نحو معجم حاسوبي للمتلازمات اللفظية في اللغة العربية المعاصرة

أيمن الغامدي[١] و إريك آتويل[٢]

كلية الحاسب، جامعة ليدز بالمملكة المتحدة

[١]**scaaa@leeds.ac.uk**

[٢]**e.s.atwell@leeds.ac.uk**

ملخص البحث

تهدف الدراسة إلى بناء قائمة للمتلازمات اللفظية كخطوة أولى لمشروع أكبر يهدف لبناء معجم حاسوبي شامل لهذه العبارات في اللغة العربية المعاصرة مبني على عدد من أكبر المدونات العربية المتاحة، وذلك ليكون مصدراً لغويا مفيداً في مجال المعالجة الحاسوبية للغة العربية، وكذلك في تعلم وتعليم اللغة العربية وخاصة لغير الناطقين بها. في هذه الورقة تقرير عن تجربة حاسوبية تم عملها لاستخراج عدد من المتلازمات اللفظية، والتي يمكن تصنيفها إلى عدة فئات بناء على المستويات اللغوية المعروفة صوتياً وصرفيا ونحويا ودلاليا. وفي التجربة الحالية استخرجت قائمة تتكون من أكثر ٦٠٠ عبارة عن طريق تطبيق نموذج حاسوبي لمعالجة المدونات العربية واستخراج هذه العبارات بطريقة شبه آلية ويتكون النموذج من ثلاث مراحل رئيسة مصحوبة بعدد من المهام الفرعية، وقد اعتمد الباحثان في هذه التجربة على الجمع بين المنهج الحاسوبي الآلي والمنهج الوصفي المبني على التدقيق اليدوي لضمان الجودة النهائية لقائمة المتلازمات اللفظية.

١ـ مقدمة

تعتبر ظاهرة المتلازمات اللفظية في اللغات الإنسانية من الظواهر التي لفتت نظر الباحثين في مختلف التخصصات التي لها علاقة بدراسة اللغة، فنجد كثيراً من الأبحاث على سبيل المثال في تخصصات: (اللسانيات، علم النفس اللغوي، تعليم تعلم اللغات ، معالجة اللغات الطبيعية ،الذكاء الاصطناعي... وغيرها) تناولت ظاهرة المتلازمات اللفظية من مختلف الزوايا العلمية، وذلك لأهميتها ودورها الحيوي في الوصول إلى فهم حقيقي شامل للغات البشرية والقدرة على تطويع الآلة (أو الحاسب) وتزويدها بمعرفة أوسع لهذه الظاهرة حتى تتعزز قدرتها على تحليل وفهم اللغات البشرية.

على سبيل المثال، في علوم اللسانيات واللسانيات التطبيقية خاصة نجد كثيرا من الأبحاث التي تخصصت في فهم وتحليل هذه الظاهرة وشرح دورها المحوري في عملية فهم اللغة وخاصة في ما يتعلق بالمعالجة الحاسوبية للغة، وكذلك تعلم وتعليم اللغة للناطقين بغيرها، فقد أكدت عدد من الأبحاث اللسانية على أن المعجم العقلي الذي يستعمله الأنسان العادي في لغة التواصل اليومية يتكون من مجموعة من المتلازمات اللفظية، وذلك بدلا من المفهوم السائد قديما والذي يدعي أصحابه أن المعرفة اللغوية ممثلة بكلمات مفردة منعزلة في العقل البشري (**Pawley & Syder, 1983; Sinclair, 1987; Kjellmer, 1990; Wray, 2002; Nesselhauf, 2000**). وبناء على هذا المفهوم ظهرت كثير من الأبحاث التطبيقية التي تحاول إثبات هذه النظرية بتطوير عدد من القوائم للمتلازمات اللفظية القائمة على الشيوع لاستعمالها كأداة تعليمية في مجال تعلم اللغات الأجنبية وكذلك المعالجة الحاسوبية للغات لتزويد ذهن متعلم اللغة الأجنبية والذاكرة الحاسوبية بهذه المعرفة، والتي يتعلمها الناطق الأصلي بطريقة طبيعية غير مباشرة في الغالب من خلال ممارسته واستعماله للغة منذ الولادة. ففي اللغة الإنجليزية على سبيل الامثال نجد عدداً من القوائم المبنية على الشيوع والتي صارت مصدراً معتمداً في تأليف مناهج تعليم اللغة واختبارات تحديد المستوى اللغوي وما إلى ذلك.

أما فيما يتعلق باللسانيات الحاسوبية فكثيرا من المهام الأساسية لتحليل وفهم اللغة حاسوبياً تعتبر ناقصة وغير مكتملة بدون إضافة قوائم أو معاجم حاسوبية للمتلازمات اللفظية، وتشمل هذه المهام كل مستويات التحليل اللغوي من المرحلة الصوتية والصرفية إلى مرحلة التحليل السيميائي المعنوي للغة. ومن هنا فقد لعبت هذه المعاجم والقوائم دوراً أساسيا في تحسين جودة النتيجة النهائية لكثير من التطبيقات في اللسانيات الحاسوبية، على سبيل المثال (الترجمة الآلية، الاستخراج الآلي للمعرفة اللغوية، التحليل الآلي لمستويات اللغة، التعرف الآلي على الأنماط اللغوية المختلفة) وإذا ما أمعنا النظر في الأبحاث في هذا المجال نجدها مطبقة على اللغة الإنجليزية، وذلك لكثرة المصادر اللغوية عالية الجودة المتاحة للباحثين، وكذلك توفر الأدوات الحاسوبية الحديثة لمعالجة وتحليل اللغة في مختلف مستوياتها، من جهة أخرى لاتزال اللغة العربية بحاجة ملحة إلى الكثير من البحث في هذا المجال، وخاصة فيما يتعلق بالأبحاث التي تركز على استعمال المتلازمات اللفظية في اللغة المعاصرة والتي تعددت تراكيبها ودخلت فيها الكثير من الكلمات الجديدة من لغات أخرى أو من اللغة نفسها وتغير معناها الدلالي بشكل كبير، وكذلك الأبحاث التي تركز على تطوير قوائم ومعاجم حاسوبية حديثة للغة العربية مبنية على الشيوع في المدونات العربية الضخمة المطورة حديثاً والتي تضم بلايين لكلمات العربية المستعملة في اللغة العربية المعاصرة. ومن هنا جاء هذا البحث ليسهم في سد هذا العجز ويساعد على ردم هذه الجفوة في المعرفة في اللسانيات الحاسوبية العربية.

