مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، المجلد الثامن والثلاثون، العدد الأول، 2023.
DOI: https://doi.org/10.35682/mjhsc.v38i1.603

وسم المدونات اللغوية: المفهوم والمجالات

عبدالله بن يحيى الفيفي*

ملخص

تستعرض هذه الورقة وسم المدونات اللغوية Corpus Tagging، وهو أحد الموضوعات التي قلما تناولتها الأدبيات العربية مع أهميتها للبحث العلمي في المجالين اللغوي والحاسوبي؛ إذ تُعرِّف هذه الورقة المدونات اللغوية ووسمها، ثم تستعرض عدداً من الدراسات الأجنبية التي تناولت وسم المدونات اللغوية، لكنّها لم تضع حداً واضحاً لأنواع الوسوم التي يمكن إضافتها، وهنا تأتي أهمية هذا البحث في التفريق بين ثلاثة من أنواع الوسم التي تضاف إلى المدونات اللغوية، وهي وسم المفردات (Tagging)، وهيكل النص (Mark-up)، والبيانات الوصفية (Metadata)، وتشرح الورقة أشكال كلّ نوع من هذه الوسوم، وآلية إضافته إلى المدونات اللغوية العربية مع أمثلة عليها، وتشرح كذلك آلية الجمع بين هذه الوسوم الثلاثة في مدونة واحدة، ما يسهم في زيادة ثرائها وفائدتها للباحثين في المجالين اللغوي والحاسوبي.

الكلمات الدالة: وسوم، معجم، تقنيات حاسوبية، كشاف سياقات، مداخل معجمية، مدونات لغوية، شيوع المفردات.

* جامعة الإمام محمد بن سعود الإسلامية، المملكة العربية السعودية.

تاريخ تقديم البحث: 2020/8/30. تاريخ قبول البحث: 2020/9/20م .

Corpus Tagging: Concept and Domains

Abdullah Alfaifi[*]

afaifi@gmail.com

Abstract

This paper reviews Corpus Tagging, a topic rarely explored in the Arab literature despite its importance in Linguistics and Natural Language Processing fields. This paper defines Corpus and Corpus Tagging then reviews several studies that investigated Corpus Tagging, which, nonetheless, did not set a clear borderline between the types of tags that can be added. Here comes the importance of this paper in distinguishing between three types of tags that can be added to corpus, which include adding linguistic tags for words (Tagging), marking-up text structure (Markup), and adding descriptive data to a corpus (Metadata). This paper also explains the forms of each type of these tags and the mechanism for adding them to Arabic language corpora accompanied with examples. It also describes the mechanism for combining these three types in one corpus, which contributes to making them more rich and useful for researchers in the Linguistics and Natural Language Processing fields.

Keywords: Tagging, dictionary, computational technologies, concordancer, lexical entries, corpora, words frequency.

[*] Imam Muhammad Ibn Saud Islamic University, Kingdom of Saudi Arabia.
Received: 30/8/2020. Accepted : 2/9/2020.

مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، المجلد الثامن والثلاثون، العدد الأول، 2023.
DOI: https://doi.org/10.35682/mjhsc.v38i1.603

المقدمة:

باتت المدونات اللغويّة (Corpora) إحدى الأدوات المهمّة في البحث اللغويّ الحاسوبيّ، كونها أحد مصادر البيانات المعيارية إذا تم بناؤها وفق معايير تصميم محددة وموثوقة. كما أنّ إضافة الوسوم إليها يزيد من فائدتها للباحثين، إذ تغدو أكثر ثراء بالبيانات من ناحية، مع تسهيل عمليات البحث فيها من ناحية أخرى. ولكن مع كثرة أنواع الوسوم وتعدد طرق إضافتها للمدونات اللغوية أصبح من الضروري إيجاد طرق علمية للتعامل معها والاستفادة منها لأقصى حدّ ممكن، خصوصاً مع كثرة المجالات التي تستفيد من هذه الوسوم سواء أكانت في الجانب اللغوي المعتمد على المدونات (Corpus Linguistics) أم في مجال معالجة اللغة الطبيعية (Natural Language Processing). ومع أنّ بعض الأدبيات السابقة في مجال بناء المدونات اللغوية واستعمالها تناولت موضوع الوسم، إلا أنّها لم تضع حداً واضحاً لأنواع الوسوم التي يمكن إضافتها إلى المدونات اللغوية، وما الطرق الممكن اتباعها لإضافة كلّ نوع لنصوص المدونات اللغوية، وهل يمكن إضافتها جميعاً؟ وما آلية ذلك؟ وهذا ما يسعى البحث الحالي للإجابة عنه، خصوصاً مع تزايد الاهتمام بموضوع وسم المدونات اللغوية، وما يضيفه الوسم من ثراء لها، وتسهيلاً لعمل الباحثين فيها. كما أنّ البحث يشير في ثناياه إلى بعض المواطن التي تحتاج إلى مزيد من الأبحاث من قبل اللغويين، ومنها جداول الوسوم التي لا زالت قاصرة عن شمول جميع أنواع المفردات العربية، إمّا بسبب نقلها من لغات أخرى، أو عدم وجود تصنيف دقيق وشامل لجميع أنواع المفردات العربية التي يمكن وسمها آلياً.

سؤال البحث:

يجيب هذا البحث على سؤالين رئيسين هما:

١. ما أهمّ أنواع الوسوم التي يمكن إضافتها إلى نصوص المدونات اللغوية؟

٢. كيف تضاف أنواع الوسوم إلى المدونات اللغويّة؟

مشكلة البحث:

من النادر أن نجد في الأدبيات – خصوصاً العربية – التي تناولت المدونات اللغوية معلومات تفصيلية حول أنواع الوسم أو آليات إدراجها في المدونات اللغوية، خصوصاً أنّ جزءاً من هذه

الوسوم يتعلق بالجوانب الحاسوبية التي عادة ما يعمل عليها الباحث في مرحلة بناء المدونة، والجزء الآخر يتعلق بالجوانب اللغوية التي يحتاجها الباحث عادة أثناء البحث في المدونة وتحليل نصوصها. ولهذا فإن البحث الحالي يهدف إلى تقديم شرح لأنواع الوسوم في المدونات اللغوية وآليات استخدامها مع تقديم أمثلة عملية تساعد على تطبيقها، خصوصاً في المدونات اللغوية العربية.

الإطار النظري:

بدأ الاهتمام بالمدونات اللغوية في أربعينيات القرن العشرين، لكنها لم تصبح طريقة فعالة لدراسة اللغة وتحليلها إلا بعد بضعة عقود عندما تطوّرت التقنيات الحاسوبية وسهلت للباحثين الاستفادة من هذه المدونات. كما أنّ هذا التطوّر سهل إضافة الوسوم آلياً إلى المدونات اللغوية، وهو ما جعلها مصادر لغوية ثرية للبحث في المجالين اللغوية واللغوي الحاسوبي. وفي هذا الجزء من البحث سنستعرض بشيء من التفصيل تعريف المدونات اللغوية وتعريف وسم المدونات.

تعريف المدونات اللغوية:

يعرف سينكلير (Sinclair, 1996) المدونة اللغوية بأنّها مجموعة من الأجزاء اللغوية التي جمعت ورتبت وفقاً لمعايير تصميم واضحة بغرض استعمالها كعينة من اللغة. ويعلل استعماله لعبارة "أجزاء" عوضاً عن "نصوص" بأنّ ذلك راجع إلى التقنيات المستعملة لجمع العينة؛ إذ إنّ العينات لو كانت جميعاً متساوية الطول فليست كلها نصوصاً، بل إنّ أغلبها أجزاءٌ من نصوص فُصلت عن سياقاتها الكاملة.

كما يُعرِّف المدونة اللغوية المحوسبة بأنّها المدونة الموسومة بطريقة معيارية ومتجانسة، والمحتوية على مهام استكتاب أو استنطاق مفتوحة (open-ended retrieval tasks)، مع توثيق الأجزاء اللغوية الواردة فيها من حيث أصولها ومصادر الحصول عليها (Sinclair, 1996).

ويمكننا تعريف المدونة اللغوية عموماً بأنّها: مجموعة حاسوبية من البيانات النصية الواقعية، التي جُمعت وفقاً لمعايير تصميم محددة، بغرض تحليل اللغة أو جزء منها ودراستها، مع وسم هذه النصوص بطريقة معيارية ومتجانسة، وتوثيق أصلها ومصدر الحصول عليها.

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

وحتى يكون التعريف أكثر وضوحاً، فسنشرح بعض مصطلحاته فيما يلي:

حاسوبية: المقصود بكونها حاسوبية أن تكون محفوظة في الحاسب الآلي بطريقة تسمح بقراءة هذه النصوص آلياً، فالمستندات النصية المدخلة على شكل صور عن طريق الماسح الضوئي مثلاً لا تندرج تحت هذا التعريف.

واقعية: كونها واقعية يعني أنها نتيجة سياق طبعي غير مصطنع، ولا بد هنا من التنويه إلى أنّه عندما يطلب من المتحدث باللغة كتابة نص معين بغرض إدراجه في المدونة – أو الحديث عن موضوع معين سواءٌ أكان في مقابلة أم محادثة أم عرض تقديمي أم غير ذلك مع تسجيله صوتياً أو عن طريق الفيديو – فإنّ لغة المتحدّث لا تكون واقعية بشكل كامل، كما لو كان يتكلم مع أحد أصدقائه هاتفياً على سبيل المثال (Granger, 2002).

معايير تصميم محددة: يقصد بهذا تحديد العناصر الأساسية لبناء المدونة، التي تتبع الهدف من إنشائها غالباً، وهذه العناصر كثيرة نذكر منها على سبيل المثال: (1) تحديد المستهدفين الذين ستُضم نصوصُهم للمدونة، (2) نوع المواد اللغوية، (3) النطاق الزمني والمكاني للمدونة، (4) حجم المدونة (5) منهجية جمعها (6) آلية الوسم ونوع الوسوم (وهو موضوع هذا البحث)، ويمكن الرجوع إلى سينكلير (Sinclair, 2005) لمزيد من التفاصيل حول معايير بناء المدونات اللغوية.

بغرض تحليل اللغة أو جزء منها ودراستها: هذا في الغالب هو الغرض النهائي من المدونات اللغوية؛ إذ تمثل المدونات اللغوية عينات مهمة جداً في البحث اللغوي، بسبب كونها واقعية ومحايدة في الغالب، ومبنية وفق معايير محددة تساعد على تعميم نتائج البحث على المجتمع اللغوي بدرجة مقبولة.

