مُدَوَّنة المُعجَم التّارِيخِيّ للّغة العَرَبِيَّة "المنهج والمُعالَجة"

*المُعتزّ بالله السَّعيد ، **مُحسن عبد الرَّازق رشوان ، *مُحمَّد حسن عبد العزيز

moataz@cu.edu.eg, mrashwan@rdi-eg.com, mh-aziz@hotmail.com

*قسم علم اللُّغة والدِّراسات السَّامِيَّة والشَّرقيَّة، كلية دار العُلوم، جامعة القاهرة

**قسم الاتِّصالات والإلكترونيَّات، كُلِّيَّة الهندسة، جامعة القاهرة

الملخص: موضُوعُ هذه الدِّراسة (مُدَوَّنَةُ المُعجَمِ التَّارِيخِيّ للُّغة العَرَبِيَّة، المنهج والمُعالَجة)، وتهدفُ إلى تقديم منهجٍ لبناء وتطوير مُدَوَّنةٍ لُغَوِيَّةٍ يُستَفاد منها في صناعة مُعجَمٍ تارِيخِيّ للُّغة العَرَبِيَّة. وقد صُمِّمَت المُدَوَّنة اللُّغَوِيَّة – موضُوع الدِّراسة – وفقَ المنهج الوَصفِيّ، إذ تُستَمَدُّ مادَّتُها من نُصُوص اللُّغة العربِيَّة المُشتَرَكة والمُؤَرَّخ لها عبرَ عُصُورِ العربِيَّة. ويتنَوَّع مجالُ البحث بينَ لُغَوِيَّات المُدَوَّنة الَّتي تُعنَى بِالبحث في الظَّواهِر اللُّغوية وتَفسِيرِها من خلال مجموعةٍ مِن النُّصُوص المُحَوسَبَة الَّتي تُمَثِّلُ الواقِعَ اللُّغَوِيّ، وعِلم اللُّغة الإحصائِيّ الَّذي يُعنى بِوَصف البيانات اللُّغَوِيَّة اعتِمادًا على أساليب التَّحلِيل الإحصائِيّ، وعِلم اللُّغة الحاسُوبِيّ الَّذي يُعنى بتَوجِيه الأنظِمة الحاسُوبِيَّة إلى فهم اللُّغات الطَّبِيعِيَّة ومُحاكاة الذَّكاء البَشَرِيّ.

الكلمات المفتاحية: لسانِيَّات المُدَوَّنة، عِلم اللُّغة الحاسُوبِيّ، مُعالَجة المُدَوَّنات اللُّغَوِيَّة، المُعجَم اللُّغَوِيّ التَّارِيخِيّ

The Historical Arabic Dictionary Corpus "Approach & Processing"

*Almoataz B. Al-Said, **Mohsen A. Rashwan, and *Muhammad H. Abdulaziz

moataz@cu.edu.eg, mrashwan@rdi-eg.com, mh-aziz@hotmail.com

*Dept. of Linguistics, Semitic and Oriental studies, Dar Al-Ulum Faculty, Cairo University

**Electronics & Electrical Communications Department, Faculty of Engineering, Cairo University

Abstract: *The subject of this paper is (The Historical Arabic Dictionary Corpus "Approach & Processing"). It is aims to Providing a methodology to processing the corpus text using the linguistic analysis tools, approval to the Arabic language nature and its writing system, as to achieve the purpose for which the corpus designed for it. The Corpus has been designed according to the descriptive approach, as it derives its materials from the common Arabic texts through the ages of Arabic. The research varies between Corpus Linguistics; which interested in research in the phenomena of language and interpretation it through a variety of computerized texts which represent the linguistic reality, Statistical Linguistics; which interested in description of linguistic data based on the methods of statistical analysis and Computational Linguistics; which interested in directing the computer systems to understand the natural languages and simulation of human intelligence.*

Keywords: *Corpus Linguistics, Computational Linguistics, Corpora Processing, Historical Dictionary*.

1. مقدمة

المُعجَم التَّاريخِيُّ مُعجَمٌ لُغَوِيٌّ عامٌّ، يَستَمِدُّ مادَّتَه مِن التُّراث الإنسانِيّ المكتوب، المُدَوَّنِ عبرَ العُصُور والأمكِنة في مُختَلِف العُلُوم والفُنُون والآداب، ويَضُمُّ مُفرَدات اللُّغة، ويُبَيِّن معانِيَها ومُشتَقَّاتِها وأساليبَها، ويُعنى بتأصِيلِها وتاريخِ استعمالِها أو إهمالِها ودِراسَة تَطَوُّر مبانِيها ومعانِيها عبرَ المراحِل الزَّمَنِيَّة المُتَعاقِبة لِلُّغة [١]. وقد كانت هناك العديد من المُحاولات السَّابقة لبناء مُعجَمٍ تاريخيٍّ للعربيَّة، منها: مُحاولة المُستشرق الألماني أوجُست فيشر August Fischer في عام ١٩٠٧، ومُحاولة الجمعيَّة الألمانيَّة للاستشراق في عام ١٩٥٧، ومُحاولة جمعيَّة المُعجَمِيَّة العربيَّة بتونس في عام ١٩٩٠، وأخيرًا.. محاولة هيئة المُعجم التَّاريخيّ بالقاهرة في عام ٢٠٠٤. إلاَّ أنَّ هذه المُحاولات قد توقَّفَت جميعًا، ولعلَّ المُشكلةَ تكمن في عدم اعتماد أيٍّ منها على مُدَوَّنةٍ لُغَوِيَّة ملائِمةٍ لطبيعة العَرَبِيَّة وتاريخِها المديد من ناحية، وعدَم النَّظَر بعَين الاعتِبار إلى اتِّساع المدى الجُغرافِيّ للعَرَبِيَّة من ناحيةٍ أخرى. أضف إلى ذلك ما أحدَثَته ثَقافات الأمم الأخرى في اللُّغة العربيَّة مِن تأثِيرٍ فِكرِيٍّ وحضارِيّ.

٢- منهج بناء مُدَوَّنة المُعجَم التَّارِيخِيّ لِلُّغة العَرَبِيَّة

تَخضَعُ صِناعَة المُدَوَّناتِ اللُّغَوِيَّةِ لِضَوابِطَ ومَعايير تُحَدِّدُها طبيعة اللُّغة والهدَفُ مِن المُدَوَّنة اللُّغَوِيَّة؛ والواقِع أنَّ منهجَ دِراسة المُدَوَّنات اللُّغَوِيَّة المُحوسَبة لا يزال جديدًا على اللُّغةِ العَرَبِيَّة، إذ لم تعرِف الطريقَ إليهِ إلا في الثمانينيَّات من القرن الماضي، مِن خلال أطروحاتٍ عِلمِيَّةٍ ومشروعاتٍ بحثِيَّةٍ مَحدودة، تَنَوَّعَت بين مشروعاتٍ لِمُدَوَّناتٍ أحادِيَّة اللُّغة، ومشروعاتٍ أخرى لِمُدَوَّناتٍ مُتَعَدِّدة اللُّغات. والواقع أنَّ بِناءَ مُدَوَّنةٍ لُغَوِيَّةٍ للمُعجَم التَّاريخِيّ أمرٌ شاقّ، يزيد من صُعوبَتِه طبيعةُ اللُّغة العربِيَّة الاشتِقاقِيَّة ونِظامُها الكِتابِيّ؛ كذلكَ فإنَّ وَضْعَ خُطَّةٍ لِبناء المُعجَم يتطَلَّبُ إحاطةً بحجم التُّراث العربِيّ المُدَوَّن، وإحصاءً لِمُدخلات المُدَوَّنة اللُّغَوِيَّة، وغير ذلك من الأمور الَّتي تُساهِم المُدَوَّنة اللُّغَوِيَّة – إلى حدٍّ كبيرٍ – في تحديدِها على وجه الدِّقَّة. ووفقًا لطبيعة المُعجَم المنشود.. فإنَّ بِناءَ مُدَوَّنةٍ لُغَوِيَّةٍ تُحَقِّقُ الغايةَ الَّتي نَصبُو إليها يتطَلَّبُ أربعَ خطواتٍ منهجِيَّةٍ [١]، نحصُرُها فيما يلي:

