspreid liggen in heel Nederland en Noord-België.

2.2 Metingen

De basis voor de experimenten in dit artikel wordt gevormd door een gegevensverzameling van Renée van Bezooijen. Deze gegevensverzameling omvat gegevens van 26 Nederlandse en Noord-Belgische dialecten plus het standaard Fries. Het standaard Fries blijkt het sterkst verwant te zijn aan het dialect van Grouw. We localiseren het standaard Fries daarom op de positie van Grouw.

Op basis van de coördinaten in longitude (lengtegraad) en latitude (breedtegraad) hebben we de hemelsbrede afstanden in kilometers tussen de 27 plaatsen berekend.[2] De inwoneraantallen voor de Nederlandse plaatsen zijn afkomstig van het *Centraal Bureau voor de Statistiek*[3] en de inwoneraantallen voor de Belgische plaatsen zijn rechtstreeks opgevraagd bij de gemeenten. De aantallen zijn gegeven in Tabel 1.

Amsterdam	742780	Borculo	10350	Uithuizen	5100
's-Gravenhage	472100	De Panne	7334	Stokkem	3635
Utrecht	275260	Rijssen	7210	Zandvliet	3575
Maastricht	121460	Grouw	7130	De Lutte	3080
Kerkrade	49560	Obdam	6790	Uitbergen	1825
Etten-Leur	39860	Meijel	5850	Ouwegem	1488
Gemert	14780	Kampenhout	5394	Hooghalen	1430
Reeuwijk	12830	Ossendrecht	5300	‘s-Heerenhoek	590
Stein	11680	Westenholte	5180	Deelen	60

Tabel 1. Aantal inwoners per plaats op 1 januari 2005 (soms 31 december 2004).[4]

Zoals beschreven in paragraaf 2.1 meten we de producten van de inwoneraantallen. Het inwonerproduct voor bijvoorbeeld het paar Amsterdam-Deelen is gelijk aan 742780 ı 60 = 44566800 inwoner-paren.

[2] Zie http://www.let.rug.nl/~kleiweg/L04/Manuals/ll2dst.html voor uitleg over het door ons gebruikte programma voor de berekening van de geografische afstanden.
[3] Zie http://www.statline.nl .
[4] Voor Uitbergen en Kampenhout was het aantal inwoners op 1 januari 2005 op het moment van schrijven niet beschikbaar. De aantallen zijn gegeven voor respectievelijk 1 januari 2006 en 1 januari 2003.

3 Meting van taalkundige afstanden

3.1 Levenshtein-afstand

In deze paragraaf introduceren we een een simpele versie van de Levenshtein-afstand, waarbij we er gemakshalve vanuit gaan 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 klanksegmenten) 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 Westenholte[5] 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	voeg ə toe	1
kniːnə		
---	---	---
		3

In dit voorbeeld hebben we aan iedere operatie één punt toegekend. 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, delen we de Levenshtein-afstand door de lengte van de oplijning.[6] Een oplijning laat zien welk segment in het ene woord correspondeert met welk segment in het andere woord, en welke segmenten in het ene woord zijn toegevoegd of verwijderd ten opzichte van het andere woord. In ons voorbeeld ziet de oplijning er als volgt uit:

1	**2**	**3**	**4**	**5**	**6**
k	ə	n	ɛː	n	
k		n	iː	n	ə
---	---	---	---	---	---
0	1	0	1	0	1

Wanneer we de Levenshtein-afstand (1+1+1=3) delen door de lengte van de oplijning (6), krijgen we een genormaliseerde woordafstand van 3/6 = 0.5, oftewel 50%.[7] Zouden we

[5] Westenholte was één van de dorpen die rond Zwolle lagen en samen de gemeente Zwollekerspel vormden. Zwollekerspel is later opgegaan in de gemeente Zwolle.
[6] Zie Heeringa 2004:130-133 voor een gedetailleerde uitleg.
[7] Bij gebruik van ongenormaliseerde afstanden wordt de *local incoherence*, een maat die de samenhang tussen geografie en taalkundige afstand op locaal niveau bekijkt, hoger (zie Nerbonne & Kleiweg 2006). Dat betekent dat het resultaat slechter wordt. Heeringa et al. 2006 vonden voor Noorse dialecten echter het omgekeerde: ongenormaliseerde metingen benaderden de waarneming van de dialectsprekers beter dan genormaliseerde metingen.

