مجلة الآداب، م29، ع2، ص ص 3-30، جامعة الملك سعود، الرياض (1438/2017هـ)

أُسُس تحليل التصاحب اللفظي في المدونة اللغوية العربية

سلطان بن ناصر المجيول
الأستاذ المساعد في لغويات المدونة الحاسوبية بقسم اللغة العربية وآدابها, كلية الآداب, جامعة الملك سعود, الرياض, المملكة العربية السعودية

(قدم للنشر في 1437/7/16هـ, وقبل للنشر في 1438/3/12هـ)

الكلمات المفتاحية : التصاحب اللفظي, التحليل اللغوي, المدونة اللغوية العربية, النصوص العربية, إحصاءات التصاحب اللفظي, أدوات معالجة المدونة العربية, محرك الاستعلام (اللغوي).

ملخص البحث: تختص هذه الدراسة بمفهوم التصاحب اللفظي في الفكر اللغوي العربي وفي الدرس اللغوي الحديث, وتركّز على ما استجدَّ من مناهج لغويات المدونة الحاسوبية corpus linguistics في آليات التحليل التصاحبي. ويرتكز مفهوم التصاحب على تفسيرات متقاربة نوعًا ما بين الفكر العربي والدرس اللغوي الحديث, غير أنَّ الأخير قد زُوِّد بإطار آلي تحليلي إحصائي مُعِينٍ على تفسيرات دقيقة لأنواع التصاحب اللفظي العربي المتمثلة في نصوص عربية ذات أوعية محددة أو متنوعة بتحديد أو تنوّع الغرض البحثي وأسئلته التي تتواءم معها. وتكون هذا النصوص مجموعة في الملفات النصية text files. ويقف البحث على برنامج أدوات معالجة المدونة العربية Arabic Corpus Processing Tools (وهي أداة مفتوحة المصدر), وأدوات "محرك التخطيط" Sketch Engine اللغوي, وأهم وظائفهما في التحليل التصاحبي, كما يقف بعد ذلك على أهم الإحصاءات التحليلية في تحليل التصاحب اللفظي, وهي المعلومات المتبادلة Mutual Information, وقياس_ت t-score وقياس الزَّهرة Dice والزهرة اللوغاريثمية Log-Dice, مع مثال تطبيقي على معاجم اللغة العربية القديمة والحديثة التي يبلغ عدد كلماتها 20 مليون كلمة تقريبًا.

Foundations of Collocation Analysis in the Arabic Corpus

Sultan Almujaiwel
An Assistant Professor of Corpus Linguistics, Arabic Language Department, College of Arts, King Saud University, Riyadh, Saudi Arabia

(Received 16/7/1437H; Accepted for publication 12/3/1438H)

Keywords: collocation;linguistic analysis; Arabic corpus; Arabic texts; collocational statistics; Arabic Corpus Processing Tools ACPTs; Sketch Engine**.**

Abstract: This paper tackles the concept of collocations in the Arabic linguistic tradition and modern linguistics with a core focus on the latest developments in approaches to corpus linguistics in terms of mechanisms of collocational analysis. The concept of collocation is somehow based on approximate interpretations between the Arabic thought and modern linguistics, however, modern linguistics has been furnished with an analytical and statistical framework using computerized applications as a tool for helping provide accurate interpretations for several kinds of Arabic collocations as reflected in the specific resources of Arabic texts by identifying the aim of research and its relevant questions. Such texts are placed in text file. The research relies on the Arabic Corpus Processing Tools as an Open-Source and Sketch Engine tools, together with their key functions in collocational analysis. The research sheds light as well on the analytics employed in the analysis of collocations e.g. Mutual Information, T-Score, Dice and LogDice, along with a case study of an example in a 20+ million-word corpus of classical and modern Arabic dictionaries .

مقدمة

أصبح الاهتمام بالحوسبة والأرقام في الدراسات الإنسانية محل اتساع وتداخل وإفادة واستفادة بين حقول العلوم الرياضية والحاسوبية والإحصائية والإنسانية, كما أنَّ ثمَّة اتجاهات في الدرس اللغوي العالمي نحو أهمية هذا التداخل الذي شكّل عدة علوم جديدة, كلغويات المدونة الحاسوبية corpus linguistics ومصادر الإنسانيات الرقمية وتقويمها Digital Resources of Humanity and Evaluation وغيرهما.

وتتمحور هذا الدراسة في التعريض بآليات التحليل للمدونة اللغوية, والوقوف على قضية التصاحب اللفظي collocation بطريقة أساسية. وتكمن أهمية هذه الدراسة في تأطير موضوع التصاحب اللفظي بمناهج لسانية حاسوبية جديدة تساعد ما انشغل به البحث اللغوي العربي في جوانبه التنظيرية التي اهتم اللغويون العرب في العصر الحديث فيها بقضية التصاحب اللفظي من جوانب تركيبية/نحوية/معجمية, مع اختلاف اللفظ المصطلحي الدال عليه كالتصاحب (عبدالعزيز 1991م: 11), أو التضام (حسان 1998م: 157), أو التلازم (عمر 2007م: 37), أو الرصف (البركاوي 1991م: 238). وكان الاهتمام بهذه القضية في البحث اللغوي العربي المعاصر من جهوية المنشغلين بالنحو العربي والتركيب أو المعجم, أمَّا في الدرس اللغوي الحديث, فقد تبلورت بدءًا من أعمال Firth (1957) التي جمعها في كتابه Synopsis of Linguistic Theory (موجز النظرية اللغوية), واتسعت مرورًا بتناول Sinclair 1991, (2004), لها في منهجه المعتمد على تحليل المدونة اللغوية corpus, وتشعّبت عند Gries 2003, 2008,) 2009, 2010) في منهجه الإحصائي القائم على كشف التوزيعات التركيبية المتجاذبة والمتنافرة لأحياز الألفاظ التركيبية في الإنجليزية.

وسيفيد هذا البحث بأسس تحليل التصاحب اللفظي في الدرس اللغوي الحديث, وبخاصة في الحقل اللغوي التطبيقي: لغويات المدونة الحاسوبية corpus linguistics.

وقُسِّمت أجزاء البحث إلى خمسة مباحث؛ هي على النحو الآتي:

المبحث الأول: مفاهيم التصاحب اللفظي بدءًا بتصنيفاته في الدرس اللغوي العربي المعاصر, ثم عند Firth (1957), وصولًا إلى تأطير Sinclair (1991, في التحليل المعتمد على المدونة الحاسوبية corpus, ولن نقف على ما اعتمد عليه Gries (2003) كونه منهجًا تُجْرَى تحليلاته على لغة البرمجة R, ويحتاج إلى بحث آخر لعدم سعة هذا البحث له, غير أنَّ مفاهيم التصاحب لديه ستُذكر كونها من أسس التحليل التصاحبي في المدونة اللغوية.

المبحث الثاني: كيفية تحديد النص المتوائم مع غرض التحليل بالطريقة المقبولة منهجيًا وأخلاقيًا وقانونيًا, وهي أفضل الطرق بدلًا من الاعتماد على المدونات الحاسوبية الشبكية المحددة نصوصها وأدواتها (انظر صالح 2015م: 67-72), و(المجيول 2015م: 266-269), حول أنواع المدونات العربية الحاسوبية في الشابكة Web), التي قد لا تتيح نطاقًا بحثيًا أوسع لفرضياتٍ وأسئلةٍ لغوية خاصة تتطلب اختصاص نصوص المدونة وأجناسها وعصور إنتاجها.

المبحث الثالث: وفيه سنعرّج على أفضل أدوات معالجة النصوص العربية, وعلى وظائفها الحاسوبية المتعلقة بمعالجة أمثلة التصاحب اللفظي وأنماطه.

المبحث الرابع: وفيه وقوف على أهم الإحصاءات المعمول بها في قياس التصاحب اللفظي, وهي قياسات مبنية خوارزميًا في الأدوات المذكورة في المبحث الثالث, ولا تكلّف الباحث اللغوي إلَّا عناء قراءتها وتفسيرها للغرض البحثي.

المبحث الخامس: تحليل تطبيقي لمثال تصاحبي لفظي على مدونة معاجم اللغة العربية البالغ عددها 22 معجمًا.

التصاحب اللفظي
collocation

اعتمد الفكر اللغوي العربي في مجمله لتفسير مفهوم التصاحب اللفظي على تحديد أربعة أوجه رئيسة له؛ الأول: الحر الذي فيه يكون التصاحب بين الكلمتين على المحور الاستبدالي paradigmatic مفتوحًا, والثاني: التضام المقيد الذي يكون بين كلمتين لا يُمكن أن يُستبدل إحداهما بالأخرى, والثالث: التعبيرات الاصطلاحية التي تتصاحب

فيها كلمتان أو أكثر لتدل على وحدة دلالية واحدة مختلفة عن دلالة كل كلمة منها. أمَّا الوجه الرابع فهو التلازم التركيبي الذي ينبني وفقًا للمعنى النحوي القياسي الذي يؤدِّي المعنى التام بتكامل تلازم أركانه التركيبية, والذي يختلف عن التصاحب اللفظي في أساس ارتباطه الخاص بالقياس النحوي كالتعليق النحوي والرتبة والتقديم والتأخير بين المتلازمات النحوية (محمد 2011م.), وسُمِّي تلازمًا خاصًا بالنحو؛ لأنَّ ركني التلازم لا يمكن أنْ ينفصلا نحويًا, فعلى سبيل المثال: تلازم الأسماء المجرورة بحروف الجر, وتلازم الحال مع صاحب الحال, وتلازم الفاعل مع الفعل, وهكذا دواليك على هذا المنوال.

أمَّا في الدرس اللغوي الحديث فإنَّ مفاهيم التصاحب قد تعددت وفقًا لعدة مناهج متلاحقة ومتطورة, وسيُعرض هنا هذه المفاهيم عند كل من (Firth 1975) و(Sinclair 1991, 2004) و Gries (2003, 2008, 2009, 2012) و Stefanowitsch and Gries (2003).

فعند الأول؛ نجده قد وضع مفاهيم الأنماط التصاحبية التي تحدد الواسم التركيبي لكل كلمة من الكلمات المتصاحبة, وتضمنت نظريته اللغوية في كتابه A Synopsis of Linguistic Theory ثلاثة أسس تكوينية للتصاحب اللفظي وهي: المقامية situational والقواعدية grammatical والتصاحبية collocational التي تستلزم أيْ معالجة للمداخل المعجمية, وتقوم هذه الأسس على النزعة السياقية التي اشتهر بها Firth نفسه. والتصاحب اللفظي عنده صارم كذلك, فهو يعتمد على النحو النمطي pattern grammar الذي يجمع أنواع التصاحب اللفظي النحوي النمطي بحسب القسم الكلمي لكل مكون من مكونات التصاحب اللفظي, على عكس Sinclair الذي لم يهتم بقضية هذا التنميط, بل جعل تحليل التصاحب اللفظي قائمًا بحسب ظهوره وظهور أمثلته من اللغة الطبيعية أو الحية (انظر المجيول 2016م.), وعليه فإنَّ المتصاحبات اللفظية الآتية: منوانٌ برًا, وصاعٌ تمرًا, وكيلا تفاحًا, ... إلخ تُرى بقياس فيرث على أنَّها نوع واحد من أنواع التصاحب اللفظي, أساسه النمطي هو (مبتدأ+ تمييز) على تأويل الضمير (هو) في أوله أو (خبر+ تمييز) على تأويل شبه الجملة (عندي) في أوله, بخلاف رؤية (سنكلير) التي تجعل من هذه الأمثلة ثلاثة أمثلة أو أكثر للتصاحب اللفظي وفقًا للاستعمال اللغوي الطبيعي في المدونة الحاسوبية corpus.

