وسم كلمات اللغة العربية باستخدام نظام ستانفورد: حروف الجر نموذجاً

Arabic Part of Speech Tagging by Using the Stanford System: Prepositions as a Case Study

ضياء الدين أبوزينة*، وتقي الدين التميمي**

Dia Eddin AbuZeina & Taqieddin Al-Tamimi

*كلية تكنولوجيا المعلومات وهندسة الحاسوب، جامعة بوليتكنك فلسطين، الخليل، فلسطين.
**جامعة فلسطين التقنية، خضوري، فرع العروب، الخليل، فلسطين

* Faculty of Information Technology and Computer Engineering, Palestine Polytechnic University, Hebron, Palestine. ** Palestine Technical University, Kadoori, Al-Aroub Branch, Hebron, Palestine

*الباحث المراسل: abuzeina@ppu.edu

تاريخ التسليم: (2019/2/21)، تاريخ القبول: (2019/7/9)

ملخص

تتناول هذه الدراسة مسألة وسم الكلمات (تحديداً حروف الجر) في النصوص العربية المكتوبة وهو ما يعرف بـ (Arabic part of speech tagging). تحتوي اللغة العربية على عدد محدد من مجموعات الحروف (particles)، مثل: حروف الجر، وحروف الجزم، وحروف النصب، الخ. وتؤدي كل مجموعة دوراً معيناً في السياق الذي ترد فيه. بشكل عام، فإن الوسم هو عملية تحديد الصيغة الصرفية لكل كلمة سواء كانت اسماً، أو فعلاً، أو حرفاً بأنواعها المختلفة وذلك بالاعتماد على السياق الذي ترد فيه. يفيد وسم النصوص في كثير من تطبيقات معالجة اللغات الطبيعية، اذ يستخدم بشكل رئيسي في تحليل النصوص (syntactic parsing) للتحقق من صحة الجملة وتوافقها مع قواعد اللغة، وكذلك يستخدم لتحليل النص لفهم المعنى المطلوب لاستخدامه في محركات البحث (search engines). تشمل تطبيقات معالجة اللغات ايضاً الترجمة الآلية (machine translation)، تأليف الكلام (speech synthesis)، التعرف على الكلام (speech recognition)، التشكيل الآلي (diacritization) وغيرها. وبذلك فإن جودة الأداء في كثير من التطبيقات اللغوية تعتمد على دقة مخرجات نظام الوسم المستخدم، فكانت هذه الدراسة التي تشير إلى نظام ستانفورد (Stanford tagger) لوسم النصوص العربية وذلك بهدف تحديد أقسام الكلام المستخدمة في النص موضع الفحص (مجموعة الصيغ الصرفية) وكذلك تهدف الدراسة الى التعرف على أداء نظام ستانفورد عند وسم حروف الجر في اللغة العربية. تناقش هذه الدراسة كذلك جوانب الضعف في نظام ستانفورد؛ فقد توصل الباحثان إلى أن هذا النظام لا يعالج مسألة الاقتران كأن يقترن حرف الجر بالكلمة، مثل (بِـ) و (لِـ)، وكذلك فإنه يعطي وسماً مشتركاً

لحروف مختلفة من حيث الوظائف اللغوية كالجزم والاستفهام، فعبر دراستنا الاستقرائية لحروف الجر لم نلحظ خلافاً في وسمها، مثل: إلى، في. ويصعب تمييز الحروف الأخرى إلا عبر السياق مثل: حتى، عدا؛ مما يشير إلى عدم الدقة في هذا الوسم والحاجة إلى تطويره لمواكبة الأنظمة المرتبطة بعملية الوسم؛ وهنا تأتي أهمية هذا البحث. تم استخدام مدونة القرآن الكريم للتعرف على أداء نظام ستانفورد عند وسم حروف الجر الواردة في القرآن الكريم. وبشكل عام فإن نتائج هذه الدراسة تدعو إلى مزيد من البحث والتمحيص في مسألة وسم الحروف الأخرى في العربية بهدف دراسة مدى توافق رموز الوسم المستخدمة في نظام ستانفورد مع الحروف المستخدمة في اللغة بشكل عام.

الكلمات المفتاحية: معالجة اللغات الطبيعية، اللغة العربية، أقسام الكلام، وسم حروف الجر، الصيغ الصرفية، مدونة القرآن الكريم.

Abstract

This paper discusses part of speech (PoS) tagging for Arabic prepositions. Arabic has a number of predefined sets of particles such as particles of *Nasb*, particles of *Jazm*, particles of *Jarr* (also called prepositions), etc. Each set has a particular role in the context in which it appears. In general, PoS is the process of assigning a tag for each word (e.g. name, verb, particle, etc.) based on the context. In fact, PoS is a beneficial tool for many natural language processing (NLP) toolkits. For instance, it is used in syntactic parsing to validate the grammar of the sentence in question. It is also beneficial to understand the required meaning via textual analysis for further processing in search engines. Many other language processing applications utilize PoS such as machine translation, speech synthesis, speech recognition, diacritization, etc. Hence, the performance quality of many NLP applications depends on the accuracy of outputs of the used tagging system. Hence, this study examines the Stanford tagger to explore its tag set in the text under examination and its performance for tagging Arabic prepositions. This study also discusses the weaknesses of the Stanford tagger, as it does not handle the merging case when a preposition joins with an adjacent word to form one single word. Another concern of the Stanford tagger is that it gives a unique tag for different particles such as *Jarr* and *Jazm* in terms of linguistic functions. Through our inductive study of prepositions in terms of linguistic functions such as *Jazm and Istifham* (interrogation), we did not note differences in tagging prepositions like "to" (إلى) and "in" (في). Other

prepositions are also difficult to distinguish unless they are contextualized; these include “until” (حتى) and “except” (عدا). This shows that this tagging system is inaccurate and the need for keeping up with tagging-related systems is vital, hence is the significance of our research. In this work, we used the Holy Quran to identify the performance of the Stanford System in tagging prepositions in the Quran. This work encourages more research on tagging other Arabic prepositions to explore the compatibility of tagging symbols employed in the Stanford System and prepositions used in the Arabic language, in general.

Keywords: Natural language processing (NLP), Arabic, part of speech, tagging, prepositions, syntactic category, Holy Quran data set.

مقدمة

حظيت حوسبة اللغة العربية في السنوات الأخيرة اهتماماً ملحوظاً بهدف الانتفاع من القدرة الهائلة التي توفرها أنظمة الحاسوب في معالجة البيانات لاستخراج المعلومات، وتشمل حوسبة اللغة كثيراً من المجالات والتطبيقات؛ فقد غدا النشر الالكتروني عبر الاعتماد على التقنيات الحديثة التي توجه عن بعد، وتوجه للعامة بقصد الإفادة من المنجزات في عمليات النشر كما أشار لذلك (Al-hoshe, 2002, p. 152)، ويرى الباحثان أنَّ ما يجمع هذه التطبيقات هو الحاجة إلى تحليل النص؛ من أجل تمكين الخوارزميات المستخدمة من إعطاء النتيجة المناسبة، وهو ما سيظهر لاحقا في هذا البحث؛ لأنَّ تحليل النص يشمل بالدرجة الأولى فهم تركيب الجملة وما تحتويه من صيغ صرفية كالأسماء والأفعال والحروف. ومن هنا يأتي دور أنظمة الوسم (taggers) في تحديد فئة الكلمة أو صنفها (the word’s tag). وبذلك فإن وسم النصوص أو الكلمات يعتبر جزءاً رئيسياً في أنظمة معالجة اللغات الطبيعية لأنه يشكل مصدرا معرفياً هاما في تطبيقات معالجة النصوص. وبالرغم من أهمية وسم النصوص لبيان دور كل كلمة في النص من الناحية الصرفية إلا أن هذه المعلومات على أهميتها تبقى قليلة مقارنة مع ما يمكن الحصول عليه من معلومات من خلال عملية التحليل الدلالي للنص (semantic analysis)، وهو موضوع خارج نطاق هذه الدراسة. إن الصنف أو الفئة وكذلك ما يمكن أن نطلق عليه وسم الكلمة يكون بناء على مجموعة أصناف معرفة مسبقاً في النظام. وبالتالي فإن كل نظام يعمل وفق ما يتم تحديده لقائمة الأصناف الممكنة في اللغة التي نرغب بوسم نصها. وبشكل عام، فإن اللغة الإنجليزية تقسم إلى ثمانية أصناف رئيسية وهي: الأسماء، الأفعال، الضمائر، حروف الجر، الظروف، حروف العطف، أسماء الأفعال، وأداة التعريف (Martin & Jurafsky, 2009, p. 137)، لكن الأعمال الحديثة في مجال حوسبة اللغة، سواء تلك التي تتعلق باللغة الإنجليزية أو غيرها من اللغات توسعت في استخدام الأصناف؛ لتلبي الحاجة الحقيقية لحوسبة اللغة، فعلى سبيل المثال استخدم (Marcus, Santorini, & Marcinkiewicz, 1993) 45 رمزا لوسم كلمات اللغة الإنجليزية في حين استخدم (Stanford

