

التعامل الآلي مع المدونات اللغوية التراثية بين خصائص لغة البرمجة وأخطاء الرقمنة

Automated Processing of Linguistic Heritage Corpus between Characteristics of the Programming Language and Digitization Errors

* أ. رادية حجار

RADIA HADJEBAR

جامعة محمد بوقرة بومرداس، الجزائر.

مخبر الممارسات اللغوية، جامعة مولود معمري تيزي وزو، الجزائر

M'hamed Bougara University Bumerds, Laboratory of linguistic practices, University of Mouloud Mammeri Tizi ousou, Algeria.

radiahadjebar@gmail.com

تاريخ النشر: 2021/06/02

تاريخ القبول: 2021/01/03

تاريخ الإرسال: 2020/11/05

ملخص البحث

إنّ الهدف من رقمنة مدونات اللغة العربية هو التعامل معها آليا؛ وذلك من خلال استرجاع المفردات والمعلومات والتنقيب في البيانات. وسنبيّن في بحثنا هذا، كيفية التعامل مع المدونات اللغوية التراثية الرقمية آليا، والتي اخترنا منها معجم لسان العرب لابن منظور، كتاب الجامع لمفردات الأدوية والأغذية لابن بيطار، ومعجم الإبل للأصمعي. ولكن كيف يتم توصيف المفردات الواردة في هذه المدونات التراثية للحاسوب؟ وهل تسمح لنا لغة البرمجة باسترجاع كل المفردات مهما كان شكلها أو بنيتها؟ وماهي اللوغاريتمات المناسبة للاسترجاع الآلي في حال كانت رقمنة المدونات غير مضبوطة؟ ولأجل الإجابة عن هذه الإشكالات، أنشأنا خوارزميات تتضمن توصيفا دقيقا للمفردة، من حيث حروفها وحركاتها، وكذا للواصق التي يحتمل أن تلتصق بها؛ لتتحقق القراءة لجميع المفردات والتراكيب المراد استرجاعها آليا من قاعدة البيانات. وننتهي من خلال الاسترجاع الآلي لبعض الصيغ والمفردات، من نماذج من المدونات اللغوية التراثية إلى وجوب إعادة النظر في رقمنة هذه المدونات؛ لأجل التعامل معها بلغات البرمجة دون إشكالات.

الكلمات المفتاحية: مدونة لغوية رقمية، لغة برمجة، استرجاع آلي.

* رادية حجار radiahadjebar@gmail.com

Abstract :

The aim of digitizing Arabic language corpuses is to deal with them automatically, as well as to accomplish the historical Arabic lexicon project, by retrieving vocabulary information and data mining. In this research, we will show how to deal with the digital linguistic heritage corpuses automatically, from which we chose the dictionary of « IBIL » of El-Asma'i, the book of «the collection of the vocabulary of medicines and food» of Ibn Baytar, and the dictionary of the « Arabic tongue » by Ibn Mandhour. But how are these vocabulary words contained in the heritage corpuses described in the computer? Does the programming language allow us to retrieve all vocabulary, regardless the shape or structure? What are the appropriate logarithms for automatic retrieval if the digitization of corpus is not accurate or inaccurate?

In order to answer these problems, we have created algorithms and instructions that include an accurate description of the single word in terms of its consonants and vowels, as well as the affixation that are likely to stick to them; so that the reading of all the vocabulary and structures to be retrieved automatically from the database is achieved. We conclude through the automatic retrieval of some formulas and vocabulary, from examples of linguistic heritage corpuses to the need to review the digitization of these corpuses in order to deal with them in programming languages without problems.

