UNIVERSITY OF LEEDS

This is a repository copy of *Arabic and Quranic computational linguistics projects at the University of Leeds* المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/81629/

Proceedings Paper:
Sharaf, A, Atwell, ES, Dukes, K et al. (8 more authors) (2010) Arabic and Quranic computational linguistics projects at the University of Leeds المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز. In: Proceedings of the workshop of Increasing Arabic Contents on the Web, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO). Workshop of Increasing Arabic Contents on the Web, 16 Oct 2010, Damascus, Syria. . (Unpublished)

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

المشاريع الحاسوبية على اللغة العربية والقرآن بجامعة ليدز

Abdul-Baquee Sharaf, Eric Atwell, Kais Dukes, Majdi Sawalha, Amal Al-Saif, Serge Sharoff and Katja Markert*
School of Computing, Leeds University, Leeds, England.
http://www.comp.leeds.ac.uk/arabic

Latifa Al-Sulaiti, Bayan Abu Shawar, Nora Abbas and Andy Roberts**
**Alumni of School of Computing, Leeds University*

ملخص الورقة

يهتم أعضاء فريق أبحاث اللغة التابع لكلية الحاسوب بجامعة ليدز بأبحاث حاسوبية لمعالجة اللغة العربية. فعندما قمنا في الماضي بعمل دراسة واسعة للأدوات المتاحة مجاناً للمعالجة النصية للغة العربية وجدنا عدداً قليلاً جداً من هذه الأدوات، ولكننا أدركنا أنه يمكن تطويع برمجيات التعلم الآلي (machine learning) للمعالجة اللغوية بشكل عام وتطبيق ذلك على اللغة العربية. وهذا الأمر يستدعي وجود نص عربي لتدريب هذه البرمجيات، ومن ثم قمنا بجمع أول مجموعة نصوص عربية (corpus) وجعلناها مفتوحة المصدر ومتاحة للتحميل المجاني وكذلك طورنا برنامج واجهة مريحة لعرض النص بشكل منسجم. هذه المجموعة اكتسبت شهرة واسعة من قبل خبراء المعالجة الحاسوبية للغة العربية واستخدموها لتدريب برامج حاسوبية أخرى وتقييم النتائج. كما أننا قمنا بتطوير أدوات لتحليل النص العربي المعاصر مثل التحليل الصرفي، التجذير والعنونة وذخائر لغوية للتغطية الواسعة للغة العربية و تطوير مجموعة نصية معنونة بالعلاقات الخطابية (discourse relations) للغة العربية على غرار العلاقات الخطابية للغة الإنجليزية الصادرة من جامعة بنسلفانيا.

ومن المجالات البحثية الرائدة التي نعتز بها في جامعة ليدز هو مشاريع حاسوبية للمعالجة اللغوية للقرآن الكريم وتعتبر إمتداد لتقنية التنقيب النصي للغة العربية التقليدية . ومن هذه الأبحاث تطوير برنامج محاورة آلية و موقع "قرآني" والذي يتيح البحث على مستوى مفاهيم معدة مسبقاً، وإطار لتمثيل المعرفة القرآنية والعنونة النحوية . ومؤخراً قمنا بتدشين موقع "ذخيرة لغوية عربية للقرآن الكريم" (Quranic Arabic Corpus) [com.quran.corpus//:http] ويعتبر مورد إلكتروني قابل للتحميل المجاني وموسم بمعلومات لغوية تفصيلية على مستوى الصرف وأقسام الكلام لكل كلمة قرآنية. ومنذ انطلاق الموقع حظي بقبول واسع واستقطب جمهوراً واسعاً من خبراء في أبحاث اللسانيات وكذلك الراغبين في تعلم لغة القرآن من غير الناطقين بالعربية، وهو الموقع الأول لدراسة قواعد القرآن على محرك جوجل للبحث ويجلب أكثر من 50 ألف زائر شهرياً. وهذا جعلنا نقترح "فهم القرآن" كتحد كبير لعلوم الحاسوب والذكاء الصناعي لعام 2010 وما بعده. وتدريجياً نقوم بتوسيع هذا المشروع حيث أننا الآن في طور إعداد شجرات بيانية تبعية لبنية الجملة القرآنية ((Quranic Arabic Dependency Treebank مستمدة من كتب إعراب القرآن الكريم وممكنة للقرآءة الآلية . كما أننا نقوم باستخلاص مفاهيم قرآنية وبناء شبكة روابط بين هذه المفاهيم وعودة ضمائر القرآن لهذه المفاهيم. ومن الخطط المستقبلية في هذا المشروع القرآني بناء ذخيرة لغوية لمفردات القرآن على غرار مشروع WordNet وكذلك تطوير ذخيرة إطارات دلالية لمفردات القرآن على غرار مشروع FrameNet وكذلك تطوير مجموعة نصية للقرآن معنونة بالعلاقات الخطابية، ومن ثم تطوير محرك بحث متكامل يتيح البحث القرآني على مستوي المفردات والأسلوب والنحو والصرف والمفاهيم وذلك من خلال الذخائر اللغوية آنفة الذكر. ومن خلال مشاريع ناجحة على القرآن الكريم نطمح تطبيق الخبرة المكتسبة في المعالجة الحاسوبية للنصوص التقليدية – مثل الحديث النبوي وكتب التراث العربي – وكذلك العربية الفصحى المعاصرة. فعلى سبيل المثال نخطط لتطوير مجموعة نصية معنونة تلبي أغراض البحث والتنقيب النصي للمختصين في مجال الإقتصاد الإسلامي وذلك من خلال التمثيل المعرفي للمفاهيم المعنونة في هذا الإختصاص. وكذلك عُرض علينا تمديد أبحاثنا الحاسوبية لتشمل كتب مقدسة في الأديان الأخرى مثل التوراة والإنجيل وهذا يمكننا من توظيف تقنيات حاسوبية لملاحظة التشابه والفروقات بين هذه المصادر.

إن مما يميز مواردنا كونها مفتوحة المصدر وغير ربحية، وهذا جعلها تكتسب شهرة واسعة وتعاد استخدامها من قبل الباحثين، على النقيض من الموارد المغلقة المطورة من قبل فرق أخرى. إن مشروع الذخيرة اللغوية للقرآن http://:com.quran.corpus/ جعلنا نلاحظ منفعة جعل الموارد مفتوحة المصدر: التغطية الإعلامية، تمكين النقاش الحر عبر ساحات الحوار، توظيف أدوات جوجل لتحليل الزوار ومعرفة مواقعهم الجغرافية، وغير ذلك من المنافع.

1. المقدمة

إن معالجة اللغات الطبيعية عن طريق الحاسوب تعد من المجالات الهامة لحقل الذكاء الإصطناعي ولها تطبيقات واسعة من أهمها المحاورة الآلية، الترجمة الآلية، برامج الإجابة الآلية للأسئلة، التلخيص الآلي، الربط الآلي للمفاهيم والتنقيب الآلي في النصوص وغير ذلك من التطبيقات. ومع مرور الزمن يشهد هذا الحقل تزايداً مستمراً في الأبحاث العلمية ونتجت عنها تطبيقات مفيدة في مجالات عديدة. والملاحظ أن اللغات الأوروبية وبالخصوص اللغة الإنجليزية وبعض اللغات الشرقية مثل الصينية واليابانية هي أكثر اللغات نضوجاً في مجال المعالجة الحاسوبية من حيث توافر أنظمة وحلول وذخائر لغوية وكذلك وجود فرق بحثية من أكاديميين وشركات تهتم بهذا المجال.

