DigiTaal

D-kwadraat: digitale databanken en digitaal gereedschap voor WBD en WLD

Folkert de vrieNd & Jos swaNeNBerG*

1 Inleiding

Het *Woordenboek van de Brabantse Dialecten* (WBD) is aan het einde van 2005 voltooid. De eerste aflevering is verschenen in 1967 en sindsdien volgden er nog 32 afleveringen. Aan het *Woordenboek van de Limburgse Dialecten* (WLD) wordt nog enkele jaren gewerkt aan de KU Leuven, maar ook dit grote lexicografische project zal binnen 2 à 3 jaar worden afgerond. In 1983 verscheen de eerste WLD-aflevering en als straks de laatste is verschenen, bestaat deze reeks uit 40 afleveringen.

Medio 2004 is een aanvang gemaakt met het project "digitale databanken en digitaal gereedschap voor WBD en WLD" (D-kwadraat). Met D-kwadraat zullen uiteindelijk alle gegevens waarop WBD en WLD zijn gebaseerd digitaal beschikbaar komen, zodat een ieder die in taalvariatie is geïnteresseerd de gegevens via het internet gemakkelijk kan raadplegen.

De websites www.ru.nl/dialect/wbd en www.ru.nl/dialect/wld kunnen gezien worden als pilotproject van D-kwadraat. In Kruijsen en Swanenberg (2003) werden deze websites, die in samenwerking met Polderland Language and Speech Technology werden ontwikkeld, al eerder in deze rubriek DigiTaal voor het voetlicht gebracht. Op deze websites kan de gebruiker ook nu al zoeken in gedeeltes van het WLD en het WBD.

Naast het digitaal toegankelijk maken van de dialectgegevens omvat D-kwadraat ook de ontwikkeling en uitwerking van digitaal onderzoeksgereedschap, zoals een breed inzetbare zoekmodule en eigentijdse cartografische middelen. Deze gereedschappen en de te ontwikkelen databasestructuur zullen tevens gebruikt kunnen worden voor de andere in het *permanent Overlegorgaan regionale Woordenboeken* (ReWo) samenwerkende woordenboekprojecten (De Vriend et al 2006).

In de volgende paragrafen gaan we achtereenvolgens in op de achtergronden van WBD en WLD, de binnen D-kwadraat doorgevoerde conversies en de manieren waarop de eindgebruiker met het gedigitaliseerde materiaal zal kunnen werken. Onder dat laatste valt ook de inzet van *Google Earth* als cartografische tool.

2 Achtergronden WBD en WLD

Tot het onderzoeksgebied van het WBD behoren de provincies Noord-Brabant, Antwerpen en Vlaams-Brabant en het hoofdstedelijk gewest Brussel. De beide provincies Limburg vormen met het noordoosten van de provincie Luik het onderzoeksgebied van het WLD. Samen met het Woordenboek van de Vlaamse Dialecten, dat Oost-, West-, Frans- en Zeeuws-Vlaanderen bestrijkt, overdekken het WBD en het WLD door eenzelfde methode van beschrijvende dialectlexicografie praktisch het hele Zuid-Nederlandse taalgebied beneden de grote rivieren.

Het WBD en het WLD zijn systematisch geordende woordenboeken. Op het hoogste niveau in

de macrostructuur worden de volgende drie delen onderscheiden: Agrarische woordenschat (I), Niet-agrarische vakterminologieën (II) en Algemene woordenschat (III). De woordenboeken zijn dus niet alfabetisch, maar op taxonomische wijze naar gebruikssfeer geordend.[1] Een woordenboekartikel vermeldt voor een bepaald concept alle gevonden dialectvormen met hun vindplaats.

