



اللغويات في الجلسات والمعجمية الآلية للغات العربية

Journal of Computational Linguistics & Arabic Language Processing



المدونة المحوسبة لكتب منتخبة من دائرة أصول الفقه الحنفي إعادة
هيكلية المحتوى لتلبية متطلبات الذكاء الاصطناعي التوليدي

تحسين استرجاع اسم التصغير في التحليل الصرفي

نحو آلية جديدة للرصد اللغوي للغة العربية

توظيف تضمينات الكلمات المُدرّبة مسبقاً في توسيم الأجزاء
النحوية لبيانات التناوب اللغوي

السنة الثانية - العدد الثالث
شوال ١٤٤٦ هـ - إبريل ٢٠٢٥ م
Vol. 2, No. 3, Apr, 2025

مجمع الملك سلمان
العالمي للغة العربية
King Salman Global Academy for Arabic Language



اللغويات الحاسوبية والمعالجة الآلية للغات العربية

Journal of Computational Linguistics & Arabic Language Processing

www.ksaa.gov.sa

مجلة علمية مُحكمة تصدر عن مجمع الملك سلمان العالمي للغة العربية





المشرف العام على المجلة

أ.د. عبدالله بن صالح الوشمي

هيئة التحرير

د. سارة الحمود

د. عريب العويشق

د. عبدالعزيز المهيوبي

د. أيمن الغامدي

إدارة التحرير

أ.د. منصور الغامدي

رئيس التحرير

الهيئة الاستشارية

أ.د. إيريك أتويل

أ.د. نزار حبش

أ.د. خالد فؤاد شعلان

أ.د. صلاح راشد الناجم

أ.د. عبد المحسن عبيد الثبيتي

أ.د. محسن عبد الرزاق رشوان

أ.د. مصطفى عرسان جرار

أ.د. هند سليمان الخليفة

د. عبد الرحمن العصيمي

مدير التحرير

د. هاجر القبيشي

أمينة المجلة



◆ قواعد النشر بالمجلة:

سياسات النشر:

تنشر مجلة اللُّغويات الحاسوبية والمعالجة الآلية للغة العربية البحوث الأصلية التي لم يسبق نشرها، والمكتوبة باللغة العربية أو الإنجليزية، والدراسات والمختصات العلمية المتخصصة في مجال حوسبة اللغة، كاللُّغويات الحاسوبية، والمعالجة الآلية للغة العربية، وكل ما له صلة بحوسبة اللغة.

تُعرض مسودة الأوراق العلمية فور وصولها على هيئة التحرير لإجراء تقييم مبدئي لها، وفي حال عدم مناسبتها للنشر تُخطر المجلة الباحث بذلك. أمّا البحوث المقبولة مبدئياً، فتُرسل إلى اثنين من المحكمين لتقييمها سريعاً، وتبذل هيئة التحرير طاقتها في التواصل مع المحكمين للوصول إلى قرار بشأن كل بحث مقدّم في أسرع وقت ممكن، ويُخطر الباحث بنتيجة التحكيم. وفي حالة قبول الورقة للنشر، فإنّ كافة حقوق النشر تؤوّل للمجلة، ولا يجوز نشرها في أي منفذ نشر آخر ورقياً أو إلكترونياً، دون إذن كتابي من رئيس هيئة تحرير المجلة. وتُرسل نسختان ورقيتان مجانيتان من المجلة إلى المؤلف عند نشر ورقته.



Journal information:

◆ بيانات المجلة:

مجلة اللُّغويات الحاسوبية والمعالجة الآلية للغة العربية Journal of Computational Linguistics & Arabic Language Processing

مجلة علمية محكمة تصدر عن A peer-reviewed bi-annual journal issued by

مجمع الملك سلمان العالمي للغة العربية King Salman Global Academy for Arabic Language

الرياض - المملكة العربية السعودية Riyadh- Saudi Arabia

ردمدم النسخة الورقية: ٩٤٧٥-١٦٥٨ ISSN paper version: 1658-9475

ردمدم النسخة الإلكترونية: ٩٤٨٣-١٦٥٨ ISSN Electronic version: 1658-9483

تعليمات عامة:

١. تُرسل الورقة المقدّمة للنشر للمجلة بصيغة وورد (word) على بريد المجلة: (jclalp@ksaa.gov.sa). على أن تُوجّه جميع المراسلات إلى: رئيس هيئة التحرير- مجلة اللُّغويات الحاسوبية والمعالجة الآليّة للغة العربيّة- مجمع الملك سلمان العالمي للغة العربيّة.

٢. يجب أن تحتوي الورقة على ما يلي:

٢-١- صفحة أولى فيها عنوان الورقة، واسم المؤلف أو المؤلفين، وجهة عملهم، وعناوينهم، وتاريخ تقديم الطلب بنشر الورقة.

٢-٢- الكلمات المفتاحية.

٢-٣- مستخلص الورقة باللّغتين العربية والإنجليزية -في حدود مئة كلمة لكلّ منهما-.

٢-٤- المصادر والمراجع.

٢-٥- الملاحق -إن وجدت-.

٣. يُراعى أن تكون الورقة العلمية مكتوبةً بلغة سليمة -سواء كانت عربية أو إنجليزية-، وأن يتراوح عدد كلماتها بين (٣٠٠٠ - ١٠٠٠٠) كلمة، بما في ذلك المراجع والهوامش والجداول والأشكال والملاحق.

٤. إذا كانت الورقة العلمية مكتوبة باللّغة العربيّة، فيجب الالتزام بكتابة ما يقابل المصطلحات العربيّة باللّغة الإنجليزيّة في المرة الأولى التي يظهر فيها المصطلح.

٥. ألا يسبق نشر البحث في أي وعاء علمي -مالم ترهئة التحرير خلاف ذلك-.

المواصفات الفنية للبحث:

- يكون مقاس الصفحة A4، وتُقسَّم إلى عمودين، وتكون المسافة بين العمودين ١ سم، وتكون الهوامش ٥,٢ سم عن يمين ويسار الصفحة، و٥٨,٣ سم من أعلى وأسفل الصفحة.
- تُترك مسافة 5.0 سم في بداية كل فقرة من فقرات المتن.
- يجب الفصل بمسافة سطر واحد بين نهاية محتوى الفقرات داخل المتن وعناوين الفقرات التي تليه.
- يجب ترك مسافة سطر واحد قبل وبعد الجدول أو الشكل والصورة والمتن السابق والتالي لهما.
- يُستخدم الخط (Sakkal Majalla) للنص العربي، والخط (Times New Roman) للنص الأجنبي وبجسم أقل بدرجتين عن الخط العربي. ويكون حجم الخط العربي كالآتي:
 - العنوان الرئيس: ١٤، عريض.
 - العناوين الفرعية: ١٢، عريض.
 - النصوص: ١٢، عادي.
 - الهوامش السفلية: ١٠، عادي.
- يجب ترقيم صفحات البحث.
- يجب ترقيم المعادلات.
- تُرتَّب المصادر والمراجع أبجدياً بنظام APA.
- تُفصل المراجع الأجنبية عن المراجع العربية في قائمة المصادر والمراجع، وتأتي المراجع/المصادر العربية أولاً في الأبحاث العربية والمراجع/المصادر الإنجليزية أولاً في الأبحاث الإنجليزية.

الجداول:

- تُدرج الجداول بعد الفقرة التي ذُكرت بها، وذلك على النحو التالي:
 - إذا كان عرض الجدول ٧,٥ سم يوضع في عمود واحد أسفل الفقرة مباشرة.
 - إذا كان عرض الجدول أكثر من ٧,٥ سم يوضع أسفل الصفحة التي ذكر بها أو أعلى الصفحة التي تليها.
- تُرقِّم الجداول بين قوسين في أقصى يمين الصفحة مسبوقة بكلمة «جدول». مثال: جدول (١).
- يُوضع عنوان الجدول في أعلى الجدول مباشرة، في أقصى اليمين.
- يجب ألا يتجاوز عرض الجدول هوامش الصفحة المذكورة في الفقرة السابقة.
- يُكتب مصدر الجدول والتعليقات المصاحبة له -إن وجدت- أسفله.
- تُضبط بيانات الجدول غير الرقمية أقصى يمين الخلية.
- تُضبط بيانات الجدول الرقمية وسط الخلية.
- إذا استمر الجدول على أكثر من صفحة، فيجب تكرار رأسه في الصفحات التالية ووضع عبارة «تابع جدول (...)» في أعلاه.

الصُّور والأشكال:

- تُدرج الصور والأشكال بعد الفقرة التي ذُكرت على النحو التالي:
 - إذا كان عرض الصورة ٧,٥ سم تُوضع في عمود واحد أسفل الفقرة مباشرة.
 - إذا كان عرض الصورة أكثر من ٧,٥ سم تُوضع أسفل الصفحة التي ذكر بها أو أعلى الصفحة التي تليها.
- تُرقِّم الصور والأشكال، ويوضع الرقم بين قوسين مسبوقة بكلمة «شكل». مثال: شكل (١).

- يُدرج عنوان الشكل أسفل الشكل مباشرةً متوسّطاً بين حدوده بحيث لا يتجاوز عنوان الشكل سطرين فقط.
- في حال وجود تعليق مصاحب على الشكل أو الصورة يجب دمجه مع الشكل وحفظه على هيئة صورة لا تقل دقتها عن (300 dpi).

مواصفات كتابة المعادلات:

- تُكتب المعادلات الرياضية باستخدام أحد البرنامجين التاليين: (Math Type) أو (Equation Editor).
- تُرقّم المعادلات، ويوضع الرقم بين قوسين في أقصى يسار العمود مقابل المعادلة على مستوى البحث.

مواصفات كتابة الاقتباس:

- إذا كان النص المقتبس ٣ أسطر أو أقل، فيجب أن يوضع بين علامتي تنصيص «...»، ويليه توثيق النص على النحو التالي: (العائلة، السنة: الصفحة).
- إذا كان النص أكثر من ٣ أسطر، فيجب أن يُميز النص المقتبس بكتابته في منتصف الصفحة، وتُترك من بداية السطر ونهايته هوامش ١ سم، ويوضع بين علامات تنصيص «...»، ويكتب بحجم ١٢ للنص العربي و١٠ للنص الإنجليزي، وينتهي بتوثيق النص على النحو التالي: (العائلة، السنة: الصفحة).

مواصفات كتابة الحواشي:

- «تُدرج الحواشي أسفل كل صفحة، مع ترقيمها آلياً حسب ورودها في كل صفحة».

◆ صفة المجلة:

مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيَّة (Journal of Computational Linguistics & Arabic Language Processing) هي مجلة علمية محكمة، نصف سنوية (ISSN:1658-9483 - ISSN:1658-9475)، وتصدر عن مجمع الملك سلمان العالمي للُّغة العربية في الرياض بنسختين: إلكترونية، وورقية (ذات إصدار محدود). وتتيح المجلة الوصول الحر (Open Access) إلى البحوث المنشورة فيها، ولا تتقاضى رسوماً على النشر. وهي مجلة تهتم بنشر الأبحاث العلمية المحكمة في مجال حوسبة اللُّغة، وتشمل: اللُّغويات الحاسوبية (Computational Linguistics)، والمعالجة الآلية للُّغة العربية (Arabic Language Processing) وما يتبعها من تخصصات متَّصلة بها. وللاطلاع على أهداف المجلة وشروط النشر وكل التفاصيل ذات العلاقة يمكنكم زيارة موقع المجلة: jclalp.ksaa.gov.sa

البحوث المنشورة بالمجلة تعبر عن رأي أصحابها، ولا تعبر بالضرورة عن رأي المجمع،
ويتحمل مؤلفوها المسؤولية كاملة عن صحة المعلومات والاستنتاجات ودقتها.

محتويات العدد

١٣

المدونة المحوسبة لكتب منتخبة من دائرة أصول
الفقه الحنفي إعادة هيكلة المحتوى لتلبية متطلبات
الذكاء الاصطناعي التوليدي

حسن مظفر الرزو

٧١

تحسين استرجاع اسم التصغير في التحليل الصرفي

هاله منيف الحري - عبد الله يحيى الفيضي

١٠٣

نحو آلية جديدة للرصد اللغوي للغة العربية

علاء ممدوح عاكف

١٢٦

توظيف تضمينات الكلمات المُدرّبة مسبقًا في توسيم
الأجزاء النحوية لبيانات التناوب اللغوي

فهد الغامدي - منى دياب

كلمة التحرير

بسم الله الرحمن الرحيم

الحمد لله رب العالمين، والصلاة والسلام على أشرف الأنبياء والمرسلين، وبعد...

عزيزي القارئ/القارئة - عزيزي الباحث/الباحثة:

السلام عليكم ورحمة الله وبركاته

يُسعدنا أن نرحّب بكم في العدد الثالث من المجلة العلمية المحكّمة: «اللُّغويات الحاسوبية والمعالجة الآليّة للغة العربيّة»، والتي تأتي ضمن إصدارات مجمع الملك سلمان العالمي للغة العربيّة، والمتخصصة في مجال حوسبة اللُّغة، خاصة اللُّغة العربيّة.

تستهدف المجلة نشر البحوث والدراسات التي تُسهم في تطوير فهمنا العميق لتطبيقات حوسبة اللُّغة العربيّة. وتعمل على رصد أحدث الابتكارات التقنية وأدوات التحليل اللُّغوي، مع التركيز على التطورات الحديثة في مجال التفاعل بين الإنسان والآلة على المستوى اللُّغوي. تسعى المجلة أيضاً إلى استشراف المستقبل وتحديد التحديات والفرص الواعدة في هذا الحقل الحيوي، بهدف دفع عجلة التقدم العلمي في مجال الحوسبة اللُّغوية.

نأمل أن يكون إطلاق العدد الثالث من المجلة في وقته، حافزاً للباحثين والمهتمين للعناية بالمجلة ومتابعة إصداراتها، وأن تكون المجلة مصدراً ثرياً للمهتمين

بالحوسبة اللغوية، وأن تُسهم في دفع عجلة الأبحاث في الحوسبة اللغوية المنشورة باللغة العربية. كما نأمل أن يكون نشر الأبحاث المترجمة حافزاً للباحثين للكتابة باللغة العربية، وإثراء المحتوى العلمي العربي في مجال الحوسبة العربية.

تحتوي المجلة في عددها الثالث أربعة أبحاث مميزة تتناول جوانب متعددة ومتخصصة في صميم اهتمامات المجلة، حيث يناقش البحث الأول خوارزميات مقترحة لتحسين استرجاع الكلمات المصغرة. أما البحث الثاني فيضع إطاراً علمياً لمدونة لأصول الفقه الحنفي. أما البحث الثالث، فيقترح آلية جديدة للرصد اللغوي للغة العربية مستقاة من تقنيات الرصد اللغوي في اللغة الصينية. أما البحث الرابع والأخير، فيدرس نتائج توظيف مضامين الكلمات في تحسين التعرف على أقسام الكلام في مجال التناوب اللغوي.

نشكر الباحثين والمحكمين الذين ساهموا في إثراء هذا العدد بمشاركاتهم القيمة، ونتمنى لك -عزيزي القارئ، عزيزي القارئة- رحلة ممتعة ومفيدة مع صفحات هذه المجلة، ونتطلع إلى الاستماع إلى آرائكم وملاحظاتكم حول محتوى المجلة، وسُبل تطويرها. ويسعدنا مشاركتكم وتفاعلكم المثمر معنا في الأعداد القادمة.

هيئة التحرير

المدونة المحوسبة لكتب منتخبة من دائرة أصول الفقه الحنفي إعادة هيكلة المحتوى لتلبية متطلبات الذكاء الاصطناعي التوليدي

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-001>

حسن مظفر الرزو^١

المعهد العالمي لحوسبة القرآن الكريم والعلوم الإسلامية، الموصل، جمهورية العراق

استقبل في	رُوجع في	قُبِل في	نُشر في
١٤٤٥/١٠/٢٧	١٤٤٥/١١/١	١٤٤٦/٠٢/٣٠	١٤٤٦/١٠/٢٨

المخلص

تهدف هذه الدراسة الى انتخاب خمسة نصوص من دائرة علم أصول الفقه الحنفي لكي تكون مورداً لتشكيل المدونة المحوسبة لأصول فقه هذه المدرسة الفقهية العريقة. وتستخدم معالجات اللغة الطبيعية لتفكيك مفردات هذه النصوص الى وحدات أولية Tokens مع استبعاد كلمات التوقف من نصوص هذه الكتب قبل مباشرة معالجات التحليل المحوسب النحوي والصرفي. وتعيد تشكيل المدونة المحوسبة لأصول الفقه الحنفي من هذه الموارد وتشكل هيكلة مفرداتها لكي تتوافق مع متطلبات معالجات اللغة الطبيعية، وكذلك نظم الذكاء الاصطناعي التوليدي، وتقوم بإحصاء عدد وحداتها الأولية وكلمات التوقف الموجودة فيها، وتمارس عمليات التحليل النحوي والصرفي المحوسب، ومعالجات تضمين الكلمات لضمان فرص استثمار مادة المدونة المحوسبة كمورد لمعالجات الذكاء الاصطناعي التوليدي المتقدمة. وتسعى الى بيان فرص استثمار مخرجات المدونة المحوسبة في تعميق فهمنا بالكثير من الجوانب الأصولية مثل الأحكام الوضعية والتكليفية، مع إلقاء الضوء على مسألة المكروه وتفرعاتها في لباب مدونتنا المحوسبة.

الكلمات المفتاحية

المدونات، معالجات اللغة الطبيعية، تضمين الكلمات، الذكاء التوليدي.

The Computational Corpus of Selected Books from the Realm of Hanafi Jurisprudential Principles: Restructuring the Content to Meet the Requirements of Generative Artificial Intelligence

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-001>

Hasan M. Al-Rizzo 

International Computing Institute for Quran & Islamic Sciences, Mosul, Iraq.

Received on	Revised on	Accepted on	Published on
06/05/2024	09/05/2024	03/09/2024	26/04/2025

ABS TRACT

This study aimed to develop a computational corpus for the principles of the Hanafi school of Islamic jurisprudence (Usul al-Fiqh Al-Hanafi). Five renowned textbooks in this field were selected as source material. The content was tokenized using natural language processing techniques, with stop words removed. Syntactic and morphological analyses were conducted computationally on the remaining text. The corpus was then reconstructed to align its vocabulary with NLP requirements. Token and stop word counts were tabulated. Word embedding methods were applied to enable utilization of the corpus for advanced generative AI pre-processing. The overarching goal was to leverage the corpus outputs to gain deeper insights into various jurisprudential principles within the Hanafi tradition. By computationally processing these authoritative sources, the study sought to unlock new avenues for understanding the nuances and applications of Hanafi juristic principles through data-driven analysis.

KEYWORDS

Computational Corpus, Natural Language Processing, Word Embedding, Generative A.I.



١- المقدمة

أضحت عملية إنشاء المدونات المحوسبة لموارد العلوم الإسلامية، تفسيراً، وحديثاً، وفقهاً، وأصولاً من المسائل المهمة بعد أن سمحت أدوات الذكاء التوليدي^(١) لنفسها بالتوغل في موارد علومنا الإسلامية، ومباشرة عملية التنقيب في النصوص Text Mining^(٢) على صعيد تعدد المعاني والدلالات التي تمتلكها المفردة الاصطلاحية في السياقات التي ترد فيها ضمن النصوص التي تتوفر بغزارة في الفضاء الرقمي. وتكمن الخطورة في هذا النمط من التغلغل للتحليل المحوسب في عدم أخذه بنظر الاعتبار المستويات المتعددة التي يمتلكها الاصطلاح داخل نطاق الخطاطة الأصولية، الأمر الذي سيسهم في توليد إجابات، أو إنشاء نصوص، أو اختصارات فارغة من روح المنطق الفقهي والأصولي؛ لتنتج لنا بعد عقد من الزمن فقهاً سيخترق شبكة المفاهيم لدى المسلم فينشأ عنه الكثير من المخاطر الجسيمة على المنظومة العقدية، والفقهية، بسبب نهج المعالجة الذي سيمارس بواسطة الذكاء المحوسب ومن خارج حدود المنطق الفقهي والأصولي.

من أجل هذا سعيينا إلى إنشاء مدونة محوسبة Computational Corpus من مصادر أصولية انتخبت بعناية للتعبير عن التغير الحاصل في الخطاب الأصولي للمذهب الحنفي خلال بضعة قرون، مع تهيئة النصوص وهيكلتها وفق معمارية منطقية تتوافق مع

(١) الذكاء الاصطناعي التوليدي Generative Artificial Intelligence حقل جديد من حقول الذكاء الاصطناعي الذي يستهدف إنشاء محتوى جديد يقارب محتوى البيانات التي مارست خوارزمياته الذكية سلسلة مكثفة من عمليات التدريب والمران المتكرر على مادتها. تركز معمارية أنموذج الذكاء التوليدي الى مجموعة من نماذج تعلم الآلة التي تمتلك القدرة على توليد المحتوى من خلال تدريب مجموعة من الشبكات العصبونية المعقدة على كم كبير جداً من البيانات، للكشف عن الأنماط السائدة في السياق، وتحليل الفضاء الدلالي للبنيات اللغوية لكي تكون قادرة على توليد محتوى جديد. ونتيجة لعمليات المران والتعلم الذي تمارسه الشبكات العصبونية يصبح أنموذج الذكاء التوليدي قادراً على إنتاج نصوص، وصور، وشفرات برمجية، وأنماط أخرى من المحتوى الصوتي والمرئي، التي يحاكي فيها المحتوى الذي ينتجه الإنسان.

(٢) عملية التنقيب في النصوص ويطلق عليها اصطلاح Text Mining والتي توظف مجموعة متنوعة من الأدوات والمعالجات التي تتدرج من إحصاء عدد الكلمات، ونسب حضورها في النص، مع تعدد الكلمات الفريدة، صعوداً الى البحث عن المفاهيم والتوجهات داخل حدود النص.

مطالب الخوارزميات الذكية التي باتت تستخدم بكثرة لاستخراج الاصطلاحات، والمفاهيم، والدلالات من النصوص، وقد عولجت مادة البيانات بواسطة معالجات اللغة الطبيعية، والمعالجات اللاحقة للذكاء المحوسب، فباشرنا التحليل النحوي والصرفي والدلالي المحوسب، مع تحليلات متقدمة هيأت محتوى المدونة؛ لكي تكون مورداً للمعالجات المتقدمة للذكاء الاصطناعي التوليدي.

٢- أهداف البحث والنهج الذي اعتمد فيه:

رغم وجود عدد كبير من البحوث التي عولجت فيها الكثير من المسائل التي تخص المدونات المحوسبة لعلوم اللغة العربية، أو مدونات تختص ببعض حقول المعرفة كالموسوعات، ومستودعات الأخبار، ومنصات تعليم اللغة العربية. فإننا نلاحظ أن جل عنايتها قد تركزت على حقل لسانيات اللغة العربية، ولم تعرّج إلى حقول المعرفة الأخرى، مثل: مدونات العلوم الإسلامية (التفسير، الحديث النبوي، الفقه، أصول الفقه، التاريخ) وغيرها من الحقول التي تمتاز ببراء ذخيرتها وفرص توفير مدونات تدعم البحث العلمي في كل حقل من هذه الحقول الغنية بمواردها. كذلك، فإن الكثير منها تنطق بلسان نقدي يتسم بمسائل لا تلتحق بالمعالجات المحوسبة التي باتت تتسارع في حقول المعالجات المحوسبة حيث نوظف الخوارزميات المعقدة، والذكاء المحوسب بمختلف مستوياته.

من أجل هذا فقد وضعنا نصب أعيننا في هذه الدراسة التوجّه صوب علم أصول الفقه الإسلامي، ووقع اختيارنا على مدرسة المذهب الحنفي؛ لتكون مورداً خصباً نمارس في محتواها سلسلة من المعالجات المحوسبة التي استثمرنا من خلالها الخوارزميات الذكية مع الالتزام بالمعالجات المحوسبة للذخيرة اللغوية لهذه الموارد، بواسطة آليات التحليل النحوي، والصرفي، والدلالي المحوسب. والسعي إلى ممارسة تحليلات متقدمة مثل: وسم هوية الكيانات (الأعلام والأماكن) (NER) Named Entity Recognition والتقاطها بواسطة خوارزميات ذكية Intelligent Algorithms، ومحاولة التوسّع على صعيد المعالجات التي توفر مادة خصبة لنظم الذكاء التوليدي، مثل: تضمين الكلمات Words Embedding، مع محاولة توليد الفضاء الدلالي Semantic Landscape لاصطلاحات منتخبة من كل علم



من هذه العلوم، في سعي نحو توفير المزيد من المدونات الرقمية التي يمكن أن تكون مورداً مفيداً لنماذج اللغة الكبيرة Large Language Models التي سينصبّ اهتمامها بلغتنا العربية، والعلوم الإسلامية المختلفة.

وجدنا أن الذخيرة المعرفية المحوسبة التي أنتجتها الخوارزميات الذكية قد تنامت، وتشعبت مفرداتها بين أقسام الكلام، والبنىات الصرفية، والفضاءات الدلالية للكلمات، وموارد أخرى يمكن أن تزداد نمواً مع إقبال الباحثين على استخدام معالجات اللغة الطبيعية المحوسبة، ومختلف أنماط معالجات الذكاء الاصطناعي التوليدي؛ لتوسيع دائرة الذخيرة المعرفية للعلوم، مع منح أنظمة الذكاء التوليدي أكثر من فرصة سانحة في توليد معانٍ أكثر دقة عند توليد الإجابات، أو معالجة مختلف أشكال المسائل المطروحة عليها.

٣- النهج المحوسب لإعادة هيكلة النصوص وتحليل مضامينها:

بصورة عامة، تتألف المدونات الخام Raw Corpus من مجموعة من النصوص المنتخبة في حقل من الحقول المعرفية، دون إجراء أي معالجة لاحقة لها باستثناء بعض أنماط الوسم اليدوي التي تمارس على بعض مفرداتها. ويمكن أن تتحول المدونة الخام إلى مدونة محوسبة بعد ممارسة سلسلة من عمليات المعالجة النحوية، والصرفية، والدلالية، ووسم الكيانات اللغوية على مفرداتها، بعد إجراء عملية تفكيك المحتوى إلى وحدات أولية Tokens بواسطة معالجات اللغة الطبيعية. (العصيمي، صالح. ١٤٣٦: ص ١٨-٢١)

ويطلق اصطلاح معالجات اللغة الطبيعية (NLP) Natural Language Processing على أحد الحقول المهمة في مجال الذكاء الاصطناعي، والذي يُعنى بمعالجة اللسانيات اللغوية وتذليل العقبات التي قد تعترض تواصل الإنسان مع الحواسيب وخوارزمياتها البرمجية؛ لكي يوفر فضاءً معرفياً للتواصل بين الحواسيب والإنسان، وفصح المجال أمام المعالجات الذكية في بناء خوارزميات برمجية ذكية تمتلك القدرة على فهم، وتفسير، الأنماط والعلاقات التي تربط بين مفردات الخطاب اللساني، والقدرة على توليد المعاني والمفاهيم من مفردات اللغة الإنسانية. (العصيمي، صالح. ١٤٣٦: ص ١٩)

تتألف معالجات اللغة الطبيعية من محورين أساسيين (Liang, Hong et.al, 2017: p4):

المحور الأول: فهم اللغة الطبيعية Natural Language Understanding.

المحور الثاني: توليد اللغة الطبيعية Natural language Generation.

وتتوجّه دفة المعالجات في المحور الأول نحو التحليل الدلالي للنصوص والكشف عن المعاني المودعة فيها؛ لغرض استثمارها في الترجمة الآلية، والإجابة عن الأسئلة المطروحة، وتوفير فضاء لتحليل الآراء والتوجهات Sentiment Analysis. أما معالجات المحور الثاني فتتوجّه دفتها نحو المعالجات التي تقوم بتوليد عبارات وجمل في نصوص مكتوبة أو منطوقة تجيب عن تساؤلات المستخدمين وتلبي حاجاتهم على صعيد حل الإشكاليات، أو صناعة مختلف أشكال المحتوى المدوّن أو المنطوق، أو المرئي.

إن الخطاطة المنهجية Paradigm التي تتعامل بها المعالجات المحوسبة، التي توظّف لإنتاج المدوّنات المحوسبة بمختلف أشكالها، ومواضيع اهتماماتها، تتعامل مع النصوص الرقمية بوصفها نسيجاً مترابطاً يتألف من كيانات متباينة Objects تتألف من حرف أو مجموعة حروف، تفصل فيما بينها الفراغات أو العلامات، أو الرموز المقيمة فيما بينها.

وتتألف المعالجات المحوسبة من مرحلتين أساسيتين: انظر الشكل (١)

المرحلة الأولى: مرحلة المعالجات الأولية (الدكروري، أيمن. ١٤٣٩: ص ٧٩-٨٧)

المرحلة الثانية: مرحلة المعالجات المتقدمة.

وتوظف في كلا المرحلتين مجموعة من الخوارزميات البرمجية الذكية التي تقوم بتفكيك مفردات النص وتجزئته عناصره إلى وحدات لغوية أولية Tokens دون النظر إلى صنفها من أقسام الكلام، أو محمولاتها من المعاني، باعتماد مبدأ المسافة الفاصلة بين المفردات أساساً للتمييز بين كيان لغوي وآخر. ثم يصار إلى إزالة الأرقام والعلامات من النص؛ لتقليص عدد



المفردات التي ستعالج في مراحل لاحقة. وتأتي مرحلة إزالة كلمات التوقف (^{١٩}) Stop Words التي تتألف من أقسام الكلام التي لا تمتلك معنى قائماً بذاته مثل الحروف والضمائر، والعوامل، لاستبقاء المفردات التي تمتلك بعداً دلاليّاً، ولكي تكون مادة للتحليل من قبل المعالجات المتقدمة (رشوان، محسن والمعتز بالله، السعيد. ١٤٤١: ص ٩٣).

أما مرحلة المعالجات المتقدمة فتستخدم فيها خوارزميات أكثر تطوراً تستثمر المنطق اللساني والبنوي، لممارسة التحليل الصرفي Morphological Analysis وتحديد جذور الكلمات Root، وجذوعها Stem، وفئة المشتقات التي تنتمي إليها، ثم إزالة اللبس الدلالي عن طريق الإفصاح عن الحركات الطارئة على المفردات والتي تلعب دوراً أساسياً في تحديد المعاني الدقيقة لكل كلمة في السياق الذي وردت فيه. ثم تمارس عملية التحليل النحوي المحوسب Syntactic Analysis فيكشف عن هوية أقسام الكلام التي تنتمي إليها كل مفردة من مفردات النص، واحتساب مستويات حضور كل فئة من فئات أقسام الكلام.

ثم يصار إلى المباشرة بالتحليل الدلالي لمفردات النص عن طريق استثمار البيانات التي أفصحت عن الحركات الطارئة Semantic Analyses على المشتقات، وما تحمل من معاني، والروابط التي تقيمها كل مفردة مع المفردات المجاورة لها في السياق. وتمارس خوارزميات وسم هوية الكيانات (N.E.R) Named Entity Recognition عملية تحديد هوية أسماء الأعلام، والأماكن، والبلدان.

(١٩) كلمات التوقف هي المفردات التي يشيع استخدامها بكثرة في توليد الجمل وتماسك السياق اللغوي الذي يرد في النصوص، غير أنها لا تمتلك معنى قائماً بذاتها، ولا يؤثر حضورها على المعاني المودعة في السياق مثل: الضمائر، وحروف الجر، واللواحق التي تطرأ على الكلمة، وعلامات الترقيم، من أمثلة كلمات التوقف: (ب، ل، عن، من، في، و، ف، ثم، هو، هي، هم، هن، ال، التعريف، إن، أن، ما، كل، حيث...). ويصار إلى استبعادها في معالجات اللغة الطبيعية بسبب تراجع أهميتها، ولتوفير فضاء = معالجة أوسع للمفردات التي تمتلك بعداً دلاليّاً مهماً. وتسهم عملية إزالة كلمات التوقف في تقليل الأبعاد الدلالية المحتملة لألفاظ النصوص التي تعالجها خوارزميات الذكاء الاصطناعي، مما يقلل من فرص تشويش المعاني، وحجم المعالجة المطلوبة للوحدات الأولية بعد عملية الإزالة. إلا أنه من الضروري الانتباه إلى أن خلوا المعنى من كلمات التوقف لا يعني غياب الدور الذي يمكن أن تمارسه في تشكيل سياق المعاني، لذا فإن عملية تحديد كلمات التوقف، ينبغي أن تكون من خلال نهج معالجات اللغة الطبيعية؛ لضمان الحفاظ على السياق وبلوغ الغاية من إزالتها في الوقت ذاته.

وأخيراً تأتي مرحلة توظيف أكثر الخوارزميات المحوسبة ذكاء، والتي تتسم باستخدام شبكات عصبونية معقدة Artificial Neural Networks لممارسة تحليل تضمين الكلمات Word Embeddings (So, Aaron, ٢٠٢٤).

شكل (١) مراحل معالجة اللغة الطبيعية المحوسبة للنصوص



ملاحظة: الشكل من إعداد الباحث.

٤- الدراسات السابقة: قراءة ومراجعة:

أولت الكثير من مراكز البحوث والدراسات اهتماماتها بإنشاء المدونات وخصصت لها مواقع الشبكة العنكبوتية العالمية World Wide Web ليسهل الوصول إليها. وقد تنوعت حقول المعرفة التي تخصصت بها هذه المدونات، إلا أن القسم الأكبر منها قد عني باللغة العربية وعلومها، وطرق تعليمها، مع وجود مدونات عني بالقرآن الكريم.



ومن هذه المدونات: مدونة القرآن الكريم لجامعة ليدز^(١)، والمدونة اللغوية العربية العالمية لمكتبة الإسكندرية^(٢)، ومدونة الكلمات العربية^(٣)، ومدونة وكالات الأنباء العربية^(٤)، ومدونة خليج ٢٠٠٤^(٥)، وغيرها كثير. ومما يلاحظ أن جلّ هذه المدونات قد أعدت يدوياً، بعيداً عن استخدام معالجات اللغة الطبيعية المحوسبة، لذلك فإن مخرجاتها افتقرت إلى المعمارية البرمجية التي تجعل منها مورداً ثميناً لتطبيقات الذكاء الاصطناعي المعاصرة، مما جعلها حبيسة المواقع، وقوائم الاختيارات التي وضعها المصممون للمستخدمين، مع انحصار سبل تطويرها بإضافة مفردات جديدة، دون وجود فرص لتطوير المحتوى، أو انفتاحه على فضاء الخوارزميات الذكية الجديدة.

بالمقابل باتت المدونات المحوسبة تشكل مورداً معرفياً ثميناً وأصبحت ترتبط بنمطين من الدراسات: الدراسات المبنية على المدونات Corpus-based Studies حيث تشكل فيه المدونات فرصة لإجراء المزيد من البحوث اللسانية على الدوائر اللغوية لتحديد مستويات شيوع كلمات أو اصطلاحات، والتثبت من حضور ظواهر لغوية محددة؛ والدراسات التي تستمد توجهاتها من المدونة Corpus-directed Studies وهي دراسات تنبثق عن المحتوى الثري والمتنوع الذي يتسم به محتوى المدونة، والذي يثمر عن كشف ظواهر تستبطن في نسيج النصوص التي تتألف منها مادة المدونة المحوسبة، والتي أضحت مادة خصبة للمعالجات المحوسبة مقارنة بالمدونات التقليدية التي تتسم بتوطنها في النص المطبوع فلا يمكن ممارسة المزيد من عمليات التنقيب في محتواها، أو الكشف عن العلاقات الدلالية التي تربط بين مفرداتها. (الدكروري، أيمن. ١٤٣٩: ص ٢٧).

وقد شهدنا ولادة الكثير من المدونات المحوسبة خلال السنوات الأخيرة، فهناك أكثر من مدونة محوسبة أنشئت في جامعات ومراكز بحث سعودية وفق نهج معالجات اللغة

(١) تتوفر على الموقع: <https://corpus.quran.com/>.

(2) <http://www.bibalex.org/fica/ar/>.

(3) <https://sourceforge.net/projects/arabicwordcorpu/files/>

(4) <http://www ldc.upenn.edu/>

(5) <https://metatext.io/datasets/khaleej-2004-corpus>

الطبيعية، وتمتلك أكثر من نمط من أنماط المعالجات اللسانية المحوسبة، منها: مدونة اللغة العربية الكلاسيكية KSUCCA التي أعدت في جامعة الملك سعود بن عبد العزيز^(١)، ومنصة فلك^(٢) التي تعدّ من المبادرات المميزة لمجمع الملك سلمان العالمي للغة العربية.

كذلك فقد نهضت جامعة ليدز بمهمة إنشاء مجموعة متنوعة من المدونات الناطقة باللغة العربية والتي طرحتها على مواقع الويب، ولعل أهمها موسوعة القرآن الكريم^(٣)، كذلك هناك مشاريع أخرى تعنى بالتحليل الصرفي المحوسب للغة العربية، وأخرى لتحليل الخطاب، ولديها مشروع ضخيم يعنى بالخزيرة القرآنية العربية (Sharaf, A, Atwell, ES, Dukes, (K et al. 2010: pp.5.6.11).

على صعيد آخر، أولت مدينة الملك عبد العزيز للعلوم والتقنية جهوداً حقيقية لترسيخ آليات ومعالجات اللغة الطبيعية المحوسبة، وفوّرت ثماني تطبيقات لسانية محوسبة (Bataineh, Ruba & Al-Wazzan, Rawan. 2014: pp.1514-1520) هي: آلة التشكيل العربي KACST Arabic Diacritiser، وقاعدة بيانات الصوتيات العربية Arabic Phonetics Database،

(١) تعد مدونة اللغة العربية الكلاسيكية (KSUCCA) التي أعدت في جامعة الملك سعود بن عبد العزيز من المدونات التي أعدت بواسطة معالجات اللغة الطبيعية المحوسبة والتي تألفت من أكثر من ٥٠ مليون كلمة انتخبت من القرون الأربع الهجرية الأولى. وقد مورست عملية التحليل الصرفي المحوسب لهذه المدونة بواسطة مكتبة MADA Tool الموجودة ضمن مكتبات لغة Python البرمجية.

Alrabiah, Al-Salman, and ES Atwell, The Design And Construction Of The 50 Million Words KSUCCA, In Proceedings of WACL2 Second Workshop on Arabic Corpus Linguistics, 2013, pages 5-8.

(٢) تعنى هذه المنصة بتوفير مادة لغوية غنية يمكن اعتمادها في تمثيل العينات اللغوية التي تصلح لدراسة الظواهر المتعلقة بالعربية، مع إتاحة أحدث الأدوات في تحليل البيانات اللغوية، واستقرانها. ويعمل المجمع على إضافة مدونات جديدة باستمرار، وكذلك أدوات حاسوبية للبحث والتحليل لتلبي الاحتياجات العلمية للباحثين في المجالات اللغوية والحاسوبية. انظر: <https://falak.ksaa.gov.sa/aboutus>.

(٣) تعد مدونة القرآن الكريم التي أعدت في جامعة ليدز، من المدونات العريقة التي أعدت يدوياً، فوسمت فيها أقسام الكلام في القرآن الكريم، مع توفير التحليل الصرفي لكلماته، وعرض الشبكة الدلالية للكلمات بالإضافة إلى المتلازمات اللفظية بحسب جذور الكلمات، وما زالت التحديثات مستمرة على محتوى المدونة للارتقاء بدقتها. ولا يلاحظ فيها أي حضور لمعالجات اللغة الطبيعية المحوسبة. انظر: <https://corpus.quran.com/>.



وكذلك النظام التجريبي لتحويل النص العربي إلى كلام منطوق KFUPM's Experimental Arabic Text-To-Speech System الذي أعدّ في جامعة الملك فهد للبترول والمعادن، والمدونة العربية KACST Arabic Corpus، ومستودع الصوت العربي السعودي The Saudi Accented Automatic Stochastic Arabic Spelling، والمصحح اللغوي العربي الآلي Arabic Voice Bank، والمحلل الصرفي العربي Correction، والمحلل الصرفي العربي Arabic Morphological Analyzer and Arabic Stemmer، والمعاجم الصوتية العربية لتمييز الكلام Arabic Phonetic Dictionaries for Speech Recognition.

ويلاحظ خلال العقد الأخير تنوع الدراسات التي أجريت لحوسبة المدونات العربية، وتنوع حقول معالجتها. بيد أن ما تظهره مراجعة الدراسات في هذا المضمار تركّز الاهتمام بمدونات اللغة العربية والقرآن الكريم، مع غياب الدراسات عن كثير من الحقول المعرفية لبقية العلوم الإسلامية في هذا المضمار مثل: التفسير، والعقيدة، والفقه، وأصوله. فقد أنشأ Jarrar وآخرون مدونة Wjood⁽¹⁾ المحوسبة (Jarrar, Mustafa et al. 2022: pp.3628) وفق معالجات اللغة الطبيعية؛ لتمييز ووسم الكيانات اللغوية العربية المتداخلة Nested Arabic N.E.R والتي ضمت ٥٥٠ ألف كلمة. أما Elhassan وآخرون (٢٠٢٣) فقد أنشؤوا أنموذجاً محوسباً وظفت فيه آليات التعلّم العميق وتضمنين الكلمات لدراسة توجهات الآراء Sentiment Analysis المودعة في منصات التواصل الاجتماعي. بينما استعرض Abumallo وآخرون (٢٠١٦) مصادر لسانيات المدونات ذات الصلة بوسم أقسام الكلام P.O.S ومراجعتها على عينة تم اختيارها يدوياً من مصادر متعددة.

وقام Ismail وآخرون (٢٠١٤) بطرح مجموعة الأدوات التي تستخدم في لسانيات المدونات لمباشرة التحليل الدلالي المحوسب لمجموعة من النصوص التاريخية لتتبع التغيرات التي طرأت على المعجم الدلالي خلال ١٥٠٠ سنة، كما يمكن أن تستخدم هذه الأدوات في حقول أخرى. واقترح Eddakroui وآخرون (٢٠٢٣) عشرة مراحل لإنشاء مدونة عربية لعلوم المكتبات، مع مناقشة الدور الذي يمكن أن تمارسه مثل هذه المدونات في إثراء علوم

(1) Jarrar, Mustafa, Khalilia, Muhammed & Sana Ghanem, Wjood: Nested Arabic Named Entity Corpus and Recognition using BERT, Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, 2025- June 2022, pp.3628.

المكتبات والمعلومات . بينما قام Bataineh وآخرون (٢٠١٤) بإجراء مسح حول أهم التطبيقات التي أعدت في مدينة الملك عبد العزيز للعلوم والتقنية بمضمار اللسانيات المحوسبة وما يمكن أن توفره من فرص لتطوير المعالجات اللسانية المحوسبة في حقول العلوم المختلفة .