(tagger, 2019) 29 وسما للكلمات العربية، واستخدم (Zeroual & Abdelhak, 2016) 11 وسماً للغة العربية وهي: فعل، مثل: (كتب)، واسم، مثل: (مدرسة)، واسم علم، مثل: (محمد)، وضمير، مثل: (هي)، وصفة، مثل: (جميل)، وظرف، مثل: (بعد، فوق)، وأداة، مثل: (إلى، ذلك، الذي)، وحروف مقطعة ورد ذكرها بداية سور القرآن الكريم، مثل: (الم، طه، كهيعص)، وحرف صوت، وتسمى في العربية أسماء الأفعال، مثل: (آه، هيهات)، وأخرى، وهي تحتاج إلى تعريب أو أن تصبح كلمات مشتقة، مثل: (أوبك، مانشستر)، وعلامة ترقيم، مثل النقطة: (.). والتقسيم السابق الذي ورد عند (Zeroual & Abdelhak, 2016) فيه نظر: إذ يصعب حصر العربية بأحد عشر وسما، كونها مليئة بالتراكيب، وأبواب الصرف متعددة، وذكر القدماء وسما لحرف الصوت، وهو أسماء الأفعال، فقد تكون اسم فعل ماض، مثل: شتان، أو اسم فعل مضارع، مثل: آه، واسم فعل أمر، مثل: حيَّ، وهذه أسماء الأفعال يصعب التمييز بينها إلا في السياق، فقد تكون كلمة "عليكم" اسم فعل أمر بمعنى إلزم، كقولنا: عليكم الاجتهاد، وقد تكون مكونة من حرف جر ومجرور، كقولنا: السلام عليكم، وقد تفسر حسب الوقف والابتداء، فإن قلنا: "قُلْ تَعَالَوْا أَتْلُ مَا حَرَّمَ رَبُّكُمْ عَلَيْكُمْ أَلَّا تُشْرِكُوا بِهِ شَيْئًا وَبِالْوَالِدَيْنِ إِحْسَانًا"، فيمكن اعتبارها اسم فعل أمر إن تم الوقف على ربكم، أو حرف جر ومجرور إن تم الوقف عليها، ولذلك فتعتبر مسألة اختيار مجموعة مناسبة لوسم النصوص من الأمور البحثية المعاصرة في مجال حوسبة اللغة العربية، فقد أشارت (Khoja, 2001) إلى أهمية هذا الأمر لتحقيق الفائدة المرجوة من الوسم؛ لأن الكلمات التي يتم وسمها سواء على مستوى الأسماء أو الأفعال أو الحروف، تتبع مجموعتين يطلق عليهما المجموعة المغلقة (closed class) والمجموعة المفتوحة (open class). والفرق بين المجموعتين هو أن المجموعة المغلقة تشمل عدداً محدداً من الكلمات، مثل: حروف الجر، وأما المجموعة المفتوحة فتقبل كلمات جديدة، مثل: كلمة ايفون (iPhone) في حالة الاسم، وفاكس (fax) في حالة الفعل. ومن المفيد أن نشير إلى أن بناء مجموعة الوسم يتعلق بالدرجة الأولى بالتطبيق الذي نريد استخدامه بعد عملية الوسم، فالأمر يرجع إلى مطور النظام ومدى الحاجة إلى تنوع الصيغ الصرفية في العمليات اللاحقة كالتحليل اللغوي مثلاً.

ولعل من صعوبات هذا البحث وسم النص العربي كون اللغة العربية تسمح بتجاهل الحركات، مثل: الفتحة والضمة والكسرة والتنوين المتعلق بالضم أو الفتح أو الكسر (أو ما يعرف بالتشكيل)، مما يؤدي إلى إمكانية أن تأخذ الكلمة الواحدة أكثر من شكل تبعاً لموقعها في السياق، فعندما نقول: (ما أحسنَ زيدٌ) فإن المعنى هنا واضح، وهو نفي الاحسان عن زيد، وعندما نقول: (ما أحسنُ زيدٍ؟) فإننا نسأل عن الشيء الذي أحسن فيه زيد، وإن قلنا: (ما أحسنَ زيدا) أفادت التعجب، واحتمال وسم الكلمة بأكثر من بنية صرفية معتمد على المعنى والسياق: مثل كلمة عليكم السابق ذكرها، واختلاف وظيفة الحرف حسب السياق، فقد تكون "حتى"، حرف جر، أو ظرفية أو حرف عطف أو حرف نصب، وقد كان من السهل فهم الجملة السابقة "ما أحسن زيد" بوجود التشكيل، وتصبح غامضة إذا تم تجاهل التشكيل، ومن مثال صعوبة البحث كذلك: فهم اللغة وخاصة عندما يتعلق الأمر بأنظمة الحاسوب، فعندما نقول "ما جاءني رجل واحد" فهل نفهم أنه جاء أكثر من رجل، أم أنه جاء الضعفاء فقط أم أن المقصود هو النفي، فما جاء رجل واحد ولا أكثر ولا قوي ولا ضعيف، وإن قلنا: مرض

سعيد وأخوه في سفر، فهل المريض سعيد أم كلاهما، وهل المسافر كلاهما أو أخو سعيد، ومن هنا فإن دور نظام الوسم هو إسناد الصيغة الصرفية الصحيحة لكل كلمة في الجملة بحيث يتم فهم المعنى المراد وإزالة الغموض قدر الإمكان، وكذلك فإن الكلمة الواحدة قد تأخذ أكثر من معنى تبعاً لاختلاف التشكيل مثل كلمة (بَر، بُر، بِر). ومن المفيد أن نشير إلى أن الشخص الناطق بالعربية يستطيع بكل سهولة ويسر أن يميز الكلمات حسب موقعها في السياق، لكن هذا الأمر غير ممكن بالنسبة لبرامج الحاسوب التي تعطي النتائج بناء على ما نزودها به من معلومات يطلق عليها بيانات التدريب.

تأتي أهمية هذه الدراسة لأنها تناولت نظام ستانفورد لدراسة أداء هذا البرنامج في وسم النص العربي وتحديداً نص القرآن الكريم. وللتوضيح فإن الشكل 1 يوضح مجموعة من الأمثلة من اللغة العربية المعاصرة (Modern Standard Arabic MSA) ويبين كيفية تعامل هذا النظام مع الكلمات تبعاً لموقعها في السياق، فتم فحص الجمل: (ذهب سعيد إلى المدرسة، حصل الفائز على قلادة من ذهب، يوم سعيد نتمناه لكم) فالشكل يوضح مخرجات نظام ستانفورد، وأما بالنسبة للرموز المستخدمة فسوف نوضحها في القسم الرابع من هذا البحث، وما نشير إليه في الرموز التالية في المثال، وهي "VBD"، "NN"،"NNP"،"JJ". إن الرمز "VBD" يستخدم للإشارة إلى الفعل الماضي، والرمز "NN" يستخدم للإشارة إلى الاسم، والرمز "NNP" يستخدم للإشارة إلى الاسم العلم، والرمز "JJ" يستخدم للإشارة إلى الصفة. فالملاحظ أن مخرجات النظام صحيحة بالنسبة لكلمة "ذهب" في الجملة الأولى وفي الجملة الثانية. وفي الجملة الأولى تم الإشارة الى كلمة "ذهب" على أنها فعل ماض من خلال استخدام الرمز "VBD" في حين تم استخدام الرمز "NN" في الجملة الثانية للإشارة إلى أنها اسم، وأما بالنسبة لكلمة "سعيد" فقد وسمت في الجملة الأولى على أنها "NNP" اسم علم، وفي الجملة الثالثة على أنها "JJ" صفة، وفي كلتا الحالتين فإن الوسم صحيح.

<table>
<tr><th colspan="3">المدخلات (مجموعة من جمل اللغة العربية المعاصرة)</th></tr>
<tr><td>الجملة الأولى</td><td>الجملة الثانية</td><td>الجملة الثالثة</td></tr>
<tr><td>ذهب سعيد الى المدرسة</td><td>حصل الفائز على قلادة من ذهب</td><td>يوم سعيد نتمناه لكم</td></tr>
<tr><td colspan="3">المخرجات (الوسم باستخدام نظام ستانفورد)</td></tr>
<tr><td>ذهب /VBD سعيد /NNP الى/IN المدرسة /DTNN</td><td>حصل/VBD الفائز/DTNN على /IN قلادة /NN من /IN ذهب/NN</td><td>يوم/NN سعيد/JJ نتمناه /VBP لكم /VBG</td></tr>
</table>

شكل (1): أمثلة على مخرجات نظام ستانفورد لوسم النص العربي.

كما هو موضح في الشكل رقم 1، فإن وسم النص يتطلب الأخذ بعين الاعتبار تسلسل الكلمات للخروج بأفضل نتيجة ممكنة لتحديد الكلمة المراد وسمها: اسم أو فعل أو حرف، وهو ما يعرف بأهمية السياق في فهم بنية الكلمة ومعناها وتحديد دلالتها، ومن المعلوم أن لكل من الاسم والفعل

والحرف تفرعات تظهر الحاجة اليها كلما كان هناك حاجة إلى تفصيل هذه التفرعات في التطبيق المراد بناؤه. إن التحدي الحقيقي في مسألة وسم النصوص يكمن في إمكانية وسم الكلمة الواحدة بأكثر من وسم – تبعاً لوظيفة الكلمة في الجملة - كما لاحظنا في كلمة "ذهب" وكلمة "سعيد" في الشكل رقم 1 أعلاه. وكما هو ملاحظ فإن وسم النص يتعلق بسلسلة من الكلمات في جملة معينة أو نص كامل ولا يأخذ كل كلمة على انفراد كما هو الحال في المحلل الصرفي الذي يعالج النص كلمة كلمة، أي بشكل مستقل ليعطي مشتقات الكلمة وجذرها أو غير ذلك من المعلومات. فالتحليل الصرفي يتعلق ببنية الكلمة من حيث: الجذر، السوابق، اللواحق، التصريفات، ووسم النص يتعلق بالجملة كاملة ودور كل كلمة فيها.