وسيقسم البحث الحالي على النحو التالي، أولاً سنحاول تقديم تعريف مبسط عملي للمتلازمات اللفظية المعنية في التجربة المقدمة في هذه الدراسة، مع إعطاء تصور مختصر عن الخلاف الموجود في الدراسات السابقة فيما يتعلق بتعريف المتلازمات اللفظية، وكذلك اختيار المصطلح الملائم لهذه الظاهرة. ثانياً: سيقدم البحث تقريرا مختصرا عن تجربة عملية لاستخراج عدد من المتلازمات اللفظية في اللغة العربية معتمدة على تقنية الاستخراج الآلي للمتتابعات اللفظية من كلمتين إلى ست كلمات (**n-gram_tool**) باستخدام نموذج محوسب مكون من ثلاث مراحل رئيسة، وسيختم البحث بمناقشة أهم النتائج الحالية وإعطاء تصور عن التجارب المستقبلية المحتملة المتعلقة بهذه الدراسة.

١,١ مفهوم المتلازمات اللفظية:

من خلال نظرة سريعة للدراسات السابقة في هذا المجال نكاد نجزم على عدم وجود إجماع بين اللسانيين والحاسوبيين على تعريف مشترك لظاهرة المتلازمات اللفظية على الرغم من إدراكهم لأهميتها ودورها المحوري في تشكيل وفهم اللغات البشرية فقد اقترحت مجموعة كثيرة من التعاريف كما في الدراسات السابقة التالية **(Baldwin, 2004, Calzolari et al., 2002, Guenthner) & (Blanco, 2004).** ومثل هذا الاختلاف يكون منطقيا إذا ما إدركنا مدى تعقيد هذه الظاهرة اللغوية وتداخلها القوي مع كل مستويات التحليل اللساني (الصوتي الصرفي والنحوي والدلالي) على سبيل المثال هناك عدد من الباحثين في اللسانيات الحاسوبية يشيرون بهذا المصطلح إلى مجموعة متشابهة من الظواهر اللغوية مثل المتلازمات الاسمية والفعلية والحرفية والحكم والأمثال المتداولة **(Sag et al., 2002; Gralinski et al., 2010)** بينما يقتصر باحثون آخرون في إطلاق هذا المصطلح على الألفاظ المتصاحبة في السياقات النصية المختلفة بناء على التحليل الإحصائي لمدونات اللغة.

فنجد في معظم الدراسات السابقة أن كل باحث حاول أن يعرف هذه الظاهرة اللسانية من الزاوية التي يوليها العناية والأهتمام في بحثه. ومن هذا الباب ولضيق المجال في الورقة الحالية عن الوصول إلى تعريف شامل وكامل لهذه الظاهرة، اعتمد الباحث على واحد من أكثر التعاريف انتشارا وأقربها دلالة على كل أنواع المتلازمات اللفظية المقصودة في البحث الحالي وهو تعريف اللسانية الانجليزية **(Wray, 2002)** التي عرفت في كتابها المطول عن ظاهرة المتلازمات اللفظية المتلازمات اللفظية في اللغات البشرية بأنها الكلمات المتتابعة بشكل متصل أو مع وجود فاصل بينها والتي يظهر أنها تعالج وتحفظ في الذاكرة كجملة واحدة مسبقة التركيب خلافاً للتراكيب اللغوية الأخرى التي يمكن فهمها وتحليلها بسهولة عن طريق قواعد اللغة وبواسطة فهم الكلمات المفردة المكونة لتركيب أو الجملة" . فكل ما يندرج تحت هذا المفهوم من المتلازمات اللفظية هو محل عنايتنا في هذا البحث.

أما فيما يتعلق بالمصطلحات المستعملة في الدلالة على هذه الظاهرة فقد تعددت وكثرت ولا يوجد مصطلح واحد متفق عليه لوصف هذه الظاهرة ويمكن أن يفسر هذا التعدد والاختلاف بنفس التفسير السابق الذي ذكرناه بخصوص تعدد واختلاف التعريفات التي اقترحت لهذه الظاهرة، ففي اللغة الإنجليزية على سبيل المثال ذكرت **Wray (2002)** أنها عثرت على أكثر من خمسين مصطلحاً كلها تشير تقريباً إلى نفس الظاهرة واقترح اللساني **Schmitt (2010)** كحل لهذه المشكلة المصطلحية اختيار مصطلح **Formulaic Sequence** ليكون مظلة تدخل تحتها كل أنواع المتلازمات اللفظية في اللغات وفي اللغة العربية كذلك نجد نفس الاختلاف حول المصطلح الأمثل لو هذه الظاهرة فقد أطلقت عدة مصطلحات في الدراسات السابقة مثل(المتتابعات اللفظية، المتصاحبات اللغوية، الحكم، الأمثال، الشواهد) وكخطوة عملية ضرورية اختار الباحث استعمال مصطلح لمتلازمات اللفظية هنا ليدل على كل الأنماط والتراكيب اللغوية المقصودة بالبحث في هذه الورقة.

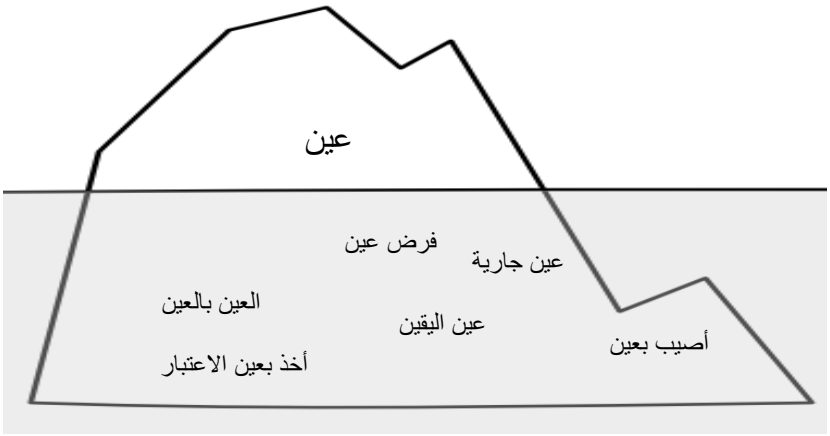
٢,١ أهداف الدراسة ودوافعها:

تهدف الدراسة الحالية إلى بناء معجم حاسوبي للمتلازمات اللفظية في اللغة العربية المعاصرة تحقيقا لغرضين أساسين هما:

- تطوير معجم محوسب للمتلازمات اللفظية مبني على عدد من أكبر المدونات العربية المتاحة للعربية المعاصرة يمكن الاستفادة منه في تحسين جودة المخرجات النهائية لأدوات المعالجة الطبيعية للغة العربية في كل المستويات اللسانية وتطبيقات معالجة اللغة المتنوعة(الترجمة الآلية، التعرف التلقائي للأنماط اللغوية، البحث الدلالي في النصوص، وغيرها).

