Elhadj, Yahya O. Mohamed; Bouziane, Rachid; Abdelai, Ahmed; Ammar, Adel "Building a Statistical system for Quran' terms recognition: tool for Morpho-syntactic features tagging". International Computing Conference in Arabic (ICCA2013). Hammamat, Tunisia, Dec 09-11,2013.

إعداد وتجهيز نظام إحصائي للتعرف الآلي على المفردات القرآنية: الخصائص والسمات الصرف- نحوية وآلية مستحدثة لوسمها

يحيى محمد الحاج[1]، رشيد بوزيان[2]، أحمد عبدالعالي[3]، عادل عمار[4]
[1] مركز تطبيقات الحاسب في المجالات الشرعية والعربية، جامعة الإمام، السعودية، yelhadj@ariscom.org
[2] قسم اللغة العربية، كلية الآداب والعلوم، جامعة قطر، قطر، rachid.bouziane@qu.edu.qa
[3] معهد قطر لبحوث الحوسبة، مؤسسة قطر، قطر، aabdelali@qf.org.qa
[4] قسم علوم الحاسب، كلية علوم الحاسب والمعلومات، جامعة الإمام، السعودية، adel.ammar@gmail.com

ملخص:

تقدم هذه الورقة جزءاً من نتائج مشروع طموح يسعى إلى تحديد الخصائص اللغوية للمفردات القرآنية آليا. وقد ركزت الورقة على تقديم البنية التحتية اللازمة لبناء النماذج الإحصائية للتعرف الآلي على الخصائص، حيث تم استخلاص أهم الخصائص والسمات الصرف - نحوية التي يمكن استخدامها في التعرف الآلي على المفردات القرآنية ، ثم توصيف الخصائص والسمات وترميزها بشكل يناسب التعاطي معها آليا، ثم إعداد بيئة حاسوبية تمكن الخبير اللغوي من إعداد عينة يدوية من التحليل القرآني لتستخدم في تدريب واختبار النماذج الإحصائية التي نسعى إلى بنائها وتطويرها لتصبح قابلة للاستخدام في تطبيقات لا تقتصر على خدمة النص القرآني وإنما تتعداه لتشمل اللغة العربية بشكل عام حيث هي لغة القرآن.

كلمات جوهرية: التعرف الآلي على الخصائص اللغوية، النماذج الإحصائية/الاحتمالية، التحليل الآلي للنص القرآني ، الذخائر اللغوية.

1. مقدمة

لقد تزايد الاهتمام في السنوات الأخيرة بما أصبح يعرف ب "الحوسبة في المجالات الشرعية" و المعالجة الآلية لمتون النصوص المنتمية إلى هذه المجالات ، وهي الجهود البحثية التي تسعى إلى استثمار التقدم الكبير الذي يشهده مجال تكنولوجيا " هندسة اللغة " ، في خدمة " النص " القرآني و متون الحديث النبوي الشريف و غيرهما من النصوص و الذخائر اللغوية ذات الصلة . هذه الورقة يرجو لها أصحابها أن تكون مساهمة في تحديث آليات التعاطي مع تحديات المعالجة الآلية للغة العربية بصفة عامة ، و كذا دعوة لتجريب مقترحات و تصورات جديدة في تشريح البنية اللسانية للنص المكتوب باللغة العربية و في المعالجة الحاسوبية لمخرجات هذا التشريح. وعلى الرغم من ظهور عدد لا بأس به من الأعمال الحاسوبية في القرآن وعلومه، وفي الحديث وطرق تخريجه واستنباطه، وفي الفقه والتشريع الإسلامي بصفة عامة، إلا أن الأعمال الحاسوبية المتعلقة بالتحليل الآلي لمفردات النص القرآني وتراكيبه، وغيرها مما عُنِي ببعض جوانب الإحصاء أو التحليل الصرفي لمفردات القرآن، لا يزال يشوبه قصور ونقص في بعض الجوانب. ويكمن ذلك في عدم الاستقصاء التام لمفردات النص القرآني كالأسماء والحروف والأفعال، أو عدمِ التصنيف الدقيق لها.

وللمساهمة في رفع مستوى حوسبة النص القرآني قمنا في مشروع سابق ، للتعلم الآلي للقرآن الكريم، ممول من مدينة الملك عبدالعزيز للعلوم والتقنية بمنحة رقم "أت-25-113" ، ببناء ذخيرة لغوية للقرآن الكريم ، تم فيها حصر مفردات النص القرآني حصرا شاملا، مع تبيان الجذر الأصلي لكل كلمة في القرآن، ومذكر كل مؤنث فيه، والتفريق بين الصيغ المتشابهة اعتمادا على المعنى، وماضي كل فعل، ومفرد كل جمع، وبناء كل فعل للمعلوم، وأصل كل كلمة حصل فيها إعلال أو إبدال، بالحذف أو الزيادة أو القلب أو الإدغام أو النقل. كما بُيِّنَت اللواحق والسوابق الزائدة على كل صيغة مجردة أو غير مجردة، وبُيِّن البسيط والمركب من الأدوات والضمائر وغيرها من صنوف الكلمات. وقد وضعت تلك الذخيرة في قاعدة بيانات متكاملة، ضمت نسخة نصية من النص القرآني بخط إملائي مشكلة تشكيلا كاملا مع تحليل صرفي لكل كلماته، حيث فككت كل كلمة و قُطِّعت إلى سوابق وجذع وجذر ثم لواحق. وتم الاحتفاظ بالكلمة في سياقها الطبيعي (الآيات) ضمن النص القرآني بصيغتين، إحداهما بالتشكيل والأخرى بدونه للاستفادة من ذلك عند الحاجة [1-3]. إلا أن الجزء الهام الذي لا يزال ينقص هذه الذخيرة هو تحديد وإضافة الخصائص اللغوية (Part of Speech Tagging) للمفردات القرآنية. ولاستكمال هذا التوجه فإننا نعمل في الوقت الراهن ، بدعم من مدينة الملك عبد العزيز للعلوم و التقنية ((منحة رقم "أت-30-199") ، على مشروع جديد يستهدف تحديد الخصائص اللغوية للألفاظ القرآنية عبر تحليل لغوي أكثر عمقا و تفصيلا ، و ذلك على محورَي المتن القرآني الرئيسين "المفردات" و"الجمل" [4]. وسيوفر ذلك مادة لغوية لجميع الدارسين في مجال اللغة العربية بكل تخصصاتها المتفرعة. كما أنه سيمهد لدراسات صرفية ونحوية عميقة ومتطورة للقرآن الكريم. وعلاوة على ذلك، سيكون بالإمكان استخدام هذه الذخيرة بشكل فعال لتدريب نماذج إحصائية لبناء تطبيقات حاسوبية تستهدف النص القرآني أصالة ، كما تستهدف اللغة العربية عموما بالتَّبَعِ و اللزوم .

وفي أولى خطوات هذا المشروع قمنا بدراسة لغوية لحصر أهم الخصائص والسمات الصرف-نحوية التي ينبغي الاعتداد بها عند التعاطي مع الجوانب اللغوية للمفردات القرآنية. وقد تم تقديمها في ورقة علمية لمؤتمر "جامعة طيبة لتوظيف تقنية المعلومات لخدمة النص القرآني وعلومه" الذي سيقام خلال الفترة 22-25 دجمبر 2013م[1]. وفي هذه الورقة سنقدم تفاصيل الخصائص والسمات التي تم استخلاصها من الدراسة اللغوية للمفردات القرآنية، ثم نستعرض نظاما ترميزيا مفصلا تم وضعه للتعاطي معها، ومن ثم نقدم بيئة حاسوبية تم إعدادها لتهييئ عينة من النص القرآني بشكل يدوي عبر خبراء لغوين ليتم استخدامها في تدريب واختبار نماذج إحصائية تكون قادرة على التعرف بشكل آلي على الخصائص والسمات المطلوبة.

2. النماذج الاحتمالية في مجال التعرف الآلي على الخصائص اللغوية للمفردات

إن استخدام مبدإ الاحتمالات بين التراكيب اللغوية في التعرف الآلي على خصائص المفردات يعود إلى الستينات من القرن الميلادي الماضي [5]. وقد تم بناء أول متعرف آلي شبه متكامل على خصائص المفردات [6] معتمد على طريقة إحصائية تستخدم خوارزمية Viterbi [7] في منتصف السبعينيات. ومن ثم توالت استخدامات الطرق الإحصائية بنماذج احتمالية متعددة.

ومن أبرز النماذج الاحتمالية المستخدمة في مجال التعرف على الخصائص اللغوية، نماذج ماركوف الخفية (Hidden Markov Models - HMMs) التي طبقت على عدة لغات من بينها العربية وأنتجت نتائج جيدة [8، 9]. وهذه النماذج تمتاز بكونها تتيح إمكانية تحديد التتابعات المنطقية للمفردات اللغوية اعتمادا على البنية التركيبية للجملة، إلا أن هذا التوجه لم يستغل بعد بالشكل الكافي في اللغة العربية. وإضافة إلى نماذج ماركوف الخفية، فإن النماذج المبنية على الشبكات العصبية (Artificial Neural Networks - ANN) أثبتت هي الأخرى نجاحها في لغات مختلفة [10، 11]. وبعض أنواع هذه الشبكات (Recurrent Neural Networks) يتيح إمكانية الاعتماد على القائمة التسلسلية، حيث طبقت في بعض الأبحاث الحديثة على مفردات اللغة الإنجليزية [12]، لكن لم يسبق استعمالها في اللغة العربية في حدود علمنا. وسنبين في ما يلي الخلفية العلمية لكل من هذين التوجهين وكيفية التعاطي معهما تمهيدا لما نسعى للقيام به في هذا العمل.

