Computationele vergelijking en classificatie van dialecten

John Nerbonne en Wilbert Heeringa*
Alfa-informatica, BCN, University of Groningen
P.O.Box 716, NL 9700 AS Groningen, The Netherlands
{nerbonne,heeringa}@let.rug.nl

5 november 1999

Samenvatting

In traditional dialectology, maps are divided into dialect areas on the basis of isoglosses or with the use of the arrow method (Daan & Blok 1969). However, the choice of isoglosses is subjective. A second shortcoming for both methods is that there is no way to characterize relations between entire varieties. Comparison is inevitable atomistic. Third, existing methods records differences in varieties without distinguishing degrees of differences. The view that some differences vary along a continuum has no analytical foundation. This paper focuses on the Levenshtein method. Using this method, pronunciation differences can be measured for corresponding pronunciations in two varieties. We can apply the method to a large sample, providing an objective foundation for further analysis. The differences can be added, which allows one to relate entire varieties, aggregating the atomic differences. After comparing dialects on the basis of their distances the dialects can be classified by clustering or multidimensional scaling. Using clustering we get a sharp classification in the form of a tree, where the dialects are the leaves. Using multidimensional scaling we get a plot on which like dialects are plotted nearby and unlike dialects are distant. When

*Alle afbeeldingen in dit artikel zijn vervaardigd met behulp van programmatuur van Peter Kleiweg. We danken hem voor het ontwikkelen en beschikbaar stellen van zijn grafische programma's. Verder danken we Joseph Kruskal voor zijn adviezen betreffende multidimensional scaling. We danken ook Hermann Niebaum, Dickie Gilbers, Tjeerd de Graaf en Wouter Jansen voor de discussies die we met hen gehad hebben. Tenslotte danken we Harrie Scholtmeijer en G. de Schutter voor de waardevolle opmerkingen en suggesties die zij gaven na het doorlezen van het manuscript.

scaling to three dimensions, a map can be colored, where the dimensions represent respectively the intensity of red, green and blue, while the areas between the dialects are colored by interpolating from the dimensions of the dialects. In that way, dialect variation is visualized as a continuum.

1 Inleiding

1.1 Overzicht

Goossens (1977) geeft in zijn 'Inleiding tot de Nederlandse Dialectologie' een overzicht van methoden met behulp waarvan men in de traditionele dialectologie tot een indeling van Nederlandse dialecten kwam.

De oudste methode is de vlakkenmethode. Op basis van kennis van dialectgeografische tegenstellingen en intuïtie werd een kaart verdeeld in vlakken. Taalgebieden werden gescheiden door lijnen, of zij worden door verschillende kleuren gekarakteriseerd. Het gebruik van intuïtie maakt dat de indeling niet volkomen objectief en controleerbaar is.

Bij de isoglossen-methode tekent men de isoglossen op een kaart. Op verschillende plaatsen op de kaart zullen isoglossen samenvallen. Sommige isoglossenstrengen worden vervolgens als dialectgrens beschouwd, anderen weer niet. De keuze die hier gemaakt wordt is subjectief. Verder vallen isoglossen niet altijd samen. Ze kunnen ook evenwijdig lopen of elkaar kruisen. In de praktijk heeft men bij het construeren van isoglossenkaarten in het Nederlandstalig gebied een keuze gemaakt uit de bekende isoglossen. Opnieuw werd daarmee een keuze gemaakt die niet objectief is.

De pijltjesmethode baseert zich op het taalbesef van de dialectsprekers. Met pijltjes worden die dialecten verbonden die volgens de sprekers (vrijwel) gelijk zijn. Stroken waar geen verbindingspijltjes doorlopen zijn de dialectgrenzen. Het Nederlandse deel van de kaart van Daan (Daan & Blok 1969) is op basis van deze methode ingedeeld. De indeling van het Belgische deel berust op gegevens van taalkundigen uit dat gebied. Volgens Goossens is de pijltjesmethode nooit consequent toegepast, maar werden correcties toegepast wanneer de gegevens niet helemaal voldeden of elkaar soms tegenspraken (blz. 167). Daarmee zouden de ontwerpers van pijltjeskaarten hun eigen methode niet betrouwbaar achten. Verder kan nog een ander nadeel genoemd worden. Met deze methode is het niet goed mogelijk dialecten uit koloniën te vergelijken met de dialecten uit het land waar de kolonisten van oorsprong afkomstig zijn. Immers het zal niet gemakkelijk zijn voor zowel de kolonisten als de inwoners uit het land van herkomst om een uitspraak

te doen over de overeenkomst van elkaars dialect omdat men elkaars dialect nauwelijks of nooit meer hoort.

Een taalgebied kan ook op structuurgeografische basis worden ingedeeld. Hierbij wordt voor elk dialect de structuur van een taalkundig aspect in kaart gebracht. Zo heeft men getracht gebieden af te grenzen op basis van verschil in foneeminventaris. Alleen wanneer alle vragen naar de structuur van foneemsystemen en naar de bezetting van hun elementen bekend is, kan met behulp van deze methode een objectieve indelingskaart vervaardigd worden.

Dialectvergelijking kan ook plaatshebben op basis van de generatief-transformationele methode. De methode blijkt voor een gering aantal dialecten geschikt, maar is tot nu toe nog niet toegepast op grotere delen van het Nederlandse taalgebied.

1.2 Doel

Dit artikel heeft als doel een nieuwe methode voor de analyse van dialectologische gegevens te presenteren. De methode is gebaseerd op een manier om de afstand tussen woorden te meten. Omdat men een grote hoeveelheid gegevens analyseert die op een objectieve manier zijn gekozen, is sprake van een objectieve basis van de analyse. De afstanden tussen individuele woorden kunnen worden opgeteld om de afstanden tussen plaatsen af te leiden. De methode biedt dus een mogelijkheid om dialectverschillen op een hoger aggregatieniveau te karakteriseren. Dit staat tegenover o.a. de methode van isoglossen waarin men beweringen kan maken over "de slot-n in Gronings [lo.pm̩]", maar geen thesen over de Groninger streektaal als geheel. Tenslotte merken we op dat de afstandsmaat uiteraard geschikt is om de mate van verschil tussen plaatsen weer te geven en daardoor de analytische basis is voor de mening van veel deskundigen dat taalvarianten beter als een 'continuüm' afgebeeld kunnen worden, in plaats van als een aantal vlakken. Het zal opvallen dat ons doel constructief is: er wordt essentieel gebruik gemaakt van bestaand werk in de dialectologie en ons werk voegt daar tegelijk wezenlijk iets aan toe.

2 Data

De data op basis waarvan het onderzoek plaatsvond zijn afkomstig uit de 'Reeks Nederlandse Dialectatlassen' (RND). Deze atlassen zijn tot stand gekomen onder leiding van Blancquaert & Peé (1925–1982). Voor 1956 dialecten zijn vertalingen gegeven van steeds dezelfde 141 zinnen. Deze verta-

lingen zijn genoteerd in fonetisch schrift. Als eerste aanzet kozen we uit de RND 104 dialecten die ongeveer regelmatig verspreid liggen over het gehele Nederlandstalig gebied (zie Figuur 1). Uit de 141 zinnen kozen we 100 items (zie Tabel 1). Aan de hand van deze 100 woorden vindt vergelijking plaats van dialecten. De items zijn zodanig gekozen dat ze ongeveer alle klinkers en medeklinkers bevatten. Het is van belang te beseffen dat de uitspraken van de items afkomstig zijn uit complete zinnen. Wanneer we bijvoorbeeld 'drinken' willen knippen uit de zinsnede [ɔsəblutrInkə], wordt dit in geïsoleerde vorm [trInkə].

Bij het vergelijken van dialecten wordt alleen de segmentele informatie van de data in de berekeningen meegenomen. De plaats van de klemtoon in een woord wordt buiten beschouwing gelaten. De data zijn beschikbaar op het World Wide Web via http://www.let.rug.nl/~heeringa/dialectology/dial.htm.

Als dialectologen willen we graag een uitspraak doen over algehele variëteiten, bijvoorbeeld dat ze verwant of juist niet verwant zijn aan de standaardtaal, of dat intervocale consonanten in deze variëteiten lenis worden uitgesproken. Het is onmogelijk om alle relevante data in een dialectatlas of een beperkte steekproef te vinden, omdat de beweringen in principe *alle* uitingen in een bepaalde variëteit betreffen, zelfs alle uitingen in alle variëteiten binnen een bepaald gebied. De aangewezen weg in dit soort gevallen is om *aselect gekozen* steekproeven te analyseren. Door middel van statistische analyse kunnen beweringen worden getoetst over de hele populatie waaruit de steekproef wordt getrokken. Wij hebben niet met aselecte steekproeven gewerkt omdat dit niet practisch was: de dialectatlassen zijn beschikbaar en nieuw veldwerk is duur. Desondanks lijken de methoden geschikt te zijn voor steekproeven, en zouden ook bij steekproeven functioneren, tenminste als de atlasdata representatief zijn.

