

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 trainingsset 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.

חקרמה

בלאחיים * העם ההולבים באשך כל הימים לא ראו אור
 וימי עולמים ושנים קדומים * ועתה הנה הגבורים
 גדולים ופגומים * ואלוהים ומוהירים כגדומים שמימים *
 ולכל בני ישראל היה אור במושבבתם לאור שבעת הימים *
 אשך השמש ומזרחיים * הן גופם אברו כלמו אברו כאל
 קישון גרשם כאל קדומים * והנה אכמו מאלמי אלומי *
 ונכונשים רשים וקסומים נשומים והיו תלומים * והיה
 הטוב הוא אשר השיג ה' לנו חמק עבד הכל כשכח כלו
 לא היה היינו כפולמים *

למספר בני ישראל

אשר העלה מירושלם
 אכירים בכונות את
 כל התלזה אשר האת' בארבות איביה' בארץ יוגוהים *
 את כל הקורות וולישת ותורדות * גדושים וטודות
 גורות גורות המרחות המרחות לא אחיה ולא אשכר
 בשעת בדידות * כי לא ידעת שפולות * כי סנכדו
 רעות במגרות * בכל מוסבותיהם מדינות פדינות
 וערים במרות * בארץ אדום כליו רשים כלו חום
 מכרות * ובארץ ישועאל כרא למוד מדרות * עד
 אשר נשמי יוהית תענגוהם כדנו הלכמן הסי' * וכן
 החרות נחאום רעות רבות נכות ומירועות רעות
 גדולות ובמרות * במראה אש בופרות * בלכתם
 בשיות שונות * כפונות של הכרים בכקעות וביערות
 ומעלות ובמרות * וכפונות בלונות שפודות בימי' וכל
 תחומות על נהרות * כי נענה בס יד ה' אש ונפלי' ורח
 כפרות * ידכה ישום בחורים ומקיים ידום טף ומשים
 ימולל אילות ויחשוק יערות * וישא העם את מנקה
 בשלמותם נהרות * בשלהבות נהלות קשורות מכות
 טריות פלע ומבמות * ומעים רעים קשים ומאומים *
 שקערותיהם חרות ומכות כבדות * זה ואמר לה' אמי
 ישיר בבקר כבקר קורא לאלוהי ישמי' קול תהלתו ימולל
 גבורות כיהו ולא אש לבני השירות * אחרות ה'
 אחרות מהרות קלמו עיניו אשתדלו' ימסכו ככבי' ששקו
 אל תועב עליו כהרות והיו למחורות נלמות ולא סדרו' *
 לא יתן תפלה לאלקים יתהלל ומימלו' * הקלות יתלונן
 ואין כפלו תוכחות ישלות הדברות * ופאו שלמו כל
 הקלות כהר עלל אחרות * וקדמות ודעיהו על דברי
 הגבורות * וזה יקרא בשם יעקב מיתחמתא דבלי ירום
 לו יתבשרה דוחות היעשר והכבוד לכח שתי הכונות
 טר

