

تحليل إحصائي لدعم التعرف الآلي على الكتابة العربية

A Statistical Analysis for Arabic Text Recognition Support

م. حسني المحتسب* د. صبري محمود* د. رامي قهوجي**

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

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

كلمات البحث: 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 [1-4]. 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 [5-8]. فقد قدم البحث [5] 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 4.3 وكذلك معدل عدد المقاطع في الكلمة العربية 2.2. 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 262647 كلمة تتكون من 1126420 حرفاً. وفي دراستنا تتضمن الإحصاءات تحليلات إضافية سنستعرضها لاحقاً كما أن حجم العينة المستخدمة أكبر. أما البحث الثاني ذا العلاقة [6] فهو يهدف إلى إنتاج معجم لعملية التعرف على الكتابة العربية وذلك في استخدامه للتعرف على مقاطع اللغة العربية وذلك تجنباً لتقطيع الحروف العربية (الذي يؤدي إلى زيادة الأخطاء في التعرف

الآلي). اعتمد البحث على تحليل 26805 ملفا بصيغة أنش تي ام ال (HTML) لمجلة الدستور الأردنية. وتحتوي تلك العينات على الأخبار المحلية وفيها أخبار زيارات دبلوماسيين أجانب والتي تتضمن كلمات أجمية منها أسماء الدول والمدن والأسماء الأجنبية. وقد احتوت العينات على 28519734 مقطعاً. وقد بينت الدراسة إحصاءات للمقاطع بطول يتراوح من حرف إلى ثلاثة عشر حرفاً. وبدراسة النتائج فإن أعداد المقاطع بطول 12 و 13 قليلة جداً وبنسبة لا تكاد تذكر والراجح أنها تتكون من مقاطع تحتوي على أخطاء. كما بينت رسالة الماجستير [7] محاولة أخرى لتحليل الكتابة العربية بغرض عمل معجم. وقد كانت العينات المستخدمة مستخلصة من شبكة الانترنت بشكل رئيس. وتحتوي العينات على 842684 كلمة من 415 عينة مختلفة. وتحتوي العينات على مواضيع مختلفة كالاقتصاد 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 [6] بأمثلة من هذه المقاطع مما يجعل هذه العينات غير دقيقة المحتوى.

وقد استخدم بعض الباحثين [8-10] 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 ونظراً لحجم العينات المستخدمة فإن نتائج هذا البحث من المنتظر أن تكون أكثر دقة. تحتوي فقرة 2 من البحث على تعريفات للرموز المستخدمة والفقرة 3 على تحليل نتائج الدراسة والفقرة 4 على النتائج والاقتراحات.

تعريفات

نقدم 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 طوله 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"، وأطولها 1 و2 و5 و1 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 "أ" 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 1. 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 1. ومثال ذلك كلمة "متحابين"، فأخر حرف في المقطع الأول هو "ا" وهو متصل، وآخر 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 2، و"حتنا" بطول 4.

التحليل الإحصائي

تم عمل الدراسة على نص كتابي صحيح البخاري وصحيح مسلم كمثال على النصوص العربية المتعلقة بالحديث الشريف. وبلغ تعداد المقاطع في الكتابين 2217178 مقطعا مع التكرار، و 18170 مقطعا بعد حذف المكرر. بينما بلغ عدد الحروف والرموز في الكتابين ما مجموعه 4405318 رمزا. ويشمل الكتابان على 1095274 كلمة مع التكرار و50367 كلمة بعد حذف المكرر.

تبين الجداول من 1 إلى 14 مجموعة من النتائج المستخلصة من الدراسة. ويمكن للباحث الراغب في الحصول على الجدول الأساس الكلي والذي يزيد على 300 صفحة الاتصال بأحد الباحثين إذ لا يمكن إرفاقه مع هذا البحث. وفي يلي وصفا تحليليا مختصرا لهذه الجداول:

يعرض جدول رقم 1 أول 348 مقطعا من المقاطع الأكثر تكرارا من مجموع 18170 مقطعا. ولاحظ أن حرف الألف المنفصل يمثل حوالي 10% من تعداد المقاطع، يليه حرف الواو المنفصل، فالألف المهموزة، فاللام. ثم المقطع الثلاثي "الله". وهكذا. ويلاحظ المستعرض للجدول أن هناك العديد من المقاطع بطول 2 و3 و4 و5 وتظهر أكثر شيوعا من مقاطع أخرى بأطوال أقل. ويعرض جدول رقم 2 تكرار المقاطع حسب طولها ونسبة وجود هذه المقاطع. ونلاحظ أن أكثر من 90% من المقاطع أطوالها ثلاثة حروف فأقل، وأن ما يصل إلى 98% من المقاطع أطوالها أربعة حروف فأقل.

