نحو تأصيل منهجي لبناء أنطولوجيا اللغة العربية

مصطفى جرار
جامعة بيرزيت، فلسطين
mjarrar@birzeit.edu

1. تقديم

خَلق الإنترنت والاتصال السهل بين الأنظمة حاجة ماسة ليس إلى تبادل البيانات فقط، بل أيضا إلى اتفاق حول معاني هذه البيانات Data Semantics. تُعتبر الانطولوجيا الحجر الأساس للتبادل السليم والفعّال للبيانات، حيث تحتوي على تعريف دقيق للمعنى الدلالي للبيانات المُراد تبادلها. حيث تكتب هذه التعريفات بلغة المنطق Formal Logic كي يستطيع أي نظامٍ فهمها وحسابها، بل والاستنتاج منها. وقد ظهر في السنوات العشر الأخيرة الكثير من التطبيقات التي تُعتبر فيها الانطولوجيا بالغة الأهمية مثل الحكومات الإلكترونية، التجارة الالكترونية، محركات البحث، المكتبات الالكترونية، وغيرها من التطبيقات.

تقدم هذه الورقة المنهجية المتبعة في مشروع ال (ArabicOntology[1]) القائم في جامعة بيرزيت-فلسطين لبناء أنطولوجيا للغة العربية، والتي تحتاج إلى العديد من السنوات والجهود لإنجازها بشكلها الشمولي. وتعتبر هذه المنهجية إطار عمل ومنصة انطلاق لأبحاث ومشاريع مستقبلية طويلة الأمد. إن هذه المنهجية بفكرتها العامة، تعتبر خطوة هامة في تاريخ اللغة العربية، فهي تؤسس لطريقة جديدة لتعريف معاني ودلالات الكلمات. وعلى مستوى المحتوى، تتيح هذه المنهجية إنتاج قاموس دلالي آلي تصويري يصنف معاني الكلمات ويشجرها، بحيث تكون هذه المعاني والعلاقات فيما بينها مؤصلة فلسفيا ولغويا وممثلة بلغة المنطق الشكلي.

ويمكن تلخيص هذه المنهجية وما تم إنجازه حتى الآن بما يلي:

(1) بناء المستويات العليا لأنطولوجيا اللغة العربية (Top Level Concepts)، والتي تشكل نواة الأنطولوجيا العربية.

(2) جمع وإستنباط تعريفات ومعاني من المعاجم العربية المتاحة (ما يقارب اربعمائة ألف مفهوم) وإعادة صياغتها وهندستها كتعريفات دلالية، بما يضمن خضوعها للضوابط التي تركز على الصفات الجوهرية المميّزة للمفهوم دون غيره، وليس الصفات العرضية.

(3) تطوير برنامج حاسوب مبنى على خوارزمية ذكية تعمل على الربط بين مفاهيم الأنطولوجيا العربية، مع مقابلاتها في أنطولوجيا اللغة الإنجليزية (WordNet)، مما يتيح إستجلاب علاقاتها الدلالية إلى الأنطولوجيا اللغة العربية.

تجدر ألإشارة إلى أن ما يميز الأنطولوجيا العربية التي نسعى لبنائها، مقارنةً مع الإنجليزية (WordNet)، أننا نسعى للوصول إلى علاقات الدلالية مؤصلة فلسفياً ومنضبطة منطقياً (Formal Logic)، وبالتالي لا يشوبها غموض دلالي. كما في ال WordNet. إضافة إلى ذلك، تركز المنهجية المتبعة علي أن تكون صياغة تعريفات المفاهيم (Glosses) محكومة بضوابط أنطولوجية في الشكل والمضمون. وأخيرا و ليس أخراَ، أن تكون المستويات العليا العربية مؤصلة فلسفيا ومنذ البداية، إعتمادا على أهم الأنطولوجيات العليا العامة (Upper Level Ontolgies) وليس ربطها ربطا خارجيا بهذه الأنطولوجيات بعد إستكمالها، كما هو الحال في الأنطولوجيا الإنجليزية (WordNet).

تقدم هذه الورقة عرضا مختصرا للمنهجية المقترحة، مقسمة على النحو التالي: يقدم القسم الثاني خلفية عامة عن الأنطولوجيا، ظهورها، فوائدها، كيفية بناءها واستخدامها، وتطبيقاتها ودورها في بناء منهجية العمل. وعرض بعض الدراسات الجارية في هذا المضمار. ويستعرض الجزء الثالث الخطوات المنهجية المقترحة لبناء الأنطولوجيا العربية. أما الجزء الرابع فيتضمن تفصيلا لمنهجية بناء الأنطولوجيا العليا للغة العربية، وأخيرا يقدم الجزء الخامس خلاصات عامة ونظرة للمستقبل.

