

اجتماع خبراء التدقيق الإملائي والنحوي والتشكيل الآلي، المعهد العالي للعلوم التطبيقية والتكنولوجيا دمشق،  
المنظمة العربية للتربية والثقافة والعلوم، دمشق 18-20 أبريل 2011.

## إثراء قاموس التدقيق الإملائي بالتفاعل مع المستخدم

طه زروقي<sup>1</sup>

المجلس الأعلى للغة العربية بالجزائر،

البريد الإلكتروني Taha.zerrouki@gmail.com

**الخلاصة.** يهدف مشروع المدقق الإملائي المفتوح المصدر إلى بلوغ هدفين هما تحسين دعم اللغة العربية في برامج التدقيق الإملائي الحرة المتداولة. وبناء قاموس التدقيق الإملائي العربي الحر.

وقد تمّ التوصل إلى أول إصدار عملي أوائل سنة 2008، ليعمل على برنامج المكتب المفتوح OpenOffice.org ومتصفح فيرفكس (Firefox)، وكثير من البرامج الحرة، وقد تُقبل بقبول حسن في عالم المصادر الحرة.

وخلال عامين منذ الإصدار الأول تتبعنا العديد من المسائل المتعلقة بالمشروع، والتي طرحها المطورون والمستخدمون، ومن بين هذه المشاكل مسألة سرعة المدقق، وتعقيد قواعد التدقيق، مما دفعنا إلى العمل على تبسيط القاموس، وتفضيل سرعة المدقق على حساب حجم البيانات، بتوليد الأشكال المتنوعة للفعل ببرنامج قطرب لتصريف الأفعال.

من جهة أخرى نسعى إلى إطلاق موقع خدمات التدقيق الإملائي، يتفاعل مع المستخدم ويساهم في إثراء القاموس بإضافة الكلمات الناقصة، أو التنبيه إلى أخطاء محتملة.

**الكلمات الجوهرية:** التدقيق الإملائي، معالجة اللغة العربية، تصريف الأفعال، التفاعل مع المستخدم

### 1 مقدمة

يكتسي التدقيق الإملائي أهمية كبيرة تتبع أهمية الكتابة نفسها، فأى نص مكتوب لابد من تدقيقه من الأخطاء البشرية والمطبعية، للحفاظ على لغة سليمة، سواء من قبل مراجع بشري، أو ب الاستعانة ببرامج التدقيق.

وتعد برامج التدقيق جزءاً مهماً يصاحب برامج الكتابة، حتى صارت وظيفة أساسية لا تتجزأ عنه، ومما يدل على أهمية برامج التدقيق الإملائي، نذكر أن شركة كولنك باعت منتجها للتدقيق العربي بقيمة مليون دولار. لشركة مايكروسوفت.

ومع ظهور برامج الكتابة والمراسلة الحرة، وشغف المستخدمين بها، ظهرت مدققات إملائية حرة، لأهميتها القصوى. وكان انعدام مدقق إملائي عربي حر، مسألة شغلت المطورين العرب والمستخدمين، ردحا من الزمن، وقد تبع ذلك العديد من المحاولات [1] لسد هذه الثغرة الكبيرة، إلى أن انتهينا من مشروعنا أيسبل لإنجاز قاموس عربي مفتوح المصدر. [1] يعمل على برنامج هانسبال [2]، ليلبي هذه الحاجة الملحة.

كما يجدر بنا أن نذكر أنّ حزمة للتدقيق الإملائي العربي (آية-سبل) لبرنامج المتصفح فيرفوكس تجاوز تحميلها مليون مرة. هذا يدل على الانتشار الواسع للقاموس خلال السنتين الماضيتين، وعن مدى استعمال المستخدمين للتدقيق الإملائي لرسائلهم على مستوى المتصفح (Firefox) بغض النظر عن نظام التشغيل المستخدم. وقد حصل هذا المشروع على عدد من التشجيعات والتقديرات في عالم المصادر المفتوحة [3][4].

وخلال الفترة الماضية منذ أول إصدار، تتبعنا الكثير من المسائل التي طرحت، إذ رغب العديد من المطورون في إعادة استعمال القاموس لأغراض أخرى نظرا لانعدام قاموس عربي مفتوح المصدر حتى الآن، وقد صادف المطورون مشاكل في تكييف قاموس التدقيق لحاجات أخرى نظرا لأنه صيغ من أجل برنامج هانسبال فقط [4].

على صعيد آخر، طلب مطورون تكييف القاموس لاستعماله مع برامج تدقيق أخرى أهمها برنامج Aspell، وتدقيق المواقع بواسطة لغة PHP ضمن مشروع مكتبة بي أتش بي واللغة العربية لخالد الشمعة. ويعود التعقيد في اعتماد القاموس أن برنامج هانسبال يتفوق على Aspell ببعض المزايا التي يعمل أسبل على إضافتها ابتداء من الإصدار 0.61 الذي لا يزال تجريبيا، وفور صدوره سيتوافق مع القاموس بإصداره الحالي [4].

## 2 مراحل المشروع

لقد تم تقسيم المشروع إلى مرحلتين :

**المرحلة الأولى** بناء القاموس اللغوي، باستعمال الخصائص المتوفرة في برنامج هانسبال، وكذا إضافة خاصية إغفال التشكيل في مرحلة أولى، وقد انتهت هذه المرحلة بإصدار القاموس ونشره على الإنترنت. وقد تحدثنا عن نتائج هذه المرحلة في مقالات سابقة يمكن الاطلاع عليها في موقع البرنامج [4].

**المرحلة الثانية** : تحسين آلية برنامج التدقيق وإضافة المزايا العربية، وإعادة هيكلة القاموس وتجريده وإثرائه. وبرمجة الأدوات المستخدمة في بناء القاموس وتوليد أجزاء منه والتحويل بين الصيغ المختلفة، ودعم البرامج المختلفة للتدقيق الإملائي. وتطوير التطبيقات ونعني بها المجالات التي تخدم التدقيق الإملائي مثل تصميم موقع خدمي للتدقيق الإملائي.

