

## Tekstretrieval in krantencollecties

Bij projecten waarin historische tekstuele bronnen worden omgezet in machineleesbare tekst, wordt vaak gebruik gemaakt van optische tekenherkenning, oftewel *Optical Character Recognition* (OCR). Meestal beschikken deze projecten niet over de middelen voor de controle van de automatisch herkende tekst. Een oplossing voor dit probleem lijkt het gebruik van tekstretrieval-software, die varianten in spelling (fuzzy-zoeken) ondersteunt. Astrid Verheusen en Rubrecht Zaat beschrijven een recent project waarin de nauwkeurigheid van deze retrievalsoftware werd getoetst.

**B**ij de Koninklijke Bibliotheek is onlangs het project *Oorlog & Revolutie* afgerond. Er zijn 76 jaargangen van vier landelijke dagbladen uit de eerste helft van de twintigste eeuw vanaf microfilm gedigitaliseerd en omgezet in machineleesbare tekst. In totaal zijn er bijna 350.000 pagina's gedigitaliseerd. Een deel van het digitaliseringsproject werd, in het kader van *Het Geheugen van Nederland*, in samenwerking met het Haags Gemeentearchief uitgevoerd. Het project is gefinancierd door het ministerie van OGenW en werd uitgevoerd in opdracht van *Metamorfoze*, het landelijk programma voor conservering van bibliotheekmateriaal. *Metamorfoze* stelde ook de microfilms ter beschikking. De films van de betreffende jaargangen waren reeds in 1998-2001 verfilmd, als onderdeel van het project *Microverfilming landelijke dagbladen 1840-1950*. Doel was het vinden van een goede methode om historische kranten vanaf microfilm te digitaliseren, zo goed mogelijk te ontsluiten en beschikbaar te stellen. Na een onderzoek van de mogelijkheden, is besloten dit uit te proberen door toepassing van OCR (zonder controle achteraf) en fuzzy-zoekmogelijkheden.

Met het oog op een optimale ontsluiting van de kranten waren de resultaten van de

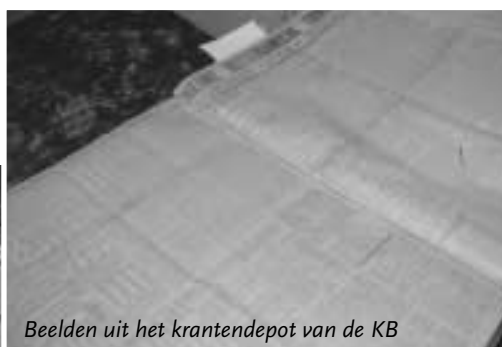
OCR van groot belang. Producenten van OCR-software beloven vaak nauwkeurigheidspercentages van nabij 100 procent. Voor historisch krantenmateriaal zijn deze percentages nauwelijks bekend.<sup>1</sup> Om de toegepaste methode voor de ontsluiting te evalueren en te achterhalen of met fuzzy-zoeken fouten in niet-gecontroleerde teksten kunnen worden gecompenseerd, wilde de Koninklijke Bibliotheek een test uitvoeren. Krijgt de gebruiker de juiste antwoorden op een zoekvraag en zijn de resultaten volledig? Is er sprake van zoekresultaten die niet aan de zoekvraag voldoen, en zo ja, in welke mate komt dit voor?

### AANPAK CONVERSIE EN ONTSLUITING

Voor het digitaliseren werd gebruik gemaakt van een Sunrise-microfilm-scanner. De hoge resolutie daarvan is equivalent met circa 300 dpi voor de papieren originelen. Na digitalisering zijn de images op verschillende manieren bewerkt. Voor de tekenherkenning is gebruik gemaakt van de *Software Development Kit* van het Russische *ABBYY FineReader*, versie 6. De hiermee verkregen full-text bestanden zijn niet gecontroleerd. Voor de fuzzy-zoekmogelijkheden is tekstretrieval-software van *Zylab* gebruikt. De keuze voor *Zylab* is mede ingegeven door de ervaring die dit bedrijf heeft opgedaan bij het ontsluiten van het historisch archief – uit de periode 1877-1940 – van het tijdschrift *De Groene Amsterdammer*.