1.2 جمع النُّصُوص

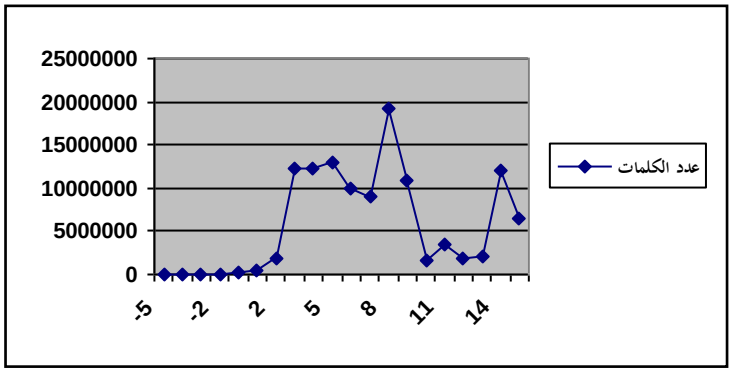
لمَّا كانَ المأمولُ مِن المُعجَم التَّاريخِيّ المنشود أن يكونَ دِيوانًا للعَرَبِيَّة، يُؤَرِّخُ لألفاظِها ومعانِيها، ويُبَيِّن ما طرأ على تلك الألفاظ من تَحوُّلٍ وتغيُّرٍ، كان لابُدَّ لمُدَوَّنة هذا المُعجَم أن تَضُمَّ قدرًا وافِيًا مِن النُّصُوص الَّتي تعكِسُ واقِعَ اللُّغة العَرَبِيَّة عبرَ عُصُورِها التَّاريخِيَّة، وفي بيئاتِها ومراكِزِها الَّتي شَهِدَت مراحِلَ نُمُوِّها وتَطَوُّر ألفاظِها وتراكِيبِها، لتكونَ بمثابة قاعِدَةِ المُعطَيَاتِ الجامِعة لآثارِها الأدَبِيَّة وتُراثِها العِلمِيّ المُدَوَّن. ويَخضَعُ اختِيارُ النُّصُوصِ – وَفقًا لمنهج بناءِ المُدَوَّناتِ اللُّغَوِيَّةِ – لإحدى الطُّرُق التَّالِية:

- **<u>الطَّرِيقةُ الأولى</u>**: تَقُومُ على **<u>الاستِبيان</u>**، حيث يَطرَحُ صُنَّاعُ المُدَوَّنة مجموعةً من الأسئِلة والاستِفسارات على أشخاصٍ يُمَثِّلُونَ المُجتَمَعَ اللُّغَوِيّ الَّذي تنتمي إليه النُّصُوص؛ وتتعلَّقُ الأسئِلةُ والاستِفساراتُ بالحُقُولِ المَعرِفِيَّة وعناوين الكُتُب وأسماء الكُتَّاب والمُصَنِّفين وأوجُه المُفاضَلَة بينها. وفي ضَوء نتائِجِ الاستِبيان يُحَدَّدُ حجمُ المُدَوَّنة اللُّغَوِيَّة والحُقُولُ المَعرِفِيَّة الَّتي تُصَنَّفُ إليها النُّصُوص ومَصادِرُ المادَّة النَّصِّيَّة والفَترَةُ الزَّمَنِيَّة الَّتي تنتمي إليها النُّصُوص والمنطِقةُ الجُغرافِيَّةُ الَّتي ينتَمِي إليها الكُتَّابُ والمُصَنِّفون. وتُستَخدَم هذه الطَّرِيقةُ – عادَةً – في بِناءِ مُدَوَّنات الدِّراسات التَّجرِيبِيَّة؛ كما تُستَخدَمُ في بناءِ المُدَوَّنات اللُّغَوِيَّةِ للمُجتَمَعاتِ الإقلِيمِيَّة ومُدَوَّنات اللَّهجات المُعاصِرة والمُدَوَّنات المَصنوعة لأغراضٍ تعلِيمِيَّة.
- **<u>الطَّرِيقةُ الثَّانِية</u>**: تَقُومُ على **<u>الحَصر الشَّامل</u>**، وتُلزِمُ صُنَّاعَ المُدَوَّنةِ بِحصرِ كُلِّ نُصُوصِ المُجتَمَعِ اللُّغَوِيّ الَّذي تُمَثِّلُه المُدَوَّنة، وتُستَخدَمُ هذه الطَّرِيقةُ – عادَةً – عند بِناءِ مُدَوَّنات الدِّراسات المَسحِيَّة، كالمُدَوَّنات المُستَخدَمة في صِناعَة مُعجَمات الأدباء. ويرى بعضُ الباحِثينَ أنَّ هذه الطَّرِيقة هي الوسيلة المُثلى لبناء مُدَوَّنة المُعجَم التَّارِيخِيّ للعَرَبِيَّة، بل يرى بعضُهُم ضَرورةَ احتواءِ المُدَوَّنةِ على جميع المُؤَلَّفاتِ العَرَبِيَّة الباقِيَةِ وعلى كُلِّ ما جَرى أو لا يزالُ يجري على ألسِنةِ النَّاطِقين بها [٢]. ويبدو هذا رأيًا غريبًا، لا يُعلَمُ حُدُوثُه في أيٍّ مِن المُعجَمات التَّارِيخِيَّة لِلُّغاتِ الأمم الأخرى، فَكيفَ باللُّغَة العَرَبِيَّة الَّتي تملكُ تراثًا مُدَوَّنًا منذ ما يَقرُبُ مِن ألفَيْ عامٍ؟!.

- **الطَّريقةُ الثَّالثة**: وتَقُومُ على **نَظَريَّة العَيّنات الإحصائيَّة** [٣]، ومِن خلالِها يقومُ صُنَّاعُ المُدَوَّنةِ باختيارِ عَيّنةٍ مِن النُّصُوصِ الَّتي تَتَّفِقُ وأهدافَهم البَحثيَّة، سواءٌ أكانَت عَيّنةً عشوائيَّةً أم غيرَ عشوائيَّة. ويَشيعُ استِخدامُ هذه الطَّريقةِ عند بناءِ مُدَوَّناتِ المُعجَماتِ العامَّة، وهِيَ الطَّريقةُ الَّتي تنتهِجُها المُدَوَّنةُ موضُوعُ الدِّراسَة.

وبِناءً على نَظَريَّة العَيّنات الإحصائيَّة، وفي ضَوءِ ما ذُكِرَ حولَ تعريف المُعجَمات اللُّغَويَّةِ التَّاريخيَّة، جُمِعَت نُصُوصُ المُدَوَّنةِ مِن التُّراثِ العَرَبِيِّ المُدَوَّن – المكتوب – فِي صُورةِ عَيّنَةٍ قَصْدِيَّةٍ مُنتَقاةٍ مِن المُصَنَّفاتِ العَرَبيَّة في مُختَلِف المَعارِف بحيثُ تُغَطِّي الفتراتِ الزَّمَنيَّةَ للعَرَبيَّةِ، وبِحيثُ تُغَطِّي الحُدُودَ الجُغرافيَّةَ الَّتي مَثَّلَت المَراكِز الثَّقافيَّةَ والحَضارِيَّة لها. واستَقَرَّت الدِّراسَةُ – في اختيارِ مادَّةِ المُدَوَّنةِ اللُّغَويَّة – على تَصنِيفِ النُّصُوص إلى مَرحَلَتَين:

- **المرحلة الأولى**: تَمتَدُّ منذُ أقدَمِ وثائِقِ المُدَوَّنة في القرن الخامس قبل الهجرة حتى نهاية القَرن الثَّاني الهِجريّ، ويَغلُبُ على نُصُوصِها الانتقالُ بالرِّواية. واشتَمَلَت المُدَوَّنةُ على كُلِّ ما أُتيحَ لها مِن نُصُوصٍ في هذه المَرحلة.
- **المرحلة الأخرى**: وتَمتَدُّ منذ بداية القَرن الثَّالِث الهِجريّ حتَّى أحدَث وثائِق المُدَوَّنة في القَرن الخامس عشر الهِجريّ، ويَغلُبُ على نُصُوصِها الانتقال بالتَّدوين. واقتَصَرَت الدِّراسةُ في اختيار مادَّة المُدَوَّنة لهذه المرحلة على أُمَّهات الكُتُب في مُختَلِف العُلُوم والفُنُون والآداب بما لا يَقِلّ عن مليون كلِمة لكُلِّ قَرنٍ زَمَنِيّ (وهُو الحَدُّ المِعياريّ الأدنى لاعتِبارِ نُصُوصِ القَرنِ الزَّمَنِيّ مُدَوَّنةً لُغَويَّة مُستَقِلَّةً)، مع مُراعاة التَّنَوُّع الجُغرافِيّ والمَوضُوعِيّ للنُّصُوص.



