كيفَ نبني مُدوَّنةً لُغَويَّةً مُوَسَّمةً تركيبيًّا للُّغة العربيَّة بطريقة نصف آليَّة

المُعتزّ بالله السَّعيد

جامعة القاهرة

ملخص البحث:

الكلمات الدَّالَّة:

المُدوَّنة اللُّغَويَّة المُوَسَّمة، التَّحليل التَّركيبيّ، أقسام الكلام، النَّحو العدديّ، الكشَّاف السِّياقيّ.

تتعلَّقُ هذه الورقة بحوسبة النَّحو العربيّ، وهو مَيدانٌ يقومُ على مجموعةٍ من الموارد اللُّغويَّة والحاسُوبيَّة. ومن هذه الموارد "المُدوَّنة اللُّغويَّة المُوَسَّمة تركيبيّاً" الَّتي تُستَخدَمُ في استكشاف أنماط الجُملة العربيَّة وتعيين مواطن الالتباس التَّركيبيّ فيها. وتَسعى هذه الدِّراسةُ إلى تقديم منهجيَّةٍ لبناء مُدوَّنةٍ لُغَويَّةٍ مُوَسَّمةٍ تركيبيّاً للُّغة العربيَّة بطريقة نصف آليَّة. وتهدفُ الدِّراسةُ إلى إيجاد وسيلةٍ لبناء هذا النَّوع من المُدوَّنات الَّتي تُعَدُّ مورداً رئيساً لبناء أدوات التَّحليل التَّركيبيّ للُّغة العربيَّة وأدوات التَّدقيق الإملائيّ لنُصُوصها. وتقوم فكرةُ الدِّراسة على توظيف تقنيات النَّحو العدديّ والكشَّاف السِّياقيّ في تحديد القرائن الدَّالَّة على أقسام الكلام العربيّ. ويأتي هذا البحثُ اختباراً لإمكانيَّة بناء مُدوَّنة لُغويَّة وتوجيه الحاسوب إلى توسيم هذه المُدوَّنة تركيبيّاً عبرَ إجراءاتٍ تقنيَّةٍ مُنتظمة. وقد جاءَت الدِّراسةُ في سبعة محاورَ أساسيَّة تتضَمَّن مُقدِّمةً ثُمَّ عرضاً لإشكالات الدِّراسة. ويلي ذلك تقديمُ المنهجيَّة المُقترَحة لبناء المُدوَّنة اللُّغويَّة المُوسَّمة، ثُمَّ التَّطبيقُ والنَّموذجُ، فنتائجُ الدِّراسة وخُلاصةُ البحث.

١. مُقدِّمة.

١,١. في المُدوَّنات اللُّغويَّة.

يُعَدُّ منهجُ البحث في لسانيَّات المُدوَّنة Corpus Linguistics حديثاً نسبيّاً؛ حيثُ ظهرَ في الولايات المُتَّحدة في السِّتِّينيَّات من القرن الماضي، ونما في الثَّمانينيَّات (نينغ، ٢٠١٦: ٣٦)، ولم تعرفهُ اللُّغة العربيَّة إلا قريباً من مطلع القرن الحادي والعشرين. ورغم محدوديَّة التَّجارب الَّتي حاولَت الإفادة من المُدوَّنات اللُّغويَّة، إلا أنَّ تأثيرَ ذلكَ أصبحَ واضحاً؛ حيثُ ساعدَ استثمارُ المُدوَّنات في تقويم نتائج البحث اللُّغويّ وتأكيد أو ترجيح كثيرٍ من الفرضيَّات حولَ اللُّغة أو نفي فرضيَّاتٍ أخرى. كذلك فقد أمكنَ استثمارُ المُدوَّنات في صناعاتٍ لُغويَّةٍ مُتعدِّدة باعتبارِها مورداً مُمثِّلاً لواقع اللُّغة الطَّبيعيَّة. وبدا الانضباطُ والتَّجانُسُ في مُخرَجات كثيرٍ من هذه الصِّناعات، لا سيَّما في ميادين مُعالجة اللُّغات الطَّبيعيَّة Natural Language Processing والصِّناعة المُعجميَّة Lexicography. ومع تطوُّر هذا المنهج من ناحية، وتشعُّب موضوعاته بينَ العُلُوم الإنسانيَّة والعُلُوم البحتة من ناحيةٍ أخرى، فقد دعَت الحاجةُ إلى بناء مُدوَّناتٍ لُغويَّةٍ كبيرة الحجم، تكونُ أكثرَ تمثيلاً وتعبيراً عن اللُّغة.

١,٢. التَّوسيم التَّركيبيّ للنُّصُوص Syntactic Tagging.

التَّوسيمُ التَّركيبيُّ Syntactic Tagging أحدُ إجراءَين لتوصيف الكلمات وفقَ النِّظام التَّركيبيّ للُّغة. ويُعنى بإضافة المعلومات التَّركيبيَّة في صُورةٍ خطِّيَّة [أفُقيَّة]. أمَّا الإجراءُ الآخرُ فهو "الإعراب Parsing" الَّذي يُعنى بإضافة المعلومات التَّركيبيَّة في صُورةٍ شجريَّة [رأسيَّة]. ويُعرَفُ الإجراءان معاً بعمليَّة "العنونة التَّركيبيَّة Syntactic Annotation"؛ ومن خلالها تتحوَّلُ المُدوَّنةُ اللُّغويَّة الخام Raw Corpus إلى مُدوَّنةٍ مُعنوَنة [مُوَصَّفة] تركيبيّاً Syntactic Annotated Corpus (Abeillé, 2012: 173). وتُعنى

هذه الدِّراسةُ بالتَّوسيم التَّركيبيّ الَّذي يُستفادُ منه في ميادين مُعالجة اللُّغات الطَّبيعيَّة. ويتمُّ التَّوسيمُ التَّركيبيُّ باستخدام ما يُعرَفُ بـ "وُسُوم أقسام الكلام Parts of Speech Tags" (Habash, 2009: 91)؛ حيثُ يُرفَقُ رَمزٌ مُعَيَّنٌ بكلِّ قسمٍ من أقسام الكلام، وفقاً لطبيعة النِّظام التَّركيبيّ للُّغة، وفي ضوء الهدف من المُدَوَّنة اللُّغَويَّة.

١,٣. مناهج التَّحليل التَّركيبيّ الحاسُوبيّ للُّغة العربيَّة Arabic Syntactic Analysis.

يُمثِّلُ التَّحليلُ التَّركيبيُّ Syntactic Analysis مُستوًى وسيطاً بينَ مُستويات التَّحليل اللُّغَويّ الَّتي تبدأ بمُستوى التَّحليل الصَّوتيّ وتنتهي بمُستوى التَّحليل الدِّلاليّ. وينبثقُ هذا المُستوى عن علم التَّركيب Syntax (Levine, 2017: 7)؛ وهو العلم الَّذي يدرس مُكوِّنات الجُملة والعلاقات بين عناصرِها (Kiss, 2015: 3\2024)، ويُعنى بدراسة أنواع الجُمَل وقواعدُ الإعراب وكيفيَّة التَّأليف بين أقسام الكلام لتكوين جُملةٍ مُنتظمة (Aoun, 2010: 12)، كما يُعنَى بتحليل الوحدات المُكَوِّنة للتَّركيب النَّحويّ. ولأنَّ وَحدةَ التَّحليل التَّركيبيّ هي الجُملة Sentence، فقد دَعَت الحاجةُ إلى توظيف الحاسُوب في تحليل عناصِرِها من خلال أدوات التَّحليل التَّركيبيّ للنُّصُوص (Soudi, 2007: 215). وثمَّةَ ثلاثةُ مناهجَ يَعتمدُ عليها الباحثُونَ في بناء أدوات التَّحليل التَّركيبيّ المُوافقة لطبيعة اللُّغة العربيَّة؛ حيثُ يقومُ المنهجُ الأوَّلُ على المعطيات اللُّغَويَّة المستمدَّة من قواعد النَّحو العَرَبيّ، سواءٌ أكانت في صُورةِ قوالبَ أم قواعدِ بيانات؛ ويقومُ المنهجُ الثَّاني على خوارزميَّة التَّحليل التَّركيبيّ الَّتي تُمثِّلُ صُورةً رياضيَّةً لقواعد النَّحو العَرَبيّ. أمَّا المنهجُ الثَّالثُ، فيقومُ على المُعالجة الإحصائيَّة لقواعد التَّركيب المُستخلَصة من المدوَّنات اللُّغَويَّة العَربيَّة باعتبارها تمثيلاً لواقع اللُّغة؛ حيثُ يتمُّ تدريبُ نُصُوص المُدَوَّنات اللُّغويَّة بهدف استخلاص الأنماط التَّركيبيَّة (Ryding, 2014: 112)، ثُمَّ تهيئة الآلة لاستقبال النَّتائج والتَّفاعل معها. وهذا المنهجُ الثَّالثُ هو الأكثرُ نجاعةً ومُناسَبةً للُّغة العربيَّة – من

وُجهة نظر الباحث – لأسباب، أهمُّها: أنَّهُ يُراعي تعدُّديَّة أنماط الجُملة العربيَّة، ويُراعي القواعدَ السَّماعيَّة التَّركيبيَّة الَّتي يصعُبُ التَّعبيرُ عنها بلُغة الآلة. ولأجل هذا، فإنَّ هذه الدِّراسةَ تسعى إلى الإبانة عن آليَّات بناء مُدَوَّنة لُغَويَّة مُوسَّمة تركيبيّاً للُّغة العربيَّة، حيثُ يُشكِّلُ وُجُودُ مثلِ هذه المُدَوَّنة نقطةَ الانطلاق إلى تطوير أدواتٍ فعَّالةٍ للتَّحليل التَّركيبيّ العَربيّ.

٢. إشكالات بناء مُدوَّنةٍ لُغَويَّةٍ مُوَسَّمةٍ تركيبيّاً للُّغة العربيَّة.

يستغرقُ بناءُ المُدَوَّنات اللُّغَويَّة المُوَسَّمة تركيبيّاً للُّغة العربيَّة وقتاً وجهداً كبيرَين، الأمرُ الَّذي يُؤدِّي إلى زيادة تكلفة إنتاج هذا النَّوع من المُدَوَّنات. أضِف إلى ذلكَ أنَّ بناءَ المُدوَّنات المُوسَّمة يستدعي زيادةَ الموارد البشريَّة العاملة، لا سيَّما إذا تعلَّقَ الأمرُ بمُدَوَّناتٍ لُغَويَّةٍ كبيرةٍ نسبيّاً. وبالنَّظَر إلى طبيعة اللُّغة العربيَّة من ناحية، وواقع صناعة المُدوَّنات اللُّغَويَّة من ناحيةٍ أخرى، نستطيعُ أن نقفَ على ثلاثة إشكالاتٍ رئيسةٍ، نعرضُها فيما يلي.

