Transcription enhancement of a digitised multi-lingual pamphlet collection: a case study and guide for similar projects

by Andrew M. Watson, Elzbieta S. Szubarczyk, Peter S. Salinger, Amy J. Howe, Steven Wright, Vanessa R. Freedman

UCL Library Services

March 2021

Transcription enhancement of a digitised multi-lingual pamphlet collection: a case study and guide for similar projects

1 Executive summary

UCL Library Services holds an extensive collection of over 9,000 Jewish pamphlets, many of these extremely rare. Over the past five years, UCL has embarked on a project to widen access to this collection through an extensive programme of cataloguing, conservation and digitisation. With the cataloguing complete and the most fragile items conserved, the focus is now on making these texts available to global audiences via UCL Digital Collections website. The pamphlets were ranked for rarity, significance and fragility and the highest-scoring selected for digitisation. Unique identifiers allocated at the point of cataloguing were used to track individual pamphlets through the stages of the project. This guide details the text-enhancement methods used, highlighting particular issues relating to Hebrew scripts and early-printed texts.

Initial attempts to enable images of these pamphlets to be searched digitally relied on the Optical Character Recognition (OCR) embedded within the software used to create the PDF files. Whilst satisfactory for texts chiefly in Roman script, it provided no reliable means to search the extensive corpus of texts in Hebrew. Generous advice offered by the National Library of Israel led to our adoption of ABBYY FineReader software as a means of enhancing the transcriptions embedded within the PDF files.

Following image capture, JPEG files were used to create multi-page PDF files of each pamphlet. Pre-processing in ABBYY FineReader consisted of: setting the language and colour mode; detecting page orientation; selecting and refining areas of the text to be read; reading the text to produce a transcription. The resultant files were stored in folders according to language of text.

The software highlighted spelling errors and doubtful readings. A verification tool allowed transcribers to correct these as required. However, some erroneous or doubtful readings were nevertheless genuine words and not highlighted; it was therefore essential to proofread the text, particularly for early-printed scripts. Transcribers maintained logs of common errors; additionally, problems with Hebrew vocalisations, cursive and Gothic scripts were noted. During initial quality checks of the transcriptions, many text searches were unsuccessful due to previously unidentified spacings occurring within words. This was generally linked to the font size being too small. Maintaining logs of font sizes used led to the adoption of a minimum of Arial 8 or Times New Roman 10 in transcribed text. The methodology was revised to include the preliminary quality-checking of one page. We concluded that it was difficult to develop a standardised procedure applicable to all texts given the variance in language, script and typography. However, we concluded that the font Arial gave the most successful accuracy ratings for Hebrew script, minimum text size 17, minimum title size 25.

ABBYY file preparation took a minimum of 1.5 hours per pamphlet; transcription correction took an average of 10.4 minutes per page; the final quality check took 30 minutes per pamphlet. On average, the work on each pamphlet took a minimum of 6 hours to complete.

As a result of the project, average accuracy ratings improved from 60% to 89%, the greatest improvement being for pre-1800 and Hebrew script publications. We are therefore inclined to focus future transcription-enhancement activity on these types of publication for the remainder of our Jewish Pamphlet Collections.

2 Description of the project

UCL Library Services holds printed, manuscript and archival collections of Hebraica and Judaica which are of national and international importance, including several significant pamphlet collections. These cover a wide range of subjects throughout the field of Jewish Studies, particularly Anglo-Jewish history, Zionism and liturgy. The pamphlets date from 1601 onwards, and are in English, Hebrew and a variety of other languages. Many of them are held in very few libraries, while some are extremely rare.

In 2014, UCL Library Services embarked on a multi-phase project to catalogue and conserve the pamphlets.[1] Hand-in-hand with these activities, a selection of the most significant pamphlets in the collection was digitised and made available for viewing online via UCL Digital Collections repository.[2] These digital copies are intended to be searchable in both Hebrew and Roman characters.

In Phase 2 of the project, a workflow for Optical Character Recognition (OCR) using the software ABBYY FineReader[3] was piloted in consultation with the National Library of Israel.[4] This produced searchable transcriptions to an acceptable standard. However, the software was unable to correctly identify some of the characters within the texts and the success rate varied considerably depending on the language of the texts and fonts employed. Thus, central to Phase 3 of the project has been a programme to enhance the transcriptions produced by the OCR software in Phase 2.

The transcription enhancement was carried out by four transcribers and a team of trained volunteers using the ABBYY software. They worked through the transcriptions page by page and in some cases, line by line, correcting any errors identified. Whilst doing so, they maintained records of their activity including logs of common problems and errors which could be used to predict recurrence and allow enhancement to be carried out more efficiently. These records form the basis of this transcription enhancement guide.

The enhanced transcriptions have been embedded within the UCL Digital Collections repository to offer academics and non-specialist audiences alike improved access to the texts and to enable this material to be used for text- and data-mining. The overall methodology is based on best-practice guidance developed by the UCL Institute of Education Archives during extensive digitisation projects, such as Digitising The Woman Teacher.[5]

3 Other similar projects

Hitherto, UCL Library Services' chief experience of Hebrew text transcription has been in the context of the *Montefiore Testimonials Digitisation Project* in 2009-2015. The 350 testimonials, part of a loan from the Montefiore Endowment (London),[6] consist mainly of manuscript tributes to Sir Moses Montefiore. They were transcribed by student volunteers from the UCL Department of Hebrew & Jewish Studies under the guidance of Dr. François Guesnet. No specific transcription software was used; rather, the texts were simply typed and the transcriptions presented as separate documents alongside the digitised images. Although these enable comprehensive searching within the testimonial texts, the search results are simply a list of the documents within which the search terms appear, and the terms are not highlighted in the digitised text or transcription. The collection can be viewed via UCL Digital Collections repository.[7]

Elsewhere, instances of projects to enhance embedded transcriptions of Hebrew script are few; notable collectors of similar material in the United Kingdom appear not yet to have embarked on such ventures. The National Library of Israel has made significant steps to enable its digitised Hebrew texts to be

searchable, both through the use of external vendors for periodicals and ABBYY FineReader for their Jewish Historical Press projects. However, in the latter case, few manual corrections have been made; instead, word-recognition machine learning is being developed. The Library's generously offered advice guided us when planning this current project and led us to choose ABBYY FineReader as our OCR software.

4 Selection of pamphlets

4.1 Source of digital images

In Phase 2 of the project, 172 pamphlets were digitised creating 6,100 images. The following criteria were considered when initially selecting the pamphlets for digitisation:

- Rarity
- Significance
- Fragility
- Availability in digital format

The purpose was to ensure that texts which were not widely available, or which could not be produced for consultation because of poor condition, would nevertheless be available online. The images were uploaded to the UCL Digital Collections as PDFs with contextualising information and metadata.

In Phase 3 of the project, 81 pamphlet transcriptions have been enhanced so far for texts in English, Hebrew, German, Greek and Italian.

4.2 Unique identifiers

It was clear at the outset of UCL's Jewish Pamphlets project that unique identifiers would be required in order to track individual pamphlets through the various processes and stages of the project: conservation assessment, cataloguing, conservation, digitisation, exhibition and display. The primary identifier chosen was the automated system number generated by the Library Management System (LMS) as each pamphlet was catalogued. This was generally a seven-digit number. However, as some assessment processes were conducted before cataloguing occurred, an alternative was also required.

Alternative identifiers were based upon the container references, in this instance, the alpha-numeric barcode affixed to each pamphlet box, for example: UCL0133510. A decimal appendage was added to this for each pamphlet according to the order within the box. For example, the third pamphlet in this particular box would be given the identifier UCL0133510.03. The alternative identifier was used in public references, for example, in exhibition captions. However, for the purposes of transcription as described in this report, only the LMS system number reference was used.

5 Methodology

5.1 Creation of digital files for the ABBYY FineReader process

The selected pamphlets were photographed on a copy stand with a Canon digital single-lens reflex camera (DSLR). Each digital page was straightened, cropped and then saved as TIFF and JPEG files. The TIFFs were archived and the JPEGs were used to create multi-page PDF files for each pamphlet. All files were named according to the unique identifier associated with each pamphlet (see Section 4.2).

5.2 Preparation of the ABBYY files for transcription

Upon opening ABBYY FineReader, there were a number of pre-processing actions to select/deselect in the Tasks section:

5.2.1 Language

This was set to the common language for each pamphlet. For pamphlets printed in multiple languages, the 'more languages' option in the dropdown menu was used and the other languages selected.

5.2.2 Colour mode setting

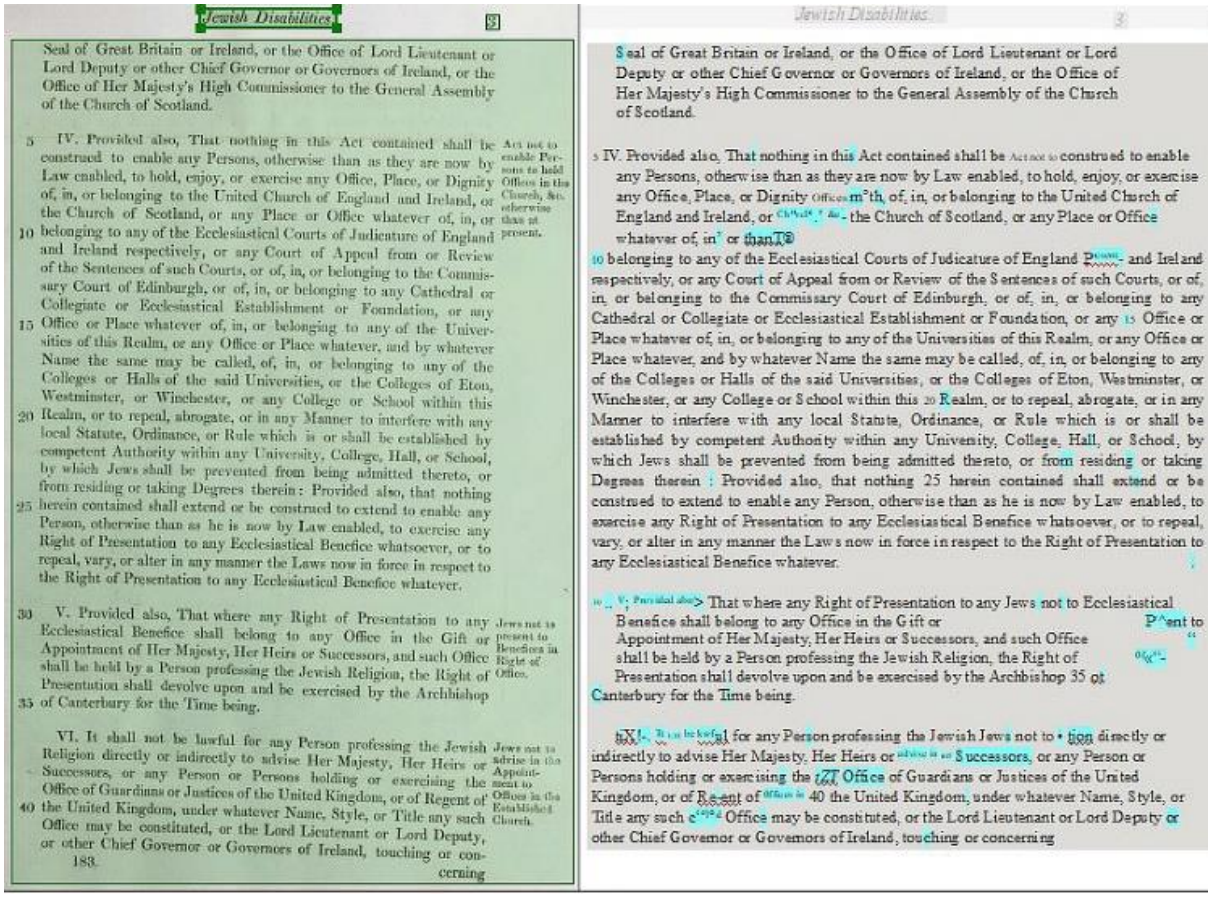
Although changing the mode to read the document in black and white sped up the processing time considerably, the visual quality of the document was inferior. Thus for all pamphlets, it was decided to keep the documents set at full-colour.