Omdat we met dialectatlasdata werken, moesten soms beslissingen genomen worden over ontbrekende woorden en lexicaal verschillende woorden, zodat de data ook verwerkt kunnen worden door woordgeoriënteerde vergelijkingsmethoden (zie sectie 3.2 en sectie 3.3). Als voor een woord twee uitspraken corresponderend met twee verschillende dialecten worden vergeleken, maar in de atlas ontbreekt bij één van beide dialecten de uitspraak voor dat woord, dan wordt dat woord buiten beschouwing gelaten in het vergelijkingsproces, omdat daaraan geen conclusie verbonden kan worden. Het kan in zo'n geval zijn dat het woord wel bestaat in die variëteit, maar niet in de uiting die op dat ogenblik werd gekozen. Waar de atlas meerdere variante uitspraken weergaf, werden alle afstanden gemeten en de gemiddelde afstand gebruikt. Waar lexicaal verschillende woorden in de atlas stonden, hebben

we deze wel mee laten wegen. Het leek zinvol om deze mee te nemen omdat de data veel lexicale variëteit bevat, en ook om te vermijden de beslissing te moeten nemen of het verschil tussen twee woorden fonetisch of lexicaal is.

3 Methoden voor vergelijking

In deze paragraaf worden drie computationele methoden voor het vergelijken van dialecten beschreven. Allereerst wordt de (foon-) en featurefrequentiemethode beschreven. Deze methode is ontwikkeld door Hoppenbrouwers & Hoppenbrouwers (1988). Verdere details kunnen worden gevonden in Hoppenbrouwers & Hoppenbrouwers (1993) en Hoppenbrouwers (1994).

Vervolgens wordt de frequentie-per-woord-methode beschreven. Dit is een tussenvorm van de frequentie-methode en de Levenshtein-afstand.

Tenslotte wordt de Levenshtein-afstand beschreven. De Levenshtein-afstand werd voor het eerst in de dialectologie toegepast door Kessler (1995). Kessler paste die toe op Ierse dialecten, wat heel succesvol bleek. Later werd deze afstandsmaat toegepast op Nederlandse dialecten door Nerbonne, Heeringa, van den Hout, van der Kooi, Otten & van de Vis (1996) en Nerbonne & Heeringa (1997), wat eveneens perspectiefvolle resultaten gaf. Een beschrijving van de Levenshtein-afstand is te vinden in Kruskal (1983).

Deze drie methoden bieden de mogelijkheid te onderzoeken of het zinvol is een woord als een taalkundige eenheid te beschouwen (frequentie vs. frequentie per woord) en of het zinvol is ook de sequentiële structuur van een woord in de beschouwing te betrekken (frequentie per woord vs. Levenshtein).

3.1 Frequentie-methode

Met de featurefrequentiemethode worden de frequentie van de fonetische features bepaald (Hoppenbrouwers & Hoppenbrouwers 1988). Twee dialecten worden met elkaar vergeleken door de featurefrequenties van die twee dialecten met elkaar te vergelijken. Een minder verfijnde methode is de foonfrequentiemethode, waarbij de frequenties van klanken worden bepaald en vergeleken. Bij het bepalen van de frequenties wordt het corpus beschouwd als niets anders dan een verzameling klanken. Woordgrenzen en klankvolgorde hebben geen betekenis. Wanneer we de frequenties van klanken of features delen door het totale aantal klanken in het corpus, krijgen we de relatieve frequenties.

3.2 Frequentie-per-woord-methode

Bij deze methode worden dialecten met elkaar vergeleken door een woord uit het ene dialect te vergelijken met het corresponderende woord in het andere dialect. Dat vergelijken gebeurt door voor beide woorden de frequenties van de klanken of van de features te bepalen, en die frequenties te vergelijken. De afstand tussen twee woorden is gelijk aan de som van de verschillen tussen de met elkaar corresponderende foon- of featurefrequenties van beide woorden. Bij deze methode worden woorden als taalkundige eenheden opgevat. De foon- of featurefrequenties van een woord worden gedeeld door het totale aantal klanken in dat woord zodat we relatieve frequenties krijgen. De volgorde van klanken in een woord wordt bij deze aanpak niet verdisconteerd. Deze methode hebben wij ontwikkeld om na te gaan wat de belangrijkste eigenschappen zijn die de frequentie-methode en de Levenshtein-methode van elkaar onderscheiden.

3.3 Levenshtein methode

Bij de Levenshtein-methode worden dialecten eveneens vergeleken door een woord uit het ene dialect te vergelijken met het corresponderende woord in het andere dialect. Echter, het vergelijken gebeurt nu door te bepalen hoe zo eenvoudig mogelijk het ene woord kan worden veranderd in het andere woord door klanken toe te voegen, te verwijderen of te vervangen. Aan de operaties worden gewichten toegekend. Toevoegen en verwijderen hebben hetzelfde gewicht, vervangen heeft als gewicht de som van het gewicht voor toevoegen en het gewicht voor verwijderen.

Stel het woord 'korst' wordt in het ene dialect uitgesproken als [kœstə] en in het andere dialect als [kɔrst]. Het veranderen van de ene variant in de andere gaat nu als volgt:

kœstə	verwijder ə	1
kœst	vervang œ door ɔ	2
kɔst	voeg toe r	1
kɔrst		
		4

Wanneer op deze manier twee woorden met elkaar vergeleken worden, zal de afstand tussen langere woorden gemiddeld genomen groter zijn dan de afstand tussen kortere woorden. Omdat dit niet overeenstemt met de idee dat een woord een taalkundige eenheid is, wordt de Levenshtein-afstand gedeeld door de som van de lengtes van beide woorden. Op die manier

krijgen we de relatieve Levenshtein-afstand tussen twee woorden. In ons voorbeeld is die afstand gelijk aan 4 / (3 + 3) = 0.67.

Omdat vergeleken wordt op basis van 100 woorden, krijgen we bij vergelijking van twee dialecten 100 relatieve Levenshtein-afstanden. De totale dialectafstand is nu gelijk aan de som van de 100 Levenstein-afstanden. Dankzij het feit dat de afstand van elk van de 100 woordparen wordt gedeeld door de som van de lengtes van beide woorden uit het woordpaar, is de invloed van een woordpaar dat bestaat uit langere woorden niet groter dan de invloed van een woordpaar dat bestaat uit kortere woorden.

We zien dat bij toepassing van de Levenshtein-afstand niet alleen woordgrenzen in acht genomen worden, maar bovendien ook de volgorde van klanken in een woord in de beschouwing betrokken wordt. Deze methode vormt het uitgangspunt voor de rest van dit artikel.

Omdat woordparen worden vergeleken in alle paren die gevormd kunnen worden uit de 104 dialecten, worden meer dan een half miljoen woordafstanden bepaald. Dit maakt de computer tot een onmisbaar hulpmiddel in het project. Daarom noemen we de aanpak 'computationeel'.

Omdat dit artikel de inzet van de Levenshtein-afstand in de dialectologie bepleit, willen we ook duidelijk zijn over de beperkingen die er aan vastzitten. Ten eerste meet de methode de afstand op basis van de segmentele representatie van woorduitspraken. Suprasegmentele eigenschappen zoals intonatie en klemtoon worden systematisch buiten beschouwing gelaten. Ons pleidooi voor het gebruik van de Levenshtein-afstand mag niet worden geïnterpreteerd als minachting van dialectverschillen die hierdoor niet goed geanalyseerd kunnen worden. Daarvoor zijn andere methoden nodig. Ten tweede zijn fonetische transcripties nodig van de uitspraken van dezelfde woorden op veel verschillende plaatsen. Het is uiteraard een voordeel dat de methode met veel materiaal uit de voeten kan, maar veel materiaal is ook *nodig* om goede resultaten te verkrijgen. Ten derde houdt de vergelijking niet speciaal rekening met fonologische processen, zoals final devoicing (verscherping) in gevallen zoals [le.və] vs. [le.f]. In dit soort gevallen wordt de afstand vergroot, zowel door de ə-deletie alsook door het verschil tussen [v] en [f], hoewel dat laatste het automatische gevolg is van het eerste, gegeven de final devoicing. Zoals we hieronder laten zien, is het verschil [v/f] echter minimaal vanwege de verfijning van de afstandsmaat door het gebruik van fonetische kenmerken.

4 Varianten van methoden voor vergelijking

Het oorspronkelijke Levenshtein-algoritme beschouwt een woord als een reeks symbolen. Daarbij zijn ten aanzien van diftongen meerdere benaderingen mogelijk (4.1). Het algoritme kan worden verfijnd door elk symbool te vervangen door een reeks van features (4.2), en bij vergelijking van klanken de features van de klanken te vergelijken (4.3). Omdat features meestal onderling afhankelijk van elkaar zijn, is sprake van redundantie. Soms wordt dit opgelost door toepassing van redundantieregels, maar het zou ook opgelost kunnen worden door elke feature te wegen, waarbij de grootte van het gewicht afhankelijk is van de onderscheidingskracht van de feature (4.4). In Figuur 2 zien we de Levenshtein-afstanden grafisch-geografisch gevisualiseerd.

