Handwritten Text Recognition for the EDT Project. Part I: Model Training and Automatic Transcription*

Joan Andreu Sánchez and Enrique Vidal

PRHLT, Uinversitat Politècnica de València (UPV) *and* tranSkriptorium AI S.L. (tS)

(jandreu,evidal)@transkriptorium.com

Abstract

Many massive handwritten text document collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried behind thousands of terabytes of high-resolution images. If perfect or sufficiently accurate text image transcripts were available, image textual content could be straightforwardly indexed for plain-text textual access using conventional information retrieval systems. But fully automatic transcription results generally lack the level of accuracy needed for reliable text indexing and search purposes. And manual or even computer-assisted transcription is entirely prohibitive to deal with the massive image collections which are typically considered for indexing. This paper explains how very accurate indexing and search can be directly implemented on the images themselves, without explicitly resorting to image transcripts. Results obtained using the proposed techniques on several relevant historical data sets are presented, which clearly support the high interest of these technologies.

1 Introduction

In recent years, massive quantities of historical handwritten documents are being scanned into digital images which are then made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive text *image* collections are available through Internet. The interest of these efforts not withstanding, unfortunately these document images are largely useless for their primary purpose; namely, exploiting the wealth of information conveyed by the text captured in the document images. Therefore, there is a fast growing interest in automated methods which allow the users to search for the relevant textual information contained in these images which is required for their needs.

In this sense, the project "European Digital Treasures (EDT): Management of centennial archives in the 21st century" [1] aims at bringing joint European heritage, especially its digital versions, major visibility, outreach and use. Three main goals are defined in the EDT project:

- To conceptualize and generate new business models that seek the profitability and economic sustainability of the digitized heritage of archives.
- To foster the development of new audiences especially focused on two groups: the young and the elderly - the latter so-called "golden-agers" or the "silver generation" made up of retirees and citizens aged 60+.
- To promote the transnational mobility of managers, historians, experts, graphic artists, industrial designers and archivists, working on the production of new technological products and interactive exhibitions that give support and visibility to three major European cultural areas.

*The second part of this publication deals with textual information search in untranscribed manuscripts and will appear in (Vidal and Sánchez, 2021).

[1] https://www.digitaltreasures.eu

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image textual content could be straightforwardly indexed for plain-text textual access. However, OCR technology is completely useless for typical handwritten text images; and fully automatic transcription results obtained using state-of-the art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes(Graves et al., 2009; Romero et al., 2012*a*; Vinciarelli et al., 2004).

An alternative to fully automatic processing is to rely on *computer-assisted* transcription. This was successfully explored empirically in (Alabau et al., 2014; Romero et al., 2012*b*; Toselli et al., 2010), following new, powerful concepts of pattern recognition-based human-machine interaction introduced in (Vidal et al., 2007) and (Toselli et al., 2011). In the last eight years, the TRANSCRIPTORIUM and READ projects[2], have further explored the capabilities of these automatic and interactive HTR (IHTR) technologies to speed-up the conversion of raw text images into electronic text.

Working conclusions from all these studies are as follows:

a) To some extent, fully automatic transcripts of text images can be useful for plain-text indexing and search purposes. However, in many historical text image collections of interest, the typical level of transcription accuracy achieved severely hinders the search *recall*; i.e., the system ability to ensure that all or most of the images where a given query text appears can actually be retrieved.

b) Similarly, the fully automatic transcription of most historical text images do not reach the level of accuracy needed for typical scholarly editions of the corresponding image collections.

c) In both cases, the required level of accuracy can obviously be obtained by means of additional user effort. If human work is to be done, rather than just letting the users to edit the noisy automatic transcripts, IHTR can be used to cost-effectively provide the desired transcription accuracy.

e) IHTR can lead to significant gains in human effort with respect to just manually editing the automatic transcripts. But the overall human effort demanded by IHTR is still substantial. Therefore, while IHTR is proving useful to produce scholarly editions of moderately sized historical collections, the required effort to deal with the kind of massive image collections, which are the typical target of indexing and search, is by all means entirely prohibitive.