ونلاحظ أن المقاطع بطول 9 هي قليلة وتبلغ 8 مقاطع أغلبها ظهر مرة واحدة، وقد جمعنا هذه المقاطع في جدول رقم 3. ويبين جدول رقم 4 كل المقاطع بطول 8 حروف وعددها 62 مقطعا وهي أيضا قليلة الظهور كما هو واضح في جدول رقم 2.

يشمل الجدول رقم 5 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 1 لأنها لا تتصل بما قبلها ولا بما بعدها. بينما حرف الميم على سبيل المثال "م" ظهر منفصلا (في المقطع بطول 1) كما ظهر في كل المقاطع ذات الأطوال المختلفة من 1 إلى 9. وملاحظة أخرى تبدو لطيفة وهي أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتما لن يظهر في مقطع أطول من ذلك المقطع.

يشابه الجدول رقم 6 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 1)، وهذه الحروف هي: ء، ا، أ، و، إ، اة، د، ذ، ر، ز، و، ي. وتتكرر هنا ملاحظة أن الحرف إذا لم يظهر في مقطع بطول معين، فإنه حتما لن يظهر في مقطع أطول من ذلك المقطع.

أما الجدول رقم 7 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 متصلا 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 8 إلى 15 تكرار الحرف والحرف الذي يليه في بيانات الدراسة، فيظهر

الجدول رقم 8 تكرار ظهور الحرف الأول والحرف الثاني، ويعرض الجدول رقم 9 تكرار ظهور الحرف الثاني والثالث، ثم جدول رقم 10 يعرض تكرار ظهور الحرف الثالث والرابع وهكذا حتى جدول رقم 15 الذي يظهر تكرار ظهور الحرف الثامن والتاسع. ويستعرض الجدول رقم 16 تكرار ظهور الحرف التاسع. ونلاحظ من هذه الجداول أن أكثر الحروف الثنائية تكرارا تقع في المقاطع ذات الأطوال 2 و3.

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.

المراجع

- [1] 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).
- [2] 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.
- [3] M. Khorsheed, "Off-line Arabic Character Recognition – A Review," Pattern Analysis & Applications, 5:31-45, 2002.
- [4] 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.
- [5] M.Z. Khedher and G. Abandah "Arabic Character Recognition using Approximate Stroke Sequence," Third International Conference on Language Resources and Evaluation (LREC2002), 2002.
- [6] 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.

- [7] Latifa Al-Sulaiti, Designing and developing a corpus of contemporary Arabic, Master thesis, School of Computing, University of Leeds, Leeds, UK, March, 2004.
- [8] M.S. Khorsheed, "Recognising handwritten Arabic manuscripts using a single hidden Markov model", Pattern Recognition Letters, 24(2003) pp. 2235-2242.
- [9] 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.
- [10] 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.

For Personal Use

جدول رقم 1. المقاطع الأكثر تكراراً: أول 350 مقطعاً.