- أداة تعليمية ومرجع في تعلم وتعليم اللغة العربية للناطقين بغيرها حيث يمكن الاعتماد عليها في تطوير مناهج تعليم اللغة والاختبارات المحوسبة لتحديد المستوى اللغوي بالإضافة إلى الأدوات الحاسوبية التي تساعد على تحديد مستوى صعوبة النصوص وملاءمتها للمستويات اللغوية المختلفة.

إن إهمال هذا النوع من التراكيب اللغوية في أي مهمة جادة للمعالجة الحاسوبية للغة العربية أو في مجال تعلم وتعليم اللغة العربية يؤدي إلى تدهور كبير في جودة المخرجات النهائية لهذه المهام كما دلت على ذلك كثير من الدراسات السانية والحاسوبية الحديثة **(e.g., Sinclair, 1987; Martinez & Murphy, 2011)** والكلمات المفردة التي يخصص لها النصيب الأكبر في الدراسات المعجمية واللسانيات الحاسوبية ماهي إلا جزء صغير من مجموعة كبيرة من التراكيب والمتلازمات اللفظية المعقدة والتي لا يمكن الوصول لفهم حقيقي للغة إلا بفهم شامل لمعانيها المتعددة في السياقات المختلفة فعلى سبيل المثال عندما يتعلم طالب اللغة العربية كلمة عين في بمعناها الحقيقي الأساسي الذي يدل على العين الباصرة يخيل له انه قد استوعب كل معاني هذه الكلمة ويمكنه بسهولة أن يدرك معناها عندما يواجهها في سياقات أخرى لكن الحقيقة أنه سيواجه عددا من الصعوبات في فهم نفس هذه الكلمة عند تغير التركيب الذي وردت فيه لأنه إنما تعلم معنى واحداً من المعاني متعددة لهذه الكلمة في السياقات المختلفة فالكلمة المفردة التي نتعلمها في الغالب ماهي إلا رأسا لجبل من الثلج لايظهر لنا للوهلة الأولى لكنه سرعان ما يظهر لنا عند التعمق في المحيط ونواجه الكلمة نفسها في سياقات لغوية مختلفة كما يظهر بشكل واضح في الصورة ١التي تبين مدى تعقيد معاني كلمة عين في السياقات اللغوية المتنوعة.



صورة ١ : صورة تظهر مدى تعقيد معاني التراكيب المتعلقة بكلمة عين في اللغة العربية

٢- الدراسات السابقة:

سيركز سردنا للدراسات السابقة في هذه الورقة على أهم المحاولات السابقة لبناء معاجم حاسوبية للمتلازمات اللفظية في اللغتين العربية والانجليزية وسنقوم بمقارنة مختصرة بين الدراسة الحالية وما سبقها من أبحاث في هذا المجال. فيما يتعلق بقوائم المتلازمات اللفظية التي طورت لاستعمالها أداة تعليمية في مجال تعليم اللغات الأجنبية نجد في اللغة الإنجليزية عددا من المحاولات أولها كانت محاولة اللساني **Leech et al. (2001)** لتطوير قائمة للمتلازمات اللفظية عندما كان مشتركا في مشروع إنشاء المدونة الوطنية للغة الإنجليزية التي تتكون من مئة مليون كلمة وتعتبر أول دراسة منشورة في هذا المجال في اللغة الإنجليزية واعتمدت الدراسة على الاستخراج الآلي للعبارات المتجمدة من المدونة بعد عملية التحليل الحاسوبي الإحصائي الأولي ولذا نجد أن هذه القائمة اعتمدت على الألفاظ المتجاورة بشكل تام فقط مثل التركيب الإنجليزي **so that** والذي يمكن فهمه ككلمة واحدة مفردة وإن تكون من لفظين متتابعين وبناء على هذه النظرة الضيقة للمتلازمات اللفظية أهملت الدراسة كل أنواع المتلازمات اللفظية الأخرى والتي قد تتكون من عدة ألفاظ متقطعة وغير متتابعة أو متجمدة مثل التركيب الإنجليزي **(e.g. write down, write it down)** ومثل ذا النوع من التراكيب قد يصعب استخراجه بالاعتماد على التعرف الحاسوبي الآلي البحت. وفي دراسة أخرى قام بها **Durrant (2009)** حاول الباحث أن يطور قائمة للعبارات الأكثر استعمالا في الكتابة الأكاديمية العلمية وكانت مبنية على مدونة للكتابات العلمية قام بتطويرها تتألف من حوالي ٢٥ مليون كلمة وكان تركيز الدراسة في اختيار البارات على معيار الشيوع فأهملت كل العبارات اللغوية غير الشائعة بحجة عدم أهميتها وهذا قد يتسبب في فقدان مجموعة كبيرة من المتلازمات اللفظية المهمة قليلة الاستعمال. وقد اعتمد المنهج الإحصائي في النموذج الحاسوبي الذي طوره لاستخراج المتلازمات اللفظية من النصوص الأكاديمية بناء على الكلمات الأكاديمية الأكثر شيوعا في المدونة. وفي دراسة أخرى من قبل **Martinez & Schmitt (2012)** حاول الباحثان تطوير قائمة للمتلازمات اللفظية في اللغة الإنجليزية العامة مبنية على الشيوع ومرتبطة بشكل خاص بقائمة الكلمات الشائعة للغة الإنجليزية التي جمعها **West 1953)** وسماها بـ**.((general service list (GSL**
وهدف الدراسة الأساسي هو استعمال هذه القائمة كأداة تعليمية في تأليف مواد لتعلم وتعليم اللغة الإنجليزية للناطقين بغيرها وقد اعتمد الباحث على الشيوع كمعيار أساسي واستعمل الأداة الحاسوبية **WordSmith tools** لاستخراج العبارات الأكثر شيوعا معتمدا على الدونة البريطانية الوطنية للغة الإنجليزية والتي تتكون من مئة مليون كلمة من النصوص المكتوبة والمسموعة.