Keywords : digital linguistic corpus , programming language, automatic retrieval,

**مقدمة:**

إنّ الهدف من رقمنة مدوّنات اللّغة العربية هو التعامل معها آلياً؛ لإنجاز مشاريع تخدم اللّغة العربية، مثل إنجاز المعجم التاريخي للّغة العربية؛ وذلك من خلال استرجاع المفردات والمعلومات والتنقيب في البيانات. وسنبيّن في بحثنا هذا، كيفية التعامل مع المدوّنات اللّغوية التراثية الرّقمية آلياً، والتي اخترنا منها معجم لسان العرب لابن منظور، وكتاب الجامع لمفردات الأدوية والأغذية لابن بيطار، ومعجم الإبل للأصمعي؛ لأجل استرجاع المفردات التي تدلّ على الحرفة من معجم لسان العرب، والمفردات التي تدل على المرض من كتاب الجامع لمفردات الأدوية والأغذية، والمفردات التي تدلّ على الصّفات من معجم الإبل للأصمعي، ولكن كيف يتم توصيف هذه

المفردات الواردة في المدونات التراثية للحاسوب؟ وهل تسمح لنا لغة البرمجة باسترجاع كل المفردات مهما كان شكلها أو بنيتها؟ وماهي اللوغاريتمات المناسبة للاسترجاع الآلي في حال كانت رقمنة المدونات غير مضبوطة وغير دقيقة؟

وقبل الإجابة عن هذه التساؤلات، لا بأس أن نقدّم مفاهيم المصطلحات ذات العلاقة بالموضوع.

1- المدونة اللغوية: هي مجموعة ضخمة من النصوص اللغوية، تمّ جمعها من مصادر مختلفة ويذكر رشاد الحمزاوي أنها "مجموعة معيّنة من النصوص المكتوبة أو المقولة، أو مجموعة من المراجع المختارة، تؤخذ سندا أسس لوضع لغة ما أو معجم أو مؤلف في موضوع من المواضيع"¹، وتنسيق هذه النصوص وتخزينها في الحاسوب سواء أكانت بصيغة (word) الورد أو بصيغة (pdf) البي دي أف لإتاحة استغلالها للاستشهاد، أو لتدريب برمجيات حوسبة اللغة، والأسهل في التناول والاستخدام هي ملفات الورد؛ لأنها تمكّننا من أخذ المعلومات كما نريدها، وتوفّر خاصية تنسيقها وتهيئتها لبيئة الحوسبة، وهذه التهيئة للمادة اللغوية هي الرقمنة*، وتمثّل هذه الأخيرة أولى خطوات الحوسبة، ولا حوسبة لأي مادة لغوية دون وجود مصادر أو مدونات رقمية متاحة أو مهيأة للحوسبة.

2- المعالجة الآلية: إنّ مفهوم مصطلح المعالجة، يستوجب تقديم مفهوم مصطلحي البرنامج وكذا لغة البرمجة أولاً، فالبرنامج هو مجموعة متتالية من الجمل وظيفتها أداء مهمة، أمّا لغة البرمجة فهي بعض القوانين والرموز والكلمات الخاصة التي تستعمل لكتابة البرنامج². وكتابة أي برنامج يسبقه فهم الإشكال أولاً، ومعرفة كيفية تحويل العمليات الفكرية إلى قوانين، وبعدها اتباع الخطوات اللازمة في كتابة ال (code) أو البرنامج.

وتكون خطوات³ كتابة أي برنامج كما يلي:

- إعطاء مدخلات (input) أو معطيات.

- إعطاء قوانين (process) ومعادلات للمعالجة.

- إعطاء أمر الإعلان عن المخرجات (output) أي النتائج المحصّل عليها وإظهارها

على الشاشة.

3- لغات البرمجة وأصنافها: إنَّ التعرف أو الاسترجاع الآلي للمفردات، والتراكيب من

النصوص الرقمية يكون باعتماد واحدة من لغات البرمجة أو اللغات الحاسوبية المختلفة والتي تنقسم إلى ثلاثة أصناف⁴:

أ- لغات برمجية تحوّل للمستعمل صناعة برامج الحاسوب، تطبيقات الجوال، مواقع إلكترونية وأنظمة تشغيل ... وتمثّل بلغة جافا (java)، سي ++ (C++)، بايثون (python)...