Given the current situation of the HTR technology that have previously been described, in the last decade the Probabilistic Indexing (PrIx) technique (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Toselli et al., 2019; Vidal et al., 2020) has emerged as a solid technique for making the searching in document images a reality. This technique provides a nice *trade-off* between the *recall* and the *precision* that allows the user to locate most of the relevant information that s/he is looking for in large image collections. The techonolgy is introduced in (Vidal and Sánchez, 2021), mainly focused in the EDT collections.

The approaches proposed here are training-based and therefore need some amount (tens to hundreds) of manually transcribed images to train the required optical and language models. In addition they may benefit from the availability of collection-dependent lexicon and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources will be more than rewarded by the benefits of accurately making the textual contents of these collections available for exploration and retrieval.

The proposed HTR and PrIx approaches have been tested in the past for many historical collections of handwritten text images. Most of the early work was carried out within the TRANSCRIPTORIUM and READ projects mentioned above. The results of these experiments can be seen in a number of recent publications.[3]. Here we will present new experiments carried out with manuscript collections researched in the EDT project.

[2] `http://transcriptorium.eu`, `http://read.transkribus.eu`

[3] See: (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Sánchez et al., 2019; Toselli et al., 2019; Vidal et al., 2020)

2 Preparation of a HTR System

Off-line automatic Handwritten Text Recognition (HTR) is a challenging problem that requires a careful combination of several advanced Pattern Recognition techniques, including but not limited to Image Processing, Document Image Analysis, Feature Extraction, Neural Network approaches and Language Modeling.

HTR has progressed enormously in the last two decades due mainly to two reasons: first, the use of holistic training and recognition concepts and techniques which were previously developed in the field of Automatic Speech Recognition (ASR); and second, the existence of an increasing number of publicly available datasets for training and testing the HTR systems.

The need for holistic techniques in HTR has been known for many years given that the processes of handwriting and speech share many similar properties and challenges (Bazzi et al., 1999; Graves et al., 2009; Toselli et al., 2004): i) in both cases the production process is sequential through time; ii) the resulting images or signals are often largely distorted and severely contaminated with different kinds of noise; iii) due to the sequential production process, it is not possible in general to accurately recognize isolated words or characters/phonemes because none of these units can be reliably and consistently segmented or isolated; and iv) handwriting images and speech signals typically exhibit similar forms of lexical and syntactical regularity and ambiguity. Because of these similarities it is not surprising that the same basic Pattern Recognition techniques which had proved successful in ASR also become successful in HTR. To name a few: hidden Markov models (HMM) and recurrent neural networks (RNN) for optical character/phoneme modeling and statistical N-gram models for language modeling. These models are trained both in ASR and HTR with identical machine learning techniques based on the use of annotated data. The availability of sufficiently large amounts of annotated data is currently one of the bottlenecks to move forward in HTR since the annotation is generally performed by human experts and is, therefore, expensive and time-consuming.

The most traditional approaches to HTR are based on N-gram language models (LM) and optical modeling of characters by means of HMMs with Gaussian mixture emission distributions (HMM-GMM) (Marti and Bunke, 2001). In the last decade, notable improvements in HTR accuracy have been achieved by using RNNs for optical modeling. As of now, the state-of-the-art optical modeling HTR technology is based on deeply layered neural network models (Bluche, 2015; Bluche et al., 2015; Graves et al., 2009). The overall architecture is often referred to as *Convolutional-Recurrent Neural Networks* (CRNN) (Shi et al., 2015).

Trained N-grams are represented as a stochastic finite-state transducer. The stochastic transducer, along with the classical Viterbi decoding algorithm (also known as "token- or message-passing"), are used to obtain an optimal transcription hypothesis of the original input line image.

The training process of the optical models is performed both with line images and their corresponding transcripts that need to be in correspondence. Consequently both layout and transcripts have to be prepared to train the HTR system. Layout annotation is related to the mark-up of relevant regions (text regions, marginalia, images, plots, etc.) and the mark-up of (base-)lines located in these regions. In this paper we assume that this process is performed only to get the baselines. Related to the transcript of the lines, they are are referred as "ground-truth" (GT) and they are usally prepared by experts. This GT preparation process is very crucial and relevant decissions have to be made that can affect further steps. Some of the decissions to be made are:

- To use or not to use modernized transcripts.
- To use or not to use all transcripts in capital letters or to use mixed case.
- To expand or not to expand abbreviated words.
- To use or not to use semantic tags.
- To use or not to use diacritics.
- To annotate or not to annotate hyphenated words.
- How to deal with dates and other figures (ages, temperatures, etc).
- To distiguish or not to distiguish between printed and handwritten characters.