أما Belinkov ومجموعة من الباحثين (٢٠١٦) فقد نجحوا باستثمار المحتوى الرقمي للموسوعة الشاملة في إنتاج مدونة تاريخية ضخمة تألفت مادتها من مليار مفردة، انتخبت من نصوص تعود إلى حقبة زمنية مختلفة . وقد أجريت مختلف أنماط معالجات اللغة الطبيعية على محتوى المدونة التي استهدفت استثمار مادتها في تحديد السياقات المتوازية للنصوص، وانتمائها إلى مختلف الحقبة التاريخية بواسطة هذا النمط من المعالجات .

٥- موارد المدونة المحوسبة لأصول الفقه الحنفي:

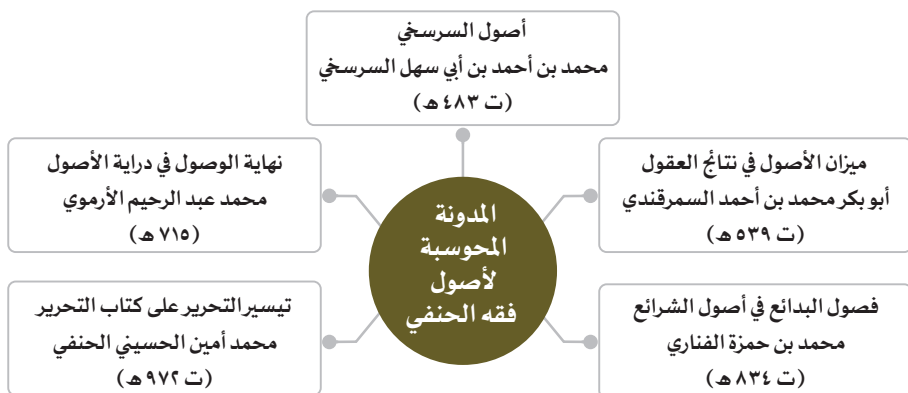
تزرخزانة أصول الفقه الحنفي بكتب ومصادر قيمة تناولت مبادئ أصول الفقه الحنفي ومسائله المختلفة . وقد حاولنا انتخاب خمسة مصادر من هذه المصنفات دوّنت خلال الحقبة الزمنية الممتدة بين ٤٨٣ هـ إلى ٩٧٢ هـ؛ لضمان استيعاب التنوع الذي طرأ على المصطلح الأصولي لهذه المدرسة الأصولية، والتي شكّلت مادة ثرية لقواعد بيانات مهمة تدعم سعيينا نحو تشكيل مدونة أصولية محوسبة تعنى بأصول الفقه الحنفي، ولتكون بادرة لشحذ همم العاملين في مجال أصول الفقه لإضافة المزيد من أمهات هذا العلم الجليل إلى مادة هذه المدونة، والسعي إلى تطوير مادة المحتوى لكي تكون مرجعاً يمكن الاعتماد عليه في النماذج اللغوية الضخمة Large Language Models التي ستنهض بمهمة تسخير خوارزميات الذكاء التوليدي Generative A.I. لتحليل الخطاب وإنتاج المعاني - انظر الشكل (٢) .

وقد حصلنا على هذه النصوص من مستودعات المكتبة الشاملة^(١)، بعد أن تم تحويلها من نظام ملفات الشاملة . bok إلى نظام ملفات Word بواسطة الأدوات التي توفرها هذه المكتبة لمستخدميها .

(١) موقع المكتبة الشاملة: <https://shamela.ws/>.



شكل (٢) موارد المدونة المحوسبة لأصول الفقه الحنفي (السرخسي، أبو بكر، ١٣٩٥)،
(السمرقندي، علاء الدين، ١٤٠٤)، (الأرموي، صفي الدين، ١٤١٦)، (الفناري، محمد، ١٤٢٧)،
و (الحنفي، محمد أمين، ١٤١٧).



ملاحظة: الشكل من إعداد الباحث.

٦- إعادة الهيكلة والتحليل اللساني المحوسب لمحتوى موارد المدونة:

في البداية بوشرت عملية معالجة كل نص من نصوص موارد المدونة الحنفية المحوسبة بمفرده. وكانت الخطوة الأولى مع تفكيك النص إلى وحداته الأولية، وتميز كلمات التوقف. ثم بوشرت عملية التحليل النحوي المحوسب فكشف النقاب عن أقسام الكلام في هذه النصوص التي ناهزت ثمانية وعشرين فئة، ثم باشرت خوارزميات التحليل الصرفي المحوسب بالتنقيب عن الجذور اللغوية، ومشتقاتها، ومحاولة كشف اللبس الدلالي حيثما وجد. وأخيراً، تم وسم هوية الكيانات في نصوص الموارد حيث حددت هوية الأعلام والأماكن مع احتساب تكراراتها (رشوان، محسن والمعتز بالله، السعيد. ١٤٤١: ٩٣).

وقد أودعنا في الجدول (١) خلاصة منتخبة من مخرجات التحليل المحوسب لنصوص موارد المدونة الحنفية، والتي يمكن أن توفر مشهداً أولياً عن السمات العامة لهذه النصوص، وتشكل مدخلاً للتحليلات اللاحقة التي ستسمح بعمق وتفصيل أكبر في تحليل محتوى نص مدونة أصول الفقه الحنفي.

جدول (١) أهم مخرجات معالجات اللغة الطبيعية على موارد مدونة أصول الفقه الحنفي

المورد	التحليل الأولي وتفكيك النص			أهم مخرجات التحليل النحوي			أهم مخرجات التحليل الصرفي		وسم هوية الكيانات	
	عدد الوحدات الأولية	التوقف	عدد كلمات التوقف	النسبة كل نص	عدد الأسماء	عدد الصفات	عدد الأفعال	عدد الجذور اللغوية	عدد الاشتقاقية	أسماء الأماكن
أصول السرخسي	٢١٩,٣٣١	٩٠,٢٢٦	%٤١,١	١٠٣,١٦٦	١٠,٠٩٧	٣٦,٤٥٦	١,٨٧٨	٤٣٨	٢٤٢	٢٣
ميزان الأصول	١٦٣,٤١٩	٨٣,٣٧١	%٥١,٠	٧١,١١٢	٨,٢٦٢	٢٣,٦٣٠	١,٨٣٦	٣٧٩	٥٩٤	١١١
نهاية الوصول	٥١٨,٧١٢	٢٦٥,١٠٤	%٥١,٠	٢١٨,٠٢٨	٣٤,٣٦٩	٧٩,٢٢٤	٢,١٧٩	٦٧٩	٥٣٠	٣٩
فصول البدائع	٢٩٢,٦٦٢	١٢٣,٧٧١	%٤٢,٣	١٤٦,٠٧٣	٢١,٨٠٠	٣٨,٨١٢	٢,٣٢٥	٤٧٦	٥٤٥	٣٣
تيسير التحرير	٥٠٧,٦٣٨	٢٣٤,٤٣٢	%٤٦,٢	٢٤,٧٠٠	٣٤,٦٢٠	٦٩,١٤٣	٢,٣٣٣	٦٤٣	٩٤٢	٦٣

ملاحظة: أعد الجدول من البيانات التي أنتجتها عمليات المعالجة المحوسبة للنصوص الأصولية المنتخبة.

٧- إنشاء المدونة المحوسبة لأصول الفقه الحنفي:

تم تشكيل المدونة المحوسبة لأصول الفقه الحنفي من مادة النصوص الأصولية^(١) التي انتخبت لهذا الغرض وبعد إجراء سلسلة العمليات ذات الصلة بمعالجات اللغة الطبيعية على هذه النصوص، وتفكيك مادتها إلى وحدات أولية Tokens، لكي تكون صالحة للمعالجات المحوسبة اللاحقة.

(١) يقصد بالنصوص الخام نصوص الكتب الأصولية التي عثر عليها من المستودع الرقمي للمكتبة الشاملة، وهي نصوص في ملفات Word لأصول هذه الكتب الأصولية، دون إجراء أي معالجة من معالجات اللغة الطبيعية على مادتها. وتعد النصوص الخام بصيغتها الرقمية جاهزة لعمليات المعالجة والتنقية التي تمارسها عليها الخوارزميات البرمجية.

وقد أثمرت عمليات إعادة تحليل نص المدونة المحوسبة بواسطة خوارزميات معالجة اللغة الطبيعية التي تتوفر في منصة برمجيات لغة Python عن تفكيك نص المدونة إلى مجموعة من الوحدات الأولية^(١)، وكلمات التوقف، التي تم حصرها، مع احتساب النسبة التي تشكلها كلمات التوقف من العدد الكلي للوحدات الأولية - انظر الشكل (٣).

شكل (٣) الخصائص العامة للمدونة المحوسبة لأصول الفقه الحنفي



ملاحظة: أعد الشكل من قبل الباحث.

٨- التحليل النحوي المحوسب لنص المدونة المحوسبة:

استخدمت خوارزمية ذكية استثمرت القدرات التي توفرها مجموعة من المكتبات البرمجية المتوفرة في البيئة البرمجية للغة Python لممارسة التحليل النحوي لنص المدونة المحوسبة. وقد نجحت خوارزميات المكتبة البرمجية CAMEL في تحديد أقسام الكلام لجميع مفردات المدونة المحوسبة، مع إحصاء مستويات حضور أقسام الكلام المختلفة، والتي ناهزت تسعة وعشرين قسماً من أقسام الكلام^(٢). انظر الشكل (٤).

(١) استخدمت مكتبة Panda لتحليل مادة النصوص إلى وحداتها الأولية واحتساب التكرارات، ومتغيرات أخرى.

(٢) استخدمنا مكتبة CAMEL وملحقاتها في عمليتي التحليل النحوي والصرفي لمحتوى المدونة الأصولية ومصادرها.

Morpheme	Frequency (approx.)
اسم	7000
فعل	2300
حرف	2100
صفة	1000
اسم موصوف	700
حرف عطف	700
حرف مشترك	600
حرف تاني	500
ضمير إشارة	300
علامة ترقيم	300
ضمير موصول	300
ضمير	200
فعل زائد	150
ضمير استنهام	150
اختصار	150
ظرف	100
حرف	50
حرف توكيد	50
حرف قبل	50
رقم	50
ظرف موصول	50
حرف استنهام	50
حرف تعجب	50
ظرف استنهامي	50
حرف نداء	50
اسم تكرار	50

وتضمنت مخرجات التحليل الصرفي لمحتوى المدونة المحوسبة لأصول الفقه الحنفي، جذر كل مفردة من مفرداتها، مع بيان ترتيب الأحرف الساكنة والمتحركة، التي تلتصق بالجذر (السوابق واللاحق)، مع بيان الشكل القاموسي لكل كلمة، وبيان كون الكلمة تصف جنساً محدداً، أو عدداً، مع تحديد الحالة الإعرابية، وبيان أداة التعريف، وتناوبات الأسماء والصفات،



والضمائر، وهوية الحركات الطارئة على الكلمة نتيجة لحضورها في السياق، وغيرها من التفاصيل التي أنتجها التحليل الصرفي المحوسب من محتوى المدونة المحوسبة.

وقد انتخبنا أنموذجاً مبسطاً لمخرجات التحليل الصرفي المحوسب للجذر اللغوي (ف.ق.هـ) حيث ذكرت جميع الاشتقاقات الصرفية لهذا الجذر التي استخدمت في السياقات الأصولية بنص المدونة - انظر الجدول (٢).

جدول (٢) أنموذج منتخب من التحليل الصرفي وكشف الالتباس عن مشتقات الجذر الثلاثي (ف.ق.هـ)

الجذر	البنية الصرفية	الكلمة	الجذر	البنية الصرفية	الكلمة	الجذر	البنية الصرفية	الكلمة
ف.ق.هـ	ال + فقه	الفقه	ف.ق.هـ	ال + فقهَاء	الفُقهاء	ف.ق.هـ	فقه	فَقِهَ
ف.ق.هـ	ال + فقه	الفقه	ف.ق.هـ	ال + فقهَاء	الفُقهاء	ف.ق.هـ	فقه	فَقِهَ
ف.ق.هـ	ال + فقه	الفقه	ف.ق.هـ	ال + فقهَاء	الفُقهاء	ف.ق.هـ	فقه	فَقِهَ
ف.ق.هـ	ال + فقه	الفقه	ف.ق.هـ	ف + ال + فقه	فَالْفَقِهَ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	يُفقه + هـ	يُفَقِّهه	ف.ق.هـ	ف + ال + فقه	فَالْفَقِهَ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	يُفقه + هـ	يُفَقِّهه	ف.ق.هـ	ف + ال + فقه	فَالْفَقِهَ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	يُفقه + هـ	يُفَقِّهه	ف.ق.هـ	ف + ال + فقه	فَالْفَقِهَ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	تَفَقَّهوا	تَفَقَّهُوا	ف.ق.هـ	نَفَقَه	نَفَقَهُ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	تَفَقَّهوا	تَفَقَّهُوا	ف.ق.هـ	نَفَقَه	نَفَقَهُ	ف.ق.هـ	أفقه	أَفَقَهَ
ف.ق.هـ	تَفَقَّهوا	تَفَقَّهُوا	ف.ق.هـ	نَفَقَه	نَفَقَهُ	ف.ق.هـ	و + ال + فقه	وَأَفَقَهَ

[illegible]



الجزر	البنية الصرفية	الكلمة	الجزر	البنية الصرفية	الكلمة	الجزر	البنية الصرفية	الكلمة
ف.ق.هـ	ال+فقيه	الفقيه	ف.ق.هـ	و+فقه	وفقهه	ف.ق.هـ	ب+فقه	يفقهه
ف.ق.هـ	ال+فقيه	الفقيه	ف.ق.هـ	فقه	فقه	ف.ق.هـ	و+أفقه	وأفقهه
ف.ق.هـ	ال+فقيه	الفقيه	ف.ق.هـ	فقه	فقه	ف.ق.هـ	و+أفقه	وأفقهه
ف.ق.هـ	ال+فقيه	الفقيه	ف.ق.هـ	فقه	فقه	ف.ق.هـ	و+أفقه	وأفقهه
ف.ق.هـ	ال+فقهاء	الفقهاء	ف.ق.هـ	فقه	فقه	ف.ق.هـ	و+أفقه	وأفقهه

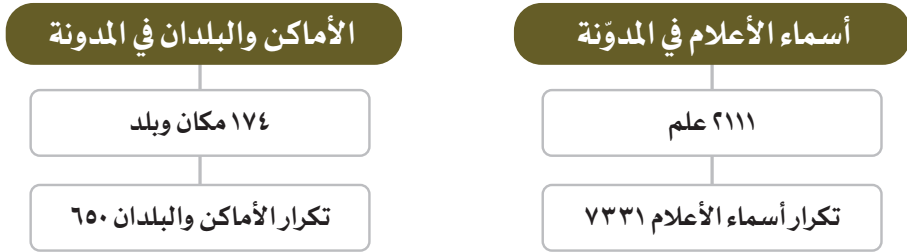
أعد الجدول من حزمة بيانات التحليل الصرفي المحوسب لمحتوى المدونة الحنفية المحوسبة

١٠- وسم هوية الكيانات Named Entity Recognition:

تعد عملية وسم هوية الكيانات من المسائل المهمة في توليد المدونات المحوسبة، حيث توفر مجالاً تنظم فيه أسماء الأعلام، والأماكن، والبلدان التي وردت في محتوى المدونة المحوسبة، والتي تعد مورداً ثميناً في إنتاج المعرفة، وتحديد مستويات حضور الكيانات (الأعلام، الأماكن والبلدان) في نسيج النص. (Najat et al. 2023: pp.448-457).

وقد استخدمت خوارزميات برمجية من مكتبة STANZA التي تعمل في البيئة البرمجية للغة Python؛ لتمييز هوية الكيانات، وتحديد انتماءاتها إلى حقول الأعلام، أو الأماكن والبلدان، ثم احتسب مستوى حضور كل منها في محتوى المدونة المحوسبة. انظر الشكل (٥).

شكل (٥) نتائج الكشف عن هوية الكيانات في المدونة المحوسبة مع بيان تكرارات حضورها



ملاحظة: أعد الباحث هذا الشكل من البيانات التي أنتجت من عملية تحليل هوية الكيانات في نص المدونة المحوسبة.

II- تحليل بعض جوانب الخطاب الأصولي في المدونة المحوسبة لأصول الفقه الحنفي:

مما لا شك فيه أن الغاية الأساسية لإنتاج المدونة المحوسبة لأصول الفقه الحنفي هي توفير مادة مهيكلة من المفردات التي وسمت نحويًا، وصرفيًا، ودلاليًا، مع تحديد هويتها؛ لتكون مادة جاهزة لمعالجات الحوسبة الذكية بصورة عامة، والذكاء الاصطناعي التوليدي بصورة خاصة.

ولما كانت اللغة العربية تحظى بمكانة متميزة في دائرة علم أصول الفقه؛ لأن الاستدلال على الأحكام الشرعية مرتتهن بفهم الخطاب الشرعي الذي لن يكون سليماً ما لم ينضبط بالضوابط اللغوية والأصول الشرعية المتعلقة بالمعاني والبيان؛ يبرز دور تراكيب اللغة العربية في الإفصاح عن الكثير من المسائل التي يعتمد عليها الأصولي عند معالجة منطق الأحكام الشرعية الوضعية والتكليفية. (بنعمر، محمد. ٢٠١٤: ص ٤).

من أجل هذا سعيينا إلى تعميق عملية التنقيب في مادة المدونة المحوسبة من خلال استخدام الجذور اللغوية للاصطلاحات التي شاع استخدامها في دائرة المصطلح الأصولي



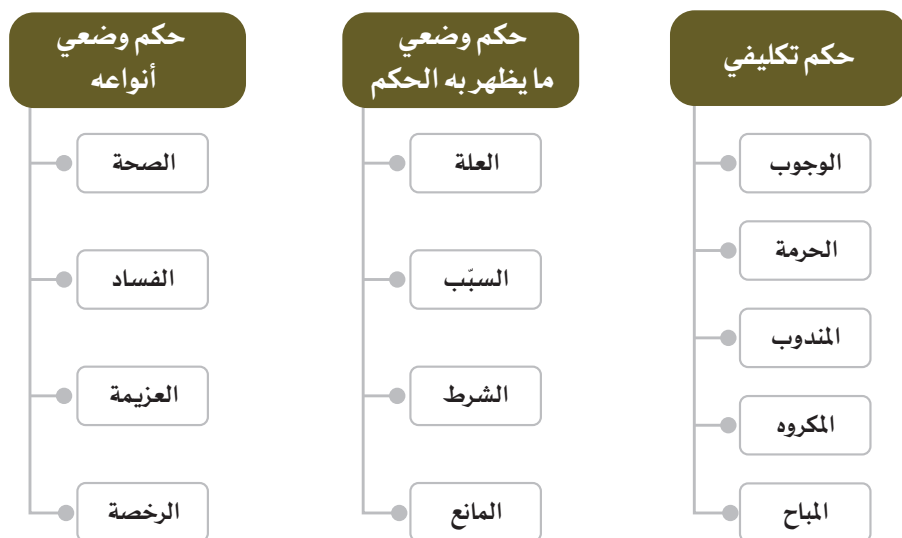
للمذهب الحنفي؛ لبيان الفرص التي ستتوفر أمام الباحثين عند اللجوء إلى توظيف أدوات الذكاء المحوسب، والذكاء الاصطناعي التوليدي لسبر مادة المدوّنة، وإنتاج معان جديدة، أو تحليل شبكات المفاهيم التي تستوطن في الخطاطة الأصولية للمذهب الحنفي.

وسنسعى في هذه الفقرة إلى بيان جزء يسير من فرص استثمار محتوى المدوّنة المحوسبة لسبر جوانب كثيرة من الخطاب الأصولي، والكشف عن تسخير اللغة لدى الأصوليين في مناقشة المسائل وترسيخ أسس الخطاب الأصولي للمذهب الحنفي.

١٢- الأحكام الوضعية والأحكام التكليفية في المدوّنة المحوسبة:

تعد مسألتنا: الأحكام الوضعية والأحكام التكليفية، من المسائل التي تستأثر باهتمام الأصوليين، كونها تشكل لباب الأحكام الشرعية. وقد سعينا إلى بيان عناصرها تين المسألتين والتعبير عنها بالشكل (٦).

شكل (٦) عناصر الأحكام الشرعية الوضعية والتكليفية



ملاحظة: أعد الشكل من قبل الباحث

وقد حاولنا استثمار مادة المدونة المحوسبة في الكشف عن نهج التعامل مع هاتين المسألتين من خلال تسخير مخرجات التحليل المحوسب، النحوي والصرفي؛ لتتبع حضور مفردات هاتين المسألتين، والاشتقاقات اللغوية التي استخدمت للتعبير عنهن، والتي تعكس بجلاء ثراء الذخيرة اللغوية لدى الأصوليين الأحناف، ورسوخ الصناعة الاصطلاحية لديهم. انظر الجدول (٣).

جدول (٣) مستويات حضور مفردات الأحكام التكليفية والوضعية في الخطاب الأصولي بالمدونة المحوسبة

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
إحكامنا / كأحكام / والإحكام / لأحكامه / لحكمنا / محكمة / وتحكمه / الحاكمين / والحاكم / ولحكمته / فالحكوم / حكيمه / لحكمة / كحكمته / فيستحكم / حكيماً / والتحكم / والمحكم / إحكاماً / الحكومة / لحكمته / تحكمه / محكمي / نحكم / بمحكمه / محكما / الحاكم / بحكمي / تحكم / فحكمهم / تحكما / بحكمهم / بحكومة / وتحكم / بمحكمته / أحكمت / الحكم / حاكمه / وكحكمه / حكميه / الحكماء / الحاكمة / وللأحكام / يحكمونك / التحكم / كالإحكام / فأحكم / حكمتهم / يحاكم / وللإحكام / حكامه / بالمحكم / فحكمتهما / أحكامهما / بالحكمة / فيحكمان / حكمها / لحكم / فالحكمة / إحكامها / وحكما / وإحكاماً / [حكمك / لأحكامها / يتحكم / فحكمته / بالحكمين / كالمحكم / الأحكام / فسيتحكم / بمحكوم / يحكموا / وكحكمته / يحكما / أحكم / محكم / لحكما / فتحكم / كالحكم / إحكام / حكمهم / محكومة / والحكومات / وإحكامه / ولحكم / أحكموا / الإحكام / بالحكام / الحكيم / حكمين / وبحكم / لمحكمه / فالمحكم / وبالحكم / وأحكام / يحكمان / فإحكام / بحكمة / حكمة / أحكام / فمحكم / حكمية / حكميته / حكماً / وحكموا / وأحكامه / المحكمات / ولحكم / حكمان / وليحكم / بحكم / كالتحكم / والمحكوم / وإحكام / أحكامي / لإحكام / وحكمهما / والحكمية / بالإحكام	٢٦٥	ح.ك.م



الأحكام التكليفية		
الجذر	عدد المشتقات	المشتقات اللغوية
ح.ك.م	٢٦٥	/ واستحكام / الأحكام / إحكامهم / لحكمين / الحكمة / بحكمين / / وأحكامه / فالأحكام / لحكمتها / حكمهن / المحاكمة / ليحكم / حكمه / حكمت / فإحكامها / للتحكم / للحاكم / حكمته / أحكامنا / لإحكامها / وبالأحكام / تحكموا / ويحكم / وحكمة / كتحييمه / كحكم / لتحكم / بأحكام / كإحكام / وحكمنا / وأحكم / أحكامهما / للحكمين / وبالإحكام / حكيم / والحكمي / كحكمها / لحكمهما / إحكامهما / الحكمين / والحكيم / بحكمهما / لحكمه / محكمه / حكمهما / حاكم / حاكمة / وأحكامها / الحكميات / إحكامي / وحكمها / حكموا / حاكما / بالحاكم / بالمحكوم / كتحييمهما / وإحكامهما / حكما / يحكمون / لإحكامها / / لمحكمة / حكم / تحكم / لإحكامه / لحكمهم / محكوم / أحكاما / والحكم / حكماء / فالإحكام / حكمكم / وأحكاما / فيحكم / وأحكامهما / للحكمة / حكمتها / فأحكامها / حكميها / كحكمه / بحكمها / المحكم / أحكامهم / أحكامه / المحكمة / للإحكام / وتحكيم / حكميها / وإحكامها / والحكمة / لأحكام / كمحكمات / ومحكم / فأحكام / للحكم / فحكمه / لحاكم / وحكم / تحكيم / الحكمان / بالحكم / فحكمنا / والحكمان / بالأحكام / حكمنا / حكمي / محكوم / فللحكم / فحكم / بحكمه / الحكمة / لحكمها / الحكمي / ومحكوم / المحكوم / والمحكمة / محكمات / فليحكم / بحكمته / حكيمة / والحكمين / حكمك / وحكمه / ويحكموا / يحكموك / محكوما / وكالحكم / بإحكام / والحكميات / للأحكام / حكام / ولحكمه / والأحكام / إحكامه / وحكمهم / الحكمتين / حكمتها / محكمان / وأحكمت / كالأحكام / وحكمته / يحكم / للمحكوم / كالحكمة / فالحكم / فحكمته / والحكماء

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
فتكليف / بكلفه / كلفه / يكلفه / تكليفهم / وتكليف / تكليفا / للمكلفين / المكلفين / نكلف / [المكلفين / يكلف / بالتكلف / والمكلف / تكليفه / كلفتين / فتكليف / الكلف / بكلفة / المكلفون / وتكلفوا / التكاليف / مكلفان / التكلف / مكلف / مكلفهم / التكاليفات / والتكاليف / ويكلف / والمكلفين / لتكلفه / فتكليفهم / لتكليفه / وتكليف / تكليف / فمتكلف / بتكلف / وكلفنا / يكلفوا / وكلفهم / كلفة / للمكلف / تكلف / وكلفة / فالمكلف / كلفتم / كتكليف /] وتكلف / كالتكليف / كلفتك / المكلف / وكلفه / تكلفنا / لكلفهن / كالمكلفين / فتكلفوا / بتكليفه / تكلفه / بمكلف / تكاليفه / بكلف / كلفهم / بتكليف / كلفتم / تكلفوا / بتكاليفه / لكلف / فالتكليف / كتكليف / بالتكاليف / فيكلف / بالمكلف / بالتكليف / فللمكلف / ومكلف / تكاليف / لتكلفه / كلفنا / يكلفنا / تكلفه / والتكليف / تكلفهم / وتكليف / وكلف / التكليف / كلف / والكلفة / المكلفة / فتكلف / يتكلف / وتكليفه / مكلفون / مكلفين / كلفوا / تكلفا / متكلف / فيكلف / مكلفه / الكلفة / مكلفة / للتكليف / مكلفا / تكلفي	١٠٣	ك.ل.ف
فرائض / افتراضه / كالمفروض / بالفرضي / أفترض / لفرضية / وفرائضها / بفرضه / فرضيهما / فرضي / والمفروض / يفرض / وفرضا / كالفرائض / تفرضوا / ويفرض / [فالمفروض / كالفرض / افتراضا / تفرض / ويفترض / فرضت / لفرضية / الافتراض / كالمفروضة / بالفرائض / الفرضية / وفرضيه / مفروضه / افتراضته / يفترض / وفرضه / المفروض / لفرض / والفريضة / لفرضيه / فرضيه / والفرضية / فرضها / يفرض / فرضه / فرائضه / كفرض / المفروضة / افتراض / بالفرضية / ففرضوا / المفروض / فريضة /	١٠٩	ف.ر.ض



الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
مفترض / فرضنا / فرضا / الفرضي / فريضته / افترض / الفريضة / فارض / ففرض / بالفرض / وفريضة / وفرض / فرضوا / بالافتراض / ففرضيه / الفرائض / والفرض / مفروض / كفريضة / وفرائض / فلنفرض / بفرضه / فرضاه / بفرض / فارضه / المفروضات / ففرضيتها / للفرض / فرضتم / فريضته / بالمفروض / بالفرض / وفريضة / نفرض / لافتراضه / فالافتراض / وفرضت / فرضناه / أفرض / فرض / الفريضتين / فالمفروض / للفرائض / الفريضة] / فروضه / فليفرض / فرضية / فرضوه / وفرضنا / مفروضة / فروض / الفرض / والفرائض / وفرضية / بفرضية / فريضتين / بفرضيه / تفترض / فرضهم .	١٠٩	ف.ر.ض
لحرمة / والتحریم / حراما / محرمي / وللحرمة / الأحرام / كحرمانه / الحرام / بحرمه / أحرم / للتحریم / بالتحریم / وحرمه / وتحريمه / حرمانه / بالحرمة / وتحريما / والإحرام / حرمة / ولحرمة / إحرامه / بحرام / لحرمتها / كتحریم / للحرام / فحرمتها / لحرمه / كالتحریم / يحرمون / بالمحرم / لحرام / والمحرم / والحرمت / وتحريمها / فالحرمت / تحرم / ومحرمه / فحرام / لحرمته / والمحارم / فحرمانه / تحريمها / والحرام / الحرمان / الاحترام / والمحرمات / ومحرم / تحرموا / محارمه / للمحارم / محترمه / يحرم / التحريم / فيحرم / الحرم / كحرمان / احتراما / للمحرم / حرمه / حرمتا / وإحرام / فالتحریم / حرمانا / ومحرم / وكحرمة / الإحرام / وحرمت / الحرمتين / فتحریم / تحريمه / المحارم / محرمان / المحترم / المحرمين / فيحرم / بمحرمة / حرمتهم / بالمحرمات / محرمين / بتحريم / حرموه / وللتحریم / وكحرم / لمحرم / وحراما / يحرمه / يحرموا / والأحرام / حرامه / وتحريم / محترم / محرمه / كحرمة / بحرمان / حرما / فحرمت / وحرمان / كحرم / المحرمة / الحرمت / تحريما / وحرمانا / أحرمه /	٢١٦	ح.ر.م

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
فحرمتموه / حرام / أحرامه / بالحرمت / [إحرامه / وكحرمه / لحرم / ومحرمه / يحرم / فكحرم / ويحرم / كالمحرم / لتحريمه / كالإحرام / محرميه / وحرمتها / محرم / والحرمة / حرمان / الحريم / حرماها / فحرمة / تحريمكم / حرمة / بجرمة / للاحترام / فحرمة / المحروم / المحرم / للحرمة / الحرمين / بالحرمة / كالمحرمة / للإحرام / كحرمت / فالمحرم / محرما / وحرمتها / للحرم / فالتحريم / وحرمة / والحرمين / بمحرمه / فتحريمه / محرمة / المحرمات / فيحرمهما / حرمت / بتحريمه / بالحرم / فبتحريمه / تحرمون / والحرم / وتحرم / ولحرم / حرمتها / وحرموا / كحرمه / لأحرام / المحترمة / تحرمه / بالأحرام / وبالحرمة / وأحرام / لتحريم / يحرمهم / احترام / يحرمها / فحرم / الحرمة / وحرمتها / ولحرمه / فحرمتها / تحريم / لحرمانه / ليحرم / إحرام / وحرمت / فتحرم / والحرمان / حرم / لإحرام / حرموا / محترمة / كالأحرام / كالأحرام / كالحرمة / بالإحرام / فيحرمون / محرمات / حرمتها / بمحرم / فالحرمة / كحرمة / لتحريمه / فحرمة / وحرمة / كمحرم / وبتحريمه / محارم / بالحرام / بجرمته / لحرمان / وحرم / تحريمهما / بتحريمها / بجرمانه / محترما / بجرمتها / وحرام / أحرام / للأحرام / والحریم .	٢١٦	ح.ر.ح
ندبا / مندوبة / وندبة / [الندب / والندوب / وندب / وأندب / بالمندوب / ومندوبا / بمندوب / وندبه / فأنتدب / الندب / ندب / فندب / بالندب / فالندب / مندوبه / المندوب / لندب / يندب / ندبه / ندبتك / المندوبة / بندب / ندهم / فمندوب / والانتداب / المندوبات / كالمندوب / المندوبة / فانتدب / ندبة / الانتداب / للندب / ندييه / وندبتك / مندوبيه / فالمندوب / النادبة / كالندب / للمندوب / أنتدب / انتدب / ويندب / فندوب / فندبا / وندبا / والندب / ندبنا / والمندوب / مندوبيه / وللندب / ندبته / فللندب / فيندب / وندبييه / ندبني / والمندوبات / مندوب / بالمندوبات / ومندوب / مندوبا	٦٣	ن.د.ب



الأحكام التكليفية		
الجزر	عدد المشتقات	المشتقات اللغوية
ك.ر.هـ	٩٨	يكرهه / وللكراهة / وكارها / إكراها / الكراهة / أكرهت / فتركه / لكراهة / كراهة / كالمكره / والمكروهات / بالمكره / للكراهة / مكرهين / كرها / كالإكراه / كرهه / كراهية / والمكره / الكاره / مكروها / فكره / الإكراه / كريبه / المكروه / مكروهه / فالمكروه / بكراهة / المكروه / مكروه / مكروهة / بإكراهه / ومكروه / كريبه / يكرههن / بالإكراه / المكروهين / المكروهة / استكرهها / وكرهوا / وكرهت / أستكرهها / والكره / كارها / بالمكروه / وكره / كراهته / كراهتهم / المكروهان / إكراههن / بإكراه / للإكراه / والإكراه / كريبه / لنكره / المكروهات / كاره / كإكراه / يكره / إكراه / ويكره / والكراهية / [بكره / المكراه / أكرهه / مكرهة / بمكروه / والمكروه / استكرهوا / يكرهن / لنكره / كره / فالإكراه / ومكروها / فكرهوا / مكرهه / فيكره / ونكره / بكره / الكره / أكره / وكراهة / بإكراه] / والكراهة / الكراهية / كرهة / للمكروه / فمكروه / إكراهه / بالكره / كمكره / نكره / بالكراهة / فكرهت / تكرهوا / فالكراهة / وأكره / مكرها
الأحكام الوضعية		
الجزر	عدد المشتقات	المشتقات اللغوية
ح.ك.م	٢٦٥	إحكامنا / كأحكام / والإحكام / لأحكامه / لحكامنا / محكمة / وتحكمه / الحاكمين / والحاكم / ولحكمته / فالحكوم / حكيمه / لحكمة / كحكمته / فيستحكم / حكيماً / والتحكم / والمحكم / إحكاماً / الحكومة / لحكمته / تحكمه / محكمي / نحكم / بمحكمه / محكما / الحاكم / بحكمي / تحكم / فحكمهم / تحكما / بحكمهم / بحكومة / وتحكم / بمحكمته / أحكمت / الحكم / حاكمه / وكحكمه / حكميه / الحكماء / الحاكمة / وللأحكام / يحكمونك / التحكم / كالإحكام / فأحكم / حكمتهم / يحاكم / وللإحكام / حكامه / بالمحكم / فحكمتهما / أحكامهما / بالحكمة / فيحكمان / حكمها / لحكم / فالحكمة / إحكامها / وحكما / وإحكاماً / [حكمك / لأحكامها / يتحكم / فحكمته / بالحكمين / كالمحكم / الحكم / فسيحكم / بمحكوم / يحكموا / وكحكمته / يحكما / أحكم / محكم / لحكما / فتحكم / كالحكم / إحكام / حكمهم /

الأحكام التكليفية		
الجزر	عدد المشتقات	المشتقات اللغوية
ح.ك.م	٢٦٥	محكومة / والحكومات / وإحكامه / وللحكم / أحكموا / الإحكام / بالحكام / الحكيم / حكمين / وبحكم / لمحكمه / فالحكم / وبالحكم / وأحكام / يحكمان / فإحكام / بحكمة / حكمة / أحكام / فمحكم / حكمية / حكميته / حكميا / وحكموا / وأحكامه / المحكمات / ولحكم / حكمان / وليحكم / بحكم / كالتحكم / والمحكوم / وإحكام / أحكامي / لإحكام / وحكمهما / والحكمية / بالإحكام / واستحكام / الأحكام / إحكامهم / لحكمين / الحكمة / بحكمين / وأحكامه / فالأحكام / لحكمته / حكمهن / الماكمة / ليحكم / حكمه / حكمت / فإحكامها / للتحكم / للحاكم / حكمته / أحكامنا / لإحكامها / وبالأحكام / تحكموا / ويحكم / وحكمة / كتحكيمه / حكم / لتحكم / بأحكام / كإحكام / وحكمنا / وأحكم / أحكامهما / للحكمين / وبالإحكام / حكيم / والحكمي / حكمها / لحكمهما / أحكامهما / الحكمين / والحكيم / بحكمهما / لحكمه / محكمه / حكمهما / حاكم / حاكمة / وأحكامها / الحكميات / إحصامي / وحكمها / حكموا / حاكما / بالحاكم / بالمحكوم / كتحكيمهما / وإحكامهما / حكما / يحكمون / لإحكامها / / لمحكمة / حكم / تحاكم / لإحكامه / لحكمهم / محكوم / أحكاما / والحكم / حكماء / فالإحكام / حكمكم / وأحكاما / فيحكم / وأحكامهما / للحكمة / حكمتهما / فأحكامها / حكميها / حكمه / بحكمها / المحكم / أحكامهم / أحكامه / المحكمة / للإحكام / وتحكيم / حكميها / وإحكامها / والحكمة / لأحكام / كمحكمات / ومحكم / فأحكام / للحكم / فحكمه / لحاكم / وحكم / تحكيم / الحكمان / بالحكم / فحكمنا / والحكمان / بالأحكام / حكمنا / حكمي / محكوم / فالحكم / فحكم / بحكمه / الحكمية / لحكمها / الحكمي / ومحكوم / المحكوم / والمحكمة / محكمات / فليحكم / بحكمته / حكمية / والحكمين / حكمك / وحكمه / ويحكموا / يحكموك / محكوما / وكالحكم / بإحكام / والحكميات / للأحكام / حكام / ولحكمه / والأحكام / إحكامه / وحكمهم / الحكميتين / حكمتهما / محكمان / وأحكام / كالأحكام / وحكمته / يحكم / للمحكوم / كالحكمة / فالحكم / فحكمته / والحكماء



الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
يوضع / أوضاعه / لتواضعه / يتواضعا / وضعية / وضعة / لوضعة / وضعوه / ووضعة / موضع / للوضع / موضوعاتها / فوضعيه / الواضعين / وأضع / الموضوعات / فوضعا / لواضع / بموضوعه / لوضعه / والمتواضع / فيوضع / والضعة / يضعوها / موضوعه / بموضوع / فموضعه / وضعيان / بواضعة / فموضوعات / وضعيا / لموضعين / موضوعيته / بوضعي / وضعك / ليوضع / الموضوعان / ومواضعه / لموضوعيته / موضوعان / فيضع / الوضع / للموضوعات / مواضعه / وأضعها / فالوضع / الواضعون / ووضعا / لموضوعها / موضوعيه / مواضع / التواضع / فموضوعها / لوضعتيهما / وموضوعيته / [الوضعية / تضعه / فالموضوعة / واضع / موضوعيهما / بالأوضاع / للوضعي / فوضع / الموضوعين / كالوضع / أوضاع / متواضع / بالوضعية / كواضعة / بالموضوع / متواضعا / بوضعة / ويضع / المواضع / ووضعت / بالوضع / واضعهما / لمواضع / والتواضع / موضوعهما / ووضعا / بأوضاع / تواضع / وموضوعات / وضعهم / كوضعا / واضعا / ولوضع / كواضعه / وكمواضع / والأوضاع / يضعها / لوضعية / كالوضعية / لموضوعاتها / للمواضع / بواضع / فالواضع / فوضعة / [بموضع / يضعوه / واضعا / لوضع / المتواضع / بموضوعيه / ومواضع / وأضعهما / يضعوا / الموضع / موضوعا / وضعه / بالوضعي / كوضعة / توضع / وضعت / وضعتها / وتضع / وضعها / وضع / فوضعت / وضعيه / وضعي / وضعوها / بوضعهم / والوضعية / وموضوعاتها / وضعوا / الموضوع / بوضع / الموضوعية / تضع / كالموضوع / وضعهما / واضعين / كوضعتيه / فموضوع / موضوعية / الموضوعة / والواضع / وموضع / للموضوع / والوضع / بمواضع / اتضاع / مواضع / والموضوع / موضوع / وضعنا / لموضوعه / يضع / موضوعين / موضعه / موضعيهما / يضعن / فموضع / وضعتيه / بالمتواضع / الوضعية / موضعا / يوضع / للموضوعين / وموضوعها / بالموضوعة / أوضاعهم / بالموضوعات / بموضوعيته / يضعه / لموضوعيهما / بضعة / بواضعه / موضعي / فوضعية / وموضوعا / الأوضاع / أضع / أوضعت / بموضوعة / وأضعه / فليضع / أوضاعها / وموضوع / كوضعه / سأضعه / والوضعي / موضوعها / وواضع / الواضع	٢٢٤	ا.ض.ع

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
/ موضوعة / موضوعات / وازعة / الوضعي / فوضعة / لموضوع / ووضعا / وأوضاعه / للواضع / ووضع / وضعها / فوضعي / مواضعها / وموضوعه / الموضوعين / ووضعه / تواضعا / للموضع / واضعه / بوضعها / وضع / بوضعه / الواضعة / موضعها / لوضعيه / وموضعا / لوضعه / فالموضوع / وموضوعة / كوضع / لوضعت / بموضوعية / لموضوعة / فمواضع / فوضعه / موضعين	٢٢٤	ا.ض.ع
بعلتين / فالعلل / يعللوا / وعلية / والمعلول / فلعلهما / كتعليل / عليلا / عللها / عللوا / وتعلل / لمعلول / تعليلنا / وتعليلهما / العلتان / أعلاء / بتعليله / بالعلل / معلوله / كالمعتل / العلل / بالمعلول / فتعليل / يتعلل / ولعله / لعليتها / كالمعلل / المعلل / بمعلول / علية / فأعلل / لمعلولييه / لعلك / كالعلتين / كالمعلول / ومعلولها / / علله / فتعلل / فالعلية / واعتل / فلعلة / بعلته / علتين / للعلل / علة / فعله / العلتين / التعليلان / معلله / كتعليله / تعلل / وعله / فيعلله / وبالتعليل / علتيهما / تعليلكم / يعتل / المعلولات / للمعلل / وأعتل / فتعليلنا / فعلنا / معلولي / لأعلاء / معلولها / [كالعلل / وأعلاء / علتاهما / فنعلل / بعليتها / فيعلل / وتعليله / تعليلها / علتها / وكعلية / كالعلة / المعلولين / العليل / تعليليه / عللا / لتعليل / لعلكم / والعلة / علتته / علتت / لعل / والمعتل / وعلل / تعليل / علتي / معلنا / وعللها / لعلهما / علتان / ولعل / معلل / وعلة / التعليلات / وعلته / عله / لعلهم / فالمعلل / وعللنا / فعلية / للتعليل / وتعليلهم / علها / فعلته / الأعلاء / المعللين / علل / ولعلهم / لعله / تعليله / والتعليل / وكتعليل / يعلل / فاعل / / كالعلل / معلولة / فعلل / واعتللا / للعلة / لعليته / لمعلولها / العلية / بعلي / فعل / يعللونه / عللنا / لعلته / ويعلله / بالعلة / تعليل / فالعلة / لعللها / تعليلي / بالتعليل / بتعليلنا / بالعلة / والتعلة / تعليلاتهم / ولعلكم / بعليته] / بعلتها / نعلل / فاعلهم /	٢١٧	ع.ل.ل



الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
معلولاتها / بعليته / فعلتها / عليتها / لعللوا / ويعللونه / فالتعليل / والتعليلات / فلعلهم / بعلة / معلولات / المعلول / لعله / معلولا / وبعلة / فعلهما / والمعلل / لعلية / فعلة / لعلنا / كعلية / علته / ومعلل / معلول / بالعلية / عليته / كأعلاء / بعلية / أعلنه / بعلى / ولعله / وبعلته / المعلولة / بمعلوليته / لعلل / تعليلهم / لعلتين / التعليل / يعللون / علتك / للعلية / وللتعليل / لتعليله / والعلية / المعتل / كالتعليل / فلعله / فلعلها / وتعليلها / وعله / علهم / عل / للمعلول / معلليه / ولعلها / فلعلها / علتها / معتل / والعلل / وبالعله / فبعليته / معللا / وأتعلل / بتعليل / يعللها / معلوليته / ومعلول / وتعليل / كعله / لعلها	٢١٧	ع.ل.ل
بسببيه / فسبب / السببان / فيتسبب / أسبابهما / فنسب / بسبب / أسبابها / بالأسباب / والسباب / والمسبب / وسببيه / أسبابا / سببته / بسببها / ومسببهما / وأسبابا / والمسببات / وأسبابها / أسبابه / الأسباب / وأسباب / السبب / بالسبب / لسبب / مسببان / تسبب / للسبب / بسببه / سي / كالسبب / نسيه / لسببية / مسبب / بسببهما / وكسب / فسببها / لسببه / لسببهما / كالتسبب / يتسبب / سبا / لسببيته / ولسبب / تسببه / فسببية / وسابت / سيه / سبة / التسبب / كسيه / فللسببية / وسببته / المسببات / المسببين / لنسبه / متسببا / لسببيتها / سبها / بسبب / بأسبابهما / لأسباب / مسببات / كسبها / وكسبة / كسببية / سبب / مسببا / بأسبابها / بالتسبب / وسببا / ومسببا / المسببة / فسببه / يسبب / للسبب / وكسبه / تسبوا / سبباها / بأسباب / بمسبب / سبببيته / سبب / وسبب / السبي / كسب / لتسببه / السببين / بأسبابه / تسبب / للمسبب / ونسب / فالأسباب / فتسببه / فنسبه / كسببيه / نسبهما / سبي / بالسببية / السببية / كالسببية / للسببية / كأسباب / والسبي / والأسباب / ونسبهم /	١٧٠	س.ب.ب

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
/ مسبباتها / سببيتها / فسبب / فسببيه / كسي / سبان / وسبي / سبيتها / وسبوا / سبيه / السبب / فالسبب / نسبها / تسبب / كسبهم / سبت / والسبب / ونسبه / وسب / السبابة / سببها / بسبيين / مسبه / وسبت / كسبة / سبية / مسبها / فتسبب / وسبه / بمسبباتها / بالمسبب / أسباب / لمسبب / يسني / سببا / سبه / والمسبة / تسببوا سبة] / المسبب / مسبة / والتسبب / ويسبب / بسببية / وسببية / سببها / فالسببية / لسببيه / نسبهم / لنسب / التسبب / وسببها / بالتسبب / ويسبه / لأسبابه / للأسباب / ونسبهم / لتسبب / [كسبها / نسب / لتسببها / تسببها / والسببية / يسبه	١٧٠	س.ب.ب
شرطت / المشروط / فشرطيه / فيشترط / مشرطا / وشرطنا / شرائط / المشروط / تشرط / [لشرط / كالشرط / والشرطية / وشرطوا / نشترط / بشرطيته / شرطي / والشروط / مشروطها / الشرطية / لشرطة / اشتراطا / لاشتراطها / باشتراط / الشرطيات / وشرطية / اشتراطهن / شرطيته / أشرطة / والشرطي / وبالشرط / فالشرط / المشروطان / وشرط / لشرطه / نشترطه / شرطوه / فالاشتراط / أشرطا / بشرائط / للشرط / شروط / لشرائطه / بمشرط / لاشتراطه / لشرائطها / أشرطه / فالشرطية / ومشرطه / وشرطيتها / اشتراطه / فشرط / واشتراطها / بشروط / لاشتراطهم / مشروطة / المشروطة / اشتراطوه / وباشتراط / وشرطة / أشرطه / شرطان / بشرط / وشرطا / للمشرط / وشرطيته / شرطهما / ويشترط / شرطه / وشروط / فاشتراطه / بشرطها / الاشتراطين / شرطها / أشرط / كشرائط / شرطها / فالشرط / وأشرطها / شرطيا / والمشرط / بالشرطين / للشرط / ليشرط / فكشرط / بشرائطه / شرطهما / للشرائط / شرطنا / شريطه / شرطا / اشتراطه / وبالشرط / الشرطيتين / شرطيه / وأشرط / بالشرط / واشترط	٢٠٢	ش.ر.ط



الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
/ اشتراطهم / اشتراطها / واشترطنا / فاشترط / فاشترط / يشترطون / كالمشروط / شرطتم / شرطوا / واشترطه / وشرطه / بمشروطه / يشترطوا / تشتط / اشتطت / يشراطه / شرائطها [/ للشرط / شرائطها / شرطية / بالاشتراط / وشرائط / لشرط / الشرائط / وشرطها / شروطه / مشروطا / اشتراطهما / نشرط / شروطا / الأشراف / فشرطه / فليشترط / فشرطية / والشرائط / الاشتراط / ولاشتراط / الشرط / يشترط / فأشترط / بالشرط / باشتراطه / يشترطه / شرطة / شرطناه / وشرائطها / اشتطوا / بشرائطها / بشرطه / بمشروطة / وأشراف / وشرطيه / اشتراط / وبشرط / وللشرط / اشتط / كشرط / فشرطنا / كاشترط / شرطين / يشترط / للشرطية / وشروطه / فتشترط / المشروطين / بشرطين / الشرط / شريطة / شرائطه / بالمشروط / يشترطه / واشترط / والشرط / بشروطه / باشتراطهم / الشرطي / شارط / لشرائط / كاشترطه / فشرطة / وشرطت / يشترطها / بشرطة / أشراف / مشروطه / يشترط / وشروطها / كشرط / بشرطها / مشروط / للأشراف / أشرط / شرط / لاشترط / ومشروطا / الشرطين / بالشرائط / وشرائطه / فلاشترط / الشرطان / وشروطا / ويشترطه / يشترطوا / ومشروطة	٢٠٢	ش.ر.ط
/ بمانعيه / ممنوعا / يمتنعون / المانع / ليمنعه / ممنعاه / ممتنعين / وأمتنع / ومنعها / بممنوع / وللمانعين / فمنعه / مانعهم / ممنوعان / منعك / منعها / يمنعه / كالمتنع / منعها / منعه / فالمانعة / الممانعة / ويمنعه / فمانعة / فيمنعه / أمتنع / وامتناعه / ممتنع / امتناعهما / فلامتناع / الممتنعين / فالمتنع / لمتنع / والامتناع / الموانع / يمتنع / يمانع / امتناع / فممنوع / فللمانع / المانعة / لمنعه / بممتنعات / الممانعات / يمتنعوا / وامتنع / فممنوعه / كالمنع / وممتنعه / ممنوعكم / بمنعه / فتمتنع / فيمنعان / بامتناعه / ومانعا	٢٥٨	م.ن.ع

الأحكام التكليفية		
المشتقات اللغوية	عدد المشتقات	الجذر
/ وامتنعنا / لموانع / ومنعنا / لممتنع / يمنعان / وموانع / موانعه / منعنا / ممانعات / فمنعهم / فلمنع / الامتناع / ومنعت / منعتني / لمنعهم / لامتنع / يمتنعان / تمنع / يمنعا / منعوا / يمنعونهم / يمنعون / وامتناعا / فامتناع / يمنعي / وممانعة / فالممنوع / لمنعها / الممنوعة / فيمنع / فمنعني / لامتناعه / يمنحك / بمنع / لامتنعوا / فيمنعونهم / ليمنع / ويمنع / المانعون / منعه / ومنعها / فمنعوا / فامتناعه / مانعين / الممتنع / وبامتناع / مانعة / كالممانعة / فأمتنع / والمنعة / ممتنعاته / يمنعان / وبمنع / منعهن / فمنع / فمنعه / مانع / وامتناع / فلمانع / ومنعهم / ويمتنع / ونمنع / لامتناع / مانعه / ممنوعه / ويمنعون / يمنعونه / مانعو / فمنعه / لمانعي / فيمتنع / تمنعها / فمنعته / كمنعه / ممتنعه / منعاه / كمنعة / تمنع / منعي / والممتنع / تمنعه / ممنوعيته / ومانع / ليمنع / / يمنعون ومانعه / وبالممانعة / مانعها / ممانعة / ممنوع / يمنعون / فالامتناع / المانعان / منعنا / يمنعن / ممتنعا / ممنوعة / فالمنع / المانع] / مانعيه / موانع / امتنعا / فيمنعون / فالمانع / بمنعهم / ومنعه / ممنوعه / وممنوع / ولوانع / بمنعة / مانعي / امتنعوا / يمنع / ولامتناع / بالامتناع / منعة / امتنعت / بالمنع / ومنع / امتناعهم / امتنع / امتناعا / ممنوعتان / والممنوع / فمنعاه / فامتنع / للمنح / كامتناعه / ومنعة / والموانع / ومانعيه / لمانع / لمنعة / المنع / ممتنعة / للامتناع / الممنوعات / بالممانعة / ومانعة / تمتنع / بامتنع / والممانعة / وبالمانع / فمانعه / ممنوعات / بمنعها / منعك / ومنعي / للممانعات / لمنع / ومنعنا / يمتنعا / ويمنعنا / منعنا / للمانعين / وممتنعة / بالمتنع / فممنوعة / والمانع / المنعة / الممتنعة / والممتنعات / فمانع / منعهم / بمانع / وللممانعة / منيع / وممتنع / للمانع / ومنعوه / فممتنع / ويمنعهم / الممنوع / كمنع / أمتنع / بامتناع / كالامتناع / بالمانع / لأمتنع / ومنعوا / مانعا / وللمنع / منعناكم / مانعان / ومنعته / والمنع / الممنوعة / المانعين / والممنوعة / تمناع / الممتنعات / امتناعه / وللمانع / فمنع / بالمنعة / لمنعه / [فامتناع / منع / منعناها / امتناعها / موانعها / تمنعه / ممتنعا	٢٥٨	م.ن.ع



الأحكام التكميلية		
المشتقات اللغوية	عدد المشتقات	الجذر
صحنا / والصحيحة / وتصح / صحيحا / صحيحها / فيصحح / صحيحهما / المصحح / فتصحح / بتصحيحه / للصحح / فلتصححه / مصححة / بمصحح / يصححه / فتصححه / بالصحة / أصحهما / صحبه / صح / لتصحح / ويصح / الأصح / بتصحيح / لصحح / كصحة / صححو / الصحيحان / وبصحة / أصح / لنصح / وصحح / كالصحيحة / نصح / مصحة / لصحته / ولصحة / صحيحهما / صحتهما / وصحاه / وتصحيح / تصحيحا / فلصحة / كالصحح / ومصححة / ومصححه / صححي / فصحة / صحا / وكصحة / بصحح / تصحيح / كالصحيحين / فصحيحهما / أصحها / وصح / ليصح / تصح / [وصحت / فصح / بصحتها] / وصحة / فصحت / وصحتها / لصحا / مصحح / بصحته / تصحيحه / فالصحح / أصححه / تصحيحها / صحتها / بصحتها / صحة / فالأصحح / لتصحيحه / والصحة / فالصحة / صحيحه / الصحة / أصح / فصحنا / يصحا / صحت / تصحح / صحح / فتصحح / كالصحة / يصحان / وصحت / والصحح / يصحح / بالأصح / صحتي / لصحت / بالصحح / يصححون / لصحتها / وصححه / صحح / الصحيحين / صحيحين / وصحت / لصح / مصححا / والتصحيح / للصحة / صحت / وصحها / وصحح / فصحته / صحيحة / ولصحته / فصحيح / فصحيحهما / والأصح / صحة / يصحوه / لصحة / وصحته / فيصح / صحيحتين / صحيحان / بصحتها / التصحيح / مصحه / الصحيحة / الصحح / وتصحيحها / ونصح / وأصح / صحت / لصحتها / بصحة / يصح / وتصحيحه / مصححه / المصححة / صححه	١٣٩	ص.ح.ج

الأحكام التكيفية		
المشتقات اللغوية	عدد المشتقات	الجذر
وكفساد / إفساد / ولفساد / ويفسد / والمفسدة / للفساد / فسد / لفسدتا / والإفساد / فاسد / لمفاسده / ففساد / ففساده / فسادا / ومفاسد / مفسدا / وإفساد / أفسدت / لفسدنا / فاسدهما / كالفساد / بفساد / كالمفسدة / وفسادها / فاسدها / وفاسدهما / تفسده / فسادها / بمفسد / مفاسده / الإفساد / مفسدة / بالإفساد / ففاسد / مفسدتها / بفسادها / فيفسده / إفساده / للمفاسد / المفسدات / مفسداته / المفسدين / مفسدات / الفاسدين / لإفساده / وفسادا / وفست / ولفساد / مفاسد / بالمفاسد / المفسدة / المفاسد / فساد / وللمفسدة / فالمفسدة / كفساد / للمفسد / ففسدت / كالفساد / بفساده / بفسادهما / مفسد / ففسد / فاسدا / [مفاسد / يفسد / وفاسد / فأفسد / المفسد / للمفسدة / لفساده / وفساد / فاسدة / يفسدان / مفسده / فالفاسد / فاسدان / فست / فالفساد / ومفسدة / والمفسد / وفاسدا / أفسد / والمفاسد / بالمفاسد / إفسادها / بالفاسدة / فاسدهما / بمفسدة / بإفساد / لإفساد / فيفسد / لمفسدته / لمفسدة / يفسدها / ومفسده / فتفسد / تفسدهما / وفاسدة / لمفسده / يفسدهما / بمفسده / فاسده / والفساد / والفاسدة / وفساد / تفسد / وفاسده / يفسدا / أفسده / الفساد / وفسد / والفاسد / لفساد / بإفسادهما / فساد / للإفساد / أفسدنا / الفاسد / أفسدها / ففاسدان / نفسدها / وأفسد / بالفساد / إفسادهما / بالمفسد / لفاسد / ومفسدته / وإفسادها / فيفسدهما / بالمفسدة / والمفسدة / يفسده / مفسدته / لفسد / وإفساده / الفاسدة / إفسادا	١٣٨	ف.س.د



الأحكام التكليزية		
المشتقات اللغوية	عدد المشتقات	الجذر
العزيمة / للعزيمة / عازما / للعزم / وعزمه / بعزيمة / العازم / وعزم / فالعزم / فعزيمتها / فعزيمة / العزائم / عزمة / عزمه / عزماء / فالعزيمة / فتعزم / عزموا / وعزيمته / فعزمة / بعزائم / وعزمة / تعزموا / وعزائم / والعزم / [عزمت / بالعزيمة / فعزمه / عزيمته / يعزم / وعزيمته] / فعزم / عزم / تعزم / عزيمة / بعزيمته / عزمي / / بالعزم / بعزمه / بعزيمة / عزمت / وعزيمة / أعزم / والعزيمة / ويعزم / العزم	٤٦	ع.ز.م
ورخصة / ترخصه / رخصة / فرخصة / للرخصة / يترخص / فيرخص / كالرخص / لرخصهما / وبالرخصة / لرخصه / ترخيص / / برخصه / كرخصه / يرخص / رخصت / ترخص / والرخص / مرخص / والمرخص / [فرخصة / ترخصا / مرخصا / والرخصة / الرخصتان / والترخيص / رخصتان / ورخصتان / بالرخصة / ومرخصا / / مرخصة / رخص] / ورخص / بالرخص / كرخصة / فرخصه / رخصه / الرخصة / فرخص / ورخصه / لرخص / برخصة / رخصتها / / للرخص / رخصته / فالمرخص / المرخص / لرخصة / كرخص / رخص / مرخصه / ورخصته / الترخيص / الرخص / برخص	٥٥	ر.خ.ص

ملاحظة: تم إعداد الجدول من البيانات التي توفرت بين أيدينا من التحليل النحوي والصرفي المحوسب
لمحتوى المدونة^(١).

لقد وقع اختيارنا على الجذور اللغوية للأحكام التكليزية والوضعية؛ لأننا على قناعة
تامة بأن الخطاطة الأصولية الحنفية مرتبطة بالوعاء اللغوي الذي يستمد مادته من الحقل
البياني للغة العربية. ويمكن أن تمارس سلسلة من التنقيبات في جذور ومشتقات مفردات

(١) وقد سعينا إلى عرض جميع المفردات في هذا الجدول؛ ليظهر حجم الذخيرة اللغوية التي يمكن توفيرها
من مادة المدونة.

الخطاب الأصولي والتي ستكشف لا محالة عن سبل فهم الخطاب الأصولي والكشف عن آلياته المعرفية؛ لأن سبر عملية التوظيف اللغوي في السياقات - حيث استخدمت اللغة بوصفها سلماً للفهم والبيان - سيسعنا إلى حد كبير بتحديد كيفية نهوض العقل الأصولي لعلماء أصول الفقه الحنفي، الذين نهلوا من دلالات الألفاظ وقواعد اللغة والخطاطة المعرفية الأصولية وتميزوا بها عن بقية المذاهب الفقهية.

١٣- المتلازمات اللفظية وفرص استثمارها:

لتوفير بيئة معرفية مناسبة من مادة المدونة المحوسبة لأصول الفقه الحنفي، قمنا بتوظيف خوارزمية لتحديد المتلازمات اللفظية Collocates، ووقع اختيارنا على اصطلاح "مكروه"، وهو من اصطلاحات الحكم الشرعي في دائرة أصول الفقه؛ وذلك لبيان ما يمكن أن توفره المدونة المحوسبة والخوارزميات الذكية التي تدعمها في تحديد طبيعة العلاقة التي تربط بين هذا الاصطلاح وبقية الاصطلاحات التي لازم حضورها معه، مع احتساب تكرار التلازم لكل حالة من الحالات، والتي تؤثر إلى العلاقة التي تربط فيما بينهم بحسب السياقات التي وردت فيها. ويمكن أن تستثمر نتائج التحليل المحوسب النحوي والصرفي مع هذه البيانات في كشف النقاب عن التراكيب اللغوية المستخدمة في نص المدونة وعلاقتها في تشكيل المفاهيم.

كما يمكن أن تُستثمر في دائرة الذكاء الاصطناعي التوليدي في استخلاص العلاقات، من خلال التدريب على مجموعة أكبر من نصوص أصول الفقه الحنفي، أو المدارس الفقهية الأخرى. وستمنح الباحثين في الوقت ذاته فرصة أكبر لفهم السياق والتنبؤ بالعلاقات استناداً إلى السياق العام الذي يميزه الخطاب الأصولي المطروح في المدونة المحوسبة.

استخراج المعلومات الكمية: يمكن استخدام المتلازمات اللفظية لاستخراج المعلومات الكمية من النصوص المكتوبة. على سبيل المثال، يمكن تحديد الأرقام والمقاييس والنسب والإحصائيات المذكورة في النصوص، واستخدامها لتحديد العلاقات بين المتغيرات المختلفة.

ويظهر في الشكل (٧) أنماط المتلازمات اللفظية لاصطلاح "مكروه" التي وردت في سياقات نص المدونة.



شكل (١) أكثر المتلازمات اللفظية حضوراً لاصطلاح "المكروه" في المدونة المحوسبة.



ملاحظة: أعد الباحث هذا الشكل من البيانات المحوسبة للمتلازمات اللفظية.

ويمكن استثمار هذه النتائج في الكشف عن أهم العبارات المتكررة، مع تحديد مستوى التكرار الذي يشير إلى أهميتها، مع فرصة تحديد العلاقة السببية بين هذه المتلازمات، والعلاقات الإشارية التي تفصح عنها الضمائر وبقية أقسام الكلام المتلازمة مع الاصطلاح.

ويمكن استخدام هذه المتلازمات اللفظية لتصنيف محتوى المدونة المحوسبة إلى فئات مختلفة، في ضوء أهم الاصطلاحات التي يتسم بها الخطاب المعرفي لدى أئمة مدرسة أصول الفقه الحنفي، حيث يمكن تحديد الأنماط اللغوية المشتركة في كل فئة واستخدامها لتصنيف النصوص الجديدة التي ستقوم بتوليدها آلة الذكاء الاصطناعي التوليدي.

١٤- تضمين الكلمات Word Embedding:

✓ تعد معالجة تضمين الكلمات لمحتوى المدونة المحوسبة لأصول الفقه الحنفي من مراحل توطئ المعالجات الذكية في فضاءه المعرفي؛ لغرض استنباط المزيد من عناصر الذخيرة

اللغوية والأصولية^(١). وإذا كانت معالجات اللغة الطبيعية السابقة قد تبنت المنطق اللساني التقليدي في إدارة معالجاتها المحوسبة الذكية، فإن توظيف معالجات تضمين الكلمات تعد خطوة أولية باتجاه نهج منطق الذكاء التوليدي في معالجات محتوى المدونة الأصولية المحوسبة، حيث سنعتمد إلى استخدام منطق احتساب متجهات معاني الكلمة في السياقات التي ترد فيها داخل حدود النص؛ لتحديد الفضاء الدلالي لمعانيها، وتحديد أكثر الكلمات قرباً منها على المستوى الدلالي، مع الوقوف على فرص استخدامها في توليد خطاب جديد، أو الإجابة عن سؤال، أو تلخيص جزء من النص (Bourahouat, et al. 2024: pp.313-325).

✓ من أجل هذا فقد سعينا إلى ممارسة سلسلة من المعالجات اللاحقة للبيانات التي حصلنا عليها من معالجة تضمين الكلمات، والتي شملت (Kusner, et al. 2015: pp.1-30):

✓ جيران الكلمة Word Neighbors، والتي يتم من خلالها تحديد أكثر الكلمات قرباً دلاليّاً وسياقياً من كل كلمة من كلمات المدونة. ويتم تحديد هذه الكلمات من على تقارب متجهات الكلمة Word Vectors في فضاء التضمين الخاص بها Embedding Space. حيث يلاحظ أن الكلمات التي تتقارب معانيها تستوطن في سياقات متشابهة.

✓ البعد الدلالي للكلمة Word or Semantic distance، والذي يعدّ مقياساً للتقارب بين الكلمات أو عدم التشابه فيما بينها. ويمكن تحديد ذلك من خلال درجة الانفصال أو التباعد بين متجهي الكلمتين^(٢).

(١) استخدمت خوارزميات مكتبة Fast Text المتوفرة في مكتبة Python البرمجية لاستخلاص تضمين الكلمات، وإنشاء مصفوفة الكلمات Word Vectors.

(٢) يعد كل من نهج احتساب Cosine Similarity والمسافة الإقليدية Euclidean Distance من المعالجات التي تستخدم لاحتساب مستوى التشابه بين متجهات الكلمات، بيد أن كل منهما يوجّه اهتمامه نحو مبدأ محدد في تحديد مستوى التشابه. فيتركز اهتمام نهج Cosine Similarity على تحديد قيمة الزاوية Angle بين متجهي الكلمات، فكلما اقتربت القيمة من ١ تأكد التشابه بين الكلمات. أما بالنسبة للمسافة الإقليدية فإن نهجها يوجّه اهتمامه نحو احتساب مسافة الخط المستقيم بين الكلمات، وكلما تناقصت المسافة ازداد التشابه. وقد عمدنا إلى استخدام نهج Cosine Similarity؛ لأنها تأخذ بعين الاعتبار مسألة تكرار حضور الكلمة بالنص بدلاً من العدد الكلي للكلمات، كما أنها توفر لنا مجاًلاً جيداً عند التحليل العنقودي للكلمات، وتحديد أهم المفاهيم المودعة في النصوص. (Zong et al. 2021: pp.127-129)

Zong, Chengqing, Xia, Rui & Jiajun Zhang, Text Data Mining, (Tsinghua: Tsinghua University Press, 2021), pp.127129-.



✓ مصفوفة المسافة بين الكلمات Distance Matrix، والتي يعبر عنها بالمصفوفة التي تحوي المسافات الفاصلة بين كل زوج من أزواج مفردات نص المدونة، والتي يعبر عنها بقيمة عددية تفصح عن مستوى التقارب فيما بينهم^(١).

✓ العناقيد الهرمية للكلمات المدونة Hierarchical Clusters، التي تصنف بموجبها عناقيد كلمات المدونة على أساس مستوى التقارب الدلالي فيما بينها، بناء على القيم التي تتمتع بها المتجهات الدلالية للكلمات^(٢).

✓ ويظهر في الشكل (٨) جزء محدود من مصفوفة المسافة بين كلمات المدونة المحوسبة، حيث تظهر قيمة متجه الفضاء الدلالي بين كل كلمتين من كلمات نصوص المدونة.

شكل (٨) جزء مقتطع من مصفوفة تضمين كلمات المدونة الحنفية المحوسبة

الكرخي	والعلو	واقبما	بالله	عين	الجمعين	جمعة	والمراد	الخبر	للمقبضين	باكد	الله	المراد	أوقى	بسيد	وتستفتح	الاعراق	الرشيد	الحفيد	المبدئ	أورخذ	شوال	الشيخ	شمس	السرخسي
٠,٣٢٥	٠,٣٢٠	٠,٢٩٦	٠,٣٨١	٠,٦٧٤	٠,٤١٢	٠,٣٤٠	٠,٢٨٥	٠,٤٧٦	٠,٣٢٨	٠,٣٤٤	٠,٤٣٧	٠,٣٤١	٠,٢٩٧	٠,٣٠٦	٠,٤٨٦	٠,٤٩٠	٠,٣٨٧	٠,٣٠٥	٠,٢٩٩	٠,٣٧٤	٠,٤٠٧	٠,٤٢٨	٠,٥٧٣	٠,٦١٢
٠,٢٦٨	٠,٣٢٧	٠,٣٠٥	٠,٣٨٥	٠,٦٦٢	٠,٣٧٩	٠,٣١٦	٠,٢٥٦	٠,٤١٢	٠,٣٦٣	٠,٢٩٨	٠,٤٣٦	٠,٢٧١	٠,٣٢٦	٠,٣١٠	٠,٤٧٨	٠,٥١٠	٠,٣٣٧	٠,٢٦١	٠,٢٧٥	٠,٣٢٩	٠,٣٧٠	٠,٤٢٧	٠,٤٧١	٠,٥٣٥
٠,١١٧	٠,٣٥٤	٠,٢٦٩	٠,٢٩١	٠,٥٤٥	٠,٣٤٧	٠,١٧٩	٠,١٣٦	٠,٢٨٦	٠,٢٤٢	٠,١٠٩	٠,٢٦٨	٠,٠٩٢	٠,١٢٢	٠,١٦٣	٠,٤٥٤	٠,٤١٧	٠,٢٣٨	٠,١١٤	٠,١٤٧	٠,١٩٢	٠,٢٦٣	٠,٣٦٥	٠,٤٧٦	
٠,٣٧١	٠,٤٤١	٠,٣٥٠	٠,٤٤٠	٠,٦٨٦	٠,٤٨٨	٠,٣٣٨	٠,٣٤٧	٠,٤٨٥	٠,٤٣٢	٠,٣٩٩	٠,٥٢٠	٠,٣٧٦	٠,٣٨٢	٠,٤١٠	٠,٥٥٩	٠,٤٨٣	٠,٤٥٨	٠,٣٧٤	٠,٣٤٣	٠,٤٣٧	٠,٤٧١	٠,٥٠٣	٠,٥٨٩	٠,٥٩٤
٠,١٩٢	٠,٤٨٩	٠,٣٣٦	٠,٢٨٥	٠,٤٤٨	٠,٢٥٨	٠,٢٩٠	٠,١٦٧	٠,٣٢٧	٠,٢٠٦	٠,٢٢٠	٠,٣٤٠	٠,١٥٩	٠,٣١٢	٠,٢٦٦	٠,٥٤٨	٠,٤٠٤	٠,٣٤١	٠,١٩٥	٠,١٥١	٠,٣٠١	٠,٢٤٧	٠,٢٦٥	٠,٤١٨	٠,٥٣٨
٠,٥١٩	٠,٧٢٥	٠,٦٤٨	٠,٣٦١	٠,٧٨٧	٠,٦٨٨	٠,٦٦٢	٠,٥٣٧	٠,٦٨٣	٠,٦٥٠	٠,٤١٨	٠,٣٠٩	٠,٤٨٤	٠,٢٤٣	٠,٥٣٠	٠,٧٤٧	٠,٦٧٢	٠,٤٨٨	٠,٥٧٤	٠,٢٠٦	٠,٥٩١	٠,٥٠٠	٠,٣٨٢	٠,٦٠١	٠,٧٦٧
٠,٥٤١	٠,٦٩٨	٠,٦٢٢	٠,٦٠٧	٠,٧١٧	٠,٦٢٢	٠,٥٧٨	٠,٥٣٠	٠,٦٢٢	٠,٥٨٠	٠,٥٣٥	٠,٦٤٠	٠,٥١٣	٠,٥٥٤	٠,٥٩٥	٠,٧٤٠	٠,٧٤٦	٠,٦١٧	٠,٥٥٨	٠,٥٢٥	٠,٥٧٨	٠,٦٢٨	٠,٦٣٥	٠,٥٠٨	٠,٧٦٤
٠,٤٠٣	٠,٥٥٨	٠,٤٨٧	٠,٥٢٧	٠,٧٣٢	٠,٤٨٠	٠,٣٥٨	٠,٤٢٨	٠,٥١٢	٠,٤١١	٠,٤٢٠	٠,٥٧١	٠,٤٠٧	٠,٤٧٠	٠,٤٥٢	٠,٦٤٤	٠,٥٤٤	٠,٥١٢	٠,٤٠٤	٠,٤٢٠	٠,٤٧٠	٠,٤٩٢	٠,٥١٢	٠,٥٩٥	٠,٧٠٢
٠,١٥٩	٠,٣٦٤	٠,٢٦٢	٠,٢٠٧	٠,٥٩٠	٠,٣٤٠	٠,٢٢٨	٠,١٦٢	٠,٣٥٤	٠,٢٣٤	٠,١٧٢	٠,٤٥٠	٠,١٤٨	٠,١٨٥	٠,١٢٠	٠,٥٠٦	٠,٤١٣	٠,١٦٩	٠,١٦٣	٠,٢٢٠	٠,٢٤٥	٠,١٩٢	٠,١٣٤	٠,٣٤٧	٠,٥١٦
٠,١٥٥	٠,٣٩١	٠,٢٧٧	٠,٢٠٧	٠,٥٤٥	٠,٢٨٨	٠,١٧٦	٠,١٢٩	٠,٢٧٤	٠,٢٠٧	٠,١٠٥	٠,٢٥٥	٠,٠٩٤	٠,٢٠٢	٠,١٥٧	٠,٤٨٤	٠,٣٨٣	٠,٢٢٠	٠,١١٥	٠,١٦٦	٠,١٩٣	٠,١٦٤	٠,١٧٦	٠,٣٦٩	٠,٤٦٩
٠,١٧٨	٠,٤٦٦	٠,٣٦٨	٠,٢٤٨	٠,٥٩٨	٠,٣٩٧	٠,٣٢١	٠,٢١٥	٠,٣٨٠	٠,٣٢٣	٠,١٤٢	٠,٢٥٨	٠,١٦٢	٠,٢٧١	٠,١٩٨	٠,٥٧٩	٠,٤٧٢	٠,٣١١	٠,١٩٩	٠,٢٥٧	٠,٢٦٥	٠,١٤٤	٠,١٤١	٠,٤٥٠	٠,٥٦٩
٠,١٠٥	٠,٣٩١	٠,٢٦٦	٠,٢٤٩	٠,٥٥٢	٠,٢٨٥	٠,١٦٦	٠,١٣٠	٠,٣٠٤	٠,٢١١	٠,١٢٥	٠,٣١٨	٠,١٨٩	٠,١٧٦	٠,١٨٨	٠,٤٧٨	٠,٣٣٩	٠,٢٨٦	٠,١٠١	٠,١٢٢	٠,٢٠٣	٠,٢٢٤	٠,٢٨٧	٠,٤١٥	٠,٤٦٣

(١) استُخدم نهج حساب Cosine Similarity لتحديد مستوى التقارب بين الكلمات.

(٢) استخدمت خوارزميات المكتبة البرمجية Scipy لبناء العناقيد الهرمية.

الترخيص	٠.٧١٢	٠.٦٢٥	٠.٥٥٢	٠.٥٠٠	٠.٤٧٢	٠.٤٨٧	٠.٤٠٠	٠.٥٥٥	٠.٦٢٨	٠.٦٢٢	٠.٤٨٢	٠.٥٠٠	٠.٤٥٥	٠.٦٣٠	٠.٤٦١	٠.٤٣٣	٠.٥٣٢	٠.٤٨٣	٠.٤٧٠	٠.٥٢٨	٠.٧٤٩	٠.٥٩٨	٠.٥٩٢	٠.٥٩٨	٠.٤٥٦
عصر	٠.٥٧٦	٠.٣٨٣	٠.١٣٠	٠.١٦٠	٠.١٤٥	٠.٢٠٨	٠.١٦٨	٠.٢٥٦	٠.٤٠٩	٠.٥٣٦	٠.١٧٢	٠.٢٣٠	٠.١٢٦	٠.٢١٣	٠.١٠٢	٠.٢٨٢	٠.٣٤٣	٠.١٦٨	٠.٢٧٩	٠.٣٢٨	٠.٥٦٥	٠.١٩٦	٠.٣٤١	٠.٤٤٦	٠.١٣٩
رمضان	٠.٧٩٥	٠.٥٤٨	٠.٥٦١	٠.٦٠٢	٠.٧٦٥	٠.٦٧٢	٠.٦٤٧	٠.٤٤٠	٠.٧٩٧	٠.٨٤٤	٠.٧٠١	٠.٧٢٩	٠.٦٠٧	٠.٦٢٣	٠.٦٣٥	٠.٦٩٤	٠.٧٠٩	٠.٤٦٨	٠.٦٠٠	٠.٧١٦	٠.٨٦٥	٠.٦٦٦	٠.٧٣٣	٠.٨٠٧	٠.٥٨٣
مفتحين	٠.٤٨٠	٠.٤٤٥	٠.٣١٢	٠.٣١٨	٠.٢٤٥	٠.٢٤٤	٠.١٧١	٠.٣١٣	٠.٤٦٥	٠.٤٤٦	٠.١٧٥	٠.٢٣٣	٠.١٩٦	٠.٢٩٢	٠.٢١٠	٠.٢٣٤	٠.٣١٢	٠.١٨٠	٠.٢٢٧	٠.٣٥٧	٠.٥٨٥	٠.٢٢٩	٠.١٩٧	٠.٢٢٧	٠.٢٠٧
مفتينه	٠.٤٦٥	٠.٤٩٦	٠.٣١٥	٠.٣٠٥	٠.٢٨٩	٠.٢١٤	٠.١٩٣	٠.٢٩٦	٠.٤٥١	٠.٣٦٦	٠.١٦٨	٠.١٩٤	٠.٢٣٦	٠.٣٠٧	٠.٢٣٢	٠.٤٦٢	٠.٤٠٥	٠.٢٢٥	٠.٤١١	٠.٦٣٢	٠.٢٥٣	٠.١٢٧	٠.١٩١	٠.٢٢٤	٠.٢٢٤
ليشي	٠.٤٦٠	٠.٤١٩	٠.٣٠١	٠.٢٥٤	٠.٢١١	٠.١٦٧	٠.١٤٤	٠.٢٦٥	٠.٣٥١	٠.٣٩٥	٠.١٩٩	٠.١٥٧	٠.١٨٦	٠.٢٣٣	٠.١٧٧	٠.٢٢٩	٠.٣٨٠	٠.١٨٧	٠.١٧٨	٠.٣٢٥	٠.٥٥٣	٠.٢٦٤	٠.١٠٤	٠.٢٥٤	٠.١٦٢
إيدرك	٠.٤٣٧	٠.٤١١	٠.٢٩٤	٠.٢٨٣	٠.٢٠٧	٠.١٥٦	٠.١١٩	٠.٢٥٣	٠.٤٣٨	٠.٣٤٢	٠.١٣٠	٠.٠٩٩	٠.١٦٢	٠.٢٧٦	٠.١٧٨	٠.١٦٧	٠.٣٢٢	٠.١٣٨	٠.١٩٠	٠.٣٠٥	٠.٥١٩	٠.٢١٤	٠.١٢٦	٠.٢٠١	٠.١٦١
مفوت	٠.٤٩٤	٠.٤٦٥	٠.٣٨٤	٠.٣١٨	٠.٢٩٤	٠.١٧١	٠.٢٠٣	٠.٣٣٨	٠.٤١١	٠.٥٣٨	٠.٢٨٤	٠.٤٤٨	٠.٢٠٢	٠.٤١٥	٠.٢٢٨	٠.٣٢٣	٠.٤٥٢	٠.٢٤٢	٠.٣٧٨	٠.٦٢٨	٠.٣٤٨	٠.٣٤٣	٠.٤٥٣	٠.٢١٣	٠.٢١٣
بالجمعة	٠.٥٦٥	٠.٤٩٩	٠.٢٨٤	٠.٢٨٠	٠.٢٢١	٠.٢٠٢	٠.١٥٨	٠.٢٥٠	٠.٤٠٤	٠.٤٧٦	٠.٢٢٩	٠.٢٢٦	٠.١٧٥	٠.٣٤٢	٠.١٩٦	٠.١٤٦	٠.٢٦٩	٠.١٦٩	٠.١٧١	٠.٥٠٠	٠.٦١٢	٠.٢٨٥	٠.٢٢١	٠.٣٥٤	٠.١٦٣
شاه	٠.٧٠٦	٠.٥٩٧	٠.٢٥٣	٠.٣٥٦	٠.٤٩٣	٠.٥٠٢	٠.٤٣٥	٠.٤٣٢	٠.٦٤٣	٠.٧٤١	٠.٣٤٦	٠.٥٣٠	٠.٤١٩	٠.٣٥٠	٠.٤٥٥	٠.٥١٢	٠.٥٤٢	٠.٤٣٠	٠.٥٢٥	٠.٥٢٨	٠.٧٤٦	٠.٣٤٥	٠.٥٤٩	٠.٦٢٠	٠.٤٢٥
والحادث	٠.٤٤٩	٠.٣٨٥	٠.٢٣٤	٠.٢١٥	٠.١٧٢	٠.١٩٠	٠.٠٨٢	٠.١٨٤	٠.٣٤٨	٠.٣٥٣	٠.١١٩	٠.١١٥	٠.٠٩٩	٠.٢٤٧	٠.١١٨	٠.١٤٤	٠.٢٦٣	٠.٣٠٥	٠.١٣٨	٠.٢٥٧	٠.٥٦٦	٠.١٨٥	٠.١٣١	٠.٤٤٤	٠.٠٨٣
كفر	٠.٤٧٦	٠.٣٨٢	٠.٢٦٢	٠.١٩٩	٠.٢٤١	٠.١٥٣	٠.٠٩٨	٠.٢٦٩	٠.٤٠٤	٠.٤٥٨	٠.١٧٧	٠.١٩١	٠.١١٩	٠.٣٣٦	٠.١٠٨	٠.٢٢٢	٠.٢٨٨	٠.١٥٠	٠.١٨٠	٠.٢٩٤	٠.٥٢٦	٠.٢٦٨	٠.٢٧٧	٠.٣٥٩	٠.١٠٢
يتضيق	٠.٦٣٩	٠.٦٩٥	٠.٦٥٥	٠.٥٦٦	٠.٥٤٢	٠.٣٨٤	٠.٤٣٢	٠.٥٣٢	٠.٥٠٤	٠.٥٨٣	٠.٥٢٢	٠.٥٠٥	٠.٤٣٢	٠.٥٨٨	٠.٤٨١	٠.٥٤٤	٠.٦٢٥	٠.٤٤٦	٠.٤٦١	٠.٦١٩	٠.٧٢٣	٠.٥٤٨	٠.٥٣٣	٠.٦٠٩	٠.٤٥١
صالح	٠.٤٦٠	٠.٣٩٥	٠.٣٠٨	٠.٢٦٤	٠.١٩٥	٠.٢٠٨	٠.١٣٢	٠.٢٨٠	٠.٤٣١	٠.٤٦٦	٠.١٧٥	٠.١٤٧	٠.١٤٢	٠.٣١٠	٠.١٥٨	٠.١٩١	٠.٢٩٦	٠.١٧٧	٠.١٦٧	٠.٣٣٢	٠.٥٦٦	٠.٢٥٥	٠.٢٥٨	٠.٣٤٢	٠.١٥٦
الخياط	٠.٤١٨	٠.٣٩٠	٠.٢٩٠	٠.٢٠٨	٠.٢٠٩	٠.١١٢	٠.٠٦٦	٠.٢٤٥	٠.٤٠٣	٠.٤٣٥	٠.٢٠٧	٠.٢٠٥	٠.٠٩٢	٠.٣٤٦	٠.١٣٦	٠.٢٣٥	٠.٣٥٦	٠.١٣٨	٠.١٦٠	٠.٢٩٢	٠.٥٥٠	٠.٢٨٤	٠.٢٥٥	٠.٣٧٩	٠.٠٧٩
ركن	٠.٤٤٢	٠.٣٥٢	٠.٢١٣	٠.١٣٨	٠.١٧٠	٠.١٣٢	٠.٠٨٣	٠.٢٢٠	٠.٣٥٤	٠.٤٨٧	٠.١٣٧	٠.١٧١	٠.٠٨٣	٠.٢٧١	٠.٠٧٤	٠.٢١٥	٠.٣٠٠	٠.١٤٠	٠.١٦٨	٠.٣٠٠	٠.٥٠٠	٠.٢١٦	٠.٢٥٠	٠.٣٧٠	٠.٠٨٢

ملاحظة: أعد الباحث الشكل من بيانات التحليل العنقودي لمحتوى النص المحوسب.

