Handwritten Text Recognition for Historical Documents in the tranScriptorium Project

Joan Andreu Sánchez
Universitat Politècnica de
València, PRHLT
46022 Camí de Vera s/n
València, Spain
jandreu@dsic.upv.es

Vicent Bosch
Universitat Politècnica de
València, PRHLT
46022 Camí de Vera s/n
València, Spain
viboscam@fiv.upv.es

Verónica Romero
Universitat Politècnica de
València, PRHLT
46022 Camí de Vera s/n
València, Spain
vromero@iti.upv.es

Katrien Depuydt
Instituut voor Nederlandse
Lexicologie
2300RA Matthias de Vrieshof
2-3, 2311 BZ
Leiden, Netherlands
Katrien.Depuydt@inl.nl

Jesse de Does
Instituut voor Nederlandse
Lexicologie
2300RA Matthias de Vrieshof
2-3, 2311 BZ
Leiden, Netherlands
Jesse.Dedoes@inl.nl

ABSTRACT

Transcription of historical handwritten documents is a crucial problem for making easier the access to these documents to the general public. Currently, huge amount of historical handwritten documents are being made available by on-line portals worldwide. It is not realistic to obtain the transcription of these documents manually, and therefore automatic techniques has to be used. tranScriptorium is a project that aims at researching on modern Handwritten Text Recognition (HTR) technology for transcribing historical handwritten documents. The HTR technology used in tranScriptorium is based on models that are learnt automatically from examples. This HTR technology has been used on a Dutch collection from 15th century selected for the tranScriptorium project. This paper provides preliminary HTR results on this Dutch collection that are very encouraging, taken into account that minimal resources have been deployed to develop the transcription system.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; I.5.4 [**Pattern Recognition**]: Applications

General Terms

Theory

Keywords

Handwritten text recognition, Historical documents

May 19-20 2014, Madrid, Spain
978-1-4503-2588-2/14/05
http://dx.doi.org/10.1145/2595188.2595193
.

1. INTRODUCTION

Transcription of historical handwritten document is a very important problem for accessing these documents that currently are being made available by on-line portals to the general public. tranScriptorium [14][1] is a three years project that aims to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images. Currently, huge amounts of these documents are being published by on-line digital libraries worldwide. For these raw digital images to be really useful, they need be transcribed. tranScriptorium aims to research on 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. These segmentation-free HTR techniques do not require to segment explicetly the characters even nor the words. Current technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition, such as Hidden Markov Models (HMMs) and N-grams [10, 15]. These models are trained from samples by using efficient techniques. A remarkable characteristic of these techniques is that few resources are needed to develop useful transcription systems.

To achieve good HTR accuracy, a combination of techniques is needed, such as layout analysis, text line extraction, pre-processing operations, lexical and language modelling, etc. Although these technologies are already providing useful results in some cases, much remains to be developed, especially for historical documents, which suffer from typical degradations [10, 13, 8].

The tools developed in tranScriptorium will be tested in real-life scenarios. Two user scenarios have been selected. In the first scenario, the technology will be made available through a content provider site. This is for users, some of which with occasional experience in transcription, who

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

access content provider portals for information or sources for research purposes, or out of interest in the cultural resources offered through these portals. For the second scenario, we have chosen to integrate the technology in an existing crowdsourcing platform for text transcription.

Section 2 introduces the HTR technology used in tranScriptorium. This technology has been applied in several collections in the project that are summarised in Section 3. The first Dutch collection selected in the project is introduced in Section 4. Finally, HTR experiments on this Dutch collection are reported in Section 5.

2. HTR TECHNOLOGY

Current state-of-the-art transcription products rely on technology for isolated character recognition developed for printed documents. But when character segmentation is just impossible in unconstrained handwritten text images like those encountered in most historical documents of interest, holistic segmentation-free *off-line* HTR technology [4, 15] has to be used.

HTR can be considered a relatively new sub-field of Pattern Recognition. This new HTR technology does not need the characters or even the words of a handwritten text image to be previously segmented or isolated. To some extent, the transcription of (historical) handwritten text images is comparable with the task of recognising continuous speech in a (significantly degraded) audio file. In fact, recent technology for HTR borrows concepts and methods from the field of Automatic Speech Recognition (ASR) such as Hidden Markov Models (HMM) and N-grams [9], where this technology is well consolidated. Currently existing HTR prototypes work at line level, mainly because efficient techniques for line detection and extraction exist [3], but it is expected in the future to work at higher levels (paragraph level or page level). Some of these HTR systems are based on HMMs [1, 10, 16] while others are based on hybrid HMM and Artificial Neural Networks [7]. 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 [6]. A detailed description of these techniques can be also seen in [13].