١. المُرُونة في نظام بناء الجُملة العَرَبيَّة.

يتمتَّعُ نظامُ بناء الجُملة العَرَبيَّة بقدرٍ كبيرٍ من المُرُونة؛ حيثُ يسمحُ بالتَّقديم والتَّأخير بينَ عناصر الجُملة، كما يسمحُ بتعدُّد أنماط الجُملة وتمدُّد عناصِرها الَّتي قد تتجاوزُ أربعينَ عُنصراً. ومن ناحيةٍ أخرى، يسمحُ نظامُ الجُملة العَرَبيَّة بتبادُل العناصِرِ التَّالية لقسم الكلام المُحَدَّد. نُلاحظُ مثلاً أنَّ الضَّميرَ المُنفصِلَ الثَّابتَ في محلِّه الإعرابيّ يقبلُ أن يلحقَ به الاسم، نحو (أنتَ مُجتَهد)، ويقبلُ أن يلحقَ به الفعل، نحو (أنتَ تجتهِد)، ويقبلُ أن تلحقَ به الأداة، نحو (أنتَ لا تجتَهد) ... وهكذا. وتُمثِّلُ هذه المُرُونةُ إشكالاً عندَ توسيم المُدوَّنات اللُّغَويَّة تركيبيّاً، لأنَّها تستدعي عملاً يدويّاً شاقّاً للبحث عن قسم الكلام الَّذي يتبعُهُ كُلُّ عُنصرٍ من عناصِرِ الجُملةِ على حِدة. وحالَ التَّدخُّل الآليّ لتوسيم المُدَوَّنة، فإنَّ نسبةَ الخطأ لن

تكونَ قليلة. وهذا يستدعي تدخُّلاً يدويّاً كبيراً لمُعالجة الأخطاء النَّاجمة عن عمل الآلة.

٢. طبيعة النِّظام الكتابيّ [الجرافيميّ] للُّغة العربيَّة.

اللُّغةُ العربيَّةُ لُغةٌ اشتقاقيَّة، يسمحُ نظامُها الكتابيُّ بأن تتشابَكَ فيها الوحداتُ الكتابيَّةُ [الجرافيمات Graphemes] بينَ مجموعة الكلمات، على النَّحو الَّذي نجدُهُ مثلاً في المجموع الكتابيّ (فَسَيَكفيكَهُم) الَّذي يتكوَّنُ من خمس وحداتٍ صرفيَّة [مورفيمات Morphemes]، هي على التَّرتيب: (الفاء) و(السِّين) و(يكفي) و(الكاف) و(هم). ولكلِّ وحدةٍ من هذه الوحدات دلالةٌ تركيبيَّةٌ تجعلُها قسماً مُستقلّاً من أقسام الكلام؛ حيثُ تدلُّ الفاءُ على الاستئناف، وتدلُّ السِّينُ على التَّسويف، ويدلُّ الفعلُ على المُضارعة والاستمراريَّة، ويدلُّ الضَّمير (الكافُ) على المُخاطَب المُفرَد المُذكَّر، ويدلُّ الضَّمير (هم) على الغائب الجَمع المُذكَّر. ومن ناحيةٍ أخرى، فإنَّ بعضَ أقسام الكلام تتماثلُ في رَسمها الكتابيّ مع اختلاف مبناها، على نحو ما نجدُ في الكلمات (من، بل، هل)؛ حيثُ تحتملُ كُلٌّ منها أن تكونَ اسماً أو فعلاً أو أداةً، بحسب ضبطِها. ووفقاً لهذا النِّظام، فإنَّ توسيم المُدوَّنات اللُّغويَّة تركيبيّاً يفرضُ الجمعَ بينَ بعض أقسام الكلام المُتشابكة، كما يستدعي ضبطَ النُّصُوص بالشَّكل تحسُّباً للالتباس المُحتَمَلِ وُقُوعُه عندَ توسيم الكلمات المُتماثلة في رَسمها.

٣. الاختلاف حَولَ أقسام الكلام العَرَبيّ Arabic POS.

تتكوَّنُ الجُملةُ العَربيَّة من مجموعةٍ من العناصر الَّتي تُعرَفُ بـ "أقسام الكلام Parts of Speech". وقد صَنَّفَ النُّحاةُ الكلامَ العَرَبيَّ قديماً إلى ثلاثة أقسامٍ، هي: الاسم Noun والفعل Verb والحَرف Particle (السَّاقي، ١٩٧٧: ٣٣). ويَحيدُ بعضُ اللُّغَويِّينَ المُعاصِرينَ عن هذا التَّصنيف، فيذهبُ فريقٌ إلى تقسيم الكلام العَرَبيّ إلى

أربعة أقسامٍ، هي: الاسم والفِعل والحَرف والضَّمير (أنيس، ١٩٧٢: ٢٧٩). ويذهبُ فريقٌ آخَرُ إلى تقسيم الكلام العَرَبيّ إلى سبعة أقسامٍ، هي: الاسم والصِّفة والفِعل والضَّمير والخالفة والظَّرف والأداة (حسَّان، ١٩٧٩: ٨٦). وعلى جانبٍ آخَرَ يلجأ العاملُونَ في حَوسَبة اللُّغة إلى ابتكار تصنيفاتٍ أخرى في محاولةٍ لتمكين الآلة من التَّعامُل مع قواعد النَّحو العَرَبيّ، على النَّحو الَّذي نجدُهُ في مجموعة الموارد اللُّغَويَّة الَّتي تُنتجُها "مُؤسَّسة البيانات اللُّغَويَّة Linguistic Data Consortium (LDC)" (Zitouni, 2014: 118). ويُمثِّلُ هذا الاختلافُ إشكالاً عندَ توسيم المُدوَّنات اللُّغَويَّة تركيبيّاً؛ حيثُ يَستدعي تحديدَ الهدف من المُدوَّنة اللُّغَويَّة المُوَسَّمة، ثُمَّ بناء المُدَوَّنة وفقَ ما يُحقِّقُ هذا الهدف، كما يُقصِرُ الإفادةَ من المُدَوَّنات اللُّغَويَّة المُوَسَّمة على جوانبَ معلومةٍ سَلَفاً دونَ غَيرِها.

٣. منهجيَّة بناء مُدوَّنةٍ لُغَويَّةٍ مُوَسَّمةٍ تركيبيّاً للُّغة العربيَّة بطريقة نصف آليَّة.

قبلَ الشُّرُوع في بناء المُدَوَّنة اللُّغَويَّة المنشُودة، ينبغي أن نُحدِّدَ الهدفَ منها، لأنَّ حجمَ المُدَوَّنة وطبيعةَ النُّصُوص الَّتي تحويها يخضعان لذلك الهدف. والواقعُ أنَّ ما ننشُدُهُ في دراستِنا هو مُدَوَّنةٌ لُغويَّةٌ يُستفادُ منها في أغراض التَّحليل التَّركيبيّ للُّغة العربيَّة المُستخدَمة فعليّاً، على النَّحو الَّذي يضمنُ صلاحيَّةَ هذه المُدوَّنة للتَّدريب والاختبار في مختلف تطبيقات التَّحليل التَّركيبيّ للُّغة العربيَّة، لا سيَّما في ميادين البحث التِّقَنيّ الأكثر نشاطاً وحاجةً إلى هذا النَّوع من المُدوَّنات، كالتَّرجمة الآليَّة، وأدوات التَّشكيل الآليّ، وأدوات فكّ الالتباس اللُّغويّ، وغيرها.

وتحقيقاً لهذه الغاية، فقد وضعَ الباحثُ ثلاثةَ ضوابطَ أساسيَّة للمُدوَّنة المنشودة:

١. استيعاب المُدوَّنة لأنماطٍ تركيبيَّة مُتعدِّدة.

٢. اعتماد المُدَوَّنةِ على نُصُوص اللُّغة العربيَّة المُعاصرة.

٣. تنوُّع المادَّة المُتَضَمَّنة لتكونَ تمثيلاً حقيقيّاً لواقع اللُّغة العربيَّة.

وفي ضوء ذلك، ستعرضُ الدِّراسةُ فيما يلي لمنهجيَّة بناء المُدَوَّنة اللُّغَويَّة المنشودة عبرَ مجموعةٍ من المراحل المُتعاقبة.

٣,١. بناء المُدوَّنة اللُّغَويَّة الخام Raw Corpus.

ثمَّةَ ثلاثةُ أساليبَ رئيسة لبناء المُدوَّنات اللُّغَويَّة المُحَوسَبة، هي: أسلوب "الحصر الشَّامل Comprehensive Inventory"، وأسلوب "الاستبانة Questionnaire"، وأسلوب "العَيِّنات الإحصائيَّة Statistical Sampling" (السَّعيد، ٢٠١١: ٣٤). ولأنَّ هذه الدِّراسةَ تنشُدُ مُدوَّنةً مُمثِّلةً لعُموم اللُّغة العربيَّة المُعاصرة، وتسعى إلى توجيه مادَّتِها لمُختلف أغراض التَّحليل التَّركيبيّ، فقد اختارَ الباحثُ العملَ وفقَ أسلوب "العَيِّنات الإحصائيَّة"؛ إذ هُو الأكثرُ مُناسبةً لأهداف الدِّراسة؛ كما أنَّهُ يتَّسمُ بقَدرٍ كبيرٍ من المُرُونة الَّتي تُساعدُ على تمثيل لُغة المُجتمَع.

ولأنَّ مرحلةَ "بناء المُدوَّنة اللُّغَويَّة الخام" تُمثِّلُ الأساسَ الَّذي تنطلقُ منهُ المُعالَجاتُ الآليَّةُ والإحصائيَّةُ الرَّاميةُ إلى التَّوظيف الأمثل للنُّصُوص، فقد صنَعَ الباحثُ مُدَوَّنةً لُغَويَّةً مُمثِّلةً للعربيَّة المُعاصِرة في صُورة عَيِّنةٍ قصديَّة [غَرَضِيَّة] Pur-posive Sampling مُنتقاةٍ من أربعة مصادرَ على النَّحو الآتي:

١. وثائق موسُوعة ويكيبيديا:

اشتملَت المُدوَّنةُ على مقالاتٍ مُختارة من موسوعة ويكيبيديا خلالَ المُدَّة من عام ٢٠٠٣ إلى ٢٠١٦. وبلغَ عددُ كلمات هذه المادَّة ٥٥٩,٦٩٥ كلمة، بنسبة ٢٥٪ من جُملة المُدوَّنة. وتأتي هذه المادَّة تمثيلاً لمُختلف أنماط العربيَّة.

