الأيام الدراسية حول المعالجة الآلية للغة العربية

المركز الجامعي بشار، ماي 2007.

الجوانب البرمجية لدعم العربية في المدقق الإملائي المفتوح المصدر هانسبال.

طه زروقي*، محمد كبداني، عمار بالة***

**المعهد الوطني للإعلام الآلي بالجزائر*

***طاقم أيسبل، المغرب.*

a_balla@ini.dz, t_zerrouki@ini.dz,kebdani1@menara.ma

.

ملخص :

يهدف مشروع أيسبل إلى إنشاء قاموس في اللغة العربية للمدققات الإملائية مثل - Ispell - Aspell Myspellأو Hunspell وخاصة لهذا الأخير لما يتمتع به من خصائص توافق نسبيا بنية اللغة العربية. و قد شرع في هذا المشروع بعد ملاحظة فقدان ساحة الإعلاميات الحرة لقاموس مبني على مواصفات تلائم المدققات الإملائية.

و سنتناول في هذه الورقة عرضا عن الأبحاث لدعم اللغة العربية في المدقق الإملائي المفتوح المصدر هانسبال. و سنتحدث خاصة عن الجوانب البرمجية من المشروع.

كلمات المفاتيح: التدقيق الإملائي، المصدر المفتوح، اللغة العربية، البرمجة.

مقدمة

إن الحاجة الملحة للتطبيقات اللغوية لم يعد أمرا خفيا، لا سيما في مجال المعلومات و تكنولوجياتها، وصار الحديث عن المعالجة الآلية للغة الطبيعية أمرا مهما في البرامج الحديثة من بواحث و قواميس والترجمة وغيرها من الحاجات اليومية للمستخدم العادي.

وقد نالت العربية حظها من البحث المكثف الدؤوب، من خلال مراكز البحث العلمي المنتشرة في بلاد العرب، و في شتى بلدان العالم.

لكن هذه الجهود الجبارة و النتائج قد تلقى مصيرا لا ثالث لهما، إمّا أن تبقى حبيسة المجامع أو المراكز، أو حكرا على المؤسسات الضخمة التي تمول هذه المشاريع، مما يجعل تكلفة إدماجها في الحلول البرمجية البسيطة أمرا غير ذي جدوى اقتصادية، فنجد قواميس و محركات بحث لا تأخذ بعين الاعتبار إمكانيات العربية و خصوصياتها.

و من هنا لاحظنا كغيرنا الجهود التي يبذلها الباحثون مرارا و تكرارا للحصول على محلل صرفي أو نحوي أو مدقق من أجل المضي قدما في أبحاث المعالجة الآلية، فيبذلون جهدا جهيدا لإعادة اختراع عجلة، عجلة قد لا تكون مستديرة تماما.

و خلال بحثنا في أطروحة الدكتوراة "المعالجة المعلوماتية للقرآن و علوم القرآن: مقاربة معنوية أنطولوجية"، و عند مرحلة إنجاز وحدة البحث في النص القرآني، واجهنا مشكل البحث الصرفي، الذي ليس من صميم بحثنا، فإذا بحثنا في هذا الموضوع نجد الكثير من برامج البحث الصرفي التجارية المكلفة، لا يمكن إدراجها حتى بعد شرائها بأثمان باهظة ضمن بحوثنا نظرا لقيود الحقوق التجارية و الملكية الفكرية. أي أننا باختصار وقعنا في فخ العجلة غير المستديرة..

لذا فإنّ فلسفة المصدر المفتوح تعطي حلا مناسبا يمكن من خلاله تقاسم المعارف و الجهود، و عدم الاختباء وراء الحقوق الملكية التي تعطل أحيانا تطور الأبحاث وتظافر الجهود.

وفي هذا السياق فإنّ المدقق الإملائي مفتوح المصدر حاجة ملحة، تظهر جليا عند العمل على خوارزميات البحث و التلخيص، و القواميس و النطق.