ب- لغات الاستعلام التي تسمح باسترجاع المعطيات ومعالجتها من نظم حاسوبية تحتوي على بيانات، وتمثّل لها بلغة SQL ولغة QUERY ولغة SPARQL.

ج- لغات واصفة تسمح بوصف المحتوى الرقمي وفق قواعد محددة وأشهرها HTML و XML الذي اخترع لتسهيل نقل وتبادل المعلومات بطريقة موحدة مهما اختلفت أنظمة التشغيل، وقد تطورت هذه اللغات الواصفة بشكل مذهل في العقد الأخير، مستفيدة من المنطق الوصفي والذكاء الاصطناعي وأشهرها RDF و RDFs و OWL.

يكون استرجاع المفردات والتراكيب آليا من المدونات الرقمية بتوصيفها للحاسوب^{*}، أو بتبيين اللواصق التي تضاف إلى الجذر، والتغييرات التي تكون على مستوى بعض الحركات، كما أنّ هذا الاسترجاع يكون من المدونات الرقمية فقط؛ بمعنى أن تكون مكتوبة على محرر النصوص الورد (word)؛ لأنه لا يستطيع أي برنامج، تشفير المدونة التي تكون على شكل بي دي ف (pdf) أو التي تم مسحها ضوئيا، وإنما يجب أن تكون مكتوبة على الورد (word) والتي تحوّل بنمط (utf8)⁵ إلى نسخة ثابته مكتوبة على (bloc- notes) أي ذات امتداد (txt) حتى يتمكن مترجم البرنامج من تشفيرها، وهي تمثّل قاعدة البيانات.

ولقد بدأت البلدان العربية بتشكيل فرق خاصة، تقوم برقمنة المدونات اللغوية العربية التي دوّنت في مختلف العصور، والتي تنتمي إلى مختلف الاختصاصات؛ لأجل استغلالها آليا، وإنجاز مشروع الذخيرة اللغوية العربية، الذي تشرف عليه الهيئة العليا التابعة لجامعة الدول العربية، وكذا مشروع المعجم التاريخي للغة العربية، الذي بدأت الدول العربية بإنجازه.

1- التعامل الآلي مع مدونة معجم لسان العرب لابن منظور:

نتعامل آليا في بحثنا هذا مع المدونة اللغوية الرقمية، لسان العرب⁶ لأبي الفضل جمال الدين محمد بن مكرم بن منظور (ت711هـ) وذلك باستخراج أو استخراج الصيغة الصرفية فعالة التي تحمل دلالة الحرفة، ونشير إلى أنّ معجم لسان العرب يمثل المدونة الرقمية الوحيدة التي وجدنا فيها رقمنة مضبوطة ودقيقة، إذ تتوفر على الحركات أو التشكيل، كما نجد فيها غياب علامات الوقف.

ونبيّن في ما يلي الخوارزمية المستخدمة، لأجل الاسترجاع الآلي من قاعدة البيانات (txt). لسان العرب) الموجودة على سطح المكتب، والتي تتوفر على ربط التعليمات بمكتبة التّصوص (re) وكذلك مكتبة خاصة بقراءة كل اللّغات⁷(codecs)