4.1 Representatie van diftongen

Diftongen worden soms beschouwd als de opeenvolging van twee afzonderlijke klanken (Moulton 1962), soms ook als één klank (Hoppenbrouwers & Hoppenbrouwers 1988). Beide benaderingen hebben we onderzocht. Beide benaderingen blijken tot goede resultaten te kunnen leiden.

4.2 Featurerepresentaties van klanken

Bij het vergelijken van dialecten op basis van fonetische symbolen wordt geen rekening gehouden met het feit dat sommige klanken heel verwant, en andere heel verschillend zijn. Bijvoorbeeld het paar [b,p] is meer verwant dan het paar [a,p]. Dit probleem kan opgelost worden door elke klank te vervangen door een reeks van featurewaarden: een featurevector. Elke feature kan worden beschouwd als een fonetische eigenschap die kan worden gebruikt voor het classificeren van klanken. Een featurevector bevat voor elke feature een waarde die aangeeft in welke mate de eigenschap geldt. Bij vergelijking op basis van symbolen werden diacritische tekens niet in de beschouwing meegenomen. Bij representatie van klanken door featurevectoren kan de invloed van diacritische tekens te gelde gemaakt worden door de featurewaarden dienovereenkomstig aan te passen.

In onze experimenten gebruikten we twee featuresystemen. Het eerste systeem is beschreven in de publicaties van Hoppenbrouwers: Hoppenbrouwers & Hoppenbrouwers (1988), Hoppenbrouwers & Hoppenbrouwers (1993), Hoppenbrouwers (1994). Het tweede systeem is te vinden in Vieregge, Rietveld & Jansen (1984).

4.2.1 Featuresysteem Hoppenbrouwers

Het systeem van Hoppenbrouwers is gebaseerd op de Sound Pattern of English van Chomsky & Halle (1968). Dit systeem is interessant omdat de ontwikkelaars het zelf voor dialectvergelijking gebruikten. Het systeem bestaat uit 21 binaire features die gelden voor zowel klinkers als medeklinkers, waardoor klinkers en medeklinkers eenvoudig met elkaar vergeleken kunnen worden (het Levenshtein-algoritme verlangt namelijk dat alle mogelijke klanken met elkaar vergeleken kunnen worden). Tabel 2 vermeldt uit welke features het systeem bestaat.

In overeenstemming met de tabel die gegeven wordt in (Hoppenbrouwers & Hoppenbrouwers 1988) lieten we de volgende diacritische tekens van invloed zijn op de waarden van één of meer features van een klank: halflang/lang (waarbij halflang=lang), halve nasalering/nasalering (waarbij halve nasalering=geen nasalering), palatalisering en vocalisering.

4.2.2 Featuresysteem Vieregge

Het systeem van Vieregge et al. is interessant vanwege het feit dat de geldigheid van het systeem van de consonanten getest is door middel van een experiment waarin proefpersonen werd gevraagd het verschil tussen een aantal consonantparen uit de drukken in een getal. Na het experiment werd het systeem zodanig aangepast dat het verschil tussen twee featurevectoren van twee consonanten overeenkwam met het getal dat de proefpersonen opgaven als verschil tussen die twee consonanten.

Het systeem van Vieregge et al. werd toegepast om de kwaliteit van fonetische transcripties van proeftranscribenten te meten. Dit is vergelijkbaar met onze taak: terwijl Vieregge et al. maten hoe sterk de transcripties van de proeftranscribenten overeenkwamen met de correcte transcriptie, willen wij de mate van overeenkomst tussen dialectuitspraken meten.

Het oorspronkelijke featuresysteem bestaat uit een systeem voor de vocalen (4 features) en een systeem voor de consonanten (10 features). Daardoor is het in beginsel niet mogelijk klinkers met medeklinkers te vergelijken. Er zijn echter verschillende redenen om de verschillen tussen klinkers en medeklinkers in de berekeningen soms op te willen nemen. Ten eerste zijn er voorbeelden waarin medeklinkers tot halfvocalen worden verzwakt, zoals 'oude' [audə/auwə], 'goede' [ɣudə/ɣu.jə] en andere voorbeelden waarin vocalen en halfvocalen alterneren, zoals 'saai' [sai/saj]. Ten tweede eist het Levenshtein-algoritme dat alle klanken met alle klanken vergeleken kunnen worden.

Om dit mogelijk te maken, hebben we beide systemen gecombineerd tot één systeem van 16 features: 4 features voor de vocalen, 10 features voor de consonanten en 2 extra features. Van de extra features geeft de eerste feature aan of het een vocaal, een consonant of geen van beide betreft, en de tweede feature geeft aan of het een monoftong of een diftong betreft. Bij vocalen worden voor de 10 consonantfeatures defaultwaarden gekozen. Hetzelfde gebeurt ook bij de consonanten voor de 4 vocaalfeatures. Het gebruik van defaultwaarden vinden we ook bij het systeem van Hoppenbrouwers. Voor de klanken æ, œ, ʌ, w, ʊ, j en voor de diftongen kozen we definities die geheel analoog zijn aan die in het systeem van Hoppenbrouwers. Tenslotte converteerden we de meerwaardige features naar binaire features, zodat het systeem gebruikt kan worden voor de frequentie-methode en de frequentie-per-woord-methode. Wanneer bijvoorbeeld een feature de waarden 1, 2 en 3 kan krijgen, worden twee features gebruikt met de waardencombinaties '0,0', '1,0' en '1,1'. Na deze conversie heeft het systeem 26 binaire features. Tabel 3 vermeldt uit welke features het systeem in z'n oorspronkelijke vorm bestaat.

We hebben alleen die diacritische tekens in de beschouwing betrokken die in het oorspronkelijke systeem van Vieregge ook een rol spelen, namelijk half lang/ lang. Bij Vieregge wordt het gebruik hiervan beperkt tot een aantal vocalen. Om de RND-data goed te kunnen verwerken is het aangepaste systeem nu zodanig opgezet dat elke vocaal kort, half lang of lang mag zijn.

4.3 Vergelijking van featurevectoren

In feite vormen de featurewaarden '+' en '-' nominale data. Door '+' te representeren als 1 en '-' als 0, kunnen featurewaarden geïnterpreteerd worden als intervaldata. We gebruiken drie methoden voor het meten van de fonetische afstand tussen featurefrequentie-histogrammen of featurevectoren.

De eerste is de Manhattan afstand, ook bekend als 'taxicab distance' of 'city block distance' (Jain & Dubes 1988). De afstand is gelijk aan de som van de featureverschillen van de corresponderende features van twee featurevectoren die elk bestaan uit n features.

$$\delta(X, Y) = \sum_{i=1}^{n} |X_i - Y_i|$$

De tweede is de Euclidean afstand (Jain & Dubes 1988). Zoals gebruikelijk is de afstand bij deze methode gedefinieerd als de wortel uit de som van de kwadraten van de featureverschillen van de corresponderende features van twee featurevectoren die elk bestaan uit n features.

$$\delta(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

De derde methode is de Pearson correlatiecoëfficiënt, r (Buys 1989). Voor het gebruik als afstandsmaat gebruikten we $1 - r(X,Y)$, waarbij

$$r(X,Y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

en waarbij z_{x_i} en z_{y_i} de genormaliseerde waarden van de vector zijn (op de i-de positie). Voor toepassingen waarbij geen onderscheid hoeft te worden gemaakt tussen positieve en negatieve correlatie gebruikt men dikwijls als afstandsmaat $1 - r(X,Y)^2$.

4.4 Redundantie

Tussen de features van een featurevector bestaan afhankelijkheden. Bijvoorbeeld een klank die als kenmerk [+vocaal] heeft, zal ook [+sonoriteit] als kenmerk hebben. Featurewaarden die op grond van regels voorspeld kunnen worden, vormen redundante informatie die niet in de beschouwing betrokken zou mogen worden. De features van deze waarden moeten als afwezig beschouwd worden en dus als waarde de waarde 0 krijgen. De regels voor het featuresysteem van Hoppenbrouwers zijn te vinden in (Hoppenbrouwers & Hoppenbrouwers 1988) en (Hoppenbrouwers & Hoppenbrouwers 1993) en worden hier gegeven in Tabel 4. Van het featuresysteem van Vieregge zijn geen redundantieregels bekend.

Een nadeel van werken met redundantie-regels is dat niet altijd in één oogopslag te zien is welke afhankelijkheden er bestaan. Zo gaven de gebroeders Hoppenbrouwers aanvankelijk in 1988 de regels 1 t/m 4, maar later bleek ook een regel 5 van toepassing te zijn die zij in 1993 gaven.

Een verdere tekortkoming van het gebruik van redundantieregels is dat ze alleen toepasbaar zijn als een featurewaarde 100% voorspelbaar is op basis van andere features. Maar sommige features zijn nauwelijks onderscheidend, hoewel ze niet 100% voorspelbaar zijn, bijvoorbeeld de feature laryngaal in het featuresysteem van Hoppenbrouwers.