أمَّا في منهج Gries وبعض أعماله التي تشارك فيها مع زميله Stefanowitsch فهي قائمة على تحليل التصاحب بمفهوم التجاور التركيبي collostruction, والتجاور التركيبي هنا يعني ما يدل عليه جزءٌ من التلازم التركيبي colligation في أمثلة التصاحب, وعلى ذلك يكون التجاور التركيبي جزءًا من التلازم النحوي colligation, ولو وضعنا مثالًا من العربية وَفْقَ هذا المفهوم, لقلنا-على سبيل المثال- إنَّ النمط النَّحوي (مصدر عامل+ مفعول به) يعدُّ مثالًا للتلازم النَّحوي, أمَّا الاحتمالات التي قد ترد منها في اللغة الطبيعية في المدونة نظرًا لغياب التشكيل فقد تكون: ضربٌ الرقابَ أو ضربُ الرقابِ, وعليه: يعدان مثاليْن للتجاور التركيبي.

كما أنَّ لـ Gries (2013: 100) تفسيرات دقيقة جدًا تتعلق بالتجاور التركيبي, حيث يقسم تحليل التجاور التركيبي إلى ثلاثة أقسام:

القسم الأول: تحليل التصاحب اللكسيمي collexeme analysis الذي به تُحوسب الكلمات المركزية nodal item ومدى قوة انجذابها لحيز تركيبي معين.

القسم الثاني: التحليل اللكسيمي المتميز distinctive collexeme analysis الذي به تُحوسب الكلمات المركزية ومدى قوة انجذابها إلى تراكيب متشابهة وظيفيًا.

القسم الثالث: التحليل اللكسيمي للتصاحب المتفاوت co-varying collexeme analysis الذي به تُحوسب الكلمات المركزية في حيز تركيبي معين وتُقاس مدى انجذابها إلى كلمات أخرى في حيز تركيبي آخر ضمن التركيب النحوي الواحد.

وفي سياق عرض تحليل التصاحب اللفظي في المبحث الثالث (أدوات التحليل), والإحصاء في التحليل (المبحث الرابع) سيكون التركيز على طريقة عمل أدوات تحليل المدونة العربية ACPTs وعمل (محرك التخطيط), والإحصاءات المتعلقة بالتصاحب اللفظي, ولن

نربط طريقة التحليل بهذه الأدوات وتلك الإحصاءات بمفهوم معين من مفاهيم التصاحب وتقسيماته؛ لأنَّ تعدد هذه الاتجاهات في تناول التصاحب اللفظي بالتحليل الآلي تلتقي ما دامت أغراضه المتعددة تصبُّ في مصلحة التحليل الآلي للتصاحب اللفظي أو التلازم النحوي أو التجاور التركيبي أو التفاضل الدلالي التي تحلل كلها في لغويات المدونة الحاسوبية corpus linguistics بـالتتابع اللفظي n-grams (النغراميّة).

وتعتمد تحليلات التتابع اللفظي في لغويات المدونات الحاسوبية على خمسة مفاهيم رئيسة, وهي على النحو الآتي:

1-العقدة node أو الكلمة المركزية nodal item التي تدل على الكلمة التي يُراد البحث عنها أو الانطلاق بالبحث الآلي منها.

2-التكرار frequency الذي ينظر إلى تكرار الكلمة المركزية في النص وتكرار المتصاحب اللفظي معها. ولهذا المفهوم في الدرس اللغوي العربي مقابل ليس قريبًا ولا بعيدًا, قد سُمِّي بالتواتر la frequence في الاستعمال الدال على رسوخ لفظة ما مع متصاحب لفظي ما في العبارات المألوفة, وهو مصطلح عربي تراثي يحمل مفهوم العرب عن كل ما يتواتر في العربية إلى أن يكون مثلًا سائرًا (عمر 2007م: 37).

3- التصاحب اللفظي collocation الذي يكون وفق قياس التتابعات اللفظية أو النغراميات n-grams, ويمتد هذا القياس من كلمة مصاحبة إلى خمس كلمات مصاحبة ترد قبل الكلمة المركزية (5>n) أو ترد بعدها (5<n).

4- الكشاف السياقي concordance الذي يستردّ سياق الكلمة آليًا من النص ويُظهر نتائج سلسلة الكلمات المتتابعة قبل الكلمة المركزية وبعدها على امتداد سياقي يبدأ من كلمتين أو تتابعين n- 2 grams حتى خمس عشرة كلمة n-grams 15.

5- المدى span الذي يدل على عدد الكلمات المتتابعة (أو النغرامية) قبل الكلمة المركزية nodal item أو بعدها.

6- الإحصاء المتعلق بالتكرار والتصاحب اللفظي في المدونة الحاسوبية.

وتتجه تحليلات التصاحب اللفظي في لغويات المدونة الحاسوبية إلى أربعة اتجاهات:

الأول: يكمن في عملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتصاحبة في النصوص, والتصاحب collocation يُعدُّ ارتباطًا بين وحدتين معجميتين معًا في سياق لغوي معين, وقد يخرج عن المتعارف عليه عند مزيد من الكشف عن مستويات النصوص اللغوية العربية الطبيعية في المدونات الحاسوبية.

الثاني: يتعلق بعملية تحليل المعاني المركبة بمجموعة من المدخلات المعجمية المتلازمة نحويًا في النصوص, والتلازم colligation يُعدُّ إيقاعًا تركيبيًا إلزاميًا لوحدتين معجميتين معًا في أيِّ سياق, مثل: الفعل اللازم وحروف الجر وما بعدها من أسماء.

الثالث: يرتبط بعملية تحليل معنى واحدًا مركبًا بمجموعة من المدخلات المعجمية المتجاورة في النصوص, والتجاور collostruction يُعدُّ إيقاعًا تركيبيًا ثابتًا لوحدات معجمية نظمية توليدية إدراكية تكون محل اهتمام في اللغويات الإدراكية cognitive linguistics واللغويات التاريخية historical linguistics واللغويات الاجتماعية sociolinguistics بمناهج المدونة الحاسوبية corpus approaches. ومِنْ أمثلة ذلك في العربية طبيعة المكملات أو المتممات complements في توليد بقية الجمل الاسمية والفعلية الأساسية من الأحوال والصفات والتمييز وأدوات الربط conjunctions, إضافة إلى المتلازمات النحوية التي تزيد عن أكثر من ثلاث كلمات؛ مثل: الفعل المتعدي, ولا النافية للجنس, والأفعال الناسخة, وأخوات إنَّ واسمها وخبرها, وظنَّ وأخواتها, والالتزامات المتتابعة بمزيد من التغاير النظمي, مثل: تنوع دلالات الكلمات الوظيفية مع الكلمات ذات المحتوى, كل ذلك في سياق تجاوري نظمي يُمكن من خلاله تحليل قواعد التركيب construction grammars (Goldberg 2009) وتغير هذا القواعد للغة عبر الزمن أو عبر جنس لغوي دون الآخر.

الرابع: ينساب في عملية تحليل المعاني بمجموعة من المدخلات المعجمية المتقاربة أو المتباعدة في المدى span والمتقاربة في الدلالة

(المجموعة الدلالية semantic set أو التفضيل الدلالي semantic preference), (انظر Price 2013), ويعدُّ انسجامًا متعارفًا عليه وتضامًا (التضام) في دلالة التراكيب, وكثرة التعارف عليه تكشفه اللغة الطبيعية المحوسبة في المدونة الضخمة large-scale corpus؛ مثل: تفضيل دلالة المراسم بالتشريف والنخب والأماكن الفارهة في اللغة الطبيعية.

ودراسة هذه الأنواع قد تكون على مستوى الكلمات أو مستوى النصوص من جهة, وثمَّة فرق بين Firth وSinclair في مدى المتصاحبات اللفظية, والفرق بينهما هو أنَّ الأول قد اهتم بالمدى span (عدد الكلمات المتتابعة تراتبيًا), بغض النظر عن الموضع position الذي اهتم به Sinclair حيث إنَّ الموضع قد يجعل من المتصاحب الثاني أو الثالث أكثر أهمية للدرس والتحليل من المتصاحب الأول للكلمة المركزية.

أيْ نص لأي تحليل للتصاحب اللفظي؟

يستلزم البحث في مسائل التصاحب اللفظي في لغويات المدونة الحاسوبية أنْ ينطلق الباحث من فرضياته اللغوية التي تُحدد بطبيعة الحال نوعَ المدونة اللغوية العربية وأجناس النصوص التي يُمكن لها أنْ تجيب عن تلك الأسئلة, أو أنْ ينظر إلى خصائص المدونة اللغوية العربية الحاسوبية الشبكية, وما توفره من أدوات من أجل أنْ ينطلق من تلك الخصائص التي تُكيِّف أصلًا أسئلة البحث اللغوي.

ومن المهم أنْ يجمع الباحث النصوص بنفسه بدلًا من الاعتماد على المدونات العربية الشبكية, مثل: المدونة اللغوية العربية الدولية (مكتبة الإسكندرية)[1] أو مدونة أرابيكوربس[2] arabiCorpus أو غيرهما (صالح 2015م.), إنْ كانت لدى الباحث أسئلة لغوية لا تتوافر إجاباتها من اختبار المدونة الشبكية. كما أنَّ الأخيرة قد لا توفر أدوات تحليلية دقيقة, ولا توفر المعطيات الإحصائية الخاصة بقوة التصاحب اللفظي من عدمه كحال أدوات معالجة المدونة العربية ACPTs (al-Thubaity et al 2013) وأدوات "محرك التخطيط" الشبكي Sketch Engine (Kilgarriff et al 2004, 2014)[3]؛ (انظر المبحث 3).