٢. وثائق الصَّحافة العربيَّة:

اشتملَت المُدوَّنةُ على مجموعةٍ مُختارة من مقالات الصُّحُف العربيَّة خلالَ المُدَّة من عام ٢٠٠٨ إلى ٢٠١٥. وبلغَ عددُ كلمات هذه المادَّة ٥٥٩,٦٨٨ كلمة، بنسبة ٢٥٪

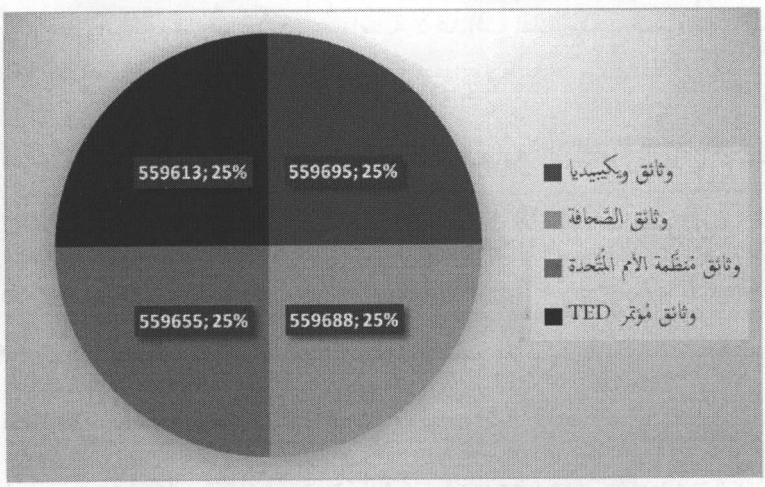
من جُملة المُدوَّنة. وتأتي هذه المادَّة تمثيلاً للُّغة العربيَّة العامَّة.

٣. وثائق مُنظَّمة الأُمم المُتَّحدة:

اشتملَت المُدوَّنةُ على مجموعةٍ من قرارات مُنظَّمة الأُمم المُتَّحدة خلالَ المُدَّة من عام ٢٠٠٠ إلى ٢٠١٠. وبلغَ عددُ كلمات هذه المادَّة ٥٥٩,٦٥٥ كلمة، بنسبة ٢٥٪ من جُملة المُدوَّنة. وتأتي هذه المادَّة تمثيلاً للُّغة العربيَّة الرَّسميَّة.

٤. وثائق مُؤتمر (تيد TED):

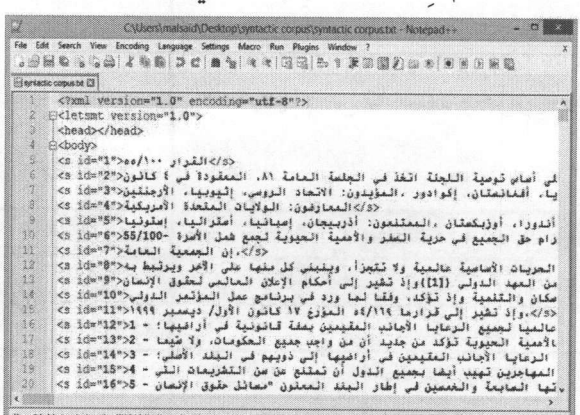
اشتملَت المُدوَّنةُ على مجموعةٍ مُترجمةٍ من مُحاضرات المؤتمر الأكاديميّ (تيد TED) خلالَ عام ٢٠١٣. وبلغَ عددُ كلمات هذه المادَّة ٥٥٩,٦١٣ كلمة، بنسبة ٢٥٪ من جُملة المُدوَّنة. وتأتي هذه المادَّة تمثيلاً للُّغة العربيَّة العلميَّة.



الشَّكل ١: مُخَطَّط بيانيّ لوثائق المُدوَّنة اللُّغويَّة

ووفقاً لهذا التَّصنيف، فقد بلغَت جُملةُ كلمات المُدوَّنة اللُّغَويَّة (٢,٢٣٨,٦٥١) كلمة. أمَّا عددُ الكلمات الفريدة Unique Words فقد بلغَ (١٥٥,٠٨٣) كلمة [قبلَ التَّنقية من الحواشي وعلامات الضَّبط]، وبلغَ (١٤٩,٠٩٩) كلمة [بعدَ التَّنقية]. ويُوضِّحُ (الشَّكل ٢) نموذجَ المُدَوَّنة اللُّغَويَّة في صُورتِها الخام Raw Corpus، قبلَ

توسيمِها تركيبيّاً، باستخدام مِنَصَّة التَّحرير المكتبيّ (Notepad++.v.7.1) .



الشَّكل ٢ : نموذج المُدوَّنة اللُّغويَّة الخام – مِنَصَّة التَّحرير (Notepad++.v.7.1)

٣,٢ . تحديد وُسُوم أقسام الكلام POS Tags .

ذكرنا آنفاً أنَّ التَّوسيمَ التَّركيبيَّ هو أحدُ إجراءَين لما يُعرَفُ بـ "العَنونة التَّركيبيَّة" ، وأنَّ الإجراءَ الآخَرَ هُو "الإعراب" . وتُعنى هذه الدِّراسةُ بالإجراء الأوَّل [التَّوسيم]؛ إذ هُو الإجراءُ الَّذي يُمكنُ توجيهُ الآلة إلى مُحاكاة ذكاء الإنسان في فهمِه. أمَّا الإجراءُ الآخَرُ [الإعراب] فيستدعي توصيفاً دقيقاً للموقع الإعرابيِّ الَّذي تشغلُهُ كُلُّ كلمةٍ على حِدة؛ وهو أمرٌ يصعبُ إدراكُهُ عبرَ الآلة، خُصُوصاً مع اللُّغة العربيَّة الَّتي تتعدَّدُ فيها أنماطُ الجُملة وأساليبُ الإعراب. ولَمَّا كانَ من أهداف الدِّراسة أن نُوجِدَ وسيلةً لإخضاع الآلة لفهم قواعد النَّحو العَرَبيّ، كانَ لزاماً أن نخرُجَ عن الأطُر التَّقليديَّة الَّتي وَضَعَها النُّحاةُ لأقسام الكلام إلى إطارٍ يُمَكِّنُ الآلةَ من استيعاب هذه الأقسام. ولأجل هذا فإنَّ الدِّراسةَ تقترحُ تقسيمَ الكلام إلى خمسة أقسامٍ رئيسة، يتفرَّعُ عنها خمسة عشر قسماً فرعيّاً على النَّحو الوارد في (الجدول ١)، مع مُلاحظة أنَّ هذا التَّقسيمَ يَضُمُّ الصِّفات إلى الأسماء، ويضُمُّ الخوالفَ إلى الأفعال، كما يُخالفُ ما جرى عليه النُّحاةُ بشأن الكلمات الدَّالَّة على

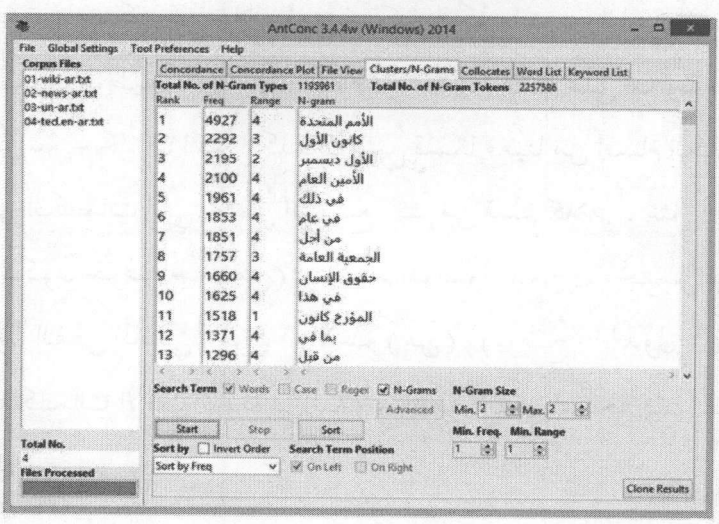
الاستفهام، حيثُ يُوزِّعُونَها بينَ الأدوات (نحو: الهمزة، هل) والأسماء (نحو: أينَ، متى، كيفَ، لماذا، ...). ومنهجُ الدِّراسة أن تُوضَعَ هذه الكلماتُ ضمنَ الأدوات لدلالتها جميعاً على الاستفهام من ناحية، وجواز إحلالِ بعضِها مكانَ بعضٍ من ناحيةٍ أخرى.