Verder bevat elk lemma een encyclopedische betekenisomschrijving van het concept en de gebruikte bronnen.[2] Sommige lemmata zijn bovendien voorzien van een kaart om de verspreiding van de vormvariatie te illustreren. Iedere aflevering van de woordenboeken wordt ontsloten door alfabetische registers. Voor een uitgebreide uitleg van de macrostructuur, de wijze waarop de reeks is opgezet en de woordenboektekst is gestructureerd, leze men Van Keymeulen (1992), of beknopter Kruijsen (1996) en Kruijsen & Van Keymeulen (1997).

3 D-kwadraat

In deze paragraaf zullen we wat dieper ingaan op de drie belangrijkste onderdelen van het project D-kwadraat; de conversie van het woordenboekmateriaal, de manieren om toegang tot het gedigitaliseerde materiaal te krijgen en de inzet van een cartografische tool.

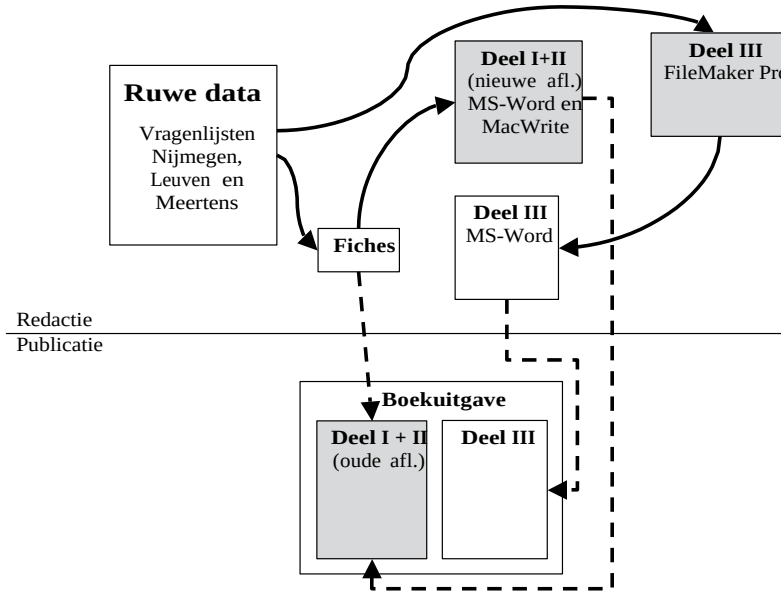
3.1 Conversie

Figuur 2 geeft grofweg weer hoe de redactie door de jaren heen te werk is gegaan. Als uitgangspunt dienden de gegevens verzameld met behulp van vragenlijsten. Deze gegevens werden in eerste instantie verwerkt door per vragenlijst de begrippen (concepten) te nemen en op aparte fiches de bijbehorende dialectwoorden en Kloekecodes te zetten.[3] De fiches werden in kaartenbakken ondergebracht. Voor de oudere afleveringen van deel I en II van de woordenboeken gold dat de kaartenbakken met fiches in zijn geheel naar de uitgever werden gebracht.

WLD III, 4.2

kikker: Siebengewald en Venlo.
kwakvors (ook *kwakvos, kwakvost, kwekvors, kwekvos):* freq. Tongerlds., Dommellds., Demerkemp., Zuidgeld.Lb. en Kleverlds., verspr. Noord.Oostlb., Maaskemp., Trichterlds., Bilzerlds., Beringerlds., Lonerlds., Truierlds., Brab.Lb. en Lommels; ook in Gronsveld, Neeritter, Grathem, Maaseik en Rotem.
kwakvros (ook *kwakfrots, kwekvros):* freq. in het zuiden van Oost.-Zuidlb., Geullds., Tongerlds. en Lonerlds., verspr. Ripuar., Zuid.Oostlb., Trichterlds. en Noord.Oostlb. en aansluitend Horns en Centr.Maaslds.
kwek: Sevenum.
kikvors (ook *kikvos):* freq. Truierlds., Getelds. en Kleverlds., verspr. Westlb.; ook in Reuver, Tessenderlo, Arcen en Venlo.
kikvros: Hoepertingen, Riemst en Zichen-Zussen-Bolder.
kikmauw: Venray Wb.
vors (ook *vos):* freq. Brab.Lb. en

Figuur 1: fragment van het lemma voor "kikker" in deel III van het WLD.



