UNIVERSIDAD POLITÉCNICA DE VALENCIA
INSTITUTO TECNOLÓGICO DE INFORMÁTICA

Confidence Measures in Interactive Handwritten Text Transcription

Master Thesis

presented by
Lionel Manuel Tarazón Alcocer

supervised by
Dr. Alberto Sanchís Navarro
and Dr. Alfons Juan Císcar

November 30, 2009

Acknowledgments

Work supported by the EC (FEDER/FSE) and the Spanish MCE/MICINN under the MIPRCV "Consolider Ingenio 2010" programme (CSD2007-00018), the iTransDoc project (TIN2006-15694-CO2-01), the FPU scholarship AP2007-02867 and the UPV grant 20080033.

CONTENTS

CHAPTER 1

Introduction

There are huge historical document collections residing in libraries, museums and archives that are currently being digitized for preservation purposes and to make them available worldwide through on-line digital libraries. The main objective is not to simply provide access to raw images of digitized documents, but to annotate them with their real informative content and with text transcriptions.

Handwritten text recognition technologies can be used to help the human transcriber in the annotation task. Documents could be recognized off-line first, and then given to a human expert for revision of incorrect parts. A more effective approach is to use interactive computer-assisted transcription tools. This way, both text recognition and human revision can be carried out interactively through a user-friendly interface. However, state-of-the-art technologies are still far from perfect, and thus post-editing automatically generated output is not clearly better than simply ignoring it.

A confidence masure is an indicator of how confident the system is that the units of the recognition hypothesis, usually words, are correct. Confidence measuring can be used to automatically validate or reject the system hypothesis, and act accordingly. This technology has been widely studied and used in speech recognition. However, its usefulness in assisted transcription of handwritten text remains almost unexplored.

The main objective of this work is to invetigate how confidence measures can be applied in interactive transcription of handwritten text to support the human transcriber. For this purpose, an interactive transcription interface with confidence measuring has been developed and evaluated. This work has been carried out within the framework of the Spanish research project "Interactive Transcription and Translation of Old Text Documents (iTransDoc)" [1].

More specifically, the contributions described in this work are the following:

GIDOC: Gimp-based Interactive transcription of old text DOCuments.
GIDOC is a user-friendly interactive transcription prototype in which word-graph based confidence measures have been developed to support and guide the human transcriber in the annotation task. This work has led to one publication submitted in international conference:

- **WEBIST-2010:** N. Serrano, **L. Tarazón**, D. Pérez, O. Ramos-Terrades and A. Juan. The GiDOC Prototype. *Procceding of 6th International Conference on Web Information Systems and Technologies (WEBIST 2010).* Valencia (Spain). April 2010. (Submitted)

Application of Confidence Measures in Interactive Transcription.
Making use of the system prototype described above, confidence measures have been evaluated in an assisted transcription environment. It has been proved they can be applied to give support for error correction and detection, reducing drastically the supervision effort. This work has led to one publication in international conference:

- **ICIAP-2009: L. Tarazón**, D. Pérez, N. Serrano, V. Alabau, O. Ramos Terrades, A. Sanchis and A. Juan. Confidence Measures for Error Correction in Interactive Transcription of Handwritten Text. *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP 2009).* Vietri sul Mare (Italy). September 2009.

GERMANA: Preparation of old text document dabase
In this work, we have collaborated in the preparation of a database of old text documents: GERMANA, a 764-page Spanish manuscript from 1891. This database is described in one article in international conference:

- **ICDAR-2009:** D. Pérez, **L. Tarazón**, N. Serrano, F. Castro, O. Ramos and A. Juan. The GERMANA database. *Proccedings of the 10th International Conference on Document Analysis and Recogntion (ICDAR 2009).* Barcelona (Spain). July 2009.

The contributions described above are the result of a collaborative work involving other authors and, in particular, authors that are also presenting their Master's Theses. Nevertheless, the author of this Thesis should be considered as the leading author of the work reported in the article denoted above as ICIAP-2009. Accordingly, the work reported in this article is described with full detail in Chapter 3. The work reported in the remaining articles is briefly described in Chapters 2 (WEBIST-2010) and 3.1.2 (ICDAR-2009). The reader is referred to the Master's Thesis by D. Pérez [2] for more details on the ICDAR-2009 article. Also, it must be noted that the basic GIDOC prototype has been developed by the author of this Thesis together with D. Pérez and N. Serrano, at the same level of dedication effort, and thus all these three authors should be considered as leading authors of the WEBIST-2010 article.

For the sake of clarity, the correspondence between Chapters and articles is summarized in Table 1.1, together with conference quality indicators (CORE rank).

Article	Status	Quality indicators	Contribution	Chapter
ICIAP-2009	Published	CORE A	Leading author	3
ICDAR-2009	Published	CORE A	Co-author	3.1.2
WEBIST-2010	Submitted	CORE C	Leading author	2

Table 1.1: Articles generated from the work described in this document.

CHAPTER 2

GIDOC: Interactive Handwritten Text Transcription

There are huge historical document collections residing in libraries, museums and archives that are currently being digitised for preservation purposes and to make them available worldwide through large, on-line digital libraries. The main objective, however, is not to simply provide access to raw images of digitised documents, but to annotate them with their real informative content and, in particular, with text transcriptions.

A direct approach to deal with transcriptions is to first apply an off-line recognition system to all documents, and then manually supervise the system output and correct it when necessary. Clearly, this approach is comfortable for the human supervisor only if minor corrections are needed; otherwise, it is not any more comfortable than simply ignoring system output. This second, uncomfortable case is precisely the most common case in collections of historical documents with handwritten text, or even with old-style printed text. Although state-of-the-art technologies for automatic text transcription are rapidly advancing in recent years, they are still far from achieving good enough results [3, 4].

A more effective approach to transcribe old text documents is to follow an interactive-predictive paradigm in which both, the system is guided by the human supervisor, and the supervisor is assisted by the system to complete the transcription task as efficiently as possible. A pioneering application of this computer-assisted transcription (CAT) approach can be found in the FP4 European research project DEBORA (Digital AccEss to BOoks of the RenAissance) [5]. An important task within this project was to efficiently transcribe printed books of the Renaissance, which are characterised by their complex page layout, rare fonts and generally unused typography.

However, conventional OCR systems were not usable at all, and thus a character-level CAT system was developed by which only 2% of all characters in a book has to be manually entered to complete its transcription. This successful application of the CAT idea has been later extended to interactive, user-driven layout analysis in the AGORA system [6].

In this work, a CAT system prototype is described for handwritten text in old documents, which implements ideas recently developed within the Spanish research project iDoc [7]. It is a first attempt to provide integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. More details on this characteristics are given in Chapters 2.2, 2.3 and 2.4.

Furthermore, state-of-the-art hypothesis verification technology has been implemented. As far as we know, GIDOC is the first interactive transcription prototype with word-graph based confidence measures. Both the system is guided by the user, and the user is guided by the system, aiming at giving support for error detection and correction, and thus reducing the transcription effort. More details on this characteristic are given in Chapter 2.5.

The reader is referred to [8] and [2] for more (technical and experimental) details on different GIDOC parts.

