De analyse van uitspraakverschillen in Nederlandse en Friese taalvariëteiten

Wilbert Heeringa, John Nerbonne, Peter Kleiweg

1. Inleiding

Schibbolets verraden de geografische herkomst van dialectsprekers. Iemand die de "slot-n" in een woord als *zitten* uitspreekt, komt vermoedelijk uit het noordoosten van Nederland of het noordwesten van België. Iemand die de "zachte g" uitspreekt in een woord als *achtentachtig* woont waarschijnlijk ergens ten zuiden van de grote rivieren. Lange tijd werden dialecten gekarakteriseerd door dergelijke eigenschappen, vaak ook (zoals de beide voorbeelden laten zien) door eigenschappen die een verschil representeren ten opzichte van het Standaardnederlands. Deze eigenschappen werden ook gevisualiseerd door een lijn op een kaart. Aan de ene kant van de lijn lag dan bijvoorbeeld het gebied waar *zitten* wordt uitgesproken als *zitt(e)n* (met slot-n), en aan de andere kant het gebied waar hetzelfde woord als *zitte* (zonder slot-n) wordt uitgesproken. Zo'n lijn heet een *isoglosse.*

Echter in recent computationeel onderzoek wordt niet zozeer gefocust op een beperkte verzameling van selectief – en daarmee subjectief – gekozen losse taalkundige verschijnselen die weergegeven worden als isoglossen, maar veel meer op de aggregatie van verschijnselen: een *groot* aantal aselect gekozen verschijnselen – honderden en soms wel duizenden - worden samengenomen oftewel *geaggregeerd*, en door analyse van het aggregaat wordt een variatiepatroon gevonden (zie Nerbonne, te verschijnen).

In deze bijdrage gaan we eerst taalkundige afstanden meten tussen dialecten op basis van een aantal bekende isoglossen, dat wil zeggen isoglossen waar de de dialectsprekers zelf zich bewust van zijn en die in de literatuur veelvoudig worden genoemd. Door het meten van afstanden wordt meteen duidelijk dat bijvoorbeeld het Groningse dialect van Roodeschool meer verwant is met het Drentse dialect van Beilen dan met de Friese taalvariëteit zoals gesproken in Grouw. Op basis van die afstanden maken we een indeling in dialectgroepen. Vervolgens gaan we uitspraakafstanden meten op basis van de aggregatie van een groot aantal verschijnselen. De verschijnselen in de aggregaat zijn min of meer willekeurig gekozen en hebben daarmee het karakter van een steekproef. De gemeten afstanden zijn dus objectief. Ook op basis van deze afstanden maken we een indeling in dialectgroepen. Vervolgens gaan we beide indelingen met elkaar vergelijken. Is er een groot verschil tussen de isoglossenindeling en de geaggregeerde indeling? Als dat zo is, moeten we concluderen dat een indeling op basis van een beperkte verzameling van selectief gekozen isoglossen geen objectief beeld geeft, ook al gaat het om bekende isoglossen. De conclusie is dan dat de aggregatie van een groot aantal verschijnselen noodzakelijk is, wil men een eerlijk beeld geven.

Dialectvariatie, dialectafstanden en dialectindelingen zijn geografisch gerelateerd. Om te komen tot de beantwoording van onze onderzoeksvraag maken we daarom gebruik van een GIS-pakket: R*u*G/L04. Dit pakket is vrij beschikbaar op het wereldwijde web en wordt internationaal door meerdere onderzoeksgroepen met succes toegepast. Het pakket kan gratis geladen worden via http://www.let.rug.nl/~kleiweg/L04/.

In paragraaf 2 bespreken we eerst de gegevensbron die de basis voor alle metingen en resultaten in deze bijdrage vormt. Vervolgens bespreken we in paragraaf 3 de isoglossen, en de manier waarop op basis van isoglossen afstanden tussen variëteiten gemeten kunnen worden. Paragraaf 4 bespreekt hoe geaggregeerde afstanden met de Levenshtein-afstand gemeten kunnen worden. In paragraaf 5 bespreken we twee classificatietechnieken. In paragraaf 6 vergelijken we de resultaten op basis van isoglossen met de geaggregeerde afstanden, en bepalen de representativiteit van een negental

isoglossen. In paragraaf 7 eindigen we met enkele conclusies.

2. Gegevensbron

De gegevensbron die we in deze bijdrage gebruiken 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. In totaal hebben 16 veldwerkers het materiaal in de atlassen verzameld. Soms waren meerdere veldwerkers bij een deel betrokken, en sommige veldwerkers hebben aan meerdere delen meegewerkt.

2.1 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 Oudgermaanse 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,[i] maar de mogelijkheid om dit te verwerken zullen we in deze bijdrage verder niet bespreken.

2.2 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 1.

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 1 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 1 aangegeven met grijze ruitjes.

Figuur 1. 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 een van de twee dialecten als een taaleiland. Het betreft Tjalleberd (meest linkse ruitje), Donkerbroek (middelste ruitje) en Appelscha (meest rechtse ruitje). De kleine cirkeltjes representeren Urk (rechtsboven) en Monnickenwerf op Marken (linksonder).

3. Isoglossen

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. In de inleiding noemden we als de slot-n-isoglosse die het noordoosten en het zuidwesten van het Nederlandse taalgebied (met de uitspraak *zitt(e)n*) afgrenst van de rest van het Nederlandse taalgebied (uitspraak *zitte*). Bij het gebruik van isoglossen zijn er twee vragen waarop niet gemakkelijk een antwoord gegeven kan worden, namelijk: 1) welke isoglossen kiezen we? en 2) hoe belangrijk is elk van de isoglossen?

3.1 Keuze van isoglossen

In 1941 verscheen een isoglossenkaart van Weijnen Op deze kaart zijn 45 isoglossen getekend. In tegenstelling tot oudere 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.

