Folkert de Vriend, Jos Swanenberg, Roeland van Hout

Dialectgebieden in Brabant. Geografische clustering op basis van de ruwe lexicale gegevens van het Woordenboek van de Brabantse Dialecten

Abstract[1]

In the project "digital databases and digital tools for WBD and WLD" (D-square) the dialect data published by the dictionary project Woordenboek van de Brabantse Dialecten (WBD) has been digitised. In this paper we analyse the WBD data using cluster analyses in order to see if we can find detailed dialect patterns in Brabant based on lexical data only. We compare the dialect patterns that we find with the detailed dialect map of Belemans & Goossens (2000). Special attention is given to the characteristics of the raw WBD data and how to cope with them.

1. Inleiding

In het project *Digitale databanken en digitaal gereedschap voor WBD en WLD* (D-kwadraat, gefinancierd door NWO, Investeringen Middelgroot) zijn alle gegevens gedigitaliseerd waarop het *Woordenboek van de Brabantse dialecten* (WBD) en het *Woordenboek van de Limburgse dialecten* (WLD) zijn gebaseerd.

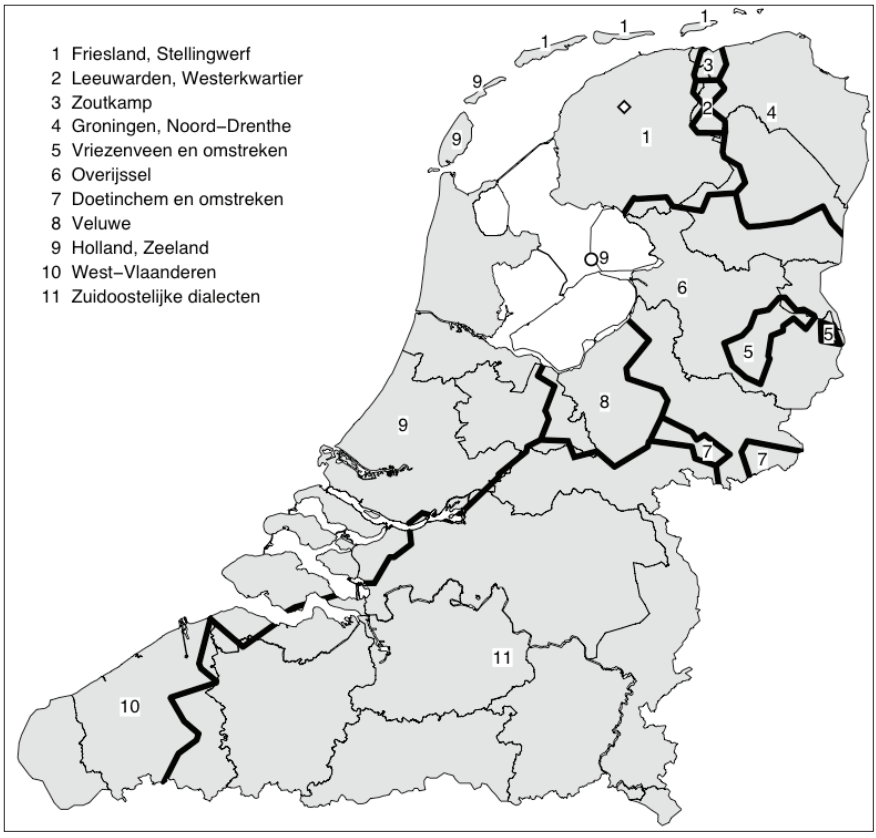
(1) We willen graag Louis ten Bosch bedanken voor zijn hulp bij het analyseren van de steekproeven, Peter Kleiweg voor een aanpassing aan RuG/L04 waardoor deze software nu ook met de WBD-data overweg kan, Jan Pieter Kunst voor zijn ondersteuning bij het gebruik van de interface waarmee symboolkaarten gegenereerd kunnen worden en Janienke Sturm voor commentaar op een eerdere versie van dit artikel.

Taal & Tongval themanr. 20 (2007), p. 83-110

Het gaat om een enorme hoeveelheid gegevens voor een groot aantal plaatsen. De digitale beschikbaarheid van deze gegevens levert de mogelijkheid om kwantitatieve onderzoeksmethoden toe te passen en om nieuwe onderzoeksvragen te stellen. Een vraag die gezien de omvang van het databestand voor de hand ligt, is of met dialectometrische technieken gedetailleerde dialectindelingen uit het materiaal zijn af te leiden. De indeling van dialecten in dialectgebieden is een onderwerp waar de dialectologie zich al van oudsher mee bezighoudt (zie bijvoorbeeld de website van het Meertens Instituut, met allerlei kaarten met indelingen van de Nederlandse dialecten). In deze bijdrage richten we ons op de indeling van de Brabantse dialecten op grond van de ruwe lexicale gegevens van het WBD.

We passen in onze analyses een bottom-up-benadering toe, een benadering vanuit de lexicale gegevens zoals ze zijn. Op grond van lexicale verschillen berekenen we steeds eerst afstanden tussen de plaatsen in het onderzoeksgebied. Hierbij moeten we rekening houden met de ruwheid van het materiaal. Met *ruw* doelen we op de aard van de gegevens zoals die zijn opgenomen in de woordenboeken; voor de lexicale varianten zijn ook de toevalligheden en wisselvalligheden opgenomen. Dit is een typische eigenschap van omvangrijke dialectwoordenboeken zoals het WBD. Maar door de ruwheid van het materiaal zijn misschien grote steekproeven benodigd om betrouwbare uitkomsten te krijgen.

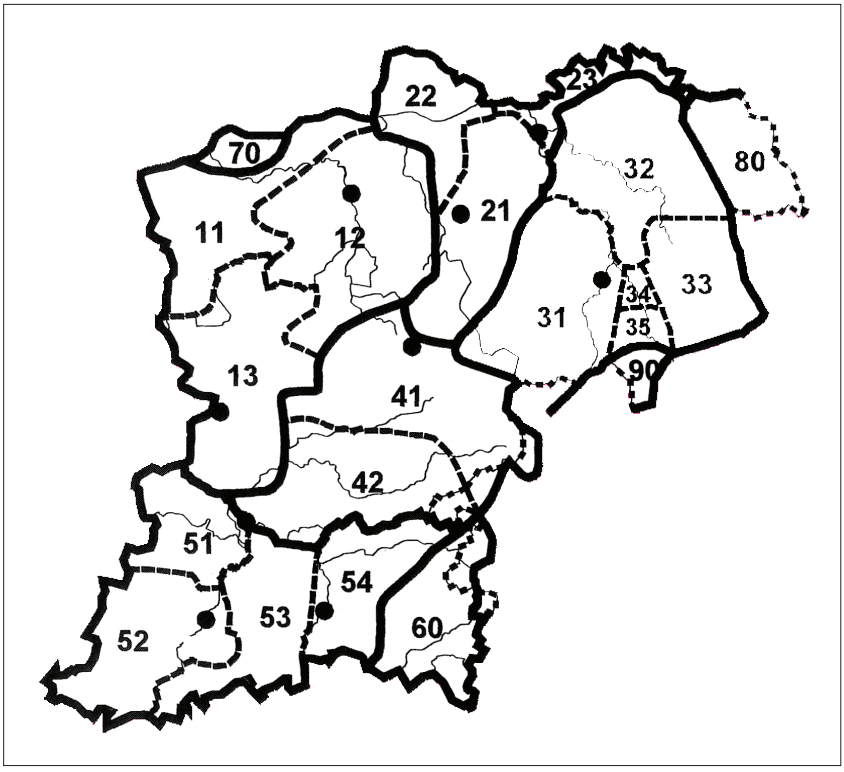
Vervolgens verkrijgen we via clusteranalyse op grond van die afstanden een indeling van de dialecten die we met een kaart kunnen afbeelden. Hierbij is het de vraag of het lexicale niveau zich eigenlijk wel goed leent om een gedetailleerd kaartbeeld voor het Brabants vanaf te leiden. De isoglossen die gebruikt worden om de dialecten te groeperen betreffen vaak fonologische of morfologische verschillen of het gaat om onderscheidingen op grond van functiewoorden (bijv. pronomina). Lexicale grenzen lijken over het algemeen veel willekeuriger en diffuser. Zo zijn in beide zuidelijke dialectwoordenboeken (WBD en WLD) zeer veel dialectkaarten opgenomen. Deze kaarten zijn vaak gekozen op grond van een interessant patroon waarbij gebiedsvorming meestal een criterium was. Praktisch alle kaarten laten ruime overlappingen met gemengde gebieden zien en bijna nergens is sprake van haarscherpe afgrenzingen.



Kaart 1: De lexicale dialectindeling van het Nederlandse taalgebied volgens Heeringa & Nerbonne (2006)

Ook de zuiver lexicale en op kwantitatief onderzoek gebaseerde indelingskaart van de Nederlandse dialecten van Heeringa & Nerbonne (2006) roept vraagtekens op over de geschiktheid van lexicaal materiaal voor het vinden van gedetailleerde grenzen in het Brabants dialect. Hun indelingskaart is afgebeeld in kaart 1 en is gebaseerd op 125 woorden uit de Reeks Nederlandse Dialectatlassen (RND). Zij selecteerden 360 plaatsen uit dit materiaal, verspreid over zestien provincies in Nederland en Vlaanderen. Dit resulteerde in een dataset van ca. 45.000 vormen (zie ook Heeringa 2004) waarop de auteurs een dialectometrische methode toepasten. Kaart 1 kent voor het hele Nederlandse taalgebied elf aaneengesloten gebieden. Van de elf gebieden zijn er al acht voor rekening van

het noordoosten, inclusief het Fries. Het Nedersaksische gebied valt uiteen in liefst zeven gebieden. Binnen Brabant is er in kaart 1 echter geen enkele sprake van differentiatie. Sterker nog, Brabant wordt samen met andere provincies (Nederlands Limburg, Belgisch Limburg, Oost-Vlaanderen) onder een enkel gebied geschaard. Dit gebied wordt in kaart 1 aangeduid als de zuidoostelijke dialecten (gebied 11). Ook al zijn er voor het bestaan van differentiatie binnen deze zuidoostelijke dialecten misschien voldoende lexicale bewijzen in de data van Heeringa & Nerbonne (2006) aanwezig, op kaart 1 komt dit in ieder geval niet tot uitdrukking.



