نَحْوَبِنَاء بَنْك شَجَرِيٍّ نَحْوِيٍّ لِلُّغَة الْعَرَبِيَّة الْفُصْحَى الْمُعَاصِرَة(*)

أحمد روبي محمد
*كلية دار العلوم، جامعة الفيوم، مصر.
Ahmedruby757@yhoo.com

المستخلص - يعدُّ البنك الشجري النحوي موردًا هامًّا لبناء التطبيقات الإحصائية لمعالجة اللغات الطبيعية NLP كما أنَّه أداة للبحث في الظواهر اللغوية التي تصفُ الواقع اللُّغوي، ومنطلقًا للتحقُّق من فرضيات النَّظريَّات اللُّغوية، فضلًا عن رسم معالم واضحة لنظام الجملة في اللُّغة المَدروسة وتحديد خصائص علاقاتها التركيبية، مبيِّنًا وجوه الائتلاف والاختلاف، التَّغَيُّر أو عدم التَّغَيُّر في بنية الجملة.فالبنك الشجري النحوي العربي Syntactic Arabic Tree Bank مجموعة من التَّحليلات النحوية لجمل عربية مستقاة من موقع إسلام أون لاين تم توصيفها في إطار نظرية العلاقات التركيبية بالتبصُّر في معطيات النَّظام اللُّغوي الذي رسمه علماء اللغة على اختلاف مناهجهم وآرائهم لاستشفاف المعلومات الأساسية التي سيقدمها التطبيق المنشود.ويختلف البنك الشجري SATB عن البنوك الشجرية الأخرى من حيث المعلومات اللغوية وطريقة تمثيلها؛ وذلك لطبيعة الهدف المنشود من بناء محلِّلٍ نحويٍّ يقوم على استنباط الوظائف النحوية؛ سعيًا لتحقيق المهمة الأساسية للتَّحليل النَّحوي الآليِّ، وهي توفير المعطيات اللَّازمة للتحليل اللُّغوي الأعمق للفهم الأتوماتي للنصوص اللغوية، منطلقًا من تمثيلَ بِنية العبارة Phrase structure representation في التحليل العلائقي مع مراعاة تحديد العلاقة النحوية لكل وحدة Token في الجملة؛ وذلك لسببين رئيسيين: الأوّل:سهولة استخلاص السمات المميزة Features Extractionللوَحدات المكونة للجملة، والآخر:الحصول على سمات هجينة بين الاعتمادية وهيكلة السين البارية X-bar Schema من حيث تحديد العلاقات النَّحويَّة التبعية بين المركبات ورؤوسها ومكمَّلاتها.
وتقدّم هذه الورقة توصيفًا لمراحل البناء المتسلسلة، ثم مقارنتها بغيرها من البنوك الشجرية الأخرى، ثم مدى الاستفادة من استخدام البنك الشجري في بناء المحلل النحويّ.

الكلمات المفتاحية:
البنوك الشجرية – التحليل النحوي – العنونة –المحلل النحوي–معالجة اللغات الطبيعية–عنونة المدونات اللغوية.

1- مقدمة

لقد أصبحت هندسة اللُّغة العربيَّة من أهم مجالات تِقنية المعلومات والاتصالات، ولقد دعا ذلك إلى تعاظم الحاجة إلى المعالجة الآلية للغة العربية بمستوياتها المختلفة الصَّوتية والصَّرفية والنَّحويَّة والدِّلاليَّة، بسبب الفَيْض المعلوماتيِّ من جانب وإلحاقها بالثورة التِّكنولوجيَّة من جانب آخرَ، وتساعد الأنظمة الخبيرة في مجال الذكاءالاصطناعيِّ على تمثيل اللغة بكافة مستوياتها وفق قواعدَ للمعرفة، وقواعد تجريبيَّة على نحو يستطيع الحاسوب التعامُلَ معها.

تمثل معالجة النَّحو آليًّا –حاليًّا على الأقلِّ- صلب اللّسانيّات الحاسوبيَّة، وتشهد ساحتها أقصى درجات الامتزاج بين اللسانيات والحاسوبيات، بجانب ذلك فالمعالجة النحوية الآلية هي قنطرة الوصل التي تعبر خلالها مسارات الاقتراض المتبادل بين علوم اللُّغة وعلوم الحاسب.[5] ويلزم لمعالجة النَّحو آليًّا وضعقواعده في صياغة

(*)بحث تكميلي لنيل درجة الماجستير في علم اللغة بإشراف: الأستاذ الدكتور فريد عوض حيدر أستاذ علم اللغة بكلية دار العلوم-الفيوم، ونائب رئيس الجامعة لشؤون البيئة وتنمية المجتمع مشرفا رئيسيا، والأستاذ الدكتور محسن رشوان أستاذ الاتصالات والإلكترونيات بكلية الهندسة جامعة القاهرة مشرفا مشاركا، والدكتور خالد أبو غالية مدرس علم اللغة بكلية دار العلوم-الفيوم مشرفا مشاركا.

رسميَّة مكتملة وفقًا للنموذج النَّحويّ المتبع[5]، وانطلاقًا من هذا سعت المؤسّسات الأوروبية والأمريكية المعنية بحَوْسَبة اللُّغات إلى بناء مدوَّنات مُعَنْوَنة نحويًّاAnnotated Parsed corporaللاستفادة من أساليب التعلم الإحصائيMachine Learning في بناء نماذج إحصائيَّة لهذه المدوَّنات اللُّغويَّة.

وتعدُّ مجموعة التَّحليلات النَّحوِيَّة للجمل المعدّة يدويًّا أو ما تسمّى بالبنوك الشجريَّةTree banksمصدرًا مهمًّا لِبناء المحلِّلات الإحصائيَّة وتقييمها بشكل عامّ،كما تستخدم هذه التحليلات أو البنوك الشجرية الغنية بالتذييل التفصيلي في العديد من التَّطبيقات منها، تجزئة النصوص، والتَّشكيل الآلي، والعَنْوَنة بأقسام الكلام، وفكّ اللَّبس الصَّرفيِّ، وتحديد أبنية المركبات، والعَنْوَنة الدَّلالِيَّة، كما تُستخدم –أيضًا - في بناء المعايير الذهبية Gold Standards؛لتقييم دِقَّة الأنظمة المحوسبة وقياسها،وكذلك لإيجاد أوجه التشابه والاختلاف في نتائج التحليل مبينة الحالات التي تتفق عليها والتي تختلف فيها الأنظمة المحوسبة.

