



اللغويات في المحتوى الديني والمعجمية الآلية للغات العربية

Journal of Computational Linguistics & Arabic Language Processing



تعزيز الدقة في المحتوى الديني المولّد بواسطة النماذج اللغوية الكبيرة
من خلال استراتيجيات كشف التحيز والتخفيف منه

الرصد الآلي للمتلازمات المعجمية في المدونة النصية العربية:
مقاربة معجمية حاسوبية

النماذج اللغوية الضخمة ومعالجة اللغة العربية: مراجعة منهجية

استخدام منهجية لسانيات المدونات الحاسوبية في تحليل اللغة
الروائية دراسة تطبيقية على الرواية السعودية

السنة الثانية - العدد الرابع
جمادى الأولى ١٤٤٧ هـ - أكتوبر ٢٠٢٥ م
Vol. 2, No. 4, Oct, 2025

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



اللغويات الحاسوبية والمعالجة اللغوية الآلية

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. محسن عبدالرزاق رشوان (مصر)

أ.د. إيريك أتويل (بريطانيا)

أ.د. نزار حبش (أمريكا)

أ.د. مصطفى عرسان جرار (فلسطين)

أ.د. عبدالمحسن عبيد الثبيتي (السعودية)

أ.د. خالد فؤاد شعلان (مصر)

أ.د. صلاح راشد الناجم (الكويت)

د. عبد الرحمن العصيمي

مدير التحرير

د. هاجر القبشي

أمينة المجلة



◆ قواعد النشر بالمجلة:

سياسات النشر:

تنشر مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيّة البحوث الأصلية التي لم يسبق نشرها، والمكتوبة بالُّغة العربيّة أو الإنجليزيّة، والدراسات والمُلخصات العلميّة المتخصصة في مجال حوسبة اللُّغة، كاللُّغويات الحاسوبية، والمعالجة الآلية للُّغة العربيّة، وكل ما له صلة بحوسبة اللُّغة.

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 المحكّمين لتقييمها سريعاً، وتبذل هيئة التحرير طاقتها في التواصل مع المحكّمين 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 المؤلّف عند نشر ورقته.



Journal information:

◆ بيانات المجلة:

مجلة اللُّغويات الحاسوبية والمعالجة الآلية للُّغة العربيّة Journal of Computational Linguistics & Arabic Language Processing

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

مجمع الملك سلمان العالمي للُّغة العربيّة King Salman Global Academy for Arabic Language

الرياض - المملكة العربيّة السعوديّة Riyadh- Saudi Arabia

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

ردمدم النسخة الإلكترونيّة: 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

تعزيز الدقة في المحتوى الديني المؤلّد بواسطة
النماذج اللغوية الكبيرة من خلال استراتيجيات كشف
التحيز والتخفيف منه

أ.د. فوزي حراق

٥٢

الرصد الآلي للمتلازمات المعجمية في المدونة النصية
العربية: مقارنة معجمية حاسوبية

د. عواطف السمعلي الزواري

٨٩

النماذج اللغوية الضخمة ومعالجة اللغة العربية:
مراجعة منهجية

د. خليوي سامر خليوي العياضي

١٤٤

استخدام منهجية لسانيات المدونات الحاسوبية في تحليل
اللغة الروائية دراسة تطبيقية على الرواية السعودية

د. طارق بن ربيع خلف الفريدي

كلمة التحرير

عزيزي القارئ/القارئة - عزيزي الباحث/الباحثة:

يُسعدنا أن نرحّب بك في العدد الرابع من المجلة العلمية المحكّمة: «اللُّغويات الحاسوبية والمعالجة الآليّة للغة العربيّة»، الصادرة عن مجمع الملك سلمان العالمي للغة العربيّة، والمتخصّصة في مجال حوسبة اللغة العربيّة.

تواصل المجلة مسيرتها العلمية الرصينة في نشر البحوث والدراسات التي تُسهم في تطوير الفهم العميق لتطبيقات الحوسبة في اللغة العربيّة، وتعمل على رصد أحدث الاتجاهات في مجال الذكاء الاصطناعي اللغوي وأدوات التحليل اللغوي الآلي. كما تُعنى المجلة باستشراف آفاق المستقبل، واستكشاف التحديات والفرص التي تواجه تطوير اللغة العربيّة رقمياً، بما يعزّز حضورها في فضاء التقنية والمعرفة.

ويأتي العدد الرابع استمراراً لهذا التوجّه العلمي البناء، حاملاً بين صفحاته أربعة أبحاث علمية متميزة تعالج موضوعات معاصرة ومتنوعة في صميم حوسبة اللغة العربيّة:

يأتي البحث الأول بعنوان "تعزيز الدقة في المحتوى الديني المولّد بواسطة النماذج اللغوية الكبيرة من خلال استراتيجيات كشف التحيز والتخفيف منه"، وهو بحث يتناول قضية حساسة تتعلق بمصداقية النماذج اللغوية في إنتاج المعرفة الدينية، ويقترح آليات علمية لضبط الجودة وتقليل الانحرافات المفهومية.

أما البحث الثاني "الرصد الآلي للمتلازمات المعجمية في المدونة النصية العربية: مقارنة معجمية حاسوبية"، فهو يستعرض أساليب تحليل العلاقات المعجمية باستخدام الخوارزميات والإحصاءات النصية، بما يثري صناعة المعاجم الذكية.

ويأتي البحث الثالث "النماذج اللغوية الضخمة ومعالجة اللغة العربية: مراجعة منهجية"؛ ليرصد أبرز الجهود العالمية والعربية في تطوير النماذج اللغوية، ويقدم قراءة نقدية لتحدياتها في البيئة العربية.

وأخيراً يمنح البحث الرابع "استخدام منهجية لسانيات المدونات الحاسوبية في تحليل اللغة الروائية: دراسة تطبيقية على الرواية السعودية" بين الأدب والحوسبة، مستعرضاً كيف يمكن للبيانات النصية الضخمة أن تفتح آفاقاً جديدة لتحليل الخطاب الروائي.

نُعرب عن بالغ الشكر والتقدير للباحثين والمحكمين الذين أسهموا في إثراء هذا العدد بجهودهم العلمية وملاحظاتهم البناءة. ونأمل أن يجد القارئ الكريم في هذا الإصدار مادة علمية ثرية تُحفّز على مزيدٍ من البحث والاكتشاف في ميدان الحوسبة اللغوية.

ونسعد دائماً بتواصلكم ومشاركاتكم، وبآرائكم التي تُعيننا على تطوير المجلة والارتقاء بمحتواها.

نسأل الله أن يجعل هذا العمل خالصاً لوجهه الكريم، وأن ينفع به الباحثين والمهتمين بحوسبة اللغة العربية.

هيئة التحرير

تعزيز الدقة في المحتوى الديني المولّد بواسطة النماذج اللغوية الكبيرة من خلال استراتيجيات كشف التحيز والتخفيف منه

المعرف الرقمي: <https://doi.org/10.60161/2521-002-004-001>

أ.د. فوزي حراق

قسم الحاسب الآلي، كلية العلوم، جامعة فرحات عباس، سطيف، الجزائر

استقبل في	رُوجع في	قُبِل في	نُشر في
١٤٤٧/٠٣/١٨	١٤٤٧/٠٣/١٩	١٤٤٧/٠٤/٢٣	١٤٤٧/٠٥/٠٩

المخلص

تظهر النماذج اللغوية الكبيرة (LLMs) تميّزاً في توليد النصوص عبر مجالات متعددة، إلا أنّ التحديات لا تزال قائمة عند التعامل مع المحتوى الثقافي والديني، وخاصة في السياقات الإسلامية. وتهدف الدراسة إلى تقييم أداء ثلاثة نماذج لغوية بارزة (ChatGPT، وIslamGPT، وJais) في الاستجابة للاستفسارات المتعلقة بالإسلام، مع التركيز على التخفيف من إنتاج المعلومات غير الدقيقة أو غير الملائمة ثقافياً، والمعروفة بالهلوسة. ونقترح إطاراً جديداً للكشف والتصحيح يدمج بين التقييم الكمي والنوعي باستخدام تقنيات التوليد المعزز بالاسترجاع (RAG) ومقاييس التشابه المتقدمة، لتحسين دقة الاستجابات وتقليل التحيز. وتُظهر نتائجنا تحسناً ملحوظاً في الموثوقية؛ حيث أدى الإطار المقترح إلى تقليص المحتوى المهلوس وتعزيز الأصالة الثقافية للمخرجات. ويمثل هذا خطوة نحو الذكاء الاصطناعي الحساس ثقافياً؛ مما يعزّز الموثوقية في التطبيقات المعتمدة على الذكاء الاصطناعي عبر سياقات متنوعة.

الكلمات المفتاحية

كشف التحيز، استراتيجيات التخفيف، النماذج اللغوية الكبيرة، المحتوى الديني، الهلوسة.

Enhancing Religious Content Generation in Large Language Models via Bias Detection and Mitigation

المعرف الرقمي: <https://doi.org/10.60161/2521-002-004-001>

Prof. Fawzi Harrak

Department of Computer Science, Faculty of Science, Ferhat Abbas University, Setif, Algeria

Received on	Revised on	Accepted on	Published on
10-9-2025	11-09-2025	15-10-2025	31-10-2025

ABSTRACT

Large language models (LLMs) have shown remarkable capabilities in generating text across domains, yet challenges persist when addressing specific cultural and religious content, particularly within Islamic contexts. This study evaluates the performance of three prominent LLMs—ChatGPT, IslamGPT, and Jais—on Islamic-related queries, with a focus on mitigating the generation of inaccurate or culturally insensitive information, known as hallucinations. We propose a novel detection and correction framework, integrating both quantitative and qualitative assessments through techniques such as Retrieval-Augmented Generation (RAG) and advanced similarity metrics, to enhance response accuracy and reduce bias. Our findings highlight significant improvements in reliability, with the proposed framework reducing hallucinated content and enhancing the cultural fidelity of generated outputs. This approach represents a step toward culturally sensitive AI, fostering trustworthiness in AI-driven applications across diverse contexts.

KEYWORDS

Bias Detection, Mitigation Strategies, Large Language Models (LLMs), Religious Content, Hallucination.



١- المقدمة

لقد أدّت التطورات الأخيرة في مجال الذكاء الاصطناعي (AI)، ومعالجة اللغة الطبيعية (NLP) إلى التطوير السريع للنماذج اللغوية الكبيرة؛ فقد أحدثت هذه النماذج تحولاً جذرياً في تطبيقات معالجة اللغة؛ حيث مكّنت من قدرات متقدمة في توليد النصوص عبر مجالات متنوعة، (براون وآخرون، ٢٠٢٠؛ شوداري وآخرون، ٢٠٢٣).

ومع ذلك، ورغم أدائها اللافت، ما تزال النماذج اللغوية المعاصرة تعاني من قيود منهجية تظهر في شكل أخطاء واقعية، وانحيازات ثقافية، ومحتوى مهلوس، وهي ظواهر تصبح أكثر وضوحاً عند تعامل هذه النماذج مع موضوعات حساسة ثقافياً وديناً، (إيدنجر وآخرون، ٢٠٢١؛ بندر وآخرون، ٢٠٢١).

ويُعرّف مفهوم (الهلوسة) في النماذج اللغوية الكبيرة بأنه: توليد لمعلومات تبدو مُقنعة من الناحية اللغوية، لكنها غير صحيحة أو غير مدعومة بالواقع، وهو ما يمثل تحدياً أساسياً أمام نشر أنظمة ذكاء اصطناعي موثوقة، (جي وآخرون، ٢٠٢٣؛ شائق وآخرون، ٢٠٢٣).

وتزداد خطورة هذه الظاهرة في السياقات الدينية؛ حيث إن التمثيل الدقيق للمفاهيم العقدية والأحداث التاريخية والممارسات الثقافية يحمل انعكاسات بالغة الأهمية على ثقة المجتمعات وتماسكها الاجتماعي، (جاهمان وآخرون، ٢٠٢١؛ عبيد وآخرون، ٢٠٢٠).

ويطرح المحتوى الإسلامي تحديات خاصة للنماذج اللغوية الحالية نتيجة عدة عوامل:

أولاً: أن التعقيد اللغوي للنصوص العربية وتنوع أشكال الترجمة الصوتية يمكن أن يؤدي إلى سوء تفسير المصطلحات الدينية، (عبد المجيد والمداني، ٢٠٢٢).

ثانياً: أن عدم توازن بيانات التدريب وهيمنة المتون الغربية يُسهم في التشويه المنهجي للمفاهيم والممارسات الإسلامية، (بلودقات وآخرون، ٢٠٢٢؛ وآخرون، ٢٠٢٠).

وقد وثّقت دراسات سابقة حالات ربطت فيها النماذج اللغوية الاستعلامات المتعلقة بالإسلام، بمشاعر سلبية أو أجوبة تتعارض مع المرجعية الشرعية الموثوقة، (عبيد وآخرون، ٢٠٢١؛ نديم وآخرون، ٢٠٢٠).

علاوة على ذلك، فإن ميل هذه النماذج إلى اختلاق الاستشهادات، أو تقديم تفسيرات افتراضية على أنها توجيهات دينية موثوقة - يُشكّل مخاطر جدية على المستخدمين الذين يبحثون عن معلومات دقيقة حول المبادئ والممارسات الإسلامية، (أم سي فارلان والقيسي، ٢٠٢٣).

وللتعامل مع هذه القيود الجوهرية، يعرض هذا البحث إطاراً شاملاً لتقييم التحيز والهلوسة، والتخفيف منهما في توليد المحتوى الإسلامي، عبر معماريات متعددة للنماذج اللغوية الكبيرة.

وتستند دراستنا إلى فحص منهجي لثلاثة نماذج بارزة، وهي:

- (ChatGPT) إصدار 3.5-GPT من OpenAI.
 - (IslamGPT) نموذجاً متخصصاً في المعرفة الإسلامية، المُدرَّب على مصادر دينية موثوقة ومفتوحة).
 - (Jais) النموذج اللغوي الكبير المتمركز حول اللغة العربية، والمطوّر من طرف MBZUA و Inception.
- وتوفّر هذه الدراسة المقارنة رؤى مهمة، حول كيفية تأثير بنية النموذج ومنهجيات التدريب والتخصّص الثقافي في الأداء، عند التعامل مع المحتوى الديني الحساس.

وتتمثل الإسهامات الرئيسية لهذا العمل فيما يلي:

١. **تقييم مقارن شامل**، في إجراء تقييم منهجي لأنماط الهلوسة والحساسية الثقافية، عبر ثلاث معماريات مختلفة بدرجات متفاوتة من التخصّص في المحتوى الإسلامي.
٢. **استراتيجية تخفيف متكاملة**؛ وذلك بتنفيذ آلية تصحيح متعددة المراحل، تشمل تعزيز البيانات والاسترجاع الانتقائي من مصادر موثوقة. وقد أظهر هذا النهج تحسينات قابلة للقياس في الدقة والملاءمة الثقافية مع الحفاظ على الطلاقة والتماسك في الاستجابات.



٣. إطار منهجي للذكاء الاصطناعي الواعي ثقافياً، عن طريق وضع منهجيات تقبل إعادة التطبيق؛ لتقييم وتحسين الحساسية الثقافية في النماذج اللغوية الكبيرة؛ مما يفتح المجال لتطبيق هذه الأساليب على سياقات دينية وثقافية أخرى.

وتسعى دراستنا للإجابة عن سؤالين بحثيين أساسيين:

- سؤال البحث الأول: كيف يختلف أداء (ChatGPT و Jais Islam GPT)، من حيث ميلها إلى توليد المحتوى المهلوس عند الرد على مطالبات متعلّقة بالإسلام؟ وكيف يمكن الكشف عن هذه الأنماط وقياسها بشكل منهجي؟
- سؤال البحث الثاني: ما مدى فعالية الإطار المقترح المعزز بالاسترجاع (RAG) في التخفيف من الهلوسة والتحيز في توليد المحتوى الإسلامي، مع الحفاظ على جودة الاستجابة وأصالتها الثقافية؟

وينتظم ما تبقى من هذا البحث على النحو التالي:

القسم الثاني: يستعرض الأعمال السابقة المتعلقة بكشف الهلوسة في النماذج اللغوية، والتحيز الثقافي في أنظمة الذكاء الاصطناعي، والمقاربات المتخصصة لمعالجة المحتوى الديني.

القسم الثالث: يستعرض منهجيتنا، مع بناء مجموعة البيانات، ومقاييس التقييم، وإطار الكشف والتخفيف المقترح.

القسم الرابع: يصف التصميم التجريبي والتفاصيل التنفيذية.

القسم الخامس: يعرض النتائج والتحليلات الشاملة، بما في ذلك المقاييس الكمية والتقييم النوعي.

القسم السادس: يختتم بالمناقشة والتوصيات والقيود وآفاق البحث المستقبلي، نحو تطوير أنظمة ذكاء اصطناعي واعية ثقافياً وموثوقة.

٢- مراجعة الأدبيات

لقد أصبح تقاطع مجالات كشف التحيز، والتخفيف من الهلوسة، وتوليد المحتوى الحساس ثقافياً بواسطة النماذج اللغوية الكبيرة - محوراً بحثياً بالغ الأهمية، مع اهتمام خاص بالسياقات الدينية والثقافية. ونستعرض في هذا القسم الأدبيات السابقة عبر ثلاثة أبعاد رئيسية هي: كشف التحيز والتخفيف منه في النماذج اللغوية الكبيرة، وآليات كشف الهلوسة وتصحيحها، والمقاربات المتخصصة لمعالجة المحتوى الديني.

وقد أظهرت أبحاث حديثة وجود تحيزات منهجية في النماذج اللغوية الكبيرة، عبر أبعاد ديموغرافية وثقافية مختلفة، فقد أجرى عبيد وآخرون (٢٠٢١) دراسة رائدة كشفت عن تحيز دائم ضد المسلمين في عدد من النماذج اللغوية البارزة؛ حيث وثّقوا كيفية ربط هذه الأنظمة المصطلحات المرتبطة بالإسلام بمشاعر سلبية وعنيفة. وقد شكّل عملهم أساساً لفهم التحيز الديني في أنظمة الذكاء الاصطناعي، رغم أن تحليلهم اقتصر على مصطلح (مسلم)، ولم يتوسع فيه ليشمل السياقات الدلالية الإسلامية الأوسع أو تنوع المصطلحات. وبناءً على ذلك، اقترح الباحثون استراتيجيات للتخفيف من التحيز، عبر تقنيات هندسة المطالبات وإعادة الضبط الدقيق للنماذج، مبرهنين على تحسينات ملموسة في تقليل التحيز.

وبالموازاة مع هذا، طوّر بهنا مهادر وميليس (٢٠٢٢) إطاراً شاملاً لتحليل التحيزات الاجتماعية، عبر عدة متغيرات من (BERT) ولغات مختلفة. وقد اعتمد نهجهم على قوالب كشف الصور النمطية واختبارات الترابط؛ لقياس التحيز عبر أبعاد مثل الجنس والعرق والدين. إلا أن منهجيتهم ركّزت بالأساس على الكشف دون التخفيف، كما أن تقييمهم اقتصر على اللغات الأوروبية؛ مما يحد من قابليته للتطبيق في السياقات العربية والإسلامية.

وعلى نطاق أوسع، شهد مجال كشف التحيز تقدماً ملحوظاً في المنهجيات. فقد قدّم بلودقات وآخرون (٢٠٢٠) مسحاً نقدياً حول التحيز في معالجة اللغة الطبيعية، مؤكدين على الحاجة إلى فهم أكثر دقةً للانحياز، باعتباره ظاهرة متعددة الأبعاد تتطلب مقاربات تقييم حساسة للسياق.



وبالمثل، سلط بندر وآخرون (٢٠٢١) الضوء على المخاطر الجوهرية للنماذج اللغوية الكبيرة، مشددين على ضرورة اعتماد مناهج منهجية لكشف المخرجات الضارة والتخفيف منها، خصوصاً في المجالات الحساسة ثقافياً.

تمثّل الهلوسة في النماذج اللغوية الكبيرة تحدياً أساسياً أمام نشر أنظمة ذكاء اصطناعي موثوقة، وخاصة في المجالات المعرفية المكثفة. وقد اعتمدت أحدث مناهج الكشف عن الهلوسة على مقاربات متعددة، إلا أن الكثير منها يواجه قيوداً في الموثوقية والنطاق. على سبيل المثال، قدّم سادات وآخرون (٢٠٢٣) إطاراً باسم (DelucionQA)؛ لكشف الهلوسة في أنظمة الإجابة عن الأسئلة المتخصصة. ويجمع هذا الإطار بين التحقق من التناسق والاستعانة بمصادر معرفية خارجية، وقد حقق نتائج واعدة عبر عدة مجالات. غير أن اعتمادهم على (ChatGPT) كآلية تحقق يثير إشكالية (الدائرية)، حيث إن نظام التحقق ذاته قد يكون عرضة للهلوسة. وبالمثل، اقترح تشن وآخرون (٢٠٢٤) إطاراً موحداً لكشف الهلوسة في النماذج اللغوية الكبيرة متعددة الوسائط. ورغم ما يقدمه نهجهم من ابتكار في التحقق عبر الوسائط المتعددة، إلا أن اعتمادهم مجدداً على نماذج لغوية كآلية تحقق قد يضخم المشكلة بدلاً من معالجتها.

وعلى النقيض، أظهرت استراتيجيات التخفيف تطورات أكثر إيجابية. فقد استكشف باتيل وآخرون (٢٠٢٣) إمكانية بناء نماذج لغوية كبيرة متخصصة تتوافق مع الرؤية الإسلامية، مقترحين تعديلات معمارية واستراتيجيات تدريب تقلل من الهلوسة في السياقات الدينية. ويبرهن عملهم على إمكانيات تطوير نماذج متخصصة، رغم أن تقييمهم اقتصر على تحليل أولي محدود.

وقد برز التوليد المعزز بالاسترجاع (RAG) كأبرز مقاربة للتخفيف من الهلوسة، (جو وآخرون ٢٠٢٠؛ لويس وآخرون، ٢٠٢٣). فمن خلال إسناد عملية التوليد إلى معلومات مُسترجعة وموثوقة، يمكن تقليل المحتوى المختلق بشكل كبير.

ومع ذلك، يشير تشانغ وآخرون (٢٠٢٣) إلى أن فعالية (RAG) تعتمد بدرجة حاسمة على جودة الاسترجاع وموثوقية المصادر، خصوصاً في المجالات المتخصصة مثل الدراسات الإسلامية.

لقد حظيت تطبيقات النماذج اللغوية الكبيرة في معالجة المحتوى الديني باهتمام متزايد؛ نظراً لخطورة النتائج المترتبة على الدقة والحساسية الثقافية في هذا المجال. فقد تناول عدد من الباحثين بشكل مباشر مسألة توليد المحتوى الإسلامي ومعالجته. فعلى سبيل المثال، طور آلان وآخرون (٢٠٢٤) نظام (MufassirQAS) للإجابة عن الأسئلة بالاعتماد على (RAG)، المصمّم خصيصاً للمعرفة الإسلامية. ويجمع هذا النظام بين النصوص الإسلامية الكلاسيكية وتقنيات معالجة اللغة الحديثة، محققاً دقةً أفضل في الإجابة عن الأسئلة الدينية، ولكن إطار تقييمهم افتقر إلى تحليل شامل للأنحياز؛ حيث ركّز أساساً على الدقة الواقعية دون مراعاة الكفاية الثقافية.

أما من الناحية اللغوية، فقد قدّم عبد العظيم وآخرون (٢٠٢٣) نظام التوليد المعزز بالاسترجاع باللغة العربية (ARAG)؛ وذلك بالتركيز على التمثيلات الدلالية لمعالجة النصوص العربية. ويسد عملهم فجوة أساسية في دعم اللغة العربية ضمن أنظمة (RAG)، رغم أن تطبيقاتهم بقيت محصورة في المحتوى العربي العام لا الديني تحديداً.

ومن زاوية أخرى، درس طلعت وآخرون (٢٠٢٢) مسألة تقييم التحيز في البيانات متعددة اللغات، مبرزين تعقيدات قياس التحيز عبر سياقات لغوية وثقافية مختلفة، ومؤكدين على الحاجة إلى أطر تقييم خاصة بالثقافات.

ورغم هذه التقدمات، ما تزال هناك فجوات كبيرة في الأدبيات، وهي:

أولاً: أن معظم الأعمال الحالية تعالج كشف التحيز أو التخفيف من الهلوسة بشكل منفصل، مع وجود قلة من الدراسات التي تدمج التحديين معاً.

ثانياً: اختلاف منهجيات التقييم اختلافاً كبيراً عبر الدراسات؛ حيث إن بعضها يفتقر إلى أطر كمية شاملة مما يصعب المقارنة، (آلان وآخرون، ٢٠٢٤؛ عبيد وآخرون، ٢٠٢١).

ثالثاً: أن التغطية اللغوية ما تزال محدودة؛ إذ يتركز أغلب العمل على النماذج الإنجليزية، حتى عند تناول موضوعات تهم المجتمعات الناطقة بالعربية. ويُعد عمل



عبد العظيم وآخرون (٢٠٢٣) استثناءً مهماً، لكنه يظل بعيداً عن التحديات الخاصة بالنصوص الدينية.

وأخيراً، فإن العديد من استراتيجيات التخفيف تعتمد على آليات تحقق غير موثوقة، نحو استخدام نموذج لغوي للتحقق من مخرجات نموذج آخر (تشن وآخرون، ٢٠٢٤؛ سادات وآخرون، ٢٠٢٣)؛ مما يثير مخاوف منهجية حول الموثوقية.

ويعالج بحثنا هذه المحدوديات من خلال تقديم إطار متكامل يدرس بشكل متزامن كشف التحيز والتخفيف من الهلوسة في توليد المحتوى الإسلامي.

وعلى خلاف الأعمال السابقة، فإننا نوفر تقييماً مقارناً شاملاً، عبر معماريات متعددة للنماذج، وذات درجات متفاوتة من التخصص الإسلامي، وذلك باستخدام مقاييس تقييم قوية تشمل التقييم الكمي والنوعي، مع تطوير منهجية هجينة للكشف، تدمج بين مقاربات تحقق متعددة لتجنّب مشكلة (الدائرية) التي سُجّلت في الأبحاث السابقة.

٣- الإطار المقترح

لمعالجة مشكلة الهلوسة والتحيز في المحتوى الإسلامي المولّد بواسطة النماذج اللغوية الكبيرة، قمنا بتطوير بنية نموذجية قوية، تتألف من وحدة كشف ووحدة تخفيف. ويوضّح هذا القسم كل مكون على حدة، بما في ذلك تقنيات الكشف، وآليات التخفيف باستخدام التوليد المعزز بالاسترجاع (ARAG)، إضافة إلى مقاييس التقييم المستعملة للتحقق من فعالية النموذج. كما نستعرض مجموعة البيانات المستخدمة، وإطار العمل المقترح بشكل منهجي في الفقرات التالية:

٣-١- مجموعات البيانات

لدعم الإطار المنهجي لهذه الدراسة، انتقينا مجموعات بيانات متخصصة من مصادر دقيقة وموثوقة، مع إعطاء الأولوية للملاءمة والأصالة والتوافق مع نطاق البحث.

وقد صُنِّفت هذه المجموعات إلى فئتين متميزتين:

٣-١-١- مجموعات الحقيقة الذهبية (Golden Truth Datasets)

تُستخدم لكشف النماذج المختارة إن كانت تُظهر تحيزاً أو مخرجات مهلوسة. وتتألف هذه المجموعات من أزواج أسئلة وأجوبة مركزة على محتوى عقدي إسلامي، على النحو التالي:

- أ. **مجموعة القرآن:** وهي مجموعة باللغة العربية أنشأها عمر السيد، وتضم أزواج أسئلة وأجوبة مستخلصة من مصدرين قرآنيين. وتغطي الأسئلة موضوعات عقدية متنوعة، بينما تتطابق الأجوبة مع نصوص قرآنية حرفية. وقد نُقِّحت هذه المجموعة بواسطة خبراء لضمان الدقة العقدية.
- ب. **مجموعة الحديث:** تضم أسئلة باللغة الإنجليزية مستوحاة من صحيح البخاري وصحيح مسلم. وأنشئت عبر دمج مطالبات من مصادر متعددة، متبوعة بعملية ترتيب تكرارية لضمان الصلة. وأجوبتها مأخوذة من أحاديث صحيحة في موضوعات عقدية وأخلاقية وفقهية.
- ج. **مجموعة الفتاوى:** استُخرجت من البوابات الرسمية لكل من دار الإفتاء السعودية ووزارة الشؤون الدينية والأوقاف الجزائرية. وهي تتضمن أحكاماً فقهية صادرة عن علماء معتمدين، بما يخرج أسئلة عقدية معاصرة وأجوبة موثوقة.

٣-١-٢- مستودعات المعرفة (Knowledge Base)

شكَّلت النصوص المرجعية في هذه المستودعات قاعدة بيانات للاسترجاع بغرض تخفيف التحيز والهلوسة، إذ توفر أساساً عقدياً راسخاً لمخرجات النموذج:

- أ. **مستودع القرآن:** ثنائي اللغة (عربي- إنجليزي)، وقد أنشأه عمران خان وأمينه عليزا، يشمل جميع الآيات القرآنية، مزودة ببيانات وصفية، كالتفسير والتعليقات الموضوعية والسياقية.
- ب. **مستودع الحديث:** أنشأه فهد وآخرون، ويجمع بين (الكتب الستة) على مدى أربعة عشر قرناً. محتوياً على النصوص العربية، والترجمات الإنجليزية، وسلاسل الرواة (الإسناد): للتحقق من الأصالة.

ج. **مستودع التفسير:** أنشأه علي زاهد راجا، ويشمل تفاسير قرآنية كلاسيكية وحديثة باللغة الإنجليزية، مرتبة آية آية، مع شروح لغوية وتاريخية، كما يتضمن روابط بين الآيات والمبادئ الفقهية والسياقات التأويلية.

ولغرض التوضيح في (جدول ١)، اعتمدنا التسميات التالية (GTQ و GTH و GTF)؛ للإشارة إلى مجموعات الحقيقة الذهبية (القرآن، الحديث، الفتاوى)، بينما ترمز KBQ و KBH و KBT إلى مستودعات المعرفة (القرآن، الحديث، التفسير على التوالي).

جدول (١) وصف مجموعات البيانات المستخدمة

مجموعة البيانات	عدد الصفوف	عدد الأعمدة	عدد الكلمات	عدد الكلمات الفريدة
GTQ	٩٧٢	٥	١٥٧٦٥٤	١٠٥٢٠
GTH	١٠٤٦٢	٥	١١٥٤٢٣٣	٢٥٢٥٣
GTF	٢٠٠	٣	٢٦٥٠٢	٦٩٩١
KBQ	٦٢٣٦	١٩	٨٣٤١٤٥	١٢٨٤٧
KBH	٢٠٩١	١٥	١٢٥٩٩٩٤	٦٩٨٥٥
KBT	٦٢٣٦	١٥	١٦٧٧٢٧٥	٢٣٥٦٦

٢-٣- بنية النموذج

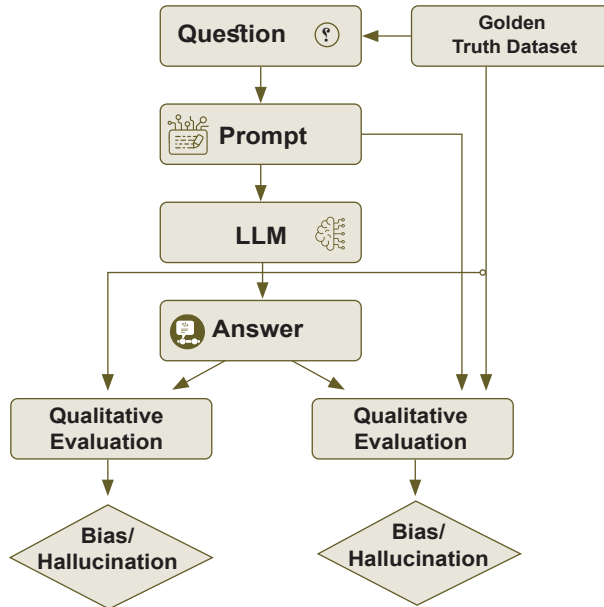
٣-٢-١- تقنيات الكشف

تركّز مرحلة الكشف في بنيتنا النموذجية، انظر (شكل ١)، على تحديد المحتوى المهلوس أو المنحاز في مخرجات النماذج اللغوية الكبيرة. وذلك بالاعتماد على مجموعة الحقيقة الذهبية المكونة من أسئلة وأجوبة، تم التحقق منها بدقة من مصادر إسلامية موثوقة، استخدمنا فيها عدداً من أساليب المطالبة لقياس أداء النماذج:

أ. المطالبة بلا أمثلة: (Zero-Shot) استخدمت لقياس أداء النماذج دون تزويدها بأمثلة مسبقة؛ بهدف رصد ميولها الأولية وتحديد خط الأداء الأساسي.

- ب. المطالبة بعدة أمثلة (Few-Shot) حيث تُزود النماذج بأمثلة قبل الإجابة. وقد ثبتت فعالية هذا الأسلوب في توجيه النماذج نحو أجوبة أكثر ارتباطاً بالسياق، عبر الاستفادة من الأسئلة والأجوبة الموجودة في مجموعة الحقيقة الذهبية.
- ج. المطالبة بالتعليمات (Instruction-Based) من خلال إدراج تعليمات صريحة مع المطالبات، فقد لاحظنا تحسناً في التزام النماذج بالمعايير العلمية الشرعية، بما يتماشى مع الممارسات المثلى التي أشارت إليها الدراسات الحديثة، (باتيل وآخرون، ٢٠٢٣).
- ولقياس دقة الأجوبة المولدة وتشابهاً مع الإجابات المرجعية، اعتمدنا على ثلاثة مقاييس، هي: تشابه جيب التمام (Cosine Similarity)، راهيتومو (٢٠١٢)، وتشابه جاكارد (Jaccard Similarity)، إيفشنكو وهونوف (١٩٩٨)، ودرجة بيرت (BERTScore)، تشانغ وآخرون (٢٠١٩). وقد مكّنتنا هذه المقاييس من تقييم التطابق التركيبي والدلالي معاً؛ لضمان شمولية الفحص عبر أبعاد متعددة من الصلة.

شكل (١) بنية نموذج الكشف المقترح



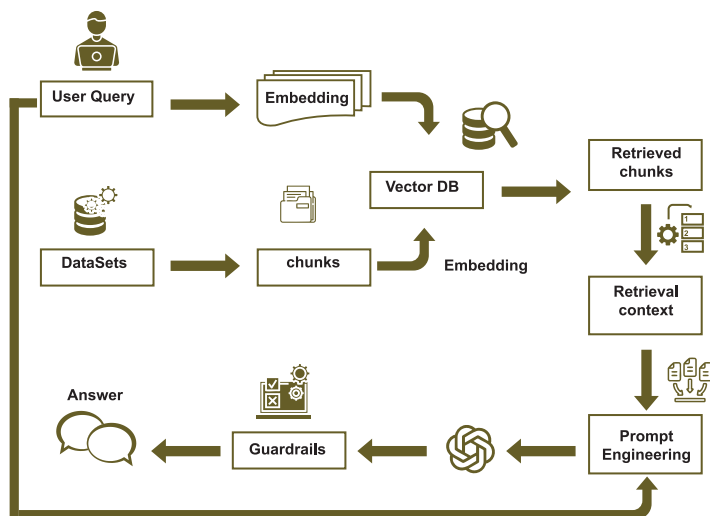
٣-٢-٢- تقنيات التخفيف

استندت وحدة التخفيف لدينا انظر (شكل ٢) أدناه إلى تقنية (RAG)؛ بهدف تحسين دقة الإجابات عبر إدماج سياق خارجي من مصادر موثوقة؛ حيث يقوم النظام باسترجاع مقاطع نصية ذات صلة من قاعدة بيانات شعاعية (vector database)، وتضم نصوصاً منظمة من مصادر إسلامية أساسية، كالقرآن والحديث. ثم تُستخدم هذه المقاطع مرجعية لبناء إجابة دقيقة، ومتسقة مع التعاليم الإسلامية.

وتبدأ عملية الاسترجاع بتحويل استعلامات المستخدم إلى متجهات (embeddings)، باستخدام نماذج لغوية متقدمة، مثل: (Universal Sentence Encoder)، و (BERT)، وهذا يتيح مطابقة دلالية فعالة داخل قاعدة البيانات.

وبعد استرجاع المقاطع الأعلى تصنيفاً، يُنشأ نموذج الإجابة اعتماداً عليها، ثم تُمرر النتيجة عبر نموذج لغوي لصياغة الاستجابة النهائية. وقد أثبت هذا الدمج بين الاسترجاع والتوليد فعاليته في تقليل الأخطاء، حيث يُخفّض من اعتماد النظام على النصوص المولدة المعرضة للتحيز، عبر ربط الإجابات بمصادر أصلية وموثوقة.

شكل (٢) بنية نموذج التخفيف المقترح



٣-٢-٣- مقاييس التقييم

اعتمدنا في عملية التقييم على فحص منهجي لجودة الإجابات في مرحلتي الكشف والتخفيف؛ حيث قورنت مخرجات النماذج القياسية مع مخرجات نموذجنا المُحسّن، باستخدام مؤشرات: تشابه جيب التمام، وتشابه جاكارد، ودرجة بيرت لقياس الجودة. كما أجرى مقيمون بشريون مراجعة نوعية؛ للتأكد من توافق الإجابات مع المعايير العلمية الصارمة المتوقعة في الخطاب الإسلامي.

يرتبط (BERTScore) بمكونات الدقة والاسترجاع من منظور تمثيلات السياق (embedding-based)، وأن F1 يستخدم هنا قياساً مجمّعاً في بعض الجداول. كما أوضحنا أننا نعرض مكونا (Recall و Precision) منفصلة؛ لإعطاء تحليل أعمق.

وبالنسبة لحساب مؤشرات الدقة والاسترجاع والقيمة المتوازنة F1 في سياق المهمتين Q/A و A/TA، فتحتسب على مستوى الوحدات الدلالية لا على الكلمات المفردة، عن طريق مطابقة العبارات المرجعية الدالة على الحقائق والمفاهيم الأساسية داخل الإجابات. كما أوضحنا أن المقارنة بين Q/A تستند إلى مدى ارتباط السؤال بالإجابة المنتجة، بينما A/TA تقيس مدى تقارب الإجابة المولدة مع المرجع الأصلي في المحتوى والمغزى. وهذا التحديد يمنح المقاييس معنىً تفسيرياً أدق عند تحليل التغيرات بعد تطبيق (RAG):

- تشابه جيب التمام (Cosine Similarity): مقياس هندسي يحدد درجة التقارب الدلالي بين المتجهات النصية.

$$\text{تشابه جيب التمام} = \frac{(أ.ب)}{(\|أ\| \cdot \|ب\|)}$$

- تشابه جاكارد (Jaccard Similarity): يقيس نسبة العناصر المشتركة بين مجموعتين مقارنة بالمجموع الكلي، ويعطي قيمةً بين ٠ (اختلاف تام)، و١ (تشابه تام).

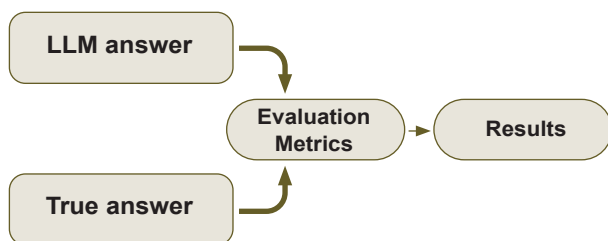
$$\text{تشابه جاكارد} = \frac{|أ \cup ب|}{|أ \cap ب|}$$

- درجة (BERTScore): يقارن الأجوبة المولدة مع المرجعية على مستوى الكلمات والسياق باستخدام تمثيلات (BERT) العميقة، جامعاً بين الدقة والاستدعاء في مقياس واحد.

$$\text{درجة بيرت (س، ر)} = \frac{1}{N} \sum_{i=1}^N \max (C_i, R_j) \text{ (التشابه جيب التمام)}$$

وبذلك، تجمع بُنيتنا بين مرحلي الكشف والتخفيف؛ لتعزيز جودة الاستجابات في السياقات الإسلامية، مما يضع معياراً جديداً في توليد المحتوى الديني بواسطة النماذج اللغوية الكبيرة. ويوضح (شكل ٣) سير عمل التقييم في نموذجنا.

شكل (٣) مخطط سير عملية التقييم



٣-٢-٤- التجارب والنتائج

أولاً: خطوة الكشف

أ. التقييم الكمي

- المطالبة بلا أمثلة (Zero-Shot Prompting):

بدأنا عملية التقييم باختيار عشوائي لـ ١٥٠ سؤالاً، و ٥٠ سؤالاً، من كل من مجموعات بيانات القرآن والحديث والفتاوى، ضمن مجموعة الحقائق الذهبية (Golden Truth Dataset)، وبالاستعانة بمقياس (BERTScore)، المعروف بقدرته على التقاط أوجه التشابه النصي الدقيقة، بما يتجاوز المقاييس التركيبية، مثل: جيب التمام (Cosine) وجاكارد (Jaccard) - أجرينا مقارنة بين (ChatGPT) و (IslamGPT) و (Jais) كما هو موضح في (جدول ٢).

جدول (٢) نتائج تقييم النماذج على مجموعات البيانات

(Query / Answer اختصار لـ Q/A، و True Answer / Answer اختصار لـ A/TA)

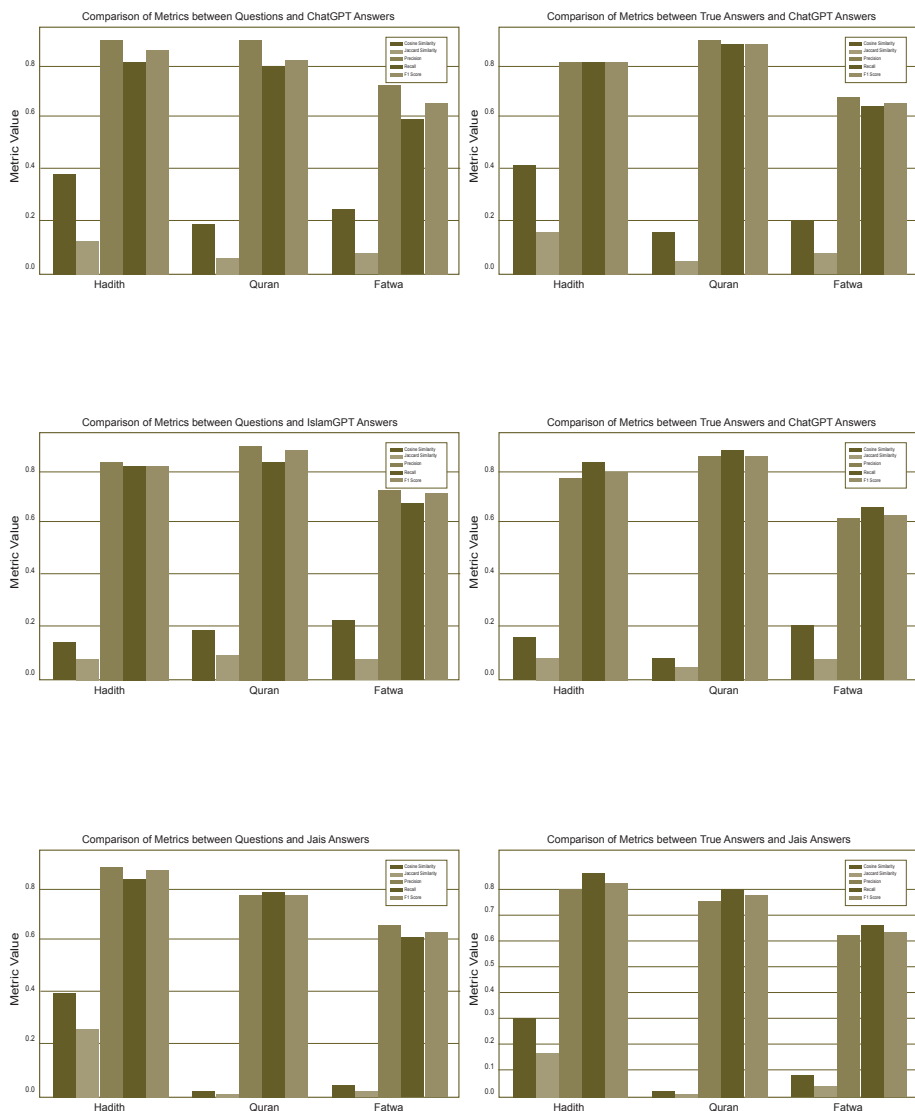
النموذج	مجموعة البيانات	خيب التمام Q/A	خيب التمام A/TA	جاكارد Q/A	جاكارد A/TA	الدقة Q/A	الدقة A/TA	الاسترجاع Q/A	الاسترجاع A/TA	مقياس F1 Q/A	مقياس F1 A/TA
ChatGPT	GTH	0.37	0.42	0.12	0.16	0.90	0.82	0.80	0.82	0.85	0.82
	GTF	0.24	0.21	0.06	0.08	0.74	0.68	0.58	0.65	0.65	0.66
	GTQ	0.19	0.16	0.05	0.05	0.89	0.90	0.80	0.89	0.84	0.89
IslamGPT	GTH	0.14	0.13	0.09	0.07	0.85	0.77	0.84	0.84	0.84	0.80
	GTF	0.22	0.13	0.12	0.07	0.74	0.64	0.65	0.67	0.70	0.65
	GTQ	0.17	0.10	0.11	0.04	0.90	0.86	0.84	0.89	0.87	0.87
Jais	GTH	0.38	0.31	0.28	0.14	0.92	0.80	0.88	0.85	0.90	0.82
	GTF	0.07	0.08	0.03	0.04	0.67	0.62	0.60	0.64	0.63	0.63
	GTQ	0.03	0.03	0.01	0.01	0.77	0.74	0.78	0.81	0.77	0.77

وقد أظهرت النتائج توازناً في كفاءة (ChatGPT)، مع تفوق ملحوظ في التعامل مع محتوى الحديث؛ مما يشير إلى قابليته للتكيف مع السرديات الإسلامية. بينما أظهر (IslamGPT) توافقاً عالياً مع أسئلة القرآن، بما يبرز تخصصه، لكنه واجه صعوبات في الحديث والفتاوى. أما (Jais)، فقد أبان عن خبرة في معالجة محتوى الحديث.

ومع ذلك، فقد قاد (ChatGPT) الأداء الإجمالي من حيث معدل الإجابات الصحيحة، مؤكداً قدرته على التكيف عبر مختلف المجالات الإسلامية، وإن كانت هناك مخاوف تتعلق بالحساسية الثقافية سيتم مناقشتها لاحقاً. ويعرض (شكل ٤) سلسلة من المقارنات بين مخرجات (ChatGPT و IslamGPT و Jais).



شكل (٤) المقارنة بين المقاييس في مختلف النماذج



يوضح (شكل ٤.أ) المقارنة بين الأسئلة الأصلية والإجابات المولدة بواسطة (ChatGPT)،
بينما يوضح (شكل ٤.ب) مقارنة بين إجابات (ChatGPT)، والإجابات المرجعية المقابلة.

أما (شكلا ٤.ج، و ٤.د)، فيعرضان مقارنات مماثلة لـ (IslamGPT) بين الأسئلة والإجابات
المولدة، وبين الإجابات المولدة والإجابات المرجعية على التوالي. وبالمثل، فإن (شكلي ٤.هـ،
و ٤.ف) يعرضان المقارنات المقابلة لـ (Jais).

أما فيما يتعلق بمدى التوافق بين الأسئلة والإجابات المولدة من (ChatGPT)، فيكشف
التحليل أن (ChatGPT) يقدم أفضل أداء في فئة الحديث عبر جميع المقاييس المُقيمة، تليه
فئة القرآن. بينما أظهرت فئة الفتاوى انخفاضاً ملحوظاً في الأداء، خصوصاً في مقاييس
الدقة والاسترجاع و F1.

وتشير هذه النتائج إلى أن (ChatGPT) قادر على توليد إجابات دقيقة وملائمة نسبياً
لأسئلة المتعلقة بالحديث والقرآن، بينما يتراجع أداؤه بشكل ملحوظ عند التعامل مع
أسئلة الفتاوى.

أما عند مقارنة إجابات (ChatGPT) بالإجابات المرجعية المقابلة، فقد سجلت إجابات
القرآن عموماً أعلى الدرجات عبر مقاييس الدقة والاسترجاع و F1.

وأما إجابات الحديث فقد أدت أداءً جيداً، في حين أظهرت إجابات الفتاوى انخفاضاً كبيراً
في جميع المقاييس المُقيمة؛ مما يشير إلى ضعف التوافق مع الإجابات المرجعية.

وبالنسبة لتوافق الأسئلة مع إجابات (IslamGPT)، أظهر التحليل تبايناً بين الفئات؛
فقد حققت إجابات الفتاوى تشابهاً أعلى باستخدام مقياس جيب التمام مع الإجابات
المرجعية، مما يشير إلى صلة سطحية، غير أن إجابات القرآن حققت درجات أعلى في الدقة
و F1؛ مما يعكس دقة أكبر وغنى بالمعلومات. أما إجابات الحديث فقد أظهرت أداءً متوسطاً
عبر جميع المقاييس.



وفيما يتعلق بمدى التوافق مع الإجابات المرجعية، فقد حققت إجابات القرآن من (IslamGPT) مرة أخرى أداءً متفوقاً عبر مقاييس الدقة والاسترجاع وF1، تلتها إجابات الحديث. بينما أظهرت إجابات الفتاوى أضعف أداءً باستمرار، خاصة في مقاييس الدقة والاسترجاع، مما يبرز التحديات في توليد محتوى يتعلق بالفتاوى بدقة واقعية.

أما بالنسبة إلى (Jais)، فقد كشفت المقارنات - سواء مع الأسئلة أو مع الإجابات المرجعية - أن النموذج يقدم أفضل أداء في أسئلة الحديث؛ إذ حقق أعلى الدرجات في جميع المقاييس. كما سجلت إجابات القرآن أداءً جيداً نسبياً، بينما تراجعت إجابات الفتاوى في جميع المقاييس، مما يدل على فعالية محدودة في هذه الفئة.

وعبر جميع النماذج والفئات، تفوّق (ChatGPT) على كل من (Jais) و (IslamGPT) في إنتاج إجابات دقيقة وذات صلة. وجاء (Jais) في المرتبة الثانية، يليه (IslamGPT) الذي أظهر أكبر قدر من التباين وأضعف أداء إجمالي، خصوصاً في مجال الفتاوى.

- المطالبة بعدة أمثلة (Few-Shot Prompting):

انعكاساً لبنية Zero-shot، قمنا بتوفير مثالين لكل نوع من الأسئلة، بهدف توجيه استجابات النماذج. واستند هذا النهج إلى إرشادات (Minne) الخاصة ببنية المطالبات. وقد استُخدم مقياس (BERTScore) مجدداً لتقييم النتائج، حيث أظهر تحسناً في التوافق النصي، بينما ظلت مقاييس جيب التمام (Cosine) وجاكارد (Jaccard) محدودة في قدرتها على كشف الفروق الدلالية الدقيقة.

ومن حيث تباين الأداء، حافظ (ChatGPT) على مستوى ثابت، في حين أظهر كل من (IslamGPT) و (Jais) انخفاضات طفيفة، كما هو موضح في (جدول ٣). وقد أبرزت هذه المرحلة محدودية تكيف النماذج في السيناريوهات الثقافية المعقدة، بما قد يؤدي إلى تمثيل خاطئ للمنظورات الدينية.

جدول (٣) نتائج التوليد بطريقة المطالبة بعد أمثلة (Few-shot)

النموذج	مجموعة البيانات	تشابه جيب التمام Q/A	تشابه جيب التمام A/TA	تشابه جاكارد Q/A	تشابه جاكارد A/TA	الدقة Q/A	الدقة A/TA	الاسترجاع Q/A	الاسترجاع A/TA	مقياس F1 Q/A	مقياس F1 A/TA
ChatGPT	GTH	0.39	0.34	0.25	0.16	0.91	0.80	0.85	0.85	0.88	0.82
	GTF	0.22	0.17	0.09	0.08	0.76	0.67	0.63	0.67	0.69	0.67
	GTQ	0.42	0.10	0.16	0.04	0.90	0.87	0.84	0.89	0.87	0.88
IslamGPT	GTH	0.0	0.01	0.0	0.1	0.80	0.74	0.82	0.82	0.81	0.78
	GTF	0.0	0.0	0.0	0.0	0.54	0.47	0.60	0.59	0.60	0.52
	GTQ	0.0	0.0	0.0	0.0	0.71	0.65	0.82	0.83	0.76	0.73
Jais	GTH	0.37	0.34	0.15	0.12	0.90	0.80	0.83	0.83	0.87	0.81
	GTF	0.06	0.07	0.02	0.03	0.66	0.63	0.58	0.63	0.62	0.63
	GTQ	0	0	0	0	0.71	0.65	0.80	0.81	0.75	0.72

فيما يتعلق بالتوافق بين الأسئلة والإجابات، قدّم ChatGPT أفضل أداء في مجال الحديث؛ حيث حقق أعلى الدرجات في تشابه جاكارد، والدقة، والاسترجاع، ومقياس F1، ومع ذلك، كان أدائه الأضعف في مجال الفتاوى؛ إذ سجّل أدنى القيم عبر جميع المقاييس. أما في مجال القرآن، فقد حقق ChatGPT أعلى درجة في تشابه جيب التمام، ودرجات تنافسية في الدقة والاسترجاع وF1، رغم أن تشابه جاكارد كان أقل مقارنة بالحديث.

أما التوافق بين الإجابات المرجعية وإجابات ChatGPT، فقد أظهر النموذج مجدداً أفضل أداء في مجال الحديث، خصوصاً في تشابه جيب التمام وجاكارد، مع تحقيق نتائج قوية في الدقة والاسترجاع، مما أدى إلى ارتفاع قيمة F1.



في المقابل، أظهر النموذج أضعف أداء في مجال الفتاوى عبر جميع المقاييس؛ مما يشير إلى ضعف الصلة والتشابه مع الإجابات المرجعية. أما في مجال القرآن، فقد برع (ChatGPT) في الدقة والاسترجاع وF1؛ مما يدل على أن إجاباته عالية الصلة والدقة، رغم انخفاض تشابه جاكارد نسبياً.

وفيما يتعلق بالتوافق بين الأسئلة والإجابات، فقد أظهر (IslamGPT) أقوى أداء في مجال الحديث، محققاً أعلى درجات الدقة والاسترجاع وF1، وهو ما يعكس صلة عالية ودقة ملحوظة. أما أدائه في مجال الفتاوى فكان الأضعف عبر جميع المقاييس. وفي مجال القرآن، قدّم النموذج أداءً متوسطاً؛ إذ كانت درجات الدقة والاسترجاع أقل مما حققه في الحديث، لكنها أعلى مما سجّله في الفتاوى، مما أدى إلى الحصول على قيمة F1 مقبولة.

أما التوافق بين الإجابات المرجعية وإجابات (IslamGPT)، فقد برع النموذج أيضاً في مجال الحديث، محققاً أعلى الدرجات في الدقة وF1، مع بعض التداخل في تشابه جيب التمام وجاكارد. أما في الفتاوى، فقد سجّل أضعف أداء عبر جميع المقاييس. وفي مجال القرآن، أظهر النموذج استرجاعاً مرتفعاً ودقة، وقيمة F1 متوسطة، رغم غياب التشابه في جيب التمام وجاكارد؛ مما يشير إلى محتوى ذي صلة، لكن مع محدودية التداخل اللفظي، أو الدلالي مع الإجابات المرجعية.

أما (Jais)، فقد حقق أفضل أداء له في مجال الحديث فيما يتعلق بالتوافق بين الأسئلة والإجابات؛ إذ سجل أعلى الدرجات في جميع المقاييس، مما يعكس صلة عالية ودقة في الإجابات. وفي المقابل، كان أدائه الأضعف في مجال الفتاوى. أما في مجال القرآن فقدّم استرجاعاً مرتفعاً ودقة، وقيمة F1 متوسطة، لكنه افتقر إلى تشابه جيب التمام وجاكارد؛ مما يشير إلى إجابات ذات صلة لكن دون تطابق لفظي أو دلالي تام.

أما التوافق بين الإجابات المرجعية وإجابات (Jais)، فقد أظهر النموذج تفوقاً في مجال الحديث، محققاً أعلى الدرجات في جميع المقاييس؛ مما يعكس إجابات دقيقة وذات صلة، مع وجود بعض التداخل اللفظي والدلالي. وفي مجال الفتاوى، حقق دقة واسترجاعاً متوسطين، لكنه سجّل أدنى القيم في تشابه جيب التمام وجاكارد؛ مما يوحي بإجابات ذات

صلة، لكن بعيدة لفظياً ودلالياً عن المرجع. أما في مجال القرآن، فقد حقق استرجاعاً مرتفعاً ودقة متوسطة، منتجاً قيمة F1 جيدة؛ لكنه أظهر تشابهاً ضئيلاً من حيث الكلمات الدقيقة والتمثيلات المتجهية.

