مباحث لغوية (١٤)

المدونات اللغوية العربية

بناؤها وطرائق الإفادة منها

تأليف

أ.د.محمود إسماعيل صالح

أ.عبدالله يحيى الفيفي

د.عبدالمحسن عبيد الثبيتي

د.عقيل حامد الشمري

د.سلطان ناصر المجيول

تحرير

د. صالح بن فهد العصيمي

الطبعة الأولى

الرياض

١٤٣٦ هـ - ٢٠١٥م

ⓒ مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٣٦ هـ

فهرسة مكتبة الملك فهد الوطنية أثناء النشر

العصيمي، صالح الفهد

المدونات اللغوية العربية بناؤها وطرائق الإفادة منها. /صالح فهد العصيمي .-
الرياض، ١٤٣٦ هـ

٣٠٠ ص؛ ١٧ × ٢٤ سم

ردمك: ٤ - ٨ - ٩٠٦٦٤ - ٦٠٣ - ٩٧٨

١- اللغات - تاريخ ٢- اللغة العربية - بحوث أ. العنوان
ديوي ٤١٠,٧٢ ١٤٣٦/٥٣٣٨

رقم الإيداع: ١٤٣٦/٥٣٣٨

ردمك: ٤ - ٨ - ٩٠٦٦٤ - ٦٠٣ - ٩٧٨

حقوق الطبع والنشر محفوظة

الطبعة الأولى

١٤٣٦هـ / ٢٠١٥م

سلسلة من الإصدارات التي تعالج قضايا لغوية متنوعة

مدير المشروع:

أ. خالد بن أحمد الرفاعي

إشراف:

د. عبدالله بن صالح الوشمي

المبحث الخامس

البحث اللغوي في المدونات العربية الحاسوبية
بين الممكن والمحتمل والمأمول

سلطان بن ناصر المجيول

قسم اللغة العربية - جامعة الملك سعود

تمهيد

ليس ثمة مناص من التداخل الاختصاصي interdisciplinary بين الحقول اللسانية والعلوم الأخرى في الدوائر المعرفية والعلمية بحثًا وتحليلًا للسانيين والحاسوبيين والمحللين للخطاب اللغوي. ومن أوجه هذا التداخل الاختصاصي: ولوج تقنيات الحاسب وبرمجياته وأدواته التحليلية والمعالجاتية في جمع النصوص في مدونات محوسبة، وتصنيفها وفق أوعيتها المعلوماتية الناقلة، وفحصها تكرارًا وجمعًا وفرزًا وتحليلًا من مكانزها الرقمية. ومن المهم البدء بأهمية البحث اللغوي الحاسوبي وقضاياه الأولية، والتي لا يمكن حصر كل مناهجه، غير أن ثمة مفاتيح أساسية للباحث اللغوي تُمكنه من البدء في الدخول إلى هذا المجال، ويوظف عن طريقها خبرته اللغوية في أثناء معالجة قضية لغوية في المدونات الحاسوبية الأحادية (انظر على سبيل المثال، ميكنري وهاردي McEnery and Hardie ٢٠١٢؛ وبيكر Baker ٢٠١٠) أو التقابلية (جرانجر Granger ٢٠٠٨؛ وجوهانسون Johansson ٢٠٠٨) أو الترجمية (جرانجر وآخرون Granger et al ٢٠٠٨). ومن الممكن أن يقال إن ما سيقدم هنا هو ما يستحثه واقع البحث اللغوي العربي الحاسوبي، وبمحاولة جديدة في الدراسات اللغويات الحاسوبية للَملمة مناهج البحث اللغوي الحاسوبي وعرضها بصورة موجزة لا تقتصر على التعريف والعرض فحسب، بل على فتح أبواب البحث اللغوي الحاسوبي التطبيقي بصورة لعلها توضح معالمها للمختصين في اللغويات والدراسات اللغوية. وتُعرّف هذه الورقة بطرائق البحث اللغوي العربي الحاسوبي بصورة أولية تَرسم تمرحلات التحليل بواسطة المعالجة الحاسوبية بشكل مبسط؛ بدءًا بالتعريف بأنواع المدونات الحاسوبية للغة الطبيعية natural language، واتجاهاتها الممكنة والمحتملة والمأمولة، كل ذلك من أجل فحص المعجم العربي الذهني التوليدي الواقعي الذي يتمثل في المدونات الحاسوبية والبحث اللغوي من حيث

اتساع هذا المعجم أو تحجيمه باتساع المدونة أو تحجيمها (علاقة البحث اللغوي بتمثيل مدونة ما the representativeness of a particular corpus للمستويات اللغوية العربية النموذجية)، مرورًا بخصائص كل طريقة ومناهجها البحثية من حيث الغرض والتحليل والمعالجة، وصولًا إلى اقتراح بعضٍ من الموضوعات البحثية المهمة في هذا المجال.

البحث اللغوي والمدونات العربية الحاسوبية

تنال الدراسات اللغوية الحاسوبية الغربية -وإرهاصات هذا الدرس في سياق اللغة العربية والحاسب والمعالجة الآلية والبحث اللغوي القائم عليها- حظاً من التداخلات الاختصاصية بالقدر الذي يحتاجه التطوير في طرائق البحث اللغوي العربي الحاسوبي، وهذه التقاطعات هي على النحو التالي: أولًا: هندسة اللغة language engineering: (استعمال أدوات التحليل الحاسوبي لمعالجة اللغة وهندستها من حيث عناصر بنائها وعلاقة هذه العناصر في علم الوجود ontology). ثانيًا: اللسانيات الحاسوبية computational linguistics ومجاله الاختصاصي المعروف بمعالجة اللغة الطبيعية natural language processing NLP: (طرائق تأسيس وبناء خوارزميات وبرامج حاسوبية تُمكن اللغة وتسعى بأدبيات متعددة ومعقدة ومتنوعة إلى محاولات تعريف الخصائص اللغوية النوعية مع الخصائص الحاسوبية الكمية والرقمية). ثالثًا: لسانيات المدونات corpus linguistics: بناء مدونات حاسوبية، وإجراء بحوث لسانية حاسوبية إحصائية تفيد مجال تقويم مصادر اللغة Language Resources LRs، بالإضافة إلى التوسيم الصرفي والنحوي والدلالي والإحالي وتحشية النصوص المكتوبة أو المنطوقة في قوائم حاسوبية تحليلية، وإعدادها لمرحلتي التدريب training والاختبار test: أي مرحلة الإعداد والتحشية والتوسيم (التدريب)، ومرحلة اختبار عمليات التوسيم على القوائم من أجل تحقيق نسبة مئوية تقل

بازدياد عدد الكلمات في أي مدونة وتزيد بقلة عدد الكلمات في أي مدونة. رابعًا: تحليل النصوص الإلكترونية electronic text analysis: (جمع النصوص، والتنقيح، والمعالجة، والإخراج في قوائم حاسوبية من أجل التحليل لغرض معين، أو تضمينها في مواقع الويب، أو إدخالها في أقراص ممغنطة، إلخ). وسيُعرّف هذا المبحث الاختصاصيين بالدرس اللساني الحديث على حقل لسانيات المدونات الحاسوبية corpus linguistics أو علم المتون دون غيرها (انظر العصيمي ٢٠١٣ في مناقشة المقابلات العربية لمصطلح corpus linguistics).

وثمة أسئلة عديدة سيُجاب عنها من خلال هذه الورقة في الجملة، ومن أهمها: هل من الممكن أن تُجيبنا المدونات عن كل أسئلة البحث اللغوي؟ وهل للبحث اللغوي المعتمِد على المدونات شروط منهجية؟ وأي مدونة؟ ولماذا؟ إن من أبرز المسارات البحثية التطبيقية للدرس اللساني الحاسوبي والتحليلي في المدونات الحاسوبية اللغوية التنوّع الوصفي synchronic variation، والتنوّع الزمني diachronic variation، ومباحث التغير اللغوي الوصفي والزمني ضمن المجالات اللسانية التركيبية والدلالية، والتغير اللغوي الوصفي في العربية المعاصرة، وذاك الزمني في التاريخ التحليلي historiography للغة العربية.

وسيُتطرق في هذا المبحث إلى عدة موضوعات؛ أولها: الفرق بين الدرس اللساني المعتمد على المدونة corpus-based وذلك الموجه بالمدونة corpus-driven. وثانيها: أنواع المدونات الحاسوبية من حيث عدد اللغة فيها إلى أحادية اللغة monolingual وتعددية اللغة multilingual، وتنقسم الأخيرة إلى تقابلية comparable (انظر: شاروف Sharoff ٢٠٠٤؛ والمجيول al-Mujaiwel ٢٠١٢)، ومتوازية parallel[٤٥] (انظر العجمي al-Ajmi ٢٠٠٣). وثالثها: الفرق بين الأدوات التحليلية للغة العربية ودورها في قراءة الرموز العربية المدعومة بالترميز الخاص بإظهار نظامها الخطي، وما يتوافر فيها من أدوات بحث ومعالجة. ورابعها: علاقة بناء المدونة اللغوية العربية الحاسوبية (أو أدوات معالجة اللغة

العربية) بطرائق الإفادة منها في البحث اللغوي. وخامسها: التصاحب اللغوي الفيرثي (نسبة إلى العالم فيرث، واضع نظرية التصاحب collocation theory في عام ١٩٥٧؛ انظر فيرث Firth ١٩٥٧؛ وروحاني Rouhani ١٩٩٤) وتحليل المتتابعات اللفظية المباشرة، والتقارب الدلالي، والعلاقة بين المواد المعجمية والقواعد اللغوية، والتفريق بين المدونة بوصفها نظرية والمدونة بوصفها منهجًا تطبيقيًا، وتطوير سنكلير Sinclair (١٩٩١، ٢٠٠٤) لهذه المفاهيم بما يتوافر في المعالجات الآلية، وترسيخ جريس Gries (٢٠٠٣، ٢٠٠٨، ٢٠٠٩، ٢٠١٠) لمفاهيم التصاحب نوعياً وكمياً من جهة التلازم القريب والبعيد، ومن جهة التلازم النحوي colligation، ومن جهة التلازم المعجمي النحوي collostruction والمجموعات الدلالية semantic sets؛ كل هذا عن طريق أهم الرزم الإحصائية المعمول بها في علم لغة المدونات (قارنها بـ لسانيات المدونات الإحصائية بـ «آر» statistical corpus linguistics with R عند جريس Gries ٢٠٠٩، ٢٠١٠). أما آخرها فيتعلق بالممكن والمحتمل والمأمول في الموضوعات الخمسة السابق ذكرها، مع إلحاق هذا الموضوع بجملة من الموضوعات البحثية المقترحة.

٥. ١. البحث اللغوي المعتمد على المدونة corpus-based والموجه بالمدونة corpus-driven

ثمة منهجان أساسيان في البحث اللغوي ومادته النصوص الرقمية (توجنيني بونيللي Tognini-Bonelli 2000)، فالأول -كما في العنوان الفرعي- ينطلق من قضية لغوية محددة بافتراضات وفرضيات وأسئلة بحثية في ذهن الباحث اللغوي، ويقوم بتفسيرها وتحليلها بالاعتماد على النصوص الرقمية في المدونة [٤٦]. أما المنهج الآخر فهو النظر إلى المدونة وتدوين ما يُمكن ملاحظته من مسائل لغوية صرفية وتركيبة ودلالية وسياقية، والتوجه من المدونة إلى صياغة فرضيات البحث وأسئلته.

