Granell, E.; Romero, V.; Martínez-Hinarejos, C. (2020). Study of the influence of lexicon and language restrictions on computer assisted transcription of historical manuscripts. Neurocomputing. 390:12-27. https://doi.org/10.1016/j.neucom.2020.01.081

Study of the Influence of Lexicon and Language Restrictions on Computer Assisted Transcription of Historical Manuscripts

Emilio Granell, Verónica Romero, Carlos-D. Martínez-Hinarejos*

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València,
Camino Vera s/n, 46022, València, Spain

Abstract

State-of-the-art Handwritten Text Recognition (HTR) systems allow transcribers to speed-up the transcription of handwritten text images. These systems provide transcribers an initial draft transcription that can be corrected with less effort than transcribing the handwritten text images from scratch. Currently, even the draft transcriptions offered by the most advanced HTR systems contain errors. Therefore, the supervision of this draft by a human transcriber is still necessary to obtain the correct transcription of the handwritten text images. This supervision can be eased by using interactive and assistive transcription systems, where the transcriber and the automatic system cooperate in the amending process.

In this paper, the draft transcription is provided by an HTR system based on Convolutional and Recurrent Neural Networks with Bidirectional Long-Short Term Memory units, and the assistive system is fed by lattices generated by using Weighted Finite State Transducers. The influence of the lexicon and language restrictions on the performance of our computer assisted transcription system is evaluated on three historical manuscripts.

The transcriptions offered by the proposed HTR system present very low error rates for the studied historical manuscripts. However, our assistive transcription system without lexicon or language restrictions is able to provide an additional reduction on the human effort required to correct the transcriptions in more than 50% over the transcriptions offered by the HTR system.

Keywords: Handwritten Text Recognition, Deep Learning, Interactive Transcription

*Corresponding author
Email address: cmartine@dsic.upv.es (Carlos-D. Martínez-Hinarejos)

1. Introduction

Transcription of handwritten documents has become an important research topic because it eases the textual access to the contents of manuscript documents. Transcription makes possible the textual search and it allows for various applications such as information retrieval or document classification [1]. In particular, the transcription of historical manuscripts is very useful for preserving their contents, which is crucial for cultural and historical reasons, and for providing access to data on cultural heritage [2].

The transcription of historical handwritten documents is usually performed by professional transcribers called paleographers, which are experts on ancient language and script. The transcription of historical handwritten documents is a hard and time consuming task. However, the Handwritten Text Recognition (HTR) technology has alleviated the transcription effort in the last years. In this context, paleographers must correct the hypothesis offered by the HTR system instead of transcribing from scratch. Additionally, the paleographer transcription effort can be even smaller by using computer assisted transcription systems [3], where the paleographer and the interactive system collaborate to obtain the correct transcript.

Currently, state-of-the-art HTR systems are based on deep learning [4]. These systems are usually composed of Convolutional and Recurrent Neural Networks (CRNN) [5]. Deep learning based HTR systems have shown to be considerably better than systems based on Hidden Markov Models with Gaussian Mixtures Models as output probability density function (HMM/GMM); for example, in a previous work for the Spanish historical manuscript Rodrigo [6] (see Section 5 for more information about this corpus) a CRNN system offered a transcript 64.8% better (in terms of word error rate) than the transcript provided by an HMM based system [7]. In spite of the better transcription quality they offer, those HTR systems are not perfect yet and the paleographer supervision is still required for obtaining good quality transcripts. Therefore, given that current HTR systems cannot replace paleographers, it is feasible that an interactive and assistive environment that uses the knowledge provided by the HTR system and the paleographer feedback would reduce the transcription work load even more. This can be possible even when the HTR system provides hypotheses with very low error rates.

HTR systems based on neural networks are usually trained at the character level. With this approach, they present some advantages over word-based systems, such as a reduced number of out-of-vocabulary units and a usually higher accuracy. As we will see in this paper, assisted transcription systems can take advantage of this potential to offer better hypotheses to the paleographers without the restrictions imposed by lexical or language models.

In this work, several contributions on the transcription of historical handwritten documents are presented. In the first term, the best result for the Rodrigo, the Cristo Salvador and the Bentham corpora reported in previous works [7, 8, 9] are improved. Additionally, results highlight the importance of assistive and interactive systems for transcribing historical manuscripts, be-

cause in spite of the fact that deep learning HTR systems provide quite accurate transcripts, interactive systems still manage to reduce the required human effort. Apart from that, in order to employ the capabilities of CRNN models to work without lexical or language constraints, the impact of using corrections at character level without language model is studied.

The rest of the paper is structured as follows: to improve the readership in the pattern recognition and neural networks community and to demonstrate the relevance of this work, recent related works are reviewed in the next section (Section 2); the proposed HTR system is detailed in Section 3; an overview of the assistive and interactive transcription system is presented in Section 4; the experimental framework is described in Section 5; the performed experiments and the obtained results are reported in Section 6; finally, the conclusions and future work lines are drawn in Section 7.

2. Related Work

Currently, the state of the art in pattern recognition systems is mainly marked by systems based on deep neural networks, which are capable of learning even the extraction of features from the original signals. This technology is known as deep learning [10, 11].

The deep learning approach is having a good reception in the scientific community because, among other reasons, it allows to develop end-to-end pattern recognition systems which can be applied to a multitude of scenarios: autonomous automotive [12, 13], action detection in videos [14], emotion recognition [15], people re-identification [16], and recognition of the human silhouette and pose [17]. However, perhaps the main reason is that deep learning has allowed to improve considerably the recognition accuracy compared with previous approaches, as it is the case for Handwritten Text Recognition (HTR) [7].

Regarding HTR, different approaches have been studied in the last years, such as multidimensional neural networks [18]. However, in a recent study [19] it was determined that best results could be achieved by HTR systems based on bidirectional neural networks at line level. This is the HTR approach followed in this work. On the other hand, HTR, as a natural language technology, has some limitations, such as those imposed by vocabulary constraints. One possible solution to overcome this limitation is the use of external textual resources to improve the linguistic models of the HTR system [20]. However, given the accuracy of HTR systems based on deep learning, it is usual to perform the recognition without lexical restrictions [21, 22].

One of the practical utilities of the HTR technology is the transcription of historical documents [23, 24, 4, 25]. This is a field of study that is currently booming for several reasons; firstly, there is a large number of historical manuscript documents that libraries and historical archives are digitalising; secondly, the historical or heritage information they contain is very important [26]; finally, the transcription of these archives and documents is needed in order to facilitate the access to its contents. The difficulty of automating the process is

an additional interesting issue. The importance of the preservation of historical manuscripts by using transcription led to the development of international projects, such as tranScriptorium[1] or READ[2].

Even the state-of-the-art HTR techniques based on deep learning [19] produce a considerable amount of errors in the transcription process of historical manuscripts, which makes necessary the supervision of the obtained results by a professional transcriber. Recently, several platforms have been developed in order to ease the transcription task (such as AnnoTate[3], Transcribe Bentham[4], or Transkribus[5]).

Human-computer interactive protocols for pattern recognition [27], together with the HTR technology, can be applied to the transcription task [28], causing the required human effort to be considerably reduced. These interactive systems depend on the technology of the HTR system on which they are based; this makes that the corrections are normally made at the word level [29, 30]. In this work, interactive transcription without lexical or language constraints is studied, which allows corrections at character level.

3. Handwritten Text Recognition

The goal of HTR systems is to find the most likely word sequence $\hat{w}$ given a feature vector sequence $x = (x_1, x_2, \ldots, x_{|x|})$ that represents a segmented handwritten text line image [23]. Usually, HTR systems are composed of three statistical models: optical, lexicon, and language. Therefore, the traditional HTR problem can be formulated as follows:

$$\begin{aligned} \hat{w} = \arg\max_{w \in W} \Pr(w \mid x) &= \arg\max_{w \in W} \Pr(x \mid w) \Pr(w) \\ &\approx \arg\max_{w \in W} \max_{c \in C_w} \Pr(x \mid c) \Pr(c \mid w) \Pr(w) \end{aligned} \tag{1}$$

where W denotes the set of available word sequences, C_w the set of different spellings of the word sequence w, $\Pr(x \mid c)$ is the probability of observing x by assuming that c is the underlying character sequence for x, which is modelled by the optical model, $\Pr(c \mid w)$ is the probability of the spelling of the word sequence w, which is modelled by the lexicon model, and $\Pr(w)$ is the probability of the word sequence $w = (w_1, w_2, \ldots, w_{|w|})$, which is modelled by a word language model.

In the lexicon model the spelling of each word w is modelled as a sequence of characters $c = (c_1, c_2, \ldots, c_{|c|})$. In this way, the lexicon model links the optical level representation with the word sequence output [31]. In this case,

[1] http://transcriptorium.eu/
[2] http://read.transkribus.eu/
[3] https://anno.tate.org.uk/
[4] http://blogs.ucl.ac.uk/transcribe-bentham/
[5] https://transkribus.eu/Transkribus/

the recognition is performed at word-level (with lexicon restrictions), because the HTR system is restricted to recognise only sequences of the words contained in the lexicon model.

Since words can be decomposed into sequences of characters, and usually characters are used as the basic linguistic units to train the optical models, the HTR can be performed at character-level, i.e. without lexicon restrictions. It would be equivalent to employ HTR at word level but using as lexicon model the list of characters. In this case, the HTR problem can be reformulated as finding the most likely character sequence $\hat{c}$ given x, and Equation (1) becomes:

$$\hat{c} = \arg\max_{c \in C} \Pr(c \mid x) = \arg\max_{c \in C} \Pr(x \mid c) \Pr(c) \tag{2}$$

where C represents the set of all permissible character sequences, and $\Pr(c)$ is the probability of the character sequence c modelled by a character language model.

Language models define all the possible sentences that can be recognised by the HTR system. In the language models, the text properties are modelled independently from the optical models [32] and they restrict the sequences of words or characters that can be recognised by the HTR system. However, the HTR decoding can also be performed without language restrictions. For instance, it can be done using zero-gram language models, where all the language elements are equiprobable; this means that $\Pr(w)$ in Equation (1) for decoding at word-level, and $\Pr(c)$ in Equation (2) for decoding at character-level, are constant and can be ignored in the maximisation.