وبوجه عام، ظل أداء (ChatGPT) ثابتاً مقارنةً بوضع (Zero-shot prompting)، بينما أظهر (Jais) انخفاضاً طفيفاً في الأداء، في حين شهد (IslamGPT) تدهوراً ملحوظاً في نتائجه مقارنة بالمرحلة السابقة.

- المطالبة بالتعليمات (Instruction-Based Prompting):

من خلال إدخال تعليمات منظمة ومتعددة اللغات، سعينا إلى التخفيف من تحيزات النماذج، وأظهرت النتائج تحسناً ملحوظاً في مقاييس جيب التمام (Cosine) وجاكارد (Jaccard) لجميع النماذج (جدول ٤)؛ حيث حافظ (ChatGPT) على أداء مستقر، في حين أظهر كل من (IslamGPT) و (Jais) تحسناً طفيفاً بعد التراجع السابق. وأشارت التحليلات إلى أن التعليمات الموجهة (Instructional Prompting) قد تسهم في تحسين أداء النماذج، على الرغم من بقاء المخاوف الأخلاقية المتعلقة بخصوصية المحتوى الديني قائمة بشكل ملحوظ.

جدول (٤) نتائج التوليد بطريقة المطالبة بالتعليمات (Instruction-based Prompting)

النموذج	مجموعة البيانات	تشابه جيب التمام Q/A	تشابه جيب التمام A/TA	تشابه جاكارد Q/A	تشابه جاكارد A/TA	الدقة Q/A	الدقة A/TA	الاسترجاع Q/A	الاسترجاع A/TA	مقياس F1 Q/A	مقياس F1 A/TA
ChatGPT	GTH	0.36	0.40	0.10	0.13	0.90	0.83	0.79	0.81	0.84	0.82
	GTF	0.22	0.20	0.06	0.08	0.73	0.68	0.59	0.65	0.65	0.66
	GTQ	0.19	0.16	0.05	0.05	0.89	0.90	0.80	0.89	0.84	0.89
IslamGPT	GTH	0.14	0.16	0.07	0.08	0.85	0.78	0.83	0.84	0.84	0.81



النموذج	مجموعة البيانات	تشابه جيب التمام Q/A	تشابه جيب التمام A/TA	تشابه جاكارد Q/A	تشابه جاكارد A/TA	الدقة Q/A	الدقة A/TA	الاسترجاع Q/A	الاسترجاع A/TA	مقياس F1 Q/A	مقياس F1 A/TA
	GTF	0.08	0.05	0.04	0.03	0.62	0.55	0.61	0.62	0.61	0.58
	GTQ	0.17	0.10	0.11	0.04	0.90	0.86	0.84	0.89	0.87	0.87
Jais	GTH	0.37	0.39	0.11	0.13	0.89	0.81	0.81	0.82	0.85	0.82
	GTF	0.10	0.15	0.03	0.06	0.70	0.67	0.59	0.65	0.64	0.66
	GTQ	0.03	0.03	0.01	0.01	0.77	0.74	0.78	0.81	0.77	0.77

فيما يتعلق بتوافق الأسئلة مع إجابات (ChatGPT)، حقق النموذج أفضل أداء في الأسئلة المتعلقة بالحديث، مسجلاً أعلى الدرجات في تشابه جيب التمام وتشابه جاكارد والدقة، إلى جانب نتائج قوية في الاسترجاع ومقياس F1، وتلاه أدائه في أسئلة القرآن، التي وإن سجلت درجات عالية في الدقة والاسترجاع وF1، إلا أنها جاءت أقل من أسئلة الحديث في مقاييس التشابه (جيب التمام وجاكارد). أما في أسئلة الفتاوى، فقد أظهر (ChatGPT) أضعف أداء، خصوصاً في الدقة والاسترجاع وF1.

أما عند مقارنة الإجابات المرجعية مع إجابات (ChatGPT) المولدة، فقد كان الأداء الأفضل في الأسئلة المتعلقة بالقرآن؛ حيث سجل أعلى الدرجات في الدقة والاسترجاع وF1، وإن كان أدائه أدنى مقارنة بالأسئلة المتعلقة بالحديث في مقاييس جيب التمام وجاكارد. وتلتها أسئلة الحديث التي حققت أعلى الدرجات في التشابه (Cosine) و (Jaccard) إلى جانب دقة واسترجاع وF1 مرتفعة، وإن كانت أقل قليلاً من القرآن في الدقة وF1، بينما أسئلة الفتاوى جاءت مجدداً بأضعف أداء في معظم المقاييس.

وفيما يخص توافق الأسئلة مع إجابات (IslamGPT)، فقد حقق النموذج أفضل أداء في الأسئلة المتعلقة بالقرآن؛ حيث حصل على أعلى الدرجات عبر جميع المقاييس، تلتها أسئلة

الحديث بدرجات مرتفعة نسبياً في الدقة والاسترجاع وF1، وإن كانت أقل من القرآن. بينما أسئلة الفتاوى أظهرت الأداء الأضعف مع درجات منخفضة باستمرار في جميع المقاييس.

وعند تقييم توافق الإجابات المرجعية مع إجابات (IslamGPT)، كان الأداء الأفضل مجدداً في الأسئلة المتعلقة بالقرآن، مسجلاً أعلى الدرجات في الدقة والاسترجاع وF1، إلا أنه جاء أدنى من أسئلة الحديث في التشابه (Cosine) و (Jaccard).

وقد أظهرت أسئلة الحديث أداءً جيداً في مقاييس التشابه، إلى جانب استرجاع وF1 مرتفعين، وإن كان بدرجة دقة أقل مقارنة بالقرآن. أما أسئلة الفتاوى فكانت الأضعف باستمرار عبر جميع المقاييس.

وبالنسبة لتوافق الأسئلة مع إجابات (Jais)، فقد كان الأداء الأفضل في الأسئلة المتعلقة بالحديث، مسجلاً أعلى الدرجات في جميع المقاييس، تلتها أسئلة القرآن التي أظهرت أداءً جيداً في الاسترجاع، ولكنها جاءت أقل من الحديث في بقية المقاييس. أما أسئلة الفتاوى فقد أظهرت مرة أخرى أضعف أداء في معظم المقاييس.

وعند تقييم توافق الإجابات المرجعية مع إجابات (Jais)، حقق النموذج مجدداً أفضل أداء في الأسئلة المتعلقة بالحديث؛ إذ حصل على أعلى الدرجات في جميع المقاييس، تلتها أسئلة القرآن بأداء جيد في الاسترجاع، ولكنه أضعف في المقاييس الأخرى. بينما سجلت أسئلة الفتاوى أدنى أداء عبر جميع المقاييس. ومن الجدير بالذكر أن هذه النتائج تمثل تحسناً مقارنةً بإعداد (Few-shot prompting) الذي شهد خلاله كل من (Jais) و (IslamGPT) تراجعاً ملحوظاً في الأداء، في حين بقي أداء (ChatGPT) مستقرًا.

وعبر استراتيجيات التوليد الثلاث (Zero-shot، Few-shot، وInstruction-based)، فقد أظهرت استراتيجية (Zero-shot) أعلى أداء إجمالي لجميع النماذج، تلتها استراتيجية (Instruction-based)، في حين كانت (Few-shot) الأقل فاعلية. علاوة على ذلك، تبين أن لغة السؤال تؤثر بشكل كبير على أداء النماذج؛ إذ تفوقت الأسئلة باللغة الإنجليزية باستمرار على تلك التي تكون باللغة العربية.



ب. التقييم النوعي

كشفت المراجعة البشرية المعمّقة عن وجود هلوسات في النماذج، وتركزت بشكل أساسي في الإجابات المتعلقة بالفتاوى، وهو مجال عُرضة لسوء الفهم الثقافي والديني. فعلى سبيل المثال، يوضح (شكل ٥) أمثلة لأنواع مختلفة من الهلوسة التي تولدها النماذج اللغوية الكبيرة. ويمثّل (شكل ١.٥) فتوى مولدة بواسطة (ChatGPT)، وقد جاءت متناقضة بشكل صارخ مع الفتوى الحقيقية. ومثل هذه التشويهات تنطوي على مخاطر أخلاقية كبيرة؛ إذ قد يؤدي إنتاج نماذج لتوجيهات دينية خاطئة إلى التأثير غير المقصود على معتقدات أو ممارسات المستخدمين. ويعرض (شكل ٥.ب) مثالاً آخر لهلوسة (IslamGPT) في الحديث؛ إذ يجسد نمطاً مميزاً للهلوسة، فيتمثل في توليد النموذج لمحتوى يبدو منطقياً دلياً، ولكنه يفتقر إلى الأساس الواقعي. فقد قدّم النموذج إجابة سطحية، تفتقر إلى التفاصيل الجوهرية، مع الإشارة إلى مرجع لا علاقة له بالإجابة الصحيحة، منسوبةً تحديداً إلى قول الله تعالى في الآية الكريمة: ﴿لَا خَيْرَ فِي كَثِيرٍ مِّنْ نَّجْوَاهُمْ إِلَّا مَنَ أَمْرٌ بَصَدَقَ أَوْ مَعْرُوفٍ أَوْ إِصْلَاحٌ بَيْنَ النَّاسِ وَمَن يَفْعَلْ ذَلِكَ ابْتِغَاءَ مَرْضَاتِ اللَّهِ فَسَوْفَ نُؤْتِيهِ أَجْرًا عَظِيمًا﴾ [النساء: ١١٤].

بينما يوضح (شكل ٥.ج) مثالاً لهلوسة (Jais) في القرآن؛ إذ يمثل ذلك خطأ واقعياً واضحاً؛ فالآية المستشهد بها لا تتطابق مع أي نص قرآني. وهذه الحالة تكشف عن تزوير للنصوص المقدسة، من خلال تقديم محتوى ملفّق باعتباره مصدراً موثقاً.

شكل (٥) أمثلة للهلوسة في النماذج اللغوية الكبيرة

(أ) هلوسة النفي في (ChatGPT)، (ب) هلوسة الإسناد في (IslamGPT)، (ج) هلوسة الحقائق في (Jais)

True Answer		ChatGPT Answer	
Question		Question	
هل الله تعالى عز وجل خلق الإنسان من عظام آدم؟		هل الله تعالى عز وجل خلق الإنسان من عظام آدم؟	
True Answer		True Answer	
No, Allah created Adam from dust and clay.		No, Allah created Adam from dust and clay.	
LLM Answer		LLM Answer	
Allah has bestowed numerous blessings and graces upon Prophet Muhammad (ﷺ), which are beyond our full comprehension. The kinds of grace and blessings that Allah bestowed upon him cannot be fully appreciated or listed. (Quran 41:49)		Allah has bestowed numerous blessings and graces upon Prophet Muhammad (ﷺ), which are beyond our full comprehension. The kinds of grace and blessings that Allah bestowed upon him cannot be fully appreciated or listed. (Quran 41:49)	
(أ)		(ب)	
True Answer		ChatGPT Answer	
Question		Question	
هل الله عز وجل خلق الإنسان من عظام آدم؟		هل الله عز وجل خلق الإنسان من عظام آدم؟	
True Answer		True Answer	
No, Allah created Adam from dust and clay.		No, Allah created Adam from dust and clay.	
LLM Answer		LLM Answer	
Allah has bestowed numerous blessings and graces upon Prophet Muhammad (ﷺ), which are beyond our full comprehension. The kinds of grace and blessings that Allah bestowed upon him cannot be fully appreciated or listed. (Quran 41:49)		Allah has bestowed numerous blessings and graces upon Prophet Muhammad (ﷺ), which are beyond our full comprehension. The kinds of grace and blessings that Allah bestowed upon him cannot be fully appreciated or listed. (Quran 41:49)	
(ج)		(د)	

أظهرت النماذج اللغوية الثلاثة جميعها - المُقيّمة في هذه الدراسة - مظاهر للهلوسة في الإجابات المولّدة ضمن سيناريوهات (Zero-shot ، Few-shot ، و Instruction-based prompting).

ولغرض تقييم منهجي لأنماط التباين الدلالي المرتبطة بالإجابات المهلوسة، أجرينا تقييماً باستخدام أسئلة مختارة عشوائياً، من مجموعات البيانات الثلاث الأساسية (الفتاوى، القرآن، الحديث). واختير عشرة أسئلة من كل مجموعة؛ لضمان تمثيل شامل لمختلف أنماط المحتوى الإسلامي. وأُجريت حسابات (BERTScore) بين الإجابات المرجعية الحقيقية ومخرجات النماذج المقابلة؛ لتحديد عتبات التشابه الدلالي الدالة على وجود هلوسة. وتشمل الأسئلة المختارة لهذا التحليل ما يلي:

س١: ما حكم الصلاة في ثياب بنصف الكم للرجل؟

س٢: حكم لبس الرجل للذهب عند خطبة النكاح؟

س٣: ما هي الشجرة الملعونة؟

س٤: ما هو ميراث الأم عن ولدها إن لم يكن له ولد؟

س٥: ما هي الآيات التي تتحدث عن موضوع الوصية في سورة المائدة؟

س٦: من هم أبناء سيدنا إبراهيم عليه السلام؟

س٧: أي نعمة أنعم الله بها على النبي محمد ﷺ؟

س٨: ما هي الشجرة التي قال عنها النبي محمد ﷺ: "تشبه المسلم، تعطي ثمارها في كل حين بإذن ربها، ولا تسقط أوراقها"؟

س٩: هل يجوز للمرأة أن تعمل؟

س١٠: هل يجب أن يقبض صداق المرأة عند تسميته أو عند العقد، أم يكفي بتسميته، ويجوز تأجيله إلى وقت لاحق بعد الزواج؟



يقارن (جدول ٥) بين التقييم الكمي والنوعي، من خلال عرض قيم (BERTScore) إلى جانب الرصد الفعلي لوجود الهلوسة في كل سؤال من الأسئلة المختارة.

جدول (٥) نتائج درجة بيرت للأسئلة العشرة

السؤال	قيمة درجة بيرت	هلوسة
س1	0.67	نعم
س2	0.67	نعم
س3	0.90	نعم
س4	0.88	نعم
س5	0.86	لا
س6	0.90	نعم
س7	0.80	لا
س8	0.83	لا
س9	0.68	لا
س10	0.63	لا

ويكشف تحليل النتائج الواردة في الجدول عن قصور جوهري في الاعتماد على المقاييس الكمية وحدها، باعتبارها أداة لاكتشاف الهلوسة؛ إذ أظهرت البيانات وجود ترابطات غير متسقة بين درجات التشابه الدلالي ووجود الهلوسة، وسُجلت حالات لقيم مرتفعة من (BERTScore) متزامنة مع محتوى غير دقيق واقعيّاً، في حين احتوت بعض الإجابات ذات الدرجات المنخفضة على عدم وجود هلوسة. وإن هذا التناقض يشير إلى أن مقاييس التشابه الدلالي وحدها غير كافية لاكتشاف الهلوسة بشكل موثوق في مجالات المحتوى الديني.

ويكشف التقييم الشامل عن تحدٍ جوهري في التقدير الآلي للهلوسة، بالرغم من إظهار جميع النماذج المدروسة لتماسك سياقي سطحي، كما تدل على ذلك القيم المرتفعة لمقياس

(BERTScore)، ولكن التحليل النوعي المنهجي أظهر أنماطاً واسعة الانتشار للهلوسة، عبر جميع فئات المحتوى وهياكل النماذج. وقد تفاوتت هذه التزويرات في شدتها وتعقيدها، ويُحتمل أن تكون ناجمة عن التعقيدات اللغوية الكامنة في الخطاب الديني أو عن التحيزات الممنهجة في بيانات التدريب. ويبرز الانفصال الملحوظ بين التشابه الدلالي والدقة الواقعية الحاجة إلى تطوير أطر تقييم متخصصة.

وبناءً عليه، نقترح في أبحاث مستقبلية تطوير مقياس مخصص لتقدير الهلوسة أطلقنا عليه اسم (HALscore)، يُصمّم خصيصاً لاكتشاف الانحرافات الواقعية في مخرجات النماذج اللغوية الكبيرة ضمن مجالات المعرفة المتخصصة.

ثانياً: خطوة التخفيف

أ. إعادة الضبط الدقيق (Fine-Tuning):

في هذه المرحلة من البحث، قمنا بتطبيق نهج الضبط الدقيق الخاص بالمجال (Domain-specific Fine-tuning)، وذلك من خلال اختيار نموذج لغوي عربي مدرب مسبقاً، وتكييفه باستخدام مجموعات البيانات المدمجة (الإصدارات: AraGPT2 Base Islamic، و AraGPT2 Medium Islamic).

وبعد ذلك، أجرينا تقييماً منهجياً للسلوكيات؛ بهدف قياس إمكانية النماذج المطوّرة المتخصصة في مجال معين في الإسهام بفعالية في التخفيف من ظاهرة الهلوسة، عند توليد المحتوى الإسلامي، مجيبين بذلك على السؤال الجوهرية: هل يُعدّ التحسين المتخصص بالمجال حلاً عملياً لتحسين الموثوقية الواقعية في تطبيقات الذكاء الاصطناعي ذات الحساسية الثقافية؟

وقد اختيرت مجموعة أسئلة الفتاوى؛ لتكون المجموعة الأساسية للتقييم بالنسبة للنموذج المضبوط بدقة؛ استناداً إلى الأدلة التجريبية المستخلصة من تحليلنا المقارن، وقد أظهرت أن الاستفسارات المتعلقة بالفتاوى كانت هي الأكثر توليداً للإجابات المهلوسة، عبر جميع النماذج المفحوصة السابقة. ولضمان الاتساق المنهجي وتمكين المقارنة المباشرة للأداء،



جرى تقييم النماذج المضبوطة، باستخدام مجموعة أسئلة الفتاوى نفسها، المستعملة في التجارب السابقة. وستُعرض النتائج المقارنة في (جدول ٦).

جدول (٦) نتائج (AraGPT2 Base Islamic، و AraGPT2 Medium Islamic)

النموذج	مجموعة البيانات	تشابه جيب التمام Q/A	تشابه جيب التمام A/TA	تشابه جاكارد Q/A	تشابه جاكارد A/TA	الدقة Q/A	الدقة A/TA	الاسترجاع Q/A	الاسترجاع A/TA	مقياس F1 Q/A	مقياس F1 A/TA
AraGPT2 Base Islamic	GTF	0.24	0.14	0.13	0.08	0.86	0.68	0.58	0.64	0.69	0.64
AraGPT2 Medium Islamic	GTF	0.26	0.18	0.13	0.09	0.86	0.69	0.60	0.65	0.70	0.65

وقد بيّنت النتائج التجريبية الموثقة في (جدول ٦) أن الضبط الدقيق المتخصص بالمدال لم يحقق التحسينات المتوقعة في الأداء، ويمكن أن يُعزى هذا المآل السلبي إلى عدة عوامل منهجية، منها: صغر حجم مجموعة بيانات الضبط الدقيق، أو إعدادات غير مثالية لمعاملات ما قبل التدريب، أو قيود جوهرية في بنية النموذج الأساسية.

بالإضافة إلى ذلك، أظهر التقييم النوعي أن إصدارات (AraGPT2)، (الأساسي والمتوسط)، قد أبانت عن تدهور ملحوظ في جودة الإجابات، مقارنة بأنماط الهلوسة المسجلة في النماذج الأخرى التي تم تقييمها؛ مما يشير إلى وجود قصور جوهري في قدرات التوليد. و(جدول ٧) أدناه، يبين النتائج النهائية لمختلف النماذج عند تطبيقها على أسئلة الفتاوى.

ومن خلال القيام بتحليل تشخيصي لعدم فاعلية الضبط الدقيق في تجربتنا، فيمكننا أن نبين أهم العوامل التالية التي رصدناها:

١. **حجم مجموعة الضبط:** كانت مجموعة الضبط المتاحة لنطاق الفتاوى محدودة نسبياً انظر (جدول ٦)؛ مما أدى إلى خطر الإفراط في التعلّم (overfitting) أو قصور في تعميم الأنماط الفقهية المتنوعة.

٢. تنوع البيانات وجودتها: تبين أن بعض حالات الضبط احتوت على إجابات مرجعية أقل تفصيلاً؛ مما يلزم لتعليم النموذج فروقاً دقيقةً بين المسائل الفقهية المختلفة.

٣. قيود بنيوية للنموذج الأساسي: بعض إصدارات (AraGPT2) المستخدمة في الضبط لم تكن قوية بما يكفي لالتقاط الفروق الدقيقة، رغم الضبط المتخصص. نتأج (جدول ٦) توضّح أن الضبط الدقيق لم يحسّن (BERTScore) بالشكل المرغوب.

جدول (٧) مقارنة بين النماذج المختلفة في أسئلة الفتاوى

النموذج	قيمة درجة بيرت
ChatGPT	0.66
IslamGPT	0.65
Jais	0.63
AraGPT2 Base Islamic	0.64
AraGPT2 Medium Islamic	0.65

ب. التوليد المعزز بالاسترجاع (RAG):

في هذا القسم، قمنا بتطبيق تقنية التوليد المعزز بالاسترجاع (Retrieval-Augmented Generation - RAG) على أربعة نماذج، هي: (AraGPT2-Medium-Islami، وChatGPT Jais، وIslamGPT)، وقد زُود كل نموذج بمقطعين نصيين ذوي صلة، يحتويان على الإجابات الصحيحة لسؤال طُرح مسبقاً، فكانت النماذج قد وُلدت له في البداية إجابات مهلوسة. وقُدّم السؤال الأصلي مع المقاطع المسترجعة باعتبارها مدخلات، وجمعت المخرجات المُنقحة للنماذج بغرض تقييمها لاحقاً، وذلك باستخدام كل من (BERTScore)، والتحليل النوعي، وستُعرض نتأج تقييم (BERTScore) في (جدول ٨).



جدول (٨) نتائج دمج تقنية الاسترجاع (RAG)

درجة بيرت		النموذج
بعد RAG	قبل RAG	
0.65	0.69	AraGPT2 Medium Islamic
0.72	0.66	ChatGPT
0.55	0.64	IslamGPT
0.58	0.64	Jais

كما هو موضح في الجدول، انخفضت قيم (BERTScore) لجميع النماذج بعد تطبيق تقنية (RAG)، باستثناء (ChatGPT) الذي أظهر زيادة ملحوظة. وعلى الرغم من هذا الانخفاض العام في قيم (BERTScore) لمعظم النماذج، إلا أن التحليل النوعي أشار إلى تحسن كبير في جودة الإجابات المولدة عبر جميع النماذج.

ومن الجدير بالذكر، أن (Jais)، و (ChatGPT)، و (IslamGPT) قدّموا إجابات دقيقة، بعد أن كانوا قد ولّدوا سابقاً محتوى مهلوساً، في حين أن (Medium-Islamic-AraGPT2) لم يُظهر أي تحسن ملحوظ في جودة الإجابات.

وفي سياق توضيح سبب الانفصال بين نتائج (BERTScore) والتقييم البشري بعد إدماج (RAG)، فقد لوحظ أن النماذج تنتج إجابات أقصر وأكثر دقة؛ مما يقلل التشابه اللفظي ويُحسن الاتساق الدلالي في الوقت نفسه. وهنا نحتاج أن نوضح أن تقييم المقيمين استند إلى جودة المعنى والمصدر، لا إلى التشابه السطحي للنصوص، وهو الذي يفسر انخفاض (BERTScore) رغم الارتفاع النوعي؛ فتقنية (RAG) تزود النموذج بمقاطع مرجعية، قد تؤدي إلى إعادة صياغة الإجابة بكلمات، ومطابقة مرجعية مختلفة، (أي إجابة أكثر حقائقية/مقتبسة من مصدر موثوق لكن بلغة مختلفة)، بينما يقيس (BERTScore) توافقاً دلاليًا/لفظيًا جزئيًا؛ لذلك قد تنخفض الدرجات العددية (BERTScore)، إذا كانت الصياغة الجديدة أقصر أو تعتمد على اقتباس مرجعي، بصياغة تختلف عن المرجع المستخدم في قياس التشابه، وبالرغم من ارتفاع الدقة الحقيقية وصحة الإشارة المرجعية كما رصده

التقييم البشري. كما أشرنا صراحةً إلى هذا التفسير عند مناقشة نتائج (جدول ٨)، وبينّا أن المثال الفعلي في الورقة يوضّح زيادة نوعية لـ (ChatGPT) مقابل انخفاض (BERTScore) لمعظم النماذج الأخرى بعد تطبيق (RAG).

كما نحب أن نوضح بالمناسبة أثر لغة السؤال على جودة النتائج؛ إذ بينت التجارب أن الأسئلة باللغة الإنجليزية أدت غالباً إلى إجابات أكثر دقة؛ بسبب توافر بيانات تدريب أوسع للنماذج في تلك اللغة. ويمكننا اقتراح مجموعة من الاستراتيجيات لتقليل هذا الأثر مثل: استخدام ترجمة مزدوجة للأسئلة، أو تطبيق (RAG) متعدد اللغات؛ لرفع أداء الإجابات العربية.

٤- الخاتمة والتوصيات المستقبلية

تقدم هذه الدراسة تقييماً شاملاً لأنماط الهلوسة، وتحليلات التحيز في النماذج اللغوية الكبيرة عند معالجة المحتوى الإسلامي، مسلطة الضوء على فجوات حرجية في تطوير الذكاء الاصطناعي الواعي ثقافياً. ومن خلال تحليل منهجي لثلاثة من أبرز النماذج (ChatGPT، و IslamGPT، و Jais)، قمنا بدراسة مدى انتشار الأخطاء الواقعية، وسوء التمثيل الثقافي في السياقات الدينية الحساسة، مع التركيز على خصائصها واستراتيجيات التخفيف منها.

وتكشف نتائجنا التجريبية عن عدة رؤى مهمة تعزز فهم موثوقية النماذج اللغوية الكبيرة في المجالات المتخصصة:

أولاً: يبيّن التحليل الكمي أن الهلوسة تحدث بشكل منهجي، عبر جميع النماذج المخصصة، مع وجود تباينات قابلة للقياس في التكرار والشدة. وإن هذا الطابع الشمولي يُبرز التحدي الجوهرية، المتمثل في ضمان الدقة الواقعية في البنى الحالية للنماذج، على الرغم من اختلاف هياكلها وأساليب تدريبها. كما أن استمرار ظهور الهلوسة عبر النماذج، بدرجات متفاوتة من التخصص في المجال الإسلامي، يوحي بأن التدريب المتخصص بالمجال وحده غير كافٍ لمعالجة هذه المخاوف المتعلقة بالموثوقية.

ثانياً: يكشف تقييمنا المقارن الشامل عن فروق دقيقة في الأداء بين النماذج التي تم فحصها. فقد أظهر (ChatGPT) أداءً متفوقاً في توليد استجابات مناسبة سياقياً،



مع الحفاظ على الاتساق الواقعي، بينما أظهر كل من (IslamGPT)، و (Jais) أنماطاً مميزة من الأخطاء، تعكس نماذج تدريبها وقيودها المعمارية الخاصة. وتسلط هذه النتائج الضوء على العلاقة المعقدة بين التخصص في المجال، وتركيبه بيانات التدريب، وموثوقية المخرجات في البيئات الحساسة ثقافياً.

ثالثاً: أدى دمج منهجية التوليد المعزز بالاسترجاع (RAG) إلى تحسينات جوهرية، عبر جميع المقاييس والنماذج. فقد أظهر الإطار المعزب (RAG) انخفاضاً كبيراً في وتيرة الهلوسة، وتحسناً في الدقة الواقعية، وتعزيزاً للملاءمة المحتوى الثقافي. وتؤكد هذه التحسينات فرضيتنا بأن ربط عملية التوليد بمصادر معرفية موثوقة يمكن أن يخفف بفعالية من تحديات الموثوقية المتأصلة في المناهج المعتمدة على النماذج البارامترية فقط.

رابعاً: نجح إطار الكشف الهجين الذي طورناه في تحديد وقياس أنماط الهلوسة بدقة واسترجاع مرتفعين؛ مما وفر قدرات تقييم آلية تتجاوز أساليب المراجعة اليدوية. وقد مكّن الجمع بين مقاييس جيب التمام (Cosine)، وجاكارد (Jaccard)، وتقييمات (BERTScore) من توصيف شامل للانحرافات الدلالية والواقعية عن المصادر الموثوقة.

وتحمل هذه النتائج انعكاسات مهمة لتطوير أنظمة ذكاء اصطناعي جديرة بالثقة في السياقات الحساسة ثقافياً؛ إذ تشير فعالية استراتيجيات التخفيف المعتمدة على (RAG) إلى مسار عملي، نحو تعزيز موثوقية النماذج، دون الحاجة إلى إعادة تصميم شاملة للمعمارية، أو عمليات إعادة تدريب باهظة التكاليف. علاوة على ذلك، يوفر منهج الكشف الذي طورناه إطاراً قابلاً للتكرار؛ لتقييم ومراقبة أداء أنظمة الذكاء الاصطناعي في المجالات المتخصصة.

ومع ذلك، هناك عدة قيود يجب الاعتراف بها؛ فقد ركز تقييمنا تحديداً على المحتوى الإسلامي، بينما تتطلب قابلية التعميم إلى سياقات دينية أو ثقافية أخرى دراسات إضافية. كما أن فعالية استراتيجيات التخفيف تعتمد بشكل حاسم على جودة قاعدة المعرفة المسترجعة وشموليتها؛ مما يبرز أهمية مجموعات المصادر المنتقاة والموثوقة.

وينبغي أن تركز الأبحاث المستقبلية على عدة مجالات رئيسية:

أولاً: توسيع إطار التقييم ليشمل سياقات دينية وثقافية أوسع للتحقق من قابلية تعميم النتائج والمنهجيات.

ثانياً: استقصاء قابلية التوسع لمقاريات (RAG) عبر سياقات لغوية متنوعة، وخاصة اللغات محدودة الموارد.

ثالثاً: تطوير خوارزميات كشف أكثر تقدماً، وقادرة على تحديد الأشكال الدقيقة؛ للتحيز الثقافي، ويمثل عدم الملاءمة السياقية فرصة حاسمة للتقدم. وأخيراً، فإن دمج آليات التغذية الراجعة المجتمعية وعمليات التحقق من قبل السلطات الدينية يمكن أن يعزز الأصالة الثقافية، وقبول المحتوى المولد آلياً.

ومن خلال النقاش التحليلي، قمنا بتفسير أسباب ضعف أداء الضبط الدقيق في بعض التجارب. فقد أظهرت الدراسة أن محدودية حجم بيانات الضبط، وتفاوت جودة النصوص المرجعية أدت إلى ضعف التعميم الإحصائي للنموذج؛ خاصة في المسائل الفقهية الدقيقة. كما بينا أن البنية الداخلية للنموذج الأساسي لا تتيح التقاط العلاقات السياقية المعقدة، رغم التدريب الإضافي.

وبناءً على ذلك، نقترح توظيف تقنيات ضبط خفيفة مثل (LoRA، QLoRA) في الأبحاث المستقبلية؛ لما توفره من كفاءة في ضبط النماذج الكبيرة، دون الحاجة إلى موارد ضخمة، مع الحفاظ على مرونة التكيف اللغوي. كما سنحاول توسيع أفاق العمل المستقبلي، باقتراح اختبار نماذج أحدث مثل (GPT-4.x)، ونماذج عربية جديدة من أجل التحقق من إذا كانت مشكلة الهلوسة المعرفية ما زالت قائمة بالحدة ذاتها.

وتتجاوز إسهامات هذا العمل المجال الخاص بمعالجة المحتوى الإسلامي؛ إذ يوفر إطاراً متكاملاً للكشف عن التحيز وتخفيف الهلوسة؛ مما يؤسس لأسس منهجية قابلة للتطبيق في تطبيقات الذكاء الاصطناعي الحساسة ثقافياً. وإن فعالية الجمع بين مقاييس تقييم متعددة، والتحقق من جدوى استراتيجيات (RAG)، وتطوير منهجيات كشف قابلة للتوسع - تسهم مجتمعة في دفع حدود المعرفة في مجال تطوير ذكاء اصطناعي جدير بالثقة.



ومع استمرار انتشار النماذج اللغوية الكبيرة في تطبيقات ومجتمعات مستخدمي متنوعة، يصبح ضمان موثوقيتها وحساسيتها الثقافية أكثر إلحاحاً. وتُظهر هذه الدراسة أنه رغم القيود المنهجية التي تعاني منها الأنظمة الحالية في معالجة المحتوى الثقافي الحساس، ولكن استراتيجيات التخفيف الموجهة يمكن أن تحقق تحسينات جوهرية في الدقة والملاءمة. ويوفر الإطار المقدم هنا أساساً لتطوير أنظمة ذكاء اصطناعي أكثر موثوقية، وتحترم السياقات الثقافية مع الحفاظ على الفاعلية وإمكانية الوصول.

ختاماً، تؤكد هذه الدراسة أن الهلوسة والتحيز في المحتوى الديني المولّد بواسطة النماذج اللغوية الكبيرة - تمثل تحديات قابلة للمعالجة، وليست قيوداً مستعصية. ومن خلال التقييم المنهجي، ومنهجيات الكشف الموجهة، واستراتيجيات التخفيف المبنية على الأدلة، يمكننا تعزيز موثوقية الأنظمة وحساسيتها الثقافية، بما يُسهم في نشرها المسؤول عبر سياقات عالمية متنوعة؛ فالطريق إلى الأمام يتطلب تعاوناً مستمراً بين التقنيين والخبراء الميدانيين وأصحاب المصلحة المجتمعيين؛ لضمان نشر أنظمة الذكاء الاصطناعي المعرفة الدقيقة، من غير تشويه ثقافي أو تضليل ديني.

المراجع الأجنبية

- Abdelazim, H., Tharwat, M., & Mohamed, A. (2023). Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG). *International Journal of Advanced Computer Science & Applications*, 14(11).
- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.
- Alan, A. Y., Karaarslan, E., & Aydin, O. (2024). A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM. *arXiv preprint arXiv:2401.15378*.
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., ... & Chen, H. (2024). Unified Hallucination Detection for Multimodal Large Language Models. *arXiv preprint arXiv:2402.03190*.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1-113.
- Elmadany, A., & Abdul-Mageed, M. (2022, May). AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)* (pp. 628-647).



- Few-shot Prompting – Guide Sydologie ChatGPT. (n.d.). Sydologie.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.
- Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vectordatabase: Storage and retrieval technique, challenge. arXiv preprint arXiv:2310.11703.
- Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence, 4(4), 799-819.
- Ivchenko, G. I., & Honov, S. A. (1998). On the Jaccard similarity test. Journal of Mathematical Sciences, 88, 789–794.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
- Milios, A., & Behnam Ghader, P. (2022). An analysis of social biases present in bert variants across multiple languages. arXiv preprint arXiv:2211.14402.
- Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456.
- Patel, S., Kane, H., & Patel, R. (2023). Building Domain-Specific LLMs Faithful To The Islamic Worldview: Mirage or Technical Possibility? arXiv preprint arXiv:2312.06652.
- Peng, R., Liu, K., Yang, P., Yuan, Z., & Li, S. (2023). Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data. arXiv preprint arXiv:2308.03107.

- Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic cosine similarity. The 7th international student conference on advanced science and technology ICAST (Vol. 4, No. 1, p. 1).
- Sadat, M., Zhou, Z., Lange, L., Araki, J., Gundroo, A., Wang, B., ... & Feng, Z. (2023). Delucionqa: Detecting hallucinations in domain-specific question answering. arXiv preprint arXiv:2312.05200.
- Talat, Z., Névéal, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ... & Van Der Wal, O. (2022, May). You reap what you sow: On the challenges of bias evaluation under multilingual settings. In Proceedings of BigScience Episode# 5--Workshop on Challenges & Perspectives in Creating Large Language Models (pp. 26-41).
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219.



نبذة تعريفية عن الباحث

Author Biodata Fouzi Harrag is a full Professor of Ethics in AI in the Department of Computer Sciences College of Sciences Ferhat Abbas University. Dr Fouzi Harrag received his PhD degree in Computer Sciences (2011) from Ferhat Abbas University. His research interests include Text and Data Mining, NLP and Ethics in AI.	بيانات الباحث: فوزي حراق، أستاذ التعليم العالي في (أخلاقيات الذكاء الاصطناعي) في (قسم الحاسب الآلي) (بكلية العلوم) في جامعة فرحات عباس بالجزائر. حاصل على درجة الدكتوراه في علوم الحاسب من جامعة فرحات عباس عام ٢٠١١، تدور اهتماماته البحثية حول. التنقيب عن النصوص والبيانات، ومعالجة اللغات الطبيعية وأخلاقيات الذكاء الاصطناعي.
معرف أوركيد (ORCID) : 7146-8348-0002-0000	
Email: fouzi.harrag@univ-setif.dz	

الرصد الآلي للمتلازمات المعجمية في المدونة النصية العربية: مقارنة معجمية حاسوبية

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

د. عواطف السمعلي الزواري

قسم العربية، الكلية العلوم الإنسانية والاجتماعية، جامعة تونس، تونس، الجمهورية التونسية

استقبل في	رُوجع في	قُبِل في	نُشر في
١٤٤٧/٠٣/١٨	١٤٤٧/٠٣/٢٢	١٤٤٧/٠٤/١٦	١٤٤٧/٠٥/٠٩

المخلص

يعد هذا البحث هو حلقة ضمن سلسلة بحوث تندرج في إطار الدراسات البينية، ونصبو من خلاله إلى استثمار التقنية في أحد مبثي الصناعة المعجمية وهو "الجمع" متخذين من المتلازمات المعجمية مادة دراسة، وتبني عناصر عملنا على مبحثين: الأول نظري نعرف فيه الحقل البيني المعجمي الحاسوبي ونضبط فيه المتلازمة المعجمية مفهوميًا، الثاني تطبيقي نُعنى فيه بالرصد الآلي لنماذج تمثيلية من المتلازمات المعجمية. وكون المتلازمة تنفرد بخصائص تميزية فإننا نقف على أهم تحديات حوسبتها الناتجة خاصة عن بنيتها اللاتوزيعة وتكلسها الجزئي سعيًا لإيجاد حلول مستنديين إلى خيارات نظرية لسانية. كما نعمل على البحث عن الاطراد النسقي لنظمته التغيرات مستأنسين بنماذج تمثيلية لم نغفل عن ترميزها رياضيًا. ونأمل أن نضع من خلال هذا العمل لبنة أولى في سبيل الرصد الآلي لمخزون معجمي مفرق في ثنايا المدونة العربية.

الكلمات المفتاحية

متلازمات معجمية، ترميز رياضي، بنية لاتوزيعة، جمع، صناعة معجمية.

Automatic Detection of Lexical Collocations in the Arabic Text Corpus: A Computational Lexicographic Approach

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

Dr. Awatef Al-Samali Al-Zouari¹⁰

Department of Arabic, College of Humanities and Social Sciences, University of Tunis, Tunis, Republic of Tunisia

Received on	Revised on	Accepted on	Published on
10-09-2025	14-09-2025	08-10-2025	31-10-2025

ABSTRACT

Through this research we aim to employ technology in one of the two fields of lexicography, “compilation”, taking lexical collocations as the subject of study. Our work is structured around two axes: The first is theoretical, in which we define the interdisciplinary field of computational lexicography and establish a conceptual definition of collocations. The second is applied, in which we focus on the automatic detection of lexical collocations. Given that collocations possess distinctive properties, we examine the challenges of their computational processing, particularly those stemming from their non-compositional structure and partial fossilization, in an effort to propose solutions grounded in theoretical linguistic approaches. We also work toward identifying the regularity that governs patterns of variation, drawing on representative models that we have encoded mathematically.

KEYWORDS

Lexical collocations, Mathematical encoding, Non-compositional structure, Compilation, Lexicography.

١- المقدمة

ينتمي بحثنا إلى مجال الدراسات البينية المعجمية الحاسوبية، ونهدف من خلاله إلى المساهمة في تجسير العلاقة بين الصناعة المعجمية والمعالجة الآلية، ذلك أنه في ظل التطور التقني المتسارع لا يمكن أن تبقى الصناعة المعجمية معزولة عن التقنيات الحديثة، ولأن مبحث الصناعة المعجمية يتصل به مبحثان ذوا فروع هما الجمع والوضع ويتخذ من الوحدات المعجمية سواء كانت في حالة الأفراد أو التضام أو التلازم؛ مادة دراسة، فإن عملنا هذا يتخذ من اللسانيات الحاسوبية إطاراً مرجعياً عاماً يفرض تعميق النظر في إحدى حالات الوحدات المعجمية وهي حالة التلازم، وخيارنا النموذجي هذا مبرر؛ ذلك أن المتلازمات المعجمية ترتبط بخصوصية من خصوصيات تجربة الجماعة اللغوية، وتنفرد بخصائص تميزية على الوجهين الشكلي والدلالي، هذا فضلاً عن ندرة المعاجم العربية الخاصة بها، ونود أن نشير إلى أن أول معاجم المتلازمات في الفرنسية هو "Lacroix" وقد ظهر سنة ١٩٥٦، وفي الإنجليزية معجم "The B B I Combinatory Dictionary" الذي ظهر سنة ١٩٨٦.

ما سبق كان دافعاً لتسليط الضوء في هذا العمل على دور الآلة في جمع المتلازمات المعجمية من المدونة العربية. ونحسب أن هذا الهدف ممتنع ما لم يعكف اللساني على استقصاء مظاهر الانتظام فيها وتوصيفها توصيفاً يلغي الحدس البشري الراجع إلى معارفنا الفطرية حتى يتمكن الحاسوب من التعرف عليها في بطون النصوص في بنيتها اللاتوزيعية. وعليه، نرى أن وضع معايير دقيقة لحد المتلازمة وتمييزها مما يمكن أن يلبسها فضلاً عن جدولتها استناداً إلى التغيرات التي قد تطرأ على عناصرها المعجمية، إضافة إلى التجريد والترميز Annotation وإثراء قواعد البيانات بملفات بيانات تؤسس لذاكرة حاسوبية، جميعها تشكل آليات تمكن الحاسوب من رصد المتلازمات في المدونات النصية العربية. لذا سنعمل في هذا البحث على حد المتلازمة المعجمية لتمييزها عما يمكن أن يلبسها والوقوف على خصائصها التي تعيق حوسبتها، ومن ثمة نقترح تصوراً مرحلياً لمعالجة آلية تعاضدها جهود بشرية تعتمد على رصد المتلازمة في بنيتها اللاتوزيعية وتتبع مظاهر الأظفار فيما يمكن أن يطرأ عليها من تغيرات ناتجة عن تكلسها الجزئي مما يسمح بالجدولة، فضلاً عن الترميز



الرياضي ملاءمةً لمقتضيات الحوسبة وإثراء قاعدة البيانات بملفات بيانات تشكل حدساً صناعياً، يعني، أن هذه الورقة تهدف إلى توصيف الكفاية اللغوية التي يصدر عنها ابن اللغة في الاهتداء إلى المتلازمة المعجمية في مظان النصوص.

بناءً على ما تقدم، سنعرض عملنا في بنية ثنائية من خلال مبحثين نجمع فيهما بين الرؤية المعجمية العميقة والمعالجة الحاسوبية الممكنة: المبحث الأول نظري، نسعى من خلاله إلى تحقيق مطلبين: الأول تعريف الحقل البيئي المعجمي الحاسوبي، والثاني ضبط مفهومي للمتلازمة المعجمية لتمييزها مما يُلبسها مع استعراض صعوبات معالجتها آلياً، الثاني تمثيل تطبيقي استثنائياً بنماذج من المتلازمات المعجمية، والعمل على تجاوز صعوبات رصدها آلياً استناداً إلى تصوّر مرحلي يهيئ للحوسبة.

تطلب منا هذا العمل ضبطاً على مستوى منهجي، فمن موجباته تضافر المنهجين التحليلي الوصفي والاستقرائي. وعليه، فورقتنا البحثية تتعلّق بإشكالية كبرى هي: ما سُبُل الاستفادة من الآلة في جمع المتلازمات المعجمية؟ وتتفرّع عنها إشكالات صُغرى نُجملها فيما يلي: ما الحدود المفهومية للمتلازمة المعجمية؟ ما تحديات حوسبتها المترتبة عن خصائصها التمييزية؟ ما الحلول المقترحة لتيسير عمل الآلة؟

ننظر في هذه الإشكالات ونحن نعلم أنه لم تعرض -حسب ما اطلعنا عليه- أي دراسة لسانية حاسوبية لتوصيف المتلازمات المعجمية بالمفهوم الذي نقرحه.

٢- مقوّمات حوسبة اللغات

تقتضي حوسبة ظاهرة لغوية ما، إطارين تمهيديين نظريين يسبقان المعالجة الآلية للعيننة التمثيلية، الأول يتعلق بمقومات المعالجة الحاسوبية للغات الطبيعية عامة، ويرتبط الثاني بالكشف عن خصائص الظاهرة اللغوية والوقوف على تحديات حوسبتها، ثم السعي لتوصيفها توصيفاً يلغي دور الإدراك الحسي الراجع إلى معارفنا الفطرية ليتلاءم في الآن ذاته مع متطلبات المعالجة الحاسوبية الرامية في بحثنا هذا إلى جمع المتلازمات المعجمية آلياً من النصوص باعتبار أن الجمع أحد مباحث الصناعة القاموسية.

رغم أن الفصل بين المستويات اللغوية ضرورة بحثية؛ لاستحالة الإلمام بالنظام اللغوي المتشعب في عمل واحد وهو ما دفع الباحثين إلى تجزئته إلى عدة مستويات هي المستوى الصوتي والمستوى الصرفي والمستوى التركيبي والمستوى الدلالي، إلا أن التوصيف وكسر الحدود بين جميع تلك المستويات من مقتضيات المعالجة الحاسوبية، ذلك أن "رسم صورة اللغة العربية للحاسوب لا يكتفي بالوصف بل يقتضي التوصيف ليعوض عن عنصر الحدس" (الموسى، ٢٠٠٠، ص. ٢) حتى تُماثل الآلة القدرة البشرية في الأداء اللغوي وتتمكن من "استكناه العمليات الإرادية التي تحدث في العقل البشري التي يمكن من خلالها إعطاء الحاسوب القدرة على فهم اللغة الطبيعية وتحليلها وإعادة إنتاجها وكيفية تشكيلها في العقل البشري" (الموسى، ٢٠٠٠، ص. ٥٧) ومن ثمة نحسب أن توصيف المستوى اللغوي الواحد يقتضي الاستناد إلى قواعده الوصفية اللسانية ثم الربط بينها وبين بقية قواعد مستويات النظام ربطاً علائقياً، لذلك نرى أن للتوصيف زيادة فضل تكمن في "الطموح إلى بلوغ النموذج الكامن في العقل الإنساني لهذه القواعد، وهو الحدس الذي يمكنه من إقامة الترابط التلقائي بين هذه القواعد في بنية متكاملة لغاية الاستفادة في مواقف الأداء". (الموسى، ٢٠٠٠، ص. ٢١) من ذلك أن حوسبة المستوى الصرفي مثلاً في المتلازمات المعجمية لا يمكن أن تتم بمعزل عن المستوى التصريفي، فالتقوانين اللغوية التي يعيها ابن اللغة وعياً حدسياً ذهنياً تغني عن تعريفه بها عند استعمال اللغة، أما بالنسبة إلى الآلة فأى تغيير - حتى إن كان نظامياً - يطرأ على ما هو مخزن في ذاكرتها تعتبره شذوذاً ما لم يكن جزءاً من ملفات قواعد بياناتها بما أن تلك القواعد هي "مجموعات منظمة تحتوي على بيانات وتكون أبسط قاعدة بيانات متكونة من العديد من الملفات تخزن وتحفظ داخل ذاكرة الحاسوب" (La plante، ٢٠٠٠، ص. ١١٧) يعني هذا أن كل مدخلات الحاسوب يجب أن تكون مقننة لا تخرج عن نسق نظامي ما، وهنا يطرح السؤال التالي: ما أهمية الأنساق النظامية في المعالجة الحاسوبية؟

ما لا جدال فيه هو أن الآلة لا تتعامل إلا مع القواعد النظامية سواءً على مستوى الشكل أو الدلالة، لذلك فالدور المحوري الموكل إلى اللساني هو تتبع مظاهر الانتظام ورصد ضوابطه واستقصاء الاطراد سواء كان ظاهراً أو مخفياً، قصد تمكين الحاسوب من التعرف على



الظاهرة اللغوية بعد تحليلها أو إقداره على توليدها في حدود ما يسمح به النظام اللغوي، يعني هذا أن بُنْيَنَ الظَّاهِرَةِ اللُّغَوِيَّةِ موضوع الحوسبة وَمَقُولَتُهَا هي السَّبِيلُ الأَوْحَدُ لمعالجتها آلياً، فمتى رصد اللساني الحاسوبي مظاهر الاطراد، وتتبع التحولات التي يمكن أن تطرأ على الظاهرة اللغوية والتغيرات الممكنة، وردها إلى النظام، تَسَيَّ له تحويلها إلى خوارزميات هي مدخلات تسبق المعالجة الآلية تُدرج في الحاسوب عن طريق وسائل إدخال Input Devices لتشكيل الذاكرة الحاسوبية. ونحسب أن النُّظْمَةَ التي يصبو اللساني الحاسوبي إلى بلوغها ممكنة مادامت اللغة نظاماً من العلامات (De Saussure, 1996, p. 26) وانتفاء الوحدة المعجمية عامة والمتلازمات خاصة إلى ذلك النظام يفترض انتظامها أيضاً، وما على اللساني الحاسوبي إلا تتبع مظاهر الانتظام، وهو ما يقتضي تجاوز الوصف إلى التَّوصيف، لملاء قواعد البيانات بملفات هي مدخلات مساعدة للحاسوب على المعالجة، وبقدر موسوعية ملفات قاعدة البيانات تكون دقة الاستجابة وسرعتها.

لئن اتسم النظام المعجمي ممثلاً في وحداته المعجمية بالحركية والانفتاح مقارنة بالنظام التركيبي، وهو ما حدا بالبعض إلى اعتبار المعجم سجناً يحوي الخارجين عن القانون لأن الكثير من مفرداته لا يقبل الجدولة (Di Sciullo & Williams, 1997, p. 3) فإننا نعتبر أن النظرة العميقة المتأنيّة يمكن أن تكشف عن الاطراد النسقي وترد ما يمكن أن يطرأ على الوحدة المعجمية من تحولات إلى الانتظام، وهو ما ييسر سبيل الحوسبة، لهذا سنُعنى في هذا العمل بتتبع نظام التغيرات التي يمكن أن تطرأ على المتلازمة المعجمية - باعتبارها خياراً النموذجي - من قبيل التحول في مقولتي الجنس والعدد أو فيهما معاً وهو ما تعدّه الآلة شذوذاً يحول دون التعرف عليها، ما لم تكن هذه التغيرات مدرجة في ملفات قواعد بياناتها لتشكيل ذاكرة الحاسوب في إطار نسقي يفي بحاجة المعالجة الآلية إلى المَقُولَةِ والانتظام. لكن السؤال المطروح في هذا المستوى من البحث هو كيف يمكن للآلة أن تتعرف على المدخلات المدرجة في ملفات قواعد البيانات؟

إن مشروعية سؤالنا متأنيّة من معرفة خصائص العقل الحاسوبي، ذلك أنه عقل صناعي رياضي " يميل إلى الرموز والأعداد والمختزلات، وذلك لطبيعته الخوارزمية من أجل تسريع البحث وتسهيله بعيداً عن الأشكال اللغوية المركبة والمعقدة بالنسبة إلى الحاسوب

لأنه لا يعقلها ولا يفهمها لذلك تتعدد لغات البرمجة وتتطور لتكون أكثر حركية ومرونة عملية استعمالية، ومن تلك اللغات (CSS) "ي آس آس" و (Java) "جافا" و (Python) "بايثون". (محمدي، ٢٠٢٢، ص. ٣٨).

بناء على ما تقدم، نرى أنه يُضاف إلى المقومين الأساسيين المذكورين سابقاً وهما التوصيف والنظمنة، مقوم أساسي آخر لا يقل عنهما أهمية في عملية الحوسبة هو الترميز الرياضي، ملاءمةً لمتطلبات العقل الصناعي، لذلك نقدّر أن تتضمن ملفات قاعدة البيانات معلومات تُدخل في الحاسوب بعد تحويلها إلى مختصرات ورموز رياضية يفهمها ويسهل عليه استرجاعها والاستفادة منها، سواء في عملية التحليل أو التوليد أو البحث الاستقصائي. لأجل ذلك سنسعى في عملنا هذا إلى تقديم اقتراح لترميز ما نفترض أن يشكل مدخلات تسبق المعالجة الآلية الرامية إلى ربط الصناعة المعجمية من حيث مبحث الجمع فيها بالتقنيات الحديثة، وذلك استناداً إلى نماذج تمثيلية من المتلازمات المعجمية باعتبارها خيارنا النموذجي في هذا العمل.

نخلص مما تقدم إلى نتيجة نعتبرها على قدر من الأهمية مفادها أن المعالجة الحاسوبية للأنظمة اللغوية في أي مستوى من مستوياتها لا يمكن أن تتم بمنأى عن الجهود البشرية، وهو ما سيتضح لنا في الجزء التطبيقي الذي سنسعى فيه إلى تقديم مقترح لحوسبة المتلازمات المعجمية.

والسؤال المطروح في هذا المستوى هو: ما الضبط المفهومي لمصطلح المتلازمة المعجمية؟
ما خصائصها التمييزية؟ وما علاقتها بالصناعة المعجمية؟

٣- الصناعة المعجمية والمتلازمات المعجمية

قبل الحديث عن الصناعة المعجمية لا بد أن نميز بين المعجم اللساني Lexicon والمعجم الصناعي أو القاموس Dictionary، فالأول هو رصيد عام تنقاسمه أفراد المجموعة اللغوية الواحدة، أما الثاني فهو رصيد جزئي يخص جزءاً مدوّناً من ذلك الرصيد العام، يعني هذا أن المعجم اللساني أعم وأشمل من المعجم الصناعي، ذلك أنه لا يمكن الإحاطة بجميع مفردات اللغة ووضعها بين دفتي كتاب، وتكمن أهمية المعجم الصناعي في أنه يحفظ



اللغة ليس باعتبارها مجرد أداة للتواصل فقط، بل باعتبارها وعاءً حاملاً للمعرفة ومرسّخاً للهويّة. (السمعلي الزواي، ٢٠١٩، ص. ١٨-٢٠) فما مباحث الصناعة المعجمية؟ وما علاقتها بالمتلازمات المعجمية؟

٣-١- في الصناعة المعجمية

ما نود الإشارة إليه هو أن المعجم الصناعي ليس مرآة عاكسة للنظام اللغوي فحسب، بل هو وعاء حافظ لهوية الجماعة اللغوية بمنأى عن المعيارية، وعليه فالصناعة المعجمية علم قائم الذات، يتخذ من الوحدات المعجمية قبل دخولها في شبكة العلاقات التركيبية مادةً دراسة؛ لتصبح مداخل معجمية خاضعة للمعالجة، ويحصر ابن منظور التأليف المعجمي في ركنين، هما الجمع والوضع وفي هذا الصدد يقول ناقدًا: "أما من أحسن جمعه فإنه لم يحسن وضعه، وأما من أجاد جمعه فإنه لم يجد وضعه، فلم يفد حسن الجمع مع إساءة الوضع، ولا نفعت إجادة الوضع مع رداءة الجمع." (ابن منظور، ١٩٩٠، ص. ٧) يعني هذا أن الصناعة المعجمية يرتبط بها مبحثان هما الجمع والوضع، أما الجمع فنقصد به "جمع الرصيد المعجمي الذي يدون في المعجم المؤلّف... وثاني المبحثين هو منهج التدوين ويسمى الوضع، وهو يشمل على مسألتين مهمتين، هما الترتيب والتعريف." (ابن مراد، ١٩٩٣، ص. ٥). وحتى لا يصبح البحث في الصناعة المعجمية متشعبًا يتجاوز حدود عملنا، اخترنا أن نهتم بمبحث الجمع دون الوضع، استنادًا إلى حالة واحدة من حالات الوحدة المعجمية، وهي حالة التلازم، فما الحدود المفهومية للمتلازمات المعجمية؟

٣-٢- في المتلازمات المعجمية وخصائصها التمييزية

إن حوسبة المتلازمات المعجمية لا بد أن تسبق بضبطها مفهوميًا وتحديد خصائصها التمييزية، إضافةً إلى رصد التحديات التي تواجه معالجتها آليًا والناجمة عما تتسم به من خصوصيات حتى نتمكن من توصيفها للكشف عن مظاهر انتظام التغيرات الناتجة عن تكلسها الجزئي.

