Une Méthode Syntaxique Pour la Reconnaissance Automatique de Caractères Amazighes Imprimés

[1]Youssef ES SAADY, [2]Ali RACHIDI, [1]Mostafa EL YASSA, [1]Driss MAMMASS

[1]IRF – SIC, Faculté des Sciences, B.P. 8106, Hay Dakhla, Université Ibn Zohr, Agadir, Maroc,
essaady2110@yahoo.fr, driss_mammass@yahoo.fr, melyass@gmail.com

[2]Ecole nationale de Commerce et de Gestion, B. P. 37/S Hay Salam, IRF – SIC, Faculté des sciences, Université Ibn Zohr, Agadir, Maroc, rachidi.ali@caramail.com

RÉSUMÉ. Nous présentons dans ce papier un système automatique de reconnaissance de caractères Amazighes imprimés isolés, basé sur une approche syntaxique utilisant les automates finis. Après des prétraitements sur l'image du caractère, des algorithmes appropriés permettent de construire la chaîne représentative du caractère à partir du codage de Freeman. La chaîne reconstruite est utilisée à l'entrée d'un automate fini qui reconnaît tous les caractères Amazighes. Cet automate global est construit à partir des automates de reconnaissance spécifique à chaque caractère Amazighe. Sur une base de données de caractères Amazighes imprimés isolés, les résultats expérimentaux montrent la robustesse de l'approche.

ABSTRACT. We present in this paper an automatic system for the recognition of the isolated printed characters Amazighes based on a syntactic approach using finite automata. After preprocessing on the image of the character, appropriate algorithms allow to build the representative chain of the character from the Freeman coding. The reconstructed chain will be used in the input of a finite automaton which recognizes all the Amazighe characters. This global automaton is built from specific automata of recognition for each Amazighe character. On a data base of isolated printed Amazighes characters, the experimental results show the robustness of the approach.

MOTS-CLÉS : Caractères amazighes, Grammaire régulière, Automate fini, Reconnaissance syntaxique.

KEYWORDS: Amazighes characters, Regular grammar, Finite automata, Syntactic recognition.

1. Introduction

Dans le domaine de la reconnaissance automatique des caractères, plusieurs recherches scientifiques ont été effectuées sur le caractère latin, arabe et chinois. Ce ci a permis le développement de plusieurs approches de reconnaissance automatique pour ces caractères. Par contre, le caractère Amazighe, appelé Tifinaghe, est très peu traité. Et pour extraire les informations Amazighes sur des supports, la reconnaissance automatique de ce caractère est devenue primordiale. Parmi les travaux qui lui ont été consacrés, on cite l'approche statistique de A. Djematen [1], l'approche géométrique proposée par A. Djematen [2] et la méthode basée sur la transformée du Hough de A. Oulamara [3]. Nous proposons, dans ce papier, une méthode automatique syntaxique de reconnaissance expérimentée sur une base de caractères Amazighes imprimés.

Nous présentons en première partie de ce papier les principales caractéristiques de la langue Amazighe. La seconde partie est consacrée à la présentation de la méthode syntaxique pour la reconnaissance de formes. Ensuite, nous présenterons, dans la troisième partie, la grammaire et l'automate fini qui reconnaît les caractères Amazighes. La quatrième partie de ce papier expose les résultats obtenus.

2. Ecriture Amazighe

2.1 Historique

L'alphabet Amazighe a subi des modifications et des variations depuis son origine jusqu'à nos jours et ce du libyque jusqu'au néotifinaghe en passant par le Tifinaghe saharien et le Tifinaghe touareg [4]. Nous retraçons ci-dessous les aspects les plus importants de chacune de ces variations.

Le libyque: Il s'agit des variétés de Tifinaghe les plus anciennes. Il existe deux formes du libyque, l'occidental utilisé le long de la côte méditerranéenne de la Kabylie jusqu'au Maroc et sans doute aux Îles Canaries. Et la forme orientale a été utilisée dans le Constantinois, en Aurès et en Tunisie.

Le Tifinaghe saharien: Cette variété est également appelée libyco-berbère ou touareg ancien. Elle contient des signes supplémentaires par rapport au libyque, plus particulièrement un trait vertical pour noter la voyelle finale /a/. Cette variété fut utilisée pour transcrire le touareg ancien mais ses inscriptions sont incomprises [5].

