القاموس العربي للتـــدقيق الإمـلائي

HUNSPELL-AR

لمشروع آيسبل

***

The arabic spell-checker dictionary

from Ayaspell project

***

ذ. محمد كبداني

27-01-2008

البطاقة التقنية

الإنجاز	القاموس العربي للتدقيق الإملائي **Hunspell-ar**
التحميل	**« Sources »:** http://sourceforge.net/project/showfiles.php?group_id=205373 http://forge.aaul.net/projects/ayaspell/ **« Binaries »:** Debian (Ubuntu): http://packages.debian.org/fr/sid/hunspell-ar Mandriva: http://rpmfind.net/linux/RPM/mandriva/2008.1/i586/media/main/release/myspell-ar_AR-1.0.2-19mdv2008.1.noarch.html Fedora (Red Hat): http://rpmfind.net/linux/RPM/fedora/devel/i386/hunspell-ar-0.20080110-1.fc10.noarch.html Firefox (Addon xpi): https://addons.mozilla.org/fr/firefox/addon/3677 Openoffice.org (Extension oxt): Not yet (if you are interested to build it, contact us please, thanks)
المواقع	http://ayaspell.sourceforge.net/ http://ayaspell.sourceforge.net/ar.html
القائمـــــة البريدية	http://groups.google.com/group/ayaspell-dic
المدونة	http://ayaspell.blogspot.com/
الدعـــــم والجوائز	جائزة الشيخ علي جابر العلي السالم الصباح (الكويت) للبرامج الحرة Cheikh Ali Jabir Al-ali Assalim Assabah (Kuwait) **Award for FLOSS (2007)** -------- الجائزة الخاصة خلال الملتقى الإفريقي الثالث للبرامج الحرة **Prix spécial des troisièmes rencontres africaines du Logiciel Libre** http://rall.logiciels-libres.org/rubrique.php3?id_rubrique=6 -------- مشروع السنة حسب الجمعية المغربية لتنمية الإعلاميات الحرة **Projet Open Source de l'année (2007), Association ADIL (Maroc)** http://81.192.48.26/linuxmaroc/modules.php?name=News&file=article&sid=143 -------- كما يحظي مشروع آيسبل بدعم مدينة الملك عبد العزيز للعلوم والتقنية (المملكة العربية السعودية) مشكورةً. http://www.kacst.edu.sa/index.php

بســـم الله الرحمـــان الرحـــــيم

الحـــيثيات

إن أهمية المدققات الإملائية **[1]** في مجال المعالجة الرقمية للغات الطبيعية من الأمور التي استرعت انتباه المهتمين بالحلول المكتبية منذ البدايات الأولى للإعلاميات، مما حفز الشركات على العمل على إنتاج هذه الأدوات والمعينات المعلوماتية - ذات المصدر المغلق في أغلبها - بوظائف متطورة أكثر فأكثر، تلبي حاجة مستعملي الحاسوب في المجال المكتبي بشكل خاص وتغني خدمات موزّعي البرامج المعلوماتية التي تدفع الملايين من أجل الحصول عليها (يناهز ثمن المدقق الإملائي العربي **[2]** المليون دولار أمريكي في السوق العالمية لسنة 2006) ...

على صعيد البرامج الحرة وإلى حدود 2006، لم يكن هناك أي مدقق إملائي عربي حر وظيفي، رغم تعدد المحاولات العربية المرتبطة بطريقة مباشرة أو غير مباشرة بمؤسسة عربآيز **[3]** Arabeyes أهمها محاولتي الأخوين محمد الزبير ببرنامج "دؤلي" **[4]** Duali ومحمد سمير ببرنامج " بغداد" **[5]** Baghdad . تأخر الحصول على دعم لغة الضاد في البرامج الحرة بشكل عام وافتقادها لمدقق إملائي بشكل خاص يرجع أساساً إلى تميزها بخصائص برمجية ولغوية معقدة نسبياً، ندرة الكفاءة المختصة وضعف الاهتمام بالبرامج الحرة في المنطقة شعبياً واقتصادياً وجامعياً. في آخر المطاف جاء الحل عبر بوابة البرامج الحرة بالتأكيد: برنامج التدقيق الإملائي هانسبل **[6]** Hunspell المعتمد من قبل مشروع الديوان المفتوح أوبن أفس OpenOffice.org **[7]** ومن برنامج أسبل **[8]** Aspell. البرنامجان مطوران أصلاً للغات اللاتينية ولكن بعد إضافة خاصية اليونيكود ودعم ثنائية الاتجاه إليهما أصبحا مؤهلين لدعم اللغات غير اللاتينية من ضمنها اللغة العربية...