بناءً على أن كل علم يفرض وجود مادة دراسة هي موضوعه، فالأصوات هي مادة دراسة علم الأصوات والبنية الصرفية هي مادة دراسة علم الصرف، والجملة هي مادة

دراسة علم النحو، فإن الوحدة المعجمية هي مادة دراسة علم المعجم، وتصنّف الوحدات المعجمية استناداً إلى بُناها الصرفية الخارجية إلى وحدات معجمية بسيطة هي المكونة من مفردة فقط، ووحدات معجمية مركبة هي القائمة على مفردتين، ويحمل التركيب معجمياً على التّضام، ووحدات معجمية معقّدة تُحمل على التضام أيضاً وهي مركبة من أكثر من مفردتين، ووحدات عبارية تمثل حالة التلازم^(١). ونود الإشارة إلى أن الرصيد الأساسي للمعجم الصناعي هو الوحدات المعجمية في حالتها الأولى، أي المفردات، لذلك كثرا لاهتمام بها، وقد اخترنا أن ننظر في هذه الورقة البحثية فيما قلّ تناوله بالدراسة وهي المتلازمات المعجمية، خاصة أن الاختلاف فيها مصطلحاً ومفهوماً شائع، فمن الباحثين من اعتبرها تعابير تحليلية (Kahane & Polguère, 2001, p. 2) ومنهم من عدها متلازمات لفظية وأخرجها من دائرة علم المعجم (بنت براك العنزي، ٢٠١٩) ومنهم من اصطلح عليها بالعبارات المسكوكة (الغانمي، ٢٠٢٤) ومنهم من ألبسها بالأمثال، (Tutin & Grossmann, 2001, pp. 5-7) هذا الاختلاف أوجب بيان حدودها بتقديم تصوّر مفهومي لمصطلح المتلازمات المعجمية يقطع مع التداخل السائد ويؤسس لتعريف جديد يبرر انتماءها إلى علم المعجم، ويسمح في الآن ذاته بمقوّلتها بالكشف عن خصائصها التمييزية في إطار العلاقات الانتلافية والاختلافية، هذا التصور لم نوجده من فراغ بل بنيناه على ما انتهت إليه كتابات معجمية في الجامعة التونسية وجمعية المعجمية بتونس انتهجت نفس المنهج، واجتمعت على إلحاق المتلازمات المعجمية بعلم المعجم، ومن أهمها أبحاث إبراهيم بن مراد وعلي الودرني وزكية السّاح دحماني وغيرهم.

متصورنا إذن يضع مصطلحاً ثابتاً ويؤسس لمفهوم جديد يقطع مع الآراء المتضاربة والتصورات المتصادمة. وعليه فالمتلازمة المعجمية مثلها مثل بقية الوحدات المعجمية قائمة على وجهين، شكلي ودلالي، ومن أهم خصائصها التمييزية نذكر:

أ. وجهها الشكلي: يقوم على عناصر معجمية هي وحدات معجمية بسيطة مختلفة الانتماءات المقولية، متلازمة تلازماً محققاً لشرط الإسناد، وهي في هذا تأتلف مع المثل.

(١) لمزيد من التعمق، ينظر: السمعلي الزواي، ٢٠١٩، ص. ٢٥-٤٥.



وما نود الإشارة إليه هو أننا نتبنى في تقسيم الكلّم موقف المدرسة التونسية ممثلاً فيما أجمع عليه أعضاء جمعية المعجمية العربية بتونس، فقد صنف إبراهيم بن مراد (١٩٩٧) وفريق بحثه المفردات حسب قابلية الإحالة أو عدمها إلى صنفين هما: الوحدات المعجمية التامة، وتنقسم حسب انتماءاتها المقوليّة إلى أربعة أقسام هي: مقولة الاسم ومقولة الفعل ومقولة الصفة ومقولة الظرف. والوحدات المعجمية غير التامة هي الأدوات وتُوظف للربط بين مقولات الصنف الأول داخل الإطار التركيبي.

ب. وجهها الدلالي: دلالة المتلازمة المعجمية، هي حصيلة عملية تأليفية بين جميع معاني عناصرها المعجمية، وهي ممتنعة بالجمع الرياضي بين معاني تلك العناصر، هذا فضلاً عن ترابط هذه العناصر بعلاقات مجازية واجبة الوجود، وهي تأتلف مع المثل في هاتين الخصيصتين.

ج. دورها الوظيفي: "ترتبط بخصوصية من خصوصيات تجربة الجماعة اللغوية التي أوجبت اجتماع العناصر المكوّنة للوحدة اجتماعاً مخصوصاً. وإذن، فإن التلازم يحدث في العبارات Locutions المعبرة عن خصوصية ما في تجربة الجماعة اللغوية." (ابن مراد، ٢٠٠٦، ص. ٢٧) يعني هذا أن للمتلازمة دور وظيفي إحالي إلى خصوصية من خصوصيات تجربة الجماعة اللغوية، هذا الدور الوظيفي يميزها من المفردات والمتضامّات لأن وظيفتها تعيينية، ومن الأمثال لأنها تعبر عن ظاهرة كونية تتجاوز الجماعة اللغوية الواحدة.

د. تكلسها: تواترت بين الجماعة اللغوية حتى تكلست تكلساً جزئياً، وهي في هذا تختلف عن المثل^(١).

لا يفوتنا في هذا الصدد التنبيه إلى أن المتلازمات المعجمية نوعان:

أ. متلازمات معجمية تحليلية: نعتبرها من المشترك الدلالي لأن لها حالتين من الوجود: الحالة الأولى هي أن تُحمل على الحقيقة، والثانية أن تُحمل على المجاز وهي في الحالة

(١) لمزيد التعمّق يُنظر، ابن مراد، ٢٠٠٦؛ الودرني، ٢٠٢١.

الأولى ليست متلازمة لغياب خصيصة المجاز فيها، أما في الحالة الثانية فهي متلازمة. ومن أمثلة المتلازمات التحليلية نذكر:

١. ألقى عصاه ← المعنى الحقيقي: رمى عصاه.
المعنى المجازي: استقر بالمكان.

ب. متلازمات معجمية اصطلاحية: لا تُحمل على الحقيقة وهي قائمة على المجاز فقط، فوحداتها لا تجتمع على هيئة معينة إلا لتؤدي معنى واحداً، ونذكر من أمثلتها:

٢. لبي داعي ربّه، مات.

ونمثل للخصائص التمييزية السابقة بالمتلازمة المعجمية التالية:

٣. أَكَلَ خُلَاتْنَهُ، "يقال: فلان يأكل خُلَاتْنَهُ: إذا كان بخيلاً. (مجمع اللغة العربية بالقاهرة، ٢٠٠٤، ص ٢٥٣).

تحقق المتلازمة السابقة شرط الإسناد شكلاً فهي جملة فعلية، والأحادية الدلالية من حيث وجهها المدلولي فهي تدل على الشحّ، وهو معنى مجازي ممتنع بضم رياضي لمعاني العناصر المعجمية القائمة على علاقة مجازية، أما تكلسها الجزئي؛ فيظهر في تلازم عناصرها المعجمية وحفظ رتب تلك العناصر تحقيقاً لوظيفتها الإحالية من ناحية، وجواز تغيير مقولتها التصريفية من ناحية أخرى فيقال "أَكَلَتْ خُلَاتْنَهَا" مثلاً أو "أَكَلُوا خُلَاتْنَهُمْ".

استناداً إلى ما سبق نحسب أن الخصائص التمييزية التي ضبطناها للمتلازمة المعجمية تشكل تحديات تعيق حوسبتها، فما تجليات هذا الرأي؟

٤- تحديات رصد المتلازمات المعجمية آلياً وحلولها

اهتم اللسانيون الحاسوبيون بالمعالجة الآلية للوحدات المعجمية البسيطة مهملين المتلازمات، وحتى من صرف جهداً لحوسبتها فقد تعثرت محاولاته، لذلك نسعى فيما يلي من هذا العمل لكشف تحديات الرصد الآلي للمتلازمات المعجمية بالنظر فيما ضبطناه من



خصائصها التمييزية، ومن ثمة نعمل على اقتراح حلول نراها تسهم في تحقيق هدفنا من هذا البحث. ونجمل التحديات التي تعيق رصد المتلازمات ألياً فيما يلي من المستويات:

أ. **مستوى البنية الصرفية:** نرى أن تعدّد العناصر المعجمية المكوّنة للمتلازمة وتحقيقها شرط الإسناد يحول دون رصدها الآلي، ذلك أن ما يرصده الحاسوب هو المتوالية اللسانية أي "ما وقع بين بياضين" (Dubois, 2002, p. 343) ويصطلح عليها بالكلمة -وقد سبق أن اقترحنا مراجعة هذا اللفظ مصطلحاً ومفهوماً فما يقع بين بياضين يمكن أن يكون مفردة أو حتى جملة^(١)- في حين أن المتلازمة المعجمية تتضمن أكثر من متوالية لسانية تقع كل واحدة منها بين بياضين ولا قيمة للمتوالية الواحدة في غياب بقية المتواليات أو العناصر المعجمية، لأن المتلازمة المعجمية غير قابلة للتجزئة إلى وحدات صغرى، لكن المُفترض هو أن الحاسوب سيرصد عناصر المتلازمة متفرقة مبثوثة في ثنايا النصوص ولن يرصدها مجتمعة، فيؤول الأمر إلى فشل المعالجة، بما أنه لا قيمة للعنصر المعجمي الواحد المكوّن للمتلازمة إلا بتلازمه مع بقية العناصر، والسؤال المطروح هنا هو كيف نجعل الحاسوب يرصد المتلازمات في ثنايا النصوص ويتعرف عليها باعتبارها وحدةً معجمية واحدة لا مجموعة وحدات؟ أي كيف نضع خوارزميات في شكل قواعد بيانات تُمكن الحاسوب من إلغاء البياضات بين وحدات المتلازمة ليعالجها على أنها في تعددها سلسلة خطيّة واحدة واقعة بين بياضين فقط؟

ب. **مستوى الدلالة:** يقوم معنى المتلازمة على الأحادية الدالية التي يحققها الارتباط الوثيق بين عناصرها. إلا أن هذه الدلالة الأحادية للعناصر المتعددة لن يهتدي إليها الحاسوب، بل ستكون الدلالة الناتجة عن المعالجة الآلية للمتلازمة المعجمية هي حصيلة الجمع الرياضي بين دلالة كل عنصر معجمي واقع بين بياضين، ويتجلى ذلك بوضوح في الترجمة الآلية، فالمتلازمة المعجمية التالية:

٤. ضرب في الأرض.

(١) لمزيد التوسّع ينظر، السّمعلي الزّواري، ٢٠٢٥، الصفحات ٦٦٥-٦٦٦

على تعدد عناصرها المعجمية تحمل معنى واحداً هو: سافر، وابن اللغة قادر على التوصل إلى هذا المعنى مستعيناً بحدسه الفطري، أما المعالجة الحاسوبية فستفترض شذوذاً دلاليًا هو نتاج الجمع الرياضي بين معاني جميع العناصر المعجمية، وأكثر ما يبرز ذلك في الترجمة الآلية للمتلازمة فعوض ترجمتها إلى الإنجليزية بـ He travels فإنها تُترجم بـ He hits the ground.

ج. **مستوى التنوع:** ذكرنا أن المتلازمات المعجمية نوعان: تحليلية واصطلاحية، ونحسب أن التحدي الذي يواجهه اللساني الحاسوبي فيما يتعلق بهذه الخاصية مضاعف، الأول يتمثل في كيفية التمييز بين المتلازمات التحليلية والاصطلاحية، ويُطرح الإشكال أكثر في مستوى المتلازمات التحليلية، فهل تستطيع الآلة أن تُقدّم مخرجات دقيقة فيما يتعلق بالتمييز بين نوعي المتلازمات التحليلية والاصطلاحية؟ أم أن الاستعانة بالجهود البشرية في هذا المستوى ضرورية؟ التحدي الثاني ونعتبره سير الحل يتمثل في ضرورة إيجاد ترميز شكلي يسمح بالتمييز الآلي بين النوعين.

د. **مستوى التكلّس:** إن شدة تكلّس المثل مقارنة بالمتلازمة المعجمية تجعل عملية رصده أيسر، أما التكلّس الجزئي الذي تختص به المتلازمة المعجمية باعتبار أن التركيب فيها يمكن أن تتغير مقولة الجنس فيه، أو العدد، أو الزمن، سواء تغييراً أحادياً أو متناسباً طردياً، يشكّل أبرز تحديات رصدها الآلي، ومثال ذلك:

٥. ليّ داعي ربّه.

قد تتغير مقولة عناصرها وصورها السطحية لتصبح على سبيل المثال "لبّت داعي ربّها" أو "لبوا داعي ربّهم" أو "سيلّي داعي ربّه". هذا التغير الحاصل لن يعيّه العقل الصناعي لذلك ستفشل عملية الرصد الآلي للمتلازمة المعجمية عند تغير صورة أي عنصر من عناصرها.

ونحسب أن خبيصة التكلّس الجزئي في المتلازمة المعجمية تبعث على الإشكالية المحورية التالية: كيف نُمكن الآلة من رصد المتلازمة المعجمية رغم تغير المقولة التصريفية فيها أو عند تغير صورة أي عنصر من عناصرها؟ الإشكالية المحورية المطروحة تتولد عنها إشكالات فرعية نذكر منها: ما ملفات قواعد البيانات التي نُزوّد بها الحاسوب حتى تكون



مدخلات يستعين بها في عملية الجمع حتى إن طرأ على المتلازمة تغيير؟ كيف يمكن ترميز معلومات ملفات البيانات حتى يبسر على الحاسوب استرجاعها؟

وقفنا فيما سبق على جميع التحديات التي تطرحها الخصائص التمييزية للمتلازمة المعجمية، ونرى أن تشخيص العلل سبيل لإيجاد الحلول، وعليه نعتبر أن تجاوز الوصف إلى التوصيف والنظمنة فضلاً عن الترميز الرياضي وإثراء قواعد البيانات بملفات بيانات هو السبيل الوحيد الذي يمكن الآلة من رصد المتلازمات في ثنایا النصوص، إلا أنه لا يمكننا ادعاء قدرة العقل الصناعي بمفرده على تقديم مخرجات دقيقة بمنأى عن العقل البشري.

0- في توصيف المتلازمات المعجمية وترميزها وجدولتها تمهيداً لحوسبتها

0-1- في توصيف المتلازمات المعجمية وترميزها

استناداً إلى ما سبق، نحسب أن رصد المتلازمات المعجمية حاسوبياً لجمعها من بطون المعاجم العامة وكذلك من ثنایا المدونات العامة تمهيداً لوضعها في المعجم الصناعي ليس بالأمر اليسير، ويجب أن يراعي أسّين، أولهما لغوي والثاني تقني، ويتمثلان في:

- أ. طبيعة المتلازمة وما تتسم به من خصائص.
- ب. ضرورة تحويل ملفات البيانات والقيود اللغوية إلى خوارزميات هي مدخلات تسبق المعالجة الآلية وتشكل الذاكرة الحاسوبية التي تستند إليها الآلة. (حمادة، ٢٠٠٩، ص. ١٥٤) ذلك أن الحاسوب لا يتعامل إلا مع خوارزميات تقنن الظاهرة اللغوية برصد دقائقها. و"الخوارزميات: مجموعة قواعد وقوانين مكتوبة، تستعمل لوصف الخطوات المنطقية المتبعة لمعالجة البيانات الداخلة للحصول على المعلومات والنتائج المطلوبة." (المهيوي، ٢٠٢٤، ص. ٦٣)

إن بلوغ هذه الغاية المنشودة وهي جمع المتلازمات المعجمية متشعب المسالك كثير الوسائل لكنه ممكن، خاصة أن المتلازمات وإن كثرت عدداً فيمكن الإحاطة بها، ومقترحنا نؤسسه على تصور مرحلي، نرى أنه لا مناص فيه من معاضدة الجهد البشري لعمل الآلة

الذي يجب أن يستند إلى أكثر من محلل آلي - المحلل النحوي والمحلل الصرفي والمحلل الدلالي - وإن كان المحلل الصرفي هو قطب الرحي. ونود الإشارة إلى أن تعقد الرصد الآلي لهذا الضرب من الوحدات المعجمية أدى إلى تعثر العديد من المحاولات ومنها محاولة Sinclair سنكلير الذي اعتمد المنهج الإحصائي مستخدماً أدوات حاسوبية ومدونات ضخمة لاستخراج كلمات تتوارد مع بعضها أكثر من المعدل العادي في السياق (English, 2005, p. 73) فأفرزت مقارنته تراكيب بعضها متلازمات وبعضها تكرارات تركيبية.

ورغم ذلك فإننا نقترح استثمار هذه المقاربة الإحصائية الآلية أي مقارنة Sinclair سنكلير باعتبارها خطوة ترشيحية أولى تمكن من استخراج الكلمات التي تتلازم في المدونات أكثر مما هو متعارف عليه إحصائياً، وهذه العملية تضمن خصيصتي التواتر والشيوع باعتبارهما من الخصائص التمييزية للمتلازمة المعجمية، إلا أن المخرجات ستكون جمعاً من المتلازمات والتراكيب التكرارية، ولنتجاوز قصور نتائج هذه المقاربة وجبت تصفية مخرجاتها قصد استبعاد التراكيب غير الإسنادية بما أن الإسناد خصيصة واجبة الوجود في المتلازمة، ونقترح الاستعانة بالمكتبة البرمجية Trankit باعتبارها تساعد على التحليل النحوي، اعتماداً على البنوك الشجرية Treebanks إلا أننا نقدّر أن المخرجات لن تكون على القدر المأمول ذلك أنه "ما تزال البنوك الشجرية Treebanks أو المدونات الموصّفة نحويًا Parsed Tagged Corpora بعيدة عن بلوغ حد الكمال لكنها في تطورات من الناحية المنهجية والنظرية." (روبي محمد بن عبد الرحمن، ٢٠٢٤، ص. ١٠٠)؛ لذلك نعتقد أن التدخل البشري للمراجعة ضرورة حتمية، -يعني هذا أن المعالجة في هذه المرحلة ستكون نصف آلية Semi-automatic- حتى نضمن تحقق شرط آخر من شروط المتلازمة المعجمية وهو الإسناد التركيبي.

ولأن الدلالة في المتلازمات المعجمية تقوم على المجاز وتقطع مع مبدأ تراكيبية المعنى - كما سبق أن بينّا - فلا بد من عرض مخرجات العمليتين السابقتين على محكّ هذا المعيار، وأمام محدودية نتائج الأتمتة وقصورها عن بلوغ حد الكمال في الكشف عن مظاهر خرق الانسجام الدلالي بين العناصر المعجمية، وحتى عند الاستعانة بأدوات حاسوبية مثل Metaphor detection فإن المخرجات لا يمكن الاطمئنان إليها اطمئناناً كلياً لتمييز المجاز من الحقيقة، لذلك نرى أن تدخل الجهد البشري ضروري لترشيح أخير يسمح بتمييز التراكيب



القائمة على المجاز من غيرها، خاصة حتى نضمن تحقق خصيصة واجبة الوجود في المتلازمة المعجمية وهي المجاز.

ولتخزين المتلازمات التي جمعناها بعد عمليات الترشيح النصف آلية حتى ييسر على الحاسوب استرجاعها والاستفادة منها نقترح ترميزها وفق ما تقتضيه المعالجة الآلية من رموز ومختصرات ويكون ذلك على مراحل هي التالية:

أ. تعليم كل متلازمة من المتلازمات المكوّنة لقاعدة البيانات تعليماً رمزياً حتى يتمكن الحاسوب من استرجاعها ورصد عناصرها بصفة لا توزيعية في نسيجها التركيبي، أي دون اعتبار البياضين بين كل عنصر من عناصرها، ذلك أن "الكلمات المركبة تمثل صعوبة بالغة عند تحليلها آلياً، وذلك بسبب وجود الفراغ الذي يفصل بين عناصر الكلمة المركبة حيث تختلط تلك الكلمات مع الكلمات غير المركبة" (المهيوي، ٢٠٢٤، ص ٦٦) وينطبق القول ذاته على المتلازمات، ونقدم فيما يلي مقترحين يمكنان الآلة من رصد المتلازمة المعجمية واعتبارها في حكم الكلمة الواحدة:

١. حصر المتلازمة بين خطين مائلين ويكون ذلك كالتالي /***/ ومثاله:

/اقتفى الأثر/

٢. وضع رموز مثل النجمة (*) مكان البياضات بين عناصر المتلازمة المعجمية،

ومثال ذلك.

اقتفى*الأثر

على أننا نميل إلى الاقتراح الأول؛ لأن الاقتراح الثاني يكرّس مفهوم الكلمة الذي نعتبره معيقاً للحوسبة لعدم دقته.

ب. ترميز كل متلازمة ترميزاً رياضياً يتلاءم مع متطلبات العقل الصناعي وحاجته للاختصار لتنظيم المدونة وتحويلها إلى بنية فهرسة، ويتم ذلك على مرحلتين:

١. إسناد رقم مرجعي لكل متلازمة معجمية باعتبارها مدخلاً مستقلاً، لتيسير الإحالة والفهرسة والاسترجاع في برنامج الرصد الآلي.

٢. تخصيص كل نوع من نوعي المتلازمات المعجمية (تحليلية/اصطلاحية) برمز رياضي أو وسم يمكن من تمييزه من النوع الآخر، بعد الاستعانة بالجهد البشري للتمييز بين النوعين.

ج. ترميز كل عنصر معجمي من عناصر المتلازمة ترميزاً رقمياً مقولياً فيعين الرقم رتبة ذلك العنصر في المتلازمة ويعين المختصر المقولي انتماءه إلى المقولة المعجمية، ونرى أن خاصية حفظ الرتب تفيد في هذه المرحلة، ذلك أن المتلازمة المعجمية ذات بنية شكلية منتظمة لثبوت رتب عناصرها فقد أوجبت الجماعة اللغوية اجتماع تلك العناصر على هيئة مخصوصة لتعبر عن خصوصية من خصوصيات تجربتها. ونكرر في هذا المستوى تكراراً وظيفياً واعياً حرصنا على ضرورة مراجعة مفهوم الكلمة باعتبارها ما يقع بين بياضين، لكن ما يقع بين بياضين يمكن أن يكون عنصراً واحداً مثل "عَصَّ" أو عنصريين مثل "يده" أو حتى أكثر، فالعربية تعتمد "دمج الأدوات والضمائر المتصلة مع الكلمة... حيث تتغير صورة الكلمة في اللغة العربية عند اتصالها بالضمير، مثال: كتبت، كتبنا، كتبوا... إلخ." (المهوي، ٢٠٢٤، ص. ٦٦) هذا الدمج ناتج عما بين النحو والصرف من روابط علائقية "أوجبت معالجتهم ألياً بأسلوب متداخل، فهناك تداخل مستمر بين النحو والصرف ينشأ في بداية التحليل الصرفي الآلي للكلمات" (سالم عبد الله آل طه، ٢٠٠٥، ص. ٢١) لذلك فتحديد عدد العناصر المعجمية المكونة للمتلازمة وضبط انتماءاتها المقولية، لا يُكتفى فيه بتجزئة المتلازمة تجزئة سطحية استناداً إلى ما يقع بين بياضين من مكوناتها، بل وجبت الاستعانة بمحلل صرفي لتجاوز البنية السطحية للمتلازمة إلى الكشف عن بنيتها العميقة وذلك بتجريد ما بين بياضين من العناصر المعجمية المدمجة كالضمائر المتصلة، ثم من الزوائد التصريفية إن وجدت، - ذلك أن الآلة لا يمكنها أن تقوم بعمليتين في الآن ذاته - ونود التنبيه في هذا المستوى من البحث إلى ضرورة وعي اللساني الحاسوبي بخصائص النظام الصرفي للغة وأهمية تبني نظرية لسانيات دقيقة واضحة المعالم لإنجاح برنامج المعالجة الآلية، وقد سبق لنا أن أشرنا إلى أننا على وعي بطبيعة البنى الصرفية



غير السلسلية في العربية، إضافةً إلى أننا نتبنى في المراحل الأولى من المعالجة التقسيم المعجمي الخماسي الذي تتبناه المدرسة التونسية وجمعية المعجمية التونسية؛ لوجاهته، ونقترح الترميز "حروف التاج" للمقولات المعجمية الكبرى تماشياً مع متطلبات الآلة حيث $V = \text{فعل}$ و $N = \text{اسم}$ و $AD = \text{صفة}$ و $FUN = \text{أداة}$ و $ADV = \text{ظرف}$ على أنه يجب الإشارة إلى أن كل مقولة كبرى تتضمن أجزاء مقولية تُرمز لها حرفياً حسب ما يقتضيه منا تحليل النماذج التمثيلية التي سنقدمها فيما سيأتي من البحث. وذلك استناداً إلى ملفات بيانات مدخلة مسبقاً نقدراً أنها تختلف باختلاف الانتماء المقولي للعنصر المعجمي.

ويكون التدرج بالتحليل للإبقاء على العنصر مفرداً. ونؤكد هنا على ضرورة الاحتكام إلى المراجعة البشرية خاصة في الحالات الملبسة.

ونُمثل لما سبق بالمتلازمتين التاليتين:

٦. أَلْقَى عَصَاهُ، أي استقر (متلازمة معجمية تحليلية) — $V2N3FUN(prm)1/1350-2$

الترميز الشكلي

٧. لَبَّى دَاعِيَ رَبِّهِ، أي مات (متلازمة معجمية اصطلاحية) — $V2AD3N4FUN(prm)1/1350-1$

الترميز الشكلي

الجدول (١) الرموز الرياضية الشكلية إحالاتها ووظائفها

الترميز الرياضي الشكلي	إحالاته ووظيفته
١	متلازمة اصطلاحية (لا تترجم حرفياً)
٢	متلازمة تحليلية (يطرح احتمال أن أحدهما فقط هو الصائب: قابلية الترجمة الحرفية، امتناع الترجمة الحرفية)
١٣٥٠	رقم مرجعي للمتلازمة المعجمية في قاعدة البيانات، واشتراك متلازمتين في الرقم المرجعي نفسه دليل على اختلافهما في النوع.

الترميز الرياضي الشكلي	إحالاته ووظيفته
//	ترميز شكلي يمكّن من رصد المتلازمة المعجمية في بنيتها الصماء التي لا تقبل التفكيك، ونحسب أن هذا الترميز الشكلي له فضل في حل الإشكال الناتج في الرصد الآلي للمتلازمات عن خاصيتها التركيبية.
1V2N3FUN(prn) 1V2AD3N4FUN(prn)	ترميز رقمي مقولي للعناصر المعجمية المكونة لكل متلازمة حسب رتبها وانتماءاتها المقولية. فالأرقام تحيل إلى رتبة العنصر المعجمي، أما الرموز الحرفية فتشير إلى الانتماء المقولي للعنصر المعجمي مفرداً: PR = سابقة تصريفية. V = فعل. AD = صفة. N = اسم fun(prn) = الجزء المقولي الضمير من مقولة الأداة.

* ملاحظة: الجدول من إعداد الباحثة

نحسب أن التصور المرحلي السابق الذي اقترحناه يفيد في رصد المتلازمات المعجمية، ما لم يطرأ على تركيبها تغيير مقولي، أي ما لم يلحق صورة أي عنصر من عناصرها تغيير، لكن المسألة ليست على هذا القدر من البساطة، ذلك أن خصيصة التكلّس الجزئي التي تميز المتلازمات المعجمية تسمح بتغيير تركيبها تغييراً مقولياً، وهو ما يفضي إلى تغيير صورة عنصر واحد أو أكثر من العناصر المكوّنة لها فتفشل الآلة في استقصائها وجمعها من بطون النصوص رغم وجودها في قاعدة البيانات، فالمتلازمة المعجمية "ألقي عصاه" بترميزها الشكلي المقترح 1V2N3FUN(prn)/1350-2 لن يتعرف عليها الحاسوب إذا تغيرت الصورة الشكلية للفعل فيها مثلاً، لتصبح "يلقي عصاه"، فالعنصر المعجمي V1 باعتباره ترميزاً شكلياً لـ "ألقي" غير موجود بالنسبة إلى الآلة؛ لأن صورته مخالفة لـ "يلقي".



٢-٥- في جدولة المتلازمات المعجمية وأهمية ملفات قواعد البيانات

نحسب أن النظر في هذه المسألة يتطلب معالجة مرحلية استرسالية نراعي فيها حاجات العقل الصناعي إلى التَّظْمَنَةِ التي نراها قائمة على الجدولة أساساً، هذا من ناحية، وإلى توسيع قاعدة ملفات البيانات وربطها بالمقولات المعجمية والأجزاء المقولية من ناحية أخرى، لذلك نقترح جدولة المتلازمات استناداً إلى تجليات التغير المفترضة فيها، على أن نميز في الترميز الشكلي العنصر المتغير من العناصر الثابتة بتغليظ الخط، ومثال ذلك:

أ. متلازمات متكلسة كلياً، مثل:

٨. بلغت القلوب الحناجر، أي اشتد الهلع.

ب. متلازمات ذات تغير بسيط، مثل:

٩. وقع في أمّ خنُور، أي في نعمة أو شِدَّة (الأعرابي، ١٩٦١، ص. ١٨٢)

يمكن أن تؤوّل إلى "وقعت في أمّ خنُور". فلا يتغير فيها إلا الضمير.

ج. متلازمات ذات تغير مركب، غير قائم على تناسب طردي، مثل:

١٠. ألقى بينهم العداوة والبغضاء، أي فتن.

يمكن أن تؤوّل إلى "أَلْقَتْ بَيْنَهُمَا الْعَدَاوَةَ وَالْبَغْضَاءَ" لكن التغير ليس من جنس "التناسب الطردي" هنا، أي أن تغير مقولة الفعل مثلاً لا يفضي إلى تغير أي عنصر آخر.

د. متلازمات ذات تغير مركب قائم على تناسب طردي، مثل:

١١. ضرب بذقنه الأرض، أي طأطأ جُبناً.

يمكن أن تؤوّل إلى "ضربت بذقنها الأرض" فيتغير الضمير المتصل تبعاً لتغير مقولة الجنس في الفعل.

على أن الاقتصار على جدولة المتلازمات المعجمية استناداً إلى تجليات التغيرات المفترضة لن يفي بالغرض على النحو المطلوب في المعالجة الآلية، وعليه نقترح جدولة جزئية ضمن

الجدولة الرباعية الكبرى السابقة نرى وجوب استنادها إلى معيار الانتماء المقولي للعنصر أو العناصر المعجمية المتغيرة.

وهذه المرحلة الثانية من الجدولة تخص المتلازمات التي يكون التغير فيها بسيطاً والمتلازمات ذات التغير المركب القائم على تناسب طردي والمتلازمات ذات التغير المركب غير القائم على تناسب طردي فحسب، لأن الصنف الأول من المتلازمات المعجمية وهي تلك التي لا يتغير أي عنصر من عناصرها لا يمكن إخضاعها إلى جدولة جزئية. ونُمثل لهذا التصوُّر بما يلي، على أننا سنقتصر على نماذج من المقولات في كل صنف حتى لا يكون بحثنا متشعباً فيخرج عن حدود عملنا.

٥-٢-١- معالجة المتلازمات ذات التغير البسيط وترميزها الشكلي

يتفرع هذا الضرب من المتلازمات نظرياً إلى خمسة أنواع استناداً إلى معيار الانتماء المقولي بعضها موجود بالقوة، وبعضها الآخر موجود بالفعل، ونقدم فيما يلي ترميزاً شكلياً لنماذج تمثيلية لمتلازمة ذات تغير بسيط مع مراعاة اختلاف الانتماء المقولي للعنصر المتغير، وننبه منذ البداية إلى أننا سنحذف فيما يلي من العمل، الأرقام المرجعية للمتلازمة لارتباطها بموقعها في المدونة، كما أننا سنحذف الرمز الرياضي المحيل إلى نوعها لعدم الحاجة إلى بيان النوع:

١٢. ضرب في الأرض — $/1V(slm)2FUN(lnk)3N/$

الترميز الشكلي

١٣. حَنَكَةُ التَّجَارِب — $/1V2FUN(prn)3N/$

الترميز الشكلي

١٤. تُضْرَبُ إِلَيْهِ أَكْبَادُ الْإِبِل — $/1V2FUN(lnk)3FUN(prn)4N5N/$

الترميز الشكلي

١٥. مشى القهقري — $/1V(naq-y)2N/$

الترميز الشكلي



في المتلازمتين (١٢) و (١٥) ما يمكن أن يتغير هو العنصر المعجمي الأول، وينتمي إلى مقولة الفعل، تتفرع مقولة الفعل إلى أربعة أجزاء مقولية هي:

١. **الفعل الثلاثي المجرد**، وينقسم إلى صحيح سالم ومهموز ومضاعف، ومعتل ينقسم بدوره إلى مثال وأجوف وناقص ولفيف بنوعيه.

٢. **الفعل الثلاثي المزيد**، وينقسم بدوره إلى ثلاثة أقسام حسب عدد حروف الزيادة.

٣. **الأفعال الرباعية المجردة.**

٤. **الأفعال الرباعية المزيدة** - لذلك ميزناه من غيره من العناصر بتغليظ رمزه وهو $1V(slm)$ في المتلازمة (١٢) و $1V(naq)-y$ في المتلازمة (١٥) وحددنا الجزء المقولي الذي ينتمي إليه وهو الصحيح السالم في المتلازمة (١٢) ورمزنا له بـ (slm) والناقص في المتلازمة (١٥) الذي رمزنا له بـ $(naq-y)$ حتى تتمكن الآلة من رصدهما والتعرف عليهما إذا طرأ على مقولتيهما التصريفتين تغيير، وذلك بالاستناد إلى ملفات بيانات هي مدخلات تسبق المعالجة الآلية.

أما في المتلازمتين (١٣) و (١٤) فالعنصر المعجمي القابل للتغيير هو الضمير المتصل، لذلك غلظنا رمزه وهو $2FUN(prn)$ في المتلازمة (١٣) وهو $3FUN(prn)$ في المتلازمة (١٤) وحددنا الجزء المقولي الذي ينتمي إليه ضمن مقولة الأداة وهو الضمير في المتلازمتين (١٣) و (١٤) ورمزنا له بـ (prn) وبذلك يمكن للآلة التعرف على المتلازمة بالعودة إلى ملفات قاعدة البيانات حتى إن وردت هذه المتلازمة على أنحاء مختلفة مثل "حَنَكْتُها التجارب" أو "حَنَكْتُهم التجارب" بالنسبة إلى المتلازمة (١٣) و "تُضَرَّبُ إليهم أكباد الإبل" أو "تُضَرَّبُ" إليه أكباد الإبل" في المتلازمة (١٣).

استبعاداً لمقتضيات الجدولة، نرى أنه رغم أن المتلازمات المعجمية الأربعة السابقة تنتمي إلى صنف المتلازمات المعجمية ذات التغيير البسيط لقابلية عنصر واحد منها فقط للتغيير، فإن المتلازمتين (١٢) و (١٣) لا يمكن إدراجهما في التصنيف الجدولي الجزئي نفسه؛ لاختلاف الانتماءات المقولية للعناصر المعجمية المتغيرة فيها، وكذلك هو الأمر بالنسبة إلى المتلازمتين (١٤) و (١٥)، في المقابل، يجب إدراج المتلازمتين (١٣) و (١٤) في الجدول الجزئي

نفسه؛ لانتلاف الانتماء الصيغي للعنصرين القابلين للتغيير فيهما، وهو مقولة الأداة. أما بالنسبة إلى المتلازمتين (١٢) و (١٥) فرغم جواز إدراجهما في الجدول الجزئي نفسه؛ لأن العنصرين القابلين للتغيير فيهما يأتلفان في الانتماء المقولي، وهو مقولة الفعل، فإننا نرى أنه يمكن مواصلة التجزئة الجدولية للتمييز بين الفعل السالم والمعتل. ونؤكد على أن الرصد الآلي عند التغير لا يتم إلا بالاستناد إلى ملفات بيانات تشكل الذاكرة الحاسوبية التي تعوضه عن الحدس الفطري عند ابن اللغة، ونود الإشارة إلى أننا سننظر في مسألة ملفات البيانات فيما سيأتي من هذا العمل.

٥-٢-٢- معالجة المتلازمات ذات التغير المركب، غير القائم على تناسب طردي وترميزها الشكلي

لن تشدّ معالجة المتلازمات ذات التغير المركب، غير القائم على تناسب طردي عن المنهج الذي اتبعناه في معالجة المتلازمات ذات التغير البسيط، ونمثل لها بما يلي:

١٦. ضرب به عرض الحائط ————— ◀ 1V(slm)2FUN(lnk)3FUN(prn)4N5N/

الترميز الشكلي

في المتلازمة السابقة، جليّ أن العنصرين المعجميين اللذين يقبلان التغير من عناصرها لا تربط بينهما علاقة تناسب طردي، أي لا يترتب عن تغير أحدهما تغير العنصر الآخر. والعنصران المتغيران هما الفعل ورمزنا له بـ 1V(slm) والضمير ورمزنا له بـ 3FUN(prn) مع تغليظ الخط لتمييزهما عن بقية العناصر، باعتبارها عناصر ثابتة، ولم نغفل عن تحديد الانتماء المقولي الجزئي للفعل فهو فعل صحيح سالم ورمزه (slm) فضلاً عن ضبط المقولة المعجمية الكبرى للضمير المتصل ورمزه (prn) وهي مقولة الأداة.

ونحسب أن التصنيف الجدولي الجزئي المتعلق بصنف المتلازمات ذات التغير المركب، غير القائم على تناسب طردي لا يقبل إدراج متلازمة يتغير فيها عنصر معجمي ينتمي إلى مقولة الاسم وآخر ينتمي إلى مقولة الصفة، على سبيل التمثيل في الجدول الجزئي الذي تنتمي إليه المتلازمة السابقة باعتبار اختلاف الانتماءات المقولية للعناصر المعجمية القابلة للتغير، إذ لا يندرج في جدولها إلا المتلازمات التي ينتمي العنصران القابلان للتغير فيها إلى مقولتي الفعل والأداة.



ولا يفوتنا في هذا الإطار التذكير بضرورة توسيع ملفات قاعدة البيانات حتى لا تفشل الآلة في رصد المتلازمة عندما يطرأ على عنصرها أي تغيير.

٥-٢-٣- معالجة المتلازمات ذات التغير المركب، القائم على تناسب طردي ورميزها الشكلي

لمعالجة هذا الصنف من المتلازمات نستند إلى النماذج التمثيلية التالية:

$$١٧. \text{لَقِيَ حَنَفَهُ} \leftarrow /1V(naq-y)2N3FUN(prn)/$$



الترميز الشكلي

$$١٨. \text{أكل خُلّالته، أي كان بخيلا} \leftarrow /1V(hf)2N3FUN(prn)/$$



الترميز الشكلي

$$١٩. \text{بضاعتهم رُدّت إليهم} \leftarrow /1N2FUN(prn)3V(+suf)4FUN(lnk)5FUN(prn)/$$



الترميز الشكلي

تتفق المتلازمات السابقة وهي التي أدرجناها ضمن صنف المتلازمات ذات التغير المركب، القائم على تناسب طردي مع المتلازمات ذات التغير المركب، غير القائم على تناسب طردي في وجود عنصرين معجميين قابلين للتغيير إلا أن هذا الصنف يختلف عن سابقه بوجود علاقة تناسب طردي بين العنصرين، وبما أن تغير العنصر الواحد يستلزم تغير العنصر الثاني فإننا تجاوزنا الاكتفاء بتعليظ رمزي العنصرين المتغيرين في المتلازمة إلى تمثيل العلاقة الترابطية بخط واصل.

بناء على المنهج المتبع في الجدولة الجزئية الذي استندنا فيه إلى معيار الانتماء المقولي نصنف المتلازمات الثلاث السابقة (١٧) و(١٨) و(١٩) في جدولين جزئيين، الأول يضم المتلازمتين (١٧) و(١٨) لاشتراكهما في التغير القائم على التناسب الطردي بين مقولتي الفعل والأداة، أما المتلازمة (١٩) فلا يمكن إدراجها في الجدول الجزئي نفسه رغم قيام التغير فيها على تناسب طردي؛ لأن هذا التناسب يربط بين عنصرين من المقولة نفسها، هي مقولة الأداة وتحديداً الجزء المقولي الضمير.

لا تفوتنا الإشارة إلى وجود متلازمات التغير فيها معقد، يجمع بين التغير القائم على غير التناسب الطردي والتغير القائم على التناسب الطردي، ونُقدّر أن هذا الصنف قليل، ونُمثّل له بالمتلازمة التالية:

$$٢٠. \text{ردّه على عَقْبِهِ} \leftarrow /1V(md)2FUN(prn) 3FUN(lnk)4N5FUN(prn)/$$

الترميز الشكليّ

في المتلازمة (٢٠) لا يثبت إلا عنصر الأداة (على) فقط، أما العنصر المعجمي الأول (ردّ) الذي ينتمي إلى مقولة الفعل وتحديدًا المضاعف والذي رمزنا له بـ $1V(md)$ فإن تغييره لا يترتب عنه تغيير أي عنصر آخر، لذلك فهو يشكّل مع بقية العناصر تغييراً غير قائم على تناسب طردي، أما $2FUN(prn)/4N/5FUN(prn)$ ، وتمثل العناصر المعجمية التالية على التوالي /هـ/ و/عقب/ و/هـ/ فإن التغير فيها قائم على تناسب طردي، فتغير العنصر الواحد يترتب عنه تغير بقية العناصر.

رغم أن الجدولة العامة والجدولة الجزئية اللتين اعتمدناهما في العمل باعتبارهما مقومين أساسيين من مقومات الحوسبة لأنهما يسمحان بنظمنة المتلازمات المعجمية، ورغم أهمية الترميز الشكلي الذي اعتمدناه استجابةً لحاجة العقل الصناعي إلى الاختصار، فإن كل ما سبق لن يفي بالغرض ولن يمكّن الحاسوب من التعرف على المتلازمات المعجمية عند تغير الصورة الشكلية لعنصر أو أكثر من عناصرها، إذ لا بد من إنشاء ملفات قاعدة بيانات يعود إليها الحاسوب حتى يتعرف على العنصر المعجمي بعد أن طرأ عليه تغيير ما. لذلك نرى أن نجاح عملية الرصد الآلي للمتلازمات المعجمية ودقتها تتحدد بالمدخلات التي تسبق المعالجة الحاسوبية وهي ملفات البيانات التي تكون الذاكرة الصناعية أو الحسّاس الصناعي -إن صح التعبير- فأى خلل أو نقص في تلك المدخلات يؤثر سلباً في المخرجات. وبما أن العناصر المعجمية المكوّنة للمتلازمات المعجمية يمكن أن تنتمي إلى مختلف المقولات المعجمية، فإننا نقترح إفراد كل مقولة بملفات بيانات معينة، وإفراد كل جزء مقولي بملفات بيانات خاصة به، إلا أننا لن نتوسع في عرض هذه الملفات؛ لأن ذلك يتجاوز حدود مجال بحثنا. لكننا مع ذلك سنبين أهمية ملفات قاعدة البيانات في الرصد الآلي استناداً إلى نموذج



تمثيلي هو متلازمة تنتمي إلى صنف المتلازمات التي يقوم التغير فيها على تناسب طردي،

وملفات البيانات المقترحة هي التالية:

- أ. ملفات بيانات مقولة الفعل: تتضمن (ملف بيانات الجذور المستعملة-ملف بيانات الجذوع الرئيسة الثلاثية مع تفرعها حسب جنس الحرف-ملف بيانات الجذوع الثلاثية الفرعية مع تفرعها حسب عدد حروف الزيادة-ملف بيانات الجذوع الرباعية الأصلية-ملف بيانات الجذوع الرباعية الفرعية مع تفرعها حسب عدد حروف الزيادة-ملف بيانات الزوائد التصريفية بمختلف أنواعها حسب مواقعها وإحالتها على مقولتي الجنس والعدد- قواعد تصريف الأفعال العربية بمختلف أنواعها. ولا تفوتنا الإشارة في هذا الصدد إلى أن التصريف العربي يتسم بالاطراد والانتظام في الأغلب الأعم، وهو ما يُيسّر سبيل الحوسبة، وحتى ما بدا شاذاً يمكن رده إلى اطراد جديد.
- ب. ملفات بيانات مقولة الاسم: يتضمن (ملف أسماء الأعلام- ملف الأسماء المركبة- ملف الأسماء المشتقة- ملف قواعد تثنية الأسماء في المحلات الإعرابية الثلاثة- ملف قواعد جمع الأسماء جمعاً مذكراً ومؤنثاً سالماً في المحلات الإعرابية الثلاثة- ملف ربط الأسماء المفردة بجموع التكسير الموافقة لها...)
- ج. ملفات بيانات مقولة الصفة: يتضمن (ملف الصفات بأنواعها-ملف قواعد تثنية الصفات-ملف قواعد جمع الصفات)
- د. ملفات بيانات مقولة الظرف: يتضمن (ملف ظروف الزمان-ملف ظروف المكان)
- هـ. ملفات بيانات مقولة الأداة: يتضمن (ملف الضمائر المنفصلة رفعاً ونصباً-ملف الضمائر المتصلة-ملف الأدوات التي تسبق الأفعال ولا تدمج فيها-ملف الأدوات التي تسبق الأفعال وتدمج فيها-ملف أدوات الربط)

بناءً على ما سبق في هذا العمل، نحسب أن الرصد الآلي للمتلازمات المعجمية بهدف جمعها والاستفادة منها في الصناعة المعجمية يجب أن يمر بمرحلتين- خاصة بالنسبة إلى المتلازمات التي يطرأ على تركيبها تغير مقولي- هما مرحلة التحليل ومرحلة التأليف

والجدولة، ولا غنى لنا في كليهما عن ملفات البيانات التي تعتمد عليها الآلة لتجاوز البنية السطحية المرصودة للمتلازمة والوصول إلى الكشف عن تغيرها النظامي باعتباره تحققاً نظامياً جديداً للمتلازمة الأصلية.

ونمثل لما سبق بالمتلازمة التالية ذات التغير القائم على التناسب الطردي:

٢١. لقي حتفه /1V(naq-y)2N3FUN(prn)/

الترميز الشكلي

يمكن أن ترد المتلازمة السابقة في ثنايا النصوص على الصورتين التاليتين "لقيت حتفها" أو "تلقيان حتفهما" فتكون معالجتهم حاسوبياً على النحو التالي:

أ. مرحلة التحليل: يُفترض في هذه المرحلة الاعتماد على المنهج الاحتمالي، كالتالي:

احتمال ١: "لقيت" و "تلقيان" كل منهما عنصر معجمي (٠)

احتمال ٢: لقيت=لقي+(ت) (*)

تلقيان=(ت)+لقي+(ان) (*)

نرمزب (٠) إلى خطأ الاحتمال ١ وب (*) إلى صحة الاحتمال ٢

التوصل إلى الاحتمال الصحيح تم بالعودة إلى ملفات قاعدة البيانات وتحديد ملف الأصول الجذعية وملف اللواحق التصريفية، ماضياً بالنسبة إلى "لقيت"، وملفي السوابق واللواحق التصريفية، مضارعاً بالنسبة إلى "تلقيان".

ثم يتم تحليل "حتفها" و "حتفهما" على النحو التالي اعتماداً على ملفات البيانات أيضاً:

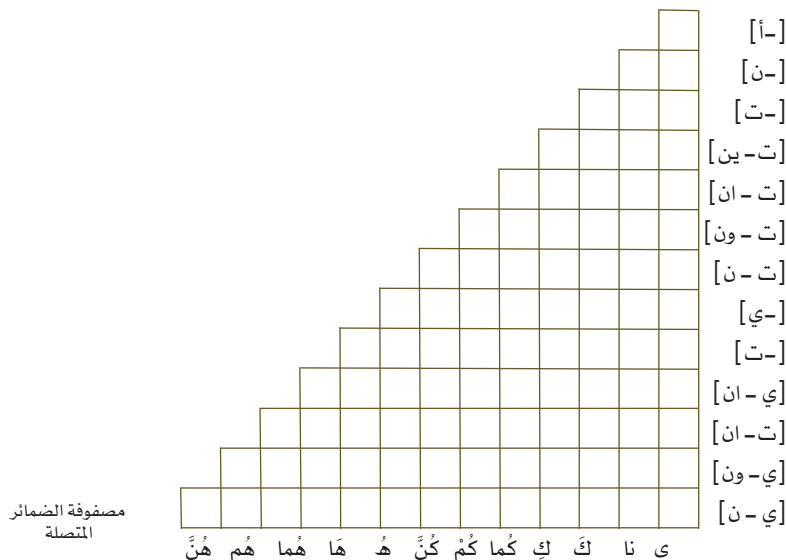
احتمال ١: "حتفها" و "حتفهما" عنصران معجميان (٠)

احتمال ٢: حتفها=حتف+ها (*)

حتفهما=حتف+هما (*)

ب. مرحلة التأليف والجدولة: تستند إلى مبدأ التناسب الطردي فعل /ضمير لتفسير العلاقة المعجمية الجوهرية بين المتلازمتين المرصودتين "لقيت حتفها" أو "تلقيان

مصنوفة اللواحق التصريفية المعبرة عن جميع المقولات التصريفية في صيغة المضارع المرفوع



*** ملاحظة: الشكل من اعداد الباحثة**

يقوم الرسمان البيانيان السابقان على تمثيل لمصفوفة الضمائر المتصلة في مستوى أفقي وتمثيل لمصفوفة الزوائد التصريفية المعبرة عن جميع المقولات التصريفية في صيغة الماضي في الرسم الأول، وفي صيغة المضارع المرفوع في الرسم الثاني لفعل صحيح سالم في مستوى عمودي، وتمكن نقاط التقاطع بين المصفوفتين لا من التعرف على العنصر الثاني المتغير في المتلازمة المعجمية بعد التعرف على التغير في العنصر الأول فحسب، بل من التنبؤ به أيضًا لما بين المصفوفتين من تناسب طردي وهو ما يؤدي إلى نجاح رصد المتلازمة آليًا رغم ما لحقها من تغير سطحي.



٦- الخاتمة

بنيت مراحل هذا العمل على الاسترسال بمفهومه العرفاني، فأفاد التطبيق اللاحق من التنظير السابق لذلك عرضنا في بداية هذا المشروع البحثي أهم مقومات حوسبة اللغات الطبيعية فتوقفنا عند ثلاث ركائز أساسية وهي النظمّة القائمة على الجدولة والتميز القائم على الاختصار وإنشاء ملفات قواعد البيانات التي تشكل الذاكرة الحاسوبية، بعث هذا على ضبط خيارنا النموذجي الواعي -المتلازمات المعجمية- ومراجعته مصطلحاً ومفهوماً بضبط خصائصه التمييزية الجامعة المانعة وتخليصه من التضارب الشائع في جل الدراسات.

ولأن معالجة الظاهرة يجب أن تسبق بالكشف عن التحديات، فقد وقفنا على أن خصائص المتلازمة في حد ذاتها تعيق حوسبتها، من هنا عملنا على معالجة هذه التحديات التي تحول دون جمعها معالجة مرحلية بالمزاوجة بين الجهد البشري، والمعالجة الآلية، فرفعنا التحدي المترتب عن بنيتها الشكلية اللا توزيعة في نسيجها التركيبي، وتجاوزنا في ضبط عناصر المعجمية البنية السطحية إلى البنية العميقة استناداً إلى خيار لساني معجمي، ثم سعينا إلى إيجاد نسق نظامي للمتلازمات المعجمية يمكّن من جدولتها جدولّة ثنائيّة مرحلية -بالنظر إلى التغيرات الممكنة- هي جدولّة عامة تستند إلى تجليات التغير فيها، ثم جدولّة جزئية تستند إلى الانتماءات المقولية للعناصر المعجمية القابلة للتغير، ولم نغفل في جميع مراحل العمل على الترميز الشكلي، باعتباره من متطلبات المعالجة الآلية مستأنسين بنماذج تمثيلية من المتلازمات المعجمية. ولأن دقة المخرجات رهينة اتساع ملفات قاعدة البيانات، فقد اقترحنا ربط كل مقولة معجمية كبرى أو جزء مقولي بملفات بيانات تشكل الحدس الصناعي والذاكرة الحاسوبية التي تستند إليها الآلة لرصد المتلازمة، وقد تغير تركيبها المقولي. انتهينا من كل ما سبق إلى أن الرصد الآلي للمتلازمات، خاصة تلك التي يطرأ عليها تغيير يقوم على ثنائية التفكيك والتأليف استناداً إلى ملفات البيانات في إطار الانتظام الجدولي فضلاً عن الترميز الشكلي؛ لذلك اخترنا هذه النتيجة مستأنسين بنموذج تمثيلي بين عناصره المتغيرة تناسب طردي، فبيناً أن المعالجة الآلية لا تفضي إلى رصد المتلازمة فحسب، بل إلى التنبؤ بالتغيرات المترتبة عن تغير عنصر فقط من عناصرها.

وعليه، فهذا العمل الذي سعيينا فيه إلى الجمع الآلي للمتلازمات المعجمية خدمة للصناعة المعجمية - نود الإشارة إلى أننا لم نعرّح حسب ما اطلعنا عليه على أي عمل عالج المتلازمات المعجمية انطلاقاً من مراجعة مصطلحية مفهومية نظيرياً وصولاً إلى تطبيق إجرائي - فانطلقنا فيه من أسس لسانية نظرية واجبة الوجود في دراسة بينية لوضع حدود جامعة مانعة للمتلازمات وعمدنا فيه إلى مجاوزة الوصف إلى التوصيف منتهجين منهجاً شكلياً صورياً هو من متطلبات العقل الصناعي انتهينا من خلاله إلى أن الرصد الآلي للمتلازمات يقوم على ثنائية التفكير والتأليف استناداً إلى ملفات بيانات موسعة، فتكون دقة المخرجات على قدر شمولية المدخلات.

أخيراً، نخلص من خلال هذا العمل إلى جملة من التوصيات التي نراها ضرورية لتطوير البحث في الصناعة القاموسية البينية، ونجملها فيما يلي:

- أ. ضرورة المراجعة المصطلحية والمفهومية للفظ "الكلمة".
- ب. ضرورة الاستناد إلى مرتكزات لسانية نظرية تسبق المعالجة الحاسوبية.
- ج. أهمية تكثيف الجهود لتجاوز وصف لغتنا إلى توصيفها بما يتلاءم والمعالجة الحاسوبية.
- د. أهمية العناية بالمتلازمات المعجمية باعتبارها ذخيرة لغوية تعكس هوية الأمة، وجب جمعها ومعالجتها قاموسياً اعتماداً على الوسائل التقنية.
- هـ. ضرورة تكاتف الجهود وتكثيفها واستثمار مخرجات الملتقيات العلمية للمزاوجة بين الصناعة المعجمية والمعالجة الآلية.

وختاماً نعتقد أن الدراسات البينية اللسانية الحاسوبية تظل أعملاً تقريبية قابلة للتطوير بمزيد استنطاق المعرفة اللغوية اللاواعية فيه لتحديد نسبة النجاح.



المراجع العربية

الأعرابي، أبو مسحل. (١٩٦١). النوار. (عزة حسن، محقق). مطبوعات مجمع اللغة العربية بدمشق.

ابن مراد، إبراهيم. (١٩٩٣). المعجم العلمي العربي المختص. دار الغرب الإسلامي.

ابن مراد، إبراهيم. (١٩٩٧). مقدمة لنظرية المعجم. دار الغرب الإسلامي.

ابن مراد، إبراهيم. (٢٠٠٦). الوحدة المعجمية بين الأفراد والتضام والتلازم. مجلة الدراسات المعجمية. (٥) ٣٢-٢٣.

ابن منظور، أبو الفضل جمال الدين. (١٩٩٠). لسان العرب. دار صادر.

الغانمي، سمر. (٢٠٢٤). في العلاقة بين البنية ودلالاتها: دراسة لنماذج من العبارات المسكوكة. مركز الأطرش للكتاب المختص.

المهيوي، عبد العزيز بن عبد الله. (٢٠٢٤). التحليل الصرفي. في عبدالله الفيافي (محرر). مدخل إلى اللسانيات الحاسوبية. (ص ص ٤٣-٧٤) مجمع الملك سلمان العالمي للغة العربية.

الموسى، نهاد. (٢٠٠٠). العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية. المؤسسة العربية للدراسات والنشر.

بنت براك العنزي، بدرية. (٢٠١٩). نحو معجم للمتلازمات اللفظية في المعاجم العربية المعاصرة: دراسة تحليلية في ضوء مدونة لغوية. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية.

حماده، سلوى. (٢٠٠٩). المعالجة الآلية للغة العربية. دار غريب للطباعة والنشر.

روبي محمد بن عبد الرحمن، أحمد. (٢٠٢٤). التحليل النحوي. في عبدالله الفيافي (محرر). مدخل إلى اللسانيات الحاسوبية. (ص ص ٧٥-١٠٣) مجمع الملك سلمان العالمي للغة العربية.

السمعلي الزواري، عواطف. (٢٠١٩). دور المكوّن الصرفي في انتظام المعجم. مركز الأطرش للكتاب المختص.

السمعلي الزواري، عواطف. (٢٠٢٥ أبريل ٢٤-٢٦). فك اللبس الدلالي في ضوء اللسانيات الحاسوبية من خلال نماذج من المشترك الصيغي [عرض ورقة]، الخلفيات العلمية للنظريات الدلالية، تونس.