Bij het genereren van de resultaten die in dit artikel gepresenteerd worden hebben we, wanneer we het featuresysteem van Hoppenbrouwers gebruikten, de redundantie-regels toegepast. We hebben echter ook reeds geëxperimenteerd met een alternatief waarbij features gewogen worden. De

zwaarte van het gewicht hangt daarbij af van de onderscheidingskracht van een feature. Het beoogde effect is dat redundante features niet of nauwelijks meetellen. De hieronder beschreven methode moet slechts beschouwd worden als een eerste stap.

Het gewicht voor elke feature wordt bepaald op basis van de informatiewinst van die feature. De informatiewinst van een feature is een getal dat de gemiddelde entropie-reductie van de feature representeert wanneer de waarde van die feature bekend is (Quinlan 1993).

Stel in de dialectdata vinden we in totaal k verschillende segementen: $S_1..S_k$. Gesommeerd over alle woorden en dialecten zijn er in totaal $|D|$ voorkomens van klanken, en dus ook $|D|$ featurevectoren. $|Sj|$ feature vectoren zijn gelijk aan segment S_j. We kunnen nu de database entropie als volgt berekenen:

$$H(D) = -\sum_{j=1}^{k} \frac{|S_j|}{|D|} \times log_2(\frac{|S_j|}{|D|})$$

Dit kan worden opgevat als het gemiddeld aantal informatiebits dat nodig is om te bepalen welk segment gebruikt wordt.

Het aantal featurevectoren waarvoor geldt dat feature f waarde v_i heeft is $|D_{[f=v_i]}|$. Nu kunnen we de gemiddelde entropie per feature berekenen:

$$H(D_{[f]}) = \sum_{v_i \epsilon V} \frac{|D_{[f=v_i]}|}{|D|} \times H(D_{[f=v_i]})$$

Het gemiddelde wordt gewogen bij het aantal keren dat een feature een bepaalde waarde aanneemt. Voor het berekenen van $H(D_{[f=v_i]})$ wordt de eerste formule gebruikt, toegepast op de verzameling van featurevectoren met waarde v_i voor feature f. De informatiewinst van feature f is nu:

$$G(f) = H(D) - H(D_{[f]})$$

Bij het bepalen van afstanden tussen featurevectoren zoals beschreven in 4.3 worden de features nu vooraf vermenigvuldigd met hun informatiewinst. Voor elk van de 21 features uit het systeem van Hoppenbrouwers hebben we $G(f)$ berekend. Deze gewichten zijn gegeven in Tabel 5.

We onderzochten ook een nog gevoeliger maat waarbij de informatiewinst genormaliseerd wordt, maar konden die maat niet toepassen. Hierbij wordt normalisatie gerealiseerd door de informatiewinst te delen door de entropie van de verdeling van de waarden van één enkele feature in de database. De

entropie van deze verdeling is gelijk aan het aantal bits dat nodig is om te bepalen wat de waarde van de feature is:

$$split(f) = -\sum_{v_i \epsilon V} \frac{|D_{[f=v_i]}|}{|D|} log_2 \sum_{v_i \epsilon V} \frac{|D_{[f=v_i]}|}{|D|}$$

$G(f)$ wordt nu gedeeld door $split(f)$. Maar deze normalisatiestap wordt gebruikt in 'machine learning' waar op basis van features geen perfecte classificatie bepaald kan worden. Aan de hand van fonetische features kan echter een perfecte classificatie bepaald worden. De genormaliseerde gewichten van de corresponderende features blijken alle dezelfde waarde te krijgen waardoor geen verschil in gewicht meer ontstaat.

De hierboven beschreven aanpak is in die zin een eerste stap dat de gewichten van de features onafhankelijk van elkaar bepaald worden. Voor elke feature wordt individueel het gewicht bepaald op basis van de onderscheidingskracht van die feature. De stap die in de toekomst genomen moet worden is bij de weging de afhankelijkheden tussen features te betrekken. In dat licht bezien is deze eerste stap weliswaar juist, maar niet voldoende.

4.5 Directe of indirecte vergelijking

Wanneer we de onderlinge afstanden van dialecten bepalen, worden de dialecten direct met elkaar vergeleken. Dialecten kunnen ook indirect met elkaar vergeleken worden. Daartoe worden alle dialecten afzonderlijk vergeleken met het standaard Nederlands. Op basis van 100 woorden kunnen we dan elk dialect definiëren als een reeks van 100 woordafstanden ten opzichte van het standaard Nederlands. De afstand tussen twee dialecten is nu gelijk aan de som van de verschillen van de waarden van de met beide dialecten corresponderende reeksen.

5 Methoden voor classificatie

Het resultaat van een vergelijkingsmethode bestaat uit een tabel, vergelijkbaar met een afstandentabel zoals die vaak te vinden is in agenda's. Echter in plaats van geografische afstanden bevat de tabel nu fonetische afstanden. Op basis van deze tabel kunnen dialecten worden geclusterd (5.1). Het resultaat is een dendrogram, een boom waarbij de dialecten de bladeren zijn. Eveneens kan men op basis van deze tabel multidimensional scaling toepassen (5.2). Dit resulteert in een abstracte visualisatie van de fonetische afstand, waarbij verwante dialecten dicht bij elkaar, en van elkaar verschillende dialecten ver uit elkaar op een kaart geplaatst worden.

5.1 Clustering

Clustering is een gangbare techniek bij historische wetenschappen ((Boonstra, Doorn & Hendrickx 1990), pp. 143 ff.), maar wordt ook toegepast in de psycholinguistiek ((Woods, Fletcher & Hughes 1986), pp. 249 ff.). Het doel van clustering is om belangrijke groepen in complexe data te identificeren. Het algoritme kan het beste aan de hand van een voorbeeld uitgelegd worden. Stel we hebben de volgende matrix:

	Assen	Delft	Kollum	Nes	Soest
Assen		68	59	66	79
Delft			84	81	75
Kollum				39	80
Nes					83
Soest					

De waarde van iedere cel (i, i) is uiteraard gelijk aan 0 (de afstand van een dialect tot zichzelf). Omdat de matrix symmetrisch is, hebben we de gegevens in de linker onderhelft niet nodig.

Clustering is nu een iteratief proces. In elke stap van de procedure zoeken we de kleinste afstand in de matrix op, en combineren de dialecten waartussen die afstand bestaat tot één cluster. Met het oog op de iteraties die op de huidige iteratie volgen bepalen we de afstand van het nieuwe gevormde cluster tot alle overige punten. Dit doen we aan de hand van een matrix-updating-algoritme. Jain en Dubes noemen zeven matrix-updating-algoritmes (Jain & Dubes 1988). Om het voorbeeld simpel te houden gebruiken we het gemiddelde ter illustratie. De afstand van k tot een nieuw gevormd cluster $[ij]$ is hierbij gelijk aan het gemiddelde van de afstand tussen i en k en de afstand tussen j en k. Voor elke k berekenen we dus:

$$d_{k[ij]} = \frac{d_{ki} + d_{kj}}{2}$$

In de afstandsmatrix hierboven blijkt de afstand tussen Kollum en Nes het kleinste te zijn. Wanneer we beide plaatsen combineren tot één nieuw cluster, moeten we de afstanden van dit nieuwe cluster tot alle overige elementen berekenen. Bijvoorbeeld de afstand tussen Assen en Kollum-Nes wordt nu als volgt berekend:

$$\begin{aligned} d_{Assen,[Kollum,Nes]} &= \frac{d_{Kollum,Assen} + d_{Nes,Assen}}{2} \\ &= \frac{59+66}{2} \\ &= 62,5 \end{aligned}$$

Nadat we ook de afstand tussen Delft en Kollum-Nes, en Soest en Kollum-Nes hebben berekend, krijgen we de volgende matrix (de nieuwe waarden zijn vetgedrukt, de oude waarden zijn in normaal lettertype gegeven):

	Assen	Delft	Kollum & Nes	Soest
Assen		68	**62.5**	79
Delft			**82.5**	75
Kollum & Nes				**81.5**
Soest				

Het proces waarbij we in elke iteratie een reductie uitvoeren herhalen we tot er geen elementen meer zijn overgebleven die tot een nieuw cluster kunnen worden gecombineerd. Het resultaat is een complete hiërarchische groepering van de dialect-varianten. Dit kan worden gevisualiseerd in de vorm van een dendrogram, een boom waarin de bladeren overeenkomen met de dialect-varianten en de lengtes van de takken corresponderen met de fonetische afstanden.

Zoals hierboven gemeld bestaan er verschillende matrix-updating-algoritmes. In onze experimenten bleek de 'Ward's method' (of 'minium variance') de beste resultaten te geven. Als de kleinste afstand gevonden wordt tussen dialect i en dialect j, vormen we een cluster $[i,j]$ en berekenen de afstand van elke k tot het nieuwe cluster:

$$\begin{aligned} d_{k[ij]} &= ((n_k + n_i)/(n_k + n_i + n_j)) * d_{ki} \\ &+((n_k + n_j)/(n_k + n_i + n_j)) * d_{kj} \\ &-(n_k/(n_k + n_i + n_j)) * d_{ij} \end{aligned}$$

waarbij n_i, n_j en n_k het aantal dialecten zijn die behoren tot respectievelijk cluster i, j en k, en d_{ij}, d_{ik} en d_{jk} de afstanden zijn tussen respectievelijk i en j, i en k, en j en k. Een resultaat is te vinden in Figuur 3.