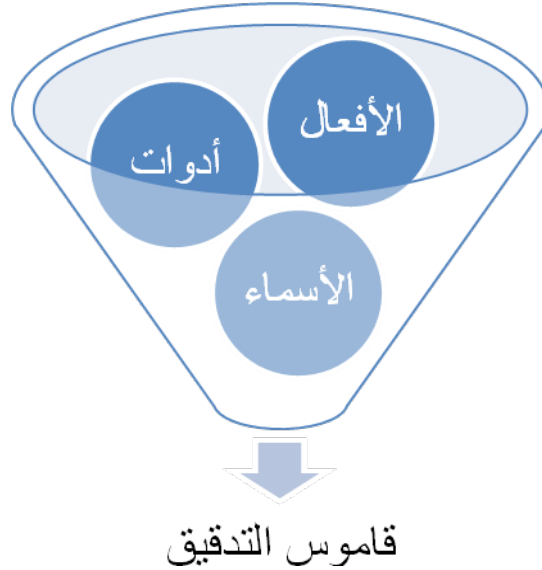
وقد انطلقنا فعلا في العمل على المرحلة الثانية من المشروع، بالعمل على تجاوز بعض عيوب القاموس، وإثرائه بالمفردات، وسنتحدث في هذا المقال، عن العمل على تسريع التدقيق الإملائي بزيادة حجم البيانات وتقليل القواعد، وذلك بإضافة مداخل متعددة للفعل الواحد، والاعتماد على برنامج قطرب لتصريف الأفعال العربية [5]. وقبل ذلك سنتحدث عن طريقة بناء القاموس في

الإصدار الأول، وأهمية المكوّن الفعلي، والعمل المبذول فيه. ثم نبرز أهمية الاعتماد على الطريقة الجديدة في التدقيق.

### 3 بناء القاموس في المرحلة الأولى

استدعى غياب قاموس عربي حر، بناء ما يلبي شروط التدقيق الإملائي بالاعتماد على المعجم اللغوية العربية المتداولة التراثية والحديثة. من هذه المعاجم أساساً معجم الأفعال العربية (مجموعة باشريل)[6]، المعجم الوسيط، المعجم الغني، معجم المحيط ولسان العرب.

يتكون قاموس التدقيق الإملائي من ملفين هما ملف المفردات المصنفة مرفقة بأسماء القواعد التصريفية الممكنة لكل واحدة منها، وملف الزوائد أو قائمة قواعد الزيادة الممكنة تطبيقها على المفردات. وقد تطلب إنشاء القاموس قرابة سنتين من النشاط المتواصل، تم خلالها تحليل آلاف المفردات وتصنيفها حسب القواعد النحوية والصرفية، وتحديد معناها لتمييز تعدي الأفعال. إجمالاً، تمت معالجة أكثر من 50.000 مفردة تتوزع على ما لا يقل عن 10.000 فعل عربي، و40.000 اسم وعشرات الحروف والأدوات النحوية وما استنتج منها.



الشكل 1. مكونات القاموس

#### 3.1 تدقيق الأفعال إملائي

مثلت معالجة الفعل الشطر الأكبر من بناء القاموس لكثرة التصريفات الممكنة لكل فعل وتعدد الحالات الإملائية فيها من إعلال وضبط للهمزة وتضعيف وإدغام، ويعتمد تدقيق الأفعال على مادة

لغوية تحتوي 10.000 فعل، أضيفت إليها الأشكال الخاصة بالإبدال والإعلال والتضعيف والإدغام وضبط الهمزة، فارتفع عدد المفردات في قاموس آيسبل Ayaspell إلى ما يقارب 15.000 ( 14523 مفردة بالضبط).

وقد وضعت القواعد التصريفية للأفعال في ملف الزوائد لتغطي كل صيغ التصريف الممكنة في اللغة العربية ما عدا صيغ المؤكد. وقد وضعنا 140 جدولاً نموذجياً لتصريف الأفعال، بالاعتماد على تصنيف الأفعال العربية حسب نموذج التصريف الملائم.

### 3.1.1 تصنيف الأفعال

بعد إنجاز هذه الجداول ومجموعات قواعدها وتصنيف الأفعال حسب منوالها، لابد من النظر إلى دلالات الفعل من ناحية اللزوم والتعدي فمثلاً الفعلان دعا ودنا لهما نفس المنوال في التصريف، لكنّ دعا فعل متعدي، أي أنه يتصل بالضمائر المتصلة في محل نصب مفعول به، مثل دعوته. بينما دنا فعل لازم، لا يتصل بالضمائر المتصلة فلا نقول دنوته.

مما يجعل لكل منوال مجموعتين من قواعد الزيادة، الأولى للمتعدي لتسمح بزيادة الضمائر المتصلة، والثانية للفعل اللازم الذي لا يتصل بها.

#### 3.1.1.1 أنواع التعدي واللزوم

وتفصيلاً للتعدي واللزوم، ينبغي النظر جيداً إلى ما يعترى الفعل من دلالات قد تؤثر فيما يجوز. أن يلتصق به من ضمائر متصلة في محل نصب مفعول به، فاحتجنا إلى وضع تصنيف دلالي للتعدي واللزوم، استقيناه من المعاجم واستعملات الأفعال والتصنيفات هي :

- 1- الفعل اللازم المطلق: [حَسُنَ] وهو الفعل الذي لا تقبل دلالاته أن يأتي منه المبني للمجهول مطلقاً، ولا الأمر، فلا يجوز أن أقول هو حُسِنَ، ولا أن أمره بأن يحسُنَ.
- 2- الفعل اللازم النسبي: [انكسر] وهو الفعل الذي لا تقبل دلالاته أن يأتي منه المبني للمجهول إلا لضمير الغائب المفرد، ويجوز أن يأتي منه الأمر والأزمنة المعلومة.
- 3- الفعل المتعدي لغير العاقل: [قال] وهو الفعل الذي لا تقبل دلالاته أن يأتي منه مفعول به عاقل، أي لا يتصرف مع الضمائر العاقلة في المبني للمجهول، فلا يجوز. أنتما قلتما في الماضي المبني للمجهول.
- 4- الفعل المتعدي للعاقل: [ضرب] وهو الفعل الذي تقبل دلالاته أن يأتي منه مفعول به عاقل، ويقبل كل الأزمنة.
- 5- الفعل المتعدي لمفعولين (الأول عاقل): [أعطى]، وهو ما يقبل مفعولين أولهما عاقل.
- 6- الفعل القلبي (متعدي عاقل): [ظنّ]، وهو الفعل الذي تقبل دلالاته أن يعود على فاعله مثل ظننتني.