### STEEKPROEF

Om te komen tot degelijke resultaten, moest een betrouwbare en representatieve steekproef worden genomen van de gedigitaliseerde pagina's. Voor het bepalen van de steekproef is gebruik gemaakt van berekeningen die zijn gemaakt door J. van Oss.<sup>2</sup> Hij stelt dat de validiteit van een steekproef niet afhangt van het percentage getrokken eenheden ten



Beelden uit het krantendepot van de KB



opzichte van het geheel, maar van hun absolute aantal. Berekend is dat bij een minimum steekproef van 1537 pagina's een foutmarge bestaat van 5 procent en een betrouwbaarheid van 95 procent. Dit aantal moet wel met regelmatige intervallen worden geselecteerd om een representatieve set te verkrijgen. Het aantal te testen pagina's en de selectie door middel van regelmatige intervallen was goed werkbaar voor het onderzoek.

In het project zijn 76 jaargangen van *Het Volk*, *Het Centrum*, *Het Vaderland* en de *Nieuwe Rotterdamsche Courant (NRC)* gedigitaliseerd. Het selecteren van 1537 pagina's hieruit kwam neer op circa 20 pagina's per jaargang. Gezien de verschillende aantallen pagina's per krant en per jaargang zijn de intervallen tussen de geselecteerde pagina's enigszins variabel. Bij het vaststellen van de te controleren pagina's rond de intervallen, is steeds een andere editie genomen en binnen een editie is gekozen voor steeds een ander volgnummer. Op deze manier zijn verschillende soorten pagina's (zoals voorpagina's, pagina's met veel advertenties, pagina's met veel cijfers zoals beursberichten) in de steekproef opgenomen. Door de variabele intervallen en door afrondingsverschillen ontstond een selectie van 1562 pagina's.

## METHODE

Er is nog geen standaardmethode met bijbehorende kwaliteitsnormen voor het controleren van elektronische historische teksten die met OCR-software zijn vervaardigd. Veel onderzoek wordt gedaan naar het automatisch corrigeren van OCR-resultaten (*post-processing*). Om het aantal pagina's dat gecorrigeerd moet worden te reduceren, wordt ook onderzoek gedaan naar het automatisch selecteren van pagina's met een laag herkenningpercentage.<sup>3</sup> In een onderzoek van de universiteit van Michigan naar de accuratesse van ongecontroleerde teksten in het grote digitaliseringsprogramma *The Making of America* is die methode toegepast.<sup>4</sup> De universiteit heeft dit gedaan met behulp van een indicatie over de mate van accuratesse die door de software zelf wordt aangegeven, de *Prime Score*. Het ging hierbij om een controle van het aantal foute tekens in de teksten.<sup>5</sup> Een andere methode werd toegepast in een onderzoek van de universiteit van Harvard. Daar werd de kwaliteit niet gemeten door het tellen van foute tekens, maar door het aantal succesvolle zoekacties in de tekst. Hiervoor waren twee redenen. Ten eerste wilde de universiteit van Harvard afbeeldingen tonen en niet de full-text bestanden, en ten tweede is het handmatig vergelijken van tekens uit het machineleesbare bestand een bijzonder tijdrovende bezigheid.<sup>6</sup> Dezelfde twee redenen speelden mee in het bepalen van een methode voor het testen van ongecorrigeerde teksten in het Nederlandse krantenproject. Alleen de afbeeldingen worden op internet getoond en er waren geen middelen voorhanden om op 1562 pagina's tekens te vergelijken. De belangrijkste reden was echter, dat de teksten niet achteraf gecorrigeerd zouden worden. Het onderzoek was vooral gericht op de vraag in hoeverre tekstretrieval-software met goede fuzzy-zoekmogelijkheden de fouten in ongecontroleerde teksten kon compenseren.

