توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية

مجدي صوالحه و إيرك أتول

جامعة ليدز – المملكة المتحدة

sawalha@comp.leeds.ac.uk, eric@comp.leeds.ac.uk

ملخص البحث

تقوم المحللات النحوية (Part-of-Speech taggers) بشكلٍ عام على تعيين التحليل الصرفي والنحوي للكلمة عن طريق البحث في القواميس المرفقة التي تحتوي على جميع التحليلات الصرفية المحتملة للكلمة، حيث تقوم المحللات النحوية بالرجوع الى سياق النصّ لاختيار التحليل الصرفي المناسب، إن هدفنا هو بناء ذخيرة لغوية معنونة (tagged corpus) بالتحليل الصرفي والنحوي لجميع كلماتها، إن تقييم المحللات الصرفية التي تم تطويرها مسبقاً سلّط الضوء على بعض القصور فيها، حيث أن ربع كلمات الذخيرة اللغوية لم يتم تحليلها بالشكل المطلوب (Sawalha & Atwell, 2008). إن عملية تعيين العناوين الصرفية والنحوية للذخائر اللغوية للعربية ليست بالأمر السهل، قد تم تطوير بعض المحللات الصرفية لاستخراج الجذر أو الجذع للكلمة العربية، ولكن هذا غير كافٍ لعمل المحللات النحوية للغة العربية بالشكل المراد تحقيقه.

تتركب الكلمة العربية من خمسة أجزاء : زوائد في بداية الكلمة ثم سوابق الكلمة ثم الجذر أو الجذع ثم لواحق الكلمة ثم زوائد في نهاية الكلمة، ويقوم المحلل الصرفي بإضافة الخصائص اللغوية المناسبة لكل جزء من أجزاء الكلمة، و عوضاً عن عنوان واحد للكلمة نحن بحاجة الى عنوان صرفي (subtag) لكل جزء من أجزاء الكلمة (أو أكثر من عنوان (multiple subtags) لكل جزء إذا احتوى على أكثر من زائدة أو سابقة أو لاحقة للكلمة).

هناك تحديات كثيرة تواجه تطوير المحللات الصرفية للغة العربية، إن الخاصية الصرفية الغنية "الجذر والوزن" والنظام المعقد للاشتقاق الكلمات من الجذر والوزن خاصةً إذا احتوت هذه الجذور على حرف أو حرفين من حروف العلّة كحروف أصلية، علاوة على ذلك، إن الخواص الإملائية كالحركات (حروف العلة القصيرة – الفتحة والضمة والكسرة)، والهمزة، والتاء المربوطة والهاء في آخر الكلمة، والياء والألف المقصورة، والتشديد (الحروف المضعَّفة)، والمد (آ) الذي يعتبر حرفاً مركباً من الهمزة متبوعاً بألف، تعتبر من التحديات الصعبة لتطوير المحللات الصرفية للغة العربية.

يوظف المحلل الصرفي الذي تمّ تطويره قواعد النحو والصرف، كما يوظف الذخائر اللغوية للتحقق من هذه القواعد المستخدمة، ولفهم المشكلة بشكل جيد؛ قمنا بتحليل وتصنيف الجذور العربية لتحديد نسبة الجذور والكلمات التي تحتوي على حرف أو حرفين من حروف العلّة أو همزة أصلية في جذور هذه الكلمات، عن طريق تحليل الكلمة وجذرها لكلمات القرآن الكريم وكلمات القاموس المختلفة الذي تم بناءه من خلال تحليل خمسة عشر معجماً عربياً، ثم قمنا باستخلاص وبناء قوائم شاملة للزوائد والسوابق واللواحق والأوزان من الكتب القيمة للنحو والصرف وقواعد اللغة العربية، وقد تم فحص هذه القوائم عن طريق تحليل الذخائر اللغوية (القرآن الكريم، الذخيرة اللغوية العربية (The Corpus of Contemporary Arabic) و الذخيرة اللغوية العربية المطورة بجامعة بنسلفانيا (The Penn Arabic Treebank) بالاضافة الى نصوص المعاجم العربية الخمسة عشر كذخيرة لغوية رابعة، يعتمد المحلل الصرفي على هذه القوائم لتحليل الكلمات، كما قمنا بتطوير خوارزمية جديدة لتحديد وزن الكلمة الصحيح إذا احتوت هذه الكلمات على حروف علّة أو همزة أو اعلال أو إقلاب.

1.المقدمة

التحليل الصرفي للكلمة يعرف بأنه عملية تعيين الخصائص الصرفية للكلمة كجذر الكلمة أو جذعها ووزنها وتحديد الخصائص اللغوية لها، كنوع الكلمة (اسم أو فعل أو حرف) والتقسيمات الاخرى التي تندرج تحت هذه الانواع الثلاثة، وتحديد العدد للكلمة (المفرد أوالمثنى أوالجمع) والحالة الاعرابية (مرفوع أو منصوب أو مجرور أو مجزوم) وغيرها من الخصائص، كما يحدد المحلل الصرفي التركيب الداخلي للكلمة (الزوائد و السوابق واللواحق والجذر أو الجذع).

هنالك أربعة أساليب رئيسية قد تم تطبيقها لتطوير المحللات الصرفية بشكل عام، الأول : التحليل الصرفي المعتمد على مقطع الكلمة (Syllable-based Morphology (SBM)) حيث يتم التحليل الصرفي اعتماداً على تحديد مقاطع الكلمة، ثانياً: التحليل الصرفي المعتمد على الجذر والوزن (Root-Pattern Morphology) حيث يتم تحديد الجذر للكلمات بالاعتماد على مطابقة الكلمات بقوائم محددة من الأوزان واللواحق والسوابق، ثالثاً: التحليل الصرفي المعتمد على الجذع (Lexeme-based Morphology (LBM)) بحيث يتم استخراج الجذع (Stem) فقط للكلمة المحللة، وأخيراً: استخدام قوائم الجذع وقواعد النحو والصرف وخصائص الكلمات في التحليل الصرفي للكلمات (Soudi et al, 2007) ، جميع هذه الأساليب الأربعة تعتمد على قوائم معدَّة يدوياً تحتوي على معلومات الجذر أو الجذع أو الأوزان، علاوة على ذلك، هناك اسلوب آخر للتحليل الصرفي يعتمد على خوارزميات الذكاء الاصطناعي للذخائر اللغوية المعنونة نحوياً لبناء قاعدة بيانات للكلمات المحللة.

أما الاساليب الإحصائية، فقد تمَّ استخدامها بكثرة لتطوير خوارزميات لاستخراج الجذر أو الجذع للكلمات، بعض هذه الخوارزميات يعتمد تطبيق تقنية اختيار الأفضل من الأكثر تكراراً من الجذور أو اللواحق، وبعضها يعتبر نهاية الكلمات الأكثر تكراراً على أنها لاحقة للكلمة، ولكن لايمكننا توقع نتائج جيدة عند تطبيق هذه الاساليب للغة العربية، وذلك لأن عملية اشتقاق الكلمات في اللغة العربية لا تعتمد على اضافة اللواحق فقط للكلمة، بعض هذه الاساليب للغة العربية جمع بين التحليل المعتمد على الكلمة (word-based) وتحليل نهايات الكلمات بطول ستة حروف (6-gram)، وكانت نتائج هذه التحليلات جيدة عند تطبيقها لعدة لغات من بينها اللغة العربية، وغيرها يعتمد على تجميع الكلمات اعتماداً على التشابه في التركيب الصرفي للكلمات، لإيجاد تصنيف يجمع هذه الكلمات بنفس الجذر، وقد تم تطبيق هذه التقنية بعد إزالة عدد صغير من اللسوابق واللواحق المستخدمة بكثرة (Thabet, 2004).

ومن المحللات الصرفية الشائعة الاستخدام للغة العربية، المحلل الصرفي للغة العربية (Tim Buckwalter Morphological Analyzer[1]) الذي يعتمد على قوائم للكلمات ومعلوماتها الصرفية تمَّ إعدادها يدوياً، هذه القوائم تحتوي على قائمة الجذع وقائمة السوابق واللواحق، وقد تمَّ إضافة جداول الصواب والخطأ لتحديد التوافيق الصحيحة التي تجمع السوابق باللواحق للكلمات (Thabet, 2004) (Buckwalter, 2004).