م	قسم الكلام	المثال	المصطلح الإنجليزي	الرمز POS Tag
	الاسم Noun			
١	الاسم الشَّائع (غير العَلَم)	إنسان	Common noun	[CN]
٢	اسم العَلَم	مُحمَّد	Proper Noun	[PN]
٣	اسم الإشارة	هذا	Determiner	[DE]
٤	الاسم الموصول	الَّذي	Relative Pronoun	[RP]
٥	العدد/الرَّقم	واحد	Cardinal Number	[CNU]
	الفِعل Verb			
٦	فِعل مُضارع	يكتُب	Imperfect Verb	[VI]
٧	فِعل ماضٍ	كَتَبَ	Perfect Verb	[VP]
٨	فعل طلب (أمر)	اكتُبْ	Request Verb	[VR]
	الأداة Particle			
٩	أداة استفهام	هَل	Question	[QU]
١٠	أداة استثناء	إلَّا	Exception	[EX]
١١	أداة رَبط	أو	Conjunction	[CO]
١٢	حَرف جرّ	على	Preposition	[PRE]
١٣	أداة أخرى	قَد	Other Particle	[PO]
	الضَّمير Pronoun			
١٤	الضَّمير	نَحنُ	Pronoun	[PRO]
	الظَّرف Adverb			
١٥	الظَّرف	حَيثُ	Adverb	[AD]

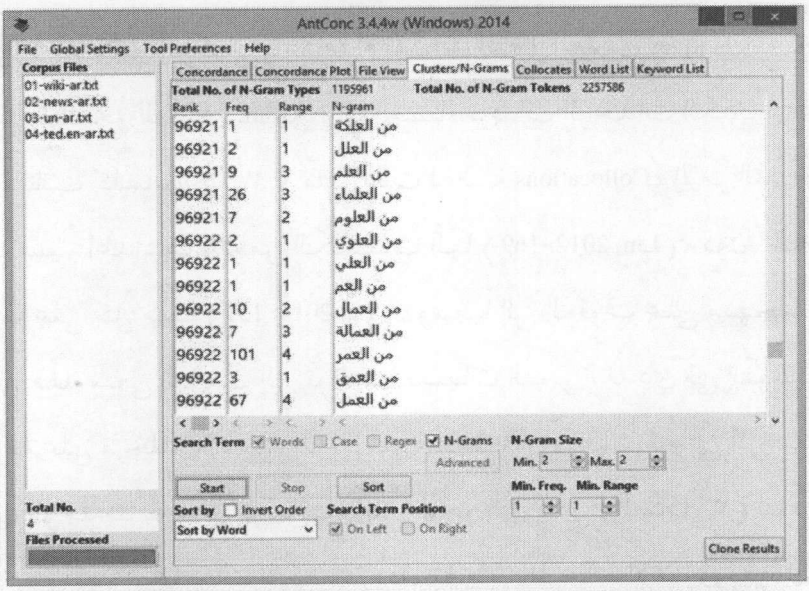
الجدول ١ : مُقترَح التَّقسيم الخُماسيّ للكلام العربيّ ووُسُوم الأقسام POS Tags المُستخدَمة في التَّوسيم التَّركيبيّ

٣,٣ . التَّوسيم التَّركيبيّ باستخدام تقنيات النَّحو العدديّ N-Gram techniques .

يُساعدُ النَّحوُ العدديُّ N-Gram في إحصاء تردُّدات الوحدات الكتابيَّة الكبرى، سواءٌ أكانت كلمات Words أم مُتلازمات لفظيَّة Collocations، الأمرُ الَّذي يُمكنُ مَعَهُ توسيمُ أعدادٍ هائلةٍ من الكلمات آليّاً (Lin, 2012: 169)، دونَ الحاجة إلى الوُقُوفِ على كُلٍّ منها (Lu, 2014: 151) . وسعياً إلى الوُقوف على منهجيَّةٍ ناجعةٍ وفعَّالة، فقد سعى الباحثُ إلى توظيف تقنيات النَّحو العدديّ في التَّوسيم الآليّ للمُدوَّنة على مرحلتَين [باستخدام برمجيَّة AntConc]؛ حيثُ قامَ في المرحلة الأولى بترتيب الوحدات بحسب تردُّداتِها، على النَّحو الوارد في (الشَّكل ٣) لتُؤدِّي الآلةُ دورَها بتوسيم الكلمات الكثر تردُّداً، وقامَ في المرحلة الأخرى بترتيب هذه الوحدات ألفبائيّاً، على النَّحو الوارد في (الشَّكل ٤) لتُؤدِّي الآلةُ دورَها بإزالة الالتباس الحادث في الكلمات الَّتي تتَّفقُ في رَسمِها وتختلفُ في قسم الكلام الَّذي تتبعُه، بما يُمكِّنُ من تمييزِها .



الشَّكل ٣ : ترتيب الوحدات الكتابيَّة الكبرى باستخدام تقنيات النَّحو العدديّ بحسب تردُّداتها – برمجيَّة AntConc 3.4



الشَّكل ٤ : ترتيب الوحدات الكتابيَّة الكبرى باستخدام تقنيات النَّحو العدديّ ألفبائيّاً – برمجيَّة AntConc 3.4

وتوضيحاً لهذه المنهجيَّة، تقترحُ الدِّراسةُ أن ننطلقَ من مُستوى النَّحو الأحاديّ Uni-gram؛ حيثُ يُمكنُ من خلالِه توسيمُ أكثرُ الكلمات دوراناً في النُّصُوص، لا سيَّما الكلمات الوظيفيَّة Function words. وبتطبيق ذلكَ على المُدَوَّنة اللُّغَويَّة – موضُوع الدِّراسة – نجدُ أنَّ أكثرَ الكلمات تتبعُ قسماً مُعَيَّناً من أقسام الكلام؛ لكنَّنا سنجدُ بعضَ الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ، مثل (من) الَّتي تحتملُ أن تكونَ حرفَ الجرِّ (مِن) أو الاسم الموصول (مَن)، وتحتملُ في حالاتٍ أقلَّ أن تكونَ الفعلَ الماضي (مَنَّ) أو الاسم (مَنّ). ويُوَضِّحُ (الجدول ٢) التَّوسيم التَّركيبيّ للكلمات الأكثر تردُّداً في المُدوَّنة اللُّغَويَّة بعدَ استخلاصِها باستخدام النَّحو الأحاديّ.

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
١	الأمم المتحدة	٤٩٢٧	[CN]الأمم [CN]المتحدة
٢	كانون الأول	٢٢٩٢	[PN]كانون [CNU]الأول
٣	الأول ديسمبر	٢١٩٥	[CNU]الأول [PN]ديسمبر
٤	الأمين العام	٢١٠٠	[CN]الأمين [CN]العام
٥	في ذلك	١٩٦١	[PRE]في [DE]ذلك
٦	في عام	١٨٥٢	[PRE]في [AD]عام
٧	من أجل	١٨٥١	[PRE]من [CN]أجل
٨	الجمعية العامة	١٧٥٧	[CN]الجمعية [CN]العامة
٩	حقوق الإنسان	١٦٦٠	[CN]حقوق [CN]الإنسان
١٠	في هذا	١٦٢٥	[PRE]في [DE]هذا
١١	المؤرخ كانون	١٥١٨	[CN]المؤرخ [PN]كانون
١٢	بما في	١٥١٨	[PRE]ب [RP]ما [PRE]في
١٣	من قبل	١٢٩٦	[PRE]من [AD]قبل
١٤	من خلال	١٢٩٣	[PRE]من [AD]خلال
١٥	الولايات المتحدة	١١٤٦	[CN]الولايات [CN]المتحدة
١٦	في المائة	١١٠٢	[PRE]في [CNU]المائة
١٧	ذات الصلة	١٠٧٥	[CN]ذات [CN]الصلة
١٨	الكثير من	١٠٦٩	[CN]الكثير [PRE]من
١٩	إلى أن	١٠٥٦	[PRE]إلى [PO]أن
٢٠	أكثر من	١٠٢٤	[CN]أكثر [PRE]من
٢١	إلى الأمين	١٠١٣	[PRE]إلى [CN]الأمين
٢٢	عن طريق	١٠١٠	[PRE]عن [CN]طريق
٢٣	يمكن أن	٩٨٥	[VI]يمكن [PO]أن
٢٤	في مجال	٩٥٨	[PRE]في [CN]مجال
٢٥	الشرق الأوسط	٩٤٦	[CN]الشرق [CN]الأوسط

الجدول ٢ : التَّوسيم التَّركيبيّ للكلمات الأكثر تردُّداً في المُدوَّنة اللُّغويَّة (النَّحو الأحاديّ)

وحتَّى نتمكَّنَ من توسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ، تقترحُ الدِّراسةُ الانتقالَ إلى التَّوسيمِ على مُستوى النَّحو الثُّنائيّ Bi-gram ثُمَّ النَّحو الثُّلاثيّ Tri-gram ...، وهكذا، إلى أن تقلَّ احتمالاتُ تعدُّد الأقسام الكلاميَّة

للكلمة الواحدة. نُلاحظُ مثلاً عندَ توسيم المُدَوَّنة اللُّغويَّة على مُستوى النَّحو الثُّنائيّ أنَّ الكلمات المُلازمة لكلمة (من) تُقلِّلُ من احتمالات تعدُّد أقسام الكلام بصورةٍ كبيرة. ومع هذا تبقى احتماليَّةُ تعدُّد الأقسام في بعض السِّياقات، كما في الثُّنائيَّات (أكثر من، عدد من، كُلّ من، ...)، وهو أمرٌ يُمكنُ معالجتُهُ باستخدام النَّحو العدديّ الثُّلاثيّ. ويُوَضِّحُ (الجدول ٣) التَّوسيم التَّركيبيّ لثُنائيَّات الكلمات الأكثر دوراناً في المُدوَّنة اللُّغَويَّة بعدَ استخلاصِها باستخدام النَّحو الثُّنائيّ.

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
١	الأمم المتحدة	٤٩٢٧	الأمم [CN] المتحدة [CN]
٢	كانون الأول	٢٢٩٢	كانون [PN] الأول [CNU]
٣	الأول ديسمبر	٢١٩٥	الأول [CNU] ديسمبر [PN]
٤	الأمين العام	٢١٠٠	الأمين [CN] العام [CN]
٥	في ذلك	١٩٦١	في [PRE] ذلك [DE]
٦	في عام	١٨٥٢	في [PRE] عام [AD]
٧	من أجل	١٨٥١	من [PRE] أجل [CN]
٨	الجمعية العامة	١٧٥٧	الجمعية [CN] العامة [CN]
٩	حقوق الإنسان	١٦٦٠	حقوق [CN] الإنسان [CN]
١٠	في هذا	١٦٢٥	في [PRE] هذا [DE]
١١	المؤرخ كانون	١٥١٨	المؤرخ [CN] كانون [PN]
١٢	بما في	١٥١٨	ب [RP] ما [PRE] في [PRE]
١٣	من قبل	١٢٩٦	من [PRE] قبل [AD]
١٤	من خلال	١٢٩٣	من [PRE] خلال [AD]
١٥	الولايات المتحدة	١١٤٦	الولايات [CN] المتحدة [CN]
١٦	في المائة	١١٠٢	في [PRE] المائة [CNU]
١٧	ذات الصلة	١٠٧٥	ذات [CN] الصلة [CN]
١٨	الكثير من	١٠٦٩	الكثير [CN] من [PRE]
١٩	إلى أن	١٠٥٦	إلى [PRE] أن [PO]
٢٠	أكثر من	١٠٢٤	أكثر [CN] من [PRE]
٢١	إلى الأمين	١٠١٣	إلى [PRE] الأمين [CN]
٢٢	عن طريق	١٠١٠	عن [PRE] طريق [CN]
٢٣	يمكن أن	٩٨٥	يمكن [VI] أن [PO]
٢٤	في مجال	٩٥٨	في [PRE] مجال [CN]
٢٥	الشرق الأوسط	٩٤٦	الشرق [CN] الأوسط [CN]

