استعمال تكنولوجيا المعلومات والاتصال الحديثة للنهوض باللغة العربية

- نضال لمغاري* سعيد رغاي** -
- مختبر الرياضيات والإعلاميات، جامعة القاضي عياض
مراكش، المغرب
* nidaliso@yahoo.fr ،** s.raghay@uca.m

مقدمة:

تعتبر اللغة أداة تواصل و تعبير عن ثقافة وهوية الشعب. وجودها في الحقل الاجتماعي و الثقافي العالميين يعتمد على جهود المجتمعات الناطقة بها. فأي لغة لا تتوفر على الوسائل الضرورية للتواصل على نطاق واسع و لا تواكب التطور السريع للتقدم العلمي والتكنولوجي مهددة بالانقراض. وتجدر الإشارة، في هذا الصدد، إلى وجود قسمين أساسيين من اللغات: لغات قوية معلوماتيا وأخرى فقيرة. ولا تستثني هذه النظرية اللغة العربية حيث يعتمد ارتقاؤها على الوسائل السياسية والمادية والتقنية الموفرة.

اللغة العربية هي واحدة من اللغات الأكثر انتشارا في العالم حيث يتكلمها أكثر من 224 مليون شخص ويكتبها أكثر من 100 مليون شخص في أكثر من 20 دولة مختلفة. لكن هذه اللغة تعتبر من بين اللغات الأكثر صعوبة في التعلم و القراءة وخاصة بالنسبة للأوروبيين الذين يستخدمون أبجدية مختلفة تماما و نظاماً لغوياً مختلفاً.

إن اللغة العربية قد حافظت على خصائصها التعبيرية والجمالية على مدى أكثر من ألف عام ولكنها أصبحت بعيدة نسبيا عن التطورات التي حدثت في مجالات المعرفة التي قادت عصر النهضة والأنوار في أروبا وأوصلت بلدانها إلى عصر التصنيع وما بعد التصنيع، ومكَّنْتها من تحقيق رصيد من التراكم في المناهج والنظريات والتطبيقات فرضت به سيطرتها على العالم منذ أكثر من ثلاثة قرون، وقد صلت اليوم إلى عولمة زاحفة من أدواتها التي لا تقهر التقدم الهائل والسريع في أنظمة الحاسوب وبرمجياتها وفي تكنولوجيا المعلومات والاتصال[1].

لقد ساهم العالم العربي ومازال يساهم في الازدهار العلمي و الثقافي للبشرية جمعاء ويتجسد هذا في تواجد رصيد هام من المخطوطات التي تمثل عظمة الحضارة العربية وتفرض علينا الحفاظ عليها والعمل على نشرها حتى ننشر الثقافة والعلم العربيين. بفضل تكنولوجيا المعلومات و الاتصال الحديثة يصبح من الممكن الحفاظ على هذا التراث المكتوب من التلف والضياع و جعله قابلا للقراءة. ويتجسد استعمال هذه الوسائل في رقمنة المخطوطات ثم إتاحتها إلى أكبر عدد من المستفيدين [2] .

في هذا البحث نهدف لتصميم وتطوير نظام للتعرف على الكتابة اليدوية للخط العربي. و يشكل هذا المجال في الواقع مصدر قلق لا يمكن إنكاره في المهام التي قد تبدو صعبة في بعض المجالات وهي المعالجة التلقائية للسجلات الإدارية العربية، رقمنة التراث الثقافي العربي الخ ...

ويعتقد أن أنظمة التعرف على الكتابة اليدوية العربية ، حتى اذا حصلت على أداء مشجع فقد أثبتت تأثيرها فقط على البيانات الأكاديمية ضمن إطار مقيدة من التطبيقات. باختصار، لا يوجد لحد الآن، نظام شمولي للتعرف على الكتابة اليدوية العربية.

نحاول في الفقرة الموالية من هذا المقال، أن نعرف بنظام التعرف على الكتابة العربية. في الفقرة الثانية، نطرح إشكالية الموضوع. بعد ذلك، نسرد بعض خصائص اللغة العربية التي تؤثر في تقدم مجال التعرف على الكتابة العربية اليدوية. و نقوم في الفقرة الثالثة من هذا المقال باستعراض للبحوث السابقة في مجال التعرف على الكتابة اليدوية والمطبوعة. نشرح، في الفقرة الرابعة المنهجية التي سنتبعها في إنشاء نظامنا الخاص بالتعرف على الكتابة اليدوية العربية. وأخيرا، نقترح طريقة بسيطة لتقسيم كلمة مكتوبة يدويا إلى حروف.