Currently available HTR technologies are far from offering error-free solutions. But the current obtained results can be useful for search, indexing and interactive transcription [13]. tranScriptorium intends to make progress in the mentioned models and in the automatic training techniques in order to achieve satisfactory accuracy.

3. CURRENT COLLECTIONS SELECTED IN TRANSCRIPTORIUM

tranScriptorium has focused on four languages: English, Spanish, German and Dutch. Several collections have been chosen for German. One of them is a collection on court decisions. The transcription of this collection is planned for the second year of the project.

Bentham manuscripts have been chosen for English [5]. 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. This transcription is currently being carried out by amateur volunteers participating in the award-winning crowd-sourced initiative, Transcribe Bentham[2].

The Bentham collection is a difficult collection from the layout analysis point of view. The images have many difficulties like marginal notes, fainted writing, stamps, skewed images, lines with different slope in the same page, slanted lines, inter-line sentences, etc. It is also difficult from the HTR point of view because it is written by several hands, it has crossed out words, hyphenated words, etc. Preliminary HTR results on a small set of 53 pages the Bentham collection were reported in [8] by using the HTR techniques mentioned in Section 2. The Word Error Rate (WER) obtained in that set was about 34% (see [8] for additional details).

Two collections have been chosen for Spanish. The first collection is a series of marriage register books [11]. Most of these books are handwritten documents, with a structure analogous to an accounting book. Each page is divided horizontally into three blocks, the husband surname's block, the main block, and the fee block and vertically into individual license records (see [11] for a more detailed description). The WER obtained with 200 pages from the same book from the 17th century was about 10% [12]. The second Spanish collection is a collection from the 17th century that has 7 books with about one thousand pages each. This collection is about Botanics. A main problem on this collection is the large vocabulary it has. The WER on an preliminary experiment with a small set of 38 pages was about 50% (see [2] for additional details).

For Dutch, a collection from the 15th century has been selected. The following section describes this collection and Section 5 describes experiments on this collection.

4. THE HATTEM COLLECTION

4.1 Manuscript

The Dutch data selected for tranScriptorium consists of a collection of late medieval manuscripts belonging to the "Artes Liberales" literature, which were mostly written in the cursive style. Most texts belong to the 15th century and belong to the medical domain. These documents are of great interest for the history of science, language and literature alike, and their transcription will be all the more valuable as only a relatively small number of documents of this type have been made available to the public. The greater part of this kind of literature still remains to be edited. Furthermore, the fact that they belong to a single subject domain will ease the task of constructing a vocabulary. At this point, we are considering a collection of about 4000 pages.

One of the selected text manuscripts is the *C5 Hattem Manuscript*[3] (15th century, 572 leaves), which has been completely transcribed by the WEMAL group[4], which makes it very suitable for experimentation. Apart from a small number of Latin and French documents, most of the texts are in Middle Dutch. The contents are very heterogeneous. There is a prose translation of the Secretum Secretorum (a Latin translation of an Arabic encyclopedia on government, health, astrology and alchemy), and a Dutch treatise on the plague, which is ascribed to the pope. The subject matter of

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

[3] Utrecht university library, MV : C5,
`http://objects.library.uu.nl/reader/index.php?obj=1874-44915\&lan=en`

[4] `http://wemal.let.uu.nl/hattem-c5.html`

the Dutch language treatises includes phlebotomy, surgery and uroscopy. There is an extensive treatment of herbalism, various recipes for medicinal potions, oils and unctions, alchemist procedures, and a specialised treatise on precious metal alloys.

