حوسبة المعاجم العربيّة

دراسة حالة

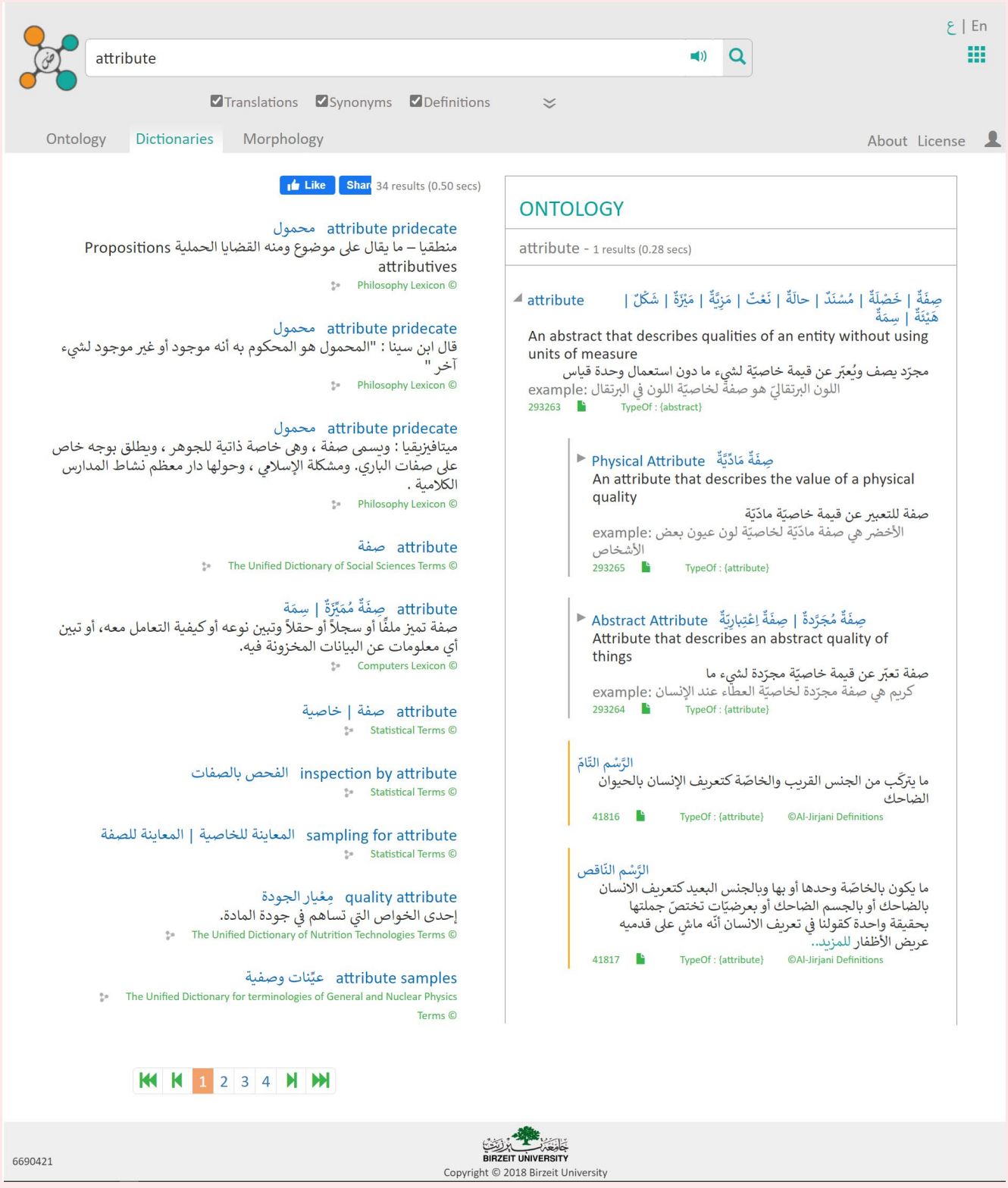
بقلم: د. مصطفى جرّار
جامعة بيرزيت، فلسطين