The decissions on these points depends on the final goal of the HTR process and can affect the final results. The GT preparation can be more or less expensive depending on the made decissions. For example, the easier

and less time consuming GT production could be to transcribe line images by using modernized transcripts, in capital letters, with all abbreviation expanded, without semantic tags, without diacritics, without annotating hyphenated words, with all dates in the same format. It is important to remark that the more richer transcripts are the more GT data is necessary and, consequently, more expensive.

In the case of the EDT project, some collections were annotated with tags and other not. The purpose of these tags was to make easier to locate words according to their semantics. We describe in more detail the annotation process for the Spain collection although other collections were annotated in a similar way.

The Spain dataset that was used to prepare the HTR system was annotated with several tags. Each transcript was processed as Figure 1 shows. The HTR model is able to learn aligning each character in the image with a character in the transcript. This figure shows that some characters (<print>, <age>) do not have a visual representation, but the optical model is able to capture some contextual information from the surrounding text. The LM also helps in the recognition process of the tags. Note also that the image contains the abbreviated forms "sol" for "soltera" (femenine form for single) and "lab" for "labores" (housekeeper).

De<print> 36<age> años<print> Estado<print> soltera<civilstate> Profesión<print> labores<job>

Figure 1: Example of a text line used for training. Both the line image (top) and the corresponding tagged transcript (bottom) are needed.

This tagging scheme lets the system learn to distinguish identical words that are used with different semantic meaning; for instance, "Juan<surname>", "Juan<name>", "Juan<place>", "Juan<residence>".

It is important to remark that the transcripts prepared for the GT are both used to train the optical model and the language model. With enough context, both the optical model and the language model are able to take into account the tags and they can be hypothesised in the recognition process.

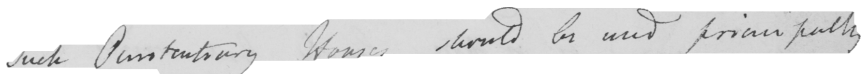
The HTR trained system is used to obtain the most probable transcript by using the Viterbi algorithm. As previously commented, the optical model and the language model are combined in a stochastic finite-state transducer. The best hypothesis uses to lack the level of accuracy needed for reliable text indexing and search purposes. An alternative solution is to obtain the N-best hypothesis transcriptions associated to a line image and to collapse them into a Word Graph (or Character Graph). When N is sufficiently large the Word Graph (WG) associated to a line is able to generalize and to include alternative solutions that were not included in the list on N best hypotheses (Toselli et al., 2016). These WG have recently used to obtain word distributions for each page image rather than just one hypotheses per line image. This idea is explained in (Vidal and Sánchez, 2021).

3 Evaluation Metrics

The most usual evaluation metrics for measuring the performance of an HTR system are the Word Error Rate (WER) and the Character Error Rate (CER). WER is defined as the minimum number of words that need to be substituted, deleted, or inserted to match the recognition output with the corresponding reference ground truth, divided by the total number of words in the reference transcripts. CER is defined in the same way but at character level. See examples in Figure 2.

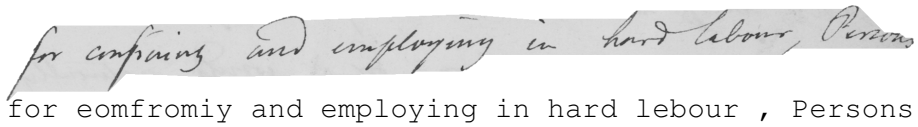
Generally speaking, WER is fairly well correlated with CER, but this correlation is not always strong or systematic. Therefore, *both* measures are important and complementary to assess the quality of an automatic transcript. A low CER but a relatively high WER reveals that the character errors are spread among many words. Conversely, a transcript with the same CER as before but lower WER indicates that errors are concentrated in few words. A good language model typically helps to achieve greater improvements in WER than in CER.