الشكل 1 جدول التوافق بين السوابق وإسناد الفعل إلى الأزمنة

السوابق	أنا	نحن	أنت	أنتما	أنتم	أنتن	هو	هي	هما مذكر	هما مؤنث	هم	هن
أنا	*	*	*	*	*	*	*	*	*	*	*	*
نحن	*	*	*	*	*	*	*	*	*	*	*	*
أنت	*	*	*	*	*	*	*	*	*	*	*	*
أنتما	*	*	*	*	*	*	*	*	*	*	*	*
أنتم	*	*	*	*	*	*	*	*	*	*	*	*
أنتن	*	*	*	*	*	*	*	*	*	*	*	*
هو	*	*	*	*	*	*	*	*	*	*	*	*
هي	*	*	*	*	*	*	*	*	*	*	*	*
هما مذكر	*	*	*	*	*	*	*	*	*	*	*	*
هما مؤنث	*	*	*	*	*	*	*	*	*	*	*	*
هم	*	*	*	*	*	*	*	*	*	*	*	*
هن	*	*	*	*	*	*	*	*	*	*	*	*

الشكل 2 جدول التوافق بين اللواحق وإسناد الفعل إلى الضمائر

### 3.2 معالجة الأسماء والحروف

بعد الانتهاء من معالجة الأفعال لصعوبتها، انكب المشروع على دراسة مشتقات الأفعال بأصنافها، اسم مفعول، اسم فاعل، مبالغة، أسماء التفضيل وصفات مشبهة. ومن ثم تصنيفها وتوليد الهيئات الصرفية الممكنة منها (المؤنث، المثني وجمع السالم) حسب القواعد العربية المعروفة. ووضعنا في القاموس مداخل تحتوي على الكلمة في صيغة المفرد المذكر أو جمع التكسير. واستثناءً على هيئة المفرد المؤنث أو جمع السالم.

### 3.3 معالجة الكلمات المستبعدة

الكلمات الشائعة (stop words) أو الكلمات المستبعدة هي الكلمات التي تستبعد قبل المعالجة اللغوية الآلية للبيانات (نصوص). وهي الكلمات التي تتكرر في النصوص مثل (في، من، إلى،...) ويستحسن تجاهلها عدم فهرستها من أجل تحسين البحث. ويرجع الفضل لهانس بيتر لوهن (من الرواد في استرجاع المعلومات) في استعمال هذا المصطلح والمفهوم في التطوير.

وتعتبر. هذه الطريقة مقارنة سلبية في التعامل مع الأدوات وحروف المعاني في الجمل. لا توجد قائمة كلمات محددة للاستبعاد، تستعملها كل المحلات الصرفية وأدوات معالجة اللغة الطبيعية. لا تستعمل كل أدوات المعالجة اللغوية قائمة كلمات شائعة. ويمكن للتحليل الصرفي التقليل من استخدام كلمات مستبعدة، لاسيما عند تحليل الجمل.

الكلمات المستبعدة قد تسبب مشاكل عند استخدام محرك بحث للبحث عن العبارات التي تدخل فيها، لاسيما في أسماء مثل 'المفعول به'، '، أو 'رسول من الله'، 'وهم'.

يعتمد الكثيرون من الباحثين [11، 12] عن المستبعدات على جردها جردا يدويا، ومن ثم توليد كل الأشكال الممكنة للكلمة وإدراجها في القائمة، مما يزيد في طول القائمة.

لكن يوجد اختلاف في اعتبار الكلمة مستبعدة، فمثلا يمكن أن تعتبر أسماء الأعداد مستبعدة في نص ما، لكنها تصبح ذات قيمة في نص اقتصادي، وكلمات مثل أسماء الشهور، لها أهمية في النصوص التاريخية، لذا توجب أن يتم تصنيف الكلمات، يمكن للمبرمج فيما بعد أن يحدد ما هي الكلمات التي يريدونها.

كما قمنا بوضع قائمة أساسية [10] مصنفة للكلمات (الشكل 3)، ومن ثم توليد (الشكل 5) كل الأشكال الممكنة (الشكل 4)، وذلك حسب التصنيف، فمثلا حروف الجر يمكن أن تتصل بالضمائر، مثل (إلى، إليه، إليهما، إليهم، إليكم...) بينما لا تتصل أسماء الأفعال بالضمائر.

كما يلجأ المبرمجون إلى إدراج الأشكال المنتظمة للكلمات مثل توحيد شكل الهمزة والتاء المربوطة، وألفات، والألف المقصورة، وهذا يمكن أن يتم برمجيا. ابتداء من القائمة الأساسية.

