توسيم أقسام الكلم العربي في المدوّنات اللّسانية العربيّة وأهمّيته في المعالجة الآلية للنّحو العربي

Tagging Arabic Word Categories in Arabic Linguistic Corpus and Its Importance in Arabic Grammar Processing.

طالب دكتوراه. عبد الوهاب معيفي[1]*،دكتورة. نسيمة قطاف[2]

[1] قسم اللّغة العربيّة وآدابها، جامعة باجي مختار عنابة، (الجزائر)، الإيميل المهني: abmaifi2014@gmail.com

[2] قسم اللّغة العربيّة وآدابها، جامعة باجي مختار عنابة ، (الجزائر)، الإيميل المهني: nassimaguettaf23@gmail.com

تاريخ الإرسال: 2022/02/08	تاريخ القبول:17 /2022/06	تاريخ النّشر:26 /2022/06

ملخّص:

رغم كلّ النّتائج التي تحقّقت في مجال المعالجة الآلية للّغات الطّبيعية، وخاصّة اللّغة الإنجليزية منها، غير أن سيرورة تطوّرها لازالت تستقطب كثيرا من الجهود البحثية اللّسانية والحاسوبيّة، وفي ذات السّياق لايزال اللّسان العربي يستدعي بذل المزيد من العطاء والجهد في ميادين البحث اللّسانية الحاسوبية، على الصّعيدين التّنظيري والتّطبيقي؛ ومن هنا تطرح الدراسة الإشكالية التالية: ما الإمكانات التي توفرها بحوث لسانيات المدونة للرقي بالمعالجة الآلية للغة العربية؟ وما الفوائد العملية لتحشية وتوسيم المدونات العربية بأقسام الكلم في بناء المعالجات النحوية العربية؟

وتهدف هذه الدراسة خصوصا لمعرفة بعض سبل توسيم وتحشية المدونات العربية، وأهمية التعرف الآلي لأقسام الكلم العربي في بناء المعالج النحوي، وعموما تحاول إبراز منافع بحوث لسانيات المدونة في المعالجة الآلية للسان العربي، وعليه سيعمل البحث على وصف طرائق تحشية المدونة العربية، وسبل استثمارها في المعالجة الآلية في المستوى النحوي.

الكلمات المفتاحية: لسانيات المدونة؛ المعالجة الآلية للغة؛ المعالج النحوي؛ التوسيم؛ التحشية.

Abstract:

Even of all the results that have been achieved in the field of automatic processing of natural languages, especially in English language, the process of its development still attracts many linguistic and computer research efforts; In this context, Arabic tongue still needs more giving and efforts in computational linguistic research, on both theoretical or applied sides. from here, this study raises the following problem: What are the possibilities offered by corpus linguistics research to advance the automated processing of the Arabic language? And what are the practical benefits of annotating and marking up the Arabic corpora with parts of speech in building Arabic grammatical processors?

* عبد الوهاب معيفي. طالب دكتوراه، و د. نسيمة قطاف

The study aims, in particular, to find out some ways to tagging and annotating Arabic corpora, and show the importance of automatically identifying the parts of Arabic speech in building the grammar processor; and generally, it tries to highlight the benefits of researches on corpus linguistics in the automatic processing of Arabic tongue. The research will then describe the methods of annotating Arabic corpora and how it can be invested in automation at the grammatical level.

Key words: Corpus linguistics; Language automatic processing; Grammatical processor; Tagging; Annotation.

1. مقدّمة

تتجه البحوث والدراسات اللسانية في العالم الغربي، وخاصة الإنجليزي، إلى الاستثمار ما أمكن في لسانيات المدونة، سواء كانت بحوثا نظرية أو تطبيقية؛ لما أثبتته هذه الأخيرة من قدرة على استخلاص النتائج الأكثر دقة وصرامة علمية، وكذا لما تمتاز به من كونها تتعامل مع المواقف الفعلية لاستعمال اللسان الطبيعي، بعيدا عن التقعيد أو التكلف وتصنع النماذج اللسانية؛ غير أن التعامل مع بحوث لسانيات المدونة يتطلب جهودا مسبقة في إعداد المدونات اللسانية محل الدراسة في مختلف مجالات الاستعمال اللساني، هذا من جهة؛ ومن جهة مقابلة العمل على إعداد برمجيات وتطبيقات حاسوبية للتعامل مع هذه المدونات بدءً ببنائها، ثم توسيمها وتحشيتها -إن تطلب الأمر-، وصولا إلى تحليلها واستخراج مختلف الأفكار والمعلومات، وربما النظريات اللسانية منها؛ ومن هنا تبرز مشكلة هذا البحث؛ والتي تركز على سبل تحشية وتوسيم المدونات العربية، وأهميته في المعالجة الآلية للسان العربي، مقتصرة على مستوى المعالجة النحوية.

مشكلة البحث وأسئلته:

تنطلق مشكلة البحث من سؤال عام لتصل إلى إشكالية محددة تتعلق بأقسام الكلم العربي وسبل تحشيتها في المدونة العربية، وأهميتها لبناء المعالج النحوي العربي، وتتضح الإشكالية من خلال السؤالين:

- ما الإمكانات التي توفرها بحوث لسانيات المدونة للرقي بالمعالجة الآلية للغة العربية؟
- وما الفوائد العملية لتحشية وتوسيم المدونات العربية بأقسام الكلم في بناء المعالجات النحوية العربية؟

أهمية البحث:

لا تكمن أهمية هذا البحث فقط في جدة موضوع لسانيات المدونة في الدرس اللساني العربي، بل إنه في الحقيقة يلامس جزئية مهمة جدا تتعلق بالمدونات اللسانية، إذ إن أي بحث في هذا المجال الحديث يتطلب بالدرجة الأولى توفير المادة اللسانية الأولية الخام التي يقوم عليها البحث، والتي تتمثل في مدونات لسانية أو نصوص من الاستعمال الحقيقي والفعلي للسان؛ على أن يتم إعدادها وفق معايير وخصائص علمية متعارف عليها فيما بين المتخصصين في هذا المجال، ومن بين الميزات التي تزيد في جودة المدونات وتيسر سبل البحث ضمنها نجد خصيصة التحشية (**Annotation**) أو التوسيم (**Tagging**)، التي تسمح بتوفير معلومات كثيرة إضافية عن أصل المدونة لفائدة البحث العلمي؛ وهو ما سيهدف البحث إلى تجلية جوانبه قدر المستطاع.

أهداف البحث:

يهدف البحث -كما تبيّن سلفا- إلى تسليط الضوء على مسألة تحشية وتوسيم المدونات أو النصوص العربية بأقسام الكلم العربي، إذ سيقدم نماذج عملية لسبل هذا التوسيم، ومن ثمة النظر في فوائد هذه العملية المهمة في بناء المدونات اللسانية من الجانب النحوي، حيث سيحاول البحث إبراز أهمية المدونات الموسمة بأقسام الكلم في المعالجة الآلية للسان العربي في مستواها النحوي، إذ لا يزال اللسان العربي يحتاج الكثير من الجهود والبحوث في مختلف الميادين والمجالات، وأولاها بالاهتمام مجال اللسانيات الحاسوبية، فيما يتعلق بالمعالجة الآلية للسان العربي، ذلك أن الحاسوب اكتسح بقوة مختلف مجالات الحياة اليومية العادية والعلمية، وعليه تلح الضرورة على مواكبة كل مستجدات البحث العلمي في هذا المجال للسماح للسان العربي ولوج عالم الحوسبة والتكنلوجيا بغايات أسمى وأهداف واقعية أقرب للتحقق في القريب العاجل.

2. مفاهيم أساسية لمصطلحات الدراسة:

يبدو أن هذه الدراسة تعج بالمصطلحات العلمية بداية بلسانيات المدونة، ثم المدونة اللسانية، مرورا بالتوسيم والتحشية، وصولا إلى المعالجة الآلية للغات الطبيعية فالمعالج النحوي؛ غير أن محدودية البحث تقف دون الإتيان عليها كلها، والإلمام بكل جوانب مصطلحات البحث، عليه سنقتصر في هذا المقام على شرح موجز لمفهوم مصطلحي: المدونة اللسانية، والمعالجة الآلية للغة الطبيعية باعتبارهما ركيزة هذا البحث.

