الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

مركز البحوث العلمية و التقنية لترقية اللغة العربية

- الجـزائر –

بحث مقدم لنيل درجة الماجستير في علوم اللسان والتبليغ اللغوي

فرع : العلاج الآلي للغة

عنوان البحث

مـن أجـل مـدقـق إمـلائي عـربي مـفـتـوح المصـدر

‘ FOR AN OPEN SOURCE ARABIC SPELLCHECKER ’

إعداد الطالب:

حجير إسماعيل.

الإشراف :

الدكتور عمار بالة.

الأستاذ طه زروقي.

2008

الإهداء

إلى الوالدين الكريمين اللذين سهرا على تربيتي وتعليمي إلى من علماني ...

إلى أخواتي في سورة ... ...

إلى إخوتي الأعزاء ... ...

إلى جميع الأحبة ... ...

إلى جميع زملائي في الجزائر ... ...

إلى جميع الأقارب ... ...

إلى جميع أبناء الفقيد ... ...

إلى كل من ساهم في إنجاز هذا العمل ... ...

إلى كل من ساعدني أو أهداني هذا العمل.

مقدمة في الاقتصاد

يتألف الاقتصاد من مجموعة من الأنشطة التي يقوم بها الأفراد والمجتمعات من أجل إشباع حاجاتهم ورغباتهم في ظل ندرة الموارد المتاحة. وتتناول هذه المقدمة أبرز المفاهيم الأساسية التي يحتاج الطالب إلى استيعابها قبل الخوض في التفصيلات اللاحقة، مع التركيز على الأسئلة الكبرى التي يطرحها علم الاقتصاد في عالم يتسم بالتغير المستمر والتعقيد المتزايد.

يقوم علم الاقتصاد على فرضية أساسية مفادها أن الموارد محدودة بينما الحاجات والرغبات لا تكاد تنتهي، الأمر الذي يفرض على الأفراد والمجتمعات ضرورة الاختيار بين بدائل متعددة. فالوقت والمال والأراضي والعمل والآلات كلها موارد نادرة لها استخدامات متعددة، وكل قرار يتخذ في سبيل تخصيص مورد معين يعني التخلي عن فرصة بديلة كان يمكن أن تحقق منفعة أخرى. هذا المبدأ البسيط في ظاهره يحمل في طياته تعقيدات كبيرة تظهر في كل مجال من مجالات الحياة اليومية والعمل والسياسة.

يقسم علماء الاقتصاد الموضوع إلى فرعين رئيسيين هما الاقتصاد الجزئي والاقتصاد الكلي. يدرس الاقتصاد الجزئي سلوك الأفراد والأسر والمنشآت والأسواق المفردة، ويبحث في كيفية تحديد الأسعار وتخصيص الموارد النادرة بين الاستخدامات المختلفة. أما الاقتصاد الكلي فينظر إلى الاقتصاد ككل، ويدرس المتغيرات الكبرى مثل الناتج المحلي والتضخم والبطالة والنمو والسياسات النقدية والمالية التي تؤثر في الأداء العام للاقتصاد.

درøTدŽ الNNjøH2دŽ ŽG'رøع ŽøŰøدŽ¦ 5ءدøدŽ üøWC دŽدŽ

: ţ øøQ#دŽ

تتناول NjذاI المذكرة مسألة التدقيق الأملائي العربي من وجهة نظر المعالجة الآلية للغة بهدف دراسة مشروع إعداد مدقق إملائي عربي مفتوح المصدر باستطاعته تدقيق النصوص العربية بطريقة آلية والقيام بتحديد الأخطاء الأملائية فيها واقتراح التصحيحات المناسبة البديلة لها.

فاللغة العربية بحاجة إلى مزيد من الاهتمام في مجال التدقيق الأملائي من خلال البرمجيات الداعمة لها، كونها تتميز عن غيرها من اللغات بالعديد من المميزات كالاشتقاقية والاشتقاقية والادغام والاعلال والابدال وغيرها من المميزات التي سنتطرق لها عبر مراحل NjذاI الدراسة.

يشمل NjذاI البحث ميدانيا دراسة شاملة لمختلف المدققات الأملائية الموجودة على ساحة البرمجيات في الوقت الراهن بمختلف أنواعها التجارية والعالمية والمفتوحة المصدر مرفقة بدراسة لمحاولات دعم اللغة العربية في مختلف NjذاI المدققات محاولين الاستفادة من إيجابياتها وتفادي سلبياتها. تم التركيز في NjذاI البحث على المدققات المفتوحة المصدر التي تمكن المطورين من الاطلاع على نصوصها المصدرية ومعرفة طريقة عملها وكيفية تنظيم قواميسها الأمر الذي ساعد المطورين والباحثين في NjذاI المجال على المدى قدما وبشكل سريع وجد فعال في تطوير NjذاI البرامج و إعطاء دفع قوي لدعم العربية فيها عكس المدققات التجارية التي لا تسهل المهمة للباحثين في NjذاI المجال والتي يمكننا أن نعتبر درجة الاستفادة منها في NjذاI المجال جد ضئيلة أو شبه معدومة.

باعتبار مشروع إعداد القاموس العربي للتدقيق الأملائي الذي يشرف عليه فريق يسمى 'Ayaspell' واحدا من أهم المشاريع في الوقت الراهن التي تعمل من أجل تطوير مدقق إملائي عربي مفتوح المصدر تم تنفيذ جزء مهم من NjذاI المذكرة لدراسة NjذاI المشروع الذي لا يزال قيد التطوير. وقد حاولنا خلال NjذاI البحث متابعة مختلف المراحل التي مر بها والنتائج التي تم الحصول عليها وأثرها على تطور NjذاI المشروع. بالإضافة إلى عرض لأهم المراحل المستقبلية المقترحة لإيجاد حلول لمختلف الحالات التي لازالت قيد الدراسة والتي تمثل بعض المعاني التي مازال يواجهها المدقق الأملائي العربي.

الكلمات المفتاحية:

مدقق إملائي، المعالجة الآلية للغة العربية، المصدر المفتوح.

Ǵor an Arabic open source spell checker

Abstract

This memorandum treats the Arabic Spellchecking from the viewpoint of natural language processing, for the purpose to studying the project of preparing an Arabic Open source Spell checker, which check the Arabic texts automatically and detect the misspelled words and suggest the suitable corrections.

The Arabic language needs more interest in the field of spellchecking through its supporter programs, specially that it has more options over the other languages like agglutination, derivation, geminating, ilal and ibdal, and more other features that we will detail through the different steps of this study.

Through this research we have conducted a comprehensive study of the various spellcheckers on the currently arena of software, the commercial ones, the International ones and the Open source Ones, this study was accompanied with the different experiences trying to add the Arabic language support to some of these spell checkers, and trying to take advantage of their positives and avoid their negatives. we concentrated in this research on the open source Spellcheckers, which enable developers to consult its source texts, get more knowledge about the way it works and the method of organization of its dictionaries, which helps the researchers in this field, to go ahead quickly and more effectively to improve this kind of programs and gives a strong push to the support of the Arabic language in these programs, on the opposite off Commercial spell checkers, which don't facilitate the task for developers and researchers in this field, and we can consider, that the degree of benefit from them in this field is very low or almost non-existent.

Considering that the project of preparing an Arabic dictionary for spellchecking, supervised by Ayaspell team, as one of the main projects in the time being that works to improve an Arabic open source spellchecker; we designate a very important part in this research for this project, which is still under development. we tried through this research to follow the different steps that is passed through, the results obtained and its impact on the project development. In addition to an expose of the main future steps suggested to this project to find solutions for some cases which is still under studying and which represent some difficulties that the Arabic spellchecker is still facing.

Keys Words:
Spell checker, Arabic Natural language processing, open source.

الفهرس

قائمة الإشارات 12

قائمة الجداول 17

I. الفصل الأول: الدراسة النظرية للتدقيق الآلي للغة العربية 22

I.1. الباب الأول: مدخل إلى العلاج الآلي للغة العربية 22

- مقدمة 22

I.1-1- تعريف اللغة 22

I.1-2- اللسانيات الحاسوبية والعلاج الآلي للغة 23

I.1-3- تعريف التدقيق الإملائي 23

I.1-4- السوابق واللواحق 23

I.1-5- الزوائد المزدوجة 24

I.1-6- الزوائد المتوسطة 25

I.1-7- الخصائص البنيوية للغة العربية المتعلقة بالتدقيق الإملائي 25

I.1-7-1. الطبيعة الاشتقاقية 26

I.1-7-2. الطبيعة الالتصاقية 26

I.1-7-3. قلة عدد حروف الحركة 26

I.1-7-4. الإدغام 26

I.1-7-5. الإعلال 27

I.1-7-6. الإبدال 29

I.1-7-7. الألف و الياء المتطرفة 29

I.1-7-8. تغيير الدلالة بتغيير مباني الكلمات 30

I.1-8- تصريف الفعل العربي 30

I.1-8-1. تعريف الصرف 30

I.1-8-2. ميدان علم الصرف 31

I.1-8-3. علاقة الصرف بالتدقيق الإملائي 31

I.1-8-4. خصائص الفعل العربي 31

- الخلاصة 32

I.2. الباب الثاني: دراسة تجارب عملية للمدققات الإملائية 33

- مقدمة 33

I.2-1- كيفية استعمال المدققات الإملائية 33

I.2-2- الوظـائـف الأساسيـة للمدقـق الآمـلائـي 33
I.2-3- طريقة عمل المدقق الآملائي 34
I.2-4- دراسـة تجـاري المدققـات الآملائيـة 37
I.2-4-1. المدققـات الآملائيـة العربيـة التجاريـة 37
I.2-4-1-1. مدقـق شركة آيـر 37
أ- تعـريف 38
ب- مـزايا 38
ج-مساوئ 39
د-تعامل مدقق آيـر مع الزيـادات السابقـة في تحـريـف الأفعـال 39
I.2-4-1-2. مدقـق كولتـ II (المستخدم من قبل شركـة مايكـروسوفـت) 40
أ- تعريـف 40
ب- مزايا 40
ج-مساوئه 40
د-التعامل مع بعض الزيادات 41
I.2-4-1-3. تقييم المدققـات التجاريـة 41
I.2-4-2. المدققـات الآملائية متـوحـة المصـدر 41
I.2-4-2-1. المصادر المتوح 42
I.2-4-2-2. أهمية المصادر المتوح في التدقيق الآملائي 42
I.2-4-2-3. قاموس بوكوولتر 43
أ- مـزايـا 43
ب-مساوئـه 44
I.2-4-2-4. المدققـات الآملائية العالمية 44
I.2-4-2-4-1. مدقـق ميسبل "Ispell" 44
أ- تاريخه 44
ب- تعريفه 44
ج- محاسنه 44
د- مساوئه 45
ز- دعم العربية في مدقق "Ispell" 45
I.2-4-2-4-2. مدقق سـبـل "Aspell" 45
أ- محاسن مدقـق سبـل "ASPELL" 45
ب- مساوئ مدقق سبـل "ASPELL" 46
ج- كيف يعمل مدقـق سبـل؟ 46
د- محاولات دعم العربية في مدقق "Aspell" 46
❖ محاولة محمد سمير 47
❖ محاولة مشروع "Ivrix" 48

❖ محاولة فريق "Google" 48
I.2-4-2-4-3. مدقق المنسبل "Hunspell" 48
أ- طريقة عمل المنسبل 48
ب- مزايا المنسبل 49
ج- محاولة دعم العربية في المنسبل 49
❖ محاولة طاقم ' Ayaspell ' 49
I-2-4-2-5. المدققات الآلية باللغة العربية 50
I.2-4-2-5-1. مدقق " Aramorph " 50
أ- تعريف 50
ب- طريقة عمل « Aramorph » 51
ج- شكل ملف القواميس 52
I.2-4-2-5-2. برنامج المدقق الإملائي " الدالي" 53
أ- تعريف 53
ب- آلية عامة 53
ج- مزايا 54
د- مساوئ 54
ه- مراحل المعالجة بالنسبة لبرنامج الدالي 54
I.2-4-2-5-3. برنامج المدقق الإملائي " بغداد" 54
أ- تعريف 54
ب- آلية عامة 54
ج- نقد 55
I.2-5- رسم توضيحي لتدرج المدققات الإملائية 55

• الخاتمة 56

II. الباب الثاني: دراسة مقارنة بين 'HUNSPELL' و 'ASPELL' الإملائيين 58
مقدمة 58

II.1. الترميز العالمي الموحد يونيكود 'Unicode' 58

II.2. مدقق عسبل (Aspell) 59
II.2-1- تعريف عسبل 59
II.2-2- مزايا عسبل 59

II.3. مدقق المنسبل 'Hunspell' 59
II.3-1- تعريف المنسبل 59

II.3-2- الملفات الأساسية لبرنامج أسبل 60
II.3-2-1. ملف القاموس (*.dic) 60
II.3-2-2. ملف الزوائد (*.aff) 62
II.3-3- مزايا أسبل 63
II.3-3-1. المزايا العامة 63
II.3-3-2. المزايا الإضافية 63

II.4. طريقة عمل المدققين أسبل وأسبل 64
II.4-1- دقة أسبل عند اقتراح الكلمات الصحيحة 64
II.4-2- طريقة إستعمال ملفات الزوائد والقاموس لضمان التدقيق الإملائي 65
II.4-2-1. طريقة تدقيق النصوص 66
II.4-2-2. طريقة توليد الاقتراحات 68

II.5. الإضافات معالجة السوابق واللواحق 68
II.5-1- الإضافات عامة 68
II.5-2- الزوائد والأفعال 69
II.5-3- الإضافات المتعلقة بالزوائد في برنامج أسبل 69
II.5-3-1. ميادين عنوان قسم الزوائد 69
II.5-3-2. ميادين قواعد الزوائد 70

II.6. مجهودات إضافة أعلام (flags) إضافية بلغات مختلفة 70
II.6-1- الإضافة 'CIRCUMFIX' في برنامج أسبل 71
أ- مزايا 71
ب- ملاحظات 72
II.6-2- الإضافة ' NEEDAFFIX' في برنامج أسبل 73
أ- مزايا الإضافة 'NEEDAFFIX' في برنامج أسبل 74

II.7. الإضافات الإضافية بالكلمات المركبة 74

II.8. نقد برنامج أسبل 76
الإضافة 77

III. الباب الثالث: التطبيقات العملية لمشروع أسبل 79
خاتمة 79

III.1. تقديم مشروع فريق أيسبل (Ayaspell) 79
III.1-1- البرامج التي استفاد فريق أيسبل من دعم العربية فيها 79

III.1-2- خارطة الطريق لمشروع سيسبل 81
III.1-3- أهم المسائل التي يعمل مشروع سيسبل على معالجتها 82

III.2. التحسينات المتعلقة بمعالجة الزوائد في كل برنامجي سسبل وهنسبل 83
III.2-1- المزايا الناتجة بمسألة إعتلال الشكل 83
III.2-1-1. المبدأ العام 83
III.2-1-2. حالة هنسبل 84
III.2-1-3. حالة سسبل 84
III.2-2- المزايا الناتجة بمسألة الأدغام 84
III.2-3- المزايا الناتجة بمسألتي الأعلال والأبدال 85
III.2-3-1. الحل الأولي المقترح لمعالجة مسألتي الأعلال والأبدال 86
III.2-3-2. الحل المطور المقترح لمعالجة مسألتي الأعلال والأبدال 88
III.2-3-2-1. التصميم الحديث لدعم الزوائد المتوسطة 88
III.2-4- المزايا الناتجة بمسألة الزوائد المزدوجة 89
III.2-4-1. بنية الكلمة العربية 89
III.2-4-2. الأرتباط المشروط للسوابق باللواحق 90
III.2-4-3. الحل الأولي لمسألة الزوائد المزدوجة 90
III.2-4-3-1. حالة سسبل "Aspell" 90
III.2-4-3-2. حالة هنسبل "Hunspell" 90
III.2-4-4. الحل المطور لمسألة الزوائد المزدوجة 91

III.3. التحسينات على القاموس 92
III.3-1- المراجع المعتمدة لبناء القاموس 93
III.3-1-1. المعجم الغني 93
III.3-1-1-1. مزايا المعجم الغني 93
III.3-1-1-2. مساوئ المعجم الغني 93
III.3-1-1-3. الحل المقترح لتغطية نقائص المعجم الغني 93
III.3-1-2. الشامل في تصريف الأفعال العربية 93
III.3-1-2-1. مميزات المعجم الشامل 93
III.3-2- العمل المنجز على المراجع المعتمدة لبناء القاموس 94
III.3-3- التصنيفات والرموز المستعملة لأعداد القاموس 94
III.3-3-1. تصنيف الأفعال 94
III.3-3-1-1. الأفعال اللازمة مطلقا 94
III.3-3-1-2. الأفعال اللازمة نسبيا 94
III.3-3-1-3. الأفعال المتعدية 95

III.3-3-2. تقنية الأسماك 96
III.3-4- أثر تحسينات على الزوائد على القاموس 96
III.3-5- مساوئ قاموس مشروع سيسبل 96

III.4. مثال تطبيقي شامل 96

III.5. التطورات المستقبلية لمدقق مشروع سيسبل 100
III.5-1- اعتبار الشكل 101
III.5-2- الإدغام عند تركيب الزوائد 101
III.5-3- تعريب على الزوائد 101
III.5-4- إستعمال الأعلام العربية بدلاً من الأعلام اللاتينية 101
الخاتمة 102

IV. الباب الرابع: التطبيق والتجارب 104
مقدمة 104

IV.1. كيفية توليد برنامج المنسبل على محرر لوين اوفيس 104
IV.1-1- توليد قاموس التدقيق الإملائي تحت نظام لينكس 104
IV.1-2- توليد قاموس التدقيق الإملائي تحت نظام ويندوز 105

IV.2. تثبيت و تجريب (Compiler) برنامج المنسبل 105
IV.2-1- تثبيت البرنامج و القاموس تحت نظام لينكس 105
IV.2-2- التجريب تحت نظام لينكس 106

IV.3. التثبيت تحت نظام ويندوز 107

IV.4. التجريب تحت نظام ويندوز 107

IV.5. ملتقى برنامج المنسبل 108

الخاتمة 109
خاتمة البحث 110
المراجع 112
الملحق 115

مشكلة الإشارات

إن معالجة اللغة الطبيعية حاسوبياً ليس بالأمر الهين بل يتطلب الكثير من الجهد والمال كما يحتاج إلى فرق بحث متخصصة ذات تطور متكامل حاسوبياً ولغوياً فمعلم الأنظمة والبرامج المجربة على اللغة الإنسانية لم تسلم حتى الآن من الكثير من المشاكل والتعقيدات سواء على المستوى المنهجي أو التطوري للغة. فالباحث في علم اللغة الحاسوبي يحتاج إلى التسلح بقسط من نظرية لسانية من ضرورة الإحاطة بجوانب تقنية تبني له الطريق أمام التطبيق والمقارنة للوصول إلى المنهج التجريبي والأكثر واقعية في وصف جزئيات الخطاب اللساني. فالعمل في حوسبة اللغة الطبيعية يتطلب التمكن من نوعين من المعرفة هما: المعرفة الدقيقة لجميع جزئيات النظام اللغوي وفق أحدث النظريات والقوانين اللسانية الحديثة كما يتطلب الإحاطة والإلمام بالمعرفة الحاسوبية ذات العلاقة بمعالجة اللغة الطبيعية لا سيما في جانبها البرمجي المنطقي فالحاسوب أو الآلة هي منظومة برمجية منطقية تقوم على مجموعة من الخوارزميات الدقيقة فلا يمكن أن نتقدم في مجال البحث في الحوسبة اللسانية حتى نجمع بين هذين النوعين من المعرفة[1].

تتمثل أهمية المعالجة الآلية للغة في وضع نماذج للمعارف اللسانية تعتمد على الرموز بالإضافة إلى التحليل (Analyse) الذي يمثل عملية الانتقال من النص المكتوبة إلى ترجمة لمعاني هذا النص نسبة إلى استعمالها في السياق الذي ورد فيه والتوليد (Génération) الذي يمثل عملية الانتقال من الأفكار التي تتعلق بمعاني النص -نسبة إلى استعمالها في السياق الذي ورد فيه- إلى النص التي تمثل الإنجاز اللساني لهذه الأخيرة[PERRIER G., 2007] . بالنسبة للغة العربية فإن العمل في مجال العلاج الآلي للغة العربية لم يرق بعد إلى مستوى العمل في مجال المعالجة الآلية والتقنية المعلوماتية للغة الأجنبية الأخرى وهذا رغم التطور التقني التي عرفها الحاسوب. الأمر الذي شكل دافعا قويا للعديد من الباحثين اللسانيين للتعاون مع الباحثين الحاسوبيين للسعي وراء حلول جذرية تعالج مختلف إشكالات اللسانيات الحاسوبية العربية فالتجربة من الضروري إعادة تياغة تراثنا من الدراسات اللغوية تياغة جديدة وذلك على ضوء الإنجازات التي تحققت في دراسات لسانيات الحاسوب في اللغات الأخرى حتى يتجاوز بالإمكان التعامل السليم مع اللغة العربية بخصائصها الذاتية وحتى نرقى بها إلى المكانة التي تليق بها كلغة عالمية في مجال تعامل الإنسان مع الحاسوب باللغة الطبيعية من خلال بناء برامج حاسوبية لسانية تمكن من خلق الحوار بين الإنسان والآلة باللغة الطبيعية[2].

تعتبر تقنيات المعالجة الآلية للغات بمثابة البناء التحتي الأساسي (Basic Infrastructure) الواجب توفيرها حيث يتم فوقها بناء كافة التطبيقات الأخرى ولا تقف تقنيات المعالجة الآلية حالياً عند حد إدخال البيانات والمعلومات واستخراجها (Input / Output) بل تمتد إلى التطبيقات اللغوية التخصصية من تدقيق إملائي (Spelling Checking) الذي يشكل موضوع هذا الدراسة وتدقيق نحوي (Grammar Checking) وتحليل صرفي[3] (Morphological Analysis) وغيرها من التطبيقات.

إن الأهمية المدققات الإملائية في مجال المعالجة الآلية للغات الطبيعية من الأمور التي استرعت انتباه المهتمين بالحلول المكتبية منذ البدايات الأولى للمعلوماتية مما حفز الشركات على العمل على إنتاج هذه الأدوات

البرمجية – المملوكة المغلقة في طبيعتها - بوصفها متطورة للغاية فتكاليف تلبي حاجات مستعملي الحاسوب في المجال المكتبي بشكل ما وتغني مصنعي البرامج المعلوماتية التي تدوم الملايين من أجل الحصول عليها.

يمكن تلخيص أهمية المدققات الإملائية في النقاط التالية:

- اكتشاف الأخطاء الإملائية واقتراح التصحيحات المناسبة البديلة لها.
- التفهم السهل والجيد للنص وما من طرف القراء.فهرسة الوصول لمحركات البحث.
- المسؤولية الجماعية للحفاظ على التراث اللغوي.

بالنسبة للتدقيق الإملائي للغة العربية فالأمر أكثر صعوبة وتعقيدا لكونها تتميز ببناء رياضي فريد لا يتوافر لأي لغة أخرى فتسلوب الاشتقاق العديد من الأفعال والأسماء والذي يتسم بنسق رياضي دقيق يتيح لها استيعاب أي مصطلح جديد والتعبير عنه بطريقة تلقائية مبسطة تفتقر من العامة قبل الخاصة بل وتطويع مثل هذا المصطلح إلى موسيقاها الخاصة بما يتفق مع الأذن العربية السليمة ويضاف إلى ذلك قوانين رياضية واضحة لإضافة السوابق واللواحق (Prefixes and Suffixes) لأي كلمة بما يضيف إلى معناها دون أن ينتقص منه شيئا[3].

وقد وُجد العديد من البرامج والتطبيقات التجارية التي تتيح التدقيق الإملائي للغة العربية والتي تتميز بكونها مغلقة ومن غير الممكن الاطلاع على أكوادها المصدرية والتعرف على طريقة عملها. من هذه البرامج نذكر مدقق شركة صخر والذي يعتبر واحدا من أوائل المدققات الإملائية التجارية العربية بالإضافة إلى مدقق (ARWOC) الذي أنتجته شركة كولتا والذي تم اختيارها من طرف شركة مايكروسوفت ليكون المدقق الإملائي العربي الرسمي لمنتجاتها منذ سنة 1997. ما يعيب هذه البرامج التجارية بصفة عامة هو أنه لا يمكن الاستفادة منها في التطوير إلا مقابل مبالغ باهظة من المال.

على صعيد البرمجيات المفتوحة المصدر والتي يندرج ضمن سياقها هذا المشروع كذلك فإن مبادئ واتفاقيات ترخيص هذه البرمجيات تتيح لأي مستخدم تعديل أو تطوير هذه البرمجيات بما يتلاءم مع حاجاته ومتطلباته دون الرجوع إلى المطور أو المصدر الأساسي وذلك عبر توفير الشيفرة المصدرية لكل برنامج وإتاحة تعديلها ضمن اتفاقية الترخيص المرفقة. ولقد استثمرت هذه الميزة من قبل الكثير من الدول والشعوب لتعديل البرمجيات المتاحة ضمن حركة البرمجيات المفتوحة المصدر وتطويعها لتتلاءم مع المتطلبات الخاصة للغاتهم وثقافاتهم مما أدى بالنتيجة إلى إثراء هذه الحركة ومنحها المرونة اللازمة لتجاوز الحواجز الثقافية واللغوية بين جميع القطاع المعمورة وإزالة أية عوائق قد تمنع أي إنسان من الاستفادة من معطيات ثورة تقنيات المعلومات سواء كانت لغوية أو ثقافية أو مادية أو سياسية.

لقد تأخر الحصول على دعم لغة الضاد في البرامج المفتوحة المصدر بشكل عام إلى حدود عام 2006 لم يكن هناك أي مدقق إملائي عربي ومفتوح المصدر رغم تعدد المحاولات العربية المرتبطة بطريقة مباشرة أو غير مباشرة بمؤسسة عيون العرب Arabeyes والتي من أهمها محاولة محمد الزبير ببرنامج "دالي" Duali ومحمد سمير ببرنامج " بغداد" Baghdad ويرجع افتقاد اللغة العربية أساسا لمدقق إملائي

مفتوح المصدر بشكل كلي إلا أن تميزها بصعوبة برمجية ولغوية معقدة نسبياً (كما ذكرنا في الفقرة السابقة) بالإضافة إلى ندرة الكفاءة المختصة وضعف الاهتمام بالبرامج المفتوحة المصدر في المنطقة.

وبعد الاطلاع على مختلف البحوث التي شملها مشهر برامج التدقيق الإملائي المفتوح المصدر ارتأينا أن تكون بداية مشروع المدقق الإملائي العربي المفتوح المصدر انطلاقا من برنامجي المنسل "Hunspell" المعتمد من قبل المجموعة المكتبية أوبن أوفيس "OpenOffice" وعسبل "Aspell". البرنامجان مطوران أصلاً للغات اللاتينية واللذين صبحا بعد إضافة صياغة اليونيكود ودعم ثنائية الاتجاه إليهما صالحين لدعم اللغات غير اللاتينية من ضمنها اللغة العربية.

ينقسم مشروع المدقق الإملائي العربي المفتوح المصدر إلى قسمين أساسيين: القسم البرمجي والقسم القاموسي ويرجع ذلك في الأساس إلى بنية المدققين السابقين (المنسل وعسبل) والتي تعتمد في الأساس على ملفين: ملف القاموس الذي يحوي قائمة الكلمات وملف الزوائد الذي يضم مختلف السوابق واللواحق التي يمكن إضافتها لجذور الكلمات المتواجدة في ملف القاموس مرفقة ببعض المزايا البرمجية التي تنظم إذن الإضافة وتضبطها وفق قواعد اللغة المستعملة.

القسم البرمجي المختص بإعادة برمجة بعض المزايا في ملف الزوائد في كل من البرنامجين لمعالجة بعض حالات التدقيق الإملائي المتعلقة بالمزايا الخاصة باللغة العربية والتي لا توجد في كل من البرنامجين الأصليين لمدققي المنسل وعسبل والتي يمكن تلخيصها فيما يلي:

- ✓ **إشكالية الهمزات (الهمزة):** وهي واحدة من المواضيع التي تواجه التدقيق الإملائي العربي كون النص العربية عادة غير مشكولة الأمر الذي يصعب المقارنة بالنسبة للمدقق في حالة وجود كلمات مشكلة في النص المراد تدقيقه مع ما هو موجود في القاموس.

:مثال

الفعل: كتب

الكلمة الموجودة في القاموس: كتب.

الكلمات المحتملة مع التشكيل: كَتَبَ كُتُبٌ كَتْبٌ.....

لذا تم العمل على إضافة صياغة إعتلال الشكل في أول الأمر مع التطلع نحو إمكانية التدقيق بواسطة الشكل.

- ✓ **الإدغام:** يُقصد الإدغام في بعض الحالات وليس في حالات الأمر الذي يستوجب التمييز بين الحالتين لضمان السير الحسن للمدقق الإملائي والحصول على نتائج تراعي قواعد الصرف العربي دون اللجوء إلى إضافة الأشكال المختلفة للكلمة الواحدة لتفادي مشكلة زيادة حجم القاموس.

:مثال

الفعل: ردّ.

الكلمات المقبولة: ردّ ردّت ردّا ردد.

- ✓ **الإعلال الإبدال:** تتعرض الكلمة العربية في بعض الحالات إلى تغيرات تمس الكلمة وتغير حدد حروفها أو تحذفها وهذا التغيرات تثبت من كلمة لأخرى الأمر الذي يستدعي إيجاد قواعد واضحة تضبط هذا التغيرات لإدراجها في برنامج المدقق حتى يتمكن هذا الأخير من التعرف على الأشكال الصحيحة للكلمات بعد تعرضها للإعلال أو الإبدال انطلاقا من الأصل الوحيد لهذا الكلمة المتواجد في القاموس.

:مثال

الفعل: قام

الكلمات المقبولة: يقوم يقمن قم....

- ✓ **التلازم بين الوحدات اللغوية:** في بعض الحالات تترابط السوابق باللواحق في الكلمة العربية وتتلازم فيما بينها تلازما مشروطا إذ أن وجود سابقة معينة يستوجب لاحقة معينة متلازمة معها. الأمر الذي يستدعي تحديد الحروف المكونة لهذا الزوائد ودراسة توافقها مع بعضها البعض وإدراج خوارزميات صرفية لمعالجتها في برامج التدقيق الإملائي حتى تتمكن هذه الأخيرة من تدقيق جذور هذه الكلمات لتقوم بعد ذلك بالتأكد من توافق السوابق مع اللواحق في هذه الكلمات.

:مثال

الفعل: جلس

الكلمات المقبولة: يجلسون = ي+جلس+ون الزائدة المزدوجة هي: ي-ون.

تجلسن= ت+جلس+ن الزائدة المزدوجة=ت-ن.

أما القسم المتعلق بالقاموس فقد تم تخصيصه لشرح المنهجية والمراحل المتبعة لإعداد القاموس العربي المفتوح المصدر الذي سيشكل مرجعا للمدققين الإملائيين السابقين فبعد حصول دعم اللغة العربية في هذين البرنامجين - هنسبل وهسبل - ظهر تحد صغير هو توفير القواميس العربية الصرفية بالتدقيق الإملائي والتي بدونها لن تتمكن هذه البرامج وبيئتها. لم يكن في الساحة إلا قاموسا عربيا واحدا حرا هو قاموس تيم بوكولتر " Tim Buckwalter" المعتمد في بنائه على مكنز لغوي مكون من مادة صحفية سياسية. لكن كانت المادة الصحفية المرجعية غير مدققة لغويا فتربى عن ذلك احتواء القاموس على مفردات خاطئة في نسبة كبيرة منها رسما ولغة مما هدر سببا على المدققات الإملائية المعتمدة عليه وجعلها لا ترقى إلى المستوى المنتظر منها وهذا ما دفع المدقق العربي-سبل "arabic-spell" ب منتوج شركة جوجل "Google" الذي يعطي نتائج رديئة جدا تدفع المستعمل إلى الاستغناء عنه منذ الوهلة الأولى. أمام هذا الوضع كان ولابد من الاعتماد على موارد أخرى: أولا تكييف المدقق الإملائي مع عادة المستعمل تجاه الحركات في كتاباته باللغة العربية أم بناء قاموس عربي مناسب مفتوح المصدر.

كانت هذه أهم النقاط التي تتمحور حولها الإشكالية المطروحة في هذا البحث والتي سنقوم بمحاولة اقتراح الحلول المناسبة لمعالجتها في سبيل المساهمة في تطوير جيل المدقق الإملائي العربي المفتوح المصدر.

hʊDŽlØØɾ hʊDŽɔØØʂDŽ

hDžløø hDžDøšDž

إن ول ما يشغل بال المرء ثناء الكتابة يا ما كانت اللغة المستعملة في هو ذا النص المكتوب من الأخطاء الإملائية حتى يكون مفهوما وذا قيمة لغوية نص الإتمام انتقل إلى مجال المعالجة الآلية للغة منذ بداية هذا التطور في مهم جدا في تدخل الحاسوب والبرمجيات المتخصصة لإنهاء الأخطاء واقتراح الحلول البديلة في حالة الخطأ الأمر الذي فتح الباب على مصراعيه لمراكز البحوث والشركات المتخصصة في إنتاج البرمجيات للتسابق من أجل الحصول على ما يسمى بمدققات إملائية فعالة تلبي حاجة مستعملي الحاسوب في هذا المجال وتنال رضاهم. وإن كانت الأعمال متقدمة بالنسبة لبعض اللغات كالإنجليزية والفرنسية إلا أنها لا تزال في بدايتها بالنسبة للغات أخرى كما هو الحال بالنسبة للغتنا العربية.

فقد تعددت البحوث في مجال العلاج الآلي للغة العربية ولكنها واجهت الشديد تصادم في أغلب الأحيان بحقوق الملكية الفكرية عند محاولتها الاستفادة من التجارب التي كان لها السبق في هذا المجال كما هو الحال بالنسبة لتجربتي شركة صخر التي أصدرت أول مدقق إملائي عربي وشركة كولتاه التي أصدرت بدورها المدقق الإملائي العربي الافتراضي لمجموعة مايكروسوفت أوفيس المكتبية الأمر الذي لم يمكن أعمال مراكز البحث الأخرى من الرقي إلى المستوى المطلوب والحصول على النتائج المرجوة وأدى إلى كبح تطورها. من هنا كان لا بد من التفكير في برامج مفتوحة المصدر في هذا المجال لأنه السبيل الوحيد للحصول على تطور ملموس ونتائج تكون في مستوى تطلعات وطموحات مستخدمي البرمجيات العربية نتيجة تضافر الجهود بين الباحثين.

يتناول الباب الأول للفصل الأول من هذا البحث بعض العموميات المتعلقة بالعلاج الآلي للغة وبعض التعاريف والشروح لمصطلحات الكلمات التي سيتم تداولها في العديد من فصول هذا البحث. فمن تعريف اللغة التي تعتبر المادة الأولية لكل دراسات العلاج الآلي للغة إلى تعريف الإنتاج الذي يندرج فيه هذا البحث ثم وهو العلاج الآلي للغة إلى ذكر لأهم مجالات تطبيق هذا الأخير ننتقل بعدها لتعريف المدقق الإملائي والذي يشكل موضوع هذا البحث والذي يعتبر واحدا من أهم مجالات تطبيق العلاج الآلي للغة يليه بعض التعريفات لكل من السوابق واللواحق والزوائد المتوسطة والمزدوجة ننتقل بعدها إلى شرح أهم المزايا الخاصة باللغة العربية كالاشتقاقية والاشتقاقية والإدغام والإعلال والإبدال بالإضافة إلى خصائص تميز الفعل العربي وغيرها من الخصائص التي تتميز بها لغتنا العربية عن سائر اللغات والتي تتطلب جهدا ومثابرة ودراسة جد معمقة للتمكن من معالجتها بطريقة آلية وفعالة في نفس الوقت. أما الجزء الثاني من هذا الفصل فسوف يتم من خلاله التطرق إلى دراسة تشمل أهم برامج التدقيق الإملائي أو المدققات الإملائية (Spell Checkers) كما يسميها أحيانا الإنتاج المتواجدة على ساحة البرمجيات منذ بداية ظهور هذا النوع من البرامج حتى يومنا هذا. خلال هذه الدراسة سنقوم بتقسيم المدققات إلى مجموعتين: الأولى تضم المدققات التجارية والتي قمنا بإخلالها بدراسة أهم وأشهر مدققين للغة العربية ويتعلق الأمر بكل من مدققي شركة صخر للبرمجيات ومدقق شركة كولتاه الذي اعتمدته شركة مايكروسوفت كمدقق افتراضي للغة العربية

لمجموعتها المكتبية مايكروسوفت "وفيت" محاولين الوقوف على إيجابيات وسلبيات كل من هذا المدقق وما قدمته للعلاج الآلي للغة العربية في مجال التدقيق الإملائي. أما المجموعة الثانية" والتي نالت الجزء الأكبر من الاهتمام والدراسة خلال هذا البحث" فهي مجموعة البرمجيات المفتوحة المصدر" والتي ارتأينا أن تكون فاتحة دراستها بالحديث عن قاموس تيم بوكوولتر " TIM BUCKWALTER " كونه القاموس العربي الوحيد المفتوح المصدر المتوفر على الإنترنيت حتى هذه الساعة (قبل الانتهاء من إنجاز قاموس التدقيق الإملائي العربي لتريك Ayaspell)" والذي أعد بطريقة منهجية ومنظمة" الأمر الذي جعل معظم محاولات التدقيق الإملائي العربي المفتوح المصدر تعتمد عليه كقاموس لمدققاتها الإملائية" إلا أنه لا يخلو من العيوب التي فتحت الباب لبعض محاولات تعديله ومحاولات إنشاء قواميس بديلة تعتمد عليها برامج التدقيق الإملائي الداعمة للغة العربية" كما هو الحال في مشروع "يسبل" بالإضافة إلى مشكلة إثراء تكمن في كبر حجم هذا القاموس بسبب احتوائه على أشكال مختلفة للكلمة الواحدة. دراسة قاموس تيم بوكوولتر سوف تتضمن أهم مزايا هذا القاموس ومساوئه ومحاولات تعديله.

سوف نتطرق بعد ذلك إلى المدققات المفتوحة المصدر والتي قسمناها بدورها إلى قسمين: المدققات العالمية والتي يتم فيها دعم العديد من اللغات العالمية" كما هو الشأن بالنسبة لمدققات: ISPELL الذي يعتبر أصل المدققات المفتوحة المصدر" و "ASPEL" الذي انحدر منه" و أيضا 'Hunspell' المنحدر من Ispell والذي يعتبر المدقق الافتراضي للغة العربية في مجموعة أوبن أوفيس المكتبية" والذي أعد في الأساس من أجل دعم اللغة المجرية" ليدعم بعد ذلك لغات إثراء كالألمانية والتركية أم العربية. دراسة هذا المدققات ستكون عبر تسليط الضوء على أهم محاسنها ومساوئها وطريقة عمل كل منها" بالإضافة إلى دراسة محاولات دعم العربية فيها" إن وجدت بطبيعة الحال" مثل محاولات كل من محمد سمير ومشروع IVRIX الذي أشرف عليه 'Dan Kenighsberg'" ومحاولة فريق 'GOOGLE' تحت إشراف 'Kevin Atkinson'" لدعم العربية في مدقق ASPELL" بالإضافة إلى محاولات طاقم فريق 'AYASPELL' لدعم العربية في مدقق النسبل" والتي كان من أهمها محاولات طه زروقي" التي تركزت بالأساس على تطوير بعض أجزاء برنامج النسبل" انطلاقا من ملء الزوائد الأصلي لهذا البرنامج. أما محاولات محمد كبداني فقد تركزت على إعداد القاموس العربي (قائمة كلمات) لهذا المدقق" لاقتناعه بعدم جدوى استعمال قاموس تيم بوكوولتر كما هو أو حتى بعد تعديله.

أما القسم الثاني لهذا المدققات العالمية والذي تم تخصيصه لدراسة المدققات الإثرائية باللغة العربية كمدققات 'ARAMORPH' ومدقق 'دلي'" الذي أنتجته شركة عيون العرب للبرمجيات والذي قام ببرمجته محمد زبير باستعمال لغة البايثون 'Python'" بالإضافة إلى مدقق 'بغداد' والذي أشرف على إعدادها محمد سمير من مصر انطلاقا من إعادة برمجة بلغة "C" لبرنامج 'دلي'.