The most traditional approaches to HTR approximate the language model by n-grams and the optical modelling of characters by means of Hidden Markov Models with Gaussian mixture emission distributions (HMM-GMM) [32]. However, in the last years, significant improvements have been achieved by using Recurrent Neural Networks (RNNs) for optical modelling. The current state-of-the-art HTR technology is based on Convolutional and Recurrent Neural Networks (CRNN) [33] for optical character modelling. This architecture is the basis of the HTR systems used in this work (see Figure 1). It consist of a stack of several convolutional layers followed by recurrent layers with Bidirectional Long-Short Term Memory (BLSTM) units [34, 35, 4, 36]. Finally, a softmax output layer computes the probabilities of each character in the training alphabet plus a non-character symbol (blank label).

The CRNN is trained by stochastic gradient descend with the RMSProp method [37] on minibatches to minimise the Connectionist Temporal Classification (CTC) cost function [38]. In order to reduce overfitting in the training process, dropout techniques [39], which have proved effectively to improve recognition accuracy [35, 36], are also used.

The CTC labelling or best path decoding is based on the assumption that the most probable path will correspond to the most probable labelling: [38]. In a first step, this method takes the best label per frame to compute the best path. Then, it removes the blank and the repeated labels from the obtained path.

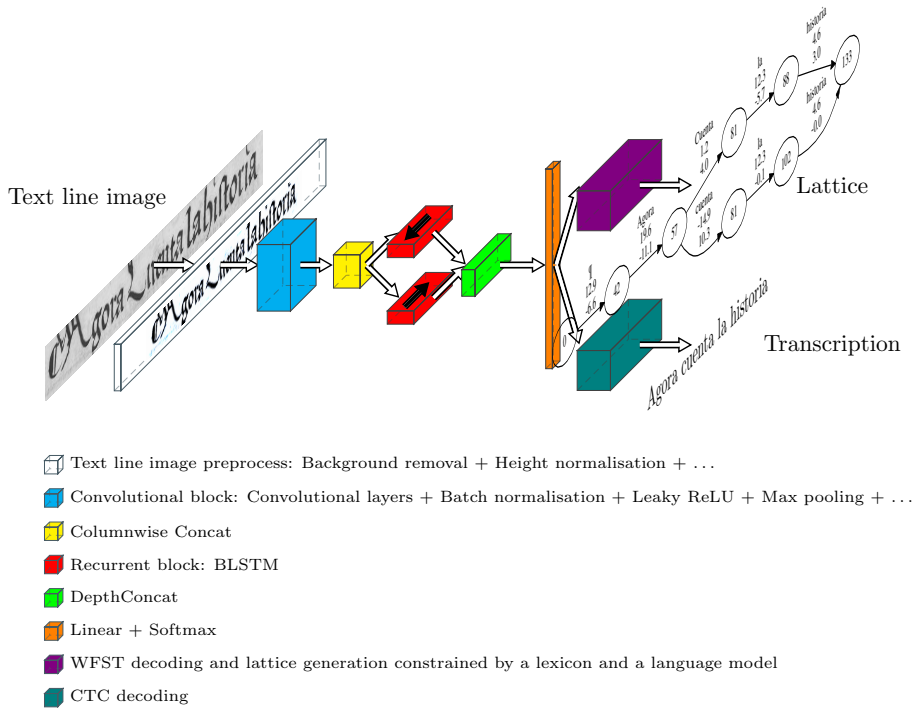
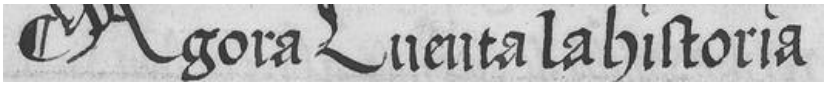


Figure 1: Convolutional recurrent neural network system architecture.

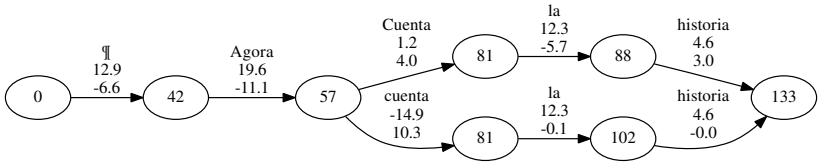
Finally, the linguistic context is explicitly modelled by means of statistical n-grams and, together with the lexical information, they are represented as Weighted Finite State Transducers (WFST). The CRNN output label probabilities, scaled with label priors, are then incorporated into the transducer edges [4].

The WFST decoding of a text line image can yield not only a single best solution, but also a huge set of best solutions compactly represented into a lattice [40]. Lattices are directed, acyclic and weighted graphs with an initial and a final node. The nodes correspond to the segmentation points between the lexical units. Links are defined as the edges between a starting node and an ending node, and each link represents a hypothesis lexical unit with the scores given by the optical and language models. Figure 2 presents an example of a text line image and the lattices generated by a decoder at word and character levels.

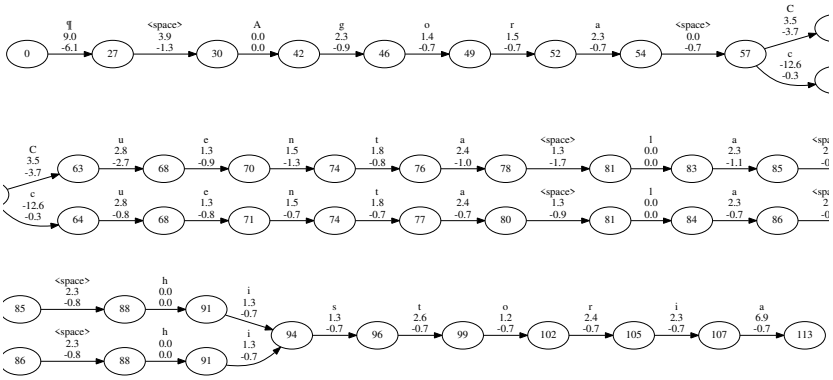
For the lattice generation, the search graph S is built by combining the following three WFSTs:



(a) Text line image.



(b) Word lattice.



(c) Character lattice.

Figure 2: A text line image and the lattices generated by a decoder at word and character levels. Character lattice is partially represented because of its excessive length.

- T: it is known as the "token" WFST, and it is designed to handle all the possible label sequences at the frame level. It allows the occurrence of the blank label along with the repetition of non-blank labels. It can map a sequence of frame-level CTC labels to a single character.
- L: it represents the "lexicon" WFST, and it models all the lexical units contained in the vocabulary as a concatenation of characters. It can map a sequence of characters to a single lexical unit.
- G: it is the "grammar" WFST, and it can be built from an n-gram language model. It restricts the decoding to the permissible sequences of lexical units.

T, L, and G are compiled independently and combined as follows:

$$S = T \circ \min(\det(L \circ G)) \tag{3}$$

Where $\circ$, det, and min denote composition, determination and minimisation, respectively. The determination and minimisation operations allow us to compress the search space, yielding to a faster lattice generation.

4. Computer Assisted Transcription Overview

As previously commented, in the last few years, the use of natural language recognition systems has allowed us to speed up the manual transcription of digitised documents, usually done by professional transcribers. However, state-of-the-art natural language recognition systems are far from being perfect, and human revision is required to produce a transcription of standard quality. Therefore, once the full recognition process of one document has finished, heavy human expert revision is required to really produce a transcription of standard quality. Such a post-editing solution is rather inefficient and uncomfortable for the human corrector.

In order to reduce the time and human effort required for obtaining the perfect transcription of digitised documents, transcribers can use interactive and assistive approaches, where the transcriber and the computer work together to obtain the perfect transcription. This is the case of Computer Assisted Transcription (CAT) of speech [41] or Computer Assisted Transcription of Text Images (CATTI) [42], where the user is directly involved in the transcription process, since he/she is responsible for validating and/or correcting the system hypothesis during the transcription process. The system takes into account the feedback provided by the user in order to propose a new, hopefully better, transcription. The corrections on this interactive transcription process can be performed at different lexical units; for instance, in this work we compare the performance of using words and characters as lexical units.

The CATTI process starts when the system proposes a full transcription $\hat{s}$ of a text line image. Then, the user reads this transcription until finding a mistake and makes a Mouse Action (MA) m (or equivalent pointer-positioning keystrokes) to position the cursor at this point. By doing so, the user is already providing some very useful information to the system: he is validating a prefix p of the transcription, which is error-free and, in addition, he is signalling that the following lexical unit e located after the cursor is incorrect. Hence, the system can already take advantage of this fact and directly propose a new suitable suffix, i.e. a new $\hat{s}$ in which in the position of the first wrong lexical unit of the previous suffix a different lexical unit is proposed. In this way, many explicit user corrections are avoided [43]. If the new suffix $\hat{s}$ corrects the erroneous lexical unit, a new cycle starts. However, if the new suffix has an error in the same position than the previous one, the user can make a new MA or can enter a lexical unit v to correct the erroneous one. This last action produces a new prefix p (the previously validated prefix followed by the new lexical unit v). Then, the system takes into account the new prefix to suggest a new suffix and a new cycle starts. This process is repeated until a correct transcription is accepted by the user.

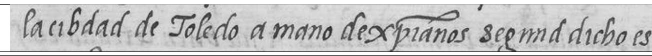
Text line image		
ITER-0	p	
ITER-1	$\hat{s}$	*la abadia de Toledo a mano de xpiānos segun el dicho es*
	m	⇑
	p	*la*
	$\hat{s}$	*deidad de Toledo a mano de xpiānos segun el dicho es*
	v	***c***
	p	*la cibdad de Toledo a mano de xpiānos segun el dicho es*
ITER-2	$\hat{s}$	*la cibdad de Toledo a mano de xpiānos segun el dicho es*
	m	⇑
	p	*la cibdad de Toledo a mano de xpiānos segun*
FINAL	$\hat{s}$	*d dicho es*
	v	#
	$p \equiv T$	*la **cibdad** de Toledo a mano de xpiānos **segund** dicho es*

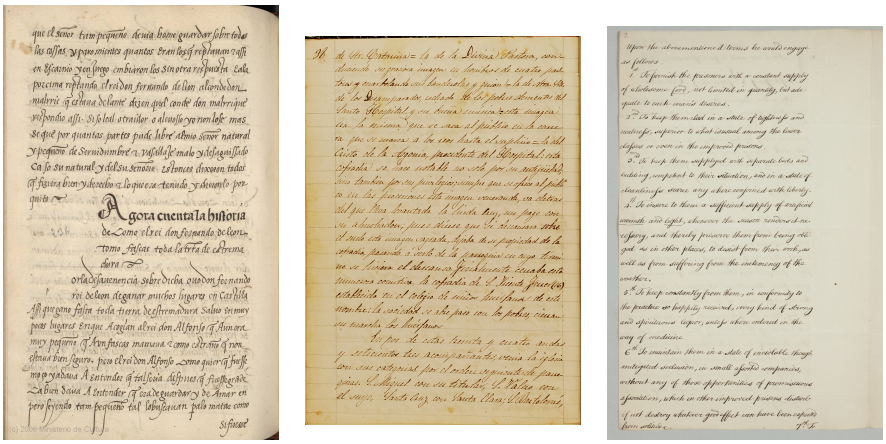
Figure 3: Example of CATTI operation using Mouse Actions at character level. Starting with an initial recognised hypothesis $\hat{s}$ from the text line image, the user validates its longest well-recognised prefix p by making a Mouse Action (MA) m, and the system emits a new recognised hypothesis $\hat{s}$. As the new suffix does not correct the mistake, the user types the correct character v, generating a new validated prefix p. Taking into account the new prefix, the system suggests a new hypothesis $\hat{s}$. As the new hypothesis corrects the first erroneous character after the new validated prefix, a new cycle starts. Now, the user validates the new longest prefix p, which is error-free, by making another MA m. The system provides a new suffix $\hat{s}$ taking into account this information. As the new hypothesis corrects the erroneous character, a new cycle starts. This process is repeated until the final error-free transcription T is obtained. The underlined boldface words in the final transcription are the two erroneous words corrected by the assistive system, where only the character ***c*** was corrected by the user. Note that in the iteration 1 two user interactions are needed (a MA and then, to type the correct character). However, in the iteration 2 only one user interaction (one MA) is needed.