Kaart 2: De dialectindeling van Brabant volgens Belemans & Goossens (2000)

Een kaart van een geheel andere aard is de kaart van Belemans & Goossens (2000). De indeling voor Brabant die zij voorstellen is afgebeeld in kaart 2.

Deze indeling biedt een samenvattend overzicht van de indelingen die voor het Brabants zijn voorgesteld op basis van kwalitatief onderzoek. Het gaat hierbij om traditionele dialectindelingen op basis van voornamelijk fonologische en morfologische en soms ook syntactische en lexicale verschillen, zoals Weijnen (1937), Pauwels (1958) en Lontie (1923). Terwijl in de kaart van Heeringa & Nerbonne (2006) Brabant opgaat in een veel groter gebied, valt Brabant in de kaart van Belemans & Goossens (2000) juist uiteen in een rijk palet aan subgebieden. Kaart 2 laat als globale indeling een indeling in negen gebieden zien. Vijf van die negen gebieden (10, 20, 30, 40 en 50) zijn ook nog eens verder opgedeeld in subgebieden. Zo ontstaat er een gedetailleerde opdeling in eenentwintig aaneengesloten dialectgebieden.

10 Noordwest-Brabants
- 11 Markizaats
- 12 Baronies
- 13 Antwerps

20 Midden-Noord-Brabants
- 21 Tilburgs
- 22 Hollands-Brabants
- 23 Maaslands

30 Oost-Noord-Brabants
- 31 Kempenlands
- 32 Noord-Meierijs
- 33 Peellands
- 34 Geldorps
- 35 Heeze-en-Leendes

40 Kempens
- 41 Noorderkempens
- 42 Zuiderkempens

50 Zuid-Brabants
- 51 Kleinbrabants
- 52 Pajottenlands
- 53 Centraal Zuid-Brabants
- 54 Hagelands

60 Getelands
70 Westhoeks
80 Cuijks
90 Budels

Ook al is de kaart van Belemans & Goossens (2000) niet enkel op lexicale gegevens gebaseerd, maar eerder op een amalgaam aan kennis over dialectvariatie in Brabant, de kaart leent zich door de hoge mate van detail erg goed om het resultaat van onze analyses tegen af te zetten. De kaart is voor Brabant bijvoorbeeld ook veel gedetailleerder dan de klassieke indeling die we kennen voor het Nederlandse taalgebied als geheel (zie bijv. Daan & Blok 1969). Een 1-op-1 overlap tussen de met onze clusteranalyses verkregen kaartbeelden en de indeling van Belemans & Goossens (2000) mogen we echter niet verwachten.

Gegeven de bovenstaande af- en overwegingen zijn er drie kernvragen die we op grond van de WBD-data met onze analyses willen beantwoorden:

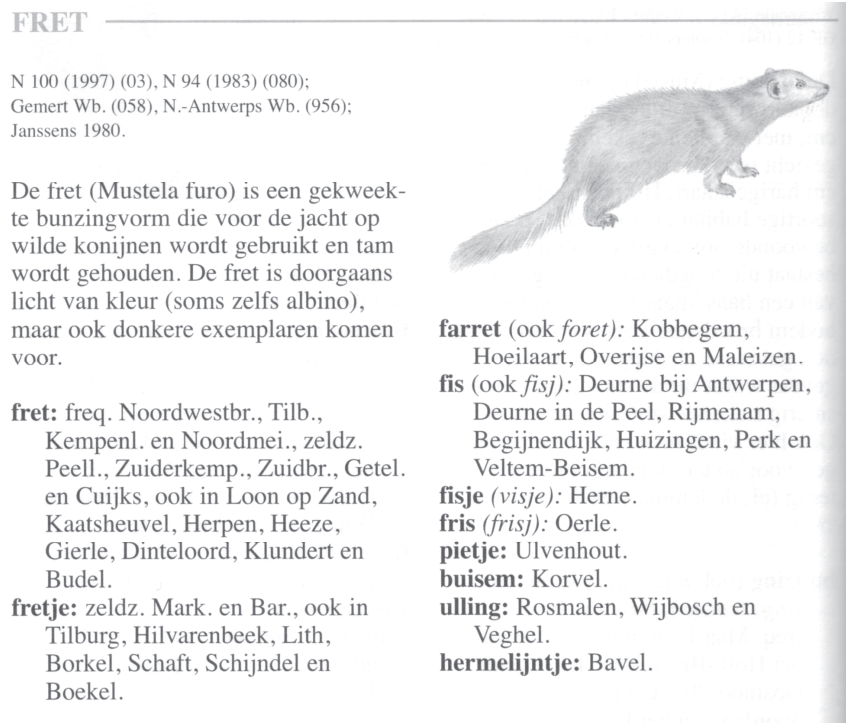
1) Hoe ruw zijn de WBD-gegevens?
2) Kunnen we op basis van lexicale gegevens gebiedsvorming voor Brabant vinden?
3) Hoe zeer sluiten de gebieden die we vinden aan bij de gedetailleerde kaart van Belemans & Goossens (2000)?

In paragraaf twee gaan we eerst in op de kenmerken van de dataset die we gebruikt hebben. Vervolgens bespreken we in paragraaf drie de methode die we hebben toegepast om plaatsen te clusteren op grond van lexicale afstanden en ze vervolgens op een kaart als dialectgebieden weer te geven. In de vierde paragraaf begint de data-analyse. We bekijken eerst hoe lexicale afstanden voor de totale dataset zich verhouden tot lexicale afstanden voor steekproeven uit de dataset, om een indruk te krijgen van de ruwheid van het materiaal. Vervolgens bespreken we de uitkomsten van de clusteranalyse voor de totale ruwe dataset. In paragraaf 5 reduceren we de ruwe data om het mogelijk storende effect van veel lege cellen in de dataset tegen te gaan en bespreken we vervolgens opnieuw de uitkomsten van de clusteranalyse. In de slotparagraaf gaan we in op enkele conclusies die we willen trekken op grond van onze ervaringen.

2. De lexicale gegevens van het WBD

Als uitgangspunt voor onze analyses hebben we deel III van het WBD genomen. Dit is het deel over de algemene woordenschat. Het in het WBD onderzochte gebied is gelijk aan het gebied van Belemans en Goossens (2000) dat is afgebeeld in kaart 2. Het onderzoeksgebied bestaat uit de provincies Noord-Brabant, Antwerpen, Vlaams-Brabant en het hoofdstedelijk gewest Brussel. In vergelijking

met deel I over de agrarische woordenschat en deel II over de niet-agrarische vakterminologieën is het materiaal van deel III van het WBD veel beter verspreid over de meetpunten van het onderzoeksgebied. De dataset is bovendien veel omvangrijker en niet gebonden aan sociaal beperkte gebruikssferen, zoals bij deel I en II. Deel III over de algemene woordenschat vormt daarom een directere afspiegeling van het lexicon. Wel moet opgemerkt worden dat functiewoorden en woorden die geen lexicale variatie vertonen niet als lemma zijn opgenomen.



FRET

N 100 (1997) (03), N 94 (1983) (080); Gemert Wb. (058), N.-Antwerps Wb. (956); Janssens 1980.

De fret (Mustela furo) is een gekweekte bunzingvorm die voor de jacht op wilde konijnen wordt gebruikt en tam wordt gehouden. De fret is doorgaans licht van kleur (soms zelfs albino), maar ook donkere exemplaren komen voor.

fret: freq. Noordwestbr., Tilb., Kempenl. en Noordmei., zeldz. Peell., Zuiderkemp., Zuidbr., Getel. en Cuijks, ook in Loon op Zand, Kaatsheuvel, Herpen, Heeze, Gierle, Dinteloord, Klundert en Budel.
fretje: zeldz. Mark. en Bar., ook in Tilburg, Hilvarenbeek, Lith, Borkel, Schaft, Schijndel en Boekel.

farret (ook *foret):* Kobbegem, Hoeilaart, Overijse en Maleizen.
fis (ook *fisj):* Deurne bij Antwerpen, Deurne in de Peel, Rijmenam, Begijnendijk, Huizingen, Perk en Veltem-Beisem.
fisje *(visje):* Herne.
fris *(frisj):* Oerle.
pietje: Ulvenhout.
buisem: Korvel.
ulling: Rosmalen, Wijbosch en Veghel.
hermelijntje: Bavel.

Figuur 1: Het WBD-lemma voor de fret

De algemene woordenschat van het WBD bevat een groot aantal lemma's. Een voorbeeld is het lemma *fret*, weergegeven in figuur 1. De figuur laat zien dat in de lemmatitel kort het concept wordt geformuleerd waarvoor in dit lemma de lexicale variatie wordt beschreven, in dit geval de fret. Onder de lemmatitel worden eerst de bronnen opgesomd waaruit de dialectvormen voor dit concept geëxcerpeerd werden. Het gaat hier hoofdzakelijk om twee vragen uit de zogeheten Nijmeegse enquête (N, in dit geval de vragenlijsten N 100 en N 94).

