# HMM based Classifier for the Recognition of Roots of a Large Canonical Arabic Vocabulary

Imen Ben Cheikh and Zeineb Zouaoui
*LaTICE Research lab,University of Tunis, ESSTT, 5.AV. Taha Hussein, BP.56, 1008, Tunisia*

Keywords: Natural Language Processing, Arabic Writing Recognition, Large Vocabulary, Hidden Markov Models, Canonical Vocabulary, Linguistic Properties, Viterbi Algorithm.

Abstract: The complexity of the recognition process is strongly related to language, the type of writing and the vocabulary size. Our work represents a contribution to a system of recognition of large canonical Arabic vocabulary of decomposable words derived from tri-consonantal roots. This system is based on a collaboration of three morphological classifiers specialized in the recognition of roots, schemes and conjugations. Our work deals with the first classifier. It is about proposing a root classifier based on 101 Hidden Markov Models, used to classify 101 tri-consonantal roots. The models have the same architecture endowed with Arabic linguistic knowledge. The proposed system deals, up to now, with a vocabulary of 5757 words. It has been learned then tested using a total of more than 17000 samples of printed words. Obtained results are satisfying and the top2 recognition rate reached 96%.

## 1 INTRODUCTION

Several approaches of writing recognition were the subject of intense research in recent decades. These approaches depend especially on writing type (printed or handwritten) and language (Arabic, Latin, Chinese, etc.). Moreover, direct application of approaches proposed for Latin seems insufficient for Arabic compared to other scripting languages due to the cursivness and the complexity of its script. Thus, several studies have focused on analyzing Arabic script topology and identifying levels of complexity of its words recognition process. The goal is to effectively implement robust recognition systems based on improvement and/or hybridation of existing approaches taking into account the specificities of the Arabic language and its cursive nature.

Moreover, because of Arabic morphological complexity, effective Arabic surface forms go past 60 billions (Cheriet and Beldjehem, 2006), what makes their automatic processing unrealistic. To deal with such problem, this number should be reduced. To this end, word morphological analysis and factorization seem to be one solution. Actually, in this context, we have already proposed, in a previous work (Ben et al., 2008); (Ben et al., 2010), a classifiers combination based approach that uses linguistic knowledge to simplify the enormous number of Arabic words. The simplification was based on word factorization in morphological entities (root, schemes and conjugation) since we deal with decomposable words. Our contribution is then, in this work, to focus on the recognition of a wide lexicon of Arabic decomposable words using an original model embedded with language skills and able to recognize roots from which words derive. This modeling is based on Hidden Markov Models (HMMs) that offer, indeed, great flexibility in modeling the Arabic script (Ben Amara, Belaïd and Ellouze, 2000). Furthermore, the use of HMMs in recognizing writing has yielded interesting results for some applications due to their ability to integrate context and absorb noise (Avila, 1996) and (Saon and Belaïd, 1997) and (El Yacoubi, 1996).

We perform the training and the recognition of roots from word global structural primitives, while proceeding 1) to simulate human reading process by operating globally first and 2) to apply natural language processing (NLP) by recognizing complementary linguistic entities. Global primitives are the result of automatic extraction of features made on 17375 printed samples corresponding to 5757 words derived from 101 roots, following 22 schemes and presenting various flexional conjugations.



Ben Cheikh I. and Zouaoui Z. (2013).
HMM based Classifier for the Recognition of Roots of a Large Canonical Arabic Vocabulary.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 244-252
DOI: 10.5220/0004335202440252

This paper is organized as follows. In section 2, we present the human reading model. The section 3 describes the main topological and morphological characteristics of Arabic script and presents the use of linguistic knowledge in literature. The section 4 is devoted to the description of the HMMs architecture and the design of the training process. Then, in section 5, experiments are conducted to evaluate the approach by revealing the recognition rates and the model limits. In section 6, we illustrate a comparison with related works. The section 7 concludes the paper and proposes some perspectives.

