دور المكون الصرفي في بناء المعجم الإلكتروني للغة العربية

"قاعدة بيانات المصادر"

محمد الحناش و رشيدة محمد القاسمي

جامعة عجمان للعلوم والتكنولوجيا

دولة الإمارات العربية المتحدة

E-mail: elhannach@yahoo.com

ملخص

سيركز هذا البحث على دور الوزن والميزان الصرفيين في بناء المعجم الإلكتروني للغة العربية، وكيف أن إغفال هذا المكون أو سوء تطبيقه أدى إلى ضعف الكثير من البرمجيات العربية، خاصة في بناء المحلل والمولد الصرفيين، وذلك على الرغم من الجهود الكبيرة التي أنفقت في حوسبة اللغة العربية. لهذا السبب لا نجد اليوم برامج حاسوبية عربية قادرة على إنجاز مهمة الترجمة الآلية بنسبة معقولة، أو قريبة من المعقولة، كما لا نجد برامج القراءة البصرية للمحارف العربية التي تفي بالغرض بشكل مريح، وغير هذه البرامج كثير، وما ذلك – في نظرنا – إلا لتقليل هذه البرمجيات من دور المكون الصرفي للغة الضاد.

كما سيركز على الطبيعة الانصهارية للغة العربية، خلافا للغات الأوروبية التي توصف بكونها لغات إلصاقية، وسترصد انتظام هذا المكون في تشكل الكلمة توليدا وتحليلا، وقدرته على توزيع المورفيمات الصرفية داخل بنية الجذور العربية لتجعل منها مداخل معجمية تصبح قابلة للتصنيف الدلالي، والتأليف التركيبي، والتنوع التداولي.

كما سنقوم بتعريف مفصل لمفهوم المعجم الإلكتروني وتطبيقاته في اللغات الطبيعية، مع التركيز على الخصائص الصرفية (الخوارزمية) للغة العربية من وجهة نظر الهندسة اللسانية، وإبراز أهمية هذا المكون وضرورته لبناء المحلل التركيبي الآلي للغة العربية، إذ تبين لنا أنه يصعب التقدم في بناء باقي المحللات الآلية للغة العربية قبل الانتهاء من توصيف المكون الصرفي توصيفا هندسيا لسانيا.

سنقترح نموذجا لسانيا حاسوبيا نطبقه على قاعدة بيانات المصادر العربية (**250.000** مدخل صرفي مشفر)، لمعالجة هذا الوضع انطلاقا من المقاربة النظرية التي تجعل من الوزن الصرفي البنية العميقة للكلمة العربية، بينما تجعل من الميزان الصرفي بنيتها السطحية.

سنختم البحث بتحليل إحصاءات جميع المداخل الصرفية المولدة صرفيا في اللغة العربية، وتشمل نحو **2.500.000** مدخل معجمي مولد من مجموع الجذور العربية المستخدمة فعليا (**12.000** جذر)، بنحو **800** صيغة صرفية، موزعة على أقسام الكلم التالية: أسماء الذوات، أفعال، مصادر، مشتقات.

مدخل تعريفي:

الهدف من هذا البحث هو التعريف بالإمكانات التي يقدمها الحاسوب لمعالجة اللغة العربية بوصفها لغة طبيعية، كما يتعرض للطريقة التي يجب اتباعها لإدماج اللغة العربية في مكونات التقانة الحديثة، بهدف إقامة الحوار بين الإنسان والآلة بلغة الضاد. وقد تبين أنه كلما تطورت تقانات الحاسوب ازداد قربا من محاكاة دماغ الإنسان في طريقة عمله، مما يسمح للغة بالقيام بدور كبير في تطور هذه التقانة، وعندما نقول معالجة اللغة في هذا السياق فإننا نقصد الهندسة اللسانية، وهذه الأخيرة لا تتوقف عن التقدم والتطور نظرية ومنهجا. وتزداد هذه العلاقة أهمية بالنسبة للغة العربية التي تنفرد بخصوصيات لسانية صورية جعلت منها، لغة قابلة للاستجابة للإجراءات الهندسية اللسانية في أغلب المستويات ذات الارتباط بالجانب الصوري وخاصة في بابي الصرف والتركيب.

يتطلب العمل في هندسة اللغة العربية التمكن من نوعين متكاملين من المعرفة ([1]): المعرفة اللسانية العميقة وصفا وتصنيفا بمختلف جزئيات النظام اللغوي على ضوء أحدث النظريات اللسانية المعاصرة، وخاصة اللسانيات الصورية، والإلمام بالمعرفة الحاسوبية ذات الصلة بمعالجة اللغات الطبيعية، وخاصة في جانبها البرمجي، وهو ما يعني التأقلم مع التفكير المنطقي الذي تقوم عليه الآلة. فالحاسوب منظومة برمجية منطقية قوامها الخوارزميات الصارمة التي لا تشتغل بالظن أو بالنسبية، ولذلك فإن القواعد التي يجب أن تصاغ لهذه الغاية يجب أن تكون صورية وحاسمة لا تقبل أكثر من تأويل واحد لكل قضية.

إن الجمع بين هذين النوعين من المعرفة هو المحك في التقدم الذي يحققه مجال البحث في الهندسة اللسانية، ولطالما تجاهل واضعو البرامج الحاسوبية الجانب المعرفي اللساني مكتفين بالمعرفة الحاسوبية، وهو ما أدى إلى تطوير برامج عربية يمكن أن نقول عنها إنها ما تزال في بداية الطريق، الأمر الذي انعكس سلبا على صناعة البرمجيات العربية فأخرها مقارنة بالتقدم الذي حصل في البرمجيات الأجنبية، كما أدى إلى تأخير الحوار بين الإنسان العربي بلغته العربية والآلة، هذه الآلة التي أصبحت تفرض نفسها علينا في جميع المجالات.

