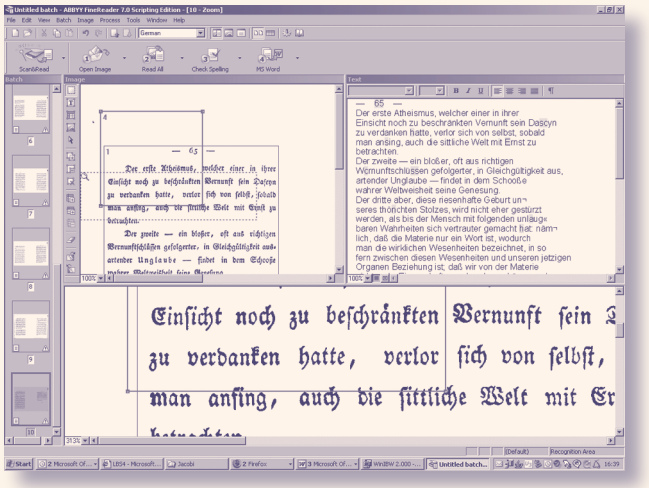
OCR van oud drukwerk

Jacob van Sluis j.van.sluis@rug.nl

en dat van Fraktur in het bijzonder



OCR-screendump

Tegenwoordig worden teksten digitaal aangeleverd, ook wanneer ze alleen op papier worden verspreid. Bij oudere teksten daarentegen bezitten we vaak alleen een papieren versie, die dan niet digitaal doorzoekbaar is. Na scannen is het als plaatje ook digitaal beschikbaar, echter slechts als plaatje op het beeldscherm en nog niet doorzoekbaar.

Met OCR-software is het mogelijk om gedrukte teksten vanaf papier of als schermbeeld om te zetten tot een digitale tekst. OCR staat voor *Optical Character Recognition*. Het OCR-inlezen van moderne teksten is vrij probleemloos, maar oud drukwerk confronteert ons met extra problemen en leidt tot veel leesfouten.

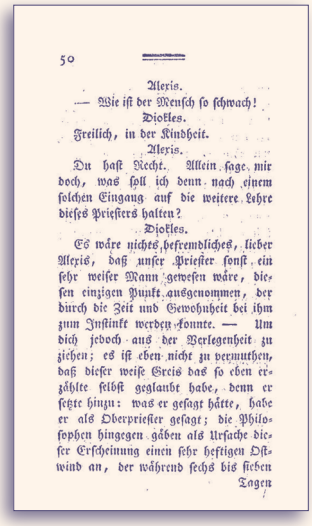
Oude teksten en hun problemen

Oude boeken kan men betrekkelijk eenvoudig scannen en via internet wereldwijd ter beschikking stellen. We kunnen dan thuis bladeren, terwijl het boek zelf verder in de kast blijft. Maar achter het scherm gezeten willen we graag de verdere mogelijkheden van de computer benutten. Bijvoorbeeld, zoeken op woorden in dit 'boek'. Maar voor de computer is een plaatje iets geheel anders dan een tekstbestand.

OCR van oude gedrukte teksten kent een aantal specifieke problemen. Bij het inlezen van moderne teksten wordt een aanvaardbaar foutpercentage gehaald. Maar bij oud drukwerk neemt het percentage aan fouten snel toe, vanwege een aantal redenen. Modern offset drukwerk is bijzonder regelmatig, maar het drukken vanuit losse loden letters kent veel meer variabelen, zeker wanneer het zetsel met de hand vervaardigd is. Voor elke losse letter is een afzonderlijk stempel gebruikt, weliswaar mechanisch vervaardigd maar toch met kleine onderlinge afwijkingen: haarscheurtjes (bramen), een wat dichtgeslibde opening, variabele slijtage, of dansende letters die niet allemaal keurig op de lijn staan. Hoe goedkoper het drukwerk, hoe minder aandacht er aan het zetsel werd besteed.

Daarnaast zijn er problemen die voortkomen uit het gebruikte materiaal dat een uiteenlopende kwaliteit kende. Een vilterige papiersoort doet letters sneller dichtslibben. Wanneer het papier in de loop de jaren donkerder is geworden (bruiner), wordt voor de inlezende scanner het contrast van de letter tegenover haar achtergrond kleiner en bijgevolg de 'leesbaarheid' kleiner.

