حمداني عبد الفتاح، اليزيدي توفيق، الحسني سعيد
معهد الدراسات والأبحاث للتعريب، الرباط

التحليل الصرفي للأسماء العربية

Résumé

Nous voulons présenter dans ce papier un modèle d'analyseur morphologique des noms Arabe se basant sur une approche linguistique et une base de données manuelle et alimentée par un ensemble d'éléments linguistiques morphologiques, lexicales et syntaxiques, et ce en proposant une nouvelle classification des noms. Nous proposons aussi une structure de base de données permettant d'obtenir des résultats précis à partir d'algorithmes génériques.

مقدمة[1]

تتزايد الحاجة يوما بعد يوم للمعالجة الآلية للغة العربية ليتزايد معها حجم الدراسات والأبحاث والمقاربات التي أنجزت وتنجز في هذا الإطار. ويعد التحليل الصرفي جزءا من هذه المعالجة قام العديد من الباحثين العرب وغير العرب، نذكر من بينهم كلا من باكوالتر ولاركي وخوجة ودرويش وغيرهم كثير، بتطويره وإغنائه باقتراحهم لمجموعة من الأنظمة التي تتباين فيما بينها من حيث الشكل أو المضمون أو هما معا. إلا أن الأسماء لم تحض بعناية كبيرة في مجال التحليل الصرفي مقارنة بالعناية الفائقة التي حضيت بها الأفعال. لذلك نود في هذه الورقة أن نعرض تصورا لسانيا يعنى بوضع قاعدة معطيات لغوية يدوية للأسماء العربية مزودة بزمرة من المعلومات الصرفية والمعجمية والتركيبية الضرورية، وتقديم تصنيف جديد لها مقرون بعدد لا يستهان به من القضايا العالقة، واقتراح تنظيم معين للقاعدة يتسم بالكفاية ويأخذ بعين الاعتبار عمليات الحذف والإبدال التي تنطبق على أسماء دون غيرها.

[1] فريق بحث معتمد من طرف الجامعة تحت إشراف الأستاذ عبد الفتاح حمداني، يعمل حاليا على إنجاز محلل صرفي للنصوص العربية، ويضم الأساتذة: الجيهاد، ع.، الحسني، س.، الخطابي، إ. اليزيدي، ت.، أوراغ، ح.، بوزبع، ك.، يوسفي، ع.

1. الأسماء العربية: ملاحظات عامة وقضايا مطروحة

للأسماء العربية خصوصيات كثيرة يتعلق البعض منها بخصائص اللغة العربية في حد ذاتها، بينما يتعلق البعض الآخر بالطبيعة المركبة للمقولات الاسمية. فهذه اللغة، كما هو معلوم، تكتب دون حركات قصيرة أثناء الكتابة، مما يصعب معالجتها آليا ويجعل الاعتماد على قاعدة معطيات غير تامة وغير دقيقة أمرا صعبا للغاية، ولا يسعفنا في تفادي الوقوع في الخطأ والإلمام بجميع النتائج المحتملة. فكلمة نحو *كتاب* ملتبسة في وضعها هذا بين الكلمات المشكولة الآتية: *كِتاب* و*كُتّاب* و*كُتّاب*.[2] الشيء نفسه يصدق كذلك على كلمة *معلمة* التي يمكن أن تكون *مُعَلِّمَة* أو *مَعْلَمَة* وهلم جرا. هذه التفاصيل ذات أهمية قصوى لكونها تبرز الإمكانات الحركية المتاحة والمعاني المختلفة التي يمكن أن تنتج عن هذه التقاليب الحركية التي لا تعكسها الكتابة. زد على هذا أن بعض الحروف الأصلية في الكلمة تكون مماثلة للواصق في حالات من قبيل *ليمون* أو *مكان* أو *ولد* أو *بلد* الخ. وهناك حالات أخرى يبلغ فيها الالتباس أشده، ومثال ذلك كلمة من قبيل *برد* التي تحتمل أن تكون *بَرْد* أو *بِرَدّ* باعتبار أن الباء سابقة (حرف جر) و*رَدٌّ* كلمة مستقلة (بمعنى جواب).