لقد أنفقنا زمنا طويلا في البحث في المعالجة الآلية للغة العربية من أجل الرقي بها إلى مصاف اللغات العالمية الفاعلة في التقانة المعاصرة، وكان تركيزنا واقعا في البداية على المستوى التركيبي، ولم نع سبب تأخرنا في تنفيذ مشروعنا إلا في السنوات العشر الأخيرة، أي عندما واجهتنا صعوبات كثيرة مصدرها أننا كنا قد بدأنا بالمركب، وكان يجب علينا البدء بالمفرد من البيانات، ولذلك ترانا نعيد النظر في خطتنا الشاملة، فعدنا إلى الاهتمام بالمستوى الأصغر من التركيب، وهو المكون الصرفي، وبذلك تقدمنا خطوة إلى الأمام في عملنا. لقد ترسخت لدينا القناعة أن إهمال وصف المكونات الصرفية للغة العربية وصفا حاسوبيا صوريا مقابل الوصف التقليدي الذي انتهى عمره الافتراضي علميا هو السبب الرئيس في تأخر إنجاز المشروعات العلمية التي أنفقت عليها الملايين من الدولارات بدون طائل.

---

[1]) انظر الملحق رقم (1)

المعاجم الإلكترونية:

يبنى المعجم الآلي للغة العربية على وصف المفردات اللغوية من وجهة نظر تصريفها conjugaison واشتقاقها derivation مع ربط هذا الوصف بالمستوى النحوي أي بالمعجم التركيبي للغة العربية. وتجب الإشارة هنا إلى أن غرضنا من وضع هذا المعجم يتمثل أساساً في تقييس الدماغ البشري لسانياً في مستوى المعجم , وذلك في محاولة منا استكناه الكفاية المعجمية عند الفرد العربي وتقعيد إنتاجية هذا المستوى معلوماتياً و/ أو خوارزمياً . ولهذا فإن بناء هذا المعجم يدخل بشكل أساسي في ما أصبح يعرف اليوم باللسانات الحاسوبية العربية .

يتألف المعجم الآلي للغة العربية من ثلاث مستويات متكاملة : الأول وهو مستوى الجذور, وفيه يصاحب كل جذر بالمعلومات النحوية والمورفولوجية التي تستخرج منه, المستوى الثاني يتألف من معجم المفردات البسيطة, حيث أثبتت كل مفردة في قاعدة البيانات بناء على المعلومات النحوية والصرفية المتعلقة بكل واحدة منها . ويتفرع عن هذه القاعدة من المفردات البسيطة قاعدة أخرى تتألف من المفردات المعربة. أما المستوى الثالث فيتأسس على قاعدة بيانات من المفردات المركبة. وهي بدورها تتفرع إلى مفردات مركبة تشمل العادي منها والمسكوك, ومفردات مركبة معربة. يمثل هذا العمل القسم الأول من قاعدة البيانات التي أنجزناها. وهناك قاعدة بيانات أخرى تتعلق بالمعجم التركيبي الآلي للغة العربية. وسيتم الربط بين القاعدتين ببرنامج يسمح بمساءلتها للقيام بالمعالجة الآلية في المستويين معاً, والهدف هو جعل الآلة تقوم بتركيب نصوص عربية سليمة، وتعرف الجمل العربية السليم منها وغيره, والقيام بمراجعة الأخطاء الإملائية والتركيبية في النصوص العربية، وتعرف الجذور اللغوية وتمييز الزائد من الأصيل في المفردات العربية، وتوليد واشتقاق المفردات صرفياً وتركيبياً، الخ .

فكيف تم الربط بين مجال الهندسة واللغة ؟

الهندسة فن التحكم في النظم، والحاسوب بشقيه العتادي والبرمجي يقوم على هذا التحكم، واللغة نظام معقد متشعب المسالك: صوتا وصواتة phonology وصرفا وتركيبا ومعجما ودلالة وتداولا pragmatics.

لقد استطاع الباحثون في لغات غربية كثيرة وضع برامج حاسوبية لسانية طبقت فيها جميع الخوارميات الصورية التي تتعرفها الآلة، وهناك برامج كثيرة تجعل الحوار بين الإنسان الغربي والآلة ميسرا بلغته الطبيعية، نذكر منها: الترجمة الآلية والتوليف الصوتي والتعرف البصري على المحارف والمدقق النحوي والإملائي ...إلخ. إلا أن التطبيق على اللغة العربية ما يزال في بداية الطريق، هذا على الرغم من أن هذه اللغة لا تختلف عن أي لغة في العالم فيما يتعلق ببرنامج الكفاية اللسانية، وما تتميز به اللغة العربية عن غيرها يجب أن يدفع بها إلى مقدمة اللغات العالمية في التعامل مع الآلة، فهي تقوم على مكونين رياضيين هما الجذر والوزن، وهما معا غير موجودين في أغلب لغات العالم.

- يعد الجذر المكون النووي للكلمة، أما الوزن فيمثل هيكلها العام، يقوم الوزن بتوزيع الحركات على مختلف حروف الجذر كما يقوم بتوزيع المورفيمات التي تضاف إلى مكونات الجدر بغرض توليد الكلمات: (سوابق ولواحق وأواسط)، يقوم هذا التوزيع على قوانين صورية تجمع في عملها بين القوانين اللسانية والرياضية.

- هذا التشكيل الرياضي للغة العربية جعل منها لغة انصهارية Fusion، خلافا للغات الأخرى التي تعد لغات إلصاقية Ensembliste فيما يتعلق بتوزيع المورفيمات داخل بنية الكلمة الأساس. والقول بالانصهارية يؤدي حتما إلى القول بالطبيعة الجبرية Algebra للغة العربية.
- أما الصوائت التي تعادل Voyelles في اللغات اللاتينية فتوزيعها مقنن بواسطة خوارزميات التطابق بين الوزن والكلمة المنتجة، لذلك لا نضعها – نحن العرب – على الكلمات، لكننا نستطيع قراءتها بدون صعوبة، لأننا نقرأ لغتنا ونتكلمها بالأوزان وليس بالحركات، والمقصود هنا بالحركات تلك العلامات التي يتولى المحلل الصرفي الطبيعي توزيعها على الحرفين الأول والثاني من الكلمة في حالة الجذر الثلاثي، والحروف الثلاثة الأولى في حالة الجذر الرباعي، أما حركة لام الكلمة فيسند توزيعها للمحلل النحوي الذي يتعامل مع اللغة بوصفها منظومة من الوظائف الصورية التي لها نظامها الخاص بها.