مقدّمة

تعرض هذه الورقة ملخّصًا لتجربة بناء أضخم قاعدة بيانات لغويّة للّغة العربيّة، وكذلك محرّك بحث للمعاجم العربيّة يحتوي على 150 معجمًا تمّت حوسبتها في جامعة بيرزيت بفلسطين.[142] ومحرّك البحث هذا متاح للعامّة ولجمهور الطلبة والباحثين والمترجمين ومتعلّمي اللغة وغيرهم، وهو يتيح للمستخدم البحث عن الكلمة في 150 معجمًا عربيًّا ومتعدّد اللغات واسترجاع تعريفاتها ومرادفاتها وكذلك ترجماتها المتخصّصة. وهذه الترجمات تعتمد على المعاجم المدقّقة والمنقّحة، وهي أدقّ بكثير من جميع المترجِمات الآليّة المتاحة (مثل جوجل للترجمة)، وتجدر الإشارة إلى أنّ محرّك البحث هو الأوّل عالميًّا، حيث لا يوجد محرّك بحث لمعاجم أيّ لغة أخرى، حتّى الإنجليزيّة. ويمكن لمطوّري البرمجيّات كذلك استعمال محرّك البحث عبر واجهة خاصّة (API) للوصول إلى الترجمات والمترادفات واستخدامها داخل برمجيّاتهم وتطبيقاتهم. وقد تمّت حوسبة المعاجم بطريقة يدويّة على مدى تسع سنوات، وبعد ذلك قمنا بتوحيد هذه المعاجم في قاعدة بيانات واحدة تشتمل على معاجم لغويّة تقليديّة وحديثة، ومسارد، ومكانز، ومعاجم ثنائيّة وثلاثيّة اللغة، وقواعد بيانات تصريفيّة واشتقاقيّة. وتعدّ الأنطولوجيا العربيّة أهمّ المصادر اللغويّة في محرّك البحث، وهي شجرة أو تصنيف لمعاني الكلمات العربيّة وتمثيل لها بلغة المنطق، بحيث يستطيع الحاسوب فهم المعاني ومعالجتها. وقد تمّ بناء هذه الشجرة كمشروع بحثيّ منفصل في جامعة بيرزيت،[143-144] ونعمل حاليًّا على ربط جميع مدخلات المعاجم وصولًا إلى شبكة لغويّة محوسبة للّغة العربيّة تُسمّى Big Linguistic Data Graph، ومن ثمّ ربط هذه الشبكة باللغات الأخرى.[145]

ويأتي هذا العمل ضمن مشروع بحثيّ غير ربحيّ وطويل الأمد لخدمة اللغة العربيّة، ولإغناء الإنترنت بمحتوى عربيّ نوعيّ، إذ سيتمكّن المستخدمون من إيجاد المصطلح العربيّ الذي يناسب متطلّباتهم وبترجمات متعدّدة، خاصّةً وأنّ قاعدة البيانات تحتوي على عدد ضخم من معاجم المصطلحات المعاصرة وفي شتّى العلوم والمجالات العلميّة والهندسيّة والتجاريّة والأدبيّة وغيرها. إضافةً إلى ذلك، فإنّ محرّك البحث متاح للعامّة مجّانًا من خلال موقع جامعة بيرزيت الإلكترونيّ (https://ontology.birzeit.edu) الذي يقدّم تعريفًا بالمحرّك وبالمشاريع الأخرى ذات العلاقة على هذا الرابط (https://ontology.birzeit.edu/about-ar)، بالإضافة إلى روابط للأوراق العلميّة التي بُني محرّك البحث على أساسها. وتجدر الإشارة إلى أنّ محرّك البحث تمّ تطويره وفق المعايير والمقاييس الصادرة عن منظّمة الشبكة العالميّة (W3C)، وخاصّةً المعايير المتعلّقة بأسلوب نشر البيانات على الإنترنت، وكذلك المعايير المتعلّقة بتمثيل وتبادل البيانات اللغويّة.

