

مجلة اللسانيات العربية، العدد 16، جمادى الآخرة، 1444هـ/يناير، 2023م

الخواص الممكنة لاستخلاص المصطلحات اللغوية من المدونات العربية المتخصصة

عبدالمحسن بن عبيد الثبيتي^أ * و داد بنت أحمد القحطاني^ب

^أ المركز الوطني لتحليل البيانات والذكاء الاصطناعي، مدينة الملك عبدالعزيز للعلوم والتقنية، الرياض، المملكة العربية السعودية
^ب قسم اللغة العربية وآدابها، كلية العلوم الإنسانية والاجتماعية، جامعة الملك سعود، الرياض، المملكة العربية السعودية

توثيق البحث APA Citation:

الثبيتي، عبدالمحسن، والقحطاني، داد. (2023). الخواص الممكنة لاستخلاص المصطلحات اللغوية من المدونات العربية المتخصصة. مجلة اللسانيات العربية، 16، 179-209

Submission Date: 26/10/2021
Acceptance Date: 26/02/2022

تاريخ الإرسال: 1443/03/20
تاريخ القبول: 1443/07/25

Abstract

This paper entitled “The enabling properties for term extraction from Arabic domain-specific corpora” presents a descriptive statistical study of some important linguistic and statistical heuristics used in term extraction from domain-specific corpora. This study relied on a domain-specific corpus manually tagged with linguistic terms, to know the common grammatical and morphological structures of these terms and their statistical properties compared to a general corpus. Knowing the common linguistic and statistical patterns helps in developing more accurate heuristics that contribute to reducing errors in extracting terms from domain-specific corpora. Also, knowing what is uncommon from the statistical structures and properties of terms helps in building tagged corpora enriched with examples of uncommon patterns, which contributes to building more accurate computational models for terms extraction.

Keywords : Terms extraction, corpus linguistics, computational linguistics.

الملخص

تقدم هذه الورقة دراسة إحصائية وصفية لبعض القواعد الاسترشادية اللغوية والإحصائية المهمة المستخدمة في استخلاص المصطلحات من المدونات اللغوية المتخصصة. وقد اعتمدت هذه الدراسة على مدونة لغوية متخصصة موسّمة بالمصطلحات في مجال اللغويات؛ لمعرفة التراكيب النحوية والصرفية الشائعة لتلك المصطلحات وخواصها الإحصائية مقارنة بمدونة لغوية عامة. حيث إنّ معرفة ما هو شائع من أنماط لغوية وإحصائية، يساعد في بلورة قواعد استرشادية أكثر دقة، ويسهم في تقليل أخطاء استخلاص المصطلحات من المدونات اللغوية المتخصصة. كما أنّ معرفة ما هو غير شائع من التراكيب والخواص الإحصائية المتعلقة بالمصطلحات، يساعد في بناء مدونات موسّمة معززة بأمثلة من الأنماط غير الشائعة، ممّا يسهم في بناء نماذج حاسوبية أكثر دقة لاستخلاص المصطلحات. الكلمات المفتاحية: استخلاص المصطلحات، لسانيات المدونات، اللسانيات الحاسوبية.

* المؤلف المراسل
Email: aalthubaity@kacst.edu.sa

1. المقدمة

تزخر شبكات الإنترنت العامة -التي يمكن الدخول إليها من عامة الناس مجاناً أو بمقابل مادي - أو شبكات الإنترنت المغلقة - الخاصة بمؤسسات وجهات حكومية أو أهلية أو تجارية - بكم هائل من النصوص بلغات مختلفة من ضمنها العربية، قابلة للقراءة آلياً ومن ثمّ الاستفادة منها وتوزيعها وتخزينها وتحليلها أيضاً آلياً أو شبه آلياً. ومن ضمن هذه النصوص ما هو مكتوب من الخبراء والعاملين في مجالات متخصّصة وموجه لأقرانهم المهتمين والعاملين في ذات التخصص مثل: الباحثين والأكاديميين في المجالات العلمية المختلفة، أو المهندسين والفنيين العاملين في المجالات الهندسية والفنية. وتشمل هذه النصوص على سبيل المثال لا الحصر: الأوراق العلميّة المحكّمة المنشورة في المجلات والمؤتمرات، والرسائل الجامعية لنيل درجتي الماجستير والدكتوراه، أو أدلة تشغيل الأنظمة والآلات وتركيبها وصيانتها.

ومن أهم ما يميّز النصوص المتخصّصة استخدامها المتكرر للمصطلحات وهي: وحدات معجمية تدلّ على المفاهيم والمناهج والأساليب، والأدوات المستخدمة في المجال المتخصص، أو ما له علاقة بالمجال من مجالات متخصّصة أخرى. ويمكن أن تكون هذه المصطلحات مفردة: مكوّنة من كلمة واحدة مثل "اللسانيّات"، و "التداوليّة"، و "الجملة"، أو مركبة: مكوّنة من كلمتين أو أكثر مثل "النحو الوظيفي"، و "النظرية التوليديّة التحويليّة"، و "الأفعال الإنجازيّة غير المباشرة".

وباستخدام النصوص المتخصّصة يمكن بناء مدوّنات لغويّة متخصّصة كبيرة الحجم، يسهم استخلاص المصطلحات منها - آلياً أو شبه آلياً - في توفير الوقت والجهد والتكلفة اللازمة لإدارة مصادر المعرفة المتخصّصة، وخدمة مجالات وتطبيقات متعددة، منها على سبيل المثال لا الحصر: بناء المعاجم المتخصّصة وتوسيعها (Kim, Choi, and Shin, 2021)، وبناء الأنطولوجيا وشبكات العلاقات بين المفاهيم (Li, 2021)، ودراسة نشوء المجالات المتخصّصة وتطورها (Eito-Brun, 2017)، تصنيف النصوص (Liu and Chen, 2019)، فهرسة واسترجاع الوثائق المتخصّصة (Fkih and Omri, 2016)، والترجمة الآليّة (Liu, 2021).

وبشكل عام، فإنّ هناك أربعة أساليب مستخدمة في استخلاص المصطلحات من المدوّنات اللغويّة المتخصّصة: الأساليب الإحصائية (Yan et al, 2020; Kosa et al, 2019)، والأساليب اللغويّة (Simon and Kešelj 2018; Luo et al, 2019)، والأساليب القائمة على منهجيات الذكاء الاصطناعي وخاصة تعلّم الآلة (Zeng, Li and Li, 2018; Rigouts Terryn et al, 2020)، والأساليب الهجينة التي تجمع أسلوبين أو أكثر من الأساليب السابقة (Deepthi et al, 2017; Oliver and Vázquez, 2020).

وتعتمد الأساليب الإحصائية على قاعدتين استرشاديتين عامتين، أولاهما: أنّ التوزيع الإحصائي للمصطلحات في المدوّنات اللغويّة المتخصّصة يختلف عن توزيعها الإحصائي في المدوّنات اللغويّة العامّة، مثل: المدوّنات اللغويّة الصحفية. فالمصطلحات تتكرر بكثرة في المدوّنات اللغويّة المتخصّصة، ولكنّها لا تظهر في المدوّنات اللغويّة العامّة أو نادرة الظهور بها، كما أنّ توزيعها الإحصائي داخل النصوص المتخصّصة يختلف عن بقية الكلمات. والقاعدة الاسترشادية الثانية: أنّ التلازم بين الكلمات المكوّنة للمصطلحات المركّبة، أو احتماليّة ظهورها مع بعضها البعض، أعلى من تلازمها واحتماليّة ظهورها مع كلمات أخرى.

وتعتمد الأساليب اللُّغويّة بشكل أساسي على أقسام الكلام التي تندرج تحتها المصطلحات. فالمصطلحات المفردة تكون أسماء مثل: ("اللسانيات"، و"التداولية" و"الصوائت")؛ أمّا المصطلحات المركّبة، فتبدأ بأسماء تليها أسماء مثل (فقه اللغة، قواعد أصول الفقه)، أو أسماء تليها صفات مثل (النحو الوظيفي، علم اللغة النّفسي). أو قد تكون الأساليب اللُّغويّة معتمدة على قواعد استرشادية أخرى، كأنّ تكون المصطلحات المتخصّصة مسبقة ومتبوعة بالكلمات الوظيفية وعلامات الترقيم. وتعتمد أساليب الذكاء الاصطناعي على المدوّنات اللُّغوية الموسّمة يدويّاً بالمصطلحات التي تتعلّم منها خوارزميّات تعلّم الآلة القواعد التي تستطيع من خلالها الكشف عن المصطلحات المفردة والمركّبة. وتكون هذه المدوّنات في العادة موسّمة آليّاً بأقسام الكلام، وتتطلب وقتاً وجهداً كبيرين لبنائها.

أمّا الأساليب الهجينة فهي تجمع في الأغلب بين الأساليب اللُّغوية والأساليب الإحصائية أو تعلم الآلة. ففي حالة الجمع بين الأساليب اللُّغوية والإحصائية، يتم أولاً تحديد المركّبات الاسمية، ثم تُرتّب بحسب مقاييس إحصائية، والأعلى رتبة من هذه المركّبات يعدّ مصطلحاً محتملاً. أمّا في حالة الجمع بين الأساليب اللُّغوية وتعلّم الآلة، فإنّ الخواصّ المستخدمة لتمثيل الكلمات في النصوص تكون أغلبها خواصّ لغوية، فهي إمّا معلومات تختص بأقسام الكلام، أو التركيب النحوي للجملة والكلمات المكونة لها، أو نماذج لغويّة مسبقة التدريب على مدوّنات لغويّة ضخمة جدّاً باستخدام التعلّم العميق.

وتواجه هذه الأساليب - ماعدا أساليب تعلّم الآلة - ثلاث عقبات رئيسية، الأولى: أنّها قائمة على قواعد استرشادية مدعومة بالحدس والملاحظات الشخصية. وبالرغم من صحتها على وجه العموم، فنحن لا نعلم على وجه الدقة إنّ كانت تنطبق على جميع المصطلحات، وفي جميع التخصصات أو لا. فقاعدة اختلاف التوزيع الإحصائي للمصطلحات في المدوّنة المتخصّصة عن توزيعها في المدوّنة العامة: هل تنطبق على كلّ المصطلحات؟ أم إنّ هناك مصطلحات لا تنطبق عليها هذه الخاصيّة؟ وقاعدة أنّ المصطلحات المفردة عبارة عن أسماء، والمصطلحات المركّبة عبارة عن تراكيب تبدأ بالأسماء وتكون بقية مركّباتها أسماء أو صفات: ما الخواصّ الصرفيّة التي تميز هذه الأسماء والصفات؟ وهل جميع الأسماء والصفات قابلة لأنّ تكون ضمن المصطلحات؟

والعقبة الثانية هي: شحّ المدوّنات اللُّغويّة المتخصّصة كبيرة الحجم في كثير من اللغات، ومنها اللغة العربيّة. فالحجم الكبير للمدونة اللُّغويّة المتخصّصة يعزّز التوزيع الإحصائي للمصطلحات، وظهور خواصها اللُّغوية المشتركة. والعقبة الثالثة: أنّ تقييم النتائج يعتمد على المراجعة البشرية المباشرة للمصطلحات المستخرجة وهذه عملية مكلفة ومجهدة. أو أنّ تُقيّم النتائج من خلال مقارنة النتائج مع ما هو موجود في القواميس المتخصّصة. ولكن المصطلحات ليست كلّها مدرجة في القواميس المتخصّصة، أو أنّ طريقة ترتيبها وعرضها موافقة للمصطلحات المستخرجة. وهذا بلا شكّ يؤثر في تقييم النتائج، وقد يبخس الطريقة المستخدمة حقها.

أمّا الأساليب المعتمدة على تعلّم الآلة، فهي تعاني من مشكلتين رئيسيتين، الأولى: شحّ المدوّنات اللُّغويّة الموسّمة يدويّاً بالمصطلحات. ولا توجد بالنسبة إلى اللغة العربيّة - في حدود علمنا - أيّ مدوّنة لغويّة عربيّة متخصّصة موسّمة بالمصطلحات. والمشكلة الثانية هي: تحديد أفضل الخواصّ التي يمكن استخدامها لتدريب خوارزميات تعلّم الآلة، وعدم وجود نماذج لغويّة مبنية على مدوّنات لغويّة متخصّصة كبيرة الحجم.

