بسم الله الرحمن الرحيم

التشريح البنائي لمشكِّل آلي عربي

لتوظيفه في نظام تخليق آلي للصوت المنطوق من النص العربي المكتوب

د. محمد عطية محمد العربي

استشاري تقنيات معالجة اللغات الحية

بالشركة الهندسية لتطوير نظم الحاسبات RDI

m_Atteya@RDI-eg.com

٠- شكر وتنويه

يتقدم المؤلف بجزيل الشكر لأسرة كلية علوم الحاسب والمعلومات "بجامعة الإمام محمد ابن سعود الإسلامية" وذلك لاستضافتها الكريمة إياه لعرض هذا العمل العلمي، ويذكر كذلك بوافر الامتنان والعرفان "الشركة الهندسية لتطوير نظم الحاسبات" RDI ; www.RDI-eg.com لدعمها المتواصل بقوة على مدى سنوات لهذا العمل المفصَّل في رسالته للدكتوراه [٢٨] والمجمَل لأول مرة بالعربية في هذا المقال، كما يخص بالثناء أستاذه الفاضل "أ.د.محسن عبد الرازق علي رشوان" المدير التنفيذي العام للشركة والأستاذ بقسم الاتصالات الكهربية والإلكترونيات "بجامعة القاهرة".

نسأل الله تعالى أن يوفق الجميع، وأن يجعل عملهم نافعًا مثمرًا لما فيه خير الناس.

١- ملخص المقال

تَبْرُزُ الحاجَةُ لِمُشَكِّلٍ صَوْتِيٍّ عَرَبِيٍّ آلِيٍّ رَفيعِ الأَدَاءِ عِنْدَ بِنَاءِ العَديدِ مِنْ نُظُمِ مُعالَجَةِ الصَّوْتِ رَقْمِيًّا؛ خاصَّةً "تخليقُ الكَلَامِ المنطوقِ مِن النَّصِّ العَرَبيِّ المكتوب".

وفِي هذا المَقَالِ نَتعرَّضُ باختصارٍ لِكَيْفيَّةِ التَّغلبِ عَلَى ما يَلِي مِنْ مَشاكِلَ فِي اللِّسانيَّاتِ العَرَبيَّةِ:

- استنباطُ التَّركيبِ الصَّرْفِيِّ بدِقَّةٍ عالِيَةٍ لِكُلِّ الكَلِماتِ العَرَبيَّةِ فِي النَّصِّ، ومِنْ ثَمَّ استنباطُ التَّشكيلِ الصَّرْفِيِّ لَهَا.
- استنباطُ عَلَاماتِ التَّشكيلِ الإعرابيَّةِ فِي المَواضِعِ الَّتِي تَتطلبُ ذلِكَ.
- استنباطُ الوَصْفِ الصَّوْتيِّ لِلْكَلِماتِ الخارِجةِ عَن النُّمُوذَجِ الصَّرْفِيِّ؛ خاصَّةً الأَسْماءَ الأَجْنَبيَّةَ المكتوبةَ بحروفٍ عَرَبيَّةٍ.
- تعديلُ الوَصْفِ الصَّوْتِيِّ المنفصلِ لِلْكَلِماتِ المتتابعةِ لِتوليدِ الوَصْفِ الصَّوْتِيِّ المُلائمِ لِنُطْقِها مُتَّصِلَةً.

وبـناءً عَلَى حُلول تِلْكَ المَسائل فإنَّنا نَعْرِضُ لِكَيْفِيَّةِ بناءِ المُشَكِّلِ الآلِيِّ العَرَبيِّ *ArabDiac* [1] وإدماجِهِ فِي عَدَدٍ مِنْ نُظُمِ تِقْنِيَّاتِ هَنْدَسةِ اللُّغَةِ العَرَبيَّةِ الموظَّفةِ فِي صِناعَةِ الاتصالاتِ والمعلوماتيَّةِ.

وأَثْـناءَ العَمَـلِ عَلَـى بُلُـوغِ تِلْـكَ الغايَـةِ فقَـدْ جَـرَى التَّأصيلُ النَّظَرِيُّ لِمَا يَلِي مِن أُطرُوحاتٍ لِسانِيَّةٍ هامَّةٍ عُمومًا، ومُهِمَّةٍ فِي العَرَبيَّةِ خُصوصًا:

- المُـزاوَجةُ فِـي إطـارِ الـذَّكَاءِ الاصـطناعِيِّ بَـيْنَ أُسْـلوبِ القـواعدِ القَطْعِـيَّةِ الحَـصْرِيَّةِ، والأسـاليبِ الاحتماليَّةِ، فِي تَناوُلِ مَسائلِ اللِّسانِيَّاتِ عالِيَةِ الالتباسِ عُمومًا والعَرَبيَّةِ مِنْها عَلَى وَجْهِ الخُصوص.
- بـناءُ عَمَلـيَّةِ العَـنْوَنةِ الـنَّحْوِيَّةِ العَـرَبيَّةِ بـناءً نَمَطِـيًّا مُطَّـرِدًا نابعًـا بشَكْلٍ طَبيعيٍّ مِنْ عَمَلِيَّةِ التَّحليلِ الـصَّرفِيِّ، ومُعَـرَّفًا بفِـئَةٍ محـدودةِ الحَجْمِ مِن العَنَاوينِ تُغَطِّي الصِّفاتِ النَّحْوِيَّةِ المُمْكِنُ استنباطُها مِن بِـنَـى الكَلِماتِ المنفصلةِ.
- توظـيفُ العَـنْوَنةِ الـنَّحْوِيَّةِ العَـرَبيَّةِ في استنباطِ عَلَاماتِ التَّشكيلِ الإعرابيَّةِ احتماليًّا لِلكَلِماتِ السَّابقِ تحليلُها صَرْفِيًّا.
- إضـافَةُ طَـبَقةٍ أَوَلِـيَّةٍ تَخْـتَصُّ بالـنَّحْوِ الـصَّوْتِيِّ إلَـى سُلَّمِ طَبَقاتِ المُعالَجةِ اللُّغَويَّةِ فِي البناءِ اللِّسانيِّ التَّجـريدِيِّ، واستخدامُ تِلْكَ الطَّبَقةِ كشَرْطٍ حَدِّيٍّ لِلأُسلوبِ الإحصائيِّ لِلْبَحْثِ باستخدامِ خُوارِزْم *A* واستخدامُ ذلِكَ لِاستنباطِ الوَصْفِ الصَّوْتِيِّ لِلأَسْماءِ الأَجْنَبيَّةِ المكتوبةِ بحروفٍ عَرَبيَّةٍ.
- صِياغةُ النَّحْوِ الصَّوْتِيِّ العَرَبيِّ رِيَاضِيًّا صِياغةً إثباتيَّةً حَصْرِيَّةً فِي فِئَةٍ محدودةٍ مِن القواعدِ الرِّياضيَّةِ بـصِيغَةِ BNF وتطبـيقُهُ فِـي اسـتنباطِ الوَصْـفِ الـصَّوْتِيِّ لِلأَسْـماءِ الأَجْنَبـيَّةِ المكتوبةِ بحروفٍ عَرَبيَّةٍ بالأسلوبِ المذكور سابقًا.

وكـذلكَ نَعْـرِضُ فِي هذا المَقَالِ تشريحَ المِعْمارِ اللِّسانِيِّ لِهذا العَمَلِ *ArabDiac*، وأَخِيرًا نَستعرِضُ شُروطَ التَّدريبِ الإحصائيَّةِ ونُحَلِّلُ نَتائجَ الأَدَاءِ فِي ظُروفِ التَّشغيلِ الواقعيَّةِ.

٢- تعريف مسألة التشكيل الآلي للنص العربي [٢٨؛ فصل ١ – قِسْم ١]

بالـرغم مـن أن اللغـة العربية بعراقتها هي لغة شديدة الثراء والحساسية تجاه تركيبها الصوتي الـذي يعـتمد بعمـق على البِـنَـى الصرفية والنحوية والدلالية للنص محل الدراسة، فإن الكتابة المعاصرة للـنص العربـي تخلـو عـادةً مـن علامـات الـضبط الصوتي (التشكيل) [٢٨؛ فصل ١ – فقرة ٢]، [٢٩]

[1] هذا هو الاسم التجاري لتقنية تشكيل النص العربي في الشركة الهندسية لتطوير نظم الحاسبات [٢٤].

وتكتفي فقط بحروف الهجاء تاركةً لفِطنة القارئ استخلاص التشكيلات الصوتية الواجبة الموائمة للبنى الصرفية والنحوية والدلالية.

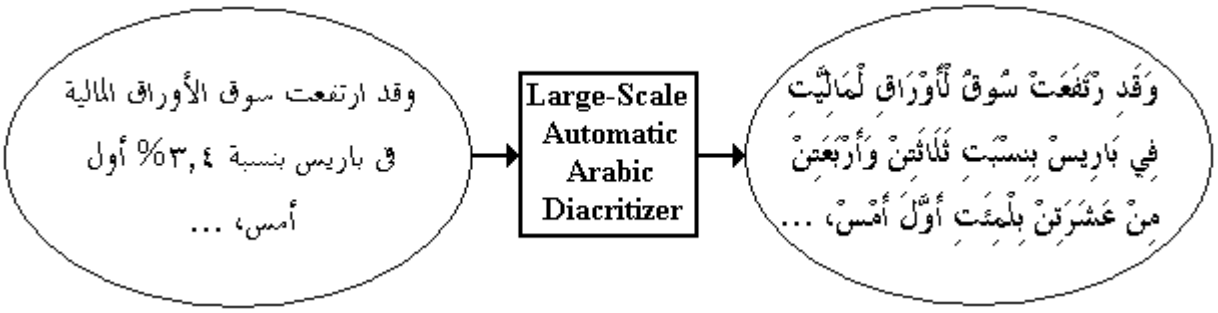
ومن المعلوم أن البون ما زال شاسعًا بين القدرات البشرية والمعالجات الحاسوبية في تحديد التراكيب اللغوية – خاصةً النحوية والدلاليّة – المناسبة لنصٍّ معيَّن في سياق معيَّن، مما يجعل بالتبعية غياب علامات التشكيل تحديًا أمام استخلاص الوصف الصوتي لنص ما حاسوبيًّا لبناء التطبيقات المتعددة المستلزِمة لذلك ومن أمثلتها الهامة؛ تخليق الكلام المنطوق من النص العربي المكتوب [٢٧]، [٤٠].

وتحتاج مثل هذه التطبيقات إلى استخلاص التوصيف الصوتي آليًّا من النص الخام في إحدى الصور القياسية المتفق عليها دوليًّا لوصف الأصوات اللغوية البشرية مثل ... SAMPA, IPA, ، ويوضح الجدول رقم ١ أدناه فقرة من نص خام، ثم توصيفها الصوتي بالكتابة الصوتية العربية، ثم أخيرًا التوصيف الصوتي المناظر في الصيغة القياسية SAMPA:

نَصٌّ خامٌ غير مشكَّل	**وقد ارتفعت سوق الاوراق المالية فى باريس بنسبة ٣,٤٪ عند الإقفال أول أمس، ...**
نفس النص مشكَّل بالكامل ومكتوب كتابةً صوتيَّةً	**وَقَدِ رْتَفَعَتْ سُوقُ لْأَوْرَاقِ لْمَالِيَّتِ فِي بَارِيسْ بِنِسْبَتِ ثَلَاثَتِنْ وأَرْبَعَتِنْ مِنْ عَشَرَتِنْ بِلْمِئَتِ عِنْدَ لْإِقْفَالِ أَوَّلَ أَمْسْ،...**
نفس النص موصوف صوتيًّا بصيغة SAMPA	**waqadi rtafa?`at su:qu l?awra:qi lma:lijjati fi ba:ri:s binisbati Tala:Tatin wa?arba?`atin min ?`aSaratin bilmi?ati ?`inda l?iqfa:li ?awwala ?ams, ...**

جدول ١: التوصيف الصوتي لعينة نصٍّ عربي بالكتابة الصوتية العربية وكذلك بصيغة SAMPA

ولما كان تحويل رموز الكتابة الصوتية العربية إلى إحدى صيغ التوصيف الصوتي الدولية القياسية يتم مباشرة عبر جدول بسيط دون أية التباسات، فإن مسألة التوصيف الصوتي للنص العربي تؤول ببساطة إلى مسألة التشكيل (الضَّبط) الصوتي الكامل لأي نصٍّ خامٍ كما هو مبين بالشكل التخطيطي أدناه:



شكل ١: رسم توضيحي لمدخلات (يسارًا) ومخرجات (يمينًا) المشكِّل الآلي للنص العربي المراد بناؤه.

٣- التحديات الأساسية [٢٨؛ فصل ١ – قِسْم ٣]

إن أيـة محاولـة جـادة لبـناء مـشكّلٍ نـصّيٍّ عربـيٍّ شـاملٍ ذي أداءٍ يُعـتمَد عليه ينبغي عليها أن تستطيع التعامل بفعالية مع التحديات التالية:

أ- كتابة النص العربي المعاصر عادة دون علامات الضبط (التشكيل). [١٣]، [١٥]، [٢٩]، [٣٠]، [٣٥].

ب- وجود العديد من الأخطاء الإملائية (الشائعة) في النصوص الخام المراد معالجتها. [٣٠]

ت- عـدم جـدوى الـتعامل مع مفردات اللغة العربية بصورة سردية حصرية (مجدولة) نظرًا لطبيعتها التوليـدية الاشـتقاقية الفائقة، ومن ثَـمَّ فإنه من دون آلية عميقة للتحليل والتركيب الصرفي قائمة علـى اللبنات البنيوية الأساسية (المورفيمات) فإن مشاكل قصور تغطية العدد الهائل من الكلمات العربية الممكن توليدها سوف تنشأ حتمًا معيقةً أية معالجات لغوية لاحقة. [٣٠]

ث- اسـتنادًا علـى دراسـاتنا الإحـصائية الموسـعة، فـإن حوالـي ٣٥٪ من ورود الكلمات في النصوص العـربية هـو لكلمات لا يعتمد وصفها الصوتي (المنفصل) إلا على بنيتها الصرفية، بينما ٦٥٪ هو لكلمات يعتمد وصفها الصوتي (المنفصل) الكامل على كلٍّ من بنيتها الصرفية وموقعها الإعرابي.