3.2 Het belang van elk van de isoglossen

De vraag welke isoglossen het meest belangrijk zijn is niet eenvoudig te beantwoorden. Dialectologen hebben een voorkeur voor isoglossen die geheel of gedeeltelijk samenvallen zodat bij het op elkaar stapelen van isoglossenkaarten isoglossenbundels ontstaan, en daardoor een heldere verdeling in gebieden (zie bijvoorbeeld de isoglossenkaart achterin *Nederlandse Dialectkunde* van Weijnen (1966)). In *Van randstad tot landrand* schreef Jo Daan onder andere:

> "De taalkundige kan alle isoglossen vaststellen en hun dichtheid en richting volgens statistische methoden vergelijken. Maar in zijn objectiviteit zal het hem ontgaan welke isoglossen belangrijk, welke onbelangrijk zijn, hij krijgt geen inzicht in de relevantie van de isoglossen als dialektscheiding." (Daan & Blok, 1969, blz. 9).

De auteur onderscheidt echter twee hoofdkenmerken: "het al of niet uitspreken van de slot-n en de *jij*/*gij*-grens." (blz. 32). Ze merkt daarbij op dat het gebied waar dialectsprekers *gij* gebruiken, ongeveer overeenkomt met het gebied waar de zachte *g* wordt uitgesproken.

In ons RND-materiaal is de slot-n-isoglosse sterk vertegenwoordigd, namelijk in maar liefst 23 woorden. In Figuur 2a wordt de geografische verdeling van *dope* versus *doopm* of *dopen* weergegeven. Het *jij*/*gij*-onderscheid vinden we – uiteraard – in precies één woord. De verdeling is te vinden in Figuur 2g. Op de kaart zien we ook een derde vorm, namelijk *ie*. *Gij*, *jij* en *ie* zijn uitspraakvarianten van elkaar. In de grijze gebieden op de kaart vinden we geen uitspraakvarianten van de drie genoemde vormen. In deze gebieden zegt men meestal *doe* of *doo* of *dou*. Zoals gezegd valt de *jij*/*gij*-grens ongeveer samen met de grens tussen de harde *g* en de zachte *g*. In het RND-materiaal kan dit onderscheid helaas niet worden teruggevonden, omdat de veldwerkers bij het noteren van de uitspraak geen onderscheiden tekens gebruiken voor de beide *g*'s.

Naast deze twee hoofdisoglossen hebben we in Figuur 2 nog zeven andere isoglossen opgenomen. Omdat we in deze bijdrage onderzoek doen naar uitspraakvariatie, zijn het uitsluitend isoglossen die de uitspraak betreffen. De gekozen woorden die de isoglossen representeren, variëren lexicaal niet of nauwelijks. We hebben de isoglossen gekozen op basis van het RND-materiaal, en dan met name die isoglossen die een tamelijk duidelijke verdeling in gebieden suggereren. Zoals we in paragraaf 2 schreven, hebben meerdere veldwerkers meegewerkt. Die noteerden de uitspraak van woorden niet altijd op precies dezelfde manier. De isoglossen in Figuur 2 betreffen taalverschijnselen die (vrijwel) volledig resistent zijn tegen veldwerkersverschillen.

De isoglossen op de kaarten b), c) en d) zijn ook te vinden op de uitvouwbare kaart achterin *Nederlandse Dialectkunde* van Weijnen (1958, 1966). Weijnen geeft ook een kaart die vrijwel hetzelfde patroon heeft als kaart f. Op die kaart wordt de uitspraakvariatie in de klinker in *lief* weergegeven. Langs de oostgrens - het gebied waar *goed* als *goud* of *good* wordt uitgesproken – is de uitspraak *leef*, *lijf* of *laif*. Kaart e. vinden we, zij het in iets andere vorm, ook in *Language*, een inleiding geschreven door Bloomfield (uitgave 1933 en 1935 of recenter). De kaart in *Language* is afgeleid van een kaart van G. G. Kloeke waarop niet alleen de uitspraak van de klinker in *huis*, maar ook die in *muis* wordt weergegeven. Op kaart i) vinden we de bekende Uerdinger lijn, een isoglosse die dialecten die de *ich*-vorm gebruiken scheidt van dialecten die de *ik*-vorm gebruiken.

3.3 Geaggregeerde isoglossenafstanden

Door het samennemen van de isloglossen kunnen we geaggregeerde afstanden meten. Het meten van geaggregeerde taalkundige afstanden tussen dialecten werd voor het eerst gedaan door Jean Séguy in het begin van de zeventiger jaren van de twintigste eeuw (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).

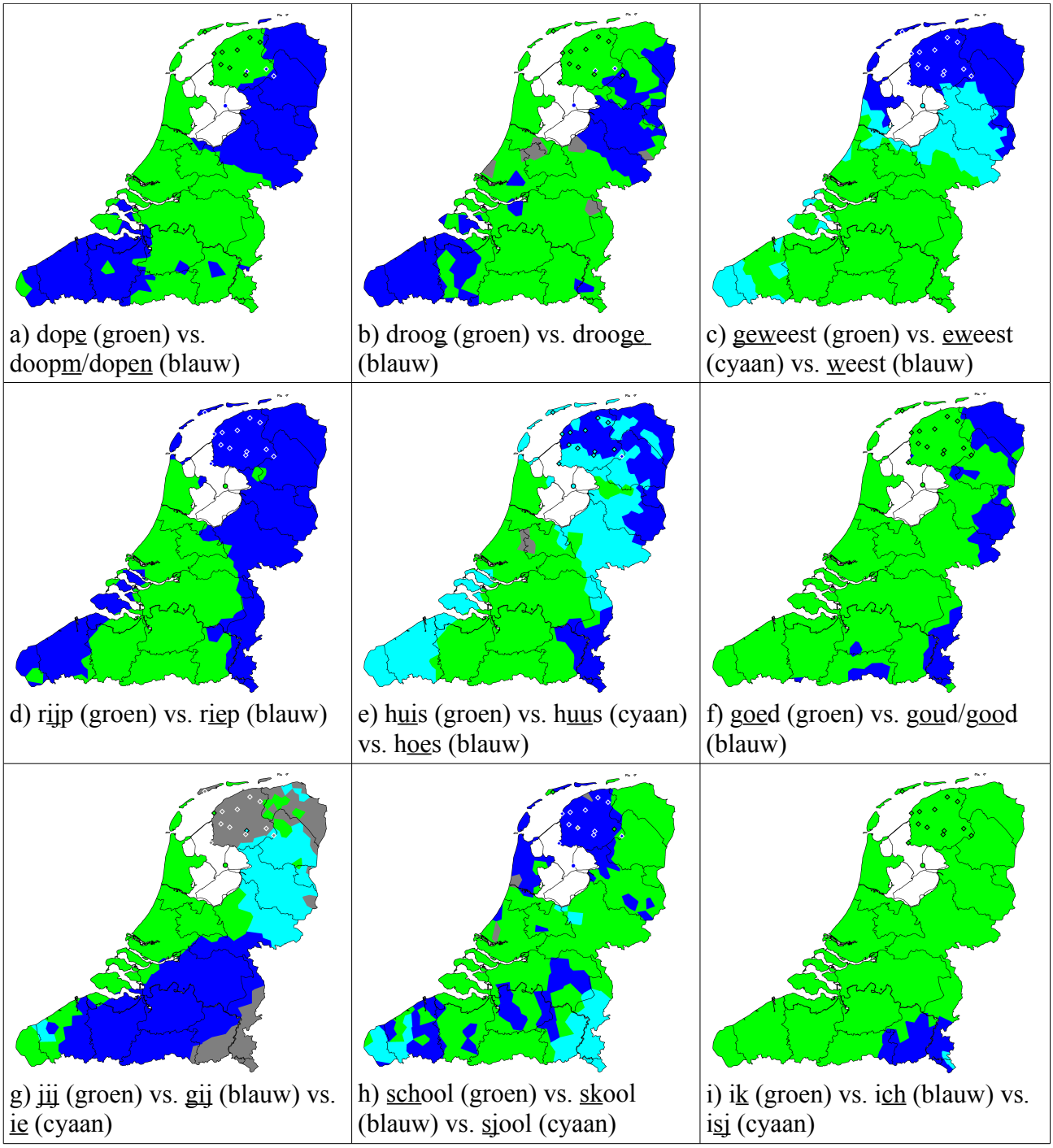
We kunnen de methodologie van Séguy gemakkelijk toepassen op onze data. Het berekenen van bijvoorbeeld de afstand tussen het Fries zoals gesproken in Grouw en het dialect van Haarlem op basis van onze negen isoglossen gaat als volgt:

isoglosse	**Grouw**	**Haarlem**	**afstand**
a)	dope	dope	0
b)	droog	droog	0
c)	weest	geweest	1
d)	riep	rijp	1
e)	hoes	huis	1
f)	goed	goed	0
g)	jij	?	?
h)	skool	school	1
i)	ik	ik	0
			4

Voor isoglosse g) kunnen we niet bepalen of Grouw en Haarlem hetzelfde zijn. In Grouw wordt geen uitspraakvariant van *gij*, *jij* of *ie* gebruikt. We delen nu het aantal isoglossen waardoor Grouw en Haarlem worden gescheiden door het totale aantal isoglossen dat we konden verdisconteren: 4 gedeeld door 8 is 0.5. Vermenigvuldigen we dit met 100, dan krijgen we een afstand van 50%.

We hebben afstanden gemeten voor alle plaats-paren. Het meten van de afstanden is goed mogelijk met het GIS-pakket R*u*G/L04. De basis is een database van 360 dialectplaatsen, waarbij voor elke plaats negen isoglossenwaarden gegeven zijn. Op basis daarvan meet het systeem de afstanden tussen de dialectplaatsen. Hier zien we meteen het unieke van R*u*G/L04 waardoor het zich onderscheidt van andere GIS-pakketten. De mogelijkheden van de standaard GIS-programma's zoals bijvoorbeeld ArcInfo en Grass beperken zich tot de analyse van gegevens die steeds betrekking hebben op *één bepaalde plaats*, bijvoorbeeld de bodemvervuiling door koper, de aanwezigheid van water, enz. We spreken dan van monadische variabelen. In de dialectologie is men echter vooral ook geïnteresseerd in de analyse van dyadische variabelen, dus de relaties *tussen plaatsen*, in ons geval het aantal isoglossen waarin twee plaatsen van elkaar verschillen. Behalve in de dialectologie is de analyse van dyadische variabelen ook gangbaar in de ethnologie en de musicologie. R*u*G/L04 is een GIS-pakket dat speciaal hiervoor ontworpen is.

Vervolgens biedt R*u*G/L04 de mogelijkheid om de afstanden weer te geven op een kaart. Deze kaart is gegeven in Figuur 3a. Hoe donkerder de lijnen, hoe kleiner de afstanden zijn. Er zijn alleen afstanden tussen plaatsen weergegeven die niet verder uit elkaar liggen dan 100 km. De verticale lijn rechtsonder in de figuur geeft deze afstanden weer. Op deze manier ontstaat een beeld dat een verdeling in verschillende gebieden suggereert.



a) dope (groen) vs. doopm/dopen (blauw)

b) droog (groen) vs. drooge (blauw)

c) geweest (groen) vs. eweest (cyaan) vs. weest (blauw)

d) rijp (groen) vs. riep (blauw)

e) huis (groen) vs. huus (cyaan) vs. hoes (blauw)

f) goed (groen) vs. goud/good (blauw)

g) jij (groen) vs. gij (blauw) vs. ie (cyaan)

h) school (groen) vs. skool (blauw) vs. sjool (cyaan)

i) ik (groen) vs. ich (blauw) vs. isj (cyaan)

Figuur 2. Negen isoglossenkaarten getekend op basis van gegevens uit de RND voor 360 variëteiten. Plaatsen waarvoor geen gegevens beschikbaar waren zijn in de kaarten grijs gekleurd. Elke isoglosse suggereert weer een andere indeling.

4 Meting van uitspraakverschillen met de Levenshtein-afstand

In paragraaf 3 eindigden we met de metingen van geaggregeerde isoglossenafstanden zoals dit begin zeventiger jaren van de twintigste eeuw door Jean Séguy werd geïntroduceerd. De negen isoglossen hadden we handmatig in het ons materiaal opgezocht en gecodeerd. In paragraaf 2 zagen we dat ons materiaal bestaat uit 125 woorduitspraaktranscripties voor elk van 360 variëteiten. In dit materiaal zitten *honderden* isloglossen verborgen, en met een handmatige analyse kunnen eventueel al die isloglossen opgezocht en gecodeerd kunnen worden.Het behoeft geen betoog dat dit een bijzonder tijdrovende klus zou worden. Dé oplossing bij uitstek om dit te automatiseren is de Levenshtein-afstand. Met de Levenshtein-afstand worden geaggregeerde afstanden gemeten waarin *alle* informatie die in de transcripties aanwezig is, verwerkt is.