Figure 3 illustrates an example of the CATTI process. In this example, without interaction with a CATTI system, a user should have to correct about three errors at word level and nine errors at character level from the original recognised hypothesis: *abadia* should be changed by *cibdad* (5 characters), *segun* by *segund* (1 character) and *el* should be deleted (3 characters counting with the previous space character). Using CATTI at character level only one explicit user-correction is necessary to get the final error-free transcription in two CATTI iterations: in the iteration 1 a single MA does not succeed and the correct character needs to be typed, but in the iteration 2 only one MA is needed to find the correct word and to remove the following erroneous word.

The CATTI framework can be defined as a traditional natural language recognition problem -Equation (1)-. In this case, in addition to the given feature sequence x, a prefix p of the transcription is available, depending on the editing operation that the user performed to correct the erroneous text. The editing operations considered are substitution, insertion, deletion, and rejection [44]. Therefore, the CATTI system should try to complete the transcription from this prefix p by searching for the most likely suffix $\hat{s}$:

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \mathrm{P}(s \mid x, p) \approx \underset{s \in S}{\operatorname{argmax}} \mathrm{P}(x \mid p, s)\mathrm{P}(s \mid p) \tag{4}$$

where S represents the set of all possibles suffixes s of p.



(a) Page 579 of Rodrigo. (b) Page 41 of CS. (c) Page 115 of Bentham.

Figure 4: Page samples of the historical manuscripts used in this work.

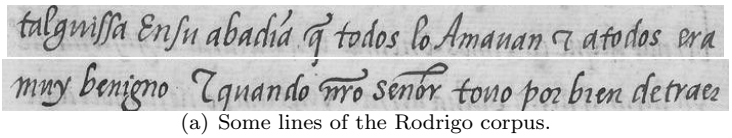
Equation (4) is very similar to first part of Equation (1), being w the concatenation of p and s. The difference is that now a part of the transcription, p, is given. As shown in [44], $\mathrm{P}(x \mid p, s)$ can be approximated by morphological models and $\mathrm{P}(s \mid p)$ by a language model conditioned by p as in conventional HTR. Therefore, the search must be performed over all possible suffixes of p. This search for s can be efficiently carried out by using the lattices obtained from a neural network based HTR system during the WFST-based decoding of the whole input signal representation x. In each interaction step, the decoder parses the validated prefix p over the lattice and then it continues searching for a suffix which maximises the posterior probability according to Equation (4). This process is repeated until a complete and correct transcription of the input text line image is obtained.

5. Experimental Framework

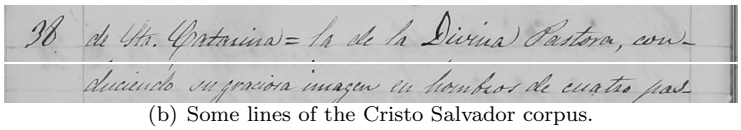
This section presents the three historical manuscripts used in the experiments, the main features of the recognition system and modules, and the evaluation metrics.

5.1. Historical Manuscripts: The Rodrigo, the Cristo Salvador, and the Bentham Corpora

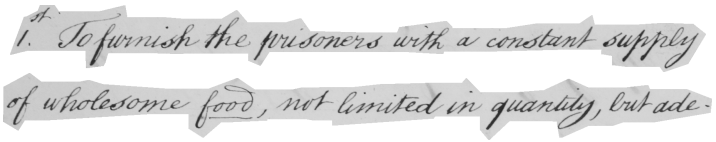
The Rodrigo corpus [6] corresponds to the digitisation of the book "Historia de España del arçobispo Don Rodrigo", which was written in old Castilian (Spanish) in 1545 by a single author. As the page presented in Figure 4(a), most pages consist of a single block of 25 well-separated lines of humanistic script, similar to the italic script but with textual Gothic influences. This book



(a) Some lines of the Rodrigo corpus.



(b) Some lines of the Cristo Salvador corpus.



(c) Some lines of the Bentham corpus.

Figure 5: Examples of separated text lines.

is composed of 853 pages that were semiautomatically segmented into lines such as the line examples shown in Figure 5(a).

The Cristo-Salvador (CS) corpus is a XIX century Spanish manuscript provided by *Biblioteca Valenciana Digital* (BiValDi). It is also a single writer book with different image features that cause some problems, such as smear, background variations, differences in bright, and bleed-through (ink that trespasses to the other surface of the sheet). It is composed of 53 pages (similar to that presented in Figure 4(b)) that were semiautomatically divided into lines (such as the lines shown in Figure 5(b)). Rodrigo and CS corpora are publicly available for research purposes on the website of the Pattern Recognition and Human Language Technology (PRHLT) research center [45].

The last corpus used in this paper corresponds to the dataset used in the HTR competition organised in 2014 within the "International Conference on Frontiers of Handwriting Recognition" (ICFHR-2014) [9]. The corresponding dataset is a small part of the Bentham Papers corpus. The Bentham Papers collection encompasses around 100,000 page images written in English by several writers. The collection includes documents authored by the philosopher and reformer Jeremy Bentham over a period of sixty years. Some of them are written by Bentham himself and other ones are fair copies handwritten by Bentham's secretarial staff. The dataset used here [6] is composed of 433 page images (similar to the page presented in Figure 4(c)) that were semiautomatically transcribed and segmented into lines as the line examples shown in Figure 5(c).

In this work, we used the same partitions for Rodrigo that were used in previous works [7]. The first 409 pages (10,000 text lines) were used for training (9000 lines for training the optical and language models and 1000 for validation),

[6] It is publicly available at `http://doi.org/10.5281/zenodo.44519`

Table 1: Description of the partitions of the Rodrigo corpus used in this work.

	Training	**Validation**	**Test**	**Total**
Lines	9000	1000	5010	15,010
Running words	102,631	11,387	56,797	170,815
Running OOV words	–	682	4313	4995
Word lexicon size	10,778	2788	6668	14,437
OOV words lexicon size	–	662	3087	3659
Running chars	493,128	54,936	272,132	820,196
Running OOV chars	–	1	14	15
Char lexicon size	105	82	91	111
OOV chars lexicon size	–	1	6	6

Table 2: Description of the partitions of the Cristo Salvador corpus used in this work.

	Training	**Validation**	**Test**	**Total**
Lines	662	78	473	1213
Running words	6680	873	5199	12,752
Running OOV words	–	219	1357	1576
Word lexicon size	2216	410	1625	3451
OOV words lexicon size	–	211	1056	1235
Running chars	32,814	4236	24,486	61,536
Running OOV chars	–	–	2	2
Char lexicon size	91	52	84	92
OOV chars lexicon size	–	–	1	1

and the last 205 pages (5010 text lines) were used for testing [7]. For Cristo Salvador we followed the directrices of the *hard* partition defined in a previous work [8]: the last 20 pages for test (473 text lines), and the rest of the pages for training. The first 30 pages (662 text lines) were used for training and the remaining 3 pages (78 text lines) for validation purposes. For the Bentham corpus we used the partition defined in the HTR ICFHR-2014 competition [9]. The 433 pages were divided into three subsets for training, development, and test, respectively encompassing 350, 50, and 33 page images. Tables 1, 2, and 3 summarise the information contained in the partitions used for Rodrigo, Cristo Salvador, and Bentham corpora, respectively. Word statistics were computed separating the words and punctuation marks.

Out-of-vocabulary (OOV) words is one of the challenges that can be found in the first two corpora. In Rodrigo, there are many rare words and words in their archaic forms, words that appear in distinct form in the training and test sets (e.g., *portugal* and *portuḡl*), abbreviations, and words hyphenated differently in

[7] It is publicly available at `http://doi.org/10.5281/zenodo.1490009`.

Table 3: Description of the partitions of the Bentham corpus used in this work. In this corpus there are not OOV chars.

	Training	**Validation**	**Test**	**Total**
Lines	9198	1415	860	11,473
Running words	89,392	13,812	8021	111,225
Running OOV words	–	778	428	1206
Word lexicon size	8235	2702	1929	9181
OOV words lexicon size	–	605	375	946
Running chars	420,029	64,929	38,540	523,498
Char lexicon size	92	85	81	92

the training and test sets, yielding a 7.6% of OOV words. Moreover, some scarce OOV characters appear in the testing partition (such as \, $\acute{p}$, $\bar{g}$, $\hbar$ and w) that do not belong to the training set. In the case of Cristo Salvador, this problem is aggravated due to the shortage of samples, reaching a 26.1% of OOV words, but very few OOV characters are present. Finally, the Bentham corpus presents only 5.3% of OOV words and no OOV characters.

In order to estimate what is the influence of the amount of data employed for estimating the lexical and language models (i.e., the lexical and language restrictions) in the final results, additional datasets for each corpora were used to increase the textual data employed for inferring these models.

In the case of Rodrigo and Cristo Salvador, we used the validation partitions data due to the difficulty of finding adequate textual information for this kind of historical documents. In order to check the effect of increasing data, we created two different datasets: training plus half of the validation dataset, and training plus the whole validation set.