٥ . ٢ . أنواع المدونات العربية الحاسوبية

تتنوع المدونات اللغوية الحاسوبية بتنوع النوع، والغرض، والعدد، والتصميم[٤٧]، والذي يهم في هذا السياق تلك الأنواع المتعلقة بالعدد، على الرغم من أهمية بقية الأنواع. وتُعزى أهمية العدد إلى كونها الأكثر شمولية لمستويات اللغة؛ بخلاف النوع والغرض اللذين تنحسر أبحاثهما على مستويات لغوية محددة، وأجناس كتابية معينة، وبخلاف التصميم الذي ينحصر اهتمامه على مجال تطوير تصميم المدونة ومعالجتها من حيث أنواع التوسيم والتحشية. ولا يعني هذا التصنيف عدم وجود مدونة حاسوبية عربية تتنوع في النوع والغرض والتصميم، فالمدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية King Abdulaziz City for Sciences and Technology's Arabic Corpus (KACSTAC) قد تطورت في نسختها الجديدة (مختلفة عن النسخة القديمة www.ksacstac.org؛ انظر الثبيتي al-Thubaity ٢٠١٤) من حيث توفّر خصائص بحثية دقيقة للكشافات السياقية والتكرارات والمتتابعات، والتي ستُمكّن من تحديد أسئلة وفرضيات لغوية أكثر دقة إن أحسن الباحث استعمالها. وعلى صعيد أنواع المدونات العربية من حيث العدد، فالنوع الأول يتضمن نصوص اللغة العربية المكتوبة والمنطوقة أو أحدهما، والـمُدخلة حاسوبيًا مع توفر محرك بحث search engine شبكي فيها من أجل البحث عن كلمة محددة في النصوص المدخلة فيها حاسوبيا، وإمكانية إظهار عدد مرات تكرارها، وإتاحة إخراج كشافاتها السياقية، كله من أجل كشف السلوك البيئي السياقي اللغوي الطبيعي لتلك الكلمة وما يرد قبلها وبعدها من وحدات معجمية نظمية syntagmatic lexical units تُعرف بمصطلح (المتصاحبات اللفظية collocations). والنوع الثاني يتضمن مدونتين منفصلتين؛ الأولى: عربية، والثانية: لغة غير العربية، ويُعمل على جمعها معًا وفق عدة معايير هي على النحو الآتي: تقابل الأوعية genres والمجالات domains من حيث النوع والحجم، وتوافق أزمنة إنتاج النصوص الحية في كلتا

اللغتين، وتقارب خيارات محرك البحث فيها (انظر ميكنري McEnery 2003: ٤٥٠؛ ميكنري وشياو McEnery and Xiao 2007: ٢٠). أما النوع الثالث من حيث العدد، فيُعرف بالمدونة المتوازية parallel corpus، وهذا النوع من المدونات ما زال في بداية انطلاقته بين العربية والإنجليزية (انظر العجمي al-Ajmi ٢٠٠٣؛ وانظر حول بداية معالجة المدونة المتوازية في: بياو Piao ٢٠٠٠، ٢٠٠٢، وبيكر وآخرين Baker et al ٢٠٠٦: ٩ فيما يتعلق بطريقة معالجة المحاذاة آليا). والفرق بين المدونة اللغوية الحاسوبية المتقابلة وتلك الموازية تكمن في أن نصوص المدونة الأولى من اللغتين لا تكون نتيجة لعمل ترجمي مسبّقا، بينما في نصوص الثانية يكون توازي النصوص مشروطًا بناتج الترجمة الفعلية الواقعية، وتُدخَل هذه النصوص آليًا في قوائم تمثل كل مُدخل تركيبي أو معجمي من اللغة المترجم منها إلى ذلك المدخل التركيبي والمعجمي في اللغة المترجم إليها، والذي يُعرف بالمحاذاة الآلية automatic alignment بين نصوص اللغة المصدر وترجماتها في اللغة الهدف[٤٨].

وبالنسبة للمدونات العربية الحاسوبية، يرى الباحث أن أكبر المدونات العربية الأحادية -كما أُشير إلى ذلك سابقًا- تتمثل في المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية KACSTAC، ومدونات العربية الشبكية Arabic Internet Corpora، ومدونة أرابيكوربوس arabiCorpus[٤٩]. ويتوفر في هذه المدونات محرك استعلام query engine شبكي يقوم من خلاله الباحث بكتابة كلمة ما والبحث عن استعمالاتها في النصوص العربية المحوسبة. وثمة في المقابل أدوات أخرى لمعالجة النصوص العربية الرقمية، ومختلفة عن تلك المتضمنة في المدونات الثلاث المذكورة آنفًا، وهي أدوات برمجية مستقلة عن المدعومة في مواقع تلك المدونات الثلاث، حيث تقوم بمعالجة النصوص العربية المحفوظة في امتدادات عديدة على غرار text و csv، و xml، و rtf وغيرها. ومن أهمها قبولا للتطبيق التحليلي في أكثر الأدوات ذكاءً حتى الآن (غواص

ACL أولًا ثم سكتش إنجن Sketch Engine). ويُرى النظر في أدوات معالجة العربية من أجل المقارنة بينها من حيث التصميم، والإتاحية، وقابلية القراءة للنصوص العربية، ونظام الاستعلام، ومؤشرات التصاحب، والنتائج الإحصائية ومقارناتها، وهي محور موضوع المبحث الآتي.

٥. ٣. أدوات معالجة النصوص العربية

تنقسم أدوات معالجة نصوص العربية إلى قسمين، أولهما: برامج تطبيقية حاسوبية حيث توفر للباحثين إمكانية تحميل نصوص عربية على امتداد ملف text أو امتداد ما يسمى بالقيم المفصولة بفواصل comma separated value csv، والتي تحول آليا في برنامج إكسل، وتوفر أيضا إمكانية البحث عن الكلمة العربية، شريطة أن تكون هذه الأدوات المخصصة لمعالجة اللغة العربية مبرمجة مسبقا من أجل دعم فك شفرات أشكال الكتابة العربية الإملائية بالترميز العربي الحاسوبي إلى أشكاله الكتابية الإملائية، أي: تكون مدعومة بأحد الترميزات الآتية: UTF-8، أو UTF-16 (الثبيتي وآخرون al-Thubaity et al 2013). ومن أنواع هذه الأدوات الداعمة لقراءة النصوص العربية: مونوكونك MonoConc[٥٠] والفهرسة الاستدراكية وأسلوب بناء الاسترجاع بلغة التوصيف الموسعة XML Aware Indexing and Retrieval Architecture Xaira[٥١] والكلمة الدالة في السياق [٥٢] KWIC (keyword in context) وأكونكورد aConCorde[٥٣] وأدوات وورد سميث WordSmith Tools[٥٤] وغواص Ghaww□□[٥٥]. ويتطلب من الباحث اللغوي عند استعمال هذه الأدوات برنامج الجافا Java من أجل تشغيل هذه الأدوات. أما ثانيهما: فهي أداة متوافرة على الشبكة العنكبوتية، وتُعرف باسم Sketch Engine[٥٦]، وتعد أداة معالجة للنصوص اللغوية (ومنها العربية) (تحدث الثبيتي عن بعض هذه الأدوات في المبحث الثالث من هذا الكتاب).

ويُتيح هذان النوعان من الأدوات نظام الاستعلام النصي عن الكلمات والعبارات وتكراراتها Frequencies في النصوص العربية واستخراج كشافاتها السياقية concordance lines. وكل أداة من هذه الأدوات تختلف عن الأخرى من حيث ما تتيحه من أنظمة استعلامية query systems، أي أن كل أداة من هذه الأدوات تتضمن أنظمة تُمكن المستخدم من تحديد النتائج المطلوب إظهارها ضمن قوائم على شاشة الحاسب، وتتوافق هذه الأدوات في بعضها، وتختلف في بعضها الآخر، وتتميز بعضها بأنظمة جديدة ليست متوفرة في الأخريات (انظر الجدول ٣٦).

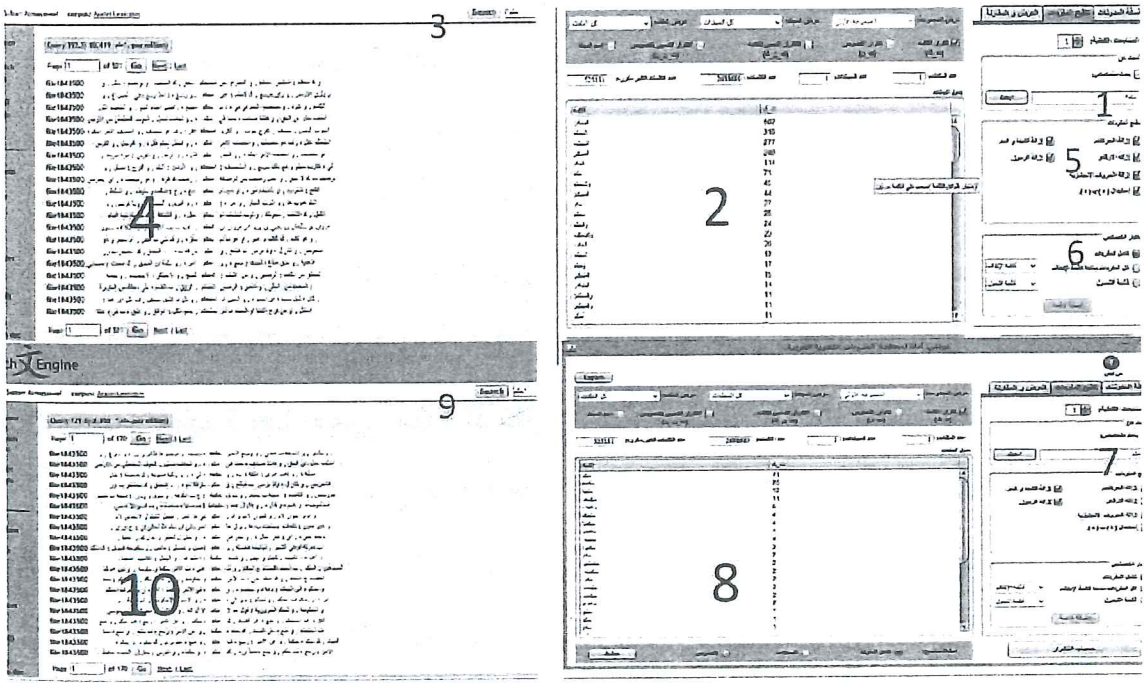
الأدوات	Mono Conc	Xiara	KWIC	aConCorde	WordSmith Tools	غواص	Sketch Engine
المصمّم	مايكل بارلو Michael Barlow	BNC	ساتورا تسوكاموتو Satoru Tsukamoto	أندرو روبرتس Andrew Roberts	مايك سكوت Mike Scott	الثبيتي وآخرون	Lexical Computing Ltd.
الإتاحية	متاح بالشراء	متاح في موقع BNC	متاح بالمجان	متاح بالمجان	متاح بالاشتراك	متاح بالمجان	متاح لتجريب مليون كلمة، ثم بالاشتراك
قابلية القراءة الآلية Machine-Readability	ترتيب الكلمات في الجملة ليس دقيقا في نتائج البحث عن كلمة	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق	الترتيب دقيق
نظام الاستعلام Query system	متوفر بدون خيارات متقدمة	دقيق ويحتاج لسرعة عالية ويتطلب الكثير من الوقت	متوفر بدون خيارات متقدمة	متوفر ويتيح البحث عن الكلمة المفتاح، أو العبارة، أو باستعمال wildcard بالرمزين (*) أو (؟)	متوفر بدون خيارات متقدمة	متوفر مع خيارات متقدمة: استعمال wildcard وإمكانية تنقيح النص من الهمزة والتاء المربوطة والأرقام والرمز الأجنبية وغيرها	متوفر مع خيارات متقدمة باستعمال wildcard بالرمز (*) ونظام استعمال phrase و Corpus Query Language