Figuur 2: het redactieproces door de jaren heen met de verschillende (half)fabricaten.

Toen de tekstverwerker zijn intrede deed op de woordenboekredacties werden de afleveringen in MacWrite en later MS-Word aangeleverd bij de uitgever. Voor de afleveringen van deel III zijn de gegevens van de vragenlijsten eerst ondergebracht in een FileMaker-database. Op basis van deze database kwamen de woordenboekartikelen tot stand.

Van belang voor D-kwadraat zijn de grijs gekleurde blokken. Dat zijn de gegevens die als uitgangspunt dienen voor de conversie. Vanwege de lange geschiedenis van de woordenboeken gaat het hierbij dus om zowel boekpagina's, tekstverwerkerbestanden als databases.

De conversie van de oudste afleveringen van deel I en II is het meest arbeidsintensief. Deze afleveringen zijn enkel in gedrukte vorm beschikbaar en zijn daarom eerst ingescand. Vervolgens is de tekst op de scans toegankelijk gemaakt met Optical Character Recognition (OCR).

Binnen de woordenboeken wordt het Genoveva-font gebruikt. Dit is een fonetisch font dat in de jaren '80 specifiek werd ontworpen voor WBD en WLD. Bij de OCR is het de software "aanleren" van dit Genoveva-font een punt van zorg geweest. Het font bleek te veel af te wijken van het soort van fonts waar de software voor was ontwikkeld. In de praktijk betekende dit dat we het trainen van Genoveva-karakters hebben moeten beperken tot enkele veel gebruikte karakters zoals de sjwa. De OCR-software wordt daarom nu in hoofdzaak door een datatypist gebruikt voor het efficiënt corrigeren van het resultaat van de automatische OCR.

De conversie van de tekstverwerkerbestanden en database-bestanden is eenvoudiger te realiseren omdat het uitgangspunt hierbij al een digitaal formaat is. De belangrijkste zorg bij de conversie van deze formaten is dat het om bestandsformaten gaat die soms met de huidige software en besturingssystemen niet meer leesbaar zijn. De tekstverwerkerbestanden zijn bij-

voorbeeld halfweg de jaren tachtig aangemaakt. In zulke gevallen wordt gebruik gemaakt van oudere hard- en software om een soort "brug-platform" te creëren. Een platform oud genoeg om de oudere formaten te kunnen interpreteren en tegelijkertijd nieuw genoeg om ze naar een formaat op te slaan dat door de huidige generatie software kan worden geïnterpreteerd.

De ca. 3 miljoen dialectvormen zullen uiteindelijk in een MySQL database worden opgeslagen. De meer hiërarchisch georganiseerde gegevens worden naar een XML-formaat geconverteerd. Hieronder valt in ieder geval de taxonomie waarop de organisatie van de woordenboeken in semantische velden is gebaseerd (De Vriend et al 2006).