5.2 Multidimensional scaling

Op basis van de coördinaten van plaatsen kan men de onderlinge afstanden bepalen. Het omgekeerde is ook mogelijk: op basis van de onderlinge afstanden kan men een optimaal coördinatensysteem bepalen met de coördinaten van plaatsen daarin. Dit laatste wordt gerealiseerd met een techniek die bekend staat als 'multdimensional scaling'. Multidimensional scaling is een wiskundige techniek die vergelijkbaar is met factoranalyse (Kruskal & Wish 1984). Op een multidimensional scaling-plot zijn sterk verwante dialecten dicht bij elkaar geplaatst, terwijl sterk verschillende dialecten juist

ver uit elkaar geplaatst zijn. Het blijkt dat de belangrijkste dimensies bij de analyse van dialectdata geografisch zijn: de eerste dimensie correspondeert met de west-oost-as, en de tweede dimensie met de noord-zuid-as.

In onze experimenten hebben we gebruik gemaakt van Mean-Field Multidimensional Scaling Version 1.1. Deze implementatie werd ontwikkeld door Hansjörg Klock en Joachim M. Buhmann (zie verder: (Klock & Buhmann 1997*a*) en (Klock & Buhmann 1997*b*). Dit programma heeft als voordeel dat het probleemloos overweg kan met 104 plaatsen, wat niet vanzelfsprekend is. Een resultaat is te vinden in Figuur 4.

Op basis van multidimensional scaling is het ook goed mogelijk een kleurenkaart te maken waarbij de mate van kleurcontrast de mate van verschil tussen dialecten visualiseert. Daartoe schalen we de onderlinge afstanden tussen de 104 dialecten naar drie dimensies. Vervolgens kan voor elk dialectpunt een kleur bepaald worden. De drie dimensies bepalen daarbij de intensiteiten van respectievelijk rood, groen en blauw.

Om ook de oppervlakte tussen de 104 punten in te kleuren moeten de waarden van de drie dimensies van de punten die liggen tussen de 104 gegeven punten bepaald worden. De waarden van de tussenpunten bepalen we door te interpoleren van de gegeven punten. Hiervoor gebruikten we Inverse Distance Weighting, een methode waarbij de invloed van een gegeven punt op een tussenpunt omgekeerd evenredig is met de afstand tussen beide punten. Een resultaat is te vinden in Figuur 6.

6 Conclusies

6.1 Keuzes in methoden

Om een indicatie te krijgen van de juistheid van de methoden hebben we de resultaten vergeleken met de kaart van Daan (Daan & Blok 1969). Op deze kaart is het Nederlandse taalgebied verdeeld in 28 gebieden. Aan de hand hiervan kunnen we nagaan welke dialecten tot hetzelfde gebied zouden moeten behoren. Dit leidde tot de volgende conclusies:

- De resultaten van de frequentie-per-woord-methode zijn beter dan van de frequentie-methode. Dit bevestigt dat het van belang is een woord als een taalkundige eenheid te zien. De Levensthein-methode is eveneens beter dan de frequentie-methode, maar is niet beter of slechter dan de frequentie-per-woord methode. We kunnen dus geen conclusies trekken over de rol van de interne structuur van een woord.

- Zowel op basis van symbolen, als op basis van het featuresysteem van Hoppenbrouwers, als op basis van het featuresysteem van Vieregge kunnen goede resultaten bereikt worden.
- Zowel de benadering waarbij een diftong als twee opeenvolgende klanken wordt beschouwd als de benadering waarbij een diftong als één klank wordt beschouwd blijkt tot goede resultaten te kunnen leiden.
- Voor het vergelijken van featurefrequentie-histogrammen of featurevectoren blijken de Manhattan-afstand en de Euclidean-afstand heel bruikbaar. Beide hebben de voorkeur boven het gebruik van de Pearson correlatiecoëfficiënt.
- Weging van features leidt niet of nauwelijks tot betere resultaten.
- Het is beter dialecten direct te vergelijken, en niet indirect via het standaard Nederlands.
- Dendrogrammen tonen een scherpe indeling (zie Figuur 3). Multidimensional scaling-plots (zie Figuur 4) laten zien hoe dialecten zich tot elkaar verhouden. Het wordt daarbij aan het oog van de kijker overgelaten om daarin groepen te ontdekken. Er kan naar twee of meer dimensies geschaald worden. Wanneer naar drie dimensies geschaald wordt en de drie dimensies de intensiteiten van rood, groen en blauw gaan representeren, kan in combinatie met interpolatie het dialectologisch landschap worden afgebeeld op de geografische kaart (zie Figuur 6). Deze werkwijze doet recht aan de gedachte dat sprake is van een continuüm.

6.2 Evaluatie

Uit sectie 1 komen als belangrijkste kritiekpunten ten aanzien van de traditionele methoden voor dialectclassificatie naar voren:

- Bij de analyse van gegevens worden subjectieve keuzes gemaakt.
- De aanpak is atomistisch, er ontbreekt een aggregatieniveau van 'variëteit'.
- Er ontbreekt een maat voor verschillen: uitspraken zijn gelijk of ongelijk, maar wanneer zij ongelijk zijn kan de mate van verschil of overeenkomst niet worden aangegeven.

De in dit artikel voorgestelde methoden dragen wezenlijk bij aan de verbetering van deze situatie.

- Ze beginnen vanuit een in principe aselect gekozen steekproef van dialectmateriaal, wat een objectieve basis garandeert.

- Ze leveren numerieke karakterisaties die in het bijzonder additief zijn. Dit biedt de mogelijkheid om observaties te aggregeren. Ze leveren de analytische basis voor uitspraken over 'Gronings' in plaats van 'de slot-[n] in de Groningse uitspraak "lopen" [lo.pm̩]'.

- Ze leveren de analytische basis om de mate van verschil te karakteriseren. Bijvoorbeeld: de standaard Nederlandse uitspraak voor 'bloemen' is [blumə]. De volgende tabel vermeldt enkele varianten, waarbij voor elke variant de Levenshtein-afstand ten opzichte van de standaarduitspraak gegeven is:

blo.mn̩	1.19
blɔ.um̩	1.13
blumn̩	1.01
blo.mə	0.75
blɔmə	0.98
blumən	0.99

  We zien dat alle mogelijke gradaties in verschil mogelijk zijn.

We karakteriseerden ons doel als constructief (sectie 1.2). Het zal duidelijk zijn in hoeverre deze karakterisering gerechtvaardigd is. Het degelijke werk van de Reeks Nederlands(ch)e Dialectatlassen levert de empirische basis voor deze methoden. De jarenlange discussie over de indeling van Nederlandse dialecten levert betrouwbaarbare ijkingspunten waarmee de methode verfijnd kan worden. Nu zijn we in een positie om op deze basis verder te bouwen, in het bijzonder om precieze uitspraken te doen over hele streektalen in plaats van over geïsoleerde verschijnselen, en over de maat van verschil in plaats van 'wel of niet verschillend'. Tenslotte kunnen we op basis van de hier voorgestelde methode een wiskundige beschrijving geven voor de visie van veel dialectologen dat we met een dialectcontinuüm te doen hebben.

6.3 Toekomstig Werk

Hoewel de resultaten tot nu toe bevredigend zijn, moet verder gewerkt worden aan een aantal punten in de technische opzet van dit werk.

gegevensbasis De dichtheid van de 104 plaatsen is laag, en de spreiding nog niet helemaal regelmatig. In de toekomst willen we rond tweehonderd plaatsen in de studie opnemen die samen een regelmatig patroon vormen.

feature weging Zoals in 4.4 uitgelegd werd zal een consequente onderlinge weging van features de betrouwbaarheid van het geheel verhogen, en in het bijzonder bescherming bieden tegen de grote redundantie in featuresystemen.

formele validatie We zijn begonnen met 40 variëteiten, hielden de keuzes van metingen vast (behandeling van diftongen, featuresysteem, maat van verschil tussen featurevectoren, behandeling van diftongen, en gebruik van weging). Een uitbreiding met 40 tot dan toe nog niet eerder geziene plaatsen bevestigde deze keuzes in een kruisvalidatie stap. Maar het is tijd om een maat van kwaliteit op deze data toe te passen om de grote reeks van mogelijke verfijningen nauwkeuriger onder de loep te nemen.

