

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 سلط 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 العلة 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 المضعّفة)، 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 العلة 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) بالإضافة 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 علة 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 الأخرى التي تندرج تحت هذه الأنواع الثلاثة، وتحديد العدد للكلمة (المفرد أو المثنى أو الجمع) 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 معدة يدوياً تحتوي على معلومات الجذر أو الجذع أو الأوزان، علاوة على ذلك، هناك أسلوب آخر للتحليل الصرفي يعتمد على خوارزميات الذكاء الاصطناعي للذخائر اللغوية المعنونة نحويًا لبناء قاعدة بيانات للكلمات المحللة.

أما الأساليب الإحصائية، فقد تم 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 لا يمكننا توقع نتائج جيدة عند تطبيق هذه الأساليب للغة العربية، وذلك لأن عملية اشتقاق الكلمات في اللغة العربية لا تعتمد على إضافة اللواحق فقط للكلمة، بعض هذه الأساليب 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 اللواحق واللواحق المستخدمة بكثرة (Thabet, 2004).

ومن المحللات الصرفية الشائعة الاستخدام للغة العربية، المحلل الصرفي للغة العربية (Tim Buckwalter Morphological Analyzer¹) الذي يعتمد على قوائم للكلمات ومعلوماتها الصرفية تم إعدادها يدوياً، هذه القوائم تحتوي على قائمة الجذع وقائمة السوابق واللواحق، وقد تم إضافة جداول الصواب والخطأ لتحديد التوافق الصحيحة التي تجمع السوابق باللواحق للكلمات (Thabet, 2004) (Buckwalter, 2004).

ومن أنظمة استخراج الجذور، نظام شيرين حوجا لاستخراج جذر الكلمة (Khoja's Stemmer²) 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 والرابعة، 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 معدة يدوياً، بل يعتمد على حسابات رياضية بتعيين أوزان رقمية لأحرف الكلمة وضرب هذه الأوزان بمواقع حروفها، بحيث تعطى الأوزان في بداية الكلمة ونهايتها أوزاناً أعلى من أوزان حروف الكلمة في وسطها، ثم يقوم النظام باختيار الحروف ذات الأقل وزناً كحروف الجذر تلك الكلمة، وقد قسّم هذا النظام الحروف العربية إلى مجموعتين، تحتوي المجموعة الأولى على الحروف التي لا تشكل أحد حروف السوابق أو اللواحق وتعطى الوزن صفر (0)، أما المجموعة الثانية فتضم الحروف التي تتكون منها السوابق واللواحق والمجموعة بكلمة (سألتمونيها) وتعطى هذه الحروف أوزاناً مختلفة.

2. المحلل الصرفي للغة العربية

لقد قمنا بتطوير محلاً 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 تعرضت نتائجاً لتحليل الصرفي للغة العربية إلا أن هذه الأنظمة غير متاحة للباحثين والمستخدمين، لذلك اقتصرنا 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)، وللمقارنة 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 المحللات 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 مختصين 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 المرجوة لاعتمادها في تطبيقات التحليل النحوي (Part-of-Speech tagging) أو استخدامها في تدقيق بنية الجملة (Parsing) حيث أن الأخطاء الناتجة من هذه المحللات الصرفية سوف تنتقل إلى 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 إلى اثنتان وعشرين مجموعة اعتماداً على التركيب الداخلي للجذر نفسه، من حيث خلو الجذر من أحرف العلة 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 مصدر آخر هو القاموس الذي تم بناءه اعتماداً على نصوص

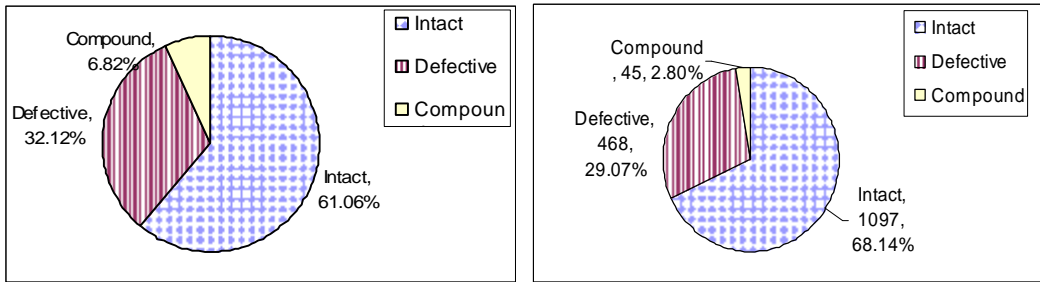
خمسة عشر معجماً عربياً اخترنا منها حوالي 376,167 كلمة مختلفة ثلاثية الجذر، الجدول (1) والجدول (2) يبينان نتائج التحليل لجميع فئات الجذور الثلاثية، وأظهرت النتائج أن حوالي 68% من الجذور الثلاثية هي جذور صحيحة (سالم أو مضعّف أو مهموز) و 61% من كلمات القرآن الكريم تنتمي لهذه المجموعة، وأن 29% من الجذور الثلاثية جذور معتلة (تحتوي على حرف أو حرفين من أحرف العلة) وتشكل الكلمات التي تنتمي إلى هذه المجموعة ما نسبته 32% من كلمات القرآن الكريم، أما المجموعة الثالثة والتي تضم الجذور الثلاثية التي تحتوي على حرف أو حرفي علة 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

