8.5 OCR op Sefer ma'ajane ha-jesju'ah (Bibl.Ros.1889E25)

Kenmerken:

- Hebreeuwse tekst zonder punten in Rashi font
- duidelijke afbeelding met weinig storende vlekken
- kleine letter, ca 25 pixels hoog, interlinie ca 30 pixels waar de stokken en staarten met ca. 20 pixels insteken. Alineas beginnen met een stukje tekst in grote letters (40 - 60 pixels hoog), die wel op de regel blijven.
- Uitvullen van de tekst gebeurt door wit tussen woorden en vooral aan het eind van een zin. Letters worden niet opgerekt.
- vrijwel geen tekst in de marges
- soms tussen de regels geschreven en woorden onderstreept, maar alleen in de pagina's die gebruikt zijn voor batchverwerking
- regelmatige bladspiegel

Resultaat:

Instelling van proLector: dirt size=5 (maximaal), accuracy=1 (minimaal).

Voor het goed herkennen van de meest linkse kolom van opening 6 (f6v) moesten 112 patronen getraind worden. Lezen van de rechterkolom van dezelfde bladzij (tweede van links van opening 6) met deze trainingsset gaf 84 onherkenbare tekens, weergegeven met ¥, dat is 4%. Om die goed te herkennen is verder getraind tot 154 patronen.

Met deze set is opening 16 automatisch gelezen met 3% niet-herkende tekens

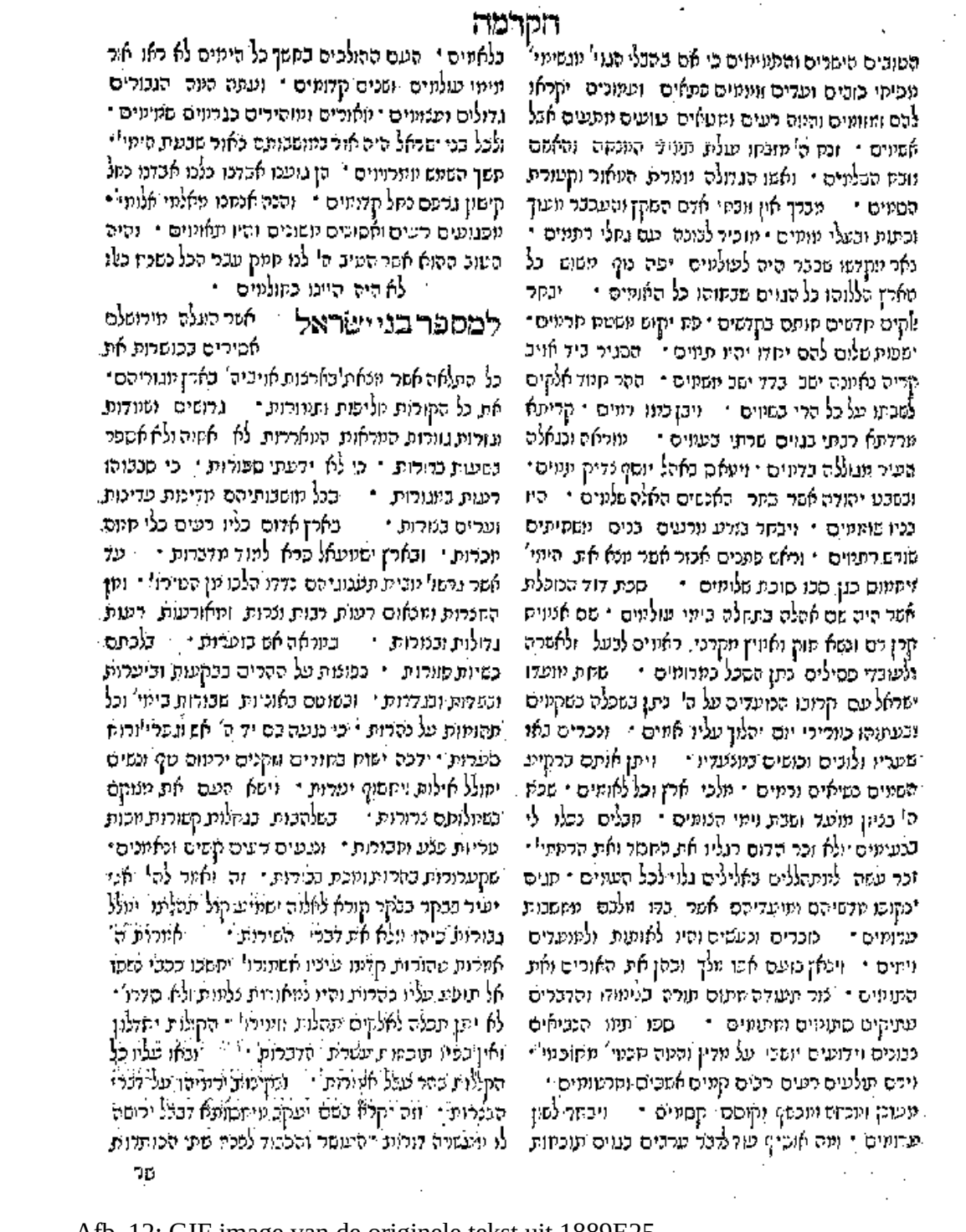
Instelling van proLector: dirt size=5 (maximaal), accuracy=2 (min=1, max=10).