[kənɛːn] (Amsterdam) vergelijken met bijvoorbeeld [kniːn] (Maastricht), dus zonder de finale [ə], dan wordt het aantal operaties gelijk aan 2 en de lengte van de oplijning wordt gelijk aan 5 (de noemer is altijd de lengte van de langste oplijning). Dit geeft een genormaliseerde afstand van 2/5=0.4.

Om ervoor te zorgen dat de Levenshtein-afstand is gebaseerd op een oplijning waarin de lettergrepen in het ene woord correct ten opzichte van de corresponderende lettergrepen in het andere woord zijn opgelijnd, is het belangrijk om niet alle mogelijke segmentcorrespondenties in een oplijning toe te staan. Onze versie van het Levenshtein-algoritme is zodanig aangepast dat een klinker alleen mag corresponderen met een klinker en een medeklinker alleen met een medeklinker. De [j] en de [w] mogen ook met een klinker corresponderen (of omgekeerd), en de [i] en de [u] met een consonant (of omgekeerd). De sjwa mag corresponderen met een sonorant. Op die manier worden onwaarschijnlijke correspondenties voorkomen.

3.2 Graduele gewichten

In dit artikel gebruiken we een verfijndere versie van het algoritme met graduele gewichten voor de drie operaties. Daarbij wordt rekening gehouden met de mate van verwantschap tussen klanken zodat uit de verf komt dat bijvoorbeeld de [ɪ] en de [e] meer op elkaar lijken dan de [ɪ] en de [ɔ]. De gewichten zijn gebaseerd op akoestische metingen tussen samples op de cassette *The Sounds of the International Phonetic Alphabet* die uitgegeven werd in 1995. Onze metingen zijn zuiver fonetisch: het doet er niet toe of een klankverschil tot een betekenisverschil kan leiden, bepalend is of er verschil in klankkleur is. Bijvoorbeeld: in tegenstelling tot bijv. de [a] van ‘maan’en de [ɑ] van ‘man’ zijn de [r] en de [ʀ] in het Nederlands niet betekenisonderscheidend, maar het verschil tussen beide klanken wordt door ons wel in rekening gebracht, evenals dat tussen [a] en [ɑ]. Voor details zie Heeringa 2004 (hoofdstuk 4).

3.3 Aggregatie

De afstand tussen twee dialecten wordt niet berekend op basis van één enkel woordpaar, maar op basis van een reeks van *n* woordparen. Stel we berekenen de afstand tussen Amsterdam en Westenholte op basis van zes woorden. De berekening ziet er dan als volgt uit:[8]

[8] Om het voorbeeld eenvoudig te houden gebruiken we hier weer geen graduele klankafstanden, maar de ruwere aanpak waarbij de drie gewichten (toevoegen, vervangen, verwijderen) altijd de waarde 1 hebben. Ook laten we diacritsche tekens (bijvoorbeeld lengte) buiten beschouwing. Een diftong wordt verwerkt als de opeenvolging van twee monoftongen.

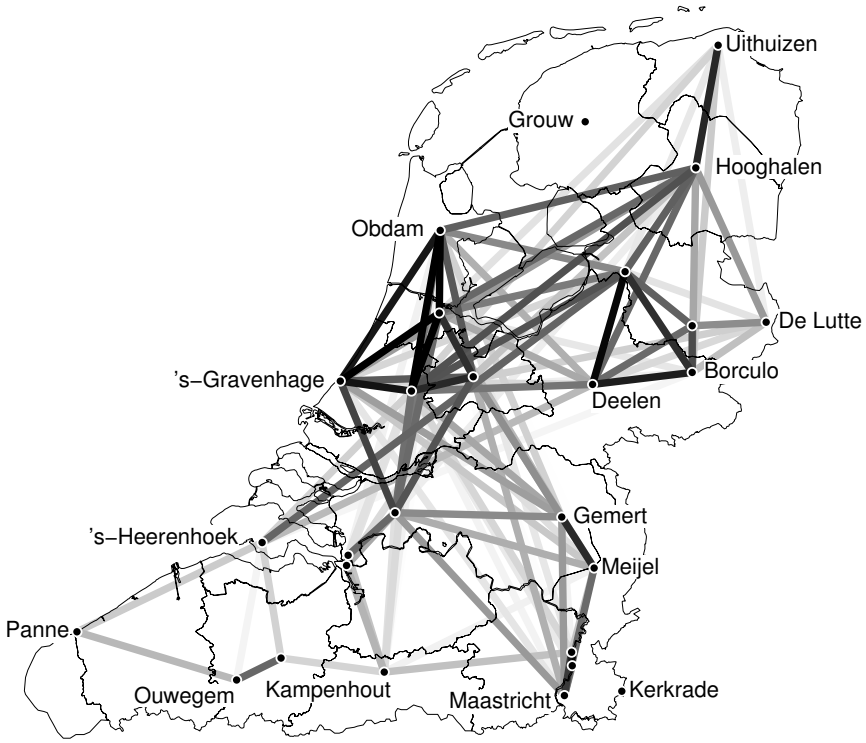
Item	Amsterdam	Westenholte	Levenshtein-afstand	lengte oplijning	genormaliseerde afstand
Dak	dɑk	dɑk	0	3	0
keuken	køkə	køkn	1	4	0.25
muis	mœys	mus	2	4	0.50
konijn	kənɛn	kninə	3	6	0.50
aardappel	ɑrəpəl	erpəl	2	6	0.33
tulp	tʏləp	tʏlpə	2	6	0.33
					1.91