[1] http://sites.birzeit.edu/comp/ArabicOntology/

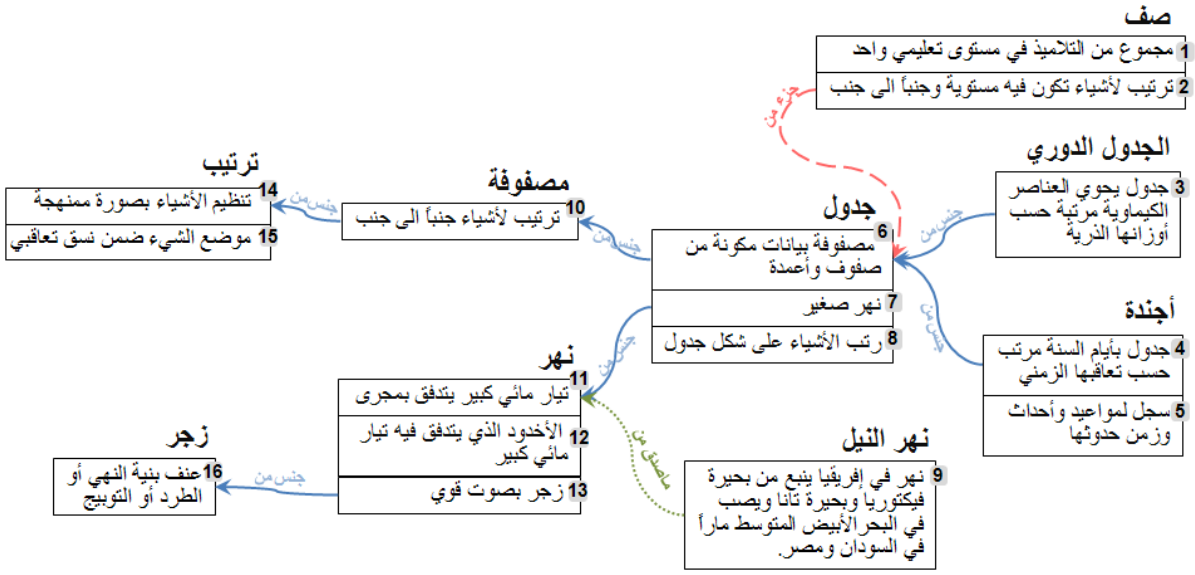
2. ما هي الأنطولوجيا العربية

كلمة الأنطولوجيا (Ontology)، هي كلمة يونانية تشير إلى فرع من فروع الفلسفة التحليلية، وتعني علم الموجود بما هو موجود [19] .أُعيد إنتاج هذا المفهوم في علم الكمبيوتر حديثاً، بشكل تبوأ صدارة الأبحاث منذ حوالي عشر سنوات. فبعد انتشار الإنترنت انتشاراً واسعاً واستخدامها في العديد من المجالات، خاصة في التجارة الإلكترونية والتعاملات اليومية، نتجت حاجة ملحة لتوحيد الأنظمة والبيانات الموجودة (System and Data Integration)، وذلك لكي تتمكن هذه الأنظمة من التعامل فيما بينها للقيام بمهمة ما[7،33]. فمثلا، قد تحتاج البنوك إلى أن تتبادل معلومات فيما بينها أو مع جهات أخرى عبر الإنترنت. إن هذا النوع من التبادل البيني (Interoperability) للبيانات بين الأنظمة يحتاج إلى حل إشكالات كثيرة لإنجازه، ليس فقط فيما يتعلّق بالسرعة والسرية، بل أيضا إلى اتفاق بين البنوك على طريقة التناقل وتركيب البيانات (Data Structure)، والأهم من ذلك هو الاتفاق على المعنى الدلالي للبيانات المتبادلة (Data Semantics). فقد يسمي بنك ما العميل "زبون"، دون التمييز إن كان هذا الزبون شخصاً أم شركة، بينما يرى بنك آخر أن التمييز بينهما ضروري، وقد يقوم بنك ثالث بالتمييز بين عميل بالغ وعميل قاصر؛ وبين شركة ربحية وغير ربحية. لحل إشكالية التعامل مع المعاني الدلالية للبيانات اقترح العلماء استخدام الأنطولوجيا كمرجع تُعرف فيها معاني الأشياء المراد وصفها [9,14]. بمعنى آخر، إن الأنطولوجيا هي تعريفات دقيقة لمعاني الأشياء المراد تداولها. بما أن الأنظمة التي تتبادل المعلومات، هي نفسها التي تستخدم الانطولوجيا -وليس الإنسان- فمن الضروري أن تكون المعاني في الانطولوجيا مكتوبة بطريقة تستطيع الأنظمة فهمها وحسابها[13] .بمعنى آخر، إن تعريف المعاني في الأنطولوجيا يُمثّل بلغة المنطق الشكلي (Formal Logic) وبالتالي يمكن حساب المعاني واستنتاجها آليا من الجمل المنطقية. إن أكثر الأنماط الشائعة لتعريف المعاني هي تصنيفها، فمثلا يمكن أن نصنف العميل البنكي إلى شخص طبيعي ومؤسسة، والمؤسسة إلى ربحية وغير ربحية ... إلخ، وكذلك، يمكن أن نضيف صفات معينة إلى كلٍ من هذه الأصناف، بحيث يرث الصنف صفات جنسه ويورث صفاته إلى الأصناف الناتجة عنه [11].

لبيان طريقة إستخدام الأنطولوجيا في مجال الحكومة الإلكترونية الفلسطينية، على سبيل المثال، يجب أن تحتوي هذه الأنطولوجيا على جميع المفاهيم المستخدمة في قواعد بيانات جميع المؤسسات الحكومية الفلسطينية، مثل مفهوم شخص، مواطن، شركة، مهنة، عملة، ضريبة الدخل، ضريبة القيمة المضافة، رخصة سياقه، رخصة مهنة، وغيرها، بحيث يتم إعادة تعريف جميع هذه المفاهيم بشكل دقيق، وضمن نسق "شجري" تصنيفي. بمعنى آخر، تحوي هذه الأنطولوجيا على شجرة (أي تصنيف) جميع المفاهيم المتداولة في الوزارات، وتعتبر هذه الأنطولوجيا المرجع الدلالي ليس فقط بين العاملين، بل والأهم من ذلك، تستعمل للربط المفاهيمي بين أنظمة المعلومات في هذه المؤسسات، بما يتيح تبادل بيني آلي للبيانات، مبني على قواعد موحدة للفهم (Meaningful Interoperation).

ومثال آخر على استعمال الأنطولوجيا في محركات البحث،-كما يجري العمل عليه حاليا في معهد الحقوق في جامعة بيرزيت لإغناء برنامج المقتفي[1]، وهو محرك بحث متخصص في مجال القانون الفلسطيني. سيتم إستخدام أنطولوجيا قانونية لإغناء قدرة المقتفي في البحث والاسترجاع بحيث يستطيع فهم الكلمات المراد البحث عنها (Meaningful Search) وليس بحثاً حرفياً (String-matching Search) كما هو الحال حالياً. أن مثل هذه الأنطولوجيا القانونية يجب أن تحوي جميع المفاهيم القانونية الواردة في القوانين والتشريعات الفلسطينية، مثل: إجازة، إجازة سنوية، إجازة عرضية... أو عقد، إتفاقية، مذكرة تفاهم... أو شخص طبيعي، شخص معنوي، وغيرها، بحيث يتم وضع تعريفات دقيقة ومصنفة، بشكل شجرة مفاهيمية. هذه الشجرة يمكن أن تسمى "أنطولوجيا القانون الفلسطيني". تستعمل هذه الأنطولوجيا في المقتفي،-كمحرك بحث (قانوني)، لإغناء عملية البحث، وجعلها أكثر دقة وتحديداً. فعلى سبيل المثال، إذا بحث شخص عن كلمة إتفاقية، ستحوي نتائج البحث ليس فقط على النصوص التي تحتوي كلمة إتفاقية، بل وأيضا، على النصوص التي تحتوي كلمة عقد، باعتبار أن كل عقد هو صنف فرعي من أصناف الإتفاقيات. بهذه الطريقة يستطيع محرك البحث إغناء عملية البحث وكذلك إسترجاع المعاني الأخص كما تم تعريفها في الأنطولوجيا. مع ضرورة الإنتباه أن التصنيف في الانطولوجيا، لا يبنى على أساس الترادف اللغوي كما في القواميس، أو المعنى الأعم والأخص كما في المكانز، ولكن التصنيف هنا يتم بناء على جنس الشيء ونوعه، ويُمثل باستخدام المنطق الشكلي، ما يتيح الاستنتاج آليا.