# 2 HUMAN READING MODELS AND NATURAL LANGUAGE PROCESSING

The basic experience of reading showed the "Word Superiority Effect" identified by Mc Clelland and Rumelhart. In order to illustrate this phenomenon, Mc Clelland and Rumelhart proposed a reading model (McClelland and Rumelhart, 1981) based on three fundamental hypotheses: 1) the perception is operated in three different processing levels, each one of them is representing a different abstraction level, 2) the perception implies parallel processing on the visual information and 3) the related processes are interactive (bottom-up and top-down). Mc Clelland and Rumelhart focus, in (McClelland and Rumelhart, 1985), on the fact that human builds a complete image of his environment by accumulating different sources of sensory data. In these various stages of decision-making, he proceeds by a general study of the problem. If this global vision is not sufficient, he seeks to go into details.

In addition to the "Word Superiority Effect", we attest that another aspect characterizes human reading models: "Word Derivation Effect". For example, given a vocabulary of decomposable words, derived from roots according to specific schemes and conjugations, the human handles with "access" and "root" letters (Cheriet and Beldjehem, 2006) ("root" letters are the three or four letters of the word root and "access" letters are those added by the scheme and the conjugation). For instance, according to Cheriet (Cheriet and Beldjehem, 2006), even if some word letters are not recognized, the human reading process can still guess about the word thanks to this distinction of letters types: if missed letters are "access", then the word recognition will be guided by the "root" letters and vice versa. Our point of view is also aligned with the human reading process. In fact, the first recognition result is not necessarily 0 or 1 but may be a quantity of significant information which needs to be completed. Thus, we find that word recognition can be performed using two independent and complementary views: recognitions of the scheme and the root. For instance, face to some words, human identifies, at first sight, the word scheme but not its root and vice versa. In these cases, the identified root or scheme helps to recognize the word.

# 3 THE INTEGRATION OF LINGUISTIC KNOWLEDGE

Many studies (Cheriet and Beldjehem, 2006), (Ben Hamadou, 1993), (Kanoun et al., 2005) and (Kammoun and Ennaji, 2004) highlight the richness and the stability of Arabic in terms of morpho-phonologic peculiar to this language. An Arabic word is decomposable or not. If it derives from a root, it is said to be decomposable in morphemes (root, prefix, infix and suffix). A word is, then, composed of root letters and access (non-root) letters. In (Cheriet and Beldjehem, 2006), Cheriet proposed to exploit this word vision using any recognition approach. He suggested analyzing errors using linguistic clues and in rejection cases, extracting the root and doing template matching to infer the rest. He affirmed that one must focus on what kind of linguistic knowledge is important and how and where it is more appropriate to incorporate. By adopting the same vision of the word, an "affixal approach", proposed by (El Yacoubi, 1996) then reconsidered by (Kammoun and Ennaji, 2004) for the recognition of Arabic typed texts, consists in the segmentation of words in letters and the recognition of their morphological entities. The authors used several linguistic concepts (Affixal and semantic restrictions) to guide the recognition process.

## 3.1 Morphological Concepts

As already mentioned, thanks to the flexibility of its derivation, with its three radical consonants, Arabic contains verb forms derived by adding affixes and intercalation. This complex and precise system contributes to the richness of Arabic verbs and also of abstract nouns that form action names (Bejaoui, 1985). We consider a decomposable word as the derivation of the root according to a conjugated scheme. This latter is the association of prefix and

suffix (letters from the conjugation: tense, gender, number...) to a brief scheme. The derivation of the tri-consonant root with the brief scheme gives rise to the radical (see Figure 1). The word changes meaning depending on the combination of the scheme: for example, derivation of the root **تجر** with the scheme **مفعل** (**م x x x**; where x denotes one of the three consonants of the root) gives the word **متجر** (shop) and with the scheme **مفاعل** (**م x ا x x**) gives the word **متاجر** (trader). Prefix and suffix are composed of letters corresponding to the flectional conjugation, while other access letters belong to the radical and depend on the scheme. Thus, according to our vision, to recognize a word, we just need to identify its root and conjugated scheme (brief scheme + conjugation), without segmenting it in letters. This is the main idea that allowed us to factorize words and to handle with a wide vocabulary while using a holistic approach to avoid effective segmentations (ICMWI 2010).