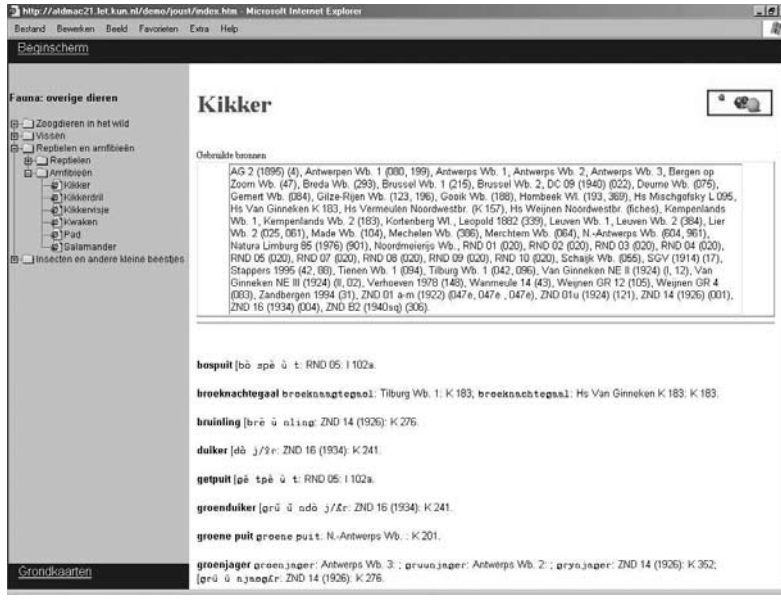
3.2 Gebruikerstoegang

De dialectgegevens zullen met behulp van een web interface toegankelijk worden gemaakt. De twee belangrijkste toegangswegen tot het materiaal zijn de zoekmachine en de taxonomie op gebruikssfeer.

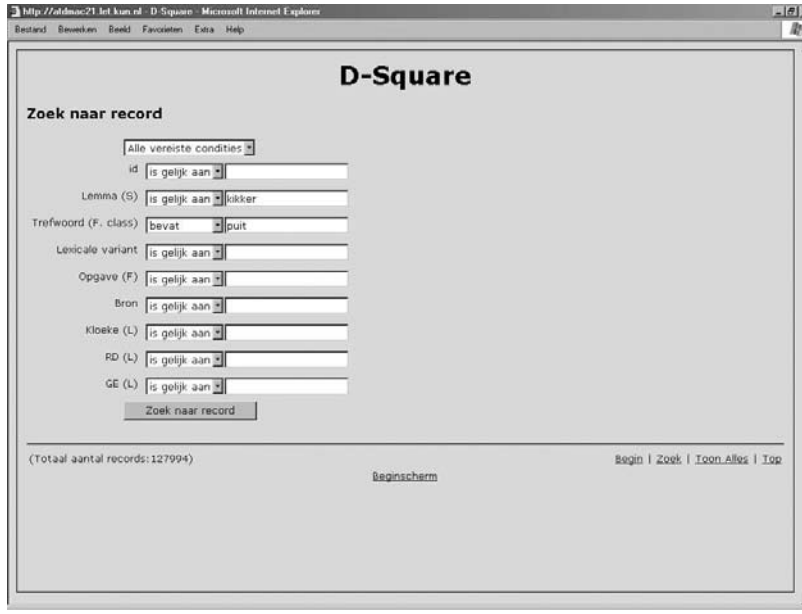
Wil men op traditionele wijze het materiaal benaderen dan klikt men door in de taxonomie tot men bij het gezochte lemma is aanbeland. In feite is dit een directe vertaling naar het digitale medium van het gebruik van een inhoudsopgave in boekvorm.

Naast de toegang met behulp van de taxonomie zal de gebruiker het materiaal uitgebreid kunnen doorzoeken met een zoekmachine.

In het voorbeeld in Figuur 4 wordt gezocht binnen het lemma voor "kikker" op alle trefwoorden die *puit* bevatten. Dit levert naast *puit* zelf o.a. ook *bospuit* en *puitje* op. Interessanter wordt het wanneer een zoekvraag over vormkenmerken wordt gesteld over het gehele semantische domein. Zo zou men bijvoorbeeld kunnen kijken in welke semantische domeinen – anders dan dat van de vogeltjes – de vorm *mus* wordt gebruikt. Men komt dan o.a. *mussekersen* (voor de lijsterbes in de flora) en *mussenpis* (voor motregen in de weersgesteldheid) tegen.



Figuur 3: toegang tot het lemma voor "kikker" in het WBD op basis van de taxonomie.



Figuur 4: voorbeeld van het gebruik van de zoekmachine.