بعد حصول دعم اللغة العربية في هذين البرنامجين - هانسبل وأسبل - ظهر للمهتمين بالشأن المعلوماتي الحر، تحدي آخر هو توفير القواميس العربية الخاصة بالتدقيق الإملائي والتي بدونها لن تؤدى هذه البرامج وظيفتها. لم يكن في الساحة إلاّ قاموساً عربياً واحداً حرّاً هو قاموس تيم بوكولتر **[9]** Tim Buckwalter المعتمد في بنائه على مكنز لغوي مكون من مادة صحفية أساساً. للأسف، كان الباحث المطور تيم بوكولتر جاهلا للغة العربية وكانت المادة الصحفية المرجعية غير مدققة لغوياً، فترتّب عن ذلك احتواء القاموس على مفردات خاطئة في نسبة كبيرة منها رسماً ولغةً مما أثر سلباً على المدققات الإملائية المعتمدة عليه، وجعلها لا ترقى إلى المستوى المنتظر منها وهذا مثل المدقق أربيك-سبل arabic-spell **[10]** ، منتوج شركة غوغل **[11]** Google، الذي يعطي نتائج جد رديئة، تدفع المستعمل إلى الاستغناء عنه منذ الوهلة الأولى.

أمام هذا الوضع، كان ولا بدّ من الاعتماد على قدراتنا الذاتية، واستثمار معرفتنا بلغة الضاد فأهل مكة أدرى بشعابها: أولاً تكييف المدقق الإملائي مع عادة المستعمل تجاهل الحركات في كتاباته باللغة العربية ثم بناء

قاموس عربي حر مناسب. من أجل بلوغ هذين الهدفين تمّ تأسيس مشروع بمواصفات مهنية حديثة تجمع كل شروط النجاح: موقع إنترنت **[12]** في نسخة أولى مؤقتة باللغة العربية ثم تلاه آخر رسمي بالإنجليزية **[13]** والعربية **[14]** ثم الفرنسية **[15]** حيث يجد المهتم آخر الأخبار ويستطيع تحميل الملفات ويطّلع على الوثائق اللغوية ويكون على علم بالمنجزات وقائمة بريدية **[16]** mailling-list حيث تناقش الاختيارات وتوضع الاقتراحات وتوضح الحلول التقنية وأخيراً مدونة **[17]** Blog حيث المقالات التي تنظر للمشروع ونصوص تفسر المقاربات وتشرح المنهجيات. كان الرهان هو تحقيق نتيجة بمستوى لا يقل قيمة عما يتداوله المستعملون لهذا الصنف من الأدوات المكتبية على الأنظمة المنغلقة وهكذا تمّ إنشاء القاموس العربي الحر للتدقيق الإملائي المفتوح Hunspell-ar أول منتوج مشروع آيسبل.

يدخل المدقق الإملائي هذا، في حقيقة الأمر، ضمن مشروع شامل، هو مشروع آيسبل Ayaspell project ، الذي يهدف توفير أدوات المعالجة الآلية للغة العربية (واللغة الأمازيغية **[18]** مستقبلاً إن شاء الله) في بيئة البرامج الحرة، بالإضافة للمدقق الإملائي، أدوات الترادف المعجمي Thesaurus **[19]**، التدقيق النحوي Grammar-checker **[20]** وقواميس الأنظمة (المندمجة) المحمولة Embedded systems **[21]** مثل الهواتف المحمولة وأجهزة PDA.

الخصائص الأساسية للقاموس

استدعى غياب قاموس عربي حر، بناء واحد يلبي شروط التدقيق الإملائي بالاعتماد على المعاجم اللغوية العربية المتداولة التراثية والحديثة. من هذه المعاجم معجم تصريف الأفعال العربية (مجموعة Bescherelle)، المعجم الوسيط، المعجم الغني، معجم المحيط ولسان العرب. هذه هي إذن الروافد المهيكلة لقاموس آيسبل الذي اصبح ثاني قاموس حر متوفر على الشبكة، حر بمعنى خضوعه للرخصة العمومية الشاملة [GPL].

تطلب إنشاء القاموس بشقيه (ملف DIC وملف AFF) أكثر من 1500 ساعة عمل على مدى قرابة سنتين من النشاط المتواصل (أبريل 2006 - يناير 2008) وتحليل آلاف المفردات من فعل واسم وأداة وحرف وتصنيفها وتوليدها حسب قواعد اللغة العربية النحوية والصرفية، ثم تحديد معناها لتمييز الفعل اللازم والمتعدي لعاقل أو غير عاقل والصفة العائدة على عاقل أو غير عاقل ومعرفة الشاذ منها والعادي. إجمالاً، تمت معالجة أكثر من 50.000 مفردة تتوزع على ما لا يقل عن 10.000 فعل عربي، 40.000 اسم وعشرات الحروف والأدوات النحوية وما استثني من هذا أو ذاك.