وفيما يلي يقدّم الشكل التالي مثالًا على طريقة استرجاع ترجمات كلمة «attribute» من عدّة معاجم ومن الأنطولوجيا، ويستطيع الباحث هنا التحكّم فيما إذا كان يرغب في استرجاع ترجمات، و/أو مترادفات، و/أو تعريفات فقط. وتظهر نتائج المعاجم في الجهة اليسرى من الشكل، ونتائج الأنطولوجيا في الجهة اليمنى:



حوسبة المعاجم

قمنا بحوسبة حوالي مائة وخمسين معجمًا عربيًّا ومتعدّد اللغات، واضطررنا لطباعة غالبيّتها يدويًّا، خاصّةً وأنّه لا توجد نسخ إلكترونيّة لغالبيّة المعاجم حتّى في حوزة مؤلّفيها. وفي المراحل التالية، عملنا على تطوير برمجيّات وخوارزميّات لإعادة تشكيل بنيويّة (re-structuring) لمدخلات المعاجم،[146] بحيث يتمّ الفصل بين المُدخَلة المعجميّة والتعريف والمشتقّات والتصريفات والترجمات والمترادفات، انتهاءً بحفظها في قاعدة بيانات واحدة.[147] وجاءت الخطوة الأخيرة ضروريةً لكون صناعة المعاجم التقليديّة تركّز في تصميمها على الاستخدام الورقيّ، بينما يحتاج الاستخدام الحاسوبيّ إلى تراكيب بنيويّة مختلفة تسهّل عمليّات الإضافة والاسترجاع والبحث.

وتشتمل قاعدة البيانات اللغويّة الحاليّة على أنواع مختلفة من المعاجم، مثل المعاجم اللغويّة التقليديّة القديمة والجديدة، وكذلك المسارد (glossaries) التي تشرح المصطلحات، والمكانز (thesauri) التي تحتوي على مترادفات، ومعاجم ثنائيّة وثلاثيّة اللغة، ومعاجم الفروق اللغويّة، وقواعد بيانات تصريفيّة واشتقاقيّة. كما تغطّي قاعدة البيانات شتّى المجالات مثل العلوم الطبيعيّة، والهندسة، والطبّ، والاقتصاد، والأدب، والإنسانيّات، والفلسفة،

والفنّ، والاقتصاد. وتجدر الإشارة إلى أنّ محرّك البحث في هذه المرحلة يُظهر الترجمات من وإلى الإنجليزيّة، إلّا أنّ قاعدة البيانات تحتوي العديد من اللغات الأخرى، خاصّةً الفرنسيّة، والتي سنعمل لاحقًا على توفيرها ضمن محرّك البحث.

حقوق الملكيّة

قمنا بالتواصل الفرديّ مع أصحاب المعاجم وحقوق الملكيّة الفكريّة للحصول على تراخيص باستخدام معاجمهم ضمن محرّك البحث، مشيرين إلى أنّ محرّك البحث، في تصميمه، يعرض اسم المعجم ورمز حقوق الملكيّة بجانب البيانات التي يتمّ استرجاعها. وعند الضغط على اسم المعجم يظهر اسم المؤلّف والناشر وروابط لصفحاتهم الإلكترونيّة ولكيفيّة شراء النسخة الورقيّة من المعجم، ما شجّع أغلبهم على منحنا الترخيص باستعمال معاجمهم. إضافةً إلى ذلك، قمنا بتوقيع مذكرة تفاهم مع منظّمة الألكسو، للتعاون في مجال حوسبة اللغة العربيّة وهندسة المعاجم، حيث منحونا الإذن باستخدام حوالي خمسين معجمًا ثلاثي اللغة صدرت عن مركز تنسيق التعريب بالرباط لحوسبتها وإضافتها إلى قاعدة البيانات الخاصّة بمحرّك البحث.

التطبيقات العلميّة والعمليّة

يتيح محرّك البحث المعجميّ، جنبًا إلى جنب مع الأنطولوجيا، خدمات لغويّة عديدة لمجموعتين رئيستَين تشمل الأولى الباحثين واللغويّين والمترجمين والطلبة والجمهور العربيّ عمومًا، إضافةً إلى متعلّمي اللغة العربيّة، وذلك لاسترجاع مترادفات وترجمات وتعريفات.[148] وتشمل الثانية الشركات التي تحتاج لبيانات لغويّة من أجل تطوير تطبيقات حاسوبيّة، مثل تطبيقات الترجمة الآليّة، أو البحث الدلاليّ، أو التحليل اللغويّ، أو التدقيق الإملائيّ، أو استخراج البيانات، وغيرها. ويمكن للعاملين في هذه الشركات الاستفادة من محرّك البحث عبر واجهة خاصّة لمطوّري البرمجيّات (API) تستطيع الشركات من خلالها استرجاع مترادفات وترجمات واستخدامها داخل برمجيّاتها وتطبيقاتها مباشرة. وفي كلتا الحالتين، فإنّ ما يقدّمه المشروع يُسهم في تقديم ترجمات وتعريفات دقيقة ومتخصّصة، تقلّل من هامش الخطأ - بنسبة كبيرة - في البحث المعرفيّ البحت، أو في التطبيقات التكنولوجيّة ذات الصلة.