1. تعريف نظام التعرف على الكتابة اليدوية:

يحظى ميدان التعرف الآلي على الكتابة اليدوية للغة في أيامنا بالاهتمام الكبير والمتزايد لدى الباحثين عبر العالم حيث يعتبر من أهم التطبيقات الحاسوبية حاليا خاصة وأن استعمال الحاسوب اكتسح جميع ميادين الحياة. وحتى نواجه كل التحديات التي تطرحها العولمة، ونرقى باللغة العربية إلى منزلة اللغات التي تتبادل بها المحتويات النصية الإلكترونية على شبكة الإنترنيت فلا بد من إنشاء برامج وتطوير تقنيات التعرف الآلي على النصوص الورقية العربية اليدوية.

نظـــام التعرف على الكتابــة اليدويــة أو المتعرِّف الضوئـــي لِلأحرف برنامجٌ قاعديٌّ ضروريٌّ تمتلكه كل لغة. ويسمح هذا النظام بتحويل نص مخطوط يدويا إلى خط آلي لإعادة إنتاجه بسهولة من قبل برامج معالجة النصوص. و تختلف التطبيقـات في هذا المجـال حسـب نوعية حـامل الكتابة. حيث نميز بين أنظـمة التعـرف المباشـر و الغير المباشر. ففي أنظمة التعرف المباشر يتم التعرف على الكتابة اليدوية مباشرة أثناء الكتابة باستخدام قلم ضوئي على شاشة خاصة (حواسيب ذكية، لوحات معلوماتية). في هذه الحالة يسهل تحديد كيفية كتابة الحرف واستخراج صفاته؛ وتكون هذه العملية متزامنة مع الكتابة. أما بالنسبة لأنظمة التعرف الغير المباشر فهي تستعمل صورا لنصوص ورقية. وتسمح هذه الأنظمة بتحويل النص المصوّر

بكاميرا أو ماسح ضوئي (سكانير) إلى نصٍّ رقميّ يمكن فتحه بناشر الكتروني (مثل «ورد»)، وأرشفته كملف على الكمبيوتر.

نهتم في هذا المقال بهذا النوع الأخير من أنظمة التعرف على الكتابة اليدوية العربية (الغير المباشر)؛ فلا يوجد حتّى اليوم قارئٌ ضوئيٌّ آليٌّ عربيٌّ يستحق أن يحمل هذا الاسم! (بيعت في الأسواق العربية برامج غير جيدة لهذا الغرض، رمى بها بعض من اشتراها في سلّة المهملات، رغم سعرها الباهظ!)[3] .

يُشكِّلُ عدم توفر اللغة العربية على متعرّف آلي للنصوص الورقية اليدوية حاجزا كبيراً بينها و بين عصر الرقمنة. فبرنامج التعرف الآلي على الكتابة اليدوية يسمح بتحويل صور صفحات الكتاب إلى نصوصٍ رقميّة يمكن قرائتها وتحيينها باستعمال برامج المعالجة النصية دون إعادة طباعة كل ما كُتِب بالعربية على الكمبيوتر.

2. الإشكالية:

على عكس اللاتينية، لا يزال التعرف على الكتابة اليدوية العربية في مستوى البحث والتجريب. ويعزى هذا التباين أساسا إلى ثلاثة أنواع من المشاكل: الأول يرجع إلى تباين واسع في أساليب الكتابة و التشابه بين كيانات الحروف والثاني مرتبط بخصائص اللغة العربية و الثالث يعزى إلى غياب قواعد بيانات مرجعية للتحقق من النتائج و إجراء تقييم موضوعي لأداء أنظمة التعرف على الكتابة.

2.1.خصائص اللغة العربية:

يكتب النص العربي من اليمين إلى اليسار، و تعتبر الكتابة العربية كتابة شبه متصلة. كما أن شكل الحرف العربي يتغير طبقاً لموضعه في الكلمة، حيث يمكن أن يكون للحرف العربي أربعة أشكال مختلفة. يبين [الجدول 1] الحرف العربي في الأشكال الأربع المختلفة.