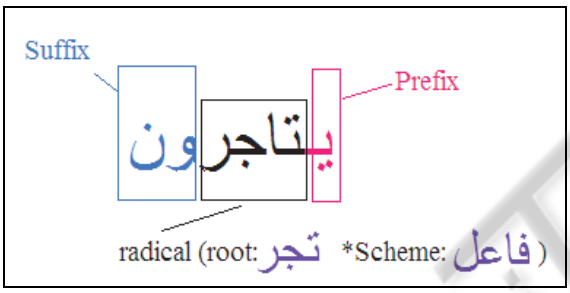


Figure 1: Our Arabic word vision.

### 3.2 Word Factorization based Recognition from Structural Primitives

▪ **Factorization.** We previously worked on a recognition system of Arabic decomposable words derived from tri-consonantal roots and belonging to a large canonical vocabulary. Note that our work does not consist in a morpho-syntaxic analyzer which deals with electronic text. It is rather about the recognition of word from sample images by 1) recognizing their morphological characteristics and then 2) reconstituting the whole world. The system is based on morphological collaboration of three classifiers specialized respectively in the recognition of roots, schemes and conjugation elements (see figure 2). Recognition takes place as follows: The system takes as input global features (see Table 1) which are structural primitives describing a word. At first, we provide word structural primitives to scheme classifier to recognize its scheme. Then we provide word structural primitives to root classifier to recognize its roots and finally, we provide word to conjugation classifier to recognize its conjugation elements. Then, the word will be reconstituted from the three morphological entities that are: the root, the scheme and the conjugation selected candidates. Remind that the phase of word root recognition is the scope of this work.

Notice that we are interesting only in the treatment of words deriving from tri-consonant roots. Dealing with tri-consonant roots, we could easily generate large vocabularies. In fact, 808 healthy tri-consonant roots can generate a lexicon of 98000 words (Kanoun, 2002). On average, 80 frequently used word can derive from a given root in various schemes (Ben Hamadou, 1993).

▪ **Structural Primitives Extraction.** Remind that our system learns and recognizes word roots from structural primitives that we extract from word images, by applying treatment illustrated in figure 3. More details are provided in (Ben Cheikh, Kacem and Belaïd, 2010). Note that a structural primitive (see table 2) is a combination of global features (see table 1).

## 4 PROPOSED APPROACH

Our contribution is to instantiate the root classifier using the Markov method for the training and the recognition of root words from their descriptions in structural primitives. Indeed, we propose a system based on 101 HMMs for the classification of 101 roots used in the generation of a vocabulary of 5757 words. It is important to attest that, this is the first work which proposes HMMs that 1) use global features instead of local ones and 2) have very personalized architectures with specific states.

### 4.1 HMMS Architecture

The proposed architecture of HMMs is personalized. It incarnates Arabic linguistic knowledge about the construction of Arabic words around the roots. This knowledge inspired us in the choice of states of consonants and affixes (prefixes, infixes, suffixes) and the choice of transitions between these states (see figure 4).

In addition, we adopted the same architecture for all HMMs, where each consists of six states for the prefix, the infix, the suffix and three other states modeling the root letters.

Moreover, our definition of affixes is not quite different from linguists one. Indeed, a prefix is any letter before the first root letter, the infix is any letter

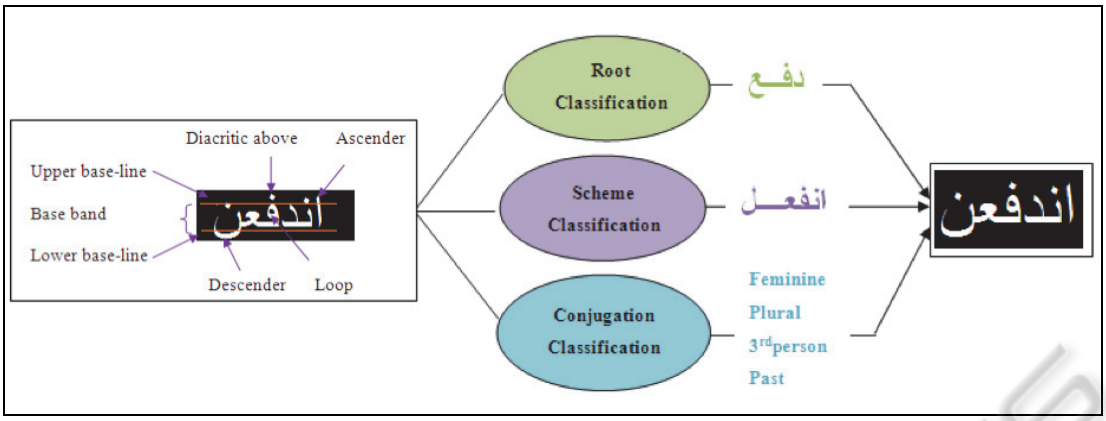


Figure 2: Classifiers collaboration based approach for the recognition of decomposable Arabic words.

