2016 15th International Conference on Frontiers in Handwriting Recognition

A Unified French/English syllabic model for handwriting recognition

Wassim Swaileh*, Julien Lerouge† and Thierry Paquet‡
LITIS Laboratory - EA 4108
Normandie University - University of Rouen
Rouen, France
**wassim.swaileh2@univ-rouen.fr,* †*julien.lerouge@insa-rouen.fr,* ‡*thierry.paquet@univ-rouen.fr*

***Abstract*—In this paper we introduce a new unified syllabic model for French and English handwriting recognition, based on hidden Markov models (HMM). The recognition system training and recognition components such as optical models, lexicons and language models are designed to be language independent. In this purpose a syllable based model is proposed for French and English. This model is evaluated and compared to n-gram character and words models. A promising performance is achieved by the syllabic model, which meets the words model performance, with the advantage of a reduced system complexity. Furthermore, the unification of likely similar scripts improves the system performance over all models considering the English and French languages. The French RIMES and the English IAM datasets are used for the evaluation.**

***Keywords*-Syllables, Handwritten Text Recognition, Language model**

I. Introduction

For a few decades, most handwriting recognition systems were designed for recognizing each language individually, even if they share the same set characters such as French and English, or Arabic and Persian. Indeed, Languages of the same origins often share their character sets and/or glyph shapes. For example the Latin-root languages share at least 21 characters [1], and Arabic, Persian and 12 other languages share at least 28 characters [2]. Inspired from this fact, almost all multilingual or multiscript recognition systems are designed to work with a unified character set [3][4][5][6].

In the literature there are different possible approaches for developing multilingual recognition systems that can be classified into two categories: selective approaches and unified approaches.

The selective approach category includes two approaches. The first approach applies different individual recognition systems, in a competitive way, for the same samples and selects the one which provides the best results. The second approach first detects the language of the samples before selecting an appropriate recognizer for the target language script detected[6]. The comparison of the scores obtained by several recognizers, which have different error ranges, and the pre-detection of the language, which is a complex task, represent the major problems of these approaches. By exploiting the multilingual MAURDOR dataset, [4] proposed an English, French and Arabic multilingual recognition system and [7] proposed and English, French multilingual recognition system, which belong to this category.

The unified approach category uses a single system for recognizing any language, possibly using multiple scripts. Several multilingual or multiscript recognition systems are proposed in the literature. In [3] the authors propose a multilingual system for Arabic and Latin handwriting recognition. In [8] the authors propose a handwriting recognition system for intermixed language scripts such as Latin, Devanagari, and Kanji. A unified network-based handwriting recognition system for Hangul and English language was proposed by [6]. These approaches have the advantage of having a single system for all recognition tasks, avoiding the problems envisaged in the selective approaches. The disadvantage of the unified approaches is the proportional increase of the system's complexity regarding to the number of languages due to their direct effect on lexicon size, which affects the overall system performance.

Almost all the proposed unified approaches are such that the unified multilingual system have lower performance in comparison to the specialised monolingual systems. This may be explained by the effect of the lexicon and language model fast expansion, due to the absence of alphabetical and/or lexical shared units between the considered languages and scripts.

In this paper we propose a unified syllabic model for French and English handwriting recognition. For one single language a syllabic model exhibit a limited complexity in comparison to a word model, while recognition performance decrease very slightly when using a syllabic model. This property allows combining multiple languages in a unified syllabic model while maintaining an acceptable complexity in terms of lexicon size and statistical language models, maintaining equivalent recognition performance. We also show that unification of similar scripts in a single recognition system leads to better trained optical character models, due to the benefit of sharing training datasets between the languages, thus providing better performance of the unified system. This is observed for any kind of model used (characters, syllables, words) on the experiments carried on the RIMES and IAM datasets.

This paper has the following organisation: the theoretical

2167-6445/16 $31.00 © 2016 IEEE
DOI 10.1109/ICFHR.2016.97

basis of written syllabic models is presented in part 2 and we also give a brief overview of the syllabification method used for the experiments. We present the structure of the recognition system in part 3. The experiments are presented and analyzed in part 4, before drawing some perspectives of this work.