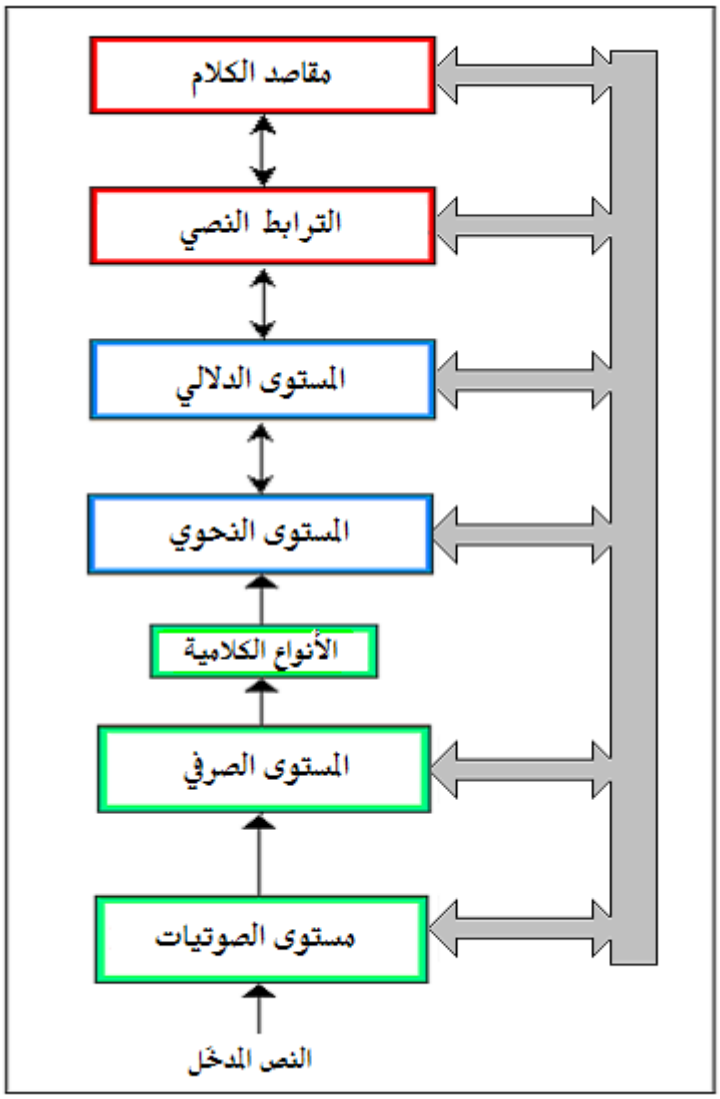
ج- إن المحلِّـلات الصرفية والمحلِّلات الإعرابية الموسعة للنص العربي تنتج حلولاً عديدة لكل مفردة في الـنص موضـع الدراسة، وعليه فإن استخلاص أنسب حل لفقرة/لجملة/لعبارة هو عملية عالية الالتباس (في الحقيقة يتوالد الالتباس أُسِّيًّا مع طول سلسلة الكلمات المطلوب إيجاد حل لها).

ح- مـثل بقـية اللغـات الحـية؛ فـإن الكلمـات الأجنبية المكتوبة بحروف عربية تتواجد في النصوص العـربية المعاصـرة بنـسبة تكرار لا يمكن التغاضي عنها (٧٫٥٪ في النصوص الإخبارية على سبيل المـثال). ومـن الواضـح أن التشكيل الصوتي لمثل تلك الكلمات ليس محكومًا بالصرف ولا بالنحو العربيين مثل المفردات العربية. [٣١]

خ- وأخـيرًا ولـيس آخِـرًا، فـإن كل ما أسلفناه من التباسات ينبغي نَظَرِيًّا أن يَتِمَّ البَتُّ فيه بالتفاعل بـين تحلـيلات النص المراد دراسته على كل المستويات (أو الطبقات) اللسانية المبينة في شكل ٢ أدنـاه بآلـية آنـية ثنائـية الاتجـاه. فالمـستوى الـنحوي – علـى سـبيل المثال – يستبعد سلاسل التراكـيب الـصرفية المتتالـية المـستحيلة نحويًّا، كما أن المستوى الدلالي بدوره يستبعد التراكيب الـنحوية غير المفيدة دلاليًّا، ...، وهكذا على طول سلم الطبقات اللسانية الاسترشادي الموضح في

الـشكل حتى يتم استخلاص حل وحيد على كل مستوى لساني يناظر مقاصد المتكلم أو الكاتب. [٣٦]

ولـسوء الحـظ فـإن قدرات العلوم اللسانية الحاسوبية المعاصرة ما زالت بعيدة عن إنتاج محلِّلات لغـوية علـى المـستويات اللـسانية العليا مقبولة الأداء تحت الظروف الواقعية، فضلاً عن قصورها عـن إنجـاز مثل تلك التفاعلية الآنية الموصوفة أعلاه بين هذه المستويات اللسانية (بما يقترب من محاكاة ذهن الإنسان فيما يتعلق بمَلَكته اللغوية)!



شكل ٢: السلم الاسترشادي للطبقات اللسانية، والتفاعلية الآنية/ثنائية الاتجاه بين تلك المستويات.

٤- حلولٌ عمليةٌ؛ نقاط التجديد والابتكار [٢٨؛ فصل ١ - قِسْم ٣]

لمجابهـة التحديات المذكورة أعلاه فقد تم تصميم وتنفيذ عدد من الحلول العملية المتضمنة لعدد من الجوانب الابتكارية نوجزها فيما يلي:

أ- تمت المزاوجة في هذا العمل بين القواعد اللغوية الْمُحْكَمة (المعروفة والقابلة للصياغة الرياضية) وبين الأساليب الإحصائية (انظر قِسْم ٦ من هذا المقال)، وذلك بعد تأسيس تنظيري عميق لهذه المزاوجة [٢٨؛ فصل ٢]، [٣٣]. ويُعَد هذا العمل (المشكِّل العربي الآلي) مثالاً واقعيًّا موسَّعًا لهذا التوجه التزاوجي الذي يفرض نفسه في هذه المرحلة من عمر اللغويات الحاسوبية كأفضل ما هو ممكن من حلول لمسائلها المعقدة.

ب- تم توظيف محلِّل صرفي عربي واسع المدى[2] *ArabMorpho*© ذي وحدات بنائية مورفيمية — وليست معجمية كجذع الكلمة أو الكلمة كاملةً مثلاً – (انظر قِسْم ٧ من هذا المقال) لمعرفة التشكيلات الصوتية الصرفية الممكنة لكلمات النص، مما يمكنه من فك وتركيب كل ما تسمح به قوانين الصرف والصوتيات العربية من ألفاظ (بمعدل فشل مقيس يقل عن ٠,٢٪). [٢٨؛ فصل ٣]، [٣٠]

ت- نظرًا لعدم توافر محلل نحوي عربي مكتمل ذي اعتمادية مناسبة (حتى الآن)، إضافةً إلى حتمية الالتباس الواسع المتوالد أسِّيًّا في مخرجات مثل ذلك المحلل النحوي (وذلك بسبب الطبيعة الشجرية للتراكيب النحوية) في غيبة محلِّل دِلالي عربي، فإننا قد اتبعنا أسلوبًا إحصائيًّا لاستنتاج التشكيلات الصوتية الإعرابية (انظر قِسْم ٨ من هذا المقال) بشكل عملي يقوم على الاقتران الاحتمالي بين الأنواع الكلامية (وهي مدخلات أي محلل نحوي على أية حال) وعلامات الضبط الإعرابية لكلمات النص. [٢٨؛ فصل ٤]، [٣٢]

ث- وأثناء السعي إلى إنجاز الأسلوب الموضح في النقطة السابقة فقد تم تطوير معنونٍ عربي للأنواع الكلامية *ArabTagger*©[3] (مبنيّ فوق المحلل الصرفي العربي السالف ذكره *ArabMorpho*©) يتميز بأن مخرجاته هي متَّجَهٌ من الأنواع الكلامية لكل كلمة عربية لاستيعاب طبيعتها المركبة، وبأن فئة الأنواع الكلامية (المحدودة الحجم) قد تم تصميمها منذ البداية خِصِّيصًا للُّغة العربية (وليست مستعارة من فئة أنواع كلامية مصمَّمة للُغاتٍ أخرى). وبينما يستعرض قِسْم ٨ من هذا المقال بإيجاز هذا المعنون للأنواع الكلامية، فإن المرجعين [٢٨؛ فصل ٤]، [٣٢] يتناولانه بالتفصيل.

ج- للتعامل مع مسألة استنتاج أنسب وصف صوتي للكلمات الأجنبية المكتوبة بحروف عربية، فإننا قد اتبعنا أسلوبًا إحصائيًّا لاستنتاج تلك التشكيلات الصوتية (انظر قِسْم ٩ من هذا المقال) يقوم

---

[2] هذا هو الاسم التجاري للمحلل الصرفي العربي في الشركة الهندسية لتطوير نظم الحاسبات [٢٥].

[3] هذا هو الاسم التجاري للمعنون العربي للأنواع الكلامية في الشركة الهندسية لتطوير نظم الحاسبات [٢٦].

علـى الاقـتران الاحتمالي بين علامات الضبط الصوتي مقيَّدًا بموافقة النحو الصوتي العربي [٢٨؛ فصل ٥]، [٣١]. وعند التأمل في هذا الأسلوب نجد أنه قابل للتطبيق مع لغات أخرى بافتراض توافر النموذج الرياضي والتنفيذ الحاسوبي للنحو الصوتي الخاص بتلك اللغة.

ح- وأثـناء الـسعي إلى إنجـاز الأسـلوب الموضـح في الـنقطة الـسابقة فقد استلزم الأمر توصيفًا رياضيًّا للقـواعد الحاكمـة للصوتيات العربية، وهو ما اصطلحنا على تسمية نموذجه الرياضي الذي تمت صـياغته تشومِـسْكِيًّا[4] بإحكـام لأول مـرة باسم "النحو الصوتي العربي" (انظر قِسْمَيْ ٩ و ١٠ من هـذا المقال) [٢٨؛ فصل ٥]، [٣١]. بحيث يسمح هذا النحو الصوتي العربي (تحليليًّا) أن يحكم بمـوافقة أو عـدم مـوافقة أيـة سلـسلة فونـيمات مطـروحة لقـواعد الـصوتيات العربية، كما يُمَكِّن (تركيبيًّا) كذلك بربط (أو لصق) التوصيفات الصوتية للكلمات المنفردة في عبارات متصلة صوتيًّا.

٥- هيكل الحل المقترح [٢٨؛ فصل ١ - قِسْم ٤]

يوضـح شـكل ٣ التخطيطـي أدنـاه الهـيكل المقـترح لبـناء المشكِّل الآلي للنص العربي والعلاقة التفاعلية بين الحلول العملية المشار إليها أعلاه والمخصَّص لكل منها قسم تالٍ من أقسام هذا المقال.

بدايـةً يـدخل الـنص الخـام المـراد تـشكيله إلى معـالج لتفـصيحه؛ بمعنـى تحـويل التعبيرات العددية، والاختـصارات الـشهيرة، وصـيغ التواريخ والوقت، ...، وما شابَهَ إلى تعبيرات نصية كلامية بالكامل، ثـم يمـرَّر الـناتج إلى المحلـل الـصرفي العربي الذي يُنْتِج أرجح المورفيمات (إحصائيًّا) المكافئة للنص وما يتصل بها من معلومات لسانية؛ وهي:

أ- التشكيلات الصوتية الصرفية للكلمات المحلَّلة.

ب- الأنـواع الكلامـية للكلمـات المحلَّلـة (بمعرفة معنونِ الأنواع الكلامية) والتي تمرر بدورها إلى آلية استخلاص علامات الضبط الإعرابي إحصائيًّا.