هكذا يبدو أن اللغة العربية لغة رياضية في أساسها مكونة من منظومة من الخوارزميات الصورية، دخلها Input الجذور مرورا بالأوزان التي تتمتع بقوة الإصهار المورفيمي المبرمج وخرجها output الكلمات والجمل، وهي كما يلي:

$$\text{الكلمة} := \frac{\text{الجذر (ف ع ل) (زوائد، أواسط، لواحق)}}{\text{الوزن / الميزان}}$$

- تنتج الكلمة في العربية عن التفاعل بين الجذر بوصفه مكونا متغيرا، من حيث إنه منظومة من المواقع الصوتية (والصواتية) كل صوت فيه يوجد في توزيع تكاملي مع منظومة الأصوات العربية المحدودة العدد والصفات والمخارج، والوزن (أو الميزان) بوصفه مكونا ثابتا، من حيث إنه يمثل البنية النظرية للكلمة، ثلاثية أو رباعية. وقيل ثنائية (بوهاس)، كما أن الوزن يمثل البنية العميقة للكلمة، أما الميزان فيمثل بنيتها السطحية ([2]).
- يمثل الجذر دخل البرنامج اللغوي في كفاية المتكلم، في البداية يتم اختيار الجذر المراد تشغيله، (ثلاثي أو رباعي)، ثم تشرع الكفاية في تطبيق خوارزميات المطابقة بين مادة الجذر اللغوي (المتغير) والمادة الصورية (ف. ع. ل) (الثابت)، وفي مرحلة لاحقة يتم تفعيل الوزن/ الميزان عن طريق خوارزميات الإقحام التي تقوم بإدراج الزوائد (سوابق ولواحق وأواسط وحركات) في البنية النظرية للجذر بهدف توليد الكلمة، أما في مرحلة التحليل فإن العملية تكون معكوسة، إذ يتم تطبيق نوع آخر من الخوارزميات تتولى إرجاع الكلمة إلى بنيتها الأساسية، أي الجذر.

[2]) لمعرفة معاني هذه المصطلحات العلمية ينظر موقعنا على الإنترنيت: www.ajman.ac.ae/hannach

- يشمل المخطط السابق مبنى الكلمة من جهتين: البساطة (الكلمة في المفرد المذكر أو المؤنث)، والإعراب (والمقصود به تصريف الكلمة في التثنية والجمع بجميع أنواعه)، كما يشمل توليد الكلمات ذات الاشتقاق الميت دلاليا (المهمل)
- لكل صيغة صرفية مقابل دلالي مخزن في الكفاية، وهذا المكون الدلالي الصرفي هو الذي يضمن ربط المستوى الصرفي بالمعجم والدلالة، وتؤمن ظاهرة الانصهار التي تنفرد بها اللغة العربية حرية حركة الكلمات داخل الجملة، أما علاقة الصرف بالمستوى الصوائي Phonologique فلا تحتاج إلى دليل، وخاصة فيما يتعلق بالأصول المعتلة. ونظرا لارتباط كل صيغة صرفية بدلالة محددة تنتظم على شكل حقول دلالية فقد أخذ الصرف نصيبا وافرا من باب الدلالة، مثال ذلك من المصادر العربية:

الجذر	وزن الفعل	المصدر الأصلي	وزنه	الهيئة	المَرّة	الميمي	الصناعي
ذأب	فَعَلَ- يَفْعَل	ذَأْباً	فَعْل	فَعْلَة	فَعْلَة	مَفْعَل	+يّة

من خلال معرفتنا بدلالة الوزن نستطيع ربطه بما يمكن أن يولده من المداخل المعجمية، فلا أحد يخطئ في أن وزن "فَعْلَة" يدل على وقوع الحدث مرة واحدة، مثل: "عظة" التي تدل على الوعظ مرة واحدة لا غير. كما نعرف أن "فَعْلَة" يدل على وقوع الحدث على شكل هيئة معينة، مثل: قَعدة، أو جَلسة، إلخ. وهذا يدل على أن المكون الصرفي يعد مرشدا إلى الدلالة، إذ يكفي وضع الكلمة في وزنها الصحيح لنخلص منها إلى معناها الذي تدل عليه، وبذلك يكون الصرف محور العمليات اللسانية الحاسوبية الذي تنبني عليه مستمويات أخرى كثيرة في منظومة التحليل الآلي للغة العربية.

أما في مجال تركيب الجمل فإننا ننظر إلى الفعل بوصفه دالة Fonction، وبقية العناصر فمتغيرات، وذلك كما يلي:

$$P \longrightarrow V\ (n1, n2 \ldots nx)$$

وعلى الرغم من قلة عدد الأوزان الفعلية المولدة للأفعال في نظام اللغة العربية، حيث لا تتجاوز **60** وزنا أصليا، فإنها تعد الأكثر إنتاجية مقارنة ببقية الأصناف الصرفية الأخرى في النظام اللغوي العربي، وهذا يؤدي إلى استخلاص نتيجة لسانية حاسوبية تتمثل في القول: كلما قل عدد أوزان الصنف اللغوي كلما كان أكثر إنتاجية وذلك كما يلي:

الباب الفعلي: فَعَل – يفعُل =: كتب – يكتب				
مصدر =: كتابة	اسم الفاعل =: كاتب	اسم المفعول =: مكتوب	مبالغة =: كتّاب	تفضيل =: أكتب

فالمصادر بأصنافها الخمسة: الأصلي، والميم، والهيئة، والمرة، والصناعي، وكذلك جميع أصناف المشتقات، وعددها ثمانية، كلها تخضع لخوارزميات توليد الأفعال، قبل أن يطبق عليها قنوها التوليدي الخاص، مما يجعل من هذين القسمين (المصادر والمشتقات) أصنافا صغرى S/classes تخضع لسانيا وحاسوبيا لقوانين توليد الأفعال.