1.2 المدونة اللسانية (Linguistic corpus):

تعتبر المدونة اللسانية أحد أهم ركيزتين تقوم عليهما بحوث لسانيات المدونة، إذ ترتكز هذه الأخيرة على المادة اللسانية الخام موضوع البحث والتحليل والدراسة، وهو ما يتجسد في شكل مدونة، ومعها الأدوات الحاسوبية من برمجيات وتطبيقات ومواقع إلكترونية وغيرها، للتعامل مع هذه المدونات بحثا وتحليلا وإحصاء وتطويرا وغير ذلك؛ وتُعرّف المدونة اللسانية بأنها: "تجميعُ نصوصٍ (جسم من اللغة) مخزنةٍ في قواعد معطيات إلكترونية، والمدونات -في الغالب- هي أجسام كبيرة لنصوص مقروءة آليا تحتوي على آلاف أو ملايين الكلمات، وتتميز المدونة عن الأرشيف -غالبا وليس دائما- في كون نصوصها منتقاة مما يمكن القول عنها بأنها تمثيل لتنوعات أو أجناس لغوية محددة؛ وفي الغالب تمتاز المدونات بتحشيتها بمعلومات إضافية مثل: وسم أقسام الكلم، أو علامات للتدليل على ميزات عرضية مرتبطة بالكلام..."[1] ، ومن خلال هذا التعريف الشامل يمكننا أن نقف على أهم خصائص المدونة اللسانية التي تتجسد في كونها:

- تجميع للاستعمال الفعلي للسان موضوع الجمع والدراسة.
- نصوصها تكون مخزنة في ذواكر الحاسوب، وقابلة للولوج إليها والتعامل معها حاسوبيا.
- أحجامها عادة تكون كبيرة وتحسب في الغالب بعدد كلماتها، وتتنوع أحجامها بتنوع صنوف المدونات والغرض من إنشائها.
- تعتبر تمثيلا عاما للسان بصورة عامة أو لأحد تنوعات استعماله أو أجناسه...
- أغلب المدونات غير موسمة، إلا أن التحشية والتوسيم من الميزات التي تضفي الجودة على المدونة مما يسهل التعامل معها.

2.2 المعالجة الآلية للغة الطبيعية (Automatic natural language processing):

نجد تعريفات متعددة في مفهوم المعالجة الآلية للغة الطبيعية، غير أنها كلها تجمع على أنها توظيف الآلة- الحاسوب- في تحليل اللسان البشري ودراسته، ومن أبسط ما نقرؤه في تعريفها: "هي تحليل اللسان البشري باستخدام الحاسوب، مثلا: التحليل الآلي للنص لتحديد أنواع البنى النحوية المستخدمة، أو معالجة المدخلات المنطوقة بالنسبة لتحليل المسموع"[2]، وفي الحقيقة فإن مثل هذا التعريف وإن كان قد أصاب في جانب من جوانب وظائف المعالجة الآلية للغة الطبيعية، إلا أنه قد أهمل الجانب الآخر المهم وهو التوليد، فالمعالجة لم تعد تقتصر على التحليل وقراءة النص اللساني المنطوق أو المسموع بل باتت تتجه نحو توليد النصوص والكلام البشري من قبل الآلة، وهي المهام التي تضطلع بها برامج الذكاء الاصطناعي التي باتت تتطور وتتزايد منتجاتها يوما بعد الآخر؛ وفي تعريف آخر نقرأ أنها: "في اللسانيات الحاسوبية، هي المعالجة الحاسوبية لمواد نصية في اللغات البشرية الطبيعية، والهدف منها هو ابتكار تقنيات تسمح بالتحليل الآلي لأحجام كبيرة من النصوص المنطوقة أو المكتوبة، بطريقة تكون موازية بشكل كبير لما يحصل عندما يضطلع الإنسان بتنفيذ هذه المهمة، وقد تحرر هذا الحقل عن الترجمة الآلية سنوات 1950، ليصبح ذا تأثير كبير في بحوث الذكاء الاصطناعي، مما جعل الجهود لاحقا تركز على البرمجيات الذكية -أو الأنظمة الخبيرة- التي سوف تحاكي جوانب السلوك البشري..."[3] ، وفيما يبدو فإن هذا التعريف قد تعرض لنقاط عديدة في أهداف وخصائص باللّسان .

المعالجة الآلية للغة الطبيعية نجمل أهمها في:

- أنها الموضوع الرئيس للسانيات الحاسوبية.
- تهتم للسان البشري منطوقا ومكتوبا ولا تقتصر على أحدهما دون الآخر.
- تهدف إلى الإبداع والابتكار في مختلف صنوف التقنيات الحاسوبية من برمجيات وتطبيقات ومواقع إلكترونية وغيرها.
- ترمي إلى محاكاة السلوك اللساني الإنساني انتاجا وفهما.
- تطوير النظم الخبيرة وبرمجيات الذكاء الاصطناعي لهذه المحاكاة؛ أي السعي لتحليل وتوليد أنظمة اللسان البشري تقليد للإنسان ذاته.
- بحوثها ودراساتها توليدا وتحليلا تشمل كل مستويات اللسان البشري، صوتا وصرفا ونحوا ودلالة ومعجما...

3. المدونات العربية وسبل توسيمها وتحشيتها

بالنظر لما تم إنجازه، وما يتم العمل عليه في اللسان الإنجليزي، يبقى اللسان العربي متأخرا جدا في مجال إنشاء المدونات اللسانية الحاسوبية لمختلف أغراض البحث العلمي، ومع هذا لا يمكننا أن نبخس حق كثير من الجهود العربية البحتة، أو الجهود المشتركة عربية وغربية، والتي سعت لإنشاء مدونات عربية أثرت قليلا رصيد اللسان العربي في هذا المجال؛ ولإن كانت في غالبها غير موسمة، ولم تتم تحشيتها بالمعلومات اللسانية، فإن هذا لا يمنع من بذل جهود لبناء مدونات موسمة، والعمل على توسيم المنجز منها؛ وفي هذا السياق فإن اللسان العربي ليس بدعا عن بقية الألسن، إذ يتبع سبيل التحشية والتوسيم الآلي بما فيه من نقائص وأخطاء، أو سبيل التحشية والتوسيم البشري بما يحتاجه من جهود مضنية ووقت طويل، وهو ما سنعرض له بعد العنصر التالي الذي سنذكر فيه نماذج عن بعض المدونات العربية الرائدة.

1.3 بعض نماذج المدونات العربية:

أحصى أيمن الدكروري على موقعه الإلكتروني (**Ayman Eddakrouri Infoguistics**) حوالي الأربعين مدونة عربية[4]، وقد قسمها إلى مجموعتين:

- الأولى ضمت المدونات في شكل مواقع إلكترونية.

- الثانية ضمت مجموعة من المدونات النصية؛ أي التي يمكن تحميلها في صيغة نصية (**txt**) مثلا، ثم العمل عليها من خلال برامج مخصصة للمدونات، على غرار برنامج نوج مثلا، أو برنامج غواص العربي، أو برنامج يونيتكس وغيرهم، حيث أن نص المدونة يكون في شكل ملف مقروء بأحد الصيغ المعروفة (**txt**) أو (**doc**) أو (**docx**)...، والتي يسهل تحميلها من موقع المدونة، وتخزينها على جهاز الحاسوب في أي مجلد أو ملف داخل ذاكرة القرص الصلب.

ومن استقراء مجموع هذه المدونات يمكننا أن نصنفها إلى ثلاثة أصناف رئيسية هي:

أ.مدونات القرآن الكريم:

تضم مختلف المدونات التي بنت دراساتها وأبحاثها ومخرجاتها على القرآن الكريم، وقد شملت عدة مدونات كلها في شكل مواقع إلكترونية، لعل أبرز مثال فيها المدونة العربية للقرآن الكريم[5] (**The Quranic Arabic Corpus**)، وهي عبارة عن موقع إلكتروني يبني قواعد معطياته على جميع آيات المصحف الشريف، وتظهر آياته بالرسم العثماني للمصحف، كما أنها تتميز عن بقية المواقع الشبيهة بتقديمها عدة مهام حاسوبية لمعالجة المدونة دفعة؛ فمنها مثلا:

- إظهار كلمات القرآن الكريم كلمة كلمة متتابعة برسم المصحف، مع إبراز موقع الكلمة من الجملة (قسمها من الكلم)، وكذا إعرابها، وأيضا ترجمة إنجليزية لها.

- معجم كلمات القرآن الكريم، الذي يمكن من خلاله اختيار أي جذر لكلمات المصحف والبحث في نطاقه لتظهر لنا نتيجة بمختلف الكلمات التي تنتمي لهذا الجذر والتي وردت في المصحف.

- البنك الشجري للتراكيب، والذي يسمح بإظهار التحليل الشجري لكل آية في القرآن من خلال ربط العلاقات التركيبية بين كلماتها، وإبراز موقعها الإعرابي من الجملة، ويعرض ذلك في شكل صورة للآية وتحتها مختلف الأسهم التي تربط كلمات الآية فيما بينها وفقا للعلاقات النحوية.