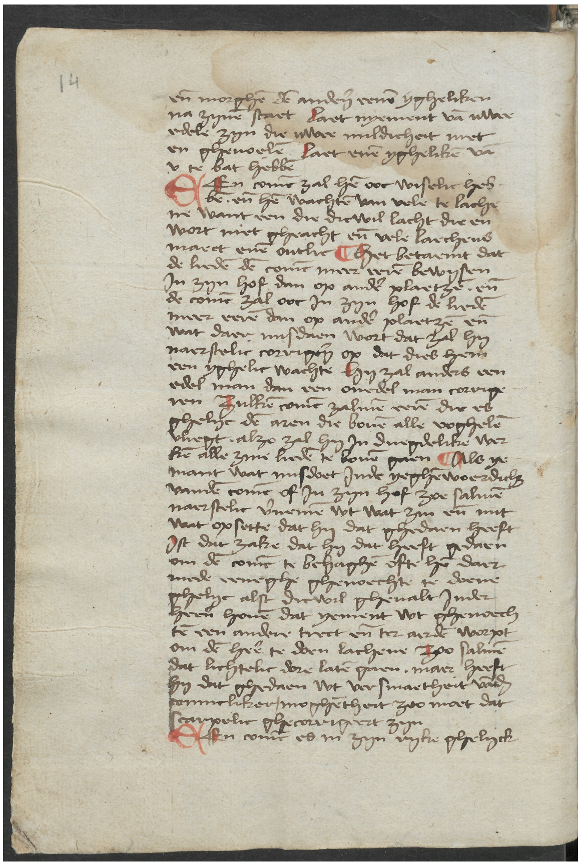


Figure 1: Sample image of the Hattem collection.

4.2 Language Modelling Issues for Dutch Medieval Manuscripts

Language Modelling (LM) is an integral part of the TRANSCRIPTORIUM holistic HTR engine. In order to arrive at a LM component which can effectively handle the challenges posed by historical language, special resources and approaches are necessary. Accordingly, we devoted a special work package in the project to the development of linguistic resources to support the HTR system, and the study of the effect of their application. In general, historical documents from some specialised domain pose problems for LM in two ways:

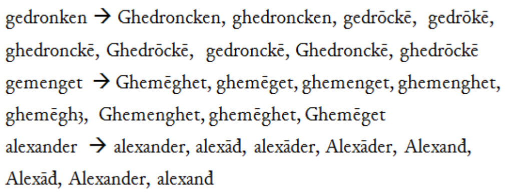
- It is usually difficult to get hold of more than a very limited amount of corpus material that is relevant for the type of language that needs to be modelled.
- (Historical) spelling variation exacerbates this problem: with a limited training corpus, even more words and word combinations will be unknown.

4.2.1 Spelling variation

Spelling variation in middle Dutch is extremely prolific. For instance, some possible realisations of the Dutch word "pelgrim" (pilgrim) are listed below:

pelegrime, pelgrime, peelgrime, peilgrime, pilgrime, pellegrime, peregrime, pelegrom, pelgrom, peelgrom, peilgrom, pilgrom, pellegrom, peregrom, pelegrum, pelgrum, peelgrum, peilgrum, pilgrum, pellegrum, peregrum, pelegrim, pelgrim, peelgrim, peilgrim, pilgrim, pellegrim, peregrim, pelegrijn, pelgrijn, peelgrijn, peilgrijn, pilgrijn, pellegrijn, peregrijn, pelegrijm, pelgrijm, peelgrijm, peilgrijm, pilgrijm, pellegrijm, peregrijm, pilgerijm, pelgerijn, pilgerijn, pillegram, pilgram, pylgram, pylgrym

Of course, within a manuscript, variation is less extreme, but one still finds may variants (see Fig. 2).



gedronken → Ghedroncken, ghedroncken, gedrōckē, gedrōkē, ghedronckē, Ghedrōckē, gedronckē, Ghedronckē, ghedrōckē

gemenget → Ghemēghet, ghemēget, ghemenget, ghemenghet, ghemēghʒ, Ghemenghet, ghemēghet, Ghemēget

alexander → alexander, alexāđ, alexāder, Alexāder, Alexanđ, Alexāđ, Alexander, alexanđ

Figure 2: Spelling variation within the manuscript

4.2.2 Abbreviations and ligatures

Medieval manuscripts contain a lot of abbreviations, which were indicated by special symbols no longer in use. According to current practice, even so-called "diplomatic" (literal) transcriptions of manuscripts and books mostly transcribe the hypothesised expansion of an abbreviation, but not the form of abbreviation used. This could be considered a problem if we are to use this material as ground truth for HTR (Fig. 3).

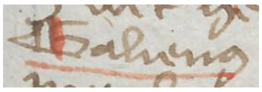


Figure 3: Example use of an abbreviation symbol

Accordingly, a typical TEI transcript of the image in Fig. 3 would be `Galien<expan>us</expan>`, which does not give us the form actually written, which is `Galien&#xF1A6;` according to the MUFI recommendations[5].