II. Syllable Based Modelling Approach

The syllable plays an important role in the organization of speech and language [9]. The name "syllable" is sometimes defined physiologically as a continuous unit of spoken language which consists of a sound or group of sounds uttered in one breath [10], [11]. The segmentation of speech into syllables can be achieved using acoustic units or phonological units [12], and syllables produced by these two models are not always compatible [10]. Most phoneticians agree that a syllable is basically composed of a rhyme that is preceded by an onset (one or more consonants "**C**" optionally comes at the beginning of the syllable). Inside a rhyme, the nucleus (usually a vowel "**V**") is the constitutive element of the syllable. This is followed by a coda (one or more consonants "**C**" at the end of the syllable) [10]. The languages differ from each other with respect to topological parameters as optionality of the onset and admissibility of the codas. For example, the onsets are mandatory in German while the codas are prohibited in Spanish [13]. In French, the nucleus is always considered as a vowel. Thus, counting the number of syllables pronounced in a French utterance should be equivalent to counting the number of pronounced vowels [10]. In English, there are syllabic consonants. Thus, counting the number of syllables pronounced in an English utterance should be equivalent to counting the number of pronounced vowels added to the number of syllabic consonants (which are limited to a few consonant types, as "l" in "bottle" or "n" in "button") [10].

Considering written languages, and according to [14], the spelling syllable may differs from the phonetic syllable depending on the language considered. Considering the French language, all "e" are considered silent when placed between two consonants or placed at the end of a word. Hyphenation rules separate the double consonants even if they are pronounced as a single consonant. For example, graphically there are three syllables in the French word **pu-re-té** even if we pronounce it as [**pyr-te**] (two phonetic syllables). The authors of [15] have classified the syllables in three different categories:

- A **phonetic syllable** is composed of a combination of phonemes that are pronounced in a single breath.
- A **graphemic syllable** represents a faithful transposition of phonetic syllabification in the spelling of the word.
- An **orthographic syllable** applies hyphenation rules that must be adhered to writing.

It seems difficult to conciliate these different views of specialists, but in any case, only graphemic or orthographic syllables provide a decomposition of writing that is likely to have an impact on a recognition system. In this study, we chose to use the orthographic representation of syllables provided by the French computerized and syllabified lexical database Lexique3 [16] and the free English language hyphenation dictionary [17] which was adapted from the Grady Ward's public domain English hyphenation dictionary.

The Lexique3 database provides an orthographic syllabic decomposition of a lexicon of almost 142,695 French words into 9,522 syllables only. It therefore constitutes a knowledge base from which our French syllabic model is developed. The free English language hyphenation dictionary contains 166,280 words decomposed into 21,991 syllables. However, despite their relatively large size, it quickly becomes out that these databases by far do not cover the French or English vocabularies. For example, the Lexique3 database covers only 69.83% of the vocabulary of the RIMES dataset, which is one of the training and reference dataset for French handwriting recognition. Similarly, the English hyphenation dictionary covers only 54.42% of the IAM dataset vocabulary (see left side of Figure 1). Therefore, we have to find a general way for generating a syllabic decomposition of any corpus.

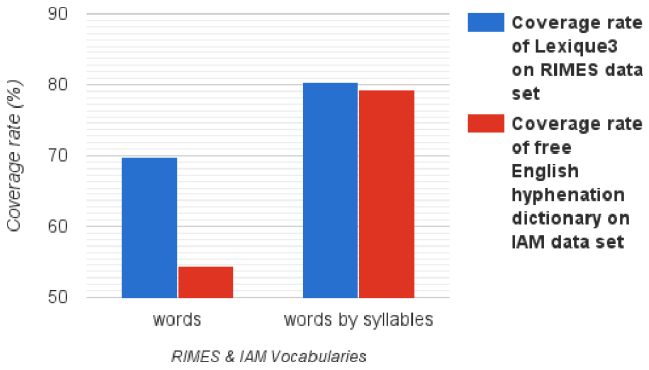

Figure 1. Coverage rates of Lexique3 and EHD on RIMES and IAM datasets respectively for words and syllables decomposition.

For this purpose it is necessary to develop an automatic syllabification method. In [18] a supervised syllabification method is proposed [1] that exploits Lexique3 or EHD in order to provide a syllabic decomposition of any word not present in these resources. The syllabification method exploits the lexical and phonetic similarities between the target word (to be decomposed into syllables) and the syllabified words that belong to the dictionaries. Right side of figure 1 shows the coverage ratios of Lexique3 and the free English hyphenation dictionary (EHD) on the target datasets RIMES and IAM respectively when using the syllabic decomposition.