وبالرغم من الاهتمام السابق والمستمر بموضوع استخلاص المصطلحات من المدوّنات المتخصّصة في لغات متعددة (Häty et al, 2020; Hua and Song, 2021; Speranza et al, 2020)، فإنّ هذا المجال لا يزال بكرةً في الأبحاث المتعلقة بالمدوّنات العربية المتخصّصة، فضلاً عن الأبحاث المكتوبة باللغة العربية. ولعلّ سبب ندرة هذه الأبحاث عائد إلى ندرة النصوص العربية المتخصّصة القابلة للقراءة آلياً، والتي يمكن أن تستخدم في بناء المدوّنات اللُّغويّة العربية المتخصّصة، إضافة إلى عدم وجود مدوّنات عربيّة متخصّصة موسّمة بالمصطلحات يمكن استخدامها لتقييم الطرق المستخدمة في الاستخلاص. ولعلّ أهم ما يساعد على تطوير الطرق المستخدمة في استخلاص المصطلحات، وتجاوز العقبات والإشكالات التي تواجهها جراء استخدام القواعد الاسترشادية، أو تطوير طرق جديدة هو: معرفة حدود القواعد الاسترشادية المستخدمة وإمكاناتها اعتماداً على الاستخدام الحقيقي للغة في المجالات المتخصّصة، ومعرفة مدى قرب المبادئ الاسترشادية المستخدمة حالياً، وبُعدها من واقع المصطلحات في المدوّنات العربيّة المتخصّصة.

تقدّم هذه الورقة مساهمتين رئيسيتين. المساهمة الأولى: مدونة لغويّة عربيّة متخصّصة بعلوم اللغة موسّمة يدوياً بالمصطلحات، وآلياً بأقسام الكلام. والمساهمة الثانية: دراسة إحصائيّة وصفيّة لبعض الخواصّ اللُّغويّة والإحصائيّة للمصطلحات في المدوّنات اللُّغويّة المتخصّصة. وتشمل هذه الدراسة وسوم أقسام الكلام، والخواصّ الصرفيّة التي تندرج تحتها المصطلحات، ودراسة التوزيع الإحصائي للمصطلحات اعتماداً على بيانات الاستعمال الحقيقي للمصطلحات. ولا يوجد في حدود معرفتنا - حتى تاريخ كتابتنا لهذه الورقة - مدونة عربيّة متخصّصة موسّمة بالمصطلحات، أو دراسة موسّعة تتحقّق من الخواصّ اللُّغويّة والإحصائيّة للمصطلحات في المدوّنات اللُّغويّة المتخصّصة.

يتناول القسم الثاني من هذه الورقة الدراسات السابقة ذات الصلة بموضوع ورقتنا هذه في. ويشرح القسم الثالث المدوّنات اللُّغويّة المستخدمة في الدراسة، وكيفية معالجتنا وتوسيمنا لها. ويوضّح القسم الرابع الطريقة التي استخدمناها لدراسة التراكيب النحوية والصرفية للمصطلحات ولتوزيعها الإحصائي. أمّا القسمان الخامس والسادس، فيستعرضان نتائج الدراسة الخاصة بالتراكيب النحوية والصرفية للمصطلحات وتوزيعها الإحصائي. وقد خصصنا القسم السابع للخلاصة والتوصيات.

2. الدراسات السابقة

اهتمت الدراسات اللُّغويّة العربية - وماتزال - بالمصطلح وعلومه من نواحي متعددة شملت على سبيل المثال لا الحصر طرق وضع المصطلحات من خلال التوليد أو التعريب (الحمزاوي، 1986؛ القاسمي، 2008؛ خسارة، 2008)، والمشاكل التي تواجه وضع المصطلح وترجمته (منتصر، 2020؛ السوداني، 2018) والتخطيط المصطلحي (العتيبي، 2021)، وتأثير المصطلح في لغة الاختصاص (بناتي، 2020؛ بوشحان، 2017). كما شملت الدراسات استخدام الحاسب، ووضع المصطلحات ونشرها (الشريدة، 2017).

ولكنّا لم نجد من الدراسات اللُّغويّة ما يهتم بدراسة الخواصّ اللُّغويّة والإحصائيّة للمصطلح بناءً على الاستخدام الواقعي للمصطلحات في اللُّغات المختصّة، أو دراسة للطرق والخواصّ الممكن استخدامها للكشف عن المصطلحات في النصوص والمدوّنات المتخصّصة. واقتصر هذا الأمر على الحاسوبيين في الأغلب، وكانت جُلّ هذه الدراسات مكتوبة بالإنجليزية. غير أننا قد وجدنا - خلال بحثنا عن دراسات تشبه دراستنا هذه - دراسة (شندول، 2008) بعنوان "بنية المصطلح في كتاب المقتضب للمبرد". وقد اقتصر الباحث في دراسته على

المصطلحات المتعلقة بالفعل والواردة في فهارس الكتاب. وقد كان عدد هذه المصطلحات 59 مصطلحاً، جاء عدد المصطلحات المفردة منها 15 مصطلحاً، والمركبة 44 مصطلحاً. وقد درس الباحث بنية هذه المصطلحات من حيث الاشتقاق، والنحت والمعجمة والتركيب. ووجد أنّ تركيب المصطلحات المركبة من كلمتين ينحصر في نوعين هما التركيب الإضافي (اسم-اسم) والتركيب الوصفي (اسم-صفة) أو (صفة-صفة) مثل "المضارع المجزوم". وأمّا المصطلحات المفردة، فقد كانت كلها أسماء معرفة بالألف واللام.

وسنوضّح في بقية هذا القسم من الدراسة شرحاً مختصراً للطرق التي استخدمت سابقاً في الدراسات الخاصة باستخلاص المصطلحات من المدوّنات اللُّغويّة المتخصّصة؛ لنقف على القواعد المستخدمة في تلك الدراسات، ولننطلق منها بعد ذلك لمعرفة حدود الطرق المقترحة وإمكاناتها لاستخلاص المصطلحات.

وكما أشرنا سابقاً، فإنّ هناك أربعة أساليب مستخدمة في استخلاص المصطلحات من المدوّنات المتخصّصة، وهي: الأساليب الإحصائية، والأساليب اللُّغوية، وأساليب تعلّم الآلة، والطرق الهجين التي تجمع بين اثنتين وأكثر من الطرق الثلاث. وكلّ الدراسات المتعلقة باستخلاص المصطلحات من المدوّنات العربية المتخصّصة اعتمدت على الأساليب الإحصائية واللُّغويّة، أو الدمج بينهما. ولم نجد أيّ دراسة تستخدم أساليب تعلّم الآلة. وسنركّز في هذا القسم على الأساليب المستخدمة في استخلاص المصطلحات المفردة والمركبة من المدوّنات العربية المتخصّصة وشرحها بشكل مبسّط.

وجدنا أربع دراسات لاستخلاص المصطلحات المفردة، كانت أولاهما دراسة الثبيتي (2007) حيث استخرج الكلمات المميزة من مدونتين عربيتين متخصصتين بعلوم الفيزياء والذكاء الاصطناعي، مستخدماً في ذلك ما يعرف بمعامل الغرابة (Weirdness Coefficient) (Ahmad and Rogers, 2001)، ثم رتّب الكلمات بحسب قيمة المعامل من الأكبر رتبة إلى الأصغر رتبة، واعتبر الكلمات التي كان معامل الغرابة لها أكبر من 100 مصطلحات مفردة. وقد استخدم مشعان وزملاؤه (Mashaan et al, 2013)، والثبيتي وزملاؤه (Al-Thubaity et al, 2014) طريقة إحصائية أخرى تعتمد على مقياس تكرار الكلمة-عكس تكرار النصوص (TFIDF)، حيث يعطي هذا المقياس كلّ كلمة في المدونة قيمة تدلّ على أهميتها، فالكلمات ذات القيم العليا تكون الأكثر أهمية في المدونة، ثم تُرتّب الكلمات بحسب أهميتها ثم يُختار الأعلى رتبة منها بكونها مصطلحات.

واستخدم الثبيتي وزملاؤه (Al-Thubaity et al, 2014) أيضاً طريقة هجينة تعتمد على قاعدة استرشادية مفادها: أنّ المصطلحات المفردة أو المركبة محاطة بالكلمات الوظيفية وعلامات الترقيم؛ حيث استخرجوا المصطلحات المفردة التي تنطبق عليها هذه القاعدة، ثم استخدموا مقياس تكرار الكلمة-عكس تكرار النصوص بنفس الطريقة الموضحة أعلاه لاختيار المصطلحات. أمّا الثبيتي والقحطاني (2022)، فقد استخدموا طريقة هجينة للكشف عن المصطلحات المفردة تعتمد على قاعدتين استرشاديتين مفادهما: أنّ المصطلحات المفردة عبارة عن: أسماء معرفة بالألف واللام، وأنّ توزيعها الإحصائي في المدونة المتخصّصة يختلف عن توزيعها في المدونة العامة. حيث استخرجوا الأسماء المعروفة بالألف واللام أولاً، ثم قارنا توزيع هذه الأسماء المعروفة مع توزيعها في المدونة العامة باستخدام معامل التشابه اللوغاريتمي، ثم جعلنا الكلمات الأعلى رتبة مصطلحات.

أمّا الدراسات الخاصة باستخلاص المصطلحات المركبة من المدوّنات العربية المتخصّصة، فقد اعتمدت جميعها على الأساليب الهجينة. واتبعت أغلبها خطوتين رئيسيتين، الأولى: هي وسم المدونة المتخصّصة بأقسام الكلام، ثم اختيار المركّبات المحتملة للمصطلحات المركبة مثل "اسم-اسم" و "اسم-صفة" للمصطلحات المركبة من كلمتين أو "اسم-اسم-اسم" و "اسم-اسم-صفة" للمصطلحات المركبة من ثلاث كلمات. والخطوة الثانية بعد ذلك اكتشاف مدى التلازم أو الارتباط بين مكونات هذه المصطلحات المركبة باستخدام مقاييس إحصائية مختلفة.

وقد شملت هذه المقاييس مقياس الاحتمال اللوغاريتمي (Log-Likelihood) (Bounhas and Slimani, 2009)، ومقياس تي (T-score)، ومقياس إف إل آر (FLR) (Boulaknadel et al, 2008)، ومقياس مربع كاي والمعلومات المتبادلة (Chi-square and Mutual Information) (Mashaan et al, 2013)، ومقياس قيمة سي (Al Khatib and Badarneh, 2010) (C-value) ومقياس مطورة من مقياس قيمة سي (El Mahdaouy, Ouatik and Gaussier, 2013; Lamrani (NC-value, NTC-value, and NLC-value) (et al, 2014; Hadni et al, 2014).

واستخدم الثبيتي وزملاؤه (Al-Thubaity et al, 2014) القاعدة الاسترشادية التي تقول: إنَّ المصطلحات يسبقها ويتبعها كلمات وظيفية أو علامات ترقيم، ونظروا في الكلمات التي تنطبق عليها هذه الصفة ومكوّنة من كلمتين أو ثلاث، ثم استخدموا مقياس تكرار الكلمة- عكس تكرار النصوص لترتيب هذه المصطلحات المحتملة، واختاروا الأعلى رتبة منها باعتبارها مصطلحات. أمّا الثبيتي والقحطاني (2022) فقد استخدموا قاعدة استرشادية مفادها: أنَّ المصطلحات المرغبة تتكون من المصطلحات المفردة التي كُشِف عنها باستخدام الطريقة الهجين الموضحة أعلاه.. فأَيّ تتابع للكلمات مكوّن من المصطلحات المفردة المعرّفة بالألف واللام، أو نفس المصطلحات بعد حذف الألف واللام من بدايتها يُعدُّ مصطلحاً مركباً. ثم قاما بتصفية النتائج باستخدام التكرار فأَيّ متتابعة يكون تكرارها أكثر من 10 مرات تُعدُّ مصطلحاً. ولكن عملهم حصر هذه المصطلحات في الأسماء فقط، ولم يتضمن الصفات.

ومن أهم نتائج الدراسات المتعلقة باستخلاص المصطلحات المفردة والمركبة، أنَّ هناك تناسباً طردياً بين دقة الاستخلاص، وارتفاع قيم المقاييس الإحصائية، والتكرار النسبي للمصطلح في المدونة المتخصصة، ولكن عدد المصطلحات المستخلصة سيكون أقل عدداً كلما ارتفعت هذه القيم. كما أنَّ دقة الاستخلاص تتناسب طردياً مع حجم المدونة المتخصصة.

وكان تقييم النتائج في الدراسات السابقة يعتمد على طريقتين. الأولى: الفحص المباشر من المتخصصين في المجال، حيث يقرر الخبير ما إذا كانت الكلمة أو المتتابعة مصطلحاً أم لا. أي أننا نتوصل لعدد المصطلحات الصحيحة من عدد المصطلحات المحتملة الناتجة من تطبيق أسلوب الاستخلاص. ولكننا لا نعرف أبداً كم من المصطلحات الصحيحة لم يوجد في القوائم المستخرجة. والطريقة الثانية تعتمد على مقارنة النتائج بالمصطلحات الموجودة في القواميس المتخصصة بالمجال. ولكن ممّا يؤثر على هذه الطريقة هو أنَّ المعاجم لا تحتوي جميع المصطلحات المستخدمة في المجال، وليست قابلة للمعالجة الآلية.