ويعود تاريخ أوَّل محاولة لبناء مُدوَّنة مُعَنْوَنة نحويًّا-للُّغة الإنجليزية- إلى ثلاثة عقود مضت، إذ كان هدفها آنذاك هو محاولة إيجاد منهج موسَّع للعَنْوَنة النَّحْويَّة يصلح لكافَّةالتطبيقات[6]،ثم تبنَّت المؤسسات المعنية بحَوْسَبة اللُّغات فكرةَ إنشاء مشروع البنوك الشَّجريَّة، وكان على رأسها مركز البيانات اللغوية(Linguistic Data Consortium) بجامعة بنسلفانيا الذي تولى مهمَّة إنشاء مشروع البنك الشجري للغة الإنجليزية (PATB) عام 2001م، ثم بدأ يتوسع ليشمل اللغة الصينية والعربية ولغات أخرى.

وهناك مجهودان كبيران في المدوّنات اللُّغويَّة العربيَّة المعُنْوَنة نحويًّا، وهما: بنك بنسلفانيا الشجري(PATB)، وبنك براغ الاعتمادي (PADB)،وهذان المجهودان قدما تمثيلات لغوية معقدة ومعلومات لغوية غنية التفاصيل؛ لتسمحَ هذه المعلومات بالبحث في التَّطبيقات العامَّة لمعالجة اللُّغة الطبيعية، إلّا أن كثيرًا من هذه المعلومات غير مستخدم حاليًّا في التَّطبيقات العربيَّة[15]، كما يتطلَّب هذا النوع من المدوَّنات وقتًا وجهدًا كبيرين؛ نظرًا لحاجته اليدوية في الترميز.

وقدم مؤخرًا مركز أنظمة التعلم الحاسوبي(Center For Computational Learning Systems) بجامعة كولومبيا بنكًا شجريًّا نحويًّا ثالثًا (CATiB) بغية تسريع عملية الترميزAnnotationوتجنب المعلومات التي لا فائدة منها، وكان غرضه التَّرجمة الآليَّة[15].

وفي هذا الإطار قدم الباحث بناءً نحويًّا آخر (SATB)، بهدف بناء محلل نحوي إحصائي يقوم على تحديد العلاقات النحوية لوحداتالجملة من خلال السَّمات الشَّجريَّة لأنظمة العلاقات التركيبة في الجملة؛ وذلك لما لأهميتها في التفهم الآلي للنصوص، وبهذا تختلف طبيعة بناء البنك الشجريِّ (SATB) عن بناء البنوك الشَّجريَّة الأخرى من حيث المعلومات اللغوية وطريقة تمثيلها، أمَّا المعلومات اللغوية فتتمثل في تجزئة النصوص، والعَنْوَنة بأقسام الكلام، والمحتوى النحوي من حيث المركبات والعلاقات النحوية، أما تمثيلها فكان باستخدام النَّظريَّة النَّحويَّة الوصفيَّة Descriptive theoryمستخدمة القوالب التخطيطيَّة ل السين البارية-X [11Bar]لعرض أبنية الجمل من خلال تمثيل شجرة بنية العبارة Phrase Structure Tree Representation .

وفي هذه الورقة،سأقدم توصيفًا لمراحل بناء البنك الشجريِّ النحويِّ، مع تزويدها بالمعلومات الإحصائية،والأدوات الحاسوبيَّة المساعدة، ثم عرض مقارنة بينه وبين البنوك الشجرية الأخرى،واستخدامه، والأعمال المستقبلية والخاتمة.

2- اختيار المدونة اللغوية

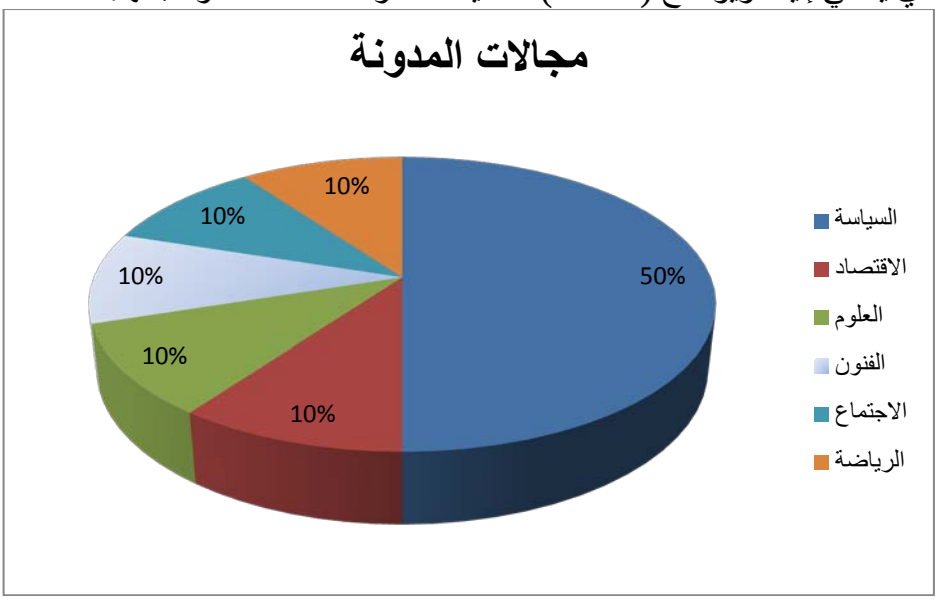
المدوَّنة اللغويَّة مجموعة من النصوص يمكن التعامل معها آليًّا، والتحكُّم في بياناتها ومدخلاتها بالإضافة أو الحذف أو التعديل من خلال قواعد البيانات التي تتعامل مع هذه النصوص[1]، كما تصف الواقع اللغوي اعتمادًا على مجموعة من النصوص التي تمثل ذلك الواقع، أو تأكيد فرضيات قائمة حول لغة معينة[1].