ومن أنظمة استخراج الجذور، نظام شيرين خوجا لاستخراج جذر الكلمة (Khoja's Stemmer[2]) حيث يقوم هذا النظام على إزالة أطول سابقة ولاحقة للكلمة، ثم يقوم بمقارنة ما تبقى من الكلمة بقائمة من أوزان الأسماء والأفعال لاستخراج جذر الكلمة، ويحتوي النظام على العديد من الملفات التي تحتوي على بيانات مفيدة كقائمة

[1] www.qamus.org
[2] http://zeus.cs.pacificu.edu/shereen/research.htm

علامات التشكيل، وعلامات الترقيم، والجذور الثلاثية والرباعية، وأدوات التعريف، وقائمة تحتوي على 168 كلمة وقف (Stop Words)، وقد استخدم هذا النظام في تطبيقات استرجاع المعلومات، ورغم العديد من الأخطاء الناتجة في تحليل الكلمات، إلا أنه قد ساعد في تحسين النتائج لأنظمة استرجاع المعلومات (Khoja, 2001) (Larkey & Connell, 2001).

وكمثال آخر لنظام استخراج الجذور، نظام استخراج الجذور الثلاثية (Al-Shalabi et al, 2003)، حيث لا يعتمد هذا النظام على أي قوائم لغوية معدَّة يدوياً، بل يعتمد على حسابات رياضية بتعيين أوزان رقمية لأحرف الكلمة وضرب هذه الأوزان بمواقع حروفها، بحيث تعطى الاوزان في بداية الكلمة ونهايتها أوزاناً أعلى من أوزان حروف الكلمة في وسطها، ثم يقوم النظام بإختيار الحروف ذات الأقل وزناً كحروف لجذر تلك الكلمة، وقد قسَّم هذا النظام الحروف العربية الى مجموعتين، تحتوي المجموعة الأولى على الحروف التي لا تشكل أحد حروف السوابق أو اللواحق وتعطى الوزن صفر (0)، أما المجموعة الثانية فتضمُّ الحروف التي تتكون منها السوابق واللواحق والمجموعة بكلمة (سألتمونيها) وتعطى هذه الحروف أوزاناً مختلفة.

2. المحلل الصرفي للغة العربية

لقد قمنا بتطوير محللاً صرفياً للغة العربية، وكان هدفنا الرئيسي بناء ذخيرة لغوية معنونة (Tagged Corpus)، لقد بدأنا بدراسة المحللات الصرفية المطورة سابقاً والمنشورة عبر الشبكة العالمية، ورغم العديد من الأبحاث التي تعرض نتائجاً للتحليل الصرفي للغة العربية إلا ان هذه الأنظمة غير متاحة للباحثين والمستخدمين، فلذلك اقتصرت دراستنا على ثلاثة محللات صرفية مختلفة، هي المحلل الصرفي للغة العربية (Tim Buckwalter Morphological Analyzer)، ونظام شيرين خوجا (Khoja's Stemmer)، ونظام استخراج الجذور الثلاثية للكلمات (Al-Shalabi et al, 2003)، ولمقارنة نتائج التحليل للأنظمة المختلفة تم تطوير المعيار الذهبي (Gold Standard) والذي يتكون من ألف كلمة من نصوص القرآن الكريم (سورة العنكبوت) و ألف كلمة أخرى من نصوص المجلات والصحف من الذخيرة اللغوية العربية (The Corpus of Contemporary Arabic) (Al-Sulaiti & Atwell, 2006)، ولقد قمنا باستخراج جذور الكلمات لهذه النصوص المختارة وتدقيقها من قبل مختصِّين باللغة العربية، ثم قورنت نتائج التحليل للأنظمة الثلاثة بنظيراتها في المعيار الذهبي مستخدمين أربعة إختبارات لمقارنة دقة هذه الأنظمة، وأظهرت النتائج أن هذه المحللات الصرفية وخوارزميات استخراج الجذور لم تحقق النتائج المرجوه لاعتمادها في تطبيقات التحليل النحوي (Part-of-Speech tagging) أو استخدامها في تدقيق بنية الجملة (Parsing) حيث أن الأخطاء النتاجة من هذه المحللات الصرفية سوف تنتقل الى الأنظمة الأخرى التي تستخدمها وبالتالي لن نحصل على نتائج جيدة (Sawalha & Atwell, 2008).

2.1 دراسة تحليلية للجذور الثلاثية لكلمات اللغة العربية

لفهم طبيعة جذور اللغة العربية، وعلاقتها بمشتقاتها من الكلمات، فلقد قمنا بتصنيف الجذور الثلاثية لكلمات اللغة العربية الى اثنتان وعشرين مجموعة اعتماداً على التركيب الداخلي للجذر نفسه، من حيث خلو الجذر من أحرف العلّة أو الهمزة أو التضعيف أو وجودها، وقامت هذه الدراسة على تحليل كلمات وجذور كلمات القرآن الكريم حيث يحتوي على 45,534 كلمة ثلاثية الجذر و ومصدر آخر هو القاموس الذي تم بناءه اعتماداً على نصوص

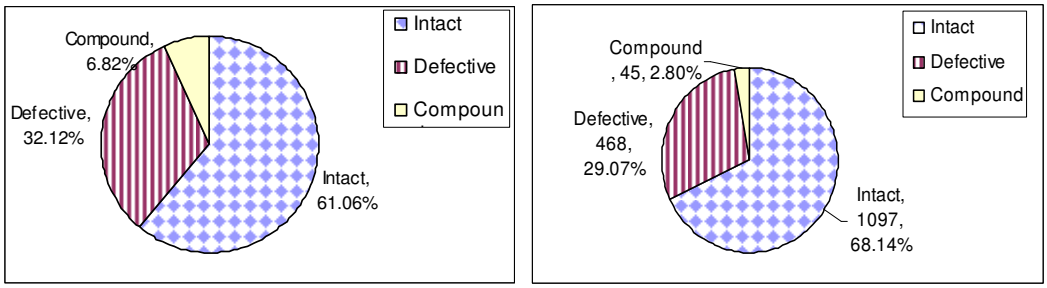
خمسة عشر معجماً عربياً اخترنا منها حوالي 376,167 كلمة مختلفة ثلاثية الجذر، الجدول (1) والجدول (2) يبينان نتائج التحليل لجميع فئات الجذور الثلاثية، وأظهرت النتائج أن حوالي 68% من الجذور الثلاثية هي جذور صحيحة (سالم أو مضعَّف أو مهموز) و 61% من كلمات القرآن الكريم تنتمي لهذه المجموعة، وأن 29% من الجذور الثلاثية جذور معتلّة (تحتوي على حرف أو حرفين من أحرف العلّة) وتشكل الكلمات التي تنتمي إلى هذه المجموعة ما نسبته 32% من كلمات القرآن الكريم، أما المجموعة الثالثة والتي تضُّم الجذور الثلاثية التي تحتوي على حرف أو حرفي علّة إضافة إلى الهمزة، فتشكل ما نسبته 3% من الجذور الثلاثية وحوالي 7% من مجموع كلمات القرآن الكريم، ويوضح الجدول (3) و الشكل (1) هذه النتائج.

جدول (2) نتائج تحليل جذور وكلمات القرآن الكريم

	الفئة		1	2	3	الجذور		كلمات القرآن الكريم	
1	صحيح	سالم	C1	C2	C3	870	54.04%	20,007	43.94%
2		مضعّف	C1	C2	C2	136	8.45%	3,814	8.38%
3		مهموز الفاء	H	C2	C3	44	2.73%	3,243	7.12%
4		مهموز العين	C1	H	C3	15	0.93%	281	0.62%
5		مهموز اللام	C1	C2	H	32	1.99%	459	1.01%
6	معتل	معتل الفاء	V	C2	C3	70	4.35%	1,252	2.75%
7		معتل العين	C1	V	C3	198	12.30%	8,162	17.93%
8		معتل اللام	C1	C2	V	167	10.37%	3,584	7.87%
9		لفيف مفروق	V	C2	V	12	0.12%	710	1.56%
10		لفيف مقرون	C1	V1	V2	19	1.18%	473	1.04%
11		لفيف مقرون	V1	V2	C3	2	0.12%	445	0.98%
12	مركب	مهموز ومضعف	H	C2	C2	7	0.43%	175	0.38%
13		معتل ومضعف	V	C2	C2	2	0.12%	40	0.09%
14		مهموز الفاء ومعتل اللام	H	C2	V	13	0.81%	958	2.10%
15		مهموز الفاء ومعتل العين	H	V	C3	6	0.37%	153	0.34%
16		مهموز الفاء ولفيف مقرون	H	V1	V2	2	0.12%	418	0.92%
17		مهموز العين ومعتل اللام	C1	H	V	2	0.12%	330	0.72%
18		مهموز العين ولفيف مفروق	V1	H	V2	0	0.00%	0	0.00%
19		مهموز العين ومعتل الفاء	V	H	C3	3	0.19%	15	0.03%
20		مهموز اللام ومعتل العين	C1	V	H	8	0.50%	998	2.19%
21		مهموز اللام ومعتل الفاء	V	C2	H	2	0.12%	17	0.04%
22		مهموز اللام ولفيف مقرون	V1	V2	H	0	0.00%	0	0.00%