أما فيما يتعلق بالدراسات العربية الحاسوبية في هذا المجال فالملاحظ بشكل عام هو قلة الدراسات بشكل ظاهر مقارنة باللغات المعاصرة الأخرى وفيما يلي سرد لأهم الدراسات التي طبقت على اللغة العربية المعاصرة في هذا المجال. من أهم الدراسات وأكثرها تداولا في هذا المجال دراسة محمد طية للمتلازمات اللفظية في اللغة العربية والتي كانت جزءا من رسالة دكتوراه تهدف إلى تعزيز وتحسين جودة المحلل الحاسوبي الصرفي الذي قام ببنائه لمعالجة المستوى الصرفي في اللغة العربية وقد وجد أن تضمين هذا النوع من العبارات اللغوية في غاية الأهمية لتحسين جودة المخرج النهائي لهذه الأداة الحاسوبية واعتمدت الدراسة في منهجيتها على طريقة شبه آلية لاستخراج المتلازمات اللفظين من مدونة للغة العربية لم تذكر أي تفاصيل عنها في الدراسة المذكورة وبناء على تصنيف سابق لهذه العبارات قام به **Sag et al. (2002** اعتمد محمد عطية هذه التصنيف وحاول تطبيقه على اللغة العربية فقسم هذه العبارات بحسب معانيها إلى أربعة أصناف بحسب مدى غموض معانيها ومرونة التراكيب النحوية التي تتكون منها.

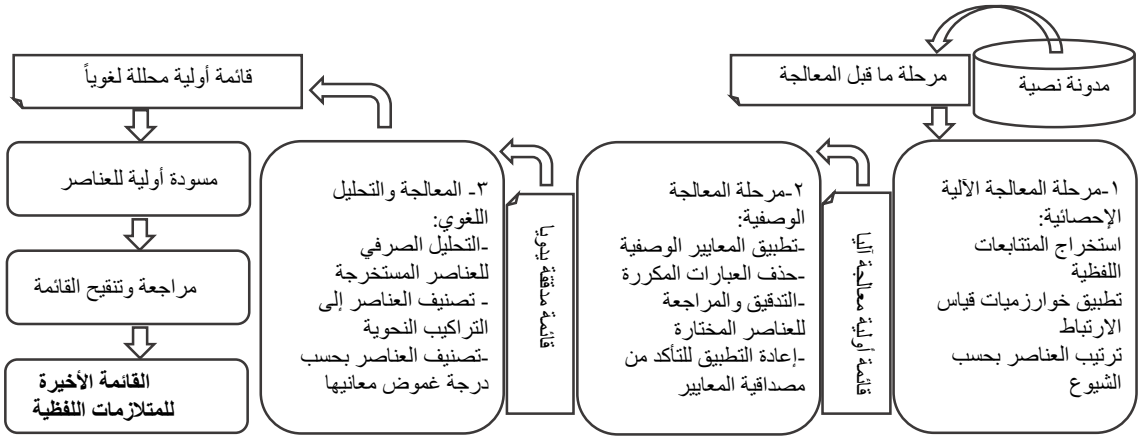
وفي دراسة أخرى قام بها **Hawwari et al. (2012)** حاول الباحثون بناء معجم حاسوبي مبني على عدد من المعاجم المنشورة سابقا للمتلازمات اللفظية في اللغة العربية كما في المعاجم المطبوعة التالية: **(Abou Saad,1987; Seeny et al., 1996; Dawod, 2003; Fayed, 2007)** وذلك لاستعماله كمادة لغوية لتدريب الحاسوب على التعرف الآلي على هذه العبارات باستخدام تقنيات وخوارزميات تعلم الآلة. وفي دراسة تالية **Hawwari et al. (2014)** حاول نفس الباحثون تطوير معجم للمتلازمات اللفظية المتداولة باللهجة المصرية واقترحوا إطارا نظريا للمعالجة الحاسوبية والتصنيف اللغوي لهذه العبارات وكلا المعجمين غير منشورين حتى كتابة هذا البحث. ومن خلال هذه النظرة المختصرة للدراسات السابقة تظهر لنا أهمية دراستنا الحالية ومدى ضرورة القيام بتطوير معجم حاسوبي مبني على المدونات اللغوية المطورة حديثا للمساعدة في تحسين عدد من مهام المعالجة الحاسوبية للغة العربية وكذلك استعماله كأداة تعليمية في تطوير مواد تعلم وتعليم اللغة العربية واختبارات اللغة وغيرها من المهام التربوية اللغوية.

٣- منهج الدراسة وطريقة البحث

اعتمدت التجربة الحالية على نموذج حاسوبي يجمع بين استعمال الطرق الإحصائية الآلية والوصفية اليدوية ويتكون من ثلاث مراحل أساسية يتخللها عدد من المهام الفرعية والتي تنظم عملية استخراج المتلازمات اللفظية من المدونة العربية وفي القسم التالي شرح مختصر لمنهجية استخراج المتلازمات اللفظية في الدراسة الحالية.

١,٣- النموذج الحاسوبي المطبق لاستخراج المتلازمات اللفظية العربية

كما يظهر بشكل واضح في النموذج أدناه تتكون عملية الاستخراج من عدة خطوات تبدأ باختيار المدونة التي يعتمد عليها في هذه التجربة ثم تمر بمرحلة ما قبل المعالجة والتي تشتمل على تصفية نصوص المدونة من الأخطاء الإملائية وتنميط الكلمات المختلفة كتابيا مثل توحيد طريقة كتابة الهمزة وألف الوصل وحذف النصوص المكررة وما إلى ذلك بعد ذلك تبدأ مرحلة المعالجة الآلية الإحصائية والتي تتضمن الاستخراج الآلي للمتتابعات اللفظية في المدونة بناء على نموذج المتتابعات الإحصائي (**n-gram model**) وتتبع هذه المرحلة مرحلة التحليل الوصفي والتي تركز على تطبيق مجموعة من المعايير الوصفية التي تم الاعتماد عليها في اختيار التراكيب المناسبة والتي سيتم توضيحها لاحقا بالتفصيل ثم تختتم عملية استخراج المتلازمات بمرحلة التحليل اللساني الآلي واليدوي لمخرجات المراحل السابقة في هذا النموذج، وسنتناول بالتفصيل جميع مراحل هذا النموذج مع الأمثلة خلال شرح التطبيق العملي للتجربة في هذه الأقسام التالية.



