Dirk Speelman & Dirk Geeraerts

De structuur van lexicale onzekerheid

Abstract

That an assumption of language-internal homogeneity is misleading when it is applied to the formal side of language has been established for quite a while now by sociolinguistics: even within linguistic communities, there is language variation, and it is structured along social lines. With regard to the conceptual side of language, however, it is only with the recent advent of Cognitive Linguistics and prototype-based models of semantic structure that linguists have become sensitive to the fact that meanings are fuzzy and flexible entities, knowledge of which may not be evenly distributed over the linguistic community.

So what effect could this conceptual heterogeneity have on the interpretation of dialectological data? In particular, can it be established that lesser known concepts exhibit more lexical variation, as a reflection of more insecurity and hesitation on the part of the informants? In other words, can we detect a negative correlation between conceptual entrenchment and lexical uncertainty?

In an initial step towards an answer to the question, we analyse the dialectological data of the Woordenboek van de Limburgse Dialecten, focusing on the lexical field of the body. We investigate to what extent lexical heterogeneity is determined by the entrenchment of the concepts in question. Using the number of different lexical types that are mentioned for a given concept as a dependent variable, we are able to establish that conceptual factors (like entrenchment, but also other semantic factors like the taboo character of the concepts) do indeed correlate with lexical uncertainty.

Taal & Tongval themanr. 20 (2007), p. 47-62

1. Vraagstelling

Wat hebben de omvangrijke dialectwoordenboeken van het Nederlands te bieden voor de lexicoloog en de lexicaal semanticus? Die vraag is niet nieuw: reeds in 1991 werd een van de Taal en Tongval-symposia aan die vraag gewijd, en ook toen werd reeds gesuggereerd dat de (toen nog recente) ontwikkelingen in de cognitieve taalkunde interessante onderzoeksmogelijkheden openden. Moerdijk & Geeraerts (1991: 66) formuleerden een centraal aspect van zo'n toepassing toen als volgt: "In het algemeen impliceert de psychologische oriëntatie van de semantiek dat het niet zal volstaan de lexicale categorieën van een taal te identificeren en te beschrijven, maar dat ook moet worden nagegaan wat de psychologische status van die categorieën is. In het bijzonder moet daarbij rekening worden gehouden met de verschillen in saillantheid die door de prototypetheorie … voor het voetlicht zijn gebracht." In het onderhavige artikel komen we terug op die vraag, maar waar Moerdijk & Geeraerts indertijd een gedetailleerde manuele analyse van het woordveld 'rund' in het Woordenboek van de Brabantse Dialecten als illustratie gebruikten, zullen we nu een ander methodologisch pad bewandelen. We zullen de vraag stellen of een grootschalige kwantitatieve analyse van het woordenboekmateriaal verschijnselen reveleert die aansluiten bij de prototypische betekenisopvatting, en we zullen daartoe een statistische analyse uitvoeren van de heterogeniteit van de naamgeving in de aflevering 'Het menselijk lichaam' van het Woordenboek van de Limburgse Dialecten.

In een ruimer verband sluit ons onderzoek aan bij de belangstelling die binnen de Cognitieve Linguïstiek begint te groeien voor de bestudering van taalinterne variatie (zie Geeraerts 2005, Kristiansen & Dirven 2008). Die voorzichtige opbloei van cognitief-linguïstisch taalvariatieonderzoek heeft echter een grotendeels sociolinguïstische en stilistische oriëntatie: ondanks studies als Swanenberg (2000), Nilsson (2001), Berthele (2002, 2004, 2006), Sharifian (2005), Szelid & Geeraerts (2008) blijft dialectologisch onderzoek marginaal binnen de Cognitieve Linguïstiek. Dat is jammer, want de inspiratie kan wederzijds zijn. Enerzijds roepen dialectologische gegevens de vraag op of taalinterne verschillen net zo goed als verschillen tussen talen gepaard kunnen gaan met verschillen in conceptualisatie en categorisatie. Anderzijds betekenen de taalgebruiksgeoriënteerde methodologische uitgangspunten van de Cognitieve Linguïstiek een uitdaging voor de traditionele methodologische focus van de dialectologie op taalstructuur veeleer dan taalgebruik.

Taal & Tongval themanr. 20 (2007), p. 47-62

We zullen nu eerst de vraagstelling van ons onderzoek stapsgewijze opbouwen; in de volgende paragrafen gaan we dan achtereenvolgens in op het design van de studie en op de resultaten. Om te beginnen is onze vraag, zeer in het algemeen, in welke mate onomasiologische variatie beïnvloed wordt door conceptkenmerken, en niet slechts door taalkenmerken, in het bijzonder door de geografische identiteit van het taalsysteem. Het ligt voor de dialectologie voor de hand om heteronymie te verklaren op geografische gronden, bijvoorbeeld op basis van contactverschijnselen tussen geografisch afgebakende taalsystemen. Maar zou de lexicale verscheidenheid niet ook gerelateerd kunnen zijn aan de kenmerken van de concepten zelf?