WER tends to be better than CER at indicating how difficult is to understand a transcript by human beings. Similarly, even if CER is low, a high WER may dramatically harm the performance of information extraction



such Penstentrary Hoases should be anid priapalty

WER = 4/7 = 57%
CER = 8/50 = 16%



for eomfromiy and employing in hard lebour , Persons

WER = 2/9 = 22%
CER = 8/52 = 15%

Figure 2: Two examples of test line images and *automatic* transcripts, along with the corresponding WER and CER. While CER is similar in both transcripts, that with higher WER may be harder to understand. *Reference* transcripts for the top and bottom line images are, respectively: "`such Penitentiary Houses should be and principally`" and "`for confining and employing in hard labour , Persons`".

or searching systems which rely on automatic transcripts. Figure 2 illustrates these facts for two samples which exhibit similar CER but different WER.

4 Datasets

Here we will present the datasets compiled in the EDT project for the experiments and results to be presented in Sec. 5. It is important to remark that collections that have been processed in the EDT project are composed by thousands of images but here we focus only in the sets that have been used for preparing the HTR systems.

EDT Hungary. This collection is composed of table images that contain Hungarian proper names in a left-most image region. Figure 3 shows examples of these images. This region is automatically detected by layout analysis techniques and only text lines in this region are detected and extracted. Significant difficulties of this collection include:

- Layout: these tables have a specific but fairly regular layout. They contain both printed and handwritten text, but most cells are typically empty. Only the proper names located in the left-most image region were interesting in the EDT project. Each proper names is preceeded by a handwritten number that had to be avoid in the line detection process to avoid recognition problems.
- Optical modeling: words contain many diacritics and the text has been written by very many hands. This leads to a very high writing style variability.
- Language modeling: the handwritten text are mainly proper names and most of the them are abbreviated in a not consistent way.

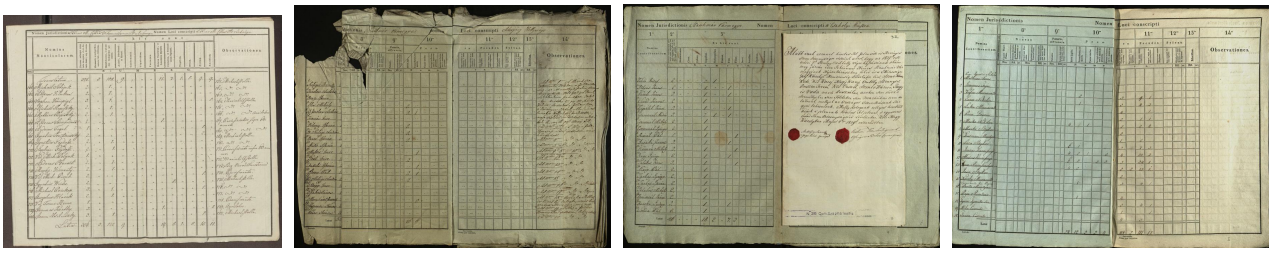


Figure 3: Examples of text images from the EDT Hungary collection.

For the ground-truth preparation, the left columns was annotated with a rectangular region. The main figures of the dataset that was used for preparing the HTR system are reported in Table 1. For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT

partitions. A list of proper names was also provided by the archive that was used to improve the training of the LM.

Table 1: Main values of the EDT Hungary dataset.

	Train	Validation	Test	Total GT
Images				410
Lines	7 000	800	672	8 472
Vocabulary	4 168	770	657	4 768
Character set				74
Running words	14 000	1 684	1 403	17 687
Running chars.	119 667	13 804	11 396	144 867

EDT Norway. This dataset is composed of images of register cards that contain mainly Norwegian proper names. Figure 4 shows examples of these images. Significant difficulties of this collection include:

- Layout: these cards have a fairly regular layout, but it is fairly complex. Both printed an handwritten text is considered for recognition. Each card contains record space for two persons, which appear in separate (left and right) image regions. However, handwritten information can be given for just one person or for both. Most card fields are typically empty and some stamps appear in many cards.
- Optical modeling: some special characters and diacritics are used and the cards have been filled by very many hands. This leads to very high writing style variability, and more so for the size of the initial capitals, which tends to be much larger than the main text body.
- Language modeling: the handwritten text contains many proper names and dates. Many date formats are used, but the archive wished to handle dates in a standard format. Only selected information items of the cards were interesting for the archive.