5.2.3 Detect page orientation

As pages were correctly rotated during the PDF creation process, we did not need to select the 'detect page orientation' option. We also left the 'enable image pre-processing' option unchecked. We discovered that this was not suitable for pamphlets when the text was not straight. The tool would straighten the text, but often leave the page at an obvious angle.

5.2.4 The processed document

During the reading or *recognition* process, the software analysed each page and put any readable text in green boxes. We found that ABBYY would often merge all text into the one box, regardless of layout, which would cause problems with the document format. In a standard text page with paragraphs, this would usually be acceptable. However, when dealing with pages with text in table format, or annotations, the software was often unable to distinguish between different groups of text and would


Jewish Disabilities 3

Seal of Great Britain or Ireland, or the Office of Lord Lieutenant or Lord Deputy or other Chief Governor or Governors of Ireland, or the Office of Her Majesty's High Commissioner to the General Assembly of the Church of Scotland.

IV. Provided also, That nothing in this Act contained shall be construed to enable any Persons, otherwise than as they are now by Law enabled, to hold, enjoy, or exercise any Office, Place, or Dignity of, in, or belonging to the United Church of England and Ireland, or the Church of Scotland, or any Place or Office whatever of, in, or belonging to any of the Ecclesiastical Courts of Judicature of England and Ireland respectively, or any Court of Appeal from or Review of the Sentences of such Courts, or of, in, or belonging to the Commissary Court of Edinburgh, or of, in, or belonging to any Cathedral or Collegiate or Ecclesiastical Establishment or Foundation, or any Office or Place whatever of, in, or belonging to any of the Universities of this Realm, or any Office or Place whatever, and by whatever Name the same may be called, of, in, or belonging to any of the Colleges or Halls of the said Universities, or the Colleges of Eton, Westminster, or Winchester, or any College or School within this Realm, or to repeal, abrogate, or in any Manner to interfere with any local Statute, Ordinance, or Rule which is or shall be established by competent Authority within any University, College, Hall, or School, by which Jews shall be prevented from being admitted thereto, or from residing or taking Degrees therein: Provided also, that nothing herein contained shall extend or be construed to extend to enable any Person, otherwise than as he is now by Law enabled, to exercise any Right of Presentation to any Ecclesiastical Benefice whatsoever, or to repeal, vary, or alter in any manner the Laws now in force in respect to the Right of Presentation to any Ecclesiastical Benefice whatever.

V. Provided also, That where any Right of Presentation to any Ecclesiastical Benefice shall belong to any Office in the Gift or Appointment of Her Majesty, Her Heirs or Successors, and such Office shall be held by a Person professing the Jewish Religion, the Right of Presentation shall devolve upon and be exercised by the Archbishop of Canterbury for the Time being.

VI. It shall not be lawful for any Person professing the Jewish Religion directly or indirectly to advise Her Majesty, Her Heirs or Successors, or any Person or Persons holding or exercising the Office of Guardians or Justices of the United Kingdom, or of Regent of the United Kingdom, under whatever Name, Style, or Title any such Office may be constituted, or the Lord Lieutenant or Lord Deputy, or other Chief Governor or Governors of Ireland, touching or concerning

Figure 1 – Initial analysis showing text merged into one box

read it all as the one text block. In the case shown in Figure 1, ABBYY's initial analysis has grouped all text together, meaning that the paragraphs have merged with the annotations. As can be seen, in the first sentence of the second paragraph, the uncorrected text reads:

'Provided also, That nothing in this Act contained shall be Act not to construed to enable any persons…'.

Once manually corrected by redrawing the boxes around each body of text as shown in Figure 2, the adjacent annotation is treated as a separate entity and the first sentence of the second paragraph correctly reads:

'Provided also, That nothing in this Act contained shall be construed to enable any persons…'

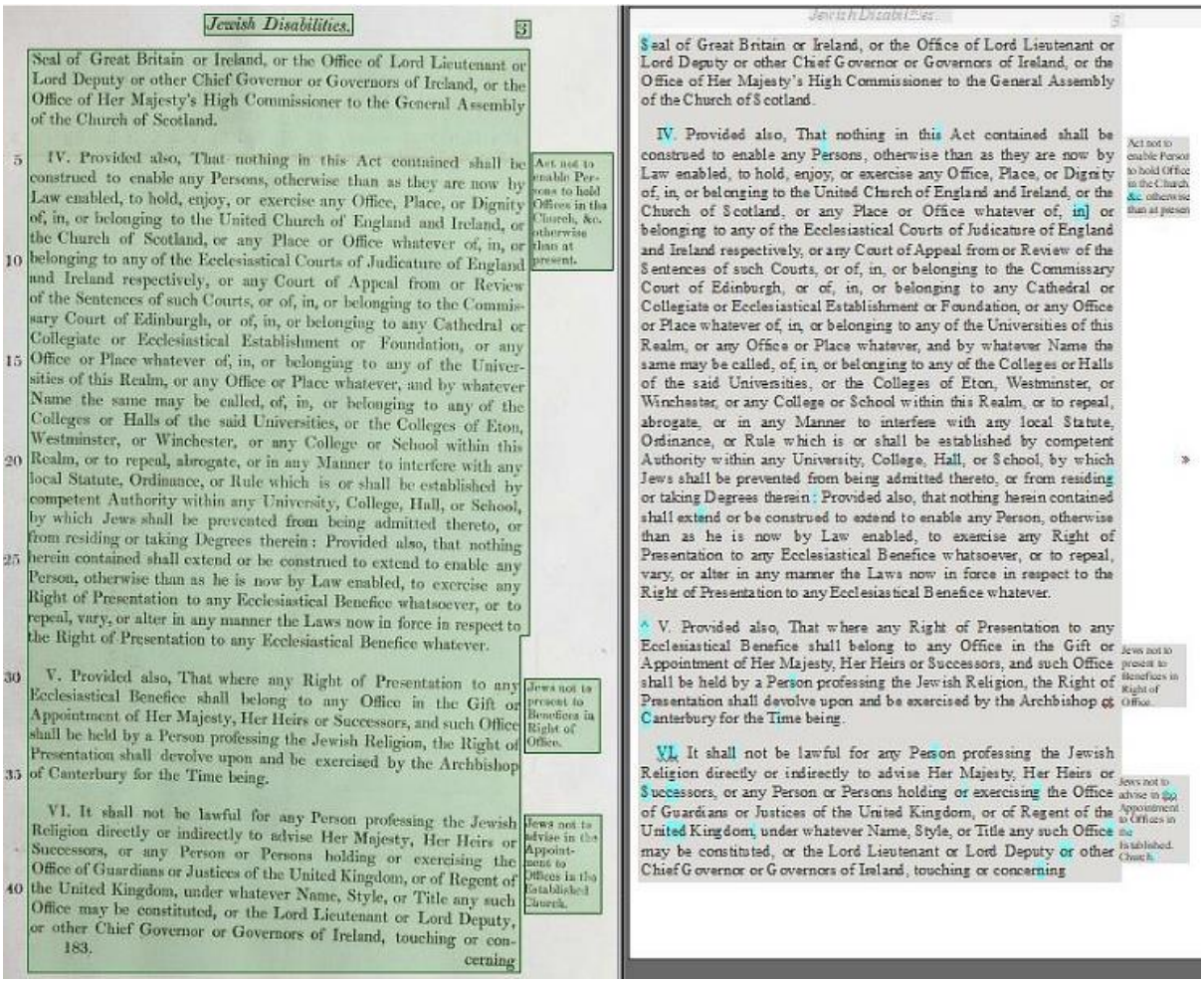


Figure 2 – The same passage showing the text edited into the correct distribution of text boxes

Another common issue encountered was the generation of text boxes on blank pages (see Figure 3), where the text from the previous page was visible though the paper. The solution was to manually tidy up each page by removing the excess boxes and redrawing them around paragraphs or main bodies of text.

The preparation generally took a minimum of 1.5 hours per average-sized pamphlet, but in reality, many took much longer due to the following factors:

- The text was not perfectly straight
- There were columns, text in the margins, illustrated letters, decorative fonts
- The paper was marked and/or dark resulting in poor definition between the text and paper
- There was bleed-through from print on the reverse.

Pamphlets exhibiting multiple instances of these could take a whole day to prepare.