[1]The source code of the syllabification method is accessible at http://swaileh.github.io/Syllabifier

Although insufficient, we can see the increased coverage when using a syllabic decomposition of the RIMES or IAM data set. The remaining words that are not syllabified will be decomposed into their character sequence, thus building a hybrid character / syllabic model.

III. The Handwriting Recognition System

Our recognition system is based on optical models of the characters based on hidden Markov models (HMM). The essential components in the design of our system are the alphanumeric character models. We have in total 100 character models for the RIMES dataset and 80 character models for the IAM dataset, by considering the white space between words as a character. The two character sets are combined in a unified French/English character set that contains 102 alphanumeric characters. Our recognition system is composed of four main stages; namely, prepossessing, optical models training, lexicon and syllable lexicon generation and language model training. The recognition step is performed following a two passes decoding algorithm.

A. Pre-processing

During the prepossessing, we proceed to the detection of text lines in the text blocks in order to improve the rectangular positioning provided in the RIMES dataset because it provides quite noisy lines. Indeed rectangular areas provide lines that overlap with the line above or below. The automatic method for line detection is described in [19]. Considering the IAM dataset, the rectangular zoning provides quite clean lines. Then line images are adjusted horizontally and vertically (deskew and deslant) and scaled to a 96 pixels height, preserving aspect ratio.

B. Optical Character Models

We define a unified character set that includes all possible French and English characters and symbols that are present in the RIMES and IAM datasets, thus amounting to a total of 102 models. Notice that the IAM data set contains two more different characters than the RIMES dataset. Every alphanumeric character is associated to an HMM model, these optical models are able to model the character variations over all character observations.

Optical models exploit HoG (Histogram of Gradients) characteristics extracted from the text line images by using a sliding window of 20 pixels width (a frame). Frame horizontal displacement is 2 pixels. Each frame is described by a 70 dimensional real valued vector. 64 features encode the HoG description, and 6 encode a geometric description of the frame. Generally, the internal structure of the character optical Models (HMM) is defined by a fixed number of hidden states and for each of them, a fixed size Gaussian mixture is also determined. We chose to use mixtures of 20 Gaussians, which guarantee a description ability fairly accurate for each frame. Determining the number of hidden states is an optimization problem. An overestimated number of the hidden states leads to over-trained models. An underestimated number of states leads to inadequate specialized models. This problem has been addressed in [20], [21].

We have been inspired by the proposed method in the first reference that is based on the Bakis method in order to optimize the number of states of each character model. Once a first training of one initial set of character models has been carried out with a fixed number of states, we compute the average number of frames T per optical model using a forced alignment of the corresponding model on the ground truth of each image on the frame sequence. The number of states E of the corresponding model is then defined as a fraction of T ($E = \alpha.T$).

A new training process (Baum-Welch) is performed for the new parametrized models that have been created, according to parameter α. Then we perform a final decoding without ground truth (no forced alignment) using the trained models and the character recognition rate (CRR) is computed. The operation is repeated for different values of α (increasing values between 0 and 1). Finally we select the most accurate models based on a criterion combining the average CRR and the alignment rates of the models on the training examples. Indeed, excessively long character models tend to maximize the recognition rate but at the expense of misalignments on shorter examples. This criterion is tested at each iteration of the Baum-Welch training procedure. Training is stopped when the criterion reaches its highest value. We then obtain optimized optical models.

Training the optical models is first performed on the RIMES 2011 training dataset, which contains 10,963 images of ground trothed text lines that are segmented from 1500 images of paragraphs written by different writers in different writing conditions. By checking the optimized number of states per characters HMM, we found that the shared characters and symbols between the French and English language gets equivalent number of state after the optimisation processing. For example the HMMs optimization process of RIMES character set attributes 10 states for the optical model of character **A** while IAM HMM's number of states optimization process attributes 12 states for the same character.

For our experiments on specialised (mono-lingual) optical models, we applied the same optimization and training procedures for both French and English specialized optical models separately. Training the unified model starts using the RIMES training dataset only. Then training is continued using the IAM training dataset which contains 11,349 ground trothed lines. Through a periodic validation process, we select the best optical models that reach optimal performance on the unified RIMES and IAM validation datasets, thus preventing the training process to specialize on the IAM dataset.