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

جدول (2) نتائج تحليل الكلمات المختلفة للقاموس

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

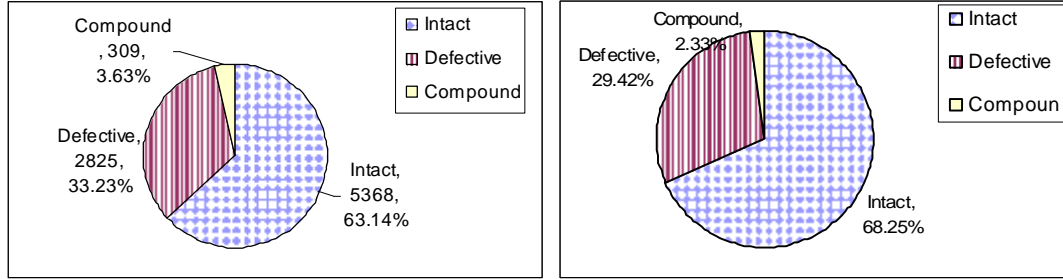


الشكل (1) توزيع الجذور (بمين) والكلمات (يسار) لجذور وكلمات القرآن الكريم

الجدول (3) توزيع جذور وكلمات القرآن الكريم

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

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 معتلة و تنتمي لهذه المجموعة 29% من كلمات القاموس، وأن حوالي 4% من الجذور هي جذور معتلة 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

الفئة	الجذور	النسبة المئوية	الكلمات المختلفة للقاموس	النسبة المئوية
الجذور الصحيحة	5368	63.14%	256,735	68.25%
الجذور المعتلة	2825	33.23%	110,666	29.42%
الجذور المعتلة 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 المدخل

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 واحدة للبيانات، فبداية يقوم برنامج خاص بتقطيع كلمات النص المدخل إلى كلمة عربية مشكولة كلياً أو جزئياً أو غير مشكولة، أو إلى رقم أو عملة أو علامة ترقيم، بعدها يقوم البرنامج بمعالجة الكلمات العربية المستخرجة من النص المدخل، وتقوم عملية المعالجة بإرجاع الحروف المضعفة والمد إلى أصلها، حيث يتم استبدال الحرف المضعف والشدة الظاهرة عليه بحرفين الأول ساكن والثاني متحرك بنفس حركة الحرف الأصلي، فمثلاً؛ كلمة (وَصَّى) تصبح (وَصَصَى)، وكذلك يتم استبدال المد (آ) بحرفي الهمزة والألف، فكلمة (آمَنُوا) تصبح (ءَامَنُوا).

ولأنه يمكن أن تظهر حركة واحدة (حرف علة قصير) على أي حرف من حروف الكلمة، تم اعتماد هيكلية موحدة للكلمات بحيث تتكون من مصفوفة أحادية تضم الحروف والحركات التي تظهر عليها إن وجدت، ويتم تخزين الحرف الأول من الكلمة في الموقع الأول للمصفوفة يليه حركة الحرف (حرف العلة القصير) في الموقع الثاني وهكذا لجميع حروف الكلمة، ويوضح الشكل (3) الهيكلية المستخدمة لتخزين كلمتي (وَصَصَى و ءَامَنُوا)، 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
وَصَصِي	و	ص	ص	°	ص	ص	ى	-				
ءامنوا	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 النص، وتبلغ نسبة هذه الكلمات في أي نص من نصوص اللغة العربية حوالي 40% من إجمالي عدد كلمات ذلك النص، ويقوم المحلل الصرفي 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
فتحن	لهؤلاء	بالذي	فعلينا	عنده
هي	في هؤلاء	للذي	عليها	عندها
فهي	فل هؤلاء	فكالذي	فعلينا	عندهما

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

2.2.3 الزوائد والسوابق واللاحق

اعتماداً على كتب قواعد اللغة العربية فقد تم حصر الزوائد في بداية الكلمة (Proclitics) (كحروف العطف والجر والنداء وأدوات التعريف)، والسوابق (Prefixes)، واللاحق (Suffixes) والزوائد في نهاية الكلمة (Enclitics) (كالضمائر المتصلة)، وقد زوّدت هذه المعلومات إلى برنامج مؤلّد، حيث تم 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 واللاحق، بلغ عدد السوابق 215 سابقة كما بلغ عدد اللاحق 127 لاحقة، ويبين الجدول (5) والجدول (6) جزءاً من هذه السوابق واللاحق مع مثال على كلٍّ منهما والعنوان الصرفي³ المعين 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 اللاحق، 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 مطابقتها السابقة واللاحقة لجزئها معاً، ويوضح الجدول (7) عملية مطابقة السوابق واللاحق لأجزاء الكلمة المحللة.