2.1 System overview

Clearly, it is a programming challenge to develop a usable, friendly Graphical User Interface (GUI) for such a prototype, and thus we decided not to start from scratch, but to build it on top of the well-known GNU Image Manipulation Program (GIMP) [9]. Apart from its high-end user interface, GIMP gives us for free many desired prototype features such as a large collection of image conversion drivers and low-level processing routines, an scripting language to automate repetitive tasks, an API for installation of user-defined plug-ins, etc. Indeed, the prototype, which will be referred to as GIDOC (Gimp-based Interactive transcription of old text DOCuments), is implemented as a set of GIMP plug-ins.

As GIMP, GIDOC is licensed under the GNU General Public License, and it can be downloaded from [7]. To run GIDOC, we must first run GIMP and open a document image. GIMP will come up with its high-end user interface, which is often configured to only show the main toolbox (with docked dialogs) and an image window. GIDOC can be accessed from the menubar of the image window (see Figure 2.1). For brevity, the reader is referred to the GIMP website for more information on it [9].

As shown in Figure 2.1, the GIDOC includes six entries: *Advanced options, 0: Preferences, 1: Block Detection, 2: Line Detection, 3: HTK Training,* and *4: Transcription. Advanced options* is a second-level menu where experimental features of GIDOC are grouped. *Preferences* opens a dialog to configure global options, as well as more specific options for preprocessing, training, recognition and confidence measures. These more specific options are discussed below together with GIDOC menu entries after *Preferences.* Regarding global options, a transcription task name has to be defined so as to create a directory where task-specific information will be saved.

Similarly, a directory name has to be defined for GIDOC to find the task document images. It is worth noting that these images are assumed to be in the GIMP's native XCF file type. This file type allows us to fully exploit all GIMP capabilities and it is general enough to include all sorts of annotations such as block and line locations, transcriptions, etc. Finally, a third global option is included to "lock" transcriptions.

2.2 Line detection

During its development, GIDOC has been mainly tested on a old book in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. An example of these pages is shown in Figure 2.1. From this example, it becomes clear that block detection (layout analysis) and line detection within textual blocks are not especially difficult. Indeed, the *Block Detection* entry in the GIDOC menu has been for now implemented to simply copy the blocks of the preceding image to the current image. This works quite well in the case of the old book mentioned above, due to its homogeneous page layout structure. As can be observed in Figure 2.1, in this case block detection amounts to only copy a minimal enclosing frame including all text lines. This frame is technically a closed path with four handlers at the "corners" which can be graphically adjusted by the user.

Given a textual block, the *Line Detection* entry in the GIDOC menu detects all its text baselines, which are marked as straight paths. The result can be clearly observed in the example of Figure 2.1. As with the block, each baseline has handlers to graphically correct its position. It must be noted, however, that the baseline detection method implemented works quite well, at least in pages like that of the example. It is a rather standard projection-based method [10]. First, horizontally-averaged pixel values or black/white transitions are projected vertically. Then, the resulting vertical histogram is smoothed and analysed so as to locate baselines accurately. Two preprocessing options are included in *Preferences*, first, to decide on the histogram type (pixel values or black/white transitions), and second, to limit the maximum number of baselines to be found.

2.3 Training

The CAT system implemented in GIDOC is based on standard techniques and tools for handwritten text preprocessing and feature extraction, Hidden Markov Model (HMM)-based image modelling, and language modelling [3, 11]. Handwritten text preprocessing applies image denoising, deslanting and vertical size normalisation to a given text (line) image. It can be configured through preprocessing options in *Preferences.* There is an option to define an alternate, custom procedure, and also two options to define (bounds for) the locations of the upper and lower lines, with respect to the baseline.

Feature extraction for HMM modelling consists in transforming the preprocessed image into a *sequence of (fixed-dimension) feature vectors.* There are two, well-known feature extraction methods available in GIDOC. The default, PRHLT method first

divides the preprocessed image into a grid of square cells whose size is a small fraction of the image height (e.g. 1/20). Then, each cell its characterised by its normalised grey level and, optionally, by its vertical and horizontal grey-level derivatives. See [3] for more details. The alternative, FKI method moves a single-column window left-to-right over the image, and extracts nine geometrical features at each position of the window [4]. The desired method can be selected in *Preferences.*

HMM image modelling is carried out with the well-known and freely available *Hidden Markov Model Toolkit (HTK).* [12]. Similarly, language modelling is implemented through the *SRI Language Modelling Toolkit (SRILM),* which is available under an open source community license [13]. Both toolkits should be made available to GIDOC for the *HTK Training* entry in the GIDOC menu to properly work.

HTK Training reads the directory of task document images and, for each image, it extracts all its transcribed text lines, if any, together with their corresponding line images. An n-gram language model is built from the transcriptions, using a SRILM command defined in the training chapter of *Preferences.* The default command generates a bigram language model with Knesser-Ney discounting. The resulting language model and vocabulary are saved into files also defined in the training chapter of *Preferences.* On the other hand, extracted line images are preprocessed and transformed into sequences of feature vectors so as to train, using their corresponding transcriptions and HTK, continuous density (Gaussian) left-to-right HMMs at character level. The list of symbols and trained HMMs are saved into files given in the training chapter of *Preferences.* This chapter also includes an option to decide on the number HMM training iterations (4 by default), and a flag to train the HMMs from scratch or re-estimate previously trained HMM.

2.4 Transcription

The *Transcription* entry in the GIDOC menu opens the GIDOC interactive transcription dialog (see Figure 2.2). It consists of two main chapters: the image chapter, in the middle part, and the transcription chapter, in the bottom part. An odd number of text line images are displayed in the image chapter together with their transcriptions, if available, in separate editable text boxes within the transcription chapter. The *current* line to be transcribed or simply supervised is selected by placing the edit cursor in the appropriate editable box. Its corresponding baseline is emphasised (in blue colour) and, whenever possible, GIDOC shifts line images and their transcriptions so as to display the current line in the central part of both the image and transcription chapter. It is assumed that the user transcribes or supervises text lines, from top to bottom, by entering text and moving the edit cursor with the arrow keys or the mouse. However, it is possible for the user to choose any order desired.

Each editable text box has a button attached to its left, which is labelled with its corresponding line number. By clicking on it, its associated line image is extracted, preprocessed, transformed into a sequence of feature vectors, and Viterbi-decoded using HTK and the models trained with *HTK training.* The *Grammar Scale Factor (GSF)* and *Word Insertion Penalty (WIP)* values to properly combine the HMM

and language models are defined in the recognition chapter of *Preferences*. Also in it there is an option to adjust the *Beam* value and thus the computational cost to perform Viterbi decoding. In this way, it is not needed to enter the complete transcription of the current line, but hopefully only minor corrections to the decoded output. Clearly, this is only possible if, first, text lines are correctly detected and, second, the HMM and language models are adequately trained, from a sufficiently large amount of training data. Therefore, it is assumed that transcription is carried out manually in early stages of a transcription task, and then is assisted as described here.

Apart from the image and transcription chapters, the dialog shown in Figure 2.2 includes a number of controls in the top part, as well as self-explanatory buttons under the transcription chapter. Regarding the controls in the top part, note that they allow the user to select the current block of the image to transcribe, the current line, the number of lines to show, etc. It is not difficult to configure the dialog so as to fulfil the task needs and user preferences.

2.5 Transcription Guided by Confidence Measures

As said in the Introduction, state-of-the-art technologies for automatic text transcription are still far from achieving good enough results, and thus considerable human effort has to be put into locating and editing systems errors, even with the advanced interactive transcription dialog described above. In order to reduce this effort, GIDOC has also an option to calculate *confidence measures* on recognised words and mark those words which have been recognised with low confidence, supporting the user in the detection and correction of recognition errors. Chapter 2.5.1 describes in detail how confidence measures are obtained.