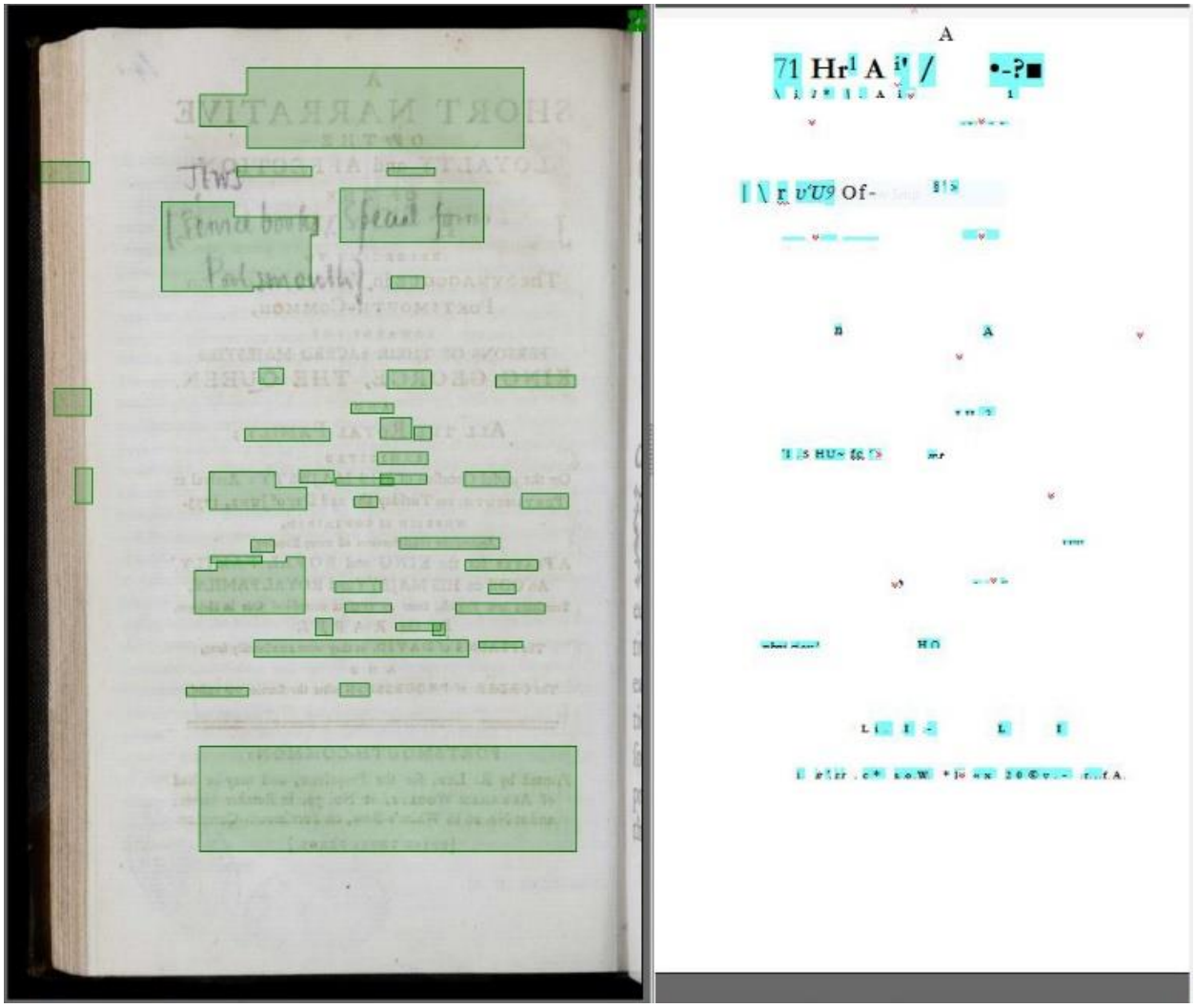


Figure 3 – Example of text boxes on blank pages

5.3 File storage and language of texts

Once the ABBYY files had been created and prepared ready for correction, they were grouped in three subfolders on the library's local shared drive according to the language skills required to correct the transcriptions. This enabled transcribers to select files to work on which matched their skills without needing to open each individual file to assess its contents.

The chief language groupings of the texts and the distribution encountered in the set chosen for this project were:

Group	Language	Percentage
1	English and other Roman script languages	57%
2	Hebrew	42%
3	Yiddish	1%

Any text containing Hebrew was allocated to group 2; of these, the majority also contained other languages, predominantly English, but also Greek, Italian and Latin. Unfortunately, the Yiddish transcriptions were not able to be completed during the project and thus are not discussed further in this guide.

Once an ABBYY file had been selected by a transcriber, it was transferred from the relevant subfolder into the transcriber's own working folder. This ensured that transcribers did not accidentally select a file which was already in the process of being corrected.

5.4 Text correction and common errors

The transcribers used ABBYY FineReader version 11 to check the transcriptions and correct any errors. The interface presents four panes as shown in Figure 4.

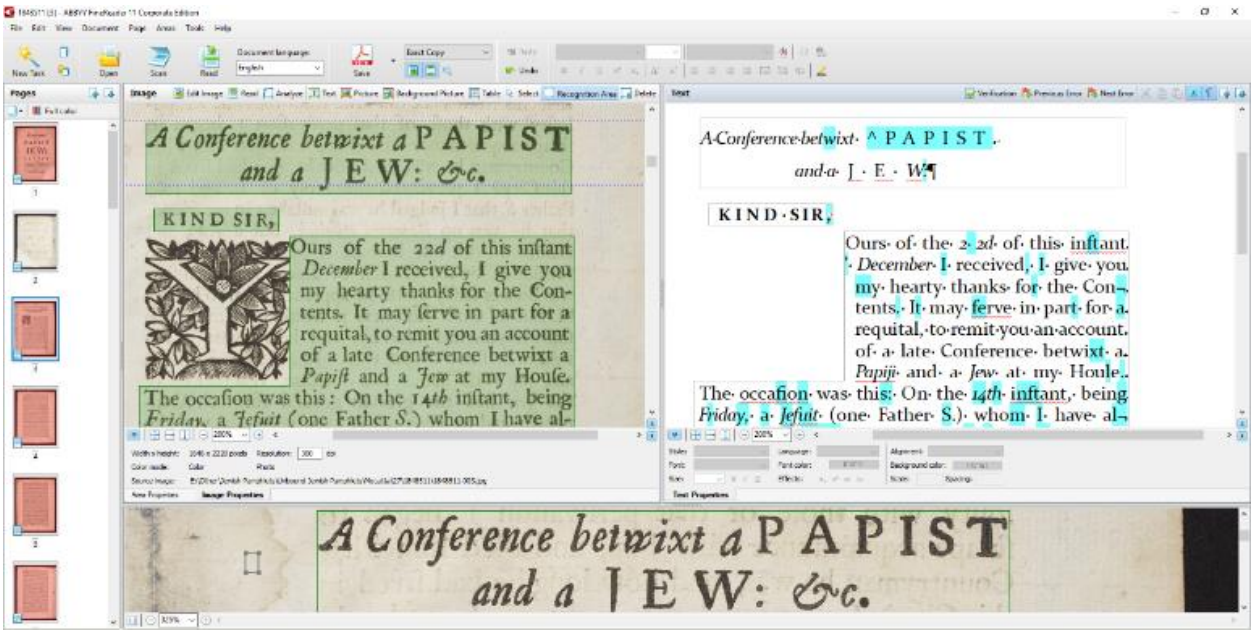


Figure 4 – ABBYY FineReader user interface

The left-hand pane displays thumbnail images of each page included in the file. In this example, page 3 has been selected. The top left-hand pane displays the appropriate image. The highlighted frames indicate which sections of the image the user has select to read. The green colouring in this case indicates that the frame contains text; red colouring would indicate a non-text area, for example, an illustration.

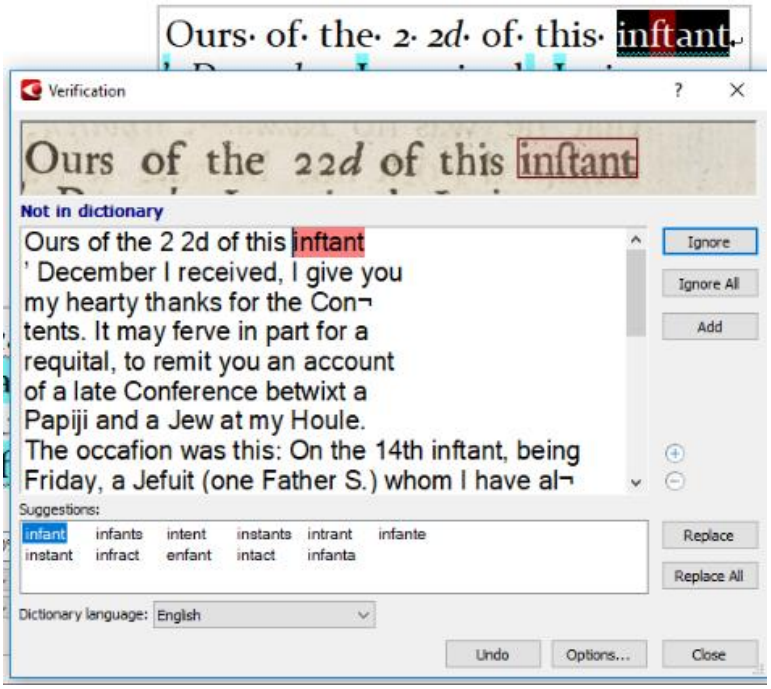
The top right-hand pane displays the transcription which the software has produced from the image. The text in this pane can be corrected as required. The blue and red highlighting will be described fully in the following section.

Lastly, a zoom tool produces a magnified view of a line of text in the lower pane. This was found to be invaluable for deciphering any characters that were difficult to read.

5.4.1 Suggested errors highlighted by software

ABBYY FineReader employs two methods of indicating potential errors: 1) instances where the reading is uncertain are highlighted in blue; 2) instances where words are not recognised as being in the software's integral dictionary of that language are underlined in red.

The verification tool offered in the toolbar options enables the correction of these blue- and red-flagged sections throughout the document. ABBYY FineReader also offers a spell-checker facility, providing a list of alternatives for non-dictionary words at the foot of the window (see Figure 5).



The correct word can be selected. Clicking "Replace" replaces the single instance of the word; alternatively, clicking "Replace All" replaces all instances of the word with the corrected form. However, for this to function successfully, the whole document must be checked using the Verification window, otherwise the instruction to automatically correct these words ceases once the window is closed.

As with other word-processing applications, there is an option to add a word to the dictionary, in which case further instances of that word will no longer be flagged as errors.

Figure 5 – Verification tool with suggested alternative words

5.4.2 Errors identified by project staff

It is important to stress that not all errors were identified by the software. This was particularly the case where the letters in some typefaces were misread for similar letters which, nevertheless created words recognised as being valid although in the context in question, they were incorrect.

