

UNIVERSITAT POLITÈCNICA DE VALÈNCIA  
MASTER IN ARTIFICIAL INTELLIGENCE,  
PATTERN RECOGNITION AND DIGITAL IMAGE ANALYSIS



Contributions to Adaptation on Automatic Speech  
Recognition and Multilingual Handwritten Text  
Recognition

Work  
presented by Miguel Ángel del Agua Teba  
and supervised by  
D. Nicolás Serrano Martínez Santos and  
Dr. Alfons Juan Císcar

September 7, 2012



# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Preliminaries</b>	<b>5</b>
2.1	Introduction . . . . .	5
2.2	Theoretical Background . . . . .	5
2.2.1	Hidden Markov Models (HMM) . . . . .	6
2.2.2	Language Models based on $N$ -grams . . . . .	11
<b>3</b>	<b>Corpora and Baseline Experiments</b>	<b>13</b>
3.1	Introduction . . . . .	13
3.2	The GERMANA Database . . . . .	13
3.2.1	Baseline Experiments . . . . .	15
3.3	The poliMedia Database . . . . .	17
3.3.1	Baseline Experiments . . . . .	17
<b>4</b>	<b>Adaptation on Handwritten Text Recognition of Multilingual Documents</b>	<b>19</b>
4.1	Introduction . . . . .	19
4.2	GSF and WIP Adaptation . . . . .	20
4.2.1	Experiments . . . . .	20
4.3	Multilingual System . . . . .	23
4.3.1	Experiments . . . . .	23
4.4	Language Identification . . . . .	23
4.4.1	Probabilistic Framework . . . . .	24
4.4.2	Experiments . . . . .	26
4.5	Dealing with OOVs: Character-based approach . . . . .	28
4.5.1	Experiments . . . . .	29

<b>5</b>	<b>Adaptation on Speech Recognition</b>	<b>33</b>
5.1	Introduction . . . . .	33
5.2	Baseline system . . . . .	34
5.3	MLLR Adaptation . . . . .	35
5.3.1	Probabilistic Framework . . . . .	35
5.3.2	Experiments . . . . .	36
<b>6</b>	<b>Matterhorn</b>	<b>39</b>
6.1	Introduction . . . . .	39
6.2	Opencast Community and the Matterhorn project . . . . .	39
6.3	Matterhorn Infrastructure . . . . .	40
6.4	Contribution to Matterhorn . . . . .	41
<b>7</b>	<b>Conclusions</b>	<b>47</b>

# CHAPTER *1*

---

## INTRODUCTION

With the arrival of the digital era, there are large amounts of information being continuously generated and stored. Automatic search engines have made possible the instant access to this information. However, information have to be completely annotated in order to enable these search engines to inspect the content. The problem is that some of these resources are expensive and hard to annotate. An example of these resources are handwritten old text documents, residing in libraries all over the world. Annotation of these documents is a time-consuming task even for paleographic experts, and it can take up to 30 minutes per page. Another example are university lectures. Many universities are currently recording lectures and storing them for posterior reference. However, searches within all lectures have to be carried out by their title or topic, because annotations of the lecturer's speech are unavailable. This master's thesis deals with some improvement in the annotation of this two related tasks, handwritten text and speech.

Natural Language Processing (NLP) is a research area that aims to develop automatic systems that are able to process and comprehend human language by means of techniques and algorithms from Machine Learning (ML). One of the most hectic sub-areas inside NLP is Automatic Speech Recognition (ASR), that deals with the automatic annotation of speech. Annotation of speech is a difficult task, as speech is a continuous signal with a high variability depending on the speaker, language, topic, among some other features. Nowadays, much progress have been performed in this area, but even state-of-the-art systems are not able to generate acceptable annotations [1] to be used by search engines. A related area to ASR is Handwritten Text Recognition (HTR), which deals with the annotation of handwritten documents.

HTR is related to ASR, as the two of them model continuous signal and the models and techniques from one can be applied into the other. In case of HTR, handwritten script is a continuous signal because handwritten word are typically written from left to right. This similarity has caused that techniques and approaches of ASR can be successfully employed in HTR [2]. However, as it happens in ASR, even the automatic transcription of the best current approaches are still far from perfect [3].

Even though automatic systems cannot be used in a fully automatic approach, they can still be used as a tool in an interactive approach, in which the system and the user collaborate to complete the task. This approach has been used successfully in both, ASR [4] and HTR [5]. Interactive approaches have to deal with several problems. The first problem is to build an user friendly interface to interact with the system. Another important difficulty is how to employ user interaction further than simply post-editing the system output. This master's thesis deals with these two problem. Concretely, in an ASR, we deal with some parts within the interactive annotation problems of video lectures. On the other hand, in HTR, we improve the interactive transcription process of multilingual documents. More specifically, the contributions described in this work are the following:

#### **Language adaptation on the transcription of handwritten text documents**

A specially appealing case is the transcription of multilingual documents, such as GERMANA [6], in which up to six different languages appear. In this task, the coexistence of languages difficulties the task, as it greatly increases the language complexity. In this work, we deal with this problem by developing a language-dependent approach, in which a different system is trained for each language. Concretely, we present to different contributions. First, we describe the implementation of a language identification method, in order to detect the language of an untranscribe line and correctly switch its corresponding language dependent HTR system. Last, we study the adaption of tuning variables on the different language dependent recogniser. These contributions have led to two publications on two international conference ranked as C, according to the CORE:

- **M. A. del Agua**, N. Serrano and A. Juan. *Language Identification for Interactive Handwriting Transcription of Multilingual Documents*. In Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011), pp 596–603. Las Palmas de Gran Canaria (Spain). 2011.
- **M. A. del Agua**, N. Serrano, J. Civera and A. Juan. *Character-based Multilingual Handwriting Recognition*. In Proc. of IBERSPEECH 2012. Madrid (Spain). 2012

#### **Adaptation in automatic speech recognition of video lectures**

Actually, many universities are digitising their lectures, creating huge repositories, in which for each lecture, users can access video recordings along with its slides. This is the case of poliMedia, a video lecture database of the “Universitat Politècnica de València” (UPV). ASR of this database entangles several difficulties, for example, the great number of different speakers and topics. In

---

this work, we present the first step on ASR of this database along with a detailed analysis. Concretely, we present results using a standard ASR system and compare them with another system in which adaptation is performed for each segment using the MLLR algorithm.

### **Extension of Matterhorn, a framework for digitising video lectures**

Matterhorn is a software framework that deals with the whole process of acquiring a lecture, which goes from its digitisation to its on-line publication. This software have been chosen by the UPV in order to record and give access to the community to its lectures. In this work, we describe the current state of development that is being carried out to deal with the poliMedia database. Concretely, the most important step is the inclusion of a ASR system inside Matterhorn to automatically transcribe the lectures speech, along with an interactive tool that will enable users to correct the ASR errors.





# CHAPTER 2

---

## PRELIMINARIES

### 2.1 Introduction

In this section, we introduce the mathematical foundations of automatic recognition of continuous signals corresponding to a sequence of words. This is the case of ASR and HTR, which are the tasks studied in this work. Current ASR and HTR systems use a statistical approach based on PR techniques. PR studies how to assign a given input data its corresponding label or class. In our case, this process is performed as a search problem of the most probable transcription given an input signal, speech in ASR, or handwritten text in HTR. Under certain assumptions and in a perfect environment, the resulting transcription can be considered the best transcription that could be obtained. Although, in real life problems this perfect conditions cannot be achieved, the resulting systems are able to deal reliably with this task.

### 2.2 Theoretical Background

Current ASR and HTR systems use a statistical approach based on PR techniques. PR is a subarea of ML, which studies how to assign to a given input its corresponding label or class. In HTR, the input corresponds to a sequence of  $N$  feature vectors  $\mathbf{x} = x_1, \dots, x_N$  representing an image, while its label corresponds to  $M$  words forming the image transcription  $\mathbf{w} = w_1, \dots, w_M$ . In case of classification tasks in which error is measured using the classification error rate (CER), i.e. the ratio of errors committed when classifying, the Bayes decision rule [7] states that, the best sequence of words

$\mathbf{w}$  for the input  $\mathbf{x}$  corresponds to the one maximizing its posterior probability

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{w} \mid \mathbf{x}) \quad (2.1)$$

This posterior probability is decomposed according to the Bayes theorem

$$\hat{\mathbf{w}} = \frac{\operatorname{argmax}_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w})p(\mathbf{w})}{p(\vec{x})} \quad (2.2)$$

The term  $p(\mathbf{x})$  remain constant for all the possible transcriptions, and thus, it is can be dropped in the maximization

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{x} \mid \mathbf{w})p(\mathbf{w}) \quad (2.3)$$

where  $p(\mathbf{x} \mid \mathbf{w})$  is the conditional probability describing how likely (or probable) is to observe  $\mathbf{x}$  for the transcription  $\mathbf{w}$ , and  $p(\mathbf{w})$  is the prior probability that expresses how likely is to observe the transcription  $\mathbf{w}$ .

As stated above, Bayes decision theory achieve the optimal decision when the evaluation metric used is CER, and the probability distribution are known. However, there are two main problems. First, the evaluation metric used in HTR is Word Error Rate (WER), which is slightly different from CER. Last, probability distributions are unknown. In this work, we assume that there is no difference between the evaluations metrics, and that the probability distributions can be modeled statistically.

In this work, the conditional probability distribution  $p(\mathbf{x} \mid \mathbf{w})$  is based on Hidden Markov Models (HMMs) [8], and prior distribution  $p(\mathbf{w})$  is modeled using  $n$ -gram language models [9].

## 2.2.1 Hidden Markov Models (HMM)

The Hidden Markov Model is a finite set of states, each of which is associated with a continuous (generally multidimensional) probability distribution of "observations". Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. Only the outcomes, not the states are visible to an external observer and therefore states are "hidden" to the outside; hence the name Hidden Markov Model.

During the past decades it has become the most successful model used in ASR. The main reason for this success is its wonderful ability to model the speech signal in a mathematically tractable way. In ASR, HMM observations are discrete time sequences of acoustic parameter vectors. Given the similarity between ASR and HTR, the HMMs have seen increased their popularity in the HTR tasks. In HTR, the HMM observations are also discrete time sequences. However, in this case, the observations represent line-image features.

HMMs can be classified according to the nature of the observations. When the observations are vectors of symbols in a finite alphabet we are speaking of discrete HMMs. Another possibility is work with continuous observations (continuous HMMs).

Finally, the third class is called semi-continuous HMMs. These models user discrete observations, but they are modelled using continuous probability density functions.

Since in this master thesis we work with continuous HMMs, the formal definition and the formulation related with this kind of HMMs is summarized on the next subsections.

### Continuous HMM

Here, a formal definition of a continuous HMM is given, using similar notations presented in [10]. We assume that the observations can only be generated at states and not in the transition. Moreover, an additional initial state, which do not emit any observations, has been defined, in a similar way as in the case of the end state.

Formally, a continuous HMM  $M$  is a finite state machine defined by the sextuple  $(Q, I, F, X, a, b)$  where:

- $Q$  is a finite set of states. In order to avoid confusions with the indexation of the different states, we are going to call the states of the model as  $q_0, \dots, q_{|Q|-1}$ , whereas the sequence of states that generates the vector sequence  $\mathbf{x}$  will be denoted as  $\mathbf{z} = z_1, z_2, \dots, z_N$ .
- $I$  is the initial state, an element of  $Q : I \in Q. I = q_0$
- $F$  is the final state, an element of  $Q : F \in Q. F = q_{|Q|-1}$
- $X$  is the real  $d$ -dimensional space of observations:  $X \subseteq \mathbb{R}^d$
- $a$  is the state-transition probability function:

$$a(q_i, q_j) = p(z_{t+1} = q_j | z_t = q_i) \quad q_i \in (Q - \{F\}), \quad q_j \in (Q - \{I\})$$

Where  $z_t = q_i$  means that the HMM is on the state  $q_i$ , at the moment  $t$ . Transitions probabilities should satisfy  $a(q_i, q_j) \geq 0$  and

$$\sum_{q_j \in (Q - \{I\})} a(q_i, q_j) = 1 \quad \forall q_i \in (Q - \{F\})$$

- $b$  is a probability distribution function:

$$b(q_i, \vec{x}) = p(x_t = \vec{x} | z_t = q_i) \quad q_i \in (Q - \{I, F\}), \quad \vec{x} \in X$$

The following stochastic constraints must be satisfied:  $b(q_i, \vec{x}) \geq 0$  and

$$\int_{\vec{x} \in X} b(q_i, \vec{x}) d\vec{x} = 1 \quad \forall q_i \in (Q - \{I, F\})$$

As the observations are continuous then we will have to use a continuous probability density function. In this case probability density function is defined as a weighted sum of  $G$  Gaussian distributions:

$$b(q_j, \vec{x}) = \sum_{g=1}^G c_{jg} b_g(q_j, \vec{x})$$

where,

$$b_g(q_j, \vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_{jg}|}} e^{-\frac{1}{2}(\vec{x}_t - \mu'_{jg}) \Sigma_{jg}^{-1} (\vec{x}_t - \mu_{jg})}$$

- $\mu_{jg}$  is the mean vector for the component  $g$  of the state  $q_j$ .
- $\Sigma_{jg}$  is the covariance matrix for the component  $g$  of the state  $q_j$ .
- $c_{jg}$  is the weighting coefficient for the component  $g$  of the state  $q_j$ , and should satisfy the stochastic constraint  $c_{jg} \geq 0$  and

$$\sum_{g=1}^G c_{jg} = 1$$

Certain assumptions should be taken into account for the sake of mathematical and computational tractability, but it is not the aim of this document. For more detail please refer to [10].