في الفصل الثاني من هذا البحث سنقوم بدراسة مفصلة للمدقق الإملائي النسبل" نصحبها دراسة لمدقق "ASPELL"" كونهما مدققين مفتوحي المصدر ومنحدرين من نفس البرنامج الأصلي "ISPELL" الذي ورثا عنه بعض المزايا ونفس المراحل المتبعة أثناء عملية التدقيق الإملائي مع إضافة في بعض المزايا. سوف يتم التركيز أكثر خلال هذا الدراسة على برنامج النسبل والذي سوف يتم الاعتماد عليه من خلال مشروع "يسبل"

كنقطة انطلاق لمشروع تطوير مدقق إملائي عربي محتوى المصدر. وذلك لما تبدأ من جائزية وما لوحظ من تشابه في بعض قواعد الألحاق للغة المجرية (التي طور الباحث من أجلها) مقارنة باللغة العربية إضافة إلى دعمه لليونيكود.

خلال دراسة هذين المدققين سنتطرق بالتفصيل إلى طريقة عملهما وأهم مميزاتهما بالإضافة إلى شرح لأهم الأعلام التي يستخدمها مدقق المنسبل مثبتة حسب طبيعة عملها وحسب أهميتها وإمكانية تطبيقها على اللغة العربية.

الفصل الثالث من هذا البحث سنخصصه للتحسينات التي أجريت على المدققين هسبل والمنسبل والذي سنفتتحه بتقديم مشروع فريق هيسبل 'Ayaspell' لإعداد قاموس المدقق الإملائي العربي ثم خلال ذكر بعض البرامج التي استفاد فريق هيسبل من دعم العربية فيها يليها تفصيل لخارطة الطريق التي تم من خلالها ذكر مختلف إدراكات برنامج المنسبل وتفصيل التحسينات التي أجراها فريق هيسبل على كل منها لننتقل بعدها إلى دراسة بعض الحالات الخاصة باللغة العربية والتي لم يتمكن برنامج المنسبل الأصلي من معالجتها عند تدقيق النصوص العربية والتي يمكن أن نلخصها في كل من المسائل التالية: مسألة التشكيل والتي يتجاهلها المدقق رغم وجود جذورها في القاموس ومسألة الإدغام إذ يجب تحديد الحالات التي يدرج فيها الإدغام من غيرها عند إضافة زوائد أو لواحق لهذا النوع من الكلمات بالإضافة إلى مسألتي الإعلال والإبدال الذي تتعرض له الأفعال العربية عند تحويلها من صيغة إلى أخرى كالتحويل من الماضي إلى المضارع ومسألة تعامل المدقق مع الزوائد المزدوجة الشائعة الاستعمال في اللغة العربية. كل هذه المسائل تجعل عملية إعداد المدقق العربي المزمع إنجازها أكثر تعقيداً وتتطلب الدراسة بعناية ودقة شديدتين حتى يتمكن المدقق من التجاوز وبشكل صحيح مع مختلف الحالات والصيغ الممكنة. إن دراسة هذه الحالات للغة العربية فتحت الباب على مصراعيه للحديث عن سلسلة من الاقتراحات والتجارب التي قام بها طه زروقي لاقتراح بعض التعديلات على مزايا البرنامجين (المنسبل وهسبل) لحل هذه المسائل والتي سنتطرق لها بالتفصيل حسب التوقيت الزمني لظهورها وتطورها. في الحالة الأولى سنتطرق إلى الخطأ التي لم يمكن من إحتمال الشكل في كل من برنامجي المنسبل وهسبل لحل مسألة التشكيل بينما نعكف في الحالة الثانية إلى تطوير الخاصية "GEMENTING" لحل مسألة الإدغام. لننتقل في هذه الدراسة إلى دراسة الزوائد المتوسطة والتي ستستعمل لحل مسألة الإعلال والإبدال ونختم في الأخير بالحلول المقترحة لحل مسألة تعامل المدقق مع الزوائد المزدوجة.

كما سنلقي الضوء على جزء لا يقل أهمية عن سابقه ألا وهو الجزء الخاص بإعداد القاموس العربي للمدقق الإملائي لمشروع هيسبل القاموس الذي يشرف على إعدادها محمد كبداني من المغرب والذي قرر إعدادها من نقطة الصفر تقريباً معتمداً على أهم المعاجم العربية كالمعجم الغني والمعجم الوسيط ولسان العرب وغيرها من المعاجم المتداولة في الأوساط اللغوية وذلك بعدما تبين له عدم نجاعة الحل القائم على استعمال قاموس تيم بوكوولتر كما هو أو حتى بعد تعديله. سيتم إعداد هذا القاموس وفق منهجية ستعتمد في البداية على تصنيف الأفعال حسب اللزوم (المطلق و النسبي) والتعدي (إلى مفعول به واحد أو مفعولين عاقل أو غير عاقل) بالإضافة إلى بعض الحالات الخاصة لبعض الأفعال كأفعال القلوب يليها تصنيف الأسماء الحروف

والأدوات. مما الإشارة إلى كون عملية تدقيق الأفعال تمثل المرحلة الأكثر أهمية وحساسية من بين مراحل إعداد هذا القانون.

الفصل الأخير من هذا البحث تم تخصيصه للحديث عن بعض الخطوات التقنية لبرنامج المنسبل مثل كيفية تحميل مدقق المنسبيل على محرر مجموعة أوبن أوفيس المكتبية تحت نظام لينكس ونظام ويندوز بالإضافة إلى كيفية تثبيت وتحضير الخادم المحلي لبرنامج المنسبل من أجل اكتمال الصورة عند المطلعين على هذا البحث ومنحهم الفرصة لتجريب هذا المدقق بصورة عملية والإطلاع عن كثب على النتائج التي يمكن الحصول عليها عند استخدامه لتدقيق النصوص العربية ومن ثمة الإدلاء بملاحظاتهم واقتراحاتهم حول هذا المدقق من أجل المضي قدما في تطويره والرقي به إلى أعلى درجات الكفاءة والفعالية.

ünjŬl üøTƏll

Ĭəúødžỵl 5ĭ5ɹøłllNj hŭĞɹøyll høƷłll

I. الإطار النظري للدراسة: اللغة العربية والتدقيق الآملائي

I.1. الكتاب النظري: مدخل إلى المعالجة الآلية للغة العربية

- **مدخل:**

سنقوم خلال هذا الباب بإعطاء بعض التعاريف والشروح اللازمة باللغة العربية ومختلف المتعلقة بالتدقيق الآملائي. البداية سوف تكون من تعريف اللغة التي تشكل المادة الأولية بالنسبة للدراسة التي يجريها المختصون في مجال المعالجة الآلية للغة ننتقل بعدها إلى تعريف المعالجة الآلية للغة التي يندرج هذا البحث ضمن ميادين تطبيقها ثم إلى تعريف التدقيق الآملائي الذي يمثل محور هذه الدراسة ثم إلى تعريف الزوائد بأنواعها المختلفة (السوابق واللواحق والزوائد المتوسطة والمزدوجة) لنعكف بعدها على شرح المختلف البنيوية للغة العربية المتعلقة بالتدقيق الآملائي كالاشتقاقية والاتباقية والإدغام والإعلال والإبدال ننتقل بعدها إلى تعريف الجذر العربي وذكر ميادينه وعلاقته بالتدقيق الآملائي كونه العلم الذي يدرس القواعد التي تعرف بها أحوال بنية الكلمة الأمر الذي يهمنا كثيرا في مجال التدقيق الآملائي للغة العربية وهو ما يقودنا في هذا الباب إلى الحديث عن مختلف الفعل العربي وتقريبه لأن الإلمام بجوانب تركيب الفعل العربي سيشكل حجر الأساس في عملية إعداد القاموس المختص بالمدقق الآملائي العربي.

I.1.1- مختلف اللغة:

هناك تعريفات عديدة للغة تتحد حينا وتختلف حينا آخر. ولعل مصدر التباين في هذا التعريف ناشئ عن منطلقات أصحابها الفكرية. فمن تعريف وضعي خارجي إلى تعريف نفسي داخلي إلى آخر يمثل ندرة فلسفية معينة لواقع الإنسان ووجوده ونشأته. هذا بعض منها حسب الإطار و حسب رأي أصحابها:

- **الإطار اللغوي:** اللغة – لغة – هي لفظة على وزن " فعلة " مثل " كرة " وأصلها " لغوة " على وزن " فعلة "، وقيل في جمعها: لغات ولغون. ومنها لغى: يلغي إذا هذى[4]. وكذلك اللغو فقد قال الله تعالى: { وَإِذَا مَرُّوا بِاللَّغْوِ مَرُّوا كِرَامًا } (سورة الفرقان الآية 72). أي مروا بالباطل. وجاء في الحديث الشريف: " من قال في الجمعة: أنصت فقد لغا "، أي تكلم[5].
- **الإطار الاصطلاحي:** هي مجموعة إشارات تصطلح للتعبير عن حالات الشعور أي عن حالات الإنسان الفكرية والعاطفية والإرادية أو أنها الوسيلة التي يمكن بواسطتها تحليل أية صورة أو فكرة ذهنية إلى أجزائها أو أصلها والتي بها يمكن تركيب هذه الصورة مرة أخرى بأنماط ذهنية وهذان غيرنا وذلك بتأليف كلماتها ووضعها في ترتيب أصلها. وهذا التعريف يتضمن وظيفة اللغة إجمالا[6] .
- **الإطار الوظيفي:** عرفها ببساطة ووضوح حين قال: أعلم أن اللغة في المتعارف هي عبارة المتكلم عن مقصوده و تلك العبارة فعل اللسان فلابد أن تصير ملكة متقررة في العضو الفاعل لها وهو اللسان وهو في كل أمة بحسب اصطلاحاتهم[7].
- **الإطار الاجتماعي:** اللغة أصلية إنسانية ترتبط بالإنسان دون الحيوان ولذلك عدها "وليام ويتني" مؤسسة اجتماعية[8].

- **د. فرج صالح:** اللغة أكثر من مجموعة أصوات وأكثر من أن تكون أداة للفكر أو تعبيرا عن عاطفة إذ هي جزء من كيان الإنسان الروحي وإنها عملية فيزيائية بسيكولوجية على غاية من التعقيد[9].

I.1-2- الأساس اللساني للمعالجة الآلية للغة [PERRIER G., 2007]:

تعتبر اللسانيات الحاسوبية والعلاج الآلي للغة تطبيقا للرياضيات وللمعلوماتية على اللسانيات فاللسانيات الحاسوبية لها توجه علمي بينما للعلاج الآلي للغة توجه نحو التطبيقات (Applications).

يتم توجيه هذين الاتجاهين بواسطة نموذجين:

- **النموذج الرمزي (Paradigme symbolique)**:

يهدف إلى وضع نماذج للمعارف اللسانية بواسطة أنظمة تعتمد على الرموز.

- **النموذج الإحصائي (Paradigme Stochastique):**

يهدف إلى استغلال المعلومة اللسانية من مجموعة من النصوص باستعمال طرق رياضية تطبق على كل ما هو عشوائي أو ما يتعلق بالاحتمالات.

إن العلاج الآلي للغة يمر بأنظمة حسب اتجاهين:

- **التحليل (Analyse):** الذي يمثل عملية الانتقال من النص إلى ترجمة لمعاني هذه الأخيرة نسبة إلى استعمالها في السياق الذي ورد فيه.
- **التوليد (Génération):** الذي يمثل عملية الانتقال من الأفكار التي تتعلق بمعاني النص -نسبة إلى استعمالها في السياق الذي ورد فيه- إلى النص التي تمثل الانجاز اللساني لهذه الأخيرة.

I.1-3- تعاريف التدقيق الإملائي:

التدقيق الإملائي (Spell Checking) هو أحد مجالات المعالجة الآلية للغة وقد شكل موضوع دراسة للكثير من البحوث في هذا المجال. يهتم التدقيق الإملائي بتحديد وتصحيح الأخطاء الإملائية في وثيقة نصية الكترونية بطريقة آلية. العديد من النظريات المتعلقة بالتدقيق الإملائي تم وضعها في علوم الآلي والتي استعملت قواعد إحصائية أو معجمية أو تقنيات مركبة. الأخطاء الإملائية تصنف بشكل واسع على أنها أخطاء غير الكلمات 'NWE' (Non word errors) وأخطاء الكلمات الحقيقية 'RWE' (Real word errors). إذا كانت سلسلة الحروف الخاطئة تمثل كلمة صحيحة في اللغة فهذا يسمى أخطاء كلمة صحيحة في الحالات الأخرى يدعى أخطاء كلمة غير صحيحة.

على سبيل المثال: في الإنجليزية الكلمة 'fun' يمكن أن تكتب بشكل خاطئ على الشكل 'gun' أو 'vun' فالأول أخطاء من نوع 'RWE' والثاني 'NWE'. من السهل تحديد الأخطاء من نوع 'NWE' لكن عملية التصحيح ليست بالهينة. من جهة أخرى فتحديد فالأخطاء من نوع 'RWE' تعتبر عملية جد صعبة لأنها بحاجة إلى تحليل صرفي ونحوي للنص . بالرغم من كون صعوبة التصحيح التي نتسها في الأخطاء من نوع 'NWE' [MONOJIT C., MARKOSE T., ANIMESH M. and ANUPAM B., 2007].

I.1-4- السوابق واللواحق:

- **السوابق:** السابقة 'Prefix' هي جزء بداية متصل بالجزء الرئيسي لهذه الكلمة (الجذر) جاء لتغيير معناها[10].

أمثلة:

- السابقة 'un' والتي عند إضافتها إلى جذر الكلمة 'kind' تحول هذه الكلمة إلى 'unkind'.
- السابقة 'K' والتي عند إضافتها إلى جذر الكلمة 'نوع' تحول هذه الكلمة إلى 'تنوع'.

- **اللواحق:** اللاحقة 'Suffix' جزء من الكلمة يأتي في آخر جذر الكلمة ويحدد جزء من أصل الكلمة[11].

أمثلة:

- اللاحقة 'ly' مضافة إلى 'immediate' (الاسم) لتشكل الكلمة 'immediately' (ظرف أو نعت).
- اللاحقة 'ي' مضافة إلى 'جبل' (الاسم) لتشكل الكلمة 'جبلي' (نعت).

I.1-5- الزوائد السياقية والمتلازمة:

الزائدة السياقية أو الزائدة المزدوجة أو المتلازمة 'circumfix' هي زيادة توضع على طرفي الجذر أو الساق. أي أنها مكونة من جزأين متلازمين: سابقة تسبق الكلمة ولاحقة تلحق بها. ومن الممكن أن تحتوي أيضا على زائدة متوسطة[12]. قد نجد هذه الزيادات في العديد من اللغات نذكر من أهمها:

- **الزوائد المزدوجة:** الزوائد المزدوجة شائعة الاستعمال في اللغة الأندونيسية.
- **الزوائد الأفعال:** تستعمل لتكوين اسم المفعول past participle للأفعال العادية.

مثال: الفعل: **spielen.**

الزائدة: **ge- -t.**

اسم المفعول:**gespielt**.

- **الزوائد الألمانية:** يوجد نظام مماثل للغة الألمانية في الهولندية.

مثال: الفعل: **spelen.**

الزائدة: **ge- -d.**

اسم المفعول: **gespeld**.

- **الزوائد الأجزاء:**

مثال: الساق: **hagdel** (يوسع). المشتق من الجذر 'gdl' (كبير).

الزائدة: **m- -etb.**

النتيجة: **magdelet** (يكبر).

- **الزوائد الظرفية:**

مثال: الساق:yomu.

الزوائد:o- -ni naru و o- -suru.

النتائج: o-yomi ni naru (محترم) و o-yomi suru (متواضع).
[BOECKX C. and FUMIKAZU N., 2004]

- **اللغات الإفريقية 'Guaraní':** تستعمل الزائدة nd- -I في صيغة النفي و nd- -mo'ãi للنفي في المستقبل.
- **اللغات الأوروبية:** فنستعمل الزائدة K-K للتأنيث.
  مثال: الساق: ساقوك.
  الزائدة: K- -K.
  المؤنث: تاقوسك.
- **اللغات السامية:** إستعمال الزوائد السياقية جد شائع في العربية لا سيما في التصريف في المضارع.
  مثال: الساق: كتب.
  الزائدة: ي- -ون.
  المضارع: يكتبون.

I.1-6- الزوائد المتوسطة:

الزائدة المتوسطة 'infix' هي زائدة 'affix' تدرج داخل الجذر أو الساق. وهي شائعة الاستعمال في العديد من اللغات الأفيانوسية والأسترالية والاندونيسية وغيرها من اللغات الآسيوية كما نجدها في بعض الكلمات بعض اللغات الأوساط[12] نذكر من أهمها:

- **اللغات الإفريقية:** كلمة 'تغالوغ' تستعمل الزائدة المتوسطة <um> التي بعد أول مقطع في التعل مثل gr**uma**duate من التعل graduate.
- **اللغات السامية:** وغيرها من اللغات السامية إذ يمكن اعتبار الإعلال(إبدال حرف علة بآخر) في اللغة العربية نوعا من الزوائد المتوسطة. كما يمكن اعتبار الاشتقاق بزيادة حروف وسط الجذر نوعا من الزيادة المتوسطة.
- **اللغات الإيرانية:** هذا النوع موجود أيضا في اللغة الانجليزية كما نلاحظ في الكلمات التالية: sing, sang, sung, song. إلا أن عدد هذا النوع من الزوائد قليل في اللغة الانجليزية مقارنة بلغات أخرى إذ يمكن اعتبارها زوائد هامشية تستعمل في طبقة لغوية معينة.
  أمثلة : - الزائدة المتوسطة <iz> و<izn> تستعمل في أوساط الهيكل house لكلمة h**iz**ouse البيت و sh**iz**nit من أجل shit قطعة.
  - الزيادة المتوسطة <ma> في بعض العائلات اللغوية تميز التوزيع الموافق من قبل اللغوي Alan C. L. Yu. مثال: sophisti**ma**cated و saxo**ma**phone و edu**ma**cation
  - التسمية الكيميائية تدرج الزيادة <pe> للدلالة على الهدرجة الكاملة (أو مشتقة من piperidine) فتعطي كلمات مثل pi**pe**coline (من picoline) و lu**pe**tidine (من lutidine). كما تدرج الزيادة المتوسطة <et> (أو مشتقة من ethyl جذر الإيثيل) فتعطي phen**et**idine (من phenidine) و xanthoxyl**et**in (من xanthoxylin).[13]

I.1-7- الزوائد الأجنبية المستعملة في اللغات الإفريقية والإيرانية:

تتميز اللغة العربية بالعديد من الخصائص البنيوية المتعلقة بالتدقيق الإملائي والتي تتميز بها عن سائر اللغات. نذكر من أهمها:

1-7-1.I. الطبيعة الاشتقاقية:

نعني بالطبيعة الاشتقاقية للغة العربية أن أغلب مادة لغوية معينة مثل (كتب) يمكن تشكيلها على هيئات مختلفة كل هيئة منها لها وزن خاص ولها دلالة خاصة نقول كاتب مكتوب مكتب كتاب كتيب وغيرها من الكلمات التي يمكن اشتقاقها من الكلمة الأصل. عدد الجذور في العربية أقل بكثير من العدد الموجود في اللغة الإنجليزية ولكن بتوليد الاشتقاق نحصل على مفردات تعد بالملايين بينما في اللغة الإنجليزية لا يتجاوز هذا العدد مئات الآلاف.

والاشتقاق هو أخذ كلمة من كلمة أخرى مع المحافظة على قرابة بينهما لفظا ومعنى مما جعل هذه اللغة يتصل بأصولها في نسيج خاص بها من غير أن تذيب معالمها فإذا أخذنا كلمة " كتب " – على سبيل المثال – و اشتققنا منها: كاتب وكتاب ومكتبة ومكتوب وجدنا أن الحروف الأصلية موجودة في كل كلمة من الكلمات و أن معناها مشتركا يجمع بينها وهو الكتابة. وذلك على عكس اللغات الأوربية حيث لا توجد – في كثير من الأحيان – أية صلة ما بين كلمات الأسرة الواحدة فكاتب في الإنكليزية : Writer وكتاب Book ومكتبة Library ومكتوب Letter ويلاحظ أن لا علاقة بين حروف هذه الكلمات[14].

2-7-1.I. الطبيعة الإلصاقية:

نعني بالطبيعة الإلصاقية للغة العربية وجود مادة لغوية يمكن تشكيل صيغ منها عن طريق إلصاق زوائد في أول المادة أو في آخرها. وباعتبار السوابق واللواحق المعمول بها في اللغة العربية يمكن للكاتب التعبير بالكلمة نفسها عن التذكير والتأنيث الأفراد والتثنية والجمع الاستفهام والنداء وغيرها من الكلمات التي يمكن الحصول عليها بإلصاق بعض الزوائد مع أول أو آخر الكلمة الأصل[15].

ملاحظة: كلمة « أسيعطيكموها » تمثل أحسن جيدة هذه الميزة.

3-7-1.I. عدد حروف الهجاء:

تتميز اللغة العربية بقلة عدد حروف الحركة 'Vowels' (الا وا ي) مقارنة باللغة الإنجليزية (a ,u ,o ,i ,e) أو باللغة الفرنسية (y, w, u, o, a, â, ê, â, û, ù, à, è, é) والاعتماد الأساسي على التشكيل (الكسرة الفتحة الضم وتنونها) لضبط المعنى وتفادي الالتباس. حركات يهملها عادة المستخدم العربي في كتاباته[16].

4-7-1.I. الإدغام:

تعريفه: إدخال حرف في حرف مثله بحيث يصيران حرفا مشددا نحو: [قال = قَلْ وعلم = عَلْ][17].

شروطه:

- يجب الإدغام إذا سكن الحرف الأول منهما وتحرك الثاني نحو: [يحجج = يحجْ] و[يحجج = يحجْ].

وبتعبير آخر أجمل وأمتن: [كلما استطاع الفاعل والإدغام وجب الإدغام].

- يجوز الفك والإدغام في فعل الأمر المبني على السكون والمضارع المجزوم بالسكون نحو:

[اشدد = شدْ (أمر مبني على السكون)] و [لم يمسسْ = لم يمسّ (مضارع مجزوم بالسكون)].

ملاحظة: الأمثلة المختارة تمثل بعض النماذج الفصيحة من الإدغام.

- قال تعالى: [يكاد زيتها يضيء ولو لم تمسسه نار] (النور 35/24)

فالإدغام في الآية على المنهاج إذ الفعل مضارع مجزوم بالسكون. ومتى كان كذلك جاز فيه الإدغام: [لم تمسسه] وجاز الفك: [لم تمسسه].

- قد نجد الإدغام والفك في كلمة تكرر في بيتين:

فمن الإدغام قوله تعالى: [ومن يشاقق الله...] (الحشر 4/59). ومن الفك قوله تعالى: [ومن يشاقق الله...] (الأنفال 13/8). والإدغام والفك في الآيتين على المنهاج. وذلك لأن الفعل فيهما مجزوم بالسكون فيجوز معه الوجهان.

- قال الحارث بن هشام (وهو أبي جهل عمرو بن هشام):

وعلمه الذي إن قاتل واحداً فقتله ولّى فراراً عدوّاي مشهدي

[يضرر]: مضارع مجزوم بالسكون إذ هو معطوف على جواب الشرط [فقتله]. ومتى كان المضارع مجزوماً بالسكون جاز الإدغام: [يضرّه] وجاز الفك: [يضرره].

- قال تعالى: [وامسحوا برؤوسكم وأرجلكم إلى الكعبين] (طه 22/20)

وقد ألغى الإدغام في الآية وذلك جائز كما أن الإدغام جائز أيضاً. ففي غير القرآن يجوز لك أن تقول: [أرجلكم إلى الكعبين]. إذ القاعدة أن الإدغام وفكه جائزان في الأمر المبني على السكون.

- وقال: [سنشد عضدك بأخيك] (القصص 35/28). والإدغام هنا واجب لأن يجوز سواكن لسكون الأول وتحرك الثاني. وليس الأمر كذلك في الآية: [واجعل لي وزيراً من أهلي هارون أخي اشدد به أزري] (طه 31/20) فهنا فعل أمر مبني على السكون فيجوز معه الفك كما هو مبين في الآية ويجوز الإدغام أي: [شدّ].

- قال تعالى: [فشدّوا الوثاق] (محمد 4/47) في هذه الحالة لا يجوز إلا الإدغام لأن الفعل - وإن كان فعل أمر هنا - فتنه ليس مبنياً على السكون بل هو مبني على حذف النون ومعلوم أن شرط جواز الوجهين أن يكون فعل الأمر مبنياً على السكون حقيقة أو يكون فعلاً مضارعاً مجزوماً بالسكون حقيقة.

1.7-1-5. الإعلال:

- **الإبدال**: هو حذف حرف علة أو قلبه أو تسكينه[57].
- **الإعلال:**
  - **الإعلال بالحذف:**

إذا التقى ساكنان والسابق منهما ياء أو واو أو ياء حذف السابق. وذلك نحو: [يا أيها الذين آمنوا] والأصل قبل الحذف: يأيها وأوامن وبيايعن فلما التقى ساكنان حذف السابق. ومثال ذلك أن تقول عن النسوة: [لن يخفن ويخفن ويبعن] والأصل: يخافنا ويخوفنا ويبيعن.

إذا كان أول الماضي واواً حذف في مضارعه ومضارعه نحو: [وعد - يعد - عدة وقف - يقف - قفة...].

- **الإعلال بالقلب:**
  - **حرف العلة** (الواو والياء) إذا تحرك وقبله فتحة نحو: [قَوَلَ - بَيَعَ] وبعد القلب: [قَالَ - بَاعَ]. فإذا كان حرف العلة طرفاً كان لقلبه الشرط الثاني فقط أي: انفتاح ما قبله نحو: [دَنَوَ الفَتَى - فرحي العَوَا] وبعد القلب: [دَنَا الفَتَى فرحاً العَوَا].
  - **قلب الواو ياء**: (ولكن اعلم من قبل أن نبسط القاعدة أن هذا إنما يكون إذا سبقت الواو كسرة).

أولاً: إذا سكنت الواو وسبقتها كسرة.

مثال: مِوْزان ⟵ مِيزان.

ثانياً: إذا وقعت الواو بين كسرة وألف شريطة أن يكون ذلك:

إما في مصدر أعلت عينه وعين فعله نحو: [قِوام: قِيام، سِواك: سِياك، انقواد: انقياد...]

فإن أفعالها معتلة العين أيضاً وهي: [قَوَمَ: قَامَ، سَوَكَ: ساكَ، انقوَدَ: انقادَ...].

وإما في جمع وزنه [فِعال] أعلت عين مفردها أو سكنت. **إعلال** في المفرد نحو: [دِوَر: دارَ] وفي الجمع: [دِوار: دِيار...]. **إعلال** في المفرد نحو: [قَوْلَ] وفي الجمع: [أقوال: أقيال].

ثالثاً: إذا اجتمعت الواو والياء والسابق منهما ساكن. وتورد كتب الصناعة كلمة: [سَيِّد] مثالاً على هذا. فتقول: الأصل فيه [سَيْوِد] فلما اجتمعت الياء والواو والسابق - وهو الياء - ساكن قلبت الواو ياء وأدغمت في الياء فقيل: [سَيِّد].

  - **قلب الواو ياء:**

أقلب الياء واواً في كلمات معدودة. نورد أشهرها فيما يلي:

قلب الياء		قلب الواو	
يَيْسَر	مَيْسَر	يَوْسَر	مَوْسَر
يَيْمَن	مَيْمَن	يَوْمَن	مَوْمَن
يَيْمَة	مَيْمَة	يَوْمَة	مَوْمَة
يَيْفَة	مَيْفَة	يَوْفَة	مَوْفَة
يَيْنَع	مَيْنَع	يَوْنَع	مَوْنَع
يَيْسَر	مَيْسَر	يَوْسَر	مَوْسَر

إعلال الحرف الصحيح المعتل بالحرف الصحيح:

أ- إعلال إذا اعتل ثالثها كان واواً في كل حال نحو: [دعوة - نشوة - تقوة] إلا في **إعلال** فيما يبقى نحو: [زيا (مدنث زيان) - وحيا (مدنث حيان)].

ب- إعلال: إذا اعتل ثالثها كان ياء في كل حال نحو: [العليا - الفتيا - الوليا (مدنث الأول) أي الأجدر والأحق] إلا في **إعلال** - فواواً يبقى نحو: [حزوى (اسم موضع)].

- **حرف العلة في الفعل:**

يُسكَّن حرف العلة إذا تطرّف وقبله متحرك. فلا تظهر عليه ضمة ولا كسرة فلا يقال مثلاً: [يدعُوُ القاضيُ بالجاثيْ] بل يسكن في كل ذلك فيقال: [يدعوُ القاضي بالجاثيْ]. وكذلك فتحاً فيُقال فيقال مثلاً: [لن ندعُوُ القاضيُ ليقضيُ في الأمر].

ويُحذف إذا المعتلان الساكنان إذا تلاقا ساكنين وذلك نحو: [يرميُ+وُن] فإن الياء تحذف لالتقاء الساكنين فيقال: [يرمُوُن].

إذا وقع حرف العلة في حشو الكلمة نقل حركته إلى الحرف الساكن الصحيح قبله. وذلك نحو: [يقوُم]. وبعد نقل الحركة: [يقُوم]. ونحو: [يبيْن] وبعد نقل الحركة: [يبَيْن].

ملاحظة: إذا لم يكن حرف العلة مجانساً حركته انقلب حرفاً يجانسها.

ففي نحو: [قوُم] إن تجانس بين الواو وفتحته فينقلب ألفاً ليجانسها: [قام]. وفي نحو: [يقوُم] إن تجانس بين الواو وكسرته فينقلب ياءً ليجانسها: [يقيم]. وفي نحو: [بيْن] إن تجانس بين الياء وفتحته فينقلب ألفاً ليجانسها: [بان].

- **الإعلال:**

- من القواعد الكلية التي تتفق في العربية أنه لا يُبدل بساكن وإن يوقف على متحرك. واستمساكاً بهذه القاعدة نظرنا إلى المفرد وعالجناه موقوفاً عليها بالسكون.
- أصول الكلمات العربية ليس فيها ألف وبناء على ذلك في اللغة حرفي علة فقط - إن ثلاثة - هما الواو والياء. فإذا رأيت الألف في كلمة فاعلم أنها منقلبة عن أحدهما أو أنها ليست أصلية. بل زائدة. فإذا قلنا في تشاعيف البحث مثلاً: (يُقلب حرف العلة أو يُحذف حرف العلة....) فإنما نعني بذلك الواو والياء.
- يمتاز وزن [فاعل] بخروجه عن قواعد الإعلال من قلب وحذف ونقل وفيما يلي بيان ذلك:

إن يُعلّ حرف العلة في [فعل] اسم تفضيل. يقال مثلاً: الخالد أطول من سعيد وأبين منه. ولو أعل لقيل: أطال من سعيد !! وأبان منه !! (وهذا لا يقال في العربية).

¨ وإن يُعلّ في [فعل] صفة مشبهة وإن في فعله أيضاً وإن في مصدره. يقال مثلاً: فلان [أحول] وقد حوِل - يحوَل - حولاً. ولو أعل لقيل: أحال!! وهو: [أيد (رافع رأسه كبراً)] وقد أيِد - يأيَد - أيداً. ولو أعل لقيل: آد!! (وهذا لا يقال في العربية)

¨ وإن يُعلّ أيضاً في [ما أفعله] و [أفعل به] تعجباً. يقال مثلاً: [ما أيدَه وأيِد به] ولو أعل لقيل: ما أدَه وأدِ به = أدْ به !! (وهذا لا يقال في العربية).

1.1-7-6. الإبدال:

- **تعريفه:** هو إحلال حرف في الكلام محل حرف كنحو: سلبلة قصب وهراق الدهن واهراق له فإنها في الأصل: سنبلة وأراق وانبرق. فالإبدال إذاً مسألة لغوية تتعلق بنطق الحروف إن بتركيب الكلام[57].

1.1-7-7. الإدغام في الإظهار والإخفاء:

تعوّد الكاتب العربي في بعض الأقطار مثل مصر على الخلط بين الألف واليائ المتطرفة وتفشّي عادة استعمال الأحرف المركبة لأم-الف (لأنّ لآن لأنّ لأنّ) التي لا محل لها من الإعراب في عصر الحاسوب والتي يعود مصلها إلى إلزامك تقنية مرتبطة بالآلة الكاتبة في بداية القرن الماضي ومن الطبع التخلص من هذه الرواسب التاريخية نظرا لاستمرار العمل بلوحة المفاتيح الموروثة عن الآلة الراقنة.

I.1-7-8. تغيّر الدلالة بتغيّر مباني الكلمات:

تتغير الدلالة في اللغة العربية بتغير مباني الكلمات و يبرز هذا كل تغيير معنى جديدا نحو : قلم - قلمان - أقلام ، كاتب – كاتبة – كاتبات . و يزاد بعض الحروف في الأفعال الثلاثية والرباعية المجردة فتأتي في كل زيادة بمعنى جديدا فاللازم يتحول متعديا نحو : كرُم _ أكرم _ كرّم . ويفيد المشاركة إذا جاء على وزن " فاعل " نحو: لعب _ لاعب. ويشتق من الأصل أوزان جديدة فتأتي بمعانٍ جديدة نحو: علم عالم (اسم فاعل) معلوم (اسم مفعول) علامة (صيغة مبالغة)..... إلخ[18].

I.1-8- تصريف الأسماء الأربعة:

I.1-8-1. تصريف الصرف:

- **الصرف في اللغة:** الصرف لغويا مأخوذ من المادة المعجمية (ص ر ف) ومن ذلك قولهم: لا يقبل منه صرف ولا عدل ... وقولهم: لأنه ليتصرف في الأمور ... وصرف الدهر حدثانه ونوائبه. والصريف: اللبن ينصرف به عن الضرع حارا إذا حلب ... والصيرف المحتال المتصرف في الأمور ... والصيرفي: الصراف من المصارفة وغيرها من التراكيب اللغوية التي تدل على معنى التحويل والتغيير والانتقال من حال إلى حال.
- **الصرف في الاصطلاح:** فهو تحويل الأصل الواحد إلى أمثلة مختلفة لمعان مقصودة لا تحصل تلك المعاني إلا بهذا التغيير. وذلك كتحويل المصدر " قطع " إلى الفعل الماضي " قطع " والمضارع " يقطع " والأمر: اقطع " وغيرها مما يمكن أن نتوصل إليه من مشتقات تتصرف عن الكلمة الأصل كاسم الفاعل واسم المفعول والصفة المشبهة وغيرها ونحو إلى جانب ذلك علم يبحث فيه عن المفردات من حيث صورها وهيئاتها أو من حيث ما يعرض لها من صحة أو إعلال أو إبدال. ولم يرد عن النحاة الأوائل تعريفا جامعا مانعا لعلم الصرف وغاية ما عرف به هذا العلم ما ورد عن ابن الحاجب في حاشيته حيث قال: " التصريف علم بأصول تعرف بها أحوال أبنية الكلم التي ليست بإعراب". وقد ناقش شارح (الرضي) في شرحه للشافية التعريف السابق وبين أوجه قصورها كما عرفه ابن جني بقوله " أن تأتي إلى الحروف الأصول فتتصرف فيها بزيادة حرف أو تحريف بضرب من ضروب التغيير ، فذلك هو التصرف فيها والتصريف لها". ليس الغرض من علم الصرف " إنتاج الكلم بمعنى الإتيان إلى الحروف الأصول والتصرف فيها " على نحو ما ذكر ابن جني، بل هو

"معرفة القوانين التي تمكن من إنتاج الكلام والقوانين المفسرة للتغيير فيها". وباختصار " إنه علم بقواعد تعرف بها أحوال أبنية الكلمة التي ليست بإعرابٍ ولا بناء"[19].

I.1-8-2. أقسام الأفعال :

يقتصر التصريف على نوعين من الكلام :

أ ـ الأفعال المتصرفة.

ب ـ الأسماء المتمكنة.

وما عدا ذلك من أنواع الكلام لا يدخل تحت طائلة الصرف وذلك كالحروف والأسماء المبنية مثل " إذا وأين وحيث" والضمائر مثل " أنا وأنت ونحن" وأسماء الإشارة كـ "هذا وهذه" وأسماء الموصول كـ " الذي والتي " وأسماء الشرط كـ " من وما ومهما" وأسماء الاستفهام كـ " من وما ومتى " والأسماء المشابهة للحرف مثل " كم " و" إذ " والأسماء الأعجمية كـ " إبراهيم وبشار وإسماعيل " والأفعال الجامدة كـ " نعم وبئس وعسى" وما كان من الأسماء أو الأفعال على حرف أو حرفين إلا ما كان مجزوما منه لأن أقل ما تبنى عليه الأسماء المتمكنة أو الأفعال المتصرفة ثلاثة أحرف[19].

I.1-8-3. علاقة الصرف بالدراسات الإملائية:

هناك علاقة وطيدة بين تصريف الأفعال العربية والمدققات الإملائية التي تدعم اللغة العربية، كون الصرف العربي يهتم ببنية الكلمة (الاسم المتمكن والفعل المتصرف)، والتي تعتبر من أهم المدققات الإملائي والذي كان بالإمكان تسميته بالمدقق الصرفي لهذا السبب عكس المدقق النحوي الذي موضوعه الجملة.

I.1-8-4. أهم الأبنية الصرفية:

يتميز الفعل العربي بالعديد من الصفات التي جعلتنا نركز على دراسته ولسنا لهذه الدراسة كل هذا الوقت والجهد لدراسة مختلف التغيرات التي قد تطرأ عليه للاستفادة منها في دراسة مشروع إعداد المدقق الإملائي العربي. فمن أهم هذه الصفات نذكر ما يلي :

- التنوع الكبير للأفعال العربية، نذكر من أهمها: الفعل الصحيح، الفعل المعتل، الفعل الثلاثي، الفعل الرباعي، الفعل المجرد، الفعل المزيد، الفعل المبني للمعلوم، الفعل المبني للمجهول، الفعل الماضي، الفعل المضارع، فعل الأمر، الفعل اللازم، الفعل المتعدين، الفعل تام التصريف، الفعل ناقص التصريف.
- التغيرات التي تحصل على حروف العلة أو الهمزة أو الإدغام والتضعيف.
- تغيرات الأفعال المبنية للمجهول.
- التغيرات المرتبطة بالنصب والجزم.

- **الإشكالية:**

تهدف هذه الدراسة التمهيدية إلى تعريف بعض المصطلحات وشرح بعض المسائل التي سيتم تناولها خلال الفصول القادمة من هذا البحث بصورة مباشرة أو غير مباشرة من خلال دراسة مشروع إعداد المدقق الإملائي العربي المحتوى المصدر. كما تهدف أيضا إلى إبراز العلاقة بين كل من اللغة العربية والعلاج الآلي للغة والصرف العربي والتدقيق الإملائي لتحرير الأرشية المشتركة بين كل هذه المجالات والتي تشكل الأرشية التي سينطلق منها هذا البحث.

I.2. الباب الأول: دراسة تجارب عالمية في التدقيق الإملائي

- **مقدمة:**

سنتطرق في هذا الفصل إلى تعريف المدقق الإملائي وذكر وظائفه كما سنقوم بدراسة بعض تجارب المدققات الإملائية العربية بمختلف أنواعها بداية من المدققات التجارية مثل مدققي شركة RT وشركة كولتا وتوني إلى المدققات المفتوحة المصدر والتي سيتم خلالها التطرق إلى المدققات العالمية مثل Ispell الذي يعتبر أول المدققات المفتوحة المصدر يليه كل من مدققي Aspell وHunspell المنحدرين منه ودعم اللغة العربية في كل منهما لننتقل بعدها إلى دراسة المدققات التي اختص باللغة العربية مثل مدقق Aramorph ومدقق دالي ومدقق بغداد.

I.2-1 - المدققات الإملائية التفاعلية:

هناك كيفيتين أساسيتين لاستعمال المدققات الإملائية:

- **المدققات الإملائية التفاعلية (interactive spelling checkers):**

وهي المدققات التي تقوم باكتشاف الأخطاء أثناء الكتابة. وتتميز بما يلي:

✓ بإمكانها وضع اقتراحات من أجل التصحيح كما يمكن أن تفعل ذلك.