ويقوم الموقع بتقديم عدة مهام أخرى لا تتسع ورقة البحث هذه للوقوف عليها جميعا.

ب. المدونات العامة:

هي جملة المدونات التي تم إنشاؤها لأغراض البحث العامة والمتعددة المجالات، إذ تستند مادتها الأساسية من الاستعمالات اللسانية في ميادين متنوعة، على أن يكون تمثيلها لهذه الميادين جيدا، كما أنها في الغالب تكون كبيرة الحجم إذ قد يصل مجموع أعداد كلماتها مئات الملايين من الكلمات، وقد تشمل اللسان المنطوق والمكتوب في آن واحد، وعليه تسمى بالمدونات المرجعية[6]، إذ تكون مرجعا للتأكد من صحة نتائج البحوث في مدونات أخرى؛ ومن أمثلتها: المدونة العالمية للغة العربية[7] (**International corpus of Arabic**)، التي تأسست وفق قواعد معطيات مكتبة الإسكندرية بمصر، ويبلغ حجمها أزيد من 100 مليون كلمة، تم تحليلها صرفيا ونحويا ودلاليا، مما يزود المطلع على مجموعة من المعلومات حول الكلمة المراد البحث عنها، وقد تنوعت مصادر المدونة إلى أربعة أنواع شملت: الصحافة، والمقالات الإلكترونية، والدراسات الأكاديمية، والكتب، واستقت نصوصها من مجالات متعددة منها: العلوم الإنسانية، العلوم التطبيقية، والدين، والأدب، والعلوم الطبيعية، والرياضة وغيرها، غير أن مثل هذه المدونات لا تسمح بالاطلاع

على النصوص الأصلية أو تحميلها، كما أنها في الغالب تحتوي وظيفة وحيدة هي الكشف السياقي (**Concordance**)؛ أي البحث عن السياقات المختلفة التي وردت فيها كلمة ما ضمن نصوص المدونة المتعددة.

ج. المدونات المتخصصة:

هي المدونات التي تقتصر في جمع مادتها اللسانية على مجال واحد متخصص من مجالات الاستعمال اللساني، فهي تحتوي على نوع محدد من النصوص لفترة زمنية محددة[8]، ومثالها: المحاضرات الأكاديمية في تخصص معين، إنتاج المتعلمين في المكتوب أو المنطوق في مدرسة ما أو مقاطعة ما وفي سنة معلومة...، وهي في الحقيقة يتم بناؤها لإجراء الدراسات والبحوث في ذلك المجال المحدد، فمثلا: مدونات المتعلمين تصلح لمختلف البحوث التربوية والتعليمية وبناء المناهج وانتقاء المحتويات وغيرها، ولا تصلح للمعالجة الآلية للغة الطبيعية أو للترجمة الآلية أو لصناعة المعاجم العامة، وعادة ما يقوم الباحث نفسه الذي يجري بحثا في موضوع متخصص ما ببناء المدونة اللازمة لذلك؛ ومن نماذجها: المدونة اللغوية لمتعلمي اللغة العربية[9]، التي سهر على بنائها الباحثان عبد الله الفيفي وإيريك أتويل (**Eric Atwell**)، اللذين قاما بجمع منتجات متعلمين للغة العربية من جنسيات مختلفة، حوالي 67 جنسية بما فيهم ناطقين أصليين باللغة العربية، وقد جمعت مادتها بين سنتي 2012 و2013، بالمملكة العربية السعودية، لإجراء بحوث ودراسات عليها في مجال تعليمية اللغة العربية للناطقين بغيرها؛ والإيجابي في هذه المدونة أنها تسمح بتحميل نصوصها في شكل ملفات بامتداد (**txt**)، مما يسمح بتصديرها لمختلف البرامج الحاسوبية المتعلقة بالمدونات وإجراء التحاليل والدراسات عليها، وبالتالي لا يقتصر الباحث على الكشف السياقي ضمن الموقع فقط، بل يمكن أن يجري عليها أي وظيفة أخرى ضمن أدوات تحليل المدونات المتنوعة التي تدعم اللسان العربي.

2.3 سبل وإشكالات تحشية وتوسيم المدونات العربية:

قبل الخوض في شرح طرق التحشية والتوسيم، ارتأينا أنه من اللزوم الوقوف على مفهومي المصطلحين باختصار، نظرا لكثرة تواردهما مترادفين وتداخلهما في البحوث العلمية والكتابات الأكاديمية.

أما التحشية (**Annotation**)، فتعرف على أنها: "عملية تطبيق معلومات إضافية على معطيات المدونة"[10]، ومن التعريف نستخلص بأن أي نوع من أنواع إدراج معلومات إضافية ليست من أصل نصوص المدونة، يعد من قبيل التحشية، وأما التوسيم (**Tagging**) فيعرف بكونه: "مصطلح أكثر ضبطا في عمليات تطبيق مستويات التحشية لمعطيات المدونة، ويحتوي الوسم عادة على رمز (**Code**) يمكن أن يلحق بفونيم أو مورفيم أو كلمة أو جملة أو مقطع طويل من النص، بطرق عديدة مثالها: استعمال عناصر لغة العلامة النموذجية العامة (**Standard Generalized Markup Language**)، أو باستعمال رمز الشرطة السفلية بين الكلمة ووسمها..."[11] ، ومن خلال هذا التعريف يتبين لنا أن التوسيم نوع أو مستوى من مستويات التحشية كما نص على ذلك التعريف صراحة، والتوسيم كما يظهر يتعلق بإدراج المعلومات اللسانية لنصوص المدونة، والتي تعمل على تقديم معلومات إما حول الكلمة المفردة، أو الجملة أو غيرها، ولهذا كثيرا ما ارتبط مصطلح التوسيم بأقسام الكلم في مختلف البحوث والكتابات الأكاديمية.

أ. سبل تحشية وتوسيم المدونات:

لا غنى للباحثين العرب والمتخصصين منهم في لسانيات المدونة عن أحد سبل ثلاثة لتحشية المدونات العربية[12]، وهي ذاتها السبل المستخدمة في مختلف اللغات، غير أن اللسان الإنجليزي مثلا، والذي شهد تطورا كبيرا في مجال بحوث لسانيات المدونة باتت عمليات التحشية والتوسيم فيه تتسم بالصبغة الآلية؛ أي أن التحشية تتم تقريبا وفق سبيل واحد هو التحشية الآلية، "وهذا هو المتبع في غالب المدونات الإنجليزية بسبب دقة مثل هذه الأنظمة الآلية في اللغة الإنجليزية؛ ولكن الحال في العربية مختلف، فالأبحاث في هذا المجال محدودة، والموجود منها لا

يراعي ما هو مستقر في النحو العربي إجمالا"[13]، فنظرا لكثرة البرامج والأدوات الحاسوبية التي تم ابتكارها لهذا الغرض في اللسان الإنجليزي، وعمليات تحيينها مرة بعد أخرى لاستدراك الأخطاء وتصحيح الهفوات البرمجية الموجودة فيها، جعلها تحقق نسبة نجاح عالية في التحشية بأنواعها؛ على عكس اللسان العربي الذي ما زال مرتبطا في غالبه بهذه البرمجيات الغربية التي تدعم اللسان العربي؛ وفيما يأتي موجز مقتضب عن سبل تحشية المدونات العربية.

الأولى: التحشية البشرية أو اليدوية الخالصة، والتي يقوم فيها الباحث بنفسه بإدخال المعلومات الإضافية واحدة تلو الأخرى ودون استغلال للإمكانات الحاسوبية، فهذه الطريقة تعتمد "كلية على محلل بشري دون أي أداة برمجية، ونظرا إلى أن هذه الطريقة مكلفة ومستنفدة للوقت، فإنه يفضل استخدامها مع المدونات اللغوية الصغيرة"[14]، ومن المؤكد بأن توسيم المدونات بطريقة يدوية يستدعي توفر متخصصين في المجال، ويتطلب جهودا بشرية كبيرة، وعليه كثيرا ما تضطلع فرق عمل بهذه المهمة، ولا يكون الجهد فيها فرديا، وقد لا تسلم هذه المدونة الموسمة فيها من الأخطاء بحكم التعب والإرهاق وقلة التركيز التي تعتري البشر، ذلك أن الذي يقوم بعملية التوسيم هذه، يتتبع كل كلمة من كلمات نص المدونة، ويضع أمامها الرمز المناسب لها، الذي يحدد موقعها في التركيب.