هناك أكثر من طريقة لوسم النصوص، ومن هذه الطرق الطريقة الإحصائية بالاعتماد على نماذج ماركوف المخفية (Hidden Markov models) وطريقة القواعد المعرفية أو يعرف بِـ (knowledge base)، أي باستخدام مجموعة من القواعد التي يحددها علماء اللغة. بالنسبة لطريقةَ نماذج ماركوف المخفية فهي طريقة فعالة في هذا المجال، حيث أنها تعتمد على بناء العلاقة الاحتمالية بين سلاسل الكلمات الموسومة (annotated words) في بيانات التدريب لاستخدامها في عمليات التنبؤ. إن العلاقة الاحتمالية في نماذج ماركوف المخفية تقوم على مبدأ حساب جميع الاحتمالات الممكنة لظهور كلمة معينة ضمن وسم معين (emission probability) واحتساب أن يأتي وسم معين بعد وسم آخر وهو ما يطلق عليه (transition probability). ولوسم الكلمة يتم استخدام الاحتمالين السابقين مع مجموعة الوسم للخروج بأعلى احتمالية واختيار الصنف بناء على ذلك. يتم الحصول على الاحتمالات المستخدمة في نماذج ماركوف المخفية من خلال عملية التدريب (training) على مجموع كبيرة من النصوص الموسومة بشكل يدوي، وأما بالنسبة للطريقة الثانية أي طريقة القواعد، فيتم استخدام قواعد خاصة يتم تعريفها مسبقاً من قبل المختصين في اللغة ولا تتطلب مجموعة تدريب كما هو الحال في نماذج ماركوف المخفية.

في هذه الدراسة سوف نستعرض أداء نظام ستانفورد لوسم النص العربي والتركيز على دقة النظام بالنسبة لحروف الجر، إذ إنها مجموعة مغلقة ومن الممكن قياس الأداء بخلاف المجموعات المفتوحة مثل الأسماء والأفعال. كما تم دراسة الوسم المستخدم لحرف الجر ومدى ملاءمته لجمل اللغة العربية. لقد تم اختيار نص القرآن الكريم لفحص نظام ستانفورد وذلك بسبب تنوع الكلمات الواردة في القرآن الكريم خصوصاً فيما يتعلق بحروف الجر، فقد أشار (Mahafdah, Omar, & Al-Omari, 2014) إلى التحديات التي تواجه أنظمة الوسم عند تطبيقها على كلمات القرآن بسبب تنوع كلمات القرآن ووجود الكثير من الكلمات التي لم يسبق للنظام التعامل معها خلال مرحلة التدريب، خصوصاً أن نظام ستانفورد يتعلق باللغة العربية المعاصرة، ونشير هنا إلى أن هذه الدراسة تتعلق بالتراكيب الصرفية (مثل الاسم، الفعل، والحرف) ولا تتطرق للتراكيب النحوية (مثل المبتدأ والخبر، الفعل والفاعل، الخ).

بالنسبة لطريقة التعامل مع الكلمات من أجل وسمها فهناك طريقتان، **الأولى** إعطاء الوسم بعد تقسيم الكلمة إلى المقاطع التي تتركب منها ثم إعطاء كل مقطع وسم معين، وأما الطريقة **الثانية** فيتم إعطاء الوسم لكامل الكلمة بما معها من إضافات كالحروف والضمائر (الاقتران) وهو ما يسمى

بالمورفيم المقيد، وبالنسبة لطريقة تقسيم الكلمة إلى مقاطع لتحديد الأوسمة لكل مقطع فقد تم استخدامها في أعمال (Diab, Hacioglu, & Jurafsky, 2004)، و(Habash & Rambow, 2005)، وقدم (Al Shamsi & Guessoum, 2005) نظاماً لوسم الكلمات العربية مع تحليل الكلمة الى المقاطع المكونة لها، وتم استخدام مجموعة خاصة مكونة من 55 وسماً، وأما فيما يتعلق بإعطاء الوصف للكلمة دون تقسيمها إلى مقاطع فقد قدم (El Hadj, Al-Sughayeir, & Al-Ansari, 2009) نظاما لوسم كلمات العربية باستخدام مجموعة مكونة من (13 وسماً، وقدم نظام جامعة ستانفورد وسما على مستوى الكلمة (Stanford tagger, 2019).

في القسم الثاني: نناقش أهمية وسم النصوص العربية والتطبيقات الممكنة، ونستعرض في القسم الثالث الدراسات السابقة ومجموعات الوسم الواردة في أبحاث سابقة، وفي القسم الرابع نستعرض مجموعة الوسم الخاصة بنظام ستانفورد مع مجموعة أمثلة على كل وسم مستخدم، وفي القسم الخامس نشرح الطريقة المقترحة للحصول على دقة النظام لحروف الجر الواردة في القرآن الكريم يليه نتائج الدراسة في القسم السادس. أما في القسم السابع فنقدم بعض التوصيات لتحسين عملية وسم حروف الجر يليه الخلاصة والأعمال المستقبلية في القسم الثامن.

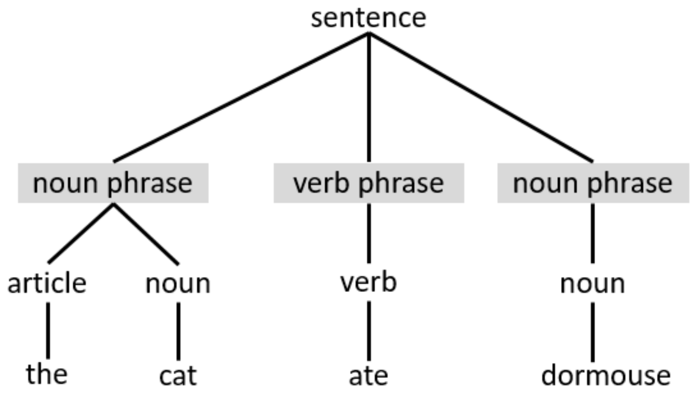
أهمية وسم النص

لوسم النصوص فوائد كثيرة في مجال معالجة اللغات الطبيعية؛ فقد أشار (Albared, Omar, & Ab Aziz, 2011) إلى أهمية دقة أنظمة وسم النصوص في أداء عمليات معالجة اللغة التي تعتمد على مخرجات أنظمة الوسم، فعلى سبيل المثال، قدم (Habash & Sadat, 2006) دراسة تشير إلى أهمية وسم النصوص وأثره الإيجابي في جودة أنظمة الترجمة الآلية، ومن فوائد وسم النصوص ما يلي:

إعطاء معلومات عن الكلمة وما يجاورها: يفيد وسم النصوص في إعطاء كمية كبيرة من المعلومات عن الكلمة وما يجاورها. إن معرفة الكلمة فيما إذا كانت اسما أو فعلا يفيد في معرفة الكلمات المتجاورة حيث إن الاسم يسبق بالمحددات مثل: "ال التعريف" كما يسبق بصفة فيما يتعلق باللغة الإنجليزية، مثل (nice car) وكما هو ملاحظ فإن الوضع مختلف بالنسبة للغة العربية، حيث يسبق الاسم الصفة (سيارة جميلة). كما يتقدم الاسم على الفعل في اللغة الإنجليزية مثل (علي يجري بسرعة == Ali runs fast) بينما يأخذ حالات مختلفة في اللغة العربية إذ يمكن أن يتقدم الاسم على الفعل كما يمكن أن يتقدم الفعل على الاسم، نظراً لأهمية المتحدث عنه، فعند التركيز على الكتابة نقول: كتب علي، وعند التركيز على علي نقول: عليٌّ كتب، فإن معرفة ما يجاور الكلمة يفيد في تشكيل المخرجات (في المدقق الاملائي او التعرف على الصوت مثلاً) بصورة أفضل حيث أن النصوص عادة ما تكون على شكل سلسلة كلمات.

يفيد وسم النصوص في عملية التحليل التركيبي (syntactic parsing) ليتم التأكد من صحة الجملة بالنسبة لقواعد اللغة. إن هذا الأمر مشابه لم يتم استخدامه في لغات البرمجة حيث يتم استخدام التحليل التركيب للتحقق من الجملة البرمجية فيما إذا كانت متوافقة مع القواعد المسموح بها في اللغة من عدمه. إن مخرجات عملية التحليل التركيبي هي عبارة عن شجرة يطلق عليها (parse tree)

أو (grammar tree) حيث تمثل هذه الشجرة العلاقات التركيبية بين عناصر الجملة. الشكل رقم 2 يمثل شجرة تحليل لنص معين في اللغة الإنجليزية.



شكل (2): شجرة تحليل لجملة في اللغة الانجليزية.

يلعب وسم النصوص دوراً مهماً في التحليل الدلالي (Semantic Role Labeling)، وهو تمثيل يعبر عن الدور المحدد الذي يأخذه كل عنصر في الجملة. من خلال التحليل الدلالي الذي يعتمد على وسم النصوص، نستطيع معرفة من الذي قام بالعمل وما هو العمل وعلى من وقع العمل وكذلك أين ومتى وقع العمل.

يمكن استخدام وسم النصوص في اكتشاف الأحداث في نص معين، وبالتالي فإن تحديد الأفعال بشكل دقيق يعتبر أمراً جوهريا في مثل هذه التطبيقات. كما يمكن استخدام وسم النصوص في اكتشاف الأسماء (Named Entity Recognition)، سواء أسماء الأشخاص أو أسماء المؤسسات.