✓ تتطلب ردا فوريا يجب أن يكون سريعا.

✓ القرار يعود للمستخدم لتحديد ما إذا كان المدقق مصيبا أم لا.

✓ إذا كانت هناك قائمة من الاختيارات قد نحتاج إلى دقة بنسبة 100% بالنسبة للكلمات المصححة.

- **المصححات الإملائية الآلية (Automatic Spelling Correctors):** [MARKUS D., 2007]

يتم في هذه الحالة تشغيل المدقق على وثيقة كاملة فيقوم باكتشاف الأخطاء ويصححها. المهمة في هذه الحالة أيسر من سابقتها.

تتميز هذه المدققات بما يلي:

✓ يتم تدقيق الوثائق المكتوبة دفعة واحدة وبطريقة آلية.

✓ يستطيع المستخدم التدخل لاحقا لتصحيح بعض النتائج (عملية اختيارية).

I.2-2 - الوظائف الأساسية للمدقق الإملائي:

للمدقق الإملائي وظيفتين أساسيتين، تتمثلان في:

✓ **اكتشاف الأخطاء (Error Detection):** الكتابات بالعثور على الكلمات الخاطئة إملائيا.

✓ **تصحيح الأخطاء (Error Correction):** تصحيح الكلمات الخاطئة إملائيا.

مثال: من السهل القول بان الكلمة 'ater' هي كلمة خاطئة. ولكن ما هي الكلمة الصحيحة: water later after هذا يعتمد على ما نريد استخدام النتيجة فيه وما نريد القيام به.

مميزات برامج التدقيق:

هناك أمور كثيرة تقوم المدققات الإملائية بتنبيه المستخدم لها، نذكر من أهمها:

- ✓ معظم المدققات الإملائية تنبه المستخدم في حالة تكرار الكلمة عند الكتابة.
- ✓ التنبيه عند استخدام الأحرف الكبيرة الشكل في مواضع غريبة. مثال: ' caPitals in pEculiar pLaces'.

ملاحظة: اكتشاف الأخطاء شرط أساسي مسبق لإمكانية التصحيح فيما بعد.

I.2.3- طرق عمل المدققات الإملائية:

المدققات الإملائية التقليدية تقوم عند تدقيقها للنص بالبحث عن كل كلمة في القاموس فإن وُجدت ينتقل المدقق إلى الكلمة الموالية، في حالة عدم وجود الكلمة في القاموس يقوم المدقق بإدراج أو حذف أو تعويض أو تبديل مواضع بعض الحروف المكونة لها، للنظر بعد ذلك إذا كان بالإمكان العثور على إحدى الكلمات في القاموس. إذن التغييرات الأربع تمثل مهام تصحيح الأخطاء. نظرا لصعوبة اقتراح الحلول بالنسبة للكلمات التي يتعدى عدد الحروف الخاطئة فيها الحرف الواحد، تقوم معظم المدققات بتحديد الأخطاء في موضع واحد في الكلمات[KERNIGHAN M., CHURCH K. and GALE W., 1990].

- **الطرق الآلية لتصحيح:**

للقيام بعملية التدقيق الإملائي، تقوم المدققات الإملائية بشكل عام بتتبع المراحل التالية:

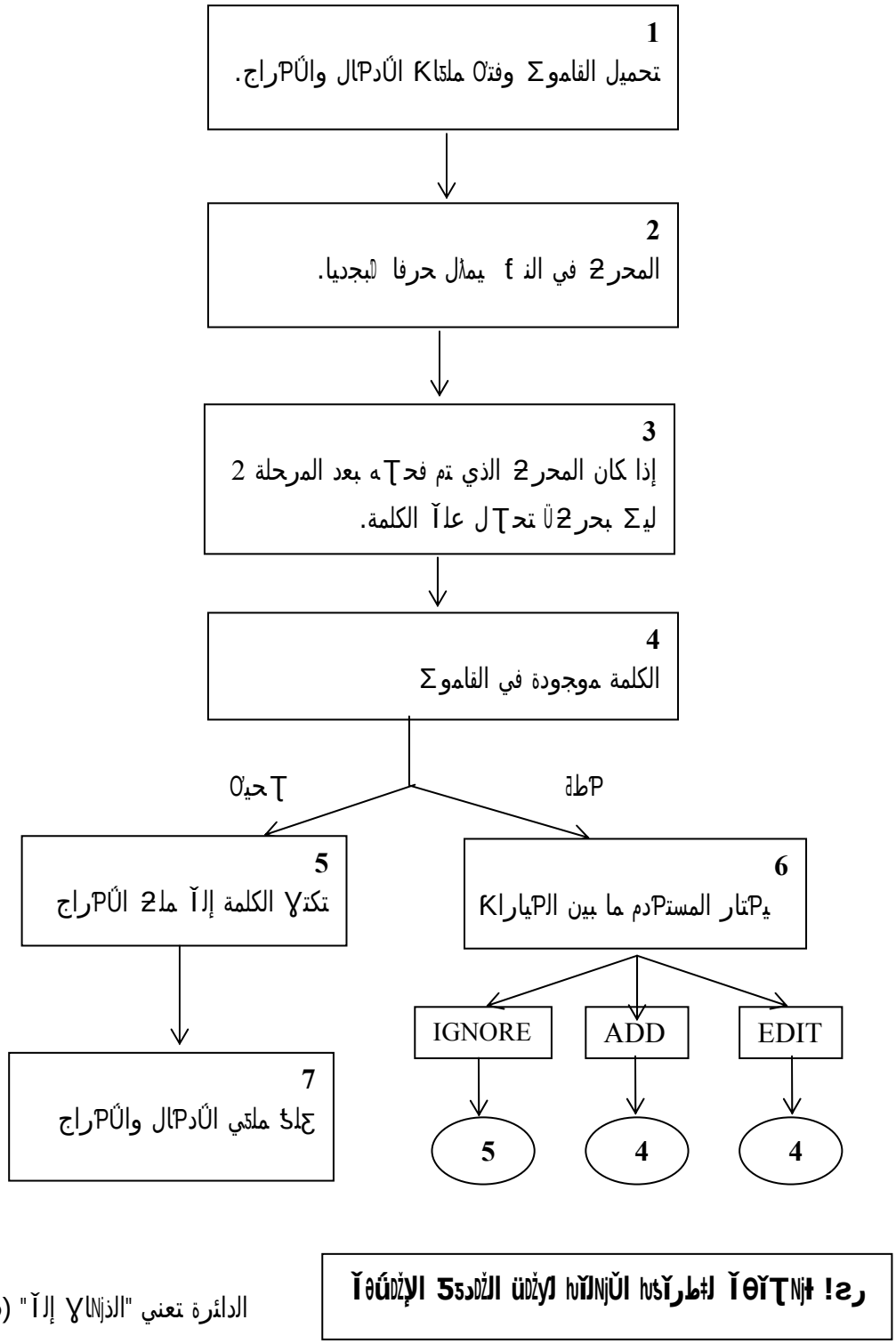
1. القراءة من ملف الإدخال (input file).
2. تجزئة كل كلمة على حدى للتأكد إذا ما كانت خاطئة إملائيا.
3. إعطاء المستخدم خيارات متنوعة.
4. كتابة الكلمة بشكلها الإملائي الصحيح إلى ملف الإخراج (Output file).

الكلمات الخاطئة إملائيا نعني بها في هذه الحالة الكلمات الغير موجودة في القاموس، أما الكلمات الصحيحة فنعني بها الكلمات الموجودة في القاموس أو تلك التي حددها المستخدم بأنها صحيحة.

من مهام برامج التدقيق الإملائي، تلك الخاصة بالخيارات المتاحة للمستخدم عند مصادفة الكلمات الخاطئة إملائيا. نذكر من بينها:

- ✓ تحرير الكلمة (Edit the word): تمكن المستخدم من إعادة كتابة الكلمة الخاطئة للحصول على النسخة المحررة المدققة إملائيا. هذا مفيد عندما يكون المستخدم غير متأكد من الشكل الإملائي الصحيح للكلمة، أو إذا ما كان يرغب بأن الكلمة تكتب بشكل خاطئ.
- ✓ تجاهل الكلمة (Ignore the word): تمكن المستخدم من كتابة الكلمة كما هي إلى ملف الإخراج والانتقال إلى تدقيق الكلمة الموالية في النص .
- ✓ إضافة الكلمة (Add the word): تمكن المستخدم من إضافة الكلمة إلى القاموس، نتيجة لذلك فكل ظهور لهذه الكلمة في المستقبل سوف يجعل المدقق يعتبرها كلمة صحيحة.

الرسم التوضيحي التالي يوضح بصورة مفصلة المراحل المتبعة من طرف المدقق الإملائي لإمكان القيام بعملية التدقيق الإملائي **[BENTLEY K.,2001]**:



الدائرة تعني "الذهاب إلى" (Go To)

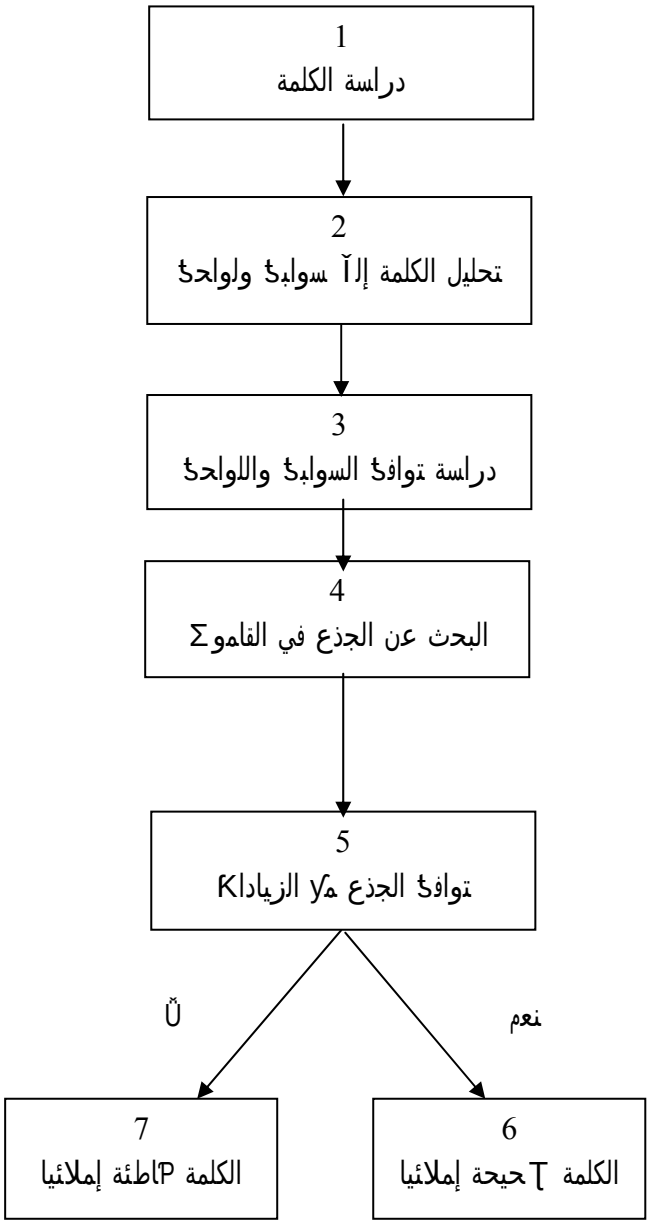
- **الطريقة المطورة للتدقيق الإملائي:**

أما الطريقة المطورة للتدقيق الإملائي والتي تعتمد على دراسة الزوائد واللواحق وتوافقها مع الجذع والتي تفيدنا أكثر عند الحديث عن التدقيق الإملائي العربي فيمكن تلخيصها في المراحل التالية:

من أجل كل كلمة يقوم المدقق بما يلي:

1. دراسة الكلمة.
2. تحليل الكلمة إلى سوابق ولواحق.
3. دراسة توافق السوابق واللواحق.
4. البحث عن الجذر في القاموس.
5. دراسة توافق الجذع مع الزيادات.
6. إمكانية رفض الكلمة لأحد الأسباب التالية:
   - الزيادات غير معتبرة (غير مبرمجة).
   - الزيادات غير متوافقة.
   - الجذع غير مذكور في القاموس (معجم غير وافي).
   - الجذع مذكور في القاموس ولكن الزيادات غير متوافقة مع الجذع. أو أن الجذع غير صحيح.
   - مع الزيادات.

الرسم التوضيحي التالي يلخص هذه المراحل:



الشكل 2-4-1: مخطط يوضح مراحل التدقيق الإملائي

I.2-4- دراسة مراحل التدقيق الإملائي:

I.2-4-1. التدقيق الإملائي في الكلمات المستوردة:

I-2-4-1-1. مدقق شركة صخر:

ج- أهم أدوات التدقيق:

يعتبر مدقق شركة أرامكس للبرمجيات أول مدقق إملائي عربي يقوم هذا المدقق بتصحيح الأخطاء الإملائية التي تحدث أثناء كتابة النص و أو العربية عن طريق إعطاء بدائل متعددة للكلمة الخاطئة ليقوم المستخدم باختيار إحداها. ولا يقتصر دور المدقق على تصحيح الأخطاء الإملائية بل يتعدى ذلك إلى تصحيح الأخطاء النحوية في النص و أو العربية إذا أراد المستخدم ذلك. يتعامل المصحح مع جميع أنواع النص و أو العربية على حد سواء وبنفس الكفاءة[20].

ب- مميزات المدقق:

من أهم مزايا مدقق أرامكس نذكر:

- **التصحيح الآلي:**

يمكننا من خلال هذه الخاصية أن نقوم باقتراح بدائل للكلمة الخاطئة بحيث يقوم المدقق بتصحيح هذه الكلمة تلقائيا للكلمة التي اقترحتها في كل مرة يجد فيها الكلمة الخاطئة.

- **تحديد عدد الاقتراحات:**

يتميز المدقق بأنه يقلل عدد الاقتراحات البديلة للكلمة الخاطئة مع مراعاة ترتيب هذه البدائل حسب علاقتها بالكلمة الخاطئة بحيث يكون الاقتراح الأول هو الأقرب إلى الصواب وهو الأمر الذي يحصر مجال الاختيار أمام المستخدم على عدد قليل من الكلمات الصحيحة .فعلى سبيل المثال إذا أخطأ كلمة "إستطاع" فسيعطيه البرنامج بديلا واحدا صحيحا لهذه الكلمة وهو "استطاع" وهي الكلمة الأقرب للكلمة التي أخطأ بها.

- **مراعاة السياق:**

يأخذ بعين الاعتبار السياق اللغوي للكلمة الخاطئة.

مثال : مائدة القطار ---- 1 عجلات القطار ---- 2 (خطأ).

مائدة الفطار ----1 عجلات القطار ----2 (صحيح).

- **التصحيح الآلي للأخطاء:**

القدرة على التعامل مع المصطلحات المكتوبة بشكل خاطئ وتصحيحها حسب المصطلحات المتعارف عليها.

مثال : الجحر الأسود ------ خاطئ

البحر الأسود أو الحجر الأسود ------ صحيح.

- **التشكيل الآلي:**

يتعامل مع علامات التشكيل في الكلمات العربية بشكل فعال ويعطي اقتراحات بديلة للكلمات التي يراها بأنها مشكلة بشكل خاطئ.

مثال : قوائمة المدفوع ---- خاطئ.

الأولى المفردة ----- صحيح.

ط - مساوئه:

من بين مساوئ مدقق صرفي نذكر:

- عدم التعامل مع بعض صيغ الأفعال القلوب.
  مثال : نحسبنا نحسبنا تحسبنا تحسبنا تحسبنا تحسبنا.
- عدم التعامل مع بعض الزيادات (النون التقرة التالية 'د').
- غياب التوكيد في إسناد جمع المذكر. مثال: يكتبنان.
- غياب تصريف الفعل المتعدي إلى مفعولين. مثال : يعطيكموها.

د- تصريف مدقق صرفي وفق الزيادات السابقة وفق صيغ الأفعال المرفوعة:

الزيادات السابقة (السوابق) هي حروف تقع في بداية الكلمة. الفعل يقبل الحروف التالية وأغلبها يغير من حركة وبنية صرفها:

" ت " : [ا الاستفهام] تدخل على الفعل المضارع والماضي (تصل الفعل المرفوع فقط): أتكتبوا أتكتبين أيكتبن

" ن" : [و العطف] تعطف الفعل على ما سبقه من رفع أو نصب أو جزم: ويكتبون وتكتبين وليكتبن.

" ف " : [ف السببية] ناصبة للفعل المضارع: فتكتبي.

" ف " : [ف العطف] للترتيب والتعقيب عاطفة للرفع والنصب والجزم: فيكتبون فتكتبين فليكتبن

" أ " : [ل التوكيد] وللابتداء يدخل ترفع الفعل المضارع والماضي (لام مزحلقة واقعة في خبر إن) وللتوكيد ايضا: ليكتبن لتكتبن ...

" أ " : [ل التعليل] و[ل الجحود] يدخل تنصب المضارع.

" أ " : [ل الأمر] تجزم فعلا مضارعا واحدا: ليكتب.

" س " : [س سوف] حرف استقبال ويختص بالمضارع لا يغير إلا في الرفع: سيكتبان.

يلاقي مدقق صرفي مشاكل كثيرة في التعامل مع الأفعال المرفقة بهذه الزيادات فبعضها يعتبرها صحيح و البعض الآخر يعتبرها أخطاء رغم أنها صحيحة لغويا.

ي: هذه قائمة ببعض حالات تصريف الفعل 'كتب':

الفعل يلي: صرف مدقق صرفي وفق الصيغ التالية: يكتبون ويكتبوا فيكتبوا ليكتبون فنكتبن وليكتبوا فليكتبون فليكتبوا ليكتبوا وليكتبوا.

الأفعال التي يعتبرها مدقق صرفي خاطئة: يكتبوا ويكتبوا فيكتبوا نكتبن ونكتبن فنكتبن ولنكتبن فلنكتبن سنكتبن وسنكتبن فسنكتبن.

I-2-1-4-2. مدققات الإملاء (Word Corrector) وشركة ARWOC:

أ- تعريف:

يعتبر 'word Corrector' ARWOC لشركة كولتا أكثر من مجرد مدقق إملائي إذ اعتبر الأسرع (أكثر من 800.000 كلمة في الثانية) والأكثر تطورا بين أنظمة تصحيح الكلمات. باستطاعة هذا المدقق تصحيح الكلمات معتمدا على البنية اللغوية بها في اللغة العربية واقتراح التبديلات الصحيحة معتمدا على أنماط الكتابة وإنشاء القواميس المختصة[21]. نظرا للنجاح الذي لقيه هذا المنتج تم اختيارها من طرف شركة مايكروسوفت ليكون المدقق الإملائي العربي الرسمي لمنتجاتها منذ سنة 1997.

ب- مزاياه:

من أهم مزايا مدقق كولتا نذكر ما يلي:

- يستند إلى تقنية كولتا للمعالجة الآلية للغة الطبيعية (NLP).
- يحترم الخصائص البنيوية لكلمات اللغة العربية.
- يولد قواميس مختصة.
- يصحح كل أنواع أخطاء الكلمات: الإملائية والصوتية والصرفية والنحوية.
- يكتشف ويصحح المسافات الزائدة أو الناقصة.
- يصحح كلمات وتحتوي وحيدة كما يصحح المستندات الكاملة.
- يصحح تلقائيا الأخطاء المتكررة.
- يوافق بين الكلمات والأعداد.
- يتضمن الاقتراحات في مجموعات: الأول منها بما لا يزيد عن 6 اقتراحات وتغطي نحو 80 ٪ من الأخطاء الشائعة.
- يمكن إدماجه في تطبيقات مختلفة.
- معجم النظام يتزامن بالإضافة إلى العربية التصحيح المعاصرة لأسماء العملات والكلمات الدخيلة وأسماء العلم (متوفر على الإنترنت).

ج- عيوبه:

من أشهر مساوئ مدقق كولتا نذكر ما يلي:

- عدم التعامل في بعض الصيغ لأفعال القلوب.

  مثال : التركيبات الآتية صحيحة لغويا لكن مدقق كولتا الإملائي يعتبرها أخطاء:

  نحسبنا تحسبين تحسبان تحسبون تحسبن تحسبك.
- عدم التعامل في بعض الزيادات (انظر الفقرة القادمة 'د').
- صياغة صيغ تعريف التوكيد.
- صياغة خاطئة حرف اللام الذي ينطق مدغما بحذف النون (اعتبر المدقق في أحد الأفعال الخمسة مصحوبا بعطف الواو مثلا: وليكتبون صحيح حسبه).

- يقبل المدقق كلمة "تعترف" ويجهل كلمة "يعترف" (تناقض في النتائج).
- كلمة "وتدري وكثير غيرها من الكلمات التي لها نفس الوزن تعتبر أخطاء في نظر المدقق.
- غياب تصريف الفعل المتعدي إلى مفعولين. مثال: يعطيكموها.

د- الإشكال في تحديد الأخطاء:

نفس المشكل الذي تم طرحه على مدقق الصرفي فإن بعض الزيادات يعتبرها مدقق كولتا الأخطاء رغم أنها صحيحة لغويا.

الحل: قائمة ببعض حالات تصريف الفعل 'كتب'.

أولا: الأفعال الخماسية المزيدة بحرف واحد:

أكتبون ويكتبون فيكتبون ليكتبون وليكتبون سيكتبون وسيكتبون فسيكتبون.

ثانيا: الأفعال الخماسية المزيدة بحرفين:

أويكتبون أفيكتبون فليكتبون أليكتبون فليكتبون أوليكتبون أفليكتبون أسيكتبون أوسيكتبون أفسيكتبون.

3-1-4-2-I. الإشكالات المتعلقة بالتدقيق الإملائي:

من خلال دراسة المدققات التجارية تظهر لنا جليا بعض النقائص التي تعاني منها المدققات السالفة الذكر والتي ينبغي تجاوزها في الإصدارات اللاحقة من هذا البرامج. كما أن كلفة استخدامها وأمنها الباهظ يعقد من إمكانية استخدامها في البحث العلمي وكذلك بسبب الحماية المفروضة عليها مما لا يخولها لأن تكون مرتكزا لأبحاث مستقبلية خارجة عن شركات تطويرها. هذا ما يدفعنا إلى التفكير في المصادر المفتوحة التي تتيح للباحثين والمطورين المضي قدما بجهود متضافرة دون خوف من القيود الفكرية.

2-4-2.I. الإشكالات المتعلقة بالمصادر المفتوحة:

لقد ساهم برامج المدققات الإملائية المفتوحة المصدر في تطور هذه الأخيرة بشكل سريع وفعال كونها تتاح الاستفادة من التجارب السابقة الأمر الذي أعطاها قابلية ومرونة في التحول لملائمة لغات مختلفة من حيث تكوين الكلمات ودراسة التغيرات التي قد تطرأ عليها.

وفيما يلي سنعرض أولا المصادر المفتوحة للتعرف أكثر على هذا النوع من البرامج على ذكر أهم الخصائص التي يتميز بها وأهميته في مجال التدقيق الإملائي لننتقل بعدها إلى الحديث عن البرامج المفتوحة المصدر والتي سنفتتحها بالحديث عن قاموس تيم بوكوولتر (TIM BUCKWALTER) تليها دراسة لأهم المدققات المفتوحة المصدر والتي سنقوم بتقسيمها إلى قسمين أساسيين: المدققات العالمية ككل من 'Ispell' و 'Aspell' و 'Hunspell' والمدققات الخاصة باللغة العربية كمدقق 'Aramorph' ومدققات شركة عيون العرب للبرمجيات وننفرد بها كل من مدققي 'الدالي' و 'بغداد'. هذا الدراسة سوف تشمل نشأة هذه المدققات وأهم مميزاتها وطريقة عملها ومحاولات دعم العربية فيها.

I-2-4-2-1. المصادر المفتوحة:

المصادر المفتوحة (Open Source) هو الحد السياسي المتبعة في إدارة عملية كتابة برمجيات الحاسوب والأنظمة وإدارتها. إذ أن البرمجيات المفتوحة المصادر توفر النص المصدري كما كتب من المبرمج على السماح بقراءة هذا النص وتوزيعه وتعديله.

حالياً يتم استعمال مصطلح البرمجيات المفتوحة المصادر في الإعلام بشكل أساسي للدلالة على البرمجيات الحرة. من الممكن اختصار تعريف البرمجيات المفتوحة المصادر بأنها البرمجيات التي تحقق الشروط التالية [12] :

- حرية إعادة توزيع البرنامج.
- توفر النص المصدري للبرنامج وحرية توزيعه.
- حرية إنتاج برمجيات مشتقة أو معدلة من البرنامج الأصلي وحرية توزيعها تحت نفس الترخيص الأصلي للبرمجيات.
- من الممكن أن يمنع الترخيص توزيع النص المصدري للنسخ المعدلة على شرط السماح بتوزيع الملفات التي تحتوي على التعديلات بجانب النص الأصلي.
- عدم وجود أي تمييز في الترخيص لأي مجموعة أو شخص .
- عدم وجود أي تحديد لمجالات استخدام البرنامج.
- الحقوق الموجودة في الترخيص يجب أن تعطى لكل من يتم توزيع البرنامج إليه.

I-2-4-2-2. أهمية المصادر المفتوحة في تطوير البرمجيات:

يعتبر المصادر المفتوحة في الوقت الراهن واحداً من أهم عوامل تطوير البرمجيات، وقد أثبتنا ذلك في الفترة الأخيرة من خلال ظهور برمجيات عالية المستوى في مختلف التخصصات من بينها برامج التدقيق الإملائي، فحسب رأي إريك رايموند (أول من استخدم مصطلح المصادر المفتوحة) فإن السماح للمستخدمين بتغيير البرامج وإعادة توزيعها أمر مستحسن كوسيلة لإنتاج برامج أكثر تفوقا من الناحية التقنية. فبداية التدقيق الإملائي المفتوح المصادر كانت مع مدقق 'Ispell' الذي ولسوء الحظ توجد أي محاولة لدعم العربية فيه، عكس مدققي 'Aspell' و 'Hunspell' المفتوحي المصادر بدورهما والمنحدرين من نفس البرنامج (Ispell)، والذين تعرضا للكثير من المحاولات الداعمة للغة العربية فيهما مثل محاولات محمد سمير ومحاولات مشروع 'Ivrix' ومحاولات فريق 'Google' في مدقق 'Aspell' ومحاولات فريق 'Ayaspell' لتطوير دعم العربية في مدقق 'Hunspell'. كل هذه البرامج يمكن الاطلاع على نصوصها المصدرية وبالتالي فهم الطريقة والمنهجية المتبعة أثناء البرمجة وكيفية استعمالها للقواميس المعتمدة أثناء التدقيق، مع إمكانية تجريبها وتعديلها إن استلزم الأمر، الأمر الذي قد يؤدي إلى تطوير أكثر سرعة وفعالية لهذه البرامج، جراء تضافر الجهود بين مختلف المطورين.

هذا الأمر يمكننا من خلاله توضيح دور وأهمية المصادر المفتوحة في تطوير المدققات الإملائية، وكون بعض اللغات تتشابه في الكثير من الخصائص المتعلقة بإضافة الزوائد للكلمات المتواجدة في القاموس

(الجذور) وبالتغييرات التي تطرأ على الكلمات عند تحويلها من صيغة إلى أخرى كما هو مدقق المنسبل الذي دعم في البداية اللغتين الألمانية والتركية للتشابه الكبير لهاتين اللغتين في الكثير من الصفات اللغة المجرية التي صمم من أجلها أم دعمه للعربية والتي عكف فريق 'Ayaspell' على تطوير المزايا المتعلقة بها مستعملا بعض الصفات الموجودة أصلا في هذا البرنامج ومطورا صفات خاصة بالمدقق كثيرا في تطوير برنامج المنسبل ودعمه للغة العربية الأمر الذي ما كان له أن يتحقق لولا المصادر المفتوحة لهذا البرنامج.

I-2-4-2-3. المدقق الإملائي "Buckwalter":

يعتبر قاموس بوكوولتر لصاحبه 'TIM BUCKWALTER'[1*] أول قاموس عربي مفتوح المصدر يرخص برخصة BSD يحتوي على آلاف الكلمات العربية ومشتقاتها وهو معد ومنظم بطريقة جيدة ومنهجية إذ نجد الكلمة وبجانبها نوعها حسب تصنيف صرفها وترجمتها وموضعها الإعرابي كما نجد في الحالتين: مشكولة وغير مشكولة. إفادة في إعداد هذا القاموس على ذخيرة لغوية مستخرجة من الصحف العربية.

تمكن الاستعانة بهذا القاموس من طرف جميع العاملين في ميدان البرامج الحرة لتطوير العديد من برامج التدقيق الإملائي نذكر من بينها:

- برنامج الدلي (مجموعة عيون العرب).
- قاموس بغداد لمحمد سمير.
- قاموس Hspell لمطورها Ivirix.
- برنامج Aramorph.
- تم تحويله بطريقة آلية إلى صيغة برنامج المنسبل 'Hunspell' من طرف مطور هذا الأخير.
- تم تعديله في بدايات مشروع ميسبل 'Ayaspell' لإعداد القاموس العربي للمدقق الإملائي (تم التطرق عنه فيما بعد).
- الكلمة مجردة من الشكل.
- الكلمة مشكولة شكلا تاما.
- التصنيف الصرفي للكلمة Morphological Categories.
- الترجمة إلى اللغة الإنجليزية.

II- مميزات المدقق:

يعتبر هذا القاموس من الناحية العلمية مستندا قيما إذ يحتوي على شروح صرفية لكل كلمة. يمكن تلخيص هذه المزايا في ما يلي[22]:

- كونه مفتوح المصدر.
- منظم بطريقة منهجية جيدة.

[1*] أرجعتني الأصل إلى صمم المحلل الصرفي العربي وواضع القاموس العربي الإنجليزي درس الإنجليزية لمدة 7سنوات في كل من بيكا لحم ونابلس والقاهرة متحصل على ماجستير في الدراسات العربية من جامعة إنديانا (1982).

- الكلمات متوفرة على شكلين: المشكّلة والمجرّدة من الشكل.
- التصنيف الصرفي للكلمة (Morphological category).
- كل الكلمات مترجمة إلى الإنجليزية.

ب- الجوانب السلبية:

- يحتوي على كلمات كثيرة غير مصرح بها في القواميس العربية.
- يحتوي على كثير من أسماء الأعلام والدول ويفتقر إلى بعض الكلمات العربية الأساسية.
- يحتوي على أشكال متعددة للكلمة الواحدة.
- كونه مكنز 'Corpus' مبني من مادة حقيقية فهو غير مؤهل للاستعمال كمحلل صرفي للكلمات العربية.

I-2-4-2-4. الأدوات المساعدة الإضافية:

نقصد بالمدققات العالمية مفتوحة المصدر البرامج المعروفة عالميا للتدقيق الإملائي متعدد اللغات والتي لم تطور من أجل لغة بعينها، بل بإمكانها تدقيق العديد من النصوص المكتوبة بلغات مختلفة شرط حصول اللغات على الدعم في برنامج هذا المدقق. سنقوم في الفقرات الآتية بدراسة أهم هذه المدققات مثل : 'Ispell' و 'Aspell' و 'Hunspell'. هذه الدراسة سوف تشمل مزايا هذه المدققات وطريقة عملها ومحاولة دعم العربية فيها.

I.2-4-2-4-1. مدقق "Ispell":

أ- التعريف:

هو برنامج جد قديم. البرنامج الأصلي تم كتابته من طرف 'R.E.GORIN' باستعمال ' PDP-10 assembly' سنة 1971. الإصدار بلغة C تم كتابته من طرف 'Willisson' من MIT. قام بعدها 'Walt Buehring' من 'Texas Instruments' بإضافة واجهة 'Emacs' و إرساله إلى الإنترنيت. فقام بعدها 'Geoff Kuenning' الدعم العالمي وإنجاز المنتج الحالي. قام 'Ken Stevens' بتحسين واجهة Emacs (Ispell.el). كثيرون أيضا ساهموا في إنجاز الإصدار الحالي. يمكن الإطلاع على قائمة جزئية لهم في ملف Contributors.

ب- الإيجابيات:

يعتبر Ispell موجّه شاشة 'oriented-screen' تدقيق إملائي يقوم باكتشاف الأخطاء في سياق الملف الأصلي بشكل سريع و يقترح التصحيحات الممكنة متى استطاع تأمينها. مقارنة مع تدقيق Unix فهو أسرع وأكثر سهولة عند الاستعمال. يستطيع Ispell أيضا أن يعالج لغات غير العربية.

تم تطوير دعم اللغات العالمية ودعم الفواصل لتصريف (Compilation) البرنامج على النسخة ويندوز في الإصدار 3.3 من برنامج Ispell الذي تم إصلاحه سد الكثير من الثغرات المتعلقة بالحماية وتصحيح عدد من الأخطاء (Bugs)[23].

ج- السلبيات:

- محتوى المصدر.

- انحدرت منه العديد من المدققات المفتوحة المصدر مثل: Aspell و Myspell و Hunspell.
- يدعم لغات عالمية غير الإنجليزية.
- السرعة في التدقيق.
- ضغط ملف الزوائد (Affix compression)
- سهولة التصريف والتثبيت تحت نظام لينكس.
- يعمل تحت أنظمة OS X أو أنظمة أكثر حداثة.

د- سلبيات:

- لا يدعم اللغة العربية.
- محدودية قائمة الاقتراحات في حالة الأخطاء.
- صعوبة التصريف والتثبيت تحت الإصدارات القديمة لنظام تشغيل MacOs.
- غير مدعوم بصفة رسمية على أنظمة مايكروسوفت 'Ms-Dos' أو 'Windows'.

هـ- دعم اللغة العربية في مدققات الإملاء:

لا توجد أي محاولة لدعم العربية في هذا المدقق.

I.2-4-2-4-2. مدققات "Aspell":

إن مدقق السبل "GNU ASPELL" هو مدقق إملائي مفتوح المصدر تم تطويره من طرف " Kevin Atkinson" والذي وضعه تحت الرخصة العمومية «GNU General Public Licence» ليعود في الأساس مدقق "ISPELL" الذي يعتبر سلف المدققات الإملائية مفتوحة المصدر. وقد تمت عدة محاولات لدعم العربية فيه اعتمدت أغلبها على قاموس تيم بولكوولتر[24].

يمكن استعماله كبرنامج مستقل أو كمكتبة حيوية لبرامج أخرى ويتميز بكونه برنامجا متعدد اللغات ويعمل على الكثير من الأنظمة ويدعم الاتصال بالعديد من البرامج.

و- سلبياته:

- كونه مفتوح المصدر.
- عرض جيد ودقيق لقائمة الكلمات المقترحة لتعويض الكلمات الخاطئة.
- يمكن أن يستعمل على شكل مكتبة "Library" في برنامج آخر أو على شكل مدقق إملائي مستقل.
- يستطيع تدقيق المستندات والملفات التي تستعمل الترميز العالمي الموحد (يونيكود).
- يأخذ بعين الاعتبار الإعدادات المحلية الحالية "Current local settings".
- يتيح استعمال عدة قواميس في وقت واحد مع التحكم التام وبشكل ذكي في القواميس الشخصية عندما تكون أكثر من وحدة برمجية "Process" مشغلة من برنامج السبل "ASPELL".
- يعمل تحت نظامي ويندوز ولينكس.
- يتوفر على قواميس مصغرة في حوالي 70 لغة.
- سهولة صياغة مثل الزوائد وأشكالها.

- يمكنه التعلم من الأخطاء المستخدمين.
- يدعم الكلمات المركبة.
- إمكانية صياغة القواميس الخاصة.
- يدعم تدقيق العديد من أنواع الملفات HTML, TeX, txt.

ب- أساليب:

من المشهور مساوئ مدقق أسبليو أن يكون خوارزم الربط التقريبي 'approximate-matching algorithm' المستعمل عند اقتراح الأخطاء أو خوارزم f ا باللغة الإنجليزية.

ج- الأليات المستخدمة في التدقيق:

يعتمد مدقق أسبل عند القيام بعملية التدقيق الإملائي على ملفين أساسيين: ملف قائمة الكلمات (ar.wl) وملف للزوائد (ar_affix.dat)، كما يحتوي على ملفات أخرى تشمل إعدادات اللغة.

يحتوي ملف الزوائد على قائمة الزوائد الممكنة وقواعد تطبيقها. أما ملف قائمة الكلمات فيحتوي على الجذوع (stems) والمفردات مرفقة بالقواعد التي يمكن تطبيقها على كل مفردة.

عند تدقيق الكلمات يقوم مدقق أسبل بإتباع الخطوات الآتية :

- تحميل ملفات اللغة.
- تحويل النص f إلى كلمات.
- البحث عن السابقة حسب الحرف الأول*.
- أخذ السوابق واللواحق المعدومة بعين الاعتبار*.
- حذف السابقة وإعادة الكلمة إلى أصلها بإضافة المحذوف*.
- البحث عن الجذع في قائمة الكلمات*.
- التحقق من مطابقة الزوائد مع الجذع في حالة وجود هذا الأخير في القاموس*.

- **أ- الخوارزمية:**

- العمليات المرفقة بالرمز (*) تتم إعادتها بنفس التسلسل من أجل كل كلمة مكونة للنص f المراد تدقيقه.
- من أسرار نجاح هذا المدقق في قيامه بتحديد f الأخطاء " prononcia tion" الكلمات المكتوبة بشكل خاطئ من أجل اقتراح التصحيحات الأكثر تلاؤما لتعويض الكلمة الخاطئة.

المثال التالي يوضح ذلك :

عند كتابة كلمة 'trubble' فمدقق Ispell يقترح فقط كلمة 'rubble' بينما يقوم Aspell باقتراح 'trouble' كأول اختيار يليها كل من 'dribble' و 'rubble' بالإضافة إلى كلمات أخرى كثيرة.

د- أدوات دعم اللغة العربية:

قامت عدة فرق بمحاولة دعم العربية في برنامج أسبليو من خلال توفير قاموس عربي f ا بذلك وقد اعتمدت هذه التجارب في جلها على قاموس بوكوولتر نذكر منها:

❖ **تركيز محاولة محمد سمير[2]* على إنجاز قائمة كلمات تستعمل كقاموس مرجعي بالنسبة لمدقق السبل بعد حصول دعم العربية في هذا الأخير.** من أهم المراحل التي مرت بها هذه المحاولة نذكر:

- تجريب مدقق سبل في تدقيق النصوص العربية بنفسه.
- محاولة استغلال قائمة كلمات ذات صيغة معينة تحتوي على مفردات اللغة العربية الحديثة المستعملة. في هذا السياق يمكننا إدراج المحاولتين التاليتين:
  - **استغلال القرآن الكريم:** استعان بالقرآن الكريم لاستغلال قائمة كلمات (wordlist).

    المساوئ:
    - ✓ قائمة كلمات مكتوبة بطرق مختلفة ليس من السهل تمييزها.
    - ✓ كلمات غير صحيحة خارج السياق الذي كتبت فيه (القرآن الكريم).
    - ✓ ملء زوائد يخم ليس من السهل إجراء تعديلات عليها و بالتالي من الصعب تجريد الكلمات المتحصل عليها.
  - **استغلال قاموس عربي:**

    المساوئ:
    - ✓ عدم التأكد من صحة هذا القاموس كونه مبني من مادة صحفية.
    - ✓ عدم القدرة على تحديد عدد الكلمات القديمة فيه.
    - ✓ مشكلة حجم هذا القاموس الذي قد يصل إلى 200 ميغا بايت.
- **استراتيجية تدقيق إملائي:**
  - ✓ الحاجة إلى قائمة كلمات تحتوي على مفردات اللغة العربية الحديثة المستعملة فالعربية حسبه مليئة بالكلمات القديمة ومن يعرف هذه الكلمات ليس بحاجة إلى مدقق إملائي.
  - ✓ إنشاء جدول يضم مشتقات الكلمات وقواعد اشتقاقها فالكلمة يمكن اشتقاقها بشكل صحيح ولكنها غير صحيحة بالنسبة للقاموس العربي.
- **نتائج المحاولة:**

لم تتوصل هذه المحاولة إلى النتائج المرجوة منها والمتمثلة أساسا في دعم تدقيق العربية في مدقق سبل من خلال إنشاء قائمة كلمات تشكل القاموس المرجعي لهذا المدقق و بالإمكان إرجاع ذلك إلى مايلي:

- ✓ الافتقاد إلى آلية جد مهمة تتمثل في الرجوع إلى ملف الزوائد لاشتقاق الكلمات.
- ✓ محدودية و فقر ملف الزوائد من ناحية قواعد الإلحاق بالنسبة لبرنامج Aspell (مقارنة مع برامج أخرى مثل هنسبل).
- ✓ فشل محاولة استغلال قائمة كلمات للغة العربية يمكن استعمالها كقاموس مرجعي للمدقق.