راجع الجزء 4 الخصائص اللغوية للكلمة العربية وتركيب العناوين الصرفية والنحوية³

الجدول (5) جزء من قائمة السوابق و العنواين الصرفية المعينة 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	والسما و	و	p--t-----	ال	r---d----d-----		
ولت	ولتجدكم و	و	p--t-----	ل	p--y-----	ت	r---s-nus-----
سن	سنجزى س	س	p--i-----	ن	r---s-npf-----		
ويم	ويمراجعة و	و	p--t-----	ب	p--r----g-----	م	r---f-----

الجدول (6) جزء من قائمة اللواحق و العنواين الصرفية المعينة لها

اللاحقة	مثال	ج1	العنوان الصرفي	ج2	العنوان الصرفي	ج3	العنوان الصرفي
اتية	ات		r---l-fp-??-?-?---	ي	r---j-----	ة	r---b-fs-??-?-?---
تموها	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	ي	عملون		تحليل مقبول
يَعْمَلُونَ	ي	يعملو	ن	تحليل غير مقبول
يَعْمَلُونَ	ي	يعمل	ون	تحليل مقبول
يَعْمَلُونَ	ي	يعم	لون	تحليل غير مقبول
يَعْمَلُونَ	ي	يع	ملون	تحليل غير مقبول
يَعْمَلُونَ	ي	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	ي	ملون		تحليل غير مقبول
يَعْمَلُونَ	ي	ملو	ن	تحليل غير مقبول
يَعْمَلُونَ	ي	مل	ون	تحليل غير مقبول
يَعْمَلُونَ	ي	م	لون	تحليل غير مقبول
يَعْمَلُونَ	ي	لون		تحليل غير مقبول
يَعْمَلُونَ	ي	لو	ن	تحليل غير مقبول
يَعْمَلُونَ	ي	ل	ون	تحليل غير مقبول
يَعْمَلُونَ	ي	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

تتم 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 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?	أَفْعَلَاوِي	nw----??-??----?qt-
فَعَلْنَا	v-p---npf---an?-st?	أَفْعِيلَال	nw----??-??----?qt-
فَعَلْتِ	v-p---mss---an?-st?	فَاعُولَاء	nw----??-??----?qt-
فَعَلْتِ	v-p---fss---an?-st?	فُعْلَمَلَان	nw----??-??----?qt-
فَعَلْتُمَا	v-p---nds---an?-st?	فُعَيْلَاء	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 المعين لهذا الوزن، ويبين الشكل (6) عملية استخراج الجذر بهذه الطريقة.

Word	أَحْسِبَ	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				
أأعلوا			أأ			من			وا				
Candidate indices list = [0, 2, 3]													
Pattern			Prefix			Stem			Suffix				
فاعلوا						أامن			وا				
Word	الْعَلِيمِ	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			علم							

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) (Alrainy, 2008) حيث تعمل هذه الخوارزمية على مطابقة الكلمة مع وزنها للكلمات المشكولة جزئياً بالحركة الظاهرة على آخر الكلمة فقط وبدون إجراء أي تحليل للسوابق واللاحق.

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، وستطبق الأوزان (فَعْل، فَعْل، فَعْل، فَعْل، فَعْل، فَعْل، فَعْل، فَعْل، فَعْل، فَعْل) 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 الصرفي
يَعْمَلُونَ	يَفْعَلُونَ	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	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-st?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ian?-at?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ipn?-tt?
يَعْمَلُونَ	يَفْعَلُونَ	v-c---mpt--ipn?-at?
كتب	فَعَلَ	v-p---mst---an?-st?
كتب	فَعَلَ	v-p---mst---an?-st-
كتب	فَعَلَ	v-p---mst---an?-st-
كتب	فَعَلَ	v-p---mst---an?-st-
كتب	فَعَلَ	v-p---mst---pn?-tt-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	nw----??-??----?st-
كتب	فَعَلَ	ny----??-??----?st-