وفي هذا المقال، سنقدم دراسة حول الحلول المعلوماتية للتدقيق اللغوي من خلال مقارنة المدققات الموجودة في السوق (صخر، مايكروسوفت) و ما يتوفر حاليا من مدققات إملائية مفتوحة المصدر، ثم نستعرض خصائص المدقق الإملائي الذي وقع عليه الاختيار، و مزاياه و مساوئه و ما بذلناه من جهود لإثرائه.

دراسة تجارب عملية للمدققات الإملائية

بعد حصول دعم اللغة العربية في المدققين الإملائيين Aspell و Hunspell انبرى الأخ محمد سمير من مصر فاستعمل – للاختبار – في أول وهلة القرآن الكريم كقاموس مرجعي لهذا المدقق الإملائي ثم بعدما لاحظ عدم صلاحية هذا الحل توجه إلى قاموس QAMUS LLC[1] لصاحبه Tim Buckwalter واعتمده قاعدة لمشروع المدقق الإملائي[2] للغة العربية إلاّ أنه تبين له بعد ذلك عدم نجاعة هذا الاختيار.

نفس الحل أخذ به مشروع Ivrix[3] للمطور الإسرائيلي Dan Kenigsberg ثم لحق بهما فريق غوغل[4] Googleالذي ضم من بين تناياه Tim Buckwalter نفسه فأصدروا إصداراً يحوي لأول مرة ملف زوائد affixe file .

نلاحظ أن المشاريع كلها لجأت لنفس القاموس بسبب غياب قواميس حرة أخرى بديلة للغة العربية.

و قد لجأنا في مبادرتنا إلى الاعتماد على خوارزمية هانسبل Hunspell وقاموس Tim Buckwalter في البداية وملف زوائد لتجربة أقصى حدود هانسبال.

و قد قام محمد كبداني بمتابعة هذه المحاولات عن كثب ولكونه كان مهتما جدا بدعم اللغة العربية في البرامج الحرة عموماً وبالمدقق الإملائي خاصة، تبين له منذ الخطوة الأولى أن الأمر جد معقد وأن الحل لا يكمن فقط في وضع لائحة بالكلمات المركبة ليصبح المدقق وظيفياً.

لاحظ كبداني أنّ كل المحاولات السابقة كان لها طابع التجريب والاختبار وتقييم ملائمة أسبل للغة العربية – باستثناء محاولتنا –

و رأى أنّه من المستحيل اعتماد قاموس Tim Buckwalter على حاله أو حتى بعد معالجته وقد تبين ذلك من خلال تجربة فريق غوغل لأنه بصفته مكنز CORPUS مبني من مادة صحفية و غير مستمد لتصريف الكلمة العربية من المنابع فلا يخول له مصدره ولا طبيعة إنشائه أن يطمح ليكون قاموس مدقق إملائي ولم تكن أبداً تلك غايته عند نشأته...

و بعد أن اطلع كبداني على طريقة عمل Ispell و Aspell وكيفية تقعيدهما لملف اللواصق affix file درس Hunspell فوجده أكثر ملائمة لمشروع مدقق إملائي للغة العربية، و بعد مراسلات حثيثة و مناقشات مستفيضة، قررنا الاعتماد عليه في مشروع آيسبل القاموس العربي للمصحح الصرفي (=المدقق الإملائي) في هذه المرحلة ولكن بالاعتماد على قائمة كلمات مختلفة عن قاموس Tim Buckwalter، قاموس نقوم بإنشائه معتمداً على القواميس الرائجة في اللغة العربية...

[1] -http://www.qamus.org/
[2] -http://home.foolab.org/cgi-bin/viewcvs.cgi/projects/arspell
[3] -http://www.ivrix.org.il/projects/arabic/
[4] -http://sourceforge.net/projects/arabic-spell/