نموذج ١ : نموذج لمراحل استخراج المتلازمات اللفظية من المدونة

لأن من الأهداف الأساسية للدراسة الحالية استخراج قائمة للمتلازمات اللفظية مفيدة لمتعلمي اللغة العربية لابد أن يكون معيار الشيوع من المعايير المهمة في اختيار العبارات في هذه التجربة لأن كثيرا من الدراسات في اللسانيات التطبيقية أكدت على ان معيار الشيوع من أهم المعايير التي يمكن الاعتماد عليها في تحديد مدى فائدة العبارة اللغوية لمتعلمي اللغة **(e.g., Nation, 2001; O'Keeffe et al., 2007)** كما أكدت كثيراً من الدراسات المطبقة على اللغة الإنجليزية أن قوائم المفردات والعبارات المبنية على الشيوع لها دور محوري في إعداد وتصميم المواد اللغوية **(e.g., Nation, 2001; O'Keeffe et al., 2007; Schmitt and Martinez, 2012)** .

وبما أن الشيوع سيكون معيارا رئيساً فلابد كذلك أن يكون عدد الكلمات في كل عبارة محدودا وقليلا لأن العبارات الشائعة غالبا ما تكون قليلة الكلمات بسيطة التراكيب وهذا ما ستركز عليه التجربة الحالية. أما فيما يتعلق بحجم القائمة أو المعجم الحالي فسيكون نطاق التجربة الحالية محدودا بحيث لا يتجاوز خمسة آلاف عبارة وذلك سبب الوقت المحدود لهذه التجربة وكذلك بناء على عدد من الدراسات السابقة التي أثبتت أن هذا الرقم مناسب لقائمة محدودة مطورة ذات أهداف محددة.

وفيما يتعلق باختيار المدونة التي سيعتمد عليها في هذه التجربة سنختار المدونة اللغوية المطورة في جامعة ليدز من قبل الباحث **Sharoff (2006)** و تحوي هذه المدونة أكثر من ١٧٦ مليون كلمة وتشتمل على أهم مواضيع الحياة اليومية وكذلك فيها مجموعة كبيرة من النصوص الأكاديمية العلمية وهذا الاختيار يعود لعدة أسباب من أهمها مراعاة الخطوات العلمية المتبعة أثناء جمع النصوص في هذه المدونة وكذلك تنوع المواضيع التي تندرج تحتها نصوص المدونة فنجد فيها على سبيل المثال نصوصا عن السياسة والرياضة والفن والعلوم النظرية والتطبيقية ومن الأسباب كذلك كبر حجم هذه المدونة والتي يساعد على تمثيل مقبول إلى حد ما للغة العربية المعاصرة.

٢,٣ معايير اختيار المتلازمات اللفظية في التجربة الحالية:

بناء على التعريف المعتمد للمتلازمات اللفظية في هذه الدراسة وكذلك النموذج الحاسوبي المقترح سيكون اختيار العبارات في المرحلة الثانية من هذه التجربة مبني على عدد من المعايير الوصفية التالية:

أ- أن يكون في العبارة نوعا من الغموض الدلالي بحيث أنه لا يمكن بوضوح فهم معنى العبارة من خلال هم معاني كلماتها فمثلا في عبارة انتقل إلى رحمة الله قد يوجد بعض الإشكال في فهمها من خلال فهم محدود للكلمات المكونة لها ويمكن معرفة هذا

النوع من العبارات بطريقة استبدال العبارة بكلمة واحدة تدل على معناها كما في العبارة السابقة يمكننا استبدالها بالفعل مات أو توفي.

ب- أن لا يكون للعبارة معنى في حد ذاتها إلا إذا وجدت في سياق معين وأغلب هذه العبارات في اللغة العربية تتكون من الأدوات كحروف الجر والنصب كما في عبارة على الرغم من وكذلك عبارة من أجل أن وغيرها من التراكيب المشابهة.

ت- أن يكثر استعمال العبارة في الدلالة على وظيفة معنوية أو اتصالية مثل العبارات المتكررة المشهورة في اللغة كألفاظ التحيات والشكر والمجاملة (السلام عليكم، مع السلامة، شكرا جزيلاً...)

٤-تطبيق التجربة:

الهدف من التجربة الحالية هو التأكد من مدى فاعلية النموذج الحاسوبي المقترح لاستخراج المتلازمات اللفظية وكذلك تعتبر خطوة أولية لجمع مادة لغوية لبناء معجم حاسوبي شامل للمتلازمات اللفظية في العربية المعاصرة مبني على المدونات اللغوية.

كما تم ذكره سابقا في قسم المنهجية في هذا البحث يعتمد تطبيق هذا النموذج الحاسوبي على عدد من المهام والمراحل التي طبقت بالتفصيل كما يلي:

٤,١- الخطوات العملية وإجراءات تطبيق التجربة:

٤,١,١ اختيار المدونة ومهام ما قبل المعالجة:

في هذه المرحلة تم الاعتماد على واحدة من أهم مدونات اللغة العربية المعاصرة والتي طورت في جامعة ليدز وبدأت بعد ذلك مهام ما قبل المعالجة التي اشتملت على المعالجة الحاسوبية لتنقية وتصفية المدونة من الأخطاء الكتابية وتفادي التكرار بتوحيد طريقة كتابة بعض الحروف العربية كحرف الألف الذي له عدد من الأشكال وذلك لتهيئة المدونة للمرحلة التالية.

٤,١,٢ مرحلة المعالجة الإحصائية الآلية:

من خلال استعمال أداة الاستخراج الآلي للمتتابعات اللفظية (**n-grm tool**) وبنافذة محددة من كلمتين إلى أربع كلمات لكل تركيب تم بطريقة آلية استخراج أكثر من خمسة آلاف عبارة في هذه المرحلة بحيث لا يقل شيوع كل تركيب عن عشر مرات في كل مليون كلمة واختيار هذا الرقم مبني على ما وجده الباحث من توصيات سابقة في الدراسات السابقة **(Church & Hanks, 1990; Stubbs, 1995).** الجدول التالي يوضح عدد من العبرات المستخرجة والمعلومات الإحصائية المرتبطة بها مرتبة بحسب درجة شيوعها في المدونة.