جدول 1: أشكال الحروف العربية

الحرف	الشكل العام للحرف	الشكل في بداية المقطع	الشكل في منتصف المقطع	الشكل في نهاية المقطع	أشكال أخرى	الحرف	الشكل العام للحروف	الشكل في بداية المقطع	الشكل في منتصف المقطع	الشكل في نهاية المقطع	أشكال أخرى
ألف	ا			ـا		صاد	ص	صـ	ـصـ	ـص	
باء	ب	بـ	ـبـ	ـب		طاء	ط	طـ	ـطـ	ـط	
تاء	ت	تـ	ـتـ	ـت	ة ـة	ظاء	ظ	ظـ	ـظـ	ـظ	
ثاء	ث	ثـ	ـثـ	ـث		عين	ع	عـ	ـعـ	ـع	
جيم	ج	جـ	ـجـ	ـج		غين	غ	غـ	ـغـ	ـغ	
حاء	ح	حـ	ـحـ	ـح		فاء	ف	فـ	ـفـ	ـف	
خاء	خ	خـ	ـخـ	ـخ		قاف	ق	قـ	ـقـ	ـق	
دال	د			ـد		كاف	ك	كـ	ـكـ	ـك	
ذال	ذ			ـذ		لام	ل	لـ	ـلـ	ـل	
راء	ر			ـر		ميم	م	مـ	ـمـ	ـم	
زاي	ز			ـز		نون	ن	نـ	ـنـ	ـن	
سين	س	سـ	ـسـ	ـس		هاء	ه	هـ	ـهـ	ـه	
شين	ش	شـ	ـشـ	ـش		واو	و			ـو	
صاد	ص	صـ	ـصـ	ـص		ياء	ي	يـ	ـيـ	ـي	

هذا ويحتوي أكثر من نصف الحروف العربية على نقاط (1، 2 أو 3) و يؤدي تغيير عدد النقاط أو مكانها إلى تغيير الحرف [الجدول2]. كما تتميز الكتابة العربيـة اليدويـة بوجـود ارتباطـات عمودية بين الحروف (تراكب الحروف) [الصورة1] .

جدول2: مختلف أشكال كتابة النقط يدويا

صورة 1: الارتباطات العمودية بين الحروف

و كما ذكرنا في السابق، فإن اللغة العربية تمتاز بحروفها شبه المتصلة. و اتصال الحروف مع بعضها البعض يختلف طبقاً لنوع الحرف نفسه وقابلية اتصاله مع الجهة اليسرى فقط أو اليمنى أو هما معا. وبنـاءاً عليه ربما تـتكـون الكلمـــة العربيـة من شبـــه كلمـــة واحـدة أو أكثر.و يمكن تعريف شبه الكلمة بأنها المقطع المفصول بما يسبقه وما يليه بفراغات داخل نفس الكلمة. في ما يلي نعتبر أن وحدة التعرف على الكتابة العربية هي شبه الكلمة (المقطع) عوضاً عن كامل الكلمة.

وفي هذا الصدد، فقد أثبتت مختلف الأبحاث في مجال التعرف على الكتابة العربية أن اتصال حروف الكتابة العربية، تعقيد شكل الحروف العربية و تراكب الحروف عموديا يشكل أهم مشاكل المعالجة الآلية لهذه اللغة. حيث تعيق هذه المشاكل مختلف المستويات لاسيما: اختيار الصفات المناسبة التي تصف شكل الحروف وطريقة تقسيم شبه الكلمة إلى حروف.

2.2. غياب قاعدة بيانات التجارب:

يعتبر توفر قاعدة بيانات مرجعية ضروريا لإجراء تقييم موضوعي لنتائج أنظمة التعرف على الكتابة. فمنذ بداية التسعينات، استفاد مجال التعرف على الكتابة اللاتينية من توفر العديد من قاعدات البيانات كقاعدة CEDAR المبرمجة سنة 1994 [10]. أما بالنسبة للغة العربية، فيختلف الوضع لأن مجال التعرف على الكتابة العربية يعاني منذ سنوات عديدة من عدم وجود

قاعدة بيانات مرجعية ستمح مقارنات موضوعية بين الأنظمة. مما يفسر تأخيرالأبحاث في مجال التعرف على كتابة هذه اللغة. توجد، في الوقت الراهن، قاعدة بيانات وحيدة مجانية وهي قاعدة البيانات التونسية IFN/ENIT التي نشرت سنة 2002 [14]. لقد تمت برمجة هذه القاعدة من طرف معهد تكنولوجيات الاتصال بتعاون مع المدرسة الوطنية للمهندسين بتونس. وتتوفر هذه القاعدة على صور أسماء مدن وقرى تونسية. باستثناء قاعدة البيانات هاته، لم يعد متاحا الولوج إلى أغلب قواعد البيانات المنشأة في مجال التعرف على الكتابة العربية، بما أنه تمت برمجتها لخدمة موضوع بحث معين [10]. و سوف نفصل هذا الموضوع في الفقرة أسفله.