وسم النصوص بطريقة معيارية ومتجانسة: يعدّ وسم النصوص أحد العناصر المرتبطة بشكل وثيق بالمدونات النصية، ومن هذا المنطلق لا بدّ أن يستند وسم النصوص في مدونات المتعلمين على إحدى المنهجيات المعيارية المتّبعة في مثل هذه العملية، إضافة إلى أهمية تطبيق هذه المنهجية بشكل متجانس على جميع أجزاء المدونة.

توثيق أصلها ومصدر الحصول عليها: يقصد به المعلومات التي تؤخذ مع كلّ مادة لغوية بغرض التوثيق وتسمى البيانات الوصفية للمدونة (Corpus metadata)، أو ترويسة المدونة

(Corpus header)، ويعرفها برنارد (Burnard, 2005: 40) بأنّها "بيانات حول البيانات"، أي معلومات إضافية حول نصوص المدونة، وهي في الغالب قسمان: الأول بياناتٌ حول مؤلف النص مثل تاريخ ميلاده، وبلده، وجنسه، ولغته الأم، ومستواه التعليمي، إلى غير ذلك. والقسم الثاني معلوماتٌ حول النصّ نفسه، مثل نوعه الأدبي (مقال، أو قصة، أو رسالة أو بحث)، وشكله (مكتوب، أو منطوق)، ومكان وتاريخ تأليفه، وعدد كلماته، ونحو ذلك.

تعريف وسم المدونات:

يُعرِّف ليتش (Leech, 1997) عملية الوسم بأنّها إضافة معلومات لغويّة تفسيريّة إلى مجموعة إلكترونية من البيانات اللغوية المكتوبة أو المنطوقة، كما يعرف الوسوم – باعتبارها المنتج النهائي لعملية الوسم – بأنّها الرموز المرتبطة بالتمثيل الإلكتروني لمواد اللغة.

ويشير إلى الفرق من حيث الواقعية بين نصوص المدونة اللغوية وما يضاف إليها من وسوم؛ إذ إنّ نصوص المدونة اللغوية في الغالب طبعية، ويمكن أن تمثل مجتمع اللغة أو جزءاً منه إلى حدّ كبير، بينما توصف الوسوم المضافة إلى المدونة اللغوية بأنّها مصطنعة سواءً من قبل الباحث أو الحاسب الآلي، ولا يمكن اعتبارها جزءاً من اللغة الطبعية المبحوثة ولو تّم التعامل معها باعتبارها جزءاً من بيانات المدونة.

ويعرف لو (Lu, 2014) وسم المدونات بأنّه إضافة معلومات لغوية إلى مدونة لغوية مكتوبة أو منطوقة، كما يشير إلى أشهر أنواع هذه المعلومات، ومنها المعلومات المعجمية، والصرفية، والنحوية، والدلالية، والتواصلية أو الاستعمالية، ومعلومات حول تحليل الخطاب، وفي حال المدونات المنطوقة فيمكن أن تضاف إليها المعلومات الصوتية، والتنغيمية.

ويمكن تعريف وسم المدونات اللغوية بأنه: إضافة معلومات لغوية أو غير لغوية إلى مواد المدونة اللغوية لإثرائها بمعلومات إضافية تزيد من فائدتها أو تسهل البحث فيها وتحليل نصوصها.

وحتى يكون التعريف أكثر وضوحاً، فسنشرح بعض مصطلحاته فيما يلي:

معلومات لغويّة أو غير لغويّة: الوسوم التي تضاف للمدونة اللغوية تكون أحياناً مفردات أو جمل أو شروحات لغوية، وفي غالب الأحيان تكون عبارة عن اختصارات أو أرقام أو رموز ذات دلالات يفهمها واضعها وكذلك يفهمها الحاسب، والغرض من هذا النوع من الوسوم تسهيل البحث والتحليل باستخدام الحاسب الآلي، كما أنّها أكثر معيارية وتجانساً.

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

إثراء المدونة بمعلومات إضافية تزيد من فائدتها: يعني أنّ الوسوم التي تضاف إلى المدونة اللغوية عبارة عن معلومات لا تقل أهمية عن المعلومات الخام الموجودة فيها قبل الوسم، ولذلك فهي مفيدة جداً للباحثين، وتزيد من قيمة المدونة اللغوية كثيراً.

تسمى المدونة اللغوية بدون وسم مدونة خام (Raw corpus) وهي التسمية الأشهر، وقد تسمى كذلك مدونة صافية أو خالصة (Pure corpus)، أمّا بعد الوسم فتسمى مدونة موسومة (Tagged corpus) أو مرمَّزة (Marked up corpus) أو ذات تحشية (Annotated corpus).

الدراسات السابقة:

مع وجود كثير من المراجع التي تشير بشكل مقتضب إلى وسم المدونات اللغوية، إلا أنه يمكن تمييز عدد من المراجع التي تناولت هذا الموضوع بشيء من التفصيل، وهي مرتبة وفق صدورها:

1. وسم المدونات اللغوية: نصوص المدونات المحوسبة مصدر للمعلومات اللغوية (Corpus Annotation: Linguistic Information from Computer Text Corpora)؛ تحرير قارساید وآخرين (Garside et al., 1997).
2. بناء المدونات اللغوية: دليل الممارسات الجيدة (Developing Linguistic Corpora: a Guide to Good Practice)؛ تحرير وين (Wynne, 2005).
3. وسم اللغة الطبيعية لتعلم الآلة (Natural Language Annotation for Machine Learning)؛ تأليف بستيافسكي وستابس (Pustejovsky & Stubbs, 2013).
4. المنهجيات الحاسوبية لوسم المدونات اللغوية وتحليلها (Computational Methods for Corpus Annotation and Analysis) للمؤلف لو (Lu, 2014).
5. المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها، تحرير صالح العصيمي (2015).

وفيما يلي نبذة موجزة عن كلّ واحد منها:

1.1 وسم المدونات اللغوية: نصوص المدونات المحوسبة مصدر للمعلومات اللغوية Corpus Annotation: Linguistic Information from Computer Text) Corpora)؛ تحرير قارسايد وآخرون (Garside et al., 1997).

يضم هذا الكتاب – الذي قام على تحريره ثلاثة من أبرز المتخصصين في المدونات اللغوية – مقدمة وستة عشر بحثاً (شارك المحررون في بعضها) وثلاثة ملاحق. ويمكن تقسيم موضوعات الأبحاث إلى قسمين رئيسين: الأول وهو الأكبر من حيث العدد يدور حول الجوانب اللغوية في وسم للمدونات، كالوسم النحوي للمدونات (إضافة السمات النحوية)، ووسم العلاقات النحوية (بناء البنوك الشجرية)، والوسم الدلالي، وتحليل الخطاب ووسم العلاقات المجازية في المدونات، ومستويات الوسم، والاتساق والدقة في وسم المدونات، وبناء معايير موحدة أو دليل لوسم المدونات. بينما يتناول القسم الثاني موضوعات تتعلق بالأدوات الحاسوبية للوسم اللغوي مثل: اختيار أدوات الوسم واستعمالها وتطويرها، وبرامج الوسم متعدد المستويات، إضافة إلى أمثلة لبعض الأدوات المستعملة في هذا السياق.

وقد حوى الكتاب شروحات وأمثلة مفيدة لعمليات الوسم اللغوي للمدونات، وجداول الوسوم Tagsets، وآليات الوسم داخل النص وفق مستويات لغوية مختلفة (صرفية ونحوية ودلالية، وتواصلية، وأسلوبية، وغيرها)، وبعدة أساليب (الرسم الشجري، الأقواس، XML، الجداول، وغيرها)، غير أنّ التطور المتسارع للتقنيات الحاسوبية وأدوات معالجة اللغة آلياً تجعله متأخراً في بعض الجوانب اللغوية، كالوسوم، وكذلك التقنيات الحاسوبية سواء من حيث المنهجيات أم الأدوات.

بناء المدونات اللغوية: دليل الممارسات الجيدة (Developing Linguistic Corpora: A Guide to Good Practice)؛ تحرير وين (Wynne, 2005).

يضم هذا الكتاب كسابقه مجموعة متميزة من الأبحاث التي تدور حول بناء المدونات اللغوية بشكل عام مع تناول وسمها بوصفه أحد أهم أجزاء البناء؛ فالبحث الأول يبدأ بالحديث عن بناء المدونات اللغوية، ومن يبنيها، ولمن، وكيف تمثل المدونة عينة للغة؟ ويتناول البحث الثاني إضافة الوسم إلى المدونة اللغوية، مع استعراض سريع لأنواعه، كالوسم الصوتي، والوسم الدلالي، والوسم التداولي، ووسم تحليل الخطاب، والوسم الأسلوبي، والوسم المعجمي، إضافة إلى شموله لمعايير الوسم ومستوياته وتقييمه. ويدور البحث الثالث حول البيانات الوصفية (Metadata)، فيعرفها،

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

ويوضح مدى الحاجة إليها وآليه تحريرها وتحليلها، كما يشير إلى أنواع البيانات الوصفية مثل بيانات التعريف أو الهوية (Corpus identification) وبيانات المصدر (Corpus derivation) وبيانات الترميز (Corpus encoding). أما البحث الرابع فيتحدث عن موضوع دقيق أقرب إلى الموضوعات الحاسوبية، وهو ترميز أو تشفير ملفات المدونة اللغوية (Encoding) وذلك لضمان إمكانية التعامل معها بسهولة من قبل الباحثين، وكذلك أدوات تحليل المدونات والبحث فيها، كما يشير إلى أبرز أنواع الترميز المستعملة، ومنها UTF بأنواعه، وهو الأشهر في التعامل مع حروف أغلب اللغات البشرية. يركز البحث الخامس على المدونات المنطوقة كأحد أنواع المدونات التي تحتاج إلى معالجة خاصة، سواء أكان ذلك في جمع البيانات، أو تحويلها إلى مكتوبة مع المحافظة على السمات النطقية فيها أو في وسمها. ويتناول البحث السادس والأخير تخزين نصوص المدونات اللغوية وأرشفتها ونشرها للاستعمال العام، وأنواع صيغ الملفات المناسبة للنشر.

ويمكن القول إنّ هذا الكتاب وإن كان يمثل دليلاً عملياً مختصراً لبناء المدونات اللغوية، إلا أنّه موجه بالدرجة الأولى لمن لديه بعض الخبرة في المجالين اللغوي والحاسوبي، ويحتاج إلى بعض الإرشادات في آليات بناء المدونات اللغوية ووسمها. وهذه الإرشادات وإن كان أغلبها عاماً إلا أنّ بعضها تخصصية ودقيقة في مجالها سواء اللغوي أم الحاسوبي.