آل طه، هدى. (٢٠٠٥). النظام الصرفي في العربية في ضوء اللسانيات الحاسوبية، مثل من جمع التكسير. [رسالة دكتوراه غير منشورة]. جامعة الأردن.

الوديني، علي. (٢٠٢١). المتلازمات المعجمية (حدّها ومعايير ضبطها). مجلة الفكر اللساني. (١) ٨٣-٩٢.

مجمع اللغة العربية بالقاهرة. (٢٠٠٤). المعجم الوسيط. مكتبة الشروق الدولية.

محمّدي، منيرة. (٢٠٢٢). إشكاليات تجذيع الكلمة حاسوبياً. [رسالة ماجستير غير منشورة]. جامعة منوبة.

رومنة المراجع العربية

Al-A' rābī, Abū Muṣṣil. (1961). Al-Nawādir ('Izza Ḥasan, Ed.). Maṭbū'āt Majma' al-Lugha al-'Arabiyya bi-Dimashq.

Ibn Murād, Ibrāhīm. (1993). Al-Mu'jam al-'Ilmī al-'Arabī al-Mukhtaṣṣ. Dār al-Gharb al-Islāmī.

Ibn Murād, Ibrāhīm. (1997). Muqaddima li-Nazariyyat al-Mu'jam. Dār al-Gharb al-Islāmī.

Ibn Murād, Ibrāhīm. (2006). Al-Waḥda al-Mu'jamiyya bayna al-Ifrād wa-al-Taḍām wa-al-Talāzum. Majallat al-Dirāsāt al-Mu'jamiyya, (5), 23–32.

Ibn Manẓūr, Abū al-Faḍl Jamāl al-Dīn. (1990). Lisān al-'Arab. Dār Ṣādir.

Al-Ghānimī, Samar. (2024). Fī al-'Alaqa bayna al-Binya wa-Dalālatihā: Dirāsa li-Namādhij min al-'Ibārāt al-Maskūka. Markaz al-Aṭrash lil-Kitāb al-Mukhtaṣṣ.



- Al-Mahyūbī, ‘Abd al-‘Azīz ibn ‘Abd Allāh. (2024). Al-Taḥlīl al-Šarfī. In ‘Abd Allāh al-Fayfī (Ed.), Madkhal ilā al-Lisāniyyāt al-Ḥāsūbiyya (pp. 43–74). Majma‘ al-Malik Salman al-‘Ālamī lil-Lugha al-‘Arabiyya.
- Al-Mūsā, Nihād. (2000). Al-‘Arabiyya naḥwa Tawsīf Jadīd fī Ḍaw’ al-Lisāniyyāt al-Ḥāsūbiyya. Al-Mu‘assasa al-‘Arabiyya lil-Dirāsāt wa-al-Nashr.
- Bint Barāk al-‘Anazī, Badriyya. (2019). Naḥwa Mu‘jam lil-Mutalāzamāt al-Lafziyya fī al-Ma‘ājim al-‘Arabiyya al-Mu‘āšira: Dirāsa Taḥlīliyya fī Ḍaw’ Madūna Lughawiyya. Markaz al-Malik ‘Abd Allāh ibn ‘Abd al-‘Azīz al-Duwalī li-Khidmat al-Lugha al-‘Arabiyya.
- Ḥamāda, Salwā. (2009). Al-Mu‘ālaja al-Āliyya lil-Lugha al-‘Arabiyya. Dār Gharīb li-al-Ṭibā‘a wa-al-Nashr.
- Rūbī Muḥammad ibn ‘Abd al-Raḥmān, Aḥmad. (2024). Al-Taḥlīl al-Naḥwī. In ‘Abd Allāh al-Fayfī (Ed.), Madkhal ilā al-Lisāniyyāt al-Ḥāsūbiyya (pp. 75–103). Majma‘ al-Malik Salman al-‘Ālamī lil-Lugha al-‘Arabiyya.
- Al-Sam‘alī al-Zuwwārī, ‘Awāṭif. (2019). Dawr al-Mukawwin al-Šarfī fī Intizām al-Mu‘jam. Markaz al-Atrash lil-Kitāb al-Mukhtaṣṣ.
- Al-Sam‘alī al-Zuwwārī, ‘Awāṭif. (2025). Fakk al-Labs al-Dalālī fī Ḍaw’ al-Lisāniyyāt al-Ḥāsūbiyya min Khilāl Namādhij min al-Mushtarak al-Šayghī: al-Khalfiyyāt al-‘Ilmiyya li-al-Nazariyyāt al-Dalāliyya (pp. 657–688). Markaz al-Atrash lil-Kitāb al-Mukhtaṣṣ.
- Āl Ṭahā, Hudā. (2005). Al-Nizām al-Šarfī fī al-‘Arabiyya fī Ḍaw’ al-Lisāniyyāt al-Ḥāsūbiyya: Mathal min Jam‘ al-Taksīr [Unpublished doctoral dissertation]. Jāmi‘at al-Urdun.
- Al-Wadrānī, ‘Alī. (2021). Al-Mutalāzamāt al-Mu‘jamiyya (Ḥadduhā wa-Ma‘āyir Ḍabṭihā). Majallat al-Fikr al-Lisānī, (1), 83–92.
- Majma‘ al-Lugha al-‘Arabiyya bi-al-Qāhira. (2004). Al-Mu‘jam al-Wasīṭ. Maktabat al-Shurūq al-Duwalīyya.
- Muḥammadī, Munīra. (2022). Ishkāliyyāt Tajdhī‘ al-Kalima Ḥāsūbiyyan [Unpublished master’s thesis]. Jāmi‘at Manūba.

المراجع الأجنبية

- Corpus Linguistics and the Description of English*. (2005). Edinburgh University Press.
- De Saussure, F. (1996). *Cours de linguistique générale* (Bally, C., & Sechehaye, A., Éds.). Paris: Payot.
- Di Sciullo, A., & Williams, E. (1987). *On the definition of word*. Cambridge, London: The MIT Press.
- Dubois, J. (2002). *Dictionnaire de linguistique*. Paris: Larousse-Bordas/VUEF.
- Kahane, S., & Polguère, A. (2001). *Un langage formel d'encodage des fonctions lexicales et son application à la modélisation des collocations*. Paris: Atala.
- La Plante, P. (2000). *Dictionary of computer science, engineering and technology*. Boca Raton, London, New York, Washington, DC: CRC Press.
- Tutin, A., & Grossmann, F. (2001). Collocations régulières et irrégulières. *Revue française de linguistique*, (1).



نبذة تعريفية عن الباحث

Author Biodata Awatef Samaali Zouari is an Assistant Professor of Theoretical and Applied Lexical Linguistics in the Department of Arabic Language at the Faculty of Human and Social Sciences of Tunis (9 Avril), University of Tunis, Republic of Tunisia. She holds a Ph.D. in Theoretical and Applied Lexical Linguistics from the University of Manouba (2018). Her research interests focus primarily on computational lexicography, with particular attention to phonological and morphological processing.	بيانات الباحث: عواطف السمعلي الزواري، أستاذة مساعدة في اللسانيات المعجمية النظرية والتطبيقية في قسم اللغة العربية في كلية العلوم الإنسانية والاجتماعية بتونس (٩ أبريل)، جامعة تونس، الجمهورية التونسية، حاصلة على درجة الدكتوراه في اللسانيات المعجمية النظرية والتطبيقية، من جامعة منوبة عام ٢٠١٨، تدور اهتماماتها البحثية حول المعالجة الآلية المعجمية عامة والمعالجة الآلية الصوتية والصرفية خاصة.
معرف أوركيد (ORCID) : 0009-0006-4847-2578	
Email: awatefsamaali@yahoo.fr	

الملاحق

ملحق (أ): المصطلحات العربية وترجماتها

المصطلح العربي	الترجمة الإنجليزية
متلازمات معجمية	Lexical collocations
ترميز رياضي	Mathematical encoding
بنية لا توزيعية	Non-compositional structure
جمع	Compilation
صناعة معجمية	Lexicography
قاعدة بيانات	Database
مدونة	Corpus
وسائل إدخال	Input Devices
المعجم اللساني	Lexicon
العبارات	Locutions
البنوك الشجرية	Treebanks
المدونات الموصفة نحويًا	Parsed Tagged Corpora
نصف آلية	Semi-automatic
القاموس	Dictionary
الرصد الآلي	Automatic detection

النماذج اللغوية الضخمة ومعالجة اللغة العربية: مراجعة منهجية

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

د. خليوي سامر خليوي العياضي

الجامعة الإسلامية، المدينة المنورة، المملكة العربية السعودية

نُشر في	قُبِلَ في	رُوجِعَ في	استقبل في
١٤٤٧/٠٥/٠٩	١٤٤٧/٠٤/١٨	١٤٤٧/٠٢/٠١	١٤٤٧/٠٢/٠٩

الملخص

يهدف هذا البحث إلى مراجعة الدراسات السابقة في ميدان النماذج اللغوية الضخمة ومعالجة اللغة العربية، بهدف الوقوف على المنجزات، وتحديد أبرز التحديات التي تواجه هذا المجال، إلى جانب استعراض الحلول المقترحة التي من شأنها تحسين أداء هذه النماذج في التعامل مع اللغة العربية. استخدم الباحث منهج المراجعة المنهجية لملاءمته طبيعة الدراسة، واقتصر على الدراسات المنشورة خلال السنوات الخمسة الأخيرة (٢٠٢٠-٢٠٢٤)، حيث بلغ عدد الدراسات التي حُلَّت (٨٢) دراسة، جُمِعت من قواعد البيانات التالية: Google Scholar و PubMed و IEEE. توصلت الدراسة إلى مجموعة من التوصيات، أهمها: ضرورة العناية بالبيانات من حيث الكم والنوع، وأهمية إشراك المتخصصين في اللغة العربية؛ نظراً لدورهم الفعال في تحسين جودة البيانات، كما شددت على أهمية دعم المؤسسات العلمية في تطوير نماذج لغوية عربية وقواعد بيانات متخصصة وشاملة.

الكلمات المفتاحية

النماذج اللغوية الضخمة، معالجة اللغة العربية، مراجعة منهجية.

Large Language Models and Arabic Natural Language Processing: A Systematic Review

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

Dr. Khalioi Samer Khalioi Al-Ayadhi

Islamic University, Madinah, Kingdom of Saudi Arabia

Received on	Revised on	Accepted on	Published on
03-08-2025	04-08-2025	10-10-2025	31-10-2025

ABSTRACT

This study aims to review recent research on large language models (LLMs) and Arabic language processing, highlighting current advancements, identifying major challenges, and exploring potential solutions to improve model performance with Arabic. A systematic review methodology was employed due to its suitability, focusing on studies published between 2020 and 2024. A total of 82 studies were analyzed, sourced from Google Scholar, IEEE, and PubMed. The review highlights key recommendations, including the need for high-quality and sufficient data, and the active involvement of Arabic language experts, who play a crucial role in ensuring linguistic accuracy. The study also emphasizes the significant role of academic institutions in addressing existing gaps by supporting the development of robust Arabic LLMs and specialized linguistic databases.

KEYWORDS

Large Language Models (LLMs), Arabic Natural Language Processing (Arabic NLP), systematic review.



١- المقدمة

يشهد العالم الحالي ثورة صناعية رابعة يقودها ويؤثر فيها الذكاء الاصطناعي الذي يُعد الأضخم في مسيرة تلك التحولات التي حدثت في العصر الحديث، وقد أشار Makridakis إلى أن التقنيات المُرتقبة لثورة الذكاء الاصطناعي خلال العقدين القادمين من المرجح أن تكون أكبر بعدة مرات من حجم تأثير الثورة الرقمية التي شهدناها بين عامي ١٩٩٥ و٢٠١٥، بل يضيف أن هذا التأثير قد يكون أكبر من تأثير الثورة الصناعية بأكملها (Makridakis, 2017, p. 47).

والغرض الذي جاء الذكاء الاصطناعي من أجله هو محاكاة الذكاء البشري في ما يملكه من قدرات في التحليل والتفسير وحل المشكلات، فانبهر المختصون لتمكين الآلة من هذا الهدف منذ منتصف القرن الماضي تحت يد آلن تورنج Alan Turing الذي طرح في أحد أرواقه المعيار الذي يمكن أن نحكم من خلاله إن كانت الآلة ذكية أم لا، وعُرف لاحقاً بما يسمى "باختبار تورنج" ويقوم على استقبال إجابتين، الأولى من إنسان، والأخرى من آلة فإن لم نستطع تحديد مصدرهما فهذا يعني أن الآلة وصلت للذكاء البشري؛ لأنها ماثلت رده (Jones & Bergen, 2024, p. 5183).

واللغة هي أحد أبرز القدرات البشرية التي سعى الباحثون إلى تمكين الآلة منها فهماً وتوليداً، وما زال المختصون والباحثون في محاولة تمكين الآلة من هذه القدرة منذ اختراع الحاسوب إلى أن جاءت النقطة المحورية التي كانت إيذاناً بتمكين الآلة من اللغة، عندما طرحت قوئل ورقة علمية بعنوان: الانتباه كل ما نحتاجه "attention all you need" عام ٢٠١٧؛ ليقدم الباحثون فيها بنية شبكية عصبية تُسمى بالمحولات "Transformers" التي أحدثت تحولاً كبيراً في مجال معالجة اللغات الطبيعية (NLP) والذكاء الاصطناعي بشكل عام (Vaswani et al., 2017).

ما يميز هذه البنية تَمَكُّن النموذج من أن يزن أهمية كل كلمة في السياق المُدخل وعلاقتها بالكلمات الأخرى، وهذا يعني أن النموذج يمكنه فهم السياق فهماً أفضل من

خلال تحديد الكلمات الأكثر صلة ببعضها، بغض النظر عن المسافة بينها، وهذا ما عالج عيوب الشبكات العصبية السابقة.

ومع هذا التحول في العتاد التقني بُنيت العديد من النماذج اللغوية الضخمة؛ لتمكين الآلة من اللغة فهما وتوليداً، فظهرت تباعاً بدءاً بـ GPT هذا النموذج الذي بُنيت عليه تقنية الـ ChatGPT عام ٢٠٢٢، وتلتها نماذج عدة في الساحة، وجاءت هذه النماذج متعددة اللغات حيث دُرِبت على لغات كثيرة، لتتجاوب مع مختلف اللغات، ولكن الملاحظ فيها، أن قدرتها على فهم وتوليد اللغة الإنجليزية تفوق بقية اللغات، خاصة تلك التي تُعد قليلة الموارد، أي أن بياناتها المتوفرة قليلة جداً، وإحدى هذه اللغات هي اللغة العربية، ويؤكد ذلك ما نشره موقع W3Techs المختص بإحصاء لغات المواقع الإلكترونية، ففي آخر إحصائية نشرها جاء ترتيب اللغة العربية في المرتبة الحادية والعشرين، مع تدني في النسبة بين سنتي ٢٠٢٣ و٢٠٢٥، ما يدل على تفوق لغات أخرى أقل منها في عدد المتحدثين والمكانة العالمية كما يشير الجدول (١):

جدول (١) نسبة المواقع حسب لغة المحتوى

المرتبة	اللغة	15- May-2023	18- March-2025
١	English	%٥٥,٥	%٤٩,١
٢	Spanish	%٥,٠	%٦,٠
٣	German	%٤,٣	%٥,٨
٤	Japanese	%٣,٧	%٥,١
٥	French	%٤,٤	%٤,٥
٦	Portuguese	%٢,٤	%٣,٩
.....			
٢١	Arabic	%٠,٧	%٠,٥

*المصدر: Adapted from Languages used on the Internet, Wikipedia (retrieved July 20, 2025)



ومع قلة موارد اللغة العربية إلا أن المختصين من أهلها وغير أهلها يسعون إلى تمكين الآلة منها؛ لما لها من مكانة عالمية. فهي إحدى اللغات الستة في اليونيسكو، وإحدى اللغات المعتمدة في الأمم المتحدة، هذا بالإضافة إلى ما لها من إرث حضاري وثقافي عريق.

وشهد ميدان تطبيق النماذج اللغوية الضخمة على اللغة العربية اهتماماً بحثياً وتطويرياً متنامياً في السنوات الأخيرة، حيث سعى الباحثون جاهدين؛ لتكييف النماذج العالمية أو بناء نماذج تراعي خصائصها، وتكون قادرة على أداء طيف واسع من المهام اللغوية. ومن تلك المهام التي عالجها الباحثون: مهام التصنيف النصي (Karajeh et al., 2023)، وتحليل المشاعر (Fang et al., 2024; Fadel et al., 2023)، والإجابة عن الأسئلة سواء كانت مفتوحة المجال (Alsubhi et al., 2022) أو مغلقة وتتطلب استدلالاً معقداً (Abdelazim et al., 2024)، وتوليد النصوص الإبداعية كالقصص (Alabdulrahman et al., 2024; Alhussain & Azmi, 2024)، والترجمة الآلية (Mohsen, 2024; Nagi et al., 2024)، وتلخيص النصوص (Einieh et al., 2022)، والتعرف على الكيانات المسماة (Mahdhaoui et al., 2023; Elchafei & Fashwan, 2024)، بل وصل الأمر إلى محاولة فهم السياقات الثقافية والقيمية الخاصة بالعالم العربي وتطوير نماذج واعية ثقافياً (Huang et al., 2023; Sengupta et al., 2023).

وتتضمن عملية تطوير هذه النماذج أو تكييفها عادةً مراحل متعددة، تبدأ بتحديد المشكلة، ثم اختيار النموذج المناسب (سواء ببنائه من الصفر (Koubaa et al., 2024) أو إعادة تدريب نموذج موجود (Alshammari & Alhumoud, 2022)، وتجهيز البيانات اللازمة (Rasheed et al., 2023) تليها مرحلة التدريب والتقييم.

وعلى الرغم من هذا الاهتمام المتزايد والجهود المبذولة، يبدو أن تطبيق النماذج اللغوية الضخمة على اللغة العربية ما يزال يواجه مساراً تطورياً يتطلب فهماً أعمق وتقييماً شاملاً. وهنا تُطرح عدة تساؤلات حول مدى كفاءة النماذج اللغوية في معالجة العربية، التي غالباً ما تكون بيانات اللغة الإنجليزية مهيمنة في تدريبها الأولي (Sengupta et al., 2023, p.3)، ومدى قدرتها في استيعاب نظام اللغة العربية وخصائصها الفريدة. كما نتقّدح تساؤلات حول

التقدم الفعلي للنماذج المطورة خصيصاً للغة العربية، ومدى قدرتها على المنافسة والأداء في ظل الخصائص اللغوية المميزة والموارد المتاحة، التي قد تكون أقل وفرة وجودة مقارنة بلغات أخرى (Almazrouei et al., 2023, p. 244; Alshahrani et al., 2023, p. 226).

إن هذه التساؤلات تستدعي نظرة فاحصة للوضع الراهن؛ للوقوف على عوامل مثل توفر البيانات العربية عالية الجودة والمتنوعة (Zhu et al., 2024, p. 1; Biswas et al., 2024, pp. 1–2)، وكفاءة المقسمات (tokenizers) المستخدمة للغة العربية (Alyafeai et al., 2024, p. 8)، وما تمثله من جوانب محورية قد تؤثر تأثيراً كبيراً على هذا التطور.

في ضوء ما سبق، تبرز الحاجة الماسة إلى إجراء مراجعات منهجية شاملة ونقدية، تجمع الجهود البحثية المبذولة في ميدان النماذج اللغوية الضخمة ومعالجة اللغة العربية. فالدراسات السابقة، رغم إسهاماتها الفردية القيمة، تناولت جوانب متفرقة من هذا المجال الواسع، وما يزال هناك نقص في رؤية تُقيّم بشكل نقدي ما أُنجِز، وتحدد بوضوح العقبات التي تعيق التقدم. إن أهمية مثل هذه المراجعات لا تكمن فائدتها فقط في الوقوف على الواقع، بل في قدرتها على توجيه الأبحاث المستقبلية، وتجنب تكرار الجهود، والمساهمة في سد الفجوة في ميدان معالجة اللغة العربية.

إن تقديم مراجعة منهجية لما كتب في هذا الميدان سيقدم للقارئ، سواء كان باحثاً، أو مطوراً، أو لغوياً، أو صانع قرار، رؤية شاملة لواقع ومستقبل معالجة اللغة العربية بواسطة النماذج اللغوية الضخمة، ويبرز الأهمية القصوى لمواصلة وتكثيف الجهود البحثية والتطويرية في هذا الاتجاه الإستراتيجي.

لهذا جاءت الدراسة الحالية لتقديم مراجعة منهجية للدراسات السابقة المتعلقة بتطبيق النماذج اللغوية الضخمة في مجال معالجة اللغة العربية، وذلك من خلال تحليل مُعمّق للأدبيات المنشورة في الفترة من عام ٢٠٢٠ إلى نهاية عام ٢٠٢٤، التي شهدت تسارعاً كبيراً في تطوير هذه النماذج وتطبيقاتها.



٢- مشكلة البحث

يشهد مجال معالجة اللغات الطبيعية (NLP) تطوراً هائلاً بفضل ظهور النماذج اللغوية الضخمة (LLMs) وقدراتها الفائقة في فهم وتوليد اللغة، وقد امتد هذا الاهتمام ليشمل اللغة العربية، حيث تتزايد الجهود البحثية لتطوير وتكييف هذه النماذج لمعالجتها وتمكين الآلة منها، وعلى الرغم من هذا النشاط البحثي المتنامي، إلا أن ميدان تطبيق النماذج اللغوية الضخمة على اللغة العربية يواجه تحديات كبيرة وفجوات معرفية أشارت لها الدراسات كثيرًا، موصيةً بمزيد من البحث لسد هذه الفجوات في هذا الميدان (Alyafeai et al., 2024, p. 8; Qarah, 2024, p. 23; Alahmadi et al., 2024, p. 17780).

وبناء عليه، تمثلت المشكلة الرئيسة في النقص الذي أشارت إليه الدراسات من جهة، مع عدم وجود رؤية متكاملة وواضحة حول الوضع الراهن للبحث والتطوير في هذا المجال الحيوي من جهة أخرى. فالدراسات المنشورة، رغم أهميتها الفردية، غالباً ما تكون متفرقة، وتركز على جوانب محددة أو مهام معينة، دون تقديم صورة شاملة تجمع بين الواقع والتحديات والفجوات البحثية. وهنا تأتي هذه الدراسة لتقوم بمراجعة منهجية لما أنجز في ميدان النماذج اللغوية الضخمة ومعالجة اللغة العربية لتحقيق هدف الدراسة.

٣- أسئلة البحث

تمثلت أسئلة البحث في النقاط التالية:

١. ما الجوانب التي تناولتها دراسات مجال النماذج اللغوية الضخمة في معالجة اللغة العربية؟
٢. ما التحديات التي تواجه النماذج اللغوية الضخمة في معالجة اللغة العربية؟
٣. ما الحلول التي تساهم في معالجة التحديات وتحسين معالجة النماذج اللغوية الضخمة للغة العربية؟

٤. الإطار النظري

٤-١- المراجعة المنهجية

تعرف المراجعة المنهجية بأنها دراسة تحليلية نوعية للمعلومات المتعلقة بسؤال بحث محدد، تُنفذ بناءً على اشتراطات ومعايير معدة مسبقاً تهدف إلى اختيار المقالات التي سيتم تضمينها في الدراسة بشكل موضوعي ودقيق (Ortiz et al., 2021, p. 47).

وتُعد أحد أهم المنهجيات البحثية التي تمكن الباحثين من تجميع الدراسات وتحليلها بصورة منظمة وعلمية، وقد اكتسب هذا المنهج مكانة مرموقة، خصوصاً في ميدان العلوم الصحية؛ ليقوم بتجميع نتائج الأبحاث السابقة في موضوع معين تجميعاً يضمن الشفافية والدقة وقابلية التكرار. وعلى الرغم من مميزاتهما، إلا أن استخدامهما ظل محدوداً في الجانب الأكاديمي الطبي، ومؤخراً بدأت تنتشر هذه المنهجية تدريجياً في الأوساط الأكاديمية في المجالات غير الطبية (Snyder, 2019, 334).

والغرض الذي تقوم عليه هذه المنهجية هو الإجابة عن مجموعة من الأسئلة من خلال استعراض ما ناقشته وبجثته الأدبيات والدراسات السابقة في موضوع معين؛ لتصل في نهاية المطاف إلى الحصول على إجابات وصورة أكثر شمولية لم تكن لتتأني إلا من خلال الوقوف على نتائج ومناقشات تلك الدراسات مجتمعة (العرفج وآخرون، ٢٠٢٢، ص. ٤١١).

وما يميز تلك الإجابات اتسامها بالعلمية والدقة مما يتيح استخدامها في دعم السياسات والممارسات للمعنيين؛ وذلك لأنها تقدم إطاراً معرفياً يقوم على تبيين الفجوات، وإبراز القضايا، والوقوف على الواقع للموضوع محل البحث (Okoli, 2015, 880). تركز المراجعة المنهجية بشكل أساسي على تحديد ما يُعرف بالبروتوكول، وهو بمثابة إطار عمل ينفذه القائمون على المراجعة. وتوضّح في هذا البروتوكول الخطوات من بداية العمل إلى نهايته، وقدم المختصون العديد من أشكال هذا البروتوكول، وقاموا بتطويرها وتحسينها لتشمل تفاصيل تبدأ من خطوة صياغة عنوان المراجعة المنهجية وصولاً إلى الإفصاح عن الجهات الممولة للدراسات إن وجدت.



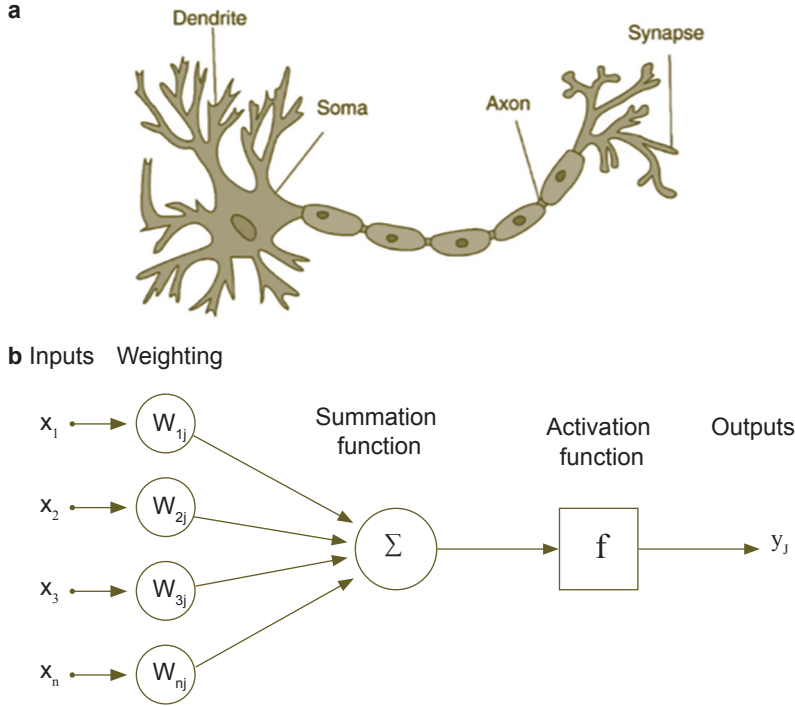
ومن أشهر أطر المراجعات المنهجية إطار "PRISMA" اختصاراً لـ "Preferred Reporting Items for Systematic reviews and Meta-Analyses"، ويعني عناصر التقرير المفضلة للمراجعة المنهجية وبيانات التحليل البعدية (Moher et al., 2010). وفيما يلي، استعراض لأحد هذه الأطر والخطوات المتبعة في المراجعات المنهجية للدراسات السابقة (Ortiz et al., 2021, p. 48):

١. تحديد أسئلة البحث.
٢. تحديد معايير القبول والاستبعاد.
٣. البحث عن الدراسات.
٤. إزالة المقالات المكررة.
٥. تطبيق معايير الفرز.
٦. تقييم جودة الأبحاث.
٧. استخراج البيانات وتحليلها.
٨. نشر نتائج المراجعة المنهجية.
٩. تحديث المراجعة المنهجية.

٤-٢- النماذج اللغوية الضخمة:

عرف Raschka النماذج اللغوية الضخمة LLM بأنها هي شبكات عصبية عميقة مصممة لفهم النصوص البشرية وتوليدها والاستجابة لها، ويتم تدريبها على كميات هائلة من البيانات النصية (Raschka, 2025, p. 2)، وظهرت هذه الشبكات كمحاولة لمحاكاة طريقة عمل العقل البشري في معالجة المعلومات، فهي تتكون من طبقات مترابطة من "العقد" التي تشبه الخلايا العصبية شكل (١):

شكل (١) محاكاة الشبكات العصبية الاصطناعية للشبكات العصبية في العقل البشري



يوضح الشكل (١) التماثل الجوهري بين العصبون البيولوجي البشري (a) والعصبون الاصطناعي (b)، حيث يُصمم الأخير ليحاكي وظائف الأول. فكما أن الزوائد الشجرية (Dendrites) في العصبون البيولوجي تستقبل الإشارات الكهروكيميائية من عصبونات أخرى، يستقبل العصبون الاصطناعي مدخلات (Inputs) متعددة. تُعطى هذه المدخلات الاصطناعية أوزاناً (Weights) تحاكي قوة المشابك العصبية (Synapses) البيولوجية، حيث تحدد هذه الأوزان مدى تأثير كل مدخل. بعد ذلك، يتم تجميع هذه المدخلات الموزونة بواسطة دالة الجمع (Summation function) في العصبون الاصطناعي، بشكل يماثل عملية تجميع الإشارات في جسم الخلية (Soma) البيولوجي. ثم يمر هذا المجموع عبر دالة التنشيط (Activation function)، التي تقرر ما إذا كان العصبون سينشط وبأي قوة، محاكاةً بذلك آلية إطلاق النبضة العصبية في جسم الخلية عند تجاوز عتبة معينة. وأخيراً، يُنتج



العصبون الاصطناعي مخرجات (Outputs) تمثل الإشارة النهائية، وهي تناظر النبضة العصبية المنقولة عبر المحور العصبي (Axon) في العصبون البيولوجي، لتصبح بدورها مدخلاً لعصبونات أخرى في الشبكة (3). (Zhang et al., 2019, p. 3).

ومع تطور هذه الشبكات، وُلدت فكرة الشبكات العصبية العميقة (Deep Neural Networks) القادرة على تعلم أنماط معقدة من البيانات كتعقيدات اللغة البشرية، فظهرت النماذج اللغوية الضخمة التي هي عبارة عن شبكات عصبونية عميقة مصممة لفهم النصوص البشرية وتوليدها والاستجابة لها. (Raschka, 2025, p. 7) لكن هذه الشبكات واجهت تحديات في فهم السياقات الطويلة والعلاقات البعيدة بين الكلمات، إلى أن نُشرت ورقة في عام ٢٠١٧ جاءت لتحل هذا الإشكال، وعرضت بنية جديدة للشبكات العصبية تُدعى بالمحولات (Transformers) (Vaswani et al., 2017)، قُدمت فيها عملية الانتباه الذاتي (Self-Attention)، مما مكن النماذج من تحليل العلاقات بين جميع الكلمات نحويًا ودلاليًا، حتى وإن تباعدت المسافة في جملة: "اشترى محمد كتابًا من المكتبة وعندما وصل إلى المنزل وضعه على المكتب" فبإمكان النماذج اللغوية المبنية على بنية المحولات أن تتوصل بأن الضمير في "وضعه" عائد على الكتاب مع تباعد المسافة بين الكتاب والضمير العائد. وبفضل هذا التقدم في البنية الهيكلية للشبكات العصبية، أصبح من الممكن تدريب نماذج اللغة الكبيرة على كميات هائلة من البيانات النصية والتقاط العلاقات التي تربط النظام اللغوي البشري؛ لتصبح قادرة على فهم اللغة البشرية وتوليدها، ونتيجة لذلك أصبحت النماذج اللغوية الضخمة قادرة على الأداء في مجموعة واسعة من مهام معالجة اللغات الطبيعية، بما في ذلك ترجمة النصوص، وتحليل المشاعر، والإجابة عن الأسئلة، وغيرها الكثير.

فبُنيت نماذج لغوية ضخمة تعتمد على بنية المحولات التي جعلت من النماذج اللغوية أكثر دقة وكفاءة في فهم السياق اللغوي، وتوليد ردود تشبه ردود البشر بدرجة مذهلة، ومن تلك النماذج التي ظهرت على الساحة النموذج التوليدي GPT الذي بني عليه تطبيق المحادثة ChatGPT عام ٢٠٢٢، وتتابع فيما بعد ظهور عدد كبير من النماذج على الساحة.

٤-٣- معالجة اللغات الطبيعية NLP

هي إحدى حقول علوم الحاسوب الذي يسعى إلى وضع المفاهيم، وإيجاد الأساليب، وبناء البرمجيات القادرة على فهم اللغات البشرية الطبيعية ودراساتها وإنتاجها؛ بهدف تمكين التفاعل البشري مع الحواسيب كتابة ونطقاً (Marie-Sainte et al., 2018, p. 7011).

اعتمدت منهجيات معالجة اللغات الطبيعية في البداية على الأساليب القائمة على القواعد (rule-based approaches)، حيث استُخدمت خوارزميات بسيطة وأُعطيت تعليمات حول العبارات والكلمات التي يجب البحث عنها في عينة من النص، مع تحديد استجابات معينة يتم إرجاعها عند العثور على هذه العبارات، ومن تلك التطبيقات ما يُعنى بالبحث عن الكلمات واسترجاعها، التي تعد شكلاً من أشكال معالجة اللغات الطبيعية، ولكن بعد ظهور التعلم العميق حدثت ثورة في حقل معالجة اللغات الطبيعية لتظهر لدينا النماذج اللغوية الضخمة التي بنيت على نهج أكثر قدرة على فهم التعقيد اللغوي وأنماطه من خلال تدريب تلك النماذج على كمية ضخمة من البيانات ليبدأ بالتعرف على العلاقات التي تربط النظام اللغوي ليصل من خلال التدريب على البيانات إلى القدرة على ذلك محاكياً في الواقع الطريقة التي يتعلم بها الطفل اللغة (Burns, 2019, p. 11).

٥. إجراءات البحث

في هذا القسم تُعرض الخطوات التي أُجريت من خلالها المراجعة المنهجية لأبحاث النماذج اللغوية الضخمة ومعالجة اللغة العربية:

٥-١- أسئلة المراجعة المنهجية

تمثلت أسئلة المراجعة المنهجية فيما يلي:

١. ما الجوانب التي تناولتها دراسات مجال النماذج اللغوية الضخمة ومعالجة اللغة العربية؟
٢. ما التحديات التي تواجه النماذج اللغوية الضخمة في معالجة اللغة العربية؟
٣. ما الحلول التي تطرحها دراسات النماذج اللغوية الضخمة في معالجة تلك التحديات؟



٥-٢- معايير قبول واستبعاد الدراسات

وضعت معايير الإقصاء والقبول بما يضمن توافقها مع غرض البحث، والدارسات التي تنطبق عليها هذه المعايير تُصنّف في قائمة الدراسات المقبولة في المراجعة المنهجية، ويوضح الجدول (٢) معايير قبول وإقصاء الدراسات التي أُتبعَت في كل نقطة. فمن حيث النطاق الموضوعي حددت الأبحاث التي تركز حصراً على اللغة العربية الفصحى سواء المستوى الكلاسيكي أم المعاصر؛ وذلك لتوحيد أساس المقارنة وتجنب التباين الكبير بين اللهجات، مما يضمن تحليلاً معمقاً للغة المعيارية المستخدمة في المصادر الأكاديمية والرسمية. ولضمان الحدّثة اقتصر الإطار الزمني على الفترة بين عامي ٢٠٢٠ و٢٠٢٤؛ للوقوف على التطورات الأخيرة في النماذج اللغوية، مع استبعاد الدراسات النظرية؛ لضمان بناء التحليل على أدلة تجريبية ونتائج قابلة للقياس. وأخيراً، حُصر البحث في الدراسات المنشورة باللغة الإنجليزية والمفهرسة ضمن قواعد البيانات المحددة؛ لأنّ جل الأبحاث في هذا الميدان منشورة باللغة الإنجليزية، مما يوفر صورة شاملة وممثّلة للمجال يمكن الاعتماد على نتائجها في المراجعة المنهجية.

جدول (٢) معايير قبول وإقصاء الدراسات

معيّار الإقصاء	معيّار القبول
تُسْتبعد الدراسات التي لا تعالج اللغة العربية أو جعلت اللغة العربية ضمن مجموعة من اللغات وليست هدفاً بذاتها في الدراسة	تقبل الدراسات التي جعلت اللغة العربية هدفها في المعالجة
تُسْتبعد الدراسات التي ركزت على اللهجات العربية في الوطن العربي	تقبل الدراسات التي جعلت اللغة العربية الفصحى بالمستويين الكلاسيكي والمعاصر هدفها في المعالجة
تُسْتبعد الدراسات التي نشرت قبل سنة ٢٠٢٠ أو صدرت بعد سنة ٢٠٢٤	تضمين الدراسات المنشورة خلال السنوات الخمسة الأخيرة لضمان حداثة المعلومات. من سنة ٢٠٢٠ وحتى نهاية ٢٠٢٤
تستبعد الدراسات المكتوبة بأي لغة غير الإنجليزية	تقبل الدراسات المكتوبة باللغة الإنجليزية

معيّار الإقصاء	معيّار القبول
تُستبعد الدراسات التي ناقشت الموضوع على نحو نظري	تقبل الدراسات التي ركزت على المعالجة اللغوية على نحو عملي
تستبعد الدراسات التي لم ترد في الأوعية والمحركات التالية: PubMed و IEEE و Google Scholer	تقبل الدراسات الموجودة في الأوعية والمحركات الآتية: PubMed و IEEE و Google Scholer

٥-٣- البحث ومصادر البيانات

أجري البحث الإلكتروني في عدد من قواعد البيانات الأكاديمية، وحُدّدت مستودعات الأبحاث التالية: Google Scholar، وIEEE، وPubMed. ولغرض أهداف البحث وطبيعة الأسئلة المطروحة اختيرت مجموعة من الكلمات المفتاحية، وهي: Large Language Models و Natural Language Processing و Arabic Natural Language Processing و Computational Linguistics Arabic و Artificial Intelligence in Language و AI. يتضح من الجدول (٣) أن عملية البحث في قواعد البيانات المُعتمَدة عن أبحاث النماذج اللغوية الضخمة ومعالجة اللغة العربية قد أسفرت عن استرجاع ٣١٠ بحث بالمكررات. وقد استحوذ محرك البحث Google Scholar على النصيب الأكبر من النتائج بواقع ١٥٠ بحثاً، مما يعكس شمولية تغطيته الأكاديمية لأنه يعد محرك بحث وليس قاعدة بيانات. وجاءت قاعدة بيانات IEEE في المرتبة الثانية بفارق ليس بأكبر، حيث قدمت ١٢٥ بحثاً، وهو ما يشير إلى الطبيعة التقنية أو الهندسية لموضوع البحث. في المقابل، كان إسهام قاعدة بيانات PubMed محدوداً بـ ٣٥ بحثاً فقط، وهو أمر متوقع؛ نظراً لأن تخصصها في العلوم الطبية، ورغم أن هذا التخصص قد لا يتقاطع مباشرة مع المحاور التقني واللغوي للدراسة، إلا أن النتائج تسلط الضوء على مجال تطبيقي واعد وحيوي. إذ تُظهر هذه الأبحاث تزايد استثمار تقنيات معالجة اللغات الطبيعية في المجالات الطبية، مثل تقييم النماذج اللغوية ومدى معرفتها في المواضيع الطبية، أو مساعدة المرضى والرد على استفساراتهم. وعليه فإن هذه الأبحاث رغم قلتها إلا أنها تمثل مؤشراً مهماً على استثمار النماذج اللغوية في قطاعات متخصصة، وهو بُعد تطبيقي يهدف البحث الحالي إلى تغطيته.



جدول (٣) نتائج البحث في قواعد البيانات

عدد الأبحاث	قاعدة البيانات
١٢٥	IEEE
١٥٠	Google Scholar
٣٥	PubMed
٣١٠	المجموع

٥-٤- فحص النتائج

بعد مراجعة النتائج التي ظهرت خلال عملية البحث، طُبقت خطوتان للاستبعاد:

أولاً: استبعاد الدراسات المكررة.

ثانياً: تطبيق معايير الإقصاء والقبول للدراسات المتبقية.

ولتطبيق معايير الإقصاء والقبول، فقد فحصت المعلومات الأساسية في كل دراسة كعنوان الدراسة وأهدافها ومنهجها ونتائجها، وأيضاً فحص مضمونها الداخلي؛ للتأكد من جودتها وانطباق الشروط عليها، وبناءً على تلك المعايير، فقد أصبح عدد الدراسات المشمولة في هذا البحث (٨٢) دراسة في جميع الأوعية، وهي موضحة في الجدول (٤):

جدول (٤) عدد الأبحاث التي انطبقت عليها المعايير في كل وعاء

عدد الأبحاث	قاعدة البيانات
٢٤	IEEE
٤٨	Google Scholar
١٠	PubMed
٨٢	المجموع

يُظهر جدول (٤) توزيع الدراسات التي انطبقت عليها المعايير في قواعد البيانات المختارة. وحصلت قاعدة بيانات Google Scholar على أعلى عدد دراسات وعددها (٤٨) دراسة ما نسبته ٥٨,٥٪، وهو أمر متوقع ومنطقي؛ لأن Google Scholar يُعد محرك بحث أكاديمي شامل يسمح بفهرسة مصادر متنوعة لا تقتصر على المجلات والمؤتمرات المحكّمة فقط، بل تشمل أيضاً الأوراق البحثية قيد المراجعة (pre-prints) مثل تلك المنشورة على منصة arXiv، بالإضافة إلى رسائل الماجستير والدكتوراه.

وتأتي قاعدة IEEE Xplore المتخصصة بالتقنية وعلوم الحاسوب في المرتبة الثانية حاصلة على (٢٤) دراسة ما نسبته ٢٩,٣٪، وهو ما يتماشى مع طبيعة المجال الذي يُعد أحد شقّيه تقنياً، فتتركز غالبية الجهود البحثية على تطوير البنى الهندسية والخوارزميات لهذه النماذج وإنتاجها واختبارها.

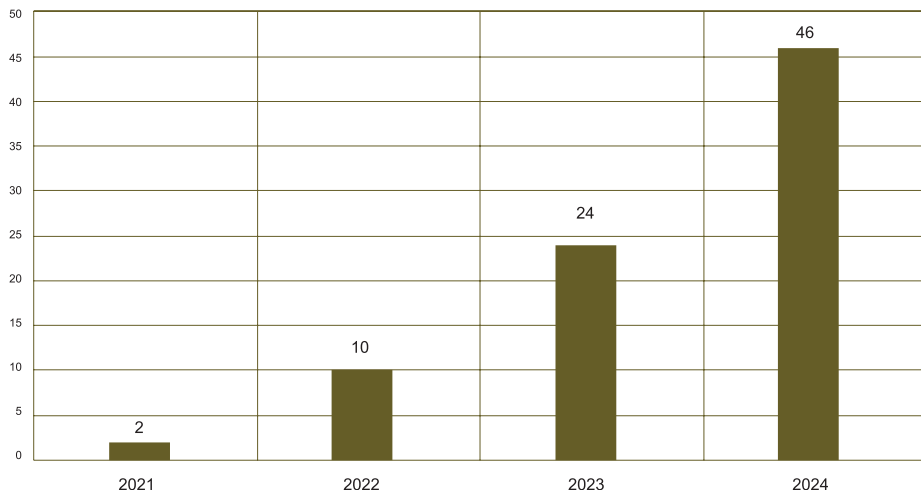
وحصلت قاعدة PubMed على أقل عدد من الدراسات المشمولة وعددها (١٠) دراسات ما نسبته ١٢,٢٪، وهو ما يشير إلى أن تطبيق النماذج اللغوية الضخمة في المجال الطبي العربي لا يزال في مراحله الأولى مقارنةً بالأبحاث التقنية الأساسية، لكنه يمثل اتجاهاً بحثياً واعدًا ومجالاً تطبيقياً ذا أهمية متزايدة.

ويعكس هذا التوزيع بشكل عام أن أبحاث النماذج اللغوية العربية تتركز حالياً في الأوساط الأكاديمية التقنية، مع بدء ظهور اهتمام بتطبيقاتها في قطاعات متخصصة مثل الطب.

وجاء توزيع الدراسات المقبولة من حيث السنة على النحو التالي كما يظهره الشكل (٢):



شكل (٢) توزيع الدراسات بناء على السنة



يتضح من الشكل (٢) النمو المتزايد في الإنتاج البحثي في مجال معالجة اللغة العربية والنماذج اللغوية الضخمة بدءاً من دراستين في عام ٢٠٢١ من الدراسات التي انطبقت عليها المعايير، ثم قفز العدد خمسة أضعاف محققاً نسبة ٤٣,٢٪ من إجمالي الدراسات (٨٢) دراسة؛ لتصل في عام ٢٠٢٤ إلى (٤٦) دراسة محققة نسبة ٥٦,٠٩٪ من إجمالي عدد الدراسات التي انطبقت عليها المعايير، أي ما يزيد عن نصف العدد.

وهو يحاكي ما كان يجري في أرض الواقع للذكاء الاصطناعي من زخم. ففي سنة ٢٠٢٠ كان الإعلان عن نموذج GPT-3 الذي يُعد ثورة في حينه في قدرته على توليد اللغة البشرية بدقة عالية محاكية قدرة البشر على ذلك، فنمت في ضوء هذا الأبحاث التي تعنى بهذا المجال، إلا أن الاهتمام باللغة العربية جاء متأخراً؛ كونها أحد اللغات قليلة الموارد، إلا أن النمو واضح ومطرّد كما يظهره الشكل (٢)، ويلاحظ أنه في عام ٢٠٢٠ ظهرت قفزة في عدد الأبحاث، وتزامن ذلك مع إطلاق الشات بوت لتلك النماذج الضخمة مثل ChatGPT، وما تلاها مثل GEMINI وغيرها، ومع ازدياد الزخم ازدادت الأبحاث؛ لما شوهده لتلك النماذج من تأثير واضح على مختلف الأصعدة، وهنا جاءت الأبحاث التي كانت هي في بادئ الأمر وراء بناء

تلك النماذج الضخمة، لتواصل الأعمال في عدة مجالات بدءاً من بناء تلك النماذج، واقتراح البنية الهندسية المناسبة لمعالجة اللغات، ومروراً بالبيانات وهيكلتها ومعالجتها لتدريب تلك النماذج، وانتهاءً بتقييمها وإنشاء الاختبارات التي تقيس دقة تلك النماذج؛ للوقوف على نقاط القوة والضعف.

٥-٥- مناقشة الأسئلة والنتائج

٥-٥-١- للإجابة عن السؤال الأول وهو: "ما الجوانب التي تناولتها دراسات مجال النماذج اللغوية الضخمة في معالجة اللغة العربية؟"

فقد حُللت الأبحاث المشمولة في الدراسة؛ للتوصل إلى الجوانب التي ناقشتها في هذا الميدان وجاءت على النحو التالي كما يوضحه جدول (٥):

جدول (٥) الجوانب التي تناولتها دراسات النماذج اللغوية الضخمة ومعالجة اللغة العربية

عدد الدراسات	الجوانب العامة	الجوانب الفرعية
٢٧	التحليل والفهم (Analysis & Understanding)	- كشف المحتوى الضار والمضلل (أخبار زائفة، سخرية، تنمر، إلخ) - تحليل المشاعر والآراء - تصنيف النصوص وتحديد سمات المؤلف - التعرف على الكيانات المسماة وتحديد تشابه النصوص
١٤	التوليد والإنتاج (Generation & Creation)	- الإجابة عن الأسئلة - تلخيص النصوص وإعادة الصياغة - توليد المحتوى الإبداعي (قصص، شعر، أكواد برمجية) - روبوتات الحوار والتفاوض
١٨	بناء النماذج (Model Building)	بناء نماذج لغوية عربية أساسية من الصفر - تكيف وتحسين النماذج العالمية للغة العربية - تحسين البنية الهيكلية للنماذج



عدد الدراسات	الجوانب العامة	الجوانب الفرعية
١٩	التقييم والقياس (Evaluation & Measurement)	- تقييم أداء وقدرات النماذج بشكل شامل - قياس وتحليل خصائص النماذج (التحيز، الأمان، الأخطاء) - دراسات مقارنة بين النماذج المختلفة
٤	بناء المصادر والبيانات (Resource & Data Building)	إنشاء مجموعات بيانات - تطوير أطر عمل للتقييم - بناء المدونات اللغوية
٨٢	المجموع	

يعرض الجدول (٥) الجوانب التي تناولتها دراسات النماذج اللغوية الضخمة ومعالجة اللغة العربية، وجاءت أهداف تلك الدراسات لتغطي خمسة جوانب على النحو التالي:

٥-١-١-٥-٥ الدراسات التي جاءت لغرض التحليل والفهم للنصوص العربية

يمثل هذا الجانب الجزء الأكبر الذي تركز عليه الدراسات فقد بلغ عدد الدراسات (٢٧) دراسة، وتهدف إلى تمكين الآلة من قراءة وفهم النصوص العربية واستخلاص معلومات محددة منها بحيث يُغذى النموذج اللغوي بالمُدخل المراد تحليله ومن ثم يصدر حكماً بناءً على غرض هذا النموذج، كأن يطلب منه أن يعرف مشاعر الشخص صاحب النص هل هو حزين أم مسرور؟

وقد شمل هذا الجانب محاور فرعية متعددة، أبرزها كشف المحتوى الضار والمضلل الذي أصبح ضرورة ملحة مع انتشار وسائل التواصل الاجتماعي لحماية المستخدم العربي، فعلى سبيل المثال، تناولت دراسة Ali et al (٢٠٢٢) مهمة كشف الأخبار الزائفة، بينما استهدفت دراسة Obeidat et al (٢٠٢٢) تحدياً أكثر دقة وهو كشف السخرية. كما برز تحليل المشاعر والآراء كأحد أكثر التطبيقات التي ناقشتها الدراسات، حيث سعت الأبحاث لتجاوز التصنيف البسيط إلى فهم أعمق، كما يظهر في دراسة Lotfy et al (٢٠٢٤) التي حللت آراء المستهلكين من خلال ما يكتبونه من مراجعات، قد تحوي الجملة فيها إلى رأي يثني على جزء

في المنتج ويزم جزءاً آخر، ومن الدراسات التي عُنيت بالتحليل كذلك دراسة Mahdhaoui et al. (٢٠٢٣) التي ركزت على تحسين التعرف على الكيانات المسماة Named Entity Recognition (NER)، وهي مهمة حيوية لأنظمة استخلاص المعلومات ومحركات البحث وغيرها؛ لما تقدمه من تحديد ما إن كانت الكلمة المدخلة علماً على بشر أم مؤسسة أم مكان، ويمكن مقابلة مفهوم التعرف على الكيانات بمفهوم المعارف في اللغة العربية.

٥-١-٢- الدراسات التي جاءت لغرض توليد النصوص العربية

بلغ عدد الدراسات في هذا الجانب (١٤) دراسة، هدفت إلى الانتقال بالآلة من مرحلة الفهم إلى مرحلة الإنتاج، بحيث يُغذى النموذج اللغوي بمُدخل معين كسؤال أو نص طويل، ويُطلب منه أن يُنشئ نصاً عربياً جديداً ومتناسكاً، وشمل هذا الجانب محاور فرعية متعددة من أبرزها الإجابة عن الأسئلة، كما في دراسة Adel et al. (٢٠٢٣) التي قدمت نموذج (AraQA) المتخصص في الإجابة عن الأسئلة المتعلقة بالنصوص الدينية. ومنها التلخيص وإعادة الصياغة كما جاء في دراسة Einieh et al. (٢٠٢٢)، حيث هدفت إلى الضبط الدقيق لنموذج AraT5 لإنتاج ملخصات عربية تجريدية في جمل جديدة لا تكون موجودة في النص الرئيس. ومن تلك المهام مهام التوليد الإبداعي، ومن الأمثلة الفريدة على ذلك دراسة Al-khafaji & Majeed (٢٠٢٤) التي اختبرت قدرة النماذج اللغوية على فهم متطلبات برمجية باللغة العربية وتوليد أكواد بايثون برمجية صحيحة.

٥-١-٣- الدراسات التي جاءت لغرض بناء النماذج اللغوية العربية

يمثل هذا المحور الدراسات التي ركزت على الجانب التقني المتعلق ببناء النماذج، وقد بلغ عدد الدراسات فيه (١٨) دراسة. وهدفت إلى هندسة وبناء النماذج اللغوية نفسها، حيث يُركّز على إنشاء وتطوير البنية الأساسية للذكاء الاصطناعي للغة العربية بدلاً من مجرد استخدامها، ويظهر في هذا الجانب اتجاهان رئيسان:

الاتجاه الأول: هو بناء النماذج من الصفر، وهو مسار يتطلب استثمارات ضخمة ويعكس سعي المنطقة لامتلاك نماذجها الخاصة المتوافقة ثقافياً، والمثال الأبرز على ذلك



هو نموذج علام (Allam) الذي أطلقته (سدايا) (Bari et al., 2024) وجيس (Jais) الذي قدمته دراسة (Sengupta et al. (٢٠٢٣ كأول نموذج لغوي ضخم مفتوح المصدر يركز على اللغة العربية، والمقصود هنا ببناء النموذج من الصفر هو بناؤه من حيث البيانات التي يُدرَّب عليها، بحيث يكون النموذج مُفرَّغاً في البداية، وقد يتضمن بعض التحسينات التي تناسب اللغة العربية، مثل دمج مقسم (Tokenizer) ملائم للعربية، ثم تدريبه على بيانات ضخمة باللغة العربية.

أما الاتجاه الثاني: وهو الأكثر شيوعاً ألا وهو التكييف والتحسين (Fine-tuning)، حيث يُؤخذ نموذج عالمي قوي ويُدرَّب تدريباً إضافياً على بيانات عربية؛ لتحسين أدائه، وتعد دراسة (Huang et al. (٢٠٢٣ التي قدمت نموذج "AceGPT" مثلاً واضحاً على توطين (Localizing) النماذج العالمية لتلبية احتياجات اللغة العربية بشكل أفضل.

٥-١-٥-٤- الدراسات التي جاءت لغرض التقييم والقياس للنماذج اللغوية العربية

مع تزايد عدد النماذج وتطور قدراتها، برز هذا الجانب كونه ضرورة علمية ملحة، وقد بلغ عدد الدراسات فيه (١٩) دراسة. وتهدف إلى ضمان أن التطور الحاصل في المجال حقيقي، ويمكن قياسه بموضوعية، بالإضافة إلى تسليط الضوء على نقاط الضعف والقوة في النماذج المبنية، ويظهر في هذا الجانب محوران رئيسان:

المحور الأول: هو بناء المعايير (Benchmarking) التي تهدف إلى إنشاء مقاييس موحدة يمكنها تقييم أداء النماذج المختلفة في مهام محددة يستهدفها المعيار، تكون مرجعاً لقياس مدى تمكن النموذج من تلك المهام، ومثال ذلك دراسة Abdelali et al. (٢٠٢٣) التي قدمت إطار عمل تقييمي شامل يُدعى (LAraBench) لاختبار أداء النماذج على مجموعة واسعة من المهام باللغة العربية، بلغت (٣٣) مهمة، ومن تلك المهام مهمة تحديد أقسام الكلام والتشكيل والتعرف على الكيانات وتحليل المشاعر وغيرها من المهام.

أما المحور الثاني: عملية التقييم، سواء التقييم الذي يقوم على المعايير أو ما سواها، كالاكتكام إلى نموذج ليقيم أداء نموذج آخر في مهمة محددة، أو من خلال الاعتماد على البشر المختصين في تقييم أداء نموذج في مهمات، كالإرشاد الطبي وغيره، وتعد دراسة (٢٠٢٤)

Alrajhi et al مثلاً بارزاً على هذا التوجه، حيث عالجت بشكل مباشر ومهم قضية قياس التحيز الجنسي في النماذج اللغوية العربية.