تجدر الإشارة إلى كون هذه النتيجة، حصلنا عليها باستثمار خصائص هانسبل Hunspell العادية فقط ولم نلجأ إلا لخاصية برمجية جديدة واحدة متمثلة في (Patch بهانسبل إصدارة 1.1.5) **[22]** تحت وظيفة IGNORE لتجاهل الحركات والتطويل (الكشيدة) في النص العربي المعالج من خلال تحويل الأخ طه زروقي،

الكود المتعلق بها المبرمج أصلاً في برنامج "دؤلي" (لغة بايتون Python) إلى برنامج هانسبل (C++). ما زالت هذه الخاصية في حاجة للتعديل **[23]** وإلى تحسين لأنها تأثر سلباً على نوعية الكلمات المقترحة في البديل الصحيح عندما تكون المفردة الخاطئة مشكولةً أو مُطولةً.

- **المُكَوِّن الفعلي**

مثلت معالجة الفعل العربي الشطر الأول من المشروع والجانب الأكثر استهلاكا للوقت واستدعت الوقوف على العشرات من المراجع اللغوية حيث يعتمد التدقيق الإملائي للفعل العربي للمشروع على مادة لغوية تحتوي على ما يفوق 10.000 فعل عربي وبعد إضافة الأشكال الخاصة بالإبدال/الإعلال والتضعيف/الإدغام وما يجري على الهمزة من تحولات، ارتفع عدد المفردات في قاموس آيسبل Ayaspell إلى ما يقارب 15.000 (14523 مفردة بالضبط).

بالنسبة للهيئات المتولدة بواسطة ملف الزيادات فإنها تغطي كل صيغ التصريف الممكنة في اللغة العربية ما عدا صيغ المؤكد وتتركب هذه الأفعال مع كل الزيادات السابقة الممكنة (سوابق Prefixes) وبأغلب الزيادات اللاحقة (لواحق Suffixes) باستثناء تلك المتعلقة بالتعدي لمفعولين.

من مميزات مدقق هانسبل Hunspell معتمداً على قاموس آيسبل Ayaspell في الجزء الخاص بالأفعال مقارنة بالمدقق الإملائي للمجموعة المكتبية MsOffice:

- ✔ اعتماد تصريف أفعال القلوب: جزئيا [مدقق MsOffice: لا] فمثلاً نقول نظننا وتظنينك ولا يجوز قول نضربنا وتضربينك.
- ✔ اعتماد التعدي إلى مفعولين: ليس بعد [مدقق MsOffice: لا] نحو يعطيكموها.
- ✔ اعتماد الأفعال النادرة: نعم [مدقق MsOffice: لا] نحو ائتثر - أوجى.
- ✔ اعتماد كامل للمبني للمجهول: نعم [مدقق MsOffice: جزئيا] مثل شوددت من (شادَّ).
- ✔ اعتماد الهيئات المتغيرة في صيغة الأمر للأفعال المهموزة والمضاعفة: نعم [مدقق MsOffice: لا] مثل "ايدب" و"فائدب" من أدب و"ود" و"ايدد" من ودّ.
- ✔ اعتماد سابقتين تتضمن همزة الاستفهام: نعم [مدقق MsOffice: لا] مثل أوتدري؟ أفتعلم؟
- ✔ اعتماد ثلاث سوابق: نعم [مدقق MsOffice: لا] نحو أفستكتبها؟
- ✔ اعتماد صيغ التوكيد: ليس بعد [مدقق MsOffice: لا] نحو ليكتبنان.
- ✔ اعتماد التعدي النسبي للأفعال اللازمة: نعم [مدقق MsOffice: لا] نحو: وكم من انتصار انتصرناه بفضل جهاد شعوبنا...!!! ؛-)

- **المُكَوِّن الاسمي والحرفي**

بالإضافة للأسماء الجامدة والمصادر وصيغ النسبة المرتبطة بهما ، انكب المشروع على دراسة مشتقات الأفعال بأصنافها، اسم مفعول، اسم فاعل، مبالغة، أسماء التفضيل وصفات مشبهة. بعد نسخها من المعاجم المرجعية، تم تصنيفها وتوليد الهيئات الصرفية الممكنة منها (المؤنث، المثنى وجمع السالم) حسب القواعد النحوية للغة العربية المعروفة. مداخل القاموس تحتوي إذن على الكلمة في صيغة المفرد المذكر أو جمع التكسير واستثناءً على هيئة المفرد المؤنث أو جمع السالم.