نقوم بإنشاء قاموس آيسبل للمدقق الإملائي من المادة اللغوية للمعجم الوسيط و التي سنغنيها بالمصادر لمعجم الهادي إلى اللغة العربية لما يتميز به في هذا الجانب. هذان المعجمان، سيكونان مع معجم الأفعال العربية لمجموعة Bescherelle نواة القاموس من الأفعال والمصادر أساساً. وفيما يرجع للأسماء فبالإضافة إلى تلك الموجودة بالمعجم الوسيط سنغني إن شاء الله المادة اللغوية لآيسبل بالأسماء المشتقة من المنجد في اللغة العربية لتمييزه عن المعاجم الأخرى بضمه للمشتقات الأكثر تداولا في العصر الحديث وأيضاً من المعجم الغني. لتحديد معاني الأفعال فيما يتعلق بالتعدي واللزوم سنعتمد أساساً على المعجم الغني، و إن أشكل معنى فعل يمكن الرجوع إلى مراجع أخرى كلسان العرب ومحيط المحيط ومعاجم أخرى .

هذه هي إذن الروافد المهيكلة لقاموس آيسبل الذي يوضع تحت الرخصة العمومية الشاملة GPL إن كان ولا بد، لأنه من غير المنطقي أن تكون قائمة كلمات لغة ما خاضعة لرخصة أياً كانت طبيعتها...

المقارنة بين المدققات الإملائية الموجودة

البرنامج	الدؤلي
المطور	محمد الزبير
لغة البرمجة	بايثون/ سي
نظام التشغيل	لينكس / متعدد الأنظمة
الترخيص	BSD
الجهة الداعمة	عيون العرب Arabeyes
الموقع	Arabeyes.org http://www.arabeyes.org/project.php?proj=Duali
القاموس	قاموس Tim BulkWalter
الوصف	دعم لغة الضاد لأول مرة.
مزايا	مدقق عربي منذ البداية.
مساوئ	توقف التطوير به، واجهة نصية..

البرنامج	بغداد
المطور	محمد سمير
لغة البرمجة	C
نظام التشغيل	لينكس All POSIX (Linux/BSD/UNIX-like OSes)
الترخيص	BSD
الجهة الداعمة	عيون العربArabeyes.org
الموقع	**http://home.foolab.org/cgi-bin/viewcvs.cgi/projects/baghdad**
القاموس	قاموس Tim BulkWalter
الوصف	تحويل لغة دؤلي من لغة برمجة بايتون إلى C++ (fork duali)
مزايا	
مساوئ	متوقف التطوير...

البرنامج	Ivrix's spell-ar
المطور	Kevin Atkinson
لغة البرمجة	C /
نظام التشغيل	لينكس / متعدد الأنظمة
الترخيص	GNU General Public licence / BSD
الجهة الداعمة	http://www.ivrix.org.il
الموقع	http://www.ivrix.org.il/projects/arabic
القاموس	قاموس Tim BulkWalter
الوصف	برنامج موجه أصلا للغة العبرية. اختبار الدعم العربي في Aspell بتوظيف قاموس Tim BulkWalter
مزايا	The word list was repackaged in Aspell's format
مساوئ	20Mo، غياب ملف الزوائد، أخطاء كثيرة.

البرنامج	Google-arspell
المطور	Google Team
لغة البرمجة	C
نظام التشغيل	– متعدد الأنظمة
الترخيص	GNU General Public License (GPL), GNU Library or (Lesser General Public License (LGPL
الجهة الداعمة	.Google inc
الموقع	https://sourceforge.net/projects/arabic-spell
القاموس	قاموس Tim BulkWalter
الوصف	قائمة كلمات و ملف زوائد.
مزايا	
مساوئ	ثقيل، مكنز، أخطاء كثيرة.

البرنامج	Hunspell
المطور	نامث لاسي، و طه زروقي للمزايا العربية
لغة البرمجة	C
نظام التشغيل	متعدد الأنظمة
الترخيص	/LGPL/MPL
الجهة الداعمة	OpenOffice
الموقع	Hunspell.sourceforge.net Ayaspell.blogspot.com
القاموس	قاموس Ayaspell
الوصف	إعداد قاموس لبرنامج هانسبال إضافة المزايا العربية لهانسبال
مزايا	خصائص إلتصاقية + خصائص اشتقاقية شبه – توليدية. اعتماد المعاجم العربية، اعتماد قواعد اللغة الصرفية والنحوية.
مساوئ	في طور الإنشاء :)