واللغة العربية بالرغم من أهميتها التاريخية وعدد الناطقين بها، إلا أن البحث العلمي في مجال الألسنيات الحاسوبية والمعالجة الآلية لا تتناسب وأهمية هذه اللغة ولا تضاهي أخواتها من اللغات الأوروبية. وهذه الملاحظة جعلنا في جامعة ليدز نصب إهتمامنا باللغة العربية. وبادئ ذي بدء أردنا أن نستطلع الأدوات والذخائر المتوفرة مجاناً للمعالجة النصية للغة العربية خصوصاً تلك التي تعتمد على الذخائر اللغوية مثل أدوات عرض النصوص (concordance) والتحليل الصرفي وعنونة أجزاء الكلام (part of speech) وأدوات للعرض المرئي للذخائر وكذلك قواميس متاحة للقراءة الآلية. ومن خلال هذه الدراسة [Atwell et al 04] أجرينا مقارنة بين مجموعة من هذه الأدوات ومثيلاتها المتوفرة للغات الأخرى وفي ختام هذه الدراسة وبناءً على الملاحظات الآنفة إقترحنا أن تعمل الفرق البحثية المتعددة لتحقيق هدف موحد وهو اختيار نص نموذجي للذخيرة النصية العربية والإتفاق على نوعية العنونة اللازمة لجعل هذه الذخيرة معياراً ذهبياً (gold standard) حتى يتسنى لأعضاء الفرق البحثية تقييم برامجهم وأبحاثهم الحاسوبية ضد هذا المعيار. واقترحنا أن يُعَدّ القرآن الكريم نموذجاً لهذا المعيار الذهبي لكونه نصاً هاماً ومتوفراً ومُشَكّلاً. وهكذا توالت أبحاث عديدة هنا في جامعة ليدز على المعالجة الحاسوبية للغة العربية والقرآن الكريم. وفي هذه الورقة سوف نعطي بشيء من التفصيل، نماذج من هذه الأبحاث، كما أننا نستعرض الأهداف والرؤي المستقبلية على ضوء المستجدات العلمية في هذا المجال ونلمس جوانب التعاون مع مؤسسات وهيئات علمية في العالم العربي.

2. مشاريع حاسوبية على اللغة العربية

2.1 تطوير ذخيرة لغوية عربية وأدوات استعراض

تلعب الذخائر اللغوية (corpus) دوراً بارزاً في العديد من تطبيقات المعالجة اللغوية وفي تعلم اللغة المرادة ودراسة خصائصها. فمثلاً من خلال التنقيب والبحث في الذخائر يستطيع المعَلّم والمتَعَلّم بحد سواء معرفة الخصائص اللغوية لكلمة معينة وسياقها في الجملة واختلاف معانيها اعتماداً على السياق والموضوع. وبناءً على هذه الملاحظات يمكن للمعلم إعداد خطة لتعليم الطلبة حسب المستويات. فعلى سبيل المثال أوضحت دراسة [van Mol 00] أجريت على ذخيرة عربية إخبارية تحوي 240 ألف كلمة من إذاعات مصرية وجزائرية وسعودية أن الإذاعة المصرية تستخدم عبارة "في الوقت نفسه" أكثر من عبارة "في نفس الوقت" على النقيض من الإذاعة الجزائرية، أما الإذاعة السعودية فهي تستخدم العبارتين بالتساوي.

ومن التطبيقات العملية الهامة للذخائر اللغوية إثراء القواميس بالخصائص اللغوية للمفردات واكتشاف الاستخدامات اللغوية الحديثة. فعلى سبيل المثال أوضحت دراسة [Braham &Ghazali00] بناءً على ذخيرة تحوي مليون ونصف كلمة أن كلمة "أَخَذَ" لها معنيان غير المعاني المتعارفة وهما: الشروع في العمل و الأخذ في عين الإعتبار. وهذا المعنى الأخير تمثل 11% من استخدامات هذه الكلمة بالرغم من عدم وجود هذا المعنى في القواميس التقليدية.

أما تلك الذخائر التي تحتوي على أكثر من لغة فلها أيضاً أهمية بالغة في التعليم حيث أنها تستعمل في إثراء أدوات لتعليم اللغات الأجنبية والهندسة اللغوية والترجمة الآلية وغير ذلك من التطبيقات. وبالرغم من توفر مثل هذه الذخائر للعديد من اللغات وخصوصاً الإنجليزية إلا أن اللغة العربية تشهد شُحّاً في هذه الثروة اللغوية.

وإنطلاقاً من هذه الفوائد والتطبيقات اللغوية للذخائر – أضف إلى ذلك كون جُلّ الذخائر العربية غير متوفرة مجاناً – سعينا في جامعة ليدز لتطوير ذخيرة للغة العربية تجمع نصوص حديثة بالعربية الفصحى وتغطي مواد إجتماعية وإقتصادية لتُلَبّي غرض مهندسي حوسبة اللغات وكذلك الراغبين في تعلم اللغة الإنجليزية لغير الناطقين بها. فكانت النتيجة إنشاء "ذخيرة لغوية للعربية الحديثة (Corpus of Contemporary Arabic)" [Al-Sulaiti and Atwell 06] والتي تحتوي على 843 ألف كلمة مأخودة من مصادر عديدة وتغطي مواد مكتوبة وبعض المواد المسموعة وتشمل مواضيع متعددة منها: قصص قصيرة، الدين، إقتصاد، سير ذاتية، مواد علمية، الصحة والطب، رياضة، مقابلات شخصية وسياسة.

وكذلك قمنا بإعداد برنامج واجهة لبحث واستعراض مواد هذه الذخيرة حيث أننا وجدنا أن برامج الاستعراض المتوفرة صُمّمَت لكي تتعامل مع النصوص اللاتينية وبالتالي تواجه مشاكل للتعامل مع النصوص العربية، أضف إلى ذلك ثراء اللغة العربية في توليد الكلمات من الجذور، كل ذلك جعلنا نَشْرع في تصميم مثل هذا البرنامج الذي سميناه (aConCorde) [Roberts et al. 06] والذي يتيح عرض نصوص عبر جذر بحثي بشكل منسجم مع إعطاء بعض الإحصائيات المفيدة والشكل التالي يعطي نموذج لهذا البرنامج.

الشكل 1 : برنامج واجهة لعرض نصوص الذخائر اللغوية

2.2 مشاريع حاسوبية للتحليل الصرفي للغة العربية

للتحليل الصرفي والعنونة الصرفية النحوية (Speech tagging and morphological analysis-of-Part) تطبيقات كثيرة. فهي متطلب سابق للعديد من تطبيقات المعالجة اللغوية لرفع كفائتها وتحسين مخرجاتها. إنَّ مشروع التحليل الصرفي والعنونة النحوية لنصوص اللغة العربية هو امتدادٌ لمشاريعِ معالجة اللغة العربية في جامعة ليدز. والهدف الأساسي من هذا المشروع هو توسيع مجال التحليل الصرفي والعنونة الصرفية والنحوية ليشملا النصوص العربية من نطاقات أوسع وأنواع وأشكال متنوعة من النصوص العربية المشكولة وغير المشكولة.

بدأ مشروع التحليل الصرفي والعنونة النحوية بدراسة المحللات الصرفية المطورة مسبقاً، وتم استخدام معياراً ذهبياً (Gold Standard for Evaluation) لتقييم نتائج التحليل الصرفي لثلاثة محللات صرفية متاحة على شبكة الإنترنت، وقد سلَّطت الدراسة الضوء على بعض القصور فيها، حيثُ أنَّ ربع كلمات الذخيرة لم يتم تحليلها بالشكل المطلوب، فلا نستطيع الاعتماد على هذه المحللات الصرفية لاستخدامها كمتطلب سابق للعنونة الصرفية والنحوية وبناء شجرة التحاليل النحوية (gspeech tagging and parsin-of-Part)، لأن الأخطاء الناتجة عن التحليل الصرفي سوف تتكرر في التطبيقات المعتمدة عليها (Atwell &Sawalha2008).

يتم تطوير محلل صرفي لنصوص الغة العربية يعتمد بشكل أساسي على المعلومات اللغوية المستخرجة من كتب قواعد العربية (Atwell &Sawalha2009b; Atwell &Sawalha 2009a)، ويعتمد أيضاً على مصدر معجمي واسع التغطية (Coverage Lexical Resource-knowledge Broad-Prior) لتحسين مخرجات التحليل الصرفي والعنونة النحوية ، تم بناء المصدر المعجمي من خلال استخراج المعلومات المعجمية من المعاجم العربية المختلفة في الترتيب، وتمَّ جَمعُ هذه المعلومات في مصدرٍ معجميٍّ واحدٍ واسع التغطية (Atwell &Sawalha2010a).