ولا يمكن للبحث المدوني الآلي الذي يختصر الوقت والجهد في معالجته لملايين الكلمات وإظهار نتائج البحث لتكرارات وكشافات التتابع اللفظي السياقي أنْ يكون كافيًا وحده, إذْ لا بدَّ من توظيف الحدس اللغوي العميق أولًا, ثم الآلة ثانيًا, ثم بهما معًا, شريطةَ أنْ يكون إعمالهما إزاء بعض منهجيًا ومقبولًا في المحصلات النهائية للتجربة, ويكونان معًا برهانًا للفرضية اللغوية المصوغة للتتابع اللفظي, والمدونة المحددة للاختبار. وعليه؛ فهل من الممكن أن تُجيبنا المدونات عن كل أسئلة البحث اللغوي؟ وهل للبحث اللغوي المعتمِد على المدونات شروط منهجية؟ وجواب (هل) هنا هو أنَّ كل مدونة لغوية حاسوبية يستحيل أنْ تجيبنا عن كل الأسئلة؛ لأنَّ الأسئلة هنا تتحدد بالمدونة بحسب نوعها أو غرضها أو عددها أو تصميمها (McEnery and Hardie 2012: 27). أمَّا الشروط المنهجية فأهمها معرفة الأدوات التي ستعرض في المبحث الثالث, ومعرفة الإحصاءات المهمة لدرجات قوة أو ضعف التصاحب اللفظي في المدونة الحاسوبية.

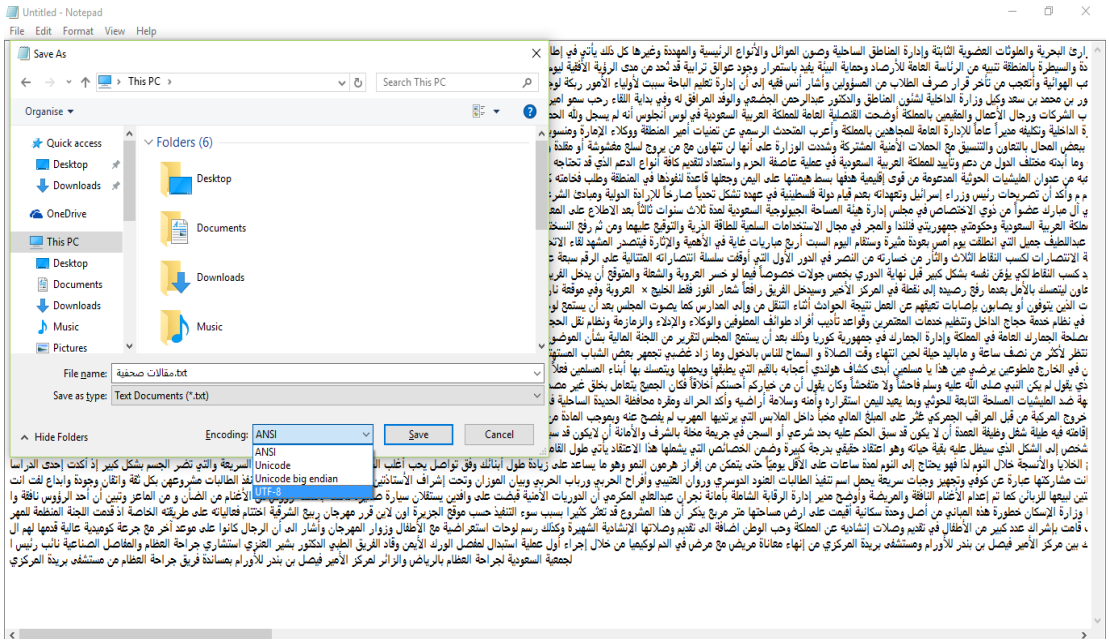
وفيما يتعلق بمدونة الباحث التي تتواءم مع ما يريد البحث عنه وما يريد أنْ يختبر بها فرضيته وأسئلته البحثية, فلو أراد الباحث -على سبيل المثال- أن يبحث عن التصاحبات اللفظية في موضوعات سياسية؛ فعليه أن يبحث عن النصوص الخاصة بها, سواءٌ كانت في مواقع

1 () انظر: http://www.bibalex.org/ica/ar/login.aspx.

2 () انظر: http://arabicorpus.byu.edu/.

(3) تُعدُّ أداة معالجة المدونة العربية ACPTs مفتوحة المصدر open source وتُعرف باسم "غوّاص" ghawwas، ويمكن تحميلها واستعمالها من موقع sourceforge من: http://sourceforge.net/projects/kacst-acptool/. وكذلك موقع محرك التخطيط Sketch Engine حيث يتيح التسجيل بالمجان مدةَ شهر واحد على هذا الرابط https://the.sketchengine.co.uk/login/. ويتيح لك استعمال ملف نصي لا يتجاوز مليون كلمة مع كامل وظائف تحليل التصاحب اللفظي، وبخاصة الأنماط النحوية لمتصاحبات النصوص العربية وأهم قياسات قوة التصاحب أو ضعفه. وحول هذه الأداة الشبكية، يمكنك الاطلاع على مزيد من وظائفها وعدد اللغات العالمية المحوسبة فيها على الرابط الآتي: http://www.sketchengine.co.uk.

الشبكة العنكبوتية أو في ملفات قابلة للنسخ والقراءة, على أنْ يتنبه إلى تلك النصوص التي تتطلب أذونات رسمية من أصحابها, فيقوم بمراسلة أصحابها لطلب إذن في استعمالها للأغراض البحثية دون التجارية, ومن بعدُ يكسب أهلية نسخها إلى ملف نصي plain text, أو أية ملفات أخرى قابلة للقراءة والتحرير الآلي, ويقوم بحفظها باسم (save as), وعند ظهور قائمة الحفظ, يختار المجلد أو الامتداد الذي سيحفظ فيه الملف, ومن ثم يختار الترميز (encoding) الخاص بالخط العربي وهو: UTF-8 (الشكل 1).



الشكل رقم (1). نسخ النص في المفكرة وحفظه باسم مع اختيار الترميز encoding (UTF-8) .

ويستلزم لأداة ACPTs أن يكون الملف النصي (أو الملفات النصية), موضوعة في مجلد folder, وعند إضافتها إلى هذه الأداة فإنَّ البحث عن الملف غير ممكن, بلْ عن المجلد, ولو قمنا بالضغط على المجلد فإنَّنا لن نجد هذه الملفات وإن كانت موضوعة فيها من قبل, وعليه فإنَّ على المحلل اللغوي أنْ يختار المجلد دون فتحه, لينقل الملف النصي (أو الملفات النصية), الموضوعة مسبقًا في المجلد, وهذا كله على عكس محرك التخطيط Sketch Engine الذي يُمكن من خلاله تحميل المدونة المحفوظة في الملفات النصية بشكل مباشر دون حاجة إلى وضعها في مجلدات.

أدوات الدراسة لمعالجة التصاحب اللفظي

اعتمد هذا البحث على برنامج أدوات معالجة المدونة العربية Arabic Corpus Processing Tools (Al-Thubaity et al. 2013) وعلى برنامج أدوات (محرك التخطيط) Sketch Engine (Kilgarriff 2004, 2010, 2014) في العرض وذكر خصائص المعالجة المتعلقة بتحليل التصاحب اللفظي آليا دون غيرها. وجدير بالذكر أنَّ البرنامج الأول المشهور باسم (غواص) لا يقبل إلَّا الملفات النصية txt* (المفكرة notepad) أو ملفات الإكسل Excel أو ما يُعرف بالقيم المفصولة بفواصل comma separated value *csv*, أمَّا البرنامج الثاني فهو يقبل العديد من الملفات ومنها ملفات PDF القابلة للقراءة والتحرير التي لا تتضمن صورًا images نصية ثابتة.

وسنتناول أهم وظائف الأدوات الأولى وبخاصة في تحليل أمثلة التصاحب من حيث قوائم تكراره, أمَّا الأدوات الثانية المتعلقة بمحرك التخطيط, فسيُقتصر البحثُ فيها عن جانب استخراج أنماط التصاحب اللفظي التركيبية أو ما

يُسمَّى بـالنحو النمطي Pattern Grammar وذلك بواسمات ومحللات ستانفورد للعربية Arabic Stanford Parsers and Taggers (حبش 2014م) المصاحبة لأدوات محرك التخطيط الشبكي. ولتوضيح معنى أمثلة التصاحب اللفظي في الأول وأمثلة أنماط التصاحب في الثاني, لو قلنا: (رغب في) و(رغب عن) فنحن هنا أمام مثالين من أمثلة التصاحب, أمَّا من جهة النمط التصاحبي فنكون أمام مثالٍ واحدٍ فقط من أنماط التصاحب؛ وهو (الفعل+ حرف الجر).

وفيما يتعلق بوظائف (غواص) الأساسية, فنُبيّن وفقًا للشكل (2), وتبعًا للترقيم السلسلي على مواضع المعالجة بشكل تراتبي, وهي على النحو الآتي:

يتضمن الرقم (1) ثلاثة وظائف, كل وظيفة تمثل واجهة جديدة من واجهات الأداة, فالوظيفة الأولى هي: (إضافة المدونة), ومن الممكن إضافة مدونة رئيسة primary corpus في موضع الرقم (2), ومدونة مرجعية reference corpus في موضع الرقم (3). أمَّا الوظيفة الثانية والثالثة فهي خيارات المعالجة (انظر الحديث عنها في مواضع الأرقام 4, و5, و6) والمقارنة (انظر الحديث عنها في الموضع رقم 8). وعودًا إلى خياريْ (إضافة المدونة الرئيسة والمرجعية) فالفرق بينهما هو أنَّ الباحث قد يريد مقارنة نص لُغوي مدوني مع نص لُغوي مدوني آخر من حيث الاختلافات بينها في تكرار الكلمات, ويُستعمل هذا المنهج عادة في محاولة الكشف عن الكلمات المميزة أو الكلمات المفتاحية في المدونة الرئيسة التي لا تظهر في المدونة المرجعية, وحريّ أنْ تكونَ المدونة المرجعية أكبر من المدونة الرئيسة من حيث الحجم وعدد الكلمات, كما أنَّ المدونة الرئيسة تتطلب في هذا المنهج أنْ تتضمن على نص من وعاء أو جنس لغوي خاص. على سبيل المثال: احتواء المدونة الرئيسة على نصوص في العلوم الإدارية واحتواء المدونة المرجعية على نصوص من أجناس متنوعة بتنوع العلوم عدا العلوم الإدارية[1].

أمَّا موضع الرقم (4) فيوفر وظيفة الاستعلام عن كلمة معينة, ومن الممكن تحديد المجلد سواء كانت المدونة الرئيسة أو المدونة الفرعية أو كليهما, ويمكن تحديد مدى span التتابع اللفظي من 1 إلى 5, كما يمكن تحديد الملفات النصية من كل مدونة من المدونتين الرئيسة والمرجعية. وتوفر أداة ACPTs ميزة البحث بواسطة المحارف البديلة wildcards, فيكتب على سبيل المثال جزءًا من كلمة ويوضع قبل هذا الجزء وبعده أو أحدهما علامة (*) ليظهر جميع الاحتمالات لبقية أجزاء الكلمة التصريفية والاشتقاقية لموضع العلامة, أمَّا العلامة (؟) فهي محرف بديل يظهر جميع الاحتمالات لحرف واحد مكمل لما قبل الكلمة أو لما بعدها أو لكليهما. على سبيل المثال: لو كتبنا (*مع*) لظهرت النتائج الآتية:

الجمعة/ الجمع/ أجمعين/ بأجمعهم/ إلخ. أمَّا لو بحثنا بالطريقة الآتية (؟مع؟) فستظهر لنا النتائج الآتية:

فجمعت/وجمعت/جمعه/تجمعت/تجمعي/إلخ.