إنَّ دراسة الخواص اللغوية والإحصائية للمصطلحات -من خلال مدونة لغوية متخصصة موسّمة بالمصطلحات، سيمكننا من فهم أفضل القواعد الاسترشادية وأدقّها، واختيار المناسب منها، أو تطويره، أو تطوير قواعد استرشادية جديدة نعرف حدودها وإمكاناتها بدقة اعتماداً على الاستخدام الحقيقي للغة. وهذه المعرفة بالخواص اللغوية والإحصائية تمكّننا من حصر المصطلحات في نطاق ضيق، اعتماداً على معرفة الأنماط الأكثر شيوعاً، والخواص الأكثر وضوحاً للمصطلحات، ممّا يؤدي أيضاً إلى توفير الوقت والجهد اللذين لتقييم نتائج الأساليب المستخدمة في استخلاص المصطلحات.

3. عينة الدراسة

استخدمنا في هذه الدراسة مدوّنتين لغويتين. المدونة الأولى تختص بعلوم اللغة تتكوّن من نصوص محكّمة منشورة في مجلات، ومؤتمرات علمية، أو رسائل علمية جامعية، جمعت نصوصها من خلال التواصل المباشر مع الباحثين أو من مواقعهم على

الإنترنت، أو مواقع الجامعات أو المؤتمرات. أمّا المدونة الثانية فهي مدونة لغوية عامة جمعت من الصحف العربية الصادرة من تسعة عشر بلدًا عام 2012م (Al-Thubaity et al, 2013) وتغطي موضوعات هذه المدونة: السياسة والاقتصاد، والمجتمع، والرياضة، والدين، والثقافة، والصحة والعلوم. والغرض من استخدام هذه المدونة هو: الوقوف على الخواص الإحصائية المتعلقة بالمدونة المتخصصة مقارنة بالمدونة العامة.

وقد استخدمنا برنامج المشذب العربي¹ لتنقيح نصوص المدونتين حيث قمنا باختيار إزالة الفراغات الزائدة، والتطويل، وفصل الأرقام والرموز عن الكلمات وإزالة التشكيل. ويوضح الجدول (1) المعلومات الإحصائية الأساسية للمدونات المستخدمة في هذه الدراسة.

ثم استخدمنا أداة التوسيم بأقسام الكلام من موقع "ألب: أداة التحليل اللغوي العربية"² (Freihat at al, 2018) التي تحوي 56 وسمًا لأقسام الكلام لتوسيم المدونة المتخصصة. ومن مميزات هذا النظام أنه يقوم بتوسيم الكلمة والمتصلات المتصلة بها. فكلية مثل "وللتداولية" -على سبيل المثال- يقوم النظام بإعطائها الوسم "حرف عطف+حرف جر+أداة التعريف+اسم مفرد مؤنث" وهذا بالتالي يساعد على معرفة الوسم النحوي للكلمة، ومعلوماتها الصرفية، وما يتصل بها من مورفيمات أخرى مثل: حروف العطف، والجر، والضمائر.

جدول 1

المعلومات الإحصائية الأساسية للمدونات المستخدمة في الدراسة

المدونة	عدد النصوص	حجم المدونة	عدد الكلمات النوعية
المتخصصة (علوم اللغة)	40	180,449	27,508
العامة (الصحف)	2,910	2,523,140	202,564

وقد قام المؤلف الثاني لهذه الورقة -بحكم التخصص- بتوسيم المدونة المتخصصة بالمصطلحات -المتعلقة بالمجال اللغوي أو بغيره إن وردت- ثم راجعها المؤلف الأول، ثم ناقشا ما اختلفا فيه، ثم استقرّا على توسيم موحد. وقد استخدمنا الوسم "B" للدلالة على المصطلح المفرد المكوّن من كلمة واحدة، أو لبداية المصطلحات المركّبة المكوّنة من كلمتين وأكثر. واستخدمنا الوسم "A" للدلالة على عناصر المصطلح المركّب التي ليست في أوله. كما استخدمنا الرمز "O" للدلالة على أنّ الكلمة ليست مصطلحًا، أو من مكوّنات المصطلح المركّب. كما راجعنا التوسيم بأقسام الكلام وتصحيح ما وجدناه غير صحيح. وقد لاحظنا أنّ نظام التوسيم الآلي يخلط في أحيان كثيرة بين الاسم والصفة، كما أنّه يتبع الصفة جميع خصائص الموصوف في كلّ الحالات، وهذا غير صحيح على إطلاقه.

ولعلّ القارئ الكريم قد لاحظ أنّ عدد النصوص في المدونة المتخصصة قليل نسبيًا ممّا انعكس على حجم المدونة. وكان السبب الرئيسي لصغر حجم المدونة هو الوقت والجهد اللازمين لتوسيم المصطلحات، ومراجعة صحة التوسيم. وكذلك تصحيح وسوم أقسام الكلام للنظام المستخدم للتوسيم بأقسام الكلام. ولكن نتائج الدراسة القائمة عليها تعطي انطباعًا عامًا

وأولياً عن الخواص اللغوية والإحصائية للمصطلحات، خصوصاً الأنماط شائعة الاستخدام في المدونة المتخصصة. وتوضح الجداول (أ2)، (ب2)، (ج2)، (د2) أجزاء من بعض الجمل في المدونة المتخصصة بعد توسيمها بأقسام الكلام والمصطلحات.

جدول أ2

مثال على التوسيم بأقسام الكلام والمصطلحات

الكلمة	لمجموعة	القوانين	السياقية	ينتج	عنه	عدم	التمائل	بين	المنطوق	والسياق
وسم أقسام الكلام	حرف جر+ اسم مفرد مؤنث	أداة التعريف+ اسم جمع تكسير	أداة التعريف+ صفة مفرد مؤنث	فعل مضارع	حرف جر+ ضمير	اسم مفرد مذكر	أداة التعريف +اسم مفرد مذكر	ظرف مكان	أداة التعريف+ اسم مفرد مذكر	حرف عطف+ أداة التعريف+ اسم مفرد مذكر
وسم المصطلح	O	B	I	O	O	O	O	O	B	B

جدول ب2

مثال على التوسيم بأقسام الكلام والمصطلحات

الكلمة	وقد	درس	دي	سوسير	أفكار	علم	اللغة	لمجموعة	من	طلابه
وسم أقسام الكلام	حرف عطف+ حرف تحقيق	فعل ماضي	شخص	شخص	اسم جمع تكسير	اسم مفرد مذكر	أداة التعريف +اسم مفرد مؤنث	حرف جر+ اسم مفرد مؤنث	حرف جر	اسم جمع تكسير+ ضمير
وسم المصطلح	O	O	O	O	O	B	I	O	O	O

جدول ج2

مثال على التوسيم بأقسام الكلام والمصطلحات

الكلمة	انتقال	علم	اللغة	العام	من	المغرب	إلى	المشرق	العربي	:
وسم أقسام الكلام	اسم مفرد مذكر	اسم مفرد مذكر	أداة التعريف+ اسم مفرد مؤنث	أداة التعريف+ صفة مفرد مذكر	حرف جر	مكان	حرف جر	أداة التعريف+ اسم مفرد مذكر	أداة التعريف+ صفة مفرد مذكر	علامة تنقيط
وسم المصطلح	O	B	I	I	O	O	O	O	O	O

جدول 2

مثال على التوسيم بأقسام الكلام والمصطلحات

الكلمة	حاز	المنهج	التوليدي	التحويلي	على	انتشار	واسع	بين	المدارس	اللغوية
وسم أقسام الكلام	فعل ماضي	أداة التعريف+ اسم مفرد مذكر	أداة التعريف+ صفة مفرد مذكر	أداة التعريف+ صفة مفرد مذكر	حرف جر	اسم مفرد مذكر	صفة مفرد مذكر	ظرف مكان	أداة التعريف+ اسم جمع تكسير	أداة التعريف+ صفة مفرد مؤنث
وسم المصطلح	O	B	I	I	O	O	O	O	B	I

وقد بلغ عدد المصطلحات الموسّمة يدويًا في المدونة المتخصصة 3,138 مصطلحًا مفردًا ومركبًا، تكررت 7,618 مرة في المدوّنة المتخصصة. ويوضح الجدول (3) توزيع المصطلحات في المدوّنة المتخصصة مع أمثلة لها. وبحسب البيانات الموضحة في الجدول (3)، فإنّ تكرار المصطلح يرتبط بعدد الكلمات المكوّنة له، فالمصطلحات المفردة هي الأكثر تكرارًا، ثم يليها المصطلحات المكوّنة من كلمتين، ثم من ثلاث كلمات، ثم من أربع كلمات. بينما لا نجد أيّ ظهور للمصطلحات المكوّنة من خمس كلمات. وقد كانت نسبة المصطلحات المفردة من مجمل المصطلحات 50.7% ثم تلتها المصطلحات المركبة المكوّنة من كلمتين بنسبة بلغت 44.0%، ثم المصطلحات المركبة المكوّنة ثلاث كلمات بنسبة بلغت 5.1%. وكانت المصطلحات المركبة المكوّنة أربع كلمات هي الأقل نسبة حيث بلغت 0.2%. وهذا يؤشر إلى أنّ التركيز على استخلاص المصطلحات المفردة والمركبة من كلمتين يؤدّي إلى نجاح عملية الاستخلاص بنسبة تفوق 94%.

جدول 3

المعلومات الإحصائية الأساسية للمصطلحات الموسّمة يدويًا

عدد كلمات المصطلح	العدد	مجموع التكرار	المصطلحات العشر الأكثر تكرارًا
1	1,017	3,860	اللغة، التداولية، المعجم، اللسانيات، النص، الدلالة، الجملة، السياق، المعنى، الأصوات
2	1,818	3,353	اللغة العربية، علم الدلالة، علم الأصوات، علم اللغة، المعجم الإلكتروني، النحو الوظيفي، الفعل المضارع، الكفاءة اللغوية، اللغة الطبيعية، الفعل التواصل
3	285	386	علم اللغة العام، النظرية التوليدية التحويلية، علم اللغة التطبيقي، علم اللغة الحديث، المعاجم العربية التراثية، علم اللغة النفسي، المدرسة التوليدية التحويلية، اللغة العربية الفصحى، علم اللغة الاجتماعي، نظرية الفعل التواصل
4	18	19	الأفعال الإنجازية غير المباشرة، المعجم العلمي والفني المختص، النظريات الدلالية الذهنية النفسية، علاقة الألفاظ بالعالم الخارجي، نظام الهجاء الخاص بالصم، تحليل العلاقات الداخلية اللغوية، المبحث النقدي الأدبي المعاصر، فعل إنجازي غير مباشر، الفعل الماضي الثلاثي المضعف، الدراسات العلمية النظرية المعرفية

وتختلف دراستنا هذه عن دراسة شندول (2008) بأنها: (أ) قائمة على مدونة متنوعة الموضوعات من اللغة العربية المعاصرة ، أمّا دراسة شندول فاعتمدت على موضوع واحد هو النحو ونص واحد هو "المقتضب في النحو" من العربية الكلاسيكية، (ب) أنّ المصطلحات التي ندرسها تتناول مجالات متعددة في اللغة، أمّا دراسة شندول فدرست المصطلحات المتعلقة بالفعل، (ج) كما أنّ عدد المصطلحات المستخرجة في دراستنا أكبر من عدد المصطلحات التي شملتها دراسة شندول بنحو 53 ضعفًا، (د) ودراستنا نظرت إلى الخواصّ اللغوية والإحصائية للمصطلحات، أمّا دراسة شندول فركّزت على الخواصّ اللغوية . ولعلّ هذه الفروقات تعطي نتائج هذه الدراسة ثقة أكبر، كما أنها أثبتت بعضًا من نتائج دراسة شندول المتعلقة بدراستنا.

4. المنهجية المقترحة

دراستنا هذه دراسة إحصائية وصفية تقوم على دراسة الخواصّ اللغويّة والإحصائية للمصطلحات؛ لمعرفة حدود القواعد الاسترشادية، وإمكاناتها المستخدمة في استخلاص المصطلحات من المدوّنات اللُّغوية المتخصصة. وفيما يلي سنقوم بذكر الخطوات التي اتخذناها لدراسة الخواصّ اللُّغوية والإحصائية للمصطلحات.