كذلك يمكن استخدام وسم النصوص لتحديد أوجه الشبه والاختلاف بين اللغات المختلفة، والتأثير والتأثر، والأخذ والاستلاف. فعندما نطبق طريقة معينة للوسم على لغة معينة ومن ثم نستخدم نفس الطريقة على لغة ثانية فإن هذا الأمر يظهر الفرق بين اللغتين.

ومن التطبيقات الأخرى، يدخل وسم النصوص في تطبيقات تلخيص النصوص، تأليف الكلام (للتفريق بين نطق الكلمات حسب نوعها) وكذلك التعرف على الكلام آلياً (كأن يتم دمج الاسم والصفة في كلمة واحدة لأنها عادة تدمج معاً عند نطقها)، والتعرف على الكلمات عديمة التأثير في نص معين حيث يمكن تجاهلها كم هو الحال في تطبيقات تصنيف النصوص (text classification) إذ تعتبر قليلة القيمة من ناحية تصنيفية (noise) ومن الأمثلة على هذه الكلمات حروف الجر وتضاف هذه الكلمات الى قائمة تسمى (stop words list). كما يفيد الوسم في تحليل الآراء في النصوص المختلفة لمعرفة حالة الاستقطاب في المجتمع وكذلك رأي الناس في منتج

معين. إضافة إلى استخدامه في هندسة البرمجيات للكشف عن متطلبات النظام في مرحلة تحليل المتطلبات. وكذلك يستخدم في الانظمة التفاعلية -الأسئلة والاجوبة (question answering)

إن تحديد الحروف في النص يفيد في عدد من الأمور منها:

- يساعد في تحديد الأسماء والأفعال في الجملة، إذ إن الأسماء تقبل دخول حرف الجر عليها في حين أن الأفعال لا تقبل دخول حرف الجر عليها.
- إن دقة وسم الحرف يفيد في نفي صفة الاسم والفعل عنه مما يزيد في فرصة فهم النص موضع الفحص.
- إن دقة وسم الحرف يفيد في نطق الكلمات المجاورة للحرف بشكل صحيح إذ ان بعضها ينصب وبعضها يجر.
- يفيد وسم الحروف وخاصة حروف الجر في تحديد الهدف من استعمالها، فنستعمل (بِـ) لتفيد الاستعانة. ومن معاني حروف الجر بشكل عام: الابتداء، السببية، التبعيض، انتهاء الغاية في الزمان، انتهاء الغاية في المكان، المجاوزة، البعدية، الاستعلاء المعنوي، الاستعلاء الحسي، الظرفية المكانية والزمانية، التعليل، الاتصال الحقيقي، الاتصال المجازي.

الدراسات السابقة

تكشف الدراسات السابقة في مسألة وسم النص العربي إلى افتقار البحث العلمي العربي إلى مجموعة معيارية تحتوي على عدد معين من عناصر الوسم تعبر عن كلمات اللغة وتنوعها كما هو الحال في اللغة الإنجليزية، إذ يوجد عدد من المجموعات الشهيرة لوسم النصوص، والحقيقة أن وجود مجموعة معيارية يساعد في مقارنة النتائج بين الباحثين بهدف تحسين دقة نتائج الوسم وبالتالي تحسين مخرجات التطبيقات التي تعتمد في الأساس على وسم النصوص، غير أننا نعتقد أن وجود مجموعة واحدة تحتوي على عدد محدد من إشارات الوسم أمر بعيد المنال في الوقت الحالي، نظراً لعوامل مختلفة، ومنها: ضعف البحث العلمي في مجال اللغة العربية، وغزارة كلمات اللغة واشتقاقاتها، وصعوبة إعداد مدونة كبيرة تحتوي على جميع الحالات الممكنة لأقسام الكلام في اللغة العربية، فعلى سبيل المثال ولتوضيح غزارة الأقسام الفرعية للصيغ الصرفية، الشكل 3 يبين أقسام الحرف في اللغة العربية (Kholly, 1989).

حروف الجر: في، بِـ، من، لِـ، لـَ، على، إلى، عن، كـَ، حتى، حروف القسم. **حروف العطف**:(و، فـَ، أو، ثم، بل، أم). **إن واخواتها:** (أنّ، لكن، ليت، لعلّ، كأنّ، لا النافية. **حروف نصب المضارع**: (أنْ، لام التعليل، حتى، فاء السببية، واو المعية، لام الجحود، لن، كي، أذن). **حروف جزم المضارع:** (لم، لماّ، لام الأمر، لا الناهية، إنْ، اذ، ما)، **ومن الحروف الأخرى:** ربما، أن المصدرية غير الناصبة، تاء التأنيث، نون الأفعال الخمسة، نون الوقاية، حروف التسويف. وحروف أخرى.

شكل (3): أقسام الحرف في اللغة العربية.

لا بد من الإشارة إلى أن أي نظام لوسم النصوص يبدأ من خلال اختيار مجموعة رموز لتعبر عن مجموعة الوسم التي ستظهر مقترنة مع كل كلمة. وقد لوحظ في بعض الأعمال محاولة محاكاة ما هو موجود في اللغة الإنجليزية، فليس بالضرورة التشابه في السياق بين اللغة العربية والانجليزية، فكل لغة لها أسرارها في السياق. وينتج كثير من الأخطاء بسبب الترجمة الحرفية. ومن الأمثلة على ذلك: المضاف والمضاف اليه. فقولنا: (كتاب الطالب) وفي اللغة الإنجليزية (Student's book). فقد أشار (Zitouni, 2014) إلى هذه المشكلة، وبين أن هذا الأمر غير مناسب عند اختلاف اللغات، فضلاً عن لغة ثرية مثل اللغة العربية. فمن الطبيعي وجود الاختلافات بين اللغات؛ مما يحتم دراسة خصائص اللغة أولاً لاختيار الرموز المناسبة لها التي تناسب التطبيق المراد بناءه إذ يمكن زيادة عدد الرموز أو تقليلها تبعاً لحاجة النظام الذي نطوره. فعلى سبيل المثال، قدم (Freeman, 2001) مجموعة وسم تحتوي على 146 رمزا للوسم، وقدمت (Diab, Hacioglu, & Jurafsky, 2004) مجموعة وسم تحتوي على 24 رمزا، وقدم (Zeroual, Lakhouaja, & Belahbib, 2017) مجموعة وسم تحتوي على 110 وسما، وفيما يلي بعض الأعمال البحثية في مجال وسم النص العربي، وتعد ضمن الدراسات السابقة لهذا البحث، أو مساندة له، وسيذكر الباحثان افتراق هذه الأبحاث أو التقائها مع هذه الدراسة:

أشار (Zeroual, Boudchiche, Mazroui, & Lakhouaja, 2017) إلى أن إرجاع الكلمة إلى جذرها (root)، أو حذف السوابق واللواحق (stem) قد أحدث تحسنا واضحا في نتائج وسم كلمات العربية، ولعل هذا الجهد من الباحث زروال هو ما يطمح إليه الباحثان؛ لأنه يصعب وسم الكلمات التي تتصل ببعضها، وهو ما يطلق عليه في علم الأصوات المورفيم المقيد: وهو الذي يتصل بغيره من الكلمات ولا يمكن أن تغير مكانه في الجملة، مثل: أل التعريف، وتاء التأنيث، والضمائر وعلامات الجمع والتثنية.

اقترح (Abdulkareem & Tiun, 2017) طريقة لوسم التغريدات (tweets) التي عادة ما تكون مكتوبة بطريقة تختلف عن قواعد اللغة مما يؤدي إلى صعوبة في وسم كلماتها، ويذكر الباحثان أن هذه التغريدات غالبا ما تكون متأثرة من اللهجات المحكية، أو قد تكون اللغة الأم فيها غريبة بسبب البعد عن الوطن أو الاغتراب الثقافي، أو فرض اللغة للأمة القوية كما يحدث في محاولة فرض الاحتلال الإسرائيلي لغته على العرب في الداخل الفلسطيني، وعن التغريدات كذلك تعد دراسة (Albogamy & Ramsay, 2015) طريقة لتعزيز وسم التغريدات من خلال الانتفاع من أنظمة الوسم الموجودة مع إجراء بعض المعالجات القبلية والبعدية (pre- and post-processing techniques) للنصوص للتقليل من عدد الأخطاء.

أشار (Albared, Al-Moslmi, Omar, Al-Shabi, & Ba-Alwi, 2016) إلى مشكلة الكلمات المجهولة، وقدم طريقة للتعامل معها من خلال قاموس يحتوي على معلومات تتعلق بالتحليل الصرفي للكلمة، ويقترب هذا الاقتراح من دراسة (Zeroual & Abdelhak, 2016) التي أشار فيها إلى مشكلة حساب احتمالية الانتقال من صنف إلى آخر في مجموعة التدريب قليلة البيانات، وتعتبر قيمة هذه الاحتمالية مهمة عند تطبيق نماذج ماركوف المخفية؛ لأن بيانات التدريب القليلة قد

تنتج قيمة صفرية لهذه الاحتمالية مما يؤدي إلى مشاكل في التطبيق لإنتاج صنف الكلمة، ولذلك تم اقتراح أشجار القرارات (decision trees) والعمل على مجموعة وسم تحتوي على 11 وسما.