Marked words have to be revised by the human user, who has to supervise the recognized words and correct the incorrect parts. Clearly, if only a few words are marked, then the effort required to supervise them is smaller than that of supervising all words, though a small number of transcription errors has to be tolerated since (hopefully minor) recognition errors may go unnoticed to confidence measures. On the contrary, if most words are marked, then it may not pay off to locate and edit system errors. Instead, it might be better to ignore system output and transcribe the whole text line manually.

In order to use this functionality, hypothesis verification has to be activated in the GIDOC preferences. Also, a appropiate threshold has to be configured (a value between 0 and 100). A low threshold will result in a optimistic system where words are rarely rejected. On the contrary, a high threshold will make the system to reject the majority of the words. Deciding which threshold to use is a very important task. If a adequate value is configured, confidence measures can efficiently guide the user in the error detection and correction, resulting in a low error rates and thus reducing drastically the transcription effort. Chapter 3 explains in detail the experiments carried out to evaluate the cost-effort impact of using confidence measures in interactive transcription.

An example of interactive transcription guided by confidence measures is shown in Figure 2.3. In this example, there are only two words for which the system is not highly confident. They are enclosed with red frames in the image chapter and also written in red colour in the transcription chapter. This double marking helps the user know eaisily which incorrect words are associated to where in the image, and thus verify eaisily if words are truly incorrect or not. In this case, it suffices to insert ", ^{260}" after "venida" and substitute "de" by "26" to obtain the correct transcription.

2.5.1 Word Posterior Confidence Estimation

In this chapter we briefly explain the estimation of word-level confidence measures. Taking advantage of the use of standard speech technology by GIDOC, we have adopted a method that has proved to be very useful for confidence estimation in speech recognition. This method was proposed in [14] and uses posterior word probabilities computed from word graphs as confidence measures.

A word graph G is a directed, acyclic, weighted graph. The nodes correspond to discrete points in space. The edges are triplets $[w, s, e]$, where w is the hypothesized word from node s to node e. The weights are the recognition scores associated to the word graph edges. Any path from the initial to the final node forms a hypothesis $\boldsymbol{f}_1^J$.

Given the observations $\boldsymbol{x}_1^T$, the posterior probability for a specific word (edge) $[w, s, e]$ can be computed by summing up the posterior probabilities of all hypotheses of the word graph containing the edge $[w, s, e]$:

$$P([w,s,e] \mid \boldsymbol{x}_1^T) = \frac{1}{P(\boldsymbol{x}_1^T)} \sum_{\substack{\boldsymbol{f}_1^J \in G : \\ \exists [w',s',e'] : \\ w'=w, s'=s, e'=e}} P(\boldsymbol{f}_1^J, [w,s,e], \boldsymbol{x}_1^T) \tag{2.1}$$

The probability of the sequence of observations $P(\boldsymbol{x}_1^T)$ can be computed by summing up the posterior probabilities of all word graph hypothesis:

$$P(\boldsymbol{x}_1^T) = \sum_{\boldsymbol{f}_1^J \in G} P(\boldsymbol{f}_1^J, \boldsymbol{x}_1^T)$$

The posterior probability defined in Eq. 2.1 does not perform well because a word w can occur in slightly different starting and ending points. This effect is represented in the word graph by different word edges and the posterior probability mass of the word is scattered among the different word segmentations (see Fig. 2.4).

To deal with this problem, we have considered a solution proposed in [14]. Given a specific word (edge) $[w, s, e]$ and a specific point in time $t \in [s, e]$, we compute the posterior probability of the word w at time t by summing up the posterior probabilities of the word graph edges $[w, s', e']$ with identical word w and for which t is within the interval time $[s', e']$:

$$P_t([w,s,e] \mid \boldsymbol{x}_1^T) = \sum_{t \in [s',e']} P([w,s',e'] \mid \boldsymbol{x}_1^T) \tag{2.2}$$

Based on Eq. 2.2, the posterior probability for a specific word $[w, s, e]$ is computed as the maximum of the frame time posterior probabilities:

$$P([w, s, e] \mid \boldsymbol{x}_1^T) = \max_{s \leq t \leq e} P_t([w, s, e] \mid \boldsymbol{x}_1^T) \tag{2.3}$$

The probability computed on Eq. 2.3 is in the interval $[0, 1]$ since, by definition, the sum of the word posterior probabilities for a specific point in time must sum to one (see Fig. 2.4). The posterior probabilities calculated as Eq. 2.3 are used as word confidence measures (see Fig. 2.5).

Using these posterior probabilities, a word is proposed to the human supervisor (see figure 2.3) if $P([w, s, e] \mid \boldsymbol{x}_1^T)$ is lower that a certain threshold τ (see chapter 3.2).

2.6 Conclusions

A computer-assisted transcription prototype called GIDOC has been described for handwritten text in old documents. GIDOC is a first attempt to provide integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription. It is build on top of GIMP, and uses standard techniques and tools for handwritten text preprocessing and feature extraction, HMM-based image modelling, and language modelling.

Furthermore, the GIDOC prototype is the first attempt of using hypothesis verification technology in interactive transcription. It has been shown confidence measures can be applied in computer-assisted environments to assist, support and guide the user in the transcription process.