تتضمن مخرجات المحلل الصرفيّ على معظم المعلومات اللغوية الصرفيَّة للكلمة، وتشمل على جذر الكلمة وجذعها و وزنها ويحدد المحلل أجزاء الكلمة (السوابق واللوحق والزوائد في بداية الكلمة ونهايتها)، ويحدد لكل جزءٍ من أجزاء الكلمة العنوان الصرفي المناسب لذلك الجزء، وتُجمع العناوين الصرفية لأجزاء الكلمة بعنوان صرفي وحيد للكلمة المحللة يحدد خصائصها الصرفية بشكل تفصيلي (Atwell &Sawalha2010b) .

تتم عملية ترميز الخصائص الصرفية للكلمة المحللة باستخدام مجموعة العناوين للخصائص الصرفية التفصيلية للغة العربية[1] (Morphological Features Tag Set for Arabic)، والتي تغطي الخصائص الصرفية الراسخة في قواعد اللغة العربية بتمثيل واضح ومدمج، يتكون العنوان الصرفي (Tag) من سلسلة رمزية تتكون من إثنين وعشرين رمزاً حرفياً، كل رمز يمثل قيمة أو متغير ينتمي إلى إحدى الخصائص الصرفية أو النحوية، ويعدُّ موقع الرمز في العنوان مهماً في تحديد هذه الخاصية، وتمثل هذه القيم أو المتغيرات برمز واحد من حروف اللغة الإنجليزية الصغيرة، فمثلا الرمز (v) في الموقع الأول من العنوان يرمز إلى الفعل، والرمز (n) في الموقع الثاني يرمز إلى اسم العلم، ويُمثَّلُ الجنسُ في الموقع السابع من العنوان حيث يرمز الحرف (m) الى المذكر والحرف (f) الى المؤنث، وإذا كانت الخاصية اللغوية غير متوافقة مع الكلمة فالرمز (–) (الشرطة) يمثلها، بينما يستخدم الرمز (؟) (علامة السؤال) لترمز إلى أن الخاصية اللغوية تنطبق على الكلمة ولكنها غير محددة.

[1] مجموعة العناوين للخاصائص الصرفية للغة العربية http://www.comp.leeds.ac.uk/sawalha/tagset/html

2.3 مشاريع حاسوبية لتحليل العلاقات الخطابية للغة العربية

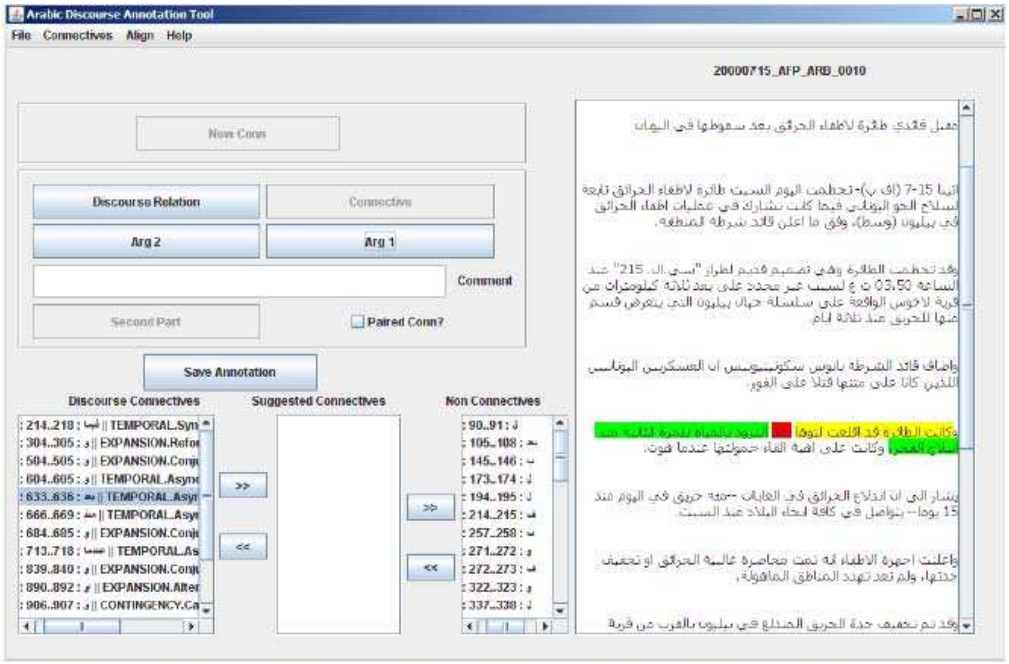
العلاقات الخطابية (discourse relations) لها دور بارز في بناء خطاب منسجم ومترابط. فمثلاً نجد أن عبارة "السيارة متطورة جداً ولكنها باهظة الثمن" حددت علاقة متناقضة (Contrast) بين جزئيتين (argument) ربطت بالرابط "لكن". وهكذا هناك علاقات خطابية أخرى مثل العلاقات السببية (Causual) التمثيل (Exemplification) الشرطية (Conditional) الخلفية (Background) الزمنية (Temporal) وغيرها.

وقد قمنا بتطوير أول ذخيرة لغوية للعلاقات الخطابية للعربية المعاصرة [Al-Saif and Markert 2010] وذلك بتحديد العلاقات بين الأجزاء المترابطة وتحديد نوعية هذه العلاقة والرابط. وتشمل هذه الذخيرة توسيم أكثر من خمسة آلالف علاقة خطابية في أكثر من خمسمائة نص إخباري. الشكل التالي يوضح مثال لعلاقة التشابه (similarity) بين جزئي الخطاب (arg1 and arg2) ومرتبط بأداة (discourse connective DC) "كما".

[ان العسكريان قتلا برصاصة في الرأس ثم شوهت
جثتهما]Arg1 [كما]DC[يحصل غالبا في عمليات
الخطف على يد القوات المسلحة.]Arg2

الشكل 2 : وسم عبارة بالعلاقة الخطابية مع تحديد جزئي العلاقة وأداة الرابط

وقد تم إعداد هذه الذخيرة عن طريق أداة للوسم اليدوي تم تطويرها خصيصاً لهذا الغرض وتُمكّن الموسم بتحديد النص واختيار نوعية العلاقة. والشكل التالي مثال لأحد النوافذ.



الشكل 3 : برنامج واجهة تساعد في الوسم اليدوي للعلاقات الخطابية

إن هذه الذخيرة هي الأولي من نوعها للغة العربية ونطمح أن تتيح المجال لخبراء المعالجة اللغوية العربية في تطوير تطبيقات مفيدة في المستقبل.