3. عرض الأعمال المنجزة سابقا في مجال التعرف على الكتابة العربية:

3.1.قواعد البيانات:

في [10] يناقش هيكل العابد و فولكر مارنيير أهمية وتنوع قواعد البيانات بالنسبة لأنظمة التعرف على الكتابة العربية، كما يعرضون وصفا موجزا لبعض قواعد بيانات مخطوطات عربية. يلخص الجـدول أسفله أهـم قواعد البيانـات التي استعملت في التعـرف على الكتابـة العربية.

قاعدة البيانات	أصلها	سنة برمجتها	خصائصها
ERIM	معهد بحوث البيئة بميشيغان	1995	تحتوي على أكثر من 750 صفحة نص مطبوع. لم تعد هذه القاعدة متوفرة.
Al-Isra	جامعة عمان	1999	تضم صور مختلفة لـ 37000 كلمة ، 10000 رقم عربي وهندي، 2500 إمضاء و 500 فقرة نص عربي.
CENPARMI	Al-Ohali	2000	تحتوي على 7000 صورة لشيكات سعودية: 1547 مبلغ بالحروف و 1547 مبلغ بالأرقام كما تضم 23325 شبه كلمة و 9865 رقم معزول.
Farsi-City	Dehghani	2001	تحتوي على 17000 صورة لـ 198 اسم مدينة إيرانية.
AHDB	Al-Ma'adeed	2002	تضم معجما من 47 كلمة، حيث تم جمع 100 مبلغ مكتوب بالحروف من طرف 100 شخص بالإضافة إلى بعض الأسطر من

			نصوص حرة.
IFN/ENIT	معهد تكنولوجيات الاتصال بتعاون مع المدرسة الوطنية للمهندسين بتونس	2002	تتكون من 32492 صورة لأسماء مدن وقرى تونسية. كتبت هذه الأسماء من طرف 1000 شخص من مختلف الأعمـار و المهن
Amin	عدنان أمين	2003	تحتوي على 4800 حرف عربي معزول
CEDARAB	Srihari	2006	تحتوي على 20000 كلمة حيث كتب 10 أشخاص 10 صفحات تضم كل واحدة مابين 150 و 200 كلمة.
Arabic-Handwriting 1.0	Société Applied Media Analysis.	2007	تتكون هذه القاعدة التجارية من 5000 صورة: 200 مخطوطة عربية (يدوية)، نقط، رسومات بيانية، قصائد شعرية، استمارات و أرقام عربية وهندية.

جدول3: مختلف قواعد البيانات للنصوص العربية

يلاحظ إذن، أن معظم قواعد البيانات المجانية التي تمت برمجتها في هذا المجال تخدم مواضيع معينة (قراءة الشيكات،مدن أو قرى دولة، أرقام..). و هذا ما يبرر عدم توفر مجال التعرف على الكتابة العربية على قاعدة بيانات موحدة عالمية تمكن من التعرف على جميع المصطلحات العربية.

3.2. أنظمة التعرف على الكتابة العربية

لقد تم تطوير أول أنظمة التعرف على الكتابة اليدوية العربية في أوائل عام 1980 . وشهدت 20 سنة من بعد تطوير عدة أنظمة للتعرف عن الكتابة العربية (المطبوعة والمكتوبة يدويا) [11]. استعملت ثلاث تقنيات في هذا المجال: منطـق الغمـوض، الشبكـات العصبيـة و نمـاذج ماركـوف المخفيـة. ولم يكن من الممكـن مقارنـة هذه الأنظمة لأنها كانـت تعتـمد على قواعد بيانات خاصـة مختلفة و محدودة المصطلحات.

في 1994، اقترح أبوهيبة مجموعة من الأشكال الهندسية الحرفية للتعرف على الحروف العربية المعزولة [4]. كما اعتمد أمين، أيضا، رسوما نحيفة في التعرف على الحروف العربية (1996) [6]. و اعتمد في نظامه على شبكة عصبية مزودة ب 5000 حـرف. وسجل معدل تعرف

يقارب 92%. و في محاولة لتطوير عمله [4]، قام أبوهيبة سنة 1998 باقتراح نظام للتعرف على النصوص المكتوبة يدويا.