بالرغم من كون معظم تطبيقات الأنطولوجيا تركز على مجالات محددة كما سلف تبيانه (particular domain ontologies)، إلا أن هناك توجهات حديثة لبناء أنطولوجيات لغوية، خاصة بعد نجاح مشروع (WordNet) والذي بُنيَ من قبل جامعة برنستون في الولايات المتحدة [14]، والذي يُعتبر الآن أنطولوجيا لغوية شاملة للغة الإنجليزية. مثل هذه الأنطولوجيات اللغوية (Linguistic Ontologies) لا تركز فقط على تصنيف المعاني،–ونركز هنا على "تصنيف المعاني" وليس "تصنيف الكلمات"، بل وتركز مثل هذه الأنطولوجيات أيضا على تحديد وتمييز تعدد معاني كل كلمة (Polysemy). بمعنى آخر يتم جمع كل كلمات لغة ما، بعد ذلك يتم تحديد مجموعة المفاهيم/المعاني التي تدل عليها كل كلمة، ويتم إعطاء رقم لكل مفهوم/معنى، وتعريفه دلاليا. بعد ذلك يتم تصنيف هذه المعاني (وليس الكلمات)، ويكون هذا التصنيف على أساس جنس الشيء، أي تعريف علاقة جنس من /جنس ل بين المفاهيم. بالإضافة لذلك يمكن أيضا أن يتم ربط المفاهيم بعلاقة جزء من /جزء ل، وعلاقة مرادف من /مرادف ل بين المفاهيم (وليس بين الكلمات). مع أهمية الملاحظة أن الأنطولوجيا تحوي علاقات دلالية وليست علاقات لغوية مثل مشتق، اسم فاعل، مصدر، صيغة مبالغة، وغيرها من العلاقات اللغوية بين الكلمات، (ولا تنطبق مثل هذه العلاقات اللغوية بين المعاني، وبالتالي هي ليست جزءا من الانطولوجيا). لتوضيح ما ورد أعلاه، نقدم المثال التالي:



مثال على تعريفات المفاهيم والعلاقات الدلالية فيما بينها

يتناول المثال في الشكل أعلاه كلمة "جدول"، حيث يحدد ويميز بين ثلاثة معان مختلفة لهذه الكلمة، ولكل معنى أعطي رقما يميزه وتعريفا دلاليا (Gloss) يحدد صفاته اللازمة والمميزة له دون غيره. لنأخذ على سبيل المثال المعنى رقم (6) لكلمة جدول: "مصفوفة بيانات مكونة من صفوف وأعمدة"، حيث يعتبر هذا المفهوم، تخصيصيا،–أي جنسا من- مفهوم رقم 10 لكلمة مصفوفة، وهو: "ترتيب لأشياء جنبا الى جنب". والمفهوم رقم (10) هو أيضا تخصيص لمفهوم رقم (14) لكلمة ترتيب. حيث أن العلاقة بين الأجناس هنا، هي علاقة بين المفاهيم المحددة لهذه الكلمات، وليست بين الكلمات نفسها، وبالتالي لا تنطبق هذه العلاقة على المفاهيم الأخرى لهذه الكلمات. كذلك يعتبر مفهوم رقم (6) لكلمة جدول هو جنس ل المفهوم رقم (3) لكلمة الجدول الدوري، وجنس ل مفهوم رقم (4) لكلمة أجندة. والعلاقة الدلالية (جنس من/جنس ل) هنا، هي علاقة مُعرّفة في علم المنطق، حيث تستوجب توارث الصفات. فعلى سبيل المثال، فإن جميع الصفات المعرفة في المفهوم رقم (14)، هي بالضرورة صفات لازمة تورث للمفهوم رقم (10)، وبالتالي فإن المفهوم رقم (6) يحمل بالضرورة صفات جميع المفاهيم التي تعلوه ضمن الشجرة المفاهيمية. وبالتالي يمكننا صياغة مفهوم رقم (6) لكلمة جدول على النحو التالي: "تنظيم البيانات بصورة ممنهجة، جنبا إلى جنب على شكل صفوف وأعمدة".

كما يُعرف الشكل أعلاه، فإن مفهوم رقم (2) لكلمة صف، هو جزء من المفهوم رقم (6). بمعنى أكثر دقة، كل جدول (حسب مفهوم رقم 6) يتكون حكما من أجزاء، أحدها يطلق عليه "صف"، حسب مفهوم رقم (2). وكذلك العلاقة الدلالية المُعرفة بين الشيء الذي يسمى نهر النيل (رقم 9)، والمفهوم رقم (11) لكلمة نهر، يطلق عليها الماصدق، بمعنى أن الشيء المحدد باسم نهر النيل يصدق جميع الصفات المُعرفة في مفهوم رقم (11) لكلمة نهر. كلمة الماصدق هنا، تعني حسب استخدام علماء المنطق العرب قديما: الشيء الذي يصدق صفات تصورية حول مفهوم ما (Instance of). مع أن المثال أعلاه قد يُعطي صورة مصغرة جداً عما يمكن أن تحويه الأنطولوجيا اللغوية، إلا أنه يجب الإشارة إلى أن هذه الأنطولوجيا سوف تحوي جميع كلمات اللغة العربية، ومفاهيمها، وارتباطاتها العلاقاتية، بشكل شجرة مفاهيمية.