ويوفر الرقم (5) وظائف ما قبل معالجة البحث, ومنها إزالة الحركات, وإزالة الشدة والمد, وإزالة الأرقام, وإزالة الرموز, وإزالة الحروف الأجنبية, واستبدال التاء المربوطة بالهاء المربوطة, واستبدال همزة القطع والمد بالألف. واختيار هذه الخيارات أو أحدها أو عدم اختيارها بالمرّة يؤثر في نتائج عدد التكرار (انظر الموضع رقم 7 حول الكلمة النوعية type والكلمة الفعلية token), ويكون-أيضا-مرهونًا بما يريده المحلل أو الباحث, فعلى سبيل المثال؛ لو أراد الباحث أن يكتشف عن المصطلحات الأجنبية المستعملة مع مقابلاتها العربية في النص فإنه يتعين عليه عدم استعمال خيار (إزالة الحروف الإنجليزية).

وفي موضع الرقم (6), فلو أراد الباحث أو المحلل أنْ يجري البحث على كامل المفردات في الملف النصي فيتعين عليه اختيار خاصية (كامل المفردات), وإنْ أراد أن يقصيَ جملة من الكلمات عن استخراجها فيتعين عليه اختيار خاصية (كل

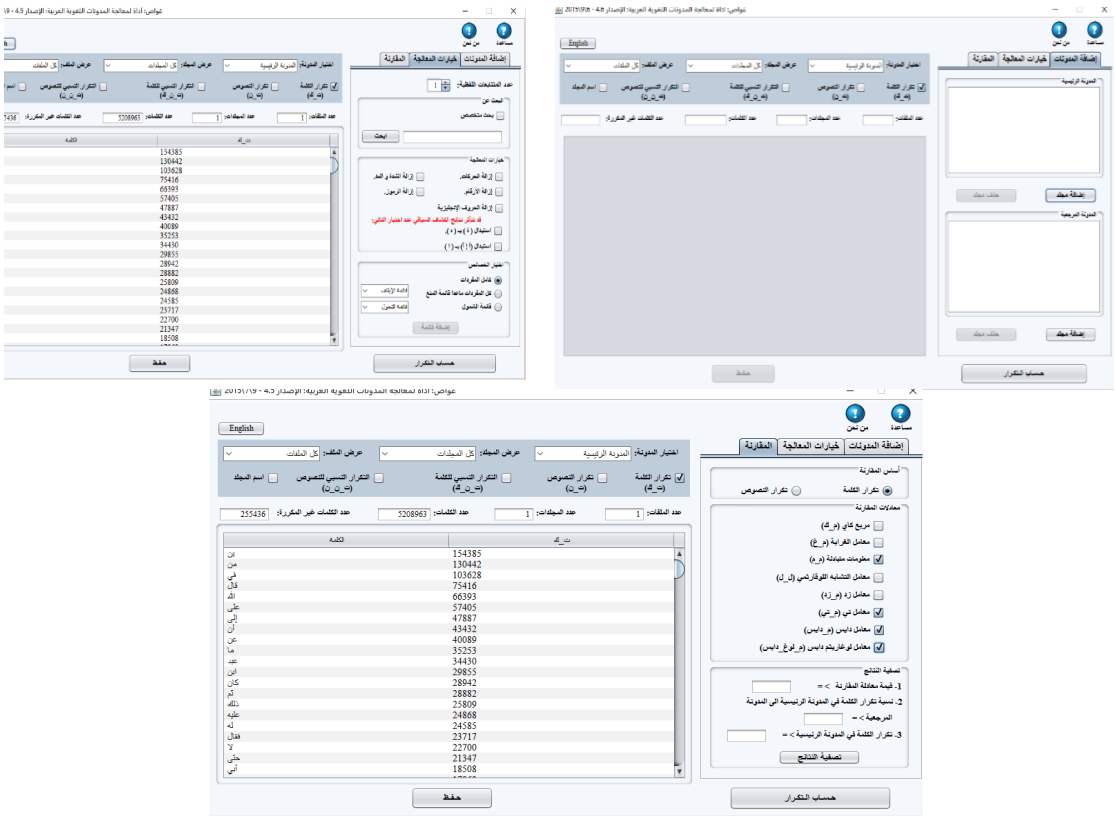
(1) توظف معامل الغرابة weirdness coefficient في لغويات المدونة الحاسوبية corpus linguistics للكشف عن الكلمات المميزة في المدونة الرئيسة والتي لم ترد، أو كان ورودها لا يُذكر، في المدونة المرجعية، وقياسها يكون بين الواحد واللانهاية infinity (انظر الشمري والثبيتي 2015م.)

المفردات ما عدا قائمة المنع), وإنْ أراد أنْ يخصص البحث عن مجموعة محددة من الكلمات فيتعين عليه اختيار خاصية (قائمة الشمول), وفي قائمة المنع أو الشمول, توضع الكلمات في ملف نصي جديد شريطة أنْ تكون كل كلمة في سطر على حدة, وأن يُحفظ الملف باسم دون اختيار الترميز UTF-8, بلْ بترك الترميز اللاتيني ANSI.

أمَّا في موضع الرقم (7) ففيه تحديد خواص عدد تكرار الكلمة أو تكرار النصوص أو التكرار النسبي أو لكليهما, وتحديد خواص إظهار اسم الملف. ويدل التكرار النسبي على حجم تكرار الكلمة أو النصوص إلى جملة الكلمات أو النصوص, وتُعبّر بالقيم فيما بين صفر و1 فعلى سبيل المثال: لو كانت نسبة التكرار النسبي للكلمة 0.01 فإنَّ ذلك يدل على أنَّ نسبة تكرار الكلمة المعنية بالبحث إلى حجم المدونة تبلغ (0.01% أو 10% قياسا على النسبة 100%). كذلك يحوي موضع هذا الرقم إفادةً بعدد تكرار الكلمات النوعية types وعدد تكرار الكلمات الفعلية tokens والأول يكون أقل بكثير من الثاني لعلة وظيفة مفهوم كل واحدة منها؛ فالكلمة النوعية هي أصل الكلمة سواء كانت الجذر أو الجذع, أمَّا الكلمات الفعلية فهي تتضمن اشتقاقات الكلمة النوعية, إضافة إلى أية مسافات أخرى في النص تحوي علامات ترقيم أو رموز أو أرقام.

وفي موضع الرقم (8), مجموعة من الحزم الإحصائية المعمول بها في التحليل المدوني الحاسوبي للمدونات, وسنتناول منها المعلومات المتبادلة Mutual Information MI وقياس_ت t_score والزَّهْرة Dice أو الزهرة اللوغارثمية LogDice لعلاقتها المباشرة بتحليل التصاحب اللفظي.

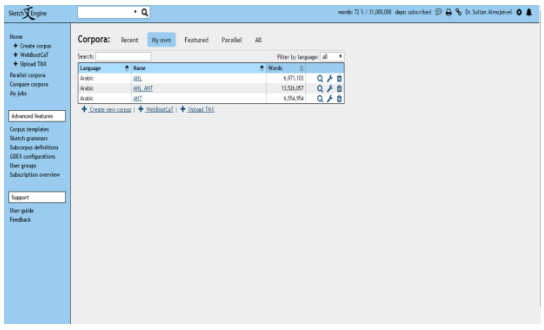
أمَّا في موضع الرقم (9) فهو محل ظهور النتائج, وتظهر النتائج تبعًا لعدد مدى التتابع اللفظي الذي يُحدّد في موضع الرقم (4) مع عدد تكرارها في المدونة.

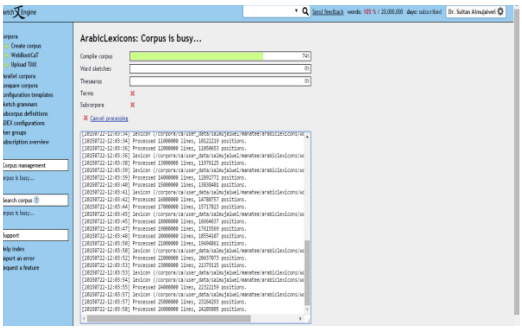


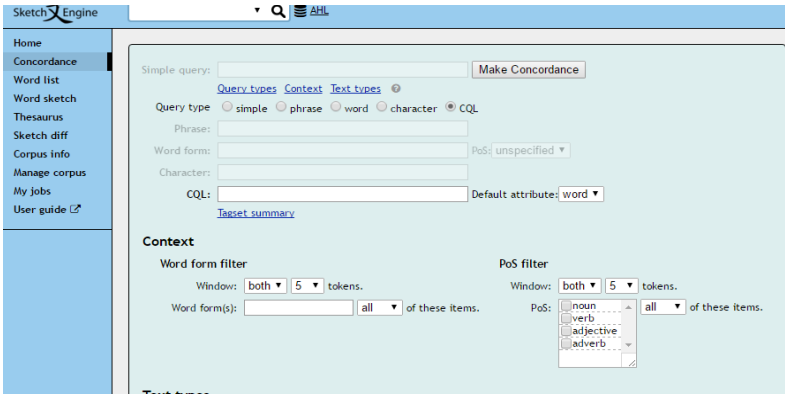
الشكل (2). واجهات خيارات (إضافة المدونة), و(وخيارات المعالجة), و(المقارنة) في برنامج ACPTs ووظائفه المتعلقة بتحليل أمثلة التصاحب.

وفي المبحث الخامس, سنجري تطبيقًا تحليليًا على معاجم اللغة العربية القديمة والحديثة لغرض تبيان طريقة هذه الأداة في تحليل مثال من التصاحب اللفظي, على عكس محرك التخطيط Sketch Engine الذي (Kilgarriff et al. 2010, 2014) سنُبين من خلاله في ذلك المبحث تطبيقًا قائمًا على النحو النمطي وإمكانية تحليله في المتصاحبات اللفظية. وفي محرك التخطيط, وبالنظر إلى واجهاته الثلاثة في الشكل (3).