الشكل(١) مُخَطَّط أعداد كلمات المُدَوَّنة اللُّغَويَّة باعتِبار القَرن الزَّمَنِيّ

2.2 تصنيف النُّصُوص

صُنِّفَت مادَّةُ المُدَوَّنةِ اللُّغَويَّة موضُوعِ الدِّراسَةِ تاريخِيًّا وجُغرافِيًّا وموضُوعِيًّا وفقًا لما يَلِي:

- **التَّصنِيف التَّاريخِيّ**: ويشتَمِل على عُصُورِ اللُّغة العَرَبِيَّةِ – وفقَ التَّقسِيم الخُماسِيّ الَّذي أقرَّهُ مجمع اللُّغة العَرَبِيَّة بالقاهرة:
  - **العَصر الجاهِلِيّ.** ويَضُمُّ وثائِقَ تنتمي إلى الفترة من عام ٤٨٠ ق.هـ / ١٥٧ م (وفاة مالِك بن فَهْم الأزدِيّ – صاحِب أقدَم وثائِق المُدَوَّنة) إلى عام ١ ق.هـ / ٦٢٢ م (العام السَّابِق للهِجرة النَّبَوِيَّة).
  - **العَصر الإسلامِيّ.** ويَضُمُّ وثائِقَ تنتمي إلى الفترة من عام ١ هـ / ٦٢٢ م (العامُ الأوَّل للهِجرةِ النَّبَوِيَّة) إلى عام ١٣١ هـ / ٧٤٩ م (العام السَّابِق لسُقُوط دَولة بني أُمَيَّة).
  - **العَصر العَبَّاسِيّ.** ويَضُمُّ وثائِقَ تنتمي إلى الفترة من عام ١٣٢ هـ / ٧٥٠ م (عام سُقُوط دَولة بني أُمَيَّة وقِيام دَولة بَني العَبَّاس) إلى عام ٦٥٥ هـ / ١٢٥٧ م (العام السَّابِق لسُقُوط دَولة بَني العَبَّاس).
  - **العَصر الوَسِيط** "عصر الدُّوَل والدُّوَيلات". ويَضُمُّ وثائِقَ تنتمي إلى الفترة من عام ٦٥٦ هـ / ١٢٥٨ م (عام سُقُوط دَولة بَني العَبَّاس وانقِسام الدَّولة العَرَبِيَّة إلى دُوَلٍ ودُوَيلات) إلى عام ١٢١٩ هـ / ١٨٠٤ م (العام السَّابِق لبداية حُكم مُحَمَّد عليّ باشا – مُؤَسِّس مِصر الحديثة).
  - **العَصر الحدِيث** "عصر النَّهضَة". ويَضُمُّ وثائِقَ تنتمي إلى الفترة من عام ١٢٢٠ هـ / ١٨٠٥ م (تَوَلِّي مُحَمَّد عليّ باشا حُكم مِصر) إلى وقت بناء المُدَوَّنة اللُّغَوِيَّة موضوع الدِّراسة (١٤٣١ هـ / ٢٠١٠ م).
- **التَّصنِيف الجُغرافِيّ**: صُنِّفَت نُصُوصُ المُدَوَّنةِ اللُّغَوِيَّة موضُوعِ الدِّراسَةِ جُغرافِيًّا، بحيث تُمَثِّل المناطِق الَّتي نَشَأَ فيها مُصَنِّفو الوَثائِق، وقُسِّمَت إلى سَبعةِ حُقُولٍ على النَّحو التَّالِي:
  - **شِبه الجزِيرة العَرَبِيَّة.** وتَضُمُّ في العالَم الحديث منطقة الخليج العَرَبِيّ (السعودية والإمارات والكويت وقطر وسلطنة عُمان ومملكة البحرين) وأرض اليَمَن.
  - **بلاد ما بينَ النَّهرَيْن** "الرَّافِدَين". وهي العِراقُ في العالَم الحديث.
  - **أرض وادي النِّيل.** وتَضُمُّ مِن المنطِقة العَرَبِيَّة في العالم الحديث مِصرَ والسُّودان.
  - **بلاد الشَّام.** وتَشمَل سوريَّا ولبنان وفلسطين والأردن والمنطقة الجنوبِيَّة مِن تركيا.

- **شِبه جزيرة الأندلس** "أيبيريا". وتَضُمُّ في العالَم الحديث إسبانيا والبرتغال وجبل طارق.
- **منطقة المَغرِب العَرَبِيّ وصِقِلِّيَة.** وتَضُمُّ في العالَم الحديث تونُس والجزائِر والمَغرِب وليبيا، وتَرتَبِطُ بأواصِرَ تارِيخِيَّةٍ مع جزيرة صِقِلِّيَة الإيطالِيَّة الَّتي تقعُ في جنوب أوروبا.
- **منطقة فارس وما وراء النَّهر** "آسيا الوُسطى". وتَضُمُّ – مِن المناطِق الَّتي انتَشَرَت فيها العَرَبِيَّة – فارِس (وهي إيران حديثًا)، وأقاليم بخارى وسمرقند وخوارزم، والحُدُود الهِندِيَّة.
- **التَّصنِيف الموضُوعِيّ**: وقد صُنِّفَت نُصُوصُ المُدَوَّنةِ اللُّغَوِيَّة موضُوعِيًّا إلى خمسة عشر حقلا، هي: (الشِّعر العربيّ، والقُرآن الكريم، والحديث النَّبَوِيّ، والكتاب المُقَدَّس، والنَّثر الأدَبِيّ، وعُلُوم العربيَّة وآدابها، والنَّثر العلميّ واصطلاحات العُلُوم، والتَّاريخ والأنساب، والطَّبقات والتراجم، والجغرافيا والرحلات والبلدان، والقَوانِين والأحكام، والمِلَل والعَقائِد، والموسوعات والمُعجَمات، والصِّحافة، والمعارِف الأخرى).

3.2 تحرير النُّصُوص

يُقصَدُ بتحرير النُّصُوص تحويلها مِن صُورَتِها الوَرَقِيَّةِ إلى صُورةٍ رَقْمِيَّةٍ مُحَوسَبة يُمكِن التَّعامُلُ معها آلِيًّا؛ وثَمَّةَ ثلاثُ وسائِلَ لتحرير نُصُوصِ المُدَوَّنات اللُّغَوِيَّة، استُخدِمَت جميعُها عندَ بِناء المُدَوَّنةِ اللُّغَوِيَّة موضُوعِ الدِّراسة [٤]:

- **الوَسِيلَةُ الأولى**: **التَّحرير الآلِيّ**. ويَتِمُّ باستِخدام أداتَين:
١. الماسِح الضَّوئِيّ. ويَقُومُ باستِقبال نُصُوص الوَثائِق المُدخَلة بعدَ استِكشافِها ضَوئِيًّا.
٢. القارِئ الآلِيّ. ووظِيفَتُه التَّعَرُّف على الحُرُوفِ المُدخَلة وتَحويلها إلى صِيغةٍ نَصِّيَّةٍ مُحَوسَبَة.
- **الوَسِيلَةُ الثَّانِية**: **التَّحرير اليَدَوِيّ**. هِو الوَسِيلةُ التَّقلِيدِيَّة المُستَخدَمة في رَقمَنة النُّصُوص. ويَتِمُّ مُباشَرَةً باستِخدام مُحرِّرات النُّصُوص.
- **الوَسِيلَةُ الثَّالِثة**: **استِخدام المادَّة المُتاحة إلكترونِيًّا**. وتُوجَد هذه المادَّة في إحدى الصُّوَر التَّالِية:
١. صِيغة الوثيقة المُتَنَقِّلة (PDF).
٢. صِيغة صَفحات الويب (HTML).
٣. صِيَغ الوثيقة النَّصِّيَّة (DOC)، و(RTF)، و(TXT).
وتُعالَجُ هذهِ الصِيَغُ باستِخدامِ مُحرِّرات النُّصُوص.

وقد رُوعِيَ عِندَ اختِيار المادَّة المُتاحة إلكترونِيًّا في نُصُوص المُدَوَّنة اللُّغَوِيَّةِ موضُوع الدِّراسة أن تكونَ خاضِعَةً لرُخصَة جنو للوَثائِق الحُرَّة [٥]، ويُسمَحُ بمُقتَضاها بإعادة نسخ النُّصُوص باعتِبارِها مِلكِيَّةً عامَّةً وفقَ ضَوابِطَ ومعايِيرَ حَدَّدَتها مُؤَسَّسَة البرمجِيَّات الحُرَّة.

4.2 ترميز النُّصُوص

ترمِيز النُّصُوص هُوَ تَحويلُها مِن صُورَتِها الأوَّلِيَّة إلى صُورةٍ مَشْرُوحَةٍ مُفَصَّلَة، أو بمعنى آخر: تحويلُها مِن نُصُوصٍ خامٍ غَيرِ مَشْرُوحَةٍ – وهُو الشَّكل الَّذي تظهَر فيه النُّصُوصُ عِندَ تحريرِها أوَّلِيًّا – إلى نُصُوصٍ مُرَمَّزَةٍ مُذَيَّلةٍ بـ:

- الخَصائِص الشَّكلِيَّة لنُصُوص المُدَوَّنة اللُّغَوِيَّة (وتَشتَمِلُ على: لُغة التَّرمِيز المُستَخدَمة، وصِيغة تحويل الرُّموز / الحُرُوف، وأنواعُ الخُطوط، وأحجام الخُطوط وألوانها).
- والمعلومات البيلوجرافِيَّة للنُّصُوص (وتَشتَمِلُ على: عُنوان الوثِيقة، واسم المُصَنِّف، واسمُ المُحرِّر، وتاريخ التَّحرير، وعدد الكلِمات "بِحِساب لُغة التَّرمِيز").