We kunnen in dat verband wijzen op een dubbele veronderstelling van homogeniteit in een klassieke structuralistische opvatting van talen. In het structuralistische ideaaltype van een taalsysteem is zowel de vorm als de betekenis van het taalteken stabiel en uniform; variatie en verandering hebben te maken met de overgang van de ene systeemtoestand naar de andere, maar zijn als zodanig geen onderdeel van het systeem zelf. De eerste veronderstelling is echter 40 jaar terug door de sociolinguïstiek ontkracht: ook binnen een taalsysteem bestaat gestructureerde variatie in de vorm van het taalteken. De tweede veronderstelling is sinds 20 jaar door de cognitieve semantiek ontkracht: betekenissen zijn ook binnen een taalsysteem flexibel en variabel. Dat de dialectologie, in aansluiting bij de ontkrachting van die eerste veronderstelling, verrijkt moet worden met een sociolinguïstisch perspectief zal wel door niemand meer betwist worden. Maar onze vraag is nu, in aansluiting bij de ontkrachting van die tweede veronderstelling: moeten we de dialectologie niet ook verrijken met expliciete aandacht voor de semantische flexibiliteit en variabiliteit van talen?

We kunnen onze initiële vraagstelling nu herformuleren: in welke mate wordt onomasiologische variatie beïnvloed door heterodoxe conceptkenmerken, van het type waar de cognitieve semantiek graag op wijst? De cognitieve semantiek legt immers de klemtoon op een aantal kenmerken die tot uiting kunnen komen in lexicale variabiliteit: betekenissen hebben onderling niet noodzakelijk scherpe grenzen (lexicale concepten kunnen vaag zijn), en betekenissen kunnen meer of minder saillant zijn (sommige concepten zijn mentaal sterker verankerd dan andere). Beide kenmerken suggereren iets over te verwachten onomasiologische variatie. Ten eerste, je kunt verwachten dat vage concepten eerder aanleiding geven tot lexicale heterogeniteit dan concepten waarover geen potentiële verwarring bestaat. Ten tweede, je kunt verwachten dat minder sterk verankerde concepten

eerder aanleiding geven tot lexicale heterogeniteit dan saillante concepten.

We kunnen daarbij in het midden laten of die heterogeniteit een enquête-effect is dan wel een structurele eigenschap. Concepten die een minder scherpe afbakening kennen en concepten die de taalgebruiker minder scherp voor de geest staan, zullen wellicht makkelijker aanleiding geven tot verwarring of overlapping met andere concepten, maar die verwarring en overlapping kunnen zowel incidenteel als structureel zijn.

2. Design: de verklarende variabelen

Hoe zullen we nu onze vraag proberen te beantwoorden? Er zijn verschillende aspecten aan het design van deze studie. Welk materiaal gebruiken we? Hoe meten we de saillantheid van concepten? Hoe meten we de vaagheid van concepten? Zijn er nog conceptkenmerken om rekening mee te houden? En hoe meten we de heterogeniteit van de naamgeving?

Wat het *materiaal* betreft, we gebruiken de aflevering "Het menselijk lichaam" van het Woordenboek van de Limburgse Dialecten, met een beperking tot de recent systematisch afgenomen enquêtes N10, N106, N107, N108 en N109 (dus met uitsluiting van het materiaal met een beperkte geografische spreiding en/of een oudere herkomst). We doen dit om de kans te maximaliseren dat alle concepten die we analyseren met hetzelfde geografische bereik geënquêteerd werden. Het ontbreken van een opgave voor een bepaald concept op een bepaalde plaats kan immers in principe twee dingen betekenen: het concept was in de enquête opgenomen maar was de informanten uit die plaats niet bekend, of het concept werd op die plaats helemaal niet bevraagd. Om die dubbelzinnigheid uit te sluiten, vertrekken we van de enquêtes die in principe dezelfde reeks plaatsen voor alle concepten bestrijken. Dat levert ons een database op met 201 plaatsen, 206 concepten, en 32591 tokens van lexicale elementen.