---

[2]* من مواليد سنة 1981 بقرية كلية الصيدلة بالقارة عمل مع شركة عيون العرب للبرمجيات وأحد مؤسسي مجموعة GNU/linux الجزائرية.

❖ **مشروع "Ivrix" لتدقيق اللغة العربية:**

مشروع Ivrix الذي طوره 'Kenigsberg Dan' [3]* بعمل يهدف إلى دعم العربية في مدقق المسبل. من أهم ما تم إنجازه خلال هذا المشروع:

تحويل قاعدة معطيات اللغة العربية لقاموس تيم بوكوولتر والتي شملت كل السوابق واللواحق إلى صيغة مسبل. الأمر الذي مكنه من تدقيق اللغة العربية في هذا الأخير.

❖ **مشروع "Google" لتدقيق اللغة العربية:**

تركز محاولات فريق 'Google' تحت إشراف 'kevin Atkinson' [4]* على ما يلي:

- ✓ إنجاز قاموس عربي كبير لمدقق المسبل.
- ✓ استخلاص قائمة الكلمات الأصلية لقاموس هذا المدقق انطلاقا من قاموس بوكوولتر (إصدار 1.0).
- ✓ إصدار يحوي لأول مرة على زوائد affixe file .

- **مميزات المشروع:**

من الأشياء التي يمكن أن نعيبها على هذا المدقق نذكر ما يلي:

- ✓ اللجوء لتدعيم القاموس (بوكوولتر) بـ سبع قواميس حرة أخرى بديلة في اللغة العربية.
- ✓ ضآلة حجم المعلومات الموجودة في قائمة الكلمات المستخلصة من قاموس بوكوولتر انعكست سلبا على أداء مدقق المسبل أثناء التدقيق.

I.2.2-4-2-3. مدقق "Hunspell":

طور المدققات العالمية أساسا لتدقيق اللغة الإنجليزية ثم تطور لتضم لغات عديدة ولكن تميز بعض اللغات بمزايا لا تتوفر في الإنجليزية دفع الكثيرين إلى إضافة مزايا جديدة وفي هذا السياق ظهر hunspell الذي يعني المدقق الهنغاري (المجري) لإضافة مزايا جديدة خاصة باللغة المجرية ولغات أخرى كالألمانية والتركية. وهو مدقق مفتوح المصدر طوره المجري نامث لاسي انطلاقا من مدقق Ispell. يعد برنامج "Hunspell" في الوقت الحالي وبعد حصول دعمه لعدة لغات المدقق الإملائي الافتراضي للمجموعة المكتبية الحرة "Open Office" وبرنامج التصفح Mozilla FireFox وبرنامج البريد Mozilla Thunderbird .

ج- طريقة عمل مدقق اللغة:

يعتمد برنامج المنسبل على ملفين أساسيين للقيام بعملية التدقيق الإملائي :

---

[3]* مساعد كاتب المدقق العبري Hspell متحصل على دكتوراه من المعهد الإسرائيلي للتكنولوجيا بمدينة حيفا في سبتمبر 2007.

[4]* من مواليد 04 مايو 1977 بميديا (الولايات المتحدة) يحضر شهادة دكتوراه في علوم الحاسوب بعنوان نظام جديد للغة البرمجة.

- **ملف الكلمات (*.dic):** يتضمن قائمة كلمات اللغة التي يكتب بها النص المراد تدقيقه (كلمة واحدة في كل سطر)، يليها علم واحد أو أعلام كثيرة تمثل زوائد أو مزايا صرفية اختيارية تستعمله الجدولة.
- **ملف القواعد (*.aff):** يُعرف معنى الأعلام الصرفية في القاموس وقد يحتوي كذلك على الكثير من المزايا الاختيارية، نذكر على سبيل المثال : SET التي تستخدم لتحديد ترميزات حمل القاموس وTRY التي تحدد حروف التغيير للاقتراحات وأيضا REP التي تحدد جدول استبدال لعدة تهجيات للحروف في حالة الاقتراح.

ب- خصائص الأداة:

- ✓ كونه مفتوح المصدر.
- ✓ ترابط السوابق واللواحق المشروط.
- ✓ تركيب الكلمات.
- ✓ دعم مزايا صرفية باللألمانية والتركية.
- ✓ دعمه لليونيكود.
- ✓ يحتوي على خيارات متنوعة للتركيب.
- ✓ التعامل مع اللواحق المزدوجة.
- ✓ يستطيع معالجة أكثر من 65 000 من أنماط الزوائد.
- ✓ خيارات متنوعة لاقتراح الحلول البديلة.
- ✓ سهولة إضافة حمل الزوائد وأشكالها.
- ✓ كونه يعمل حاليا على وسع للتحليل الصرفي.
- ✓ كونه المدقق الإملائي الافتراضي لطقم أوبن أوفيس المكتبي مفتوح المصدر.
- ✓ يمكن استعماله كمدقق افتراضي لبرنامج abiword ومكتب غنوم Gnome.

ج- **خصائص أداة دعم الإملاء في اللغة العربية:**

❖ **أداة طلب إملاء ' Ayaspell ':**

بدأ طاقم ميسبل[5]* بالعمل على دعم العربية في المدقق الإملائية مفتوحة المصدر، وبعد دراسة مستفيضة تم اختيار المنسبل ليكون المنطلق بالنظر إلى مزاياه الكثيرة التي سنتطرق إليها بالتفصيل في المحور الثاني، من غيره من المدققات ليكون مدققا إملائيا للغة العربية وقرر الاستفادة من مزاياه.

وقد تمحور العمل على محاولة إعداد قاموس عربي لهذا المدقق، ومن جهة أخرى محاولة تطوير البرنامج نفسه لاستيعاب بعض المزايا الجديدة الصرفية باللغة العربية.

وقد تم تقسيم العمل إلى محورين:

- **إعداد الكلمات وضبط الصرفيات وتحديد الزوائد:**

[5]* تأسس في مارس 2006، مجال الاهتمام: البرامج الحرة، بدء مشروع ميسبل:إنشاء قاموس في اللغة العربية لمدقق المنسبل.

قام فريق بيسبل بعدة محاولات لتطوير قاموس عربي مفتوح المصدر يمكن للمدققات الإملائية الداعمة للغة العربية أن تعتمد عليه أثناء قيامها بالتدقيق الإملائي. من أهم هذه المحاولات نذكر ما يلي:

- **"Tim Buckwalter" محاولة تيم بولكوولتر:**

لتعديل قاموس تيم بولكوولتر "Tim Bulkwalter" وإدراجه في برنامج المدقق المنسبل اقترح طه زروقي[6]* إجراء سلسلة من التعديلات على هذا القاموس نلخصها فيما يلي :

- ✓ استعمال الكلمات الغير مشكولة.
- ✓ حذف الكلمات الغير عربية.
- ✓ حذف الأشكال المتعددة للكلمة الواحدة.
- ✓ حذف الحروف مثل "إلي" و"علي" والإبقاء على "على" و"إلى".
- ✓ حذف همزة القطع والإبقاء على همزة الوصل.

- **محاولة عرفان سادات:**

بعد أن تبين للعديد من الباحثين في مجال التدقيق الإملائي العربي عدم ملاءمة قاموس بوكوولتر قرر طاقم بيسبل العمل على تطوير قاموس عربي جديد مفتوح المصدر ومستمد من عدة قواميس رائجة الاستعمال في الأوساط اللغوية العربية.

- **محاولات الفريق أحمد فرحات لإصلاح سادات:**

المحاولات التي تكفل بها طه زروقي والتي جاءت من أجل معالجة بعض المسائل التي وجد برنامج المنسبل الأصلي صعوبة في التعامل معها والتي تتلخص في ما يلي:

- ✓ إشكال الشكل.
- ✓ إشكال التطويل.
- ✓ الإدغام.
- ✓ الإعلال والإبدال.
- ✓ الزوائد المزدوجة.

هذه المزايا ساهمت في تكيف هذا البرنامج مع الكثير من خصائص اللغة العربية وقواعد الصرف العربي. وهذه الأعمال هي محل موضوع بحثنا لذا سنتحدث عنها بالتفصيل في الفصول القادمة.

I-2-4-2-5. الأدوات المستخدمة في التدقيق الإملائي:

هناك بعض برامج التدقيق الإملائي التي طورت خصيصا من أجل اللغة العربية نذكر منها برنامج **Aramorph** و برنامجي الدالي و بغدادي الذين أنتجتهما شركة عيون العرب للبرمجيات.

I.2.4.2.5.1. "Aramorph" برنامج:

- **أ- تعريفه:**

---

[6]* مستشار في الإعلام الآلي متحصل على شهادة ماجستير في الإعلام الآلي من المعهد الوطني للإعلام الآلي. بالجزائر يحضر شهادة دكتوراه حول موضوع "بنية نظام المستندات الإلكترونية للقرآن وعلومه:مقاربة أنطولوجية".

يعتبر ARAMORPH إعادة برمجة بلغة (JAVA) للمنتوج المشابه المطور بلغة Perl من طرف TIM BUCKWALTER لصالح LDC (Linguistic DATA Consortium) والذي يمكن معاينته على الويب على العنوان التالي: http//www.LDC.UPENN.edu/catalog/catalog enrg.JSP ?catalog Id=LDC2002 L49

يحتوي المشروع على برنامج بلغة "JAVA" التي تتيح التحليل الصرفي لملفات النصوص العربية. يمكن استعمال ترميزات مختلفة UTF-8 وISO8859-6 وCP 1256.

ب- طريقة عمل « Aramorph »:

قبل كل شيء يجب معرفة أن "Aramorph" كما كان برنامجه الأصلي بلغة " Perl " يعمل على الكتابة اللاتينية للكلمة العربية التي تستعمل بصفة طبيعية نظام الكتابة اللاتينية لتيم بوكوولتر.

فعملية التحليل الصرفي لكلمة 'كتاب' على سبيل المثال تشمل المراحل الآتية:

1. تحول الكلمة إلى كتابتها اللاتينية أي (ktab) قبل تحليلها صرفيا (هذه العملية ليست ضرورية لأن Java يعمل بنظام اليونيكود للتشكيل).
2. يقوم "Aramorph" باستعمال خوارزم جد فعال لتجزئة الكلمة إلى سلسلة متكونة من سابقة وجذر ولاحقة على النحو التالي:

سابقة	جذر	لاحقة
ktAb	Ø	Ø
ktA	b	Ø
ktA	Ø	b
kt	Ab	Ø
kt	A	b
kt	Ø	Ab
k	tAb	Ø
k	tA	b
k	t	Ab
k	Ø	tAb
Ø	ktAb	Ø
Ø	ktA	b
Ø	kt	Ab
Ø	k	tAb
Ø	Ø	ktAb

3. يبحث "Aramorph" عن وجود كل من العناصر السابقة في ثلاثة قواميس:

السابقة في: dictPrefixes

الجذر في: dictStems

اللاحقة في: dictSuffixes

4. في حالة الحصول على نتيجة يقدم Aramorph المعلومة الصرفية لكل من العناصر السابقة.

- **مخرجات:**

السوابق واللواحق ∅ لها معنى من الناحية الصرفية. في حالة التشكيل يبحث Aramorph إذا ما كان صرف كل من العناصر تتلاءم مع بعضها البعض وذلك بتحديد ثلاثة جداول للتوافق:

- ✓ بين السابقة والجذر في: tableAB.
- ✓ بين السابقة واللاحقة في:tableAC .
- ✓ بين الجذر واللاحقة في: tableBC.

الكلمة الصحيحة في النتيجة هي الكلمة التي يكون فيها:

- ✓ كل من السابقة والجذر واللاحقة له مدخل في القاموس.
- ✓ كل من السابقة والجذر واللاحقة متوافقون صرفيا فيما بينهم.

بالنسبة لكلمة "كتابي" في المثال السابق هناك ثلاثة حلول ممكنة كما نراها في الأعلى.

- **الخوارزمية:**

- ✓ بعض المعلومات الموجودة في قاموس الجذور هي في الحقيقة معلومات ذات صلة بالسوابق أو اللواحق.
- ✓ عندما يقترح "Aramorph" التصحيحات يحاول تغيير هذه المعلومات من أجل السوابق أو اللواحق.
- ✓ من الممكن أن يكون للكلمة الواحدة عدة سوابق أو لواحق.

4- إضافة السوابق:

- **الهدف:**

لنتحقق إذا الجزء من قاموس السوابق: dictPrefixes

```
; conjunctions
w       wa      Pref-Wa and <pos>wa/CONJ+</pos>
f       fa      Pref-Wa and;so <pos>fa/CONJ+</pos>
```

نلاحظ بأن التعاليق تسبق بالرمز ";" بينما الأعمدة مفصولة بواسطة الجدولة وهي تمثل على الترتيب :

- ✓ البنية الصائتية " squelette consonnatique " للسابقة حسب نظام الكتابة اللاتينية لتيم بوكوولتر.
- ✓ تشكيل السابقة "vocalisation" حسب نظام السامي.
- ✓ التصنيف الصرفي للسابقة.
- ✓ ترجمة واحدة أو عدة ترجمات للسابقة متبوعة بتصنيف نحوي واحد أو عدة تصنيفات نحوية. ثم بأن الرمز '+' يوضح بأن هناك جذر منتظر بعد السابقة.

- **مخرجات:**

بعض المعلومات الاختيارية المثال الأكثر وضوحا هو مثال السابقة التاريخية :

```
; The first category is the null prefix (has a null gloss as well):
Pref-0
```

حين يظهر فقط التحليل الصرفي.

- **الملف:**

لنتحقق الآن هذا الجزء من قاموس اللواحق : dictSuffixes

```
; perfect verb, null suffix: banA-h, daEA-h
h          hu          PVSuff-0ah  he/it <verb> it/him        <pos>+(null)/PVSUFF_SUBJ:3MS+hu/PVSUFF_DO:3MS</pos>
hmA        humA        PVSuff-0ah  he/it <verb> them (both)   <pos>+(null)/PVSUFF_SUBJ:3MS+humA/PVSUFF_DO:3D</pos>
hm         hum         PVSuff-0ah  he/it <verb> them          <pos>+(null)/PVSUFF_SUBJ:3MS+hum/PVSUFF_DO:3MP</pos>
hA         hA          PVSuff-0ah  he/it <verb> it/them/her   <pos>+(null)/PVSUFF_SUBJ:3MS+hA/PVSUFF_DO:3FS</pos>
hn         hun~a       PVSuff-0ah  he/it <verb> them          <pos>+(null)/PVSUFF_SUBJ:3MS+hun~a/PVSUFF_DO:3FP</pos>
k          ka          PVSuff-0ah  he/it <verb> you           <pos>+(null)/PVSUFF_SUBJ:3MS+ka/PVSUFF_DO:2MS</pos>
k          ki          PVSuff-0ah  he/it <verb> you           <pos>+(null)/PVSUFF_SUBJ:3MS+ki/PVSUFF_DO:2FS</pos>
kmA        kumA        PVSuff-0ah  he/it <verb> you (both)    <pos>+(null)/PVSUFF_SUBJ:3MS+kumA/PVSUFF_DO:2D</pos>
km         kum         PVSuff-0ah  he/it <verb> you           <pos>+(null)/PVSUFF_SUBJ:3MS+kum/PVSUFF_DO:2MP</pos>
kn         kun~a       PVSuff-0ah  he/it <verb> you           <pos>+(null)/PVSUFF_SUBJ:3MS+kun~a/PVSUFF_DO:2FP</pos>
ny         niy         PVSuff-0ah  he/it <verb> me            <pos>+(null)/PVSUFF_SUBJ:3MS+niy/PVSUFF_DO:1S</pos>
nA         nA          PVSuff-0ah  he/it <verb> us            <pos>+(null)/PVSUFF_SUBJ:3MS+nA/PVSUFF_DO:1P</pos>
```

المبدل هو نفسه رغم أن المثال أكثر تعقيد نسبيا. لدينا في الواقع تسلسل مزدوج للواحق الأول هو اللاحقة "ـضمير الغائب للمفرد المذكر في الماضي والثاني هو نسبة إلى ضمير متصل في محل نصب مفعول به. نسجل الرمز "+" للربط مع اللاحقة "∅".

- **الشكل:**

الشكل مختلف نوعا ما لأن السطر المسبوق بـ ";;" يعطي تعريفا للجذر. الباقي تقريبا متشابه.

- **الخلاصة:** يمكن من الأمثلة السابقة أن نلاحظ ما يلي:
  - ✓ التحليل النحوي مفقود في الغالب لأنه يمكن استنباطه من التحليل الصرفي.
  - ✓ يجب تحديد التحليل النحوي أحيانا لأن التحليل الصرفي مثل النسبة اسمية صرفيا 'morphologicalement nominale' يمكن أن تتيح المجال لاستعمالها نعتيه.
  - ✓ المحور الدائم للكتابة اللغوية حسب نظام تيم بوكوولتر.

I.2-5-2-4-2. تطبيق الإملاء على الدارجة " الجزائرية":

أ- الهدف:

سمي بهذا الاسم نسبة إلى مقعد قواعد اللغة العربية الشهير "أبو الأسود الدؤلي"d.688". انطلق هذا المشروع بتاريخ 31 جانفي 2002 وكان هدفه الأساسي يتمثل في تطوير مدقق إملائي عربي في اللغة العربية (مع إمكانية التطوير لملائمة لغات أخرى غير عربية).

ب- الفريق العامل:

- ✓ المطور: محمد الزبير.
- ✓ لغة البرمجة: بايثون / C .
- ✓ نظام البرمجة: لينكس / متعدد الأنظمة.
- ✓ الترخيص : BSD.
- ✓ الجهة الداعمة: عيون العرب Arabeyes.

- ✓ الموقع: http : //www.arabeyes.org//project.php ?proj=Duali
- ✓ القاموس: TIM BULKWALTER
- ✓ الوصف: دعم لغة الضاد لأول مرة.

ج- **خصائص:**

- ✓ مدقق عربي منذ البداية.
- ✓ تم الاستفادة منه لاستنباط بعض الأفعال المتعلقة بقضايا التشكيل والتطويل.
- ✓ مبرمج بلغة بايثون.
- ✓ يستعمل مكنز بولكولتر.

د- **مراحل التطوير:**

- ✓ متوقف التطوير.

ذ- **آلية عمل المدقق الإملائي:**

عند معالجة ملف ما يقوم مدقق دالي بإتباع الأنماط التالية:

- ✓ فصل الزوائد (السوابق واللواحق) عن الكلمة المراد تصحيحها.
- ✓ فحص الكلمة المتصل عليها للتأكد من كونها جذرا صحيحا.
- ✓ مقارنة الجذر مع ما هو موجود في المعجم (المعجم يحتوي فقط على الجذور بالإضافة إلى كلمات ثابتة (أدوات حروف) fixed terms وأفعال معربة arabized terms).
- ✓ في حالة عدم وجود المقارنة يتم إعادة تركيب الزوائد والمقارنة من جديد مع ما هو موجود في المعجم.
- ✓ في حالة عدم الوجود مرة أخرى أن الكلمة على أنها كلمة خاطئة.

I.2-2-4-5-3. **مراحل المدقق الإملائي " أوفيس 2007 لمدقق الإملائي":**

أ- الإصدار:

هو عبارة عن إعادة برمجة لبرنامج الدالي من لغة البايثون إلى لغة C++ قام بها محمد سمير.

ب- أدوات التطوير:

من أهم أدوات مدقق بغداد نذكر ما يلي:

- ✓ **المبرمج:** محمد سمير.
- ✓ **لغة البرمجة:** C .
- ✓ **نظام التشغيل:** لينكس / متعدد الأنظمة.
- ✓ **الترخيص:** BSD .
- ✓ **الداعم المالي:** عيون العرب Arabeyes.
- ✓ **الموقع:** http : //www.arabeyes.org/cgi-bin/viecvs.cgi/projects/baghdad
- ✓ **المشرف:** TIM BULKWALTER.

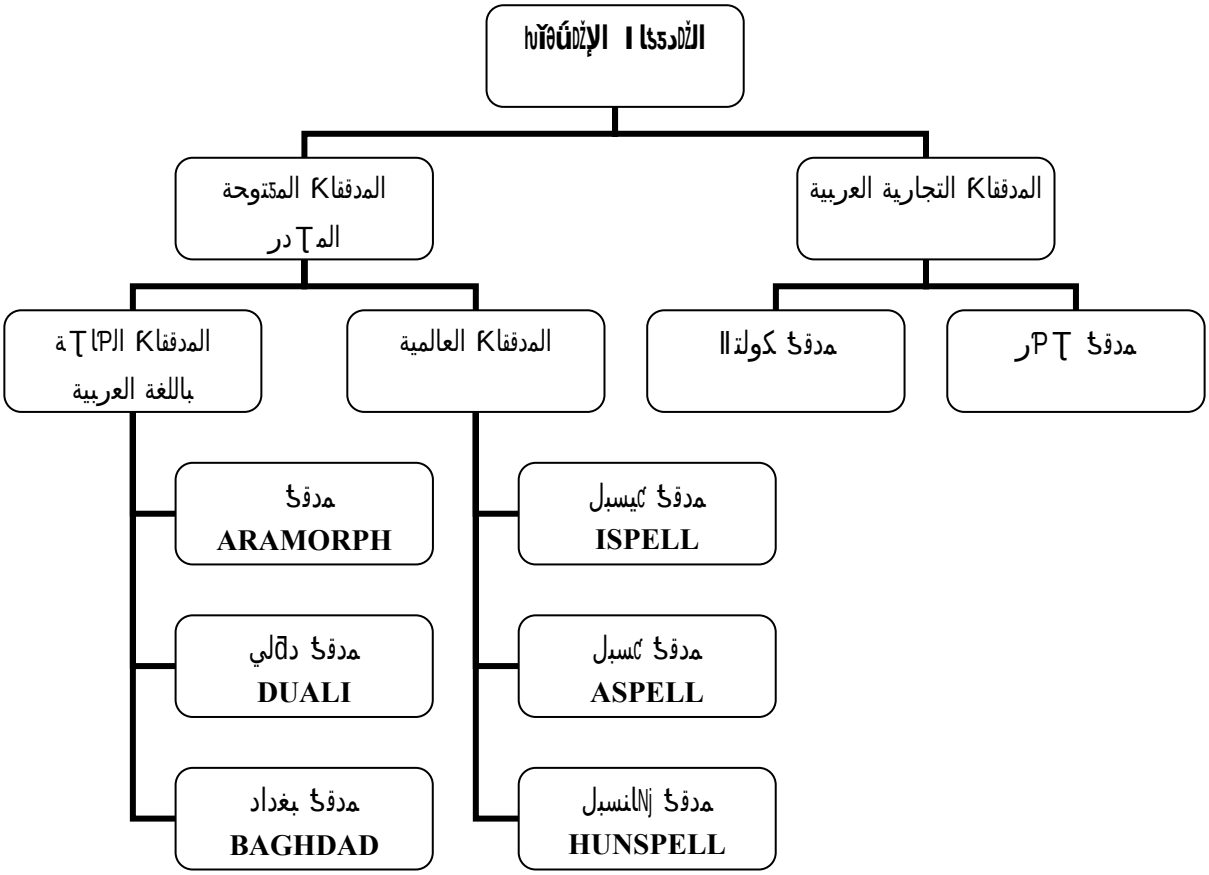
- ✓ **مستوى التطوير**: متوسط.

أ - المعالجة:

مثل بقية المدققات الداعمة للغة العربية واجه مدقق بغداد مشكلة القاموس (قائمة الكلمات) وعدم التلاؤم التام بين هذا الأخير من مزايا البرمجية لذلك يعتبر محمد سمير هذا المدقق أكثر جاهزية ليكون مدققا نحويا من أن يكون مدققا إملائيا. لأنه يتميز ببعض المزايا التي يمكن استغلالها من اجل دراسة الكلمة حسب موقعها في الجملة أو حسب السياق الذي وردت فيه أكثر من الأخطاء التي تتعلق ببنية الكلمات والتي تعتبر من اختصاص المدقق الإملائي.

5-2.I - رصد أهم المدققات الإملائية:

الرسم التوضيحي التالي يوضح تدرج المدققات الإملائية التي تم التطرق لها خلال هذا الفصل:



- **الإشكالية:**

من خلال دراسة المدققات الإملائية للغة العربية تبين لنا أن استخدام المدققات التجارية لا يفي بمتطلبات البحث والتطوير على الرغم من قدراتها العالية بسبب القيود على الملكية الفكرية لذا فإننا إلى التوجه نحو البرمجيات المفتوحة المصدر.

إننا بعد المحاولات الجيدة في مجال المعالجة الآلية للغة العربية مفتوحة المصدر والتي درسنا بعضا منها والتي اعتمدت كثير منها على مكنز تيم بوكوولتر كما نوه حتى بعد تعديله ماعدا محاولة طاقم يسبل لإعداد المدقق العربي المفتوح المصدر. هذه الأخيرة التي يجدر التنويه بها لأنها ترمي إلى إعداد قاموس عربي مرجعي أول له مصادرها وطريقة إنشائه لأن يكون قاموس المدقق الإملائي العربي المزمع تطويرها. من هنا يتضح لنا جليا أهمية هذا النوع من البرامج (المفتوحة المصدر) وفائدة في هذا المجال الذي يتطلب عملا جبارا وتواصلا لمختلف الجهود للوصول إلى نتيجة تلبي حاجيات مستعملي اللغة العربية وتنال رضاهم.

ومن خلال دراسة المدققات الإملائية المفتوحة المصدر فضلنا التركيز على برنامج "Hunspell" وبدرجة أقل برنامج "Aspell" للاعتماد عليهما كنقطة انطلاق لمشروع إعداد المدقق الإملائي العربي المفتوح المصدر لذا سنخصص الفصل القادم للحديث عن أهم مميزات ومكونات وطريقة عمل كل منهما مع مزيد من التفاصيل.

الفصل الثاني

دراسة مدققي الهنسبل 'Hunspell' وأسبل 'Aspell' الإملائيين

II. الفصل الثاني: دراسة مدققي المنسبل 'Hunspell' ومسبل 'Aspell' الإملائيين

- دراسة:

بعد دراسة التجارب العملية لمختلف المدققات الأملائية المفتوحة المصدر بمختلف أنواعها وبعد توضيح مزايا ومساوئ كل منها والتعرف على طريقة عملها حاولنا الاعتماد على كل من برنامجي المنسبيل 'Hunspell' (المدقق الافتراضي للمجموعة المكتبية OpenOffice) وبدرجة أقل برنامج مسبل 'Aspell' المفتوحي المصدر والمنحدرين من نفس البرنامج الأصلي ميسبل 'Ispell' والذي يعتبر هو المدقق المفتوحة المصدر لتكوين النواة الأساسية لمشروع تطوير المدقق الإملائي العربي المفتوح المصدر (هدف هذا الدراسة) معتمدين على بعض المزايا الموجودة في برنامجي المنسبل ومسبل الأصليين هذه المزايا التي شجعتنا أكثر للاعتماد على هذين البرنامجين كنقطة انطلاق لهذا المشروع بالإضافة إلى بعض المزايا التي تمت إضافتها (في مدقق المنسبل) لتطوير البرنامج للتلائم مع بعض خصائص اللغة العربية والتي سنتحدث عن تفاصيلها في الفصل القادم.

افتتاحية هذا الفصل ستكون بالحديث عن ترميز اليونيكود الذي يدعمه كلا المدققين ننتقل بعدها إلى التعريف بالمدققين من خلال ذكر مزايا كل منهما وأسمائهما ومحتوياتهما ملفاتهما القواميس والزوائد في كل منهما وطريقة عملهما عند تدقيق النص و ف تليها دراسة مفصلة لأهم ف وأعلام برنامج المنسبل وحالات استعمال كل منها مع التركيز على بعض الف التي تتلائم مع مزايا اللغة العربية والتي يمكن الاستفادة منها في إنجاز هذا المشروع. لنختم هذا الفصل بنقد شامل لبرنامج المنسبل يتم من خلاله تحديد النقائص المكتشفة في هذا البرنامج عند تدقيق النص و ف العربية والتي استدعت تطويرها فيما بعد.

II.1. الترميز الإلكتروني موحد الشيفرات 'Unicode':

تم تطوير هذا الترميز بالتعاون بين كل من كونسورتيوم يونيكود (Consortium Unicode) والمنظمة الدولية للتوحيد القياسي (ISO). يعتبر يونيكود (نظام الترميز العالمي الموحد) محاولة لتوحيد الجديات ورموز (iodeographs) لغات العالم في مجموعة محارف عالمية واحدة. يركز اليونيكود على المحارف نفسها بدلاً من اللغات. وهكذا فالحرف المشترك بين اللغة الإنجليزية والروسية على سبيل المثال سوف يكون له نفس المحرف في ترميز اليونيكود. كمعيار متعدد اللغات يأتي يونيكود للمطورين إنجاز التطبيقات دون تغيير الوقت في عملية إنجاز النسق المحلية لكل لغة.

معظم مجموعات محارف اللغات الغربية هي محارف 7 بت (مثل: US ASCII) أو 8 بت (مثل: Latin-1) يحددهم على التوالي 128 محرفا أو 256 محرفا. هذا التقييد إلى ظهور عدد كبير من المجموعات المختلفة لكل لغة. بالنسبة للغات مثل الصينية والكورية واليابانية التي تستعمل بكثرة رموز (التي تعتمد على محتوى الكلمة بدلاً من الأصوات التي تكونها) أنظمة كتابة تشمل آلاف المحارف فمجموعات المحارف 7بت أو 8بت التقليدية ليست مناسبة. من أجل ذلك ولاحتواء مجموعات محارف أنظمة

الكتابة العالمية الرئيسية يستعمل يونيكود أوليا مجموعة 16 بت التي تسمح لـ 65536 محرفا. نتيجة لذلك اليونيكود يشغل مساحة القرص مقارنة بالنسخة الذي يستعمل مجموعة محارف 8بت.

بوصفه مجموعة محارف لا يعير اليونيكود الاهتمام بالمظهر المرئي أو بالرمز المنقوش (Glyph) لمحرف معين. بدلا من ذلك فهو يحتوي فقط على رمز (code) واسم (name) لكل محرف.

الخطوط الفردية تعنى بمهام تحويل المحارف إلى رموز منقوشة بالمظاهر الدقيقة للرموز المنقوشة والذي يتطلب من الخط إلى آخر. بينما الطريقة لا يدعمها اليونيكود في الكثير من الأحيان بين النص البسيطة (plain text) والغنية (Rich text) بدلا من السماح للتطبيقات بتطبيق معالجات ومهيئات نصوصها الخاصة بها[25].

II.2. مدققو الهجاء (Aspell):

II-2-1- أهداف مدققو الهجاء:

يعنى المسبل 'GNU Aspell' مدقق إملائي وريث لبرنامج Ispell يمكن استعماله كبرنامج مستقل أو كمكتبة حيوية لبرامج أخرى ويتميز بكونه برنامجا متعدد اللغات ويعمل على الكثير من الأنظمة ويدعم الاتصال بالعديد من البرامج[26].

يحتوي برنامج المسبل على ملف قائمة الكلمات (القاموس) 'ar.wl' و ملف الزوائد 'ar_affix.dat' بالإضافة إلى ملفات خاصة بإعداد اللغة.

II-2-2- مزايا مدققو الهجاء:

من أهم مزايا المنسبيل نذكر ما يلي:

- كونه مفتوح المصدر.
- سهولة صياغة ملف الزوائد وأحرائها.
- يدعم العديد من اللغات والتطبيقات.
- يمكنه التعلم من أخطاء المستخدمين.
- يمكن استعماله كمكتبة برمجية حيوية لتطبيقات أخرى.
- يمكنه استعمال عدة قواميس في الوقت نفسه.
- يدعم الكلمات المركبة.
- استخدام الترميز العالمي الموحد اليونيكود.
- إمكانية صغط القواميس الأصلية.
- يدعم تدقيق العديد من أنواع الملفات HTML, TeX, txt.

II.3. مدققو الهجاء 'Hunspell':

II-3-1- أهداف مدققو الهجاء:

المدقق الإملائي المنسبل هو البرنامج المطور الذي يلي برنامج Myspell والذي تم تحسينه من أجل دعم مزايا إضافية من أجل اللغات الأوربية لا سيما اللغة المجرية hungarian التي طور من أجلها بداية لغة المبرمج الأصلي وكذلك لغات أخرى كالألمانية والتركية[27].

II-2-3- الخطوات العملية لإنشاء مدقق إملائي:

يتطلب برنامج المنسبل ملفين أساسيين لتحديد اللغة التي هو بصدد تدقيقها إملائيا. الملف الأول هو عبارة عن قاموس يحتوي على مفردات اللغة والثاني هو ملف زوائد يُعرّف معاني الأعلام الخاصة في القاموس.

II.1-2-3. تحضير الملف الأساسي (*.dic):

يتضمن قائمة من الكلمات واحدة في كل سطر. السطر الأول من القواميس (باستثناء القواميس الشخصية) يتضمن إحصاء تقريبي للكلمة (للاستغلال الأمثل للذاكرة). كل كلمة يمكن اختياريا أن يليها الفاصل ("/") بالإضافة إلى علم واحد أو أعلام كثيرة تمثل زوائد أو مزايا خاصة.

يرمز للعلم افتراضيا بحرف واحد (عادة أبجدي). في ملف قاموس المنسبل المنا يضاف ميادين معطيات اختيارية 'Optional data fields' تستعمل الفراغات أو الجدولة.

يمكن تمثيل هذه الميادين على النحو التالي:

- ❖ المعطيات الحرفية الافتراضية والمحددات الأخرى (المستعملة في الاقتراحات استئصال الجذع 'stemming' والتوليد الحرفي).
- **Ph**: الكتابة اللاتينية البديلة من أجل اقتراحات أفضل. هذا مفيد من أجل الكلمات الأجنبية النطق (قاموس يقوم على الاقتراحات الصوتية).

مثال:

Marseille ph:maarsayl

- **St**: الجذع (stem). اختياري: الجذع الافتراضي هو أحد مواد القاموس المستعملة في التحليل الحرفي ميدان الجذع مفيد من اجل الجذوع الوهمية ' virtuel stems' (كلمات القاموس لها علم Needaffix) والاستثناءات الحرفية بدل القواعد الحرفية الجديدة المستعملة بشكل فردي.

مثال:

feet st:foot is:plural
mice st:mouse is:plural
teeth st:tooth is:plural

- ❖ أشكال الكلمات المتعددة الجذوع والتي تحتاج عدة مواد من القاموس.

مثال:

lay po:verb st:lie is:past_2
lay po:verb is:present
lay po:noun

- **al**: `Allomorph(s)`. إن أي مادة من مواد القاموس تمثل جذعا لكل الكلمات المشتقة 'allmorphms' من هذه المادة. التعريف يحتاج: الجذع الكلمات المشتقة وميادين الزوائد.

مثال:

sing al:sang al:sung
sang st:sing

- **po**: جزء من نمط الإملاء.
- **ds**:'derivationnel suffixes'. استئصال الجذع 'Stemming' لا يحذف اللواحق الاشتقاقية. التوليد الصرفي يتعامل بترتيب ميادين اللاحقة.

في قواعد الزوائد 'Affix rules':

SFX Y Y 1
SFX Y 0 ly . ds:ly_adj

في القاموس 'Dictionary':

ably st:able ds:ly_adj
able al:ably

- **is**:Inflectional suffix(es) (اللواحق الأعرابية): كل اللواحق الأعرابية يتم حذفها بواسطة استئصال الجذع. التوليد الصرفي يتعامل بترتيب ميادين اللاحقة.

مثال:

feet st:foot is:plural

- **ts**:Terminal suffix(es) (اللاحقة الطرفية): ميادين اللواحق الطرفية هي ميادين اللواحق الأعرابية المحذوفة بواسطة اللواحق الأضافية 'additional suffixes' (وليس الطرفية).

مفيدة من اجل الكلمات المعدومة 'zero morphemes' والزوائد المحذوفة بواسطة قواعد التقسيم.

مثال:

work/D ts:present

SFX D Y 2
SFX D 0 ed . is:past_1
SFX D 0 ed . is:past_2

المثال النموذجي بالنسبة للاحقة الطرفية هي والكلمة المعدومة في الحالة الأسمية.

- **sp**:Surface prefix (السابقة السطحية): حل مماثل لأضافة سوابق للجذع وتوليد لأشكال الكلمة.
- **pa**: لأجزاء الكلمات المركبة: الميادين الناتجة عن التحليل الصرفي لأستئصال الجذع.
- **dp**: سابقة اشتقاقية. (قيد التخطيط).
- **ip**: سابقة إعرابية. (قيد التخطيط).
- **tp**: سابقة طرفية. (قيد التخطيط).

II.3-2-2. ملف الإعدادات (*.aff):

قد يحتوي على الكثير من الخيارات الإختيارية. من أهم استعماليته نذكر ما يلي:

- إعداد ترميز محارف الزوائد وملفات القاموس.
- تحديد الحروف البديلة بالنسبة للاقتراحات.
- تحديد جدول للتبديلات لتصحيح العديد من المحارف في وقت الإقتراحات.
- تعريف أقسام السوابق واللواحق مسماة بأعلام الزوائد.

مثال1:

إذا المثال يستخدم كلا من الملفات f الآتية:

- SET : لتحديد ترميزات حروف الزوائد و ملفات القاموس.
- TRY :لتحديد حروف التغيير للاقتراحات.
- REP : لتحديد جدول استبدال لعدة تهجيات للحروف في حالة الإقتراح.
- PFX و SFX : لتعريف أقسام السوابق واللواحق المسماة من أعلام الزوائد.

```
SET UTF-8
TRY esianrtolcdugmphbyfvkwzESIANRTOLCDUGMPHBYFVKWZ’

REP 2
REP f ph
REP ph f
PFX A Y 1
PFX A 0 re .

SFX B Y 2
SFX B 0 ed [^y]
SFX B y ied y
```

ملف الزوائد السابق مثلا يعرف ما يلي:

- نظام تشفير الحروف UTF-8 .
- اقتراحات ’TRY’ التي تختلف عن الكلمة الخاطئة بحرف انكليزي واحد أو فاصلة 'Apostrophe'.
- تعاريف الخاصية 'REP' التي تتيح لبرنامج المنسبل اقتراح الشكل الصحيح للكلمات عندما تتضمن الكلمة الخاطئة 'f' بدلا من 'ph' والعكس بالعكس.

هناك فئتان من الزوائد في القاموس:

- الفئة 'A' تعرف السابقة '-re'.
- الفئة 'B' تعرف اللاحقتي '-ed'.

أول لاحقة يمكن أن تكون إلا كلمة ماثلة عندما يكون الحرف الأخير يختلف عن 'Y'. أما اللاحقة الثانية فيمكن أن تكون فقط إلا الكلمات التي تنتهي بحرف 'Y'.

مثال:

مثال القاموس التالي يستخدم هذه الأقسام من الزوائد.

```
hello
try/B
work/AB
```

الكلمات المقبولة من هذا المثال هي:

reworked، rework، worked، work، tried، try، hello

II.3-3-3- تنسيق ملف الزوائد:

II.3-3-1. الزوائد الأمامية:

من أهم المزايا العامة لبرنامج المنسبل يمكننا أن نذكر:

- كونه مفتوح المصدر.
- استخدام الترميز العالمي الموحد اليونيكود.
- سهولة إضافة مثال الزوائد وأمثلتها.
- كونه يعمل حاليا على وينداوز للتحليل الصرفي.
- كونه المدقق الإملائي الافتراضي لطقم أوبن أوفيس المكتبي مفتوح المصدر.
- يمكن استعماله كمدقق افتراضي لبرنامج abiword ومكتب غنوم Gnome.
- يدعم إمكانية التلازم الشرطي للسوابق باللواحق.

