



أدب/ ترجمة , البحوث , لغة

جامعة بيرزيت تطلق قاموساً إلكترونياً للغة العربية

إيوارد فوكس / 2018-11-28



مصطفى جرار مع ما يقرب من 150 من القواميس العربية التي تم مسحها ضوئياً لتندخل في قاعدة بيانات مشروع أنطولوجيا اللغة العربية (الصورة: إياد جاد الله - جامعة بيرزيت)

صمم أحد أساتذة علوم الكمبيوتر في جامعة بيرزيت أداة على الإنترنت أطلق عليها اسم "أنطولوجيا اللغة العربية"، وهي عبارة عن قاموس شامل للغة العربية ونظام يتيح تصميم برامج جديدة باللغة العربية، بما في ذلك توفير ترجمة آلية أفضل.

عمل مصطفى جرار، الأستاذ المساعد في قسم علوم الكمبيوتر في بيرزيت، في الضفة الغربية، لمدة ثماني سنوات على إنشاء الأداة الجديدة. لا تعمل الأداة كمعجم عربي قابل للبحث يمكن استخدامه كقاموس ومعجم للمرادفات فحسب، لكنها تعمل بمثابة نظام منطقي - علم الأنطولوجيا - من شأنه أن يميز الخصائص الفريدة للغة العربية. حيث يمكن أن تجد الأداة العلاقات بين معاني الكلمات العربية بطريقة عربية أصيلة للمرة الأولى في التاريخ.

توفر أداة أنطولوجيا اللغة العربية إمكانية الحصول على نتائج أكثر دقة من بحث محرك البحث جوجل باللغة العربية، وترجمة إلكترونية أفضل للنصوص العربية، وأفاق جديدة على اللغة لطلاب وعلماء الأدب العربي.

تم الإعلان عن الأداة الجديدة، وهي متاحة للاستخدام الشخصي على الموقع <http://ontology.birzeit.edu>، في حفل أقيم في بيرزيت يوم 25 أيلول/ سبتمبر. وتمتلك جامعة بيرزيت حقوق الطبع والنشر.

قال جرار "إنه محرك البحث الأول من نوعه بالنسبة للغة واحدة." بمعنى أن محرك البحث يقدم نتائج من 150 قاموس عربي. أضاف، "تخيل أن يكون لديك قاموس أكسفورد الإنجليزي، وقاموس ميريام فيبستر وكل القواميس الأخرى مجتمعة في مكان واحد، متكاملة وموحدة في قاعدة بيانات واحدة".

تشير كلمة "علم الأنطولوجيا" في اسم المشروع إلى ذلك المفهوم في علم اللسانيات الذي يعني طريقة تصنيف المعاني والعلاقة بين الكلمات في اللغة، وهو في الأصل مصطلح فلسفي يعني دراسة الوجود، أي طريقة تنظيم الأشياء الموجودة.

قال جرار "لقد قمنا بأخذ الكلمات العربية وتصنيف معانيها على شكل شجرة. إن علم أنطولوجيا اللغة العربية بمثابة شجرة لمعاني الكلمات العربية". تنتمي المفاهيم المجردة الأوسع لجذع الشجرة، في حين تمثل الفروع أشياء محددة وترمز الأوراق لخصائص شيء محدد.

وصف جرار الأنطولوجيا بكونها وريثة اللغة العربية في مشروع اسمه WordNet، وهو مشروع بدأ في عام 1985 في جامعة برينستون في الولايات المتحدة. يعمل WordNet بمثابة معجم للكلمات الإنجليزية قابل للبحث، ويعمل كقاموس ومعجم للمرادفات.

يكن الفرق بين أنطولوجيا اللغة العربية وWordNet، بحسب جرار، في أن نظامه يقوم بإصلاح معاني الكلمات العربية وفق استخدامها من قبل المتخصصين، فيما يقوم قاموس WordNet بتعريف معنى الكلمة وفقاً لما هو متفق عليه بخصوص استخدامها من قبل المتحدثين الأصليين بتلك اللغة. قال، "عندما كنت أقوم بتصميمه، استشرت علماء الطبيعة والفيزياء والكيمياء والأحياء وما إلى ذلك، لأنهم من يحدد معاني الكلمات في تخصصاتهم".

قالت كريستيان فيليوم أستاذة علوم الكمبيوتر في جامعة برينستون والمديرة الحالية لـ WordNet، بأن مشروع جرار بمثابة خطوة إلى الأمام بالنسبة للفكرة الأصلية لـ WordNet. قالت فيليوم "إنه يختلف عن WordNet في كونه يرغب ببناء علم الأنطولوجيا. تعود هذه الفكرة إلى فلسفة أرسطو، والتي نظمت [في أعمال مثل الفئات والميتافيزيقا (ما وراء الطبيعة)] المعرفة من حيث الفئات العامة والضيقة، فضلاً عن العلاقات الكاملة والجزئية. يتمثل ابتكار مصطفى في علم الأنطولوجيا، الذي يخلق فهماً للعلاقات بين معاني الكلمات".