وتتخذ بعض الأسماء العربية الأخرى صورا (سطحية) مختلفة متأثرة في ذلك إما بطبيعة الحركة الإعرابية الواردة، أو بطبيعة العناصر الملحقة بها. فالحرف الأخير من الاسم المهموز يكتب بأشكال متعددة بحسب محله من الإعراب: ماء أو مائه (في حالة الجر) أو ماؤه (في حالة الرفع)، كما أن نفس الحرف في الاسم المقصور يتحول إلى ألف (ا) أو ياء نحو: *مرمى* التي تصير *مرماه* أو *مرمياه* عندما

[2] كلمة *كُتّاب* يمكن أن تكون جمعا ل *كاتِب*، ويمكن أن تعني مكان تعليم الصغار وجمعها *كتاتيب*.

يتم إقحام اللواحق، أو عندما تكون مصرفة في المثنى، في حين يحذف الحرف الأخير من الاسم المنقوص نحو: *محامي-محام-محامون*. ومن ناحية أخرى تصير التاء المربوطة مبسوطة عندما تتلوها لاحقة، وتحذف في حالة جمع التكسير مثل: *ورقة-ورقته-ورقات*، كما أن اللاحقة ***لل*** لا تعبر عن أصلها الذي هو ***ل+ال*** في الحالات العادية، إذ عندما تكون اللام الثانية أصلية، كما في كلمة من قبيل للغة، تقرأ بكيفيتين: لِلُّغَة أو لِلُغَة. اللام الأولى في الكلمة الأولى هي في الأصل ***ل+ال***، وفي الكلمة الثانية تبقى على حالها.

الملاحظ أن بعض اللواصق الاسمية في اللغة العربية تتشابه من حيث الشكل وتختلف من حيث الوظيفة. فالواو يمكن أن يكون للعطف أو للابتداء أو للحال أو للمعية، كما أن الياء يمكن أن تكون للنسبة أو ضمير ملكية. أما الألف (ا) في آخر الكلمة فهو إما يمثل لاحقة العدد المثنى في حالة الإضافة (وبالتالي يمكن أن يكون متلوا بلاحقة)، أو يمثل الألف الحامل للتنوين الذي يجب أن يكون في نهاية الكلمة. التاء المربوطة بدورها يمكن أن تكون صنيفة أو علامة دالة على التأنيث أو غير دالة عليه، ولاحقة العدد المثني وجمع المذكر السالم تفقد النون في حالة الإضافة.

2. التصنيف الاسمي المعتمد: الموارد والأسس

هناك عدة تصنيفات لغوية وغير لغوية للأسماء العربية مستمدة من التصنيفات التقليدية أو الحديثة على اختلاف أنواعها ومصادرها وخلفياتها النظرية، لكن لا ينبغي أن ننسى أننا بصدد المعالجة الآلية للغة العربية، وأن أي تصنيف عليه أن يراعي هذه الخصوصية. أصناف الأسماء التي اعتمدناها في القاعدة محصورة فيما

يلي: اسم الجنس، واسم العلم، واسم الآلة، واسم المكان، واسم الزمان، والتصغير، واسم الحدث، واسم الفاعل، واسم المفعول، وصيغة المبالغة.

الأصناف الستة الأولى هي أسماء ذوات باستثناء اسم الزمان،[3] وهي تتقاسم الخصائص الآتية: أ) تقبل لاحقتي المثنى وجمع المؤنث السالم، ب) تجمع جمع تكسير فقط، ج) التاء المربوطة ليست لاحقة، ولا تنفصل عن الكلمة، ولا تعبر عن التأنيث بالضرورة، د) تختفي التاء المربوطة من الكلمة عند ما تتلوها لاحقة أو عندما تكون مجموعة جمع مؤنث سالم، ه) تقبل النسبة. وفي مقابل ذلك، يبدي اسم العلم سلوكا مخالفا من حيث أنه غالبا ما يرد في صورة صرفية واحدة قلما تتغير، لكنه لا يخرج عن إطار الخصائص السابقة الذكر شأنه في ذلك شأن اسم الحدث الذي لا يدل على الذوات، والذي لا يقبل أحيانا أن يثنى أو يجمع. وتختلف الأصناف المذكورة عن الثلاثة الأخيرة في كونها تملك الخصائص ذاتها: أ) التاء المربوطة لاحقة للتأنيث دائما، ب) النسبة غير ممكنة، ج) جمع المذكر السالم ممكن.