Vervolgens wordt een korte uitleg van het concept gegeven. In deel III van het WBD is ervoor gekozen om de uitspraakvariatie niet te publiceren maar alleen de lexicale variatie. Daarmee vallen klankverschillen weg. Wel worden verschillen in afleiding onderscheiden, met als gevolg dat bijvoorbeeld basisvorm en verkleinvorm worden onderscheiden (zie bijvoorbeeld het trefwoord *fretje* in figuur 1). In een WBD-lemma worden de lexicale vormen als trefwoord in vet weergegeven. Achter de lexicale vormen vindt men tot slot de dialectgebieden waar deze vorm gevonden werd en een frequentieaanduiding: *zeldz. Mark.* wil bijvoorbeeld zeggen dat deze vorm verhoudingsgewijs weinig (zeldzaam) is opgetekend in het Markizaatse dialectgebied. Indien de vormen in een bepaald gebied maar in één of twee plaatsen werden opgetekend, worden de namen van die plaatsen opgesomd in plaats van het dialectgebied.

Voor onze analyses zijn we niet uitgegaan van de data zoals die in de woordenboeken gepubliceerd zijn, maar van de database waarin de ruwe gegevens zijn opgeslagen en waarvan de gepubliceerde woordenboeklemma's zijn afgeleid. Van deze database hebben we op basis van twee criteria onze dataset afgeleid. Ten eerste werken we voor onze analyse enkel met de datacategorieën *concept*, *lexicale vorm* en *plaats* (zie ook: De Vriend, Boves, Van den Heuvel, Van Hout, Kruijsen & Swanenberg 2006). Als we bijvoorbeeld het lemma in figuur 1 nemen dan zijn we enkel geïnteresseerd in het gegeven dat voor het concept *fret* de lexicale vorm *farret* is gevonden in de plaats *Kobbegem*. De overige informatie in het lemma is niet van belang voor onze analyse.

Ten tweede hebben we de dataset beperkt tot data uit de Nijmeegse enquête. Materiaal uit andere bronnen zoals andere enquêtes en lokale woordenboeken hebben we buiten beschouwing gelaten omdat die bronnen niet ons gehele onderzoeksgebied bestrijken. Zo is de enquête van Schrijnen, Van Ginneken en Verbeeten uit 1914 bijvoorbeeld alleen in oostelijk Noord-Brabant afgenomen.

Onze totale ruwe dataset bestaat zo uit 4229 concepten, 639 Brabantse plaatsen en 614.941 lexicale vormen. In figuur 2 is een uitsnede van de dataset te zien. Deze heeft de vorm van een matrix. Op de horizontale as staan allereerst de plaatsen aangegeven. Deze zijn in de data als Kloeke-codes gecodeerd (Kloeke & Grootaers 1934): i 057a staat bijvoorbeeld voor Nieuw-Vossemeer en i 078 voor Halsteren. Op de verticale as zijn de concepten uitgezet: *gemartel*, *gemelijk*, etc. De cellen van de matrix bevatten de lexicale vormen. Figuur 2 geeft de matrix netjes in geordende kolommen weer. In de feitelijke database zijn de

velden door tabs gescheiden.

Ook al bestrijkt de Nijmeegse enquête het gehele onderzoeksgebied, niet voor iedere vraag en iedere vragenlijst is ook voor iedere plaats een antwoord opgetekend. Dit zien we terug in figuur 2 door de verschillende lege cellen. Verderop in dit artikel zal nog blijken dat deze eigenschap van belang is. Daarnaast komt het ook veelvuldig voor dat er voor een plaats juist meerdere vormen zijn opgetekend voor een concept. Zo zien we in figuur 2 dat bijvoorbeeld voor plaats i 057a en het concept *gerookt vlees* zowel de vorm *paardenspiertje* als *rundsspiertje* zijn opgetekend.

	i 057a	**i 057b**	**i 078**	**i 078a**
Gemartel	een heel gesjouw	Sukkelen	gesukkel	-
gemelijk	-	Grimmig	-	-
gemet	-	Gemet	gemet	-
gemoed	gemoed; ziel	Gemoed	gemoed	-
generale biecht	generale biecht	-	-	-
genezen (beter)	Beter	Beter	beter	-
geniepige plager	Pestkop	-	-	-
genoegen	Contentigheid	Plezier	-	-
geplooide kanten boord	-	Ruchetjes	-	-
geraamte	Geraamte	Geraamte	-	geraamte
gereed	Afgedaan	Af	af; klaar	-
gering aantal, een paar	-	paar; stukjes	paar	-
gerookt vlees	paardenspiertje; rundsspiertje	Rookvlees	rookvlees	-
gerookte haring	bruine bukkem; bukkem	Bukkem	bukkem	-
gerookte panpaling	gerookte paling	-	-	-
geruite jurk	Ruitenkleedje	Ruitenkleedje	-	-

Figuur 2: Uitsnede van de totale ruwe dataset met lexicale gegevens

3. De methode

De methode die we toepassen valt uiteen in drie hoofdstappen. Eerst berekenen we de lexicale afstand voor alle plaatsparen in het onderzoeksgebied. Vervolgens worden de plaatsen geclusterd op basis van deze lexicale afstanden. Tot slot wordt het resultaat van de clustering vertaald naar een kaartbeeld.

3.1. Lexicale afstanden

Lexicale afstanden geven aan hoe ver twee plaatsen lexicaal van elkaar afliggen, gebaseerd op de overeenkomsten en verschillen tussen het op elk van de plaatsen gebezigde lexicon. Een eenvoudige manier om lexicale afstand te berekenen is met een binaire maat voor al dan niet overeenkomst. Deze aanpak werd geïntroduceerd door Jean Séguy (Chambers & Trudgill 1998). In tabel 1 wordt geïllustreerd hoe met zo'n binaire maat de afstand tussen plaats A en plaats B wordt berekend voor elk concept in de dataset. De afstand is 1 voor concept A omdat plaats A en plaats B ieder een andere vorm (vorm A en vorm B) gebruiken voor dat concept. Voor concept B is de afstand 0 omdat plaats A en B dezelfde vorm voor dat concept gebruiken (vorm C). Uiteindelijk resulteert de optelsom van de verschillen, gewogen voor het aantal vergeleken concepten, in de lexicale afstand tussen twee plaatsen.

	plaats A	**plaats B**	**binaire afstand tussen plaats A en B**
concept A	Vorm A	vorm B	1
concept B	Vorm C	vorm C	0

Tabel 1: Lexicale afstand op basis van een binaire maat voor al dan niet overeenkomst

Heeringa & Nerbonne (2006) hebben met hun lexicale materiaal veelal betere gebiedsindelingen gevonden door gebruik te maken van een gewogen maat welke geïntroduceerd is door Goebl (1984): de *gewichteter Identitätswert* (GIW). Met de GIW-maat wordt de lexicale afstand tussen plaatsen die als énige in het onderzoeksgebied een vorm voor een concept gemeen hebben kleiner dan de lexicale afstand tussen plaatsen die een vorm voor een concept gemeen hebben met veel ándere plaatsen in het onderzoeksgebied. In tabel 2 wordt geïllustreerd hoe de afstand tussen twee plaatsen berekend wordt met behulp van de GIW-maat. Voor concept A hebben plaats A en B ieder een verschillende vorm. Met de GIW-maat wordt de afstand in die gevallen eenvoudigweg 1, net zoals bij de binaire maat. Voor concept B hebben plaats A en B echter dezelfde vorm. In dat geval is de afstand gelijk aan N'/N. Hierbij is N' het totaal aantal keer dat vorm C gevonden wordt voor concept B. N is het totaal aantal vormen dat gevonden wordt voor concept B. Wanneer plaats A en B een identieke vorm hebben wordt de afstand daarom een waarde tussen de 1 en de 0. De waarde zal meer richting de 0 gaan als de vorm maar weinig op andere plaatsen dan A en B voorkomt. Spruit (2008) merkt hierover nog op dat GIW infrequente woorden dus zwaarder

meetelt dan frequente woorden en dat dit ingaat tegen de in de kwantitatieve linguïstiek vaak gehoorde aanname dat infrequente woorden veelal juist ongewenste ruis zouden zijn.

	plaats A	**plaats B**	**GIW-afstand tussen plaats A en B**
concept A	vorm A	vorm B	1
concept B	vorm C	vorm C	N'/N

Tabel 2: Lexicale afstand op basis van de GIW-maat (gewichteter Identitätswert)

Voor het berekenen van zowel binaire afstanden als GIW-afstanden hebben we gebruik gemaakt van RuG/L04, een softwarepakket voor dialectometrie en cartografie dat ontwikkeld is door Peter Kleiweg en dat vaker is toegepast in recent dialectometrisch onderzoek naar de Nederlandse dialecten (zie bijv. Heeringa 2004 of Spruit 2008). Een groot voordeel van RuG/L04 is dat het gemakkelijk met categoriale variabelen, zoals dialectvormen, overweg kan.