Le Tifinaghe touareg: Il existe au sein du Tifinaghe touareg quelques divergences dans la valeur attribuée aux signes qui correspondent aux variations dialectales

touarègues. Si d'une région à une autre, la forme et le nombre des signes peuvent changer, les textes restent en général mutuellement compréhensibles.

Le néotifinaghe: Le néotifinaghe désigne les systèmes d'écriture développés pour représenter les parlers Amazighes du Maghreb. La première variante fut celle proposée à la fin des années 60 par l'Académie berbère sur la base de lettres Tifinaghes touarègues. Elle est largement diffusée au Maroc et en Algérie.

2.2. Tifinaghe : L'alphabet Amazighe

L'IRCAM (Institut Royal de la Culture Amazighe) a proposé à l'Organisation de Standarisation Internationale (21/06/2004) [6] l'Alphabet Tifinaghe et ce dernier a été confirmé le 05/07/2004. La figure 1 ci-dessous illustre l'alphabet Tifinaghe et le plan Unicode associé attribuée par l'ISO et Unicode.

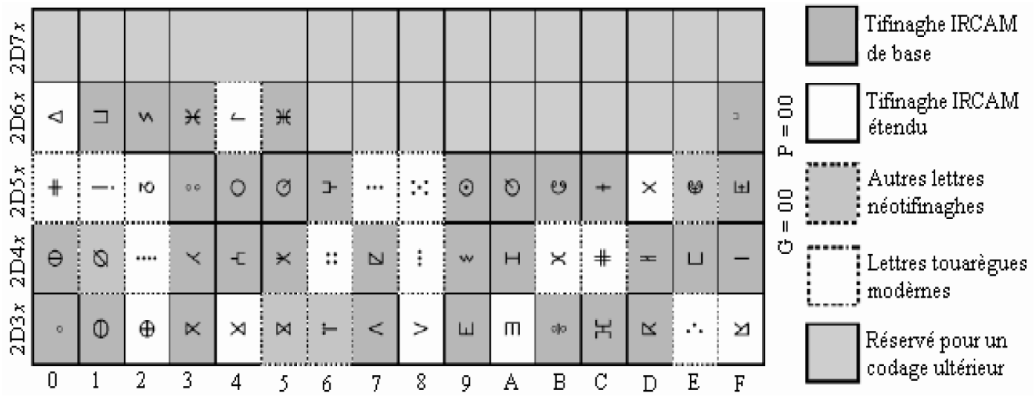

Figure 1 : *l'alphabet Tifinaghe et leur code Hexadécimal dans le format Unicode*

Avant de présenter notre système de reconnaissance de ces caractères, nous proposons tout d'abord un bref aperçu sur les méthodes syntaxiques.

3. Méthodes syntaxiques pour la reconnaissance de formes

Les méthodes orientées vers un fort pouvoir d'explicitation des informations de structure peuvent se regrouper sous la rubrique de méthodes structurelles. Elles englobent les méthodes syntaxiques, ancrées dans les langages formels. La notion de structure peut se ramener à l'existence d'un tout décomposable en un ensemble de parties et des relations entre ces parties.

Du point de vue des formalismes de représentation, les méthodes syntaxiques adoptent des représentations de type grammaire. Le vocabulaire des grammaires est associé à des composantes de la forme. Les relations entre ces composantes sont explicitées par des règles de production de ces grammaires.

La reconnaissance syntaxique d'une forme se compose de trois étapes principales :

– Prétraitement, qui améliore la qualité d'une image, par exemple filtrage, normalisation, squelettisation, perfectionnement, etc.;

– Représentation structurelle de formes, qui segmente l'image et représente les éléments structurels par un modèle grammatical;

– Analyse syntaxique, qui regroupe les deux tâches d'apprentissage et de décision. En effet, à partir de la représentation structurelle de la forme, elle tente, dans un premier temps, d'attribuer à chaque forme un modèle de référence (apprentissage) et dans un deuxième temps, elle décide si cette forme appartient à une classe donnée [7].