أما الشكل (٩)، فيعرض مقطعاً من المعمارية الهرمية لكلمات المدونة التي قامت الخوارزمية البرمجية بتقسيمها إلى عناقيد أساسية، وفرعية، بحسب التقارب الدلالي بين الكلمات المتحركة بكل عنقود من عناقيدها.

في ضوء التقارب في دلالاتها



ويمكن لمن يريد أن ينشئ نظاماً للذكاء الاصطناعي التوليدي الذي سيستمد موارده من محتوى المدونة المحوسبة لأصول الفقه الحنفي، أن يستثمر البيانات التي حاولنا الحصول عليها، ومعلومات أخرى ستفرضها طبيعة النظام الذكي؛ ليكون قادراً على ممارسة مجموعة من المهام (Choudhary, Bharat, 2020)، مثل: توليد المحتوى، تلخيص النصوص، تشكيل أنموذج للمفاهيم الأصولية في دائرة الفقه الحنفي، تقييم رصانة النصوص التي ستتولد عن النظام التوليدي الذكي من خلال استثمار البيانات وفق الخطاطة التي أدرجناها في الجدول (٤).

جدول (٤) الثمار التي يمكن جنيها من نتائج معالجات تضمين الكلمات والمعالجات اللاحقة على محتوى المدونة المحوسبة لأصول الفقه الحنفي

المحور	تضمين الكلمات	قياس الكلمات المتجاوزة	مصفوفة المسافات	العناوين الهرمية للكلمات
توليد المحتوى	توفير فرصة لتوسيع نطاق نصوص المدونة	توفير فرصة لتوليد محتوى جديد	تحديد معايير التماثل والتباين بين مفردات نصوص المدونة	تحديد أهم العناوين السياقات التي يمكن أن تتولد من نص المدونة
تلخيص النصوص	تعزيز عملية اختصار النص	تقييم مستوى الفروق بين النصوص المتولدة والنصوص الأصلية	قياس المسافة الدلالية بين الكلمات لانتخاب أفضل جملة معبرة	تبويب الملخصات المتقاربة في المعنى لغرض المراجعة وانتخاب الأفضل
تشكيل أنموذج للمفاهيم	العثور على أهم العلاقات الموضوعية في نصوص المدونة	تحديد أكثر النصوص تشابهاً في مضامينها لتحليل المواضيع	احساب مقدار التشابه بين النصوص التي تشكلت منها المدونة	الكشف عن المعيارية الهرمية للمواضيع المنبئة في محتوى المدونة



المحور	تضمين الكلمات	قياس الكلمات المتجاوزة	مصفوفة المسافات	العناقيد الهرمية للكلمات
استرجاع المعلومات	الارتقاء بكفاءة استرجاع المعلومات من خلال استثمار ظاهرة التشابه الدلالي لفضاء الكلمات	تحديد أكثر نصوص المدونة التي يمكن أن تستخدم في استرجاع المعلومات	حساب أفضل سمات التشابه بين نصوص المدونة بناء على المسافة الفاصلة بين مفردات نصوصها	قابلية نتائج البحث في نصوص المدونة في ضوء المعمارية الهرمية التي تتسم بها كل منها
اكتشاف الشذوذ	اكتشاف النصوص التي لا تنظم في بعض جوانبها مع المحتوى الكلي لنصوص المدونة	تحديد التضمين غير المتوافق لكلمات بعض النصوص مع المحتوى الكلي للمدونة	تحديد مستويات شذوذ محتوى بعض نصوص المدونة عن المحتوى الكلي	اكتشاف المجاميع والعناقيد التي يحيد محتواها عن المحتوى الكلي لمادة المدونة
تقييم النصوص المتولدة	تقييم جودة وتماسك النص المتولد من مادة المدونة المحوسبة	تحديد مستوى التماثل في النص المتولد مع نصوص معيارية منتخبة	تحديد مؤشرات تقارب المعاني بناء على قيم المسافات بين الكلمات	تقييم جودة وتماسك العناقيد التي ولدت أو المواضيع التي تعبر عنها

١٥- الخاتمة

تحليل نتائج المعالجات المحوسبة وبيان فرص استثمارها:

أثمرت عملية معالجة محتوى الموارد الأصولية وإنشاء المدونة المحوسبة لأصول الفقه الحنفي عن إنتاج كم كبير من البيانات التي شملت: حصر الوحدات اللغوية الأولية التي تألفت منها نصوص الموارد الأصولية، مع احتساب مستويات تكراراتها، وحصر كلمات التوقف، مع مخرجات التحليل النحوي والصرفي المحوسب، إضافة إلى تحليل بعض الجوانب الدلالية ذات الصلة بالخطاب الأصولي المطروح في مادة المدونة. ثم استكمال

شوط المعالجات المحوسبة مع معالجات تضمين الكلمات، التي استطعنا من خلالها رسم الفضاء الدلالي لكلمات المدونة مع المعمارية الهرمية لورود المفردات الأصولية في مادة المدونة المحوسبة.

بيد أن هذا الكم الكبير من البيانات سيكون بحاجة إلى مستويات عدة لضمان استثماره بطرق سليمة.

المستوى الأول: استثمار البيانات في الفضاء المعرفي البيني بين اللسانيات وأصول الفقه الحنفي.

المستوى الثاني: توظيف البيانات لتكون قاعدة معرفية تستمد منها نماذج اللغة الكبيرة، ونماذج الذكاء التوليدي في إنتاج محتوى جديد يتوافق مع الخطاطة الأصولية، ودلالات اصطلاحاتها، ومضامين السياقات التي يؤمل أن تكون أكثر التزاماً بمعايير أصول الفقه الإسلامي وثوابته.

وقد سعينا إلى بيان تفاصيل كل مستوى من هذين المستويين في جدول مستقل؛ لتتضح لنا فرص استثمار محتوى المدونة الأصولية التي عكفنا على إعدادها. انظر الجدولين (٥، ٦)^(١).

جدول (٥) فرص استثمار مادة المدونة في الفضاء البيني المعرفي بين اللسانيات وأصول الفقه

المعالجة	فرص الاستفادة	نماذج البيانات	التطبيقات المتاحة
التجزئة إلى الوحدات الأولية	تجزئة النص إلى المفردات	بيانات الجدول (١)	تحديد الذخيرة اللغوية، تبويب مفردات النص، دعم عملية البحث في النص

(١) نود التنبيه إلى أن حقل نماذج البيانات يوضح شريحة محدودة جداً من البيانات العملاقة التي تمتلكها المدونة المحوسبة والتي يمكن أن تشكل مورداً خصباً يمكن استثماره في أكثر من حقل معرفي.



المعالجة	فرص الاستفادة	نماذج البيانات	التطبيقات المتاحة
استخراج جذور الكلمات وجذوعها	اختزال الكلمات إلى الجذور التي اشتقت منها	بيانات الجدول (٣)	دعم عمليات التنقيب في النص، تحليل المفاهيم، توزيع الجذور اللغوية في النص
تحديد أقسام الكلام	تحديد الفئات النحوية بالنص	بيانات الشكل (٤)	وسم الكيانات النحوية، المراجعات النحوية، دعم نظم تحويل النص إلى كلام منطوق، التنوع اللغوي.
التحليل الصرفي	تحليل البنيات اللغوية للكلمات	بيانات الجدول (٢)	تقويم الأخطاء الصرفية، دعم أنظمة تعليم اللغة، الكشف عن الثراء اللغوي
تحديد هوية الكيانات	تحديد هوية الأعلام والأماكن	بيانات الشكل (٥)	الكشف عن الأعلام، والأماكن والبلدان، بيان علاقة الكيانات مع المفاهيم المطروحة، ترسيم جغرافية المفاهيم
تضمين الكلمات	وصف الفضاء الدلالي للكلمات	بيانات الشكل (٨)	تحليل التوجهات، التحليل الدلالي، تحديد ارتباطات معاني المفردات في مختلف أنماط السياق، استخلاص العلاقات بين الكلمات، كشف اللبس الدلالي

جدول (٦) استخدام بيانات المدونة كمورد لنماذج اللغة الكبيرة

والذكاء الاصطناعي التوليدي

أنموذج الذكاء التوليدي	فرص الاستفادة	البيانات	التطبيقات المتاحة
توليد النصوص	استخدام مصفوفات تضمين الكلمات كمورد للنماذج الذكية	بيانات الشكل (٨)	التأليف الإبداعي في مجال أصول الفقه الحنفي، التوليد الذاتي للنصوص في وسائل الإعلام، ومنصات التواصل الاجتماعي

أنموذج الذكاء التوليدي	فرص الاستفادة	البيانات	التطبيقات المتاحة
تحليل المحتوى	استخدام بيانات تضمين الكلمات في تحليل العلاقة بين الكلمات المتتالية	بيانات الشكل (٨)	تلقين مبادئ أصول الفقه، وكيفية بناء السياق الأصولي
نظم الحوار	استثمار مخرجات تضمين الكلمات في روبوتات الدردشة	بيانات الشكل (٨)	توفير خدمات ذات صلة بالخطاب الأصولي إلى فئات متعددة من طلبة العلم
تلخيص النصوص	استخدام تضمين الكلمات في تلخيص النصوص	بيانات الشكل (٨)	توليد الملخصات الأصولية
تصنيف مادة المحتوى	استخدام معمارية التصنيف المتولد عن تضمين الكلمات	بيانات الشكل (٩)	تحديد توجهات التيارات الأصولية، واتفاق الأقوال، تتبع الآراء التي لا تتوافق مع الخطاطة الأصولية الحنفية

ملاحظة: الجدول من إعداد الباحث.

يبد أن ما ذكر من فرص متعددة لاستثمار الكم الهائل من البيانات التي استطعنا الحصول عليها من تفكيك مفردات موارد مدونة أصول الفقه الحنفي، لا يعني أن الطريق سيكون سهلاً في بلوغ تطبيقات أصولية رصينة، ما لم نأخذ بعين الاعتبار العقبات المعرفية المحتملة التي ستشخص قبالة آلة الاستدلال والتوليد في نماذج الذكاء الاصطناعي التوليدي، والتي ستتأثر بجملة من العوامل البنيوية، وبطبيعة البيانات التي مورس التدريب والمران على مادتها، ومسائل أخرى ستجبرنا على التزام الحذر في التعامل معها. انظر الجدول (٧).



جدول (٧) أهم العقبات التي تشكّل تحدياً للذكاء الاصطناعي التوليدي عند تبني معالجاته الذكية لمادة النصوص العربية والإسلامية

المحور	العقبات والتحديات المحتملة
التحديات اللسانية	❖ الثراء والتعقيد البنيوي الذي تتصف به منظومة صرف اللغة العربية، والتي ستنعكس آثارها على الخوارزميات المستخدمة في تفكيك النص إلى وحداته الأولية، والكشف عن الجذور اللغوية، والتعامل مع اللواصق (السوابق واللاحق). ❖ حضور الحركات في مفردات الكلم العربية، نتيجة الدور الذي تمارسه المفردة في أقسام الكلام، سيشكل عائقاً أمام الذكاء التوليدي ما لم يتم تدريب أنموذج اللغة الكبير على نصوص متعددة، لكي تتوفر لديه فرصة تمييز المعاني والدلالات. ❖ وجود سمة الالتباس الدلالي Semantic Ambiguity والدور الذي يمارسه السياق في الحقل اللساني العربي في بيان الفضاء الدلالي للكلمات والاصطلاحات الأصولية التي تحمل معان عدة، بحسب السياق الذي وردت فيه، الأمر الذي يتطلب وجود معالجات عميقة لفك الالتباس الدلالي؛ لضمان إنتاج نصوص أصولية تتماسك معاني مفرداتها مع السياق الذي وردت فيه. ❖ اتجاه الكتابة من اليمين إلى اليسار يشكل عقبة أمام خوارزميات اللغة الطبيعية التي أنشئت لنصوص تتجه من اليسار إلى اليمين.

المحور	العقبات والتحديات المحتملة
التحديات المعرفية	❖ السمات التي تتسم بها الخطاطة المعرفية الأصولية، والنهج الذي يسود كل حقل من حقولها المعرفية (المدرسة المالكية، المدرسة الشافعية، والمدرسة الحنفية، والمدرسة الحنبلية، والمدرسة الظاهرية) تفرض وجود نماذج استنبطت من مادة كل مدرسة من هذه المدارس الأصولية دون معالجتها دفعة واحدة، الأمر الذي قد يتسبب في إنتاج محتوى لا يتوافق مع ثوابت الشريعة الإسلامية ومبادئها. ❖ صعوبة التحقق من هوية الكيانات وتثبيت أسمائها N.E.R بسبب بنية الأسماء باللغة العربية وعدم استخدام الحروف الكبيرة في بداياتها كما في اللغات الأخرى، الأمر الذي سيتطلب إنشاء خوارزميات أكثر ذكاءً في معالجة النصوص والكشف عن هويتها. ❖ ضرورة توفير خوارزميات قادرة على التعامل مع العمق الدلالي للاصطلاحات الأصولية، والقدرة على فهم الفروق الدلالية والمعاني السياقية، والتي ستعكس على دقة الخطاب التوليدي. ❖ الغنى السياقي لمادة الخطاب الأصولي والتي استثمرتها المدارس العلمية لأصول الفقه في إنتاج خطاطاتها المعرفية. الأمر الذي يتطلب المزيد من المرن والتدريب على نصوص متعددة، وبحق زمنية مختلفة.
التحديات الأخلاقية	❖ إن شحّة البيانات الأصولية المحوسبة، أو توجّه العاملين على نماذج اللغة الكبيرة نحو مذهب، أو تيار علمي دون آخر، بسبب وفرة النصوص، أو التحيز غير المعلن للقائمين على تدريب النماذج الذكية؛ سيولد لا محالة نمطاً من التحيز غير المعلن في الخطاب التوليدي. ❖ ضرورة التأكد من غياب سمة التحيز في الخطاب التوليدي نتيجة لممارسة التدريب على نصوص تعود إلى مدرسة فقهية دون أخرى. ❖ حصر مادة الخطاب التوليدي في نطاق ثوابت العلوم الإسلامية وعدم السماح لنماذج اللغة الكبيرة في التوسع باستنطاق مادة الخطاب التي تتعارض أخلاقياً مع ثوابت الخطاب الإسلامي.

ملاحظة: الجدول من إعداد الباحث.



هذه المسائل مجتمعة تؤكد ضرورة المباشرة بإنشاء معمارية أصولية لسانية تؤسس لحضور مستوى رصين من المعالجات اللاحقة لمحتوى المدونة المحوسبة لأصول الفقه الحنفي، بما يضمن استثمار البيانات من هذا العمل، وأعمال أخرى قائمة، وأخرى ستري النور قريباً لكي تتلاحم مخرجاتها مع نماذج اللغة الكبيرة التي ستشكلُ العمود الفقاري لآلة الاستدلال التوليدي مع ضمان توافقها مع ثوابت علم أصول الفقه، وبقيّة مرتكزات العلوم الإسلامية.

المراجع العربية

- الأرموي، صفى الدين محمد. (١٤١٦). *نهاية الوصول في دراية الأصول* (صالح بن سليمان اليوسف وسعد بن سالم السويح، محققان). المكتبة التجارية.
- بنعمر، محمد. (١٤١٧). *الدرس اللغوي عند علماء أصول الفقه*. أوراق نماء (٥٨) مركز نماء للبحوث والدراسات، <http://nama-center/article/details/30555>.
- الحنفي، محمد أمين (أمير بادشاه). (١٤١٧). *تيسير التحرير على كتاب التحرير في أصول الفقه الجامع بين اصطلاحى الحنفية والشافعية*. دار الفكر.
- الدكروري، أيمن. (١٤٣٩). *المدونات اللغوية ودورها في معالجة النصوص العربية*، دراسات (١٢). مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية.
- رشوان، محسن. (١٤٤١). *مدخل إلى حوسبة اللغة*. في محسن رشوان والمعتز بالله السعيد (محرران). مقدمة في حوسبة اللغة العربية. (٥٥). مركز الملك عبد بن عبد العزيز لخدمة اللغة العربية، ١٥-٥٩.
- السمرقندي، علاء الدين أبو بكر محمد بن أحمد. (١٤٠٤). *ميزان الأصول في نتائج العقول*. (محمد زكي عبد البر، محقق). مطابع الدوحة.

رومنة المراجع العربية

- Al-Dakrouri, A. (1439 H). *Al-mudawwanat al-lughawiyah wa dawruha fi muaalajat al-nusus al-Arabiyyah. Dirasat*, (12). Markaz al-Malik Abd Allah bin Abd al-Aziz li-Khidmat al-Lughah al-Arabiyyah.
- Al-Fanari, M. b. H. Sh. al-D. (1427 H). *Fusul al-badaa'i fi usul al-sharai* (M. H. Muhammad, Ed.). Dar al-Kutub al-Ilmiyyah.
- Al-Hanafi, M. A. (Amir Badshah). (1417 H). *Taysir al-tahrir ala Kitab al-Tahrir fi usul al-fiqh al-jami bayna istilah al-Hanafiyyah wa al-Shafiyyah*. Dar al-Fikr.



- Al-Usaymi, S. (1439 H). *Lisaniyyat al-matoun: Qadaya asassiyah fi al-ta'sil, al-tatbiq wa al-manhaj*. Markaz al-Malik Abd Allah bin Abd al-Aziz li-Khidmat al-Lughah al-Arabiyyah.
- Al-Armuwi, S. al-D. M. (1416 H). *Nihayat al-wusul fi dirayat al-usul* (S. S. Al-Yousef & S. S. Al-Suwaih, Eds.). Al-Maktabah Al-Tijariyyah.
- Al-Samarqandi, A. al-D. A. B. M. b. A. (1404 H). *Mizan al-usul fi nataij al-uqul* (M. Z. Abd al-Barr, Ed.). Matabi al-Doha.
- Ben Omar, M. (1417 H). *Al-dars al-lughawi inda ulama usul al-fiqh. Awrak Nama*, (58). Nama Center for Research and Studies. <http://nama-center/article/details/30555>
- Rashwan, M. (1441 H). *Madkhal ila hawsabat al-lughah*. In M. Rashwan & A. B. Al-Saeed (Eds.), *Muqaddimah fi hawsabat al-lughah al-Arabiyyah* (pp. 15–59). Markaz al-Malik Abd Allah bin Abd al-Aziz li-Khidmat al-Lughah al-Arabiyyah.
- Tuayman, A. b. A. (1433 H). *Usul al-Sarakhsi al-musamma: Tamhid al-fusul fi al-usul: Min bidayat al-qiyas ila nihayat al-kitab dirasatan wa tahqiqan* [Unpublished doctoral dissertation]. Jamiat Umm al-Qura.

المراجع الأجنبية

- Abumalloh, Rabab., Al-Sarhan, Hassan. & Abu-Ulbeh, Waheeb. (2016), Building Arabic Corpus Applied to Part-of-Speech Tagging. *Indian Journal of Science and Technology*, 9(46), 1-9.
- Alrabiah, A., Al-Salman, A., & Atwell, ES. (2013 July 22). The Design And Construction Of The 50 Million Words KSUCCA, In Proceedings of *WACL2 Second Workshop on Arabic Corpus Linguistics*. The WACL'2 Second Workshop on Arabic Corpus Linguistics, Lancaster University, UK. The University of Leeds.
- Bataineh, Ruba. & Al-Wazzan, Rawan Abdulrahman. (2014). Major Arabic Computational Linguistics Tools in Saudi Arabia: An Exploratory Study. *International Journal of Computer and Information Technology*, 3 (6), 1514-1520.
- Belinkov, Yonatan., Magidow, Alexander, Romanov, Maxim, Shmidman, Avi. & Koppel, Moshe.

- (2016 December 11-17). Shamela: A Large-Scale Historical Arabic Corpus. In Erhard Hinrichs, Marie Hinrichs, Thorsten Trippel (Chairs), Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka, Japan.
- Bialy, Asmaa A., Gaheen, Marwa A., ElEraky, R. M., ElGamal, A. F. & Ewees, Ahmed A.(2020), Single Arabic Document Summarization Using Natural Language Processing Technique. In Mohamed Abd Elaziz, Mohammed A. A. Al-qaness, Ahmed A. Ewees & Abdelghani Dahou (Eds), *Recent Advances in NLP: The Case of Arabic Language* (pp. 17-37). Springer.
- Bourahouat, Ghizlane., Abourezq, Manar & Daoudi, Najima. (2024). *Word Embedding as a Semantic Feature Extraction Technique in Arabic Natural Language Processing: An Overview. The International Arab Journal of Information Technology*, 21(2), 313-325.
- Choudhary, Bharat. (2020 August 4). *Introduction to Word Embeddings*. Medium. <https://pub.towardsai.net/introduction-to-word-embedding-5ba5cf97d296>
- Darwish, Kareem., Habash., Nizar., Abbas, Mourad., Al-Khalifa, Hend., Al-Natsheh, Huseein., El-Beltagy, Samha., Bouamor, Houda., Bouzoubaa, Karim., Cavalli-Sforza, Violetta., El-Hajj, Wassim., Jarrar, Mustafa. & Mubarak, Hamdy. (2021). A Panoramic Survey of Natural Language Processing in the Arab World. *Communications of the ACM*, 64(4), 72-81.
- Eddakrouri, Ayman. (2023). Arabic Corpus of Library and Information Science: Design and Construction, *Egyptian Journal of Language Engineering*, 10(1), 1-9.
- Elhassan, Nasrin., Varone, Giuseppe., Ahmed, Rami., Gogate, Mandar., Dashtipour, Kia., Almoamari, Hani., El-Effendi, Mohammed., Al-Tamimi, Bassam Naji., Albalwy, Faisal & Hussain, Amir. (2023), Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning, *Computers*, 12(6), 126.
- Ismail, Omaima., Yagi, Sane. & Hammo, Bassam. (2014). Corpus Linguistic Tools for Historical Semantics in Arabic. *International Journal of Arabic-English Studies (IJAES)*, 15, 135-152.



- Jarrar, Mustafa., Khalilia, Muhammed. & Ghanem, Sana. (2022 June 20-25). *Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT*, The 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille.
- Kusner, Matt., Sun, Yu., Koklin, Nicholas. & Weinberger, Kilian. (2015 July 7-9). *From Word Embeddings to Document Distances*. The 32nd International Conference on Machine Learning, Lille, France.
- Liang, Hong., Sun, Xiao., Sun, Yunlei. & Gao, Yuan. (2017) Text Feature Extraction Based on Deep Learning: A Review. *EURASIP Journal on Wireless Communications and Networking*, 2017.
- Mohamed, Najat., Rassem, Malak., Han, Lifeng. & Nenadic, Goran. (2023 September 4-6). *AlphaMWE-Arabic: Arabic Edition of Multilingual Parallel Corpora with Multiword Expression Annotations*. The 14th International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria.
- Sharaf, Abdul-Baqee., Atwell, Eric., Dukes, Kais., Sawalha, Majdi., Al-Saif, Amal., Sharoff, Serge., Markert, Katja., Al-Sulaiti, Latifa., Shawar, Bayan Abu., Abbas, Nora. & Roberts, Andy. (2010 October 16), Arabic and Quranic Computational Linguistics Projects at The University Of Leeds, In *Proceedings of the workshop of Increasing Arabic Contents on the Web*, Workshop of Increasing Arabic Contents on the Web, Arab League Educational, Cultural and Scientific Organization (ALECSO), Damascus, Syria.
- So, Aaron. (2024). Introduction to Word Embeddings: Problems and Theory. *Learn Data Science*. <https://www.learndatasci.com/tutorials/intro-to-word-embeddings-problems-theory/>
- Zong, Chengqing., Xia, Rui. & Zhang, Jiajun. (2021). *Text Data Mining*, Springer.

نبذة تعريفية عن الباحث

Author Biodata Hasan M. Al-Rizzo is a Senior Research Scientist in Computational Simulation, a Senior Advisor for Strategic Planning in Central Government. His research interests include Computational Intelligence, Computational Linguistics & Generative A.I Models and applications.	بيانات الباحث: حسن مظفر الرزق ، رئيس باحثين علميين، (محاكاة نظم طبيعية بالحاسب) مستشار حكومي سابق لشؤون التخطيط الاستراتيجي بالعراق. حاصل على درجة الدكتوراه في المحاكاة بالحاسوب الآلي من جامعة الموصل عام ١٩٨٧. تدور اهتماماته البحثية حول معالجات الحوسبة الذكية وتطبيقاتها، واللسانيات المحوسبة، ونماذج الذكاء التوليدي.
معرف أوركيد (ORCID) : 0000-0002-8089-2066	
Email: halrizzo@gmail.com	



الملاحق

ملحق (أ) قائمة ترجمة الاصطلاحات العلمية

الاصطلاح	الترجمة العربية
Arabic Phonetics Database	قاعدة بيانات الصوتيات العربية
Artificial Intelligence	الذكاء الاصطناعي
Artificial Neural Network (A.N.N)	شبكات اصطناعية عصبونية
Computational Corpus	مدوّنة محوسبة
Corpus	مدوّنة
Corpus Directed Studies	دراسات تستمد توجهاتها من المدوّنات
Corpus-Based Studies	الدراسات المبنية على المدوّنات
Diacritizer	آلة التشكيل
Distance Matrix	مصفوفة المسافة
Embedding Space	فضاء التضمين
Generative Artificial Intelligence	الذكاء الاصطناعي التوليدي
Hierarchical Clusters	العناقيد الهرمية
Intelligent Algorithm	خوارزمية ذكية
Large Language Model (L.L.M)	نماذج اللغة الكبيرة
Morphological Analysis	التحليل الصرفي
Named Entity Recognition	وسم هوية الكيانات
Natural Language Generation	توليد اللغة الطبيعية
Natural Language Processing (N.L.P)	معالجات اللغة الطبيعية

الاصطلاح	الترجمة العربية
Natural Language Understanding	فهم اللغة الطبيعية
Paradigm	خطاطة / منهجية
Raw Corpus	مدونة خام
Root	جذر الكلمة
Semantic Analysis	التحليل الدلالي
Semantic Distance	البعد الدلالي
Semantic Space	فضاء دلالي
Sentiment Analysis	تحليل الآراء والتوجهات
Stem	جذع الكلمة
Stop Words	كلمات التوقف
Syntactic Analysis	التحليل النحوي
Text Mining	التنقيب في البيانات.
Word Embedding	تضمين الكلمات
Word Neighbors	جيران الكلمة
Word Vector	متجه الكلمة

تحسين استرجاع اسم التصغير في التحليل الصرفي

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-002>

هاله منيف الحربي*^{ORCID} عبد الله يحيى الفيافي^{ID}

مجمع الملك سلمان العالمي للغة العربية، الرياض، المملكة العربية السعودية

استقبل في	رُوجع في	قُبِل في	نُشر في
١٤٤٦/٠٤/٠٢	١٤٤٦/٠٤/١٧	١٤٤٦/٠٩/٢٧	١٤٤٦/١٠/٢٨

المخلص

تناقش هذه الدراسة آلية تحسين استرجاع الكلمات المصغرة من خلال صياغة قواعد حاسوبية تصنف صيغ التصغير، وتمييزها عن غيرها من الكلمات. اعتمدت الدراسة على سياقات كلمات العينة التي جمعناها ووصفنا قواعدها اللغوية، لتسهيل عملية التعامل معها حاسوبياً. وللتأكد من صحة هذه القواعد، تحققنا من ذلك بالاعتماد على مدونة لغوية خام، يزيد عدد كلماتها عن خمسة وثلاثين مليون كلمة، واستدعاء الكلمات المصغرة، ثم التأكد من مطابقة سياق الكلمات المصغرة للقواعد اللغوية المقترحة. وقد أظهرت نتائج التقييم مطابقة هذه السياقات للقواعد الموصفة بنسبة ٩٩٪.

الكلمات المفتاحية

اسم التصغير، التعبيرات النمطية، التحليل الصرفي، المدونة اللغوية، الصرف.

Optimizing the Retrieval of Diminutive Forms in Morphological Analysis

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-002>

Halah Al-Harbi^{*ID} Abdullah Alfaihi^{ID}

King Salman Global Academy for Arabic Language

Received on	Revised on	Accepted on	Published on
05/10/2024	20/10/2024	27/03/2025	30/04/2025

ABS TRACT

This study aims to improve the retrieval of diminutive words by developing computational rules that classify and distinguish diminutive forms from other words. It is based on analyzing the contexts of a collected sample of words to define their linguistic rules, thereby facilitating their computational processing. To ensure the accuracy of these rules, we validated them using a raw corpus of over thirty-five million words and retrieved diminutive words. We then verified whether the contexts of the retrieved diminutive words aligned with the linguistic rules. The evaluation results showed a 99% accuracy in matching these contexts with the proposed rules.

KEYWORDS

Diminutive Noun, Regular Expression, Morphological Analysis, Corpus, Morphology.

*Corresponding author: Halah Al-Harbi (Email: hmuneef@ksaa.gov.sa).

١- المقدمة

تعتمد التطبيقات الحاسوبية في معالجة اللغات الطبيعية على التحليل الصرفي للكلمات، وهو العملية التي تُحدّد من خلالها جميع التحليلات الصرفية الممكنة لكلمة معينة، وعادة ما تُعرف هذه التحليلات كتابياً، ويتضمن كل تحليل اختياراً واحداً من أجزاء الكلام الأساسية مثل الاسم أو الفعل (Habash, 2010).

ولأن اللغة العربية لغة اشتقاقية، فإن التحليل الصرفي في بعض أبوابها الصرفية يعد صعباً إلى حدّ ما. ومن أبرز هذه الأبواب باب التصغير؛ فنظراً لاشتقاقية اللغة العربية، فإن التغيرات التي تحدث عندما نريد تصغير كلمة ما تكون في دواخل الكلمة interfixes دون استخدام السوابق أو اللواحق كما في لغات أخرى؛ فالتغيير يكون في نطاق الحرف الأول، حيث يُنقل الحرف الأول من حركته الأصلية إلى الضم، وتُزاد الياء بعد حرفين أو قبل آخر حرف (Zivingi & Sulaiman, 2022). كما أنه قد لا يُكتفى بإضافة حرف التصغير (الياء) إلى الاسم المصغر، بل إنه قد يحدث تغيير في الأحرف الأساسية في الكلمة، فمثلاً كلمة (خاتم) تُصَغَّر فتصبح (خَوَيْتِم). إضافة إلى ذلك، تتطلب بعض الأسماء المصغرة إضافة علامة تأنيث، حتى لو كانت الكلمة الأساسية مذكرة. فمثلاً كلمة (بحر) تُصَغَّر فتصبح (بُحَيْرَة). من التحديات أيضاً أنه قد تُحذف بعض الأحرف من الأسماء الخماسية، فمثلاً كلمة (عنكبوت) تُصَغَّر فتصبح (عُنَيْكِب) بحذف بعض الأحرف (Nasser, 2020).

من هنا جاءت هذه الدراسة لتتّرح منهجية لتحسين استرجاع الكلمات في باب التصغير. وقد اخترنا باب التصغير على وجه الخصوص لقلة طرقه في الدراسات اللغوية الحاسوبية، ولأن استرجاعه أيضاً يعد صعباً أو محدوداً؛ لشبهه في رسمه الإملائي بصيغ أخرى من صيغ العربية، أبرزها الصفة المشبهة.

وتهدف هذه الدراسة إلى تحسين استرجاع الأسماء المصغرة من خلال توصيف قواعد لغوية لهذه الأسماء وتميزها حاسوبياً؛ ثم تقييم هذه القواعد نظرياً وتحكيمها، واختبارها تطبيقياً من خلال استرجاع أسماء التصغير في مدونة لغوية ومقارنة سياقاتها بالقواعد المقترحة.

٢- الإطار النظري

٢-١- التصغير

التصغير في اللغة من صغر، والصَّغَر: ضد الكِبَر (ابن منظور، ١٩٩٤). وهو في الاصطلاح: "تغيير صيغة الاسم لأجل تغيير المعنى، تحقيراً أو تقليلاً أو تقريباً أو تكريماً أو تعظيماً" (الجرجاني، ١٩٨٣، ص ٦٠).

وللتصغير صيغ ثلاث ذكرها سيبويه في قوله: "التصغير على ثلاثة أمثلة: فُعِيل، فُعِيلِل، فُعِيلِيل" (سيبويه، ١٩٨٨)، كما في: نُحْيِم من نجم، ولُوْعِب من لاعب، ومُسَيِّكِين من مسكين. وليس المقصود أن تتطابق الأسماء المصغرة مع الميزان الصرفي حرفاً بحرف، وإنما المقصود بها أنها القالب الذي يخرج على أساسه الاسم المصغر، بحيث يتساوى مع الصيغة في عدد الحروف وفي نوع الحركة والسكون (الراجحي، د ت).

ويجمع التصغير وسيلتين من وسائل التعبير اللغوي، هما: اللصق والصيغة، فهو يُوجب إضافة ياء ساكنة في وسط الكلمة، كما يُغيّر في أصواتها بضمّ الأول وفتح الثاني، وهو بهذا يرتبط بالناحية الصوتية أكثر من الناحية الشكلية، وهما مُهمّان في تكوين الاسم المُصغّر (الروابدة، ٢٠١٠).

وللتصغير شروط في الكلمة المراد تصغيرها، هي (الأزهري، ٢٠٠١):

- أ. أن تكون اسماً، فلا يصغر الفعل ولا الحرف. وشذ "ما أحيسنه" عند البصريين.
- ب. أن يكون الاسم متمكناً، فلا تصغر المضمرات، ولا أسماء الإشارة، ولا الأسماء الموصولة، ولا مَنْ وكيف ونحوهما كآين ومتى.
- ج. ألا يكون الاسم على صيغة تصغير، فلا يصغر نحو: "كَمِيت"؛ لأنه على صيغة التصغير. ولا "مبيطر"؛ لأنه على صيغة تشبه صيغة التصغير.
- د. أن يكون الاسم قابلاً للتصغير، فلا تصغر الأسماء المعظمة كأسماء الله تعالى وأنبيائه وملائكته ونحوها، ولا جمع الكثرة، وكلّ، وبعض، ولا أسماء الشهور، والأسبوع عند سيبويه، وغير، وسوى، والبارحة، والغد، والأسماء العاملة.



٢-٢- طرق معالجة الكلمات العربية صرفياً

يعد التحليل الصرفي الآلي من أهم المجالات في معالجة اللغات الطبيعية، حيث يساعد على تحليل الكلمات إلى جذورها وبنياتها المختلفة بهدف فهمها وتحليلها حاسوبياً. وقد طُوِّرت العديد من الطرق لمعالجة الكلمات العربية وتحليلها صرفياً، ومن بينها: طريقة قوائم الكلمات المخزنة، والطريقة اللغوية، والطريقة الرياضية، وطريقة الأوزان. وتختلف هذه الطرق فيما بينها من حيث آلية العمل والدقة والسرعة في تحليل الكلمات. وفيما يأتي تفصيل ذلك (الزامل، ٢٠٠٧):

- أ. طريقة قوائم الكلمات المخزنة، وتعتمد هذه الطريقة على تخزين كلمات اللغة العربية جميعها في قوائم ضمن مكوناتها الصرفية على شكل جداول كبيرة تتضمن قواعد بيانات ضخمة، ويتم البحث فيها عند تحليل الكلمات.
- ب. الطريقة اللغوية، وتعتمد هذه الطريقة على توصيف قواعد اللغة العربية الصرفية بشكل خوارزميات حاسوبية، بحيث يحاكي البرنامج عمل اللغوي عند تحليل الكلمات وإرجاعها.
- ج. الطريقة الرياضية، وتقوم هذه الطريقة على تحليل الكلمات وفقاً لنظام التجربة والخطأ، حيث تُجرَّب الجذور المحتملة للمفردة بناءً على مقارنتها بجذور مخزنة في قاعدة بيانات. تُجرَّب ثلاثة أحرف متتالية حتى يتوصل إلى الجذر الصحيح.
- د. طريقة الأوزان، وتعتمد هذه الطريقة على القواعد الصرفية الآلية من خلال مقارنة قائمة كبيرة من الكلمات مع ما يقابلها من مصادرها وأوزانها، ما يساعد على معرفة مصدر الكلمة بشكل أكثر دقة. فإذا طبقت الكلمة على أكثر من قاعدة، تُرجَّح القاعدة الأكثر تكراراً.

وبعد دراسة الطرق المختلفة للتحليل الصرفي الآلي، اخترنا الطريقة اللغوية؛ لأنها تعتمد على توصيف قواعد اللغة العربية الصرفية في شكل خوارزميات حاسوبية، ولأنها أكثر منهجية وأقرب إلى طريقة التحليل البشري. بالإضافة إلى ذلك، فإن هذه الطريقة تعالج بعض عيوب الطرق الأخرى، مثل الحاجة إلى قواعد بيانات ضخمة كما في طريقة القوائم

المخزنة، أو الاعتماد على التجربة والخطأ كما في الطريقة الرياضية. وبهذا توفر الطريقة اللغوية دقة أكبر في معالجة الكلمات، إلى جانب استخدام التعبيرات النمطية التي تُعد خطوة أولى في استخراج الأنماط اللغوية وتحليلها، وغالباً ما تستخدم عند البحث عن أمثلة معينة في المدونات الضخمة (الزامل، ٢٠٠٧).

٣- الدراسات السابقة

هذه الدراسة تعد امتداداً لعدد من الدراسات التي هدفت إلى تحسين استخراج الكلمات في أبواب صرفية مختلفة من خلال وضع قواعد لغوية وتوصيفها حاسوبياً، ومن أبرز هذه الدراسات: دراسة فهمي (٢٠٢١) التي هدفت إلى تطوير خوارزمية حاسوبية لمعالجة الصيغ الصرفية لاسم الفاعل والصفة المشبهة. واستندت إلى القواعد النظرية الخاصة باشتقاق هاتين الصيغتين من المصادر الصرفية القديمة والحديثة، ثم رُمزت القواعد لتيسير برمجتها حاسوبياً بهدف إنشاء قاعدة بيانات قابلة للاستخدام في برمجيات تحليل وتوليد اسم الفاعل والصفة المشبهة. وقد كشفت نتائج الدراسة عن كفاءة الخوارزمية المطورة في التعامل مع عينات الاختبار، حيث أظهرت دقة عالية في تقديم النتائج الصحيحة لمختلف أنواع الأفعال.

أما الدراسة الثانية فهي دراسة القبشي (٢٠١٩) التي هدفت إلى بناء خوارزمية حاسوبية تُسهّم في تقليل من اللبس الصرفي في جموع القلة. واستعرضت الباحثة فيها الموضوعات المُلبسة في تحليل جموع القلة، وهي مشكلة وسم الكلمة بجنسها الصحيح، ومشكلة الوسم بالعدد، وكذلك مشكلة استخراج أصول كلمات الوزن السماعي مثل: فَعَلَة. وقدّمت الدراسة الحلول لهذه المشكلات من خلال وضع مجموعة من القواعد اللغوية التي أسهمت في تقليل مواضع اللبس الصرفي الحاسوبي محلّ الدراسة. أما دراسة الدوسري (٢٠١٠) التي حصر فيها الباحث القواعد الخاصة بالمنسوب بياء النسب ووصفها، ثم حوّلها إلى قواعد حاسوبية يمكن للحاسوب فهمها، وتوصلت الدراسة إلى بناء أكثر من مئة قاعدة حاسوبية للنسب بالياء، وإلى معالجة كثير من الكلمات الشاذة.



أما الدراسات المتعلقة بمعالجة اسم التصغير على وجه الخصوص فدراستان، الأولى دراسة ناصر (Nasser, 2020)، فقد قدمت نموذجاً صرفياً حاسوبياً لتوليد الأسماء المصغرة في اللغة العربية باستخدام نظرية الصرف الإيقاعي Prosodic Morphology. وقد اعتمدت على أداة "نوج (NooJ)" لتنفيذ النموذج، الذي يتكون من قاموس وقواعد تحويلية قادرة على توليد الأسماء المصغرة من جذور مختلفة (ثنائية، ثلاثية، رباعية، وخماسية الحروف الساكنة). واختبر النموذج، وأظهر قدرة على التعامل مع تعقيدات الصرف في اللغة العربية. والفرق الرئيس بين هذه الدراسة ودراستنا هو أن الدراسة تركز على توليد الأسماء المصغرة، بينما تركز دراستنا على تحسين استرجاعها من خلال تحليل السياقات اللغوية. أما الدراسة الثانية فدراسة ثلجي (Thalji, 2018)، التي هدفت إلى تطوير خوارزمية حاسوبية لتصغير الأسماء في اللغة العربية باستخدام قواعد التصغير. فقد جمعت بيانات اختبارية صغيرة تضمنت أسماء بأحرف مختلفة (ثنائية، ثلاثية، رباعية، وأكثر)، واختبرت الخوارزمية على هذه البيانات. وقد أظهرت النتائج دقة عالية في تصغير الأسماء الثنائية (١٠٠٪)، بينما واجهت صعوبات مع الأسماء الأطول بسبب وجود أحرف متحركة تحتاج إلى معالجة خاصة، وبلغت الدقة الإجمالية ٩٤,٧٩٪. ولعل أبرز اختلاف بين هذه الدراسة ودراستنا هو أن هذه الدراسة تسعى إلى تصغير الأسماء، بينما تسعى دراستنا إلى تحسين استرجاع الأسماء المصغرة.

٤- منهجية الدراسة

انتهجت هذه الدراسة المنهج الوصفي؛ إذ تعتمد على توصيف قواعد اسم التصغير وتميزها، والمنهج الإحصائي في اختبار دقة هذه القواعد، وفيما يأتي تفصيل ذلك.

أُخذت عينة الكلمات التي اعتمدنا عليها في توصيف القواعد اللغوية من كتب الصرف العربي. فقد حُصرت الألفاظ المصغرة الواردة في شرح باب التصغير في بعض كتب الصرف؛ لتكون عينة الدراسة، على أن تكون الكتب معاصرة وتحتوي على عدد كافٍ من الأمثلة، وهذه الكتب هي:

- الصرف الكافي. تأليف: أيمن أمين عبد الغني.
- الصرف العربي أحكام ومعاني. تأليف: محمد فاضل السامرائي.
- الواضح في الصرف. تأليف: حسان بن عبد الله الغنيمان.

بعد حصر الألفاظ استُخرجت سياقاتها من مدونتين لغويتين، وانتهت العينة إلى ١٦٢ سياقاً (انظر الجدول ١). وتجدر الإشارة إلى أن بعض الكلمات لم تظهر في هاتين المدونتين، أو ظهرت في سياقات مصطنعة؛ مثل أن تُذكر الكلمة المصغرة مثلاً على صيغة تصغير، وبذلك لم تُضَف إلى العينة.