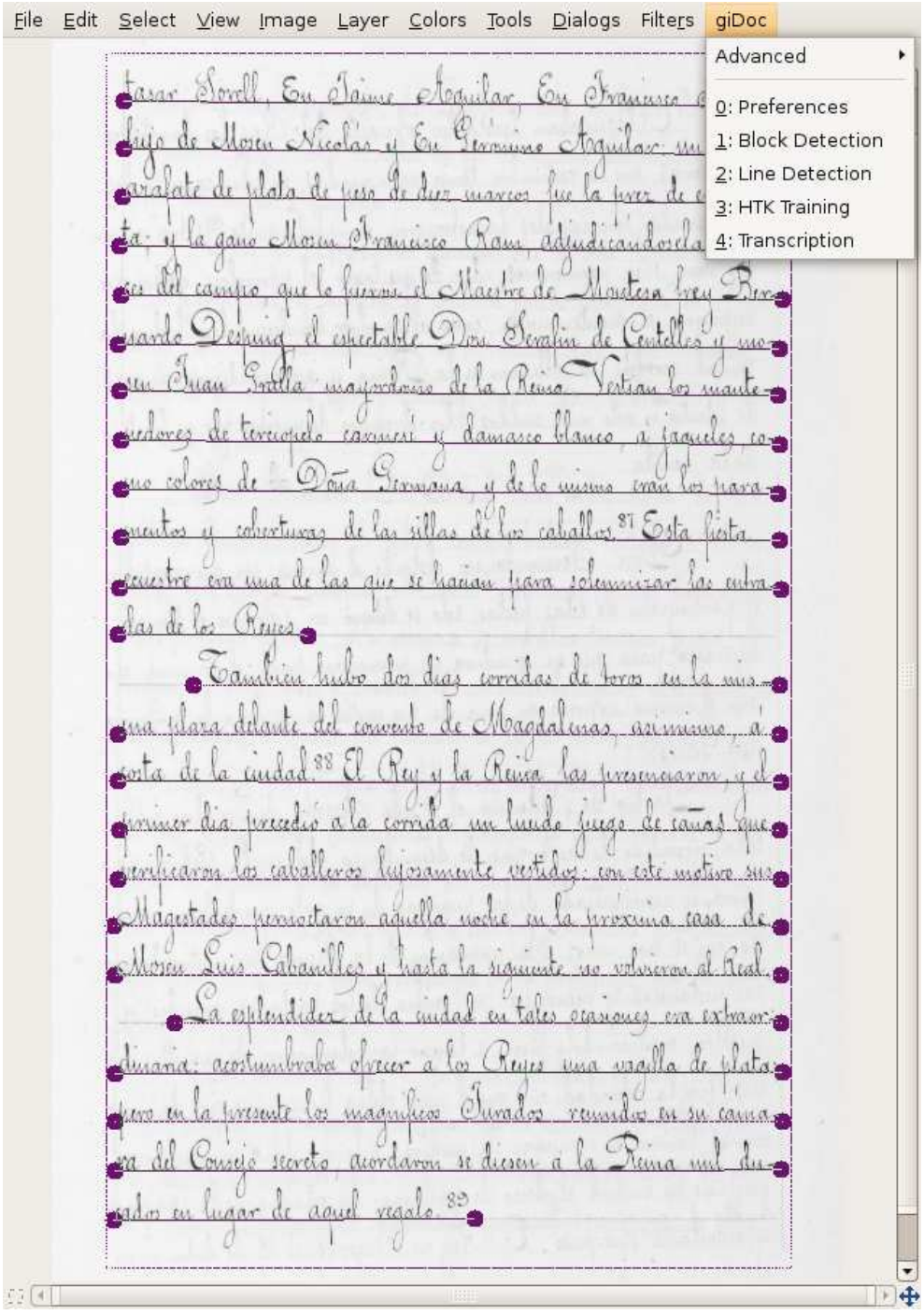


Figure 2.1: Image window and GIDOC menu.

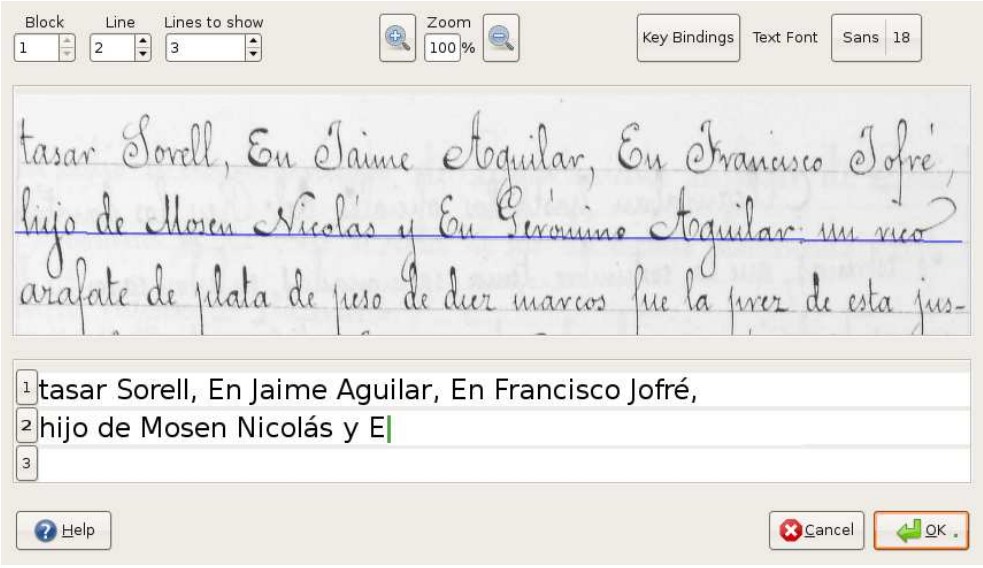


Figure 2.2: Interactive transcription dialog.

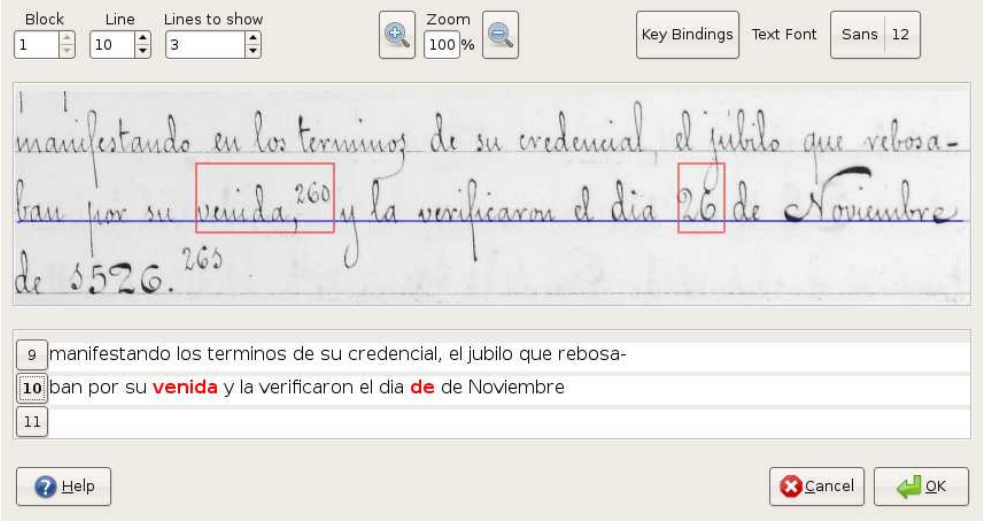


Figure 2.3: Interactive transcription guided by confidence measures.

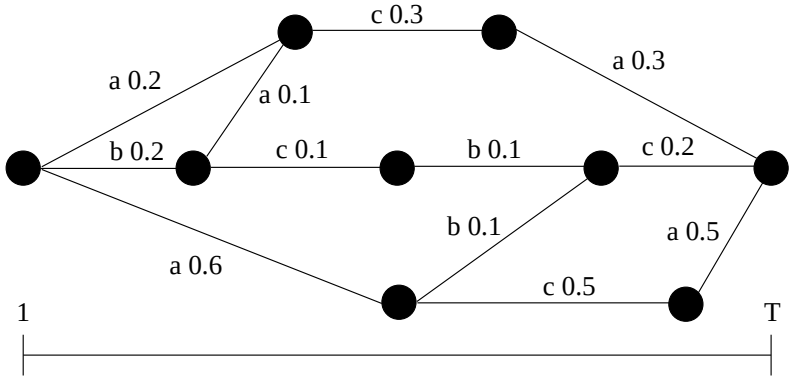


Figure 2.4: Word graph with the word posterior probabilities computed as Eq. 2.1.

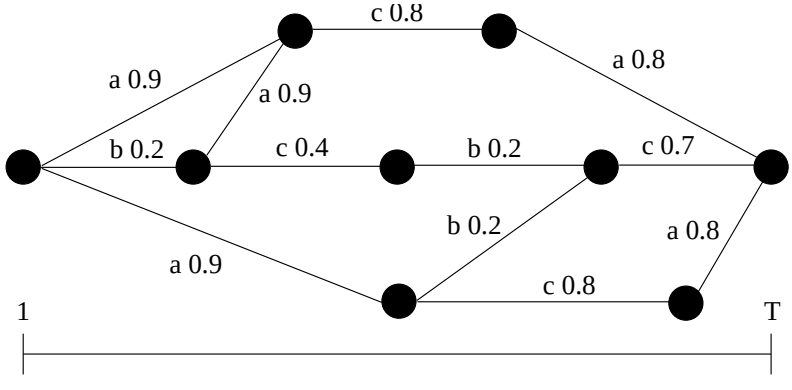


Figure 2.5: Word graph with the word posterior probabilities computed as Eq. 2.3.