Behalve over verfijning van de methode, is het goed om wat over de mogelijke toepassingen van de methode te zeggen. Gegeven een maat van afstand tussen taalvariëteiten zijn enkele vragen preciezer te beantwoorden dan voorheen. Deze geven verdere motivatie aan het hierboven beschreven werk.

convergentie of divergentie Gegevens van verschillende tijdstippen en dezelfde plaatsen kunnen worden onderzocht om te zien of afstanden groter of juist kleiner werden. Het is onze bedoeling onderzoek te doen naar divergentie en convergentie van dialecten in Nederland op basis van data uit (Winkler 1874) en Scholtmeijer (die het werk van Winkler in 1996 herhaalde voor de dialecten in Nederland).

politiek invloeden Een verdere klassieke vraag binnen de Germaanse dialectologie betreft het precieze effect van het trekken van een politieke grens. Met de Levenshtein-afstand hebben we goed gereedschap in handen om deze vraag te bestuderen.

economische en culturele invloeden De dialectologie is van oudsher interessant als een spiegel van ontwikkelingen op economisch en cultureel vlak zoals migratie, overwinning, mobiliteit en de opkomst van media. De Levenshtein-afstand levert een nieuwe manier om op zulke vragen een preciezer antwoord te geven.

Vanzelfsprekend zouden toepassingen op andere taal- of dialectgebieden interessant zijn net zoals toepassing op andere, bijvoorbeeld sociale parameters van variatie.

Referenties

Blancquaert, E. & W. Peé (1925–1982), *Reeks Nederlands(ch)e Dialectatlassen*, De Sikkel, Antwerpen.

Boonstra, O., P. Doorn & F. Hendrickx (1990), *Voortgezette statistiek voor historici*, Coutinho, Muiderberg.

Buys, A. (1989), *Statistiek om mee te werken*, Stenfert Kroese, Leiden en Antwerpen.

Chomsky, N. & M. Halle (1968), *The Sound Pattern of English*, Harper & Row, New York.

Daan, J. & D. P. Blok (1969), *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

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

Hoppenbrouwers, C (1994), 'De indeling van de zuidoostelijke steektalen', *TABU: Bulletin voor taalwetenschap* **24**(2), 37–63.

Hoppenbrouwers, C. & G. Hoppenbrouwers (1988), 'De featurefrequentiemethode en de classificatie van nederlandse dialecten', *TABU: Bulletin voor taalwetenschap* **18**(2), 51–92.

Hoppenbrouwers, C. & G. Hoppenbrouwers (1993), 'De indeling van noordoostelijke dialecten', *TABU: Bulletin voor taalwetenschap* **23**(4), 193–217.

Jain, A. K. & R. C. Dubes (1988), *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, New Yersey.

Kessler, B. (1995), Computational Dialectology in Irish Gaelic, *in* 'Proceedings of the European Association for Computational Linguistics', EACL, Dublin, pp. 60–67.

Klock, J. & J. M. Buhmann (1997*a*), Data visualization by multidimensional scaling: A deterministic annealing approach, Technical Report IAI-TR-96-8, Rheinische Friedrich-Wilhelms-Universität, Institut für Informatik III, Bonn. Beschikbaar als: http://www.cs.uni-bonn.de/III/forschung/publikationen/tr/IAI-TR-96-8.abstract-en.html.

Klock, J. & J. M. Buhmann (1997*b*), 'Multidimensional scaling by deterministic annealing', *EEMMCVPR '97* pp. 245–260. Beschikbaar als: http://www-dbv.informatik.uni-bonn.de/abstracts/klock.emmcvpr.html.

Kruskal, J. B. (1983), An overview of sequence comparison, *in* D. Sankoff & J. Kruskal, eds, 'Time Warps, String edits, and Macro molecules; the sequence theory and practice of sequence comparison', Addison-Wesley, Massachusetts, pp. 1–40.

Kruskal, J. B. & M. Wish (1984), *Multidimensional Scaling*, Sage Publications, Beverly Hills and London.

Moulton, W. (1962), 'The vowels of dutch: Phonetic and distributional classes', *Lingua* **11**, 294–312.

Nerbonne, J. & W. Heeringa (1997), Measuring Dialect Distance Phonetically, *in* J. Coleman, ed., 'Workshop on Computational Phonology', Madrid. Beschikbaar als: http://odur.let.rug.nl/~nerbonne/paper.html.

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, eds, 'CLIN VI, Papers from the sixth CLIN meeting', University of Antwerp, Center for Dutch Language and Speech, Antwerp, pp. 185–202. Beschikbaar als: http://odur.let.rug.nl/~nerbonne/paper.html.

Quinlan, J. Ross (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.

Vieregge, W. H., A. C. M. Rietveld & C. I. E. Jansen (1984), A Distinctive Feature Based System for the Evaluation of Segmental Transcription in Dutch, *in* 'Proceedings of the 10th International Congress of Phonetic Sciences', Dordrecht, pp. 654–659.

Winkler, J. (1874), *Algemeen Nederduitsch en Friesch Dialecticon*, Martinus Nijhoff, 's-Gravenhage.

Woods, A., P. Fletcher & A. Hughes (1986), *Statistics in Language Studies*, Cambridge University Press, Cambridge.

kippen	knie	bed	kindje	bakken
mijn(bzvnw)	ragebol	optillen	dochtertje	eieren
vriend	pet	metselaar	ladder	markt
bloemen	paddestoel	boterham	mond	eikels
spinnen(ww)	brede	jaar	weg	hooi
machines	breder	water	liedje	groen
werk	breedste	potten	kelder	boompje
schip	standbeeld	maart	ossebloed	huis
beschimmeld	duivel	kaars	broer	melk
brood	meester	paard	karnemelk	koe
timmerman	keelpijn	zwaluwen	Italie	kruiwagen
splinter	steel	kaas	bergen	Duitsers
vinger	bezem	motor	vuur	blauw
fabriek	neen	dag	spuwen	geslagen
vier	geroepen	avond	brug	sneeuw
bier	peer	jongetje	veulen	nieuwe
twee	geld	brief	deur	stad
drie	vrouw	kar	naaien	doopvont
knuppel	zwemmen	koning	gras	soldaten
ik	sterk	rozen	brouwer	gebonden

Tabel 1: De 100 woorden op basis waarvan dialecten met elkaar vergeleken werden.

Figuur 1: De 104 dialecten op basis waarvan het onderzoek plaatsvond.

naam	**waarden**	**betekenis**
vokaal	-,+	geen vokaal, vokaal
voor	-,+	niet voor, voor
achter	-,+	niet achter, achter
rond	-,+	niet rond, rond
laag	-,+	niet laag, laag
polair	-,+	niet polair, polair
lang	-,+	kort, lang
perifeer	-,+	niet perifeer, perifeer
diftong	-,+	geen diftong, diftong
nasaal	-,+	niet nasaal, nasaal
consonant	-,+	geen consonant, consonant
anterieur	-,+	niet anterieur, anterieur
coronaal	-,+	niet coronaal, coronaal
posterieur	-,+	niet posterieur, posterieur
laryngaal	-,+	niet laryngaal, laryngaal
sonoriteit	-,+	geen sonoriteit, sonoriteit
stemhebbend	-,+	stemloos, stemhebbend
hoog	-,+	niet hoog, hoog
continuant	-,+	geen continuant, continuant
lateraal	-,+	niet lateraal, lateraal
syllabisch	-,+	niet syllabisch, syllabisch

Tabel 2: Het featuresysteem van Hoppenbrouwers.

Vocalen

naam	**waarden**	**betekenis**
advancement	2, 4, 6	voor, midden, achter
high	1, 2, 3, 4	laag, midden laag, midden hoog, hoog
long	1, 2, 3	lang, half lang, kort
rounded	0, 1	gespreid, rond

Consonanten

naam	**waarden**	**betekenis**
place	1,	bilabiaal/labiodentaal,
	2,	dentaal/alveolair/palato-alveolair,
	3,	palataal,
	4,	velair/uvulair,
	5	glottaal
voice	0, 1	stemloos, stemhebbend
nasal	0, 1	niet nasaal, nasaal
stop	0, 1	geen ploffer, ploffer
glide	0, 1	geen glijder, glijder
lateral	0, 1	niet lateraal, lateraal
fricative	0, 1	geen fricatief, fricatief
flap	0, 1	geen flap, flap
high	0, 1	niet hoog, hoog
distributive	0, 1	niet distributief, distributief