الجدول ٣: التَّوسيم التَّركيبيّ لثُنائيَّات الكلمات الأكثر تردُّداً في المُدوَّنة اللُّغَويَّة (النَّحو الثُّنائيّ)

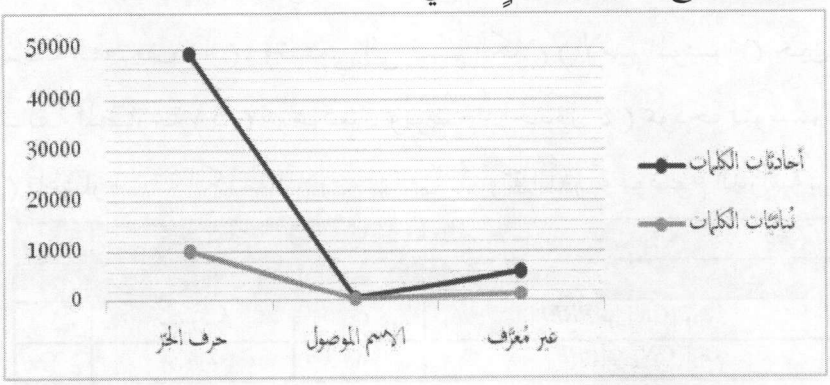
وبتطبيق تقنيات النَّحو العدديّ على كلمات المُدوَّنة، نلمسُ نتيجةً حقيقيَّة عندَ توسيم الكلمات المُتردِّدة الَّتي لا تحتملُ أكثرَ من قسمٍ كلاميٍّ، سواءٌ على مُستوى النَّحو الأحاديّ أم النَّحو الثُّنائيّ. ومع هذا، يبقى إشكالُ توسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميٍّ قائماً، إذ ينبغي أن نتحقَّقَ من قسم الكلام الصَّحيح لكلِّ سياقٍ على حِدة. وسعياً إلى مُعالجة هذا الإشكال، تقترحُ الدِّراسةُ إعادةَ ترتيب ثُنائيَّات الكلمات ألفبائيّاً، ثُمَّ بناء خوارزميَّة التَّوسيم آليّاً بالنَّظر إلى سوابق الكلمات المُلازمة للكلمة الَّتي ننشُدُ توسيمَها. وعلى سبيل المثال، سنُلاحظُ أنَّ كلمة (من) تنتمي إلى قسم الكلام [حرف الجرّ] حينَ تلحقُ بها سابقةُ التَّعريف (الـ)، وتنتمي إلى قسم الكلام (الاسم الموصول) حينَ تلحقُ بها سابقة المُضارَعة (يتـ)، وهكذا. ويُوَضِّحُ (الجدول ٤) نموذجاً لتوسيم كلمة (من) باعتبار سابقة الكلمة المُلازمة لها في المُدَوَّنة اللُّغويَّة موضوعِ الدِّراسة.

م	الكلمة	التَّردُّد	التَّوسيم التَّركيبيّ
سابقة الكلمة المُلازمة (الـ)			
١	من الوصول	٣٥	[PRE]من [CN]الوصول
٢	من الوضع	٨	[PRE]من [CN]الوضع
٣	من الوضوح	٢	[PRE]من [CN]الوضوح
٤	من الوعي	٦	[PRE]من [CN]الوعي
سابقة الكلمة المُلازمة (يتـ)			
٥	من يتصف	١	[RP]من [VI]يتصف
٦	من يتعاون	١	[RP]من [VI]يتعاون
٧	من يتقدم	١	[RP]من [VI]يتقدم
٨	من يتكلم	١	[RP]من [VI]يتكلم

الجدول ٤: نموذج التَّوسيم التَّركيبيّ لكلمة (من) باعتبار سابقة الكلمة المُلازمة في المُدَوَّنة اللُّغويَّة

لقد وردَت كلمة (من) في المُدَوَّنة اللُّغويَّة في (٥٥,٨٣٤) سياق، تضمُّهم (١١,٧٨٧) ثُنائيَّة. وقامَ الباحثُ بالتَّوسيم التَّركيبيّ للكلمة آليّاً باستخدام تقنيات النَّحو العَدَديّ، فكانت النَّتيجةُ أن أمكنَ التَّعرُّفُ على قسم الكلام الَّذي

تتبعُهُ الكلمة في ٨٨,٨ ٪ من السِّياقات، منها ٨٧,٨ ٪ تنتمي إلى [حرف الجرّ PRE]، بواقع (٤٩,٠٤٥) سياق، تضمُّهم (١٠,٠٧٦) ثُنائيَّة، و ١٪ تنتمي إلى [الاسم الموصول RP]، بواقع (٦١٩) سياق، تضمُّهم (٣٢٥) ثُنائيَّة. وفي مُقابل ذلك لم تسمح تقنياتُ النَّحو العدديّ بالتَّعرُّف على ١١,٢٪ من السِّياقات، بواقع (٦,١٧٠) سياق، تضمُّهم (١,٣٨٦) ثُنائيَّة، على النَّحو الوارد في (الشَّكل ٥)؛ وهُوَ ما يعني إمكانيَّةَ الإفادة من النَّحو العدديّ لتوسيم ٨٨,٨٪ من السِّياقات الَّتي وَرَدت فيها كلمة (من) بصورةٍ آليَّة؛ وقِس على المنهجيَّة مجموعةَ الكلماتِ الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميّ.

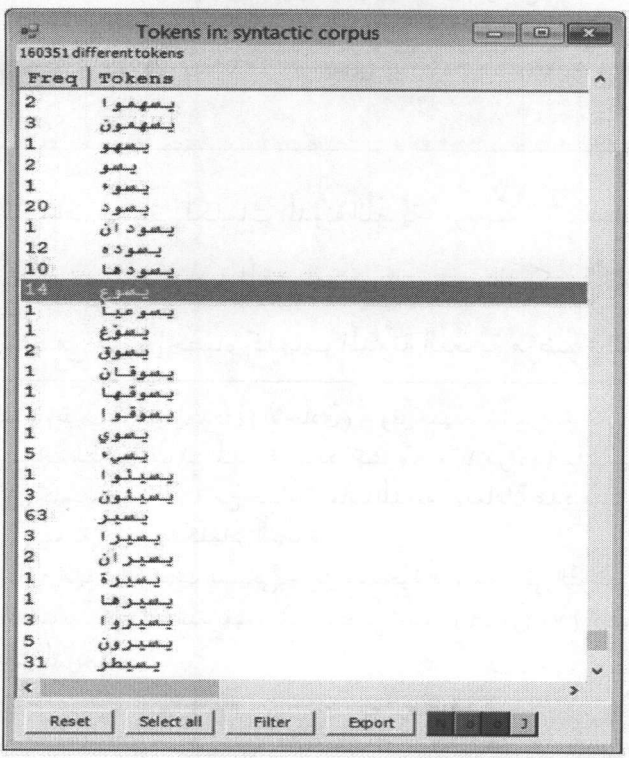


الشَّكل ٥: مُخَطَّط بيانيّ خطّيّ لنتائج التَّوسيم التَّركيبيّ لكلمة (من) في المُدَوَّنة اللُّغَويَّة باستخدام تقنيات النَّحو العدديّ

٣,٤. التَّوسيم التَّركيبيّ باستخدام الكشَّاف السِّياقيّ Concordancer.

تزدادُ قُدرةُ النَّحو العَدديّ على التَّوسيم التَّركيبيّ في الكلمات الأكثر تردُّداً، لكنَّها قد لا تكونُ مُجديةً بصورةٍ كبيرة في الكلمات الأقلّ تردُّداً. لهذا، تقترحُ الدِّراسةُ إكمالَ توسيم المُدَوَّنة باستخدام "الكشَّاف السِّياقيّ Concordancer"؛ وهُوَ تطبيقٌ مُتمِّمٌ لعمل المُفهرس الآليّ للنُّصُوص Text Indexer، يَسمحُ باستكشاف جميع كلمات المُدوَّنة في سياقاتها (Lehmann, 2016: 169)، بما يُتيحُ تعقُّبَ

الوحدات الكتابيَّة الَّتي تُلزمُ الكلمةَ قسماً كلاميّاً مُعيَّناً. ويُساعدُ الكشَّافُ السِّياقيُّ أيضاً في مُراجعة المُطابقة بينَ كلمات المُدوَّنة وأقسام الكلام الَّتي تُوسَمُ بها إذا احتملَت الكلمةُ التَّوسيمَ بأكثر من قسمٍ كلاميّ، من خلال الكشف عن سياقات كُلِّ كلمةٍ على حِدة. وعلى سبيل المثال، يُلزمُ الجرافيمان المُتتاليان (ي، س) في بداية الكلمة قسماً كلاميّاً هُو (الفعل المُضارع). وباستخدام الكشَّاف السِّياقيّ، نستطيعُ الكشفَ عن كلمة (يسوع) الَّتي خالَفَت القاعدة لتتبع القسمَ الكلاميّ (اسم العَلَم) على النَّحو المُوضَّح في (الشَّكل ٦)، ونستطيعُ الكشفَ عن سياقات كلمة (يسير) الَّتي تحتملُ أن تكونَ اسماً أو فعلاً، على النَّحو المُوضَّح في (الشَّكل ٧) [برمجيَّة Nooj Concordance 3.2.]