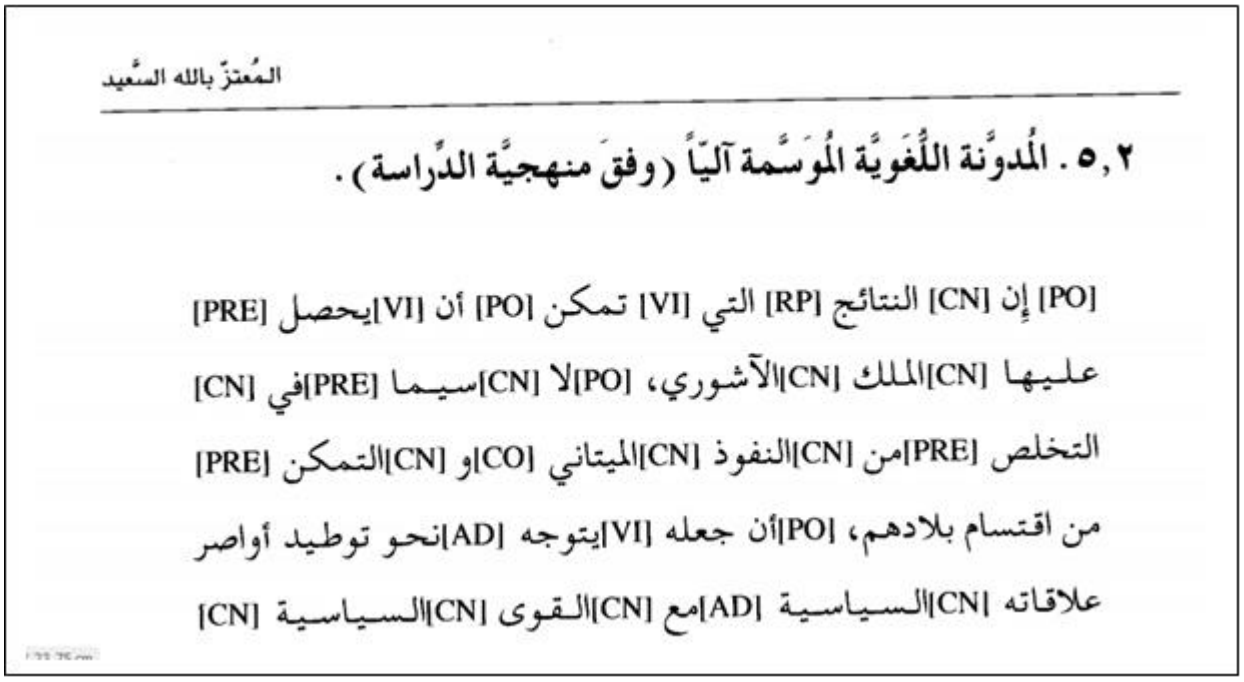
الثانية: التحشية الآلية البحتة، وهي التي تقوم الآلة فيها بهذا الدور من خلال البرامج الحاسوبية التي تدعم هذه الوظيفة في معالجة المدونات، إذ "تقوم البرمجيات بالتحشية بناء على قواعد وخوارزميات تم إعدادها مسبقا بواسطة مبرمجين"[15]؛ أي أن مهندسي الحاسوبيات رفقة لسانيين متخصصين يقومون بتزويد برامج وتطبيقات حاسوبية بالخوارزميات والقواعد التركيبية للغة، مما يجعل الحاسوب فيما بعد يطبق هذه الخوارزميات على نص المدونة، ويضع أمام كل كلمة أو جملة الرموز المناسبة للتوسيم بها، وتعد هذه العملية مكلفة من الناحية المادية، فهي تتطلب توفير الأموال الكافية، وأيضا تستغرق بعض الوقت لإنجاز وبناء هذه البرامج، غير أنه "بمجرد الانتهاء من هذه الأداة البرمجية فإنه يمكن تحشية كم ضخم من النصوص بالسرعة والاتساق المطلوبين"[16]، وعلى الرغم من أن هذه العملية توفر الوقت والجهد البشري، وتتسم بالدقة التقنية، إلا أنها لا تسلم من الأخطاء، خاصة فيما يتعلق باللسان العربي، كما أن نتائج أدائها ضعيفة، وتتوفر العديد من البرمجيات الحاسوبية التي تؤدي هذه الوظيفة.

الثالثة: التحشية شبه الآلية، وهي التي يتوزع فيها الدور بين الإنسان والآلة في عمليات التحشية، حيث تقوم الآلة بالعمل الأولي ثم يقوم الإنسان بإجراء تعديلات وتحسينات على ما قامت به الآلة، "فيتدخل البشر في إجراء التصويبات المطلوبة، حيث توفر بعض برمجيات التحشية إمكانية التدخل البشري لحل حالات الخطأ أو اللبس التي تستعصي على البرامج الآلية. وينتج عن هذه الطريقة شبه الآلية نتائج مسترجعة أكثر موثوقية من نظيرتها في الطريقة الآلية كلية"[17]، وتبدو هذه الطريقة الأنجع لكسب الوقت وتوفير الجهود وتحقيق نتائج عملية وأكثر دقة.

تستخدم في عمليات التحشية بكل صنوفها رموز محددة متعارف عليها في لغات البرمجيات الحاسوبية، مثل "لغة الترميز المعممة القياسية (SGML)، أو باستخدام لغة الترميز القابلة للتمديد (XML)"[18] ، هاتين اللغتين المستخدمتين بكثرة لبناء صفحات الإنترنت؛ بينما قد تستخدم بعض البرمجيات أسلوبا آخر في التحشية بوضع رمز الشرطة السفلية (_) بين الكلمة ورمز توسيمها.

وفيما يأتي صورة لمثال عن توسيم نص عربي.

الشكل رقم (01): صورة لنموذج عن نص عربي موسم آليا بوضع رمز وسم كل كلمة بين عارضتين



المُعتزّ بالله السَّعيد

٥,٢ . المُدوَّنة اللُّغَويَّة المُوَسَّمة آليّاً (وفقَ منهجيَّة الدِّراسة) .

[PO] إن [CN] النتائج [RP] التي [VI] تمكن [PO] أن [VI]يحصل [PRE]
عليها [CN]الملك [CN]الآشوري، [PO]لا [CN]سيما [PRE]في [CN]
التخلص [PRE]من [CN]النفوذ [CN]الميتاني [CO]و [CN]التمكن [PRE]
من اقتسام بلادهم، [PO]أن جعله [VI]يتوجه [AD]نحو توطيد أواصر
علاقاته [CN]السياسية [AD]مع [CN]القوى [CN]السياسية [CN]

المصدر: المعتز بالله سعيد، (2017)، ص 325.

ب. إشكالات تحشية المدونات العربية:

تفرض طبيعة وخصائص اللسان العربي تحديات وإشكالات تقف في وجه توسيم المدونات العربية، هذا إذا استثنينا الجهود البشرية المتخصصة المطلوبة، وتجاوزنا مشكلة قلة البرمجيات الحاسوبية المتخصصة في اللسان العربي، والتي تأسست أساسا بناء على معطيات معالجة اللسان العربي، إذ إن غالبية البرمجيات تأسست أصلا لمعالجة اللسان الإنجليزي أو الفرنسي أو غيرهما، وتكون مدعومة بخاصية قابلية دعم اللسان العربي؛ ومن بين الإشكالات اللسانية البحتة التي تزيد من صعوبة عمليات تحشية المدونات العربية[19]:

أولا. مرونة نظام الجملة العربية:

إن التركيب العربي يختلف كثيرا عن نظيره في مختلف اللغات الأوربية أو غيرها، حيث إن التركيب العربي سلس في الكثير من أنظمته، إذ تكثر فيه ظواهر التقديم والتأخير، وكذا الإضمار والإظهار، وتتنوع الجملة فيه بين اسمية وفعلية، وقد تكون الجملة فيه مركبة من العديد من العناصر مما يجعلها طويلة وصعبة الوقوف على مختلف وحدات الكلم فيها وقواعدها الإعرابية؛ "وتمثل هذه المرونة إشكالا عند توسيم المدونات اللغوية تركيبيا، لأنها تستدعي عملا يدويا شاقا للبحث عن قسم الكلام الذي يتبعه كل عنصر من عناصر الجملة على حدة، وحال التدخل الآلي لتوسيم المدونة، فإن نسبة الخطأ لن تكون قليلة، وهذا يسدعي تدخلا يدويا كبيرا لمعالجة الأخطاء الناجمة عن عمل الآلة"[20].