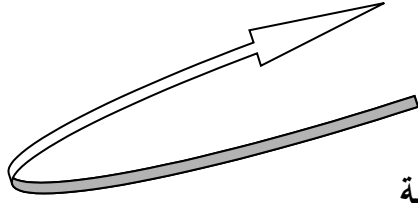
الكلمة	نوعها	صنف						
بيد	حرف	استثناء	ع	*	*	*	*	*
سوى	اسم	استثناء	ع	*	*	*	*	*
غير	اسم	استثناء	ع	*	ل	*	ض	س
لاسيما	حرف	استثناء	ع	*	*	*	*	*
أنى	اسم	الاستفهام	ع	*	*	*	*	*

الشكل 3 جدول تصنيف الكلمات المستبعدة وزواندها

بيد	بيد
وبيد	وبيد
فبيد	فبيد
سوى	سوى
وسوى	وسوى
فسوى	فسوى
غير	غير
ب-غير	بغير
ك-غير	كغير
ل-غير	لغير

	فيغير	ف-ب-غير
	وكغير	و-ك-غير
	فكغير	ف-ك-غير
	ولغير	و-ل-غير
	فلغير	ف-ل-غير
	ولغير	و-ل-غير
	فلغير	ف-ل-غير
	أغير	أ-غير
	أبغير	أ-ب-غير
	أكغير	أ-ك-غير

الشكل 4 الأشكال المولدة من الكلمات المستبعدة



## الكلمات المستبعدة مصنفة

الكلمة	نوعها	صنف	الطلب	التعرف	التر	الإضافة	الاستخدام	المصرف	إلى التسم	معر	ملاحظة
بيد	حرف	استثناء	ع	ع	ع	ع	ع	ع	ع	ع	ع
موى	اسم	استثناء	ع	ع	ع	ع	ع	ع	ع	ع	ع
غير	اسم	استثناء	ع	ع	ع	ع	ع	ع	ع	ع	ع
لاسمما	حرف	استثناء	ع	ع	ع	ع	ع	ع	ع	ع	ع
أش	اسم	اسم الاستثناء	ع	ع	ع	ع	ع	ع	ع	ع	ع

بيد  
ويبد  
فبيد

تدقيق وإثراء

وسوى	وسوى
فسوى	فسوى
سوى	سوى
غير	غير
ب-غير	بغير
ك-غير	كغير
ل-غير	لغير
ل-غير	لغير
و-غير	وغير
ف-غير	فغير

الأشكال المولدة

الشكل 5 توليد الأشكال المختلفة للكلمات المستبعدة

العدد	
449 كلمة	الكلمات المستبعدة المصنفة
13015 كلمة	كل الأشكال المولدة

### 3.3.1 تطبيق الكلمات المستبعدة في التدقيق

تعتبر المستبعدات عادة من الكلمات الشائعة، لذا رأينا أن تدرج الأشكال متنوعة مباشرة في القاموس دون تطبيق قواعدها، وذلك للتخفيف على البرنامج وتسريع التدقيق دون الحاجة إلى تحليل الكلمات. كما أنّ أي إثراء لمشروع الكلمات المستبعدة، أو التدقيق الإملائي، يمكن أن يثري المشروعين معاً.

## 4 تقييم نتائج المرحلة الأولى

في نهاية المرحلة الأولى، أطلقنا الإصدار النهائي الأول لقاموس التدقيق الإملائي، في بداية سنة 2008 دام العمل على تدقيق الأفعال فترة طويلة جداً، وعددا هائلا من مجموعات القواعد، لكن معظم هذه المجموعات متشابهة جداً، مما يؤدي إلى بطء برنامج التدقيق. نظرا لكثرة الخيارات والقواعد التي يمكن تطبيقها.

أضف إلى ذلك خاصية تراكب الزوائد في العربية إلى عدة مستويات مثل فسيفعلونها (ف-س-ي- فعل-ون-ها)، بينما لا يدعم برنامج هانسبال [7][8] إلا مستويين اثنين في السوابق واللواحق..

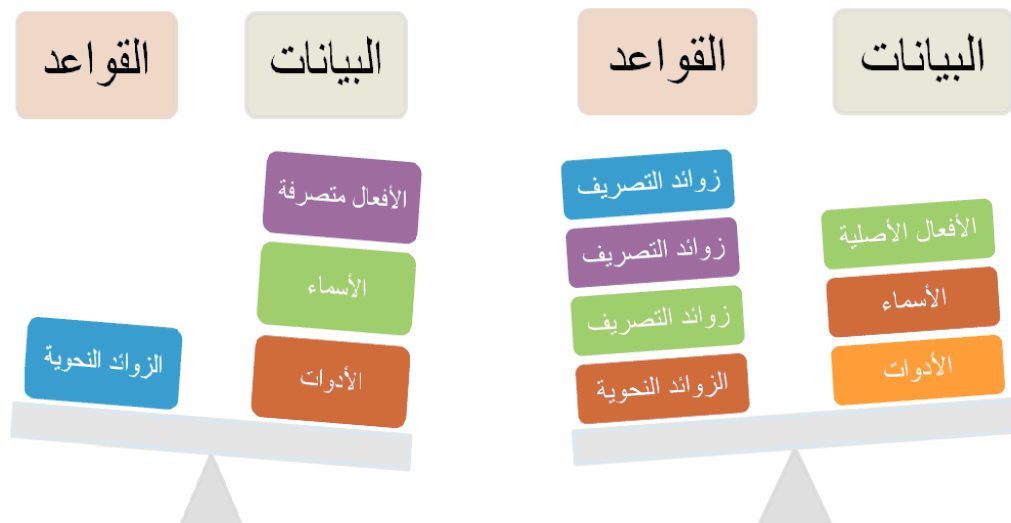
كل هذا تسبب في :

- صعوبة تصريف فعل ما، بل يجب دراسته أولاً لمعرفة منوال فعل ما، مما يتطلب عملاً يدوياً جهيداً.
- تعقيد تمثيل التغييرات الداخلية التي تطرأ على الفعل المعتل والمهموز. أدى إلى كثرة القواعد وتشابهاها [3][4].

لذا فكرنا في إعادة بناء القاموس الإملائي والموازنة بين حجم البيانات وتعقيد القواعد. وهو ما سنتحدث عنه في الفقرة الموالية.

#### 4.1 بناء قاموس التدقيق الإملائي بين حجم البيانات وحجم القواعد

إنّ القاموس الخاص بالتدقيق يبني على قسمين أساسيين هما قائمة المفردات وقائمة القواعد، ويحدث بينهما تجاذب في الحجم والقوة والتعقيد، وهذا ما استقيناه من خبرتنا في بناء الإصدار الأول، فإذا كانت القواعد بسيطة سهلة كانت قائمة المفردات متعددة المداخل لمفردة واحدة، مما يصعب صيانة قائمة المفردات، أمّا إذا كانت القواعد أكثر تعقيداً شاملة مختلف التغييرات صعبت صيانة قائمة القواعد وأخذت حجماً كبيراً، بينما يكون القاموس أبسط في عدد مفرداته، أعقد في وسوم تلك المفردات، وقد استخدمنا هذه الطريقة في الإصدار الأول، وتجلت عيوبها في بطء توليد الاقتراحات والتعرف على الأخطاء لكثرة القواعد وتعقيدها. ونبين فيما يلي شرحاً للطريقتين ولما لهما وما عليهما (الشكل 6).



