

5. OCR

5.1 Inleiding

OCR staat voor Optical Character Recognition. Toepassing van OCR is belangrijk omdat dan de plaatjes van teksten omgezet worden naar teksten zelf. Dit betekent dat de tekst niet langer uit één beeld van zwarte en witte puntjes bestaat, maar uit een aantal ASCII tekens, zodat de tekst gebruikt kan worden in een tekstverwerker voor aanvullingen, mutaties, verbeteringen of annotaties. Tevens kan een zoekprogramma deze ASCII tekst doorzoeken.

Er zijn grofweg twee soorten OCR-software. De een werkt volgens het principe van automatische font-herkenning, dat wil zeggen het programma herkent bestaande fonts en zet ze om in digitale tekst. Het andere werkt volgens het principe van patroonherkenning, dat wil zeggen dat het programma geen bestaande fonts herkent, maar zo geprogrammeerd is dat er ruimte is om allerlei patronen (lettervormen) te leren, dus in het geheugen op te slaan en daarna te herkennen.

Op dit moment is er geen OCR-programma dat 17^e eeuwse teksten zonder meer kan herkennen. De bestaande OCR-programma's zijn ontwikkeld om eigentijds drukwerk te behandelen. De moeilijkheid van de Menasseh collectie is in de eerste plaats het feit dat het hier gaat om 17^e eeuwse teksten, d.w.z. dat de drukkers werkten met handgemaakte loden letters. Het gevolg hiervan is dat de letters onderling verschillen omdat ze van verschillende afgietsels afkomstig kunnen zijn.

De OCR-programma's die volgens het principe van automatisch font-herkenning werken, herkennen alleen de variëteit die in het geheugen ligt opgeslagen. Als er van elke letter meer dan 30 variëteiten kunnen optreden, kunnen de programma's dat niet bolwerken.

Voor de experimenten met de Hebreeuwse teksten was het nodig te onderzoeken wat representanten van beide soorten programma's konden presteren. Bij toepassing van OCR-technieken op Hebreeuwse letters treedt de beperking op dat alleen die programma's voor OCR in aanmerking komen die voor Hebreeuws ontworpen zijn zoals ERA, ART en Ligature of programma's die zelflerend zijn zoals proLector en in mindere mate Omnipage Pro.

5.2 Resultaten Hebreeuwse OCR-programmatuur

ERA: via e-mail is naar Israël een deel van de testcollectie gestuurd, waar men het OCR-programma ERA heeft toegepast. Helaas waren de resultaten niet bevredigend. ERA is gebaseerd op Wordscan. Dit programma, waarover de UB beschikt, heeft bij de digitalisering van teksten in het Latijns schrift al een hoger foutgehalte dan bijvoorbeeld een programma als Omnipage.

ART: via Menachim Nir te Bney Brak, Israël is hiernaar gekeken. Ook hier waren de resultaten verre van bevredigend, omdat te weinig letters door het programma herkend werden.

Voor OCR toepassingen op Hebreeuws is Ligature, opgericht in 1989 in Israël, wereldwijd het meest gebruikt. In 1996 waren er meer dan 1 miljoen gebruikers van Ligature's OCR. Via Gideon Ben-Zvi in Jerusalem, werd een test gedaan.

Evenals de eerder genoemde Hebreeuwse OCR-software is dit een programma dat werkt volgens automatische font-herkenning, en het font werd te slecht herkend door de onderlinge verschillen. In de woorden van Michal S. Berger van de Customer Service afdeling van Ligature:

“Once again, regarding the TIFF you sent, there is still not enough consistency in the shapes of the specific letters nor enough distinction between the different letters, i.e., the same letter looks different each time, and the different letters resemble each other. It is therefore impractical to try and teach such a font as the result would be unacceptable to you ”.

De moeilijkheid ligt dus inderdaad uitsluitend aan de geaardheid van 17^e eeuwse letters. De moderne letters, die door de computer zijn vervaardigd en elk identiek zijn, geven heel wat minder problemen.

Aangezien de Hebreeuwse OCR-programma's geen bevredigende resultaten opleverden, is het meest gebruikte programma ter wereld, Omnipage Pro 7.0, als representant van een programma dat werkt volgens de automatische font-herkenning, gekozen. Als vertegenwoordiger van programma's die werken volgens patroon-herkenning is proLector 1.20D van Improx aangeschaft. Beide OCR-programma's zijn uitgebreid geëvalueerd.

5.3 Evaluatie Omnipage en ProLector

Testmateriaal:

1. Een pagina met duidelijke moderne Hebreeuwse tekst met HP DeskScan II gescanned.
2. Een Engelstalig boekje uit de Menasseh collectie (To his Highnesse The Lord Protector of, sign. 19F5), font vergelijkbaar met 32 pnts Times^{*)}, met een aantal afwijkende letters, de regels zijn regelmatig maar 'springerig', frequent cursief, geïllustreerd met versierde letters, kwaliteit van de film nog niet optimaal waardoor veel 'gebroken' letters voorkomen.
3. Een Hebreeuws boek uit de Menasseh collectie (ׁׂ׃ׅׄ׆ ׇ׈׉׊ ׋׌׍ - Sefer ma'ajane ha-jesju'ah, sign. 1889E25), pagina's in twee kolommen gezet, tekst in twee fonts, kleine letter^{*)} (vgl. 32 pnts Times) met blokken grote letters (vgl. 40 -72 pnts Times), geen illustraties, redelijke papierkwaliteit.