ثانيا. نظام كتابة اللغة العربية:

يمتاز نظام الكتابة العربي أو وحداته الكتابية (**Graphemes**)، بطبعته الإلصاقية، كما يميز اللسان العربي خصيصة الاشتقاق، مما يجعل المصدر الواحد ترتبط به الكثير من المشتقات، وتختلف صور كتابة حروفها بحسب الصيغة الصرفية لكل مشتق، وهذه الخصائص تجعل من التركيب العربي صعب التحليل والتجزيء، إذ قد تجد سلسلة من الحروف المتلاصقة تحوي في طياتها مجموعة من الكلمات تنتمي لأقسام كلم متباينة، ولا أبرز للمثال على هذا من كلمتي القرآن الكريم الشهيرتين: "أنلزمكموها"، و"سيكفيكهم"؛ فالأولى ضمت: أداة الاستفهام (أ)، وحرف الزيادة للدلالة على الزمن المضارع (ن)، الفعل (لزم)، والضمير المتصل الدال على جماعة المخاطَبين (كم)، الضمير المتصل العائد على المؤنث الغائب (ها)؛ ومن هنا نلاحظ أن سلسلة حروف متلاصقة في كلمة واحدة، ضمت خمسة كلمات لكل واحدة منها حالتها الإعرابية، وقسمها الخاص من الكلم العربي، وعليه "فإن توسيم المدونات اللغوية تركيبيا يفرض الجمع بين بعض أقسام الكلام المتشابكة، كما يستدعي ضبط النصوص بالشكل تحسبا للالتباس المحتمل وقوعه عند توسيم الكلمات المتماثلة في رسمها"[21].

ثالثا. الاختلاف حول أقسام الكلم العربي:

اختلفت الدراسات وتباينت قديما وحديثا حول أقسام الكلام العربي، فقد أجمع الفكر اللساني القديم كله تقريبا على أن تقسيم عناصر الكلام العربي يتوزع على ثلاثة أقسام هي: الاسم، والفعل، والحرف؛ غير أن بعض الآراء اللسانية الحديثة قد حادت عن هذا الإجماع وتباينت فيما بينها حول أقسام الكلم العربي، "فيذهب فريق إلى تقسيم الكلام العربي إلى أربعة أقسام، هي: الاسم، والفعل، والحرف، والضمير(...)، ويذهب فريق آخر إلى تقسيم الكلام العربي إلى سبعة أقسام، هي: الاسم، والصفة، والفعل، والضمير، والخالفة، والظرف، والأداة (...) ويمثل هذا الاختلاف إشكالا عند توسيم المدونات اللغوية تركيبيا"[22]، إن مثل هذا الإشكال يجعل من عملية توسيم المدونة صعب التحقق في ظل عدم ضبط أقسام الكلم العربي، وعليه لا بد للباحث الذي يريد توسيم مدونة ما، أن يضبط هدفه من المدونة أولا، ويحدد أقسام الكلم مسبقا لوضع الرموز المقابلة لكل قسم منها حسب لغة البرنامج الحاسوبي، حتى يتمكن فيما بعد من وضع كل رمز أمام الكلمة التي تنتمي لذلك القسم، سواء يدويا أو باستعمال الحاسوب.

كانت هذه بعض إشكالات توسيم التركيب العربي، وإذا ما أسقطناها على المعالجة الآلية في المستوى النحوي سنجدها نفس الإشكالات تطرح نفسها، فمعالجة الجملة العربية آليا سواء من جهة التحليل أو التوليد تتعثر أمام نفس هذه الإشكالات، وهو ما يحيلنا لشرح نموذج عملي بسيط لتوسيم المدونة العربية.

4. نموذج عملي لتوسيم المدونة العربية بأقسام الكلم:

في هذا النموذج الذي سنمثل به لكيفية توسيم المدونة العربية، وفائدة ذلك في المعالجة الآلية في مستواها النحوي، اخترنا نصوصا من مدونة الملخصات العربية إسكس (Essex Arabic Summaries Corpus)، حيث اتبعنا الخطوات التالية في عملية التوسيم.

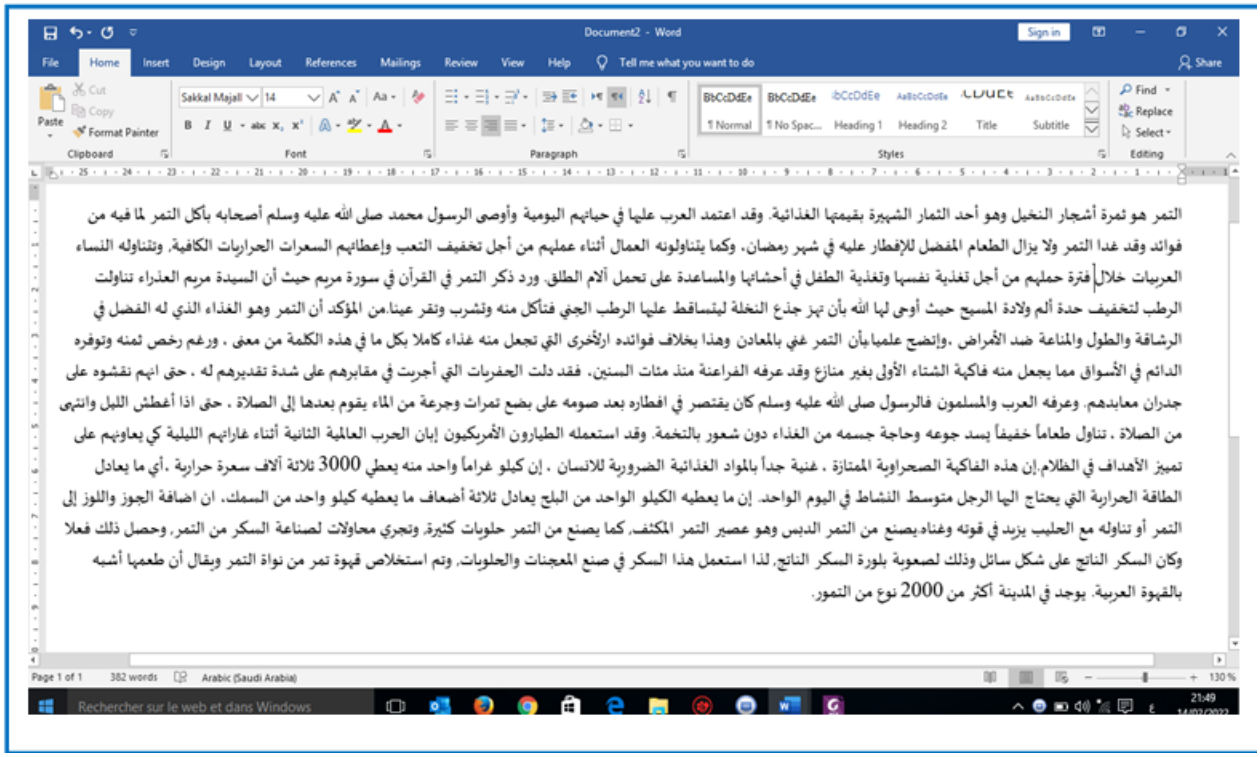
1.4 تحضير المدونة الخام (نموذج التطبيق):

تم اختيار مدونة النموذج العملي هذا، من ضمن مجموعة من النصوص سهر على تجميعها وترتيبها محمد الحاج، بجامعة إسكس البريطانية سنة 2013، وتحتوي في مجموعها على ثلاثة وخمسين ومائة (153) ملخص نص لمقالات صحفية، من مصادر ثلاثة رئيسية هي: جريدة الوطن السعودية، جريدة الرأي السعودية، وموقع ويكيبيديا، وقد تنوعت مجالات النصوص المنقولة إلى عشرة مجالات منها الرياضة والتعليم والبئة والصحة والتكنلوجيا وغيرها،

وهي مقسمة بشكل منفصل كل نص ضمن ملفه الخاص بامتداد (*txt*)، ومعنون بنوع المجال ورقم النص مثاله: ((Art and Music (1)، وهكذا مع بقية النصوص المائة والخمسة والثلاثين هنا[23].

وقد اخترنا لهذا النموذج ثلاثة نصوص من مجموعة نصوص مدونة (EASC)، وهي التي يوضحها الجدول رقم (01)، وأسميناه نموذج التطبيق

الشكل رقم (02): صورة توضح النص المقتطف من مدونة (EASC) مفتوح على صفحة وورد (Word)



رقم (01): معلومات نصوص المدونة الخام (نموذج التطبيق) المستوردة من مدونة (EASC)

رقم النص	مجاله	عنوانه في المدونة الأصل	موضوعه	مصدره	حجمه
01	التعليم	النص رقم 03	الإصلاح	ويكيبيديا	465 كلمة
02	البيئة	النص رقم 03	المطر		294 كلمة
03	علوم وتكنلوجيا	النص رقم 03	سيارة مرسيدس		276 كلمة
حجم المدونة الخام (عدد كلماتها)					1035 كلمة

كانت هذه لمحة مختصرة عن مدونة الدراسة ومعلوماتها مقتضبة، وفيما يأتي مراحل التعامل مع كلمات المدونة وتوسيمها نحويا بأقسام الكلم العربي.