كما ورد سابقا، تعتبر أنطولوجيا اللغة الانجليزية (WordNet) بداية لإنشاء الأنطولوجيات اللغوية، وأخذت زخما كبيرا [206،16،6] ليس فقط في التطبيقات، ولكن في مجال الأبحاث، حيث لحق بذلك عددا كبيرا من اللغات الأخرى مثل (الفرنسية، والألمانية، والإيطالية، والعبرية...)، والجدير بالذكر أن هذه الأنطولوجيات ليست مشجرات مفاهيمية منفصلة عن بعضها، بل لقد تم الربط بينها، وذلك من خلال ربط المفاهيم المتطابقة عبر اللغات المختلفة، أي نفس المفهوم يعطى رقما واحداً للتعبير عنه بغض النظر عن لغة المنشأ.

بالنسبة للغة العربية، لم تتم أي جهود جادة لبناء أنطولوجيا لغوية حتى الآن، ما عدا مشروعا يسمى "Arabic WordNet" قامت به وكالة الاستخبارات الأمريكية (CIA) منذ ثلاثة أعوام، حيث وظفت الوكالة بضعة باحثين لبنائها [1]. لم يستطع هذا المشروع أن ينجز سوى بضعة آلاف من المعاني ولم يتم تصنيفها بالكامل، إذ إتبع الباحثون أسلوب ترجمة أنطولوجيا اللغة الانجليزية، بما يقابلها باللغة العربية. وقد استنتج العاملون في هذا المشروع،–كما ذكروا في مقالاتهم [18،4]، بأن منهجية الترجمة لا يمكن أن تقدم نموذجا فعالا، وذلك كون المفاهيم اللغوية تعبر عن أنساق فكرية وثقافية، لا يمكن أن تتطابق في بنية إنتاجها، ودلالاتها بشكل حرفي، وذلك بالرغم من اشتراك اللغات في عدد كبير من المفاهيم. وعليه، فإن الاستنتاج هنا أن مثل هذا المشروع يستغرق سنوات عديدة، ولا يمكن إنجازه بشكل سريع عن طريق الترجمة.

من الجدير بالذكر أن هذا النوع من الأبحاث لا يعتبر جزءا من علم اللغة فقط، بل أنه في جوهره يعتبر ذو نزوع فلسفية منطقية، حيث يركز على تحديد الصفات اللازمة والجوهرية (Intrinsic Properties) وليس العرضية (Extrinsic Properties) واللغوية لمفهوم ما، وكذلك رسم العلاقات الدلالية بين هذه المفاهيم، هو موضوع فلسفي يقتضي فهم عميق لعلاقة هذه المفاهيم. بالإضافة إلى ذلك تعتبر منهجية التصنيف، وتمثيل هذه المعرفة آليا فرع من فروع المنطق الوصفي في علم الذكاء الصناعي.

1. المنهجية المتبعة لبناء الأنطولوجيا العربية

كما أسلفنا سابقاً فإنه وحسب معرفتنا لا توجد أي جهود جادة لبناء أنطولوجيا للغة العربية. من هنا جاءت الحاجة إلى بناء إطار دلالي تأسيسي للغة العربية بحيث يُمَكِّنها من اللحاق بركب باقي لغات العالم والتي قطعت شوطاً كبيراً في هذا الإتجاه والإندماج معها وتسهيل التفاعل والتبادل فيما بينها. كذلك الحاجة إلى إغناء تطبيقات الحاسوب التي تتعامل وتُعالج مصطلحات اللغة العربية كمحركات البحث والترجمة الآلية وغيرها وجعلها أكثر دقة وشمولية.

إن عملية بناء أنطولوجيا اللغة العربية تُعتبر مهمة طويلة الأمد، وتحتاج إلى جهودٍ متضافرة ومتعددة الاختصاصات المعرفية. ومن هنا فإن مشكلة البحث هي تأسيس هذا الحقل المعرفي، بما يسمح لإطلاق الأبحاث والمشاريع والمساهمات المستقبلية لبناء أنطولوجيا شاملة للغة العربية.

لبناء أنطولوجيا اللغة العربية، نقترح منهجية بحث وصفية تحليلية، تعتمد في مبناها الرئيسي على تطبيق معايير "وصفية منطقية" على المفاهيم قيد الدراسة. وفيما يلي عرض لمنهجية البحث، ضمن خمسة مكونات تشكل خطوات البحث، وتحمل كل منها عناصر منهجية، وتقييمية خاصة بها للتأكد من صحة النتائج والأدوات المستخدمة:

الخطوة الأولى: جمع وإستنباط تعريفات ومعاني من القواميس العربية المتاحة، سواء كانت متخصصة أو عامة، وذلك لجمع أكبر عدد ممكن من المصطلحات ومعانيها في عدة مجالات، حيث أن الكلمة يمكن أن يتعدد معانيها، وهذا ما نهدف إليه في هذه الخطوة. أي حصر كافة المعاني لكل مصطلح عربي. والجدير بالذكر أن غالبية قواميس اللغة العربية لا يمكن إستخدامها كونها تركز على تصريف الكلمات، وغالبا ما تخلط التصريف اللغوي بالمعنى الدلالي، بل وإن معظمها يدلل على المعنى بأمثله إيمائية، ولا يحدد المعنى تصريحا مباشرا. إن المعايير المتبعة في اختيارنا للقواميس,-كمصدر للمعاني الدلالية- تتلخص بما يلي:

(1) أن يكون القاموس غير مُعتمِد في بناءه على تصريف الكلمات، حيث أن الانطولوجيا العربية لا تهتم بتصريف الكلمات بل بالمعاني الدلالية وتعدد هذه المعاني لكل مصطلح.

(2) أن يكون المعنى الذي يحدده القاموس واضح، أي صريح غير مخلوط بمعاني أخرى، وأن يكون التعبير عن المعنى فقط بكلمات توضح هذا المعنى، دون الإشارة إليه أو خلطه بأمثلة من الشعر أو الأمثال والحكم وغيرها.

(3) جودة التعريف وطريقة تركيبه، إذ أن أحد الأمور المهمة لفهم المعنى بصورة صحيحة غير قابلة للخطأ هو أن يكون مكتوب بطريقة واضحة، ذات جودة عالية، وبكلمات صحيحة.