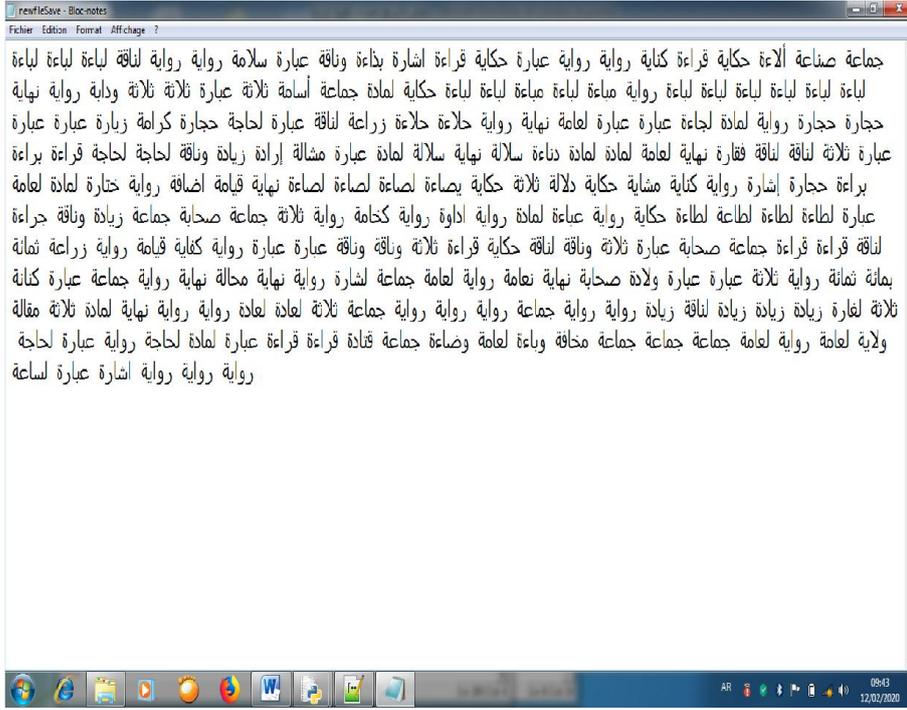
```
import codecs
import re
file =
codecs.open("C:/Users/Client/Desktop/لسانالعرب.txt",
"r",encoding= "utf-8")
fileSave =
codecs.open("D:/newfileSave.txt","w",encoding= "utf-8")
com= re.compile ( r" \w \w \w \w \w \w ")
for ligne in file:
    tab = com.findall(ligne)
    for mot in tab:
        print (mot)
        fileSave.write(mot+" ")
fileSave.close()
```

ونشير إلى أن هذه الخوارزمية تمثل القراءة الآلية للقلب أو الصيغة الصرفية فعالة، الذي تأتي عليه دلالة الحرفة، من المدونة الرقمية لسان العرب لابن منظور، كما هو وارد في التعليمة التي جاءت في السطر السابع من الخوارزمية وهي:

```
com= re.compile( r" \w {2} \w ")
```


- مفردات مضمومة الفاء مثل: أسامة، سلالة...

ونحن نعلم أنّ الحركات في اللّغة العربيّة تمكّننا من تمييز صيغة عن أخرى، وإنّ ترك الشكل في اللّغة العربيّة ليس أصلا من أصولها، ولا ضرورة محتومة، بل ربما كان العكس هو الصّحيح، كما ينطق بذلك الخط العربي⁸ وترك الشكل في المدوّنات الرقمية أدى إلى ظهور صيغ تحمل نفس الشكل، لكن بحركات مختلفة في نتائج التعرف الآلي. وهذا ما تبيّنه هذه الصورة التوضيحية لنتائج الاسترجاع الآلي:



شكل 2: صورة لنتائج الاسترجاع الآلي لصيغة فعالة دون إضافة الحركات

ونقترح بأن يكون الاسترجاع الآلي للصيغة الصرفية فعالة، من المدوّنة الرقمية لسان العرب، أوّلا بإضافة حركة الكسرة في التعليمية، ونحتفظ بالنتائج المتحصّل عليها، وبعدها تغيير التعليمية بحذف حركة الكسرة منها؛ لأجل أن تظهر في النتائج المفردات التي جاءت على القالب فعالة، دون تشكيل أو دون حركات، لنقوم بعد ذلك بتصفية النتائج؛ أي نختار المفردات التي جاءت بالكسر فقط.

ومردّ هذا الإشكال هو أخطاء في الرقمنة؛ أي عدم وجود الحركات في كلّ مفردات المدوّنة، إذ إنّ بعضها جاءت مشكّلة، وبعضها جاءت غير مشكّلة، كما أنّ من خصائص لغة البرمجة أن تسترجع تماما ما ورد في التعليمات لا غير.