הטובים שימרים והתנוחים כי אם צהנלי הגני' ונשמי'
 מביחו כותים ועדים וזומים פתאים ועמוכים יקראו
 להם זחומים והיה רעים וזמאים טועים ותשש אצל
 אשמים * זכר ה' מזכרו עלות תני' המקה והאש
 נוכח הכלמים * ואשו הנחלה נומרת המורד וקטורת
 הסמים * מדרך און ופתי אדם המקן וטעבכר ועוד
 וכמות ובעלי זומים * מוכיר לכותה עם גמלי רתמים *
 נאר וקדשו טכר היה לעולמים יפה טוף מטוב כל
 הארץ הללוהו כל היום שנעוהו כל האומים * יבקר
 לקיים קדשים קותם בקדשים * פת יקוש מטתה מרמים *
 ישות שלום להם יקחו יהיו תמים * הכגור ביד אויב
 קדיה כלכותה יטב כדד יטב משמים * החר קמוד אלקים
 לשבתו על כל הרי בשמים * ויבן כונו רמים * קדימא
 מרדיתא רנתי בניים שרתו בעמים * מוראה ונאלה
 הפיר תגללה בלומים * ויעלה באל' יוסף נדיק ינונים *
 וכשכע יקדח אשר בחר האלשים האלה שלמים * היו
 כנו שומים * ויבקר בזיע נרשים בניים נשיתים
 שורש רתמים * ואלש פתכים אכור אשר מלא את הימי'
 יחמום כן סכו סוכת שלמים * סכת דוד הכוללת
 אשר היה שם אלה בתחלה בימי עולמים * שם אנונים
 קרן דם ופאו טוף ואנון מקרני' ראונים לבצל זלשורה
 זלעובדי ססילים כתי הסכל כהומים * שמת מועדו
 ישראל עם קרובו המועדים על ה' נתן טפלה כשקמים
 ובעתהו כומרי יום יהלך עליו אחים * וככרים נאו
 שפרי' ולומים וכשים כומעני' * ויתן אותם כרקיב
 השמים כשאים ורמים * מלכי ארץ וכל לאומים * שכל
 ה' כנין מועד ושבת ימי הנומים * חבלים ככלו לי
 כבעמים ולא זכר הדום רגליו את החסור ואת הרמתי' *
 זכר עשה לנתהלים כאלילים גלוי לכל העמים * תום
 ינקשו מדיהם ומיעדיהם אשר כדו מלכס ממסכות
 פדומים * כוכרים ונשים והיו לאומות ולמועדים
 יחיים * ויאלץ כועם אבו מלך וכהן את האוכים ואת
 העומים * אור תעלה חתום תורה בנימוה והדברים
 עתיקים סתומים וחתומים * ספו יתנו הכנאים
 ככמים וידעים ונשי' על מדין והיה כמיו' מחוכמי'
 וידם תלשים רעים רכים קמים אשכים ומדעומים *
 מעוקן ווכוז ומכסף וקוסם קסומים * ויבחר לשון
 פדומים * ומה אוכיף טרד לזכר ערבים כמים תוכחות

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

בלאיווי י העם התולכים בקשכל הימים לא ראו אור
 חיחי עולמים ושנים קדוחים. ועתה היה הגבורים
 גדולים ועצויים. וואורים ויוהירים כנרחיס שמחים.
 ולכל בני ישראל היה אורבחושבותם כאורשבתיחי.
 ששך השחש יומרוויים. הוגועני אבדנו כלכו אבדנו נמל
 קישון גרפסנסל אדוחים. והנהאנמנו חאלחי אלויי.
 יופגועים רעיסואסויים יושונים והיו תאוויים. יהיה
 הטוב ההוא אשרהטיב הי לנו סחק עבר הכל נשאלו כליו
 לא היה היינו כסולויים.
 אשרהגולה מירושלם
 אסירים בכושרותא
 כל התלאהאשר יוצאת בארצותאויביה, בא. צמגוריהם.
 את כל הקורות קליפות ותחורות. גרושים ושיות
 וגורותגורית היירות החאררות לא אמלהולאספר
 בשעותבארות. כי לא דיגתי ספורות. כי סבבוהו
 רעות בחגורות. בכל מושבותיהם מדינותעדינות
 וערים בצורות. כארץאדום כליי רעים כלי סיוס
 חכרות. ובארץ ישועאל פרא לחוד מדברות. ע
 אשר גרשו יוביתתענוניהם נדדו הלכו חן הטירו י וח
 הקצרותחוצאום רעות רבותוצרות וחארעות רעות
 גדוליתובצורות. כיוראהאש בוער. בלכתם
 כשיותפורות. כפוצותעל ההרים בבקעות וביעיות
 זבדותובנאות. ובשוטם באוניות שבורותביחי וכל
 תיהחות על נהרות. ככי נעיהבם די. אש וגפריורוס
 סערות. ידכה ישוק בסזרים ו זקנים יריוום טף ונשים
 יחולל אילואיסשוף יירית. וישא העם אתמצוקם
 בשחלותם נרורית. בשלהבות בג ילותקשורותיוכות
 טריותפעצומבירות. ינגעים רעים קשים ונאחנים.
 שקערורותבהרותוחכת בכרות. וה ואור ליה א
 יעירבבקר
 בב יר אורא. אלוה ישמי. אוןול אלתו יחלל
 גבורותפיהו חלא אדברי אשירות. אחראחה
 אחריתאורות קדליו עיניו אשחורוי ימשאו ככבי נשפ
 אל תופעלעליו נהרותוהיו לחאו אנתצלחותולא סדרו
 לא יאן תפלה לא ל צים תהלוא וחירי. הקולוא יסדלואן
 ואין בליו תכארא עשרת הדארו. א אובאו אליו
 הין
 בהאכל אנת. אוןינהאוריהוהעלל
 הברות. אהיקר
 בשם יע אן. אהחיק אוןאכל ירוה
 לו ח אשאה דורות. העושר והאבוד לאכיו שתי הכוא