1.4. الخواصّ اللُّغوية

- أ. استخراج أقسام الكلام المتعلقة بالمصطلحات المفردة والمركبة من المدونة الموسّمة.
 - ب. إحصاء تكرار وسوم أقسام الكلام المستخدمة في المصطلحات المفردة والمركبة.
 - ج. إحصاء الوسوم النحويّة للكلمات السابقة واللاحقة للمصطلحات.
- وقد استخدمنا في الخطوات السابقة برنامجًا كتبناه لهذا الغرض باستخدام لغة البرمجة "بايثون 3.7".

2.4. الخواصّ الإحصائية

- أ. مقارنة المئة كلمة الأكثر تكرارًا في المدونة المتخصصة مع المئة كلمة الأكثر تكرارًا في المدونة العامة. وقد استخدمنا لاستخراج الكلمات الأكثر شيوعًا برنامج "غواص" النسخة 34.6 المخصص لتحليل المدونات اللُّغوية من مدينة الملك عبدالعزيز للعلوم والتقنية (Al-Thubaity et al., 2013).
- ب. استخراج الكلمات المميزة للمدونة المتخصصة باستخدام مقياس الاحتمال اللوغاريتمي (لمزيد من المعلومات حول استخراج الكلمات المميزة انظر الثبيتي (2016)). وقد استخدمنا أيضًا برنامج "غواص" لهذا الغرض. والنتائج من هذه الخطوة هي قائمة بالكلمات مرتبة بحسب أهميتها في المدونة المتخصصة، كما يظهرها مقياس الاحتمال اللوغاريتمي. وقد استثنينا من القائمة الكلمات الوظيفية، وعلامات التقييم، والكلمات الأجنبية.
- ج. اختيار الكلمات الأكثر تميّزًا في القائمة المستخرجة باستخدام شرطين رئيسيين هما:

— أن تكون قيمة الاحتمال اللوغاريتمي للكلمات المستخرجة أعلى من حد معين. ويعتمد هذا الحد على مستوى الثقة الإحصائي للنتائج المستخرجة. وقد اخترنا أربعة قيم هي 3.841 و 5.024 و 6.365 و 10.828. وهذه القيم تقابل مستويات الثقة 95% و 98% و 99% و 99.99%. ويعتبر مستوى الثقة 95% هو أقل المستويات المقبولة في دراسات العلوم الإنسانية.

— أن يكون التكرار النسبي للكلمة في المدونة المتخصصة أكبر بضعفين من تكرارها النسبي في المدونة العامة.

د. مقارنة المصطلحات الموسمة يدويًا في المدونة المتخصصة مع قوائم الكلمات المميزة.

هـ. كررنا الخطوات ب، ج، د للمصطلحات المركبة.

وسوف نركز في هذه الدراسة على المصطلحات المفردة، والمصطلحات المركبة المكونة من كلمتين؛ لأنها هي الأكثر استخدامًا وشيوعًا في المدونة، حيث بلغت نسبتها كما هو موضح في جدول الجدول (3) أكثر من 94.7%. كما أن عدد الصفحات المتاح لنشر البحث لا يسمح بدراسة كل التراكمات الممكنة للمصطلحات.

5. الخواص اللغوية للمصطلحات

كما يتضح من الجدول (2) فإن نظام التوسيم بأقسام الكلام يقوم بوضع الوسم الكامل لكل المورفيمات التي تتكون منها الكلمة، مع وضع علامة "+" للفصل بين وسوم كل مورفيم بالكلمة. وفي دراستنا هذه سوف نهمل المتصلات التي تأتي في أول المصطلح كحروف الجر مثل "الباء واللام"، وحروف العطف مثل "الواو والفاء"، وغير ذلك ما عدا أداة التعريف "الألف واللام"؛ لأن حذفها يغير من تصنيف الكلمة، فهي جزء لا يمكن فصله عنها مثلها مثل "الألف والتاء" في جمع المؤنث السالم على سبيل المثال. وسنهمل الضمائر المتصلة التي تأتي في آخر المصطلحات. وسوف نبقي المتصلات في حال كانت في وسط المصطلح مثل: "التعريف بالأمثلة"؛ لأن معنى المصطلح لا يتم إلا بها، وكذلك لو كان الضمير في آخر المصطلح مثل: "المسند إليه" لذات السبب. ومن ثم فإن مصطلحًا مثل: "والبناء للمفعول" سيعد مصطلحًا مركبًا من كلمتين عند وسمه بالمصطلحات، وتوسم الكلمة الأولى منه (والبناء) - بحسب نظام التوسيم - بالوسم "حرف عطف + أداة التعريف + اسم مفرد مذكر". وتوسم الكلمة الثانية منه (للمفعول) بالوسم "حرف جر + أداة التعريف + اسم مفرد مذكر". ولكن عند دراسة التركيب النحوي لهذا المصطلح سوف نعهده مكوّنًا من الوسوم الثلاثة التالية: "أداة التعريف + اسم مفرد مذكر" و "حرف جر" و "أداة التعريف + اسم مفرد مذكر". فحذف حرف الجر من بداية الكلمة الثانية سيجعل التركيب غير مستقيم المعنى. وفي مثل هذه الحالات سنعهده مكوّنًا من ثلاث كلمات وليس اثنتين. وقس على هذا في الحالات المشابهة.

1.5 المصطلحات المفردة

يوضح الجدول (4) وسوم أقسام الكلام المقابلة للمصطلحات المفردة. وكما تشير بيانات التوسيم بأقسام الكلام فإن المصطلحات المفردة أتت تحت قسم رئيسي واحد من أقسام الكلام هو الاسم. وكان عدد المصطلحات المعرّفة بالألف واللام

3509 مصطلحًا شكّلت 90.9% من تكرار المصطلحات المفردة. بينما بلغ عدد المصطلحات المفردة غير المعرّفة بالألف واللام 351 مصطلحًا شكّلت 9.1% من تكرار المصطلحات المفردة، وقد وجدنا منها 270 مصطلحًا (تشكل 77.0% من تكرار المصطلحات المفردة غير المعرّفة بالألف واللام) لها ما يقابلها من المصطلحات المعرّفة بالألف واللام. وهذا يعزّز القاعدة الاسترشادية التي استخدمها الثبيتي والقحطاني (2020) في استخلاص المصطلحات المفردة من المدوّنات المتخصصة، والتي تقول: إنّ المصطلحات المفردة عبارة عن أسماء معرّفة بألف. ويوضح الجدول (5) الكلمات العشر الأكثر تكرارًا في المصطلحات المفردة غير المعرّفة بالألف واللام، وما يقابلها من مصطلحات مفردة معرّفة بالألف واللام.

جدول 4

وسوم أقسام الكلام المقابلة للمصطلحات المفردة

الوسم	التكرار	النسبة	أمثلة
أداة التعريف+اسم مفرد مذكر	1431	37.1%	المعجم، النص، السياق، المعنى، النحو، الاتصال، اللفظ، الجمل، السبك، الفعل
أداة التعريف+اسم مفرد مؤنث	1418	36.7%	اللغة، التداولية، الدلالة، الجملة، البنيوية، العلامة، السيميوطيقا، الكلمة، الإحالة، المدونة
أداة التعريف+اسم جمع مؤنث سالم	382	9.9%	اللسانيات، اللغات، العلامات، الأبجديات، المتلازمات، الإشارات، المسكوكات، اللغويات، المصطلحات، الكلمات
أداة التعريف+اسم جمع تكسير	247	6.4%	الأصوات، المعاجم، الجمل، القواعد، الروابط، العناصر، الضمائر، الألفاظ، النصوص، الصوائت
اسم مفرد مذكر	138	3.6%	نص، مصطلح، معجم، معنى، نظام، دال، سياق، علم، حرف، منطوق
اسم مفرد مؤنث	120	3.1%	لغة، تداولية، جملة، دلالة، كلمة، نظرية، إحالة، مدونة، علامة، حركة
اسم جمع تكسير	44	1.1%	أصوات، صوائت، معاجم، عناصر، أسماء، كلم، معاني، مبادئ، قوانين، أنظمة
اسم جمع مؤنث سالم	36	0.9%	لغات، كلمات، تفسيرات، مورفيمات، تداوليات، رسومات، علامات، وحدات، تأويلات، مصطلحات
أداة التعريف+اسم جمع مذكر سالم	31	0.8%	السيميوطيقيين، اللغويين، السيميوطيقيين، البنيويين، البنيويين، التداوليين، اللسانيين، السيميولوجيين، السيميوطيقيين، السيميوطيقيين، اللسانيون
اسم مثنى مؤنث	11	0.3%	لغتين، لغتان
اسم جمع مذكر سالم	2	0.1%	تداوليين
المجموع	3860	100%	

جدول 5

أمثلة للمصطلحات المفردة غير المعرّفة بالألف واللام، وما يقابلها من مصطلحات مفردة معرّفة بالألف واللام.

مصطلح مفرد غير معرّف بالألف واللام.		مصطلح مفرد معرّف بالألف واللام.	
لغة	28	اللغة	242
لغات	10	اللغات	34
تداولية	8	التداولية	209
لغتين	7	اللغة	242
نص	7	النص	95
جملة	6	الجملة	65
دلالة	6	الدلالة	83
مصطلح	6	المصطلح	10
صوائت	5	الصوائت	4
كلمة	5	الكلمة	21

وقد توزعت المصطلحات المفردة بين الإفراد والجمع والتثنية حيث بلغت نسبة المصطلحات التي انضوت تحت وسوم الاسم المفرد 80.5%، والمصطلحات التي كانت تحت وسوم الجمع 19.2%. بينما كانت المصطلحات المفردة التي تنتهي لوسوم الاسم المثنى نادرة جداً بنسبة 0.3%. وبالنسبة للتذكير والتأنيث في المصطلحات المفردة، فقد كانت المصطلحات الواردة بصيغة التأنيث هي الأعلى بنسبة 50.0% تليها المصطلحات المذكرة بنسبة 42.4%، ثم أخيراً المصطلحات التي لا يمكن الحكم على تذكيرها أو تأنيثها من خلال الوسم النحوي الآلي مباشرة - أي جموع التكسير - فقد كانت نسبتها 7.5%.

2.5 المصطلحات المركبة من كلمتين

توضح بيانات التوسيم بأقسام الكلام للمصطلحات المركبة من كلمتين أنّها انضوت تحت 45 تركيباً لأقسام الكلام. ويوضح الجدول (6) وسوم أقسام الكلام الأكثر شيوعاً للمصطلحات المركبة من كلمتين. وقد شكلت هذه التراكيب 95.1% من تكرار هذه المصطلحات المركبة في المدوّنات المتخصصة. وتشير بيانات التوسيم بأقسام الكلام إلى أنّ جميع الكلمات الأولى المكوّنة للمصطلح المركّب المكوّن من كلمتين عبارة عن أسماء (99.5%) ماعداً أربعة تراكيب بدأت بحرفي النفي "غير" و "لا" كان مجموع تكرارها 18 مرة (0.5%) مثل "لا صائتي" و "غير نطعي". وقد كانت نسبة الأسماء في الكلمة الثانية 21.3% ونسبة الصفات 78.7%.

وكانت نسبة الأسماء المعرّفة بالألف واللام في الكلمة الأولى للمصطلح المركّب المكوّن من كلمتين 70.7% وغير المعرّفة بالألف واللام 28.7%. وكانت نسبة الأسماء والصفات المعرّفة بالألف واللام في الكلمة الثانية من المصطلحات المركّبة المكوّنة من كلمتين 91.1% وغير المعرّفة 8.9%. ويوضح الجدول (7) توزيع مكوّنات المصطلح المركّب من كلمتين بالنسبة للتعريف بالألف واللام والاسمية والصفة. وكما توضح البيانات فإنّ أكثر التراكيب شيوعاً للمصطلح من كلمتين هي "اسم معرّف - صفة معرفة" (70.7%)، ثم "اسم غير معرّف - اسم معرّف" (20.0%).