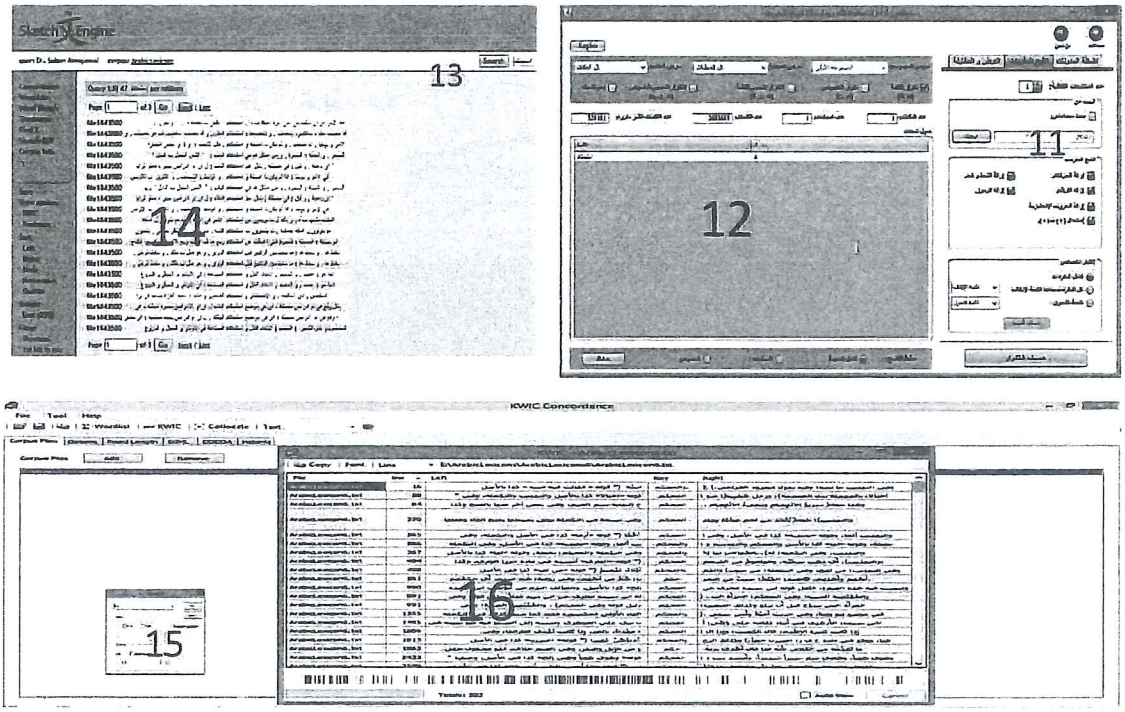
مؤشرات التصاحب Collocation indicators	متوفرة بدون دقة	متوفرة	غير متوفرة	متوفرة بدون دقة	متوفرة بدون دقة	متوفرة ودقيقة مع تحديد مواقع امتداد التصاحب إلى مدى ١٥ كلمة	متوفرة، لكن مع عدم إمكانية تحديد مواقع امتداد التصاحب
الرُزمة الإحصائية الخاصة بالمدونة Corpus-Based Statistics Package	غير مدعومة	مدعومة بـ Mutual Infor- mation و z-score	غير مدعومة	غير مدعومة	مدعومة بـ: Chi-Square Likelihood logDice	مدعومة بالرزم الآتية Chi-Square Weird Coefficients Mutual Information Likelihood t-score z-score LogDice	مدعومة بالرزم الآتية t-score Mutual Information logDice

الجدول (٣٦) أدوات معالجة نصوص اللغة العربية وخصائصها

بالنظر إلى هذا الجدول، نلحظ أن الإتاحية وقابلية القراءة تبدو واضحة، بخلاف نظام الاستعلام، ومؤشرات التصاحب، والرزمة الإحصائية الخاصة بالمدونة. فنظام الاستعلام يراد به ما يمكن توافره من خيارات بحث تتيح للمستخدم إمكانات تتباين من حيث العموم (التطابق الكلي) والدقة (التطابق الكلي والجزئي ونوع الكلمة)؛ فالأدوات MonoConc، و KWIC، و WordSmith Tools لا تتضمن إمكانية البحث بطريقة المطابقة الجزئية باستعمال خاصية wildcard، أي: باستعمال الرمز (*) أو الرمز (؟). ونجد هذه الخاصية في الأداتين غواص و Sketch Engine، وتمكّن هذه الخاصية البحث عن التطابق الجزئي، فعلى سبيل المثال: لو أدخلنا (حكم*) فإن نتائج البحث ستظهر جميع المتطابقات للكمة التي تبتدئ بالجذر (حكم) وجميع وحداتها الصرفيمية السابقة المقيدة بها، ومن أمثلة النتائج: أحكم/استحكم/ المحكم/أراحكم/إلخ. (الكلمة: أراحكم ليست من جذر /ح/، /ك/، /م/ إلا أن طريقة المعالجة بالرمز × تظهر أي تطابق جزئي شكلي). أما لو أدخلنا (*حكم) فإن نتائج البحث ستظهر جميع المتطابقات الجزئية للجذر نفسه مع جميع وحداتها الصرفيمية

اللاحقة المقيدة بها، ومن أمثلة النتائج: حكمة/ حكمك/ حكمنا/ حكمهم/ إلخ. (انظر إلى الأرقام ١، و٢، و٧، و٨ في واجهة غواص، والأرقام ٣، و٤، و٩، ١٠ في واجهة Sketch Engine في الشكل ١٤) ويُرى هنا أن خاصية المحرف البديل wildcard character لا تضمن استخراج المتطابقات للوحدات الصرفية المقيدة الداخلة، فكلمة (استحكام) مثلا تحتاج إلى أن يُبحث عنها لوحدها، وكذا الحال لكل كلمة اشتقاقية طابعها النسقي الصرفيمي غير سلسلي non-concatenative (انظر إلى الأرقام ١١، و١٢، و١٣، و١٤ في الشكل ١٤). وما يلفت الانتباه في هذا السياق هو أن أداة KWIC Concordance برغم عدم توفر هذا الخاصية فيها، إلا أنه بمجرد كتابة الجذر (حكم) في خيار KWIC، فإن جميع النتائج تتضمن متطابقات هذه الكلمة الكلية والجزئية ذات النسق الصرفي السلسلي فقط (انظر إلى الرقمين ١٥ و١٦ الشكل ١). أما الرمز الثاني (؟) المتعلق بخاصية wildcard فإن وضعه قبل الكلمة المراد البحث عنها أو بعدها يعين على استخراج تطابقات هذه الكلمة الجزئية بإظهار حرف واحد فقط، على سبيل المثال: (حكم؟) = حكمة/حكمت/حكمك/ إلخ. و(؟حكم) = أحكم/يحكم/نحكم/ إلخ.

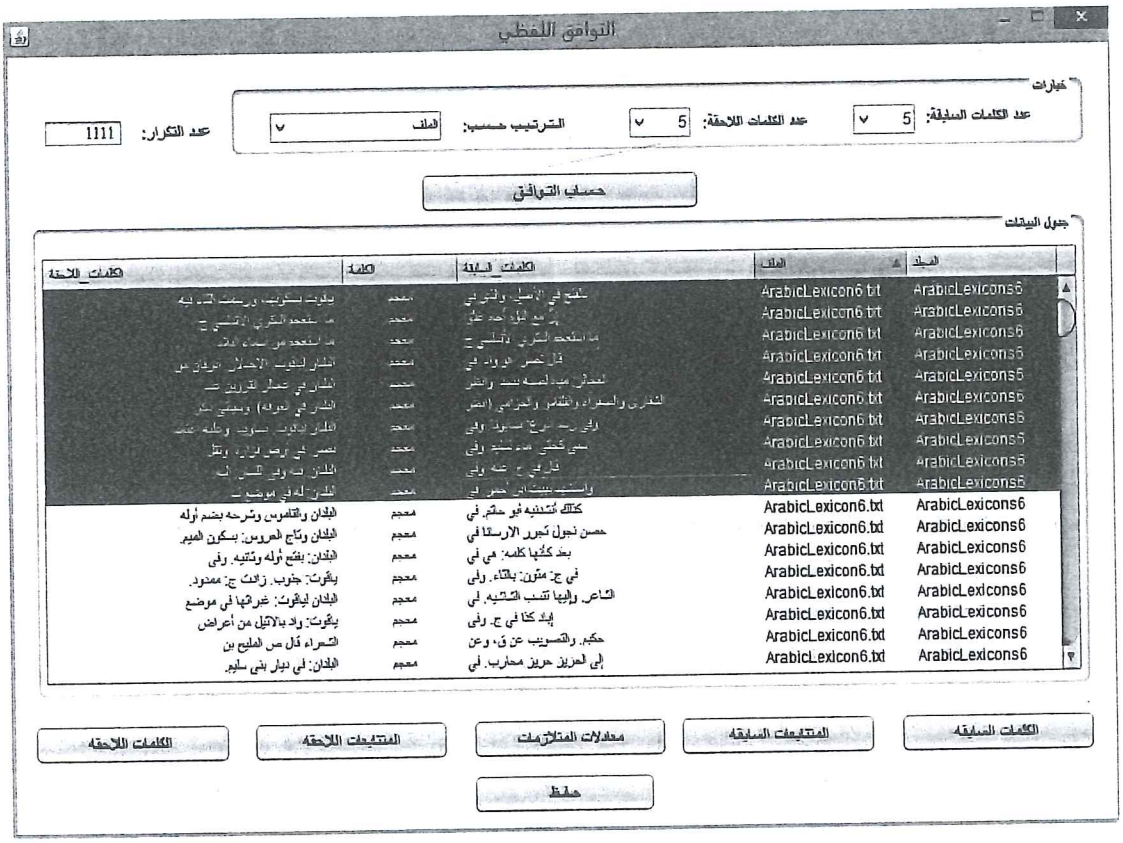




١. محرك البحث بـ (حكم*)
٢. نتائج البحث لـ (حكم*)
٣. محرك البحث بـ (حكم*)
٤. نتائج البحث لـ (حكم*)
٥. خواص تنقيح المفردات في غواص
٦. خواص إضافية (قوائم الإيقاف(المستثناة) والشمول)
٧. محرك البحث بـ (حكم*)
٨. نتائج البحث لـ (حكم*)
٩. محرك البحث بـ (حكم*)
١٠. نتائج البحث لـ (حكم×)
١١.نتائج البحث لـ (استحكام)
١٢.نتائج البحث لـ (استحكام)
١٣. نتائج البحث لـ (استحكام)
١٤. نتائج البحث لـ (استحكام)
١٥. نافذة جديدة للبحث
١٦.نتائج البحث للجذر (حكم)

الشكل ١٤: واجهة غواص وSketch Engine وKWIC Concordance ونتائج بحث كلمة (حكم)

أما مؤشرات التصاحب collocational indicators فهي عملية معقدة، ويرمز لها بـ N-gram، وتشير إلى المتتابعات اللفظية السابقة أو اللاحقة، والتي تمتد من ١ إلى ١٥، وتنطلق من المادة العنقودية Nodal item التي يُبحث عنها في الأصل بوساطة أدوات التحليل المذكورة آنفا (انظر: ٢. ٥.)، وتظهر المتصاحبات (أو المتتابعات) اللفظية للمادة العنقودية على شكل كشافات سياقية concordance lines (الشكل ١٥ وحساب التوافق لكلمة معجم في مدونة فرعية للمعاجم العربية ArabicLexicon6).



الشكل (١٥) الكشافات السياقية للكلمة الأساس (معجم) بامتداد ٥ متتابعات لفظية سابقة ولاحقة

ويُرى أن حساب توافق الكلمة النوعية type والكلمة الفعلية token تختلف بين غواص وSketch Engine، فالأول يختلف عن الثاني في الحساب، ولزامًا على الباحث أن يُدرك سبب الاختلاف، والعائد إلى دقة غواص في تحليل مفهومي الكلمة النوعية والكلمة الفعلية والمسافات بينهما التي عادة ما تكون محشوة بعلامات الترقيم، وعليه لو وضعنا مدونة عربية في صيغة text، وقمنا بتحميلها في أداة غواص والأداة الشبكية Sketch Engine، فإن حساب التطابق في الثاني سيكون أعلى نظرًا لاعتباره المسافات المملوءة بعلامات الترقيم وغيرها من الرموز. ولكن يتميز Sketch Engine بإمكانية معالجة النصوص العربية المتوفرة على المواقع الشبكية من خلال خاصية WebBootCat (انظر http://www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat).