جدول (١) جُزء من عينة الدراسة

الكلمة	السياقات
رجل	بعد أن كانوا من طائفة الأحرار: رجيل يدعى بمعاذ بن يوسف.
	فقال لبعض رجاله اذهبوا وأحضروا ذاك الرجل.
نهر	وتجري أحداثها في المدينة الحدودية "تشاتونج" التي تقع على ضفة نهر عند ملتقى حدود ثلاث مقاطعات صينية.
	أسواق بطراز معماري حديث وحدائق غناء... وخير نهر صغير مجاور يترك أذنيه.
جبل	خبت من رمل في وسطه جبيل أسود شديد السواد يقال له طفيل.
	ولا يوجد في واحة الأحساء كلها قارة أو جبيل فيه هذه الميرة سوى تل صغير يُسمى حتى الآن برأس القارة.
فوق	التوسط بمقدار أربع حركات، وفوق التوسط بمقدار خمس حركات.
	فتقدم الكلام على أنها تخرج من المخرج التاسع من الفم مما يلي طرف اللسان وفوق الثنايا السفلى.
دار	انتقلت تلك الخزانة إليها وزاد عليها بناية خاصة سميت (دويرة الكتب).
	وبنى دويرة للصوفية وهي أول ما بني في الإسلام.

وحيث أثبتت المدونات اللغوية corpora منذ ظهورها قدرتها على إعادة صياغة منهجيات الأبحاث العلمية في كثير من الجوانب اللغوية؛ وذلك لما توفره من إمكانيات كبيرة جعلتها أساساً لمادة البحث اللغوي الحديث (الفيفي، ٢٠١٥، ص ٩٧)، فقد استخدمنا ثلاث مدونات لغوية في استخراج سياقات عينة الكلمات المستخرجة من الكتب واختبار القواعد اللغوية،



وهي: المدونة العربية، ومدونة arTenTen، ومدونة الجاسم. وقد اختيرت هذه المدونات نظراً إلى حجمها الكبير الذي يُمكن من رصد معظم صيغ التصغير، بالإضافة إلى سهولة البحث في أدوات كلٍّ من المدونتين الأولى والثانية. إذ اعتمدت الدراسة على هاتين المدونتين في استخراج سياقات كلمات العينة، فيما اختيرت الثالثة لتقييم القواعد اللغوية.

وتحتوي المدونة العربية مليار كلمة مما دُوّن بالعربية بدءاً من العصر الجاهلي وحتى العصر الحديث ومن مختلف المناطق والبلدان، مع الأخذ في الحسبان طبيعة النشاط الفكري لكل فترة وحجمه وتنوع أوعية النشر فيها، والسائد من المجالات العلمية والفكرية المختلفة. وقد أُتبعت عدة معايير عند جمع المدونة وهي البعد الزمني والجغرافي والتصنيف الموضوعي وأوعية النشر والمجال المعرفي (منصة فلك للمدونات اللغوية، د.ت).

فيما جُمعت نصوص مدونة arTenTen العربية من الإنترنت وتحتوي على نصوص عربية متنوعة تشمل الفصحى واللهجات العامية. وقد اعتمد الباحثان على الفصحى منها. ويمكن الوصول إليها من خلال منصة Sketch Engine واستخدام أدوات البحث المتاحة فيها (Sketch Engine, n.d.).

أما مدونة الجاسم (٢٠٢٢) فهي مدونة لغوية خام، يزيد عدد كلماتها عن ٣٥ مليون كلمة، وتنوع مجالاتها لتشمل: الطب والعلوم الصحية، العلوم والتقنية والهندسة، التربية، العلوم الدينية، العلوم اللغوية والاجتماعية، القانون والعلوم السياسية، الإدارة والاقتصاد. وتشمل أوعية النشر: الكتب، الرسائل العلمية، الدوريات المحكمة.

٥- الإطار التطبيقي

٥-١- توصيف القواعد اللغوية وترميزها

تستند العديد من تطبيقات معالجة اللغات الطبيعية على وجود قواعد لغوية واضحة وقابلة للتفسير الحاسوبي، ومن أهمها قواعد التحليل الصرفي. فتوصيف القواعد اللغوية وفق منهج منظم ومدرّس يعد أساساً لتطوير أدوات لغوية قادرة على التعامل مع الأبواب الصرفية المختلفة، ومنها باب التصغير الذي يشترك مع أنماط صرفية أخرى في البنية

الشكلية مما يؤدي إلى اللبس في استرجاعها وتصنيفها. لذا فإن وضع قواعد لاسترجاع هذا النوع من الكلمات، مع الأخذ في الاعتبار المواضع التركيبية التي ترد فيها، يُسهم في رفع جودة التحليل الصرفي وتقليل اللبس.

٥-٢- توصيف القواعد اللغوية لصيغ التصغير

لتوصيف القواعد اللغوية لكل وزن من أوزان التصغير، أحصينا سياقات كل وزن على حدة تمهيداً لصياغة قواعدها وترميزها؛ وذلك بالاعتماد على المدونة العربية، ومدونة arTenTen، فبدأنا بملاحظة السياقات التي تحتوي على صيغ التصغير؛ للوصول إلى السمات التي تميزها ونستطيع من خلالها أن نسن قواعدها اللغوية، مع تحديد مجال الملاحظة والتحليل، وهو ثلاث كلمات تسبق الكلمة الهدف، وثلاث كلمات تليها؛ حيث إنَّ هذا العدد من الكلمات السابقة واللاحقة - وبعد الاطلاع على سياقات العينة - يرتبط بالكلمة الهدف، ويُسهم في تحديد دلالتها وصيغتها الصرفية.

وسنقسّم السياقات بحسب صيغ التصغير موضحين في كلٍّ منها القواعد اللغوية التي صغناها بناءً على هذه السياقات، ومنتهين بترميز هذه القواعد باستخدام التعبيرات النمطية regular expression. وهي طريقة للبحث في النصوص، تعتمد على رموز خاصة تُمكن من استخراج أجزاء محددة من تلك النصوص، وتدعم استعمالها اللغات البرمجية المختلفة، انظر الجداول (٢-١٣).

جدول (٢) التعابير المستخدمة في توصيف القواعد اللغوية

التعبير النمطي	دلالاته
\b	يُمثّل هذا التعبير حدود الكلمة، أي: بدايتها ونهايتها. (Word Boundary).
\s	يُمثّل هذا التعبير مسافة فارغة (Space).
[أ-ي]	يُقصد بهذا التعبير وجود أحد الحروف من الألف إلى الياء في النص.
[أ-ي][أ-ي] ي [أ-ي]	يُمثّل هذا التعبير كل كلمة على وزن (فُعِيل) مثل (جَبِيل)، حيث إنَّ الياء (ي) حرف ثابت، بينما يُمثّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).



التعبير النمطي	دلالاته
[أ-ي] [أ-ي] ي [أ-ي] [أ-ي]	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْل) مثل (مَكَيْتَب)، حيث إنَّ الياء (ي) حرف ثابت، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] ي	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْل) مثل: (قَنِيدِل)، حيث إنَّ الياء (ي) حرف ثابت، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] اء	يُمثِّل هذا التعبير كل كلمة على وزن (فُعِلَاء) مثل: (حَمِيرَاء) تصغير (فعلاء)، حيث إنَّ (ي) و (اء) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] ي	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْلِي) مثل: (صَغِيرِي) تصغير (فعلى)، حيث إنَّ (ي) و (ي) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] [ان]	يُمثِّل هذا التعبير كل كلمة على وزن (فُعِلَان) مثل: (عَطِيشَان) تصغير (فعلان)، حيث إنَّ (ي) و (ان) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
أ [أ-ي] ي [أ-ي] ا [أ-ي]	يُمثِّل هذا التعبير كل كلمة على وزن (أَفْعِيَال) مثل: (أَبْيَطَال)، تصغير (أفعال)، حيث إنَّ (أ) و (ي) و (ا) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] ة	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْلَة)، مثل: (شَجِيرَة)، حيث إنَّ (ي) و (ة) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
[أ-ي] [أ-ي] ي [أ-ي] ي	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْلَة)، بزيادة بتاء تأنيث بعد أربعة أحرف، مثل: (مَدِيرَة)، حيث إنَّ (ي) و (ة) حروف ثابتة، بينما يُمثِّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).

التعبير النمطي	دلالاته
[أ-ي] [أ-ي] ي [أ-ي] [أ-ي]	يُمثِّل هذا التعبير كل كلمة على وزن (فُعَيْلاء)، وهل للكلمات المختومة بألف التأنيث الممدودة ومدتها خامسة، مثل: (كهيراء) حيث إنّ (ي) و(ة) حروف ثابتة، بينما يُمثَّل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل).
أ [أ-ي] ي [أ-ي] [أ-ي] [أ-ي] ة	يُمثِّل هذا التعبير كل كلمة على وزن (أَفْعِيلة) مثل: (أغيلمّة)، التي تمثل شواذ التصغير لجموع التكسير على وزن (فَعْلَة)، حيث تمثل (أ)، و(ي)، و(ة) حروف ثابتة، وتمثل [أ-ي] الحروف غير الثابتة، أي الحروف المقابلة لـ (ف-ع-ل)..
V	فعل
N	اسم
Adj	صفة
ال	ال التعريف
[إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و]	حروف الجر
أم أوبل ثم حتى ف لكن لا و]	حروف العطف
[ذا هذا]	أسماء الإشارة
[أنتَ، أنتِ، أنتمَا، أنتم، أنتن]	ضمائر المخاطب
[هو هي هما هن هم]	ضمائر الغائب
[ي ناك كما كم كن هها هما هم هن ت نا واي]	الضمائر المتصلة



التعبير النمطي	دلالاته
[كما]	حرف الجر "الكاف" متصلاً بـ "ما" المصدرية
[لا]	لا النافية للجنس
[ثم أو حتى أم إما لكن بل لن كي كيلاً أن إذن إذًا لنلا لا لم لما إن إذ ما من متى أي أين أيان حيثما مهما لولولا لوما هلا لا سوف قد ل]	الأدوات السابقة للفعل المضارع
؟	يُقصد بهذا التعبير أن وجود ما يسبقه اختياري

جدول (٣) توصيف القواعد اللغوية وترميزها لوزن (فُعِيل)

توصيف القاعدة وترميزها	مثال
أن تكون الكلمة التالية للكلمة الهدف اسمًا، ويمكن أن يتصل هذا الاسم بأل التعريف، أو أن يفصل بينه وبين الكلمة الهدف حرف جر، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
تُحيت عنوان، صُنيم من صنيحات، بييت بوسط، تحيت سيطرة، دوين مائة، قبيل أسبوعين، قبيل بدء، سييد خالد، فوقيق التوسط، فوقيق الثنايا، بعيد الاجتماع، قريب المسجد، قريب الساعة، بعيد القرار، دوين السحاب، مييت الجاهلي، مويه خلف. \b [أ-ي] [أ-ي] ي [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ ال ؟ \b [N]	
أن تكون الكلمة التالية للكلمة الهدف صفة، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [أ-ي] [أ-ي] ي [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ ال ؟ \b [adj] نهير صغير، غصين صغير، جبيل أسود.	

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة السابقة للكلمة الهدف اسمًا ويمكن أن يفصل بينهما حرف عطف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
صاحب قليم، قارة أوجيل، ضوء قمير، ضفة نهير، شبه مييت.	$\backslash b [N] \backslash s$ أم أو بل ثم حتى ف لكن لا و ؟ $\backslash s [أ-ي] [أ-ي]$ $\backslash b [أ-ي]$
أن تكون الكلمة السابقة للكلمة الهدف فعلًا، ويمكن أن يسبقه حرف عطف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
وغاب قمير.	$\backslash b [V] \backslash s$ أم أو بل ثم حتى ف لكن لا و ؟ $\backslash s [أ-ي] [أ-ي]$ $\backslash b [أ-ي]$
أن تكون الكلمة التالية للكلمة الهدف مباشرة فعلًا، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
رجيل يدعى.	$\backslash b [أ-ي] [أ-ي] [أ-ي] \backslash s [V]$
أن تكون الكلمة السابقة للكلمة الهدف اسم إشارة، ويمكن اتصال هذا الاسم بضمير، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل ماعدا (ال) التعريف.	
ذاك الرجل.	$\backslash b [أ-ي] [أ-ي] [أ-ي] \backslash s$ ذا هذا ؟ $\backslash b [أ-ي] [أ-ي] [أ-ي]$
أن تكون الكلمة التالية للكلمة الهدف فعلًا، ويمكن أن يفصل بينهما حرف الجر "الكاف" متصلًا بما المصدرية، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
سُييد كما تقول.	$\backslash b [أ-ي] [أ-ي] [أ-ي] \backslash s [كما] \backslash s [V]$
أن تكون الكلمة السابقة للكلمة الهدف اسمًا، ويمكن أن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
جليل في كتيب.	$\backslash b [N] [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] \backslash s [أ-ي] [أ-ي] [أ-ي] \backslash b [أ-ي]$
أن تكون الكلمة التالية للكلمة الهدف مباشرة فعلًا، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	



مثال	توصيف القاعدة وترميزها
مويل يشقّف، مويل يقع.	\b [أ-ي] [أ-ي] ي [أ-ي] \s [V] \b
أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن اتصاله بضمير متصل، أو أن يفصل بينهما حرف عطف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
ولادته وعيب، أصحابك فلا عيب.	\b [N] [ي ناك كما كم كن ها هما هم هن] \s [أم أو بل ثم حتى ف لكن لا و] \s [أ-ي] [ي -ي] ي [أ-ي] \b
أن تكون الكلمة التالية للكلمة الهدف مباشرة اسماً، ويمكن اتصاله بأل التعريف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
بويب الكوخ.	\b [أ-ي] [أ-ي] ي [أ-ي] \s [ال] \b [N]
أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن اتصاله بضمير متصل أو أن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
ربه من بويب.	\b [N] [ي ناك كما كم كن ها هما هم هن] \s [إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و] \s [أ-ي] و ي [أ-ي] \b
أن تكون الكلمة السابقة للكلمة الهدف فعلاً، ويمكن أن يفصل بينهما حرف جر متصل بضمير، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
يطلق عليها تويج.	\b [V] \s [إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و] [ي ناك كما كم كن ها هما هم هن] \s [أ-ي] [أ-ي] ي [أ-ي] \b
أن تكون الكلمة التالية للكلمة الهدف صفة، ويمكن اتصال الكلمة الهدف بحرف جر.	
عصية صغيرة، بعصية صغيرة.	\b [إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و] [أ-ي] [أ-ي] ي [أ-ي] \s [adj] \b

جدول (٤) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْل)

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة التالية للكلمة الهدف اسمًا ويمكن أن يفصل بينهما حرف عطف أو أن يتصل الاسم بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [أ-ي] أم أو بل ثم حتى ف لكن لا و [N] ؟ [ي ناك كما كم كن هها هما هم هن] \b ؟
لييلة وضحاها، أحيمق فوالله.	
أن تكون الكلمة التالية للكلمة الهدف اسمًا ويمكن باتصاله بال التعريف أو أن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
مُليعب للكرة، مُكيتب في أحد الطوابق، مُجِليس الإدارة، مغيرب يوم، مغيرب الشمس، كويكب آخر، لييلة الميلاد.	\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [أ-ي] إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و [؟ \s ال ؟ \b [N]
أن تكون الكلمة التالية للكلمة الهدف صفة، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
مكيتب صغير.	\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [أ-ي] إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و [؟ \s ال ؟ \b [adj]
أن تكون الكلمة التالية للكلمة الهدف اسمًا ويمكن أن يتصل بال التعريف وأن يفصل بينهما حرف عطف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
كويكب والكواكب.	\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [أ-ي] أم أو بل ثم حتى ف لكن لا و [N] ؟ [ي ناك كما كم كن هها هما هم هن] \b ؟
أن تكون الكلمة التالية للكلمة الهدف فعلًا، ويمكن باتصال الفعل بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
منيزل لبسنا.	\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [V] [ي ناك كما كم كن هها هما هم هن] \b ؟



مثال	توصيف القاعدة وترميزها
	أن تكون الكلمة التالية للكلمة الهدف مباشرةً فعل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
أحيق قم.	\b[أ-ي][أ-ي]ي[أ-ي][أ-ي]\s[V]\b
	أن تكون الكلمة التالية للكلمة الهدف اسماً ويمكن أن يفصل بينهما أحد الضمائر، ويمكن أن يفصل بين الاسم والضمير حرف جر متصل بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
منيزل أنت به حريّ.	\b[أ-ي][أ-ي]ي[أ-ي][أ-ي]\s[أنت]\s[إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و؟]ي ناك كما كم كن ها هما هم هن\b[N]
	أن تكون الكلمة التالية للكلمة الهدف فعلاً، ويمكن باتصال الفعل بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
منيزل يحله.	\b[أ-ي][أ-ي]ي[أ-ي][أ-ي]\s[V]ي ناك كما كم كن ها هما هم هن\b
	أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن أن يفصل بينهما حرف عطف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
ملعب أو مليعب.	\b[N]\s[أم أو بل ثم حتى ف لكن لا و؟]\s[أ-ي][أ-ي]ي[أ-ي]\b
	أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن أن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
جليل في كتيب.	\b[N]\s[إلى ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا مذ من منذ و؟]\s[أ-ي][أ-ي]ي[أ-ي]\b
	أن تكون الكلمة التالية للكلمة الهدف اسماً ويمكن أن يتصل بأل التعريف وأن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.

مثال	توصيف القاعدة وترميزها
كُتِبَ عن الصيام.	\b [أ-ي][أ-ي] ي [أ-ي] \s [إي ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا منذ من منذ و] \s [إي] ؟ \b [N]
أن تكون الكلمة التالية للكلمة الهدف اسماً ويمكن اتصاله بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
عُجِيزٌ عَارِضُهَا.	\b [أ-ي][أ-ي] ي [أ-ي] \s [N] [ي ناك كما كم كن ه ها هما هم هن] \b
أن تكون الكلمة التالية للكلمة الهدف فعلاً، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
عجيز تحلف.	\b [أ-ي][أ-ي] ي [أ-ي] \s [V] \b
أن تكون الكلمة التالية للكلمة الهدف مباشرة أداة خاصة بالفعل المضارع، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
غزِيل لم تزل.	\b [أ-ي][أ-ي] ي [أ-ي] \s [ثم أو حتى أم إما لكن بل لن كي كيلاً أن إذن إذا لئلا لم لما إن إذا من متى أي أين أيان حيثما مهما لو لولا لوما هلا ألا سوف قد ل] \s [V] \b
أن تكون الكلمة التالية للكلمة الهدف فعلاً، ويمكن اتصال الفعل بحرف عطف أو ضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
خويدم فقتله. رويجل فغضب.	\b [أ-ي][أ-ي] ي [أ-ي] \s [أم أو بل ثم حتى ف لكن لا و] [V] [ي ناك كما كم كن ه ها هما هم هن] ؟ \b
أن تكون الكلمة التالية للكلمة الهدف اسماً ويمكن اتصاله بال التعريف أو أن يفصل بينهما حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
خويدم العلماء. صويحب المقالة، كويتب المؤتمر، عويلم بكلام.	\b [أ-ي][أ-ي] ي [أ-ي] \s [إي ب ت حتى حاشا خلا ربّ عدا على عن في ك كي ل لولا منذ من منذ و] \s [إي] ؟ \b [N]



مثال	توصيف القاعدة وترميزها
	أن تكون الكلمة التالية للكلمة الهدف صفة، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
طوينر أبقع، شويعر بنديء، رويجل ضعيف.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ \b [adj] ؟
	أن تكون الكلمة التالية للكلمة الهدف فعلاً، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
كويتب يكتب.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [V] ؟
	أن تكون الكلمة السابقة للكلمة الهدف فعلاً، ويمكن اتصاله بضمير متصل، وأن يفصل بينهما حرف جر متصل بضمير، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
يوصل لنا صويحب، يسبقك إليه شويعر.	\b [V] ؟ [ي ناك كما كم كن هها هما هم هن] ؟ [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] [ي ناك كما كم كن هها هما هم هن] ؟ \s [أ-ي] [أ-ي] [أ-ي] \b [أ-ي] [أ-ي] [أ-ي] [أ-ي]
	أن تكون الكلمة التالية للكلمة الهدف اسماً ويمكن أن يفصل بينهما لا النافية للجنس، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
طوينر لا وجود.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [لا] ؟ \b [N] ؟
	أن تكون الكلمة السابقة للكلمة الهدف اسماً ويمكن أن يفصل بينهما حرف عطف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.
فقيه أو عويلم.	\b [N] ؟ أم أو بل ثم حتى ف لكن لا و \s [أ-ي] [أ-ي] [أ-ي] [أ-ي] [أ-ي] \b [أ-ي] [أ-ي] [أ-ي] [أ-ي]

جدول (٥) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْعِل)

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة التالية للكلمة الهدف اسماً، ولا بأس باتصاله بضمير متصل، أو أن يفصل بينه وبين الكلمة الهدف حرف جر، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي] [أ-ي] ي [أ-ي] ي [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ [N] ي ناك كما كم كن هـها هما هم هن ؟ \b
عُصيفير في فراشه، تلميذ متطفل، منيديل كلينكس، دنينير واحد.	
أن تكون الكلمة التالية للكلمة الهدف صفة، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي] [أ-ي] ي [أ-ي] ي [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ [adj] ي ناك كما كم كن هـها هما هم هن ؟ \b
تلميذ صغير، مسيكين ضعيف.	
أن تكون الكلمة التالية للكلمة الهدف مباشرةً فعلاً ماضياً، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي] [أ-ي] ي [أ-ي] ي [أ-ي] \s [V] \b
عصيفير هوى.	
أن تكون الكلمة التالية للكلمة الهدف مباشرةً فعل أمر ولا بأس باتصاله بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي] [أ-ي] ي [أ-ي] ي [أ-ي] \s [V] \b [ي ناك كما كم كن هـها هما هم هن ؟ \b
مُسيكين دعه.	
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن اتصاله بضمير متصل، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	\b [أ-ي] [أ-ي] ي [أ-ي] ي [أ-ي] \s [N] \b [ي ناك كما كم كن هـها هما هم هن ؟ \b
مُوزين حسناك.	



جدول (٦) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْلاء) تصغير (فعلاء)

توصيف القاعدة وترميزها	مثال
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلاء)، تكون الكلمة السابقة لها اسماً، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [N] \s [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] \b	حلة حميراء.

جدول (٧) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْلى) تصغير (فعلى)

توصيف القاعدة وترميزها	مثال
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلى)، تكون الكلمة السابقة لها اسماً، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [N] \s [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] \b	فتاة صغيرة، أهمية كبرى، شجرة عظيمة، طيوراً عظيمة، جريمة كبرى، طفلة صغيرة.

جدول (٨) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْلة) تصغير (فعيلة)

توصيف القاعدة وترميزها	مثال
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن أن يفصل بينه وبين الكلمة الهدف حرف جر، ويمكن باتصال الكلمة الهدف بحرف جر.	
\b [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] \s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ \b [N]	غريفة بمصعد.
أن تكون الكلمة التالية للكلمة الهدف صفة، ويمكن اتصال الكلمة الهدف بحرف جر.	

مثال	توصيف القاعدة وترميزها
شجيرة معمرة، غريفة سوداء، وريدة صغيرة، شجيرة صغيرة. ثميرة وردية، بصيغلة زاهية.	\b [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و؟ [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ [adj] \b
أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن اتصاله بضمير متصل، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
كسرة تميرة، فصّه وريقة، وريقاتها وريقة.	\b [N] [ي ناك كما كم كن هـ ها هما هم هن] ؟ \s [أ-ي] [أ-ي] [ة] \b
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن اتصاله بأل التعريف، أو أن يفصل بينه وبين الكلمة الهدف حرف جر، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
عيينة صغيرة، شميصة رمضان، أذينة مخفية، أريضة هريضة، دويرة الكتب، دويرة للصوفية.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ [N] \b
أن تكون الكلمة التالية للكلمة الهدف صفة، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
عيينة صغيرة.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و] ؟ [adj] \b
٧- أن تكون الكلمة التالية للكلمة الهدف مباشرة فعلاً، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
أريضة تنبت، نويرة تقد، نويرة تخبو.	\b [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [V] \b
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن أن يفصل بينهما اسم إشارة مسبوق بحرف عطف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	



مثال	توصيف القاعدة وترميزها
شميسة فهي شمس .	\b [أ-ي] [أ-ي] ي [أ-ي] [ة] \s [أم أو بل ثم حتى ف لكن لا و]؟ [هو هي هما هن هم] \s [N] \b
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن أن يتصل بأل التعريف أو أن يفصل بينهما جرف جر، ويمكن أن تتصل الكلمة الهدف بحرف جر.	
غيبية دائرية، سويعة من المشي، بسويعة على هيئة.	\b [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ \s [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] /s [إلى ب ت حتى حاشا خلا ربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ [N] \b

جدول (٩) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْلان) تصغير (فعلان)

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة السابقة للكلمة الهدف فعلاً، ويمكن اتصاله بحرف عطف أو بضمير متصل، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
وصرت عطيشان.	\b [أم أو بل ثم حتى ف لكن لا و]؟ [V] [ت ن ا و ي] \s [أ- ي] [أ-ي] ي [أ-ي] [ان] \b

جدول (١٠) توصيف القواعد اللغوية وترميزها لوزن (أفيعال) تصغير (أفعال)

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن اتصاله بأل التعريف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
أصيحاب المناصب.	\b [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] ا [أ-ي] \s [N] \b
أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن أن يفصل بينهما ظرف، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
ثالثة مع أصحباب له.	\b [N] [مع] \s [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] ا [أ-ي] \b

جدول (١١) توصيف القواعد اللغوية وترميزها للكلمات المختومة بألف التأنيث الممدودة ومدتها خامسة

توصيف القاعدة وترميزها	مثال
أن تكون الكلمة التالية للكلمة الهدف فعلاً، ويمكن أن يفصل بينه وبين الكلمة الهدف حرف عطف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
\b [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] [أ-ي] \s [ء] \s أم أو بل ثم حتى ف لكن لا و]؟ [b [V]	كهيرياء ونستمطر.

جدول (١٢) توصيف القواعد اللغوية وترميزها للكلمات المختومة بتاء تأنيث بعد أربعة أحرف

توصيف القاعدة وترميزها	مثال
أن تكون الكلمة التالية للكلمة الهدف اسمًا ويمكن اتصاله بأل التعريف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
\b [أ-ي] [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] [أ-ي] \s [ة] \s قنيطرة جسرية، مديرة التربية، مريحلة تاريخ.	
أن تكون الكلمة السابقة للكلمة الهدف فعلاً ويمكن أن يسبقه أداة خاصة بالفعل المضارع، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
\b [ثم أو حتى أم إما لكن بل لن كي كيلاً أن إذن إذا لئلا لا لم لما إن إذ ما من متى أي أين أيا ن حيثما مهما لو لولا لوما هلا ألا سوف قد ل]؟ [b [V] \s [أ-ي] [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] [أ-ي] [ة] \b لنقل مديرة.	
أن تكون الكلمة التالية للكلمة الهدف اسمًا ويمكن أن يفصل بينهما لا النافية للجنس، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
\b [أ-ي] [أ-ي] [أ-ي] ي [أ-ي] [أ-ي] [أ-ي] \s [ة] \s مريحلة لا تهاون.	



جدول (١٣) توصيف القواعد اللغوية وترميزها للكلمات المصغرة تصغيراً شاذاً على وزن (أفيعلة)

مثال	توصيف القاعدة وترميزها
أن تكون الكلمة التالية للكلمة الهدف اسماً، ويمكن اتصاله بأل التعريف، على ألا تتصل الكلمة الهدف بأي حرف سابق أو ضمير متصل.	
أصبيبة صغار، أغيلمة التحقيق، أغيلمة بني هاشم.	$[N] \text{ أ } \backslash b \text{ [أ-ي] [أ-ي] [أ-ي] \text{ ة } \backslash s \text{ [أ-ي] } [N]$

٣-٥- تقييم القواعد اللغوية لصيغ التصغير نظرياً وتطبيقياً

٣-٥-١- تحكيم القواعد اللغوية نظرياً

للتأكد من صحة القواعد اللغوية المقترحة نظرياً استعان الباحثان بثلاثة من الأساتذة المتخصصين في اللغويات واللغويات الحاسوبية لإبداء مرئياتهم ومقترحاتهم العلمية حول القواعد بشقها اللغوي الذي يشمل القواعد اللغوية بناءً على جمل العينة، وشقها الحاسوبي الذي يشمل توصيف القواعد وترميزها. واتفق المحكمون الثلاثة على مناسبة القواعد والتوصيف.

٣-٥-٢- تحكيم القواعد اللغوية تطبيقياً

من الناحية التطبيقية، اعتمدنا على مدونة لغوية خام، يزيد عدد كلماتها عن ٣٥ مليون كلمة، وهي مدونة الجاسم (٢٠٢٢)؛ لاستدعاء الكلمات المصغرة، والكلمات الثلاث التي تسبق هذه الكلمات وتليها، وذلك اعتماداً على القواعد اللغوية التي وصفناها. وتم ذلك من خلال استخدام حزم بايثون re لقراءة المدونة والبحث عن التعبيرات النمطية regular expression التي تُمثل كل قاعدة لوزن من أوزان الكلمة المصغرة. بعد ذلك فُصلت هذه الأنماط، واستُدعيت مع سياقاتها، ثم روجعت السياقات يدوياً، واستُبعدت منها الأسماء غير المصغرة التي تكون مشابهة لأسماء التصغير في الرسم الإملائي. وكذلك

استُبعدت أسماء الأعلام التي جاءت على أوزان التصغير مثل (نُجيم ومحمد). وفي الجدول (١٤)، عرضنا عدد الأسماء المصغرة لكل نمط من القواعد ونسبة الدقة accuracy لدى مطابقة النتائج للقواعد اللغوية ذات العلاقة. وتجدر الإشارة إلى أن هناك قواعد لم نجد لها أسماء مصغرة في المدونة، وهي: القواعد اللغوية لوزن (أُفيلة) للكلمات المصغرة تصغيراً شاذاً، والقواعد اللغوية للكلمات المختومة بتاء تأنيث بعد أربعة أحرف، والقواعد اللغوية للقواعد اللغوية للكلمات المختومة بألف التأنيث الممدودة ومدتها خامسة، والقواعد اللغوية لوزن (فُعيلان) تصغير (فعالان).

وتعرض نتائج الدقة في الجدول (١٤) وجود ثلاثة أنماط تقل فيها نسبة الدقة عن ١٠٠٪. ولتحسين الاسترجاع ورفع نسبة الدقة أضفنا قواعد لغوية بناء على سياقات مدونة الاختبار، انظر الجداول (١٥ - ١٧).

جدول (١٤) نتائج التقييم التطبيقي للقواعد اللغوية

نسبة الدقة لدى مطابقتها للقواعد	الكلمات المصغرة	القاعدة اللغوية
١٠٠٪	١٤١	القواعد اللغوية لوزن (فُعيل)
١٠٠٪	١٣٨	القواعد اللغوية لوزن (فُعيلع)
١٠٠٪	١	القواعد اللغوية لوزن (فُعيلعيل)
٥٠٪	١٢	القواعد اللغوية لوزن (فعيلاء) تصغير فعلاء
٥٠٪	٢	القواعد اللغوية لوزن (فُعيلي) تصغير فعلي
١٠٠٪	٢	القواعد اللغوية لوزن (أفيعال) تصغير أفعال
٩٧٪	١٨٩٩	القواعد اللغوية لوزن (فُعيلة)



جدول (١٥) توصيف القواعد اللغوية وترميزها على وزن فعيلاء تصغير (فعلاء) بعد تحديث القواعد

توصيف القاعدة وترميزها	مثال
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلاء)، تكون الكلمة السابقة لها اسماً، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [N] \s [أ-ي][أ-ي] ي [أ-ي][أ-ي] \b	حلة حميراء.
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلاء)، تكون الكلمة التالية لها اسماً، ويمكن باتصالها بأل التعريف أو ضمير متصل، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [أ-ي][أ-ي] ي [أ-ي][أ-ي] \s [أ-ي][أ-ي] \b	سويداء القلوب، سويداء قلوبهم.

جدول (١٦) توصيف القواعد اللغوية وترميزها لوزن (فُعَيْلى) تصغير (فعلى) بعد تحديث القواعد

توصيف القاعدة وترميزها	مثال
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلى)، تكون الكلمة السابقة لها اسماً، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.	
\b [N] \s [أ-ي][أ-ي] ي [أ-ي][أ-ي] \b	فتاة صغيرة، أهمية كبيرى، شجرة عظيمى، طيوراً عظيمى، جريمة كبيرى، طفلة صغيرة.
إذا كانت الكلمة المُصَغَّرَة على وزن (فعيلى)، تكون الكلمة السابقة لها فعلاً مضارعاً ويمكن باتصاله بضمير متصل، ويمكن أن تتصل الكلمة الهدف بأل التعريف.	
\b [V] [ي ناك كما كم كن هـ ها هما هم هن ت ن ا و ا ي] \s [أ-ي][أ-ي] ي [أ-ي][أ-ي] \b	يمشين الهوينى.

جدول (١٧) توصيف القواعد اللغوية وترميزها لوزن (فَعِيلَة) بعد تحديث القواعد

مثال	توصيف القاعدة وترميزها
	أن تكون الكلمة التالية للكلمة الهدف اسم، ويمكن أن يفصل بينه وبين الكلمة الهدف حرف جر.
غريفة بمصعد	\b [إلى ب ت حتى حاشا خلاربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [إلى ب ت حتى حاشا خلاربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ \b [N]
	أن تكون الكلمة التالية للكلمة الهدف صفة، ويمكن باتصال الكلمة الهدف بحرف جر.
شجيرة معمرة، غريفة سوداء، وريدة صغيرة، شجيرة صغيرة. ثميرة وردية، بصبيغة زاهية.	\b [إلى ب ت حتى حاشا خلاربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ [أ-ي] [أ-ي] [أ-ي] [أ-ي] \s [ة] [إلى ب ت حتى حاشا خلاربَّ عدا على عن في ك كي ل لولا مذ من منذ و]؟ \b [adj]
	أن تكون الكلمة السابقة للكلمة الهدف اسماً، ويمكن اتصاله بضمير متصل، على ألا تتصل الكلمة الهدف بحرف سابق أو ضمير متصل.
كسرة تميرة، فصّه وريقة، وريقاتها وريقة.	\b [N] [ي ناك كما كم كن هها هما هم هن]؟ \s [أ-ي] [أ-ي] [أ-ي] [أ-ي] \b [ة]

٦- الخاتمة

خلصت هذه الدراسة إلى اقتراح بعض الحلول لتحسين استرجاع الكلمات المصغرة وتصنيفها تصنيفاً صحيحاً، من خلال استخراج قواعدها اللغوية وتوصيفها حاسوبياً. وقد حكم هذه القواعد نظرياً أساتذة متخصصون في اللغويات واللغويات الحاسوبية. أما تطبيقياً فقد تحققنا من صحة القواعد بالاعتماد على مدونة لغوية خام، يزيد عدد كلماتها عن خمسة وثلاثين مليون كلمة، واستدعاء الكلمات المصغرة من خلال استخدام كود برمجي بلغة بايثون Python، ثم التأكد من مطابقة سياق الكلمات المصغرة للقواعد اللغوية المقترحة. وبهذا تكون الدراسة قدّمت عدداً من الحلول المقترحة والمحكمة لتحسين استرجاع هذه الكلمات، ووصفت خمساً وستين قاعدة لغوية لأوزان التصغير. وقد أظهرت نتائج تقييم القواعد اللغوية مطابقتها لما ورد في الدراسة بنسبة ٩٩٪.

المراجع العربية

- الأزهري، محمد بن أحمد. (٢٠٠١). تهذيب اللغة. دار إحياء التراث العربي.
- الجاسم، حوراء علي. (٢٠٢٢). الألفاظ الأكاديمية في المدونات العربية العلمية [رسالة ماجستير غير منشورة]. جامعة الملك سعود.
- الجرجاني، علي بن محمد بن علي الزين الشريف. (١٩٨٣). التعريفات. دار الكتب العلمية.
- الدوسري، مسفر محماس. (٢٠١٠). برمجة الاسم المنسوب ببناء النسب في العربية حاسوبياً. [رسالة دكتوراه غير منشورة]. جامعة اليرموك.
- الراجحي، عبده. (د.ت). التطبيق الصرفي. دار النهضة العربية.
- الروابدة، محمد. (٢٠١٠). التصغير في اللغة العربية. نظرة في الدلالة والتحليل الصوتي. مجلة مجمع اللغة العربية الأردني، (٧٩)، ١١-٤٧.
- الزامل، عبد الله عبد الرحمن. (٢٠٠٧). العلاقة الصرفية بين الجذور والأوزان (تصنيف جديد لجذور اللغة العربية). مدينة الملك عبد العزيز للعلوم والتقنية.
- زايد، محمد. (٢٠٠٩، أبريل ٢٦-٢٨). تقرير في المحللات الصرفية للغة العربية. في ندوة المحللات الصرفية الحاسوبية (٢٠٠٩). دمشق.
- السامرائي، محمد فاضل. (٢٠١٣). الصرف العربي أحكام ومعان. دار ابن كثير.
- عبد الغني، أيمن أمين. (١٩٩٩). الصرف الكافي. مكتبة ابن خلدون.
- الغنيان، حسان عبد الله (د، ت). الواضح في الصرف. جامعة الملك سعود.
- منصة فلك للمدونات اللغوية. (د.ت). المدونة العربية. مجمع الملك سلمان العالمي للغة العربية. <https://falak.ksaa.gov.sa/corpora/arabiccorpus>
- فهيم، رفيدة أحمد. (٢٠٢١). بناء خوارزمية حاسوبية للتوليد والتحليل الصرفيين في بابي اسم الفاعل والصفة المشبهة. [رسالة ماجستير غير منشورة]. جامعة الإمام محمد بن سعود الإسلامية.

الفيضي، عبد الله، يحيى. (٢٠١٥). *مدونات المتعلمين*. في صالح فهد العصيمي (محرر)، *المدونات اللغوية العربية بناؤها وطرق الإفادة منها*. مجمع الملك سلمان العالمي للغة العربية.
القيشي، هاجر عبد الرحمن. (٢٠١٩). *بناء خوارزمية لفك اللبس الصرفي الحاسوبي في جموع القلة*. [رسالة ماجستير غير منشورة]. جامعة الإمام محمد بن سعود الإسلامية.
ابن منظور، محمد بن مكرم علي. (١٩٩٤). *لسان العرب*. دار صادر.

رومنة المراجع العربية

- Abdul Ghani, Ayman Amin. (1999). *Al-Serf Al-Kafee*. Mektabet Ibn Khaldun. (In Arabic)
- Al-Azhari, Muhammad ibn Ahmad. (2001). *Tahtheeb Al-Lughah*. Dar Ihya' Al-Turath Al-Arabi. (In Arabic)
- Al-Dosari, Misfir Mehmas. (2010). *Barmajat Al-Ism Al-Mansoob bi yaa Al-Nasab hasoobian*. [risalat dukturah ghayr manshuratin]. Jameat Al-Yarmouk. (In Arabic)
- Al-Fifi, Abdullah, Yahya. (2015). *Mudawwanat Al-Muta'allemeen. fi salih fahd aleasaymi (mhrrr), almudawwanat allughawiat alearabiat binawuha waturuq al'iifadat minha*. Mujama'a Al-Melek Salman Al-'Aalemy Le Al-Lughah Al-'Arabiah. (In Arabic)
- Al-Ghuniman, Hassan Abdullah (d.t.). *Al-Wadhih fi Al-Serf*. Jame'at Al-Melek Saud. (In Arabic)
- Al-Jassem, Hawra Ali. (2022). *Al-Alfadh Al-Academiah fi Al-Mudawwanat Al-'Arabiah Al-'Ilmiah*, [risalat majistir ghayr manshuratin]. Jameat Al-Melek Saud. (In Arabic)
- Al-Jurjani, Ali bin Muhammad bin Ali Al-Zain Al-Sharif. (1983). *Al-T'areefat. Dar Al-Kotob Al-'Ilmiyah*. (In Arabic)
- Al-Qubaishi, Hajar Abdul Rahman. (2019). *Bena'a Khawarizmiah Li Fak Al-Labs Al-Serfi Al-Hasoobi Fi Jomo'o Al-Qellah*. [risalat majistir ghayr manshuratin]. Jame'at Al-Imam Muhammad ibn Saud Al-Islamiah. (In Arabic)
- Al-Rajhi, Abdo. (n.d.) *Al-Tatbeeq Al-Sarfi*. Dar Al-Nahda Al-'Arabiya. (In Arabic)
- Al-Rawabdeh, Muhammad. (2010). *altasghir fi allughat alearabiat. nazrat fi aldilalat waltahlil alsawti. majalat majmae allughat alearabiat al'urduniyati*, (79), 11-47. (In Arabic)



- Al-Samarrai, Muhammad Fadhel. (2013). *Al-Serf Al-'Arabi Ahkam wa Ma'anen*. Dar Ibn Kathir. (In Arabic)
- Al-Zamil, Abdullah Abdul Rahman. (2007). *Al-'Alaqah Al-Saerfiah bayna Al-Jothoor wa Al-Awzan (Tasneef jadeed li Jothoor Al-Lughah Al-'Arabiah)*. Madeenat Al-Malek Abdulaziz li Al-'Uloom wa Al-Taqniah. (In Arabic)
- Fahmy, Rafidah Ahmed. (2021). *Bena Khawarezmiah Hasoobiah Li Al-Tawleed Wa Al-Tahleel Al-Sefeyen Fi Baby Ism Al-Fael Wa Al-Sefeh Al-Mushabbaheh*. [risalat majistir ghayr manshuratin]. Jame'at Al-Imam Muhammad ibn Saud Al-Islamiah. (In Arabic)
<https://falak.ksaa.gov.sa/corpora/arabiccorpus>
- Ibn Manzur, Muhammad ibn Makram Ali. (1994). *Lisan al-Arab*. Dar Sadir. (In Arabic)
- Manassat Falak Li Al-Mudawwanat Al-Lughawiah. (n.d.). *Al-Mudawwanah Al-'Arabiah*. Mujamm'a Al-Melek Salman Al-A'alemy Li Al-Lughah Al-'Arabiah. (In Arabic)
- Zayed, Mohammed. (2009, 26-28 'abril). taqir fi almuhalilat alsarfiah lilughat alearabiati. *fi nadwat almuhalilat alsarfiah alhasubia (2009)*. dimashqa. (In Arabic)

المراجع الأجنبية

- Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00277ED1V01Y201008HLLT010>
- Naaser, M. S. A. (2020). A computational morphological model for Arabic diminutive generation. *International Journal of Linguistics and Translation Studies*, 1(2).
- Sketch Engine. (n.d.). *arTenTen: Corpus of the Arabic Web* <https://www.sketchengine.eu/artenten-arabic-corpus/>
- Thalji, Z. (2018). A new algorithm to minimize names in the Arabic language. *International Journal of Applied Engineering Research*, 13(18), 13950-13960. Research India Publications.
- Zivingi, S., & Sulaiman, R. (2022). Comparative morphological analysis of diminutive and augmentative in English, Kurdish, and Arabic. *International Journal of Communication*, 9(2), 132.

