الفصل السابع

مذكرة حول علم الدلالة العربية

A Note on Arabic Semantics

يعرف علم الدلالة (Semantics) على أنه دراسة معنى التعابير اللغوية (linguistic expressions). ويعتبر عدد أبحاث النماذج الحاسوبية (computational models) في علم الدلالة أقل بكثير من مجالات أخرى في حقل معالجة اللغة الطبيعية. وقد يعود السبب في ذلك لطبيعة المجال المعقد والدقيق، وبالطبع لا تختلف الأبحاث في علم الدلالة في مجال معالجة اللغة العربية عن هذا الحال.

في هذا الفصل، سنبدأ بمذكرة موجزة حول المصطلحات المستخدمة في علم الدلالة. يليها استعراض لمجموعة من المصادر التي طورت للنمذجة الحاسوبية الدلالية للغة العربية وبعض تطبيقاتها المرتبطة. ولن نتطرق للنقاشات المختلفة حول النظريات الدلالية وتمثيل الدلالة في هذا الكتاب.

٧,١ مذكرة موجزة حول المصطلحات

تميل المصطلحات المستخدمة في الحديث عن علم الدلالة إلى استقلاليتها لغوياً، ليست كما في الصرف والهجاء. على سبيل المثال، تعتبر المفاهيم الأساسية للمشترك

اللفظي أو ما يطلق عليه بالإنجليزية الهومونيمي (homonymy)[١] والترادف (synonymy)[٢] والأدوار الدلالية (Semantic roles)[٣] إلى حد كبير هي نفسها عند استخدامها للغة العربية أو الإنجليزية. ولمعلومات أكثر اطلع على [159، 160، 161].

وبقولنا هذا، نجد أن خصوصيات اللغة العربية، مثل غناها الصرفي وغموضها الهجائي بسبب التشكيل الاختياري، يمكن أن تؤدي إلى عدد كبير من الجناس وبذلك تصبح اللغة أكثر غموضاً قياساً بما يمكن إيجاده في اللغة الإنجليزية. بالإضافة إلى ذلك، وبلا اختلاف عن اللغات الأخرى، فإن الكلمات العربية تُمثل وتفرق بين الجوانب المختلفة للمعنى بامتياز، وهي بذلك لا تختلف عن اللغات الأخرى. على سبيل المثال، كلمة قلم (*qalam*) تستخدم للدلالة على (pen) و(pencil) في اللغة الإنجليزية. أما كلمة صلاة (*SlAħ*) فهي تستخدم في العبادة وليس للطلب كما يحدث في اللغة الإنجليزية[٤]. من الشائع والرائع للغة العربية احتواؤها على عدد كبير من الكلمات

---

(١) المشترك اللفظي (أو الجناس): هي حالة وجود كلمتان متشابهتان في الصورة اللفظية (نفس الإملاء والنطق) لكنهما مختلفتان في المعنى، على سبيل المثال، كلمة بيت (bayt) تعني إما المكان الذي يسكن فيه أو بيت الشعر. إذا كانت هذه الكلمات لها نفس الإملاء ولكن بنطق مختلف، تسمى في هذه الحالة جناس خطي (homographs)، على سبيل المثال الكلمة "حب Hb" بدون تشكيل يمكن أن تنطق حُب /Hubb/ أو حَب /Habb/. لكن إذا كانت الكلمتان لهما نفس النطق ولكن بإملاء مختلف تسمى في هذه الحالة التطابق اللفظي أو الهوموفون (homophones)، مثال ذلك: عصى (çaSaý) وعصا (çaSaA) كلاهما لهما نفس النطق /çaSa/. وعليه فإن الهومونونيمي لابد أن يكون جناساً ومتطابقاً لفظياً في نفس الوقت.

(٢) الترادف هو حالة كلمتين لهما معنى مماثل ولكن بشكلين مختلفين، مثال ذلك: بيت (bayt) ودار (dAr).