### Basic algorithms for HMMs

Once we have an HMM, there are three problems of interest. The evaluation problem, the decoding problem and the learning problem.

- **The Evaluation Problem:** This problem consist on computing the probability  $p(\mathbf{x}|\mathcal{M})$ . Given an HMM  $\mathcal{M}$  and a sequence of observations  $\mathbf{x} = x_1, \dots, x_N$  with  $x_i \in \mathbb{R}^d$ , this is, the probability that the observations are generated by the model. This problem could be tackled with the Forward and Backward algorithms.
- **The Decoding Problem:** Given a model  $\mathcal{M}$  and a sequence of observations  $\mathbf{x}$ , the problem consist on find the most likely state sequence in the model that produced the observations. In other words, the problem consist on find the hidden part of the HMM. In order to achieve the solution, we shall use the Viterbi algorithm.
- **The Learning Problem:** Given a model  $\mathcal{M}$  and a sequence of observations  $\mathbf{x}$ , how should we adjust the model parameters  $\mathcal{M}$  in order to maximize the probability  $p(\mathbf{x}|\mathcal{M})$ . This problem could be addressed with the Baum-Welch algorithm.

### Forward and Backward Algorithms

Let  $\mathbf{x} = (x_1, \dots, \vec{x}_N)$  with  $x_i \in \mathbb{R}^d$  a sequence of real vectors and  $Z = \{\mathbf{z} = z_1, \dots, z_T : z_k = q_i \in (Q - \{I, F\}), 1 \leq i \leq |Q| - 2\}$  a set of state sequences associated with the vector sequence  $\mathbf{x}$ . Then, then probability that  $\mathbf{x}$  be generated by the model  $\mathcal{M}$  is:

$$p(\mathbf{x}|\mathcal{M}) = \sum_{z \in Z} \left( \prod_{i=1}^T a_{z_{i-1}} b_{z_i}(x_i) \right) a_{z_T} F$$

where  $z_0$  is the initial state  $I : z_0 = q_0 = I$ .

Direct calculation of this probability involves  $|Q|^N$  calculations, which is extremely large even when the length of  $\mathbf{x}$  is moderate.

The **Forward** algorithm is an efficient mean to compute  $p(\mathbf{x}|\mathcal{M})$ . The time complexity order of this algorithm is  $O(|Q|^2N)$ , but using a left-to-right HMM the complexity falls to  $O(|Q|N)$ .

The forward function  $\alpha_{j(t)}$  for  $0 < j < N$ , is defined as the probability of the partial observation sequence  $x_1, \dots, x_t$ , when it terminates at the state  $j$ . Mathematically,  $\alpha_{j(t)} = P(\mathbf{x}_1^t, q_j)$  and it can be expressed in a recursive way:

$$\alpha_{j(t)} = \begin{cases} a_{0j}b_j(x_1) & x = 1 \\ (\sum_{i=1}^{N-1} \alpha_i(t-1)a_{ij})b_j(x_t) & 1 < t \leq N \end{cases}$$

with the initial condition that  $\alpha_0(1) = 1$ . Using this recursion we can calculate the probability that the sequence  $\mathbf{x}$  be emitted by the model  $\mathcal{M}$  as:

$$P(\mathbf{x}|\mathcal{M}) = P(\mathbf{x}_1^N|\mathcal{M}) = \alpha_N(N) = \sum_{i=1}^{N-1} \alpha_i(N)a_{iN}$$

Similarly, the **Backward** function  $\beta_i(t)$  for  $0 < i < N$ , as the probability of the partial observation sequence  $x_{t+1}, \dots, x_N$ , given that the current state is  $i$ . Mathematically,  $\beta_i(t) = P(\mathbf{x}_{t+1}^N|q_i)$  and it can be expressed on a recursive manner:

$$\alpha_{j(t)} = \begin{cases} a_{iN} & t = N \\ (\sum_{j=1}^{N-1} a_{ij}b_j(x_{t+1})\beta_j(t+1)) & 1 \leq t < N \end{cases}$$

with the initial condition that  $\beta_N(N) = 1$ . Using this recursion the probability that the sequence  $\mathbf{x}$  be emitted by the model  $\mathcal{M}$  can be calculated as:

$$P(\mathbf{x}|\mathcal{M}) = P(\mathbf{x}_1^N|\mathcal{M}) = \beta_0(1) = \sum_{j=1}^{N-1} a_{0j}b_j(x_1)\beta_j(1)$$

## Viterbi Algorithm

In this case we want to find the most likely state sequence,  $\mathbf{z} = z_1, \dots, z_N$ , of the model  $\mathcal{M}$ , for a given sequence of observations,  $\mathbf{x} = x_1, \dots, x_N$ . The algorithm used here is commonly known as the Viterbi algorithm. This algorithm is similar to the forward algorithm, but replacing the sum by the dominating term.

$$\alpha_{j(t)} = \begin{cases} a_{0j}b_j(x_1) & x = 1 \\ (\max_{i \in [1, N-1]} v_i(t-1)a_{ij})b_j(x_t) & 1 < t \leq N \end{cases}$$

with the initial condition  $v_0(1) = 1$ . The probability of the sequence  $\mathbf{x}$  to be emitted by the model  $M$  is computed as:

$$v_N(T) = \max_{i \in [1, N-1]} v_i(TN) a_{iN} \leq \sum_{i=1}^{N-1} \alpha_i(N) a_{iN} = \alpha_N(N)$$

The time complexity of the Viterbi algorithm is:  $O(|Q|^2 N)$ , and using a left-to-right HMM the complexity falls to  $O(|Q|N)$ .

### Baum-Welch Algorithm

The learning problem is how to adjust the HMM parameters  $(a_{ij}, b_i(x), c_{jg}, \mu_{jg}, \Sigma_{jg})$ , so that a given set of observations (called training set) is generated by the model with maximum likelihood. The Baum-Welch algorithm (also known as Forward-Backward algorithm), is used to find these unknown parameters. It is an expectation-maximization (EM) algorithm.

Let  $E = \{\mathbf{x}_r = (x_{r1}, x_{r2}, \dots, x_{rT_r}) : x_{rk} \in X\}$  for  $1 \leq k \leq T_r \wedge 1 \leq r \leq R$  a set of  $R$  vector sequences, used to adjust the HMM parameters. The basic formula to estimate state-transition probability  $a_{ij}$  is:

$$\hat{a}_{ij} = \frac{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_i^r(t) a_{ij} b_j(x_{rt+1}) \beta_j^r(t+1)}{\sum_{r=1}^R \frac{1}{P_r} \sum_{t=1}^{T_r} t = 1 \alpha_i^r(t) \beta_i^r(t)}$$

where  $0 < i < N, 0 < j < N$  and  $P_r = p(\mathbf{x}_r | \mathcal{M})$  is the total probability of the sample  $r$  from the set  $E$ .

If the probability density function of each state on the HMM is approximated by a weighted sum of  $G$  Gaussian distributions we must find the unknown parameters  $c_{jg}, \mu_{jg}$  and  $\Sigma_{jg}$ . With this purpose we define  $L_{jg}^r(t)$  as the probability that the vector  $x_{rt} \in \mathbb{R}^d$  be generated by the Gaussian component  $g$  in the  $q_j$  state:

$$L_{jg}^r(t) = \frac{1}{P_r} U_j^r(t) c_{jg} b_{jg}(x_{rt}) \beta_j^r(t)$$

where

$$U_j^r(t) = \begin{cases} a_{0j} & \text{if } t = 1 \\ \sum_{i=1}^{N-1} \alpha_i^r(t-1) a_{ij} & \text{otherwise} \end{cases}$$

Taking into account the previous definitions, the parameters  $c_{jg}, \mu_{jg}, \Sigma_{jg}$  can be estimated as:

$$\begin{aligned} \hat{\mu}_{jg} &= \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t) x_{rt}}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t)} \\ \hat{\Sigma}_{jg} &= \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t) (x_{rt} - \hat{\mu}_{jg})(x_{rt} - \hat{\mu}_{jg})'}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t)} \\ c_{jg} &= \frac{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t)}{\sum_{r=1}^R \sum_{t=1}^{T_r} L_{jg}^r(t)} \end{aligned}$$

In terms of time complexity, one iteration of the Baum-Welch algorithm is:  $O(R|Q|^2N)$ . But using a left-to-right HMM the complexity falls to  $O(R|Q|N)$ . This algorithm is iterated until some convergence criterion is reached.

## 2.2.2 Language Models based on $N$ -grams

Language Models (LMs) are used to model text properties like syntax and semantic independently from morphological models. They are used in many natural language applications such as speech recognition, machine translation or handwritten recognition. These models try to capture the properties of a language, and are used to predict the next word in a word sequence. Language models assign probability to sequence of  $M$  words  $\mathbf{w} = w_1, \dots, w_M$ , which can be expressed using the chain rule as

$$p(\mathbf{w}) = p(w_1) \cdot \prod_{i=2}^M p(w_i | \mathbf{w}_1^{i-1}) \quad (2.4)$$

where  $p(w_i | \mathbf{w}_1^{i-1})$  is the probability of the word  $w_i$  when we have already seen the sequence of words  $w_1 \dots w_{i-1}$  (history).

In practice, estimating the probability of sequences is a very difficult task due to the high number of possible sentences that can appear and the lack of sufficient training data to estimate them all. In fact, for a vocabulary with  $|V|$  different words, the number of different histories is  $|V|^{i-1}$ . So, the estimation of  $p(\mathbf{w})$  can be unworkable. For that reason these models are often approximated using smoothed  $n$ -gram models which obtains surprisingly good performance although they only captures short term dependencies.

An  $n$ -gram defines a function:  $\phi_n : V^* \rightarrow V^{n-1}$  in which, all sequences finishing with the same  $n - 1$  words belong to the same equivalence class. Now,  $p(\mathbf{w})$  can be approximated as:

$$p(\mathbf{w}) \approx \prod_{i=1}^M p(w_i | \Phi_n(\mathbf{w}_1^{i-1})) = \prod_{i=1}^M p(w_i | \mathbf{w}_{i-n+1}^{i-1}) \quad (2.5)$$

Owing to the fact that, for the first  $n - 1$  words in  $\mathbf{w}$ ,  $i - n \leq 0$ , the Equation (2.5) must be written as:

$$p(\mathbf{w}) \approx p(w_1) \cdot \prod_{i=2}^{n-1} p(w_i | \mathbf{w}_1^{i-1}) \cdot \prod_{i=n}^M p(w_i | \mathbf{w}_{i-n+1}^{i-1}) \quad (2.6)$$

Given a vocabulary  $V$  and a transcribed training data or text corpora represented by  $\mathbf{w} = w_1, \dots, w_l$  the estimated probability of the word  $v \in V$ , having seen a sequence of  $n - 1$  words  $\mathbf{v} \in V^{n-1}$ , is computed as:

$$P(v | \mathbf{v}) = \frac{C(\mathbf{v}v)}{C(\mathbf{v})}$$

where  $C(\mathbf{v})$  is the number of times that the sequence  $\mathbf{v}$  has appeared on the training sequence  $\mathbf{w}$ . This is a maximum likelihood (ML) estimate.

Since not all possible  $n$ -grams have typically been seen in training, some smoothing method must be used to allow for unseen  $n$ -grams in the recognition phase. Two main smoothing techniques were used in this work: interpolation and "Back-off". However, it is not the aim of this master's thesis to develop these techniques for smoothing, accordingly the reader is referred to [11] for an extended overview.



# CHAPTER 3

---

## CORPORA AND BASELINE EXPERIMENTS

### 3.1 Introduction

In this chapter, the main features of the different corpora that have been used thoughtfully on this master's thesis are exposed, along with the results of a first baseline approach. The first one, the GERMANA database [6], is an off-line handwritten text manuscript obtained as a result of annotating and digitising a 764-page manuscript entitled "*Notícias y Documentos relativos a Doña Germana de Foix, última reina de Aragón*", written in Spanish up to page 180 and from there it coexists with Catalan, French, Italian, Latin and German until the end. It was written by Vicent Salvador, the Cruïlles' marquis in 1891. It has approximately 21K text lines manually marked and transcribed by paleographic experts and in terms of running words it is comparable to other databases.

By the other hand, the poliMedia repository [12] is a speech corpus obtained by transcribing 704 video lectures from the *Universitat Politècnica de València*, corresponding to 115 hours, so as to provide in-domain data set for training, adaptation and internal evaluations in Spanish, within the transLectures project.