والرزم الإحصائية الخاصة بالمدونة مجموعة من القياسات الإحصائية statistical measurements التي تقدم دلالات رقمية يكون منها القبول وعدم القبول في قياس قوة الارتباط بين الكلمات المتصاحبة (كلمة بصحبة كلمة) أو المتجاورة (كلمة بجوار كلمة يفصل بينهما كلمة أو كلمتين أو أكثر). ويُرى هنا تلخيص وظائف أهم هذه الإحصاءات المتعلقة بالتحليل الآلي للنصوص العربية، والتي تفيد الباحث اللغوي بالدرجة الأولى، وهي على النحو الآتي:

أولاً: ما يتعلق بالتحليل بين مجلدين (مدونتين)، ومن أهمها: مربع كاي Chi-Square (x2) حيث تفيد في قياس توزيع تكرارات النصوص أو الكلمات بين مجلدين بحساب التكرارات الملحوظة observed frequencies مع التكرارات المتوقعة expected frequencies (والأخيرة تُستخلص آلياً من الأولى) من أجل دعم الفرضية الصفرية null hypothesis. وأداة غواص تقوم بحساب مربع كاي بشكل آلي، ولا يُفضل استعماله لمعرفة قياس تكرار الكلمات في المجلد الواحد. والسؤال المهم هنا: كيف يُمكن للباحث اللغوي الذي سيحلل نصا رقميا باستعمال أداة غواص أو Sketch Engine أن يستفيد من قراءة الأرقام الإحصائية التي تقدمها هاتان الأداتان فيما يتعلق بهذا القياس؟ وكيف لهما أن تدعما أو ترفضا الفرضية الصفرية التي يصوغها الباحث؟ عند مقارنة مدونتين (مجموعتين) من حيث النظر، على سبيل المثال، إلى تكرارات كلمة ما، فإنه كلما كانت الفرضية الصفرية null hypothesis محددة فإنها لن تتوافق إلا بطبيعة المدونة المحللة، والمدخلة في أدوات التحليل. بمعنى آخر: لو قمنا بجمع نصوص المعاجم العربية كلها قديما وحديثا (٢٤ معجمًا)، واستخلصنا فرضية لغوية صفرية مفادها: أن اللواصق التصريفية inflectional suffixes أكثر من اللواصق الاشتقاقية derivational suffixes فيها بخلاف مدونة الصحف السعودية حيث اللواصق الاشتقاقية فيها أكثر من اللواصق التصريفية، فإن العينة هنا يجب أن تتوافق مع أركان الفرضية الصفرية، وهي أن الفروق بين

هذين النوعين من اللواصق واقعٌ طبيعيًا بين المدونتين بمحض الصدفة، وإن لم يكن وقوعهما بمحض الصدفة by chance، فإن ذلك يعني أن الفرضية لا يُمكن قبولها. ولكن كيف يُمكن قراءة الأرقام المتعلقة بهذا القياس؟ كلما اختلف حجم كلٍّ من المجموعتين (كل مجموعة تتضمن ملفًا لنصوص رقمية عربية)، اختلفت القياسات المتعلقة بمربع كاي؛ وعليه تختلف الفرضية الصفرية التي تُدعم بقيمة كل من التكرارات الملحوظة observed الواقعية والتكرارات المتوقعة expected، والتي تُستخلص بشكل آلي وفقاً لمعطيات المعادلة الآتية: (□ E-O)2/E وحساب هذه المعادلة ببساطة هو استخلاص نتيجة التكرار المتوقع expected frequency عن طريق جمع معطيات التكرارات الملحوظة (الواقعية فعليًا) من كل مجلد مستقل، واستخلاص نتائج جمعها عمودياً وأفقياً (الجدول ٣٧).

	تكرار اللواصق الاشتقاقية للفعل (ذهب) الملحوظة	تكرار اللواصق التصريفية للفعل (ذهب) الملحوظة	المجموع
مدونة المعاجم العربية	٦	١١	١٧
مدونة الصحف السعودية	١٢	١٠	٢٢
المجموع	١٨	٢١	٣٩

الجدول (٣٧) المرحلة الأولى لطريقة حساب التكرارات المتوقعة

في المرحلة الثانية، تُضرب كل قيمة لمجموع كل من التكرارات الملحوظة، ثم تقسم على مجموع نتائج التكرارات العمودية والأفقية: 18*17/(39)= 7.8 و18*22/(39)= 10.1، و21*17/(39)= 9.1، و21*22/(39)= 11.8 (الجدول ٣٨).

ثم تقسم نواتج هذه القيم باستعمال المعادلة الآتية: (O-E)2/E من أجل استخلاص أربع قيم (انظر الخانات المضللة في الجدول ٣٩)، ثم تجمع هذه

القيم لاستخلاص مربع كاي بشكل نهائي (مجموع حاصل (O-E)2/E من كل خانة: 1.51 =0.303+0.396+0.396+0.415 مربع كاي).

	تكرار اللواصق الاشتقاقية للفعل (ذهب) الملحوظة O	تكرار اللواصق التصريفية للفعل (ذهب) الملحوظة O	تكرار اللواصق الاشتقاقية للفعل (ذهب) المتوقعة E	تكرار اللواصق التصريفية للفعل (ذهب) المتوقعة E	المجموع
مدونة المعاجم العربية	٦	١١	٧,٨	٩,١	١٧
مدونة الصحف السعودية	١٢	١٠	١٠,١	١١,٩	٢٢
المجموع	١٨	٢١	١٨	٢١	٣٩

الجدول (٣٨) التكرارات الملحوظة observed frequencies الفعلية والتكرارات المتوقعة الآلية expected frequencies

	مجموع طرح قيمة الملحوظة على المتوقعة (O-E) للواصق الاشتقاقية	(O-E)2/E	مجموع طرح قيمة الملحوظة على المتوقعة (O-E) للواصق التصريفية	(O-E)2/E
مدونة المعاجم العربية	١,٨	٠,٤١٥	١,٩	٠,٣٩٦
مدونة الصحف السعودية	١,٩	٠,٣٩٦	١,٩	٠,٣٠٣

الجدول (٣٩) القيم المحسوبة calculated values للتكرارات الملحوظة والمتوقعة

هذه القيمة المحسوبة calculated value تعد أقل من القيمة الواقعية critical value؛ أي عدد الأسطر والأعمدة من المتغيرات في حساب مربع كاي في المربع في الجدول (٣٩). وهنا فإن كل عملية حسابية للمتغيرات (أو ما يُصطلح عليه

باسم درجة التكرار degree of frequency، والذي يمثل قيمة ٣ في المثال المذكور آنفا؛ أي: الخانة الأفقية الأولى الخاصة بالفرضية مع خانتي القيم الأفقية للمدونتين كما هو في الجدول ٣٩) تتضمن القيمة المحسوبة والقيمة الواقعية، وكل فرضية لغوية تُحسب بوساطة مربع كاي تكون نتائجها وفق معيارين؛ الأول: تُرفض الفرضية اللغوية الصفرية null linguistic hypothesis عندما تكون القيمة المحسوبة calculated value أعلى من ٠,٠٥ والثاني: تُقبل الفرضية اللغوية الصفرية عندما تكون القيمة المحسوبة أقل من ٠,٠٥، وعليه فإن القيمة النهائية لمربع كاي ١,٥١ تُعد أعلى من القيم ذات الدلالات الإحصائية: ٠,٠٥، أو ٠,٠١، أو ٠,٠٠١، وعليه إن الفرضية التي صُغناها (اللواصق الاشتقاقية والتصريفية في المدونتين مختلفة) غير صحيحة.

ثانيًا: فيما يتعلق بتحليل الكلمة المبحوث عنها ومتصاحباتها (الكلمة السابقة أو اللاحقة مباشرة) أو متجاوراتها (الكلمة السابقة أو اللاحقة غير المباشرة؛ أي: في الموضع الثاني أو الثالث أو الرابع أو الخامس). ومن أهم الإحصاءات التي تفيد دلالة التصاحب هنا هي: المعلومات المتبادلة Mutual Information (MI)، وقياس-ت T-Score، وقياس-ز Z-Score، ومعامل دايس Dice واللوج دايس logDice.

فالمعلومات المتبادلة قد وضحها تشرتش وهانكز Church and Hanks (١٩٩٠) وأواكس Oakes (١٩٩٨) على أنها تفيد بالكشف عن احتمالات تكرار وقوع كلمتين يكونان متصاحبين معًا مرة، ووقوعهما منفصلتين. ومعادلة هذه الإحصائية هي: $\log2((P(x,y)/P(x)P(y))$ حيث إن P هو عدد تكرار الكلمة، أما x و y فهما المتصاحبان اللذان يُراد اختبارهما. ولو قمنا على سبيل المثال باختبار تكرار كلمة (عاصفة أو عاصف) مع الصفتين (شديدة) و(قاشرة)، فإن المعطيات الإحصائية ستكون على النحو الآتي: عدد كلمات مدونة المعاجم العربية ٢٠,٤٣٢,٢١٢ كلمة، وعدد تكرار كلمة (عاصفة أو عاصف) فيها ١٢٠

مرة، وعدد تكرار (عاصفة قاشرة) ٤ مرات، وعدد تكرار (عاصفة شديدة) ٥ مرات، وعدد تكرار الصفة (قاشرة) لوحدها ١٢ مرة، وعدد تكرار الصفة (شديدة) لوحدها ١٥١٩ مرة. ولقياس كل متصاحب collocate من الصفتين (أي: شديدة وقاشرة) على حدة مع المادة العنقودية nodal item (أي: عاصفة أو عاصف)، فإن المعادلة تتمثل على النحو الآتي:

أولا: معادلة المعلومات المتبادلة للتصاحب «عاصفة أو عاصف شديدة» وناتجهما هي:

log2((5 * 20,432,212)/(120 * 1519))= 9.13

ثانيًا: معادلة المعلومات المتبادلة للتصاحب «عاصف أو عاصفة قاشرة» وناتجهما هي:

log2 ((4 * 20,432,212)/(120 * 12))= 15.7

بنتائج كل معادلة، نلحظ أن قوة التصاحب بين المادة العنقودية والصفة (قاشرة) أعلى من تلك الواقعة بينها والصفة (شديدة)؛ وقوة التصاحب هنا يُشير إلى أن التزام الصفة الأولى أكثر دلالة من حيث الاستعمال المنسجم من الصفة الثانية. وفيما يتعلق بدلالات المعلومات المتبادلة إحصائيا، فإن الناتج الذي يكون أعلى من ٣، يكون مقبولًا من حيث الدلالة الإحصائية، وإن كان أقل فإن دلالة التصاحب تكون حينها غير مقبولة.

وقياس-ت t-score يتشابه مع المعلومات المتبادلة غير أنه يقوم بإظهار مقاييس التشتت لاحتمالات تكرارات التطابق للمادة العنقودية nodal item ومصاحبها collocate بالاستناد إلى عدد الكلمات في المدى الواردتين فيه (سكوت Scott ٢٠١٠، وانظر هانستون Hunston ٢٠٠١، ٢٠٠٢؛ وبرايس Price ٢٠١٣). ويتكون هذا القياس من المعادلة الآتية: $((x/n)-x)/\sqrt{j}$ حيث إن n يعبر عن العدد الكلي للكلمات في المدونة، و j يعبر عن حاصل ضرب التكرار المشترك بين

المادة العنقودية ومصاحبتها؛ حيث إن x يشير ببساطة إلى F1*F2 (أي: ضرب عدد تكرار الكلمة الأولى مع عدد تكرار الكلمة الثانية المصاحبة للكلمة الأولى). ويُستفاد من هذه العملية في تحليل الكلمات ذات المعاني المتعددة polysems. والناتج الآلي الذي يُستخرج بهذه المعادلة يجب أن يكون من ٢ فما فوق من أجل ضمان قياس إحصائي ذي دلالة قوية.

أما قياس-ز z-score، فقد شرحه بيري-روجي Berry-Rogghe (١٩٧٣: ١٣١) بالمعادلة الآتية:

Z: العدد الكلي للكلمات في المدونة

A: المادة العنقودية وعدد مرات تكراراها

B: المصاحب للمادة العنقودية وعدد مرات تكرارها

K: عدد مرات تكرار A و B معًا

S: المدى وعدد الوحدات المعجمية قبل المادة العنقودية وبعدها

تقوم الأداة المدعومة بهذه الرزمة الإحصائية باستخراج قيمة z-score بين المادة العنقودية والمتصاحب أياً كان موقعه (مدى خمس كلمات5-word span)، وكل متصاحب لفظي مع المادة العنقودية تكون قيمته أعلى من ٣ يُعد دالًا إحصائيا بشكل مقبول.

أما آخر وأهم قياس إحصائي بين المادة العنقودية ومصاحبها فهو الدايس Dice (كيلجارف وآخرون2004 Kilgariff et al)، وأساس معادلته هو 2fAB/ (fA+fB) حيث إن fAB يدل على تكرار الوحدة المعجمية الهدف مع المتصاحب المعني بالتحليل، و fA+fB يدل على حاصل جمع تكرار المادة العنقودية لوحدها مع تكرار المتصاحب المعني بالتحليل. ومشكلة هذا القياس هو أنه يقدم قيمًا صغيرة جدًا (أقل من ٠,٠)، أما معامل اللوج دايس logDice فيضاف إلى

معادلة الدايس الآتي:logD +14 (2fAB/(fA+fB . فلو نظرنا إلى المثال الذي ذكرناه آنفا حول المادة العنقودية (عاصف أو عاصفة) مع الصفتين اللتين تُصاحبها في مدونة المعاجم العربية: (شديدة وقاشرة)، فإن الحسبة الآلية تكون على النحو الآتي:

تكرار الصفة المتصاحبة (شديدة) للكلمة الأساس (عاصف أو عاصفة) وفقًا لتكرارات كل واحدة على حدة مع تكرارهما معًا وبناء على الدايس ومعادلته (2fAB/(fA+fB تكون النتيجة: ٥/(١٢٠+١٥١٩) = ٠,٠٠٣٠٥ وقيمة اللوج دايس ١١,٥، أما تكرار الصفة المتصاحبة (قاشرة) للكلمة الأساس (عاصف أو عاصفة) وفقا لتكرارات كل واحدة على حدة مع تكرارهما معًا وبناء على معادلة الدايس (2fAB/(fA+fBتكون النتيجة 0.0303 =(120+12)/4. وتكون نتتيجة اللوج دايس ١٢,٥. وبناء على النتيجة الأولى والثانية نجد أن الثانية الأعلى من الأولى ذات قيمة عليا، وينظر للأعلى أو الأقل بحسب ما يريده الباحث من الفرضية من حيث قوة التصاحب من عدمه[٥٧].