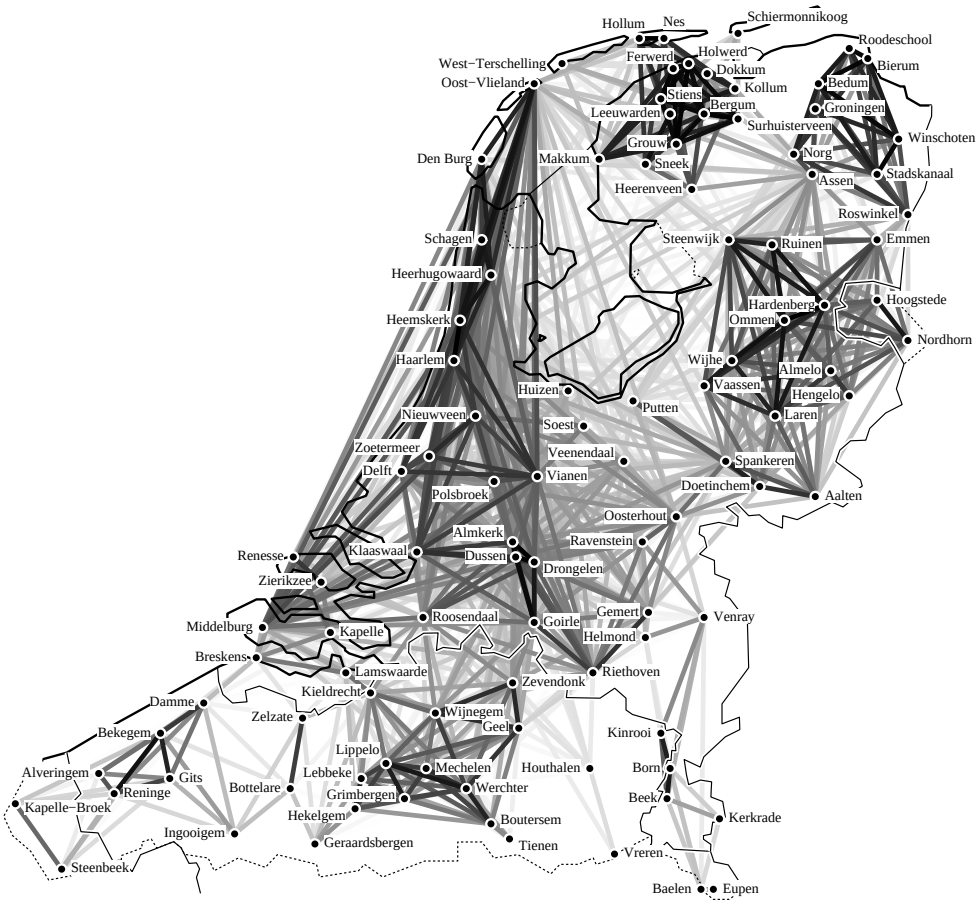
Tabel 3: Het featuresysteem van Vieregge.

1a	[+voor]	$\longrightarrow$	[+vok]
1b	[+acht]	$\longrightarrow$	[+vok]
1c	[+rond]	$\longrightarrow$	[+vok]
1d	[+laag]	$\longrightarrow$	[+vok]
2a	[+vok]	$\longrightarrow$	[+son]
2b	[+vok]	$\longrightarrow$	[+stem]
2c	[+vok]	$\longrightarrow$	[+cont]
2d	[+vok]	$\longrightarrow$	[+syll]
3	[+dift]	$\longrightarrow$	[+lang]
4	[+post]	$\longrightarrow$	[+hoog]
5	[+lar +stem]	$\longrightarrow$	[+cont]

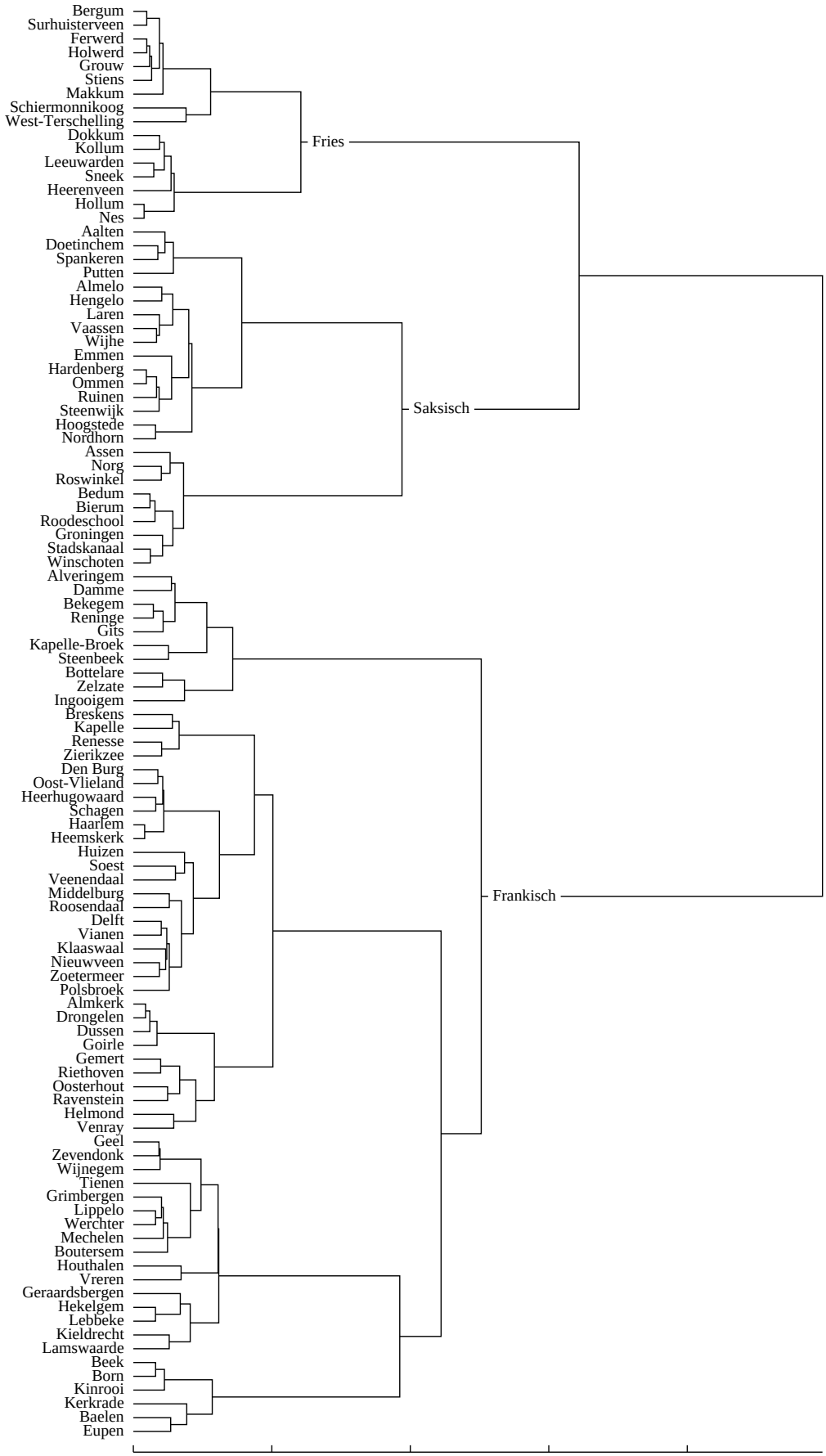
Tabel 4: Redundantie-regels voor het featuresysteem van Hoppenbrouwers.

feature	**gewicht**
vokaal	0.89560734
voor	0.58237173
achter	0.43238033
rond	0.47408821
laag	0.47575129
polair	0.47817399
lang	0.48153834
perifeer	0.36583266
diftong	0.00000000
nasaal	0.47261042
consonant	0.97052877
anterieur	0.98730285
coronaal	0.91274756
posterieur	0.44166573
laryngaal	0.04171495
sonoriteit	0.88995002
stemhebbend	0.76535414
hoog	0.56775535
continuant	0.87666806
lateraal	0.27770453
syllabisch	0.93374642

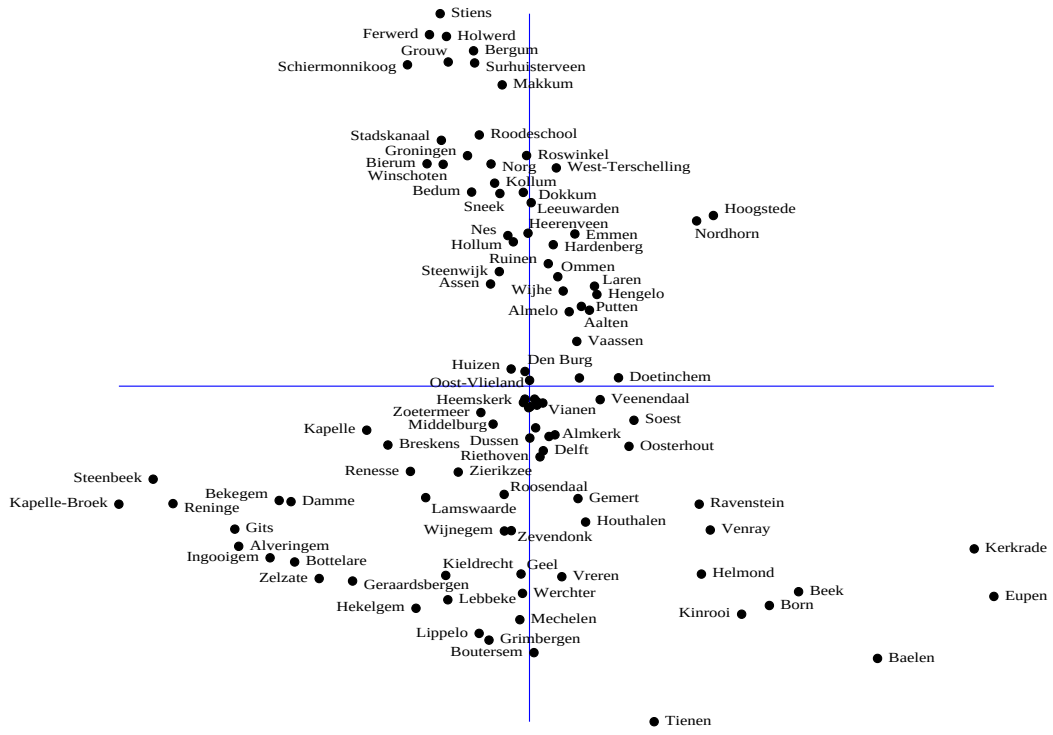
Tabel 5: Gewichten van de features van het systeem van Hoppenbrouwers.