3. مشاريع حاسوبية على القرآن الكريم

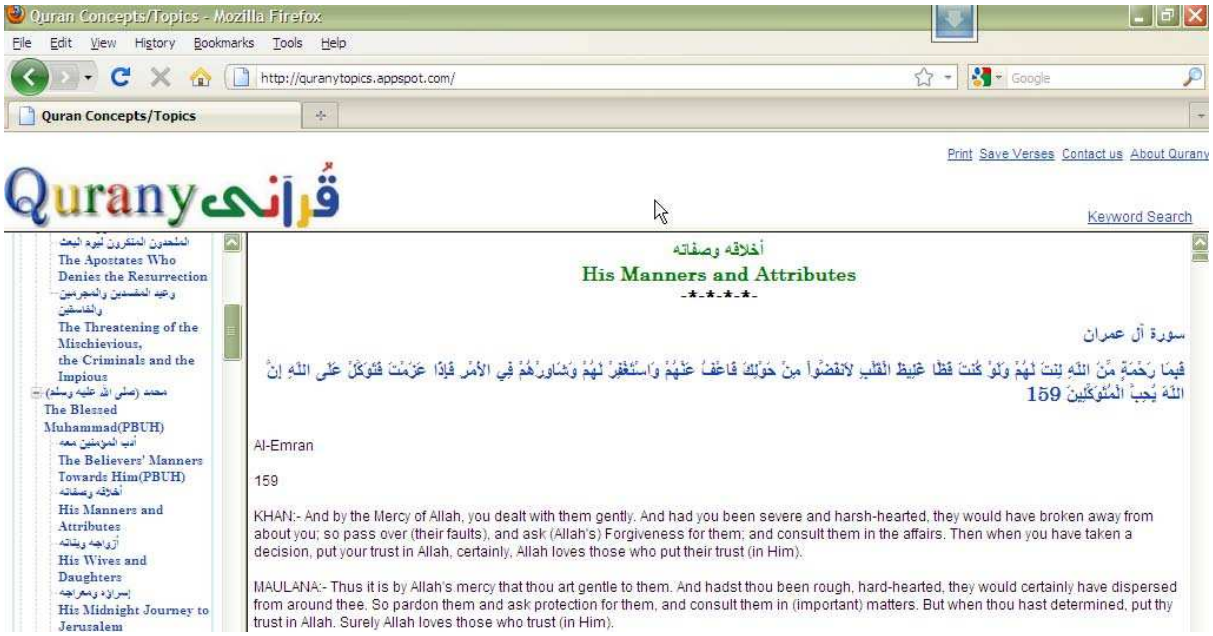
لقد أدركنا أهمية القرآن الكريم كنص للمعالجة الحاسوبية وذلك لعدة أسباب أهمها: كون القرآن نص مفتوح ومتوفر مجاناً، وهو نص مُشَكّل وذو حجم معتدل. أضف إلى ذلك كون القرآن نص مشهور ومتداول ويستقطب اهتماماً بالغاً من فئات شتى. وكذلك بسبب وجود مئات المؤلفات القديمة والحديثة في تحليل القرآن من كتب التفاسير والمعاجم وغيرهما يجعل أمر العنونة وتصنيف المفردات والتثبت من صحتها أسهل.

ولكن في الجانب الآخر ندرك أوجه الصعوبة في التعامل مع نص القرآن في المعالجة الحاسوبية. فكون القرآن كتاب ديني يستلزم التثبت من صحة المعالجة الحاسوبية وتجنب الخطأ في النتائج. وكذلك بسبب كون نص القرآن يتكون من نحو 80 ألف كلمة فقط قد لا يتناسب مع دوال البرمجيات الإحصائية على الذخائر اللغوية التي تتطلب ذخائر ذات مليون كلمة فما فوق، ولكن يمكن تمديد نص القرآن بإلحاق نصوص تقليدية أخرى مشابهة مثل الحديث النبوي وبعض كتب التراث العربي. كما يمكن جعل القرآن معياراً ذهبياً يمكن توظيفه في برمجيات معالجة أخرى.

وفيما يلي تفصيل لبعض الأبحاث والمشاريع المنجزة والجارية في جامعة ليدز على القرآن الكريم.

3.1 برنامج "قرآني" للبحث عن المواضيع القرآنية

قد يكون من العسير لغير المتخصصين البحث عن موضوع معين في القرآن الكريم. ومجرد البحث عن طريق كلمة بحثية لا يضمن ارجاع جميع النتائج المرجوة. ومن باكورة المشاريع القرآنية في ليدز مشروع الماجستير الذي قامت به الطالبة نورهان عباس حيث عمدت إلى تصنيف موضوعي متداول وهو "مصحف التجويد" [Mushaf at-Tajweed 1420H] وصممت موقع على الإنترنت يتيح البحث عن المواضيع باللغتين العربية والإنجليزية مع توفر عدد من التراجم الإنجليزية للقرآن. والشكل التالي نموذج لإحدى الصفحات.



الشكل 4: نافذة من موقع "قرآني" للبحث في المواضيع القرآنية

3.2 برمجيات المحاورة والإجابة التلقائية

برامج المحاورة الآلية (dialogue systems) تفترض إعطاء إجابة آلية لأسئلة مفتوحة والتحدي بالنسبة للحاسوب هو مدى مقدرتها على الإجابة واستمرارها في الحوار بشكل طبيعي وكأنّ المتحاورَيْن بَشَر. يشَكّل هذا النوع من البرامج تعقيداً كبيراً لخبراء الذكاء الأصطناعي وغالبية الحلول الموجودة تعتمد على تخزين عدد كبير من أنماط الأسئلة (Question patterns) المتداولة وأجوبتها. ولكن مثل هذه الحلول قد لا تتناسب مع القرآن الكريم ولا يمكن الإعتماد عليها في استخلاص أجوبة من القرآن. وقد عملنا في إعداد برنامج [04Shawar and Atwell] مبدئ للحوار يعطي آيات قرآنية تحوي بعض الكلمات المدخلة، أما إذا كان المدخل آية قرآنية معينة فإنها تعطي الآية التالية حيث أننا نلاحظ أنه في بعض الأحيان تأتي آية قرآنية لتجيب عن سؤال وارد في آية سابقة. الشكل التالي يعطي بعض الأمثلة.

Human: أحَـدٌ
Qur'an28-30: 1قُـلْ هُـوَ اللّـهُ أحَـدٌ (سورة الإخلاص)
26وَلَـا يُـوثِقُ وَثَـاقَـهُ أحَـدٌ (سورة الفجر)

Human: قُـلْ هُـوَ اللّـهُ أحَـدٌ
Qur'an28-30: 2اللّـهُ الصَّـمَدُ (سورة الإخلاص)

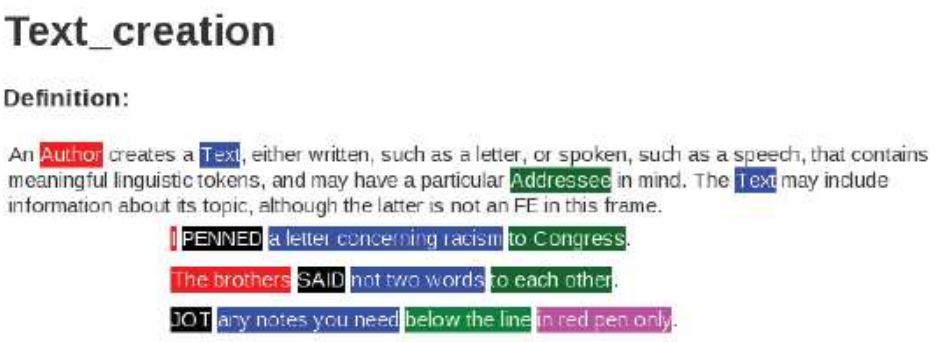
Human: عَـابِدُونَ أنـتُم
Qur'an28-30: 3لَـا أنـتُمْ عَـابِدُونَ مَا أعْـبُدُ (سورة الكافرون) وَ
5وَلَـا أنـتُمْ عَـابِدُونَ مَا أعْـبُدُ (سورة الكافرون)

الشكل 5 : مثال لحوار قرآني

من أجل الوصول لبرنامج أكثر كفاءة للأجوبة التلقائية يجب إثراء النص المرغوب بطبقات عديدة من المعلومات اللغوية والمعرفية التي تخص ذلك النص. فبالنسبة للقرآن يجب تغذية الذخيرة القرآنية بالتحليل الصرفي والنحوي لكل كلمة إضافة لمعلومات مفرداتية وخصوصيات قرآنية مستقاة من كتب التفسير حتى يكون برنامج الحوار أكثر كفاءة. ومن أجل ذلك أردنا أن نكون أولاً هذه الطبقات من الوسم اللغوي والمعرفي على القرآن الكريم قبل البدء في تصميم تطبيقات مفيدة.

