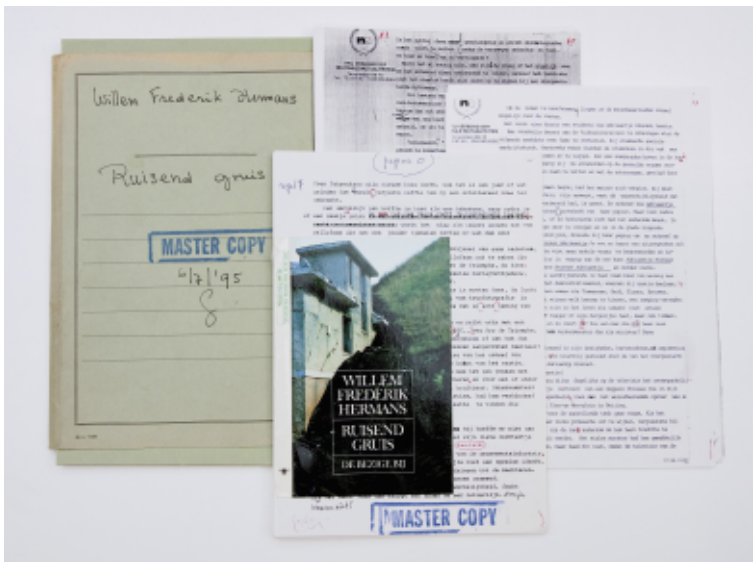

Bronnenmateriaal 'Mandarijnen op zwavelzuur', W.F. Hermans archief *credits wfhermansvolledigewerken.nl*

Nieuwe software helpt bij collationeren

CollateX vergelijkt automatisch teksten

Tientallen versies van een tekst met elkaar vergelijken is een arbeidsintensieve klus. Bij Huygens ING experimenteerden onderzoeker en ontwikkelaar met software die hierbij helpt.

Marieke Willems

Door verschillende versies van een tekst te vergelijken, kunnen onderzoekers te weten komen wat de oorspronkelijke brontekst van bijvoorbeeld de bijbel is. Ook voor onderzoek naar de totstandkoming van een tekst is dit 'collationeren' een geschikte methode, die letterkundigen bijvoorbeeld gebruikten in Volledige Werken, een Huygens ING-project over het werk van W.F. Hermans. Daarnaast kan collationeren de kwaliteit van getranscribeerde handgeschreven brieven verbeteren wanneer verschillende transcripties elkaar aanvullen.

Basistekst kiezen

Wanneer een onderzoeker handmatig collationeert, kiest hij eerst een basistekst, een versie van de tekst waartegen hij alle andere versies vergelijkt. In het geval van honderden versies is dat een moeilijke keuze die een bias met zich meebrengt. Een internationaal team van ontwikkelaars ontwierp daarom CollateX, software die meerdere versies onderling vergelijkt zonder keuze van een basistekst. Een computer kan tientallen tekstversies met elkaar vergelijken in enkele seconden of minuten en doet dat consistent en zonder bias. De gebruiker kan de keuze voor een basistekst zo uitstellen tot een later moment in het onderzoek of zelfs helemaal achterwege laten.

Training onderzoeker

Het is belangrijk dat onderzoekers begrijpen hoe de software werkt, stelt Ronald Haentjens Dekker, lead engineer in Huygens ING en ontwikkelaar van CollateX. "Indien de computer en de gebruikte tools een zogenaamde 'blackbox' blijven, is het onduidelijk welke aannamen gedaan zijn tijdens het onderzoek en hoe betrouwbaar het resultaat is." Haentjens Dekker trainde daarom onderzoekster Elli Bleeker van de Universiteit van Antwerpen. Bleeker onderzocht hoe de computer kan worden ingezet voor tekstgenetisch onderzoek, waarbij het schrijfproces belangrijker is dan het eindproduct. Dit resulteert in zeer complexe transcripties met meerdere 'schrijflagen'. Haentjens Dekker en Bleeker experimenteerden binnen CollateX met het collationeren van twee XML-files, inclusief alle auteurscorrecties, schrijflagen en andere tags. De resultaten waren veelbelovend en worden verder uitgewerkt. Bleeker concludeert: "Al met al heeft mijn training en onderzoek in CollateX zeer waardevolle inzichten opgeleverd. Niet alleen op het gebied van geautomatiseerde tekstcollatie, maar evengoed over het belang van het kennen én begrijpen van de software die je gebruikt, en over de ideale vorm van samenwerking tussen onderzoeker en ontwikkelaar."

huygens.knaw.nl/collate-x
github.com/interedition/collatex

GELEZEN

Peer Review Quality and Transparency of the Peer-Review Proces in Open Access and Subscription Journals, J.M. Wicherts

