

تحليل إحصائي لدعم التعرف الآلي على الكتابة العربية

م. حسني المحتسب ود. صبري محمود

قسم علوم الحاسب والمعلومات، جامعة الملك فهد للبترول والمعادن، الظهران، المملكة العربية السعودية

د. رامي قهوجي

قسم التصوير الإلكتروني والاتصال، كلية علم المعلومات، جامعة برادفورد، برادفورد، بريطانيا

٦٣

الكلمات المفتاحية: مقاطع كلمات اللغة العربية، تكرار الحروف العربية، التعرف الآلي على الكتابة العربية.

ملخص: تقدم هذه الدراسة ملخصاً لنتائج دراسة إحصائية لأعداد ظهور حروف ومقاطع الكلمات في اللغة العربية. وتشمل النتائج المعروضة تكرار كل حرف من الحروف العربية في كل مقطع من المقاطع، وتكرار الحرف والحرف الذي يليه في المقاطع المختلفة لكل الحروف. كما تشمل الدراسة على إحصائيات استخدام الحروف والمقاطع ونسبة استخدام كل منها في حالات الاستخدام المختلفة في اللغة العربية. وقد تم تطبيق الدراسة على كتابي صحيح البخاري ومسلم. وتفيد الدراسة في المساعدة في عملية التعرف الآلي على الكتابة العربية، كما تفيد في عملية تصحيح الأخطاء بعد عملية التعرف.

مقدمة

اتسمت السنوات الخمس الأخيرة بزيادة اهتمام الباحثين بالحوسبة العربية وتطبيقاتها المختلفة. ويلاحظ هذا الاهتمام في العديد من تطبيقات الحوسبة العربية، ومنها التعرف الآلي (أو ما يعرف بالتعرف الضوئي) على الكتابة العربية، سواء كانت كتابة مطبوعة أو بخط اليد. يقدم هذا البحث تحليلاً لمقاطع اللغة العربية ولورود الحروف العربية في مواقعها المختلفة من المقاطع في محاولة لدعم التعرف الآلي على الكتابة العربية. وحسب علمنا فإن هنالك محاولات قليلة تمت في هذا الاتجاه ولكنها مختلفة من حيث النتائج وبعضها في النتائج والأهداف. فقد قدم البحث إحصائيات لمقاطع اللغة العربية ونسب المقاطع المتكونة من حرف إلى ثمانية حروف وكذلك نسبة الحروف في بداية الكلمة ووسطها وآخرها للمقاطع المتكونة من حرفين إلى ثمانية حروف. كما عرض البحث معدل عدد الحروف في مقطع الكلمة ٢، ٤، وكذلك معدل عدد المقاطع في الكلمة العربية ٢، ٢. وقد عرض البحث استخدام هذه المعلومات في التعرف الآلي على الكتابة العربية. وذكر الباحثون أن الكتابة العربية التي جرى تحليلها تم جمعها بشكل عشوائي من كتب قديمة وكتب حديثة وجرائد ومعلومات تم جمعها من شبكة الإنترنت. وتتكون العينات في بحثهم من ٢٦٢٦٤٧ كلمة تتكون من ١١٢٦٤٢٠ حرفاً. وفي دراستنا تتضمن الإحصاءات تحليلات إضافية سنستعرضها لاحقاً كما أن حجم العينة المستخدمة أكبر. أما البحث الثاني ذا العلاقة فهو يهدف إلى إنتاج معجم لعملية التعرف على الكتابة العربية وذلك في استخدامه للتعرف على مقاطع اللغة العربية وذلك تجنباً لتطبيع الحروف العربية (الذي يؤدي إلى زيادة الأخطاء في التعرف الآلي). اعتمد البحث على تحليل ٢٦٨٠٥ ملفاً بصيغة أتش تي إم ال (HTML) لمجلة الدستور الأردنية. وتحتوي تلك العينات على الأخبار المحلية وفيها أخبار زيارات دبلوماسيين أجانب والتي تتضمن كلمات أعجمية منها أسماء الدول والمدن والأسماء الأجنبية. وقد احتوت العينات على ٢٨٥١٩٧٢٤ مقطعاً. وقد بينت الدراسة إحصاءات للمقاطع بطول يتراوح من حرف إلى ثلاثة عشر حرفاً. ودراسة النتائج فإن أعداد المقاطع بطول ١٢ و ١٣ قليلة جداً ونسبة لا تكاد تذكر والراجح أنها تتكون من مقاطع تحتوي على أخطاء. كما بينت رسالة الماجستير محاولة أخرى لتحليل الكتابة العربية بغرض عمل معجم. وقد كانت العينات المستخدمة مستخلصة من شبكة الإنترنت بشكل رئيس. وتحتوي العينات على ٨٤٢٦٨٤ كلمة من ٤١٥ عينة مختلفة. وتحتوي العينات على مواضيع مختلفة كالإقتصاد وقصص الأطفال والتعليم والصحة والطب ومقابلات وسياسة... الخ. وقد بينت الدراسة إحصاءات لمقاطع تتراوح بين حرف واحد وسبعة عشر حرفاً. وتحتوي العينات على أخطاء مطبعية ووجود مقاطع غير حقيقية في اللغة نتيجة اندماج أكثر من مقطع كما بين الباحث بأمثلة من هذه المقاطع مما يجعل هذه العينات غير دقيقة المحتوى.

وقد استخدم بعض الباحثين احتمالات تتابع الحروف العربية في التعرف الآلي على الكتابة العربية باستخدام نماذج ماركوف المخفية (Hidden Markov Models) وذلك بتحليل الكتابة العربية للعينات المستخدمة في التدريب واستخدام تلك النتائج في مرحلة الفحص. وحيث أن العينات المستخدمة في التدريب محدودة نوعاً ما فإن احتمالات تتابع الحروف قد لا تكون دقيقة. وقد حسب الباحثون احتمالات وجود حرف يتبعه آخر واحتمالات وجود حرف قبله حرف وبعده حرف. إلا أن هذه الاحتمالات تستخدم داخلياً في نظام التعرف الآلي على الكتابة العربية هذا ولم يعرض الباحثين تلك النتائج.



إن نتائج هذا البحث وخاصة فيما يخص احتمالات تتابع الحروف العربية يمكن استخدامه في التعرف الآلي على الكتابة العربية دون الحاجة إلى تحليل عينات التدريب في كل مرة. كما أنه ونظرا لحجم العينات المستخدمة فإن نتائج هذا البحث من المنتظر أن تكون أكثر دقة.

