2014 11th IAPR International Workshop on Document Analysis Systems

Ground-Truth Production in the tranScriptorium Project

B. Gatos and G. Louloudis
Inst. of Inf. and Telecommunications
National Centre for Scientific
Research "Demokritos"
Athens, Greece.
{bgat, louloud}@iit.demokritos.gr

Tim Causer and Kris Grint
University College London
Bentham House - Endsleigh Gardens
London, England
{t.causer, k.grint}@ucl.ac.uk

V. Romero, J. A. Sánchez,
A. H. Toselli and E. Vidal
Dpto. de Sist. Informáticos y Computación
Universitat Politècnica de València
Valencia, Spain
{vromero, jandreu, ahector, evidal}@dsic.upv.es

***Abstract*—tranScriptorium is a 3-years project that aims to develop innovative, cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using Handwritten Text Recognition (HTR) technology. The production of ground-truth (GT) of a dataset of handwritten document images is among the first tasks. We address novel approaches for the faster production of this GT based on crowdsourcing and on prior-knowledge methods. We also address here a novel low-cost semi-supervised procedure for obtaining pairs of correct line-level aligned detected/extracted text line images and text line transcripts, specially suitable for training models of the HTR technology employed in tranScriptorium.**

I. Introduction

Huge amounts of handwritten historical documents are being published by on-line digital libraries world wide. However, these raw digital images need to be annotated with informative content. The tranScriptorium[1] project [3] aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using modern, holistic Handwritten Text Recognition (HTR) technology.

For typical handwritten text images of historical documents, traditional Optical Character Recognition (OCR) is simply not usable since characters can not be isolated automatically in these images. Therefore, holistic, segmentation-free HTR techniques are needed [4]. Currently, these segmentation-free techniques run at line level, but it is expected in the future to work at higher levels (paragraph level or page level). Current technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMMs) and N-grams [5].

To achieve good HTR accuracy, a combination of techniques is needed, such as layout analysis, text line detection and extraction, preprocessing operations, lexical and language modelling, HMM training, etc. Although these technologies are already providing useful results in some cases, much remains to be developed, especially for historical documents.

The models used in segmentation-free HTR are trained using already well known, powerful learning techniques, most of them based on the Expectation-Maximisation algorithm. Therefore, training data is needed to build these models.

tranScriptorium has focused on four languages: Spanish, German, English and Dutch. Bentham manuscripts have been chosen for English [6]. Bentham collection is a large set of documents that were written by the renowned English philosopher and reformer Jeremy Bentham (1748-1832) about different topics. The enormous influence of Bentham writings on his time makes the transcription of this collection very interesting[2]. This transcription is currently being carried out by amateur volunteers participating in the award-winning crowd-sourced initiative, Transcribe Bentham[3]. Currently, more than 6,000 documents have been transcribed.

One of the first tasks defined in the tranScriptorium project was to create a ground-truth (GT) for each dataset. This is necessary both for training HTR models and for testing the developed techniques. The transcripts produced in the Transcribe Bentham initiative are a very valuable information, but they could not be used in their existing format. From an HTR perspective, for these transcripts and their corresponding images being useful, an adequate GT should be prepared.

This GT consisted mainly on, first, to add to the transcripts relevant information that was no present in the initial transcripts (hyphened words, catch words, etc). Second, to annotate the geometric information of the main layout parts, that is text blocks and lines. Third, to put in correspondence the transcribed lines and the physical lines in the images. Different critical problems were foreseen in this third part: first, just an additional line in the transcripts or in the images could produce a shift in the remaining lines. Second, even if the number of transcribed lines and line images is the same, the pairing could not be correct. Given the large amount of transcribed data, it was not feasible to perform these pairings manually, and therefore an automatic procedure was defined. The GT generated was finally recorded in PAGE format [7].

This paper describes how this automatic procedure was carried out. Section II describes how the Bentham dataset used in tranScriptorium was organised. Section III describes the GT creation of the line images, while Section IV describes the GT creation of the line transcripts. Section V describes the semi-automatic process for pairing the line transcripts and the lines images, and Section VI presents preliminary HTR results on a small set with this semi-automatic pairing process.