2- التعامل الآلي مع كتاب الجامع لمفردات الأدوية والأغذية لابن البيطار:

إنّ التعامل الآلي مع المدوّنة التراثية الجامع لمفردات الأدوية والأغذية، لضياء الدين أبي محمد عبد الله بن أحمد الأندلسي الملقب المعروف بابن البيطار⁹، في هذا البحث يكون باسترجاع المفردات التي تدلّ على مرض، والتي يأتي أغلبها على صيغة (فُعال).
وبتطبيق الخوارزمية السابقة على مدوّنة ابن بيطار، ولكن من أجل استرجاع الصيغة الصرفية فُعال، نغيّر التعليمات لتصبح كما يلي:

com= re.compile(r” \w {2} \ \w ”)

ونتحصّل في النتائج على 7571 مفردة جاءت على القلب فُعال، وبما أنّه لا وجود للحركات في المدوّنة الرّقمية، تظهر:

- كلمات مضمومة الفاء: كراث، غراب، رعاف، ذباب، زجاج، سماق، هوام، سعال

خناق، صداع، نحاس، غبار...

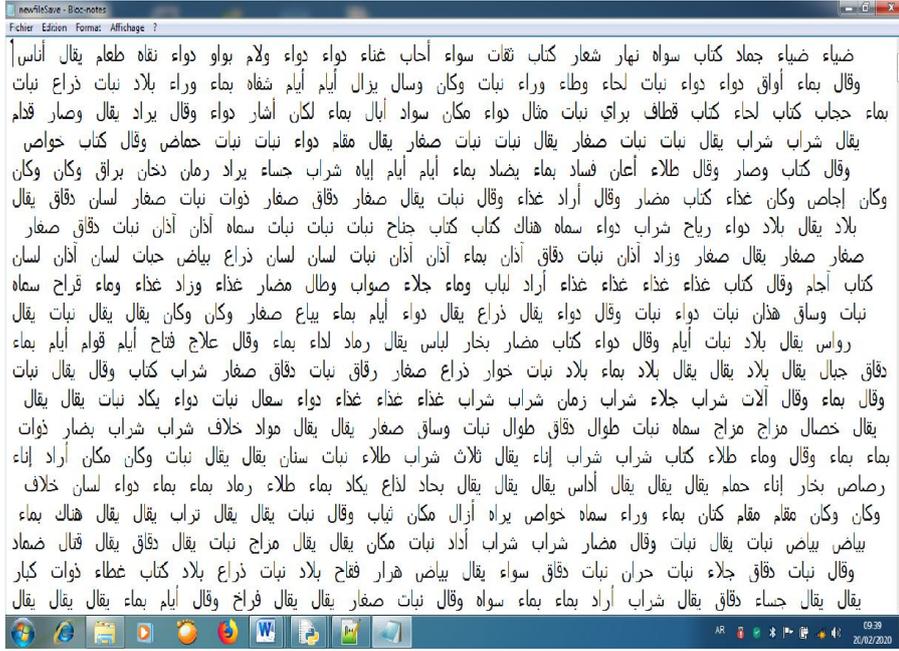
- كلمات مفتوحة الفاء: بيان، دواء، هلاك، نبات، سذاب، حصاة، عباس، بياض،

مذاق رصاص، طعام، هواء، وباء، لهاء، سواد، غزال، حمام...

- كلمات جاءت بكسر الفاء: كتاب، ديار، لسان، جبال، قلاع، مياه، ثياب، دماغ،

طحال إحصاء، غشاء، جماع، عراق، طلاء...

وتوضّح الصورة الموالية هذه النتائج المتحصّل عليها:



شكل 3: صورة لنتائج الاسترجاع الآلي لصيغة فُعال من مدونة ابن البيطار.

