

Automatische tekstvergelijking maakt wetenschappelijke editie mogelijk

Hermans blijvend leesbaar

PETER KEGEL EN BERT VAN ELSACKER

Ruim een jaar geleden verscheen het eerste deel van de Volledige Werken van Willem Frederik Hermans, met daarin de romans *Conserve* (1947) en *De tranen der acacia's* (1949). Dat was de officiële start van een grootschalig editieproject, waarin vierentwintig verzamelbanden verschijnen van zo'n 800 pagina's, tweemaal per jaar tot in 2016.

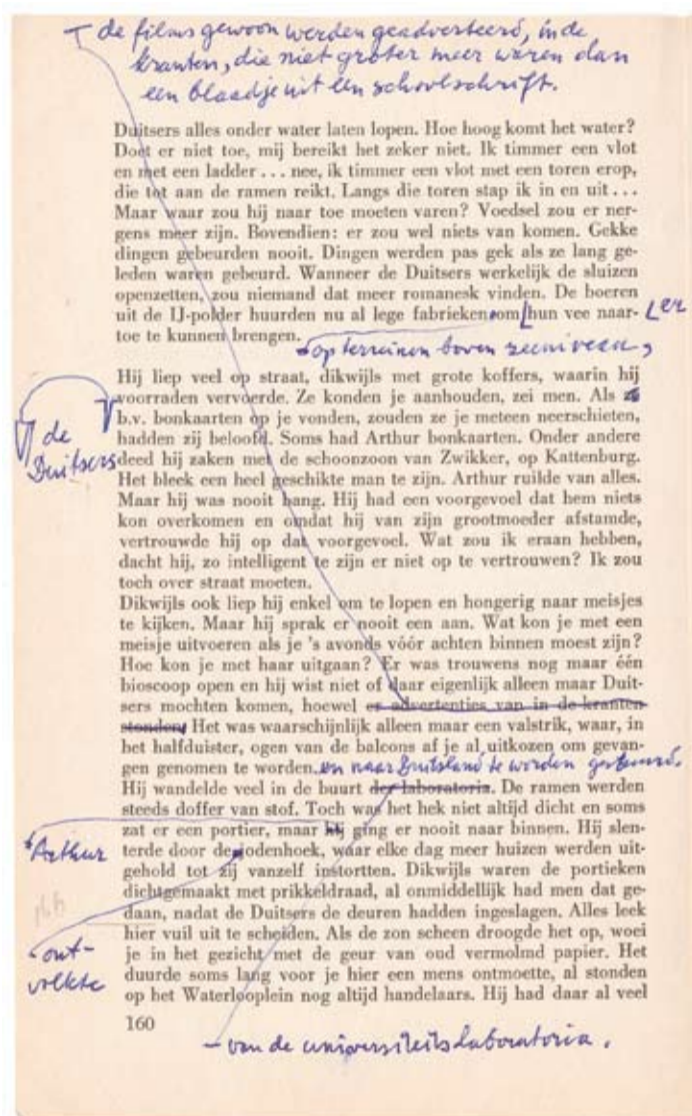
Inmiddels zijn drie delen gepubliceerd: afgelopen voorjaar deel 7, met de verhalenbundels *Moedwil en misverstand* (1948), *Paranoia* (1953) en *Een landingspoging op Newfoundland en andere verhalen* (1957); deze maand kwam daar deel 12 bij, dat *Boze Brieven van Bijkaart* (1977) en *Houten leeuwen en leeuwen van goud* (1979) bevat.

De wetenschappelijke editie is een samenwerkingsproject van het Huygens Instituut – verantwoordelijk voor de feitelijke bezorging van de tekst – het Willem Frederik Hermans Instituut en Hermans' vaste uitgever De Bezige Bij. De editie bevat de definitieve teksten van Hermans' werk: uitgangspunt is steeds de laatste tekstversie die Hermans zelf nog onder ogen kreeg en voor heruitgave goedgekeurd.