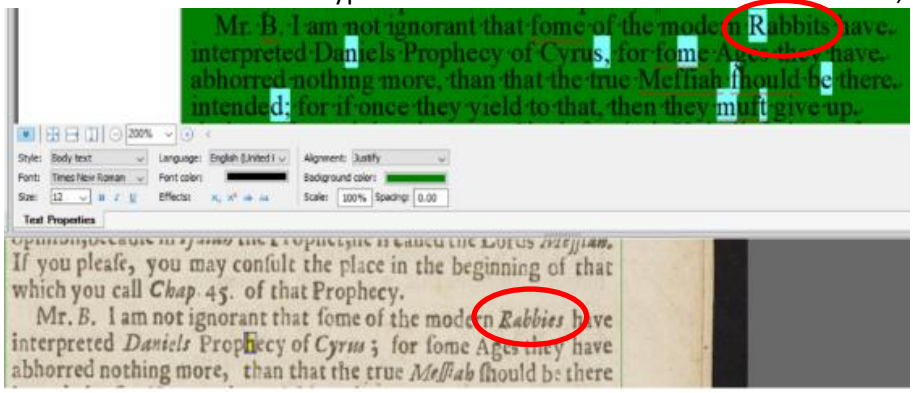


Figure 6 shows a typical example where "Rabbies" has been misread as "Rabbits" but, being a valid word, has not been recognised as error. It can

Figure 6 – Example of an incorrect reading not being recognised as an error

also be seen that the long “s” has been misread as an “f”. Whereas “fome” has been recognised as an error in this instance, “fame”, the reading for “same” would not be. This can be seen in Figure 7 along with “fort”, the reading for “sort”.

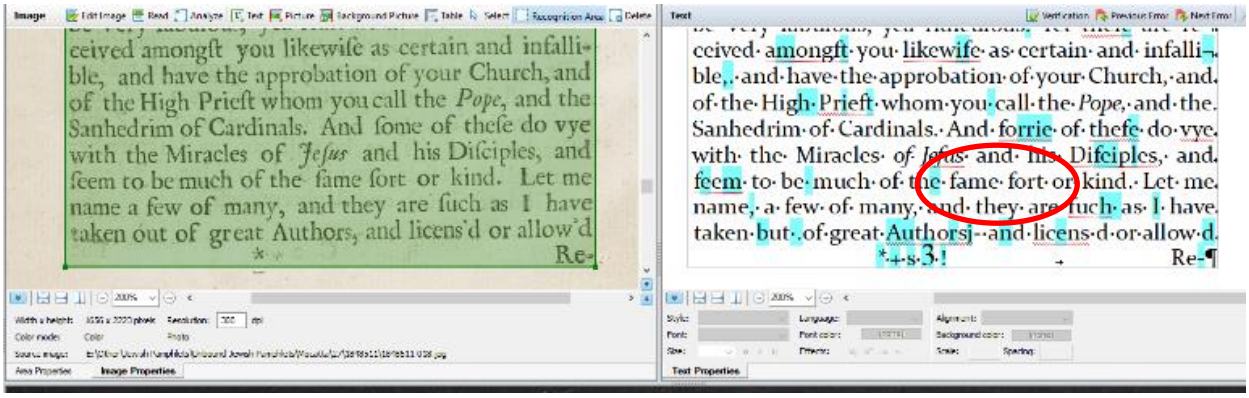


Figure 7 – Example of long “s” misread as “f” and not recognised as an error

However, in some cases, the software itself introduced inaccuracies, such as replacing spaces by extraneous characters as shown in the title in Figure 8.

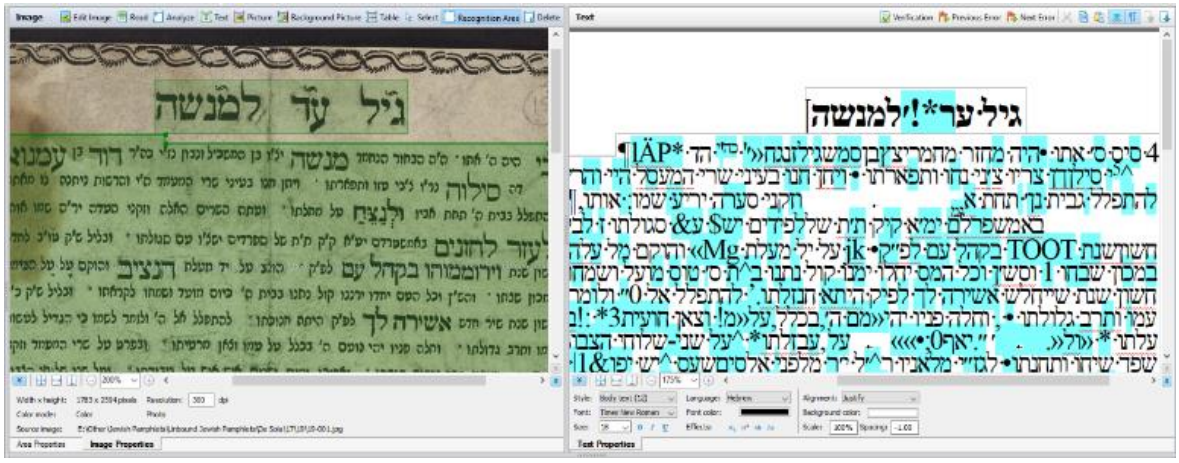


*Figure 8 – Example of error in Hebrew text being introduced by the **software***

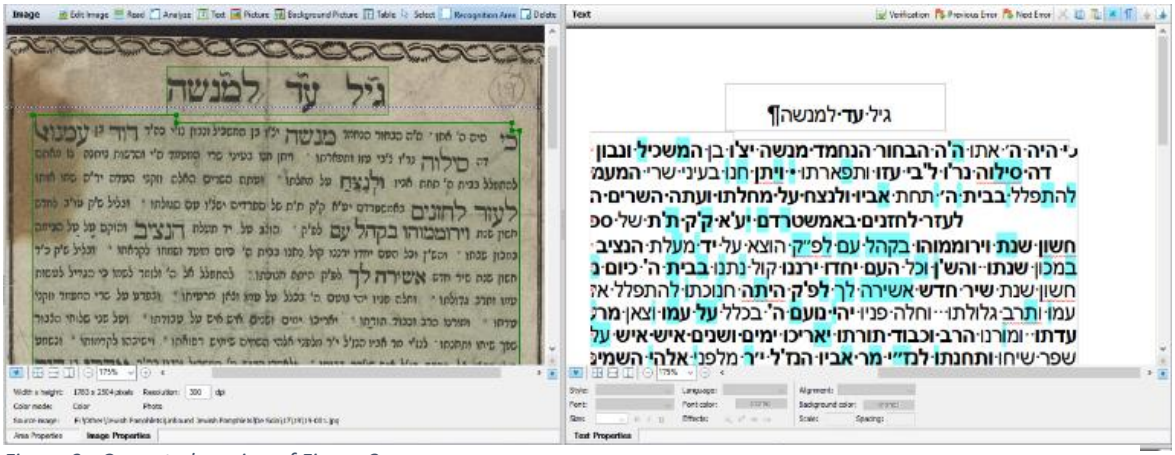


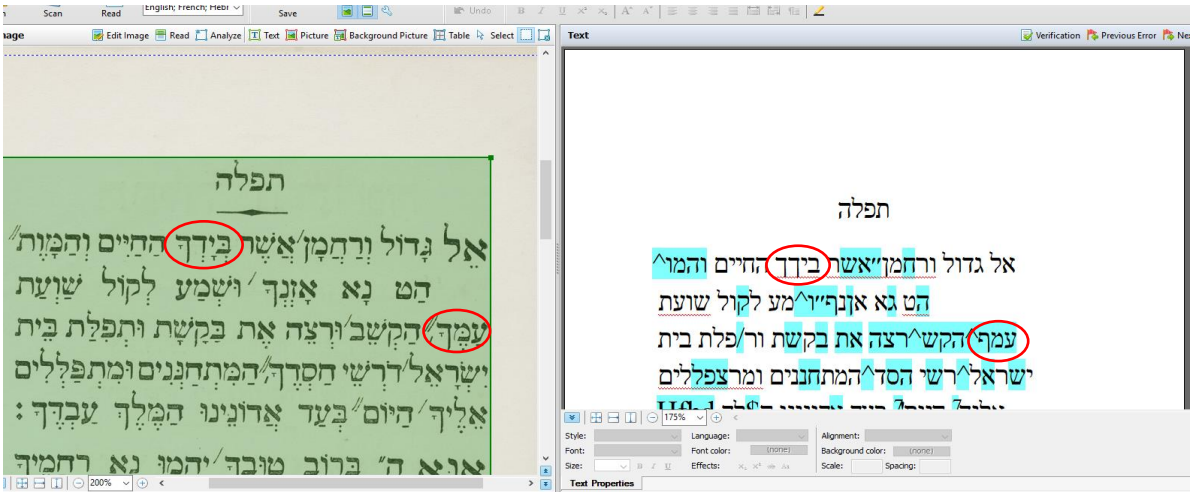
Figure 9 - Corrected version of Figure 8

Thus, it was essential that transcribers proofread the document in order to correct any misreadings that may have been missed. At the beginning of the project, transcribers were instructed to highlight any such corrected misreadings in bold in the transcription in order that the discoverability of these corrections could be checked in the final PDF file. Later in the project, an alternative means of checking the accuracy of transcriptions was employed, namely, the copying of transcribed text from the PDF into Microsoft Word. This process is described more fully in Section 5.5.1.

5.4.3 Common errors

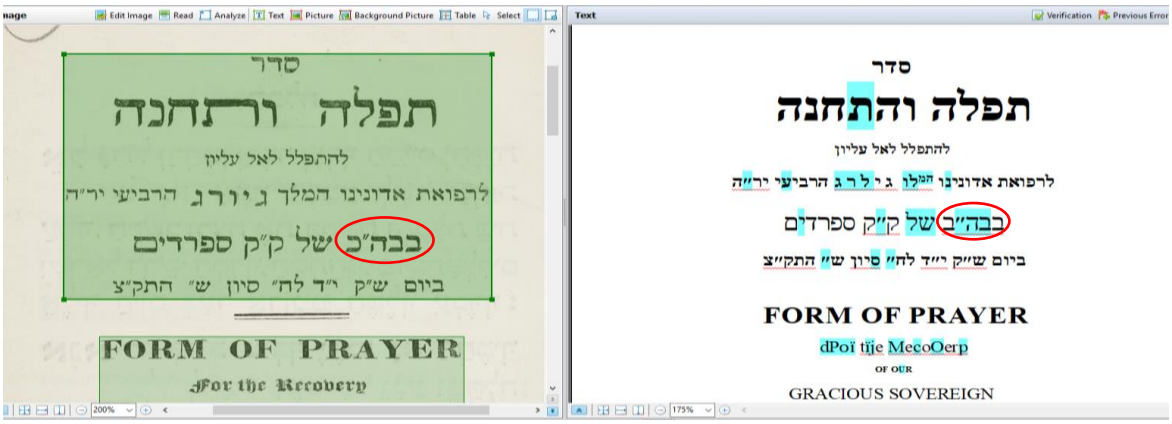
The transcribers also kept records of common errors found in the transcriptions in order to build up a picture of the types of error being encountered. This enabled them to anticipate where potential errors could lie according to language, typescript, font, and so on.