Milo van de Pol

Veel wetenschappelijke Open Acces Journals nemen het niet zo nauw met het peer-reviewproces. OA-uitgevers zouden economisch gewin een belangrijker afweging vinden dan wetenschappelijke betrouwbaarheid. Althans, dat is de mening van heel wat wetenschappers. Een berucht voorbeeld dat vaak wordt aangehaald is het hoax artikel van wetenschapsjournalist John Bohannon dat door bijna de helft van alle OA-journals zonder meer werd geaccepteerd.
Jelte Wicherts, psycholoog en UD aan de Tilburg University, zocht uit of bovenstaande mening wel klopt. Hij ontwikkelde een vragenlijst waarmee de transparantie van het peer-reviewproces bij zowel OA als bij traditionele wetenschappelijke tijdschriften kan worden gemeten.
Wicherts' onderzoek toont aan dat enerzijds veel OA-tijdschriften zich inderdaad schuldig maken aan een falend peer-reviewproces, anderzijds zijn er ook traditionele, subscription based Journals die er een potje van maken. In die laatste groep bevinden zich ook enkele High Impact bladen.
De tool van Wicherts om de kwaliteit van peer review bij zowel Open Acces als trationele wetenschappelijke tijdschriften te meten is beschikbaar via: ***qoam.eu***

COLUMN

#Digilemma

Ooit had ik een grote fysieke bibliotheek. Zo'n veertig kasten vol boeken en tijdschriften, vooral op het gebied van taal- en letterkunde. Maar die collectie werd me tot last en ik maakte er niet optimaal gebruik van.
Daarom heb ik me erin verdiept hoe je boeken moet digitaliseren en indexeren, zodat je grote collecties in één keer kunt doorzoeken. Met vallen en opstaan heb ik dat onder de knie gekregen.
Nu heb ik een grote digitale bibliotheek. Ruim honderdzestigduizend boeken en afleveringen van tijdschriften, het overgrote deel in pdf-formaat. Veel heb ik zelf gescand, of laten scannen door studenten, maar ik heb ook het nodige van internet geplukt.
Met regelmaat word ik benaderd door wetenschappers met de vraag of ik ze bepaalde bronnen digitaal kan leveren. Of een corpus. Bij de Digitale Bibliotheek voor de Nederlandse Letteren zijn, vanwege de beperkingen van de Auteurswet, nauwelijks literaire bronnen van na 1950 voorhanden. Digitaliseren voor 'thuisgebruik' mag van die wet. Ik heb juist duizenden boeken van na 1950 gedigitaliseerd, want die zijn doorgaans makkelijker te scannen dan oude boeken.

Die verzoeken van wetenschappers stellen me voor een dilemma en ik hoor graag uw mening (het liefst via Twitter, @ewoudsanders #digilemma).
Hoe moet ik op dergelijke verzoeken reageren?
Als ik me aan de Auteurswet wil houden, luidt het antwoord altijd: nee. Ja, ik heb alle spellinggidsen van het Nederlands gedigitaliseerd en ja, ik zou je aan een corpus van duizend romans uit de jaren zestig kunnen helpen. Maar nee, ik kan dat niet doen, want daarmee overtreed ik de Auteurswet.

Overigens valt het me op hoe 'kleinschalig' sommige verzoeken zijn. Er is mij weleens gevraagd of ik vijfentwintig titels kon leveren voor een onderzoek naar vertelstructuren in naoorlogse Nederlandse romans. Aangezien er sindsdien tienduizenden romans zijn gepubliceerd, snap ik niet hoe zo'n piepkleine selectie ooit representatief zou kunnen zijn, maar de details van dat onderzoek zijn me niet bijgebleven.

foto Leo van Velzen

Ik zou dus altijd nee moeten antwoorden, maar in de praktijk zeg ik vaak ja. Ik leef van mijn pen en ik begrijp het belang van de Auteurswet, maar mijn hart gaat uit naar de wetenschap. Dus toen ik onlangs werd benaderd door een wetenschapper die dolgraag samen met zijn studenten enkele naamkundige naslagwerken digitaal wilde kunnen doorzoeken, ben ik meteen gaan kijken of ik die toevallig al digitaal beschikbaar had.

Als ik iemand digitale bronnen lever, zeg ik er altijd bij dat ze alleen voor 'thuisgebruik' zijn en dat ze nooit zomaar op internet terecht mogen komen, maar ik begrijp goed dat ik daar geen controle over heb. Bij grote verzoeken ('Het liefst alle duizend streekromans!') antwoord ik: je kunt ze bij mij thuis komen raadplegen.
In feite vind ik dit gerommel in de marge. Het zou niet nodig moeten zijn dat wetenschappers zich wenden tot een goedbedoelende particulier. De belangrijkste standaardwerken op alle vakgebieden zouden onder bepaalde voorwaarden voor wetenschappers en studenten beschikbaar moeten worden gesteld, simpelweg omdat dit de wetenschap vooruit helpt. Hetzelfde geldt voor goed gestructureerde, representatieve literaire corpora. Ik hoor graag ideeën over hoe dat zou kunnen worden gerealiseerd, ondanks de beperkingen van de Auteurswet, die dateert uit 1912 en mijns inziens hoognodig aan een update toe is.

Ewoud Sanders
Taalhistoricus en journalist. Sanders is vaste medewerker van onder meer NRC Handelsblad en Onze Taal.