Omvangrijke wijzigingen

Hermans was een auteur die ook lang na eerste publicatie nog aan zijn teksten bleef schaven. Zelf wilde hij het liefst dat zijn lezers alleen de laatste druk tot hun beschikking hadden: 'Ik zou willen dat alle oude drukken van boeken die in verbeterde vorm herdrukt zijn, als bij toverslag tot stof uiteenvielen, ook al gaat het maar om een komma', schreef hij in het voorwoord bij de bibliografie met zijn verspreide publicaties *Schrijven is verbluffen*. Dat het niet alleen om komma's ging maar vaak om omvangrijke wijzigingen, blijkt uit de tekstgeschiedenis van *De tranen der acacia's*, een roman die Hermans continu zou blijven herzien. De vierentwintigste druk van deze roman uit 1993, uitgangspunt voor de tekst van de editie, toonde uiteindelijk duizenden verschillen ten opzichte van de eerste (gedeeltelijke) publicatie, in 1946 en 1947 in het tijdschrift *Criterium*.

Bij het maken van een wetenschappelijke editie is het een eerste voorwaarde om de werkwijze van Hermans te kennen en inzicht te krijgen in de aard van de herzieningen die hij in zijn teksten aanbracht. Onderzoek naar de tekstgeschiedenis van *De tranen der acacia's* heeft uitgewezen dat het Hermans vooral ging om een blijvende leesbaarheid van zijn teksten. Om dat te bereiken bracht hij veelvuldig, maar niet altijd consequent, veranderingen aan in spelling, stijl en woordgebruik. Bovendien voorzag hij, waar hij dacht



Hermans' correcties bij *De tranen der acacia's*

dat dat nodig was, zijn teksten van toelichtende informatie. Ten slotte bracht hij, om de compositie van zijn teksten te versterken, herhaaldelijk meer verteltechnische wijzigingen aan.

Een op drie drukken

Het tekstkritische onderzoek voor de editie steunt op twee pijlers. Eén daarvan is het archiefonderzoek. De bezorgers van de editie hebben volledige toegang tot het archief-Hermans, dat zich in bruikleen in het Letterkundig Museum bevindt. Het archief bevat onder veel meer enkele manuscripten, typoscripten, talrijke correctie-exemplaren en drukproeven. Dankzij dit materiaal, aangevuld met de uitgebreide auteurscorrespondentie, is het mogelijk ten minste een gedeeltelijke reconstructie te maken van de totstandkoming van Hermans' publicaties, en de respectievelijke inbreng daarbij van auteur, uitgever, zetter, redacteur en/of corrector. Maar voor een verantwoorde wetenschappelijke editie is meer nodig. Een volledig beeld van alle veranderingen in Hermans' teksten, vereist een gedetailleerde

bestudering van alle tekstversies met onderlinge verschillen. Bij Hermans is grofweg een op de drie drukken van zijn teksten herzien, zodat meer dan vijftig duizend gedrukte pagina's vergeleken moeten worden. Dergelijk onderzoek is ondenkbaar zonder het tweede fundament onder de editie: automatische tekstvergelijking.

Al bij de voorbereiding van de *Volledige Werken* was duidelijk dat de bezorging van de editie zonder geautomatiseerde collatie onhaalbaar zou zijn. Betrouwbare digitale bronbestanden waren daarvoor een vereiste. Om dat te realiseren heeft het Huygens Instituut samengewerkt met twee partners. Specialisten van de Koninklijke Bibliotheek verzorgden de microverfilming van het bronnenmateriaal, dat vervolgens door het Nederlands Instituut voor Wetenschappelijke Informatiediensten werd gedigitaliseerd en met OCR-software omgezet naar computerleesbare tekst in RTF (Rich Text Format). In een fase daarna zijn bij het Huygens Instituut de digitale bestanden grondig gecontroleerd aan de hand van een referentieverisie, onder meer met gebruikmaking van het programma Araxis. Ten

slotte zijn, om de teksten geschikt te maken voor automatische collatie, formele tekstkenmerken expliciet gecodeerd met behulp van speciaal voor dit doel ontwikkelde scripts. Na dit complexe voorbereidingstraject zijn de teksten klaar voor automatische tekstvergelijking.