واختيار المدوَّنة اللُّغويَّة المعنيَّة بالتَّحلِيل يعتمد على الهدف المنشود منها،كمايرتبط بالزمان والمكان والمستوى اللغوي المطلوب، وفي معظم الحالات تتكون هذه النُّصوص من الصحف أو الجرائد أو المجلّات المعاصرة، و في إطار تزايُد أهمية الصحافة الإلِكْتُرُونيَّة وقع الاختيار على صحيفة إسلام أون لاين– للمدَّة الزمنية المحددة من 2002 إلى 2010 - مصدرًا للمدوَّنة اللُّغوية المعنيةبالتحليل(*)لعدَّة أسباب منها:

- إسلام أون لاين صحيفة إلكترونية، كما أن الصحافة الإلِكْتُرُونيَّة أصبحت من أهم الوسائل الإعلامية المعاصرة، ويتم نشرها بين كافة الشرائح المجتمعية بصورة متسارعة.
- التزامها بالكتابة العربية الفصحى غير الهجينة بالعاميَّة.
- معظم المحررين العاملين بها من اللُّغوِيِّين، وهذا يعكس نقاء المفردات والأساليب اللغوية وخلوها من الأخطاء الإملائية والنحوية.
- كثرة عدد زوّار موقع الصحيفة عن المواقع الأخرى.

ويخضع اختيار النُّصوص – وفقًا لبناء المدوَّنة اللُّغويَّة – على نظريَّة العينات الإحصائيَّة Statistical Sampling Theory، ومن خلالها يقوم صنّاع المدوَّنات اللُّغويَّة باختيار عينة من النُّصوص التي تتفقوأهدافهم البحثية سواء أكانت عينة عشوائية(Probabilistic Samples(Random، أم عينة غير عشوائية-Non Probabilistic Samples (Non-Random)[1].

وتم تصنيف هذه النُّصوص حسب المجالات التي تنتمي إليها، وبلغ عدد كلماتها 100,000كلمة،وبلغ عدد الواحدات000, 123وحدة، وعدد المقالات 98 مقالا، وتم تحريرها في ملفات نصية،كما تم تسمية كل ملف باسم المجال الذي ينتمي إليه، ويوضح (الشكل1) تصنيف المدونة المستخدمة وحجمها.



الشكل رقم (1)
تصنيف المدونة المستخدمة إلى مجالات.

(*)هذه النصوص مأخوذة من مدونة الشركة الهندسية لتطوير النظم الرقمية RDI، والتي يبلغ عدد كلماتها مليون كلمة.

3- تجزئة النصوصTokenization

تعد عملية تجزئة النصوص- آليًّا- مرحلة أساسية قبل المعالجة الآلية للنُّصوص اللُّغويَّة، وفي ظلِّ التطوَّر السريع لمعالجة اللُّغات الطبيعيَّة أصبحت عمليَّة تجزئة النصوص خطوة حاسمة في تنقيب النصوص واستخلاص المعلومات. وتقول الحكمة الشائعة في معالجة اللُّغة الطبيعية "إن تجزئة النَّصِّالعربيِّ إلى كلمات من خلال التجريد وتقليص الاحتمالات الهجائيةOrthographic Normalizationمفيد للعديد من التطبيقات مثل نمذجة اللغة واسترجاع المعلومات والترجمة الآلية الإحصائية"[15].

تعمل معظم تطبيقات معالجة اللغات الطبيعية مثل مُعَنْوَنات أقسام الكلام Part-Of-Speech taggers والمحلّلات النحوية Syntactic ParsersوالتجذيعStemming على النصّ المقسّم إلى أجزاء/عناصر، وتشمل هذه الأجزاء الكلمات والأرقام وعلامات الترقيم والرموز، وغيرها من الواحدات المكوّنة للنصّ.

تتوقف دِقَّةُ هذه التَّطبيقات وفقًا لدِقَّة عملية تجزئة النصوص؛لأنها مرحلة أوليَّة أو أساسيَّة في التَّحليل اللُّغويِّ تقوم عليها بقية المراحل التحليلية، فهي مطلب أساسيٌّ للتَّحليل النَّحويِّ الذي ترسي دعائمه تلك الوحدات النحوية، ويوضح (الشكل2) مراحل البناء الشَّجريِّ.

الشكل رقم (2)
مراحل بناء البنك الشجري

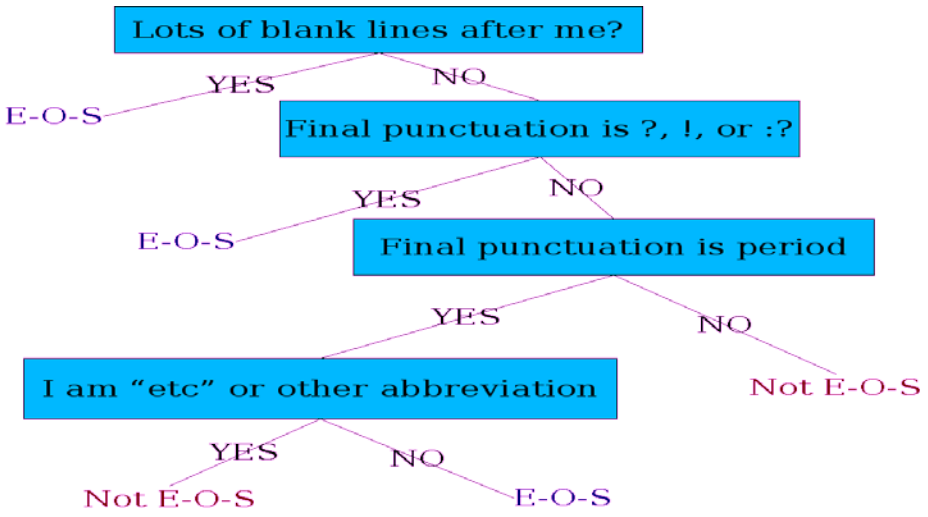
تجزئة النصوصTokenizationهي عملية تقسيم تُجرى على النص لتقسيمه إلى كلمات أو وحدات أو جمل،حتى تتمكن آليات معالجة اللغات الطبيعية من معالجة النصوص حسب ما تهدف إليه طبيعة التطبيقات المنشودة[8]، وتتمُّ عملية تجزئة النُّصوص –بالنسبة للُّغة العربيَّة- على ثلاثة مستويات:

1) التجزئة على مستوى الجملة

ينجز التقسيم -في تحديد أبعاد الجملة في المدوَّنة اللُّغوية موضوع الدراسة- حسب الإسناد والتركيب التامّ المفيد، وما بين الجمل من علاقات الربط بواسطة أدوات الاستئناف والعطف[3]، فكانت علامات الترقيم خير سبيل لتحديد هذه الأبعاد الجملية -تحديدًا شكليًّا- في النَّصّ.

وتتمّ عملية تجزئة الجمل أو تحديدها -آليًّا- في النّصّ من خلال وجود فواصل الأسطر وعلامات الترقيم، وتستطيع آلية تجزئة النصوص Tokenizerأن تحدد حدود الجملة من خلال علامة الترقيم النقطة(.) التي توضع في نهاية الجملة، وكذلك بعض الجمل التَّكميلية مثل الجملة الاستفهاميَّة التي تنتهي بعلامة استفهام(؟)، والجملة التعجبية التي تنتهي بعلامة تعجب(!)، إلّا أن هذه الآلية التي تعتمد على هذه المدخلات تواجه عدَّة مشكلات في

تحديد حدود الجمل بسبب تعدد وظائف بعض علامات الترقيم مثل النقطةالتي توضع في نهاية الجملة، وتوضع أيضا بين الاختصارات مثل أ.د، ص.د.ب وغيرها، وكذلك الفاصلة (,) التي تعد ملحمًا مميزًا؛ للفصل بين الوحدات أوالمكونات، توضع أيضا في حال الأرقام العشرية مثل 11,4، ومع ذلك يتغلَّب اللسانيون الحاسوبيون على هذه المشكلات عن طريق بناء مصنفات ثنائيةbinary classifiersمثل أشجار القرار Decision Trees، التي تعتمد على تحديد خواص العلامات وإعادة ترتيبها في صورة شجرية متدرِّجة، حيث يشكل التنسيق بين هذه الخواص تصورًا يؤدي إلى تحديد الفئات من حيث التجزئة أو عدم التجزئة، ويبيّن (الشكل 3) الخواص التي يعتمد عليها مصنِّف أشجار القرار في تحديد الكلمة النهائية للجملة.



الشكل رقم (3)
تحديد الكلمة النهائية في الجملة باستخدام أشجار القرار.

استعانت الدراسة بآليَّة تجزئة النصوصTokenizerالمدرجة في المحلل النحويStanford parserالتابع لفريق معالجة اللغات الطبيعية بجامعة ستانفورد، إلّا أنَّه لم يستخدم هذه التقنيات في إزالة لبس حدود الجملة، ونتج عن ذلك أخطاء في تحديد حدود الجمل، قمت بمعالجتها يدويًّا، ومن هذه الأخطاء:

-التجزئة في الحالات الآتية مثل: د.كلثم، 2,4 سم، باعتبار أن النقطة دليلًا على نهاية الجملة، والفاصلة على أنها بين مركبين.

-التجزئة في حالة تعدد النقاط الدالة على "الكلام المحذوف" ...، أي بجعل كل نقطة في سطر جديد.

- لا تميّز بين حد السطر والذي يليه سوى بتلك العلامات (.) أو (؟)، مما يؤدّي إلى تداخل الأسطر مع بعضها البعض.

2) التجزئة على مستوى الوحدات/العناصر الرئيسية

العنصر اللغوي Token هو أصغر وحدة نحوية، يمكن أن تكون كلمة أو جزءًا من الكلمة، أو تعبيرًا اصطلاحيًّا، أو مركبًا، أو علامة ترقيم،ومادامت العناصر اللغوية الرئيسية هي الجزء الملموس من التحليل فيمكن أن نطلق عليها أيضا"وحدات التحليل النحوي"[2]تلك الوحدات الرئيسية التي تعتبر عنصرًا أساسيًّا في النص اللغوي، فالوحدة الرئيسة هي البناء اللغوي المتكامل سواء أكانت كلمة أو علامة أو رقمًا.[8]

فالتجزئة على مستوى الوحدات أو العناصر الرئيسية تشمل ثلاثة متسويات:

أ- الكلمة

الكلمة هي أصغر وحدة مستقلة في النَّصِّ، ولعل أشهر من عرف الكلمة من علماء اللُّغة المحدثين هو العالم الأمريكي"بلومفيلد" Bloomfield، الذي قال " الكلمة أصغر وحدة حرة"، ومعنى هذا أن الكلمة عنده هي أصغر وحدة لغوية يمكن النطق بها معزولة[2]، كما يمكن استعمالها لتركيب جملة أو كلام، وينبغي أن تتكون من مورفيم حر Free Morpheme على الأقل. وتعتبر عنصرًا تحليليًّا بسيطًا على المستوى النَّحويِّ في بعض الأنحاء رغم تركيبها مع مورفيمات أخرى.

ب- المركب غير الكلامي

هو انضمام كلمة إلى كلمة فأكثر، وتكون بحكم المفرد نحويا ودلاليا مثل: عبد_الله، أبو_عيد، إسلام_أون_ لاين، الصهيو_أمريكي، الجيو_ إستراتيجية.

ت- الرمز أو العلامة

يشمل جميع الرموز المستخدمة في النص العربي، مثل علامات الترقيم والأرقام، وغيرها من الرموز.

