محلل صرفي عربي موجَّه بالتطبيقات

An Application Oriented Arabic Morphological Analyzer

م. رياض سنبل د. ندى غنيم د. محمد سعيد دسوقي

يقوم نظام التحليل الصرفي بتحليل الكلمة العربية والمعالجة الصرفية لها. وهو يقبل في مدخله كلمةً مشكولةً كلياً أو جزئياً أو غير مشكولة، ويعطي في مخرجه جميع الوجوه الممكنة للسوابق واللواحق والجذور والجذوع والأوزان، ويرتِّب النظام المنجز هذه الحلول حسب "جودتها" بعد أن يقدِّر درجة وثوقيَّة معينة لكل حل.

يعتمد النظام على خوارزمية مبتكرة جعلته يصل إلى وثوقية تتجاوز 97.7% دون الاعتماد على مواد مخزنة، وبسرعة معالجة يمكن أن تصل إلى أكثر من خمسة أضعاف الأنظمة المشابهة (تصل إلى أكثر من ٥٠ ألف كلمة بالثانية في بعض الحالات). ويحقق النظام المنفَّذ مرونة كبيرة تجعله قابلاً للاستخدام في مختلف المجالات ومختلف أنواع التطبيقات، مثل تطبيقات المعالجة اللغوية كالمعاجم وأنظمة تعليم اللغة العربية، وتطبيقات البحث عن المعلومات، ومحركات البحث، وتطبيقات التنقيب في النصوص والمعطيات، وتطبيقات الترجمة الآلية، وغيرها. وقد تم تحقيق ذلك من خلال التحكم بنقطة توازن النظام بين سرعة المعالجة ووثوقية النتائج وشموليَّة الحلول (مدى الحصول على كل التحليلات الممكنة للكلمة)، وتفعيل أو عدم تفعيل معالجة الحالات الخاصة كالإعلال (بالقلب والحذف) والإدغام (الشدة).

يلخِّص الجدول التالي وسائط التحكم المستخدمة في المحلل الصرفي وعددها ٩، والتي ينتج عنها ١٨ نسخة من نظام التحليل الصرفي. تضيف كل نسخة إمكانية معالجة جديدة على النسخة السابقة لها في الترقيم. فالنسخة **Rn** في الجدول تعني أن الوسائط من ١ .. **n** مفعَّلة (تأخذ قيم **true**). أما الإشارة **+** فهي تعني أنَّ النسخة تأخذ الحل الأفضل (ذو الاحتمال الأعلى) من الحلول الممكنة.

وصف النسخة	1	2	3	4	5	6	7	8	+	اسم النسخة
	Stp	Frg	Pat	Rot	Ebd	Ela	Shd	Elm	Bst	
النسخة الأولى دون معالجة الحالات الخاصة										R0
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R0										R0+
إضافة اختبار الكلمات الخاصة stop words										R1
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R1										R1+
إضافة اختبار الكلمات الأجنبية والمعرَّبة										R2
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R2										R2+
إضافة اختبار وجود الجذر في قائمة الجذور										R3
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R3										R3+
إضافة اختبار الوزن										R4
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R4										R4+
إضافة معالجة الإبدال										R5
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R5										R5+
إضافة معالجة الإعلال بالقلب										R6
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R6										R6+
إضافة معالجة الإدغام (الشدة)										R7
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R7										R7+
إضافة معالجة الإعلال بالحذف										R8
اختيار الحل الأفضل من بين الحلول الممكنة في النسخة R8										R8+

يجب أن لا ننسى أنَّ هناك معايير أخرى مثل سرعة المعالجة، وعدم الاعتماد على الموارد اللغوية المخزنة، إضافةً إلى جودة الحلول الناتجة ووثوقية النتائج. وهذا ما دفعنا إلى اعتماد مبدأ وسائط التحكم التي نستطيع بها الموازنة بين هذه المعايير واختيار الأفضل منها حسب التطبيق.