جدول 6

أقسام الكلام الأكثر شيوعاً للمصطلحات المركبة من كلمتين

الأولى	الثانية	التكرار	النسبة	أمثلة
أداة التعريف+اسم مفرد مذكر	أداة التعريف+صفة مفرد مذكر	865	25.8%	المعجم الإلكتروني، النحو الوظيفي، الفعل المضارع، الفعل التواصلي، المعجم العربي، البحث المعجمي، الحقل الدلالي، النص الأدبي، التحليل الدلالي، الصوت اللغوي
أداة التعريف+اسم مفرد مؤنث	أداة التعريف+صفة مفرد مؤنث	823	24.5%	اللغة العربية، الكفاءة اللغوية، اللغة الطبيعية، المقاربة التداولية، الألسنة الطبيعية، الدراسة التداولية، البنية السطحية، المدونة اللغوية، الفنولوجيا التوليدية، الظاهرة اللغوية
أداة التعريف+اسم جمع تكسير	أداة التعريف+صفة مفرد مؤنث	364	10.9%	الأصوات اللغوية، المعاجم العربية، المعاجم الإلكترونية، التراكيب النحوية، الأفعال الكلامية، المعاجم الحديثة، المداخل المعجمية، التعابير الاصطلاحية، المعاجم اللغوية، الظواهر الصوتية، العناصر الإشارية
أداة التعريف+اسم جمع مؤنث سالم	أداة التعريف+صفة مفرد مؤنث	307	9.2%	الدراسات الصوتية، الوحدات المعجمية، الأدوات النحوية، اللسانيات الحديثة، اللغات الطبيعية، اللسانيات الوظيفية، الافتراضات المسبقة، الدراسات اللغوية، البنيات الذهنية، اللغويات التطبيقية
اسم مفرد مذكر	أداة التعريف+اسم مفرد مؤنث	184	5.5%	علم الدلالة، علم اللغة، اسم الإشارة، نحو الجملة، فقه اللغة، اكتساب اللغة، علم التربية، استعمال اللغة، علم الألسنة، اختبار الكفاءة،
اسم مفرد مذكر	أداة التعريف+اسم مفرد مذكر	159	4.7%	نحو النص، تحليل الخطاب، علم الاتصال، سياق الحال، اسم الفاعل، جسم اللسان، علم الصرف، علم النحو، مورفيم الإعراب، سياق الموقف
اسم مفرد مذكر	صفة مفرد مذكر	88	2.6%	مركب اسمي، معجم عربي، فعل تواصلي، مركب فعلي، مدخل معجمي، معجم إلكتروني، حقل دلالي، حقلاً دلالياً، معجم أساسي، تحليلاً نحوياً
اسم مفرد مؤنث	صفة مفرد مؤنث	80	2.4%	وحدة لغوية، مدونة لغوية، أداة نحوية، لغة طبيعية، بنية سطحية، ظاهرة لغوية، ظاهرة أسلوبية، حركة طويلة، نظرية لغوية، معجمية عامة
اسم مفرد مذكر	أداة التعريف+اسم جمع تكسير	66	2.0%	علم الأصوات، مذهب الذرائع، جهر الأصوات، تنجيم الأبراج، علم المعاني، تحليل النصوص
اسم مفرد مؤنث	أداة التعريف+اسم مفرد مذكر	61	1.8%	مهارة الاستماع، فلسفة التواصل، نظرية النظم، نظرية التواصل، حوسبة المعجم، نظرية السياق، دلالة اللفظ، دلالة المنطوق، ميتافيزيقيا النص، مقصدية النص،
اسم جمع تكسير	أداة التعريف+اسم مفرد مؤنث	53	1.6%	أصوات اللغة، أصوات العربية، قواعد اللغة، أسماء الإشارة، علوم اللغة، قواعد الزيادة، صيغ المبالغة، علوم المناظرة، فلاسفة اللغة، مصادر اللغة
اسم مفرد مؤنث	أداة التعريف+اسم مفرد مؤنث	51	1.5%	مهارات القراءة، مهارة الكتابة، نظرية المناسبة، بنية اللغة، نظرية المحاكاة، صيغة المبالغة، بنية العبارة، دلالة الجملة، نظرية الدلالة، دلالة الصيغة،

اسم جمع تكسير	صفة مفرد مؤنث	49	1.5%	جمل نحوية، معايير تفسيرية، مداخل معجمية، مكانز مصطلحية، قواعد معجمية، معاجم مصطلحية، صوائت مجهورة، أصوات أنفية، أصوات مغورة، أصوات مطبقة
اسم جمع مؤنث سالم	صفة مفرد مؤنث	38	1.1%	وحدات معجمية، لسانيات إنسانية، علامات لغوية، مصطلحات علمية، مصطلحات أعجمية، احتياطات منهجية، وحدات دلالية، سياقات مقالية، سياقات حالية، تخطيطات قواعدية

جدول 7

توزيع مكونات المصطلح المركب المكون من كلمتين بالنسبة للتعريف والاسمية والوصفية

الكلمة الأولى	الكلمة الثانية	النسبة	أمثلة
اسم معرف بالألف واللام	اسم معرف بالألف واللام	0.0%	
	صفة معرفة بالألف واللام	70.7%	المعجم الإلكتروني، النحو الوظيفي، الكفاءة اللغوية، اللغة الطبيعية، التراكيب النحوية، الأفعال الكلامية، اللسانيات الوظيفية، الافتراضات المسبقة، البنيات الذهنية، اللغويات التطبيقية
	اسم غير معرف	0.0%	
	صفة غير معرفة	0.0%	
اسم غير معرف بالألف واللام	اسم معرف بالألف واللام	20.3%	نحو الجملة، فقه اللغة، تحليل الخطاب، علم الاتصال، علم الأصوات، مذهب الذرائع، نظرية النظم، نظرية التواصل، نظرية المحاكاة، صيغة المبالغة
	صفة معرفة بالألف واللام	0.0%	
	اسم غير معرف بالألف واللام	0.5%	جهاز تسجيل، حرف عطف، فعل ماض،
	صفة غير معرفة بالألف واللام	7.9%	معجم استعمال، تسجيل صوتي، نموذج لغوي، وسيط لغوي، مركب اسمي، وحدة لغوية، مدونة لغوية، أداة نحوية، لغة طبيعية، بنية سطحية
حرف نفي	اسم معرف بالألف واللام	0.1%	غير الرنانة، غير التصريفية، غير التحليلية، غير اللساني
	صفة معرفة بالألف واللام	0.0%	
	اسم غير معرف بالألف واللام	0.4%	لا صائتي، لا صامتي، غير نطعي، غير أمامي، غير مرتفع، غير خلفي، غير مستدير، غير موزع، غير استمرارية، غير لغوية
	صفة غير معرفة بالألف واللام	0.0%	

ومن حيث التذكير والتأنيث، كانت نسبة الأسماء المذكورة في الكلمة الأولى من المصطلح المركب من كلمتين 42.0%، ونسبة الكلمات المؤنثة 41.9% و 16.1% كانت لكلمات لا يمكن الحكم على تذكيرها أو تأنيثها (جمع التكسير وحروف النفي). وكانت نسبة الأسماء والصفات المذكورة في الكلمة الثانية من المصطلح المركب من كلمتين 37.8%، والمؤنثة 59.5%، وكان الباقي

(2.7%) لجموع التكسير. ويوضح الجدول (8) توزيع تراكيب المصطلح المركب من كلمتين على التذكير والتأنيث. وكما توضح البيانات في الجدول 8، فإنّ الأنماط الأكثر شيوعاً هي التي يكون فيها توافق في جنس الكلمة الأولى والثانية. وقد شكّل الاسم المفرد من الكلمة الأولى للمصطلح المركب المكون من كلمتين 72.5%، والجمع 26.9%. في حين شكّل المثنى نسبة ضئيلة بلغت 0.1%. أمّا الكلمة الثانية من المصطلح المركب فقد كانت فيها نسبة الأسماء والصفات المفردة 96.3%، والجمع 3.6%. وأمّا المثنى، فقد كان لحالة واحدة (0.03%). ويوضح الجدول (9) توزيع تراكيب المصطلح المركب المكون من كلمتين على الإفراد والتثنية والجمع.

جدول 8

توزيع تراكيب المصطلح المركب من كلمتين على التذكير والتأنيث.

الكلمة الأولى	الكلمة الثانية	النسبة	أمثلة
مذكر	مذكر	33.8%	البحث المعجمي، الحقل الدلالي، التحليل الدلالي، المعجم الذهني، البحث الدلالي، علم سياق الحال، اسم الفاعل، جسم اللسان، منهج التنوع، علم الصرف،
	مؤنث	6.2%	علم الدلالة، علم اللغة، اسم الإشارة، اكتساب اللغة، اختبار الكفاءة، الاسم النكرة، علم العلامات، علم الصوتيات، نظام العلامات، علم المفردات، نظام الاشارات
	غير محدد	2.0%	علم الأصوات، مذهب الذرائع، علم الفونيمات، جهر الأصوات، تحليل النصوص، نحو النصوص، علم الأصول، تناسل الوقائع، منطق القضايا،
مؤنث	مذكر	2.7%	مهارة الاستماع، فلسفة التواصل، حوسبة المعجم، نظرية السياق، دلالة اللفظ، مقصدية النص، جملة التعجب، أداة استفهام، دلالة التزام، دلالة رابط،
	مؤنث	39.0%	المقاربة التداولية، الدراسة التداولية، البنية السطحية، المدونة اللغوية، الفنولوجيا التوليدية، المقاربة الأدنوية، اللسانيات الوظيفية، البنيات الذهنية، بنية العبارة، دلالة الصيغة،
	غير محدد	0.3%	صناعة المعاجم، فلسفة العلوم، دراسة الأصوات، لغة النصوص، قاعدة البيانات، دلالة الألفاظ، الحركات القصار
غير محدد	مذكر	1.4%	حروف الجر، قواعد الحذف، قواعد التعويض، أفعال الكلام، معاجم الترادف، أصول الفقه، قواعد الاشتقاق، أدوات الجر، ضمائر المتكلم، جمل الاستفهام
	مؤنث	14.4%	التراكيب النحوية، الأفعال الكلامية، المداخل المعجمية، التعابير الاصطلاحية، العناصر الإشارية، الأصوات الصامتة، الأصوات المجهورة، الأصوات المهموسة، القواعد الوجيهية، الرموز الصوتية
	غير محدد	0.4%	مخارج الحروف، معاجم الطلاب، أسماء الأعلام، أقسام الكلم، معاجم المعاني، أسماء الأجناس، الأسماء الأعلام،

3.5 الكلمات السابقة واللاحقة للمصطلح

إحدى الطرق التي استخدمت في التعرف على المصطلحات اعتمدت على قاعدة استرشادية مفادها: أنَّ الكلمات السابقة واللاحقة للمصطلحات يغلب عليها أن تكون كلمات وظيفية، أو علامات ترقيم. وبالنظر لوسوم الكلمات السابقة للمصطلح المفرد وجدنا أنَّ 34.8% من هذه الكلمات عبارة عن كلمات محتوى (اسم وصفة وفعل) والبقية (65.2%) كانت كلمات وظيفية، أو علامات ترقيم تنتهي لقوائم مغلقة يمكن حصرها بسهولة مثل: حروف المعاني، وعلامات الترقيم، ومحددات الكمية (مثل: كل، وبعض، وبضعة...). وكانت أكثر كلمات المحتوى وروداً قبل المصطلح المفرد لها علاقة بالبحث العلمي مثل: "دراسة، مفهوم"، "مجال"، "حق"، "موضوع". ومثل هذه الكلمات يمكن أن يحصر الأكثر تكراراً منها، وتستخدم ضمن قوائم للكشف عن المصطلحات. واتضح علاقة الكلمات الوظيفية وعلامات الترقيم بالمصطلح المفرد في الوسوم المتعلقة بالكلمات التي ترد بعده، حيث كانت نسبتها 77.7%، بينما بلغت نسبة كلمات المحتوى 22.3% مثل "بوصفها"، "بوصفه"، "فرع"، "دراسة"، "جزء".

جدول 9

توزيع تراكيب المصطلح المركب المكون من كلمتين على الإفراد والتثنية والجمع.