وسم اللغة الطبيعية لتعلم الآلة (Natural Language Annotation for Machine Learning)؛ تأليف بستيافسكي وستابس (Pustejovsky & Stubbs, 2013).

يشير المؤلف في بداية كتابه إلى أنّ هذا الكتاب مُصمّم ليكون مرجعاً للأشخاص المهتمين باستخدام أجهزة الحاسب للمساعدة في معالجة اللغة الطبيعية. وهو يتكون من اثني عشر فصلاً يبدأ أولها بشرح بعض الأسس، مثل أهمية الوسم وطبقاته، وتعريف المدونات اللغوية وتاريخها. يليه الفصل الثاني الذي يتحدث عن تحديد هدف الباحث وجمع عينته من البيانات لبنـــاء المدونة اللغوية. ويتحدث الفصل الثالث عن بعض أسس تحليل المدونة اللغوية، مثل تحـــليل الشـــيوع والإنقرام (N-Gram) وغيرهما. ويتناول الفصل الرابع بناء نماذج الوسم والتصنيف وأنواعها واستعمالاتها، بينما يتحدث الفصل الخامس عن آلية تطبيق نماذج الوسم على المدونات اللغوية، ويشرح الفصل السادس كيفية الوسم والحكم عليه من خـــلال بعض الاختبارات الإحصـــائية، مثل Kappa (k). يتحدث الفصل السابع عن مرحلة التدريب في تعلم الآلة (Machine Learning)، وعن بعض تقنيات تعلم الآلة المستعملة في وسم المدونات اللغوية، وكذلك بعض خوارزميات

التصنيف، ويكمل الفصل الثامن الحديث عن اختبار الأدوات وتقييم النتائج، وكذلك الفصل التاسع يعرض إلى مراجعة العمل وتعديله وعرض تقارير النتائج للمهتمين من الباحثين. ويتناول الفصلان العاشر والحادي عشر أداة TimeML لوسم المدونات كمثال على ما تم شرحه في الفصول السابقة، ويختم الفصل الثاني عشر الكتاب بالحديث عن مستقبل تقنيات الوسم المستعملة لتدريب خوارزميات تعلم الآلة التي قد تحدث تغييراً في مستقبل معالجة اللغة الطبيعية.

يتّضح ممّا سبق أنّ هذا الكتاب وإن كان يستعرض بعض الجوانب اللغوية في أوله إلا أنّه يركز أكثر على الجوانب الحاسوبية، إذ إنّه موجه في المقام الأول لتطوير تقنيات تعلم الآلة (Machine Learning)، وهي تقنيات حاسوبية مستعملة في عدة أغراض منها معالجة اللغة الطبيعية. وهو يعد من المراجع المتخصصة في موضوعه، ويناسب المتخصصين في معالجة اللغات الطبيعية آلياً أكثر من المتخصصين في اللغويات الحاسوبية؛ لما يتضمنه من موضوعات متقدمة في المعالجة الحاسوبية، وإن كان يتخللها ما يسندها من موضوعات لغوية.

المنهجيات الحاسوبية لوسم المدونات اللغوية وتحليلها (Computational Methods for Corpus Annotation and Analysis) للمؤلف لو (Lu, 2014).

يشير المؤلف في مقدمة كتابه إلى أنّ الهدف الرئيس لهذا الكتاب إعطاء مقدمة منهجية سهلة لأحدث الأنظمة الحاسوبية والبرامج التي يمكن استعمالها لوسم المدونات اللغوية آلياً أو بشكل شبه آلي، وكذلك تحليل نصوص المدونات اللغوية بعدة مستويات متنوعة.

كما يشير إلى أنّه ليس كتاباً لشرح لغويات المدونات (Corpus Linguistics)، ولذلك فإنّه لا يستهدف تعريف المدونات اللغوية، أو استعراض تاريخها، أو الحديث عن أسس تصميم المدونات اللغوية وبنائها، أو سرد المدونات التجارية المتاحة، أو الحديث عن أنواع شيوع المفردات، أو المتصاحبات، أو الأسلوبية، أو التحليل النحوي المعجمي الذي يمكن إجراؤه على المدونات اللغوية غير الموسومة باستعمال كشافات السياقات، ولا يهدف إلى الحديث عن مناهج التحليل المختلفة المستعملة سابقاً في أبحاث المدونات اللغوية.

ويحتوي الكتاب على ثمانية فصول شاملة المقدمة والختام، يتحدث الفصل الثاني عن معالجة النصوص باستعمال سطر الأوامر (Command Line) وفيه استعراض لبعض الأوامر البسيطة للعمل على الملفات النصية ومنها ملفات UTF-8، وكذلك لبعض أدوات معالجة النصوص من خلال سطر الأوامر مثل egrep، وRegular Expressions، وsed ، وawk ، وغيرها. ويتناول

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

الفصلان الثالث والرابع الوسم المعجمي (Lexical Annotation) والتحليل المعجمي (Lexical Analysis) على التوالي، ويشمل ذلك وسم أقسام الكلام (Part-of-Speech) وتصنيفات أقسام الكلام (PoS tagsets) وبعض برامج وسم أقسام الكلام (Stanford PoS Tagger)، وكذلك استخراج الأصول المعجمية (Lemmatization)، وتحليل قوائم الشيوع (Frequency Lists) والإنقرام (N-Gram) والثراء المعجمي (Lexical Richness). بنفس الطريقة خصص المؤلف الفصلين الخامس والسادس للوسم النحوي (Syntactic Annotation) والتحليل النحوي (Syntactic Analysis) مع الحديث عن بعض أدوات الوسم النحوي وتحليل العلاقات النحوية وقياس التعقيد النحوي وتحليله. ويتناول الفصل السابع التحليل الدلالي والتداولي وتحليل الخطاب، ويتحدث عن مجموعة من أدوات التحليل في هذه الجوانب الثلاثة. ويختم المؤلف كتابه في الفصل الثامن بالحديث عن بعض الاتجاهات المستقبلية في التحليل الحاسوبي للمدونات اللغوية.

المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها، تحرير صالح العصيمي (2015).

يمكن اعتبار هذا الكتاب أول إصدارات مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية حول المدونات اللغوية، وقد حوى الكتاب خمسة مباحث لخمسة مؤلفين، وأكثر ما قد يهمنا هي المباحث الثلاثة الأولى التي أوردت معلومات حول وسم المدونات اللغوية، فالمبحث الأول منها تناول تعريف المدونات اللغوية، وأنواعها، وطرق الإفادة منها بشكل عام، مع إيراده لأمثلة من المدونات، وكذلك لأنواع الوسم الممكن إضافتها للمدونات، مثل التحشية (Annotation)، وترميز هيكل المدونة (markup)*. والمبحث الثاني تحدث عن مدونات المتعلمين (Learner Corpora)، وذكر أنّ أهمّ أنواع الوسم المستعملة في هذا النوع من المدونات هو وسم الأخطاء، وأورد آلية وسمها مع أمثلة لها. أما المبحث الثالث فقد تحدث عن معايير تصميم المدونات اللغوية وبنائها، ومنها التحشية التي قسمها إلى ثلاثة أقسام: معلومات عن النص، ومعلومات عن بنية النص، ونتائج التحليل اللغوي للنصوص. وتناول المبحث الرابع نماذج تطبيقية لتحليل المدونات مطبقة على لغة الصحافة العربية. ويدور المبحث الخامس حول طرائق البحث اللغوي في المدونات العربية الحاسوبية.

* يمكن كذلك الوصول إلى نسخة مشابهة من المبحث من خلال مدونة المؤلف على شبكة الإنترنت (صالح، 2014) عن طريق الرابط الآتي: http://dr-mahmoud-ismail-saleh.blogspot.com/2014/04/blog-post_5.html.

التعليق على الدراسات السابقة:

قد يلاحظ المطلع على الدراسات الأربع التي تم استعراضها – وهي من أهم المراجع التي تناولت موضوع وسم المدونات اللغوية بشيء من التفصيل – أنّها لم تضع حداً واضحاً للتفريق بين أنواع الوسوم التي يمكن إضافتها، مع توضيح المنهجيات الممكنة لاستعمال كل نوع منها، وتطبيق ذلك كله على اللغة العربية. وهنا تأتي أهمية هذا البحث في التفريق بين ثلاثة من أنواع الوسم التي تضاف إلى المدونات اللغوية وهي الوسم (Tagging) أو التحشية (Annotation)، والترميز (Mark-up)، والبيانات الوصفية (Metadata)، مع بيان آليات إضافتها متفرقة، وكذلك الجمع بينها في مدونة واحدة لزيادة ثرائها وفائدتها للبحث اللغوي، إضافة إلى أمثلة على تطبيق هذه الآليات على نصوص عربية.

أنواع الوسم وآلياته:

على الرغم من أنّ كثيراً من المراجع في مجال المدونات اللغوية لم تفرق بين أنواع البيانات التي تضاف إلى المدونات اللغوية أو تزودنا بحدود واضحة لها، إلا أنّه من خلال ما هو متوفر من أدبيات تعرضت لهذا الموضوع، وكذلك من خلال الاستعراض العملي للبيانات المضافة إلى عدد كبير من المدونات اللغوية المكتوبة والمنطوقة، يمكننا التمييز بين ثلاثة أنواع رئيسة من الوسم وهي:

- الوسم (Tagging) أو التحشية (Annotation)
- الترميز (Mark-up)
- البيانات الوصفية (Metadata)

وفيما يلي شرح لهذه الأنواع الثلاثة وبيان بعض الآليات المناسبة لإضافتها لنصوص المدونات اللغوية:

النوع الأول: الوسم (Tagging) أو التحشية (Annotation)

تعريف الوسم

يعرف وين (Wynne, 2005) الوسم Tagging بأنه إضافة معلومات لغوية تفسيرية إلى النص اللغوي نفسه. لكننا عند النظر إلى هذه الوسوم من خلال تعريف Wynne نجد أنّ الوسوم أحياناً لا تظهر في شكل لغوي معروف ومفهوم ضمن سياق النص، بل تكون على شكل رموز، وقد تكون مكتوبة بأحرف لغوية أو أرقام أو أي رموز أخرى، وهي جميعاً ذات دلالات لغوية محددة

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

ومشروحة من خلال جدول خاص يفسر دلالات هذه الرموز. ومن هنا يمكننا إعادة تعريف الوسم بأنّه إضافة علامات – نصية أو غير نصية – إلى نصوص المدونة اللغوية؛ لإثرائها بمعلومات إضافية تزيد من فائدتها، أو تسهل البحث فيها وتحليل نصوصها.