هناك فرق بين تجزئة النصوص للُّغات ذي النظام الألفبائي واللغات ذي النظام الفكري مثل الصينية، فعادة في اللُّغات ذي النظام الألفبائي يتم الفصل من خلال حدود الكلمات أي الفراغات أو المساحات البيضاء، أما في اللُّغات ذي النظام الفكري فهي لا تحتوي على معلومات حول حدود الكلمة؛ لذلك التجزئة أصعب بكثير من النظام الألفبائي.

تعتمد آليَّة تقطيع النصوص Tokenizer على الفراغات البيضاء وعلامات الترقيم والأرقام كعلامات مميَّزة لتجزئة هذه الوحدات الرئيسية[9]، أما الوحدات التي تحتوي على كلمتين أو أكثر مثل عبدالله، قمت بوضع شرطة بدلا من الفراغات البيضاء لتصبح : عبد_الله.

3) التجزئة على مستوى الوحدات/العناصر الفرعية

يمكن أن نعرّف العنصر اللُّغوي أيضا بأنه" بناء لُّغوي يحدده مستوى التحليل" ، إذ نجد أن العنصر اللُّغوي الرئيسي قد يكون مكونًا من مورفيم / عنصر فرعي واحد أو أكثر من مورفيم [2]، فعلى سبيل المثال، حيث يمكن للكلمة المفردة (العنصر الرئيسي) أن تشمل على ما يصل إلى أربع وحدات فرعية سواء سوابق أو لواحق[8].

تتوقف حدود عملية تجزئة العناصر الرئيسية إلى عناصر فرعية إلى طبيعة الغرض من البحث، أي ما العناصر الفرعية المراد تجزئتها من العناصر الرئيسية؟وللإجابة عن هذا السؤال، نبيّن -أولا- أنواع المورفيمات اللصقية Concatenative Morphemes في اللغة العربية،فهناك ثلاثة أنواع من المورفيمات المتسلسلة وهي : الجذع (Stem) واللواصق(affixes) والزوائد (Clitics).

أ- الجذع Stem

أساس الكلمة بعد حذف الزوائد واللواصق منها، ووجوده ضروري لكل كلمة، ومن أمثلته: الجذع (كتب) الذي تكون عنه التركيب في(وسيكتبونها) والجذع (مكتب) في صيغة الجمع(المكتبات).

ب- اللَّواصق Affixes

هي مورفيمات تتعلق بجذع الكلمة، وهناك نوعان من اللَّواصق:
1) السَّوابق(Prefixes): هي مورفيم يسبق الجذع في أوله ومن أمثلته: نون في الفعل المضارع في "نفعل-نعمل-نشكر".
2) اللَّواحق(Suffixes): هي مورفيم يلحق الجذع في آخره ومن أمثلته: الواو والنون في جمع المذكر السالم في" المسلمون-العاملون".

ت- الزَّوائد Clitics

هي مورفيمات نحوية تكون مقيدة بكلمات أخرى، و تتعلق بجذع الكلمة بعداللواصق،وهناك نوعان من الزَّوائد:
1) الزَّوائد في بداية الكلمة (Proclitics)فهي تشبه اللواصق، ولكنها تختلف اختلافًا واضحًا عن اللواصق التي تمثل جزءًا من الكلمة صوتيًّا وبنيويًّا، ومن أمثلتها:حروف العطف، وحروف الجر، والنداء.
2) الزَّوائد في نهاية الكلمة(Enclitics)وهي التي تعقب الكلمة، مثل الضمائر المتصلة.

قد يكون القرار مربكًا أحيانًا في جعل المورفيم لاصقة صرفية أو زائدة نحوية، ومع ذلك نقول –عمومًا- أن اللواصق تحمل ملامح صرفية نحوية مثل(الزمن–الشخص–الجنس–العدد)، بينما الزوائد تخدم الوظائف النحوية مثل(النفي–التعريف– العطف أو الجر).

وطبيعة ما نصبو إليه يجعلنا نقف أمام الوحدات النحوية (الزائدة بنوعيها) باعتبارها عنصرًا مستقلًّا للتحليل النحوي، فهي عنصر تحليلي ذات علاقات نحوية نظمية بغيرها، وسنبيّن منهجنا في التجزئة من خلال الأمثلة التالية:

- للمدرسة ----> ل+المدرسة
- وسيكتبونهم ------>و+سيكتبون+هم
- مستشفاهم ------> مستشفى+هم
- سمائه -------> سماء+ه
- مكتبتنا ------> مكتبة+نا

واستعانت الدراسةبآلية MADA+TOKANلتجزئة النصوص وفقًالمنهج ATBفي التجزئة كخطوة أوليَّة للتحليل، ثم قمت بتعديل الخارج يدويًّا؛ ليتناسب مع المنهج المقترح.

4- العَنْوَنة بأقسام الكلام POS Tagging

تعدُّ عنونة الأقسام الكلاميَّة عملية أساسية من عملية التحليل اللساني، حيث تهدف إلى استخلاص الانواع الكلامية، وهي تلك السمات التي تحمل الخواص النحوية البدائية المميزة لكل كلمة منفردة بمعزل عن سياقها الإعرابي في النَّصِّ محل الدراسة[7].

ومن البديهي أن الأقسام الكلامية لكلمات نص ما هي من أهم المدخلات الابتدائية لأي عملية تحليل نحوي لهذا النص، فلا نستطيع مثلا على الإطلاق أن نعرف أن كلمة(تحليل) في الجملة السابقة مضاف إليه دون أن نعرف أولا أنها اسم[7]؛ ولذلك تقوم المحللات النَّحويَّة بشكل عامٍّ على التحليل الصرفي للكلمات.

فالعنونة بأقسام الكلام هي مهمَّة تعيين السمات الصرفية والنحوية لكل وحدة في النص[10] عن طريق إلحاق كل مفردة بالنص برمز أو عدة رموز تشير إلى سماتها الصرفية والنحوية.ويمكن أن تكون سمات أقسام الكلام للغة العربية كبيرة جدًّابسبب غنى اللغة العربية صرفيًّا، ومع ذلك يفضل كثيرمن الباحثين العاملين في معالجة اللغة العربية العمل على مجموعات أصغر حجمًا؛ حتى يمكن التنبؤ بها بدقَّة عن طريق أساليب التعلم الإحصائي.