Wat de *saillantheid* van de concepten betreft, die meten we aan de hand van drie verschillende factoren. We gebruiken de volgende drie verklarende variabelen als indicatoren van (niet-)saillantheid: de ongebruikelijkheid van het concept, het aantal plaatsen zonder antwoorden, en het aantal meerwoordige antwoorden. We zullen deze drie factoren nu achtereenvolgens toelichten. We gaan telkens in op de onderliggende logica die de opname van de factor motiveert en op de wijze waarop de variabele geoperationaliseerd zal worden; daarbij geven we dan ook enkele voorbeelden.

De *ongebruikelijkheid* van een concept willen we meenemen in de analyse omdat minder gebruikelijke concepten de kans op onzekerheid bij de taalgebruikers kunnen verhogen, en dus ook de kans op uniformering over gebruikers en dialecten heen kunnen verminderen. Ongebruikelijkheid hebben we geoperationaliseerd aan de hand van een beperkte enquête onder 7 leden van onze onderzoeksgroep, aan wie we gevraagd hebben om aan de 206 concepten uit de materiaalverzameling een gebruikelijkheidsscore toe te kennen op een vijfpuntsschaal. De resultaten van deze enquête zijn plausibel: aan de onderkant van de schaal vinden we concepten als KNOKKELKUILTJES, BLOEDWEI, LEVEND VLEES ONDER DE HUID, VOORVOET, AFHANGEND KUIFJE (BIJ KORTGEKNIPT HAAR) in contrast met KEEL, KNIE, MIDDELVINGER etc.

Het *aantal plaatsen zonder antwoorden* voor een gegeven concept is interessant omdat het aantal blanco antwoorden een indicatie kan zijn van de onbekendheid van het concept. We moeten er inderdaad rekening mee houden dat de dekking van concepten over plaatsen niet zo systematisch is als wat we met de keuze voor de systematisch afgenomen enquêtes N10, N106, N107, N108 en N109 hadden willen bereiken: ook in dit materiaal zijn nog hiaten te vinden. Volgens de logica die we voor het design van het onderzoek hanteren, zouden we die hiaten kunnen interpreteren als indicaties van onbekendheid, maar een gebrekkige of onvolledige enquêtering kunnen we natuurlijk niet uitsluiten. Voorzichtigheid bij de interpretatie van deze variabele is dus geboden. We hebben de factor geoperationaliseerd aan de hand van het aantal plaatsen (op 201 plaatsen) waarvoor het concept niet geattesteerd werd. Concepten met weinig antwoorden zijn bv. SLECHT GROEIEN, GELUIDLOZE WIND, KAAKGESTEL EN HUIG.

Opname van het *aantal meerwoordige antwoorden* wordt gemotiveerd door de vaststelling dat basisconcepten, wanneer ze over een hele reeks talen heen bekeken worden, meestal met kortere lexicale elementen aangeduid worden: dat is in het onderzoek naar de 'basic level hypothesis' terug te vinden. Meerwoordige antwoorden kunnen bovendien wijzen op verlegenheidsopgaven, en het aantal verlegenheidsopgaven kunnen we natuurlijk (zoals het aantal blanco antwoorden) interpreteren als een gevolg van een lage conceptsaillantheid. Hoewel we in principe ook het onderscheid tussen simplicia en gelede woorden in het onderzoek zouden kunnen betrekken, beperken we ons voor de operationalisering tot het onderscheid tussen afzonderlijke woorden (geleed of niet) en meerwoordige uitdrukkingen. Voor de operationalisering kiezen we voor de proportie (op tokenniveau) van meerwoordige antwoorden in de totale set van antwoorden voor een

concept. Dergelijke meerwoordige antwoorden voor het concept BORSTELIG HAAR zijn bijvoorbeeld *haar wie een stekelvarken, haar wie stro, steil haar, stijf haar, ruspeltig haar, pekelig haar, weers haar, stekkerig haar.*