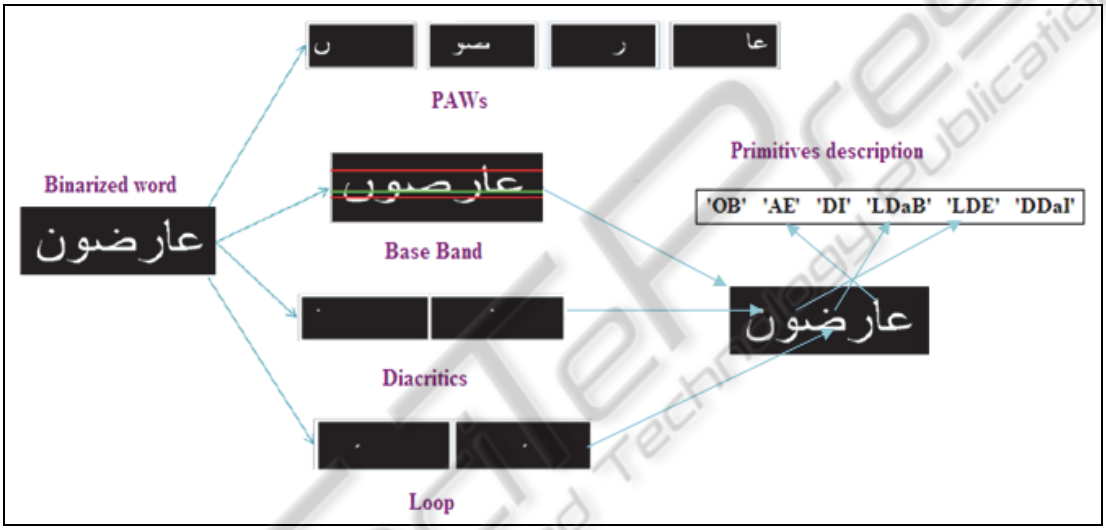


Figure 3: Extraction steps for word "عارضون".

Table 1: Global features (Ben et al., 2010).

| Feature | Designation | Description | Feature | Designation | Description |
|---|---|---|---|---|---|
| A | Ascender | Ascending characteristic over the baseline | O | Otherwise | No primitives listed above |
| D | Descender | Descending characteristic below the baseline | B | Beginning | Primitive position in the beginning |
| L | Loop | Intersection between 2 closed boundaries | M | Middle | Primitive position in the middle |
| Da | Diacritics above | Diacritical points over the word body | E | End | Primitive position at the end |
| Db | Diacritics below | Diacritical points below the word body | I | Isolated | Isolated primitive |

between the first and second root letter or between the second and third root letter and a suffix is any letter after the third root letter. Figure 4 describes the architecture for the HMM of the root بعد (i.e: he walked away). States emitted symbols are primitives extracted from word images. We also specified the beginning probabilities of each state.

Table 2: Examples of structural primitives.

| Primitive | Example | Primitive | Example |
|---|---|---|---|
| AB | ل | $D_bI$ | ب |
| AI | ا | OB | ح ع |
| AE | ﻟ | OI | د |
| DI | ع ح ر | ADI | ل |
| DE | ى | ALI | ط |
| LB | ص ه | $AD_aI$ | أ |
| LI | ٥ | $AD_bI$ | إ |
| LM | ﻫ ﻫ | $DD_aE$ | خ |
| LE | ٩ | $DD_aI$ | ش |
| $D_aB$ | خ ز ژ ؤ | DLE | و |
| $D_aI$ | ذ | DLI | ص |
| $D_bB$ | ﺑ ﻳ | $DD_bI$ | ج |