٥. ٤. علاقة بناء المدونة اللغوية العربية الحاسوبية وأدوات التحليل بطرائق البحث اللغوي

على كل باحث في بادئ ذي بدء أن ينطلق من فرضياته اللغوية ثم يُوجد المدونة اللغوية العربية الحاسوبية أو الأدوات اللغوية أو بهما معا، والتي يُمكن لها أن تجيب عن تلك الأسئلة، أو أن ينظر إلى خصائص المدونة اللغوية العربية الحاسوبية من أجل أن ينطلق من تلك الخصائص التي تُكيف أصلا أسئلة البحث اللغوي. ومن كلا الجهتين ينطلق الحدس اللغوي العميق أولًا، ثم الآلة ثانيًا، ثم بهما معا شريطة أن يكون إعمالهما إزاء بعض منهجيًا ومقبولًا في المحصلات النهائية للتجربة، ويكونان معًا برهانًا للفرضية اللغوية الصفرية المصوغة، والمدونة المجددة للاختبار. وهنا تكون إجابات الأسئلة التي ذكرناها

في بداية المبحث؛ ومفادها: هل من الممكن أن تُجيبنا المدونات عن كل أسئلة البحث اللغوي؟ وهل للبحث اللغوي المعتمِد على المدونات شروط منهجية؟ ولماذا؟ فجواب (هل) هنا هو أن كل مدونة لغوية حاسوبية يستحيل أن تجيبنا عن كل الأسئلة؛ لأن الأسئلة هنا تتحدد بماهية المدونة بحسب نوعها أو غرضها أو عددها أو تصميمها (ميكنري وهاردي McEnery and Hardie 2012: 27).

أما الشروط المنهجية فتتحدد بطريقة يُوفّقها الباحث اللغوي مع العلماء المختصين بمعالجة اللغة الطبيعية Natural Language Processing. وأما السؤال (لماذا؟) فهو أساس البحث اللغوي العام أو الخاص، أي: أساس المنهج المتبع بتتبع نوع المدونة، وتتبع نتائج الرزم الإحصائية، وتوافق تلك النتائج بالفرضيات اللغوية.

٥. ٥. تحليل التصاحب اللغوي

يجب تبسيط مفهوم التصاحب أولًا، وهو مفهوم مركب يمكن تقديمه على النحو الآتي[٥٨]:

أولاً: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتصاحبة في النصوص، فالتصاحب collocation يُعد ارتباطًا تركيبيًا أوليا لوحدتين معجميتين معًا في سياق لغوي معين، مثل: يستقل/يركب/يستعمل القطار، وقد يخرج عن المتعارف عليه عند مزيد من الكشف عن مستويات النصوص اللغوية العربية الرقمية.

ثانياً: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتلازمة في النصوص، فالتلازم colligation يُعدّ إيقاعا تركيبيًا إلزاميا لوحدتين معجميتين معًا في أي سياق، مثل: الفعل اللازم والفعل المتعدي وحروف الجر وما بعدها من أسماء معرفة بـ (أل) وغيرها من المتلازمات.

ثالثاً: في عملية تحليل المعاني المركبة نصًا بمجموعة من المدخلات المعجمية المتجاورة في النصوص، فالتجاور collostruction يعد إيقاعا تركيبيا متغيّرا لوحدات معجمية نظمية توليدية في سياقات عديدة؛ مثل: المكملات أو المتممات complements في توليد بقية الجمل الإسمية والفعلية الأساسية من الأحوال والصفات والتمييز وأدوات الربط conjunctions التي تولد مزيدًا من التوليد، والتلازمات النحوية التي تزيد عن أكثر من ثلاث كلمات؛ مثل: الفعل المتعدي، ولا النافية للجنس، والأفعال الناسخة، وأخوات إن واسمها وخبرها، وظن وأخواتها، والالتزامات المتتابعة بمزيد من التجاور النظمي، كل ذلك في سياق تجاوري نظمي يُمكن من خلاله تحليل قواعد التركيب construction grammar (جولدبيرج Goldberg 2009).

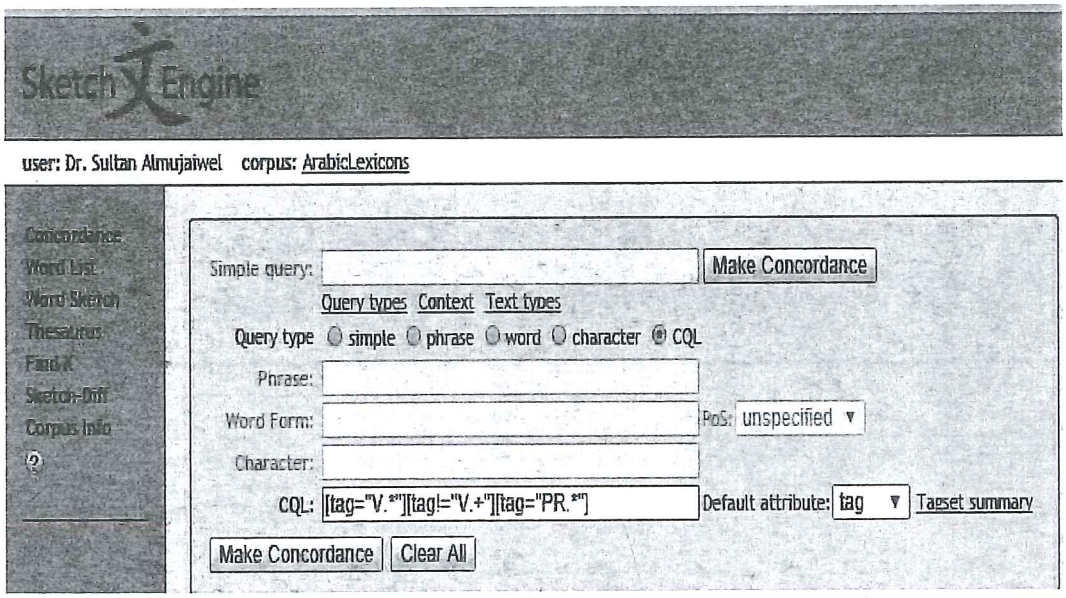
رابعًا: في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتباعدة في المدى span والمتقاربة في الدلالة (المجموعة الدلالية semantic set أو التقارب الدلالي semantic preference) يُعدُّ التضام هو المحك، وكثرة التعارف عليه تكشفه اللغة الطبيعية المحوسبة في المدونة الضخمة large-scale corpus؛ مثل: اصطادَ السمك من قارب أو زورق (لا يحدث في العادة اصطياد السمك من متن باخرة أو سفينة).

ودراسة هذه الأنواع قد تكون على مستوى الكلمات أو مستوى النصوص من جهة، وعلى مستوى التصاحب مع كلمة في نص وتوافق هذا التصاحب من عدمه في نص آخر من جهة أخرى. وأول من اهتم بنظرية التصاحب هو فيرث Firth (١٩٥٧)، وجاء من بعده سنكلير Sinclair (١٩٩١، ٢٠٠٤)، والفرق بينهما هو أن الأول قد اهتم بالمدى span بغض النظر عن الموضع position، واهتم الثاني بالموضع بغض النظر عن المدى. أما مفاهيم جريس Gries التي ذكرناها آنفًا فهي أكثر دقة من كلٍّ من فيرث وسنكلير.

وقوائم التصاحب اللفظي التي يمكن الحصول عليها بأدوات معالجة العربية قد لا تُجيبنا كثيرًا عن السؤال اللغوي العام: ما فائدة هذه القائمة من الناحية اللغوية؟ ولعل أعم إجابة يُمكن أن تبرهن عن فائدة واحدة لهذا القائمة هو أن بعض التكرارات المتعلقة بأدوات الربط conjunctions والحروف particles (حرف الجواب والنفي والشرط والتوكيد والتمني والصلة والترجي والنداء والأمر والجر والنهي والجزم والحروف المشبهة بليس والحروف المشبهة بالفعل إلخ)، وأسماء الإشارة والضمائر قد تعطينا ملمحًا عن نتائج هذه المدونة المحددة بحدود ماهيتها من حيث النوع والعدد والحجم، وهو ملمح قد لا يتجاوز فكرة طريقة التركيب والاستعمال التركيبي بين كل هذه الأدوات وما يرد بعدها وقبلها، وما تفيده في تطبيقات التحشية الإحالية anaphoric annotation من جهة، وأشكال التنوع variation الاستعمالي من وجهة نظر لغوية سوسيولوجية.

وثمة مناحٍ أخرى جديرة أوليًا بالاهتمام في مجال التحليل التصاحبي في أي مدونة، ومنها: قضية التصاحب collocation والتجاور collostruction. ولو أخذنا على سبيل المثال ما يوفره Sketch Engine في نظام لغة استعلام المدونة Corpus Query Language CQL (ياكوبيتشيك وآخرون Jakubíček et al، ٢٠١٠) والذي يُتيح البحث في المدونة عن طريق خواص الترميز [attribute=»value»]لنوع صيغة الكلمة أو بالاعتماد على رموز التوسيم للغة العربية (مجموعة وسوم ستانفورد ومنى دياب لأقسام الكلام Stanford's Mona Diab's Part-Of-Speech Tagging Sets: ومنى دياب يُعرف بـ AMIRA وهو مطور من توسيم MADA؛ انظر: دياب Diab ٢٠٠٧، ٢٠٠٩ وانظر حول مدى MADA عند: حبش وآخرين Habash et al 2009). ولو أراد الباحث على سبيل المثال استعمال هذه الخاصية من أجل الكشف عن البحث على استعلامات الضِمّن within and containing queries، أو على استعلامات الاتحاد meet

and union queries فستُتيح الخواص الترميزية لكل مشغل على حدة فرصة استخراج بحث عالٍ من الدقة من حيث التوافق في نظام استعلام CQL (انظر الشكل ١٦)، وهذا كله يُسمى بـ «مشغل الضمن أو الاتحاد withingand containing or meet and union operator». (الشكل ١٦).