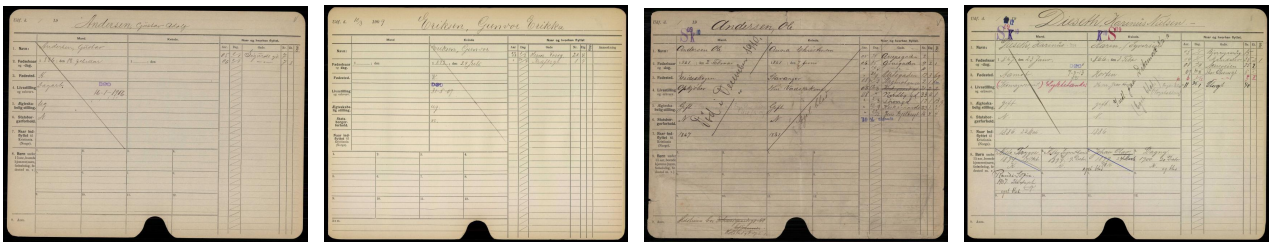


Figure 4: Examples of text images from the EDT Norway collection.

For the ground-truth preparation, the left column of each card was annotated with a rectangular region. Then, the space for two persons were also annotated with rectangular regions along with the proper name in the header. This made easier the training of a specific layout analysis system. This made easier the training of a specific layout analysis system. Lines were detected only inside these regions. The main figures of this dataset are reported in Table 2.

For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions. Printed and handwritten characters were modelled without distinction.

EDT Portugal. This dataset is composed of grayscale images with running text written in Portuguese. Figure 5 shows examples of these images. Significant difficulties of this collection include:

Table 2: Main values of the EDT Norway dataset.

	Train	Validation	Test	Total GT
Images				36
Lines	5 000	243	200	5 443
Vocabulary	1 588	180	131	1 676
Character set				74
Running words	12 069	581	328	13 100
Running chars.	65 405	3 195	2 479	71 079

- Layout: baseline detection is difficult because text lines generally exhibit a great amount of warping, along with very variable slopes and slant. Layout becomes often complex because of plenty marginalia and other more complex layout structures. Many images include parts of adjacent pages.
- Optical modeling: Text include many diacritics and has been written by several hands leading to a significant amount of writing style variability.
- Language modeling: The text contains many proper names and dates and a greet amount of abbreviations and hyphenated words.

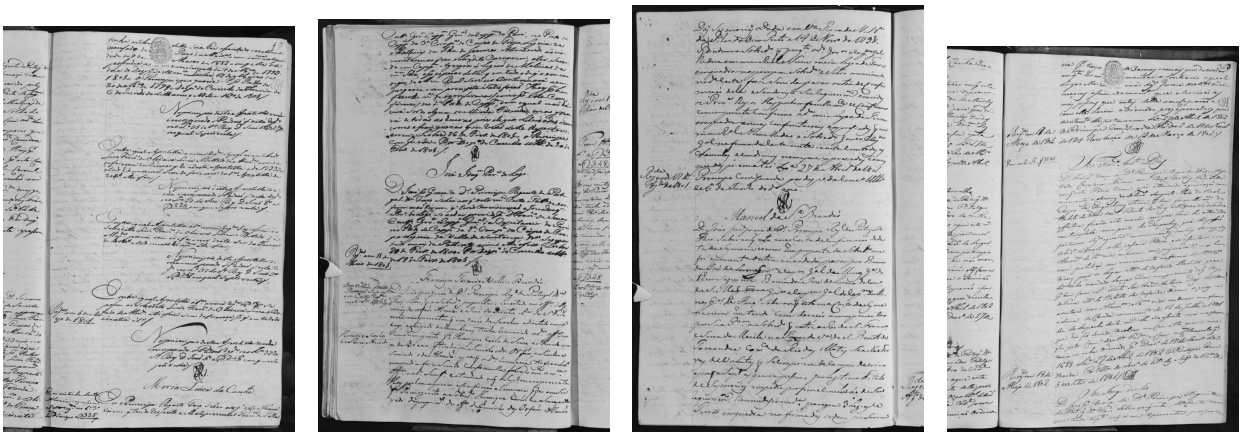


Figure 5: Examples of text images from the EDT Portugal collection.