We followed the same approximation in Bentham, but in this case it was possible to employ, apart from the validation partition, an additional corpus [46]. The statistics for the finally used datasets are presented in Tables 4, 5 and 6, for Rodrigo, Cristo Salvador, and Bentham, respectively. In summary, adding validation data increments the number of samples by 11.1% for Rodrigo and 11.8% for Cristo Salvador with respect to the original training data. In the case of Bentham, the increments from the training data are about 15.4% when using only the validation data and more than a 200% when adding the extra dataset.

5.2. System Setup

The HTR systems used in this work are based on the technology described in [19]. This approach has shown that using BLSTM provides a handwriting recognition performance comparable to that obtained when using multi-dimensional LSTM. This approach is implemented in the HTR Laia Toolkit [19], which is based on the Torch machine learning platform.

While raw images can be directly accepted as input, results can be better if images are previously preprocessed. In this work, the employed preprocessing consists of slope and slant correction, background removal, 64 pixels in height

Table 4: Description of the partitions with additional data for inferring the language and lexical models for the Rodrigo corpus.

	Training	**Training + 50% Validation**	**Training + Validation**
Lines	9000	9500	10,000
Running words	102,631	108,403	114,018
Word lexicon size	10,778	11,107	11,442
Running chars	493,128	520,908	548,064
Char lexicon size	105	106	106

Table 5: Description of the partitions with additional data for inferring the language and lexical models for the Cristo Salvador corpus.

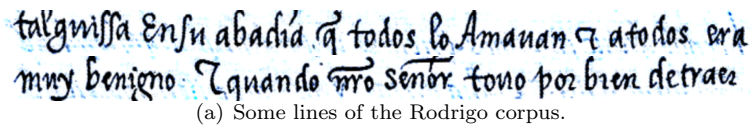
	Training	**Training + 50% Validation**	**Training + Validation**
Lines	662	701	740
Running words	6680	7124	7553
Word lexicon size	2216	2218	2429
Running chars	32,814	34,956	37,050
Char lexicon size	91	91	91

Table 6: Description of the partitions with additional data for inferring the language and lexical models for the Bentham corpus.

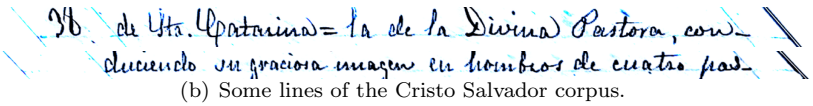
	Training	**Training + 50% Validation**	**Training + Validation**	**Training + Extra**
Lines	9198	9905	10,613	29,598
Running words	89,392	95,535	103,204	271,801
Word lexicon size	8235	8562	8842	12,292
Running chars	420,029	448,942	484,958	1,559,034
Char lexicon size	92	92	92	97

Table 7: Architecture of the optical models.

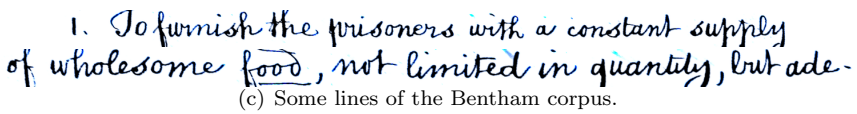
Parameters	Rodrigo & Bentham	Cristo Salvador
CNN Layers	5	3
Filters	{16,32,48,64,80}	{16,32,48}
Kernel size	3×3	3×3
MaxPool size	2×2	2×2
Dropout	{0,0,0.2,0.2,0.2}	{0,0.2,0.2}
RNN Layers	5	3
BLSTM Units	256	256



(a) Some lines of the Rodrigo corpus.



(b) Some lines of the Cristo Salvador corpus.



(c) Some lines of the Bentham corpus.

Figure 6: Examples of preprocessed text lines.

normalisation, and noise removing [47, 48, 49, 50, 51]. This preprocessing has been carried out by using the `textFeats` tool [8]. As an example, Figure 6 presents the resulting images of preprocessing the lines shown in Figure 5.

The optical models are based on CRNN, which consist of a convolutional (CNN) block and a recurrent (RNN) block with the architecture detailed in Table 7 for each corpus. The CRNN consists of convolutional layers with filters composed of different features maps with kernel sizes of 3×3 pixels and horizontal and vertical strides of 1 pixel. LeakyReLU is used as the activation function, and the output of the convolutional layers is fed to a maximum pooling layer with non-overlapping kernels of 2×2 pixels (only at the output of the first three layers for Rodrigo and Bentham, and at the output of the first two layers for Cristo Salvador). After that, the recurrent blocks are composed of different recurrent layers composed of 256 Bidirectional Long-Short Term Memory (BLSTM) units. Finally, a linear fully-connected layer is used after the recurrent block. The CTC training is carried out with minibatches of 16

[8] `https://github.com/mauvilsa/textfeats`

samples, using a base learning rate of 0.0003. All the hyper parameters, such as the number of convolutional and recurrent layers, were set up on the validation sets.

The lexicon models are in HTK lexicon format [52]. For the experiments with lexicon restrictions (decoding at word level), each word from the training partition was modelled as a concatenation of characters. On the other hand, for the experiments without lexicon restrictions (decoding at character level), the lexicon models contain only the set of characters contained in the training partition.

The language models were estimated as n-gram with Kneser-Ney back-off smoothing [53] directly from the transcriptions of the text lines included on the training partition using the SRILM *ngram-count* tool [54]. For the experiments with language restrictions at word level 2-gram word language models were estimated for the three corpora. However, for the experiments with language restrictions at character level a 7-gram character language model was estimated for the Cristo Salvador corpus and 8-gram character language models were estimated for the Rodrigo and Bentham corpora. On the other hand, for the experiments without language restrictions, zero-gram language models were used at both word and character level. The lattice generation was performed by using the EESEN decoder [55], which is based on WFST.

With respect to the user interaction in the CATTI system, in this work the best performance, over the validation sets, was obtained with a limit of 2 Mouse Actions. Thus, the assistance in all interactive-assistive experiments was limited to 2 Mouse Actions.

5.3. Evaluation Metrics

The transcriptions quality is assessed using the Levenshtein edit distance [56] with respect to the reference text, which allows us to obtain a good estimation for the transcriber post-edition effort at both the word and the character levels. In this framework, the Character Error Rate (CER) value is especially interesting, since transcription errors are usually corrected at the character level. The CER is the Levenshtein edit distance at character level and it can be defined as the minimum number of substitutions, deletions, and insertions needed to transform the transcription into the reference text, divided by the number of characters in the reference text:

$$\mathrm{CER} = \frac{s + d + i}{n} \cdot 100\% \tag{5}$$

where s is the number of substitutions, d is the number of deletions, i is the number of insertions, and n is the total number of characters in the reference text. Similarly, Word Error Rate (WER) is this edit distance calculated at word level.

The quality of the lattices can be defined as the quality of the best hypothesis contained in each of them, and it is known as the oracle error rate. Then, the quality of the word lattices is estimated by the oracle WER, which represents the smallest WER that can be obtained from the word sequences contained in

them. In the same way, the quality of the character lattices is estimated by the oracle CER, which represents the smallest CER that can be obtained from the character lattices.

In the interactive approach, Word Click Rate (WCR) and Character Click Rate (CCR) are used to assess the number of additional Mouse Actions (MA) per word or character that the user has to do in order to obtain the best transcription from the interactive system. The definition of CCR is:

$$\mathrm{CCR} = \frac{c}{n} \tag{6}$$

where c is the number of MA carried out at character level and n is, again, the number of reference characters. The definition of the Word Click Rate (WCR) is analogous to the previous one, but substituting characters by words.

The overall interactive performance is given by the Word Stroke Ratio (WSR) and the Character Stroke Ratio (CSR), which can be also computed by using the reference text. After each hypothesis is proposed by the system, the longest common prefix between the hypothesis and the reference text is obtained, and the first error from the hypothesis is corrected in an interaction action by the user. This process is iterated until a full match is achieved.

Therefore, the CSR can be defined as the number of (character level) user interaction actions (a) that are necessary to achieve the reference transcriptions of the text images considered, divided by the total number of reference characters (n):

$$\mathrm{CSR} = \frac{a}{n} \cdot 100\% \tag{7}$$

The definition of WSR is analogous but at word level. This definition makes comparable the WER with the WSR, and the CER with the CSR. The relative difference between the recognition error and the stroke ratio gives us the effort reduction (EFR), which is an estimation of the reduction of the transcription effort that can be achieved by using the interactive system.

The statistical significance of the experimental results is estimated by means of confidence intervals of probability 95% ($\alpha = 0.025$) calculated by using the bootstrapping method with 10,000 repetitions [57].

6. Experimental Results

This section presents the experimental results for the three corpora. Firstly, the transcription given by the CTC decoding is evaluated to use it as a reference with which to compare in the following experiments. Next, four experiments for studying the influence of the lexicon and language restrictions in the lattice generation and the computer assisted transcription are performed. Specifically, two experiments (Section 6.2 and Section 6.3) at word level (i.e. with lexicon restrictions) and two experiments (Section 6.4 and Section 6.5) at character level (i.e. without lexicon restrictions). For both levels, word and character, one

Table 8: Quality of the CTC transcription given by the CRNN system.

Measure	Rodrigo	Cristo Salvador	Bentham
WER	$8.4\% \pm 0.2$	$25.5\% \pm 1.4$	$11.9\% \pm 0.9$
CER	$1.75\% \pm 0.05$	$8.15\% \pm 0.60$	$3.30\% \pm 0.29$

experiment (Section 6.2 and Section 6.4) was performed with language restrictions imposed by language models and the other without language restrictions (Section 6.3 and Section 6.5).

In the experiments with lexical or language restrictions (Sections 6.2, 6.3, and 6.4) the use of additional textual information for training the lexical and language models was used in order to test the performance with additional data, as described in Section 5.1.

6.1. Transcription given by the CTC decoding

Table 8 presents the results obtained by the CTC decoding. As it can be seen in this table, despite the large number of out-of-vocabulary words (7.6% for Rodrigo, 26.1% for Cristo Salvador, and 5.3% for Bentham), the transcription given by the CTC decoding presents a quite good quality. It should be noted that the transcription obtained for Cristo Salvador presents a WER smaller than the limit established by the percentage of out-of-vocabulary words. It is possible because the CTC decoding provides character sequences not restricted by the lexicon or the language model.

The CTC decoding in state-of-the-art handwritten recognition systems based on deep neural networks allows us to obtain transcriptions with very low error rates. However, from these systems much more knowledge can be extracted. One way to extract it is by using WFST decoding for obtaining a set of hypotheses that can be compactly stored in form of lattices and that are richer than the single hypothesis solution given by the CTC transcription.