2.4 ضبط أقسام الكلم التي سيتم اعتمادها للتوسيم:

شرحنا سابقا أن الاختلاف قائم في اللسان العربي حول ضبط أقسام الكلم العربي، بين القدماء والمحدثين، والحقيقة أن المسألة تزيد تعقيدا عندما نعلم أن كل قسم من هذه الأقسام توجد أقسام فرعية أخرى؛ فقسم الاسم مثلا: تنطوي تحته الصفة، والاسم العلم، والمؤنث، والمذكر، والجمع، والمثنى، والمفرد، واسم الإشارة، والاسم الموصول، واسم العدد...، والحقيقة أن هذه المسألة لا ينفرد بها اللسان العربي عن غيره، فحتى اللسان الإنجليزي رغم التوافق على تصنيف أقسام كلمه إلى ثمانية أقسام إلا أنها تتفرع بدورها إلى ما ذكرناه سابقا، مع اختلافات أكيدة تميز لسانا عن آخر في هذه التفرعات؛ وعليه كان لزاما علينا ضبط أصناف أقسام الكلم وتحديد رموزها مسبقا للانطلاق منها في ضبط توسيم نموذج المجونة الخام الذي اقترحناه لهذه الدراسة، وفيما يأتي جدول بضبط أقسام الكلم وفروعها، انطلقنا فيه من تقسيم القدامى، والتفريع الذي يمكن أن يكثر استعماله، على أننا لم نتوسع فيه ليشمل كل ما ذكر في كتب النحو؛ وعليه يكون قد اجتمع لنا ثلاثة أقسام كلم رئيسية، وخمسة عشر قسما فرعيا.

الجدول رقم (02): أهم أقسام الكلم التي سيتم التركيز عليها في توسيم المدونة الخام (نموذج التطبيق)

قسم الكلم	فرعه	مثاله	مقابله الإنجليزي	رمز توسيمه
اسم (Noun)	الاسم المشترك	تعليم	Common Noun	CN
	الاسم العلم	علي	Proper Noun	PN
	الصفة	جديد	Adjective	Adj
	الظرف	تحت	Adverb	Adv
	اسم الاشارة	هذا	Determiner	DE
	الاسم الموصول	الذي	Relative Pronoun	RPRO
	اسم العدد	خمسون	Cardinal Number	CNU
	الضمير	أنتم	Pronoun	PRO
فعل (Verb)	ماض	عَلِم	Past Verb	PV
	مضارع	يتعلَّم	Present Verb	PRV
	أمر (طلب)	تعلَّمْ	Request Verb	RV

حرف أو أداة (Article)	حرف العطف	و	Conjunction	CONJ
	حرف الجر	في	Preposition	PRE
	أداة الاستفهام	هل	Question Article	QA
	أداة أخرى	إن، قد، إلا	Other Article	OA

3.4 الكشف عن قائمة كلمات المدونة وسياقات استعمالها:

قمنا بتحليل نص المدونة الخام على تطبيق آنت كونك (**AntConc 3.5.8**)، وهو برنامج يحتوي على مجموعة من الوظائف لتحليل بيانات ومعلومات المدونات النصية، أين استظهرنا قائمة كلمات المدونة مرتبة من الأكثر ترددا إلى الأقل ترددا، فظهرت النتيجة كما تظهره الصورة.

الشكل رقم (03): صورة توضح جانبا من قائمة كلمات المدونة نموذج التطبيق مرتبة حسب درجة ترددها

تظهر الصورة المعلومات الأولية عن المدونة الخام، فيما يتعلق بعدد كلماتها كلية، وعدد كلماتها دون تكرار وترتيبها حسب كثرة ترددها، وهو ما يلخصه الجدول التالي:

الجدول رقم (03): يلخص المعلومات الأولية لمحتوى المدونة نموذج التطبيق

عدد هياكل كلمات المدونة (Word Tokens)	1026
عدد الكلمات دون تكرار (Word Types)	693
الكلمات العشر الأولى الأكثر ترددا	من، و، في، على، أن، إلى، التي، لا، هي، الذي
الكلمات العشر الأخيرة الأقل ترددا	ينفذه، ينظم، ينخرط، ينتظر، ينبغي، يمكن، يمتاز، يقيد، يقع، يفوق

ويمكن تسجيل بعض الملاحظات الأولية هنا:

- عدد كلمات المدونة يتم احتسابه من خلال فاصل المسافة الموجود بين الكلمة والتي تليها، وليس بحسب جذر كل كلمة وفصلها عن سوابقها ولواحقها.

- الكلمات الأكثر ترددا في المدونة تنتمي لقسم الحروف والأدوات، وبعضها لقسم الأسماء ضمن فرع الأسماء الموصولة، ولا يظهر ضمنها أصناف الأسماء الأخرى والأفعال.

- الكلمات الأقل ترددا تنتمي لقسم الأفعال.

وهذا ما يعطينا فكرة أولية عن عملية التوسيم.

وفيما يأتي صورة توضيحية عن سياقات استعمال أحد كلمات المدونة، باستعمال نفس البرنامج الحاسوبي.

الشكل رقم (04): صورة توضح سياقات استعمال كلمة (قطرات) في المدونة نموذج التطبيق

كما توضحه الصورة فإن كلمة قطرات وردت في نص المدونة ستة مرات بنفس الصيغة، ويمكننا أن نستظهر السياقات التي وردت فيها هذه الكلمة، وبنفس الطريقة يمكن تتبع كل سياقات كلمات المدونة التي وقع لنا لبس في معرفة قسم كلمها، إن كانت اسما أو فعلا أو أداة، لأن بعض الكلمات قد تلتبس حال انفرداها وانعزالها عن السياق، مثل: كلمة (أحمد)، فهي تحتمل أن تكون اسم علم، أو تكون فعلا مضارعا؛ وكذا كلمة (من) فهي تحتمل أداة الاستفهام، وتحتمل حرف الجر، وتحتمل الاسم؛ وهكذا المسألة في كثير من الكلمات التي لا يتجلى وجه استعمالها، وقسمها من الكلم إلا من خلال تعرف سياقها.

وبعد إحصاء كلمات المدونة وتعرف سياقاتها وأيضا متصاحباتها اللفظية قبلها وبعدها، من خلال استخدام وظيفة النحو العدد (N-Grams) في البرنامج لنزداد معرفة بكل كلمة في المدونة أداة أو حرفا أو اسما وما يمكن أن يلحقها في التركيب، أو يسبقها، فتتكون لنا فكرة عن أهم عناصر الكلم في المدونة والأكثرها ترددا؛ قمنا بتجربة توسيم آلي للمدونة الخام، لنرى نتائج التوسيم ودقته.

4.4 تجربة التوسيم الآلي باستعمال برنامج (TagAnt):

باستخدام برنامج (تاغ آنت) (**TagAnt**) وهو من نفس زمرة البرمجيات التي ينتمي لها برنامج التحليل سابق الاستعمال (آنت كونك) الذي يظهر في الشكلين (03) و(04)، باعتبار أن حزمة برامج (آنت) تدعم اللغة العربية في أغلبها، عدا برنامج التوسيم هذا -رغم أنه يدعم اللغة العربية- إلا أنه غير مزود بخوارزميات التحليل العربية، وعليه يتعامل معها على أنها من بقية اللغات.

وبالقيام بتجربة التوسيم الآلي تظهر لنا الصورة التالية للمدونة بعد توسيمها آليا ببرنامج (**TagAnt**).

الشكل رقم (05): صورة المدونة نموذج التطبيق بعد توسيمها آليا بالبرنامج الحاسوبي

يمكن أن نسجل ملاحظات عديدة حول هذه العملية الآلية للتوسيم:

1-البرنامج لم يستخدم الرموز المختصرة للتوسيم والتي ذكرنا نماذج عنها في الجدول رقم (02)، رغم أن البرنامج مزود بلائحة من الرموز التوسيمية بلغت في مجملها ثمانية وخمسين رمزا[24]، شملت مختلف أصناف أقسام الكلم الإنجليزي وهي نفسها المطبقة على بقية الألسن التي يدعمها البرنامج، وبدلا عنها فقد أشار في الغالب بكلمات تعبر عن قسم الكلم، مثل: (**Verb**) الفعل، (**PROPN**) الاسم العلم، (**NOUN**) الاسم...