The main figures of this dataset that are used for preparing the HTR system are reported in Table 3.

For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions. It is worth mentioning that the amount of handwritten text for training is reasonable, as it surpasses the desirable amount of at least 10K words.

Table 3: Main features of the EDT Portugal dataset.

	Train	Validation	Test	Total GT
Images				36
Lines	1 200	122	92	1 414
Vocabulary	2 034	526	424	2 281
Character set				78
Running words	11 338	1 131	816	13 285
Running chars.	62 785	6 359	4 685	73 829

EDT Spain. This dataset is composed of images that contain visa records of Spanish citizens for traveling worldwide. They were issued between years 1936 and 1939 in a Spanish consulate based in Buenos Aires (Argentina). Figure 6 shows examples of these images. Significant difficulties of this collection include:

- Layout: Each image contains four visa forms and each form includes one (or several) picture(s) associated to each visa. Some dates are stamped and others are handwritten. Text lines often exhibit extreme slope.
- Optical modeling: Forms include printed text and are filled with handwritten text written by many hands, leading to a very high writing style variability.
- Language modeling: the handwritten text contains many proper names (given names, surnames, town and state names, countries, etc.) and dates. A large amount of words are heavily abbreviated. All the textual information in each visa was relevant for the archive.

Figure 6: Examples of text images from the EDT Spain collection.

A small set of 99 images of the whole collection were selected for ground-truth annotation. For layout analysis, each image was annotated with four rectangular regions to isolate each visa, and then the geometric data of each photograph region and all the baselines were annotated. Then each line was manually transcribed and annotated word by word with "semantic" tags.

The tags that have been used in this dataset and their meanings are:

- `<print>`: printed word of the form
- `<date>`: date, both stamped or handwritten
- `<gname>`: given name
- `<surname>`: surname (two surnames are used in Spanish)
- `<state>`: province
- `<country>`: country
- `<civilstate>`: civil state (single, married, etc.)
- `<residence>`: place of residence
- `<place>`: location (city, village, etc.)
- `<job>`: occupation
- `<age>`: years old

The main values of this dataset are reported in Table 4. Since the amount of handwritten text and printed text can significantly affect the recognition results, this table provides the amount of both types of text. Experiments will be reported without taking into account this difference. Since the tags are considered special characters, the same word with two different tags was considered two different words. This fact explains the large vocabulary that can be observed in the "Hand" columns.

For the training, the lines were shuffled at collection level and therefore lines from the same pages may be included in the training and in the test partitions. It is worth mentioning that the amount of handwritten text for training is reasonable, as it surpasses the desirable amount of at least 10K words.

Table 4: Main features of the EDT Spanish dataset. The vocabulary is shown both with and without numbers

	Train		Validation		Test		Total GT	
	Print	Hand.	Print	Hand.	Print	Hand.	Print	Hand.
Images							99	
Lines	6 500		250		261		7 011	
Vocabulary w numbers	317	2 900	48	254	49	276	333	3 057
Vocabulary w/o numbers	53	2 040	36	203	37	202	53	2 149
Character set							85	
Running words	14 662	10 530	575	390	602	455	15 839	11 375
Running chars.	108 226	72 019	4 329	2 585	4 503	3 046	117 058	77 650

EDT Malta. This collection is composed of grayscale images with lists of proper names and the name of the flight or the ship in which the people arrived to Malta Figure 7 shows examples of these images. Significant difficulties of this collection include:

- Layout: the lines have quite slope, slant, skew and warping. Many images are physically degraded with many holes because of woodwarm and other insects. Many images include parts of adjacent pages.
- Optical modelling: there are fainted text and several hands, and many flights and ship names are replaced with quotes (“).
- Language modeling: the text contains mainly proper names and dates.