3.3 تطوير إطارات دلالية لأفعال القرآن

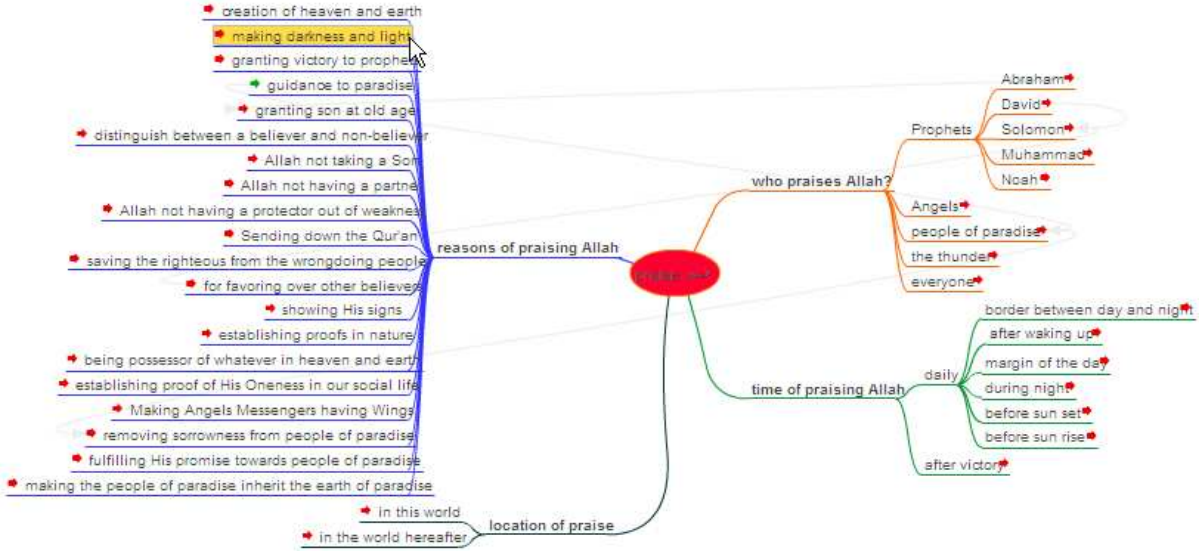
الإطارات الدلالية (Semantic Frames) تعنى بوصف الأدوار الدلالية الممكنة التي يمكن ان تأتي فيها الفعل أو الكلمة. فمثلاً يمكن للفعل "كتب" أن يُعَيّن في العبارة الواردة فيها مجموعة من الأدوار الدلالية مثل "النص المكتوب" و "الكاتب" و "الأشخاص المعنيين بالكتابة" وغير ذلك. وهذه الطريقة ابتكرها فيلمور [76Fillmore] في الستينات من القرن الماضي ثم أطلق مشروع FrameNet [2005Ruppenhofer et al] والذي يهتم بتطوير قاعدة بيانات مفتوحة على الإنترنت تحتوي على مجموعة من الإطارات الدلالية وتوزيع الكلمات الإنجليزية على هذه الإطارات. فمثلاً الشكل التالي يعطي تعريفاً لإطار دلالي باسم "إنشاء النص" (text creation) ويمكن إطلاق هذا الإطار عبر كلمات مثل الفعل "كتب" و"قال" و "خطّ" وغيرها.



الشكل 6: إطار دلالي يصف الأدوار الدلالية لإنشاء نص

ويتكون الموقع من نحو 900 إطار دلالي يمكن إطلاقها من خلال 10 آلاف كلمة من بين فعل وإسم وصفة. ومنذ إطلاق المشروع استقطب انتباه خبراء المعالجة اللغوية واستخدموا هذه الإطارات في تطبيقات لغوية عديدة، كما قام فرق أخرى بعمل مشاريع مشابهة للغات أخرى مثل الألمانية والإيطالية واليابانية وغيرها.

وقد عملنا بإجراء دراسة مبدئية عن إمكانية إجراء قاعدة بيانات مشابهة لأفعال القرآن الكريم [09Atwell &Sharaf] ووجدنا أنه وإن كان هناك إشتراك كبير مع إطارات FrameNet إلا أن الدلالات القرآنية العربية قد تكون لها استخدامات جديدة. فمثلاً نجد أن الفعل "كتب" في القرآن يأتي بمعنى "قَدّر" غير وارد في قاعدة بيانات rameNetF في مثل قوله تعالى "كتب في قلوبهم الإيمان". وهذا يَسْتدعي تَكْوين مشروع كبير يضم خبراء لغة وتفسير بتنسيق مع خبراء معالجة اللغات للقيام بدراسة المفردات القرآنية وملاحظة استخداماتها في الآيات القرآنية وبناءً على ذلك تعريف إطارات دلالية خاصة بالقرآن الكريم. وكتقريب لهذا المعنى فإن الشكل التالي محاولة لتحديد إطار دلالي قرآني يصف مفهوم "الحمد" من خلال سياقه القرآني.



الشكل 7 : محاولة لتحديد إطار دلالي قرآني لمفهوم الحمد

3.4 توظيف "التعلم الآلي" في معرفة خصائص السور المكية والمدنية

يعتمد "التعلم الآلي" (Machine Learning) على تحديد مجموعة من الخصائص المعرّفة على البيانات في النموذج المراد التعلم منها ثم جعل دوال التعلم الآلي "يتعلم" من هذا النموذج ليستطيع تحديد الخواص آلياً في نماذج جديدة. وهناك أدوات متوفرة لإجراء مثل هذه التجارب مثل برنامج WEKA المفتوح المصدر.

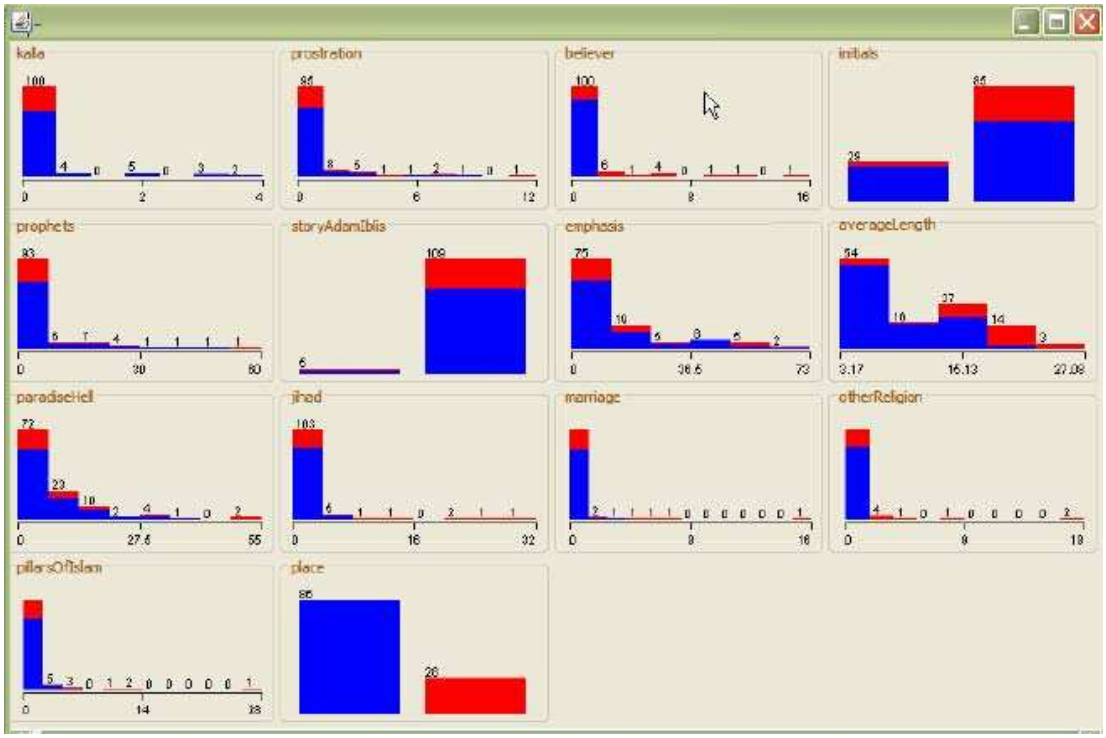
وقد قمنا بالإطلاع على خصائص السور المكية والمدنية من مؤلفات العلماء والمفسرين ثم قمنا بإعداد جدول يحدد خاصية كل سورة إزاء هذه المعايير. وبعد ذلك قمنا بإدخال هذه البيانات إلى Weka ومن ثم تطبيق دوال التعلم الآلي.