[1] http://www.transcriptorium.eu

[2] http://www.ucl.ac.uk/Bentham-Project/

[3] http://blogs.ulcc.ac.uk/td/transcribe-bentham

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.23

II. Bentham dataset

The Bentham collection has more than 80,000 documents, most of them digitised. From the digitised documents, more than 6,000 have been transcribed with the crowd-sourcing platform previously mentioned. In this platform the registered users can transcribe the documents and the transcripts are recorded in TEI-compliant XML format. Given the nature of the transcription process, the transcripts produced by the amateur volunteers are not completely consistent, and therefore the transcripts are finally reviewed by experts transcribers. It is important to remark that in the annotation process no geometric information is registered, which means that there is not information between the transcripts and their corresponding geometric position in the images. Note that this information is necessary for training HTR models. Therefore, the problem that we also studied in this paper is how to pair the lines and their corresponding geometric information.

The physical documents in the Bentham collection are written without any standard and they have a very variable layout with different difficult degrees. The collection has also documents with different physical geometry. In order to produce the GT as much automatically as possible, the images were classified according to the expected difficulty in the GT production, for being processed in increasing order of difficulty. Three criteria have been used for classifying the documents: first, a score was computed for each image according to the TEI elements that appear in the transcripts. Thus, a document image with many deletions, additions, stroke-out words, etc. had lower score than a document image that had few of these elements. Second, after sorting the images according to this score, they were classified according to their physical geometry. For the first round, only the images that had 2731 × 4096 pixels were chosen. Third, the resulting images were automatically clustered in three classes: images that had a main text block in one column and left margin, images that had a main text block in one column and right margin, and other. Horizontal projections were used for performing this clustering. Finally, the initial chosen images were the images that belonged to the first two clusters and which score were above a given threshold. The initial set is composed by 798 pages. The GT production of these document images is explained in the following sections.

III. Production of text line images GT

The production of text line segmentation GT is accomplished by involving a two-step sequential procedure. First, all text blocks are correctly detected and then, all text lines for each text block are correctly segmented.

A. Text block detection

The GT production requires as first step, the segmentation of text areas from non-textual ones. Historical handwritten documents do not have strict layout rules and thus, a layout analysis method needs to be invariant to layout variability. Most of the state-of-the-art methods focus on modern handwritten documents and only a few deal with historical handwritten documents [8], [9].

Although we have not yet automated the text block detection for complex documents and the work is done almost manually with the help of Aletheia tool[4], some help is provided for the documents having rule lines (see Fig. 1a). For those cases, we automatically detect horizontal and vertical lines with the help of enhanced profiles to localise text areas.

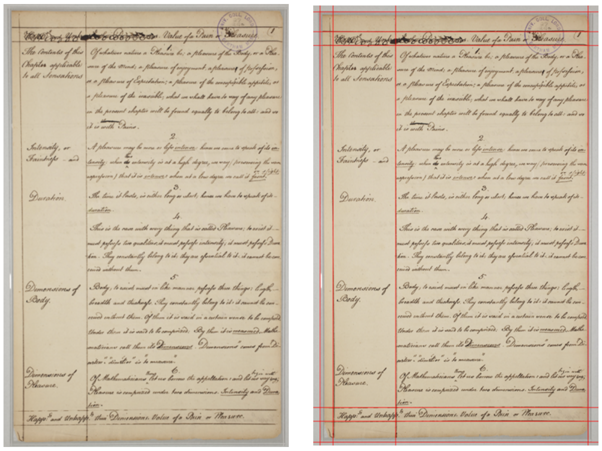
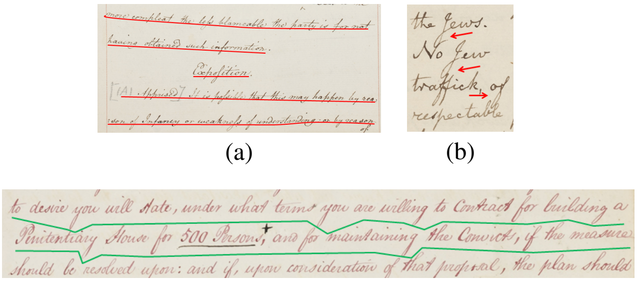

(a) (b)

Fig. 1. (a) A handwritten document with rule lines. (b) Horizontal and vertical lines detected.

B. Text line segmentation

Text line segmentation refers to the process of defining the region of every text line on a document image. Many challenges need to be addressed for text line segmentation which include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching (see Fig. 2).


the Jews.
No Jew
traffick of
respectable
to desire you will state, under what terms you are willing to Contract for building a
Penitentiary House for 500 Persons, and for maintaining the Convicts, if the measure
should be resolved upon: and if, upon consideration of that proposal, the plan should
(a) (b)

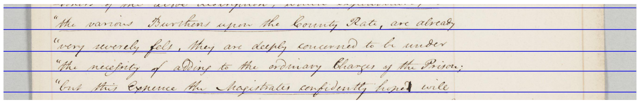
(c)

Fig. 2. Text line segmentation challenges: (a) skew angle differences, (b) touching text lines, (c) overlapping text lines.