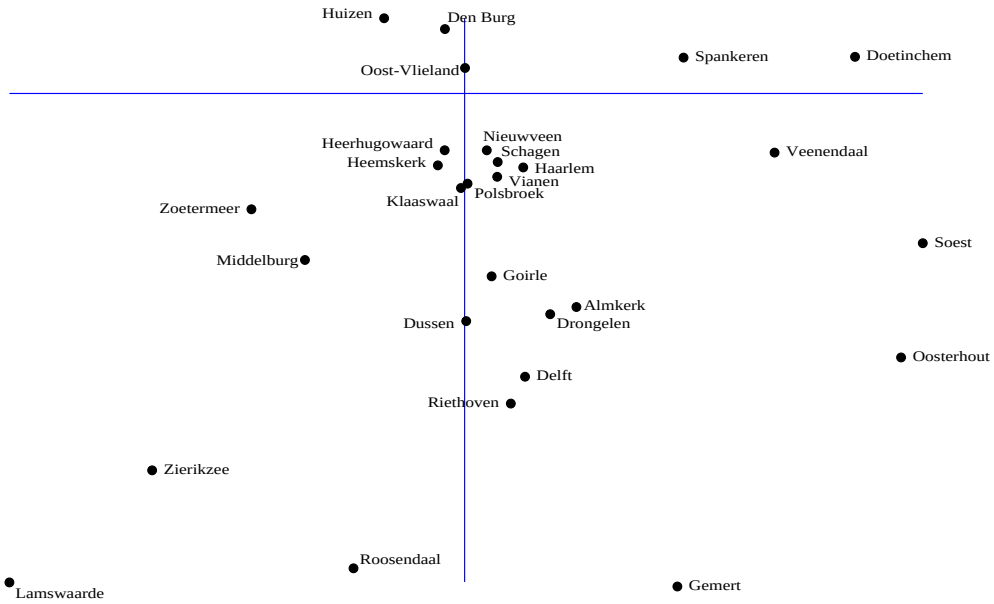
Figuur 2: De gemiddelde Levenshtein-afstanden tussen de dialecten. Diftongen zijn beschouwd als twee opeenvolgende segmenten, het featuresysteem van Hoppenbrouwers is gebruikt, de Euclidean-afstand tussen featurevectoren is toegepast, de features zijn niet gewogen, de dialecten werden direct met elkaar vergeleken. Hoe donkerder de lijn tussen twee punten, hoe kleiner de afstand tussen die twee punten is. Pearson's correlatiecoëfficient met de geografische afstanden is gelijk aan 0.6792 (significant).


Bergum
Surhuisterveen
Ferwerd
Holwerd
Grouw
Stiens
Makkum
Schiermonnikoog
West-Terschelling
Dokkum
Kollum
Leeuwarden
Sneek
Heerenveen
Hollum
Nes
Aalten
Doetinchem
Spankeren
Putten
Almelo
Hengelo
Laren
Vaassen
Wijhe
Emmen
Hardenberg
Ommen
Ruinen
Steenwijk
Hoogstede
Nordhorn
Assen
Norg
Roswinkel
Bedum
Bierum
Roodeschool
Groningen
Stadskanaal
Winschoten
Alveringem
Damme
Bekegem
Reninge
Gits
Kapelle-Broek
Steenbeek
Bottelare
Zelzate
Ingooigem
Breskens
Kapelle
Renesse
Zierikzee
Den Burg
Oost-Vlieland
Heerhugowaard
Schagen
Haarlem
Heemskerk
Huizen
Soest
Veenendaal
Middelburg
Roosendaal
Delft
Vianen
Klaaswaal
Nieuwveen
Zoetermeer
Polsbroek
Almkerk
Drongelen
Dussen
Goirle
Gemert
Riethoven
Oosterhout
Ravenstein
Helmond
Venray
Geel
Zevendonk
Wijnegem
Tienen
Grimbergen
Lippelo
Werchter
Mechelen
Boutersem
Houthalen
Vreren
Geraardsbergen
Hekelgem
Lebbeke
Kieldrecht
Lamswaarde
Beek
Born
Kinrooi
Kerkrade
Baelen
Eupen
Fries
Saksisch
Frankisch
0.0
1.0
2.0
3.0
4.0

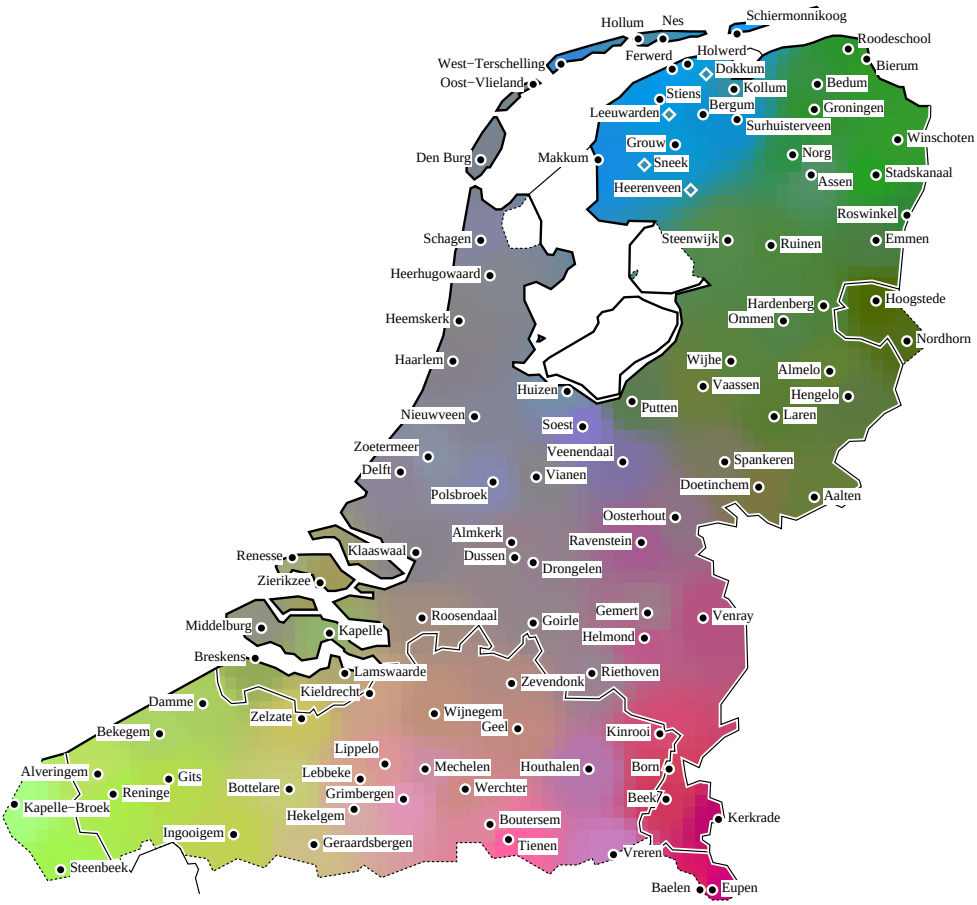
Figuur 3: Een dendrogram gemaakt op basis van de Levenshtein-afstanden. De drie hoofdgroepen zijn de Friese (boven), Saksische (midden) en Frankische dialecten (onder). Dit komt overeen met resultaten uit de traditionele dialectologie.



Figuur 4: Met behulp van multidimensional scaling worden 104 dimensies gereduceerd tot twee dimensies. De x-coördinaat representeert de tweede dimensie, en de y-coördinaat de eerste dimensie. Van boven naar onderen zien we achtereenvolgens de Friese, de Saksische en de Frankische dialecten. Merk op dat de horizontale as in sterke mate “correleert met” de west-oost-as van de geografische kaart, en de verticale as met de noord-zuid-as. Om overlap van labels in het middengebied te voorkomen, zijn er een aantal weggelaten. In Figuur 5 wordt dit gebied gedetailleerder weergegeven en zijn deze labels wel geplaatst.



Figuur 5: De dialecten uit het middengebied van Figuur 4.



Figuur 6: Met behulp van multidimensional scaling worden 104 dimensies gereduceerd tot drie dimensies. De drie dimensies bepalen in volgorde van belangrijkheid respectievelijk de intensiteit van rood, groen en blauw. Interpolatie vindt plaats met behulp van Inverse Distance Weighting. Het vloeiende verloop in de kleuren weerspiegelt de geleidelijke overgangen tussen de dialecten. De Friese steden doen niet mee in het interpolatie-proces. Dit is gevisualiseerd door de positie van de plaats met een diamant te markeren en alleen deze diamant in te kleuren.