2- وقع البرنامج في الكثير من الأخطاء في توسيم أقسام الكلم، وهذا كما قلنا لعدم تزويده بخوارزميات التحليل العربي، إذ بالرغم من أن حزمة البرامج التي طورها لورنس أنطوني تحت مسمى (**Ant**)، والتي اشتغلنا على برنامجين منها هنا تدعم اللغة العربية، إلا أن هذا البرنامج لا زال يحتاج الدعم سواء من المطورين العرب، أو من صاحب البرنامج نفسه لتدعيم توسيم اللسان العربي؛ ونذكر أمثلة على هذه الأخطاء للتوضح فقط: (وصلت_**PROPN**) فعل موسوم بأنه اسم علم، (لا_**PROPN**) أداة موسومة بأنها اسم علم، (في_**VERB**) حرف جر موسوم بأنه فعل.

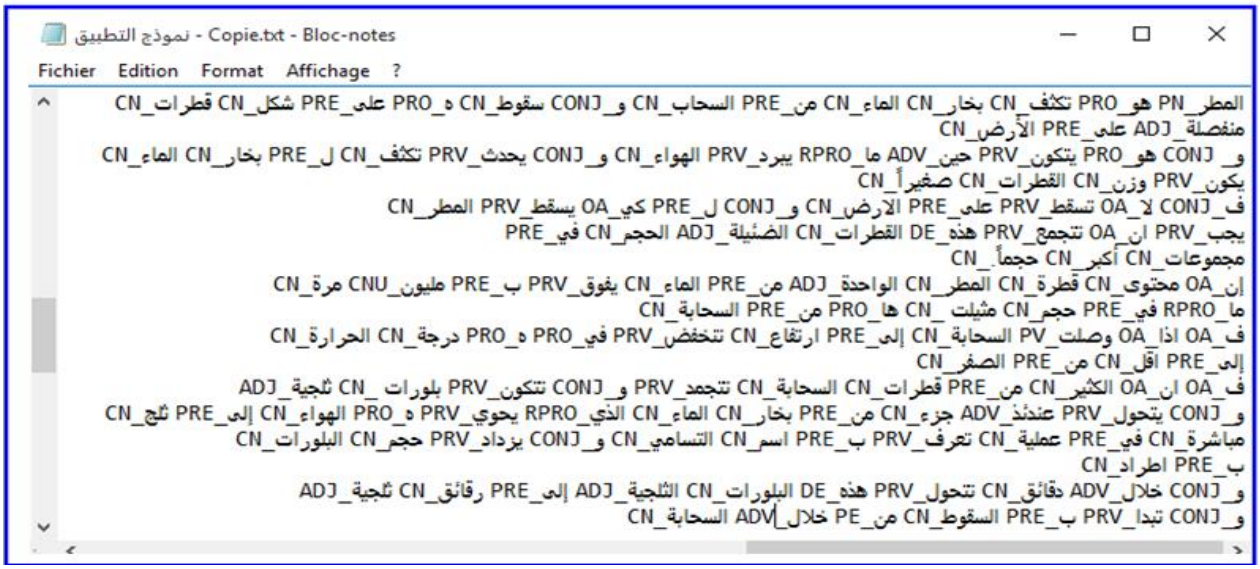
3- يجب الاعتراف أيضا أن البرنامج نجح في كثير من عمليات التوسيم أيضا رغم افتقاره لخوارزميات التوسيم العربي، ونذكر أمثلة أيضا للتوضيح: (السحب_**NOUN**) اسم، (واسعة_**ADJ**) صفة، (القطرات_**NOUN**) اسم؛ والحقيقة أن أغلب توسيماته كانت تحت قسم الاسم، وعليه جاءت نجاحاته أغلبها مشيرة إلى الأسماء.

4- لم يستطع البرنامج التعامل مع الكلمات التي تتشكل من سوابق ولواحق ملتصقة بأصل الكلمة، وعليه عاملها كلها ككلمة واحدة ودخلت معها سوابقها ولواحقها في نفس رمز الوسم، في حين أن كثيرا من هذه السوابق هو عبارة عن أدوات أو حرف جر أو عطف وغيرها، ومثالها: (وفي_**INTJ**) وسم الحرفين معا بأنهما أداة تعجب وهو خطأ، في حين أن أحدهما حرف عطف والثاني حرف جر؛ أي أن توسيمهما السليم يكون على الشكل (و_**CONJ** في_**PER**).

5.4 التوسيم اليدوي لتصحيح واستدراك أخطاء التوسيم الآلي:

بعد عرض نتائج التوسيم وفق البرنامج الآلي سابقا قمنا بإعادة التوسيم يدويا بالمحافظة على الجوانب السليمة والإشارة إليها برمزها المختصر، وتصحيح الجوانب الخاطئة، بعد تعرف سياقات كل كلمة كما رأينا سابقا، للتأكد من تصنيف قسمها وفرعه الذي تنتمي إليه فبرزت المدونة كما في الصورة التالية:

الشكل رقم (06): صورة المدونة نموذج التطبيق بعد توسيمها يدويا إثر التوسيم الآلي



في هذه الصورة تتضح نتيجة التوسيم اليدوي ويظهر رمز كل وسم لقسم من أقسام الكلمة بجانب الكلمة الموافقة له تفصل بينهما شرطة سفلية، وهو الأسلوب البسيط الذي تتبعه كثير من البرامج الحاسوبية سواء في عملية التوسيم ذاتها، أو في قراءة وتحليل المدونات الموسمة، لاستخراج قوائم كلماتها، وتعرف مواقعها من التركيب، لتقديم مختلف

المعلومات الإحصائية بدقة أكبر ونتائج تحليل أفضل؛ وقد قمنا بوسمها وفق الاختصارات المحددة سابقا، والتي تستعملها مختلف البرامج الحاسوبية وتتعرفها بسهولة، وبهذه العملية الأخيرة نكون قد سلكنا سبيل التوسيم شبه الآلي الذي يجمع بين استثمار الموارد الحاسوبية في الحصول على مختلف معلومات كلمات المدونة أولا، ثم استغلالها لتسريع عملية التوسيم مع ما تتضمنه من أخطاء، وبعدها الاستدراك على نتائج هذه الموارد الحاسوبية باستغلال التوسيم اليدوي لضبط الأمور وتصحيحها، ومن هنا نربح الوقت والجهد في توسيم المدونات العربية ونقدم موارد لسانية عربية محوسبة بشكل جيد لاستغلالها في مختلف عمليات المعالجة الآلية.

6. سبل استثمار المدونات الموسمة تركيبا في بناء المعالج الآلي:

قد يتساءل الكثير من الباحثين اللسانيين التطبيقيين عن جدوى التوسيم بأقسام الكلم للمدونات العربية، وهذا ما يدفعنا للتأكيد على أن التركيب العربي يتشعب كثيرا فيختلط على كثير من الباحثين إدراك عناصره اللسانية ووحداته البسيطة بدقة وصحة، ولأن عمليات تحليل النصوص العربية آليا تحتاج إلى تعرف على عناصر التركيب وسبل ترابطها في علاقات نحوية، وإعطاء التوصيف الدقيق للتركيب العربي بنوعيه الاسمي والفعلي، ورسم العلاقات الترابطية بين مختلف الكلمات العربية في التركيب.

ذلك أن التوسيم السليم لأقسام الكلم يعطينا تحليلا دقيقا للنصوص العربية من خلال برامج تحليل المدونات، وباستغلال وظيفة استخراج المتصاحبات اللفظية (N-Grams) في برامج تحليل المدونات بتغيير العدد (N) حسب حاجاتنا لدرجات الترابط والتلاصق اللفظي؛ أي معرفة كلمتين أو ثلاث أو أربع أو خمس كلمات من التي تسبق كلمة معينة أو تلحقها، فيمكننا بذلك استخراج العلاقات الترابطية لهذه الكلمة في التركيب، وبالتالي تقديم توصيف جيد لهندسة التراكيب العربية.