(٣) الدور الدلالي هو العلاقة الكامنة بين المسند والمسند إليه بغض النظر عن التعبير النحوي للمسند إليه. وتسمى أيضاً الأدوار الدلالية بالأدوار الموضوعية أو أدوار ثيتا (theta). مثال ذلك جملة "كتب علي كتابا" (kataba çaliy~ũ kitAbAã)، كتب هو الفعل المتوقع وعلي هو الفاعل والكتاب هو المفعول به.

(٤) معنى كلمة (Word sense) هو مصطلح تقني يشير إلى معنى محدد للكلمة.

لكلمة مثل (جمل). غير أن أغلب الناطقين باللغة العربية لا يعرفون أكثر من كلمتين، خاصة "جمل" (*jamal*) و"ناقة (*nAqah*) (أنثى الجمل) و"إبل" (*Ăibil*) (جمع جمل). الكلمات الأخرى لكلمة "جمل" تعتبر جزءاً من لغة مربي الإبل والمختصين بها، مثال ذلك كلمة "حوار" (*HuwAr*) و"لبون" (*labuwn*) و"خلوج" (*xaluwj*). هذا الوضع مشابه للكلمات العديدة المستخدمة مع كلمة "حصان" في اللغة الإنجليزية التي يستخدمها مربو الخيول، على سبيل المثال: كلمة (*foal*) وتعني الحصان الصغير بجانب أمه وكلمة (*gelding*) وتعني الحصان المخصي. وفي هذا الشأن، فإن اللغة العربية لا تختلف عن اللغات الأخرى، كما ذكرنا آنفاً.

٧,٢ بنك أبنية الحمل/الإسناد العربية

يعتبر بنك أبنية الحمل/الإسناد (Proposition Bank (propbank) نوعاً من المدونات المحشاة دلالياً. ففي بنك أبنية الحمل/الإسناد تُوضع حواشي القضايا (prepositions) وإسناداتها على هيئة معلومات بصيغة (المسند- المسند إليه) (predicate-argument) مع تسميات للأدوار الدلالية تضاف فوق بنك شَّجريّ نحوي (syntactic treebank) قائم مسبقاً [162]. وتتضمن المدونات (corpora) الهامة المحشاة دلالياً شبكة الأطر (Framenet) [163].

وتطور جامعة كولورادو الأمريكية حالياً بنك أبنية الحمل/الإسناد العربية Arabic Propbank (APB) باتباع منهجية للتطوير مماثلة لتلك المستخدمة لعمل بنك أبنية الحمل/الإسناد للغة الانجليزية والصينية [162، 164]. وقد بنيت APB على البنية النحوية لبنك بنسلفانيا الشجري للتحاليل النحوية (PATB) مع الالتزام بها [151]. كما

أن APB لها القدرة على الوصول لتوسيمات الكلمة المعجمية (Lemma annotations) و(dashtags) الدلالية الموجودة في PATB.

يحدد بنك أبنية الحمل/الإسناد قائمة بمجموعات الأطر (framesets) لكل فعل. وتحدد مجموعة الأطر الواحدة المعنى المتنبأ للفعل وعدد ودور إسناداته[١]. عادة لا تضمن الملحقات (Adjuncts)، التي توسع من معنى الجملة، وليست ضرورية للفعل الإسنادي (predicate verb)، ضمن مجموعة الأطر. الشكل رقم (٧,١) يوضح بالأمثلة مجموعة الأطر الخمسة المرتبطة مع الفعل قام (*qAm*).