٥-١-٥-٥ الدراسات التي جاءت لغرض بناء المصادر والبيانات العربية

على الرغم من قلة عدد الدراسات في هذا الجانب (٤) دراسات، إلا أنه يُعد الأكثر أهمية من الناحية الإستراتيجية؛ لأن البيانات هي وقود الذكاء الاصطناعي، وبدونها لا يمكن تدريب النماذج أو تقييمها بفعالية. وتهدف هذه الدراسات إلى إنشاء بيانات عالية الجودة ومتاحة للجميع لدفع عجلة البحث في اللغة العربية قدماً. وتعد دراسة Ali et al (٢٠٢٣) مثلاً واضحاً على هذا التوجه، حيث ساهمت مساهمة مباشرة في هذا الجانب عبر بناء وإتاحة مدونة عربية متنوعة تُستخدم لتدريب نماذج اللغة الأساسية، وقد تجاوز حجمها ٥٠٠ جيجابايت من النصوص المجهزة، جُمعت من ٢٠ مصدرًا مختلفًا لضمان تنوع المجالات (أخبار، وكتب، ونصوص دينية، وشعر)، بالإضافة إلى المحتوى الذي تضمّن العربية الفصحى، كان هناك جزء من المدونة حوى لهجات عربية. كما قدمت دراسة Abdallah et al (٢٠٢٤) مجموعة بيانات (ArabicQA) مصممة خصيصاً لدعم مهمة الإجابة عن الأسئلة باللغة العربية، مما يوفر مورداً قيماً للباحثين لتدريب واختبار نماذجهم على هذه المهمة. ويمكن تقسيم الدراسات بحسب المستوى اللغوي إلى خمسة مستويات كما يظهره الجدول (٦):

جدول (٦) توزيع الدراسات بحسب المستوى اللغوي

عدد الدراسات	الهدف المحدد من الدراسة	طبيعة المهمة	المستوى اللغوي
١	تصحيح الأخطاء الإملائية والكتابية.	توليد	المستوى الصرفي-الإملائي
١	التشكيل (إضافة الحركات الصحيحة للنص)	تحليل	
١	تصحيح الأخطاء النحوية، وعلامات الترقيم، وتكوين الجمل	توليد	المستوى النحوي
٣	التعرف على الكيانات المسماة (NER) وتحديد ها	تحليل	



عدد الدراسات	الهدف المحدد من الدراسة	طبيعة المهمة	المستوى اللغوي
٥	الإجابة عن الأسئلة في مجالات متعددة	توليد	المستوى الدلالي
١	إعادة الصياغة		
٢	التلخيص		
٢	الترجمة الآلية بين العربية والإنجليزية		
٦	تحليل مشاعر وآراء المستخدمين	تحليل	
٢	تصنيف النصوص حسب الموضوع (سياسي، وتاريخي، إلخ)		
١	تصنيف النصوص متعدد التسميات (رومانسي، وكوميدي، ودرامي)		
١	استخلاص العلاقات بين الكيانات (مثل: مؤسس، وشريك)		
١	التحقق من صحة المنطق في الجمل		
١	الاستدلال باللغة الطبيعية وتحديد التناقض		
١	قياس التشابه الدلالي بين النصوص أو الأسئلة		
١	بناء روبوتات للتفاوض التجاري		
٤	اكتشاف الأخبار الكاذبة والمزيفة	تحليل	
٢	اكتشاف الشائعات والادعاءات غير الموثوقة		
٢	اكتشاف التقنيات الدعائية في النصوص		
٢	اكتشاف السخرية		
١	اكتشاف النصوص المهينة، أو التي تحتوي على لغة مسيئة		
١	تصنيف النص إلى ذاتي أو موضوعي		

المستوى اللغوي	طبيعة المهمة	الهدف المحدد من الدراسة	عدد الدراسات
المستوى الأدبي والأسلوبي	توليد	توليد القصص الإبداعية (مثل قصص الأطفال)	١
	تحليل	تحليل القصائد العربية وتحديد خصائصها (البحر، والقافية... إلخ)	١
		تحديد جنس المؤلف من خلال أسلوب الكتابة	١

يعرض الجدول (٦) توزيعاً للدراسات البحثية بناءً على المستوى اللغوي الذي تعالجه كل دراسة، سواء كان في جانب التوليد اللغوي أو التحليل اللغوي. يتضح من الجدول أن التركيز البحثي الأكبر انصب على المستويات اللغوية العليا، وتحديدًا المستوى الدلالي والمستوى التداولي والخطابي ففي المستوى الدلالي، ركزت مهام التوليد بشكل كبير على الإجابة عن الأسئلة في سياقات متخصصة، مثل المجال الديني (Adel et al., 2023) والمجال الطبي (Sallam et al., 2024)، بالإضافة إلى مهام الترجمة الآلية (Mohsen, 2024) والتلخيص التجريدي (Einieh et al., 2022) وفي مهام التحليل، كان النصيب الأكبر لتحليل المشاعر والآراء حيث بلغت (٦) دراسات، سواءً في التغريدات (Al Wazrah & Alhumoud, 2021) أو في مراجعات المنتجات (Fadel et al., 2023) أما في المستوى التداولي، فقد هيمنت مهام التحليل الهادفة إلى حماية المستخدم وكشف التضليل، مثل اكتشاف الأخبار الكاذبة (Ali et al., 2022) والتقنيات الدعائية (Hasanain et al., 2024)، إلى جانب كشف السخرية (Obeidat et al., 2022) هذا التوجه يعكس الطلب الكبير على تطبيقات عملية قادرة على فهم المعنى المباشر والنوايا الكامنة وراء النصوص العربية.

على النقيض من ذلك، يُظهر الجدول (٦) نقصاً واضحاً في الدراسات التي تستهدف المستويات اللغوية التي تشكل الهيكل الأساسي للغة. فحظي المستوى الصرفي-الإملائي بتمثيل ضئيل للغاية، حيث اقتصر على دراسة في التحليل لمعالجة تحديات التشكيل (Alshammari & Elleithy, 2024) وأخرى في التوليد لتصحيح الأخطاء النصية (Abdelaal et al., 2024) وكان المستوى النحوي أقل تمثيلاً، حيث اقتصر على دراسة واحدة في جانب التوليد لمعالجة



الأخطاء النحوية (Kwon et al., 2023)، بينما غابت الدراسات التي تتعمق تعمقاً كاملاً في التحليل النحوي المعقد. وهذا يؤكد أن الأبحاث التي تعالج بنية الكلمة والجملة معالجة مباشرة ما تزال قليلة.

أما المستوى الأدبي والأسلوبي فقد جاءت ثلاث دراسات لتعالج بعض مهامه، مثل التحليل الحاسوبي للقوائد العربية (Qarah, 2024)، وتوليد القصص الإبداعية (Alhussain & Azmi, 2024)، وتحديد سمات المؤلف من خلال أسلوبه الكتابي (Bassem et al., 2024).

ويلاحظ غياب المستوى الصوتي في هذه الدراسة، لأنه يعد خارج نطاقها الذي يهتم بالنماذج اللغوية التي تعالج النصوص المكتوبة. ويمكن تقسيم الدراسات بحسب المجالات غير اللغوية إلى تسعة مجالات كما يظهره الجدول (٧):

جدول (٧) توزيع الدراسات بحسب المجالات غير اللغوية

عدد الدراسات	مجال الدراسة
١١	- بناء نماذج لغوية
٤	- بناء نماذج متخصصة لمهام محددة
١	- توليد المعلومات المعزز بالاسترجاع (RAG)
١٣	- بناء أطر تقييم ومعايير (Benchmarks)
٢	- بناء متون لغوية ضخمة (Corpora)
٣	- اكتشاف ما إذا كان النص مكتوباً بالذكاء الاصطناعي
١	- قياس التحيز الجنسي في مخرجات النماذج
١	- خداع النماذج (Adversarial Attacks)
١	- استخدام النماذج للمساعدة في توصيف البيانات (Annotation)

يعرض الجدول (٧) توزيع الدراسات التي لم تستهدف مستوى لغوياً محدداً في معالجتها اللغة العربية، بل انصب تركيزها على النموذج نفسه وأدائه وما يتعلق ببنيتها

التحتية، والبيانات وإستراتيجيات التعامل معها، وتقع هذه الدراسات البالغ عددها (٣٦) دراسة، ضمن فئة الدراسات غير اللغوية التي لا تركز في هدفها على تحليل معنى النص بشكل مباشر، بل تهدف إلى بناء الأدوات الأساسية (البنية التحتية)، وتقييمها، وتحليل خصائصها الداخلية، ويمكن تقسيم هذا الجانب الحيوي إلى عدة مجالات رئيسية:

أولاً: بناء البنية التحتية الأساسية للنماذج

يمثل هذا الجانب الجهود الرامية إلى التركيز على بناء نماذج لغوية وإطلاقها، وقد بلغ عددها (١١) دراسة قامت على توفير نماذج عربية قوية ومنافسة، مثل نموذجي جيس (Sengupta et al., 2023) وعلام (Bari et al., 2024)، أو تكييف النماذج العالمية لتناسب السياق العربي كما في دراسة (Huang et al., 2023).

ثانياً: بناء مصادر وقواعد البيانات

تضمن هذا الجانب بناء المصادر والبيانات، وجاءت دراستان في هذا الموضوع، واعتنت بمصادر البيانات سواء كانت تلك البيانات مدونات لغوية ضخمة يمكن استخدامها في التدريب (Ali et al., 2023) أو مجموعات لبيانات موسمة لاستخدامات محددة.

ثالثاً: بناء نماذج وتقنيات متخصصة

إلى جانب النماذج العامة، ظهرت جهود لبناء نماذج متخصصة أو التركيز على إستراتيجيات متخصصة للتعامل مع البيانات، وقد بلغ عدد الدراسات في هذا الجانب (٤) دراسات هدفت إلى التركيز على بناء نموذج لحل مشكلة محددة بكفاءة عالية، مثل نموذج "Dhati" لتحديد ذاتية النصوص (Nehar et al., 2023)، كما تندرج تحت هذا المجال دراسات تستكشف إستراتيجيات أخرى مثل التوليد المعزز بالاسترجاع (Khaled et al., 2024) (RAG)، وهي تقنية تجمع بين قدرتي الاسترجاع والتوليد لحل المشكلات المعقدة.

رابعاً: تحليل وتقييم النماذج المتخصصة

مع تزايد عدد النماذج، أصبح من الضروري تقييمها علمياً، ويشمل هذا المحور بناء أطر تقييم ومعايير موحدة Benchmarks وقد بلغت الدراسات في هذا الجانب (١٣) دراسة، ومن



تلك الدراسات دراسة (Abdelali et al., 2023; Khondaker et al., 2023)، التي توفر اختبارات شاملة لمقارنة أداء النماذج بدقة. كما تتضمن عملية التقييم الحكم على مخرجات النماذج، ومن ذلك اكتشاف ما إذا كان النص مولدًا بالذكاء الاصطناعي أو هونص بشري، وبلغت الدراسات في هذا الجانب (٣) دراسات كما في دراسة (Alhayan & Himdi, 2024)، أو الحكم هل هناك تحيز في مخرجات النموذج مثل قياس التحيز الجنسي التي عالجتها دراسة (Alrajhi et al., 2024)، أو الحكم على مقدرة اكتشاف النموذج للأوامر التي تحوي على استفسارات لا يُسمح للنموذج بتقديم إجابات عنها كما في دراسة (Nakhleh et al., 2024). وهذا التنوع في الأبحاث في المستوى غير اللغوي يؤكد أن المجال لا يتقدم فقط على مستوى التطبيقات، بل يشهد نضجًا كبيرًا في بناء أدواته الأساسية وتقييمها نقدياً للوقوف على نقاط القوة والضعف والعمل في ضوء ذلك.

٥-٥-٢- للإجابة عن سؤال البحث الثاني، وهو: "ما التحديات التي تواجه معالجة اللغة العربية من خلال النماذج اللغوية الضخمة؟"

فقد أظهرت نتائج المراجعة المنهجية عددًا من التحديات منها ما ذكرتها الدراسات مباشرةً ونصت عليها وهناك عدد من التحديات أظهرها واقع تلك الدراسات في معالجتها للغة العربية، وقد تمثلت تلك التحديات فيما يلي:

٥-٥-٢-١- ندرة البيانات العربية عالية الجودة ومحدوديتها

يُعد شح البيانات العربية الرقمية عالية الجودة أحد أبرز التحديات وأكثرها تكرارًا في الأدبيات. فقد أشار Sengupta et al. (2023, p. 3) إلى أن التحدي الأساسي في تطوير نموذج لغوي ضخم عربي هو محدودية البيانات العربية عالية الجودة، كما أكد Qian et al. (2024, p. 1) أن المحتوى العربي لا يشكل سوى جزء صغير فقط من مجموعة بيانات الويب المستخدمة للتدريب الأولي، حيث يمثل ٠,٦٪ فقط من البيانات، ولا يقتصر هذا التحدي على حجم البيانات وجودتها، بل يمتد إلى توازنها، حيث حذر Aboulola & Umer (2024, p. 3) من التحيزات الكامنة في خوارزميات التعلم الآلي وندرة مجموعات البيانات المصنفة. وفي محاولة للتغلب على هذه الندرة، يلجأ العديد من الباحثين إلى ترجمة مجموعات البيانات، وهو نهج حذرت

منه دراسات عدة؛ فقد أوضح Alahmadi et al (2024, p. 17780) أن الاعتماد على هذا النهج قد يكون له أثر على دقة استجابات النماذج.

أما عن الأثر المترتب عن البيانات ذات الجودة المتدنية فقد أشار Bari et al إلى أنه أثناء عملية تدريب نموذج allam لوحظ أن وجود ١٪ من البيانات المشوشة أثرت تأثيراً ملحوظاً على جودة النموذج (2024, p. 9).

٥-٥-٢- خصائص اللغة العربية ونظامها اللغوي

تتميز اللغة العربية بخصائص صرفية ونحوية ودلالية فريدة تشكل تحدياً كبيراً للنماذج التي غالباً ما تُصمم للغات أبسط تركيبياً. وقد وصفت العديد من الدراسات هذه الطبيعة بأنها تشكل تحدياً جوهرياً، حيث أشار Alkaoud (2024, p. 3) إلى أن اللغة العربية بقواعدها النحوية المعقدة وطبيعتها الاشتقاقية العالية تشكل تحدياً لعلماء اللغويات الحاسوبية، وهو ما يلاحظ في اللغة العربية فيعد النظام التركيبي فيها ذومرونة عالية فيمكن أن يبدأ بالفعل أو الاسم أو الحرف، وكذا يقبل النظام التقديم والتأخير، بل قد يؤثر على مجالات أخرى تتجاوز القواعد التركيبية بحد ذاتها، ومن ذلك ما أشار له Alrajhi et al. فيما يتعلق بالجنس النحوي (التذكير والتأنيث) الذي يجعل قياس التحيزات الاجتماعية أكثر صعوبة مقارنة بالإنجليزية (2024, p. 77406).

ويتفاقم هذا التعقيد عند غياب التشكيل الذي يعتري معظم النصوص الحديثة، مما يولد غموضاً دلالياً كبيراً، وهو ما وصفه Al-Duwais et al (2024, p. 3) بأنه أحد أسباب الغموض في أنظمة التعرف على الكيانات المسماة. وقد عرض Alhussain & Azmi (2024, p. 9) مثالاً عملياً لهذا الإشكال عندما أخطأ النموذج في فهم كلمة "شعر" بسبب تعدد معانيها المحتملة بناء على اختلاف التشكيل.

٥-٥-٣- إشكالية التقسيم (Tokenization)

تعد عملية التقسيم خطوة أساسية في ميدان معالجة اللغات الطبيعية فهي تقوم على تجزئة النصوص إلى سلسلة من الوحدات (Tokens)، وهذه العملية تمكن النموذج من فهم



اللغة المُعالجة على نحو أفضل. فعلى سبيل المثال تقطع كلمة writer باللغة الإنجليزية إلى [write,er]، وهذه المقسمات في الأساس مصممة للغات اللاتينية، وغالبًا تكون غير فعالة مع اللغة العربية. وقد أشار Abdelazim et al. (2024, p. 74) إلى أن السبب يعود إلى طبيعة مقسمات التشفير الإحصائي، فهي تُفَضِّل اللغات ذات الموارد العالية على حساب اللغات ذات الموارد المنخفضة، وينتج عن ذلك تمثيل غير فعال للنصوص العربية، حيث لوحظ أن بعض النماذج تفك ترميز الكلمات العربية إلى سلاسل من الحروف الأبجدية بدلاً من وحدات أكثر كفاءة (Zhu et al., 2024, p. 1) ويوضح Ammar et al. (2024, p. 4) هذه المشكلة بالإشارة إلى أن المقسم قد يجزئ كلمة بسيطة مثل "مرحباً" إلى حروف فردية لتصبح الكلمة بعد التقسيم هكذا [م، ر، ح، ب، ا، ء]، هذا التقسيم المفرط لا يزيد فقط من طول السلسلة ويبطئ المعالجة فحسب، بل يضعف أيضاً قدرة النموذج على الفهم العميق للروابط الصرفية والنحوية مما يقلل من فعالية المعالجة (Koubaa et al., 2024, p. 1).

٥-٢-٤-٥- قصور التقييم والمعايير المرجعية

يواجه الباحثون صعوبة في تقييم أداء النماذج العربية بدقة وموثوقية؛ بسبب نقص المعايير المرجعية (Benchmarks) المصممة خصيصاً للغة العربية. فقد أشار Alkaoud (2024, p. 1) إلى أنه ما يزال تقييم النماذج المدربة على اللغة العربية محدوداً. وحتى عند وجود معايير عربية، فإنها غالباً ما تكون غير كافية أو منحازة؛ حيث أشار Fang et al (2024, p. 507) إلى أن معايير تحليل المشاعر العربية تميل إلى أن تكون قديمة، بينما حذر Chouikhi et al (2024, p. 14) من أن العديد من المقاييس المستخدمة تُظهر تحيزاً يركز على اللغات الأجنبية. علاوة على ذلك، فإن مجرد ترجمة المعايير الإنجليزية قد يؤدي إلى أنواع إضافية من التحيزات غير المرغوب فيها (Alyafeai et al., 2024, p. 8)، وأشار Abdallah et al (2024, p. 8) إلى أن المقاييس التقليدية قد لا تكون مناسبة لتقييم إجابات النماذج التي تميل إلى الإسهاب.

٥-٢-٥-٥- تحديات البنية والموارد التقنية

بعيداً عن اللغة والبيانات، تبرز تحديات تتعلق بالبنية التحتية والقيود التقنية. فقد أشار Aljamel et al (2024, p. 62) إلى أن نقص البنية التحتية الحاسوبية الكافية قد يجبر الباحثين

على استخدام نماذج أصغر أو تدريبها بشكل غير كافٍ، مما يؤثر على النتائج. كما أن النماذج نفسها لديها قيود، مثل محدودية طول السياق الذي يمكن معالجته في المرة الواحدة (Alyafeai et al., 2023, p. 5)، وهنا يقع المختصون بين محاولة توفير نماذج صغيرة الحجم؛ لما لها من مميزات تشمل قلة التكلفة والسرعة، وكفاءة الأداء وجودته التي في الغالب تتطلب نماذج ذات معلمات ضخمة، وهي ما تتطلب تكاليف باهظة، فكلما زاد حجم النموذج (عدد المعاملات)، زادت الحاجة إلى ذاكرة وقوة معالجات لتشغيله. هذا يعني أن تشغيل النماذج الكبيرة يتطلب خوادم باهظة الثمن ومجهزة بوحدات معالجة قوية (Alyafeai et al., 2024, p. 1).

٥-٢-٦- تحديات الموارد البشرية

إلى جانب التحديات التقنية، تبرز عقبات جوهرية تتعلق بالموارد البشرية اللازمة لتجهيز البيانات وتقييم النماذج. فقد أشار Biswas et al. (2023, p. 2) إلى أن عملية توصيف البيانات، خاصة بيانات وسائل التواصل الاجتماعي، هي عملية شاقة ومكلفة وتتطلب جهداً بشرياً كبيراً. ويزداد الأمر تعقيداً بالحاجة الماسة إلى التنقيح اليدوي لضمان جودة البيانات (Boughorbel & Hawasly, 2023, p. 2).

ولا يقتصر الدور البشري المحوري على مرحلة الإعداد فقط، بل يمتد إلى مرحلة التقييم، خاصة وأن مقاييس التقييم الآلية أثبتت في سياقات عديدة أنها معقدة وغير موثوقة (Ammar et al., 2024, p. 2)، وتزداد الحاجة إلى مزيد من الكوادر البشرية مع اختلاف المجالات التي تُدرب عليها النماذج، كالدين والطب وغيرها، حيث لا يقتصر الأمر على إتقان اللغة، بل يتطلب تقييماً دقيقاً من خبراء متخصصين؛ لضمان صحة وموثوقية مخرجات النماذج، وهو ما يمثل تحدياً كبيراً في توفير الموارد البشرية المؤهلة (Adel et al., 2023, p. 239; Alnefaie et al., 2023, p. 129).

وهذا التحدي الذي أشارت إليه الدراسات سببه أن نماذج الذكاء الاصطناعي يتطلب تدريبها وزيادة كفاءتها على بيانات ضخمة بحق، وهذا يتطلب فرقاً كبيرة جداً من المتخصصين، الذين لا يقوم عليهم حجم البيانات فقط، بل أيضاً جودتها.



٥-٣-٥ - للإجابة عن سؤال البحث الثالث، وهو: "ما الحلول المقترحة لتحسين معالجة

النماذج اللغوية الضخمة للغة العربية؟"

عرضت الدراسات عدداً من الحلول المقترحة للتحسين من معالجة النماذج اللغوية الضخمة للغة العربية في عدة جوانب، وتمثلت تلك الحلول فيما يلي:

٥-٣-٥ - زيادة مصادر البيانات العربية الأصيلة والمتخصصة

لمواجهة التحدي الجوهرى المتمثل في ندرة البيانات، يتفق الباحثون على أن الحل الأساسى يكمن في بناء وتوسيع مصادر البيانات العربية عالية الجودة. تؤكد الدراسات أن الحل لا يكمن فقط في زيادة حجم البيانات، بل في ضمان كونها ذات جودة عالية وأصيلة ليست مصطنعة، حيث لا يكفي أي من الشرطين بمفرده لتحقيق أداء جيد (Alshahrani et al., 2023, p. 226)، ويتجلى هذا الحل عملياً في جهود مثل دراسة (Ali et al., 2023) التي قامت بتجميع أكبر مدونة لغوية عربية متنوعة بحجم يتجاوز ٥٠٠ جيجابايت، من خلال جمع نصوص من ٢٠ مصدرًا متنوعًا شملت الأخبار والكتب والشعر والمحتوى الرقمي، مما أدى إلى تفوق النموذج المدرب عليها. ولا يقتصر الحل على النصوص الخام، بل يشمل بناء مجموعات بيانات للتعليمات المتخصصة (instruction datasets)، التي تعد نادرة في اللغة العربية (Chouikhi et al., 2024, p. 1)، وكمثال على قوة هذا الحل، قامت دراسة (El-Sheikh et al., 2024) بإنشاء ٨ ملايين أمر عربي، وعند استخدامها لضبط نموذج بحجم ٧ مليارات معامل، تفوق هذا النموذج الصغير على نموذج آخر بحجم ٧٠ مليار معامل، مما يبرهن على القوة الهائلة لبيانات الضبط الدقيق. وللتغلب على شح البيانات، تقترح الدراسات حلولاً عملية مثل توليد بيانات اصطناعية (Al-Shameri & Al-Khalifa, 2024, p. 121896)، على سبيل المثال، استخدام تقنية الترجمة العكسية (Back-translation) بترجمة الجمل العربية إلى الإنجليزية ثم إعادتها للعربية لإنشاء صيغ جديدة ومتنوعة، أو كما فعل (Zakraoui et al., 2022). بتوليد نصوص صحية طويلة من عناوين المقالات لزيادة حجم بيانات التلخيص في مجال ذي موارد محدودة، بالإضافة إلى استخدام النماذج اللغوية التي تقوم بتوليد نصوص جديدة، هذا كله يضعه الباحثون في عين الاعتبار عندما يضطرون لذلك في حال نقصان البيانات

فيما يسمى عملياً بزيادة البيانات "Data augmentation"، مع اتفاقهم على أن البيانات كلما كانت أصيلة كان ذلك أفضل لتمثيلها الحقيقي لاستخدام اللغة على أرض الواقع.

ومن هنا تتضح أهمية العناية بالبيانات كونها إحدى ركني قوة الذكاء الاصطناعي، وهي إحدى الأسباب الجلية التي جعلت النماذج المدربة على بيانات باللغة الإنجليزية ذات أداء أفضل؛ وهذا بسبب جودة البيانات وضخامتها (Boughorbel & Hawasly, 2023, p. 5).

٥-٥-٣-٢- تطوير نماذج ومنهجيات مخصصة للغة والسياق العربي

أحد التحديات التي ناقشتها الأبحاث ضعف النماذج اللغوية في توليد وتحليل العربية في مقابل اللغة الإنجليزية، وأحد الحلول المقترحة لعلاج هذا التحدي هو تطوير نماذج ومنهجيات تتناسب مع الطبيعة الفريدة للغة العربية وسياقها الثقافي. وقد أشارت الأبحاث إلى أهمية بناء نماذج مخصصة للغة العربية؛ لتحسين مستوى المعالجة، فقد أظهرت دراسة Bari et al. (2024, p. 4) أن النموذج المبني من الصفر على بيانات عربية أظهر تحسناً كبيراً مقارنة بالنموذج الإنجليزي الذي درب فيما بعد على العربية، ومن ذلك تفوق نموذج AraT5 المخصص للعربية في مهام محددة دُرّب عليها، على نموذجي LLaMA-3 و ChatGPT وهما نموذجان عامان (Khondaker et al., 2024, p. 288)، وهو ما أكدته Sengupta et al. (2023, p. 24) بقوله إن النماذج متعددة اللغات تميل إلى الأداء بشكل أسوأ بكثير من النماذج أحادية اللغة المخصصة.

ویدخل في هذا السياق تحسين المقسمات (Tokenizers) فبدلاً من تقطيع كلمة بسيطة إلى حروف فردية (Ammar et al., 2024, p. 4)؛ مما يسبب ضعفاً في الأداء وزيادة في استهلاك الموارد، توصي الدراسات بتكييف المقسم أو تدريب مقسم جديد ليكون أكثر كفاءة في اللغة الهدف (Alyafeai et al., 2024, p. 8) أو استخدام استراتيجيات مثل "التوسع التدريجي للمفردات" التي طبقتها دراسة (Zhu et al., 2024, p. 2) وأدت إلى نماذج تفك ترميز النصوص العربية بسرعة بلغت ثلاثة أضعاف سابقاتها، وتقوم فلسفة هذه الطريقة على الكيفية التي يكتسب فيها متعلم اللغة الأجنبية اللغة، من خلال البدء بالكلمات الأساسية ثم زيادة الكلمات الجديدة تدريجياً، هذه الطريقة أعطت كفاءة لمقسم النموذج في فهم اللغة وتوليدها.



٥-٥-٤- إستراتيجيات زيادة كفاءة النماذج اللغوية الضخمة في معالجة العربية

إن أحد التحديات المشاهدة التي تواجه النماذج اللغوية الضخمة في معالجة العربية تدني كفاءتها في مقابل معالجتها للغات تتمتع بوفرة الموارد كاللغة الإنجليزية، وهنا تبرز عدد من الإستراتيجيات التي أشارت إليها الدراسات ويمكن تلخيصها فيما يلي:

٥-٥-٤-١- نقل التعلم عبر اللغات (Cross-lingual Transfer)

وفيه تُدرب النماذج على مزيج من البيانات العربية والإنجليزية، مما يسمح لها بنقل قدرات الاستدلال والمعرفة المكتسبة من البيانات الإنجليزية الوفيرة إلى السياق العربي. حيث أشارت دراسة (Sengupta et al., 2023) إلى أن إضافة اللغة الإنجليزية إلى بيانات التدريب حسّنت النتائج للغة العربية، وهي إستراتيجية أساسية نجحت في نماذج مثل Jais، ومثل هذا ما قام به Al-Duwais et al حين درّبوا نموذجاً على مهمة التعرف على الكيانات باللغة الإنجليزية مثل الأشخاص والأماكن والمؤسسات، ثم قاسوا التعرف على الكيانات باللغة العربية ليجدوا أن النموذج استطاع التوصل إلى تحديد أنواع الكيانات في النص العربي المدخل مطبقاً القواعد التي تعلمها من الإنجليزية (Al-Duwais et al., 2024)، والذي ساعد في هذه الخطوة، أن البيانات باللغة الإنجليزية تميزت بالوفرة والجودة ما مكّن النموذج عند التعامل معه بالعربية من الاستفادة من القدرات التي اكتسبها من اللغة الإنجليزية.

٥-٥-٤-٢- هندسة الأوامر (Prompt Engineering)

يقصد بهندسة الأوامر النصوص التي يطلب من النموذج تنفيذها، فقد أظهرت الدراسات أن طريقة صياغة الأمر للنموذج تتحكم في جودة مخرجاته، وأشارت إلى عدد من الطرق عند صياغة الأمر ومنها:

كتابة أمر النموذج باللغة الإنجليزية

هذه العملية يمكن أن تحسن أداء النموذج في العربية، فقد لاحظ Khondaker et al. (2023, p. 3) أن تقديم الأمر باللغة الإنجليزية أفضل من تقديمه باللغة العربية من حيث الأداء، مما دفعهم لتصميم قالب أوامر عالمي بالإنجليزية. وكمثال أكثر تقدماً، استخدم Gupta et al. (٢٠٢٥)

تقنية التلقين عبر اللغات (CLP)، حيث يتم أولاً شرح منهجية الحل للنموذج بالإنجليزية ثم يُطلب منه تطبيقها على الادعاء العربي، مما رفع دقة التحقق من الادعاءات العربية إلى ٨٦,٥٥٪.

إضافة أمثلة توضيحية في الأمر

أظهرت الدراسات أن تعزيز الأوامر (prompts) بأمثلة توضيحية، يؤدي إلى تحسن تدريجي ومباشر في الدقة، وتزيد تلك الدقة مع زيادة عدد الأمثلة، وهذا ما ذكره Biswas et al حيث لاحظ ارتفاع متوسط الدقة من ٨٢,١٤٪ في اختبارات التعلم بدون أمثلة إلى ٨٣,٨٥٪ عند استخدام مثال واحد وصولاً إلى ٨٥,٥٧٪ مع استخدام أمثلة أكثر (Biswas et al., 2023, p. 1)، وفسروا هذا الأداء المتفوق بأن النموذج يتعلم من الأمثلة المقدمة مباشرة ضمن سياق الأمر، مما يمكنه من فهم المهمة المطلوبة فهماً أفضل وتقديم استجابات أكثر دقة وموثوقية (Biswas et al., 2023, p. 3).

إستراتيجيات لتحديات نقص الموارد البشرية

طبقت الدراسات عدداً من الإستراتيجيات التي عالجت فيها نقص الموارد البشرية عند معالجة اللغة العربية ومن هذه الإستراتيجيات:

٥-٤-٣- النموذج المعلم teacher model

جاءت هذه الإستراتيجية كحل لندرة البيانات المصنفة التي يُحتاج فيها إلى عدد من المختصين ليصنفوا البيانات التي دُرِبَ عليها النماذج، حيث يتم تدريب "نموذج معلم" على بيانات مصنفة قليلة ليقوم بعد ذلك بتصنيف البيانات غير المصنفة، ثم تؤخذ هذه البيانات المصنفة لتستخدم في تدريب نموذج آخر عادة ما يسمى النموذج الطالب (Almasri et al., 2023, p. 1).

٥-٤-٤- النموذج الحكم judge model

وهنا يُحتكم إلى نموذج يكون في العادة ذا كفاءة عالية في نقطة معينة، ليحكم على مخرجات نموذج آخر، وهذا ما قام به Boughorbel & Hawasly حيث قدما للنموذج الذي درياه مجموعةً



من الأسئلة ثم أُرسلت إجاباته إلى النموذج الحكم، وفي هذا المثال كان النموذج الحكم هو GPT-4 الذي يحكم على جودة الإجابة بناءً على معايير مثل الفائدة، والصدق، وعدم الضرر، والطلاقة، ويعطيها درجة (Boughorbel & Hawasly, 2023).

٥-٥-٤-٥- بناء منظومة تقييم عربية أصيلة وموثوقة

سبق ذكر النقص الذي يواجه مقاييس تقييم النماذج العربية وأن الاعتماد على ترجمة مقاييس التقييم الإنجليزية قد يكون مضللاً ومتحيزاً؛ لذا يكمن الحل في بذل جهود حثيثة لتطوير وبناء معايير مرجعية (Benchmarks) مصممة خصيصاً للغة العربية وسياقها الثقافي، ومن تلك المعايير التي أطلقت معيار AlGhafa، وهو معيار تقييم لنماذج اللغة العربية عبر مجموعة متنوعة من المهام التي حُولت إلى صيغة اختيار من متعدد، وُجمعت هذه المهام من قواعد بيانات مختلفة بلغ عددها (١١) قاعدة، وعولجت ودققت بشرياً، وشمل هذا المعيار مهام مثل: الفهم القرائي، وتصنيف صحة الادعاءات، وتحليل المشاعر، والإجابة عن أسئلة، وتوليد الأسئلة (Almazrouei et al., 2023).

٦. التوصيات والمقترحات

يُعدّ تمكين النماذج اللغوية الضخمة من معالجة اللغة العربية وفهمها بعمق ركناً أساسياً في مواكبة متطلبات الثورة الرابعة، وتعزيز الاقتصاد المعرفي في العالم العربي. فإتقان هذه النماذج للغة العربية يمثل وسيلة إستراتيجية للحفاظ على الهوية اللغوية في العصر الرقمي، وتمكين أكثر من ٤٠٠ مليون ناطق بها من الوصول إلى المعرفة العالمية والمشاركة فيها بفعالية، فضلاً عن إثراء المحتوى الرقمي العربي وفتح آفاق اقتصادية وإبداعية واسعة. ومع ذلك، فإن هذا الميدان الواعد يواجه تحديات جوهرية، كما أظهرت نتائج هذه المراجعة المنهجية، حيث تتنوع التحديات بين ندرة البيانات العربية عالية الجودة، والتعقيد الكامن في بنية اللغة العربية صرفياً ونحوياً، وقصور أدوات التقييم والمعايير المرجعية، بالإضافة إلى فجوة الأداء الملحوظة مقارنة باللغة الإنجليزية.

وتُشكّل البيانات والموارد العربية المتخصصة الركيزة الأساس التي يقوم عليها تقدم هذا المجال برمته، حيث إن جودة النماذج وقدراتها تعتمد بشكل مباشر على جودة وكمية البيانات التي تُدرب عليها. وهذه التحديات تفرض على المؤسسات البحثية والجامعات

والشركات التقنية وصناع السياسات أن توحيد جهودها وتكاتف لبناء منظومة متكاملة تضمن تطوير نماذج لغوية قادرة على فهم اللغة العربية بعمق وتلبية احتياجاتها الفريدة. ولتحقيق هذا الهدف، توصي الدراسة الحالية بالتركيز على الجوانب التالية:

١. بناء وتطوير بنية تحتية متكاملة للبيانات والمعايير العربية: وذلك من خلال إطلاق مبادرات لجمع وتصنيف مجموعات بيانات عربية ضخمة ومتنوعة وعالية الجودة، تغطي مختلف المجالات مثل: الصحة، والتعليم، والقانون، واللهجات والأساليب اللغوية، كما دعت إليه دراسات عدة (Nagi et al., 2024; Mohammad et al., 2024) ويجب أن يصاحب ذلك إنشاء وتوحيد معايير تقييم (Benchmarks) عربية مفتوحة المصدر، لا تقتصر على المهام التقليدية بل تمتد لتقييم القدرات المعقدة كالاستدلال والفهم الثقافي، كما أشار (Sengupta et al., 2023).

٢. الاستثمار في تطوير نماذج عربية خالصة ومعماريات مخصصة: فبدلاً من الاعتماد الكلي على تكييف النماذج متعددة اللغات، يجب دعم الأبحاث التي تهدف إلى بناء نماذج لغوية ضخمة تتمحور حول اللغة العربية من الأساس، كما فعلت دراسات مثل (Koubaa et al., 2024)، ويشمل ذلك تطوير مقسمات (tokenizers) مُحسَّنة للغة العربية لزيادة الكفاءة (Alyafeai et al., 2024).

٣. ينبغي تشجيع التعاون الوثيق بين علماء الحاسوب واللغويين وخبراء المحتوى في مختلف المجالات؛ لضمان أن تكون النماذج المطورة دقيقة لغوياً وملائمة سياقياً، وتجدر الإشارة إلى أن التعاون بين الحاسوبيين واللغويين لم يكن ظاهراً فيما وقف عليه البحث من دراسات، بل العجيب أن الأبحاث تشير في باب التحديات إلى تعقيد نظام اللغة العربية، وهو ما يخالف آراء اللغويين الحاسوبيين من العرب مثل نبيل علي ونهاد الموسى وغيرهم الذين يشيرون إلى إمكانية معالجة اللغة حاسوبياً.

٧. النتائج

قدمت هذه المراجعة المنهجية بعد تحليلها لـ ٨٢ دراسة علمية، مجموعة من النتائج والإسهامات المعرفية التي ترسم ملامح ميدان معالجة اللغة العربية



بواسطة النماذج اللغوية الضخمة، ويمكن تلخيص أبرز ما أنجزته هذه الدراسة في النقاط التالية:

١. بناء خريطة معرفية للمجال

أنجزت الدراسة مسحاً منهجياً يوثق ويصنف الأبحاث المنشورة في مجال النماذج اللغوية العربية خلال الفترة بين (٢٠٢٠-٢٠٢٤). وقد أسفر هذا المسح عن خريطة معرفية كشفت عن توجهات الأبحاث والواقع الذي عالجته.

٢. تحديد التحديات التي تواجه معالجة النماذج اللغوية الضخمة للغة العربية

أظهرت نتائج المراجعة المنهجية وجود تحديات جوهرية متعددة تواجه النماذج اللغوية الضخمة في معالجتها للغة العربية. يأتي في مقدمة هذه التحديات ندرة البيانات العربية عالية الجودة ومحدوديتها، حيث لا يمثل المحتوى العربي سوى نسبة ضئيلة من بيانات الشبكة العالمية.

٣. طرح للحلول في ميدان معالجة اللغة العربية والنماذج اللغوية الضخمة

سلطت الدراسة الضوء على عدد من الحلول في مقابل التحديات التي تواجه ميدان معالجة النماذج اللغوية الضخمة للغة العربية، ومن أهم تلك الحلول الاهتمام ببناء وتوسيع مصادر البيانات العربية الأصيلة وعالية الجودة، إلى جانب تطوير نماذج ومنهجيات مخصصة تتناسب مع الخصائص الفريدة للغة، كما تضمنت الحلول إمكانية تطبيق عدد من الإستراتيجيات لزيادة كفاءة النماذج اللغوية مثل نقل التعلم وهندسة الأوامر.

المراجع العربية

العرفج، عبد الحميد بن عبدالله، المعيلي، فهد علي، الحربي، مشعل بن سعد، الرشيد، مسفر بن ناصر، والبشر، ماهر بن عبدالرحمن. (٢٠٢٢). مراجعة منهجية للأدبيات في مجال تربية الموهوبين خلال الفترة بين ٢٠١٨-٢٠٢٠. *مجلة القراءة والمعرفة*، (٢٥٢)، ٤٠٣-٤٨٣.

<https://doi.org/10.21608/mrk.2022.267463>

رومنة المراجع العربية

Al-'Urfaj, 'Abd al-Hamid bin 'Abd Allah, Al-Mu'ayli, Fahd 'Ali, Al-Harbi, Mish'al bin Sa'd, Al-Rashidi, Misfir bin Nasir, & Al-Bishr, Mahir bin 'Abd al-Rahman. (2022). A systematic review of the literature in the field of gifted education during the period 2018–2020. *Reading and Knowledge Journal*, (252), 403–483. <https://doi.org/10.21608/mrk.2022.267463>

المراجع الأجنبية

Abdallah, A., Kasem, M., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., & Jatowt, A. (2024, July). Arabicaqa: A comprehensive dataset for arabic question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2049-2059).

Abdelaal, A., Medhat, A. A., Elsayad, M., Foad, M., Khaled, S., Tamer, A., & Medhat, W. (2024). Text Correction for Modern Standard Arabic. *Procedia Computer Science*, 244, 371-377.

Abdelali, A., Mubarak, H., Chowdhury, S. A., Hasanain, M., Mousi, B., Boughorbel, S., ... & Alam, F. (2023). LARA-Bench: Benchmarking Arabic AI with large language models. *arXiv preprint arXiv:2305.14982*.

Abdelazim, H., Begemy, T., Galal, A., Sedki, H., & Mohamed, A. (2024). Multi-hop arabic llm reasoning in complex qa. *Procedia Computer Science*, 244, 66-75.



- Aboulola, O., & Umer, M. (2024). Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. *Scientific Reports*, 14(1), 30463.
- Adel, Y., Dorrah, M., Ashraf, A., ElSaadany, A., Mohamed, M., Wael, M., & Khoriba, G. (2023, December). *AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text*. In 2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC) (pp. 235-239). IEEE.
- Al Wazrah, A., & Alhumoud, S. (2021). Sentiment analysis using stacked gated recurrent unit for arabic tweets. *IEEE Access*, 9, 137176-137187.
- Al-Duwais, M., Al-Khalifa, H., & Al-Salman, A. (2024). A Benchmark Evaluation of Multilingual Large Language Models for Arabic Cross-Lingual Named-Entity Recognition. *Electronics*, 13(17), 3574.
- Al-Khafaji, N. J., & Majeed, B. K. (2024, August). *Evaluating Large Language Models using Arabic Prompts to Generate Python Codes*. In 2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA) (pp. 1-5). IEEE.
- Al-Shameri, N. A., & Al-Khalifa, H. S. (2024). Arabic paraphrase generation using transformer-based approaches. *IEEE Access*, 12, 121896-121914.
- Alabdulrahman, M., Khayyat, R., Almowallad, K., & Alharz, Z. (2024, March). *Sarid: Arabic Storyteller Using a Fine-Tuned LLM and Text-to-Image Generation*. In 2024 16th International Conference on Computer and Automation Engineering (ICCAE) (pp. 1-5). IEEE.
- Alahmadi, M. D., Alharbi, M., Tayeb, A., & Alshangiti, M. (2024). Evaluating large language models' proficiency in answering Arabic GAT exam questions. *Engineering, Technology & Applied Science Research*, 14(6), 17774-17780
- Alhayan, F., & Himdi, H. (2024). Ensemble learning approach for distinguishing human and computer-generated Arabic reviews. *PeerJ Computer Science*, 10, e2345.
- Alhussain, A., & Azmi, A. M. (2024). Beyond event-centric narratives: Advancing Arabic story generation with large language models and beam search. *Mathematics*, 12(10), 1548.

- Ali, A. R., Siddiqui, M. A., Algunaibet, R., & Ali, H. R. (2023). A large and diverse Arabic corpus for language modeling. *Procedia Computer Science*, 225, 12–21.
- Ali, S. B., Kechaou, Z., & Wali, A. (2022, December). *Arabic fake news detection in social media Based on AraBERT*. In 2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (pp. 214-220). IEEE.
- Aljamel, A., Khalil, H., & Aburawi, Y. (2024). Comparative study of fine tuned BERT-based models and RNN-based models. case study: Arabic fake news detection. *The International Journal of Engineering & Information Technology (IJEIT)*, 12(1), 56-64.
- Alkaoud, M. (2024). A bilingual benchmark for evaluating large language models. *PeerJ Computer Science*, 10, e1893.
- Almasri, M., Al-Malki, N., & Alotaibi, R. (2023). A semi supervised approach to Arabic aspect category detection using Bert and teacher-student model. *PeerJ Computer Science*, 9, e1425.
- Almazrouei, E., Cojocaru, R., Baldo, M., Malartic, Q., Alobeidli, H., Mazzotta, D., ... & Noune, B. (2023, December). AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023* (pp. 244-275).
- Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023, October). *Is gpt-4 a good islamic expert for answering quran questions?*. In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023) (pp. 124-133).
- Alrajhi, W., Al-Khalifa, H. S., & Al-Salman, A. S. (2024). Quantifying gender bias in arabic pre-trained language models. *IEEE Access*, 12, 77406-77420.
- Alshahrani, S., Alshahrani, N., Dey, S., & Matthews, J. (2023, December). Performance Implications of Using Unrepresentative Corpora in Arabic Natural Language Processing. In *Proceedings of ArabicNLP 2023* (pp. 218-231).
- Alshammari, H., & Elleithy, K. (2024). Toward robust Arabic AI-generated text detection: Tackling diacritics challenges. *Information*, 15(7), 419.
- Alshammari, W. T., & AlHumoud, S. (2022). TAQS: an Arabic question similarity system using transfer learning of BERT with BiLSTM. *IEEE Access*, 10, 91509-91523.



- Alsubhi, K., Jamal, A., & Alhothali, A. (2022). Deep learning-based approach for Arabic open domain question answering. *PeerJ Computer Science*, 8, e952.
- Alyafeai, Z., Alshaibani, M. S., AlKhamissi, B., Luqman, H., Alareqi, E., & Fadel, A. (2023). Taqyim: Evaluating arabic nlp tasks using chatgpt models. *arXiv preprint arXiv:2306.16322*.
- Alyafeai, Z., Pieler, M., Teufel, H., Tow, J., Bellagente, M., Phung, D., ... & Riquelme, C. (2024). Arabic Stable LM: Adapting stable lm 2 1.6 b to arabic. *arXiv preprint arXiv:2412.04277*.
- Ammar, A., Koubaa, A., Benjdira, B., Nacar, O., & Sibae, S. (2024). Prediction of arabic legal rulings using large language models. *Electronics*, 13(4), 764.
- Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., ... & Khan, H. (2024). Allam: Large language models for arabic and english. *arXiv preprint arXiv:2407.15390*.
- Bassem, B., Abraham, A., & Mounir, Z. (2024). Transformer Models for Authorship Profiling in Arabic Social Media Texts. *International Journal of Computer Information Systems and Industrial Management Applications*, 16(3), 13-13.
- Biswas, M. R., Mohsen, F., Shah, Z., & Zaghouani, W. (2023, November). *Potentials of chatgpt for annotating vaccine related tweets*. In 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-6). IEEE.
- Biswas, M. R., Shah, Z., & Zaghouani, W. (2024). MemeMind at ArAIEval Shared Task: spotting persuasive spans in arabic text with persuasion techniques identification. *arXiv preprint arXiv:2408.04540*.
- Boughorbel, S., & Hawasly, M. (2023). Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. *arXiv preprint arXiv:2310.14819*.
- Burns, S. (2019). *Natural Language Processing: A Quick Introduction to NLP with Python and NLTK (Step-by-Step Tutorial for Beginners)*. Amazon KDP Printing and Publishing C.
- Chouikhi, H., Aloui, M., Hammou, C. B., Chaabane, G., Kchaou, H., & Dhaouadi, C. (2024). GemAr: Enhancing LLMs Through Arabic Instruction-Tuning. *arXiv preprint arXiv:2407.02147*.

- Einieh, Y., Almansour, A., & Jamal, A. (2022, December). *Fine tuning an AraT5 transformer for Arabic abstractive summarization*. In 2022 14th international conference on computational intelligence and communication networks (CICN) (pp. 194-198). IEEE.
- El-Sheikh, A., Elmogtaba, A., Darwish, K., Elmallah, M., Elneima, A., & Sawaf, H. (2024). Creating Arabic LLM prompts at scale. *arXiv preprint arXiv:2408.05882*.
- Elchafei, P., & Fashwan, A. (2024). Arabic NER Evaluation: Pre-Trained Models via Contrastive Learning vs. LLM Few-Shot Prompting. *Procedia Computer Science*, 244, 229-237.
- Fadel, A. S., Abulnaja, O. A., & Saleh, M. E. (2023). Multi-Task Learning Model with Data Augmentation for Arabic Aspect-Based Sentiment Analysis. *Computers, Materials & Continua*, 75(2).
- Fang, Y., Xu, C., Guan, S., Yan, N., & Mei, Y. (2024, November). *Advancing Arabic Sentiment Analysis: ArSen Benchmark and the Improved Fuzzy Deep Hybrid Network*. In Proceedings of the 28th Conference on Computational Natural Language Learning (pp. 507-516).
- Gupta, A., Singhal, A., Law, T., Rao, V., Duan, E., & Li, R. L. (2025, January). Can LLMs Verify Arabic Claims? Evaluating the Arabic Fact-Checking Abilities of Multilingual LLMs. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script* (pp. 104-113).
- Hasanain, M., Hasan, M. A., Ahmed, F., Suwaileh, R., Biswas, M. R., Zaghouani, W., & Alam, F. (2024). ArAIEval shared task: propagandistic techniques detection in unimodal and multimodal Arabic content. *arXiv preprint arXiv:2407.04247*.
- Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., ... & Xu, J. (2023). Acegpt, localizing large language models in arabic. *arXiv preprint arXiv:2309.12053*.
- Jones, C., & Bergen, B. (2024, June). *Does GPT-4 pass the Turing test?*. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 5183-5210).
- Karajeh, O., Al-Kabi, M. N., & Fox, E. A. (2023, December). *Fusing AraBERT and graph neural networks for enhanced Arabic text classification*. In 2023 24th International Arab Conference on Information Technology (ACIT) (pp. 1-8). IEEE.



- Khaled, S., Mohamed, E. H., & Medhat, W. (2024). Evaluating large language models for Arabic sentiment analysis: A comparative study using retrieval-augmented generation. *Procedia Computer Science*, 244, 363-370.
- Khondaker, M. T. I., Naeem, N., Khan, F., Elmadany, A., & Abdul-Mageed, M. (2024, August). *Benchmarking llama-3 on arabic language generation tasks*. In Proceedings of The Second Arabic Natural Language Processing Conference (pp. 283-297).
- Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.
- Koubaa, A., Ammar, A., Ghouti, L., Najar, O., & Sibae, S. (2024). Arabiangpt: Native arabic gpt-based large language model. *arXiv preprint arXiv:2402.15313*.
- Kwon, S. Y., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). *Beyond English: Evaluating LLMs for Arabic grammatical error correction*. In Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023) (pp. 101-119). Association for Computational Linguistics.
- Lotfy, A., Saleh, K., Mohamed, S., Lorange, J., Yehia, E., Mohammed, K., ... & Yasser, T. (2024, March). *Sentiment analysis for arabic product reviews using llms and knowledge graphs*. In 2024 6th International Conference on Computing and Informatics (ICCI) (pp. 411-417). IEEE.
- Mahdhaoui, H., Mars, A., & Zrigui, M. (2023, September). *Optimizing Arabic named entity recognition through active learning and AraBERT*. In 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA) (pp. 1-5). IEEE.
- Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90, 46-60.
- Marie-Sainte, S. L., Alalyani, N., Alotaibi, S., Ghouzali, S., & Abunadi, I. (2018). Arabic natural language processing and machine learning-based systems. *IEEE Access*, 7, 7011-7020.
- Mohammad, R., Alkhnbashi, O. S., & Hammoudeh, M. (2024). Optimizing large language models for Arabic healthcare communication: A focus on patient-centered NLP applications. *Big Data and Cognitive Computing*, 8(11), 157.

- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery*, 8(5), 336-341.
- Mohsen, M. (2024). Artificial intelligence in academic translation: a comparative study of large language models and google translate. *Psycholinguistics*, 35(2), 134-156.
- Nagi, K. A., Alzain, E., & Naji, E. (2024). Informed Prompts and Improving ChatGPT English to Arabic Translation. *Al-Andalus Journal for Humanities & Social Sciences*, 11(98).
- Nakhleh, S., Qasaimeh, M., & Qasaimeh, A. (2024, August). *Character-level adversarial attacks evaluation for arabert's*. In 2024 15th international conference on information and communication systems (ICICS) (pp. 1-6). IEEE.
- Nehar, A., Bellaouar, S., Souffi, S., & Bouameur, M. (2023, October). *Dhati: a Fine-tuned Large Language Model for evaluating Subjectivity in Arabic Textual Data*. In 2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-7). IEEE.
- Obeidat, R., Bashayreh, A., & Younis, L. B. (2022, June). *The impact of combining arabic sarcasm detection datasets on the performance of bert-based model*. In 2022 13th International Conference on Information and Communication Systems (ICICS) (pp. 22-29). IEEE.
- Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Communications of the association for information systems*, 37.
- Ortiz, A. F. H., Camacho, E. C., Rojas, J. C., Camacho, T. C., Guevara, S. Z., Cuenca, N. T. R., ... & Malo, R. G. (2021). A practical guide to perform a systematic literature review and meta-analysis. *Principles and Practice of Clinical Research*, 7(4), 47-57.
- Qarah, F. (2024). AraPoemBERT: A pretrained language model for Arabic poetry analysis. *arXiv preprint arXiv:2403.12392*.
- Qian, Z., Altam, F., Alqurishi, M., & Souissi, R. (2024). Camelevel: Advancing culturally aligned arabic language models and benchmarks. *arXiv preprint arXiv:2409.12623*.
- Raschka, S. (2025). *Build a large language model (from scratch)*. Manning Publications.
- Rasheed, A. F., Zarkoosh, M., Abbas, S. F., & Al-Azzawi, S. S. (2023, December). *Arabic offensive language classification: Leveraging transformer, lstm, and svm*. In 2023 IEEE

International Conference on Machine Learning and Applied Network Technologies (ICMLANT) (pp. 1-6). IEEE.

Sallam, M., Al-Mahzoum, K., Alshuaib, O., Alhajri, H., Alotaibi, F., Alkhurainej, D., ... & Egger, J. (2024). Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC Infectious Diseases*, 24(1), 799.

Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., ... & Xing, E. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of business research*, 104, 333-339.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. *Advances in neural information processing systems*, 30.

Zakraoui, J., AlJa'am, J. M., & Salah, I. (2022, December). Domain-specific text generation for Arabic text summarization. In 2022 International Conference on Computer and Applications (ICCA) (pp. 1-4). IEEE.

Zhang, Q., Yu, H., Barbiero, M., Wang, B., & Gu, M. (2019). Artificial neural networks enabled by nanophotonics. *Light: Science & Applications*, 8(1), 42.

Zhu, J., Huang, H., Lin, Z., Liang, J., Tang, Z., Almubarak, K., ... & Xu, J. (2024). Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion. *arXiv preprint arXiv:2412.12310*.

المواقع الإلكترونية

W3Techs. (2025, July 20). Usage statistics of content languages for websites. W3Techs.com. https://w3techs.com/technologies/overview/content_language

Wikipedia contributors. (2025, July 20). Languages used on the Internet. Wikipedia. https://en.wikipedia.org/wiki/Languages_used_on_the_Internet

نبذة تعريفية عن الباحث

Author Biodata Khalawi Samir Al-Ayyadi is an Associate Professor of Teaching Arabic for Non-Native Speakers in the Department of Teacher Preparation and Training, Institute of Teaching Arabic to Non-Native Speakers, at the Islamic University in the Kingdom of Saudi Arabia. Dr. Al-Ayyadi received his PhD degree in Teaching Arabic for Non-Native Speakers in 2017 from the Islamic University. His research interests include Teaching Arabic to Non-Native Speakers, Computational Linguistics, and Machine Learning.	بيانات الباحث: خليوي سامر خليوي العياضي، أستاذ مشارك في تعليم اللغة العربية لغير الناطقين بها في (قسم إعداد وتدريب المعلمين) (بمعهد تعليم اللغة العربية لغير الناطقين بها) في الجامعة الإسلامية بالمملكة العربية السعودية. حاصل على درجة الدكتوراه في تعليم اللغة العربية لغير الناطقين بها من الجامعة الإسلامية عام ٢٠١٧. تدور اهتماماته البحثية حول تعليم اللغة العربية لغير الناطقين بها، واللسانيات الحاسوبية، وتعليم الآلة. لديه خبرة في البرمجة لأكثر من ١٥ سنة في عدة لغات وعمل على العديد من أطر العمل سواءً في تطبيقات الجوال أو المواقع الإلكترونية.
معرف أوركيد (ORCID) : 0009-0006-7379-1052	
Email: khlewe@gmail.com	



الملاحق

ملحق (أ): قائمة بعناوين الدراسات التي أجريت عليها المراجعة المنهجية

1. Abd El-Mageed, S. M., Aboutabl, A. E., & Mohamed, E. H. (2024, July). ArFakeDetect: A Deep Learning Approach for Detecting Fabricated Arabic Tweets on COVID-19 Vaccines. In 2024 Intelligent Methods, Systems, and Applications (IMSA) (pp. 103-108). IEEE.
2. Abdallah, A., Kasem, M., Abdalla, M., Mahmoud, M., Elkasaby, M., Elbendary, Y., & Jatowt, A. (2024, July). Arabicaqa: A comprehensive dataset for arabic question answering. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2049-2059).
3. Abdelaal, A., Medhat, A. A., Elsayad, M., Foad, M., Khaled, S., Tamer, A., & Medhat, W. (2024). Text Correction for Modern Standard Arabic. Procedia Computer Science, 244, 371-377.
4. Abdelali, A., Mubarak, H., Chowdhury, S. A., Hasanain, M., Mousi, B., Boughorbel, S., ... & Alam, F. (2023). LAraBench: Benchmarking Arabic AI with large language models. arXiv preprint arXiv:2305.14982.
5. Abdelazim, H., Begemy, T., Galal, A., Sedki, H., & Mohamed, A. (2024). Multi-hop arabic llm reasoning in complex qa. Procedia Computer Science, 244, 66-75.
6. Aboulola, O., & Umer, M. (2024). Novel approach for Arabic fake news classification using embedding from large language features with CNN-LSTM ensemble model and explainable AI. Scientific Reports, 14(1), 30463.
7. Adel, Y., Dorrah, M., Ashraf, A., ElSaadany, A., Mohamed, M., Wael, M., & Khoriba, G. (2023, December). AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text. In 2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC) (pp. 235-239). IEEE.
8. Al Deen, M. M. S., Pielka, M., Hees, J., Abdou, B. S., & Sifa, R. (2023, December). Improving Natural Language Inference in Arabic Using Transformer Models and Linguistically Informed Pre-Training. In 2023 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 318-322). IEEE.