جدول (2) نتائج تحليل الكلمات المختلفة للقاموس

	الفئة		1	2	3	الجذور		كلمات القاموس	
1	صحيح	سالم	C1	C2	C3	4147	48.78%	201,385	53.54%
2		مضعّف	C1	C2	C2	446	5.25%	32,007	8.51%
3		مهموز الفاء	H	C2	C3	289	3.40%	10,449	2.78%
4		مهموز العين	C1	H	C3	216	2.54%	3,909	1.04%
5		مهموز اللام	C1	C2	H	270	3.18%	8,985	2.39%
6	معتل	معتل الفاء	V	C2	C3	386	4.54%	19,219	5.11%
7		معتل العين	C1	V	C3	1115	13.11%	43,512	11.57%
8		معتل اللام	C1	C2	V	1151	13.54%	41,295	10.98%
9		لفيف مفروق	V	C2	V	45	0.08%	2,372	0.63%
10		لفيف مقرون	C1	V1	V2	106	1.25%	4,057	1.08%
11		لفيف مقرون	V1	V2	C3	22	0.26%	211	0.06%
12	مركب	مهموز ومضعف	H	C2	C2	30	0.35%	888	0.24%
13		معتل ومضعف	V	C2	C2	29	0.34%	463	0.12%
14		مهموز الفاء ومعتل اللام	H	C2	V	74	0.87%	2,111	0.56%
15		مهموز الفاء ومعتل العين	H	V	C3	47	0.55%	892	0.24%
16		مهموز الفاء ولفيف مقرون	H	V1	V2	7	0.08%	135	0.04%
17		مهموز العين ومعتل اللام	C1	H	V	42	0.49%	1,041	0.28%
18		مهموز العين ولفيف مفروق	V1	H	V2	2	0.02%	52	0.01%
19		مهموز العين ومعتل الفاء	V	H	C3	15	0.18%	292	0.08%
20		مهموز اللام ومعتل العين	C1	V	H	42	0.49%	1,590	0.42%
21		مهموز اللام ومعتل الفاء	V	C2	H	21	0.25%	1,302	0.35%
22		مهموز اللام ولفيف مقرون	V1	V2	H	0	0.00%	0	0.00%

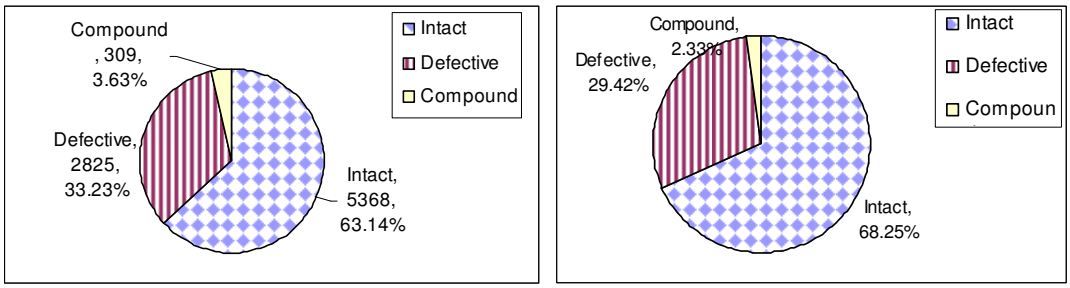


الشكل (1) توزيع الجذور(يمين) والكلمات(يسار) لجذور وكلمات القرآن الكريم

الجدول(3) توزيع جذور وكلمات القرآن الكريم

	الجذور		كلمات القرآن الكريم	
الفئة	المجموع	النسبة المئوية	المجموع	النسبة المئوية
الجذور الصحيحة	1097	68.14%	27,804	61.06%
الجذور المعتلّة	468	29.07%	14,626	32.12%
الجذور المعتلّة والمهموزة	45	2.80%	3,104	6.82%
المجموع	1610	100.00%	45,534	100.00%

وجاء تحليل الكلمات المختلفة (Word Types) والمخزنة في القاموس مشابهاً لنتائج تحليل جذور وكلمات القرآن الكريم، حيث أظهرت النتائج أن 63% من الجذور هي جذور صحيحة و 68% من الكلمات تنتمي لهذه المجموعة، وأن 33% من الجذور هي جذور معتلّة و تنتمي لهذه المجموعة 29% من كلمات القاموس، وأن حوالي 4% من الجذور هي جذور معتلّة ومهموزة معاً وحوالي 2% من هذه الكلمات تنتمي لهذه المجموعة، ويوضح الجدول (4) والشكل (2) هذه النتائج.



الشكل (2) توزيع الجذور(يسار) والكلمات(يمين) لجذور وكلمات القاموس

الجدول(4) توزيع جذور وكلمات القاموس

	الجذور		الكلمات المختلفة للقاموس	
الفئة	المجموع	النسبة المئوية	المجموع	النسبة المئوية
الجذور الصحيحة	**5368**	**63.14%**	**256,735**	**68.25%**
الجذور المعتلّة	**2825**	**33.23%**	**110,666**	**29.42%**
الجذور المعتلّة والمهموزة	**309**	**3.63%**	**8,766**	**2.33%**
المجموع	**8502**	**100.00%**	**376,167**	**100.00%**

2.2 مواصفات المحلل الصرفي العربي

2.2.1 المُدخَل

يقبل المحلل الصرفي النصوص أو الكلمات المفردة سواء أكانت مشكولة كلياً أو جزئياً أو غير مشكولة كمدخلات للنظام، وللتعامل مع هذه الأنواع المختلفة من النصوص تم إقتراح هيكلية واحده للبيانات، فبداية يقوم برنامج خاص بتقطيع كلمات النص المدخل إلى كلمة عربية مشكولة كلياً أو جزئياً أو غير مشكولة، أو إلى رقم أو عُملة أو علامة ترقيم، بعدها يقوم البرنامج بمعالجة الكلمات العربية المستخرجة من النص المدخل، وتقوم عملية المعالجة بإرجاع الحروف المضعَّفة والمد إلى أصلها، حيث يتم استبدال الحرف المضعَّف والشَّدة الظاهرة عليه بحرفين الأول ساكن والثاني متحرك بنفس حركة الحرف الأصلي، فمثلا؛ كلمة (وَصَّى) تصبح (وَصْصَى)، وكذلك يتم استبدال المد (آ) بحرفي الهمزة والألف، فكلمة (آمَنُوا) تصبح (ءامَنُوا).

ولأنه يمكن أن تظهر حركة واحدة (حرف علّة قصير) على أي حرف من حروف الكلمة، تم اعتماد هيكلية موحدة للكلمات بحيث تتكون من مصفوفة أُحاديَّة تضُّم الحروف والحركات التي تظهر عليها إن وجدت، ويتمُّ تخزين الحرف الأول من الكلمة في الموقع الأول للمصفوفة يليه حركة الحرف (حرف العلّة القصير) في الموقع الثاني وهكذا لجميع حروف الكلمة، ويوضح الشكل (3) الهيكلية المستخدمة لتخزين كلمتي (وَصْصَى و ءامَنُوا)، كما تستخدم هذه الهيكلية لتحليل الأوزان لتسهيل عملية مطابقة الكلمة مع الوزن المناسب لها.

الموقع	1	2	3	4	5	6	7	8	9	10	11	12
وَصْصَى	و	َ	ص	ْ	ص	َ	ى	–				
ءامَنُوا	ء	–	ا	–	م	َ	ن	ُ	و	–	ا	–

الشكل (3) الهيكلية المستخدمة في تخزين الكلمات المحللة

2.2.2 كلمات الوقف (Stop Words)

يحتوي النظام على قائمة بكلمات الوقف (Stop Words) مكونة من 1,368 كلمة، تتميز هذه الكلمات بأنها تحمل تحليلاً صرفياً واحداً أينما وجدت في سياق النَّص، وتبلغ نسبة هذه الكلمات في أي نص من نصوص اللغة العربية حوالي 40% من إجمالي عدد كلمات ذلك النَّص، ويقوم المحلل الصَّرفي بالبحث في هذه القائمة عن الكلمة المراد تحليلها، فإذا وجدت هذه الكلمة ضمن كلمات الوقف تُعطى التحليل الصرفي المخزن في القائمة، ويتم الانتقال إلى الكلمة التالية ليتم تحليلها، يتضمن الشكل (4) بعض هذه الكلمات.