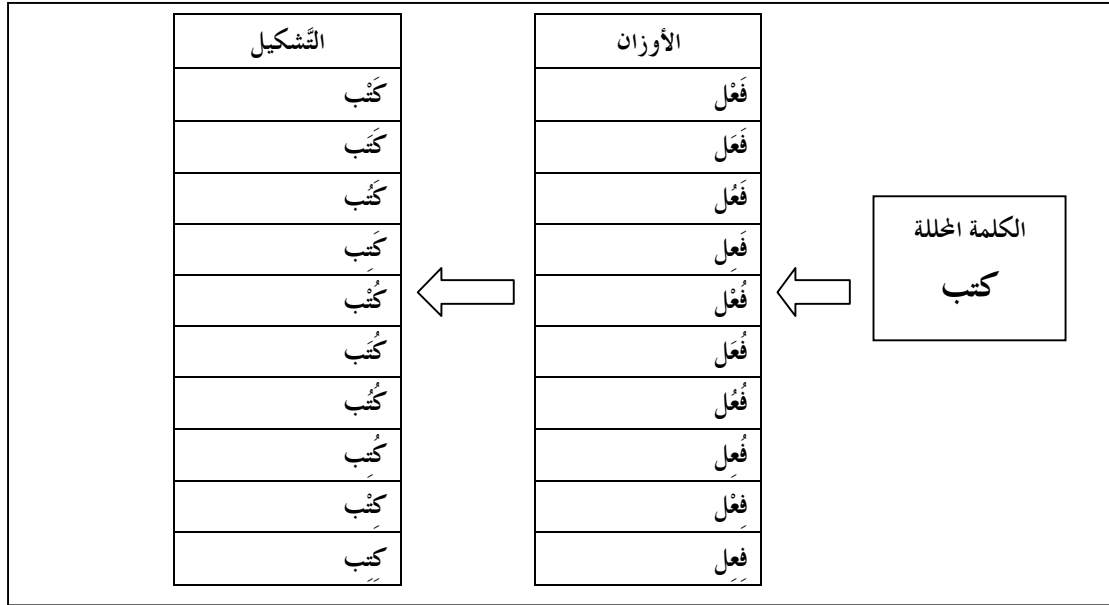
الشكل (7) عملية استخراج الأوزان للكلمات باستخدام الطريقة الثانية.

2.2.6 التَّشْكِيل

يعتبر التَّشْكِيل أحد الخصائص المهمة للكلمة العربية حيث يفيد التشكيل في تحديد خصائص لغوية أخرى للكلمة، فوجود التَّشْكِيل 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 اللبس 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 إلى 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

قسّم 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 الجرّ والنسخ والعطف والجزم والتّصّب 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، والتّوكيد (مؤكد أو غير مؤكد)، والفعل اللازم أو المتعدي لمفعول واحد أو مفعولين إثنين أو ثلاثة مفاعيل، ومن الخصائص الأخرى التي تمَّ تحديدها للكلمة والتي تصف بنيتها، الجرّد 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 الرئيسية
2	أقسام الكلام الفرعية (الاسم)
3	أقسام الكلام الفرعية (الفعل)
4	أقسام الكلام الفرعية (الحرف)
5	أقسام الكلام الفرعية (أخرى)
6	أقسام الكلام الفرعية (علامات الترقيم)
7	الجنس
8	العدد
9	الشخص
10	الحالة الإعرابية للاسم
11	التعريف والتنكير
12	الحالة الإعرابية للفعل
13	المعلوم والمجهول
14	التوكيد
15	التعدية
16	العاقل وغير العاقل
17	المجرد والمزيد
18	عدد الحروف الأصلية
19	تركيب أحرف الفعل