تحتوي فقرة ٢ من البحث على تعريفات للرموز المستخدمة والفقرة ٢ على تحليل نتائج الدراسة والفقرة ٤ على النتائج والاقتراحات.

تعريفات

تقدم تحت هذا العنوان مجموعة تعريفات للمصطلحات المستخدمة في الدراسة. وهذه التعريفات هي: مقطع، ومنفصل، ومتصل، وأول حرف في المقطع، وآخر حرف في المقطع وطول المقطع.

مقطع: نعرف المقطع على أنه مجموعة الحروف المتصلة ببعضها في كلمة واحدة. فمثلا كلمة "مقطع" مكونة من مقطع واحد طولها أربعة حروف، وكلمة "مجموعة" مكونة من مقطعين هما "مجمو" بطول أربعة حروف و "عة" بطول حرفين، وكلمة "الاستحسان" مكونة من أربعة مقاطع هي: "ا" و "لا" و "ستحسا" و "ن"، وأطوالها ١ و ٢ و ٥ و ١ على الترتيب.

منفصل: نقول أن الحرف منفصل إذا كان غير مرتبط بما بعده، وغير مرتبط بما قبله، أي أنه يكون مقطعا لوحده. فمثلا في كلمة "أن" ذات المقطعين يمثل كل مقطع منها حرفا منفصلا، أي أن "أ" حرف منفصل في هذه الكلمة و "ن" حرف منفصل أيضا فيها، بينما "ا" في كلمة "لأنه" ذات المقطعين يمثل حرفا متصلا، وكذلك حرف "ن" في المقطع الثاني مثلا على الحرف المتصل.

متصل: نقول أن الحرف متصل إذا كان مرتبطا بما بعده، أو مرتبطا بما قبله، أو بكليهما. ومثال ذلك في كلمة "مرتبطا" كل حرف من حروفها مثلا على الحروف المتصلة.

أول حرف: هو الحرف الأول في المقطع. ويمكن أن يكون غير متصل بما بعده فيكون مقطعا لوحده بطول ١. ومثال ذلك كلمة "أول" تتكون من ثلاثة مقاطع، كل حرف منها هو أول حرف في المقطع (وهو آخر حرف في المقطع)، بينما تمثل كلمة "متحابين" مقطعين، أول حرف في المقطع الأول هو "م" وأول حرف في المقطع الثاني هو "ب".

آخر حرف: هو الحرف الأخير في المقطع. وهو غير متصل بما بعده. وإذا لم يكن متصلا بما قبله فيكون مقطعا لوحده بطول ١. ومثال ذلك كلمة "متحابين"، فأخر حرف في المقطع الأول هو "ا" وهو متصل، وأخر حرف في المقطع الثاني هو "ن" وهو أيضا متصل.

طول المقطع: عدد الحروف التي تكون المقطع. ومثال ذلك كلمة "فرحتنا" تتكون من مقطعين: "فر" بطول ٢، و "حتنا" بطول ٤. التحليل الإحصائي

تم عمل الدراسة على نص كتابي صحيح البخاري وصحيح مسلم كمثال على النصوص العربية المتعلقة بالحديث الشريف. وبلغ تعداد المقاطع في الكتابين ٢٢١٧١٧٨ مقطعا مع التكرار، و ١٨١٧٠ مقطعا بعد حذف المكرر. بينما بلغ عدد الحروف والرموز في الكتابين ما مجموعه ٤٤٠٥٣١٨ رمزا. ويشمل الكتابان على ١٠٩٥٢٧٤ كلمة مع التكرار و ٥٠٣٦٧ كلمة بعد حذف المكرر.

تبين الجداول من ١ إلى ١٤ مجموعة من النتائج المستخلصة من الدراسة. ويمكن للباحث الراغب في الحصول على الجدول الأساس الكلي والذي يزيد على ٣٠٠ صفحة الاتصال بأحد الباحثين إذ لا يمكن إرفاقه مع هذا البحث. وفي يلي وصفا تحليليا مختصرا لهذه

الجدول

يعرض جدول رقم ١ أول ٢٤٨ مقطعا من المقاطع الأكثر تكرارا من مجموع ١٨١٧٠ مقطعا. ولاحظ أن حرف الألف المنفصل يمثل حوالي ١٠٪ من تعداد المقاطع، يليه حرف الواو المنفصل، فالألف المهموزة، فاللام. ثم المقطع الثلاثي "لله". وهكذا. ويلاحظ المستعرض للجدول أن هناك العديد من المقاطع بطول ٢ و ٣ و ٤ و ٥ وتظهر أكثر شيوعا من مقاطع أخرى بأطوال أقل.

ويعرض جدول رقم ٢ تكرار المقاطع حسب طولها ونسبة وجود هذه المقاطع. ونلاحظ أن أكثر من ٩٠٪ من المقاطع أطوالها ثلاثة حروف فأقل، وأن ما يصل إلى ٩٨٪ من المقاطع أطوالها أربعة حروف فأقل. ونلاحظ أن المقاطع بطول ٩ هي قليلة وتبلغ ٨ مقاطع أغلبها ظهر مرة واحدة، وقد جمعنا هذه المقاطع في جدول رقم ٣. ويبين جدول رقم ٤ كل المقاطع بطول ٨ حروف وعددها ٦٢ مقطعا وهي أيضا قليلة الظهور كما هو واضح في جدول رقم ٢.

يشمل الجدول رقم ٥ تكرار كل حرف من حروف العربية في كل مقطع ظهر به حسب طول المقطع. لاحظ أن الهمزة المنفصلة ” ء “ لم تظهر إلا في المقطع الذي طوله ١ لأنها لا تتصل بما قبلها ولا بما بعدها. بينما حرف الميم على سبيل المثال ” م “ ظهر منفصلاً (في المقطع بطول ١) كما ظهر في كل المقاطع ذات الأطوال المختلفة من ١ إلى ٩. وملاحظة أخرى تبدو لطيفة وهي أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتماً لن يظهر في مقطع أطول من ذلك المقطع.