De laatste kolom geeft de genormaliseerde Levenshtein-afstanden. Deze genormaliseerde Levenshtein-afstanden aggregeren we. De afstand tussen Amsterdam en Westenholte wordt nu gelijk aan (1.91/6)*100=31.8%.

3.4 Metingen

Voor elk van de dialecten zijn de vertalingen van 100 zelfstandige naamwoorden opgevraagd en getranscribeerd in fonetisch schrift. Het gaat om voorwerpen en begrippen uit het alledaagse leven, zodat de woordenlijst mag worden beschouwd als een tamelijk representatieve steekproef uit de woordenschat van de dialecten.

Zoals beschreven in paragraaf 3 worden afstanden tussen woorden gemeten met de Levenshtein-afstand. Omdat we 100 woorden hebben, is de afstand tussen twee dialecten gelijk aan het gemiddelde van 100 Levenshtein-afstanden. De afstanden zijn weergegeven op de kaart in Figuur 3. Met name in de Randstad vinden we een groep van relatief sterk aan elkaar verwante dialecten. Verder ook in Groningen/Noord-Drenthe en Zuid-Overijssel/Noord-Gelderland. Het Fries (de stip in het noordwesten) heeft alleen maar grote afstanden ten opzichte van de andere variëteiten. Die afstanden zijn zo groot dat het Fries eigenlijk door witte lijnen verbonden is met de andere variëteiten. Maar die witte lijnen zijn in de kaart niet zichtbaar omdat ze wegvallen tegen de witte achtergrond van de kaart. Hetzelfde geldt voor Kerkrade, helemaal in het zuidoosten vlak tegen de grens met Duitsland.



Figuur 3. Gemiddelde Levenshtein-afstanden tussen 27 dialecten in Nederland en het noorden van België. Hoe donkerder de lijn, hoe kleiner de afstand.

4 Resultaten

4.1 Correlatie met de geografie

Een fundamenteel inzicht in de dialectologie is dat dialectvariatie samenhangt met de geografie (Nerbonne & Kleiweg 2006). Wanneer tussen dialecten die geografisch ver uit elkaar liggen een grote taalkundige afstand bestaat, terwijl dialecten die vlak bij elkaar liggen taalkundig heel verwant zijn, zeggen we dat geografie en taalkundige afstand met elkaar correleren. Een correlatiecoëfficiënt drukt de mate van samenhang uit en kan variëren van -1 (hoe lager de geografische afstanden, hoe hoger de taalkundige afstanden) tot 1 (hoe hoger de geografische afstanden, hoe hoger de taalkundige afstanden). Wanneer we de correlatiecoëfficient *r* kwadrateren en vermenigvuldigen met 100, krijgen we de variantie in de taalkundige afstanden die verklaard wordt door de geografische afstanden. Variantie is een maat voor variatie. De waarden van een reeks gegevens – in ons geval de taalkundige afstanden – worden vergeleken ten opzichte van hun gemiddelde. De variantie is ongeveer gelijk aan de gemiddelde kwadratische afwijking van een waarde ten opzichte van het gemiddelde. We kregen de volgende resultaten:

transformatie geografie	correlatie geografie	verklaarde variantie geografie
kwadratisch	0.49	24%
geen	0.58	33%
wortel	0.61	37%
logaritmisch	0.62	39%