CHAPTER 3

Confidence Measures in Interactive Transcription

Using confidence measures for offline handwritten text line recognition is not new, and it is well known it can be used efficiently for hypothesis verification (see [15] and the references therein). Nevertheless, small efforts have been made in studying how they can be applied in an interactive environment to support the human supervisor.

This chapter explains in detail how standard speech confidence measures technology (see [14, 16]) can be used in interactive handwritten text transcription to give support for error location and correction. For instance, if a small number of transcription errors can be tolerated for the sake of efficiency, then he/she might validate the system output after only checking those (few) words, if any, for which the system is not highly confident. On the contrary, if at a first glance no significant portion of the text line seems to be correctly recognised, then he/she might ignore system output and transcribe the whole text line manually.

By contrast to previous works [15], here confidence measures are based on *posterior word probabilities* estimated from *word graphs* since, at least in the case of speech recognition, experimental evidence clearly shows that they outperform alternative confidence measures, and even posterior word probabilities estimated from *N-best lists* [14, 16].

3.1 Corpora

In order to evaluate experimentally de use of confidence measures in interactive transcription of handwritten text two copora were used: The IAM-DB 3.0 database [17] and the GERMANA database [18]. The first one is well known by the handwritten

	IAM 20K Voc.			GERMANA 9.4K Voc.		
	Train	Val.	Test	Train	Val.	Test
Pages	747	116	336	94	36	38
Lines	6.161	920	2.781	2.131	811	811
Run.Words (K)	53.8	8.7	25.4	23.7	9.4	9.1
out-of-voc (%)	-	6.6	6.4	-	17.5	18.6

Table 3.1: Basic statistics of IAM and GERMANA databases.

text recognition community, and the second one has been presented recently to the text recognition community. In the following subchapters both corpora are described in more detail.

3.1.1 IAM

The IAM-DB 3.0 database is a collection of handwritten English texts compiled by the *Computer Vision and Artificl Intelligence (FKI)* investigation group of the *Computer Science and Applied Mathematics (IAM)* institute of Bern. This database is publicly available and free for investigation purposes. It is a manual transcription of part of the *Lancaster Oslo/Bergen Corpus* (LOB) [19], transcribed by more than 600 people. No restrictions were stablished in the writing styles of the writers, therefore, even though the text was previously determined, it is a spontaneous text concerning writing style. Its texts are available segmented at word level, line level and phrase level. In this work the line level segmentation has been used.

This database contains 1.199 pages scanned at 300dpi resolution and saved as PNG images with 256 gray levels. This makes a total of 9.862 lines and $87.9K$ running words. For this job a training set with 6.161 lines was made, making aproximately $54K$ runing words. Validation and test sets contained 920 lines ($8.7K$ running words) and 2.781 lines ($25.4K$ running words) respectively. Feture extraction was performed using geometric-based method. HMMs were trained with lineal topology composed of 7 states with a mixture of 16 16 gausians per state. A WER of 35.5% was achieved on the test set.

3.1.2 GERMANA

GERMANA is a new handwritten text database, presented to facilitate comparison of different approaches to text line extraction and handwritten recognition. The GERMANA database is the result of digitising and annotating a 764-page Spanish manuscript entitled *"Noticias y documentos relativos a Doña Germana de Foix, última Reina de Aragón"* written in 1891 by Vicent Salvador, the Cruïlles' marquis. This manuscript was carefully scanned by experts from the Valencian Library at 300dpi in true colors. It has approximately $21K$ text lines manually marked and transcribed by palleography experts. This database was collected, analyzed and presented by the *Pattern Recognition and Human Language Technology (PRHLT)* investigation

Lang.	Pages	Lines	Words (K)	Lexicon Size (K)	Lexicon Sing. (%)	Char set
Spanish	595	16599	176.8	19.9	55.6	111
Catalan	87	2417	26.9	4.6	63.2	86
Latin	29	951	8.3	3.4	69.2	87
French	8	266	3.0	1.1	71.1	82
German	8	228	1.5	0.6	52.7	71
Italian	2	68	0.8	0.3	67.3	59
None	35	0	0.0	0.0	0.0	0
All	764	20529	217.2	27.1	57.4	115

Table 3.2: Basic statistics of GERMANA (Sing=Singletons, words occurring only once).

group of the *Instituto Tecnológico de Informática (ITI)* of Valencia, a task in which I collaborated.

GERMANA is a single-author book on a limited-domain topic. Most pages contain nearly caligraphed text written on ruled sheets of well-separated lines. It goes without saying that text lines extraction and off-line handwriting recognition on GERMANA is, by contract, not particullarly easy. It has typical characteristics of historical documents that make things difficult: spots, writing from the verso appearing on the recto, unusual characters and words, etc. Also, the manuscript includes many notes and appended documents that are written in languages different from Spanish, mainly Catalan, French and Latin. It is also worth noting that 68% of language model words occur once (singletons), and abbreviations appear in many different ways. Furthermore, 33% of them are incomplete words since they are at the beginning or the end of lines.

GERMANA entails and appropiate trade-off between task complexity and amount of data. To owr knowledge, it is the first publicly available database for handwriting research, mostly written in Spanish and comparable in size to standard databases. Due to its sequential book structure, it is well-suited for realistic assessment of interactive handwriting recognition systems. More details on the GERMANA database can be found in [2].

The experiments have been performed using only the first 179 pages, which correspond to well structured pages written only in Spanish. This makes a total of 3.753 lines and $42.4K$ running words. For this job a training set with 2.131 lines was made, containing aproximately $23.7K$ runing words. Validation and test sets had both 811 lines, with $9.4K$ and $9.1K$ running words respectively. Feature extraction was performed using geometric-based method. HMMs have lineal topology composed of 6 states with amixture of 64 gaussians per state. A WER of 42% was achieved on the test set.

3.2 Evaluation Measures

Let us assume that, after the recognition output is obtained, the system produces C correctly recognised words and I incorrectly recognised words. Using confidence measures (see chapte 2.5.1), only words with confidence below the decision threshold are rejected by the system, and therefor suggested to the human supervisor for correction.

Each suggested (rejected) word can be either a *True Rejection* or a *False Rejection*. A True Rejection is an incorrectly recognized word which was correctly rejected by the system, and a False Rejection is a correctly recognized word which was incorrectly rejected by the system. For evaluation purposes, the number of Truly and Falsely rejected words, and therefor suggested to the human supervisor for correction, are measured:

- *True Rejections* (TR): number of incorrectly recognised words suggested for correction.
- *False Rejections* (FR): number of correctly recognised words suggested for correction.

Every hypothesis verification system tries to use an appropiate decision threshold τ that performs as many True Rejections as possible while keeping low the number of False Rejections. The problem is that reducing τ makes both True Rejections and False Rejections increase, whilst increasing τ makes both values decrease. Therefor an agreement is needed.