Certain letters were frequently confused, see some examples below:

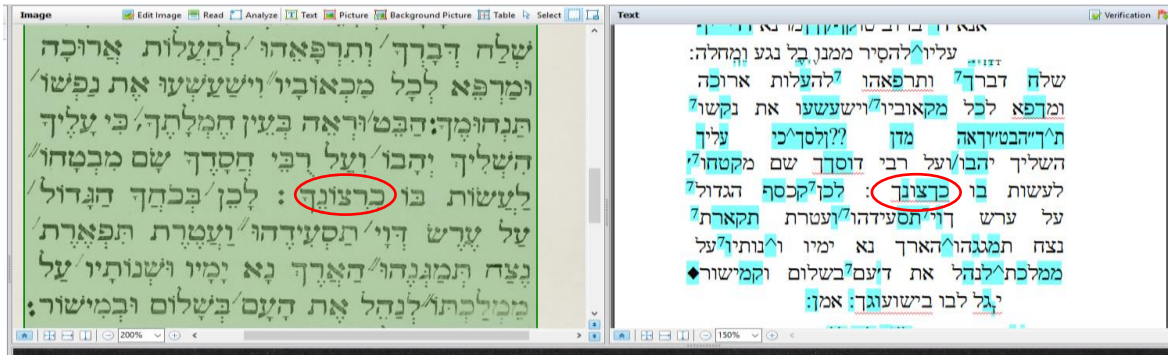


Dalet ד – read as Khaf sofit ך

Khaf sofit ך – read as Fe sofit ף



Kaf כ – read as Bet ב



Resh ר → read as Khaf sofit ך

Some other common OCR errors included:

Hebrew script

ה ↔ ת

ו ↔ י

ג ↔ נ

נ ↔ ג

אלהים ↔ אליהים

כל ↔ בל

ו ↔ 1

ט ↔ ש

Roman script

Capital I → 1

Capital I → lower case l

O → 0

E ↔ F

M → H, /'\

W → V, \./, M

H → FI, PI

B ↔ R

F ↔ R ↔ P

c ↔ e

h → n

h ↔ li

tl → d

is → b

e ↔ o ↔ c

the → die

rn → m

G ↔ C ↔ O

Italics: f → / (backslash), or l (italic l)

The above list is not exhaustive; many instances depended on the typeface used and could occur extensively in one document but not another.

5.4.4 Hebrew vocalisations

The decision was taken at the outset of the project not to include vocalisations in the transcriptions of the Hebrew texts. The chief reason for this was that the vocalisations were read separately to the letters to which they related and therefore were not reproduced correctly in a printed transcript. More importantly, they were not necessary in order to search the transcription successfully.

What was observed when ABBYY FineReader read the script was that the line of text was followed by a line containing the vocalisations, in other words, they were read as separate lines. However in the transcriptions, these lines could be superimposed leading to a confused image, especially if the font size was enlarged during the correction process (see Figure 10).

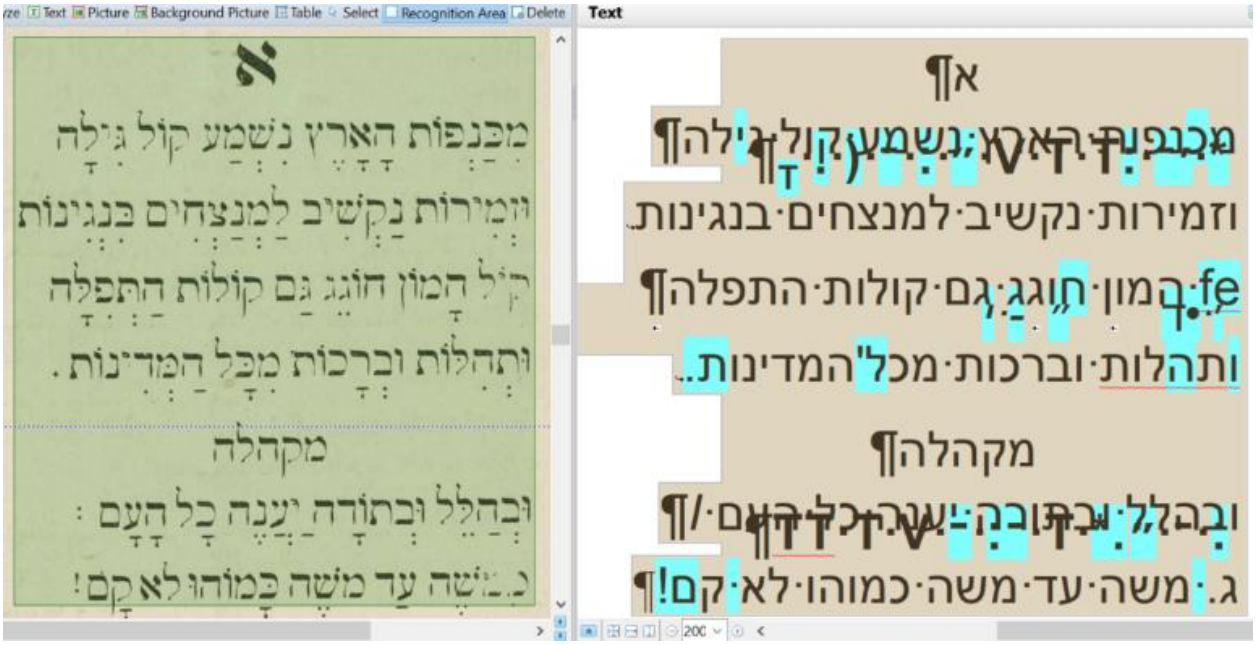


Figure 10 - Readings of vowel points overlapping consonantal text; font: Arial 16

Transcribers found that the superimposition rendered deletion of vocalisations almost impossible. However, having sought Library Services' Digital Curation Team's advice regarding adjusting font size in transcriptions (see Section 5.5.1), it was discovered that reducing the text to a maximum font size of 8 would separate the lines sufficiently to enable the the line of vocalisations to be deleted. Figure 11 shows the same passage of text reduced in size to Arial 6. The first line of vocalisations has been highlighted to demonstrate the distinction between the lines ensuring that, upon deletion, the consonantal text will not be affected.

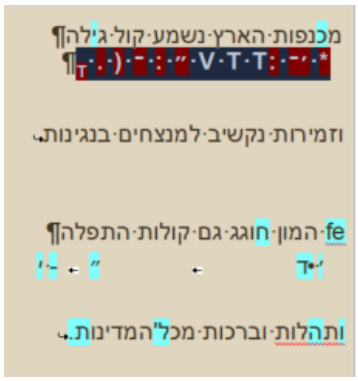


Figure 11- Font changed to Arial 6

The lines of poetry in Figure 12 are one of the many examples of vocalisations occurring in Hebrew text:

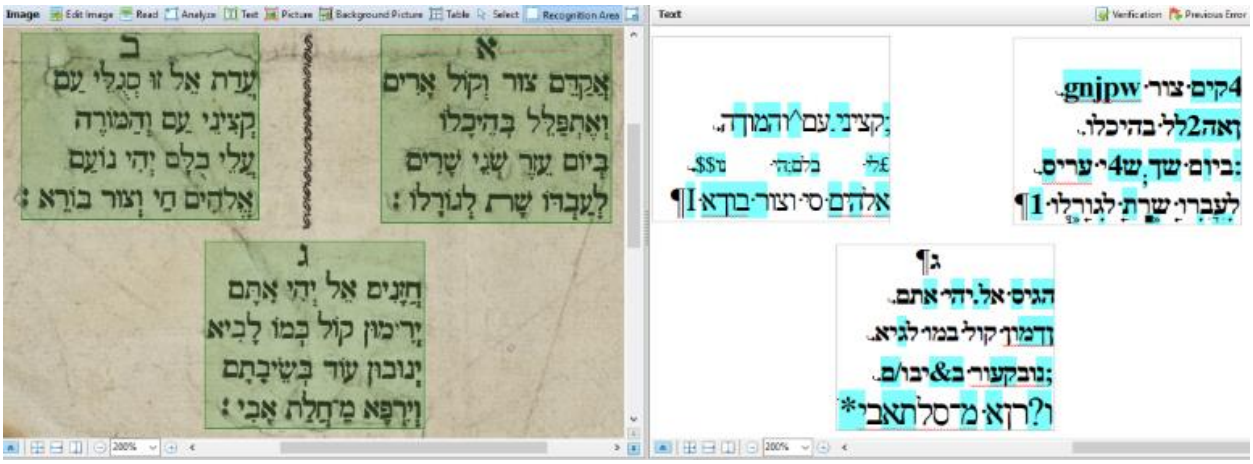


Figure 12 – Vocalisations in Hebrew text shown before correction

In Figure 13, the transcription has been corrected and the vocalisations removed.