يشابه الجدول رقم ٦ سابقه غير أنه يستعرض أول المقطع فقط، ولذا نلاحظ أن الحروف التي لا تتصل من اليسار لا تظهر كحرف أول إلا في منفصلة (طول المقطع ١)، وهذه الحروف هي: ء، ا، أ، و، إ، اء، د، ذ، ز، و، ي. وتتكرر هنا ملاحظة أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتماً لن يظهر في مقطع أطول من ذلك المقطع.

أما الجدول رقم ٧ فيتعامل مع الحرف الأخير من المقطع. لاحظ أنه في المقطع المنفصل بطول حرف واحد يكون الحرف هو الأول وهو الأخير. ومن فوائد البيانات في هذا الجدول معرفة نسبة ظهور الحرف متصلاً في آخر الكلمة، ومثال ذلك: بالنظر إلى الهاء نرى أنها تأتي متصلة بما قبلها عشرة أضعاف ظهورها منفصلة. بينما تأتي التاء المربوطة منفصلة أكثر من ضعفي ظهورها متصلة بما قبلها. وترصد الجداول من ٨ إلى ١٥ تكرار الحرف والحرف الذي يليه في بيانات الدراسة، فيظهر الجدول رقم ٨ تكرار ظهور الحرف الأول والحرف الثاني، ويعرض الجدول رقم ٩ تكرار ظهور الحرف الثاني والثالث، ثم جدول رقم ١٠ يعرض تكرار ظهور الحرف الثالث والرابع وهكذا حتى جدول رقم ١٥ الذي يظهر تكرار ظهور الحرف الثامن والتاسع. ويستعرض الجدول رقم ١٦ تكرار ظهور الحرف التاسع. ونلاحظ من هذه الجداول أن أكثر الحروف الثنائية تكراراً تقع في المقاطع ذات الأطوال ٢ و٣.

النتائج

عرضنا في هذه الدراسة مجموعة من نتائج تحليل كتابي البخاري ومسلم بحيث بينت الدراسة نسبة استخدام حروف اللغة العربية ومقاطعها. ويمكن الاستفادة من هذه النتائج في التعرف الآلي على الكتابة العربية حيث تفيد في زيادة دقة التعرف على الحروف كما تساعد في المعالجة التي تتم بعد التعرف في تصحيح الأخطاء. ويمكن استخدام هذه النتائج لحساب احتمالات ظهور حرف بعد حرف آخر، أو الاستفادة من احتمالية ظهور المقاطع كاملة، كما يمكن الاستفادة منها وهي متوفرة لدى الباحثين كنسخة الكترونية يمكن الاستفادة منها لمن يرغب بذلك. كما يمكن استخدامها في التعرف الآلي على قراءة الكتابة العربية باستخدام نماذج ماركوف المخفية (Hidden Markov Models).

وينوي الباحثين تطوير الدراسة باستخدام عينات أخرى لكتب متنوعة ومواضيع مختلفة ومجلات ومعلومات متوفرة على الشبكة الإنترنت. كما ينوي الباحثون استخدام نتائج هذه الدراسة في التعرف الآلي على الكتابة العربية.

شكر

يتقدم الباحثان الأول والثاني بالشكر لجامعة الملك فهد للبترول والمعادن على الدعم المتواصل.

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
ا	9.6243	و	5.0401	أ	4.6884	ل	3.0741	لله	2.6366	ر	2.5324
بن	2.4033	حد	1.9598	ن	1.8796	فا	1.8777	عن	1.794	ثنا	1.432
()	1.3445	()	1.3443	إ	1.3385	عليه	1.054	صلى	0.9907	سلم	0.9786
د	0.8547	ة	0.7936	من	0.7781	لا	0.7274	سو	0.7138	ب	0.7009
بي	0.6725	م	0.6691	[]	0.6575	[]	0.6575	ما	0.6419	با	0.618
ذ	0.6102	عبد	0.5851	في	0.5614	ء	0.5367	نا	0.5082	ه	0.4862
لى	0.4743	فقا	0.4485	ي	0.442	بو	0.4281	لا	0.4192	لنبي	0.4033
ز	0.389	ل	0.3751	خبير	0.3578	2	0.3426	كا	0.3291	علي	0.3285
عمر	0.3153	لر	0.3111	لك	0.2902	ثم	0.2846	ني	0.283	محمد	0.279
ت	0.2767	فأ	0.2725	لو	0.272	له	0.2646	س	0.2632	ير	0.2598
يا	0.2569	مر	0.2566	يو	0.2552	بد	0.2516	ثني	0.2469	5	0.244
6	0.2431	بر	0.2419	3	0.2417	4	0.2416	تعا	0.234	عا	0.233
ق	0.2239	هر	0.215	ضي	0.2082	7	0.2014	}	0.2008	9	0.1981
{	0.1961	لت	0.192	8	0.1907	لنا	0.1907	0	0.1891	يحيى	0.1861
نه	0.1833	جا	0.1828	حتى	0.1806	لم	0.1797	سعيد	0.1746	كر	0.1711
قد	0.1709	فا	0.1697	لد	0.1643	ها	0.1605	آ	0.1602	يقو	0.1556
بكر	0.1528	فإ	0.1525	جل	0.152	هو	0.1499	هذ	0.1472	ح	0.1469
نس	0.1452	لز	0.1446	خر	0.1432	مو	0.1432	ك	0.1387	به	0.1352