وقد بني المعجم الإلكتروني لتراكيب اللغة العربية على هذا الأساس.

- ونظرا للطبيعة الجبرية algebra لنظام اللغة العربية المتمثلة في منظومة الخوارزميات اللسانية التي تجمع بين مختلف مكونات النظام، فإن الربط بين هذه المستويات أمر بالغ التعقيد لا يمكن أن تقوم به الآلة إلا إذا تم تزويدها بالقواعد اللسانية الصورية، أي بالخوارزميات اللغوية، بنوعيها التوليد و التحليل..
- إذا قارنا هذا العمل بمكونات الآلة نجد بينهما تشابها كبيرا في خطوات العمل وفي طبيعة النتائج التي يتوصل إليها البحث في كلا المجالين، وهي تسير وفق ما يلي:

بيانات مشفرة ⟸ ترجمة الخوارزميات ⟸ اللغة الطبيعية

تمثل الخطوة الأولى لغة صورية (=: الجذر) تتبع تركيبا معينا في بناء الموضوع، وفي الخطوة الثانية يتم ترجمة خوارزميات البرمجة إلى مقدمة لبناء النص باللغة الطبيعية (=: الوزن)، ثم تأتي في المرحلة الثالثة اللغة الطبيعية (=: الكلمة) التي تعد النتيجة التي يتم من خلالها التأكد من صحة البرنامج من عدمه.

التطبيق:

يقوم تصورنا للعمل في المكون الصرفي على أن الكلمة تتولد بإضافة الزوائد والحركات (الثوابت Constant) بشكل انصهاري إلى الجذر (المتغير Variable)، أما الوزن فيعد الهيكل الذي تنتظم فيه الثوابت والمتغيرات، ويتم ذلك بموجب تفعيل ثلاثة أنواع من الخوارزميات ([3]):

أولا: خوارزميات المطابقة، وتقوم بالمقابلة بين المكون الصوتي (الجذر) وأحد الهياكل الصرفية التي تجمع في (ف ع ل) أو (ف ع ل ل) استعدادا لإدراج الزوائد والحركات في بنيته الصرفية حسب ما تسمح به القوانين الصواتية phonological rules للجذر root، ويتم ذلك وفق نظام تطريزي prosodic محكم توظف فيه كافة الخوارزميات الصورية التي تراعي التكوين الداخلي للجذر، تقوم هذه الخوارزميات بمهمة الاستبدال والمطابقة بين الجذر بوصفه متغيرا، ومكونات الوزن بوصفه ثابتا، وذلك وفق مقتضيات المحتوى الصوتي والصوائي للجذر، مثال ذلك ([4]):

الدخل =: الجذر	خوارزميات المطابقة	الخرج =: الكلمة
ك ت ب	/ك/ =: ف، /ت/=: ع، / ب/ =: ل/	كتب

ثانيا: خوارزميات الاشتقاق، وتتولى توزيع الحركات والزوائد (سوابق ولواحق وأواسط) داخل الهيكل الصوتي (الجذر الثلاثي أو الرباعي) باستخدام الهيكل الصرفي (ف ع ل (ل))، وتتم عملية التوزيع وفق مبادئ

[3]) اللغة العربية والتقنيات العلوماتية المتقدمة، د/ محمد الحناش، 1996

[4]) Traitement automatique du mot: Etat de l`art, Eric Laporte, IGM, 2001

الانصهار التي تميز النظام الصرفي للغة العربية عن غيره من أنظمة العديد من اللغات وخاصة اللغات الأوروبية ([5]).

وقد تبين لنا أن توليد أقسام الكلم الأساسية من الجذر تقوم على منظومات من الخوارزميات تراعي التعدد المقولي للغة، فلكل مقولة نحوية مولدة من الهيكل الصوتي (الجذر) منظومتها الخوارزمية الخاصة بها، مع اتفاق في بنية الهيكل الصرفي الأساسية.

فللأسماء البسيطة خوارزمياتها الخاصة بها التي تولدها من الجذر، وللأفعال البسيطة خوارزميات أخرى خاصة بها. تتميز خوارزميات أسماء الذوات بكونها تولد بنيات اسمية مغلقة، حيث لا تتفرع عنها إلا بعض البنيات الاسمية الداخلية التي تعد من صميم التكوين الإعرابي للاسم، بينما تتميز خوارزميات الأفعال بالانفتاح على نفسها أولا (أزمنة الأفعال وتصريفها مع مختلف الضمائر)، ثم على بنيات أخرى متعلقة بها اشتقاقيا، والمقصود بها المصادر والمشتقات بجميع أنواعها.

ثالثا: خوارزميات التوليد، والمقصود بها العمليات الصورية التي تولد البنيات المُعربة، والمقصود بالإعراب هنا مختلف التغيرات الاشتقاقية التي تعرفها الكلمة داخل مجالها التصريفي، فالاسم يتفرع إلى مفرد ومثنى وجمع تكسير (قلة وكثرة)، وتعريف وتنكير وتصغير ونسب، ... إلخ، ولكل مقولة فرعية نظامها الخاص بها فيما يخص توزيع الزوائد والحركات بشكل انصهاري. أما الفعل فيعد أكثر إنتاجية من حيث التصريف، إذ تتفرع عنه مجموعة من البنيات التصريفية يكون أغلبها على مستوى الزمن، وقد ضاعف وجود الزمن في الفعل من تعقيد مهمة الخوارزميات، ففيه الفعل المضارع والماضي والأمر، ولكل منها نظامه الاشتقاقي الخاص به، كما أن لكل منها جدولا خاصا يتوزع فيه الفعل مع مختلف الضمائر. أما اشتقاق المصادر والمشتقات فتعد أعباء إضافية يتحملها الفعل الذي يفتح الطريق أمام توليدها، إذ بدون الفعل لن نتمكن من الوصول إليها بحال وفق فرضيتنا العلمية التي نشتغل بها ([6]).