استخدم (Nabil, Atiya, & Aly, 2015) وسم الكلمات لاستخراج المفاهيم الرئيسية من النص وهو ما يطلق عليه (Key phrases extraction)، بحيث يستخدم في تطبيقات أخرى، مثل: محركات البحث، العنقدة (clustering)، تلخيص النصوص، وتحليل الآراء، ولعل هذا الوسم الذي قدمه (Nabil, Atiya, & Aly, 2015) يقارب دراسة (Aliwy, 2015) الذي بيَّن طريقة لوسم الكلمات التي تستخدم أكثر من نظام للوسم، وبعد ذلك يتم دراسة مخرجات جميع الأنظمة للخروج بأنسب وسم للكلمة.

أشار (Al-Arfaj & Al-Salman, 2015) إلى إمكانية استخدام وسم النصوص في استخراج المفاهيم في النص، وربطها مع بعضها البعض وهو ما يطلق عليه (ontology)، وهو ما يطلق عليه أهمية السياق عبر ربط اللاحق بالسابق ليتم فهمه.

قدم (AbuZeina & Alsaheb, 2013) طريقة لاكتشاف القواعد التركيبية في القرآن الكريم من خلال وسم الكلمات، ومن ثم استخدام تقنيات التنقيب عن البيانات لاكتشاف التراكيب اللغوية الأكثر شيوعاً في القرآن الكريم، وبيَّن كذلك (Al-Mashhadani & Omar, 2015) أن وسم الكلمات في التعرف على الأسماء التي تحتوي مقطعين أو أكثر، واستخدم (Hussein, 2015) وسم النصوص في كشف الانتحال (plagiarism detection)، ويظهر عبر استقرائنا للدراسات السابقة المحاولات المختلفة لوسم الكلمات عبر بيان جذورها الصرفية، غير أن هذه الدراسات تفترق عن دراسة الباحثين في اتصافها بالعموم، ودراسة الباحثين قائمة على الخصوص عبر وسم حروف الجر فقط، ويظهر كذلك افتراق هذه الدراسات عن دراسة الباحثين في اعتبار الدراسات السابقة قائمة على المورفيم المقيد لا المورفيم الحر وتحاول بعض الدراسات السابقة تجزئة المورفيم المقيد، في حين إن دراسة الباحثين قائمة على اعتبار الكلمات ذات مورفيم حر داخل السياق، فإن قلنا: عصفور على الشجرة، أو على الشجرة عصفور فإن حرف الجر "على" يستخدم بحرية داخل الجملة، ويمكن تقديمه أو تأخيره؛ مما يعطي دقة في وسمه، ولذلك فإن هذه الدراسة تعد استمرارا للجهود المبذولة في وسم الحروف غير أن نظرة الباحثين أن التعمق في وسم حروف معينة كحروف الجر أو النصب أو الجزم يعطي صدقا أكثر في الدراسة، ويقود إلى دراسات مستقبلية تتصف بالدقة أكثر من الدراسة على العموم.

ومن أشهر الأعمال في مجال وسم النصوص العربية مشروع تري بانك (Universal Dependencies, 2019) حيث قدم مدونة تحتوي على 7,664 جملة تعالج 282,384 كلمة، وتم استخدام الرمز (ADP) للإشارة إلى حرف الجر وحروف العطف، بالإضافة إلى التشكيل ووسم الكلمات، فإن مشروع تري بانك يقدم الوسم لحروف الجر في حالة الاتصال مع الكلمة (حالة الاقتران)، والجدول رقم 1 يبين عددا من الأمثلة لحروف الجر لمدونة تري بانك.

جدول (1): وصف مدونة تري بانك.

الوسم لحرف الجر المظلل	مثال من مدونة تري بانك مع التشكيل	حرف الجر
ADP	عشرين سنة **مِن** حياة التشرد	من
ADP	متوجهاً **إِلَى** ولاية	إلى
ADP	قد زار **حَتَّى** الآن الكويت	حتى
	لا يوجد مثال في مدونة تري بانك	خلا
	لا يوجد مثال في مدونة تري بانك	حاش
ADP	ما **عَدَا** هذا الحادث	عدا
ADP	كل شيء تغير **فِي** حياة المشرد	في
ADP	قد أسفر **عَن** مقتل شخصين	عن
ADP	بناء **عَلَى** خلفية	على
	لا يوجد مثال في مدونة تري بانك	مذ
ADP	الأحوال الجوية **مُنذُ** اسبوعين	منذ
	لا يوجد مثال في مدونة تري بانك	رب
ADP	للحصول على ← لِ الحصول على ← **لِ** حصول على	اللام
لِكَي ← CCONJ	لكي نعقب على بعض ← **لِكَي** عقب على بعض	كي
وَ ← CCONJ	ولا تنتظر تفويضاً ← **وَ** لا انتظر تفويضاً	واو
	لا يوجد مثال في مدونة تري بانك	تاء
كَ ← ADP	متى بدأت مهنتك كقاض ← متى بدأت مهنتك **كَ** قاض	الكاف
بِـ ← ADP	لتبلغه بأنه ورث ← لتبلغه **بِـ** أنَّ ه ورث	الباء
	لا يوجد مثال في مدونة تري بانك	لعل
متى ← DET	ومتى سيتم تسليمهم ← و **مَتَى** سيتم تسليمهم	متى

وبالنظر الى الجدول 1 أعلاه، لا بد من الإشارة إلى أن الحروف التي استخدمها "تري بانك" فيها نظر، إذ يظهر جليا عدم ذكر نماذج لبعض حروف الجر؛ علما أن العربية مليئة بمثل هذه النماذج، فيمكن اعتبار الحروف: عدا، خلا، حاشا، حروف جر إن جاء بعدها اسم مجرور، كقولنا: حضر الجمهور عدا خالدٍ، ويمكن اعتبارها أفعالا، أو حروف استثناء وتفيد المصدر كقولنا (ما عدا) ويعتمد ذلك على السياق داخل الجملة. وكلمة: "رُبَّ"، فيمكن القول: رُبَّ أخٍ لك لم تلده أمك، وما نراه حسب الجدول 1 اعتبار بعض الحروف نماذج لحرف الجر علما أنها ليست من حروف الجر، ومنها: الحرف "كي"، فقد ذكرنا قبلا أنها حرف نصب، وكذلك "لعل"، فهي من أخوات "إنَّ"،

و"متى"، اسم استفهام وليست حرفا، ويجب التمييز بين اللام التي اعتبرها حروف جر، فقد تكون مفتوحة، كقولنا: لَك، وقد تكون مكسورة، كقولنا: لِي.

اما أحدث الاعمال البحثية في مجال وسم الكلام العربي فقد قدمت من قبل (AbuZeina & Abdalbaset, 2019) حيث تم استخدام نظام ستانفورد لوسم كلمات القرآن الكريم للتعرف على أداء النظام عند وسم فعل الامر في اللغة العربية، حيث بينت تلك الدراسة ان أداء نظام ستانفورد ينخفض بشكل كبير عند استخدامه لوسم نص تراثي ديني مقارنة مع ما تم تدريب نظام ستانفورد عليه الا وهو نص حديث يتعلق باللغة العربية المعاصرة.

الصيغ الصرفية في نظام ستانفورد

كما أشرنا في مقدمة هذا البحث أننا سوف نستخدم نظام وسم الكلمات المقدم من جامعة ستانفورد، ويعد هذا النظام مبينا على أقسام الكلام التي تم اقتراحها في (Stanford tagger, 2019)، ويحتوي النظام على 29 وسم كما هو موضح مع الأمثلة في الجدول رقم 2.

جدول (2): مجموعة الاوسمة لنظام ستانفورد.

#	Tag	Meaning with examples	#	Tag	Meaning with examples
1	DTJJ صفة معرفة	DT + Adjective النفطية، الجديد، الأبيض، العزيز	16	PRP ضمائر مفرد	Personal pronoun هي، هو، نحن
2	DTJJR صفة للمقارنة معرفة	DT + Adjective, comparative الكبرى، العليا، الوسطى	17	PRP$ ضمائر جمع	Possessive pronoun هم
3	DTNN اسم معرف	DT + Noun, singular or mass المنظمة، العاصمة، المال، العلم	18	RB ظرف زمان	Adverb هناك، حيث
4	DTNNP اسم علم معرف	DT + Proper noun, singular العراق، القاهرة، المسيح	19	RP حرف نفي	Particle لم، لا، لن

... تابع جدول رقم (2)

#	Tag	Meaning with examples	#	Tag	Meaning with examples
5	DTNNS اسم جمع	DT + Noun, plural السيارات، الولايات، الثمرات	20	VB فعل امر	Verb, the imperative form ادخلوا، ادع، قل، خذ
6	IN حرف جر	Preposition or subordinating conjunction حرف جر مثل: في | حرف مصدري مثل: أنْ	21	VBD فعل ماض	Verb, past tense أعلن، قالت، كان
7	JJ صفة	Adjective جديدة، قيادية، سميع، شديد	22	VBG مصدر مشتق من فعل	Verb, gerund or present participle نية، اعتبار، قول
8	JJR صفة تتعلق بالمقارنة	Adjective, comparative أدنى، كبرى، أربى	23	VBN مبنى للمجهول	Verb, past participle يقام، يعد، تتلى
9	NN اسم نكرة	Noun, singular or mass إنتاج، نجم، أمة	24	VBP فعل مضارع	Verb, non-3rd person singular تتزايد، يعمل، يشكر present
10	NNP اسم علم	Proper noun, singular أوبك، لبنان، إبراهيم	25	VN اسم مفعول	Verb, 3rd person singular present مسجلة، مدعومة