أنا	هؤلاء	الذي	على	عند
فأنا	فهؤلاء	فالذي	فعلى	عندي
نحن	بهؤلاء	كالذي	علينا	عندنا
فنحن	لهؤلاء	بالذي	فعلينا	عنده
هي	فبهؤلاء	للذي	عليها	عندها
فهي	فلهؤلاء	فكالذي	فعليها	عندهما

الشكل (4) أمثلة على كلمات الوقف (Stop Words)

2.2.3 الزوائد والسوابق واللواحق

اعتماداً على كتب قواعد اللغة العربية فقد تمَّ حصر الزوائد في بداية الكلمة (Proclitics) (كحروف العطف والجر والنداء وأدوات التعريف)، والسوابق (Prefixes) ، واللواحق (Suffixes) والزوائد في نهاية الكلمة (Enclitics) (كالضمائر المتصلة)، وقد زوِّدت هذه المعلومات إلى برنامج مولِّد، حيث تمَّ توليد جميع الزوائد الممكنة وكان عددها كبيراً، ولحصرها تم التحقق من صحة هذه القوائم من خلال فحصها عن طريق تحليل كلمات الذخائر اللغوية: القرآن الكريم، والذخيرة اللغوية العربية (Corpus of Contemporary Arabic (CCA)) ، والذخيرة اللغوية العربية المطورة بجامعة بنسلفانيا (Penn Arabic Treebank) ونصوص معاجم اللغة العربية المستخدمة في بناء القاموس كذخيرة لغوية رابعة. وقد حصلنا على قائمتين من السوابق واللواحق، بلغ عدد السوابق 215 سابقة كما بلغ عدد اللواحق 127 لاحقة، ويبين الجدول (5) والجدول (6) جزءاً من هذه السوابق واللواحق مع مثال على كلٍّ منهما والعنوان الصرفي[3] المعيَّن لكل سابقة ولاحقة.

كما يقوم النظام بتقسيم الكلمة إلى ثلاثة أجزاء بأطوال مختلفة، ثم يبحث عن الجزء الأول في قائمة السوابق ويبحث عن الجزء الثالث بقائمة اللواحق، وعند العثور على السابقة أو اللاحقة في القوائم يتم تعيين التحليل الصرفي المرفق في القوائم لهذه الأجزاء من الكلمة، ثم يتم اختيار تحليلات الكلمة التي تمت مطابقة السابقة واللاحقة لجزئيها معاً، ويوضح الجدول (7) عملية مطابقة السوابق واللواحق لأجزاء الكلمة المحللة.

[3] راجع الجزء 4 الخصائص اللغوية للكلمة العربية وتركيب العناوين الصرفية والنحوية

الجدول (5) جزء من قائمة السوابق و العنوانين الصرفية المعيَّنة لها

السابقة	مثال	ج1	العنوان الصرفي	ج2	العنوان الصرفي	ج3	العنوان الصرفي
ف	فقام	ف	p--t----------				
فبال	فبالصدق	ف	p--t----------	ب	p--r-----g---------	ال	r---d-----d--------
فست	فستذكرون	ف	p--t----------	س	p--i----------	ت	r---s-nus----------
وال	والسماء	و	p--t----------	ال	r---d-----d--------		
ولت	ولتجدنهم	و	p--t----------	ل	p--y----------	ت	r---s-nus----------
سن	سنجزي	س	p--i----------	ن	r---s-npf---------		
وبم	وبمراجعة	و	p--t----------	ب	p--r-----g---------	م	r---f----------

الجدول (6) جزء من قائمة اللواحق و العنوانين الصرفية المعيَّنة لها

اللاحقة	مثال	ج1	العنوان الصرفي	ج2	العنوان الصرفي	ج3	العنوان الصرفي
اتية		ات	r---l-fp-??----?---	ي	r---j----------	ة	r---b-fs-??--------
تموهما		تم	r---&-mps??----h---	و	r---l-mp-n?----?---	هما	r---&-ndt??----h---
هما		هما	r---&-ndt??----h---				
يون		ي	r---j----------	ون	r---l-mp-n?----?---		
هم		هم	r---&-mpt??----h---				
ها		ها	r---&-fst??----h---				
هن		هن	r---&-fpt??----h---				

الجدول (7) مثال على عملية اختيار التحليل الذي يطابق السوابق واللواحق.

الكلمة	الجزء الأول	الجزء الثاني	الجزء الثالث	تحليل السوابق واللواحق
يَعْمَلُونَ		**يعملون**		**تحليل مقبول**
يَعْمَلُونَ		يعملو	ن	تحليل غير مقبول
يَعْمَلُونَ		**يعمل**	**ون**	**تحليل مقبول**
يَعْمَلُونَ		يعم	لون	تحليل غير مقبول
يَعْمَلُونَ		يع	ملون	تحليل غير مقبول
يَعْمَلُونَ		ي	عملون	تحليل غير مقبول
يَعْمَلُونَ	ي	**عملون**		**تحليل مقبول**
يَعْمَلُونَ	ي	عملو	ن	تحليل غير مقبول
يَعْمَلُونَ	ي	**عمل**	**ون**	**تحليل مقبول**
يَعْمَلُونَ	ي	عم	لون	تحليل غير مقبول
يَعْمَلُونَ	ي	ع	ملون	تحليل غير مقبول
يَعْمَلُونَ	يع	ملون		تحليل غير مقبول
يَعْمَلُونَ	يع	ملو	ن	تحليل غير مقبول
يَعْمَلُونَ	يع	مل	ون	تحليل غير مقبول
يَعْمَلُونَ	يع	م	لون	تحليل غير مقبول
يَعْمَلُونَ	يعم	لون		تحليل غير مقبول
يَعْمَلُونَ	يعم	لو	ن	تحليل غير مقبول
يَعْمَلُونَ	يعم	ل	ون	تحليل غير مقبول
يَعْمَلُونَ	يعمل	ون		تحليل غير مقبول

2.2.4 الجذر أو الجذع

يستخدم النظام قائمة لجذور اللغة العربية الثلاثية والرباعية والخماسية حيث احتوت هذه القائمة على أكثر من 12,000 جذراً، تم استخراجها من تحليل خمسة عشر معجماً عربياً، وبعد اختيار تحليلات الكلمة من الخطوة السابقة والتي تم فيها مطابقة السوابق واللواحق، يقوم النظام بالبحث في قائمة الجذور ومطابقة الجزء الثاني من الكلمة المحللة، ويبين الجدول (8) التحليلات المقبولة في الخطوة السابقة وعملية اختيار التحليل الذي يطابق السوابق واللواحق والجذر معاً.

الجدول (8) مثال على عملية اختيار التحليل الذي يطابق السوابق واللواحق والجذر معاً.

الكلمة	الجزء الأول	الجزء الثاني	الجزء الثالث	تحليل الزوائد واللواحق	تحليل الزوائد واللواحق والجذر معاً
يَعْمَلُونَ		يعملون		تحليل مقبول	تحليل غير مقبول
يَعْمَلُونَ		يعمل	ون	تحليل مقبول	تحليل غير مقبول
يَعْمَلُونَ	ي	عملون		تحليل مقبول	تحليل غير مقبول
يَعْمَلُونَ	**ي**	**عمل**	**ون**	**تحليل مقبول**	**تحليل مقبول**

2.2.5 وزن الكلمة

تتمُّ عملية اشتقاق الكلمات المختلفة من الجذر الثلاثي أو الرباعي أو الخماسي من خلال اتباع أوزان محددة، تحمل هذه الأوزان خواصاً لغوية هي نفسها للكلمة المشتقة، من هذا الأساس؛ وقد قمنا بتزويد المحلل الصرفي بقائمة من الأوزان التي تم استخراجها من كتب قواعد اللغة العربية والنحو والصرف، تحتوي قائمة أوزان الأفعال على 2730 وزناً وتحتوي قائمة أوزان الأسماء على 390 وزناً، ولقد تم تعيين التحليل الصرفي لكل وزن، ويبين الجدول (9) بعض هذه الأوزان والعناوين الصرفية لها.