This problem can be seen in a different way from a the interactive paradigm point of view: When the human supervisor completes the revision of the suggested corrections, we are interested in evaluating the human effort along with the improvement achieved. For this purpose, this two measures are used:

- *Supervision* (Sup): The ratio of recognized words that have been rejected by the system, and therefor revised by the human supervisor.

$$Supervision = \frac{TR + FR}{I + C}$$

- *Accuracy* (Acc): Measures the accuracy achieved after the suggested word have been revised.

$$Accuracy = \frac{TR + C}{I + C}$$

To provide an adequate overall estimation of these two measures, we need to compute both values for all possible decision threshold τ. This can be easily achieved based on a *Receiver Operating Characteristic* (ROC) curve. ROC curves are typically used to evaluate the performance of confidence measures. A ROC curve represents the *True Rejection Rate* (TRR) against the *False Rejection Rate* (FRR) for all possible values of τ. TRR and FRR are computed as:

$$TRR = \frac{TR}{I} \qquad FRR = \frac{FR}{C}$$

Let (frr,trr) be a point of the ROC curve, we can compute the Supervision and Accuracy measures for this decision threshold, as:

$$Sup(frr, trr) = \frac{trr \cdot I + frr \cdot C}{I + C} \qquad Acc(trr) = \frac{trr \cdot I + C}{I + C}$$

Computing the Supervision and Accuracy as a function of the ROC curve allows to evaluate the impact of confidence measures over the trade-off accuracy-effort.

3.3 Experimental Results

The proposed approach has been tested using GIDOC toolkit along with the IAM and GERMANA corpora (described in Sec. 3.1).

For both corpus, a bigram language model and character-level HMMs have been obtained using the training set. Upper and lower case words were distinguished and punctuation marks were modelled as separate words. The validation set has been used to adjust the Grammar Scale Factor (GSF) and Word Insertion Penalty (WIP) recognition parameters. For confidence estimation, a parameter to scale the language model probabilities has been also optimized using the validation set. This scaling has an important impact on the performance of word posterior probabilities as confidence measures [14]. The optimized parameters have been used in the test phase.

The improvements on the transcription accuracy as a function of the ROC curve are shown in Figure 3.3. We have emphasised the Supervision needed to achieve 80%, 90% and 95% of transcription accuracy.

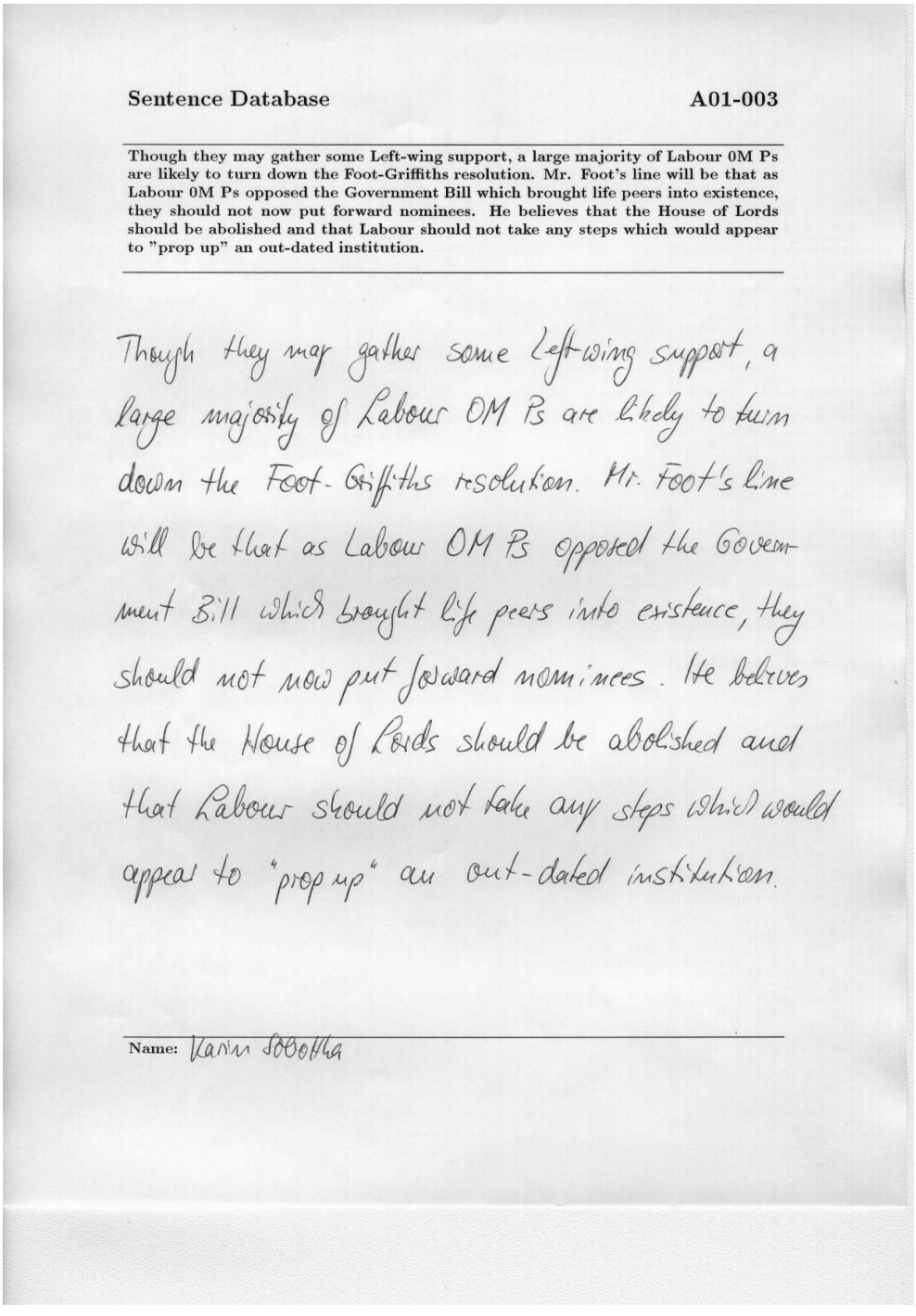
The transcription accuracy baseline (without supervision) for the IAM corpus is about 69%. Confidence estimation allows us to improve it up to an 80% by supervising only 15% of recognised words. This value increases to a nearly optimal 99% by supervising 69% of recognised words. In absolute terms, this implies a saving of $7k$ words to be supervised. Another view is that, when a small number of transcription errors can be tolerated for the sake of efficiency, the use of confidence measures can help to reduce drastically the supervision effort. For the IAM, a 97% of accuracy is achieved by supervising half of the recognized words.

Similar results have been obtained on the GERMANA corpus. The accuracy baseline (67%) is improved to an 80% by supervising only 16% of recognised words. Also, an accuracy of 96% is achieved by supervising half of the recognized words.

3.4 Conclusions

Confidence estimation has been presented to reduce human supervision effort in interactive transcription of handwritten text. Posterior probabilities computed from word graphs have been used as confidence measures. The approach proposed has been tested using the GIDOC toolkit along with the IAM and GERMANA databases. It

has been stated how the use of confidence measures can help to reduce drastically the supervision effort whilst improving the transcription accuracy. Experimental results show that the transcription accuracy can be higher than 95% while the number of words is reduced to the half. interactive paradigm. likely to be recognition errors.