2.1. نماذج ماركوف الخفية في مجال التعرف على الخصائص اللغوية

يعتمد مبدأ استخدام نماذج ماركوف الخفية (HMMs) [13] ،في التعرف الآلي على خصائص المفردات ،على تحديد جملة من الرموز (tags) تمثل السمات اللغوية التي يراد التعرف عليها، ومن ثم تستخدم هذه الرموز لتكوين حالات أو عقد النموذج (HMM states). يتم تحديد روابط الانتقال بين حالات النموذج (HMM topology) من خلال التتابعات اللغوية الصحيحة بين المفردات اللغوية. ويتم ذلك عبر حساب احتمالات التوالي بين المفردات في نص مجهز سلفا وموسوم بالسمات المراد التعامل معها، ويسمى بذخيرة التدريب. ويمكن الاطلاع على تفاصيل استخدام نماذج مركوف الخفية في مجال تحديد الخصائص اللغوية بشكل عام في البحوث [14، 15]. وفي ما يلي تلخيص لأهم تلك الجوانب للتسهيل على القارئ.

2.1.1. بنية النموذج

عند بناء نماذج ماركوف الخفية، لا بد من تحديد أربعة أشياء أساسية:

أ. **<u>حالات النموذج (model states: S):</u>** وهي مجموعة العقد التي تشكل البنية الأساسية للنموذج، ويتم الانتقال بينها وفقا لمبدإ الاحتمالات. وفي هذا العمل، ستمثل الحالات الرموز اللغوية (tags) التي ستوضع لتمثيل السمات المراد التعرف عليها وذلك بناءا على تصنيف المفردات العربية. وسيكون عدد الحالات في النموذج نفس عدد الرموز اللغوية، بالإضافة إلى حالة ابتدائية وأخرى نهائية لتحديد نقطة الدخول في النموذج والخروج منه. وإذا رمزنا للحالة ب s_i، ولعدد الحالات بـ M، فان حالات النظام تكتب على الشكل $S = (s_i)_{0 \le i \le M+1}$.

ب. **<u>الاحتمالات بين الحالات (transition probabilities: A):</u>** وهي مصفوفة تضم قيما عددية تمثل احتمالات الانتقال الممكنة والمسموح بها بين حالات النموذج. وفي عملنا سيتم حسابها من خلال بنية الجملة العربية التي تحدد الشكل العام للنموذج (انظر الجزء الثاني من التقرير: منهج العمل). وإذا رمزنا لاحتمال الانتقال بين الحالة t_i إلى الحالة t_j بالرمز $P(t_j \mid t_i)$ حيث $a_{ij} = P(t_j \mid t_i)$ ، فإن مصفوفة الاحتمالات يمكن أن تكتب على الشكل $A = (a_{ij})_{1 \le i,j \le M}$. ولحساب هذه القيم لابد من وجود ذخيرة لغوية يتم ترميزها مسبقا في سياقها اللغوي.

[1] http://www.taibahu.edu.sa/pages.aspx?pid=11437

ت. **الملاحظات(observations):** ونعني بها الأشياء التي نحاول التعرف عليها. وفي هذا العمل ستمثل هذه الملاحظات المفردات (يرمز لها ب: w_i)، كما سيتم تصنيفها في الجانب اللغوي (كلمات محللة صرفيا)، التي نريد تصنيفها أو التعرف على خواصها من خلال تحديد الرمز المناسب لكل منها. وسنرمز للملاحظات بالرمز $O=(o_i)_{1\le i\le N}$، حيث $o_i = w_i$ وN تمثل طول السلسلة المدخلة.

ث. **احتمالات الملاحظة (observation probabilities):** وهي قوانين احتمالية (منفصلة discrete أو متصلة continous) من خلالها يتم التعرف على الأشياء المراد تحديدها عند المرور بكل حالةفي النموذج. وفي هذا العمل سنستخدم احتمالات منفصلة في كل حالة من حالات النموذج تمثل احتمال أن تكون للكلمة المدخلة الرمز الموجود بالحالة. وإذا استخدمنا التعاريف السابقة، فإن القيمة الاحتمالية تكتب على الشكل $P(w_i \mid t_i)$ حيث $b_i(w_i)=P(w_i \mid t_i)$، وعليه يمكن كتابة القوانين الاحتمالية المرفقة بحالات النظام على الشكل $B=(b_i(.))_{1\le i\le M}$. وهنا نحتاج كذلك إلى ذخيرة لغوية يتم ترميزها سلفا لإجراء إحصائيات عليها لتحديد هذه القيم.

2.1.2. التعرف على الخصائص أو السمات: سلسلة الرموز الأكثر احتمالا

تمثل هذه الخطوة مرحلة التعرف الآلي على الخصائص. إذا رمزنا للنص المدخل (سلسلة من الكلمات محللة صرفيا) ب $W=(w_i)_{1\le i\le n}$، فان السؤال يصبح: كيف نحدد سلسلة الرموز اللغوية $T=(t_i)_{1\le i\le n}$ التي تحدد أفضل تصنيف لقائمة الكلمات المدخلة ؟ أي أننا نبحث عن حل للمعادلة الرياضية: $\max_T[P(T \mid W)]$. ولحل هذه المعادلة لا بد أن نقوم أولا بتحويلها إلى الصيغة التالية باستخدام قانون بايس الاحتمالي: $P(T \mid W)=\frac{P(W \mid T)*P(T)}{P(W)}$. وبما أن $P(W)$ هي قيمة ثابتة مهما غيرنا سلسلة الرموز اللغوية T، فإن المعادلة التي نريد حلها تصبح: $\max_T[P(W \mid T)*P(T)]$، وهي تضم جزئين هامين. يمثل الجزء الأول $P(T)$ احتمال تتابع الرموز في السلسلة T، ويمكن حسابه من خلال مصفوفة احتمالات الانتقال بين حالات النموذج ($A=(a_{ij})_{1\le i,j\le M}$) كما وصفناها سابقا. أما الجزء الثاني $P(W \mid T)$ فيمثل احتمال مطابقة النص المدخل W مع سلسلة الرموز T، وهو ما يمكن حسابه عن طريق القوانين الاحتمالية المرفقة بحالات النظام ($B=(b_i(.))_{1\le i\le M}$) كما وصفناها أيضا سابقا.

أ. **حساب $P(T)$:** تمثل $P(T)$ احتمال تتابع الرموز اللغوية في السلسلة T، ويمكن حسابه عن طريق المعادلة $P(T=t_1t_2\cdots t_n)=P(t_1)*P(t_2 \mid t_1)*\ldots*P(t_n \mid t_1\ldots t_{n-1})$. ولتسهيل حساب هذه القيمة عمليا، يتم استخدام النموذج اللغوي n-gram بحيث يكتفى باحتساب آخر n-1 رمزا في السلسلة لتحديد الرمز الموالي. يتم تدريب هذا النموذج على ذخيرة لغوية تضم تراكيب لغوية صحيحة. ونشير إلى أنه كلما قل عدد الرموز المحتسب في السلسلة ، فقدنا السياق، لكن تحديد القيم الاحتمالية يصبح أسهل. وفي المقابل كلما زاد عدد الرموز زادت معه نسبة الاحتفاظ بالسياق ، لكن حساب الاحتمالات يكون أصعب ، حيث إن ذخيرة التدريب قد لا تضم كل التراكيب اللغوية الممكنة. من هذا المنطلق،يتم الاكتفاء بقيم وسطية للنموذج، حيث عادة ما يستخدم نموذج " الحلقات ثنائية التكوين " (=bi-gram) أو نموذج " الحلقات ثلاثية التكوين " (=tri-gram) فإذا استخدمنا مثلا نموذج " الحلقات ثلاثية التكوين "(tri-gram)، يمكن كتابة المعادلة السابقة على الشكل $P(T=t_1t_2\cdots t_n)=\prod_{i=1}^{n}P(t_i \mid t_{i-2}t_{i-1})$، حيث إن t_{-1} وt_0 هي رموز افتراضية توضع في بداية الجملة لتمكين حساب الاحتمال لأول كلمة في الجملة. وهذه الرموز يمكن أن تضاف في القاموس الترميزي للكلمات. ولحساب القيمة $P(t_i \mid t_{i-2}t_{i-1})$، فإننا نحتاج إلى القيام بإحصائيات على الذخيرة اللغوية الخاصة بالتدريب والمرمزة مسبقا، أي أننا نقوم بإحصائيات لتحديد عدد مرات ورود التركيبة الثلاثية $t_{i-2}t_{i-1}t_i$ (سنرمز لها ب: $f(t_{i-2}t_{i-1}t_i)$) والتركيبة الثنائية $t_{i-2}t_{i-1}$ (سنرمز لها ب: $f(t_{i-2}t_{i-1})$) حسب السياقات اللغوية المختلفة. ويتم حساب القيمة من خلال المعادلة $P(t_i \mid t_{i-2}t_{i-1})=\frac{f(t_{i-2}t_{i-1}t_i)}{f(t_{i-2}t_{i-1})}$. والمشكلة هنا تكمن في كون ذخيرة التدريب قد لا تضم كل التراكيب اللغوية الممكنة، حيث يمكن أن تأتي في مرحلة اختبار جملة تضم سلسلة من الكلمات لم تظهر سلفا في

مرحلة التدريب، وعليه ستكون القيمة في المعادلة السابقة صفرا، مما يؤدي إلى تجاهل هذا التركيب!!. ولحل هذه المشكلة، يتم عادة استخدام ما يعرف بتقنيات التمليس (smoothing) والتي تدمج بين النموذج tri-gram والنماذج الأقل منه درجة، bi-gram وuni-gram [16]، من خلال التركيبة التالية: $\lambda_1 * P(t_i \mid t_{i-2}t_{i-1}) + \lambda_2 * P(t_i \mid t_{i-1}) + \lambda_3 * P(t_i)$ ، حيث إن $\lambda_1 + \lambda_2 + \lambda_3 = 1$. وهناك عدة طرق لحساب القيم $\lambda_3, \lambda_2, \lambda_1$ ، لعل أكثرها استخداما الخوارزمية المبينة في الشكل 1 [17]. ترمز الخوارزمية بـ N لعدد الكلمات في الذخيرة، وتعتبر أن القيمة في أي من الحالات الثلاث مساوية للصفر إذا كانت قيمة البسط تساوي الصفر.

```
set λ1 = λ2 = λ3 = 0
foreach trigram t1, t2, t3 with f(t1, t2, t3) > 0
    depending on the maximum of the following three values:
        case (f(t1,t2,t3)−1)/(f(t1,t2)−1):  increment λ3 by f(t1, t2, t3)
        case (f(t2,t3)−1)/(f(t2)−1):  increment λ2 by f(t1, t2, t3)
        case (f(t3)−1)/(N−1):  increment λ1 by f(t1, t2, t3)
    end
end
normalize λ1, λ2, λ3
```

الشكل 1: خوارزمية لحساب القيم اللازمة لدمج النماذج حسب [17].