وتَتَّضِحُ أهمِّيَّةُ تَرمِيزِ المُدَوَّنات اللُّغَوِيَّة للعَرَبِيَّة عندَ التَّعامُلِ معها برمَجِيًّا أو تحليلِ نُصُوصِها باستِخدام أدوات المُعالَجة الآلِيَّة، إذ لا تتعرَّفُ بعضُ لُغاتِ البرمَجة على التَّرمِيزاتِ التَّقلِيدِيَّة للعَرَبِيَّة، الأمر الَّذي يُؤدِّي إلى نُزوح هذه اللُّغات وما بُنِيَ عليها مِن أدوات المُعالَجة الآلِيَّة عن الاستِجابة للحُرُوف العَرَبِيَّة الَّتي تظهَرُ بصُورةٍ مُشَوَّهةٍ تُعِيقُ الآلةَ عن أداء عملها. مِن ناحِيةٍ أخرى يُساعِدُ التَّرمِيزُ فِي التَّعامُل مع المُدَوَّنة اللُّغَوِيَّة والبَرمَجِيَّات المُساندَة لَها عبرَ الشَّابِكة العنكبوتِيَّة من خلال استِدعاء بيانات الويب، كما يُمَكِّن من التَّعامُل المُباشِر مع تِقنِيات التَّنقيب عن البيانات، وما يَتَفَرَّعُ عنها كالبَحث في النُّصُوص والبحث في مُستودعات البيانات.

ويَتِمُّ تَرمِيز النُّصُوصِ عبرَ مَرحَلَتَين رئِسَتَين:

المرحلة الأولى: تشفِير النُّصُوص، وقد التَزَمَت الدِّراسَةُ بتَشفِير نُصُوص المُدَوَّنة اللُّغَوِيَّة بِصِيغة التَّرمِيز 8-Bit UTF-8) (UCS/ Unicode Transformation Format، وهِيَ صِيغة تحوِيل نِظام الحُرُوف الدَّوْلِيّ المُوَحَّد بِقُوَّة (8 Bit).

وتتوافَقُ هذه الصِّيغة مع المعايِير القِياسِيَّة العالَمِيَّة، كما تدعَمُ العدِيدَ من أبجَدِيَّات اللُّغات الطَّبِيعِيَّة، وإن كانَ يَعِيبُها كِبَر المِساحة التَّخزِينِيَّة الَّتي تشغلُها، إذ يتِمُّ من خلالِها تشفير الرَّمز الواحِد في مِساحةٍ تتراوَح بين ١ بايت و ٤ بايت (1 Byte : 4 Byte / 8 Bit : 32 Bit).

المرحلة الأُخْرى: تَوصِيف النُّصُوص، واستُخدِمَت لِذلك لُغة التَّرمِيز القابِلة للامتِداد (XML)، وتُعتَبَرُ لُغَةً تَوصِيفِيَّةً مِثالِيَّةً، إذ تَدعَمُ نِظامَ الحُرُوف الدَّوْلِيّ المُوَحَّد وتَعمَلُ كقاعِدة مُعطَيَاتٍ

مُرَمَّـزة يَسـهُل تناقُلُهـا عبـرَ صَـفحات الوِيـب. واسـتُخدِمَ لتحريـر نُصُـوصِ المُدَوَّنـة اللُّغَوِيَّـة موضُـوع الدِّراسـة بِلُغـة التَّرمِيـز القابِلـة للامتِداد XML بَرنامَجُ التحرير (مفتوح المَصدَر) XML Copy Editor 1.1.0.5.

٣- إحصاء بيانات المُدَوَّنة اللُّغَوِيَّة

1.3 بيانات المُدَوَّنة

- **وفقَ التَّصنِيف التَّارِيخِيّ**

م	العَصر	عَدد الكلمات	%	الوَثائِق	%
1	الجاهِلِيّ	161,502	00.14	100	11.51
2	الإسلامِيّ	650,819	00.56	101	11.62
3	العَبَّاسِيّ	55,150,687	47.37	383	44.07
4	الوَسِيط	40,132,211	34.47	147	16.92
5	الحدِيث	20,319,145	17.45	138	15.88
-	Σ	**116,414,364**	99.99	**869**	100.00

- **وفقَ التَّصنِيف الجُغرافِيّ**

م	المنطقة	عَدد الكلمات	%	الوَثائِق	%
1	الجزيرة	6,169,857	05.30	234	26.93
2	الرَّافدان	31,605,843	27.15	211	24.28
3	الوادي	30,142,049	25.92	121	14.04
4	الشَّام	20,438,704	17.56	123	14.15
5	الأندلس	6,353,048	05.46	58	06.67
6	المغرب	4,074,958	03.47	25	02.76
7	فارس	17,629,905	15.14	97	11.16
-	Σ	**116,414,364**	100.00	**869**	099.99

- **وفقَ التَّصنِيف المَوضُوعِيّ**

م	الحقل	عَدد الكلمات	%	الوَثائِق	%
1	الشِّعر	6,312,207	05.42	370	42.58
2	القرآن	78,119	00.07	1	00.11
3	الحديث	4,654,659	04.00	9	01.04
4	العهدان	436,807	00.38	2	00.23
5	نثر أدبيّ	1,548,818	01.33	36	04.14
6	العربيَّة	30,417,127	26.12	186	21.40
7	نثر علميّ	5,909,418	05.08	37	04.26
8	التَّاريخ	29,100,016	25.00	70	08.06
٩	طبقات	11,986,418	10.30	33	03.79
١٠	جغرافيا	2,466,419	02.12	20	02.30
١١	قوانين	2,586,860	02.22	37	04.26
١٢	عقائد	2,604,845	02.24	22	02.53
١٣	موسوعات	7,591,941	06.52	13	01.50
١٤	صِحافة	8,933,422	07.67	10	01.15
١٥	أخرى	1,787,288	01.54	23	02.65
-	Σ	**116,414,364**	100.01	**869**	100.00

2.3 الوَسَط الحِسابِيّ لبيانات المُدَوَّنة اللُّغَوِيَّة

ويُمكِن التَّعبِير عنه رِياضِيًّا بالمُعادَلة:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

أي أنَّ الوَسَـطَ الحِسـابِيّ μ يُسـاوي مجمـوع قِـيَم عناصِـر المُجتَمع (المُدَوَّنة اللُّغَوِيَّة) X_i مقسُومًا على عدد العَناصِر (المُدَوَّنات الفَرعِيَّة) N. وعليه، يكُون مُتَوَسِّط بيانات المُدَوَّنة كالتَّالِي:

م	التَّصنِيف	مُتَوَسِّط كلمات	مُتَوَسِّط وثائق
1	التَّارِيخِيّ	23,282,872.8	173.8
2	الجُغرافِيّ	16,630,623.4	124.1
3	المَوضُوعِيّ	7,760,957.6	57.9
مُتَوَسِّط عدد الكلمات لِعدد الأعوام "1911" = ٦٠٩١٨ كَلِمَة.			

3.3 نسبة أخطاء التَّحرير في المُدَوَّنة

لِتَعيين نِسبة الخطأ اللُّغَوِيّ في نُصُوص المُدَوَّنة مَوضُوعِ الدِّراسَة، سُحِبَت مِنها عَيِّنة عشوائِيَّة طَبَقِيَّة مُكَوَّنةٌ مِن 116,414 كلمة (0.1 % مِن حجم المُدَوَّنة)، بِنِسبَة 0.1 % مِن كُلِّ حقلٍ موضُوعِيّ، وباستِخدام المُعادَلة:

حيث $\bar{X}_E$ الوَسَط الحِسابيّ للأخطاء، و X_i عدد الأخطاء، و n عدد آلاف العَيِّنة، $\bar{X}_E = \frac{\sum X_i}{n}$ يَكُون متوسط الأخطاء لكل 1,000 كلمة = 2.3. وبتَعمِيم نتائِج العَيِّنة الطَّبَقِيَّة، تَكُون نِسبة أخطاء التَّحرير في المُدَوَّنة اللُّغَوِيَّة = 00.23 %.