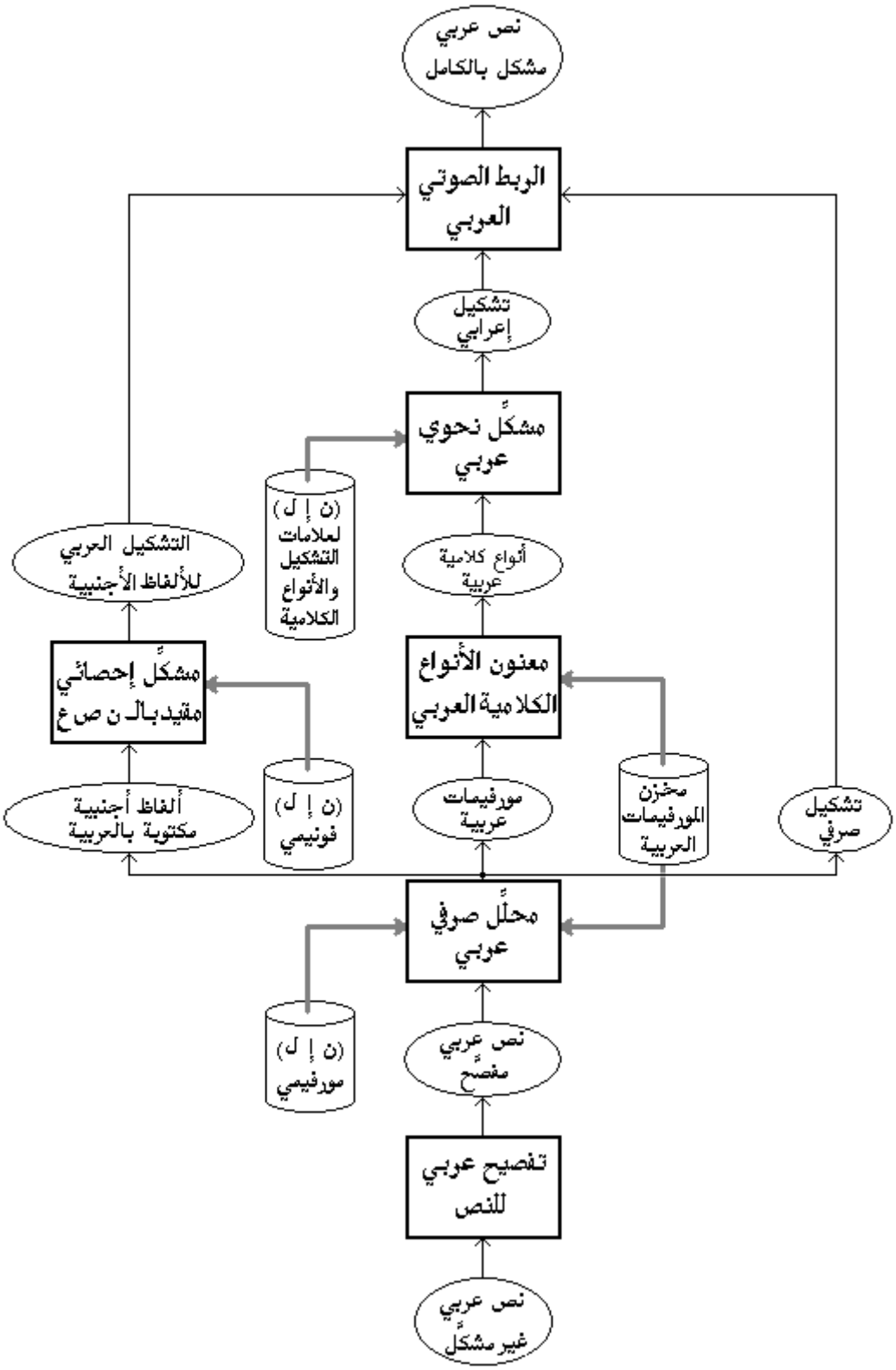
ت- الألفـاظ الـتي قـرر المحلِّل الصرفي أنها أجنبية مكتوبة بالحروف العربية، والتي تمرر بدورها إلى آلية تشكيل تلك الألفاظ إحصائيًّا.

تمـرَّر كـل التشكيلات الناتجة للكلمات المفردة من أيٍّ من هذه الأنواع الثلاثة إلى آلية الرَّبْط (أو اللَّصْق) الصوتي للكلمات المنفردة في عبارات متصلة صوتيًّا.

ويلاحـظ أن الآليات الإحصائية الثلاث تستند إلى قاعدة بيانات إحصائية تُدْعَى نموذجًا إحصائيًّا لغويًّا (ن إ ل) تتكوَّن أثناء تدريب كل آلية تدريبًا إشرافيًّا Supervised Training قبل استخدامها.

---

[4] نِسبةً إلى رائد اللغويات/اللسانيات الحاسوبية "نُوَام تُشُومِسْكِي" (ويُنْطَقُ اسْمُهُ أحيانًا؛ نَعُوم تُشُومِسْكِي).



شكل ٣: هيكل الحل المقترح لبناء المشكِّل الآلي العربي.

٦- فك الالتباس الإحصائي [٢٨؛ فصل ٢]، [٣٣]

يعرض هذا القسم بإيجاز للآلية العامة لفك الالتباس الإحصائي المستخدَمة في ثلاثة مواضع حيوية من نظامنا للتشكيل الصوتي العربي ألا وهي؛ فك الالتباس الصرفي، واختيار أرجح علامة ضبط إعرابي، واستنتاج أنسب تشكيل صوتي للكلمات الأجنبية المكتوبة بحروف عربية.

ونود البدء في هذا المقام بالتأكيد على أن استخدام الآليات الإحصائية لمعالجة مسائل اللسانيات حاسوبيًّا هو حتميةٌ ناشئةٌ عن العوامل التالية: [٣٦]، [٤٨]، [٥٢]

أوَّلاً؛ أنَّ أغلب المعلومات المتعلقة بالنص (مكتوبًا كان أو منطوقًا) موجودة خارج النص في السياقات (أو بيئة النص) المحيطة به، في حين أن النص يُفِيدُ (غالِبًا) التغيُّرَ (التفاضلي) فقط في واحدٍ أو بَعْضٍ من هذه السياقات.

ثانياً؛ عِلْمِنا غير المكتمل بالقوانين العميقة الحاكمة لديناميكيات اللغات البشرية خاصَّةً في طبقاتها اللسانية العليا.

ثالثاً؛ عَدَمِ التوصل لميكانيكيةٍ متكاملة للتفاعل ثنائي الاتجاه والآني بين هذه الطبقات اللسانية المشار إليها في قِسْم ٤ عالِيَهُ.

رابعاً؛ الالتباس الحقيقي الكامن في كلام الكاتب/المتحدِّث عن قصد أو عن غير قصد والذي لا يستطيع الإنسان القارئ/السامع من أهل لغته فَكَّهُ دون إعادة السؤال عن مَقْصِد الكلام.

وبالرغم من كونها تعويضيةً، فلا يَظُنَّنَّ قارئ هذه المقال أن الآليات الإحصائية مجرَّدَ حساباتٍ سطحيةٍ عاجزةٍ، بل إنها في الواقع قد حققت إنجازات معتبَرةً ارتقت بالعديد من المعالجات اللسانية الحاسوبية إلى تقنيات صناعية يُعتَمد عليها. وفضلاً عن ذلك فإن تِلْكُمُ الآليات ذات جذور عميقة في السلوك التَّعلُّميّ الإنساني اللغوي الذي تسهل ملاحظته مثلاً؛ في تعلم الأطفال لِلُّغةِ، وعند محاولة غير المتخصص لقراءة نصوص متخصِّصة في غير تخصُّصِه، ...، إلخ. [٤١]، [٤٩]

وانطلاقًا مما سبق فقد جاءت الاستعانة بالآليات الإحصائية تكاملاً مع ما هو معروف وقطعي من قواعدَ محْكَمة، وفي حين أنه توجد عدة توجهات (مدارس) أخرى في هذا الصدد (انظر [٤٧] على سبيل المثال، وكذلك [٤١]، [٤٩])، فإن الاختيار في هذا العمل قد وقع على منهجية اختيار الحل ذي الأَرْجحية القُصوى (المنبثقة أصلاً عن نموذج "قناة الاتصال المشوشَرة" السائد في علوم الاتصالات الإلكترونية [٤٤]) وذلك بعد صياغة مسألة الالتباس في صيغتها العامة التالية الموضحَّة في شكل ٤ أدناه والتي تَسْمَح بإعادة استخدام تلك الآلية في أية مسألة يمكن إخضاعها لتلك الصياغة.

w_1	w_2	$\cdots$	w_L
$q_{1,1}\bullet$	$q_{2,1}\bullet$	$\cdots$	$q_{L,1}\bullet$
$q_{1,2}\bullet$	$q_{2,2}\bullet$	$\cdots$	$q_{L,2}\bullet$
$\vdots$	$\vdots$		$\vdots$
$q_{1,J_1}\bullet$	$q_{2,J_2}\bullet$	$\cdots$	$q_{L,J_L}\bullet$
$\underline{q}_1$	$\underline{q}_2$	$\cdots$	$\underline{q}_L$

شكل ٤: الصياغة العامة لمسألة الالتباس اللساني.

تعبِّر w_i في الشكل ٤ أعلاه عن أية وحدة (فونيم، مورفيم، ...، كلمة، ..) في النص المطلوب تحليله $w_1w_2...w_L$ ، كما تعبر $\underline{q}_i$ عن فئة الحلول (التأويلات) الممكنة لتلك الوحدة (على أي مستوى لساني) والتي تتولد بعد تطبيق القواعد اللغوية القَطْعِيَّة المعروفة، وحسب منهجية الأرجحية القصوى فإن المطلوب هو اختيار سلسلة الحلول $\{q_{1,j_1}, q_{2,j_2},..., q_{L,j_L}\}$ (وتُكْتَبُ كذلك على الهيئة المختصرة q_{1,j_1}^{L,j_L}) ذات أعلى أرجحية من بين كل سلاسل الحلول الممكنة خلال الشبكة في شكل ٤، وهو ما يعبَّر عنه رياضيًّا بالصيغة:

$$\begin{aligned}\underline{\hat{Q}} &= \arg\max_{s}\left\{P\left(q_{1,j_1}^{L,j_L}\right)\right\}\\ &= \arg\max_{s}\left\{\prod_{i=1}^{L} P\left(q_{i,j_i} \mid q_{(i-N),j_{(i-N)}}^{(i-1),j_{(i-1)}}\right)\right\}\\ &= \arg\max_{s}\left\{\sum_{i=1}^{L} \log P\left(q_{i,j_i} \mid q_{(i-N),j_{(i-N)}}^{(i-1),j_{(i-1)}}\right)\right\}\end{aligned}$$

حيث P هي الدالة الاحتمالية وحيث N هو طول الأُفُق الارتباطي؛ ويُقْصَدُ به عدد الوحدات السابقة على أية وحدة والتي تؤثر إحصائيًّا بدرجة ملموسة على تأويل هذه الوحدة (والحالة المثالية هي أن تكون $N = i$ وعمليًّا لا يمكن تحقيق ذلك ممكنًا على الدوام، فتُختار أكبر قيمة لـ N تسمح بها القدرات الحاسوبية المتاحة).

وحيث أن عدد سلاسل الحلول الممكنة هو في الظروف الواقعية عدد هائل (يتوالد أسِّيًّا مع عرض شبكة الحلول L)، فإنه يستحيل استكشافها جميعًا لاختيار السلسلة الأعلى أرْجحيةً، ولذلك يتم استخدام خُوارِزم A^* الشهير لضمان هذا الاختيار بأقل تكلفة حاسوبية ممكنة [٤٦]. ويُعبَّر عن دالَّة تكلِفةِ مَسَارٍ ما المميِّزةِ لهذا الخُوارِزْم بالصيغة الرياضية:

$$f^*\left(k, q_{k,j_k}, L\right) = g\left(k, q_{k,j_k}\right) + h^*\left(k, q_{k,j_k}, L\right)$$

حيث g هي دالة التكلفة للجزء السالف من المسار، والتي تؤول تبعًا لمنهجية الأرجحية القصوى إلى الصيغة الرياضية:

$$g\left(k, q_{k,j_k}\right) = \sum_{i=1}^{k} \log P\left(q_{i,j_i} \mid q_{(i-N+1),j_{(i-N+1)}}^{(i-1),j_{(i-1)}}\right)$$

ولحساب الاحتمال الشَّرْطِيّ في هذه الصيغة فإننا نستخدم الأسلوب المدْمَج لتقديرات "بايِز–جُود تِيُورِنج–الارتداد الخلفي" [٣٣]، [٤٣]، [٤٥] الذي يتطلب بناء قاعدة بيانات إحصائية تُدْعَى نموذجًا إحصائيًّا لغويًّا (ن إ ل) تتكوَّن مُسبَّقًا أثناء التدريب الإشرافيّ Supervised Training من كميات ضخمة – قدْر الإمكان – من التحليلات المراجَعة بواسطة لُغويِّينَ متخصِّصِين.

أما دالَّة التكلفة المتوقَّعة h^* للجزء المتبقِّي من المسار فإن حسابها التقديري الآمن (الضامن للحصول على المسار ذي الأرجحية القصوى) يتم حَسْب الصيغة الرياضية:

$$h^*\left(k, q_{k,j_k}, L\right) = \begin{cases} \sum_{i=k+1}^{L} \log(P_{\max,N}) = (L-k)\cdot\log(P_{\max,N}); & L \geq N, k \geq N-1 \\ \sum_{i=N}^{L} \log(P_{\max,N}) + \sum_{i=k+1}^{N-1} \log(P_{\max,i}) & \\ = (L-N+1)\cdot\log(P_{\max,N}) + \sum_{i=k+1}^{N-1} \log(P_{\max,i}); & L \geq N, k < N-1 \\ \sum_{i=k+1}^{L} \log(P_{\max,i}); & L < N \end{cases}$$

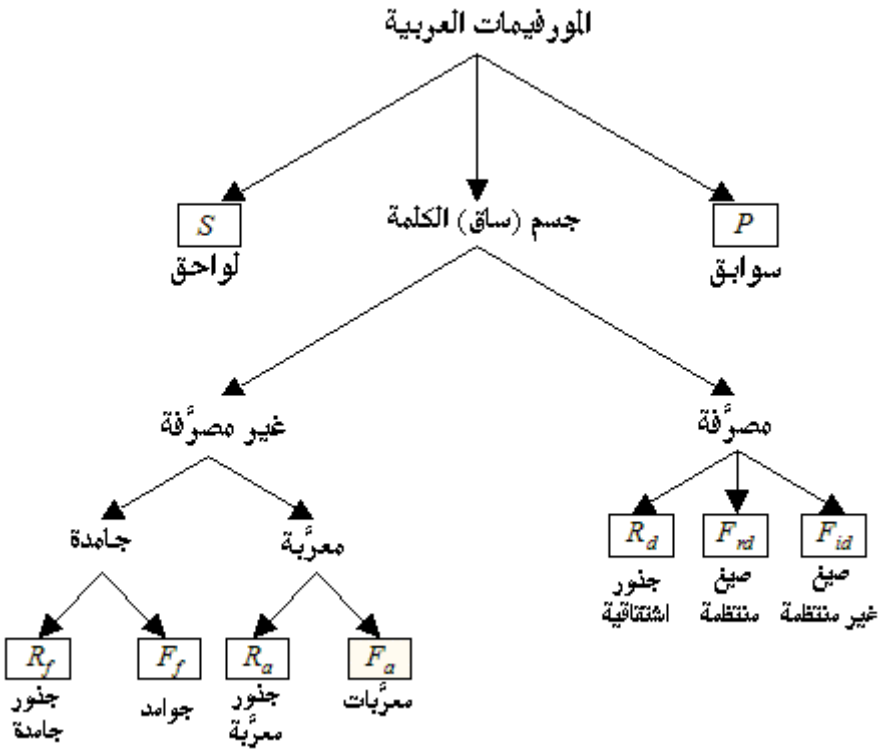
٧– المكنز المورفيمي العربي، التحليل الصرفي، واستنباط التشكيل الصرفي [٢٥]، [٢٨؛ فصل ٣]، [٣٠]