رقم الإطار	تعريف الإطار	مثال
F1	لتنفيذ أو لإجراء to carry out or to undertake Arg0: implementer (المنفذ) Arg1: implemented التنفيذ))	Arg1[ب+ رسم الصورة] Arg0[الفنان] Pred [قام] [*qAm*]*Pred* [*AlfnAn*]*A1g0* [*b+ ism AlSwrh*]*A1g1* [*The artist*]*A1g0* [*undertook*]*Pred* [*the painting of the picture*]*A1g1*
F2	لبدء أو ليحدث to start or to happen Arg1: event (حدث)	Arg1[الحرب] Pred [قامت] [*qAmt*]*Pred* [*AlHrb*]*A1g1* [*The war*]*A1g1* [*started*]*Pred*
F3	القيام أو تحديد موقع to stand or be located Arg1: thing standing (الشـــيء الواقف) Arg2: location (موقع)	Arg2[ب+ جانب الكنيسة] Arg1 [المسجد] Pred [يقوم] [*yqwm*]*Pred* [*Alm.jd*]*A1g1* [*b+ jAnb Alknys.h*]*A1g2* [*The mosque*]*A1g1* [*is located*]*Pred* [*next to the church*]*A1g2*

الشكل رقم (٧,١). مجموعة الأطر المختلفة المرتبطة بالفعل "قام" مع الأمثلة.

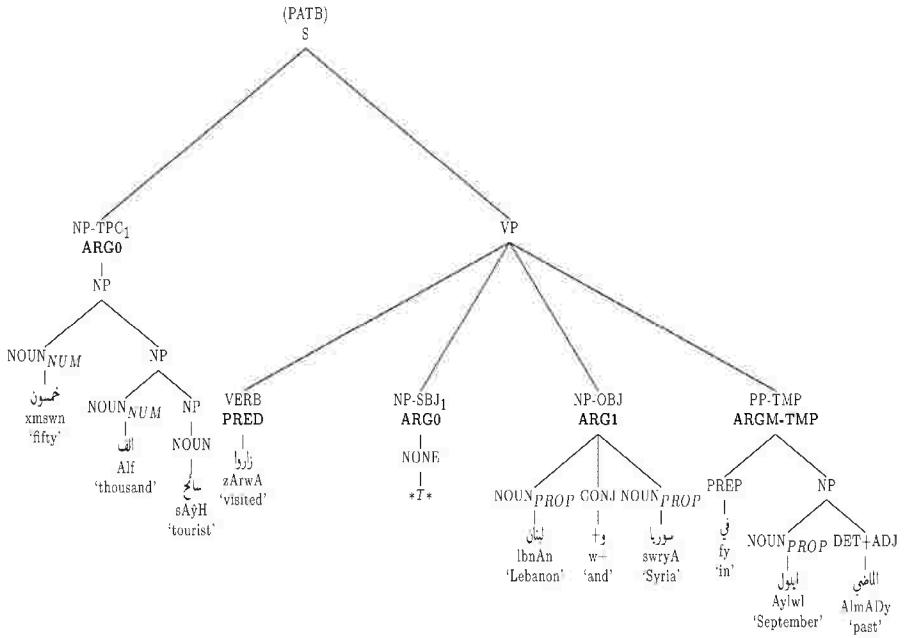
(١) http://verbs.colorado.edu/propbank/framesets-arabic/

F4	يقف to stand up Arg1: person standing (شخص واقف)	Arg1[الرجل] Pred [قام] [*qAm*]*Pred* [*Alrjl*]*A1g*1 [*The man*]*A1g*1 [*stood up*]*Pred*
F5	يتكون من to consist of Arg1: whole (الكل) Arg2: parts (الجزء)	Arg2[على أربعة مراحل] Arg1 [المشروع] Pred [يقوم] [*yqwm*]*Pred* [*Almšrwç*]*A1g*1 [*çlY Arbç mrAHl*]*A1g*2 [*The project*]*A1g*1 [*consists*]*Pred* [*of four phases*]*A1g*2

تابع الشكل رقم (٧,١).

أيضاً يعرف APB أربعة وعشرين نوعاً من الإسنادات (arguments)، التي تتضمن خمسة إسنادات رئيسية مرقمة كالتالي (ARG0, ARG1, ARG2, ARG3, ARG4) وتسعة عشر إسناداً مساعداً تتضمن المساعد المؤقت (ARGM-TMP)، ومساعد النفي (ARGM-NEG). كما يسمح استخدام الإسنادات المرقمة لبنك الأبنية الحمليّة من التقاط التعميمات الخاصة بمجموعات الأطر لفعل معين من دون الحاجة للاختيار بين مجموعة محدودة من الأدوار الموضوعية/الدلالية. الشكل رقم ٧,٢ يوضح مثالاً كاملاً على شجرة محشاة[1].