وللتَّأكُّد مِن إمكانِيَّة تطبِيق نتائِج العَيِّنَة على مُجتَمَع الدِّراسَة (المُدَوَّنة اللُّغَوِيَّة) نقُومُ – أوَّلاً – بتَعيين المُتَوَسِّط الحِسابِيّ لأحجام حُقُول العَيِّنة المُختارة باستِخدام المُعادَلة:

$$\bar{X}_S = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\sum X_i}{n}$$

حيث $\bar{X}_S$ الوَسَط الحِسابيّ لأحجام حُقُول العَيِّنة، و X_i عدد الكَلِمات، و n عدد الحُقُول، ثُمَّ نقُومُ بتَعيين الانحِراف المِعيارِيّ للعَيِّنة، باستِخدام المُعادَلة:

$$S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_S\right)^2} = \sqrt{\frac{\sum\left(X_i - \bar{X}_S\right)^2}{n-1}}$$

حيث S_D الانحِراف المِعيارِيّ للعَيِّنة الطَّبَقِيَّة، و X_i قِيَم عناصِر العَيِّنة، و $\bar{X}_S$ المُتَوَسِّط الحِسابِيّ لأحجام حُقُول العَيِّنة، و n عدد كلمات العَيِّنة. ثُمَّ نَقُوم بتَعيين الخطأ المِعيارِيّ للعَيِّنة (ويُقصَدُ بِه دَرَجة الاختِلاف بينَ العَيِّنة ومُجتَمَع الدِّراسَة)، وذلك باستِخدام المُعادَلة [٦]:

$$S_E = \sqrt{(N-n)/(N-1)} * (SD \wedge 2/n)$$

حيث S_E الخَطأ المِعيارِيّ للعَيِّنة، و N حجم المُجتَمَع (عدد كلِمات المُدَوَّنة)، و n حجم العَيِّنة (عدد كلِمات العَيِّنة)، و S_D الانحِراف المِعيارِيّ للعَيِّنة؛ فتكون نتيجَتُه:

وعَليه.. يُمكِن التَّقدِيرِيَّة باستِخدام المُعادَلة: $\mathrm{CI} = \bar{X}_E \pm (Z * S_E)$ حِساب فترة الثِّقة

حيث CI فترة الثِّقَة لنِسبَة أخطاء التَّحرير، و $\bar{X}_E$ الوَسَط الحِسابيّ للأخطاء (لكُلِّ ١٠٠٠ كلِمَة)، و Z قِيمَة مِعيارِيَّة ثابِتة (تُساوي 1.96 لمُستَوى الثِّقة ٩٥ %)، و S_E الخَطأ المِعيارِيّ للعَيِّنة. وهُوَ ما يَعني أنَّ: **نسبَة أخطاء التَّحرير في المُدَوَّنة اللُّغَوِيَّة موضُوع الدِّراسَة تتراوَح بين 2.1 إلى 2.5 لكل ١٠٠٠ كلمة، بدَرَجة ثِقة ٩٥ %**.

٤- المُعالَجة الآلِيَّة لنُصُوص المُدَوَّنة اللُّغَوِيَّة

للوُقُوفِ على منهجٍ مُناسِبٍ في المُعالَجَةِ الآلِيَّة لنُصُوصِ المُدَوَّنةِ مَوضُوع الدِّراسَة وتَعيين الآلِيَّات الَّتي تُحَقِّقُ الغايَةَ منها، ينبغي أن نَقِفَ أوَّلا على المَعلُومات الأساسِيَّة الَّتي سيُقَدِّمُها المُعجَم التَّارِيخِيّ المنشود، ثُمَّ نُحَدِّدَ المعلومات الَّتي سَنَستَقِيها مِن المُدَوَّنة اللُّغَوِيَّة للمُعجَم. وفي ضَوءِ مفهُوم المُعجَم التَّارِيخِيّ – المَذكور آنِفًا – يُمكِن القَول: إنَّ المُدَوَّنةَ اللُّغَوِيَّةَ للمُعجَم التَّارِيخِيّ هي مَصدرُ المَداخِل والوحدات المُعجَمِيَّة، وهِي مَصدَرُ المعانِي الوظِيفِيَّة لهذه الوحدات؛ كما أنَّها مَصدر المعانِي المُعجَمِيَّة والشَّواهِد الَّتي تُؤَرِّخ لاستِعمال المُفرَدات والتَّطَوُّر الحادِثِ فيها. وعليه.. أُخضِعَت نُصُوصُ المُدَوَّنة للمُعالَجة الآلِيَّة على مُستَوَيين، هما: المُعالَجة على مُستوى المباني، و المُعالَجة على مُستوى المعاني.

1.4 المُعالَجة على مُستوى المباني

1.1.4 آليَّة فهرسة النُّصُوص

تعتَمِدُ آليَّة فَهرَسة النُّصُوص – في الأساس – على التَّحلِيل التَّصريفِيّ، حيثُ تُعنى المُفَهرِسات الآليَّة بتحلِيل النُّصُوص إلى فِقراتٍ وجُمَلٍ، ثُمَّ تحليل الجُمَلِ إلى كلماتٍ، ثُمَّ تحليل الكَلِمات إلى فُرُوعٍ تُكَوِّنُ الوَحَداتِ المُعجَمِيَّة المُتَفَرِّعَةِ عن المَداخِل. وتَخضَعُ عَمَلِيَّةُ التَّفرِيع لِطَبِيعةِ اللُّغةِ المُفهرَسة، فتَزِيدُ قُدرةُ المُفهرِسات الآليَّة على تَفرِيع الكلِمات عندَ فَهرَسة نُصُوص اللُّغات الإلصاقِيَّة (كالألمانيَّة والإنجلِيزيَّة والفَرَنسِيَّة)، وتَقِلُّ عند فَهرَسة نُصُوص اللُّغات الاشتِقاقِيَّة (كالعَرَبِيَّة والعِبرانِيَّة والحَبَشِيَّة)؛ لهذا.. تُستَخدَمُ أدواتٌ مُساعِدةٌ للمُفهرِساتِ الآليَّةِ عندَ تحلِيل النُّصُوص العَرَبِيَّة كالمُحَلِّلات التَّركِيبِيَّة والتَّصرِيفِيَّة.

وبالإضافَةِ إلى ما تَتَمَتَّعُ به المُفَهرِساتُ الآلِيَّةُ مِن قُدرةٍ على إدارةِ النُّصُوص، وحصر تَرَدُّداتِها، وإعادة تشكِيلِها في قواعِدَ مُعطَياتٍ مُنتَظِمَةٍ، فإنَّها تَقُومُ بِحَصرِ السِّياقات والتَّراكِيب الَّتي تَرِدُ فِيها المُفرَدات، كما تُتِيحُ عددًا مِن خِيارات البَحثِ والتَّرتِيب. واستُخدِمَت في فهرسة نُصُوص المُدَوَّنة آلِيَّتان، هما آليَّة الفهرسة الجذعيَّة، وآليَّة الفهرسة الألفبائِيَّة.

1	وبرز في علم القراءات وعمر مدرسة للقراء وسماها دار القرآن
2	الكريم الوحيدة قبلي المدرسة العصرونية والمسرورية وغيري
3	ابن المقير وجعل عبد الجماعة ودرس بالسمارية وكان صدرًا محترمًا
4	الحمام وبناه سكنًا للشيخ المدرس بها انتهى. وقال الذهبي في
5	وهمذان ودمشق وحران من حلائق ودرس بالقدس الشريف في الصلاحية
6	من معلم القدس انتهى. ثم درس بدمشق في الشامية الجوانية
7	ثلاث عشرة سنة انتهى. ثم درس بالرواحية وهو أول من درس بها
8	ثم درس بالرواحية وهو أول من درس بها واشتغل وأفتى وكانت
9	وبرع فيه وتقدم وأفتى وناظر ودرس وناب عن أبيه في الحكم واشتغل
10	بعد أبيه مدة قليلة ثم عزل ودرس بالعزالية مدة كما سيأتي وناظر
11	وفي جمادى الآخرة منها درس الشيخ شهاب الدين أبو شامة عبد
12	أنه لم يراجع شيئًا حتى أورد درسه ومثله لا يستكثر عليه ذلك