## الشكل 6 بناء القاموس التدقيق بين حجم البيانات وتعقيد القواعد

### 4.1.1 تخفيف حجم البيانات على حساب تعقيد القواعد

إنّ الحفاظ على بنية مقروءة للقاموس، وصيانة سهلة له وتسهيل إثرائه بالمفردات، يتطلب تقليل حجمه، وعدم تعدد مداخله للكلمة الواحدة، مما يعقد صياغة قواعد الزيادة وضبطها ضبطاً جيداً. وقد لاحظنا هذا في المرحلة الأولى، من خلال رغبتنا في توحيد مداخل القاموس، فأصبحت قواعد الزيادة معقدة جداً، حتى صارت صيانتها صعبة على مصممها، مستحيلة على غيره.

### 4.1.2 تبسيط القواعد على حساب حجم البيانات:

تستعمل هذه الطريقة لزيادة سرعة البحث والتصحيح، مع وجود مساحة كبيرة، ويمكن وضعها على خادم ويب، أو جهاز قوي ذي ذاكرة كافية، ويمكن في هذه الحالات تبسيط القواعد إلى شكل محدود بسيط جداً، بينما يحتوي القاموس على كلمات أصلية وكلمات مركبة مثل الأفعال في أشكالها المتصرفة، فنضع مثلاً قواعد إلحاق الزوائد التركيبية للأفعال فقط، بينما نضع الأفعال المتصرفة جميعها في القاموس.

**محاسن:** تقليل عدد القواعد يزيد سرعة التدقيق، ويمكن استعماله في الأوساط التي تتطلب استجابة سريعة، وبها مساحة كافية. بينما يلقي على عاتق برنامج تصريف الأفعال توليد المداخل المتعددة للفعل في القاموس.

**مساوئ:** كثرة المداخل الخاصة بكلمة واحدة يصعب صيانة قائمة المفردات،

ولتجاوز هذا العائق نقترح أنه يمكن توليد البيانات بواسطة برامج خاصة مرة واحدة، ومن ثم العمل عليها بواسطة قواعد مبسطة. ويمكن عمل ذلك من خلال برنامج قطرب لتصريف الأفعال [5][9]. كما أنّ برنامج هانسبال يوفر خاصية لضغط ملف القاموس، مما يقلل من حجمه.

وبالمقارنة بين الطريقتين، نعتقد أنّ توافر البيانات والبرامج المناسبة للتحويل من قاموس مجرد يمكننا من استعمال الطريقة الثانية لتبسيط القواعد على حساب البيانات، وأيضاً توليد العديد من الصيغ المناسبة لمختلف برامج التدقيق الإملائي.

## 5 تسريع التدقيق الإملائي

نسعى من خلال هذا التحسين إلى تسريع التدقيق الإملائي، بتخفيف الضغط على برنامج التدقيق وتقليل العمليات التي يجريها. للوصول إلى الكلمة الصحيحة، من خلال مضاعفة عدد مداخل الكلمة في القاموس، ونخص بالاهتمام الأفعال، لما يعترها. من تصريف يغيّر من بنيتها الداخلية.

لذا فكرنا في البحث في مسألة تصريف الأفعال، من أجل فهم الآلية التي تتصرف بها الأفعال العربية، وصولاً إلى تسهيل التعرف عليها. وقد استمرّ بحثنا عدة سنوات ليثمر برنامج قطرب لتصريف الأفعال العربية، والذي لا يعد مجرد برنامج للتصرف الآلي، بل خوارزمية جديدة كلياً

للتصريف بسيطة للغاية، وسنطبق هذا البرنامج في مجال التدقيق الإملائي، بعد أن نعرف التصريف الآلي، وطرائقه.

## 5.1 خوارزمية قطرب لتصريف الأفعال

من أجل تصريف الفعل العربي آليا، نعتقد أن الأمر يتعلق بتصريف الفعل صوتيا. وليس إملائيا. لأن الأصل في اللغة الأصوات والكتابة الإملائية صورة، قد تكون غير مكتملة للأصوات. كما أنّ التصريف العربي يخضع لقواعد صوتية ونغمية أكثر من القواعد الكتابية.

فالعربي عند تصريفه الفعل فإنه يستعمل قواعد صوتية وليس إملائية ومن ذلك أن الفعل المهموز لا يختلف عن غيره في التصريف عند إغفال كتابته الإملائية. كما أن حرف ألف الوصل لا تعتبر عند التصريف إلا من حال منعه للابتداء بساكن. وقد كتبنا مقالا مفصلا عن الخوارزمية في مقال مستقل في. وما سننظر في هذا المقال هو تطبيق التصريف الآلي في التدقيق الإملائي [5]

[9]

وقد صمنا برنامج قطرب لتصريف الأفعال العربية أسميناه تيمنا بالعالم النحوي أحمد بن المستنير. المكنى بقطرب لأنه أول من ألف رسالة مستقلة في الأفعال. وتم برمجة التطبيق بلغة بايثون، التي تساعد في توفير إصدارات مختلفة واجهات متعددة للبرنامج على أوضاع مختلفة:

- تطبيق الويب على الموقع <http://qutrub.arabeyes.org> في الشكل 7.
- تطبيق مكتبي يعمل على نظام وندوز. ولينكس موضح في الشكل رقم 4 متوفر للتحميل على الموقع نفسه.

والبرنامج متوفر بترخيص مفتوح المصدر، ويمكن الاستفادة منه في العديد من التطبيقات، في معالجة اللغة العربية، وتعليمها. وسنعمل على تطبيقه في مجال التدقيق الإملائي.