(١) يُحتفظ بتوسيمات APB في ملف منفصل عن شجرة PATB التي وسعوها. ويهدف الشكل الذي نستخدمه هنا للتوضيح.



الشكل رقم (٧,٢). تحشية بنك أبنية الحمل/الإسناد العربية لبنك بنسلفانيا الشجري للتحاليل النحوية لجملة (خمسون الف سائح زاروا لبنان وسوريا في ايلول الماضي) (xmswn Alf sA^yH zArwA lbnAn wswryA fy Aylwl AlmADy). الفعل المسند الأساسي "زار" لديه مجموعة أطر واحدة فقط بإسنادين هما الكيان الزائر ARG0 والكيان الذي تم زيارته ARG1. وقد تم إسناد الطرف NP-TPC للقيمة ARG0 بطريقة غير مباشرة من خلال مؤشر مشترك NP-SBJ.

وقد تم بالفعل استخدام بنك الأبنية الحمليّة العربية من قبل الباحثين لعملية عنونة الأدوار الدلالية Semantic Role Labeling (SRL) [165, 166] .

٧,٣ شبكة الكلمات العربية (ARABIC WORDNET)

شبكة الكلمات (wordnet) هي قاعدة بيانات معجمية قابلة للقراءة بواسطة الآلة، تعمل على جمع الكلمات على شكل مجموعة من المترادفات يطلق عليها اسم المجموعات المترادفة (synsets). ويمكن اعتبار كل مجموعة مترادفة تمثل معنى لكلمة فريدة من نوعها (للمعنى أو المفهوم). وعادة ما توفر شبكة الكلمات تعريفات عامة وأمثلة للمجموعات المترادفة وتضمين العلاقات الدلالية فيما بينها. العلاقات الدلالية التي تتضمن علاقات التضمين أو ما يسمى الهايبونمي (hyponymy)[١] والاشتمال أو ما يسمى الهايبرنمي (hypernymy)[٢]، تسمح لشبكة الكلمات بالتشكل على شكل هرمي من التصنيفات أو الأنطولوجيا المعجمية (lexical ontology).

تعتبر شبكة كلمات برينستون الإنجليزية (Princeton English WordNet) أول شبكة كلمات منتجة [167]. وقد تبعتها جهود كثيرة مماثلة وامتدادات لها من أبرزها شبكة الكلمات الأوربية (EuroWordNet) [168] . كما نُسقت العديد من الجهود بين شبكات الكلمات المختلفة لتشمل روابط تبادلية فيما بينها (cross links)، مما يسمح باستخدامها ليس فقط باعتبارها مكنزاً حاسوبياً متطوراً أحادي اللغة ولكن كقواميس أيضاً.

في عام ٢٠٠٦م بدأ تطوير شبكة الكلمات العربية (ArabicWordNet)[٣] بتضافر جهود العديد من الجامعات والشركات [169، 170]. ويستند تصميم ومحتوى AWN

---

(١) الهايبونمي: تضمين كلمة في مجال دلالي إلى كلمة أخرى، مثال ذلك: كلمة رجل تندرج تحت مجال الإنسان.

(٢) الهايبونمي: اشتمال كلمة في مجال دلالي لكلمة أخرى، مثال ذلك: كلمة إنسان يندرج تحتها كلمة رجل.

(٣) تختصر (AWN).

على شبكة كلمات برينستون. وقد رُبطت مجموعات المترادفات العربية مع مجموعات المترادفات في شبكة كلمات برينستون وأيضاً مناظرتها لمجموعات المترادفات في شبكة الكلمات الأوربية. كما تُمثل الكلمات العربية في AWN بكلماتها المعجمية (Lemma) وذلك لتجريدها من التغيرات الصرفية.