كما أَلْمَحْنا في قِسْم ٣ عالِيَهُ من هذا المقال؛ فإن الطبيعة التوليدية فائقة الاشتقاقية للمفرَدات في اللغة العربية تجعل من الأشْمَل والأنْجَع والأكثر اقتصاديَّةً أن نتعامل معها عبْرَ الفئة المحدَّدة والمحدودة لِلَبِناتِها البنائية الأساسية (المورفيمات)، وذلك بدلاً من حَصيلة مفْرَداتِها المُمْكِن توليدُها ذات الحجم الهائل (عشرات الملايين) التي تصعب (أو تستحيل) السيطرة عليها. وانطلاقًا من هذا الأسلوب المورفيمي النَّزْعة فإن البنية المعيارية لبِنْية أية كلمة عربية w تَبَعًا لنموذجنا الصرفي العربيّ يمْكِن التعبير عنها بالصيغة الرُّبَاعية:

$$w \rightarrow \underline{q} = (t : p, r, f, s)$$

وهـي ببـساطة تعـني أننا نحلِّل (أو نـَرُدُّ) أية كلمة عربية إلى أربعة عناصر بِـنْيَويَّة (مُورْفِيمات)؛ سابقة (ورَمْزُها p)، وجِـذر (ورَمْزه r)، وصِيغَة صرفيّة (ورَمْزُها f)، ولاحقة (ورمزها s).

وبوجـود نـوع الكلمـة t وهـو الـذي يمكـن أن يكـون إمـا "اشـتقاقية منـتظِمة"، أو "استثناء صرفي"، أو "جامـدة"، أو "أَعْجَمـيّة معرَّبة" فإن المورفيمات التي تكوِّن كُلّ الرباعيات التي تُمَثِّـل بشكل فريد (غير ملتبِس) أية كلمة عربية ممكِنة يمكِنُ تصنيفُها إلى تسعة أصْناف كما يُبَيِّن شكل ٥ أدناه



شكل ٥: تصنيف الأنواع التسعة للمورفيمات في المكنز المورفيمي العربي الخاص بـ *ArabMorpho*©.

وبِـغَضِّ النظر عن الكلمات الأجنبية المكتوبة بحروف عربية، فإن قياساتنا المطوَّلَة تبيِّن أن معدَّل تغطية هـذا الـنموذج الـصرفي لمفـردات اللغة العربية السليمة (بعيدًا عن العامِّيَّات) يتجاوز ٩٩,٨٪ تأسيسًا على قاعـدة مَعْـرِفيَّة (المَكْنَـز المورفيمي) عِمادُها ما مجموعه فقط ٧٨٠٠ مورفيم[5] من كل الأنواع التسعة المبيَّنة عالِـيَهُ، وتحـتوي هـذه القاعـدة علـى وَصْـف لُغَوِيٍّ مستقِلٍّ لكل مورفيم من عدة نواحٍ (هجائيًّا، صوتيًّا، صَـرْفِيًّا، مـن حـيث الأنـواع الكلامـية، ...، إلخ) مَـصُوغةً بأسـلوب "العَـوَامِل البَـرْمَجِـيّة الطَّلِـيقة Software Agents" [٤٨]، [٥٢] القابلة للتفاعل المتبادَل دون معرفة مسبَّقة ببعضها البعض.

[5] تمت الاستعانة خلال ما يزيد على سِتِّ سنوات بالعديد من المصادر العربية المعتبَرة (انظر على سبيل المثال المراجع: [١]، [٢]، [٦]، [٩]، [١١]، [١٢]، [١٦]، [١٧])، والعديد من الباحثين اللُّغويين المتميِّزين، وذلك لجمع المادة الخام لبناء تلك القاعدة المعرفية.

ويعرِض شَكْل ٦ أدناه عَيِّناتٍ من مَكْنزِنا المورفيمي العربي تحتوي على وَصْف ثلاثة مورفيمات هي على الترتيب (صيغة "فَاعِل" السالِمة، السابِقة "بِالـ"، اللاحقة "ـينَ" لجمع المذكَّر حالَ نَصْبِه أو جَرِّه).

```
// ****
Spell     (mrfX1    , mrfALEF , mrfZ2    , mrfZ3    ) ;
Shadda    (mrfFALSE, mrfFALSE, mrfFALSE, mrfFALSE) ;
Voc       (mrfFTHA , mrfVWL  , mrfKSRA , mrfUNDT ) ;
Cgroup    (mrfNNBS , mrfCONS , mrfWYx  , mrfWYx  ) ;
Must1     (mrfNULL ) ; IDvec1(mrfpVORN, mrfpNOUN) ;
Must2     (mrfNULL ) ; IDvec2(mrfsVORN, mrfsNOUN) ;
POStagVec(mrfPOS_Esm, mrfPOS_EsmFa3el) ;
Flag1     (mrfBpfNMSR ) ;
Flag2     (mrfBsfNOUN ) ;
Next      () ; // 0805
// ----
```

```
Spell     (mrfBAA  ,mrfALEF ,mrfLAM  ) ;
Voc       (mrfKSRA ,mrfBYPS ,mrfBPSK ) ;
IDvec1    (mrfpNOUN,mrfpHGRR,mrfpGRRb,mrfpDEF ) ;
POStagVec(mrfPOS_7arfGarr, mrfPOS_Ta3reefAL) ;
Action1   (mrfActpDFGR) ;
Next      () ; // 035
```

```
Spell     (mrfYAA  ,mrfNOON ) ;
Shadda    (mrfFALSE,mrfFALSE) ;
Voc       (mrfVLSK ,mrfFTHA ) ;
IDvec1    (mrfsVERB,mrfsDMRF,mrfsMODR,mrfsMFmd,mrfsRFmd) ;
POStagVec(mrfPOS_Modhare3, mrfPOS_Marfou3, mrfPOS_DhameerRaf3) ;
Action2   (mrfActsDRYY) ;
Next      () ; // 0372
```

شكل ٦: عَيِّنات من المكْنَز المورفيمي العربيّ الخاص بـ ©*ArabMorpho*.

وفي حين أن الصِّيَاغات التفصيلية للمكنز المورفيمي مذكورة في [٢٥؛ فصل ٥]، فإن خُوارِزْمات التحليل/التركيب الصرفي العاملة على هذا المكنز المورفيمي مذكورة بالتفصيل في [٢٥؛ فصلَيْ ٦، ٧]، وبناءً على ذلك يمكن الحصول على نواتج التحليل الصرفي للكلمات العربية التي يعْرِض جدول ٢ أدناه (على الصفحة التالية) عددًا من أمْثِلَتِها...

الكَلِمة المِثال	نوع الكَلِمة	السابقة ورَمْزها	الجِذر ورَمْزه	الصِّيغة ورَمْزها	اللاحقة ورَمْزها
فَمَا	جامدة	فَـ ٢	اَلَّذِي ٨٧	مَا ٤٨	– ٠
تَتَنَـاوَله	اشتقاقيَّة منتظِمة	تـ ٨٦	ن و ل ٤٠٧٧	تَفَاعَلَ ١٧٦	ـه ٨
اَلْكِتَابَات	اشتقاقيَّة منتظِمة	الـ ٩	ك ت ب ٣٣٥٤	فِعَال ٦٨٤	ـات ٢٧
اَلْعِلْمِيَّة	اشتقاقيَّة منتظِمة	الـ ٩	ع ل م ٢٧٥٤	فِعْل ٨٤٢	ـيَّة ٢٨
مِنْ	جامدة	– ٠	مِنْ ٦٣	مِنْ ١١٨	– ٠
مَوَاضِيع	اشتقاقيَّة منتظِمة	– ٠	و ض ع ٤٣٣٩	مَفَاعِيل ٩٣	– ٠
مُتَّخَذة	استثناء صَرْفِيّ	– ٠	أ خ ذ ٣٩	مُتَّخَذ ١٣	ـة ٢٦

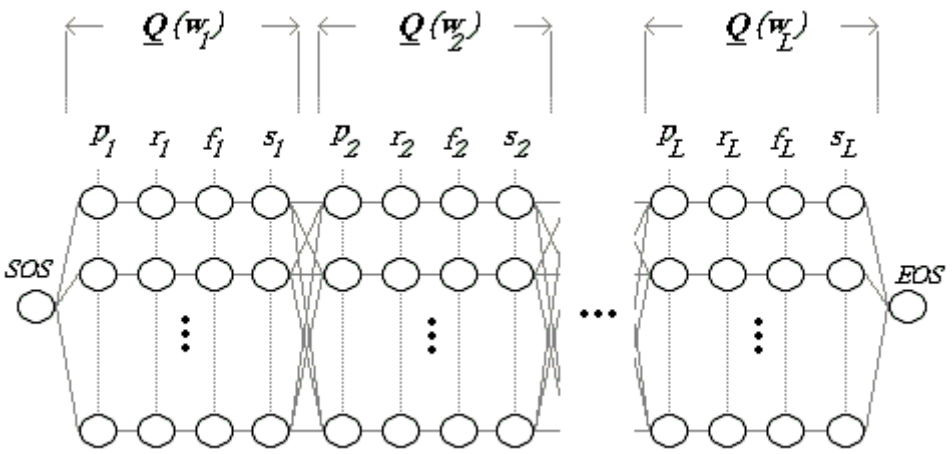
جدول ٢: أمثلة للبنية الصرفية في هيئتِها المعياريّة لعدد من الكلمات العربية.

ولتصوير الالتباس الذي لا يستهان به على المستوى الصرفي، فإننا نعْرِض في جدول ٣ أدْناهُ (على الصفحة التالية) التحليلات (التأويلات) الصرفية العديدة الممكِنة (على القِياس) والتي يستطيع نموذجنا للتحليل الصرفي توليدَها لكلمة (بطين) كمثالٍ (معتدِلٍ) على هذا الالتباس.

وعلى ذلك فإننا بعد استنفاد تطبيق نموذجنا الصرفي المشار إليه، يَتحصَّل لدينا بشكل عام حلول صرفية متعدِّدة لكل كلمة من كلمات النص المطلوب تحليله، وهو ما يمكن إخضاعه في إطار ما سَلَفَ ذِكْرُهُ من صياغةٍ عامَّةٍ لفك الالتباس اللساني (راجِعْ شكل ٤) والمخصَّصَة هنا لِفَكِّ الالتباس الصرفي بشبكة المَسَارَات (بالطوبولوجية) الموضَّحة في شكل ٧ أدناه (على الصفحة بعد التالية) والتي يُسْمَحُ فيها فقط بالمسارات على الخطوط المتصلة.

اللاحقة ورَمْزها	الصِّيغة ورَمْزها	السابقة ورَمْزها	الجِذر ورَمْزه	نوع الكَلِمة	الكَلِمة مشكَّلةً
– ٠	فَعِيل ٦٧٣	ب ط ن ٣٥٢	– ٠	اشتقاقيَّة منتظِمة	بَطِين
– ٠	فُعَيْل ٧٨٩	ب ط ن ٣٥٢	– ٠	اشتقاقيَّة منتظِمة	بُطَيْن
ـيْنِ ٨٠	فَلّ ٨٢٠	ب ط ط ٣٤٨	– ٠	اشتقاقيَّة منتظِمة	بَطَّيْنِ
– ٠	فِعل ٨٤٧	ط ي ن ٢٥٦٣	بـِ ١٥	اشتقاقيَّة منتظِمة	بـِطِين
– ٠	فِعِّيل ٦٧٠	ب ط ن ٣٥٢	– ٠	اشتقاقيَّة منتظِمة	بـِطِّين
– ٠	فُعَيِّل ٧٨٨	ب ط ن ٣٥٢	– ٠	اشتقاقيَّة منتظِمة	بُطَيِّن
– ٠	فَعْل ٨١٩	ط ي ن ٢٥٦٣	بـِ ١٥	اشتقاقيَّة منتظِمة	بـِطَيْن
ـيْنِ ٨٠	فُلّ ٨٣٤	ب ط ط ٣٤٨	– ٠	اشتقاقيَّة منتظِمة	بُطَّيْنِ
ـيْنِ ٨٠	فِلّ ٨٤٣	ب ط ط ٣٤٨	– ٠	اشتقاقيَّة منتظِمة	بـِطَّيْنِ
– ٠	فِعَل ٨٥٠	ط ي ن ٢٥٦٣	بـِ ١٥	اشتقاقيَّة منتظِمة	بـِطِيَن
– ٠	فُعَّل ٧٣٩	ط ي ن ٢٥٦٣	بـِ ١٥	اشتقاقيَّة منتظِمة	بـِطُيَّن
– ٠	فَعِّل ٦٧٥	ط ي ن ٢٥٦٣	بـِ ١٥	اشتقاقيَّة منتظِمة	بـِطَيِّن

جدول ٣: مثال على تعدُّد الحلول الصرفيّة الممكنة للكلمة العربية.