لماذا هانسبال

بعد دراسة المدقق الإملائي هانسبال، و خصائصه المميزة، وجدناه أكثر ملاءمة من غيره لعدة أسباب:

- كونه مفتوح المصدر
- استخدام الترميز العالمي الموحد اليونيكود.
- سهولة صياغة ملف الزوائد و ثرائها.
- كونه يدعم التحليل الصرفي.
- كونه المدقق الإملائي الافتراضي لطقم الأوبن أوفيس المكتبي مفتوح المصدر.
- يمكن استعماله كمدقق افتراضي لبرنامج abiword و مكتب غنوم Gnome.

و لهذه الأسباب و غيرها قررنا الاستثمار في هذا المدقق و اتصلنا بمطوره الأصلي الذي رحب كثيرا بمساهمتنا و ساعدنا في مراحل كثيرة على المضي قدما.

كيف يعمل المدقق الإملائي هانسبل Hunspell

- يتكون قاموس ayaspell-dic من ملفين: الملف الأول هو ملف dic والملف الثاني هو ملف aff . يحتوي ملف dic على الكلمات التي ستمثل الجذور Racine/root وليس من الضروري أن تكون جذراً بالمفهوم المتداول في اللغة العربية. لكون Hunspell غير قادر ، بعد، على التوليد والاشتقاق بطريقة ملائمة للغة العربية، وجب وضع الكلمة الأصل و مشتقاتها في ملف dic . لنأخذ مثال "كتب" : نجد في ملف dic المداخل المتعلقة بالأفعال التالية:

- كتب { مفردة، ستمثل كَتَبَ وايضا كَتَّبَ بتضعيف عين الفعل وهذا بسبب عدم أخذنا بعين الاعتبار للتشكيل في المرحلة الأولى من إنشاء القاموس}
- كاتب
- تكتب
- تكاتب
- أكتب
- اكتتب
- استكتب

كان ممكناً اشتقاق الأفعال المزيدة: كاتب وتكتب وتكاتب وأكتب واكتتب واستكتب من الجذر كتب بتوظيف قواعد الإلحاق لملف aff كما يلي: كاتب = كتب (–ك +كا) وتكتب = كتب (+ت) وتكاتب = كتب (–ك +تكا) وأكتب = كتب (+أ) واكتتب = كتب (–ك +اكت)

إلاّ أن هذا الحل غير مجدي ويثقل كثيرا ملف aff لاختلاف فاء الفعل{ك} من جذر إلى آخر. للأسف مصدر hunspell غير قادر، بعد، على استبدال الحروف الوسطى للكلمة دون ذكرها بعينها (infix)... وإذا تعلق الأمر بفعل معتل أجوف يكون من الضروري وضع الأشكال مختلفة عين الفعل في ملف dic مثل قال و قول وقيل وقل وباع وبيع وبع كما يتوقع إضافة المبني للمجهول للثلاثي المزيد بحرفي التاء والألف: تفاعل – تفوعل...

بناء ملف aff في الأفعال:

تنقسم زيادات Affixes الفعل العربي إلى صنفين، صنف يسبق الفعل وصنف يليه. اختير مصطلح السوابق للصنف الأول Prefixes ومصطلح اللواحق للصنف الثاني Suffixes. هذه الزوائد مرتبطة بتصريف الفعل وتحمل اسم حروف المضارعة (حروف أنيت) و الضمائر المتصلة في كتب النحو المتداولة نحو ت، نا، ما،تم، تن، ين، وا....

هناك زيادات أخرى تضاف قبل السوابق وهي الزيادة السابقة Proclitique مثل واو العطف، الفاء، همزة الاستفهام و أخريات تلي اللواحق هي الزيادة اللاحقة Enclitiques وهي متعقلة بالتعدي مثل ـكما، ـهما...

أسيكتبونها = أ + ي + كتب + ون + ها

أ = Proclitique زيادة سابقة ي = Prefixe سابقة كتب = root/Racine جذر ون = Suffixe لاحقة ها = Enclitique زيادة لاحقة.