بناء على هذه الفرضية:

يمكن القول إن الوزن والميزان يمثلان أحد تجليات النظام الخوارزمي في شقيه الاشتقاقي والتوليدي، فالوزن بوصفه الهيكل الصرفي الذي يخطط لبناء الكلمة قبل أن تخرج إلى الوجود يعد المجال الحقيقي لبناء المفردات، وبدونه لا يمكن أن نصل إلى المفردة أبدا، ولهذا حرصنا على تمثيل أغلب البيانات بالأوزان في قاعدة البيانات التي سنتحدث عنها لاحقا.

الوزن هو العمدة في بناء الكلمات، فهو الحامل لمادة الجذر وفيه تنصهر الصوائت القصيرة والطويلة مع مكونات الجذر الصامتة، وهو الذي يسمح بتطبيق جميع القوانين الصوائية للغة العربية، وهو الذي يسمح لنا أيضا

[5] Les bases de données du LADL, Analyse automatique des langues naturelles, Aspects technologiques, Paris, 1989

[6] تختلف نظرية الاشتقاق التي نعمل بها عن نظرية القدماء في الموضوع نفسه.

بكتابة اللغة العربية وقراءتها وهي بدون حركات. نخلص من هذا إلى أننا نكتب اللغة العربية ونقرأها بالأوزان، نكتفي بكتابة الصوامت مع الصوائت الطويلة، ولكننا نقرأ الكلمة بصوامتها وصوائتها (بنوعيها)، فأين تختفي الصوائت القصيرة إذن في كتابتنا العربية ؟، وما الذي يجعلنا نلحقها بالصوامت؟ إنه الوزن، والوزن فقط، وهذا الأخير يقوم على منظومة من الخوارزميات التي تسمح له بالتعامل مع الهيكل الصوتي، أي الجذر. أما من حيث الدلالة فإن الوزن وحده يحمل دلالة صرفية تعد لوحدها بمثابة القالب العام الذي تفرغ فيه الدلالة المعجمية. ويتم ذلك على مراحل كما في هذا المثال:

1) الجذر (ثلاثي) =: ك ت ب (الفراغات الموجودة بين الحروف الأصلية لإدخال الحركات القصيرة والطويلة)
2) الوزن (المصدر) =: فِ عا لَ ة (الفراغات مخصصة لحركات الوزن المعروفة سلفا)
3) تطبيق الوزن على الجذر =: كِ تا ب َ ة : كِتابَة

لا يمكن الوصول إلى الخرج (كتابة) إلا عبر المراحل الثلاث الأولى التي تسبق الخرج، وبدونهما يستحيل على أي متكلم عربي أن يصل إلى هذه النتيجة في اللغة العربية بشكل صحيح.

وقد احتجنا في هذه العملية البسيطة إلى البرامج الطبيعية التالية:

1. منظومة من القواعد الفونولوجية التي تسمح بتأليف الصوامت مع بعضها في بنية الجذر، وبدونه لا يمكن أن يستقيم الدخل ويسمح بإدراجه في المنظومة اللسانية العربية.
2. مولد صرفي طبيعي للكلمات يربط بين الجذر وتوجه التوليد Orientation de generation (هنا المصدر الأصلي)، وهو يختلف عن أنواع أخرى من المصادر العربة: الميمي، الصناعي، المرة، والهيئة.
3. مشكل طبيعي يضع الحركات المناسبة على الصوامت التي تتألف منها المتوالية اللسانية (الكلمة =: المصدر).
4. محلل دلالي يسمح بالتأكد من المحتوى الخبري للخرج.

فإذا انتقلنا من الكفاية الطبيعية إلى الكفاية الحاسوبية المراد نسخها على الحاسوب، سنرى أن الأمور تتعقد أكثر، لأننا سنكون في حاجة إلى الانتهاء أولا من الأمور الآتية:

1. بناء قاعدة بيانات الجذور العربية التي تقبل أن تكون دخلات لتوليد المصادر العربية الصحيحة، أي بناء معجم إلكتروني يحتوي على جميع الجذور المستعملة في اللغة العربية وتشفيرها بشكل صحيح، وهو ما يمكن أن يترجم بمصطلح الاستقراء التام، درءا للوقوع في الأخطاء التي يؤدي إليها أسلوب الاستدلال القائم على منطق الفرضيات والاحتمالات.
2. ربط كل جذر بشبكة الأفعال التي تتولد عنه، لأن الفعل يعد دخلا سابقا لكل مصدر، وما ذلك إلا للانضباطية التي يتميز بها الفعل، فلكل فعل مصدر (تستثنى الأفعال الجامدة وبعض الأفعال الناقصة).
3. بناء قاعدة بيانات الأوزان والموازين الصرفية التي تتولد بها المصادر في اللغة العربية، وهو ما يمكن أن يترجم بالمولد الصرفي العربي الخاص بالمصادر.

4. ضبط منظومة القواعد الصواتية التي تتحكم في الصوامت المعتلة (ا و ي) عند الانتقال من مرحلة توليد الفعل إلى مرحلة توليد المصدر.
5. تطوير مشكل آلي يتولى توزيع الحركات المناسبة على الصوامت داخل الوزن، وهذا قد يكون ميسرا نظرا لأن الوزن نفسه يتضمن الحركات.

ومن الجدير بالذكر أن الوزن في اللغة العربية يعد الحجر الأساس في تطوير أي برنامج للتشكيل الآلي للنصوص، لأن تشكيل الكلمة يخضع لتعرف بينتها الداخلية التي تنتج عن معرفة الوزن الصرفي، إلا أن دور الوزن لا يتجاوز تشكيل الصوامت الأولى للوزن ، لأنه لا يتحكم في الحركة الأخيرة، فهذه من اختصاص المولد النحوي، ولا أظن أننا سنطلب في المرحلة الأولى من أي برنامج صرفي، سواء كان مولدا أو محللا، أن يهتم بالحركة الأخيرة، ولهذا السبب ترانا أهملناها في المثال الذي قدمناه سابقا.