De test is uitgevoerd door een team van medewerkers van de Koninklijke Bibliotheek en het Haags Gemeentearchief. Ieder teamlid kreeg een aantal pagina's uit de



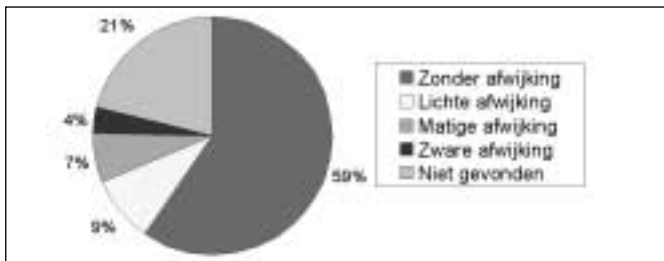
steekproef toebedeeld. De bijbehorende afbeelding werd gezocht via een door Zylab ontwikkelde website. Op de pagina werd een woord of woordcombinatie geselecteerd. Op basis van een analyse van de tekst waren verschillende categorieën tekst gedefinieerd. Er is gezocht in gewoon zetsel, cursieve tekst, tekst die geheel uit hoofdletters bestaat en woorden in advertenties. De woorden of woordcombinaties waarnaar is gezocht, varieerden in omvang van 2 tot 27 letters.

De geselecteerde term werd gezocht via de zoekpagina, waarbij de zoekactie beperkt werd tot de betreffende pagina. Een aantal veel voorkomende woorden, zoals lidwoorden, was uitgesloten. Indien de term werd gevonden op de pagina zonder gebruik te maken van fuzzy-zoekmogelijkheden werd dit genoteerd. Het ging daarbij wel om het voorkomen van de term op de juiste plek. Indien de zoekactie niet succesvol was, werd dezelfde zoekactie herhaald maar nu met fuzzy-graad 'licht' waarbij één teken kan afwijken. Indien de term nu wel werd gevonden, werd dit genoteerd. Indien de term nog niet werd gevonden, werd de zoekactie herhaald met fuzzy-graad 'matig' en 'zwaar', waarbij respectievelijk twee of drie tekens in een woord kunnen afwijken of ontbreken. De software van Zylab houdt daarbij rekening met de lengte van een woord. Aanpassing volgt wanneer met een zware fuzzy-graad wordt gezocht op korte woorden. Om een overdaad aan zoekresultaten te voorkomen, wordt een zoekopdracht met zware fuzzy-graad dan automatisch teruggebracht tot een lagere fuzzy-graad.<sup>7</sup>

Indien de term helemaal niet werd gevonden, werd dat ook genoteerd. Met fuzzy-zoeken kunnen verkeerde zoekresultaten worden gevonden (ruis). Het aantal verkeerde zoekresultaten werd separaat geïnventariseerd per fuzzy-graad.

## ANALYSE

In 59 procent van de zoekacties werd de gevraagde zoekterm gevonden zonder gebruik te maken van de fuzzy-



Figuur 1. Resultaten van zoekacties met gebruikmaking van de fuzzy-optie. Totaal (1562 pagina's)

optie. Figuur 1 laat zien dat 20 procent van de zoekacties slechts resultaat opleverde wanneer gebruik werd gemaakt van een van de drie fuzzy-opties (lichte-, matige- of zware afwijking). De zoekterm werd helemaal niet gevonden in 21 procent van de gevallen.

Tussen de verschillende kranten is er een significant verschil te bespeuren. De NRC uit de periode 1910-1929 leverde de slechtste resultaten (zie figuur 2). Minder dan de helft van alle zoekopdrachten (48 procent) had direct succes. Fuzzy-zoeken kon dit teleurstellende resultaat slechts ten dele tenietdoen (20 procent). In 32 procent van alle zoekopdrachten werd de zoekterm niet gevonden.

Het dagblad *Het Vaderland*, uit de periode 1920-1945, leverde de beste scores (zie figuur 3). In 73 procent van de zoekopdrachten werd de zoekterm direct gevonden zonder gebruik van de fuzzy-optie. 10 procent van de zoekopdrachten leverde geen resultaat op. In 17 procent werd de zoekopdracht succesvol uitgevoerd met behulp van de fuzzy-zoekopdracht.

De scores van de andere dagbladen (*Het Centrum* en *Het Volk*) lagen tussen deze twee uitersten in.