تتمُّ عملية ترميز الخصائص الصرفية للكلمة المحللة باستخدام مجموعة العناوين للخصائص الصرفية للكلمة المحللة POS Tag set ، ويمكن أن تصل السمات الصرفية للغة العربية نظريًّا إلى 330 ألف سمة، وقد اعتمد باكولتر على آلاف السمات (للنص غير المقطع)، بينما اعتمد على مجموعة من السمات(للنص المقطع) تصل إلى حوالي 500 سمة، وهذا ما استخدمه بنك بنسلفانيا الشجريّ.

اعتمدت الدراسة على مجموعة العناوين الصرفية POS Tag set - وبلغ عدد سمات هذه المجموعة 62 سمة-التي اقترحها الباحث محمد عطية في رسالته للدكتوراه لتطوير آلية التشكيل، لما أراه مناسبًا لعملية التحليل النحوي، كماتتميز بأنها ليست مستعارة من اللغات الأخرى، بل صُمّمت خصيصًا للغة العربية.وتم استخدام آلية ©Arab Tagger[*]لعنونة الأقسام الكلاميةللمدونة اللغوية المعنية بالتحليل، ثم مراجعة الخرْج يدويًّا، ومن أمثلة التحليل لمعنون الأقسام الكلامية كما في (الشكل 4):

{(وَ); (NullPrefix) (Conj) (NullSuffix) }

{(جَاءَ); (NullPrefix) (Active) (Verb) (Past) (NullSuffix) }

{(اَلْعُدْوَانُ); (Definit) (Noun) (NullSuffix) }

{(اَلْإِسْرَائِيلِيُّ); (Definit) (Noun) (NoSARF) (RelAdj) }

{(اَلْأَخِيرُ); (Definit) (Noun) (ExaggAdj) (NullSuffix) }

{(عَلَى); (NullPrefix) (Prepos) (NullSuffix) }

{(لُبْنَانَ); ((NullPrefix) (Noun) (NoSARF) (Femin) (Single) (NullSuffix) }

{(عَلَى); (NullPrefix) (Prepos) (NullSuffix) }

{(قَاعِدَةِ); (NullPrefix) (Noun) (SubjNoun) (Femin) (Single) }

{(اِسْتِمْرَارِ); (NullPrefix) (Noun) (NounInfinit) (NullSuffix) }

{(حِرَاكِ); (NullPrefix) (Noun) (NullSuffix) }

{(اَلْجَبْهَةِ); (Definit) (Noun) (Femin) (Single) }

{(اَلْمَفْتُوحَةِ); (Definit) (Noun) (ObjNoun) (Femin) (Single) }

{(فِي); (NullPrefix) (Prepos) (NullSuffix) }

{(اَلْجَنُوبِ); (Definit) (Noun) (ExaggAdj) (NullSuffix) }.

الشكل رقم (4): تحليل POS باستخدام معنون أقسام الكلام ©Arab Tagger

(*)الاسم التجاري للمعنون العربي للأقسام الكلامية في الشركة الهندسية لتطوير النظم الرقمية**RDI.**

5- التحليل النحوي Syntactic Parsing

وتظلُّ المهمَّة الرئيسية للتَّحليل النحوي الآلي، هي توفير المعطيات اللَّازمة للتَّحليل اللُّغوي الأعمق، ألا وهو الفهم الأتوماتي للنصوص اللغوية[5]، وتقوم هذه المرحلة على عدَّة خطوات أساسية في التَّحليل، وهي التقويس النحوي أو التمثيل النحوي، والنظرية النحوية، والمحتوى النحوي.

1) التَّقوِيس النحوي Syntactic bracketing

هو نموذج رياضي ينظم الكلمات بطريقة متماسكة، بحيث يظهر العلاقات بين الكلمات في الجملة[13]، وهو ما يشبه صيغة باكوس نور BNF، وقداعتمده تشومسكي في إعادة كتابة القواعد النحوية الشكلية من خلال النحو المتحرر من السياق[12].

النظام الرياضي المستخدم في العنونة النحوية للمدوَّنةهو النحو المتحرر من السياق اعتمادًا على مخطط السين الباريةX-barمن حيث الرأس والمكمل والوصف، مع عدم الاعتماد على الفئات الفارغة، وهذا ما يعتمده محلل ستانفورد النحوي، ومحلل بايكلBikel's parser مع الاعتماد على الفئات الفارغة.

واستعانت الدراسةبآلية Stanford Parserكأدة مساعدةفي تحليل نصوص المدونة اللغوية، ثم فحص الخرْجيدويًّاباستخدام التعبيرات النمطيةRegular Expression، ونجد في (الشكل 5) مثالًا لجملة تمّ تحليلها باستخدام Stanford Parser.

```
(ROOT
  (S و
    (PP في
      (NP
        (NP اكتوبر)
        (NP 2006)))
    (VP وقع
      (NP الاختيار)
      (PP على
        (NP عهدية
          (NP احمد)))
      (PP ك
        (NP رسمية متحدثة))
      (PP عن
        (NP
          (NP الانتخابات)
          (ADJP النيبابية
            و
            البلدية)
          (ADJP البحرينية))))
      .))
```

الشكل رقم (5)
جملة محلل باستخدام Stanford Parser

ويبيّن (الشكل 6) الجملة السابقة بعد التعديل اليدويّ، إذ تمّ في ضوء تحديد المحمول وبنية العوامل Predicate-argument structure.

```
(S   وَ
  (PP   فِـي
    (NP   أُكْـتُـوبَـرَ
      (NP   2006)))
  (VP   وَقَـعَ
    (NP   اَلِاخْـتِـيَـارُ)
    (PP   عَـلَـى
      (NP   أَحْـمَـدَ   عَـهْـدِيَّـةِ))
    (PP   كَ
      (NP   رَسْـمِـيَّـةٍ   مُـتَـحَـدِّثَـةٍ))
    (PP   عَنْ
      (NP   اَلْـبَـحْـرَيْـنِـيَّـةِ   الْـبَـلَـدِيَّـةِ   وَ   اَلْـنِـيَـابِـيَّـةِ   اَلِانْـتِـخَـابَـات)))
  .)
```

الشكل رقم (6)
نتائج Stanford parser مجردة من POS بعد التعديل اليدوي.