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	قا	1.8777	عن	1.794	ثنا	1.432
()	1.3445	()	1.3443	!	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	1	0.3751	خير	0.3578	2	0.3426	كا	0.3291	على	0.3285
عمر	0.3153	لر	0.3111	لك	0.2902	ثم	0.2846	ني	0.283	محمد	0.279
ت	0.2767	فا	0.2725	لو	0.272	له	0.2646	س	0.2632	ير	0.2598
يا	0.2569	مر	0.2566	يو	0.2552	يد	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	ضى	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	أ	0.1602	يقو	0.1556
بكر	0.1528	فا	0.1525	جل	0.152	هو	0.1499	هد	0.1472	ح	0.1469
نس	0.1452	لز	0.1446	خر	0.1432	مو	0.1432	ك	0.1387	به	0.1352
عنه	0.1348	سمعت	0.134	ج	0.1338	ى	0.1294	لذ	0.1266	قو	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	قر	0.0937	يز	0.0935
حا	0.0888	سفيا	0.0884	ع	0.0884	شعبة	0.0884	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	سى	0.0752	غير	0.0734	تر	0.0732
شبية	0.0726	بين	0.072	بعد	0.0714	ض	0.0705	عنهما	0.0687	قلما	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
لما	0.0608	نز	0.0608	لا	0.0608	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	ثا	0.0527	شر	0.0525	قيل	0.0522	ثلا	0.052
يت	0.0519	لها	0.0517	عنها	0.0497	جد	0.0497	هير	0.0494	قتبية	0.0493
نو	0.0493	شي	0.0489	سليما	0.0488	منه	0.048	هد	0.0479	لمثنى	0.0476
سا	0.0472	كذ	0.0471	قتا	0.0467	يعني	0.0467	تو	0.0466	يا	0.046
عة	0.0458	جعفر	0.0457	طا	0.0457	با	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	لمد	0.0415	لحد	0.0414	عثما	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	بنا	0.0383	نما	0.0382
حين	0.0382	بني	0.0381	صحا	0.0373	فلم	0.0371	عطا	0.0371	لقا	0.0371
قة	0.0367	جو	0.0367	حميد	0.0367	للهم	0.0363	لمر	0.0362	و	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	للبيث	0.034	مثل	0.034	كو	0.0334
لح	0.0333	بيج	0.0332	حمد	0.0332	لمو	0.0331	تي	0.033	بيد	0.0325
حب	0.0324	بنت	0.0322	سف	0.032	يذ	0.032	مي	0.0318	نك	0.0316
نمير	0.0316	هي	0.0311	نسا	0.0309	لقو	0.0308	سم	0.0306	تى	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	0.0266	حو	0.0258
فيفو	0.0257	مسلم	0.0257	نحو	0.0257	صم	0.0256	كثير	0.0256	لحسن	0.0254
ثو	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	ليبت	0.0232	لعا	0.0231	منهم	0.0231	نهم	0.0231	تقو	0.0231
تد	0.023	علم	0.023	لليل	0.0229	يه	0.0229	لسلا	0.0227	فكا	0.0226

جدول رقم 2. أطوال المقاطع وتكرارها ونسبتها.

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 رقم 3. كل المقاطع بطول 9 ونسبتها.

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
لمستضعفين	0.00086	فليقطعهما	0.00009	قسطنطيني	0.00009
ليثنيهما	0.00005	فانقتانهم	0.00005	يستنيهما	0.00005
مستقبلها	0.00005	فليانمساها	0.00005		

جدول رقم 4. كل المقاطع بطول 8 ونسبتها.

مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة	مقطع	نسبة
مستخافكم	0.00005	مستخافين	0.00005	ستطعمتها	0.00005	فليحملها	0.00005	للمطفين	0.00005
ملتصقتين	0.00005	فسقيتهما	0.00005	لمجيبتين	0.00005	لمستقبلة	0.00005	لمطمنين	0.00005
فهيجتها	0.00005	بحبيبتيه	0.00005	فليحليها	0.00005	تستطيعها	0.00005	فتستقبله	0.00005
فجعلتها	0.00005	تحفيتهما	0.00005	يستحسنها	0.00005	فغمستها	0.00005	يستكملها	0.00005
فقبضتها	0.00005	لقطعتكما	0.00005	ليبتليكم	0.00005	فليمسكها	0.00005	فليمتها	0.00005
يستعملها	0.00005	فمنعنيها	0.00005	فجمعتهما	0.00005	لمقتسمين	0.00009	فليستهما	0.00009
يستقبلكم	0.00009	ففظعتهما	0.00009	ليخلعها	0.00009	فليجعلها	0.00009	فليستجر	0.00009
فليستغفر	0.00009	فليستنشق	0.00009	فليتحلله	0.00009	فليمتحها	0.00009	يستلمها	0.00009
فتبيعنيه	0.00009	لمتخافين	0.00009	بمغيبتين	0.00009	مستضعفين	0.00009	ستقبلها	0.00009
تسليمتين	0.00009	تصليتهما	0.00014	لخليقتين	0.00014	لجهنمين	0.00014	لينعلمها	0.00014
لمتشبهين	0.00018	فليستنثر	0.00018	تستعملني	0.00018	تستعنيها	0.00018	للملحقين	0.00027
لمتكافين	0.00027	فليستعفف	0.00027	بسيقيهما	0.00032	ففختها	0.00036	ليقطعها	0.00036
فليطلقها	0.00045	للمسلمين	0.00086						