فمثلاً من خصائص السور المكية الإكثار من ذكر الآخرة وأهوال يوم القيامة والجنة والنار. وبناءً عليه قمنا بتحديد مجموعة من المؤشرات والكلمات الدالة على هذه الخاصية مثل (الجنة، النار، جهنم، سعير، قيامة، وعذاب) وقمنا بتعداد عدد تكرار هذه الكلمات ومشتقاتها في كل سورة. وهكذا حددنا قائمة تحتوي على 13 خاصية للسور المكية والمدنية. ويسمح برنامج Weka بإعطاء رسومات توضيحية مفيدة عن البيانات المدخلة فمثلاً يوضح الشكل رقم 8 إحصائية عن كل خاصية من الخواص الثلاثة عشر موزعة بين السور المكية (اللون الأزرق) والسور المدنية (اللون الأحمر). كما يمكن توليد رسومات تعقد مقارنة بين خاصيتين أو أكثر.

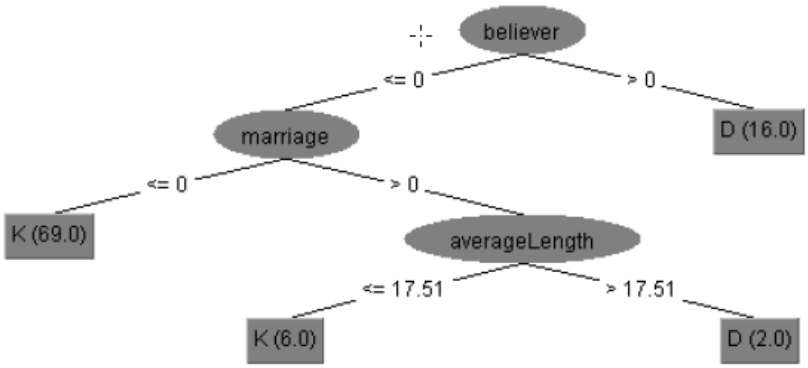
ثم يأتي التصنيف الآلي بناءً على دوال التعلم الآلي الذي يلاحظ الخواص في البيانات المدخلة ويحاول استخراج شجرة قرارات (decision tree) مناسبة مثل تلك الموضحة في شكل 9 التالي. توضح هذه الشجرة القرارات اللازمة لتحديد السور المكية (K) والمدنية (D) . فمثلاً فإذا كانت السورة تخلو من عبارة "يا أيها اللذين آمنوا" وتخلوا من مؤشرات الحديث عن العلاقات الزوجية فهي سورة مكية.

وهكذا أجرينا هذه التجارب على 93 من السور التي اتفق علي تحديدها العلماء، ثم اختبرنا مدى فاعلية هذه الدوال في تحديد ماهية السور الـ 21 المختلف فيها وقارننا ذلك مع المنشور في مصحف المدينة ووجدنا الإختلاف فقط في ست سور من ضمن هذه السور الـ 21.

كما يمكن تكوين مجموعات متشابهة (clustering) من السور آلياً بناءً على هذه الخواص فمثلاً اعتبرت دوال التصنيف أن سور البقرة والأعراف تتميزان بخصائص مشابهة فشَكّلتا مجموعة واحدة. وهكذا يوضح شكل رقم 10 بقية المجموعات.



الشكل 8 : رسومات بيانية مولدة من برنامج weka يوضح عدد السور ضد كل خاصية: السور المكية باللون الأحمر والمدنية باللون الأحمر



الشكل 9 : شجرة قرارات (decision tree) توضح مزايا السور المكية والمدنية

1		2		3		4		5		6		7		8	
2	D	6	K	15	K	74	K	8	D	1	K	31	K	3	D
7	K	10	K	19	K	83	K	24	D	50	K	35	K	4	D
		11	K	21	K	96	K	33	D	53	K	36	K	5	D
		12	K	23	K					55	D	44	K	9	D
		13	D	26	K					56	K	45	K		
		14	K	30	K					69	K	46	K		
		16	K	32	K					72	K	47	D		
		17	K	34	K					76	D	49	D		
		18	K	37	K					77	K	51	K		
		20	K	43	K					79	K	52	K		
		22	D	54	K					84	K	57	D		
		25	K	68	K					86	K	58	D		
		27	K							88	K	59	D		
		28	K							90	K	60	D		
		29	K							91	K	61	D		
		38	K							92	K	62	D		
		39	K							93	K	63	D		
		40	K							94	K	64	D		
		41	K							95	K	65	D		
		42	K							97	K	66	D		
		48	D							99	D	67	K		
										100	K	70	K		
										101	K	71	K		
										103	K	73	K		
										105	K	75	K		
										106	K	78	K		
										108	K	80	K		
										109	K	81	K		
										110	D	82	K		
										111	K	85	K		
										112	K	87	K		
										113	K	89	K		
										114	K	98	D		
												102	K		
												104	K		
												107	K		

الشكل 10 : تصنيف السور القرآنية إلى مجموعات (clustering) متشابهة والرقم يبين رقم السورة و السور المكية (K) والمدنية (D)

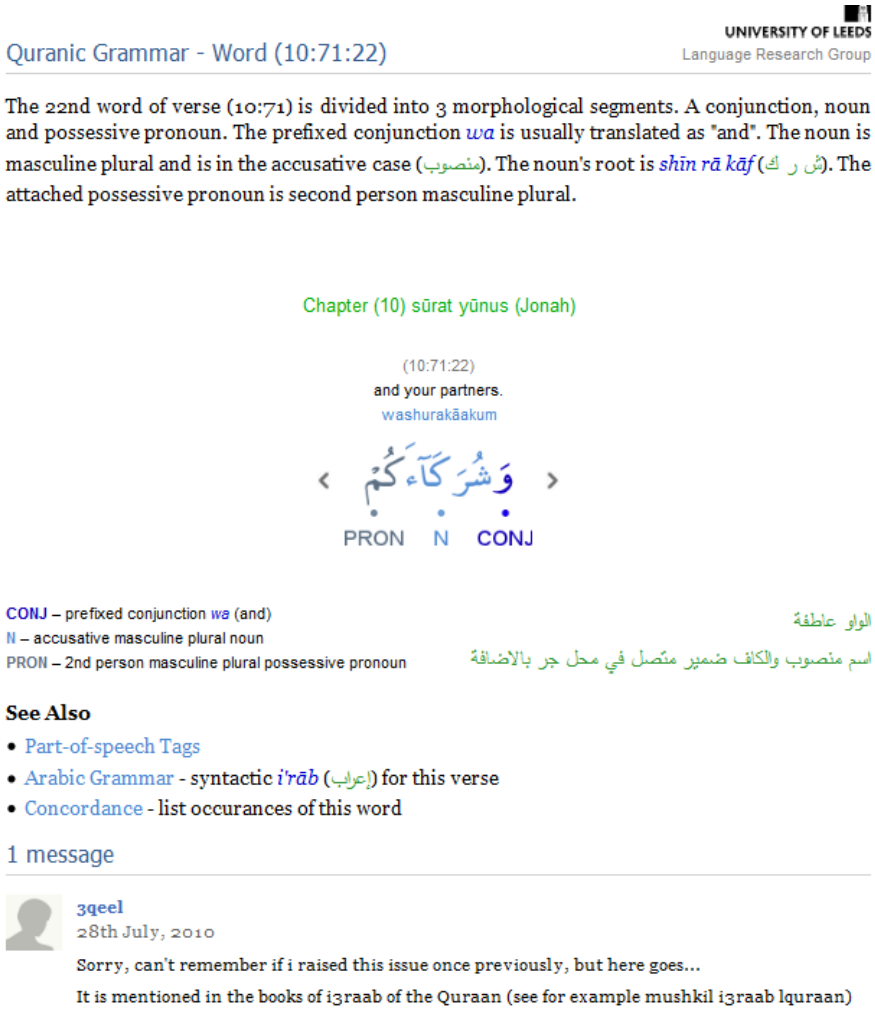
إن هذه التجارب توضح إمكانية توظيف تقنيات التعلم الآلي في اكتشاف خصائص وارتباطات مخفية في النص القرآني.

3.5 الذخيرة القرآنية العربية (Quranic Arabic Corpus)

لقد قمنا في جامعة ليدز بتدشين موقع "الذخيرة القرآنية العربية" على الشبكة العنكبوتية وسرعان ما اكتسب شهرة عالمية و أصبح الموقع المفضل للراغبين في معرفة لغة القرآن الكريم والباحثين في المعالجة الحاسوبية لنص القرآن. ويرتاد الموقع أكثر من خمسين ألف زائر شهرياً.

