Schrijfsters in een databank

SUZAN VAN DIJK

Op 21 september 2007 vond in het Huygens Instituut de afsluiting plaats van het NWO-digitaliseringsproject *The International Reception of Women's Writing*, waaraan van 2004 tot 2007 met vier assistenten was gewerkt. Opgeleverd werden 17 duizend records met informatie over de ontvangst in Nederland van Nederlandse en buitenlandse schrijfsters uit de achttiende en negentiende eeuw. In het verlengde van de grote bloemlezing uit 1997 over Nederlandstalige schrijfsters uit de periode 1550-1850, *Met en zonder lauwerkrans*, kan nu het onderzoek naar het aandeel van vrouwen in de literaire wereld grootschalig worden voortgezet.

De records zijn onderdeel van de database *WomenWriters*. Deze bevat informatie over de ontvangst door tijdgenoten en vroege literatuurhistorici (tot ca. 1900) van het werk van vrouwelijke schrijvers, ontleend aan verschillende grootschalige bronnen zoals tijdschriften, bibliotheekcatalogi en privé-correspondentie. De records verwijzen en linken soms door naar deze bronnen; een enkele keer is de tekst van bijvoorbeeld een artikel of brief integraal overgenomen. De informatie is afkomstig uit zowel online toegankelijke als papieren bronnen.

Idealiter zullen ooit databanken zoals *WomenWriters* fungeren als distributieapparaten, waar steeds kan worden doorgeklikt naar achterliggende bronnen. Het is nog niet zo ver, maar dankzij de DBNL (Digitale Bibliotheek der Nederlandse Letteren) was het mogelijk om voor enkele tijdschriften (bijv. *De Gids*, *De Nieuwe Gids*) een directe toegang te maken naar artikelen over schrijfsters die nu veelal onbekend zijn, en waarover belangrijke negentiende-eeuwse critici als Conrad Busken Huet uitgebreid schreven – in positieve of negatieve zin. Met subsidie vanuit het academische repositoryproject DARE is samen met het Huygens Instituut het tijdschrift *Vaderlandsche Letteroefeningen* online gezet. Vanuit *WomenWriters* zijn hyperlinks gemaakt naar de daar gevonden artikelen.

De database biedt dus onderzoekers, studenten en belangstellenden informatie uit bronnen die die informatie, indien gezocht op individueel niveau, niet gemakkelijk prijsgeven. Bovendien is het informatie waarnaar onderzoekers niet gemakkelijk op zoek gaan: *buitenlandse* onderzoekers beseffen niet altijd dat kleine landen in receptie-historisch opzicht interessant zijn, en *Nederlanders* zijn zich er niet per se van bewust dat meer dan zevenhonderd Nederlandse vrouwen actief zijn geweest als vertaalster/journaliste/schrijfster, én dat hun werk enigermate de aandacht heeft getrokken.

Er is zo een voorraad aan gegevens verzameld die eerder nauwelijks betrokken werden bij het onderzoek naar het vrouwelijk auteurschap. Nadere bestudering en interpretatie ervan zullen nieuw licht werpen op de rol van vrouwelijke auteurs in het Nederlandse literaire veld tot circa 1900, bezien in de Europese context. In beoogde vervolgprojecten zullen ook gegevens uit Portugal, Rusland, België en andere landen worden ingevoerd. Vervolgens zal aan de hand van *WomenWriters* in een Europees samenwerkingsproject een 'internationale vrouwenliteratuurgeschiedenis' worden voorbereid.

Daarbij laten wij ons onder andere inspireren door de ideeën van de literatuur- en cultuurhistorici Linda Hutcheon en Mario Valdés, die in *Rethinking Literary History* (2002) uitgaan van de *'literary text as a historical event of production and subsequently of reception'* (p. 67). De benadering van het literaire veld als continu verwikkeld in dialoog tussen auteur en lezers vormde ook het uitgangspunt voor de database *WomenWriters*, met dien verstande dat de aandacht zich richt op de plaats die vrouwen zich wisten te verwerven, bij een vrouwelijk zowel als een mannelijk contemporain publiek. Zoals Margaret Ezell enige tijd geleden al betoogde in *Writing Women's Literary History* (1993), was deze plaats in de eeuwen vóór de negentiende dikwijls veel groter dan de in die tijd ontstane canon deed vermoeden.

www.womenwriters.nl.
www.databasewomenwriters.nl

MARIE-PIERRE LEGRAND

Leden van de projectgroep 'New approaches to European Women's Writing before 1900' in overleg over de mogelijkheden van een nieuwe literatuurgeschiedenis, juli 2007

Volkstelling 1947

JAN JONKER

Herkenning van handgeschreven tabellen

In een pilotproject in het kader van de digitalisering van de Volkstellingen zijn de mogelijkheden onderzocht van optische tekenherkenning bij het verwerken van de handgeschreven tabellen. Jan Jonker was bij dat project betrokken en doet verslag van de bevindingen.

Sinds 1997 hebben DANS en zijn voorgangers samengewerkt met het Centraal Bureau voor de Statistiek (CBS) om de uitkomsten van de Nederlandse volkstellingen digitaal toegankelijk te maken. Na de gedrukte publicaties van de volkstellingen vanaf 1795 tot en met de laatst gehouden telling van 1971, is het digitaliseren van de bij het CBS aanwezige transparanten en lichtdrukken aangepakt. Zo zijn onder meer in 2005 de circa dertig duizend beschikbare transparanten van de Volks- en beroepstelling 1947 gescand en als images toegankelijk gemaakt op www.volkstellingen.nl.