De berekeningen met RuG/L04 leveren voor elk plaatskoppel in het onderzoeksgebied een lexicale afstand op. Figuur 3 toont een uitsnede van een door RuG/L04 opgeleverde matrix met daarin lexicale afstanden. In dit geval betreft het afstanden op basis van de binaire maat. Op de horizontale en de verticale as zijn de 639 plaatsen uit onze dataset uitgezet. In de cellen staan de afstanden die zijn berekend op basis van alle 614.941 lexicale vormen in de ruwe dataset. Verder is de matrix gespiegeld en bestaat de diagonaal uit enkel nullen. De afstand tussen een plaats en zichzelf is immers altijd nul. De maximale ongelijkheid (ofwel afstand) in de matrix van figuur 3 is 1 en de minimale ongelijkheid is 0. Wanneer er sprake is van meerdere vormen per plaats dan berekent RuG/L04 eenvoudigweg het gemiddelde van de afzonderlijke waardes. Als een vergelijking niet mogelijk is omdat voor één of allebei de plaatsen in een plaatspaar een vorm ontbreekt dan telt die vergelijking niet mee. Wanneer bij een plaatspaar voor álle concepten in de dataset blijkt dat dit het geval is, dan wordt dit in de matrix aangegeven met *NA* (*'not available'*). Zo zien we in Figuur 3 bijvoorbeeld dat voor het plaatspaar i 078a - i 102 om deze reden geen afstand berekend kon worden.

	i 057a	i 057b	i 078	i 078a	i 079	i 102	i 102a
i 057a	0	0.557414	0.601638	0.592415	0.645127	0.528001	0.64319
i 057b	0.557414	0	0.356028	0.646959	0.497905	0.423266	0.524859
i 078	0.601638	0.356028	0	0.611111	0.434753	0.425115	0.467366
i 078a	0.592415	0.646959	0.611111	0	0.531585	NA	0.563596
i 079	0.645127	0.497905	0.434753	0.531585	0	0.46354	0.53636
i 102	0.528001	0.423266	0.425115	NA	0.46354	0	0.338649
i 102a	0.64319	0.524859	0.467366	0.563596	0.53636	0.338649	0

Figuur 3: Matrix met lexicale afstanden tussen de plaatsen op basis van de binaire maat

3.2. Clustering

Op grond van de verkregen lexicale afstanden kunnen de plaatsen worden geclusterd. Het doel van clusteranalyse is om een indeling te krijgen naar elementen die bij elkaar geplaatst worden omdat ze veel op elkaar lijken (maximale gelijkenis), waarbij de clusters onderling maximaal verschillen (minimale gelijkenis). De elementen zijn in ons geval de plaatsen. Met clusteranalyse komen we dus te weten welke plaatsen lexicaal op elkaar lijken. Ook voor de clusteranalyses hebben we gebruik gemaakt van RuG/L04. De verschillende clusteralgoritmen die L04 aanbiedt worden beschreven in Jain & Dubes (1988) en zijn van het type *hierarchical aglomerative*. De algoritmen van dit type vertrekken *bottom up* vanaf de afzonderlijke plaatsen waarbij elke plaats wordt beschouwd als een cluster. Vervolgens worden de clusters samengevoegd in steeds groter wordende clusters met steeds meer plaatsen. Bij het zoeken van oplossingen voor het clusteren betrekken de algoritmen geen informatie uit andere data. Het resultaat van de clusteranalyse is een toekenning van de plaatsen aan clusters.

Om nog voor het stadium van kartering te kunnen bepalen welke parameterinstellingen voor afstandsmaat en clusteranalyse de beste gebiedsindeling opleveren maken we gebruik van een programma in RuG/L04 waarmee op basis van informatie over de lengte- en breedtegraden van de plaatsen in het onderzoeksgebied de *local incoherence* gemeten kan worden; het gebrek aan geografische samenhang op lokaal niveau. Over het algemeen wil een lagere waarde voor de *local incoherence* zeggen dat er sprake is van een betere meting. (zie: http://www.let.rug.nl/kleiweg/L04/Tutorial/t06.html.nl)

3.3. Kartering

Om de resultaten van de clusteranalyse vervolgens te kunnen interpreteren, vertalen we met een semi-automatische procedure het bestand waarin de clusters worden gedefinieerd naar een symboolkaart welke door de geo-browser *Google Earth* kan worden afgebeeld. Hiertoe hebben we met RuG/L04 telkens eerst het resultaat van de clusteranalyse opgedeeld in een aantal sets met plaatsen. Vervolgens dienen deze sets als input voor een kaartmodule die is ontwikkeld op het Meertens Instituut (zie www.meertens.knaw.nl/kaart onder "XML-RPC interface") en welke symboolkaarten kan genereren in het formaat van de geobrowser *Google Earth*. De Google-Earth-uitbreiding op de kaartmodule is ontwikkeld binnen het project D-kwadraat (De Vriend & Swanenberg 2006). Op de symboolkaarten krijgen plaatsen uit hetzelfde cluster telkens hetzelfde symbool toebedeeld. Op deze manier kunnen de mate van geografische differentiatie en de overeenkomst tussen het resultaat van de clusteranalyse en de indeling van Belemans & Goossens (2000) visueel geïnspecteerd worden. Om de vergelijking tussen onze resultaten en de indeling van Belemans & Goossens (2000) te vergemakkelijken, projecteren we de symboolkaart in *Google Earth* telkens over de indeling van Belemans & Goossens (2000) heen.

4. De analyse van de ruwe data

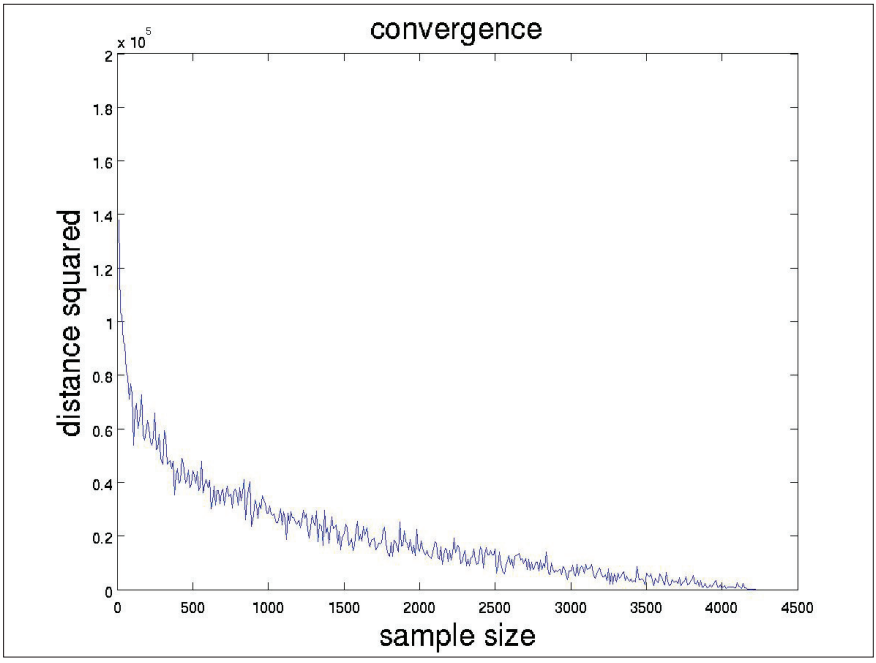
4.1. De steekproefomvang en de convergentie van de uitkomsten

Om een indruk te krijgen van de ruwheid van de totale dataset bekijken we eerst hoe de lexicale afstanden voor de totale dataset van 4229 concepten zich verhouden tot de lexicale afstanden voor steekproeven uit de dataset. Als het materiaal erg ruw en heterogeen is dan zullen de afstanden voor kleine steekproeven sterk afwijken van de afstanden voor de hele dataset. Ook zullen dan grote steekproeven benodigd zijn voor het krijgen van betrouwbare uitkomsten.

In totaal hebben we 422 steekproeven getrokken uit de dataset, oplopend met intervallen van 10, te beginnen met een trekking van 10 concepten: 10, 20, ..., 4210, 4220. De steekproeven zijn random en met teruglegging getrokken. Voor de random-trekking hebben we gebruikgemaakt van de Perl-functie *rand* welke random getallen genereert. Elk van de 422 steekproeven resulteert in een subset van de totale dataset. Op basis van de lexicale data in elk van deze subsets is vervolgens een afstandsmatrix berekend met gebruikmaking van de meest eenvoudige van de twee besproken maten; de binaire maat. Zo zijn uiteindelijk

422 afstandsmatrices verkregen welke op een steeds grotere steekproef zijn gebaseerd.

Om te bepalen hoe de afstanden behorend bij de steekproeven convergeren naar de afstanden voor de totale dataset hebben we een maat genomen welke wordt berekend tussen elk van de afstandsmatrices en de afstandsmatrix behorend bij de gehele dataset. De maat is berekend als de som van de gekwadrateerde verschillen over alle cellen tussen de afstandsmatrix voor de gehele dataset en de afstandsmatrix behorend bij een steekproef. Figuur 4 toont het resultaat.



Figuur 4: Convergentie van de afstand tussen steekproef en de totale dataset, oplopend in grootte van de steekproef

De grafiek toont wat het verloop is vanaf de kleinste steekproefomvang (10 concepten) naar de totale dataset (de 4229 concepten; in dit geval te beschouwen als de populatie). We zien dat zowel de fluctuatie als het verschil tussen steekproef en de populatie duidelijk afneemt bij toenemende steekproefgrootte. Kleinere random steekproeven kunnen echter nog behoorlijk verschillen ten opzichte van de populatie en dit is een indicatie dat ons materiaal inderdaad behoorlijk ruw

is. Hierdoor kunnen we pas bij grotere steekproeven betrouwbare uitkomsten verwachten. Voor de verdere analyse van de ruwe data werken we daarom in eerste instantie met de totale dataset en niet met een van de steekproeven.