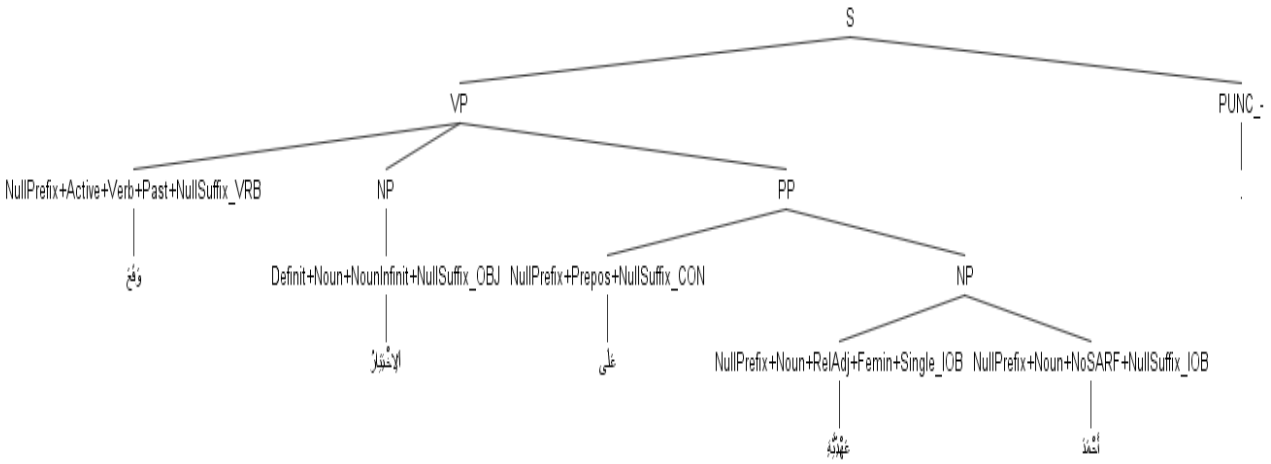
2) الرموز الوظيفية Functional Tags

ولا يستغنى التَّحليل اللُّغوي عن البيانات التصنيفيَّة والوظيفيَّة؛ لأنَّ المكونات مع نفس خصائصها التصنيفية يمكن أن تقع في علائق وظيفية أخرى، والعلائق الوظيفية نفسها يمكن أن تنطبق على مكونات أخرى مع اختلاف خصائصها التصنيفية[4]، وانطلاقًا من الهدف المنشود جُعلت هذه الرموز الوظيفية لوحدات الجملة وليس مركباتها، وهذه الرموز مصنفة كما في (الجدول1):

الجدول رقم (1)
الرموز الوظيفية المقترحة

الفئة	المصطلح العربي	الرمز	المصطلح الإنجليزي
الوظائف النحوية Grammatical functions	-المسند	SBJ	Predicate
	-المسند إليه	PRD	Subject
	-الموضوع	TPC	Topic
	-المفعولية	OBJ	Object
	-فعل	VRB	Verb
	-تمييز	TMZ	Tamyiyz
	-وصف	MOD	Modifier
	-إضافة	IDF	IDafa
	-مفعول غير مباشر	IOB	Indirect object
الأدوار الدلالية semantic roles	-الزمان والمكان	TMB	Locative Temporal
	-السببية	PRB	Purposive
	-الحركة	MOV	movement
	-الاتجاه	DIR	direction
	-الإيضاح	IDH	IDah
التلازم التركيبي Phrase depended	-المتلازم	DPN	Depending
	-فعل_متلازم	VRD	Verb_depend
	-ربط	CON	Conjunction
	-	-	For punc and sym

وباستخدام المراحل البنائية الثلاث، يظهر التحليل كما (بالشكل7)، مستخدمًا أداة عرض التحليل الشجريstanford-tregex.



الشكل رقم (7)
التحليل الشجري باستخدام أداة العرض الشجري.

6- أوجه التشابه والاختلاف بينSABT والبنوك الشجرية الأخرى

يحتوي (الجدول 2) على أوجه التشابه والاختلاف بين البنوك الشجرية العربية.

الجدول رقم (2)
أوجه التشابه والاختلاف بين البنوك الشجرية العربية

البنك النحوي	المدونة اللغوية	التمثيل النحوي	سمات أقسام الكلام	عددالوظائف النحوية المعنونة للكلمات	عدد الوظائف النحوية المعنونة للمركبات	الفئات الفارغة
ATB	صحيفة النهار	شجرة بنية العبارة	500 سمة	×	20	✓
PADT	صحيفة وكالة فرانس برس+وكالة أنباء الحياة	شجرة بنية التبعية	أكثر من 500 سمة	20	×	×
CATiB	صحيفة وكالة فرانس برس+وكالة أنباء الحياة+الحياة	شجرة بنية التبعية	6 سمات	7	×	×
SATB	صحيفة إسلام أون لاين	شجرة بنية العبارة	62 سمة	18	×	×

7- ترميز المدونة باستخدام لغة التوصيف XML

تم ترميز البنك النحوي بلغة الترميز القابلة للامتداد Extensible Markup LanguageباستخدامDTDالخاص بالبنك الشجريّ الألمانيTIGER،حيث يحولنظام التقويسbracketingإلى تنسيقNegra export format،وهذا التنسيق يجعل البيانات على شكل مصفوفة سهلة القراءة؛ممَّا يساعدنا على سهولة استخلاص السماتFeatures Extractionمن البناء الشجري.