بناءا على ما سبق، قمنا بتجميع بضعة أصناف اسمية في طبقات كلما تشابهت في قبولها لنفس العدد من اللواحق، فكان أساس هذا التقسيم مستمدا من كون الأصناف الاسمية الموجودة في اللغة العربية تنتمي بالضرورة لإحدى الطبقات الثلاث الرئيسية الآتية:

[3] لقد أدرجنا اسم الزمان تجاوزا ضمن الصنف الخاص باسم الذات، على أن نقوم بفصله عنه مستقبلا.

•الطبقة الأولى، وهي التي لا تقبل لاحقة التأنيث ***ة*** ولاحقة الجمع المذكر السالم ***ون/ين***، تتضمن الأصناف الاسمية التالية (وهي تشكل نسبة كبيرة من قاموس الأسماء): اسم الذات/الحدث/العلم/جمع التكسير.

•الطبقة الثانية، وهي التي لا تقبل لاحقة الجمع المذكر السالم ***ون/ين***، ولاحقة المثنى ***ان/ين***، تتضمن اسم النوع فقط، وهو إما اسم ذات أو حدث.

•الطبقة الثالثة، وهي الطبقة التي لا تقبل لاحقة النسبة، تتضمن الأصناف الاسمية التالية: اسم الفاعل/المفعول/صيغة المبالغة.

لقد أدرجنا جمع التكسير إلى جانب اسم الذات والحدث والعلم في الطبقة الأولى لكونه لا يخضع لقاعدة مطردة (في حالة الثلاثي)، ولكونه يتقيد بخصائص الطبقة التي ينتمي إليها. هناك بعض الأصناف الاسمية التي وردت في طبقتين، ونخص بالذكر هنا اسمي الذات والحدث الواردين في الطبقتين الأولى والثانية، ويرجع ذلك إلى أن بعض الأسماء المدرجة ضمن هذين الصنفين تقبل التاء المربوطة التي تكون صنيفة الوحدة، وهي لاحقة كما في قولنا: *شجر-شجرة*، *رقص-رقصة*. هذه الأسماء ينبغي أن تجمع في طبقة خاصة لا تتضمن اسم الوحدة.

3. الموارد اللسانسة لقاعدة المعطيات اليدوية

1.3. القواميس

هناك ثلاثة قواميس يتضمن كل واحد منها عنصرا من العناصر الثلاثة المكونة للكلمة السطحية (أو الكلمة الدخل): السوابق واللواحق والجذوع. فقاموسا السوابق واللواحق يحتويان على كل اللواصق الاسمية البسيطة والمركبة الموجود في اللغة العربية، وهي مشكولة وغير مشكولة، مفككة وغير مفككة ومصحوبة بمصفوفة من المعلومات المتعلقة بها. ويقوم قاموس الأسماء على تقسيم ثلاثي الطبقات تندرج ضمنه كل المركبات الاسمية التي يمكن العثور عليها في اللغة العربية، وهو مرتب ألفبائيا على أساس الجذر (الصامتي) الذي يحتضن قائمة من الكلمات المصصفة أسفله حسب الترتيب الألفبائي أيضا. هذه الكلمات مشكولة وغير مشكولة ومقرونة هي الأخرى بالمعلومات المتعلقة بالطبقة والصنف والصنف الفرعي، فضلا عن سمتي الجنس والعدد.

وضمن الأصناف الاسمية المعتمدة نجد أسماء ذات طبيعة خاصة لكونها تظهر في أشكال مختلفة كما أشرنا إلى ذلك من قبل، يتعلق الأمر هنا بالمهموز والمنقوص والمقصور والأسماء الخمسة التي تشكل ما نسميه بالأصناف الفرعية.[4]

2.3. الجنس والعدد

يوسم الجنس المذكر على مستوى القاعدة بالنسبة لجميع الطبقات باستثناء الحالات التي تكون فيها الكلمة السطحية مصرفة في المثنى المذكر أو جمع المذكر

[4] نظرا لقلة الأسماء الخمسة في اللغة العربية سنزود قاموس الأسماء بكل تمظهراتها الممكنة.