4.2. Clustering van de ruwe data

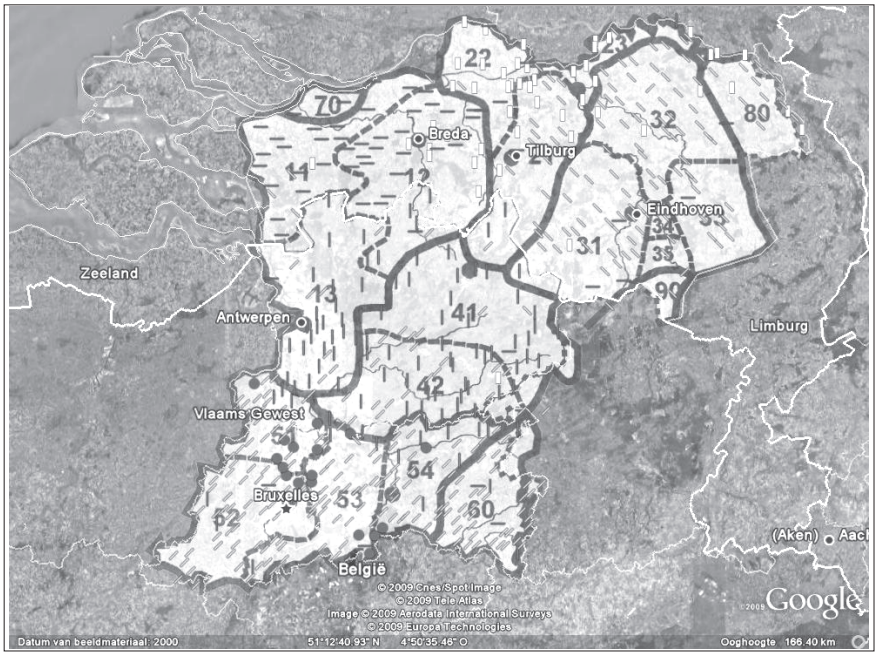
Op basis van de totale ruwe dataset bestaande uit 614.941 lexicale vormen hebben we afstanden berekend met zowel de binaire maat als de GIW-maat. Door vervolgens naar de *local incoherence* te kijken hebben we bepaald welke van de twee afstandsmaten de beste is voor onze dataset. De *local incoherence* voor de binaire maat blijkt 13.8227 te zijn en die voor de GIW-maat 10.173. De GIW-maat blijkt dus de beste maat te zijn voor onze dataset. Met de matrix met GIW-afstanden zijn we daarom verder gegaan.

Bij de beschrijving van de lexicale gegevens in paragraaf 2 wezen we al op het opvallend grote aantal lege cellen in onze dataset. In onze dataset zijn voor sommige plaatsen vele dialectvragenlijsten ingevuld, maar voor andere plaatsen slechts een enkele. Van de in totaal 2.702.331 cellen van de matrix in figuur 2 blijken er maar liefst 2.250.522 leeg te zijn. Dat komt neer op wel 83.3%. Dit is het natuurlijk gevolg van de wijze waarop de gegevens voor de zuidelijke dialectwoordenboeken verzameld zijn. Voor de in de woordenboeken toegepaste methodologie is dit geen probleem geweest. Daar was het hoofddoel om de vormdifferentiatie voor concepten te inventariseren, waarbij het er op zich niet veel toe doet uit welke plaats een dialectvorm nu precies afkomstig was. Voor de door ons toegepaste methodologie is dit echter wel een probleem. De vele lege cellen zorgen er voor dat het relatief vaak voorkomt dat voor een plaatspaar geen afstand berekend kan worden. Dit werd in de afstandsmatrix aangegeven met *NA* (zie paragraag 3.2). In totaal vinden we in de afstandsmatrix 43.407 *NA*'s. Dit komt neer op 10,6 % van de in totaal 408.321cellen in de afstandsmatrix. Het probleem zit hem er nu in dat de clusteralgoritmen die RuG/L04 aanbiedt niet overweg kunnen met ontbrekende afstanden. Het pakket biedt als oplossing voor het probleem een programma aan waarmee voor de ontbrekende afstanden plausibele waardes ingevuld kunnen worden ("*imputeren"*). De ontbrekende afstanden worden ingevuld op basis van de lexicale afstand tot geografisch nabije plaatsen. Welke plaatsen als geografisch nabij worden beschouwd wordt afgeleid uit een bestand waarin we voor elke plaats in ons onderzoeksgebied de lengte- en de breedtegraad hebben gedefinieerd. Met deze extra stap hebben we een afstandsmatrix gemaakt waarin voor elk plaatskoppel in het onderzoeksgebied een afstand is gedefinieerd. Met deze matrix zijn we vervolgens gaan clusteren.

Clustering is van nature erg instabiel (Jain, Murty & Flynn 1999). Zo kunnen kleine verschillen in de afstandsmatrix die voor de clusteranalyse gebruikt wordt, een groot effect hebben op de uitkomst. Om te voorkomen dat de uitkomst van de clusteranalyse te veel afhangt van toevalligheden in de afstandsmatrix stellen Kleiweg, Nerbonne & Bosveld (2004) voor om meerdere keren te clusteren met toevoeging van ruis. We hebben voor onze data dezelfde aanpak gehanteerd en ons voor de parameterinstellingen laten leiden door Nerbonne, Kleiweg, Manni & Heeringa (2008). Zij gebruikten voor een dataset bestaande uit Duitse dialectgegevens een ruiswaarde van 0,5 keer de standaarddeviatie van de afstanden en herhaalden de clusteranalyse vervolgens minimaal 100 keer.

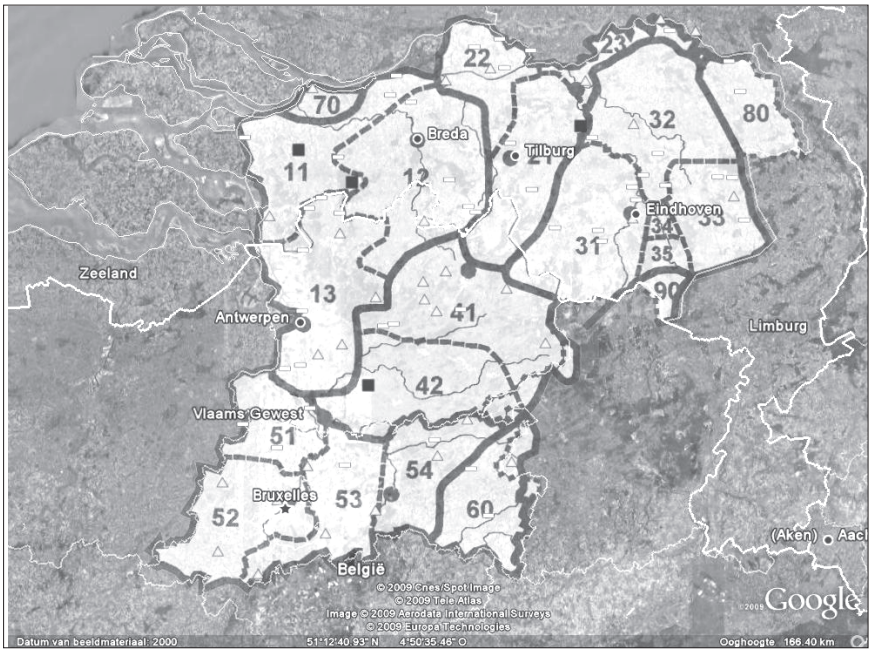
Opdeling van het resultaat van de clusteranalyse in slechts twee clusters levert duidelijk twee zwaartepunten op in het onderzoeksgebied; één cluster aan Nederlandse zijde van de staatsgrens en één cluster aan Vlaamse zijde. Beide clusters bevatten echter ook enkele plaatsen aan de andere kant van de staatsgrens en van een scherpe afbakening is dus geen sprake. Ook al is de rijksgrens van een jongere datum dan de vele oude dialectgrenzen, het is wel een belangrijke dialectgrens. De tweedeling van Brabant langs de rijksgrens zien we bijvoorbeeld ook terug in de analyses van Spruit (2008), welke op syntactische kenmerken gebaseerd zijn (de gegevens van de Syntactische Atlas van de Nederlandse Dialecten). Het is dus bemoedigend dat deze verdeling zich als eerste aandient. Vervolgens zijn we verder op gaan delen. Bij opdeling in 3 tot en met 8 clusters vallen er twee dingen op. Allereerst zien we steeds meer gebieden verschijnen die min of meer aansluiten bij de indeling van Belemans & Goossens (2000). Ten tweede zien we dat er telkens een cluster verschijnt dat bestaat uit exact dezelfde 53 plaatsen en dat niet aansluit bij Belemans & Goossens (2000) maar het gehele gebied overdekt. Vanaf opdeling in 9 clusters splitst dit gebiedsoverdekkende cluster zich in een groot gebiedsoverdekkend cluster van 49 plaatsen en een klein, maar tevens wijd verspreid, cluster van 4 plaatsen. Het resultaat van de opdeling in 9 groepen is weergegeven in kaart 3 en 4.[(2)] Kaart 3 toont de zes gebiedsvormende clusters bij deze opdeling en kaart 4 de drie gebiedsoverdekkende clusters. In beide kaarten zijn de clusters over de indeling van Belemans & Goossens (2000) heen geprojecteerd.

[(2)] De KML-bestanden voor kaart 3 tot en met 7 vindt u op de website http://dialect.ruhosting.nl/d2. Voor deze bestanden heeft u het programma Google Earth nodig dat gratis is te downloaden op http://earth.google.nl.