وتتيح الأداة إمكانية إنشاء مدونة (Create Corpus), خاصة بالباحث (الشكل 3) ويتطلب تتبع التعليمات البسيطة لنقل الملف النصي حتى بلوغ مرحلة بناء المدونة compile corpus وفي أثناء ذلك ستظهر عملية الإنشاء كما في موضع الرقم (2), وبعد تمام العملية تظهر المدونة (أو المدونات) التي بُنِيَت كما في الموضع (1), وتتوافر للباحث خواص البحث في المدونة search corpus, والخاصية التي تتعلق بالبحث عن أنماط التصاحب هي خاصية CQL (الموضع رقم 3) التي توفر البحث التخطيطي عن الأنماط التركيبية التصاحبية (انظر Jakubíček et al 2010), وخواص واسمات ستانفورد وخواص كتابة قواعد التخطيط sketch grammars وبخواص مُشغّلات الضمن *within* and *containing* operators ومشغّلات الاتحاد *union operators* and *meet*. وتتطلب كتابة قواعد التخطيط استعمال واسمات أقسام الكلام العربية الخاصة بواسمات ومحللات ستانفورد للعربية, وهذه الواسمات تبلغ 33 واسمًا, مقسمة ما بين الكلمات الوظيفية function words, وتفاصيل أنواع الأسماء, والأفعال وأنواعها, والصفات وتفريعاتها, وعبارات العاطفة, وعلامات الترقيم, والكلمات الأجنبية التي قد ترد في النصوص العربية. وكل هذه الواسمات قد بُيّنت في الجدول رقم (1) بذكر رمز الواسم النحوي ومن ثَمَّ التعريف به بمصطلحه الإنجليزي الموضوع له, وتلا ذلك شرح المصطلح الإنجليزي بشكل موجز مع ذكر الأمثلة التي ظهرت في أول نتائج بحث بخاصية لغة استعلام المدونة في مدونة عربية معاصرة صغيرة الحجم







الشكل رقم (3). ثلاث واجهات أساسية لمحرك التخطيط Sketch Engine الشبكي ووظائفه المتعلقة بتحليل أنماط التصاحب.

الجدول رقم (1). واسمات ستانفورد.

#	الواسم	التعريف به	أمثلته في النتائج
1	FW	Foreign Words	كلمات أجنبية
2	CC	Coordinating Conjunctions	أدوات الربط (و, ف, أو, ثم, بل, لكن, أم, كما, لا, أمَّا, كيما, إلخ)
3	RB	Adverbs	الظروف (هناك, ثمَّة, هنا, إلخ)
4	WRB	Wh-Advbers	أدوات الاستفهام (كيف, لماذا, أين, متى, إلخ)
5	CD	Cardinal Numbers	الأعداد
6	DT	Demonstrative Pronouns	أسماء الإشارة (هذا, أولئك, تلك)
7	PRP	Personal Pronouns	ضمائر متصلة/منفصلة (هو, هي, نا, كم, نحن, إلخ)
8	PRP$	Possessive Personal Pronouns	ضمائر ملكية متصلة (ها, هم, نا, كم, كن, إلخ)
9	WP	Relatives	الأسماء الموصولة (التي, الذي, الذين, إلخ)
10	IN	Subordinating Conjunction	حروف الجر (ل, ب, في, من, عن, إلى, إلخ)
11	RP	Particle	أدوات (لا, قد, لم, س, هل, يا, لقد, إلخ)
12	UH	Interjections	تعابير عاطفية (نعم, اللهم, كلا, أجل)
13	PUNC	Punctuations	علامات الترقيم

14	VBG	Verbal Particles	مصدر عامل/ أداة فعل (قول, اعتبار, منح, إلخ)
15	VBD	Perfect Verbs	فعل تام (قال, صلى, سلّم, إلخ)
16	VBN	Passive Verbs	فعل مبني للمجهول (قيل, يقال, ولد, إلخ)
17	VBP	Imperfect Verbs	فعل مضارع (يكون, يسكن, يقول, يجب, إلخ)
18	VB	Infinitive Verbs	فعل أمر (انظر, قم, أضف, خذ, إلخ)
19	VN	Verbal Noun	أسماء تشبه الفعل (مشيرا, مؤكدا, مضيقا, موضحا, إلخ)
20	NN	Common Nouns	أسماء شائعة (سلام, كلام, خلال, إلخ)
21	NNS	Common Nouns (Pl.)	اسم مثنى أو جمع (سنوات, عمليات, معلومات, خدمات)

تابع الجدول رقم (1).

#	الواسم	التعريف به	أمثلته في النتائج
22	DTNN	Determined Nouns	اسم معرف/ مفرد (الناس, العمل, اليوم, إلخ)
23	DTNNS	Determined Nouns (Pl.)	اسم معرف مثنى أو جمع (المسلمين, المعلومات, الولايات, إلخ)
24	NOUN	NOUN	أسماء التوكيد (كل, بعض, جميع, نصف, إلخ)
25	NNP	Proper Noun (Sing.)	اسم علم مفرد (محمد, أحمد, علي, إلخ)
26	NNPS	Proper Noun (Pl.)	اسم علم مثنى أو جمع (طالبات, جامعات, بنايات, إلخ)
27	DTNNP	DT_Proper Noun	اسم علم معرف مفرد (الإنترنت, المغرب, القاهرة, إلخ)
28	DTNNPS	DT_Proper Noun (Pl.)	اسم علم معرف مثنى أو جمع (الإمارات, الروحانيات, البرازيليين, إلخ)
29	JJ	Adjectives	الصفات (آخر, خاصة, واحدة, كبيرة, إلخ)
30	ADJ	Adjective_Numeric	الصفات العددية (الأول, الثاني, الثالث, إلخ)
31	DTJJ	DT_Adjectives	الصفات المعرفة (العربية, العامة, السياسية, الوطنية)
32	DTJJR	DT_Comparative Adjectives	الصفات المعرفة للمقارنة (الأقل, الأوسط, الأكبر, إلخ)
33	JJR	Comparative Adjectives	صفات المقارنة (أفضل, أكبر, أقل, أعظم, إلخ)

وتُجرى خطوات البحث عن أنماط التصاحب في وظيفة CQL طبقًا لواسمات ستانفورد لأقسام الكلام لقواعد اللغة العربية على النحو الآتي:

لو أراد المحلل اللغوي أنْ يتصدى لأنماط

متصاحبات لفظية معينة, على سبيل المثال: البحث عن تطابقات نمط الواسمين (فعل تام+ اسم علم مفرد), فإنَّ واسم الفعل التام هو VBD والواسم NNP يخصُّ اسم العلم المفرد (كما هو مبين في الجدول 1). وبإضافة هذين الواسمين كوسيلة بحث عن تطابقهما في وظيفة لغة استعلام المدونة CQL في محرك التخطيط Sketch Engine اللغوي فإنَّ صيغة المشغل لها تكون على النحو الآتي[1]:

[tag="VBD.*"][tag="NNP.*"]

ومن الممكن توسيع هذه الصيغة لوضع ثلاثة أو أربعة احتمالات لواسمات نحوية يروم الباحث بها إلى كشف محدد عن أنماط التصاحب اللفظي في مدونته.

وثمَّة صيغ أخرى تساعد على بحث أكثر دقة عن طريق مشغلات الضمن والاتحاد في وظيفة لغة استعلام المدونة CQL, وتتضمن صيغتها أكواد خاصة؛ وهي على النحو الآتي:

- لربط كل أسماء الأعلام بخاصية الضمن *within* في سلسلة تركيبية تبدأ بفعل تام وتنتهي بفعل تام:

```
[tag="NNP.*"]+ within [tag="VBD.*"] []{0,5}
    [tag="VBD.*"]
```

- لربط سلسلة تركيبية تبدأ وتنتهي بفعل تام تتضمن اسم علم واحد بخاصية الضمن *containing*:

```
[tag="VBD.*"] []{0,5} [tag="VBD.*"] containing
    [tag="NNP.*"]
```

- لربط كل اسم علم محاط بفعل تام على مدى span سياقي يمتد إلى ثلاث كلمات من قبلُ ومن بعــدُ (-3/+3):

```
(meet [tag="NNP.*"] [tag="VBD.*"] -3 3)
```

- لتوسيع نتائج البحث السابق باستخراج أنماط كل الصفات (وواسمها JJ كما هو في الجدول 1) المحاطة بالفعل التام في مدى span سياقي يمتد إلى -2/+2

```
(union (meet [tag="NNP.*"] [tag="VBD.*"] -3 3)
(meet [tag="JJ.*"] [tag="VBD.*"] -2 2))
```

إحصائيات التحليل التصاحبي في أدوات الدراسة[2]

إنَّ السؤال الذي يروم إلى توضيح قراءات إحصاءات التصاحب اللفظي بالاستناد إليه يتطلب الإجابة عن فائدة أرقام هذه الإحصاءات, وكيفية قراءتها في سياق تحليل التصاحب اللفظي, وبما تفيد المحلل أو الباحث اللغوي به. ووضح كل من (Church and Hanks 1990), و(Oakes 1998: 63) المعلومات المتبادلة Mutual Information على أنَّها تفيد بالكشف عن احتمالات تكرار كلمتين تكونان متصاحبتين معًا مرة وتكرار كل واحدة منهما وحدها لقياس التصاحب اللفظي. كما يُفيد هذا النوع من الاختبار الإحصائي-من غير قياس التصاحب اللفظي- قياس مدى ارتباط كلمة في مدونة اللغة المصدر بكلمة في مدونة اللغة الهدف, وهو قياس مفيد في تحليلات نظرية المعرفة information theory في لغويات المدونة الحاسوبية corpus linguistics, وبخاصة بين مدونة لغوية للغة الأصل ومدونة لغوية للغة الهدف.

وفي إحصائيات قياس_ت t_score تحسب الكلمة المركزية (nodal item) إلى مجموع الكلمات النوعية (tokens) في المدونة, ويساعد هذا القياس

(1) انظر: حول مزيد من الشروحات عن كتابة مشغلات الضمن والاتحاد وقواعد التخطيط
https://www.sketchengine.co.uk/corpus-querying/#Usingwithinandcontainingoperators.-Grammars Sketch.

(2) الإحصائيات في لغويات المدونة الحاسوبية corpus linguistics عديدة، وقد اخترنا المعلومات المتبادلة Mutual Information وقياس_ت t-score والزَّهْرة Dice أو الزهرة اللوغارثمية LogDice لفوائدها في قياس درجات القوة والضعف التصاحبيْن بين المتصاحبات اللغوية. وثمَّة قياسات إحصائية أخرى، مثل: مربع كاي Chi-Squared الذي تقيس مدى تشتت التوزيع بين مدونتين ومدى دلالتها على كونها مقبولة لفرضيات البحث من حيث إنَّ تكرار الكلمات النوعية وتكرارات التصاحبات اللفظية تمُثل عادةً توزيعًا عشوائيًا يحمل دلالة الاقتناع بالصدفة لها التي ينتج عنها قبول الفرضية أو دلالة عدم الصدفة الذي ينتج عنه رفض الفرضية (انظر المجيول 2015م). ولمزيد من التعرف على الإحصاءات في لغويات المدونة الحاسوبية المتعلقة وبخاصة الرجعة المنطقية logistic regression التي تعدُّ أحد أهم الإحصاءات في لغويات المدونة الحاسوبية كونها تقيس التشتت والارتباط بين عدة بيانات لغوية (عدة مدونات أو عدة أوعية لغوية)، انظر Oakes (1998).