An interactive and assistive transcription system can take advantage of the knowledge stored in lattices and help to reduce the transcriber workload and time to obtain the correct transcription. In the next experiments, the influence of lexicon and language restrictions in computer assisted transcription are studied.

6.2. Computer assisted transcription with lexicon and language restrictions

In this work, the best results when decoding the validation set with lexicon and language restrictions were obtained by using 2-gram language models, for the three corpora. Thus, these models were used in the test set decoding.

As it can be seen in the first part of Table 9, the 1-best hypotheses offered by the lattices generated with lexicon and language restrictions for the Rodrigo corpus present a WER equal to $16.7\% \pm 0.3$. This value is statistically significant better than the best result ($17.9\% \pm 0.4$) presented in a recent previous work [7],

Table 9: Results for the computer assisted transcription with lexicon and language restrictions (2-gram for all corpora). EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	Measure	Rodrigo	Cristo Salvador	Bentham
Lattice generation	**WER**	$16.7\% \pm 0.3$	$46.5\% \pm 1.4$	$23.3\% \pm 1.1$
	CER	$3.98\% \pm 0.08$	$17.32\% \pm 0.67$	$5.56\% \pm 0.3$
	Oracle WER	10.0%	32.7%	10.6%
CATTI	**WSR**	$12.4\% \pm 0.3$	$40.2\% \pm 1.3$	$14.9\% \pm 0.8$
	WCR	0.31 ± 0.01	0.90 ± 0.03	0.37 ± 0.01
	MA	$17,625$	4698	2960
	$\textbf{EFR}_{\textbf{WFST}}$	25.7%	13.6%	36.1%
	$\textbf{EFR}_{\textbf{CTC}}$	-47.6%	-57.6%	-25.2%

although worse than the value obtained by the CTC decoding (see Table 8) because of the low amount of data employed to generate the language model. In any case, the obtained word lattices contain hypotheses that reach an oracle WER equal to 10.0%.

Regarding the Cristo Salvador corpus, the performance of our system at word-level is quite poor. This is because it is a relatively small data set (only 662 samples were used for training in this work) and training and test sets present a high disparity, reflected in the 26.1% of out-of-vocabulary words. Again, the scarce data used to infer the language model explains why these results are worse than the CTC results. The generated lattices present a 1-best WER equal to $46.5\% \pm 1.4$ and an oracle WER equal to 32.7%.

In the case of Bentham, the generated word lattices present worse results than the obtained by the CTC decoding, specifically, a 1-best WER equal to $23.3\% \pm 1.1$ and an oracle WER equal to 10.6%.

The second part of Table 9 presents the results obtained by the assistive system (CATTI) working at word-level with lexicon and language restrictions. As it can be observed, the assistive system allows us to decrease significantly the number of words to correct over the most probable hypotheses contained in the word lattices. Concretely, for Rodrigo it presents a WSR equal to $12.4\% \pm 0.3$, which represents a 25.7% of significant human effort reduction (WER equal to $16.7\% \pm 0.3$). For Cristo Salvador it presents a WSR equal to $40.2\% \pm 1.3$, which represents 13.6% of significant human effort reduction (WER equal to $46.5\% \pm 1.4$). Finally, for Bentham it presents a WSR equal to $14.9\% \pm 0.8$, which represents 36.1% of significant human effort reduction (WER equal to $23.3\% \pm 1.1$).

Nevertheless, given that the performance of CATTI is limited by the quality of the hypotheses contained in the lattices, only in the case of the Bentham corpus some human effort reduction could be expected with respect to CTC decoding. However, this is not the case, and no human effort reduction can be

Table 10: Results for the computer assisted transcription with lexicon and language restrictions (2-gram for all corpora) with additional data for the Rodrigo corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$16.5\% \pm 0.3$	$16.2\% \pm 0.3$
	CER	$3.93\% \pm 0.09$	$3.87\% \pm 0.08$
	Oracle WER	9.7%	9.6%
CATTI	**WSR**	$12.3\% \pm 0.3$	$12.1\% \pm 0.3$
	WCR	0.31 ± 0.01	0.30 ± 0.01
	MA	$17,380$	$17,195$
	$\textbf{EFR}_{\textbf{WFST}}$	25.5%	25.3%
	$\textbf{EFR}_{\textbf{CTC}}$	-45.9%	-44.3%

Table 11: Results for the computer assisted transcription with lexicon and language restrictions (2-gram for all corpora) with additional data for the Cristo Salvador corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$46.5\% \pm 1.4$	$45.8\% \pm 1.4$
	CER	$17.34\% \pm 0.65$	$16.69\% \pm 0.64$
	Oracle WER	32.7%	31.9%
CATTI	**WSR**	$40.2\% \pm 1.3$	$39.3\% \pm 1.3$
	WCR	0.90 ± 0.03	0.89 ± 0.03
	MA	4698	4597
	$\textbf{EFR}_{\textbf{WFST}}$	13.5%	14.2%
	$\textbf{EFR}_{\textbf{CTC}}$	-57.6%	-54.1%

considered when comparing the obtained WSR with the transcription offered by the CTC decoding for the three corpora (see Table 8).

Finally, regarding the number of clicks or additional Mouse Actions (MA), in the case of Rodrigo $17,625$ MA where performed for transcribing the 5010 text lines, resulting in 3.5 MA per line, 0.31 MA per word, and 0.065 MA per character. In the case of Cristo Salvador, 4698 MA where performed for transcribing the 473 text lines, resulting in 9.9 MA per line, 0.90 MA per word, and 0.192 MA per character. Finally, for Bentham 2960 MA where performed for transcribing the 860 text lines, resulting in 3.4 MA per line, 0.37 MA per word, and 0.077 MA per character. In the case of Cristo Salvador these figures are really bad, but it was expected because of the high error rates and the high percentage of OOV words.

In order to see the influence of the training data size on the performance of the system, experiments with the same test set but training the lexical and

Table 12: Results for the computer assisted transcription with lexicon and language restrictions (2-gram for all corpora) with additional data for the Bentham corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	Measure	Training + 50% Validation	Training + Validation	Training + Extra
Lattice generation	**WER**	$22.9\% \pm 1.1$	$22.8\% \pm 1.0$	$17.0\% \pm 0.9$
	CER	$5.43\% \pm 0.29$	$5.32\% \pm 0.30$	$3.69\% \pm 0.24$
	Oracle WER	10.1%	10.0%	4.28%
CATTI	**WSR**	$14.5\% \pm 0.8$	$14.2\% \pm 0.8$	$8.6\% \pm 0.6$
	WCR	0.36 ± 0.02	0.35 ± 0.02	0.22 ± 0.01
	MA	2881	2839	1726
	$\textbf{EFR}_{\textbf{WFST}}$	36.7%	37.7%	49.4%
	$\textbf{EFR}_{\textbf{CTC}}$	−21.8%	−19.6%	27.8%

language models with the data described in Section 5.1 were performed.

The results for Rodrigo, Cristo Salvador, and Bentham are presented in Tables 10, 11, and 12, respectively. From these results it is clear that adding more data for training the lexical and language models is beneficial in any case for the recognition, which is reflected in the decrease of WER, CER, and oracle WER. The tendency shows that, the more data, the better results. This tendency appears as well in the assistive system results. However, results are only significantly better compared with those obtained without additional data when a huge amount of additional data is available (i.e., Bentham case with the extra dataset), while providing only a moderate amount of additional data has no practical effect. Moreover, when studying the effort reduction when using the assistive system, the same conclusions are obtained: only using a huge amount of training data for lexical and language models provides a better effort reduction than the only use of the CTC transcription postedition (last row and column cell of Table 12.

Consequently, we can conclude that the pure CTC approximation offers a better performance than the assistive system that employs lexical and language restrictions when the amount of training data is moderate, something that it is quite usual when dealing with historical texts because of the difficulty of finding comparable data to improve the lexical and language models.

6.3. Computer assisted transcription with lexicon restrictions but without language restrictions

The first part of Table 13 shows the results obtained by the word lattices generation without language restrictions (using zero-gram word language models). For the Rodrigo corpus, the best hypotheses offered by these lattices presents a WER equal to $18.8\% \pm 0.4$, which is statistically significant worse than the

Table 13: Results for the computer assisted transcription with lexicon restrictions but without language restrictions. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	Measure	Rodrigo	Cristo Salvador	Bentham
Lattice generation	**WER**	$18.8\% \pm 0.4$	$55.5\% \pm 1.8$	$27.3\% \pm 1.3$
	CER	$3.80\% \pm 0.08$	$16.09\% \pm 0.63$	$5.83\% \pm 0.31$
	Oracle WER	10.5%	34.1%	10.4%
CATTI	**WSR**	$12.0\% \pm 0.3$	$42.0\% \pm 1.4$	$17.3\% \pm 0.9$
	WCR	0.33 ± 0.01	0.96 ± 0.03	0.40 ± 0.02
	MA	18,745	4964	3170
	$\textbf{EFR}_{\textbf{WFST}}$	36.2%	24.3%	36.6%
	$\textbf{EFR}_{\textbf{CTC}}$	−42.9%	−64.7%	−45.4%

obtained in the previous experiment. However, the oracle WER (10.5%) it is only slightly worse.

In the case of the Cristo Salvador corpus, the obtained lattices present a 1-best WER equal to $55.5\% \pm 1.8$ and an oracle WER equal to 34.1%. These results are, once again, worse than the results obtained in the previous experiment.

For the Bentham corpus, the generated lattices present a 1-best WER equal to $27.3\% \pm 1.3$ and an oracle WER equal to 10.4%. In this case, the WER is worse than the obtained in the previous experiment but the oracle WER is slightly better.

In the second part of Table 13 the results obtained by the CATTI system working at word-level without language restrictions are presented. As it can be observed, for Rodrigo it presents a WSR equal to $12.0\% \pm 0.3$, which is slightly better than the obtained in the previous experiment. For Cristo Salvador it presents a WSR equal to $42.0\% \pm 1.4$, which is slightly worse than the obtained in the previous experiment. Finally, for Bentham it presents a WSR equal to $17.3\% \pm 0.9$, which is statistically significant worse than the obtained in the previous experiment. Therefore, as in the previous experiment, no human effort reduction can be considered when comparing the obtained WSR with the transcription offered by the CTC decoding.

Finally, regarding the number of interactions, in the case of Rodrigo 18,745 MA where performed, resulting in 3.7 MA per line, 0.33 MA per word, and 0.069 MA per character. In the case of Cristo Salvador, 4964 MA where performed, resulting in 10.5 MA per line, 0.95 MA per word, and 0.203 MA per character. For Bentham, 3170 MA where performed, resulting in 3.7 MA per line, 0.39 MA per word, and 0.082 MA per character.