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 symbolen 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. Bij vergelijking van taal- en dialectvariëteiten worden fonetische transcripties van woorduitspraken met elkaar vergeleken. 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.

4.1 De Levenshtein-afstand

Zoals we hierboven al meldden, introduceren we in deze paragraaf een simpele versie van de Levenshtein-afstand om het principe beter te kunnen uitleggen. We gaan er daarbij gemakshalve van uit dat klanken óf hetzelfde óf 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 Fries zoals gesproken in de plaats Grouw wordt *melk* uitgesproken als [mɔlkə]. In het dialect van Haarlem wordt hetzelfde woord uitgesproken als [mɛlək].[ii] De ene uitspraak zou je kunnen veranderen in de andere op de volgende manier:

mɔlkə	verwijder ə	1
mɔlk	vervang ɔ door ɛ	1
mɛlk	voeg toe ə	1
mɛlək		
		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.

4.2 Oplijning

De procedure zoals we die hierboven bespraken is tamelijk abstract. Het is daarom goed die ook te beschrijven vanuit een andere gezichtshoek, namelijk die van de oplijning (Engels: alignment). 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 komt de oplijning er zo uit te zien:

	1	2	3	4	5	6
Grouw	m	ɔ	l		k	ə
Haarlem	m	ɛ	l	ə	k	
		1		1	1	

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.

4.3 Graduele gewichten

In de twee vorige paragrafen introduceerden we een simpele versie van de Levenshtein-afstand waarbij klanken óf gelijk óf ongelijk aan elkaar zijn. Voor de berekeningen verderop in deze bijdrage 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).

4.4 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 illustreren dit aan de hand van een voorbeeld. In dat voorbeeld berekenen we de afstand tussen Grouw en Haarlem op basis van zes woorden. De berekening ziet er dan als volgt uit:[iii]

item	Grouw	Haarlem	Levenshtein-afstand
drinken	drɪŋkə	drɪŋkə	0
geroepen	rʊpm	ɣərupə	4
ladder	ljɛdər	lɑdər	2
melk	mɔlkə	mɛlək	3
schip	skɪp	sxɪp	1
werk	ʊʏrk	ʊɛrək	2
			12

De laatste kolom geeft Levenshtein-afstanden. Deze Levenshtein-afstanden *aggregeren* we. De afstand tussen Grouw en Haarlem wordt nu gelijk aan 12 Levenshtein-operaties. Op deze manier meten we afstanden tussen alle plaats-paren.

In paragraaf 3.3 lieten we zien dat met R*u*G/L04 geaggregeerde isoglosenafstanden gemeten kunnen worden. Het GIS-pakket biedt daarnaast de mogelijkheid om geaggregeerde Levenshein-afstanden te meten. Ook deze afstanden kunnen eenvoudig op een kaart worden weergegeven. De kaart is te vinden in Figuur 4a. Hoe donkerder de lijnen, hoe kleiner de afstanden zijn. Net als in paragraaf 3 zijn alleen afstanden tussen plaatsen weergegeven die niet verder uit elkaar liggen dan 100 km.

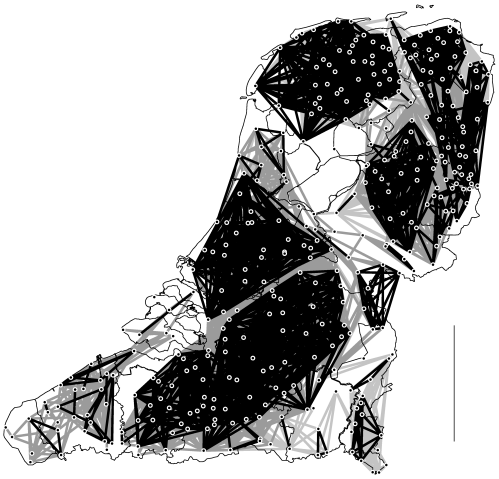
Vergelijken we nu de kaart met de geaggregeerde isoglossenafstanden (Figuur 3a) met de kaart waarin de Levenshtein-afstanden zijn weergeven (Figuur 4a), dan zien we belangrijke verschillen. In beide kaarten zien we bijvoorbeeld in het noordoosten twee groepen, maar in Figuur 3a strekt de noordelijke groep zich verder langs de Nederlandse/Duitse grens zuidwaarts uit. Verder zien we in Figuur 3a een scherpe grens tussen enerzijds de Hollandse dialecten (Noord- en Zuid-Holland en Utrecht) en anderzijds de Brabantse (Noord-Brabant, Antwerpen, Vlaams-Brabant) en Oost-Vlaamse dialecten. In Figuur 3b vinden we deze grens zó niet terug. Duidelijk komt dus uit de verf dat de isoglossenmetingen tot een ander beeld leiden dat de Levenshtein-afstandsmetingen.

5. De classificatie van de variëteiten

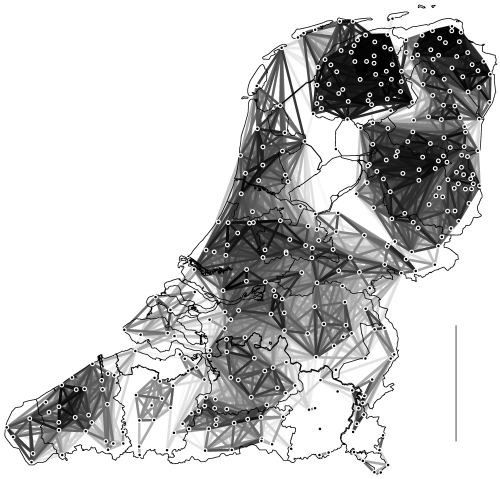
5.1 Clusteranalyse

In de paragrafen 3 en 4 bespraken we hoe we afstanden kunnen meten tussen variëteiten. Omdat we 360 dialecten hebben (zie paragraaf 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). Zo'n boom noemen we een dendrogram. De takken in het dendrogram representeren de afstanden tussen dialecten en clusters. De afstanden tussen de dialecten (de bladeren) zoals gesuggereerd door het dendrogram, worden cofenetische afstanden genoemd.