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، فمثلاً 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) إلى المذكر والحرف (f) إلى 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 كليهما 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 عطف منصوب
وَصَّى	v-p---npf--iano-at&	فعل فعل ماضي مؤنث/مذكر جمع سالم المُخاطَب الرَّفْع مَبْنِيٍّ لِلْمَعْلُومِ غَيْرِ مُؤَكَّدٍ مُتَعَدِّيٍّ إِلَى مَفْعُولٍ وَاحِدٍ مَزِيدٍ بِحَرْفِ ثَلَاثِي اللَّفِيفِ الْمَفْرُوقِ
نَا	p--&---p-n-----	حرف جمع سالم مرفوع
الْإِنْسَانَ	nq----np-ad-----bt-	اسم اسم جنس مؤنث/مذكر جمع سالم منصوب مَعْرِفَةٌ مَزِيدٌ بِحَرْفَيْنِ ثَلَاثِي
بِ	p--r-----g-----	حرف جر مجرور
وَالَّذِي	nw---nd-gd-----at-	اسم اسم معرب مؤنث/مذكر مثنى مجرور مَعْرِفَةٌ مَزِيدٌ بِحَرْفِ ثَلَاثِي
هَ	p--&-----g-----	حرف مجرور
حُسْنًا	no----nu-ai-----st-	اسم مصدر مؤنث/مذكر غير محدد منصوب نَكْرَةٌ مُجَرَّدٌ ثَلَاثِي
المثال الثاني جملة من الذخيرة اللغوية العربية (Penn Arabic Treebank)		
تم	v-p---mst---ano-stb	فعل فعل ماضي مذكر مفرد الْعَائِبِ مَبْنِيٍّ لِلْمَعْلُومِ غَيْرِ مُؤَكَّدٍ مُتَعَدِّيٍّ إِلَى مَفْعُولٍ وَاحِدٍ مُجَرَّدٌ ثَلَاثِي مَضَعَفٌ
اعداد	no----ms-ni-----bt-	اسم مصدر مذكر مفرد مرفوع نَكْرَةٌ مَزِيدٌ بِحَرْفَيْنِ ثَلَاثِي
الوثائق	nw---fb-ad---ntt-	اسم اسم معرب مؤنث جمع تكسير منصوب مَعْرِفَةٌ غَيْرِ عَاقِلٍ مَزِيدٌ بِثَلَاثَةِ حُرُوفٍ ثَلَاثِي
المتوفرة	ns---fs-gd-----tt-	اسم صفة مشبهة مؤنث مفرد مجرور مَعْرِفَةٌ مَزِيدٌ بِثَلَاثَةِ حُرُوفٍ ثَلَاثِي
ب	p--r-----g-----	حرف جر مجرور
كثرة	ns---fb-gi-----at-	اسم صفة مشبهة مؤنث جمع تكسير مجرور نَكْرَةٌ مَزِيدٌ بِحَرْفِ ثَلَاثِي
حول	nh---nu-ai-----st-	اسم ظرف مؤنث/مذكر غير محدد منصوب نَكْرَةٌ مُجَرَّدٌ ثَلَاثِي
أول	nm---nu-gi-----st-	اسم اسم عدد مؤنث/مذكر غير محدد مجرور نَكْرَةٌ مُجَرَّدٌ ثَلَاثِي
رحلة	nw---fs-gi-----at-	اسم اسم معرب مؤنث مفرد مجرور نَكْرَةٌ مَزِيدٌ بِحَرْفِ ثَلَاثِي
طيران	no---nu-gi-----bt-	اسم مصدر مؤنث/مذكر غير محدد مجرور نَكْرَةٌ مَزِيدٌ بِحَرْفَيْنِ ثَلَاثِي
عثمانية	nr---fu-gi---htq-	اسم اسم منسوب مؤنث غير محدد مجرور نَكْرَةٌ عَاقِلٍ مَزِيدٌ بِثَلَاثَةِ حُرُوفٍ رِبَاعِي
فوق	nh---nu-ai-----st-	اسم ظرف مؤنث/مذكر غير محدد منصوب نَكْرَةٌ مُجَرَّدٌ ثَلَاثِي
البلاد	nw---fb-gd---nat-	اسم اسم معرب مؤنث جمع تكسير مجرور مَعْرِفَةٌ غَيْرِ عَاقِلٍ مَزِيدٌ بِحَرْفِ ثَلَاثِي
العربية	nr---fu-gd---hbt-	اسم اسم منسوب مؤنث غير محدد مجرور مَعْرِفَةٌ عَاقِلٍ مَزِيدٌ بِحَرْفَيْنِ ثَلَاثِي

الشكل (9) أمثلة لجملتان تمت عنونة كليهما 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 عدّة أنظمة أو خوارزميات طورت لحل مشكلة معيّنة، وتُظهر المعايير الحالات التي تنتج أو تفشل الأنظمة المُقيّمة بتحديد التحليل المناسب للمدخلات، وتستخدم المعايير الذهبية لإيجاد أوجه الشبه أو الاختلاف في نتائج التحليل مبينة الحالات التي تتفق عليها والتي تختلف فيها الأنظمة المحوسبة.

ولبناء معيار ذهبي لتقييم الأنظمة الحوسبية، يجب علينا بداية تحديد موضوع المشكلة التي تقوم هذه الأنظمة على حلها، كما يجب تحديد الذخيرة اللغوية التي ستستخدم لبناء المعيار الذهبي، وتحديد تنسيق 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 معظمها على نصوص مقتبسة من المجالات والصحف، ولكن لبناء معيار ذهبي واسع التطبيق يجب علينا اختيار نصوص عربية من مصادر وأشكال ومجالات متعددة، ومن نصوص مشكولة كلياً أو جزئياً أو غير مشكولة، ومن الذخائر اللغوية المقترحة 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 مجلداً 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 مكوّن من نصوص القرآن الكريم كاملاً، لاستخدامه لفحص المحللات الصرفية في مسابقة (Morphochallenge 2009⁴) 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)، وبين الشكل (10) جزءاً من هذا المعيار الذهبي.

