HET IDEALE WOORDENBOEK

Marc van Oostendorp

In de dertiende druk van de grote Van Dale (uit 1999) worden 85 soorten noten genoemd, van 'aardaker' tot 'zeepnoot'. Wie alle definities van die nootsoorten op een rijtje zet, merkt dat ze nogal willekeurig zijn. Zo wordt van de pistache wél vermeld dat hij gegeten kan worden, terwijl die informatie bij de aardnoot ontbreekt. Bij de walnoot wordt opgemerkt dat hij soms zwart is, terwijl over het uiterlijk van de hazelnoot helemaal niets gezegd wordt.

Die inconsistenties in het woordenboek waren tot voor kort onvermijdelijk. Zo'n boek is een enorm weefsel, waarvan de man achter het getouw onmogelijk alle honderdduizenden draden afzonderlijk in het oog kan houden. Pas sinds kort, sinds de komst van de computer, kunnen we automatisch heel snel het hele woordenboek doorzoeken op alle woorden waarin *noot* voorkomt, zowel trefwoorden als woorden die in de definities staan. Pas zo wordt het mogelijk om woorden met een verwante betekenis ook daadwerkelijk bij elkaar te zetten en te vergelijken.

Uitgevers van woordenboeken als Van Dale hebben daarom belangstelling voor de mogelijkheden die nieuwe technieken bieden. Eind jaren negentig bood Van Dale de Universiteit Utrecht een door de uitgever te betalen promotieplaats aan. Enkele jaren mocht de toen net afgestudeerde neerlandicus Oele Koornwinder onderzoek doen naar de mogelijkheden om de computer in te zetten bij het verbeteren van het woordenboek. Op het resultaat van dat onderzoek promoveerde hij onlangs aan de Universiteit Utrecht.

Foto: Frank Fahrner

Taalwetenschapper Oele Koornwinder met het schilderij dat prijkt op het omslag van zijn proefschrift: de werkelijkheid opgedeeld in stukjes.

In stukjes

Koornwinder onderzocht de woordbouw, de manier waarop woorden in kleine samenstellende delen kunnen worden opgehakt: *reanimatie* bestaat uit de stukjes *re* (dat we ook terugvinden in bijvoorbeeld *reïntegreren*), *anim* (dat bijvoorbeeld terugkomt in *geanimeerd*) en *atie* (dat we terugvinden in *demonstratie*). Juist met dit soort ontledingen kun je gemakkelijk woorden met een verwante betekenis opsporen. Koornwinder: "Van Dale gebruikte al methoden om automatisch woorden in stukjes te delen, maar die waren minder geschikt voor bijvoorbeeld woorden van buitenlandse oorsprong. Ze hadden iemand nodig die een database zou maken met alle woorddelen van het Nederlands."

Een andere toepassing van zo'n database is de automatische correctie van fouten. Koornwinder: "We weten dat vrijwel alle woorden met de uitgang *-ing* een meervoud hebben op *-en: verenigingen*, *bewegingen*, enz. Als je dat weet, kun je in één keer het woordenboek doorzoeken om te kijken of er ergens per vergissing het meervoud *-s* is gegeven."

De eerste jaren werkte Koornwinder bij de uitgever op kantoor. Daarna kreeg hij een kamer op de universiteit. Zo leerde hij twee werelden kennen: "Een uitgeverij is een bedrijf; alles is gericht op efficiëntie en het halen van deadlines. De database die ik maakte, werd al meteen ingezet voor een programma dat woorden uitspreekt op de cd-rom van Van Dale, en ook voor een nieuwe editie van Van Dale hedendaags Nederlands – het kleine broertje van de grote Van Dale. In een dergelijk bedrijf werk je veel samen, bijvoorbeeld met de mensen die zo'n nieuw woordenboek maken. Op de universiteit werkt iedereen meer voor zichzelf en aan fundamenteler onderzoek. Er zijn geen klanten die op je resultaten zitten te wachten, en de deadlines zijn daardoor ook minder strikt. Uiteindelijk heeft de wetenschappelijke kritiek me wel geprikkeld om mijn gedachten preciezer op te schrijven."

Tien pagina's

Wanneer gaat Van Dale Koornwinders proefschrift gebruiken? "Omdat ze in 2005 weinig tijd hadden vanwege de introductie van de nieuwe spelling en het verschijnen van een nieuwe druk, hebben ze er nog niet precies naar kunnen kijken. Maar ze gaan er zeker iets mee doen." Met behulp van Koornwinders methode kan Van Dale een stapje dichter komen bij het 'ideale woordenboek', waarin de definities duidelijk en precies zijn. "Maar helemaal sluitend zal een woordenboekdefinitie nooit zijn, want dan heb je voor elk woord zeker tien pagina's nodig. Daar heeft niemand wat aan."

Inmiddels werkt Koornwinder voor het Amsterdamse bedrijfje GridLine, dat zich ook richt op dit soort taaltechnologie. Hij bouwt onder andere aan een programma dat uit een grote verzameling gespecialiseerde teksten automatisch alle vaktermen filtert, bijvoorbeeld om daar een woordenboek van te maken. En ooit keert hij misschien toch ook terug naar de wetenschap: "Ik ben bij verschillende subsidieaanvragen in Utrecht betrokken."