### 3.2 The GERMANA Database

As said, the GERMANA, is an off-line handwritten text manuscript obtained as a result of annotating and digitising a 764-page of 1981. GERMANA is not a particularly difficult task for several reasons. First, it is a single-author manuscript on a limited-

domain topic. Also, the original manuscript was well-preserved and most pages only contain nearly calligraphic text written on ruled sheets of well-separated lines. Moreover, the manuscript comprises about 217K running words from a vocabulary of 30K which, apparently, is a reasonable amount of data for single-author handwriting and language modeling.

However, text line extraction and off-line handwriting recognition on GERMANA is not particularly easy. It has the typical properties of historical documents that make things difficult: spots, writing from the verso appearing on the recto, unusual characters and words, etc. Also, the manuscripts includes many notes and appended documents. In addition, GERMANA possesses a high language complexity due to the appearance of multiple languages.

Due to its sequential book structure, it is also well-suited for realistic assessment of interactive handwriting recognition systems. Moreover, it can be used as well to test approaches for language identification and adaptation from single author handwriting as it is used in this masters' thesis.

The manuscript was carefully scanned by experts from the Valencia Library at 300dpi in true colours. Then, the whole manuscript was transcribed line by line, by paleographic experts, in accordance with the following transcription rules:

- Page and line breaks were copied exactly.
- Blank space was only used to separate words.
- No spelling mistakes were corrected.
- No case or accentuation change was done.
- Punctuation signs were copied as they appeared.
- Words abbreviations were first copied verbatim, except for subindices and superindices, which were written in L<sup>A</sup>T<sub>E</sub>X-like notation as  $\_ {sub}$  and  $\hat {super}$ , respectively. Then, they were followed by the corresponding word between brackets.

Also, to facilitate language-dependent processing of the manuscript, each transcribed line was manually labelled in accordance with its dominant language. In table 3.1 on the next page contains some basic statistics drawn from GERMANA. These statistics were computed after applying the following preprocessing steps in order to reduce the language modeling complexity:

- Substitution of abbreviations by their corresponding words.
- Concatenation of hyphenated words at line ends with their remainders.
- Isolation of punctuation signs.

Note that Spanish part of GERMANA comprises about 17K text lines and 177K running words from a lexicon of 20K words. It is also worth noting that 56% of the

Language	Lines	Words	Lexicon	Singletons	Perplexity
All	20151	217K	27.1K	57.4%	289.8±17.0
Spanish	80.9%	81.4%	19.9K	55.6%	238.1±27.7
Catalan	11.8%	12.4%	4.6K	63.2%	112.9±61.6
Latin	4.6%	3.8%	3.4K	69.2%	211.1±51.3
French	1.3%	1.4%	1.1K	71.1%	88.3±21.0
German	1.1%	0.7%	0.6K	52.7%	92.1±29.2
Italian	0.3%	0.3%	0.3K	67.3%	63.3±14.4

**Table 3.1:** Basic statistics of GERMANA.

words only occur once (singletons). Regarding the other, non-Spanish parts, it is clear that it is difficult to reliably estimate independent models for them (c.f. HMMs and  $n$ -gram language models). In terms of running words, Spanish comprises about 81% of the document, followed by Catalan (12%) and Latin (4%), while the other three languages only account for less than a 3%. Similar percentages also apply for the number of lines. In terms of lexicons, it is worth noting that Spanish and, to a lesser extent, Catalan and Latin, have lexicons comparable in size to standard databases such as IAM [13].

Also note that the sum of individual lexicon sizes (29.9K) is larger than the size of the global lexicon (27.1K). This is due to presence of words common to different languages, such as common words in Spanish and Catalan. On the other hand, singletons, that is, words occurring only once, account for most words in each lexicon (55% – 71%). It goes without saying that, as usual, language modelling is a difficult task. To be more precise, in Table 3.1 we have included the global perplexity and the perplexity of each language, as given by a bigram model on a 10-fold cross-validation experiment. Perplexity is an information theory metric that is typically used to evaluate language models. Perplexity can be understood as the mean number of words that can follow a given word. the lower the perplexity is the lower the complexity of the language, as there is a lower uncertainty.

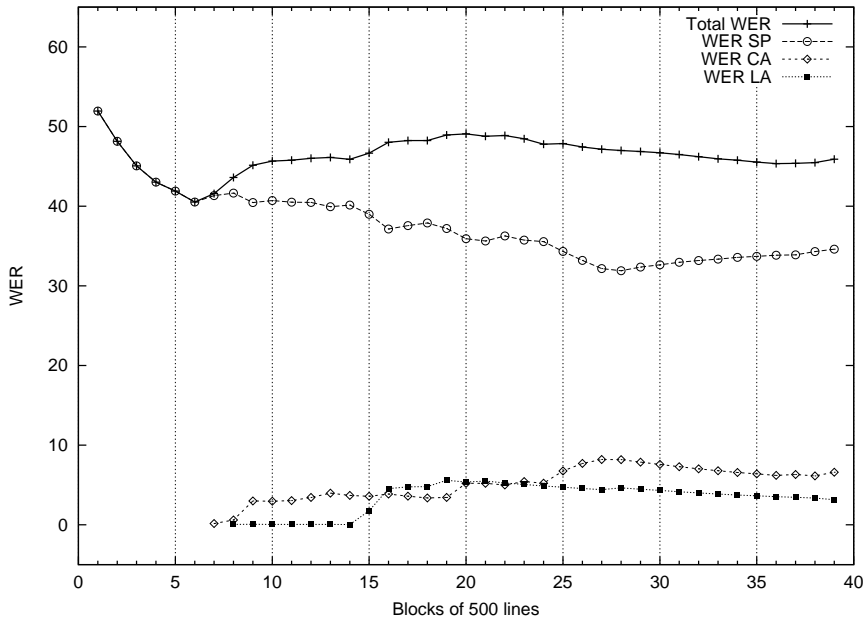
### 3.2.1 Baseline Experiments

As indicated below, GERMANA is a single-author manuscript written up to 6 different languages, but mainly in Spanish, Catalan and Latin. Our main goal is to study the results of transcribing of the whole GERMANA database using a first baseline system to be used as baseline in the next chapter. In the baseline system, which is referred as *monolingual* in the following, we assumed that all lines belong to the same language, and thus only require one language model. The image models, HMMs, are also trained from all available transcriptions.

In our experiments, we simulated the sequential transcription process of GERMANA. We divided GERMANA into 40 blocks of 500 lines each. The first two blocks were fully transcribed and an initial system was trained from the first and adapted in the second. This adaptation resulted in HMM models were 64 components per Gaus-

sian mixture with 4 states each, while the language model resulted in an interpolated 2-gram with modified Kneser-Ney discount [9]. These parameters remain unaltered for the rest of the experiments. Then, from block 2 to 40, each new block was, first recognized by the system, then evaluated in terms of WER and completely supervised, finally a new system is trained from all available transcriptions so far. The software used to train the HMMs was HTK [14] and SRILM [15] for language model training, while the recognition was also carried out by HTK.

The performance of the system, in terms of WER, can be observed in Fig. 3.1. For each block, we represented the WER of all recognised block so far. This is to represent an overall mean error at its stage of the transcription because the individual error of each block highly depends on its language and structure.



**Figure 3.1:** WER as a function of the blocks lines trained and obtained recognising the next. Furthermore, WER by language has been included for Spanish, Catalan and Latin.

As observed, the monolingual system achieves a final WER of 45.9. Even though, user interaction could be used to improve the transcriptions, this baseline error is too high to really improve from the manual transcription. A further analysis of the results revealed that, each time a language appears, the system gets worse. The main cause of this effect is the increment of out-of-vocabulary (OOVs) words, which are words that cannot be recognised by the system because they were not present in the training. In addition, each different language follows a different structure, which is not well estimated in a monolingual model. In summary, GERMANA presents two

main problems: Multilinguality and Out-Of-Vocabulary words. These two problems will be treated in the next chapter.

### 3.3 The poliMedia Database

Recently, an innovative service for creation and distribution of multimedia educational content has been developed at the *Universitat Politècnica de València* (UPV) under the name of poliMedia [12]. Its purpose is to allow UPV professors to record lectures on videos lasting of 10 minutes at most. Video lectures are accompanied with time-aligned slides and recorded at specialised studios under controlled conditions so as to ensure maximum video and audio quality and homogeneity. For the time being, poliMedia catalogue includes almost 8000 videos accounting for more than 1000 hours of lectures. Authors retain all intellectual property rights and not all videos are publicly available. More precisely, only about 2000 videos can be accessed freely.

“poliMedia” along with Videlectures.NET <sup>a</sup>, are the two repositories planned to be fully transcribed in the framework of the European project transLectures<sup>b</sup>. To this purpose, 704 video lectures in Spanish corresponding to 115 hours were manually transcribed using the tool Transcriber [16], so as to provide in-domain data sets for training, adaptation and internal evaluations in the transLectures project. These transcribed video lectures were selected according to the open access permissions granted by the authors, which guarantees that the corpus can be used by the research community beyond the scope of the transLectures project.

Most of the transcribed videos were annotated with its corresponding speaker, topic and keywords. More precisely, 94% of the videos were assigned a topic and 83% were described with keywords. However, these topics and keywords were not derived from a thesaurus, such as EuroVoc.

#### 3.3.1 Baseline Experiments

In this section, we described the first baseline experiments to assess the availability of ASR tasks. We divided the poliMedia corpora in three speaker-independent partitions: training, development and test. The statistics of this partition can be found in Table 3.2. Topics included in development and test sets range from art studies such as marketing or law, to technical studies such as chemistry or statistics. On the other hand, this topics are also included in the training set among many other ones, hence, this partitions is not topic independent.

To carry out the baseline experiments, the RWTH ASR [17] software was used for acoustic modeling and SRILM [15] for language model training. First, The baseline system, including acoustic, lexicon and language models was trained on the training set. Then, system parameters were adapted in terms of WER on the development set. Acoustic models were trained using triphones because it is well known that they outperforms monophonemes due to its context knowledge. Triphoneme models were

---

<sup>a</sup><http://videlectures.net>

<sup>b</sup><http://translectures.eu>

	Training	Development	Test
Videos	655	26	23
Speakers	88	6	5
Hours	117.6	3.8	3.5
Sentences	39K	1.4K	1.1K
Vocabulary	27K	4.5K	4K
Running Words	948K	34K	28K
OOVs	-	4.7%	5.3%
PPLs	-	212	221

**Table 3.2:** Basic statistics on the poliMedia partition.

inferred using conventional CART model using 2001 leaves. System adaptation on the development set resulted in a acoustic mode, in which each HMMs has 5 state with no loop-back, and each of the emits a Gaussian of  $2^9$  components. The best language model according to the system adaptation is an interpolated trigram model with Kneser-Ney discount.

The result obtained with these parameters, and adapting on the development set was 00 in terms of WER. It will be discussed how to improve this result, by means of speaker adaptation such as Maximum Likelihood Linear Regression (MLLR) in the following chapters.

# CHAPTER 4

---

## ADAPTATION ON HANDWRITTEN TEXT RECOGNITION OF MULTILINGUAL DOCUMENTS

### 4.1 Introduction

As shown in the previous chapter, HTR has gained much interest nowadays. The reason is that there are large volumes of old handwritten documents that need to be transcribed in order to preserve and quickly access the contents. The problem is that, even in state-of-the-art documents [3], automatic transcription are still far from perfect. In the previous chapter, we performed an HTR experiment on the GERMANA database, which corresponds to a single-author handwritten text documents of 1891. From the results, we observed that quality of automatic transcriptions was quite low. This was mainly caused by two important features of GERMANA: multilinguality and out-of-vocabulary (OOV) words.

In this chapter, we introduce some improvements in order to solve the commented problems. First, as new supervised words are generated after the recognition of each block, we studied the adaptation of some recognition parameters dealing with the language model. Next, we consider the multilinguality of the document by performing a language-dependent approach. In this approach, we also developed a method for automatically classifying the language of a line, as it is required to recognise it with its corresponding language dependent system. Finally, the OOVs problem is approached by means of building a character-based model, rather than the typical word-based models.

## 4.2 GSF and WIP Adaptation

The aim of this task is to fully transcribe the GERMANA database as it would in a real interactive scenario. In that case, GERMANA must be transcribed sequentially from the beginning to the end. As said, GERMANA is not uniform, and the document and language change from one part to another. Hence, in order to better adapt to these changes, recognition parameters can be optimized once new data is incorporated to train. More specifically, in our case, these parameters will be adapted on the last block added to the training partition, differently from the first experiments performed, in which the parameters were only optimized on the second block and remained unchanged.