نبذة تعريفية عن الباحثين

Author Biodata Dr. Halah Al-Harbi serves as a Senior Specialist in Corpus Development at King Salman Global Academy for Arabic Language. She earned her Ph.D. in Applied Linguistics from Imam Mohammad Ibn Saud Islamic University in 2024. Her research interests lie in computational and applied linguistics.	بيانات الباحث: د. هاله الحربي ، أخصائي أول بناء المدونات اللغوية في مجمع الملك سلمان العالمي للغة العربية، حاصلة على شهادة الدكتوراة في اللغويات التطبيقية من جامعة الإمام محمد بن سعود الإسلامية عام ٢٠٢٤. تدور اهتماماتها البحثية حول اللغويات الحاسوبية والتطبيقية.
معرف أوركيد (ORCID) : 0009-0005-6707-6848	
Email: halah.m.alharbi@gmail.com	

Author Biodata Dr. Abdullah Alfaifi is an Associate Professor of Computational Linguistics and the Head of the Computational Linguistics Department at King Salman Global Academy for Arabic Language. He holds a Ph.D. in Natural Language Processing from the School of Computing at the University of Leeds, United Kingdom. He has published over 20 peer-reviewed research papers in scientific journals and international conferences. He has also co-authored several books on artificial intelligence techniques in Arabic language processing. He serves as a member of several advisory boards and scientific committees for specialized academic journals.	بيانات الباحث: د. عبد الله الفيافي ، أستاذ مشارك في اللغويات الحاسوبية، ورئيس قطاع الحوسبة اللغوية في مجمع الملك سلمان العالمي للغة العربية. حاصل على درجة الدكتوراة في المعالجة الآلية للغة الطبيعية من قسم الحاسب الآلي بجامعة Leeds في المملكة المتحدة. نشر أكثر من ٢٠ بحثاً علمياً في مجلات علمية ومؤتمرات، كما شارك في تأليف عدد من الكتب حول تقنيات الذكاء الاصطناعي في حوسبة اللغة العربية. عضو في عدد من اللجان والهيئات الاستشارية والعلمية في مجلات علمية متخصصة.
معرف أوركيد (ORCID) : 0000-0001-5819-431X	
Email: afaifi@gmail.com	

نحو آلية جديدة للرصد اللغوي للغة العربية

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-003>

علاء ممدوح عاكف

كلية اللغات الأجنبية، جامعة بكين، بكين، الصين

نُشر في	قُبِل في	رُوجع في	استقبل في
١٤٤٦/١٠/٢٨	١٤٤٦/٠٤/٠٣	١٤٤٦/٠٣/٠٣	١٤٤٦/٠٢/٢٨

المخلص

تطرح ورقتنا البحثية آلية للرصد اللغوي للغة العربية، تلك آلية مستقاة من نظريات وتقنيات الرصد اللغوي في اللغة الصينية. وللتأكد من فاعلية تلك الآلية، قمنا أولاً ببناء مدونة لغوية ديناميكية للصحف العربية المتداولة، تمثل خمس لهجات عربية (الخليجية والعراقية والشامية والمصرية ولهجة بلاد المغرب)، وتحتوي على مليار كلمة و٤ مليون نص إخباري، وتغطي الفترة من ٢٠٠٥ وحتى فبراير ٢٠٢٤. ثانياً، استخدمنا الطرق اللغوية الحاسوبية المستخدمة في اللغة الصينية والتي طورناها لحوسبة واستخراج الكلمات العامة من مدونة الصحف السابقة، وهي الكلمات التي تمثل الاستقرار اللغوي للمفردات العربية. كما قمنا باستخراج الكلمات التي تمثل التغير الديناميكي الزماني والمكاني للمفردات من نفس المدونة، مثل الكلمات الخاصة التي تعبر عن خصوصية الاستخدام اللغوي للكلمة في مجتمع ما، والكلمات المميزة التي تعبر عن أهم الأحداث السياسية والاقتصادية والرياضية المحلية أو الدولية خلال فترة زمنية معينة.

الكلمات المفتاحية

اللغة العربية؛ الرصد اللغوي؛ المدونة الديناميكية متنوعة الأوعية اللغوية؛ الاستعمال المعجمي؛ التقنيات الحاسوبية.

Towards a New Mechanism for Arabic Language Monitoring

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-003>

Alaa Mamdouh Akef^{ID}

Department of Arabic Language, School of Foreign Languages, Peking University.

Received on	Revised on	Accepted on	Published on
01/09/2024	05/09/2024	06/10/2024	26/04/2025

ABS TRACT

This paper proposes an innovative mechanism for monitoring of the Arabic language, inspired by the theories and techniques used in Chinese linguistic Monitoring studies. To verify the effectiveness of this mechanism, we first constructed a dynamic Arabic newspaper corpus, comprising major newspapers representing five major Arabic dialects (Gulf, Iraqi, Levantine, Egyptian, and Maghrebi). This corpus includes 1 billion words and 4 million news articles, covering the period between 2005 and February 2024. Secondly, we applied computational linguistic methods -either adopted from Chinese linguistic practices or developed by us- to computationally extract general words from the corpus. These words represent lexical stability of Arabic lexical. Concurrently, we identified words that reflect the spatiotemporal dynamism of Arabic lexical, such as unique words, which signify culturally specific lexical usage in each community, and distinctive words, which highlight major political, economic, and sports-related events on both local and international levels within a specific timeframe.

KEYWORDS

Arabic Language; Language Monitoring; Dynamic Circumstance Corpus; Lexical Usage; Computational Methods.

١- مقدمة

تُعرّف العالمتان اللغويتان هومين ويانغ أرهونغ (٢٠١٥) الرصد اللغوي (Language monitoring) في ورقتهما البحثية "عشر سنوات على دراسات الرصد اللغوي الصيني" التي تستعرضان فيها البنية العلمية ونتائج الرصد اللغوي في الصين في الفترة بين ٢٠٠٥ و٢٠١٥ والمنشورة في مجلة "اللغويات التطبيقية" بأن الرصد اللغوي هو دراسة لغوية اجتماعية واسعة النطاق، تضم علماء متعددي التخصصات يسترشدون بنظريات علم اللغة واللغويات التطبيقية وغيرها من النظريات المتعلقة باللغة، ويستخدمون تقنيات معالجة البيانات والمسح الميداني بشكل رئيسي وأساسي. وتوضحان في ورقتهما المشتركة أن الرصد اللغوي يتضمن ثلاثة عناصر أساسية هي: اللغة، والإحصاء، والمجتمع، أو يشمل استخدام طرق الإحصاء المختلفة لوصف اللغوي والتحليل المجتمعي.

ومن التعريف السابق نجد أن الرصد الذي تقصده العالمتان اللغويتان في ورقتهما هو مستوى متقدم من الإحصاء اللغوي. مستوى لا يستعرض مجموعة أرقام لظواهر لغوية فحسب، أو يقوم بتجربة طرق إحصائية مختلفة وحساب نسب نجاحها فقط، ولكنه يتخذ من نتائج الإحصائيات اللغوية أساساً يعكس التغير اللغوي (المكاني والزمني) للغة الواحدة، ويحلل في ضوءها التغيرات اللغوية والمجتمعية، في ربط جديد بين علم اللغة والمجتمع، تدعمه بيانات إحصائية لغوية، حُسبت بناء على مدونات لغوية (Corpus) ضخمة وديناميكية (Dynamic) ومتنوعة البيانات.

وتستعرض الباحثتان في ورقتهما العلمية قيام مراكز لغوية بحثية صينية بدعم من وزارة التعليم الصينية بإنشاء مدونات لغوية ديناميكية (Dynamic corpus)، متخذة من الصحف الصينية، والإعلام المرئي الصيني، ومدونات الإنترنت، والمحتويات التعليمية الصينية، وغيرها من المصادر اللغوية المختلفة مادة لها، وقد صممت تلك المراكز البحثية اللغوية منصات لرصد التغيرات التي تتطراً على الاستخدام اللغوي الصيني على صعيد الرموز، والمفردات، والعبارات، والجمل رصداً آلياً. كما وضحت الباحثتان البنية الهيكلية لتلك المنصات، وكيفية تحليل مخرجات تلك المنصات من الإحصائيات اللغوية التي يقوم

فريق من أفضل خبراء اللغة الصينية في الصين بتحليلها، وتقديم صورة شاملة عن تطور اللغة الصينية يومياً، وتأثير هذا التطور اللغوي على المجتمع الصيني، في تقرير سنوي يسمى "الحالة العامة للغة الصينية"، يضم إحصائيات عن استخدام الرموز والكلمات والعبارات المسكوكة الصينية، ويضم أيضاً مقالات تحليلية لتلك الإحصائيات المتنوعة.

وغني عن الذكر ذلك التطور الهائل الذي حققته اللغة الإنجليزية في مجال الرصد اللغوي، إذ يكفي أن نرى موقع المرصد اللغوي العالمي^(١) (Global Language Monitor) الذي يستعرض كل عام الكلمات الإنجليزية الجديدة، والكلمات الإنجليزية الأكثر استخداماً خلال العام، وكلمات العام (Word of the Year)، أو الكلمة التي تلخص أحداث العام في كلمة واحدة. هذا بخلاف استخدام الرصد اللغوي لخدمة نشر اللغة الإنجليزية وتعليمها، وأشهر تلك المحاولات هو معجم كولينز (Collins Dictionary)، الذي كان ثمرة لمشروع قاعدة البيانات اللغوية العالمية بالتعاون بين كولنز وجامعة برمنجهام "Collins Birmingham University International Language Database"، الذي يُعرف اختصاراً باسم "COBUILD"، ومعجم لونجمان للإنجليزية المعاصرة "Longman Dictionary of Contemporary English"، الذي استعان بمدونات لغوية إنجليزية ضخمة ومتعددة تتيح للمستخدم التعرف على الاستخدام الفعلي للكلمات في الحياة اليومية، عن طريق تقديم إحصائيات لاستخدام المفردات والمصاحبات اللفظية والتعبير المسكوكة وغيرها من الظواهر اللغوية في المدونات اللغوية، كما توضح الفروق في استخدام الكلمات عند متحدثي الإنجليزية بصورها المختلفة في الدول الناطقة بها.

أما على صعيد الرصد اللغوي في اللغة العربية، فعلى الرغم من وجود محاولات عديدة من دول عربية ومراكز بحثية من دول أخرى لإنشاء مدونات لغوية مفتوحة المصدر تضم بيانات لغوية ضخمة ومتنوعة، تتيح للمستخدم البحث عن عدد مرات تكرار الكلمة ومعدل تكرارها في المدونة، وتقدم أيضاً خاصية الكشف السياقي والتصاحب اللفظي للكلمة، مثل "المدونة العربية"، و"المدونة اللغوية العربية العالمية لمكتبة الإسكندرية"، أو تدعم بعضها عرض الأشكال الصرفية المختلفة للكلمة الواحدة، كما في "مدونة اللغة العربية"

(1) <https://languagemonitor.com/category/new-words/>



التابعة لجامعة برجهام ينج (Brigham Young University Arabi Corpus)، على الرغم من ذلك إلا أنَّ محتوى تلك المدونات لا يتغير ولا يُحدَّث باستمرار، أي أنها مدونات لغوية إستاتيكية (Static corpus)، وهذا يؤثر على نتائج الإحصائيات اللغوية التي تقدمها تلك المدونات. من جهة أخرى، فإنَّ الإحصائيات اللغوية القائمة على تلك المدونات اللغوية -على الرغم من أنها محاولات مهمة للغاية- إلا أنها كانت محدودة النطاق، ولم تُطبق على المدونة بأكملها، بل على عينات لغوية منها فقط، كما أنَّ الدراسات العربية والإنجليزية التي تقوم بالإحصائيات اللغوية العربية القائمة على مدونات لغوية والتي رصدتها دراستنا لم تركز على تأثير اللهجات العربية على الاستخدام اللغوي العربي في الدول العربية المختلفة، وهو الاختلاف الذي يظهر بوضوح في ظهور بعض المفردات العامية على سبيل المثال في قائمة كلمات وعاء الصحف العربية في "المدونة العربية". كما كانت هناك إحصائيات في بعض الدراسات تشير إلى وجود اختلاف في الاستخدام اللغوي للغة العربية في الوعاء اللغوي نفسه في دول عربية مختلفة، أو ظاهرة تأثير البعد الجغرافي التي لاحظها الدكتور عبد المحسن بن عبيد الثبيتي في "المعالجة الآلية للكلمات المميزة (key words) للمدونات: قضايا تقنية"، وهو عنوان دراسة تقع في الفصل الثالث من كتاب "لغويات المدونة الحاسوبية تطبيقات تحليلية على العربية الطبيعية، ٢٠١٦"، حيث لاحظ فيها أنه "عند مقارنة الكلمات المميزة للصحف المصرية وسعودية ومغربية، كان عدد الكلمات المميزة المشتركة ٥٩٧ كلمة، أي نحو ٦٥٪ من الكلمات المميزة عندما كانت المدونة المرجعية هي الصحافة المصرية. ولكن عند النظر لأكثر مائتي كلمة تميزاً نجد أن عدد الكلمات المميزة المشتركة هي ١١٩ كلمة، أي ما نسبته ٦٠٪ تقريباً، وهذه النسبة أقل بكثير من حالات التشارك التي رأيناها في تجارب حجم المدونة، وهذا المؤشر يعود تأثيره بذلك إلى تأثير المنطقة الجغرافية على الكلمات المميزة". وتلك الظاهرة أوردتها الدكتورة عبد المحسن في أثناء محاولاته للمعالجة الآلية للكلمات المميزة، وعُقب عليها قائلاً: إنَّ "هذا (التأثير الجغرافي) بحاجة لدراسة أعمق ليس هنا مكانها. والخلاصة أنَّ للمنطقة الجغرافية تأثيراً على الكلمات المميزة، فيمكن أن تظهر الاختلافات السياسية والاجتماعية وتتضح الفروق في التفضيلات اللغوية". وقد أشارت دراسات أخرى للظاهرة نفسها، وكانت الإحصائيات اللغوية فيها تشير إلى وجود تأثير

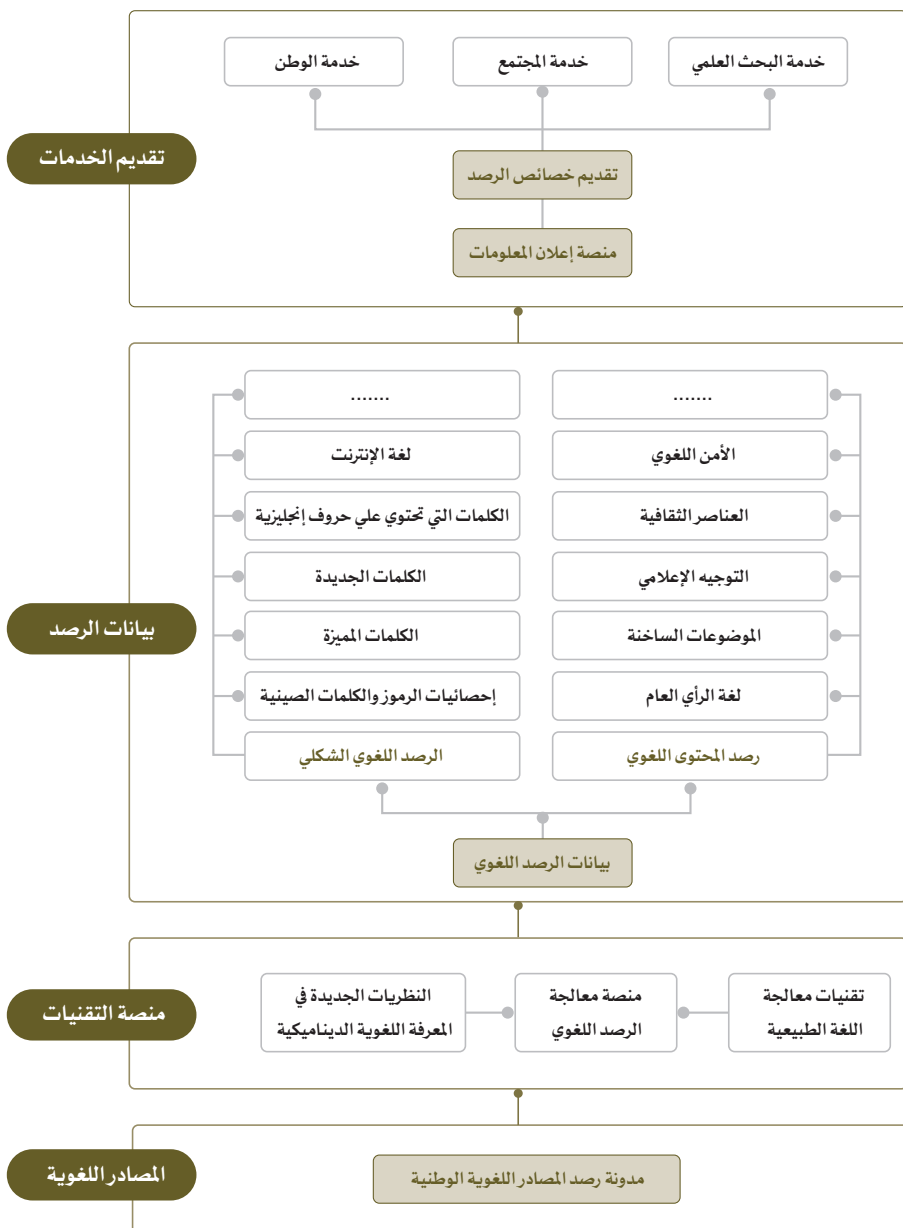
جغرافي بسبب التنوع الجغرافي للبيانات اللُّغوية لتلك المدونات، ولكنها لم تتوقف لتحليل تلك الظاهرة وبيان أسبابها، وربطها باختلاف التركيبات المجتمعية العربية. وبالإضافة إلى ما ذكرناه سابقاً من أهمية البعد الجغرافي والمجتمعي (العامل المكاني) وتأثيره على الاستخدام اللُّغوي، هناك عامل آخر لا يقل عنه أهمية، ألا وهو العامل الزمني، وتأثيره على الاستخدام اللُّغوي، في ضوء إحصائيات قائمة على مدونات لغوية ضخمة ومتنوعة وديناميكية.

وبناء على ما سبق، تهدف الدراسة إلى تحقيق التالي:

١. تصميم مخطط لعملية الرصد اللُّغوي للصحف العربية، وتوضيح كيفية معالجة البيانات، والإحصائيات اللُّغوية المستخرجة منها، وكيفية استخدام تلك الإحصائيات في تقديم الخدمات اللُّغوية المختلفة (الشق النظري لدراستنا).
٢. بناء مدونة لغوية عربية ضخمة وديناميكية للصحف العربية المتداولة تمثل اللهجات العربية الخمس: اللهجة الخليجية، واللهجة العراقية، واللهجة الشامية، واللهجة المصرية، واللهجة المغاربية، والقيام بإحصائيات لغوية تضع في الحسبان عناصر مثل: الزمان والمكان والاستقرار والتغيير اللُّغوي، واستعراض نتائج تلك الإحصائيات.
٣. تصميم مخطط لعملية الرصد اللُّغوي للصحف العربية.

نستعرض في هذا الجزء من دراستنا تصوراً لتصميم مخطط لعملية الرصد اللُّغوي للصحف العربية، وهو التصميم الذي استلهمناه من توصيف العالمتين اللُّغويتين هو مين ويانغ أرهونغ (٢٠١٥) في بحثهما المشترك "عشر سنوات على دراسات الرصد اللُّغوي الصيني" لمخطط آلية الرصد اللُّغوي بأكملها في الصين، وتوصيفهما هذا جاء شاملاً لكل مراكز الرصد اللُّغوي، باختلاف بياناتها اللُّغوية، سواء كانت مراكز لرصد لغة الصحافة الصينية، أو لغة الإنترنت، أو لغة المحتويات التعليمية، أو غيرها من المصادر اللُّغوية، ويتكون توصيفهما من أربعة أجزاء رئيسية هي: المصادر اللُّغوية (Language resource)، وتقنيات معالجة تلك المصادر، والظواهر اللُّغوية المرصودة، وكيفية استخدام بيانات الرصد اللُّغوي في خدمة الوطن والمجتمع والبحث العلمي كما هو موضح في شكل ١.

شكل (١) مخطط لعملية الرصد اللغوي في الصين



وأهم جزء في هذا المخطط - من وجهة نظرنا - هو تحديده لـ "الظواهر اللُّغوية المرصودة"، ففي هذا الجزء، جمعت الباحثان أهم ما ترصده منصات المراكز البحثية الصينية من ظواهر لغوية، حيث قسمتها إلى جزء يرصد الظواهر اللُّغوية، سواء كانت رموز كلمات أو عبارات بأشكالها كافة، وجزء ثانٍ أسمته "رصد المحتوى اللُّغوي"، وهو قائم على تحليل بيانات الظواهر اللُّغوية المرصودة في الجزء الأول، وتحليل مدلولها في لغة الرأي العام والتوجه الإعلامي والثقافي للمجتمع.

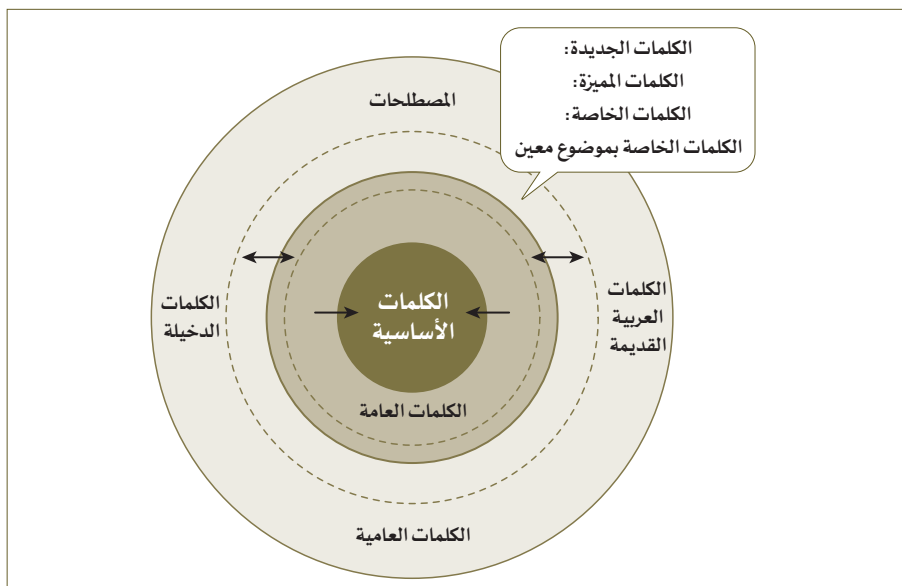
أما بالنسبة إلى "الرصد اللُّغوي الشكلي"، أُرصد وإحصاء الظواهر اللُّغوية المثلة في الرموز والكلمات والعبارات الصينية، التي ذكرتها الباحثان في دراستهما، فقد جاءت نتائجها على إثر جهود ضخمة ومضنية، بدأت منذ عام ٢٠٠٠ وما زالت مستمرة حتى وقتنا هذا. وقد قام بها فريق من باحثي الدكتوراه بقيادة عالم اللُّغويات الصيني تشانغ بو، والذي يعد رائد الرصد اللُّغوي الصيني، فهو وفريقه البحثي، أول من قام ببناء مدونة لغوية ديناميكية لأشهر عشر صحف صينية متداولة في ذلك الوقت (وصلت لأكثر من ١٠٠ صحيفة صينية الآن). وهم أول من بدأ في تحديد الظواهر اللُّغوية التي سيتم رصدها، وماهية تلك الظواهر، وكيفية اكتشافها ورصدها، وأول من عمل على تطوير معادلات وخوارزميات لاكتشاف ورصد تلك البيانات اللُّغوية آلياً. وقد جمع العالم اللُّغوي الصيني تشانغ بوتلك الدراسات التي قام بها هو وفريقه البحثي في كتابه الشهير "دراسات جديدة في المعرفة اللُّغوية الديناميكية" (٢٠٠٧)، والتي بينَ فيها تلك الجهود.

ومن بين محتويات ذلك الكتاب جاء رسم توضيحي للمستويات المختلفة للمفردات، التي تُكوّن لغتنا التي نكتبها ونتحدثها، وهي عبارة عن دائرة، مركزها الكلمات الأساسية، التي تمثل بنية المفردات الأساسية للغة، وهي أساسية لأنه لا غنى عنها في التواصل اللُّغوي، وتأتي الكلمات العامة (General Words) التي تُستخدم في كل التخصصات المعرفية، وفي الفترة الزمنية نفسها في أماكن مختلفة سواء في شمال البلاد وجنوبها، أو شرقها وغربها في حلقة أكبر من المفردات، والحلقتان السابقتان تمثلان الاستقرار اللُّغوي للمفردات في أي لغة،



والتي غالباً لا يطولها التغيير بسبب الزمان والمكان. أمّا بالنسبة للحلقة الخارجية فهي تمثل التغيير اللغوي، إذ تتغير بتغير الزمان أو المكان أو كليهما، وتحتوي على الكلمات الجديدة (New Words) التي تظهر في اللغة في فترة زمنية ما. والكلمات المميزة لفترة زمنية معينة (Distinctive words)، والتي تعكس أهم الأحداث السياسية والاقتصادية والرياضية المحلية أو الدولية خلال فترة زمنية معينة. والكلمات الخاصة (Unique Words) التي تعبر عن خصوصية الاستخدام اللغوي للكلمة في مجتمع ما. والكلمات الخاصة بموضوع معين، والتي تصاحب الحديث عن موضوع معين. وبين هذه الحلقة وحلقة الكلمات العامة، هناك نوعان من الكلمات ينتقل بين الحلقتين، وهي الكلمات القديمة للغة، والكلمات الدخيلة، وهما النوعان اللذان يتأثران بتغير الزمان والمكان. أما الحلقة الأخيرة فهي الحلقة التي يحدث فيها التغير بتغير مكان الاستخدام اللغوي والذي تمثله الكلمات العامة، أو التغير في استخدام المفردات الذي يصاحب اختلاف المجالات الذي تمثله المصطلحات، انظر شكل ٢.

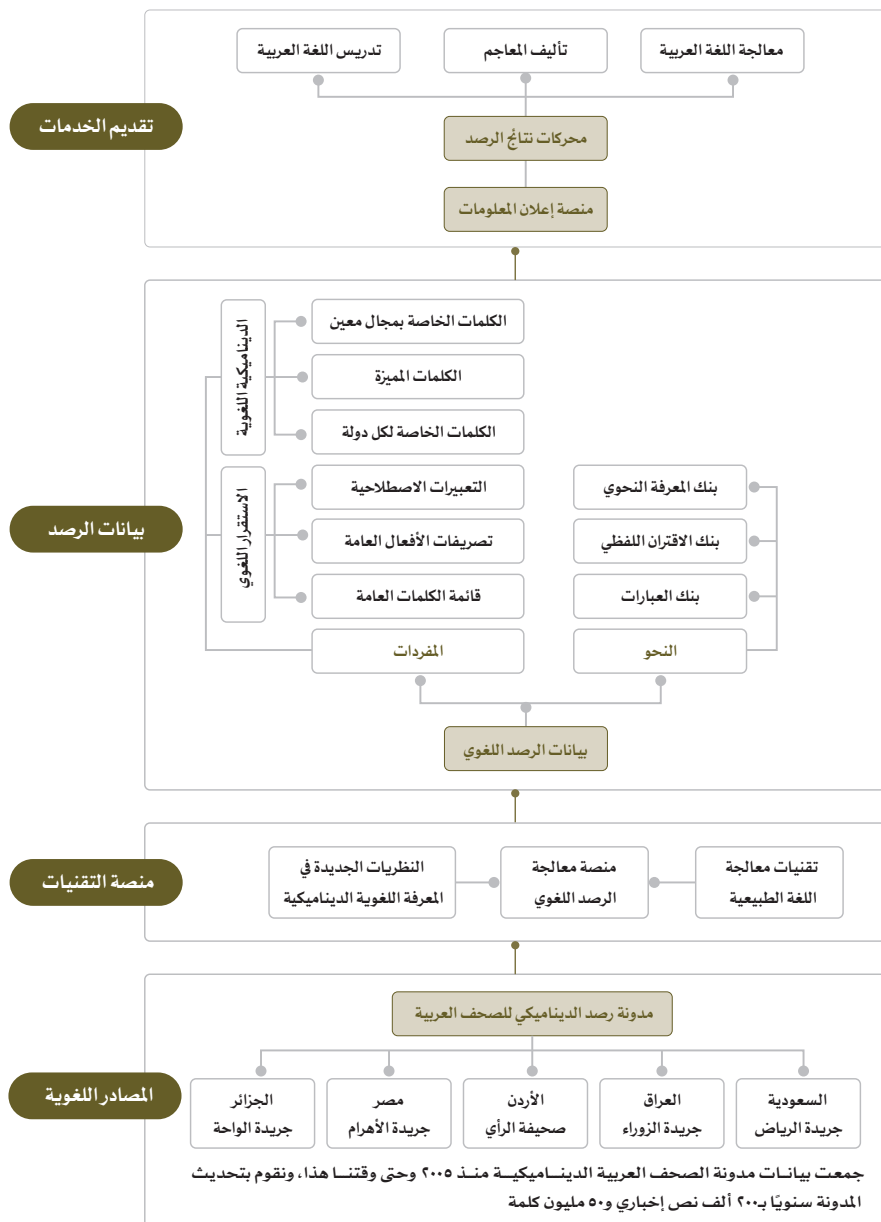
شكل (٢) المستويات المفرداتية المختلفة للغة العربية في ضوء كتاب "دراسات جديدة في المعرفة اللغوية الديناميكية"



وبناء على ما سبق، قمنا في دراستنا بتصميم مخطط لعملية الرصد اللُّغوي للصحف العربية مسترشدين بالمخطط الصيني للرصد اللُّغوي في الصين. ومخططنا يتكون من أربعة أجزاء رئيسية مماثلة للمخطط الصيني، ولكنها تختلف عنه في شكل الظواهر اللُّغوية التي قمنا برصدها، حيث قسّمنا تلك الظواهر تقسيماً لغوياً قائماً على مستويين، هما: المستوى المعجمي، والمستوى النحوي. والمستوى المعجمي ينقسم إلى جزأين، يستهدف الأول منهما رصد الاستقرار اللُّغوي ممثلاً في رصد الكلمات العامة وأشكال الأفعال الصرفية في الصحف العربية، وأهم التعبيرات الاصطلاحية أو المسكوكة التي تستخدم في تلك الصحف. أمّا الآخر ديناميكي متغير، يتمثل في الكلمات الخاصة التي تظهر في صحيفة كل دولة في المدونة اللُّغوية الديناميكية، والكلمات التي ظهرت في فترة زمنية محددة في صحيفة كل دولة، والكلمات المميزة لمجال معين (الاقتصاد، السياسة). أمّا المستوى النحوي فيهدف إلى رصد الظواهر النحوية للتركيب والتصاحب اللفظي (الاقتران) وصولاً لبناء بنك معرفة نحوي. يأتي هذا في تصور مبدئي عززنا في بعض مواضعه نتائج الإحصاء اللُّغوي بتحليل لأسبابه ومدلولاته في ضوء ما استخلصناه من المدونة الديناميكية للصحف العربية المتداولة، في خطوة لتحليل أوسع في المستقبل لأوجه التشابه والاختلاف في الاستخدام اللُّغوي لتلك الظواهر اللُّغوية في اللهجات العربية المختلفة.

أمّا على صعيد استخدام نتائج الرصد اللُّغوي، فقد كان تركيزنا المبدئي أن تكون نتائج هذه الإحصائيات موجهة لخدمة معالجة اللُّغة العربية، وتأليف المعاجم العربية المزودة بإحصائيات توضح الفروق في استخدام الكلمات عند متحدثي اللُّغة العربية بصورها المختلفة في الدول العربية، وتكون معيناً في عملية تأليف المناهج التعليمية سواء للطلاب العرب أو الأجانب، انظر شكل ٣.

شكل (٣) مخطط الرصد اللغوي للصحف العربية



٢- المدونة الديناميكية للصحف العربية المتداولة ونتائج الرصد اللُّغوي

٢-١- بناء مدونة ديناميكية للصحف العربية المتداولة

قمنا في دراستنا ببناء مدونة ديناميكية للصحف العربية المتداولة طبقاً لمخطط الرصد اللُّغوي للصحف العربية الذي قمنا بتصميمه، ووضعنا في الاعتبار أثناء اختيار الصحف أن تكون صحفاً رسمية ومتداولة وتحظى بشعبية كبيرة في بلادها، فوقع اختيارنا على جريدة الرياض السعودية (اللهجة الخليجية)^(١)، وجريدة الزوراء العراقية (اللهجة العراقية)^(٢)، وصحيفة الرأي الأردنية (اللهجة الشامية)^(٣)، وجريدة الأهرام المصرية (اللهجة المصرية)^(٤)، وجريدة الواحة الجزائرية (اللهجة المغاربية)^(٥). ويبلغ حجم المدونة ١,٦ مليون نص إخباري، وتحتوي على أكثر من ٤٠٠ مليون كلمة. وتغطي البيانات الفترة من ٢٠٠٥ وحتى عام ٢٠١٨م^(٦)، وقد قمنا بحفظ البيانات المستخدمة في المشروع في قاعدة بيانات (MongoDB)^(٧)، كما هو موضح في شكل ٤.

(1) <http://www.alriyadh.com/>

(2) <http://alzawraapaper.com/>

(3) <http://alrai.com/>

(4) <http://www.ahram.org.eg/>

(5) <http://elwaha-dz.com/ar/>

(٦) يبلغ حجم المدونة الديناميكية للصحف العربية المتداولة مليار كلمة و٤ مليون نص إخباري، وتغطي الفترة من ٢٠٠٥ وحتى فبراير ٢٠٢٤، ولكن اخترنا عرض جزء من البيانات حتى ٢٠١٨، لأنها الفترة التي قمنا فيها بإحصاء الاستقرار اللُّغوي والتغيير الديناميكي فيها، وجارٍ تحديث هذه الإحصائيات.

(٧) قام الباحثون بالحصول على البيانات السابقة لأغراض علمية وبخفية، وبما لا يخالف قوانين حقوق الملكية الفكرية لجمهورية الصين الشعبية.

شكل (٤) طريقة حفظ البيانات في قاعدة بيانات (MongoDB)

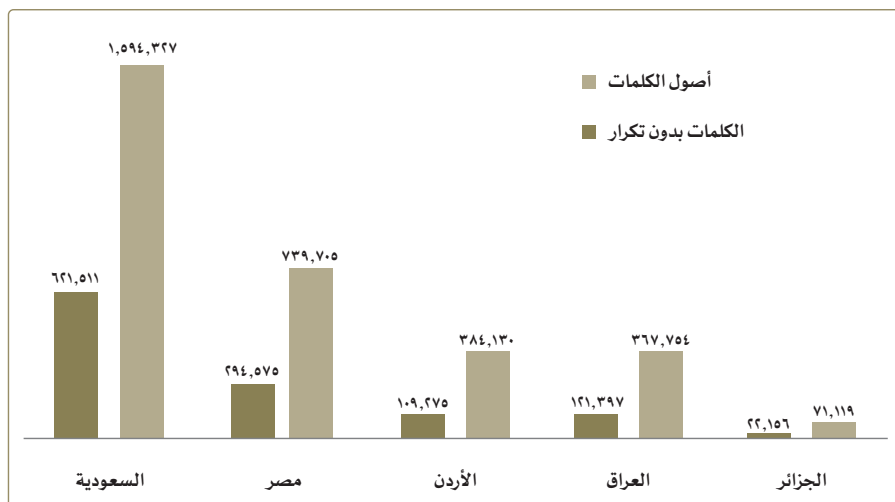
```
{'_id': ObjectId('5c276c5d764cc073dc34b2e9'), ←
  'URL': 'http://www.ahram.org.eg/NewsQ/440734.aspx', ←
  'TIME': '2015-10-4', ←
  'TITLE': 'قنديل رئيساً لنادى هيئة النيابة الإدارية', ←
  'DOMAIN': 'مصر', ←
  'TEXT': 'أعلنت اللجنة المشرفة على انتخابات نادى مستشارى هيئة النيابة الإدارية، برئاسة المستشار سلطان السلموني،
  نتيجة انتخابات التجديد الكلى لرئاسة وعضوية مجلس إدارة النادي، والتي أسفرت عن فوز المستشار عبد الله قنديل نائب رئيس
  الهيئة، بمنصب رئيس النادي لدورة جديدة، وذلك بحصوله على 967 صوتاً' ←
```

وقبل أن نقوم بحفظ البيانات اللغوية في قاعدة البيانات، قمنا بإجراء الخطوات التالية:

١. حذف جميع علامات التشكيل، وعلامات (الكشيدة) المضافة لتطويل الحروف لأسباب جمالية.
٢. حذف علامة الترقيم المكررة، مثل علامات الاقتباس والمسافات.
٣. تغيير التاريخ من الشكل الكتابي للتقويم العربي الهجري إلى التنسيق الرقمي (YYYY-MM-DD).
٤. تجنب استخدام اللغة العربية في روابط الأخبار.
٥. حذف التعليقات والروابط الواردة في بعض مواقع الويب بعد الأخبار وليست من نصها، مثل عبارة "التعليقات هي كما يلي: " وغيرها من العبارات المشابهة.
٦. حذف الأخبار التي تحتوي على صور فقط وليس فيها نصوص كتابية.
٧. حذف عناوين الصور في الأخبار.

ولتجنب وجود أخطاء إملائية قمنا بتسوية الاحتمالات الهجائية (Normalization)، والتي من شأنها تقليل الأخطاء الإملائية في مدونتنا اللغوية، وأيضاً حذف علامات الترقيم غير العربية، وتحويل الأرقام الهندية (٣٢١) إلى أرقام عربية (٣٢١). وللتقطيع الصرقي، والحصول على أصول الكلمات (Lemmatization) استخدمنا محلل فُراسة الذي قام بتطويره فريق بحثي من معهد قطر لبحوث الحوسبة التابع لجامعة حمد بن خليفة^(١). يظهر أدناه العدد الإجمالي للكلمات العربية وأصول الكلمات لكل صحيفة.

(1) <https://farasa.qcri.org/#contact>



٣- نتائج الرصد اللغوي

٣-١- الكلمات العامة للصحف الخمس

قمنا في دراسة سابقة بعنوان "استخراج الكلمات العربية العامة للصحف واستخداماتها" والتي نشرت في مجلة "دراسات الشرق الأوسط في جامعة بكين" (٢٠٢٥) باللغة الصينية - باستخدام معادلات استخراج الكلمات العامة للغة الصينية آلياً للحصول على الكلمات العامة للصحف العربية في مدونتنا اللغوية، والمعادلة تضع في الحسبان أن تكون الكلمة شائعة الاستخدام في الصحف العربية الخمس في فترة زمنية معينة، وتستبعد الكلمات التي قد تكون شائعة الاستخدام في صحيفة ما دون صحيفة أخرى. بعد ذلك قمنا بحساب درجة عمومية الكلمات، ووجدنا أن ٧٧٤٤ كلمة عامة تغطي نسبة ٩٥٪ من كلمات المدونة الديناميكية للصحف العربية المتداولة، وقمنا بعمل قائمة بكل كلمة وأشكال كتابتها المختلفة في المدونة، انظر جدول ١. وقمنا بمقارنة قائمة الكلمات بقائمة كلمات (المدونة العربية) لوعاء الصحف، ووجدنا أن أغلب كلمات قائمة وعاء الصحف (المدونة العربية) موجودة في قائمة الكلمات العامة للصحف العربية، وكانت الكلمات التي لم تظهر في قائمة



كلماتنا وعددها ٦٦٥ عبارة عن أسماء أعلام، وكلمات عامية ودخيلة. وفي نهاية ورقتنا البحثية، قمنا بترشيح بعض من الكلمات العامة لتدخل المعجم العربي الصيني لجامعة بكين، ووضحنا أسباب اختيار تلك الكلمات، وقواعده.

جدول (١) الكلمات العامة وأشكال كتابتها في المدونة العربية

الترتيب الكلمة	أصل الكلمة	عدد الأشكال	الشكل الأول	الشكل الثاني	الشكل الثالث	الشكل الرابع	الشكل الخامس
١	في	٢	في	في			
٢	من	١	من				
٣	ان	٤	آن	أن	ان	إن	
٤	علي	٢	علي	على			
٥	الي	٥	الي	آلي	آلي	الي	إلى
٦	الذي	١	الذي				
٧	هذا	١					
٨	كان	٢					
٩	عن	١					
١٠	مع	١					
١١	ما	١					
١٢	عام	١					
١٣	ذلك	١					
١٤	لا	١					
١٥	عمل	١					

٣-٢- أشكال الأفعال الصرفية في الصحف الخمس

بعدما قمنا باستخلاص قائمة الكلمات العامة للصحف العربية، قمنا باستخراج الأفعال الموجودة بها، والبحث عن تصريفات تلك الأفعال في الصحف العربية، في خطوة مهمة، لتسهيل تدريس الأفعال العربية لغير الناطقين بها، خاصة الذين لا توجد في لغتهم الأم بنية صرفية مثل اللغة العربية، انظر جدول ٢.