ب. **حساب $P(W \mid T)$:** تمثل $P(W \mid T)$ احتمال مطابقة النص المدخل W مع سلسلة الرموز T ويمكن حسابها ببساطة من خلال حساب احتمال أن تأخذ كل كلمة في النص الرمز المقابل في السلسلةكماتبينه المعادلة $P(W \mid T) = \prod_{i=1}^{n} P(w_i \mid t_i)$. ولحساب القيمة $P(w_i \mid t_i)$ ، فإننا نحتاج إلى القيام بإحصائيات على الذخيرة اللغوية الخاصة بالتدريب والمرمزة مسبقا، أي أننا نقوم بإحصائيات لعدد المرات التي رمزت فيها الكلمة w_i بالرمز t_i (سنرمز له ب: $f(w_i, t_i)$) حسب السياقات المختلفة وعدد المرات التي استخدم فيها الرمز t_i مع أي كلمة في الذخيرة (سنرمز له ب: $f(t_i)$). ويتم حسب القيمة من خلال المعادلة $P(w_i \mid t_i) = \frac{f(w_i, t_i)}{f(t_i)}$. هنا كذلك يمكن أن تظهر مشكلة في مرحلة الاختبار عندما نصادف كلمة جديدة لم تظهر في ذخيرة التدريب على غرار ما حصل في الجزء السابق من غياب بعض التراكيب اللغوية. هناك عدة حلول لهذه المشكلة، أبسطها يعتمد على اعتبار الكلمة التي لم ترد في ذخيرة التدريب كلمة حصل فيها غموض، وبالتالي لها نفس الاحتمالات على كل الرموز [18] ويترك للسياق تصحيح الوضع لا حقا. هناك حلول أخرى أكثر دقة و تفصيلا ، تعتمد على السوابق واللواحق للكلمة لتحديد نوعها العام وبالتالي حساب احتمالها من خلال هذا النوع [17].

ت. **حساب $\max_T [P(W \mid T) * P(T)]$:** إن إيجاد سلسلة الرموز اللغوية $T = (t_i)_{1 \le i \le n}$ التي تحدد أفضل تصنيف لقائمة الكلمات المدخلة تحتاج إلى خوارزمية فعالة للتعاطي مع عملية البحث في كل التراكيب الممكنة. ولعل خوارزمية "فيتربي " (Viterbi) التي تعتمد على البرمجة الديناميكية (dynamic programming) هي الأكثر استخداما في هذا المجال عندما يتم التعاطي مع نماذج مركوف الخفية. وهذه الخوارزمية طبقت لأول مرة في مجال التعرف الآلي على الكلام في نهاية الستينات [7]، ومن ثم توالت استخداماتها في عدة مجالات أخرى. ولتطبيق خوارزمية "فيتربي " (Viterbi) ، نحتاج إلى مصفوفة الاحتمالات بين حالات النموذج والقوانين الاحتمالية المرفقة بالحالات، وهو ما تم توضيح كيفية حسابه في الأجزاء السابقة.

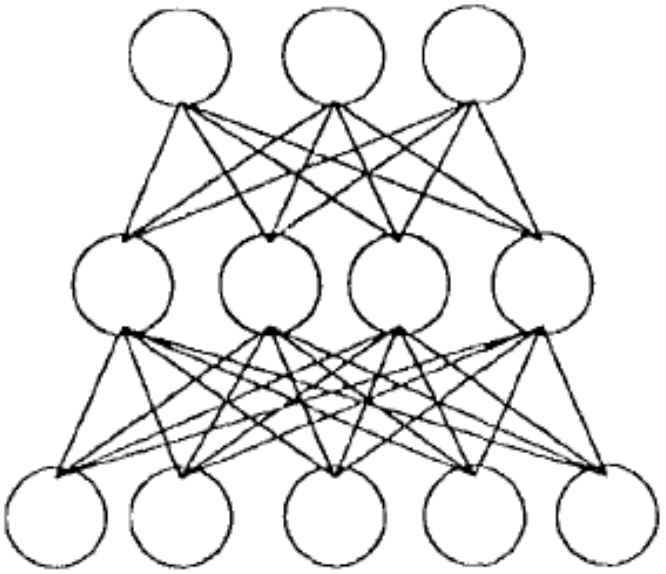
2.2. الشبكات العصبية في مجال التعرف على الخصائص اللغوية

ظهر أول استخدام للشبكات العصبية الاصطناعية في مجال التعرف على الخصائص اللغوية في أواسط التسعينات [10] مطبقا على اللغة الإنجليزية، ثم توالى استخدامها في لغات أخرى منها البرتغالية [19] والصينية [20] والهندية [21]، حيث أثبتت نجاحها. لكن استخدامها في اللغة العربية لا يزال محدودا ولم يجرب إلا في نطاق محدود [22]. وسنصف هنا نموذجا شبيها بالذي طبقه [10] على اللغة الإنجليزية.

2.2.1. بنية النموذج

تتكون الشبكات العصبية الاصطناعية من عدد كبير من الوحدات الحسابية البسيطة تسمى العصبونات (neurons). وهذه الوحدات بينها روابط مباشرة، ولكل رابط بين وحدتين i و j وزن W_{ij}. ولكل وحدة دالة تفعيل.

في الشبكات العصبية التي من نوع MLP (Multi-Layer Perceptrons) وهي الأكثر استعمالا، تكون الوحدات منظمة في شكل طبقات متتالية (انظر الشكل 2): طبقة المدخلات، طبقة المخرجات، وبينهما طبقة واحدة أو أكثر تسمى طبقات مخفية. ولا تكون الروابط إلا بين وحدات تقع في طبقات متجاورة. وينتقل التفعيل من وحدة إلى أخرى عن طريق الروابط لينتقل من طبقات المدخلات إلى طبقة المخرجات عبر الطبقات المخفية.



الشكل 2 : مثال لشبكة عصبية من نوع MLP

كل وحدة j تجمع مخرجات وحدات الطبقة السابقة مضروبة في أوزان الروابط $(a_{i*}w_{ij})$ وتضيف قيمة ثابتة b_j على النحو التالي:

$$u_j = \sum_i a_i w_{ij} + b_j$$

ثم تمر هذه القيمة عبر دالة تفعيل العصبون (والدالة المستخدمة غالبا هي من نوع sigmoid) ليصير مخرج الوحدة j كالآتي:

$$a_j = \frac{1}{1+e^{-u_j}}$$

تتعلم الشبكة العصبية من خلال تعديل أوزان الروابط بين الوحدات، حتى تحصل على المخرج المطلوب. وقيمة التعديل Δw_{ij} تكون على حسب المعادلة التالية:

$$\Delta w_{ij} = \eta . a_{pi} \delta_{pj}$$

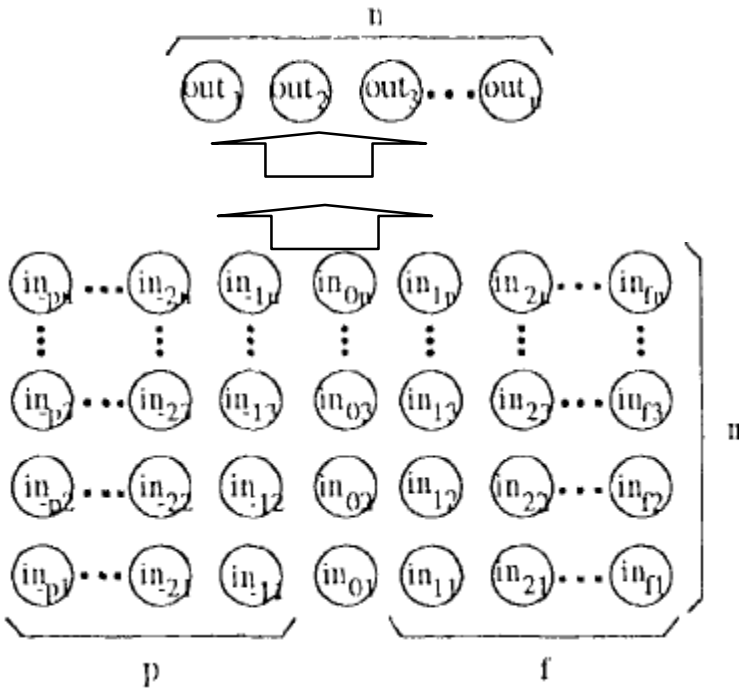
حيث: $\delta_{pj} = a_{pj}(1-a_{pj})(t_{pj}-a_{pj})$ إذا كانت الوحدة j من طبقة المخرجات

و: $\delta_{pj} = a_{pj}(1-a_{pj})\sum_k \delta_{pk} w_{jk}$ إذا كانت الوحدة j من طبقة مخفية

أما t_p فهو المتجه الصحيح الذي ينبغي أن يكون في المخرج، وهو الذي تتدرب عليه الشبكة.

2.2.2. كيفية استخدام نموذج الشبكات العصبية للتعرف على الخصائص اللغوية

للتعرف على الخصائص اللغوية نحتاج شبكة عصبية من نوع MLP (انظر الشكل 3) وذخيرة محللة يدويا لتكون قاعدة للتعلم.



الشكل 3 : شبكة عصبية من نوع MLP للتعرف على الخصائص اللغوية للكلمات (مستفادة من [10])

تتكون طبقة المخرجات في هذه الشبكة من وحدات تمثل كل واحدة منها سمة من السمات اللغوية الممكنة (tagset). وخلال التدريب على ذخيرة رُمزت يدويا، تتعلم الشبكة تفعيل الوحدة التي تمثل السمة الصحيحة، وعدم تفعيل الوحدات الأخرى. وهكذا، بعد عملية التدريب، تدلنا الوحدة التي لها أكبر قيمة من بين المخرجات على السمة التي ينبغي أن تُسند إلى الكلمةَ المستهدفة بالتحليل.