متلازمات لفظية ثنائية	الشيوع	خوارزميات الارتباط			
		T-score	**MI**	**log likelihood**	**Log Dice**
بشكل	**68964**	**258.296**	**5.182**	**451,430**	**9.208**
من خلال	**59900**	**233.999**	**4.507**	**289,756**	**9.131**
بسبب	**49771**	**225.534**	**5.071**	**329,927**	**8.829**
بالنسب	**47091**	**214.019**	**5.407**	**341,640**	**8.667**
من أجل	**37889**	**193.911**	**5.378**	**263,642**	**8.566**
وسائل الإعلام	**5998**	**77.936**	**11.250**	**86,598**	**12.221**
سبيل المثال	**9452**	**97.299**	**12.252**	**156,386**	**12.924**
يوم القيامة	**8323**	**91.442**	**10.299**	**113,011**	**11.182**

جدول ١: أمثلة من العبارات الثنائية المستخرجة خلال المرحلة الأولى للتجربة

٤,١,٣ مرحلة تطبيق المعايير الوصفية:

استحوذت هذه المرحلة على الوقت الأكبر خلال إجراء هذه التجربة لأنها اعتمدت على التدقيقي اليدوي للعبارات المستخرجة من المرحلة السابقة ومحاولة تطبيق المعايير الوفية المحددة مسبقا وذلك لتنقية القائمة المطورة من كل التراكيب غير المرغوب فيها وتحسين جودة العناصر المستخرجة لتوافق الأهداف المحددة لهذه التجربة. وقد حذف خلال هذه التجربة عدد كبير من العبارات المستخرجة وذلك لمخالفتها للمعايير المعتمد عليها لاختيار المتلازمات اللفظية في هذه الدراسة ومن الأمثلة على العبارات المستبعدة ما يلي:

- العبارات التي تدل على اختصارات أو كلمات غير مفهومة أو أسماء أعلام وشركات أو أرقام مكتوبة.
- العبارات المكررة والتي تستعمل للدلالة على نفس المعنى مع اختلاف بسيط في طريقة التركيب.
- العبارات التي لا يصح أن تمثل اللغة العربية المعاصرة وذلك لاحتوائها على ألفاظ غير عربية أو محاكاة لأحد اللهجات العربية المحلية.
- العبارات المكونة من عدد من الحروف وليس لها معنى مستقل واضح كتركيب مفيد مثل من إلى وغيرها.
- العبارات التي لا يمكن أن يوجد أي غموض في فهم معناها الدلالي وذلك لتطابق معاني المفردات مع معنى التركيب الذي يتكون منها مثل عبارة وسائل الإعلام والجدول التالي يظهر بعض الأمثلة للعبرات المستبعدة في هذه المرحلة

تعديل أي من حقول ملفك الشخصي باستثناء	اسم المستخدم	,ف هو ثابت لك منذ تسجيلك . إذا كان الأمر
علي مسرح القومي . . فقرا في بعض يعني	مش عايز	تشوف مسرحية او حتي تعدي كده من جنب المسرح
الغربة غربة الروح أكثر ما هي غربة المكان و	بالنسبة	صفات الغرب الحميدة . لا أبالغ حين أقول
قليلة من الدولارات , و الآن يتقاضى عشرين	مليون دولار	ل يبحث عن كنز +ه القومي . ستة أصفار أشقاء
واسعة في وسائل الاعلام المطبوعة . حيث ان	وسائل الإعلام	في المملكة العربية السعودية تخضع ل سيطرة
عشرة ترجمة ل نهج البلاغة . و هذا ما يوضح	إلى حد ما	مكانة الكتاب و قيمت +ه بين المسلمين . هناك
و تهدف ل زيادة طاقة التفريغ و التحميل من	وعلى	السفن . و قال بان عمليات الانشاء تمت على
. عضو هيئة التحرير ل مجلة عالم الغذاء ب	المملكة العربية السعودية	عضو لجنة الإعجاز العلمي ب المجلس الأعلى
حسناء في عيون أهل بيت +ها . . . و كل ذلك	من أجل أن	تكون وفق رضا +هم . ذ هل هي امرأة خدومة .

جدول ٢: أمثلة للعبارات المستبعدة في مرحلة تطبيق المعايير الوصفية

وفي نهاية هذه المرحلة وبعد مراجعة دقيقه لكل العبارات المستخرجة قام بها الباحث بالاعتماد على المعايير الوصفية المحددة في هذه التجربة تم استبعاد القسم الأكبر من العبارات المستخرجة في المرحلة الأولى وبقي حوالي ٦٠٨ متلازمة لفظية انطبقت عليها المعايير المحددة في هذه المرحلة. وقد عمل الباحث على تأكيد وتوثيق طريقة تطبيق المعايير الوصفية بتطبيقها مرة أخرى من قبل شخص محايد ليس له علاقة بالبحث الحالي وهو أحد الطلاب المختصين في اللسانيات وكانت النتيجة الإحصائية متقاربة في تطبيق المعايير الوصفية وهذا بدوره عزز من الاعتماد على المعايير السابقة.

كما شملت هذه المرحلة على استخراج أمثلة لكل متلازمة لفظية من المدونة وكان التركيز على الأمثلة التي تبرز الاستعمال الحي للغة العربية المعاصرة كما يظهر ي الجدول التالي:

أمثلة	المتلازمات اللفظية
كان من أبطال الفيلم الأساسيين على الرغم من بعض العثرات في تقليد اللهجة.	على الرغم من
التعرض ل أشعة الشمس لفترات طويلة خطر يهدد جميع الأعمار بغض النظر عن كون من يتعرض لها رضيع أو طفل أو شاب.	بغض النظر عن
أن تناول طعام صحي وسليم مع اللعب النشيط يؤدي إلى صحة جيدة، وبالتالي فان الصحة الجيدة تؤدي إلى شهية جيدة.	وبالتالي
يجب أن تكون التكاليف في الإسلام حسب طاقة المكلف فلا يطلب منه على سبيل المثال إلا اثنين ونصف بالمائة من ربحه السنوي الصافي كزكاة.	على سبيل المثال

جدول ٣: المتلازمات اللفظية مع أمثلة استعمالها

٤,١,٤ مرحلة التحليل والتصنيف اللغوي:

في هذه المرحلة تم تمرير العناصر المستخرجة من المراحل السابقة على مجموعة من مراحل التحليل اللغوي أولها التحليل الصرفي الآلي والذي اشتمل على تقطيع الكلمات صرفيا ووسم كل كلمة بنوعها (اسم – فعل – حرف) ثم تحديد أجزاء الكلام لتراكيب العناصر المستخرجة وأخيرا التحليل الدلالي الذي من خلاله صنفت الكلمات إلى عدة أصناف بناء على درجة غموض المعنى في المتلازمات اللفظية. في هذه المرحلة اعتمد الباحث أداة مدى أميرة للتحليل الصرفي **MADAAMIRA** والجدول التالي يوضح التصنيف الصرفي الأولي للمتلازمات اللفظية المستخرجة في هذه التجربة