شكل ٧: طوبولوجيَّة فَكِّ الالتباس الصَّرْفِيّ.

٨- الأنواع الكلامية، واستخلاص التشكيل الإعرابي [٢٦]، [٢٨؛ فصل ٤]، [٣٢]

تُعَدُّ عَنْوَنة الأنواع الكلامية عملية أساسية من عمليات التحليل اللساني حيث تهْدِف إلى استخلاص الأنواع الكلامية وهي تلك السِّمَات التي تحمِل الخواص النحوية البدائية المميِّزة لكل كلمة منفردةً بمعزل عن سِياقِها الإعرابي في النص محل الدراسة [٤٩] (اسْم، فِعْل، حَرْف، مضارع، ماضي، مؤنث، ...، ناصب، جازم، ...)، ومن البَدَهيّ أن الأنواع الكلامية لكلمات نَصٍّ ما هي من أَهَمِّ المُدْخَلات الابتدائية لأي عملية **تحليل** إعرابي لهذا النص. فلا نستطيع مثلاً على الإطلاق أن نعرف أن كلمة (تحليل) في الجملة السابقة مضافٌ إليه دون أن نعرف أولاً أنها اسْمٌ.

وفي سبيل تصميم فئة الأنواع الكلامية العربية ضِمْنَ هذا العمل، فقد كانت تتوجَّب علينا مَهَمّةٌ (شِبْهُ) مستحيلة تقضي باستقراء الخصائص الصرف_نحوية لكل كلمة عربية مُمكِنة! وبدلاً من ذلك واستنادًا على ما سَلَف ذِكْرُه في القِسم ٧ عالِيَهُ حَوْل تغطية نموذجنا الصرفي العربي وعملية التحليل الصرفي للمفردات العربية، فإن تلك المهمة المستحيلة تؤول إلى مهمة متيسِّرة باستقراء الـ ٧٨٠٠ مورفيم في مَكْنَزِنا المورفيمي العربي للحصول عبر عدة جولات من التحليل والاختصار على أصغر فئة ممكنة من الأنواع الكلامية المعبِّرة عن أية مفردة عربية ممكنة.

ويمكن تلخيص المنهجية التي اتُّبِعَت أثناء عمليات التصميم والاستقراء والاختصار تلك في؛ الاكتمال (حيث لا توجد صفة نحوية غير سياقية في أية كلمة عربية لا يُناظِرُها نوعٌ كلاميّ في فئة الأنواع الكلامية)، والأَحَديَّة (حيث لا يُوجَدُ في فئة الأنواع الكلامية نوعٌ كلاميٌّ يكافئه نوع كلامي آخر أو أكثر)، والتَّأكُّديّة (حيث يمكن التأكد قَطْعِيًا من تحقُّق أو عدم تحقُّق أي نوع كلامي في أيٍّ من

مورفـيمات مَكْنَـزِنا المورفيمـي). هـذا ويمكن الاطِّلاع على فئة الأنواع الكلامية العربية التي توصلنا إليها كاملةً في [٢٨؛ ص ٤٥] أو [٣١] وهي تتكون من ٦٢ نوعًا كلاميًّا.

بعـد ذلـك يتم تسمية كل مورفيم في مَكْنَـزِنا المورفيمي بالأنواع الكلامية التي تستوفي خصائصه النحوية، ويعرض جدول ٤ أدناه أمثلة على تلك التسمية.

اسم المورفيم	نوعه ورمزه	متَّجَه العناوين الكلامية
الـ	سابقة؛ ٩	[ال التعريف]
سَيـ	سابقة؛ ١٢٥	[استقبال، مضارع، مبني للمعلوم]
مُفَاعِل	صيغة اشتقاقية منتظمة؛ ٤٨٢	[اسم، اسم فاعل]
اسْتِفْعَال	صيغة اشتقاقية منتظمة؛ ٦٧	[اسم، مصدر]
مَلَائِك	استثناء صرفي؛ ٢٩	[اسم، ممنوع من الصرف، جمع]
هُوَ	جامدة؛ ٨	[اسم، مذكر، مفرد، ضمير رفع]
ذُو	جامدة؛ ٣٩	[اسم، مذكر، مفرد، مضاف، مرفوع]
ـات	لاحقة؛ ٢٧	[مؤنث، جمع]
ـونَهُمْ	لاحقة؛ ٤٢٧	[مضارع، مرفوع، ضمير رفع، ضمير نصب]
ـيَّتَانِ	لاحقة؛ ١٩٥	[نسب، مؤنث، مثنى، غير مضاف، مرفوع]

جدول ٤: أمثلة على تسميات الأنواع الكلامية لعدد من المورفيمات.

وتوجد هنا ثلاث نقاط جديرة بالملاحظة:

- أنـه في أيـة كلمـة عـربية محلَّلـة حـسب نموذجنا الصرفي (انظر قِسْم ٧ السابق في هذا المقال) تنبع الأنـواع الكلامية لجسمها (أو "لِسَاقِها" بتعبير آخر) من تسمية الأنواع الكلامية للمكوِّن f المعبِّر عن صِيغتِها، في حـين تنـبع الأنـواع الكلامـية لِلَواصِقِها من تسْميتَي الأنواع الكلامية للمكوِّنَـيْنِ p و s المعبِّرَيْنِ عن سابِقَتِها ولاحِقَتِها على الترتيب. وعلى ذلك فإن مورفيمات الجذور (من كل الأنواع) لا تلعـب دورًا في اسـتنباط الأنـواع الكلامـية للكلمـة العربية ومن ثَـمَّ فلا حاجة (ولا معنى) لتسميتها بأنواع كلامية في المكنز المورفيمي.
- نظـرًا لأَحَديَّة الأنواع الكلامية من ناحيةٍ، وللبنية المركَّبة للمورفيمات العربية من ناحية أخرى فإن تـسمية الأنواع الكلامية لأي مورفيم عربي هي — كما هو واضحٌ من الجدول ٤ أعلاه – على العموم

مُـتَّجَهٌ (بالمـصطَلح الرياضـي) مـن الأنـواع الكلامية وليست اسْمًا بسيطًا مكونًا من نوعٍ كلاميٍّ واحدٍ (كما هو غالب في الإنجليزية مثلاً، انظر [٤٩]).

- يحـتوي مـتجه الأنـواع الكلامـية في تـسمية كـل مورفـيم عربـي فقـط على الأنواع الكلامية المؤكَّد حـصولها في هـذا المورفـيم، فـإذا كان المورفيم يحتمِل (ولا يحتِّم) نوعًا كلاميًّا معيَّنًا (ولو بقوَّة) فإن هذا النوع الكلامي لا يُتَضَمَّن في متَّجَه تسمية المورفيم.

والآن إذا كـان لديـنا نـَصٌّ عربـيٌّ تم تحليله وفك التباسه صرفيًّا (كما في قِسْم ٧ السابق) بحيث تُمَثَّـل كـل كلمـة تمثـيلاً فـريدًا بالـصيغة الـرباعية $\underline{q} = (t : p, r, f, s)$ ، فإنـنا للحـصول علـى متَّجَه الأنواع الكلامـية لكـل كلمـة نـستعيد أوَّلاً مـن المكنـز المورفيمـي تـسميات الأنـواع الكلامـية الـثلاث؛ لِلسَّابقة *APoS(p)*، وللصيغة *APoS(f)*، ولِلَّاحقة *APoS(s)*، ثـُمَّ تُدْمَج الثلاث تسميات عَبْرَ دالة وَصْل متجهات الأنواع الكلامية *Concat* كما يلي:

$$APoS(\mathbf{w})=Concat(APoS(p), APoS(t: f), APoS(s))$$

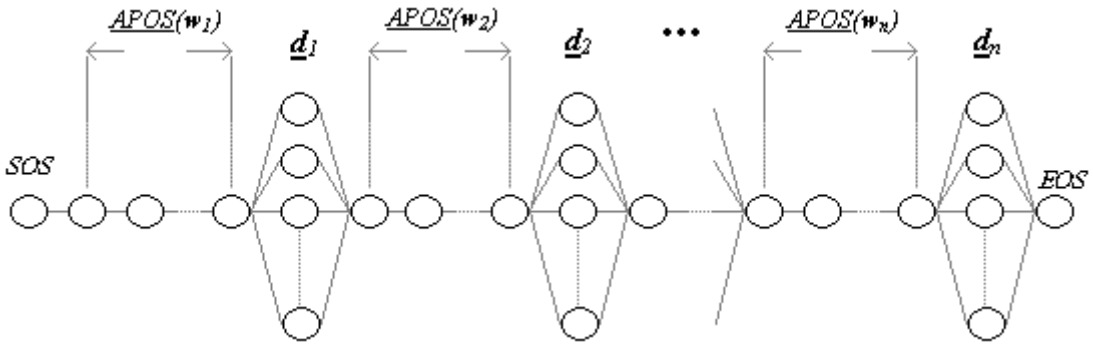
حـيث تقـوم الدالـة *Concat* بعـد حـذف التكـرارات ومعالَجَة التأثيرات المتبادلة بين الأنواع الكلامية بوصـل المـتجهات الجـزئية في مـتجه واحـد للأنـواع الكلامـية لكامل الكلمة، والجدول ٥ أدناه يعرض متجهات الأنواع الكلامية الناتجة عن هذه العملية عند تطبيقها على بعض الأمثلة من الكلمات العربية.

الكلمة	متَّجَه الأنواع الكلامية المناظر للكلمة
فَمَا	[عطف، اسم، اسم موصول، لا لاحقة]
تَتَنَاوَله	[مضارع، مبني للمعلوم، فعل، ضمير نصب]
اَلْكِتَابَات	[ال التعريف، اسم، جمع، مؤنَّث]
اَلْعِلْمِيَّة	[ال التعريف، اسم، نسب، مؤنَّث، مفرد]
مِنْ	[لا سابقة، حرف، لا لاحقة]
مَوَاضِيع	[لا سابقة، اسم، ممنوع من الصرف، جمع، لا لاحقة]
مُتَّخَذة	[لا سابقة، اسم، اسم مفعول، مؤنَّث، مفرَد]

جدول ٥: متجهات الأنواع الكلامية لبعض الأمثلة من الكلمات العربية.

وبعـد التحلـيل الصرفي، وفَكِّ التباساته، وتَمَامِ الحصول على متجهات الأنواع الكلامية لكلمات النص، وحـيث أن الأنـواع الكلامـية هـي مـن أهـم المـدْخَلات الابتدائـية لعملية التحليل الإعرابي، وحيث أن التباس أي محلِّل إعرابي حَتْمِيٌّ بدون المدْخَلات الدِّلالية (بكافة مُسْتَوَياتِها) له وهي غائبةٌ عَمَلِيًّا (حتى

الآن على الأقل)، فإننا نختصر الطريق ونقْرِن إحصائيًّا بين متَّجَه الأنواع الكلامية لكل كلمة $APoS(w_i)$ وبين علامات الضبط الإعرابية المُمْكِنة لها $\underline{d}_i$ (والتي هي إحدى مُخْرَجات التحليل الصرفي)، وهو ما يمكن إخضاعُه في إطار ما سَلَفَ ذِكْرُهُ من صياغةٍ عامَّةٍ لفك الالتباس اللساني (راجِعْ شكل ٤) والمخصَّصَة هنا لِفَكِّ التباس علامات الضبط الإعرابية بشبكة المَسَارَات (بالطوبولوجية) الموضَّحة في شكل ٨ أدناه والتي يُسْمَحُ فيها فقط بالمسارات على الخطوط المتصلة.



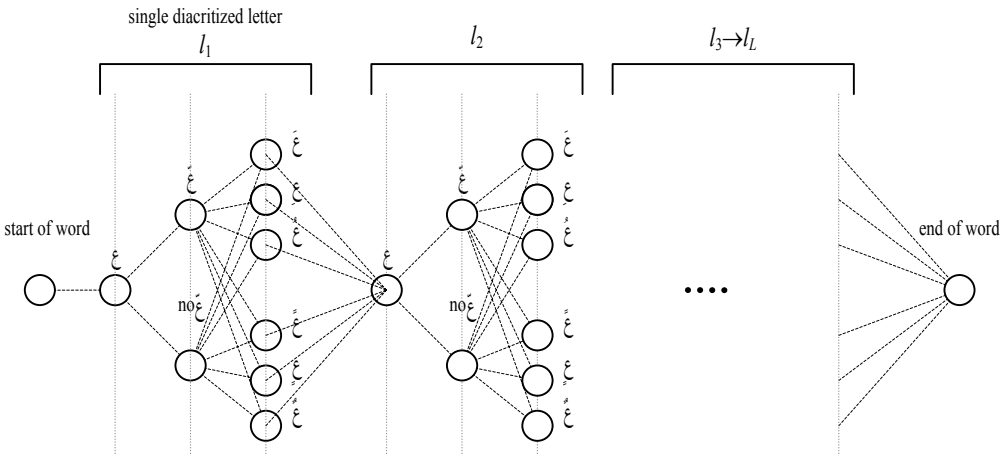
شكل ٨: طوبولوجيَّة فَكِّ التباس التشكيل الإعرابي.

٩- النحو الصوتي العربي، واستنتاج التشكيل الصوتي للكلمات الأجنبية المكتوبة بالحروف العربية [٢٨؛ فصل ٥]، [٣١]

كما أسلفنا في قِسْم ٣-فقرة ح من هذا المقال؛ فإن الكلمات الأجنبية المكتوبة بحروف عربية (مثل؛ شنغهاي، يوجوسلافيا، كسمايو، المكارثية، ...) تتوارد بمعدل حوالي ٧,٥٪ في النصوص العربية الحديثة (خاصَّةً الصحفية/الإخبارية) وهو معدَّل معتبَرٌ لا يمكن تجاهله. وفي الوقت نفسه فإنه من غير المُجْدِي محاوَلةَ حصر تلك الكلمات في جدول؛ فهي يمكن أن تنبع من أية لغة في العالم، ومجموعات الألفاظ محل الاهتمام منها تتغير باستمرار مع تغير الحوادث في الزمن، كما أنه بطبيعة الحال لا توجد طريقة موحَّدة لتهجئتها، ومما يزيد الأمر صعوبةً أنه من المعتاد إضافة اللواصق الصرفية العربية لها (مثل؛ الكلينتونية، الإمبريالي، ...)، وعليه تبرز المعالجة الإحصائية كخيارٍ ذي وجاهة للتعامل مع تلك الصعوبات.