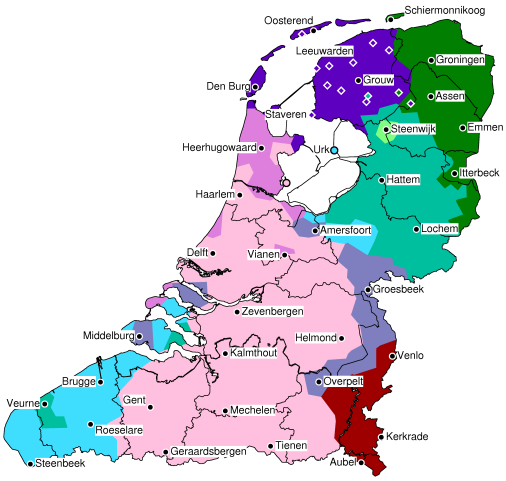
Er bestaan verschillende clustermethoden. In R*u*G/L04 kunnen we kiezen uit maar liefst zeven verschillende methoden. Wij gebruikten UPGMA (Unweighted Pair Group Method using Arithmetic Averages). Het bleek dat de cofenetische afstanden in de boom die met déze methode gemaakt werd, de originele afstanden – dus de afstanden tussen de 360 dialecten op basis waarvan de clusteranalyse werd uitgevoerd – het meest nauwkeurig weerspiegelen (zie Heeringa 2004, blz. 150-153).



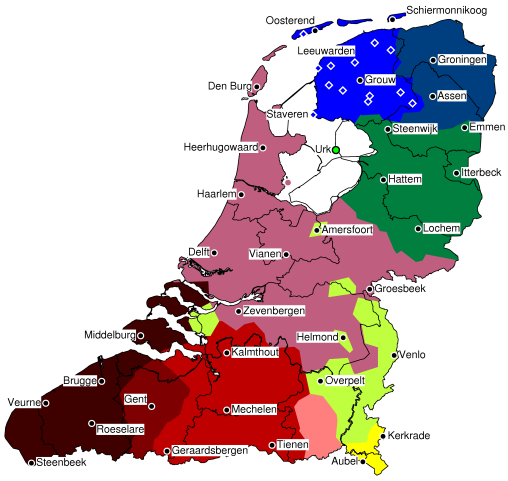
a. Isoglossenafstanden.



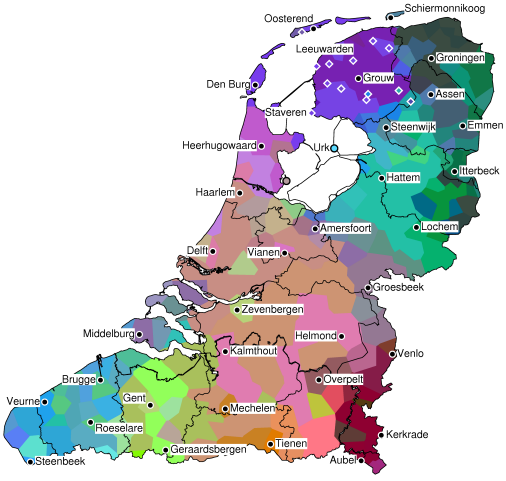
a. Levenshtein-afstanden.



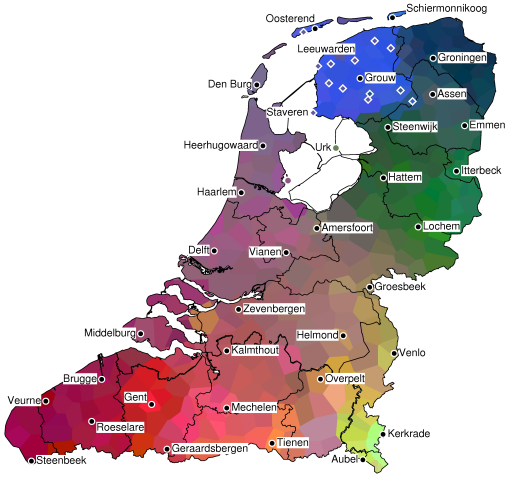
b. De negen natuurlijke groepen op basis van isoglossenafstanden.



b. De elf natuurlijke groepen op basis van Levenshtein-afstanden.



c. Dialectcontinuüm op basis van isoglossenafstanden.



c. Dialectcontinuüm op basis van Levenshtein-afstanden.

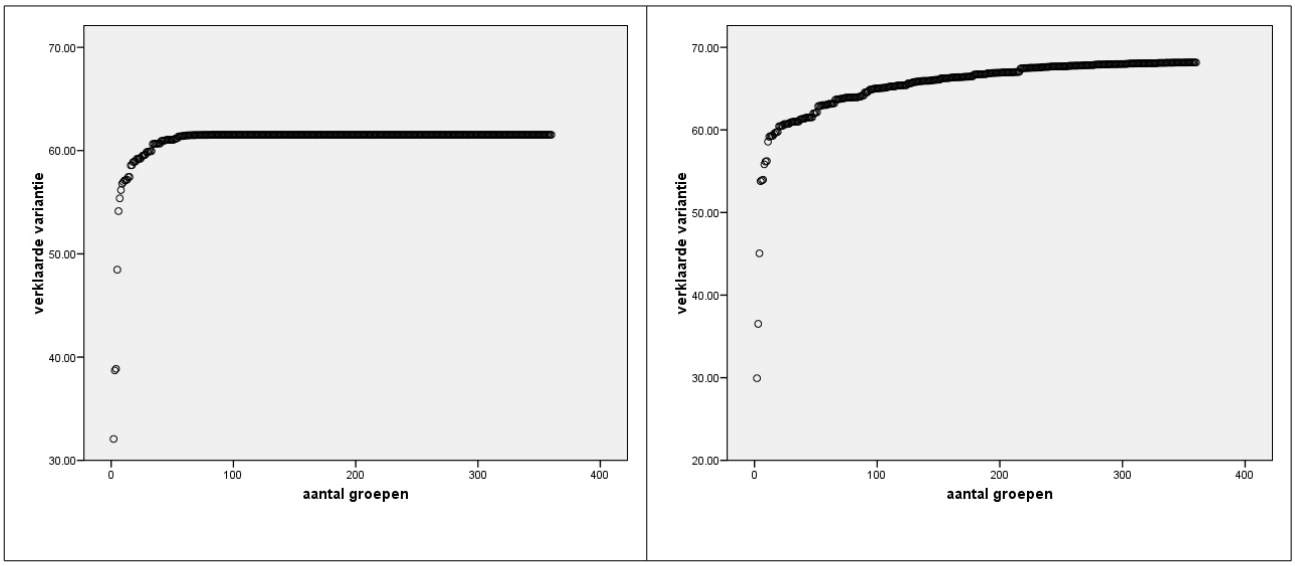
Figuur 3. De afstanden tussen variëteiten zijn gemeten als het aantal isoglossen waarin de variëteiten overeenkomen (a.). Op basis van deze afstanden is clusteranalyse (b.) en multidimensionale schaling (c.) uitgevoerd.