وتتكون طبقة المدخلات من جميع الخصائص اللغوية المتاحة عن الكلمة الحالية، وعن عدد p من الكلمات السابقة وعدد f من الكلمات اللاحقة. فيكون لكل سمة من الرموز اللغوية pos_j ولكل كلمة من كلمات السياق (وعددها p+f+1)، وحدة in_{ij} في طبقة المدخلات، تمثل قيمتها احتمال كون الكلمة i تحمل السمة j:

$in_{ij}=P(pos_j \mid word_i)$ ، إن كان $i\geq 0$

حيث إن هذه هي المعلومة الوحيدة التي نعلمها عن الكلمة الجاري ترميزها والكلمات اللاحقة. وتحسب هذه الاحتمالات اعتمادا على الذخيرة من خلال المعادلة:

$$P(pos_j \mid word_j) = \frac{f(pos_j, word_j)}{f(word_j)}$$

أما الكلمات السابقة فإننا يمكن أن نستخدم نتيجة الترميز الذي أعطتنا مخرجات الشبكة فيكون:

$in_{ij}=out_j(t+i)$ ، إن كان $i<0$

لكن نسخ قيمة مخرجات الشبكة وإعادتها إلى المدخلات (recurrence) يعقد عملية التدريب. لأن المخرجات لا تكون صحيحة في بداية التعلم. لذلك لا نستخدم المخرج نفسه، ولكن نستخدم متوسطا وزنيا بين مخرج الشبكة والمخرج الصحيح (الموجود في قاعدة التعلم). في بداية عملية التدريب يكون وزن المخرج الصحيح كبيرا ثم يتناقص تدريجيا إلى أن يصل إلى الصفر في نهاية التدريب.

بعد انتهاء عملية التدريب، تستخدم الشبكة لترميز كلمات جديدة على النحو التالي:

أ. تنسخ احتمالات سمات الكلمة الحالية واحتمالات سمات الكلمات اللاحقة، مع نتائج ترميز الكلمات السابقة إلى مدخلات الشبكة.

ب. ينتشر تفعيل الوحدات عبر الطبقة المخفية ليصل إلى المخرجات.

ت. المخرج الذي تكون قيمته أكبر هو الذي يمثل السمة التي ستُسند إلى الكلمة الجديدة.

ث. إن كان هناك مخرج ثان قيمته قريبة من قيمة المخرج الأكبر فإنه قد يقدّم باعتباره حلّا ثانيا للترميز .

3. الأعمال البحثية في مجال التعرف الآلي على المفردات اللغوية

سنقتصر هنا على ما وقفنا عليه من أبحاث في مجال التحليل الآلي للنص القرآني على وجه الخصوص، واللغة العربية عموما بصفتها الوعاء اللغوي الحامل لهذا النص ، وكذلك الأبحاث التي يمكن تطبيق نتائجها على النص القرآني وإن لم يعتمد أصحابها هذا النص مضمارا تطبيقيا رئيسا . ومن الجديد بالذكر أن هذه الأعمال في مجملها تُعدُّ قليلة جدا إذا ما قورنت بما تم في لغات أخرى كالإنكليزية مثلا.

إن أول تقدم ملحوظ في اتجاه إنشاء نظام للتحليل الآلي لخصائص المفردات العربية هو الخطوة التي قدمها الباحث أبو ليل وزميله إفنس [23] كمرحلة أولى لتحليل النصوص الصحفية. وكذلك المحلل شبه الآلي الذي قمه الباحثان الكاره والأنصاري [24]. أما أول محلل شبه حقيقي فهو الذي قدمته الباحثة شرين [25]، وإن كان لا يزال يحتاج إلى تطوير. ثم ظهرت بعد ذلك محللات أخرى، منها ما اشتملت عليه الأبحاث [26، 27، 28، 29، 30].

وقد بُنيت أغلب المحللات العربية الموجودة بهدف تحليل نصوص غير مشكولة، باستخدام ذخيرة لغوية محللة يدويا سابقا، وتحتوي على جميع الأقسام الصرفية الممكنة. لكن – خلافا للغة الإنجليزية –فإن اللغة العربية لا تزال تنقصها ذخيرة ذات حجم كبير مجهزة يدويا، يمكن أن يستخلص منها قواعد تعلم كبيرة. ولأجل التعاطي مع هذا الموضوع، فقد قام البارد وزملاءه [31] ببحث يهدف إلى إيجاد الخطة المثلى لإنشاء محلل إحصائي، يجمع بين الدقة والسرعة، في حال ندرة الذخائر التي يمكن استخدامها للتعلم الآلي. وهذه نقطة أساسية عند التعامل مع اللغة العربية التي تندر فيها النصوص المحللة أو لا تكون متاحة للاستخدام. وقد درس الباحثون في عملهم هذا استعمالات مختلفة لمحلل يعتمد على نماذج ماركوف المخفية: ثنائية (bigram) أو ثلاثية الكلمات (trigram)، بالإضافة إلى دراسة تقنيات " تشذيب " (smoothing) مختلفة. وقد قاموا بعدة تجارب باستخدام قاعدتي تعلم - إحداهما هي الذخيرة القرآنية (Quranic Arabic Corpus)[2] - لقياس فاعلية كل واحدة من هذه المحللات.

وكذلك مراعاة لحال ندرة الذخائر، قام القريني وزملاءه [32] ببناء محلل صرفي مبني على وزن الكلمات ولا يعتمد على أي ذخيرة محللة يدويا، لكن بشرط أن تكون الكلمات التي يتعامل معها مشكولة جزئيا (حركة آخر الكلمة فقط). وقد قاموا بإنشاء خوارزمية جديدة لإيجاد التحليل المناسب لكل كلمة اعتمادا على قاعدة " بطاقات وسمية " (tags) أعدوها مسبقا. وقد حصلوا على نسبة نجاح في التحليل تصل إلى 91% عند تجربتها على قاعدة بيانات تحتوي على 5000 كلمة (فعل أو اسم) مستخرجة من مقررات مدرسية. وأغلب الأخطاء كانت ناتجة عن أسماء الأعلام والأسماء المعربة. إلا أن هذه الدراسة لم تطبق على النص القرآني، رغم أنها قد تكون مناسبة له من حيث كونه نصا مشكولا. وقد نشر القريني وعياش [33] قاعدة " البطاقات الوسمية " (tags) التي استخدموها لكل الحركات الممكنة في آخر الكلمة.

ومن التوجهات البحثية التي اهتمت بالتحليل الصرفي للقرآن الكريم، وإن كانت غير خاصة بنصوصه، ما نشره الباحثان صوالحة واتوال [34]. فقد قام هذا الفريق بمسح للمحللات الصرفية العربية الموجودة، وقارنوا نتائج أربعة منها على نموذج من نصوص المتن القرآني . ثم قاموا بتوصيف المحلل الصرفي الذي أنشؤوه والذي يتميز بدقة أقسامه الصرفية (Fine-Grain) باستخدام ذخيرة تجمع بين 23 معجما عربيا. وقد جربوا هذا المحلل على الذخيرة العربية على الشبكة (Web Arabic Corpus) والتي تحتوي على 100 مليون كلمة[3]. واقترحوا كذلك معيارا لمقارنة نتائج المحللات الصرفية/النحوية.

ومن بين الأعمال البارزة التي تخص النص القرآني، نذكر ما قام به فريق بحثي في جامعة حيفا من بناء نظام حاسوبي لتحليل النص القرآني صرفيا وتحديد خصائص مفرداته بشكل آلي، للأغراض البحثية والتعليمية [35، 36]. وقد عمد الباحثون إلى استخدام تقنيات ألـ FST (Finite State Automaton) لتمثيل الظواهر الفونولوجية والنحوية في النص القرآني . واستخدمت هذه الأدوات مع النص القرآني للحصول على تحليل لكلماته بشكل آلي. وتم تخزين نتائج التحليل هذه في قاعدة بيانات لتستخدم عبر واجهة بيانات تمكن من الاستفسار عن بعض المعلومات اللغوية ذات الصلة. وقد قام الباحثون أولا بتحويل النص القرآني من كتابته العادية إلى كتابة "فونيمية" باستخدام الحروف اللاتينية للتخفيف - حسب ما يذكرون- من الغموض الذي يحصل في بعض الكلمات، وليتمكنوا كذلك من تخزينه بصيغة ASCII. ثم قاموا بتقسيم الألفاظ القرآنية إلى ثلاثة أقسام: قسم للكلمات ويضم الضمائر والحروف والكلمات الوظيفية وما إلى ذلك، أما القسمان الآخران فيخصان الصيغ الاسمية والفعلية.

ورغم أهمية هذا التوجه، إلا أن هناك جملة من المآخذ من بينها أن الباحثين قاموا بكتابة النص القرآني بصيغة "فونيمية" بحروف لاتينية وهذا قد يستحيل معه الاحتفاظ بكل الخصائص الكتابية والنطقية للكلمات القرآنية. يضاف إلى ذلك أيضا أن الباحثين اعتمدوا على بعض المراجع لحصر مكونات الأقسام الكلامية التي ذكروها، والخوف هنا من عدم الدقة في التوزيع والشمولية في التصنيف.