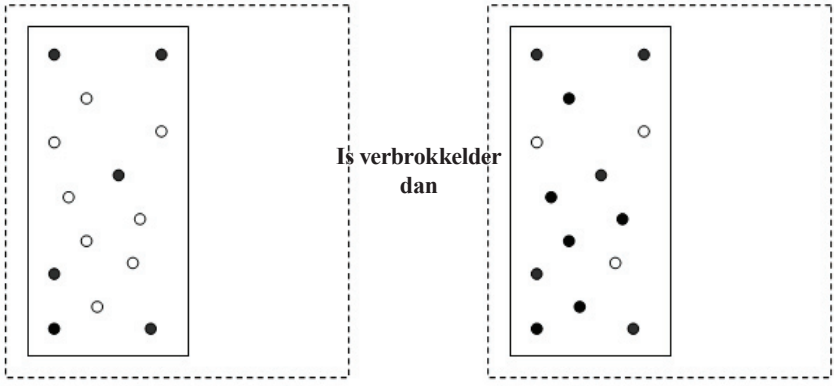
Naast de saillantheid van de concepten willen we ook kijken naar de vaagheid van de begrippen. Hiervoor gebruiken we slechts één indicatie, nl. de mate van lexicale niet-uniekheid van de concepten, d.w.z. de mate waarin ze benoemd worden met lexicale elementen die ook bij andere concepten een rol blijken te spelen. De motiverende logica is deze: naarmate de woorden die naar een doelconcept verwijzen, ook voorkomen in de naamgeving van andere concepten, is het doelconcept minder makkelijk van andere concepten te onderscheiden. We operationaliseren de factor aan de hand van het aantal attestaties (op type-niveau) van het gebruik van een term voor het benoemen van een ander concept. Iets formeler gesteld gaat het om het aantal verschillende koppels (N,C) waarbij C een ander concept is dan het onderhavige concept en waarbij het woordtype N niet alleen geattesteerd is voor het benoemen van het onderhavige concept, maar ook voor het benoemen van C. Een alternatieve operationalisering op token-niveau is vanzelfsprekend niet uitgesloten, maar die zullen we niet meenemen in de hier gepresenteerde initiële analyse; in vervolgonderzoek zullen we deze alternatieve invulling wel onderzoeken. Voor het concept LIES vinden we de vormen *lies, de dun, vlim, lende, liest, hees, lee, liesje, var*. Vijf van deze negen types komen ook bij andere begrippen voor: lies bij BEKKENHOLTE, vlim bij WIMPER, lende en lee bij LENDE, en hees bij KNIEHOLTE. Daarom krijgt dit concept de waarde vijf voor lexicale niet-uniekheid. Dit voorbeeld laat zien dat de lexicale overlapping vermoedelijk niet altijd ook een semantische overlapping inhoudt: de aanwezigheid van *wimper* zal eerder als homonymie dan als vaagheid of polysemie geïnterpreteerd moeten worden. We hebben echter, gezien de omvang van de database, geen poging gedaan deze gevallen van homonymie weg te zuiveren; hiermee zal natuurlijk wel rekening moeten worden gehouden bij de interpretatie van de variabele.

Naast saillantheid en vaagheid, als bij uitstek cognitief-semantische factoren, zouden nog andere, meer traditionele conceptkenmerken een rol kunnen spelen bij de heterogeniteit van de naamgeving. Dat geldt m.n. voor de *gevoelswaarde* van de concepten: taboebegrippen vertonen meestal een ruime synonymie en met taboe moeten we zeker rekening houden in het geval van het menselijk lichaam, met domeinen als voortplanting en ontlasting. We beperken ons overigens niet tot taboe, maar breiden de categorie uit naar negatief affect in het algemeen: het

Woordenboek van de Limburgse Dialecten bevat vragen waarin expliciet naar spotnamen gevraagd wordt, bijvoorbeeld bij NEUS, GEZICHT, HOOFD. Ook deze vorm van negatieve gevoelswaarde nemen we mee in het onderzoek. We hebben ook deze factor geoperationaliseerd met behulp van een enquête onder 7 leden van de onderzoeksgroep, waarbij op een vijfpuntsschaal scores toegekend moesten worden voor de negatieve lading van de begrippen. En ook hier zijn de resultaten plausibel. Negatief gewaardeerd worden bijvoorbeeld de begrippen AARSSPLEET, GELUIDLOZE WIND, KWIJL, PAPPERIG PERSOON.

3. Design: de responsvariabele

Welke verschijnselen kunnen nu, aan de kant van de afhankelijke variabelen, wijzen op heterogeniteit? We willen daarbij twee dingen in rekening brengen: lexicale *diversiteit*, gedefinieerd als het bestaan van verschillende woorden voor de naamgeving van een concept, en geografische *verbrokkeling*, gedefinieerd als niet-homogeniteit in de geografische distributie van die verschillende woorden. We definiëren *heterogeniteit* als het product van lexicale diversiteit en geografische verbrokkeling, maar uiteraard hebben we eerst nog een operationalisering nodig van deze beide factoren. Voor diversiteit is dat eenvoudig: we tellen het aantal verschillende types in de naamgeving voor een concept.