Bij het inlezen van teksten uit de vroege negentiende eeuw stuitte ik op nog een ander bijkomend probleem. Zetwerk vervaardigd met lettertypen uit de Bodonifamilie – toen zeer populair – is slecht herkenbaar omdat de letters gekenmerkt worden door


50

Alexis.
— Wie ist der Mensch so schwach!
Diokles.
Freilich, in der Kindheit.
Alexis.
Du hast Recht. Allein sage, mir doch, was soll ich denn nach einem solchen Eingang auf die weitere Lehre dieses Priesters halten?
Diokles.
Es wäre nichts befremdliches, lieber Alexis, daß unser Priester sonst ein sehr weiser Mann gewesen wäre, diesen einzigen Punkt ausgenommen, der durch die Zeit und Gewohnheit bei ihm zum Instinkt werden konnte. — Um dich jedoch aus der Verlegenheit zu ziehen; es ist eben nicht zu vermuthen, daß dieser weise Greis das so eben erzählte selbst geglaubt habe, denn er setzte hinzu: was er gesagt hätte, habe er als Oberpriester gesagt; die Philosophen hingegen gäben als Ursache dieser Erscheinung einen sehr heftigen Ostwind an, der während sechs bis sieben Tagen

Een pagina van Hemsterhuis in Duitse vertaling in Fraktur-letters, met matig resultaat voor OCR.

stevige verticalen met dunne, soms met haarfijne horizontale lijnen. Het onderscheid tussen de letters c/e, b/h, f/i/l/l en n/u, maar ook tussen punt en komma, of tussen m/n, n/m, nin/mn en nog veel meer van dergelijke combinaties, wordt voor de computer onhaalbaar.

Een slecht resultaat behoeft echter niet tot wanhoop of tot moedeloosheid te leiden. Er bestaat een aantal trucs om het foutpercentage positief bij te stellen, door met 'zoek & vervang' een aantal lettercombinaties weg te halen. Bijvoorbeeld in Nederlandse teksten valt de woorduitgang c+n+spatie te vervangen door e+n+spatie of e+h door c+h, of in Franse teksten q+n door q+u. Of een correctie in twee stappen: eerst elke H te vervangen door dubbel-l, vervolgens elke spatie+l+l door spatie+H.
Het teleurstellende eerste resultaat van OCR kan zo tamelijk snel verbeterd worden. Het is bijzonder bevredigend om correctierondes via 'zoek & vervang' te bedenken en uit te voeren. Dergelijke handelingen lonen meer naarmate de gescande tekst langer is. Hoe slim je ook werkt met zoek & vervang-formules, uiteindelijk blijft er veel handmatige correctie over. Het toepassen van OCR op oud drukwerk blijft arbeidsintensiever dan dat bij moderne teksten.

Extra problemen bij gotische letters

Bij OCR van gotische letters of 'Fraktur' zijn de problemen weer groter. Allereerst moet het OCR-programma in staat zijn om dit type letter te herkennen en dit is voorbehouden aan de duurdere programma's, waarover later meer. De varianten tussen verschillende lettertypen van Fraktur is zo mogelijk nog groter dan bij de gewone Romeinse letter, vaak afhankelijk van de streek van herkomst. Fraktur wordt per definitie gekenmerkt door sterke verticale en zwakke horizontale lijnen. Fraktur werd vooral gebruikt voor drukwerk in de volkstaal, daarmee gericht op een groter maar minder kapitaalkrachtig publiek, waardoor ook snel gekozen werd voor haastig zetwerk en goedkoper papier.
Ook Fraktur kent letters die onderling sterk op elkaar lijken, bijvoorbeeld wederom c/e en n/u. Er komen nieuwe probleemletters bij. De stok van de letter-d wijkt terug tegen de leesrichting in, waardoor er verwarring met de letter-b ontstaat. De veel gebruikte stok-s, die overigens tot in diep in de achttiende eeuw ook in Romeins zetwerk voorkomt, zorgt voor verwarring met de letter-f. Hoofdletters worden gekenmerkt door zwierige krullen, wat leidt tot bijvoorbeeld een verwarring rond B/G/S. Brede letters als -m en -w worden bij inlezing gemakkelijk opgedeeld in kleinere letters als n/u/i.
Want OCR-programmatuur leest letters als patronen gescheiden door verticale onderbrekingen; wanneer een horizontaal verbindingsstreepje door een gebrek aan contrast wegvalt, dan valt de betreffende letter uiteen. Patroonherkenning berust nu eenmaal op een volledig patroon. Gebrek aan contrast, niet alleen van de letter tegen haar achtergrond, maar ook tussen verticale en horizontale lijnen binnen de letter, leidt tot een veelheid aan leesfouten.