---

[2] http://corpus.quran.com

[3] http://smlc09.leeds.ac.uk/query-ar.html

ومن أبرز التوجهات البحثية الحديثة الخاصة بالقرآن، مشروع الذخيرة العربية للقرآن الكريم (Quranic Arabic Corpus) بجامعة ليدز الإبريطانية والذي يسعى إلى بناء موارد الكترونية تمكن من تحليل عميق للقرآن الكريم مبني على النماذج اللغوية للنحو التقليدي أو ما يعرف بإعراب القرآن. وهذه الذخيرة اهتمت في البداية بإعطاء تفاصيل لغوية على مستوى التحليل الصرفي وأقسام الكلام للمفردات القرآنية [37]، ومن أجل ذلك تم استخدام المحلل الصرفي لبك ولتر (Backwalter Arabic Morphological Analyzer)[4] [38] على الكلمات القرآنية للحصول على تحليل آلي وقد تلى ذلك مرحلة تصحيح يدوي. بعد ذلك بدأ أصحاب المشروع ببناء ما يعرف بالأشجار الترابطية (depency treebank) لتوضيح إعراب الآيات القرآنية [39] أعتمادا على كتب إعراب النص القرآني حسب ما ذكر. وبناء على هذه الذخيرة، قام قيس وزملاؤه [40] مؤخرا بتصميم طريقة للتحليل اللغوي عن طريق تقنية تعاونية (تحت الإشراف) متعددة المراحل. تتلخص هذه المراحل في تحليل آلي يعتمد على مجموعة من القواعد اللغوية مع تحقق يدوي مبدئي ثم تدقيق لغوي تعاوني تحت الإشراف. وقد جربوا طريقتهم هذه وقاسوا فاعليتها على الذخيرة القرآنية.

ونظرا لأهمية مشروع الذخيرة العربية للقرآن الكريم وما يسعى إليه من أهداف تشمل تحديد الخصائص اللغوية للمفردات القرآنية التي نحن بصدد العمل عليها، وإن كانت بشكل يدوي لا آلي خلافا لما نسعى إليه، فقد طلبنا من بعض المتخصصين اللغويين ومن ضمنهم عضو الفريق المختص مراجعة الموقع وتحليل محتواه لمعرفة مدى صحة ما يحتويه من معلومات وللنظر في آلية تعاطيه مع خصائص المفردات والتراكيب اللغوية في النص القرآني. وقد كانت النتيجة أن هذا العمل، وعلى الرغم من أهميته البالغة وطموحه الكبير، وقع في بعض الأخطاء من العيار الثقيل جدا و التي نذكر من بينها: أ) عدم الدقة في التصنيف والترميز اللغوين للكلمات حيث يتم المزج بين ماهو حرف مبني وماهو اسم معرب كحال أسماء الاستفهام، ب) الخلط بين المقولات النحوية (التركيبية) والمقولات الصرفية التي تخص المفردات، ج) الأخطاء الكبيرة في الإعراب وعدم الدقة في تفصيل الحالة الإعرابية لبعض الحروف، إلخ. وهناك أمثلة في موقع المشروع تبين ما ذكرناه، وقد قمنا بحصر حالات متعددة منها ضمن الورقة التي تم تقديمها لمؤتمر "جامعة طيبة لتوظيف تقنية المعلومات لخدمة النص القرآني وعلومه"[؟؟؟].

وقد لا حظنا أن السبب الرئيس في تلك الأخطاء ينبع من عدم وجود أسس لغوية متينة ترتكز عليها الجوانب الحاسوبية إضافة إلى ضرورة أن يتم العمل تحت أعين متخصصين، تجنبا للأخطاء اللغوية وبشكل خاص ما يؤدي للتحريف في معاني النص القرآني سواء عن قصد أو غير قصد. والمعضلة الكبيرة أن أغلب الأعمال المتعلقة بالمعالجة الآلية للغة العربية تتبع من بيئات غير عربية لا تتوافر فيها بالضرورة الخبرات اللغوية المطلوبة في مثل تلك الأعمال. ولأجل ذلك وتفاديا لأكبر قدر ممكن من الأخطاء اللغوية بدأنا في توجهنا البحثي المتعلق بالنص القرآني بدراسة لسانية عميقة لـ "الخصائص الصرفية والنحوية" المستهدفة بهذا العمل وذلك على محورَي المتن القرآني الرئيسين "المفردات" و"الجمل". وقد تم تقديم تفاصيل هذه الدراسة في جزئها المتعلق بالمفردات في الورقة المذكورة آنفا المقدمة لمؤتمر "جامعة طيبة لتوظيف تقنية المعلومات لخدمة النص القرآني وعلومه"[؟؟؟] المشار إليه سابقا. وفي الفقرة التالية سنقوم بتفصيل النظام الترميزي الذي تم وضعه لاستيعاب الخصائص والسمات الصرف-نحوية المستقاة من الدراسة اللغوية المذكورة.

4. ترميز الخصائص اللغوية للمفردات العربية

سنقدم في هذا الجزء من الورقة تلخيصا لأصناف الألفاظ العربية وسماتها الصرفية النحوية بناء على ماتم تقديمه في الدراسة اللغوية ومن ثم سنقترح نظاما ترميزيا لتلك الأصناف والسمات. وقبل الشروع في ذلك، نشير إلى أننا تبنينا تقسيما لغويا يصنف الألفاظ العربية إلى سبعة أقسام رئيسة، وهي مبينة في الجدول التالي (جدول 1):

الجدول 1: أصناف الألفاظ العربية

الصنف	الدلالة
الأفعال	تعني الألفاظ الدالة على الأحداث مقترنة بالأزمنة
الأسماء	تعني الألفاظ الدالة على "الذوات" Entities بنوعيها الحسية وغير الحسية
الصفات	تعني الألفاظ الدالة على الذوات متلبسة بالأحداث أو الهيئات ...
المصادر	تعني الألفاظ الدالة على الأحداث مجردة من الزمن
الظروف	مقولات تخصص الكتل الإسنادية بتقييد زمانها ومكانها
الحروف	لا تدل على ذوات ولا أحداث ولا أزمنة
العلامة	أداة دالة على التنكير والتعريف، التذكير والتأنيث، العدد، الحالة الإعرابية ... إلخ.

[4] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

وقد اعتمدنا في هذا التصنيف على طرح نظري يمثل آخر وأجود وأقوى ما توصلت إليه الدراسات اللسانية الحديثة في هذا الشأن، وهو الطرح القائم على أن المفردات في كل اللغات الطبيعية تندرج قطعا تحت أحد نوعين مقوليين اثنين: المقولات المعجمية والمقولات الوظيفية (Lexical categories vs functional categories)، وأن كل مقولة (معجمية أو وظيفية) تمثل اختزالا لمصفوفة من السمات شطر منها مشترك بينها وبين مقولة أو مقولات أخرى وشطر منها (قد يكون سمة واحدة فقط) مميز لها عن غيرها. أما الأولى أي المقولات المعجمية فتندرج تحتها كل الأصناف التقليدية المعروفة: الأسماء والأفعال والحروف والظروف والصفات، وأما الثانية فتندرج تحتها أصناف غير تقليدية لا تصنف في الأنحاء الكلاسيكية عربيِّها وغربيَّها على أنها من أقسام الكلام، وهي العناصر الدالة على معاني "الزمن" (Time) و"الجهة" (Aspect) و"الوجه" (Mood) و"الإعراب" و"التطابق" (Agreement) بمختلف مقولاته: "الشخص" و"العدد" و"الجنس" ...إلخ. وعلى هذا الأساس، فإن منْحَنا "العلامة" في اللغة العربية صفة صنف مستقل أو قسم قائم بذاته ينبغي أن يفهم في هذا الإطار النظري الذي اختصرنا بعض ملامحه في الأسطر السابقة.

إن هذا التمييز بين المقولات المعجمية والمقولات الوظيفية طرح يحظى بإجماع واسع بين المتخصصين في العلوم اللسانية واللغويات الحاسوبية في عدد من الجامعات والمعاهد العالمية المرموقة كجامعة كامبريدج ومعهد MIT. ولمزيد من التفاصيل حول نقاش هذه المسألة يرجى الرجوع إلى الفصل الثاني والثالث من الكتاب[41]، كما يمكن الإطلاع على عرض مختصر للقضية في الصفحة 24 من الكتاب [42].

4.1. ألفاظ اللغة العربية: أصنافها وسماتها الصرفية-النحوية الأساسية

الجدول 2: أصناف الألفاظ العربية وسماتها الصرفية النحوية الأساسية

النـــوع الأساسي	التفريع الأول	التفريع الثاني	حصر (إن أمكن) لمكونات الصنف (إن وجدت)	السمات الأساسية
اسم	ضمير	ضمائر شخصية	أنا وأخواتها (ضمائر الرفع)، إياي وأخواتها (ضمائر النصب)	1) الحالة الإعرابية: رفع، نصب، جر. 2) المنع من الصرف. 3) البناء والإعراب: مبني على الضم، مبني على الفتح، مبني على الكسر، مبني على السكون، معرب. 4) العدد: مفرد، مثنى، جمع سالم، جمع تكسير، جمع جموع، منتهى الجموع، إسم جنس. 5) الجنس: مذكر، مؤنث، محايد (ما يستوي فيه المذكر والمؤنث). 6) الحد: نكرة، معرفة. 7) الإفراد/التركيب: مفرد، مركب.
		ضمائر موصولة خاصة	الذي، اللذان، اللذين، التي، اللتان، اللتين، اللاتي، اللوائي، اللائي، الأولى	
		ضمائر موصولة مشتركة	من، ما، ذا، أيّ، ذو	
		ضمائر الإشارة	ذا، ذان، تين، ذه، ته، تين، أولائ، أولى، ذلك، تلك	
		ضمائر الاستفهام	من، من ذا، ما، ماذا، متى، أيان، أين، كيف، أنى، كم (الاستفهامية)، أيْ.	
		ضمائر الكناية	كم (الخبرية)، كذا، كأيّن	
	غير ضمير	علم	علم شخص، علم جنس، علم مفرد، علم مركب	
		غير علم	الأسماء الدالة على ذات (حسية أو معنوية)، الأسماء الدالة على الأزمنة، الأسماء الدالة على الأمكنة	
صفة	خبري	صفة الفاعل، صفة المفعول	*	1) الوزن الصرفي (وهو السمة المميزة للصفة) + 2) كل سمات الأسماء
	إنشائي	صفة المبالغة، الصفة المشبهة، اسم التفضيل	*	
ظرف	زمان	ظرف زمان وظيفي	اذا، الآن، الحالي، إذ، أمس، حين، صباحا، غدا، قبل، مساء، يوم، الليلة، إلخ.	1) (ليس لها سمات تخصها بعينها) 2) من سمات الأسماء: الحالة الإعرابية، البناء والإعراب.
		ظرف زمان معجمي		
	مكان	ظر مكان وظيفي	أمام، بعد، بين، تحت، حول، حيث، خلف، دون، عند، فوق، لدى، نحو، هنا، هناك، وبين، يسار، يمين، الخ.	
		ظر مكان معجمي		
مصدر	مصدر ميمي	مصدر ميمي	*	1) هناك بعض السمات الخاصة (أنظر ص 19) حسب كل نوع + 2) من سمات الأسماء: الحالة الإعرابية، البناء والإعراب، العدد، الجنس، الحد. 3) من سمات الصفات: الوزن الصرفي 4) من سمات الأفعال: كل السمات إلا سمة الزمن.
	مصدر غير ميمي	مصدر قياسي من فعل ثلاثي مجرد، مصدر سماعي من فعل ثلاثي مجرد، مصدر من فعل ثلاثي مزيد (أوله تاء زائدة)، مصدر من فعل ثلاثي مزيد (أوله ليس تاء زائدة)، مصدر العدد، مصدر الهيئة، مصدر مراد به المفعول، مصدر	*	