يوضح الشكل رقم (٧,٣) الأشكال المختلفة لمجموعات المترادفات لفعل "زار" (*zAr*) وترجمته للغة الإنجليزية (*visit*). حيث يمثل كل صف في الشكل مجموعتين من مجموعات المترادفات التي رُبطت ببعض (للغة العربية والإنجليزية). وتعتبر شبكة الكلمات الإنجليزية أكبر من شبكة الكلمات العربية التي تحتوي على ثلاث مجموعات إضافية للمترادفات ليس لها نظير حالياً في AWN. الاقتران الموضح في الجدول رقم ٧,٣ يسلط الضوء على بعض الاختلافات المهمة بين الكلمتين. فالفعل زار (*zAr*) لا يتضمن معنى الإلحاق الممكن مع الفعل (*visit*). كما يمكن القول أيضاً، أن فعل "زار" أكثر عمومية وذلك لأنه يرتبط بمجموعة المترادفات لكلمة (tour)، التي تشتمل فعل (Visit). وفي الجدول رقم ٧,٣ لا تظهر علاقة الاشتمال (hypernymy). وللمعلومية يمكن الوصول إلى شبكة الكلمات العربية AWN وشبكة الكلمات الإنجليزية وتصفحهما من شبكة الانترنت.

استخدمت AWN كمرجع معجمي لتقييم أنظمة فك اللبس الدلالي للعربية word sense disambiguation (WSD) [166] . ففي أنظمة فك اللبس الدلالي، تُوسم الكلمات مع معانيها المحددة في السياق باستخدام تعريفات المعنى من مصدر معجمي محدد مسبقاً.

مجموعة المترادفات في شبكة الكلمات العربية	مجموعة المترادفات في شبكة الكلمات الإنجليزية
زار zAr, تجول tjwl, جال jAl, جاب jAb, دار dAr, طاف TAf	tour
زار zAr, رأى rÂý, شاهد šAhd	visit, see
زار zAr, سافر إلى sAfr ĂlY	visit, travel to
زار zAr	visit, call in, call
وجه wjh, فرض frD, صب Sb, أبتلى Abtlý, أزعج ÂzEj, أنزل Ânzl, أصاب ÂSAb	visit, inflict, bring down, impose
تسامر tsAmr, تحادث tHAdθ, دردش drdš	visit, chew the fat, shoot the breeze, chat, confabulate

الشكل رقم (٧,٣). اقتران لمجموعات المترادفات من شبكة الكلمات العربية مع نظيرتها الإنجليزية.

٧,٤ المصادر العربية لاستخلاص المعلومات

الهدف من استخلاص المعلومات (Information Extraction (IE)) هو استخراج معلومات منظمة ومعرفة دلالياً بشكل آلي من الوثائق غير المنظمة والنصوص الخام ، مثال ذلك: تحديد أسماء المواقع الجغرافية في النص. وعادة ما تتضمن المهام الفرعية لاستخلاص المعلومات التالي: الكشف عن الإشارة (mention detection)، وحل مشكلة الإحالة (coreference resolution) واستخراج العلاقة (relation extraction). وتتضمن عملية الكشف عن الإشارة تحديد ثلاث فئات من تواردات الكيان (entity mentions) وهي: اسم العلم (مثال: باراك اوباما *bAraAk AwbAmA* – Barack ʻ Obama’)، والمصطلح الاسمي (مثال: الرئيس الأمريكي *Alrŷys AlÂmryky* - ‘the – American President’)، والضمير (مثال: هو – *hw* – ‘he’ʻ).

ويمكن تصنيف كل إشارة إلى واحدة من خمسة أنواع هي: الشخص (PER)، والمنظمة (ORG)، والمكان (LOC)، والكيان الجيوسياسي (GPE)، والوسيلة (FAC).