9. Al Mraikhat, O. R., Hamoud, H., & Zaraket, F. A. (2024, August). AREEj: Arabic Relation Extraction with Evidence. In Proceedings of the Second Arabic Natural Language Processing Conference (pp. 67-72).
10. Al Wazrah, A., & Alhumoud, S. (2021). Sentiment analysis using stacked gated recurrent unit for arabic tweets. IEEE Access, 9, 137176-137187.
11. Al-Bashabsheh, E., Al-Khazaleh, H., Elayan, O., & Duwairi, R. (2021, December). Commonsense validation for Arabic sentences using deep learning. In 2021 22nd International Arab Conference on Information Technology (ACIT) (pp. 1-7). IEEE.
12. Al-Duwais, M., Al-Khalifa, H., & Al-Salman, A. (2024). A Benchmark Evaluation of Multilingual Large Language Models for Arabic Cross-Lingual Named-Entity Recognition. Electronics, 13(17), 3574.
13. Al-Khafaji, N. J., & Majeed, B. K. (2024, August). Evaluating Large Language Models using Arabic Prompts to Generate Python Codes. In 2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA) (pp. 1-5). IEEE.
14. Al-Shameri, N. A., & Al-Khalifa, H. S. (2024). Arabic paraphrase generation using transformer-based approaches. IEEE Access, 12, 121896-121914.
15. Alabdulrahman, M., Khayyat, R., Almowallad, K., & Alharz, Z. (2024, March). Sarid: Arabic Storyteller Using a Fine-Tuned LLM and Text-to-Image Generation. In 2024 16th International Conference on Computer and Automation Engineering (ICCAE) (pp. 1-5). IEEE.
16. Alahmadi, M. D., Alharbi, M., Tayeb, A., & Alshangiti, M. (2024). Evaluating large language models' proficiency in answering Arabic GAT exam questions. Engineering, Technology & Applied Science Research, 14(6), 17774–17780
17. Albalawi, Y., Nikolov, N. S., & Buckley, J. (2022). Pretrained transformer language models versus pretrained word embeddings for the detection of accurate health information on Arabic social media: Comparative study. JMIR Formative Research, 6(6), e34834.



18. Albared, D., Hamoud, H., & Zaraket, F. (2023, December). Arabic topic classification in the generative and AutoML era. In Proceedings of ArabicNLP 2023 (pp. 399-404).
19. Alderazi, F., Algosaibi, A., Alabdullatif, M., Ahmad, H. F., Qamar, A. M., & Albarrak, A. (2024). Generative artificial intelligence in topic-sentiment classification for Arabic text: a comparative study with possible future directions. PeerJ Computer Science, 10, e2081.
20. Alhayan, F., & Himdi, H. (2024). Ensemble learning approach for distinguishing human and computer-generated Arabic reviews. PeerJ Computer Science, 10, e2345.
21. Alhussain, A., & Azmi, A. M. (2024). Beyond event-centric narratives: Advancing Arabic story generation with large language models and beam search. Mathematics, 12(10), 1548.
22. Ali, A. R., Siddiqui, M. A., Algunaibet, R., & Ali, H. R. (2023). A large and diverse Arabic corpus for language modeling. Procedia Computer Science, 225, 12–21.
23. Ali, S. B., Kechaou, Z., & Wali, A. (2022, December). Arabic fake news detection in social media Based on AraBERT. In 2022 IEEE 21st International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (pp. 214-220). IEEE.
24. Aljamel, A., Khalil, H., & Aburawi, Y. (2024). Comparative study of fine tuned BERT-based models and RNN-based models. case study: Arabic fake news detection. The International Journal of Engineering & Information Technology (IJEIT), 12(1), 56-64.
25. Alkaoud, M. (2024). A bilingual benchmark for evaluating large language models. PeerJ Computer Science, 10, e1893.
26. Almasri, M., Al-Malki, N., & Alotaibi, R. (2023). A semi supervised approach to Arabic aspect category detection using Bert and teacher-student model. PeerJ Computer Science, 9, e1425.
27. Almazrouei, E., Cojocaru, R., Baldo, M., Malartic, Q., Alobeidli, H., Mazzotta, D., ... & Noune, B. (2023, December). AlGhafa evaluation benchmark for Arabic language models. In Proceedings of ArabicNLP 2023 (pp. 244-275).

28. Alnefaie, S., Atwell, E., & Alsalka, M. A. (2023, October). Is gpt-4 a good islamic expert for answering quran questions?. In Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023) (pp. 124-133).
29. Alrajhi, W., Al-Khalifa, H. S., & Al-Salman, A. S. (2024). Quantifying gender bias in arabic pre-trained language models. IEEE Access, 12, 77406-77420.
30. Alshammari, H., & Elleithy, K. (2024). Toward robust Arabic AI-generated text detection: Tackling diacritics challenges. Information, 15(7), 419.
31. Alshammari, H., El-Sayed, A., & Elleithy, K. (2024). Ai-generated text detector for arabic language using encoder-based transformer architecture. Big Data and Cognitive Computing, 8(3), 32.
32. Alshammari, W. T., & AlHumoud, S. (2022). TAQS: an Arabic question similarity system using transfer learning of BERT with BiLSTM. IEEE Access, 10, 91509-91523.
33. Alsubhi, K., Jamal, A., & Althothali, A. (2022). Deep learning-based approach for Arabic open domain question answering. PeerJ Computer Science, 8, e952.
34. Alyafeai, Z., Alshaibani, M. S., AlKhamissi, B., Luqman, H., Alareqi, E., & Fadel, A. (2023). Taqyim: Evaluating arabic nlp tasks using chatgpt models. arXiv preprint arXiv:2306.16322.
35. Alyafeai, Z., Pieler, M., Teufel, H., Tow, J., Bellagente, M., Phung, D., ... & Riquelme, C. (2024). Arabic Stable LM: Adapting stable lm 2 1.6 b to arabic. arXiv preprint arXiv:2412.04277.
36. Ammar, A., Koubaa, A., Benjdira, B., Nacar, O., & Sibace, S. (2024). Prediction of arabic legal rulings using large language models. Electronics, 13(4), 764.
37. Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., ... & Khan, H. (2024). Allam: Large language models for arabic and english. arXiv preprint arXiv:2407.15390.
38. Bassem, B., Abraham, A., & Mounir, Z. (2024). Transformer Models for Authorship Profiling in Arabic Social Media Texts. International Journal of Computer Information Systems and Industrial Management Applications, 16(3), 13-13.



39. Biswas, M. R., Mohsen, F., Shah, Z., & Zaghouani, W. (2023, November). Potentials of chatgpt for annotating vaccine related tweets. In 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-6). IEEE.
40. Biswas, M. R., Shah, Z., & Zaghouani, W. (2024). MemeMind at ArAIEval Shared Task: spotting persuasive spans in arabic text with persuasion techniques identification. arXiv preprint arXiv:2408.04540.
41. Boughorbel, S., & Hawasly, M. (2023). Analyzing multilingual competency of LLMs in multi-turn instruction following: A case study of Arabic. arXiv preprint arXiv:2310.14819.
42. Chouikhi, H., Aloui, M., Hammou, C. B., Chaabane, G., Kchaou, H., & Dhaouadi, C. (2024). GemmAr: Enhancing LLMs Through Arabic Instruction-Tuning. arXiv preprint arXiv:2407.02147.
43. Einieh, Y., Almansour, A., & Jamal, A. (2022, December). Fine tuning an AraT5 transformer for Arabic abstractive summarization. In 2022 14th international conference on computational intelligence and communication networks (CICN) (pp. 194-198). IEEE.
44. El-Sheikh, A., Elmogtaba, A., Darwish, K., Elmallah, M., Elneima, A., & Sawaf, H. (2024). Creating Arabic LLM prompts at scale. arXiv preprint arXiv:2408.05882.
45. Elchafei, P., & Fashwan, A. (2024). Arabic NER Evaluation: Pre-Trained Models via Contrastive Learning vs. LLM Few-Shot Prompting. Procedia Computer Science, 244, 229-237.
46. Fadel, A. S., Abulnaja, O. A., & Saleh, M. E. (2023). Multi-Task Learning Model with Data Augmentation for Arabic Aspect-Based Sentiment Analysis. Computers, Materials & Continua, 75(2).
47. Fang, Y., Xu, C., Guan, S., Yan, N., & Mei, Y. (2024, November). Advancing Arabic Sentiment Analysis: ArSen Benchmark and the Improved Fuzzy Deep Hybrid Network. In Proceedings of the 28th Conference on Computational Natural Language Learning (pp. 507-516).

48. Gupta, A., Singhal, A., Law, T., Rao, V., Duan, E., & Li, R. L. (2025, January). Can LLMs Verify Arabic Claims? Evaluating the Arabic Fact-Checking Abilities of Multilingual LLMs. In Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script (pp. 104-113).
49. Hajj, A., Adla, Y. A. A., Albast, S., Hajj, H., Elbassuoni, S., El-Hajj, W., & Shaban, K. (2024, October). Design and Comparison of Arabic Negotiation Bots Using LLMs versus Seq2Seq Models with Reinforcement Learning. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024) (pp. 304-318).
50. Hasanain, M., Hasan, M. A., Ahmed, F., Suwaileh, R., Biswas, M. R., Zaghouani, W., & Alam, F. (2024). ArAIEval shared task: propagandistic techniques detection in unimodal and multimodal Arabic content. arXiv preprint arXiv:2407.04247.
51. Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., ... & Xu, J. (2023). Acegpt, localizing large language models in arabic. arXiv preprint arXiv:2309.12053.
52. Karajeh, O., Al-Kabi, M. N., & Fox, E. A. (2023, December). Fusing AraBERT and graph neural networks for enhanced Arabic text classification. In 2023 24th International Arab Conference on Information Technology (ACIT) (pp. 1-8). IEEE.
53. Khaled, S., Mohamed, E. H., & Medhat, W. (2024). Evaluating large language models for Arabic sentiment analysis: A comparative study using retrieval-augmented generation. Procedia Computer Science, 244, 363-370.
54. Khondaker, M. T. I., Naeem, N., Khan, F., Elmadany, A., & Abdul-Mageed, M. (2024, August). Benchmarking llama-3 on arabic language generation tasks. In Proceedings of The Second Arabic Natural Language Processing Conference (pp. 283-297).
55. Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. arXiv preprint arXiv:2305.14976.
56. Koubaa, A., Ammar, A., Ghouti, L., Najar, O., & Sibae, S. (2024). Arabiangpt: Native arabic gpt-based large language model. arXiv preprint arXiv:2402.15313.



57. Kwon, S. Y., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. In Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023) (pp. 101–119). Association for Computational Linguistics.
58. Liang, J., Cai, Z., Zhu, J., Huang, H., Zong, K., An, B., ... & Xu, J. (2024). Alignment at pre-training! towards native alignment for arabic LLMs. Advances in Neural Information Processing Systems, 37, 13872-13896.
59. Lotfy, A., Saleh, K., Mohamed, S., Lorange, J., Yehia, E., Mohammed, K., ... & Yasser, T. (2024, March). Sentiment analysis for arabic product reviews using llms and knowledge graphs. In 2024 6th International Conference on Computing and Informatics (ICCI) (pp. 411-417). IEEE.
60. Mahdhaoui, H., Mars, A., & Zrigui, M. (2023, September). Optimizing Arabic named entity recognition through active learning and AraBERT. In 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA) (pp. 1-5). IEEE.
61. Mahdi, A. F., Farhan, R. N., & Al-Rawi, S. S. (2022, August). Improvement of semantic-BERT model with bidirectional LSTM for Arabic language. In 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM) (pp. 202-207). IEEE.
62. Mohammad, R., Alkhnbashi, O. S., & Hammoudeh, M. (2024). Optimizing large language models for Arabic healthcare communication: A focus on patient-centered NLP applications. Big Data and Cognitive Computing, 8(11), 157
63. Mohsen, M. (2024). Artificial intelligence in academic translation: a comparative study of large language models and google translate. Psycholinguistics, 35(2), 134-156.
64. Nagi, K. A., Alzain, E., & Naji, E. (2024). Informed Prompts and Improving ChatGPT English to Arabic Translation. Al-Andalus Journal for Humanities & Social Sciences, 11(98).
65. Nakhleh, S., Qasaimeh, M., & Qasaimeh, A. (2024, August). Character-level adversarial attacks evaluation for arabert's. In 2024 15th international conference on information and communication systems (ICICS) (pp. 1-6). IEEE.

66. Nassif, A. B., Elnagar, A., Elgendy, O., & Afadar, Y. (2022). Arabic fake news detection based on deep contextualized embedding models. *Neural Computing and Applications*, 34(18), 16019-16032.
67. Nehar, A., Bellaouar, S., Souffi, S., & Bouameur, M. (2023, October). Dhati: a Fine-tuned Large Language Model for evaluating Subjectivity in Arabic Textual Data. In *2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS)* (pp. 1-7). IEEE.
68. Nfaoui, E. H., & Elfaik, H. (2024). Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(2), 1118–1141.
69. Obeidat, R., Bashayreh, A., & Younis, L. B. (2022, June). The impact of combining arabic sarcasm detection datasets on the performance of bert-based model. In *2022 13th International Conference on Information and Communication Systems (ICICS)* (pp. 22-29). IEEE.
70. Qarah, F. (2024). AraPoemBERT: A pretrained language model for Arabic poetry analysis. *arXiv preprint arXiv:2403.12392*.
71. Qian, Z., Altam, F., Alqurishi, M., & Souissi, R. (2024). Camelevel: Advancing culturally aligned arabic language models and benchmarks. *arXiv preprint arXiv:2409.12623*.
72. Rasheed, A. F., Zarkoosh, M., Abbas, S. F., & Al-Azzawi, S. S. (2023, December). Arabic offensive language classification: Leveraging transformer, lstm, and svm. In *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)* (pp. 1-6). IEEE.
73. Sadek, A., Khalil, M. I., & Salama, C. (2023, October). Detection of Cyberbullying in Arabic using Machine Learning and ChatGPT. In *2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (pp. 242-245). IEEE.
74. Salem, A., Medhat, W., & GamalElDin, N. (2024, October). Exploring State-of-the-Art Models in Arabic NLP: Insights into Multi-Label Text Classification. In *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)* (pp. 533-538). IEEE.



75. Sallam, M., Al-Mahzoum, K., Alshuaib, O., Alhajri, H., Alotaibi, F., Alkhurainej, D., ... & Egger, J. (2024). Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. BMC Infectious Diseases, 24(1), 799.
76. Samaan, J. S., Yeo, Y. H., Ng, W. H., Ting, P. S., Trivedi, H., Vipani, A., ... & Ayoub, W. S. (2023). ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab Journal of Gastroenterology, 24(3), 145-148.
77. Sayah, B., Nehar, A., Cherroun, H., & Bellaouar, S. (2023, December). AraBERT and mBert: Insights from Psycholinguistic Diagnostics. In Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023) (pp. 159-167).
78. Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., ... & Xing, E. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149.
79. Tamer, A., Hassan, A. A., Ali, A., Salah, N., & Medhat, W. (2023, December). Fine Tuning of large language Models for Arabic Language. In 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA) (pp. 1-4). IEEE.
80. Wafa'Q, A. J., Mustafa, A. M., & Ali, M. Z. (2022, June). Sarcasm detection in Arabic short text using deep learning. In 2022 13th International Conference on Information and Communication Systems (ICICS) (pp. 362-366). IEEE.
81. Zakraoui, J., AlJa'am, J. M., & Salah, I. (2022, December). Domain-specific text generation for Arabic text summarization. In 2022 International Conference on Computer and Applications (ICCA) (pp. 1-4). IEEE.
82. Zhu, J., Huang, H., Lin, Z., Liang, J., Tang, Z., Almubarak, K., ... & Xu, J. (2024). Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion. arXiv preprint arXiv:2412.12310.

استخدام منهجية لسانيات المدونات الحاسوبية في تحليل اللغة الروائية دراسة تطبيقية على الرواية السعودية

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

د. طارق بن ربيع خلف الفريدي

قسم اللغويات، كلية اللغة العربية والدراسات الإنسانية، الجامعة الإسلامية بالمدينة المنورة، السعودية

نُشِرَ في	قُبِلَ في	رُجِعَ في	استقبل في
١٤٤٧/٠٥/٠٩	١٤٤٧/٠٤/٢٨	١٤٤٧/٠٣/٢٢	١٤٤٧/٠٣/١٧

الملخص

يسعى هذا البحث إلى تجسير العلاقة بين لسانيات المدونات الحاسوبية والدراسات الأدبية العربية السردية المتمثلة في الرواية بوصفها النمط السردى الأكثر شهرة في العصر الحاضر، ويأتي هذا التجسير في صورتين: الأولى: جاءت على هيئة قراءة للجهود الغربية في هذا المجال والوقوف على طرائقهم في تطبيق منهجية لسانيات المدونات في تحليل النصوص الروائية والكشف عن أبرز النتائج التي توصلوا إليها. وجاءت الصورة الثانية على هيئة نموذج تطبيقي أجريته فيه تحليلاً لمدونة الروايات السعودية بغية الكشف عن عدد من الأنماط التي تصف سلوك اللغة المستعملة في الخطاب الروائي السعودي. ومن خلال تطبيق المنهج الاستكشافي والوظائف المزودة في أداة معالجة اللغة العربية (غواص)، فقد توصلت في هذا البحث إلى عدد من النتائج التي تكشف عن الكلمات والمتابعات الأكثر شيوعاً في المدونة وما يتعلق بها من أنماط تركيبية ودلالية وسياقية، كما عرضت لكلمة المحتوى الأكثر شيوعاً في المدونة بالتحليل المفصل وهي كلمة "نفس" ورصدت أنماطها الدلالية ومتصاحباتها اللفظية للكشف عن سلوكها اللغوي في الرواية السعودية.

الكلمات المفتاحية

المدونات اللغوية، الرواية السعودية، الأسلوبية، قوائم الشبوع، المتابعات اللفظية.

The Application of Corpus Linguistics to the Analysis of Narrative Language: A Case Study of Saudi Novels

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

Dr. Tareq Rabee' Khalaf Al-Furaidei

Department of Linguistics, College of Arabic Language and Humanities,
Islamic University of Madinah, Saudi Arabia

Received on	Revised on	Accepted on	Published on
09-09-2025	14-09-2025	20-10-2025	31-10-2025

ABSTRACT

This research aims to bridge the gap between computational corpus linguistics and literary, narrative studies, specifically focusing on the novel as the most prominent narrative form. This aim was accomplished in two ways: The first involved a review of Western efforts in this field, examining their methodologies for applying corpus linguistics to the analysis of narrative texts and identifying their key findings. The second involved conducting an empirical analysis of a Saudi Novels Corpus to uncover various patterns related to the linguistics and stylistic behavior within Saudi narrative discourse. By employing this exploratory approach and utilizing the computational functions provided by the Arabic language processing tool (Ghawwas), this research identified several key findings. These results reveal the most frequent words and sequences of words in the corpus, along with their associated syntactic, semantic, and contextual patterns. The study also presented a detailed analysis of the most common content word in the corpus, which is the word “nafs”, observing its semantic patterns and lexical collocations to illuminate its linguistic behavior within Saudi novels.

KEYWORDS

Corpus Linguistics, Saudi novel, Stylistics, frequency lists, Sequence of words.

١- مقدمة

مع التقدّم الهائل والملاحظ في مجال الحاسب الآلي وتطبيقاته في معالجة اللغات الطبيعية والتعامل مع البيانات اللغوية الضخمة؛ ظهر حقل لسانيات المدونات بوصفه منهجيةً تتعامل مع النصوص بحثاً وجمعاً وتحليلاً؛ بغية الكشف عن الأنماط اللغوية في الخطاب، كشفاً مقرونًا بالأدلة الملموسة. وتعدّ هذه المنهجية من المنهجيات التجريبية الناجعة في تحليل أنماط اللغة في سياق استعمالها؛ كونها تتيح النظر إلى اللغة من عدة زوايا مختلفة (مثل: الكشف السياقي والتصاحب اللفظي) يصعب على الباحث إدراكها عندما يعتمد على الطريقة التقليدية في تحليل النصوص. ليس هذا فحسب، بل استثمرت هذه المنهجية في عدد من الحقول التي تتقاطع مع اللغة، مثل: تحليل الخطاب (Baker, 2006) ودراسات الترجمة واللسانيات التقابلية (Mikhailov and Cooper, 2016) واللسانيات الاجتماعية (Baker, 2010) واللسانيات الجنائية (Wright, 2021) وتعليم اللغات (Jones and Durrant, 2010). ويعدّ حقل الدراسات الأدبية من أحدث الحقول التي أفادت من هذه المنهجية في تحليل النصوص والكشف عن أنماط اللغة المستعملة فيها تركيبياً ودلالياً وسياقياً (Stubbs, 2005; Fischer-Starcke, 2010; Maiwald, 2011)، يأتي ذلك بعد أن كان الاعتماد الكلي على الحدس والتذوق المبنيين على ذاتية الباحث يشكل الإطار الأساس والسمة الأبرز في تحليل النصوص الأدبية وتفسيرها (Fischer-Starcke, 2010, p.32,35).

وفي سياق الدراسات العربية، فإنه على الرغم من ظهور الاهتمام بتطبيق منهجية لسانيات المدونات في حقول معرفية مختلفة، فإن الحقل الأدبي، وحقل الرواية على وجه الخصوص، ظلّ بمنأى عن تطبيق هذه المنهجية، فلا أعلم دراسةً عربية استثمرت لسانيات المدونات في تحليل النص الروائي، وقد سبق إلى الإشارة إلى هذه الفجوة المجهول (٢٠١٨، ص. ٧٤). فالمحوظ أن الدراسات التي تجعل من الرواية العربية مجالاً لها تنتهج الطريقة التقليدية التي تعتمد على قراءة النص أو النصوص المراد تحليلها سطرًا سطرًا، ثم تسجيل الملاحظات وتصنيفها يدوياً ثم تحليلها تحليلًا يعتمد على الحدس والخلفية المعرفية والخبرة الذاتية لدى الباحث، وهذه المنهجية على الرغم من أهميتها إلا أنها تفتقر إلى قدرتها على التعامل مع البيانات الضخمة التي تمثّل عدداً كبيراً من الروايات.

ولعل من أبرز أسباب غياب تطبيق منهجية لسانيات المدونات في تحليل النصوص الروائية العربية:

١. عدم وجود مدونة متخصصة في الرواية العربية توفر البيانات وتتيحها للبحث والتحليل والمعالجة الحاسوبية، حتى لو أردنا الاعتماد على المدونات العامة التي تشتمل على نصوص روائية فإنها لا تخلو من إشكالات تعيق إجراء دراسات على الروايات من خلال الاعتماد على لسانيات المدونات (Alfraidi et al., 2022)^(١)
٢. القصور في التعريف بأهمية هذه المنهجية وجدواها في تحليل النصوص الأدبية المتمثلة في الرواية، بل إن جُلّ الجهود اتجهت إلى التجسير بين لسانيات المدونات وصناعة المعاجم (القاسمي، ٢٠٠٦، الهرساني، ٢٠١٨) وتحليل النص القرآني (Abdul-Baqee and Atwell, 2009) وتحليل الخطاب الإخباري (الشمري والثبتي، ٢٠١٩؛ الشمري والمحمود، ٢٠١٦) وتعليم اللغة العربية (الفيضي، ٢٠١٦) ودراسات الترجمة (العتيبي، ٢٠١٦) واللغة الأكاديمية (العصيمي والثبتي، ٢٠١٨).

وهنا يأتي الدافع الأساس وراء كتابة هذا البحث الذي يسعى إلى الكشف عما يمكن أن تسهم فيه لسانيات المدونات الحاسوبية في مجال الدراسات الأدبية، خاصة في تحليل النص الروائي تحليلاً لغوياً. ولتحقيق هذا الهدف، سعت أولاً إلى مراجعة الأدبيات المكتوبة بالإنجليزية وعرضتها عرضاً يجلي كيفية الإفادة من لسانيات المدونات في تحليل لغة الرواية والكشف عن أنماطها، ثم سعت ثانياً إلى تقديم نموذج تطبيقي للإفادة من هذه المنهجية في تحليل الرواية السعودية مستعيناً بالمدونة السعودية للروايات السعودية التي تحتوي على نصوص روائية معالجة حاسوبياً (Alfraidi et al., 2022) ومستعيناً كذلك بالبرمجيات المتقدمة التي تتعامل مع النصوص بدقة عالية.

(١) مثلاً، بعض المدونات العربية العامة لا توضح نوع النص عندما تضع قسماً خاصاً بالأدب العربي في المدونة، فلا يعلم هل الكلمة أو الكشاف السياقي يمثل رواية أو قصة قصيرة أو مسرحية أو كتاباً نقدياً، وبعضها يعتمد مكان نشر الرواية مؤشراً للجنسية عند تقسيم الروايات حسب الدول، فلو كانت دار طباعة لبنانية وطُبعت رواية لكاتب سعودي فستظهر النتائج على أن الرواية لبنانية. وهناك إشكاليات أخرى ليس المقام تفصيلها. ينظر للاستزادة: (Alfraidi et a, 2022).

وهذا البحث هو بحث استكشافي، يهدف إلى رصد الأنماط اللغوية والدلالية التي تشكل اللغة في الرواية السعودية بصورة عامة، أسعى فيه إلى تقديم نموذج عملي يفيد الباحثين في مجال الدراسات السردية العربية من خلال الإفادة من منهجية لسانيات المدونات وأدواتها الحاسوبية التي تمكنهم من رصد ما لا يسهل رصده من الأنماط عند الاعتماد على الطريقة التقليدية.

وقد اقتضت طبيعة البحث إلى أن يُقسَّم إلى ثلاثة مباحث من دون اعتبار المقدمة والخاتمة، يأتي المبحث الأول معرِّفاً بلسانيات المدونات وخصائصها، ويختص المبحث الثاني بعرض طبيعة العلاقة بين لسانيات المدونات وحقل الدراسات السردية؛ إذ تتبعت فيه جهود الغربيين في هذا المجال واستخلصت أهم السمات التي نهضت بهذه العلاقة ودورها في تجسير العلاقة بين المدونات اللغوية والدراسات الأسلوبية، وجاء المبحث الثالث مخصصاً للدراسة التطبيقية لللسانيات المدونات الحاسوبية في تحليل الرواية السعودية، واشتمل هذا على وصف المدونة المدروسة وأدوات المعالجة والتحليل، وأبرز النتائج التي توصلنا إليها مع محاولة تفسيرها.

٢- لسانيات المدونات: المفهوم والخصائص

يعد مصطلح لسانيات المدونات ترجمة للمصطلح الإنجليزي "Corpus Linguistics". وهناك ترجمات أخرى أخذ بها بعض الباحثين، مثل: لسانيات المتون (العصيمي، ٢٠١٨)، علم/لسانيات الذخيرة النصية (الحاج صالح ١٩٩٩)، وبعضهم يضيف مفردة "الحاسوبية" فيقول "لسانيات/لغويات المدونة الحاسوبية" (المجيول ٢٠١٦)، وهي إشارة صريحة إلى الطبيعية الحاسوبية للمدونات والتي لا تنفك عنها. والذي يظهر أن لفظة "المدونات" هي الأكثر تداولاً وشيوعاً؛ لذا رأيت الأخذ به^(١).

(١) انظر ما ذكره صالح والعصيمي حول المصطلحات المستخدمة مقابل "Corpus Linguistics" وإشارتهم كذلك إلى غلبة لفظة "المدونات" من حيث الاستعمال على بقية المقابلات العربية. (صالح، ٢٠١٥، ص ٢٠؛ العصيمي، ٢٠١٨، ص ٢٢-٢٣).



وللتعريف بهذا المصطلح يلزمنا التفريق بين مستويين :

المستوى الأولي: مصطلح "المدونة اللغوية (الحاسوبية)" ، **والمستوى الثاني:** مصطلح "لسانيات المدونات". ففي المستوى الأول يشير المصطلح إلى مصدر البيانات المدروسة، بينما يشير في المستوى الثاني إلى المنهجية المستخدمة في تحليل بيانات المدونة.

أما مصطلح "المدونة اللغوية" فيدل على مجموعة من النصوص الإلكترونية التي تعكس الواقع اللغوي والقابلة للمعالجة الحاسوبية والتي جُمعت ضمن معايير محددة لتحقيق أغراض بحثية لغوية محددة (ينظر: الثبيتي، ٢٠١٥، ص. ١٥٠؛ العصيمي، ٢٠١٨، ص. ١٩). ومن أمثلتها: المدونة العربية التابعة لمدينة الملك عبدالعزيز للعلوم والتقنية (Al-Thubaity, 2015)، والمدونة العربية الدولية التابعة لمكتبة الإسكندرية (Alansari and Nagi, 2014). وتعد منصة فلك من أحدث المنصات التي توفر عدداً من المدونات العربية الفرعية^(١).

أما مصطلح "لسانيات المدونات" فيشير إلى تلك الممارسة المنهجية التي تدرس الظاهرة اللغوية من خلال النصوص المحوسبة في المدونات للكشف عن الأنماط التي تقف خلف استعمال تلك النصوص من خلال تكرار الكلمات والمتتابعات والكشاف السياقي والتصاحب اللفظي وغيره (ينظر: العصيمي، ٢٠١٨، ص. ٢٢)^(٢). ويُجز ذلك عبر استخدام برامج حاسوبية مزودة بعدد من الوظائف، مثل: استخراج قوائم الشيع، الكشاف السياقي، التصاحب اللفظي وغيرها. والغرض الأساسي من تطبيق هذه المنهجية هو رصد الأنماط (patterns) التي تتكشف لنا من خلال مخرجات البرامج الحاسوبية التي تملك الإمكانيات لتحقيق هذا الغرض. وعادة ما يُربط بين النمط والتكرار، فما يتكرر في اللغة ضمن سياقات معينة يعد نمطاً (Hunston, 2010, p.153). وسنلاحظ لاحقاً أن مرتكز الدراسات التي طبقت هذه المنهجية هو تتبع الأنماط التي تشكل النصوص على طريقة لغوية معينة.

(١) يمكن الاطلاع على منصة فلك عبر الرابط التالي: <https://falak.ksaa.gov.sa/>

(٢) فقد نقل عدداً من التعريفات والتي أفدت منها في صياغة التعريف أعلاه.

ويمتاز البحث القائم على منهجية لسانيات المدونات بعدد من الخصائص (Biber et al., 1998, p.4; Stefanowitsch, 2020, p.23-30):

١. أنه بحث تجريبي يعتمد على الاستعمال الفعلي للغة وتشكلها داخل النصوص، فاللغة المدروسة هي ما يمارسه الناس منطوقاً أو مكتوباً في حياتهم.
٢. أنه يمكن الباحث من التعامل مع البيانات الضخمة وصفاً وتحليلاً.
٣. أنه يتسم بإمكانية إطلاق الأحكام وتعميم النتائج عندما تكون المدونة ممثلة للمجال المدروس أو قريبة من التمثيل.
٤. أنه يمثل أداة فاعلة لاختبار الفرضيات وفحص الأحكام المسبقة المبنية على الحدس والتأمل الذاتي. يقول حسين (Hussein, 2015, p.75) "يمكن للتأمل الذاتي أن يكون مفيداً... ولكن من الضروري فحص [نتائجه] من خلال الاعتماد على أدلة قوية مستخرجة من بعض المدونات الممثلة".
٥. أنه يعتمد على تقنيات حاسوبية متقدمة وذات دقة عالية في التحليل.

وقد شاع تطبيق هذه المنهجية البحثية في ستينيات القرن العشرين، وكان التركيز في بداية تطبيقه على الحقل المعجمي، حين استُعين بالمدونات والبرامج الحاسوبية في بناء المعاجم الإنجليزية، وعقب ذلك استخدام المدونات في بناء كتب القواعد (صرفاً ونحواً) من خلال رصد الاستعمال الفعلي للغة وقياس مدى شيوعها، ثم تطورت هذه الممارسة المنهجية تدريجياً حتى أصبحت مجاًلاً من المجالات البحثية البينية نتيجة تطبيقها في حقول مختلفة تتقاطع مع اللغة (Biber, 2011, p.16). ويعد الحقل الأدبي من أحدث الحقول التي تناغمت مع لسانيات المدونات، وهذا ما سوف أسلط عليه الضوء في المبحث الآتي.

٣- لسانيات المدونات والدراسات الأدبية: جسور التواصل:

تشكل اللغة على صور متعددة وأجناس مختلفة من النصوص، وعلى الرغم من اشتراك هذه الأجناس في السمات العامة المنبثقة من اللغة الواحدة إلا أن لكل جنس خصائصه التي تميزه عن غيره، ويشار في الدراسات الغربية إلى تلك القوالب بمصطلح genre والذي



يمكن ترجمته بـ "جنس أو نوع النص". وتتمثل تلك الخصائص على هيئة أنماط معجمية وتركيبية ودلالية ونصية وتواصلية (Biber and Conrad, 2009, p.16; Fiarclough, 2003, p.67). وتنوع تلك الأجناس وتعدد، مثل: النص الإعلامي والنص الديني والنص المسرحي والنص الروائي والنص الأكاديمي، وقد أفضى هذا التفريق بين أنواع النصوص إلى ظهور حقل ما يعرف بـ genre analysis، وهو يشير إلى الدراسات التي تهدف إلى تحليل كل جنس في سياق معين ورصد أنماطه وتفسيرها وإجراء مقارنات فيما بينها، ففي كل مجتمع - حسب فيركولف - يُتعارف على استعمال اللغة في جنس أدبي محدد بطريقة أو طرائق معينة (Fairclough, 2003, p.67)، وتقف خلف هذه الطرائق أغراض تواصلية مرتبطة بالسياق الثقافي والاجتماعي للنص. فاللغة الصحفية طرائقها في مخاطبة جمهور القراء، فتسلك مسالك معينة لإيصال الفكرة أو إقناع القارئ بتلك الفكرة، مثل: استخدام الفعل المبني للمعلوم في سياقات، والفعل المبني للمجهول في سياقات أخرى؛ لتحقيق مآرب تواصلية معينة. ويمكن قول ذلك على بقية أجناس النصوص.

وتعد اللغة السردية متمثلة في الرواية ذات خصائص تميزها عن بعض أجناس النصوص (كاللغة الصحفية والأكاديمية)، مثل: احتضانها للغة الإبداعية والأحداث المتخيلة ووجود شخصيات تشارك في عرض تلك الأحداث، والاستناد إلى تقنيات الحوار وتوظيفها توظيفاً معيناً وعرض الأحداث وتسلسلها بطريقة معينة، وكل من ذلك يتطلب أدوات لغوية محددة لعرضه على هيئة نص (Siepmann, 2015, p.365)^(١) فاللغة في السرد الروائي هي التي تصوّر الحدث وتجسد الحركة والزمن وتصور الأماكن وتصف الأشياء وتدمج ذلك كله في وحدة واحدة تؤدي الغرض من الرواية (مرتاض، ١٩٩٧، ص. ٤٩٠). ويقول بايبر وكونراد (Biber and Conrad, 2009, p.138): "إن الغرض التواصلية من سرد قصة ما ينعكس على خصائص نوع النص السردية وعلى طريقة عرض الخطاب المرتكز على الخيال، مثل: [استخدام] الأفعال الماضية وضمان الغائب والظروف الزمانية وأفعال القول". ويشير الحازمي (٢٠٠٦، ص. ٤٩٤-٥٢٥) إلى أن اللغة التي تحتضنها الرواية ليست على نمط واحد،

(١) ينظر في عناصر البناء السردية للرواية في: مرتاض (١٩٩٧، ص. ٢٤)، وفي آليات الحوار في الرواية ووظائفه في: قسومة (٢٠٢٢، ص. ٦١-٩٢).

بل تتشكّل على هيئة أنماط مختلفة، فالنمط التقريري يتسم بالمباشرة في سرد الأحداث وتقرير وقوعها؛ لذا تسيطر عليه الأفعال الماضية وأدوات تقرير الحدث، (مثل: قد)، بينما يسعى النمط التصوري -الذي تعتمد عليه اللغة السردية اعتماداً كبيراً- إلى تجسيد الحركة وتصويرها في عرض مشهد حي؛ لذا تكثر فيه الأفعال الماضية التي تصطبغ بصبغة الحاضر، في حين يعبر النمط التعبيري عما في النفس من مشاعر وأحاسيس وعواطف. وكل نمط يستدعي المفردات والعبارات التي تحقق مبتغاه، ويؤكد الحازمي (٢٠٠٦، ص. ٥٥٢) أن الرواية السعودية تمتاز بتنوع أنماط اللغة السردية.

ويشار إلى المجال الذي يدرس الخصائص اللغوية للنص الأدبي ويجسّر العلاقة بينها وبين المعنى بالأسلوبية (Stylistics). وهو منهج يستهدف البحث في النصوص الأدبية ونقدها من خلال نافذة التحليل اللغوي. (Leech and Short, 1982, p.11 ; Fischer-Starcke, 2010, p.2). ويعد الدمج بين التحليل الأسلوبي ولسانيات المدونات ممارسة منهجية حديثة نسبياً، ويُطلق على هذه الممارسة بـ "Corpus Stylistics".

يقول بايبر (Biber, 2011, p.16) في هذا السياق: "على مدى العقود العديدة الماضية، أنجز الكم الهائل من الدراسات الأسلوبية للغة الأدبية التي تنطوي على خصائص التحليل المدوني باستخدام أساليب التحليل الحاسوبي/الكمي للكلمات في النصوص الأدبية. ولكنه في السنوات الأخيرة فقط تم دمج مثل هذه الدراسات بشكل صريح مع لسانيات المدونات". وقد أكدت ماهلبيرق هذا بكلام مشابه (Mahlberg, 2007; p.2; 2012, p.77). وهذا يعني أن استخدام أدوات الحاسب الآلي وتوظيف تقنياته في التعامل مع النصوص الأدبية معالجة وتحليلاً يعد ممارسة منهجية قديمة مشابهة لمنهجية لسانيات المدونات، وقد أرجعها بايبر إلى تاريخ ١٩٦٤م (Biber 2011, p.20)، ولكن لم يكن يُشير إلى كون ذلك يمثل حقلاً من حقول لسانيات المدونات ومجالاً ذا خصائص منهجية معينة في الدراسات الأسلوبية حتى صُرّح بذلك مؤخراً نتيجةً لتطبيق أدوات لسانيات المدونات (مثل: قوائم الشيوع والكشاف السياقي والتصاحب اللفظي) في تحليل النص الأدبي (خاصة النص الروائي) بطريقة منهجية سعياً إلى رصد الأنماط التي تقف خلف تشكّله بتلك الصورة.



وجاء هذا الدمج بين التحليل الأسلوبي للنصوص الأدبية ولسانيات المدونات نتيجة لطبيعة المجالين المرتكزة على التجريبية التي تعتمد على التعاطي مع اللغة في الاستعمال وتحليل سياقاتها (Wynne, 2006:1)، هذا من جهة، ومن جهة أخرى، ما تحظى به لسانيات المدونات من خاصية الاستفادة من التقنيات الحاسوبية وإمكاناتها العالية والدقيقة في رصد الأنماط اللغوية في النصوص المدروسة والتي يصعب رصدها دون الاستعانة بتلك التقنيات (Mahlberg, 2014, p.387; Mahlberg et al. 2015, p.345). وقد قويت أواصر الدمج بشكل خاص بعد الانتقادات التي وُجّهت إلى الدراسات الأسلوبية كونها تعتمد على الحدس والانطباعات والأحكام الذاتية (subjectivity) والتي تعتمد غالباً على نماذج انتقائية للنصوص، فجاءت مشاركة لسانيات المدونات للتخفيف من وطأة هذه السلطة الذاتية التي تتلبّس ذلك النوع من الدراسات، وذلك من خلال سلوك آليات تجعل الأحكام أقرب إلى الموضوعية (Fischer-Starcke, 2010, p.16-19,167; Wijitsopon 2013, p.52). ومن صور الموضوعية في لسانيات المدونات استخراج بيانات النصوص المحوسبة، وقيام الأدوات الحاسوبية بتوليد قوائم الشيوخ والكشافات السياقية والمتصاحبات اللفظية وعرضها كما تتحقق في النصوص، وإجراء الإحصاءات اللازمة، وتحليلها تحليلاً قابلاً للاستعادة والاختبار (الشمري والمحمود، ٢٠١٦، ص. ٢٠٨). ويشير مفهوم القابلية للاستعادة (replicability) إلى إمكانية الحصول على النتائج نفسها عند تكرار نفس طرق آليات التحليل المدوني، بينما يشير مفهوم القابلية للاختبار (testability) إلى قدرة منهجية لسانيات المدونات على اختبار الفرضيات بصورة تجريبية تعتمد على الأدلة الملموسة؛ فينتج عن ذلك دعم تلك الفرضيات أو دحضها (McEnery and Hardie, 2012, p.14-16; Stefanowitsch, 2020, p.134-135) وعلى الرغم من ذلك فإن البحث القائم على المدونات لا يمكن وصفه ووصف نتائجه بالموضوعية المطلقة. تقول فيشر ستاركي (Fischer-Starcke, 2010, p.19) بعد أن ناقشت هذه النقطة نقاشاً مطولاً: "تُظهر المناقشة أعلاه أنه لا يمكن تحقيق تحليل موضوعي مطلق يستند إلى أدوات لسانيات المدونات". بل إن الموضوعية والذاتية مكملان لبعضهما في لسانيات المدونات، فما يستخرجه الحاسب من نتاج بطريقتهم موضوعية تعقبه قراءة وتفسير مبنيان على الحدس وذوق الباحث وخبرته المعرفية حتى تكتمل صورة التحليل للنص الأدبي؛ لذا تقول كذلك (Fischer-Starcke, 2010, p.18) في هذا السياق "[إن]

العناصر الذاتية والموضوعية مترابطة في منهجية لسانيات المدونات، ويؤثر كلٌّ منهما في الآخر ولا ينفصلان". وكذلك تؤكد ما هلبيرق (Mahlberg, 2012, p.77) على أن التحليل الأسلوبي الذي يستعين بلسانيات المدونات يعدّ منهجاً موضوعياً نسبياً توجّه نتائجه بتفسيرات ذاتية نسبياً.

وقد اهتمت الدراسات الأجنبية بتحليل الروايات من خلال عدسة لسانيات المدونات وأفادت منها كثيراً، ونتج عن هذا كمٌّ هائل من الدراسات المتنوعة، منها ما استهدف رواية معينة ومنها ما استهدف روايات كاتب معين، ومنها ما استهدف روايات بيئة جغرافية معينة أو حقبة زمنية معينة، ومنها ما اتجه إلى المقارنة بين روايات كُتّاب معينين، وفي الأسطر التالية عَرَضُ لبعض منها؛ لنبرز الأثر المهم الذي أحدثته لسانيات المدونات في مجال الدراسات السردية:

درس ستبز (Stubbs, 2005) رواية الكاتب البولندي-الإنجليزي Joseph Conrad المعنونة بـ Heart of Darkness وحللها مطبقاً منهجية لسانيات المدونات. وقد دفعه إلى تطبيق هذه المنهجية أن هذه الرواية لم تحظ بتحليل أسلوبي دقيق يكشف عن خصائصها اللغوية وأنماطها الدلالية، وقد أراد أن يثبت أن هذه المنهجية مثمرة للاتجاه الأسلوبي في تحليل النصوص الأدبية، خاصة وأن هذا الاتجاه لم يكن يحظى بثقة اللغويين ودارسي الأدب. وقد أسهمت هذه المنهجية في الكشف عن أنماط لم تكن لتُرى فيما لو اعتمد على الطرق التقليدية وإمكاناتها في التحليل، مثل: قراءة الرواية كاملة وتدوين الملاحظات والاعتماد على الحدس الذاتي في رصد النتائج. من الناحية التطبيقية، تتبّع ستبز شيوع عدد من الكلمات والعبارات التي تحتويها الرواية وحلّل الكلمات المميزة لها من خلال مقارنتها بمدونات أخرى، ورصد كذلك أنماط التصاحب اللفظي لعدد من الكلمات. ومما توصل إليه أن الفعل say (قال) ومشتقاته يعدّ الأكثر شيوعاً مقارنة ببقية الأفعال، كما لاحظ أن الأفعال التي تدل على الإدراك (مثل: know) تتكرر تكراراً بارزاً في الرواية. وفيما يتعلق بالتصاحب اللفظي، فقد رصد الكلمات التي غالباً ما تصاحب كلمة grass (عشب)، وهي كلمات تدل على الفناء death (موت) و desolation (دمار). وقد مكّنت لسانيات المدونات ستبز من دحض فرضية أن التكرار يشكل سمة أسلوبية غالبية في هذه الرواية ومميزة لها عن غيرها (مثل:



تكرار الألفاظ العاطفية)، وهي فرضية تبناها بعض النقاد السابقين؛ مؤكداً أن أسلوبية Conrad تتسق مع الإطار العام للكتابة الروائية من دون تمييز. وقد ختم دراسته بأن لسانيات المدونات تزود الباحثين بعدة أدوات منهجية لتحليل النصوص الأدبية، وينتج عن هذا التوصل إلى رؤى نحو تلك النصوص لم تكن لتُدرك لولا تطبيق هذه المنهجية، مما يمكن الباحث من مناقشة النقد حول أحكامهم التي تعتمد على القراءة التقليدية للنصوص واستعراضها يدوياً.

قدّمت فيشر-ستاركي (Fischer-Starcke 2010) دراسة مطولة حول أسلوبية الروائية الإنجليزية جين أوستن Jane Austen، وقد اتخذت من روايتها Northern Abbey مجالاً للتطبيق، ومن خلال مقارنة هذا الرواية مع روايات أخرى لأوستن وغيرها وتحليلها من منظور لسانيات المدونات، وقد تمكنت فيشر-ستاركي من استخراج الكلمات والعبارات المميزة للرواية وحللت الكشافات السياقية لها، فنتج عن هذا أن وقفت على عدد من الأنماط التي تنطوي عليها لغة أوستن في الرواية المدروسة، من ذلك ما توصلت إليه من تمييز الكلمات التي تدل على العواطف. بالإضافة، فقد ظهر تميّز كلمتي heroine (وتعني البطل) وnovel (وتعني الرواية)، وقد فسرت ظهور كلمة novel ضمن قائمة الكلمات المميزة بأنه خلال الفترة التي عاشتها أوستن لم تكن قراءة الروايات من الأنشطة الشائعة بين الناس، واللانقطة بذوي الطبقة الوسطى في المجتمع، فقد كان الاهتمام يتجه نحو قراءة الكتب العلمية؛ لذا سعت أوستن إلى التركيز على ظاهرة قراءة الروايات وتصويرها تصويراً إيجابياً؛ سعياً منها إلى تصحيح هذا المفهوم. وهذه النتائج لم تكن لتُرى لولا الأدوات التي يتيحها تطبيق منهجية لسانيات المدونات، مثل: استخراج الكلمات المميزة، وقراءة الكشافات السياقية لها.

وفي دراسة مشابهة، درست ويجتسوبن (Wijitsopon, 2013) ستاً من روايات أوستن دراسة مدونية، وقد هدفت إلى الكشف عن الأنماط المعجمية والنصية التي تميز أسلوبية أوستن عن غيرها، مؤكداً على أن تطبيق منهجية لسانيات المدونات تساهم في التحقق من فرضيات النقد وأحكامهم حول كتابة أوستن من منظور مقاربي تستند إلى الأدلة الكمية والكيفية. وعند قراءة الجانب التطبيقي لدراستها نجد -وبشكل بارز- أن هذه المنهجية

قد مكّنتها من رصد عدد من الأنماط، مثل: الحقول الدلالية التي تنتمي إليها طائفة من الكلمات الشائعة والمميزة لروايات أوستن (مثل: الكلمات التي تشير إلى المرأة والعلاقات الأسرية، وتلك التي تشير إلى المشاعر، وضميري الغائب للمفرد المذكر والمؤنث). ونجدها تقرر أن تطبيقها لهذه المنهجية أسهم في الكشف عن أنماط لم تكن تُرصد لولا هذه المنهجية، فقد لاحظت -مثلاً- أن تلك الكلمات التي تدل على مقدار درجة حدوث الفعل أو درجة الصفة المتلبس بها (مثل: very engaging) أو تلك التي تدل على بلوغ درجة عالية من التلبس بالصفة (مثل: entirely occupied) تشكّل نمطاً أسلوبياً مميزاً لروايات أوستن، مستنتجةً أن التعبير عن المبالغة (exaggeration) يعدّ أحد خصائص لغتها. ومن النتائج التي توصلت إليها ما يتعلق بتصاحب الكلمات، من ذلك ما وقفت عليه من تصاحب المتتابعة "could not" مع أفعال الإدراك والقول للتعبير عن الصعوبة التي تواجه الشخصيات النسائية في فهم أو التعبير عن بعض القضايا المعينة.

كما أجرى ميفلد (Maiwald, 2011) دراسة استكشافية قائمة على المدونات لروايات الروائي والشاعر الإنجليزي Goerge MacDonald. وقد رصد الكلمات الأكثر شيوعاً في المدونة المدروسة، ملاحظاً أن الأفعال تشكل الفئة الأغلب في قائمة الكلمات الخمس والعشرين الأكثر شيوعاً. وقد تصدر الفعل الدال على القول (Say) القائمة مما يدل على الحضور المكثف للحوار المباشر بين شخصيات الروايات. كما رصد الهيمنة البارزة للشخصيات الذكورية وذلك من خلال مقارنة الكلمات التي تحيل إلى الذكر (مثل: man, father, he) بتلك التي تحيل إلى الأنثى (مثل: woman, mother, she) مقارنة إحصائية. كما لاحظ ظهور كلمتي (God) و (Jesus) في قائمة الكلمات المميزة للغة هذا الكاتب مع تصدر واضح لكلمة (God)، مستنبطاً أن الحضور الديني متمثلاً في الثقافة النصرانية وما يرتبط بها مهيمناً على بقية الموضوعات (القيمات) في روايات هذا الكاتب، خاصة في الروايات الواقعية (realistic novels). وختم ميفلد هذه الدراسة بتحليل الحقول الدلالية المميزة (semantic key domains) لأسلوبية MacDonald الروائية، ومما لاحظته هيمنة الكلمات التي تنتمي إلى المخلوقات الحية (مثل: الطيور والحيوانات) والأمكنة (مثل: الجبال والأنهار) والكون (مثل: القمر)، معللاً هذه النتيجة بكون مصب اهتمام الكاتب كان متجهاً -وبشكل بارز- نحو الطبيعة بوصفها مسرحاً لأحداث الروايات.



وقد حظي الروائي الإنجليزي تشارلز ديكنز Charles Dickens بنصيب وفير من التحليل المدوني، وأبرز من اهتم بذلك الباحثة الألمانية ميكايل ماھلبيرق والتي أنجزت عدداً من الدراسات في هذا المجال (Mahlberg et al 2019; Mahlberg et al. 2016; Mahlberg et al. 2013; Mahlberg, 2007a,b; وسبرت أغوار لغة ديكنز من خلال منهجية لسانيات المدونات. وتعد دراستها (Mahlberg, 2013) أوسع هذه الدراسات والتي ركزت فيها على رصد سلوك المتتابعات اللفظية (clusters) وأنماطها الدلالية (semantic patterns). فقد لاحظت -مثلاً- أن المتتابعات المميزة في لغة ديكنز تشمل على الإشارة إلى أعضاء الجسد، وبعضها تنطوي على الدلالة على الزمان أو المكان. ومن أبرز المتتابعات التي تشمل على عضو الجسد عبارة "his eyes on the ground" والتي تسبق كثيراً بفعلي الحدث "stand and walk"، وقد لاحظت أن هذه العبارة تشير إلى سلوك التأمل أحياناً والحزن أحياناً أخرى. كما رصدت شيوع المتتابعات اللفظية التي تبتدئ بـ "as if" (كما لو) والتي تعبر عن تشبيه حدث أو سلوك بأخر لم يقع أو لن يقع (افتراضي)، كما استطاعت رصد سلوك الكلمات التي تصاحب "as if" وتنميطها دلالياً، مثل مصاحبها لأفعال الحركة والحدث والأسماء الدالة على السلوك. ويمكن القول إنه قد كان يصعب على ماھلبيرق الوصول إلى مثل هذه النتائج لولا الاعتماد على الوظائف الحاسوبية التي تمتلكها لسانيات المدونات مثل: قوائم الشيوع والكشاف السياقي والتصاحب اللفظي. وختمت ماھلبيرق الدراسة بالتنصيص على أهمية الأدوات الحاسوبية في رصد أكبر قدر من الأنماط التي تكمن خلف استعمال اللغة في النص الروائي، وأشارت إلى أنه ممن الممكن كذلك تتبع تطور اللغة في النمط الكتابي لدى ديكنز من خلال مقارنة أعماله المبكرة بالتأخرة.

ومن الدراسات التي لم تخصص بروايات كاتب معين أو رواية معينة دراسة قرين (Green 2017) حول مدونة الأدب الغربي الرفيع "The Canon of Western Literature"، وهي مدونة مرجعية للأدب الإنجليزي الرفيع وتحتوي على أكثر من ٧٠ مليون كلمة، وتشكل الروايات جزءاً منها. ويشير هذا النوع من الأدب -حسب قرين- إلى تلك النصوص الأدبية التي تشتمل على عناصر جمالية ولغة أدبية رفيعة المستوى، وتمثل في الوقت نفسه عاملاً مؤثراً في الفكر والثقافة والنصوص الأدبية الأخرى، وقد اتخذ الباحث لسانيات المدونات منهجاً له في تحليل هذه المدونة الكبيرة. ويمثل ما أطلق عليه "culturomic and stylistics"

"إطاراً نظرياً للدراسة. ويعني بـ "culturomics" دراسة سلوك اللغة الإنسانية وسماتها الثقافية والاجتماعية وما يطرأ على ذلك من تغيرات من خلال النصوص الرقمية (Michel, 2011, p.176; Green, 2017, p.382). وقد توصل الباحث إلى عدد من النتائج التي تتعلق بالكلمات والمتابعات اللفظية المميزة، ومما لاحظته هيمنة الضمائر المستخدمة من الشخصيات بُغية التعبير عن المشاركة في سرد الأحداث. كما لاحظ أن الألفاظ والضمائر التي تحيل إلى الذكور أكثر من تلك التي تحيل إلى الإناث، مما يدل على التفريق الجندري في استعمال اللغة الأدبية، مستنتجاً أن هيمنة شخصيات الذكور ومشاركتهم في سرد الأحداث الروائية يعد علامة مميزة في المدونة المدروسة. كما لاحظ قرين الصبغة الدينية للنصوص الأدبية التي تمثل العصر الثيوقراطي Theoretic Age، فتشيع فيها الكلمات الدينية مثل: God (الرب)، soul (الروح)، lord (السيد) شيوعاً أكثر من العصور التي أتت بعده، مما يدل على التغير في استعمال لغة الأدب الرفيع خلال العصور الماضية حتى القرن العشرين، ويعكس هذا التغير جوانب من الثقافة التي كانت تبناها المجتمعات الغربية في تلك العصور وما صاحبها من تغيرات. وختم دراسته بأهمية دراسة هذا الأدب دراسة عميقة ترصد عدداً من الظواهر، مثل: الاختلافات بين الدول والأقاليم في استعمال اللغة في مثل هذا النوع من الأدب، ورصد الاختلافات بين المؤلفين الذكور والإناث من حيث المواضيع (القيمات) واختيار الشخصيات وطريقة السرد؛ إذ إن دراسته تعد تمهيدية لتقرير أهمية هذا النوع من الدراسات.