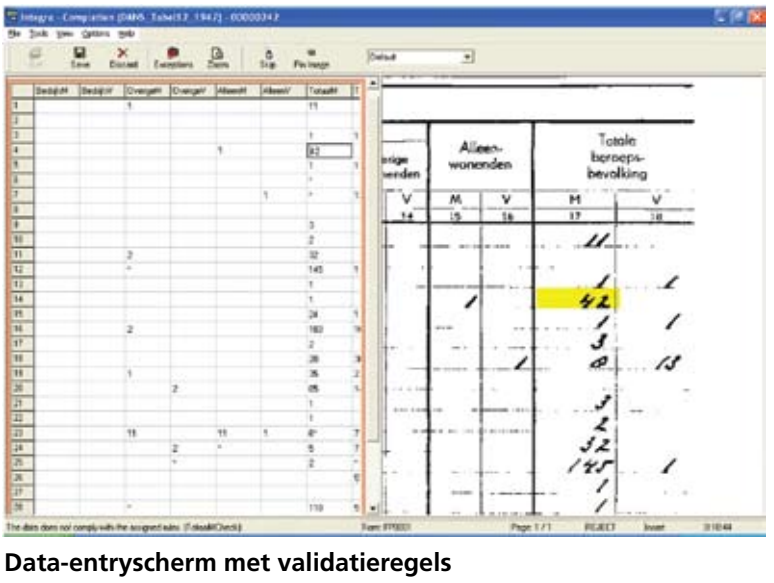
Doel van het digitaliseringproces is steeds om de cijfermatige gegevens van de tellingen in verwerkbare vorm beschikbaar te stellen voor nieuw statistisch onderzoek. Dat vereist na het scannen een stap van 'herkenning' van de gegevens, uitlopend op representatie daarvan in tabellen, bijvoorbeeld in Excel. In de praktijk is dat meestal gedaan door een vorm van data-entry van de gegevens, al dan niet rechtstreeks in Exceltabellen.

Voor Tabel 12 van de Volkstelling van 1947 zijn de actuele mogelijkheden onderzocht van optische tekenherkenning (*optical character recognition* of OCR). Dat is gedaan in een pilotonderzoek met gebruik van *eFlow*, een OCR-applicatie van Top Image Systems Benelux (TIS), uitgevoerd door Xerox Global Services. Uitdagend detail daarbij was dat de statistische gegevens in handschrift zijn ingevuld op de transparanten van deze Volkstelling.

In *TIS eFlow* worden automatisch de images van Tabel 12 geselecteerd uit de totale verzameling images van telling 1947. Images die niet voldoende worden herkend, worden alsnog handmatig geclassificeerd. Per saldo zijn er meer dan duizend images van Tabel 12, namelijk voor alle gemeenten uit 1947 en voor de wijken van grotere gemeenten. Voor de images van Tabel 12 worden vervolgens de gegevens uit de tabelcellen bepaald door weging van de resultaten van een aantal OCR-*engines*. In een daarop volgende *completion* fase worden de OCR resultaten visueel gecontroleerd, aangevuld en/of gecorrigeerd om *false positives* uit te sluiten. Hierbij wordt gebruik gemaakt van validatieregels voor regel- en kolomtotalen om de kwaliteit van de herkende en/of ingevoerde data te valideren.

De mogelijkheden van *TIS eFlow* voor dit doel zijn op zichzelf veelbelovend. In de eerste proeven van Xerox was zelfs sprake van 84% automatisch correct herkende tabelcellen. Wel bleken de resultaten bij een controle achteraf bij het data-instituut DANS nog relatief veel *false positives* te bevatten. Een van de oorzaken daarvan zijn kaderlijntjes om de tabelcellen in de transparanten. In samenwerking tussen DANS, TIS en Xerox is de *eFlow*-applicatie verfijnd, waarbij alle voor Tabel 12 mogelijke controletellingen over regels en kolommen van de tabel zijn gerealiseerd.

Daardoor zijn de resulterende Exceltabellen een zeer nauwkeurige weergave van de cijfers in de originele transparanten. De herkenningsprocedure vergt op deze manier wel meer interactief werk in de *completion* fase voor validatie van de uitkomsten. Toch werd nog circa 70% van de tabelcellen geheel automatisch correct herkend.



Data-entryscherm met validatieregels

Het pilotonderzoek met OCR van images van de Volks- en beroepstelling 1947 heeft laten zien dat een exacte beschrijving vooraf van de voor herkenning aangeboden tabellen onmisbaar is om de vereiste nauwkeurigheid te bereiken. Dat geldt in het bijzonder voor de per tabel mogelijke controletellingen. Op basis daarvan zou de *TIS eFlow* applicatie voor elke tabel in detail moeten worden gespecificeerd. Het laatste kost in de huidige situatie nog relatief veel gespecialiseerde arbeid. De OCR aanpak is daardoor nog niet kosteneffectief. Om die reden wordt voor de herkenning van de images uit dit volkstellingjaar alsnog gebruik gemaakt van 'traditionele' data-entry.