جدول رقم 6. تكرار الحرف الأول في المقطع مع طول المقطع

مجموع	تكراره كأول حرف في مقطع بطول								الحرف الأول	
	9	8	7	6	5	4	3	2		1 (منفصل)
11900									11900	ء
3552									3552	أ
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	هـ
111749									111749	و
2869									2869	ى
86024	1	7	88	1188	3017	14698	17471	39753	9801	ي

جدول رقم 7. آخر حرف: حرف متصل في آخر المقطع أو منفصل

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

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	إ	و	أ	آ	الحر ف الأول		
78655	2334	1825 1	181	1294	109 9	1344	1495	5608	135 3	1330 0	341	267	383 4	79	437	111 1	907	315	129 5	206	257 3	906	923	722	672	138 5	68	204 1	336	677	3762	46	338 1	6	604 1	65	ف	
69603	4964	1590	31	2738	362	187	305	3023	55	10	209		427	24	534	444	287	18	214	54	207 8	43	3789		47	2	340 5	814	2303	4163 1				15		ق		
32502	3075	1457	14	741	348	2196	2404	3876	24	47	554	5	133 3	2	8	21	10	61	279	22	379 3	104 5	57	9	52	4	103 3	122 4	120	958	7297	6	12		412	3	ك	
31039 3	68159	9661	1051 6	6031	879 6	1660 2	1785 8	6334 8	961 8	5632	208 2	106 2	569 3	388	126 1	482	357 6	269 1	428 0	320 7	689 8	280 6	3642	255 5	775 0	343 3	113 9	738 5	435	3814	1612 7	216	258 6	40	929 4	133 0	ل	
95101	14835	1984	142	3176	147 5	2359 8	718	1445	116 6	488	112	238	595 4	40	288	602	282	323	348 6	196	569 0	39	345	425	703 7	679	157 3	137 5	245 3	157	1423 3	234		230	82	1	م	
91279	41675	7271	199	1092	661 4	274	1696	26	135 1	503	171 3	58	155 8	392	627	119	185 7	228	451 0	134 9	185	156	184	220	117 4	370	74	246 5	650	1235	1126 7	68		28	91		ن	
44267	10779	4527	58	3323	197	623	3612	3779	300	35				103	5	3	156 6		61		476 7	326 4	1063			110		92	52	2136	3559		251			ه		
11174 9	11174 9																																				و	
2869	2869																																					ى
86024	9801	95		5659	142 2	5917	2160	1380	189 2	5930	885	520	374 0	120	461	631	132 3	969	242 8	207 2	576 0	710	5579	909	626 8	173 4	276 3	327 6	238 5	2014	5695	33		473	102 0		ي	

جدول رقم 9. تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثاني – الحرف الثالث

الحرف الثاني	أ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع					
أ																																			1459	1459				
أ																																				20129	20129			
ؤ																																				1257	1257			
إ																																				6277	6277			
ئ				19		31	402					17	219	76	1	40			8							108	168	161	2	10	14	98	1374	13983	13983					
ا																																				3	3			
ب	111	4	1	96	4502	106	291	806	96	36	678	90	1328	45	945	70	362	139	126	112	187	3	121	133	362	439	222	6	7	431	618	441	15	5805	4486	46774				
ة																																					10867	10867		
ت	399	6		21	1969	108	78	141	47	255	723	261	267	41	937	187	359	167	283	47	224	71	518	273	385	782	820	121	7	907	1909	759	4100	2830	9949	33473				
ث					488	63	141	2				3	10		512						19	4	202		8	164	7	1094	128	428	226	49	711	2173	7909					
ج	19	3	131	1770	383	221	117	7	21	141	3	817	38	100	2	217	45	17	2		915		21	22	558	729	124	5	729	355	326	14	376	1014	10529					
ح				4003	671	119	391	47	986	8		1860	98	123	3	86	876	115	159	142	63	5		128	427	497	981	7451	643	98	972	79	4810	1432	28380					
خ				352	289		169	3				423	322	112	3	102	38	85	77	66	701		155				682	373	119	151		598	56	5884						
د																																					65359	65359		
ذ																																						11690	11690	
ر																																							53029	53029
ز																																							9087	9087
س	5	816	74	30	2464	704	62	136	512	197	106	513	860		9						149				210	544	224	2	1610	862	323	503	56	1001	3726	20162				
ش		57	14		2634	122	217	235	171	21	19	142		200	5	2	6			219	47	464	46			214	115	1	582	27	389	62	1282	386	11727					
ص					1794	787	86	59			136	49	750	858				24		560	42	169	9		17	8	299	4	275	264	50	434	52	740	166	10323				
ض	3		1	734	266	86	54	26		357	26	9	714						9	16	197	48					562	68	8	19	58	222	513	357	4336					
ط	76	19	6	1508	224	23	33	67		43	67	1	630			38	18			228	2	577	13			14	900	62	220	173	519	178	923	563	7070					
ظ			1	79	9	64						392								4		22	25				160	4	149	84	322	12	120	47	1492					
ع				7869	316	580	146	680	245			231	3387	129	653	134	529	452	104	375	124				118	727	208	291	2882	210	1690	633	5641	5890	46144					