الجدول (9) بعض الأوزان والعناوين الصرفية لها

فَعَلْتُ	v-p---nsf---an?-st?	أُفْعُلاوَى	nw----??-??----?qt-
فَعَلْنَا	v-p---npf---an?-st?	اِفْعِيلال	nw----??-??----?qt-
فَعَلْتَ	v-p---mss---an?-st?	فاعُولاء	nw----??-??----?qt-
فَعَلْتِ	v-p---fss---an?-st?	فُعَلْعُلان	nw----??-??----?qt-
فَعَلْتُمَا	v-p---nds---an?-st?	فُعَيْلاء	nw----??-??----?qt-

تتميز هذه الأوزان بأنها مشكولة مما يتيح إضافة التشكيل للكلمات المحللة وغير المشكولة، ويستخدم النظام طريقتين لايجاد الوزن المناسب للكلمة اعتماداً على التحليلات السابقة للسوابق واللواحق والجذور.

أ- الطريقة الأولى (الكلمة وجذرها)

تعتمد الطريقة الأولى لاستخراج وزن الكلمة على الكلمة نفسها وجذرها كمدخل للبرنامج، حيث يتم اختيار التحليلات التي تتطابق فيها السوابق واللواحق والجذر معاً، حيث يتم استبدال حروف الجذر في الكلمة بالحروف (ف، ع، ل)، ولكن لاتتم هذه العملية بهذه السهولة؛ فبعض حروف الجذر قد يطرأ عليها تغيير كالإدغام والإقلاب والإعلال والإبدال، وعلى البرنامج أن يستخرج الوزن الصحيح لهذه الكلمات ومعالجة هذه الحالات، وفي النهاية

يتم البحث عن الوزن المستخرج في قوائم الأوزان، فإن وجد تعطى الكلمة التحليل الصرفي المعيين لهذا الوزن، ويبين الشكل (6) عملية استخراج الجذر بهذه الطريقة.

Word	أَحَسِبَ	Letters	أ	ح	س	ب		Root	حسب	ح	س	ب	
		Index	0	1	2	3				1	2	3	

Root letters indices		
First letter (ح) = [1]	Second letter (س) = [2]	Third letter (ب) = [3]

Candidate indices list = [1,2,3]

Pattern	Prefix	Stem	Suffix
أفعل	أ	حسب	

Word	آمَنُوا	Letters	أ	ا	م	ن	و	ا	Root	أمن	أ	م	ن	
		Index	0	1	2	3	4	5			1	2	3	

Root letters indices		
First letter (أ) = [-1, 0]	Second letter (م) = [2]	Third letter (ن) = [3]

Indices [-1, 2, 3] , [0, 2, 3]

Candidate indices list = [-1 , 2 , 3]

Pattern	Prefix	Stem	Suffix
أاعلوا	أا	من	وا

Candidate indices list = [0, 2, 3]

Pattern	Prefix	Stem	Suffix
فاعلوا		أامن	وا

Word	الْعَلِيم	Letters	ا	ل	ع	ل	ي	م	Root	علم	ع	ل	م	
		Index	0	1	2	3	4	5			1	2	3	

Root letters indices		
First letter (ع) = [2]	Second letter (ل) = [1,3]	Third letter (م) = [5]

Candidate indices list = [2 ,1 , 3] False [2,3,5] True

Pattern	Prefix	Stem	Suffix
الفعيل	ال	عليم	

الشكل (6) عملية استخراج الجذر بالطريقة الأولى (الكلمة وجذرها)

ب– الطريقة الثانية لاستخراج وزن الكلمة

تعتمد الطريقة الثانية لايجاد وزن الكلمة بشكل أساسي على قوائم الأوزان، مسترشدين بخوارزمية مطابقة الأوزان (Pattern Matching Algorithm (PMA)) (Alqrainy, 2008) حيث تعمل هذه الخوارزمية على مطابقة الكلمة مع وزنها للكلمات المشكولة جزئياً بالحركة الظاهرة على آخر الكلمة فقط وبدون إجراء أي تحليل للسوابق واللواحق.

أما في هذا المحلل الصرفي، يقوم النظام بالبحث عن جميع الأوزان المساوية في الطول للكلمة المحللة بعد أن تم إزالة الزوائد من بدايتها ونهايتها، فمثلاً كلمة (كتب) طولها 6 حسب الهيكلية المتبعة لتخزين الكلمة، سواء أكانت مشكولة كلياً أو جزئياً أو غير مشكولة يبقى طولها 6، وستطابق الاوزان (فَعْل، فَعَل، فَعُل، فَعِل، فُعْل، فُعَل، فُعُل، فُعِل، فِعْل)، وفي الخطوة الثانية يتم استبدال حروف الكلمة المقابلة للحروف (ف، ع، ل) في الوزن، وبعد

ذلك يتم البحث عن هذه الأوزان الناتجة عن دمج الكلمة والوزن معاً في قائمة الأوزان، فإن وجدت يكون هذا وزناً محتملاً للكلمة ويمثل التحليل الصرفي المرفق مع الوزن في القائمة تحليلاً صرفياً لهذه الكلمة، وإذا لم يتم العثور على الوزن في قائمة الأوزان فلا يعتبر تحليلاً محتملاً للكلمة، ويبين الشكل (7) عملية استخراج الأوزان للكلمات بهذه الطريقة.

الكلمة	الوزن	العنوان الصَّرفي
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعِلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعِلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يُفْعِلُونَ	v-c---mpt--ian?-at?
يَعْمَلُونَ	يُفْعَلُونَ	v-c---mpt--ipn?-tt?
يَعْمَلُونَ	يُفْعَلُونَ	v-c---mpt--ipn?-at?
كتب	فَعَلَ	v-p---mst---an?-st?
كتب	فَعِلَ	v-p---mst---an?-st-
كتب	فَعُلَ	v-p---mst---an?-st-
كتب	فَعِلَ	v-p---mst---an?-st-
كتب	فُعِلَ	v-p---mst---pn?-tt-
كتب	فَعْل	nw----??-??----?st-
كتب	فَعَل	nw----??-??----?st-
كتب	فَعَل	nw----??-??----?st-
كتب	فَعِل	nw----??-??----?st-
كتب	فُعْل	nw----??-??----?st-
كتب	فُعَل	nw----??-??----?st-
كتب	فُعَل	nw----??-??----?st-
كتب	فُعِل	nw----??-??----?st-
كتب	فِعْل	nw----??-??----?st-
كتب	فِعل	nw----??-??----?st-
كتب	فَعِل	ny----??-??----?st-

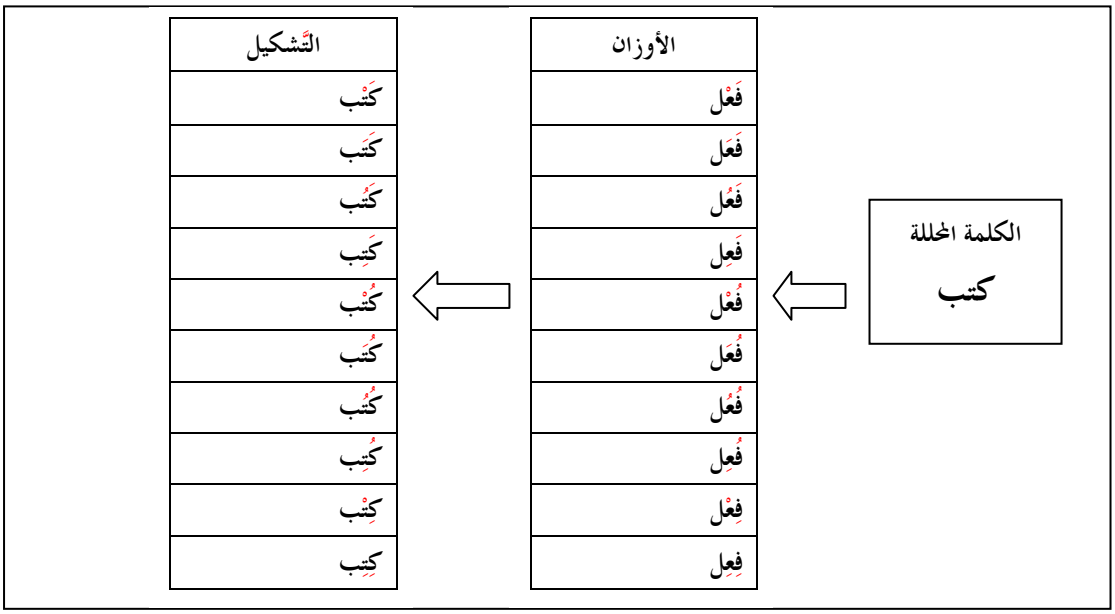
الشكل (7) عملية استخراج الأوزان للكلمات باستخدام الطريقة الثانية.