ونشير إلى أنّ ورود بعض المفردات التي جاءت على النمط الصرفي فُعال متصلة بلواصق

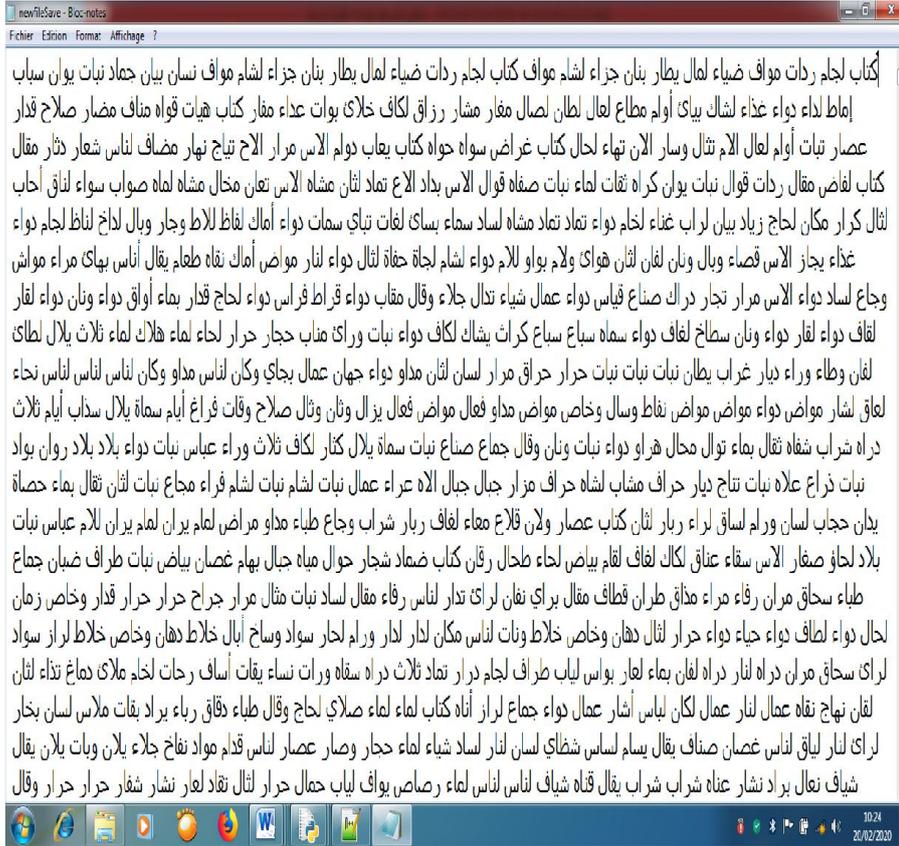
(كالألف واللام، وعلامات الوقف المختلفة) جعلنا نغيّر التعليمات إلى هذا الشكل:

("w" \ {2} \ "w") re.compile("com=

ويعني هذا أنّه سيتم استرجاع الصيغة الصرفية فُعال مهما التصق بها، سواء في بدايتها أو

نحابتها، لتكون نتائج القراءة أو الاسترجاع 61421 مفردة، وهذه صورة من نتائج الاسترجاع

الآلي:



شكل 4: صورة للاسترجاع الآلي لصيغة فُعال المتصلة بلواصق.

ونلاحظ من خلال هذه الصورة ورود مفردات جاءت على صيغة فُعال، إضافة إلى كلمات تشبه صيغة فُعال شكلا فقط، وإنما لا تدلّ على المرض، وأخرى أنصاف كلمات؛ مثل: وقال يقال، أراد، بماء، وكان، وراء، شراب، تعان، الاس، يلان... وهذا يرجع إلى افتقار المدونة الرقمية للحركات.