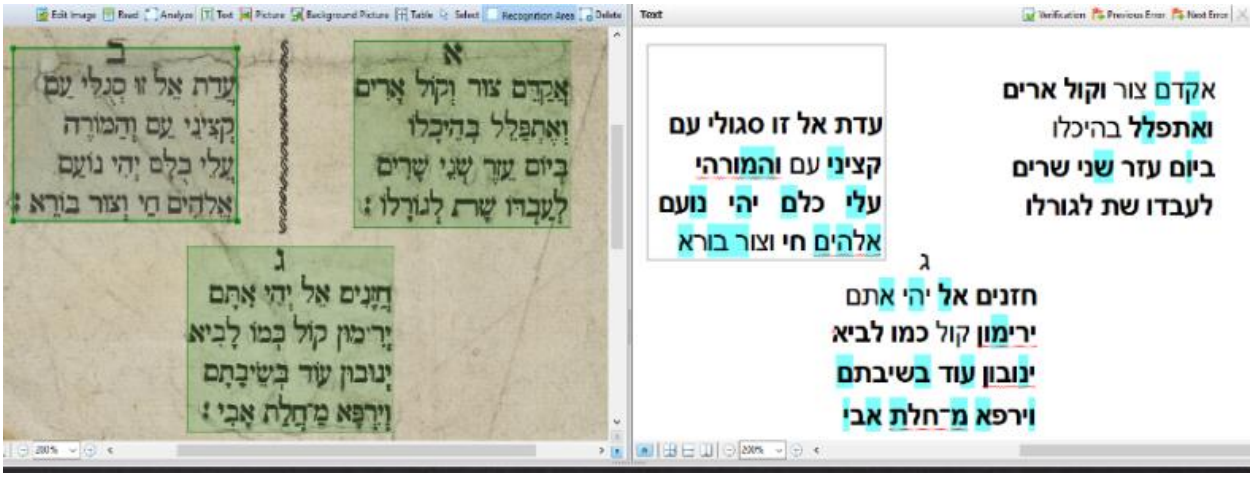


Figure 13 – The corrected text without vocalisations

5.4.5 Cursive Hebrew and Gothic Latin scripts

The Hebrew texts often contained cursive and semi-cursive scripts as show in in Figure 14. In these instances, the software frequently had difficulty transcribing any of the text correctly. In such cases, it would have been more time-efficient to transcribe the text manually, in other words, overwriting the existing text completely and retyping it rather than by using software to correct the text on a letter-by-letter basis. However, such over-written additions, although searchable in the final PDF, would not have been linked to the corresponding text in the document image. For this reason, this practice was not adopted.

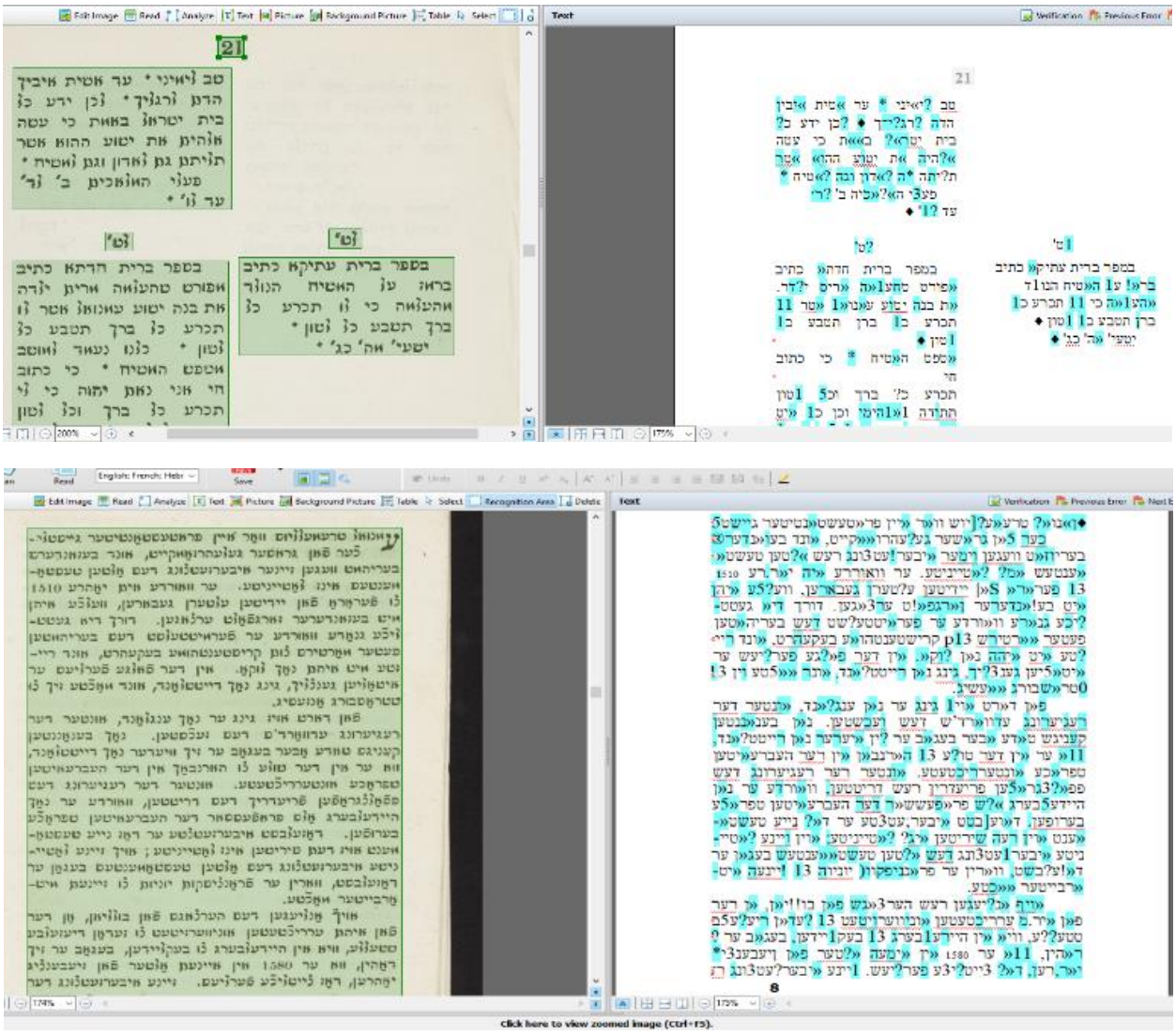


Figure 14 – Examples of cursive and semi-cursive Hebrew scripts not read correctly by the software

Similarly, Gothic script could not be read satisfactorily by the software as shown in Figure 15. Specific software is available to deal with such text and external services can also be used to provide corrections on a cost-per-page basis. In this project, however, Gothic text was only encountered occasionally, for example in titles. It was therefore considered more practical and cost-effective to correct such instances manually as they occurred.

Figure 15 – Example of Gothic script

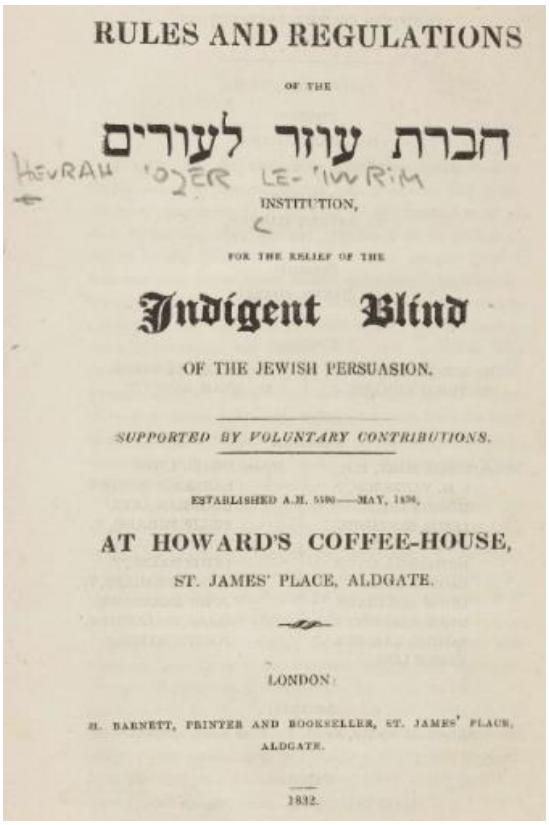
5.5 Document checking and quality control

5.5.1 Preliminary investigations

Once the ABBYY FineReader document had been saved, it was then re-saved as an archival PDF document (PDF/A). Initial testing of the searchability of these PDF documents showed that many of the corrections noted in bold as described in Section 5.4.2 were not discoverable in the PDF text. Moreover, sections of the text which already appeared to be correct in the transcriptions were also failing to be discoverable in the completed PDFs. Following advice from the Digital Curation Team, transcribers highlighted areas of the PDF text, copying and pasting these into a Word document which was set to show all hidden formatting symbols by selecting the ¶ icon. This led to the discovery of instances of spacing appearing within words, often separating each letter (see Figure 16); these had not been present in the original text or the transcriptions. Clearly, this was severely hindering discovery as words were no longer searchable as strings but as individual letters.

Figure 16 – Example of spacing occurring within words

This "interspacing" phenomenon did not occur consistently through passages but in an apparently random way as can be seen. The text "SUPPORTED BY VOLUNTARY CONTRIBUTIONS" is entirely interspersed with spaces save after "I" whereas the word "Indigent" has only one spacing interposed, in this case, after the "I". A comparison with the original image shown in Figure 17 gives an indication of the large variety of typefaces used within the publication.



RULES AND REGULATIONS

OF THE

חברת עוזר לעורים

HEVRAH 'OZER LE-'IWRIM

INSTITUTION,

FOR THE RELIEF OF THE

Indigent Blind

OF THE JEWISH PERSUASION.

SUPPORTED BY VOLUNTARY CONTRIBUTIONS.

ESTABLISHED A.M. 5590——MAY, 1830,

AT HOWARD'S COFFEE-HOUSE,

ST. JAMES' PLACE, ALDGATE.

LONDON:

H. BARNETT, PRINTER AND BOOKSELLER, ST. JAMES' PLACE, ALDGATE.

1832.

Figure 17 - Rules and regulations of the Ḥevrat 'Ozer la-'Iyrim Institution, for the relief of the indigent blind of the Jewish persuasion. London : H. Barnett, printer and bookseller, St. James' Place, Aldgate. 1832. De Sola Pamphlets vol. 4, no. 2.

ABBYY FineReader seeks to emulate these font styles and sizes as far as possible when reading the text. This led to an enormous variety of styles being used within transcriptions with font sizes ranging from 5 to 30 or more. Comparing a more detailed analysis of the text styles with the corresponding layout in the transcription pointed to common areas where spacing inaccuracies occurred:

- Large fonts in block capitals
- Font sizes smaller than 8
- Texts in italics

Typically, these texts corresponded to specific publication areas, that is to say:

- Titles
- Imprint statements
- Colophons
- Footnotes

Titles in particular were seen as a major concern as, being the chief source of information, it was considered imperative that the title should be fully searchable. It is worth stressing that the majority of the pamphlets displayed similarly large variations in typeface on their title pages as does the one in Figure 17.

5.5.2 Guidance on font sizes

Following these discoveries, additional advice from the Digital Curation Team was sought. They recommended applying a standard font and minimum font size to ensure consistency of the text. For Arial and Times New Roman fonts, the minimum sizes suggested to achieve accurate results were:

Font	**Minimum size**
Arial	8
Times New Roman	10

Consistency could be ensured in two ways:

1. On a page-by-page basis: sections of text could be highlighted and the font and size selected from the toolbar
2. At document level: from the menu Tools option, the Style editor could be selected (see Figure 18) and the fonts set for the whole document.