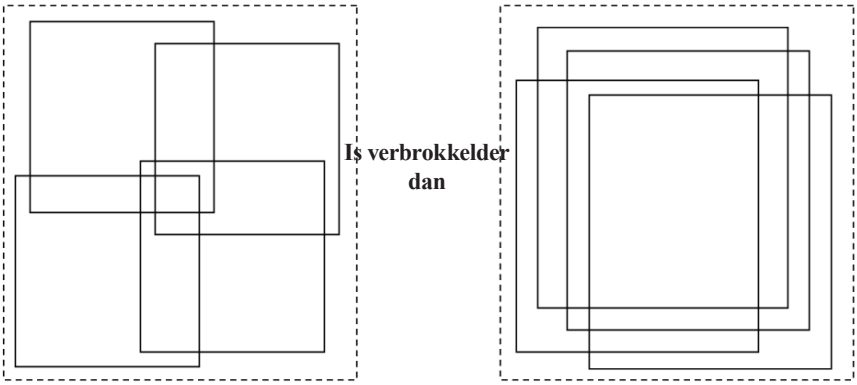


Figuur 1. Schematische voorstelling van dispersie

Voor de meting van verbrokkeling moeten we echter nog enkele extra stappen nemen. Stellen we om te beginnen vast dat verbrokkeling enerzijds te maken met de *dispersie* van de geografische distributie van de termen (gaten in de spreiding wijzen op een grotere verbrokkeling) en anderzijds met het *bereik* van

de termen (een gemiddeld kleiner bereik wijst op grotere verbrokkeling). Figuur 1 laat zien wat we met dispersie bedoelen: een geringere of grotere dekking van het aantal meetpunten binnen een gegeven geografisch gebied. De stippellijn stelt het hele onderzochte gebied (dat van de Limburgse dialecten) voor. De volle lijn geeft aan in welk deelgebied daarvan een bepaald concept voorkomt. De zwarte punten zijn plaatsen waar voor het concept in kwestie een opgave aanwezig is; de open punten zijn de geografische plaatsen waarvoor dat niet zo is. De situatie in de rechterhelft van het schema, met een grotere hoeveelheid meetpunten mét attestaties voor een bepaalde term, is minder verbrokkeld dan die in de linkerhelft van het schema.

In Figuur 2 volgt een schematische voorstelling van wat we met 'bereik' bedoelen. We kijken naar het maximale verspreidingsgebied van de verschillende termen die voor een gegeven concept aanwezig zijn. In de situatie in de linkerhelft van het schema hebben de termen ieder een beperkt verspreidingsgebied; in de situatie rechts vinden we hetzelfde aantal termen, maar met een groter bereik. De situatie rechts beschouwen we als minder verbrokkeld dan de situatie links.



Figuur 2. Schematische voorstelling van geografisch bereik

Hoe implementeren we deze maten nu in de praktijk ? Dispersie drukken we uit aan de hand van een verhouding tussen twee gemiddelde kortste afstanden: enerzijds de gemiddelde afstand van elke attestatie van een bepaalde term voor een bepaald concept tot de dichtstbijzijnde andere attestatie van dezelfde term voor hetzelfde concept, en anderzijds de gemiddelde afstand van elke attestatie van een bepaalde term voor een bepaald concept tot de dichtstbijzijnde andere attestatie van het concept als dusdanig, ongeacht de betrokken termen. Wanneer

we deze verhouding tussen beide gemiddelde afstanden voor één enkele term berekenen, door de eerste gemiddelde afstand door de tweede te delen, dan hebben we een maat voor de dispersie van die term. Hoe groter de verkregen ratio, hoe groter de dispersie: er zitten tussen de attestaties van de term veel witte vlekken.

Maar de dispersie van een enkele term zegt nog niet veel over de heterogeniteit in de naamgeving van het concept als geheel, en daarom moeten we de procedure herhalen voor alle termen waarmee een concept voorkomt. Van de dispersiegraad voor ieder van die termen nemen we vervolgens nog het gemiddelde, zij het dan wel een gewogen gemiddelde. De wegingsfactor die we toepassen is het gewicht van een term in het onomasiologische profiel van het concept, waarbij het onomasiologische profiel het geheel van de termen is waarmee een bepaald concept benoemd wordt, gedifferentieerd volgens hun relatieve frequentie (zie Geeraerts, Grondelaers & Speelman 1999, en Speelman, Grondelaers & Geeraerts 2003). Op die manier garanderen we dat termen die hoogfrequent zijn in de naamgeving van een concept, ook sterker doorwegen bij de berekening van de dispersiegraad van dat concept.