From the obtained results, it can be observed that removing the language restrictions when working at word-level in the CATTI system increases the required number of user interactions for obtaining similar or even worse results.

In order to see the influence of the training data size on the performance of

Table 14: Results for the computer assisted transcription with lexicon restrictions but without language restrictions with additional data for the Rodrigo corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$18.5\% \pm 0.4$	$18.1\% \pm 0.4$
	CER	$3.76\% \pm 0.08$	$3.87\% \pm 0.08$
	Oracle WER	10.3%	10.1%
CATTI	**WSR**	$11.8\% \pm 0.3$	$11.7\% \pm 0.3$
	WCR	0.33 ± 0.01	0.32 ± 0.01
	MA	$18,475$	$18,148$
	$\textbf{EFR}_{\textbf{WFST}}$	36.2%	35.4%
	$\textbf{EFR}_{\textbf{CTC}}$	-39.7%	-38.8%

Table 15: Results for the computer assisted transcription with lexicon restrictions but without language restrictions with additional data for the Cristo Salvador corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$55.4\% \pm 1.9$	$55.5\% \pm 1.8$
	CER	$16.08\% \pm 0.64$	$16.10\% \pm 0.64$
	Oracle WER	34.1%	34.1%
CATTI	**WSR**	$42.0\% \pm 1.4$	$42.0\% \pm 1.4$
	WCR	0.95 ± 0.03	0.95 ± 0.03
	MA	4964	4964
	$\textbf{EFR}_{\textbf{WFST}}$	24.2%	24.3%
	$\textbf{EFR}_{\textbf{CTC}}$	-64.6%	-64.6%

the system, experiments with the same test set but training the lexical models with the data described in Sections 5.1 were performed.

The results for Rodrigo, Cristo Salvador, and Bentham are presented in Tables 14, 15, and 16, respectively. From these results it can be seen that adding more data for the lexical model inference has a low positive impact (that is not statistically significant with respect to those results obtained without the additional data) in recognition performance when the provided additional data is similar to the test data, which is not the case of the large set of extra data used in Bentham (last column of Table 16, WER and CER results).

However, the impact on the assisted transcription is positive or neutral in all cases; even in the case that mismatching data is provided for increasing the lexicon, the interactive results are slightly worse, but differences are not significant at WSR level. This can be explained by the low oracle WER obtained

Table 16: Results for the computer assisted transcription with lexicon restrictions but without language restrictions with additional data for the Bentham corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**	**Training + Extra**
Lattice generation	**WER**	$26.5\% \pm 1.2$	$26.1\% \pm 1.2$	$38.1\% \pm 1.1$
	CER	$5.67\% \pm 0.31$	$5.60\% \pm 0.29$	$8.50\% \pm 0.31$
	Oracle WER	9.9%	9.7%	4.29%
CATTI	**WSR**	$16.7\% \pm 0.9$	$16.6\% \pm 0.9$	$18.0\% \pm 0.8$
	WCR	0.38 ± 0.02	0.38 ± 0.02	0.53 ± 0.02
	MA	3087	3050	4221
	$\textbf{EFR}_{\textbf{WFST}}$	37.0%	36.4%	52.8%
	$\textbf{EFR}_{\textbf{CTC}}$	−40.6%	−39.2%	−51.2%

in this case in the recognition process, that means that the correct words are present at more or less the same position in the list of alternatives in all cases.

In any case, the addition of the lexical restriction does not provide effort reduction with respect to using pure CTC decoding, as the EFR_{CTC} results shows for all corpora with additional data.

6.4. Computer assisted transcription without lexicon restrictions but with language restrictions

When decoding the validation partition using character-level language models, i.e. with restrictions imposed only by the character language model, the best results were obtained by using 8-gram character language models in the cases of Rodrigo and Bentham, and 7-gram in the case of Cristo Salvador. Thus, these models were employed as well for the test decoding. Since decoding is at character level, the most important evaluation values should be those related to character (i.e., CER, oracle CER, CSR), instead of the word evaluation measures we paid attention to in the previous experiments.

The first part of Table 17 presents the quality of the generated character lattices with language restrictions. In the case of Rodrigo, these lattices present a CER equal to $2.15\% \pm 0.06$, which represents a statistically significant improvement over the best result ($3.01\% \pm 0.07$) presented in a previous work [7], but again worse than the result obtained by the CTC decoding ($1.75\% \pm 0.05$, see Table 8). However, the quality of the hypotheses contained in these lattices reach an oracle CER equal to 0.36%, which represents a relative improvement of 79.4% over the CER present on the transcription offered by the CTC decoding.

Regarding Cristo Salvador, the character lattices presented a CER equal to $9.75\% \pm 0.68$ (worse than the CTC result) and an oracle CER equal to 2.63%. This oracle CER represents a relative improvement of 67.7% over the CER present on the transcription given by the CTC decoding (8.15%, see Table 8).

Table 17: Computer assisted transcription without lexicon restrictions but with language restrictions (7-gram for Cristo Salvador and 8-gram for Rodrigo and Bentham). EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (CER in second row) and CTC decoding (CER in Table 8), respectively.

Experiment	Measure	Rodrigo	Cristo Salvador	Bentham
Lattice generation	**WER**	$11.0\% \pm 0.3$	$30.7\% \pm 1.5$	$11.9\% \pm 0.9$
	CER	$2.15\% \pm 0.06$	$9.75\% \pm 0.68$	$3.10\% \pm 0.27$
	Oracle CER	0.36%	2.63%	1.03%
CATTI	**CSR**	$0.67\% \pm 0.04$	$4.59\% \pm 0.51$	$1.70\% \pm 0.20$
	CCR	0.032 ± 0.001	0.161 ± 0.013	0.050 ± 0.005
	MA	8802	3955	2058
	EFR_{WFST}	68.8%	52.9%	45.2%
	EFR_{CTC}	61.7%	43.7%	48.5%

For Bentham, the character lattices generation gives us a CER equal to $3.10\% \pm 0.27$, which is better than the result obtained by the CTC decoding (a CER equal to $3.30\% \pm 0.29$), and an oracle CER equal to 1.03%, which represents a relative improvement of 68.8%.

Given the oracle CER reached by the hypotheses contained in the generated character lattices, an outstanding performance in our computer assisted transcription system can be expected when working at character level.

The obtained results for the CATTI experiments are presented in the second part of Table 17. A CSR equal to $0.67\% \pm 0.04$ for Rodrigo corpus was achieved. It represents a 68.8% of significant human effort reduction over the most probable hypotheses contained in the character lattices. Moreover, thanks to the knowledge contained in the character lattices, our assistive system is able to offer 61.7% of statistically significant human effort reduction over the transcription offered by the CTC decoding, which presented a very low CER ($1.75\% \pm 0.05$, see Table 8). A similar behaviour was observed for Cristo Salvador and Bentham. In the case of Cristo Salvador, a CSR equal to $4.59\% \pm 0.51$ was achieved, which represents a statistically significant human effort reduction of 52.9% over the most probable hypotheses contained in the character lattices, and 43.7% over the transcription offered by the CTC decoding (CER equal to $8.15\% \pm 0.60$, see Table 8). In the case of Bentham, a CSR equal to $1.70\% \pm 0.20$ was achieved, which represents a statistically significant human effort reduction of 45.2% over the most probable hypotheses contained in the character lattices, and 48.5% over the transcription offered by the CTC decoding (CER equal to $3.30\% \pm 0.29$, see Table 8).

Regarding the number of clicks or additional Mouse Actions (MA), for Rodrigo 8802 MA where performed, resulting in 1.8 MA per line, 0.16 MA per word, and 0.032 MA per character. In the case of Cristo Salvador, 3955 MA where performed, resulting in 8.4 MA per line, 0.76 MA per word, and 0.162 MA per character. In the case of Bentham, 2058 MA where performed, resulting

Table 18: Results for the computer assisted transcription without lexicon restrictions but with language restrictions (8-gram) with additional data for the Rodrigo corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (CER in second row) and CTC decoding (CER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$10.8\% \pm 0.3$	$10.7\% \pm 0.3$
	CER	$2.13\% \pm 0.06$	$2.11\% \pm 0.06$
	Oracle CER	0.36%	0.36%
CATTI	**CSR**	$0.66\% \pm 0.04$	$0.67\% \pm 0.04$
	CCR	0.031 ± 0.001	0.031 ± 0.001
	MA	8409	8401
	$\textbf{EFR}_{\textbf{WFST}}$	69.0%	68.2%
	$\textbf{EFR}_{\textbf{CTC}}$	62.6%	61.7%

Table 19: Results for the computer assisted transcription without lexicon restrictions but with language restrictions (7-gram) with additional data for the Cristo Salvador corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (WER in top row) and CTC decoding (WER in Table 8), respectively.

Experiment	**Measure**	**Training + 50% Validation**	**Training + Validation**
Lattice generation	**WER**	$29.9\% \pm 1.6$	$29.9\% \pm 1.6$
	CER	$9.52\% \pm 0.67$	$9.57\% \pm 0.72$
	Oracle CER	2.64%	2.66%
CATTI	**CSR**	$4.56\% \pm 0.52$	$4.55\% \pm 0.51$
	CCR	0.158 ± 0.013	0.158 ± 0.013
	MA	3880	3875
	$\textbf{EFR}_{\textbf{WFST}}$	52.1%	52.5%
	$\textbf{EFR}_{\textbf{CTC}}$	44.0%	44.1%

in 2.4 MA per line, 0.26 MA per word, and 0.053 MA per character. All these figures represent an improvement in the use of the assistive system with respect to those obtained when working with lexicon restrictions (at word level).

As it can be observed from the presented results, our computer assisted transcription system working at character-level (without lexicon restrictions) not only allows us to reduce significantly the number of characters to be corrected over the transcription given by the CTC decoding, but it also gets it with a minimal interaction with the user.

In order to see the influence of the training data size on the performance of the system, experiments with the same test set but training the language models (at char level) with the data described in Section 5.1 were performed.

The results for Rodrigo, Cristo Salvador, and Bentham are presented in Tables 18, 19, and 20, respectively. As happened in the previously presented

Table 20: Results for the computer assisted transcription without lexicon restrictions but with language restrictions (8-gram) with additional data for the Bentham corpus. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (CER in second row) and CTC decoding (CER in Table 8), respectively.