الشَّكل ٦: نموذج كلمات المُدوَّنة اللُّغويَّة مُفهرَسة آليّاً – برمجيَّة Nooj Concordance 3.2

الشَّكل ٧ : الكشَّاف السِّياقيّ لكلمات المُدوَّنة اللُّغَويَّة– برمجيَّة Nooj Concordance 3.2

وحتَّى نتبَيَّنَ جدوى المنهجيَّة المُقتَرَحة في التَّوسيم التَّركيبيّ، فإنَّنا نقفُ على الإجراءات الآتية، في ضوء إحصاء كلمات المُدوَّنة اللُّغَويَّة موضُوع الدِّراسة:

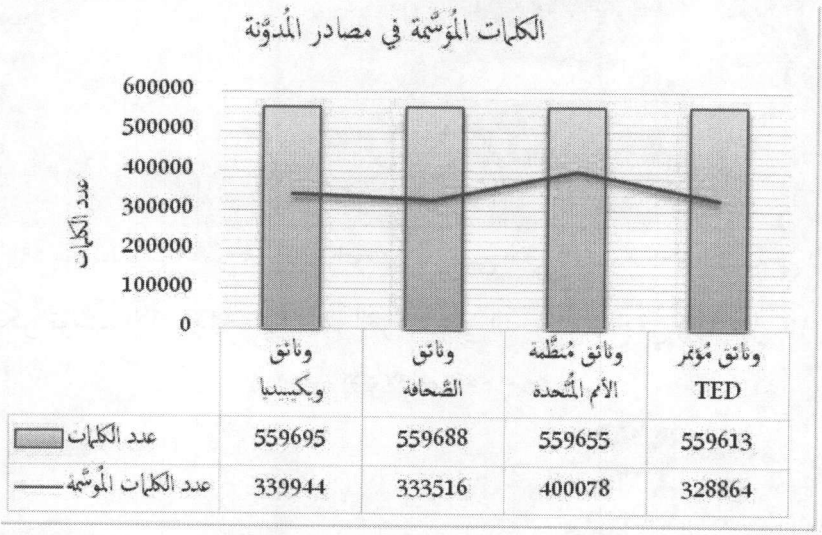
١. إذا قُمنا بتوظيف النَّحو العدديّ (الأحاديّ) – وفقَ منهجيَّة الدِّراسة – في التَّوسيم الآليّ لمجموعة الكلمات الَّتي ورَدَت في المُدوَّنة بتردُّد أكثر من ١٠٠ مرَّة، فإنَّنا نستطيعُ توسيم (١,٤٥٧,٦٩٢) كلمة، بنسبة ٦٥٪ من جُملة كلمات المُدوَّنة، علمًا بأنَّ عددَ هذه الكلمات بدون تكرار (٣,١٩٣) بنسبة ٢ ٪ من الكلمات الفريدة.

٢. إذا قُمنا بتوظيف الكشَّاف السِّياقيّ – وفقَ منهجيَّة الدِّراسة – في التَّوسيم الآليّ لمجموعة الكلمات الَّتي بدأت بسابقة (الـ) التَّعريفيَّة، فإنَّنا نستطيعُ توسيم (٧٠٢,٠٤٩) كلمة، بنسبة ٣١٪ من جُملة كلمات المُدوَّنة.

٣. إذا جَمَعنا بينَ الإجراءَين السَّابقَين، فإنَّنا نستطيعُ توسيم (١,٦٩٧,٩٢٧) كلمة، بنسبة ٧٦٪ من جُملة كلمات المُدوَّنة. وهي نسبةٌ قابلةٌ للزِّيادة – بصورةٍ كبيرة – حالَ توسيع النَّماذج المُوَسَّمة في المُدوَّنة اللُّغَويَّة الخام، سواءٌ على مُستوى النَّحو العدديّ أم على مُستوى الكشَّاف السِّياقيّ.

٤ . التَّطبيق.

بتطبيق المنهجيَّة على مِئتَي نموذجٍ في المُدوَّنة اللُّغَويَّة الخام، بواقع ١٠٠ نموذج لمُستوى النَّحو العدديّ و ١٠٠ نموذج لمُستوى الكشَّاف السِّياقيّ، أمكنَ توسيمُ (١,٤٠٢,٤٠٢) كلمة، بنسبة ٦٣٪ من جُملة كلمات المُدوَّنة. وكانت نتائجُ التَّوسيم على مُستوى مصادر المُدوَّنة على النَّحو المُوضَّح في (الشَّكل ٨).

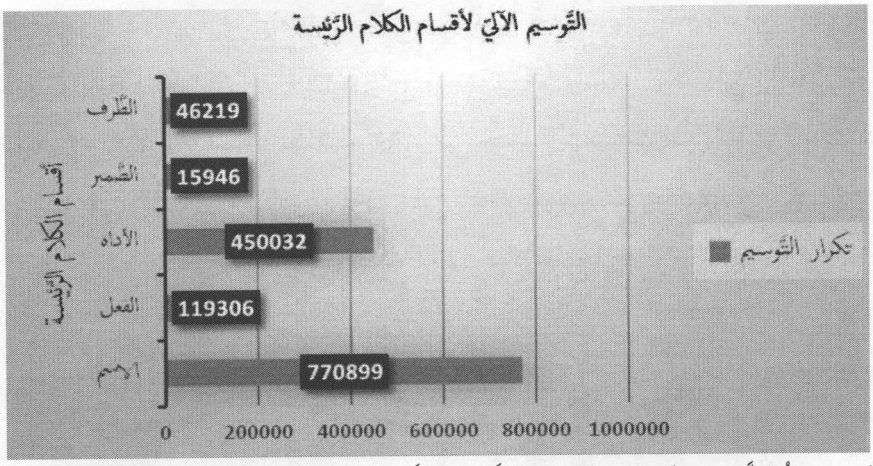


الشَّكل ٨: مُخَطَّط بيانيّ عموديّ خطِّيّ لنتائج التَّوسيم التَّركيبيّ الآليّ للكلمات في مصادر المُدوَّنة

تُشيرُ النَّتائجُ إلى أنَّ (وثائق مُنظَّمة الأمم المُتَّحدة) الَّتي تُعَبِّرُ عن [العربيَّة الرَّسميَّة] كانت أكثرَ اتِّساقاً وقابليَّة للتَّوسيم الآليّ؛ حيثُ أمكنَ توسيمُ ٧١,٥٪ منها. وتلتها على التَّرتيب: (وثائق ويكيبيديا) بنسبة ٦٠,٧ ٪، و (وثائق الصَّحافة) بنسبة ٥٩,٦ ٪، و(وثائق مُؤتمر TED) بنسبة ٥٨,٨٪. وتعكسُ هذه النِّسَبُ قُدرةَ الآلة على تمييز أقسام الكلام العربيّ من ناحية، وقُدرتَها على استكشاف أنماط الجُملة من ناحيةٍ أخرى. ولعلَّ السَّبَبَ في تقلُّص قُدرة الآلة على توسيم الوثائق المُعبِّرة عن العربيَّة العلميَّة يعودُ إلى غِناها بالمُصطلحات والمُسمَّيات غير الشَّائعة الَّتي تُخالفُ نسَقَ اللُّغة العربيَّة.

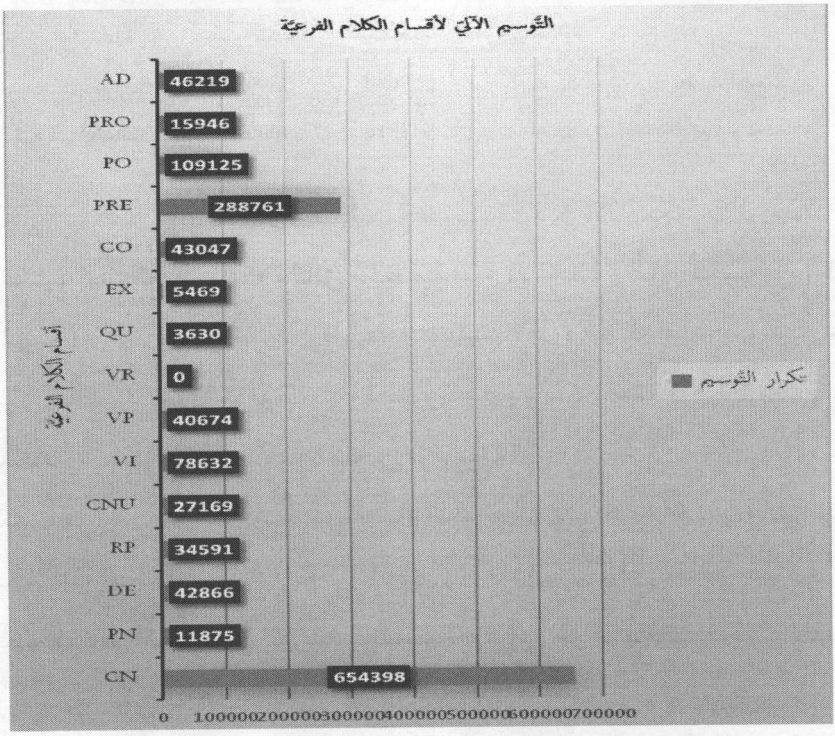
أمَّا على مُستوى أقسام الكلام، فقد أمكنَ توسيمُ الأقسام الخمسة الرَّئيسة على

النَّحو المُوضَّح في (الشَّكل ٩).



الشَّكل ٩ : مُخَطَّط بيانيّ عموديّ لنتائج التَّوسيم التَّركيبيّ الآليّ لأقسام الكلام الرَّئيسة في المُدوَّنة

وأمكنَ توسيمُ الأقسام الخمسةَ عشرَ الفرعيَّة على النَّحو المُوضَّح في (الشَّكل ١٠).



الشَّكل ١٠ : مُخَطَّط بيانيّ عموديّ لنتائج التَّوسيم التَّركيبيّ الآليّ لأقسام الكلام الفرعيَّة في المُدوَّنة

بتحليل نتائج التَّوسيم في الشَّكلَين (٩، ١٠)، نُلاحظُ – على مُستوى أقسام الكلام الرَّئيسة – توالي النِّسَب المئويَّة لتمييز (الاسم) ثُمَّ (الأداة) ثُمَّ (الفعل) ثُمَّ (الظَّرف) ثُمَّ (الضَّمير) على التَّرتيب. ونُلاحظُ – على مُستوى أقسام الكلام الفرعيَّة – ارتفاع نسبة توسيم (الاسم الشَّائع) من بين الأسماء، و(حرف الجرّ) من بين الأفعال، و(الفعل المُضارع) من بين الأفعال؛ في حين لم تتمكَّن الآلة من توسيم صيغة (فعل الأمر) في نُصُوص المُدوَّنة.

وتعكسُ هذه النَّتائجُ أمرَين:

١. فاعليَّة المنهجيَّة في تمييز أقسامِ الكلام (الرَّئيسة والفرعيَّة).

٢. دوران بعض أقسام الكلام في النُّصُوص العربيَّة المكتوبة، ونُدرة بعضها الآخَر.

٥. نموذج المُدوَّنة اللُّغويَّة.