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
عنه	0.1348	سمعت	0.134	ج	0.1338	ى	0.1294	لذ	0.1266	فو	0.1235
عر	0.1167	لا	0.1166	فر	0.116	عبا	0.1154	جر	0.113	سا	0.1116
مة	0.1106	لي	0.1096	هيم	0.1094	حر	0.1083	ية	0.1076	فع	0.1068
ين	0.1064	ثشة	0.1056	حمن	0.1056	ف	0.1053	ث	0.1019	هل	0.1013
علي	0.1005	صا	0.0972	هم	0.0964	يث	0.0943	فر	0.0937	يز	0.0935
حا	0.0888	سفيا	0.0884	ع	0.0884	شعبة	0.0876	سحا	0.0872	خا	0.0869
بيه	0.0867	عبيد	0.0866	عند	0.0866	عو	0.0825	سما	0.0812	هب	0.0776
كل	0.0767	نت	0.0763	قلت	0.0753	سي	0.0752	غير	0.0734	تر	0.0732
شبية	0.0726	بين	0.072	بعد	0.0714	ض	0.0705	عنهما	0.0687	فلما	0.0674
شها	0.0665	ته	0.0663	كم	0.0654	فيه	0.0638	مع	0.0634	هما	0.0631
معا	0.0631	هشا	0.0622	فلا	0.0621	خذ	0.0617	سلمة	0.0615	عيل	0.0615
لا	0.0608	نز	0.0608	لا	0.06	بها	0.0593	فو	0.0584	فقلت	0.058
سعد	0.0566	سمع	0.056	عد	0.0556	بهذ	0.0548	لصلا	0.0541	خل	0.0539
نها	0.0537	سنا	0.0533	نا	0.0527	شر	0.0525	قبل	0.0522	نلا	0.052
يت	0.0519	لها	0.0517	عنها	0.0497	جد	0.0497	هير	0.0494	قتيبة	0.0493
نو	0.0493	شى	0.0489	سليما	0.0488	منه	0.048	هد	0.0479	لثني	0.0476
سأ	0.0472	كذ	0.0471	فتا	0.0467	يعنى	0.0467	نو	0.0466	يا	0.046
عة	0.0458	جعفر	0.0457	طا	0.0457	بأ	0.0455	شا	0.045	نصا	0.0449
بة	0.0447	بد	0.0442	كنت	0.044	صلا	0.0437	تا	0.0433	ليه	0.0429
بير	0.0425	فيها	0.0422	نعم	0.0422	لهم	0.0421	عمش	0.0419	لجنة	0.0419
خير	0.0417	فد	0.0416	لد	0.0415	لحد	0.0414	عشا	0.0414	فد	0.0409
لحا	0.0404	ليس	0.0401	جلا	0.04	لقر	0.0397	سر	0.0396	صد	0.0396
شينا	0.039	بت	0.0388	ينة	0.0384	حما	0.0383	ينا	0.0383	نما	0.0382
حين	0.0382	بني	0.0381	صحا	0.0373	فلم	0.0371	عطا	0.0371	لقا	0.0371
فة	0.0367	جو	0.0367	حميد	0.0367	للهم	0.0363	لر	0.0362	ؤ	0.0357
مه	0.0353	لقد	0.0352	معمر	0.0352	لسا	0.035	منها	0.035	كما	0.0347
كنا	0.0342	منا	0.0341	يب	0.0341	ليث	0.034	مثل	0.0335	كو	0.0334
لح	0.0333	يج	0.0332	حمد	0.0332	لؤ	0.0331	تي	0.033	بيد	0.0325
حب	0.0324	بنت	0.0322	سف	0.032	بذ	0.032	مي	0.0318	نك	0.0316
نمير	0.0316	هي	0.0311	نسا	0.0309	لقو	0.0308	سم	0.0306	تي	0.0295
معه	0.0295	للفظ	0.0293	نة	0.0293	بشا	0.0292	فما	0.0292	تم	0.029
بما	0.0285	بنا	0.0285	بشر	0.0285	جميعا	0.0284	ليد	0.0284	صو	0.0282
سلا	0.0281	لمسجد	0.028	جع	0.0279	عشر	0.0275	لعز	0.0273	منصو	0.0272
لكم	0.0272	كلا	0.027	شد	0.0267	لتي	0.0266	بك	0.0258	حو	0.0258
فيقو	0.0257	مسلم	0.0257	نحو	0.0257	صم	0.0256	كثير	0.0256	لحسن	0.0254
نو	0.0252	ضى	0.0252	لقيا	0.0252	يحد	0.0251	بيع	0.0243	سه	0.0243
يقا	0.0242	لسما	0.0239	عليها	0.0238	يصلى	0.0235	عز	0.0234	تا	0.0233
ثة	0.0232	لبيت	0.0232	لعا	0.0231	منهم	0.0231	نهم	0.0231	تقو	0.0231
ند	0.023	علم	0.023	لليل	0.0229	به	0.0229	لسلا	0.0227	فكا	0.0226

جدول ١ :
المقاطع الأكثر
تكرارا: أول ٣٥٠
مقطعا

طول المقطع	عدد	تكرار	نسبة
1	57	942097	42.49079
2	537	638359	28.79149
3	3192	418885	18.89270
4	6778	169938	7.664604
5	4896	39026	1.760164
6	2118	7361	0.331998
7	522	1322	0.059625
8	62	162	0.007306
9	8	28	0.001262
المجموع	18170	2217178	100

جدول ٢ :
أطوال المقاطع
وتكرارها
ونسبتها

جدول ٣ : كل المقاطع بطول ٩ ونسبتها

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
لستضعفين	0.00086	فليقطعهما	0.00009	قسطنطينية	0.00009
ليثينهما	0.00005	فلتقتلهم	0.00005	يستينهما	0.00005
مستقبليها	0.00005	فليلتمسها	0.00005		

٦٧

جدول ٤ : كل المقاطع بطول ٨ ونسبتها

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
	0.00005	فليحملها	0.00005		0.00005		0.00005		0.00005
	0.00005	لستقبله	0.00005	لمجنبتين	0.00005		0.00005		0.00005
	0.00005		0.00005	فليحلبها	0.00005	بحبيبتيه	0.00005		0.00005
	0.00005		0.00005		0.00005		0.00005		0.00005
	0.00005	فليمتهما	0.00005	لبيتلكم	0.00005		0.00005		0.00005
	0.00009	لمقتسمين	0.00009		0.00005	فمنعنيها	0.00005		0.00005
	0.00009	فليجعلها	0.00009		0.00009		0.00009		0.00009
	0.00009	فليمحنها	0.00009	فليتحلله	0.00009		0.00009		0.00009
	0.00009		0.00009	بمغنين	0.00009	لنخلفين	0.00009	فتبعيه	0.00009
	0.00014	لينعلهما	0.00014	لجهنمين	0.00014	لخليفين	0.00014	تسليمين	0.00009
	0.00027	للمحلقين	0.00018	تستعينها	0.00018		0.00018	فليستنثر	0.00018
	0.00036		0.00036	فنفضتهما	0.00032		0.00027		0.00027
							0.00086		0.00045