بناء على هذه المعايير قمنا بطباعة ورقمنة لحوالي مائة معجم، حتى الآن، وتوحيدها جميعا في قاعدة بيانات واحدة. لقد شارك في طباعة هذه القواميس حوالي ثلاثمائة طالب من جامعة بيرزيت كعمل تعاوني. كما تجدر الإشارة إلى اننا قد استنبطنا حوالي نصف مليون مفهوم وتعريف من هذه المعاجم.

الخطوة الثانية: إعادة صياغة وهندسة التعريفات (glosses) المستنبطة في الخطوة الأولى،حيث أن إعادة الصياغة هنا تتم يدويا باستخدام الضوابط الأنطولوجية التي تركز على الصفات الجوهرية المميزة للمفهوم دون غيره. وذلك بإتباع الضوابط كما عرفتها الورقة البحثية [10]، بعنوان "Towards The Notion of Gloss In Ontology Engineering". وتتلخص هذه الضوابط المنهجية بما يلي:

(1) بدء التعريف بالجنس الأعلى للمفهوم المراد تعريفه؛ فمثلا يجب أن يبدآ تعريف "مصفوفة" ب "ترتيب..." وتعريف "ترتيب" ب "تنظيم..." كما هو مبين في المثال السابق.

(2) ذكر الصفات الجوهرية المميزة للمفهوم، وليس الصفات العرضية أو الإشتقاقات اللغوية. فمثلا، تعرف المصفوفة ب "ترتيب لبيانات على شكل صفوف وأعمدة"، أي أن أية بيانات مرتبة على شكل صفوف وأعمدة تعتبر مصفوفة دون سواها.

(3) أن تكتب الصفات المدرجة بطريقة تصورية تقود لإستنباط المفهوم.

(4) الإشارة إلى صحة حالات شاع الإعتقاد بخطئها، وخطأ حالات شاع الإعتقاد بصحتها.

(5) تآلف التعريف المقدم مع موقعه (أي تصنيفه) ضمن الشجرة المفاهيمية.

(6) أن يتسم التعريف بالوضوح والإختصار (أي قاطعا مانعا ما أمكن).

بهذه الطريقة، نكون قد أنتجنا تعريفات دلالية ذات منطق صارم. ويمكن التأكد من صحة ذلك عند إعادة قراءة التعريفات. فمثلا، يمكن إعادة قراءة تعريف مصفوفة بأنها "تنظيم لبيانات مرتبة جنبا إلى جنب على شكل صفوف وأعمدة"، وذلك بدمج مفهوم "ترتيب" في هذا التعريف. كما يمكن أيضا قراءة تعريف مصفوفة بدمج مفهوم "تنظيم" أيضا، ليصبح التعريف "تنظيم البيانات بصورة منهجية جنبا إلى جنب على شكل صفوف وأعمدة".

تجدر الإشارة إلى أننا،-بالإضافة إلى الضوابط المقدمة أعلاه، نقوم أيضاً بإستعمال المنهجية التي تم تطويرها ونشرها مؤخراً بعنوان "نحو منهجية لبناء هندسة الأنطولوجيات -التصنيف بالصفات" [5] والتي تساعد على معرفة ما هي الصفات الجوهرية الدالة لمفهوم ما.

الخطوة الثالثة: ربط التعريفات المنتجة في الخطوة السابقة بما يقابلها في أنطولوجيا اللغة الإنجليزية (WordNet) إن وجد هذا المقابل. ولتحقيق ذلك، فقد قمنا بتطوير برنامج ذكي للقيام بهذا الربط بشكل آلي (بدقة وصلت إلى 90%). حيث يأخذ البرنامج التعريفات الدلالية من اللغة العربية (التعريفات المنتجة ضمن الخطوة السابقة) ويقوم بالبحث عن مقابلها الدلالي من قائمة تعريفات المفاهيم في انطولوجيا اللغة الإنجليزية. وتتلخص خوارزمية هذا البرنامج بخطوات أهمها:

(1) ترجمة آلية للتعريف العربي باستخدام (Google Translate) أو محركات الترجمة الأخرى المتاحة.

(2) إضافة جميع الإشتقاقات والمترادفات لكل كلمة واردة في ترجمة التعريف الناتجة عن الخطوة السابقة، كذلك إضافة جميع الكلمات ذات العلاقة الدلالية أو اللغوية الإشتقاقية المرتبطة. بمعنى آخر، يتم إنتاج سلة من الكلمات تحوي على الكلمات الواردة في الترجمة بالإضافة إلى كلمات أخرى ذات علاقة. فمثلا، قد يصل عدد الكلمات في السلة الى مئات وأحيانا آلاف.

(3) مقارنة جميع كلمات هذه السلة بكلمات كل تعريف ورد في قائمة تعريفات الانطولوجيا الإنجليزية، حيث تعطى علامة كل مرة تتطابق فيها الكلمات.

(4) حساب مدى تقارب التعريف (الأصلي) المدخل بكل تعريف إنجليزي بناء على نتيجة المقارنة السابقة. ضمن الحسابات الناتجة يُرشح التقارب الأعلى نسبة لأن يكون هو المقابل الإنجليزي للمفهوم العربي، والعكس صحيح. وقد وصلت دقة الربط لهذا البرنامج إلى 90%، بعد إدخال بعض التحسينات.

تجدر الإشارة هنا إلى أننا نستخدم هذا البرنامج ليس فقط لربط التعريفات والمفاهيم العربية بمقابلاتها الإنجليزية. بل أيضا للبحث عن تعريفات مكررة إو متداخلة في التعريفات العربية نفسها، وهذه حالة شائعة كون التعريفات مستنبطة آليا من عدة مصادر.