السالم، في حين أن الجنس المؤنث يوسم على مستوى القاعدة بالنسبة للطبقة الأولى وعلى مستوى اللاحقة بالنسبة للطبقتين الثانية والثالثة في المفرد والمثنى المؤنث وجمع المؤنث السالم.[5] وفي ما يخص العدد نشير إلى أن جميع الكلمات موسومة بالعدد المفرد على مستوى القاعدة، ما عدا تلك المدرجة في الصنف الرابع، يعني جموع التكسير التي تكون موسومة على مستوى القاعدة دائما على أنها جمع. وعندما تكون الكلمة السطحية مصرفة في المثنى والجمع السالم بنوعيهما، تستخلص سمة العدد المناسبة من نفس السمة الموجودة في اللاحقة (وهي سمة متضمنة في قاموس اللواحق).[6]

لدينا، إذن، ثلاث طبقات اسمية، وضمن كل طبقة هناك أصناف وأصناف فرعية موزعة على النحو الآتي: (بالنسبة للصنف: 1= اسم الذات، 2= اسم الحدث، 3= اسم العلم، 4= جمع التكسير، 5= اسم الفاعل، 6= اسم المفعول، 7= صيغة المبالغة. بالنسبة للصنف الفرعي: 1= المقصور، 2= المنقوص، 3= المهموز. بالنسبة للجنس: 1= المذكر 2= المؤنث. بالنسبة للعدد: 1= المفرد، 2= جمع التكسير.)

نموذج من قاموس الكلمات المعتمد

الجذر	الكلمة	الكلمة المشكولة	الطبقة	الصنف	الصنف الفرعي	العدد	الجنس

[5] فكلمة من قبيل *مدرسة* أو *دار* موسومة في القاعدة على أنها مؤنث، بيد أن سمة الجنس المؤنث في كلمة نحو معلمة أو شجرة أو معلمتان أو شجرات مستخلصة من نفس السمة في اللاحقة الواردة.

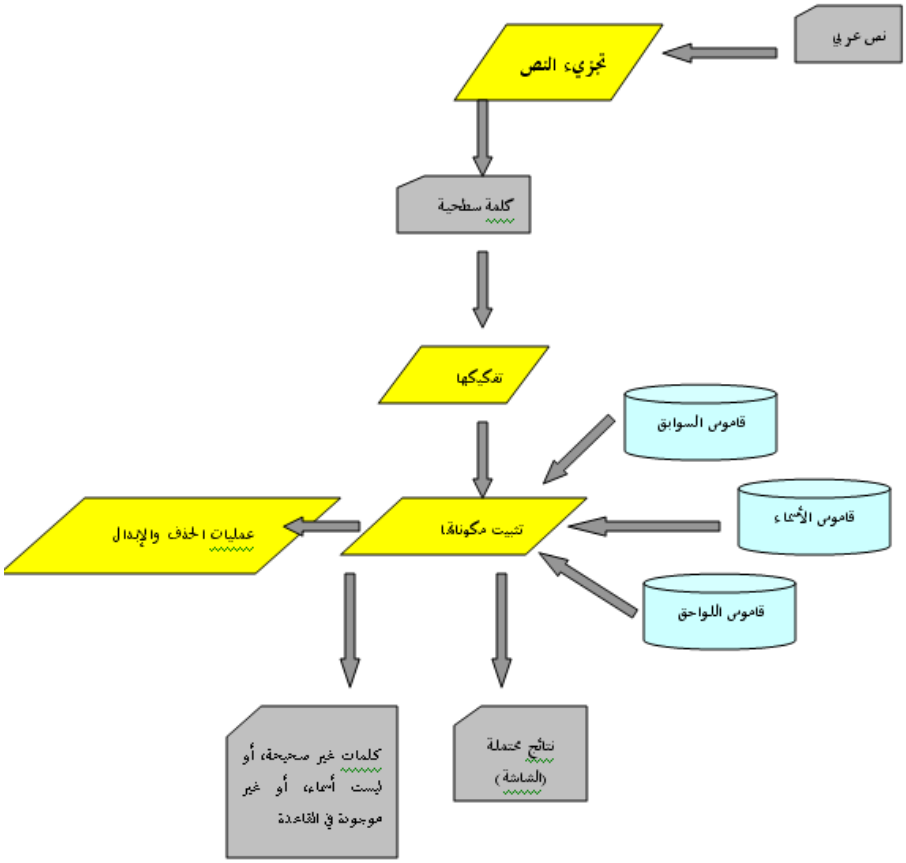
[6] الهدف هو تقديم أكبر عدد ممكن من المعلومات حول عناصر الكلمة الدخل المفككة، والاستغلال المستقبلي لهذه المعلومات في إنجاز محللات تركيبية أو دلالية أو في وضع مترجمات آلية الخ.