Verrijken met meta-informatie

Vooral binnen de informatica wordt al sinds de jaren zeventig veel praktisch en theoretisch onderzoek naar automatische tekstvergelijking gedaan. Informatici maken hiervan bijvoorbeeld gebruik om verschillen in de opeenvolgende versies van een programmacode te analyseren en op die manier bugs op te sporen. Dit onderzoek heeft relatief snel geleid tot een aantal standaardalgoritmes en -toepassingen, waarvan het Unix/Linux-hulpmiddel 'diff' het bekendste is. Enigszins los van deze ontwikkelingen zijn er ook in de wereld van de editiewetenschap initiatieven geweest om computers in te zetten voor tekstvergelijking en voor de productie van edities. De meest bekende voorbeelden zijn

COLLATE van Peter Robinson en TUSTEP van Wilhelm Ott. Een nieuw en interessant product op dit gebied is Juxta, een initiatief van het Amerikaanse NINES (Networked Interface for Nineteenth-century Electronic Scholarship)

Bij de voorbereidingen van de *Volledige Werken* is tot nu toe vooral gewerkt met COLLATE en diff-toepassingen. Om optimaal gebruik van de variantenoverzichten mogelijk te maken, worden deze via scripts omgezet naar XML-TEI gecodeerde documenten. Het basisidee van XML (Extensible Markup Language) is om tekst niet te beschouwen als een reeks tekens, maar als een structuur, die expliciet gecodeerd wordt zodat ze computerleesbaar is. Op die manier is het mogelijk een tekst te verrijken met meta-informatie, die als een gegevensbank kan worden doorzocht en bewerkt. Voor het gebruik van XML-codes bestaan verschillende standaarden. Het Text Encoding Initiative (TEI) heeft uitgebreide richtlijnen opgesteld voor toepassingen in de humaniora, met specifieke modellen voor proza, toneel en poëzie, maar ook voor bijvoorbeeld transcripties van spraak, voor woordenboeken, en voor kritische edities.

Het eindresultaat van het digitaliseringstraject bestaat uit full-text digitale overzichten van Hermans' teksten. Deze XML-bestanden staan aan de basis van het tekstkritische onderzoek door de editor. De

verslaglegging van dat onderzoek wordt ook weer gedocumenteerd in de XML-bestanden. De editor legt via TEI-codes zijn correcties en de verantwoording daarvan vast in de te editeren tekst. Een specifieke markering krijgen ook varianten die illustratief zijn voor bijvoorbeeld het herzienings- of productieproces van een bepaalde titel, of die om tekstinhoudelijke redenen van belang zijn, net als tekstplaatsen met een bijzondere typografie en eigenaardigheden die specifiek zijn voor Hermans' werkwijze. Het oorspronkelijke variantenoverzicht groeit zo uit tot een systematische onderzoeksdocumentatie.

Juist de digitale beschikbaarheid van het onderzoeksmateriaal kan leiden tot nieuwe inzichten in de ontstaansgeschiedenis van Hermans' teksten. Zo bleek bijvoorbeeld uit zoekacties in de gecodeerde digitale gegevens dat Hermans, anders dan hij later in diverse interviews meldde, vóór publicatie van de eerste druk van *De tranen der acacia's* al uitzonderlijke veel herzieningen aanbracht ten opzichte van de eerder in *Criterium* verschenen versie.

Relatief korte periode

Doordat de XML-TEI-bestanden alleen maar bestaan uit tekst en codes, dus vrij zijn van opmaak, blijven ze ook op lange termijn computerleesbaar. De XML-data, die tevens gebruikt worden als kopij voor de editie, kunnen daardoor worden hergebruikt voor latere gedrukte uitgaven of voor digitale edities. De codering in XML-TEI schept bovendien de mogelijkheden voor nieuwe, dynamische vormen van tekstpresentatie en -analyse, die bijvoorbeeld kunnen ingaan op compositie, intertekstuele, cultuurhistorische of meer interpretatieve aspecten van de onderzochte tekst. Er is nog een laatste, belangrijk voordeel. De digitale beschikbaarheid van het onderzoeksmateriaal maakt het voor de editor mogelijk om in een relatief korte tijd duizenden pagina's tekst nauwkeurig te onderzoeken. Daardoor komen in relatief snelle opeenvolging de teksten van Hermans in hun definitieve vorm en voorzien van een uitgebreid commentaar beschikbaar voor alle liefhebbers, docenten en studenten. Een speciale website: www.wfhermansvolledigewerken.nl bevat de wetenschappelijke verantwoording. Deze site wordt ook het platform voor digitale publicaties op basis van de XML-TEI-onderzoeksgegevens.

Volledige Werken: www.wfhermansvolledigewerken.nl
 Collate: www.itsee.bham.ac.uk/software/collate/
 TUSTEP: www.zdv.uni-tuebingen.de/tustep/tustep_eng.html
 Juxta: www.patacriticism.org/juxta/
 Diff: www.gnu.org/software/diffutils/
 Araxis: www.araxis.com
 TEI: www.tei-c.org