الكلمة الأولى	الكلمة الثانية	النسبة	أمثلة
مفرد	مفرد	69.61%	الفعل الإنجازي، التيار الهوائي، المعنى الدلالي، السياق التداولي، الجملة الفعلية، البنية التصورية، الدلالة المعجمية، الدلالة النحوية، العبارة المنطقية، القدرة المعجمية.
	مثنى	0.00%	-
	جمع	2.92%	علم المعاني، تحليل النصوص، علم الأصوات، علم العلامات، علم الصوتيات، نظام العلامات، علم المفردات، نظام الاشارات، قاعدة البيانات، دلالة الألفاظ
مثنى	مفرد	0.03%	صيغتا المبالغة
	مثنى	0.03%	الوترين الصوتيين
	جمع	0.00%	-
جمع	مفرد	26.16%	الأصوات الأمامية، الأصوات الخلفية، التعابير السياقية، الأوزان الصرفية، الشواهد اللفظية، سياقات مقالية، انحرافات أسلوبية، ملابسات عرفية، مصوتات طويلة، استلزمات وضعية
	مثنى	0.00%	-
	جمع	0.72%	مخارج الحروف، معاجم الطلاب، أسماء الأعلام، أقسام الكلم، السيميوطيقيين البنيويين، السيميوطيقيون الاجتماعيون، حقائق الموجودات، معاجم المصطلحات، الحركات القصار
حرف نفي	مفرد	0.54%	غير الرنانة، غير التصريفية، غير التحليلية، غير اللساني، لا صائتي، لا صامتي، غير نطعي، غير أمامي، غير خلفي، غير مستدير
	مثنى	0.00%	
	جمع	0.00%	

وتشابه توزيع أقسام الكلام السابقة واللاحقة للمصطلح المركب المكون من كلمتين مع توزيعها بالنسبة للمصطلح المفرد في أغلب الحالات (انظر على سبيل المثال توزيع الاسم والصفة). وقد كانت نسبة كلمات المحتوى التي تسبق المصطلحات المركبة 36.4%، ونسبة الكلمات الوظيفية وعلامات الترقيم 63.6%. وبالنسبة لكلمات المحتوى التي ترد بعد المصطلح المركب من كلمتين فقد كانت نسبتها 19.6%، وكانت نسبة الكلمات الوظيفية وعلامات الترقيم 80.4%. كما تشابهت كلمات المحتوى التي تسبق المصطلح المركب أو تليه مع تلك التي تسبق المصطلح المفرد أو تليه.

يوضح الجدول (10) توزيع وسوم أقسام الكلام للكلمات السابقة واللاحقة للمصطلح المفرد والمركب من كلمتين. وكما يتضح من البيانات الموضحة في الجدول 10 وما وجدناه من تشابه في توزيع أقسام الكلام للكلمات السابقة واللاحقة للمصطلح المفرد والمركب، وغلبة أن تكون الكلمات الوظيفية وعلامات الترقيم سابقة ولاحقة للمصطلحات، دليل يؤكّد صلاحية القاعدة الاسترشادية التي استخدمها الثبيتي وزملاؤه (Al-Thubaity et al, 2014).

جدول 10

توزيع وسوم أقسام الكلام للكلمات السابقة واللاحقة للمصطلح المفرد والمركب

المصطلح المركب من كلمتين		المصطلح المفرد		الوسم
بعد	قبل	بعد	قبل	
14.1%	27.6%	16.9%	27.4%	اسم وصفة
5.5%	8.8%	5.4%	7.3%	فعل
3.7%	1.6%	3.6%	1.9%	ظرف
30.7%	43.0%	32.8%	43.4%	حرف
40.2%	12.3%	34.8%	10.1%	علامة تنقيط
1.6%	3.3%	2.0%	4.7%	ضمير
0.2%	1.5%	0.1%	3.3%	اسم إشارة
0.1%	1.2%	0.4%	1.3%	محدد كمية
0.3%	0.4%	0.2%	0.4%	عدد
0.1%	0.3%	0.3%	0.1%	صفة عددية
3.6%	0.0%	3.5%	0.03%	اسم موصول

6. الخواصّ الإحصائيّة للمصطلحات.

إن استخدام خاصية اختلاف التوزيع الإحصائي لكلمات المدوّنة المتخصصة عن توزيعها في المدوّنة العامة خطوة أوليّة ومهمة للمساعدة في الكشف المصطلحات. ولا يزعم أحد أنها كافية أو أنّ الكلمات التي يختلف توزيعها الإحصائي في المدوّنة المتخصصة تُعدّ مصطلحات، بل تُعدّ كلمات وتعايير محتملة بشكل كبير لتكوّن مصطلحات. وتكون عندئذ مهمة المتخصص الخبير النظر

فيها ، وتقدير ما هو مصطلح منها. ولعلّ استخدام التوزيعات الإحصائية لأقسام الكلام الموضّحة سابقاً في هذا البحث، تكون عوناً وسنداً كبيراً واختصاراً للوقت والجهد في تحديد نطاق الكلمات المميزة المحتملة. وتعدّ الكلمات الأكثر تكراراً في أيّ مدوّنة لغوية بصمة لغوية تحدّد طبيعة المدوّنة وموضوعها. الجدول (11) والجدول (12) يوضحان الكلمات المئة الأكثر تكراراً في المدوّنة العامة والمدونة المتخصصة على التوالي. وكانت أهم نقاط الاختلاف بين الكلمات الأكثر تكراراً الموضحة في الجدولين ولها علاقة بموضوع دراستنا هي:

أ. أنّ نسبة الكلمات المئة الأكثر تكراراً في المدوّنة المتخصصة (29.45%) أكبر من نسبتها في المدوّنة العامة (23.86%) بمقدار 5.6%.

ب. كان عدد كلمات المحتوى في الكلمات المئة الأكثر تكراراً للمدوّنة المتخصصة أكبر منه في الكلمات المئة الأكثر تكراراً في المدوّنة العامة بمقدار 16 كلمة.

ج. أنّ كلّ كلمات المحتوى الأكثر تكراراً في المدوّنة المتخصصة -وعدها 39 كلمة- عبارة عن مصطلحات مثل: "اللغة"، و "المعجم"، والدلالة"، أو أجزاء من مصطلحات مثل "علم"، و "نظرية"، ما عدا ست كلمات عبارة عن تنوعات للفعل كان ولفظ الجلالة "الله" و "مثل". بينما كان عدد كلمات المحتوى في المدونة العامة 23 كلمة فقط لها دلالات سياسية أو اقتصادية.

د. نلاحظ في الكلمات الأكثر تكراراً في المدوّنة المتخصصة استخداماً واضحاً للأحرف مثل: "و"، والتي استخدمت في أغلب حالتها كسابق لكلمات وضعت بين قوسين أو علامتي تنصيص، والحرف "ص" الذي يرمز لرقم الصفحة، والحرف "ب" الذي كان حاله مثل الحرف "و"، والحرف "د" الذي استخدم كلقب علمي قبل اسم الباحث ومثله كذلك الحرف "أ" الذي استخدم للدلالة على أستاذ. أمّا الحرف "م" فقد كان للدلالة على السنة الميلادية. في حين نجد في المدوّنة العامة حرفاً واحداً فقط هو الواو. وقد كان استخدامه في نصوص المدوّنة العامة مماثلاً لاستخدامه في المدوّنة المتخصصة. هـ. كما نلاحظ في الكلمات الأكثر تكراراً في المدوّنة المتخصصة ورود خطأ إملائي وهو "في" (من غير نقط الياء)، والصحيح هو حرف الجر "في". وقد يكون السبب في مثل هذا الخطأ هو ترميز النصوص في مواقع الإنترنت التي جلبنا منها النصوص، أو الانتقال من ترميز إلى آخر أثناء معالجة نصوص المدونة المتخصصة.

إنّ هذه الاختلافات الواضحة جدّاً بين الكلمات الأكثر تكراراً في المدوّنة المتخصصة والمدوّنة العامة تدل على أنّ التوزيع الإحصائي للمصطلحات المفردة والمركبة -لأنّ كثيراً من المصطلحات المفردة أتت كمكوّن للمصطلحات المركبة مثل "المعجم اللغوية"- يختلف بشكل كبير بين المدوّنتين اللغويتين المتخصصة والعامة.

جدول 11

المئة كلمة الأكثر تكرارًا في المدوّنّة العامة

النسبة	الكلمات	
11.85%	في، من، على، أن، إلى، التي، عن، ما، الذي، لا	10
3.10%	مع، هذا، الله، أو، هذه، ان، و، بين، كان، الى	20
2.11%	كل، ذلك، بعد، كما، خلال، لم، هو، إن، حيث، عليه	30
1.42%	وفي، ولا، وهو، قد، حتى، قبل، له، ومن، العام، كانت	40
1.22%	هي، قال، بن، أنه، أي، غير، إلا، بعض، علي، فيها	50
1.03%	وقد، وقال، محمد، بها، به، عام، ثم، العالم، الذين، فيه	60
0.92%	أكثر، يكون، العربية، اليوم، يمكن، وهي، هناك، لها، لكن، رئيس	70
0.80%	تلك، مثل، العربي، إذا، منذ، فإن، منها، بشكل، بل، الأول	80
0.74%	دون، عند، ولكن، عبد، الناس، العمل، تم، فقد، بأن، أيضا	90
0.68%	الرئيس، عليها، خاصة، تكون، حول، أخرى، مليون، وما، عدد، أما	100
23.86%	المجموع	

جدول 12

المئة كلمة الأكثر تكرارًا في المدوّنّة المتخصصة

النسبة	الكلمات	
12.49%	في، من، على، أن، إلى، و، أو، اللغة، التي، ما	10
4.67%	هذا، هذه، لا، عن، العربية، بين، الذي، هو، علم، ذلك	20
2.61%	في، اللغوية، المعنى، كما، إن، المعجم، الدلالة، ص، التداولية، يمكن	30
2.09%	مع، غير، <u>كان</u> ، كل، ومن، هي، عند، النص، الأصوات، ب	40
1.71%	وهو، وقد، م، أي، فإن، ثم، اللغوي، قد، لم، مثل	50
1.46%	به، البحث، أما، <u>يكون</u> ، بها، بعض، اللسانيات، إلا، دراسة، الكلمات	60
1.25%	حيث، المعجمية، أنه، نظرية، إذا، خلال، عليه، أ، فيه، بل	70
1.16%	معنى، منها، د، له، <u>تكون</u> ، وهي، أخرى، <u>الله</u> ، النظرية، فيها	80
1.04%	الخطاب، أكثر، ولا، لها، الكلمة، السياق، كانت، فقد، المعاجم، العربي	90
0.97%	المتكلم، تلك، اللغات، الجملة، الفعل، هنا، الدراسات، النحو، معجم، لغة	100
29.45%	المجموع	

ولو نظرنا للمتتابعات الخمسين الأكثر تكراراً في المدونة المتخصصة، التي لا تحتوي أيّ كلمات وظيفية عدا كلمتي "لا" و"غير" كما يوضحها الجدول (13)، لوجدنا أنّ 70% منها عبارة عن مصطلحات، بل إنّ المصطلحات تتبوأ الصدارة في هذه القائمة (انظر السطر الأول جدول 13). وهذا مؤشر آخر يدعم اختلاف التوزيع الإحصائي ودلالته في الكشف المصطلحات).

جدول 13

المتتابعات الخمسين الأكثر تكراراً في المدونة المتخصصة.

المتتابعات من كلمتين	
اللغة العربية، علم اللغة، علم الدلالة، علم الأصوات، البحث اللغوي، قوله تعالى، باللغة العربية، تعليم اللغة، امتحان الكفاءة	10
سبيل المثال، للغة العربية، فروق دالة، المعجم العربي، دالة إحصائية، القرآن الكريم، الكفاءة باللغة، توجد فروق، النحو الوظيفي	20
مرجع سابق، لسان العرب، آفاق جديدة، المعجم الإلكتروني، المعاجم العربية، المقاربة التداولية، الفعل التواصل، الألسنة الطبيعية، الفعل المضارع	30
تفسير أبي، قسم اللغة، اللفظ والمعنى، الدرجة الكلية، تحليل الخطاب، اللغة العام، سياقه القرآني، الكفاءة اللغوية، أهمية علم	40
جهاز النطق، المعجمية العربية، الأصوات اللغوية، المعنى الأصلي، تحصيل الطلاب، العربية للناطقين، الرسم البياني، اللسانيات والتداولية، تعلم اللغة	50

والنظر في الكلمات المميزة للمدونة المتخصصة يوضح أهمية اختلاف التوزيع الإحصائي للكشف عن المصطلحات المفردة أو المركبة. يوضح الجدول (14) الكلمات المئة الأكثر تميزاً في المدونة المتخصصة. وتشير البيانات الموضحة في الجدول (14) إلى: أ. 96 من الكلمات المميزة إمّا مصطلحات مفردة، أو أسماء وصفات يمكن أن تكون الكلمة الثانية للمصطلح المركب. ب. البقية (4 كلمات) كان اثنان منها أسماء أعلام في المجال اللغوي وهما "هابرماس" و"سوسير"، وكلمة تستخدم للإحالة في الكتابة العلمية وهي: "انظر"، والرابعة كلمة "دار" لتحديد دار النشر في المراجع العلمية.