*) De kwalificatie 'kleine letter' staat wat vreemd naast 32 punts Times, maar de verklaring is: beide boeken zijn van kwarto formaat (4°). 'Klein' slaat op de originele letters, vgl. 9 punts Times, in het boek. Voor OCR zijn de opnames omgezet in TIF-images van gemiddeld 4000 × 3000 pixels. Dit is een aanzienlijke vergroting. Voor kleine boeken uit de collectie is die vergrotingsfactor nog aanzienlijker, waardoor de GIF-images goed leesbaar worden. De grote boeken houden relatief kleine letters.

Vorbereiding

Eerst moest uitgezocht worden hoe Hebreeuwse teksten op een webpagina aangeboden en gelezen kunnen worden, uitgaande van een Engelse Windowsversie en een U.S.-international toetsenbord.

Via Internet (URL: http://www1.snunit.k12.il/heb_pc.html) zijn drie zogeheten eurofonts voor Hebreeuws opgehaald, te weten:

- WebHebrew AD, d.i. het meest gebruikte font Hadassa vergelijkbaar met Times
- ElroNet pro, vierkanter en dikker, vgl. schreefloze Univers
- Netextpro vgl Arial

Kenmerkend van deze fonts is dat de Hebreeuwse tekens weergegeven worden door ASCII-codes tussen 224 en 250. De overige codes genereren het Latijnse schrift. Op deze manier kan een webpagina zowel Nederlandse, Engelse als Hebreeuwse tekst bevatten. In onderstaande tabel staan de ASCII codes en de bijbehorende letter in het Hebreeuws. Het gebruikte font is WebHebrew AD.

Resultaat in	Code	Resultaat in	Code	Resultaat in	Code
Hebreeuws	Num+ALT+0+	Hebreeuws	Num+ALT+0+	Hebreeuws	Num+ALT+0+
alef	א 224	bet	ב 225	gimel	ג 226
dalet	ד 227	he	ה 228	waw	ו 229
zajin	ז 230	chet	ח 231	tet	ט 232
jod	י 233	sluitchaf	ך 234	kaf	כ 235
lamed	ל 236	sluitmem	ם 237	mem	מ 238
sluitnoen	ן 239	noen	נ 240	samech	ס 241
ajin	ע 242	sluitfe	ף 243	fe/pe	פ 244
sluittsadi	צ 245	tsadi	צ 246	koef	ק 247
resj	ר 248	sjin	ש 249	taw	ת 250

ASCII-codes kunnen ingevoerd worden via het numerieke keyboard met Num Lock aan. Druk ALT in, type 0 gevolgd door de gewenste code, houd de ALT-toets steeds ingedrukt.

Het merendeel van de ASCII-codes tussen 224 en 250 zijn ook rechtstreeks via het US-international toetsenbord in te typen door letters met accenten.

Op de volgende pagina staat dezelfde tabel in font Times.

Resultaat in	Code	Resultaat in	Code	Resultaat in	Code
Hebreeuws	Num+ALT+0+	Hebreeuws	Num+ALT+0+	Hebreeuws	Num+ALT+0+
alef	à 224	bet	á 225	gimel	â 226
dalet	ã 227	he	ä 228	waw	å 229
zajin	æ 230	chet	ç 231	tet	è 232
jod	é 233	sluitchaf	ê 234	kaf	ë 235
lamed	ì 236	sluitmem	í 237	mem	î 238
sluitnoen	ï 239	noen	ð 240	samech	ñ 241
ajin	ò 242	sluitfe	ó 243	fe/pe	ô 244
sluittsadi	ø 245	tsadi	ö 246	koef	÷ 247
resj	ø 248	sjin	ù 249	taw	ú 250

Zoeken in Hebreeuwse tekst met een standaard toetsenbord gaat ook via de ASCII-codes.

Resultaten:

- ad. 1 De tekst bestond uit 1596 tekens.
 In proLector was training van 42 patronen (accuracy 2), waaronder één tweetal, nodig voordat automatisch lezen perfect ging, inclusief de spatiëring. Dit was ca 15 min. werk.
 Om hetzelfde resultaat met Omnipage te bereiken was een training van 86 patronen nodig, waaronder 19 tweetallen. Doordat voor de training van één teken meerdere handelingen moeten worden verricht duurde dit ca. 45 min. Een afdruk van de resultaten zijn vergeleken met het origineel. Op het oog waren het perfecte kopieën.
 Op pagina 15 staan de originele tekst en het resultaat in Hebreeuws naast elkaar, op pagina 16 staat het origineel naast het OCR resultaat in Times.