الرئيسية | توثيق | روابط | تحميل | اتصل بنا | من هو قطرب؟

# قطرب

تصريف الأفعال العربية

الفعْل

كل الأمانة  متعدي  خيارات أكثر << [مساعدة]

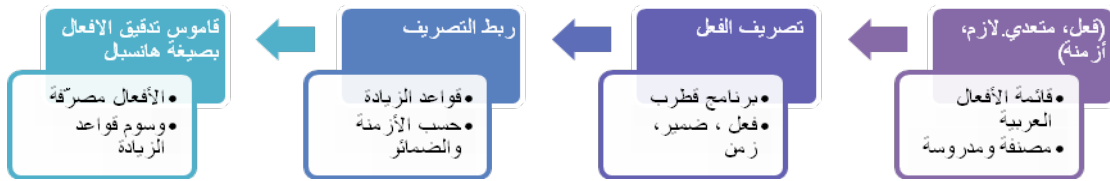
أخبار: إصدار جديد على نظام وندوز، للتحميل

للاتصال المجموعة البريدية بدعم من التتبعين العرب arabeyes.org

الشكل 7 تطبيق قطرب على الويب <http://qutrub.arabeyes.org>

## 5.2 تطبيق التصريف الآلي في التدقيق الإملائي

بعد كل ما عايناه من جهد مبذول لضبط القواعد المختلفة للتصريف حسب برنامج هانسبال للتدقيق الإملائي، أردنا أن نركز على الزوائد النحوية فقط مثل حروف العطف والضمائر. المتصلة في محل نصب مفعول به، بينما نمثل الزوائد التصريفية من خلال تعدد المداخل في القاموس، بينما يوئد المصرف قطرب الحالات المتعددة للفعل، ويربطها. بما يوافقها من زوائد حسب جدول توافق السوابق واللواحق (الشكل 1 والشكل 2).



### الشكل 8 مخطط توليد قاموس التدقيق للأفعال

وقد استفدنا من جميع البيانات الموضحة أعلاه المستعملة في المرحلة الأولى، إذ يستغل المصرف بيانات تصنيف الأفعال (التعدي وأنواعه، والأزمنة المحددة للفعل)، لتوليد التصريفات الصحيحة إملائيًا، ودلاليًا، للأفعال المصنفة.

وتكون النتيجة عبارة عن الفعل مسندا إلى ضمير. معين، في زمن معين، وباعتبار الضمير. يمكننا تحديد قائمة اللواحق التي تتصل به، مع اعتبار تعديده إلى العاقل أو غير العاقل، وكذا كونه فعلا من افعال القلوب، وذلك حسب الشكل 2، كما يمكننا تحديد السوابق حسب الزمن الذي يقع فيه الفعل حسب الشكل 1. وتسمى هذه المرحلة ربط التصريف بقواعد الزيادة.

وفي الأخير نحصل على قاموس تدقيق الأفعال العربية (الشكل 8)، فندمجه مع الأسماء والحروف لنحصل على القاموس النهائي للتدقيق الإملائي. ويوفر برنامج التدقيق هانسبال يوفر آلية لضغط حجم وسوم القواعد المطبقة على مداخل القاموس، مما يقلل حجمه.

## 6 تقييم عمل المدقق

من أجل تقييم الإصدار. الثاني بالنسبة للإصدار. الأول نعتبر جانبين هما:

**حجم القاموس:** والذي نلاحظ أنه تضاعف في الإصدار الثاني، وازداد عدد المداخل للفعل الواحد إلى 38.54 مدخلا، بينما كان 1.23 في الإصدار الأول، ولكن على الرغم من تخوفنا. في البداية من الحجم الكبير لملف البيانات، إلا أنه تضاعف أربع مرات فقط، وهو ما أزال مخاوفنا. من الحجم.

الإصدار الأول	الإصدار الثاني	
10489	10489	عدد الأفعال
12903	404262	عدد مداخل الأفعال في القاموس
1.23 مدخل للفعل	38.54 مدخل /الفعل	معدل المداخل للفعل الواحد
2167 كيلوبايت	6914 كيلوبايت	حجم الملف
13689 سطر	2278 سطر	عدد الأسطر في ملف القواعد

**سرعة التدقيق:** بما أن عملنا انصب على تسريع المدقق، فإننا نأخذ عامل السرعة بعين الاعتبار أولا، فقد جربنا المدقق في إصداره الأول على عينة فيها 100000 كلمة، ليتم تدقيقها خلال 5دقائق. و 16 ثانية دقائق. بمعدل 318 كلمة في الدقيقة، بينما تمكّن المدقق في إصدار الثاني من تدقيق 1.4 مليون كلمة في 1.20 دقيقة، فلا مجال للمقارنة.

تتكون العينة من 1.44 مليون كلمة غير مكررة، تم جمعها من مكتبة ثواب التي تحتوي على مئات الكتب العربية. قائمة الكلمات مرتبة ترتيبا تنازليا حسب تواتر الكلمات في النصوص. وقد جربنا الإصدار الأول مع المئة ألف كلمة الأولى، لذا تحصلنا على نسبة تعرف جيدة تصل إلى 96% من الكلمات.

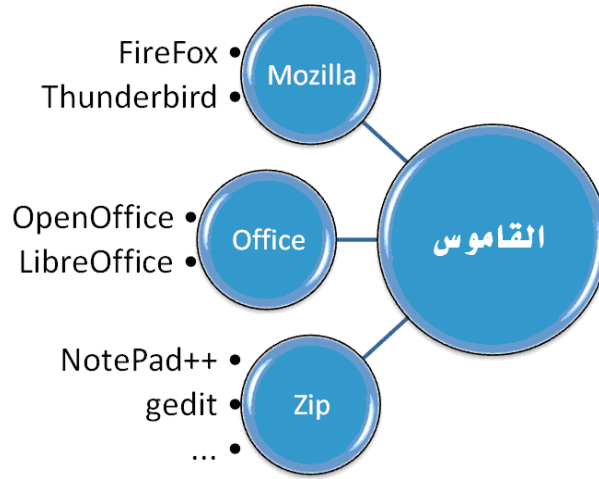
أما الجهاز الذي جربنا عليه العينة، فهو معالج أنتل سيليرون 2.4 جيجا هرتز، بذاكرة حية قدرها 1 جيجا بايت، على نظام أوبنتو. 10.04.