C. lexicons and language models

The third step in building our system is the definition of the vocabularies and language models to be used by the system during decoding. A unified French & English corpus is collected from the texts of the RIMES and the IAM dataset (learning set and validation set). From these corpora, we generated the unified vocabularies and n-gram back-off language models. The first model is a model without lexicon, it is a n-gram character model. The unified n-gram models are estimated by using the MIT language model toolkit using Kneser-Ney smoothing. The second model is a word n-gram language model. The third model is a syllable n-gram model. The vocabularies are vocabularies of syllables obtained through the syllabification method presented in [18], and the language models are N-grams of syllables estimated on the same unified corpora (French/RIMES & English/IAM).

D. Recognition step

Our system is characterized by a tow pass decoding. The first pass processes the test sample by performing decoding according to the Viterbi algorithm with pruning over time (time synchronous Viterbi beam search). The optical models are used for the character model, or they are concatenated to form words or syllables of the lexicon to be used, depending on the recognition scenario.

According to the scenario, the decoding algorithm uses character, syllable or word bi-gram, to produce a network of characters, syllables or words hypotheses. Two important parameters guide this first decoding pass: they are the language model scaling parameter γ and the word insertion penalty parameter β that controls the insertion of too frequent short words. These two parameters need to be optimized for optimum coupling of the optical models with the considered language model, because these two models are estimated independently from each other during training. The second decoding pass analyses the hypotheses network provided by the first pass using a language model of higher order n-gram which allows re-weighting the first hypotheses. This last step provides the final output recognition of the text lines.

When decoding we seek the $\hat{W}$ word sequence that maximizes the posterior probability $P(W|S)$ among all possible sentences W. Using Bayes's formula and introducing the two hyper-parameters defined above, we finally arrive to the formula given in equation 1 which governs the decoding step. In this formula, S represents the sequence of observations extracted from the image and $P(S|W)$ represents the likelihood that the characteristics S are generated by the sentence W, it is deduced from the optical model. $P(W)$ is the prior probability of the sentence W, it is deduced from the language model.

$$\hat{W} = argmax_w P(S|W)P(W)^{\gamma}\beta^{length(W)} \quad (1)$$

IV. Evaluation

To optimize and test the performance of our specialized (mono-lingual) systems, we used the RIMES validation data set that contains 764 lines retrieved from 100 images of paragraphs written by different writers. Half of this validation set was used for optimizing the decoding parameters and the other half was used for testing French models (character, word and syllables) recognition performance. Similarly, we used 1033 lines of the IAM dataset extracted form 120 documents written by different writers for evaluating the performance of the English characters, words and syllables models. We combined the evaluation and test sets of RIMES & IAM datasets in order to optimize the decoding parameters and testing the performance of the unified characters, words and syllables models.

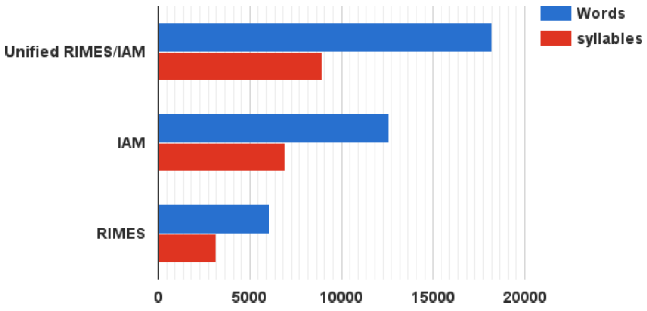

Figure 2. Size of syllables lexicons derived from RIMES & IAM word lexicons.

We studied the behavior of the proposed unified syllabic model in comparison with characters and words model performances from two different points of view; model complexity and recognition performance. The model complexity is evaluated as a function of lexicon size. The statistical analyses of RIMES and IAM lexicon sizes shows the compact size of the syllabic lexicon which is considerably reduced compared to the words lexicon (52.02% of RIMES word lexicon, 54.86% of IAM word lexicon and 49.09% of the unified RIMES & IAM word lexicon). Figure 2 illustrates the reductions ratios of the lexicon size between RIMES, IAM and their unified words and syllables lexicons. After mixing RIMES and IAM lexicons and eliminating duplicated words, the unified word/syllable lexicons are reduced regarding to the total sum of RIMES and IAM word lexicons by 2.4% and by 11.3% from the total sum of the two datasets syllable lexicons. The lexicon size affects directly the recognition system performance while the system complexity increased proportionally to the lexicon size. Thus, the compact size of the syllabic lexicon leads to a recognition system of limited complexity.