Kaart 3: Kaart op basis van de ruwe data; opdeling in 9 groepen, alleen de 6 gebiedsvormende groepen zichtbaar

We kunnen van de zes clusters in kaart 3 zeggen dat ze zeer globaal de indeling van Belemans & Goossens (2000) volgen. Ze tonen daarmee in ieder geval aan dat met onze methode op basis van lexicale gegevens gebiedsvorming voor Brabant is te vinden. Toch is het resultaat niet erg sterk. De zes clusters zijn erg diffuus en bevatten plaatsen die geografisch vaak ook erg ver uit elkaar liggen. Van lexicale grenzen hebben we in de inleiding gezegd dat ze meestal wel diffuus zijn, maar dit is wel erg extreem. Bovendien hebben we nog te maken met de drie clusters in kaart 4 welke zelfs het gehele gebied overdekken en waarvoor we dus helemaal geen geografische gebiedsvorming zien.



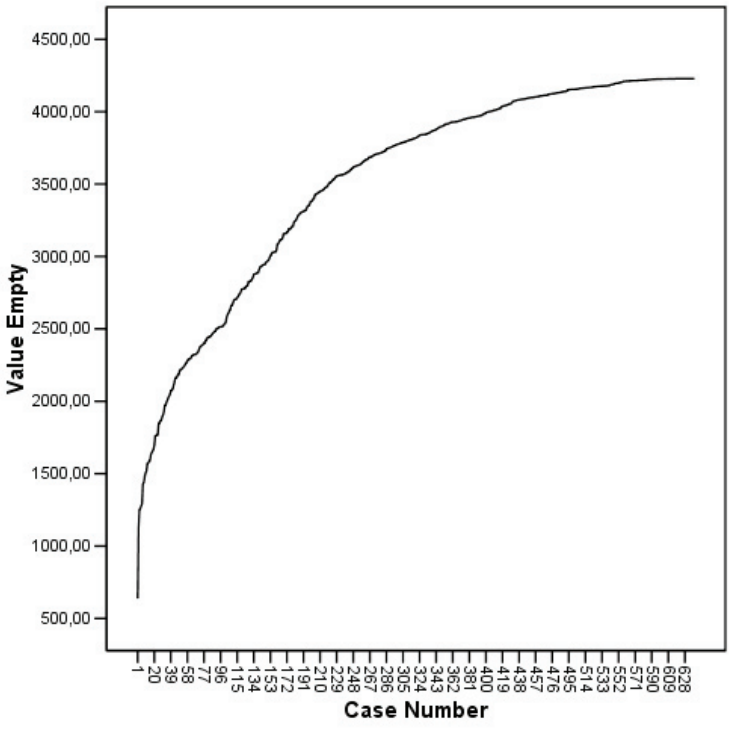
Kaart 4: Kaart op basis van de ruwe data; opdeling in 9 groepen, alleen de 3 niet-gebiedsvormende groepen zichtbaar

Omdat we vermoeden dat de tegenvallende resultaten die we zien in kaart 3 en 4 hun oorsprong vinden in de vele lege cellen en de manier waarop we die vervolgens hebben aangepakt met imputatie, proberen we de lege cellen in de volgende paragraaf op een fundamentelere manier aan te pakken.

5. Reductie van de ruwe data

In de vorige paragraaf hebben we de ontbrekende lexicale afstanden ingevuld op basis van de lexicale afstand tot geografisch nabije plaatsen. Een nadeel van imputatie op basis van geografie is dat we in zekere zin onze afstanden met oneigenlijke, want niet op lexicale gegevens gebaseerde, afstanden vervuilen en dat hierdoor de invloed van de geografie op onze indeling wordt versterkt. In deze sectie pakken we de ontbrekende afstanden op een fundamentelere manier aan. We reduceren het percentage ontbrekende gegevens drastisch door plaatsen en concepten uit de dataset te verwijderen waarvoor weinig of geen lexicale gegevens beschikbaar waren.

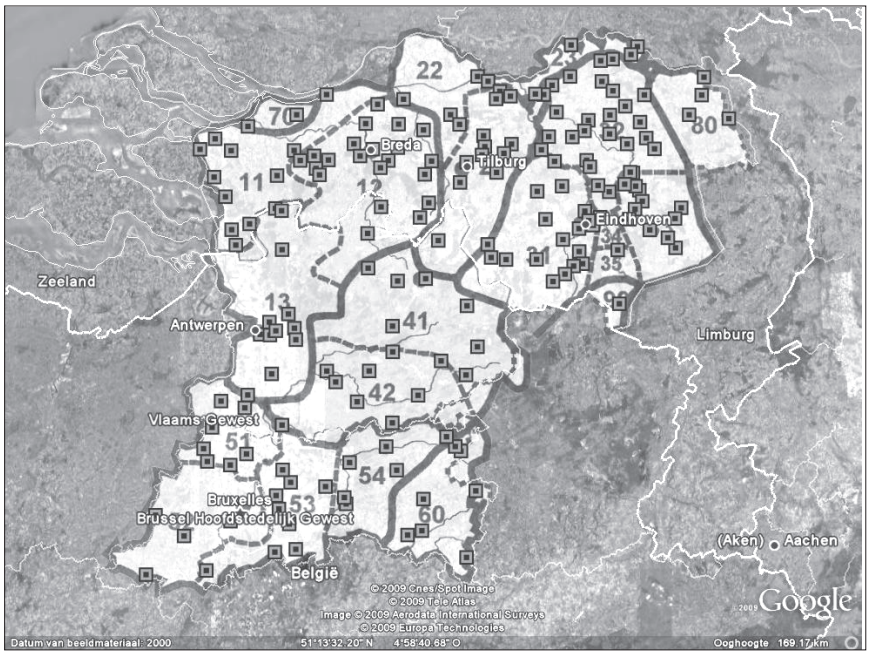
We zijn begonnen met de plaatsen. In figuur 5 is het aantal lege cellen (verticaal) uitgezet tegen de plaatsen (horizontaal). De plaatsen zijn oplopend genummerd van 1 tot en met 639, waarbij een hoger nummer een hoger aantal lege cellen inhoudt.



Figuur 5: Het aantal lege cellen met dialectgegevens afgezet tegen plaats

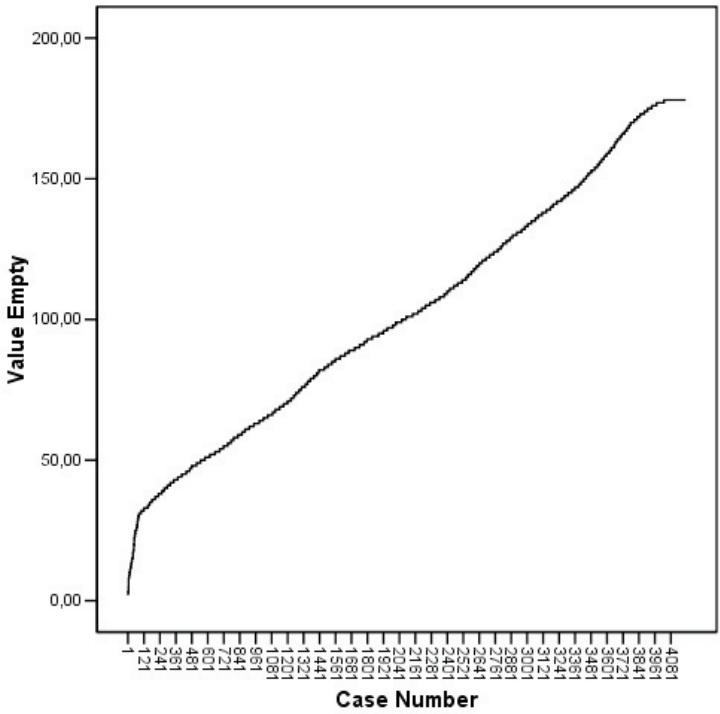
Figuur 5 laat een regelmatig stijgende curve zien, waaruit blijkt dat voor veel plaatsen relatief weinig gegevens beschikbaar zijn. De plaats met het maximum aantal lege cellen is Neerhespen (plaatsnummer 639) met 4228 lege cellen op een totaal van 4229 cellen. Voor deze plaats hebben we dus maar een enkele cel met gegevens in de dataset. De plaats met het kleinst aantal lege cellen is Roosendaal (plaatsnummer 1) met 638 lege cellen. Dit komt altijd nog neer op 15.1% lege cellen. We hebben besloten om verder te gaan in de analyse met die plaatsen waarvoor minstens 1000 gevulde cellen beschikbaar waren, omdat zo rond de waarde van 3200 lege cellen de stijging in het aantal plaatsen begint af te nemen (het duidelijkste omslagpunt ligt ongeveer bij 3500). Bij het criterium

van 3200 blijven er van de 639 plaatsen nog 179 over. Dat is een forse reductie, maar kaart 5 laat zien dat met deze 179 plaatsen de overdekking van het onderzoeksgebied nog steeds voldoende is.



Kaart 5: De geografische verdeling van de 179 plaatsen waarvoor 1000 of meer lexicale gegevens beschikbaar zijn

Vervolgens hebben we ons gericht op de concepten. Het totaal aantal concepten is na verwijdering van de 639-179 = 460 plaatsen geen 4229 meer, maar 4192. We hebben op basis van de nieuwe datamatrix (bestaande uit 179 bij 4192 cellen) bekeken hoeveel lege cellen er nog per concept aanwezig waren. Eén van de concepten met het grootst aantal lege cellen (178) was *zwavel*. Aangezien het onderzoeksgebied was teruggebracht tot nog maar 179 plaatsen, betekent dit dat we in onze datamatrix nog maar één opgave hebben voor dit concept. Het concept met het minst aantal lege cellen was *wieg* met maar twee lege cellen. In onderstaande grafiek is het aantal lege cellen (verticaal) uitgezet tegen de 4192 concepten (horizontaal). De concepten zijn oplopend genummerd van 1 tot en met 4192, waarbij een hoger nummer een hoger aantal lege cellen inhoudt.