An important advantage of the segmentation-free approach to HTR is that we are not restricted to a one-to-one mapping of characters in the recognised or ground-truth text to locations in the manuscript page. Accordingly, there are two ways of training and testing the HTR system for this type of data:

1. Both ground truth and recognition results transcribe the text in a "hyperdiplomatic" way, using the symbols found on the manuscript page. The advantage of this is that the text is transcribed exactly as found on the page. The disadvantage is that language models and lexical data need to account for the spelling variants resulting from the use of abbreviations.

2. Both ground truth and recognition results transcribe the text using the hypothesised expansions. The advantage is that the transcribed text is easier to use in retrieval, and LM training material which is not in "hyperdiplomatic" form can be used without problems

[5] `http://www.mufi.org`

To be able to investigate both possibilities, we have adopted a TEI-conformant ground truth transcription scheme where both abbreviation and expansion are transcribed. Note that the current study follows the first scenario.

4.2.3 Hyphenated words

Typically, the TRANSCRIPTORIUM historical data contains many hyphenations, which break words into two parts, as well as instances of words which are split without any explicit graphical indication (see Fig. 4).

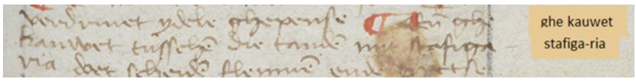


Figure 4: Two consecutive lines ending with incomplete words. The second case (stafiga-ria) has an explicit hyphen, the first case (ghe|kauwet) is not marked in any way.

Currently, the HTR system is still line-based, so the LM in the TRANSCRIPTORIUM HTR system cannot take hyphenated words as single words. If the parts of a hyphenated word have not been encountered in the training set for the language model, they lead to unknown words, even if the single word is in the vocabulary. Furthermore, there is no sharing of probability information between hyphenated and non-hyphenated occurrences of words. These issues will be tackled in the near future.

5. EXPERIMENTS

This section describes the HTR experiments that were carried out with the Hattem database described in Section 4. For the HTR experiments 40 pages were selected from the complete collection. The selected pages are not consecutive in the book but they are representative of the page images that appear along all the book. Figure 1 shows an example of the selected images. The comments in this section are referred to the selected pages.

5.1 Experimental Setup

From an image pre-processing point of view, the selected images do not have too much bleed-through, but some pages have smear zones. There are some drop capitals in red colour, and some special marks also in red that denote paragraphs. The skew in the pages is very small; some lines have a small slope degree. Some pages are slightly warped due to the scanning process. In some pages the ink colour is slightly pale. But no correction on these aspects was carried out for the experiments in this section.

From the layout point of view, the document images have a single column. The lines are very close each other (see Fig. 6). There are many occurrences of touching ascenders and descenders in consecutive lines. In this way the line detection and the line extraction seem very difficult. Some lines start with a drop capital. In general, there are few short lines (in words) as Fig. 5 shows.

For the experiments reported in this section, the line detection was carried out manually. A baseline was manually annotated as a polyline for each line. It is important to remark that this information can be obtained with little effort, but automatic methods could also be used [3]. The drop capitals were ignored. Then, polygon around this baseline was defined by selecting a fixed number of pixels above and below the manually annotated polyline. Figure 6 shows some lines in detail, with the manually annotated baseline and the polygon obtained automatically. Overlap between polygons was not taken into account in the experiments and they were allowed.

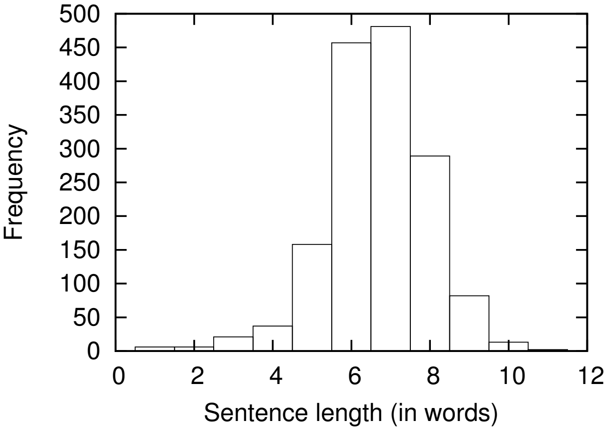


Figure 5: Histogram of the sentence lengths (in words).