In het zwaartekrachtmodel worden gekwadrateerde geografische afstanden gebruikt, maar we zien dat die het slechtste correleren. Séguy 1971 gebruikte de wortel van de geografische afstanden, en we zien dat die in ons geval beter correleren dan de niet-getransformeerde en de gekwadrateerde geografische afstanden. Nog iets beter correleren de logaritmische geografische afstanden die zelfs statistisch significant beter correleren dan de gekwadrateerde afstanden (p=0.04).[9] Het logaritmische model suggereert dat op kleinere schaal een toename van geografische afstand samengaat met een significante toename in taalkundige afstand. Maar naarmate de geografische afstand groter wordt, wordt de toename in taalkundige afstand steeds kleiner en minder betekenisvol. Bij grote geografische afstanden is vooral het feit *dat* dialecten heel verschillend zijn belangrijk, en veel minder belangrijk is de *mate* waarin de dialecten taalkundig van elkaar verschillen (vergelijk Heeringa & Nerbonne 2001).

4.2 Correlatie met de producten van de inwoneraantallen

Inwoneraantallen blijken 6% van de variatie in de taalkundige afstanden te verklaren. Preciezer geformuleerd: de producten van de inwoneraantallen in Trudgill's formule (zie paragraaf 2.1) blijken 6% van de variantie in de taalkundige afstanden te verklaren. De correlatie is negatief: -0.24. Dat wil zeggen: hoe groter de producten van de inwoneraantallen, hoe kleiner de taalkundige afstanden. Dit is in overeenstemming met het zwaartekrachtmodel. Immers dit model voorspelt dat dialecten verwanter zullen zijn naarmate het aantal inwoners van de respectieve plaatsen groter is. Hoe meer inwoners, hoe meer de dialecten op elkaar lijken, hoe kleiner de taalkundige afstand. Het model voorspelt dus een negatieve correlatie tussen de producten van de inwoneraantallen en de taalkundige afstanden.

In Tabel 1 zagen we dat Amsterdam, 's-Gravenhage, Utrecht en Maastricht de grootste plaatsen zijn. Wanneer we de vier grootste plaatsen weglaten, verklaren de producten van de inwoneraantallen nog slechts 1% van de variantie in de taalkundige afstanden. De correlatie is nu positief: 0.12. Deze correlatie is nog significant (p=0.02).

We zien dus een negatieve correlatie (-0.24) als de grote steden in de beschouwing betrokken worden, en een positieve correlatie (0.12) als we de grote steden weglaten. De negatieve correlatie mèt de vier grootste steden, en de positieve correlatie zònder de grote steden is opmerkelijk. Een negatieve correlatie betekent dat intensief contact leidt tot minder taalkundige verschillen, waarschijnlijk door attractie: de

[9] Om te bepalen of de ene correlatiecoëfficiënt significant (d.i. niet maar toevallig, maar echt betekenisvol) hoger is dan de andere gebruiken we een speciale toets: de Mantel-toets. Deze wordt uitvoerig uitgelegd in Heeringa 2004 (p. 74/75).

dialectsprekers nemen elementen van elkaar over zodat de dialecten meer op elkaar gaan lijken. Een positieve correlatie betekent dat intensief contact leidt tot differentiatie: dialectsprekers willen zich waarschijnlijk van elkaar blijven onderscheiden en houden onderscheidende elementen daarom vast of voegen die zelfs toe. De beide correlaties – negatief en positief – suggereren dat attractie vooral plaats heeft tussen en ten opzichte van de vier grootste steden, en differentiatie tussen en ten opzichte van de middelgrote steden.

We onderzochten ook of de correlatie met de producten van inwoneraantallen misschien verbetert wanneer alleen paren van plaatsen in de beschouwing betrokken worden die niet te ver uit elkaar liggen. We gebruikten hierbij alle 27 plaatsen. De correlatie bleek inderdaad beter te worden. We vonden een verbetering van -0.24 (alle paren) tot -0.31 (alleen paren met afstanden van kleiner dan 60 kilometer, totaal 45 paren).[10] De verbetering is echter niet significant.