نسبة	مجموع	تكراره في المقطع الذي بطول									الحرف	
		9	8	7	6	5	4	3	2	1		
0.27013	11900										11900	ء
0.11445	5042					3	6	22	1459	3552		أ
2.87396	126607				2	38	430	2057	20129	103951		أ
0.07041	3102					3	55	995	1257	792		ؤ
0.81715	35998						1	43	6277	29677		إ
0.22566	9941			2	4	71	298	1289	1374	6903		ئ
11.56709	509567	6	75	475	2251	10664	29471	113405	139833	213387		ا
5.30854	233858			8	44	1199	6578	23260	46774	155995		ب
1.24127	54682	2	1	40	625	5350	11377	8824	10867	17596		ة
2.23285	98364			1	203	2205	11861	15133	33473	35488		ت
1.48396	65373			4	37	77	1540	3554	7909	52252		ث
0.94608	41678				18	58	968	3320	10529	26785		ج
2.45047	107951			2	50	288	1729	4373	28380	73129		ح
0.69021	30406				4	31	248	1186	5884	23053		خ
3.02850	133415			7	76	1401	16728	30893	65359	18951		د
0.65771	28974			7	18	102	426	3202	11690	13529		ذ
3.90578	172062		8	20	420	2392	15879	44166	53029	56148		ر
0.48201	21234				11	52	499	2960	9087	8625		ز
2.60038	114555			1	48	273	2325	9913	20162	81833		س
0.70592	31098			2	3	36	466	3060	11727	15804		ش
1.07436	47329				11	101	692	3605	10323	32597		ص
0.42031	18516			2	9	107	697	3008	4336	10357		ض
0.37089	16339				7	155	1108	3082	7070	4917		ط
0.10260	4520				18	143	944	958	1492	965		ظ
4.63717	204282			9	181	684	3947	14845	46144	138472		ع
0.24620	10846				6	36	515	1604	2932	5753		غ
2.28596	100704		6	33	86	584	2098	6785	12457	78655		ف
2.48772	109592		2	1	88	350	2352	6591	30605	69603		ق
1.51982	66953			14	101	710	3832	6168	23626	32502		ك
12.11075	533517			21	253	1499	9909	26657	184785	310393		ل

تحليل
إحصائي
لدعم التعرف
الآلي

جدول ه :
تكرار الحروف
في مقاطعها

نسبة	مجموع	تكراره في المقطع الذي بطول									الحرف
		9	8	7	6	5	4	3	2	1	
5.40935	238299	1	8	148	799	4783	14581	54796	68082	95101	م
7.37929	325081	19	56	401	638	2906	9593	21691	198498	91279	ن
4.63272	204086		9	127	956	4291	32520	85099	36817	44267	ه
4.39562	193641			73	588	2261	7272	15456	56242	111749	و
1.47871	65142				39	1659	6550	35409	18616	2869	ى
6.31825	278339		25	114	1279	3387	20342	79313	87855	86024	ي
3.45775	152325										أخرى

مجموع	تكراره كأول حرف في مقطع بطول									الحرف الأول
	9	8	7	6	5	4	3	2	1 (منفصل)	
11900									11900	ء
3552									3552	آ
103951									103951	أ
792									792	ؤ
29677									29677	إ
6903					45	160	3602	2912	184	ئ
213387									213387	ا
155995		10	29	483	2456	6286	24975	106216	15540	ب
17596									17596	ة
35488		15	60	471	1885	4191	13122	9610	6134	ت
52252				13	121	743	39515	9600	2260	ث
26785			4	64	1160	2993	5296	14302	2966	ج
73129			7	67	624	3682	14474	51019	3256	ح
23053			22	77	636	1487	11693	8895	243	خ
18951									18951	د
13529									13529	ذ
56148									56148	ر
8625									8625	ز
81833		3	110	328	2897	13131	34475	25053	5836	س
15804			1	15	121	6015	4529	4793	330	ش
32597			6	40	270	1097	25817	4886	481	ص
10357				10	83	349	793	7559	1563	ض
4917				28	161	916	1374	2025	413	ط
965				8	17	146	529	225	40	ظ
138472			23	231	4676	31891	46091	53599	1961	ع
5753			1	32	180	612	3346	1365	217	غ
78655	4	58	331	1340	3133	11738	24425	35292	2334	ف
69603	2		2	68	1773	1920	7924	52950	4964	ق
32502			21	51	743	2661	8273	17678	3075	ك
310393	20	64	528	2181	12144	44072	95723	87502	68159	ل
95101	1	5	61	376	1694	15878	14292	47959	14835	م
91279			28	281	1076	4324	12424	31471	41675	ن
44267				9	114	948	8722	23695	10779	ه
111749									111749	و
2869									2869	ى
86024	1	7	88	1188	3017	14698	17471	39753	9801	ي

جدول ٦ :
تكرار الحرف
الأول في المقطع
مع طول المقطع

مجموع	منفصل	متصل في آخر المقطع	الحرف الأخير
11900	11900		ء
5042	3552	1490	آ
126607	103951	22656	أ
3102	792	2310	ؤ
35998	29677	6321	إ
369	184	185	ئ
509567	213387	296180	ا
25466	15540	9926	ب
54682	17596	37086	ة
33174	6134	27040	ت
6629	2260	4369	ث
4461	2966	1495	ج
6645	3256	3389	ح
507	243	264	خ
133415	18951	114464	د
28974	13529	15445	ذ
172062	56148	115914	ر
21234	8625	12609	ز
13067	5836	7231	س
1976	330	1646	ش
1377	481	896	ص
3044	1563	1481	ض
1584	413	1171	ط
995	40	955	ظ
13135	1961	11174	ع
699	217	482	غ
5988	2334	3654	ف
8460	4964	3496	ق
16976	3075	13901	ك
94308	68159	26149	ل
77093	14835	62258	م
172437	41675	130762	ن
133188	10779	122409	ه
193641	111749	81892	و
65142	2869	62273	ى
81907	9801	72106	ي