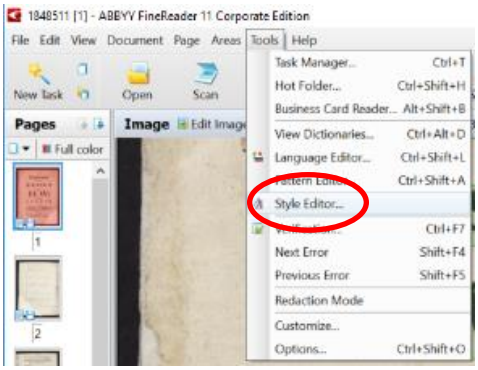


Figure 18 –Style Editor selection

The Style Editor window shown in Figure 19 demonstrates the large variety of font size used in the transcription of a single pamphlet.

Style Editor

New Merge...

Style name	Font name	Font size	Scale	Spacing	B	I	Aa
Heading #4	Times Ne…	14	100%	0	B	I	Aa
Heading #3 (2)	Times Ne…	15	100%	-1	B	I	Aa
Heading #2 (3)	Times Ne…	10	100%	0	B	I	Aa
Heading #1 (2)	Times Ne…	25	100%	-1	B	I	Aa
Body text (17)	Courier New	5	100%	0	B	I	Aa
Body text (16)	Times Ne…	8	100%	0	B	I	Aa
Body text (14)	Times Ne…	14	100%	0	B	I	Aa
Body text (13)	Times Ne…	10	100%	0	B	I	Aa
Body text (12)	Times Ne…	18	100%	-1	B	I	Aa
Body text (3)	Times Ne…	13	100%	0	B	I	Aa

Figure 19 – ABBYY FineReader Style Editor

5.5.3 Revised methodology

In accordance with this new guidance, a new methodology for working on the transcriptions emerged. This entailed carrying out a preliminary check of a sample page of the final transcript for accuracy before proceeding with the rest of the work, as shown in this summary of the instructions:

1. Correct the first page or one selected page only; save the FineReader document; right-click on the page thumbnail image in the left-hand pane and save that page as a PDF in your own working folder, adding the font size at the end of the name, e.g. 1840991-8.pdf.
2. Open the PDF; highlight the text, copy this and paste into a blank Word document; click the ¶ icon.
3. Check the text for accuracy in the Word document; in particular, check for irregular spacing issues.
4. If the Word text is satisfactory, proceed to correct the ABBYY document; otherwise, return to the ABBYY file and try selecting a different font; re-save as a PDF and check the text in Word as before; if the text continues to have too many errors, discontinue working on this transcription and select another one.

In the majority of cases, following this procedure led to a significant improvement in the discoverability of text within transcriptions. However, instances of spacing within words were not entirely eliminated. Contrary to what one might have expected from the guidance cited in Section 5.5.2, titles remained the chief area where unwanted spacing occurred despite having font sizes well above the recommended levels.

The final part of the revised methodology, therefore, was for the transcribers to keep records, or Transcription-Checking Logs, of the font sizes used during the transcription process. Additionally, the quality of all PDFs was monitored upon completion of the transcription and, where possible, remedial action undertaken on a case-by-case basis to improve discoverability, details of which were recorded in a Transcription Monitoring Log. These logs will be described in more detail in the following section.

5.6 Record keeping and evaluation

5.6.1 Transcription-checking logs

The purpose of these logs was for transcribers to identify problematic fonts and font sizes, and to keep a record of those which were successful. This information was then used to anticipate potential problematic sections of text and take pre-emptive steps to correct them before carrying out a preliminary page sample check as outlined in Section 5.5.3. This not only saved a considerable amount of time but also made the transcription process a more streamlined and satisfying experience. The template shown in Table 1 includes examples highlighted in red.

Table 1 – Transcription-Checking template

Transcription-Checking Template

Language
E=English
H=Hebrew

			Didn't work		Did work		
PDF no.	**Section**	**Language**	**Font**	**Size**	**Font**	**Size**	**Notes**
2087198	*title*	*E*	*Times New Roman*	*12*	*Times New Roman*	*18*	*In part of the title 'a discourse', the smallest font that didn't cause spaces was 36; for the part in Italics, the largest font that worked was 8 as anything larger than this made the text completely unsearchable*

2087198	*title*	*H*	*n/a*	*n/a*	*Arial*	*14*	*Ariel 14 worked well on first attempt*
2087198	*text*	*E*	*Times New Roman*	*8*	*Times New Roman*	*10*	


RULES·AND·REGULATIONS¶
OF·THE¶
לעורים·עזר·חכרת¶
■HU·BARNETT,·PRINTER·AND·BOOKSELLER,·ST.·JAMES'·PLACE,¶
ALDGATE.¶
INSTITUTION,¶
FOR·THE·RELIEF·OF·THE¶
Indigent·Blind¶
OF·THE·JEWISH·PERSUASION.¶
SUPPORTED·BY·VOLUNTARY·CONTRIBUTIONS.¶
ESTABLISHED·A.M.·5590·MAY,·1830,¶
AT·HOWARD'S·COFFEE-HOUSE,¶
ST.·JAMES'·PLACE,·ALDGATE.¶
LONDON:¶
1832.¶

Figure 20 - Example of larger font sized used to eliminate spacing within words

By judicious use of this data, it was possible to judge more accurately the font sizes required, particularly in titles which, as have been noted, were the least successful elements of the transcriptions in terms of discoverability. Figure 20 shows this procedure applied to the earlier example in Figure 16.

Note that all the text below the Hebrew script has been enlarged considerably.

5.6.2 Time-monitoring logs

In order to monitor progress of the project, a record of the time taken per pamphlet was recorded in the template shown in Table 2. It can be seen that the time taken per page varied considerably. In General, Hebrew texts took longer to correct. However, it was also noted that the earlier the date of publication, the greater the average time required per page. This was generally due to the unevenness of the typescript and the more frequent instances of the text from the previous page being visible though the paper as described in Section 5.2.4.

Table 2 – Time-monitoring template

Unique ID	Languages	Date transcribed	Time per pamphlet (mins)	No. of pages	Average time per page
2046561	Hebrew	10-10-18	91	7	13
2087157	Hebrew	17-10-18	47	6	8
2111670	Hebrew	24-10-18	65	1	65
2087219	English	07-11-18	152	24	6

5.6.3 Transcription monitoring log

Once the transcriptions had been corrected by the transcribers, they were assessed for overall accuracy as outlined in Section 5.5.3. The process was similar: the PDF of the completed transcription was opened and the text copied and pasted into Word. This copied text was then given a searchability rating by selecting a representative passage containing 100 words and assessing the number correctly transcribed. The percentage was assigned a value according to the scale shown in Table 3.

Table 3 – Transcription searchability rating

Rating	% accuracy	Status	Action
0	0-20%	Problematic	Re-correct transcription in entirety
1	20-40%	Problematic	Re-correct transcription in entirety
2	40-60%	Problematic	Correct selected passages
3	60-80%	OK	Correct selected passages if time
4	80-95%	OK	Minor correction - chiefly title
5	95-100%	OK	None

The Digital Curation Team advised that an accuracy rating of over 60% could be considered satisfactory, therefore actions for further enhancement were graded accordingly. Thus, for texts rated 3 or higher, further correction was limited chiefly to titles and proper nouns occurring within the text in accordance with UCL Institute of Education Archives' best-practice guidelines. This assessment process was recorded in the template shown in Table 4.

Table 4 – Transcription monitoring and assessment template

Transcription Monitoring

				Problem areas				
Unique ID	Title	Issues	Language	Title	Text	Colophon	Searchability rating	PDF status
1841010	An appeal on behalf of the Jews scattered in India, Persia and Arabia	Some tablature structure lost, but text searchable	E				5	OK
2086145	Discourse on the Passover Festival.	Some spacing issues, esp in title and Hebrew text	EH	x	x	x	4	OK
2085862	Rules and Regulations of the Hevrat Ozer la-Ivrim Institution, for the Relief of the Indigent Blind of the Jewish Persuasion.	Spacing in titles and small fonts	E	x	x	x	2	Problematic

6 Analysis of data

The data gathered in the various logs referred to in Section 5.6 served three purposes: firstly, they led to refinements in the methodology as has already been outlined; secondly, they highlighted certain features of the selected pamphlet literature hitherto unrecorded; and thirdly, they confirmed several observations made about the transcription process during the project. In addition, reports generated from the LMS enabled further analysis of the corresponding bibliographic data.

Language

During the cataloguing process, the chief language of multi-lingual texts had been recorded in the 008 Fixed-Length Data field of the MARC record rather than "multiple languages" which was generally felt to be unhelpful, particularly when one language often predominated. In reality, the pamphlets selected for this project were rarely in only one language. The Transcription Monitoring Log highlighted instances of bi- and multi-lingual texts as represented by overlapping cells of the Venn diagram in Figure 21.