زيادة سابقة Proclitique	سابقة Prefixe	الجذر Racine - Root	لاحقة Suffixe	زيادة لاحق Enclitique
س	ي	كتب	و	ها

- **توليد الاقتراحات:**

عندما يكتشف المدقق مفردة غير موجودة في القاموس (ملف dic وملف aff) يعتبرها خاطئة فيقترح مجموعة من البدائل الممكنة و ذلك بالعمل على تغيير حرف واحد فقط ليتوافق مع مفردة القاموس بمفردها أو بتركيبها مع زيادة، وذلك إمّا بتبديل موضعي permutation [قدص/قصد] أو استبدال

حرف بحرف changement [قدص/قرص أو قدس] أو حذف حرف suppression [قدص/قص أو قد] أو إضافة حرف ajout [قق/حقق] أو إدراج فراغ [البتطريق/البت طريق] .

الخصائص المطلوبة

بعد دراسة خصائص المدقق الإملائي و ما يوفره من مزايا، ظهرت بعض المتطلبات التي لا يمكن توليدها بما يتوفر من وظائف، بل يجب إضافتها مثل : إغفال التشكيل و التطويل، و معالجة الإعلال و الإبدال و الإدغام.

المزايا الجديدة لبرنامج هانسبال

إنّ المصدر المفتوح لبرنامج هانسبال المكتوب بلغة C++ أمكننا من إجراء التعديلات و الإضافات الخاصة باللغة العربية و الاستفادة مباشرة من الوظائف الموجودة سلفا دون الاضطرار إلى إعادة برمجتها. و قد قمنا بإضافة عدد من الوظائف الجديدة التي تسهل علينا العمل في إعداد القاموس الخاص بهذا المدقق. و قد قمنا ببرمجة الوظائف التي تحل المسائل الآتية:

- مسألة التشكيل
- مسألة الإدغام.
- مسألة الإبدال و الإعلال.
- مسألة الزوائد المزدوجة.

1- مسالة التشكيل :

بما أنّ معظم النصوص العربية غير مشكولة، فالمدققات الإملائية عموما تغفل التشكيل و التطويل، و وبما أنّ الحركات و التطويل تدخل ضمن الكلمة، وجب إضافة هذه الميزة للبرنامج. فالكلمة الآتية المشكولة "كَتَبَ" غير صحيحة حسب البرنامج، على الرغم من وجود كلمة كتب في القاموس.

الإثراء:

قمنا بإضافة الخاصية الجديدة NOHARAKAT إلى البرنامج لإغفال الحركات و التطويل. و يتم ذلك بتفعيل اللغة العربية في ملف الزوائد LANG ar، ثم إضافة الخاصية NOHARAKAT. و لقد سميناها لا حركات NOHARAKAT بدلا عن NODIACRITICS لأنها خاصة بالعربية فقط.

أمثلة:

كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
كـــــــــتب كَتَبَ	كتب	SET UTF8 LANG ar NOHARAKAT

كُتِـــــــــــب ـــــــــكتبـــ		

2- مسالة الإدغام :

من بين المسائل التي واجهتنا خلال تعاملنا مع المدقق مسألة الإدغام، مثلا : أنا رددت، هي ردّت، و التي للتعامل معها نضطر إلى زيادة مدخل آخر للفعل في القاموس.

مثال:

كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
ردت رددت	رد/T ردد/T	SET UTF8 LANG ar NOHARAKAT SFX T Y 1 SFX T 0 ت

الإثراء

لحل هذه المسألة قمنا بإضافة خاصية الإدغام من خلال وضع علم خاص في ملف الزوائـــد، و هكـــذا ليس علينا إضافة مدخل جديد في القاموس.

الصيغة: GEMINATING G

الكلمة GEMINATING تعني أن العلم الذي بعدها سيستعمل لفرض الإدغام في أي قاعدة زيـــادة يكتب فيها. ثم استعماله في قاعدة الزيادة لفرض الإدغام. الإدغام يتم بحذف الحرف الأخير مهما كان، و من ثم إضافة الزيادة. مثل ردد + ت/G= رد[د] +ت=>ردت.