Extra functionaliteit wordt geboden door de mogelijkheid om het resultaat van een zoekvraag te exporteren naar een CSV-bestand (Comma Separated Value). Op deze manier worden gebruikers in staat gesteld zelf selecties uit het materiaal te extraheren, bijvoorbeeld om daarmee een lokaal dialectwoordenboek samen te stellen; "geef alle records behorend bij plaats X". Een CSV-bestand kan vervolgens eenvoudig worden bewerkt in bijvoorbeeld Microsoft Excel.

3.3 Cartografie

Een van de redenen waarom de woordenboeken destijds systematisch zijn opgezet is de mogelijkheid om bij de lemma's kaarten op te nemen met daarop de vormvariatie (De Tollenaere & Weijnen 1963). In de woordenboeken vind je dan ook veel van zulke kaarten terug. Binnen D-kwadraat willen we ook extra aandacht schenken aan de mogelijkheden die cartografie biedt voor het inzichtelijk maken van de variatie die er bestaat in de dialecten, zowel binnen de vorm als binnen de betekenissen. We hebben daartoe voor de geo-browser *Google Earth* (verder; "GE") gekozen. GE blijkt uitermate geschikt te zijn als cartografische tool voor de dialectgegevens. De standaardversie van GE kan gratis worden geïnstalleerd en fungeert vervolgens als zelfstandig programma, d.w.z. los van de binnen een webbrowser functionerende interface. Figuur 5 laat zien hoe de dialectgegevens uitgezet kunnen worden in GE. Het voorbeeld toont de drie meest frequente dialectvormen voor "kikker"; *puit*, *vors* en *kikvors* in het Brabants dialectgebied.

Het formaat waarin de gegevens aan GE moeten worden aangeboden heet Keyhole Markup Language (KML), een speciaal op GE toegesneden XML-formaat. Wanneer de gebruiker de gegevens van een bepaald lemma wil zien in GE dan klikt hij op de "GE-knop" bij het lemma.[4] Vervolgens wordt het bijbehorende KML-bestand automatisch geopend in GE.

Met de intuïtieve 3D interface van GE kan de leek of onderzoeker letterlijk vanuit verschillen-

Figuur 5: de verspreiding van de drie meest frequente trefwoorden voor "kikker" in GE.

de invalshoeken tegen de gegevens aankijken. Ook wordt met het gebruik van GE de waarde van cartografie als onderzoeksinstrument verhoogd omdat verschillende linguïstische kaartbeelden relatief eenvoudig met elkaar of met andere gebiedsinformatie gecombineerd kunnen worden. Men kan bij dat laatste denken aan de combinatie met historische kaarten, kaarten met isoglossen, landschappelijke kaarten, kaarten met informatie over bevolkingsdichtheid, over de verspreiding van religie, etc. Het in GE combineren van deze gegevens levert interessante inzichten op over de oorsprong van de patronen in de dialectgegevens.

Het is de bedoeling dat de gebruiker uiteindelijke voor elk lemma in de database een bijbehorende kaart kan bekijken in GE.

4 Tot besluit

Het project D-kwadraat loopt nog tot zomer 2007 aan de Radboud Universiteit Nijmegen en is gedeeltelijk gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).[5] Geïnteresseerden kunnen tussentijds op de projectwebsite terecht; www.ru.nl/dialect/d2. Naast achtergronden over het project vindt men daar demo's van de in deze bijdrage beschreven functionaliteit waaronder het getoonde kikkerlemma in *Google Earth* en verschillende met dit lemma in *Google Earth* te combineren grondkaarten.

Op het moment van schrijven is nog niet duidelijk waar de binnen D-kwadraat gedigitaliseerde gegevens en ontwikkelde gereedschappen uiteindelijk zullen worden ondergebracht. De Centrale voor Taal- en Spraaktechnologie (www.tst.inl.nl) ligt het meest voor de hand als toekomstig beheerder. Gegevens en gereedschappen zullen in ieder geval vrijelijk beschikbaar worden gesteld, zowel voor wetenschappelijke als niet-wetenschappelijke doeleinden.