4. Reconnaissance syntaxique de caractères Amazighes imprimés

4.1 Caractéristiques des caractères Amazighes

Tout d'abord, nous introduisons les différents modèles de l'alphabet Amazighe qui comporte trente trois lettres (cf Figure 2). A la différence des caractères latins et arabes, l'écriture Amazighe n'est jamais cursive [1], ce qui facilite toute opération de segmentation. La majorité des modèles graphiques des caractères est composée de points, de petits cercles, et/ou de segments. De plus, les segments sont tous verticaux, horizontaux, ou diagonaux. Le problème est donc très différent de celui posé pour la reconnaissance des caractères latins et arabes qui comportent des courbes et des boucles.

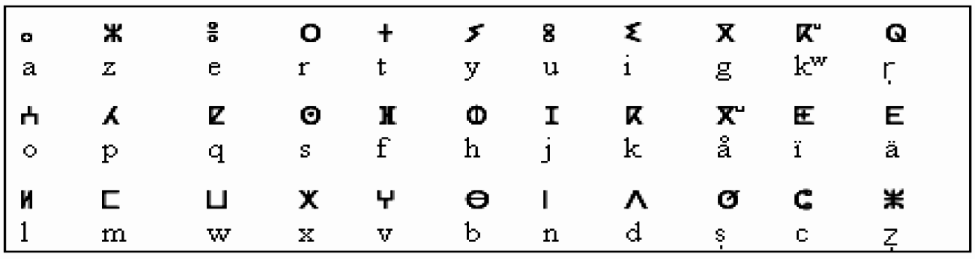

Figure 2: *Alphabet Tifinaghe IRCAM et sa correspondance latine*

4.2 Modélisation d'un caractère Amazighe imprimé

La représentation d'un caractère Amazighe par une grammaire régulière se fait en trois étapes. L'étape initiale est le prétraitement, où l'on construit le squelette du caractère. Ainsi, cette opération de squelettisation est invariable à la mise en échelle et robuste au bruit grâce aux caractéristiques des caractères traités citées ci-dessus. La seconde permet d'extraire des points caractéristiques du caractère étudié et enfin, la construction de la chaîne représentative du caractère est réalisée. En effet, nous avons utilisé l'algorithme de Zhang-Suen pour construire le squelette du caractère [10]. La figure 3, ci-dessous, présente quelques exemples de caractères et leurs squelettes obtenus par cet algorithme.

Une fois le squelette obtenu, nous cherchons à le décomposer en un ensemble de segments élémentaires. Tout d'abord, trois types de points caractéristiques seront extraits du squelette du tracé [11]: les points d'extrémité, les points de croisement et les points d'inflexion.

Pour cela, on analyse les 8 voisins de chaque pixel du squelette et on compte le nombre de transitions 0→1, selon le sens horaire (tab.1).

– Si le nombre de transitions est 1, alors on est à une extrémité.

– Si le nombre de transitions est ≥3, alors on est à un croisement.

– Si le nombre de transitions est 2, alors on est au milieu d'une continuité ou d'inflexion. Dans ce cas, on regarde la position des pixels de transition pour déterminer l'angle d'inflexion. Cet angle se calcule à partir de l'écart entre les positions des transitions 0→1 (Tab. 2).

a	b	c
h	**X**	d
g	f	e

Tab. 1: *Voisins de X (sens horaire de parcours): a, b, c, d, e, f, g, h*

Écart	**Angle**
1,7	45°
2,6	90° (angle droit)
3,5	135°
4	180° (angle plat)

Tab. 2: *Correspondance en angle*

Un algorithme de suivi de squelette permet de construire les segments d'un caractère. En effet, pour transformer le squelette en segments, on recherche uniquement les "extrémités" et les "croisements" et on les relie en passant par les "inflexions".

Le point de départ du premier segment est la première extrémité, appartenant au squelette. Le deuxième point de ce segment est ensuite recherché dans le voisinage immédiat du premier. Cette recherche faite selon le processus d'écriture du caractère Amazighe.