على إزالة غموض الكلمات البوليزمية (polysems) ومدى تشتت هذه الكلمات مع المتصاحبات, ومدى تعدد معاني التصاحب لكل من الكلمة المركزية البوليزمية والمتصاحب معها. على سبيل المثال: كلمة (ضرب), حيث لو نظرنا إلى متصاحباتها collocates مع الباب, والعملة, والطريق, والمثال (أي نوع المثال), إلخ. فإنَّ نتائجها في برنامج من برنامجي الدراسة ستتنوع, وكل نتيجة لهذا القياس تكون أعلى من 2.00 فإنَّها بذلك تحمل عادة دلالة إحصائية مقبولة, ولو افترضنا أنَّ هذه الأمثلة قد أظهرت لنا قياسات_ت في مدونة ما على النحو التتابعي الآتي: 11.13 و6.00 و3.00 و1.40 فإنَّ دلالتها متدرجة من حيث القوة, فأكثرها ارتباطًا هو التتابع (ضرب الباب), وأقلها ارتباطًا هو التتابع (ضربُ المثل), وتتنوع درجات القوة من عدما بتنوع المدونة ونوعية النصوص فيها, إذْ تختلف هذه القياسات بلا محالة لو قمنا بتطبيقها على مدونة عربية تتضمن نصوصًا اقتصادية, حيث قد ينعدم وجود التتابع (ضربُ الباب), و(ضربُ الطريق), و(ضربُ المثل) بينما تخرج تطابقات التتابع (ضرب العملة) بقيمة عالية لقياس_ت[1].

أما قياس الزهرة LogDice والزّهرة Dice (Kilgarriff et al. 2004) فالأول يكون قيمته ما بين الصفر و14, ويدل على أنَّ التصاحب اللفظي كلما اقتربت قيمته إلى 14 دلَّ على قوة الارتباط, وتدل القيمة. على أنَّ التوارد co-occurrence بين الكلمة المركزية nodal item والمتصاحب لها يكون بواقع مرة واحدة في كل 16 ألف مرة لتكرار الكلمة المركزية وحدها. أمَّا قياس الزهرة Dice, فقيمها تكون بين الواحد والصفر, وقوة التكرار بين التوارد تأتّى في كل قيمة تكون الأصغر, فمتصاحبٌ تكون قيمة الزهرة له مع كلمة مركزية ما على سبيل المثال (0.01) وآخر بقيمة (0.001), يكون أقل قوة من حيث الارتباط الكلي. والفرق الجوهري بين هذه القياسات لغرض التحليل التصاحبي يتمثّل في السؤال أو الفرضية البحثية التي تكون أساس الغرض من التحليل, ويمكن إيجاز ذلك على النحو الآتي:

أولًا: لو كانت قيمة المعلومات المتبادلة بين كلمة مركزية ومتصاحب لفظي ما معها في مدونة ما هي الأعلى, فإنَّ ذلك لا يعني بالضرورة أنْ يكون تكرار المتصاحب وحده هو الأعلى أيضًا, وكثيرٌ ما يكون تكرار المتصاحب اللفظي مع الكلمة المركزية (nodal item) الأقل هو الأعلى في قيمة المعلومات المتبادلة, وهذا ما سنراه في التحليل التطبيقي في المبحث الخامس.

ثانيًا: إذا كانت قيمة قياس_ت أعلى من (2) كان لذلك دلالة إحصائية, أمَّا إنْ كانت أقل من (2) فإنَّ الدلالة الإحصائية منعدمة, وقد تُؤخذ في بحث التحليل التصاحبي بعين الاعتبار في سياق الأقل استعمالًا في مدونة ما.

ثالثا: قياس الزهرة Dice والزهرة اللوغارثمية Log Dice سواء كانت الدلالةُ الإحصائية الأقوى للأول للقيم الأكثر صغرًا من الواحد, أو كانت الدلالة الإحصائية الأقوى للثاني أقرب إلى 14 من الصفر, فإنَّ الأقل قد يتميّز عن الأكثر لو كان السؤال البحثي حول التصاحب اللفظي عن ندرة الاستعمال, وقد يتميز الأكثر قيمة إنْ كان السؤال البحثي حول التصاحب اللفظي عن الأكثر تصاحبًا من حيث التكرار.

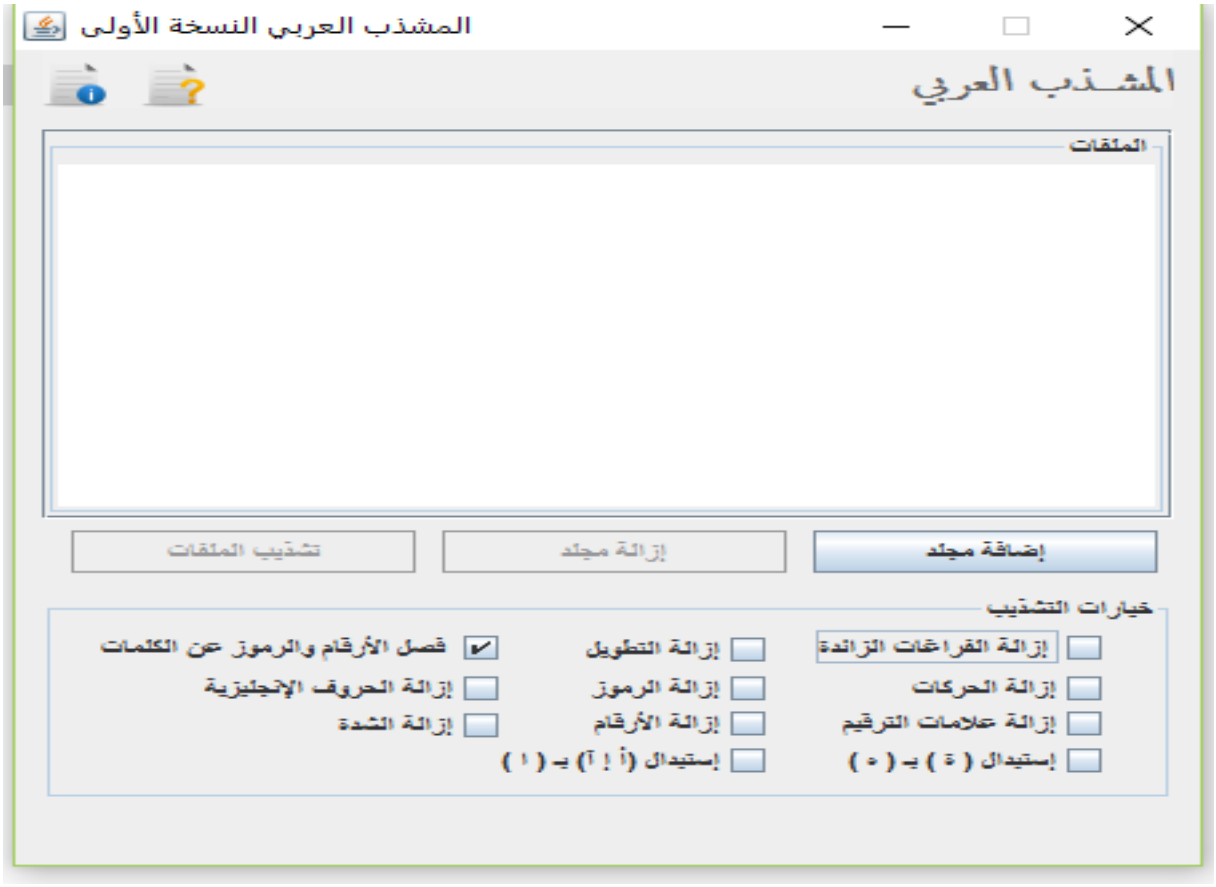
مثال تطبيقي على معاجم اللغة العربية

جُمعت المعاجم العربية البالغ عددها 22 معجمًا من موقع الشاملة[2], ومن ثَمَّ وُضعت في الملف النصي plain text, وقد شُذبت النصوص

(1) ثمَّة قياس ليس من ضمن بحثنا هنا يعرف بقياس_ز z-score الذي لا يختلف عن قياس_ت؛ فالدلالة في المعطى واحدة في كون النتيجة من 2 فأعلى دالّة؛ غير أنَّه يُفضل استعمالها على المدونات الضخمة (Oakes 1998: 7، 8).

(2) يتيح هذا الموقع تصدير وحفظ كثير من الكتب المنسوخة على ملفات الوورد word على صيغ doc* وdocx* مما يسهل من عملية نسخها وقابلية قراءتها وتحريرها. انظر: http://www.almeshkat.net/books/index.php. أمَّا هذه المعاجم فهي على النحو الآتي: العين، والبحر المحيط، والصحاح في اللغة، والعباب الزاخر، والمحكم والمحيط الأعظم، وأساس البلاغة، ولسان العرب، والمحيط في اللغة، والمخصص، ومعجم الجيم، ومقاييس اللغة، ومعجم ما استعجم، وتاج العروس، ومجمل اللغة، وجمهرة اللغة، وتهذيب اللغة، وتاج العروس، ومختار الصحاح، والمنجد، والقاموس المحيط، والمصباح المنير، ومعجم العربية المعاصرة.

بمعالج التنقيح والتنظيف, وهو برنامج يُعرف باسم (المُشذّب العربي) (الشكل رقم 4), وهو غير مفتوح المصدر. ويوفر هذا البرنامج إمكانية تشذيب النص لجعله نصًا محكمًا تتابع فيه الكلمات دون وجود أكثر من مسافة واحدة بين الكلمات ودون وجود علامات ورموز وحروف غير عربية, ودون وجود أرقام أو تطويلات تؤثر في حسابات التكرار, ويُقصد بالتطويل أو الكشيدة clitics المسافات أو المسافة الخطية التي تضاف إلى الحرف, والتي تضاف في نظام لوحة المفاتيح الخاصة بـ Windows (QWER) بمفتاح عالي Shirt ومفتاح الحرف (ت).



الشكل رقم (4). واجهة برنامج المشذب العربي ووظائف التشذيب المتوفرة فيه.

وبهذا البرنامج؛ فقد عُولجت مدونة المعاجم العربية الأربعة والعشرين؛ لأجل معالجتها في برنامجي ACPTs ومحرك التخطيط Sketch Engine, ومن أجل أنْ تكون معالجة التصاحب اللفظي إحصائيًا دقيقة, تتعامل مع سلسلة تراكيب الوحدات المعجمية النظمية syntagmatic lexical units النقية من التشكيل ومن أيّ رسم خطي غير رسم الحروف العربي في تلك الوحدات.

وحري القول إنّ ما سيُشرح من طريقة استخراج أدوات معالجة التصاحب اللفظي في هذا البحث لا يُجرى بشكل يدوي؛ لأنَّ أدوات برنامج ACPTs التي سنطبق عليها المعالجة لمثال تصاحبي محدد وأدوات محرك الاستعلام Sketch Engine التي سنُطبق عليها المعالجة لأنماط ذلك المثال التصاحبي المحدد تقوم باستخراج القيم الإحصائية للمعلومات المتبادلة Mutual Information وقياس_ت t-score والزهرة Dice والزهرة اللوغارثمية LogDice بشكل آلي. ولكن: رأيت

شرحها بشكل رياضي حتى يعطي ذلك تفسيرًا منطقيًا لكيفية عمل هذه الخوارزميات الإحصائية في قياس التصاحب اللفظي ومعالجته في مدونات اللغة العربية.