أما نجوى بنعمارة، فاهتمت بتطبيق الأساليب الماركوفية المستعملة في التعرف على الكتابة اللاتينية [7] . حسب نفس الباحثة، فإن هذه الأساليب تعطي نتائج مشجعة في حالة الكتابة العربية اليدوية كما هو الحال بالنسبة لنظيرتها اللاتينية.

في [11] يقدم ليانا لوريكو و فونو كوفاندراجو دراسة مهمة في مجال التعرف على اللغة العربية مستعرضين بذلك أغلب المشاكل المرتبطة بهذا المجال. و يوضح الباحثون أن نتائج أنظمة التعرف على اللغة العربية تكون جيدة فقط في حالة تطبيقات خاصة (عدد صغير من المصطلحات، أشكال محددة من المصطلحات..)، كما يؤكدون أن نجاعة و نجاح الأنظمة في هذا المجال يتمثل في قدرتها على التعرف على نصوص حرة (لا تقتصر على مجال معين). غير أن هذه التطبيقات تتطلب نماذج لغوية متطورة.

و في [9] يطرح محمد شرييط، من جهته، عدة تساؤلات مفتوحة مصرا على أهمية دمج محلل مورفولوجي:

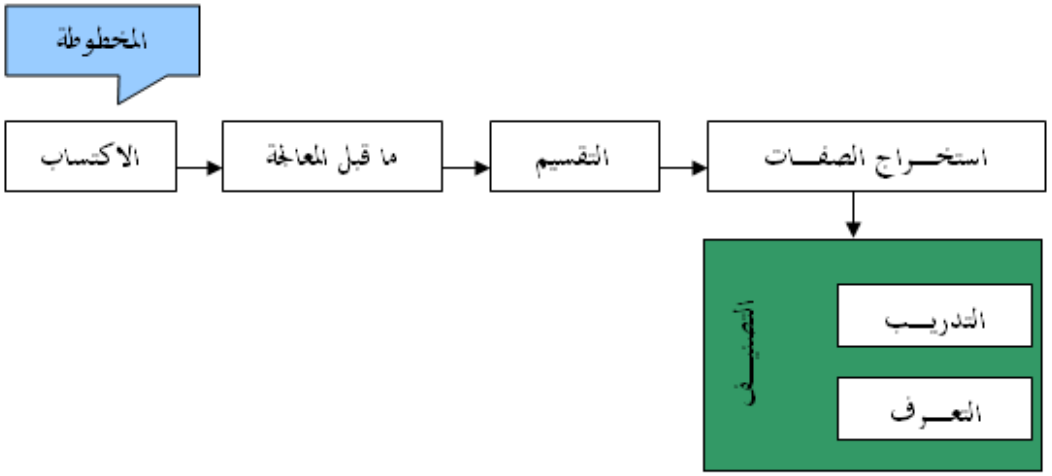
- هل من الضروري تقسيم الكلمة؟
- أية تقنية للتعرف يجب أن نستعمل؟
- هل يجب القيام بمرحلة بعد المعالجة أو محاولة دمج القواعد المورفولوجية ضمن سلسلة عملية التعرف؟
- هل يجب خلق بنيات جديدة أو الاعتماد على الأساليب المستعملة؟

ويتساءل محمد شرييط أيضا عن كيفية تشكيل معجم لتدليل نظام التعرف على كتابة جميع المصطلحات. وينهي الباحث دراسته بأهمية استخدام القرآن كدليل لأنظمة التعرف على الكتابة.

ويعتبر باحثون آخرون [13] أن غياب نظام تجاري للتعرف على الكتابة العربية اليدوية يعزى إلى إشكالية تقسيم الكلمة إلى حروف. كما يعرضون عددا من الأعمال التي تهدف إلى تحسين و تطوير مجال التعرف على الحروف العربية اليدوية الناتجة عن عملية تقطيع يديوية للكتابة العربية اليدوية. حيث تعتبر هذه الأعمال أنه من المستحيل أن تنجح عملية تقطيع كلمة، مكتوبة يدويا، أوتوماتيكيا إلى حروف صحيحة تجعلها قابلة للتعرف عليها من طرف نظام للتعرف على الحروف المعزولة.

4. منهجية العمل:

يبدو مهما أن نحاول اتباع نفس المنهجية الكلاسيكية للنظام التقليدي للتعرف على الكتابة اليدوية. و تتجلى هذه المنهجية في المراحل التالية: مرحلة الاكتساب أي تحويل المخطوطة الورقية إلى صورة رقمية ، مرحلة ما قبل المعالجة، مرحلة التقسيم، مرحلة استخراج الصفات ثم مرحلة التصنيف التي تتجلى في مرحلتين : التدريب و التعرف على الحرف [الصورة 2] .