Bij deze procedure kunnen we nog opmerken, ten eerste, dat de berekeningswijze relatief is. Doordat we rekening houden met de gemiddelde afstand tussen de attestaties van het concept, maakt het niet uit of we met een dun- of dichtbevolkt gebied te maken hebben. Dat is uiteraard wat we willen: de dispersie die we berekenen zou niet beïnvloed mogen worden door de vraag of de meetpunten in een bepaald gebied toevallig dicht bij elkaar liggen of niet. Ten tweede, de dispersiemaat is onafhankelijk van het aantal verschillende lexicale types waarmee een concept benoemd wordt. Doordat we het (gewogen) gemiddelde nemen van de dispersiegraden van de afzonderlijke termen, maakt het niet uit of we voor een bepaald concept veel of weinig verschillende types hebben. Ook dat is een gewenst resultaat: de dispersie die we berekenen zou niet beïnvloed mogen worden door de vraag of voor een gegeven concept veel of weinig typevariatie bestaat. Die hebben we nl. al in onze heterogeniteitsberekening opgenomen onder de noemer ‘diversiteit’, en we willen met dispersie een vorm van verbrokkeling meten die onafhankelijk is van die diversiteit.

In vergelijking met de berekening van de dispersiemaat is het bepalen van de maat voor het bereik van de concepten relatief eenvoudig. We maken daarbij gebruik van het maximale oppervlak dat het verspreidingsgebied van een term

inneemt, en we vergelijken dat met het verspreidingsgebied van het concept. Voor een gegeven term berekenen we het bestreken oppervlak aan de hand van de lengte- en de breedtegraad van de meest extreme meetpunten waarin die term voorkomt, en we berekenen vervolgens de verhouding tussen dat oppervlak en het verspreidingsgebied van het concept als geheel, ook weer vertrekkend van de lengte- en de breedtegraad van de meest extreme meetpunten waarin het concept voorkomt. Die verhouding tussen oppervlaktes bepalen we nu voor alle termen die het begrip benoemen, en van die verhoudingen nemen we dan het gewogen gemiddelde (met dezelfde wegingsfactor als bij de dispersiemaat).

Anders dan de dispersiemaat is de bereikmaat omgekeerd evenredig met de verbrokkeling van een concept, zoals een blik op Figuur 1 en Figuur 2 kan leren: als het gemiddelde door een term bestreken oppervlak groter wordt, is de verbrokkeling kleiner, maar de verbrokkeling stijgt daarentegen als de gemiddelde afstand tussen de attestaties van een term toeneemt. Verbrokkeling kunnen we daarom definiëren als de verhouding tussen dispersie en bereik. Voor de heterogeniteit als geheel komen we dan uit bij de volgende formule:

$$heterogeniteit = diversiteit \times \frac{dispersie}{bereik}$$

3. Analyse

We onderwerpen de relatie tussen de vijf verklarende variabelen en heterogeniteit als responsvariabele aan een meervoudige lineaire regressie. De vijf factoren zijn met de volgende labels in de resultaten van de regressieanalyse terug te vinden:

ongebruikelijkheid	ongebruikelijkheid
ontbrek.plaatsen	aantal ontbrekende plaatsen
meerwoord.antw	aantal meerwoordige antwoorden
niet.uniekheid	lexicale niet-uniekheid
negatief.affect	negatief affect

Voor we de resultaten bekijken, willen we volledigheidshalve wijzen op enkele technische ingrepen die we hebben doorgevoerd. (Deze alinea is alleen relevant voor wie enigszins vertrouwd is met de technische facetten van een regressieanalyse.) Om te beginnen hebben we als responsvariabele om analysetechnische

redenen (niet normaal verdeelde residuwaarden) niet de pure heterogeniteit, maar wel de log van de heterogeniteit genomen. Ten tweede, om gevallen van extreme dataschaarste te vermijden hebben we de analyse uitgevoerd op basis van de concepten die op minstens tien plaatsen geattesteerd zijn; hieraan voldoen slechts 186 van de 206 concepten. Ten derde, er moet melding gemaakt worden van twee interacties in de invloed van de bestudeerde verklarende variabelen op heterogeniteit: er is een interactie tussen niet-uniekheid en ongebruikelijkheid en een interactie tussen niet-uniekheid en negatief affect. De eerste interactie bestaat erin dat ongebruikelijkheid heterogeniteit alleen bevordert bij lage tot middelmatige niet-uniekheid en geen uitgesproken effect meer heeft bij extreem hoge niet-uniekheid. De tweede interactie is analoog: negatief affect bevordert heterogeniteit alleen bij lage tot middelmatige niet-uniekheid en heeft geen uitgesproken effect meer heeft bij extreem hoge niet-uniekheid. Omdat beide interacties noch op technisch niveau noch op het niveau van de interpretatie een substantieel ander licht werpen op het onderzoek, achten we het in dit geval, voor de rechtlijnigheid van het betoog, legitiem om te rapporteren over het model zonder interacties, ook al is het model met interacties accurater. Ten vierde, er zijn 3 outliers en nog 19 andere 'influential observations'. Het weglaten van deze 22 observaties levert een iets beter model op dan dat in Tabel 1 (adjusted R squared van 0.7173 en ook een iets lagere standard error voor residuals), maar omdat dit noch technisch noch op het niveau van de interpretatie van de modellen wezenlijke verschillen zijn, rapporteren we hier over het model mét deze 22 observaties.