قالت فيليوم "إنه شخص غير عادي وشجاع، وهو يستحق دعماً هائلاً".

بناء قاعدة البيانات

ويهدف إنشاء معجم عربي مشترك، توجب على جرار إدخال محتويات 150 قاموس عربي يدوي في قاعدة بيانات. كان ذلك عملاً شاقاً في البداية، حاول جرار التقاط المعلومات من الكتب باستخدام الماسح الضوئي. لكن استخراج البيانات المفيدة من الكتب يتطلب برنامج التعرف الضوئي على الحروف والمعروف باسم OCR، القادر على قراءة اللغة العربية.

قال جرار، "حاولت استخدام برنامج OCR، لكنه لم ينجح". وأضاف بأن برنامج OCR في اللغة العربية لا يزال ضعيفاً للغاية، موضحاً بأن "مقدار التصحيحات التي يتعين عليك القيام بها أكثر بكثير من العمل الذي ستقوم به إذا ما قمت بإدخال النص من الصفر بشكل يدوي."

عوضاً عن ذلك، قام جرار بحشد طلاب جامعة بيرزيت للقيام بهذه المهمة.

تطالب جامعة بيرزيت الطلاب بأداء 120 ساعة خدمة مجتمعية قبل أن يكون في إمكانهم التخرج. يهدف البرنامج إلى تحسين تواصل الجامعة مع المجتمع الفلسطيني خارج الحرم الجامعي. في العادة، يقوم طلاب بيرزيت الذين يقومون بالخدمة المجتمعية بقطف الزيتون في المزارع المحلية، أو مساعدة كبار السن في منازلهم.

اعتبرت إدارة جامعة بيرزيت العمل في مشروع جرار بمثابة نشاط مناسب لخدمة المجتمع، بسبب قيمته للثقافة العربية والمجتمع ككل، بحسب جرار، وبذلك أصبح في إمكان الطلاب الوفاء بواجب الخدمة المجتمعية عن طريق كتابة محتويات صفحات القواميس العربية في قاعدة بياناته. ولتحسين دقة العمل، يقوم جرار بإعطاء ذات الصفحة لأكثر من طالب واحد للقيام بنسخها.

في نهاية المطاف، قام باختيار الطلاب الذين يمكنهم القيام بهذا العمل على مستوى عالٍ، ومن دون أخطاء، وقام بإشراك هؤلاء الطلاب للعمل مقابل أجر. استغرق إنجاز العمل ثماني سنوات.

أكد فادي زراقل، الأستاذ المشارك في هندسة الكمبيوتر في الجامعة الأميركية في بيروت والمتخصص في المعالجة اللغوية الطبيعية للغة العربية، على قيمة مشروع جرار في الطريقة التي يتبع من خلالها الخصائص المميزة للغة العربية، والعلاقات الطبيعية بين الكلمات العربية.

من المحتمل أن تشترك كلمتان عربيتان مشتقتان من ذات الجذر الثلاثي المؤلف من الحروف الساكنة في المعنى، بشكل مباشر أو غير مباشر. أوضح زراقل بأن أنطولوجيا اللغة العربية ستكون قادرة على تحديد هذه العلاقات بطريقة لا يستطيع قاموس WordNet القيام بها.

قال "سيساعدنا ذلك على اكتشاف علاقات دلالية جديدة بين الكلمات العربية."

بإمكان النظام أيضاً التعرف على مورفولوجيا "شكل" الكلمات العربية، أي الطرق التقليدية لكتابتها. قال، "في اللغة العربية، بإمكان الناس قراءة الكلمات حتى لو لم تُكتب حروف العلة القصيرة "الحركات"، لكن الأدوات الآلية التي لا تستخدم تحليلاً يستند إلى المعرفة، كذلك الموجود لدى أنطولوجيا اللغة العربية، ستواجه وقتاً عصيباً في فهم الكلمات العربية بدون حروف العلة."

من شأن دمج برنامج أنطولوجيا اللغة العربية مع مترجم جوجل، على سبيل المثال، أن يحسن من جودة الترجمة من وإلى اللغة العربية، بحسب زراقل. تستخدم خدمة الترجمة من جوجل طريقة إحصائية للقيام بالمهمة: فهي تنتج ترجمة من خلال تحليل كميات هائلة من البيانات، ولكن نتائجها تفتقر إلى الدقة في بعض الأحيان. قال زراقل، "ينظم برنامج مصطفي مفاهيم اللغة العربية. إذا ما استخدمت ذلك بالإضافة إلى التقنيات الإحصائية، ستزداد الدقة بشكل كبير، وهناك حاجة ماسة لتحقيق ذلك."

يرى زراقل بأن مشروع جرار بمثابة مساهمة في دراسة اللغة العربية نفسها، فضلاً عن كونه أداة بحث مفيدة. قال، "أنطولوجيا اللغة العربية توثيق للقدرة الفلسفية لما يمكن أن تعبر عنه اللغة العربية."