الخطوة الرابعة: بناء العلاقات الدلالية بين هذه التعريفات، للوصول إلى الشجرة المفاهيمية. تعتمد هذه الخطوة على الخطوتين السابقتين، حيث أن النجاح في تطبيق الضوابط الأنطولوجية في النقطة (1) في الخطوة الثانية وهي بدء التعريف بالجنس الأعلى منه تقود غالبا إلى تعريف علاقة جنس من/ جنس ل (SuperType/SubType) بين مفهومين أو أكثر، بمعنى أن لكل مفهوم "أب" واحد، ولكل أب أكثر من "ابن". كما أن النجاح في تطبيق الخطوة الثالثة أعلاه، أي الربط بين المفهوم العربي ومقابله الإنجليزي يقود إلى إستجلاب وإستنباط معظم العلاقات الدلالية من الانطولوجيا الإنجليزية، ولتوضيح ذلك: إذا وجد أن مفهوم (A) هو جنس ل مفهوم (B) في الأنطولوجيا الإنجليزية، وتم ربط مفهوم (س) باللغة العربية بمقابله (A)، وكذلك ربط مفهوم (ص) باللغة العربية بمقابله (B)، فهنا يمكننا الإستنتاج أن العلاقة بين (س) و(ص) هي ذاتها بين مقابليهما في اللغة الإنجليزية. مع ضرورة التأكيد أن هذه المنهجية ليست ترجمة بين الانطولوجيتين، وإنما ربط دلالي، وبالتالي فإن الإستنتاج هنا هو إستنتاج منطقي بحت.

ولتدقيق صحة التصنيف والشجرة المفاهيمية الناتجة يتم أتباع منهجية ال OntoClean المعروفة بصرامتها [21,8]المنطقية والفلسفية، وكذلك بإستعمال الشجرة العليا للغة العربية التي سنبيّنها لاحقا.

الخطوة الخامسة: ربط المفاهيم والتعريفات المنتجة في الخطوة الثانية وكذلك العلاقات المنتجة في الخطوة الرابعة بالمفاهيم العليا (Top-Level Concepts) للغة العربية، والتي تم بناءها بشكل منفصل. بمعنى آخر، لقد قمنا ببناء شجرة مصغرة تسمى الشجرة العليا (سيتم شرحها لاحقا) وتتكون من عشر مستويات (حوالي 400مفهوم) حيث تعتبر هذه المفاهيم العليا هي (إمهات معاني الكلمات العربية)، وتستخدم هذه المفاهيم للأهداف التالية:

(1) تستخدم كنواة يتم ربط جميع المفاهيم الأخرى بها. أي أن كل مفهوم منتج في الخطوة الثانية أعلاه يتم ربطه بإحدى المفاهيم في المستوى الأخير من الشجرة العليا. تجدر الإشارة إلى أن المقصود هنا ليس إدراج جميع مفاهيم اللغة العربية في مستوى إضافي، بل أن المستوى الأخير في الشجرة العليا يجب أن يعلو جميع المستويات المنتجة في الخطوة الرابعة.

(2) تستخدم كأداة للتحقق من صحة التعريفات والعلاقات المنتجة سابقا. أي أن هذه الشجرة العليا تستخدم للتحقق (آليا) من صحة العلاقات الناتجة في الخطوة الرابعة أعلاه. بمعنى آخر لضمان الجودة العالية في عملية التشجير (تصنيف المعاني) يجب أن تكون الأنطولوجيا على شكل شجرة،-وليس شبكة- من المعاني. أي يجب مراعاة ألا يكون للمعنى أكثر من جنس واحد يعلوه، ما أمكن ذلك. وذلك لأن تعدد الأجناس لمفهوم ما، ينجم غالبا عن عدم فهم لهذا المفهوم. إن وجود شجرة عليا صحيحة تعلو جميع المستويات يُمكّن من التحكم في صحة هذه العلاقات.

1. بناء المستويات العليا للأنطولوجيا العربية (الأنطولوجيا النواة)

لقد قمنا، وضمن مشرع ال (ArabicOntology) في جامعة بيرزيت، ببناء الشجرة العليا للمفاهيم العربية وذلك لتكون هذه الشجرة (وهي مكونه من عشر مستويات، أي ما يقرب 400 مفهوم) هي المستويات العليا للأنطولوجيا العربية النهائية. لقد إعتمدنا في بناء هذه الشجرة العليا على أنطولوجيات عليا أجنبية وهي (SUMO, KOYOTO ,BFO, DOLCE) وغيرها. حيث تمت دراستها وفهمها بصورة دقيقة وعميقة للتحقق من ملائمتها للغة العربية. لبناء الأنطولوجيا العليا العربية قمنا بالخطوات التالية:

أولاً :تحديد المفاهيم العليا للغة العربية:

1- تمت ترجمة (BFO, DOLCE, SUMO, KOYOTO) كلٍ على حده، حيث حددنا لكل مفهوم من هذه الأنطولوجيات عدد (ثلاث إلى خمسة) من الكلمات العربية التي تعتبر الأقرب تعبيرا عن هذا المفهوم. تجدر الإشارة أن عدد المفاهيم في DOLCE هو 80 مفهوم مصنفة في 7 مستويات، و تحوي SUMO على 700 مفهوم مصنفة في 15 مستوى. لقد بلغ عدد الكلمات العربية الناتجة عن ترجمة هذه المفاهيم (لكلتا الانطولوجيتين) ما يقارب 1200 كلمة. تهدف هذه الخطوة الى الحصول على سلة واسعة من الكلمات العربية التي قد تدلل على معاني عليا في اللغة العربية.