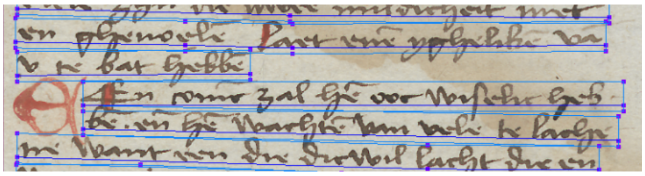


Figure 6: Sample images of the Hattem collection with detail of the baselines (polylines) and polygons obtained from these baselines for several lines.

From the HTR point of view, this is a single writer collection. Another characteristic is that there are no punctuation symbols and the writing is very clean, without crossed-out words. Paragraphs are marked with a special symbol (in red) but it was considered as a usual word (see red symbol in Fig. 4).

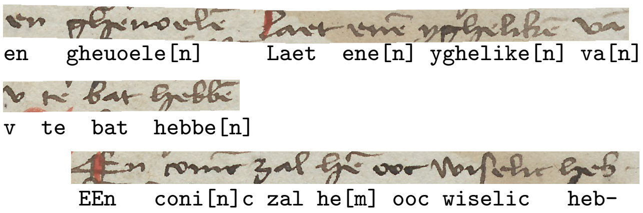
As discussed in Section 4.2.2, the manuscript has many abbreviations. For this experiment, the ground truth transcription refer to the diplomatic transcript of the abbreviations, that is, ignoring expansions. The polygon surrounding each line was used in order to obtain the corresponding line images. Figure 7 shows some lines and the corresponding transcript. The pair line image-line transcript (removing expansions) is the input for training the system.

An important characteristic of these documents is the large number of hyphened words. The total number of hyphened words was 532 in total, which means that 34% of the lines had a hyphened word at the end of the sentence (and consequently 34% of the lines had a suffix at the beginning of the sentence). The hyphened words that were explicitly registered in the line images (with a dash line at the end) was 78% of the line images and without any mark the rest 22% of the line images. Last column in Table 1 summarises the basic statistics of these pages.

The 40 pages were divided into 8 blocks of 5 pages each,

Table 1: Basic statistics of the different partitions in the selected 40 pages of the Hattem dataset.

Number of:	P0	P1	P2	P3	P4	P5	P6	P7	Total
Pages	5	5	5	5	5	5	5	5	40
Lines	186	200	195	185	191	191	203	201	1,552
Run. words	1,280	1,351	1,318	1,227	1,237	1,230	1,344	1,343	10,330
Run. OOV	244	276	264	261	250	285	306	248	-
Lex. OOV	219	241	228	234	210	221	264	225	-
Lexicon	551	575	562	581	554	544	611	570	2,602
Character set size	45	43	44	46	43	52	44	45	60
Run. Characters	5,148	5,506	5,446	5,063	5,242	5,006	5,656	5,645	42,712



`en gheuoele[n] Laet ene[n] yghelike[n] va[n]`

`v te bat hebbe[n]`

`EEn coni[n]c zal he[m] ooc wiselic heb-`

Figure 7: Sample line images of the Hattem collection with their corresponding transcripts. The diplomatic word can obtained by removing the square brackets and their content.

aimed at performing cross-validation experiments. That is, we carried out 8 rounds, with each of the partitions used once as test data and the remaining seven partitions used as training data. Table 1 contains some basic statistics of the different partitions defined. The number of running words for each partition that did not appear in the other seven partitions is shown in the running out-of-vocabulary (Run. OOV) row. The percentage of running OOV is about 20%. The pre-processing techniques that were applied and features used for representing the line images were the same that are described in [15].

5.2 HTR Results

Two different groups of experiments were carried out, one with open vocabulary (OV) and one with closed vocabulary (CV). In the OV experiments, only the words that appeared in the training set were included in the language model (LM). The words that appeared in the test set and not in the training set were Out-of-Vocabulary (OOV) running words. Note that these words in the test set were sure errors in the recognition step. In the CV experiments, the words that did not appear in the test were included in the LM as an additional lexicon. Note that the CV experiment is a very optimistic evaluation but allowed us to study the influence of the availability of a lexicon for a given task: it gives a lower bound for the error rate that could be obtained by the availability of a better lexicon. These two lexicon settings represent extreme cases with respect to real use of HTR systems. For evaluation we used the Word Error Rate (WER) and the Character Error Rate (CER).