2.2.6 التَّشكيل

يعتبر التَّشكيل أحد الخصائص المهمة للكلمة العربية حيث يفيد التشكيل في تحديد خصائص لغوية أخرى للكلمة، فوجود التَّشكيل في آخر الكلمة (الفتحة أو الضمة أو الكسرة أو السكون) يفيد في تحديد الحالة الإعرابية للكلمة، ووجود الحركة في بداية الكلمة (الضمة أو الفتحة أو الكسرة) يفيد في تحديد بناء الفعل للمعلوم أو المجهول، أما الحركات الأخرى كالشَّدة فظهورها على الكلمة يحلُّ بعض اللُّبس في تحليل الكلمات.

بعد تحديد الأوزان المطابقة للكلمة في الخطوة السابقة، وبما أن الأوزان مشكولة كلياً، يقوم النظام بإضافة الحركات التي ظهرت على الوزن المطابق الى الكلمة المحللة سواء أكانت مشكولة جزئياً أو غير مشكولة، وبهذا

تكون الكلمة قد تمَّ شكلها بالحركات الصحيحة المحتملة لهذه الكلمة، ويوضح الشكل (8) عملية شكل هذه الكلمات.



الشكل (8) إضافة حركة الوزن الى الكلمة المحللة غير المشكولة

4. الخصائص اللغوية للكلمة العربية وتركيب العناوين الصرفية والنحوية

قسَّم علماء اللغة العربية وقواعدها الكلمة إلى ثلاثة أقسام رئيسية هي الاسم والفعل والحرف، وتم وصف وتفصيل وتحديد خصائص هذه الأقسام بدقة، فضمت الاسماء المصادر والضمائر والصفات واسماء الإشارة والأسماء الموصولة واسم العلم واسماء الزمان والمكان وغيرها، كما ضمت الأفعال الفعل الماضي والفعل المضارع وفعل الأمر، وضمت الحروف؛ حروف الجرِّ والنسخ والعطف والجزم والنَّصب والنداء وغيرها، وقد تمَّ تحديد الخصائص اللغوية للكلمة كالجنس (مذكر أو مؤنث) والعدد (مفرد أو مثنى أو جمع)، والإسناد (متكلم أو مخاطب أو غائب) والحالة الإعرابية للاسم (مرفوع أو منصوب أو مجرور)، والمعرفة والنكرة، وحالة الإعراب أو البناء للفعل (مرفوع أو منصوب أو مجزوم)، وبناء الفعل للمعلوم أو المجهول، والتَّوكيد (مؤكد أو غير مؤكد)، والفعل اللازم أو المتعدي لمفعول واحد أو مفعولين إثنين او ثلاثة مفاعيل، ومن الخصائص الأخرى التي تمَّ تحديدها للكلمة والتي تصف بنيتها، المجرَّد والمزيد، وعدد حروف الكلمة الأصلية، وتركيب أحرف الفعل الثلاثي من حيث الصحة والإعلال.

اعتماداً على هذه الخصائص اللغوية تمَّ تصميم مجموعة العناوين الصرفية والنحوية (Morphological features part-of-speech tag set) لإستخدامها في بناء المحلل النحوي للغة العربية، لعنونة الذخائر اللغوية بشكل تفصيلي يعكس الخصائص اللغوية والصرفية لكلماتها، ولتمكيننا من دراسة وتحليل وتقييم نتائج المحللات الصرفية والنحوية للغة العربية بشكل مباشر مستخدمين مجموعة العناوين (Tag set) نفسها، وقد صممت مجموعة العناوين بتجميع تسعة عشر خاصية لغوية للكلمة في عنوان واحد، ويوضح الجدول (10) الخصائص اللغوية المستخدمة في بناء مجموعة العناوين الصرفية والنحوية للغة العربية، تحتوي قائمة المرفقات على العناوين الصرفية والنحوية للخصائص اللغوية.

الجدول (10) الخصائص اللغوية المستخدمة في بناء مجموعة العناوين الصرفية والنحوية للغة العربية

الموقع		الخصائص اللغوية
1	أقسام الكلام الرئيسيَّة	Main Part-of-Speech
2	أقسام الكلام الفرعيَّة (الاسم)	Subcategory (Noun)
3	أقسام الكلام الفرعيَّة (الفعل)	Subcategory (Verb)
4	أقسام الكلام الفرعيَّة (الحرف)	Subcategory (Particle)
5	أقسام الكلام الفرعيَّة (أخرى)	Subcategory (Residual)
6	أقسام الكلام الفرعيَّة (علامات الترقيم)	Subcategory (Punctuation)
7	الجنس	Gender
8	العدد	Number
9	الشخص	Person
10	الحالة الإعرابية للاسم	Case
11	التَّعريف والتَّنكير	Definiteness
12	الحالة الإعرابية للفعل	Mood
13	المعلوم والمجهول	Active / Passive
14	التَّوكيد	Emphasized
15	التَّعدية	Transitivity
16	العاقل وغير العاقل	Humanness
17	المجرَّد والمزيد	Stripped or augmented
18	عدد الحروف الأصلية	Root letters
19	تركيب أحرف الفعل	Internal Structure

يتكون العنوان الصرفي والنحوي من تسعة عشر رمزاً، كل رمز يمثل قيمة أو متغير ينتمي إلى إحدى الخصائص الصرفية أو النحوية، ويعدُّ موقع الرمز في العنوان مهماً في تحديد هذه الخاصية، وتمثل هذه القيم أو المتغيرات برمز واحد من حروف اللغة الإنجليزية الصغيرة، فمثلا الرمز (v) في الموقع الأول من العنوان يرمز إلى الفعل، والرمز (n) في الموقع الثاني يرمز إلى اسم العلم، ويمثل الجنس في الموقع السابع من العنوان حيث يرمز الحرف (m) الى المذكر والحرف (f) الى المؤنث، وإذا كانت الخاصية اللغوية غير متوافقة مع الكلمة فالرمز (–) (الشرطة) يمثلها، بينما يستخدم الرمز (؟) (علامة السؤال) لترمز إلى أن الخاصية اللغوية تنطبق على الكلمة ولكن غير محددة.

وتتم ترجمة العنوان الصرفي والنحوي للكلمة المعنونة بالرجوع إلى الرمز و موقعه في العنوان لتحديد الخاصية اللغوية التي ينتمي لها الرمز، وبعدها يتم تجميع هذه الأجزاء في وصف صرفي واحد للكلمة، ويبين الشكل (9) أمثلة لجملتان تمت عنونة كلماتهما باستخدام مجموعة العناوين الصرفية والنحوية هذه وترجمة هذه العناوين.