De resultaten van de analyse zijn in Tabel 1 terug te vinden. De tabel leert, ten eerste, dat het onderzochte model, met meer dan 61% verklaarde variatie, als goed mag worden bestempeld. (Voor wie niet vertrouwd is met een weergave als in Tabel 1: het betreft hier het cijfer dat wordt weergegeven bij Adjusted R-squared.) Ten tweede, alle factoren blijken een hoge mate van significantie te bezitten. (Dat is te zien aan de significantiecodes in de laatste kolom, weergegeven met asterisken.) De analyse detecteert m.a.w. een effect van de verklarende variabelen op de heterogeniteit, en de significantiecijfers laten zien dat de kans dat dat gedetecteerde effect aan het toeval te wijten is, te verwaarlozen is. Ten derde, op **ontbrek.plaatsen** na hebben alle verklarende variabelen een positieve correlatie met de responsvariabele. (De richting van de correlatie blijkt uit het teken van het cijfer dat in de kolom Estimate bij de verklarende variabelen vermeld wordt.) Die positieve correlatie betekent dat een stijging van de waarde van de verklarende variabelen ook een stijging van de heterogeniteit inhoudt.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.061618   0.350465   3.029  0.00281 **
ontbrek.plaatsen -0.005888   0.001984  -2.968  0.00341 **
ongebruikelijkheid 0.740298  0.142952   5.179 5.94e-07 ***
meerwoord.antw    2.782169   0.428651   6.491 8.04e-10 ***
niet.uniekheid    0.053341   0.007283   7.324 7.78e-12 ***
negatief.affect   0.540066   0.120095   4.497 1.23e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 1.216 on 180 degrees of freedom
Multiple R-Squared: 0.6232,     Adjusted R-squared: 0.6128
F-statistic: 59.55 on 5 and 180 DF,  p-value: < 2.2e-16
```

Tabel 1. Resultaten van de meervoudige lineaire regressie

De positieve correlaties tussen de verklarende variabelen en de reponsvariabele zijn in overeenstemming met de verwachtingen. De heterogeniteit neemt toe naarmate een concept minder gebruikelijk is, meer meerwoordige antwoorden telt, meer overlapt met andere concepten, en meer negatief geladen is: dat is precies wat we konden verwachten op grond van de a priori redenering waarmee we de opname van een bepaalde factor gemotiveerd hebben.

Echter, voor de variabele **ontbrek.plaatsen** is de vastgestelde negatieve correlatie tegengesteld aan de verwachtingen: de heterogeniteit neemt af naarmate er meer ontbrekende plaatsen zijn, terwijl we ervan uitgingen dat de heterogeniteit juist zou toenemen naarmate er minder attestatieplaatsen zouden zijn. Tegelijkertijd hebben we echter opgemerkt dat een beperkt aantal attestatieplaatsen niet alleen het gevolg zou kunnen zijn van de onbekendheid van het concept, maar ook van een minder systematische enquêtering. Van onbekendheid namen we aan dat er een verband met een verhoogde heterogeniteit zou kunnen zijn, maar van een onsystematische enquêtering kan dat niet zo makkelijk gezegd worden. Bovendien speelt bij een beperkter aantal attestatieplaatsen een mogelijk mathematisch effect: als de steekproef kleiner is, dan is de kans om meer lexicale types aan te treffen vanzelf kleiner. Het onverwachte resultaat voor 'aantal ontbrekende plaatsen' is dus niet zo geheel onverwacht wanneer we rekening houden met de dubbelzinnige status van de factor in kwestie.

4. Conclusie

We hebben met deze studie de vraag willen beantwoorden of heterogeniteit in het materiaal van een dialectwoordenboek als het Woordenboek van de Limburgse dialecten samenhangt met conceptkenmerken en niet slechts met geografische of taalstructurele gegevens. Meer specifiek hebben we de vraag gesteld of een grootschalige kwantitatieve analyse van het woordenboekmateriaal verschijnselen reveleert die aansluiten bij conceptuele saillantheid en conceptuele vaagheid, als verschijnselen die bij uitstek voor het voetlicht gebracht zijn door een prototypische betekenisopvatting.

Een regressieanalyse op de aflevering 'Het menselijk lichaam' van het Woordenboek van de Limburgse Dialecten laat onomstotelijk zien dat zulks het geval is: niet alleen de verschillende factoren waarmee we vaagheid en saillantheid geoperationaliseerd hebben blijken een effect te hebben op lexicale heterogeniteit, maar dat geldt ook voor negatief affect (een 'traditioneel' conceptkenmerk dat niet specifiek met een prototypische betekenisopvatting samenhangt).