بمعرفة الحروف الهجائية العربية للكلمة الأجنبية المطلوب تشكيلها صوتيًّا، وبمعرفة أن كل حرف l_i يمكن أن يكون مشدَّدًا أو لا، وكذلك تحتمل علامة ضبطه أن تكون إحدى علامات التشكيل الصوتي المعروفة (فتحة، كسرة، ...)، فإنه يمكِنُنا صياغة مسألة تشكيل مثل تلك الكلمات في إطار الصياغة

العامَّة لفك الالتباس اللساني (راجِعْ شكل ٤) والمخصَّصَة هنا بشبكة المَسَارَات (بالطوبولوجية) الموضَّحة في شكل ٨ أدناه والتي يُسْمَحُ فيها بكافة المسارات.



شكل ٩: طوبولوجيَّة تشكيل الكلمات الأجنبية المكتوبة بحروف عربية.

وبالرغم من أن الكلمات الأجنبية المكتوبة بحروف عربية لا توجد عليها قيود لسانية عربية سواء كانت صرفية أو نحوية، فإن القيد الفونولوجي الممثل بطبقة الصوتيات في أدنى السُّلَّم الاسترشادي للطبقات اللسانية (انظر شكل ٢ في قسم ٣ من هذا المقال) يظل باقيًا، خاصَّةً إذا كان المستهدف هو بناء عمل نظام آلي لتخليق الكلام العربي المنطوق وذلك لأن الخروج على هذا القيد الفونولوجي يؤدي إلى انهيار ذلك النظام أثناء مرحلة التقسيم الصوتي المقطعي [٢٧]، [٣٤].

وفضلاً عن ذلك، فإن تطبيق ذلك القيد الفونولوجي يؤدي إلى تهدئة الالتباس الإحصائي لتلك المسألة باستبعاد المسارات غير المطابقة فونولوجيًّا أثناء عَمَل خُوارِزْم A^* (انظر قِسْم ٦ من هذا المقال).

ولتنفيذ ذلك التقييد الفونولوجي فإنه قد تم مراجعة ما أقرَّه اللغويون الكلاسيكيون عن الفونولوجيا العربية ذات الصلة بهذه المسألة [٥]، [٨]، [١٠] ومن ثَمَّ إعادة صياغته تشومسكيًّا في هيئة نَحْو عِيارِيّ محْكَم يمكن تنفيذه حاسوبيًّا بسهولة بحيث يمكن الحكم على أي تتابع صوتي (فونيمي) عربي مقترَح بالموافقة أو عدم الموافقة للقيود الفونولوجية العربية. وهذا النحو الذي أطْلَقْنَا عليه اسم النحو الصوتي العربي مبيَّن في صيغة BNF الرياضية (مع تعليق توضيحي بالعربية) في جدول ٦ أدناه[6] (على الصفحة التالية).

[6] في الحقيقة هذا هو أحد جزأين من النحو الصوتي العربي وهو الجزء الداخلي (الحاكم للتتابعات الفونيمية داخل حدود الكلمة الواحدة)، أما الجزء الآخر وهو الجزء البَيْنِيّ (بين الكلمات المتتابعة) فيُناقَش في القِسْم التالي (قِسْم ١٠) من هذا المقال.

$\mathbf{w} := \mathbf{y}_{start}[\mathbf{y}_{mid}\#][\mathbf{y}_{end}]$

أية كلمة مكوَّنةٌ من ثلاثة مقاطع صوتية؛ البادئ، والأوسط (اختياري)، والخاتم (اختياري)

$\mathbf{y}_{start} := \mathbf{c}_{start}\ \mathbf{f}_{vowel}$

المقطع البادئ مكون من صامت يليه صائت

$\mathbf{y}_{mid} := \mathbf{y}_{mid,regular}|\mathbf{y}_{mid,sokoon}|\mathbf{y}_{mid,silent}$

المقطع الأوسط هو إما مقطع أوسط منتظم، أو مسكَّن، أو غير منطوق

$\mathbf{y}_{end} := \mathbf{y}_{end,sokoon}|\mathbf{y}_{end,silent}|\mathbf{y}_{end,layyina}|\mathbf{y}_{end,tanween}$

المقطع الخاتم مكون من مقطع خاتم ساكن، أو غير منطوق، أو ذي ألِف لَيِّنةٍ، أو منوَّن

$\mathbf{y}_{mid,regular} := \mathbf{c}_{mid}[SHADDA]\mathbf{f}_{vowel}$

المقطع الأوسط المنتظم مكون من صامت أوسط ثم شدَّة (اختيارية) ثم صائت

$\mathbf{y}_{mid,sokoon} := \mathbf{c}_{mid}\ SOKOON\ \mathbf{c}_{mid}\ \mathbf{f}_{vowel}$

المقطع الأوسط المسكَّن مكون من صامت أوسط ساكِن ثم صامت أوسط ثم صائت

$\mathbf{y}_{mid,silent} := \mathbf{c}_{mid}\ BYPASS$

المقطع الأوسط غير المنطوق مكون من صامت أوسط غير منطوق

$\mathbf{y}_{end,sokoon} := (\mathbf{c}_{end}\ SOKOON)|(\mathbf{c}_{mid}\ SOKOON\ \mathbf{c}_{end}\ SOKOON)|(\mathbf{c}_{mid}\ SHADDA\ SOKOON)$

المقطع الخاتم المسكَّن مكون من صامت خاتم ساكِن،أو صامت أوسط ساكن فصامت خاتم ساكن، أو صامت أوسط ساكن مشدَّد

$\mathbf{y}_{end,silent} := \mathbf{c}_{mid}\ (SOKOON|\mathbf{f}_{vowel}|\mathbf{f}_{tanween}|(SHADDA\ \mathbf{f}_{tanween}))\ \mathbf{c}_{end}\ BYPASS$

المقطع الخاتم غير المنطوق مكون من صامت خاتم ساكِن أو صائت أو منوَّن أو مشدَّد بالتنوين، ثم صامت خاتم غير منطوق

$\mathbf{y}_{end,layyina} := \mathbf{c}_{mid}[SHADDA]\mathbf{f}_{layyina}$

المقطع الخاتم ذي الألِف اللينة مكون من صامت أوسط، ثم شَدَّة (اختيارية)،ثم حركة ألِف لَيِّنةٍ

$\mathbf{y}_{end,tanween} := \mathbf{c}_{end}[SHADDA]\mathbf{f}_{tanween}$

المقطع الخاتم المنوَّن مكون من صامت خاتم، ثم شَدَّة (اختيارية)،ثم تنوين

$\mathbf{c}_{start} := (HMZA|BAA|TAA|...|HA|WAW|YAA)|(ALIF|HMZe)$

الصامت البادئ هو أحد الحروف الأبجدية (من ألف إلى ياء) أو ألف المد أو الألف ذات الهمزة السفلى

$\mathbf{c}_{mid} := (\mathbf{c}_{start} - \{ALIF,HMZe\})|(HMZs|HMZy|HMZw)$

الصامت الأوسط هو صامت بادئ (عدا ألف المد والألف ذات الهمزة السفلى) أو همزة على السطر أو على ياء أو على واو

$\mathbf{c}_{end} := \mathbf{c}_{mid}|Yend|TAAM$

الصامت الخاتم هو صامت أوسط أو ياء حقيقية أو تاء مربوطة

$\mathbf{f}_{vowel} := (FATEHA[ALIF\ VWL])|(KASRA[YAA\ VWL])|(DHAMMA[WAW\ VWL])$

الصائت هو فتحة (قد يليها مد بالألف)، أو كسرة (قد يليها مد بالياء)، أو ضمة (قد يليها مد بالواو)

$\mathbf{f}_{layyina} := FATEHA\ YAA\ YAAL$

حركة الألف اللينة هي فتحة تليها ياء غير منقوطة تليها علامة الألف اللينة

$\mathbf{f}_{tanween} := TNWa|TNWo|TNWe$

حركة التنوين هي إما تنوين بالفتح أو تنوين بالكسر أو تنوين بالضم

جدول ٦: النحو الصوتي العربي (الداخلي).

ويبين الجدول ٧ أدناه أمثلة على التشكيلات الصوتية الناتجة من تطبيق تلك الطريقة على بعض الكلمات الأجنبية المكتوبة بحروف عربية مع حكم السامعين على جودة الكلام المخلَّق وفقًا لها.

الكلمة المراد تشكيلها	ناتج التشكيل (بالكتابة الصوتية العربية)	جودة التشكيل
الديمقراطيون	اَدِّيمُقْرَاطِيُّونْ	بلا أخطاء (ممتاز)
ماساشوستس	مَاسَاشُوسِتْسْ	بلا أخطاء (ممتاز)
بوليفارد	بُلِفَارْدْ	بلا أخطاء (ممتاز)
تويوتا	تُيُتَّا	بلا أخطاء (ممتاز)
فالكروموسومات	فَالْكَرُّومَسُومَاتْ	٤ أخطاء – غير مفهوم (رَدِيء)
للبلياردو	لُلْبِلِيَارْدُ	خطأ واحد (جيد جدًّا)
التراجيدية	اَتِّرَاجِيدِيَّهْ	بلا أخطاء (ممتاز)
انطونيو	اَنْطُنْيُو	بلا أخطاء (ممتاز)
شنغهاى	شِنْغَهَايْ	بلا أخطاء (ممتاز)
رونالدو	رُنَّادُّو	ثلاثة أخطاء – مفهوم بالكاد (مقبول)

جدول ٧: أمثلة على نواتج التشكيل الصوتي الإحصائي المقيد بالنحو الصوتي الداخلي، لبعض الكلمات الأجنبية المكتوبة بالعربية.

١٠- الربط الصوتي بين الكلمات المتتابعة للقراءة المتصلة [٢٨؛ فصل ٦]

من المعلوم أن نهاية أية كلمة عربية وبداية الكلمة التالية لها مباشَرةً تتفاعلان بتأثيرات صوتية متبادَلةً عند نطق النص متَّصِلاً (دون فواصل أو سَكَتاتٍ) وهو ما ندعوه بالرَّبْط الصوتي العربي، وكمثالٍ على هذا يوضح جدول ٨ أدناه (على الصفحة التالية) فقرةً من نص عربي تم تشكيل كل كلمة فيها تشكيلاً صوتيًّا كاملاً، ثم يبيِّن تأثُّر هذا التشكيل عند اتصال الكلمات في عبارات متَّصِلة صوتيًّا.

ولتنفيذ ذلك الربط يُسْتَخْدَمُ الجزءُ البَيْنِيُّ من النحو الصوتي العربي، وقد تمت صِياغتُهُ رياضيًّا بإحكام على هيئة خريطة التدفُّق المبيَّنة في شكل ١٠ (على الصفحة بعد التالية) والتي تؤدِّي إلى تطبيق واحدة من تسعة قواعد حَصْرِيّة مشروطة. ويمكن قراءة هذا الجزء البيني من النحو الصوتي إذا اعتبرنا أن الكلمتين المتتاليتين محل البحث ذواتَيْ تهجئة $\{s_{1,1}s_{1,2}...s_{1,L_1-1}s_{1,L_1}\}$ و $\{s_{2,1}s_{2,2}...s_{2,L_2-1}s_{2,L_2}\}$، وحالات تضعيف (تشديد) $\{h_{1,1}h_{1,2}...h_{1,L_1-1}h_{1,L_1}\}$ و $\{h_{2,1}h_{2,2}...h_{2,L_2-1}h_{2,L_2}\}$، وعلامات ضبط صوتيّ $\{v_{1,1}v_{1,2}...v_{1,L_1-1}v_{1,L_1}\}$ و $\{v_{2,1}v_{2,2}...v_{2,L_2-1}v_{2,L_2}\}$.