الصورة 2: منهجية نظام التعرف على الكتابة اليدوية

كشفت لنا قراءة الدراسات والأبحاث السابقة عن أصعب مشكل في عملية التعرف على الكتابة العربية اليدوية. يتعلق الأمر بمرحلة تقسيم الكلمة أو شبه الكلمة إلى حروف. فمن الصعب جدا تقسيم الكلمة المرتبطة الحروف. ومن ثم، فإن نجاح نظام التعرف على الكتابة العربية اليدوية يرتبط بنجاح عملية التقسيم والتجزئة.

ولمواجهة مشاكل التقسيم الناتجة عن تعقيد الكتابة العربية، اقترح الباحثون أساليب محددة: الأساليب شبه العامة، الأساليب التحليلية و الأساليب اللسانية العصبية. نظرا لما أثبتته معظم الأساليب المستعملة في تقسيم الكتابة العربية اليدوية، نقترح فيما يلي طريقة بسيطة لتجزئة الكلمة العربية المكتوبة يدويا.

5. طريقة التقسيم المقترحة

تمر عملية تقسيم وثيقة عبر عدة مراحل. ففي البداية يجب تقسيم الصفحات إلى سطور ثم تقسيم هذه السطور إلى كلمات وأخيرا تقسيم الكلمات إلى حروف [10]. وتعتبر عملية التقسيم إلى حروف أصعب من عملية التقسيم إلى أسطر أو كلمات.

في بحثنا هذا نهتم بتقسيم الكلمة إلى حروف. وفي هذا الصدد، اخترنا أن نبني طريقتنا على الأساليب الشبه العامة، حيث تهتم هذه الأساليب بمعالجة شبه الكلمة أو المقطع. نحاول أولا

تقسيم الكلمة إلى مقاطع (شبه كلمات) ثم نقسم كل شبه كلمة أو مقطع إلى حروف. نستغل، في طريقتنا هذه، الاختلاف الموجود بين ارتباط الحروف في اللغة اللاتينية و العربية. حيث تتصل حروف شبه الكلمة العربية فيما بينها بواسطة خطوط "أفقية". تتطلب هذه الطريقة معالجة دقيقة لصورة الكلمة المراد التعرف عليها.

5.1.معالجة صورة الكلمة:

لا نتطرق في هذه الفقرة إلى المشاكل العادية التي تخص معالجة الصور. نفترض أننا نشتغل على صورة نظيفة و أن الكلمة منحّفة. نهتم في هذا المقال بتقنيات تصحيح منحدر الأسطر، تقنيات تصحيح منحدر الأحرف، وتقنيات العثور على خط الأساس. و تتطلب طريقة التجزئة التي نعتمدها استرداد المحاذاة الأفقية لاتصالات حروف شبه الكلمة . و تتميز هذه الاتصالات بانتمائها أو قربها من خط الأساس [الصورة3].

الصورة 3: استرداد المحاذاة الأفقية لاتصالات الحروف

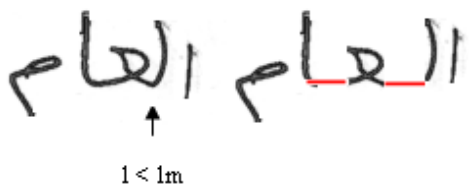
ولتجنب تقسيم الحرف الأخير من شبه الكلمة أو بعض الحروف ك: "س" و "ش" لا نقوم بمعالجة:

- الاتصالات الموجودة بين أسنان الحرفين "سـ" و "شـ"
- الحرف الأخير من شبه الكلمة

ونظرا لوجود اتصالات قصيرة بين الحروف (l)، في بعض الحالات، فإننا نعمل على إطالة هذه الاتصالات حتى يصبح طولها يساوي الحد الأدنى المطلوب (lm) [الصورة4].