Figuur 4. De Levenshtein-afstanden tussen de variëteiten (a.). Ook op basis van deze afstanden is clusteranalyse (b.) en multidimensionale schaling (c.) uitgevoerd. De verschillen met de kaarten in Figuur 3 zijn evident.

Dendrogrammen zijn binaire bomen. Binnen een dendrogram kunnen verschillende niveaus van gedetailleerdheid worden onderscheiden. Beperken we ons tot de wortel, dan krijgen we een verdeling in twee groepen. Gaan we echter iets dieper in de boom, dan blijkt een van de twee groepen op zijn beurt ook weer uit twee groepen te bestaan. Op dat diepteniveau vinden we dus een verdeling in drie groepen. Helemaal onderin de boom vinden we de bladeren. Hier vinden we een verdeling met het grootste aantal groepen, in ons geval 360 groepen met elk één variëteit. In de boom hebben we dus 359 niveaus, van 2 groepen t/m 360 groepen. Voor elk niveau berekenen we nu de variantie in de oorspronkelijke afstanden die door de cofenetische afstanden van de boomstructuur tot op dat niveau verklaard wordt. De variantiewaarden worden vervolgens in een grafiek uitgezet tegen het aantal groepen dat door elk van de 359 niveaus onderscheiden wordt. In zo'n grafiek zien we meestal eerst een snelle stijging van de variantie, gevolgd door een knikpunt – de elleboog – en daarna stijgt de grafiek niet of nauwelijks meer.

In figuur 5 vinden we een elleboogcurve op basis van zowel de isoglossenafstanden als de Levenshtein-afstanden. In de elleboogcurve op basis van de isoglossenafstanden vinden we een knikpunt – de elleboog - bij negen groepen. Na 9 groepen stijgt de curve niet zo sterk meer. We concluderen dat er 9 natuurlijke groepen zijn. De groepen zijn weergegeven in Figuur 3b. De cofenetische afstanden in het dendrogram tot het niveau van 9 groepen verklaren 57% van de variantie in de oorspronkelijke isoglossenafstanden.

Wanneer we kijken naar de elleboogcurve op basis van Levenshtein-afstanden, dan vinden we het knikpunt bij 11 groepen. Na 11 groepen is er geen sterke stijging meer. Op basis van de Levenshtein-afstanden kunnen we dus 11 natuurlijke groepen onderscheiden. Op de kaart in Figuur 4b worden de 11 groepen weergegeven. De cofenetische afstanden in het dendrogram tot het niveau van 11 groepen verklaren 59% van de variantie in de oorspronkelijke gemiddelde Levenshtein-afstanden.



Figuur 5. Links de elleboogcurve op basis van de isoglossenafstanden. Na 9 groepen stijgt er grafiek niet zo sterk meer. Rechts de elleboogcurve op basis van Levenshtein-afstanden. Na 11 groepen is er geen sterke stijging meer.

Een andere manier om te bepalen wat de belangrijke clusters zijn, is te letten op stabiliteit van clusters. De stabiliteit van clusters kan bepaald worden door middel van *clusteranalyse met ruis* of door middel van *bootstrapclusteranalyse*. Bij clusteranalyse met ruis worden de afstanden tussen de variëteiten met een willekeurige waarde verhoogd. Deze waarde ligt tussen 0 en *max*, waarbij *max* gewoonlijk een halve standaarddeviatie is. De individuele ruiswaarden zijn uniform verdeeld tussen

0 en max. Op basis van deze nieuwe afstanden worden de variëteiten geclusterd. Herhalen we dit bijvoorbeeld honderd keer, dan krijgen we honderd dendrogrammen. Hoe vaker nu een cluster in de honderd dendrogrammen voorkomt, hoe stabieler dat cluster is (zie Kleiweg, Nerbonne & Bosveld, 2004).

In ons onderzoek gebruiken we 125 woorden per variëteit, zodat twee variëteiten vergeleken worden op basis van 125 woordparen. Bij gebruik van bootstrapclusteranalyse worden aselect 125 woorden gekozen, waarbij hetzelfde woord meerdere keren gekozen kan worden en andere woorden niet gekozen worden. Op basis van deze nieuwe selectie van woorden worden afstanden gemeten en clusteranalyse uitgevoerd. Ook dit herhalen we bijvoorbeeld honderd keer, zodat we honderd dendrogrammen krijgen. Clusters die in (bijna) alle dendrogrammen terugkeren, zijn stabiele clusters. Nerbonne, Kleiweg, Heeringa & Manni (2008) hebben aangetoond dat het effect van bootstrapclusteranalyse hetzelfde is als van clusteranalyse met ruis.

In R*u*G/L04 laten de resultaten van clusteranalyse zich gemakkelijk omzetten naar cartografische weergave's. Figuur 3b laat de clusterindeling op basis van geaggregeerde isoglosenafstanden zien, en Figuur 4b toont de indeling op basis van de geaggregeerde Levensthein-afstanden. Vergelijken we nu beide figuren, dan vinden we belangrijke verschillen. Net als bij de vergelijking van Figuur 3a met Figuur 4a in paragraaf 4.4 vinden we ook hier een verschil met betrekking tot het verloop van de grens die de twee groepen in het noordoosten van elkaar scheidt. Verder zijn het Oost-Vlaams en het Antwerps/Vlaams-Brabants in Figuur 4b in Figuur 4a niet meer als aparten groepen onderscheiden. Anderzijds vinden in in Figuur 4a het Westfries (in de kop van Noord-Holland) als aparte groep, wat in Figuur 4b niet het geval is.