We have considered two recognition parameters: *Grammar Scale Factor* (GSF) and *Word Insertion Penalty* (WIP). These parameters are introduced in the recognition to perform a trade-off between image and language model scores. More concretely, GSF is the amount by which the language model probability is scaled before being added to each token as it transits from the end of one word to the start of the next. And on the other hand, the WIP parameter is a fixed value added to each token when it transits from the end of one word to the start of the next. These parameters are introduced in Eq. 2.1 as follows

$$\begin{aligned}\hat{w} &= \underset{\mathbf{w}}{\operatorname{argmax}} p(\mathbf{x}|\mathbf{w})p(\mathbf{w}) \\ &\approx \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{x}|\mathbf{w}) + \log p(\mathbf{w}) \\ &\approx \underset{\mathbf{w}}{\operatorname{argmax}} \log p(\mathbf{x}|\mathbf{w}) + \alpha \cdot \log p(\mathbf{w}) + \beta\end{aligned}$$

where  $\alpha$  is the GSF and  $\beta$  is the WIP.

As said, the main idea of this adaptation is finding the parameter combination that minimizes the WER on the last supervised block, considering that two consecutive blocks may share common characteristics.

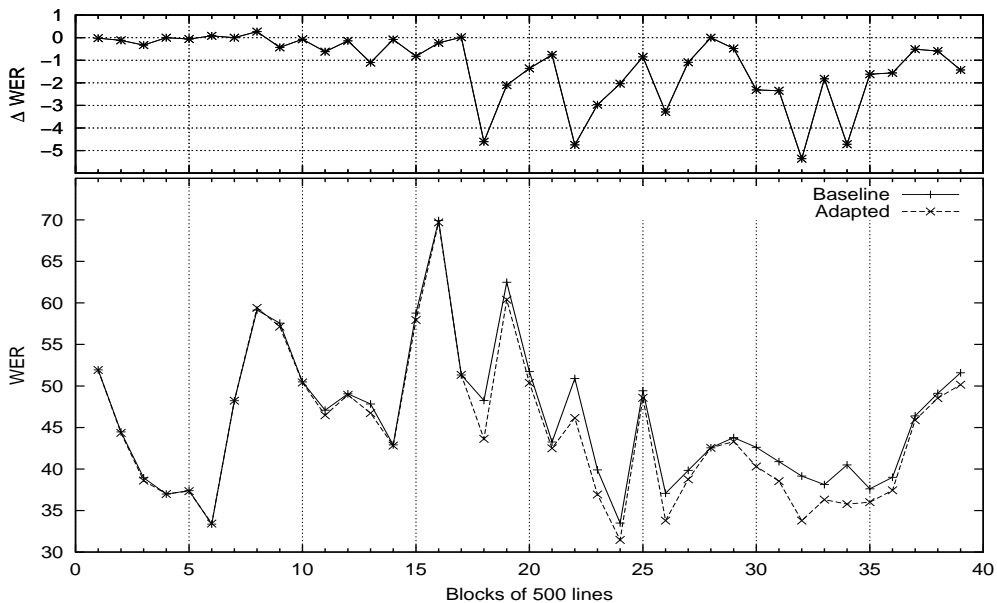
### 4.2.1 Experiments

In this experiment, we follow the same process introduced in Sec. 3. The GERMANA database was divided into 40 blocks of 500 lines each. The first two blocks were already transcribed from which an initial system was trained and adapted. Then, from block 2 to 40, each block is recognised, evaluated in terms of WER, supervised, added to the training set, and finally the system is re-trained from all supervised block so far. However, in this experiment, recognition parameters are adapted on the last supervised block. For the sake of clarity figure 4.1 is presented. As observed, adaptation on the last added block is based on the idea that two consecutive blocks might share more similar structure, writing or style than two separate blocks.

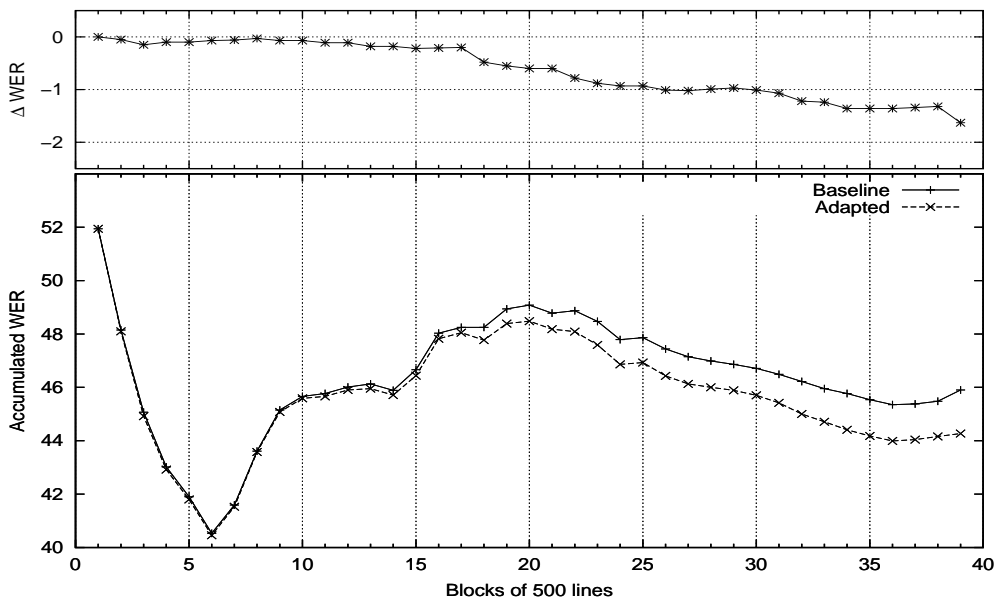
Results in terms of WER for each individual block are presented in Fig. 4.2. In this figure, the previous “Baseline” approach, in which no adaptation is used, is compared with the current system, which is “Adapted” on the last block. As observed, the adapted system works slightly better than the non-adapted until block 14 when the







**Figure 4.2:** Bottom: WER obtained as a function of the trained blocks for adapted and baseline systems. Top: WER increment between both systems.



**Figure 4.3:** Bottom: WER on all recognised block so far obtained as a function of the trained blocks for adapted and baseline systems. Top: WER increment between both systems.

## 4.3 Multilingual System

In this section, we deal with the problem of the multilinguality in GERMANA. Even though the book is written by a single author, we can take advantage of treating each language separately. As it is written by a single-author, an image model can be shared between languages. However, since each language holds its own vocabulary, they will differ on its lexicon and even the language model due to its different sentence structure. Language-dependent models are likely to better model the language than a global one. However, each language-dependent model will be correctly estimated if sufficient data is available. In addition, the training cost of multiple language-dependent recogniser compared to a single monolingual model has also to be considered.

Therefore, our main target is to sequentially transcribe all the GERMANA as it was performed in previous experiment, but, taking into account the language label of each sentence. In the current approach, before a line is recognised, its language has to be known in order to recognise it with its corresponding language-dependent recogniser. In the first set of experiment, we consider that the user specifies the correct language. In the second set of experiments, the language is detected using a language identification algorithm.

### 4.3.1 Experiments

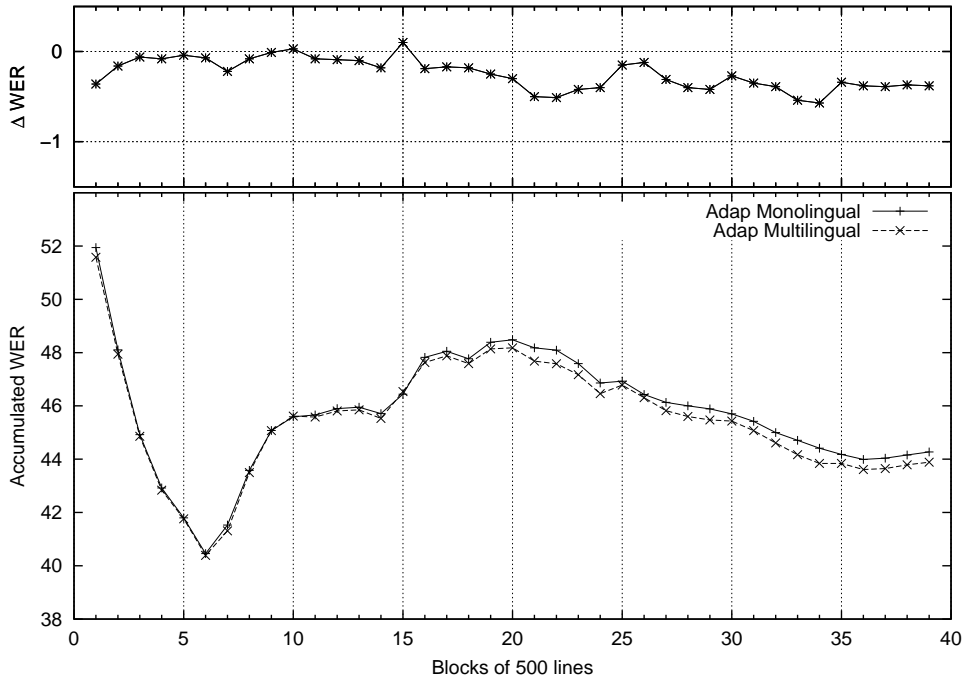
As mentioned, these experiments are aimed at elucidating the suitability of training independent models for each language. In this first experiment, the language label is known, thus we will obtain the best possible results in a multilingual approach. It is worth noting that this system also employs the adaptation method discussed in the previous section, adapting GSF and WIP on the last block shown, and also depending on the language.

In Fig. 4.4, we compare the recognition results from the monolingual and multilingual system, when language class of each line is known. Results are evaluated in term of WER on recognised blocks so far. As observed, the multilingual approach slightly improve the results. Even though there is not a significant enhancement, we think that in other multilingual books it could be greater. It must be taken into account the fact that GERMANA consist of 6 different languages arranged in a very inequitable proportion. This leads to a very poor recognition results in languages whose content is not enough to reliably estimate its models, for example German or Italian. For a detailed analysis of this results, the interested reader is referred to [18].

In the following section, it will be studied three different techniques to predict the language label of a given sentence and will be compared in terms of WER and IER (*Identification Error Rate*).

## 4.4 Language Identification

As stated above, in a multilingual context it becomes necessary to define a technique for language identification.



**Figure 4.4:** Bottom: Accumulated WER comparison between the adapted monolingual system and the adapted multilingual as a function of the blocks lines trained. Top: WER difference between both systems.

#### 4.4.1 Probabilistic Framework

The Probabilistic Framework will be presented supposing the case of word-based language models. Let  $t$  be the number of the current text line image to be transcribed, and let  $x_t$  be its corresponding sequence of feature vectors. The task of our system is to predict for each text line image first its language label,  $l_t$ , and then its transcription,  $w_t$ . We assume that all preceding lines have been already annotated in terms of language labels,  $l_1^{t-1}$ , and transcriptions,  $w_1^{t-1}$ .

By application of the Bayes decision rule, the minimum-error system prediction for  $l_t$  is:

$$\begin{aligned}
 l_t^*(x_t, l_1^{t-1}) &= \underset{\tilde{l}_t}{\operatorname{argmax}} p(\tilde{l}_t | x_t, l_1^{t-1}) \\
 &= \underset{\tilde{l}_t}{\operatorname{argmax}} p(\tilde{l}_t | l_1^{t-1}) p(x_t | \tilde{l}_t)
 \end{aligned} \tag{4.1}$$

where in Eq. (4.1), it is assumed that  $x_t$  is conditionally independent of all preceding language labels,  $l_1^{t-1}$ , given the current line language label,  $\tilde{l}_t$ . For the term  $p(x_t | \tilde{l}_t)$ , we marginalise over all possible word-based transcriptions for language  $l_t$ , that is,  $W(\tilde{l}_t)$ .

$$p(x_t | \tilde{l}_t) = \sum_{\tilde{w}_t \in W(\tilde{l}_t)} p(\tilde{w}_t | \tilde{l}_t) p(x_t | \tilde{l}_t, \tilde{w}_t) \quad (4.2)$$

$$\approx \max_{\tilde{w}_t \in W(\tilde{l}_t)} p(\tilde{w}_t | \tilde{l}_t) p(x_t | \tilde{l}_t, \tilde{w}_t). \quad (4.3)$$

Eq. (4.3), the Viterbi (maximum) approximation to the sum in Eq. (4.2), is applied to only consider the most likely transcription.