Op f6v (3571 tekens) 200 patronen getraind.
Met deze set f5r (opening 6 de twee kolommen rechts, zie pagina 40) gelezen: 3554 tekens waarvan 212 niet-herkend, d.i. 6%

De meeste ¥'s zijn te vinden onderaan de linker kolom waar veel spikkels de tekst vertroebelen. Zie pagina 41.
Verder werd een blokje grote letters, die nauwelijks getraind waren, niet herkend

Door onervarenheid met het Rashi font komen er ook fouten voor die veroorzaakt zijn door verkeerde training van sommige letters. Verder klopt de spatiëring niet altijd, maar dat kan in proLector beter ingesteld worden.

Door de regelmatige bladspiegel leent dit boek zich ook voor batchverwerking. Met dezelfde trainigsset van 200 patronen zijn de pagina's 61 tot en met 90 automatisch gelezen. Dit resulteerde in 20% niet-herkende tekens, grotendeels veroorzaakt door met pen onderstrepen van woorden en tussen de regels schrijven. Ook grote letters werden niet herkend.



Afb. 12: GIF image van de originele tekst uit 1889E25

הפובים היש¥יסוה¥ויייוים ¥י אם בהבליהגויי יוגשיחי¥ו
יופיקי כובים ועדים וויוחים פתאים ועחונים י¥ראו
להם וחוויוים והיוה רעים יסטאים טועים חתי¥ים¥אבל
'¥שיוים . זנק י ' חזבקו עולת תיויד היונסה והאשם
יובס השלחים י ואשי הגדולה יונורת היואור וקטורת
הסחים . חברך און זובקי אדם השקיוהעכבר יועו¥
וכתותיב¥י יוייוים .חוכירלבונה ¥ם גקלי רוחים .
נאר ייקדשו שכבר היה לעולחים יפה נוף חשוש כל
הארץ הללוהי כלהגיים שבסיהו כל האויוים . יבסר
י¥קים סדשים סיתם בקדשים .פס יקיש יושטס סריוים.
ישפותשלים להם יסדו יהיו תיחים. הסגיר ביד אייב
¥¥¥ריה נאחנה ישב ברד ישב משחים . ההר סחדאלקים
לשבתו על כל הרי כשחים . ויבוכמו רחים . קריתא
יורדתא רבתי בגוים שרתי בעיוים . חוראה ונגאלה
י¥עיר חגוללה בדיוים . וי¥אס באהל יוסיפצדיק תיוים.
ובשבט יהודהאשר במר האנשים האלהשליוים . היו
בניו שומחיס . ויבסר ב¥רע יורעים בנים חשסיתים
שורשרתחים . וראש פתנים אכוראשר חנא את תיחי,
¥¥סחום כגן סבו סוכתשלוחים . סכת דו¥ הנ¥פלת
¥שר היה שם אהלה בת¥מלה ביחי ¥ולחים . שם אניויס
¥¥ירן רם ונשא סזק יאיויקח"¥רני¥ ראיוים לבעל זלאשרה
¥לעובדי פסילים נתן הסכל כמרומים . שסתיוועדו
.שראלעם קרובו הכועדיסעל ה' נתן בשפלה כשקיוים
¥בעתוהו כחרירי יום יהלוקעליו אחי¥ . ונכרים באו
ש¥יו ולובים וכושים בחצעדיו¥. ייתן אותם ברקיע
השיייים כשיאים ורחים . יולכי ארץוכללאוחים . שכס
הי בצי¥ן חוער ושבת ויחי הנוחים . קכלים נפלי לי
בנעיתים.ולא וכר הרום רגליו א¥הססד ואת הרסיויי.
¥כר עששה ליותיהללים באלילים גלוי לכל העיוים . סגים
ינקופו קדשיהם וח¥עדיהם אשר בדי יילבם ממשבות
¥רוחים. נוכרים ונע¥שים והיו לאותות וליוועדים
ויחים . וינאיכועם א¥ו יולך וכהן את האורים ואת
התויוים . נור ת¥עודהקתום תורה כ¥ייודו יהדברים
עתיקים סתויוים יסתויוים . ספי תיוו הנביאים
נבונים וירועים יושבי על חדין וחחה סכיוי יומוכיו¥¥¥
וירם תולעים רעים רבים קיים¥שפים וקרטוûים .
¥ועונן
ויינ¥וש וחכשו.¥ ו¥¥וסס¥ קסחים . וי¥ב¥¥¥לשון ¥
¥¥ג¥וחים . וחה אזסי¥¥ עוד¥בר ערבים כגוים תוכסות ¥