... تابع جدول رقم (2)

#	Tag	Meaning with examples	#	Tag	Meaning with examples
11	NNS اسم جمع نكرة	Noun, plural توقعات، طلبات، درجات	26	WP اسم موصول	Wh-pronoun الذي، اللذين
12	NOUN_QUANT	Noun, quantity الربع، ثلثي، كل، بعض	27	WRB ظرف مكان	Whadverb حيث، كيف، كلما
13	CC حروف العطف	Coordinating conjunction ثم ، و، كما، بل	28	ADJ_NUM المتعلق بالعدد	Adjective, Numeric السابع، الرابعة، الثالث
14	CD ارقام	Cardinal number مئة، ألفين، ثلاث، سبع	29	UH غامض	Interjection unusual kind of word اللهم، كلا، نعم
15	DT الـ ، أسماء الاشارة	Demonstrative pronouns هذه، ذلك، هذا			

الطريقة المقترحة

نعرض في هذا القسم خطوات العمل للحصول على وسم نص القرآن الكريم، ومن ثم تقييم دقة وسم حروف الجر لتشمل الخطوات الآتية:

- الحصول على نسخة نصية الكترونية من مصحف المدينة المنورة من موقع مجمع الملك فهد لطباعة المصحف الشريف (Quran Printing Complex, 2019).
- تحميل برنامج وسم النصوص ستانفورد من الموقع (Stanford tagger, 2019) (النسخة الكاملة = Full version) لأنها النسخة التي تدعم وسم النص العربي لمجموعة كبيرة من الجمل وإعطاء النتيجة دفعة واحدة.
- تنفيذ عملية الوسم لمدونة القران الكريم التي تشمل نص القرآن الكريم كاملاً.
- تحديد جميع الكلمات التي حصلت على الوسم (IN) إذ إنه الرمز المستخدم لوسم حرف الجر.

- اجراء بعض العمليات الإحصائية لمعرفة الحالات التي تم فيها وسم حرف الجر بصورة صحيحة من أجل قياس دقة عملية الوسم. الشكل رقم 4 يوضح ملف الإدخال حيث يحتوي على نص القرآن الكريم (للاختصار، الشكل رقم 4 يعرض سورة الفاتحة وأول أربع آيات فقط من سورة البقرة)، والشكل رقم 5 يوضح مخرجات نظام وسم النصوص وهو عبارة عن ملف نصي بجميع كلمات القرآن إلا أن كل كلمة حصلت على الوسم الخاص بها، وهذا الملف هو محور العمل للحصول على النتائج في القسم التالي، قسم النتائج. نشير هنا الى أننا عرضنا في الشكل 5 الوسم لبعض الآيات الظاهرة في الشكل 4 وليس لكل كلمات وآيات القرآن.

```
الرحيم الرحمن الله بسم
العالمين رب لله الحمد
الرحيم الرحمن
الدين يوم مالك
نستعين وإياك نعبد إياك
المستقيم الصراط اهدنا
الضالين ولا عليهم المغضوب غير عليهم أنعمت الذين صراط
الم
للمتقين هدى فيه ريب لا الكتاب ذلك
ينفقون رزقناهم ومما الصلاة ويقيمون بالغيب يؤمنون الذين
يوقنون هم وبالآخرة قبلك من أنزل وما إليك بما يؤمنون والذين
```

شكل (4): جزء من مدونة القرآن الكريم.

```
بسم/NNP الله/NNP الرحمن/DTNNP الرحيم/DTNNP
الحمد/DTNNP لله/NN رب/NN العالمين/DTNNS
الرحمن/DTNNP الرحيم/DTNNP
مالك/NNP يوم/NN الدين/DTNN
إياك/VBD نعبد/VBP وإياك/NN نستعين/NNS
اهدنا/VBD الصراط/DTNN المستقيم/DTJJ
```

شكل (5): جزء من مدونة القرآن الكريم بعد عملية الوسم.

نتائج الدراسة

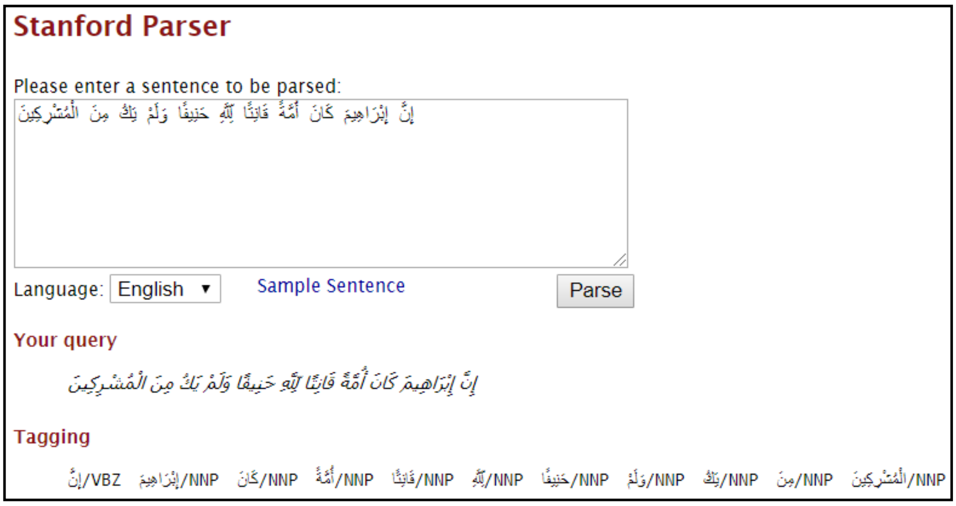
في هذا القسم تم عرض النتائج التي حصلنا عليها حسب مخرجات نظام وسم الكلمات ستانفورد، ونشير هنا إلى أن نظام ستانفورد لوسم النص العربي لا يعالج حالة الاقتران، وبذلك فإنها لن تدخل في حساب دقة حرف الجر كما هو موضح في الخطوات التالية:

- بعد وسم نص القرآن كاملاً، تم استخراج قائمة بجميع الكلمات التي حصلت وسم حرف الجر (IN) وكان عددها 7907 كلمة.
- تم إزالة التكرار والحصول على الحالات التي وسمت بحرف جر (IN) وكان العدد 32 كلمة (بعد إزالة التكرار) وهي: {من/IN، على/IN، الا/IN، في/IN، الى/IN، ان/IN، لما/IN، عن/IN، ما/IN، لو/IN، حتى/IN، فيما/IN، لولا/IN، اذ/IN، اذا/IN، للذين/IN، جمعناهم/IN، الي/IN، علي/IN، لمن/IN، جاءتهم/IN، وما/IN، ربي/IN، ربك/IN، أرأيتم/IN، عدا/IN، كي/IN، جاءه/IN، أذاقهم/IN، ربهم/IN، جاءوها/IN، بلى/IN}.
- بعد تحليل النتائج المبينة في الخطوة السابقة تبين أن عدد حروف الجر التي وسمت بشكل صحيح بلغ 8 حالات وتشمل: {من، على، في، عن، الى، حتى، عدا}، وأما الحالات التي حصلت على وسم حرف جر بشكل خاطئ فبلغت 24 حالة وتشمل المجموعة التالية: {إلا، إن، لما، ما، لو، فيما، لولا، إذ، إذا، للذين، جمعناهم، اليَ، عليَ، لمن، جاءتهم، وما، ربي، ربك، أرأيتم، جاءه، أذاقهم، ربهم، جاءوها، بل}.
- بالرغم من أن النظام قد وسم أغلب حروف الجر المنفصلة الواردة في القرآن الكريم بشكل صحيح، إلا أن النظام أخطأ في وسم عدد من الكلمات، كأن يعطي الاسم وسم حرف الجر (IN) مثل كلمة (ربي)، كما يعطي لبعض الأفعال وسم حرف الجر (IN) مثل (جاءه). الشكل رقم 6 يوضح وسم بعض الآيات القرآنية لبيان وسم كلماتها حسب نظام ستانفورد.

يحسبه/VBP الظمآن/DTNN ماء/NN حتى/IN اذا/IN جاءه/IN لم/RP يجده/VBP شيئا/NN
إنما/NN علمها/NN عند/NN ربي/IN لا/RP يجليها/VBP لوقتها/NN إلا/NN هو/PRP

شكل (6): وسم عدد من الكلمات باستخدام نظام ستانفورد.

هناك طريقتان لوسم الكلمات حسب نظام ستانفورد؛ إما تحميل البرنامج كاملا على جهاز خاص ومن ثم وسم عدد كبير من الجمل دفعة واحدة وهو ما تم تطبيقه في هذا البحث، وإما الطريقة الثانية فتشمل وسم عدد محدد من الجمل على موقع جامعة ستانفورد كما هو موضح في الشكل رقم 7. إن الوسم باستخدام الموقع الالكتروني يفيد في حالة الجمل القصيرة للتأكد من وسم كلمة معينة مثلاً، أما إجراء الدراسات الإحصائية فلا بد من تحميل البرنامج على جهاز خاص لوسم مجموعة كبيرة من الجمل دفعة واحدة.



شكل (7): وسم جملة واحدة على موقع جامعة ستانفورد.