ليكن طول الا تصال بين الحروف = l و الحد الأدنى المطلوب بين حرفين = lm

إذا كانت l < lm إذن l = lm



الصورة4: إطالة اتصال قصير

خوارزمية استرداد المحاذاة الأفقية بين الحروف:

تكرار

لكل اتصال قريب أو فوق سطر الأساس

إذا كان الاتصال قصيرا جدا < lm

إذا وجدنا بعد هذا الاتصال سنين آخرين

إذن يبتعلق الأمر ب "س" أو "ش"

لا نقسم بين الأسنان

نهاية إذا

وإلا يعوض الاتصال بخط أفقي طوله = lm

نهاية إذا

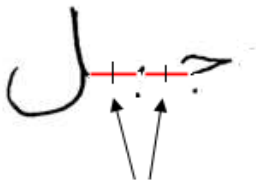
يعوض الاتصال بخط أفقي

نهاية لكل

إلى الوصول إلى الحرف ما قبل الأخير

5.2.طريقتنا في تقسيم شبه الكلمة إلى حروف:

بعد تسوية ومعالجة الاتصالات بين الحروف يمكن أن نقسم شبه الكلمة إلى حروف.
نعتبر أن منتصف الاتصالات الخطية هي نقط التجزئة.



نقط التجزئة

خوارزمية تقسيم شبه الكلمة

تكرار

لكل اتصال خطي أفقي

قياس طوله l

قياس منتصفه l/2

تقسيم الخط في منتصفه

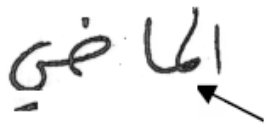
نهاية لكل

إلى الوصول إلى نهاية شبه الكلمة

في نهاية مرحلة التقسيم، نقوم بإخضاع الحروف الناتجة عن هذه العملية إلى محرك للتعرف على الحروف العربية اليدوية و ذلك من أجل التحقق من صحة نقاط التجزئة المحصل عليها.

6. حالات خاصة لا تخضع للتقسيم :

كما ذكرنا أعلاه، فاللغة العربية تمتاز بتراكب بعض الحروف عموديا (حرفان إلى أربعة أحرف). تنتج عن هذه الظاهرة أشكال هندسية مختلفة تماما عن الحروف التي تكونها [8].مما يجعل عملية التقسيم إلى حروف مستحيلة باستعمال طريقتنا. و توضح الصورة أسفله تراكب حرفين "ل" و "م".

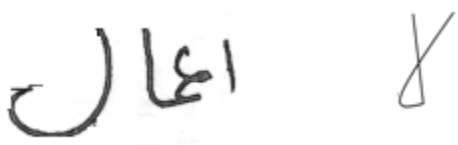


لحل هذا المشكل، يُعرّف ف.مناصري [12] أبجدية حروف عربية جديدة [الجدول4]. باستعماله هذه الحروف يتجنب الباحث تقسيم الحروف المتراكبة ويتعرف عليها النظام كوحدة كاملة.

ا	ر	ه	ه	ح	ح	د	ر	س	س
ص	ص	ط	ء	ع	ع	ع	و	ك	ل
ل	م	م	ن	ه	ه	و	ى	Tail 1 : ب	
لا	لـ	لـ	لـ	لـ					

جدول4: قائمة الرموز الأبجدية ال

من خلال دراستنا لهذه الرموز، تبين لنا أن بعض الأحرف المتراكبة غير ممثلة في أبجدية ف.مناصري [12]. وتوضح الصورة أسفله مثالا لذلك:



ولتجنب تقسيم خاطئ للحروف المتراكبة عموديا، نعتمد نهج ف.مناصري ونحاول إضافة الرموز الغير الممثلة في أبجدية الباحث ف.مناصري.

الخاتمة

مما لا شك فيه، فإن التعرف على الكتابة اليدوية العربية هو موضوع مهم جداً. فهذا المجال يمكن أن يخدم العرب و غير العرب من المهتمين باللغة العربية. يحتاج البحث في هذا المجال إلى وقت مهم للحصول على النتائج المرجوة. لذلك وجب تكتل و تظافر جهود الباحثين في الجامعات العربية المختلفة ومراكز البحوث، كما وجب أيضا إسهام المنظمات المهتمة في دعم التقدم الحضاري والعلمي في العالم العربي.

حاولنا في هذا المقال أن نقوم بدراسة تمهيدية في مجال التعرف على الكتابة اليدوية العربية. كما سردنا خاصيات مهمة في أساليب الكتابة العربية اليدوية التي تختلف عن اللغة اللاتينية. سمحت هذه الأساليب بالكشف لنا عن أهمية التعامل مع شبه الكلمة عوض الكلمة. بعد ذلك حاولنا إيجاد حل لأهم مشكلة في مجال التعرف على الكتابة اليدوية العربية. ألا وهي عملية تقسيم الكلمة أو بالأحرى شبه الكلمة إلى حروف. قمنا بتقديم طريقة سهلة لتقسيم شبه الكلمة تختلف عن باقي الطرق المستعملة في هذا المجال. ولا تزال العديد من المهام، مثل استخراج خصائص الحروف المحصل عليها أثناء عملية التقسيم ، التعرف عليها وتعليمه للنظام.