مثال ذلك أن نضع الرمز V للدلالة على الأفعال، ونضيف له الرمز P للدلالة على الفعل الماضي، أو الرمز I للدلالة على الفعل المضارع، أو الرمز IV للدلالة على فعل الأمر، كما في الجدول الآتي:

الجدول (1) مثال لتحديد وسوم صرفية على مستويين

النوع العام	النوع الخاص	الوسم
الفعل (V)	الفعل الماضي (P)	VP
	الفعل المضارع (I)	VI
	فعل الأمر (IV)	VIV

تهدف الوسوم إلى اختصار المعلومات التي تضاف إلى المدونات اللغوية، وذلك من خلال بناء جدول يصنف هذه الوسوم ودلالاتها، وقد يشرح آلية استعمالها في بعض الحالات، مثل حالات اللبس، أو التداخل بين الوسوم، أو عند استثناء بعض الحالات من وسم محدد.

مع ملاحظة أنّ هذه الوسوم قد تكون باللغة العربية أو الإنجليزية أو أيّة لغة أخرى ما دامت مفهومة الدلالة لدى المستعمل، أو لدى الحاسب، ولكن شاع استعمال الحروف اللاتينية لسهولة التعامل معها بواسطة لغات البرمجة في الحاسب.

وإضافة إلى إثراء المدونة اللغوية وتسهيل البحث فيها، فإن الوسم اللغوي يعد جسراً مهماً لتطوير تقنيات ذكية لمعالجة اللغة البشرية، فهو خطوة مهمة في عملية تدريب أجهزة الحاسب على فهم الكلام البشري، والتعامل مع مهام مثل الإجابة عن الأسئلة، والترجمة الآلية، والتلخيص (Pustejovsky & Stubbs, 2013).

ويعد الوسم أشهر أنواع البيانات التي تضاف إلى المدونات اللغوية لعدة أسباب منها:

1- أنّها بيانات تضاف داخل النص نفسه؛ لإثرائه وزيادة الفائدة اللغوية منه.

٢- وجود قوائم جاهزة لكثير من الوسوم اللغوية التي يمكن استعمالها في المدونات اللغوية.

سهولة إضافتها في الغالب مع وجود برامج حاسوبية لبعضها؛ لتسهيل عملية إدراجها بالطريقة المناسبة.

الفرق بين الوسم والتحشية:

في الغالب يسمى هذا النوع وسماً (Tagging) وأحياناً يطلق عليه التحشية (Annotation)، ومع شيوع استعمالهما بنفس المعنى إلا أنّه قد يفرق بينهما بأنّ الوسوم عادة ما تكون عبارات أو كلمات أو رموزاً مختصرة، وذات دلالات لغوية متّفق عليها، أو موضّحة في جدول خاص يعرف بجدول الوسوم، كما أشرنا إلى ذلك سابقاً، أما التحشية، فهي في الغالب معلومات لغوية مفهومة الدلالة بشكل مباشر، وبالتالي فلا تحتاج إلى جدول يفسر دلالتها كما هو الحال مع الوسوم. وفي حال سلمنا بهذا التفريق فإنّه يجمعهما في الغالب كونهما يدلان على إضافات داخل النص نفسه بغرض إثراء المدونة وتسهيل استعمالها من قبل الباحثين أو من قبل الحاسب كما ورد في التعريف السابق، وفي هذه الحال يمكن اعتبار الرموز التي تلي المفردات في المثال الآتي وسوماً، واعتبار (السيارة=Ref) تحشية أضيفت لتوضيح عائد الاسم الموصول.

أقبل−VP خالد−NP بالسيارة−PPR+DT+NC التي−NPRRS−(السيارة=Ref)
اشتراها−VP+NPRP من−PPR صديقه−NC+NPRP

أنواع الوســم:

تستعمل في المدونات اللغوية عدة أنواع من الوسوم، لكن أشهرها الوسوم الصرفية أو النحوية لتوفر المحللات الصرفية التي تسهل وسم المدونات آليـــاً، كما تسهل تصنيف المفردات اللغوية وفق معايير صرفية أو نحوية محددة، فمثلاً جدول الوسم الذي أنشـــأه مجدي صـــوالحة (Sawalha, 2011) – وأسماه SALMA – يقسم المفردات إلى خمسة أنواع: (1) اسم (2) وفعل (3) وحرف (4) وأخرى (5) وعلامة ترقيم، وتحت كلّ نوع عدة أقسام فرعية، إذ يقسم الاسم مثلاً إلى أربعة وثلاثين نوعاً، مثل: المصدر، اسم الإشارة، الاسم الموصول، اسم الاستفهام، اسم الشرط، اسم الفاعل، اسم المفعول، اسم المكان، اسم الزمان، اسم الآلة، اسم العلم، وهكذا. كما أنّه يضيف إلى وسم هذه الأنواع وسم صفاتها مثل: التذكير والتأنيث، العدد، التعريف والتنكير، الإعراب، التصريف، وغيرها. وفيما يلي مثال لوســـم قوله تعالى: ﴿وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا﴾ (العنكبوت: 8) باستعمال وسم SALMA:

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

Word	Morphemes		Tag
wa waaṣṣaynā	وَ *wa*	*And*	p--c----------
And We have	وَصَّيْنَا وَصَّ *waṣṣay*	*Have enjoined*	v-p---mpfs-s-amohvtt&-
enjoined	نَا *nā*	*We*	r---r-xpfs-s----hn----
al-'insāna	الْإِنسَانَ الْ *al-*	*The*	r--d----------
(on) man	إِنسَانَ *'insāna*	*man*	nq----ms-pafd---htbt-s
bi- wālidayhi	بِوَالِدَيْهِ بِ *bi*	*To*	p--p----------
His parents	وَالِد *wālida*	*Parents*	nu----md-vgki---htot-s
	يْ *y*	*Both*	r---r-xdts-s---------
	هِ *hi*	*His*	r---r-msts-k---------
ḥusn^{an}	حُسْنًا حُسْن *ḥusn*	*kindness*	ng----ms-vafi---ndst-s
Kindness	اً *an*		r---k------f---------

الشكل 1: مثال للوسم الصرفي باستعمال SALMA (Sawalha, 2011)

وقد استقصى محمد (Mohammad, 2017) في كتابه البنك الشجري النحوي، مجموعة من الوسوم المبنية خصيصاً لوسم النصوص العربية، وإن كان أغلبها قد تم تطويره خارج الوطن العربي، لكنّها استعملت من قبل الكثير من الباحثين في مجالات حوسبة اللغة المختلفة، ومن هذه الوسوم على سبيل المثال:

1. وسم خوجة (Khoja, 2001).
2. وسم باكوالتر (Buckwalter, 2004).
3. وسم بييز BIES (Diab, 2007).
4. وسم بادت PragueArabic DependencyTreebank (PADT) (Smrž et al., 2008).
5. وسم القريني (Alqrainy, 2008).

وقد تكون دراسة هذه الوسوم ومقارنتها مجالاً خصباً للأبحاث العلمية أو لرسائل الدراسات العليا، إذ إنّ كثيراً منها يحتاج إلى مزيد من البحث والدراسة للتأكد من مناسبته للنصوص العربية، وشموله لجميع أنواع المفردات فيها، خصوصاً إذا عرفنا أنّ بعضها منقول من لغات أخرى، أو أنّ واضعه غير متخصص في الجوانب اللغوية، ممّا يوجب على المتخصصين اللغويين والحاسوبيين العمل معاً لتدقيقها وتيسير الاستفادة العملية منها، بما يعود بالفائدة المنشودة من استعمالها لوسم المدونات اللغوية العربية.

يمكن القول بأنّه من النادر إضافة الوسوم يدوياً إلى المدونات الكبيرة أو ما يعرف بالبيانات الضخمة (Big Data)؛ بسبب الوقت الذي تستغرقه هذه العملية، كما يصعب ضمان الاتساق بين الوسوم في جميع مفردات المدونة خصوصاً في حالة تعدد الأشخاص القائمين على إضافتها، ولهذا تستعمل في العادة المحللات الصرفية التي يمكنها تحليل المفردات وإضافة الوسوم الصرفية المناسبة لها، وهناك العديد من المحللات الصرفية العربية، منها على سبيل المثال:

1. محلل الخليل (Boudchiche et al., 2017).
2. محلل باكوالتر Buckwalter (Buckwalter, 2004).
3. محلل مداميرا MADAMIRA (Pasha et al., 2014).

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

وغيرها من المحللات الصرفية المصممة خصيصاً للغة العربية، أو لعدة لغات، منها العربية. هذه المحللات – كما أشرنا بالنسبة للوسوم – هي كذلك مجال خصب للدراسة والمقارنة، مع دراسة آليات تصنيفها للمفردات العربية، سواء ضمن الأبحاث العلمية أم الرسائل الجامعية.

```
INPUT STRING: وَوَصَّيْنَا
LOOK-UP WORD: wwSynA
* SOLUTION 1: (wawaS~ayonA) [waS~aY_1] wa/CONJ+waS~ay/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS): and + recommend/advise + we <verb>
  SOLUTION 2: (wawaSiy~nA) [waSiy~_1] wa/CONJ+waSiy~/NOUN+nA/POSS_PRON_1P
     (GLOSS): and + authorized agent/trustee + our

INPUT STRING: الْإِنسَانَ
LOOK-UP WORD: Al<nsAn
* SOLUTION 1: (Al<inosAn) [<inosAn_1] Al/DET+<inosAn/NOUN
     (GLOSS): the + human being +

INPUT STRING: بِوَالِدَيْهِ
LOOK-UP WORD: bwAldyh
  SOLUTION 1: (biwAlidiy~h) [wAlidiy~_1] bi/PREP+wAlidiy~/ADJ+hu/POSS_PRON_3MS
     (GLOSS): by/with + parental + its/his
* SOLUTION 2: (biwAlidayohi) [wAlid_1]
               bi/PREP+wAlid/NOUN+ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
     (GLOSS): by/with + parents/father and mother + his/its two

INPUT STRING: حُسْنًا
LOOK-UP WORD: HsnA
  SOLUTION 1: (Hasun~A) [Hasun-u_1] Hasun/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + be beautiful/be good + we <verb>
  SOLUTION 2: (HasunA) [Hasun-u_1] Hasun/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + be beautiful/be good + they (both) <verb>
  SOLUTION 3: (Has~an~A) [Has~an_1] Has~an/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + improve/decorate + we <verb>
  SOLUTION 4: (Has~anA) [Has~an_1] Has~an/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + improve/decorate + they (both) <verb>
* SOLUTION 5: (HusonAF) [Huson_1] Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good/beauty + [acc.indef.]
  SOLUTION 6: (HasanAF) [Hasan_2] Hasan/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good + [acc.indef.]
  SOLUTION 7: (HasanA) [Hasan_2] Hasan/NOUN+A/NSUFF_MASC_DU_NOM_POSS
     (GLOSS):  + good + two
  SOLUTION 8: (HasanAF) [Hasan_2] Hasan/ADV+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + well + [acc.indef.]
  SOLUTION 9: (Has~anA) [Has~-i_1] Has~/VERB_PERFECT+a/PVSUFF_SUBJ:3MS+nA/PVSUFF_DO:1P
     (GLOSS):  + feel + he/it <verb> us
  SOLUTION 10: (Has~nA) [Has~_1] Has~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + perception/feeling + our
  SOLUTION 11: (His~nA) [His~_1] His~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + sensation/perception + our
```

الشكل 2: مثال لمخرجات المحلل الصرفي باكوالتر Buckwalter (Sawalha, 2011)

آليات وسم النصوص اللغوية

يمكن إضافة الوسوم إلى نصوص المدونات اللغوية بأكثر من طريقة، ولكن ينبغي التنبه إلى أهمية وجود منهجية واضحة تساهم في اتساق عملية الوسم، وتسهيل التعرف على الوسوم وتمييز مدلولاتها آلياً، وذلك من أجل تسهيل البحث في المدونة اللغوية وتحليل نصوصها.