ك ت ب							
	كتب	كَتْب	1	2		1	1
	كتب	كُتُب	1	4		2	2
	كتاب	كِتاب	1	1		1	1
	كتاب	كُتّاب	1	4		2	1
	كتاب	كُتّاب	1	1		1	1
	كتابة	كِتابَة	1	2		1	2
	كتيب	كُتَيِّب	1	1		1	1
	مكتب	مَكْتَب	1	1		1	1
	مكتبة	مَكْتَبَة	1	1		1	2
	اكتتاب	اكْتِتاب	1	2		1	1
	مكاتب	مَكاتِب	1	4		2	2
	كاتب	كاتِب	3	5		1	1
	مكتوب	مَكْتوب	3	6		1	1
	مكاتبة	مُكاتَبَة	1	2		1	2
	كتبة	كِتْبَة	1	2		1	2
ك ت ح							
	كتح	كَتْح	1	2		1	1

4. كيفية اشتغال النظام

يقوم هذا المحلل في بداية الأمر بتجزيء النص العربي (المشكول أو الغير المشكول) إلى كلمات سطحية حاملة لكل العناصر الملصقة بها ويتعامل معها

الواحدة تلو الأخرى، إذ يفككها إلى عناصرها الأساسية ويثبت مكوناتها المفككة بالاعتماد على القواميس الثلاثة: قاموس الأسماء، وقاموس السوابق، وقاموس اللواحق، وعلى عمليات الحذف والقلب التي تلعب دورا مركزيا في التعرف على الكلمة وتسويغ النتائج المحتملة. ويقدم المحلل في نهاية المطاف النتائج المحتملة التي يتم العثور عليها. أما إذا تعذر عليه ذلك، فذلك يعني أن هذه الكلمة السطحية غير صحيحة أو أنها ليست اسما (وإنما هي فعل أو حرف أو ما شابه ذلك)، أو أنها غير موجود في قاعدة المعطيات اللغوية (وفي هذه الحالة ينبغي ضمه إليها). وعليه، فالهيكلة الداخلية لهذا المحلل هي الواردة في الخطاطة الآتية:



يتم البحث عن كل كلمة سطحية في قاموس الأسماء أولا قبل أن تفكك بالنظر إلى أحد القواميس المعتمدة، ثم تفكك على أساس قاموس السوابق (لنحصل على: سابقة/كلمة محتملة)، وتتكرر هذه العملية حتى تشمل جميع السوابق الواردة، وفي كل مرة يتم البحث من جديد عن ما تبقى من الكلمة في قاموس الأسماء. بعد ذلك تنطبق قاعدة التفكيك مرة أخرى، ولكن على أساس قاموس اللواحق هذه المرة (لنحصل على: كلمة محتملة/لاحقة) فيتم اللجوء إلى قاموس الأسماء للمرة الثالثة، وبشكل متكرر، للبحث عن ما تبقى بعد إزالة اللواحق بكيفية تدريجية على الشاكلة

المشار إليها قبل قليل. وعند القيام بعملية التفكيك الثالثة والأخيرة سيتجرد الجذع من كل اللواصق العالقة به باعتماد القواميس الثلاثة (لنحصل على: سابقة/كلمة محتملة/لاحقة). في هذه المرحلة يتم احتساب كل التأليفات الممكنة بين العناصر الثلاثة المكونة للكلمة السطحية بالعودة المتكررة لقاموس الأسماء للتحقق من وجودها فيه.

5. عمليات القلب والحذف

لمعالجة جملة من التغيرات التي تطرأ على الكلمات السطحية نقترح هذه العمليات التي من شأنها أن تسعفنا في الإلمام بجل السيناريوهات الممكنة:

1) في حالة وجود ا (الألف) بعدها ثلاثة حروف على الأكثر تقلب ى (الألف المقصورة) شريطة أن يكون ما بعدها موجودا في قاموس اللواحق، مما يعني أن البحث سينصب على الكلمة *مرمى* بدلا من *مرما* الواردة في كلمة سطحية من قبيل *مرماهم*. أما في حالة وجود ياء بعدها ألف بمفرده أو متلو بحروف أخرى (أقصاها ثلاثة)، يحذف ما بعد الياء وتقلب ألفا مقصورة، وينصب البحث عن هذه الكلمة في قاموس الكلمات (وتحديدا في الصنف الفرعي1 المتعلق بالمقصور). فإذا ما وجدت، تكون عملية القلب صحيحة ويتم التحقق من أن ما بعد الياء هو بالفعل لاحقة استنادا إلى قاموس اللواحق، وإذا لم يكن موجودا، تلغى عملية القلب مباشرة.