جدول (٢) الأفعال العامة وأشكالها الصرفية في مدونة الصحف العربية

مسلسل	الفعل	عدد أشكاله	الشكل الأول	الشكل الثاني	الشكل الثالث	الشكل الرابع	الشكل الخامس
١	كان	٥	كان	كانت	يكون	تكون	يكن
٢	عمل	٥	تعمل	يعمل	عملت	يعملون	نعمل
٣	قال	٤	قال	قالت	يقول	تقول	
٤	بين	٢	بينت	يبين			
٥	تم	٢	تم	يتم			
٦	أول	٤	أولت	أُوول	تأول	يُوول	
٧	غير	٤	تغير	يغير	غيرت	نغير	
٨	قبل	٥	يقبل	تقبل	قبل	نقبل	قبلت
٩	عدد	٣	عدد	عددت	أعدد		
١٠	أكد	٤	أكد	أكدت	يؤكد	تأكد	
١١	شكل	٤	تشكل	يشكل	شكلت	شكل	
١٢	قدم	٥	قدم	يقدم	تقدم	قدمت	قدموا
١٣	أكثر	٣	تكثر	يكثر	أكثر		



الشكل الخامس	الشكل الرابع	الشكل الثالث	الشكل الثاني	الشكل الأول	عدد أشكاله	الفاعل	مسلسل
		يمثلون	تمثل	يمثل	٣	مثل	١٤
قاموا	قامت	قام	يقوم	تقوم	٥	قام	١٥

٣-٣- الخاصة لكل صحيفة من الصحف الخمس

كما وضحنا سابقاً، فإن كل مجتمع يختلف في استخدامه اللغوي للمفردات والتراكيب عن المجتمعات الأخرى، وإن جمعتهم لغة واحدة مشتركة، وهو ما يُعرف بالتأثير الجغرافي في الاستخدام اللغوي. وفي هذا الجزء من دراستنا قمنا باستخلاص الكلمات الخاصة بكل دولة من الدول الخمس، وذلك وفقاً لثلاثة مستويات من التكرار: الكلمات الأكثر تكراراً، والكلمات المتوسطة التكرار، والكلمات الأقل تكراراً في الصحف العربية الخمس. ثم قمنا بتحليل الفروقات في استخدام هذه الكلمات، وتفسير الأسباب التي جعلتها "كلمات خاصة". ويُعرض في الجدول أدناه بعض هذه المفردات. وقد تبين - كما في الجدول الآتي - أن الفروقات الأكثر وضوحاً في الكلمات الخاصة بين الدول الخمس تظهر في استخدام أسماء الأعلام. كما تبين أن أهم العوامل المؤثرة في اختلاف استخدام هذه الكلمات الخاصة هي الأحداث الاجتماعية البارزة، والتركيبية السكانية، والبيئة الجغرافية، انظر جدول ٣.

جدول (٣) قائمة الكلمات الخاصة للصحف العربية الخمسة

السعودية		مصر		الأردن		الجزائر		العراق	
الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل
سعودية	١٨	مصر	١٦	أردني	١٦٤	غرداية	٢١	عراقي	٥٦
سعودي	١١١	قاهرة	٤٢	أردن	١٢٣	جزائر	٦٨	بغداد	٦٥
رياض	٢٦٩	أهلي	٥٢	عمان	٨١	أوربي	٢٤	عراق	٦٤

السعودية		مصر		الأردن		الجزائر		العراق	
مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة	مسلسل	الكلمة الخاصة
٢٥٨	معللة	٣٠٥	هرم	٦٦	دينار	١٦٥	بريطاني	١٧٥	موصل
١٩٤	ريال	٩٦	جنيه	٢١٧	حسين	٢١٥	روسيا	٢٥٥	قضاء
٣٢	عبد العزيز	٢٩٣	رئاسة	١٥٢	جلالة	١٣٥	روسي	٢٤٣	إقليم
١٣٤	سعود	٢٨٢	سيسي	٢١٥	إسرائيلي	٢٤٧	فرنسا	٧٥	داعش
٨٦	سلمان	٢٧٦	كام	٣٠٧	سوريا	٢٠٧	واشنطن	٢٠٣	تحرير
٢٠٦	فيصل	٣٠٨	محمود	١٧٥	قدس	٨٠	هاتف	١٣٢	نفط
٢٠٣	خالد	٣٢١	حسن	١٩٠	بلدية	١٠١	مستخدم	٣١٧	فنان

٣-٤- الكلمات المميزة للصحف الخمس

قمنا في دراسة سابقة بعنوان "الكلمات المميزة لصحيفة الرياض السعودية.. دراسة قائمة على مدونة لغوية ديناميكية" (٢٠٢٤)، والمنشورة في مجلة "الدراسات اللغوية" بتطوير منهجية آلية لرصد الكلمات المميزة في فترة زمنية معينة، وتحليل تلك الكلمات وعلاقتها بالمجتمع وتطوره، وفي جدول ٤ بعض من الكلمات المميزة التي ظهرت من عام ٢٠٠٥ وحتى عام ٢٠٠٨، مثل "بلدية، انتخاب" والتي تشير إلى الانتخابات البلدية السعودية التي حدثت في عام ٢٠٠٥، وهي أول انتخابات جرت في المملكة العربية السعودية منذ اعتماد نظام المجالس البلدية. كما ظهر اسم "رفيق الحريري" في عام ٢٠٠٥ بسبب حادثة اغتياله في ذلك العام، واسم صدام حسين في ٢٠٠٦ بسبب تنفيذ حكم الإعدام فيه. كما أنَّ ظهور كلمات مثل: "حرب، لبنان، قصف" تشير إلى العدوان الإسرائيلي الغاشم على لبنان في عام ٢٠٠٦. وظهور كلمات مثل "انتخابات، أمريكي" تشير إلى الانتخابات الأمريكية، و"بترو، أزمة، اقتصادي" تشير إلى الأزمة الاقتصادية التي شهدتها العالم عام ٢٠٠٨. وقد قمنا بتقديم شرح مفصل لدلالات تلك الكلمات في ورقتنا البحثية.



جدول (٤) مجموعة الكلمات المميزة لجريدة الرياض في الفترة بين ٢٠٠٥ - ٢٠٠٨

العام	مجموعة الكلمات المميزة
٢٠٠٥	بلدية، انتخاب، عراق، دستور، مكافحة، إرهاب، حريري، تضامن، لعبة، فهد، قمة، مواجهة، عضوية
٢٠٠٦	فلسطين، حماس، بورصة، انهيار، صدام، إسرائيل، لبنان، حرب، قصف، جازان، قدم، كرة، كأس
٢٠٠٧	قمة
٢٠٠٨	بترو، ارتفاع اقتصادي، أزمة، غرة، سابك، جازان، مهرجان، انتخاب، أمريكي

٣-٥- الكلمات المتعلقة بموضوع معين

من الظواهر اللغوية الهامة التي ركزنا عليها في دراستنا هو استخراج الكلمات المميزة عند الحديث عن موضوع معين، فقمنا باستخراج مقالات صحيفة الأهرام التي تغطي "قضية سد النهضة"، في الفترة بين عامي ٢٠١٦ و٢٠٢٤، وقمنا بإحصاء الكلمات الشائعة في تلك المقالات، ومقارنتها بقائمة الكلمات العامة للصحف العربية، وحساب الفرق بين درجة عمومية الكلمة، وعدد مرات تكرار الكلمة؛ للحصول على الكلمات المتعلقة بموضوع "سد النهضة" ألياً، كما هو موضح في جدول ٥.

جدول (٥) الكلمات المتعلقة بأخبار "سد النهضة" في صحيفة الأهرام المصرية

الفرق	ترتيب الكلمة في قائمة الكلمات العامة	الكلمة	الترتيب
٨٠٤٧	٨٢٠٥	تعنت	١٥٨
٥٧٤٢	٥٨٤٣	أحادي	١٠١
٤٨٢٨	٤٨٦١	ملء	٣٣
٤١٢٣	٤٢١٥	ملزم	٩٢

٢٢	ري	٢٧٨٢	٢٧٦٠
١١١	تفاوض	٢٧٠٢	٢٥٩١
٢٥	مائي	١٩٦٧	١٩٤٠
١٩٧	نية	١٩٤٢	١٧٤٥
١٣	نهضة	١٦٨٨	١٦٧٥
٦	سد	١٣٣٢	١٣٢٦
١٧٠	ضرر	١٤٩٦	١٣٢٦
٥٠	مفاوضة	١٢٨٤	١٢٣٤
٧٥	توصل	١١٥٥	١٠٨٠

٤- الخاتمة

في ورقتنا البحثية استرشدنا بمخطط الرصد اللُّغوي الصيني، وقمنا بتصميم مخطط رصد لغوي للصحف العربية، وإنشاء مدونة لغوية ديناميكية للصحف العربية المتداولة في الفترة من ٢٠٠٥ وحتى ٢٠٢٤. كما قمنا باستخراج الظواهر اللُّغوية المختلفة من تلك المدونة في الفترة من ٢٠٠٥ وحتى ٢٠١٨. وسنستمر في تحسين تخطيط الرصد للصحف العربية، وسنستكمل الإحصائيات في بقية المدونة من عام ٢٠١٨ وحتى وقتنا الحالي، وفي الوقت نفسه سنقوم بمقارنة الطرق الإحصائية المختلفة، للتأكد من فاعلية الطرق الإحصائية التي استخدمناها أو طورناها. كما سنقوم بمقارنة قوائم الكلمات التي حصلنا عليها بالمعاجم العربية المتاحة، من أجل بناء معجم عربي عصري يوفر لستخداميه الإحصائيات المتعلقة، ويوضح الاستخدامات المختلفة للمضردات والتراكيب العربية.



المراجع العربية

- الدكروري، أيمن. (٢٠١٨). المدونات اللغوية ودورها في معالجة النصوص العربية. مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.
- العصيمي، صالح بن فهد. (٢٠١٨). لسانيات المتون قضايا أساسية في التأصيل والتطبيق والمنهج. مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.
- العصيمي، صالح بن فهد. (٢٠١٩). نحو قائمة للألفاظ الأكاديمية في العربية المعاصرة: دراسة قائمة على المدونات. مجلة علوم الشرعية واللغة العربية، ٤ (١)، ٤٠٣-٤٣١.
- المجيول، سلطان بن ناصر. (٢٠١٦). لغويات المدونة الحاسوبية تطبيقات تحليلية على العربية الطبيعية. مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.
- صالح، محمود إسماعيل. (٢٠١٥). المدونات اللغوية العربية بناؤها وطرائق الاستفادة منها. مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية.

رومنة المراجع العربية

- aldikruri, 'ayman. (2018). al mudawanat allughawiat wadawruha fi muealajat al nusūs al arabīati. markaz al malik eabdallah bin eabdialeaziz alduwlii likhidmat allughat al arabīati.
- aleusaymi, salih bin fihd. (2018). lisaniaat almutun qadaya 'asasiat fi altaasil waltatbiq walmanhaji. markaz al malik eabdallah bin eabdialeaziz alduwlii likhidmat allughat al arabīati.
- aleusaymi, salih bin fihd. (2019). nahw qayimat lil'alfaz al'akadimiati fi al arabīati al mueasirati: dirasat qayimat ealaa al mudawinati. majalat eulum al shareiat wallughat al arabīati, 4 (1), 403-431.
- almijyuli, sultan bin nasir. (2016). lughawyaat al mudawanat al hasubiat tatbiqat tahliliat ealaa al arabīati al tabīciati. markaz al malik eabdallah bin eabdialeaziz alduwlii likhidmat allughat al arabīati.

saliha, mahmud 'iismaeil. (2015). almudawanat allughawiat alearabiat binawuha watarayiq al'iifadat minha. markaz almalik eabdallah bin eabdialeaziz alduwlii likhidmat allughat alearabiati.

المراجع الأجنبية

Abdelali. A., Darwish. K., Durrani. N., Mubarak, H. (2015 June 12-17). *Farasa: A fast and furious segmenter for Arabic*. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, California, USA.

Akef, Alaa Mamdouh., 王莹莹, 杨尔弘. (2024). 基于动态语料库的阿拉伯语报纸年度特色词探究——以沙特阿拉伯《利雅得报》为例. *语言学研究*, 1, 166-180.

Akef, Alaa Mamdouh., 王莹莹, 付志明, 杨尔弘. (2025). 阿拉伯语通用词表的研制及教学应用研究. *北大中东研究*, 1.

侯敏, 杨尔弘. (2015). 中国语言监测研究十年. *语言文字应用*, 3, 12-21.

Habash, Nizar Y. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool.

Pimentel, J. (2001). Sociolinguistic reflections of privatization and globalization: The Arabic of Egyptian newspaper advertisements. Michigan university.

张普. (2009). *动态语言知识更新研究*. 商务印书馆.



نبذة تعريفية عن الباحث

Alaa Mamdouh Akef is an Assistant Professor affiliated with the Department of Arabic Language, School of Foreign Languages, Peking University. His research interests focus on Artificial Intelligence-assisted Language Learning. He has published papers in Chinese and English academic journals and has contributed to the translation of several major works from Chinese into Arabic, including The Encyclopaedic Dictionary of Chinese Culture and A History of Sino-Omani Relations. He has received multiple awards from both China and Arab countries in the fields of education and translation. He is also the founder of the research team “SIRAJ” and is leading the project titled Chinese-Arabic Multi-Resources Parallel Database (CAMP Database).	علاء ممدوح عاكف، أستاذ مساعد، يعمل في قسم اللغة العربية بكلية اللغات الأجنبية بجامعة بكين. تتركز اهتماماته البحثية على تعلم اللغات بمساعدة الذكاء الاصطناعي. له أبحاث منشورة في مجلات محكمة باللغات الصينية والإنجليزية، وشارك في ترجمة العديد من المؤلفات من الصينية للعربية، من بينها "المعجم الثقافي الصيني الموسوعي" و"تاريخ العلاقات بين الصيني وعمان". حاصل على جوائز مختلفة من الصين والدول العربية في مجال التعليم والترجمة. ومؤسس فريق "سراج". من مشاريعه البحثية قاعدة البيانات المتوازية الصينية العربية متعددة المجالات (Chinese Arabic Multi-Resources Parallel Database, CAMP Database).
معرف أوركيد (ORCID) : 0009-0006-9072-7597	
Email: alaa_eldin_akef@pku.edu.cn	

توظيف تضمينات الكلمات المُدرّبة مسبقاً في توسيم الأجزاء النحوية لبيانات التناوب اللغوي^(١)

المعرف الرقمي: 004-003-002-10.60161/doi.org/https://

فهد الغامدي^{أ*}، منى دياب^ب

أ قسم علوم الحاسبات، كلية الحاسبات والمعلومات، جامعة الباحة، المملكة العربية السعودية
ب معهد تقنيات اللغة، كلية هاينز، جامعة كارنيجي ميلون، بيتسبرغ، الولايات المتحدة الأمريكية

استقبل في	رُوجع في	قُبِل في	نُشر في
١٤٤٦/٠٨/٠٢	٢٠٢٥/٠٨/١١	١٤٤٦/٠٩/٢٠	١٤٤٦/١٠/٢٨

المُلخَص

التناوب اللغوي ظاهرة يتنقل فيها متحدثون متعددون اللغات بين لغات أو لهجات متعددة ضمن محادثة واحدة. تبحث هذه الدراسة تحديات وسم الأجزاء النحوية في البيانات المتضمنة لتناوب اللغات داخل الجملة الواحدة، عبر اختبار عدة بنى عصبية وتضمينات لغوية مسبقة التدريب لأربعة أزواج لغوية: الإسبانية-الإنجليزية، الهندية-الإنجليزية، العربية الفصحى-المصرية، العربية الفصحى-الشامية. أظهرت النتائج أن التضمينات متعددة اللغات ترفع دقة الوسم للأزواج اللغوية المتقاربة (الفصحى-المصرية، الفصحى-الشامية) وتُضعفها في الأزواج المتباعدة (الإسبانية-الإنجليزية، الهندية-الإنجليزية)، بينما حسّن نموذج متعدد المهام (وسم نحوي-تحديد هوية اللغة) الأداء للأزواج المتباعدة، متجاوزاً النماذج السابقة ومبرزاً أهمية اختيار التضمينات الملائمة..

الكلمات المفتاحية

التناوب اللغوي، توسيم الأجزاء النحوية، التضمينات، نماذج متعددة المهام.

(١) هذا البحث ترجمة عربية لبحث منشور باللغة الإنجليزية للباحث نفسه. وإثراء للمحتوى العلمي العربي، تقبل المجلة نشر الأبحاث بعد ترجمتها، وفق شروط خاصة

*الباحث المراسل: فهد الغامدي (البريد الإلكتروني: fghamdi@bu.edu.sa) (Email:).

Leveraging Pretrained Word Embeddings for Part-of-Speech Tagging of Code-Switching Data

المعرف الرقمي: <https://doi.org/10.60161/2521-002-003-004>

Fahad AlGhamdi^{A*} Mona T. Diab^B

A Department of Computer Science, College of Computing and Information Technology, Al-Baha University, Saudi Arabia.

B Language Technologies Institute, School of Computer Science, Carnegie Mellon University, United States

Received on	Revised on	Accepted on	Published on
01/02/2025	10/02/2025	20/03/2025	26/04/2025

ABS TRACT

Linguistic Code Switching (CS) occurs when multilingual speakers alternate between languages or dialects within a single conversation. This study investigates challenges in Part-of-Speech tagging for intra-sentential CS data, evaluating neural architectures and diverse pre-trained embeddings across four language pairs: SPA-ENG, HIN-ENG, MSA-EGY, and MSA-LEV. Results indicate merged multilingual embeddings improve accuracy for related pairs (MSA-EGY, MSA-LEV) but reduce performance for distant pairs (SPA-ENG, HIN-ENG). Incorporating a multitask POS-language-identification model resolves performance issues for distant pairs. The proposed methods surpass previous state-of-the-art tagging for MSA-EGY, highlighting the importance of embedding choice in CS scenarios..

KEYWORDS

Code-Switching, Embeddings, POS-Tagging, Multitask Model.

*Corresponding author: Fahad AlGhamdi (Email: fghamdi@bu.edu.sa.com).

١- المقدمة

يُعَدُّ التناوب اللغوي (Code Switching) ظاهرةً لغوية شائعة، يتمّ خلالها استخدام لغتين أو أكثر، أو لهجاتٍ متعددة بالتناوب في الكلام أو الكتابة. ويظهر هذا التناوب عادةً على مستويات عدّة من البنية اللغوية: فإمّا أن يكون على مستوى الجُمْل (inter-sentential)، أو ضمن العبارة نفسها بمزج لغتين أو أكثر في السياق ذاته (intrasentential)، أو حتى على مستوى الكلمات أو الصيغ الصرفية. وتبرز ظاهرة التناوب اللغوي بشكل لافت في البلدان التي تضم مجموعات كبيرة من المهاجرين، مما ينجّم عنه انتشار الثنائية اللغوية. وفي العادة، يتقن الأشخاص الذين يمارسون التناوب اللغوي لغتين (أو أكثر): لغة أولى مشتركة (lang1)، ولغة ثانية سائدة (lang2). وقد تكون هاتان اللغتان مختلفتين تماماً كالماندريين والإنجليزية أو الهندية والإنجليزية، أو قد تكونان مجرد تنويعات من اللغة نفسها، كما في حالة اللغة العربية الفصحى الحديثة (Modern Standard Arabic, MSA) واللهجات الإقليمية العربية (مثل اللهجة المصرية-EGY). ومع أنّ التناوب اللغوي ينتشر تقليدياً في الأوساط الشفوية، فإنّ الانتشار الواسع للإنترنت ووسائل التواصل الاجتماعي جعله شائعاً أيضاً في النصوص المكتوبة وأنواعها (Vyas et al., 2014a; Danet and Herring, 2007; C'ardenas-Claros and Isharyanti, 2009). وقد أفضى هذا الواقع الجديد إلى تدفّق كبير وغير مألوف من النصوص المكتوبة المتضمّنة لظاهرة التناوب اللغوي على شبكة الإنترنت. ورغم ما يتيح ذلك من فرص، فإنّه يفرض تحدياتٍ جدّية على مهام متعدّدة في مجال معالجة اللغات الطبيعية (NLP)، إذ إنّ الأساليب التقليدية التي دُرِّبَت للتعامل مع لغة واحدة تميل إلى الانهيار حين تصادف نصوصاً تتضمّن لغتين أو أكثر. كما يتراجع أداء النماذج المُستخدمة حالياً في تقنيات يُفترض أن تقدّم نتائج جيّدة مثل Part-of-Speech Tagging (POS) بمعدّل يتناسب مع مقدار المزج اللغوي ومستواه، ويعود السبب في ذلك إلى الكلمات غير المعروفة (-out-of-vocabulary) في إحدى اللغتين، أو إلى البنى النحوية الهجينة الجديدة، أو إلى الكلمات المشتركة أو الغامضة التي قد توجد في كلا المعجمين.



تُعَدُّ مهمّة (Part-of-Speech Tagging (POS) مكوناً أساسياً في أيّ نظام لفهم اللغة الطبيعية، وغالباً ما تشكّل إحدى أوائل المهام التي يجري توظيفها عند معالجة البيانات؛ فهي تقنية تمهيدية توفر مخرجات ضرورية لأدوات عالية المستوى في مجال المعالجة اللغوية مثل نماذج تقسيم العبارات (chunkers) والمحلّلات النحوية والدلالية وتحليل الخطاب. وتستخدم هذه الأدوات في تطبيقات متنوعة تشمل تحليل المشاعر والرأي، وتلخيص النصوص، واستخراج المعلومات، والتعرّف التلقائي على الكلام، والترجمة الآلية وغيرها. ومن هنا تبرز أهمية أن تكون مؤسّمات (taggers) الأجزاء النحوية قادرة على معالجة البيانات النصية المتضمّنة للتناوب اللغوي.

في هذه الورقة، نركّز على مشكلة (Part-of-Speech Tagging (POS في النصوص المتضمّنة للتناوب اللغوي على مستوى العبارات (intrasentential) ولأزواج لغوية متعدّدة. نستكشف أثر استخدام تشكيلات مختلفة من نماذج التضمينات (embeddings) فضلاً عن بنى متنوعة للشبكات العصبية، سعياً للحدّ من مشكلة ندرة البيانات المصنّفة سلفاً الخاصة بالتناوب اللغوي. نقترح مجموعة من الأساليب في تمثيل الكلمات (word embedding) قد تسهم في تحسين أداء مهمة POS في البيانات المتضمّنة للتناوب اللغوي.

ولتحقيق قابلية التعميم على مختلف أزواج اللغات، نجري دراستنا على بيانات أربع مجموعات خاصّة بالتناوب اللغوي، تشمل أربعة أزواج لغوية: اللغة العربية الفصحى الحديثة مع اللهجة المصرية (MSA-EGY)، والعربية الفصحى الحديثة مع اللهجة الشامية (MSA-LEV)، والإسبانية مع الإنجليزية (SPA-ENG)، والهندية مع الإنجليزية (-HIN-ENG). ونستخدم المجموعة نفسها من العلامات النحوية في جميع الأزواج اللغوية، وهي Universal POS tag set (Petrov et al., 2011). وتتلخص إسهاماتنا فيما يأتي:

أ. نستخدم شبكة عصبية متقدمة ثنائية الاتجاه (bidirectional) من فئة recurrent neural networks؛

ب. نستكشف إستراتيجيات مختلفة للاستفادة من الموارد النصية الخام في إنشاء تمثيلات مسبقة (pretrained embeddings) لمهمة POS في البيانات المتضمّنة للتناوب اللغوي؛

ج. ندرس أثر تضمين مؤشرات اللغة (language identifiers) لأداء مهمة مشتركة تجمع بين تحديد اللغة ووسم الأجزاء النحوية؛ (د) نقدّم أول تقييم تجريبي لمهمة POS عبر أربعة أزواج لغوية مختلفة، علماً بأن جميع الأبحاث السابقة ركّزت على زوج لغوي واحد أو زوجين اثنين لا أكثر.

٢- الأعمال ذات العلاقة

لتحقيق قابلية التعميم على مختلف أزواج اللغات، نجري دراستنا على بيانات أربع مجموعات خاصّة بالتناوب اللغوي، تشمل أربعة أزواج لغوية: اللغة العربية الفصحى الحديثة مع اللهجة المصرية (MSA-EGY)، ويُعدّ تحليل المحتوى الذي ينتجه المستخدمون وتلبية احتياجات المجتمعات متعدّدة اللغات أمراً بالغ الأهمية (Vyas et al., 2014b). في الدراسات السابقة التي تناولت مشكلة توسيم الأجزاء النحوية (POS) في بيانات التناوب اللغوي (CS)، غالباً ما يُشرع أولاً في تحديد اللغة الصحيحة لكل كلمة قبل إدخالها في مُوسّم أحادي اللغة (Solorio and Liu, 2008; Vyas et al., 2014a; AlGhamdi et al., 2016). وكما هو الحال في كثير من تطبيقات المعالجة اللغوية الطبيعية (NLP)، تعاني هذه السلاسل الإجرائية (pipelines) من مشكلة تراكم الأخطاء؛ إذ إنّ إخفاق مرحلة تحديد اللغة ينعكس سلباً على دقة تنبؤ التوسيم النحوية. ومن جهة أخرى، اعتمدت بعض المقاربات نماذج خاضعة للإشراف مُدرّبة على بيانات تناوب لغوي تحمل وسم الأجزاء النحوية، لكنّ إعداد مثل هذه البيانات مُكلف ويندر وجوده لغالبية أزواج اللغات (AlGhamdi et al., 2016; Solorio and Liu, 2008; Jamatia et al., 2015; Barman et al., 2016).

قدّم Solorio and Liu (2008) أول مقارنة إحصائية لمهمة توسيم الأجزاء النحوية في سياق التناوب اللغوي، حيث استخدموا عدة قواعد عملية (heuristics) لدمج مُوسّسات أحادية اللغة، لكنّ نجاحها كان محدوداً؛ إذ بلغ مستوى الدقة ٨٦٪ عند اختيار مخرجات واسم أحادي اللغة استناداً إلى هوية اللغة من المعجم لكل رمز لغوي. غير أنّ النموذج المعتمد على آلة متجهات الدعم (SVM) المُدرّبة على مخرجات المُوسّسات الأحادية تفوّق على ما يُعرّف بـ "المُستشرف المثالي" (oracle)، محققاً دقة بلغت ٩٣,٤٨٪. وفي سياق آخر، قدّم



Royal Sequiera (2015) مقارنة قائمة على التعلّم الآلي مع مجموعة ميزات جديدة لتوسيم HIN-ENG (تغريدات تويتر ومحدثات فيسبوك)، حيث تأخذ هذه المجموعة الجديدة من الميزات في الحسبان مشكلة transliteration الشائعة في وسائل التواصل الاجتماعي، وبلغت دقة النظام ٨٤٪.

أما (Jamatia et al. (2015 فقد استخدموا مجموعتيّ وسم نحوي إحداهما مفصلة والأخرى مبسطة، وطرحوا مقارنة بين أداء دمج عدّة مُوسّسات خاصة بكل لغة على حدة، وبين أداء مقارنة قائمة على التعلّم الآلي تستخدم طيفاً من الميزات المختلفة. وخلصوا إلى أن المقارنة المعتمدة على التعلّم الآلي لم تتفوّق على دمج المُوسّسات الخاصّة بكل لغة. وفي سياق آخر، درس (alGhamdi et al. (2016 سبع إستراتيجيات مختلفة لتوسيم الأجزاء النحوية بهدف استثمار الموارد الأحادية المتوافرة للتعامل مع بيانات التناوب اللغوي، وطبقوها على زوجي لغات: MSA-EGY وSPA-ENG. قُسّمت هذه الإستراتيجيات إلى ما يُعرّف بشروط الدمج (combined conditions) وشروط الدمج المتكامل (integrated conditions). شملت ثلاثة من شروط الدمج تشغيل مُوسّسات أحادية اللغة إلى جانب مُحدّدات اللغة (language ID taggers) بترتيب مختلف، ودمج المخرجات ضمن توقّع واحد متعدّد اللغات، فيما تمثّل الشرط الرابع في تدريب نموذج SVM استناداً إلى مخرجات المُوسّسات الأحادية. أما المقاربات الثلاث المتكاملة فقد اعتمدت تدريب مُوسّم أحادي اللغة متقدّم على: (أ) المتن الخاص بالتناوب اللغوي (CS corpus)، أو (ب) اتحاد متّين أحاديّين للغتين المتضمّنيتين في التناوب، أو (ج) اتحاد المتّين المستخدمين في الخطوتين (أ) و(ب). وقد تفوّقت جميع هذه الشروط (سواء الدمج أو الدمج المتكامل) على الأنظمة الأساسية، وحقّق نهج SVM نتائج أفضل باستمرار من المقاربات المتكاملة.

وفي إطار الزوج اللغوي SPA-ENG، اقترح (Soto and Hirschberg (2018 مقارنة جديدة لتوسيم الأجزاء النحوية في سياق التناوب اللغوي بالاعتماد على الشبكات العصبية التكرارية (RNN)، بهدف توفير أدوات أكثر فعالية لمعالجة بيانات التناوب اللغوي، بما في ذلك أنظمة التوسيم النحوي. استخدم المؤلفان مقارنة تجريبية تتضمن توظيف شبكة عصبية تكرارية ثنائية الاتجاه (bi-directional RNN) لدراسة أثر تضمين محدّدات اللغة

على نحو موسّع. وفيما يخص الميزات المنطقية (Boolean)، فقد اعتمدت المقاربة شبكة عصبية متقدمة من نوع BiLSTM تستخدم النهايات (suffix)، والبدایات (prefix)، وتمثيلات الكلمات (word embeddings). وأظهرت النتائج أن النموذج المقترح يتفوق نسبياً على أحدث نُظُم التوسيم الخاصة بالتناوب اللغوي، إذ بلغ مستوى الدقة ٩٦,٣٤٪، في حين حقق التوسيم النحوي المشترك مع تعرّف اللغة دقة ٩٦,٣٩٪ و ٩٨,٧٨٪ على التوالي (Soto and Hirschberg, 2018).

وقد طُرِحَت مقاربات لنمذجة اللغة (Language Modeling, LM) في نصوص HIN-ENG من (Chandu et al., 2018). واقترح (Pratapa et al., 2018) بدورهم آلية حسابية لتوليد بيانات مُصطنعة من التناوب اللغوي تتمتع بصلاحية نحوية، مستندةً إلى نظرية التكافؤ المقيد (ECT: Equivalence Constraint Theory)، وذلك للتعامل مع مشكلة نقص بيانات التدريب الخاصة بنماذج اللغة في سياق التناوب اللغوي. واعتمد (Pratapa et al., 2018) تقنيتين من تمثيلات ثنائية اللغة: Correlation Based و Bilingual Compositional Model (BiCVM) (Faruqui and Dyer, 2014; Blunsom and Hermann, 2014). وقد أدّت التضمينات المعتمدة على الكلمات إلى تحسين مهام المعالجة الدلالية والنحوية في سياق التناوب اللغوي؛ إذ إن BiCVM يعمل جيداً على المستوى الدلالي فقط، وكذلك BiCCA تؤدي أداءً ملحوظاً في المهام الدلالية نظراً لاعتمادها على محاذاة الكلمات. إضافةً إلى ذلك، تفوّقت skip-gram embedding على كلٍّ من BiCVM و BiCCA، وحقّقت أداءً أفضل في المهام النحوية والدلالية على حدّ سواء. وأظهرت دراسة (Pratapa et al., 2018) أن استخدام تمثيلات مُسبقة التدرّب (pretrained embeddings) مشتقة من بيانات التناوب اللغوي يفوق في جودته تمثيلات ثنائية اللغة المُتعارف عليها. من جانب آخر، قدّم Bhat et al. (2018) محلاًّ للاعتمادية (dependency parser) مصمماً خصيصاً لبيانات التناوب اللغوي في HIN-ENG. واستندوا إلى بنية التراصّ العصبي المقترحة في (Zhang and Weiss, 2016; Chen et al., 2016) لتعلّم كلٍّ من توسيم الأجزاء النحوية والتحليل النحوي، ونقل المعرفة من النماذج ثنائية اللغة المُدرّبة على المتن الموحد (UD treebanks) لكلٍّ من الهندية والإنجليزية. وقد حقّق المُوسّم المعروف بـ "stack-prop tagger"، والمقترح في (Zhang



(and Weiss, 2016)، أعلى دقة بلغت ٩٠,٥٣٪. ولتحديد اللغة، قاموا بتدريب شبكة إدراك عميقة متعدّدة الطبقات (MLP) فوق شبكة تكرارية ثنائية الاتجاه من نوع (Bi-LSTM). ووصلت دقة النظام الناتج إلى ٩٧,٣٩٪.

٣- المنهجية

تُتيح تمثيلات الكلمات المُدرَّبة مسبقاً (pretrained word embeddings) للنماذج استثمار البيانات النصيّة الخام المتوفرة بجميع اللغات، والتي تفوق في حجمها البيانات الموسّمة (annotated). وقد برز في الآونة الأخيرة اهتمامٌ بمناهج التضمينات (embeddings) متعددة اللغات وثنائية اللغة؛ حيث تُنسب تمثيلات أحادية اللغة للغتين إلى مساحة تمثيلات مشتركة. وتتمثّل الفائدة الأساسية لتمثيلات ثنائية ومتعدّدة اللغات في حل مهام تستلزم الربط بين لغتين، مثل الترجمة الآلية (Zou et al., 2016; Vulic and Moens, 2016) (MT) (al., 2013)، إلى جانب إتاحة نقل النماذج المُعلّمة على لغة غنيّة بالموارد إلى لغة شحيحة الموارد (Adams et al., 2017). ومن التطبيقات المحتملة لهذه التضمينات ثنائية ومتعدّدة اللغات، معالجة اللغة المتضمّنة للتناوب اللغوي (code switching). في هذا القسم، نُجري مقارنةً بين استخدام بنياتٍ متعددة للشبكات العصبية في توسيم الأجزاء النحوية (POS) لبيانات التناوب اللغوي (CS). إضافةً إلى ذلك، نستكشف تشكيلات متنوعة للتمثيلات (embedding setups) بغية التعرّف على أفضل السبل للتعامل مع توسيم الأجزاء النحوية في بيانات التناوب اللغوي. أولاً، نوضّح الأداة المستخدمة في تدريب طبقة التضمينات. ثم نعرض نماذج الشبكات العصبية. بعدها نورد تشكيلات التضمينات التي استخدمناها.

٣-١- نموذج تمثيل الكلمات المُدرَّبة مسبقاً (Pretrained Word Embedding Model)

تعتمد أكثر الأساليب نجاحاً في تعلّم تمثيلات الكلمات (Mikolov et al., 2013c; Pennington et al., 2014; Bojanowski et al., 2017) على فرضية التوزيع (distributional hypothesis) (et al., 2013b)، أي إنّ الكلمات التي تَرِد في سياقات متشابهة تميل إلى امتلاك دلالات متقاربة.

ومن بين جميع تقنيات تمثيل الكلمات، اخترنا أداة FastText التي طوّرتها شركة فيسبوك (Bojanowski et al., 2017). ويعود سبب الاختيار إلى أنّ شبكات التواصل الاجتماعي تُعدّ المصدر الأساسي لبياناتنا غير الموسومة، حيث يتّسم هذا النوع من البيانات بتنوّعاتٍ إملائية كبيرة وأخطاءٍ كتابية عديدة. يستفيد FastText من المعلومات المستقاة من الأجزاء الفرعية للكلمات (subword) مثل n-grams؛ إذ يخلق تمثيلاتٍ مُتجهية (vector representations) لمرشحات تصحيح الأخطاء الإملائية غير الموجودة أصلاً في فضاء التضمينات المُدرَّب، وذلك بجمع المتجهات الخاصّة بالأحرف المكوّنة للـ n-grams. وتهدف مختلف تقنيات تمثيل الكلمات إلى التقاط علاقاتٍ بين المفردات التي تظهر في سياقات متشابهة، وهي علاقاتٌ تحدث عادةً على مستوى الجُمْل في معظم تطبيقات المعالجة اللغوية الطبيعية المُدرّبة على بيانات أحادية اللغة. لكنّ نقص السياق الكافي في بيانات التناوب اللغوي يجعل تعلّم هذا النمط من العلاقات أمراً صعباً؛ إذ يحدث التناوب اللغوي على مستوى أقلّ من الجُمْلَة (intrasentential). ومن ثمّ، لجأنا إلى استخدام FastText نظراً لاستناده بالأساس إلى استثمار معلومات الأجزاء الفرعية (subword). إذ تحلّ مقارنة n-gram مشكلة نمذجة اللغات التي تشهد تصنيفاتٍ نادرة (rare word inflections) بالاعتماد على نُطق الأحرف (character n-grams). وفي المقابل، تفتقر تقنيات تمثيل أخرى، مثل word2vec و glove، إلى معلومات الأجزاء الفرعية، ما يجعلها أقلّ فعالية مع اللغات الغنية صرفياً كالعربية، والبيانات المليئة بالضوضاء كالتّي في تويتر (Mikolov et al., 2013a; Pennington et al., 2014).

في تجاربنا، نستخدم ثلاث بُنى مختلفة للشبكات العصبية:

- أ. بنية BiLSTM-CRF شبيهة بما اقترحه (Reimers and Gurevych, 2017) لأغراض ٧ توسيم أجزاء الكلام نحويّاً (POS tagging)؛
- ب. نموذج تعلّم متعدد المهام يهدف إلى تعلّم توسيم أجزاء الكلام نحويّاً (POS tagging) لأزواج لغوية متقاربة في آنٍ واحد؛
- ج. نموذج تعلّم متعدد المهام يهدف إلى تعلّم توسيم أجزاء الكلام نحويّاً (POS tagging) وتحديد هوية اللغة (Language ID) بشكلٍ متوازٍ.



- في النموذج الأول، نقوم بتدريب شبكة BiLSTM مع هدف دالة حقل عشوائي شرطي (conditional random field) على غرار (Reimers and Gurevych, 2017)، حيث نحصل على التوزيع الاحتمالي لجميع التصنيفات من خلال نمذجة احتمالية تسلسل العلامات النحوية بأكمله. نُهيئ طبقة التضمينات (embedding layer) بتضمينات FastText المُدرّبة مسبقاً، ثم نُمرّر المخرجات المتتالية لهذه الطبقة إلى طبقة BiLSTM. تحتوي الطبقة الخفية (hidden layer) في BiLSTM على ٢٠٠ وحدة في كل اتجاه، مع معدل حذف (dropout) قدره ٠.٥. ونستعين بآلية الإيقاف المبكر (early stopping) (Chollet et al., 2015) استناداً إلى الأداء على مجموعات التحقق (validation sets). نشير إلى هذا النموذج لاحقاً بـ BiLSTM-CRF POS Tagger في بقية المقال والجداول.
- أما النموذج الثاني، فهو نموذج تعلّم متعدد المهام يتعلّم بالتوازي توسيم الأجزاء النحوية (POS) لأزواج لغوية متقاربة في سياق التناوب اللغوي. يشترك هيكل هذا النموذج مع بنية (Reimers and Gurevych, 2017) BiLSTM-CRF، لكننا نضيف طبقة CRF ثانية لإنتاج العلامات النحوية الخاصة بزواج لغوي معيّن (على سبيل المثال MSA-LEV)، بينما تختص الطبقة الأولى بتصنيف العلامات النحوية للزوج اللغوي الآخر (على سبيل المثال MSA-EGY). وتعمل الطبقتان على نحو مستقل، إذ ترتبط كل منهما بمصفوفة الأوزان في الطبقة الخفية، ويُعطى كلا دالتي الخسارة (loss functions) الوزن نفسه. نُطلق على هذا النموذج لاحقاً اسم MTL-POS Tagger في بقية المقال والجداول.
- النموذج الثالث هو أيضاً نموذج تعلّم متعدد المهام، لكنّه يتعلّم بالتوازي توسيم الأجزاء النحوية (POS) وتحديد هوية اللغة (language id)، بهدف تعزيز أداء مهمة توسيم الأجزاء النحوية. يقوم هيكل هذا النموذج على أساس بنية MTL-POS Tagger نفسها، مع فارقٍ وحيد هو تخصيص طبقة مخرجات واحدة لتوسيم الأجزاء النحوية، وطبقة مخرجات أخرى لمهمة تحديد هوية اللغة. نشير إلى هذا النموذج باسم MTL-POS+LID Tagger.

٣-٢- الظروف التجريبية (Experimental Conditions)

٣-٢-١- تضمين أحادي اللغة (خط الأساس) (Monolingual embedding (baseline)

في هذه الحالة، نقوم بتدريب تضمينات الكلمات بالاعتماد على المدونات الأحادية اللغة لكل لغة من اللغات المشاركة في الأزواج اللغوية الأربع. وينتج عن هذه الطريقة ستة تضمينات مستقلة مُدرّبة مسبقاً، وهي مخصصة للـ MSA، و EGY، و LEV، و ENG، و SPA، و HIN. ولكل لغة، ندرّب نموذج BiLSTM-CRF بالاعتماد على واحدٍ من التضمينات الستة المُدرّبة مسبقاً. ونعدّ هذه النماذج بمثابة الأنظمة الأساسية (baseline systems). أمّا أداء خط الأساس (baseline performance) فيتمثّل في دقة توسيم الأجزاء النحوية (POS tagging) التي تُحقّقها النماذج الأحادية اللغة، دون أي تدريبٍ مخصّص لبيانات التناوب اللغوي (CS).

٣-٢-٢- التضمينات ثنائية اللغة المدمجة (Merged Bilingual embeddings)

المدونات الأحادية اللغة المُجمّعة والمُرشّحة - Combined filtered monolingual corpora (CFM): بغية الاستفادة من النمط الخاص بالتناوب اللغوي على مستوى الجُمْل (inter-sentential)، نقوم بتدريب نموذج باستخدام تضمين مُدرّب مسبقاً للكلمات (pretrained word embedding) بالاعتماد حصرياً على البيانات الأحادية اللغة. يفترض هذا التصوّر أن بيانات MSA تقتصر على الفصحى الحديثة وحدها، وأن بيانات الـ EGY تقتصر على اللهجة المصرية فحسب. وينتج عن ذلك تضمينات مُدرّبة مسبقاً للتناوب اللغوي على مستوى الجُمْل (inter-sentential CS). ولا تحتوي أيٌّ من هذه الجُمْل على بيانات تناوب لغوي داخل العبارة نفسها (intra-sentential).

٣-٢-٣- المجموعات النصية ذات التبديل اللغوي الصرف - Pure Code switched corpora (PCS)

للاستفادة من نمط التبديل اللغوي داخل الجملة (intra-sentential CS)، قمنا بتدريب نموذج باستعمال تضمينات كلمات مُسبقة التدريب مبنية حصرياً على بيانات تبادل



لغوي (CS). يفترض هذا الإجراء أنَّ البيانات المستخدمة في بناء التضمينات تجسّد ظاهرة التبدل اللغوي ذاتها، مما يؤدي إلى الحصول على نمط التبادل اللغوي (CS) داخل الجملة حصراً؛ إذ لا تتضمّن أي من الجمل تبديلاً لغوياً بين الجمل (inter-sentential CS).

٣-٤-٤- المدونات الأحادية المدمجة مع المدونات ذات التبدل اللغوي - (Pseudo)

Combined monolingual and CS corpora (PseudoCS)

لمعالجة كلا نمطي التبدل اللغوي، داخل الجملة وبين الجمل، ندمج المدونات ذات التبدل اللغوي الصرف (PCS) مع المدونات الأحادية المُجمّعة والمُرشّحة (CFM) لتكوين مدونة موحّدة تُستخدم في تدريب تضمينات كلمات موحّدة.