أمثلة	نوع الكلمة الأولى
نظراً لـ	اسم
متعلقة بـ	صفة
هنا وهناك	ظرف
ما يلي	اسم موصول
يؤدي إلى	فعل
لابد	أداة
في إطار الوصول	حرف
وبـالتالي	حرف عطف
سبحان الله	اسم مصدر

جدول ٤: أمثلة لأهم الأصناف التي تنتمي لها أول كلمة في المتلازمات اللفظية

أما بالنسبة لتحديد أجزاء الكلام فقد ظهر في هذه المرحلة أن المتلازمات اللفظية تنتمي إلى أغلب أنواع تركيب الجملة في اللغة العربية كما يظهر بالتفصيل في الجدول التالي لأهم تراكيب المتلازمات اللفظية في اللغة العربية:

أمثلة	التركيب
بمناسبة	حرف + اسم
ردًا على	اسم +حرف
في نهاية المطاف	حرف + اسم + اسم
سبحان الله	اسم مصدر + اسم
وبالتالي	حرف عطف +حرف جر+ اسم
رحمه الله	فعل +ضمير + اسم
من الضروري	حرف + صفة
مرتبط بـ	صفة +حرف
هنا وهناك	ظرف + حرف عطف + ظرف

جدول ٥: أمثلة لبعض التراكيب النحوية التي صنفت لها المتلازمات اللفظية في التجربة

وركزت الخطوة الأخيرة في التحليل اللغوي على محاولة تصنيف المتلازمات اللفظية إلى عدة فئات بناء على درجة غموض المعاني التي تدل عليها فكانت هناك ثلاث فئات في هذا المجال غموض شديد ومتوسط وبسيط ومع وجود تداخل قوي في هذا التصنيف إلا نه مفيد جدا وخاصة عند محاولة استعمال هذه القائمة في مهام المعالجة الحاسوبية للغة فالباحث حينئذ لا يحتاج إلا للعبارات الأكثر غموضا والتي لا يمكن بأي حال من الأحوال فهم معانيها من خلال فهم معاني الكلمات المكونة لها.

٥-ملخص وخاتمة

في هذه التجربة المختصرة حاولنا تطبيق نموذج حاسوبي معتمد على الطرق الإحصائية والوصفية لاستخراج قائمة للمتلازمات اللفظية في اللغة العربية المعاصرة وذلك كخطوة أولية لبناء معجم حاسوبي شامل لهذا النوع من العبارات بحيث يمكن الإفادة منه في مجال معالجة اللغة العربية حاسوبيا وكذلك تصميم وإعداد المواد التعليمية للغة العربية وخاصة لغير الناطقين بها. وقد اشتمل النموذج المطبق على ثلاث مراحل رئيسة يتخللها مجموعة من المهام والتي أسهمت أخيرا في إنشاء قائمة مراجعة ومنقحة للمتلازمات اللفظية.

وفي التجارب المستقبلية نرجو أن نتمكن من توسيع نطاق التجربة الحالية لتطبق على المدونات العربية الضخمة وكذلك محاولة تطبيق التجربة بطريقة آلية في كل الخطوات وذلك لتوفير الوقت والجهد وزيادة الدقة في مراحل استخراج المتلازمات اللفظية. وكذلك يمكن أن يكون هذا المعجم نواة لبيئة حاسوبية متكاملة لتعليم العربية على شبكة المعلومات الإنترنت.

المراجع:

احمد ابو سعد, & ظبية عبد الله محمد السليطي. (1991). معجم التراكيب والعبارات الاصطلاحية العربية القديم منها والمولد، دار العلم للملايين.

Ackermann, K. and Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. Journal of English for Academic Purposes. 12(4), pp.235-247.

Attia, M. (2006). Accommodating Multiword Expressions in an Arabic LFG Grammar. In T. Salakoski et al. (Eds.): Advances in Natural Language Processing. FinTAL 2006, Lecture Notes in Computer Science. Vol. 4139, pp. 87 - 98, 2006. Springer-Verlag Berlin Heidelberg.

Attia, M., Tounsi, L., Pecina, P., van Genabith, J., & Toral, A. (2010). Automatic extraction of Arabic multiword expressions.

Baldwin, T. (2004). Multiword Expressions, an Advanced Course. Paper presented at The Australasian Language Technology Summer School (ALTSS 2004), Sydney, Australia.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. Applied linguistics, 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). Longman grammar of spoken and written English (Vol. 2). MIT Press.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. Language teaching research, 10(3), 245-261.

Butt, M. (1999). *A Grammar Writer's Cookbook*. Stanford, CA: CSLI.

Calzolari, N., Lenci, A., & Quochi, V. (2002). Towards Multiword and Multilingual Lexicons Between Theory and Practice. na.

Capel, A. (201)0. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. English Profile Journal. 1(01), pe3.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics. 16(1), pp.22-29.

Coxhead, A.(2000). A new academic wordlist. TESOL Quarterly, 34(2), pp.213-238.

Davies, M. and Gardner, D.(2013). A Frequency Dictionary of American English: Word Sketches, Collocates and Thematic Lists. Routledge.

Dawood, Mohammed. (2003). A Dictionary of Arabic Contemporary Idioms (mu'jm alta'bir alastlahiat).Dar Ghareeb.

Diab, M.T. and Krishna, M. (2009)a. Handling sparsity for verb noun MWE token classification. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics: Association for Computational Linguistics, pp.96-103.

Diab, M.T. and Krishna, M. (2009)b. Unsupervised classification of verb noun multi-word expression tokens. Computational Linguistics and Intelligent Text Processing. Springer, pp.98-110.

Dipper, Stefanie. (2003). Implementing and Documenting Large-Scale Grammars – German LFG, Institut f¨ur maschinelle Sprachverarbeitung, Institut f¨ur maschinell Sprachverarbeitung der Stuttgart University: Ph.D.

Dorgeloh, H. and Wanner, A. (2009). Formulaic argumentation in scientific discourse. Formulaic language. 2, pp.523-44.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. English for Specific Purposes. 28(3), pp.157-169.

Durrant, P.L. (2008). High frequency collocations and second language learning. thesis, University of Nottingham.

EbdAlrzAq, O. (2007). AlmtlAzmAt AllfZyp fy Allgp wAlqwAmys AlErbyp. twns: mjmE AlOTr$.

Ellis, N. C., SIMPSON-VLACH, R. I. T. A., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. Tesol Quarterly, 42(3), 375-396.