בלאיוים י העם ההולכים בקש¥כל הימים לא ראו אור
חיחי עולחים ושנים קדוחים .¥ ועתה היוה הגבורים
גדולים ועצייוים .יואורים ויווהירים כגרחיס שמיחים .
ולכל בני ישראל היהאורבחושבותם כאורשב¥תהיחי¥.
ששך השחש יומרויוים . הוגוועני אבדנו כלכו אבדנו נמל
קישון גרפסנסל ¥דוחים . והנהאנמנו חאלחי אלויוי¥.
יופגועים רעיסואסו¥ים יושונים והיו תאויוים . יהיה
הטוב ההוא אשרהטיב הי לנו סחק עבר הכל נשנ¥ו כי¥ו
לא היה היינו כסוליוים .
¥¥¥¥¥י¥¥י ¥¥¥¥¥ אשרהוגלה מירושלם
אסירים בכושרותא¥
כל התלאהאשר יוצאת'בארצותאויביה, בא¥.ץמגוריהם.
את כל הקורות קליפות ותחורות. גרושים ושיודות
וגזרותגזורית היוראות התאררות לא אמ¥הולאאספר
בשעותבר¥רות . כי לא יד¥נתי ספורות. כי סבבוהו
רעות בחגורות . בכל מושבותיהם מדינותעדינות
וערים בצורות. כארצא¥דום כליי רעים כלי סיוס
חכרות. ובארץ ישיועאל פרא לחוד מדברות . ע¥
¥שר גרשו' יוביתתענוגיהם נדדו הלכו חן הטירו' י וח¥
הקצרותוחצאום רעות רבותוצרות וחאורעות רעות
גדוליתובצורות . כיוראהאש בוער¥. בלכתם¥
כשיותפוורות . כפוצותעל ההרים בבקעות וביערית
זב¥¥דותובנ¥¥ות. ובשוטם באוניות שבורותביחי' וכל
תיהוחות על נהרות.¥כי נגעיהבם יד ¥¥, אש וגפרייורוס
סערות. ידכה ישוק בסזרים ו זקנים יריוום טף ונשים
יחולל אילו¥ויסשוף י¥ירית . וישא העם אתמצוקם
בשחלותם נרורית.¥ בשלהבות בנ ילותקשורותיוכות
טריותפצעומבירות. ינגעים רעים קשים ונאחנים.
שקערורותבהרותוחכת בכ¥רות. וה וא¥ור ליה' א¥¥
יעירבבקר
בב י¥ר ¥¥ורא¥¥¥,¥לוה ישמי.¥¥ון¥ ¥¥¥לתו יחל¥ל
גבורותפיהו חלא ¥תדברי ¥שירות. אחר¥¥¥הה
אחרית¥הורות קד¥יו עיניו אשחורוי ימש¥ו ككבי נשפ
אל תופע¥עליו נהרותוהיו לחאו ¥ותצלחותולא סדרו ¥
לא י¥ן תפלה לא ל ¥ים תהלו¥¥ וזחירו' . הקולו¥ יסדלו¥ן
ו¥ין ב¥יו ת¥כ¥ו¥ ע¥שרת הד¥רו¥ .¥ ¥ ¥ובאו ¥ליו ¥¥
הין¥¥¥¥¥ו¥
בה¥¥¥¥ל א¥ו¥ת¥. ¥¥¥ינ¥¥¥¥רי¥יהו¥עלל¥¥¥¥י
הב¥¥רות.. ¥¥¥ה¥יקר¥
בשם יע ¥ו¥.¥¥¥¥¥חיק"¥ו¥¥¥ ¥כ¥ל ירו¥ה
לו ח ¥¥ש¥¥ה דורות .¥העושר וה¥בוד ל¥כיו שתי הכו¥ו¥