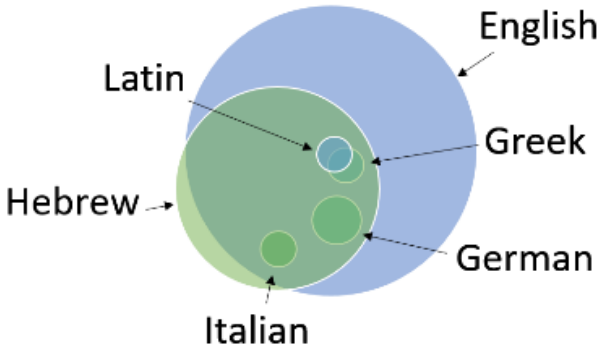
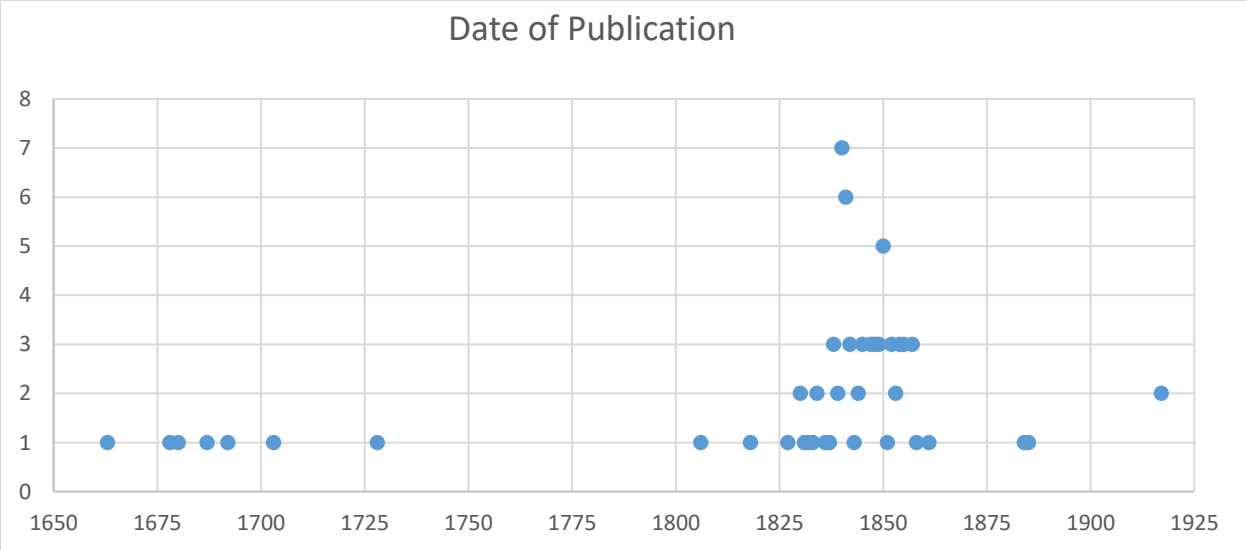


Figure 21 - Venn diagram showing distribution of languages encountered in the sample

Date of publication

Table 5 shows the distribution of the dates of publication of the literature which stem chiefly from two periods: late 17th century and mid-19th century.

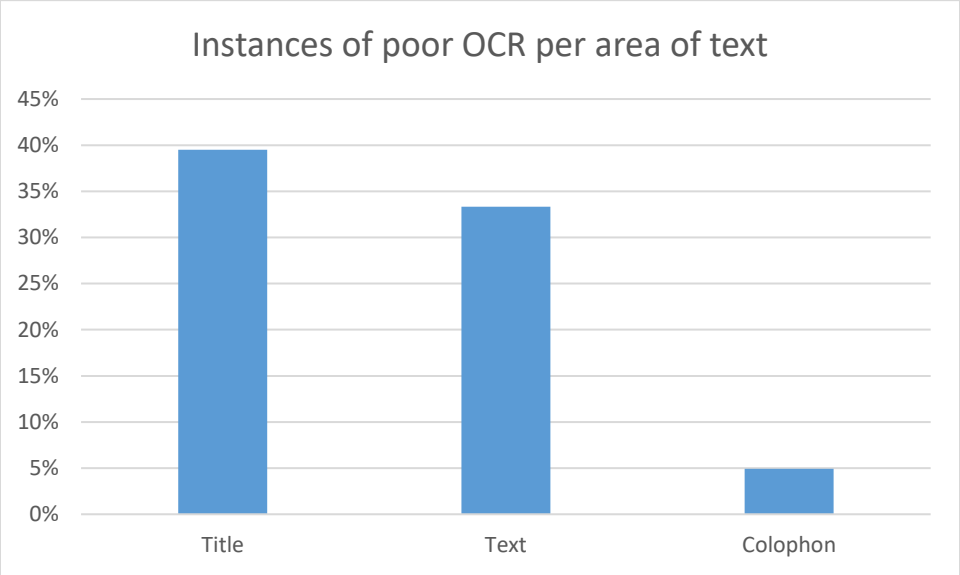
Table 5 – Scattergram showing distribution of publication dates encountered in the sample



Areas of text where OCR was substandard

In the transcription monitoring log, the areas of the text where significant problems were encountered in the OCR text were recorded. This supported the general observation that titles were more problematic for the software to read than text or colophon areas as show in Table 6.

Table 6 – Problematic areas of text



As noted earlier, this was attributed to the large variety of typography encountered in titles as well as the unconventional layouts requiring multiple text boxes to preserve the integrity of the text.

Font size

In the transcription-checking log, detailed notes were made of the font sizes which had led to spacing within words in the transcription, and the changes in font size required to eliminate this and improve the accuracy of the transcription. Table 7 shows the average font sizes that were deemed to be the minimum required for successful searching and the average increase in font size required for problematic areas of text.

Table 7 – Changes to font size

Language	Font	Section	Size	Average increase required
English	Times New Roman	Text	11	125%
		Title	20	
Hebrew	Arial	Text	17	150%
		Title	25	

Thus, if a portion of an English title in font Times New Roman was not read correctly at size 18, for example, increasing it approximately 125% to size 22 would generally ensure success.

Accuracy rating
The improved accuracy of the transcriptions resulting from the work carried out in this project was calculated from the transcription monitoring log and is summarised in Table 8.

Table 8 – Overall improvement in accuracy

	Average accuracy rating		
Attribute	**Before project**	**after project**	**Overall improvement %**
Pre-1800 publication	58%	91%	57%
Post-1800 publication	62%	90%	45%
Roman Script	73%	90%	23%
Hebrew Script	47%	85%	81%

The figures were calculated based on the number of correct words identified in a representative passage of 100 words. The increase in rating was compared with the initial rating to calculate the overall improvement percentage.

The figures confirm observations made in practice: namely, that transcriptions for pre-1800 texts and texts in Hebrew script required the greater amount of correction. It is therefore highly likely that in further work of this nature, these texts would be prioritised for attention. However, the corrections required in Roman scripts frequently corresponded to proper nouns not recognised in the software's integral dictionaries; therefore, there is still merit in correcting these texts when such nouns are likely to be key search terms.

Transcriptions completed and time taken
Lastly, the time monitoring log enabled the time requirements for this type of transcription-enhancement work to be calculated (see Table 9).

Table 9 – Time requirements

Activity	**Metrics**
Transcriptions completed	81
Number of pages corrected	1,779
Minimum time required to prepare ABBYY file per pamphlet	1.5 hours
Average time taken to correct one page	10.4 minutes
Average time taken to assess final transcription and carryout remedial corrections per pamphlet	30 minutes
Minimum average time required per pamphlet	6 hours

7 Conclusion
This was a pioneering project for UCL Library Services. With little practical support and guidance available from the library community, it was clear from the outset that we needed to adopt a flexible, exploratory approach, revising the methodology as necessary, and adjusting expectations of the project outcomes as a consequence.

The pilot conducted in Phase 2 of the Jewish Pamphlets Project left us in no doubt as to the painstaking nature of this work: correcting the transcripts was repetitive, often frustratingly slow to carry out and

required unremitting attention to detail. As such, it was difficult to envisage working continuously at this task for lengthy periods. The literature itself was also known to present a number of challenges: the formats and typography were diverse, publications dates wide-ranging and instances of multi-lingual texts considerable.

This leads to consideration of the combination of skills sets required for this type of work. Beyond the essential: linguistic skills; the ability to maintain speed and accuracy while undertaking repetitive tasks; and attentiveness to detail, experience of working with rare printed material is highly desirable. But perhaps the overarching quality required for successful delivery is persistence; this, when applied to voluntary roles, translates into commitment. The student volunteers who worked on this project, although highly skilled, were unable to commit to the lengthy periods required to bring individual transcriptions to completion. This was partly our fault: we had offered them a choice of texts, some quite lengthy. In future, we would be much more judicious in selecting texts which they could complete within a shorter time frame.

The ABBYY FineReader software was suited in many ways to dealing with this literature. The text-selection tools were sophisticated and able to demarcate the diverse areas as required: tables, annotations, footnotes, colophons. They also allowed the deselection of initials and repeated punctuation which semantically separated sections of text from one another. In addition, the verification tool and integral multi-lingual dictionaries offered valuable means of highlighting and correcting errors rapidly and, crucially for this project, recognised Hebrew script, although in our experience, the spellcheck facility for Hebrew proved to be less effective than that for English and other major European languages.

However, the software had difficulty reading some of the many fonts occurring in early publications as well as Gothic and cursive Hebrew scripts. In such cases, it would have been much quicker to re-type the text from scratch but the link between text and image would then have been lost. Similarly, copying and pasting texts, such as Biblical passages, was ineffectual. Additionally, superfluous spacings appearing in the completed transcriptions proved detrimental to searching and required a trial-and-error approach to be ensure elimination.

From the outset, it was clear that evaluation of the methodology and analysis of the results needed to be an on-going process in order that lessons learned could be instantly applied to the workflow. The transcription checking log devised in response to the spacing phenomenon proved to be a valuable tool for identifying areas of text which were likely to be problematic in the final transcription. Logs allowed transcribers to predict the appropriateness of fonts for specific portions of text thereby minimising the need for extensive remedial correction during the final quality check. However, attempts to standardise procedures and thereby minimise the time required on each transcription were rendered difficult as:

- We could not identify a standard font or font size which could be applied in all circumstances to ensure successful searchability
- Common errors identified in logs proved not to be widespread or consistent enough to form the basis of global search-and-replace procedures

However, our analyses of data did allow us to establish certain generalities:

- Titles were the area of text least successfully searched and the most problematic to correct

- For Hebrew script, the font Arial appeared to be more successful than Times New Roman, minimum text size 17 and minimum title size 25
- For Roman script, the font Times New Roman proved successful, minimum text size 11 and minimum title size 20
- For Hebrew script, using a font smaller than size 8 enabled additional lines of vowel points to be removed easily

We also established that items benefiting from the greatest overall improvement as a result of work in this project were:

- Pre-1850 publications (57% improvement)
- Hebrew script publications (81% improvement)

This, and the fact that it took on average 6 hours' work on each pamphlet, leads us to consider focusing future transcription-enhancement activity in these areas for the remainder of our Jewish Pamphlet Collections.

---

[1] Uncovering UCL's Jewish Pamphlet Collections https://blogs.ucl.ac.uk/library-hebrew/2014/07/uncovering-ucls-jewish-pamphlet-collections
[2] UCL Digital Collections repository: Jewish Pamphlets https://www.ucl.ac.uk/library/digital-collections/collections/jewish-pamphlets
[3] ABBYY FineReader PDF https://pdf.abbyy.com
[4] National Library of Israel https://web.nli.org.il/sites/nli/english/pages/default.aspx
[5] Digitising The Woman Teacher https://nuwtarchiveioe.wordpress.com/2013/11/12/digitising-the-woman-teacher
[6] Montefiore Endowment https://www.montefioreendowment.org.uk/collections/testimonials
[7] UCL Digital Collections repository: Tributes to Sir Moses Montefiore https://www.ucl.ac.uk/library/digital-collections/collections/montefiore