تجدر الإشارة إلى أن الوزن الصرفي يشكل العمود الفقري البرامج الحاسوبية التالية:

1. التشكيل الآلي للنصوص، سواء كان تاما أو ناقصا.
2. التعرف البصري على المحارف العربية، إذ لا يعقل أن يشتغل برنامج OCR خارج نطاق الصرف
3. الترجمة الآلية، إذ بدون ضبط الوزن لا يمكن أن تقوم أي ترجمة آلية
4. برامج النطق الآلي للنصوص، بدون معرفة الوزن يتعذر تطوير برامج التوليف الصوتي وتخليق الكلام.

وغيرها من البرامج المتعلقة بمعالجة اللغات الطبيعية في عصر تقانة المعلومات.

قسمت قاعدة بيانات المصادر إلى مكونين لغويين أساسيين:

الأول: البيانات المشتركة التي تمثل دخل المعجم، أي الأدوات الأساسية التي نظمت على شكل معاجم فرعية:

- معجم الجذور (الوحدات الذرية) والأفعال التي تعد المرحلة الوسيطة في توليد المصادر بجميع أصنافها.
- معجم الأوزان (الخوارزميات) التي تولد المصادر عن طريقها، إذ إن الكلمة التي ليس لها وزن تصنف خارج بنية النظام الصرفي العربي، وتكون إما معربة أو دخيلة.

الثاني الحصيلة اللغوية التي توصلنا إليها، التي تشكل العتاد اللساني الذي نشير إليه عادة بمصطلح قاعدة البيانات اللغوية، ومجموعها 2.50.000 مدخل مصدري موصوف ومشفر ومصنف فيخمسة أصناف صغرى.

يلخص الشكل (2) مختلف المراحل التي يمر بها اشتقاق المفردات العربية مع إحصاء يكاد يكون دقيقا لمختلف المداخل المعجمية المولدة بهذين النوعين من الخوارزميات، كما يبين العلاقة الاشتقاقية والترابطات المتشعبة بين المداخل ([7]):

[7] Dictionnaires électroniques et analyse automatique de textes, Max Silberztein, Masson, Paris, 1993)

طريقة بناء قاعدة البيانات

صممت قاعدة البيانات بطريقة تسمح للمبرمجين بقراءتها بشكل ميسر، فقد وضعت على شكل جداول يمكن قراءتها بأي برنامج مختص في قواعد البيانات، مثل Access أو Excel. وتقوم على أعمدة وصفوف، بحيث تضبط المعلومة المطلوبة عند نقطة الالتقاء بينهما. أما عدد الأعمدة فهو كما يلي:

1- حقل الجذور التي تتفرع عنها الأفعال والمصادر (مثبتة معجميا).
2- حقل وزن الفعل (فعل – يفعل) الذي يشتق منه المصدر.
3- حقل المصدر الأصلي (مثبت معجميا).
4- حقل وزن المصدر الأصلي.
5- حقل وزن مصدر الهيئة.
6- حقل وزن مصدر المرة.
7- حقل وزن المصدر الميمي.
8- حقل وزن المصدر الصناعي.

لقد ركزنا في جمع البيانات وتصنيفها على الجانب الصوري دون الجانب الصوتي المفعّل معجميا، ولم نثبت معجميا في قاعدة البيانات إلا الجذر والمصدر الأصلي موضوع هذا البحث، أما بقية الاشتقاقات التي تتم من الجذر فقد اكتفينا بإثبات أوزانها يقينا منا أن الوزن هو العمدة التي تبنى عليها البيانات اللغوية في جميع المراحل الحاسوبية اللاحقة. ونظرا لأهمية التشكيل في بناء قاعدة البيانات فقد ضبطنا جميع المواد التي تشكل متن المعجم الإلكتروني للمصادر العربية، وذلك تحسبا لإمكانية استغلاله في تطوير برامج الترجمة الآلية والتخاطب الصوتي بين الإنسان والآلة، وبرامج القراءة الآلية للنصوص العربية، إلخ.

خلاصة:

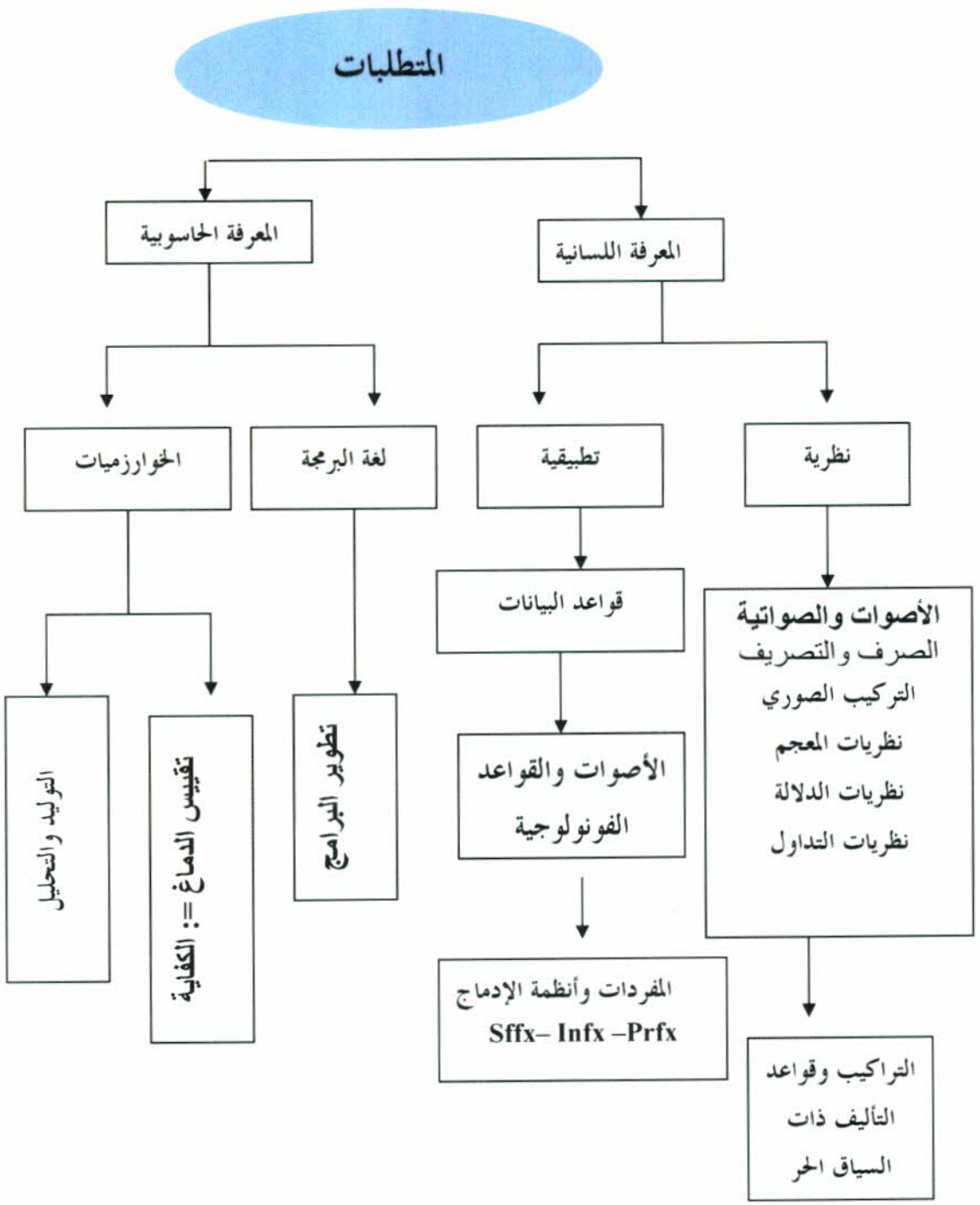
لقد تبين لنا من خلال هذا العرض السريع لبنية الكلمة العربية أن الوزن الصرفي يعد العمدة في بنيتها اللسانية، كما تبين لنا أن معرفة الوزن آليا تحتاج إلى ضبط عدد من المكونات اللسانية التي تنضبط في قواعد مطردة على شكل خوارزميات صورية تتولى توليد الكلمة في المعجم. كما تبين لنا أن سبب إهمالنا للحركات في الكتابة هو كوننا نقرأ العربية صرفيا، إذ لمجرد أن تقع أعيننا على الكلمة نرسلها إلى المحلل الصرفي الطبيعي الذي يتولى وضعها في صيغتها الصرفية المناسبة، وبذلك يتم تعرفها آليا وقراءتها بشكل سليم، بعد أن نسند إليها المحتوى الخبري الذي يناسبها.

كما عرفنا أن أي برنامج حاسوبي يتعامل مع اللغة العربية بوصفها لغة طبيعية لا يمكن أن ينجح إلا إذا فهم آلية بناء الكلمات صرفيا، ومن هنا كان المكون الصرفي عمدة في معظم البرمجيات العربية الناجحة، منها الترجمة الآلية، والتعرف البصري على المحارف، والنطق الآلي للنصوص، والتخاطب الآلي بين أنظمة الإرسال المتطورة، والتخاطب بين البشر والآلية.

كما طبقنا هذا المفهوم على المصادر العربية، وقد تبين أن هذا المكون الصرفي في اللغة العربية يعتمد كليا على خوارزميات صورية تولده من الأفعال، ولذلك قمنا بربط كل مصدر بالفعل الذي يشتق منه تيسيرا لعلمية الحوسبة التي نعتزم الشروع فيها مستقبلا.

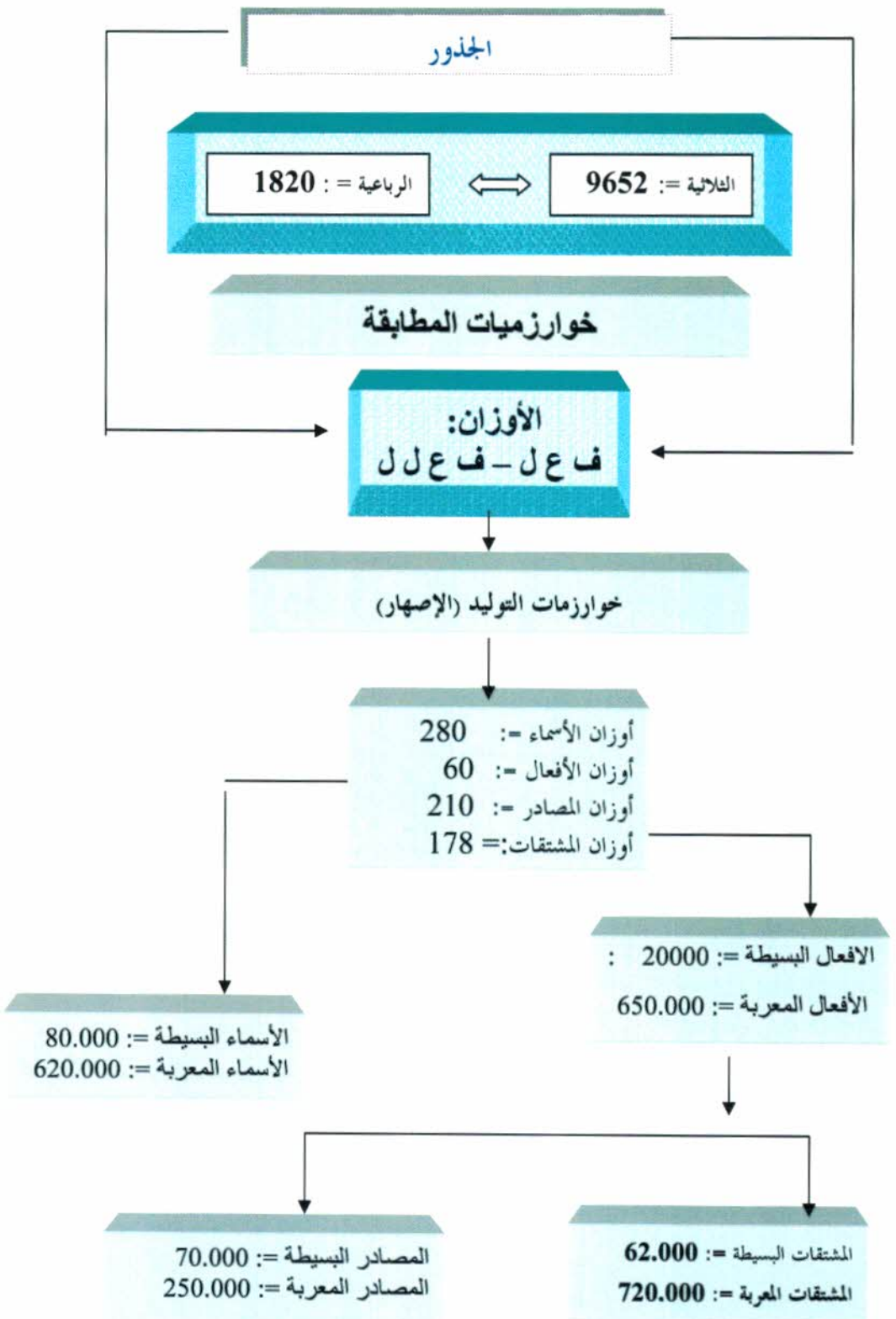
ملحق (1)