إن هذه الهندسة النحوية السليمة للتركيب العربي من طرف اللسانيين التطبيقيين الذين يشتغلون على لسانيات المدونة يسهل كثيرا على بناء الخوارزميات الحاسوبية الصحيحة من طرف مهندسي تطوير البرامج الحاسوبية لبناء المعالجات النحوية؛ وهذا من أجلى وأوضح سبل استثمار التوسيم في بناء المعالج النحوي العربي، إذ أنه يسهم في:

- تعرف مكونات التركيب العربي بنوعيه؛
- وتوضيح العلاقات الترابطية التركيبية بين كلمات التركيب العربي،
- وتعرف قسم كل كلمة في التركيب العربي، وبالتالي تسهيل معرفة حركاته الإعرابية،
- ومعرفة أقسام الكلام الأكثر شيوعا واستعمالا في التراكيب العربية، والنادرة منها؛ للتركيز في بناء المعالج النحوي على عناصر الكلام التي يشيع استعمالها،
- وبناء توصيف سليم للتركيب العربي ومن خلاله هندسة التراكيب العربية في شكل معادلات رياضية ومخططات تشجيرية تسهل تعرف مختلف العلاقات النحوية بين الكلمات العربية للتراكيب السليمة،
- ومساعدة وتسهيل عمل المبرمجين الحاسوبيين في إيجاد الخوارزميات الحاسوبية الصحيحة لبناء المعالج النحوي العربي.

7. خاتمة:

في ختام هذا البحث المختصر عن سبل توسيم أقسام الكلم العربي في المدونات اللسانية العربية، والذي رأينا فيه كيفية استغلال الموارد الحاسوبية لتوسيم شبه آلي للنصوص العربية بأقسام الكلم فيها؛ ومن خلال المدونة الموسمة أشرنا إلى بعض فوائد هذا التوسيم في بناء المعالج النحوي العربي؛ نصل إلى الخلاصة التي تؤكد على مدى أهمية بناء المدونات اللسانية العربية وتوسيمها في مختلف ميادين البحث اللساني، وخصوصا منها ميدان المعالجة الآلية للغة العربية؛ ونؤكد على أن فوائدها الجلية في بناء المعالجات الحاسوبية للسان العربي وعلى رأسها المعالج النحوي من خلال توسيم هذه المدونات بأقسام الكلم العربي. كما لا يفوتنا الإشارة إلى قلة المدونات العربية، مما يعيق إجراء البحوث والدراسات لتطوير البرمجيات العربية؛ وندرة البرمجيات الحاسوبية العربية التي تتفاعل مع المدونات تحليلا أو توسيما، وأن المتوفر منها تم تطويره في جامعات ومراكز بحث غربية، مما يجعلها تفتقد في بنائها للتركيز على خصائص اللسان العربي.

وعليه يقدم البحث توصيات عامة مهمة:

1- زيادة الاهتمام ببحوث لسانيات المدونة ودراساتها، وتشجيع تدريس هذا التخصص في مختلف الجامعات العربية؛ وإن استدعى الأمر إرسال بعثات بحثية للاستفادة من تجارب العالم العربي في هذا المجال خاصة ما تعلق باللسان الإنجليزي.

2- بعث المشاريع العربية في مختلف أقطار الوطن العربي وتمويلها لبناء مدونات عربية خام وموسمة، وتوفيرها بصورة مجانية بيد الباحثين لاستغلالها في مختلف البحوث والدراسات اللسانية، والتي قد تحقق نتائج بحث أو رؤى لسانية جديدة تفيد اللسان العربي.

3- التنسيق مع مختلف أقسام الإعلام الآلي في الجامعات العربية، وتشجيع التواصل بين اللسانيين التطبيقيين والمبرمجين الحاسوبيين لبناء البرمجيات والتطبيقات الحاسوبية العربية الصنع، التي تشتغل على المدونات والنصوص العربية تحليلا وتوسيما وإحصاء وبيانات وغيرها.

8. الهوامش والإحالات :

1- Paul Baker, Andrew Hardie & Tony McEnery; (2006); A glossary of corpus linguistics; Edinburgh University Press Ltd; UK; P48.
2- Jack C. Richards & Richard Schmidt; (2010); Longman dictionary of language teaching and applied linguistics; 4th edition; Pearson education; UK; P388.
3- David Crystal; (2008); A dictionary of linguistics and phonetics; 6th edition; Blackwell publishing; USA & UK; P322.

4- ينظر: موقع أيمن الدكروري للمعلومات اللسانية، قسم: لسانيات المدونة، عنوان : المدونات العربية، على الرابط:

- https://sites.google.com/a/aucegypt.edu/infoguistics/directory/Corpus-Linguistics/arabic-corpora

5- يمكن تصفح المدونة على الرابط:

- https://corpus.quran.com/

6- ينظر: أيمن الدكروري، (2018)، المدونات اللغوية ودورها في معالجة النصوص العربية، الطبعة الأولى، مركز الملك عبد الله بن عبد العزيز الدول لخدمة اللغة العربية ودار وجوه للنشر والتوزيع، الرياض السعودية، ص 57.

7- ينظر لمعلومات وافية حول المدونة، مع إمكانية طلب التسجيل للدخول، الرابط:

- http://www.bibalex.org/ica/ar/About.aspx

8- ينظر: أيمن الدكروري، (2018)، مرجع سابق، ص 58.

9- ينظر لمعلومات وافية حول المدونة، الرابط:

- https://www.arabiclearnercorpus.com/about-the-corpus-en

10- Paul Baker, Andrew Hardie & Tony McEnery; (2006); A glossary of corpus linguistics; P13.

11- Ibid; P154.

12- ينظر لمزيد من الشرح والتفصيل في هذه المسألة:

- Mohamed Zakaria Kurdi; (2016); Natural Language Processing and Computational Linguistics 1: Speech, Morphology and Syntax; 1st edition; ISTE Ltd, John Wiley & Sons; UK & USA; PP18-22.

وينظر أيضا: أيمن الدكروري، (2018)، مرجع سابق، ص ص 67-78.

13- محمود صالح إسماعيل وآخرون، (2015)، المدونات اللغوية العربية بناؤها وطرائق الإفادة منها، الطبعة الأولى، منشورات الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض السعودية، ص 166.

14- المرجع نفسه، ص 70.

15- المرجع نفسه، ص 70.

16- المرجع نفسه، ص 70.

17- المرجع نفسه، ص 70.

18- أيمن الدكروري، (2018)، مرجع سابق، ص 72.

19- ينظر: المعتز بالله سعيد، (2017)، كيف نبني مدونة لغوية موسمة تركيبيا، مجلة الدراسات اللغوية، مركز الملك فيصل للبحوث والدراسات الإسلامية، الرياض السعودية، المجلد 19، العدد 03، الصفحات (313 - 343)، ص 316.

20- المرجع نفسه، ص 316.

21- المرجع نفسه، ص 317.

22- المرجع نفسه، ص 318.

23- يمكن مطالعة معلومات مدونة (EASC) على موقع جامعة إسكس، على الرابط:

- https://www1.essex.ac.uk/linguistics/research/arabic/arabiccorpora/easc.aspx

24- يمكن تحميل لائحة الرموز التي يشتغل عليها برنامج (TagAnt) على الموقع الرسمي للباحث لورنس أنطوني مطور حزمة برامج (Ant) على عنوان الرابط التالي:

- https://www.laurenceanthony.net/software/tagant/

9.قائمة المصادر والمراجع:

- المعتز بالله سعيد، "**كيف نبني مدونة لغوية موسمة تركيبيا**"، مجلة الدراسات اللغوية، مركز الملك فيصل للبحوث والدراسات الإسلامية، الرياض السعودية، 2017، المجلد 19، العدد 03.
- أيمن الدكروري، **المدونات اللغوية ودورها في معالجة النصوص العربية**، مركز الملك عبد الله بن عبد العزيز الدول لخدمة اللغة العربية ودار وجوه للنشر والتوزيع، الطبعة الأولى، الرياض السعودية، 2018.
- محمود صالح إسماعيل وآخرون، **المدونات اللغوية العربية بناؤها وطرائق الإفادة منها**، منشورات الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الطبعة الأولى، الرياض السعودية، 2015.

- David Crystal; **A dictionary of linguistics and phonetics**; Blackwell publishing; 6th edition; USA & UK; 2008.
- Jack C. Richards & Richard Schmidt; **Longman dictionary of language teaching and applied linguistics**; Pearson education; 4th edition; UK; 2010.
- Mohamed Zakaria Kurdi; Natural **Language Processing and Computational Linguistics** 1: Speech, Morphology and Syntax; 1st edition; ISTE Ltd, John Wiley & Sons; UK & USA; 2016.
- Paul Baker, Andrew Hardie & Tony McEnery; A **glossary of corpus linguistics**; Edinburgh University Press Ltd; UK; 2006.
- https://corpus.quran.com/
- http://www.bibalex.org/ica/ar/About.aspx
- https://www.arabiclearnercorpus.com/about-the-corpus-en
- https://www1.essex.ac.uk/linguistics/research/arabic/arabiccorpora/easc.aspx
- https://www.laurenceanthony.net/software/tagant/
- https://sites.google.com/a/aucegypt.edu/infoguistics/directory/Corpus-Linguistics/arabic-corpora