الجدول رقم 3 يوضح تفصيل النتائج الخاصة بالحروف في القرآن الكريم بعد وسمها باستخدام نظام ستانفورد. ونلحظ انه تم وسم حرف الجر "إلى" دون اقتران بغيرها من الكلمات التي يتصل بها رسما مثل الضمائر، مثل: إليك، إليكم، إليهم، وغيرها، وهي حروف جر ولم يشر النظام الى ذلك. ويشار إلى الأمر نفسه كذلك لحرف الجر: في، وعن، وعلى. وما ورد من وسم "عدا" على انه حرف جر غير صحيح فقد ورد "عدًّا"، بتشديد الدال منصوبة، وهي اسم. وكذلك اعتبار "كي" حرف جر غير سليم لأنها حرف نصب وليست حرف جر.

جدول (3): نتائج وسم حروف الجر في القرآن الكريم.

الحرف	عددها وصحة وسمها
من	وردت 2763 مرة، وسمت حرف جر (IN) 2737 مرة، وأما بقية الحالات فقد اخذت وسم الاسم الموصول (WP) وحرف النفي (RP).
إلى	وردت 405 مرة، وسمت حرف جر (IN) 405 مرة، أي أن الوسم صحيح لجميع الحالات.
حتى	وردت 142 مرة، وسمت حرف جر (IN) 134 مرة، وأما بقية الحالات وعددها 8 فقد أخذت وسم وحرف النفي (RP).
حاش	وردت مرتان، وسمت اسم علم NNP)) في حالة ووسمت اسم (NN) في الحالة الثانية، أي ان النظام يعطي وسماً خاطئاً لهذه الكلمة.
عدا	لم ترد عدا وانما وردت عدّا وهي اسم منصوب. فهي ليست حرف جر (IN).

... تابع جدول رقم (3)

الحرف	عددها وصحة وسمها
في	وردت 1185 مرة، وسمت حرف جر (IN) 1185 مرة. أي ان الوسم صحيح لجميع الحالات.
عن	وردت 223 مرة، وسمت حرف جر (IN) 223 مرة. أي ان الوسم صحيح لجميع الحالات.
على	وردت 670 مرة، وسمت حرف جر (IN) 670 مرة. أي ان الوسم صحيح لجميع الحالات.
مذ	لم ترد في مدونة القرآن الكريم.
منذ	لم ترد في مدونة القرآن الكريم.
رب	لم ترد بضم الراء وانما بفتحها للدلالة لعل لفظ الجلالة. فقد وردت 129 مرة، وسمت اسم (NN) في جميع الحالات فالوسم صحيح.
اللام	مقترنة لا يعالجها نظام ستانفورد.
كي	ليست من حروف الجر وانما هي من حروف النصب. وردت 4 مرات، وسمت حرف جر (IN) في جميع الحالات.
واو	مقترنة لا يعالجها نظام ستانفورد.
تاء	مقترنة لا يعالجها نظام ستانفورد.
الكاف	مقترنة لا يعالجها نظام ستانفورد.
الباء	مقترنة لا يعالجها نظام ستانفورد.
لعل	من اخوات ان. وردت 3 مرات، وسمت فعل مضارع (VBP) في جميع الحالات. أي ان التصنيف خاطئ في جميع الحالات.
متى	وردت 9 مرات، وسمت ظرف مكان (WRB) في جميع الحالات. أي أن التصنيف خاطئ في جميع الحالات، وهي اسم استفهام.

كما تم تحليل أداء نظام ستانفورد بالنسبة للحرف (إنَّ) لإحصاء عدد المرات التي تم وسم فيها وسمه حرف جر (IN) وكان عدد هذه الحالات 1370 حالة من أصل 1604 حالة للحرف "إن" في القرآن الكريم. مع الأخذ بعين الاعتبار أن فحص الحرف "إن" لم يأخذ بعين الاعتبار التشكيل لأن نظام ستانفورد لوسم الكلام العربي لا يراعي التشكيل أثناء وسم الكلمات، والحالات الأخرى لوسم الحرف "إن" كان ضمن الرمز (VBP)، أي فعل مضارع، كما تم فحص نظام ستانفورد لحرف الجزم "لما" اذ وردت 74 مرة وتم وسمها في جميع الحالات على انها حرف جر "IN".

تم فحص أداء النظام لعدد من الضمائر حيث تم قياس الأداء للضمير المفرد "هو" الذي يحمل الوسم (PRP) حيث تم وسم 265 حالة بشكل صحيح. أما الضمير في حالة الجمع "هم" (PRP$) فقد تم وسم 151 حالة من أصل 261 حالة وردت في القرآن الكريم، والبقية تم وسمهم بالرمز (PRP) أي ضمير مفرد. اسم الاشارة "هذا" تم وسمه بالرمز (DT) في 190 حالة وردت في القرآن الكريم، وهي جميع الحالات. الاسم الموصول "الذي" تم وسمه بالرمز (WP) في جميع الحالات وعددها 268 حالة. كلمة "مع" تم وسمها اسم (NN) في جميع الحالات وعددها 58 حالة.

التوصيات

يوصي الباحثان بأهمية التمييز بين الحروف في اللغة العربية؛ عبر وسمها برموز خاصة، وهذا من شأنه الخروج من الإشكال التي تقع في هذا النظام، وذلك بوسم الحروف التي تؤدي وظيفة مشتركة بوسم خاص لا كما يصفها النظام، ولعل الباحثين يقترحان أن يكون الوسم كما يلي:

وسم خاص لحروف النصب؛ كونها محصورة بحروف بعينها، ومعروفة في علم اللغة العربية، وهذا من شأنه أن يبعد الخطأ في اعتبار حرف النصب "كي" حرف جر كما حصل في تحليل هذه الدراسة.

وسم حروف الاستفهام بوسم خاص؛ كونها تعطي معنى يتطلب الإجابة عليه، وهذا يحصل بوضع علامة خاصة بالسؤال بعدها، وهي علامة "؟"، وبذلك نخرج من الإشكال الذي حصل في هذا التحليل من اعتبار اسم الاستفهام "متى" بظرف مكان، فلا علاقة لها بظرف المكان.

يرى الباحثان أن توسم الكلمات التي لم ترد إلا مرة واحدة بوسم خاص؛ للخروج من الإشكال الذي ينقل هذه الكلمات إلى حقل دلالي جديد، لا علاقة لها به، ومن ذلك كما حصل من اعتبار الاسم: "عدًّا" حرف جر، ولا علاقة لها بذلك، فهي اسم، وتوسم في علم النحو بمفعول مطلق، وتندرج ضمن الأسماء المنصوبة.

يقترح الباحثان أن توسم الكلمات الجامدة في اللغة العربية بوسم خاص، للخروج من الإشكال الذي وقع فيه النظام، وهذا ينطبق على كلمة: "حاش، إذ اعتبرها النظام اسم علم، وهي ليست كذلك؛ كونها تفيد التنزيه، وهي في أصلها "فعل" بمعنى: أعْزِل، ووردت في قوله تعالى: " وقلن حاش لله ما هذا بشرا" في سورة يوسف، في الآية: 31، ولعل هذه التوصية يسيرة جدا، ويمكن تطبيقها كون اللغة العربية من اللغات المتصرفة، وليست من اللغات الجامدة، وما ورد من ألفاظ جامدة فهي محدودة جدا، وليست مطلقة.

الخلاصة والاعمال المستقبلية

بعد إنجاز هذا البحث والذي استعرضنا من خلاله أداء نظام ستانفورد لوسم حروف الجر في اللغة العربية (تم استخدام نص القرآن الكريم)، تبين لنا أن هذا الأمر بحاجة الى مزيد من البحث والتمحيص، حيث بلغت دقة النظام 100% في بعض الحروف الجر المنفصلة مثل "في" و "عن" و "على"، بشرط ورودها مجردة غير متصلة بما قبلها ولا بعدها، وهو ما يطلق عليه المورفيم

المقيد، أما حالة الاقتران فلم يتم قياسها لأن نظام ستانفورد لا يأخذها بعين الاعتبار، وانما يعطي الوسم للكلمة بغض النظر عما اقترن بها من حروف جر مثل اللام والباء، إن اللغة العربية تتميز بحالة الاقتران بخلاف اللغة الإنجليزية وهذا يستدعي معالجة هذا الأمر بصورة مناسبة خلال عملية الوسم، كما أن النظام لا يميز بين حرف الجر "مِن" واسم الاستفهام "مَن" وهي اسم موصول، وقد تكون اسم استفهام، اذ تأخذ وسما واحداً حسب نظام ستانفورد بسبب تجاهل التشكيل أثناء عملية الوسم، وكذلك الحال بالنسبة لـ "أن" و "إن". كما ان النظام لا يميز بين "حتى" كحرف جر و "حتى" الظرفية. وكذلك فلا يمكن اعتبار حروف الجر وحروف النصب ذات وسم واحد فهناك فرق بين المعنى وتوظيفها داخل السياق، ودخولها على الكلمات. فحروف الجر تدخل على الأسماء وحروف النصب تدخل على الأفعال. كما تبين أن هناك مجموعات مختلفة لرموز الوسم في الاعمال السابقة مما يدعو إلى توحيد رموز الوسم من أجل تعزيز البحث العلمي في هذا المجال وإعطاء المجال للباحثين لمقارنة النتائج وقياس دقة الأداء في الأعمال البحثية المختلفة، كما تخلص هذه الدراسة إلى أن وسم حروف الجر لا يختلف اختلافاً جوهريا سواء كانت المادة النصية من اللغة العربية المعاصرة أم إنها من النصوص التاريخية، أم أنها كانت من الشعر أو النثر، أو من اللغة الفصحى أو من اللغة العامية، ومن الملاحظات على نظام ستانفورد أنه يخلط بين المستوى النحوي والمستوى الصرفي، كما أنه بحاجة إلى التمييز بين حالات الإفراد، والتثنية، والتذكير والتأنيث، والمعرفة، والنكرة، وظرف الزمان وظرف المكان. الملحق 1 يوضح أقسام الكلام في اللغة العربية بحيث يمكن الانتفاع به في الأعمال المستقبلية (Kholly, 1989). نقترح أن يتم العمل على دراسة أداء أنظمة الوسم للنصوص المختلفة سواء التاريخية او المعاصرة كأن يتم استخدام نظام ستانفورد لدراسة اشعار العرب القديمة للوصول الى دلالة الكلمات التي أكثر من استخدامها الشعراء.