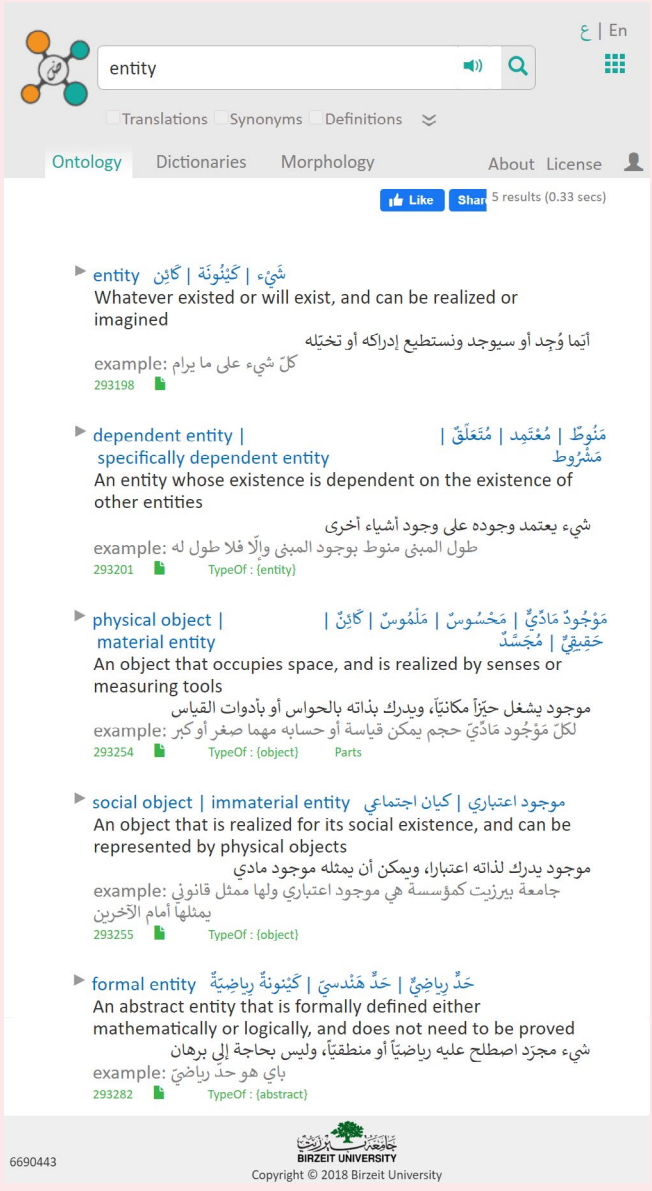
تصميم محرّك البحث

من أهمّ سمات محرّك البحث اعتماده على المعايير والمقاييس الصادرة حديثًا عن منظّمة الشبكة العالميّة (W3C) والتزامه بها، خاصّةً تلك المتعلّقة بأسلوب نشر المعلومات والمعروفة بـ (Best Practices for Publishing Linked Data). وبالتالي، فقد تمّ تصميم الروابط لتكون (Cool URLs, Simple, Stable, Manageable, Linkable)، وهي المعايير الحديثة لنشر البيانات على الإنترنت.[149] إضافةً إلى ذلك، تمّ الالتزام بالمعايير والمقاييس المتعلّقة بتمثيل البيانات اللغويّة والمعروفة بـ (W3C Lemon model)،[150] حيث يتيح محرّك البحث الوصول إلى البيانات المسترجعة ممثلةً باستخدام (Lemon model).[151] انظر، مثلًا، تمثيل معنى كلمة "موضع" في المعجم الفلسفيّ:
https://ontology.birzeit.edu/lemon/lexicalconcept/300000117