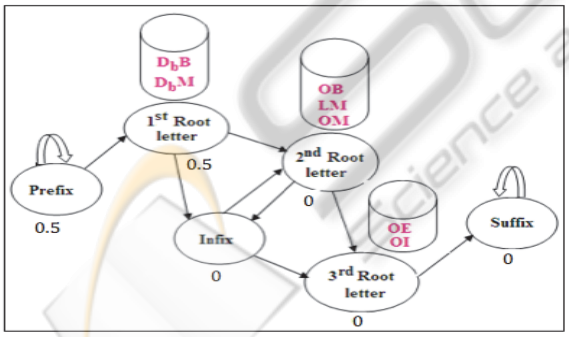


Figure 4: Architecture of a HMM_Root بعد (i.e: he walked away).

## 4.2 Training

We integrate linguistic knowledge, not only in the architecture using states for consonants and affixes, but also in the training, mainly in probabilities initializations. The training supervision is as follows:

- The starting state must be either the prefix or the first consonant; hence, we set to zero the probabilities of beginning at other states. Figure 4 shows the behavior of the HMM of the root" بعد ", which starts in the prefix or the first consonant " ب ".
- For initial emission probabilities of symbols, we have, depending on the corresponding state, supported some symbols (primitives) against others. Thus, we have estimated high initial probabilities for symbols describing the consonants of each root, prefixes and suffixes and we set near-zero probabilities for the primitives that may never appear for a root.
- Initial transition probabilities are closely chosen so that they incarnate Arabic properties of letters organization in a word: the order of prefix, suffix, infix and the three consonants (see transition in figure 4).
- Finally, we applied the same principle for all HMMs and used the Baum-Welch algorithm for training the 101 models.

It is important to mention that we propose a new mode of training, since one root HMM is not trained on samples of this root but through hundreds of different word samples, presenting various inflectional conjugations, following different schemes but generated from the same root (see figure 5). For example, the HMM of the root "دفع " in Figure 5. (b) is learned through several samples of different words ( دفع, مدافع, دفعتا, يدفعون, اندفعت, دافعت ), derived from the same root " دفع " under different schemes namely ( انفعل, مفاعل ، فعل , and فاعل ) and having different conjugations (past, present, singular).

Note that if we want to expand the vocabulary by hundreds of words derived from a root, we just need to create a new HMM for the new root; train it on samples of new words and add it to the system without affecting the other already trained HMMs.

## 4.3 Recognition

For recognition, we use Viterbi algorithm on trained HMMs. Indeed, the system works as follows: each HMM takes as input the word structural primitive, in order to calculate the probability of recognizing this word by every HMM. Then, the system sorts these probabilities to choose the maximum and retains the corresponding HMM root. Figure 6 illustrates this process for the word " تدرس ".

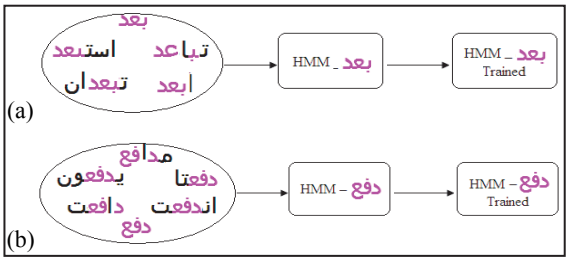


Figure 5: (a): Training of the HMM of the root **بعد** (i.e: to walk away), (b): Training of the HMM of the root **دفع** (i.e: to push).

# 5 EXPERIMENTATION

## 5.1 Training

The experiments were conducted on a database of more than 17000 samples of words of a large vocabulary generated from 101 Arabic tri-consonant roots. We used 11600 samples for training and the rest for recognition. The samples are written in 3 Arabic fonts (Regular AGA_Abasan, AF_Buryidah and Arial) characterized by variability of the forms of letters (disappearance of loops, missing legs, overlapping, etc.). Note that, the extraction of primitives has succeeded at 97%. Table 3 describes the configuration of our experiments (training corpus size per output, etc.).