٣-٤-٥- تضمينات متعددة اللغات مدمجة - Merged multilingual embeddings

نقوم بتدريب تضمين متعدد اللغات لأزواج اللغات التي تشترك في لغةٍ محورية واحدة (Pivot language). ولتحقيق ذلك، ندمج المدونات المُستخدمة في تدريب تضمينات الكلمات للأزواج التي تتقاسم لغةً مشتركة. في زوجي اللغات MSA-LEV و MSA-EGY تمثل اللغة الفصحى الحديثة (MSA) اللغة المشتركة، في حين تُعد اللغة الإنجليزية (ENG) لغةً محورية بين SPA-ENG و HIN-ENG. للاستفادة من هذا القاسم المشترك في كلّ من هذين الزوجين، ندمج جميع المدونات السابقة - PCS و Mono و PseudoCS - الخاصة بكل زوج لغوي لتشكيل مدونة موحّدة تُستخدم في تدريب نموذج تضمين متعدد اللغات مدمج للغات MSA-EGY-LEV، ومدونة أخرى لأزواج SPA-HIN-ENG. تهدف هذه التضمينات إلى التقاط أنماط استعمال الكلمات في سياق كل لغة على حدة، مع الحدّ من الالتباس الناجم عن الكلمات ذات البنية السطحية المتطابقة في أكثر من لغة.

٣-٤-٦- التضمين الثنائي المُسقط - Projected Bilingual Embedding

تمثّل التضمين الثنائي المُسقط (Projected bilingual embeddings) متجهات (vector representations) للغتين تُسقط في فضاءٍ مشترك، بحيث تصبح متجهات أزواج الكلمات المترجمة شبيهة ببعضها. هناك ثلاثة مناهج لتعلم التضمينات الثنائية للغة: (١) عن

طريق ربط فضائيّ التضمينات أحادية اللغة في فضاء مشترك واحد؛ ٢) تكييف أحادي اللغة لفضاء تضمينات لغة ما ليناسب لغة أخرى؛ ٣) التدريب الثنائي اللغة عن طريق الاستفادة من التضمينات الهدف المستخلصة من فضاء تضمينات لغة مصدر مدرب تدريباً جيداً.

في عملنا، نقوم بتدريب كلٍّ من نماذج التضمين الخاصة بالتناوب اللغوي (CS) والأحادية اللغة بشكل مستقل، قبل إسقاطها في فضاء مشترك واحد. ولتنفيذ ذلك، نستخدم نموذج MUSE (Conneau et al., 2017)، وهو إحدى أحدث التقنيات في إنشاء تمثيل ثنائي اللغة مسقط (projected bilingual embedding) بالاستناد إلى تقنية monolingual adaptation لتشكيل فضاء مشترك. يتميز MUSE بإمكانية التعلّم سواءً بالإشراف (supervised learning) أو دون إشراف (unsupervised learning). وفي دراستنا، نستعين بالإصدار غير الخاضع للإشراف (unsupervised) من MUSE.

٤- التقييم

٤-١- مجموعات البيانات (Datasets)

خلال تجاربنا، اعتمدنا مجموعة تقييم واحدة لكل زوج لغوي، إلى جانب مدونات متعدّدة لتدريب طبقة التضمين (embeddings). يوضّح الجدول توزيع مجموعات البيانات المُخصّصة للتقييم.

MSA-EGY: للتقييم، استخدمنا في هذه الدراسة على LDC Egyptian Arabic Treebanks (Maamouri et al., 2012) (ARZ1-5) 1-5، والتي تتألف بياناتها من نصوص مندييات حوارية مكتوبة أساساً باللهجة المصرية (EGY). يبلغ إجمالي الجمل في هذه المدونة ١٣,٦٩٨ جملة، ويصل عدد كلماتها إلى ١٧٤,٩٦٧ كلمة.

ولبناء التضمينات المُسبقة التدريب، قمنا بجمع تغريدات مصرية من حسابات بعض الشخصيات العامة المصرية على منصة X. أما بقية النصوص الخام باللهجة المصرية فقد استُقيت من المصادر الآتية: مدونة التعليقات الإخبارية المصرية على الإنترنت



(Zaidan and Callison-Burch, 2011)؛ تغريدات مصرية استُخدمت في المهام المشتركة للتبديل اللغوي (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018)؛ جزء من MSA Gigaword (Parker et al., 2011)؛ إضافة إلى مدونات LDC Arabic Treebank الخاصة بـ MSA (Maamouri et al., 2004; Diab et al., 2013).

ولتمييز هوية اللغة على مستوى الوحدات اللغوية (tokens) بين MSA و EGY، استخدمنا أداة (Al-Badrashiny et al., 2015) Automatic Identification of Dialectal Arabic (AIDA2) لتنفيذ التصنيف اللغوي على مستوى الكلمة في السياق.

MSA-LEV: اعتمدنا في بيانات التبديل اللغوي النصي بين العربية الفصحى (MSA) واللهجة الشامية (LEV) على Curras Corpus للغة العربية الفلسطينية (Jarrar et al., 2017)، وهي إحدى لهجات العربية الشامية. تضم المدونة نحو ٥٧,٠٠٠ كلمة؛ نصفها مأخوذ من تفرغ برنامج تلفزيوني، فيما يأتي النصف الآخر من مصادر متنوعة كمنصة Facebook والمنديات الحوارية وتغريدات منصة X والمدونات. وقد وُسمت المدونة صريفاً بواسطة (Eskander et al. 2016) وفق الإرشادات نفسها المعتمدة في وسم مدونة ARZ المصري. قمنا بوسم مجموعة التقييم الخاصة بـ MSA-LEV بهوية اللغة مستخدمين الإرشادات والأداة المقترحة في (Diab et al., 2016; AlGhamdi and Diab, 2018).

ولبناء التضمينات المسبقة التدريب، جمعنا تغريدات من حسابات شخصيات عامة شامية، بالإضافة إلى نصوص خام بالفصحى والشامية من مصادر متعددة: مدونة التعليقات الإخبارية على الإنترنت (Zaidan and Callison-Burch, 2011)، مدونات COLABA (Diab et al., 2010)، التعليقات والتغريدات في (Cotterell and Callison-Burch 2014)، الجزء الشامي من مجموعة (PADIC (Meftouh et al., 2015)، جزء من MSA Gigaword (Parker et al., 2011)، إضافة إلى مدونات LDC Arabic Treebank الخاصة بالفصحى (Maamouri et al., 2004; Diab et al., 2013).

SPA-ENG: اعتمدنا في بيانات التبديل اللغوي بين الإسبانية والإنجليزية على مدونة (Miami Bangor (MB) corpus) وهي عبارة عن حوارات مسجلة مع متحدثين ثنائيي اللغة (الإسبانية-الإنجليزية) يقيمون في ميامي بولاية فلوريدا. تضم المدونة ٥٦ محادثة مسجلة

لعدد ٨٤ متحدثاً (Soto and Hirschberg, 2017) ، ويتكوّن من ٢٤٢,٤٧٥ كلمة (٣٣٣,٠٦٩ كلمة عند احتساب علامات الترقيم) تغطّي ما يقارب ٣٥ ساعة من التسجيلات الصوتية. وقد وُسمت هوية اللغة لكل كلمة في هذه المدونة يدوياً.

لتدريب التضمينات المسبقة لكلّ من الإسبانية والإنجليزية، استخدمنا English Universal Dependencies (UD) corpus (Silveira et al., 2014) و McDonald Spanish UD corpus (et al., 2013). علماً بأنّ مشروع Universal Dependencies (UD) يهدف إلى توفير أرشفة نحوية شجرية متّسقة (treebank) عبر اللغات. إضافةً إلى ذلك، اعتمدنا بيانات أحادية اللغة (إنجليزية) من مصادر متنوعة يبلغ حجمها نحو ٢٥٠ مليون جملة.

HIN-ENG : تستند شجرة التراكيب (treebank) الخاصة بالتبديل اللغوي بين الهندية والإنجليزية إلى تغريدات تحتوي على تبديل لغوي (CS) كتبها متحدثون متعدّدو اللغات بالهندية والإنجليزية (غالبيتهم من الهند) (Bhat et al., 2017). وُسمت الشجرة يدوياً وفق مخطط Universal Dependencies (UD). تضم المدونة بيانات مأخوذة من منصة X تتألف من ١,٨٥٢ تغريدة تحتوي على ٢٦,٢٢٤ وحدة لغوية (tokens).

ولإنشاء تضمينات مُسبقة التدريب لكلّ من اللغتين الهندية والإنجليزية، اعتمدنا مدونة English UD (Silveira et al., 2014) ومدونة Hindi UD (Bhat et al.; Palmer et al., 2009)، إلى جانب بيانات أحادية اللغة هندية جُمعت من مصادر متنوعة. جرى تعلّم تمثيلات الكلمات باستخدام نموذج Skip-gram مع التعيين السلبي (negative sampling) كما هو مُنفذ في حزمة FastText لجميع اللغات.

٤-٢. نتائج خطّ الأساس (Baseline Results)

يشير أداء الخطّ الأساس (baseline performance) إلى دقة توسيم الأجزاء النحوية (POS tagging) في النماذج الأحادية اللغة دون تدريب خاصّ على بيانات التبديل اللغوي (CS) معيار الأداء المرجعي لدينا. وعليه، تُعدّ نماذج الشبكات العصبية المُدرّبة بالاعتماد على التضمينات الأحادية (monolingual embeddings) هي خطوط الأساس في تجاربنا. وفي حال



لم تطرح بيانات التناوب اللغوي (CS data) أي صعوبة إضافية على مُوسّسات الأجزاء النحوية الأحادية اللغة، فلا تتوقّع عندئذٍ تراجعاً كبيراً في مستوى الأداء. يوضّح الجدول ٢ أداء مختلف أنظمة وسم الأجزاء النحوية (POS) الأساسية على بيانات الاختبار. لكل زوج لغوي يوجد نظامان مرجعيان:

- في زوج MSA-EGY سجّل النموذج الأساس دقّة ٨٥,٤٠٪ عند استخدام تضمينات MSA المسبقة التدريب، مقابل ٨١,٠٦٪ عندما استند نموذج BiLSTM-CRF إلى تضمينات EGY المسبقة التدريب.
- وبالمناهجية ذاتها تقدّم نتائج خطوط الأساس لأزواج اللغات MSA-LEV و SPA-ENG و HIN-ENG، كما هو مبين في الجدول ٢.

جدول (١) توزيع مجموعات البيانات للأزواج اللغوية الأربع

اختبار - Test	تطوير - Dev	تدريب - Train	مجموعات البيانات (Datasets)
20,464	21,146	133,357	ARZ (MSA-EGY)
5,779	5,749	45,167	LEV (MSA-LEV)
67,114	67,114	268,464	Miami Bangor (SPA-ENG)
3,190	3,339	19,695	UD-HIN-ENG

٣-٤ النتائج

في هذا القسم نعرض نتائج تجاربنا باستخدام نماذج الشبكات العصبية ومنهجية التضمينات المبينة في القسم ٣، بالاعتماد على مجموعات البيانات الواردة في القسم ١,٤. كما نستعرض أداء ثلاثة نماذج عصبية عند تهيئة التضمينات عشوائياً. يبيّن الجدول ٢ دقّة وسم أجزاء الكلام (POS) لجميع أزواج اللغات، في حين يوضّح الجدول ٣ دقّة تحديد هوية اللغة (LID)

جدول (٢) نسبة دقة وسم أجزاء الكلام POS (%) على المجموعات النصية الأربعة

Embedding – حالة التضمين – Condition	MSA-EGY	MSA-LEV	SPA-ENG	HIN-EN
BiLSTM-CRF Tagger (1) + Random-Initi-Embed	89.12	88.90	95.33	69.87
(1) + mono (MSA/SPA/HIN) (Baseline)	85.40	83.25	50.30	63.30
(1) + mono (EGY/LEV/ENG) (Baseline)	81.06	80.21	71.21	67.11
(1)+ Merged Bilingual CFM	90.00	89.41	95.40	85.98
(1)+Merged Bilingual PCS	89.06	88.94	94.22	83.31
(1)+Merged Bilingual PseudoCS	91.96	91.92	96.55	86.01
(1)+ Merged Multilingual Embeddings(pivot)	92.81	92.91	94.81	85.87
(1)+Projected Bilingual	89.24	87.60	91.31	83.02
MTL-POS Tagger (2) + Random-Initi-Embed	89.91	90.02	92.89	79.51
(2)+Merged Bilingual PCS	90.01	90.51	93.21	84.30
(2)+Merged Bilingual PseudoCS	92.90	92.92	94.14	84.33
MTL-POS+LID Tagger (3) + Random-Initi-Embed	88.96	89.79	92.65	78.51
(3)+ Merged Bilingual CFM	90.00	89.01	95.42	84.01
(3)+Merged Bilingual PCS	90.41	90.48	95.29	84.41
(3)+Merged Bilingual PseudoCS	91.89	91.92	96.50	85.91
(3)+Projected Bilingual	88.61	88.09	92.91	82.39
State-of-the-art	90.56	--	96.63	91.90



يُظهر الجدول المتوسط المحسوب عبر خمس تجارب باستخدام بذور عشوائية مختلفة. يُشير الخط العريض المائل إلى أفضل نتيجة ضمن تجاربنا، في حين يدل الخط العريض وحده على أفضل نتيجة تتفوّق على الأنظمة الأحدث (state-of-the-art). نُرّمز إلى BiLSTM-CRF Tagger بـ(١)، وإلى MTL-POS Tagger الذي يتعلّم توسيم POS للغاتٍ متقاربة بـ(٢)، وإلى MTL-POS+LID الذي يتعلّم بالتوازي كلاً من توسيم POS وتحديد هوية اللغة بـ(٣). أمّا Random-Initi-Embed فيُقصد به التمثيلات المهيّأة عشوائياً.

لكل زوج عند استخدام MTL-POS+LID Tagger. ولتقييم الأداء، نحسب الدقّة في كل شرط تجريبي عبر مقارنة الوسوم الناتجة بالوسوم الذهبية المتاحة في كل مجموعة بيانات. نطبّق جميع الشروط التجريبية على مجموعة الاختبار نفسها داخل كل زوج لغوي: بالنسبة إلى MSA-EGY نُقدّم النتائج على مجموعة اختبار MSA-EGY، وبالنسبة إلى MSA-LEV على مجموعة اختبارها، ولـ SPA-ENG على مجموعة اختبار مدونة Miami Bangor، وأخيراً لـ HIN-ENG على مجموعة الاختبار UD-HIN-ENG. أعلى دقتين لزوجي MSA-EGY و MSA-LEV بلغت ٩٢,٩٠٪ و ٩٢,٩٢٪ على التوالي، وقد تحققتا باستخدام MTL-POS Tagger مع تضمينات Merged Bilingual PseudoCS.

جدول (٣) دقّة تحديد هوية اللغة (LID) (٪) في المجموعات النصية الأربعة

حالة التضمين - Embedding Condition	MSA-EGY	MSA-LEV	SPA-ENG	HIN-ENG
MTL-POS+LID Tagger: (3) + Random-Initi-Embed	80.71	79.29	96.42	78.41
(3) + mono (MSA/SPA/HIN) (baselines)	77.41	78.08	88.11	71.46
(3) + mono (EGY/LEV/ENG) (baselines)	71.51	76.37	85.09	74.80
(3)+ Merged Bilingual CFM	80.06	78.39	95.49	92.54
(3)+Merged Bilingual PCS	81.33	79.28	95.02	93.20

حالة التضمين - Embedding Condition	MSA-EGY	MSA-LEV	SPA-ENG	HIN-ENG
(3)+Merged Bilingual PseudoCS	82.15	81.32	97.20	94.92
(3)+Projected Bilingual	78.17	79.11	90.01	87.69
State-of-the-art	--	--	98.78	97.39

يوضّح الجدول ٣: دقّة تحديد هوية اللغة (LID) (%) في المجموعات النصية الأربعة. يُعرّض المتوسط المحسوب عبر خمس تجارب بيدور عشوائية مختلفة. يدل الخط العريض المائل على أفضل نتيجة ضمن تجاربنا، بينما يشير الخط العريض وحده إلى أفضل أداء مقارنة بالأنظمة الأحدث. يُرمز إلى نموذج - MTL-POS+LID - الذي يتعلّم وسم الأجزاء النحوية POS وتحديد الهوية اللغوية في آن واحد (LID) - بالرقم (٣). ويقصد ب-Random Embed-Initi التضمينات التي تُهيأ أوزانها عشوائياً.

يتفوّق نموذجنا الأفضل ل-MSA-EGY على أحدث الأنظمة بنحو ٢٪، AlGhamdi et al., (2016). ولغياب أنظمة سابقة على مجموعة بيانات MSA-LEV بعد تحويل وسوم Buckwalter إلى وسوم UD، تعدّ إجراء مقارنة مباشرة. أما تضمينات BiLSTM-CRF+Merged Bilingual PseudoCS فأظهرت أعلى دقّة مع زوجي SPA-ENG و HIN-ENG؛ إذ بلغ أداء نموذج SPA-ENG ٩٦,٥٥٪، وهو قريب من النظام الأحدث (Soto and Hirschberg, 2018) ٩٦,٦٣٪، بينما سجّل نموذج HIN-ENG ٨٦,٠١٪، أي أدنى من النظام الأفضل (Bhat et al., 2018) ٩١,٩٠٪. وفي مهمة LID جاءت جميع نماذجنا دون مستوى الأنظمة الأحدث؛ ونظراً لغياب أنظمة مرجعية لزوجي MSA-EGY و MSA-LEV، قارّنا أداءنا بخطوط الأساس فقط.

٥- المناقشة

تُسهّم التضمينات المتعدّدة اللغات (multilingual embedding) مثل تلك الخاصة بزوجي MSA-EGY و MSA-LEV في تحسين أداء اللغات ذات الصلة الوثيقة أو المتقاربة (EGY / LEV)، في حين توفّي إلى إضافة قدر من الضوضاء عند التعامل مع لغات متباعدة (SPA /



(HIN). وبالمثل، أظهر التدريب المشترك على وسم أجزاء الكلام (POS) للغات متقاربة أفضل دقة في زوجي MSA-LEV و MSA-EGY مقارنةً باللغات المتباعدة. لقد سجّلت نتائج الدقة لهذين الزوجين أعلى القيم عبر جميع الإعدادات التجريبية، ويُعزى هذا التحسّن على الأرجح إلى العدد الكبير من الكلمات المتشابهة شكلياً - بعضها من أصل مشترك - التي تتقاطع بين هاتين اللغتين.

يختلف سلوك التناوب اللغوي (CS) تبعاً لوسيلة التواصل، والموضوع، والمتحدثين (أو الكتاب)، واللغات المتداخلة، إلى جانب عوامل أخرى. بناءً على ذلك، نرى أنّ تباين نوعية (genre) مجموعات بيانات التقييم في زوجي اللغات SPA-ENG و HIN-ENG هو أحد الأسباب المحتملة لعدم استفادتهما من التمثيلات اللغوية المتعددة (multilingual embedding) أو التعلّم المشترك لمهمة POS tagging. على النقيض، يُظهر MTL-POS-LID Tagger - الذي يتعلّم في آن واحد توسيم أجزاء الكلام النحوية (POS tags) وهوية اللغة (language id) بهدف تحسين مهمة POS tagging - فائدة أكبر للغات المتباعدة (SPA/HIN) مقارنةً باللغات المترابطة (EGY/LEV). نعرّف نقاط التناوب/التبديل اللغوي باعتبارها المواقع داخل الجملة التي تختلف فيها لغة الكلمتين المجاورتين. وقد لاحظنا ارتفاعاً مفاجئاً في الدقة في مدونة SPA-ENG؛ ونرجّح أنّ العامل الأساس في ذلك هو الانخفاض الملحوظ في نسبة نقاط التناوب اللغوي (~6٪)، في حين بلغت النسبة 38,78٪، 30,12٪، و 15,17٪ في مجموعات بيانات كلّ من MSA-LEV و MSA-EGY و HIN-ENG على التوالي. كما أنّ انخفاض نسبة نقاط التناوب في مدونة SPA-ENG يدفع النماذج التي تتعامل مع نمط التناوب اللغوي على مستوى الجُمْل (inter-sentential) - مثل (BiLSTM-CRF+Merged BilingualMerged) و (Bilingual CFM) و (MTL+Merged BilingualMerged Bilingual CFM) - إلى تحقيق ثاني أعلى النتائج في الدقة، إذ بلغت 95,40٪ و 95,42٪ على التوالي.

تتجلّى ميزتان محورتان لتضمينات (Merged Bilingual PseudoCS) والتضمينات المتعددة اللغات؛ الأولى أنّها تمكّن التضمينات المتعلّمة من التقاط واستيعاب التفاعلات بين الكلمات عبر لغات مختلفة، والثانية أنّها ترصد استعمال الكلمة في سياق كلّ لغة على حدة، فتقضي بذلك على اللبس المصاحب للكلمات ذات البنية السطحية المتطابقة في أكثر

من لغة. ونتيجة لذلك ينخفض كلٌّ من معدّل الكلمات غير المشمولة في القاموس (OOV) وغموض الكلمات. فعند استخدام التضمينات المتعدّدة اللغات لزوجي MSA-EGY و MSA-LEV تراجعت نسبة الـ OOV من ١٠٪ إلى ١٣٪ و ٨٪ إلى ١٠٪ على التوالي، كما انخفضت في زوجي SPA-ENG و HIN-ENG من ١٢٪ إلى ١٥٪ و ٩٪ إلى ١١٪ على الترتيب.

جدول (٤) أكثر الأخطاء شيوعاً في أفضل الأنظمة لجميع الأزواج اللغوية (Gold-POS > Predicted-POS)

MSA-EGY		MSA-LEV		SPA-ENG		HIN-ENG	
Error Type	%	Error Type	%	Error Type	%	Error Type	%
ADJ > NOUN	19%	ADJ > NOUN	21%	NOUN > ADJ	10%	NOUN > VERB	27%
VERB > NOUN	15%	VERB > NOUN	19%	NOUN > PRON	8%	VERB > NOUN	24%
NOUN > VERB	11%	NOUN > ADJ	16%	VERB > NOUN	7%	NOUN > ADJ	19%
NOUN > ADJ	8%	NOUN > VERB	9%	ADJ > PRON	5%	ADJ > VERB	19%

يُبين الجدول ٤ أكثر الأخطاء شيوعاً لدى أفضل الأنظمة في كلّ زوج لغوي. نلاحظ تشابهاً شبه تام في الأنماط الخطأ بين زوجي MSA-EGY و MSA-LEV، بينما تختلف الأخطاء الشائعة نسبياً في زوجي SPA-ENG و HIN-ENG.

أحد الأنماط الشائعة للتناوب اللغوي (CS) على مستوى العبارة (intra-sentential) التي لاحظناها في مجموعات بياناتنا هو نمط الإدراج (insertion patterns). يتلخّص هذا النمط في إدراج عنصر معجمي أو حتى مكوّن تركيبّي كامل من لغةٍ ما داخل بنية اللغة الأخرى (Muysken et al., 2000). ولتقييم وقياس أثر هذا النمط، عرّفنا ما يُسمّى بمقطع التناوب اللغوي (CSF) CS Fragment في الجمل التجريبية. وعرّفنا CSF بأنه أصغر امتداد متصل من الكلمات يظهر فيه تناوب لغوي (Soto and Hirschberg, 2018). يبلغ متوسط طول قطع التناوب اللغوي (CSF) في مجموعة الاختبار الخاصّة بـ SPA-ENG نحو ١٦,٢، بينما يصل إلى ١٦,٦ في مجموعة HIN-ENG. وفي المقابل، يبلغ متوسط الطول ١٠,٥ في مجموعة MSA-LEV و ٨,٥ في مجموعة MSA-EGY. لوحظ أن طول مقطع التبدّل يؤثّر بصورة مباشرة في أداء المصنّفات: فالمقاطع القصيرة غالباً ما تُربك أفضل النماذج في جميع أزواج اللغات. وقد تبين



أنَّ معظم الجمل التي يقتصر التبديل فيها على عنصر معجمي واحد أو عنصرين اثنين قد أُسيء تصنيفها من قِبل معظم النماذج في جميع أزواج اللغات.

يُظهر استخدام التضمينات ثنائية اللغة تفوقاً على أنظمة خط الأساس (baseline systems)، غير أنَّه لا يحقق أعلى النتائج التي سجّلتها النماذج المقترحة الأخرى في أزواج اللغات جميعها. وتشير النتائج التجريبية إلى توجّه واحد نحو تطوير تضمينات ثنائية أكثر مواءمة لمهام CS. ونعتقد أنَّ السبب في هذا الأداء يعود إلى التفاوت الكبير في توزيع المعاني (sense distribution) للكلمات متعدّدة الدلالة بين المدونة الأحادية (mono) والمدونة التي تتضمّن تمثيلات مشتركة مثل Merged Bilingual PseudoCS. فعلى سبيل المثال، يمكن لكلمة bank أن تحمل عدة معانٍ:

أ. ضفّة أو منحدر أرضي،

ب. مؤسّسة مالية،

ج. صفّ أو سلسلة من أشياء متشابهة.

أمّا في جملة أو مدونة يهيمن عليها الإسبانية، فيغلب استعمال bank إن لم يقتصر على المعنى (أ) فقط. تُبيّن تجاربنا عموماً أنَّ التضمينات الثنائية المعيارية ليست ملائمة عموماً لمهام التناوب اللغوي (CS)؛ إذ تُعطي التضمينات المتعلّمة من بيانات التناوب اللغوي (CS) أداءً أفضل، وهو ما يتّسق مع ما توصّل إليه (Pratapa et al., 2018)

٦- الخاتمة

في هذه الورقة، قدمنا دراسة مفصّلة لعددٍ من الإستراتيجيات المتعلقة بمهمّة توسيم الأجزاء النحوية (POS tagging) في بيانات التناوب اللغوي (CS) لأربعة أزواج لغوية. استقصينا أثر التضمينات المسبقة التدريب (pretrained embeddings) على أداء POS tagging في بيانات التناوب اللغوي (CS). فتبيّن أنَّ الأزواج اللغوية المتقاربة (مثل MSA-EGY و MSA-LEV) تستفيد من التعلّم المشترك لمهمّة POS tagging إضافةً إلى استخدام التضمينات المدمجة متعدّدة اللغات (merged multilingual embeddings)، أو ما يُعرّف بتضمينات

التضمين المحوري (pivot embedding)؛ في المقابل، تستفيد الأزواج اللغوية المتباعدة (مثل SPA-ENG و HIN-ENG) من نموذج تعلّم متعدّد المهام (multi-task learning model) يتعلّم مهمّتين مختلفتين، كـ POS tagging وتحديد هوية اللغة (language identification).

علاوةً على ذلك، أجرينا مقارنةً بين نتائجنا وأحدث النماذج المنشورة لتوسيم الأجزاء النحوية في الأزواج اللغوية SPA-ENG و HIN-ENG و MSA-EGY، فبيّنا تفوّق مصنّفاتنا على النظام المتقدّم في MSA-EGY في كل الإعدادات (AlGhamdi et al., 2016)، بينما جاءت دقّة نموذج Soto and Hirschberg BiLSTM-CRF+Merged Bilingual PseudoCS مماثلة لما حققه (2018). وسنواصل العمل مستقبلاً في اتجاهين رئيسين: أولاً، دراسة الجوانب النظرية لتعلّم التضمينات، وثانياً اختبار وتوظيف فعالية التضمينات المقترحة في تطبيقات أخرى من مجالات المعالجة اللغوية الطبيعية، مثل التقسيم النحوي (segmentation) والتحليل النحوي التركيبي (parsing).



المراجع الأجنبية

- Adams, Oliver., Makarucha, Adam., Neubig, Graham., Bird, Steven. & Cohn, Trevor. (2017 April 3-7). *Cross-lingual word embeddings for low-resource language modeling*. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain.
- Aguilar, Gustavo., AlGhamdi, Fahad., Soto, Victor., Diab, Mona., Hirschberg, Julia. & Solorio, Thamar. (2018). Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, Julia Hirschberg (Chairs), *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Association for Computational Linguistics, Melbourne, Australia.
- Al-Badrashiny, Mohamed., Elfardy, Heba., & Diab, Mona T. (2015 July 30-31). *Aida2: A hybrid approach for token and sentence level dialect identification in arabic*. In Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China.
- AlGhamdi, Fahad. & Diab, Mona. (2018 May 7-12). *Wasa: A web application for sequence annotation*. In Proceedings of the 11th Language Resources and Evaluation Conference, Miyazaki, Japan.
- AlGhamdi, Fahad., Molina, Giovanni., Diab, Mona T., Solorio, Thamar., Hawwari, Abdelati., Soto, Victor. & Hirschberg, Julia. (2016 November 1). Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching@EMNLP 2016*, Austin, Texas, USA.
- Bali, Kalika., Sequiera, Royal. & Choudhury, Monojit. (2015 December 11-14). *POS tagging of Hindi-English Code Mixed Text from Social Media: Some Machine Learning Experiments*. In Proceedings of International Conference on NLP. NLP AI, Trivandrum, India.
- Barman, Utsab., Wagner, Joachim., & Foster, Jennifer. (2016 November 1). Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In

*Proceedings of the Second Workshop on Computational Approaches to Code Switching@
EMNLP 2016, Austin, Texas, USA.*

Bhat, Irshad Ahmad., Bhat, Riyaz Ahmad., Shrivastava, Manish., & Sharma, Dipti Misra. (2018). *Universal dependency parsing for Hindi-English codeswitching*. arXiv preprint, arXiv:1804.05868.

Bhat, Irshad., Bhat, Riyaz A., Shrivastava, Manish., & Sharma, Dipti. (2017 April 3-7). *Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data*. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain.

Bhat, Riyaz Ahmad., Bhatt, Rajesh., Farudi, Annahita., Klassen, Prescott., Narasimhan, Bhuvana., Palmer, Martha., Rambow, Owen., Sharma, Dipti Misra., Vaidya, Ashwini., Vishnu, Sri Ramagurumurthy., & Xia, Fei. *The hindi/urdu treebank project*. In *Handbook of Linguistic Annotation*. In Nancy Ide, James Pustejovsky (Eds), Handbook of Linguistic Annotation (pp 659–697). Springer Press.

Blunsom, Phil. and Hermann, Karl Moritz. (2014). Multilingual models for compositional distributional semantics. arXiv preprint, arXiv: 1404.4641.

Bojanowski, Piotr., Grave, Edouard., Joulin, Armand., & Mikolov, Tomas. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Chandu, Khyathi., Manzini, Thomas., Singh, Sumeet. & Black, Alan W. (2018 July 19). Language informed modeling of code-switched text. In Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, Julia Hirschberg (Editors), *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Melbourne, Australia.

Chen, Hongshen., Zhang, Yue., & Liu, Qun. (2016 November 1-5). *Neural network for heterogeneous annotations*. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas.



- Chollet, François. et al. (2015). *Keras*. <https://github.com/fchollet/keras>.
- Conneau, Alexis., Lample, Guillaume., Ranzato, Marc'Aurelio., Denoyer, Ludovic., & J'égou, Hervé. (2017). *Word translation without parallel data*. arXiv preprint, arXiv:1710.04087.
- Cotterell, Ryan. & Callison-Burch, Chris. (2014 May). *A multi-dialect, multi-genre corpus of informal written Arabic*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland
- Danet, Brenda. & Herring, Susan C. (2007). *The multilingual Internet: Language, culture, and communication online*. Oxford University Press.
- Diab, Mona., Ghoneim, Mahmoud., AlGhamdi, Fahad., AlMarwani, Nada., & Al-Badrashiny, Mohamed. (2016 May 23-28). *Creating a large multi-layered representational repository of linguistic code switched arabic data*. In 10th International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia.
- Diab, Mona., Habash, Nizar., Rambow, Owen. & Roth, Ryan. (2013). *LDC Arabic treebanks and associated corpora: Data divisions manual*. arXiv preprint, arXiv:1309.5652.
- Diab, Mona., Habash, Nizar., Rambow, Owen. Altantawy, Mohamed. & Benajiba, Yassine. (2010 May 17-23). COLABA: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, Malta.
- Eskander, Ramy., Habash, Nizar., Rambow, Owen. & Pasha, Arfath. (2016 December 11-16). *Creating resources for dialectal Arabic from a single annotation: A case study on egyptian and levantine*. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan.
- Faruqui, Manaal. & Dyer, Chris. (2014 April 26-30). *Improving vector space word representations using multilingual correlation*. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden.
- Jamatia, Anupam., Gambäck, Björn., & Das, Amitava. (2015 Sep 7-9). *Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages*. Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria.

Jarrar, Mustafa., Habash, Nizar., Alrimawi, Faeq., Akra, Diyam. & Zalmout, Nasser. (2017).

Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

M´onica Stella C´ardenas-Claros and Neny Isharyanti. (2009). Code-switching and code-mixing in internet chatting: Betweenyes, 'yes', 'ya', and 'si' a case study. *The Jalt Call Journal*, 5(3):67-78.

Maamouri, Mohamed., Bies, Ann., Buckwalter, Tim. & Mekki, Wigdan. (2004). The penn Arabic treebank: Building a large-scale annotated Arabic corpus. In NEMLAR conference on Arabic language resources and tools, Cairo.

Maamouri, Mohamed., Bies, Ann., Kulick, Seth., Tabessi, Dalila. & Krouna, Sondas. (2012). *Egyptian Arabic Treebank pilot*. Linguistic Data Consortium.

McDonald, Ryan., Nivre, Joakim., Quirmbach-Brundage, Yvonne., Goldberg, Yoav., Das, Dipanjan., Ganchev, Kuzman., Hall, Keith., Petrov, Slav., Zhang, Hao., Tackstrom, Oscar., Bedini, Claudia., Castello, Nuria Bertomeu. & Lee, Jungmee. (2013 4 – 9 August). *Universal dependency annotation for multilingual parsing*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria.

Meftouh, Karima., Harrat, Salima., Jamoussi, Salma., Abbas, Mourad. & Smaili, Kamel. (2015 October 30- November 1). *Machine translation experiments on padic: A parallel Arabic dialect corpus*. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China.

Mikolov, Tomas., Chen, Kai., Corrado, Greg. & Dean, Jeffrey. (2013) *Efficient estimation of word representations in vector space*. arXiv preprint, arXiv:1301.3781.

Mikolov, Tomas., Le, Quoc V. & Sutskever, Ilya. (2013) *Exploiting similarities among languages for machine translation*. arXiv preprint, arXiv:1309.4168.

Mikolov, Tomas., Sutskever, Ilya., Chen, Kai., Corrado, Greg S. & Dean, Jeff. (2013) Distributed representations of words and phrases and their compositionality. In C.J. Burges and L.



Bottou and M. Welling and Z. Ghahramani and K.Q. Weinberger (Eds), *Advances in neural information processing systems*, (pp. 3111–3119), Bradford Books.

Molina, Giovanni., AlGhamdi, Fahad., Ghoneim, Mahmoud., Hawwari, Abdelati., ReyVillamizar, Nicolas., Diab, Mona. & Solorio, Tamar. (2016). *Overview for the second shared task on language identification in code-switched data*. In Proceedings of the Second Workshop on Computational Approaches to Code Switching@EMNLP 2016, Austin, Texas, USA.

Muysken, Pieter. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Palmer, Martha., Bhatt, Rajesh., Narasimhan, Bhuvana., Rambow, Owen., Sharma, Dipti Misra. & Xia, Fei. (2009). *Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure*. In Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India.

Parker, Robert., Graff, David, Chen, Ke., Kong, Junbo. & Maeda, Kazuaki. (2011). *Arabic Gigaword fifth edition LDC2011t11*. Linguistic Data Consortium.

Pennington, Jeffrey., Socher, Richard. & Manning, Christopher. (2014 October 25-29). *Glove: Global vectors for word representation*. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar.

Pratapa, Adithya., Choudhury, Monojit. & Sitaram, Sunayana. (2018 October 31-November 4). *Word embeddings for code-mixed language processing*. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium.

Reimers, Nils. & Gurevych, Iryna. (2017 September 7-11). *Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging*. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark.

Silveira, Natalia., Dozat, Timoth., Marneffe, Marie-Catherine de., Bowman, Samuel., Connor, Miriam., Bauer, John. & Manning, Chris. (2014 May). *A Gold Standard Dependency Corpus for English*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). Reykjavik, Iceland.

- Solorio, Thamar & Liu, Yang. (2008 October 25-27). *Part-of-speech tagging for English-Spanish code-switched text*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii
- Solorio, Thamar., Blair, Elizabeth., Maharjan, Suraj., Bethard, Steven., Diab, Mona., Gohneim, Mahmoud., Hawwari, Abdelati., AlGhamdi, Fahad., Hirschber, Julia., Chang, Alison. & Fung, Pascale. (2014 October 25). Overview of the first shared task on language identification in code-switched data. In Mona Diab, Julia Hirschberg, Pascale Fung, Thamar Solorio (Eds.), *First Workshop on Computational Approaches to Code Switching*, EMNLP-2014, Doha, Qatar.
- Soto, Victor. & Hirschber, Julia. (2017 August 20-24). *Crowdsourcing Universal Part-Of-Speech tags for CodeSwitching*. In Interspeech 2017, Stockholm, Sweden
- Soto, Victor. & Hirschberg, Julia. (2018 July 19). Joint part-of speech and language id tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic CodeSwitching*, Melbourne, Australia
- Vulic, Ivan. & Moens, Marie-Francine. (2016). Bilingual distributed word representations from documentaligned comparable data. *Journal of Artificial Intelligence Research*, 55, 953–994.
- Vyas, Yogarshi., Gella, Spandana., Sharma, Jatin., Bali, Kalika., & Choudhury, Monojit. (2014 October 25-29) *Pos tagging of English-Hindi code-mixed social media content*. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.
- Zaidan, Omar F. & Callison-Burch, Chris. (2011 June 19-24). *The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA.
- Zhang, Yuan. & Weiss, David. (2016). *Stack-propagation: Improved representation learning for syntax*. arXiv preprint, arXiv:1603.06598.
- Zou, Will Y., Socher, Richard., Cer, Daniel. & Manning, Christopher D. (2013 October 18-21). *Bilingual word embeddings for phrase-based machine translation*. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA.



نبذة تعريفية عن الباحثين

Author Biodata 1: Dr. Fahad Ali Rashed AlGhamdi is an Assistant Professor of Natural Language Processing in the Department of Computer Science, College of Computing and Information Technology, Al-Baha University, Saudi Arabia. He earned his PhD in Computer Science from The George Washington University (USA) in 2020. His research interests center on Natural Language Processing (NLP). He works on linguistic code-switching analysis and examines social-media data using machine-learning and deep-learning techniques to understand text and extract relevant information. He also focuses on low-resource languages and multilingual or cross-lingual processing, with particular emphasis on training and deploying Large Language Models (LLMs) to enhance interaction among different languages and advance linguistic understanding.	بيانات الباحث ١: فهد علي راشد الغامدي، أستاذ مساعد في (معالجة اللغات الطبيعية) في (قسم علوم الحاسبات بكلية الحاسبات والمعلومات في جامعة الباحة) بالمملكة العربية السعودية، حاصل على درجة الدكتوراة في علوم الحاسب من جامعة جورج واشنطن الأمريكية عام ٢٠٢٠م تدور اهتماماته البحثية حول معالجة اللغة الطبيعية (NLP). حيث يعمل على تحليل التبدل اللغوي في اللسانيات، وتحليل وسائل البيانات المستخرجة من مواقع التواصل الاجتماعي مستخدماً تقنيات تعلم الآلة والتعلم العميق لفهم النصوص واستخراج المعلومات ذات الصلة. كما أنه يعمل بمعالجة اللغات ذات الموارد المحدودة، والمعالجة اللغوية المتعددة اللغات والمتقاطعة مع التركيز على تدريب واستخدام النماذج اللغوية الكبيرة (LLMs) لتحسين التفاعل بين اللغات المختلفة وتعزيز الفهم اللغوي.
معرف أوركيد (ORCID) : 0000-0002-4161-2658	
Email: fghamdi@bu.edu.sa	

Author Biodata 2:

Prof. Mona T. Diab is a Professor of Natural Language Processing at the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, United States. She earned her PhD in Computational Linguistics from the University of Maryland, College Park, in 2003. Prof. Mona T. Diab is a leading researcher in Natural Language Processing (NLP) whose work spans several core areas of the field. Her research focuses on semantic representation (lexical semantics and meaning in context), cross-lingual and low-resource language processing, and computational sociolinguistics, particularly as they relate to Arabic and other linguistically diverse contexts en.wikipedia.org lti.cs.cmu.edu . She is especially known for advancing Arabic NLP and multilingual systems, emphasizing the rich diversity of dialects rather than treating languages as monolithic entities cmu.edu lti.cs.cmu.edu . In recent years, Prof. Diab has also placed a strong emphasis on fairness and ethics in AI, aiming to develop responsible language technologies that are culturally aware and inclusive

بيانات الباحث ٢:

منى دياب، أستاذة معالجة اللغات الطبيعية في (معهد تقنيات اللغة) في (كلية علوم الحاسب)، جامعة كارنيجي ميلون، الولايات المتحدة الأمريكية. حصلت على درجة الدكتوراه في اللسانيات الحاسوبية من جامعة ميريلاند (كوليدج بارك) عام ٢٠٠٣. تُعد البروفيسورة دياب من أبرز الباحثين في مجال معالجة اللغات الطبيعية (NLP)؛ إذ تمتد إسهاماتها إلى مجالات جوهرية عدة، أبرزها: تمثيل الدلالة (الدلالات المعجمية ومعنى الكلمة في السياق)، المعالجة بين اللغات ومعالجة اللغات محدودة الموارد، إضافة إلى اللسانيات في وسائل التواصل الاجتماعي، ولا سيّما في سياق اللغة العربية واللغات المتنوعة الأخرى. تُعرف بدورها الريادي في تطوير تقنيات NLP العربية والأنظمة المتعددة اللغات، مع إبراز ثراء اللهجات بدلاً من النظر إلى اللغات ككيانات أحادية. وفي السنوات الأخيرة، أولت البروفيسورة دياب اهتماماً كبيراً بقضايا العدالة والأخلاقيات في الذكاء الاصطناعي، ساعية إلى ابتكار تقنيات لغوية مسؤولة تراعي البعد الثقافي وتُعزز الشمولية.

معرف أوركيد (ORCID) : 0000-0002-7696-1436

Email: mtdiab@gmail.com



هذه الطبعة
إهداء من المجمع
ولا يسمح بنشرها ورقياً أو تداولها تجارياً