من الآليات البسيطة المتبعة في عملية وسم النصوص – التي يمكن استعمالها آلياً للنصوص الطويلة أو يدوياً للنصوص القصيرة – الفصل بين المفردة والوسم برمز مخصص مثل:

- الشرطة المتوسطة Hyphen (-)
- الشرطة السفلية Underscore (_)
- الشرطة المائلة (/) Slash

ويوضح الجدول الآتي شكل النص بعد وسمه باستعمال هذه الآلية:

الجدول (2) مثال للفصل بين المفردة والوسم برمز مخصص

الفاصل	مثال			
شرطة متوسطة	أقبل-VP	خالد-NP	مع-PPR	أخيه-NC+NPRP
شرطة سفلية	أقبل_VP	خالد_NP	مع_PPR	أخيه_NC+NPRP
شرطة مائلة	أقبل/VP	خالد/NP	مع/PPR	أخيه/NC+NPRP

ويمكن استعمال الأقواس بأنواعها لتمييز الوسم: كالأقواس الدائرية Brackets ()، أو الأقواس المربعة Square brackets []، أو الأقواس المثلثة Angle brackets <>، أو الأقواس المزخرفة Braces {}، وفي هذه الحال يمكن الفصل أو عدم الفصل بين المفردة والوسم بفراغ.

مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، المجلد الثامن والثلاثون، العدد الأول، 2023.
DOI: https://doi.org/10.35682/mjhsc.v38i1.603

الجدول (3) مثال لاستعمال الأقواس بأنواعها في وسم المفردات

الفاصـــل	مثـــــــال
أقواس دائرية	أقبل (VP) خالد (NP) مع (PPR) أخيه (NC+NPRP)
أقواس مربعة	أقبل [VP] خالد [NP] مع [PPR] أخيه [NC+NPRP]
أقواس مثلثة	أقبل <VP> خالد <NP> مع <PPR> أخيه <NC+NPRP>
أقواس مزخرفة	أقبل {VP} خالد {NP} مع {PPR} أخيه {NC+NPRP}

وفي حال وجود أكثر من وسم للمفردة الواحدة، بحيث يمثل كل وسم جزءاً من هذه المفردة، فتضاف هذه الوسوم بفاصل مختلف عن الفاصل الأصلي، كعلامة الزائد (+)، مثل:

أقبل–VP خالد–NP مع– PPR أخيه–NC+NPRP بالسيارة–PPR+NC الجديدة–DT+NC

وقد لا تضاف في حال أمكن التعرف على كل وسم على حده كالوسوم المميزة عن بعضها بحيث لا تشتبه، كما في الجدول الآتي.

الجدول (4) مثال لعدم استعمال الفاصل بين الوسوم المتميزة عن بعضها

جاء–VPMSg3 الرجل–NCMSgND
شرح الوسوم: جاء = V(فعل) P(ماض) M(مذكر) Sg(مفرد) 3(غائب) الرجل = N(اسم) C(عام) M(مذكر) Sg(مفرد) N(مرفوع) D(معرفة)

أو في حال استعمال الوسوم المكونة من رمزين فقط، كما في المثال الآتي:

الجدول (5) مثال لعدم استعمال الفاصل بين الوسوم المكونة من رمزين فقط

وسيكتبونها–PCPFPAVCPMNP
شرح الوسوم: و = PC (حرف: واو العطف) س = PF (حرف: سين الاستقبال) ي = PA (حرف: ياء المضارعة) كتب = VC (فعل: مضارع) ون = PM (حرف: علامة الجمع) ها = NP (اسم: ضمير)

من آليات الوسم الشائعة كذلك الطريقة العمودية، وذلك بتقسيم المفردات على عمود واحد، تأخذ كل مفردة سطراً مستقلاً فيه، ويضاف الوسم بعد ذلك في عمود آخر، مع مقابلة كلّ مفردة بوسمها، وتفصل بينهما مسافة جدولة Tab.

يقرأ	VI
الطالـ	DT+N
ب	C
الكتا	DT+N
ب	C
الجد	DT+N
يد	C

يميز هذه الطريقة إمكانية إضافة أكثر من عمود، وبالتالي أكثر من وسم للمفردة الواحدة، ومن ذلك إضافة عدة تحليلات صرفية للمفردة مثل الجذر Root أو الجذع Stem أو الأصل المعجمي Lemma أو غيرها من خصائص الكلمات كما في الجدول التالي.

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

الجدول (6) مثال لإضافة أكثر من وسم في الطريقة العمودية

المفردة	الجذر	الأصل المعجمي	الوسم الصرفي	وسم الحالة الإعرابية
يقرأ	ق.ر.أ	قرأ	VI	N
الطالب	ط.ل.ب	طالب	DT+NC	N
الكتاب	ك.ت.ب	كتاب	DT+NC	A
الجديد	ج.د.د	جديد	DT+NC	A

من الآليات المستعملة لوسم المدونات أيضاً لغة الترميز القابلة للامتداد Extensible Markup Language (XML)، وهي لغة معيارية تستعمل لوصف البيانات وتخزينها وفق قواعد محددة تُسهِّل استرجاع هذه البيانات عند الحاجة، وتساعد لغة XML في وصف المعلومات وخصائصها. وعند استعمالها في وسم النصوص ستبدو بشكل مشابه لما يلي:

```
</word>أقبل<word id="1" tag="VP">
</word>خالد<word id="2" tag="NP">
</word>مع<word id="3" tag="PPR">
</word>أخيه<word id="4" tag="NC+NPRP">
```

الشكل (3) مثال لنص موسوم باستعمال لغة XML

ويمكن شرح أجزاء هذا الوسم كما يلي (من اليمين لليسار):

<نوع المعلومة عنوان الخاصية الأولى ="قيمة الخاصية الأولى" عنوان الخاصية الثانية ="قيمة الخاصية الثانية"><المعلومة </نهاية المعلومة>

وهذا يعني أن هذه الآلية تسمح بإضافة أي عدد من الوسوم (الأوصاف) للمعلومة الواحدة، التي قد تتضمن رقم المفردة (ترتيبها في النص)، أو وسمها الصرفي أو النحوي أو الدلالي، أو الأصل المعجمي للمفردة، أو جذرها، أو جذعها، أو غير ذلك من الوسوم. ويمكن مثلاً استعمال لغة XML لإعادة وسم الجملة الواردة في الجدول (7) لتصبح كالتالي:

```
" pos="VI"قرأ " lemma="أ.ر.ق <Token id="1" root="
</Token>يقرأcase="N">
" pos="DT+NC"طالب" lemma="ط.ل.ب <Token id="2" root="
</Token>الطالبcase="N">
" pos="DT+NC"كتاب" lemma="ك.ت.ب <Token id="3" root="
</Token>الكتابcase="A">
" pos="DT+NC"جديد" lemma="ج.د.د <Token id="4" root="
</Token>الجديدcase="A">
```

الشكل (4) مثال لإضافة أكثر من خاصية في الوسم باستعمال لغة XML

تقييم دقة الوسم:

يركز تقييم الوسوم على تقييم المحللات نفسها التي تضيف هذه الوسوم للمفردات، وذلك من خلال عنصر رئيس وهو دقة وضع الوسوم الصحيحة لكل مفردة، إذ تُعطى المحللاتُ العينة نفسها من النصوص أو المدونات، ثم تُحلّلُ مخرجاتها – إمّا يدوياً من خلال متخصصين، أو آلياً من خلال عينة موسومة بشكل دقيق سلفاً – وتُحسبُ النتائج الصحيحة لكلّ محلل وتُقارنُ بالمحللات الأخرى.

ومن أمثلة تقييم الوسوم تقييم الباحثة الربيعة (Alrabiah, 2014) لاثنين من المحللات: الخليل (AlKhalil v.1) ومدى (MADA v3.2)، إذ وجدت أنّ دقة محلل الخليل في استخراج الجذوع تصل إلى 75.1% وفي وسم المفردات 77.6%، بينما كانت دقة محلل مداميرا في استخراج الجذوع 84.9%، وفي وسم المفردات 83.4%. من الأمثلة كذلك، المقارنة التي قام بها الباحثان العصيمي وأتويل (Alosaimy & Atwell, 2017) لسبعة من المحللات الصرفية هي: AraMorph، وAlKhalil، وAraComLex، وALMORGEANA، وElixir FM، وSarf from Arabic Toolkit Service، وQutuf، وكذلك لثمانية من أدوات وسم أقسام الكلام (Part-of-Speech) وهي: MADA+TOKAN suite، وAMIRA Toolkit، وMADAMIRA suite، وStanford POS tagger and segmenter، وMarMoT، وArabic Toolkit Service POS Tagger، و Segmentor and Part-of-speech tagger for Arabic، وFarasa. وكانت نتائج التحليل كما يلي:

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

الجدول (7) نتائج مقارنة المحللات الصرفية

وأدوات وسم أقسام الكلام في دراسة العصيمي وأتويل (Alosaimy & Atwell, 2017)

المحللات الصرفية	AraMorph	AlKhalil	AraComLex	ALMORGEANA	Elixir FM	Sarf from Arabic Toolkit Service	Qutuf
دقة التحليل	٪88	٪90	٪56	٪88	٪84	٪82	لا ينطبق
أدوات وسم أقسام الكلام	MADA+ TOKAN suite	AMIRA Tool kit	MADAMIRA suite	Stanford POS tagger and segmenter	MarMoT	Arabic Toolkit Service POS Tagger	Segmentor and Part-of-speech tagger for Arabic
دقة الوسم	٪70	٪79	٪71	٪78	٪67	٪68	٪69

Farasa
٪75

كما قارنت الباحثة القبيشي (Alqubaishi, 2020) بين اثنين من المحللات الصرفية: مداميرا (MADAMIRA) وفراسة (FARASA)، وذلك في عدة محاور: التقطيع الصرفي، ونوع الكلمة، وأصلها، والجنس، والعدد. وأظهرت النتائج تفوق محلل فراسة على مداميرا، وعللت الباحثة ذلك بوجود العديد من المعاجم المساندة له، كمعجم الجذوع، ومعجم الكلمات الوظيفيَّة، وغيرهما (ص103).