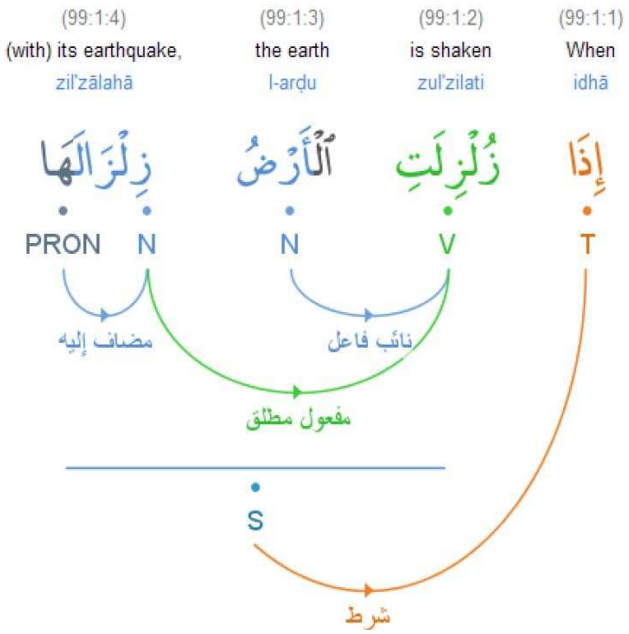
تعطي هذه الذخيرة تفاصيل صرفية ونحوية لكل كلمة في القرآن الكريم. وهذه التفاصيل مستقاة مبدئياً بشكل آلي من محلل باكوالتر الصرفي (Buckwalter Arabic Morphological Analyzer) ثم بعد ذلك تم إجراء تصحيح يدوي قبل نشره في الموقع[Dukes and Habash10]. وبعد ذلك سمح لمستخدمي الموقع بالتعليق وتصحيح الأخطاء إن وجد

وهكذا اكتسبت هذه الذخيرة درجة عالية من الصحة والتوثيق مع مرور الأيام. الشكل التالي نموذج للتحليل الصرفي والنحوي لكلمة قرآنية.



الشكل 11 : تفاصيل صرفية ونحوية لكلمة قرآنية ويلاحظ وجود تعليق للمستخدمين على هذه الكلمة

وكمرحلة ثانية بدأنا بتطوير تقنية الشجرات التبعية (dependency treebank) لتوضيح إعراب الآيات القرآنية، وهكذا قربنا الإعراب التقليدي للأذهان وجعلناها قابلة للمعالجة الحاسوبية [Dukes et al. 10]. الشكل التالي نموذج لإعراب آية قرآنية.



الشكل 12 : شجرة تبعية (dependency treebank) معدلة تضم إعراب آية قرآنية

تمتاز هذه الذخيرة بكونها مفتوحة المصدر وقابلة للتحميل المجاني، وبدأ يشكل مورد نصي مهم للعديد من التجارب اللغوية في أنحاء العالم، ولا زلنا نستقبل الكثير من التعليقات والتشجيع والإقتراحات من المستفيدين من هذه الذخيرة.

3.6 تطوير قاعدة بيانات للمفاهيم القرآنية العائدة من الضمائر

لقد لاحظنا أن الضمائر تستخدم بشكل كبير في القرآن الكريم بما يشكل نحو 30% من مجموع الكلمات. وكثيراً ما نجد أن القرآن يشير لمفاهيم بصيغة الضمائر وهذا يجعل الوصول لهذه الآيات عسيراً بمجرد البحث عن الكلمات. فمثلاً نقرأ الآية الكريمة (إنا أنزلناه في ليلة القدر) حيث أن الضمير يعود للقرآن الكريم وكذلك قوله تعالى (إنا أنزلناه في ليلة مباركة).

وبناءً على هذه الملاحظه بدأنا في مشروع لعَنْوَنة جميع الضمائر البارزة في القرآن ووسمها بأحد المفاهيم القرآنية مستعينين في ذلك بكتب التفاسير المعتمدة. وهذا المجهود لم يسبق إليه أحد حسب علمنا ويتوقع أن يكون لها استخدامات حاسوبية عديدة وتُعِين في استخلاص محتويات القرآن بشكل لا يمكن التوصل إليها بدون هذه الذخيرة. وقد انتهينا من نصف القرآن الكريم. والشكل التالي يعطي نموذج من آية قرآنية معنونة بالمفاهيم العائدة من الضمائر.

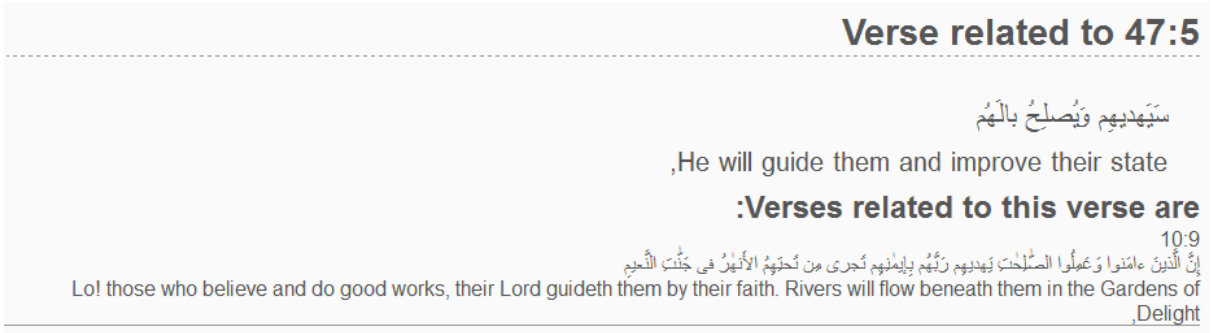
Antecedent #	Pronoun context	Verse	Concept	gloss
1	إِلَّا هُوَ ٱلْحَىُّ ٱلْقَيُّومُ نَزَّلَ عَلَيْكَ ٱلْكِتَٰبَ بِٱلْحَقِّ مُصَدِّقًا لِّمَا بَيْنَ يَدَيْهِ	3:3	محمد	Prophet Muhammad
2 ٱلْكِتَٰبَ	ٱلْكِتَٰبَ بِٱلْحَقِّ مُصَدِّقًا لِّمَا بَيْنَ يَدَيْهِ وَأَنزَلَ ٱلتَّوْرَىٰةَ وَٱلْإِنجِيلَ مِن قَبْلُ هُدًى	3:3	القرآن	the Qur'an

شكل 13 : نموذج لعنونة الآيات بالمفاهيم العائدة من الضمائر البارزة في القرآن

3.7 تطوير قاعدة بيانات للآيات المرتبطة

نلاحظ أن آيات القرآن تتشابه في الألفاظ والمعاني والقصص تتكرر والمعنى المبهم في آية يوضحها آية أخرى كما قال تعالى (الله الذي نزل أحسن الحديث كتاباً متشابهاً مثاني). وبناءً على ذلك أردنا تطوير قاعدة بيانات يمكن الوصول لجميع الآيات المترابطة

للآية المدخلة. وقد بدأنا كمرحلة أولية باستخلاص الآيات المترابطة من تفسير إبن كثير وطورنا برنامج شبه آلي لإستخلاص الآيات المترابطة من موقع مجمع الملك فهد لطباعة القرآن الكريم. والشكل التالي نموذج من الآيات المترابطة.



الشكل 14 : الآيات المترابطة من تفسير إبن كثير

وفي المستقبل يمكن استخلاص مجموعات مترابطة أخرى من تفاسير أخرى. ويمكن تطوير مثل هذه الذخيرة للآيات المترابطة بعدة لغات بناءً على التراجم القرآنية المتوفرة. ومثل هذه الذخيرة ستلعب دوراً رائداً في أبحاث المعالجة اللغوية التي تعنى باكتشاف الترابط آلياً. ونهدف في المستقبل أن نكتشف خصائص عدة يمكن من خلالها توظيف دوال التعلم الآلي في اكتشاف الإرتباطات آلياً من القرآن الكريم.