المعيار	الإصدار الأول	الإصدار الثاني
عدد الكلمات في العينة	100000	1444516
زمن التدقيق	5د 16 ثانية	1د و 17 ثانية
نسبة التعرف على الكلمات الصحيحة	96%	84.40%
عدد الكلمات الخاطئة	3233 كلمة	225303
سرعة التدقيق	318 كلمة/الثانية	18760 كلمة /الثانية

أما نسبة التعرف فكانت 85% من الكلمات التي تم التعرف عليها من مدونة فيها 1.4 مليون كلمة غير مكررة. وسنعمل على دراسة النسبة الباقية من الكلمات غير المعروفة، لإثراء قاموس التدقيق.

## 7 تطبيقات القاموس

وبعد الانتهاء من بناء القاموس، نحزمه في حزم خاصة بالبرامج الحرة (الشكل 9)، وعلى راسها برنامج الكتابة الحر openOffice.org، وشقيقه LibreOffice، وكذلك برنامج التصفح Mozilla Firefox و برنامج البريد الإلكتروني Mozilla Thunderbird، وتوفير القاموس في صيغة عامة يمكن أن تستفيد منها برامج أخرى مثل Gedit.



الشكل 9 تطبيقات القاموس في البرامج الحرة

## 8 تصميم موقع خدمة التدقيق الإملائي

بعد اكتمال العمل على الإصدار الأول والثاني قاموس التدقيق العربي، نفكر في تصميم موقع للتفاعل مع المستخدم (الشكل 10)، لإفادة المستخدم، وتمكينه من المساهمة في تطوير القاموس من خلال إضافة المفردات الناقصة، والإبلاغ عن الأخطاء، و

### - الخدمات المقدمة

- خدمة التدقيق الإملائي باللغة العربية.
- واجهة برمجية لتضمين خدمة التدقيق في المواقع الأخرى.

○ عدم الحاجة لتنصيب المدقق على الجهاز أو المتصفح، ويفيد في حالات استخدام أجهزة الغير أو العامة.

○ خدمة التصحيح التلقائي، للكلمات مثل همزات الوصل والقطع، وبعض الأخطاء الشائعة التي يمكن تصحيحها تلقائياً.

#### - مساهمة المستخدم

○ اقتراح مفردات جديدة لا يتعرف عليها المدقق حالياً.

○ الإبلاغ عن الأخطاء في القاموس.

○ جمع الكلمات الخاطئة لدراستها.

○ دراسة شيوع الكلمات المستخدمة.

#### - المسائل التقنية

○ تطوير واجهة وموقع تعمل بتقنية ajax وتتفاعل مع المدقق الإملائي هانسبال.

○ يمكن تطبيق هذا الموقع على لغات أخرى.

○ تصميم برامج معالجة البيانات المدخلة وتحويلها إلى صيغ خاصة بالتدقيق الإملائي.

○ تصميم واجهة لتلقي مساهمات المستخدمين.



الشكل 10 موقع خدمة التدقيق وتفاعل المستخدمين

## 8.1 إثراء قاموس التدقيق

نضع تحت تصرف المستخدم، والمطورين، نموذجاً للبيانات التي يمكن إدخالها لإضافة كلمة، للمساهمة في إثراء القاموس. ويمكن تقسيمها إلى قسمين أساسيين الأول للأفعال، والثاني للأسماء. ويمكن أن يكون هذا التفاعل نواة لبناء معجم تفاعلي عربي حر، على أن تخضع قرارات قبول المفردات في قاموس التدقيق للجنة من الخبراء.

### 8.1.1 اقتراح مادة جديدة في الأفعال

نضع تحت تصرف المستخدم، والمطورين، نموذجاً للبيانات التي يمكن إدخالها لإضافة كلمة، للمساهمة في إثراء القاموس.

إذا كان المستخدم لا يعرف كل التفاصيل يمكن أن ينبّه فقط للفعل. بعدها يتم التحقق من البيانات، وعند صحتها يمكن توليد مداخل للكلمة الجديدة في قاموس التدقيق الإملائي.

الفعل	كتابة الفعل مشكولاً
الجزر	تحديد جذر الفعل
نوع الفعل	ثلاثي
	غير ثلاثي
التعدي/ اللزوم	متعدي
	متعدي لعاقل
	متعدي لغير العاقل
	متعدي لمفعولين أولهما عاقل
	متعدي لمفعولين أولهما غير عاقل
	لازم
	مطلق/ نسبي
	متعدي بحرف
الأزمنة	من إلى في ب ل عن
	كل الأزمنة
شرح	بعض الأزمنة
	الماضي المضارع الأمر المبني للمجهول
درجة الشبوع	شرح المفردة
	شائع، غير شائع، نادر
نوع النص الوارد فيه	ديني
	قرآن، سنة
	تراثي
	شعر جاهلي، نصوص قديمة

معاصر.	صحف، مجلات، أدب، علوم
--------	-----------------------

### 8.1.2 اقتراح مادة جديدة في الأسماء

نضع تحت تصرف المستخدم، والمطورين، نموذجاً للبيانات التي يمكن إدخالها لإضافة كلمة، للمساهمة في إثراء القاموس.

إذا كان المستخدم لا يعرف كل التفاصيل يمكن أن ينبّه فقط للفعل. بعدها يتم التحقق من البيانات، وعند صحتها يمكن توليد مداخل للكلمة الجديدة في قاموس التدقيق الإملائي.