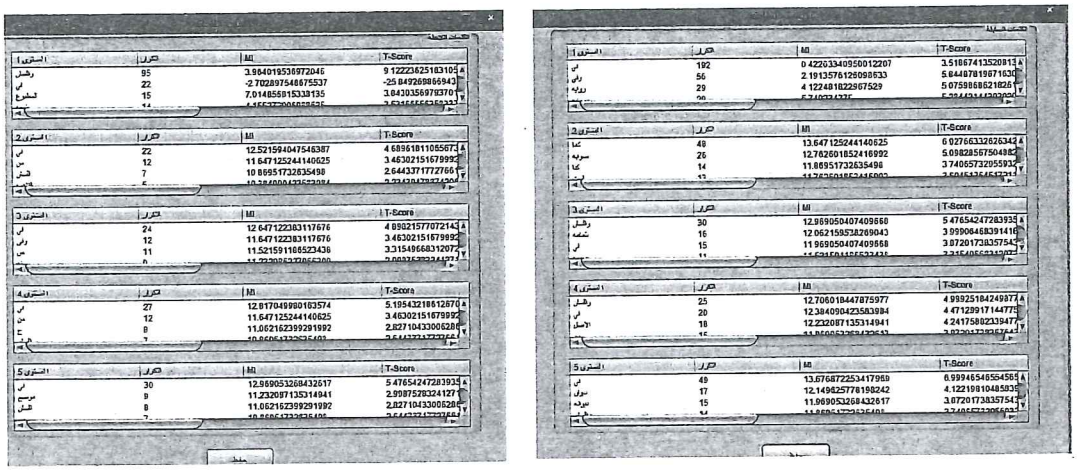
الشكل (١٦) برنامج المستخدم للنظام: نافذة حوار نظام الاستعلام في Sketch Engine

في الشكل (١٦) نلحظ أن المدونة المستعملة هي مدونة المعاجم العربية (٢٤ معجمًا عربيا قديما وحديثا). وبالنظر إلى هذا الشكل، نلحظ خاصية الاستعمال CQL التي تعتمد على أساس خواص الترميز لكل مشغل مذكور آنفًا. ومن الممكن للباحث اللغوي هنا أن يتصدى للحروف والأحوال وأدوات الربط والتعابير الاصطلاحية والجمل الفعلية، وذلك بنظام التوسيم الذي تقوم عليه Sketch Engine (ستانفورد STANFORD وأميرة AMIRA؛ انظر: دياب Diab 2009). وفيما يلي نشير إلى مثال تطبيقي لمشغل الضِمّن أو الاتحاد:

within	[tag=»PR.*»] within [tag=»V.*»] [tag!=»V.+»] [tag=»PR.*»] توسيم يُظهر الأفعال اللازمة في مدونة المعاجم العربية، ونتيجته: ٤٢٣,٤٦٦ من أصل ٢٠,٤٣٢,٢١٢ كلمة
الضِمْن	
Containing	
meet	[tag=»V.*»][tag!=»V.+»][tag=»PR.*»]
الاتحاد	توسيم يظهر الفعل العربي وامتداده في نصوص مدونة المعاجم العربية، :
Union	ونتيجته ٤٢٢,٠٥٠ من أصل ٢٠,٤٣٢,٢١٢ كلمة

وهذان الترميزان أنموذجان لطريقة البحث في Sketch Engine، وثمة العديد من الترميزات التي تتيح الكشف عن أنواع مواد الكلمة المعجمية lexical word-items (الجمل الفعلية والمتصاحبات الاسمية من الأعلام وأسماء المؤسسات والأفعال اللازمة والإضافة والتعابير الاصطلاحية والتضام الدلالي المنسجم؛ انظر إلى نظام توسيم أقسام الكلم للعربية الخاص بأميرة في: حبش ٢٠١٠: ١٧٥)، والتي يستطيع الباحث اللغوي أن يستعين بها من قبل المختصين بلسانيات المدونات.

وتُمكّن أداة غواص لتحليل مفاهيم التصاحب عن طريق المادة العنقودية nodal item وما يتبعها من متصاحبات مع إمكانية تقديم تفسيرات إحصائية لقوة التصاحب في مواضع positions أربع: الكلمة الأولى، والكلمة الثانية، الكلمة الثالثة، والرابعة، وهي مزية تُحسب لأداة غواص، وتتميز بها، ولا توفرها أداة Sketch Engine (الشكل ١٧).



الشكل (١٧) نتائج حساب التطابق من حساب موضع الكلمات السابقة واللاحقة إحصائيا بـ MI وt-score وz-score و logDice

وبالنظر إلى هذا الشكل، سنلحظ أن التتابع السابق يعني: المتتابعات السابقة، والتتابع اللاحق يعني: المتتابعات اللاحقة، والمستوى يعني: الموضع position الذي يعكس موقع التتابع بوصفه المتتابع الأول (المستوى الأول)، أو المتتابع الثاني (المستوى الثاني)، أو المتتابع الثالث (المستوى الثالث)، أو المتتابع الرابع (المستوى الرابع). وبهذه المزية المعالجاتية، فإن للباحث اللغوي القدرة على كشف مواضع التتابع المباشر وغير المباشر مع قراءة إحصاءاتها المتعلقة وفق ذلك الموضع.

الاتجاهات البحثية الممكنة

لا يمكن حصر كل الاتجاهات الممكنة، ولكن ثمة ما قد يفتح للباحث اللغوي آفاقا ما إن يُعمل بها فإنها ستدفع بما سيشغله في الدرس اللساني. وتعد الاتجاهات البحثية اللغوية والترجمية والمعجمية بوساطة الاستناد إلى مدونة لغوية عربية حاسوبية[٥٩] أولية من جهة الواقع الحيوي للوحدات الجذعية وما يلتصق بها من لواصق اشتقاقية وتصريفية مستجدة، والتراكيب اللغوية وما يتغاير فيها من حيث الأهمية لمواد الكلمات المعجمية lexical word-items، والتقارب الدلالي

semantic preference، والكشف عن بيئات هذه الوحدات المعجمية الاستعمالية من حيث الاشتقاقات والتصاحب والدلالات اللغوية وغير اللغوية (اجتماعية، أو نفسية، أو أكاديمية، إلخ).

الاتجاهات البحثية المحتملة

من الاتجاهات البحثية المحتملة: إعادة النظر من جديد في مفاهيم التصاحب اللغوي، وتلك التقسيمات المذكورة آنفا (انظر ٥. ٥.) التي تحتاج أبحاثًا على مستوى المؤسسات والأفراد، وسيفرض هذا الاتجاه رسم الأبعاد الاستعمالية للوحدات المعجمية في المعجم الذهني العربي، والوصول لقواعد بيانات تتضمن صورًا وأشكالًا لتصاحب الألفاظ عن طريق تصاحبها الحر والمقيد لفظيًا، وتلازمها نحويا، وتجاورها تركيبيا، وتضامها دلاليا.

إن قواعد بيانية حاسوبية لهذه الأحوال التصاحبية الأربعة ستفتح آفاقًا لمعرفة لغوية عربية حية تتولد من عقول مستعملي اللغة، وتعيش بعيدة عن البعثرة من جهة الحدس، وأي لغة لها تاريخ استعمالي تاريخي طويل (كحال العربية) يتحتم فيها أن تُمثَّل في تلك القواعد البيانية في ثلاثة اتجاهات؛ الأول: تعاقبي diachronic والثاني: آني (تزامني) synchronic، والثالث: التوقيف compromise (اللغة التوقيفية compromise language؛ أو اللغة الوسيطة) [٦٠] التي تتوسط بين التغيرين الزمني والوصفي من جهة والتقليدي واللهجي من جهة ثانية والعمل المؤسسي على تحديد المستويات النموذجية standard levels من جهة ثالثة (انظر: مول Moll ٢٠٠٨-٢٠١٠، وانظر الحمزاوي ١٩٨٦)؛ كل ذلك من أجل السعي إلى التمثيل البياني المقبول وصفًا واستعمالًا نموذجياً للغة العربية.

الاتجاهات البحثية المأمولة

من المأمول مستقبلاً توسيع دوائر معالجة اللغة العربية بأدوات حاسوبية يُمكّن لها أن توفر للباحث اللغوي العربي مزيدًا من التحليل، وذلك عن طريق تطوير معالجات التوسيم النحوي part-of-speech tagging (أو التوسيم القواعدي grammatical tagging) واستحداث بدايات جادة للتوسيم الدلالي semantic tagging والإحالي anaphora والنبري prosodic، والتي ستُفيد الباحث اللغوي بمزيد من وقائع التحليل اللغوي لوقائع اللغة العربية الحية.

ومن المأمول أيضًا أن يكون هناك اهتمام بالمدونات المتقابلة comparable corpora والمدونة المتوازية parallel corpus، حيث تفيد الأولى في الكشف عن المتقابلات السياقية اللغوية الاستعمالية بين مدونتين تتوافقان في المعطيات الزمنية والوعائية والمجالاتية والموضعية والسياقية. وهذا المرحلة كلما زادت جدّتها زاد عونها للمختصين في معالجة اللغات الطبيعية بفهم أعمق لآليات الترجمات الآلية؛ أما الثانية فستفيد في معرفة مزيد من المقابلات الترجمية السياقية الطبيعية، والتي سيكون لها دور في فهم أيّ المقابلات من اللغة الهدف أولى استعمالا من حيث الأكثر تكرارًا من جهة السياق.

موضوعات مقترحة في البحث اللغوي الحاسوبي

يعتمد البحث اللغوي على تحليل معلمَيْن أساسين من وجهة نظر المعجميات التركيبية structural lexicology (انظر ليبكا Lipka 2002)، وهما: المعلم الضملغوي (تحليل اللغة بالنظر إلى مكوناتها الداخلية بنيويا من الصوت إلى الدلالة) والمعلم الفولغوي (تحليل اللغة بالنظر إلى مكوناتها الخارجية وجوديا من النفس فالتربية فالتعليم فالثقافة فالمعرفة وصولا إلى التكوين الأم: علم الوجود).

وستُقترح هنا موضوعات في البحث اللغوي الحاسوبي تكون مادتها اللغة الحية (أي مدونة لغوية حاسوبية محفوظة في امتداد txt أو csv)، وأدوات تحليل هذه المادة اللغوية (وأفضلها: غواص وSketch Engine).

وأذكر هنا بعض الموضوعات التي يُمكن أن تُقترح، ليس من أجل بحثها بل من أجل تقريب الصورة الأولية لمنطلقات البحث اللغوي في المدونات العربية الحاسوبية وتحديد نوع المدونة. ومن ثم فالفرضية تكون صياغتها بداية تحديد البنية الصورية thematic structure لأية ورقة بحثية لغوية معتمدة على المدونات العربية وأدوات تحليلها ومعالجتها آليا.

نوع التحليل	المدونة المقترحة	الفرضية المقترحة
أنواع التراكيب النحوية من حيث التصاحب والتلازم	معاصرة (صحف/ سرديات وروايات)	التراكيب المتنوعة بتنوع المجالات
أنواع التراكيب من حيث التجاور والتضام	معاصرة (صحف/ سرديات وروايات)	التراكيب المتنوعة بتنوع الأوعية
أنواع التراكيب الاسمية من حيث أنواع مواد الكلمات المعجمية lexical word-items	مقارنة بين مدونة تراثية ومدونة معاصرة	مواد الكلمات المعجمية وأنواعها في العربية
تحليل الوجود للمواد المحسوسة	الكتب التراثية المتعلقة بالرحلات	مسميات الموجودات والمحسوسات في الكتب العربية التراثية
تحليل التعبيرات الخاصة بدقائق المواد الثقافية	الكتب التراثية	أسماء أجزاء كل محسوس أو موجود على حدة

التغاير بين التماثل المعجمي isolexical والتماثل النصي isotexutal	الكتب التراثية/ أو الصحف العربية القُطرية	نسبة التطابق والتغاير بين الوحدات المعجمية ومجالاتها الموضوعية
الروايات	الرواية العربية أو المحددة عربيا من حيث الجغرافيا السياسية	الاتجاهات الاستعمالية للوحدات المعجمية
الكلمة في السياقات الشعرية	الشعر العربي القديم والحديث	التنوع في الدلالات الشعرية
تحليل توافق الثقافة واختلافها للوحدات المعجمية	المدونات المتقابلة	بين لغتين فيهما أثر التشابه أو الاختلاف بين الدوال اللغوية ومدلولاتها
التطابقات لجميع الأفعال الزائدة عن الجذر root والجذع stem	إعمال tokenization و lemmatization لأجل التصدي لها	الاشتقاقات القديمة والجديدة
السير	السير الذاتية	الاختصاص الدلالي للوحدات المعجمية مكانيًا/ زمانيًا/ ثقافيا
تحليل الأخطاء من حيث الأعمار/ الخلفية الثقافية/ البيئة التعليمية إلخ.	الفيفي وأتويل al-Faifi and Atwell ٢٠١٤ (مدونة متعلمي العربية Arabic Learner Corpus)	معامل الارتباط بين نوع الأخطاء والخلفيات العمرية و/أو الثقافية و/أو العمرية لغير العرب

خاتمة

عُرِضَ آنفًا أهم الأدوات اللغوية التحليلية للنصوص العربية، ومفاهيم المدونة العربية الحاسوبية وأهم أنواعها، ومفهوم التصاحب اللفظي من وجهة نظر لسانية مدونيّة حاسوبية، وأهم الطرائق الإحصائية من حيث قياس القوى والمدى والتجاور والتضام، وأثر هذه القياسات الإحصائية في دعم الفرضيات اللغوية من عدمه. ولا غرو أن نلحظ أهمية هذه الاتجاهات في الدراسات اللغوية العالمية، واللغة العربية ليست بمنأى عن المستجدات البحثية اللغوية لخدمة لغة الضاد. ولقد وقف الباحث على ما يزيد عن ٤٠٠ دراسة أجنبية، وقد كانت أدوات وورد سميث WordSmith Tools منطلقَها الأساس في تحليل فرضياتها المتعددة. إن دخول المهتمين في اللغة والأدب إلى هذا المضمار سيبني أُطرًا جديدة لمزيد من التفسير والتحليل الآلي والإحصائي والنوعي الحذر، وسيضفي آفاقًا رحبة تعين على مزيد من استيعاب واقع اللغة العربية بوصفها لغةً حيّة مُنتجَة بالصور القياسية لمستويات العربية النموذجية كلها.

