Google wil graag iets terugdoen

Jon Orwant is engineering manager bij Google en werkt vooral aan Google Books. Hij was in Nederland op uitnodiging van NWO. E-data sprak met hem.

Peter Boot

Eén ergernis voor de niet-Amerikaanse gebruikers van Google Books is dat we sommige boeken niet te zien krijgen, terwijl het auteursrecht toch echt verstreken is.
Orwant: “We kennen niet altijd de precieze publicatiedatum van een boek – we verzamelen boekgegevens uit heel veel bronnen, en het is een heel werk om die te matchen. Het volgende probleem is dat we geen betrouwbaar register van overlijdensdata van auteurs hebben. We moeten voorzichtig zijn, want mensen procederen graag tegen Google - zeker omdat in de VS bij schendingen van het auteursrecht schadevergoedingen tot $150 000 worden toegekend.”

Bij veel digitaliseringsprojecten laat de kwaliteit van de OCR (Optical Character Recognition) te wensen over. Werken jullie daar aan?
“Ja, we hebben een OCR-team, dat trouwens de software als open source ter beschikking stelt. Bij de Gotische letters en klein afgedrukte tekst hebben we de laatste tijd veel vooruitgang geboekt. “

Werkt Google ook aan manieren om tekststructuur in een boek te analyseren?
“Ja. Ik denk dat we onze boeken al kunnen taggen in TEI Lite (TEI: Text Encoding Initiative, standaard voor codering van teksten). Veel verder in de analyse zouden we nu niet durven gaan. We zouden liever een basistekst leveren, en dan de wetenschapper de mogelijkheid geven om de tekst te annoteren. Maar de infrastructuur daarvoor vergt nogal wat. Hoe weten we dat iemand geen onzin typt? Moeten we anoniem commentaar toestaan? Wat moet de eenheid van annotatie zijn: een boek, een bladzijde, een hoofdstuk, een woord? Mag iemand zijn eigen annotaties nog wijzigen?”

Wordt dat een betaalde dienst?
“Nee, het gaat Google goed, en willen iets teruggeven aan de wereld. We verdienen nu wel iets aan de boeken, door advertenties en het verkopen van digitale versies, maar dat haalt het niet bij wat het scannen gekost heeft. In dezelfde spirit zijn de corpora van onze n-gram viewer (een website waar het gebruik van woorden in de tijd kan worden gevolgd – PB) open source beschikbaar. Iedere onderzoeker kan de bestanden downloaden en herpubliceren met een eigen interface.”

Wanneer komt de Nederlandse versie?
“Die komt, maar wanneer weet ik nog niet. We willen eerst zeker weten dat de collectie voldoende representatief is. Van de 3,1 miljoen Nederlandstalige boeken hebben we er nu 168.000 gescand, maar overwegend ouder materiaal. Dat is nog niet genoeg. Houd er trouwens rekening mee dat we voor deze bestanden regelmatig met nieuwe versies zullen komen, omdat we nieuwe boeken hebben gescand en onze software weer beter is geworden. Analyses op de bestanden zullen dus altijd moeten aangeven met welke versie ze zijn uitgevoerd. Daarmee wordt onderzoek in de digitale geesteswetenschappen repliceerbaar. En omdat de zichtbaarheid van onderzoek met digitale bronnen enorm toeneemt, willen we er zeker van zijn dat het onderzoek is dat de toets der kritiek kan doorstaan.”

http://books.google.com/

http://ngrams.googlelabs.com/
http://googlebooks.byu.edu/