قام سيبمن (Siepmann, 2015) بدراسة استكشافية للروايات التي نُشرت بعد الحرب العالمية الأولى (post-war fiction) من خلال مدونة عدد كلماتها حوالي ١٦٠ مليون تمثل روايات كُتّاب من بقاع جغرافية مختلفة، وقد ركّز دراسته على استكشاف الكلمات المميزة للمدونة مقارنة بالمدونات التي تشتمل على نصوص حوارية وتلك الخاصة بالنصوص الأكاديمية. وقد هدف على وجه الخصوص إلى رصد الأنماط المعجمية والتركيبية والدلالية للكشف عن الكيفية التي أنتجت بها تلك النصوص من جهة، ولاختبار بعض الفرضيات التي تبناها عدد من نقاد أدب تلك الفترة من جهة أخرى. ومن النتائج التي توصل إليها الشيوخ البارز للألفاظ ذات الدلالة الوصفية عند سرد الأحداث، مثل: shake (يصفح)،



gaze (ينظر)، causally (بشكل اعتيادي). بيد أنه لم يقف عند هذا الحد من البحث، فقد تعمق أكثر لدراسة سياقات بعض الألفاظ، ومما لاحظته أن هناك ألفاظاً ذات دلالات خاصة باللغة الروائية، مثل: كلمة pause بمعنى التوقف لفترة قصيرة قبل التحرك، بينما تلك التي تكون بمعنى التوقف عن الكلام فهي ليست خاصة باللغة الروائية. فيما يتعلق بالأنماط الدلالية فقد صنف المتتابعات اللفظية المميزة للمدونة إلى عدة أنماط، مثل: الدلالة الزمنية والمكانية والتبليغية وتلك التي تدل على الحركة والتفكير والإدراك أو تشتمل على الإشارة إلى عضو من أعضاء الجسد، وغيرها من الدلالات التي يرى الباحث أنها من سمات لغة الروايات المدروسة والمتمثلة لأدب تلك الفترة من تاريخ الأدب الغربي. وختم دراسته بتتبع السلوك التركيبي/النظمي "semantic prosody" المميز لبعض الألفاظ الواردة في المدونة مركزاً على ثلاث كلمات وهي: thought (فكرة)، sun (شمس)، jerk (يهز رأسه)، فمن خلال وظيفة التصاحب اللفظي "collocation" المزودة في الأداة الحاسوبية Sketch Engine، رصد الباحث الكلمات والعبارات التي تصاحب هذه الكلمات الثلاث تصاحباً يجعل منه نمطاً خاصاً وسمّة بارزة لأدب تلك الفترة، وقد استعان بالاختبارات الإحصائية لتحقيق هذه النتيجة. فمثلاً، وجد الكلمات المتصاحبة مع كلمة thought يمكن تقسيمها إلى ستة حقول دلالية (مثل: الكلمات ذات الدلالة العاطفية)، ووجد الاستخدام الشائع لأفعال معينة تسبق كلمة sun، وهي أفعال مأخوذة من الألمانية القديمة (Old Germanic)، كما لاحظ أن التركيبي النظمي (at the very thought of + Noun/Verb-ing) يمثل سلوكاً تركيبياً خاصاً بالمدونة الأدبية المدروسة يكاد يغيب في اللغة الإنجليزية العامة. هذه النتائج وأمثالها لم تكن لتيسّر لولا الاستعانة بمنهجية لسانيات المدونات الحاسوبية.

ومن الدراسات ما يتجه نحو المقارنة بين مدونتين روائيتين، من ذلك دراسة راشد محمد وزملائه (Mahmood et al, 2013) الذي درس تمثيل الرجل والمرأة في الروايات المكتوبة بالإنجليزية في بريطانيا وباكستان، سعيًا منه إلى رصد جوانب التشابه والاختلاف في الكتابة الروائية بين البلدين اللذين يمثلان ثقافتين مختلفتين، بمعنى أدق، أراد تتبع الكيفيات التي يمثل بها الرجل والمرأة في كلتا الثقافتين من خلال الرواية بوصفها من أهم النوافذ الثقافية وتفسير تلك الكيفيات. وقد اعتمدت على المنهجين الكمي والكيفي لتحقيق

هدف الدراسة، مستعيناً بلسانيات المدونات الحاسوبية، مطبقاً عدداً من أدواتها مثل: الكشف السياقي وقوائم الشيوع والكلمات المميزة. ومن بين النتائج التي توصل إليها أن الرواية الباكستانية المكتوبة باللغة الإنجليزية تعرض الرجل مقروناً بالصفات الإيجابية (حوالي ٧٠٪ من الصفات)، مثل: القوة والشجاعة والذكاء، كما أن ثيمة "سيطرة الرجل على المرأة" جاءت مهيمنة في الخطاب الروائي، بينما السمة الغالبة في وصف المرأة جاءت مقروناً بالصفات السلبية (أكثر ٧٥٪ من الصفات)، مثل: الضعف والغباء والقبح، كما جاءت ثيمة "وجوب طاعة الرجل وعدم استقلالها عنه" مهيمنة. وهذا يعكس -حسب الباحثين- ثقافة المجتمع الباكستاني نحو الرجل والمرأة على النتاج الأدبي، مما يدل أن الأدب مرآة للمجتمع. وهذا السلوك اللغوي للخطاب الروائي الباكستاني جاء مختلفاً عن الخطاب الروائي البريطاني، خاصة فيما يخص المرأة، فجاءت أغلب الصفات المقرونة بالمرأة متضمنة السمة الإيجابية، مثل: وصفها بحقها في التعليم والعمل، وبالاعتماد على النفس وبالقوة. ويختتم محمد دراسته بالأهمية التي تتسم بها منهجية لسانيات المدونات في دراسة الأدب ودورها في الوصول إلى نتائج ورؤى جديدة نحو الخطاب الروائي.

من الدراسات الحديثة دراسة شبير وشاه زادي Shabbir and Shahzadi (٢٠٢٥) التي هدفت إلى تحليل ثلاث روايات مكتوبة بالانجليزية للروائي البريطاني-الباكستاني محمد حنيف من منظور لسانيات المدونات، مع التركيز على الوحدات المعجمية المميزة (من خلال مقارنة كلمات المدونة بالمدونة البريطانية الوطنية BNC) ودورها في تشكيل أسلوب لغته الروائية. كما استعانت الدراسة بلسانيات هاليداي الوظيفية في تفسير اختياراته المعجمية. وكان حجم المدونة المدروسة ٢٠٤,٧١٣ كلمة. من النتائج التي رصدها كاتبها الدراسة تميز الوحدات المعجمية التي تدل على بعض السياقات الثقافية الخاصة (مثل: مجيء كلمة man في رأس قائمة الكلمات المميزة واستعمالها في سياق التعبير عن السلطة). كما رصدا شيوع بعض الأفعال التي تدل على العنف وتلك التي تشير إلى معتقدات (أيديولوجيات) الشخصيات المتخيلة وانفعالاتها العاطفية ودورها السرد في تصوير سمات تلك الشخصيات. كما رصدا شيوع كلمة own الدالة على الملكية للتعبير عن قوة الشخصية وتعزيز الهوية. واختتمت الدراسة بأن اختيارات الوحدات المعجمية المميزة في روايات محمد حنيف ترتبط ارتباطاً وثيقاً بمفهوم السلطة السياسية والنقد الاجتماعي والتعددية الثقافية.



هذا العرض للدراسات الأجنبية يكشف عن الاهتمام الكبير من الباحثين الغربيين بمنهجية لسانيات المدونات وتطبيقها في تحليل النص الروائي وتتبع سلوك اللغة المستعملة في الروايات المكتوبة باللغة الإنجليزية. ويمكننا بعد هذا العرض أن نسجل بعض ما يمكن أن تقدمه لسانيات المدونات لتحليل النص الروائي تحليلاً لغوياً:

١. فحص الفرضيات والنتائج السابقة التي يتبناها النقاد نحو أسلوبية رواية / روايات معينة، فإما أن تؤكد أو تُحطى أو تعدّل، وغالباً ما تكون تلك الفرضيات تعتمد على الحدس والاستنتاج الذاتي المبنيين على قراءة تلك الروايات.
٢. الوصول إلى نتائج ذات أدلة محسوسة تصطبغ بصبغة قريبة من الموضوعية، وتنأى - قدر الإمكان - عن الأحكام الذاتية المبنية على الانطباع والحدس.
٣. تيسير التعامل مع البيانات الضخمة الممثلة لعدد كبير من الروايات، وقد ينتج عنه إمكانية تعميم النتائج في حال كانت النصوص المضمنة في المدونة تتسم بـ "التمثيل".
٤. تيسير الكشف عن الأنماط التي تقف خلف استعمال اللغة في النصوص الروائية، وهذا الأمر هو الغاية الأهم من تطبيق لسانيات المدونات في تحليل النصوص.
٥. إمكانية إجراء دراسات مقارنة بين أسلوبية كُتاب معينين أو روايات تنتمي إلى بيئات جغرافية مختلفة ضمن إطار منهجي منضبط.
٦. إمكانية دراسة التغير التاريخي لاستعمال اللغة الأدبية في الرواية وتتبع تطورها لدى كاتب معين أو فئة معينة.
٧. بناء الجسور بين عدة تخصصات: اللسانيات والأدب؛ متمثلاً في الرواية من جهة، والبرمجيات الحاسوبية والإحصاء من جهة أخرى. وهذه الممارسة العلمية تندرج تحت الحقل المعرفي المعروف بالإنسانيات الرقمية (Digital Humanities) والذي يُعرّف بالحقل البيئي الذي يبحث في العلوم الإنسانية من خلال التقنية، فهو يجعل من التقنيات الرقمية والوظائف الحاسوبية منهجاً للتعاطي مع العلوم الإنسانية المختلفة، وهذا يشمل على - سبيل المثال - جمع البيانات ومعالجتها ونشرها وتحليلها وترميزها باستخدام التقنيات الحاسوبية (Cuddon, 2013, p.204).^(١)

(١) للاستزادة حول هذا المصطلح وما يتعلق به من تعريفات وآليات تطبيقه ينظر: شعبان (٢٠٢١)، ص. ٥٣٣-٥٤٥.

ويأتي السؤال هنا: ما واقع الدراسات العربية السردية من تطبيق منهجية لسانيات المدونات؟ أقول ابتداءً: إننا عندما نولي وجوهنا شطر الدراسات العربية نجد الغياب الواضح لهذه الممارسة المنهجية في تحليل النص الروائي، ولعل ذلك يعود إلى أمرين:

١. أن المدونات اللغوية بشكل عام ومنهجيتها بشكل خاص، لم تعط حقها من التعريف للباحث العربي وبيان أهميتها، خاصة الباحث في مجال الدراسات السردية، فما زالت الكتب والأبحاث المكتوبة باللغة العربية في هذا المجال دون المأمول، فغياب التعريف يقود بلا شك إلى غياب الممارسة المنهجية وإدراك أهميتها.

٢. عدم قناعة الباحث العربي في مجال الدراسات السردية بنجاعة المناهج الكمية ودورها في تحليل النص الروائي، والسمة الكمية (quantitative) تعد من أهم سمات لسانيات المدونات وتعتمد عليها اعتماداً كبيراً؛ كونها تهتمّ بتتبُّع تكرار الأنماط وتحليلها إحصائياً (كاحتساب التكرار النسبي وتطبيق الاختبارات الإحصائية)، ولا تلقي الضوء على جوانب أخرى مثل: المعنى وسياقاته والملاحم النصية التي تشكّله والإيحاءات العاطفية^(١)، وهذه الإشكالية قد كان لها صدى من قبل في الدراسات الأدبية لدى الغربيين، يقول ويني (Wynne, 2006, p.5) "قضية عدم الاعتداد بالاتجاه الكمي في تحليل النصوص الأدبية ناتجة عن الاعتقاد بكونه قد يؤدي إلى إغفال النصية والمعنى". ومثله يذكر محمود (Mahmood, 2013, p.97) أنه "غالباً ما يقال إن آليات لسانيات المدونات لا يمكن تطبيقها في تحليل النص الأدبي". وعلى الرغم من أن هذه الانتقادات تتسم بشيء من الحقيقة لكون لسانيات المدونات لا تعطي إجابة عن جميع الأسئلة التي تنطوي عليها الدراسات السردية، فإن هذا لا يقلل من أهمية تطبيق هذه المنهجية؛ لأن الأدوات الحاسوبية المتطورة قادرة -وبشكل كبير- على التعامل مع البيانات الضخمة من النصوص واستخراج النتائج الكمية الدقيقة، وهذا مما لا يتاح عند التعامل مع البيانات الضخمة بطريقة تقليدية -، إلا

(١) ينظر ما ذكره د. صلاح فضل من الانتقادات نحو استخدام الإحصاء في الدراسات الأسلوبية في: فضل (١٩٩٨، ص. ٢٢٧).



أنه يجب أن يُتم ذلك بدور الباحث في تفسير تلك النتائج، فتكون منهجية لسانيات المدونات بذلك قد جسّرت للعلاقة بين الموضوعية والذاتية، فكلاهما مكمل للآخر -كما أشرنا سابقاً، فما لا تدركه الآلة يمكن إدراكه من خلال القدرة البشرية على التحليل والتفسير.

ولعلي أقف قبل نهاية هذا المبحث عند دراسة مكتوبة باللغة الإنجليزية وكان جزءاً من المدونة المدروسة روايتان عربيتان، وهي دراسة منى شلبي (Sahalby, 2016) التي بحثت في ظاهرة التعبير الشفوي في الحوارات المضمنة في الروايات، وهي دراسة مقارنة بين أربع روايات إنجليزية وعربية، وقد عُقدت المقارنة على مستويين: بين روايات كل لغة على حدة، وبين الروايات الإنجليزية من جهة والروايات العربية من جهة أخرى. ويُقصد بالتعبير الشفهي: تلك العبارات التي تحاكي اللغة المنطوقة في الحياة الحقيقية للمتحدثين. وقد مكّنت منهجية لسانيات المدونات الباحثة من رصد عدد من الظواهر المتعلقة بأخذ الأدوار عند التحوّل، والتحويل إلى العامية في بعض السياقات، وطريقة عرض الحوار، والعبارات المسكوكة والتغير الأسلوبي. وسأذكر بعض النتائج المتعلقة بالروايات العربية فقط. منها أن الباحثة رصدت أن رواية شيكاغو هي أقل الروايات استخداماً للحوار، ورصدت ذلك من خلال قلة استخدام الأفعال التي تشير إلى الحوار (مثل: قال - قلت)، وهذا يعكس ميول الكاتب إلى سرد الأحداث أكثر من إقحام حوارات بين الشخصيات. كما لاحظت أن شيوع الألفاظ التي تستخدم في الخطاب المباشر وهيمنتها في قوائم الشيوخ، خاصة ضميري المتكلم "أنا" والمخاطب "أنت". وهي ظاهرة ليست خاصة بالروائيتين العربيتين، بل رصدت شيوعها كذلك في قوائم الشيوخ الخاصة بالروائيتين الإنجليزيتين.

ومن خلال مقارنة قوائم الشيوخ الخاصة بالروائيتين العربيتين بمدونة arTenTen، وهي مدونة عربية مرجعية يزيد حجمها عن خمسة مليار كلمة (Belinkov et al., 2013)، تمكّنت الباحثة من تسجيل عدد من الملاحظات التي تميّز كل رواية عن الأخرى. فمثلاً، لاحظت اعتماد رواية (تاكسي) على استراتيجية الحوار بشكل أكبر مقارنة برواية شيكاغو، فقد ظهرت تلك الألفاظ التي تعبر عن الخطاب المباشر في قائمة الكلمات المميزة لرواية (تاكسي)، مثل: "أقولك"، مما يعني أن هذه الرواية تستند إلى التفاعل المباشر بين شخصياتها من

خلال إستراتيجية الحوار. في المقابل، لاحظت غياب تلك الألفاظ التي تشير إلى الخطاب المباشر في رواية شيكاغو، في حين ظهرت في أعلى قوائم الشيوخ عدد من المفردات التي تشير إلى الكلمات الوظيفية وأسماء الأعلام والأماكن، وقد أكد هذا فرضيتها السابقة من أن هذه الرواية لا تستند كثيراً إلى إستراتيجية الخطاب المباشر مقارنة برواية (تاكسي).

وقد أسهمت المنهجية التي اعتمدت عليها الباحثة في تتبع الظواهر الأسلوبية في الروايات المدروسة والوقوف على بعض الفروقات، فحين كان كاتب رواية (تاكسي) يراوح بين مستويي العربية المعاصرة (الفصحح والعامية المصرية) في الحوارات التي يجريها على ألسنة الشخصيات، كان المستوى اللغوي الطاغوي في رواية شيكاغو هو الفصحح، فمستوى لغة السرد يكاد يكون مماثلاً لمستوى لغة الحوار - على حد تعبير الباحث -.

وهناك ظواهر متعددة درستها الباحثة في الروايات من خلال تقنيات لسانيات المدونات التي زودت الباحثة باليات وأدوات يسرت الوصول إلى نتائج مصحوبة بالأدلة الموضوعية. على أنه من الملحوظ أن الباحثة اعتمدت على مدونة صغيرة جداً مما أدى إلى محدودية النتائج؛ فيصعب إطلاق أحكام خارج إطار تلك الروايات، فلا يمكن تعميم هذه النتائج على أسلوبية الرواية العربية أو المصرية أو حتى أسلوبية مؤلفي الروايتين العربيتين. وهذا عكس ما لاحظناه في عدد من الدراسات الأجنبية التي تسلط الضوء على روايات كاتب معين أو حقبة زمنية معينة، فتكون المدونة المدروسة كبيرة إلى حد يمكن الباحث من رصد الأنماط وإطلاق الأحكام العامة. وهذا ما أنوي تطبيقه في هذه الدراسة التي تعتمد على مدونة كبيرة نوعاً ما، تحتوي على عدد من الروايات المكتوبة من مؤلفين مختلفين حتى نستطيع التقاط صورة كبيرة للظواهر اللغوية والسلوكيات السياقية.

وعندما نصّيق مجال البحث ونركزه في مجال الرواية السعودية، فإننا نجد الغياب التام لتطبيق لسانيات المدونات الحاسوبية بوصفها منهجاً لإجراء التحليل، فلا أعرف حسب اطلاعي دراسة قامت بذلك وأفادت من هذه المنهجية؛ لذا سأقدم نموذجاً تطبيقياً للسانيات المدونات في تحليل جوانب من لغة الروايات السعودية بغية الوقوف على مدى جدوى تطبيقها في هذا المجال.

٤- الدراسة التطبيقية: تطبيق على الرواية السعودية

٤-١- وصف المدونة المدروسة وأدوات المعالجة والتحليل

يُعنى هذا المبحث بوصف بيانات المدونة موضوع الدراسة وهي مدونة خاصة بالروايات السعودية، مع الإشارة إلى الأدوات الحاسوبية المستخدمة في المعالجة القبلية للنصوص وتلك التي استخدمت في توليد القوائم والبيانات المراد تحليلها.

اعتمدت في هذا البحث على المدونة اللغوية للروايات السعودية، وهي مدونة متخصصة تحتوي على ٥٣ رواية يبلغ حجمها ٣,٠٢٧,٤٥٣ كلمة، تمثل روايات كتبت من كلا الجنسين (الروائيين والروائيات) وصدرت خلال الأعوام ١٩٣٠ - ٢٠١٩ (Alfraidi et al. 2022). ومن المعالجات القبلية التي أُجريت على نصوص المدونة أنها شُذِّبت بواسطة المشذب العربي^(١)، وذلك لتقليل تشتتها، فقد أزيلت الفراغات الزائدة والتطويل (الكشيدة)، وفُصلت الأرقام والرموز وعلامات الترقيم عن الكلمات. ولما كانت نصوص المدونة الحالية غير مقطعة لأسباب تتعلق بالواسم المستخدم في تحشية كلماتها (Alfraidi et al., 2022)، فقد قُطعت كلمات المدونة الخام (غير الموسومة) بواسطة الأداة الحاسوبية Arabic Linguistics Pipeline^(٢) (Freihat et al. 2018)، وذلك بفصل اللواحق والسوابق الصرفية (المورفيمات)، فقامت الأداة -مثلاً- بفصل "ال" التعريف عن الأسماء بعدها، وفصل حروف الجر الأحادية عما بعدها (فليذهب تصبح ل يذهب، وكالشمس تصبح ك ال شمس)، وهذه الخطوة مهمة؛ إذ تمكننا من رصد تكرار هذه الكلمات بدقة حتى ولو كانت ذات سوابق ولواحق في الأصل، وقد نتج عن تقطيع كلمات المدونة أن زاد حجمها إلى ٤٢٠٩٣٢٧، أي بنسبة ٣٩٪.

وقد استعنت في بعض مواضع البحث بمدونة أخرى لغرض المقارنة، وهي مدونة الصحافة العربية ٢٠١٢ (Al-Thubaity et al., 2013)، وهي تضمّ نصوصاً إخبارية مستخرجة

(١) يمكن تحميل المشذب العربي من خلال الرابط الآتي: https://sourceforge.net/projects/ghawwasv4/files/Almoshatheb_Alarabi%28V_4.0.0%29_09_12_18.jar/download

(٢) الوصول إلى الأداة يكون عبر الرابط الآتي: <http://arabicnlp.pro/>

من صحف عربية متنوعة من حيث الدول الصادرة عنها ومن حيث المجالات، وعدد كلمات هذه المدونة ٢,١ مليون كلمة.

أما الأداة المستخدمة في التحليل فهي برنامج (غواص ٦,٤)، وهو برنامج تم تطويره لتسهيل عملية تطبيق لسانيات المدونات في تحليل النصوص العربية، فيمكن الباحثين من حساب تكرار كلمات المدونات تحت الدراسة وحساب النسب المتعلقة بذلك، واستخراج الكلمات المميزة وتزويدهم بالكشافات السياقية والمتصاحبات اللفظية. وحسب دراسة (Alfaifi and Atwell, 2016) فإن هذا البرنامج يعد الأفضل للبحث في النصوص العربية مقارنة بغيره من البرامج، مثل: Sketch Engine و AntConc و aConcord، مستندين إلى عدد من المعايير.

٤-٢- التحليل والنتائج

٤-٢-١- قوائم الشيوخ

يعد البحث في قوائم الشيوخ (frequency lists) من أهم الخطوات التي تُنفذ عند تطبيق منهجية لسانيات المدونات في دراسة النصوص وتحليلها، بل إن البعض عدّ هذه الخطوة والكشاف السياقي أساساً لتطبيق هذه المنهجية (Hussein, 2020, p.78)، فبواسطتها نتعرّف على الكلمات الأكثر تكراراً في المدونة المدروسة وتوزيعها الإحصائي. ومن مميزات هذه الخطوة أنها تعطي الدارس صورة عامة حول السلوك اللغوي للمدونة (Hussein, 2020, p.79)، فيتمكّن من التعرّف على طبيعة الكلمات المستخدمة من حيث نوعها (كالأسماء والأفعال) والصيغة (كالمفرد والجمع، والمذكر والمؤنث) ودلالاتها (كانتمائها إلى حقول دلالية معينة). يعرض جدول ١ قائمة الكلمات المئة الأكثر شيوعاً في مدونة الروايات السعودية:



جدول (١) الكلمات المئة الأكثر شيوعاً في مدونة الروايات السعودية

ترتيب الكلمات	أ	أن	؟	التي	إن	ثم	شيء	غير	رأس	هناك
الكلمات	و	ي	لم	"	حتى	نفس	قال	مرة	آخر	كأن
	.	على	هم	الذي	ذلك	عند	—	أي	أمام	الآن
	،	ك	ني	لكن	يا	أم	حين	بعض	أمر	أخرى
	هـ	لا	!	مع	أو	كنت	مثل	أنت	كم	دون
	ب	ما	عن	قد	س	الله	صوت	ان	رجل	مكان
	ها	ف	كان	بعد	بين	إلا	وجه	كيف	قلت	أكثر
	ل	—	نا	هو	أنا	قبل	أحد	يد	لو	باب
	في	:	كل	كانت	هي	هل	يوم	بيت	)	أب
	من	إلى	هذا	هذه	تلك	هما	إلى	إذا	(	ماذا
%	٣٦,٣	٩,١	٤,٢	٢,٦	٢	١,٥	١,١	٠,٩	٠,٨	٠,٧
	٥٩,٢									

يمكننا تسجيل بعض الملاحظات حول ما يتضمنه جدول ١ من ظواهر تعكس طبيعة اللغة المستعملة في الروايات السعودية. نلاحظ أولاً أن هذه الكلمات المئة تمثل حوالي ٥٩٪ من كلمات المدونة، وهي نسبة كبيرة، مما يدل على أهمية هذه الكلمات ودورها في تشكيل عدد من الأنماط اللغوية في المدونة. فمثلاً نجد أن غالبية هذه الكلمات تنتمي إلى ما يعرف بالكلمات الوظيفية، وهي كلمات تؤدي وظائف نحوية وليس لها دلالات معجمية عندما تعزل عن التركيب، مثل: حروف الجر، أدوات الاستفهام والضمائر وأسماء الإشارة (Kestemont, 2014, p.60)، وتشكّل هذه الفئة من الكلمات ما نسبته ٧٣٪ من القائمة. ونرى كذلك ورود بعض علامات الترقيم في الجدول، وهذا يدل على حرص الكُتّاب أو الناشرين على التقيد بتلك العلامات نظراً لدورها في تأدية بعض المعاني خاصة حين تغيب الأدوات التي تؤدي تلك المعاني أو حين تغيب القيم الصوتية في الكتابة، مثل: علامة الاستفهام في حال غياب الأداة، وعلامة التعجب في حال غياب التنغيم، وكذلك علامتا الشرطة (-) والتنصيص (") للدلالة على الحديث الذي يرد على لسان شخصيات الروايات.

أما فيما يتعلق بكلمات المحتوى، وهي الكلمات ذات الدلالة المعجمية، مثل: الأسماء والأفعال (Kestemont, 2014, p.60)، فإننا نلاحظ غيابها في الخمسين كلمة الأولى؛ فقد وردت أول كلمة من هذه الفئة في المرتبة ٥٢ وهي "نفس"، يتلوها لفظ الجلالة في المرتبة ٥٦ ثم "شيء" في المرتبة ٦١. وتمثل كلمات المحتوى ما نسبته ١٧٪ من القائمة، وتشكّل الأسماء فيها النسبة الأغلب؛ إذ لم يرد من بين كلمات المحتوى إلا فعلين يشيران إلى دلالة معجمية وزمنية واحدة مع اختلاف الضمير في كلٍّ منهما، وهذان الفعلان هما: "قال وقلت". وسبب شيوع فعل القول على وجه الخصوص هو طبيعة اللغة الروائية التي تحتوي على حوارات بين شخصيات الرواية بوصف ذلك جزءاً رئيساً من البنية السردية التي تعتمد على التفاعل؛ إذ يشكّل الحوار صيغة أساسية من صيغ العرض في البنية السردية (قسومة، ٢٠٠٢، ص. ١٧؛ زعلة، ٢٠١٥، ص. ١٧٦). ويرى ميفلد (Maiwald, 2011, p. 53) في سياق تحليله لرواية الكاتب الإنجليزي George MacDonald "أن شيوع هذا الفعل لا يعد موطن تعجّب في سياق استعماله في الخطاب الروائي؛ كونه يؤدي دور الوسيط الذي من خلاله يُنقل الكلام المباشر الوارد على لسان الشخصيات". وهذا يعني أن شيوع الأفعال التي تدل على الحوار في مدونة الروايات السعودية لا يمثل خصيصة من خصائص النصوص المضمنة فيها ولا يعني تميزها عن غيرها في هذا الجانب، بل يعني العكس، وهو اتساق البنية الروائية السعودية مع الإطار العام لبنية الرواية التي تستند كثيراً إلى الحوار بين الشخصيات.

عوداً على الكلمات الوظيفية، فإننا نلاحظ شيوع أداتي الشرط "إذا" ثم "لو"، وهذا الشيوع متسق مع ما أشار إليه عدد من اللغويين المعاصرين من أن "إذا" أصبحت تستعمل بوصفها أداة للشرط الخالص استعمالاً شائعاً في العربية المعاصرة بل إن شيوعها يفوق شيوع "إن"، وكذلك ال مع "لو"، ويعدون هذه الظاهرة إحدى ظواهر تطور العربية في صورتها المعاصرة، مخالفة في ذلك الأصل الذي كانت عليه العربية في صورتها التراثية التي قررها النحاة العرب من كون "إن" هي الأصل في أدوات الشرط، وأن الأصل في "إذا" هو الدلالة الظرفية^(١) (Badawie et al., 2004, p.636; Holes, 2004, p.296)، وهذا التطور له

(١) هناك نقاش طويل في المدونة النحوية التراثية حول عدم إفادة "إذا" للشرط وخلصها معنى الظرفية، ومتى أُعتدَّ بها من ضمن أدوات الشرط، ولكن ليس المقام نقاشه. ينظر: الجملة الشرطية عند النحاة العرب، الشمسان، إبراهيم (١٩٨١، ص. ١٩٩-٢٣٢).



تفسيران- كما أشار بعض الباحثين-: الأول: أنه أثر من أثار الترجمة من اللغات الأوروبية خاصة الإنجليزية والفرنسية، والتي عمد فيها المترجمون إلى ترجمة if إلى "إذا"، ثم أثر شيوعها على الكتاب جيلاً بعد جيل (Sartori, 2015, p.3). الثاني: أنه أثر من أثار اللغة المنطوقة الدارجة التي تشيع فيها "إذا" للتعبير عن الشرط أكثر من "إن" (Holes, 2004, p.299)^(١). وعلى الرغم من أن ما ورد في تلك الدراسات من أحكام وتفسيرات اعتمد فيها أصحابها على مدونات عربية عامة أو روايات عربية إلا أن ما توصلنا إليه يؤكد أن لغة الرواية السعودية تتسق مع الإطار العام للعربية الفصحى المعاصرة في هذا النمط من الاستعمال، أي: أن شيوع هاتين الأداتين "إذا ولو" في الروايات السعودية لا يعدّ خصيصة تميّزها عن غيرها من الروايات العربية أو الأجناس الأدبية الأخرى، بل هو ناتج عن تناغمها مع النمط الكتابي العام للعربية الفصحى المعاصرة.

ومن الظواهر البارزة التي نلاحظها في الجدول، شيوع الفعل الماضي الناسخ "كان" وبعض تنوعاته الصرفية "كانت" و"كُنْتُ"، وكذلك "قال وقلتُ". وهذا يدل على أن طبيعة السرد الروائي السعودي تنكس كثيراً على الإحالة على الأحداث والأحوال في الزمن الماضي، بيد أن هذه السمة لا تعني بالضرورة أنها خاصة بالرواية السعودية، فقد أشار غير واحد من المتخصصين إلى أن السرد الروائي -عموماً- يشيع فيه استعمال الفعل الماضي نتيجة لكونه يحتضن بين طيّاته الإحالة إلى الأحداث والأحوال الماضية كثيراً (Biber and Conrad, 2009, p.138؛ الحازمي، ٢٠٠٦، ص. ٣٨٨). وقد قمت في خطوة أخرى بحصر الأفعال المئة الأكثر شيوعاً في المدونة، فكان نصيب الأفعال الماضية منها تجاوز ٥٠٪، مما يؤكد الحضور الطاعني للزمن الماضي. وهذه النتيجة تتسق مع ما توصل إليه بايبر وكونراد (Biber and Conrad, 2004, p. 47) للذان قارنا بين شيوع الفعل الماضي وشيوع الفعل المضارع في الروايات الإنجليزية، حيث لاحظنا تفضيل استعمال الفعل الماضي في سرد الأحداث أكثر من الفعل المضارع.

ويتضح سلوك اللغة في الرواية السعودية وضوحاً بارزاً عندما ننظر إلى قائمة الشيوع في مدونة أخرى تختلف عن جنس الرواية. جدول ٢ يعرض الكلمات المئة الأكثر شيوعاً في مدونة الصحافة العربية ٢٠١٢:

(١) عبّر عنه هولز بالاحتمال الأقرب.

جدول (٢) الكلمات المئة الأكثر شيوعاً في مدونة الصحافة العربية ٢٠١٢

ترتيب الكلمات	أ	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز
الكلمات	ال	و	،	في	ب	ل	.	من	هـ	ها	
	على	أن	ما	إلى	"	التي	ف	هم	عن	لا	
	ك	هذا	الذي	مع	ان	:	ذلك	)	(	هذه	
	-	نا	إن	ي	س	هو	كل	أو	بين	بعد	
	قد	كان	خلال	عام	لم	الى	هي	لكن	,	حيث	
	رئيس	غير	بعض	قال	كانت	عمل	يوم	قبل	؟	حتى	
	عند	فريق	أي	نسبة	يمكن	م	أول	أكثر	يكون	خاصة	
	نفس	هناك	دولة	شكل	مجلس	عالم	أخرى	الله	دور	/	
	نظام	محمد	مثل	دول	"	عدد	مباراة	بن	جميع	وقت	
	عربية	تكون	دون	تم	منطقة	شركة	تلك	منتخب	إلا	منذ	
مجموع التكرار النسبي	٥٣,٦	٠,٦	٠,٦	٠,٧	٠,٨	١	١,٦	٢,١	٢,٨	٥,٩	٣٧,٥
%											

أول ما نلاحظه في جدول ٢ هو شيوع كلمات المحتوى بنسبة أكبر مما لاحظناه في الجدول السابق (جدول ١) الخاص بمدونة الروايات السعودية، فقد شكّلت كلمات المحتوى حوالي ٣٢٪ من القائمة الواردة في الجدول، وجاءت كلمة "رئيس" الأكثر شيوعاً من بين كلمات المحتوى، وهي مفردة سياسية، وهذا يدل على هيمنة الإحالة على المناصب ذات الطابع السياسي.

ومما يمكن ملاحظته في الجدول أن عدداً من كلمات المحتوى تختلف عن تلك التي ظهرت في جدول قائمة مدونة الروايات السعودية من حيث الحقل الدلالي، فمثلاً نجد شيوع كلمات تشير إلى السياق السياسي (رئيس، مجلس، نظام، دول) وإلى السياق الاقتصادي (شركة، نسبة) وإلى السياق الرياضي (منتخب، فريق، مباراة)، وهذا يعكس طبيعة المجالات والموضوعات التي تتطرق لها النصوص المضمّنة في المدونة، بينما لم نلاحظ هذا في الجدول الخاص بمدونة الروايات السعودية حيث لا تساعدنا القائمة على



استنتاج الموضوعات أو المجالات التي تتطرق لها الروايات. ويمكن تفسير هذا الاختلاف بأن لغة الصحافة تنسم بالمرجعية المباشرة للواقع وتغطية الأحداث الجارية في العالم الحقيقي فتتكرر فيها الإحالات إلى كيانات معينة، بينما تعتمد لغة الرواية على الإبداع الذي يعتمد على وصف التجربة الإنسانية في عوالم متخيلة والتركيز على تفاصيل نفسية وحوارية وسردية متنوعة حسب ما يرد في مخيلة الكاتب التي تنبثق من تجربته الخاصة؛ فتقل فيها نسبة الإحالة إلى كيانات معينة.

ويعدّ مجيء الفعل "تم" ضمن الكلمات الأكثر شيوعاً في مدونة الصحافة العربية ٢٠١٢؛ من العلامات الكمية الفارقة بين اللغة المستعملة في الصحافة العربية والروايات السعودية. فبينما ورد هذا الفعل في المرتبة ٩٤ من حيث التكرار في مدونة الصحافة العربية فقد ورد في المرتبة ٧٤٨ في مدونة الروايات السعودية. وقد نصّ عدد من اللغويين على أن مجيء الفعل "تم" متبوعاً بالمصدر (مثل: تم الاتفاق) هو من ظواهر العربية المعاصرة في المجال الصحفي (Holes, 2004, p.317)، ويمكن تفسير استعماله في الرواية السعودية بأنه تأثر باللغة الصحفية التي يهemin فيها هذا التركيب تجنّباً لصيغة المبني للمجهول، ومزيد من البحث والمقارنة بين المجالين (الصحفي والروائي) من خلال الاعتماد على مدونات أكبر ربما يكشف لنا مدى صحة هذه الفرضية.

وعلى نقيض ما رصدناه في الجدول الخاص بمدونة الروايات السعودية، نلاحظ غياب أداتي الشرط "إذا" و "لو" عن قائمة المئة كلمة الأكثر شيوعاً في المدونة الصحفية. بل وردت في مراتب متأخرة نوعاً ما، "إذا" في المرتبة ١٠٨، و "لو" في المرتبة ٣٥٠. ولعل هذا ناتج عن طبيعة اللغة الإخبارية التي تعتمد على وصف الحقائق والأحداث، بينما تعتمد اللغة الروائية على التخيل والافتراضات، وهو معنى يتناسب مع أدوات الشرط التي تحتضن غالباً الأحداث الافتراضية، سواء كانت محتملة الوقوع أو غير محتملة.

ما عرضناه في الأعلى قد أعطانا صورة عامة حول طبيعة كلمات مدونة الروايات السعودية الأكثر شيوعاً، لكن هذا لا يجلي لنا الأنماط الدلالية التي تتكرر في المدونة؛ بسبب قلة كلمات المحتوى في قائمة المئة؛ لذا اتخذت خطوة أعمق للغوص في قوائم

الشيوع، واستخرجت كلمات المحتوى المئة الأولى في مدونة الروايات السعودية بعد إقصاء الكلمات الوظيفية؛ إذ إن كلمات المحتوى تمثل أهمية كبرى في الكشف عن البناء المعجمي للغة والموضوعات (التيماث) التي يدور حولها النص أو تحتضنها النصوص المدرجة في المدونة (Hussein, 2015, p.79). جدول ٣ يعرض لنا البيانات المستخرجة الخاصة بكلمات المحتوى في المدونة:

جدول (٣) كلمات المحتوى الأكثر شيوعاً في مدونة الروايات السعودية

١٠-١	٢٠-١١	٣٠-٤١	٤٠-٣١	٥٠-٤١	٦٠-٥١	٧٠-٦١	٨٠-٧١	٩٠-٨١	١٠٠-٩١
نفس	مرة	أكثر	شيخ	يمكن	شعر	تعرف	أخذ	فجأة	شعرت
الله	يد	باب	ناس	ذات	حديث	موت	زوجة	كثير	جدة
شيء	بيت	أب	طريق	ابن	عين	أبو	جديد	واحد	أخذت
قال	رأس	قلب	شينا	حيات	صغيرة	أهل	سيارة	بدأت	أبي
مثل	آخر	وقت	أول	اسم	مدينة	ليل	واحدة	يجب	كلام
صوت	أمر	أيام	ليلة	عبد	ماء	يعرف	عمر	عمل	صباح
وجه	رجل	حياة	أعرف	لحظة	صمت	صدر	امراة	عالم	سيد
أحد	قلت	جسد	حب	كلمات	قرية	جاء	أريد	زوج	وجدت
يوم	أخرى	أرض	يقول	كثيرا	رجال	جميع	نظر	سبب	قليلا
غير	مكان	داخل	قالت	عيني	غرفة	تقول	نساء	منزل	ظهر

إذا نظرنا إلى جدول ٣ نظرة عامة من حيث أنواع الكلمات، نجد أن أغلبها ينتمي إلى فئة الأسماء، مشكّلة ما نسبته حوالي ٧٣٪، في حين شكّلت الأفعال ما نسبته حوالي ١٧٪، أما بقية الكلمات (وهي التي ميّزتها بخط تحتها) فهي محل التباس لا يزيله إلا تحليل الكشاف السياقي لهذه الكلمات وتتبع استعمالاتها، مثل كلمة "شعر" التي وردت فعلاً مشتقاً من



الشعور بمعنى الإحساس، ووردت اسماً (مثل: شعر الرأس). ونلاحظ كذلك أن كلمة "نفس" هي الأكثر شيوعاً من بين كلمات المحتوى، يتلوها لفظ الجلالة ثم الكلمة الدالة على التنكير التام "شيء".

وعند النظر إلى الأسماء الواردة في الجدول، نستطيع وبسهولة أن نستخرج عدداً من الأنماط التي تشكل حقولاً دلالية:

- الألفاظ المتعلقة بالجسد، مثل: عين، وجه، يد، جسد، قلب، ومن أمثلة ذلك في المدونة:
كانت عين مها مخضلة بالدمع
وقد لاحظ التغيير الذي طرأ على وجه أستاذه
وما إن يستكنّ جسد صالح حتى تغفو بسرعة عجيبة
- الكلمات الدالة على الزمن، مثل: وقت، ليل، صباح، ليلة، يوم. ومن أمثلة ذلك في المدونة:
وفي ليل بهيم، سلمني أبي لغد غائم
وتذكرت يوم وفاة زهرة
- الكلمات الدالة على المكان، مثل: منزل، بيت، غرفة، أرض، طريق، قرية، مدينة، مكان. ومن أمثلة ذلك في المدونة:
كانت مدعوة إلى الغداء إلى بيت إحدى أخواتها
مع الإشارة إلى أن الطبيب في قرية طيبة الوادي رفض القدوم إلى السجن.
- الكلمات الدالة على أفراد العائلة، مثل: زوج، زوجة، أب، أم. ومن أمثلة ذلك في المدونة:
وتعيش وحيدة مع زوج متسلط
أمي لا تقلقي

- فعل القول بتصريفاته، مثل: قال، قالت، قلت، يقول، تقول، وهذا يعكس الطبيعة الحوارية للخطاب المباشر^(١) بين الشخصيات الواردة في الروايات وحضوره المكثف.
- الفعل ذو الدلالة الإدراكية "عرف" وتصريفاته مثل: يعرف، أعرف، تعرف. كما جاءت الأفعال الواردة في الجدول إما ماضية أو مضارعة في بنيتها الصرفية، مع غياب واضح لفعل الأمر.
- الكلمات الدالة على الكم، مثل: أكثر، كثيراً، قليلاً. ومن أمثلة ذلك في المدونة:
أذكر أنني أخبرته بقراري فلم يجادلني كثيراً
فربما سأقسو عليك قليلاً أمام بصرهم
- الكلمات الدالة على الموجهية "Modality"، مثل: يمكن، يجب، يريد. ومن أمثلة ذلك في المدونة:
هذا الحارس الليلي كان آخر أمل يمكن أن يوصلني إلى بغيتي.
- كما نلاحظ شيوع عدد من الثنائيات، مثل: ثنائية الموت والحياة، وثنائية الصمت والكلام، وثنائية القرية والمدينة.

وشيوخ هذه الحقول يدل على مركزيتها في الخطاب الروائي السعودي واعتماد البنية السردية عليها في أغلب الأحوال. فالإشارة إلى أعضاء الجسد والزمان والمكان وأفراد العائلة تؤدي أدواراً رئيسية في البنية الروائية وصياغة المشهد الروائي السعودي. وهذا انعكاس جلي لطبيعة الخطاب الروائي بشكل عام والذي تتكوّن بنيته من عناصر رئيسية، ومنها: الزمان والمكان والشخصيات، فالرواية فن زمني؛ إذ تتكئ على الزمن في سرد الأحداث وتتابعها، ويمثل المكان على اختلاف أحيائه سمة رئيسية في السرد؛ إذ يمثل الفضاء الذي يحتضن الأحداث بينما تشكل الإشارة إلى أفراد العائلة جزءاً فاعلاً في تشكيل الشخصيات^(٢)؛

(١) الخطاب المباشر هو "قول يصدر عن إحدى الشخصيات فيُنقل بحرفيته". ينظر: معجم السرديات، محمد القاضي وآخرون (ص. ١٨٦).

(٢) ينظر في عناصر بنية الرواية: بحراوي، تشكيل الخطاب الروائي (١٩٩٠، ص. ١٩-٢٧)، قاسم، بناء الرواية (٢٠٠٤، ص. ٣٧، ١٠٦)، الحازمي، البناء الفني للرواية السعودية (٢٠٠٦، ص. ٣٢).



فنستنبط من ذلك أن التركيبة الاجتماعية للشخصيات المتمثلة في أفراد العائلة شائعة في الخطاب الروائي السعودي. كما أن رصد الثنائيات في قوائم الشيوخ يؤدي دوراً في الكشف عن عدد من الموضوعات (القيمات) التي تسيطر على المشهد الروائي السعودي.

وعلى الرغم من قدرة قوائم التكرار على مساعدة الباحثين في تتبع الأنماط الدلالية للكلمات الأكثر شيوعاً في المدونة، إلا أنه ينبغي علينا أن نكون حذرين من إطلاق تعميمات حول تلك الأنماط الدلالية، ووصفها دون النظر إلى السياق الذي وردت فيه. فلو أخذنا، مثلاً، ألفاظ الجسد، فإننا لا نستطيع أن نصف الأنماط الدلالية المنبثقة من استعمالها في المدونة إلا إذا تتبعنا سياقاتها؛ فلا نستطيع أن نحكم أو نبني فرضية مفادها أن استعمال تلك الألفاظ تدل حصراً على الإشارة إلى لغة الجسد فقط دون غيرها من الأنماط الممكنة، فالأمثلة الآتية والمستخرجة من المدونة تشير إلى دلالات أخرى (مجازية) لا علاقة لها بلغة الجسد، مثل:

الوقوف في وجه هذه المقاومة.

إنّ ما تقوم به هو عين الخطأ.

لم يكن الباب موجوداً في جسد البيت الأصلي.

ولتأكيد ذلك، فقد تتبعنا سياقات استعمال كلمة "عين" في المدونة عبر وظيفة الكشف السياقي التي يتيحها برنامج (غواص). فبعد استخراج جميع السياقات (حوالي ١٨٦٣ سطرًا) وتحليل المئة سطر الأولى، تمكنت من تحديد عدة أنماط دلالية لكلمة "عين"، وهي مع أمثلة من المدونة:

١. العين الباصرة: (كانت تسترق النظر إلى عين سلمى).
٢. عين الماء. (أرادت أن تزور تل التوبة بعد أن تغتسل في العين المباركة)
٣. العين بمعنى التأكيد. (نعم هو بعينه).
٤. العين الحاسدة. (بالتأكيد هي عين وأصابني)

٥. العين بوصفها جزء من تراكيب مسكوكة لها معانٍ خاصة، مثل: عين الخير، أثراً بعد عين، قرير العين.

٦. لبّ الشيء وجوهره. (هو عين الصواب).

٧. جزء من اسم مركب. (قررنا الاستراحة في قرية الرصد).

فنخلص من ذلك إلى أن أدوات التحليل التي توفرها الأنظمة والبرمجيات التي تعالج وتبحث في المدونات - مثل: غواص - يجب أن تتظافر للوصول إلى وصف أعمق لبيانات المدونة المدروسة.

٤-٢-٤ المتتابعات اللفظية

يستخدم مصطلح المتتابعات اللفظية بوصفه مقابلاً للمصطلح الإنجليزي "Clusters"، ويمكن تعريفها بالمتواليات التي تتوالى بامتداد معين، كأن تتوالى كلمتان أو ثلاثة أو أربعة وهكذا (الشمري والثبيتي ٢٠١٩، ص. ١٠٩). ويشار إليها بمصطلح "lexical bundles" عندما يستند رصدها إلى معايير كمية؛ فينتج عن ذلك أن تعدّ المتتابعات المستخرجة من مدونة ما، هي الأكثر شيوعاً من منظور إحصائي (Biber, 2006, p.134; Hyland, 2018, p.383).

وتقوم فكرة النظر في المتتابعات اللفظية والبحث فيها على أن نسق التركيب اللغوي وتتابع كلماته بطريقة معينة لا يقع عشوائياً، بل ينطوي على سلوك لغوي نظامي يتناغم مع أنواع النصوص وسياقاتها (Biber, 2006, p.174, Senthok et al., 2019, p.14)، بل يعبر تكرارها عن أنماط تكشف عن سلوك اللغة المستعملة في مدونة ما. وقد عرضتُ لبعض من نماذج البحث في هذه الظاهرة في الرواية الإنجليزية في المبحث الثاني. ويشير هايلان (Hyland, 2018, p.383) وسِنْثوك وزملاؤه (Senthok et al., 2019, p.14) إلى أن المتتابعات اللفظية تؤدي وظيفة مهمة وهي التمييز بين أنواع النصوص ومجالاتها، فيمكن إدراك أي نوع/ مجال ينتسب نص/ نصوص ما، وذلك من خلال رصد المتتابعات التي تتكرر تكراراً يشكّل بصمة بارزة وخصيصة من خصائص ذلك النوع، كأن تؤدي دوراً في التمييز بين اللغة المستعملة في الصحافة وتلك المستعملة في الرواية أو في الكتابة الأكاديمية، أو تمييز بين الأبحاث في المجال التاريخي عن تلك التي أنجزت في



المجال التريوي. فالمتابعات الشائعة تمثل اللبنة التي تبني ملامح النص وما يحيط به من سياقات (Mahlberg, 2012, p.78; Conrad and Biber, 2005, p.58).

ولاستخراج المتابعات اللفظية المكونة من ثلاث كلمات '3-gram clusters' من مدونة الروايات السعودية، فقد حددت معيارين، **المعيار الأول**: هو تحديد الحد الأدنى للتكرار المعياري للمتابعات (normalized frequency)، وهو ألا يقل تكرارها المعياري عن ١٥ مرة في المليون^(١)، فما دون ذلك لا ينظر إليه. **المعيار الثاني**: تحديد الحد الأدنى للنصوص (الروايات) التي تتكرر فيها المتابعة، وهو ١٠ نصوص، وذلك لتقليل احتمالية أن يكون تكرار متباعدة ما خاصة بكتاب معين (Biber, 2006, p.134, Hyland and Jiang, 2018, p.386)^(٢).

وقد نتج عن تطبيق المعيارين المذكورين أعلاه جدول ٤:

جدول (٤) المتابعات اللفظية في مدونة الروايات السعودية

٩٣-٩١	٩٠-٨١	٨٠-٧١	٧٠-٦١	٥٩-٥١	٥٠-٤١	٤٠-٣١	٣٠-٢١	٢٠-١١	١٠-١
لم أكن أعرف	من هذا النوع	في أي مكان	منذ زمن بعيد	ما تبقى من	على كل حال	وفي كل مرة	لم يكن هناك	في اليوم التالي	لا بد أن
على وشك أن	في هذه الحياة	ألم أقل لك	لم أستطع أن	من أجل أن	في مثل هذا	في ذلك اليوم	أكثر من ذلك	أكثر من مرة	لا بد من

(١) يشير مصطلح التكرار المعياري (normalized frequency) إلى تحويل التكرار الواقعي إلى تكرار محدد بمقياس معياري تحسب من خلاله نسبة التكرارات، كأن تحدد نسبة تكرار كلمة في مدونة حجمها ٥٠٠٠ كلمة بالنظر إلى تكرارها الافتراضي / النسبي في مدونة حجمها مليون كلمة. ينظر:

Biber et al., (1998, p.33); McEnery and Hardie (2012, p.247).

(٢) لم أجد اتفاقاً حول هذين المعيارين وآلية تحديدهما إحصائياً، فعندما ننظر في الأدبيات التي بحثت في المتابعات اللفظية وآلية اختيارها للدراسة فإننا نجد أن من الباحثين من حدد الأدنى للتكرار المعياري بـ ٤ (Biber, 2006, p.134, Mahlberg, 2013, p.52)، ومنهم من حدده بـ ٢٠ (Hyland and Jiang, 2018, p.386)، ومثل ذلك يقال في تحديد الأدنى للنصوص، والأمر في ذلك فيما يبدو اعتباطياً إلى حد ما كما نص على ذلك بايبروهايلاند (Biber, 2006, p.134; Hayland 2008, p.8).

٩٣-٩١	٩٠-٨١	٨٠-٧١	٧٠-٦١	٥٩-٥١	٥٠-٤١	٤٠-٣١	٣٠-٢١	٢٠-١١	١٠-١
كان يريد أن	يوما بعد يوم	كيف يمكن أن	كان لا بد	كل ما في	ما يمكن أن	لا أدري كيف	من الممكن أن	في تلك الليلة	في تلك اللحظة
	في هذا المكان	لم يكن من	في كل حين	على قيد الحياة	لا لا لا	على كل شيء	في ذلك الوقت	في كل مرة	من غير أن
	في بعض الأحيان	في هذه المدينة	على ما يرام	لا يستطيع أن	فلا بد أن	في الوقت الذي	من كل شيء	في مثل هذه	لا يمكن أن
	هل يمكن أن	ما شاء الله	لا أعرف لماذا	في تلك الأيام	عن كل شيء	على أية حال	في كل شيء	إلى هذا الحد	في كل مكان
	في حالة من	غير قادر على	المرّة الأولى التي	وهو ينظر إلى	يوم من الأيام	بدمن أن	لا أستطيع أن	ولا بد أن	على الرغم من
	صباح اليوم التالي	في وقت لاحق	في حاجة إلى	في الوقت نفسه	خوفاً من أن	يمكن أن يكون	في هذه اللحظة	كل شيء في	من دون أن
	حتى لو كان	في أحيان كثيرة	مرّة أخرى إلى	وعلى الرغم من	لم يكن في	لا تريد أن	لا أدري لماذا	لا أعرف كيف	لا أريد أن
	لا يلبث أن	ولكن سرعان ما	كان يجب أن	في يوم من	لأول مرة في	كنت أريد أن	لا يريد أن	إلا بعد أن	إن شاء الله

جدول ٤: يعرض المتتابعات اللفظية المستخرجة من المدونة. وبالنظر فيه، يمكن تسجيل
الملحوظات الآتية:

أ. أن عدد المتتابعات اللفظية المكونة من ثلاث كلمات وتقع ضمن نطاق
المعياريين هو ٩٣ متتابعة^(١) تشكّل في مجملها ملامح تركيبية ودلالية
للخطاب الروائي السعودي.

(١) قمت بحذف ثلاث متتابعات من الجدول على الرغم من أن تكرارها كان أعلى من الحد الأدنى للتكرار
المعياري لكن لم يتحقق فيها الحد الأدنى للنصوص، والقاسم المشترك بين هذه المتتابعات أنها تشير إلى
أسماء شخصيات في الروايات المذكورة، وهذا يدل على أنها لا تشكّل خصيصة تميز الرواية عن غيرها من
الأنواع الأخرى للنصوص، فهي أقرب إلى اختيارات مؤلفي تلك الروايات.



- ب. أن من المتتابعات ما يمثل عبارة مكتملة (مثل: "في تلك اللحظة"، "في الوقت نفسه")،
ومنها ما يمثل جزءاً من بنية أكبر (مثل: "من دون أن"، "من الممكن أن").
- ت. نلاحظ شيوع المتتابعات التي تؤدي دور الربط بين أجزاء النص الواحد لتحقيق
سمة السبك النصية (مثل: "في تلك اللحظة"، "في تلك الليلة"، "في تلك الأيام")،
والسمة المشتركة بين هذه المتتابعات هي الإحالة إلى زمن من أزمنة الرواية.
- ث. بالنظر إلى الوظائف الدلالية التي يمكن لتلك المتتابعات أن تعبّر عنها، فقد رصدت
عدداً من الأنماط، وهي:

١. الدلالة الموجهية (مثل: لا بد أن، لا يمكن أن، لا يستطيع أن).
٢. الدلالة على الزمان (مثل: في تلك اللحظة، في اليوم التالي، في الوقت الذي).
٣. الدلالة على المكان (مثل: في كل مكان، في هذه المدينة).
٤. الدلالة التعميمية (في كل شيء، في كل حال، في كل مرة).
٥. الدلالة على التكرار (مثل: في كل مرة، أكثر من مرة، أكثر من ذلك).
٦. العبارات المسكوكة (مثل: إن شاء الله، ما شاء الله، رأساً على عقب).
٧. الدلالة على التمثيل (مثل: في مثل هذا، في مثل هذه).
٨. الدلالة الرغمية^(١) (مثل: على الرغم من، حتى لو كان).

ونخلص من هذا أن هذه المتتابعات تؤدي وظائف دلالية داخل النص الروائي ولا تتوالى
توالياً عشوائياً، وهذا يتسق مع ما ذهب إليه ماهلبيرق (Mahlberg, 2012, p.78) من أن
تكرار هذه المتتابعات يُنبئ عن تصاحبها مع وظائف تخدم النص. بالإضافة، فإن تكرارها
يشكّل عدداً من السمات التركيبية - الدلالية للخطاب الروائي السعودي.

(١) التركيب ذو الدلالة الرغمية (concessive clause) هو التركيب المكون من جملتين تعبّر إحداهما عن
وقوع الحدث بغض النظر عن ارتباطها سببياً بالأخرى، فمحتواها واقع أو صحيح لا محالة، فلا علاقة
سببية حقيقة بين الجملتين؛ بل العلاقة السببية افتراضية لها وظيفتها التداولية لغرض تقوية
المعنى، ولهذا التركيب أدوات في العربية من أشهرها "على الرغم"، ويؤدي كذلك بأدوات الشرط المركبة
مع "حتى والواو" مثل: حتى ولو، وإن، وتسمى الواو "واو الرغم". ينظر حول هذا التركيب ودلالته في:
الشمسان (١٩٨١)، ص. ٣٥٨-٣٦١؛ (Badawi et al., 2004, p.669).

وأود هنا لفت النظر إلى أن المتتابعات التي تنضوي تحت نمط الدلالة الموجهية هي الأكثر شيوعاً في الجدول؛ إذ تمثل أكثر من ٢٠٪ من مجموع المتتابعات المئة؛ لذا سأقف قليلاً للتعليق على هذا النمط ووظائفه.