شكر وتقدير

يتقدم الباحثان بالشكر والتقدير لجامعة بوليتكنك فلسطين ولجامعة فلسطين التقنية - خضوري - فرع العروب على توفيرها بيئة العمل المناسبة لإنجاز هذا البحث.

References (Arabic & English)

- AbuZeina, D. & Abdalbaset, T. M. (2019). Exploring the Performance of Tagging for the Classical and the Modern Standard Arabic. *Advances in Fuzzy Systems*.
- Al-Arfaj, A. & Al-Salman, A. (2015, March). Arabic NLP tools for ontology construction from Arabic text: An overview. In *2015 International Conference on Electrical and Information Technologies (ICEIT)* (pp. 246-251). IEEE.
- Albared, M. Omar, N. & Ab Aziz, M. J. (2011, April). Developing a competitive HMM arabic POS tagger using small training corpora. In

Asian Conference on Intelligent Information and Database Systems (pp. 288-296). Springer, Berlin, Heidelberg.

- Albared, M. Al-Moslmi, T. Omar, N. Al-Shabi, A. & Ba-Alwi, F. M. (2016). PROBABILISTIC ARABIC PART OF SPEECH TAGGER WITH UNKNOWN WORDS HANDLING. *Journal of Theoretical & Applied Information Technology*, *90*(2).
- Abdulkareem, M. & Tiun, S. (2017). COMPARATIVE ANALYSIS OF ML POS ON ARABIC TWEETS. *Journal of Theoretical & Applied Information Technology*, *95*(2).
- Aliwy, A. H. (2015, March). Combining POS taggers in master-slaves technique for highly inflected languages as Arabic. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1-5). IEEE.
- Albogamy, F. & Ramsay, A. (2015). POS tagging for Arabic tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 1-8).
- Al-Mashhadani, M. & Omar, N. (2015). EXTRACTION OF ARABIC NESTED NOUN COMPOUNDS BASED ON A HYBRID METHOD OF LINGUISTIC APPROACH AND STATISTICAL METHODS. *Journal of Theoretical & Applied Information Technology*, *76*(3).
- AbuZeina, D. E. M. A. & Alsaheb, M. H. (2013, December). Capturing the Common Syntactical Rules for the Holy Quran: A Data Mining Approach. In *2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*. (pp. 670-680). IEEE.
- Al-hoshe, M. (2002). Altaqniya alhadeetha. Egypt, Daralfajer Lenasher Watawzeea.
- Diab, M. Hacioglu, K. & Jurafsky, D. (2004, May). Automatic tagging of Arabic text: From raw text to base phrase chunks. In *Proceedings*

of HLT-NAACL 2004: Short papers (pp. 149-152). Association for Computational Linguistics.

- Al Shamsi, F. & Guessoum, A. (2006, April). A hidden Markov model-based POS tagger for Arabic. In *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France*. (pp. 31-42).
- Freeman, A. (2001). Brill\'s {POS} tagger and a Morphology parser for {Arabic}.
- Habash, N. & Sadat, F. (2006, June). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 49-52). Association for Computational Linguistics.
- Hussein, A. S. (2015, May). Arabic document similarity analysis using n-grams and singular value decomposition. In *2015 IEEE 9th international conference on research challenges in information science (RCIS)* (pp. 445-455). IEEE.
- Zeroual, I. & Abdelhak, L. (2016, March). Adapting a decision tree based tagger for Arabic. In *2016 International Conference on Information Technology for Organizations Development (IT4OD)* (pp. 1-6). IEEE.
- Martin, J. H. & Jurafsky, D. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River: Pearson/Prentice Hall.
- Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at NAACL* (pp. 20-25).
- Kholly, M. (1989). *Altarakeeb Alshaeah*, Jordan, Dar Elfalah Lenasher Watawzeea.

- Mahafdah, R. Omar, N. & Al-Omari, O. (2014). Arabic Part of speech Tagging using k-Nearest Neighbour and Naive Bayes Classifiers Combination. *JCS*, *10*(9), 1865-1873.
- Marcus, M. Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank.
- Nabil, M. Atiya, A. F. & Aly, M. (2015, April). New approaches for extracting arabic keyphrases. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)* (pp. 133-137). IEEE.
- Habash, N. & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 573-580).
- Quran Printing Complex. (2019). Retrieved on 16 June 2019 from: https: //www.qurancomplex.org/
- Zeroual, I. Lakhouaja, A. & Belahbib, R. (2017). Towards a standard Part of Speech tagset for the Arabic language. *Journal of King Saud University-Computer and Information Sciences*, *29*(2), 171-178.
- Zeroual, I. Boudchiche, M. Mazroui, A. & Lakhouaja, A. (2017, March). Developing and performance evaluation of a new Arabic heavy/light stemmer. In *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications* (p. 17). ACM.
- Stanford tagger, Retrieved on 16 June 2019 from: https://nlp.stanford.edu/software/tagger.shtml.
- Universal Dependencies. (2019). Retrieved on 16 June 2019 from: https://universaldependencies.org/
- Zitouni, I. (Ed.). (2014). *Natural language processing of Semitic languages* (pp. 299-334). Berlin: Springer.
- El Hadj, Y. Al-Sughayeir, I. & Al-Ansari, A. (2009, April). Arabic part-of-speech tagging using the sentence structure. In *Proceedings of*

the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt.

ملحق رقم 1

الصيغ الصرفية للغة العربية (Kholly, 1989)

الأسماء	الأفعال	الحروف
في حالة الجمع، في حالة التثنية، في حالة الافراد، المبني، المعرب، المذكر، المؤنث، المحايد، المعرف بال، النكرة المضاف الى معرفة، الضمير، الاسم الموصول، اسم الاشارة، العلم، الاسم النكرة، الاسم المعرفة، المصدر، اسم المرة، اسم الهيئة، اسم الفاعل، الصفة المشبهة، اسم المفعول، افعل التفضيل، صيغة المبالغة، اسم المكان، اسم الزمان، اسم الالة، ضمير المتكلم، ضمير المخاطب، ضمير الغائب، الضمير المتصل، الضمير المنفصل، جمع المذكر السالم، جمع المؤنث السالم، جمع التكسير، الاسماء الخمسة، النسبة، التصغير، المقصور، المنقوص، الممدود، الاسماء المرفوعة، أسماء المنصوبة، الاسماء المجرورة، الفاعل المذكر، الفاعل المؤنث، نائب الفاعل، المبتدأ، خبر المبتدأ، ضمير الشأن، ضمير الفصل، اسم كان، اسم كاد، خبر أن، خبر لا النافية للجنس، التابع المرفوع، المفعول المطلق، نائب المفعول المطلق، المفعول به المذكر، المفعول به المؤنث، المفعول به لعامل مذكر، المفعول به على التحذير، المفعول به على الاغراء، المفعول به على الاختصاص، المفعول به على الاستغاثة، المفعول به على الندبة، المنادى، المرخم، المفعول لأجله، ظرف الزمان، ظرف المكان، المفعول معه، الحال، التمييز، خبر كان وأخواتها، اسم ان واخواتها، اسم لا النافية للجنس، التابع المنصوب، المضاف اليه، المجرور بحرف الجر، التابع المجرور، البدل، الصفة، التوكيد اللفظي، التوكيد المعنوي، المعطوف	الفعل الناقص، الفعل التام، الفعل الماضي، الفعل المضارع، فعل الأمر، الفعل اللازم، الفعل المتعدي، الفعل المعلوم، الفعل المجهول، الفعل ذو الفعل الظاهر، الفعل ذو الفعل المستتر، الفعل المتعدي لمفعول واحد، الفعل المتعدي لمفعولين، الفعل المتعدي لثلاثة مفاعيل، الفعل المضارع المنصوب، الفعل المضارع لمجزوم، الفعل المضارع المرفوع، الفعل الثلاثي المجرد، الفعل الرباعي المجرد، الفعل الرباعي المزيد، الفعل الصحيح، الفعل المعتل، كان وأخواتها، كاد وأخواتها، افعال المدح والذم، الفعل المؤكد بنون التوكيد، اسم الفعل،	حروف الجر: في، بِ، من، لِ، لـَ، على، الى، عن، كـَ، حتى حروف القسم حروف العطف:(و، فـَ، أو، ثم، بل، أم) ان وأخواتها: (أنّ، لكن، ليت، لعلّ، كأنّ) لا النافية، قد حروف نصب المضارع: (أنْ، لام التعليل، حتى، فاء السببية، واو المعية، لام الجحود، لن، كي، أذن) حروف جزم المضارع: (لم، لما، لام الأمر، لا الناهية، إنْ، اذ، ما) ربما، أن المصدرية غير الناصبة تاء التأنيث نون الأفعال الخمسة نون الوقاية حروف التسويف حروف اخرى