2) بالنسبة للمنقوص (*كمحام ومحامون*)، يتم البحث للوهلة الأولى، وكالعادة، في قاموس الأسماء الذي لا يتوفر على ما يبدو على كلمات من هذا القبيل. عندئذ تتم إضافة ي إلى هذه الكلمة بعد إزالة اللاحقة (إذا كانت موجودة) وقبل البحث عنه في الصنف الفرعي2 الخاص بالمنقوص. وإذا ما تم العثور على الكلمة في قاموس الأسماء منذ الوهلة الأولى، فلن تضاف الياء إلى الجذع مطلقا.

3) تبديل الشكل **ؤ** والشكل **ـئـ** بالشكل الأصلي **ء** والبحث عن الكلمة في الصنف الفرعي3 المتعلق بالمهموز إذا كانا متلوين بحروف موجودة في قاموس اللواحق، وإلا فهي أصلية ولا ينبغي تبديلها.

4) في حالة الكلمات التي تبدأ بـ **ال**، تحذف اللام الأولى ويتم البحث عن البقية في قاموس الأسماء، فإذا وجد فهذا يعني أن هذه اللام يمكن أن تكون **ل** بمفردها أو **لل** (التي تفكك استنادا إلى قاموس السوابق على النحو الآتي: **ل/ال**)، وأن اللام الثانية هي أصلية. أما إذا تعذر وجودها، تحذف **لل** ويتم التحقق من وجود الباقي في قاموس الأسماء.

5) كل كلمة تنتهي بالتاء المبسوطة **ت** بعد تجريدها من جميع اللواصق يتم البحث عنها في بداية الأمر في قاموس الأسماء، وسواء وجدت أو لم توجد تحول **ت** إلى **ة**، مما يتطلب القيام ببحث جديد عن هذه الكلمة الجديدة في قاموس الأسماء. وفي حالة جمع المؤنث السالم الذي ينصب على الأسماء التي تنتمي للطبقتين الأولى والثانية، يتم حذف لاحقة العدد المؤنث من الكلمة وإضافة تاء مربوطة لما تبقى من الكلمة قبل البحث عنها في قاموس الأسماء.

6) تكون عملية التفكيك مشروعة فقط إذا كان ما تبقى من الكلمة بعد إزالة اللواصق لا يقل عن حرفين.

7) أية كلمة تنتهي بالياء ينطبق عليها ما يلي: إذا كانت موجودة في الصنف الفرعي2، فإن هذه الياء أصلية ولا يمكن أن تكون لاحقة.

خاتمة

أثناء المعالجة الآلية للغة العربية لا بد من مراعاة الدفة وتوخي الحذر لكثرة تمظهراتها الناتجة بالدرجة الأولى عن كونها غير مرفقة بالحركات القصيرة. وقد قدمنا في هذه الورقة ما نعتقد أنه أساس لساني متين مكننا من التوصل إلى تنظيم معقول للعناصر المدرجة في قاعدة المعطيات اليدوية لبلوغ الكفاية والشمولية والإحاطة بكل الاحتمالات الممكنة. فكان النظام المقترح أدنويا من ناحيتين:

أ) اعتماد قاموس مصنف للأسماء لا يتضمن الكلمات السطحية التي يمكن اشتقاقها بإضافة لواصق إليها، ب) اعتماد النظام على إجراءين تكراريين مرتبطين بعدد محدد من عمليات الحذف والإبدال، هما: التفكيك والعودة للقواميس (قاموس الأسماء على وجه الخصوص).

مراجع

Buckwalter, T. Qamus: Arabic lexicography. http/www.qamus.org/

Darwish, K.: 2002, Building a shallow morphological analyser in one day. ACL 2002 Workshop on computational Approaches to Semitic languages.

Diab, M.: 2004, Arabic SVM Tools. http://www.stanford.edu/~mdiab/software/arabicSVMTools.tar.gz

Khoja, S. and Garside, R.: 1999, Stemming Arabic test. Computing Department, Lancaster University, Lancaster. http://www.comp.lans.ac.uk/computing/users/khoja/stemmer.ps

Larkey, L. S. and Connell, M. E.: 2001, Arabic information retrieval at UMass in TREC-10. in TREC 2001.Gaithersburg: NIST.

LDC, Linguistic Data Consortium. Buckwalter Morphological Analyser. Version **1.0**, LDC2002L49, 2002.