ونفسر النتائج التي توصلنا إليها من خلال الاسترجاع الآلي لصيغة فُعال من المدونة الرقمية لابن البيطار، إلى أخطاء في الرقمنة؛ كخلو الكلمات من الضبط الحركي (الإعجام) في المدونة وإلى خصائص لغة البرمجة، فالبرنامج يفتقر إلى الكفاءة المعرفية؛ إذ يتعرف على الكلمات

داخل النص، من خلال الفراغات الموجودة بينها، كما أنه يظهر في النتائج ما هو مطابق تماما لما وارد في التعليمات، وهذا ما يجعلنا نقوم بتصفية النتائج بعد كل قراءة آلية، والذي يتطلب جهدا ووقتا كبيرين، ولتختيل معي القارئ مدى الإحباط الذي يصاب به الباحث، حين ظهور تلك النتائج التي تُجره على إعادة تصنيفها يدويا. ورغم توفر البرمجة الآلية لتشكيل النصوص تلقائيا (مثل برنامج مشكال الذي يعتمد في تشكيل النصوص المراد إعرابها آليا) إلا أن التشكيل اليدوي أو إضافة الحركات من خلال لوحة المفاتيح أثناء الرقمنة هو الحل الوحيد لهذا الإشكال.

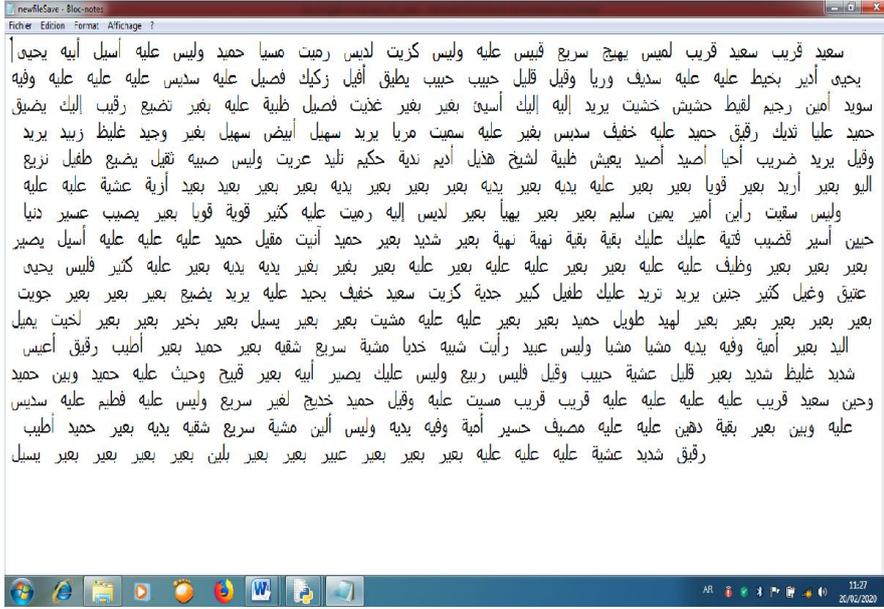
3- التعامل الآلي مع معجم الإبل للأصمعي:

اخترنا مدونة لغوية تراثية أخرى للتعامل معها آليا، وهي كتاب الإبل¹⁰ للأصمعي أبو سعيد عبد الملك بن قريب بن علي بن أصمع (ت 216هـ) لأجل استخراج المفردات أو الصيغ التي تدل الصفات، والتي تأتي أغلبها على وزن فاعل. وتتحقق القراءة الآلية للصيغة الصرفية فاعل بالخوارزمية نفسها، مع تغيير على مستوى التعليمات التي تصبح بهذا الشكل:

com= re.compile(r" \w {2} ي \w ")

ونتحصل من القراءة الآلية للصيغة فاعل دون لواصل على 337 مفردة، كما هو مبين في

هذه الصورة التالية:



شكل 5: صورة للاسترجاع الآلي لصيغة فعيل من مدونة معجم الإبل للأصمعي.