The decision rule (4.1) requires a *language identification model* for  $p(\tilde{l}_t | l_1^{t-1})$  and, for each possible language  $\tilde{l}_t$ , a  $\tilde{l}_t$ -dependent *word-based language model* for  $p(\tilde{w}_t | \tilde{l}_t)$  and a  $\tilde{l}_t$ -dependent *image model* for  $p(x_t | \tilde{l}_t, \tilde{w}_t)$ . As done in language modeling for monolingual documents, the language models in the multilingual case, both for identification and transcription, can be implemented in terms of *n-gram language models* [19]. Those for language-dependent transcription can be implemented as usual in the monolingual case though, in our case, each language  $\tilde{l}_t$  will have its own *n-gram language model*, trained only from available transcriptions labeled with  $\tilde{l}_t$ . Regarding the *n-gram language identification model*,  $p(\tilde{l}_t | l_1^{t-1})$ , as commented below we propose and compare three rather simple techniques:

1. A *bigram* model estimated by relative frequency counts:

$$\hat{p}(\tilde{l}_t | l_{t-1}) = \frac{N(l_{t-1}\tilde{l}_t)}{N(l_{t-1})} \quad (4.4)$$

2. A *unigram* model also estimated by relative frequency counts:

$$\hat{p}(\tilde{l}_t | l_{t-1}) = \frac{N(\tilde{l}_t)}{t-1} \quad (4.5)$$

3. And a “*copy the preceding label*” (*CPL*) bigram model:

$$\hat{p}(\tilde{l}_t | l_{t-1}) = \begin{cases} 1 & \tilde{l}_t = l_{t-1} \\ 0 & \tilde{l}_t \neq l_{t-1} \end{cases} \quad (4.6)$$

where  $N(\cdot)$  denotes the number of occurrences of a given event in the preceding lines, such as the bigram  $l_{t-1}\tilde{l}_t$  or the unigram  $\tilde{l}_t$ . Note that (4.4) and, especially (4.6), assume that consecutive text lines are usually written in the same language. This is not necessarily true though, in this kind of manuscripts (applications) we have in mind (e.g. GERMANA), it is a reasonable assumption.

Also as in the monolingual case, the *image models* for the different languages can be implemented in terms of *character HMMs* [19]. Moreover, if only a single script is used for all the languages considered (e.g. Latin), then a single, shared image model for all of them might produce better recognition results than a separate, independent model for each language. Clearly, this can be particularly true for infrequent languages.

Finally, it is often useful in practice to introduce scaling parameters in the decision rule so as to empirically adjust the contribution of the different models involved. In our case, the decision rule given in Eq. (4.1) can be rewritten as

$$l_t^*(x_t, l_1^{t-1}) \approx \operatorname{argmax}_{\tilde{l}_t} p(\tilde{l}_t | l_1^{t-1})^\beta \max_{\tilde{w}_t \in W(\tilde{l}_t)} p(\tilde{w}_t | \tilde{l}_t)^{\alpha_{\tilde{l}_t}} p(x_t | \tilde{l}_t, \tilde{w}_t) \quad (4.7)$$

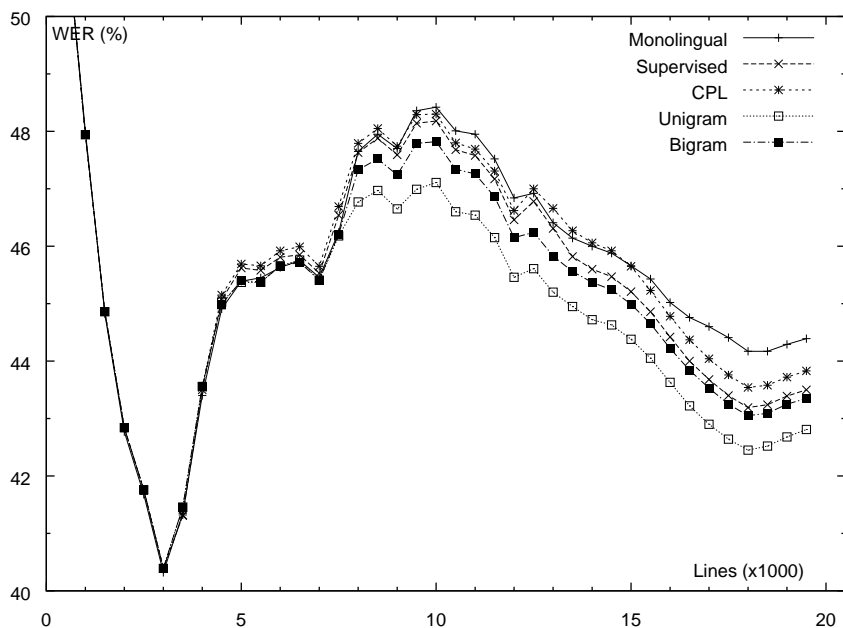
where we have introduced an *Identification Scale Factor (ISF)*  $\beta$  and, for each language  $\tilde{l}_t$ , a language-dependent *Grammar Scale Factor (GSF)*  $\alpha_{\tilde{l}_t}$ . Obviously, Eq. (4.7) does not differ from Eq. (4.1) when all these scaling parameters are simply set to unity.

## 4.4.2 Experiments

Up to now, only two systems have been compared: adapted monolingual and adapted multilingual. Henceforth, we are going to compare its performance with three different multilingual systems that only differ in the way they identify the language of the current line: *supervised* (manually given), *bigram* (using Eq. (4.4)), *unigram* (using Eq. (4.5)) and *CPL* (using Eq. (4.6)). Clearly, in all these multilingual systems, a different language (transcription) model was required for each of the 6 languages in GERMANA. However, as suggested at the end of the preceding section, a single, shared image model was used instead of a separate, independent image model for each language in GERMANA. The results are plotted in Fig. 4.5, in terms of WER of the recognized text up to the current line.

As expected, the multilingual systems achieves better results than the monolingual system. Also as expected, the correct language identification label (supervised) produces better results than an automatic, error-prone technique such as CPL. Surprisingly, however, the unigram and, to a lesser extent, the bigram identification techniques achieve better results than manual supervision. In other words, it is sometimes preferable not to use the correct, but probably poorly-trained language (transcription) model, and use instead a well-trained model for a different yet close language (e.g. Catalan and Spanish). On the other hand, it can be also observed that there are certain blocks at which the WER curve abruptly changes from a (smooth) decreasing tendency to a rapid increase. This was studied carefully in [18] by decomposing the (total) WER curve into its corresponding language-dependent WER curves. It was found that these abrupt changes are due to the occurrence of text from previously unseen languages, most notably Catalan (from line 3500) and Latin (from line 4000).

Although optimal (supervised) language identification does not necessarily lead to better recognition results than those obtained with suboptimal (imperfect) identification techniques, it is still important to have an identification technique of minimal error, maybe to just minimize user effort while correcting identification errors. Table 4.1 shows the Identification Error Rate (IER) of the proposed techniques for all and each language in GERMANA and both, in absolute and relative terms.



**Figure 4.5:** WER in GERMANA as a function of the number of recognized lines.

Language	Lines	IER (absolute)			IER (%)		
		2-gram	1-gram	CPL	2-gram	1-gram	CPL
All	19500	1290	2183	488	6.6	11.2	2.5
Spanish	15725	243	312	224	1.5	2.0	1.4
Catalan	2414	534	1136	181	22.1	47.1	7.5
Latin	951	255	409	49	26.8	43.0	5.2
French	266	116	182	31	43.6	68.4	11.7
German	76	74	76	2	97.4	100.0	2.6
Italian	68	68	68	1	100.0	100.0	1.5

**Table 4.1:** Identification Error Rate (IER) on GERMANA for the techniques proposed.

From the results in Table 4.1, it becomes clear that the simplest technique, CPL, is also the most accurate. It achieves an IER of 2.5%, that is, on average, only 3 identified labels out of 100 need to be corrected by the user. In contrast, the 1-gram and 2-gram techniques clearly fail in identifying languages other than Spanish. This might be due to the fact that scaling parameters were adapted to minimize the WER instead of the IER and, indeed, these techniques provided better results than CPL in terms of WER.

## 4.5 Dealing with OOVs: Character-based approach

Previous results exploiting multilinguality on the GERMANA database proved the benefits of explicitly modelling language identification at the line level in a interactive transcription scenario. However, these results are far from allowing an effective interactive transcription. In that work, the supervision effort would be excessively high, and the user might prefer to ignore the automatically generated output and transcribe the manuscript from scratch. An error analysis revealed that most of these errors were due to out-of-vocabulary (OOV) words. In fact, 53% to 71% of the words in the GERMANA database are singletons, words occurring only once in the lexicon of each language. Another important problem was the scarce resources available for some languages in the GERMANA database, so as to train their corresponding word-based language models.

The treatment of OOV words is an open problem in different areas of NLP. In speech recognition, which is closely related to handwritten text recognition as far as modelisation is concerned, notable efforts has been deployed over the last decades to deal with OOV words. In [20], the original lexicon is extended with words from external resources that are represented as a sequence of characters (graphemes, to be more precise) converted into phonemes. In [21], several sub-word based methods for spoken term detection task and phone recognition are presented to search OOV words. Phone and multigram-based systems provide similar performance on the phone recognition task, superseding the standard word-based system.

Regarding handwriting text recognition, the authors in [22] compared the performance of a conventional word-based language model to that of a character-based language model in the context of a German offline handwritten text recognition task. However, character-based language models were not superior to their word-based counterparts. A hybrid approach between a standard character-based n-gram language model and a character-based connectionist language model is proposed in [23], which obtain similar results to word-based systems on the IAM corpus [13].

To the best of our knowledge, character-based language models has not been able so far to supersede word-based language models in handwritten text recognition. Our hypothesis is that tasks tackled in previous work did not contain a significant number of OOV words compared to the figures of the GERMANA database<sup>a</sup>. In GERMANA, the problem of OOV words is aggravated by its multilingual nature, since the presence of languages such as Latin, French, German and Italian is less than 4% of the total number of words. Therefore, the estimation of word-based language models is notably poor, and it is necessary to fall back to adequate character-based language models.

Our main objective is to study the use of character-based models in GERMANA. As it has been said, the utilization of character-based models is motivated by two main features of GERMANA: the high number of OOVs, and the resource scarcity to train robust word language models. In addition, we analyze the performance of the language identification techniques presented in previous section.

---

<sup>a</sup>For example, the IAM corpus only contains about 7% of OOV words.



System	CPL	Unigram	Bigram
Character-based	2.5	14.2	4.0
Word-based		15.9	5.0

**Table 4.2:** Language identification results on GERMANA

### 4.5.1 Experiments

As in the rest of the experiments, we followed an interactive transcription framework, where the user supervises the output of a system, which is continuously retrained. To this purpose, we divided GERMANA in blocks of 500 lines, numbered from 1 to 40. First, blocks number 1 and 2 were fully transcribed and used to build an initial system and tune the training and recognition parameters. Training parameters, such as number of mixture components and states per HMM, remains unchanged in all experiments. It is worth noting that, in character-based models, the optimisation of the language model results in a 9-gram, instead of the 2-gram model of the word-based approach. Then, starting from block number 2 to the last. First, the language of each is identified (if needed) and its transcriptions is recognised by the corresponding language dependent system. Next, its transcription and language label is supervised. Finally, after a full new block is supervised, the system is re-trained from all supervised blocks and adapted on the last supervised block. It must be noted that, HMMs image modeling is carried out by the RWTH ASR toolkit [24] and language modeling by SRILM toolkit [15]. This software change is due to the fact that HTK cannot handle  $n$ -grams over order 2.

We performed two different sets of experiments on the described framework. The objective of the first set was to study the performance of the language identification methods proposed. On other hand, the objective of the second set was to study the transcription accuracy of the system when using each different language identification method.

In the first set of experiments, we compared the three different approaches for language identification presented in Sec. 4.4 but using a character-based system. We performed the interactive transcription of GERMANA using described framework for each of the approaches. Each time a block is recognised, we measured the number of errors committed by the language identification method used. It must be noted that, in this set of experiments, recognition parameters were tuned to minimise the number of language identification errors. Table 4.2 shows the results in terms of language identification error-rate (IER) for the whole document. We also included the results on the same framework of the word-based approach presented in previous section.

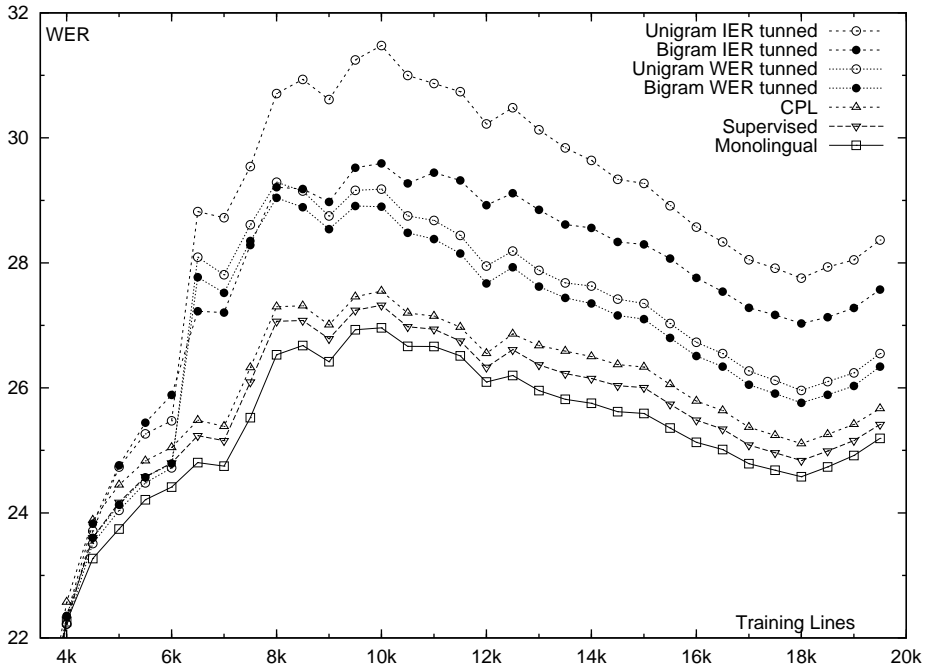
From the results in Table 4.2, it can be observed that CPL achieved the best performance. CPL took fully advantage of document sequentially and it only committed errors when the language changed from line to line, which only occurs a few times in GERMANA. In both, character and word based systems, the bigram approach tuned its parameters to ignore the language dependent recogniser probability and it forces

the system to only rely on the language model probability of language labels. In this case, the bigram approach identifies the language only using the bigram probability. However, the bigram approach only adapts its parameters each time a block is supervised, and thus, it fails to identify all lines of a language when it appears the first time in the transcription process. On the other hand, the character-based unigram approach achieved slightly better results than its word-based version.