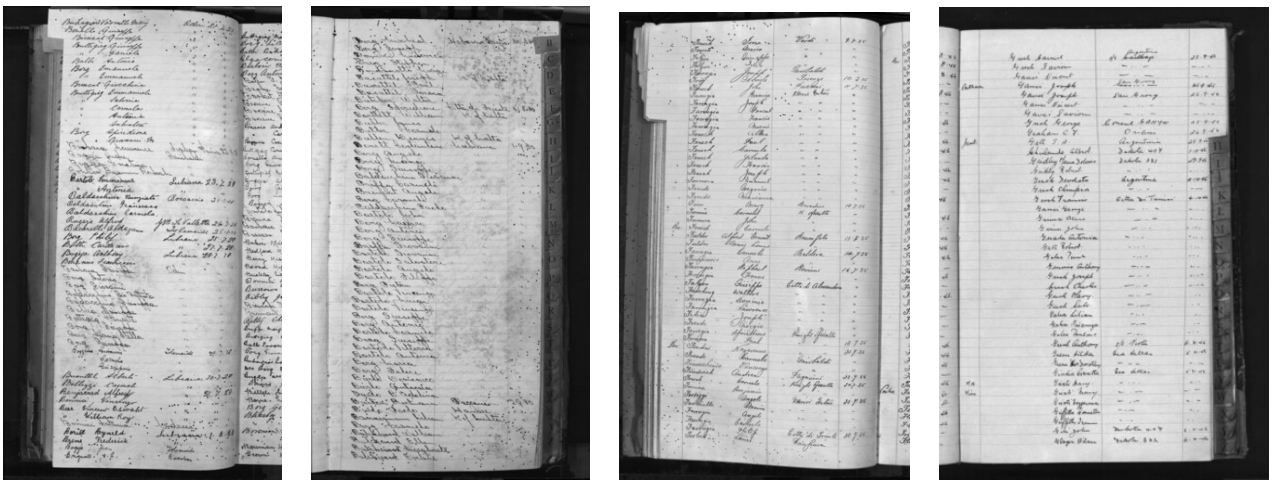


Figure 7: Examples of text images from the EDT Malta collection.

The main figures of this dataset used for preparing the HTR system are reported in Table 5. For training, the lines were shuffled at collection level and therefore lines from the same pages may be included in all GT partitions.

Table 5: Main features of the EDT Malta dataset.

	Train	Validation	Test	Total GT
Images				49
Lines	2 200	230	101	2 531
Vocabulary	3 249	551	274	3 614
Character set				77
Running words	9 996	1 037	448	11 481
Running chars.	65 376	6 829	2 971	75 176

5 Experiments and Results

The datasets introduced in Section 4 were used to prepare an HTR system for each collection. Theses HTR systems were evaluated using the partitions described for each collection. CER and WER were computed and the results are shown in Table 6. Several values of N were used for the N-gram LM but only the results obtained with $N = 5$ are shown since this was the best value or other values of $N > 5$ did not get a significant improvements.

Table 6: CER and WER obtained for each collections.

Dataset	CER	WER
Hungary	8.8	25.7
Norway	5.6	13.4
Portugal	14.5	35.4
Spain	9.4	20.8
Malta	12.8	36.6

We observe that both CER and WER have large difference among collections. The worst result are obtained for the Portugal and Malta collections. In the case of the Portugal collection, the bad WER results are due mainly to the large amount of abbreviated forms. This WER could be decreased by using additional GT. In the case of the Malta collection, the problem is again with the amount of training data, data shoul clearly increased.

6 Conclusion and outlook

We have introduced experimental results for the EDT collections. The obtained results reveal that additional GT data should be prepared for improving the training of the models. For future work, we expect to add additional training material produced in the EDT project with a crowdsourcing initiative.

7 Acknowledgments

References

Alabau, V., Martínez-Hinarejos, C., Romero, V. and Lagarda, A. (2014), 'An iterative multimodal framework for the transcription of handwritten historical documents', *Pattern Recognition Letters* **35**, 195–203. Frontiers in Handwriting Processing.

Bazzi, I., Schwartz, R. and Makhoul, J. (1999), 'An omnifont open-vocabulary OCR system for English and Arabic', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(6), 495–504.