Afb. 13: Resultaat van OCR

9. Bijlage 2: Nieuwe ontwikkelingen

Fuzzy logic search met ZyINDEX van ZyLAB

Zoeken in een OCR-tekst met gebruikmaking van de zogeheten ‘fuzzy logic’ technologie vergroot het aantal hits aanmerkelijk. ZyLAB heeft een demo-versie van haar programma ‘ZyINDEX’ waarmee met deze techniek gezocht kan worden. Echter, omdat het Hebreeuws gebruikt maakt van de ASCII-codes 224-250 of letters met accenten, kon deze demo-versie niet toegepast worden omdat hiervan de 'mapping' zodanig is ingesteld dat bijvoorbeeld alle a's, ongeacht het accent, als één en hetzelfde teken worden geschouwd. De versie van ZyINDEX die te koop is kan men echter naar wens instellen. Helaas heeft het programma wel een NT4.0 server nodig, omdat niet verder voor UNIX wordt ontwikkeld. De prijs voor de ZyINDEX software bedraagt ƒ 3.000.

Mogelijkheden van ZyIMAGE en Smart Image

ZyIMAGE
Met ZyIMAGE kan gezocht worden in een OCR-tekst, waarna de hits in het oorspronkelijke plaatje worden weergegeven. Helaas kunnen de resultaten van proLector niet op deze manier doorzocht worden: proLector geeft niet voldoende informatie door om de hits in het plaatje op te zoeken. Het kan wel met Recognita, Omnipage en de ZyLABocrmachine (Scanocx), maar die zijn alle drie ongeschikt voor Hebreeuws en ook niet of niet voldoende te trainen voor andere, oude, 'drukkunsten'. Met de firma van Ligature was moeilijk contact te krijgen.

Er zijn wel ontwikkelingen op dit terrein: ZyLAB en Improx, de ontwikkelaar van proLector, werken nauw samen. Improx heeft inmiddels al een nieuw OCR systeem ontwikkeld waarmee ZyIMAGE tezijnertijd wel kan samenwerken. Dit product is nog niet verkrijgbaar.

ZyIMAGE kost ƒ 17.500. Het werkt al wel voor o.a. het Russisch.
ScanOCX kost ƒ 4.500

Smart Image Technology
De voor 1997 aangekondigde ‘Smart Image Technology’ heeft vertraging opgelopen. Momenteel zijn er tests voor 1997 geplanned in de USA. In Europa zullen de eerste bètaversies niet vóór 1998 verschijnen. Smart Image Technology maakt gebruik van fuzzy-logic retrieval, waarbij, evenals bij ZyIMAGE, de resultaten in het plaatje, en niet de OCR-tekst getoond worden.

Voor het Menasseh Ben Israel project vallen deze ontwikkelingen buiten de projecttermijn. In ieder geval bieden de mogelijkheden voor retrieval van OCR-teksten in de toekomst interessante vooruitzichten.

Informatie over ZyLAB produkten (4 sept.'97) van de heer Scholtens, hoofd produktontwikkeling van ZyLAB, Amsterdam, tel. 6919550.