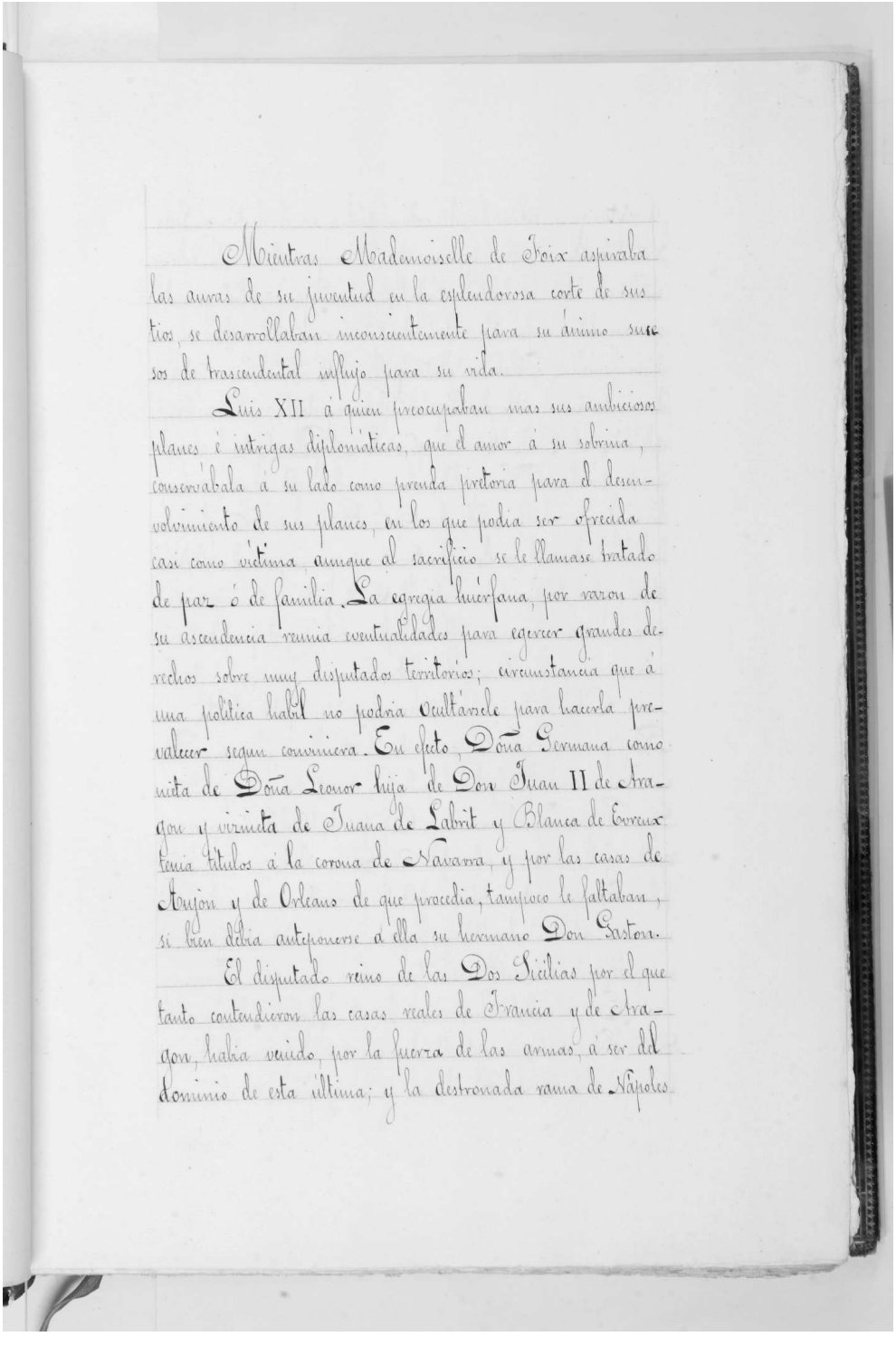
Sentence Database A01-003

Though they may gather some Left-wing support, a large majority of Labour 0M Ps are likely to turn down the Foot-Griffiths resolution. Mr. Foot's line will be that as Labour 0M Ps opposed the Government Bill which brought life peers into existence, they should not now put forward nominees. He believes that the House of Lords should be abolished and that Labour should not take any steps which would appear to "prop up" an out-dated institution.

Though they may gather some Left-wing support, a large majority of Labour OM Ps are likely to turn down the Foot-Griffiths resolution. Mr. Foot's line will be that as Labour OM Ps opposed the Government Bill which brought life peers into existence, they should not now put forward nominees. He believes that the House of Lords should be abolished and that Labour should not take any steps which would appear to "prop up" an out-dated institution.

Name: Karin Sobottka

Figure 3.1: Example of a IAM-DB 3.0 database page.



Mientras Mademoiselle de Foix aspiraba las auras de su juventud en la esplendorosa corte de sus tios, se desarrollaban inconscientemente para su ánimo sucesos de trascendental influjo para su vida.

Luis XII á quien preocupaban mas sus ambiciosos planes é intrigas diplomáticas, que el amor á su sobrina, conservábala á su lado como prenda pretoria para el desenvolvimiento de sus planes, en los que podia ser ofrecida casi como víctima, aunque al sacrificio se le llamase tratado de paz ó de familia. La egregia huérfana, por razon de su ascendencia reunia eventualidades para egercer grandes derechos sobre muy disputados territorios; circunstancia que á una política hábil no podria ocultársele para hacerla prevalecer segun conviniera. En efecto, Doña Germana como nieta de Doña Leonor hija de Don Juan II de Aragon y viznieta de Juana de Labrit y Blanca de Evreux tenia títulos á la corona de Navarra, y por las casas de Anjou y de Orleans de que procedia, tampoco le faltaban, si bien debia anteponerse á ella su hermano Don Gaston.

El disputado reino de las Dos Sicilias por el que tanto contendieron las casas reales de Francia y de Aragon, habia venido, por la fuerza de las armas, á ser del dominio de esta última; y la destronada rama de Nápoles

Figure 3.2: Example of a GERMANA database page.