الشكل(٢) نموذج لمادَّة مُفَهرَسة جذعيًّا

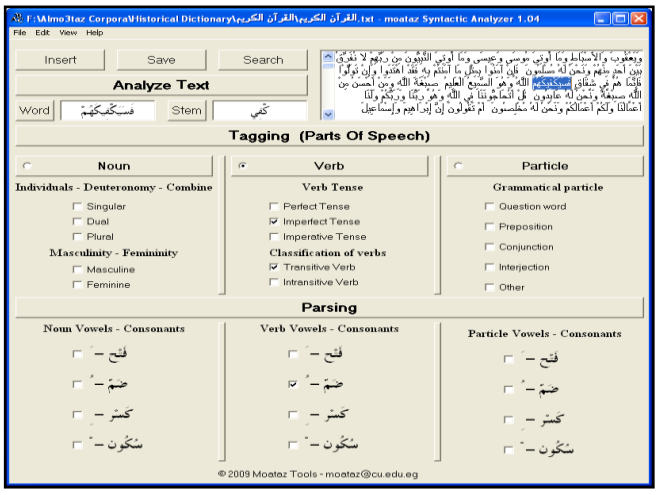
أَمِينًا أَمَامَكَ، وَقَطَعْتَ مَعَهُ	الْعَهْدَ	أَنْ تُعْطِيَهُ أَرْضَ الْكَنْعَانِيِّينَ
الشَّرَائِعِ، عَبَرُوا الْفَرِيضَةَ، وَكَسَّوا	الْعَهْدَ	الأَبَدِيَّ. لِذلِكَ لَعْنَةٌ أَكَلَتْ
آلِهَةً أُخْرَى. وَلاَ تَنْسَوْا	الْعَهْدَ	الَّذِي قَطَعْتُهُ مَعَكُمْ، وَلاَ
وَيَصْغَى إِلَى الَّذِينَ تَرَكُوا	الْعَهْدَ	الْمُقَدَّسَ. وَتَقُومُ مِنْهُ أَذْرُعٌ
آبَاؤُنَا قَطَعَ الرَّبُّ هذَا	الْعَهْدَ	بَلْ مَعَنَا نَحْنُ الَّذِينَ
بَادَ عَابِرُ السَّبِيلِ. نَكَثَ	الْعَهْدَ	رَذَلَ الْمُدُنَ. لَمْ يَعْتَدَّ
حَقِيرَةً وَلاَ تَرْتَفِعِ، لِتَحْفَظْ	الْعَهْدَ	فَتَثْبُتَ. فَمَرَّدَ عَلَيْهِ بِإِرْسَالِهِ
مُحْتَرِفَاتٍ. وَلكِنَّهُمْ كَآدَمَ تَعَدَّوْا	الْعَهْدَ	هُنَاكَ غَدَرُوا بِي. خَلْعًا
يَحْفَظُ لَكَ الرَّبُّ إِلهُكَ	الْعَهْدَ	وَالإِحْسَانَ الَّذِي أَقْسَمَ لآبَائِكَ،
اللهُ، الإِلهُ الأَمِينُ، الْحَافِظُ	الْعَهْدَ	وَالإِحْسَانَ لِلَّذِينَ يُحِبُّونَهُ وَيَحْفَظُونَ

الشكل(٣) نموذج لمادَّة مُفَهرَسة ألفبائِيًّا

2.1.4 آلِيَّة التَّحليل التَّركِيبِيّ

وَفقًا لطَبِيعَةِ المُعجَم التَّارِيخِيّ المَنشُود، يَتَطَلَّبُ تَعيين المَعانِي الوَظِيفِيَّة للوحدات المُعجَمِيَّة – على المُستَوى التَّركِيبِيّ – إخضاعَ نُصُوص المُدَوَّنة للمُعالَجة التَّركِيبِيَّة عبرَ مَرحَلَتَين، يُحَدَّدُ مِن خِلالِهما التَّوصِيفُ النَّحوِيّ (التَّركِيبِيّ) لوحداتِ المُعجَم:

- **المَرحَلَة الأولى: تَعيين أجزاء الكلام**. وتُعنَى بتَعيين أجزاء الكلام ووظائِفِها في الجُملة العَرَبِيَّة، ومِن تَطبِيقاتِها في صِناعَة المُعجَم:
  - **بالنِّسبَة للاسم**: تَعيينُ نَوعِه مِن حيثُ الإفراد والتَّثنِية والجَمع، وتَعيين جِنسِه مِن حيث التَّذكِير والتَّأنِيث.
  - **وبالنِّسبة للفعل**: تَعيين زَمَنِه (ماضٍ، أو مُضارِع، أو طَلَبِيّ)، وَتَعيين نَوعِه (مُتَعَدٍّ ، أو لازِم).
  - **وبالنِّسبة للأداة**: تَعيين أدوات الاستِفهام، وحُرُوف الجَرّ، والعَطف، وأدوات التَّعَجُّب.
- **المَرحَلَة الأُخرى: إعراب أجزاء الكلام**. وتُعنَى بتَعيين حَركات أجزاء الكلام في حالَتَي الإعراب والبِناء، وتَكُونُ على النَّحو التَّالِي:
  - **الاسم**: يأتِي مُعرَبًا (وحالاتُه: الرَّفع، والنَّصب، والخَفض)، ويأتِي مَبنِيًّا (وحالاتُه: بِناء الضَّمّ، وبِناء الفَتْح، وبِناء الكَسر، وبِناء السُّكُون).
  - **الفعل**: يأتِي مُعرَبًا (وحالاتُه: الرَّفع، والنَّصب، والجَزْم)، ويأتِي مَبنِيًّا (وحالاتُه: بِناء الضَّمّ، وبِناء الفَتْح، وبِناء الكَسر، وبِناء السُّكُون).
  - **الأداة**: تأتِي مَبنِيَّةً (وحالاتُها: بِناء الضَّمّ، وبِناء الفَتْح، وبِناء الكَسر، وبِناء السُّكُون).



الشكل(٤) آلِيَّة التَّحلِيل التَّركِيبِيّ Syntactic Analyzer 1.04

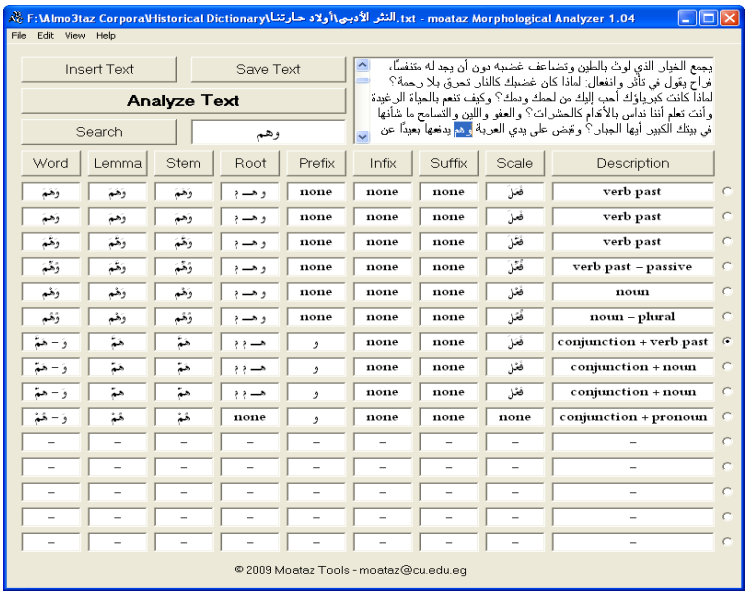
3.1.4 آلِيَّة التَّحليل التَّصريفِيّ

وَحْدَة التَّحليل التَّصريفِيّ هِيَ المُورفِيم الَّذي يُعرَفُ – في اصطِلاح اللُّغَويّين – بأنَّهُ أصغَرُ وحدة لُغَويَّة مُجَرَّدة لها معنى [٧]، أو بِمعنى آخر: هُوَ الصِّيغَة المُجَرَّدة كجُزء مِن النِّظام اللُّغَوِيّ. وتَتَداخَلُ المُورفِيمات نَتيجَةَ التَّألُف بينَ عناصِر الكَلِمَة (الجِذر، والجذع، والفرع، والسَّابقة واللاحِقة) [١].

ووَفقًا لهذِه العَناصِر، تَخضَعُ آلِيَّة التَّحليل التَّصريفِيّ لِثلاثِ إجراءات:

- **التَّجذير**. ويُقصَدُ بِه تَحليل الكَلِمات إلى جُذُور.
- **التَّجذيع**. ويُقصَدُ بِه تَحليل الكَلِمات إلى جُذُوع.
- **التَّفريع**. ويُقصَدُ بِه تَحليل الكَلِمات إلى فُرُوع (وَحدَات المُعجَمِيَّة).

وقد ساعَدَ النِّظامُ التَّصريفِيّ القِياسِيّ للعَرَبِيَّة على تَطوير خَوارزمِيَّة التَّحليل التَّصريفِيّ لَها وتحسين أداء المُحَلِّلات التَّصريفِيَّة لمُفرَدَاتِها، إذ يَقُومُ على مُعطَيَاتٍ ثابِتَةٍ لِقَواعِد التَّصريف العَرَبِيّ؛ ومِن خِلال هذه المُعطَيَات تُبنَى الآلِيَّة لتَكُونَ قادِرةً على تَوليد الأبنِيَة مِن عَناصِرها الأوَّلِيَّة (الجُذُور والجُذُوع والفُرُوع) وتَحليل الأبنِيَة إلى عَناصِرها الأوَّلِيَّة وفقَ ما تَسمَحُ بِه طَبِيعَةُ اللُّغَة العَرَبِيَّة الاشتِقاقِيَّة.