Experiment	Measure	Training + 50% Validation	Training + Validation	Training + Extra
Lattice generation	**WER**	$11.6\% \pm 0.8$	$11.6\% \pm 0.8$	$9.2\% \pm 0.7$
	CER	$3.07\% \pm 0.27$	$3.04\% \pm 0.28$	$2.65\% \pm 0.26$
	Oracle CER	1.02%	1.01%	0.86%
CATTI	**CSR**	$1.68\% \pm 0.20$	$1.67\% \pm 0.21$	$1.36\% \pm 0.19$
	CCR	0.051 ± 0.005	0.052 ± 0.005	0.041 ± 0.005
	MA	1980	1990	1589
	$\textbf{EFR}_{\textbf{WFST}}$	45.3%	45.1%	48.7%
	$\textbf{EFR}_{\textbf{CTC}}$	49.1%	49.3%	58.8%

cases, the use of additional data for training the models has a positive impact in recognition, although differences are only significant at word level (WER) when a huge set of additional data is used (last column of Table 20), but at character level (CER) differences are not significant with respect to the results obtained without additional data in any case. When the CTC decoding is considered, only in the case of the huge set of additional data differences are significant at both word and character level.

The impact in the oracle CER follows the same tendency, which is reflected as well in the assisted transcription results, where no significant differences are obtained when additional data is used. As happened when no additional data is used, in this case a positive effort reduction with respect to CTC decoding is obtained, and the larger the dataset used, the higher is this effort reduction.

6.5. Computer assisted transcription without lexicon or language restrictions

In this last experiment, character lattices were generated without lexicon or language restrictions (using zero-gram character language models). In the first part of Table 21 the quality of the generated character lattices is presented. In the case of Rodrigo, these lattices present a CER equal to $2.84\% \pm 0.06$, which represents a statistically significant deterioration over the result obtained in the previous experiment. However, the quality of the hypotheses contained in these lattices reach an oracle CER equal to 0.17%, which represents a relative improvement of 52.8% over the oracle CER obtained in the previous experiment.

The character lattices obtained for Cristo Salvador presented a CER equal to $8.34\% \pm 0.63$ and an oracle CER equal to 1.40%. In this case, both results are better than those obtained in the previous experiment with language restrictions.

Regarding Bentham, the character lattices generation gives us a CER equal to $3.80\% \pm 0.33$, which is worse than the results obtained in the previous experi-

Table 21: Computer assisted transcription without lexicon or language restrictions. EFR_{WFST} and EFR_{CTC} refer to effort reduction with respect to WFST lattice generation (CER in second row) and CTC decoding (CER in Table 8), respectively.

Experiment	Measure	Rodrigo	Cristo Salvador	Bentham
Lattice generation	**WER**	$18.4\% \pm 0.4$	$26.6\% \pm 1.5$	$15.9\% \pm 1.2$
	CER	$2.84\% \pm 0.06$	$8.34\% \pm 0.63$	$3.80\% \pm 0.33$
	Oracle CER	0.17%	1.40%	0.78%
CATTI	**CSR**	$0.49\% \pm 0.03$	$4.05\% \pm 0.47$	$1.55\% \pm 0.19$
	CCR	0.033 ± 0.001	0.139 ± 0.012	0.056 ± 0.005
	MA	9025	3427	2143
	EFR_{WFST}	82.7%	51.4%	59.2%
	EFR_{CTC}	72.0%	50.3%	53.0%

ment and the obtained by the CTC decoding. However, these lattices presented an oracle CER equal to 0.78%, which represents a relative improvement of 24.3% over the oracle CER obtained in the previous experiment.

Given the improvement in the oracle CER reached by the lattices generation without lexicon or language restrictions, a better performance of our computer assisted transcription system can be expected.

The obtained results for the CATTI experiments are presented in the second part of Table 21. A CSR equal to $0.49\% \pm 0.03$ for Rodrigo corpus was achieved. It represents 72.0% of significant human effort reduction over the transcription offered by the CTC decoding. A similar behaviour was observed for Cristo Salvador and Bentham. In the case of Cristo Salvador, a CSR equal to $4.05\% \pm 0.47$ was achieved, which represents a significant human effort reduction of 50.3% over the transcription offered by the CTC decoding. In the case of Bentham, a CSR equal to $1.55\% \pm 0.19$ was achieved, which represents a significant human effort reduction of 48.5% over the transcription offered by the CTC decoding.

Regarding the number of interactions, for Rodrigo 9025 MA where performed, resulting in 1.8 MA per line, 0.16 MA per word, and 0.033 MA per character. In the case of Cristo Salvador, 3427 MA where performed, resulting in 7.3 MA per line, 0.66 MA per word, and 0.139 MA per character. In the case of Bentham, 2143 MA where performed, resulting in 2.5 MA per line, 0.27 MA per word, and 0.056 MA per character.

As it can be observed from the presented results, our computer assisted transcription system working at character-level, without lexicon or language restrictions, provides a better estimated human effort reduction (EFR) than the obtained in the previous experiment at character-level with language restrictions obtained in all conditions, except for the use of the huge extra dataset in the Bentham corpus. However, only in the case of Rodrigo this improvement is statistically significant ($0.49\% \pm 0.03$ over $0.67\% \pm 0.04$). With respect to Mouse Actions, they increase in Rodrigo and Bentham (possible because of the large amount of training data, that makes character language models much more

Table 22: Summary of the estimated effort reduction (EFR) in the computer assisted transcription experiments over the transcriptions given by the CTC decoding when using only the training data for obtaining the lexicon and language models.

Restrictions		Rodrigo	Cristo Salvador	Bentham
Lexicon	**Language**		**EFR$_{\mathbf{CTC}}$**	
Yes	Yes	−47.6%	−57.6%	−25.2%
Yes	No	−42.9%	−64.7%	−45.4%
No	Yes	61.7%	43.7%	48.5%
No	No	72.0%	50.3%	53.0%

Table 23: Summary of the best estimated effort reduction (EFR) in the computer assisted transcription experiments over the transcriptions given by the CTC decoding when using the available additional data for obtaining the lexicon and language models.

Restrictions		Rodrigo	Cristo Salvador	Bentham
Lexicon	**Language**		**EFR$_{\mathbf{CTC}}$**	
Yes	Yes	−44.3%	−54.1%	27.8%
Yes	No	−38.8%	−64.6%	−39.2%
No	Yes	62.6%	44.1%	58.8%
No	No	72.0%	50.3%	53.0%

reliable) but they decrease for Cristo Salvador (which has much less training data).

6.6. Estimated effort reduction in the assisted transcription experiments

Table 22 presents a summary of the estimated effort reduction (EFR) in the computer assisted transcription experiments over the transcriptions given by the CTC decoding for the three corpora when using only the training data. It can be observed that no human effort reduction can be considered when working with lexical restrictions (working at word level). However, when working without lexicon restrictions, the estimated effort reduction reaches the 40%. Moreover, the estimated effort reduction increases with a minimum of 50% when working without lexicon or language restrictions.

When additional data is available, conclusions are similar. Table 23 presents a summary of the best estimated effort reduction (EFR) when this additional data is available. Lexicon restrictions are only better when a huge amount of data is available (i.e., the Bentham extra data corpus, that provides three times more data), but in any other case the results with these restrictions do not improve those obtained with no lexical restrictions. When language restrictions are present without lexical restrictions, similarly only huge amounts of data (the same case in the Bentham corpus) provide a better result in assisted transcription with respect to that obtained without any restriction, but at the cost of

Table 24: Summary of the additional Clicks of Mouse Actions performed in the computer assisted transcription experiments per word (WCR) and per character (CCR).

Restrictions		**Rodrigo**		**Cristo Salvador**		**Bentham**	
Lexicon	**Language**	**WCR**	**CCR**	**WCR**	**CCR**	**WCR**	**CCR**
Yes	Yes	0.31	0.065	0.90	0.192	0.37	0.077
Yes	No	0.33	0.069	0.95	0.203	0.39	0.082
No	Yes	0.16	0.032	0.76	0.162	0.26	0.053
No	No	0.16	0.033	0.66	0.139	0.27	0.056

needing much more data that, in general, it could be difficult and/or expensive to obtain.

The main reason for this behaviour is that, in general, having no restrictions provides a richer set of alternatives in the resulting lattice of the WFST decoding. Moreover, this set of alternatives is of better quality, as the oracle CER value demonstrates. For results with lexical restrictions, CER values are quite higher, thus even if oracle CER values were comparable, much effort would be necessary to obtain the final reference (as the EFR values demonstrate).

6.7. Mouse actions performed in the assisted transcription experiments

The cost of point signalling (by clicks or mouse actions) the erroneous words or characters in computer assisted transcription is usually so small as to be not worth considering. However, in this work we considered interesting to study how it is influenced by the lexicon and the language restrictions.

Table 24 presents a summary of the additional clicks or mouse actions performed in the computer assisted transcription experiments per word and per character for the three corpora. It can be observed that the lexicon restrictions (working at word level) increase considerably the additional mouse actions. However, they are slightly reduced when using the language restrictions (except for the Cristo Salvador corpus, possibly because of the scarce data).

The number of MA demonstrates that, when no lexical restrictions appear, less actions are necessary to get the final reference. In the case of absence of language model when no lexical model is present, the number of actions is comparable to those of having lexical model, which means that the depth of the best solutions in the lattices are similar, but in the case of no restrictions at all, they are better (lower oracle CER), which implies a higher effort reduction.

7. Conclusions

State-of-the-art Handwritten Text Recognition systems based on Convolutional and Recurrent Neural Networks provide high quality draft transcriptions by CTC decoding. However, it is still necessary to supervise them by professional transcribers.

In this work we have studied the influence of the lexicon and language restrictions on computer assisted transcription for reducing the human transcription effort. From the experimentation with three different historical manuscripts, we concluded that our interactive approach adds an additional reduction to the required human effort over the best transcription provided by the CTC decoding when working at character-level, i.e. without lexicon restrictions.

Working at character level, independently of using language model or not, allows our interactive approach to reduce significantly the human transcription effort. However, the best results were obtained when working without language restrictions, except in the case were a huge training set for the language model is available, which is not common in historical manuscripts. Concretely, the transcriptions provided by the CTC decoding present a CER equal to 1.75%, 8.15%, and 3.30% for each one of the three manuscripts used on the experimentation. When using our assistive transcription system only 0.49%, 4.05%, and 1.55% of the characters, for each manuscript, has to be effectively corrected by the human transcriber, allowing to reduce the human effort in more than a 50%. This behaviour can be attributed to the richer set of alternatives provided in the lattice that results from the WFST decoding on the CTC results, which allows to find the correct transcription with less effort than using the postedition approach.