الكلمة	العنوان الصرفي	ترجمة العنوان الصرفي
المثال الأول جملة من القرآن الكريم		
وَ	p--t-----a---------	حرف حرف عطف منصوب
وَصَّيْ	v-p---npf--iano-at&	فعل فعل ماضي مؤنث/مذكر جمع سالم المُخاطَب الرَّفع مَبْنِي للمَعْلُوم غَيْر مُؤكَّد مُتَعدِّي الى مَفعُول واحِد مَزِيْد بِحَرف ثلاثي اللفيف المفروق
نَا	p--&---p-n---------	حرف جمع سالم مرفوع
الْإِنسَانَ	nq----np-ad-----bt-	اسم اسم جنس مؤنث/مذكر جمع سالم منصوب مَعْرِفَة مَزِيْد بِحَرفَيْن ثلاثي
بِ	p--r-----g---------	حرف حرف جر مجرور
وَالِدَيْ	nw----nd-gd-----at-	اسم اسم معرب مؤنث/مذكر مثنى مجرور مَعْرِفَة مَزِيْد بِحَرف ثلاثي
هِ	p--&-----g---------	حرف مجرور
حُسْنًا	no----nu-ai-----st-	اسم مصدر مؤنث/مذكر غير محدد منصوب نَكِرَة مُجَرَّد ثلاثي
المثال الثاني جملة من الذخيرة اللغوية العربية (Penn Arabic Treebank)		
تم	v-p---mst---ano-stb	فعل فعل ماضي مذكر مفرد الغَائِب مَبْنِي للمَعْلُوم غَيْر مُؤكَّد مُتَعدِّي الى مَفعُول واحِد مُجَرَّد ثلاثي مضعف
اعداد	no----ms-ni-----bt-	اسم مصدر مذكر مفرد مرفوع نَكِرَة مَزِيْد بِحَرفَيْن ثلاثي
الوثائق	nw----fb-ad----ntt-	اسم اسم معرب مؤنث جمع تكسير منصوب مَعْرِفَة غَيْر عَاقِل مَزِيْد بِثَلاثَةِ حُرُوف ثلاثي
المتوفرة	ns----fs-gd-----tt-	اسم صفه مشبهة مؤنث مفرد مجرور مَعْرِفَة مَزِيْد بِثَلاثَةِ حُرُوف ثلاثي
ب	p--r-----g---------	حرف حرف جر مجرور
كثرة	ns----fb-gi-----at-	اسم صفه مشبهة مؤنث جمع تكسير مجرور نَكِرَة مَزِيْد بِحَرف ثلاثي
حول	nh----nu-ai-----st-	اسم ظرف مؤنث/مذكر غير محدد منصوب نَكِرَة مُجَرَّد ثلاثي
أول	nm----nu-gi-----st-	اسم اسم عدد مؤنث/مذكر غير محدد مجرور نَكِرَة مُجَرَّد ثلاثي
رحلة	nw----fs-gi-----at-	اسم اسم معرب مؤنث مفرد مجرور نَكِرَة مَزِيْد بِحَرف ثلاثي
طيران	no----nu-gi-----bt-	اسم مصدر مؤنث/مذكر غير محدد مجرور نَكِرَة مَزِيْد بِحَرفَيْن ثلاثي
عثمانية	nr----fu-gi----htq-	اسم اسم منسوب مؤنث غير محدد مجرور نَكِرَة عاقِل مَزِيْد بِثَلاثَةِ حُرُوف رباعي
فوق	nh----nu-ai-----st-	اسم ظرف مؤنث/مذكر غير محدد منصوب نَكِرَة مُجَرَّد ثلاثي
البلاد	nw----fb-gd----nat-	اسم اسم معرب مؤنث جمع تكسير مجرور مَعْرِفَة غَيْر عَاقِل مَزِيْد بِحَرف ثلاثي
العربية	nr----fu-gd----hbt-	اسم اسم منسوب مؤنث غير محدد مجرور مَعْرِفَة عاقِل مَزِيْد بِحَرفَيْن ثلاثي

الشكل (9) أمثلة لجملتان تمت عنونة كلماتهما باستخدام مجموعة العناوين الصرفية والنحوية هذه وترجمة هذه العناوين

5. النتائج والتقييم

5.1 المعيار الذهبي (gold standard) لتقييم نتائج المحللات الصرفية والنحوية

تستخدم المعايير الذهبية لتقييم وقياس دقة الأنظمة المحوسبة، كما يمكن استخدامها للمقارنة بين عدَّة أنظمة أو خوارزميات طورت لحل مشكلة معيَّنة، وتُظهر المعايير الحالات التي تنجح أو تفشل الأنظمة المُقيَّمة بتحديد التحليل المناسب للمدخلات، وتستخدم المعايير الذهبية لإيجاد أوجه الشبه أو الإختلاف في نتائج التحليل مبينة الحالات التي تتفق عليها والتي تختلف فيها الأنظمة المحوسبة.

ولبناء معيار ذهبي لتقييم الأنظمة المحوسبة، يجب علينا بداية تحديد موضوع المشكلة التي تقوم هذه الأنظمة على حلها، كما يجب تحديد الذخيرة اللغوية التي ستستخدم لبناء المعيار الذهبي، وتحديد تنسيقة أو ترتيبه وحجمه، كما يجب تحديد قواعد الكتابة والترجمة ومراحل بناء المعيار الذهبي.

5.1.1 موضوع مشكلة البحث

سيتم بناء معيار ذهبي لاستخدامه في تقييم المحللات الصرفية والنحوية للغة العربية، فلذلك يجب أن يتوفر التحليل الصرفي والنحوي لجميع كلمات المعيار الذهبي.

5.1.2 الذخيرة اللغوية

تستخدم الذخائر اللغوية لبناء المعايير الذهبية، حيث يوجد العديد من الذخائر اللغوية العربية التي تم بنائها مسبقاً، ويعتمد معضمها على نصوص مقتبسة من المجلات والصحف، ولكن لبناء معيار ذهبي واسع التطبيق يجب علينا اختيار نصوص عربية من مصادر واشكال ومجالات متعددة، ومن نصوص مشكولة كلياً أو جزئياً أو غير مشكولة، ومن الذخائر اللغوية المقترح استخدامها لبناء المعيار الذهبي، نصوص القرآن الكريم المشكولة وغير المشكولة، ونصوص الذخيرة اللغوية العربية (Corpus of Contemporary Arabic (CCA)) والتي تتكون من مليون كلمة تم تجميعها من الصحف والمجلات وتشمل أربعة عشر مجالاً كالسير الذاتية، والقصص القصيرة، وقصص الأطفال، ومقالات في الإقتصاد والتعليم والصحة والطب والسياسة والدين والرياضة والسياحة والسفر ووصفات الطعام ونصوص علمية أخرى (Al-Sulaiti & Atwell, 2006).

5.1.3 تنسيق المعيار الذهبي

سيحتوي المعيار الذهبي على التحليل الصرفي والنحوي لكل كلمة من كلمات الذخيرة اللغوية المستخدمة في بنائه، بحيث يظهر التحليل النحوي والصرفي والكلمة معاً في سطر واحد، ويمكن إضافة جذر الكلمة ووزنها لهذا التحليل.

5.1.4 حجم المعيار الذهبي

يتم اختيار المعيار الذهبي بحجم كبير نسبياً بحيث يغطي معظم الحالات المتوقعة من المحللات الصرفية والنحوية أن تكون قادرة على تحليلها، ويقاس حجم المعيار الذهبي بعدد الكلمات التي يحتويها.

لقد قمنا بتطوير معيار ذهبي مكوَّن من نصوص القرآن الكريم كاملاً، لاستخدامه لفحص المحللات الصرفية في مسابقة (Morphochallenge 2009[4]) لبناء محلل صرفي لعدة لغات من ضمنها اللغة العربية، ويبلغ حجم هذا المعيار 78,004 كلمات مزودة بالتحليل الصرفي الكامل للكلمة، حسب التحليل الصرفي لكلمات القرآن الكريم في قاعدة البيانات الصرفية للقرآن الكريم المطورة بجامعة حيفا (Dror et al, 2004)، وبيين الشكل (10) جزءاً من هذا المعيار الذهبي.