We evaluated the language models performance through their perplexity measures (see table I). The language models perplexities are calculated for the specialized and unified language models on RIMES and IAM corpora separately

for the mono-lingual language models and together for the unified language models.

	Data sets		
Language models	IAM	RIMES	RIMES + IAM
Specialized words model	19.63	12.23	—
Specialized syllables model	17.25	9.29	—
Unified word	21.27	19.28	19.8
Unified syllables model	19.08	16.39	17.31

Table I
LANGUAGE MODELS PERPLEXITY EVALUATIONS ON RIMES, IAM AND RIMES+IAM DATA SETS

We considered only the word and syllables language models perplexities because they reflect similar behavior for the two models (words and syllables) when the language models perplexities are measured for each unified and specialized models on unified and specialized datasets. From table I, we note that the perplexity of the unified words and syllables models have an equivalent increase compared to the specialized words and syllables models when evaluating the perplexities on the mono-lingual datasets. Thus, we can conclude that the syllabic model represents an alternative model for the word model with the advantage of reduced recognition system complexity when using syllabic model.

The recognition performance was evaluated by using the word error rate (WER%). We carried out all evaluation tests under the same condition without considering any out of vocabulary words (OOV). The evaluation was performed in three phases, the first phase contains the evaluations of the specialized models that contain specialized (mono-lingual) optical and language models. The evaluation of the unified models (unified optical and language models) was performed in the second phase. In the third phase we evaluated mixed structure recognition systems that contains unified optical models operating with specialized language models.

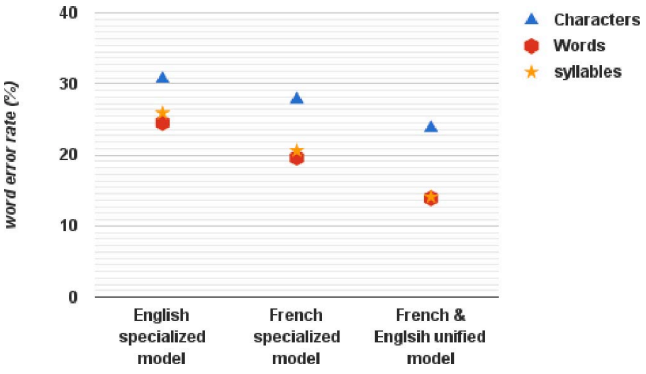

Figure 3. Test results of the specialized (French and English) models and the unified models

Every specialised French/English model (characters, words and syllables) was tested on its corresponding RIMES/IAM test dataset. Figure 3 illustrates the performance of the unified models versus the specialized model.

The test results of the specialised models shows that the word models achieves the best performance with a minimum word error rate in both French and English tests (19.6% WER on RIMES and 24.5% WER on IAM). It also shows that the syllabic model achieves very similar performance to the word models with 20.6%WER and 25.9%WER for RIMES and IAM syllabic models respectively. The worst performance are registered for the character models for both French (27.8% WER) and English (30.7% WER).

The tests results of the unified models shows superior performance than the specialized models over all unified models; characters, syllables or words. This time, the unified syllabic model achieves almost similar performance to the unified word models with a difference of 0.2% only.

The third step of the evaluation was to consider the contribution of the unified optical model (trained on both the RIMES and IAM datasets) to the performance of the specialized models (French or English language models). The idea was to replace the specialized optical models of the specialized models by the unified one and doing the test on the RIMES and IAM test datasets separately. These specialized recognition systems belong to the selective approaches category because they are specifically selected for the corresponding language.

The results reported in table II show a significant performance increase with the mixed English recognition system which uses the unified optical character models. Surprisingly the mixed French system exhibits slightly lower performance than the specialized systems and for any type of model (word, syllable, character). This may be explained by the different amount of training data between the RIMES and the IAM datasets, where the French RIMES dataset size is only $1/3$ of the English IAM dataset size. The unified optical model may have shifted towards English character during training.