تقييم الاتساق بين الواسمين:

من المهم عند وسم عينات النصوص أو المدونات اللغوية الصغيرة يدوياً من قبل عدة أشخاص إجراء ما يعرف بقياس مدى الاتساق بين الواسمين Inter-Annotator Agreement Measurement، ويمكن حساب الاتفاق عن طريق النسبة المئوية للوسوم التي اتفق عليها المشاركون في الوسم، لكن هذه الطريقة لا تأخذ في الحسبان أنّ الاتفاق قد يكون عن طريق الصدفة بسبب تخمين الواسم أحياناً للوسم الصحيح، ومن هنا أتى المقياس الإحصائي المعروف بكوهينز كابا Cohen's Kappa (Cohen, 1960) الذي يقيس مدى اتفاق الوسوم بين اثنين أو أكثر من الواسمين مع الأخذ في الاعتبار نسبة حالات الاتفاق عن طريق المصادفة، ويرمز لهـــذا المقياس بالحرف (k)، وله عــدة صيغ رياضية لحســاب قيمته، أبسطها الصيغة التي اقترحها كابا (Cohen, 1960) وهي:

$$K = \frac{P^o - P^e}{1 - P^e}$$

- الرمز P^o يشير إلى حالات التطابق الملحوظة بين المرمزين.
- الرمز P^e يشير إلى حالات التطابق المحتملة عن طريق المصادفة.

وتكون قيمة k من صفر إلى واحد، ويمكن تقسيم هذه القيمة إلى عدة مستويات كما يلي:

- 0 = لا يوجد تطابق
- من 0.01 إلى 0.20 = تطابق ضعيف جداً
- من 0.21 إلى 0.40 = تطابق ضعيف
- من 0.41 إلى 0.60 = تطابق متوسط
- من 0.61 إلى 0.80 = تطابق كبير
- من 0.81 إلى 0.99 = تطابق كبير جداً
- 1 = تطابق تام

وتكون هذه القيم ذات دلالة إحصائية عند مستوى ($p < 0.001$).

النوع الثاني: ترميز هيكل النص (Mark-up)

وهو إضافة رموز مع معلومات نصية إلى ملفات المدونة اللغوية لوصف هياكل النصوص الواردة فيها، مثل تحديد عنوان النص، أو بداية الفقرات ونهايتها، أو بداية الجمل ونهايتها، أو تحديد فواصل الصفحات. وفي المدونات المنطوقة يستخدم الترميز لبيان بداية النطق ونهايته، أو التداخل بين الأصوات، أو ترميز الأصوات غير اللغوية كالضوضاء، أو لإخفاء المعلومات الشخصية كالأسماء ونحوها.

وتستعمل في الغالب رموز محددة لوصف هيكل النص ليسهل على المتخصصين وكذلك برامج الحاسب الآلي التعرف عليها وتمييز مدلولاتها، ومن ذلك على سبيل المثال لغة الترميز الممتدة XML التي تعتبر – كما ذكرنا سابقاً – لغة معيارية لوصف محتويات الملفات وفق قواعد محددة لجعل هذه الملفات مقروءة بشرياً وحاسوبياً.

وفيما يلي مثال لترميز هيكل النص يوضح بداية ونهاية كل من العنوان (Title) والفقرات (Paragraphs) التي يشار إليها اختصاراً بالرمز (P):

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

```
<doc ID="S938">
<text>
</title>الرحلة إلى القرية<title>
<P pid="1">اعتدت الذهاب إلى قريتي الموجودة في غرب بلدي في
الإجازات الصيفيّة، وكنت أقابل أصحابي القدامى خلال هذه الزيارات وأفرح بلقائهم والحديث
معهم، وفي أحيان كثيرة نخرج للتنزه في الأماكن التي تعودنا زيارتها أثناء صغرنا، إنها
ذكريات جميلة لا أنساها كلما زرت قريتي في الإجازة الصيفية</P>
<P pid="2"> ... </P>
<P pid="3"> ... </P>
</text>
</doc>
```

الشكل (5) مثال لترميز العنوان والفقرات باستعمال XML

ونلاحظ في المثال السابق إضافة خاصية للفقرات في الترميز وهي رقم الفقرة (Paragraph ID)، والذي يشار إليه اختصاراً بالرمز (pid)، وقد سبق الحديث عن إمكانية إضافة صفات للمعلومات اللغوية في لغة XML عند الحديث عن الوسم.

في المثال الآتي ترميز يوضح بداية ونهاية كل من العنوان والفقرات إضافة إلى الجمل (Sentences) التي يشار إليها اختصاراً بالرمز (S)، مع إضافة خاصية رقم الجملة (Sentence ID)، والذي يشار إليه اختصاراً بالرمز (sid):

```
<doc ID="S938">
<text>
</title>الرحلة إلى القرية<title>
<P pid="1">
<S sid="1">اعتدت الذهاب إلى قريتي الموجودة في غرب
بلدي في الإجازات الصيفيّة.</S>
<S sid="2">وكنت أقابل أصحابي القدامى خلال هذه
الزيارات وأفرح بلقائهم والحديث معهم.</S>
<S sid="3">وفي أحيان كثيرة نخرج للتنزه في الأماكن
التي تعودنا زيارتها أثناء صغرنا.</S>
<S sid="4">إنها ذكريات جميلة لا أنساها كلما زرت
قريتي في الإجازة الصيفية.</S>
</P>
</text>
</doc>
```

الشكل (6) مثال لترميز العنوان والفقرات والجمل باستعمال XML

أمّا المثال الآتي فهو لترميز ملف يوضح بداية ونهاية كلّ من العنوان والفقرات والجمل، إضافة إلى الكلمات (Words)* التي يشار إليها اختصاراً بالرمز (w)، مع إضافة خاصية رقم الكلمة (Word ID)، الذي يشار إليه اختصاراً بالرمز (wid):

*عند تقسيم الكلمات في نصوص المدونة اللغوية فإنّ الحد المعتبر في الغالب بين الكلمات هو الفراغ الفاصل بين كلمتين؛ إذ إنّ التطبيقات الحاسوبية في الغالب لا تجيد التعامل مع النصوص إلا على هذا الأساس رغم احتواء بعض الكلمات على أكثر من مورفيم. ويشير الثبيتي (Althubaiti, 2015) إلى ذلك عند حديثه عن حساب أحجام المدونات اللغوية بناءً على هذا الأساس بقوله: "فالكلمة هنا تعني أي مجموعة متتابعة من الرموز لا يفصل بينها فراغ، وبالتالي فإنّ بعض الكلمات – حسب هذا التعريف– قد تكون كلمة معروفة وصحيحة، مثل "كتاب"، أو قد تكون أرقاما "9730"، أو كلمات ليس لها معنى "ععخعج" أو كلمات تحوي

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

```
<doc ID="S938">
<text>
<title>
</w>الرحلة<w wid="1">
</w>إلى<w wid="2">
</w>القرية<w wid="3">
</title>
<P pid="1">
<S sid="1">
</w>اعتدت<w wid="4">
</w>الذهاب<w wid="5">
</w>إلى<w wid="6">
</w>قريتي<w wid="7">
</w>الموجودة<w wid="8">
</w>في<w wid="9">
</w>غرب<w wid="10">
</w>بلدي<w wid="11">
</w>في<w wid="12">
</w>الإجازات<w wid="13">
</w>الصيفيّة<w wid="14">
</w>.<w wid="15">
</S>
<S sid="2">
</w>وكنت<w wid="16">
</w>أقابل<w wid="17">
```

أخطاء طباعية "كعنبوت" أو كلمات تمت إضافة الكشيدة في وسطها "بســـم". وبالتالي فإن "بســـم" و"بسم" تعدان كلمتين مختلفتين بالنسبة لأدوات معالجة المدونات لاختلاف شكليهما، على الرغم من كونهما كلمة واحدة. مثل هذه الأمثلة موجودة في أغلب المدونات خصوصاً الكبيرة منها، ونسبتها إلى إجمالي حجم المدونة لا يذكر، ولكنّ الإشارة إليها لازمة لفهم معنى الكلمة التي بناء عليها يتم قياس حجم المدونة" (p. 156).

```
</w>أصحابي<w wid="18">
</w>القدامى<w wid="19">
</w>خلال<w wid="20">
</w>هذه<w wid="21">
</w>الزيارات<w wid="22">
</w>وأفرح<w wid="23">
</w>بلقائهم<w wid="24">
</w>والحديث<w wid="25">
</w>معهم<w wid="26">
</w>.<w wid="27">
</S>
<S sid="3">
</w>وفي<w wid="28">
</w>أحيان<w wid="29">
</w>كثيرة<w wid="30">
</w>نخرج<w wid="31">
</w>للتنزه<w wid="32">
</w>في<w wid="33">
</w>الأماكن<w wid="34">
</w>التي<w wid="35">
</w>تعودنا<w wid="36">
</w>زيارتها<w wid="37">
</w>أثناء<w wid="38">
</w>صغرنا<w wid="39">
</w>.<w wid="40">
</S>
<S sid="4">
</w>إنها<w wid="41">
</w>ذكريات<w wid="42">
```

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

```
</w>جميلة<w wid="43">
</w>لا<w wid="44">
</w>أنساها<w wid="45">
</w>كلما<w wid="46">
</w>زرت<w wid="47">
</w>قريتي<w wid="48">
</w>في<w wid="49">
</w>الإجازة<w wid="50">
</w>الصيفية<w wid="51">
</w>.<w wid="53">
</S>
</P>
</text>
</doc>
```

الشكل (7) مثال لترميز العنوان والفقرات والجمل والكلمات في ملف نصي باستعمال XML

يفيد ترميز هيكل النص في الوصول إلى أجزاء محددة من النص والبحث فيها، وذلك إمّا لتسريع عمليات البحث، أو عند الرغبة في التركيز على أجزاء محددة مثل: عنوان النص، أو الفقرة الأولى أو الأخيرة من النص، أو المفردات الأولى في كلّ فقرة، وهكذا.