Table 3: Training and test corpus sizes.

| Number of trained outputs | Training set size | Test set size |
|---|---|---|
| 101 roots | 11600 | 5757 |
| 1 root | Around 114 | Around 57 |

## 5.2 Recognition

In the test phase, we obtained the following overall recognition rate: top1= 88.18%, top2 = 96.92% and top3 = 98.26%. Table 4 presents some word recognition probabilities and the corresponding Viterbi paths. Table 5 illustrates some root classification rates.

Table 4: Word recognition probability.

| Word | Best path | Best probability | Recognized root |
|---|---|---|---|
| يطرق | P R1 R2 R3 | $3.1\ 10^{-4}$ | طرق |
| فرطوا | R1 R2 R3 S S | $1.582\ 10^{-5}$ | فرط |
| إفساد | P R1 R2 In R3 | $2.946\ 10^{-4}$ | فسد |

Table 5: Recognition rate by root HMM.

| Root | Recognition rate |
|---|---|
| قطع | 100 % |
| فجر | 97.5 % |
| خلف | 100 % |
| ركل | 95.91 % |

## 5.3 Collisions Analysis

To study and identify collisions, we evaluated our system in terms of recall. As shown in figure 8, most of the words were well classified. In fact, the majority of samples that have not been well recognized, derive from similar roots (that have the same structural primitives). For example, the words **حرقت** and **عرضت** deriving respectively from roots" **حرق** " (ie: to burn) and " **عرض** " (ie: to display), have exactly the same structural primitives: OB-AE-LDaB-DaE (see figure 7). This is the same for more than eight couples of roots such as " **كبس** " (ie. to press) and " **لبس** " (ie. to wear) (see figure 7). Even though many roots (consequently, many words) are similar, ambiguities could be easily resolved by the use, in addition to global primitives, of some local features to distinguish similar letters.

# 6 COMPARISON WITH RELATED WORKS

Our approach always produces higher success rates than the current approaches with regard to the vocabulary size as illustrated in table 6.

Note that we are comparing our approach to other approaches which 1) use linguistic knowledge like (Kanoun et al., 2005) and (Kammoun and Ennaji, 2004) and/or 2) are conceived for large vocabularies like (Kanoun et al., 2005), (Kammoun and Ennaji, 2004) and (Touj et al., 2007), even though the experimentation lexicon of some of them is not wide enough.

# 7 CONCLUSIONS

This paper describes our approach based on the Markov method for the training and the recognition of roots of Arabic words, from their descriptions in global structural primitives. Indeed, we have proposed a system based on 101 HMMs for the classification of 101 roots of 5757-sized vocabulary.

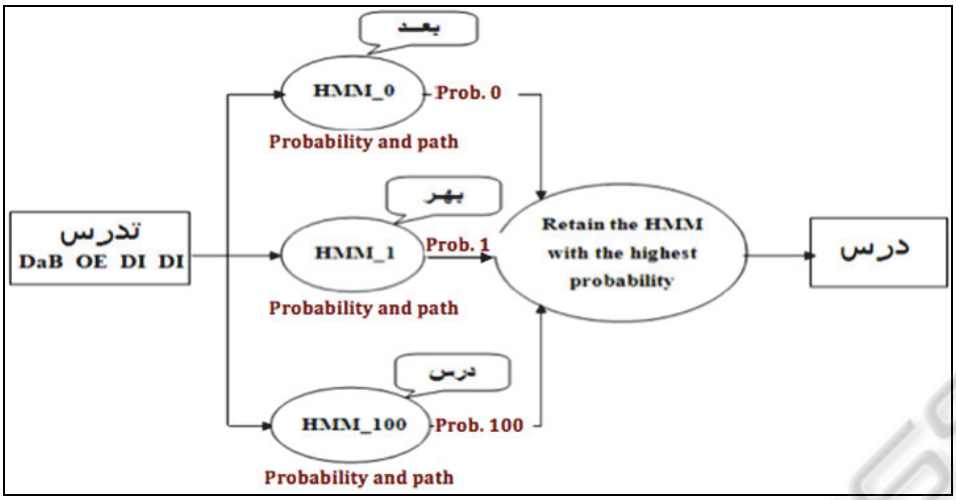


Figure 6: General architecture of word root recognition system.