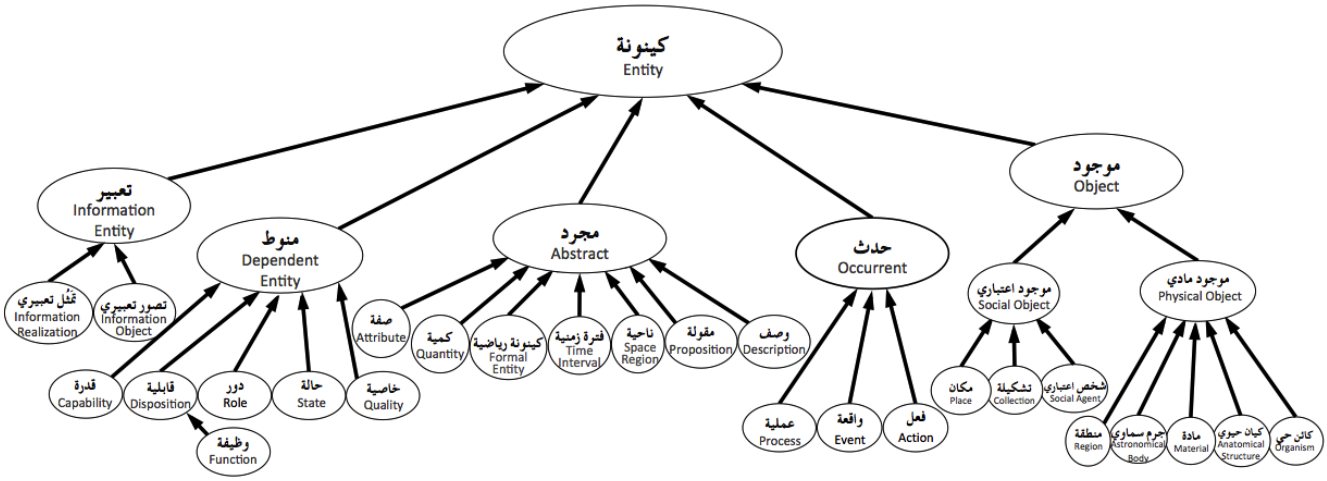
2- لكل كلمة عربية (ضمن ال 1200) تم تحديد جميع مفاهيم هذه الكلمة، حيث بلغ عدد معاني جميع الكلمات حوالي 5000 معنى، أي بواقع 4.2 معنى لكلٍ من ال 1200 كلمة. وقد إعتمدنا في تحديد معاني هذه الكلمات على أمهات وأصول المعاجم العربية ودراستها بشكل موسع آخذين بعين الإعتبار الأبعاد الفلسفية والتاريخية لهذه المعاني. وكذلك تمت عملية صياغة تعريفات هذه المعاني (Glosses) بناءً على الضوابط المبينة سابقا. تهدف هذه الخطوة إلى الحصول على سلة واسعة تحوي المفاهيم الأكثر عمومية وشمولية من غيرها.

3- اختيار معنى واحد مُعبر لكل مفهوم من المفاهيم الواردة في كلتا الانطولوجيتين (SUMO و DOLCE)، وذلك إعتماداً على المفاهيم ال (5000) الناتجة من الخطوة السابقة. وتجدر الإشارة هنا إلى أن هذه الخطوة هي ربط دلالي (Semantic Mapping) بين المفاهيم،-وليست ترجمة، حيث أن عملية تحديد معاني الكلمات العربية (في الخطوة السابقة) كانت عملية مستقلة عن الكلمات الإنجليزية.

ثانياً :بناء الشجرة العليا للمفاهيم العربية: بما أننا حددنا المفاهيم العربية المقابلة للمفاهيم الموجودة في كلٍ من (SUMO و DOLCE)، وبما أن كلٍ منهما تحوي علاقات تجنيس بين المفاهيم (والتي تعتبر عليا ومستقلة عن التطبيقات واللغات) قمنا وإعتماداً على هذه العلاقات بإستنباط العلاقات بين المفاهيم العربية، أي الشجرة العليا للأنطولوجيا العربية. أنظر الشكل أدناه

والذي يبين بعض المستويات العليا لهذه الشجرة. وقد إعتمدنا على العلاقات الموجودة في (DOLCE) لإنشاء المستويان الأولين في الشجرة، وذلك لإعتقادنا أنها أكثر صحة وعمقا في الفهم، إلا أننا قمنا بإدخال بعض التعديلات الصغيرة، وخاصة فيما يختص بالمفاهيم متعددة الأجناس (Multiple Inheritance) إذ تم فصلها إلى عدة مفاهيم وذلك للحصول على شجرة (Proper Subtypes). وإضافةً إلى المفاهيم الموجودة في (DOLCE)، قمنا بدراسة كل مفهوم في (SUMO) وإضافته إلى شجرتنا قيد الدراسة. حيث تم تبني بعض المفاهيم وتجاهل بعضها كوننا نعتقد أنها مفاهيم خاصة وليست عليا.

ثالثاً :التحقق من صحة الشجرة العليا الناتجة، وذلك لتقييم ما إذا كانت الشجرة العليا الناتجة من الخطوة السابقة هي فعلاً عليا، أي أنه لا يوجد مفهوم أخر في اللغة العربية يعلو أحد المفاهيم الواردة فيها. وعليه فقد قمنا بتجنيس جميع المفاهيم ال 5000 الناتجة في الخطوة الأولى أعلاه. بمعنى آخر، قمنا بربط (علاقة "جنس") كل مفهوم من ال 5000 بإحدى المفاهيم في المستوى الأخير من الشجرة العليا. والمقصود هنا هو التحقق من أن المفاهيم في المستوى الأخير (المستوى الأدنى في الشجرة العليا) هي أجناس عليا لجميع المفاهيم الأخرى. تنبع أهمية هذه التجربة من كون ال 5000 مفهوم هنا تمثل المفاهيم الأكثر عمومية في اللغة العربية، كما بّينا سابقا.



جزء من المستويات العليا للأنطولوجيا العربية

2. ملخص الإستنتاجات والخطة المستقبلية

لقد نمت الحاجة مؤخراً لوجود أنطولوجيا للغة العربية، وذلك بسبب زيادة التوجه لإستخدام اللغة العربية في كثير من التطبيقات والمجالات، مما يستدعي الحاجة لوجود وسيط يقوم بتعريف المصطلحات العربية بمفاهيم دلالية تتيح المجال لفهم، مشاركة وتبادل البيانات بصورة واضحة ودون أي غموض. وعليه، إنطلق مشروع بناء انطولوجيا اللغة العربية في جامعة بيرزيت، متبعاً منهجية صارمة منطقيا ومؤصلة فلسفيا. ومنذ إنطلاقه، حقق المشروع عدة إنجازات يمكن تلخيصها بما يلي:

(1) بناء المستويات العليا لأنطولوجيا اللغة العربية، والتي تشكل أساس الأنطولوجيا العربية، وسيتم ربط كافة مصطلحات اللغة العربية بها، بدافع إيجاد علاقات مفاهيمية مؤصلة بين المعاني العربية، والتحقق من صحة هذه العلاقات.

(2) تم طباعة ورقمنة لحوالي مائة معجم، وتوحيد جميعا في قاعدة بيانات واحدة، تحوي ما يقارب نصف مليون مفهوم، مما يشكل اللبنة الأساسية للانطولوجيا العربية.