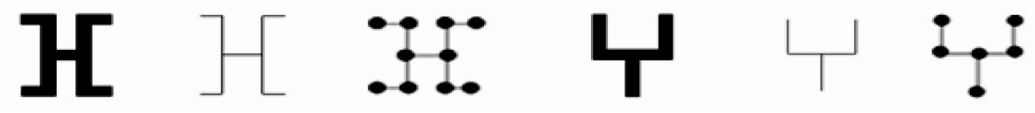

Figure 3*: Exemples de caractères Amazighes imprimés et leurs squelettes ainsi les points caractéristiques extraits*

Chaque caractère Amazighe est alors représenté par un ensemble de primitives. Pour représenter les caractères Amazighes, nous avons retenu les 8 directions de codage de Freeman. Dans l'exemple de la figure 3, le code relatif de la chaîne de Freeman pour le caractère ⵀ serait: 1775131315771, et celui du caractère ⵖ serait: 717313. Nous retenons le vocabulaire terminal X = {1, 2, 3, 4, 5, 6, 7} pour représenter les caractères Amazighes segmentés. Chaque caractère est représenté par une grammaire régulière. Par exemple, le caractère ⵀ de la figure 4 se décrit par la grammaire suivante: **G=(X, V, S, P)** où : X = {1, 3, 5, 7}, V = {A, B, C, D, E, F, G, H, I, J}, S est l'axiome de départ et les règles de productions se décrivent comme suit:

P:
S → 1A
A → 7B|7A
B → 5C
C → 1D
D → 3E
E → 1F
F → 3G
G → 1H
H → 5I
I → 7J|7I

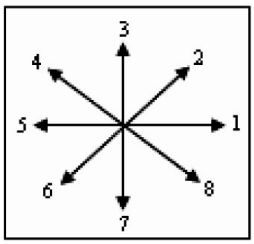

Figure 4 : *Les 8 directions de Freeman*

Une fois que les caractères sont représentés par des grammaires régulières, on peut les représenter par les automates finis. Par exemple, pour la grammaire précédente, on obtient l'automate fini de la figure 5 ci-dessous.

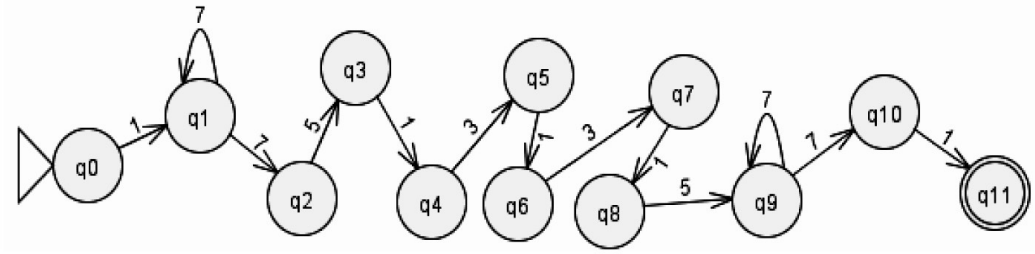

Figure 5: *L'automate fini qui reconnaît le caractère ⵀ*

Une fois qu'on a les automates de tous les caractères, on génère un automate global pour reconnaître tous les caractères Amazighes.

4.3 Construction d'Automate Canonique Maximal

Le plus grand automate pour l'échantillon d'apprentissage I_+ (l'échantillon qui représente les caractères Amazighes), s'appelle l'automate canonique maximal (ACM (I_+)). Il est construit en réalisant l'union de l'ensemble des automates acceptant chacun un caractère de l'échantillon d'apprentissage. Cet automate réalise un apprentissage par cœur de l'échantillon d'apprentissage. A titre d'exemple, l'automate représenté ci-dessous (figure 6) présente l'automate canonique maximal qui reconnaît les quatre caractères (ⵢ, ⵃ, ⵜ, ⵍ).

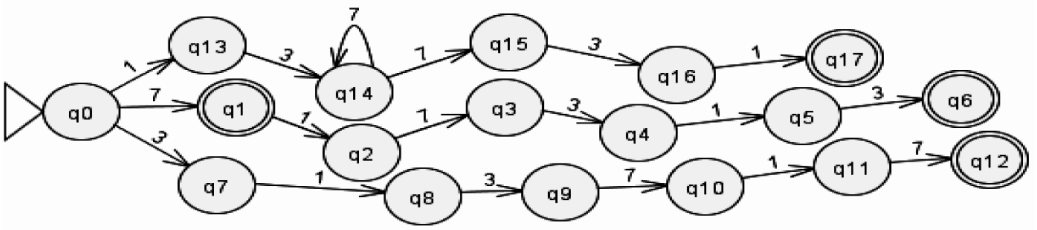

Figure 6: *ACM relatif à l'échantillon I+ qui représente les caractères (ⵢ, ⵃ, ⵜ, ⵍ)*

5. Expérimentations et résultats

Pour pouvoir expérimenter notre approche, nous avons créé une base de caractères Amazighes imprimés isolés, de différentes tailles, et sous forme d'images brutes. Elle est composée de 240 caractères (30 caractères pour chaque classe).