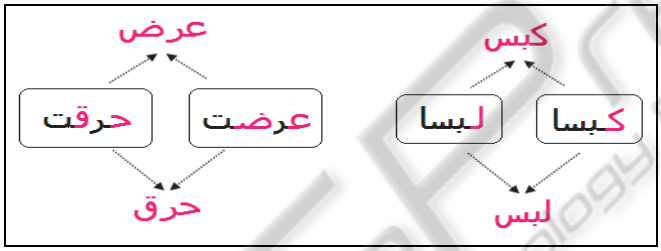


Figure 7: Example of different words with similar roots (having the same primitives).

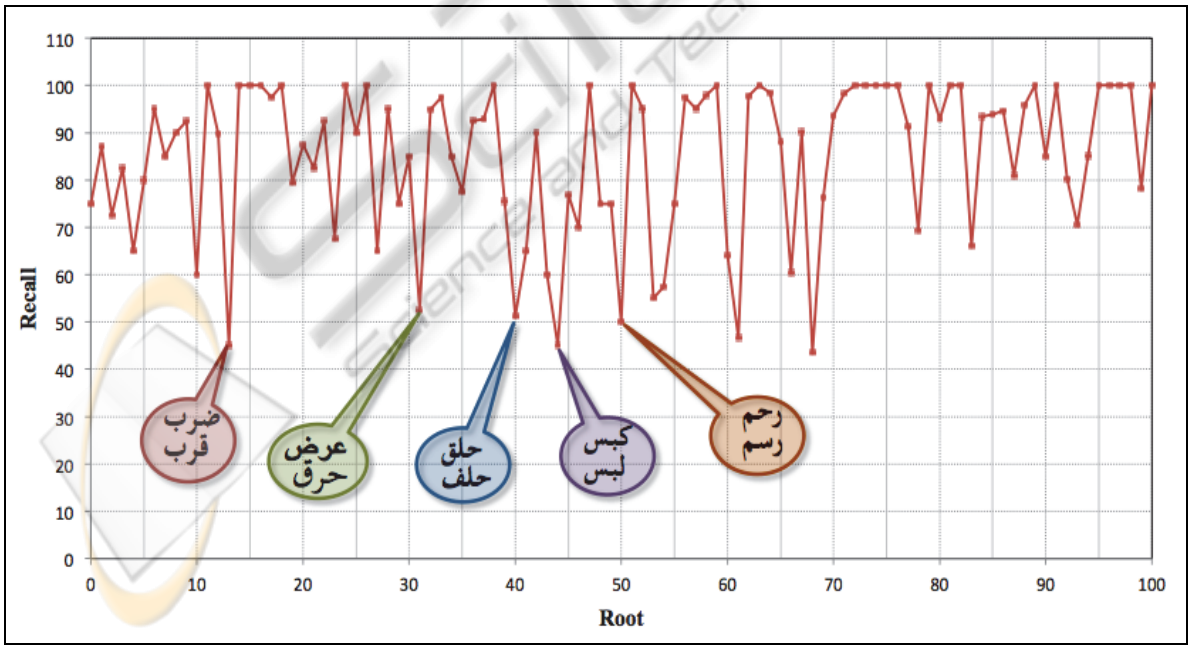


Figure 8: Scores of root HMMs.

Table 6: Proposed approach vs. current approaches.

| Approach | Writing | Vocabulary size | Top1 | Top2 | Top3 |
|---|---|---|---|---|---|
| Analytic (Kanoun et al., 2005) | Printed | 1000 | 74 | 81.2 | 83.9 |
| Analytic (Kammoun and Ennaji, 2004) | Printed | 1423 | 81.3 | 95.7 | 99.7 |
| Analytic (Touj et al., 2007) | Handwritten | 25 | 88.7 | - | - |
| Holistic (Our approach) | Printed | 5757 | 88.1 | 96.9 | 98.2 |

The proposed architecture of HMMs is personalized. It embodies knowledge about the linguistic properties of construction of Arabic words around the roots. This knowledge inspired us in the choice of states of consonants and affixes (prefix, infix, and suffix) and the choice of transitions between these states. The learning was performed on a database of over 11000 samples of words. The training of one HMM of a given root is performed via hundreds of words derived from this root following various schemes and different conjugations. Very satisfactory rates of recognition (top1=88% and top2=96.92%) were obtained in a phase of test made on more than 5700 samples.