II.3-3-2. الزوائد الأمامية:

يتميز برنامج مدقق المنسبال ببعض المزايا الهامة التي شجعتنا للاعتماد عليه ليكون مدققا إملائيا عربيـا في مستوى تطلعات مستخدمي البرمجيات العربية في جميع أنحاء العالم. هذه المزايا هي مزايا برمجية في الأساس، مكننا المصدر المفتوح لبرنامج المنسبل من الاطلاع عليها عن كثب وفهـم طريقـة اسـتعمالها وإمكانية تطويرها إن لزم الأمر. من أهم هذه المزايا نذكر:

- السوابق المركبة (**ComplexPrefixes**) : تحدد سابقة ثنائية النوع (لكن اللاحقة أحادية النوع) و هي ميزة قد بعض اللغات مثل القبطية، في البداية اعتقد أنها ملائمة للعربية، لكن طه زروقي يرى أن هذا غير صحيح.
- الزوائد السياقية : و هي الزوائد المماثلة بالعلم **CIRCUMFIX** قد تكون على كلمة عندما تكون هذه الكلمة لها أيضا سابقة من العلم CIRCUMFIX والعكس بالعكس.
- تجاهل الحروف (**IGNORE Characters**): تجاهل حروف من قاموس الكلمات واللواحق و الكلمات المدخلة. مفيدة للحروف الاختيارية، مثل علامات التشكيل في اللغة العربية (الحركات).
- تعريف لاحقة مزدوجة : الهدف من زم الأصلي لآيسبل 'Ispell' يعرف لاحقة واحدة فقط. يستطيع المنسبل أن يعرف لاحقة أخرى. تعريف اللواحق المزدوجة يشكل تحسنا كبيرا في التعامل مع عدد هائل من اللواحق، التي تميز اللغات الالتصاقية.

مثال:

سيتم إضافة لاحقة ثانية (اللاحقة Y سوف يطبق المتمم للاحقة 'able').

ملف الزوائد (ar.aff):

```
SFX Y Y 1
SFX Y 0 s . +PLUR

SFX X Y 1
SFX X 0 able/Y . +ABLE
```

ملف القاموس (ar.dic):

```
drink/X   [VERB]
```

ملف التجارب (test file):

```
drink
drinkable
drinkables
```

التجربة (test):

```
$ hunmorph test.aff test.dic test.txt
 drink:       drink[VERB]
 drinkable:   drink[VERB]+ABLE
 drinkables: drink[VERB]+ABLE+PLUR
```

كان هذا عرضا لأهم الفلاتر برنامج المنسبل التي لوحظ بأنه يمكن الاستفادة منها لمعالجة بعض حالات التدقيق الإملائي العربي ولا يزال هناك الكثير الفلاتر والأعلام التي يستعملها هذا الأخير والتي سندرسها بالتفصيل في الفقرات القادمة على شكل مجموعات من الأعلام كل على حدة حسب نوع هذه الأعلام والأدوار التي تؤديها إلى الجمهور واستعمالها والدور الذي تلعبه كل منها.

II.4. طرق عمل المدققين الإملائيين وتقييمهم:

سنتحدث في هذه الفقرة عن طريقة عمل المدققات الإملائية على العموم وفي بالذكر المدققان الإملائيان مسبل والمنسبل المفتوحي المصدر. المدققات المفتوحة المصدر تمكننا من الاطلاع على كيفية عملها مباشرة من مصدرها (code source) وإمكانية تغييرها وكذلك تعطينا جميع المعلومات من أجل بناء قاموس لغة معينة. أما المدققات التجارية فلا تسمح لنا بالاطلاع على خوارزميتها ولا على كيفية عملها وقواميسها عادة إلا صيغة معينة غير قابلة للاطلاع [BOYD A., 2006]. هذه المميزات بالمدققات المفتوحة المصدر شكلت عاملا تحفيزيا هاما لدراسة هذين المدققين هناك هذا البحث.

II.4-1- دور مكتبة مدقق مسبل في توليد الكلمات المقترحة للتصحيح الإملائي:

يمكن إستعمال مكتبة مدقق مسبل لتوليد الكلمات المقترحة للتصحيح الإملائي كما يمكن حساب رتبة 'rank' الكلمة الصحيحة في القائمة التي يقترحها مدقق مسبل. تتميز دقة مدقق مسبل عندما يتعلق الأمر برتبة الكلمة

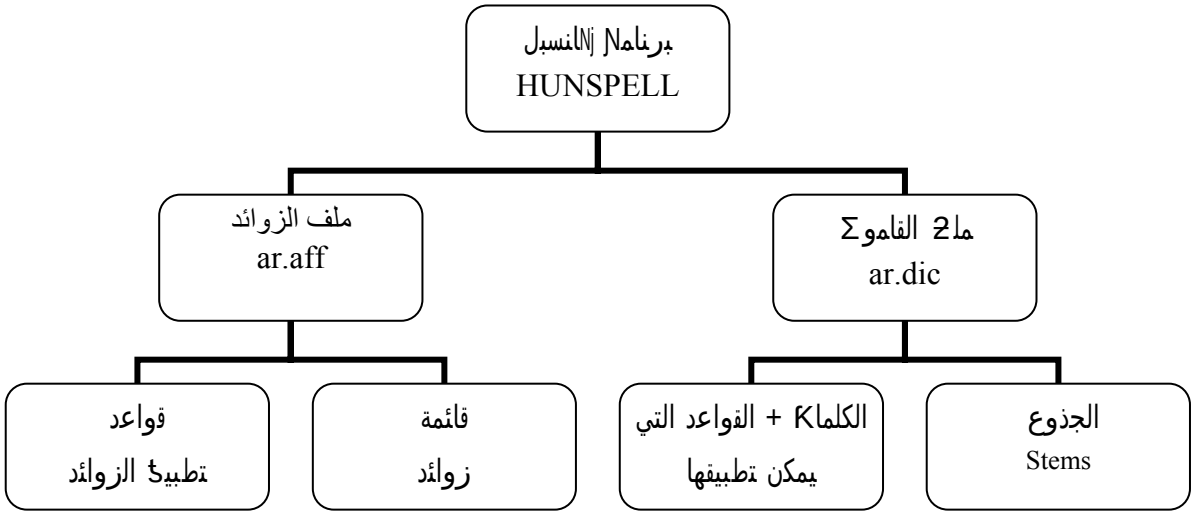
الصحيحة في قائمة الاقتراحات حسب نوع الخطأ. بينما التجارب في حالة الخطأ الإضافة ' insertion errors' التي يجد مدقق مسبل مكثر سهولة للتعامل معها إذ مسبل الكلمة الصحيحة أولى بنسبة 43.1 % من عدد المحاولات وفي حالة الخطأ الحذف 'deletion errors' فهو إذ الكلمة الصحيحة أولى بنسبة 33.1 % أما بالنسبة لأخطاء التعويض 'substitution errors' فيقوم عندما بتصحيح الكلمة الصحيحة بنسبة 30.9 % من عدد المحاولات [CARLSON A. and FETTE I., 2007].

II.4-2- طرق إصلاح نتائج المدققين الإملائيين المنسبل Hunspell والمسبل Aspell:

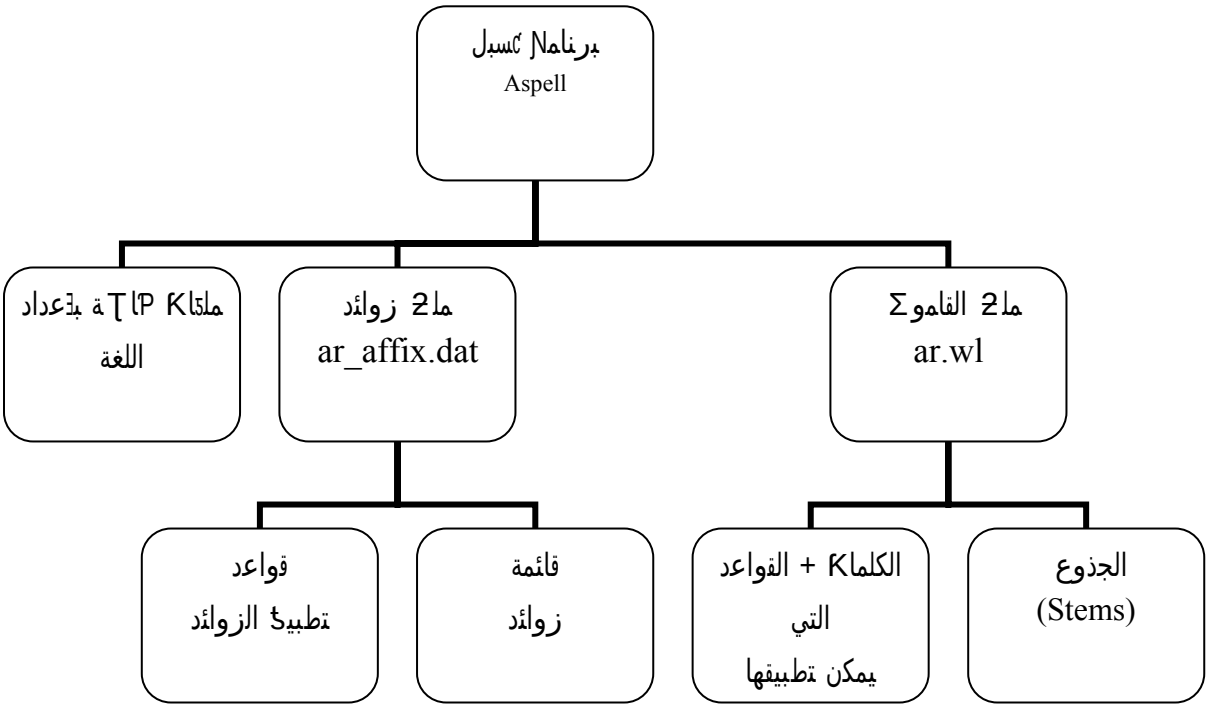
يتكون قاموس اللغة عادة من ملفين أساسيين أما ملف قائمة الكلمات وملف للزوائد بالنسبة لهذا مسبل تحتوي حزمة اللغة على ملفين أما قائمة الكلمات ar.dic وملف الزوائد ar.aff أما مسبل فيسمي ملف قائمة الكلمات ar.wl وملف الزوائد ar_affix.dat ويحتوي ملف على أرقام لأعداد اللغة.

وللبرنامجين نفس صيغة الملفين في الغالب لأنهما ورثا عن برنامج ispell ويختلفان فقط في بعض المزايا.

- **ملف الزوائد:** يحتوي ملف الزوائد على قائمة الزوائد الممكنة وقواعد تطبيقها ولكتابة قاعدة زيادة نفس الصيغة في البرنامجين الصيغة التي ورثا عن برنامج Ispell.
- **ملف الجذوع:** يحتوي ملف قائمة الكلمات على الجذوع والكلمات مرفقة بالقواعد التي يمكن تطبيقها على كل مفردة.



رسم توضيحي لمكونات الملفات المستعملة من طرف المنسبل أثناء التدقيق الإملائي



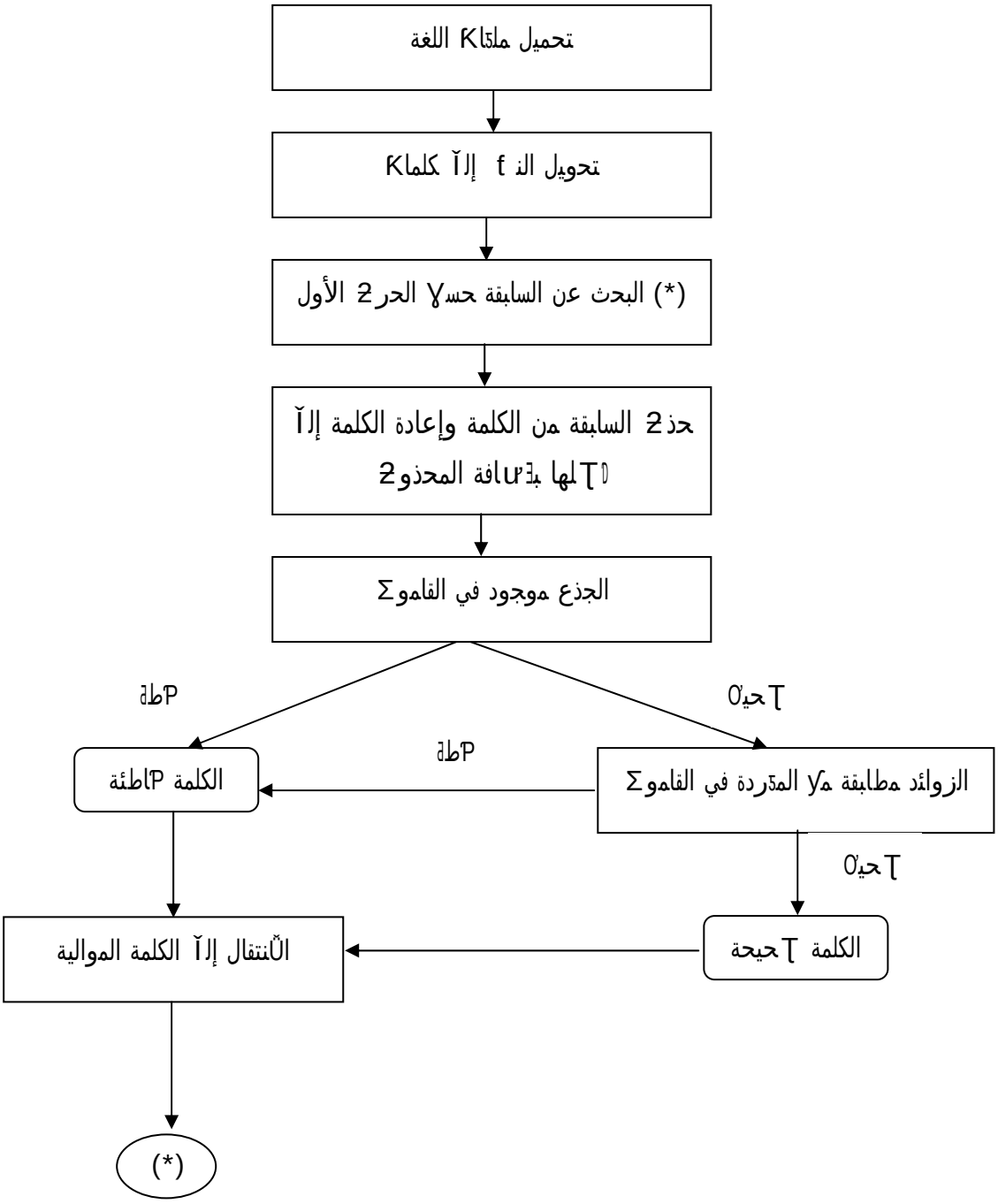
رسم توضيحي لمكونات الملفات المستعملة من طرف مسبل المنافي التدقيق الإملائي

II.4-2-1. طرق التدقيق الإملائي:

عند تشغيل البرنامج للقيام بعملية التدقيق الإملائي يقوم البرنامج بإتباع سلسلة العمليات التالية:

- تحميل ملفات اللغة.
- تحويل النص إلى كلمات.
  - من أجل كل كلمة يقوم البرنامج بما يلي:
    - البحث عن السابقة حسب الحرف الأول (يمكن أن تكون السابقة أو اللاحقة معدومة).
    - حذف السابقة من الكلمة وإعادة الكلمة إلى أصلها بإضافة المحذوف.
    - البحث عن الجذع في قائمة الكلمات.
    - في حالة وجود الجذع يجب التحقق من مطابقة الزوائد مع المدرجة في القاموس.

رابعا: خوارزمية التدقيق:



الدائرة تعني "الذهاب إلى" (GOTO)

رسم توضيحي لطريقة عمل مدققي مسبل والمنسبل

مثال:

ملف الزوائد	ملف القاموس	كلمات قائمة	كلمات غير معروفة
# لاحقة لزيادة التاء في الماضي SFX T Y 1 SFX T 0 ة . # سابقة لزيادة ياء المضارع PFX Y Y 1 PFX Y ا ي .	استعمل/TY	استعملة يستعمل	تستعمل استعملي

لنأخذ كلمة "يستعمل" ونقوم بتدقيقها بإتباع المراحل المذكورة سابقاً فنحصل على ما يلي:

- نبحث عن السوابق التي تبدأ بأول حرف من الكلمة حرف الياء.
- نجد السابقة Y لا نقوم بتطبيق القاعدة بطريقة عكسية.
- نحذف السابقة يستعمل -> ستعمل
- نعيد القسم المحذوف إلى الكلمة ستعمل -> استعمل.
- نبحث عن هذا المدرد في القاموس لا نجدها لا ثم نتأكد أن السابقة Y قابلة للتطبيق عليها.
- نستنتج أن الكلمة صحيحة.

II.4-2-2. طرق التدقيق الإملائي:

عندما يكتشف المدقق مفردة غير موجودة في القاموس (ملف dic وملف aff) يعتبرها أخطاء فيقترح مجموعة من البدائل الممكنة و ذلك بالعمل على تغيير حرف واحد فقط ليتوافق مع مفردة القاموس بمفردها أو بتركيبها مع زيادة وذلك إما بتبديل موضعي [قد f /قدر] أو استبدال حرف بحرف [قد f /قر f أو قدم] أو حذف حرف [قد f /ق f أو قد] أو إضافة حرف [قف/حقف] أو إدراج فراغ [البتطريق/البك طريق] .

II.5. تطبيق خوارزمية هاشينغ:

II.5-1- تطبيق على:

تعتبر الطبيعة الاشتقاقية واحدة من الخصائص البنيوية للغة العربية والمتعلقة بالتدقيق الإملائي حيث توجد مادة لغوية يمكن تشكيل صيغ منها عن طريق إضافة زوائد في أول المادة أو في آخرها. وباعتبار السوابق واللواحق المعمول بها في اللغة العربية يمكن للكاتب التعبير باستعمال جذر الكلمة نفسها عن التذكير والتأنيث الأفراد والتثنية والجمع الاستفهام والنداء ... إلخ. كلمة « سيعطيكموها » تمثل صيغة جيدة لهذه الميزة.

يمكننا إستعمال كلمة زوائد (Affixes) والزائدة هي إما سابقة أو لاحقة متصلة بجذور كلمات لإنشاء كلمات جديدة. يتم تطبيق هذه الزوائد على كل كلمة بها في معجم المدقق الإملائية يحتوي على قائمة الزوائد الممكنة وقواعد تطبيقها كما هو الشأن بالنسبة لملف 'ar.aff' في برنامج الهنسبل أو ملف ' ar_affix.dat ' في مدقق

منسبل. كما يفتقر لها مدلول الصلة في المعاجم الحديثة عدا المعاجم العربية لاحتواء لغتنا العربية على تصريف صرفية دلالية (تحدث دلالة الكلمة نفسها) تختلف عن السوابق واللواحق.

إن أنواع الصلة من الزوائد سنتطرق له بالتفصيل في الفقرات اللاحقة مثل الزوائد المتوسطة والتي تدرج داخل جذور الكلمات والزوائد السياقية والتي تتكون من جزئين متلازمين أحدهما في بداية جذر الكلمة والآخر في نهايته.[28]

II.5-2- الزوائد السابقة:

فيما يخص الأفعال فهناك زيادات تأتي قبل السوابق وهي الزيادة السابقة Proclitique مثل واو العطف الثلاثية المزة الاستفهام وتأتي تلي اللواحق هي الزيادة اللاحقة Enclitiques وهي متعلقة بالتعدي مثل 'ـكما' و'ـهما' [ABBES R., DICHY J. and HASSOUN M., 2004].

مثال:

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

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

II.5-3- الصيغة العامة للقواعد المتعلقة بالزوائد في ملف المنسبل الأشكال التالية:

الصيغة العامة للقواعد المتعلقة بالزوائد في ملف المنسبل على أحد الأشكال التالية:

PFX	علم	منتج مركب	عدد.
PFX	علم	شرط تجريد سابقة	وقف صرفي.
SFX	علم	منتج مركب	عدد.
SFX	علم	شرط تجريد لاحقة	وقف صرفي.

الزائدة هي عبارة عن سابقة أو لاحقة متصلة بجذور كلمات لتكوين كلمات أخرى[29].

تشمل القواعد المتعلقة بالزوائد في برنامج المنسبل ما يلي:

- أقسام الزوائد (Affix classes): وتعرف بعدد من قواعد الزيادة ويشار إليها بأعلام الزوائد.
- العنوان: ويمثل السطر الأول في تعريف فئة الزوائد.

II.5-3-1. تعريف فئة الزوائد:

(0): اسم الفئة (PFX or SFX)

(1) : علم (اسم فئة الزوائد)

(2) : منتظم مركب (السماح بجمع السوابق واللواحق). القيم الممكنة : Y (نعم) أو N (لا).

(3) عدد أسطر القواعد التالية.

II.5-3-2. قواعد الزوائد:

(0) اسم الآلية.

(1) علم.

(2) تجريد الحروف من البداية (في قواعد السوابق) أو من النهاية (في قواعد اللواحق) للكلمة.

(3) زائدة (اختياري أو إعلام فئة الاستمرارية يتم إلحاقها عن بعدها البعد الرمز '/').

(4) شرط.

الشرط الافتراضي (Zero Condition) يشار إليه بالنقطة. الشرط هو عبارة منتظمة (Regular Expression) مبسطة (شبه نموذج) لا يجب تحققه قبل تطبيق الزائدة. (النقطة تمثل حرفا اعتباطيا.الأحرف في الحاسوب تمثل حروفا اعتباطية من مجموعة حروف الاستبدال. الرمز (-) ليس له معنى أما ^ ولكن العلامة (^) بعد أول حاسوبة تحدد مجموعة الحروف المكملة أو المتممة).

(5) ميادين اختيارية يتم إلحاقها فراغا أو جدولا.

II.6. قواعد الإلحاق (flags) !عكس قواعد الإلحاق :

لقد طور برنامج المنسبل في الأساس من اجل أن يكون مدققا إملائيا يحترم قواعد الإملاء والآلية اللغة المجرية وكل الأعلام التي استخدمها عند أهور برنامج المنسبيل كانت لا في هذا الانتجاز سنذكر بعضا منها لاحقا. ولكن بعد دعم برنامج المنسبل للغات غير المجرية كالألمانية والتركية اضطر العاملون على تطوير برنامج المنسبيل إلى إضافة أعلام أخرى لتلبية حاجة مستخدمي البرامج الناطقين بتلك اللغات واحترام قواعد لغتهم.

يمكننا أن نلاحظ أن الجهود المبذولة لإضافة أعلام إلى برنامج المنسبل والآلية بمختلف اللغات في مجموعة من الأعلام يمكن تلخيصها فيما يلي :

- **complexPrefixes**: علم يخص السوابق المركبة وهو يمثل سابقة ثنائية النوع (لكن اللاحقة أحادية النوع). هذا العلم مهم أساسا من أجل طلب أما باللغة القبطية وإن كان يبدو أنه مفيد للغة العربية فهو ليس كذلك لأنه يسمح فقط بتركيب سابقين ولاحقة واحدة بينما نرى في العربية أن اللواحق أكثر تعقيدا من السوابق. مثلا : س+ي+عط+ون+كم+هم.
- **Lang langcode**: علم يحدد رموز اللغة إذ يتيح المنسبل إستعمال حروف لغة معينة تحددها الآلية 'lang code'. نذكر على سبيل المثال: TR_tr للغة التركية أو hu_HU للغة المجرية.
- **Ignore characters**: يستعمل هذا العلم لتجاهل حروف من قاموس الكلمات واللواحق والكلمات المدمجة. وهو جد مفيد لمعالجة مشكل الحروف الاختيارية مثل علامات التشكيل في اللغة

العربية (الحركات). هذه الآلية تستعمل فعلا إلا البرنامج انطلاقا من الأدار 1.1.5 بطلب طه زروقي وبرمجته.

- **Break:** يستعمل لتحديد نقاط الفصل بالنسبة للكلمات المفصلة وتتحد أجزاء الكلمة مفصلة. وهو علم لا نحتاجه عند استعمال اللغة العربية، إلا أنه جد مفيد عند التعامل مع عدة لغات أجنبية تستخدم الكلمات المركبة مثل الإنجليزية والألمانية والمجرية.
- **Compound Rule:** يعتبر أسلوب لمعالجة التراكيب التي تجمع حروفا أو سلسلة من الحروف. ونستعمله بدلا منه 'break' عندما لا تكون هناك إمكانية لإجراء وصل دقيق لقواعد التركيب.
- **OnlyInCompound**: اللواحق المشار إليها بالعلم ONLYINCOMPOUND تكون فقط داخل التراكيب (Fuge-elements باللغة الألمانية، fogemorphemes باللغة السويدية).
- **CompoundRoot :** العلم COMPOUNDROOT يشير إلا التراكيب في القاموس (يستخدم الآن أساسا للغة المجرية).
- **CHECKCOMPOUNDREP :** يمنع التراكيب، إذا كانت أخطاء فالكلمة المركبة قد لا تكون مركبة بشكل جيد بسبب الأخطاء الناتجة عن آلية الاستبدال (REP).
- **COMPOUNDSYLLABLE :** يستعمل في حالة آلية من قواعد التركيب في اللغة المجرية. العامل الأول هو القيمة القصوى لعدد المقاطع، التي قد تكون في التركيب، إذا كانت الكلمات في التراكيب أكثر من COMPOUNDWORDMAX. العامل الثاني هي قائمة حروف العلة (لحساب المقاطع).
- **SYLLABLENUM :** يستعمل في حالة آلية من قواعد التركيب في اللغة المجرية.
- **WORDCHARS :** WORDCHARS يطيل محدد الكلمات في سطر الأوامر لهانسبل بتحرف كلمات إضافية. على سبيل المثال ، النقطة ، الواثبة ، ن-واثبة ، الأرقام ، رمز النسبة المئوية هي أحرف كلمة في اللغة المجرية.
- **CHECKSHARPS** : زوج الحروف 'SS' في حالة أحرف كبيرة (في اللغة الألمانية) للكلمات قد تكون حالة حرف كبير الشكل ل 'S' حاد(β). المنسبل يمكنه معالجة هذه الحالة الآلية مع علم CHECKSHARPS.

II-6.1- معالجة 'CIRCUMFIX' وأخواتها :

عملية إلحاق الزوائد المرفقة بالعلم 'CIRCUMFIX' يمكن تسميتها بالزيادة الدائرية، فالزوائد المماثلة بالعلم CIRCUMFIX قد تكون على كلمة عندما تكون هذه الكلمة لها أيضا سابقة مع العلم CIRCUMFIX والعكس بالعكس.

أ- الإلحاق:

- إجبار ربط مجموعة من السوابق بمجموعة أخرى من اللواحق.

مثال : في اللغة المجرية، عملية إضافة السابقة 'leg ' متزامنة مع إضافة اللاحقة 'bb' إلا قاعدة الفصل.

- الإشارة إلى الأشكال الممنوعة للكلمة المزيدة.
- استخراج مجموعة فرعية من مجموعة الكلمات المقبولة المزيدة والمركبة.

ب- الإشكالية:

- تعتبر تلقائية معالجة حالات الزيادة الدائرية من أهم الآثار الجانبية لتعدد خطوات تجريد الزوائد.
- المشكلة في بنية المستوى الواحدي تكمن في انه ليس المطلوب طريقة لوصف شجرة f معجمي f بالسوابق واللواحق المترابطة وبالتالي الأشكال غير الصحيحة تعرف على أنها صحيحة بمعنى '*legv?fR = leg + v?fR ' 'قديم'.
- الإشارة الترابط سابقة -- لاحقه موجودة في كل مكان في فئة أنماط التغير الاشتقاقي (راجع الإنكليزية payable إلى non-payable لكن *non-pay أو drinkable إلى undrinkable لكن *undrink). و بعبارة بسيطة المناط السابقة 'un' مشروعة فقط في حالة ما إذا كانت القاعدة 'drink ' متبوعة باللاحقة 'able'.
- الزوائد المشروطة الموضوعة عن طريق استمرار الصيغ ليست كافية للزيادة الدائرية لأن الزيادة الدائرية هي صرف زائدة واحدة. نحن بحاجة إلى لمزايا الزيادة الدائرية من أجل تحليل صرفي صحيح.

الملف 1 :

```
---------- ملف الزوائد ar.aff ----------
CIRCUMFFIX X
PFX   Y Y 2
PFX   Y 0   ي/X

PFX   W  Y 1
PFX   W 0 ك/x
SFX   N  Y  2
SFX   N 0   و ن/XY
SFX   N 0 ن/XW
---------- ملف القاموس ar.dic ----------
فعل / N
---------- كلمات صائبة txt.good ----------
يتعلون
تتعلون
تتعلن
-------- كلمة خاطئة txt.wrong --------
يتعل
```

فعلون

فعلن

تفعل

يفعلن

مثال 2:

```
# circumfixes: ~ obligate prefix/suffix combinations
# superlative in Hungarian: leg- (prefix) AND -bb (suffix)
# nagy, nagyobb, legnagyobb, legeslegnagyobb
# (great, greater, greatest, most greatest)
CIRCUMFIX X

PFX A Y 1
PFX A 0 leg/X.

PFX B Y 1
PFX B 0 legesleg/X.

SFX C Y 3
SFX C 0 obb . +COMPARATIVE
SFX C 0 obb/AX . +SUPERLATIVE
SFX C 0 obb/BX . +SUPERSUPERLATIVE
```

الملف .2 :

```
1
nagy/C
[MN]
```

الإخراج :

```
> nagy
nagy[MN]
> nagyobb
nagy[MN]+COMPARATIVE
> legnagyobb
nagy[MN]+SUPERLATIVE
> legeslegnagyobb
nagy[MN]+SUPERSUPERLATIVE
```

II.6-2- استخدام 'NEEDAFFIX' لإنشاء كلمات مركبة:

إذا العلم يشير إلى الجذور الافتراضية في القاموس. فقط الأشكال المرفقة بزوائد لهذا الكلمات ستكون مقبولة من طرف المنسبل. إلا إذا كانت كلمة القاموس لها لاحقة متجانسة أو الزائدة تمر 'zero affix'. NEEDAFFIX تعمل على السوابق وعلى سابقة + ترتيبات اللاحقة. تستعمل هذه الآلية عند الحاجة إلى زائدة إضافية لتشكيل كلمة صحيحة.

مثال :

```
---------- ملف القاموس *.aff ----------
        COMPOUNDRULE sde
        NEEDAFFIX x
        COMPOUNDWORDMAX 3
```

```
COMPOUNDMIN 1
---------- ملف القاموس *.dic ----------
 start/s
end/e
-/xd
```

هذا البرنامج يسمح بتشكيل الكلمة 'start-end' بينما يعتبر الكلمة 'startend' كلمة خاطئة.

-ج- الأوامر الخاصة بـ 'NEEDAFFIX' وشرحها:

- **الأوامر المتعلقة بالإدراج:**

لقد سامح KaNjiza الخاصية في حل مشكلة الأدغام حين يتم استبدال المفردة المدغمة مثل (شدّ) بهيئتها المساعدة (شدّد) على مستوى القاموس إذ أن المنا بعض المفردات التي قد تكون خاطئة في حالتها المساعدة.

لذلك تم تفعيل الخاصية (NEEDAFFIX X0) على مستوى القاموس مثلا نحو (مسس .X0/).

- **الأوامر الخاصة بالزوائد الحرف الأخير شرط وجود علامة الشدة:**

لحل مشكل الأدغام تم تمييز الأفعال المساعدة في ملف القاموس بالشدة في مكرر (ملحق) وعلى مستوى ملف الزوائد ربط الشدة بتعريف الحرف الأخير على النحو التالي (تعريف الحرف الأخير شرط وجود علامة الشدة في مكرر المفردة).

II.7. الأوامر الخاصة بالتدقيق والتصويب:

يستعمل برنامج المنسبل بعض الخاصية بمعالجة الكلمات المركبة إذ يتطلب هذا النوع من الكلمات إلى تدقيق خاص بها إذ تتكون هذه الكلمات من سلاسل من الحروف تفصلها علامة الواصلة '-' (dash) يضطر المدقق إلى تقسيمها إلى عدة أقسام دون الاعتماد على رمز الواصلة كنقطة فاصلة لأن هذا الحل قد لا يكون مجديا في الكثير من الأحيان لاحتواء الكلمات المركبة على أجزاء كلمات غير صحيحة بذاتها بل باستعمال خاص تمكنه من تحديد نقاط الفصل مثل الخاصية (Break) ومن ثم تدقيق الأقسام المحصل عليها.

من هذه الخاص التي تستعمل لتدقيق الكلمات المركبة نذكر ما يلي:

- **COMPOUNDMIN:**

تحدد القيمة الدنيا لطول الكلمات في الكلمات المركبة. القيمة الافتراضية هي 3 حروف.

- **COMPOUNDFLAG:**

الكلمات المشار إليها بـ COMPOUNDFLAG قد تكون في الكلمات المركبة (إلا إذا كانت كلمة اقصر من COMPOUNDMIN). الزوائد ذات COMPOUNDFLAG تسمح لها بتركيب الكلمات المرفقة بزوائد.

- **COMPOUNDBEGIN:**

الكلمات المشار إليها بـ COMPOUNDBEGIN (أو ذات زائدة مسجلة) قد تكون العناصر الأولى في الكلمات المركبة.

- **COMPOUNDLAST**: الكلمة المشار إليها بـ COMPOUNDLAST (أو هي زائدة مسجلة) قد تكون العناصر الأخيرة في الكلمة المركبة.
- **COMPOUNDMIDDLE**:
الكلمة المشار إليها بـ COMPOUNDMIDDLE (أو هي زائدة مسجلة) قد تكون في منتصف عناصر الكلمة المركبة.
- **ONLYINCOMPOUND**:
اللواحق المشار إليها بالعلم ONLYINCOMPOUND قد تكون فقط داخل التراكيب (Fuge-elements باللغة الألمانية، fogemorphemes باللغة السويدية).
- **COMPOUNDPERMITFLAG**:
السوابق يسمح بها في بداية المركبة، اللواحق يسمح بها افتراضيا في نهاية المركبة. اللواحق التي تحتوي COMPOUNDPERMITFLAG قد تكون داخل المركبة.
- **COMPOUNDFORBIDFLAG**:
اللواحق هي إذا العلم لا تسمح بتركيب كلمة مرفقة بزوائد.
- **COMPOUNDROOT**:
العلم COMPOUNDROOT يشير إلى المركبة في القاموس (يستخدم الآن فقط في القانون الخاص باللغة المجرية).
- **COMPOUNDWORDMAX**:
تحدد القيمة القصوى لعدد الكلمة في كلمة مركبة. (الأفتراضي غير محدود).
- **CHECKCOMPOUNDDUP**:
يمنع تكرار الكلمة في التراكيب (مثل foofoo).
- **CHECKCOMPOUNDREP**:
يمنع التركيب إذا كانت (سيئة غالبا) الكلمة المركبة قد تكون كلمة غير مركبة من خطة الاستبدال (REP). مفيد بالنسبة للغة من 'تراكيب دقيقة' الأملاء.
- **CHECKCOMPOUNDCASE**:
يمنع اعالي حالة الحروف عند تطويل الكلمة في التراكيب.
- **CHECKCOMPOUNDTRIPLE**:
تمنع التركيب إذا كانت الكلمة المركبة تحتوي على كلمة ثلاثية الحروف (مثال foo|ox أو xo|oof).

الاشارات:

فقدان الدعم للحروف متعددة البايت في نظام التشفير UTF-8 (يعمل فقط حروف نظام ASCII 7 بت).

```
CHECKCOMPOUNDPATTERN number_of_checkcompoundpattern_definitions

 CHECKCOMPOUNDPATTERN endchars beginchars
```

يمنع التركيب في حالة ما إذا كانت أول كلمة في التركيب تنتهي بأحد حروف النهاية 'endchars' وكانت الكلمة المقبلة تبدأ بأحد حروف البداية 'beginchars'.

- **COMPOUNDSYLLABLE max_syllable vowels:**

تستعمل في حالة من قواعد التركيب في اللغة المجرية. العامل الأول هو القيمة القصوى لعدد المقاطع التي قد تكون في التركيب إذا كانت الكلمات في التراكيب أكثر من COMPOUNDWORDMAX. العامل الثاني هي قائمة حروف العلة (لحساب المقاطع).

- **SYLLABLENUM flags:**

تستعمل في حالة من قواعد التركيب في اللغة المجرية.

II.8. تحديد برنامج المدقق:

من بين الأمور التي يمكن أن نعيبها على المنسبل يمكننا أن نذكر ما يلي :

- لم يطور في الأساس من أجل دعم اللغة العربية و إنما من أجل بعض اللغات الأوروبية وبالتالي اللغة المجرية والتي تمثل لغة مطور هذا البرنامج.
- عدم جاهزية قاموس عربي بحت متاح ومُعد بطريقة منهجية يمكن أن يعتمد عليه مدقق المنسبل (العمل الذي يعكف عليه طاقم مصبل).

فمن بين أهم المشاكل التي واجهها محمد كبداني منذ شروعه في هذا المشروع نذكر :

- اضطر إلى إضافة الأشكال الخاصة بالإعلال والإبدال والتصريف والإدغام وما يجري على الهمزة من تحول رافعا بذلك قائمة جذور الأفعال العربية من 10000 إلى 15000 جذر. بسبب غياب خاصية تعالج هذا النوع من التغيرات في برنامج المنسبيل الأصلي. الشيء الذي أدى إلى رفع حجم القاموس.
- مسألة التشكيل (الحركات): لا بد من تجاهل التشكيل أحيانا وذلك لتشابه بعض المفردات فيما بينها بالرغم من اختلاف معانيها. الشيء الذي قد يؤدي إلى اقتراح حلول خاطئة عند التحيين.
- عدم التعامل مع ما يسمى بالزيادات المتوسطة INFIXES التي تمس جذر الكلمة. في العربية مثلا يمكننا التعبير عن الإرتي الإعلال (إبدال حرف علة بآخر) والإشتقاق والذي يتم أحيانا بزيادة حروف في وسط الكلمة بزوائد متوسطة.
- صعوبة اكتشاف الخطأ: هذا المشكل مشترك بين كل اللغات لأن المدقق يتعامل مع المفردة في معزل عن جوارها و في غياب اعتبار المدقق للسياق الذي تأتي فيه المفردة تبقى فائدة المدقق الإملائي جد محدودة من الناحية الدلالية كما أن تعود الكاتب العربي على تجاهل التشكيل لا يزيد الطينة إلا بلة لأن النتيجة الحتمية هي تشابه المفردات فيما بينها بالرغم من اختلاف معانيها.

مثال: ألا كلا قد تعني كلمة (الفعل الماضي) وقد تعني (كلا) جمع كتاب)[30].

- صعوبة اقتراح البديل التحيين: من عوائق التخطئة البنيوية للغة العربية في مرحلة التحيين الأساسية التي يقترح المدقق من خلالها المفردات البديلة بإضافة حرف أو حذفه أو باستبداله بآخر أو بتقسيم الكلمة الخاطئة إلى كلمتين ارتفاع عدد المفردات المرشحة

للتصحيح وبالنسبة المقارنة باللغات الأخرى هي العدد المتوسط للكلمات المقترحة هو 3 في اللغة الإنجليزية و 3,5 في الفرنسية و26,5 بالنسبة لمفردات اللغة العربية غير المشكلة[31].

- غياب دعم الحروف المركبة مثل : ئئ لآ ئئ لأ .....الخ.
- لإحداث معلام مستخدمي الهنسبيل البطء على برنامج هذا المدقق و الخطأ عند اقتراح الحلول البديلة وذلك لتعقيد ملف الزيادات هنارك لتعقيد تصريف الأفعال العربية وكثرة النماذج التصريفية.

- **الخوارزمية:**

تتاح لنا هذه الدراسة التعرف على نظام الترميز العالمي الموحد 'Unicode' والفائدة الكبيرة التي جنتها برامج التدقيق الإملائي الداعمة للغة العربية كغيرها من البرامج الداعمة للعربية إثر ظهور هذا النظام وشيوع استخدامه في العديد من البرمجيات.

كما مكنتنا دراسة كل من برنامجي الهنسبل وأسبل المفتوحي المصدر من التعرف على مزايا هذين المدققين المنحدرين من نفس المدقق (Ispell) الذي ورثا عنه العديد من المزايا والتعرف على طريقة عملهما وكيفية استعمالهما لملفي القاموس والزوائد لحدهما يخص قائمة الكلمات (القاموس) والآخر يخص قائمة الزوائد وطريقة تطبيقها.