(3) إنشاء برنامج حاسوب مبنى على خوارزمية ذكية تعمل على الربط بين مفاهيم الأنطولوجيا العربية، مع مقابلاتها في أنطولوجيا اللغة الإنجليزية.

أن من أهم صفات الأنطولوجيا العربية التي نسعى لبنائها، مقارنةً مع الإنجليزية، أن (1) العلاقات الدلالية مؤصلة فلسفياً ومنضبطة منطقياً (Formal Logic)، وبالتالي لا يشوبها غموض دلالي. (2) صياغة تعريفات المفاهيم (Glosses) تحكمها ضوابط أنطولوجية في الشكل والمضمون. (3) المستويات العليا العربية مؤصلة فلسفيا ومنذ البداية، اعتمادا على أهم الأنطولوجيات العليا العامة (Upper Level Ontolgies) وليس ربطها ربطا خارجيا بهذه الأنطولوجيات بعد استكمالها، كما هو الحال في الأنطولوجيا الإنجليزية (WordNet).

وكخطوة مستقبلية، نسعى إلى مضاعفة حجم قاعدة البيانات لتضم عدد أكبر من المصطلحات العربية وتعريفاتها الدلالية، حيث سنتابع البحث عن مصادر وقواميس ليتم إستنباط معاني مصطلحاتها تبعاً للأسس التي ذكرت سابقاً. وسوف نعمل على تطوير البرنامج بحيث تزيد نسبة دقة المطابقة عن 90%، لنرقى إلى نتائج أفضل. وبطبيعة الحال سوف يتم ربط أية مصطلحات جديدة -تضاف إلى قاعدة البيانات، بالمستويات العليا للأنطولوجيا، لنحصل على أكبر كم ممكن من المصطلحات العربية التي تربطها علاقات دلالية، أي بمعنى آخر، لنحصل على الشجرة المفاهيمية.

شكر وعرفان

بداية أنوه، أن هذا البحث ممول جزئيا من جامعة بيرزيت. وهنا أتقدم بالشكر والتقدير لكل من أسهم في وضع اللبنات الأولى لهذا المشروع البحثي. أتقدم بالشكر الجزيل للباحث وسيم أبو فاشة، وذلك لمساهمته في إثراء البحث لغويا وفلسفيا. وشكري الجزيل للطالبة رنا رشماوي، لمساهمتها في بناء الأنطولوجيا العليا للغة العربية، عبر رسالتها في الدراسات العليا في الموضوع ذاته. كما أتقدم للعديد من الباحثين والزملاء والطلبة بالشكر على مساهماتهم المختلفة، وهم: أ.د. مهدي عرار، أ. جمال ضاهر، د. هنادا خرمة، وأنطون دعيق.

المراجع

1. Al Muqtafi Page http://muqtafi2.birzeit.edu/ (Visited ،September 2010)
2. Arabic WordNet http://www.globalwordnet.org/AWN (Visited ،September 2010)
3. Berners-Lee T. ،Fischetti M.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. Harper ،San Francisco. (1999)
4. Elkateb S. ،Black W. ،Vossen P. ،Farwell D. ،Rodriguez H. ،Pease A. ،Alkhalifa M.: Arabic WordNet and the Challenges of Arabic. In Proceedings of Arabic NLP/MT Conference (2006)
5. Daher J. ،Jarrar M.: نحو منهجية لبناء هندسة الانطولوجيات - التصنيف بالصفات. In proceedings of the 3rd Palestinian International Conference on Computer and Information Technology (PICCIT 2010). Hebron ،Palestine. March 2010.
6. Global WordNet http://www.globalwordnet.org/ (Visited ،September 2010)
7. Gruber ،T.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies ،43(5/6) (1995)
8. Guarino N. ،Welty C.: Evaluating Ontological Decisions with OntoClean. Communications of ACM 45(2):61-65. (2002).

9. Guarino ،N.: Formal Ontology in Information Systems. Proceedings of FOIS'98 ،IOS Press ،Amsterdam. pp. 3–15(1998)
10. Jarrar M ،Ayesh S ،Al-Badawi M ،Samara H.: Towards Building An Arabic Ontology. Technical Report. Faculty of Information Technology ،Birzeit University. 2010.
11. Jarrar M. ،Meersman R.: Ontology Engineering -The DOGMA Approach. Book Chapter in "Advances in Web Semantics";. Volume I ،LNCS 4891 ،Springer.ISBN:978-3540897835. (2008).
12. Jarrar M: Towards the notion of gloss ،and the adoption of linguistic resources in formal ontology engineering. In proceedings of the 15th International World Wide Web Conference (WWW2006). Edinburgh ،Scotland. Pages 497-503. ACM Press. ISBN: 1595933239.)2006(.
13. Jarrar M.: Towards methodological principles for ontology engineering. PhD Thesis. Vrije Universiteit Brussel. (May 2005)
14. Meersman R: Ontologies and Databases: More than a Fleeting Resemblance. OES/SEO Workshop ،Luiss Publications (2001)
15. Miller G. ،Beckwith R. ،Fellbaum F. ،Gross D. ،Miller K.: Introduction to WordNet: an on-line lexical database. International Journal of Lexicography ،3(4). (1990) pp. 235–244
16. Multi WordNet http://multiwordnet.fbk.eu/english/home.php (Visited ،September 2101)
17. Pease A. ،Niles I.: IEEE standard upper ontology: a progress report. In The Knowledge Engineering Review 17(01). Pages 65-70. Cambridge Univ Press. 2002
18. Rodriguez R. ،Farwell D. ،Farreres J. ،Bertran M. ،Alkhalifa M. ،Marti A. ،Black W. ، Elkateb S. ،Kirk J. ،Pease A. ،Vossen P. ،Fellbaum C.: Arabic WordNet: Current State and Future Extensions. Proceedings of The Fourth Global WordNet Conference ،Szeged ، Hungary. January 22-25. (2008)
19. Smith B.: An Introduction to Ontology: From Aristotle to the Universal Core. Online Training course. http://ontology.buffalo.edu/smith/IntroOntology_Course.html (Visitedو September 2010)
20. Vossen P. (eds.): EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers ،Dordrecht. (1998)
21. Welty C. ،Guarino N.: Support for Ontological Analysis of Taxonomic Relationships. J. Data and Knowledge Engineering. 39(1):51-74. 2001.