Line segmentation can become easier and more robust if the baseline of each text line is previously detected with sufficient accuracy. Most traditional techniques hypothesise the vertical position of each line by detecting relevant picks in the text region horizontal projection profile [10]. Recently, a new approach has been developed which yields significantly more robust and accurate results [12]. It uses HMMs to model multiple horizontal projection profiles computed for several vertical slabs of the text region. A representative result of this method is shown in Fig. 3.

Preliminary text line segmentation results have been obtained on a subset of the Bentham dataset [6], using a

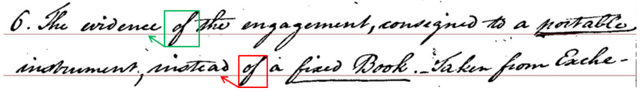
[4] http://www.primaresearch.org/tools.php



"the various Burthens upon the County Rate, are already
"very severely felt, they are deeply concerned to be under
"the necessity of adding to the ordinary Charges of the Prison;
"but this Expence the Magistrates confidently hoped will

Fig. 3. Representative result of the baseline estimation method for a portion of a document image.

novel methodology that takes into consideration the baselines produced using [12]. In more detail, after calculating the connected components of the image, a grouping procedure is applied based on the distances of the connected components from the baselines. Each connected component is assigned to the closest baseline (see Fig. 4).



6. The evidence of the engagement, consigned to a portable
instrument, instead of a fixed Book. Taken from Eache-

Fig. 4. Assignment of connected components to the closest baseline.

The text line segmentation output is a set of polygons (one for each text line) which includes all relevant text line entities (e.g. characters, words, punctuation marks, accents).

After evaluating this method using 433 images (from the 798 previously mentioned) which include 11,235 text lines from [6], we have observed that 9,157 text lines have been correctly detected (81.5%). We recorded that for the manual creation of the text line segmentation GT for one page the user needs about 700 seconds in average. We also estimated that the average time needed for the user to correct a text line segmentation error was 20 seconds using the Aletheia tool. We also took into account an average time for visually checking the generated text line segmentation result which was 40 seconds per image. Based on all the above mentioned observations, the total time needed for producing the text line segmentation GT by checking and correcting the results produced by the automatic procedure is estimated to 16 hours while the complete manual procedure would take 84 hours. This leads to a reduction of about 80% in terms of time needed.

IV. Production of transcripts GT at line level

A. Crowdsourcing

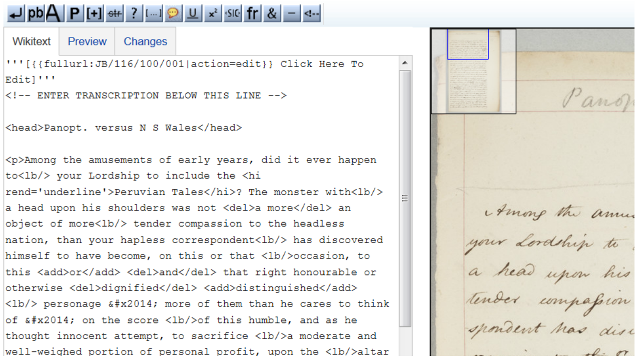
Crowdsourcing is a relatively recent phenomenon in the cultural and heritage sector, in which an organization or project makes an open call for assistance for online volunteers to assist in large-scale ventures such as tagging, commenting, rating, reviewing, text-correcting, and the creation and uploading of content. This activity is harnessed in order to improve the quality of, and widen access to, online collections. Perhaps the most famous examples are Galaxy Zoo and the National Library of Australia's historic newspaper programme[5,6].

Transcribe Bentham was launched to the public in 2010 in order to assess whether untrained, amateur volunteers were capable to transcribing complex manuscripts, and whether their work would be of a suitable standard for uploading to a digital repository for access and searching, and for editing as part of the ongoing work to produce a critical edition of Bentham's works. As of 11 October 2013, volunteers have transcribed or partially-transcribed 6,345 manuscripts, or an estimated 3.2 million words. Of these transcripts, 6,041 (95%) have met the project's quality-control standards.