⁴ <http://www.cis.hut.fi/morphochallenge2009/>

None +Noun +ProperName +Gen +Def	الله
Triptotic +Adjective +Sg +Masc +Gen +Def+Noun	الرَّحْمَنُ
Triptotic +Adjective +Sg +Masc +Gen +Def+Noun	الرَّحِيمُ
Triptotic +Sg +Masc +Nom +Def+Noun	الْحَمْدُ
None None ل +Prep+Noun +ProperName +Gen +Def	لِلَّهِ
Sg +Masc +Gen+Noun +Triptotic+رب فعل رب +Triptotic +Sg +Masc +Pron +Dependent +1P +Sg+Noun	رَبِّ
Triptotic +Pl +Masc +Obliquus +Def+Noun	الْعَالَمِينَ
Triptotic +Adjective +Sg +Masc +Gen +Def+Noun	الرَّحْمَنُ
Triptotic +Adjective +Sg +Masc +Gen +Def+Noun	الرَّحِيمُ
Triptotic +ActPart +Sg +Masc +Gen+Verb	مَالِكِ
Triptotic +Sg +Masc +Gen+Noun	يَوْمِ
Triptotic +Sg +Masc +Gen +Def+Noun	الدِّينِ
None None ياء +Particle +Pron +Dependent +2P +Sg +Masc	إِيَّاكَ
Act +1P +Pl +Masc/Fem+Verb +Imp	تَعْبُدُ
None None و +Particle +Conjunction+ياء +Particle +Pron +Dependent +2P +Sg +Masc	وإِيَّاكَ
P +Pl +Masc/Fem1+Verb +Imp +Act+عن يستعمل لمتعين +Verb +Imp +Act +1P +Pl +Masc/Fem+عن يستعمل لمتعين	نَسْتَعِينُ
Imperative +2P +Sg +Masc +Pron +Dependent +1P +Pl+Verb	اهدنا
Triptotic +Sg +Masc +Acc +Def+Noun	الصِّرَاطِ
Verb +Triptotic +ActPart +Sg +Masc +Acc +Def	المُسْتَقِيمِ
Triptotic +Sg +Masc +Acc+Noun	صِرَاطِ
None None للذين +Pron +Relative +Pl +Masc	الَّذِينَ
Perf +Act +2P +Sg +Masc+Verb	أَنْعَمْتَ
None None علي +Particle +Pron +Dependent +3P +Pl +Masc	عَلَيْهِمْ
Triptotic +Sg +Masc +Gen+Noun	غَيْرِ
Triptotic +PassPart +Sg +Masc +Gen +Def+Verb	المَغْضُوبِ
None None علي +Particle +Pron +Dependent +3P +Pl +Masc	عَلَيْهِمْ
None None و +Particle +Conjunction+ل +Particle +Negative	وَلَا
Triptotic +ActPart +Pl +Masc +Obliquus +Def+Verb	الضَّالِّينَ

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 وحوارزميات 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 للجذور الثلاثية لكلمات القرآن الكريم والكلمات

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

لقد قمنا بتطوير محلل صرفي للغة العربية بالإعتماد على قوائم معدة مسبقاً للزوائد والسوابق والوواحق 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 لتحليل الكلمات وتحديد خواصها اللغوية، وقد ميّزنا 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 والأشكال بحيث تحتوي على نصوص مشكولة كلياً و جزئياً وأخرى غير مشكولة.

المراجع

- AL-GHALABINI, A.-S. M. "الغلابيني ا.م.", Jami' Al-Duroos Al-Arabia "جامع الدروس العربية", Saida - Lebanon, Al-Maktaba Al-Asriyah "المكتبة العصرية".
- ALQRAINI, 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 العربي - 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 Computational 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.

الملحق 1: مجموعة العناوين الصرفية والنحوية للغة العربية (Morphological features part-of-speech tag set)

1. أقسام الكلام الرئيسية

Table 1: Main part-of-speech category attributes and tags

Position	Feature Name		Tag
1	Main Part-of-Speech	أقسام الكلام الرئيسية	
	Noun	اسم	n
	Verb	فعل	v
	Particle	أداة	p
	Residual	أخرى	r
	Punctuation	علامة ترقيم	u

2. أقسام الكلام الفرعية (الإسم)

Table 2. Part-of-Speech subcategory of noun and its attributes and tags

Position	Feature Name		Tag
2	Part-of-Speech of Noun	أقسام الكلام الفرعية (الإسم)	
	Original noun	مصدر	o
	Personal noun	ضمير	p
	Demonstrative noun	اسم إشارة	d
	Conjunctive noun	اسم موصول	c
	Interrogation noun	اسم استفهام	g
	Conditional noun	اسم شرط	b
	Allusive noun	كناية	a
	Adverb	ظرف	h
	Agent noun	اسم فاعل	j
	Patient noun	اسم مفعول	k
	Adjective	صفة مشبهة	s
	Superlative	افعل التفضيل	v
	Noun of place	اسم مكان	l
	Noun of time	اسم زمان	t
	Noun of instrument	اسم آلة	i
	Proper noun	اسم علم	n
	noun of genus	اسم جنس	q
	Numeral noun	اسم عدد	m
	Verbal noun	اسم فعل	&
	Five nouns	الأسماء الخمسة	f
	Declined noun	اسم معرب	w
	Relative noun	اسم منسوب	r
	Noun of Diminution	اسم تصغير	x
	Noun/Form of superlative	صيغة مبالغة	y
	Noun of place/time	مكان/زمان	z
	Called noun	اسم منادى	*
	Noun of "Indeed"	اسم ان	\$
	Noun of "to be"	اسم كان	b
	Excluded noun	اسم مستثنى	e