Eenvoudige en betere OCR-programma's

Bij aankoop van een scanner wordt doorgaans met de software ook een eenvoudig OCR-programma meegeleverd. Dat werkt bij recent geproduceerde teksten doorgaans heel bevredigend. Maar voor ouder drukwerk schieten deze programma's al snel te kort. Er zijn ook weinig aanvullende mogelijkheden om fijner af te stemmen op de te lezen tekst en – nog belangrijker – er ontbreekt een trainingsoptie. De betere programma's beschikken over deze functie om de graad van herkenning beter af te stemmen op het specifieke drukwerk dat men wil verwerken. Een afwijkende letter, die elders weinig voorkomt en die anders tot een consequente fout leidt, kan men bij inlezen definiëren tot de juiste letter. Zo is het mogelijk om het onderscheid tussen probleemletters als c/e en n/u te verbeteren.
Een verdere mogelijkheid bestaat in het leren herkennen van ligaturen: hieronder wordt verstaan dat bepaalde vaste lettercombinaties aan elkaar geschreven zijn, d.w.z. zonder blanco verticale strook ertussen. Zo smelten in ouder drukwerk veelvuldig de letters f+i samen tot en f+l tot , of ontstaan er lettercombinaties via een overbruggende krul tussen c+t en s+t. De Duitse Ringel-s is te lezen als een ligatuur van tweemaal een stok-s, zoals het &-teken te herleiden is tot het Latijnse 'et' met een overbruggende krul. Via een trainingsoptie kan het programma leren om zulke ligaturen te herkennen en om te zetten in de juiste lettercombinatie.
Dat bijleren gaat als volgt. Na het inschakelen van deze optie wordt het document stapsgewijs doorlopen en bij elk letterteken dat het programma niet kan thuisbrengen, maakt het een stop, doet een suggestie voor een omzetting en wacht tot er een bevestiging of een verbetering is ingetikt. Gaandeweg worden steeds minder tussenstops gemaakt.

Wat bij een eenvoudig OCR-programma ook ontbreekt, is de mogelijkheid om een onderscheid te maken tussen verschillende lettertypen die kunnen worden herkend, bijvoorbeeld: 'normal', 'typewriter' en zelfs een meegeleverde 'gothic'. Het toepassen van een dergelijke optie zorgt voor een betere afstemming en verhoogt de herkenning aanmerkelijk. Dat wordt zelfs opvallend wanneer in de Fraktur-tekst een passage in Romeinse letters voorkomt, bijvoorbeeld een citaat in het Latijn: plotseling kan het programma vanuit de gothic-

optie niet meer goed de gewone letters herkennen. Een dergelijke keuzemogelijkheid in combinatie met de trainingsoptie kan een goede OCR-functionaliteit opleveren, die bij de eenvoudige programma's volstrekt onhaalbaar is.

Scannen en herkennen

Wat een OCR-programma doet is de weg afleggen van inlezen van beeldtekens en die omzetten tot lettertekens vergelijkbaar met die welke na toetsaanslagen op het scherm komen. Op basis van een patroonherkenning wordt de 'meest gelijkende' ASCII-code erbij gezocht. Anders gezegd: eerst is er de beeldherkenning (scannen) en vervolgens de omzetting (herkennen).

De ervaring leert dat de mate van OCR-herkenning aanmerkelijk verhoogd kan worden door beide niveaus goed te onderscheiden en optimaal op elkaar af te stemmen. Heel effectief is om zodanig te scannen dat het herkennen met zo weinig mogelijk ruis wordt geconfronteerd. Nog voordat het OCR-programma op de tekst wordt losgelaten, probeert men zo duidelijk mogelijk te scannen.

Een voorbeeld om dit te verduidelijken: rafelige letters leiden tot vele OCR-fouten, en je zou kunnen menen dat de herkenning van deze letters het best kan worden ondervangen via een verdere fijnafstemming in de bijleerfunctie. Veel beter is het echter om te zorgen dat dergelijke letters bij het inscannen niet rafelig worden ingelezen. Dat vraagt enig experimenteren met de instelling van de software van de scanner, door voor de helderheid en het contrast andere waarden dan de standaardwaarden uit te proberen. Het contrast van de letters tegen hun achtergrond moet optimaal zijn. De ervaring leert dat OCR het beste werkt wanneer men scant op zwart/wit (of aldus wegschrijft voor OCR-bewerking), en dus niet op kleur of op grijswaarden. Het kan raadzaam zijn om een digitaal plaatje van een aangeleverde tekst eerst uit te printen en die vervolgens opnieuw in te scannen zonder grijswaarden. Want digitale beeldbewerking, bijvoorbeeld contrastversterking via Photo Express, blijkt weinig effectief.

De conclusie is dat het scannen voorafgaand afgestemd moet zijn op het herkennen. De letters moeten zo contrastrijk en helder mogelijk worden, en idealiter moet de achtergrond van het papier verdwijnen. Zoveel mogelijk vlekjes en puntjes van het papier moeten via de zeef van de juiste instelling van de scanner weggefilterd worden, zodat de OCR-fase niet struikelt over moeilijk identificeerbare 'tekens'. En dat alles op een wijze dat de letters niet zo bleek worden dat de tekenherkenning daardoor weer lastig wordt. Doorgaans geldt dat de achtergrond als storende factor voor een belangrijk deel automatisch weggezuiverd kan worden. Een scan geschikt voor OCR ziet er dan ook veel schoner uit dan het origineel, maar is weer niet geschikt om als waarheidsgetrouwe illustratie te dienen.