الشكل(٥) آلِيَّة التَّحليل التَّصريفِيّ Morphological Analyzer 1.04

2.4 المُعالَجة على مُستوى المعاني – آلِيَّة التَّحليل الدِّلالِيّ

تُعتَبَرُ آلِيَّةُ التَّحليل الدِّلالِيّ إحدى تطبيقات الذَّكاء الاصطِناعِيّ المُتَقَدِّمة، حيثُ تَرمِي إلى إخضاع الآلة لفهمِ مُفرَدات اللُّغَةِ في جُملَةٍ من السِّياقات والأنماط التَّركيبِيَّة بالاعتِمادِ على مُعطَياتٍ سابِقَةٍ يُحَدِّدُها الهَدَفُ الَّذي يُبنَى لأجلِه المُحَلِّلُ الدِّلالِيّ. ووَفقًا لطبيعَةِ المُدَوَّنة اللُّغَوِيَّةِ مَوضُوع الدِّراسَةِ والهَدَفِ منها، فإنَّنا سنكُونُ بحاجةٍ إلى مُحَلِّلٍ قادِرٍ على فَكِّ اللَّبس الدِّلالِيّ للكَلِمات، أو بعبارةٍ أخرى: سنكُون بحاجَةٍ إلى أداةٍ قادِرةٍ على تَعيين دِلالات المُفرَدات بتحليل سِياقاتِها في هذا القَدر مِن النُّصُوص [٨].

وَتَستَمِدُّ آلِيَّةُ فَكِّ اللَّبس الدِّلالِيّ فِكرَتَها مِن المُتَصاحِبات اللَّفظِيَّةُ الَّتي تَتَكَوَّنُ عن سِلسِلَةٍ مِن كلِمَتَين أو أكثَر، تَتَلازَمُ مُفرَداتُها في عَلاقَةٍ تَركِيبِيَّة [٩]، كَعَلاقَة الفِعل والفاعِل (مِثل: صاحَ الدِّيك، سادَ الصَّمْت) وعَلاقَة الفِعْل والمَفعُول (مِثل: قَدَّمَ استِقالَةً، أحرَزَ هَدَفًا) وعَلاقَة الإضافة (مِثل: عابِر سَبِيل، قاطِع طَرِيق).

لكنَّ الآلِيَّةَ تأخُذُ بُعدًا آخَرَ إذ يَكُونُ التَّلازُمُ على مُستوى السِّياقِ بأكمَلِه؛ فالكَلِمَةُ العَرَبِيَّةُ تَرتَبِطُ في دِلالاتِها بِكلِماتٍ أخرى في حَيِّز الجوار، سواءٌ أكانَت سابِقَةً لها أم لاحِقَةً بِها، وسواءٌ أكانَت مُصاحِبَةً لها أم مُنفَصِلةً عنها.

ويَتَكَرَّرُ التَّصاحُبُ بينَ الكَلِمَة ومُتَصاحِباتِها بصُورةٍ كبيرَةٍ نسبِيًّا، الأمر الَّذي يجعلُنا قادِرين على الرَّبط بين الكلمات الوارِدةِ في مجموعةٍ من السِّياقات، وبالتَّالي.. يَقُودُنا ذلك إلى تَعيين المعانِي المُعجَمِيَّة لمُفرَدات النُّصُوص.

وعليه.. تَقُومُ الفِكرَةُ الأساسِيَّةُ لآلِيَّة فَكِّ اللَّبس الدِّلالِيّ على استِدعاءِ مَعنى الكَلِمَة مُتَعَدِّدَة الدِّلالات من خِلال الكَلِمات المُصاحِبَة لها في السِّياق، إذ يَغلُبُ على هذه الكلِمات أن تَشتَرِكَ مَعَها في حَقلٍ دِلالِيٍّ واحِد [١٠].

وتَقُومُ الخوارزمِيَّات المُستَخدَمَةُ في بِناءِ آلِيَّة فَكِّ اللَّبس الدِّلالِيّ بدَورِ فَعَّالٍ في إحصاء تَرَدُّد الكَلِماتِ المُصاحِبَة للكلِماتِ مُتَعَدِّدَة الدِّلالات في سِياقاتِها لمعرفة مدى استِجابَتها للتَّصاحُب مع كلماتٍ مُعَيَّنَة، ومَعرِفَة نِسبَةِ ورودِها مع مُتَصاحِباتٍ دونَ غَيرِها، ثُمَّ مَعرِفة مدى استِجابة الكَلِمة ومُتَصاحِباتِها للارتِباطِ بِدِلالةٍ مِن الدِّلالات المُعَبِّرة عنها.

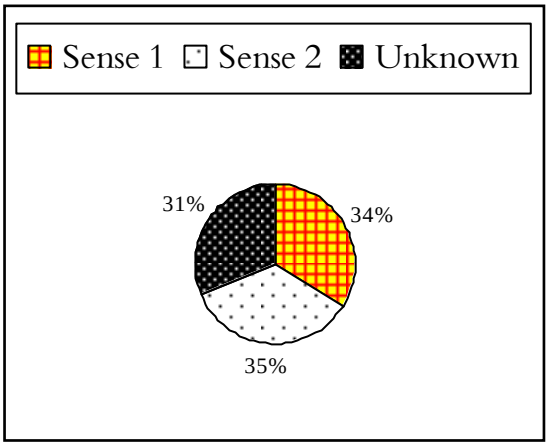
وقد اعتمدت الدِّراسة – في فكّ اللَّبس الدِّلالِيّ لمُفردات المُدَوَّنة – على خوارزميَّة Lesk الَّتي قَدَّمَها المُبرمِج الأمريكِيّ مايكل لِيسك Michael Lesk في عام ١٩٨٦م؛ وتفترض هذه الخوارزميَّة وُجُودَ عَلاقةٍ بينَ الكلِمة ودِلالاتِها مِن ناحِية، ودِلالات الكلِمة وتَعريفاتِها المُفَصَّلة لها من ناحِيةٍ أخرى. وعليه.. تَقُومُ الخوارزمِيَّة على ثلاثة إجراءاتٍ رَئِيسِيَّة:

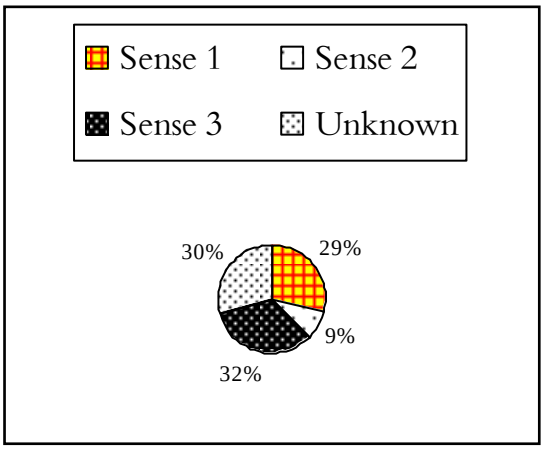
١. إعداد قائِمة بدِلالات الكلِمة.
٢. تَعيين التَّعريفات المُفَصّلة لدِلالات الكَلِمة.
٣. مُطابَقة التَّعريفات بنُصوص المُدَوَّنة اللُّغَويَّة الحاوِية لسياقات الكلِمة.

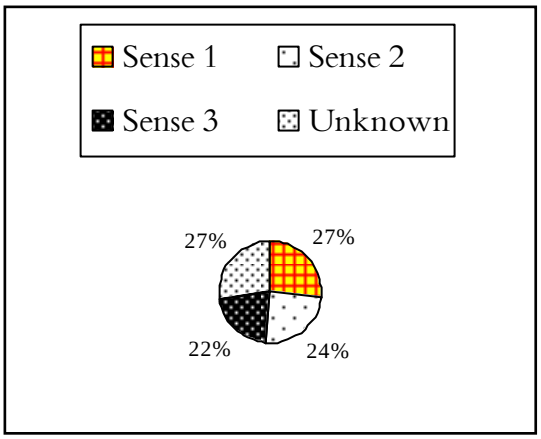
وبتطبيق الخوارزميَّة على اللُّغة العَرَبِيَّة.. يُمكِنُ القَول: إنَّ خَوارِزميَّة فَكّ اللَّبس الدِّلالِيّ للعَرَبِيَّة هي نَتيجَةُ اتِّحادِ السِّياقات المُستَمَدَّة مِن المُدَوَّنة اللُّغَوِيَّة مع العَلاقات الدِّلالِيَّة المُستَمَدَّة مِن شَبَكة الكَلِمات العَرَبِيَّة. وللوُقُوفِ على أهمِّيَّة فَكّ اللَّبس الدِّلالِيّ في اللُّغَة العَرَبِيَّة ودَورِهِ في مُعالَجة نُصُوص المُدَوَّنة اللُّغَوِيَّة للمُعجَم التَّارِيخِيّ المَنشُود في ضَوء خوارزمِيَّة Lesk، قُمنا باختِيارِ ثلاث كَلِماتٍ مُتَعَدِّدَة الدِّلالَة (جدول، حاجب، سَرَطان)، ثُمَّ قاموا باختِيار عَيِّنَةٍ عشوائِيَّةٍ من نُصُوص المُدَوَّنة اللُّغَوِيَّة موضُوع الدِّراسَة تحوي عددًا من السِّياقات المُعَبِّرَةِ عن دلالات الكلمة، بواقِعِ مائَتَي سِياق لكُلِّ دلالة على حِدَة، كما يُوَضِّحُها الجدوَل التَّالِي:

م	الكَلِمَة	S	الدِّلالة (المَعنى المُعجَمِيّ)	No.
١	**جدول**	١	مَجرًى صَغِير للمِياه	٢٠٠
		٢	صَفحَة يُخَطّ فيها خُطُوط مُتَوازِيَة ومُتَقاطِعَة	٢٠٠
٢	**حاجب**	١	الشَّعْر النَّابِت فَوقَ العَيْن	٢٠٠
		٢	خازن الباب وحارِسه	٢٠٠
		٣	عَلَم (من أسماء الذُّكور) / وَظِيفَة	٢٠٠
٣	**سَرَطان**	١	حيوانٌ بحريٌّ من القشريَّات	٢٠٠
		٢	أحد بروج السماء، بين الجوزاء والأسد	٢٠٠
		٣	ورم خبيث يتولد في الخلايا الظاهرية الغُدّية	٢٠٠

وبتَحلِيل سِياقات الكَلِمات الثَّلاثَة على مُستوى الجِذع، كانت النَّتائِج على النَّحو التَّالي:







الشكل(٦) النِّسَب المِئَوِيَّة لِدِلالاتِ العَيِّنَة المُحَلَّلَة دِلالِيًّا – خوارِزمِيَّة Lesk

٥- الخاتمة

وبعد... فقد سَعَت الدِّراسَةُ إلى بِناءِ مُدَوَّنَةٍ لُغَوِيَّةٍ مُحَوسَبَة لِمُعجَمٍ تـارِيخِيٍّ لِلُّغَـة العَرَبِيَّـة، صُنِّفَت نُصُوصُـها تارِيخِيًّـا لتَشـمَلَ عُصُـورَ العَرَبِيَّـةِ الأدَبِيَّـة، وجُغرافِيًّـا لتَشـمَلَ مَراكِزَهـا الثَّقافِيَّـةَ والحَضـارِيَّة، ومَوضُـوعِيًّا لتَشـمَلَ خَمسَـةَ عشـرَ حقـلا مَعرِفِيًّـا؛ وتَوَصَّلَت الدِّراسَةُ – مِن خِلالِ جمعِ النُّصُوصِ وتَصنِيفِها وتَحلِيلِها – إلى عددٍ مِن النَّتائِج على النَّحو التَّالِي:

١. بَلَغَ عَدَدُ كَلِماتِ المُدَوَّنة 116,414,364 كَلِمَة، تَضُمُّها 869 وَثِيقَة، وتَنتَمِي نُصُوصُها إلى الفترة مِن عام ٤٨٠ ق.هـ / ١٥٧ م : ١٤٣١ هـ / ٢٠١٠ م.
٢. اشتَمَلَت المُدَوَّنةُ اللُّغَوِيَّةُ على خمس مُدَوَّناتٍ فَرعِيَّة تُمَثِّلُ العُصُورَ الأَدَبِيَّةَ لِلُّغَة العَرَبِيَّة: العَصر الجاهِلِيّ (161,502 كلمة، بنسبَة 0.14 %)، والعَصر الإسلامِيّ (650,819 كلمة، بنسبَة 0.56 %)، والعَصر العَبَّاسِيّ (55,150,687 كلمة، بنسبَة 47.37 %)، والعَصر الوَسِيط (40,132,211 كلمة، بنسبَة 34.47 %)، والعَصر الحدِيث (20,319,145 كلمة، بنسبَة 17.45 %).
٣. بَلَغَت نِسبَةُ النُّصُوص المَشكولة كُلِّيًّا إلى النُّصُوص المَشكُولَة جُزئِيًّا إلى النُّصُوص غير المَشكُولَة 1 : 1.3 : 7.7 على التَّرتِيب.
٤. بلَغَ عددُ الأعوام الَّتي تُغَطِّيها المُدَوَّنة – وَفقَ التَّارِيخ الهِجرِيّ – 1911 عامًا.
٥. بَلَغَ مُتَوَسِّط عدد الكلمات لِعدد الأعوام الَّتي تُغَطِّيها المُدَوَّنة اللُّغَوِيَّة 60,918 كَلِمَة.
٦. تَراوَحَت نِسبَة أخطاء التَّحرير في المُدَوَّنة اللُّغَوِيَّة بين 2.1 : 2.5 لكل ١٠٠٠ كلمة، بِدَرَجة ثِقة ٩٥ %.
٧. أمكَنَ إخضاعُ نُصُوصِ المُدَوَّنة لآلِيَّة الفهرَسة – الألفبائِيَّة والجذعِيَّة، وأفادَ منها الباحِثُون في تَعيِين المداخِل والوحدات المُعجَمِيَّة للنَّموذَج المُعجَمِيّ.
٨. أمكَنَ إخضاعُ نُصُوصِ المُدَوَّنة لآلِيَّة التَّحلِيل التَّركِيبِيّ عبرَ مَرحَلَتَيْها (تَعيِين أجزاء الكلام، وإعراب أجزاء الكلام)، وأفادَ منها الباحِثُون في تَعيِين المعاني الوظِيفِيَّة – التَّركِيبِيَّة – للوحدات المُعجَمِيَّة.
٩. أمكَنَ إخضاعُ نُصُوصِ المُدَوَّنة لآلِيَّة التَّحلِيل التَّصرِيفِيّ وفقَ الإجراءات الثَّلاثة (التَّجذِير والتَّجذِيع والتَّفرِيع)، وأفادَ منها الباحِثُون في تَعيِين المعاني الوظِيفِيَّة – التَّصرِيفِيَّة – للوحدات المُعجَمِيَّة.
١٠. أمكَنَ إخضاعُ نُصُوصِ المُدَوَّنة لآلِيَّة التَّحلِيل الدِّلالِيّ، وأفادَ منها الباحِثُون في فَكّ اللَّبس الدِّلالِيّ للكلمات؛ وأبانَت الآلِيَّة عن إمكانِيَّة دِراسة التَّطَوُّر اللُّغَوِيّ للكلِمات، كما أبانَت عن مدى تأثُّرِ الكلمة بمتُصاحِباتٍ لُغَوِيَّةٍ – على مُستَوى السِّياق – تتغيَّرُ وتَتَبَدَّلُ وفقًا للحُدُود الزَّمانِيَّة والمَكانِيَّة لِلُّغَة بما يُؤَدِّي إلى دلالاتٍ مُتَنَوِّعة للكلِمةِ محلّ الدِّراسة.

٦- المراجع

[1] السَّعيد (المُعتزّ بالله): مُدَوَّنة مُعجَم تاريخيّ لِلُّغة العربيَّة "مُعالَجة لُغَوِيَّة حاسوبِيَّة"، دكتوراه، جامعة القاهرة، ٢٠١١م.

[2] كورينطي (فيديريكو): دَورُ العامِّيَّة والسَّامِيَّات في المُعجَمِ العَرَبِيّ التَّارِيخِيّ، مَجَلَّة المُعجَمِيَّة، العدد الخامس، وقائع ندوة "المُعجَم العَرَبِيّ التَّارِيخِيّ: قضاياه ووسائِل إنجازه"، تُونُس، ١٩٨٩م.

[3] Sampath, S. (1996). Sampling theory and methods. CRC Press.

[4] السَّعيد (المُعتزّ بالله): مُدَوَّنة مُعجَم عربِيّ مُعاصِر "مُعالَجة لُغَوِيَّة حاسوبِيَّة"، رسالة ماجستير، جامعة القاهرة، ٢٠٠٨م.

[5] GNU Free Documentation License. From: http://www.gnu.org/copyleft/fdl.html.

[6] Kalton, G. (1983). Introduction to survey sampling. SAGE.

[7] Bloomfield, L. (1984). Language. University of Chicago Press.

[8] Armstrong, A. & Armstrong, S. (1999). Natural language processing using very large corpora. Springer.

[9] Firth, J. R. (1957). Papers in linguistics: 1934-1951. Oxford University Press.

[10] Eid, S. M. & Al-Said, A. & Wanas, N. M. & Rashwan, M. A. & Hegazy, N. H. (2010). A Comparative Study of Rocchio Classifier Applied to supervised WSD Using Arabic Lexical Samples. 10th Conference on Language Engineering. Ain Shams University. Cairo.