من الأمور المهمة التي نحتاج إلى أخذها في الحسبان أنّه يمكننا الجمع بين الوسم (Tagging) والترميز (Marking-up) في ملف واحد؛ وهذا يسهل علينا توحيد ملفات نصوص المدونة التي تضاف إليها كل هذه المعلومات، ولكن الوسيلة الوحيدة لذلك هي باستعمال لغة XML في كليهما – الوسم والترميز – إذ لن يكون استعمال الطرق الأخرى في الوسم فعالاً كالفصل بين المفردة والوسم برمز مخصص أو الطريقة العمودية؛ لأنّه سيصعب التعرف عليها حينئذٍ داخل ملف تم ترميز هيكله باستعمال XML ما لم تُستعمل أدوات بحث مخصصة لذلك.

فيما يلي مثال يجمع بين الوسم والترميز في ملف واحد:

```
<doc ID="S938">
<text>
<title>
</w>الرحلة" pos="DT+NC">رحلة<w wid="1" lemma="
</w>إلى" pos="PPR">إلى<w wid="2" lemma="
</w>القرية" pos="DT+NC">قرية<w wid="3" lemma="
</title>
<P pid="1">
<S sid="1">
</w>اعتدت> " pos="VP+NPRPS"اعتاد<w wid="4" lemma="
</w>الذهاب" pos="DT+NC">ذهب<w wid="5" lemma="
</w>إلى" pos="PPR">إلى<w wid="6" lemma="
</w>قريتي" pos="NPRPS">قرية<w wid="7" lemma="
</w>الموجودة" pos="DT+NC">وجد<w wid="8" lemma="
</w>في" pos="PPR">في<w wid="9" lemma="
</w>غرب" pos="NC">غرب<w wid="10" lemma="
</w>بلدي" pos="NPRPS">بلد<w wid="11" lemma="
</w>في" pos="PPR">في<w wid="12" lemma="
</w>الإجازات" pos="DT+NC">إجازة<w wid="13" lemma="
</w>الصيفيّة" pos="DT+NC">صيف<w wid="14" lemma="
</w>." pos="PU">.<w wid="15" lemma="
</S>
</P>
</text>
</doc>
```

الشكل (8) مثال للجمع بين الوسم والترميز في ملف واحد باستعمال لغة XML

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

النوع الثالث: البيانات الوصفية (Metadata)

وهو إضافة معلومات حول بيانات المدونة، أو كما يعرفها برنارد (Burnard, 2005, p 30) بأنّها بيانات عن البيانات (data about data)، مثل: قائل النص، وتاريخه، ومصدره، ونوعه (مكتوب أو منطوق)، وأية معلومات أخرى، وتفيد هذه المعلومات الباحثين عند تحليل نصوص المدونة اللغوية، وذلك عند تركيز البحث على فئة محددة من النصوص، مثل: البحث في النصوص المكتوبة في سنة معينة، أو المكتوبة من قبل جنســـية محددة، أو المكتوبة في عام معين، وهكـــذا. وفيما يلي مثال للبيانات الوصفية المأخوذة من المدونة اللغوية لمتعلمي اللغة العربية (Alfaifi, 2015):

Text ID: S002_T1_M_Pre_NNAS_W_C
Learner Profile
Age: 25
Gender: Male
Nationality: Russian
Mother tongue: Russian
Nativeness: NNAS
No of languages speak: 5
No of years learning Arabic: 5
No of years in Arabic countries: 5
General level: Pre-university
Level of study: Diploma course
Year/Semester: Second semester
Educational institution: Arabic Inst. at Imam Uni

Text Profile
Genre: Narrative
Where produced: In class
Year of production: 2012

Country of production: Saudi Arabia

City of production: Riyadh

Timed: Yes

References use: No

Grammar book use: No

Monolingual dictionary use: No

Bilingual dictionary use: No

Other references use: No

Mode: Written

Medium: Written by hand

Length: 294 words

Text title: رحلة الحج المباركة

Text:

كتب الله لي أن أحج إلى بيته الحرام السنة الماضية. فلله الحمد والمنة على هذه النعمة العظيمة؛ لأنّ كثيراً من الناس محرومون عن هذه النعمة؛ إمّا بسبب المرض وإمّا لعدم القدرة المالية. فكانت بداية رحلتي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعمئة وألف. صباح يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكنت واحداً من بينهم. وكان بعض الطلاب مع عوائلهم. ففي المسجد تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة. بعد ذلك ركبنا الحافلة وبدأنا في السير وكان وقت الضحى. في الحافلة قلنا دعاء السفر ونصحنا مسؤول الرحلة نصائح مفيدة. ومضينا في السير حتى قرب وقت صلاة العصر فنزلنا وصلينا الظهر والعصر جمع تأخير، وتغدينا واسترحنا، ثم ركبنا الحافلة مرة أخرى واستمررنا في السير حتى وصلنا ميقات أهل نجد السيل الكبير. هناك صلينا المغرب والعشاء وفعلنا سنن الإحرام ودخلنا في الإحرام ثم اتجهنا إلى مكة المكرمة. وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الآتين من روسيا حتى استرحت عندهم. وقبل الزوال اتجهت مع أخي إلى المسجد الحرام لقضاء صلاة الجمعة. بعد الصلاة رجعنا إلى الفندق واسترحنا حتى جاءنا الليل. ففي الليل اتجهنا إلى عرفة. ووصلنا إلى عرفة قبل صلاة الفجر فأخذنا راحتنا حتى الفجر. فصلينا الفجر ثم نمنا إلى الضحى وبعد الاستيقاظ تناولنا الطعام وانتظرنا صلاة

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

الظهر. بعد صلاة الظهر دعونا لأنّه وقت فضيلة واستجابة الدعاء كما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس ركبنا الحافلة وذهبنا إلى المزدلفة فصلينا المغرب والعشاء جميعا بعد نزولنا. فبتنا بمزدلفة وبعد الفجر يوم العاشر اتجهنا إلى المنى. بعد وصولنا فعلنا أفعال اليوم العاشر واسترحنا. والأيام الثلاثة القادمة رمينا الجمار واستمعنا إلى نصائح ومواعظ مفيدة فازددنا إيمانا بحمد الله تعالى، فكانت هذه الرحلة مباركة لأنّها رحلة الطاعة والتقوى وازدياد الخير. لم أذكر كثيراً من تفاصيل الرحلة لضيق الوقت.

الشكل (9) مثال لإضافة البيانات الوصفية في ملف نصي (Alfaifi, 2015)

كما يمكن استعمال لغة XML في إضافة البيانات الوصفية، وهو ما يساعد على تنظيمها وتسهيل البحث فيها عند التحليل الآلي لنصوص المدونات اللغوية، وفيما يلي مثال للبيانات الوصفية باستعمال لغة XML:

```
<doc ID="S002_T1_M_Pre_NNAS_W_C">
<header>
<learner_profile>
<age>25</age>
<gender>Male</gender>
<nationality>Russian</nationality>
<mothertongue>Russian</mothertongue>
<nativeness>NNAS</nativeness>
<No_languages_spoken>5</No_languages_spoken>
countries> <No_years_Arabic_countries>5</No_years_Arabic_
<general_level>Pre-university</general_level>
<level_study>Diploma course</level_study>
semester> <year_or_semester>Secondsemester</year_or_
<educational_institution>Arabic Inst. at Imam
Uni</educational_institution>
</learner_profile>
```

<text_profile>
<genre>Narrative</genre>
<where>In class</where>
<year>2012</year>
<country>Saudi Arabia</country>
<city>Riyadh</city>
<timed>Yes</timed>
<ref_used>No</ref_used>
<grammar_ref_used>No</grammar_ref_used>
<mono_dic_used>No</mono_dic_used>
<bi_dic_used>No</bi_dic_used>
<other_ref_used>No</other_ref_used>
<mode>Written</mode>
<medium>Written by hand</medium>
<length>294</length>
</text_profile>
</header>
<text>
</title>رحلة الحج المباركة<title>

<text_body>كتب الله لي أن أحج إلى بيته الحرام السنة الماضية. فلله الحمد والمنة على هذه النعمة العظيمة؛ لأنّ كثيرا من الناس محرومون عن هذه النعمة؛ إمّا بسبب المرض وإمّا لعدم القدرة المالية. فكانت بداية رحلتي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعمئة وألف. صباح يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكنت واحداً من بينهم. وكان بعض الطلاب مع عوائلهم. ففي المسجد تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة. بعد ذلك ركبنا الحافلة وبدأنا في السير وكان وقت الضحى. في الحافلة قلنا دعاء السفر ونصحنا مسؤول الرحلة نصائح مفيدة. ومضينا في السير حتى قرب وقت صلاة العصر فنزلنا وصلينا الظهر والعصر

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

جمع تأخير، وتغدينا واسترحنا، ثم ركبنا الحافلة مرة أخرى واستمررنا في السير حتى وصلنا ميقات أهل نجد السيل الكبير. هناك صلينا المغرب والعشاء وفعلنا سنن الإحرام ودخلنا في الإحرام ثم اتجهنا إلى مكة المكرمة. وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الآتين من روسيا حتى استرحت عندهم. وقبل الزوال اتجهت مع أخي إلى المسجد الحرام لقضاء صلاة الجمعة. بعد الصلاة رجعنا إلى الفندق واسترحنا حتى جاءنا الليل. ففي الليل اتجهنا إلى عرفة. ووصلنا إلى عرفة قبل صلاة الفجر فأخذنا راحتنا حتى الفجر. فصلينا الفجر ثم نمنا إلى الضحى وبعد الاستيقاظ تناولنا الطعام وانتظرنا صلاة الظهر. بعد صلاة الظهر دعونا لأنّه وقت فضيلة واستجابة الدعاء كما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس ركبنا الحافلة وذهبنا إلى المزدلفة فصلينا المغرب والعشاء جميعا بعد نزولنا. فبتنا بمزدلفة وبعد الفجر يوم العاشر اتجهنا إلى المنى. بعد وصولنا فعلنا أفعال اليوم العاشر واسترحنا. والأيام الثلاثة القادمة رمينا الجمار واستمعنا إلى نصائح ومواعظ مفيدة فازددنا إيمانا بحمد الله تعالى، فكانت هذه الرحلة مباركة لأنّها رحلة الطاعة والتقوى وازدياد الخير. لم أذكر كثيراً من تفاصيل الرحلة لضيق الوقت<text_body/>.