Wat kan gezegd worden over het nut van de fuzzy-optie? Het aantal zoekopdrachten waarbij de fuzzy-optie met succes werd aangewend, lag tussen 17 (*Het Vaderland*) en 22 procent (*Het Centrum*). Het zoekresultaat werd met gemiddeld 20 procent verbeterd. Het gebruik van een zwaardere fuzzy-graad leverde meer ruis in de zoekresultaten. De *precision* (het aantal relevante zoekresultaten in verhouding tot het aantal irrelevante zoekresultaten) nam duidelijk af.<sup>8</sup> Bij sommige zoekopdrachten leverde de zware fuzzy-graad meer dan tien ruiswoorden op. Daarbij dient dan nog te worden opgemerkt dat de hoeveelheid ruis beperkt was, omdat er slechts gezocht werd naar een bepaalde term op een specifieke pagina en niet in het hele bestand. De ruis maakte het gebruik van deze optie tot een minder goed instrument om snel specifieke zoekopdrachten uit te voeren. Bijkomend nadeel was dat de performance van het zoeken afnam wanneer gezocht werd met de fuzzy-mogelijkheid.

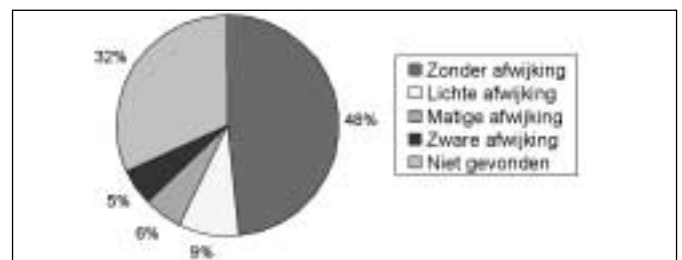
## VERKLARINGEN

Voor de verschillen tussen de vier dagbladen is een aantal oorzaken aan te wijzen. Ten eerste is de kwaliteit van het bronmateriaal van belang. De meest recente krant is *Het Vaderland* uit de periode 1920-1945. De zoekopdrachten in *Het Vaderland* waren het meest succesvol en de *recall* (de mate waarin een systeem relevante zoekresultaten genereert in verhouding tot het totaal aantal relevante resulta-

ten) was het hoogst. De lay-out van 'modernere' kranten speelt hierbij een belangrijke rol.

De NRC uit de periode 1910-1929 leverde de slechtste scores. Het bronmateriaal is vele malen slechter dan bijvoorbeeld *Het Vaderland*, met veel scheuren, aanwezigheid van verschillende bijlagen en uitlopen en doordrukken van de inkt (*bleeding ink*). De lay-out van de NRC zorgde voor verschillende problemen: vele – nauwelijks gescheiden – kolommen tekst en weinig variatie op de pagina met betrekking tot drukletter of opmaak. Illustraties of foto's komen niet of nauwelijks voor. Op de originele pagina's van de NRC is tot slot vaak sprake van enorme contrastverschillen, waardoor het soms moeilijk is met het blote oog de tekst te herkennen.

Naast de kwaliteit van het bronmateriaal, speelt de wijze van verfilming een rol bij de tekenherkenning. In 1999 werd het krantenmateriaal op 35 mm microfilm gezet met conservering als belangrijkste doel. De films zijn van hoge technische kwaliteit, de belichting en het contrast zijn goed. Bij de verfilming is echter vooral aandacht gegeven aan conservering; zaken die voor digitalisering en ontsluiting van belang zijn, speelden toen geen rol. Zo is om het



Figuur 2. Resultaten van zoekacties met gebruikmaking van de fuzzy-optie. De NRC (407 pagina's)

origineel te sparen het materiaal ingebonden verfilmd, hetgeen veel weglappende tekst in de band heeft opgeleverd. Ook is er sprake van schaduwwerking als gevolg van bolting van het papier in de band (*gutter shadow*).

Een ander euvel is de glasplaat die gebruikt is bij het afdekken/gladstrijken van het materiaal. De bovenkant van de glasplaat is regelmatig zichtbaar op de film en dus op de digitale afbeelding. Ook deze kan voor een hinderlijke schaduwwerking zorgen. Soms staan pagina's enigszins scheef op de film. Voor gebruikers van de films is dat niet ernstig, voor het goed herkennen van tekst is het belangrijk dat de tekst recht staat.