3. أقسام الكلام الرئيسية (الفعل)

Table 3. Part-of-Speech subcategory of verb and its attributes and their tags

Position	Feature Name		Tag
3	Part-of-Speech of Verb	أقسام الكلام الفرعية (الفعل)	
	Perfect/Past verb	فعل ماضي	p
	Imperfect/Progress verb	فعل مضارع	c
	Imperative verb	فعل امر	i

4. أقسام الكلام الرئيسية (الحروف)

Table 4. Part-of-speech category of particles attributes and their tags

Position	Feature Name		Tag
4	Part-of-Speech of Particles	أقسام الكلام الفرعية (الحروف)	
	Letter of elision	حرف جزم	e
	letter of opening	حرف نصب	o
	letter of reduction	حرف جر	r
	letter of annulment	حرف نسخ	a
	Letter of attraction	حرف عطف	t
	Letter of partial opening	حرف النصب الفرعي	p
	Letter of call	حرف نداء	c
	letter of exclusion	حرف استثناء	x
	Interrogative letter	حرف استفهام	g
	Introduction letter	حرف استقبال	i
	Letter of Causality	حرف تعليل	y
	Letter of Negation	حرف نفي	n
	Letter of Oath	حرف قسم	h

5. أقسام الكلام الرئيسية (أخرى)

Table 5. Part-of-speech category of Residuals attributes and their tags

Position	Feature Name		Tag
5	Subcategory of Residuals	أقسام الكلام الفرعية (أخرى)	
	Prefix / part of the word's pattern	زيادة في اول الكلمة	f
	Suffix / part of the word's pattern	زيادة في آخر الكلمة	u
	Relative Pronoun	ضمير متصل	&
	Plural Letters	حروف الجمع	l
	Dual Letters	حروف الثني	m
	Taah Marbouta	تاء مربوطة	b
	Relative yaa	ياء النسبة	j
	Tanween	تنوين	k
	Feminine Taa	تاء التانيث	q
	Doar Taa	تاء الفاعل	w
	Protection Noon	نون الوقاية	*
	Emphasize Noon	نون التوكيد	z
	Progress letter	حرف مضارعة	s
	Imperative letter	حرف امر	v
	Definitive Article	اداة تعريف	d

6. أقسام الكلام الرئيسية (علامات الترقيم)

Table 6. Part-of-speech category of punctuation attributes and their tags

Position	Feature Name		Tag
6	Subcategory of Punctuation	أقسام الكلام الفرعية (علامة ترقيم)	
	Full stop	نقطة (.)	s
	Comma	فاصلة (,)	c
	Colon	نقطتان (:)	n
	Semi colon	فاصلة منقوطة (;)	l
	Parenthesis	القوسان (())	p
	Square brackets	القوسان الحاصرتان ([])	b
	Quotation mark	علامة الاقتباس (" ")	t
	Dash	الشرطة المعترضة (-)	d
	Question mark	علامة سؤال (?)	q
	Exclamation mark	علامة تعجب (!)	e
	Elimination mark	علامة حذف (...)	i

7. الجنس

Table 7. Gender category attributes and tags

Position	Feature Name		Tag
7	Gender	الجنس	
	Masculine	مذكر	m
	Feminine	مؤنث	f
	Neuter	مؤنث/مذكر	n

8. العدد

Table 8. Morphological feature of number category attributes and their tags

Position	Feature Name		Tag
8	Number	العدد	
	Singular	مفرد	s
	Dual	مثنى	d
	Sound Plural	جمع سالم	p
	Broken Plural	جمع تكسير	b
	undefined	غير محدد	u

9. الإسناد

Table 9. Morphological feature of person category attributes and their tags

Position	Feature Name		Tag
9	Person	الشخص	
	First Person	المخاطب	f
	Second Person	التكلم	s
	Third Person	الغائب	t

10. الحالة الإعرابية للاسم

Table 10. The morphological feature of case category attributes and their tags

Position	Feature Name		Tag
10	Case	الحالة الإعرابية للاسم	
	Nominative	مرفوع	n
	Accusative	منصوب	a
	Genitive	مجرور	g

11. المعرفة والتكررة

Table 11. The morphological feature of definiteness category attributes and their tags.