4.3 Toegevoegde waarde van inwoneraantallen

4.3.1 Zwaartekrachtmodel

In Trudgill's zwaartekrachtmodel worden de producten van de inwoneraantallen gedeeld door de gekwadrateerde geografische afstanden. Wanneer we de uitkomsten van deze delingen correleren met de taalkundige afstanden, krijgen we r=-0.22. Dit is geen verbetering ten opzichte van de correlatie met alleen de producten van inwoneraantallen (r=-0.24) of met alleen de gekwadrateerde geografische afstanden (r=0.49). Omdat de correlatie met logaritmische geografische afstanden significant beter is dan de correlatie met gekwadrateerde geografische afstanden, onderzochten we een alternatief zwaartekrachtmodel waarbij gedeeld wordt door de logaritmische geografische afstanden. Dit gaf r=-0.24, dezelfde waarde die we krijgen wanneer we correleren ten opzichte van de producten van de inwoneraantallen afzonderlijk. Deze correlatie is niet sterker dan de correlatie met de logaritmische geografische afstanden afzonderlijk (r=0.62). Deze resultaten suggeren dat het zwaartekrachtmodel in dit opzicht in ons geval niet het juiste model is.

4.3.2 Meervoudige regressieanalyse

Een andere manier om de factoren geografie en inwonersaantalproduct te combineren is meervoudige regressie-analyse, een statistische techniek die het verband tussen variabelen zo nauwkeurig mogelijk in een formule uitdrukt. Het idee daarbij is in ons geval dat de taalkundige afstanden voorspeld kunnen worden op basis van geografische afstanden en inwonersaantalproducten. Omdat taalkundige afstanden dus door twee factoren voorspeld worden, gaat het hier om 'meervoudige' regressieanalysie. In paragraaf 4.1 gebruikten we vier transformaties voor de geografische afstanden:

[10] Als de correlatiecoëfficiënt gelijk is aan 0, is er geen correlatie. Naarmate de correlatiecoëfficiënt verder verwijderd ligt van 0, is de correlatie sterker. Omdat in ons geval de correlatiecoëfficienten negatief zijn, geldt dat de laagste correlatiecoëfficiënt de sterkste correlatie representeert.

kwadratisch, geen transformatie, de wortel en logaritmisch. We hebben daarom vier meervoudige regressie-analyses uitgevoerd, voor elke transformatie één. Dit gaf de volgende resultaten:

transformatie geografie	**correlatie geografie**	**correlatie inw. prod. + geografie**	**verklaarde variantie geografie**	**verklaarde variantie inw. prod + geografie**
kwadratisch	0.49	0.53	24%	29%
geen	0.58	0.60	33%	36%
wortel	0.61	0.63	37%	40%
logaritmisch	0.62	0.65	39%	42%

In alle vier gevallen correleren de door het meervoudige regressiemodel voorspelde waarden iets beter met de taalkundige afstanden dan de geografische afstanden afzonderlijk, maar de verbetering is in geen van de vier gevallen significant. We kunnen hier dus geen bewijs vinden dat het product van inwoneraantallen een verklarende factor vormt voor dialectvariatie.

5 Conclusie

Waarom zijn sommige dialecten sterk verwant aan elkaar, terwijl andere dialecten juist sterk van elkaar verschillen? Waarom lijken de dialecten van Amsterdam en Utrecht relatief veel op elkaar, maar zijn de dialecten van Amsterdam en Maastricht heel verschillend? Dit blijkt vooral bepaald te worden door de geografische ligging van de plaatsen. In het hierboven voorgestelde onderzoek blijkt geografie 33% van de taalkundige variatie van 27 variëteiten in het Nederlandse taalgebied te verklaren. Eenvoudig gezegd: 33% van de variatie in de Nederlandse dialecten is het gevolg van geografie.

Sociaal contact wordt niet alleen bepaald door geografie, maar ook door inwoneraantallen. Tussen plaatsen met veel inwoners zal meer contact bestaan dan tussen plaatsen met maar heel weinig inwoners. Toch bleken inwoneraantallen maar 6% van dialectvariatie te verklaren. Formeler gezegd: de producten van de inwoneraantallen verklaren slechts 6% van de variantie in de taalkundige afstanden.

Wat gebeurt er als we geografie en inwonersaantalproducten combineren? We bekeken daarvoor het zwaartekrachtmodel. Maar zeker in vergelijking met geografie blijkt dit model helemaal geen goede voorspeller te zijn: het verklaart maar 5% van de variantie in de taalkundige afstanden.