فمعادلة إحصائية المعلومات المتبادلة MI هي: log2((P(x,y) /P(x)P(y)) حيث إن P يدل على الاحتمالية probability لهذه الخوارزمية الإحصائية, أمَّا x وy فهما المتصاحبان اللذانِ يُراد اختبارهما. ولو قمنا على سبيل المثال بتحليل التصاحب اللفظي (ريح شديدة), والتصاحب اللفظي (ريح صرصر) واستخراج تكراراتها من مدونة المعاجم العربية التي أنشأناها, فإنَّ المعطيات التكرارية هي على النحو الآتي: عدد كلمات مدونة المعاجم العربية 20.432.212 كلمة, وعدد تكرار كلمة (ريح) فيها 1891 مرة, وعدد تكرار (ريح شديدة) معًا 28 مرة, وعدد تكرار (ريح صرصر) معًا 5 مرات, وعدد تكرار الصفة (شديدة) وحدها 1318 مرة, وعدد تكرار الصفة (صرصر) وحدها 53 مرة. ولقياس كل متصاحب collocate من الصفتين (أيْ: شديدة وصرصر) على حدة مع الكلمة المركزية nodal item (ريح), فإن معادلة المعلومات المتبادلة تتمثل على النحو الآتي:

معادلة المعلومات المتبادلة للتصاحب الأول

log2((28*20,432,212)/ (1891*1318))

572,101,936/2,492,338=229.54

log2 (229.54) = **7.84**

معادلة المعلومات المتبادلة للتصاحب الثاني:

log2 ((5*20,432,212)/ (1891*53))

102,161,060/100,223=1019

log2 (1019) = **9.99**

وبنتائج كل معادلة, نلحظ أنَّ ناتج المعلومات المتبادلة للتصاحب اللفظي بين الكلمة المركزية المعنية (ريح) والصفة (صرصر) هي MI (9.99) وهي أعلى من تلك الواقعة بين الكلمة المركزية وبين الصفة (شديدة) التي نتيجتها MI (7.84)؛ وقوة التصاحب هنا يُشير إلى أنَّ قوة ارتباط الصفة الأولى أعلى من حيث الاستعمال المنسجم لمفهوم التصاحب اللفظي في لغويات المدونة الحاسوبية من الصفة الثانية, وعليه, فإنَّ معالجة هذا التصاحب في معجم عربي يُراد تأليفه بمناهج لغويات المدونة الحاسوبية تُقدّم التصاحب اللفظي (ريح صرصر)-لدلالة استعماله الأقوى-على التصاحب اللفظي (ريح شديدة).

وقياس-ت t-score يتشابه مع المعلومات المتبادلة غير أنَّه يقوم بإظهار مقاييس التشتت لاحتمالات تكرارات التطابق للمادة العنقودية nodal item ومتصاحبها collocate (Scott 2010, وانظر Hunston 2001, 2002 وPrice 2013). ويتكون هذا القياس من المعادلة الآتية: ((x/n)-x)/√j حيث إنَّ n يعبر عن العدد الكلي للكلمات في المدونة, وj يعبر عن حاصل ضرب التكرار المشترك بين الكلمة المركزية nodal item ومصاحبتها؛ حيث إنَّ x يشير ببساطة إلى F1*F2 (أي: ضرب عدد تكرار الكلمة المركزية مع عدد تكرار الكلمة المتصاحبة). ويُستفاد من هذه العملية في تحليل الكلمات ذات المعاني المتعددة polysems. والناتج الآلي الذي يُستخرج بهذه المعادلة يجب أنْ يكون من 2 فما فوق من أجل ضمان قياس إحصائي ذي دلالة قوية. فلو طبقنا هذه المعادلة على التصاحب اللفظي (ريح صرصر), فإن معادلة تتابع قيم المعادلة بهذا القياس ستكون كما هو آتٍ:

((1891*53/20,432,212)-1891*53)/√j

((100,223/20,432,212)-100,223)/√j

قياس_ت (أقل من الواحد)

وقيمة هذا التصاحب ليست أعلى من القيمة 2.00 وإذا كانت القيمة أقل من واحد فإنَّما هو بسبب أنَّ متوسط التكرار بين ريح (1891 مرة), وصرصر (53 مرة), هو 972, والانحراف المعياري من المتوسط 42.871, ونتيجة قسمة ما هو أقل من الواحد على هذا الانحراف المعياري هي (0). وعليه فإنَّ إطار تحليل الكلمة المركزية (ريح) بوصفها كلمة متعددة المعنى polysemic بقياس_ت هو الأنسب, وستتضمن معالجة التحليل التصاحبي بالنظر إلى عدد التكرار ونسبه مع معطى قياس_ت له؛ لِتُخبر عن اطّراد التصاحب اللفظي للكلمة المركزية مع المتصاحبات الأخرى التي تُكوّن معنىً سياقيًا محددًا بدلالة خاصة مع كلمة (ريح), وفي الجدول (2) نتائج هذا الاطراد مع عدد تكراره ونتائج معادلة قياس_ت لكل تصاحب لفظي. وقد جُعل قياس المتصاحبات مع كلمة (ريح) ذات المعاني الدلالية المختلفة مرتبة وفقًا للأكثر ارتباط, ورتبت من أعلى قيمة

لقياس_ت إلى أقلها, كما أنَّ متصاحبات هذه الكلمة للمعنى الدلالي الواحد متفاوتة في القياس, ولكنَّها جُمعت في قياس واحد لتحديد قيمة المعنى الخاص للكلمة المركزية. ونُضيف هنا إلى أنَّ ندرة المعنى المكتسب, الذي عادةً ما ينحاز إلى الاستعمالات المجازية, تكون قيمها هي الأقل في المدونات العامة, وبالأخص في مدونة المعاجم العربية المفحوصة هنا لغرض هذا البحث كونها معاجمًا عامة.

أمَّا القياس الإحصائي بين الكلمة المركزية ومتصاحبها بالزهرة Dice والزهرة اللوغارثمية LogDice (Kilgarriff et al 2004) فأساس معادلة الأول هو $\frac{2fAB}{fA+fB}$ حيث إنَّ fAB يدل على تكرار الكلمة المركزية مع المتصاحب المعني بالتحليل, و$fA+fB$ يدل على حاصل جمع تكرار الكلمة المركزية وحدها مع تكرار المتصاحب المعني بالتحليل. أمَّا أساس معادلة الثاني فهو $14+\log 2\frac{2fAB}{fA+fB}$.

الجدول رقم (2). قياس_ت t-score وتوزيعات المعاني المتعددة للكلمة المركزية (ريح) ومتصاحباتها.

ريح+التصاحب	ت_تصاحب	قياس_ت	المعنى
ريح+(صفة الهواء[خجوج]/[مريضة]/[عقيم]/[عرية]/[زعزوع] خجوج	57	2.8	ريح خجوج: دائمة الهبوب والالتواء/ومريضة: ضعيفة الهبوب/وعقيم: ا تجلب مطرا ولا تنفع أرضا/عرية: باردة/ زعزوع: شديدة
ريح+(حاسة الشم والتذوق [قطنة]/[ذفرة]/[الخزامى])	15	1.7	قطنة: الشواء/ الخزامي: نبات/ ذفرة: نبات مر أو طعم لبن مرّ
ريح+(نوع المرض[الحدب]/[الماء])	19	1.2	الحدب: يصيب فقرات الظهر أو قرحة تتكون داخل العنق/ الماء: سبب للإغماء
ريح+(صوت[سهيج])	3	0.45	ريح سهيج: صوت النأج والتضرع منه أو الصياح
ريح الموت	5	0.26	ريح الموت: دنو الأجل

الجدول رقم(3). تتابعات إجراء حساب الزهرة والزهرة اللوغارثمية للكلمة المركزية (ريح) ومتصاحبيْها (شديدة) و(ريح).

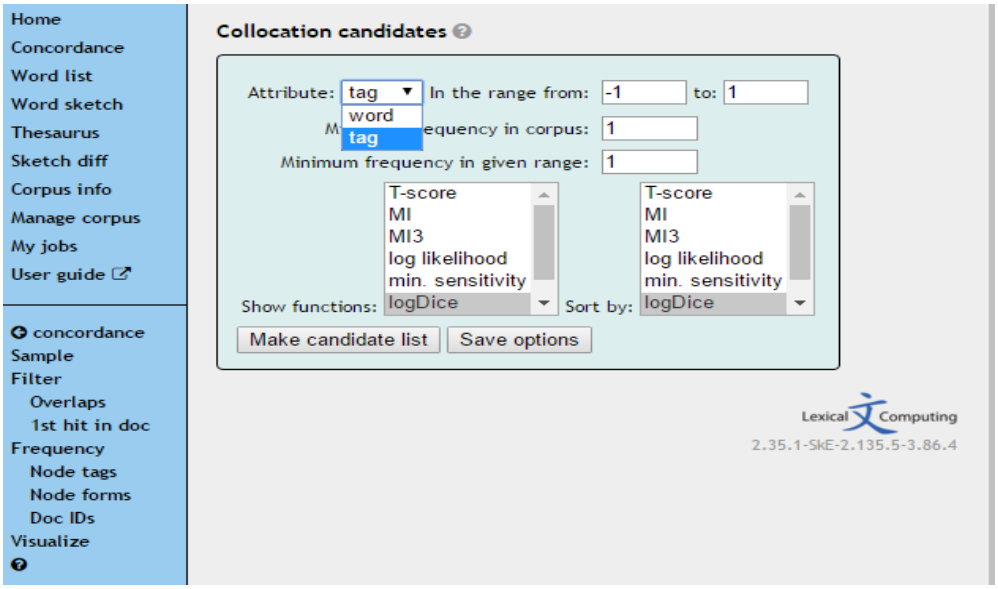
الزهرة	الزهرة اللوغارثمية
[ريح شديدة] 28 مرة/ [ريح] 1891 مرة/ [شديدة] 1318 مرة	
Dice $\frac{2fAB}{fA+fB}$	LogDice $14+\log 2\frac{2fAB}{fA+fB}$
$\frac{28}{1891+1318}$	$14+\log 2\frac{28}{1891+1318}$
ناتج الزّهرة: 0.0087	ناتج الزّهرة اللوغارثمية: 7.1216
[ريح صرصر] 5 مرات/ [ريح] 1891 مرة [صرصر] 53 مرة	
Dice $\frac{2fAB}{fA+fB}$	LogDice $14+\log 2\frac{2fAB}{fA+fB}$

$\frac{5}{1891+53}$	$14+\log_2\frac{5}{1891+53}$
ناتج الزّهرة: 0.0026	ناتج الزّهرة اللوغارثمية: 5.4127

وفي الجدول رقم (3) حساب تكرارات التصاحب اللفظي (ريح شديدة), والتصاحب اللفظي (ريح صرصر), وفقًا لهاتين المعادلتين. ففي ناتج حساب الزهرة Dice للتصاحب اللفظي الأول نجده 0.0087 أمَّا ناتج التصاحب اللفظي الثاني فهو 0.0026, وقيمة الثاني أصغر من قيمة الأول, وهي دلالة على أنَّ ارتباط التصاحب اللفظي الثاني أقوى, كما هو حال ناتجه في المعلومات المتبادلة, أمَّا قيمة التصاحب اللفظي الأول بالزهرة اللوغارثمية فهي 7.1216 والتصاحب اللفظي الثاني هي 5.4127, وقيمة الأول أقرب إلى القيمة 14, ويُؤخذ هذا القياس في قياس التصاحب اللفظي على نوعية فرضية البحث وسؤاله في المدونة, فإنْ كان السؤال عن الأكثر تميزًا يُؤخَذ بالأقل قيمة, وإنْ كان السؤال عن الأكثر تصاحبًا يُؤخذ بالقيمة.