To improve our solution, we conducted an analysis of the collisions which allowed us to encircle the problems of ambiguity between roots aiming to resolve them at the level of post-treatment phase. Indeed, the most striking problem affects the similar roots that present the same global primitives although their letters are different. Hybridizing the approach, by a local refinement of these particular letters, could significantly increase the scores of the system. Finally, despite that models of hidden Markov are known by their robustness absorption of writing variability and despite that they are usually used in analytical approaches (with local primitives), we used global primitives and we were able to reach an interesting rate of recognition, thanks to the integration of morphological knowledge.

Our approach could be also directly applied on handwriting since global primitives are easy to extract. Then, in the mean run, since we have been already reassured regarding the linguistic based structures of HMMs used just on global features, we could switch to the use of local primitives such as densities, invariable moments, Hu moments, etc.

# REFERENCES


Avila, M. (1996). Optimisation de modèles markoviens pour la reconnaissance de l'écrit. *PHD Thesis*, University of Rouen.

Ben Amara, N., Belaïd A., Ellouze, N. (2000). Utilisation des modèles markoviens en reconnaissance de l'écriture arabe : Etat de l'art. *Colloque International Francophone sur l'Écrit et le Document (CIFED)*. Lyon, France, pp 181-191.

Ben Cheikh, I., Kacem, A. and Belaïd, A. (2010). A neural-linguistic approach for the recognition of a wide Arabic word lexicon. *17th Document Recognition and Retrieval Conference*, part of the *IS&T-SPIE Electronic Imaging Symposium*, San Jose, CA, USA, January 17-22, 2010, SPIE Proceedings, pp 1-10.

Ben Cheikh, I., Belaïd, A. and Kacem, A. (2008). A novel approach for the recognition of a wide Arabic handwritten word lexicon. *19th International Conference on Pattern Recognition (ICPR)*, IEEE, Tampa, Florida, USA, pp. 1-4.

Bejaoui, M. (1985). Etude et réalisation d'un système expert appliqué à l'analyse morpho-syntaxique de phrases en langue arabe: méthode ascendante. *PHD Thesis*, University Paul Sabatier, Toulouse (Sciences), France.

Ben Hamadou, A. (1993). Vérification et Correction Automatiques par Analyse Affixale des Textes Ecrits en Langage Naturel. *PHD Thesis*, Faculty of Sciences of Tunis, Tunisia.

Cheriet, M., Beldjehem M. (2006). Visual Processing of Arabic Handwriting: Challenges and New Directions. *Summit Arabic and Chinese Handwriting Recognition,* Springer, India, September 27-28, pp 1-21.

El Yacoubi, A. (1996). Modélisation markovienne de l'écriture manuscrite, application à la reconnaissance des Adresses postales. *PHD Thesis*, University of Rennes 1.

George Saon, Abdel Belaïd (1997). High Performance Unconstrained Word Recognition System Combining HMMs and Markov Random Fields. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), Volume 11(5)*, pp 771-788.

Kammoun, W., Ennaji, A. (2004). Reconnaissance de Textes Arabes à Vocabulaire Ouvert. *Colloque International Francophone sur l'Écrit et le Document (CIFED)*, France.

Kanoun, S., Alimi, A., Lecourtier, Y. (2005). Affixal Approach for Arabic Decomposable Vocabulary Recognition: A Validation on Printed Word in Only One Font. *Eighth International Conference on Document Analysis and Recognition (ICDAR)*, IEEE

Computer Society, Seoul, pp 1025-29.
Kanoun, S. (2002). Identification et Analyse de Textes Arabes par Approche Affixale. *PHD Thesis*, University of Science and Technology of Rouen.
McClelland, J. L., Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. In *Psychological Review*, 88: pp. 375-407.
McClelland, J. L., Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. In *Journal of Experimental Psychology: General*, pp.159-188.
Touj, S., Ben Amara, N., Amiri, H. (2007). A Hybrid Approach for Off-line Arabic Handwriting Recognition Based on a Planar Hidden Markov Modeling. *9th International Conference on Document Analysis and Recognition (ICDAR)*, IEEE Computer Society, Brazil, pp 964-968.