جدول رقم 12. تكرار ظهور الحرف والذي يليه في المقطع: الحرف الخامس – الحرف السادس

الحرف الخامس	أ	أ	ب	ة	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	منفصل	مجموع			
أ																																	3	3		
أ																																		38	38	
ؤ																																		3	3	
ئ																																		71	28	
ا																																		10664	10664	
ب																																			1199	777
ة																																			5350	5350
ت																																			2205	1415
ث																																			77	49
ج																																			58	8
ح																																			288	170
خ																																			31	23
د																																			1401	1401
ذ																																			102	102
ر																																			2392	2392
ز																																			52	52
س																																			273	167
ش																																			36	3
ص																																			101	41
ض																																			107	72
ط																																			155	68
ظ																																			143	95
ع																																			684	335
غ																																			36	21
ف																																			584	216
ق																																			350	157
ك																																			710	449
ل																																			1499	902
م																																			4783	3262
ن																																			2906	1727
ه																																			4291	2857
و																																			2261	2261
ى																																			1659	1659
ي																																			3387	2259

جدول رقم 13. تكرار ظهور الحرف والذي يليه في المقطع: الحرف السادس – الحرف السابع

الحرف السادس	ا	ب	ة	ت	ث	ح	د	ذ	ر	س	ش	ض	ع	ف	ق	ك	ل	م	ن	ه	و	ي	منفصل	مجموع
أ																							2	2
ئ																							4	4
ا																							2251	2251
ب	1								2								2		1	4	5		44	29
ة																							625	625
ت	3					1						2				3		9		18		14	203	153
ث	1								7									2			2		37	27
ج	9											2					2	2	1			1	18	1
ح	1																		2	4			50	43
د								1	1				1										4	1
ذ																							76	76
ر																							18	18
س																							420	420
ش	5	1							1								11			3			11	11
ص	8								3														48	27
ض									1														3	
ع																							11	1
ف																				1	1		9	7
ق																					2		7	1
ك																							18	14
ل																			3	10	18	1	181	113
م															25								6	4
ن															2								86	58
ه	1								5							1							88	49
و	1																1			10	9	7	101	72
ي	6	1																24					101	72
	6																						253	168
	105																						799	653
	69	5																					638	481
	247																						956	582
																							588	588
	9	1																					39	39
																							1279	843

جدول رقم 14. تكرار ظهور الحرف والذي يليه في المقطع: الحرف السابع – الحرف الثامن

الحرف السابع	ا	ة	ر	ف	ق	م	ن	ه	ي	منفصل	مجموع
ئ										2	2
ا										475	475
ب										8	8
ة										40	40
ت										1	1
ث			4							4	4
ح										2	2
د										7	7
ذ										7	7
ر										20	20
س								1		1	1
ش					2					2	2
ض										2	2
ع										9	9
ف			2	6					19	6	33
ق										1	1
ك						4				10	14
ل								3		17	21
م			2							98	148
ن								1	6	394	401
ه						4				96	127
و										73	73
ي							56	4		54	114

جدول رقم 15. تكرار ظهور الحرف والذي يليه في المقطع: الحرف الثامن – الحرف التاسع

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

جدول رقم 16. تكرار ظهور الحرف التاسع (أطول مقطع 9 حروف)

الحرف التاسع	مجموع
ة	2
ا	6
م	1
ن	19