Eigen ervaringen met Fraktur

De door mij gestelde taak was om ongeveer 800 pagina's tekst uit de *Vermischte philosophische Schriften*, verschenen in de jaren 1782-1797, van de filosoof Frans Hemsterhuis in te lezen. Het betrof Fraktur-drukwerk in het kleine octavo-formaat, met kleine letters dus, en op papier dat bij bewerking van matiger kwaliteit bleek dan op het eerste gezicht: vaak schemerden de letters vanuit de andere kant van het papier door en die zorgden voor storingen. Bij de OCR-werkzaamheden bleek de taak lastiger dan voorzien.

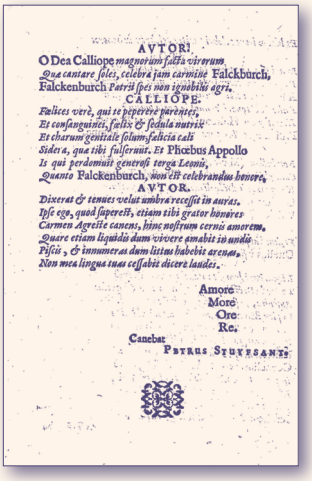
Ik beproefde twee programma's, beide geleverd door Abbyy, een belangrijk leverancier voor dergelijke programma's. Abbyy Finereader XIX (Engelse versie) heeft een module voor 'Gothic', met bovendien een trainingsoptie. Daarnaast probeerde ik de Nederlandse versie van Finereader 8.0, weliswaar met leeroptie maar zonder module voor Fraktur.

In beide programma's moet de in te lezen taal worden ingesteld, in mijn geval Duits; dit is nodig omdat anders de Ringel-S ingelezen wordt als een hoofdletter-B en de Umlaut tot verwarring leidt. De mogelijkheid om tevens een spellingcontrole te laten meelopen (waaronder de mogelijkheden van Duits en Nederlands) was onbruikbaar, omdat de achttiende-eeuwse spelling te zeer afwijkt van de huidige. Beide programma's waren bruikbaar, al viel het resultaat me in beide gevallen tegen: alleen met slim zoek & vervang kon ik het resultaat aanvaardbaar maken. Vanwege de grote hoeveelheid tekst (ruim 80.000 woorden, meer dan een half miljoen aanslagen) was het proces desondanks lonend.

Overigens, een andere Frakturtekst (F.H. Jacobi, *Werke*, verschenen in 1819, in een herdruk uit 1968) leverde een zeer acceptabel resultaat op in Finereader XIX: een grotere maat letters en (dankzij de reprint) een veel betere zwart/wit-balans.

Conclusie

OCR van oude teksten en Fraktur is mogelijk, maar verre van volmaakt. Rechtstreeks ingelezen is het resultaat naar foutpercentage gezien uiterst teleurstellend. Maar naast deze harde conclusie is een nuancering mogelijk. Laat men het inlezen gepaard gaan met een gedegen voorbereiding (in de vorm van een scans met een optimaal contrast) en met een sluwe nazorg (om de regelmaat in de foute lezingen via routines te corrigeren), dan kan men zich veel werk in de vorm van overtypen besparen. Maar zelfs dan geldt: na het mechanisch OCR-inlezen van de tekst volgt onvermijdelijk een visuele en letterlijke (!) controle, over de volle lengte.


AVTOR!
O Dea Calliope magnorum fabra virorum
Qua cantare soles, celebra jam carmine Falckburch,
Falckenburch Patrii spes non ignobilis agri.
CALLIOPE:
Felices verè, qui te peperere parentes,
Et consanguinei, fulix & fedula nutrix
Et charum genitale solum felicia cali
Sidera, que tibi fulserunt. Et Phœbus Appollo
Is qui perdomuit generosi terga Leonis,
Quanto Falckenburch, non est celebrandus honore,
AVTOR.
Dixerat & tenues velut umbra recessit in auras.
Ipse ego, quod superest, etiam tibi grator honore
Carmen Agreste canens, hinc nostrum cernis amorem.
Quare etiam liquidis dum vivere amabit in undis
Piscis, & innumeras dum lustra habebit arenas,
Non mea lingua tuas cessabit dicere laudes.
Amore
More
Ore
Re.
Canebat
PETRUS STUYVESANTS

Een Latijns gedicht van Peter Stuyvesant uit 1630, dat een zeer slecht OCR-resultaat geeft. (collectie Tresoar Leeuwarden)