8- استخدام البنك الشجريSATB

وتحقيقًا للغرض من البناء، قمت ببناء محلل نحوي قائم على الوظائف النحوية Functional Tagsباستخدام أساليب التعلم الإحصائي الموجهةSupervised learningمستخدمًا مصنّف الحقول العشوائية المشروطة(Conditional random fields (CRFs، وذلك في ضوء استخراج سمات البناء الشجريمن خلال البناء الموصَّف بلغةXML، وكانت السماتFeatures هي الأقسام الكلامية والفئة النحوية للكلمة من حيث كونها مركبًا اسميا أو فعليا...، ووضْع الكلمة في البناء الشجري من حيث التسلسل العائلي(الأب-الأخوات)، نوع الكلمة في هيكلة السين البارية من حيث الرأس والمكمل والوصف، أما المصنفات Classes هي الرموز الوظيفية.

وبعد تدريبمصنف الحقول العشوائية المشروطة++CRF على المدوَّنة الموصَّفة نحويًّا(البنك الشجري) كانت نتيجة الإحصائيات كالآتي، كما هو مبيّن في (الجدول 3).

الجدول رقم (3)
إحصائيات عن البنك الشجري والتدريب الآلي

Words	Tokens	Sentence	Training	Testing	Class error rate
100,000	123K	5,045	%90	%10	%11,2

9- الخاتمة

عرضنا في هذا البحث، موردًا لغويًّا جديدًا بهدف بناء محلل نحوي قائم على الوظائف النحوية،لتحقيق المهمة الأساسية للتحليل النحوي الآلي وهي الفهم الأتوماتي للنصوص اللغوية، ثم ذكرت مراحل البناء من اختيار المدونة اللغوية، وتجزئة النصوص، والتحليل النحوي، ثم بيّنت استخدام البنك الشجري في بناء محللٍ نحويٍّ، وآمل أن تزيد عدد كلمات البنك الشجري ليصل إلى مليون كلمة.

10- المراجع

[1] السعيد، المعتزبالله (2010): مدونة معجم تاريخي للغة العربية، معالجة لغوية حاسوبية، أطروحة دكتوراه، كلية دار علوم-القاهرة.

[2] شمس الدين، جلال (د ت): الأنماط الشكلية لكلام العرب، مؤسسة الثقافة الجامعية-الإسكندرية.

[3] عاشور، المنصف (1991): بنية الجملة العربية بين التحليل والنظرية، منشورات كلية الآداب بمنوبة-تونس.

[4] عبادة، محمد (2007): الجملة العربية مكوناتها-أنواعها-تحليلها، مكتبة الآداب-القاهرة.

[5] علي، نبيل (1988): اللغة العربية والحاسوب، تعريب-القاهرة.

[6] **Abeillé**, **A**.(2003) Building and Using Parsed Corpora, Springer.

[7] **Attia, Mohamed**.(2004) Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications, Faculty of Engineering, Cairo University

[8] **Attia, Mohammed**.(2007) Arabic Tokenization System in Proceedings of 5th Workshop on important unresolved Matters.

[9] **Habash, Nizar and Farag, Reem, and Roth, Ryan**.(2009) Syntactic Annotation in Columbia Arabic Treebank, In proceedings of the 2nd International Conference on Arabic Language Resources Tools.

[10] **Habash , Nizar**.(2010), Introduction to Arabic Natural Language Processing, A Publication in the Morgan & Claypool Publishers series.

[11] **Jurafdky, Daniel** & **Martin, James H**. (2006). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recongnition, Formal Grammars of English.

[12] **Slonneger**, **Kennth** and **others**, (1995). Formal Syntax and Semantics of Programming Languages.

[13] **Pustejovsky, James and Stubbs, Amber**. (2012) Natural Language Annotation for Machine Learning, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol.

[14] **Rambow, Owen**. (2010) The Simple Truth about Dependency and Phrase Structure Representations, In proceedings of the 2010 Annual Conference Chapter of the ACL.

السيرة الذاتية

أحمد روبي محمد عبد الرحمن

باحث لغوي في معالجة اللغة العربية آليا بإحدى شركات تقنية اللغات الطبيعية بالقاهرة، درس اللغة العربية التراثية والمعاصرة في مرحلة الليسانس بكلية دار العلوم جامعة الفيوم،وفي مرحلة الماجستير ركز على اللغويات بشكل عام واللغويات الحاسوبية بشكل خاص، كما عمل علىمشروعات عدة في معالجة اللغة العربية آليًّا، منها في معالجة الكلام آليًّا(تحويل المكتوب إلى منطوق، وتحويل المنطوق إلى مكتوب)، والتشكيل الآلي، والمدقق الإملائي للغة العربية، والتعرُّف على الأسماء، ويهتم بالنظريات اللغوية الحاسوبية.

Building Syntactic Treebank for Modern Standard Arabic

Ahmed Ruby Mohammed
Faculty of Dar Al Uloom, Fayoum University, Egypt
Ahmedruby757@yhoo.com

Abstract-***The syntactic Treebank is an important resource for building applications for statistical natural language processing (NLP) and used to study syntactic phenomena and also provide evidence of coverage and support the discovery of new, unanticipated, grammatical phenomena. Syntactic Arab Treebank (SATB)is a set of grammatical analysis of Arabic sentences, has been based on Islam on line corpus as a source of texts to be analyzed, and SATB differs from other tree banks in terms of linguistic information and the method of representation; and that the nature of the objective of building Parser based on semantic grammatical functions; to achieve the primary task of the automated analysis of grammar, which provide the necessary linguistic analysis to Natural Language Understanding, I used the Phrase Structure Tree representation in the analysis taking into account to determine the grammatical relations per Token in the sentence, and so for two main reasons: first easily extract the distinctive features of tokens constituent of the sentence, and the other is to get a hybrid attributes between dependency and structure X-bar schema in terms of identifying grammatical dependency relations between the heads and complements phrases. This paper is a description of the stages of the constructions sequent and then compare it to other tree banks and how to use it.***

Keywords:
Tree banks – Syntactic analysis – Annotation – Syntactic Parser – Natural Language processing – Corpus Annotation