الحواشي

[٤٥] تتفرع المدونات إلى ما هو أكثر تفصيلًا من هذه التفريعات التي تتعلق بعدد لغة المدونة، وهنا أذكر -بشكل موجز- التقسيمات المعمول بها وفق النوع والغرض والعدد والتصميم إلخ. فعلى صعيد النوع، هناك مدونات لغوية حاسوبية للغات الكلاسيكية، وهناك ما هو للغة المعاصرة، وهناك ما يتضمن أوعية genres معينة مثل مدونة الصحف وهكذا دواليك. وثمة مدونات من حيث النصوص تتضمن نصوصًا ليست كاملة وتعرف بمدونات العينة sample corpus، ونصوصا حية كاملة من حيث الإنتاج، وتعرف بالمدونات كاملة النصوص full-text corpora. وهناك مدونات من حيث الوعاء الناقل للغة الحية، والذي

يكون نصا مكتوبا، أو منطوقا يُحول إلى نص مكتوب. وتفترق المدونات أيضا من حيث التنوع الزمني، فهناك مدونة لغوية معاصرة مثل مدونة السليطي al-Sulaiti (٢٠٠٤) (مدونة العربية المعاصرة Contemporary Arabic Corpus، انظر أيضا: السليطيal-Sulaiti 2009)، ومدونة زمنية متعلقة بالتطور الزمني اللغوي مثل مدونة هيلسنكي للنصوص الإنجليزية the Helsinki Corpus of English Texts التي تتضمن مليون ونصف المليون كلمة للإنجليزية القديمة والوسطي وبداية العصر الحديث (١٩٩١). كما تتمايز المدونات أيضا من حيث مستعملي اللغة إلى مدونات اللغة الأم ومدونات اللغة الثانية أو الأجنبية للمتعلمين (مثل مدونة أبوحكيمة Abuhakema 2009 ومدونة الفيفي وأتويل ٢٠١٢). أما من حيث الغرض، فهناك مدونات حاسوبية لأغراض عامة general-purpose corpora، ومدونات محددة المجال domain-specific corpora. أما التقسيم الثالث؛ فهو ما يتعلق بالعدد، أي: بعدد اللغة في المدونة، وتنقسم إلى مدونة أحادية وتقابلية ومتوازية. أما آخر صنف تتمايز به المدونات فهو ما يتعلق بالتصميم (بمعنى آخر: بالتوسيم النحوي Part of Speech Tagging والتحشية annotation) حيث إن هناك مدونات لم يعمل لها توسيم وتحشية، أو ناقصة من حيث عناصر التحشية: فواصل الصفحات، والفقرات، والفوارق والتشاكل الإملائيين، والأرقام، والزوائد، والرموز الأجنبية المختلفة عن رموز لغة المدونة الأصلية، وتواريخ النصوص، والمؤلفين، والأوعية، والمجالات، والموضوعات، والمستويات اللغوية، والسجل اللغوي (لمزيد من الأمثلة حول أنواع المدونات الغربية وأمثلتها؛ انظر: بيكر وآخرين Baker et al 2006: ٤٩، ٧١، ١٢٦، ١٤٢).

ووفقاً لهذا التصنيف، فإنه لا يمكن أن يقوم باحث لغوي -على سبيل المثال- بتحليل أنواع تراكيب الجمل الفعلية، أو وظائف حروف الجر الدلالية، أو تحليل الخطاب الديني في الثقافة العربية، أو تفسير الموجودات والمحسوسات والمواد الثقافية بعلم الوجود في اللغة العربية باستعمال مدونة عربية معاصرة (مثل

مدونة السليطي) أو باستعمال مدونة متعلمي اللغة العربية (مثل مدونة الفيفي وأتويل) إلا وأن يكون مدركاً بأية مدونة يكون لها أساس علمي يُجيب عن تلك الأنواع من التحليلات.

[٤٦] التوسيم النحوي Part-of-Speech Tagging (ويُسمى أيضا بـ Grammatical Tagging؛ انظر: ميكنري وآخرون McEnery et al 2006)، والتوسيم الدلالي Semantic Tagging، والتوسيم النبري Prosodic Tagging، والتحشية Annotation؛ كلها إجراءات حاسوبية مطلوبة لتطوير أي مدونة حاسوبية، وكل هذه التوسيمات، وبالأخص التوسيم النحوي، يصعب إنجازها في المدونة العربية الحاسوبية لسبب يُعزى إلى حاجتها إلى تطوير التوسيم النحوي بصورة يدوية لأقسام الكلم شكلًا ووظيفةً من أجل ضمان تطابق أكبر وأدق. ولو قلنا -مثلا-مدونة مكونة من مليون كلمة، فإن توسيم أقسام الكلام فيها يدويا يحتاج إلى فريق عمل يصل إلى المائة ربما، ويُنجز التوسيم بشكل يومي يدويا، وعلى مدى سنوات عديدة على أقل تقدير. وثمة توسيمات نحوية مقترحة وعديدة، ومعظمها يُمكن أن تحقق نسبًا مرضية من التطابق، ويرى الباحث أن نسب التطابق في المدونات العربية لا يُمكن أن تتجاوز نسبة ٧٠٪ للمدونات بسبب صعوبة ضبط كل وظائف الكلمات العربية المتنوعة بتنوع أنظمتها الخطية والصوائتية والتصريفية والنوعية والوظيفية. ويرى الباحث أيضا أن أي منجز أو اقتراح أو عمل تكون جهوياته منصبة على محاولات تطوير التوسيم النحوي في العربية، ولا يكون مقبولا بافتراضاته في اختباره وفحصه آليا، فإن نطاق تطوير التوسيم النحوي العربي سيكون منحسرًا ولن يُكتب له التطوير.

[٤٧] انظر إلى الحاشية رقم [٤٥].

[٤٨] ثمة مدونة متوازية ليست متوفرة، وصُممت من قبل جامعة جون هوبكنز John Hopkins University، وهي مدونة توازي مقوّمات constituents الآيات القرآنية بمقابلاتها الترجمية في الإنجليزية، أي أنها مدونة تتضمن آيات القرآن

الكريم وما يقابلها في الترجمة الفعلية؛ حيث لكل مقوم وعنصر لفظي في كل آية ما يوازيه (أو يحاذيه aligning؛ من محاذاة alignment) من مقابلات في اللغة الإنجليزية. وحيث إن هذه المدونة ليست متاحة، فلا يُعلم ما إن كانت النصوص المترجمة دقيقة للنص القرآني أم لا.

[٤٩] ينقسم معظم المدونات العربية الحاسوبية من حيث التوسيم إلى قسمين؛ الأول: شبه الموسّمة semi-tagged (بخلاف partially tagged الموسّم جزئيًا) كحال مدونة أرابيكوربوس ١٧٣,٦٠٠,٠٠٠ مليون كلمة arabicCorpus http://arabicorpus.byu.edu، والمدونة العربية KACSTAC بنسختها الجديدة. أما القسم الثاني فهو غير الموسّم non-tagged كحال معظم المدونات العربية العامة، ومنها مدونات العربية الشبكية Arabic Internet Corpora (٣١٧ مليون كلمة) لـ سيرج شاروف Serge Shroff (انظر: http://smlc09.leeds.ac.uk/query-ar.html). ولا توجد مدونة عربية حاسوبية موسومة بصورة كاملة كحال، على سبيل المثال، مدونة بيرمنغهام الإنجليزية WordBanks Online (البنك الإنجليزي Bank of English http://wordbanks.harpercollins.co.uk/Docs/Help/guide.html). ثمة مدونات عربية تتضمن نصوصا عربية رقمية، ولا يُرى إضافتها هنا نظرًا لاختصاصها بسجلات لغوية وأوعية لغوية خاصة للعربية، هذا بالإضافة إلى عدم تجاوز عدد كلماتها ١٠٠ مليون كلمة (انظر: http://www.kacstac.org.sa/osact/UsefulResources.html)، ومنها المُجذّع جزئيا partially tokenized والموسم جزئيا partially tagged (وليس شبه الموسم semi-tagged): Arabic Gigaword Corpus Firth Edition (انظر الثبيتي al-Thubaity 2014).

[٥٠] انظر (بارلو http://www.monoconc.com (Barlow 2014/

[٥١] IT Servieces' Oxford University: برنامج يعالج النصوص وملحق بالمدونة الوطنية البريطانية British National Corpus؛ انظر: http://www.

natcorp.ox.ac.uk/tools/index.xml ويمكن تحميل البرنامج من هذا الرابط http://sourceforge.net/project/showfiles.php?group_id=١٣٠٢٨٩. ويدعم هذا البرنامج تحليل مصادر لغوية ضخمة ومدونات لغوية طبيعية على امتداد ملفات XML.

[٥٢] انظر (تسوكاموتو Tsukamoto n.d.) http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/index_e.html

[٥٣] انظر (روبرتس Roberts 2014) http://www.andy-roberts.net/coding/aconcorde

[٥٤] انظر (سكوت Scott 2012) WordSmith Version 6 http://www.lexically.net/wordsmith/version6/index.html

[٥٥] انظر الثبيتي وآخرين al-Thubaity et al (ACP Tool) 2013 http://sourceforge.net/projects/kacst-acptool

[٥٦] انظر: https://the.sketchengine.co.uk/login

[٥٧] هذه القياسات الإحصائية كافية فيما يراه باحث هذه الورقة، وهي أساس البحث اللغوي في أي مدونة لغوية حاسوبية. وثمة حزم أخرى لا تختلف عن هذه الأساس، ونقاط اختلافها تكمن في قراءات خوارزمية أخرى تُوصل إلى ما يقربُ من فوائد المعطيات الأساسية ذاتها (أي: قبول قوة التصاحب من عدمه).

[٥٨] يقوم هذا التقسيم أساسا على نقل مفاهيم ستيفان جريس Stefan Gries (٢٠١٠) حول إعادة نظرته إحصائياً لمصداقية التحليل الحاسوبي لأي مدونة، ويلخص من خلال تحليليه أن أية مدونة من حيث الأصناف الأربعة: النوع أو الغرض أو العدد أو التصميم، فإن على الباحث فيها أن يلتزم في تقديم

الإحصاءات بما يتوافق وعينة تلك الأصناف الأربعة، وبشكل دقيق فيما يخص التصاحب أو التلازم أو التجاور أو التضام (التقارب).

أما هذه المقابلات العربية فيُرى ضبطُها بصورة لعلها تكون مقبولة مستقبلا، وهي قابلة للدحض بلا شك؛ وشكل ضبطها هو على النحو الآتي: التصاحب collocation، والتلازم colligation، والتجاور collostruction، والتضام (الدلالي) semantic set. والمعاني اللغوية لكل مقابل عربي هي: مفهوم التصاحب (أي: القريب اللفظي المباشر) أو التلازم النحوي (القريب النحوي المباشر)، أو التجاور (التجاور اللفظي تركيبًا وغير المباشر)، أو التضام الدلالي (أي التقارب الدلالي semantic preference) أي: التضامات الدلالية التي تشترك في العلاقات الوجودية للحقول الدلالية ontological relationships of Arabic semantic fields (انظر الحاشية ٤٩). وهذا الأخيرة هي حلم الدرس اللغوي العربي، ويتحتّم أن تكون واقعًا مأمولًا لكل مختص عربي في الدرس اللساني الحديث، وعلى كل باحث لغوي أن يأخذ في تخصصه شيئا من التقنيات التي تُوفرها أدوات معالجة اللغة آليا، والكيفية النموذجية في بحث سيُكوّن مستقبلًا كمًا يقدم لقرائه مزيدًا من التفتّق المعرفي الذي يأتي به باحثون جدد حول هذا الموضوع. إن هذه القضية وما ذُكر هنا هو شرط العلوم البنائية في العلوم الحديثة.