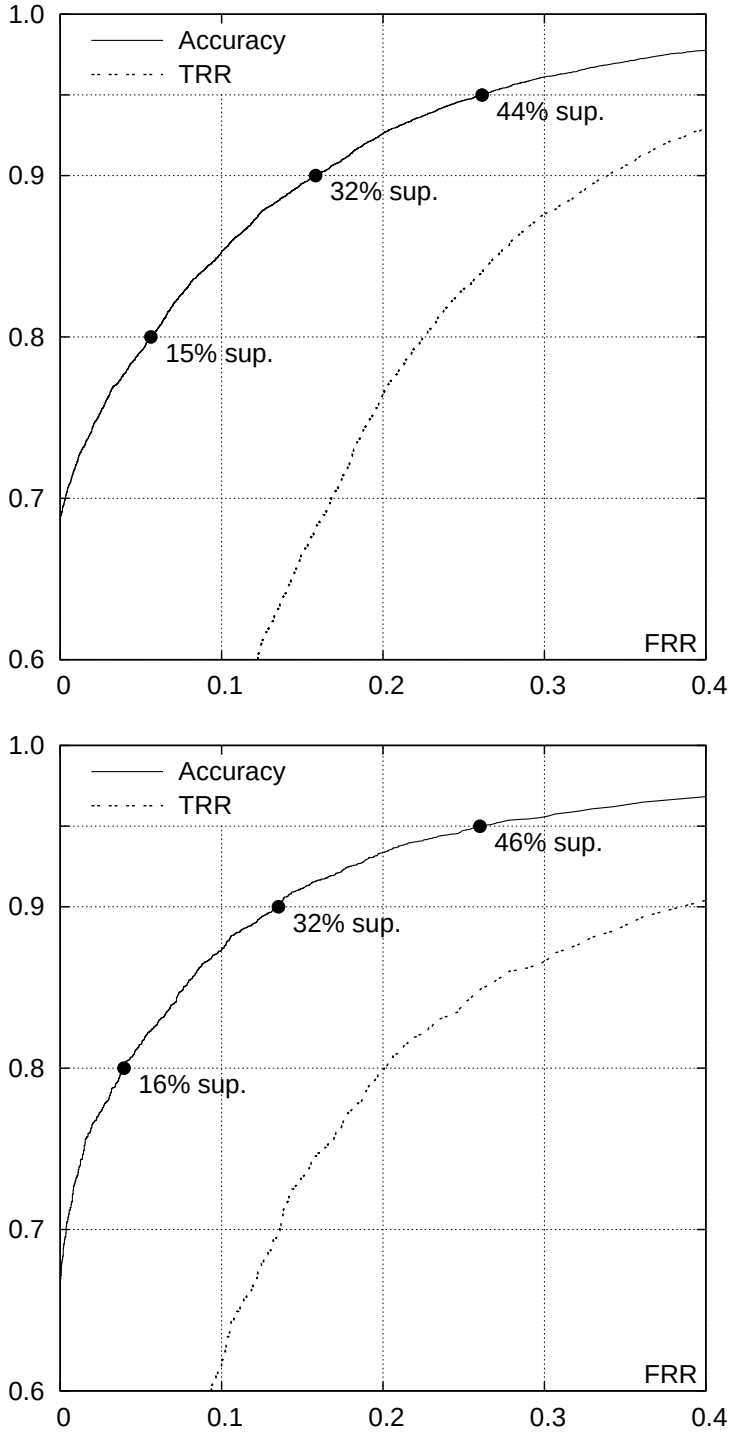


Figure 3.3: ROC curve and Accuracy on IAM (up) and Germana (down) databases.

CHAPTER 4

Conclusions

This work has contributed to the development of advanced techniques and interfaces for interactive transcription of handwritten text. More importantly, it has been demonstrated confidence measures can be applied in computer-assisted transcription to reduce the human effort. More specifically, the contributions described in this work are the following:

GIDOC: Gimp-based Interactive transcription of old text DOCuments.
GIDOC is a user-friendly interactive transcription prototype in which word-graph based confidence measures have been developed to support and guide the human transcriber in the annotation task. This work has led to one publication submitted in international conference:

- **WEBIST-2010:** N. Serrano, **L. Tarazón**, D. Pérez, O. Ramos-Terrades and A. Juan. The GiDOC Prototype. *Procceding of 6th International Conference on Web Information Systems and Technologies (WEBIST 2010).* Valencia (Spain). April 2010. (Submitted)

Application of Confidence Measures in Interactive Transcription.
Making use of the system prototype described above, confidence measures have been evaluated in an assisted transcription environment. It has been proved they can be applied to give support for error correction and detection, reducing drastically the supervision effort. This work has led to one publication in international conference:

- **ICIAP-2009: L. Tarazón**, D. Pérez, N. Serrano, V. Alabau, O. Ramos Terrades, A. Sanchis and A. Juan. Confidence Measures for Error Correction in Interactive Transcription of Handwritten Text. *Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP 2009)*. Vietri sul Mare (Italy). September 2009.

GERMANA: Preparation of old text document dabase

In this work, we have collaborated in the preparation of a database of old text documents: GERMANA, a 764-page Spanish manuscript from 1891. This database is described in one article in international conference:

- **ICDAR-2009:** D. Pérez, **L. Tarazón**, N. Serrano, F. Castro, O. Ramos and A. Juan. The GERMANA database. *Proccedings of the 10th International Conference on Document Analysis and Recogntion (ICDAR 2009)*. Barcelona (Spain). July 2009.

As said in the Introduction, it must be noted that the contributions described above are the result of a collaborative work involving other authors. The interested reader is referred to Table 1.1, for further information in the work attribution.

BIBLIOGRAPHY

[1] "iTransDoc: Interactive Transcription and Translation of Old Text Documents." `prhlt.iti.es/itransdoc.php`., 2010.

[2] D. P. i Cardona, "Preparació de corpus i desenvolupament de prototips en reconeixement de text manuscrit," Master's thesis, Dep. de Sistemes Informàtics i Computació, València, Spain, Dec 2009. Advisor(s): A. Juan and M. Pastor.

[3] A. H. Toselli, A. Juan, D. Keysers, *et al.*, "Integrated handwriting recognition and interpretation using finite-state models," *IJPRAI*, vol. 18, no. 4, pp. 519–539, 2004.

[4] R. Bertolami and H. Bunke, "Hidden Markov model-based ensemble methods for offline handwritten text line recognition," *Pattern Recognition*, vol. 41, pp. 3452–3460, 2008.

[5] F. L. Bourgeois and H. Emptoz, "DEBORA: Digital AccEss to BOoks of the RenAissance," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 193–221, 2007.

[6] J. Y. Ramel *et al.*, "User-driven page layout analysis of historical printed books," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 243–261, 2007.

[7] A. Juan *et al.*, "iDoc: Interactive Analysis, Transcription and Translation of Old Text Documents." Please visit `prhlt.iti.es`.

[8] N. S. M. Santos, *Multiple Contributions to Interactive Transcription and Translation of Old Text Documents*. PhD thesis, Dep. de Sistemes Informàtics i Computació, València, Spain, December 2009. Advisor: Dr. Alfons Juan Císcar.

[9] S. Neumann, M. Natterer, *et al.*, "GIMP: GNU Image Manipulation Program." Please visit `www.gimp.org`.

[10] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 123–138, 2007.

[11] A. H. Toselli, V. Romero, L. Rodríguez, and E. Vidal, "Computer Assisted Transcription of Handwritten Text," in *Proccedings of the 10th International Conference on Document Analysis and Recognition*, pp. 944–948, 2007.

[12] S. Young *et al.*, *The HTK Book*. Cambridge University Engineering Department, 1995.

[13] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. of ICSLP*, pp. 901–904, 2002.

[14] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Conf. measures for large vocabulary speech recognition.," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 3, pp. 288–298, 2001.

[15] R. Bertolami, M. Zimmermann, and H. Bunke, "Rejection strategies for offline handwritten text recognition," *Pattern Recognition Letter*, vol. 27, pp. 2005–2012, 2006.

[16] A. Sanchis, *Estimación y aplicación de medidas de confianza en reconocimiento automático del habla*. PhD thesis, Univ. Politécnica de Valencia, Spain, 2004.

[17] U. V. Marti and H. Bunke, "The IAM-database: an English sentence database for off-line handwriting recognition," *International Journal of Document Analysis and Recognition (IJDAR)*, pp. 39–46, 2002.

[18] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos, and A. Juan, "The GERMANA database," in *Proccedings of the 10th International Conference on Document Analysis and Recognition*, (Barcelona (Spain)), pp. 301–305, 2009.

[19] S. Johansson, G. Leech, and H. Goodluck, "Manual of information to accompany the lancaster-oslo/bergen corpus of british english, for use with digital computers.," tech. rep., Department of English, University of Oslo, 1978.

LIST OF FIGURES

LIST OF TABLES