Hebben inwoneraantallen dan geen enkele toegevoegde waarde ten opzichte van geografie als verklaring voor dialectvariatie? Om die vraag te kunnen beantwoorden gebruikten we een speciale statistische techniek: meervoudige regressieanalysie. Toepassing van deze techniek maakte duidelijk dat de producten van inwoneraantallen wel een verbetering geven, namelijk van 3%, maar deze verbetering bleek niet significant te zijn.

De taalkundige afstanden die we gebruikten zijn gebaseerd op lexicale, fonetische en morfologische variatie. In verder onderzoek zou het interessant zijn deze taalkundige

niveaus elk afzonderlijk te onderzoeken, en bovendien ook het syntaktische en prosodische niveau te bekijken. Daarbij zouden ook andere dialectgegevens gebruikt kunnen worden met een groter oppervlak en/of een grotere dichtheid. Ook is het misschien zinvol om te zoeken naar alternatieven voor de meting van sociaal contact, bijvoorbeeld metingen van verkeersstromen tussen plaatsen, of de dagelijkse frequentie van de openbaarvervoerverbindingen tussen plaatsen.

In dit onderzoek onderzochten we de rol van geografie en inwoneraantallen in de verklaring van dialectvariatie. Het zou interessant zijn beide factoren ook te onderzoeken voor variatie in de spelling van middeleeuwse documenten. In documenten van steden waartussen veel sociaal contact bestond is wellicht een vergelijkbare spellingstraditie gehanteerd. Kempken (2005) laat zien dat verschillen in spelling kunnen gemeten worden met de Levenshtein-afstand, de afstandsmaat die we in dit artikel uitgebreid besproken hebben.

Behalve variatie in spelling, zou ook onderzocht kunnen worden in welke mate variatie in stijl verklaard kan worden door sociaal contact. Het onderzoek zoals gepresenteerd in dit artikel is dus ook van belang voor het tekstanalytisch onderzoek van bijvoorbeeld middeleeuwse documenten. Naast geografie en inwoneraantallen zouden ook andere factoren zoals historische en politieke verschillen in de beschouwing betrokken kunnen worden.

Bibliografie

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

Goebl 2002 – H. Goebl: 'Analyse dialectométrique des structures de profondeur de l'ALF'. In: *Revue de linguistique Romane*, 66 (2002). Strasbourg: Société de linguistique Romana, p. 1-63.

Goebl 2003 – H. Goebl: 'Regards dialectométriques sur les données de l'*atlas linguistique de la france (ALF)*: relations quantitatives et structures de profondeur'. In: *Estudis Romànics*, XXV (2003). Barcelona: Institut d'estudis Catalans, p. 59-121.

Goebl 2005 – H. Goebl: 'La dialectométrie corrélative: un nouvel outil pour l'etude de l'aménagement dialectal de l'espace par l'homme'. In: *Revue de linguistique Romane*, 69 (2005). Strasbourg: Société de linguistique Romana, p. 321-367.

Gooskens 2004 – Ch. Gooskens: 'Norwegian dialect distances geographically explained'. In: B.-L. Gunnarson, L Bergström, G. Eklund, S. Fridella, L. H. Hansen, A. Karstadt, B. Nordberg, E. Sundgren & M. Thelander (red.): *Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2, June 12-14, 2003*. Uppsala, 2004, p. 195-206.

Gooskens & Heeringa 2004 – Ch. Gooskens & W. Heeringa: 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data'. *Language variation and change*, 16 (2004), p. 189-207. Beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/papers/.

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

Heeringa et al. 2006 – W. Heeringa, P. Kleiweg, Ch. Gooskens & J. Nerbonne: 'Evaluation of String Distance Algorithms for Dialectology'. In: J. Nerbonne & E. Hinrichs (eds.), *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, p. 51-62.

Heeringa & Nerbonne 2001 – W. Heeringa & J. Nerbonne: 'Dialect areas and dialect continua'. In: *Language Variation and Change*, 13 (2001), p. 375-400. Beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/papers/.