The produced transcripts have two main purposes. First, they will be uploaded to UCL's free-to-access digital repository of Bentham's manuscripts, for research purposes and to ensure the long-term digital preservation of this priceless collection[7]. The volunteers' transcripts will also feed into scholarship, making new discoveries about Bentham's life and thought, and feeding directly into the production of the critical edition of Bentham's works[8].

B. Acquisition of transcripts and protocols

Volunteers access the manuscripts and the transcription interface via the "Transcription Desk" website, which is a customized installation of MediaWiki developed by the University of London Computer Centre[9,10]. As volunteers are also requested to encode their transcripts in TEI-compliant XML, a "transcription toolbar" was developed in order to allow them to add basic TEI formatting to their work without necessarily having to learn the minutiae of mark-up (see top of left figure in Fig. 5).



```
'''[{{fullurl:JB/116/100/001|action=edit}} Click Here To
Edit]'''
<!-- ENTER TRANSCRIPTION BELOW THIS LINE -->

<head>Panopt. versus N S Wales</head>

<p>Among the amusements of early years, did it ever happen
to<lb/> your Lordship to include the <hi
rend='underline'>Peruvian Tales</hi>? The monster with<lb/>
a head upon his shoulders was not <del>a more</del> an
object of more<lb/> tender compassion to the headless
nation, than your hapless correspondent<lb/> has discovered
himself to have become, on this or that <lb/>occasion, to
this <add>or</add> <del>and</del> that right honourable or
otherwise <del>dignified</del> <add>distinguished</add>
<lb/> personage &#x2014; more of them than he cares to think
of &#x2014; on the score <lb/>of this humble, and as he
thought innocent attempt, to sacrifice <lb/>a moderate and
well-weighed portion of personal profit, upon the <lb/>altar
```

Fig. 5. The "Transcription Desk" transcription interface.

In practice, the volunteer is presented with a zoomable and navigable manuscript image, a plain text box into which they type their transcript, and the transcription toolbar (see Fig. 5). Transcribers can highlight a piece of text and click a button on the toolbar to identify a particular characteristic of the chosen portion. These include spatial and organizational features such as line breaks, page breaks, headings, and paragraphs; linguistic features like marginal notes, unusual spellings, and non-English text; compositional features such as additions and deletions; and interpretive decisions regarding questionable readings or illegible text.

Since the aim was to make adding TEI mark-up as straightforward as possible, and to avoid obscuring users' transcripts with too much code, minimal mark-up has been employed, using only element names, and avoiding attributes and attribute values where possible. For instance, where a word is broken

[5] http://www.galaxyzoo.org

[6] http://trove.nla.gov.au/newspaper

[7] http://www.ucl.ac.uk/library/bentham

[8] http://www.ucl.ac.uk/Bentham-Project/publications/collected_works

[9] http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham

[10] http://www.ulcc.ac.uk

over two lines, as in the case of "insanity" in the example in Fig. 6, Transcribe Bentham volunteers are asked to complete that word before adding a line-break, hence:

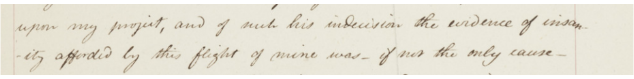
```
    ... the evidence of insanity<lb/>
    afforded by this flight of mine ...
```

It would have been possible to encode this as

```
... the evidence of insan-<lb break="no"/>
-ity afforded by this flight of mine ...
```

in order to indicate that the line-break does not mark the start of a new word, but introducing two forms of line-break would be unnecessarily confusing for volunteers who had little or no experience of mark-up prior to participating in Transcribe Bentham. However, as previously mentioned, for the purposes of generating the GT data for TRANSCRIPTORIUM, some changes are required in the transcription. In the case of broken words the line-breaks have been restored as follows:

```
  ... the evidence of insan-<lb/>
  -ity afforded by this flight of mine ...
```



upon my project, and of such his indecision the evidence of insan-
-ity afforded by this flight of mine was – if not the only cause –

Fig. 6. Bentham manuscript JB/116/100/001.

In addition to the hyphened words, another relevant information, that was no present in the initial transcript provided by the volunteers, have been added for HTR purposes. Between this information stand out the catch words or the page numbers.