ونظرا لورود بعض المفردات التي جاءت على صيغة فعيل ملتصقة بلواحق، نحقق إمكانية

قراءتها بتغيير التعليمات كما يلي:

com= re.compile(r''\w {2} \w'')

وتكون نتائج القراءة أو الاسترجاع الآلي: 1608 مفردة كما توضح هذه الصورة:

خلصنا من خلال بحثنا إلى أنّ اللغة العربية مرنة وقابلة للتعامل معها حاسوبيا، إلا أنّ تحقيق التعامل الآلي مع المدوّنات اللّغوية الرقمية، يحتاج إلى:

- وجوب إعادة النّظر في رقميتها؛ لأنّه مهما يكن البرنامج المعتمد، فإن من خصائصه أنه يُظهر في نتائج البحث تماما ما يطابق التعليمات التي قدّمت له، وعليه يجب أن تُكتب بطريقة دقيقة ومضبوطة.

- ضرورة تكثيف الجهود المبذولة في مجالات الأبحاث والتطوير لتقنيات المعلوماتية من خلال تفعيل آليات التعاون بين علماء اللغة، وكافة المتخصصين في مجالات المعلوماتية؛ من أجل تطوير هذه التقنيات للتعامل الكفؤ والفعال مع اللغة العربية.

- وإنّ الأخذ بهذه المقترحات في رقمنة المدوّنات هو تيسير للتعامل معها آليا، وتحقيق لمشروع الذخيرة اللّغوية، وكذا مشروع المعجم التاريخي، كما هو مخطط له، وأفضل مشروع جاهز للتعامل معه آليا ويحتذى به، مدوّنة براون (Brown) الإنجليزية التي تمثل مكنزا، ومدوّنة ضخمة تمّ رقميتها بطريقة صحيحة ومنظمة وجاهزة، للتعامل معها آليا.

هوامش:

- ¹ - عبد الكريم مجاهد مرداوي، مناهج التأليف عند العرب، معاجم المعاني والمفردات، ط1. عمان: 2010، دار الثقافة، ص11.
- * الرقمنة هي إتاحة المادة اللغوية وهيئتها على وسيط إلكتروني عوضا عن الوسيط الورقي، أمّا الحوسبة فهي تصميم خوارزميات حاسوبية وبرامج للتعامل مع المادة اللغوية المرقمنة، ومعالجتها آليا.
- ² - محمد إبراهيم الدسوقي. «your first program in c++, variables, data types» عن موقع www.youtube.com بتاريخ: الأربعاء 3 جوان 2015.
- ³ - محمد إبراهيم الدسوقي، الموقع نفسه، التاريخ نفسه.
- ⁴ - طارق المالكي، أنطولوجيا حاسوبية للنحو العربي، نحو توصيف منطقي ولساني حديث، ط1. المغرب: 2015، دار التابعة للنشر والتوزيع، ص6.
- * يوجد فرق بين الوصف والتصنيف، والأصل في ذلك هو متلقي العمل، فإذا كان المتلقي هو الإنسان نسمي العملية بالوصف، أمّا إذا كان الحاسوب هو الذي سيتلقى العمل فالمصطلح المناسب لهذه العملية هو التوصيف. طارق المالكي أنطولوجية حاسوبية للنحو العربي، ص8.

⁵ -Gérard swinnen, apprendre à programmer avec python, Paris: 2009, Editions Eyrolles 61, bd saint-germain 75240, p37.

⁶ - أبو الفضل جمال الدين محمد بن مكرم بن منظور، لسان العرب، ط1. بيروت: دت، دار صادر.

⁷ -Tarek ziadé, programmation python conception et optimisation, 2e édition. Paris:2009, Editions Eyrolles 61, bd saint-germain 75240, p105

⁷ - نبيل علي، العرب وعصر المعلومات، دط. الكويت: 1994، المجلس الوطني للثقافة والفنون والآداب، ص137.

⁸ - ابن بيطار، كتاب الجامع لمفردات الأدوية والأغذية، دط. لبنان: 2001، دار الكتب العلمية.

⁹ - الأصمعي أبو سعيد عبد الملك بن قريب بن علي بن أصمغ، الإبل، ط1. دمشق: 2003، دار البشائر، تح: حاتم صالح الضامن.