جدول ٧ :
آخر حرف:
حرف متصل في
آخر المقطع أو
منفصل

مجموع	منفصل	ي	ى	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ة	ب	ا	ى	إ	و	أ	أ	الحرف الأول			
6903	184	378	2	6	412	104	642	488	323	32	165	9	45		113	144	1	2347	10	95	180	89	264		39	6	4	141	373	299	8						أ			
213387	213387																																					أ		
155995	15540	23447	175	9491	6336	56482	2050	1709	4933	789	140	355	5005	44	467	91	487	1683	681	80	5364	404	980	275	693	170	168	1518	991	258	13703	115	298	6	1009	58	أ			
17596	17596																																					ب		
35488	6134	2194	655	1033	2091	919	1532	1000	1159	1449	542	248	6038	49	333	171	681	324	1019	349	1623	150	510	601	1174	474	30	360	14	968	961	10	176	517		ب				
52252	2260	317	10	559	522	37838	6704	1251	156	132	2	4	25							298		68	2		1			294	515	124	1169	1				ت				
26785	2966	311	24	813	1541	892	1512	5186	60	53	2750								47	81	228	2506	86	1101	143	3	17	738	257	1209	4052	196	4	9		ت				
73129	3256	2207	92	573	127	306	5984	986	272	548	832			37	58	180	419	137	1229	187	2402	283	43452		1109	50	4300	157	1978	1968						ج				
23053	243	1808	39	394	8	53	505	3119	1	248		1			387	81	242	265	87	85	3176	1367	220				27	474	14	8282	1927					ح				
18951	18951																																					ح		
13529	13529																																					د		
56148	56148																																					د		
8625	8625																																						ر	
81833	5836	1063	1668	15827	1423	1736	7819	26121	276	415	2990	5270		188				5	879		247			67	2029	426											ر			
15804	330	4089	8	154	2465	49	333	6	232	310	255	33	3019	68				4	5	1164		592	15	60	294	1	475	37	372	997	240	20	1047	2		س				
32597	481	564	41	626	162	357	620	23704	11	2	523	114	133	65						193		877	63	1014				10	54	828	2155						س			
10357	1563	790	4616	420	157	50	75	91	21		6	9	801	95					399		2	21	216	8				153	94	44	329	20	17	360		ص				
4917	413	426	10	419	179	34	268	1002	17	1	95	10	512					5	18		226			12				16	35	149	1013	16	2	39		ظ				
965	40	15	2	369	200	6	190				44	5									18								3	10	60	3					ظ			
138472	1961	3543	135	1829	686	48249	9684	37005	406	982	307	1	317	1510	56	224	697	164	518	2587	338	1233			323	938	1038	1016	17560	5165							ع			
5753	217	1867		42	26	488	100	442			359			67	198	24	51	372	373	270	4	231					22	174	3	143	280							ع		
78655	2334	18251	181	1294	1099	1344	1495	5608	1353	13300	341	267	3834	79	437	1111	907	315	1295	206	2573	906	923	722	672	1385	68	2041	336	677	3762	46	3381	6	6041	65	ف			
69603	4964	1590	31	2738	362	187	305	3023	55	10	209	427	24	534	444	287	18	214	54	2078	43	3789	47		47		2	3405	814	2303	41631			15			ف			
32502	3075	1457	14	741	348	2196	2404	3876	24	47	554	5	1333	2	8	21	10	61	279	22	3793	1045	57	9	52	4	1033	1224	120	958	7297	6	12				ق			
310393	68159	9661	10516	6031	8796	16602	17858	63348	9618	5632	1062	5693	388	1261	482	3576	2691	4280	3207	6898	2806	3642	2555	7750	3433	1139	7385	435	3814	16127	216	2586	40	9294	1330	ل				
95101	14835	1984	142	3176	1475	23598	718	1445	1166	488	112	238	5954	40	288	602	282	323	3486	196	5690	39	345	425	7037	679	1573	1375	2453	157	14233	234	230	82	1		ل			
91279	41675	7271	199	1092	6614	274	1696	26	1351	503	1713	58	1558	392	627	119	1857	228	4510	1349	185	156	184	220	1174	370	74	2465	650	1235	11267	68	28	91		ن				
44267	10779	4527	58	3323	197	623	3612	3779	300	35				103	5	3	1566		61	4767	3264	1063				110		92	52	2136	3559			251			ه			
111749	111749																																						و	
2869	2869																																							و
86024	9801	95		5659	1422	5917	2160	1380	1892	5930	885	520	3740	120	461	631	1323	969	2428	2072	5760	710	5579	909	6268	1734	2763	3276	2385	2014	5695	33	473	1020		ي				

جدول ٩ :
تكرار ظهور
الحرف والذي
يليه في المقطع:
الحرف الثاني
- الحرف
الثالث

مجموع	منفصل	ي	ى	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ة	ب	ا	ى	إ	و	أ	أ	الحرف الثاني	
1459	1459																																				آ	
20129	20129																																				أ	
1257	1257																																				ؤ	
6277	6277																																				أ	
1374	98	14		10	2	161		168	108							8			40	1	76	219	17								19				أ			
139833	139833																																				أ	
46774	4486	5805	15	441	618	431	7	2226	439	362		133	1216	3	187	112	126	139	362	70	9453	45	13281	90	678	36	96	806	291	106	4502	96	4	1	111	ب		
10867	10867																																				ب	
33473	9949	2830	4100	759	1909	907	1459	1217	820	782	385	273	518	71	224	47	283	167	359	187	937	41	267	261	723	255	47	141	78	1082	1969	21	6	399	ت			
7909	2173	711	49	226	428	128	1094	1647	8	202	4	19									512	10	3													ت		
10529	1014	376	14	326	355	1245	729	558	22	21		915					2	17	45	217	1002	38	817	3	141	21	7	117	221	383	1770	131	3	19		ث		
28380	1432	4810	79	972	98	643	7451	981	497	427	128			5	63	142	159	115	876	86	1233	98	1860		8	986	47	391	119	671	4003					ج		
5884	56	598		151		119	373	682		155					701	66	77	85	38	102	1123	322	423														ح	
65359	65359																																					ح
11690	11690																																					د
53029	53029																																					د
9087	9087																																					ذ
20162	3726	1001	56	503	323	862	1610	2242	544	210	301	3	919		149						860	513	106	197	512											ذ		
11727	386	1282	62	127	389	27	582	1	115	214	219	47	464		46					2	2005	142	19	21	171											س		
10323	166	740	52	434	50	264	275	2994	8	17	560	42	169		9		24				858	750	49	136												س		
4336	357	513	222	58	19	8	68	562			9	16	197		48						714	9	9	357	26											ص		
7070	563	923	178	519	173	220	62	900	14	11	228	2	577		13						630	1	1	43	67											ص		
1492	47	120	2	12	322	84	149	160	4		22		25								392																ط	
46144	5890	5641	35	633	1690	2109	2882	2915	208	727	1189			124	375	1043	452	529	134	653	1295	231	3387														ظ	
2932	36	494	6	66	1	258	40	329			187			52	195	10	65	234	108	281	5	183															ظ	
12457	1281	2653	27	306	148	119	6	274	20	324	27		585	156	212	419	630	18	1190	87	1323	19	53	58	48	252	15	682	405	1026	38	20	36		ع			
30605	1153	2254	138	5018	138	5018	271	124	381	2055	18	6	75	316	16	426	290	316	39	261	14	1805	57	1637													ع	
23626	9015	641	26	865	76	1346	1867	952	14		415		228	1							4176	452	29														ف	
184785	11626	34037	29486	1275	61098	912	27334	136	930	1419	1329	493	402	40	78	6	123	36	344	54	75	106	283	33	881	67	82	2637	671	887	7641	9	38	19	185	ف		
68082	17512	3564	113	764	650	4879	117	1698	532	445	139	555	5808	26	268	101	264	1506	2786	450	8988	10	1753	162	374	243	1853	632	552	190	10376	7	1	735	28	1	ل	
198498	113822	8874	306	659	9244	134	201	15	1137	392	697	6	873	496	107	331	906	91	646	746	26	175	2363	213	412	238	16	2444	1346	9845	41704					ل		
36817	18224	833	439	648	401	550	3722	972	49	5				3	14						139	917	1328	1311												ن		
56242	56242																																					هـ
18616	18616																																					و
87855	41500	599	4	684	6834	6161	4387	2973	681	1048	699	30	1614	20	121	234	229	168	2344	41	5485	56	1804	179	254	83	535	1811	986	2885	2163	949	87	207		ي		