Bibliografie

De Tollenaere, F. & A. Weijnen (1963). Het dialectwoordenboek. *Woordenboek en dialect. Lezingen gehouden voor de Dialectcommissie der Koninklijke Nederlandse Akademie van Weten-*

schappen op 4 november 1961 door Dr. F. de Tolle- naere en prof. Dr. A. Weijnen. Amsterdam: N.V. Noord-Hollandsche Uitgevers Maatschappij.

De Vriend, F., L. Boves, H. van den Heuvel, R. van Hout, J. Kruijsen & J. Swanen- berg (2006). A Unified Structure for Dutch Dialect Dictionary Data. *proceedings of The fifth international conference on Language resources and Evaluation (LREC 2006), Genoa, Italy.*

Google Earth (2005-2006). earth.google.com.

Kruijsen, J. (1996). De Nijmeegse dialectlexico- grafische projecten. *Trefwoord* 11, 93-108. **Kruijsen, J. & J. Van Keymeulen (1997).** The Southern Dutch Dialect Dictionaries. *Lexicos* 7 *(AFRILEKS-reeks* 7), 207-228.

Kruijsen, J. & J. Swanenberg (2003). Lim- burgse en Brabantse dialectdatabanken op internet. *Nederlandse Taalkunde* 8-2, 158-162.

SW (1994-2004). *Stellingwarfs Woordeboek.* Oos- terwolde: Stichting Stellingwarver Schrie- versronte.

Van Keymeulen, J. (1992). *De algemene woor- denschat in de grote dialectwoordenboeken (WBD, WLD, WVD). Een methodologische reflectie.* Uni- versiteit Gent.

WALD (1984--). *Woordenboek van de Achterhoekse en Liemerse Dialecten.* Doetinchem: Staring- Instituut.

WBD (1967-2005). *Woordenboek van de Brabantse Dialecten.* Assen: Van Gorcum / Groningen: Gopher.

WLD (1983--). *Woordenboek van de Limburgse Dialecten.* Assen: Van Gorcum / Groningen: Gopher.

WVD (1979--). *Woordenboek van de vlaamse Dia- lecten.* Universiteit Gent.

WZD (1959-1964). *Woordenboek der Zeeuwsche dialecten.* Amsterdam: Elsevier.

Noot

* Afdeling Taalwetenschap, Radboud Universiteit, Nijmegen, Postbus 9103, 6500 HD Nijmegen.
E-mail: f.devriend@let.ru.nl; j.swanenberg@let. ru.nl

1 De in WBD en WLD gebruikte taxonomie is gebaseerd op *Begriffssystem als Grundlage für die Lexikographie: versuch eines Ordnungsschemas* van R. Hallig & W. Von Wartburg (1952).

2 Dit zijn voornamelijk schriftelijke enquêtes en lokale woordenboeken.

3 De wijze van coderen en ook de toekenning van de Kloekecodes aan de plaatsen in het WBD- en WLD- gebied staat uitgebreid beschreven in de Inleiding bij deel III van het WBD en de Inleidingen van het WLD. Zie WBD, deel III, *Inleiding* en *Klankgeografie* (2000): 25-68, WLD, deel I, *Inleiding* en Afl. 1, *Bemesten en ploegen* (1983), 7-22 en deel III, *Inleiding* en Afl. 4.1, *Fauna: vogels* (2001: XVIII-XLI)). De codes werden geïntroduceerd door Grootaers en Kloeke in het "Systematisch en alfabetisch register van plaatsnamen" uit 1962.

4 De "GE-knop" is ook te zien in de rechter bovenhoek van figuur 3.

5 Het project wordt uitgevoerd door Jos Swanenberg en Folkert de Vriend. De onderzoeksgroep bestaat uit Roeland van Hout (projectleider), Henk van den Heuvel (CLST), Theo van de Heuvel (Polderland Language & Speech Technology B.V.) en Joep Kruijsen (redacteur van het WLD).