</text>
</doc>

الشكل (10) مثال لإضافة البيانات الوصفية باستعمال لغة XML (Alfaifi, 2015)

وكما ذكرنا عند الحديث عن ترميز هيكل النص (Mark-up) حول إمكانية الجمع بين الوسم والترميز في ملف واحد، فإنّ الجمع كذلك بين الوسم (Tagging) والترميز (Marking-up) والبيانات الوصفية (Metadata) في ملف واحد ممكن عن طريق استعمال لغة XML، وفيما يلي مثال لذلك:

```
<doc ID="S002_T1_M_Pre_NNAS_W_C">
<header>
<learner_profile>
<age>25</age>
<gender>Male</gender>
<nationality>Russian</nationality>
<mothertongue>Russian</mothertongue>
```

<nativeness>NNAS</nativeness>
<No_languages_spoken>5</No_languages_spoken>
countries> <No_years_Arabic_countries>5</No_years_Arabic_
<general_level>Pre-university</general_level>
<level_study>Diploma course</level_study>
semester> <year_or_semester>Secondsemester</year_or_
<educational_institution>Arabic Inst. at Imam
Uni</educational_institution>
</learner_profile>
<text_profile>
<genre>Narrative</genre>
<where>In class</where>
<year>2012</year>
<country>Saudi Arabia</country>
<city>Riyadh</city>
<timed>Yes</timed>
<ref_used>No</ref_used>
<grammar_ref_used>No</grammar_ref_used>
<mono_dic_used>No</mono_dic_used>
<bi_dic_used>No</bi_dic_used>
<other_ref_used>No</other_ref_used>
<mode>Written</mode>
<medium>Written by hand</medium>
<length>294</length>
</text_profile>
</header>
<text>
<title>
</w>الرحلة" pos="DT+NC">رحلة<w wid="1" lemma="

</w>إلى" pos="PPR">إلى<w wid="2" lemma="

</w>القرية" pos="DT+NC">قرية<w wid="3" lemma="
</title>
<P pid="1">
<S sid="1">
</w>اعتدت> " pos="VP+NPRPS"اعتاد<w wid="4" lemma="

</w>الذهاب" pos="DT+NC">ذهب<w wid="5" lemma="

</w>إلى" pos="PPR">إلى<w wid="6" lemma="

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

```
</w>قريتي" pos="NPRPS">قرية<w wid="7" lemma="
</w>الموجودة" pos="DT+NC">وجد<w wid="8" lemma="
</w>في" pos="PPR">في<w wid="9" lemma="
</w>غرب" pos="NC">غرب<w wid="10" lemma="
</w>بلدي" pos="NPRPS">بلد<w wid="11" lemma="
</w>في" pos="PPR">في<w wid="12" lemma="
</w>الإجازات" pos="DT+NC">إجازة<w wid="13" lemma="
</w>الصيفيّة" pos="DT+NC">صيف<w wid="14" lemma="
</w>." pos="PU">.<w wid="15" lemma="
</S>
</P>
</text>
</doc>
```

الشكل (11) مثال للجمع بين الوسم والترميز والبيانات الوصفية في ملف واحد باستعمال لغة XML

الختام:

قدم هذا البحث تفاصيل حول أحد أهم الموضوعات المتعلقة بالمدونات اللغوية (Corpora)، وهو الوسم، خاصة مع قلة ما كتب عنها باللغة العربية؛ إذ بدأ بتعريف المدونات اللغوية، وتعريف مصطلح الوسم، ثم عرّج على أهم الدراسات السابقة التي تناولت موضوع وسم المدونات اللغوية بشيء من التفصيل، مستدركاً عليها أنها لم تضع حداً فاصلاً للتفريق بين أنواع الوسوم التي يمكن إضافتها للمدونات، وهنا تأتي الإضافة العلمية لهذا البحث الذي يسعى للتفريق بين ثلاثة من أنواع الوسم التي تضاف إلى المدونات اللغوية وهي: الوسم (Tagging) أو التحشية (Annotation)، والترميز (Mark-up)، والبيانات الوصفية (Metadata)، إذ عرّف البحث كل واحد من هذه الأنواع، وقدّم شرحاً مفصلاً لآليات إضافته لنصوص المدونات اللغوية، مع أمثلة مطبقة على اللغة العربية، وبيّن آلية قياس دقة الوسوم، وكيفية تقييم الاتساق بين الواسمين في حال وسم المدونة من أكثر من واحد. كما بيّن آلية الاستفادة من لغة الترميز القابلة للامتداد Extensible Markup Language (XML) في إضافة الأنواع الثلاثة –الوسم والترميز والبيانات الوصفية– في مدونة

واحدة؛ لزيادة ثرائها، وتسهيل استعمالها في البحث والتحليل، والرفع من قيمة نتائج الأبحاث المبنية على المدونات الموسومة بها.

وقد يلاحظ القارئ في ختام هذا البحث ما لوسم المدونات اللغوية من أهمية كبيرة في تطوير البحث اللغوي والحاسوبي، خصوصاً مع تعدد أنواع الوسوم، وتعدد طرق إضافتها واستعمالها في المدونات، والأهم من ذلك كله الثراء الذي يضيفه وجود أكثر من وسم في المدونة نفسها. وهنا لا بد من التأكيد على أهمية الأبحاث في مجال تطوير المدونات اللغوية، وتطوير الوسوم المستعملة فيها، وأثر ذلك على الأبحاث اللغوية المستقبلية والتطبيقات الحاسوبية في مجال حوسبة اللغة والذكاء الاصطناعي اللغوي؛ ولهذا السبب أشار البحث في ثناياه إلى بعض المواطن التي تحتاج إلى مزيد من الدراسات، خصوصاً من قبل اللغويين، ومنها جداول الوسوم التي لا زالت قاصرة عن شمول جميع أنواع المفردات العربية، إمّا بسبب نقلها من لغات أخرى، أو عدم وجود تصنيف دقيق وشامل لجميع أنواع المفردات العربية التي يمكن وسمها آلياً.

المراجع العربية

الثبيتي، عبدالمحسن عبيد (2015). *تصميم المدونات اللغوية وبناؤها*. في كتاب: صالح بن فهد العصيمي (محرراً)، المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها. (ص147–178). الرياض، السعودية: مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.

صالح، محمود إسماعيل (2014). *لسانيات المدونات اللغوية: مقدمة للقارئ العربي*. في موقع: مدونة الدكتور محمود إسماعيل صالح. تاريخ الدخول 11 سبتمبر 2020. http://dr-mahmoud-ismail-saleh.blogspot.com/2014/04/blog-post_5.html

العصيمي، صالح فهد (محرراً)، (2015). *المدونات اللغوية العربية: بناؤها وطرائق الإفادة منها*. الرياض، السعودية: مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.

القبيشي، هاجر (1441هـ) *بناء خوارزمية لفك اللبس الصرفي الحاسوبي في جموع القلة*، رسالة ماجستير غير منشورة، جامعة الإمام محمد بن سعود الإسلامية، الرياض.

محمد، أحمد روبي (1438هـ) *البنك الشجري النحوي: بناؤه وتوظيفه في إطار تقنيات الذكاء الاصطناعي*، مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة، الرياض.

Reference:

Alfaifi, A. (2015). Building the Arabic Learner Corpus and a System for Arabic Error Annotation. Unpublished Ph.D Thesis, University of Leeds.

Alosaimy, A. & Atwell, E. (2017). Tagging Classical Arabic Text using Available Morphological Analysers and Part of Speech Taggers. The Journal for Language Technology and Computational Linguistics (JLCL). 32 (1), 1-26.

Alqrainy, S. (2008). *A Morphological - Syntaical Analysis Approach For Arabic Textual Tagging*. Unpublished Ph.D Thesis. De Montfort University.

Alqubaishi, H. (2020). *An algorithm for Morphological Disambiguating in the Arabic oligarchs*. Unpublished master's dissertation. IMSIU.

Alrabiah, Maha Sulaiman (2014). Building A Distributional Semantic Model for Traditional Arabic & Investigating its Novel Applications to The Holy Quran. Unpublished Ph.D Thesis. King Saud University. Riyadh.

Althubaiti, A. (2015). Designing and Building Corpora. In S. Alosaimi (Ed.), Arabic Corpora: How to Build and Utilise (pp. 147–178). Riyadh: Kabaical.

Boudchiche, M., Mazroui, A., Ould Abdallahi, M., Lakhouaja, A., Boudlal, A. (2017). AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. *Journal of King Saud University – Computer and Information Sciences*. 29(2), 141-146.

Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0. Lingistic Data Consortium*, University of Pennsylvania, 2004. LDC Catalog NO: LDC2004L02.

Burnard, L. (2005). *Metadata for corpus work. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice* (pp. 30–46). Oxford, UK: Oxbow Books.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Diab, M. (2007). *Improved Arabic Base Phrase Chunking with a new enriched POS tag set*. In: Proceedings of the 5th Workshop on

DOI: https://doi.org/10.35682/mjhsc.v38i1.603

Important Unresolved Matters, Association for Computational Linguistics (ACL), Prague.

Garside. R., Geoffrey, L. & Tony, M. (Eds.) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. New York: Routledge.

Granger, S. (2002). *A bird's-eye view of computer learner corpus research*. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 3–33). Amsterdam, the Netherlands: Benjamins.

Khoja, S. (2001). *APT: Arabic Part-of-speech Tagger*. In: Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania.

Leech, G. (1997). *Introducing Corpus Annotation. In Roger Garside*, Geoffrey Leech & Tony McEnery (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora (pp. 1-18). New York: Routledge.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis.* New York: Springer.

Mohammad. A. (2017). *Grammatical Tree Bank: construction and employment in the context of artificial intelligence techniques*. Riyadh: KABAICAL.

Pasha, A., Mohamed A., Mona, D., Ahmem E., Ramy, E., Nizar, H., Manoj, P., Owen, R. & Ryan M. (2014). *Madamira: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic*. LREC.

Pustejovsky, J. & Stubbs, A. (2013). Natural language annotation for machine learning. Sebastopol, CA: O'Reilly Media.

Sawalha, M. (2011) O*pen-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora*. Unpublished Ph.D Thesis, University of Leeds.

Sinclair, J. (1996). *EAGLES. Preliminary recommendations on corpus typology*. Retrieved 11 April 2013 from http://www.ilc.cnr.it/ EAGLES/ corpustyp/ corpustyp.html

Sinclair, J. (2005). *Corpus and text - basic principles*. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1–16). Oxford, UK: Oxbow Books.

Smrž, Otakar et al., (2008). *Prague Arabic dependency treebank: A word on the million words*. In: Proceedings of the Workshop on Arabic and Local Languages (LREC) 2008.Marrakech, Morocco. European Language 2008.Marrakech, Morocco. European Language Resources Association.

Wynne, M. (Ed.) (2005). *Developing linguistic corpora: A guide to good practice*. Oxford, UK: Oxbow Books.