وهذا مؤشر واضح على أنّ الكتابة العلمية أو المتخصصة تعتمد على دوران المصطلحات وتكرارها. ولو نظرنا للمصطلحات المفردة الأكثر تكراراً على سبيل المثال وهي: اللغة، التداولية، المعجم، اللسانيات، النص، الدلالة، الجملة، السياق، المعنى، الأصوات، سنجد أنها جميعاً في الجدول 14. وكذلك الحال في المصطلحات الأكثر تكراراً والمكوّنة من كلمتين مثل اللغة العربية، علم الدلالة، علم الأصوات، علم اللغة، الكفاءة اللغوية، الفعل التواصل. ولأنّ المصطلحات المفردة عبارة عن أسماء، فقد حصرنا الأسماء في المدونة المتخصصة من خلال المدونة الموسومة بأقسام الكلام، ثم نظرنا فيما يوجد منها ضمن الكلمات المميزة، ثم أوجدنا نسبة المصطلحات منها. الجدول (15) يوضح عدد الكلمات المميزة، ونسبة المصطلحات منها ومن العدد الكلي للمصطلحات بحسب مستوى الثقة.

من الجدول (15) نلاحظ أنّ عدد الكلمات المميزة ينقص بازدياد مستوى الثقة، ولكن نسبة المصطلحات تزداد مع ازدياد مستوى الثقة. كما نلاحظ أنّ 78.1% فقط من المصطلحات ظهرت عند أدنى حد لمستويات الثقة (95%) مما يعني أنّ 21.9% من المصطلحات لم تظهر ضمن الكلمات المميزة. وبمراجعة هذه الكلمات وجدنا أنها إمّا شائعة ومستخدم في المجال العام بمعناها العام الذي يتشابه مع معناها الخاص في المدونة المتخصصة مثل: "الأسلوب" و"الماضي" و"الإسناد" و

"التصغير"، أو أنّها صور مختلفة من المصطلح المعرف بالألف واللام بعد دخول حروف الجر أو العطف عليها، وكان تكرارها قليلاً في المدونة المتخصصة، وكثيراً في المدونة العامة مثل: "والاستثناء" و "والفصح" و "بالمسكوكات"، أو قد تكون اسم علم مثل "المولد" وهو اسم عائلة للاعب كرة قدم سعودي. ويوضح الجدول (16) أدناه بعضاً من هذه المصطلحات.

جدول 14

الكلمات المئة الأكثر تميزاً بحسب مقياس الاحتمال اللوغاريتمي.

الكلمات المميزة	
اللغة، اللغوية، المعجم، التداولية، علم، الدلالة، المعنى، الأصوات، اللسانيات، العربية	10
المعجمية، اللغوي، النص، الكلمات، المتكلم، المعاجم، معجم، الدلالية، نظرية، النظرية	20
الجملة، المعجمي، دلالة، اللغات، الدلالي، النحوية، معنى، اللسانية، النحو، لغوية	30
اللفظ، اللسان، مصطلح، الخطاب، البحث، الفعل، الصوتية، الكلمة، التحليل، المصطلحات	40
هابرماس، دراسة، النطق، المفردات، للغة، التداولي، اللساني، النحوي، السياق، المنطقية	50
البنوية، لغة، المعاني، الدراسات، الألفاظ، الإحالة، الكفاءة، التعريف، والمعنى، التواصل	60
التوليدية، المعاصرة، الجمل، تحليل، الدلالي، الاستعمال، الصرفية، سوسير، التواصلية، تعليم	70
والتداولية، الكلام، الصوتي، اللفظية، المدونة، المداخل، العلامات، للتداولية، الرابط، مدخل	80
مهارة، معاجم، الألسنة، انظر، دلالية، الروابط، باللغة، السيميوطيقا، المقاربة، دار	90
كلمة، والدلالة، لغوي، معجمية، الدرس، البلاغة، اللغويين، التعريفات، التركيب، الحنك	100

جدول 15

عدد الكلمات المميزة ونسبة المصطلحات منها ومن العدد الكلي للمصطلحات بحسب مستوى الثقة

مستوى الثقة	عدد الكلمات المميزة	النسبة من الكلمات المميزة	النسبة من المصطلحات
%95	5,895	%13.5	%78.1
%98	5,531	%14.0	%75.9
%99	2,803	%21.0	%57.8
%99.99	2,154	%24.2	%51.3

جدول 16

أمثلة من المصطلحات المفردة التي لم تظهر ضمن قائمة الكلمات المميزة

الأسلوب، الفاعلية، الماضي، ومدلولاتها، والسوسيولوجيين، الإيماءات، التصغير، والانثروبولوجيا، الإسناد، والاستثناء، والفصح، والفلسفة، الإسناد، والفلسفية، بالمسكوكات، والمولد، أصوات، مدلولات، الحوسبة، المضاف، للتواصل، والمعرفة، والنفعية، صوامت، للمبالغة، لمشتقات، التناظر، والإضمار، التبادلية، والأمر، بالإشارة، والاستفهام، الترميز، والوسيط، الصيرورة، فالنص، بالفصحى، للعامل، والغائب

وتوضح البيانات بأنّ 48,758 متتابعة لفظية من كلمتين كانت ضمن المتتابعات المميزة للمدوّنة المتخصصة عند أدنى مستوى مقبول للثقة (95%). ويوضح الجدول (17) المتتابعات المئة الأكثر تميّزاً المكوّنة من كلمتين. وتشير البيانات أنّ 59 من المتتابعات المئة الأكثر تميّزاً كانت من المصطلحات المركبة من كلمتين. وهذا دليل آخر على أنّ التوزيع الإحصائي للمصطلحات في المدوّنات المتخصصة يختلف عنه في المدوّنات العامة. أمّا بقية المتتابعات الأكثر تميّزاً وكان عددها 41 متتابعة توزعت على ست مجموعات رئيسة يوضحها الجدول (18) مع أمثلة لها.

جدول 17

المتتابعات المئة الأكثر تميّزاً المكونة من كلمتين

المتتابعات المميزة	
اللغة العربية، علم اللغة، علم الدلالة، علم الأصوات، البحث اللغوي، امتحان الكفاءة، فروق دالة، المعجم العربي، دالة إحصائية، العربية المعاصرة	10
الكفاءة باللغة، توجد فروق، النحو الوظيفي، اللغوي المعاصر، تعليم اللغة، المعجم الإلكتروني، اللغة العربية، المقاربة التداولية، الفعل التواصل، الألسنة الطبيعية	20
الفعل المضارع، باللغة العربية، آفاق جديدة، الدرجة الكلية، تحليل الخطاب، اللغة العام، المعاجم العربية، لسان العرب، سياقه القرآني، الكفاءة اللغوية	30
أهمية علم، المعجمية العربية، الأصوات اللغوية، المعنى الأصلي، تحصيل الطلاب، اللفظ والمعنى، اللغة الطبيعية، جهاز النطق، الرسم البياني، اللسانيات والتداولية	40
البحث المعجمي، دي سوسير، مهارة القراءة، الدلالة يحملها، العربية للناطقين، قسم اللغة، الدراسات المعجمية، اللغة التطبيقية، البحث الدلالي، تحصيل طلاب	50
درسوا اللغة، لغة النص، دراسة الأصوات، لصالح الطلاب، الأصلي للجذر، التوليدية التحويلية، العربية لمدة، أخلاقيات النقاش، اكتساب اللغة، دلالة الفروق	60
أعضاء النطق، مهارة الكتابة، جدول دلالة، علم العلامات، وعلم اللغة، اختبار الكفاءة، نحلة محمود، أصوات اللغة، دراسة اللغة، محمود أحمد	70
اللغة الحديث، الدراسات اللغوية، استعمال اللغة، الدراسات الصوتية، مستعمل اللغة، القدرة التواصلية، إسناد الإحالة، دراسة المعنى، الناطقين بغيرها، المجموعات الكلي	80
الاختبارات اللغوية، تعليم اللغات، إستراتيجيات الخطاب، مختار عمر، الفنولوجيا التوليدية، الفونيتيك والفنولوجيا، المعجمية والمعجمية، المتوكل أحمد، المعجمي الحديث	90
أحمد آفاق، الوحدات المعجمية، التحليل التداولي، للناطقين بغيرها، الكلمات المفتاحية، الدراسي للطلاب، فرانسواز المقاربة، المصطلحات المعجمية، المعنى الأساسي، اللسانيات الوظيفية	100

جدول 18

المتتابعات التي لم تكن من المصطلحات ضمن المتتابعات المئة الأكثر تكرارًا .

المجموعة	العدد	أمثله
أسماء أعلام	7	دي سوسير، نحلة محمود، محمود أحمد، مختار عمر، المتوكل أحمد، أحمد آفاق، فرانسواز المقاربة
تحتوي أفعالاً	4	توجد فروق، الدلالة يحملها، درسوا اللغة، يتحدثها الطالب
تحتوي ضميرًا متصلًا في نهاية الكلمة الأولى / حرف جر أو عطف في بداية الكلمة الثانية	11	الكفاءة باللغة، سياقه القرآني، اللفظ والمعنى، العربية للناطقين، الأصلي للجذر، العربية لمدة، الناطقين بغيرها، للناطقين بغيرها، الدراسي للطالب
جزء من مصطلح مركب	6	اللغوي المعاصر، اللغة العام، اللغة التطبيقي، التوليدية التحويلية، اللغة الحديث، المعجمي الحديث،
تركيب/ مرجع أكاديمي	7	فروق دالة، دالة إحصائية، الدرجة الكلية، لسان العرب، قسم اللغة، دلالة الفروق، جدول دلالة،
أخرى	6	آفاق جديدة، أهمية علم، تحصيل طلاب، لصالح الطلاب، اخلاقيات النقاش، المجموعات الكلي

وتوضح البيانات أنّ المجموعات الثلاث الأولى يمكن التعامل معها أو تلافيها في كلّ الكلمات المميزة، باستبعاد الكلمات التي تحوي أسماء الأعلام، أو الأفعال أو الضمير المتّصل في نهاية الكلمة الأولى، أو حرف الجر أو العطف في بداية الكلمة الثانية. أمّا المجموعة الرابعة التي تتكون من المتتابعات التي تكون جزءًا من مصطلح مركب، فإنّه يمكن دمجها مع متتابعات تنتهي بنفس الكلمة الأولى للمتتابعة، أو تبدأ بنفس الكلمة الثانية للمتتابعة مثل: "علم اللغة" و "اللغة العام"، لتصبح مصطلحًا مركبًا من ثلاث كلمات "علم اللغة العام". كما يمكن استخدام أقسام الكلام لو كانت المدوّنّة موسّمة، فنأخذ بعين الاعتبار المتتابعات التي تتكون من اسم-اسم، أو اسم-صفة.

وبالنظر في نسبة المصطلحات المركّبة التي توجد في المتتابعات المميزة المكونة من كلمتين، نجد أنّ نسبة كبيرة من المصطلحات المركبة (أكثر من 92%) توجد في المتتابعة المميزة ضمن مستوى ثقة أعلى أو يساوي 98%، ولكن عدد الكلمات المميزة كبير جدًا للفحص والنظر البشري (أكثر من 98900 متتابعة)، ويتناقص عدد المصطلحات ضمن نطاق المتتابعات المميزة مع ازدياد مستوى الثقة، ويقلّ كذلك عدد المتتابعات المميزة بشكل كبير أيضًا.

ويوضح الجدول (19) عدد المتتابعات المميزة، ونسبة المصطلحات منها، ومن العدد الكلي للمصطلحات بحسب مستوى الثقة. وبمقارنة الجدول (19) مع الجدول (15) نرى بوضوح أنّ وجود المصطلحات المفردة في الكلمات المميزة أعلى عند مستويات الثقة الأعلى مع عدد محدود نسبيًا بالنسبة لعدد الكلمات المميزة. فاستخدام المتتابعات المميزة للكشف عن المصطلحات يحتاج إلى عمليات تصفية تستخدم فيها أقسام الكلام، وقوائم محددة سلفًا لاستبعاد الكلمات الوظيفية التي قد تكون ضمن الأسماء، وقياس للتلازم بين الكلمات، واستبعاد ذات التلازم المنخفض والتكرار القليل.

جدول 19

عدد المتتابعات المميزة ونسبة المصطلحات منها ومن العدد الكلي للمصطلحات بحسب مستوى الثقة

مستوى الثقة	عدد المتتابعات المميزة	النسبة من الكلمات المميزة	النسبة من المصطلحات
%95	49,054	%3.5	%92.8
%98	48,970	%3.5	%92.7
%99	4,604	%14.4	%36.1
%99.99	4,300	%15.0	%35.1

وكما هو الحال في المصطلحات المفردة، فقد كان هناك مصطلحات مركبة من كلمتين لم تظهر في المتتابعات المميزة المكونة من كلمتين، وقد شكّلت هذه المصطلحات 6.6% من المصطلحات المركبة من كلمتين. وتميزت هذه المصطلحات بتكرارها المنخفض، ولم يظهر أغلبها سوى مرة واحدة في المدوّنة المتخصصة. وقد كانت هذه المصطلحات مستخدمة في المجال العام والشأن الأدبي والثقافي أو الديني في الصحافة العربية مثل: "أصول الفقه" و"علم التجويد" و"علم الاجتماع" و"علم النفس" و"النقد الأدبي" و"النص الشعري".

