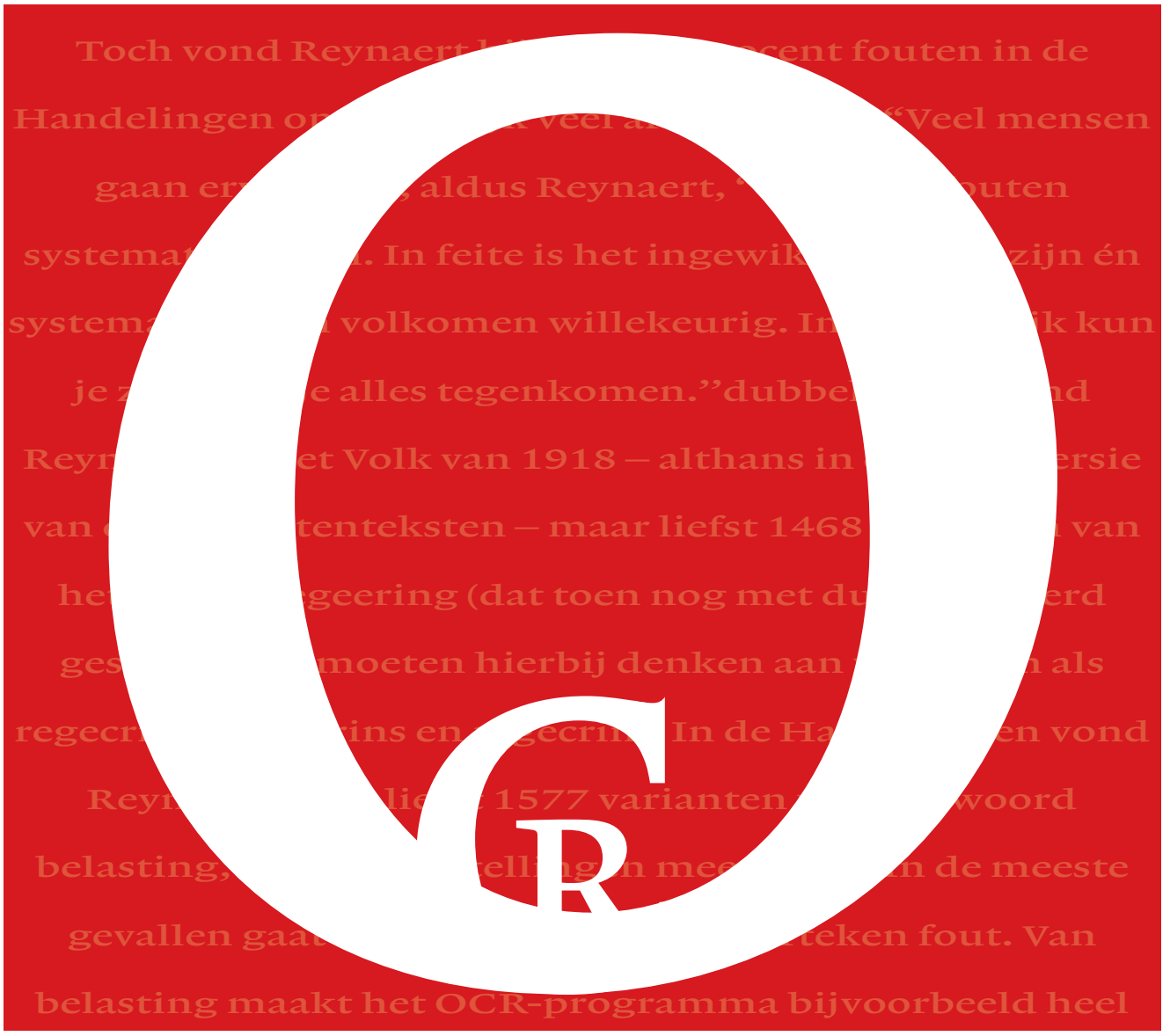
PROGRAMMA INVENTARISEERT DRUK- EN LEESFOUTEN IN OUDE TEKSTEN

1468 varianten van regeering

De leesfouten van de automatische digitalisering van gedrukte tekst zijn niet langer een ramp. Martin Reynaert kan makkelijk zoeken op álle varianten van een woord.

Ewoud Sanders



NAVLOOIEN VAN TEKSTEN zit Martin Reynaert in het bloed. Als pas afgestudeerde vertaler werkte hij ooit als corrector bij de *Gazet van Antwerpen* en later legde hij een grote verzameling van veelvoorkomende tikfouten aan. Ruim vijftigduizend tikfouten telt die verzameling nu, en er komen dagelijks nieuwe bij. Reynaert (44) heeft nu veel baat bij die collectie, want inmiddels is hij als computertaalkundige gespecialiseerd in de automatische detectie van fouten in grote tekstverzamelingen.

Het belang van deze specialisatie is de laatste jaren sterk toegenomen, vooral doordat allerlei bibliotheken en archieven zijn overgestapt op massadigitalisering. Bij massadigitalisering wordt gebruik gemaakt van OCR – *optical character recognition*, optische tekenherkenning. Men scant een boek, tijdschrift of krant en vervolgens wordt dit door een OCR-programma ‘gelezen’, dat wil zeggen: omgezet in digitale tekens, waardoor de tekst volledig doorzoekbaar wordt.