Figuur 6: Het aantal lege cellen voor de 179 plaatsen afgezet tegen de concepten

Figuur 6 laat aan het begin een scherp oplopende lijn zien. Dat betekent dat er snel een groter aantal lege cellen is voor een groot aantal concepten. Het is moeilijk om ergens een grens te trekken, maar zo rond concept 500 loopt het aantal ontbrekende gegevens richting 50 en dat is rond de 30%. We hebben de dataset verder ingeperkt door alleen de 500 concepten te gebruiken met het minst aantal lege cellen. Deze 500 concepten bevinden zich aan de linkerkant van de grafiek in figuur 6 en hebben maximaal 48 lege cellen.

Door onze bewerkingen voor plaatsen en concepten is het totaal aantal lexicale vormen teruggebracht van 614.941 naar 117.286 en is de datamatrix teruggebracht tot een formaat van 179 x 500 = 89.500 cellen. Gemiddeld bevat elke gevulde cel 1,64 (117.286/71.362) lexicale vormen. Van de in totaal 89.500 cellen zijn er 18.138 leeg. Vergeleken met de totale ruwe dataset is het percentage lege cellen nu teruggebracht van 83.3% naar 20.3%. We hebben met onze bewerkingen de

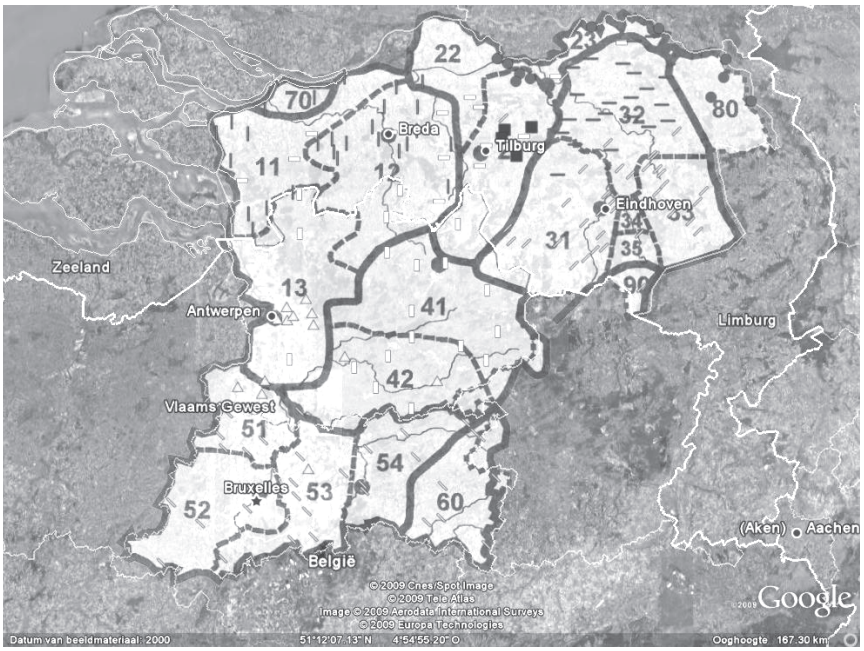
dataset dus zodanig ingeperkt dat we een veel beperkter percentage ontbrekende gegevens hebben, terwijl er nog altijd sprake is van een grote hoeveelheid gegevens en een redelijke spreiding van de plaatsen over het onderzoeksgebied.

Op basis van de 117.286 vormen zijn we vervolgens opnieuw afstanden gaan berekenen met de binaire maat en de GIW-maat. Onze criteria voor inperking blijken succesvol want het komt nu niet meer voor dat er voor een plaatspaar geen afstand berekend kan worden en dus hoeven er ook geen afstanden meer geïmputeerd te worden op basis van geografie. De *local incoherence* is voor de GIW-maat 1.60774 en voor de binaire maat 2.07088. Ook voor de gereduceerde dataset is GIW dus de meest succesvolle van de twee maten. Door vervolgens alleen vormen mee te nemen die minimaal 5 keer voorkomen in de dataset blijkt de *local incoherence* voor de GIW-maat nog verder omlaag te gaan van 1.60774 naar 1.42852 en wordt de dataset verder gereduceerd van 117.286 naar 100.277 lexicale vormen.

Met de op deze manier berekende afstanden zijn we vervolgens weer gaan clusteren. Dit doen we op dezelfde manier als in de vorige paragraaf; 100 keer met een ruiswaarde van 0,5 keer de standaarddeviatie van de afstanden. In kaart 6 is de opdeling van het clusterresultaat in negen groepen afgebeeld. Ook nu weer hebben we de clusters over de indeling van Belemans & Goossens (2000) heen geprojecteerd.

Duidelijk is te zien dat er gebiedsvorming is en dat de negen gebieden beter verdeeld zijn over het onderzoeksgebied dan bij de totale ruwe dataset het geval was. Geheel gebiedsoverdekkende clusters komen ook niet meer voor. Vervolgens bekijken we opnieuw de mate van aansluiten van de clusters bij de negen hoofdgebieden in de kaart van Belemans & Goossens (2000). Dit keer doen we dat wat uitgebreider en bespreken we de clusters een voor een, op volgorde van groot (38 plaatsen) naar klein (4 plaatsen).

Het eerste en grootste cluster op de kaart bevat 38 plaatsen en wordt weergegeven door schuine streepjes (van links onder naar rechts boven). Dit cluster beperkt zich tot de zuidelijke helft van het Oost-Noord-Brabants gebied (30), met uitzondering van drie plaatsen die net in het Midden-Noord-Brabants gebied (20) liggen.



Kaart 6: Kaart op basis van de gereduceerde dataset; opdeling in 9 groepen

Het tweede cluster, bestaande uit 31 plaatsen en weergegeven door een omgekeerd schuin streepje (van rechts onder naar links boven), beperkt zich tot het Zuid-Brabants gebied (50) en het Getelands (60), met uitzondering van Mechelen dat er tegenaan ligt maar door Belemans & Goossens (2000) tot het Zuiderkempens (42) wordt gerekend, een subgebied van het Kempens (40).

Het derde cluster bevat 27 plaatsen en is aangegeven met een verticaal streepje. Dit cluster bevindt zich binnen de grenzen van het Markizaats (11) en het Baronies (12), twee subgebieden van het Noordwest-Brabants (10).

Het vierde cluster van 22 plaatsen is gemarkeerd met een horizontaal streepje. Dit cluster vult het eerste cluster (schuine streepjes) aan, want het bestrijkt de noordelijke helft van het Oost-Noord-Brabants gebied (30). Daarnaast heeft het nog een plaats in het aansluitende Maaslands (23).

Het vijfde cluster bevat 19 plaatsen, is gemarkeerd met rechtopstaande rechthoeken en verspreidt zich over het Kempens (40) en het zuiden van het Noord-

west-Brabants (10).

Het zesde cluster wordt aangegeven door liggende rechthoeken en laat een patroon zien dat zich minder aan de indeling van Belemans & Goossens (2000) houdt. De 14 plaatsen centreren zich rond Tilburg in het Midden-Noord-Brabants (20) maar liggen daarbuiten in een vrij rechte lijn vanaf Bergen op Zoom in het uiterste Westen van het Noordwest-Brabants (10) tot Oss dat aan de Noordelijke grens van het Oost-Noord-Brabants (30) ligt. Om precies te zijn bevat het cluster plaatsen in de subgebieden Markizaats (11), Baronies (12), Tilburgs (21), Hollands Brabants (22), Maaslands (23) en Noord-Meierijs (32). Het cluster doorkruist dus ook verschillende van de andere clusters.

Het zevende cluster bestaat uit 13 plaatsen en wordt weergegeven met driehoeken. De belangrijkste concentratie van plaatsen in dit cluster ligt in het Antwerps subgebied (13). Verder houdt ook dit cluster zich iets minder aan de indeling van Belemans & Goossens (2000), met plaatsen in zowel het Noordwest-Brabants (10), het Zuid-Brabants (50) als het Kempens (40).

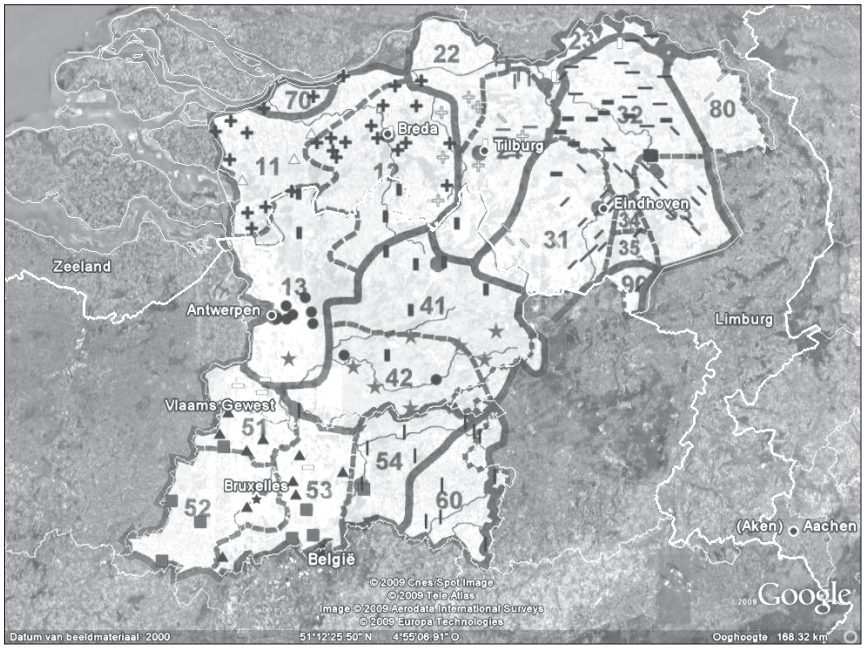
Het achtste cluster bestaat uit 11 plaatsen en is weergegeven met cirkels. Het gebied bevat plaatsen in 80 en het noorden van de drie subgebieden van het Midden-Noord-Brabants (20). Het gebied lijkt hierbij sterker aan te sluiten bij de stroomgang van de rivier de Maas dan bij de indeling van Belemans & Goossens (2000).