7. الخاتمة

استعرضنا في هذه الورقة - باختصار - الأساليب المستخدمة في استخلاص المصطلحات المفردة والمركبة من المدوّنات اللغويّة العربية المتخصصة. وقد اعتمدت جميع تلك الدراسات على قواعد استرشادية لغوية، أو إحصائية، أو عليهما معاً. وقد درسنا في هذه الورقة حدود هذه القواعد الاسترشادية وإمكاناتها اعتماداً على مدوّنة لغوية متخصصة بعلوم اللغة، موسّمة يدوياً بالمصطلحات، وآلياً بأقسام الكلام يبلغ حجمها أكثر من 180 ألف كلمة، احتوت على 3,138 مصطلحاً مفرداً ومركباً. تكررت 7,618 مرة. وبالرغم من صحة القواعد الاسترشادية المستخدمة في استخلاص المصطلحات من المدوّنات اللغوية المتخصصة كما توضّح الدراسة، فإن أهم ما توصّلت إليه الدراسة هو: معرفة الأنماط الأكثر شيوعاً في القواعد اللغويّة، وأوجه القصور الموجودة في القواعد الإحصائية؛ حيث إنّ معرفة أهم الأنماط اللغوية شيوعاً وحدود الطرق الإحصائية وإمكاناتها، يمكّن الباحثين من صياغة قواعد استرشادية أكثر تخصيصاً ودقة، وتسهم في زيادة دقة الأساليب المستخدمة في استخلاص المصطلحات من المدوّنات اللغويّة المتخصصة.

وبالرغم من تغطية هذه الدراسة لأهم القواعد الاسترشادية في المجال، فلازنا بحاجة لدراسات أوسع تشمل على سبيل المثال: حدود التلازم والترابط بين مكونات المصطلح المركب، وما أفضل المقاييس الإحصائية للكشف عن التلازم بين مكونات المصطلح المركب؟ وما الأنماط الشائعة في المصطلحات المركبة من ثلاث وأربع كلمات؟ وما حجم المدوّنة اللغويّة المتخصصة الكافي لاستخلاص المصطلحات؟ وما حجم نصوص المدوّنة المرجعيّة التي يمكن استخدامها وطبيعتها؛ للكشف عن التوزيع الإحصائي المميز للمصطلحات؟

عبدالمحسن الثبيتي & وداد القحطاني، الخواصّ الممكنة لاستخلاص المصطلحات اللُّغويّة من المدوّنات العربية المتخصّصة

كما أنّ مجال الدراسة بحاجة إلى زيادة حجم المدوّنة المستخدمة، وتغطيتها كذلك لمجالات متخصّصة أخرى تتميز بها الثقافة العربية مثل: أصول الفقه، وعلوم الحديث، وغيرها. وليكون العمل في هذا المجال أكثر دقة، فنحن بحاجة إلى واسمٍ لِيّ لأقسام الكلام مُدرَّب على مدوّنة عربيّة متخصّصة.

الهوامش

1. https://sourceforge.net/projects/ghawwasv4/files/Almoshatheb_Alarabi%28V_4.0.0%29_09_12_18.jar/download
2. <http://arabicnlp.pro/ar/>
3. https://sourceforge.net/projects/ghawwasv4/files/Ghawwas_V4.6.jar/download

المراجع العربية

- الثبتي، عبدالمحسن، والقحطاني، وداد. (2022). طريقة مقترحة لاستخلاص المصطلحات من المدونات اللغوية العربية المتخصصة. *مجلة اللسانيات العربية*. ع 14، 35-56.
- الحمزاوي، محمد رشاد. (1986). *المنهجية العامة لترجمة المصطلحات وتوحيدها وتنميطها* (ط1) دار الغرب الإسلامي، بيروت.
- خسارة، ممدوح. (٢٠٠٨). *علم المصطلح وطرائق وضع المصطلحات في العربية* (ط1). دار الفكر، دمشق.
- السوداني، حسين. (2018). ترجمة المصطلح وتوطين اللسانيات: مثال الترجمات العربية الخمس لدروس فردينان دي سوسير. *مجلة اللسانيات العربية*. ع 7، 6-57.
- الشريدة، صفاء. (2017). توظيف الحاسوب في وضع المصطلح العربي ونشره. *مجلة دراسات العلوم الإنسانية والاجتماعية*. ع 3 (44)، 73-84.
- شندول، محمد (2008). بنية المصطلح اللُّغويّ في كتاب المقتضب للمبرّد. *مجلة المعجميّة*. ع 24، 169-186.
- عبدالرحيم، منتصر. (2020). مشكلات المصطلح والترجمة في معجم "اللغويات الاجتماعية". *مجلة اللسانيات العربية*. ع 11، 321-344.
- العتيبي، لينا. (2021). تخطيط المصطلح الطبي: (مصطلحات ابن سينا في الصيدلة نموذجاً). *مجلة التخطيط والسياسة اللغوية*. ع 12، 51-96.
- القاسمي، علي. (2008). *علم المصطلح أسسه النظرية وتطبيقاته العملية* (ط1). مكتبة لبنان ناشرون، بيروت.

المراجع الأجنبية

- Ahmad, K., & Rogers, M. (2001). Corpus linguistics and terminology extraction. In Sue Ellen Wright and Gerhard Budin (Eds.) *Handbook of terminology management*. (pp.725-760). Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Al Khatib, K., & Badarneh, A. (2010). Automatic extraction of Arabic multi-word terms. *Proceedings of the International Multiconference on Computer Science and Information Technology* (pp. 411-418). IEEE.
- Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. *2013 International Conference on Asian Language Processing (IALP)* (pp. 67-70). IEEE.
- Al-Thubaity, A. M., Khan, M., Alotaibi, S., & Alonazi, B. (2014). Automatic arabic term extraction from special domain corpora. *2014 International Conference on Asian Language Processing (IALP)* (pp. 1-5). IEEE.
- Boulaknadel, S., Daille, B., & Aboutajdine, D. (2008). A multi-word term extraction program for Arabic language. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)* (pp. 1485–88). European Language Resources Association.
- Bounhas, I., & Slimani, Y. (2009). A hybrid approach for Arabic multi-word term extraction. *2009 International Conference on Natural Language Processing and Knowledge Engineering* (pp. 1-8). IEEE.
- Deepthi, S., Rajesh, B., Vyshnavi, N., & Sushma, K. M. (2017). Automatic Key Term Extraction from Research Article using Hybrid Approach. *International Journal of Computer Applications*, 166(6),17-21.
- Eito-Brun, Ricardo (2017, April). Terminology extraction and the identification of research areas. *Systèmes d'organisation des connaissances et humanités numériques: Actes du 10ème colloque ISKO 2015* (pp. 114-128). ISTE Group.
- El Mahdaouy, A., Ouatik, S. E. A., & Gaussier, E. (2013). A Study of Association Measures and their Combination for Arabic MWT Extraction. *10th International Conference on Terminology and Artificial Intelligence* (pp. 45-52).
- Fkih, F., & Omri, M. N. (2016). IRAFCA: an O (n) information retrieval algorithm based on formal concept analysis. *Knowledge and Information Systems*, 48(2), 465-491.

- Freihat, A. A., Bella, G., Mubarak, H., & Giunchiglia, F. (2018). A single-model approach for Arabic segmentation, POS tagging, and named entity recognition. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1-8). IEEE.
- Hadni, M., Lachkar, A., & Ouatik, S. A. (2014). Multi-Word Term Extraction Based on New Hybrid Approach for Arabic. *Proceeding of Second International Conference on Computational Science and Engineering* (pp. 109–120).
- Hätty, A., Schlechtweg, D., Dorna, M., & im Walde, S. S. (2020). Predicting degrees of technicality in automatic terminology extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2883-2889). ACL.
- Kim, T. H., Choi, K. N., & Shin, Y. J. (2021). Development of a management system for an R&D terminology dictionary. *Ilkogretim Online*, 20(3), 662-670.
- Kosa, V., Chaves-Fraga, D., Dobrovolskyi, H., & Ermolayev, V. (2019). Optimized term extraction method based on computing merged partial c-values. *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications* (pp. 24-49). Springer, Cham.
- Lamrani, E.K., Marzak, A. and Ballaoui, H. (2014). Mixed method for extraction of domain terminology from text: Linguistic and statistical filtering. *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)* (pp. 291-295). IEEE.
- Li, X. F. (2021). Domain Concept Extraction and Ontology Visualization. *Journal of Physics: Conference Series* (Vol. 1994, No. 1, p. 012014). IOP Publishing.
- Liu, J. (2021). Research on Translation Processing of Complex Long Sentences Based on Multi-strategy Analysis Algorithm. *Journal of Physics: Conference Series* (Vol. 1881, No. 4, p. 042092). IOP Publishing.
- Liu, K., & Chen, L. (2019). Medical social media text classification integrating consumer health terminology. *IEEE Access*, 7, 78185-78193.
- Luo, H., Li, T., Liu, B., Wang, B., & Unger, H. (2019). Improving aspect term extraction with bidirectional dependency tree representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(7), 1201-1212.

- Mashaan, A., Tiun, S., & Albared, M. (2013). Arabic Term Extraction Using Combined Approach on Islamic Document. *Journal of Theoretical & Applied Information Technology*, 58(3), 601-8.
- Oliver, A., & Vázquez, M. (2020). Termeval 2020: Using TSR filtering method to improve automatic term extraction. *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 106-113). European Language Resources Association.
- Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020). Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. *6th International Workshop on Computational Terminology (COMPUTERM 2020)* (pp. 85-94). European Language Resources Association (ELRA).
- Shao, W., Hua, B., & Song, L. (2021). A pattern and POS auto-learning method for terminology extraction from scientific text. *Data and Information Management*, 5(3), 329-335.
- Simon, N. I., & Kešelj, V. (2018). Automatic term extraction in technical domain using part-of-speech and common-word features. *Proceedings of the ACM Symposium on Document Engineering 2018* (pp. 1-4). ACM.
- Speranza, G., Manna, R., di Buono, M. P., & Monti, J. (2020). The Archaeo-Term Project: Multilingual Terminology in Archaeology. *Proceedings of the Seventh Italian Conference on Computational Linguistics* (pp. 387-393).
- Yan, J., Qin, Y., Yu, Y., & Xu, H. (2020). Terminology extraction in the field of water environment based on rules and statistics. *Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education* (pp. 144-147).
- Zeng, W., Li, X., & Li, H. (2018). Study on Chinese term extraction method based on machine learning. *International Conference of Pioneering Computer Scientists, Engineers and Educators* (pp. 128-135). Springer, Singapore.

بيانات الباحث

AUTHOR BIODATA

Abdul Mohsen Al-Thubaity is a Full Professor of natural language processing in The National Center for Data Analytics and Artificial Intelligence, King Abdul-Aziz City for Science and Technology. Prof. Al-Thubaity obtained his PhD degree in computer sciences (2004) from University of Surry. His research interests include language resources and evaluation, Arabic NLP, computational linguistics, and information extraction.

عبدالمحسن عبيد الثبتي، أستاذ في معالجة اللغات الطبيعية في المركز الوطني لتحليل البيانات والذكاء الاصطناعي في مدينة الملك عبدالعزيز للعلوم والتقنية بالمملكة العربية السعودية. حاصل على درجة الدكتوراه في علوم الحاسب من جامعة ساري عام 2004م. تدور اهتماماته البحثية حول المدونات اللغوية وتقويمها، ومعالجة اللغة العربية، واللسانيات الحاسوبية، واستخلاص المعلومات من النصوص.

معرف أوركيد (ORCID): 0000-0003-2376-0849

Email: aalthubaity@kacst.edu.sa

بيانات الباحثة

AUTHOR BIODATA

Wedad Ahmed Alqahtani is an Associate Professor of syntax and morphology in the Department of Arabic Language and Literature, College of Humanities and Social Sciences, King Saud University. Dr Alqahtani obtained her PhD degree in syntax and morphology (2005) from Umm Al-Qura University. Her research interests include syntax, morphology, semantic and corpus linguistics.

وداد أحمد القحطاني، أستاذ مشارك في النحو والصرف في قسم اللغة العربية وآدابها، بكلية العلوم الإنسانية والاجتماعية، في جامعة الملك سعود، حاصلة على درجة الدكتوراه في النحو والصرف من جامعة أم القرى عام 2005م. تدور اهتماماتها البحثية حول قضايا النحو والصرف والدلالة ولسانيات المدونات.

معرف أوركيد (ORCID): 0000-0001-9318-6278

Email: walqahtany@ksu.edu.sa