يوفر النظام عدداً متنوعاً من الحالات، فمثلاً الحالات (**R3 ... R6**) تتوافق مع متطلبات الأنظمة العالية الوثوقيّة (كالمراحل الأخرى من المعالجة الحاسوبيَّة، الشكل الآلي للنصوص،...). وقد بينت الاختبارات أن هذه الحالات تحقق وثوقيّة تُجاري أو تفوق وثوقيّة الأنظمة الموجودة في هذا المجال، إضافة لكونها تتمتع بسرعات معالجة تصل لما بين ضعف إلى خمسة أضعاف الأنظمة الأخرى (التي تمّت دراستها). مع وجود احتمال كبير لوجود الحل الموافق للسياق كأوّل حل من الحلول الناتجة.

أما بالنسبة لمجال تطبيقات التنقيب في النصوص، فإن الحالات **+Rn** تعتبر الحل الأكثر مثالية بين الحلول المطروحة، لأنها سرعات معالجة عالية تصل إلى أكثر من ٥٠ ألف كلمة بالثانية في بعض الحالات. ومن المميزات الهامة لهذه الحالات أن الحل الأوّل يتوافق مع السياق في أكثر من ٩٥% من الكلمات المختبرة.

أما بالنسبة للتطبيقات اللغويّة (المعاجم الحاسوبيَّة، أنظمة تعليم اللغة العربيَّة،...) فهي تطبيقات تتطلب عادةً وثوقيّة "شبه كاملة"، وسرعة المعالجة ليست ذات أولويّة فيها، لذلك يمكن أن نلجأ لمشاركة النظام مع قاعدة معطيات تحوي جذور الجذوع العربيّة، وبالتالي فإن ما يفترض أن يقدّمه نظام التحليل الصرفي هو وجود كل الحلول الممكنة للكلمة، وإن لم يكن ذلك بوثوقيّة عاليّة لأن وجود قاعدة المعطيات سيرفع الوثوقية للمستوى المطلوب. وبناء على هذا التوصيف فالنظام المنجز يُخدم هذا النوع من التطبيقات أيضاً (مثل الحالة **R7, R8**).

النتائج:

نورد في الجدول التالي النتائج التي تعطيها النسخة **R8** على الكلمات الواردة في وثيقة " مواصفات نظام التحليل الصرفي في اللغة العربية" للأستاذ مروان البواب. وهي –كما أسلفنا– النسخة الأفضل للنظام من حيث شمولية الحلول (توليد جميع الحلول الممكنة) وهي الحالة الأهم في تطبيقات المعالجة اللغوية كالمعاجم وأنظمة تعليم اللغة. علماً أنَّ الحل الأفضل (إحصائياً) هو الحل الأول، ويمكن الحصول عليه مباشرةً باستخدام النسخة **+R8**.

يعطي المحلل الصرفي هذه النتائج دون الاعتماد على قاعدة معطيات المعجم أي باستخدام أقل تخزين ممكن للمعطيات. مع إمكانية الربط مع قاعدة معطيات المعجم (عند توفره).

الكلمة	الحل الأول	الحل الثاني	الحل الثالث	الحل الرابع	الحل الخامس	الحل السادس	الحل السابع	الحل الثامن
سمع	**سمع**							
تعلم	**علم**							
سحب	**سحب**							
وهم	**وهم**	همم	هيم					
فرمت	**فرم**	رمم	ورم	رمي				
وليكتب	**كتب**	كبب						
لسعت	**لسع**	سوع	سعي					
ضربت	**ضرب**							
ضربها	**ضرب**							
ضربتها	**ضرب**							
زوجناكها	**زوج**	زجج						
دلك	**دلك**	دلل	دول	دلا				
أخذنا	**أخذ**							
افتقر	**فقر**							
ينكمش	**كمش**							
ابتعد	**بعد**							
أكرم	**كرم**	رمم	رمي					
يسر	**يسر**	سرر	سور	سير	سرا			
اصطادوا	**صيد**	صيد						
نصر	**نصر**	صرر	صور	صير				
علم	**علم**							
تصافح	**صفح**							
استقبل	**قبل**							
عرقل	**عرقل**							
تزلزل	**زلزل**							
اطمأن								
علم	**علم**							
استقام	**سقم**	قوم	قمم					
يقال	**قول**	قيل	قلل					
يرث	**ورث**	رثث	روث	ريث	رثا			