مفصل مجموع	ي	ي	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ة	ب	ا	ئ	ا	و	أ	أ	الحرف الثالث				
22																																								
2057																																								
995																																								
43																																								
1289	8		23	30	17	1	19	13								4	1	1	66	16																				
113405																																								
23260	2295	9654	26	299	677	261	531	86	56		78	313	53	99	13	99	324	26	1014	56	594	17	347																	
8824																																								
15133	6772	1009	10	223	1340	422	742	710	292	210	120	130	185	64	152	13	67	52	140	42	332	6	85	63	443	81	8	73	44	323	856	10	4	110						
3554	936	75	5	38	64	1183	118	762	36	7	1	3								134																				
3320	421	163	15	199	11	61	119	172	2	4	134									47	710	5	400	3	3	1	45	89	118	543	1	1	21							
4373	645	267	109	260	44	88	169	120	165	76	81		1	5	9	41	30	41	5	613	3	239																		
1186	83	84		22	14	18	17	291			21	17								33	190	76	5																	
30893	30893																																							
3202	3202																																							
44166	44166																																							
2960	2960																																							
9913	2068	1115	397	290	493	738	349	915	271	83	53																													
3060	1091	289	88	54	110	26	22	10	10	26	73	3	13		27					1	775	31	5	3																
3605	578	164	14	675	25	69	80	257		25	117	8	53		9	24					642	1111	39																	
3008	755	216	86	48	369	34	64	270	57		2	9	39		3	2					451	26	2	32	4															
3082	247	245	27	109	91	88	47	199	2	8	57	2	287		6						389		2	90	42															
958	103	53	2	9	40	22	50	90	5			8									494																			
14845	3388	515	32	783	392	455	670	1276	67	105	80			4	47	101	34	6	33	13	425	49	438																	
1604	213	386	5	56	31	149	14	52	24		127				8	31	5	10	6	220	15																			
6785	1331	427	63	192	248	82	2	124	29	225	46	1	167	707	32	107	34	3	175	4	1708	9	4	53	20	7	15	323	63	553	4	2	25							
6591	1192	545	161	1268	287	63	249	261	67	9	126			22	9	107	59	100	7	26	4	315	5	307	16															
6168	2346	371	69	177	87	119	1310	116	28	77		5																												
26657	8153	1480	724	1147	1507	494	2524	77	883	952	383	259	153	2	65																									
54796	32716	789	101	701	359	578	28	746	97	18		10	1067		16	18	22	23	807	23	2117	5	52	4	14	779	1672	5249	21	23	20									
21691	10213	2194	334	200	889	167	381		378	60	135	393	135	48	3	42	34	57	170	143	145	10	53	18	49	383	1407	661	2944	39	1	5								
85099	72280	237	153	359	26	324	5016	95	5	24	18																													
15456	15456																																							
35409	35409																																							
79313	13599	56	4129	140	25386	4135	2609	2816	1315	468	577	15	1055	22	387	181	255	111	300	52	3998	44	7008	64	258	44	931	1744	1043	2804	3528	164	16	57	2					

جدول ١١ :
تكرار ظهور
الحرف والذي
يليه في المقطع:
الحرف الرابع
- الحرف
الخامس

مجموع	منفصل	ي	ى	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ة	ب	ا	ى	و	ا	الربع	الحرف	
6	6																																			ا	
430	430																																			ا	
55	55																																			ا	
1	1																																			ا	
298	25	7		10	19	24	3		4										2	2	4	37					45	68		48					ا		
29471	29471																																			ا	
6578	2331	701	2	171	248	88		157	44	13		3	73		19	16	6	4	32	3	448	24	21	22	7	140	1583	2	415	2	3			ب			
11377	11377																																			ب	
11861	8750	165	12	134	1243	138	409	101	202	52	17	9	62	6	16	21	12	13	19	15	82		24	5	109	8	4	26	30	58	106	6	7		ب		
1540	1184	77		26	31	40	27	15	6												32						34	13	3	52					ب		
968	51	18	1	16	3	15	15	46					52							9	15		641				10	4	29	41	2				ب		
1729	1097	55	10	74	68	28	32	44	7	18	27				3	3	7	9	1	41		23				2	46	44	25	65					ب		
248	101	4		4	8	4	5	22	2	14		1		10	4					17	2	4				25	12	7	2					ب			
16728	16728																																			ب	
426	426																																			ب	
15879	15879																																				ب
499	499																																				ب
2325	1243	72	12	103	148	53	74	116	47	21	18	3		4					4		60		5	1	11	13	131	56	28	97					ب		
466	166	47	4	18	11	1	2	3	5	17	16	1		13						33		8					11	81	10	17					ب		
692	110	39	1	32	17	16	42	82	4	6	12			1						69		28		2		29	138	35	29						ب		
697	288	20	3	41	64	13	56	68				6		2						9			4			28	32	20	40						ب		
1108	292	147	6	52	27	11		38	4	6	13	93								41			1	1	1	35	43	11	284	2					ب		
944	696	2	1	3	12	4	1	103												93							7	10		12						ب	
3947	1439	143	2	215	269	110	61	73	65	9	14			6					3	2	15	6	23		12	21	273	476	15	695					ب		
515	208	109	5	41	33	29	3	2			48				2				3	3	9						16			4					ب		
2098	756	256	7	119	60	39		17	10	41	25		69	3	3	3	1			13	133	2		3	3	1	171	244		114		1	1		ب		
2352	942	77	57	201	90	98	16	85	1	3			12	101	22	1	22		4	1	57	12	11				176	86	86	190					ب		
3832	2009	88	2	27	47	62	1009	26	16		49	3	2						1	73		114	2	15	74	163	33	11	1	5					ب		
9909	5283	288	133	487	824	166	753	33	58	84	120	22	12	1	15	8			49			7	3	21	8	350	556	205	390	26	2	3	2		ب		
14581	8012	312	8	339	252	134	48	152	32	9		147	2	3	37	1			48	169		85	6	2	2	167	812		3789	13					ب		
9593	4050	686	1337	98	315	26	88	62	3	20	1	36	13	8					5	1	4	3			8	2	6	9	11	72	732	39	1954	3	1		ب
32520	28364	26	11	40	19	158	1729	25	7					1						10	31	81		3		2	14	13	1986							ب	
7272	7272																																				ب
6550	6550																																				ب
20342	13847	48	45	10	483	1649	410	291	134	68	191	1	114	15	30	26	39	5	85	3	961	19	318	10	104	3	15	337	153	580	323	16	9		ب		