Models	French models		English models		Unified models
	specialized	mixed	specialized	mixed	
Characters	27.8	30.8	30.7	23.2	23.8
Words	19.6	22.3	24.5	10.4	13.9
Syllables	20.6	24.2	25.9	12.1	14.1

Table II
WER (%) FOR THE UNIFIED, SPECIALISED AND MIXED MODELS

V. CONCLUSION

In this study we proposed a unified French/English syllabic model for handwriting recognition. This model offers many advantages over the character model which models badly the words and over the word model which models only a limited number of words that are found in the vocabulary

of the training corpus. The advantages of this model lies in its limited complexity, since it works with a reduced syllables lexicon. It follows an n-gram model of syllables which is itself more compact, so better parametrized, and therefore easier to optimize. To generate the syllabic model we stand on the lexique3 and on the free English hyphenation dictionary which propose the orthographic modelling of syllables for the French and English language respectively. The unification of the French and English syllables lexicon shows and interesting size reduction compared to the size of the two syllables lexicons. This is because the two languages have a lot of shared syllables knowing that they have different syllabication rules. It will be interesting to find optimal syllabification rules that maximize this criteria, thus optimum compact lexicon could be found. Moreover, it should be interesting to know if this syllabic model can offer the same interest for modelling some other languages for the same latin script. Exploring the syllabic model for some other scripts such as arabic script for example is also another perspective.

ACKNOWLEDGMENTS

We thank Ms. Elise Ryst and Mr. Christopher Coupeur who offered us their advice as language specialists.

REFERENCES

[1] D. Diringer, *The alphabet: a key to the history of mankind*, 1951.

[2] V. Märgner and H. El Abed, *Guide to OCR for Arabic scripts*. Springer, 2012.

[3] Y. KESSENTINI, T. PAQUET, and A. BENHAMADOU, "A multi-stream hmm-based approach for off-line multi-script handwritten word recognition," *a a*, vol. 1, p. 1, 2008.

[4] B. Moysset, T. Bluche, M. Knibbe, M. F. Benzeghiba, R. Messina, J. Louradour, and C. Kermorvant, "The a2ia multi-lingual text recognition system at the second maurdor evaluation," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 297–302.

[5] M. Kozielski, M. Matysiak, P. Doetsch, R. Schloter, and H. Ney, "Open-lexicon language modeling combining word and character levels," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, 2014, pp. 343–348.

[6] J. J. Lee and J. H. Kim, "A unified network-based approach for online recognition of multi-lingual cursive handwritings." IWFHR, 1997.

[7] M. Kozielski, P. Doetsch, M. Hamdani, and H. Ney, "Multilingual off-line handwriting recognition in real-world images," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE, 2014, pp. 121–125.

[8] A. Malaviya, C. Leja, and L. Peters, "Multi-script handwriting recognition with fohdel," in *Fuzzy Information Processing Society, 1996. NAFIPS., 1996 Biennial Conference of the North American*. IEEE, 1996, pp. 147–151.

[9] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in rwth's system for off-line handwriting recognition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 935–939.

[10] E. Ryst, "Syllabation en anglais et en français: considérations formelles et expérimentales," Ph.D. dissertation, Paris 8, 2014.

[11] K. Brown, "Encyclopedia of language and linguistics," 2006.

[12] R. Ridouane, Y. Meynadier, and C. Fougeron, "La syllabe: objet théorique et réalité physique," *Faits de langue 37*, pp. 225–246, 2011.

[13] S. Bartlett, G. Kondrak, and C. Cherry, "On the syllabification of phonemes," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 308–316.

[14] D. Flipo, B. Gaulle, and K. Vancauwenberghe, "Motifs français de césure typographique," *Cahiers gutenberg n*, 1994.

[15] S. Roekhaut, S. Brognaux, and R. Beaufort, "Syllabation graphémique automatique à l'aide d'un dictionnaire phonétique aligné," 2012.

[16] B. New, C. Pallier, M. Brysbaert, and L. Ferrand, "Lexique 2: A new french lexical database," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 516–524, 2004.

[17] M. Hindson, " Free English language hyphenation dictionary ," http://hindson.com.au/info/free/free-english-language-hyphenation-dictionary/, 2016, [Online; accessed 21-Dec-2015].

[18] W. Swaileh, K. A. Mohand, and Paquet, "un modle syllabique pour la reconnaissance de l'criture," in *Ecrit et Document (CIFED), 2016 Colloque International Francophone sur lcrit et le Document 2016*. under submission, 2016.

[19] W. Swaileh, K. A. Mohand, and T. Paquet, "Multi-script iterative steerable directional filtering for handwritten text line extraction," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 1241–1245.

[20] K. Ait-Mohand, T. Paquet, and N. Ragot, "Combining structure and parameter adaptation of hmms for printed text recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 9, pp. 1716–1732, 2014.

[21] M. Zimmermann and H. Bunke, "Hidden markov model length optimization for handwriting recognition systems," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 369–374.