5.2 Multidimensionale schaling

Met multidimensionale schaling (MDS, zie Heeringa 2004, blz. 156-163) kunnen de variëteiten worden geplaatst in een driedimensionele ruimte. Hoe meer twee variëteiten op elkaar lijken, hoe dichter ze bij elkaar worden geplaatst. De coördinaten in de driedimensionele ruimte kunnen worden vertaald naar kleuren, bijvoorbeeld de x-coördinaat bepaalt de intensiteit van rood, de y-coördinaat de intensiteit van groen en de z-coördinaat de intensiteit van blauw. Zo krijgt iedere variëteit z'n eigen unieke kleur. Variëteiten die veel op elkaar lijken, krijgen bijna dezelfde kleur, maar variëteiten die heel verschillend zijn, krijgen ook heel verschillende, sterk ten opzichte van elkaar contrasterende kleuren.

Zowel op basis van de isoglossenafstanden als op basis van de Levenshtein-afstanden hebben we een driedimensionele MDS toegepast. MDS is standaard opgenomen in R*u*G/L04. Op basis van een MDS-analyse biedt R*u*G/L04 de mogelijkheid om kleurenkaarten te maken met gebruikmaking van de techniek die we zojuist beschreven. Ieder dialectpunt krijgt z'n eigen unieke kleur, en de ruimtes tussen dialectpunten worden ingekleurd door te interpoleren vanuit de dialectpunten. Halverwege een rood en een geel dialectpunt zal de kaart dus oranje zijn.

De kleurenkaarten op basis van de MDS-uitvoer zijn te vinden in respectievelijk Figuur 3c en Figuur 4c. Omdat we het aantal dimensies in de data (d.w.z. de afstanden tussen de variëteiten) met de MDS-techniek reduceren tot drie, is het waarschijnlijk dat bepaalde details verloren gaan. Om een idee te krijgen van de mate waarin dat het geval is, berekenen we hoeveel variantie in de oorspronkelijke afstanden verklaard wordt door de drie MDS-dimensies. Voor de isoglossenafstanden blijkt dit 85% te zijn en voor de gemiddelde Levenshtein-afstanden is dit 90%.

Clusteranalyse en MDS zijn twee verschillende technieken die elkaar goed aanvullen. Vergelijken we de clusteranalysekaarten (Figuren 3b en 4b) met de MDS-kaarten (Figuren 3c en 4c), dan zien

we dat de clusteranalysekaarten een helderder beeld geven: er is een duidelijke verdeling in gebieden. De MDS-kaarten daarentegen doen meer recht aan het feit dat een dialectlandschap een continuüm vormt, met meer of minder geleidelijke overgangen. Daar waar MDS-kaarten een vrij abrupte overgang laten zien, zal in de clusteranalysekaarten een grens tussen twee gebieden gevonden kunnen worden.

6. De representativiteit van isoglossen

In de inleiding stelden we de vraag inhoeverre een indeling op basis van isloglossen zou verschillen van een indeling op basis van de aggregaat van een steekproef van heel veel taalkundige verschijnselen. In paragraaf 3 introduceerden we negen isoglossen, en maten afstanden tussen variëtieten op basis van deze isoglossen. In paragraaf 4 berekenden we geaggregeerde Levenshtein-afstanden. Daarmee worden een heel groot aantal taalkundige verschijnselen voor wat betreft de uitspraak van woorden, verdisconteerd, niet alleen verschijnselen die een tamelijk heldere verdeling in gebieden geven (zoals de negen isoglossen doen), maar ook verschijnselen die een bijzonder grillig verloop hebben. Nemen we het aggregaat van al deze verschijnselen, dan ontstaat een helder beeld (zie de kaarten in de Figuren 4b en 4c).

In paragraaf 5 vonden we dat de indeling op basis van isoglossen en de indeling op basis van Levensheinafstanden wezenlijk van elkaar verschillen. In deze paragraaf willen we nog een stap verder gaan door elk van de negen isloglossen afzonderlijk te vergelijken ten opzichte van de geaggregeerde Levenstheinafstanden. Op die manier kunnen we voor elk van de isoglossen bepalen hoe representatief ze zijn. In de traditionele dialectologie worden het al of niet uitspreken van de "slot-n" (isoglosse a) in paragraaf 3) en de uitspraak van de g (isoglosse g) in paragraaf 3) als de belangrijkste isoglossen gezien.

Om nu de representativiteit te bepalen van de isoglossen, bereken we afstanden tussen de variëteiten op basis van elk van elk van de negen isoglossen afzonderlijk. Tussen variëteiten die gescheiden worden door de betreffende isoglossen is de afstand 1, maar worden ze niet gescheiden door die isoglosse, dan is de afstand 0. We krijgen nu negen afstandstabellen. Daaraan voegen we nog een tiende toe, namelijk de tabel met metingen op basis van alle negen isoglossen samen (zie het laatste deel van paragraaf 3). We correleren de metingen in de tabellen nu met de geaggregeerde Levenshtein-afstanden. Wanneer we de correlaties kwadrateren en vermenigvuldigen met 100, krijgen we de variantiepercentages. We zien nu hoeveel variantie in de Levenshtein-afstanden door elk van de isoglossen en de combinatie van de negen isoglossen verklaard wordt. De resultaten worden gegeven in Tabel 1. Voor iedere isoglosse wordt ook het aantal dialecten gegeven waarvoor de waarde van de isoglosse bekend is.

isoglosse	**verklaarde variantie**	**aantal dialecten**
a) bakk<u>e</u> vs. bakk<u>(e)n</u>	11%	360
b) droo<u>g</u> vs. droo<u>ge</u>	6%	354
c) <u>ge</u>weest vs. <u>e</u>weest vs. <u>w</u>eest	20%	360
d) r<u>ij</u>p vs. r<u>ie</u>p	8%	360
e) h<u>ui</u>s vs. h<u>uu</u>s vs. h<u>oe</u>s	10%	358
f) <u>goe</u>d vs. <u>gou</u>d/<u>goo</u>d	6%	360
g) <u>jij</u> vs. <u>gij</u> vs. <u>ie</u>	26%	271
h) <u>sch</u>ool vs. <u>sk</u>ool vs. <u>sj</u>ool	6%	356
i) i<u>k</u> vs. i<u>ch</u> vs. i<u>sj</u>	7%	360
alle negen isoglossen samen	44%	360

Tabel 1. Representativiteit van de isoglossen in Figuur 2 in de geaggregeerde Levenshtein-afstanden. Voor ieder isoglosse wordt het aantal dialecten gegeven waarvoor de waarde van de isoglosse bekend is. In de onderste rij wordt de representativiteit gegeven van de isoglossenafstanden die we gemeten hebben in paragraaf 3.