الاسم	كتابة الاسم مشكولا
الجذر	تحديد جذر الاسم إذا كان مشتقا
نوع الاسم	اسم علم
	اسم علم أجنبي
	جامد
مشتق	اسم فاعل، اسم مفعول، اسم المرة، مصدر صناعي، مصدر ميمي، اسم زمان، اسم مكان... صيغة مبالغة
العدد	حالته الأصلية
	مفرد
	مثنى
	جمع
تذكير/تأنيث	مذكر
	مؤنث
ممنوع من الصرف	يقبل
شرح	شرح المفردة
	شائع، غير شائع، نادر
درجة الشبوع	الإفراد، الجمع
نوع النص الوارد فيه	مذكر
	مؤنث
	ديني
تراثي	الإفراد، التثنية
	شعر جاهلي، نصوص قديمة
معاصر.	صحف، مجلات، أدب، علوم

يمكن استعمال هذه البيانات بعد تدقيقها ودراستها. في توليد المداخل اللازمة لإثراء قاموس التدقيق، وإطلاق إصدارات محدثة.

## 9 الخاتمة

بعد الحاجة الملحة للتدقيق الإملائي في عالم المصادر المفتوحة، وعلنا الحثيث على إصدار القاموس الأول، وانتشار استعماله، بعد كل هذا بدأنا التفكير بنأ في تحسين المدقق وإصلاح ما صادفناه من مشاكل وصعوبات. في الإصدار الأول. وقد تضمن المشروع ثلاثة أقسام قسم للأفعال وهو الأكبر والأصعب لما يعترض الأفعال من تصاريح مختلفة وتغييرات على أطرافها. ووسطها، أما الأسماء والحروف، فكانت أقل جهداً.

وقد كانت مرحلة الأفعال أصعب مرحلة لما يطرأ عليها من تغييرات داخلية (غير مدعومة حالياً من برنامج التدقيق الذي يدعم السوابق واللواحق فقط). مما اضطرنا إلى العمل على خوارزمية خاصة لتصريف الأفعال العربية، توصلنا من خلالها إلى تصميم برنامج قطرب، في عدد من الأشكال، كموقع وب، وتطبيق مكتبي يعمل على أنظمة تشغيل مختلفة.

وقد قمنا باستعمال خوارزمية قطرب للتصريف الآلي، لتوليد تصاريح الفعل المختلفة، ومن ثم إدراجها في قاموس التدقيق الإملائي، وبالتالي تقليل عدد قواعد الزيادة إلى الزيادات النحوية فقط، مما أدى إلى تسريع المدقق الإملائي لينتقل من 21 كلمة في الثانية إلى 24000 كلمة في الثانية.

من جهة أخرى نريد العمل على تطوير موقع خدمي تفاعلي يقدم خدمة التدقيق الإملائي، وما يتعلق بها من خدمات مرافقة، مثل التصحيح التلقائي للنصوص، وتمكين المواقع من ربط هذه الخدمة إليها، كما نمكّن المستخدمين من المساهمة في إثراء القاموس والاستفادة بعد ذلك من الإصدارات الجديدة. كما يفيد الموقع في رصد الكلمات الخاطئة والأخطاء الشائعة، وتواتر الكلمات المستخدمة.

## المراجع

[1] إسماعيل حجير، "من أجل مدقق إملائي مفتوح المصدر". مقدم لنيل درجة الماجستير في علوم اللسان و التبليغ اللغوي . فرع العلاج الآلي للغة، مركز البحوث العلمية والتقنية لترقية اللغة العربية 2009.

[2] مشروع القاموس العربي للمدقق الإملائي هانسبال  
<http://ayaspell.sourceforge.net>  
[البواب، 1996]:

- [3] طه زروقي، محمد كبداني، "مشروع أية -سبل، القاموس العربي للتدقيق الإملائي مفتوح المصدر، آفاق المرحلة القادمة". ورشة إثراء المحتوى الرقمي العربي، المنظمة العربية للتربية والثقافة والعلوم، وزارة التعليم العالي سوريا، دمشق 16-19 أكتوبر. 2010.
- [4] طه زروقي، محمد كبداني، "مشروع أية -سبل، القاموس العربي للتدقيق الإملائي مفتوح المصدر، تجربة وآفاق". اجتماع الخبراء، مشروع المعجم العربي التفاعلي، المنظمة العربية للتربية والثقافة والعلوم، دمشق 26-28 مايو 2009
- [5] طه زروقي، عمار بالة، "برنامج قطرب لتصريف الأفعال العربية: خوارزمية جديدة مبسطة"، الندوة الدولية لعلوم وهندسة الحاسوب السادسة، ياسمين الحمامات، تونس 19-21 ماي 2010.
- [6] Ammar S. Dichy J. Al-Shamil, in Arabic verb Conjugation, Collection Bescherelle, Bescherelle (Paris, France : 1997) Hatier, 1999
- [7] Taha Zerrouki, Amar Balla, "Implementation of infixes and circumfixes in the spellcheckers", 2nd International Conference on Arabic Language Resources and Tools, Cairo (Egypt), 22 - 23 April 2009
- [8] طه زروقي، الجوانب البرمجية لدعم العربية في المدقق الإملائي مفتوح المصدر هانسبال. الأيام الدراسية حول المعالجة الآلية للغة العربية، المركز الجامعي بشار، ماي 2007.
- [9] طه زروقي، عمار بالة، "برنامج قطرب لتصريف الأفعال العربية، أداة تعليم التصريف"، ندوة منهاج تعليم العربية وتعلمها: الاختيارات النظرية وكيفية تطبيقها، مركز البحث في ترقية اللغة العربية وترقيتها، الجزائر 17-18 ماي 2010.
- [10] أنطوان الدحداح، معجم قواعد اللغة العربية، في جداول ولوحات، مكتبة لبنان ناشرون، 2008.
- [11] Leah S. et al Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis, Univ. of

Massachusetts, Dept. of Computer Science, 2002  
Ibrahim Abu El-Khair, Effect of stop words elimination for [12]  
Arabic information retrieval: a comparative study, Dept of  
library and information science, faculty of arts, Minia  
University-Egypt, iabuelkhair@gmail.com, international journal  
of computing & information sciences, vol. 4 No. 3 December  
.2006, On-Line