[4] http://www.cis.hut.fi/morphochallenge2009/

بِسْمِ	سم None ب +Prep سم +Noun +Triptotic +Sg +Masc +Gen
اللَّهِ	None None لله +Noun +ProperName +Gen +Def
الرَّحْمَـٰنِ	رحم فَعلَان رحمَان +Noun +Triptotic +Adjective +Sg +Masc +Gen +Def
الرَّحِيمِ	رحم فَعِيل رَحِيم +Noun +Triptotic +Adjective +Sg +Masc +Gen +Def
الْحَمْدُ	حمد فَعل حَمد +Noun +Triptotic +Sg +Masc +Nom +Def
للَّهِ	None None لله +Prep +Noun +ProperName +Gen +Def
رَبِّ	ربب فَعل رَبّ +Noun +Triptotic +Sg +Masc +Pron +Dependent +1P, ربب فَعل رَبّ +Noun +Triptotic +Sg +Masc +Gen
الْعَالَمِينَ	علم فَاعَل عَالَم +Noun +Triptotic +Pl +Masc +Obliquus +Def
الرَّحْمَـٰنِ	رحم فَعلَان رحمَان +Noun +Triptotic +Adjective +Sg +Masc +Gen +Def
الرَّحِيمِ	رحم فَعِيل رَحِيم +Noun +Triptotic +Adjective +Sg +Masc +Gen +Def
مَـٰلِكِ	ملك فَعَلَ مَالك +Verb +Triptotic +ActPart +Sg +Masc +Gen
يَوْمِ	يوم فَعل يَوم +Noun +Triptotic +Sg +Masc +Gen
الدِّينِ	دين فَعل دِين +Noun +Triptotic +Sg +Masc +Gen +Def
إِيَّاكَ	None None إيا +Particle +Pron +Dependent +2P +Sg +Masc
نَعْبُدُ	عبد فَعَلَ نَعبُد +Verb +Imp +Act +1P +Pl +Masc/Fem
وإِيَّاكَ	None None وَ +Particle +Conjunction إيا +Particle +Pron +Dependent +2P +Sg +Masc
نَسْتَعِينُ	عون يَستَفعِلُ نَستَعِين +Verb +Imp +Act +1P +Pl +Masc/Fem, عين يَستَفعِلُ نَستَعِين +Verb +Imp +Act +2P +Pl +Masc/Fem
اهدِنَا	هدي فَعَلَ هد +Verb +Imperative +2P +Sg +Masc +Pron +Dependent +1P +Pl
الصِّرَاطَ	صرط فِعال صِرَاط +Noun +Triptotic +Sg +Masc +Acc +Def
الْمُسْتَقِيمَ	قوم يَستَفعِلُ مُستَقِيم +Verb +Triptotic +ActPart +Sg +Masc +Acc +Def
صِرَاطَ	صرط فِعال صِرَاط +Noun +Triptotic +Sg +Masc +Acc
الَّذِينَ	None None لَذِينَ +Pron +Relative +Pl +Masc
أَنعَمتَ	نعم يَفعَلُ ءَنعمتَ +Verb +Perf +Act +2P +Sg +Masc
عَلَيهِمْ	None None عَلَي +Particle +Pron +Dependent +3P +Pl +Masc
غَيرِ	غير فَعل غَير +Noun +Triptotic +Sg +Masc +Gen
الْمَغضُوبِ	غضب فَعَلَ مَغضُوب +Verb +Triptotic +PassPart +Sg +Masc +Gen +Def
عَلَيهِمْ	None None عَلَي +Particle +Pron +Dependent +3P +Pl +Masc
وَلَا	None None وَ +Particle +Conjunction لَا +Particle +Negative
الضَّالِّينَ	ضلل فَعَلَ ضَالّ +Verb +Triptotic +ActPart +Pl +Masc +Obliquus +Def

الشكل (10) جزءاً من هذا المعيار الذهبي المستخدم في مسابقة (Morphochallenge 2009)

كما يمكن استخدام المعيار الذهبي لمعرفة خصائص المحللات الصرفية والنحوية، وذلك عن طريق مقارنة نتائج تحليلاتها مع الخصائص الصرفية والنحوية للمعيار الذهبي، ومعرفة أي الخصائص الصرفية أو النحوية للكلمة يستطيع المحلل الصرفي أو النحوي تحديدها، وهذه طريقة أخرى لتقييم المحللات الصرفية والنحوية بطريقة وصفية ودون حساب دقتها.

6. الخاتمة

في هذا البحث، قمنا بدراسة المحللات الصرفية والنحوية لمعرفة إمكانية استخدامها في بناء ذخيرة لغوية معنونة بالعناوين الصرفية والنحوية للكلمات، وبدأ البحث بعرض نتائج مقارنة محللات صرفية وخوارزميات استخراج الجذور اعتماداً على معيار ذهبي احتوى ألف كلمة من نصوص القرآن الكريم، وألف كلمة من نصوص الصحف والمجلات أُقتبست من الذخيرة اللغوية العربية (Corpus of Contemporary Arabic (CCA))، ولم تكن نتائج المقارنة إيجابية، حيث أن هذه الأنظمة لم تستطع تحديد التحليل الصحيح لحوالي ربع كلمات الإختبار ، وعليه؛ بدأنا بالبحث عن طرق أخرى لتطوير محلل صرفي قادر على تجاوز الأخطاء التي وقعت بها المحللات الصرفية التي تمت مقارنتها، ولفهم المشكلة بشمولية، فقد أجرينا تحليلاً للجذورالثلاثية لكلمات القرآن الكريم والكلمات

المختلفة المخزنة في القاموس، وأظهرت الدراسة أن حوالي 40% من هذه الجذور الثلاثية هي جذور معتلّة تشكل تحدياً للمحللات الصرفية.

لقد قمنا بتطوير محلل صرفي للغة العربية بالإعتماد على قوائم معدَّة مسبقاً للزوائد والسوابق واللواحق والجذور والأوزان، تم استخراجها من كتب قواعد اللغة العربية و كتب النحو والصرف، ولقد تم التحقق من هذه القوائم عن طريق فحصها ومقارنتها بكلمات الذخائر اللغوية العربية مثل القرآن الكريم والذخيرة اللغوية العربية (Corpus of Contemporary Arabic (CCA)) والذخيرة اللغوية العربية (Penn Arabic Treebank) و نصوص المعاجم العربية الخمسة عشر كذخيرة لغوية رابعة، واحتوت قائمة السوابق على 215 سابقة وقائمة اللواحق على 127 لاحقة، كما احتوت قائمة الأوزان على 2730 وزناً للأفعال و390 وزناً للإسماء.

ولقد طور المحلل الصرفي للتحليل الكلمات وتحديد خواصها اللغوية، وقد ميَّزنا بين العديد من الخصائص اللغوية للكلمات، نطمح أن تكون المحللات الصرفية قادرة على تحديدها، وقد قمنا بتطوير مجموعة العناوين الصرفية والنحوية للغة العربية (Morphological features part-of-speech tag set) والتي يمكن استخدامها في المحللات الصرفية والنحوية وعنونة الذخائر اللغوية، وتتكون عناوين هذه المجموعة من متسلسلة من الرموز طولها تسعة عشر رمزاً، بحيث يمثل كل رمز في موقع معين في العنوان الصرفي والنحوي قيمة لخاصية لغوية للكلمة المحللة.

ولتقييم نتائج المحللات الصرفية المختلفة ومقارنة نتائجها، تم إقتراح بناء المعيار الذهبي لهذا الغرض، بحيث يتكون هذا المعيار من حجم مناسب من الكلمات تغطي معظم الحالات التي يجب على المحللات الصرفية والنحوية على تحديد خصائصها، ويتم اختيار النصوص المكونة للمعيار الذهبي من نصوص متعددة المصادر ولأشكال بحيث تحتوي على نصوص مشكولة كلياً و جزئياً وأخرى غير مشكولة.

المراجع

AL-GHALAYYNI, A.-S. M. الغلاييني", ا. م. (2005) *Jami' Al-Duroos Al-Arabia "جامع الدروس العربية",* Saida - Lebanon, Al-Maktaba Al-Asriyiah "المكتبة العصرية".

ALQRAINY, S. (2008) *A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging. 2008.* Leicester, UK, De Montfort University.

AL-SHALABI, R., KANAAN, G., & AL-SERHAN, H. (2003, December). *New approach for extracting Arabic roots*. Paper presented at the International Arab Conference on Information Technology (ACIT'2003), Alexandria, Egypt.

AL-SULAITI, LATIFA & ATWELL, ERIC (2006). *The design of a corpus of contemporary Arabic*. International Journal of Corpus Linguistics, vol. 11, pp. 135-171. 2006.

BUCKWALTER, T. (2004) Buckwalter Arabic Morphological Analyzer Version 2.0. *Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.*

DAHDAH, A. (1987) *A dictionary of Arabic Grammer in Charts and Tables " معجم قواعد اللغة العربيه – في جداول ولوحات",* Beirut, Lebanon, Librairie du Liban publisher.

DAHDAH, A. (1993) *A dictionary of Arabic Grammatical nomenclature Arabic – English " معجم لغة النحو العربي عربي-انكليزي",* Beirut, Lebanon, Librairie du Liban publishers

DROR JUDITH, SHAHARABANI DUDU, TALMON RAFI & WINTNER SHULY. (2004) *Morphological Analysis of the Qur'an.* Literary and Linguistic Computing, 19(4):431-452, 2004.

LARKEY LEAH. S. AND CONNELL MARGRATE. E. (2001). *Arabic information retrieval at UMass*. In Proceedings of TREC 2001, Gaithersburg: NIST, 2001.

MAAMOURI, M. AND BIES, A. (2004) Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).*

SAWALHA, M. & ATWELL, E. (2008) Comparative evaluation of Arabic language morphological analysers and stemmers. *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics.*

SOUDI, A., BOSCH, A. V. D. & NEUMANN, G. (Eds.) (2007) *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*, Springer Netherlands.

THABET, N. (2004) Stemming the Qur'an. *COLING 2004, Workshop on computational approaches to Arabic script-based languages.August 28,2004.*