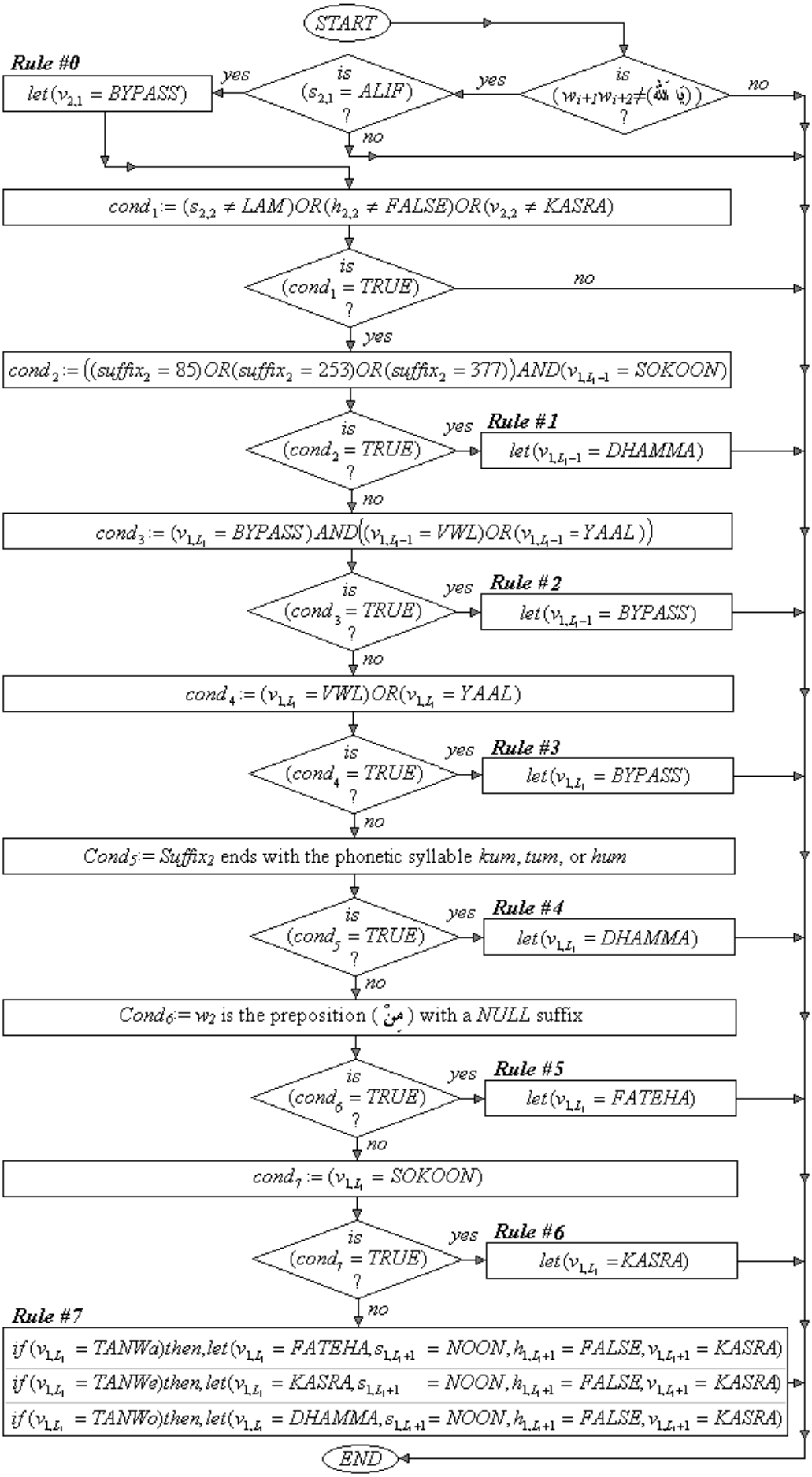
نَصٌّ عربي مشكَّل صوتيًّا بالكامل قبل ربْطه صوتيًّا	
أَيُّهَـا اَلـسَّادَةُ اَلْكِـرَامُ/مِـنْ اَلْمُـسَلَّمِ بِهِ أَنَّ اِسْتِثْمَارَاتِكُمْ اَلضَّخْمَةَ فِي اَلْمَشْرُوعَاتِ اَلْكُبْرَى اَلَّتِـي اِسْـتَطَعْتُمْ اَلـتَّقَدُّمَ مِـنْ خِلَالِهَا إِلَى أَسْوَاقِهِمْ اَلْوَاسِعَةِ قَدْ فَتَحَتْ اَلْبَابَ وَاسِعًا اَلْآنَ نَحْوَ اَلنُّمُوِّ اَلِاقْتِصَادِيِّ اَلْمَأْمُولِ/	1
نَصٌّ عربي مشكَّل صوتيًّا بالكامل بعد إتمام ربْطه الصَّوْتيّ	
أَيُّهَ_سَّادَتُ_لْكِرَامْ/مِنَ_لْمُسَلَّمِ بِهِ أَنَّ_سْتِثْمَارَاتِكُمُ_ضَّخْمَتَ فِ_لْمَشْرُوعَاتِ_لْكُبْرَ_لَّتِ_سْتَطَعْتُمُ_تَّقَدُّمَ مِنْ خِلَالِهَا إِلَا أَسْوَاقِهِمُ_لْوَاسِعَتِ قَدْ فَتَحَتِ_لْبَابَ وَاسِعَنِ_لْآنَ نَحْوَ_نُّمُوِّ_لِقْتِصَادِيِّ_لْمَأْمُولْ/	2
نَصٌّ عربي مشكَّل صوتيًّا بالكامل بعد إتمام ربْطه الصَّوْتيّ مع بيان رقم قاعدة الربط الصوتي عند موضع تطبيقها	
أَيُّهَ_*Rule#3*_سَّادَتُ_*Rule#0*_لْكِرَامْ/ مِنَ_*Rule#5*_لْمُسَلَّمِ بِهِ أَنَّ_*Rule#0*_سْتِثْمَارَاتِكُمُ_*Rule#4*_ضَّخْمَتَ فِ_*Rule#3*_لْمَشْرُوعَاتِ_*Rule#0*_لْكُبْرَ_*Rule#3*_لَّتِ_*Rule#3*_سْتَطَعْتُمُ _*Rule#4*_تَّقَدُّمَ مِنْ خِلَالِهَا إِلَا أَسْوَاقِهِمُ_*Rule#4*_لْوَاسِعَتِ قَدْ فَتَحَتِ _*Rule#6*_لْبَابَ وَاسِعَنِ_*Rule#7*_لْآنَ نَحْوَ_*Rule#0* _نُّمُوِّ_*Rule#0*_لِقْتِصَادِيِّ_*Rule#0* _لْمَأْمُولْ/	**3**

جدول ٨: أمثلة على نواتج التشكيل الصوتي الإحصائي المقيد بالنحو الصوتي الداخلي.

فالقاعدة رقم ٣ – *Rule#3* – في خريطة التدفق في شكل ١٠ على سبيل المثال تؤدي إلى:

((تجـاوُز نـُطْـق آخـر حـرف مـن الكلمة الأولى إذا كانت علامة ضبطه الصوتي مَدًّا (بالألف أو الواو أو الـياء) أو ألفًـا لَيِّـنَـةً، بشرط أن لا تكون الكلمتان المتتابعتان (يا الله) وأن يكون الحرف الأول في الكلمة الثانية أَلِف وَصْل وألا يكون الثاني لامًا مكسورةً غير مشدَّدة وألا تنطبق أيٌّ من القاعدتين ١ ولا ٢))

وهكـذا لبقية القواعد السبعة في الجزء البيني، المكمِّل للجزء الداخلي من النحو الصوتي العربي الذي تم عرضه في القِسْم ٩ السابق من هذا المَقَال.



شكل ١٠ : خريطة التدفُّق المعبِّرة عن الجُزْء البَيْنِـيّ من النحو الصوتي العربي.

١١- تفصيح النص العربي الخام [٢٨؛ فصل ٦]

كـثيرًا مـا تحتوي النصوص العربية المعاصرة على العديد من التعبيرات العددية، والاختصارات الـشهيرة، وصـيغ الـتواريخ والـوقت، ...، وما شابَهَ، والتي ينبغي تحويلها إلى تعبيرات نصية كلامية بالكامـل قـبل الولـوج إلى عملية التشكيل الآلي، ونسمِّي هذه العملية "تفصيح[7] النص" والجدول ٩ أدناه يعرض مثالاً لهذا التفصيح المطلوب.

النـَّصّ الخامُ قبل تفصيحه	النـَّصّ الخامُ بعد تفصيحه
كمـا قـام **أ.د.**محـسن رشوان يوم السبت الموافق **٢٥-١٢-٢٠٠٥م** بمناقـشة **م**/محمـد عطـية في رسـالته للدكـتوراه، وموضـوعها **Automatic Arabic Phonetic Transcription** في جلـسة استمرت حوالي **٥٠** دقيقة من **١٢ م** حتى **١٢:٥٠ م**	كمـا قـام **اَلْأُسْـتَاذ اَلدُّكْـتُور** <pause> محسن رشـوان يـوم الـسبت الموافـق **اَلْخَامِس وَالْعِشْرُونَ** <pause> **اَلـشَّهْر الثَّـانِـي عَشَر** <pause> **عَامَ أَلْفَيْنِ وَخَمْسَة مِيلَادِيَّة** <pause> بمناقشة **اَلْمُهَـنْدِس** <pause> محمـد عطـية في رسالته للدكـــتوراه <pause> **Automatic Arabic Phonetic Transcription** <pause> وموضـوعها <pause> في جلـسة اسـتمرت حوالـي **خَمْـسُونَ** دقـيقة مـن **اَلـسَّاعَة الثَّـانِـيَة عَشَرة مَسَاءً** <pause> حتى **اَلسَّاعَة الثَّـانِـــيَة عَـــشَرَة وَخَمْـــسُونَ دَقِـــيقَة مَـــسَاءً** <pause>

جدول ٩: مثال على تفصيح النص العربي.

وتـتم عملـية تفـصيح النص من خلال نَـحْو عِيارِيّ محْكَم (بنفس المفهوم الخاص بالنحو الصوتي العربي الداخلـي في شكل ٦ السالف إيراده في هذا المقال) يتم تنفيذه حاسوبيًّا ويؤدي إلى اكتشاف الموضع الذي يـتطلب التفـصيح ويستبدله بـ/يُخَلِّق له التعبير النصي المناسب. ويوضح جدول ١٠ أدْناهُ (على الصفحة التالية) الهيكل العام لهذا النحو العياري في صيغة BNF الرياضية.

[7] يعود اقتراح هذه التسمية كتعريب لمصطلح Text Normalization إلى الزميل العزيز م/جلال خلَّاف جلال.

$\mathbf{T} := \mathbf{T}_{Arabic_Other} | \mathbf{T}_{Other_Arabic}$

$\mathbf{T}_{Arabic_Other} := \mathbf{P}_{Arabic}[\mathbf{P}_{Other}\ \mathbf{P}_{Arabic}]\#[\mathbf{P}_{Other}]$
$\mathbf{T}_{Other_Arabic} := \mathbf{P}_{Other}\ [\mathbf{P}_{Arabic}\ \mathbf{P}_{Other}]\#[\mathbf{P}_{Arabic}]$

$\mathbf{P}_{Arabic} := (\mathbf{u}_{Arabic}|\mathbf{d})\#$
$\mathbf{P}_{Other} := \mathbf{c}_{Other}[\mathbf{c}_{Other}|\mathbf{s}]\# \rightarrow$ *Stream out*

$\mathbf{u}_{Arabic} := \mathbf{f}_{Arabic}|\mathbf{w}_{Arabic}$
$\mathbf{d} := \mathbf{s}\#$

$\mathbf{w}_{Arabic} := \mathbf{c}_{Arabic}\#$
$\mathbf{f}_{Arabic} := \mathbf{f}_1|\mathbf{f}_2|\ldots|\mathbf{f}_N$

$\mathbf{f}_1$:= (/أ)|(.أ)|(أ) → (اَلْأُسْتَاذ)

$\mathbf{f}_2$:= (/م)|(.م)|(م) → (اَلْمُهَنْدِس)

$\mathbf{f}_3$:= (ذ.م.م)|(ذ م م) → (ذَات مَسْؤُولِيَّةٍ مَحْدُودَةٍ)

.
.
and all the other common Arabic acronyms

$\mathbf{f}_{N_{Acronyms}+1} := \mathbf{n}\# \rightarrow$ *verbalized integral number*
$\mathbf{f}_{N_{Acronyms}+2} := [\mathbf{f}_{N_{Acronyms}+1}].\mathbf{f}_{N_{Acronyms}+1} \rightarrow$ *verbalized floating point number*
.
.
and all the other common Arabic numeral formats

$\mathbf{c}_{Arabic}$ *is any Arabic alphabetical character.*
$\mathbf{c}_{Other}$ *is any non-Arabic alphabetical character.*
n *is any numeral character.*
s *is any non-alphabetical and non-numeral character.*

جدول ١٠: الهيكل العامّ للنحو العِياري المستخدَم لتفصيح النص العربي.

١٢- تقويم أداء نظام التشكيل الآلي ومكوِّناتِهِ [٢٨؛ فصل ٨]

من بين مكوِّنات نظام التشكيل الصوتي الآلي النص العربي الذي عرضناه، فإنه بالتزاوج مع القواعد القَطْعيّة الحَصْريّة توجد ثلاثة مُكوِّنات تحتوي على عنصر معالَجة إحصائية وتحتاج إلى قياس دِقَّة أدائها، ألا وهي؛ عَمَليّة فك الالتباس الصرفي (ويُرْمَز لدقتها بـ A_L)، وعَمَليّة استنتاج علامة الضبط الإعرابي (ويُرْمَز لدقتها بـ A_S)، وعَمَليّة استنتاج التشكيل الصوتي للكلمات الأجنبية المكتوبة بالعربية (ويُرْمَز لدقتها بـ A_T).