In the second set of experiments, we compared five different approaches in terms of Word Error Rate (WER) on recognised transcriptions. WER is defined as the ratio between the minimum number of editing operations to convert the recognised words into the reference, and the number of reference words. In the first approach, we built a *monolingual* system, where we assume all lines to belong to the same language. This approach is considered the baseline, as language identification step is not needed and it is the simplest approximation to the problem. Next, motivated from the results of the previous section, we also built the same four different language dependent systems *supervised*, *CPL*, *bigram*, and *unigram*. It must be noted that, in this case, all approaches adapted their parameters to optimize the WER on last block. As the unigram and bigram approaches can be optimized for WER or IER, we also compared the results of both optimizations when transcribing, as the transcriptions produced are different. The results are represented in Fig. 4.6, in terms of WER of the recognized text up to the current line.

On the contrary, as it happened in section 4.4 on page 23, all multilingual systems achieved worse results than the monolingual system. However, even though there is not significant difference between the three best approaches, as corroborated by a bootstrap evaluation [25]; the monolingual approach is considered the best as it is easier to build and it does not need a language identification step in recognition. In error mean terms, even in the supervised approach, where the language is given, the use of language dependent recognizers could not outmatch the monolingual approach. The main cause of the monolingual performance is produced by the origin of all languages but German in GERMANA. Most languages in this document are *Romance* languages, which come from the same original language, sharing a common underlying language structure. For instance, the lexeme of many words can be correctly estimated from the Spanish part in order to recognise other similar romance languages, such as Catalan. In fact, the main responsible of the monolingual result is the high order (9-grams) character-based language model, which was able to estimate the common lexeme structure of all romance languages.

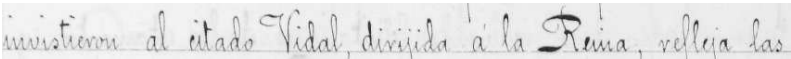
In language dependent approaches, it can be observed that, even though both supervised and CPL approaches achieved the best transcription results, the system performance did not always depend on the language identification performance. On one hand, there is not always a direct relationship between IER and WER. For instance, the unigram and bigram IER optimised approaches achieved a IER of 14.2 and 4.0, respectively, while the WER results were 28.36 and 27.57. On the other hand, as observed from the difference between the different optimizations of unigram and bigram approaches, a system with a worse IER can obtain a better WER results. For example, the bigram WER optimised approach obtained 26.34 of WER from a IER of 8.5, while optimising the IER on the same approach achieved 27.57 of WER



**Figure 4.6:** WER in GERMANA as a function of the number of recognized lines for the monolingual and language-dependent approaches. Results are presented from line 3500, in which a different language apart from Spanish appears.

from a IER of 4. These results corroborate our previous conclusions, in which we observed that a language is better recognised using a different language dependent recogniser. However, as said, the monolingual approach achieved better recognition results because the improvement from better estimated languages is already included in the character-based language model.

In terms of transcription performance, in our previous work [26], we also dealt with the complete transcription of GERMANA, but using word-based models. In that case, the monolingual approach obtained 44.39% of WER, however, in this work the same approach obtains 25.19%. This improvement is caused by two factors. On one hand, the RWTH recogniser improved the results due to a new feature extraction method. On the other hand, further error analysis revealed that, as expected, most of this improvement is due to the correct recognition of OOVs words, and punctuation signs. In Figure 4.7, we can observe the performance of both models in the recognition of a line, concretely, in this example, word-based errors (“estado”, “Viuda”, and “reflejasen”) occurred due to OOVs words (“citado”, “Vidal”, and “refleja”). On the other hand, punctuation signs (“,” after “Vidal” and “Reina”), are successfully recognized in the character-based approach, whereas, the word-based approach failed to recognize this signs due to its scarcity in the training dataset.

Image										
Character-based	invirtieron	al	citado	Vidal,	dirijida	á	la	Reina,	refleja	las
Word-based	invirtieron	al	estado	Viuda	dirijida	á	la	Reina	reflejasen	

**Figure 4.7:** Comparison of word-based and character-based recognition.

# CHAPTER 5

---

## ADAPTATION ON SPEECH RECOGNITION

### 5.1 Introduction

Nowadays the access to information is becoming an increasing challenge. Automatic search engines have made possible the instant access to large amounts of information obtained from very different contexts. Until now, we have talked about data becoming from old handwritten text documents and the need of annotating them to allowing its indexing and ease the access through digital libraries. But there are many other sources that need an annotation process in order to facilitate its search and dissemination, it is the case of videos. More specifically in the case in question, for the same reason a document needed to be transcribed, a video should be transcribed to allow its video indexing as well as its in-video content. Transcription of videos is an important time-consuming that is being carried out by universities which are currently recording lectures and storing them for posterior reference.

To reduce this effort, automatic speech recognition (ASR) techniques will have a major role. The object of ASR is to capture an acoustic signal representative of speech and determine the words that were spoken. In this chapter, our objective is to transcribe the previously presented poliMedia database obtaining the best possible results. In section 4, we tried to take advantage of the multilinguality feature of GERMANA, and we concluded that treating each language separately was the best option because of a better model adaptation. However, the results were only slightly better due to there was not enough data to train reliably models.

In ASR, instead of defining language-dependent systems, we could define speaker-dependent systems. Several studies have proved [27] that speaker-dependent (SD)

systems are typically performing, in terms of WER, from two to three times better than their equivalent speaker-independent (SI) counterparts. Since a large amount of speaker-specific data is needed for training SD, SI adaptation techniques must be applied. Thus, in this chapter a baseline system without adaptation will be defined and compared to an adapted system, trained with the well-known *Maximum Likelihood Linear Regression* (MLLR) transform.

## 5.2 Baseline system

The baseline system has been trained with the RWTH ASR [17] toolkit, along with the SRILM [15] toolkit. The RWTH ASR toolkit includes state-of-the-art speech recognition technology for acoustic model training. It also includes speaker adaptation, speaker adaptive training, unsupervised training, a finite state automata library, and an efficient tree search decoder. SRILM toolkit is a widespread language modeling toolkit which have been applied to many different natural language processing applications. Recognition is also carried out by the RWTH ASR toolkit.

Audio data was extracted from videos and preprocessed to extract the normalized acoustic features obtaining the Mel-frequency cepstral coefficients (MFCCs). Then, triphoneme acoustic models based on a prebuilt CART tree were trained using the training set, adjusting parameters such as number of states, the number of Gaussian components, number of CART leaves, etc. on the development set. The lexicon model was obtained in the usual manner by applying a phonetic transliteration to the training vocabulary. Thereafter, an  $n$ -gram language model was trained on the transcribed text after filtering out unwanted symbols such as punctuation marks, silence annotations and so on.

Finally, as it has been proved in [28], in order to enrich the language model, we have added an external resource in the language model estimation. More specifically, the final language model is the result of linearly combining an in-domain language model (training of poliMedia), with an external large out-domain language model computed on the Google N-Gram corpus [29]. To estimate the trade-off between such models, a  $\lambda$  parameter has been optimised so as to minimise the perplexity on the development set. It goes without saying that the lexicon has been extended to 50000 most frequent words present in Google N-Gram, in order to alleviate the OOV words appearance.

According to the partition of poliMedia established in 3 on page 13, the final results, in terms of WER, can be observed in the table 5.1. As expected, the extended language model works better than the first one due to its larger lexicon (less OOVs) and to its more precise probabilities estimation. It is worth emphasizing that the extended system will be referred as the baseline system, as well as the language model will be the same for the rest of the experiments.

System	WER
poliMedia	46.3
poliMedia + Google $N$ -grams	39.8

**Table 5.1:** Comparison between in-domain system versus in-domain extended with Google  $N$ -grams.

## 5.3 MLLR Adaptation

In order to improve the proposed baseline, we consider an MLLR adaptation. Adaptation techniques fall into two main categories: Speaker normalization in which the input speech is normalized to match the speaker that the system is trained to model, and model adaptation techniques in which the parameters of the model set are adjusted to improve the modelling of the new speaker. An important issue with both approaches is its effective operation with a limited amount of adaptation data. For a system with a large number of models and a small amount of adaptation data, some models will not be observed in the data. On the other hand, adaptation techniques only update the parameters of models which are observed in the adaptation data [30].

MLLR model adaptation uses a set of regression-based transforms to tune the HMM mean parameters to the new speaker. Each of the transformations is applied to a number of HMM mean parameters and estimated from the corresponding data. Using this sharing of transformations and data, the method can produce improvements with small amounts of adaptation data. If only a small amount of adaptation data is presented, a global transform is used for all models in the system; and if more data is available, the number of transforms is increased. This ensures that all model states can be adapted even if no model-specific data is available. For further information, please refer to [31].

### 5.3.1 Probabilistic Framework

The main idea is to apply a transformation matrix  $W$  to the Gaussian means on the state HMMs. For a specific Gaussian  $s$ , the transformation matrix  $W_s$  is applied in this way:

$$\hat{\mu}_s = W_s \cdot \mu_s + w_s \quad (5.1)$$

where

- $\mu_s$  is the mean of the Gaussian  $s$ .
- $w_s$  is the offset vector for  $s$ .
- $\hat{\mu}_s$  is the new mean for  $s$ .

For the sake of clarity, offset vector is introduced into mean vector:  $\tilde{\mu}_s = [w_s : \mu_s]$

$$\hat{\mu}_s = \tilde{W}_s \cdot \tilde{\mu}_s \quad (5.2)$$

MLLR estimates the regression matrices  $W_s$  that maximises the likelihood of on an adaptation set. The derivation of the MLLR estimate is not the aim of the present masters' thesis, but the reader is referred to [31] for further details.

When regression matrices are tied across mixtures components, each matrix is associated with many mixture components. This is achieved by defining a set of regression classes where each class contains all the mixture components associated with the same regression matrix.

In the tied approach, in order to be effective, it is desirable to consider an equivalence class for all the mixture components that use similar transforms. However, since we have no *prior* knowledge of the transforms, the mixture components will be compared using the likelihood as a measure.

### 5.3.2 Experiments

Our experiments objective is to study the improvement achieved by means of the MLLR transformation. The software used has been the RWTH ASR [17] toolkit, which is a state-of-the-art speech recognition that include utilities for speaker adaptation (such as MLLR).

We have carried out an unsupervised adaptation by firstly training a speaker-independent system with only the training set. Secondly, it was adapted on the development set in terms of WER, by trying different values of GSF and WIP parameters. Then a first recognition of the test set performed, whose result was our adaptation target.

In the next step, target classes within the test has to be considered, but the reference, and thus, the speakers are unknown. Instead of speaker-oriented adaptation, we have considered different classes by clustering the segments obtained in the first pass recognition. This segment clustering was performed by means of a bottom-up clustering, which used the Bayesian Information Criterion (BIC) as the stop criterion. As a result, target classes were obtained for the MLLR adaption. For the sake of clarity, we enumerate the steps of the described process:

1. Train with the whole Training set.
2. Adapt recognition parameters on Development.
3. First pass recognition of Test.
4. Segment clustering.
5. Estimate adaptation matrices depending on the specified number of regression classes.
6. Apply the transformation matrices in a second pass recognition.

Finally, the number of regression classes (sets of Gaussian which shares a common transformation matrix  $W_c$ ) were set automatically by specifying the minimum number of observations for each class. Thus, the experiments reported below compare the WER performance of a non-adapted (baseline) system and an adapted with MLLR:



System	WER
Baseline	39.8
MLLR	<b>33.9</b>

**Table 5.2:** Comparison between non-adapted and adapted systems

As it can be observed in table 5.2, thanks to the MLLR adaptation a reduction of 15% over the baseline is achieved. It must be noted that both systems were trained with the same lexicon and language model. In fact, the model used in the first pass recognition was the same as the baseline. These results confirm that MLLR adaptation is a good approximation to apply for speaker adaptation. In the future, it could be used as well as MLLR adaptation, adaptive training or even vocal tract length normalisation.