		التأكيد، اسم المصدر، المصدر الصناعي		
فعل	ما يدل على حدث أو حالة	معجمي	*	1) <u>الزمن</u>: ماضي، حال، مستقبل. 2) <u>الوجه</u>: خبر، أمر. 3) <u>التعدي واللزوم</u>: متعدي، لازم، متعدي ولازم. 4) <u>الصحة</u>: صحيح (للفعل المعجمي) 5) <u>الاعتلال</u>: معتل الأول (مثال)، معتل الوسط (أجوف)، معتل الآخر (منقوص). 6) <u>التجرد</u>: مجرد 7) <u>الزيادة</u>: مزيد فيه حرف واحد، مزيد فيه حرفان، مزيد فيه ثلاثة أحرف. 8) <u>الجمود والتصرف</u>: جامد، متصرف. + من سمات الأسماء: <u>الحالة الإعرابية</u>، <u>البناء والإعراب</u> + من سمات الصفات: الوزن (يحدد حالة البناء للمعلوم/المجهول **<u>للفعل المعجمي</u>**)
	ما لا يدل على حدث	وظيفي	*	
حرف	مختص بالفعل	مختص بالفعل للنصب	**<u>حروف النصب</u>**: أنْ، لن، إذن، كي (كذلك لام كي)، حتى. ويضاف للنواصب فاء السببية (لا تأتي إلا بعد اسلوب طلبي: الاستفهام، الأمر، النهي).	تأخذ سمة البناء (على الفتح، على الكسر، على الضم، على السكون) من سمات الأسماء.
		مختص بالفعل للجزم	**<u>حروف الجزم</u>**: لم, لما, لام الأمر (ل: التي في نحو "لينفقْ ذو سعة من سعته، لْ: بعد الواو والفاء), لا الناهية, إنْ (الشرطية), إذما, ما (الشرطية), من (الشرطية), مهما, متى (الشرطية), أيان (الشرطية), أينما, أين(الشرطية), أنّى, حيثما, كيفما, أيُّ (الشرطية).	
		مختص بالفعل للرفع (عامل التجرد)	مكان فارغ قبل الفعل يمكن أن يوضع فيه الناصب أو الجازم أو يبقى فارغا.	
	مختص بالاسم	مختص بالاسم للنصب	1) **<u>حروف النداء</u>**: أ, أي, يا, ا, أيا , هيا, وا. 2) **<u>إنَّ وأخواتها</u>**: إنّ، أنّ، لكنّ(للاستدراك)، كأن، ليت، لعل، لا (النافية للجنس). 3) واو المعية	
		مختص بالاسم للجر	**<u>حروف الجر</u>**: الباء, من, إلى, عن, على, في, ك, واو القسم, تاء القسم, مذ, منذ, رُبّ, حتى, خلا, عدا, حاشا, ل (يمكن أن تكون لام التعليل، لام الجحود، لام العاقبة).	
		مختص بالاسم للرفع	**<u>عامل التجرد</u>**: إذا تجرد الإسم من عوامل الجر وعوامل النصب، فإنه يكون مرفوعا.	
	غير مختص	ربطي	و, ف, ثم, حتى, أو, أم, بل, لا, لكنْ (بتسكين النون), ف (الاستئنافية).	
		استفهامي	أ، هل	
		شرطي	لو، لولا، إذا، كلّما، لمّا	
		جوابي	نَعَم، بلى، أجَلْ، لا	
		استفتاحي	ألا	
علامة	مختصة بغير بالأفعال إلا الحروف	علامة مختصة بالاسم والصفة والمصدر والظرف	الضمة، الألف، الواو، الفتحة، الياء، الكسرة، العلامة المقدرة (مكان فارغ توضع فيه العلامة المقدرة، أي ما يشبه عامل التجرد)، لام التعريف، التنوين، إلخ.	تأخذ سماتها من سمات الأسماء.
	مختصة بالأفعال	علامة مختصة بالفعل	الضمة، الفتحة، السكون، ثبوت النون (في الأفعال الخمسة)، التجرد من العلامة، حذف النون (في الأفعال الخمسة)، حذف حرف العلة (في الفعل المعتل من نوعي الحال والمستقبل).	

وبناء على هذا الجدول، فإن أقسام الكلام التي ستستهدف بالترميز هي تلك التي وضعت في العمود الثالث من الجدول تحت عنوان التفريع الثاني. أما السمات المستهدفة بالترميز فهي تلك الموجودة في العمود الأخير من الجدول مع مراعات عدم التداخل بينها. ولإظهار قائمة السمات التي سترمز، قمنا بحصرها في الجدول 3.

الجدول 3: قائمة السمات التي ينبغي ترميزها

الصنف	السمات
الأسماء	1) **الحالة الإعرابية:** رفع، نصب، جر. 2) **المنع من الصرف**. 3) **البناء والإعراب:** مبني على الضم، مبني على الفتح، مبني على الكسر، مبني على السكون، معرب. 4) **العدد:** مفرد، مثنى، جمع سالم، جمع تكسير، جمع جموع، منتهى الجموع، اسم جنس. 5) **الجنس:** مذكر، مؤنث، محايد (ما يستوي فيه المذكر والمؤنث). 6) **الحد:** نكرة، معرفة. 7) **الإفراد/التركيب:** مفرد، مركب
الأفعال	**1) الزمن:** ماضي، حال، مستقبل. **2) الوجه:** خبر، أمر. **3) التعدي واللزوم:** متعدي، لازم، متعدي ولازم. **4) الصحة:** صحيح (لاينطبق على الفعل الوظيفي). 5) **الاعتلال:** معتل الأول (مثال)، معتل الوسط (أجوف)، معتل الآخر (منقوص). **6) التجرد:** مجرد. **7) الزيادة:** مزيد فيه حرف واحد، مزيد فيه حرفان، مزيد فيه ثلاثة أحرف. **8) الجمود والتصرف:** جامد، متصرف. **9) من سمات الأسماء:** الحالة الإعرابية، البناء والإعراب. **10) من سمات الصفات:** الوزن الصرفي (لتحديد حالة البناء للمعلوم/المجهول للفعل المعجمي).
الصفات	**1)** الوزن الصرفي. **2) من سمات الأسماء:** كل السمات.
الظروف	**1) من سمات الأسماء:** الحالة الإعرابية، البناء والإعراب.
المصادر	1) **من سمات الأسماء:** الحالة الإعرابية، البناء والإعراب، العدد، الجنس، الحد. 2) **من سمات الصفات:** الوزن الصرفي. 3) **من سمات الأفعال:** كل السمات إلا سمة الزمن.
الحروف	1) **من سمات الاسماء:** البناء (على الضم، على الفتح، على الكسر، على السكون).
العلامات	1) تأخذ سماتها من سمات الأسماء.

4.2. ترميز أصناف الألفاظ العربية وسماتها الصرفية-النحوية الأساسية

بناء على الجداول السابقة قمنا بترميز أصناف الألفاظ وسماتها كل على حدة. ومن الجدير بالذكر أننا نستهدف بالترميز الأقسام الصرفية للكلمة بمعنى أن الكلمة تمر أولا على محلل صرفي ليقوم بتقسيمها إلى أجزاء وكل جزء يتم وسمه على حدة بما ينطبق عليه من أصناف وسمات. وعلى هذا الأساس فإننا نضيف صيغة لترميز الكلمة في شكلها الإجمالي ليكون النظام الترميزي شاملا، بمعنى أنه يتيح إمكانية توصيف أجزاء الكلمة بما ينطبق عليها من أصناف وسمات وكذلك توصيف الكلمة بشكل إجمالي عند الحاجة لذلك.
مثال: المسلمون، ستقسم أولا إلآ أربعة أقسام هي: أل، مسلم، و، ن. وكل جزء منها سيأخذ الصنف المناسب والسمات التي تنطبق عليه، ومنه يتم استخراج السمات العامة التي تخص الكلمة في مجملها.