[٥٩] الإشارة هنا إلى إحدى المدونات العربية المحوسبة التي تُمكّن نمطًا من أنماط البحث التمكينيّة؛ والمتوفرة في نظام الاستعلام query system، أو بالاستناد إلى مدونة عربية يقوم الباحث بجمعها في ملف text ومن ثمّ إضافتها إلى أداة من أدوات تحليل اللغة العربية وأفضلها غواص و Sketch Engine. وتُربط هذه الخصائص البحثية بثلاث مدونات عربية رئيسة، كل واحدة منها تُوفّر أدوات تحليلية خاصة بها. ولزاما على كل باحث لغوي أن ينظر إلى مستجدات تلك المدونات العربية الثلاث: المدونة العربية (مدينة الملك عبدالعزيز

للعلوم والتقنية) http://www.kacstac.org.sa / ، ومدونة أرابيكوربوس (جامعة بريغهام يونغ الأمريكية) http://arabicorpus.byu.edu/، ومدونة جامعة ليدز لمحتوى العربي الشبكي من بعض الصحف وكامل الويكيبيديا http://corpus.leeds.ac.uk/internet.html. وفي كل مدونة من هاته المدونات أدوات بحث خاصة تتطور، غير أن أكثر الأدوات الواعدة تطورًا فيها هي المدونة العربية، ومن ثمة مدونة أرابيكوربوس كونها قد زادت من عام ٢٠١٠ حتى عام ٢٠١٢ بما يزيد عن ٩٠ مليون كلمة مع إمكانية البحث عن الجذر اسميًا وفعليًا، وعن الجذع ولواصقه الاشتقاقية والتصريفية (word forms)، وعن إمكانية عرض النص الأصلي كاملًا بتطابقات الكلمة المبحوث عنها.

[٦٠] هذا الدراسة الأجنبية ودراسة الحمزاوي (١٩٨٦) من قبلها تُعدّان المحك الأساس والتحدي الأكبر لقضايا اللغة المتنوعة زمنيا ووصفيًا من جهة. ومن جهة ثانية: التحدي الآخر الذي يتطلّب من المؤسسات أن تجعل من التوقيف (أو التوسيط) للمتغيرات اللغوية استعمالا الإطارَ العام لـ: التوصيف الدقيق لمواد الثقافة العربية بالوحدات المعجمية العربية، والاستعمالات التركيبية التوليدية المستجدة، والوحدات المعجمية العربية المعربة والدخيلة، والأساسَ لها جميعًا في بناء مواد العربية التعليمية ومعاجمها متعددة الأغراض.

المراجع

المراجع العربية

الثبيتي، عبد المحسن وآخرون. غواص، KACST, ACP Tools، http://sourceforge.net/error-404.html?project=kacst-acptool.

حبش، نزار. مقدمة في المعالجة الطبيعية للغة العربية، ترجمة: هند بنت سليمان الخلفية. دار جامعة الملك سعود للنشر، الرياض، ٢٠١٤.

الحمزاوي، محمد رشاد. العربية والحداثة أو الفصاحة فصاحات، دار الغرب الإسلامي، بيروت، ١٩٨٦.

العصيمي، صالح بن فهد. علم المتون وعلوم اللغة. مجلة كلية الآداب والعلوم الإنسانية، كلية الآداب والعلوم الإنسانية، فاس (المملكة المغربية)، العدد (١٩)، ٢٠١٣، ص. ٣٧-٦٧.

الفيفي، عبدالله وأتويل، إريك. المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية. المؤتمر الدولي لعلوم وهندسة الحاسوب باللغة العربية (الدورة الثامنة)، ٢٦-٢٨ ديسمبر، جامعة القاهرة، ٢٠١٢.

المراجع الإنجليزية

Abuhakema, G., Feldman, A. and Fitzpatrick, E. ARIDA: An Arabic Interlanguage Database and Its Applications: A Pilot Study. Journal of the National Council of Less Commonly Taught Languages, 7, 2009, pp. 161-184.

Al-Ajmi, H. Compiling an English-Arabic parallel text corpus. In: Proceedings of Asian Association for Lexicography, August 27-29, Urayasu (Japan): Meikai University, 2003, pp.51-54.

Al-Faifi, Abdullah and Atwell, Eric. Arabic Learner Corpus and Its Potential Role in Teaching Arabic to Non-Native Speakers. In the Proceedings of the Seventh Biennial IVACS conference, 19 - 21 Jun 2014. Newcastle (UK), 2014.

Al-Mujaiwel, S. Contrastive Lexicology and Comparable English-Arabic Corpora-Based Analysis of Vague and Mistranslated Arabic Equivalence. Ph.D. thesis, Exeter University, 2012.

Al-Sulaiti, L. Designing and Developing a Corpus of Contemporary Arabic, MSc dissertation. Leeds: Leeds University's School of Computing, 2004.

Al-Sulaiti, L., A Survey of Arabic Corpora. 2009. http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

Al-Thubaity, A., et al. New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. In the Proceedings of the International Conference on Asian Language Processing in Urumqi, China, 2013.

Al-Thubaity, et al. ACP Tool. Available for free use in: http://sourceforge.net/projects/kacst-acptool/, 2013.

Al-Thubaity, A. O. A 700M+ corpus: KACST Arabic corpus design and construction. Language resources and evaluation, 2014. DOI 10.1007/s10579-014-9284-1. Arabic

Baker, P. et al. A Glossary of Corpus Linguistics. Edinburgh: Edinburgh University Press, 2006.

Baker, P. Sociolinguistics and Corpus Linguistics. Edinburgh University Press, 2010.

Barlow, M. MonoConc, version 2.2, 2014. http://www.monoconc.com/.

Bernard, Lou and Dodd, Tony. Xaira: an XML aware tool for corpus searching. In: Archer, D., Rayson, P., Wilson, A., and McEnery, T., eds., the Proceedings of the Corpus Linguistics Conference, 16, 2003, pp. 142–144, UCREL, University of Lancaster.

Berry-Rogghe, Godelieve. The Computation of Collocations and Their Relevance in Lexical Studies. In: Aitken, A., J., Bailey, R., W., and Hamilton-Smith, N., the Computer and Literary Studies. Edinburgh University Press, 1973, pp. 103-112.

Church, Kenneth, and Hanks, Patrick. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1), 1990, pp., 22-29.

Diab, Mona. Improved Arabic Base Phrase Chunking with a New Enriched POS tag set. In Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007, pp. 89-96.

Diab, Mona. Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In: Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009.

Firth, J., R. A Synopsis of Linguistic Theory 1950-1955: Studies in Linguistic Analysis. Blackwell, Oxford, 1957.

Goldberg, Adele E. The Nature of Generalization in Language. Cognitive Linguistics, 20(1), 2009, pp. 93-127.

Granger, S. et al., eds. Corpus-based Approaches to Contrastive Linguistics and Translation Studies, Rodopi, Amsterdam, 2008.

Granger, S. The Corpus approach: a common way forward for Contrastive Linguistics and Translation Studies. In: Granger, S. Lerot, J. and Petch-Tyson, S., eds., Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Rodopi, Amsterdam, 2008, pp. 18-29.

Gries, St. Th. Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics, 8(2), 2003, pp. 209-243.

Gries, St., Th. "Useful Statistics for Corpus Linguistics." In: A. Sánchez and M. Almela, eds., a Mosaic of Corpus Linguistics: Selected Approaches. Frankfurt: Peter Lang, 2010, pp. 269–291.

Gries, St., Th. Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 2008, pp. 403–437.

Gries, St., Th. Quantitative Corpus Linguistics with R: A Practical Introduction, Routledge, London, 2009.

Habash, Nizar, et al. MADA+TOKEN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In Proceedings of the second International Conference on Arabic Language Resources and Tools, Cairo, 2009.

Hunston, S. Corpora in Applied Linguistics. Cambridge: Cambridge University Press, 2002.

Hunston, Susan., Colligation, Lexis, Pattern, and Text. In: Scott, Mike, and Thompson, eds., Patterns of Text: In Honour of Michael Hoey. John Benjamins, Amsterdam, 2001, pp. 14-33.

IT Services's Oxford University. British National Corpus and Xaira, http://projects.oucs.ox.ac.uk/xaira/ and http://www.natcorp.ox.ac.uk/tools/index.xml.

Jakubíček, M., et al. Fast syntactice searching in very large corpora for many languages. PACLIC 2010, Japan. http://www.sketchengine.co.uk/documentation/wiki/SkE/DocsIndex.

Jakubíček, Miloš, et al. Fast syntactic searching in very large corpora for many languages, Japan, PACLIC, 2010, pp. 741-746.

Johansson, S. Contrastive linguistics and corpora. In: Granger, S. Lerot, J. and Petch-Tyson, S., eds., Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Rodopi, Amsterdam, 2008, pp. 31-44.

Kilgariff, Adam, et al. A quantitative Evaluation of Word Sketches. EURALEX, the Netherlands, Leeuwarden, July 2010.

Kilgarrif, Adam, et al. The Sketch Engine (Lexical Computing Ltd.), https://the.sketchengine.co.uk/login/.

Kilgarrif, Adam, et al. The Sketch Engine: ten years on. Lexicography, 1(1), 2014, pp. 7-36.

Kilgarriff, Adam, et al. The Sketch Engine. In: Proceedings of EURALEX, Lorient, France, 2004, pp. 105-116, http://www.sketchengine.co.uk.

Lipka, L. English Lexicology: Lexical Structure, Word Semantics, and Word-Formation. Tübingen: Narr Studienbücher, 2002.

McEnery, Anthony. Corpus Linguistics. In: Mitkov, R., ed., Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press, 2003, pp. 448-463.

McEnery, Tony and Hardie, Andrew. Corpus Linguistics. Cambridge University Press, Cambridge, 2012.

McEnery, Tony and Xiao, Zhonghua. Parallel and comparable corpora: What is happening? In: Anderman, G. and Rogers, M., eds., Incorporating Corpora: The Linguist and the Translator. Clevedon: Multilingual Matters, 2007, pp. 18-3.

McEnery, Tony, et al. Corpus-Based Language Studies, an advance resource book. Routledge, Oxford, 2006.

Moll, Clàudia, P. When diachrony meets synchrony. How linguistic variation sheds light on the origin of synchronic phonological processes and theories of language change. Dialect Laboratory. Dialects as a testing ground for theories of language

change. Studies in Language Companion Series, 128, 2012, pp.197-225.

Oakes, M. Statistics for Corpus Linguistics, Edinburgh: Edinburgh University Press, 1998.

Piao, S., Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University, 2000.

Piao, S. Word alignment in English-Chinese parallel corpora. Literary and Linguistic Computing 17(2), 2002, pp. 207-30.

Price, Todd L. Structural Lexicology and the Greek New Testament: Applying Corpus Linguistics for Word Sense Possibility Delimitation Using Collocational Indicators. Ph.D. thesis. Middlesex University, 2013.

Roberts, A., al-Sulaiti, L., Atwell, E. aConCorde: Towards an open-source, extendable concordancer for Arabic. Corpora, Edinburgh University Press, 1, 2006, pp. 39-57.

Roberts, Andrew. aConCorde. 2014. http://www.andy-roberts.net/coding/aconcorde.

Rouhani, Jameela. An Applied Research Into the Linguistic Theory of Collocation: English-Arabic Dictionary of Selected Collocations and Figurative Expressions with an Arabic Index. Ph.D. thesis, Exeter University, 1994.

Rychly, Pavel. A Lexicographer-Friendly Association Score. In: Proceedings of 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN, Masaryk University, Brno, 2008.

Scott, Mike. WordSmith Tools 5.0. Lexical Analysis Software, 2010.

Scott, Mike. WordSmith Tools, version 6. Lexical Analysis Software, Liverpool, 2012. http://www.lexically.net/wordsmith/version6/index.html.

Sharoff, S. Harness¬ing the lawless: using comparable corpora to find translation equiva¬lents. Journal of Applied Linguistics, 2004, 1(3), pp. 333-350.

Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, 1991.

Sinclair, J. Reading Concordances: An Introduction. London: Pearson, 2003.

Sinclair, J. Trust the Text: Language, Corpus and Discourse. Edited with Ronald Carter. Routledge, London, 2004.

Tognini-Bonelli, E., Lexis in contrast. In: Granger, S. and Altenberg, B., eds., Studies in Corpus Linguistics. Benjamins, Amsterdam, 2000, pp. 3-48.

Tsukamoto, Satoru. KWIC Concordance, (n.d.) http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/index_e.html.