Bluche, T. (2015), Deep Neural Networks for Large Vocabulary Handwritten Text Recognition, PhD thesis, Ecole Doctorale Informatique de Paris-Sud - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur. Discipline : Informatique.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H. and Vidal, E. (2017), Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project, *in* 'Int. Conf. on Document Analysis and Recognition (ICDAR)', Vol. 01, pp. 311–316.

Bluche, T., Ney, H. and Kermorvant, C. (2015), The LIMSI/A2iA Handwriting Recognition Systems for the HTRtS Contest, *in* 'International Conference on Document Analysis and Recognition', pp. 448–452.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009), 'A Novel Connectionist System for Unconstrained Handwriting Recognition', *IEEE Transaction on Pattern Analysis and Machine Intelligence* **31**(5), 855–868.

Lang, E., Puigcerver, J., Toselli, A. H. and Vidal, E. (2018), Probabilistic indexing and search for information extraction on handwritten german parish records, *in* '2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)', pp. 44–49.

Marti, U.-V. and Bunke, H. (2001), 'Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system', *International Journal of Pattern Recognition and Artificial Intelligence* **15**(1), 65–90.

Puigcerver, J. (2018), A Probabilistic Formulation of Keyword Spotting, PhD thesis, Univ. Politècnica de València.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2020), Advances in handwritten keyword indexing and search technologies, *in* A. Fischer, M. Liwicki and R. Ingold, eds, 'Handwritten Historical Document Analysis, Recognition, And Retrieval-State Of The Art And Future Trends', Vol. 89, World Scientific, pp. 175–193.

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

Romero, V., Toselli, A. and Vidal, E. (2012*b*), *Multimodal Interactive Handwritten Text Recognition*, Vol. 80 of *Machine Perception and Artificial Intelligence*, World Scientific.

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M. and Vidal, E. (2019), 'A set of benchmarks for handwritten text recognition on historical documents', *Pattern Recognition* **94**, 122–134.

Shi, B., Bai, X. and Yao, C. (2015), 'An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition', *CoRR* **abs/1507.05717**.

Toselli, A. H., Juan, A., Keysers, D., González, J., Salvador, I., Ney, H., Vidal, E. and Casacuberta, F. (2004), 'Integrated handwriting recognition and interpretation using finite-state models', *International Journal of Pattern Recognition and Artificial Intelligence* **18**(4), 519–539.

Toselli, A. H., Romero, V., Vidal, E. and Sánchez, J. A. (2019), Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, *in* '15th Int. Conf. on Document Analysis and Recognition (ICDAR)'.

Toselli, A. H., Vidal, E., Romero, V. and Frinken, V. (2016), 'HMM word graph based keyword spotting in handwritten document images', *Information Sciences* **370-371**, 497–518. Information Sciences 370-371 (2016) 497-518.

Toselli, A., Romero, V., i Gadea, M. P. and Vidal, E. (2010), 'Multimodal interactive transcription of text images', *Pattern Recognition* **43**(5), 1814–1825.

Toselli, A., Vidal, E. and Casacuberta, F. (2011), *Multimodal Interactive Pattern Recognition and Applications*, 1st edition edn, Springer.

Vidal, E., Rodríguez, L., Casacuberta, F. and García-Varea, I. (2007), Interactive pattern recognition, *in* 'International Workshop on Machine Learning for Multimodal Interaction', Springer, pp. 60–71.

Vidal, E., Romero, V., Toselli, A. H., Sánchez, J. A., Bosch, V., Quirós, L., Benedí, J. M., Prieto, J. R., Pastor, M., Casacuberta, F., Alonso, C., García, C., Márquez, L. and Orcero, C. (2020), The carabela project and manuscript collection: Large-scale probabilistic indexing and content-based classification, *in* '17th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)', pp. 85–90.

Vidal, E. and Sánchez, J. A. (2021), Handwritten text recognition for the EDT project. Part II: Textual information search in untranscribed manuscripts, *in* M. A. Bermejo et al., ed., 'Proc. of the EDT Alicante workshop', To appear.

Vinciarelli, A., Bengio, S. and Bunke, H. (2004), 'Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 709–720.