C. Quality control

When the volunteer is happy that their transcript is complete, they submit it for assessment by a Transcribe Bentham editor with experience in reading and transcribing Bentham's manuscripts, who checks the transcript for textual accuracy and consistency of encoding. Changes are made to the text and mark-up, if necessary, with the key question being whether any appreciable improvement is likely to be made through further crowdsourcing, and if the transcript is of the requisite quality for public viewing and searching, and as a basis for editorial work. If approved-i.e. if there are few or no unclear words or gaps in the text-the transcript is locked. If there are a number of gaps in the text, or the text is only partially transcribed, then the manuscript remains available for editing. Though an unavoidable impressionistic and subjective judgment, this process ensures that locked transcripts, and those used in this experiment, are a reliable guide to the contents and layout of the manuscripts [1], [6]. It currently takes an average of around five minutes to check a submitted transcript-considerably less than had we transcribed the manuscript ourselves-though there are great variations depending on the length and complexity of the original manuscript [2].

V. SEMI-SUPERVISED ALIGNMENT BETWEEN GT OF LINE TRANSCRIPTS AND GT OF LINE IMAGES

As commented in previous sections, for each page image of the Bentham collection the production of GT sets, corresponding to the line transcripts and the line images, were made independently from each other. Hence, there was not guarantee that the number of line transcripts and the number of line images coincided, as well as the elements of both GT sets were correctly paired. Correct pairing of the elements of both GT sets is actually of vital importance to the adequate training of the HMMs employed by the HTR technology in the TRANSCRIPTORIUM project.

The causes that lead to a such disparity between both GT sets are due to different issues taking place during the production of them. Thus, with respect to the line transcripts GT production, the mainly causes are the missing line breaks and the forgetfulness of transcribing small pieces of text as catch-words, page numbers, etc. Likewise, concerning to the line images GT production, we have the undetected text additions-deletions, which appear marked in the line transcripts GT set in TEI format, but are quite difficult to detect using the approach described in Section III. Once all these issues are detected (usually by visual inspection) in each of the corresponding GT sets, they are manually corrected.

To speed-up the detection process of such mentioned issues, an approach called "morphology alignment" has been addressed to find the best alignment between the elements of both GT sets, associated to each page image of the Bentham collection. In more detail, for a given page image, this approach takes as inputs the sequence of estimated lengths (in terms of characters) of the segmented text lines and the sequence of transcripts lengths (also in characters), both of them produced for that page, and find the best possible matching between the lengths of these sequences by using dynamic programming. The minimisation of the sum of absolute differences between the matched length elements of the sequences was utilised as the objective function to find the best alignment. This alignment serves as a map to pair up the GT of segmented line images with the GT of line transcripts, and also points out through insertions/deletions and high-cost substitutions which are (hopefully) the unpaired elements of one GT set with respect to the other to directly proceed with the visual inspection of them. It is worth noting that in the case that some insertion, deletion or high-cost substitution has taken place to get the best alignment, this does not ensure in any way that the correct pairing between both GT sets is actually obtained.

Preliminary results of the morphological alignment previously described have been obtained. At the moment of testing it, the extracted text line images of 137 pages were available along with the line-level transcripts (ready for HTR GT) and thus, the validation experiment were carried out using just these pages. After executing the morphological filter obtained GT of each of such 137 pages, we observed that 53 did not have any alignment error. That meant that only in these 53 pages the number of segmented text line images and the number of line transcripts coincided. For the remaining pages, a manual revision was required in order to correct the detected errors.

As before mentioned, concerning to these 53 non-errors detected pages, it is important to note that they might still had pairing errors between the elements of both GTs. In order to check how many of these pages actually had paring errors, we checked and corrected the proposed pairings manually. From this revision we observed that from the 1,342 lines contained in the 53 pages, 45 of them had alignment errors, i.e. 3.4% of the total number of lines.

VI. PRELIMINARY HTR EXPERIMENTS ON THE BENTHAM DATA

In this section, we provide baseline results using standard techniques and tools for HTR. Most specifically, we have used a HTR system based on HMMs, where each character is modelled by a continuous density left-to-right HMM, with 8 states and 64 gaussians per state. HMM parameters were trained from line images and their corresponding transcripts employing the forward backwrd or Baum-Welch algorithm. Words are modelled by stochcastic finite-sate automatons which represent all poossible concatenations to compose words. Finally, the concatenation of words into text lines or sentences is modelled by an N-gram language model with Kneser-Ney back-off smoothing. More details of this system can be seen in [4].

The experiments presented here have been carried out with the 53 pages that have error 0 in the previous section. In order to check how the 3.4% of the alignment errors detected in these pages affect to the HTR results, we carried out experiments with both kinds of alignments (with and without errors). These 53 pages contain 1,342 lines with nearly of 12,000 running words and a vocabulary of more than 2,000 different words. The upper part of Table I summarises the basic statistics of these pages.