4.2.1. ترميز الأصناف الأساسية للألفاظ العربية:

الجدول 4: رموز الأصناف الأساسية للألفاظ العربية

النوع الأساسي	التفريع الأول	التفريع الثاني	الرمز	اسم الرمز بالإنكليزي
الإسم	غير ضمير	إسم	NNN	Noun
		إسم علم	NNP	Proper Noun
	ضمير	ضمير شخصي	NPP	Personal Pronoun
		اسم اشارة	NPD	Demonstrative Pronoun
		اسم إستفهام	NPI	Interrogative Pronoun
		اسم كناية	NPE	Euphemism Pronoun
		اسم موصول خاص	NPR	Relative Pronoun
		اسم موصول مشترك	NPC	Common Relative Pronoun
المصدر	ميمي	مصدر ميمي	VNM	Verbal Noun Starts with "m"
	غير ميمي	مصدر قياسي من فعل ثلاثي مجرد	VTT	Verbal Noun templatic from triletral base verb
		مصدر سماعي من فعل ثلاثي مجرد	VST	Verbal Noun soundex from triletral base verb
		مصدر من فعل ثلاثي مزيد (أوله تاء زائدة)	VAT	Verbal Noun from Augmented verb starting with "t"
		مصدر من فعل ثلاثي مزيد (أوله ليس تاء زائدة)	VAX	Verbal Noun from Augmented verb starting with not "t"
		مصدر العدد	VNU	Number Verbal Noun
		مصدر الهيئة	VSS	State Verbal Noun
		مصدر مراد به المفعول	VNO	Object Verbal Noun
		مصدر التأكيد	VNA	
		اسم المصدر	VNN	

		المصدر الصناعي	VNI	
الفعل		وظيفي	VBF	Functional Verb
		معجمي	VBL	Lexical Verb
الصفة	صفة خبرية	صفة الفاعل	AJV	Subject Adjective
		صفة المفعول	AJO	Object Adjective
	صفة إنشائية	صفة المبالغة	AJS	Superlative Adjective
		الصفة المشبهة بإسم الفاعل	AJP	Participle Adjective
		اسم التفضيل	AJC	Comparative Adjective
الظرف	زمان	وظيفي	ATF	Functional Temporal Adverb
		معجمي	ATL	Lexical Temporal Adverb
	مكان	وظيفي	ALF	Functional Location Adverb
		معجمي	ALL	Lexical Location Adverb
الحرف	مختص بالاسم	رافع	PNN	Noun Nominative Particle
		ناصب	PNA	Noun Accusative particle
		جار	PNG	Noun Genitive Particle
	مختص بالفعل	رافع	PVN	Verb Nominative Particle
		ناصب	PVA	Verb Accusative Particle
		جازم	PVJ	Verb Jussive Particle
	غير مختص	ربطي	PUC	Conjunction Particle
		إستفهامي	PUI	Interrogative Particle
		شرطي	PUD	Conditional Particle
		جوابي	PUA	Answer Particle
		إستفتاحي	PUT	Introduction Particle
العلامة	مختص بالفعل	علامة مختصة بالفعل	MVB	Verbal Mark
	مختص بغير الفعل	علامة مختصة بالاسم والصفة والمصدر	MNV	Non-Verbal Mark

4.2.2. رموز السمات الصرفية-النحوية الأساسية

الجدول 5: الرموز المقترحة للسمات الصرفية-النحوية الأساسية

السمة بالعربية	السمة بالإنكليزي	الرمز	اسم النوع بالعربي	اسم النوع بالإنكليزي
العدد	Number	NS	مفرد	Singular
		ND	مثنى	Dual
		NP	جمع سالم	Plural
		NL	جمع جموع	Plural of Plurals
		NM	منتهى الجموع	Plurals of Multitude
		NB	جمع تكسير	Broken Plural
		NA	اسم جنس	Abstract Noun
التركيب	Compound	CU	مفرد	Non Compound
		CP	مركب	Compound
الحد	Definiteness	DF	معرفة	Definite

Indefinite	نكرة	DI		
Feminine	مؤنث	GF	Gender	الجنس
Masculine	مذكر	GM		
Neutral	محايد(مايستوي فيه المذكر والمؤنث)	GN		
Past	الماضي	TP	Tense	الزمن
Present	الحاضر	TR		
Future	المستقبل	TF		
Predicative	خبر	MP	Mode	الوجه
Subjunctive	أمر	MS		
Accusative	النصب	CA	Case	الحالة الإعرابية
Genitive	الجر	CG		
Jussive	الجزم	CJ		
Nominative	الرفع	CN		
Active	مبني للمعلوم	VA	Voice	صيغة البناء
Passive	المجهول	VP		
First	المتكلم	P1	Person	الشخص
Second	المخاطب	P2		
Third	الغائب	P3		
Transitive	متعدي	TS	Transitivity	التعدي
Intransitive	لازم	TI		
Declinable	المعرب	DC	Declension	البناء
Indeclinable Accusative	المبني على الفتح	DA		
Indeclinable Genitive	المبني على الكسر	DG		
Indeclinable Jussive	المبني على السكون	DJ		
Indeclinable Nominative	المبني على الضم	DN		
Mark Implicit	ظاهرة	MI	Declension Mark	علامة الإعراب
Mark Explicit	مقدرة	MX		
Pattern	الوزن	PA	Patterns	الوزن
Sound	صحيح	SS	Soundness	الصحة
Assimilated Verb (first-weak)	مثال	SA		
Hollow Verb (second-weak)	أجوف	SH		
Defective Verb (third-weak)	منقوص	SD		
Deficient Noun Ending with alif	مقصور	SL		
Deficient Noun Ending with ya	الناقص	SY		
Deficient Noun Ending with Alif-Hamza	الممدود	SZ		
Form 1	مجرد	F0	Form	التجرد
Augmented with one letter	مزيد حرف واحد	F1		
Augmented with two letters	مزيد حرفان	F2		
Augmented with three letters	مزيد ثلاث أحرف	F3		
Diptote	جامد/لاينصرف	ID	Invariability	الجمود
Triptote	متصرف	IT		

4.2.3. بنية الرمز الإجمالي للكلمة

كما أسلفنا، فإن الكلمة يتم تقسيمها أولا إلى أجزائها الصرفية الأساسية، ومن ثم يتم تصنيف كل جزء منهاوتحديد سماته الرئيسة بناء على ما تقدم من توصيفات لغوية والترميزات التي تقابلها. ومن هذه العملية تنتج مصفوفةمن الخصائص والسمات لكل قطعة على حدة؛ وهذه المصفوفات يمكن دمجها في مصفوفة واحدة تعكس التوصيف اللغوي للكلمة في مجملها مع مراعات أن بعض الخصائص في الأجزاء المكونة للكلمة يمكن أن يحصلفيهانوع من التعارض. وبالتالي سيكون هناك رمز للكلمة يتألف من سلسلة من الخانات كل منها تمثل سمة لغوية معينة.

4.3. تطبيق: استخدام التوصيف اللغوي في تحليل عينة من النص القرآني

بعد الانتهاء من حصر أصناف الألفاظ العربية وتحديد سماتها الصرفية-النحوية الرئيسة وترميزها بدأنا نعد لاستخدامها الفعلي في تحليل عينة من النص القرآني بغية التعرف عن قرب على مدى قدرتها على وصف الخصائص والسمات الأساسية للمفردات العربية. وسيتيح ذلك لنا فرصة مراجعة التوصيف وتحسينه عند الحاجة ليكتمل وينضج. كما أننا نهدف أيضا من خلال هذا التحليل اليدوي إلى تجهيز عينة تستخدم في تدريب النماذج الإحصائية التي نسعى لبنائها من أجل التعرف على الأصناف والسمات بشكل آلي تماشيا مع أهداف في هذا التوجه.

ولأجل ذلك قمنا بإعداد نظام حاسوبي يسمح للخبير اللغوي المتخصص بالقيام بالتحليل والتوصيف المطلوبين وفقا لما تم وضعه من خصائص وسمات صروفو-نحوية. ونعتمد في ذلك على نسخة من النص القرآني مشكلة ومنقحة تم إعدادها في الجزء السابق من مشروع النص القرآني (رقم "أت-25-113") [1-3].

يقوم النظام بعرض سور القرآن المسندة للخبير اللغوي (محدد باسم مستخدم وكلمة سر) ثم يتيح له إمكانية تقطيع الكلمات إلى أجزاء عبر نافذة متحركة يتحكم فيها بسهولة من خلال أزرار للتنقل في جوانبها (الشكل 4). وعند عزل قطعة لغوية من الكلمة يتاح للخبير اللغوي إمكانية توصيفها وفقا للخصائص والسمات التي تم وضعهافي الدراسة اللغوية ثم حفظها في قواعد بيانات النظام والانتقال إلى المقطع الذي يليها في نفس الكلمة التي تتم معالجتها مع إمكانية التحديث في أي وقت يريده الخبير (الشكل 5). وعند انتهاء جلسة العمل يقوم النظام بوضع مؤشر عند آخر ما وصل إليه الخبير في التحليل، وعند الدخول إلى النظام في المرة التالية يتم عرض البيانات من آخر نقطة توقف عندها الخبير وذلك لتسريع عمله وتسهيل المهمة عليه.

يتم الآن العمل على إضافة جانب من الذكاء في عمل نظام التحليل الحاسوبي هذا، من خلال البحث في قاعدة البيانات عن الحلول المخزنة لمقاطع مشابهة تم تحليلها في مراحل سابقة. فإن وجدت مقاطع مطابقة للمقطع الذي يتم توصيفه يعرض للخبير التوصيف المخزن ويترك له الخيار في اعتماد الأنسب منه للمقطع الحالي أو تجاهله ووضع توصيف يدوي جديد. هذه الطريقة ستسرع عمل الخبير اللغوي بشكل ملحوظ إضافة إلى ضمان مستوى من التناسق في التحليل. وقد قمنا بتصميم النظام كتطبيق ويب[5] ليتمكن عدة خبراء من العمل في نفس الوقت وليستفيد كل منهم من خبرة الآخر من خلال خاصية الحلول الآلية التي يتم استخراجها من قاعدة البيانات الموحدة في النظام والذي يتم فيها تخزين نتائج التحليل لكل الخبراء.

[5] http://www.ariscom.org/POST

الشكل 4 : واجهة استخدام النظام الخاصة بوسم المقاطع اللغوية

الشكل 5 : واجهة استخدام النظام لعرض الأجزاء اللغوية والتنقل عبرها

5. خاتمة

لقد قدمنا في هذه الورقة البنية التحتية لنظام حاسوبي للتعرف الآلي على الخصائص والسمات اللغوية للمفردات القرآنية ضمن أنشطة مشروع يخص النص القرآني ممول من مدينة الملك عبد العزيز للعلوم والتقنية بمنحةرقم "أت-30-199".