Tot slot speelt het formaat van de pagina's een rol: een pagina van *Het Volk* is 34 x 49 centimeter, bij de NRC is dit formaat 43 x 57 centimeter. Bij het verfilmen is het van belang om maximale beeldvulling van de film te bereiken. Hoe groter het origineel, hoe groter de verkleiningsfactor.

Deze onvolkomenheden op de film hebben grote gevolgen voor digitalisering en tekenherkenning. Deskundigen beweren dat digitalisering van film dezelfde kwaliteit kan opleveren als digitalisering van het origineel: *it is possible to get equally good full text search from texts, which are scanned from microfilm as from scanning from the originals.*<sup>9</sup> Hierbij

dient voorafgaand aan het verfilmen echter bekend te zijn dat de films gedigitaliseerd worden, zodat hiermee rekening kan worden gehouden.

Ook de kwaliteit van digitalisering is van belang voor het succes van OCR. Resolutie, contrast, belichting en kwaliteit van de scanapparatuur kunnen het resultaat van de tekenherkenning sterk beïnvloeden. In het project *Oorlog & Revolutie* was de kwaliteit van de images hoog en daarom niet van wezenlijke invloed op het eindresultaat.

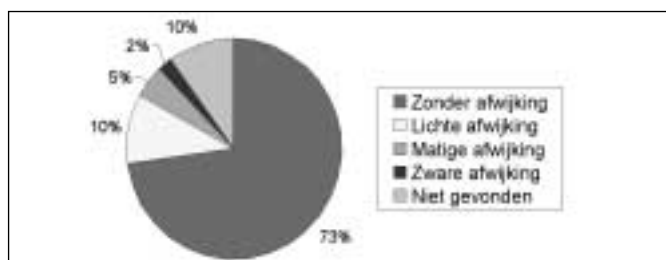
### CONCLUSIES EN AANBEVELINGEN

Hoewel het uitgevoerde onderzoek beperkt van omvang was en er geen absolute gegevens zijn verkregen over herkenningspercentages, heeft het inzicht gegeven in de mogelijkheden om grote historische tekstuele bestanden te ontsluiten. Ook al is de OCR niet van goede kwaliteit, de full-text biedt hoe dan ook betere zoekmogelijkheden dan de papieren bronnen doen. Daarbij leidt het zoeken met behulp van spellingsvarianten tot nog meer zoekresultaten. Geconcludeerd kan worden dat bij omvangrijke projecten met beperkte middelen een tekstretrieval-systeem vooralsnog de best mogelijke ontsluiting biedt.

Wel is duidelijk geworden dat fuzzy-zoeken in ongecontroleerde teksten beslist geen 100 procent garantie biedt voor succesvolle zoekacties. Ruim een vijfde deel van de zoektermen werd helemaal niet gevonden; fuzzy-zoeken leverde veelvuldig zoekresultaten op die van geen belang zijn voor de gebruiker en de zoekopdrachten duurden daarbij ook langer. Voor projecten waarin hoge eisen worden gesteld aan de accuratesse, is deze aanpak dus niet geschikt.

De oplossing kan gedeeltelijk worden gezocht in andere retrievalssystemen met betere zoekmethoden. Maar hoe geavanceerd de zoekmethoden ook zijn, de echte oplossing lijkt toch vooral te liggen in het verbeteren van de kwaliteit van de teksten waarin gezocht wordt, en dus in de kwaliteit van de OCR. Hiervoor zijn een aantal mogelijkheden. Het is aan te bevelen digitaliseringsprojecten met OCR vooraf te laten gaan door een uitgebreide materiaalanalyse van de bronnen, waarbij gelet moet worden op de wijze van inbinden, schaduwwerking, de kwaliteit van de tekst en van het papier. De gegevens die bij deze analyse worden verzameld, kunnen worden gebruikt bij de instructie voor verfilming en digitalisering. Voor kleinschaligere projecten met ruimere middelen is *rekeying* (het overtypen van de tekst) of een combinatie van OCR en correctie een optie.