أما مهمة تمييز الأعلام (Named Entity Recognition (NER فتركز على الكشف عن الإشارة وتصنيف الكيانات الاسمية فقط. وتعمل مهمة حل مشكلة الإحالة على ربط الإشارات المختلفة في النص بعضها مع بعض، مثال ذلك: تمييز الضمير هو (*hw*) على أنه راجع للاسم باراك اوباما (*bArAk AwbAmA*) وليس لشخص آخر. وتحدد استخراج العلاقات نوع العلاقة بين كيانين مختلفين، مثال ذلك شخص معين (PER) موجود في كيان جيوسياسي محدد (GPE).

وقد طُورت المصادر العربية لاستخلاص المعلومات كجزء من برنامج الاستخلاص التلقائي للمحتوى Automatic Content Extraction (ACE) [171][1]. ويتضمن ذلك نصوصاً مُحشاة لمختلف عمليات استخلاص المعلومات، التي يمكن استخدامها لتطوير واختبار أنظمة استخلاص المعلومات [173, 129, 128, 172]. الشكل رقم (٧,٤) يوضح مثالاً على الناتج من نظام تمييز الأعلام.

زار <PER>الملك حسين<GPE> </PER><GPE>لبنان</GPE> في العام الماضي.

zAr <PER>*Almlk Hsyn*</PER> <GPE>*lbnAn*</GPE> *fy AlçAm AlmADy.*

<PER>King Hussein</PER> visited <GPE>Lebanon</GPE> last year.

الشكل رقم (٧,٤). مثال على جملة باللغتين العربية والإنجليزية مع وسوم تمييز الأعلام باستخدام XML.

(١) إرشادات التحشية ببرنامج الاستخلاص التلقائي للمحتوى للغة الإنجليزية والعربية والصينية موجودة في هذا الموقع http://projects.ldc.upenn.edu/ace/data/

٧,٥ المزيد من القراءات

في هذا القسم، سنقدم قائمة مختصرة من الإشارات لجهود جديرة بالذكر في النمذجة الدلالية للعربية. كما أن هناك روابط إلى موسوعات وقواميس ومكانز في ملحق ب و ج.

- في [174] نجد وصفاً وتقييماً لمنهجية تطوير المصادر ونظاماً لفك اللبس الدلالي للعربية.
- نظام (OntoNotes) هو جهد لتوسيم النص الإنجليزي والعربي والصيني لمختلف أنواع نظم الجملة، والبنية الإسنادية، ومعنى الكلمة وإحالتها [152].
- بحث مشروع مدونة التحشية بين اللغات لنص متعدد اللغات (Interlingual Annotation for Multilingual Text Corpora (IAMTC) تمثيل مشترك لتوصيف الظواهر الدلالية المتزايدة في سبع لغات هي (العربية، والهندية، والإنجليزية، والإسبانية، والكورية، واليابانية، والفرنسية) [175].
- بنك براغ للتحاليل الشجرية النحوية التبعية للغة العربية (Prague Arabic Dependency Treebank) يتضمن بعض التحشيات من نوع تكتيكات نحوية (Tectogrammatics)[1] ، ونظماً للجملة التي تعكس المعنى اللغوي من الكلام [138].
- مدونة المفاهيم الوسيطة (Conceptual Interlingua) التي تستخدم كمصدر في استرجاع المعلومات التي تعتمد على شبكة كلمات برينستون الإنجليزية تم توسعتها لتتضمن مصطلحات عربية كما في [176].

---

(١) هي الطريقة التجريدية التي تُبنى فيها العلامات اللغوية (توضيح من المترجمة).

- مدونة تحشية فهم اللغة (The Language Understanding Annotation Corpus) هي مدونة تجريبية مكونة من نصوص بالإنجليزية والعربية محشاة بإحالات لحقائق، وأحداث وكيانات، وأعمال الحوار والعلاقات الزمانية.
- في [14] بُنيت مدونة بتحشية للتعبيرات العددية الطبيعية واستخدمت لتقييم نظام للكشف التلقائي عن هذه التعبيرات.