Table 2 shows the best results obtained both in the OV and CV experiments. The values of the parameters that gave the best results were with 8 states per HMM, 64 Gaussian Densities (NG) per each HMM state, 90 of grammar scale factor (GSF) and $= -130$ of word insertion penalty (WIP). The language model in all experiments was a bi-gram. The morphological character models were trained by using pairs of line image-line transcript. HTK[6] was used for HMM training and SRILM was used for LM training[7].

Table 2: Best WER(%) and CER(%) results on Hattem dataset, for HMMs with 8 states and for different number of Gaussian Densities per each HMM state (NG) and optimised parameter values of grammar scale factor (GSF) and word insertion penalty (WIP).

			CV		OV	
NG	GSF	WIP	WER(%)	CER(%)	WER(%)	CER(%)
32	70	-130	25.3	13.7	43.0	21.6
64	90	-130	24.7	13.2	42.0	21.1

Figure 8 shows the normalised histogram of the length of words that took part in the error events for the experiment in which the NG was 64, the GSF was 90, the WIP was -130 and with OV experiments. The normalised histogram of the length of words in all the corpus is represented with dashed lines. For the errors, only substitutions and deletions from the correct transcript were considered. Note that most of the errors were concentrated in short words.

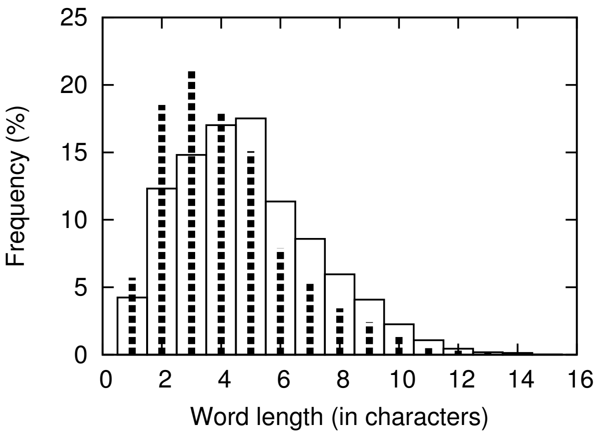


Figure 8: Normalised histogram of the length of words that took part in error events in Hattem dataset, and normalised histogram of the length of words in all the corpus that is represented with dashed lines.

Figure 9 shows in boxes the normalised histogram of the positions of the errors in the correct line transcript. We used

[6] `http://htk.eng.cam.ac.uk`

[7] `http://www-speech.sri.com/projects/srilm/`

the lines with more than three words in the reference line transcript for computing this histogram. Note that most of the errors were concentrated in the initial and last parts of the lines.

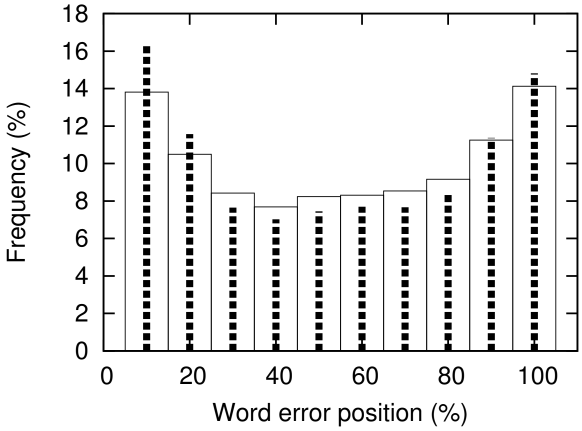


Figure 9: Normalised histogram of the positions in the lines in which the errors are located in the CV experiment.

5.3 Discussion

An important aspect to remark in these experiments is the obtained results because of several reasons. First, the required transcription resources for obtaining good results was really few. Note that only 35 pages were necessary for obtaining a WER below 42% in an OV experiment and a WER below 25% in an CV experiment taking into account the difficulty of the task. Although these results could seem poor, the CER reveals that the obtained transcripts are readable. Table 3 shows the transcripts obtained in the OV experiment and in the CV experiment of the three samples of Fig. 7. In these examples, word "gheuoele" in first sentence was not observed in training, however the system provided a similar word (not semantically but morphologically) that was in the training. Something similar happened with word "wiselic" in the third sentence.