يعرضُ الباحثُ فيما يأتي نموذجاً للمُدوَّنة اللُّغويَّة – موضوع الدِّراسة – في ثلاثة أشكالٍ؛ حيثُ يُمثِّلُ الشَّكلُ الأوَّلُ نموذجَ المُدوَّنة الخام قبلَ توسيمها، وتردُ الكلماتُ فيه بلا تظليل؛ ويُمثِّلُ الشَّكلُ الثَّاني نموذجَ المُدوَّنة اللُّغويَّة بعدَ توسيمها آليّاً؛ وتردُ الكلماتُ المُوَسَّمةُ فيه مُظلَّلةً دونَ غيرها؛ ويُمثِّلُ الشَّكلُ الثَّالثُ نموذجَ المُدوَّنة المُوسَّمة آليّاً بعدَ التَّدخُّل اليدويّ لإكمال التَّوسيم؛ وتردُ فيه جميعُ الكلمات مُظلَّلةً كونها صارت مُوسَّمةً تركيبيّاً في الصُّورة النِّهائيَّة.

٥,١ . المُدوَّنة اللُّغَويَّة الخام.

إن النتائج التي تمكن أن يحصل عليها الملك الآشوري، لا سيما في التخلص من النفوذ الميتاني والتمكن من اقتسام بلادهم، أن جعله يتوجه نحو توطيد أواصر علاقاته السياسية مع القوى السياسية الفاعلة، حيث أقدم على الزواج من ابنة الملك الكاشي الذي كان يفرض نفوذه على بابل. وقد حظيت مملكة آشور بملوك خلفوا آشور أوبلط وكانوا على مستوى المسؤولية وانتهجوا ذات الأسلوب الذي سار عليه ليثمر عن ذلك خلال القرن التاسع ق.م بلوغ مستوى الامبراطورية الآشورية بكل قوتها ونفوذها السياسي. من الملوك الآشوريين البارزين شلمنصر الأول الذي دام حكمه ١٢٦٦ -١٢٤٣ ق.م، وتطلع إلى توجيه العديد من الحملات العسكرية وعمل على استبدال العاصمة آشور بمدينة نمرود. أما العمل الأبرز فكان على يد الملك توكلتي نورتا ١٢٤٣ - ١٢٢١ ق.م، الذي تمكن من السيطرة على بلاد بابل، وتوسيع سلطانه في الجهات الشرقية والغربية. لكن بعد وفاة هذا الملك دخلت آشور في مرحلة الضعف السياسي، نتيجة لوصول ملوك ضعاف الشخصية، غير قادرين على إدارة مقاليد الحكم، واستمرت هذه الفترة حوالي مائة عام، حتى بلوغ الملك تجلات بلاسر الأول ١١١٦ - ١٠٩٠ ق.م إلى سدة الحكم، لتكون هذه الفترة مليئة بالإنجازات العسكرية الكبيرة، حيث تمكن من تحقيق الانتصارات المتوالية في الأصقاع البعيدة، في البحر الأسود وسواحل آسيا الصغرى والمدن الفينيقية على الساحل السوري.

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] الفاعلة، [AD]حيث [VI]أقدم [PRE]على [CN]الزواج [PRE]من ابنة [CN] الملك [CN]الكاشي [RP]الذي [VP]كان [VI]يفرض نفوذه [PRE] على بابل. [CO]و [PO]قد [VI]حظيت مملكة آشور بملوك خلفوا آشور أوبلط [CO]و [VP]كانوا [PRE]على مستوى [CN]المسؤولية [CO]و [VP] انتهجوا ذات [CN]الأسلوب [RP]الذي [VP]سار [PRE]عليه [PO] ل [VI]يثمر [PRE]عن [DE]ذلك [AD]خلال [CN] القرن [CNU] التاسع [AD]ق.م بلوغ مستوى [CN] الإمبراطورية [CN] الآشورية[PRE]ب [CN]كل قوتها ونفوذها [CN] السياسي. [PRE] من [CN]الملوك [CN]الآشوريين [CN]البارزين شلمنصر [CNU] الأول [RP]الذي [VP]دام حكمه [AD] [CNU]1243 ? [CNU]1266 ق.م، [CO]و [VI]تطلع [PRE]إلى توجيه [CN]العديد [PRE]من [CN] الحملات [CN]العسكرية وعمل [PRE]على [CN]استبدال [CN] العاصمة آشور [PRE]ب [CN]مدينة نمرود. [PO]أما [CN] العمل [CN]الأبرز [CO]ف [VP]كان [PRE]على يد [CN]الملك توكلتي نورتا [CNU]١٢٤٣[CNU] - ١٢٢١[AD] ق.م، [RP] الذي

[VP] تمكن [PRE]من [CN] السيطرة [PRE]على بلاد بابل، وتوسيع سلطانه [PRE]في [CN] الجهات [CN]الشرقية [CO]و [CN]الغربية. [PO] لكن [AD]بعد وفاة [DE]هذا [CN]الملك [VP]دخلت آشور [PRE]في مرحلة [CN]الضعف [CN]السياسي، نتيجة لوصول ملوك ضعاف [CN] الشخصية، [EX]غير قادرين [PRE]على إدارة مقاليد [CN]الحكم، [CO] و [VP]استمرت [DE]هذه [CN]الفترة [AD]حوالي [CNU]مائة عام، [CO] حتى بلوغ [CN]الملك تجلات بلاسر [CNU]الأول [CNU]١١١٦ – [CNU] ١٠٩٠ [AD] ق.م [PRE]إلى سدة [CN]الحكم، [PO]ل [VI] تكون [DE]هذه [CN]الفترة مليئة [PRE]ب [CN]الإنجازات [CN] العسكرية [CN]الكبيرة، [AD]حيث [VP]تمكن [PRE]من تحقيق [CN] الانتصارات [CN] المتوالية [PRE]في [CN]الأصقاع [CN] البعيدة، [PRE]في [CN]البحر [CN]الأسود وسواحل آسيا [CN] الصغرى [CO]و [CN]المدن [CN]الفينيقية [PRE]على [CN] الساحل [CN]السوري.

٥,٣ . المُدوَّنة اللُّغَويَّة المُوَسَّمة آليّاً-يدويّاً .

[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] الفاعلة، [AD]حيث [VI]أقدم [PRE]على [CN]الزواج [PRE]من ابنة [CN] الملك [CN]الكاشي [RP]الذي [VP]كان [VI]يفرض نفوذه [PRE] على بابل. [CO]و [PO]قد [VI]حظيت مملكة آشور بملوك خلفوا آشور أوبلط [CO]و [VP]كانوا [PRE]على مستوى [CN]المسؤولية [CO]و [VP] انتهجوا ذات [CN]الأسلوب [RP]الذي [VP]سار [PRE]عليه [PO] ل [VI]يثمر [PRE]عن [DE]ذلك [AD]خلال [CN] القرن [CNU] التاسع [AD]ق .م بلوغ مستوى [CN] الإمبراطورية [CN] الآشورية[PRE]ب [CN]كل قوتها ونفوذها [CN] السياسي . [PRE] من [CN]الملوك [CN]الآشوريين [CN]البارزين شلمنصر [CNU] الأول [RP]الذي [VP]دام حكمه [AD] 1243 [CNU] ? 1266 [CNU] ق .م، [CO]و [VI]تطلع [PRE]إلى توجيه [CN]العديد [PRE]من [CN] الحملات [CN]العسكرية وعمل [PRE]على [CN]استبدال [CN] العاصمة آشور [PRE]ب [CN]مدينة نمرود. [PO]أما [CN] العمل [CN]الأبرز [CO]ف [VP]كان [PRE]على يد [CN]الملك توكلتي نورتا [CNU]١٢٤٣[CNU] - ١٢٢١[AD] ق .م، [RP] الذي

[VP] تمكن [PRE]من [CN] السيطرة [PRE]على بلاد بابل، وتوسيع سلطانه [PRE]في [CN] الجهات [CN]الشرقية [CO]و [CN]الغربية. [PO] لكن [AD]بعد وفاة [DE]هذا [CN]الملك [VP]دخلت آشور [PRE]في مرحلة [CN]الضعف [CN]السياسي، نتيجة لوصول ملوك ضعاف [CN] الشخصية، [EX]غير قادرين [PRE]على إدارة مقاليد [CN]الحكم، [CO] و [VP]استمرت [DE]هذه [CN]الفترة [AD]حوالي [CNU]مائة عام، [CO] حتى بلوغ [CN]الملك تجلات بلاسر [CNU]الأول [CNU]١١١٦ – [CNU] ١٠٩٠ [AD] ق.م [PRE]إلى سدة [CN]الحكم، [PO]ل [VI] تكون [DE]هذه [CN]الفترة مليئة [PRE]ب [CN]الإنجازات [CN] العسكرية [CN]الكبيرة، [AD]حيث [VP]تمكن [PRE]من تحقيق [CN] الانتصارات [CN] المتوالية [PRE]في [CN]الأصقاع [CN] البعيدة، [PRE]في [CN]البحر [CN]الأسود وسواحل آسيا [CN] الصغرى [CO]و [CN]المدن [CN]الفينيقية [PRE]على [CN] الساحل [CN]السوري.

٦. نتائج الدِّراسة.

١. أبانَت الدِّراسةُ عن ماهيَّة المُدوَّنات اللُّغَويَّة وأثرِها في الصِّناعات اللُّغَويَّة، وعَرَضَت لمفهوم "التَّوسيم التَّركيبيّ" الَّذي يُعَدُّ أحَدَ إجراءَين لما يُعرَفُ بـ "العنونة التَّركيبيَّة"، وأبانَت كذلكَ عن ثلاثة مناهج للتَّحليل التَّركيبيّ الحاسُوبيّ للُّغة العربيَّة؛ حيثُ يعتمدُ المنهجُ الأوَّلُ على المعطيات اللُّغَويَّة المستمدَّة من قواعد النَّحو العَرَبيّ؛ ويعتمدُ المنهجُ الثَّاني على خوارزميَّة التَّحليل التَّركيبيّ الَّتي تُمثِّلُ صُورةً رياضيَّةً لقواعد النَّحو العَرَبيّ؛ ويقومُ المنهجُ الثَّالثُ على استخلاص قواعد النِّظام التَّركيبيّ من المدوَّنات اللُّغَويَّة العَربيَّة باعتبارها تمثيلاً لواقع اللُّغة.

٢ . أبانَت الدِّراسةُ عن إشكالات بناء مُدوَّنةٍ لُغَويَّةٍ مُوَسَّمةٍ تركيبيّاً للُّغة العربيَّة؛ وتمثَّلَت هذه الإشكالاتُ في : المُرُونة في نظام بناء الجُملة العَرَبيَّة، وطبيعة النِّظام الكتابيّ للُّغة العربيَّة، والاختلاف حَولَ أقسام الكلام العَرَبيّ .