Het hier voorgestelde onderzoek is slechts een eerste stap in de richting van een meer systematische analyse van het woordenboekmateriaal vanuit een up-to-date visie op lexicale betekenis. Mogelijkheden om het onderzoek uit te breiden liggen daarbij voor het grijpen. Zo kunnen we onderzoeken of de onderscheiden verklarende variabelen in dezelfde richting werken: wanneer we het effect van de variabelen nagaan op de verschillende factoren waarmee we de responsvariabele gedefinieerd hebben, zien we dan hetzelfde patroon bij diversiteit en bij verbrokkeling? We kunnen de resultaten uitsplitsen over geografische gebieden: werken de factoren op dezelfde manier in Nederland en in Vlaanderen? En als we het Woordenboek van de Vlaamse Dialecten en het Woordenboek van de Brabantse Dialecten bekijken, zien we dan een verschil met de hier gepresenteerde resultaten? We kunnen ook alternatieve operationaliseringen van de factoren uittesten, bijvoorbeeld door bereik en dispersie te meten met behulp van aantallen meetpunten i.p.v. oppervlaktes en afstanden, of door te vertrekken van een tokengebaseerde in plaats van een typegebaseerde meting van lexicale niet-uniekheid. En we kunnen het onderzoek natuurlijk ook uitbreiden naar andere conceptvelden (lees 'afleveringen'): werken vaagheid en saillantheid op dezelfde manier wanneer we het hebben over artefacten en cultuurbepaalde concepten als wanneer we het hebben over het menselijke lichaam?

Het onderzoeksdomein dat hiermee afgebakend wordt op de grens van cognitieve semantiek en dialectologie is voor beide disciplines nieuw en spannend. Voor de dialectlexicologie biedt het de kans om aan te sluiten bij de recente ontwikkelingen in de theoretische semantiek, en voor de cognitieve semantiek biedt het de kans om de interactie tussen cognitie, cultuur en taal te bestuderen vanuit het oogpunt van taalinterne taalvariatie.

Bibliografie

Berthele, Raphael

(2002). Learning a second dialect: A model of idiolectal dissonance. *Multilingua* 21: 327-344.

(2004). The typology of motion and posture verbs: A variationist account. In Bernd Kortmann (ed.), Dialectology Meets Typology. Dialect Grammar from a Cross-*Linguistic Perspective*. 93-126. Berlin / New York: Mouton de Gruyter.

(2006). *Ort und Weg. Die sprachliche Raumreferenz in Varietäten des Deutschen, Rätoromanischen und Französischen*. Berlin / New York: Walter de Gruyter.

Geeraerts, Dirk

(2005). Lectal variation and empirical data in Cognitive Linguistics. In Francisco J. Ruiz de Mendoza and Sandra Peña Cervel (eds.), *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*. 163-189. Berlin / New York: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers and Dirk Speelman.

(1999). *Convergentie en divergentie in de Nederlandse woordenschat. Een onderzoek naar kleding- en voetbaltermen*. Amsterdam: Meertens Instituut.

Kristiansen, Gitte & René Dirven (eds.)

(2008). *Cognitive Sociolinguistics*. Berlin / New York: Mouton de Gruyter.

Moerdijk, Alfons & Dirk Geeraerts

(1991). Het systematische dialectwoordenboek en de lexicologische gebruiker. Taal en Tongval 43: 58-104.

Nilsson, Anneli

(2001). Semantic shift among adjectives in the Southern-Swedish dialect of Scania. *Travaux de l'Institut de Linguistique de Lund* 39 (2): 227-240.

Sharifian, Farzad

(2005). Cultural conceptualisations in English words: A study of Aboriginal children in Perth. *Language and Education* 19 (1): 74-88.

Speelman, Dirk, Stefan Grondelaers & Dirk Geeraerts
(2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities* 37: 317-337.

Swanenberg, Jos
(2000). Lexicale variatie cognitief-semantisch benaderd. Over het benoemen van vogels in Zuid-Nederlandse dialecten. Dissertatie Katholieke Universiteit, Nijmegen.

Szelid, Veronika & Dirk Geeraerts
(2008). Usage-based dialectology. Emotion concepts in the Southern Csango dialects. *Annual Review of Cognitive Linguistics*. Geaccepteerd voor publicatie 6: 23-49.

Taal & Tongval themanr. 20 (2007), p. 47-62