وبقياس أنماط التصاحب لهذه الكلمات في محرك الاستعلام (اللغوي) Sketch Engine وذلك بتوظيف واسمات التخطيط القواعدي الخاصة باللغة العربية (الجدول 1), فلو قسنا نمط مدى ارتباط الأسماء الشائعة مع الصفات مبتدئة بحرف جر, فإن خاصية الضمن within تكون هي الأنسب, وعليه تكون صيغة لغة استعلام المدونة على النحو الآتي: corpus query language CQL
[tag="IN.*"]+ within [tag="NN.*"] []{0,5} [tag="JJ.*"]

وفي نتائج البحث سيظهر عدد تطابق هذا النمط التصاحبي بتكرار يبلغ 140.235 مرة في مدونة معاجم اللغة العربية مع أمثلته في كشافات سياقية concordances, وبقياس التصاحب اللفظي لهذا النمط مع أنماط أخرى ومدى تشتتها بقياس الزهرة اللوغارثمية, فإنَّه يتعين علينا اختيار وظيفة collocations وتحديد مرشّحات التصاحب اللفظي collocation candidates بخيار (الواسم) tag من وظيفة الخاصيّة Attribute وجعل المدى من -1 إلى +1 واختيار قياس LogDice (الزهرة اللوغارثمية) من أجل إظهار قياس أكثر المرشحات على هذا المدى لنتائج النمط التصاحبي لصيغة استعلام المدونة المذكورة أعلاه.



الشكل رقم (5). تحديد مدى مرشحات أنماط المتصاحبات للسلسلة النمطية التصاحبية في مدونة

المعاجم العربية :

[tag="IN.*"]+ within [tag="NN.*"] []{0,5} [tag="JJ.* "]

الجدول رقم (4). بقية واسمات التخطيط وتكرارها ومدى قوة ارتباطها بالنمط التصاحبي على امتداد -1/+1 بقياس الزهرة اللوغارثمية

[tag="IN.*"]+ within [tag="NN.*"] []{0,5} [tag="JJ.* "].

الواسم	التكرار	الزهرة اللوغارثمية	الواسم	التكرار	الزهرة اللوغارثمية	الواسم	التكرار	الزهرة اللوغارثمية
أنماط الكلمات الوظيفية			أنماط الأسماء			أنماط الأفعال		
PRP	21.734	9.016	DTNN	21.483	7.989	VBD	8.941	7.118
JJR	5.743	10.011	NNP	2.802	6.164	VBP	3.641	6.883
DT	2.065	7.639	NNP	2.802	6.164	VBN	791	5.987
WP	1.142	6.623	NNP	2.802	6.164	VBG	445	6.217
RP	962	5.606	NNS	1.730	7.911	VB	48	3.116
CC	554	2.655	DTNNP	1.186	6.820	VN	31	2.620
CD	209	4.582	DTNNS	698	6.723			
WRB	67	3.840	DTJJ	205	3.797			
RB	26	2.069						

وفي الجدول (4) نتائج مرشحات أنماط التصاحب للنمط التصاحبي المُختبر, وحري أنْ تُحدَّد النتائج وفقًا للمعطيات الإحصائية لها, فعلى مستوى الأسماء, نجد أنَّ DTNN (اسم مفرد معرف بأل), أكثر أنواع الأسماء تصاحبًا مع النمط, وأقلها كان DTNNS (اسم مثنى أو جمع معرف بأل), وعلى مستوى الأفعال جاء الفعل الماضي VBD الأكثر تصاحبًا, بخلاف فعل الأمر VB والمصدر VN, ويُقاس على ذلك مع بقية واسمات المتصاحب التي تضامّت مع نتائج سلسلة النمط التصاحبي المُختبر (حرف جر+اسم شائع أو ظرف+صفة). أمَّا من حيث الكلمات الوظيفية, فالضمائر المتصلة والمتصلة PRP جاءت ضمن أكثر المرشحات, أمَّا أقلها في الظروف RP. وهذا الطريقة من البحث تساعد على تطبيق نظرية الأنماط التصاحبية لفيرث (Firth 1957) كما أنَّها توفر فهما أكبر لطبيعة النحو التركيبي construction grammar لهذا الأنماط. والبحث عن هذه الأنماط بهذه الخاصية يوفر إمكانية أمثلتها باختيار الكلمة word بديلا عن الواسم tag في وظيفة الخاصيّة attribute الموضحة في الشكل (5).

ويتراوح ما توفره خاصية لغة الاستعلام CQL من نسب تطابق البحث عن الأنماط بواسمات التخطيط القواعدي للغة العربية في أيّة مدونة لغوية عربية ما بين 90 و95% (Green and Manning 2010) كما أنَّ وظائف البحث بالضمن within وغيرها المذكورة في نهاية المبحث الثالث والمتعلقة بمحرك الاستعلام Sketch Engine تتنوع في نتائجها بالطريقة التي يروم الباحث اللغوي إلى تحديدها وفقًا لأسئلته اللغوية وفرضياته التمحيصيّة لتحليل التصاحب اللفظي.

الخاتمـــة

هدف هذا البحث إلى دراسة مفهوم التصاحب اللفظي collocation في الحقل اللغوي التطبيقي في لغويات المدونة الحاسوبية corpus linguistics الذي تطورت أدبياته منذ منتصف القرن العشرين حتى وقت كتابة هذا البحث, وقد نلنا من هذا التطور ما يسع لكيفيات بحثية وطرائق تحليلية تمكّن من فهم أعمق وأدق لطبيعة التصاحب اللفظي وعلاقة قوة ارتباطه في مدونة معينة. وقوة التصاحب مرهونة بطبيعة النص المدوني المبحوث فيه, وليس بطبيعة التصاحب اللفظي بين الكلمتين بشكل مطلق, كما أنَّ هذه الدراسة قد طبقت هذه الآليات على مدونة المعاجم العربية, وعلى أمثلة تصاحبية لأجل الكشف عن الوسائل الآلية المتاحة في عرضها ونقدها وتحليلها نوعيًا وكميًا.

ففي منهج تحليل أمثلة التصاحب بأداة ACPTs

تبيّن أنَّ قياس التصاحب بين المعلومات المتبادلة وقياس ت يختلفان في أنَّ الأول يقيس مدى القوة والارتباط بغض النظر عن شيوع تكرار أحد ركني التصاحب, أمَّا القياس الثاني فهو يعتمد على قياس التنوع الدلالي للكلمة المركزية بوصفها كلمة بوليزمية بفعل تتابعها مع الكلمة المصاحبة سياقيًا. وفي قياس الزهرة اللوغارثمية وجدنا أنَّ القيمة بين الصفر و14 متعلقة بالكثرة والقلة من حيث توزع ركني التصاحب في مدونة المعاجم العربية القديمة والحديثة, ويُستفاد من هذا القياس للأكثر في الشيوع وللأقل في الندرة الاستعمالية. أمَّا في تحليل الأنماط, فإنَّ محرك الاستعلام Sketch Engine وبخاصية واسمات التخطيط القواعدي فإنَّ تنوع أنماط التصاحب ممكن كشفه وكشف أمثلته بوظائف محددة في لغة استعلام المدونة, وهي وظائف يحددها الباحث لصيغة سلسلية تركيبية تجمع بين ركني التصاحب اللفظي أو أركان التصاحب اللفظي الممتد إلى 5 كلمات متتابعة.

شكر وتقدير:

يشكر الباحث مركز بحوث كلية الآداب بجامعة الملك سعود على دعم مشروع هذا البحث.

المراجع العربية

البركاوي, عبدالفتاح. *دلالة السياق وعلم اللغة الحديث*. دار المدار, القاهرة, 1991م.

حبش, نزار. *مقدمة في المعالجة الطبيعية للغة العربية*, ترجمة: هند بنت سليمان الخليفة. دار جامعة الملك سعود للنشر, الرياض, 2014م.

حسان, تمام.*اللغة العربية معناها ومبناها*. الهيئة المصرية العامة للكتاب, ط2, 1997م.

عبدالعزيز, محمد. *المصاحبة في التعبير اللغوي*, دار الفكر العربي, القاهرة, 1990م.

عمر, عبدالرزاق. *المتلازمات اللفظية في اللغة والقواميس العربية*, مجمع الأطرش, تونس, 2007م.

محمد, جودة مبروك. *ظاهرة التلازم التركيبي*: دراسة في منهجية التفكير النحوي. مجلة مجمع اللغة العربية الأردني, المجلد (15), العدد (31), 2011 م. الصفحات 111-146.

صالح, محمود إسماعيل. *المدونات اللغوية وكيفية الإفادة منها*. في: المدونات اللغوية العربية بناؤها وطرق الإفادة منها, تحرير: صالح العصيمي, مركز الملك عبدالله الدولي لخدمة اللغة العربية, الرياض, 2015م, الصفحات 17-93.

المجيول, سلطان.*مناهج التهيئة المعجمية في تعليم العربية لغير الناطقين بها*, الأعمال الكاملة للمؤتمر الدولي الثاني (اتجاهات حديثة في تعليم العربية لغة ثانية), دار جامعة الملك سعود للنشر, الرياض, 2016م, الصفحات 601-633.

المجيول, سلطان.*البحث اللغوي في المدونات العربية الحاسوبية بين الممكن والمحتمل والمأمول*, في: المدونات اللغوية العربية بناؤها وطرق الإفادة منها, تحرير: صالح العصيمي, مركز الملك عبدالله الدولي لخدمة اللغة العربية, الرياض, 2015م, الصفحات 235-279.

المراجع الأجنبية:

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.

Green,*Spence, and Manning*, Christopher, D. Better Arabic Parsing: Baselines, Evaluations and Analysis. In: COLING 10 Proceeding of the 23rd International Conference on Computational Linguistics, 2010, pp.394-402.

Gries, S.*Th. Collostructions: investigating the*

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

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

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

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

Price,T. 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.

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.

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

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

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

Stefanowitsch, A. and Gries,*St. Th. 'Collostructions: investigating the interaction between words and constructions'*,International Journal of Corpus Linguistics 8(2), 2003, pp. 209–43.

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

interaction between words and constructions. International Journal of Corpus Linguistics, 8(2), 2003, pp. 209-243.

Gries, S. *Th. Data in Construction Grammar*. In: Graham Trousdale& Thomas Hoffmann (eds.), The Oxford Handbook of Construction Grammar, pp. 93-108. Oxford: Oxford University Press, 2013.

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

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

Hunston,*Susan. 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.

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

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

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