De isoglossen g) en a) blijken qua representativiteit respectievelijk de eerste en derde plaats in te nemen. Beide isoglossen worden dus terecht als hoofdkenmerken gezien. Isoglosse c) neemt de tweede plaats in. Het valt dus goed te verdedigen om ook deze isoglosse onder de hoofdkenmerken te rangschikken.

7. Conclusie

In deze bijdrage hebben we negen isoglossenkaarten bekeken en geaggregeerde isoglosenafstanden tussen 360 taal- en dialectvariëteiten berekend. We hebben ook Levenshtein-afstanden berekend. De Levenshtein-afstand is een eenvoudige techniek waarbij keuze van isoglossen vooraf niet nodig is, maar alle informatie die in de transcripties van woorduitspraken – in ons geval 125 transcripties voor elk van 360 variëteiten – vervat is, wordt volledig verwerkt. Een ontrafeling van het materiaal in honderen aparte isoglossen – wat zeer tijdsintensief is – is dankzij deze techniek niet nodig.

De isoglossen vormen een beperkte en subjectief gekozen gegevensverzameling. Met de Levenshtein-afstand wordt echter een grote hoeveelheid aselect gekozen materiaal verwerkt, wat leidt tot objectieve resultaten. We vonden dat de negen isoglossen – als het ware een kleine aggregaatje - samen 44% van de Levenshtein-afstanden – het grote aggregaat - verklaren. De isoglossenafstanden suggeren dan ook een andere gebiedsindeling dan de Levenshteinafstanden. We zien dit als we de kaarten in Figuur 3 vergelijken met de kaarten in Figuur 4. Onze conclusie is dat het niet genoeg is om een gebiedsindeling te baseren op een klein aantal selectief gekozen isoglossen. Om een objectief beeld te krijgen zal de dialectoloog zich moeten baseren op een grote hoeveelheid aselect gekozen materiaal. Met de Levenshein-afstand kan dit materiaal snel en efficiënt worden verwerkt.

Vervolgens zijn we nog een stap verder gegaan en hebben we de representativiteit van de elk van de negen isoglossen afzonderlijk ten opzichte van de grote aggregaat bepaald. De slot-n en de jij/gij/ie-onderscheiding zijn inderdaad relatief belangrijke isoglossen zoals in de traditionele dialectologie gesuggereerd wordt, maar de uitspraak (of weglating) van de prefix in bijvoorbeeld *geweest* lijkt belangrijker te zijn dan het wel of niet uitspreken van de slot-n.

De metingen, analyses en cartografische weergave van geaggregeerde afstanden bleek met het GIS-pakket R*u*G/L04 op een adequate manier mogelijk. Veel GIS-pakketten ondersteunen alleen de analyse van monadische variabelen. Elk punt heeft eigenschappen, en die worden geanalyseerd en gevisualiseerd. R*u*G/L04 biedt de mogelijkheid om relaties *tussen* de punten, namelijk geaggregeerde taalkundige afstanden, te berekenen, te analyseren en te visualiseren. Dit GIS-pakket bleek een onmisbaar instrument te zijn bij de beantwoording van de vraag inhoeverre dialectindelingen op basis van één of meer isoglossen een representatief beeld geven.

Bibliografie

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

E. Blancquaert en W. Pée (eds.), *Reeks Nederlands(ch)e dialectatlassen*. De Sikkel (Antwerpen

1925-1982).

L. Bloomfield, *Language*. Henry Holt & Co (New York 1933).

R. Bolognesi en W. Heeringa, 'De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten', *Gramma/TTT; tijdschrift voor taalwetenschap*, 9 (2002), 45-84.

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

J. Daan en D. P. Blok, *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, N.V. Noord-Hollandsche uitgevers maatschappij (Amsterdam 1969).

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

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

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

W. Heeringa, W., *Measuring dialect pronunciation differences using Levenshtein distance*. Proefschrift rijksuniversiteit Groningen (Groningen 2004).

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

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

P. Kleiweg, J. Nerbonne en L. Bosveld, 'Geographic Projection of Cluster Composites', in: Alan Blackwell, Kim Marriott en Atsushi Shimojima (eds.), *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004. Cambridge, UK, March 2004 / Lecture Notes in Artificial Intelligence 2980*. Springer (Berlin 2004), 392-394.

J. Nerbonne, 'Data-driven Dialectology', aangenomen om te verschijnen in: *Language and Linguistics Compass*.

J. Nerbonne, W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten en W. van de Vis, 'Phonetic distance between Dutch dialects', in: G. Durieux, W. Daelemans en S. Gillis (eds.), *CLIN VI, Papers from the sixth CLIN meeting*. Universiteit Antwerpen, centrum voor Nederlandse taal en spraak (Antwerpen 1996), 185-202.

J. Nerbonne, P. Kleiweg, W. Heeringa en F. Manni, 'Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering', in: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt en Reinhold Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society.* Springer (Berlin 2008), 647-654.

J. Nerbonne en C. Siedle, 'Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede', *Zeitschrift für Dialektologie und Linguistik,* 72 (2005), 129-147.

J. Séguy, ‘La relation entre la distance spatiale et la distance lexicale’, *Revue de Linguistique Romane*, 35 (1971), 335-357.

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

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

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

A. Weijnen, *Nederlandse dialectkunde.* Van Gorcum & Comp (Assen 1966).

[i] 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/.

[ii] Om het voorbeeld eenvoudig te houden zijn diacritische tekens hier buiten beschouwing gelaten, maar worden in de berekeningen verderop in deze bijdrage wel verwerkt. In deze inleidende bijdrage voert het te ver daar dieper op in te gaan, maar we verwijzen de lezer naar Heeringa (2004) waarin de mogeljikheden tot verwerking van diacritische tekens bij gebruik van de Levenshtein-afstand uitvoerig worden besproken.

[iii] 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.