Position	Feature Name		Tag
11	Definiteness	المعرفة والتكررة	
	Definite	معرفة	d
	Indefinite	نكرة	i

12. الحالة الإعرابية للفعل

Table 12. The morphological feature of mood category attributes and their tags

Position	Feature Name		Tag
12	Mood	الحالة الإعرابية للفعل	
	Indicative	الرّفع	i
	Subjunctive	النّصب	s
	Imperative/Jussive	الجزم	j

13. المبني للمعلوم و المبني للمجهول

Table 13. The morphological feature of voice category attributes and their tags.

Position	Feature Name		Tag
13	Voice	المعلوم والمجهول	
	Active voice	مبني للمعلوم	a
	Passive voice	مبني للمجهول	p

14. المؤكّد وغير المؤكّد

Table 14. The morphological feature of emphasize category attributes and their tags

Position	Feature Name		Tag
14	Emphasize	المؤكّد وغير المؤكّد	
	Not emphasized	غير مؤكّد	n
	Must be emphasized	يجب التأكيد	m
	Emphasizing is allowed	مسموح التأكيد	a

15. اللازم والمتعدي

Table 15. The morphological feature of transitivity category attributes and their tags

Position	Feature Name		Tag
15	Transitivity	اللازم والمتعدي	
	Intransitive	لازم	i
	Transitive to one object	متعدي الى مفعول واحد	o
	Transitive to two objects	متعدي الى مفعولين	b
	Transitive to three objects	متعدي الى ثلاثة مفاعيل	t

16. العاقل وغير العاقل

Table 16. The morphological feature of humanness category attributes and their tags

Position	Feature Name		Tag
16	Humanness	العاقل وغير العاقل	
	Human	عاقل	h
	Non-human	غير عاقل	n

17. المجرد والمزيد

Table 17. The stripped or augmented category attributes and their tags.

Position	Feature Name		Tag
17	Stripped or augmented	المجرد والمزيد	
	Stripped	مجرد	s
	Augmented by one letter	مزيد بحرف	a
	Augmented by two letters	مزيد بحرفين	b
	Augmented by three letters	مزيد بثلاثة حروف	t

18. عدد أحرف الجذر

Table 18. The root letters category attributes and their tags.

Position	Feature Name		Tag
18	Root letters	عدد أحرف الجذر	
	Tri-literal	ثلاثي	t
	Quad-literal	رباعي	q
	Quint-literal	خماسي	f

Table 19. The verb internal structure category attributes and their tags.

Position	Feature Name	Tag
19	Verb Internal Structure	بُنية الفعل
	Intact	صحيح
	doubled	مضعف
	1 st letter Hamza	مهموز الفاء
	1 st letter hamza & doubled	مهموز الفاء مضعف
	1 st & 3 rd letters hamza	مهموز الفاء ومهموز اللام
	2 nd letter hamza	مهموز العين
	3 rd letter hamza	مهموز اللام
	1 st letter waw	المثال الواوي
	1 st letter waw & doubled	المثال الواوي المضعف
	1 st letter waw & second letter hamza	المثال الواوي مهموز العين
	1 st letter waw & 3 rd letter hamza	المثال الواوي مهموز اللام
	1 st letter yaa'	المثال اليائي
	1 st letter yaa' & doubled	المثال اليائي المضعف
	1 st letter yaa' & 2 nd letter Hamza	المثال اليائي مهموز العين
	2 nd letter waw	الاجوف الواوي
	2 nd letter waw & 1 st letter Hamza	الاجوف الواوي المهموز الفاء
	2 nd letter waw & 3 rd letter Hamza	الاجوف الواوي المهموز اللام
	2 nd letter yaa'	الاجوف اليائي
	2 nd letter yaa' & 1 st letter Hamza	الاجوف اليائي المهموز الفاء
	2 nd letter yaa' & 3 rd letter Hamza	الاجوف اليائي المهموز اللام
	3 rd letter waw	الناقص الواوي
	3 rd letter waw & 1 st letter Hamza	الناقص الواوي المهموز الفاء
	3 rd letter waw & 2 nd letter Hamza	الناقص الواوي المهموز العين
	3 rd letter yaa'	الناقص اليائي
	3 rd letter yaa' & first letter Hamza	الناقص اليائي المهموز الفاء
	3 rd letter yaa' & 2 nd letter Hamza	الناقص اليائي المهموز العين
	Adjacent vowels	اللفيف المقرون
	2 nd letter & 3 rd letters are vowels & 1 st letter Hamza	اللفيف المقرون المهموز الفاء
	1 st letter & 3 rd letter are vowels	اللفيف المفروق
	1 st letter & 3 rd letter are vowels & 2 nd letter Hamza	اللفيف المفروق المهموز العين