Ellis, N.C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. The Quarterly Journal of Experimental Psychology: Section A. 49(1), pp.234-250.

Ellis, R.S.-V.a.N.C. (2010). An Academic Formulas List: New Methods in Phraseology Research Applied Linguistics. 31, pp.487-512.

Erman, B. and Warren, B. (2000). The idiom principle and the open choice principle. text-the hague then amsterdam then Berlin-. 20(1), pp.29-62.

Fayed, Wafaa Kamel. (2007). A Dictionary of Arabic Contemporary Idioms (mu'jm alta'bir alastlahiat).Abu Elhoul.

Fellbaum, C. (1998). WordNet. Blackwell Publishing Ltd.

Firth, J. (1951). Papers in Linguistic [s J]. Oxford University Press.

Garside, Roger. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson (Eds.), The Computational Analysis of English (30-41). London: Longman.

Gralinski, F., Savary, A., Czerepowicka, M., & Makowiecki, F. (2010). Computational Lexicography of Multi-Word Units: How Efficient Can It Be?. In Workshop Multiword Expressions: from Theory to Applications.

Guenthner, F., & Blanco, X. (2004). Multi-lexemic expressions: an overview. Lexique, Syntaxe et Lexique-Grammaire. Papers in Honour of Maurice Gross, 239-252.

Habash, N., & Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In Proceedings of the Conference of American Association for Computational Linguistics (pp. 578-580).

Hawwari, A., Attia, M., & Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal arabic. ANLP 2014, 48.

Hawwari, A., Bar, K., & Diab, M. (2012). Building an Arabic multiword expressions repository. Proc. of the 50th ACL, 24-29.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge University Press.

Hyland, K.(2008). As can be seen: Lexical bundles and disciplinary variation. English for specific purposes. 27(1), pp.4-21.

Kjellmer, Goran. (1990). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), English corpus linguistics: Studies in honour of Jan Svartvik (pp. 111-127). London: Longman.

Lee, D.(2002). Notes to accompany the BNC WORLD edition (bibliographical) index. Unpublished manuscript, last edited. 30.

Leech, G., & Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Leech, G., Garside, R., & Atwell, E. S. (1983). The automatic grammatical tagging of the LOB corpus. ICAME Journal: International Computer Archive of Modern and Medieval English Journal, 7, 13-33.

Li, W., Zhang, X., Niu, C., Jiang, Y., & Srihari, R. (2003, July). An expert lexicon approach to identifying English phrasal verbs. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 513-520). Association for Computational Linguistics.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. English for Specific Purposes. 31(1), pp.25-35.

Martinez, R. (2011). The development of a corpus-informed list of formulaic sequences for language pedagogy. thesis, University of Nottingham.

Martinez, R. and Murphy, V.A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. TESOL Quarterly. 45(2), pp.267-290.

Martinez, R. and Schmitt, N. (2012). A Phrasal Expressions List. APPLIED LINGUISTICS. 33(3), pp.299-320.

Meghawry, S. (2015). Semantic extraction of Arabic multiword expressions.

Mel'ćuk, I.(1998). Collocations and lexical functions. 2001 [1998]. pp.23-54.

Milton, J. (2009). Measuring second language vocabulary acquisition. Multilingual Matters.

Nation, I.S.P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. Vocabulary: Description, acquisition and pedagogy. 14, pp.6-19.

Nerima, L., Seretan, V., & Wehrli, E. (2003, April). Creating a multilingual collocation dictionary from large text corpora. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2 (pp. 131-134). Association for Computational Linguistics.

Nesselhauf, N. (2005). Collocations in a learner corpus. Amsterdam: John Benjamins.

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. Formulaic language. 2, pp.387-404.

O'keeffe, A., McCarthy, M., & Carter, R. (2007). From corpus to classroom: Language use and language teaching. Cambridge University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), Language and communication (pp. 191-226). New York: Longman.

Roth, R., Rambow, O., Habash, N., Diab, M., & Rudin, C. (2008, June). Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers (pp. 117-120). Association for Computational Linguistics.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002, February). Multiword expressions: A pain in the neck for NLP. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 1-15). Springer Berlin Heidelberg.

Sawalha, M. and Atwell, E. (2011). Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus. Advanced Research Computing Open Event.

Schmitt, N. (2004). Formulaic sequences: Acquisition, processing, and use. John Benjamins Publishing.

Schmitt, N. and Martinez, R.(2012). A Phrasal Expressions List. Applied Linguistics. 33(3), p299.

Schmitt, N.(2010). Researching vocabulary: a vocabulary research manual. Basingstoke: Palgrave Macmillan.

Scott, Mike. (1999). WordSmith Tools users help file. Oxford: Oxford University Press.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. WaCky. pp.63-98.

Shin, D.a.N., P. (2008). Beyond single words: The most frequent collocations in spoken English. ELT Journal 62(4), pp.339-348.

Sieny, Mahmoud Esmail, Mokhtar A. Hussein and Sayyed A. Al-Doush. (1996). A contextual Dictionary of Idioms (almu'jm alsyaqi lelta'birat alastlahiah). *Librairie du Liban Publishers*.

Sinclair, J. (1987). Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English language dictionary. London: Collins ELT.

Sinclair, John M. (1987). Collocation: a progress report. In R. Steele & T. Threadgold (Eds.), Language topics: Essays in honour of Michael Halliday (Vol. 2, pp. 319-331). Amsterdam: John Benjamins.

Siyanova-Chanturia, A. et al.(2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. Second Language Research. 27(2), pp.251-272.

Smadja, F. et al. (1996). Translating collocations for bilingual lexicons: A statistical approach. Computational linguistics. 22(1), pp.1-38.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. Functions of language. 2(1), pp.23-55.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge University Press Cambridge.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), Formulaic sequences: acquisition, processing and use (pp. 249-268). Amsterdam: John Benjamins.

Wray, A. (2008). Formulaic language: Pushing the boundaries. Oxford: Oxford University Press.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. Formulaic language. 1, pp.27-51.

Wray, A. (2013). Formulaic language. Language Teaching. 46(03), pp.316-334.

Wray, A. and Namba, K. (2003). Formulaic language in a Japanese-English bilingual child: a practical approach to data analysis. Japan Journal for Multilingualism and Multiculturalism. 1(9), pp.24–51.

Wulff, S., Swales, J. M., & Keller, K. (2009). "We have about seven minutes for questions": The discussion sessions from a specialized conference. English for specific purposes, 28(2), 79-92.