Het negende en kleinste cluster, met maar vier plaatsen, wordt weergegeven door vierkanten en is midden in het Tilburgs gebied (21) gesitueerd, een subgebied van het Midden-Noord-Brabants (20).

De clusters zijn dus allemaal mooi gebiedsvormend en sluiten overwegend vrij goed aan bij de negen hoofdgroepen in Belemans & Goossens (2000). Wel zijn er nog clusters (met name de liggende rechthoeken) die zich duidelijk niet houden aan Belemans & Goossens (2000) en een sterke menging laten zien.

Omdat sommige van de clusters ook aan subgebieden van de negen hoofdgebieden waren toe te kennen gaan we vervolgens nog een stap verder. We hebben bekeken of ook een verdere opdeling van het resultaat van de clusteranalyse aansluit bij Belemans & Goossens (2000). De negen hoofdgebieden bevatten in totaal eenentwintig subgebieden. Om die reden hebben we ook het resultaat

van de clusteranalyse opgedeeld in eenentwintig clusters. De bijbehorende kaart is afgebeeld als kaart 7.



Kaart 7: Kaart op basis van de gereduceerde dataset; opdeling in 21 groepen

Ook al is er geen sprake van een verregaande overeenkomst met de 21 subgebieden van Belemans & Goossens (2000), het resultaat vertoont op veel plaatsen nog altijd een duidelijke overlap. Verder valt op dat er nu effecten van de verzamelmethode zichtbaar beginnen te worden. Zo zijn bijvoorbeeld de twee kleinste clusters (Aarle en Rixtel, Beek en Donk) het gevolg van de beslissing van de WBD-redactie om voor de ene plaats uit het cluster de data te gebruiken die voor de andere plaats uit het cluster was verzameld. Dit is destijds gedaan omdat de kernen van deze plaatsen al lange tijd geleden zijn samengesmolten, zodat er vanuit kon worden gegaan dat de gegevens voor de ene plaats ook gelden voor de andere. Wanneer we teruggaan naar kaart 6 dan kunnen we nu ook het kleinste cluster aldaar verklaren: twee van de vier plaatsen in dat cluster zijn Berkel en Enschot. Ook voor deze plaatsen geldt dat de kernen al lange tijd geleden zijn samengesmolten.

6. Conclusies

In deze bijdrage hebben we gekeken naar lexicale gegevens uit het WBD. Kernvraag 1 was hoe ruw de gegevens zijn. In 4.1 zagen we aan het convergerende karakter van de steekproeven dat we pas bij grotere steekproeven betrouwbare uitkomsten kunnen verwachten. De ruwheid van het materiaal wordt voor een belangrijk deel bepaald door de hoeveelheid ontbrekende gegevens, een typische eigenschap van de zuidelijke woordenboeken. Omdat de methode die we gebruikten hier niet mee overweg kan hebben we de ontbrekende lexicale afstanden die hier het gevolg van zijn eerst aangepakt door te imputeren op basis van geografische gegevens. Het resultaat van de clusteranalyse toonde aan dat we voor Brabant op basis van lexicale gegevens in ieder geval gebiedsvorming kunnen vinden (kernvraag 2). Maar het resultaat was nog niet erg bevredigend omdat de kaart grote menggebieden bevatte en ook enkele clusters die zelfs het gehele gebied overdekten. Omdat we het vermoeden hadden dat de tegenvallende resultaten hun oorsprong hebben in de vele ontbrekende gegevens en de manier waarop we die vervolgens hebben aangepakt met imputatie, hebben we deze vervolgens op een meer fundamentele manier aangepakt. Concepten en plaatsen met veel ontbrekende gegevens hebben we uit de dataset verwijderd. Hierdoor bleek imputatie voor de afstandsmatrix niet meer nodig en de uitkomst van de clusteranalyse een veel beter kaartbeeld op te leveren. Deze kaart sloot op veel punten aan bij de indeling van Belemans & Goossens (2000) en bevatte geen clusters meer die het gehele gebied overdekten (kernvraag 3).

Concluderend kunnen we stellen dat de data van het WBD door het grote percentage ontbrekende gegevens problemen oplevert voor de clusteranalyse en voor het vinden van een gedetailleerde indeling. Door de data volgens weloverwogen criteria sterk te reduceren blijkt het echter toch mogelijk om van de ruwe lexicale gegevens van het WBD een gedetailleerde indeling van het Brabants af te leiden, welke sterke overeenkomsten vertoont met de indeling van de Brabantse dialecten van Belemans & Goossens (2000).

Voor toekomstig onderzoek zou het interessant zijn om verklaringen te vinden voor juist die gebieden die duidelijk afwijken van Belemans & Goossens (2000). Zo zagen we in kaart 6 een cluster dat zich breed verspreide over subgebieden van het Noordwest-Brabants en het Midden-Noord-Brabants en dat ontbreekt in de kaart van Belemans & Goossens (2000). Ook zouden we de dataset nog verder kunnen polijsten in de hoop een nauwkeuriger indeling te krijgen. Naast ontbrekende gegevens kent de dataset nog meer ruwe eigenschappen welke

van invloed kunnen zijn maar waar we verder niet op in hebben kunnen gaan. Zo komt het veelvuldig voor dat er voor een plaats juist méérdere vormen zijn opgetekend voor een concept. Doordat elk van de vormen even zwaar meetelt gaan plaatsen hierdoor meer op elkaar lijken. Ten slotte willen we nog wijzen op de vele morfologische onderscheiden en meerwoordsexpressies die het WBD bevat en welke in onze methode, strikt genomen ten onrechte, als afzonderlijke lexicale elementen zijn behandeld.

Bibliografie

Belemans, R. & J. Goossens
(2000). *Woordenboek van de Brabantse Dialecten. Deel III. Inleiding en Klankgeografie.* Assen, Van Gorcum.

Chambers, J.K. & P. Trudgill
(1998). *Dialectology*. Cambridge, CUP.

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

De Vriend, F. & J. Swanenberg
(2006). D-kwadraat: digitale databanken en digitaal gereedschap voor WBD en WLD. In: *Nederlandse Taalkunde* 11, 366-372.

De Vriend, F., L. Boves, H. van den Heuvel, R. van Hout, J. Kruijsen & J. Swanenberg.
(2006). A Unified Structure for Dutch Dialect Dictionary Data. In: *Proceedings of The fifth international conference on Language Resources and Evaluation* (LREC 2006, Genoa). Paris, European Language Resources Association, 1660-1665.

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

Heeringa, W.
(2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen: dissertatie Rijksuniversiteit Groningen.

Heeringa, W & J. Nerbonne
(2006). De analyse van taalvariatie in het Nederlandse dialectgebied: methoden en resultaten op basis van lexicon en uitspraak. In: *Nederlandse Taalkunde* 11, 218-257.

Jain, A.K. & R.C. Dubes
(1988). *Algorithms for Clustering Data*. Englewood Cliffs NJ, Prentice Hall.

Jain, A.K., M.N. Murty & P.J. Flynn
(1999). Data clustering: A review. In: *ACM Computing Surveys* 31, 264–323.

Kleiweg, P., J. Nerbonne & L. Bosveld
(2004) Geographic Projection of Cluster Composites. In: A. Blackwell, K. Marriott and A. Shimojima, *Diagrams 2004. Lecture Notes in Computer Science*, Berlin, Springer-Verlag, 392-394.

Kloeke, G.G. & L. Grootaers
(1934). *Dr. L. Grootaers' en Dr. G.G. Kloeke's systematisch en alfabetisch register van plaatsnamen voor Noord-Nederland, Zuid-Nederland en Fransch-Vlaanderen*. 's-Gravenhage, Nijhoff.

Lontie, R.
(1923). *Het dialect van Lubbeek en dialectgeographie van West-Hageland.* Leuven, Licentiaatsverhandeling KU Leuven.

Nerbonne, J., P. Kleiweg, W. Heeringa & F. Manni
(2008) Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: C. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker, *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society* Berlin, Springer, 647-654.

Pauwels, J.L.
(1958). *Het dialect van Aarschot en omstreken*. Tongeren, Belgisch Interuniversitair Centrum voor Neerlandistiek.

RUG/L04, Software for dialectometrics and cartography. http://www.let.rug.nl/~kleiweg/L04/.

Spruit, M.
(2008). *Quantitative perspectives on syntactic variation in Dutch dialects.* Utrecht, LOT.

WBD
(1967-2005). *Woordenboek van de Brabantse Dialecten*. Assen, Van Gorcum, Amsterdam, Gopher.

Weijnen, A.A.
(1937). *Onderzoek naar de dialectgrenzen in Noord-Brabant in aansluiting aan geografie, geschiedenis en volksleven.* Fijnaart, dissertatie K.U. Nijmegen.

WLD
(1983-2008). *Woordenboek van de Limburgse Dialecten*. Assen, Van Gorcum, Amsterdam, Gopher.

Taal & Tongval themanr. 20 (2007), p. 83-110