חפובים הישוהיוויים אס בהבליהגויי יוגשיחיו
 יופיקי כובים ועדים וויוחים פתאים ועחונים יראו
 להם וחווויים והיוה רעים יסטאים טועים חתאיםאבל
 אשויים. זק י חובקו עולת תיוד היונסה והאשם
 יובס השלחים י ואשי הגדולה יונרת היואור וקטורת
 הסחים. חברך און זובקי אדם השקוהעכבר יועא
 וכתותיבלי יויים. חוכירלבוה אם גקלי רתחים.
 נאר ייקדשו שכבר היה לעולחים יפה נוף חשוב כל
 הארץ הללוהי כלהגיים שבסיהו כל האויים. יבסר
 אקים סדשים סיתם בקדשים. פס יקיש יושטס סריוים.
 ישפותשלים להם יסדו יהיו תיחים. הסגיר ביד אייב
 איה נאחנה ישב ברד ישב משחים. ההר סחדאלקים
 לשבתו על כל הרי כשחים. ויבכמו רחים. קריתא
 יורדתא רבתי בנוים שרתי בעויים. חוראה ונגאלה
 יעיר חגולה בדיוים. וילאס באהל יוסףצדיק תיוים.
 ובשבת יהודהאשר במר האנשים האלהשלוים. היו
 בניו שומחים. ויבסר בארע יורעים בנים חששיתים
 שורשרתחים. וראש פתנים אכוראשר חנא את היחי,
 אסחום כגן סבו סוכתשלוחים. סכת דוא תלפלת
 אשר היה שם אהלה בתאמלה ביחי אולחים. שם אניים
 אירן רם ונשא סזק יאיויחאריני ראויים לבעל זלאשרה
 לעובדי פסילים נתן הסכל כמרומים. ששתיוועדו
 שראלעם קרובו הכועדיעל ה' נתן בשפלה כשקויים
 בענתוהו כחרירי יום יהלונעליו אחי. ונכרים באו
 שאיו ולובים וכושים בחצעדיו. ייתן אותם ברקיע
 השיים כשיאים ורחים. יולכי ארזוכללאוחים. שכס
 הי בציאן חוער ושבת ויחי הנוחים. קכלים נפלי לי
 בעניחים. ולא וכו הרום רגליו אהסטד ואת הרסיווי.
 אכר עששה ליותיהללים באלילים גלוי לכל העויים. סגים
 ינקופו קדשיהם וחאעדיהם אשר בדי יילבם ממשבות
 ארוחים. נוכרים ונעאשים והיו לאותות וליוועדים
 ויחים. וינאקעום אלו יולך וכהן את האורים ואת
 התוויים. נור תאעודהקתום תורה כאיוודו יהדברים
 עתיקים סתוויים יסתוויים. ספי תיו הנביאים
 בנוים וירועים יושבי על חדין וחחה סכיוי יומוכיווי
 וירם תולעים רעים רבים קיייםשפים וקרטוים.
 אועונן
 וינאוש וחכשו. וואוסם אקחים. ויבאאלשון
 ארוחים. וחה אסי. עודאבר ערבים כגיים תוכסות

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 f 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 tezamen wel kan samenwerken. Dit product is nog niet verkrijgbaar.

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