إجمالاً، نجد في القاموس : 10328 اسم-جامد، 13372 مصدر، 8406 اسم-الفاعل، 1807 اسم-مفعول، 2066 مبالغة-اسم-الفاعل، 1017 صفة-مشبهة، 378 اسم التفضيل، 862 اسم منسوب، بالإضافة إلى مفردات أخرى بعدد 4248 تتوزع بين الصفة والنسبة والاسم الجامد نسخت من معاجم مختلفة ومكانز ونصوص متنوعة. الحصيلة هي إذن: 42484 مفردة أما عدد الحروف والأدوات النحوية وما استثني من هيئات صرفية في قاموس آيسبل وصل إلى 611.

من حيث التصريف، يتضمن القاموس صيغاً لا تدعمها المدققات الإملائية المغلقة كصيغة الإضافة اللفظية (مثل: المقيمي الصلاة)، وصيغ التعدي بحرف أو بظرف (مثل المذهوب بعقله أو المجلوس عليه) وصيغ كثيرة متعلقة بالسوابق مثل أوكاتبُ (أوكاتبون؟) أو وَلَكاتبُ (ولكاتبون) التي أهملتها المدققات الإملائية الأخرى لندرة استعمالها في الكتابات الحديثة على ما يبدو.

تميزٌ آخر في قاموس آيسبل هو تعيين وتحديد تنوين النصب "صراحة" لتفادي الأخطاء المتعلقة بالممنوع من الصرف وتكيفاً مع تعود الكاتب العربي على رسم هذه الحركة بالرغم من إغفال حركات التشكيل في غالب الأحيان.

مستقبل المشروع

من المظاهر السلبية في عمل المدقق الإملائي المعالج للنص العربي، البطء الكبير في اقتراح البديل الصحيح وضعف الدقة والحل المنتظر تجسيده في الإصدارات القادمة إن شاء الله، قصد تحسين فعالية المدقق الإملائي من حيث السرعة والنجاعة، هو إعادة بناء القاموس بالاقتصار على المستعمل من المفردات والتركيز على التصريف المتداول فعلاً في الكتابات الحديثة حسب المقاربة المبينة في الورقة الخاصة بها في مدونة المشروع والمعنونة بقاموس آيسبل بين "المستعمل والمهمل" في اللغة العربية **[24]**.

هذا من جهة، ومن جهة أخرى العمل على هيكلة القاموس بطريقة تتماشى مع المدقق النحوي العربي **[25]**

الذي بدأ الإخوة في عربآيز التفكير في برمجته ليعمل ضمن المجموعة المكتبية أوبن أوفس Openoffice.org.

والله نسأل أن يجعل أعمالنا خالصة لوجهه،
والله وحده ولي التوفيق.

الروابط المذكورة حسب ترتيبها في التقرير لمن أراد التفاصيل:

[1] http://en.wikipedia.org/wiki/Spellchecker
[2] http://www.ameinfo.com/ar-69687.html
[3] http://www.arabeyes.org/
[4] http://www.arabeyes.org/project.php?proj=Duali
[5] http://home.foolab.org/cgi-bin/viewcvs.cgi/projects/baghdad/
[6] http://hunspell.sourceforge.net/
[7] http://www.openoffice.org/
[8] http://aspell.sourceforge.net/
[9] http://www.qamus.org/
[10] http://sourceforge.net/projects/arabic-spell/
[11] http://www.google.com/intl/en/about.html
[12] http://perso.menara.ma/~kebdani/ayaspell-dic/
[13] http://ayaspell.sourceforge.net/
[14] http://ayaspell.sourceforge.net/ar.html
[15] http://forge.aaul.net/projects/ayaspell/
[16] http://groups.google.com/group/ayaspell-dic
[17] http://ayaspell.blogspot.com/
[18] http://ayaspell.sourceforge.net/am.html
[19] http://en.wikipedia.org/wiki/Thesaurus
[20] http://en.wikipedia.org/wiki/Grammar_checker
[21] http://en.wikipedia.org/wiki/Embedded_system
[22] http://sourceforge.net/project/shownotes.php?release_id=494764&group_id=143754
[23] http://sourceforge.net/tracker/?group_id=205373&atid=993378
[24] http://ayaspell.blogspot.com/2007/09/blog-post.html
[25] http://wiki.arabeyes.org/المدقق_النحوي_العربي