# CHAPTER 6

---

## MATTERHORN

### 6.1 Introduction

Matterhorn is a free, open-source platform to support the management of educational audio and video content. Institutions will use Matterhorn to produce lecture recordings, manage existing video, serve designated distribution channels, and provide user interfaces to engage students with educational videos.

The main idea of this chapter, is to integrate a speech recognition system as well as other tools developed within the framework of the transLectures project into Matterhorn, so as to enable real-life evaluation. In what follows, after a brief description of the Opencast Community and the Matterhorn project, we provide some technical details about Matterhorn infrastructure, its development, architecture and services.

### 6.2 Opencast Community and the Matterhorn project

Driven by the development of "pod-casting" technology, the increased quantity, quality and use of lecture recording have highlighted video management as a strategic imperative for universities in years to come. Founded in 2007, the Opencast Community is a global community addressing all facets of this domain, thus providing a framework for institutions to look for guidance, best practice and exchange of experience. It is open to all interested institutions and individuals including commercial providers. Its mailing list and communications infrastructure have encouraged their long-term cooperation and coordination and, indeed, over 300 organisations have al-

ready expressed interest in Opencast and more than 600 people have joined its mailing list. The Opencast Community also supports and guides a number of projects with the overall goal of facilitating and further developing the management of audiovisual content.

In 2008, the core of the Opencast Community consisted mainly of Universities that were already implementing their own video lectures broadcasting system. Nevertheless, the evaluation of these solutions and the discussions conducted within the framework of the Opencast Community had shown that none of the systems presented was able to fulfill the needs of at diverse international universities. Taking advantage of this circumstance, the Opencast community launched its first project: Matterhorn.

Matterhorn is a collaboration between North American and European institutions, funded in part by The Andrew W. Mellon and The William and Flora Hewlett foundations. The following 12 institutions constitute the "Matterhorn Partners" and also comprise the primary membership of the transLectures consortium through Knowledge for All Foundation (K4A): UC Berkeley, ETH Zurich, University of Nebraska-Lincoln, University of Osnabrück, Northwestern University, Indiana University, University of Vigo, University of Catalonia, University of Saskatchewan, University of Copenhagen, University of Toronto, and Jozef Stefan Institute (JSI). As a matter of principle, the Matterhorn project is open for collaboration with any interested persons and institutions. The project's governance model of "meritocracy" means that the role and influence of the participating institutions are predicated exclusively on their contributions. Key access points are the project's mailing list<sup>a</sup>, wiki and issue tracker<sup>b</sup>, code repository and public virtual meetings that are recorded and documented.

## 6.3 Matterhorn Infrastructure

Matterhorn provides a framework of services around the management of academic video that institutions can customise to meet their individual needs. Its architectural design and software principles allow for it to support transLectures tools. Fig. 6.1 shows a diagram of the Matterhorn architecture which includes its main components and dependencies among them.

Matterhorn is an open source; this means that the product is fully based on open source products. The members of the Opencast Community have selected Java as programming language to create the necessary applications and a Service-Oriented Architecture (SOA) infrastructure. The overall application design is highly modularised and relies on the OSGI (dynamic module system for Java) technology. The OSGI service platform provides a standardised, component-oriented computing environment for cooperating network services. Matterhorn is as flexible and open as possible and further extensions should not increase the overall complexity of building, maintaining and deploying the final product. To minimise the coupling of the components and third party products in the Matterhorn system, the OSGI technol-

---

<sup>a</sup>[matterhorn-users@opencastproject.org](mailto:matterhorn-users@opencastproject.org)

<sup>b</sup><http://opencast.jira.com>

ogy provides a service-oriented architecture that enables the system to dynamically discover services for collaboration.

In Fig 6.2 it is exposed the workflow of Matterhorn and for further details its points are described below:

1. **Prepare & Capture.** At the beginning of the recording process it must be determined what is to be recorded, where and what form. Matterhorn is open to both the learning management systems and administrative data bases so as to setting the Campus data and allowing the system to automatically schedule recordings.
2. **Process.** At the end of the recording the tracks are sent to an "inbox" to be processed. The inbox also serves as "ingest" for other video objects to be integrated in the subsequent work flows of Matterhorn. The different recording tracks (audio, content, video) are bundled to a media package, content-indexed (at first through optical character recognition of the slide, later certainly through audio recognition also) and if necessary archived in the most native formats. They are encoded according to the specified distribution parameters.
3. **Distribute.** The distribution module copes not only with the heterogeneous distribution formats (RSS, Atom, Web service interfaces), but also with the recording formats specified at the beginning which are transmitted in homogeneous form to external services and platforms. In addition, the distribution channel re-transmits the information necessary for statistical analysis and user data .
4. **Engage.** This module is closely linked to the distribute module since it must also manage presentation and use of the objects. To make sure that the produced material will be used, Matterhorn video and audio player components are easily integrated in existing course websites, wikis, and blog systems. In this module, barrier-free accessibility is more than a catch phrase; components are designed to support captions, screen readers and keyboard navigation. The possibility of integrating existing applications in Matterhorn is one of the its main properties. Taking advantage of this feature, it will be presented in this master's thesis an application demonstrating how would work the integration of an interactive speech recognition system.

## 6.4 Contribution to Matterhorn

The main target in transLectures is to develop tools and models for the Matterhorn platform that can obtain accurate transcriptions by intelligent interaction with users. For that reason, an HTML5 media player prototype has been built in order to provide a user interface to enable interactive edition and display of video transcriptions (Fig. 6.3). This prototype offers a main page where available poliMedia Videolectures are listed according to some criteria such as author or topic. Automatic video transcriptions are obtained from the ASR system when playing a particular video.

Since automatic transcriptions are far from not being in need of supervision, an interactive transcription editor facilitates user interaction to improve transcription quality. However, as users may have different roles while watching a video, the player offers two working environments depending on the user function: simple user or collaborative viewer.

Simple users will have a very restricted player which only allow them to assess the transcription quality. On the other hand, collaborative users may provide richer feedback to correct transcriptions. As shown in Figure 6.3, collaborative users have an *edit transcription* button available on the player control bar that enables the transcription editor panel. The editor panel is situated next to the video. It basically contains the transcription text, which is shown synchronously with the video playback. Clicking on a transcription word or sentence enables the interactive content modification. User corrections are sent to the speech recognition module through a web service, so corrections are processed and new transcription hypothesis are offered back to the user.

The current working HTML5 prototype<sup>c</sup> is a proof-of-concept version that works with pre-loaded transcriptions, however the version currently being developed communicates with the ASR system through a web service implemented for that purpose.

The next step is to integrate the developed interactive ASR system into the Matterhorn infrastructure. There are many different approaches to perform this integration. Our proposal lets an external system manage all the transcriptions, so there will not be necessary neither to add nor store them in any way into the current Matterhorn system<sup>d</sup>.

Moreover it is necessary to define a new Matterhorn workflow operation to transfer the audio data of the new media to the ASR system through a REST service, so as to obtain automatic transcriptions for every recording uploaded to the Matterhorn platform. This task will involve the implementation of a new Matterhorn service.

And finally, the Matterhorn Engage Player must be replaced or adapted to enable transcription edition. The player must obtain and transmit every transcription-related information through the REST Web Service in a similar way as the HTML5 prototype did. Here the main problem is the addition of new features to the Flash-based Matterhorn player, since it is not straightforward to implement the transcription functionalities provided by the HTML5-based player. The proposed solution is to use an alternative open-source Matterhorn engage player based on HTML5 called Paella Engage Player<sup>e</sup>.

---

<sup>c</sup><http://translectures.eu/player>

<sup>d</sup><http://opencast.jira.com/wiki/display/MH/MediaPackage+Overview>

<sup>e</sup><http://unconference.opencast.org/sessions/paella-html5-matterhorn-engage-player>

**MATTERHORN**  
 Architectural draft  
 Version 1.0

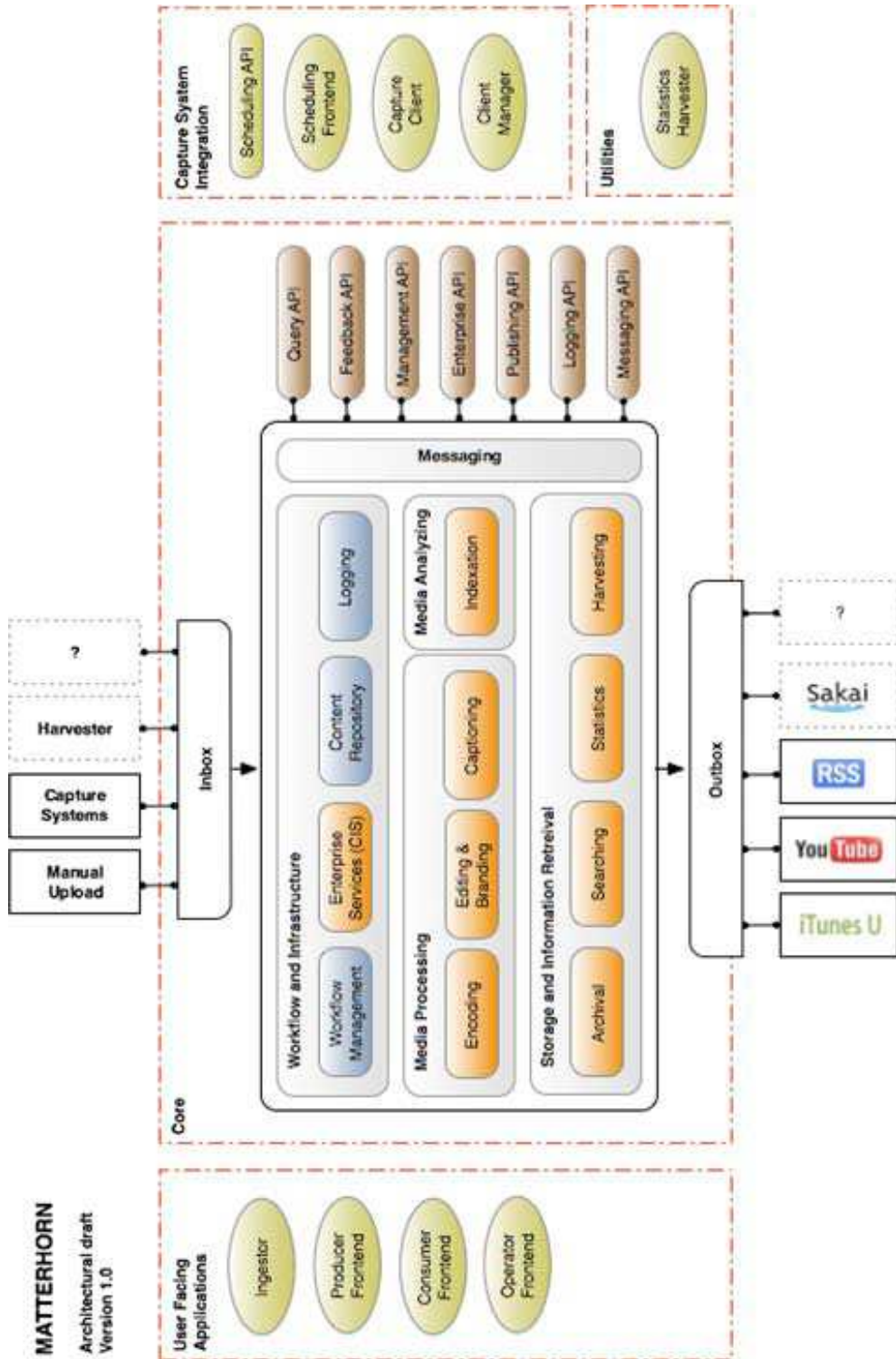
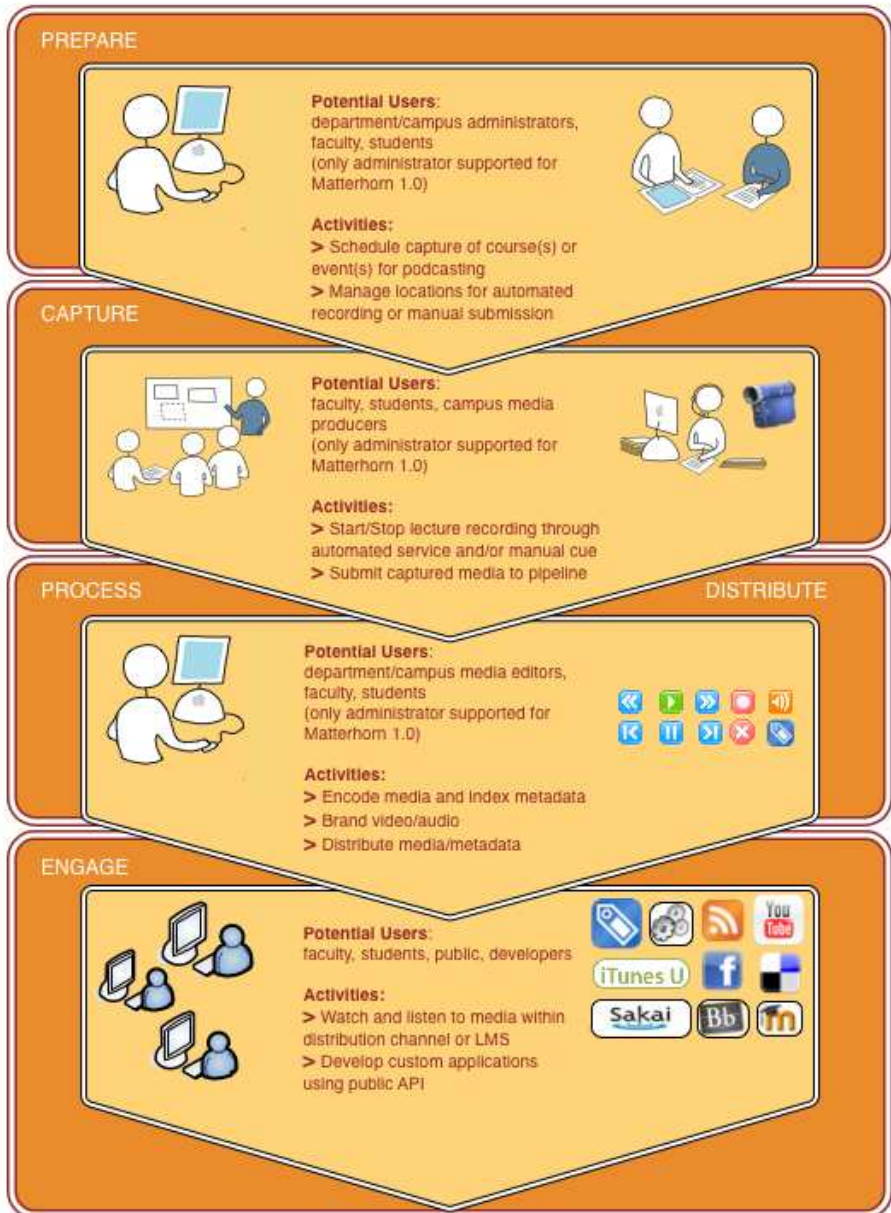