Pour valider notre approche, et dans premier temps, nous l'avons testé sur les caractères Amazighes imprimés (ⵅ, ⵢ, ⵃ, ⵜ, ⵡ, ⵛ, ⵍ, ⵎ) extraits de la base de caractères, déjà décrite précédemment. Ces caractères portent des caractéristiques tenues en compte par notre approche. Nous avons obtenu des résultats encourageants. En effet, sur les 240 caractères lus, 232 ont été reconnus, soit un taux de reconnaissance de 96,6%. Le tableau ci-dessous présente les taux des mauvaises affectations et de mauvais rejets. Ces erreurs proviennent de la forme de certains caractères non reconnus dont le squelette comporte des segments non orthogonaux.

Taux d'affectations à tort	**1,25 %**
Taux de rejets à tort	**2,08 %**

Tab. 3: *Pourcentages des erreurs de reconnaissance*

6. Conclusion et perspectives

Dans cet article, nous avons présenté un système pour la reconnaissance des caractères Amazighes imprimés, utilisant une approche syntaxique basé sur les automates finis. Des résultats encourageants ont été obtenus. Dans les futurs travaux, nous allons améliorer cette approche pour tenir compte des autres caractères non traités. En effet, nous utiliserons des opérateurs prétopotologiques pour repérer les points caractéristiques dans le squelette des caractères dont la structure n'est pas vectorielle. Ainsi nous comptons appliquer des algorithmes d'apprentissage pour générer l'automate final. En plus, nous traiterons le cas de caractères Amazighes manuscrits.

BIBLIOGRAPHIE

[1] A. Djematen, B. Taconet, A. Zahour: Une méthode statistique pour la reconnaissance de caractères berbères manuscrits*; CIFED'98*, p 170-178.

[2] A. Djematen, B. Taconet, A. Zahour: A Geometrical Method for Printing and Handwritten Berber Character Recognition**.** icdar, p. 564, *Fourth International Conference Document Analysis and Recognition* (ICDAR'97), 1997.

[3] A. Oulamara, J Duvernoy: An application of the Hough transform to automatic recognition of Berber characters. *Signal Processing*, vol. 14, 1988, 79-90.

[4] A. Rachidi, D. Mammass: Informatisation de La Langue Amazighe: Méthodes et Mises En Œuvre, SETIT 2005 3rd International Conference*,* March 27-31, 2005 – TUNISIA.

[5] Institut Royal de la Culture Amazighe, Centre de l'Aménagement Linguistique , Graphie de la langue Amazighe, Coordinateur El Mehdi Iazzi, Publications de l'IRCAM, Rabat, 2004.

[6] Proposition d'ajout de l'écriture Tifinaghe au répertoire de l'ISO/CEI 10646 (format Unicode), 21/06/2004, centre CEISIC, IRCAM, Rabat, Maroc.

[7] S. Njah, A. Triki, A. M.Alimi: Système de reconnaissance de code postal. *18éme Conférence Tunisienne d'Electrotechnique et d'Automatiqu*e, Novembre 1998, Hammamet, Tunisie.

[8] N. Chomsky: Three models for the description of language. *IRE Trans. On Information Theory*, vol. 2, n° 3, 1956.

[9] C. de la Higuera: Characteristic sets for polynomial grammatical inference. *Machine Learning*, 27(2):125–138, 1997.

[10] T.Y. ZHANG et C.Y .SUEN: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3): 236–240, mars 1984.

[11] A. Elbaati, M. Kherallah, A. M. Alimi, A. Ennaji: De l'Hors-Ligne Vers un Système de Reconnaissance En-Ligne: Application à la Modélisation de l'Écriture Arabe Manuscrite Ancienne; ANAGRAM'06, septembre 2006.