We also verified that adding more information during the training of the lexical and language models allows us to obtain better results. However, in this case this improvement (a CSR equal to 1.36% for Bentham) is not statistically significant compared to the results obtained without language restrictions.

Regarding the number of additional mouse actions, working without lexicon restrictions reduces this number considerably. However, an additional reduction may be achieved when working with the language restrictions imposed by a character language model trained with abundant data.

Future work lines include the exploration of new assistive and interactive strategies, the use of multimodal interaction, and the experimentation with other datasets.

ACKNOWLEDGMENTS

Work partially supported by the BBVA Foundation through the 2017–2018 Digital Humanities research grant “Carabela” and by the Ministerio de Ciencia/AEI/FEDER/EU through the MIRANDA-DocTIUM project (RTI2018-095 645-B-C22).

References

[1] C. D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[2] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, M. Stolz, Automatic Transcription of Handwritten Medieval Documents,

in: Proceedings of the 15th International Conference on Virtual Systems and Multimedia (VSMM '09), 2009, pp. 137–142. doi:10.1109/VSMM.2009.26.

[3] V. Romero Gómez, Multimodal Interactive Transcription of Handwritten Text Images, Ph.D. thesis, Universitat Politècnica de València (2010).

[4] T. Bluche, Deep Neural Networks for Large Vocabulary Handwritten Text Recognition, Ph.D. thesis, Université Paris Sud-Paris XI (2015).

[5] T. Wang, D. J. Wu, A. Coates, A. Y. Ng, End-to-End Text Recognition with Convolutional Neural Networks, in: Proceedings of the 21tst International Conference on Pattern Recognition (ICPR 2012), 2012, pp. 3304–3308.

[6] N. Serrano, F. Castro, A. Juan, The RODRIGO Database, in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), 2010, pp. 2709–2712.
URL http://aclweb.org/anthology/L10-1330

[7] E. Granell, E. Chammas, L. Likforman-Sulem, C. D. Martínez-Hinarejos, C. Mokbel, B.-I. Cîrstea, Transcription of Spanish Historical Handwritten Documents with Deep Neural Networks, Journal of Imaging 4 (1) (2018) 15.

[8] V. Romero, A. H. Toselli, L. Rodríguez, E. Vidal, Computer Assisted Transcription for Ancient Text Images, in: M. Kamel, A. Campilho (Eds.), Image Analysis and Recognition (ICIAR 2007), Vol. 4633 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 1182–1193. doi:10.1007/978-3-540-74260-9_105.

[9] J. A. Sánchez, V. Romero, A. H. Toselli, E. Vidal, ICFHR2014 competition on handwritten text recognition on tranScriptorium datasets (HTRtS), in: International Conference on Frontiers in Handwriting Recognition, 2014, pp. 181–186.

[10] J. Schmidhuber, Deep learning in neural networks: An overview, Neural networks 61 (2015) 85–117.

[11] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[12] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, K. Zieba, End to end learning for self-driving cars, arXiv preprint arXiv:1604.07316.

[13] H. Xu, Y. Gao, F. Yu, T. Darrell, End-to-end learning of driving models from large-scale video datasets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2174–2182.

[14] S. Yeung, O. Russakovsky, G. Mori, L. Fei-Fei, End-to-end learning of action detection from frame glimpses in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2678–2687.

[15] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE Journal of Selected Topics in Signal Processing 11 (8) (2017) 1301–1309.

[16] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Transactions on Image Processing 26 (7) (2017) 3492–3506.

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik, End-to-end recovery of human shape and pose, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7122–7131.

[18] A. Graves, J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks, in: Advances in neural information processing systems, 2009, pp. 545–552.

[19] J. Puigcerver, Are multidimensional recurrent layers really necessary for handwritten text recognition?, in: International Conference on Document Analysis and Recognition, Vol. 01, 2017, pp. 67–72.

[20] C. Oprean, L. Likforman-Sulem, A. Popescu, C. Mokbel, Using the web to create dynamic dictionaries in handwritten out-of-vocabulary word recognition, in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, IEEE, 2013, pp. 989–993.

[21] P. Kumar, R. Saini, P. P. Roy, U. Pal, A lexicon-free approach for 3d handwriting recognition using classifier combination, Pattern Recognition Letters 103 (2018) 1–7.

[22] A. Ahmed, Y. Hifny, K. Shaalan, S. Toral, End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks, in: Computational Linguistics, Speech And Image Processing For Arabic Language, Vol. 4 of Series on Language Processing, Pattern Recognition, and Intelligent Systems, World Scientific, 2018, Ch. 11, pp. 231–248.

[23] A. Fischer, Handwriting Recognition in Historical Documents, Ph.D. thesis, University of Bern (2012).

[24] V. Frinken, A. Fischer, C.-D. Martínez-Hinarejos, Handwriting Recognition in Historical Documents using Very Large Vocabularies, in: Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing (HIP '13), 2013, pp. 67–72.

[25] V. Romero, V. Bosch, C. Hernández, E. Vidal, J. A. Sánchez, A historical document handwriting transcription end-to-end system, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer, 2017, pp. 149–157.

[26] J. I. Toledo, M. Carbonell, A. Fornés, J. Lladós, Information extraction from historical handwritten document images with a context-aware neural model, Pattern Recognition 86 (2019) 27–36.

[27] J. Calvo-Zaragoza, J. Oncina, An efficient approach for interactive sequential pattern recognition, Pattern Recognition 64 (2017) 295–304.

[28] A. H. Toselli, E. Vidal, F. Casacuberta, Multimodal Interactive Pattern Recognition and Applications, Springer Science & Business Media, 2011.

[29] N. Serrano, A. Giménez, J. Civera, A. Sanchis, A. Juan, Interactive handwriting recognition with limited user effort, International Journal on Document Analysis and Recognition (IJDAR) 17 (1) (2014) 47–59.

[30] E. Granell, C.-D. Martínez-Hinarejos, V. Romero, Improving transcription of manuscripts with multimodality and interaction, in: Proceeding of the IberSPEECH 2018, 2018, pp. 92–96.

[31] M. Adda-Decker, L. Lamel, The Use of Lexica in Automatic Speech Recognition, in: Lexicon Development for Speech and Language Processing, Text, Speech and Language Technology, Springer, 2000, pp. 235–266.

[32] U.-V. Marti, H. Bunke, 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 (01) (2001) 65–90.

[33] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, CoRR abs/1507.05717. arXiv:1507.05717.

[34] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 855–868.

[35] B. Moysset, T. Bluche, M. Knibbe, M. F. Benzeghiba, R. Messina, J. Louradour, C. Kermorvant, The A2iA multi-lingual text recognition system at the second Maurdor evaluation, in: International Conference on Frontiers in Handwriting Recognition, 2014, pp. 297–302.

[36] T. Bluche, H. Ney, C. Kermorvant, The LIMSI/A2iA Handwriting Recognition Systems for the HTRtS Contest, in: International Conference on Document Analysis and Recognition, 2015, pp. 448–452.

[37] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural networks for machine learning 4 (2).

[38] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: International Conference on Machine Learning, 2006, pp. 369–376.

[39] V. Pham, C. Kermorvant, J. Louradour, Dropout improves recurrent neural networks for handwriting recognition, CoRR abs/1312.4569. arXiv:1312.4569.

[40] A. Ljolje, F. Pereira, M. Riley, Efficient General Lattice Generation and Rescoring, in: Proceedings of the 6$^{\text{th}}$ European Conference on Speech Communication and Technology (Eurospeech'99), 1999.

[41] L. Rodríguez, F. Casacuberta, E. Vidal, Computer Assisted Transcription of Speech, in: J. Martí, J. M. B. A. M. Mendonça, J. Serrat (Eds.), Pattern Recognition and Image Analysis (IbPRIA 2007), Vol. 4477 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, pp. 241–248.

[42] A. Toselli, V. Romero, L. Rodríguez, E. Vidal, Computer Assisted Transcription of Handwritten Text Images, in: Proceedings of the 9$^{\text{th}}$ International Conference on Document Analysis and Recognition (ICDAR'07), Vol. 2, 2007, pp. 944–948.

[43] V. Romero, A. H. Toselli, E. Vidal, Using Mouse Feedback in Computer Assisted Transcription of Handwritten Text Images, in: Proceedings of the 10$^{\text{th}}$ International Conference on Document Analysis and Recognition (ICDAR'09), 2009, pp. 96–100.

[44] V. Romero, A. H. Toselli, E. Vidal, Multimodal Interactive Handwritten Text Transcription, Vol. 80 of Machine Perception and Artificial Intelligence, World Scientific Publishing, 2012.

[45] PRHLT, Pattern Recognition and Human Language Technology Research Center, accessed on 5 June 2018 (2018).
URL https://www.prhlt.upv.es

[46] J. Snchez, A. Toselli, V. Romero, E. Vidal, ICDAR 2015 Competition HTRtS: Handwritten Text Recognition on the tranScriptorium Dataset (Jan. 2017). doi:10.5281/zenodo.248733.
URL https://doi.org/10.5281/zenodo.248733

[47] P. Roeder, Adapting the RWTH-OCR handwriting recognition system to French handwriting, Ph.D. thesis, RWTH Aachen University, Aachen. Germany (2009).

[48] D. S. Bloomberg, G. E. Kopec, L. Dasari, Measuring document image skew and orientation, SPIE 2422 (1995) 302–316.

[49] R. Buse, Z. Liu, T. Caelli, A structural and relational approach to handwritten word recognition, IEEE Transactions on Systems, Man, and Cybernetics, Part B 27 (5) (1997) 847–861.

[50] A. Vinciarelli, J. Luettin, A new normalization technique for cursive handwritten words, Pattern Recognition Letters 22 (9) (2001) 1043 – 1050.

[51] M. Villegas, V. Romero, J. A. Sánchez, On the modification of binarization algorithms to retain grayscale information for handwritten text recognition, in: R. Paredes, J. Cardoso, X. Pardo (Eds.), Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings, 2015, pp. 208–215.

[52] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.

[53] R. Kneser, H. Ney, Improved backing-off for M-gram language modeling, in: Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'95), Vol. 1, 1995, pp. 181–184.

[54] A. Stolcke, SRILM-an extensible language modeling toolkit., in: Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech), 2002, pp. 901–904.

[55] Y. Miao, M. Gowayyed, F. Metze, EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 167–174.

[56] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (8) (1966) 707–710.

[57] M. Bisani, H. Ney, Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04), Vol. 1, 2004, pp. 409–412.