متطلبات بناء المعجم الإلكتروني للغة العربية



ملحق (2)

إحصائيات التوليد الصرفي في العربية



ملحق (3): نموذج قاعدة بيانات المصادر العربية:

الجذر	وزن الفعل	المصدر الأصلي	وزنه	الهيئة	المَرّة	الميمي	الصناعي
ذأب	فَعَلَ- يَفْعَل	ذَأْباً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	فَعِلَ- يَفْعَل	ذَأَباً	فَعَل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	فَعُلَ- يَفْعُل	ذَآبَةً	فَعالَة	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	أَفْعَلَ	إِذْآباً	إِفْعال	–	+ة	مُفْعَل	+يّة
	فَعَّلَ	تَذْئِيباً	تَفْعِيل	–	+ة	مُفَعَّل	+يّة
	تَفَعَّلَ	تَذَؤُباً	تَفَعُّل	–	+ة	مُتَفَعَّل	+يّة
	تَفاعَلَ	تَذاؤُباً	تَفاعُل	–	+ة	مُتَفاعَل	+يّة
	اسْتَفْعَلَ	اسْتِذْئاباً	اسْتِفْعال	–	+ة	مُسْتَفْعَل	+يّة
ذأت	فَعَلَ- يَفْعَل	ذَأْتاً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
ذأج	فَعَلَ- يَفْعَل	ذَأْجاً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	فَعِلَ- يَفْعَل	ذَأِجاً	فَعَل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	انْفَعَلَ	انْذِئاجاً	انْفِعال	–	+ة	مُنْفَعَل	+يّة
ذأح	فَعَلَ- يَفْعَل	ذَأْحاً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
ذأذأ	فَعْلَلَ	ذَأذَأَةً	فَعْلَلَة	–	+واحدة	مُفَعْلَل	+يّة
	تَفَعْلَلَ	تَذَأذَؤاً	تَفَعْلُل	–	+ة	مُتَفَعْلَل	+يّة
ذأر	فَعَلَ- يَفْعَل	ذَأْراً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	فَعِلَ- يَفْعَل	ذَأِراً	فَعَل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	أَفْعَلَ	إِذْآراً	إِفْعال	–	+ة	مُفْعَل	+يّة
	فاعَلَ	مُذآرَةً	مُفاعَلَة	–	+واحدة	مُفاعَل	+يّة
ذأط	فَعَلَ- يَفْعَل	ذَأْطاً	فَعْل	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
ذأف	فَعَلَ- يَفْعَل	ذَأْفاً- ذَأَفاً- ذَأَفاناً- ذُؤافاً	فَعْل- فَعَل- فَعَلان- فُعال	فِعْلَة	فَعْلَة	مَفْعَل	+يّة
	أَفْعَلَ	إِذْآفاً	إِفْعال	–	+ة	مُفْعَل	+يّة
	انْفَعَلَ	انْذِئافاً	انْفِعال	–	+ة	مُنْفَعَل	+يّة

ملحق (4)

نموذج تردد Frequency أوزان المصدر في اللغة العربية

وزن المصدر	مجموع المصادر الأصلية المولدة	النسبة المئوية
فعْل	5976	17.997
فعَل	3381	10.182
إفعال	**1817**	5.472
فُعُول	1738	**5.234**
تَفْعِيل	**1664**	**5.011**
تَفَعُّل	**1533**	**4.616**
افْتِعال	**1149**	**3.460**
فعْلَلة	**965**	**2.906**
فعَالة	916	**2.758**
مُفاعَلة	**818**	**2.463**
تَفاعُل	**673**	**2.026**
فعِيل	660	**1.987**
فِعَال	656	**1.975**
فُعُولة	582	**1.752**
فُعَال	579	**1.743**
اسْتِفْعال	**533**	**1.605**
انْفِعال	**408**	**1.228**
فعَلان	357	**1.075**
فعْلة	317	**0.954**
فعَال	299	**0.900**
فِعَالة	274	**0.825**
فُعْلة	273	**0.822**
المجموع	**33205**	**100**

مصادر البحث ومراجعه:

- اللغة العربية والتقنيات العلوماتية المتقدمة، د/ محمد الحناش، 1996
- اللغة العربية والحاسوب: الهندسة اللسانية العربية، مجلة اقرأ، العدد السابع، الشارقة، 2004
- Traitement automatique du mot: Etat de l'art, Eric Laporte, IGM, 2001
- Les bases de données du LADL, Analyse automatique des langues naturelles, Aspects technologiques, Paris, 1989
- Dictionnaires électroniques et analyse automatique de textes, Max Silberztein, Masson, Paris, 1993