Second, it is important also to remark that other existing sophisticated pre-processing techniques and/or line extraction were not used in these experiments, like removing descenders or ascenders from the lines that or above or below a given sentence [10].

Another aspect to remark in these experiments was the main reason of the errors. Note that a large percentage of errors were due to the OOV words as the comparison between the OV experiment and the CV experiment reveals. With a good lexicon, the WER could be substantially reduced, from 42.0% to 24.7% (at best) in the CV experiments.

Finally, but not less important in this collection is that the hyphenated words have also large influence in the errors. Figure 9 shows that errors are concentrated at the beginning and at the end of the sentences, where the hyphened words are located. Figure 9 also shows with dashed lines this behaviour for the OV experiment. Note that in this case the relative error at the beginning of the line increases 2 points due to the absence of the context provided by the LM.

Table 3: Examples of transcript results. Sentence Ref. is the reference while OV and CV are the transcripts obtained in the OV experiment and CV experiment, respectively.

Ref.:	en gheuoele Laet ene yghelike va
OV:	en ghemeen laet ene ghelike va
CV:	en gheuoele laet ene yghelike va
Ref.:	v te bat hebbe
OV:	va te bat hebbe
CV:	va te bet hebbe
Ref.:	EEn conic zal he ooc wiselic heb-
OV:	En come zal he ooc wisene heb-
CV:	En come zal he ooc wiselic heb-

6. CONCLUSIONS

Transcription of historical document employing modern holistic HTR technology is one of the main goals in tranScriptorium. This paper presented recent results on a Dutch collection. It is important to remark that good results were obtained by using minimal resources to develop the transcription system. Therefore, we are firmly convinced that this is the HTR technology to be used historical handwritten documents.

For future work we intend to develop efficient language modelling techniques for dealing with several problems mentioned in previous sections, like hyphened words, abbreviations, and development of specific vocabularies for historical documents.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600707 - tranScriptorium and the Spanish MEC under the STraDa (TIN2012-37475-C02-01) research project.

8. REFERENCES

[1] I. Bazzi, R. Schwartz, and J. Makhoul. An Omnifont Open-Vocabulary OCR System for English and Arabic. *IEEE Trans. on PAMI*, 21(6):495–504, 1999.

[2] V Bosch, Isabel Bordes-Cabrera, P. Cuenca Muñoz, C. Hernández-Tornero, L.A. Leiva, M. Pastor V. Romero, A.H. Toselli, and E. Vidal. Transcribing a XVII manuscript from scratch using computer-assisted transcription technology. In *(submitted to this conference)*, Madrid, Spain, 2014.

[3] V. Bosch, A.H. Toselli, and E. Vidal. Statistical text line analysis in handwritten documents. In *ICFHR*, pages 201–206, Bari, Italy, 2012.

[4] H. Bunke. Recognition of cursive roman handwriting-past, present and future. In *ICDAR*, page 448, Washington, DC, USA, 2003. IEEE Computer Society.

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

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, ser. B. 39:1–38, 1977.

[7] S. España-Boquera, M.J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez. Improving offline handwriting text recognition with hybrid HMM/ANN models. *IEEE Trans. on PAMI*, 33(4):767–779, 2011.

[8] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J.A. Sánchez, A.H. Toselli, and E. Vidal. Ground-truth production in the tranScriptorium project. In *DAS*, Tours, France, April 2014. (accepted).

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

[10] U.-V. Marti and H. Bunke. Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. *IJPRAI*, 15(1):65–90, 2001.

[11] V. Romero, A. Fornés, N. Serrano, J.A. Sánchez, A.H. Toselli, V. Frinken, E. Vidal, and J. Lladós. The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6):1658–1669, 2013.

[12] V. Romero and J.A. Sánchez. Category-based language models for handwriting recognition of marriage license books. In *ICDAR*, pages 788–792, 2013.

[13] V. Romero, A.H. Toselli, and E. Vidal. *Multimodal Interactive Handwritten Text Transcription.* Series in Machine Perception and Artificial Intelligence (MPAI). World Scientific Publishing, 2012.

[14] J.A. 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*, pages 227–228, 2013.

[15] 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*, 18(4):519–539, June 2004.

[16] A.H. Toselli, A. Juan, and E. Vidal. Spontaneous Handwriting Recognition and Classification. In *ICPR*, volume 1, pages 433–436, August 2004.