TABLE I. BASIC STATISTICS AND TRANSCRIPTION WORD ERROR RATE (WER) WITH AND WITHOUT CORRECTING THE ALIGNMENT ERRORS. VALUES OF RUNNING/LEXICON-SIZE OUT-OF-VOCABULARY WORDS (OOV) AVERAGED OVER THE 10 CROSS-VALIDATION FOLDS ARE ALSO REPORTED.

Number of:	Total
Pages	53
Lines	1,342
Running words	11,935
Average Running OOV	140
Average Lex-size OOV	128
Lexicon	2,181
Characters	63,400
WER with alig. err. (%)	**34.6**
WER without alig. err.(%)	**33.7**

The 53 pages have been divided into ten blocks of 5 or 6 pages each, aimed at performing cross-validation experiments. The last two rows in Table I show the results obtained with both kind of alignments (with and without errors). The quality of the automatic transcriptions obtained with the HTR system is measured by means of the WER. To carry out the experiments only the words seen in the training transcriptions were included in the recognition lexicon and, in the same way, bi-grams were estimated only from the training transcripts.

According to the results, the obtained WER using the transcriptions with alignment errors is only slightly higher than that obtained using the manually revised transcriptions. It is important to remark here that the system used in both cases is the same. Taking into account this small difference in the results, the process of GT generation could be accelerated by manually revising only those pages for which the morphological filter give some error.

VII. REMARKS AND CONCLUSIONS

Transcription of historical document employing HTR technology is one of the main goals in TRANSCRIPTORIUM project and thereby, the creation of adequate GT for such documents is of primary importance. Two kinds of GT are required by the HTR technology: one corresponding to the line images extracted from each document page, and the other to the line-level transcripts.

For one of the TRANSCRIPTORIUM documents: "Bentham's collection", novel approaches for the production of both GTs for each of its page images are presented. On the one hand, an approach based on previously-detected baselines is employed to speed-up and improve the text line segmentation process. On the other hand, crowdsourcing techniques are utilized for low-cost obtention (in time and resources) of the required line-level transcripts. Finally, a semi-supervised procedure to detect possible disparities between the GT elements (text line images and transcripts) for a given page is described.

The reported HTR results, despite of the difficulty of the task, are really encouraging. However, they should be considered preliminary, as significant improvements are expected when more GT data will be available for training both the language model and the HMMs.

VIII. ACKNOWLEDGMENTS

Work supported by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 600707 - tranScriptorium.

REFERENCES

[1] T. Causer, J. Tonra, and V. Wallace. Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham, Literary and Linguistic Computing, 27(2):119â137, 2012.

[2] T. Causer and M. Terras. "Many hands make light work. Many hands together make merry work": Transcribe Bentham and crowdsourcing manuscript collections, Crowdsourcing Our Cultural Heritage, ed. M. Ridge. Forthcoming, 2014.

[3] J. Sánchez, G. Mühlberger, B. Gatos, P. Schofield, K. Depuydt, R. Davis, E. Vidal, and J. de Does, "tranScriptorium: an European project on handwritten text recognition," in *DocEng*, 2013, pp. 227–228.

[4] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta, "Integrated handwriting recognition and interpretation using finite-state models," *IJPRAI*, vol. 18, no. 4, pp. 519–539, June 2004.

[5] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[6] T. Causer and V. Wallace, "Building a volunteer community: results and findings from Transcribe Bentham," *Digital Humanities Quarterly*, 2012, (in press).

[7] S. Pletschacher and A. Antonacopoulos, "The PAGE (page analysis and ground-truth elements) format framework," in *Proc. ICPR*, 2010, pp. 257–260.

[8] V. Malleron and V. Eglin, "A mixed approach for handwritten documents structural analysis," in *ICDAR*, Beijing, China, Sept. 2011, pp. 269–273.

[9] J.-Y. Ramel, S. Leriche, M. Demonet, and S. Busson, "User-driven page layout analysis of historical printed books," *IJDAR*, vol. 9, no. 2-4, pp. 243–261, 2007.

[10] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *IJDAR*, vol. 9, no. 2-4, pp. 123–138, 2007.

[11] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line and word segmentation of handwritten documents," *Pattern Recognition*, vol. 42, no. 12, pp. 3169–3183, 2009.

[12] V. Bosch, A. Toselli, and E. Vidal, "Natural language processing framework for handwritten text line detection in legacy documents," in *LaTeCH*, 2012.