أما المرحلة الأخيرة من هذه الدراسة والخطة بنقد برنامج الهنسبل فقد ساهم في تحديد الخطوط العريضة لدراسة الخطة الرامية إلى دراسة مشروع إعداد مدقق إملائي عربي مفتوح المصدر وذلك بالوقوف على أهم حالات التدقيق الإملائي العربي والتي لازال برنامج الهنسبل الأصلي يجد صعوبة في التعامل معها والتي سيتم تناولها في الفصل القادم من خلال دراسة مشروع فريق أيسبل 'Ayaspell' الذي يهدف إلى إعداد قاموس المدقق الإملائي العربي والتحسينات التي قام بإنجازها على ملف زوائد برنامج الهنسبل وملف القاموس الخاص بمفردات اللغة العربية الذي تم إعادة إنجازه بالكامل.

الفصل الثالث

التحسينات المقترحة لتطوير برنامجي الانسبل وتسبل

III. الفصل الثالث: التحسينات المقترحة لتطوير برنامج المنسبل وعسبل

- **مقدمة:**

إن النقائص المكتشفة في كل من برنامجي المنسبل وعسبل والتي يمكن تلخيصها في كل من مسألة إشكال الشكل والتطويل الناتج التدقيق وغياب صيغ ونماذج واضحة في مالخ الزوائد للتعامل مع حالات الإدغام والإعلال والإبدال وعدم التعامل مع الزوائد المزدوجة بالإضافة إلى مشكلة افتقاد كلا البرنامجين إلى قائمة كلمات (قاموس مرجعي) معدة بطريقة منهجية واضحة ومستنبطة من أهم وأشهر معاجم اللغة العربية هذا الطريق للحديث عن المزايا الجديدة والتي سأهم المصدر المفتوح لهذين البرنامجين في تسهيل عملية تطويرهما التي هي وبشكل جد فعال إلى حل العديد من المشاكل التي واجهها هذا البرنامج عند تدقيقه للنص و في العربية.

سنتناول هذا الفصل بتقديم شامل لمشروع فريق عيسبل لتطوير المدقق الإملائي العربي المفتوح المصدر كونه الفريق المشرف على التحسينات التي أجريت على كل من برنامجي المنسبل وعسبل لتطوير مزاياهما للتلاءم مع صيغ اللغة العربية إذ يقوم طه زروقي بالتنسيق مع المطور الأصلي لبرنامج المنسبل بإعادة برمجة بعض الخصائص في مالخ الزوائد بينما يتكفل محمد كبداني[1]* بإنجاز القاموس المرجعي الخاص بالمدقق الإملائي العربي ننتقل بعدها إلى عرض شامل لأهم المسائل التي يطمح فريق عيسبل إلى إيجاد حلول لها من خلال تطوير أو إضافة بعض المزايا في كلا البرنامجين مثل مسألة التشكيل والإدغام والإعلال والإبدال وأخيرا مسألة التعامل مع الزوائد المزدوجة يليها عرض شامل ومفصل لمختلف المزايا المطورة في كلا البرنامجين من أجل اقتراح حلول لكل من المسائل السابقة لنعرج بعد ذلك على العمل الذي قام به محمد كبداني من اجل إنشاء القاموس الخاص بالمدقق الإملائي العربي المفتوح المصدر من خلال ذكر مختلف المعاجم التي اعتمد عليها لإنجاز هذا القاموس مثل المعجم الغني والمعجم الوسيط وعمله على هذه المعاجم لاستخلاص قائمة الكلمات وتهيئته بعد ذلك للأفعال والأسماء من أجل وضع منهجية لطريقة إعداد هذا القاموس لاستغلاله عند التدقيق الإملائي على أحسن وجه ننتقل بعدها إلى مثال تطبيقي يشمل مختلف المزايا المطورة ويوضح طريقة استعمالها والنتائج التي يمكن الحصول عليها لنختم هذا الفصل ببعض الأفكار والتصورات المستقبلية لمشروع فريق عيسبل.

III.1. أداة التصحيح الإملائي (Ayaspell):

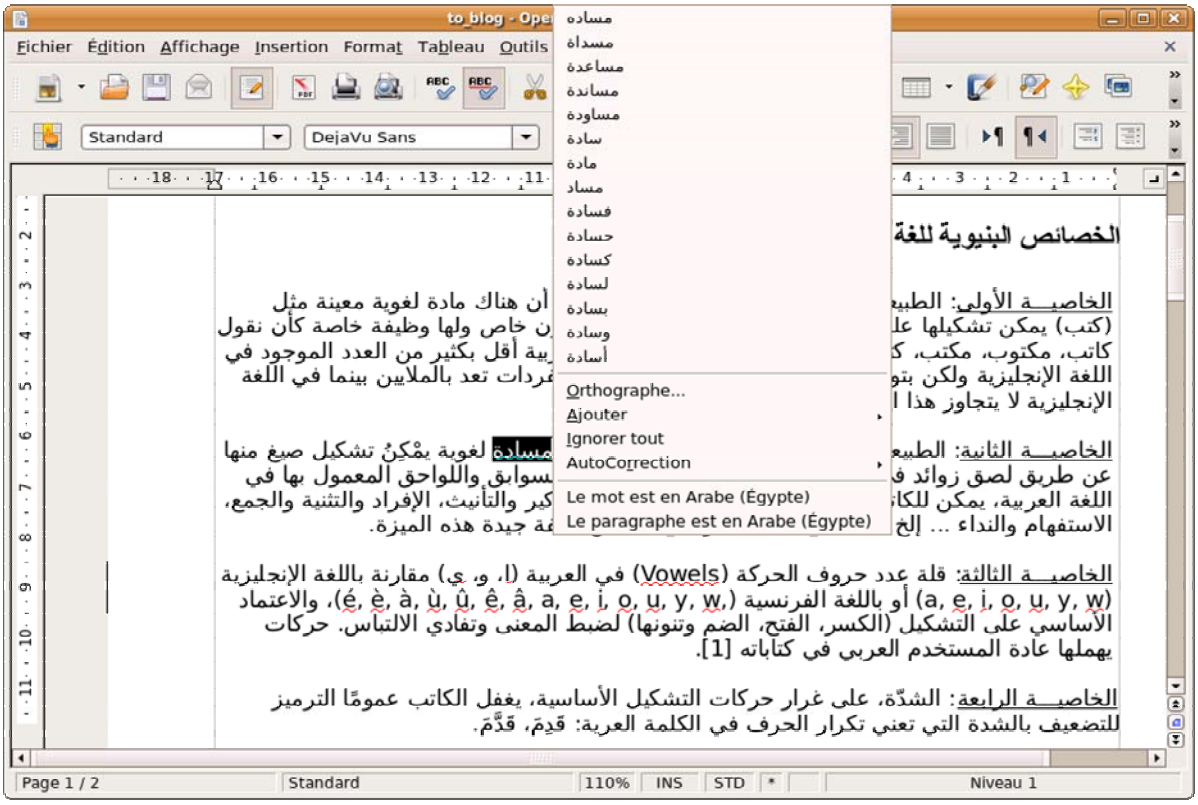
يهدف مشروع فريق عيسبل 'Ayaspell' إلى إنشاء قواميس في اللغة العربية للتطبيقات المكتبية معتمدا في الأساس على برنامج المنسبل المطور لدعم المزايا العربية والذي سنتطرق إلى التقنية في الفقرات القادمة و بدرجة أقل على برنامج عسبل الذي لا يزال بحاجة إلى تطوير أكثر للتلاءم مع المزايا العربية.

III.1-1- المكتبات التي اعتمد عليها فريق عيسبل لدعم اللغة العربية:

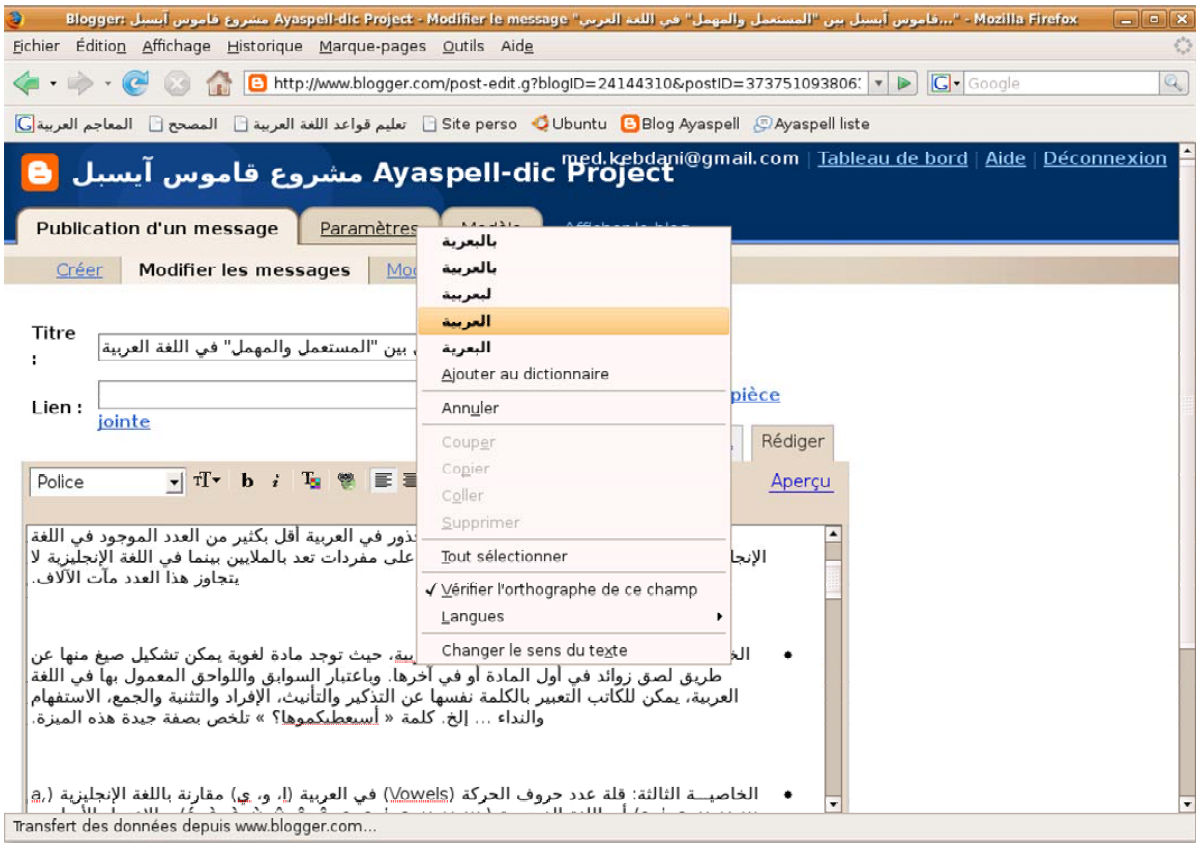
من أهم هذه البرامج المكتبية التي استفاد فريق عيسبل من دعم اللغة العربية فيها نذكر:

[1]* مهتم بالإعلامية الحرة

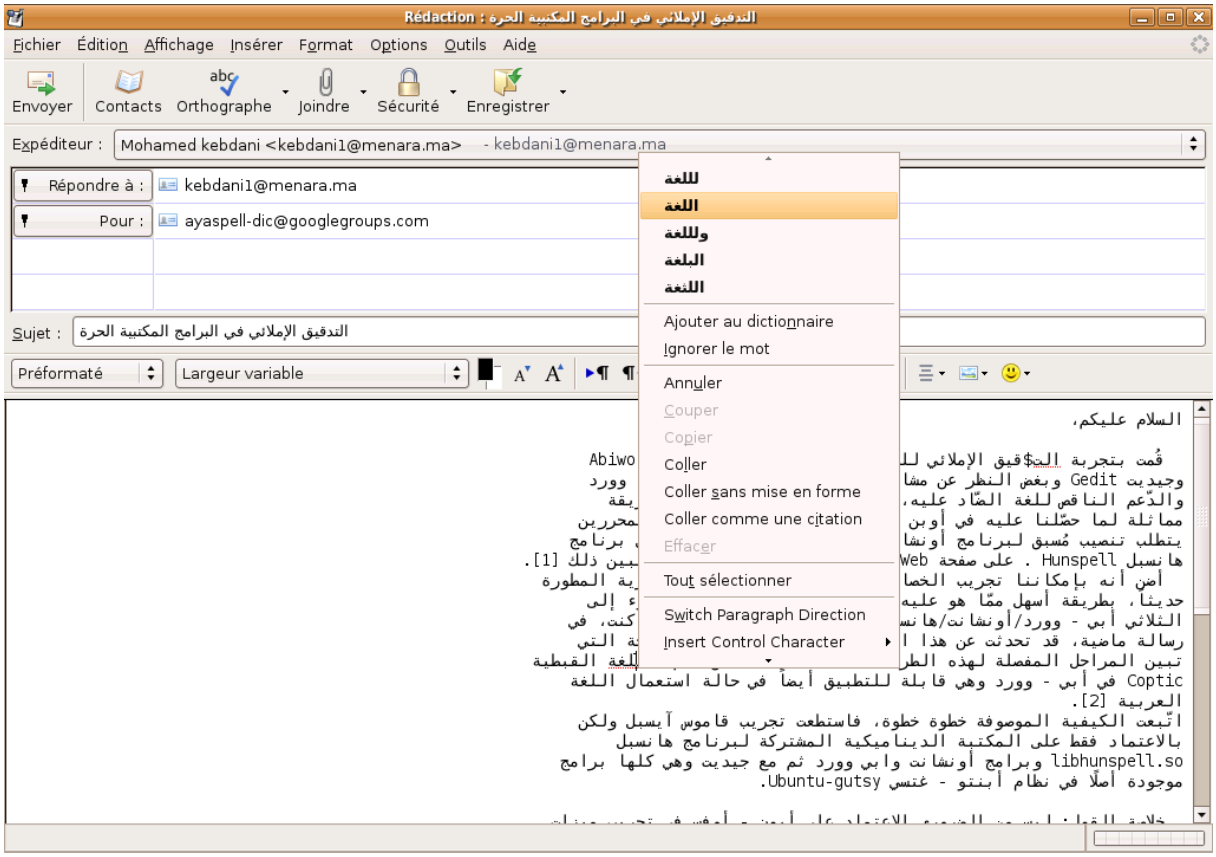
ج- الإضافة الخاصة بـ 'OpenOffice.org':



ب- الإضافة الخاصة بـ 'Fire Fox':



ب - إدراج التدقيق الإملائي في برنامج 'Thunderbird':



III.1-2- الخارطة الطريق لمشروع التحسين:

تسمى الخارطة التي أعدت بتاريخ 22 مايو 2007 بالاطلاع على التقدم الحاصل بمشروع يسبل من خلال التطورات التي رافقت مراحل إعداد وتطوير قاموس المدقق الإملائي العربي انطلاقا من برنامج المنسبل.

شملت الخارطة تقييم ما تم إنجازه في بعد الإدارات السابقة بالإضافة إلى التطورات التي رافقت إدارات القاموس العربي للمدقق الإملائي انطلاقا من الإدار بيطا1 إلى الإدار بيطا2 ثم الإدار المحسن يليه الإصدار المستقر الحالي بالأنظمة المحمولة ووتو إلى الإدار المندمج.

عرض الخارطة:

- **الإدارات السابقة (بيطا):**
  - **الإدار بيطا1:**
    - القاموس غير تام.
    - قواعد الإلحاق غير كاملة.
    - آلية مقترحة التحيين التدريجي REP الخاطئة.
    - غياب دعم الحروف المركبة لآن لأ.

- **الإصدار الثاني:**
  - إضافة المصادر الناقصة من المراجع الأساسية.
  - إضافة أسماء الفاعل الناقصة باشتقاقها من الأفعال.
  - إضافة أسماء المفعول الناقصة باشتقاقها من الأفعال.
  - إضافة أسماء الآلة المشبهة الناقصة باشتقاقها من الأفعال.
  - إضافة الأسماء الناقصة من المعجم الوسيط والمعجم المحيط.
  - تفعيل آلية التماثل الحرفي/الموضعي: MAP.
  - إعادة بناء آلية التبادل التضاعلي: REP.
  - تفعيل آلية تجاهل التشكيل: IGNORE .
- **الإصدار الثالث (الإضافي/الكامل):**
  - حذف سوابق ولواحق الأفعال نادرة الاستعمال.
  - حذف الزيادات السابقة المتضمنة لحروف أو تركيبات حرفية نادرة الاستعمال.
  - إعادة بناء اللواحق والزيادات اللاحقة لتصريف المتعلقة بإشكالية التعدي (حصر القاعدة بالأفعال مهموزة اللام فقط أو تحويل هيئاتها إلى ملف القاموس).
  - إضافة دعم الحروف المركبة لا لا لأ.
  - استكمال تصنيف عاقل/غير-عاقل نكرة/تعريف للمفردات الاسمية.
  - إضافة تصريف المفكد لتصريف المستعملة بكثرة.
  - إضافة التعدي إلى مفعولين لتصريف المستعملة بكثرة.
  - إضافة مفردات ناقصة.
  - حذف المفردات المكررة.
  - حذف المفردات قليلة الأهمية المشوشة على مقترحات التصحيح (ة/ه/ى/ء ؤ/ئ).
  - حذف الأفعال والأسماء النادرة.
- **الإصدار الخفيف**:

  تطبيق بالأنظمة المحمولة (Embedded version).
- **الإصدار الشامل:**
  - التشكيل الكامل للقاموس.
  - الفهرسة النحوية للمفردات.

III.1-3- تحليل المشاكل المتعلقة بإشكالية التعدي:

من أهم المسائل المتعلقة بملف الزوائد التي يعمل فريق ميسبل على دراستها ومحاولة اقتراح حلول لها من أجل تطوير طريقة عمل المدقق الإملائي العربي نذكر ما يلي:

- **المشكلة الأولى:**

ويتعلق الأمر بالكلمات المشكلة والتي يعتبرها كل من مدققي التنسيق والتسبيق الخاطئة رغم وجود جذورها في القاموس وذلك لكون الجذور المسجلة في القاموس غير مشكلة الأمر الذي استوجب تجاهل الحركات أثناء التدقيق حتى تسهل المقارنة مع الجذور في القاموس.

- **المشكلة الثانية:**

ويتعلق الأمر بالكلمات التي بها حرف ساكن علامته الشدة والتي استوجبت وضع قواعد واضحة تفصل بين حالتي فرض الإدغام من عدمه والتي يمكن عند تطبيقها من تحديد قائمة الكلمات الصحيحة إملائيا التي يمكن الحصول عليها انطلاقا من جذع الكلمة الأصلية عند وصله بمختلف السوابق واللواحق الممكنة مع فرض الإدغام أحيانا وتركه أحيانا أخرى.

- **الإعلال والإبدال:**

ويتعلق الأمر بالكلمات التي تطرأ عليها تغيرات تمس جذع الكلمة وذلك من خلال تبديل أحد الحروف (حرف علة في حالة الإعلال أو حرف صحيح في حالة الإبدال) أو حذفه أحيانا الأمر الذي استوجب إستعمال بعض القواعد الخاصة بالزيادات المتوسطة لضبط التغيرات التي قد تطرأ على هذا النوع من الكلمات عند تعرضها للإعلال أو الإبدال.

- **الروابط الداخلية:**

ويتعلق الأمر بالكلمات التي تترابط فيها السوابق باللواحق الأمر الذي استوجب وضع قواعد خاصة بهذا النوع من الكلمات لضبط هذا التلازم بين السوابق واللواحق.

III.2. التحسينات المقترحة لتطوير برنامج التنسيق والتسبيق:

III.2-1- الإضافات المستوجبة لمعالجة الشاذ:

يتجاهل معظم مستخدمي اللغة العربية الحركات عند كتابة النصوص مما أدى إلى تجاهل هذه الحركات أثناء التدقيق الإملائي بالنسبة لمعظم المدققات على الرغم من أن لهذه الحركات دورا لا يمكن تجاهله فبإمكانها أن تغير معنى الكلمات و تنيتها بشكل خاص للانتباه.

مثال: كَتَبَ ---------> فعل الماضي .

كُتُبٌ ---------> جمع كتاب.

الكلمتين السابقتين تعتبران خاطئين بالنسبة لكل من برنامجي التنسيق والتسبيق رغم وجود كلمة 'كتب' في القاموس.

اقترح فريق تيسبيق إضافة هذه الخاصية إلى برنامج المدقق حتى تكتمل وظائفه على أحسن وجه.

III.2-1-1. الإدغام:

تم اقتراح خاصية جديدة في خاص إعداد القاموس واللغة في كل مدقق والاعتماد على تغيير برمجة التدقيق إذ عند قراءة كلمة ما يقوم المدقق بحذف الحروف المحددة فتكون النتيجة كلمة غير مشكولة ثم يواصل عمله كالمعتاد.

III.2-1-2. الأمر IGNORE:

تضيف الآلية الجديدة **'IGNORE'** إلى البرنامج تأجيل الحركات والتطويل. وذلك بتفعيل الآلية في ملف الزوائد 'IGNORE' والحركات التي سيتم إجتثاثها أمامها. الآلية يمكن أن تستعمل من أجل لغات أخرى غير العربية. حتى الآن لم نعلم أن أحدا من استعملها لغير العربية.

مثال:

ملف الزوائد ar.aff	ملف القاموس ar.dic	كلمات نصية txt.good
SET UTF8 IGNORE ب	كتب	كـــــــتب كَتَبَ كـــــــاتب ـــــكتب

III.2-1-3. الأمر DIACRITICS:

تضيف الآلية الجديدة **'DIACRITICS'** إلى البرنامج تأجيل الحركات والتطويل. وذلك بتفعيل الآلية 'DIACRITICS' في ملف إعداد اللغة 'ar.dat' متبوعة بالحروف والحركات التي سيتم إجتثاثها. لم أستعمل الآلية 'Ignore' لأن أحد الآليات لها نفس الاسم أستعمل في حالة الحروف ذات العلامة 'accent'. قد تتبنى الآلية من أجل لغات غير العربية.

مثال:

ملف إعداد اللغة ar.dat	ملف القاموس ar.wl	كلمات نصية txt.good
SET UTF8 diacritics ب	كتب	كـــــــتب كَتَبَ كـــــــاتب ـــــكتب

III.2-2- الأمر GEMENATING وتأثيره على الإدخال:

يتم مواجهة الآلية المسألة عند تعامل المدقق مع بعض الكلمات مثل: لنا **ردد** التي **ردد** والتي كانت تؤدي في السابق إلى زيادة مدخل في القاموس مما يجعل الأمر للتعقيد والكبر حجما. الآمر الذي استوجب إضافة الآلية الأدغام **'GEMENATING'** من أجل وضع علم لها في ملف الزوائد دون اللجوء إلى إضافة مدخل جديد في القاموس.

- **GEMENATING**: وتعني بأن العلم الذي سيأتي بعدها سيستعمل لتركيب الأدغام في كل قاعدة يكتب فيها ثم يتم استعماله في قاعدة الزيادة لتركيب الأدغام.

الآلية: GEMENATING G

الأدغام يتم بحذف الحرف الأخير مهما كان ومن ثم إضافة الزيادة.

ملاحظة: ردد + K = رد [د] + K <== ردK.

---------- ملف الزوائد ar.aff ----------

```
SET UTF-8
LANG ar
NOHARAKAT
GEMENATING G
SFX  T Y 4
SFX T 0 K
SFX T 0 تما
SFX T 0 K/G
SFX T 0 I/G
```

---------- ملف الأفعال ar.dic ----------

رد / T

---------- ملف الكلمات الصحيحة txt.good ----------

ردد I

رددت

رد I

ردا

---------- ملف الكلمات الخاطئة txt.wrong ----------

رددا

ملاحظة:

المخرجات الخاصة تعمل على العربية واللاتينية.

III.2-3- الإعلال الآلي للأفعال الصحيحة:

تعتبر واحدة من مهام ومعالجة المسائل التي تواجه التدقيق الإملائي المبني على سوابق ولواحق وقاعدة في اللغة العربية لأنها تتعلق بدراسة التغيرات الداخلية التي تطرأ على جذع الكلمة.

ملاحظة:

الفعل المعتل الأجوف: قام -------> قمK (حذف الألف) ---------> يقوم (إعلال الألف إلى واو).

في السابق كان لا بد من إضافة مداخل الكلمة في القاموس حسب تغيراتها أي إضافة كل من:

- قام
- قم
- قوم

الجدول التالي يوضح شكل ملف القاموس وملف الزوائد وكل من لائحة الكلمات الخاطئة والصائبة في هذه الحالة:

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

:خوارزمية

نلاحظ في هذه الحالة بأن هذا الحل سوف يتسبب في زيادة حجم القاموس وكثرة القواعد وتعقيدها.

III.2-3-1. الخوارزمية المقترحة لتصحيح الإعتلالات:

اعتمد طه زروقي على الخاصية 'Swap' عند بداياته في معالجة مسألة الأعلال والأبدال لكنه قرر تجاوزها فيما بعد من أجل حل أكثر دقة وشمولية سنتطرق إليه بعد ذكر مميزات هذه الخاصية وطريقة استعمالها.

- **SWAP:** الخاصية والأعلام التي تستخدمها للتعبير عن الأبدال أي إبدال حرف بحرف أخر أو حذفه أو إدراجه في موضع معين.

  أولا يجب تعريف جدول الأبدال وعددها وتسمية كل قاعدة إبدال.

  - تعريف جدول الأبدال في ملف الزوائد:

    الصيغة: كلمة swap أمامها عدد الأبدال الآتية. مثال : SWAP 2

  - تعريف قاعدة الأبدال بعد تعريف جدول الأبدال:

    الصيغة: كلمة SWAP أمامها:

    اسم العلم: وتعني أن العلم يقرر استبدال الحرف بحرف أخر. مثلا W.

    الإبدال الأول: رقم يدل على أن الحرف في هذا الموضع يجب أن يكون مطابقا للحرف المبدل به وسيتم إبداله بالحرف المبدل. مثلا 2.

    الحرف الأول: هو الحرف الذي سيحذف. مثلا A

    الحرف الثاني: هو الحرف الذي سيدرج. مثلا B

    إذا كان الحرف المبدل آخرا 0 فهذا يعني بأن الحرف المبدل به يحذف.

:أمثلة

القاعدة الآتية تستبدل بالألف الحرف واو في الموضع الثاني من بداية الكلمة: ا و SWAP W 2

موضع الأبدال: يتحدد موضع الأبدال بالنسبة إلى بداية الكلمة مثل : 2 الذي يمثل موضع الألف في كلمة 'قام'أو من نهاية الكلمة بكتابة رقم سالب: مثلا -2 هو موضع الألف في استقال أي الحرف الثاني من الأخير.

ملاحظة: علينا أن نعلم بأن تحديد موضع الحرف يتم بعد حذف الجزء المحذوف:

أي عند التحول من استقال إلاّ يستقيلان فموضع الألف الثانية لنا بعد حذف همزة الوصل هو 4 بدلاً من 5 وهذا ينطبق في الانتجازات أي على السوابق واللواحق.

ملاحظة:

---------- ملف الزوائد ar.aff ----------

```
SET UTF-8
LANG ar
NOHARAKAT
GEMENATING G
```

جدول الإبدال

(a) # ابدال W 2 ا و

(b) # ابدال Z 2 ا 0

```
SWAP 2
SWAP W 2 ا و
SWAP W 2 ا 0
```

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 ي/W
```

---------- ملف القاموس ar.dic ----------

المدخل واحد فقط

قام /T Y

---------- كلمات ثابتة txt.good ----------

قام

قم

يقوم

---------- كلمات خاطئة txt.wrong ----------

يقام

الشرح:

لدينا علمان:

العلم W: يقوم بحذف الألف في الرتبة الثانية وتعويضها بواو فتصبح: ي+قام ---> يقوم.

العلم Z: يقوم بحذف الألف في الرتبة الثانية فتصبح: قام + ت ---> قمت.

الإشكالات:

- يمكن إستعمال الخاصية لتصريف الفعل الماضي المبني للمجهول. مثال : عاقب ---> عوقب.
- يمكن إجراء إبدالات متعددة.

مثال : نيم -----> لوئم. وذلك بتطبيق ابدالين على نفس القاعدة :

الابدال الأول: w 2 وا

الابدال الثاني: Z 3 ى

- لا بد من الاشارة إلى الفرق بين SWAP وREP الموجودة سلفا في البرنامج والتي تستبدل بعض الحروف من اجل توليد الاقتراحات بينما تجري SWAP تغييرات على جذع الكلمة.
- قرر فريق تيسبل الاستغناء عن خاصية الخاص في الاصدارات الحديثة لبرنامج المنسبل والعمل على دعم المزايا الخاصة بالزوائد المتوسطة 'Infixes'.

III.2-3-2. الخطوات المنهجية لتطوير الإعتبارات الإضافية:

إن تعقيد هذا النوع من التغيرات وتعدد حالاته وأنواعه واختلافه عن تغيرات كونه يمس جذع الكلمة جعل فريق تيسبل يفكر في اقتراح حل يضمن شمولية وفعالية من سابقه فتم إضافة بعض القواعد الخاصة بالزوائد المتوسطة (Infixes) في ملفي الزوائد الخاصين ببرنامجي المنسبل وتسبل لضبط هذه التغيرات بمختلف أنواعها وفق قواعد التدقيق الإملائي العربي.

III.2-3-2-1. الخطوات الأساسية لدعم الزوائد المتوسطة:

قام طه زروقي بوضع تصميم جديد خاص بالزوائد المتوسطة يعتمد على ما يلي:

- يعبر في السابقة على أي حرف باستعمال النقطة فالسابقة ي.و تعني أن نضيف الياء في أول الكلمة أم نضيف الواو بعد أول حرف فنحصل مثلا على يقوم من قم.
- اللاحقة ي. تعني أن نضيف الياء إلى ما قبل الحرف الأخير فنحصل على يستقيل من يستقل.
- يمكن استعمال الحذف بنفس الطريقة فحذف .ا يعني حذف حرف الألف الثاني في الكلمة فنحصل على قل من قال. وفي اللاحقة يمكن حذف ا. يعني حذف الألف ما قبل الأخير من الكلمة فنحصل على استقل من استقال.

الجدول:

كلمات خاطئة txt.wrong	كلمات صائبة txt.good	ملف القاموس ar.dic	ملف الزوائد ar.aff
يقام	قام قم يقوم	#مدخل واحد فقط قام/TY	SET UTF8 # لزيادة تاء الماضي SFX T Y 1 SFX T 0 ت لاحقة T ا. ت . # لزيادة ياء المضارع PFX Y Y 1 سابقة Y . ا ي.و

ملاحظة:

يمكن استغلال الخاصية من أجل لغات خاصة فيمكننا مثلا توليد الكلمات swam, swum من swim بالانجليزية.

III.2-4- الاقتراحات المتعلقة بالمدقق الإملائي:

إن تركيب السوابق واللواحق لكلمة معينة (جذر في القاموس) من أجل الحصول على كلمات صحيحة في الغالب لقيود تنظم هذا التركيب وتحترم صيغا اللغة المراد تدقيقها.

III.2-4-1. الأخطاء الإملائية:

تبين من خلال الدراسات التي تناولت اللغة العربية أن الكلمة العربية تتشكل مما يلي : [ABBES R., DICHY J. and HASSOUN M., 2004]

- **الملواصق (Proclitics):** تتضمن أدوات الربط الأحادية مثل 'و' أو الأدوات التي تفيد التعليل مثل 'ل' أو حروف الجر مثل 'ب'.
- **السوابق (Prefix):** هذه الفئة تحتوي فقط على سوابق المضارع مثل السابقة 'ي' التي تمثل سابقة الكلمة ضمير الغائب.
- **الجذع (Stem):** الذي يمكن أن يكون ممثلا في كلمات الجذر'Root' (سلسلة مرتبة من ثلاثة أو أربعة حروف) والنموذج (قالب من المقاطع مكون من الحروف التي تشكل الجذر الذي تضاف إليه الزوائد).

مثال: الجذع ' تكبّر' (takabbar) مكون من حروف الجذر 'ك ب ر' (k b r) ومن النموذج (taR1aR2R2aR3) حيث R1 وR2 وR3 تمثل حروف الجذر 1 و2 و3 وبالإمكان تعويضها بالترتيب على النحو التالي (R1=k R2=b R3=r).

ملاحظة: الأسماء التي لا يمكن تحليلها إلا جذور ونماذج تم الاتفاق على تسميتها بأشباه الجذوع (Quasi-stems).

مثال: *kahramân*, ‘amber’ ، *’ismâ‘îl*, ‘Ishmael’.

- **اللواحق (Suffixes):** مثل نهايات الأفعال والحالات الاسمية والنهايات الاسمية للمؤنث 'ات' وغيرها.

- **الزيادات اللاحقة (Enclitics):** في اللغة العربية تعتبر الزيادات اللاحقة ضمائر متممة.

مثال:

كلمة: س + ي + فعل + ون + ها.

تشمل مختلف الجزاءات بنية الكلمة العربية المشروحة سابقا.

III.2-4-2. الارتباط المشروط للسوابق واللواحق:

من الأمثلة التي تبين بأن ارتباط السوابق واللواحق مشروط نذكر:

الكلمة: لا + ال + فعل + ين ة فهي كلمة مقبولة

بينما: لا + ال + فعل + ان ة مع 'ان' للتثنية فهي كلمة غير مقبولة.

بالنسبة للفعل المضارع فهو يقبل سابقة ولاحقة مشروطة:

- فعل + ي-ون ---> يفعلون.
- فعل + ت-ن ---> تفعلن.

III.2-4-3. الزوائد المزدوجة المرتبطة بالسوابق واللواحق:

بعض الزوائد تتلازم فيما بينها مثل ي-كتب-ون سواء في العربية أو في غيرها من اللغات كما سلفنا في التعريف بالزوائد المزدوجة. وللتعامل مع هذا النوع من الزوائد عند إستعمال برنامج المنسبل تم الاستعانة بالتقنية Circumfix بينما استعملت حلول أخرى في الإصدارات السابقة لبرنامج تسبل قبل ظهور الإصدار 0.61 الذي أدرج تقنية تدعم الزوائد المزدوجة حملت اسم 'Circumfix'.

III.2-4-3-1. Aspell طلب:

- الإصدارات السابقة لبرنامج تسبل (قبل الإصدار 0.61) لم تكن تحتوي على تقنية لدعم الزوائد المزدوجة الأمر الذي كان يستوجب إضافة كلمات جديدة للقاموس واستعمالها كجذوع قاموسية.

الحلول:

للحصول على يفعلون يجب أن يحتوي القاموس على كلمة فعلون أم نضيف إليها السابقة ي-ت أو المفردة يفعل أم نضيف إليها اللاحقة –ون.

الإشكالات:

تضخم حجم القاموس.

الحل الأمثل:

إستعمال تقنية ضغط القاموس في برنامج تسبل.

- ابتداءا من الإصدار 0.61 لبرنامج تسبل تم إدراج التقنية التي تدعم الزوائد المزدوجة وتم تسميتها بـ 'Circumfix' مثل نظيرتها في برنامج المنسبل.

III.2-4-3-2. طلب مشترك:

في البرنامج الأصلي لهانسبل الزوائد المماثلة بالعلم CIRCUMFIX قد تكون على كلمة عندما تكون إذا الكلمة لها أيضا سابقة مع العلم CIRCUMFIX والعكس بالعكس. و بذلك يتم إجبار ربط السوابق باللواحق.

مثال:

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

في هذه الحالة لدينا القاعدة:

SFX N 0 ون/XY

و التي تعني بان العلم X يقرر أن ترتبط اللاحقة 'ون' بالسابقة Y. فنحصل على 'يفعلون' و'تفعلون'.

و لدينا القاعدة:

SFX N 0 ن/XW

و تعني بأن العلم X يقرر أن ترتبط اللاحقة 'ون' بالسابقة W.

فنحصل على 'تفعلان' فقط وليس 'يفعلان'.

ملاحظة: نلاحظ انه كلما تنوعت السوابق كلما استعملنا تقنية الكثر.

III.2-4-2-4. الطريقة المقترحة لتطوير المنسبل:

بعض السوابق واللواحق التي تتردد دائماً سنجمع في ظهور ما يسمى بالسوابق واللواحق المزدوجة والتي جمعت بين السوابق واللواحق في قاعدة واحدة.

مثال:

ي-ون ، ت-ون ، ت-ن كلها زوائد مزدوجة للفعل.

العلامة '-' (نقاط ليس تطويل) تدل على الزوائد المزدوجة والتي تستبدل بجذع الكلمة.

ي-ون + فعل ----> يفعلون.

تفعيل الآلية يتم بواسطة الأمر SPLITAFFIX في بداية ملف الزوائد.

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

الحل المقترح:

- السوابق التي لا تحتوي على العلامة فاصلة "-" تعامل كسوابق عادية.
- يمكن استعمال السوابق المزدوجة مع الخانات الجديدة (البدال).
- يمكن تركيب السوابق المزدوجة باللواحق المزدوجة.
- يمكن إعطاء اقتراحات حسب الزوائد المزدوجة.
- الفرق بين السوابق المزدوجة واللواحق المزدوجة هو جهة الحذف.

مثال:

سابقة Y | ي-ون

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

لاحقة Y | ي-ون

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

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

استفاد => يستفادون

بتركيب السابقة

سابقة Y | ي-ون/S ← تحذف الأحرف الأول.

لاحقة S ă 0 ← تحذف الأحرف المفاورة.

III.3. الحل المقترح لتطوير المنسبل:

اعتمد مدقق المنسبل في بداياته على قاموس تيم بوكوولتر 'TIM BUCKWALTER' لكن بعد تقدم مشروع فريق ميسبل تبين عدم جدوى هذا الحل نظرا للنقائص العديدة التي تعيق هذا القاموس والتي تم تناولها بالتفصيل في الفصل الأول. وحتى محاولات التعديل التي اقترحها طه زروقي والتي أجريت على هذا القاموس لقاق نقص المصدر. بعد كل هذا تبين بجلاء أن بد من التفكير في إعداد قاموس جديد مفتوح المصدر وهو الأمر الذي اقتضى بضرورته فريق ميسبل وقام بالإشراف على إنجازه محمد كبداني من المغرب بهدف إنشاء قاموس منظم بطريقة منهجية ومستنبط من أهم وأغنى المعاجم المتداولة على الساحة

اللغوية العربية مثل المعجم الغني والمعجم الوسيط ولسان العرب وغيرها من المعاجم التي تعتمد على شروط وافية وتقنيات دقيقة وجد مهمة الفاعة فيما يخص قائمة الأفعال.

III.3-1- الدراسة التحليلية لمشروع المنسبل:

لإنجاز قاموس مشروع منسبل اعتمد محمد كبداني في الأساس على مرجعين أساسيين دون أن يهمل بقية المراجع محاولا الاستفادة من كل ما هو إيجابي في هذا المراجع للحصول على تقنية شامل وهام بمختلف أنواع الأفعال دون إهمال عملية تحليل وتقنية قائمة الأسماء الموجودة في هذا القواميس.

III.3-1-1-1. الإطار العام:

تأليف عبد الغني أبو العزم و هو مرجع قد ارتبط اسمه بمنتجات شركة خاص للبرمجيات.

III.3-1-1-1. الإطار العام للمشروع:

يتميز هذا المعجم بما يلي :

- تقنية الأفعال إلى حوالي 10000 فعل.
- طريقة تقنية شاملة وممنهجة.
- تقنية الأفعال حسب التعدي واللزوم (الفاعلية أغلبتها المعاجم الأخرى).

III.3-1-1-2. الإطار الموضوعي:

يعاب على هذا المعجم بعض المسائل التي يمكن تلخيصها في النقاط التالية:

- عدم تناول مسألة التعدي إلى مفعول به واحد أو مفعولين.
- عدم تمييز نوع المفعول به (عاقل أو غير عاقل).
- عدم وضع تقنية خاصة ببعض الأفعال مثل الأفعال ناقصة التي تأخذ أفعال القلوب.

III.3-1-1-3. الأدوات المستخدمة في الإطار العام:

قرر محمد كبداني الاستعانة ببعض المراجع الأخرى (كالمعجم الوسيط والمعجم المحيط ولسان العرب وغيرها من المعاجم العربية الرائجة التداول والاستعمال) لاستكمال التقنية السابقة الذكر حتى تكتمل المادة اللغوية و تفي بالأهداف المرجوة من هذا المشروع.

III.3-1-2. الشاشة الأولى لإطار المشروع:

أنشأه كل من سام عمار و يوسف ديشي (مؤسسة الفاتيه – باريس - 1999). ارتبط اسم هذا المرجع بالمشروع الأوروبي 'DIINAR' (القاموس المعلوماتي العربي).

III.3-1-2-1. الإطار العام للشاشة:

- اعتمادها على منهجية جد دقيقة.
- اعتمد في تأريخ الفعل العربي على نتائج دراسة لأفعال الواردة في المعجم الوسيط (10.000 فعل يومها).
- تقنين الأفعال حسب الأنواع والأوزان التي تتبع قواعد اللغة العربية.
- حصر نماذج تأريخ الأفعال العربية رغم تنوعها إلى 127 نموذجا من بينها نماذج مركبة تجمع بين العلة في التوسيع والإدغام في الهمزة.

III.3-2- الأفعال النموذجية المعجمية المرجعية لإعداد الدراسة:

من بين ما استفاده محمد كبداني من الذين المرجعين بالدرجة الأولى بالإضافة إلى المراجع الأخرى السالفة الذكر والتي استعان بها لإعداد الذي القاموس نذكر ما يلي:

- اختيار الأفعال النموذجية من بين تلك التي تفي بمعيار التحريف إلى كل الصيغ والي:
  - الأفعال المتعدية لعاقل.
  - الأفعال المتعدية لغير العاقل.
  - الأفعال اللازمة نسبيا.
  - الأفعال اللازمة مطلقا.
- إضافة ثلاثة نماذج جديدة للأفعال التي لأمها تاء أو نون لما يحمل لها من إدغام عند التحريف.
- التخطيط لإضافة التوكيد بالنون والثقيلة في إدراج اللاحقة.

III.3-3- الخطوات المتبعة لإعداد الدراسة:

يتكون المعجم المرجعي من عدة أجزاء مثبتة حسب الحروف الأبجدية. تحتاج الذي المادة المعجمية والمتكونة من أفعال وأسماء وحروف مشروحة ومصحوبة بأمثلتها الصرفية ولزومها وتعديها في بعض الأحيان والمتوفرة بشكل إلكتروني على موقع شركة صخر للبرمجيات إلى عملية معالجة تتمثل في عزل المطردات وتعليمها برموز لها معاني أو أمثلة صرفية قد نحتاجها عند بناء ملف الزوائد.

III.3-3-1. تصنيف الأفعال:

تصنف الأفعال بشكل عام حسب علاقتها بالمفعول به من حيث التعدي واللزوم لكننا في التصنيف الذي سنقدمه لاحقا والذي اعتمد عليه محمد كبداني لإنشاء قاموس المدقق الإملائي العربي سنقترح بعض المزايا الإضافية كاللزوم النسبي والمطلق والتعدي إلى مفعول به واحد أو مفعولين عاقلا كان أم غير عاقل لضرورتها في الذي المجال من أجل الحصول على تصنيف أكثر دقة يساهم في تسهيل وتبسيط عملية التدقيق بالنسبة للمدقق الذي سيعتمد على الذي القاموس.

III.3-3-1-1. الأفعال اللازمة:

معيار أن تكون على وزن : ' فَعَلَ ' – ' يَفْعُلُ '. الذي الأفعال لا يشتق منها الأمر (حسب بعض المراجع عكس ما يذهب إليه صاحب المعجم الغني) و لا المبني للمجهول ولو في صيغة المطرد الغائب.

مثال: حسُنَ - يحسُنُ ، وَجَلَ – يَوجَلُ .

في الذي الحالة نعلم الفعل بالرمز **'5'** بعد ضبط مائل ثم نحذف الشروط فتكون النتيجة المحصل عليها كما في المثال الآتي :

حَسُنَ / ـُ

III.3-3-1-2. الأفعال المتعدية:

معيار لا بد أن تكون على وزن مفتعل عن ' فَعَلَ ' – ' يَفْعَلُ '. هذا النوع من الأفعال لا يأتي منه في المبني للمجهول إلا صيغة الغائب المفرد و لكنه يصير في الأمر.

في هذه الحالة يُرمز للفعل الرمز **"ذ"** بعد الضبط المائل أم نحذف الشروط فتكون النتيجة كما في المثال الآتي:

:ثلاثي

سأل / ن

:شروط

في حالة اللزوم لا تهمنا الشروط لتحديد نوع اللزوم إذ يكفي الاطلاع على وزن الفعل لمعرفة نوع اللزوم.

III.3.3-1-3-3. الأفعال المتعدية:

عكس حالة اللزوم ففي حالة التعدي التي تتطلب التمعن في الشرط لمعرفة ما إذا كان التعدي إلى مفعول واحد أو مفعولين وما إذا كان عاقلا أم لا. بعض الحالات تستوجب الاستعانة بمراجع وقائمة في حالة الشك في إمكانية وجود عاقل مجازي.

في حالة افتراض معنى غير الذي تعنيه الشروط المواكبة توضع العلامة المتغيرة بين قوسين {} حتى نعود إلى التفنيد مرة أخرى للتأكد وهكذا.

استعان محمد كبداني لتحديد هذه الأفعال بالرموز الآتية:

- إذا كان الفعل متعديا إلى مفعول به واحد غير عاقل يُرمز الرمز **'د'.**
- إذا كان الفعل متعديا إلى مفعول به واحد عاقل بمعنى (حقيقي أو مجازي) يُرمز الرمز **'ع'.**
- إذا كان الفعل متعديا لمفعولين يُرمز الرمز **' R'.**

:شرح

- نَزَلَ - [ن ز ل]. (ع:نزل). نَزَلَ يَنْزِلُ . ف م . نَزَلَ يَنْزِلُ . 1." نَزَلَ الرَّجُلُ " : حَلَّ بالمكان. 2. " نَزَلَ الناقة " : ذبحها بالنَّازِل.

الشكل <---------- نزل / د {ع}.

- قرأ – [ق ر أ].(ع.لغ.متعدي.بحرف). قرأ يَقْرَأُ . ف م . قَرَأَ يَقْرَأُ . 1." قرأ القصيدة " : نطق بكلماتها المكتوبة. 2." قرأ كتابا من القرآن " : نطق بألفاظها عن ظهر قلب أو عن حفظ. 3. " يقرأ الكتاب " : يتتبع كلماته و لا ينطق بها. 4. " قرأ عليه السلام " : بلغه إياه.

الشكل <---------- قرأ / د .

- أَلَمَّ – [أ ل م]. (ع:رباعي.متعد). أَلَمَّ يُلِمُّ أَلَمًّا . ف م .أليم. 1."أَلَمَّ الرَّجُلُ" : رماه بالألم و نسبه إليه. 2. "أَلَمَّ القارئ" : أخذ من أَلَمِه.

الشكل <---------- ألم / ع.

- كسا – [ال ع و]. (ع.ث.متعد). كسوة. كسا f م .كسوة. 1."كسا إبنه ثوبا" : ألبسه إياه.
2. "كسا احبه ثوبا" : أعطاه إياه. 3."كسا الثلج الجبال" : غطاها. "سوف يأتي الربيع ليكسوكم أوراقا خضراء".

الإجابة ----------> ثوب / ج.

III.3-3-2. الإحصاءات:

المطلوب بالنسبة للأسماء حذف الشروح والأبقاء فقط على التأنيث والجموع إن وجدا.

:

- أرض / ج: أراض.
- سماء / ج – ج: ونون _ ات.
- أبيض / بيضاء – ج: بيضاوان بيض.
- كاتب الكاتب/ ة.
- كسا / ا

III.3-4- تطوير الإحصاءات المذكورة سابقا:

لقد سامت التحسينات المذكورة سابقا في حل بعض المشاكل التي واجهها مشروع فريق تيسير لأعداد قاموس المدقق الأملائي العربي كغيرها من المدققات الأملائية التي دعمت اللغة العربية. من أهم هذه المشاكل نذكر مشكلة حجم القاموس والتي سامت تحسينات ملتاة الزوائد في برنامجي تسيير والمخزون السابقة الذكر في تقليصها بشكل كبير لكل حسب وتيرتها التي أنجزت من اجلها كما هو موضح في ما يلي:

- حل مسألة إعتال الشكل والتي أدت إلى تقليص حجم القاموس والأكتفاء بجذر الكلمة غير المشكول في ملف القاموس. بالإضافة إلى تفادي اقتراح حلول خاطئة عند التصحيح.
- حل مسألة الأدغام والتي أدت كذلك إلى تقليص حجم القاموس إذ يتم وضع جذر الكلمة المدغمة فقط في القاموس ويتم وضع قواعد الألحاق الخاصة بالأدغام للحصول على الأشكال الممكنة لهذه الكلمة عند تطبيق هذه القواعد.
- نسى الشيء بالنسبة لمسألتي الأعلال والأبدال والتين سامتا بدورهما وبشكل كبير ليضا في تقليص حجم القاموس وخاصة قائمة الأفعال الجوفاء المعتلة والتي تتعرض لتحولات على مستوى الجذور كما رأينا في الأمثلة السابقة على الفعل 'قام'.

III.3-5- تحليل نتائج التحسينات:

اشتكى بعض مستخدمي هذا القاموس على شبكة الأنترنت من البطء الذي ظهر على البرنامج عند اقتراح الحلول والذي يعود سببه إلى تعقيد ملف الزوائد وكثرة النماذج التصريفية للفعل العربي. الأمر الذي ينال اهتمام محمد كبداني الذي وعد بإيجاد حل لهذه المسألة في إصدار 2008 موضحا المنهجية التي سيتم إتباعها من خلال خارطة طريق هذا المشروع.

III.4. خاتمة الفصل:

في هذا المثال سنحاول تشكيل قاموس يتم بعد الكلمات التي اقترحها محسن التنيتات المذكورة سابقا بالنسبة للأفعال والأسماء حاولنا أخذ كلمة واحدة (فعل أو اسم) من كل جذر لتشكيل هذا القاموس نحاول بعدها تطبيق أهم الأفكار التي درسناها سابقا والتي طورت أيضا من أجل اللغة العربية على هذا القاموس لنحصل على قائمة من الكلمات المقبولة وأخرى للكلمات الغير مقبولة بالنسبة للمدقق الإملائي العربي لمشروع ويسبل.

--------- ملف القاموس ar.dic ----------

T1 / حسان

T2 / جلس

T3 / مد

T4 / سال

T5 / استمر

T6 / سار

T7 / كسا

T8 / استدعى

T9 / قرأ

M1 / درس

M2 / باقي

M3 / كسال

---------- ملف الزوائد ar.aff ----------