Figure 6.1: Matterhorn architecture



**Figure 6.2:** Phases of the Matterhorn Workflow

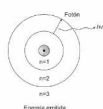


playing "estructura atómica"

**ESTRUCTURA ATÓMICA**

**ABSORCIÓN Y EMISIÓN DE ENERGÍA EN EL ÁTOMO**

- SI UN ELECTRÓN CAE A CAPA O NIVEL ENERGÉTICO MENOR, SE EMITE UNA CANTIDAD DISCRETA DE ENERGÍA (CUANTO) EN FORMA DE RADIACIÓN ELECTROMAGNÉTICA (FOTÓN).



1:00:31



Quando un electrón cae a una capa a un nivel energético de menor energía, se emite una cantidad discreta de energía, en forma de /radiación/radiación/ electromagnética.

[sonido de fondo]

Si el electrón sube, a una capa /nivel su objetivo/de nivel energético/ mayor, se absorberá un cuanto en forma de fotón.

[sonido de fondo]

El cuanto es la cantidad de energía, y el fotón es la forma de relación electromagnética.

[sonido de fondo]

Si Esto se puede expresar en una ecuación, que la variación de energía es igual a la  $h$  por  $\nu$ .

[sonido de fondo]

$h$  es la constante de Planck, cuyo valor es seis coma sesenta y tres por diez a la menos treinta y cuatro julios por segundo, y  $\nu$  es la frecuencia de este fotón.

Figure 6.3: Web Player and interactive transcription editor



# CHAPTER 7

---

## CONCLUSIONS

There are large amounts of information being continuously generated and stored. However, information have to be completely annotated in order to enable its content search by search engines. The problem is that some of these resources are hard and expensive to annotate. An example of such resources are old handwritten text documents and videos. Both are different, but the theoretical background of its automatic annotation is shared.

This work has contributed to improve the recognition performance of old text documents with a multilingual nature. More concretely, the contributions in this area has been the following:

### **Language adaptation on the transcription of handwritten text documents**

A specially appealing case is the transcription of multilingual documents, such as GERMANA [6], in which up to six different languages appear. In this task, the coexistence of languages difficulties the task, as it greatly increases the language complexity. In this work, we have dealt with this problem by developing a language-dependent approach, in which a different system is trained for each language. Concretely, we presented two different contributions. First, we described the implementation of a language identification method, in order to detect the language of an untranscribe line and correctly switch its corresponding language dependent HTR system. Last, we studied the adaption of tuning variables on the different language dependent recogniser. These contributions led to two publications on two international conference ranked as C, according to the CORE:

- **M. A. del Agua**, N. Serrano and A. Juan. *Language Identification for Interactive Handwriting Transcription of Multilingual Documents*. In Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011), pp 596–603. Las Palmas de Gran Canaria (Spain). 2011.
- **M. A. del Agua**, N. Serrano, J. Civera and A. Juan. *Character-based Multilingual Handwriting Recognition*. In Proc. of IBERSPEECH 2012. Madrid (Spain). 2012

As a future work, it can be improved the recognition of the multilingual system by combining linearly the language models from each language.

Regarding the improving of speech recognition, it has been applied the well-known MLLR adaptation technique to the corpus poliMedia:

### **Adaptation in automatic speech recognition of video lectures**

Actually, many universities are digitising their lectures, creating huge repositories, in which for each lecture, users can access video recordings along with its slides. This is the case of poliMedia, a video lecture database of the “Universitat Politècnica de València” (UPV). ASR of this database entangles several difficulties, for example, the great number of different speakers and topics. In this work, we present the first step on ASR of this database along with a detailed analysis. Concretely, we present results using a standard ASR system and compare them with another system in which adaptation is performed for each segment using the MLLR algorithm.

In the future, the application of adaptive training or vocal tract length normalisation could be applied to better adapt the acoustic models.

And finally, in an effort to apply a speech recogniser in a real scenario, it has been presented an HTML5 video player which allows to interactively transcribe videos:

### **Extension of Matterhorn, a framework for digitising video lectures**

Matterhorn is a software framework that deals with the whole process of acquiring a lecture, which goes from its digitisation to its on-line publication. This software have been chosen by the UPV in order to record and give access to the community to its lectures. In this work, we described the current state of development that is being carried out to deal with the poliMedia database. Concretely, the most important step had been the inclusion of a ASR system inside Matterhorn to automatically transcribe the lectures speech, along with an interactive tool that enable users to correct the ASR errors.

As a future work, the developed interactive ASR system will be integrated into the Matterhorn infrastructure, so as to enable the users to interactively correct automatic speech transcriptions. Moreover, it will be extended to allow interactive translation.

## BIBLIOGRAPHY

- [1] L. L. et al., “Speech recognition for machine translation in quæro,” in *International Workshop on Spoken Language Translation*, (San Francisco, California, USA), pp. 121–128, 2011.
- [2] H. Bunke, S. Bengio, and A. Vinciarelli, “Offline recognition of unconstrained handwritten texts using hmms and statistical language models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 709–720, 2004.
- [3] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, “A novel connectionist system for unconstrained handwriting recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.
- [4] S. Luz, M. Masoodian, and B. Rogers, “Interactive visualisation techniques for dynamic speech transcription, correction and training,” in *Proceedings of the 9th ACM SIGCHI New Zealand Chapter’s International Conference on Human-Computer Interaction: Design Centered HCI*, (New York, NY, USA), pp. 9–16, 2008.
- [5] N. Serrano, A. Giménez, A. Sanchis, and A. Juan, “Active learning strategies in handwritten text recognition,” in *Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, no. 86, (Beijing (China)), 2010.
- [6] D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos-Terrades, and A. Juan., “The GERMANA database,” in *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, (Barcelona (Spain)), pp. 301–305.
- [7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing ed., 2007.

- [8] L. R. Rabiner in *Readings in speech recognition*, ch. A tutorial on hidden Markov models and selected applications in speech recognition, pp. 267–296, 1990.
- [9] S. F. Chen, “An empirical study of smoothing techniques for language modeling,” tech. rep., 1998.
- [10] V. Romero, *Multimodal Interactive Transcription of Handwritten Text Images*. PhD thesis, Universidad Politécnic de Valencia, 2010. Advisors: Enrique Vidal and Alejandro H. Toselli.
- [11] A. H. Toselli, *Reconocimiento de Texto Manuscrito Continuo*. PhD thesis, Departamento de Sistemas Informáticos y Computación. Universidad Politécnic de Valencia, 2004.
- [12] UPV, “poliMedia.” <https://polimedia.upv.es/catalogo/>, 2008.
- [13] U. V. Marti and H. Bunke, “The IAM-database: an English sentence database for off-line handwriting recognition,” *IJDAR*, pp. 39–46, 2002.
- [14] P. Woodland, C. Leggetter, J. Odell, V. Valtchev, and S. Young, “The 1994 htk large vocabulary speech recognition system,” in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, pp. 73–76 vol.1, 1995.
- [15] A. Stolcke, “SRILM – an extensible language modeling toolkit,” in *Proc. of IC-SLP’02*, pp. 901–904, 2002.
- [16] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, “Transcriber: Development and use of a tool for assisting speech corpora production,” *Speech Communication*, vol. 33, no. 1–2, pp. 5 – 22, 2001.
- [17] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, “Rasr - the rwth aachen university open source speech recognition toolkit,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, (Hawaii, USA), 2011.
- [18] M. A. del Agua, “Multilingualidad en el reconocimiento de texto manuscrito.” Final Degree Project, 2010.
- [19] N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades, and A. Juan, “The GI-DOC prototype,” in *Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010)*, (Funchal (Portugal)), pp. 82–89.
- [20] M. Bisani and H. Ney, “Open vocabulary speech recognition with flat hybrid models,” in *Proc. of the European Conf. on Speech Communication and Technology*, p. 725–728, 2005.
- [21] I. Szoke, L. Burget, J. Cernocky, and M. Fapso, “Sub-word modeling of out of vocabulary words in spoken term detection,” in *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pp. 273 –276, 2008.

- [22] A. Brakensiek, J. Rottl, A. Kosmala, and G. Rigoll, “Off-Line handwriting recognition using various hybrid modeling techniques and character N-Grams,” in *In 7th International Workshop on Frontiers in Handwritten Recognition*, p. 343–352, 2000.
- [23] F. Zamora, M. J. Castro, S. España, and J. Gorbe, “Unconstrained offline handwriting recognition using connectionist character n-grams,” in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–7, 2010.
- [24] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Löff, R. Schlüter, and H. Ney, “The RWTH aachen university open source speech recognition system,” in *Interspeech*, (Brighton, U.K.), pp. 2111–2114, 2009.
- [25] B. Efron and R. J. Tibshirani, *An Introduction to Bootstrap*. Chapman & Hall/CRC, 1994.
- [26] M. A. del Agua, N. Serrano, and A. Juan, “Language identification for interactive handwriting transcription of multilingual documents,” in *Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011)*, (Las Palmas de Gran Canaria (Spain)), pp. 596–603, 2011.
- [27] C.-H. Lee, C.-H. Lin, and B.-H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden markov models,” *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 806–814, 1991.
- [28] J. D. V. et. al, “Integrating a state-of-the-art asr system into the opencast matterhorn platform.” IberSpeech2012, 2012. Submitted.
- [29] J. B. Michel *et al.*, “Quantitative analysis of culture using millions of digitized books,” *Science*, vol. 331, no. 6014, pp. 176–182.
- [30] J. luc Gauvain and C. hui Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.
- [31] C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.





## LIST OF FIGURES

3.1	Monolingual System: WER as a function of the block lines used in training, obtained on the next block and accumulated to the previous.	16
4.1	Experiment methodology . . . . .	21
4.2	Bottom: WER obtained as a function of the trained blocks for adapted and baseline systems. Top: WER increment between both systems. . .	22
4.3	Bottom: WER on all recognised block so far obtained as a function of the trained blocks for adapted and baseline systems. Top: WER increment between both systems. . . . .	22
4.4	Bottom: Accumulated WER comparison between the adapted monolingual system and the adapted multilingual as a function of the blocks lines trained. Top: WER difference between both systems. . . . .	24
4.5	WER in GERMANA as a function of the number of recognized lines. .	27
4.6	WER in GERMANA as a function of the number of recognized lines for the monolingual and language-dependent approaches. Results are presented from line 3500, in which a different language apart from Spanish appears. . . . .	31
4.7	Comparison of word-based and character-based recognition. . . . .	32
6.1	Matterhorn architecture . . . . .	43
6.2	Phases of the Matterhorn Workflow . . . . .	44
6.3	Web Player and interactive transcription editor . . . . .	45



## LIST OF TABLES

3.1	Basic statistics of GERMANA. . . . .	15
3.2	Basic statistics on the poliMedia partition. . . . .	18
4.1	Identification Error Rate (IER) on GERMANA for the techniques proposed. . . . .	27
4.2	Language identification results on GERMANA . . . . .	29
5.1	Comparison between in-domain system versus in-domain extended with Google <i>N</i> -grams. . . . .	35
5.2	Comparison between non-adapted and adapted systems . . . . .	37