ونأمل أن نساهم بدراستنا في تقدم البحث في هذا المجال و أن تستفيد لغتنا من نظام للتعرف على كتابتها اليدوية يرقى بها إلى المستوى التي تستحقه.

المراجع

[1] منشورات المجلس الأعلى للغة العربية بالجزائر 2002، اللغة العربية في تكنولوجيات المعلوميات تطور واعد...وتطوير متواصل.
http://www.csla.dz/mjls/index.php?option=com_remository&Itemid=55&func=fileinfo&id=31

[2] حافظي زهير، دور تكنولوجيا المعلومات في حفظ المخطوطات العربية.
http://www.journal.cybrarians.org/index.php?option=com_content&view=article&id=421:2009-08-02-08-08-50&catid=137:2009-05-20-09-51-17&Itemid=56

[3] د حبيب عبد الرب سروري، اللغة العربية في الزمن الرقميّ: ستُّ فجائع، وثلاثة مقترحات!
http://my-last-articles-and-texts.blogspot.com/2009/03/blog-post_23.html

[4] I. S. I. Abuhaiba, S. A. Mahmoud, and R. J. Green, "Recognition of handwritten cursive Arabic characters," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 16, pp. 664-672, 1994.
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=295912&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel1%2F34%2F7321%2F00295912.pdf%3Farnumber%3D295912

[5] B. Al-Badr , S.A. Mahmoud : « Survey and bibliography of Arabic optical text recognition ». Signal processing, vol. 41, pp. 49-77, 1995.
http://www.sciencedirect.com/science/article/pii/016516849400090M

[6] A. Amin, H. Al-Sadoun, and S. Fischer, "Hand-printed Arabic character recognition system using an artificial network," Pattern recognition , vol. 29, pp. 663-675, 1996
http://www.sciencedirect.com/science/article/pii/0031320395001107

[7] N. Ben Amara, A. Belaïd, and N. Ellouze. Utilisation des modèles markoviens en reconnaissance de l'´ecriture arabe état de l'art. Colloque International Franco-phone sur l'Ecrit et le Document (CIFED'00), Lyon, Lyon, France, juillet 2000.
perso.telecom-paristech.fr/~lauli/Articles/cifed_V6.doc

[8] A. BOUKHAROUBA, thèse de doctorat : Contribution à la segmentation et à la reconnaissance de l'écriture arabe manuscrite, 2011.

[9] M. Cheriet. Strategies for visual arabic handwriting recognition: issues and case study. In ISSPA 2007, International Symposium on Signal Processing and its Applications, 12 – 15 February 2007, Sharjah, United Arab Emirates, Feb 2007.
http://books.google.co.ma/books?id=ufXlw_vRI1UC&pg=PA9&dq=Strategies+for+visual+arabic+handwriting+recognition:+issues+and+case+study&hl=fr&sa=X&ei=8Xj-UryAFpGZ0QXG2IH4Ag&ved=0CCsQ6AEwAA#v=onepage&q=case%20study&f=false

[10] H. El Abed and V. Märgner. Base de Données et Compétitions – Outils de Développement et d'Évaluation de Systèmes de Reconnaissance de Mots Manuscrits.
http://hal.archives-ouvertes.fr/docs/00/33/44/03/PDF/paper-22.pdf

[11] L. M. Lorigo and V. Govindaraju. Offline Arabic Handwriting Recognition : A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 28(5), pp. 712 –724, 2006.

[12] F.Menasri, thèse de doctorat :Contributions à la reconnaissance de l'écriture arabe manuscrite, 2008.

[13] Pavithra Babu Sargur Srihari, Harish Srinivasan, Chetan Bhole. Handwritten Arabic word spotting using the cedarabic document analysis system. In 2005 Symposium on Document Image Understanding Technology, The Marriott Inn and Conference Center, University Maryland University College, Adelphi, Maryland, November 2-4, 2005.

[14] M. Pechwitz, S. Snoussi Maddouri, V. Märgner, N. Ellouze, H. Amiri. IFN/ENIT Database of Handwritten Arabic Words. In CIFED'02, 2002.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.1245&rep=rep1&type=pdf