مثال:

كلمـــــات خاطئـــــة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
رددا	رددت رددتما ردت ردا	رد/T	SET UTF8 LANG ar NOHARAKAT GEMINATING G SFX T Y 4 SFX T 0 ت

			SFX T 0 تما SFX T 0 ت/G SFX T 0 ا/G

الإعلال و الإبدال

من أصعب الأمور في التدقيق الإملائي المبني على السوابق و اللواحق هو التحكم فيما يعتري جذع الكلمة من تغيرات داخلية: فمثلا الفعل المعتل الأجوف قام => قامت، => قمت (حذف الألف)،=>يقوم(إعلال الألف إلى واو).

ولتمثيل هذه المسألة نلجأ إلى إضافة مداخل للكلمة في القاموس حسب تغيراتها (قام، قم، قوم)/ مما يعني زيادة حجم القاموس، و كثرة القواعد و تعقيدها.

مثال:

كلمات خاطئة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
يقام	قامت قمت يقوم	قام/T قم/T قوم/Y	SET UTF8 LANG ar NOHARAKAT GEMINATING G # لزيادة تاء الماضي SFX T Y 1 SFX T 0 ت # لزيادة ياء المضارع PFX Y Y 1 PFX Y 0 ي

الإثراء

و لحل هذه المسألة قمنا بإدخال تقنية جديدة في البرنامج سميناها تقنية الإبدال، أي إبدال حرف بحرف آخر، حذفه أو إدراجه في موضع معين.

في البداية علينا تعريف جدول الإبدالات و عددها، و تسمية كل قاعدة إبدال

تعريف جدول الإبدالات في رأس الزوائد :

الصيغة : كلمة SWAPأمامها عدد الإبدالات الآتية.

SWAP 2

تعريف قاعدة الإبدال بعد تعريف جدول الإبدالات

الصيغة : كلمة SWAPأمامها:

SWAP W 2 A B

- اسم العلم: و تعني أنّ العلم يفرض استبدال الحرف بحرف آخر. مثلا W
- موضع الإبدال: رقم يدل على أنّ الحرف في هذا الموضع، يجب أن يكون مطابقا للحرف المبدل به، و سيتم إبداله بالحرف المبدل. مثلا 2.
- الحرف المبدل به: هو الحرف الذي سيحذف. مثلا A
- الحرف المبدل: هو الحرف الذي سيدرج. مثلا B

إذا كان الحرف المبدل صفرا 0 فهذا يعني أنّ الحرف المبدل به يحذف.

القاعدة الآتية تستبدل بـ الألف الحرف واو في الموضع الثاني من بداية الكلمة

SWAP W 2 و ا

(لوضوح أكثر لاتجاه الإبدال) إبدال W 2 ا و

موضع الإبدال

- يتحدد موضع الإبدال بالنسبة إلى بداية الكلمة مثل : 2 هو موضع الألف في قام.
- أو من نهاية الكلمة بكتابة رقم سالب: مثلا –2 هو موضع الألف في استقال، أي الحرف الثاني من الأخير.
- و يجب أن نعلم أنّ مرتبة الحرف يتم بعد حذف الجزء المحذوف:

أي أنّ استقال–> يستقيل، فموضع الألف الثانية هنا بعد حذف همزة الوصل هو 4 بدلا عن 05.

و هذا ينطبق في الاتجاه، أي على السوابق و اللواحق.

مثال:

كلمات خاطئة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
يقام	قامت قمت يقوم	#مدخل واحد فقط قام/TY	SET UTF8 LANG ar NOHARAKAT GEMINATING G #جدول الإبدالات # إبدال W 2 ا و # إبدال Z 2 ا 0 SWAP 2

SWAP W 2 و ا SWAP W 2 0 ا # لزيادة تاء الماضي SFX T Y 1 SFX T 0 ت SFX T 0 ت/Z # لزيادة ياء المضارع PFX Y Y 1 PFX Y 0 ي/W			

لدينا علمين للإبدال:

- العلم W يقوم بحذف الألف في الرتبة الثانية و تعويضها بواو، فتصبح ي +قام=> يقوم.
- العلم Z يقوم بحذف الألف في الرتبة الثانية ، فتصبح قام+ ت => قمت.