4. نحو إنشاء مشروع كبير للمحتوى التراثي للأديان

في شهر إبريل 2010 استضافت جامعة إدنبره مؤتمراً عالمياً لنقاش التحديات الحاسوبية المستقبلية (Grand Challenges in Computing Research for 2010 and beyond). وكان لنا مشاركة في هذا المؤتمر باقتراح كون "فهم القرآن والنصوص الدينية عن طريق الحاسوب" من التحديات القادمة لعلماء الحاسوب [Atwell et al 2010]. وكان هذا الأقتراح من طرفنا بسبب كون النصوص الدينية تتميز بقوة تأثيرها في الأفراد والمجتمعات. وكذلك كون النصوص الدينية التراثية لا سيما التوراة والإنجيل والقرآن تمثل وحدة موضوعية متشابهة ويمكن تطوير معايير موحدة وبرمجيات وأدوات تخدم جميع هذه النصوص.

وانطلاقاً من كوننا قد قمنا بنجاح بإجراء تجارب وأبحاث رائدة للمعالجة الحاسوبية للنصوص العربية والقرآن فإننا نأنس من أنفسنا المقدرة والكفاءة للمشاركة في مشروع كبير مدعوم من جهات معنية وبالمشاركة مع فرق بحثية أخرى. بإمكاننا في جامعة ليدز أن نضع خبرتنا الطويلة في المعالجة الحاسوبية على الذخائر اللغوية ونشرف على مشاريع منبثقة منها في تطوير تطبيقات وأدوات حول هذه الذخائر. ومن أجل تطوير تطبيقات ذات كفاءة وجودة عالية لا بد من إنشاء ذخائر معنونة بطبقات عدة من المعلومات اللغوية والمعرفية التي تخص طبيعة النص المرغوب مثل القرآن والحديث والتوراة والإنجيل وغيرها. وهذا يستدعي تكوين فرق بحثية منفصلة من خبراء في هذه النصوص ولكن يعمل الجميع تحت معايير موحدة تجمع الوحدة الموضوعية لهذه النصوص.

نتطلع في هذا المشروع الكبير تطوير أدوات وموارد لغوية مفيدة للغة العربية الفصحى تغطي محتوى القرآن وبعض كتب الحديث. وهذه الأدوات والموارد يمكن تمديدها لتغطي نصوص من اللغة العربية المعاصرة. ونعني بالموارد تلك المتوفرة في اللغة الإنجليزية مثل معجم WordNet لتصنيف العلاقات بين المفردات، وكذلك معجم FrameNet لتطوير إطارات دلالية وشبيه منها مشروع PropBank. كما أننا نستطيع تطوير شجرات بيانية (Treebank) للبنية النحوية والإعراب لهذه النصوص على غرار ما بدأناه في إعراب القرآن الكريم. وفي مرحلة متقدمة يمكن إنشاء تصنيف متكامل للمفاهيم (Conceptual

Ontology) من هذه النصوص. ومثل هذه الموارد تساعد في تكوين روابط علاقات بين النصوص المتناثرة وتلعب دوراً رائداً في تطبيقات التنقيب والتحليل النصي (Text Mining).

5. الخاتمة

عرضنا في هذه الورقة نظرة سريعة على المشاريع الحاسوبية في جامعة ليدز على اللغة العربية والقرآن الكريم. يتبين من خلال هذه الجولة أننا في ليدز قد اكتسبنا خبرة طويلة في العمل على اللغة العربية وكانت غالبية المشاريع ناتجة من أبحاث ماجستير ودكتوراة ونتجت عنها أدوات وذخائر لغوية ومواقع إلكترونية. وتتميز أبحاثنا بكونها مفتوحة المصدر ومتاحة مجاناً وهكذا نشجع أن يجرب الخبراء هذه الذخائر والأدوات وأن يوافونا بإقتراحات لتحسين هذه الموارد. وبما أن العمل على اللغة العربية يستلزم التعاون مع خبراء في اللغة في توثيق النتائج وتصحيح الأخطاء فإننا على تنسيق مع قسم اللغة العربية والدراسات الشرق أوسطية في جامعة ليدز بشكل مستمر.

إن تجربتنا الناجحة مع موقع الذخيرة القرآنية العربية جعلنا نتشجع بتكرار هذا النجاح وإعتماد نفس آلية العمل في المشاريع القادمة: وهي البداية بتوظيف الأدوات والموارد المتوفرة لتكوين المرحلة الأولى وتجنب البداية من الصفر. ثم تحسين المورد يدوياً أو شبه آلياً قدر المستطاع، ثم في مرحلة ثالثة وضع النتائج في موقع إنترنت وتشجيع التنسيق مع الخبراء والزوار في تحسين المورد بشكل مستمر.

إننا ندرك أهمية توظيف هذه الخبرة الطويلة في المشاركة الفعالة في مشاريع اللغة العربية في الوطن العربي. ولا شك أن معيار نجاح أبحاثنا على اللغة العربية هو قبولها واستفادتها من قبل مستخدمي اللغة العربية في العالم العربي وإلا يبقي ذخائرنا وأداتنا حبيسة الأبحاث الأكاديمية التقليدية وهذا ما لا نرغب فيه على الإطلاق.

المصادر

Abu Shawar, Bayan; Atwell, Eric. An Arabic chatbot giving answers from the Qur'an in: Bel, B & Marlien, I (editors) Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles, Volume 2, pp. 197-202 ATALA. 2004.

Al-Saif, A; Markert, K. 2010. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. (pdf) In: *Proc. of the conference on Language Resources and Evaluation. Malta, 2010.*

Al-Sulaiti, Latifa; Atwell, Eric. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171. 2006.

Eric Atwell, Kais Dukes, Abdul-Baquee Sharaf, Nizar Habash, et al.(2010) **Understanding the Quran: A new Grand Challenge for Computer Science and Artificial Intelligence.** Grand Challenges for Computing Research (2010). British Computer Society Workshop. Edinburgh

Fillmore, C. (1976). "Frame Semantics and the nature of language." Annals of the New York Academy of Science.

Ghazali, S. & Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL. Toulouse, France.

Kais Dukes and Nizar Habash. Morphological Annotation of Quranic Arabic. The seventh international conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta, 2010.

Kais Dukes, Eric Atwell and Abdul-Baquee M. Sharaf. Syntactic Annotation Guidelines for the Quranic Arabic Treebank. The seventh international conference on Language Resources and Evaluation (LREC-2010). Valletta, Malta, 2010

Kais Dukes and Tim Buckwalter. A Dependency Treebank of the Quran using Traditional Arabic Grammar. Submitted to the 7th international conference on Informatics and Systems. Cairo, Egypt, 2010

Mushaf at-Tajweed 1420H. – مصحف التجويد الملون – دار المعرفة الطبعة الرابعة

Roberts, Andrew; Al-Sulaiti, Latifa; Atwell, Eric. aConCorde: Towards an open-source, extendable concordancer for Arabic. Corpora journal, vol. 1, pp. 39-57. 2006.

Ruppenhofer, J., M. Ellsworth, M. Petruck, and C. Johnson (2005). "FrameNet: Theory and Practice.

Sawalha, Majdi and Atwell, Eric (2008). Comparative evaluation of Arabic language morphological analysers and stemmers. *Proceedings of COLING 2008 22nd International Conference on Computational Linguistics.*

Sawalha, Majdi and Atwell, Eric (2009a). Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguuistics Conference CL2009* Liverpool, UK.

Sawalha, Majdi and Atwell, Eric (2009b). عـد النحـو والصـرف فـي بنـاء محلـل صـرفي للغـةتوظيـف قـوا العربيـــة (Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy.* Damascus, Syria.

Sawalha, Majdi and Atwell, Eric (2010a). Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta.

Sawalha, Majdi and Atwell, Eric (2010b). Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta.

Sharaf, A. and Atwell, E. (2009) A Corpus-based computational model for knowledge representation of the Qur'an. 5th Corpus Linguistics Conference, Liverpool

van Mol, M. (2000b). Exploring annotated Arabic corpora: preliminary results. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/tunis.pdf)