مجموع	منفصل	ي	ى	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ة	ب	ا	ئ	أ	الحرف الخامس			
3	3																																	ا			
38	38																																	ا			
3	3																																	ؤ			
71	28	2		1		13	1		1												6	2					3	11		3			ئ				
10664	10664																																	ا			
1199	777	87		52	34	11		77	6			12		2	1	1	3	23	2	37	2	2	6			11	44		11				ب				
5350	5350																																	ة			
2205	1415	104		37	395	33	74	7	40	2	1	2	23	1				11	5	6		2	6	2		1		4	33	1			ت				
77	49	3	3	1	1	2		3												12							1		5				ث				
58	8	8		4	1			16				1								1	16								3				ج				
288	170	15		14	14	4	2	8	1	7	1							1		10	1	1				1	27	2	10				ح				
31	23										3	2								2														خ			
1401	1401																																		د		
102	102																																		ذ		
2392	2392																																		ر		
52	52																																		ز		
273	167	2		8	27	8	6	11	16	1										11							1		8	5				س			
36	3	3		3	4					1										13															س		
101	41	11								5	2									17	9	1						2	11		1				س		
107	72			1	2		2					20																							ص		
155	68	9	1	17	5	3	1	1	1			16								7								1	3	2	18	1			ص		
143	95			6	2	4	2	3												16								12		3					ظ		
684	335	55	3	43	46	34	7	36	2	3																2	17	25	29		39				ع		
36	21	1	1	4	1	2			1	4																										غ	
584	216	8	1	41	24	7		1	5	2	14	7								67						5	14	127		36	4				غ		
350	157	37		21	25	11	7	11	1			1	11	3						6	1	2					3	31	3	19				ف			
710	449	5		5	30	4	178	2	1	8										9							9	4	3						ق		
1499	902	100	3	95	94	45	36	9	2	40	34	4	1															40	45	6	23				ق		
4783	3262	424		214	50	24	1	48	1	3		8								74	4						18	110		523					ل		
2906	1727	384	17	20	80	2	4		16	1	1	3	1														3	5	85		551				م		
4291	2857	5	10		2	62	411	4												3	12						1		1	923					ن		
2261	2261																																				ه
1659	1659																																				و
3387	2259	19		1	119	368	68	16	9	24	15		87	6	1	1	1	3	4	129	8	14					3	59	90	17	33				ى		

مجموع	منفصل	ن	م	ة	ا	الحرف الثامن
75	75					ا
1	1					ة
8	8					ر
6	6					ف
2	2					ق
8	4				4	م
56	56					ن
9	6		1		2	ه
25	4	19		2		ي

جدول ١٥ : تكرار ظهور
الحرف والذي يليه في المقطع:
الحرف الثامن - الحرف
التاسع

مجموع	الحرف التاسع
2	ة
6	ا
1	م
19	ن

جدول ١٦ : تكرار
ظهور الحرف التاسع
(أطول مقطع ٩ حروف)

المراجع

- Badr Al-Badr, Sabri A. Mahmoud. "Survey and Bibliography of Arabic Optical Text Recognition." J. of Signal Processing, Vol. 41, No.1, pp.49-77 (Jan. 1995).
- Asim Nabawi, Sabri Mahmoud. "Arabic Optical Text Recognition: A Classified Bibliography." Engineering Research Bulletin, Minufiyah University, Egypt, Vol. 23, No. 1, Jan. 2000, pp. 79-131.
- M. Khorsheed. "Off-line Arabic Character Recognition - A Review." Pattern Analysis & Applications, 5:31-45, 2002.
- Liana M. Lorigo, Venu Govindaraju. "Offline Arabic Handwriting Recognition: A Survey". IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 712-724, May 2006.
- M.Z. Khedher and G. Abandah "Arabic Character Recognition using Approximate Stroke Sequence." Third International Conference on Language Resources and Evaluation (LREC2002), 2002.
- Yousof S. ElArian, A Lexicon of Connected Components for Arabic Optical Text Recognition, Master thesis, Jordan University of Science and Technology, Amman, Jordan, August, 2006.
- Latifa Al-Sulaiti, Designing and developing a corpus of contemporary Arabic, Master thesis, School of Computing, University of Leeds, Leeds, UK, March, 2004.
- M.S. Khorsheed. "Recognising handwritten Arabic manuscripts using a single hidden Markov model". Pattern Recognition Letters, 24(2003) pp. 2235-2242.
- Issam Bazzi, Richard Schwartz, John Makhoul. "An Omnifont Open-Vocabulary OCR System for English and Arabic". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 6, June 1999.
- R. Schwartz, C. LaPre, J. Makhoul, C. Raphael, and Y Zhao. "Language-Independent OCR Using a Continuous Speech Recognition System." Proc. Int. Conf on Pattern Recognition, Vienna, Austria, pp. 99-103, August 1996.