وإذا سَمَّيْنا معدل ورود الكلمات الأجنبية المكتوبة بحروف عربية f_T، وسَمَّيْنا نسبة الكلمات العربية المحتاجة إلى علامة ضبط إعرابي f_S ، فإنه يمكن ربط الأداء الكُلِّيّ للنظام بهذه الكميات بالصيغة:

$$A_{Strict} = (1 - f_T) \cdot A_L \cdot \left((1 - f_S) + f_S \cdot A_S\right) + f_T \cdot A_T$$

وإذا مـا اعْتَبَرْنا ما قَرَّره المَقَال سابقًا أن $f_T=0.075$ ، وأن $f_S=0.65$ ، فإن هذه الصيغة تؤول في التقويم المتشدِّد حيث يُحاسَب على أخطاء التشكيل الإعرابي إلى:

$$A_{Strict} = 0.32375 \cdot A_L + 0.60125 \cdot A_L \cdot A_S + 0.07500 \cdot A_T$$

وتؤول في التقويم المتساهِل حيث لا يُحاسَب على أخطاء التشكيل الإعرابي إلى:

$$A_{Lenient} = 0.925 \cdot A_L + 0.07500 \cdot A_T$$

ويعـرض جـدول ١١ أَدْنـاهُ لخلاصـة قياساتنا لأداء تلك المكونات الثلاث تحت آخر ما لدينا من ظروف تدريب إشرافي إحصائي:

الكمية المقيسة	حجم عينة التدريب الإشرافي	حجم عينة الاختبار	طول الأفق الارتباطي N	دقة الأداء	ملاحظات
A_L	٠,٥ مليون كلمة ٢ مليون مورفيم	١٠ آلاف كلمة ٤٠ ألف مورفيم	١٢ مورفيمًا	٪٩٦	–
A_S	٠,٤ مليون كلمة ٦,٤ مليون وحدة	١٠ آلاف كلمة ٤٠ ألف وحدة	١٦ وحدة (نوع كلامي/علامة تشكيل)	٪٨٠	–
A_T	٢٠ ألف كلمة ٣٠٠ ألف فونيم	ألف كلمة ١٥ ألف فونيم	١٥ فونيمًا	٪٩٤	تُعَدُّ خطأً الكلماتُ التي يحكم السامعون بعدم فهم تشكيلها أو فهمه بالكاد.
A_{Strict}	تقويم الأداء الكلي المتشدِّد			٪٨٤,٣	
$A_{Lenient}$	تقويم الأداء الكلي المتساهِل			٪٩٥,٨٥	

جدول ١١: تقويم أداء المكونات الإحصائية في نظام التشكيل الآلي للنص العربي.

١٣- قائمة المراجع:

أولاً؛ المراجع بالعربية:

[١] دِيـوانُ الأَدَبِ، أَوَّلُ مُعْجَـمٍ عَرَبِـيٍّ مُـرَتَّبٍ حَـسْبَ الصِّيَغِ الصَّرْفِيَّةِ، أَبُو إبْراهيمَ إسْحاقَ ابْنُ إبراهيمَ الفارابيّ، تحقيقُ د/أَحْمَد مختار عُمَر، ١٩٧٤م.

[٢] مَقاييسُ اللُّغَةِ، أبُو الحُسَيْنِ أحْمَدُ ابْنُ فارِسٍ ابْنُ زَكَرِيَّا، تحقيقُ أ/عَبْد السَّلَامِ محمَّد هارون، الطَّبْعَةُ الثَّانِيَةُ، ١٩٦٩م.

[٣] لِسانُ العَرَبِ، أبُو الفَضْلِ جَمَالُ الدِّين محمَّدٌ ابْنُ منظور، دارُ فاضِل، بَيْروت.

[٤] مُعْجَمُ مَتْنِ اللُّغَةِ، أَحْمَد رِضَا، مَكْتَبَةُ الحَيَاةِ، بيْروت، ١٩٦٠م.

[٥] دِراسةُ الصَّوْتِ اللُّغَوِيّ، د/أَحْمَد مُخْتار عُمَر، عالَم الكُتُبِ، مِصْر، ١٩٩٠.

[٦] مُعْجَمُ قَواعِدِ العَرَبيَّةِ العالَمِيَّةِ، أنْطُوانُ الدَّحْداح، الطَّبْعة الأُولَى، ١٩٩٠م، مَكْتَبة لُبْنان.

[٧] القاموسُ الطَّويلُ لِلْقُرْآنِ الكَرِيمِ، إبراهيمُ أَحْمَد عَبْد الفَتَّاح، مَجْمَعُ البُحوثِ الإسلاميَّةِ، ١٩٨٣م.

[٨] الأصوات اللُّغَويَّة، د/إبراهيم أَنِيس، مَكْتَبة الأَنْجِلُو المِصْرِيَّة–القاهرة، ١٩٧١م.

[٩] فِـي النَّصِّ الأَدَبيِّ، دِراسةٌ أُسْلوبيَّةٌ إحصائيَّةٌ، د/سَعْد مصلوح، النَّادِي الأَدَبيُّ الثَّقَافيُّ جِدَّة، الطَّبْعةُ الأُولَى، ١٩٩١م.

[١٠] فونولوجـيا العربية، د/سَلْمَان حَسَن العانِي، ترجمة ياسِر المَلَّاح، دار النَّادي الأَدَبيِّ بجِدَّة–المَمْلَكة العَرَبيَّة السَّعوديَّة، ١٩٨٣م.

[١١] الحُـقولُ الدَّلَالِيَّةُ الصَّرْفيَّةُ لِلأَفْعالِ العَرَبيَّةِ، سُلَيْمان فَيَّاض، دارُ المَرِّيخِ بالرِّياضِ، ١٩٩٠م.

[١٢] التَّطبيقُ الصَّرْفِيُّ، د/عَبْدُهُ الرَّاجِحيّ، دارُ المَعْرِفةِ الجامِعِيَّةِ، الإسْكَنْدريَّةُ، ١٩٩٣م.

[١٣] الخِطاطة (الكِتابة العَرَبيَّة)، عَبْد العَزيز الدَّالِي، مَكتبة الخانْجِيّ، مِصْر، ١٩٩٨م.

[١٤] رَسْمُ المُصْحَفِ؛ دِراسةٌ لُغَوِيَّةٌ تاريخيَّةٌ، د/غانِم قَدُّوريّ الحَمَد، الطَّبْعةُ الأُولَى، بَغْداد، ١٩٨٢.

[١٥] نَـشْـأة وتَطـوُّر الكِـتابة الخَطِّـيَّة العَـرَبيَّة ودَوْرُهـا الثَّقافيّ والاجتماعيّ، فَوْزيّ سالِم عَفيفيّ، الطَّبْعة الأُولَى، وِكالةُ المطبوعات، الكُوَيْت.

[١٦] المُعْجَمُ الوَسيطُ، مَجْمَعُ اللُّغَةِ العَرَبيَّةِ بالقاهِرةِ، الطَبْعةُ الثَّالِثَةُ، ١٩٨٥م.

[١٧] مُعْجَمُ أَلْفاظِ القُـرْآنِ الكَرِيمِ، مَجْمَعُ اللُّغَةِ العَرَبيَّةِ بالقاهِرةِ، طَبْعةٌ مُنَقَّحةٌ، ١٩٩٠م.

[١٨] مـا تَـمَّ مِـنْ مـشروعِ المُعْجَمِ الكَبيرِ، مَجْمَعُ اللُّغَةِ العَرَبيَّةِ بالقاهِرةِ، حَرْفُ الأَلِفِ، الطَّبْعَةُ الأُولَى، ١٩٩٠م، حَرْفُ الباءِ، الطَّبْعَةُ الأُولَى ١٩٩٠م، حَرْفا التَّاءِ والثَّاءِ، الطَّبْعَةُ الأُولَى، ١٩٩٢م.

[١٩] مُخـتارُ الـصِّحاحِ، مُحَمَّدُ ابْنُ أَبي بَكْرٍ ابْنُ عَبْدِ القادِرِ الرَّازِيّ، دارُ الجِيلِ، بَيْروت، طَبْعةٌ حَديثةٌ مُنَقَّحةٌ.

[٢٠] المُغْنِي فِي تصريفِ الأَفْعالِ، مُحمَّد عَبْدُ الخالِقِ عُضَيْمة، دارُ الحَديثِ.

[٢١] تـاجُ العَـروسِ مِـنْ جَواهـرِ القاموسِ، السَّيِّد/مُحمَّدُ مُرْتَضَى الحُسَيْنِيُّ الزَّبيديُّ، طَبْعةُ وَزَارةِ الإعلامِ الكُوَيْتيَّةِ.

[٢٢] فَـنُّ الخَـطِّ العَرَبـيّ؛ مَـوْلِدُهُ وتَطَـوُّرُهُ حَتَّـى العَـصْرِ الحاضِرِ، مُصْطَفَى أَغْوَر دُرْمان، تَرْجَمة؛ صالِح سَعْداوِيّ، الطَّبْعةُ الأُولَى، إستانبول، ١٩٩٠م.

ثانيًا: المراجع بالإنجليزية:

[**23**] Al-Wa'er, M., Toward a Modern Theory of Basic Structures in Arabic, Ed. Arab School of Science and Technology, Damascus, Syria, 1983 .

[**24**] ArabDiac©; an online trial version of the Arabic diacritizer described in this thesis; RDI 's ArabDiac© is found at: http://www.RDI-eg.com under the sub menu item *Arabic NLP* under the main menu item *Lang Tech*, 2004.

[**25**] ArabMorpho©; an online trial version of an Arabic Lexical Analyzer; RDI 's ArabMorpho© upon which the Arabic diacritizer described in this thesis; RDI 's ArabDiac© is built, can be reached at: http://www.RDI-eg.com under the sub menu item *Arabic NLP* under the main menu item *Lang Tech*, 2000.

[**26**] ArabTagger©; an online trial version of an Arabic POS Tagger; RDI 's ArabTagger© upon which the Arabic diacritizer described in this thesis; RDI 's ArabDiac© is built, can be reached at: http://www.RDI-eg.com under the sub menu item *Arabic NLP* under the main menu item *Lang Tech*, 2004.

[**27**] ArabTalk©; an online trial version of an Arabic Text-To-Speech system; RDI 's ArabTalk© which is based on the Arabic diacritizer described in this thesis; RDI 's ArabDiac© is found at: http://www.RDI-eg.com under the sub menu item *Speech* under the main menu item *Lang Tech*, 2004.

[**28**] Attia, M., 2005, Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications. PhD thesis, Department of Electronics and Electrical Communications, Cairo University. http://www.RDI-eg.com/RDI/Technologies/paper.htm

[**29**] Attia, M., Arabic Orthography vs. Arabic OCR, Multilingual Computing & Technology magazine www.Multilingual.com, 2004.

[**30**] Attia, M., A Large-Scale Computational Processor of The Arabic Morphology, and Applications, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000. This thesis is also downloadable from the following web pages: http://www.NEMLAR.org/ScientificPapers/Index.htm, and http://www.RDI-eg.com/RDI/Technologies/paper.htm.

[**31**] Attia, M., Rashwan, M., Khallaaf, G., A Formalism of Arabic Phonetic Grammar and Application on the Automatic Arabic Phonetic Transcription of Transliterated Words, NEMLAR int'l conference in Cairo, Sept. 2004. This paper is downloadable from http://www.RDI-eg.com/RDI/Technologies/paper.htm.

[**32**] Attia, M., Rashwan, M., A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, NEMLAR int'l conference in Cairo, Sept. 2004. This paper is downloadable from http://www.RDI-eg.com/RDI/Technologies/paper.htm.

[**33**] Attia, M., Rashwan, M., Khallaaf, G., On Stochastic Models, Statistical Disambiguation, and Applications on Arabic NLP Problems, The Proceedings of the 3rd Conference on Language Engineering; CLE'2002, the Egyptian Society of Language Engineering (ESLE). This paper is downloadable from http://www.NEMLAR.org/ScientificPapers/Index.htm and http://www.RDI-eg.com/RDI/Technologies/paper.htm.

[**34**] Dutoit, T., An Introduction to Text-To-Speech Synthesis, Kluwer Academic Publishers, 1996.

[**35**] Fatehy, N., An Integrated Morphological and Syntactic Arabic Language Processor Based on a Novel Lexicon Search Technique, master thesis, Faculty of Engineering, Cairo University, 1995.

[**36**] Grosz, B.J., Jones, K.S., and Webber, B.L., Readings in Natural Language Processing, Morgan Kauffman publishers, 1986.

[**37**] Hamza, W.M., A Large Database Concatenative Approach for Arabic Speech Synthesis, PhD thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 2000.

[**38**] Hassan, H. M., Maximum Entropy Framework for Natural Language Processing Applied on Arabic Text Classifications, MSc. thesis, Dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 2002.

[**39**] Haung, X., Acero, A., Hon, H.W., Spoken Language Processing, Prentice Hall PTR, 2001.

[**40**] Hifny, Y., Qurany, S., Hamid, S., Rashwan, M., Attia, M., Ragheb, A., Khallaaf, G., ArabTalk©; An Implementation for Arabic Text To Speech System, The proceedings of the 4th Conference on Language Engineering; CLE'2003, the Egyptian Society of Language Engineering (ESLE), and published also in the News Letter of Evaluation of Language Resources and Distribution Agency (ELDA), May 2004 issue. This paper is downloadable from http://www.NEMLAR.org/ScientificPapers/Index.htm and http://www.RDI-eg.com/RDI/Technologies/paper.htm.

[**41**] Jurafsky, D., Martin, J. H., Speech and Language Processing; An Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing, Prentice Hall, 2000 .

[**42**] Kapur, J. N., Saxena, .H.C., Mathematical Statistics, 7th edition, S. Chand & Co. (Pvt.) LTD, 1972.

[**43**] Katz, S. M., Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-35 no. 3, March 1987.

[**44**] Lathi, B.P., Modern Digital and Analog Communication systems, 2nd edition, Holt, Rinehart and Winston Inc., 1989.

[**45**] Nadas, A., On Turing's Formula for Word Probabilities, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-33 no. 6, December 1985.

[**46**] Nilsson, N.J., Problem Solving Methods in Artificial Intelligence, McGraw-Hill, 1971.

[**47**] Ratenaparkhi, A., Maximum Entropy Models for Natural Language Ambiguity Resolutions, PhD thesis in Computer and Information Science, Pennsylvania University, 1998.

[**48**] Rich, E., Knight, K., Artificial Intelligence 2nd edition, McGraw-Hill, 1991.

[**49**] Schütze, H., Manning, C.D., Foundations of Statistical Natural Language Processing, the MIT Press, 2000 .

[**50**] Sproat, R., Multilingual Text-To-Speech Synthesis, Kluwer Academic Publishers, 1998.

[**51**] Van Santen, J. P. H., Sproat, R.W., Olive, J.P., Hirschberg, J., Progress in Speech Synthesis, Springer, 1998.

[**52**] Winston, P.H., Artificial Intelligence 3rd edition, Addison Wesley, 1992.