Hinskens 1992 – F. Hinskens: *Dialect levelling in Limburg. Structural and sociolinguistic aspects*. Proefschrift Katholieke Universiteit Nijmegen, Nijmegen, 1992.
Hinskens 1993 – F. Hinskens: 'Dialectnivellering en regiolectvorming'. In: *Taal en Tongval*, 6 (1993), p. 40-61.
Kempken 2005 – S. Kempken: *Bewertung Historischer und Regionaler Schreibvarianten mit Hilfe von Abstandsmaßen*. Doctoraalscriptie Universiteit Duisburg-Essen, Duisburg, 2005.
Kessler 1995 – B. Kessler: 'Computational dialectology in Irish Gaelic'. In: *Proceedings of the 7th conference of the European chapter of the association for computational linguistics*. Dublin, 1995, p. 60-67.
Nerbonne et al. 1996 – J. Nerbonne, W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten & W. van de Vis: 'Phonetic distance between Dutch dialects'. In: G. Durieux, W. Daelemans & S. Gillis (red.): *CLIN VI, Papers from the sixth CLIN meeting*. Antwerpen, 1996, p. 185-202. Beschikbaar via: http://www.let.rug.nl/~heeringa/dialectology/papers/.
Nerbonne & Heeringa 2006 – J. Nerbonne & W. Heeringa (2006): 'Geographic distributions of linguistic variation reflect dynamics of differentiation'. In: S. Featherstone en W. Sternefeld (red.), *Linguistic Evidence*. 2006. Geaccepteerd.
Nerbonne & Kleiweg 2003 - J. Nerbonne & 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 (2003), p. 339-357. Beschikbaar via: http://www.let.rug.nl/~nerbonne/paper.html.
Nerbonne & Kleiweg 2006 – J. Nerbonne & P. Kleiweg (2006): 'Toward a dialectological yardstick'. In: *Quantitative Linguistics*, 13 (2006). Geaccepteerd.
Nerbonne & Siedle 2005 - J. Nerbonne & C. Siedle: 'Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede'. In: *Zeitschrift für Dialektologie und Linguistik*, 72 (2005), p. 129-147. Beschikbaar via: http://www.let.rug.nl/~nerbonne/paper.html.
Séguy 1971 – J. Séguy: 'La relation entre la distance spatiale et la distance lexicale'. In: *Revue de Linguistique Romane*, 35 (1971), p. 335-357.
Trudgill 1974 – P. Trudgill: 'Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography'. In: *Language in Society*, 2 (1974), p. 215-246.
Trudgill 1983 – P. Trudgill: *On Dialect. Social and Geographical Perspectives*. Oxford: Basil Blackwell, 1983.
Van Gemert 2002 – I. van Gemert: *Het geografisch verklaren van dialectafstanden met een geografisch informatiesysteem (GIS).* Afstudeerscriptie Rijksuniversiteit Groningen, Groningen, 2002. Beschikbaar via: http://www.let.rug.nl/~alfa/scripties.html.
Wikipedia-bijdragers 2006 - Wikipedia-bijdragers: 'Zwaartekracht'. In: *Wikipedia, de vrije encyclopedie.* Opgehaald op 25 juli 2006. Beschikbaar via: http://nl.wikipedia.org/wiki/Zwaartekracht.
Wikipedia-bijdragers 2006a - Wikipedia-bijdragers: 'Graviteitsmodel in de geografie'. In: *Wikipedia, de vrije encyclopedie.* Opgehaald op 25 juli 2006. Beschikbaar via: http://nl.wikipedia.org/wiki/Graviteitsmodel_in_de_geografie.
Zipf 1946 – G. K. Zipf: 'The P_1P_2/D Hypothesis: On the Intercity Movement of Persons'. In: *American Sociological Review*, 11 (1946), p. 677-686.

Adressen van de auteurs

Wilbert Heeringa, Rijksuniversiteit Groningen, vakgroep Alfa-Informatica, postbus 716, 9700 AS Groningen, w.j.heeringa@rug.nl
John Nerbonne, Rijksuniversiteit Groningen, vakgroep Alfa-Informatica, postbus 716, 9700 AS Groningen, j.nerbonne@rug.nl
Renée van Bezooijen, Radboud Universiteit Nijmegen, vakgroep taalwetenschap, postbus 9103, 6500 HD Nijmegen, r.v.bezooijen@let.ru.nl
Marco René Spruit, Meertens-Instituut, postbus 94264, 1090 GG, Amsterdam, marco.rene.spruit@meertens.knaw.nl