Eind vorig jaar ontstond veel heisa over het plan van de Koninklijke Bibliotheek om boeken los te snijden om het scannen ervan te vergemakkelijken. Alleen zo zou een collectie snel te digitaliseren zijn.

VEROUDERING Bij dit OCR’en worden echter leesfouten gemaakt, of exacter: herkenningsfouten. De vuistregel is: hoe ouder de teksten, hoe meer herkenningsfouten. Door veroudering van het papier is het contrast minder en de oude druktechnieken zijn minder precies. En hoe meer herkenningsfouten, hoe minder goed de ontsluiting van een tekstbestand.

Althans, zo was het tot nu toe.

Reynaert, die is verbonden aan de Universiteit van Tilburg, heeft nu een programma ontwikkeld waarmee je het probleem van de herkenningsfouten grotendeels kunt omzeilen – een programma dat hij onlangs heeft gepresenteerd op een congres van computertaalkundigen in Haifa.

Hoe werkt dit programma, dat TICCL heet, wat een afkorting is van Text-Induced Corpus Clean-up? Reynaert: “Simpel gezegd komt het hierop neer. Eerst maak je een lijst van alle woorden die in een tekstverzameling voorkomen, gesorteerd op frequentie. Neem vervolgens per letter het getal dat de computer intern gebruikt voor die letter en verhef dat getal tot de vijfde macht. Tel die waarden samen voor dat woord. Je krijgt een groot getal. Alle woorden die bestaan uit dezelfde letters, krijgen hetzelfde getal. Kat en tak krijgen dus dezelfde waarde.

“Vervolgens gebruik je die getalswaarden als index”, gaat Reynaert verder, “Met die getallen kun je nu systematisch gaan optellen en aftrekken. Door op deze manier te rekenen vind je automatisch alle varianten van een woord in een bepaald corpus. Op basis van een lijst met correct gespelde woorden kun je nu uit de tekstverzameling alle afwijkende vormen zeven. Dat kan trouwens ook al zonder zo’n controlelijst.”

Hoe ouder de teksten, hoe meer herkenningsfouten er optreden

Om zijn systeem te testen kreeg Reynaert toegang tot twee grote bestanden van de Koninklijke Bibliotheek: alle kranten uit een proefproject voor de digitalisering van historische kranten, plus zes jaargangen van de Handelingen (de letterlijke verslagen) van de Eerste en Tweede Kamer. Uit die kranten nam Reynaert één jaargang, namelijk *Het Volk* uit 1918, een bestand met acht miljoen woorden. Van de Handelingen bewerkte hij zes jaargangen uit de periode 1989-1995, samen goed voor 125 miljoen woorden.

De OCR-kwaliteit van beide bestanden loopt sterk uiteen. Scans van oude kranten zijn door een computerprogramma moeilijk te lezen en volgens de KB was het bij dit proefproject bij ongeveer 70 procent van de tekst goed gegaan. Bij de Handelingen, teksten van recenter datum die beter zijn gedrukt, is 99 procent van de tekst foutloos door de computer gelezen.

Toch vond Reynaert bij dat ene procent fouten in de Handelingen opmerkelijk veel afwijkingen. “Veel mensen gaan ervan uit”, aldus Reynaert, “dat OCR-fouten systematisch zijn. In feite is het ingewikkelder: ze zijn én systematisch, én volkomen willekeurig. In de praktijk kun je zo’n beetje alles tegenkomen.”

DUBBEL ‘E’ Zo vond Reynaert in Het Volk van 1918 – althans in de OCR-versie van deze krantenteksten – maar liefst 1468 varianten van het woord ‘regeering’ (dat toen nog met dubbel ‘e’ werd gespeld). We moeten hierbij denken aan varianten als ‘regecring’, ‘regecrins’ en ‘regecrinf’. In de Handelingen vond Reynaert maar liefst 1577 varianten van het woord ‘belasting’, alle samenstellingen meegeteld.

In de meeste gevallen gaat het slechts bij één letterteken fout. Van belasting maakt het OCR-programma bijvoorbeeld heel vaak ‘belastmg’, ‘beiasting’ of ‘belastlng’.

FOUT GESPELD Het grote nut van TICCL is dat dergelijke tikfouten straks niet meer zoveel kwaad kunnen. Wie nu belasting opzoekt in de digitale versie van de Handelingen, mist alle plaatsen waar dit woord fout is gespeld. Het programma van Reynaert brengt geen correcties aan in de OCR (dat zou grotendeels met de hand gedaan moeten worden), maar in de index waarmee je de tekstverzameling benadert. Omdat TICCL automatisch heeft berekend dat met belastmg eigenlijk belasting is bedoeld, krijg je ook de passages te zien waar het woord op deze manier verminkt is.

Is Reynaert, nu hij zoveel herkenningsfouten heeft gezien, voorstander geworden van de aanpak van de Digitale Bibliotheek voor de Nederlandse letteren (DBNL), die duizenden boeken laat uittikken, juist om OCR-fouten te voorkomen? Reynaert: “Momenteel ben ik de kwaliteit van die teksten aan het onderzoeken. Foutloos zijn ze zeker niet. Als je er bijvoorbeeld ‘hoosd’ opzoekt, krijg je een zestal hits waarbij duidelijk is dat het ‘hoofd’ had moeten zijn. En je vindt er ook ‘Amfteldam’, ‘Amfterdam’ en ‘fchrijven’. Hoe dan ook, TICCL kan dat op termijn allemaal oplossen.”