الأنطولوجيا العربيّة

الأنطولوجيا هي فرع من فروع الفلسفة وتعني علم الوجود، وهي تعلو الإبستمولوجيا في نظريّة المعرفة، ولكنّ هذا المصطلح أصبح رائجًا مؤخرًا في علم الحاسوب وهندسة المعرفة، ويعني وصف المعرفة مفاهيميًّا ودلاليًّا، وتمثيلها بلغة المنطق.[152-153] والأنطولوجيا العربيّة هي مشروع آخر ومستقلّ عن محرّك البحث المعجميّ، نعمل عليه في دائرة علم الحاسوب بجامعة بيرزيت،[154-155] ويهدف إلى وصف وتصنيف مفاهيم الكلمات العربيّة. أي أنّ الأنطولوجيا العربيّة وبمعنًى آخر، هي شجرة مفاهيم ومعاني الكلمات العربيّة، وليس الكلمات نفسها، حيث يتمّ وصف كلّ مفهوم/معنى لكلّ كلمة عربيّة بحدوده وبصفاته الجوهريّة المُميّزة، وإعطاؤه رقمًا فريدًا، ثمّ يتمّ تصنيف المفهوم إلى أجناسه الأدنى، وهكذا، بحيث يكون التصنيف تصنيفًا مانعًا. وقد تمّ استخدام المنطق الوصفيّ الحديث (Description Logic) كلغة لتمثيل الأنطولوجيا العربيّة ولتمكين الحاسوب من فهمها والاستنتاج منها. وبالرغم من أنّ الأنطولوجيا العربيّة مصمّمة لاستخدامها في التطبيقات الحاسوبيّة، فإنّه يمكن استخدامها كمعجم؛ ولكن تجدر الإشارة إلى أنّه، إضافةً لتصنيفها للمعاني، فإنّ ما يميّز الأنطولوجيا عن المعاجم أيضًا، أنّه يتمّ التحقّق من تعريفات وتصنيفات الأنطولوجيا بالاعتماد على ما وصلت إليه العلوم، خلافًا للمعاجم في اعتمادها على شيوع الدلالة المستخدمة بين المتحدّثين. هذا بالإضافة إلى أنّ المفاهيم في الأنطولوجيا هي كلّيّات/أصناف لأفراد.

ومشروع الأنطولوجيا مشروع طويل الأمد، وقد تمكّنّا من إنجاز ما يزيد عن ألف مفهوم، وهي المفاهيم الأعلى والأكثر تجريدًا في اللغة العربيّة، بالإضافة إلى عشرة آلاف مفهوم آخر تمّ تعريفها وتصنيفها جزئيًّا. ويمكن الوصول إلى شجرة الأنطولوجيا كاملةً، واستكشاف المفاهيم الأدنى عبر محرّك البحث الذي يمكن الوصول إليه عن طريق الرابط:
https://ontology.birzeit.edu/concept/293198



أهمّ التحدّيات

1- عدم توفّر مصادر لغويّة محوسبة

بالرغم من وجود كمّ كبير من المعاجم العربيّة المطبوعة ورقيًّا، فإنّ عدد المصادر اللغويّة المحوسبة والمتاحة للبحث العلميّ ولمطوّري البرمجيّات ضئيل، ممّا يحدّ من دعم اللغة العربيّة في معظم التطبيقات الحاسوبيّة. وعلى الرغم من أنّ سوق التقنيّات في العالم العربيّ كبير، فإنّ مطوّري البرمجيّات يواجهون تحدّيات كبيرة في إيجاد مصادر لغويّة يمكن استعمالها بسهولة وبتكلفة معقولة.

2- ضعف الصناعة المعجميّة وتوانيها عن مواكبة النظريّات اللسانيّة الحديثة

ثمّة نقص في أنواع جديدة من المعاجم، مثل: مكانز المترادفات، ومعاجم الفروق الدلاليّة، والشبكات الدلاليّة، ومعاجم السلوك التصريفيّ للأفعال، ومعاجم المصطلحات والترجمات الحديثة والكلمات العربيّة المستجدّة. وقد أصبحت الحاجة إلى مثل هذه الأنواع من المعاجم، والتي لم تكن موجودة أو منتشرة قديمًا، ملحّة مع تطوّر احتياجات المجتمع، ووسائل تعليم اللغة، وظهور برمجيّات ذكيّة أصبحت اللغة ركنًا أساسيًّا فيها.