```
Set UTF-8
LANG ar

IGNORE ـَـُـِـْـٌـٍـً

GEMENATING G
SFX T3 Y 5
SFX T3 0 ت
SFX T3 0 تما
SFX T3 0 ت / G
SFX T3 0 ا / G
SFX T3 0 ي / G

CIRCUMFIX X
PFX Y Y 2
PFX Y 0 ي / X
PFX Y 0 ت / X

PFX W Y 1
```

PFX W 0 ت / X

SFX T2 Y 2
SFX T2 Y 0 ون / XY
SFX T2 W 0 ن / XW

SPLITAFFIX PFX T2 Y 5
PFX T2 0 ي
PFX T2 0 ت
PFX T2 0 ي-ون
PFX T2 0 ت-ون
PFX T2 0 ت-ن

PFX T5 Y 1
PFX T5 ا ي-ون # (استفعل=> يستفعلون) حذف الألف من بداية الكلمة و إدراج السابقة المزدوجة ي-ون

PFX T7 Y 1
SFX T7 ا ي-ون # (دعا=> يدعون) حذف الألف من نهاية الكلمة و إدراج اللاحقة المزدوجة ي-ون

SFX T6 Y 1
SFX T6 ى ي-ون

PFX T8 Y 2
S/ون-ي ا T8 سابقة # إدراج زائدة مزدوجة مع حذف من الجهتين
0 ة S لاحقة

(c) # الفعل الأجوف المعتل

SFX T4 Y 1
SFX T4 0 K
K.ا T4 لاحقة
PFX T4 Y 2
ي.ي ا. T 4 سابقة
ي.K ا. T 4 سابقة

SPLITAFFIX T9 Y 4
S/ون-ي T9 سابقة
S/ون-K T9 سابقة
ن-K T9 سابقة
ة ا S لاحقة

PFX M1 Y 2
S / ل M1 سابقة
ا. S لاحقة
PFX M1 Y 1
ا.ل M1 سابقة

SFX M 2 Y 2
SFX M2 0 ة
SFX M2 ي 0

SFX M3 Y 2
SFX M3 0 ان
SFX M3 0 ă
و. M3 لاحقة

---------- الكلمات الخائبة txt.good ----------

حسان
حسانة
حسنة
ب
كب
بب
ي
كبب
تكببب
يجاب
يجلسون
تجلسون
تجلسان
سالك
يسيل
تسيل
يستمرون
يسرون

يكسون

يقرأون

تقرأون

تقرأن

أتار

أفر

باقية

باك

كسلان

كسلا

كسول

---------- الكلمات الخاطئة txt.wrong ----------

مددا

مددي

جلسون

جلسن

يسال

استمرون

يسرة

سرون

يكسا

يقرأون

قرأون

يقرأن

أتار

باقة

كسلا

III.5. الإثراء الآلي لمدقق الإملائي العربي:

يسعى فريق ميسبل إلى التفكير في إيجاد حلول لبعض المسائل التي مازالت تعيق عمل المدقق العربي بالشكل الذي ينبغي أن يكون عليه بالإضافة إلى تطوير البعض الآخر في محاولة من فريق ميسبل للرقي بهذا المدقق

إلى أعلى الدرجات ليكون في مستوى طموحات مستخدمي البرمجيات العربية. من هذا المنطلق التطورات المستقبلية لمشروع ميسبل نذكر ما يلي:

III.5-1- اعتبار الشائع:

يتم التفكير حاليا في محاولة اخذ الشكل بعين الاعتبار عند تدقيق الكلمات بدلا من إهماله الشيء الذي قد يساهم في اقتراح الكلمات البديلة الملائمة وذلك عن تقليل عدد الكلمات المكونة لقائمة الكلمات المقترحة في حالة الخطأ.

III.5-2- الإدراج! عدد اختبارات الخطأ:

الحل الذي تم عرضه في الفقرات السابقة والمتعلق بمسألة الاستخدام يتيح حالة إلحاق الفعل بلاحقة واحدة فقط ولكن هناك حالات كثيرة تعقيدا عندما يتعلق الأمر بتركيب زوائد (سوابق ولواحق في آن واحد) . لا تزال هذه الحالات قيد الدراسة والتحليل لمعالجة مختلف الحالات الممكنة.

III.5-3- اختبار صحة الخطأ:

تم التفكير في مسألة تعريف ما الزوائد بسبب المشكلة التي تمت مصادفتها عند صياغة ما الزوائد والتي تتمثل في مشكلة الاتجاه من اليمين إلى اليسار التي تظهر عليها القواعد عند الكتابة.

مثال : PFX Y a b

إذا عوضنا حرف 'a' بـ "ة" و حرف 'b' بـ 'ي' فستحصل على :

PFX Y ي ا

مما قد يجعلنا نخطئ في قراءة القاعدة التي تعني :

سابقة Y ا ي

الخطأ الناتج عن الحاجة:

بعد الملاحظة تبين أن برنامج المنسبيل لا يعتبر كلمتي PFX و SFX في سطر تعريف القاعدة ولكنه يعتبرها فقط في تعريف عائلة القواعد.

```
SFX X Y 1
SFX X 0 a
```

في هذا المثال يمكن تعويض كلمة SFX في السطر الثاني بأي كلمة غير معدومة مهما كانت وعليه فالمثال التالي صحيح:

```
SFX X Y 1
TAHA X 0 a
```

وعليه فالكتابة التالية تعتبر صحيحة:

```
SFX X Y 1
لاحقة X ة ي [ ة ]
```

III.5-4- إضافة الأخطاء! الأخطاء الناتجة عن الأخطاء!:

هذا الأمر ممكن حاليا لكن فريق ميسبل يود استغلاله على النحو التالي:

- **00!** : للماسي
- **00 f** : للمسارع
- **12ت** : للعطل

أي باستخدام حرف أبجدي عربي يليه رقمين لكن هذا غير ممكن في الوقت الراهن. الشيء الممكن هو ما يلي:

- حرف يونيكود واحد. مثل: ة ت ـ
- رقم واحد من 1 إلى 65000. مثل : 100ة 2500
- حرف لاتيني و رقم واحد . مثل.: P1 ة X5.

- **الإضافات:**

قمنا في هذا الفصل باستعراض أهم التحسينات التي أجراها فريق تيسبل على برنامجي المنسبل وتسبلة على مستوى الملفين المشكلين لكلا البرنامجين ونقصد كلا من ملف الزوائد وملف القاموس.

فعلى مستوى ملف الزوائد تم التطرق إلى أهم المزايا التي تم تطويرها في البرنامجين من أجل إيجاد حلول لبعض المسائل التي شكلت نقاط ضعف هذا المدقق كتعقيد الشكل والانسجام والانفصال والتبادل والتعامل مع الزوائد المزدوجة.

أما على مستوى القاموس فرغم كون قاموس مشروع تيسبل لم يصل إلى مرحلة الكمال بعد إذ لا تزال هناك العديد من الأمور قيد البحث و التجريب. لكن الشيء الأكيد هو كون هذا المشروع يتقدم ويتطور بخطى ثابتة ويلاقي ترحيبا وتجاوبا كبيرين من طرف مستخدمي اللغة العربية على الأنترنيت والذين لا يدخرون جهدا للتأكد بملاحظاتهم واقتراحاتهم التي تساهم بشكل فعال في تقدم هذا المشروع.

يجدر بنا التنويه في نهاية هذا الفصل بتلسكة المصدر المفتوح التي ساهمت في تطور وارتقي مستمر للبرامج المفتوحة المصدر بما فيها تلك الداعمة للغة العربية بفضل تكلفة مادية وبشرية ممكنة.

في الفصل القادم سنتطرق إلى تفصيل بعض الجوانب التقنية لبرنامج المنسبل من خلال شرح بعض الطرق والمراحل العملية الواجب إتباعها لتجريب برنامج المنسبل بصورة صحيحة والاطلاع بشكل تجريبي وعملي على أهم المزايا التي يوفرها هذا البرنامج المفتوح المصدر.

الجدول الرابع

الملائمة التقنية لبرنامج انسبل

IV. الفصل الرابع: الهندسة التقنية لبرنامج المنسبل وتحليل مميزاته

- **مميزاته:**

بعد التعرف على الهندسة العامة لبرنامج المنسبل وعلى طريقة عمله ومميزاته والمهمة التي تم تطويرها أو إعادة برمجتها لدعم تدقيق اللغة العربية سنقوم في هذا الفصل الأخير من هذه المذكرة بالحديث عن الهندسة التقنية لبرنامج المنسبل حتى تكتمل الصورة عند المتابعين لهذا المشروع (مشروع تطوير مدقق إملائي عربي مفتوح المصدر) و لكي يتمكن من يريد تجربة هذا المدقق من معرفة المراحل الواجب إتباعها لتجريب هذا المدقق بشكل صحيح وليستفيد الجميع من قريب أو من بعيد فيما كان تخصصهم مهتمين بالبرامج الحرة كانوا أو مبرمجين أو لغويين أو مجرد فضوليين غيورين على لغتهم وراغبين في المساهمة في تطوير البرمجيات الداعمة لها و لو بمجرد رأي.

سنتناول في هذا الفصل كيفية توليف برنامج المنسبل من محرر أوبن أوفيس تحت كل من نظامي لينكس وويندوز ثم كيفية تثبيت البرنامج وتحريره (Compilation) تحت كل من النظامين السابقين لنختم هذا الفصل بذكر أهم الجزاء والملفات ومكتبات برنامج المنسبل.

IV.1. تحميل محرر أوبن أوفيس تحت نظام التشغيل لينكس:

بالإمكان تحميل آخر إصدار تجريبي لحد الساعة (الإصدار Beta2 الذي تم إصداره بتاريخ 2007/09/05) من برنامج المدقق الإملائي المنسبل f ليا باللغة العربية و الذي يمثل آخر ما توصل إليه فريق يسبل من أجل إعداده لمشروعه الهادف إلى إنشاء قواميس في اللغة العربية للتطبيقات المكتبية من صفحة الإنترنت التالية :

http://sourceforge.net/project/downloading.php?groupname=ayaspell&filename=hunspell-ar_20070905_beta2.tar.gz&use_mirror=switch

بعد تحميل برنامج المنسبل هذه الآن سلسلة المراحل الواجب إتباعها لتوليف القاموس على محرر المجموعة المكتبية أوبن أوفيس تحت كل من نظامي لينكس و ويندوز.

IV.1-1- تثبيت المدقق الإملائي الإصدار الثاني لمحرر أوبن أوفيس:

هذه سلسلة المراحل الواجب إتباعها من اجل تثبيت وتوليف المدقق الإملائي الإختباري المنسبل على محرر المجموعة المكتبية أوبن - أوفيس OpenOffice.org تحت نظام لينكس (Gnu/Linux):

- **الأولى**: تثبيت التدقيق إملائي لأية لغة تريد على أوبن-أوفيس حتى تنشط آلية المدقق الإملائي وتحمل المركبات والبرامج الضرورية له.
- **الثانية**: من اللازم إغلاق Openoffice إن كان مشغلاً وتحييد برنامج Quickstarter إن كان منشطاً من قبل على النظام.
- **الثالثة**: نسخ الملفين ar.dic و ar.aff في الدليل /usr/share/myspell/dicts/

cp verb_ar.dic/ usr/share/myspell/dicts/ar.dic

cp verb_ar.aff/ usr/share/myspell/dicts/ar.aff

و كذلك التثبيت حسب تشجير النظام (openoffice.org2.1/share/dict/ooo) مثلاً.

- **راجع:** تحرير /etc/openoffice/dictionary.lst بإضافة السطر الآتي مثلاً (لن تنسى حذف الأحتياط):
DICT ar TN ar
- **ملاحظة:** تشغيل برنامج أوبن - أوفيس من خلال الأعداد (دواك أم كياراك) عين اللغة المراد تدقيق إملائها: عربية (تونس) في المثال.
- **ملاحظة:** افتح ملف الأختبار: verb_conj.odt (ملف تجريبي الأفعال العربية).

IV.1-2- تثبيت وتوظيف المدقق الأملائي الأختباري المنسبل على محرر المجموعة المكتبية أوبن- أوفيس:

يتم تثبيت وتوظيف المدقق الأملائي الأختباري المنسبل على محرر المجموعة المكتبية أوبن- أوفيس (OpenOffice.org) تحت نظام مايكروسوفت ويندوز بإتباع المراحل التالية :

- **الأولى:** تثبيت التدقيق إملائي لأية لغة تريد على أوبن-أوفيس حتى تنشط آلية المدقق الأملائي وتحمل المركبات والبرامج الضرورية له.
- **الثانية:** من اللازم إغلاق Openoffice إن كان مشغلاً وتحييد برنامج Quickstarter إن كان منشطاً من قبل على النظام.
- **الثالثة:** نسخ الملفين ar.dic و ar.aff في الدليل :

Program Files\OpenOffice.org 2.1\share\dict\ooo\

- **راجع:** تحرير Program Files\OpenOffice.org 2.1\share\dict\ooodictionary.lst\ بإضافة السطر الآتي مثلاً (لن تنسى حذف الملف الأصلي من باب الأحتياط):
DICT ar TN ar
- **ملاحظة:** تشغيل برنامج أوبن - أوفيس من خلال الأعداد (دواك أم كياراك) عين اللغة المراد تدقيق إملائها: عربية (تونس) في المثال.
- **ملاحظة:** افتح ملف الأختبار: verb_conj.odt

IV.2. تثبيت المترجم (Compiler) وتوظيفه:

IV.2-1- تثبيت برنامج المترجم الخاص بالمنسبل:

لتثبيت برنامج المنسبل بإضافة إلى القاموس يجب إتباع الخطوات التالية :

- تأكد أولاً بأن لديك الحد الأدنى لمتطلبات المنسبل الحالية مثبتة.
- عليك بتثبيت كل من برنامج المنسبل بإضافة إلى قاموس معين. تجدهم موزعين على شكل حزمة متفرقة.

- يمكن نسخ القاموس في أي مكان منذ البداية المسألة الكامل لكل من الملفات (.dic و .aff) سوف يعطي لهانسبل.

حزمة الهنسبل و القاموس تحتوي على ملفات للشرح نجد فيها تعليمات التثبيت الآن نذكرة عامة على محتواها:

يجب في البداية تثبيت الهنسبل أم تثبيت ملف القاموس و في الأخير تثبيت وحدة واجهة 'Perl' لهانسبيل.

مثال تثبيت الهنسبيل :

1. فك الضغط tar xvfz hunspell.tgz.
2. cd hunspell
3. ./configure
4. ./make
5. cd ./src/hunspell
6. ./configure
7. make
8. copy makeso.sh file here
9. sh makeso.sh أم قم بعملية النسخ يدويا بالإضافة إلى تعليمتي ربط كما هو مشروح في ملف makeso.sh.
10. cd ../tools
11. ./configure
12. make

إذا استعملت تقنية سابقة لتثبيت الهنسبيل في موقع غير نظامي عليك بتحيين المسألة الذي لديك ليحتوي '$PREFIX/bin'. برنامج الأعداد للقاموس يحتاج إلى برامج مثبتة في المرحلة السابقة.

لحد هذه النقطة بإمكانك تنفيذ الهنسبيل في حالته التفاعلية (Interactive) . على سبيل المثال :

```
cd hunspell/src/tools
./example full_path_aff full_path_dic file_to test
```

IV.2-2- الواجهة Perl لهانسبيل:

تحرير و تثبيت هذه الوحدة لواجهة 'Perl' الهنسبيل :

```
$ perl Makefile.PL
$ make
$ make test
# make install
```

عند تثبيت الهنسبيل في موقع غير نظامي (مثلا إذا لم يكن لديك حق الوصول إلى الجذر) عليك بان توجه لـ ' Makefile.PL ' مكان تواجد المكتبة.

ملاحظة:

إذا كان الهنسبل مثبتا في: $HOME/local (-- prefix=$HOME/local)

و وحدة 'Perl' يجب ان تكون مثبتة في مكتبة 'Perl' : $HOME/perl_lib

فنتحصل على:

```
$ perl Makefile.PL PREFIX=$HOME/perl_lib \
CCFLAGS=-I$HOME/local/include \
LIBS="-L$HOME/local/lib -lhunspell"
```

```
$ LD_RUN_PATH=$HOME/local/lib make
$ make test
$ make install
```

IV.3. التثبيت على الأنظمة الأخرى:

للتثبيت على أرضية (win32)لا يجب أن تحصل أولا على برنامج التثبيت الكامل (Full Installer) القابل للتنفيذ (Exucutable) من الموقع : http://hunspell.net/win32/

إذا سيقوم بتثبيت المنسبل في موقع ما مثل: 'C:\Program Files\Hunspell'.

ستحتاج كذلك إلى جلب وتثبيت على الأقل واحدا من القواميس المعرفة مسبقا والمتواجدة على نفس الصفحة.

تتأكد من أن المسلك إلى دليل 'bin' لبرنامج المنسبل (مثال: C:\Program Files\Hunspell\bin\ ) موجود في مسلك متغيرات المحيط لديك. من أجل المساعدة على تحديد موقع المسلك يمكنك الاطلاع على 'تحديد متغيرات البيئة' (set environment variables) من خلال أداة المساعدة (Help Utility) لنظام ويندوز.

عليك بإعادة تشغيل النظام بعد إعداد المسلك.

ملف (.dll) يجب أن يتم تحديد موقعه قبل إستعمال برنامج المنسبل على نظام ويندوز.

بعد ذلك لتثبيت برنامج المنسبل على نظام ويندوز اكتب في سطر الأوامر 'Prompt DOS ' (الكل على سطر واحد) التعليمة الآتية :

```
C:\> ppm install http://theoryx5.uwinnipeg.ca/ppms/Text-Hunspell.ppd
```

بالنسبة لـ دار ' ActivePerl 8xx '.

أو التعليمة:

```
C:\> ppm install http://theoryx5.uwinnipeg.ca/ppmpackages/Text-Hunspell.ppd
```

بالنسبة لـ دار ' ActivePerl 6xx '.

IV.4. التركيب على الأنظام الأخرى:

إذا كنت تريد تركيب برنامج المنسبل انطلاقا من الشفرة المصدري بنفسك فقط بحاجة إلى مترجم لغة 'C' (C Compiler) والذي يجب أن يكون نفسه الذي بنيت بواسطته وحدة Perl [1]* لديك (بالنسبة ل ActivePerl فهذا يعني VC++ 6).

إذا افترضنا بأن لديك إذا بالإضافة إلى برنامج التحميل الكامل 'Full installer ' الثنائي الحزمة (Binary Package)لا من الموقع : http://hunspell.net/win32/ ، ستحتاج لهما إلى تحميل وتثبيت حزمة 'Hunspell-dev' الموجودة لهما على الموقع : http://hunspell.net/win32/ التي تحتوي على الملفات الضرورية (**.h** و **.lib**) المطلوبة لتركيب وحدة 'Perl'.

[1]* لغة برمجة النشرية 'Larry Wall' عام 1987لا لعادة إستعمال ولفائدة لغة 'C' ولغات 'scripts' مثل : sed و awk و shell (sh).

قبل تحرير برنامج المنسبل يجب التأكد من أن مسار متغير المحيط يشير إلى موقع الملف (.dll) لبرنامج المنسبل المنزل قبل البدء بعملية التحرير.

بعد ذلك تتم عملية التثبيت بنفس الطريقة التي تم وصفها من أجل إصدار Unix:

```
perl Makefile.PL
nmake
nmake test
nmake install
```

مع بعض المتطلبات الإضافية لتمرير التعليمات الأساسية (مثل INC و LIBS) لتحديد موقع ملفات العناوين الرئيسية (header files) وملفات المكتبات (lib files) في حالة ما إذا تم تثبيت الملفات في غير مواقعها النظامية. تأكد في حالة إستعمال موقع غير نظامي بأنه تم إضافة الأمر إلى مسار محيط المتغير الذي لديك قبل إجراء التجارب.

IV.5. تجارب الهندسة التقنية:

يمكن تحميل آخر إصدار للنسخة المصدري لبرنامج المنسبل (الإصدار 2.1) من الموقع التالي: ' http://cpan.waldonet.net.mt/authors/id/E/EL/ELEONORA/ ' ثم لنحمل بعدها على الدليل المضغوط ' text_hunspell_1.2.tar.gz '. بعد فك ضغط هذا الدليل يمكننا معاينة مكوناته والتي يمكن تلخيصها في ما يلي:

- **الدليل 't'**: و يتكون مما يلي:
  - ملفات تحمل التابعة '.t'.
  - الملفين : test.dic و test.aff
  - الملف test1.pl
- الملف: 'changes'
- الملفات: 'Hunspell.h' و'Hunspell.xs' و'Hunspell.pm'
- الملفين: 'makefile.pl' و 'test.pl'
- الملف: ' Makeso.sh '
- الملف: ' MANIFEST '
- الملف: ' META.yml '
- الملف: ' Perlobject.map '
- ملفي التعليمات: 'README' و 'README_1'
- الملف: 'typemap'

- **الهدف:**

مكنتنا هذه الدراسة للهندسة التقنية لبرنامج المنسبل من التعرف بكيفية توظيف مدقق المنسبل للعمل على المجموعة المكتبية (Open Office) تحت نظامي لينكس وويندوز كما أتاحت لنا شرح طريقة تثبيت وتحرير البرنامج (Compilation) تحت نظامين انطلاقا من النسخة المصدري لهذا الأخير كونه برنامجا مفتوح المصدر والتعريف بأهم مكوناته من خلال عرض قائمة لأهم الملفات والمكتبات المكونة لهذا البرنامج.

الهدف من هذه الدراسة يكمن في شرح عملية تجريب برنامج المنسبل بطريقة عملية صحيحة حتى يتمكن مستخدمو هذا البرنامج والمطلعين على هذا البحث عند تجريبه من التعرف عن كثب على أهم مميزاته التقنية وتقييم النتائج التي يمكن الحصول عليها عند تجريبه على برنامج المجموعة المكتبية أوبن أوفيس لتدقيق النصوص العربية لتأكد بعدها بآرائهم واقتراحاتهم عند ملاحظة بعض المسائل التي لا يزال تعاطي هذا البرنامج معها منقوصا في بعض الحالات أو منعدما في حالات أخرى الأمر الذي سيؤدي حتما إلى تطوير مردود هذا البرنامج. هذه الطريقة في التطوير شائعة الاستعمال بالنسبة للبرامج المفتوحة المصدر وقد أثبتت فعاليتها في الكثير من مراحل مشروع فريق تيسبل الهادف إلى إنشاء قاموس في اللغة العربية للمدقق الإملائي المنسبل.

Køθ¢JI hø∂žtløq

لقد تناول هذا البحث مسألة التدقيق الإملائي في البرامج الداعمة للغة العربية وبالأخص المفتوحة المصدر منها لما توفرها على المطورين من جهد ومال بتلسنتها التي ساعدت كثيرا على تطويرها وتطوير دعم العربية فيها.

فلقد تمثلت نقطة انطلاق هذا البحث في محاولة إيجاد حل لمشكلة يعاني منها معظم مستعملي برمجيات اللغة العربية وهي مشكلة التدقيق الإملائي التي ولتأسف الشديد لا تزال تشكل عائقا لمستعملي البرامج الداعمة للغة العربية والتي قد يعود السبب الأساسي لعدم تطورها إلى أن الشركات المنتجة لهذه البرامج لا تعير هذا الاهتمام الذي ينبغي أن تناله عندما يتعلق الأمر بدعم العربية فيها سواء بسبب غياب المنافسة أو بسبب مشكلة حقوق الملكية الفكرية المفقودة في معظم دولنا العربية.

وزعم أن مشروع فريق يسبل الذي سلطنا عليه الضوء من خلال هذا العمل وركزنا على دراسة مختلف مراحل تطوره والنتائج التي تم التوصل إليها لم يصل بعد إلى مرحلة الكمالية إلا أنه مشروع يتطور باستمرار وقد تمكن من إيجاد الكثير من الحلول لبعض الحالات التي استعصت عليه في البداية بالإضافة إلى العمل الجبار الذي يقوم به محمد كبداني على مستوى القاموس الذي أعاد بناءه من نقطة الصفر تقريبا معتمدا على قواميس عربية معروفة ومتداولة كثيرا في الأوساط اللغوية.

تمكنت من خلال هذا البحث من الاطلاع على كافة المشاريع التي تناولت مسألة التدقيق الإملائي العربي والتي رغم أنني أحسست بعد دراستها بصعوبة المهمة التي تنتظر مشروع فريق يسبل لأنشاء قاموس المدقق الإملائي العربي إلا أنها أعطتني دافعا قويا في نفس الوقت لدراسة هذا المشروع والحديث عنه والمساهمة في تطويره في المراحل القادمة إن شاء الله حتى تسترجع اللغة العربية مكانتها بين اللغات وترقى البرمجيات الداعمة لها إلى مستوى باقي البرمجيات العالمية والمتوفرة بلغات مختلفة.

كما تمكنت من معرفة الكثير عن البرامج المفتوحة المصدر وعن طريقة عملها وعن مزاياها المتعددة التي ساهمت وبشكل كبير في تطور هذا النوع من البرامج والتي يرى فيها الكثيرون مستقبل البرمجة لما فيها من تكامل للجهود وربح للوقت وسرعة وفعالية في التطوير.

إن مشروع فريق يسبل يستحق كل التشجيع والمتابعة وهو بحاجة إلى مساهمة الجميع مبرمجين ولغويين ومهتمين بالبرامج الحرة للوصول به إلى درجة عالية من الرقي والاحترافية حتى يصل هذا المشروع إلى بر الأمان وينجز أول مدقق عربي مفتوح المصدر قابل للمناقشة والتجريب من خلال الإصدارات المختلفة لهذا المدقق التي يتم تطويرها ثم تجربتها وبعد ذلك نشرها على موقع المشروع ثم شرح لمختلف المميزات التي

يتم إضافتها ومعالجة الحالات التي يتم معالجتها والأخطاء التي يتم تفاديها حتى يتسنى للمهتمين بهذا المشروع الاطلاع عن كثب على النتائج المتحصل عليها عند استخدامه لتدقيق النصوص العربية الآلي بعد ذلك بخرائطهم واقتراحاتهم والتي قد تكون جاهزة في الأخطاء التي مازال يتجاهلها هذا المدقق المتطور باستمرار حتى تلبية جميع متطلبات معايير المدققات العالمية.

نود الإشارة في الأخير إلى كون هذه الدراسة تعتبر أول بحث أكاديمي يتناول مراحل تطور وإنجاز فريق مشروع يسبق لإعداد قاموس المدقق الإملائي العربي وهدفها الأساسي يتمثل في دراسة معالجة تجارب المدققات الإملائية المفتوحة المصدر الداعمة اللغة العربية بالإضافة إلى محاولات كل من طه زروقي لبرمجة بعض الأخطاء في برنامج المنسق وتطويره للتلاؤم مع خطأ اللغة العربية ومحاولات محمد كبداني لإعداد القاموس العربي من أجل تطوير مدقق إملائي عربي مفتوح المصدر يلبي حاجات مستخدمي البرمجيات العربية كما يلبي المعايير التي أنجزت وفقها معالجة المدققات العالمية المشهورة.

الملاحق

الملاحق المصطلحية:

[1] د.عبدالعزيز عبدالله المهيوبي "المصطلح اللساني في الخطاب العربي المعاصر (دراسة في المفاهيم والتأسيس)" جامعة الإمام محمد بن سعود 12 نوفمبر 2007.

[2] اقتباس عن: د. مولاي الحسن امراني علوي " اللسانيات في الخطاب العربي المعاصر: مساءلة المنهج والمصطلح والمفاهيم (دراسة في الأصول المعرفية والإبستمولوجية)" ربيع 2007.

[3] علاء الدين الزحل العجماوي " الإشكالية المصطلحية في الخطاب العربي المعاصر" المحاضرة الثانية 24 أبريل 2001.

[4] ابن منظور "لسان العرب" مادة (ل ز و).

[5] ابن جني "الخصائص" 1 / 31.

[6] د.عبد المجيد "الخطاب العربي" ص 15.

[7] ابن خلدون "المقدمة" ص 546.

[8] يوسف غازي ومجيد النعيمي "النظريات اللسانية الكبرى" ترجمة منقولة عن فرديناند دي سوسور ص 21.

[9] د.لبنى فريحة "اللسانيات العربية" ص 9.

[12] ، [22]، [28] طه زروقي "الترجمة المصطلحية في الخطاب اللساني العربي المعاصر" ندوة " البرمجيات باللغة العربية: خطوات نحو الإدارة الإلكترونية" المجلس الأعلى للغة العربية الجزائر 09-10 ديسمبر 2007.

[14] الجندي "الأدب العربي الحديث" نقلاً عن بحث لعمر الدسوقي ص 17.

[15] محمد كبداني "مشروع تيسير اللغة العربية" مشروع تيسير 10 سبتمبر 2007.

[16][30] علاء الدين الزحل العجماوي "الإشكالية المصطلحية في الخطاب العربي المعاصر" المحاضرة الثانية مجمع اللغة العربية الأردني 24 نيسان 2001.

[17] يوسف الحمادي "الأعداد (الخلاف)" دار الفكر دمشق.

[18] مجلة المشرق بيروت السنة الأولى عدد 699 في 1898 م / 1316 هـ.

[19] د. مسعد زياد "الصراع في اللسانيات".

[22] طه زروقي "الأعداد في اللسانيات" مشروع قاموس تيسير 2006/07/10.

[31] د.محمد عابد الجابري "الخطاب والمصطلح" 2003/11/19.

المصطلحات اللسانية:

[10] http://www.library.chisholm.vic.edu.au/coil/site/glossary.htm
[11] http://www.maine.gov/education/lres/review/revised_mlr_ela_glossary.htm
[13] http://www.linguistics.sil.org

[20] http://www.sakhr.com/Technology_a/Correction/features.aspx
[21] http://www.coltec.net/ArabicProofingTool.aspx
[23] http://fmg-www.cs.ucla.edu/geoff/ispell.html
[24, 26] http://aspell.net
[25] http://kb.iu.edu/data/aems.html
[27, 29] http://hunspell.sourceforge.net

المصادر والمراجع:

[ABBES R., DICHY J. and HASSOUN M., 2004] ABBÈS RAMZI, DICHY JOSEPH and HASSOUN MOHAMED. **The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source prog.** COLING'04, 20th international conference on computational linguistics. Proceeding of the workshop computational approaches to Arabic script-baseslanguages., 28 august 2004., Genova.

[BENTLEY K., 2001] BENTLEY KATIE. **Spell Checker**. Easy MSc Programming Techniques project. Cogs University of Sussex, 2001.

[BOECKX C. and FUMIKAZU N., 2004] BOECKX, CEDRIC and FUMIKAZU NIINUMA. **Conditions on Agreement in Japanese**. Natural Language and Linguistic Theory, 2004.

[BOYD A., 2006] BOYD ADRIANE. **Language and computers.** (ling 384). topic4: Writer's Aids (Spelling and Grammar Correction), Winter 2006.

[CARLSON A. and FETTE I., 2007] CARLSON ANDREW and FETTE IAN. **Memory-Based Context-Sensitive Spelling Correction at Web Scale**. School of Computer Science Carnegie Mellon University Pittsburgh, December 2007.

[KERNIGHAN M., CHURCH K. and GALE W., 1990] KERNIGHAN MARK D, CHURCH KENNETH W and GALE WILLIAM A. **A spelling correction program based on a noisy channel model.** Proceedings of the Thirteenth International Conference on Computational Linguistics, 1990. Pages 205–210.

[MARKUS D., 2007] MARKUS DICKINSON. **The computer and natural language.** Ling (445/515). Topic 4: Writer's aids (Spelling and Grammar Correction). Dept. of Linguistics. Indiana, Autumn 2007.

[MONOJIT C., MARKOSE T., ANIMESH M. and ANUPAM B., 2007] MONOJIT CHOUDHURY, MARKOSE THOMAS, ANIMESH MUKHERJEE and ANUPAM BAS. **How Difficult is it to Develop a Perfect Spell-checker? .**A Cross-linguistic Analysis through Complex Network Approach. 1Department of Computer Science and Engineering. IIT Kharagpur. India, 2007.

[PERRIER G., 2007] PERRIER GUY. **Initiation au traitement automatique des langues.** Université Nancy2, 2007.

الملحق

الملحق

هذه الوثيقة تمثل وثائقا شاملا لمدقق هنسبل ومزاياه تحتاجها عليها من الموقع الرسمي للبرنامج: http://hunspell.SourceForge.net وقمنا بترجمتها إلى اللغة العربية.

I. هيكل ملفات :

يتطلب هنسبل ملفين لتحديد اللغة التي هو بصدد تدقيقها إملائيا. الملف الأول هو عبارة عن قاموس يحتوي على مفردات اللغة والثاني هو ملف زوائد يُعرف معنى الأعلام الإملائية في القاموس.

ملف القاموس (*.dic) يتضمن قائمة من الكلمات واحدة في كل سطر. السطر الأول من القواميس (باستثناء القواميس الشخصية) يتضمن إحصائيا تقريبي للكلمة (الأمثل لبعثرة حجم الذاكرة). كل كلمة يمكن اختياريا أن تليها الشريحة ("/") بالإضافة إلى علم واحد أو أعلام كثيرة التي تمثل زوائد أو مزايا إملائية.

قاموس الكلمات يمكن أن يحتوي ليس على الشرائح ("/") مثل "" لغوي. العلم الافتراضي هو حرف واحد (عادة البجدي). في ملف قاموس هنسبل هناك ليس مجال إرفي اختياري تتألف الجدولة. الأولى الإرفية لها الشكال إملائية.

ملف الزوائد (*.aff): قد يحتوي على الكثير من الإستراتيجيات الاختيارية. على سبيل المثال SET تستخدم لتحديد ترميز حروف الزوائد وملفات القاموس. TRY تحدد حروف التغيير للاقتراحات. REP تحدد جدول استبدال لعدة تهجيات للحروف في حالة الاقتراح.

PFX وsfx تعرفان أقسام السوابق واللواحق المسماة مثل أعلام الزوائد.

ملف الزوائد التالي مثلا يعرف نظام تشفير الحروف UTF-8 . إقتراحات 'TRY' تبدأ عن كلمة سيئة مثل حرف انكليزي أو فاصلة 'Apostrophe'. مثل هذه التعاريف لـ'REP' يمكن أن يقترح هنسبل الشكل الصحيح للكلمة عندما تتضمن الكلمة الإملائية 'f' بدلا من 'ph' والعكس بالعكس.