فقد قدمنا أهم النماذج الإحصائية المستخدمة في مجال التعرف الآلي على المفردات اللغوية ووضعنا تصورا لكيفية استخدامها والخلفية العلمية اللازمة لذلك، ثم استعرضنا أدبيات البحث العلمي في المجال مع التركيز على الجهود التي تخص النص القرآني. وقد أوضحنا ضرورة إيجاد خلفية لغوية متينة تتبنى عليها المعالجة الآلية على محوري المفردات والجمل، وهو ما قمنا به كمرحلة أولى من مراحل تنفيذ المشروع. وبعد انتهاء الدراسة اللغوية على مستوى المفردات، قمنا باستخلاص قائمة الخصائص والسمات المميزة للمفردات ثم وضعنا لها نظاما

ترميزيا يسهل التعاطي معها أثناء المعالجة الآلية. ومن ثم عكفنا على تهييئ بيئة حاسوبية للتعامل مع تلك الخصائص والسمات لتسهيل وتسريع عمل الخبراء اللغويين العاملين عليها. ونحن نسعى من خلال هذا النظام الحاسوبي إلى تهييئ أرضية مناسبة لتطبيق نتائج الدراسة اللغوية من جهة، ومن جهة أخرى إلى تحليل عينة من النص القرآني بشكل يدوي عبر خبراء لغويين ليتم استخدامها في تدريب واختبار نماذج إحصائية تكون قادرة على التعرف بشكل آلي على الخصائص والسمات المطلوبة. وستنصب جهودنا في المرحلة القادمة على بناء النماذج الإحصائية وتدريبها على العينة اليدوية ثم المفاضلة بين هذه النماذج باعتبار معيار قوة الأداء و جودته ثم اختيار الأجود و الأكفى و الأمثل .

6. شكر

هذا العمل يقدم بعضا من نتائج مشروع "تحديد الخصائص اللغوية للمفردات القرآنية" الممول من مدينة الملك عبدالعزيز للعلوم والتقنية بمنحة رقم "آت-30-199" والذي ينفذ بجامعة الإمام محمد بن سعود الإسلامية. فلهما الشكر على الدعم والتسهيلات.

7. المراجع

[1] يحيى الحاج، عماد الصغير، محمد الكنهل، منصور الغامدي، يوسف العوهلي، عبدالله الأنصاري. التعلم الآلي للقرآن الكريم. التقرير الفني النهائي، مدينة الملك عبد العزيز للعلوم والتقنية،الرياض - السعودية، 2010.

[2] يحيى الحاج، عماد الصغير، احمد خرصي، عبد الله الأنصاري. التحليل الصرفي للقرآن الكريم. المجلة الدولية لعلوم وهندسة الحاسوب باللغة العربية، المجلد 3، العدد 1، 2010.

[3] يحيى الحاج، عماد الصغير، احمد خرصي، عبد الله الأنصاري. قاعدة بيانات مفهرسة لكامل النص القرآني. سجلات المؤتمر الدولي الخامس حول ممارسة علوم الحاسب باللغة العربية، الرباط – المغرب، 10-13 مايو 2009م.

[4] يحيى الحاج، عادل عمار، رشيد بوزيان. تحديد الخصائص اللغوية للمفردات القرآنية. التقرير الفني للسنة الأولى، مدينة الملك عبد العزيز للعلوم والتقنية، الرياض – السعودية، 2013.

[5] Stolz W. S., Tannenbaum P. H., Carstensen F. V. A stochastic approach to the grammatical coding of English, Communications of the ACM, 8(6), pp. 399–405, 1965.

[6] Bahl L. R., Mercer R. L. Part of speech assignment by a statistical decision algorithm, Proceedings IEEE International Symposium on Information Theory, pp. 88–89, 1976.

[7] Vintsyuk, T. K. Speech discrimination by dynamic programming. Cybernetics, 4(1), 52–57. Russian Kibernetika, 4(1), pp. 81-88, 1968.

[8] Schmid H., Laws F. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. COLING'08 Manchester, UK, 2008.

[9] Sameti H., Bahrani M., Babaali B. Segmental HMM-based Part-of-speech tagger. International Conference on Audio Language and Image Processing (ICALIP), 2010.

[10] Schmid H. Part-of-speech tagging with neural networks. Proc. Int. Conf. on Computational Linguistics, Japan, pages: 172–176, 1994.

[11] Zamora-Martinez F., Castro-Bleda M.J., Espana-Boquera S., Tortajada-Velert S. Adding morphological information to a connectionist part-of-speech tagger. In Proceedings of the Current topics in artificial intelligence, and 13th conference on Spanish association for artificial intelligence (CAEPIA'09), Pedro Meseguer, Lawrence Mandow, and Rafael M. Gasca (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 191-200, 2009.

[12] Pérez-Ortiz J.A., Forcada M.L. Part-of-speech tagging with recurrent neural networks. Proceedings of the International Joint Conference on Neural Networks, IJCNN 2001 (Washington D.C., USA), pp. 1588-1592, 2001.

[13] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77 (2), pp. 257-286, 1989.

[14] Y.O.M. Elhadj. Statistical Part-of-Speech Tagger for Traditional Arabic Texts. Journal of Computer Science, Vol. 5, No 11, pp: 794-800, 2009.

[15] Y.O.M. Elhadj, I.A AlSughayeir, A.M. Alansari. Arabic part-of-speech tagging using the sentence structure. Proceeding of the 2nd International Conference on Arabic Language Resources and Tools, pp: 241-245, Cairo, Egypt, April 22-23, 2009.

[16] Jurafsky D. and J. H. Martin. Speech and Language Processing: An Introduction to Speech Recognition, Computational Linguistics and Natural Language Processing. Prentice Hall, ISBN: 10: 0131873210, 2008.

Brants, T. TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on [17] applied natural language processing, Seattle, Washington, pp. 224–231. Morgan Kaufmann Publishers Inc, 2000.

Baayen R. H. and Sproat R. Estimating lexical priors for low-frequency morphologically [18] ambiguous forms. Commputational Linguistics, 22(2), pp. 155–166, 1996.

Marques, N.C. and Gabriel, P.L. Using neural nets for Portuguese part-of-speech tagging. [19] Proceeding of the Fifth International conference on The Cognitive Science of Natural Language Processing, 1996.

Ma, M. S. Q. and Isahara, H. A multi-neuro tagger applied in Chinese texts. In Proceedings of [20] 1998 International conference on Chinese information processing, pp 200-207, 1998.

Parikh, A. Part-Of-Speech Tagging using Neural Network. Proceedings of ICON-2009: 7th [21] International Conference on Natural Language Processing, 2009.

Jabar, H. Yousif, Tengku, M. and Tengku Sembok. Design and Implement an Automatic Neural [22] Tagger Based Arabic Language for NLP Applications. Asian Journal of Information Technology, Vol. 5, pp. 784-789, 2006.

S. Abuleil and M. Evens. Discovering lexical information by tagging Arabic newspaper text. [23] Proceedings of the workshop on Semitic Language Processing, COLING-ACL98, University of Montreal, Montreal, PQ, Canada, pp. 1–7, 1998.

El-Kareh and Al-Ansary. An Arabic interactive multi-feature pos tagger. Proceeding of the [24] international conference on Artificial and Computational intelligence for Decision Control and Automation in engineering and Industrial Application (ACIDCA) conference, Tunisia, pp. 83–88, 2000.

S. Khoja. Apt: Arabic part-of-speech tagger. Proceedings of the Student Workshop at the Second [25] Meeting of (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania, 2001.

N. Habash and O. Rambow. Arabic tokenization, part-of-speech tagging and morphological [26] disambiguation in one fell swoop. The Association for Computer Linguistics, 2005.

M. Diab, K. Hacioglu, and D. Jurafsky. Automatic tagging of Arabic text: From raw text to base [27] .phrase chunks. Proceedings of HLT-NAACL, 2004.

E. Marsi, A. van den Bosch, and A. Soudi. Memory-based morphological analysis generation and [28] part-of-speech tagging of Arabic. ACL-05. Computational Approaches to Semitic Languages. Workshop Proceedings, University of Michigan, Ann Arbor, Michigan, USA, 2005.

F. A. Shamsi and A. Guessoum. A hidden Markov model based pos tagger for Arabic. [29] Proceedings of 8th International Conference on the Statistical Analysis of Textual Data, France, 2006.

H. M. Harmain. Arabic part-of-speech tagging. Proceedings of the Fifth Annual U.A.E. [30] University Research Conference, Al-Ain, U.A.E, 2006.

M. Albared, N. Omar, and M. J. Abd Aziz. Developing a Competitive HMM Arabic POS Tagger [31] Using Small Training Corpora. Proceedings of ACIIDS (1): pp. 288-296, 2011.

S. Alqrainy, H. Mauidi, and A. Ayesh. Pattern-Based Algorithm for Part-of-Speech Tagging [32] Arabic Text. Proceedings of the International Conference on Computer Engineering and System (ICCES), pp. 119-124, 2008.

S. Alqrainy and A. Ayesh. Developing a tagset for automated pos tagging in Arabic. WSEAS [33] TRANSACTIONS on COMPUTERS, vol. 5, no. 11, pp. 2787–2792, 2006.

M. Sawalha, and E.S. Atwell. Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for [34] Arabic Text. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 1258 – 1265, 2010.

Rafi T, Shuly W. Morphological Tagging of the Quran. Proceedings of the Workshop on Finite- [35] State Methods in NLP, an EACL'03 Workshop, Budapest, Hungary, April 2003.

Judith D, Dudu S, Rafi T, Shuly W. Morphological Analysis of the Quran. Literary and [36] Linguistic Computing, 19(4), pp. 431-452, 2004.

K. Dukes and N. Habash. Morphological Annotation of Quranic Arabic. Proceedings of [37] Language Resources and Evaluation Conference (LREC), Valletta, Malta, 2010.

Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data [38] Consortium, Philadelphia, 2002.

K. Dukes and T. Buckwalter. A Dependency Treebank of the Quran using Traditional Arabic Grammar. Proceedings of the 7th International Conference on Informatics and Systems (INFOS), Cairo, Egypt, 2010. [39]

K. Dukes, E. Atwell and N. Habash. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. Language Resources and Evaluation Journal (LREJ). Special Issue on Collaboratively Constructed Language Resources, 2012. [40]

N. Chomsky. The Minimalist Program. Cambridge Mass: The MIT press, 1995. [41]

A. Radford. Minimalist Syntax. Cambridge: Cambridge University Press, 2004. [42]