قل	**قلل**	قول	قيل	قلا				
قِ	**وقي**							
يعد	**وعد**	عدد	عود	عدا				
لاحظت	**لحظ**	حظظ	حظا					
تحسن	**حسن**	حسس	حسا					
يغزون	**غزو**	غزز	غزو					
يختار	**ختر**	خور	خير					
اصطف								
يتدحرجان	**دحرج**							
يحسن	**حسن**	حسس	حسا					
وكتاب	**كتب**	توب	تبب					
الكتاب	**كتب**							
والكتاب	**كتب**							
وبالكتاب	**كتب**							
أفبالكتاب	**كتب**							
بطاقات	**بطق**	طوق						
فكه	**فكه**	فكك						
بناء	**بني**	نوأ	بنن	بين				
باب	**بوب**							
آلام	**ولم**	يلم	لوم	لمم				
سيِد	**سود**	ودد	سدد	يدي				
أبي	**أبا**	أوب	أبا					
عدو	**عدا**	عدد	عود	عدا				
أكف	**أكف**	كفف	كيف	كفي				
ميزان	**ميز**	مزا	وزن					
اسم	**وسم**	سمم	سما	سمو				
تؤدة	**أدد**	أيد	أدا					
ميناء	**مون**	منن						
اضطرار	**ضرر**							
حيوان	**حيا**	حوا						
اضطرار	**ضرر**							
دية								

تراث	**روث**	ريث	رثث					
آلاء	**أول**							
مآب	**أوب**							
كتاب	**كتب**	توب	تبب					
مجهول	**جهل**							
مغرب	**غرب**							
مفتاح	**فتح**	فوح						
جِلسة	**جلس**							
مذهب	**ذهب**							
جبار	**جبر**							
شعب	**شعب**							
عالم	**علم**							
عظيم	**عظم**							
مكتب	**كتب**	كبب						
فرح	**فرح**	روح	رحي					
إنسانية	**أنس**	سنا	سنن	نوس	نسا			
نظرة	**نظر**							
محمد	**حمد**							
كبرى	**كبر**	برا	برر	بور				
عسل	**عسل**							
تفاح	**فوح**							
مدارس	**درس**							
عطشان	**عطش**							
إبراهيم	**كلمة_جامدة**							
عثمان	**كلمة_ جامدة**							
سوداء	**سود**	سدد						
قمر	**قمر**							
أرض	**أرض**	رضض	روض	رضا	رضو			
طريق	**طرق**							
شمس	**شمس**							
يمين	**يمن**	مون	يمم	منن	يوم			

عنق	**عنق**							
خضراوان	**خضر**							
أعلون	**علو**	علن	علل	عول	عيل	علو		
صحراوات	**صحر**							
عصوان	**عصا**	عوص						
أَرضون	**أرض**	رضو	رضض	روض	رضا			
أمهات	**أمه**	أمم	موه	هوا				
معطيات	**عطا**							
أعماوي	**عما**	عمم	عوم					
لغية	**لغا**	لغا	غيا					
عصية	**عصا**	عوص						
ظبي	**ظبي**	ظبي						
دمي	**دما**	دمم	دوم	ديم				
خلوي	**خلا**	خلل	خول	خيل				
بيضاوِي	**بيض**							
فتوِي	**فتي**	فتت	فوت					
أبا	**أبا**	أوب						
فلاحون	**فلح**	لحن	لوح	حين	لحح	حنن	حوا	حنا
بمصابيح	**صبح**							
للفتى	**لفت**	فتي	لفف	فتت	ليف	لفا	فوت	
من	**كلمة_خاصة**	منن	مون	منا				
حذار	**حذر**							
كيف	**كلمة_خاصة**	كيف	كفف					
المدرسة	**درس**							
حيث	**كلمة_خاصة**	حثث						
سوف	**كلمة_خاصة**	سوف	سفف					
بل	**كلمة_خاصة**	بلل	بول	بلا				
وعدت	**وعد**	عدد	عود	عدا				
ومفصلك	**فصل**							
فمنكم	**كلمة_خاصة**	منن	كمم	يمن	مون	منا	كوم	كمي