3- عدم وجود منهجيّات موحّدة ومعياريّة في الصياغة المعجميّة العربيّة

فمثلًا، لا يوجد اتّفاق على كيفية صياغة ووسم المدخلات المعجميّة، ويظهر ذلك جليًّا في التباين الحاصل بين صياغة المعاجم التي صدرت في السنوات الأخيرة.

4- عدم توفّر قواعد بيانات لغويّة شاملة

فعلى سبيل المثال، لا توجد قائمة شاملة وموحّدة بالجذور، أو بالأوزان الصرفيّة والاشتقاقيّة، أو بفروع الكلمات (lemmas)، أو بالاشتقاقات، أو غيرها. ويشكّل هذا النقص تحدّيًا لكلّ من يريد المساهمة في تأليف معجم، أو إنشاء قاعدة بيانات لغويّة، أو بناء تطبيق حاسوبيّ.

5- عدم توفّر ميزانيّات للبحث العلميّ الرصين والموجّه نحو حوسبة اللغة العربيّة لدعمها في التطبيقات الحاسوبيّة.

رؤية مستقبليّة: بناء شبكة لغويّة شاملة Big Linguistic Data Graph

تتركّز أبحاثنا في جامعة بيرزيت حاليًّا في ثلاثة مسارات:

الأوّل: إغناء المحرّك المعجميّ بمزيد من التعريفات والشروحات اللازمة، وتوسيع الأنطولوجيا العربيّة بإضافة المزيد من المفاهيم في المستويات الدنيا من الشجرة الدلاليّة.

الثاني: ربط مدخلات المعاجم العربيّة مع شبكة المفردات الإنجليزيّة (Wordnet)، وهي شبكة مفردات تمّ تطويرها في جامعة برنستون وربطها بغالبيّة اللغات في العالم،[156] وبالتالي، فإنّ ربطها بالمدخلات المعجميّة العربيّة يؤدي تلقائيًّا إلى ربط المدخلات العربيّة بسائر اللغات العالميّة الأخرى.

الثالث: وهو المشروع المستقبليّ الذي شرعنا بأولى خطواته، ويقوم على ربط المعاجم وجميع البيانات اللغويّة المتوفّرة لدينا ببعضها، وصولًا إلى بناء شبكة بيانات لغويّة شاملة من حيث الكمّ ومن حيث المستويات اللغويّة الدلاليّة والتصريفيّة والاشتقاقيّة، والمقابل في اللغات الأخرى. وبدوره، فإنّ هذا المسار يتضمّن ثلاثة مستويات مترابطة:

- تصريفيًّا. نعمل على وصف كلّ مدخلة معجميّة بجذعها (lemma)، وبالتالي تصبح كلّ مدخلات المعاجم مترابطة تصريفيًّا، بغضّ النظر عن السوابق واللواحق في المدخلات، أو الصيغ المختلفة من نفس المدخلة في المعاجم الأخرى.
- اشتقاقيًّا. نقوم بالربط بين المدخلات التي تختلف عن بعضها اشتقاقيًّا (derivation)، إن كانت فعلًا، أو اسم فاعل، أو اسم مفعول، أو مصدرًا، أو اسم تفضيل، ... إلخ.
- دلاليًّا. نستعمل الأنطولوجيا العربيّة لتقعيد وربط المعاجم مفاهيميًّا، حيث نقوم بربط كل دلالة لغويّة (lexical concept) لكلّ مدخلة معجميّة في كلّ معجم بمفهومٍ مقابل في الأنطولوجيا (ontological concept)، وبذلك، تصبح مدخلات المعاجم مترابطة دلاليًّا فيما بينها.

وبالرغم من إدراكنا للجهود الكبيرة التي يتطلّبها بناء شبكة لغويّة شاملة ومحوسبة للّغة العربيّة، فإنّ من شأنها - حين تتحقّق - أن تُحدِث نقلة جذريّة في المعرفة اللغويّة والدلاليّة وتؤسّس لمنهجيّات جديدة للبحث النظريّ والتطبيقيّ.