مزايا أخرى

- يمكن استعمال هذه الميزة لتصريف الفعل الماضي المبني للمجهول

إبدال W 2 ا و ؛ مثلا: عاقب => عوقب.

إبدال W 4 أ ؤ ؛ مثلا: استأجر => استؤجر.

- يمكن إجراء إبدالات متعددة:

بتطبيق إبدالين على نفس القاعدة مثلا: لاءم => لوءم.

إبدال W 2 ا و ؛ مثلا: لاءم => لوءم

إبدال Z 3 ء ئ ؛ مثلا: لوءم => لوئم

لاحقة y 0 0/WZ .

التبديلات و الإبدالات

لابد من الإشارة إلى الفرق بين SWAP و REP الموجودة سلفا في البرنامج،

- REP تستبدل بعض الحروف من اجل توليد الاقتراحات، فمثلا إذا كتبت "كباعة" الخاطئـــة، نعلم أنّ الطاء قرب الكاف على لوحة المفاتيح، لذا يقترح المدقق التصحيح "طباعة".
- SWAP يجري تغييرات على جذع الكلمة.

4- مسألة الزوائد المزدوجة

من خلال الدراسات التي تناولت اللغة العربية (رمزي عباس، و محمد كبداني)

نجد أنّ الكلمة العربية تقبل سابقتين و لاحقتين

مثلا : س+ي+فعل+ون+ها

من جهة أخرى، نلاحظ أن ارتباط بعض السوابق واللواحق مشروط

مثلا:

- ب+ال+فعل+ين : مقبولة
- ب+ال+فعل+ان : ان للتثنية، غير مقبولة

كما أنّ الفعل المضارع يستلزم سابقة و لاحقة مشروطة

فعل +ي-ون => يفعلون.

فعل + ت-ن=> تفعلن.

الحل العادي

يتم إجبار ربط السابقة باللاحقة بواسطة خاصية الزيادة الدائرية CIRCUMFFIX X

مثال

كلمات خاطئة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
يفعل	يفعلون	فعل/N	CIRCUMFFIX X PFX Y Y 2
فعلون	تفعلون		PFX Y 0 ي/X
فعلن	تفعلن		PFX Y 0 ت/X
تفعل			
يفعلن			PFX W Y 1
			PFX W 0 ت/X
			SFX N Y 2
			SFX N 0 ون/XY
			SFX N 0 ن/XW

في هذه الحالة:

- لدينا القاعدة:

- SFX N 0 ون/XY
  - تعني أنّ العلم X يفرض أن ترتبط اللاحقة ون بالسابقة Y.
  - فتعطينا يفعلون، تفعلون.
  - لدينا القاعدة:
- SFX N 0 ن/XW
  - تعني أنّ العلم X يفرض أن ترتبط اللاحقة ون بالسابقة W.
  - فتعطينا تفعلن، فقط و ليس يفعلن.

و هكذا نلاحظ أنه كلما تنوعت السوابق كلما استعملنا تصانيف أكثر.

الإثراء

بما أنّ بعض السوابق و اللواحق تترادف دائما، فكرنا في إدماجها في قاعدة واحدة.

فنجعل من ي–ون، ت–ون، ت–ن، سوابق مزدوجة للفعل.

السوابق المزدوجة تحتوي على علامة "–" (علامة ناقص، و ليس تطويلا) تستبدل بجذع الكلمة.

ي–ون + فعل =>يفعلون.