٣ . اقترَحَت الدِّراسةُ منهجيَّةً لبناء مُدوَّنةٍ لُغَويَّةٍ مُوَسَّمةٍ تركيبيّاً للُّغة العربيَّة بطريقة نصف آليَّة عبرَ أربع خُطوات، بدءاً ببناء المُدوَّنة اللُّغويَّة الخام في ضوء أساليب التَّحليل الإحصائيّ، ومرُوراً بتعيين أقسام الكلام بما يتوافقُ مع الهدف المنشود ومنطق الآلة، ثُمَّ التَّوسيم التَّركيبيّ باستخدام تقنيات النَّحو العدديّ [الأحاديّ، والثُّنائيّ، والثُّلاثيّ، ...]، وانتهاءً بالتَّوسيم التَّركيبيّ باستخدام الكشَّاف السِّياقيّ . وتطبيقاً لهذه المنهجيَّة فقد صنَعَ الباحثُ مُدوَّنةً لُغويَّةً لتكونَ مادَّةً للدِّراسة، واستمدَّ مادَّتَها من أربعة مصادرَ أساسيَّة، هي : وثائق موسُوعة ويكيبيديا، ووثائق الصَّحافة العربيَّة، ووثائق مُنظَّمة الأُمم المُتَّحدة، ووثائق مُؤتمر (تيد TED)، بنسبة ٢٥٪ لكلّ مجموعة .

٤ . اقترَحَت الدِّراسةُ تقسيمَ الكلام العربيّ إلى خمسة أقسامٍ، هي : الاسم (ويتفرَّعُ عنه : الاسم الشَّائع، واسم العلم، واسم الإشارة، والاسم الموصول، والعدد)، والفعل (ويتفرَّعُ عنه : الفعل المُضارع، والفعل الماضي، وفعل الطَّلَب)، والأداة (ويتفرَّعُ عنها أدوات : الاستفهام، والاستثناء، والرَّبط، والجرّ، والأدوات الأخرى)، والضَّمير، والظَّرف .

٥ . أبانَت الدِّراسةُ عن إمكانيَّة توظيف النَّحو العدديّ في توسيم الكلمات الأكثر تردُّداً في المُدوَّنات، واقترَحت حلولاً لتوسيم الكلمات الَّتي تحتملُ أن تتبعَ أكثرَ من قسمٍ كلاميّ، كما أبانَت الدِّراسةُ عن جدوى توظيف الكشَّافات السِّياقيَّة في التَّوسيم التَّركيبيّ للكلمات الأقلّ تردُّداً، وأبانَت كذلك عن إمكانيَّة الإفادة من الكشَّاف السِّياقيّ في مُراجعة المُطابقة بينَ كلمات المُدوَّنة وأقسام الكلام الَّتي تُوسَمُ

بها إذا احتملَت الكلمةُ التَّوسيمَ بأكثر من قسمٍ كلاميّ.

٦. قامَ الباحثُ بتطبيق المنهجيَّة على مِئتَي نموذجٍ في المُدوَّنة اللُّغَويَّة الخام، بواقع ١٠٠ نموذج لمُستوى النَّحو العدديّ و ١٠٠ نموذج لمُستوى الكشَّاف السِّياقيّ. وخَلُصَ إلى توسيم ٦٣٪ من جُملة كلمات المُدوَّنة، تتنامى بزيادة عدد النَّماذج الخاضعة للمنهجيَّة.

٧. خلُصَت الدِّراسة إلى أنَّ (وثائق مُنظَّمة الأمم المُتَّحدة) الَّتي تُعَبِّرُ عن [العربيَّة الرَّسميَّة] كانت أكثرَ قابليَّة للتَّوسيم الآليّ؛ حيثُ أمكنَ توسيمُ ٧١٫٥ ٪ منها. وتلتها: (وثائق ويكيبيديا) بنسبة ٦٠٫٧ ٪، و (وثائق الصَّحافة) بنسبة ٥٩٫٦٪، و (وثائق مُؤتمر TED) بنسبة ٥٨٫٨٪ على التَّرتيب.

٨. أبانَ التَّطبيقُ عن قُدرة الآلة على توسيم أقسام الكلام الرَّئيسة: (الاسم) و(الأداة) و(الفعل) و(الظَّرف) و(الضَّمير) على التَّرتيب. وعلى مُستوى الأقسام الفرعيَّة، أبانَ التَّطبيقُ عن ارتفاع نسبة توسيم (الاسم الشَّائع) من بين الأسماء، و(حرف الجرّ) من بين الأفعال، و(الفعل المُضارع) من بين الأفعال.

٩. قامَ الباحثُ بعرض نموذجٍ للمُدوَّنة اللُّغَويَّة في ثلاثة أشكالٍ، هي: المُدوَّنة اللُّغَويَّة الخام، والمُدوَّنة اللُّغَويَّة المُوَسَّمة آليّاً (باستخدام منهجيَّة الدِّراسة)، والمُدوَّنة اللُّغَويَّة المُوَسَّمة آليّاً–يدويّاً (بالجمع بينَ المنهجيَّة والتَّوسيم اليدويّ).

٧. الخُلاصة.

يستغرقُ بناءُ المُدَوَّنات اللُّغَويَّة المصنوعة لأغراض التَّحليل التَّركيبيّ في العربيَّة وقتاً وجهداً كبيرَين؛ وهُو أمرٌ يُؤدِّي إلى زيادة تكلفة بناء هذا النَّوع من المُدَوَّنات. ووُقُوفاً على طبيعة اللُّغة العربيَّة وحاجتها إلى المُدَوَّنات اللُّغويَّة بهدف توظيفِها في تطبيقات حَوسَبة النَّحو العربيّ، لا سيَّما في بناء أدوات التَّحليل التَّركيبيّ وتطوير أدوات التَّدقيق الإملائيّ، فإنَّ هذه الدِّراسةَ تَسعى إلى تقديم منهجيَّة لبناء مُدَوَّنة

لُغَويَّة مُوَسَّمة تركيبيّاً للُّغة العربيَّة بطريقة نصف آليَّة. وينطلقُ الباحثُ في دراسته من تقنيات النَّحو العَدديّ N-Gram الَّتي تسمحُ بإحصاء وترتيب الكلمات والمُتلازمات وفقَ نَسَقٍ يُساعدُ على إيجاد القرائن الدَّالَّة على البنية التَّركيبيَّة، وتقنيات الكشَّاف السِّياقيّ الَّتي تُساعدُ على تعقُّب الوحدات الكتابيَّة الَّتي تُلزمُ الكلمةَ قسماً كلاميّاً مُعيَّناً، وتُساعدُ أيضاً في مُراجعة المُطابقة بينَ كلمات المُدوَّنة وأقسام الكلام، من خلال الكشف عن سياقات كُلِّ كلمةٍ على حِدة. ويسعى الباحثُ إلى ضبط منهجيَّته باستخدام القواعد القياسيَّة للنَّحو العَرَبيّ على مُستوى أقسام الكلام بما يضمنُ بناءَ المُدَوَّنة في صُورةٍ تُحقِّقُ الإفادةَ القُصوى من مادَّتها. وقد سعى الباحثُ في المنهجيَّة والتَّطبيق إلى التَّأليف بينَ منطق الآلة الَّذي يضمنُ إخضاعَها لأهداف البحث، ومنطق اللُّغة الَّذي يضبطُ قواعدَها وقوانينَها.

مراجع الدِّراسة

أوَّلاً: المراجع العربيَّة

١. أنيس، إبراهيم (١٩٧٢). مِن أسرار اللُّغة، مكتبة الأنجلو المِصْرِيَّة، القاهِرة، ط٤.

٢. حسَّان، تمَّام (١٩٧٩). اللُّغة العَرَبيَّة "مَعناها وَمَبناها"، الهَيئة المِصرِيَّة العامَّة للكتاب، القاهِرة، ط٢.

٣. السَّاقي، فاضل مُصطفى (١٩٧٧). أقسام الكلام العربيّ من حيثُ الشَّكل والوظيفة، مكتبة الخانجي، القاهرة.

٤. السَّعيد، المُعتزّ بالله (٢٠١١). مُدَوَّنة مُعجَم تاريخيّ للُّغة العربيَّة "مُعالَجة لُغَويَّة حاسوبيَّة"، أطروحة دكتوراه، كُلِّيَّة دار العُلُوم، جامعة القاهرة.

٥. نينغ، خوانغ تشانغ، و: تزي، لي جوان (٢٠١٦). علم الذَّخائر اللُّغَويَّة، ترجمة: هشام موسى المالكي، المركز القوميّ للتَّرجمة، القاهرة، ط١.

ثانياً: المراجع الأجنبيَّة

6. Abeillé, A. (2012). Treebanks: Building and Using Parsed Corpora. Springer Science & Business Media.

7. Aoun, J., Choueiri, L., Benmamoun, E. (2010). The Syntax of Arabic. Cambridge University Press.

8. Beeston, A. (2016). The Arabic Language Today. Routledge.

9. Habash, N. (2009). Arabic Natural Language Processing. Morgan & Claypool Publishers.

10. Kiss, T. (2015). Syntax - Theory and Analysis. Walter de Gruyter GmbH & Co KG.

11. Lehmann, M., Lugossy, R., Horv?th, J. (2016). Empirical Studies in English Applied Linguistics. Lingua Franca Csoport.

12. Levine, R. D. (2017). Syntactic Analysis. Cambridge University Press.

13. Lin, Yuri; Michel, Jean-Baptiste; Aiden, Erez Lieberman; Orwant, Jon; Brockman, Will and Petrov, Slav. (2012). Syntactic annotations for the Google Books Ngram Corpus. ACL '12 Proceedings of the ACL 2012 System Demonstrations.

14. Lu, Xiaofei. (2014). Computational Methods for Corpus Annotation and Analysis. Springer.

15. Ryding, K. C. (2014). Arabic: A Linguistic Introduction. Cambridge University Press.

16. Soudi, A., Bosch, A., (2007). Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer.

17. Zitouni, Imed. (2014). Natural Language Processing of Semitic Languages. Springer Science & Business.

ثالثاً: المواقع الإلكترونيَّة

*http://ar.wikipedia.org.

* http://opus.lingfil.uu.se/.

* http://www.laurenceanthony.net/.

* http://www.nooj4nlp.net.