Bij het gebruik van intermediairs bij het digitaliseringsproces – zoals microfilms – dient te worden nagegaan of deze voldoen aan de eisen die digitalisering en ontsluiting stel-



Figuur 3. Resultaten van zoekacties met gebruikmaking van de fuzzy-optie. *Het Vaderland* (527 pagina's)

len. Een test kan uitwijzen of de kwaliteit volstaat.

Bij het project *Oorlog & Revolutie* speelde de kwaliteit van de microfilms een grote rol in de kwaliteit van de OCR en dus in het aantal succesvolle zoekacties. Het is interessant te onderzoeken hoe succesvol dezelfde zoekacties zouden zijn in scans vanaf het origineel. Een kleine test hiermee gaf aan dat digitalisering vanaf de originele pagina's betere scans en dus betere zoekresultaten opleveren. Voor een uitgebreidere test zou het materiaal van *de Groene Amsterdammer* geschikt zijn, aangezien dit tijdschrift vanaf het origineel is gedigitaliseerd. Ook zou onderzocht kunnen worden hoe de zoekresultaten zouden zijn bij microfilms die met het oog op digitalisering zijn gemaakt. De vraag doet zich voor of microverfilming ten behoeve van conservering te verenigen is met microverfilming voor digitalisering, tenzij aan beide kanten aan de kwaliteitseisen wordt toegegeven.

De ontwikkelingen op het gebied van tekenherkenning staan niet stil. Nieuwe versies van pakketten zorgen voor verbeterde nauwkeurigheid. Voor historisch bronmateriaal, waaronder kranten, is er echter nog veel werk voor de boeg. Vooralsnog lijkt meer te verwachten van het onderzoek naar automatische correctie van elektronische teksten en naar methoden om pagina's met slechte OCR snel te kunnen selecteren. Op dit punt is internationale samenwerking en overleg gewenst met partners die reeds dergelijke technieken gebruiken.

### Noten

1. Uitzondering hierop is het rapport in het kader van het Tidenproject. *Report on Microfilming and OCR-reading tests* (Helsinki University Library, Finland, The Royal Library, Sweden, z.j.). Zie hiervoor: <http://tiden.kb.se/microfilm.pdf>.
2. J. van Oss m.m.v. R. Rutgers, 'Onder behandeling. De reductie van het cliëntenarchief van de afdeling geestelijke gezondheidszorg van de GG&GD Amsterdam', in: Paul M.M. Klep (red), *Steekproeven uit massale archiefbestanden ter wille van het historisch belang* (Den Haag 1997), 56-57 en 62.
3. Zie bijvoorbeeld Prateek Sarkar, Henry S. Baird, John Henderson, 'Triage of OCR Results Using Confidence Scores', in: Proc., *9th IS&T/SPIE Document Recognition & Retrieval Conf.* San Jose, CA, (january 2002).
4. Zie [www.hti.umich.edu/m/moagrpf/](http://www.hti.umich.edu/m/moagrpf/). In The Making of America waren halverwege 2004 zo'n 8.500 boeken en 50.000 tijdschriftenartikelen gedigitaliseerd en door middel van OCR omgezet in full-text.
5. Douglas A. Bicknese, *Measuring the Accuracy of the OCR in the Making of America* (1998).
6. LDI Project Team1, *Measuring Search Retrieval Accuracy of Uncorrected OCR: Findings from the Harvard-Radcliffe Online Historical Reference Shelf Digitization Project* (Harvard University Library, August 2001).
7. De fuzzy-graad wordt berekend door de waarde van de instelling en 0,5 maal de woordlengte. Als bijvoorbeeld de fuzzy-graad 4 is en de zoekopdracht uit 6 tekens bestaat, dan is de werkelijke fuzzy-graad  $0,5 \times 6 = 3$  i.p.v. 4.
8. Voor meer informatie over *precision* en *recall* zie: Richard Jizba, *Measuring Search Effectiveness*, Creighton University Health Sciences Library and Learning Resources Center (Omaha, Nebraska, 2000).
9. IFLA rapport, *Microfilming for Digitization and Optical Character Recognition, supplement to guidelines* (december 2002).

Astrid Verheusen en Rubrecht Zaat zijn beiden Projectleider Afdeling Research & Development bij de Koninklijke Bibliotheek.