تفعيل الخاصية يتم بواسطة الأمر SPLITPREFIX في بداية ملف الزوائد

سنتبدل لاحقا هذا الأمر بكلمة أخرى SPLITAFFIX

كلمات خاطئة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
فعلون	يفعلون	فعل/N	SPLITPREFIX PFX Y Y 5
فعلن	تفعلون		PFX Y 0 ي–ون
يفعلن	تفعلن		PFX Y 0 ت–ون
ي–ونفعل	يفعل		PFX Y 0 ت–ن
ت–ونفعل	تفعل		PFX Y 0 ي
ت–نفعل			PFX Y 0 ت

ملاحظات

- السوابق التي لا تحتوي على علامة ناقص "–" تعامل و تعمل كسابقة عادية.
- يمكن استعمال السوابق المزدوجة مع الخصائص الجديدة (الإبدال، الإدغام).
- يمكن تركيب السوابق المزدوجة باللواحق المزدوجة.
- يمكن إعطاء اقتراحات حسب الزوائد المزدوجة.
- الفرق بين السوابق المزدوجة و اللواحق المزدوجة هو جهة الحذف.

مثلا:

سابقة Y ا ي–ون

تحذف الألف من بداية الكلمة مثل : استفعل=> يستفعل.

لاحقة Y ا ي–ون

تحذف الألف من نهاية الكلمة مثل : دعا=> يدعون.

- إذا كان الفعل يحتاج إلى حذف من الجهتين، يجب تركيب الزوائد المزدوجة

استقصى => يستقصون

بتركيب السابقة

سابقة Y ا ي–ون/S ؛ تحذف الألف الأول.

لاحقة S ى 0 ؛ تحذف الألف المقصورة.

خاتمة

استعرضنا في هذا المقال الحاجة الملحة لمدقق عربي إملائي مفتوح المصدر، و قد تم اختيار المدقق هانسبال نظرا لمزاياه العديدة، و بعد دراسة معمقة لوظائفه، تقرر إضافة بعض المزايا الجديدة الخاصة باللغة العربية. هذه الوظائف خصصت لإغفال التشكيل و التطويل، و معالجة مسائل الإعلال و الإبدال و الإدغام.

و هكذا يمكننا الآن التفرغ من أجل تجريب المزايا الجديدة لهانسبال و تصميم قاموس ملائم، بعد أن تم تذليل العديد من الصعوبات.

ولقد تناولنا في هذا المقال الجوانب البرمجية لدغم اللغة العربية و تحسينها في المدقق الإملائي المفتوح المصدر هانسبال، و لم نتناول كيفية تصميم القاموس و مختلف الخطوات المحققة.

المراجع

1. موقع المشروع http://ayaspell.blogspot.com
2. موقع برنامج هانسبال http://hunspell.sourceforge.net
3. المعجم الوسيط – مجمع اللغة العربية – المكتبة الإسلامية للطباعة والنشر والتوزيع – الطبعة الثانية – استنبول – تركيا
4. معجم الغني www.sakhr.com/lexicons
5. رمزي عباس La conception et la réalisation d'un concordancier élèctronique pour l'arabe المعهد الوطني للعلوم التطبيقية بليون – فرنسا.
6. معجم الأفعال المتعدية بحرف – دار العلم للملايين – موسى بن محمد بن الملياني الأحمدي – مارس 1986 – بيروت – لبنان – الطبعة الثانية
7. مجموعة بيشرال – الأفعال العربية / الشامل في تصريف الأفعال العربية – سام عمار و يوسف ديشي – 1999 – هاتييه – باريز
8. معجم تعدي الأفعال في اللغة العربية – سلسلة المتقن – أنطون قيقانو – دار الراتب الجامعية – بيروت – لبنان
9. معجم الأفعال وتصريف الأفعال – سلسلة المتقن – غريد الشيخ – دار الراتب الجامعية – بيروت – لبنان
10. Ispell المدقق الإملائي http://www.gnu.org/software/ispell/ispell.html
11. Aspell المدقق الإملائي http://aspell.sourceforge.net/ .
12. Hunspell . http://hunspell.sourceforge.net
13. Duali : http://www.arabeyes.org/project.php?proj=Duali
14. Mohemde Sameer : araspell project, http://www.foolab.org/projects/arspell
15. Ivirix Project : http://www.ivrix.org.il/projects/arabic/
16. Google team for arabic-spell : http://sourceforge.net/projects/arabic-spell/
17. Tim Buckwalter: aramorph project : http://www.qamus.org