```
SET UTF-8
TRY esianrtolcdugmphbyfvkwzESIANRTOLCDUGMPHBYFVKWZ'

REP 2
REP f ph
REP ph f

PFX A Y 1
PFX A 0 re .

SFX B Y 2
SFX B 0 ed [^y]
SFX B y ied y
```

اللغتان فئتان من الزوائد في القاموس. الفئة 'أ' تعرف السابقة '-re'. الفئة 'ب' تعرف اللاحقتي '-ed'. أول لاحقة يمكن أن تعرف إلا كلمة ما إذا كان الحرف الأخير يختلف عن 'Y'. اللاحقة الثانية يمكن أن تعرف إلا الكلمات التي تنتهي بحرف 'Y'. (انظر التذييل في وقت لاحق). ما القاموس التالي يستخدم هذه الأقسام من الزوائد.

```
hello
try/B
work/AB
```

كل الكلمات المقبولة من هذا المثال هي : hello Ü try Ü tried Ü work Ü worked Ü rework Ü reworked .

II. **قارئ عائلة :**

1- **تنسيق الترميز**:

يضبط ترميز حروف الكلمات والوحدات اللغوية الأساسية في ملفات القاموس والزوائد. القيم الممكنة:
UTF-8, ISO8859-1 - ISO8859-10, ISO8859-13 - ISO8859-15, KOI8-R, KOI8-U, microsoft-cp1251, ISCII-DEVANAGARI.

2- **نوع العلم** :

تحدد نوع العلم. النوع الافتراضي هو حرف (8-bit) extended ASCII. عوامل 'UTF-8' تحدد أعلام الحروف

'UTF-8' تشفير يونيكود. قيمة 'long' تحدد قيمة مزدوجة لنوع علم حروف ' Extended AscII' أما 'num' فتحدد نوع علم العدد العشري. الأعلام العشرية مرقمه من 1 الا 65535 وفي مجالات العلم تتجاهلها فالله.

ملاحظة : نوع العلم UTF-8 لا يعمل من منهاج (برنامج) ARM.

3- **السوابق الأحادية (complexPrefixes)** :

تحدد سابقة ثنائية النوع (لكن اللاحقة أحادية النوع) للغات الاشتقاقية من نظام كتابة من اليمين إلا اليسار.

4- **رمز اللغة (langcode Lang)** :

تحدد رموز اللغة. في المنسبل يجوز رموز لغة معينة تتيحها الآلية lANG code. وفي الوقت الحاضر اللغات : az_AZ Ü hu_HU Ü TR_tr رموز آلة في المنسبل (انظر برنامج المدرب).

5- **تجاهل الأحرف (Ignore characters)**:

تجاهل حروف من قاموس الكلمات واللواحق والكلمات المدمجة. مفيدة للحروف الاختيارية مثل علامات التشكيل في اللغة العربية (الحركات).

AF number_of_flag_vector_aliases

AF flag_vector

يستطيع المنسبل تعويض علم اللاحقة بعدد طبيعي في قواعد الزيادة (صيغ الاسم المستعار). المثال الأول من صيغ الاسم المستعار:

```
3
hello
try/1
work/2
```

تعاريف 'AF' في ملف الزوائد :

```
SET UTF-8
TRY esianrtolcdugmphbyfvkwzESIANRTOLCDUGMPHBYFVKWZ'
AF 2
AF A
AF AB
```

انظر ميزات المثلة الاختبارية / الاسم المستعار*.

ملحوظة : إذا كان ملف الزوائد يحتوي على العلم 'parameter' عرفه قبل تعاريف الـ'AF'.

الصيغ المختصرة : استخدم وسيلة 'Makealias' في توزيع المنسبل لصيغ الملتقاة Aff وdic.

```
AM number_of_morphological_description_aliases

AM morphological_description
```

يستطيع المنسبل ميزات تعويض وق ق رفي بعدد طبيعي في قواعد الزوائد (صيغ الاسم المستعار).

III. الخيارات الإضافية :

- **TRY characters**

المنسبل يمكن أن يقترح أشكال الكلمات الصحيحة عندما تخطئ الأذن الأميرة عن الشكل الخاطئ باستعمال 'TRY character ' واحد. العامل TRY جد حساس.

- **NOSUGGEST flag**

الكلمات المشار إليها بالعلم NOSUGGEST ليست مقترحة. العلم المقترح للكلمات المبتذلة والبذيئة.

- **MAXNGRAMSUGS num :**

تحدد عدد إقتراحات الـ 'n_gram'. القيمة 0 تغلق إقتراحات الـ 'n_gram'.

- **NOSPLITSUGS :**

تعطيل اقتراحات كلمة مشطورة.

- **SUGSWITHDOTS :**

تحديد نقطة (نقاط) إلى الاقتراحات إذا كانت الكلمة المدخلة تنتهي في نقطة (نقاط). (الأمر ليس كذلك بالنسبة لقواميس penoffice.org لأن openoffice.org له آلية تلقائية لتوسيع النقطة).

```
REP number_of_replacement_definitions

REP what replacement
```

يمكننا تحديد المعلومات الخاصة لتوابع اللغة في ملف الزوائد (.aff) باستعمال جدول استبدال. أول استبدال أو رمز هذا الجدول يليه سطر واحد أو عدة أسطر لمعطيات الاستبدال. بمساعدة هذا الجدول يستطيع المنسبل اقتراح الأشكال الصحيحة لأخطاء الإملائية المثالية عندما يختلف الشكل الخاطئ للكلمة عن الشكل الصحيح بأكثر من حرف واحد .

مثال : جدول تعريف استبدال انكليزي لمعالجة الحروف الساكنة الخاطئة :

```
REP 8
REP f ph
REP ph f
REP f gh
REP gh f
REP j dg
REP dg j
REP k ch
REP ch k
```

: الأولوية

1. من المفيد جدا تحديد البدائل لأخطاء ذات الحرف الواحد الأكثر مثالية لأولى: في الخاصية 'REP' يمكننا إضافة أولوية أعلى لمجموعة فرعية من اقتراحات 'TRY' (قائمة الاقتراحات تبدأ باقتراحات 'REP').
2. جدول الاستبدال يمكن استخدامه لتحديد أكثر خرامة لكلمة مركبة (من توليد الكلمات المركبة كما لو أنها مجرد كلمات بسيطة في أخطاء مثالية انظر CHECKCOMPOUNDREP).

```
MAP number_of_map_definitions

MAP string_of_related_chars
```

يمكننا تحديد اللغة التي تعتمد على معلومات عن الحروف التي ينبغي اعتبارها متماثلة (بمعنى أقرب من سائر الحروف الغير موجودة في المجموعة. في ملف الزوائد (.aff) باستعمال جدول خريطة حروف. في هذا الجدول لا يمكن أن يقترح المنسبل الأشكال الصحيحة للكلمات التي تختار بشكل غير صحيح الحرف الخاطئ من مجموعة ذات ثلاثة أكثر من مرة في كلمة واحدة.

فعلى سبيل المثال يمكن رسم خرائط لألمانية 'umlauted us' ال 'u' المنتظمة كلمة 'Frk' ' ينبغي حقا أن تكتب باستعمال 'umlauted u's' ' و ليس المنتظمة منها.

```
MAP 1
MAP u.RE
```

IV. الإرشادات الإضافية:

```
BREAK number_of_break_definitions
BREAK character_or_character_sequence
```

تحديد نقاط الفصل بالنسبة للكلمات المنفصلة وتفكك الجزئي الكلمة منفصلة.

السبب : طريقة جد مفيدة للتركيب من سلسلة من الحروف أو الرموز (مثلا كلمة ' hyphen' باللغة الإنكليزية و الألمانيه أو ' hyphen ' و ' n-dash' بالمجرية). الوثائق غالبا ما تكون نقاط فصل سيئة لتحديد العلامات لأن المركبات من الوثائق قد تحتوي على الجزئي غير صحيحة (ليس). من ' BREAK' يستطيع المنسبل تدقيق جانبي إذا المركبات تقسيم الكلمات إلى وثبة أو ن-وثبة :

```
BREAK 2
BREAK -
BREAK --     # n-dash
```

التقسيمات متكررة لا بحيث foo-bar و bar-foo و foo-foo--bar-bar ستكون مركبات صالحة.

الإستعمال:

1. ' COMPOUNDRULE' تعتبر الفصل (أو ستكون الفصل) لمعالجة الوثائق وغيرها من المركبات التي تجمع حروفا أو سلسلة من الحروف. استخدم 'BREAK' إذا أردت التحقق من كلمات بها وثائق أو حروف وصل ويتطلب وقت أو لمكانيه لو 2 دقيق لقواعد التركيب من COMPOUNDRULE) COMPOUNDRULE تعامل فقط من الألحاق الأخير للكلمة المركبة سابقا).
2. بالنسبة لأوامر التدقيق الأملائي حدد عوامل WORDCHARS:
   WORDCHARS (انظر مثال التجارب / *.break).

```
COMPOUNDRULE number_of_compound_definitions

 COMPOUNDRULE compound_pattern
```

المثال يعرف نماذج تقليدية للمركبات باستعمال تركيبة ' regex-like'. أول 'compoundrule' وهي عبارة عن عنوان رئيسي من عدد تعريفات 'compoundrule' القادمة. نماذج التركيب تتألف من أعلام مركبة وحروف نجمة أو علامة استفهام. العلم المتبوع بـ '*' يربط سلسلة مكونة من 0 أو عدة روابط للكلمات المشار إليها بعلم التركيب إذا. العلم المتبوع بـ ' ?' يربط سلسلة مكونة من 0 أو رابطة واحدة (1) للكلمات المشار إليها بعلم التركيب إذا. انظر إلى أمثلة اختبارات / التركيب *.*

3. الرموز '*' و ' ?' لا تعمل إلا من الحرف الافتراضي 8 بت و لأنواع العلم UTF-8.

4. أعلام COMPOUNDRULE لم تكن متلائمة مع COMPOUNDFLAG, COMPOUNDBEGIN, ... الأعلام المركبة سابقا (استخدم إذا الأعلام على كلمات مختلفة).

- **COMPOUNDMIN num**

القيمة الدنيا لطول الكلمات في الكلمات المركبة. القيمة الافتراضية هي 3 حروف.

- **COMPOUNDFLAG flag**

الكلمات المشار إليها بـالعلم COMPOUNDFLAG قد تكون في الكلمات المركبة (إلا إذا كانت كلمة أقصر من COMPOUNDMIN). الزوائد مع COMPOUNDFLAG تسمح ليتا بتركيب الكلمات المرفقة بزوائد.

- **COMPOUNDBEGIN flag**

الكلمات المشار إليها بـالعلم COMPOUNDBEGIN (أو مع زائدة مسجلة) قد تكون العناصر الأولى في الكلمات المركبة.

- **COMPOUNDLAST flag**

الكلمات المشار إليها بـالعلم COMPOUNDLAST (أو مع زائدة مسجلة) قد تكون العناصر الأخيرة في الكلمات المركبة.

- **COMPOUNDMIDDLE flag**

الكلمات المشار إليها بـالعلم COMPOUNDMIDDLE (أو مع زائدة مسجلة) قد تكون في منتصف عناصر الكلمات المركبة.

- **ONLYINCOMPOUND flag**

اللواحق المشار إليها بالعلم ONLYINCOMPOUND قد تكون فقط داخل التراكيب (Fuge-elements باللغة الألمانية fogemorphemes باللغة السويدية). العلم ONLYINCOMPOUND يعمل ليتا مع كلمات (انظر اختبارات / onlyincompound.*).

- **COMPOUNDPERMITFLAG flag**

السوابق يسمح بها في بداية المركبات اللواحق يسمح بها افتراضيا في نهاية المركبات.
اللواحق التي تحتوي COMPOUNDPERMITFLAG قد تكون داخل المركبات.

- **COMPOUNDFORBIDFLAG flag**

اللواحق مع هذا العلم لا تسمح بتركيب كلمة مرفقة بزوائد.

- **COMPOUNDROOT flag**

العلم COMPOUNDROOT يشير إلى المركبات في القاموس (يستخدم الآن فقط في القانون الخاص باللغة المجرية).

- **COMPOUNDWORDMAX number**

تحدد القيمة القصوى لعدد الكلمات في كلمة مركبة. (الافتراضي غير محدود).

- **CHECKCOMPOUNDDUP**

يمنع تكرار الكلمات في التراكيب (مثل foofoo).

- **CHECKCOMPOUNDREP**

يمنع التركيبات إذا كانت (سيئة غالبا) الكلمة المركبة قد تكون كلمة غير مركبة مع خطة الاستبدال (REP). مفيد بالنسبة للغات مع 'تراكيب دقيقة' الأملاء.

- **CHECKCOMPOUNDCASE**

يمنع الأعالي حالة الحروف عند تطويل الكلمات في التراكيب.

- **CHECKCOMPOUNDTRIPLE**

تمنع التركيبات اذا كانت الكلمة المركبة تحتوي على كلمات ثلاثية الحروف (مثل foo|ox or xo|oof).

المشكل : فقدان الدعم للحروف متعددة البايت في نظام التشفير UTF-8 (يعمل فقط حروف نظام ASCII 7 بت).

```
CHECKCOMPOUNDPATTERN
number_of_checkcompoundpattern_definitions

 CHECKCOMPOUNDPATTERN endchars beginchars
```

يمنع التركيبات في حالة ما إذا كانت أول كلمة في التركيب تنتهي بأحد حروف النهاية 'endchars' و كانت الكلمة المقبلة تبدأ بأحد حروف البداية 'beginchars'.

- **COMPOUNDSYLLABLE max_syllable vowels**

يستعمل في حالات خاصة من قواعد التركيب في اللغة المجرية. العامل الأول هو القيمة القصوى لعدد المقاطع التي قد تكون في التركيبات إذا كانت الكلمات في التراكيب اكثر من COMPOUNDWORDMAX. العامل الثاني هي قائمة حروف العله (لحساب المقاطع).

- **SYLLABLENUM flags**

يستعمل في حالة من قواعد التركيب في اللغة المجرية.

V. القائمة إشتقاق الزوائد:

العلم PFX عدد منتظم مركب.

العلم PFX يقوم بتعريف شرط السابقة الواجب الظرفي.

العلم SFX عدد منتظم مركب.

العلم SFX يقوم بتعريف شرط اللاحقة الواجب الظرفي.

الزائدة إما سابقة أو لاحقة متصلة بجذور كلمات أنشأت كلمات أخرى. يمكننا تحديد أقسام الزوائد بعدد تعسفي من قواعد الزوائد. أقسام الزوائد يشار إليها بأعلام الزوائد. السطر الأول من تعريف قسم زوائد إما العنوان. مجالات عنوان قسم زوائد:

(0): معيار الاسم (PFX or SFX)

(1) : علم (اسم قسم الزوائد)

(2) : منتظم مركب (إذن الجمع بين السوابق واللواحق). القيم الممكنة : Y (نعم) أو N (لا).

(3) عدد الأسطر للقواعد التالية.

: عدد القواعد الأساسية

(0) اسم المعيار

(1) علم

(2) تجريد الحروف من البداية (في قواعد السوابق) أو من النهاية (في القواعد اللواحق) للكلمة.

(3) زائدة (اختياريا مع أعلام استمرار الأقسام يفصل بينها الرمز '/').

(4) شرط.

تجريد الجذر أو لاحقة يشار إليهما بالجذر. شرط الجذر يشار إليه بنقطة. الشرط إما نموذج مبسط لشبه عبارة منتظمة يجب توافرها قبل تطبيق الزائدة. (النقطة تمثل حرف تعسفي.الأحرف في الحاشية تمثل حرفا تعسفيا من مجموعة حروف الاستبدال. الواثبة ليس لها معنى ما ولكن العلامة (^) بعد أول حاشية تحدد مجموعة الحروف المكملة أو المتممة).

(5) وجه الظرفي عادي.

VI. القائمة الزوائد:

- **CIRCUMFIX flag**

الزوائد المماثلة بالعلم CIRCUMFIX قد تكون علامة كلمة عندما تكون إذا الكلمة لها أيضا سابقة مع العلم CIRCUMFIX والعكس بالعكس.

- **FORBIDDENWORD flag**

هذا العلم يوضح الشكل الممنوع للكلمة. لأن الأشكال التي بها زوائد هي ليست ممنوعة يمكننا أن نطرح مجموعة فرعية من قبل مجموعة من الكلمات المقبولة التي بها زوائد والكلمات المركبة.

- **KEEPCASE flag**

تمنع الأشكال الكبيرة للحروف والأشكال الأساسية للكلمات المشار إليها بالأعلام KEEPCASE. مفيد لحالات إملائية. (القياسات والعملات تحتوي غالبا بحالاتها في نص و f حروفها كبيرة الشكل) أو ونظام الكتابة (مثل الاحتياط بالأشكال الصغيرة للحروف الأكاديمية).

:هوية ألمانية

من إعلان CHECKSHARPS الكلمات التي بها حرف 's' حاد والعلم KEEPCASE يمكن أن تكون أساسية وفي حالاتها العادية ولكن أشكال الحالات العادية لهذه الكلمات لا يجب أن تتضمن حرف 's' حاد. فقط SS. انظر مثلا التركيب الألماني germancompounding في دليل التجارب لتوزيع المنسبل.

- **LEMMA_PRESENT flag**

عموما المنا قاموس كلمات في نتيجة التحليل الصرفي. أحيانا كلمات القاموس ليست جذورا حقيقية (Lemmas) ولكن يمكنها أن تبدأ زوائد إلى الجذور (غير الحقيقية) والجذور الافتراضية. في هذه الحالة الجذور الحقيقية في حاجة إلى أن تتوضع في الوصف الصرفي وتمنع الجذور الغير حقيقية في التحليل الصرفي مبدئيا علم LEMMA_PRESENT إلى قاموس الكلمات.

- **NEEDAFFIX flag**

هذا العلم يشير إلى الجذور الافتراضية في القاموس. فقط الأشكال المزيدة لهذه الكلمات ستكون مقبولة من طرف المنسبل. إلا إذا كانت كلمة القاموس لها لاحق متجانس أو الزائدة صفر 'zero affix'. NEEDAFFIX تعمل أيضا من السوابق ومن سابقة + ترتيبات اللاحقة (انظر التجارب / pseudoroot5.*).

- **PSEUDOROOT flag**

مستنكر. (الاسم السابق للمعيار NEEDAFFIX).

- **حروف WORDCHARS**

WORDCHARS يطيل محدد الكلمات في سطر الأوامر لهانسبل بتحرف كلمات إضافية. على سبيل المثال أو النقطة أو الواصلة ن-واصلة الأرقام أو رمز النسبة المئوية هي أحرف كلمة في اللغة المجرية.

- **CHECKSHARPS**

زوج الحروف 'SS' في حالة الحرف كبيرة (في اللغة الألمانية) للكلمات قد تكون حالة حرف كبير الشكل ل 'S' حاد(β). المنسبل يمكنه معالجة هذه الحالة الخاصة من إعلان CHECKSHARPS (انظر أيضا علم KEEPCASE والتجارب / التركيب في الألمانية مثلا) وفي كل من الأمثلة والأقتراح.

VII. الخاتمة:

تحتوي قواعد الألحاق في المنسبل على مجال وتفرعي تياري. وثمة مجال التياري مماثل في ملف القاموس تستعملهما الجدولة:

كلمة / العلام الفرع

تحدد مورد بسيط بواسطة المعلومات الفرعية.

ملف الزوائد :

```
SFX X Y 1
SFX X 0 able . +ABLE
```

ملف القاموس:

```
drink/X   [VERB]
```

ملف الاختبار :

```
drink
drinkable
```

الاختبار:

```
$ hunmorph test.aff test.dic test.txt
 drink:     drink[VERB]
 drinkable: drink[VERB]+ABLE
```

يمكننا أن نرى في المثال أن المحلل يقوم بتحليل المجالات الفرعية في البند و أسلوب الترتيب.

VII. الاختبار والتصحيح :

الحوار زم الأتلي لايسبل يعرف لاحقة واحدة فقط. المنسبل يمكن أن يعرف لاحقة أخرى حتى الآن. تعريف اللواحق المزدوجة يشكل تحسنا كبيرا في التعامل مع عدد هائل من اللواحق التي تميز اللغات الاشتقاقية. تمديد المثال السابق بإضافة لاحقة ثانية (تنفيذ اللاحقة Y سيكون تنفيذ الاستمرار للاحقة 'able') :

```
SFX Y Y 1
SFX Y 0 s . +PLUR

SFX X Y 1
 SFX X 0 able/Y . +ABLE
```

ملف القاموس:

```
drink/X   [VERB]
```

ملف الاختبار:

```
drink
drinkable
drinkables
```

الاختبار:

```
$ hunmorph test.aff test.dic test.txt
 drink:      drink[VERB]
 drinkable:  drink[VERB]+ABLE
 drinkables: drink[VERB]+ABLE+PLUR
```

نظريا في تجريد اللاحقة المزدوجة نحتاج فقط إلى الجذر التربيعي لعدد قواعد الالحاق بالمقارنة في ما يقوم به المنسبل. من خلال ممارستنا تمكنا من إعداد صرفي نحوي للغة المجرية في تجريد اللاحقة المزدوجة.

ملاحظة : في معالج القواعد 'Hunlex' يمكننا استخدام ليس فقط اللاحقة المزدوجة ولكن اللاحقة المتعددة الجوانب.

VIII. كشاف الملفات الخاصة :

يستطيع المنسبل معالجة أكثر من 65 000 من ثنائي الزوائد. هناك عبارتين جديدتين لإعطاء الأعلام في ملفات الزوائد والقاموس.

التعليمة FLAG long تحدد أعلام 2-حروف :

```
FLAG long
SFX Y1 Y 1
SFX Y1 0 s 1
```

القاموس سجل في y1 Ü Z3 Ü F Þ الأعلام :

```
foo/Y1Z3F?
```

التعليمة FLAG num تحدد الأعلام الرقمية تستعملها فاصلة :

```
FLAG num
SFX 65000 Y 1
SFX 65000 0 s 1
```

مثال القاموس :

```
foo/65000,12,2756
```

IX. الملفات الخاصة :

يمكن أن يحتوي قاموس المنسبل على تكرار العناصر ذات الألفاظ المتجانسة :

```
work/A    [VERB]
work/B    [NOUN]
```

ملف زوائد:

```
SFX A Y 1
 SFX A 0 s . +SG3

SFX B Y 1
 SFX B 0 s . +PLUR
```

ملف الاختبار:

```
works
```

الاختبار:

```
  > works
 work[VERB]+SG3
 work[NOUN]+PLUR
```

هذه الميزة تعطي للباحث طريقة لمنح عمليات السوابق/اللواحق الغير مشروعة في الحالات الصعبة.

x. الزيادة الدائرية – circumfixes:

من أهم الآثار الجانبية لتعدد الخطوات التجريدية هي تلقائية معالجة حالات الزيادة الدائرية 'circumfixes'. فعلى سبيل المثال في اللغة المجرية أغلب الدرجات تشكلها بإضافة السابقة 'leg' متزامنة مع إضافة اللاحقة 'bb' إلى قاعدة الصفة.

المشكلة مع بنية المستوى الواحد هي انه ليس هناك طريقة لوصف تركيب f معجمي بالسوابق واللواحق المترابطة وبالتالي لإشكال غير صحيحة تعرف على أنها صحيحة بمعنى '*legv?fR = leg + v?fR' 'قديم'. الآن عملية إدخال مجموعات معاملة الصفة لدرجات الصفة قد تكون جد مثيرة برنامج المنسبل السابق.

هذا قد يكون مشروعا لحالة واحدة ولكن في الواقع الزيادة الترابط سابقة -- لاحقه موجودة في كل مكان في فئة أنماط التغير الاشتقاقي (راجع الانكليزية payable ، non-payable لكن *non-pay و drinkable ، undrinkable لكن *undrink). و بعبارة بسيطة هناك السابقة 'un' مشروعة فقط في حالة ما إذا كانت القاعدة 'drink' متبوعة باللاحقة 'able'. إذا كانت كل من هذه الأنماط متعامل بها في قواعد الزيادة على الانترنت و كانت قواعد الزيادة تدقق عن طريق القاعدة فقط وليس هناك طريقة للتعبير عن هذه التبعية والنظام سوف يقوم بالضرورة بتنفيذ العملية أو إعادة الانشاء 'undergenerate'.

في المثال التالي تنفذ اللاحقة 'R' على السابقة تنفذ 'استمرار' ('P' تنفذ).

```
PFX P Y 1
PFX P   0 un . [prefix_un]+

SFX S Y 1
SFX S   0 s . +PL

SFX Q Y 1
SFX Q   0 s . +3SGV
```

```
SFX R Y 1
SFX R   0 able/PS . +DER_V_ADJ_ABLE
```

القاموس:

```
2

 drink/RQ
  [verb]

 drink/S
  [noun]
```

التحليل الصرفي:

```
> drink
drink[verb]
drink[noun]
> drinks
drink[verb]+3SGV
drink[noun]+PL
> drinkable
drink[verb]+DER_V_ADJ_ABLE
> drinkables
drink[verb]+DER_V_ADJ_ABLE+PL
> undrinkable
[prefix_un]+drink[verb]+DER_V_ADJ_ABLE
> undrinkables
[prefix_un]+drink[verb]+DER_V_ADJ_ABLE+PL
> undrink
Unknown word.
> undrinks
Unknown word.
```

xi. الزوائد المشروطة:

الزوائد المشروطة الموسوعة عن طريق استمرار صيغة ليست كافية للزيادة الدائرية لأن الزيادة الدائرية تعني صرف زائدة واحدة. نحن بحاجة ليسا لمزايا الزيادة الدائرية من أجل تحليل صرفي صحيح.

```
# circumfixes: ~ obligate prefix/suffix combinations
# superlative in Hungarian: leg- (prefix) AND -bb (suffix)
# nagy, nagyobb, legnagyobb, legeslegnagyobb
# (great, greater, greatest, most greatest)

CIRCUMFIX X

PFX A Y 1
PFX A 0 leg/X .

PFX B Y 1
```

```
PFX B 0 legesleg/X .

SFX C Y 3
SFX C 0 obb . +COMPARATIVE
SFX C 0 obb/AX . +SUPERLATIVE
SFX C 0 obb/BX . +SUPERSUPERLATIVE
```

الملف 2 :

```
1
nagy/C
[MN]
```

الناتج:

```
> nagy
nagy[MN]
> nagyobb
nagy[MN]+COMPARATIVE
> legnagyobb
nagy[MN]+SUPERLATIVE
> legeslegnagyobb
nagy[MN]+SUPERSUPERLATIVE
```

XII. الإشتقاق (Compounds):

السماح بحرية نتائج التركيب يقلل من دقة التعرف لتحديد جذور الكلمات والتحليل الصرفي. رغم أن التبديلات اللغوية تم وضعها في تركيب قواعد عن طريق تسلسل وإذا يتم مقيدا بما فيه الكفاية. على سبيل المثال :

```
# affix file
COMPOUNDFLAG X
 2
foo/X
bar/X
```

فإذا البرنامج "foobar" و "barfoo" هي كلمات مقبولة.

إذا قد تحسن كثيرا فإن إدخال التركيب المراعي للاتجاهات أي أن الملامح المعجمية يمكن أن تحدد بشكل مستقل ما إذا كان يمكن لقاعدة أن تنطبق أو اليسار أو اليمين التقسيمي في المركبات. وفي ذلك لا تزال غير كافية لمعالجة أنماط معقدة من المركبات يتم تحديد (ولغة معينة) لقواعد الشرط.

البرمجيات المنسبل يسمح حاليا بأي شكل من أشكال الكلمات المزيدة والتي هي مسجلة معجميا على أنها أصول محتملة في المركبات. المنسبل حسن إذا و قواعدها للتحقق المتكرر للمركبات تجعل من الممكن تنفيذ الاشتقاقات الإملائية المعقدة في المركبات المجرية. فعلى سبيل المثال استخدام خيارات COMPOUNDROOT COMPOUNDSYLLABLE COMPOUNDWORDMAX SYLLABLENUM يمكن أن تحدد بشكل واضح القاعدة المجرية '6-- 3'. مثال آخر في اللغة المجرية

اشتقاق الواحد كثيرا ما يغير طائفة المركبة. هنالك سبل يسمى بعلامة التركيب على الواحدة وهنالك علمين طائفين (COMPOUNDPERMITFLAG وCOMPOUNDFORBIDFLAG) بسمات أو من تركيب الاشتقاق.

الواحد من هذا العلم لا تسمى بتركيب الكلمات المزيدة.

نحتاج إلى عدة طائفة لهانسبل لمعالجة التركيب في اللغة الألمانية :

التركيب الألماني

تحديد لغة لمعالجة حالة طائفة من الحرف الألماني الحاد 's'

```
LANG de_DE
```

علامة المركبة

```
COMPOUNDBEGIN U
COMPOUNDMIDDLE V
COMPOUNDEND W
```

السوابق مسموح بها في بداية المركبة.

الواحد مسموح بها في نهاية المركبة بشكل افتراضي :

(سابقة)+ (جذر) + (لاحقة)

اللاحقة من COMPOUNDPERMITFLAG قد تكون داخل المركبة.

```
COMPOUNDPERMITFLAG P
```

من اجل المقاطع في اللغة الألمانية (Fuge-element)

تلميح : ONLYINCOMPOUND ليست مطلوبة في كل مكان لكن

التحقق سيكون أسرع قليلا معها.

```
ONLYINCOMPOUND X
```

تمنع الحروف الكبيرة الشكل عند حدود الكلمة المركبة.

```
CHECKCOMPOUNDCASE
```

لمعالجة العناصر من الواجبات 'Arbeits-'

الواجبة سوف تتبع كلمة طائفة.

```
COMPOUNDMIN 1
WORDCHARS -
```

خيارات التركيب و المقاطع بالنسبة ل 'Arbeit'

```
SFX A Y 3
SFX A 0 s/UPX .
SFX A 0 s/VPDX .
SFX A 0 0/WXD .

SFX B Y 2
SFX B 0 0/UPX .
SFX B 0 0/VWXDP .
```

لاحقة لكلمة 'Computer'

```
SFX C Y 1
```

```
SFX C 0 n/WD .
```

لمنع الاستثناء (*Arbeitsnehmer)

```
FORBIDDENWORD Z
```

سابقة الواثبة للمركبة التي بها واثبة (Arbeits-Computer)

```
PFX - Y 1
PFX - 0 -/P .
```

سابقة decapitalizing

circumfix للوع في المركبة

```
PFX D Y 29
PFX D A a/PX A
PFX D ?X  .
 .
PFX D Y y/PX Y
PFX D Z z/PX Z
```

مثال (قاموس):

```
4
Arbeit/A-
Computer/BC-
-/W
Arbeitsnehmer/Z
```

المركبة المقبولة للكلمة المركبة من البرنامج السابق هي:

```
Computer
Computern
Arbeit
Arbeits-
Computerarbeit
Computerarbeits-
Arbeitscomputer
Arbeitscomputern
Computerarbeitscomputer
Computerarbeitscomputern
Arbeitscomputerarbeit
Computerarbeits-Computer
Computerarbeits-Computern
```

المركبة الغير مقبولة:

```
computer
arbeit
Arbeits
arbeits
ComputerArbeit
ComputerArbeits
Arbeitcomputer
```

ArbeitsComputer
Computerarbeitcomputer
ComputerArbeitcomputer
ComputerArbeitscomputer
Arbeitscomputerarbeits
Computerarbeits-computer
Arbeitsnehmer

هذا الحل لا يزال غير مثالي و سوف يعود مستند نمط الوارزم فحص المركبات والمتكامل مباشرة في ذاكرة الادخال المدققة للترميز 'Tokenization'. الأنماط التي تحلل المركبات تقتضي علاً لنها مصدر إدخال منخفض والتي قد تعود إلى المستوى العالي لخلية الأجزاء المكونة (كزيادة عدد المقاطع أو أعلام الزوائد ومحتوى الوحدات). الأنماط تقترن بالتقسيمات المحتملة للمركبات لتقييم أحسن التكوينات.

XIII. استخدام الهوامش:

طبيعه المشاكل في تشفير 8 بت.

كل من 'Ispell' و'Myspell' يستعمل تشفير الحروف 8 بت لسكي 'ASCII' أو وهو نقص كبير عندما يتعلق الأمر بالسهولة. ورغم أن اللغة المجرية لها مجموعة حروف لسكي قياسيه (ISO 2-8859) أ فإنه لا يسمح بالتنفيذ الكامل للاتفاقيات الأملائية المجرية. فعلاً سبيل المثال الرمز '–' (ن – وأنبه) مفقود من مجموعة الحروف هذا لا يعني حقيقة انه ليس فقط الرمز الرسمي لتحديد عبارات اعتراضية في اللغة ولكن يمكن أن تكون في الكلمات المركبة علاً شكل واصلة أصلة كبيرة.

Myspell حصل علاً بعض جداول التشفير 8 بت ولكن هناك لغات بدون التشفير القياسي 8 بت أيضا. علاً سبيل المثال أ الكثير من اللغات الأفريقية لها حروف غير لاتينية أو تمديدا لحروف لاتينية.

ونفس الشيء باستعمال التهجي الأصلي لبعض الأسماء الأجنبية مثل 'gstrR' أو ' ? Moli' أ تشجعه القاعدة الأملائية المجرية وحتى الحروف '' و '' ? ليست جزءا من 'ISO 8859-2' أ عندما تتوحد في الأصوات التي تحتوي علاً الحرف الا في 'ISO 8859-2' (مثلا ro=l حادة مزدوجة) أ هذا النتيجة في كلمات (مثل gstr=l أو Moli ?-to=l) والتي لا يمكن تشفيرها باستخدام أي مخطط تشفير لسكي موحد.

المشاكل تكمن في كون الترميز ASCII 8 بت معروف منذ فترة طويلة من طرف مقترحي نظام اليونيكود. الشيء المدسوس هو كون التبديل إلى نظام اليونيكود UTF-16 يحتاج إلى مقدار ضخم من التحسين لهذا النظام وهو ما قد يؤدي إلى التأثير علاً فعالية الوارزم.

الوارزم DÖMÖLKI المستعمل في تدقيق شروط إضافة الزوائد يستعمل جدول حروف ' Byte 256 ' و الذي سوف يكبر حجمه حتى يصل إلى ' Ko 64 ' في نظام الترميز يونيكود. حتى عملية إضافة الزوائد علاً الانترنيت بالنسبة للغة غنية اشتقاقية يمكن أن يكون لها وبكل بساطة بعض المئات من الجداول المماثلة (في حالة المصادر المجرية القياسية نستعمل هذا العدد 'is ca'. 300 أو أكثر حتى التخزين المكرر

لنماذج زوائد ميكليا متطابقة يحسن الكتابة) وعملية التحول إلى اليونيكود سوف تتحمل ارتفاع تكاليف الموارد.

ومع ذلك فمن الواضح أن تجارة الكتابة لاستقلالية الترميز لها مزاياها عندما يتعلق الأمر حقا ببرنامج متعدد اللغات ولذا كان من بين ما خططنا له لفترة طويلة هو تمديد البنية في هذا الاتجاه.

خلاصة:

مررنا جربنا بنجاح ذاكرة ومعالجة كتابة يونيكود معا. في ترميزها الحروف المختلفة عن نظام UTF-8 لا يعمل متسلسل مع الخوارزمية الأصلية ذات الـ 8 بت ولكن مع الفاموس ومع الزوائد المشتركة بواسطة نظام UTF-8 يستخدم متسلسل سلسلة مركبة للتلاعب وشرط التحقق لدعم اليونيكود: الزوائد والكلمات مخزنه في نظام UTF-8 أثناء التحليل يتم التعامل في معظم الحالات مع UTF-8 في حالة فحص واقتراح تحول إلى UTF-16).

خوارزم DÖMÖLKI تستخدم لتخزين وفحص حروف الشرط 7 بت لسكي (ISO 646) حرفا شرطا وتخزن قوائم UTF-16 من أجل حروف يونيكود أخرى من نماذج الشرط.

متسلسل دعم فقط الحروف الـ '65536' الأولى (مخطط قاعدي متعدد اللغات) لليونيكود القياسي لحد الآن.