يُقصد بهذا النمط هو: اشتغال المتتابعة على فعل من أفعال الموجهات (مثل: يجب، يمكن) أو أداة من أدواته (مثل: قد، ربما) أو عبارة من عباراته (مثل: إنه من الممكن)، وهي صيغ ترد للتعبير عن موقف المتكلم من فحوى الجملة التي وردت في سياقها، (يشار إليها بالقضية)، فتعبّر عن ما يعتقده المتكلم تجاه وقوع الحدث أو الفكرة المناقشة إما إمكاناً أو شكاً أو تأكيداً أو قبولاً أو رفضاً أو إيجاباً أو إرادة أو معرفة أو افتراضاً (نحلة، ٢٠٠٩، ص. ١٣٩؛ المتوكل، ٢٠١٣، ص. ١٧٣)؛ وللمعاني التي تؤديها هذه الأفعال تقسيمات مختلفة تندرج في مجملها تحت قسمين: الموجهات المعرفية (مثل: تلك التي تعبّر عن احتمالية وقوع الحدث) والموجهات الإلزامية (مثل: تلك التي تعبّر عن وجوب وقوع الحدث أو الاستعداد للقيام به) (الفريدي، ٢٠٢١، ص. ٤٣٥). وتؤدي هذه المتتابعات دوراً مهماً في الكشف عن أفكار المتكلم/ الكاتب ومعتقداته ومقاصده وما تنطوي عليه وظيفة التواصل بين المتحدثين من أغراض (الفريدي، ٢٠٢١، ص. ٤٣٥). وفي النص الروائي، يجري الكاتب هذه المتتابعات على لسانه سارداً أو على لسان إحدى الشخصيات متحدثاً، فالمتتابعات ذات الدلالة الموجهية تمثل نافذة تكشف لنا مواقف وأفكار الكاتب أو الشخصيات الروائية، وشيوع هذه المتتابعات يدل على أهميتها في صياغة الخطاب الروائي السعودي ودورها في تشكيل اللغة المستعملة وما تحمله من مضامين وتصوير صفات الشخصيات الناطقة بها. وهذا الدور لا يمكن إدراكه بصورة جليّة إلا بعد النظر في الكشف السياقي الخاص بالمتتابعات، ولنأخذ المثالين التاليين نموذجين لتوضيح هذه النقطة:

المثال الأول: "ولذلك لا بد أن أسجل اعتراضى على هذه التجاوزات أولاً، ولا بد من التقييد بالنظام الداخلي في كل خطوة".

تؤدي المتتابعة "لا بد أن" هنا وظيفة الموجهات الإرادية، فمن خلالها عبرت الشخصية عن عزمها على تسجيل الاعتراضات، وهو عزم شكلته الحمولة الدلالية التي تختزنها المتتابعة من خلال "لا بد أن" التي تدل على معنى الإلزام والالتزام نحو القيام بحدث ما.



المثال الثاني: "كلا إنني رغم وطأة عذابي الخاصة تبقى لي مبادئ وقيمي التي لا يمكن أن أتعداها أو أسمح لأي كان بتجاوزها أو خدشها".

تعبّر المتتابة "لا يمكن أن" عن رأي الشخصية المتحدثة في عدم وجود أي احتمال يؤدي إلى تخليها عن مبادئها. فنلاحظ كيف شكّلت المتتابة دوراً وظيفياً مهماً في نقل الفكرة كما أرادت الشخصية وفي تصوير الشخصية بكونها تتميز بالثبات على مبادئها، وهذه إشارة واضحة إلى المعتقد الذي تتبناه وتؤمن به.

ولا يقتصر دور الكشف السياقي على هذا فحسب، بل له دور مهم في إزالة اللبس الذي يمكن أن يغشى المتتابة التي تحتوي على مكون ذي دلالة موجهية. فبعض تلك المكونات لها القدرة على التعبير عن أكثر من دلالة موجهية حسب ما يتيح نظام اللغة التي تنتمي إليها (Jeyanthi, Zalina, Afida and Arbaayah, 2018, p.2404). لنأخذ عبارة "من الممكن أن" الواردة في جدول ٤ مثلاً لتوضيح هذه النقطة، فهذه العبارة قد تدل - حسب السياق الواردة فيه - على احتمال وقوع الحدث أو على السماح والإذن بإيقاعه، ولكن لا يمكننا التيقن من ذلك إلا بعد فحص الكشف السياقي الخاص بها، ولننظر الآن إلى المثالين اللذين استخرجنا عن طريق أداة الكشف السياقي المزودة في نظام "غواص":

يمكن أن تكون هذه مجرد مصادفة! هل من الممكن أن تكون هي ذاتها.

لديه جرح في رقبته، لكن حالته مستقرة، اطمئنوا لا شيء يدعو للقلق - هل من الممكن أن يكلم ابنته؟ هي منهارة.

في المثال الأول، تدل المتتابة على وجود احتمال، بينما تدل في المثال الثاني على طلب الإذن والسماح بالحديث مع الابنة، فتكون دلالة العبارة على الدلالة الموجهية المعرفية في المثال الأول، وعلى الدلالة الموجهية الإرادية في المثال الثاني، وهذا التفريق لم نكن لندركه لولا استعراض الكشف السياقي لذات المتتابة اللفظية.

ج. على الرغم من أن قائمة الشيوخ الواردة في الجدول قد مكّنتنا من تحديد دلالات عدد من المتتابعات ثم تنميطها فإن دلالات بعضها ما يزال غامضاً، ولا يمكن إزالة ذلك

الغموض إلا بعد تحليل الكشف السياقي الخاص بها، ولنأخذ مثلاً واحداً لتوضيح هذه النقطة، تكررت المتتابعة "كل ما في" ٦٣ مرة في المدونة، وعند استعراض الكشف السياقي لهذه المتتابعة نجد أنها تشكل عدداً من الأنماط الخاصة بها، وسأورد الأنماط الثلاثة الأكثر شيوعاً، وهي:

١. "كل ما في + مكان" (٣٣٪). ومن أمثلة هذه المتتابعة في المدونة: "كل ما في المدينة"، "كل ما في الغرفة".
٢. "كل ما في + الأمر" (٢٨,٥٪). وهي متتابعة تبدو مسكوكة متداولة للتعبير عن اختصار الأمر المتحدث عنه وعرضه بالمجمل. ومن أمثلتها في المدونة: "ما رأيته هو مجرد آثار ما بعد العملية مصحوبا بتغيرات نتجت عن الالتهاب الرئوي الذي أصيب به فؤاد، هذا كل ما في الأمر". ومما يُلاحظ أن هذه المتتابعة تشيع في سياق الحوار بين أشخاص، وليس في سياق السرد الذي يجيء على لسان السارد.
٣. "كل ما في + عضو من أعضاء الجسد" (٩,٥٪). ومن أمثلة هذه المتتابعة "أريد أن تترك كل ما في يديك الآن".

ويبدو أن المتتابعة "كل ما في الأمر" تمثل سمة من سمات اللغة السردية في الرواية السعودية مقارنة باللغة الصحفية؛ إذ لم ترد في مدونة الصحافة العربية ٢٠١٢ إلا مرة واحدة.

من المتتابعات التي تلفت النظر هنا هي "لا لا لا"، وهي عبارة النفي التأكيدية، وقد تكررت في ١٠ نصوص مختلفة؛ وهذا يعطي مؤشراً على تفضيلها لدى عدد من الكتاب / الكاتبات في الرواية السعودية من خلال إجراء هذه المتتابعة على لسان الشخصيات المتخيلة. ومن جهة أخرى، يشكّل استعمالها ملمحاً من ملامح تجسير العلاقة بين الحوار في العالم المتخيل الوارد في الرواية والحوار في العالم الحقيقي، وهي استراتيجية سردية ينتهجها بعض الكتاب للتقريب بين العالَمَين: المتخيل والحقيقي (Shalaby, 2016, p.209)، ومثل



ذلك يقال في شيوع المتتابعين المسكوكتين "إن شاء الله" و "ما شاء الله" اللتين تجريان في السياق المنطوق للغة العربية في الحياة اليومية^(١).

ح. مما نلاحظه في الجدول (٤) ندرة المتتابعات التي تشير إلى شخصيات الروايات أو الإحالة إليها؛ إذ لم يرد إلا مثال واحد يحتوي على ضمير يحيل إلى إحدى الشخصيات، وهو "وهو ينظر إلى". وغياب هذا النمط في الروايات السعودية يختلف عما أشارت إليه ماهلبيرق (Mahlberg, 2013, p.66-70) من أن هذه الظاهرة شائعة في الروايات الإنجليزية البريطانية، وأطلقت عليها مسمى "Label" والتي يمكن ترجمتها بوظيفة "التسمية".

خ. تبدو بعض المتتابعات أنها من الظواهر الأسلوبية للعربية المعاصرة الشائعة أو المستحدثة، مثل عبارة "في الوقت نفسه" التي تكررت في المدونة ٦١ مرة. ومما يدل على أنها أقرب إلى الأسلوب المعاصر للعربية أنها لم ترد في المدونة العربية (Al-Thubaity, 2015)^(٢) إلا إحدى عشرة مرة في نصوص ما قبل ١٨٠٠م، بينما تكررت ٣٢,٧٨٨ في نصوص ما بعد هذا التاريخ الذي يعد البداية النسبية للعربية المعاصرة^(٣). ومثل هذا يقال في عبارة "في تلك اللحظة" وعبارة "على ما يرام" وعبارة "لا أدري لماذا"، فهي عبارات يغلب الظن أنها أقرب إلى أسلوب العربية المعاصرة منها إلى العربية التراثية، وقد اصطُغت بها لغة الرواية السعودية.

ويمكن القول كذلك أن اللغة السردية في مدونة الروايات السعودية تفضل التركيب "في الوقت نفسه" على التركيب "في نفس الوقت"، كما يعرضه جدول ٥:

(١) قد يمثل استخدام مثل هذه العبارات ظاهرة عامة في لغة الروايات العربية، فقد أشارت منار شلبي (Shalaby, 2016, p. 246-248) إلى أن الإحالة إلى لفظ الجلالة "الله" أو الرب في عدد العبارات العربية المسكوكة المنطوقة ظاهرة تشيع في الروايتين العربيتين اللتين حلفتها من خلال منهجية لسانيات المدونات، وذكرت أن عبارة "إن شاء الله" على وجه الخصوص شائعة في كلتا الروايتين.

(٢) وهي مدونة مرجعية تحتوي على أكثر من مليار كلمة وتغطي فترة تتجاوز ١٥ قرناً من تاريخ العربية. ويمكن زيارة موقع المدونة العربية عبر منصة فلك من خلال الرابط التالي: <https://falak.ksaa.gov.sa/corpora/arabiccorpus>

(٣) ينظر حول تحديد التاريخ النسبي للعربية المعاصرة في: (Newman (2013, p.472)؛ Versteegh (2014, p.221)

جدول (٥) تكرار المتتابعين "في الوقت نفسه" و "في نفس الوقت" في المدونة

المتابعة اللفظية	مدونة الروايات السعودية	
	التكرار الفعلي للمتتابعة	الترتيب في قائمة الشبوع
في الوقت نفسه	٦١	٦٢
في نفس الوقت	٢٦	٣٢١

وتتجلى سمات المتتابعات اللفظية الخاصة بالروايات السعودية أكثر من خلال مقارنتها بتلك الخاصة بمدونة الصحافة العربية ٢٠١٢. جدول ٦ يعرض المتتابعات المئة الأكثر شيوعاً في مدونة الصحافة العربية:

جدول (٦) المتتابعات اللفظية لمدونة الصحافة العربية ٢٠١٢

١٠-١	٢٠-١١	٣٠-٤١	٤٠-٥١	٥٠-٥٩	٦٠-٦٩	٧٠-٧٩	٨٠-٨٩	٩٠-٩٩	١٠٠-١٠٩
مشيرا إلى أن	بما في ذلك	وعلى الرغم من	خلال النصف الأول	من دون أن	الناج المحلي الإجمالي	التي تمر بها	من المتوقع أن	في عدد من	٤٦ و
على الرغم من	لا يمكن أن	عدد كبير من	لافتا إلى أن	في الوقت الراهن	في ما يتعلق	٤٦ و	إلى أن هذه	مشيرة إلى أن	في الوطن العربي
في هذا المجال	الله عليه وسلم	في هذا الصدد	في كثير من الأحيان	بعض الأول من	الربع	من شأنها أن	على أرض الواقع	في منطقة الشرق	في هذا القطاع
في الوقت الذي	هذا النوع من	وذلك من خلال	التي يمكن أن	في التعامل مع	في أكثر من	في الشوط الثاني	مليون دأ	الولايات المتحدة الأمريكية	مسجلا زما قدره
صلى الله عليه	الأول من العام	المملكة العربية السعودية	منطقة الشرق الأوسط	رئيس مجلس الوزراء	في وقت سابق	في مثل هذه	في نهاية المطاف	إلى حد كبير	في الكثير من



١٠٠-٩١	٩٠-٨١	٨٠-٧١	٧٠-٦١	٥٩-٥١	٥٠-٤١	٤٠-٣١	٣٠-٢١	٢٠-١١	١٠-١
يجب أن تكون	يمكن أن يكون	في هذا الشأن	بن عبد العزيز	الشرق الأوسط وشمال	على مستوى العالم	بغض النظر عن	في الوقت نفسه	النصف الأول من	في الولايات المتحدة
المجلس الوطني السوري	خادم الحرمين الشريفين	إلى أكثر من	المحكمة الدستورية العليا	يجب أن يكون	في الوقت الحالي	في كل من	في الشرق الأوسط	لا بد من	في المائة من
وهو الأمر الذي	الربع الثاني من	اف ب	في نفس الوقت	على حد سواء	في المئة من	في هذا السياق	في حين أن	من العام الماضي	دول مجلس التعاون
ما إذا كان	الله صلى الله	خلال شهر رمضان	من شأنه أن	في المباراة النهائية	في دول مجلس	الرغم من أن	مجلس التعاون الخليجي	وأشار إلى أن	من جهة أخرى
صندوق النقد الدولي	٢٦ و	في الوقت ذاته	ثورة ٢٥ يناير	من العام الجاري	في هذه المرحلة	من العام الحالي	في العديد من	في العالم العربي	على سبيل المثال

على الرغم من وجود تشابهات في تكرار بعض المتتابعات ذاتها في كلتا المدونتين (مثل: "على الرغم من"، "في الوقت نفسه"، "لا يمكن أن"، "مثل هذه")، فإنه يمكننا تسجيل بعض الاختلافات. مثلاً، تشيع الإشارة إلى أسماء الأشخاص (مثل: "منصور بن زايد") والكيانات الدولية (مثل: "دول مجلس التعاون"، "الولايات المتحدة") في مدونة الصحافة العربية ٢٠١٢، وهو مما لم نلاحظ شيوعه في مدونة الروايات السعودية. كما تشيع الإشارة إلى منطقة الشرق الأوسط بوصفها فضاء مكانياً ومسرحاً للأحداث السياسية في تلك الفترة. كما نلاحظ شيوع المتتابعات التي تتركب مع كلمة "خلال" (مثل: "ذلك من خلال"، "خلال الفترة من") وهي مما يعتقد أنها من الظواهر التركيبية المستحدثة في العربية المعاصرة (الشمري والثبيتي، ٢٠١٩، ص. ١١٠)، وشيوع المتتابعات ذات الدلالة الزمنية التي تشير إلى فترات زمنية من السنة (مثل: "النصف الأول من"، "الربع الأول من"، "الربع

الثاني من")، وهذا عكس ما رصدناه في الجدول الخاص بالمتتابعات في الروايات السعودية حيث تكون الدلالة الزمنية الأكثر شيوعاً هي تلك التي تشير إلى الزمن القريب من سرد الأحداث والمحيط بها (مثل: "في تلك الليلة"، "صباح اليوم التالي").

ختاماً لهذا المبحث، ندرك أهمية هذا النوع من التحليل المعتمد على منهجية لسانيات المدونات ودوره في تتبع خصائص اللغة الروائية على المستوى التركيبي والمستوى الدلالي، ودوره كذلك في التفريق بين أنواع النصوص "genres"، فاللغة المستعملة ترتبط ارتباطاً وثيقاً بنوع النص الذي تحتضنها وتتشكل وفقاً لسياقاته وأغراضه التواصلية. في المبحث التالي، سأواجه بوصلة البحث نحو كلمة المحتوى الأكثر شيوعاً في مدونة الروايات السعودية وأحلل سلوكها الاستعمالي من خلال الكشف السياقي والتصاحب اللفظي للكشف عن الأنماط الدلالية التي تتشكل من خلالها.

٣-٤- كلمة "نفس": أنماطها الدلالية ومتصاحباتها اللفظية

عرضت سابقاً الكلمات الأكثر شيوعاً في مدونة الروايات السعودية وبيّنت أن كلمة "نَفْس" (وصورها المختلفة، مثل: نفسي ونفسه، ونفسها، والنفس) هي كلمة المحتوى الأكثر تكراراً، وهذا يدل على أنها مركزية دلاليّاً وتركيبياً في اللغة الروائية السعودية؛ إذ تكررت هذه الكلمة في مدونة الروايات السعودية ٦٩٤٣ مرة في جميع الروايات البالغ عددها ٥٣؛ لذا سأسلط الضوء على سلوك هذه الكلمة، محاولاً استكشاف أنماطها من خلال أداتي الكشف السياقي والتصاحب اللفظي المتاحين في نظام (غواص).

٣-٤-١- الكشف السياقي

الكشف السياقي هو قائمة من الأسطر التي تعرض الكلمة المركزية (المبحوث عنها) مصحوبة بكلمات سابقة ولاحقة، وتفيد هذه القائمة في تتبع سلوك الكلمة الدلالي من خلال السياق المحيط بها (Baker et al., 200643).

فيما يتعلّق بكلمة "نفس" فأشير-ابتداءً- إلى المعاني التي أوردتها المعاجم لهذه الكلمة قبل فحص كشفها السياقي من منظور دلالي. جاء في المعجم الوسيط على أن هذه الكلمة -أصالة-



على الروح، يقال: خرجت نفسه وجاد بنفسه، أي: مات. (المعجم الوسيط، ص ٩٤٠).
لكن لهذه الكلمة ومشتقاتها معانٍ دلالية أخرى، ومن هذه المعاني (عمر، ٢٠٠٨، ص. ٢٢٥٥):
:(Wehr, 1960, p.985

١. معنى الدم.
٢. الإنسان، ويقال لها: النفس الناطقة.
٣. العين والحسد.
٤. حقيقة الأمر وذاته وجوهره.
٥. الجسد.
٦. العقل.
٧. الإشارة إلى علم النفس وما يتصل به مثل المرض النفسي.
٨. التوكيد المعنوي.

ولاستخراج الأنماط الدلالية لهذه الكلمة في المدونة فقد حُلّت الكشاف السياقي الخاص
بها، ولما كان من الصعب تتبع جميع التكرارات سياقياً، فقد قسمت البيانات إلى أقسام
طول كل قسم ٥٠٠ سطر، ثم فقد اخترتُ أول ٥٠ سطرًا في بداية كل ٥٠٠؛ فتكوّن لديّ ٧٠٠ سطر
تمثّل البيانات التي استهدفت بالتحليل. شكل ١ يعرض عينة من الكشاف السياقي كما
استخرجه برنامج (غواص).

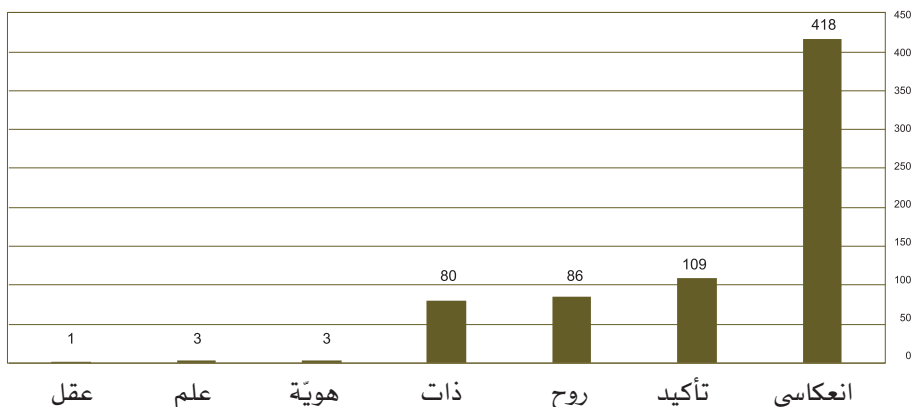
شكل (١) عينة من الكشف السياقي لمفردة "نفس" كما يعرضها برنامج (غواص)

خيارات	
عدد الكلمات السابقة: 15 ▾	عدد الكلمات اللاحقة: 15 ▾
الترتيب حسب: ▾	الماد: 6943

الكشف السيفي

وقد كشف التحليل لهذه الأسطر المختارة عن الأنماط الدلالية لهذه الكلمة وسلوكها الإحصائي. شكل ٢ يوضح تلك الأنماط مصحوبة بإحصاءاتها:

شكل (٢) الأنماط الدلالية لكلمة "نفس" وإحصاءاتها في مدونة الروايات السعودية





نلاحظ في هذا الشكل أن استخدام كلمة "نفس" في الرواية السعودية لم يكن على نمط واحد من حيث الوجهة الدلالية، وفيما يلي توضيح لهذه الأنماط:

النمط الأول: جاء استخدام كلمة "نفس" بوصفها ضميراً منعكساً - أو لنقل متصلة بالضمير المنعكس - هو الاستخدام الأكثر شيوعاً، فقد تكرر هذا الاستخدام ٤١٨ مرة مشكلاً ما نسبته حوالي ٥٩٪ من مجموع الأمثلة التي اختيرت عشوائياً من الكشف السياقي. والضمير المنعكس (reflexive pronoun) هو الضمير الواقع مفعولاً (ما حل محله) ويتفق في إحالته على ما يحيل عليه الفاعل، بعبارة أخرى، يحيل الفاعل وضمير المفعول على ذات الشخص أو الشيء، ومن أشهر صوره أن يتصل ضمير المفعول بكلمة "نفس"، مثل: "جرح محمد نفسه" و "كافأت هند نفسها" (ينظر: نخلة، ١٩٩٠، ص. ٤-١٢).

ومن أمثلة ذلك من المدونة:

أخذ يحدث نفسه كيف تبدل الحال بالمسلمين.

في الأيام الأولى، كنت أسأل نفسي مئات المرات.

قبل أن يلفظ أنفاسه الأخيرة كان يجاهد نفسه.

ويجدر لفت النظر إلى أن كلمة "نفس" هنا استعملت استعمالاً وظيفياً لتحقيق غاية الربط والإحالة إلى الفاعل الذي يتفق في إحالته مع الضمير المتصل بالنفس الواقعة مفعولاً، فيصح تجاوزاً أن نقول: إن الفاعل والمفعول به يحيلان على ذات واحدة؛ لذا - والذي يظهر لي - أن كلمة "نفس" قد قُرِغت من محتواها المعجمي الذي يدل أصالة على الروح أو الجسد - كما بينا سابقاً - واستُعملت استعمال الكلمات الوظيفية التي تؤدي وظائف نحوية، وهو الربط من خلال وظيفة الإحالة التي اكتسبتها من الضمير المتصل بها.

النمط الثاني: الدلالة على التوكيد، ويأتي هذا في المرتبة الثانية من حيث الشيوخ (١٥,٥٪)، وله صورتان تركيبيتان:

الأولى: صورة التوكيد المعنوي^(١)، كما قررها النحاة في كتبهم، ومن أمثلتها في المدونة:

وأفعل الشيء نفسه في قصائدي.

وأنا أعيش مع الموت نفسه.

الثانية: أن تقع كلمة "نفس" مجرورة بالباء:

سأعتذر للشيخ بنفسي.

لقد أشرف بنفسي على بناء قصر الضيافة.

النمط الثالث: الدلالة على الروح (١٢,٢٪)، من أمثلته في المدونة:

أثار في نفسها اضطراباً أقرب للضيق.

وقد غادرت نفسه ثم اختفت في السماء.

النمط الرابع: الدلالة على ذات الشيء (١١,٣٪)، وفي هذا النمط تضاف كلمة "نفس" إلى ما بعدها، ومن أمثلته في المدونة:

وبعد فترة صمت، تابع يقول بنفس النبوة.

ولذلك أصبحت تبسّم وتردد نفس الجملة.

(١) هناك اختلاف بين اللغويين حول تصنيف التوكيد بالنفس "جاء الرئيس نفسه"، فمنهم من يجعله صنفاً من الضمائر المنعكسة وصورة من صورته التركيبية (Mkhitarayan and Bazikyan, 2016; p.428)، ومنهم من يجعله قسماً قائماً بذاته لسبب تركيبى، وهو أن النفس عندما تستعمل بوصفها ضميراً منعكساً (مثل: جرح محمد نفسه) تكون عنصراً إجبارياً لا يمكن حذفه، بينما عندما تستعمل في سياق التوكيد؛ فإنها تكون عنصراً اختيارياً يمكن حذفه من الناحية التركيبية. (نحلة، ١٩٩٠، ص. ٤٥؛ ٦١٢، p.612، Celce-Murcia, Larsen-Freeman and Williams, 1983). وقد سرت على الرأي الثاني؛ لكون هذه الدراسة تعنى بالأنماط الدلالية في المقام الأول، ولا خلاف في وجود التباين الدلالي بين التركيبين.



النمط الخامس: الدلالة على فرع من فروع العلوم، وهو علم النفس (٤٢، ٠٪)، وتقع
"نفس" مضافاً إليها كلمة "علم"، ومن أمثلة ذلك في المدونة:

اجتازت الماجستير في علم النفس.

النمط السادس: الدلالة على الهوية (٤٢، ٠٪)، ومن أمثلة ذلك في المدونة:

من أنت حتى نسألك عن نفسك.

النمط السابع: الدلالة على العقل (١٤، ٠٪)، ومن أمثلة ذلك في المدونة:

حدّثته نفسه بأن يقف وجهاً لوجه مع شبح عائشة.

هذه النتيجة التي عرضناها تختلف جزئياً عن النتيجة التي توصلت إليها مجري (Mejri, 2022) فيما يتعلق بالأنماط الدلالية لاستعمال كلمة النفس في الصحافة العربية، وقد اعتمدت في تحليلها على ٥٠٠ سطر مختارة عشوائياً من الكشف السياقي لمدونة Arabic Timestamped JSI Web Corpus 2014–2021، وهي إحدى المدونات الفرعية لمدونة Timestamped JSI Web Corpus والتي تحتوي على عدد من المدونات الفرعية لبعض اللغات. والمدونة الخاصة بالعربية هي مدونة خاصة بالصحافة العربية وتحتوي على أكثر من ٤,٧ مليار كلمة^(١). ويرتكز الاختلاف على أن مجري رصدت أن دلالة كلمة النفس بوصفها كلمة محتوى ذات دلالة معجمية (تدل على الشخص /الأشخاص) هي الأكثر شيوعاً مقارنة باستعمالها ضميراً منعكساً أو توكيداً معنوياً. بالإضافة إلى أن الدالتين الأخيرتين (الضمير المنعكس والتوكيد المعنوي) قد تقاربتا إحصائياً عندها، وهذا عكس ما نلاحظه في مدونة الروايات السعودية التي تهيم فيه تأدية النفس وظيفة الضمير المنعكس. تجلي لنا هذه النتيجة المهمة ملمحاً من الملامح التمييزية بين أسلوبية اللغة العربية المستعملة مجال الرواية السعودية وتلك المستعملة في مجال الصحافة العربية.

(١) يمكن الاطلاع على مزيد من المعلومات عن هذه المدونة من خلال الرابط التالي:

<https://www.sketchengine.eu/timestamped-arabic-corpus/>

٤-٣-٢- التصاحب اللفظي

التصاحب اللفظي (collocation) هو تلك الظاهرة اللغوية التي تكشف عن تصاحب كلمة مع كلمة أخرى تصاحباً شائعاً ضمن سياقات معينة (Baker et al., 2006, p. 36-37)، فالتصاحب إذاً يدل على المجاورة الاعتيادية والسياقية لكلمتين معنيتين. من أمثلة ذلك في العربية (سرب الطيور، قفل راجعاً)، ولكن يجب الأخذ بالاعتبار إلى أن التصاحب لا يعني بالضرورة تجاور الكلمتين بلا فاصل، فقد يقع التصاحب حتى ولو فُصل بين الكلمة المركزية وما يصاحبها بكلمات أخرى (McIntyre and Walker, 2019, p. 41)؛ لذا يلزم تحديد المدى الذي تقع خلاله الكلمة المصاحبة عند إجراء التحليل القائم على المدونات (مثل: تحديد المدى به كلمات قبل المركزية وه كلمات بعدها. وللتصاحب اللفظي قياسات إحصائية تكشف عن مدى قوة التصاحب بين الكلمات وضعفه (المجول، ٢٠١٧، ص. ٨). وتكمن أهمية تحليل التصاحب اللفظي في رصده للعلاقة غير العشوائية بين وحدة معجمية ووحدات معجمية أخرى، تلك العلاقة التي تخلق معنى أو معانٍ داخل النص (Wijitsopon, 2013, p. 45).

فيما يتعلق بالمتصاحبات مع كلمة "نفس" في مدونة الروايات السعودية، فقد أجريت التحليل مع الأخذ بالمعايير التالية:

- أ. البحث عن المتصاحبات على مدى ٥ كلمات قبل كلمة "نفس" وه كلمات بعدها،
- ب. تحديد الحد الأدنى لتكرار الكلمة المصاحبة وهو ١٠ تكرارات،
- ج. تحديد الحد الأدنى للتكرار في السياق (أي: تكرار التصاحب ذاته) وهو ١٠ تكرارات،
- د. تحديد الاختبار الإحصائي (الاحتمالي اللوغاريتمي) Log-Likelihood مقياساً لرصد المتصاحبات، (هـ) تحديد قيمة المعادلات بـ ٢٤، وهذا يعني أن نسبة الثقة بالاختبار الإحصائي أكثر من ٩٩,٩٩٩٩٪.

وقد نتج عن التحليل القائم على هذه المعايير ٨٠ متصاحبة لفظية لكلمة "نفس". وجدول ٧ يعرض تلك المتصاحبات ويرتبها حسب القيمة الإحصائية الأعلى فالأدنى:



جدول (٧) متصاحبات كلمة "نفس"

٨٠-٧١	٧٠-٦١	٥٩-٥١	٥٠-٤١	٤٠-٣١	٣٠-٢١	٢٠-١١	١٠-١
أقف	تجد	كنت	حدثت	اعتداد	ثقة	يخاطب	ي
تابع	بعض	من	أدافع	تمالكت	لك	وقت	هـ
وجه	طريقة	علم	يهيئ	أنا	يجد	أجد	وجدت
انتبهت	دفاع	هدأت	سألت	أخاطب	يتمالك	هم	قلت
منع	جابر	تلوم	ألقت	يلوم	في	ى	وجد
تلقي	تعرض	لماذا	قال	أسأل	تتمالك	أتمالك	ال
أحضان	رأس	أقنع	هكذا	تحدث	ألوم	ب	ها
نفس	عناء	يد	إقناع	كأن	تحامل	نا	ل
همست	أحد	لو	يكلف	هيأت	عهدا	يحدث	قرارة
سمحت	يا	يعتبر	اطمأنت	أقتل	تمالك	أقول	:

وبالنظر في جدول ٧ يمكن تسجيل عدد من الملاحظات:

أ. تعدّ الأفعال أكثر الكلمات المتصاحبة مقارنة ببقية أنواع الكلام، فقد شكّلت أكثر من ٥٠٪ من عدد المتصاحبات. لكن عندما نمعن النظر نجد أن بعض المتصاحبات تعود إلى جذر واحد، مثل: (وجدت، وجد، يجد، أجد)، فهي تعود جميعاً إلى الجذر "وجد"؛ لذا ينبغي أن ينظر إلى هذا ويؤخذ بالاعتبار عند تحليل التصاحب اللفظي في حال كون مرتكز التحليل على الجذر لا الصيغة الصرفية.

ب. تعدّ الأفعال المشتقة من الجذر "وجد" هي الأقوى تصاحباً مع كلمة "نفس" والأكثر شيوعاً، والوجود هنا ليس وجوداً حسيّاً، بل هو إدراكي ونفسي، (مثل: وجدت نفسي عاجزاً أمام إصراره العتيد)، فالمتحدث هنا يشير إلى إدراكه حالة الحزن التي ألمّت به. وهذه الملاحظة تتسق مع النتيجة التي توصلت إليها مجري (Mejri, 2022:92) من أن "وجد" بوصفه فعلاً نفسياً هو الأكثر شيوعاً من حيث

تصاحبه مع كلمة "نفس" في مدونة الصحافة العربية Timestamped JSI Web Corpus 2014-2021 التي اعتمدها في دراستها؛ وهذا يقودنا إلى استنتاج، وهو أن هذا التصاحب لا يعد سمة من سمات اللغة الروائية السعودية، بل من المحتمل أن تكون من سمات اللغة العربية المعيارية المعاصرة.

ومما يلاحظ كذلك في هذا التصاحب أن كلمة "نفس" تستخدم ضميراً منعكساً أو ضميمة إلى الضمير المنعكس العائد على الفاعل، فالتصاحب بين الكلمتين يبدو أنه مرتبط بهذه الوظيفة لكلمة "نفس".

ج. يمكن تقسيم الأفعال التي تتصاحب مع كلمة "نفس" إلى ثلاثة أنماط دلالية:

١. الأفعال الإدراكية والنفسية، مثل: وجد، يعتبر، اطمأنت، ومن أمثلتها في المدونة:

وفجأة وجد نفسه يبكي

همام يعتبر نفسه عريباً

واطمأنت نفسه إليها

تما لكت هدى نفسها

ولكن نفسه هدأت في النهاية

٢. أفعال القول، مثل: قلت، يخاطب، يحدث، ومن أمثلتها في المدونة:

قلت لنفسي بتألم حقيقي:...

قال خالد وقد أخرج نفسه من المسألة:...

تابع وكأنه يخاطب نفسه: أنت لم تمت يا عساف.

أخذ يحدث نفسه وهو في غيبوبة.

٣. أفعال الحدث، مثل: أدافع، أقف، تعرض، ومن أمثلتها في المدونة:

في تلك اللحظة، لم أكن أدافع عن نفسي.

فلماذا تعرض نفسها للعقاب؟



فجأة مررت أنت بنفس الممرالذي كنت أقف فيه .

ثم يجد المرء نفسه مثل ثمرة أُلقت بها شجرة .

وعلى الرغم من أن مجري (Mejri, 2022, p.93) قد رصدت تصاحب الأفعال النفسية وأفعال الحدث مع كلمة "نفس" في سياق لغة الصحافة العربية، إلا أنها لم تشر إلى تصاحبها مع أفعال القول، مما قد يدل على غياب هذا التصاحب في مدونتها الصحفية أو ندرته . ويمكن تفسير ذلك بأن لغة الرواية تعتمد على الحوار، ومن صور ذلك الحوار حوار الشخصية مع نفسها وحديثها مع ذاتها .

د . يلاحظ أن الفعل "وجد" مسنداً إلى تاء الفاعل (وجدتُ) هو الأقوى تصاحباً مع "نفس" مشكلة عبارة "وجدتُ نفسي" ؛ لذا سأقف عندها للنظر في سلوكها في لغة الرواية السعودية . من الناحية الكمية، فإن عبارة "وجدت نفسي" تكررت ١٥٨ مرة في ٣١ رواية من الروايات المضمنة في المدونة . وقد قمت باستخراج قائمة الكشاف السياقي لها ونظرت في الوظيفة التي تؤديها هذه العبارة، ويمكن القول إن هذه العبارة تستعمل في الرواية السعودية في سياق السرد الوصفي للأحداث والأحوال التي تواجه الشخصية الناطقة بها، وتشكّل هذه العبارة صورة من صور حديث الشخصية مع نفسها للتعبير عن الوجود الإدراكي (لا الحسي) لحالتها الراهنة في الرواية ووعيها بالمحيط الذي يحتضنها زماناً ومكاناً، ويبدو أن هذه العبارة تأتي في سياق الإدراك المفاجئ الذي يعتري المتكلم، ومما يدل على ذلك هو مجيء مفردة "فجأة" من بين الكلمات المتصاحبة لعبارة "وجدت نفسي" وتكرار وقوعها قبلها في عدد من السياقات المختلفة، وفيما يلي نماذج من استعمال هذا التركيب حتى تتضح دلالاته ووظيفته السياقية :

كلما عبرت الذكرى أو لاح وجه المحبوب ولو من وراء الغيوم وجدتُ نفسي لا شعورياً
أتعاطف مع هذا الشاب .

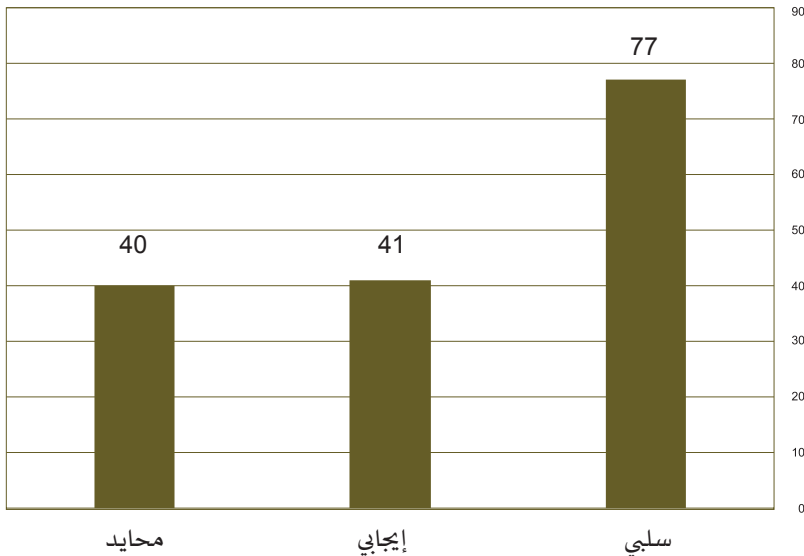
وجدت نفسي على مشارف الأربعين أبحث عن عمل جديد .

وفجأة وجدت نفسي أغرق في البكاء .

كانت أفكار مضطربة وقلبي يرتجف وفجأة وجدت نفسي أقف دون أفكر.

ولا تقتصر دلالة عبارة "وجدت نفسي" على مجرد الوصف، بل يمتد الأمر إلى وقوعها الاعتيادي في سياقات دلالية أو تداولية معينة تحمل إحياءات سلبية أحياناً وإيجابية أحياناً أخرى، وهذه الظاهرة تُعرف في حقل لسانيات المدونات بالنظم الدلالي "semantic prosody"، وتُعرف بتلك العلاقة القائمة والمتكررة استعمالاً بين كلمة /عبارة وكلمات /عبارات أخرى تتقارب دلالياً. بمعنى آخر، يتشكل النظم الدلالي من خلال تكرار كلمة أو عبارة في سياقات معينة، وتؤدي هذه العلاقة وظيفة الكشف عن مضامين إيجابية (سلبية أو إيجابية أو محايدة) (Liu, 2020, p.283; Zhou and Zhang, 2020, p. 247)، ويعدّ البحث في هذه الظاهرة من الغايات التي تسعى لسانيات المدونات إلى الكشف عنها، كونها تؤدي دوراً في تجلية اتجاهات الكاتب وميوله (العصيمي، ٢٠١٩، ص. ٣٨). وفي سياق اللغة الروائية، فإن هذه العبارة تجلي بعضاً من سمات الشخصية المتخيلة في رواية معينة، وذلك من خلال تتبع حديثها عن نفسها. وبعد فحص الكشف السياقي لعبارة "وجدت نفسي" في الرواية السعودية وجدت أنها غالباً ما تستعمل في السياقات ذات الدلالات السلبية وغير المرغوبة كما يعرضه شكل ٣:

شكل (٣) النظم الدلالي للتصاحب بين "وجدت نفسي" و "نفسى".





وتشمل الدلالات ذات الإيحاء السلبي التعبير عن الشعور بالعجز والخوف والحزن
والياس والحيرة والوحدة والغضب وغير ذلك، ومن أمثلة ذلك في المدونة:
وعندما حاولت أن أكتب جواباً عن سؤال أميرها وجدت نفسي عاجزاً وتتقاصر
كلماتي وتتضاءل.

وعندما غادرتهم وجدت نفسي حائراً.

وجدت نفسي وحيداً بلا أب بلا أقارب.

وليت الأمر قد انتهى عند هذا الحد، بل بعدها وجدت نفسي بلا مأوى.

ودون أن أفكر في شيء وجدت نفسي أبكي.

ولم يقتصر الأمر على الدلالة السلبية، فقد وردت العبارة في سياق الدلالة الإيجابية
والمحايدة (أي: مجرد وصف الحدث أو الحال الراهنة) لكن بنسب أقل، ومن أمثلة
ذلك في المدونة:

الإيجابية:

وجدت نفسي أذود عنهم أذى أهل القرية.

وجدت نفسي نشيطاً خفيف الحركة.

المحايدة:

وجدت نفسي في غرفة تشبه غرفة مكتبي في لندن.

وإذا وجدت نفسي خالياً مشيت إلى دار الحكمة.

بذلك نستطيع أن نستنتج أن هذا النظم الدلالي المنبثق عن التصاحب بين الفعل
"وجد" ومفردة "نفس" يشكل سمة من سمات اللغة السردية في مدونة الروايات
السعودية، ويمثل ركيزة تعبيرية لا يكاد الروائيون يستغنون عنها للتعبير عن معانٍ

محددة وأحوال معينة تحيط بالشخصيات الروائية، وقد لاحظنا فاعلية منهجية لسانيات المدونات وأدواتها في رصد هذه الظاهرة والتي لم تكن لترصد -أولنقل كان من الصعوبة أن تُرصد- لولا هذه المنهجية. وقد أكد غير واحد من المتخصصين على أن تتبّع مثل هذه الظواهر الدلالية يعتمد اعتماداً كبيراً على المدونات اللغوية الحاسوبية (Adolphs and Carter, 2002, p. 7; Hunston, 2002, p. 142; Liu, 2020, p. 283).

٥- الخاتمة

هدفتُ في هذا البحث إلى إلقاء الضوء على أهمية استخدام منهجية لسانيات المدونات الحاسوبية في تحليل اللغة الروائية، وذلك من خلال عرض جهود الغربيين في استثمار هذه المنهجية في تتبع عدد من الأنماط الموضوعية والأسلوبية في الروايات المكتوبة باللغة الإنجليزية. ثم عقبْتُ بعد ذلك بتطبيق هذه المنهجية وآلياتها في استكشاف ملامح اللغة المستعملة في الرواية السعودية تطبيقاً استكشافياً يبتعد قدر الإمكان عن الانطباع والاستنتاج الذاتي، وقد أثبت البحث أن هذه المنهجية لها إمكاناتها الحاسوبية المتقدمة في رصد الأنماط الخفية والمتعددة للغة الرواية السعودية وتشكّلها على تلك الصورة. وتمتاز المنهجية بأنها تعتمد على الأدلة الملموسة ذات السمة الإحصائية، وتسهم في الابتعاد -قدر الإمكان- عن الاعتماد الكلي على الانطباعات الشخصية والحدس التي قد تعتري الدراسات الأسلوبية.

وقد كشف لنا هذا البحث عن الأنماط التركيبية والدلالية للكلمات ذات الدلالة المعجمية (كلمات المحتوى) والمتتابعات اللفظية الأكثر شيوعاً، ومن تلك الأنماط، شيوع الفعل الدال على القول والأفعال الماضية، والمتتابعات ذات الدلالة الموجهية التي تجلّي موقف الشخصيات المتخيلة من الأحداث التي تدور حولها الروايات.

كما كشف هذا البحث عن السلوك التركيبي والتركيبى -المعجمي لكلمة "نفس"، الكلمة الأكثر شيوعاً في قائمة كلمات المحتوى في المدونة، فقد تجلّي لنا أن هذه الكلمة تتجه اتجاهاً شائعاً وملحوظاً إلى تأدية وظيفة الضمير المنعكس، مفرّغة من محتواها المعجمي الدالة



عليه أصالة كما تقرره المعاجم العربية. كما استبان لنا أن هذه الكلمة تتصاحب تصاحباً شائعاً مع الفعل "وجد" للدلالة على الوجود الإدراكي والنفسي، وهذا التصاحب يميل -كما كشف التحليل السياقي- إلى التعبير عن المعاني السلبية أكثر من تعبيره عن المعاني الإيجابية أو المحايدة.

وأختم هذا البحث بالتوصية بإجراء دراسات تعتمد على منهجية التحليل المدوني تمكننا من رصد الكلمات والمتتابعات اللفظية المميزة للروايات السعودية، وذلك من خلال مقارنتها بمدونات أخرى تختلف عنها في المجال والحجم، والمأمول أن ينتج عن مثل تلك الدراسات أننا سنقف على الملامح اللغوية والأسلوبية التي تبني الحدود الفارقة بينها وبين غيرها من أجناس النصوص. كما أوصي بإجراء دراسات معمقة لروائيين وروائيات سعوديين معينين؛ لرصد ملامح اللغة التي يستعملونها وتميزهم عن غيرهم، مما يساهم في الكشف عن بصماتهم الأسلوبية في اللغة السردية. ومن المقترحات البحثية التي أوصي بها تطبيق هذه المنهجية في تتبع أنماط اللغة الروائية السعودية في سياق عدد من المفاهيم (القيمات)، مثل: مفهوم الحب، ومفهوم البداوة، ومفهوم الاغتراب، وفي سياق تشكيل الشخصيات الروائية (characterization) كتتبع لغة الجسد لها.

المراجع العربية

- بجراوي، حسين. (١٩٩٠). تشكيل الخطاب الروائي. المركز الثقافي العربي.
- الحاج صالح، عبدالرحمن. (١٩٩٩). ورقة حول مشروع الذخيرة اللغوية، مجلة اللسان العربية. (٤٧) ١٠٧-١١٧.
- الحازمي، حسن. (٢٠٠٦). البناء الفني في الرواية السعودية: دراسة نقدية تطبيقية. مطابع الحميضي.
- زعلة، علي. (٢٠١٥). الخطاب السرد في روايات عبدالله الجفري. إصدارات النادي الأدبي الثقافي.
- شعبان، جمال. (٢٠٢١). الإنسانيات الرقمية: بحث في الماهية والمناهج، التقنيات والبنية التحتية. مجلة آفاق للبحوث والدراسات، المجلد ٤ (٢) ٥٣٣-٥٤٥.
- الشمري، عقيل والثبيتي، عبدالمحسن. (٢٠١٩). استخدام المدونات الحاسوبية في التحليل اللغوي: نماذج من لغة الصحافة العربية. مجلة القراءة والمعرفة، المجلد ١٩ (٢٠٧) ٨٩-١٣٧.
- الشمري، عقيل والمحمود، محمود. (٢٠١٦). المعالجة الآلية لوعاء الأخبار: تحليل الخطاب النقدي المعتمد على المدونة الحاسوبية. في سلطان المكيول (محرر) لغويات المدونة الحاسوبية: تحليلات تطبيقية على العربية الطبيعية (ص ص. ١٩٨-٢٥٢). مركز الملك عبدالله لخدمة اللغة العربية.
- الشمسان، إبراهيم. (١٩٨١). الجملة الشرطية عند النحاة العرب. مطابع الدجوي.
- صالح، محمود. (٢٠١٥). المدونات اللغوية وكيفية الإفادة منها. في صالح العصيمي (محرر) المدونات اللغوية العربية وطرق الإفادة منها (ص ص. ١٧-٩٤). مركز الملك عبدالله لخدم اللغة العربية.



العتيبي، هند. (٢٠١٦). المعالجة الآلية للمدونات المتوازية واستخداماتها في تعليم اللغات وتدريب المترجمين. في سلطان المجيول (محرر) لغويات المدونة الحاسوبية: تحليلات تطبيقية على العربية الطبيعية. (ص ص. ١٧٠-١٩٧). مركز الملك عبدالله لخدمة اللغة العربية.

العصيمي، صالح (٢٠١٩). النظم الدلالي والتفضيل الدلالي: دراسة تأصيلية وتطبيقية. مجلة العلوم الإنسانية، جامعة البحرين، (٣٣) ٣٦-٦٥.

العصيمي، صالح. (٢٠١٨). لسانيات المتون: قضايا أساسية في التأصيل والتطبيق والمنهج. مركز الملك عبدالله لخدمة اللغة العربية.

الفيضي، عبدالله. (٢٠١٦). المعالجة الآلية لمدونات المتعلمين: تحليل الأخطاء في المدونة اللغوية لتعلمي اللغة العربية. في سلطان المجيول (محرر) لغويات المدونة الحاسوبية: تحليلات تطبيقية على العربية الطبيعية. (ص ص. ١٣٦-١٦٤). مركز الملك عبدالله لخدمة اللغة العربية.

قاسم، سيزا. (٢٠٠٤). بناء الرواية. مطبعة مكتبة الأسرة.

القاسمي، علي. (٢٠٠٦). لسانيات المدونة الحاسوبية وصناعة المعجم العربي. المؤتمر السنوي الخامس: اللغة العربية في عصر المعلوماتية. ١-٤٦.

القاضي، محمد وآخرون. (٢٠١٠). معجم السرديات. دار محمد علي للنشر.

قسومة، الصادق. (٢٠٢٢). الحوار في الخطاب القصصي: خلفياته وآلياته وقضايا. مسكيلياني للنشر والتوزيع.

المتوكل، أحمد. (٢٠١٣). قضايا اللغة العربية في اللسانيات العربية. منشورات دار ضفاف.

المجيول، سلطان. (٢٠١٧). أسس تحليل التصاحب اللفظي في المدونة اللغوية العربية. مجلة الآداب بجامعة الملك سعود. المجلد: ٢٩ (٢) ٣-٣٠.

المجبول، سلطان. (٢٠١٨). البحث اللغوي الآلي في المدونة الحاسوبية واللغة العربية. مجلة التواصل اللساني، (١٨) ٥٣-٧٨.

مرتاض، عبد الملك. (١٩٩٧). في نظرية الرواية. عالم المعرفة.

نحلة، محمود. (٢٠٠٩). علم اللغة النظامي: مدخل إلى النظرية اللغوية عند هاليداي. دار المعرفة الجامعية.

الهرساني، عبد الرحمن. (٢٠١٨). معجم سياقي لألفاظ الطعام في العربية المعاصرة للناطقين بغير العربية (من حرف الهمزة إلى حرف الصاد). [أطروحة دكتوراه غير منشورة]، معهد تعليم اللغة العربية لغير الناطقين بها، الجامعة الإسلامية بالمدينة المنورة.

رومنة المراجع العربية

Bahrawi, H. (1990). *Tashkil al-khitab al-riwai*. Al-Markaz al-Thaqafi al-Arabi.

Al-Hajj Salih, A. (1999). Waraqa hawla mashru' al-dhakhira al-lughawiyya. *Majallat al-Lisan al-'Arabiyya*, (47), 107-117.

Al-Hazimi, H. (2006). *Al-bina' al-fanni fi al-riwaya al-Sa'udiyya: Dirasah naqdiyya tatbiqiyya*. Matabi' al-Humaidi.

Za'lah, A. (2015). *Al-khitab al-sardi fi riwayat 'Abd Allah al-Jifri*. Isdarat al-Nadi al-Adabi al-Thaqafi.

Sha'ban, J. (2021). Al-insaniyyat al-raqamiyya: Bahth fi al-mahiyya wa-al-manahij, al-taqniyyat wa-al-bunyah al-tahtiyya. *Afaq li-al-Buhuth wa-al-Dirasat*, 4(2), 533-545.

Al-Shammari, A., & Al-Mahmud, M. (2016). Al-mu'alajah al-aliyya li-wi'a' al-akhbar: Tahlil al-khitab al-naqdi al-mu'tamid 'ala al-maduna al-hasubiyya. In S. Al-Mujaywul (Ed.), *Lughawiyyat al-maduna al-hasubiyya: Tahlilat tatbiqiyya 'ala al-'Arabiyya al-tabi'iyya* (pp. 198-252). Markaz al-Malik 'Abd Allah li-Khidmat al-Lugha al-'Arabiyya.

Al-Shammari, A., & Al-Thubayti, A. (2019). Istikhdam al-madunat al-hasubiyya fi al-tahlil al-lughawi: Namadhij min lughat al-sahafa al-'Arabiyya. *Majallat al-Qira'a wa-al-Ma'rifa*, 19(207), 89-137.



- Al-Shamsan, I. (1981). *Al-jumla al-shartiyya 'inda al-nuhat al-'Arab*. Matabi' al-Dajawi.
- Salih, M. (2015). Al-madunat al-lughawiyya wa-kayfiyyat al-ifada minha. In S. Al-'Asimi (Ed.), *Al-madunat al-lughawiyya al-'Arabiyya wa-turuq al-ifada minha* (pp. 17–94). Markaz al-Malik 'Abd Allah li-Khidmat al-Lugha al-'Arabiyya.
- Al-'Utaybi, H. (2016). Al-mu'alajah al-aliyya li-al-madunat al-mutawaziyya wa-istikhdamatuhā fi ta'lim al-lughat wa-tadrib al-mutargimin. In S. Al-Mujaywul (Ed.), *Lughawiyyat al-maduna al-hasubiyya: Tahlilat tatbiqiyya 'ala al-'Arabiyya al-tabi'iyya* (pp. 170–197). Markaz al-Malik 'Abd Allah li-Khidmat al-Lugha al-'Arabiyya.
- Al-'Asimi, S. (2018). *Lisanayat al-mutun: Qadaya asasiyya fi al-ta'sil wa-al-tatbiq wa-al-manhaj*. Markaz al-Malik 'Abd Allah li-Khidmat al-Lugha al-'Arabiyya.
- Al-'Asimi, S. (2019). Al-nizam al-dilali wa-al-tafdil al-dilali: Dirasah ta'siliyya wa-tatbiqiyya. *Majallat al-'Ulum al-Insaniyya, Jami'at al-Bahrayn*, (33), 36–65.
- Al-Fifi, A. (2016). Al-mu'alajah al-aliyya li-madunat al-muta'allimin: Tahlil al-akhta' fi al-maduna al-lughawiyya li-muta'allimi al-lugha al-'Arabiyya. In S. Al-Mujaywul (Ed.), *Lughawiyyat al-maduna al-hasubiyya: Tahlilat tatbiqiyya 'ala al-'Arabiyya al-tabi'iyya* (pp. 136–164). Markaz al-Malik 'Abd Allah li-Khidmat al-Lugha al-'Arabiyya.
- Qasim, S. (2004). *Bina' al-riwaya*. Matba'at Maktabat al-Usra.
- Al-Qasimi, A. (2006). Lisanayat al-maduna al-hasubiyya wa-sina'at al-mu'jam al-'Arabi. *Al-Mu'tamar al-Sanawi al-Khamis: Al-Lugha al-'Arabiyya fi 'Asr al-Ma'lumatiyya* (pp. 1–46).
- Al-Qadi, M., et al. (2010). *Mu'jam al-sardiyyat*. Dar Muhammad 'Ali lil-Nashr.
- Qasuma, S. (2022). *Al-hiwar fi al-khitab al-qasasi: Khalfiyyatuhu, aliyatuhu wa-qadaya-hu*. Maskiliyani lil-Nashr wa-al-Tawzi'.
- Al-Mutawakkil, A. (2013). *Qadaya al-lugha al-'Arabiyya fi al-lisanayat al-'Arabiyya*. Manshurat Dar Difaf.
- Al-Mujaywul, S. (2017). Usus tahlil al-tasahub al-lafzi fi al-maduna al-lughawiyya al-'Arabiyya. *Majallat al-Adab, Jami'at al-Malik Sa'ud*, 29(2), 3–30.

- Al-Mujaywul, S. (2018). Al-bahth al-lughawi al-ali fi al-maduna al-hasubiyya wa-al-lugha al-
‘Arabiyya. *Al-Tawasul al-Lisani*, (18), 53–78.
- Murtad, ‘A. (1997). *Fi nazariyyat al-riwaya*. ‘Alam al-Ma‘rifa.
- Nahla, M. (2009). ‘*Ilm al-lugha al-nizami: Madkhal ila al-nazariyya al-lughawiyya ‘inda Halliday*. Dar al-Ma‘rifa al-Jami‘iyya.
- Al-Harsani, ‘A. (2018). *Mu‘jam siyaqi li-alfaz al-ta‘am fi al-‘Arabiyya al-mu‘asira li-al-natiqin bi-ghayr al-‘Arabiyya (min harf al-hamza ila harf al-sad)* [Unpublished doctoral dissertation]. Ma‘had Ta‘lim al-Lugha al-‘Arabiyya li-Ghayr al-Natiqin Biha, al-Jami‘a al-Islamiyya bi-al-Madina al-Munawwara.

المراجع الأجنبية

- Abdul-Baqee, S., & Atwell, E. S. (2009). Knowledge representation of the Quran through frame semantics: A corpus-based approach. In Proceedings of the Fifth Corpus Linguistics Conference. University of Liverpool.
- Alansary, S., & Nagi, M. (2014). The international corpus of Arabic: Compilation, analysis and evaluation. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 8–17). Doha, Qatar.
- Alfaifi, A., & Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, 19, 347–357.
- Alfraidi, T., Abdeen, M. A. R., Yatimi, A., Alluhaibi, R., & Al-Thubaity, A. (2022). The Saudi Novel Corpus: Design and compilation. *Applied Sciences*, 12, 1–12.
- Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). New language resources for Arabic: Corpus containing more than two million words and a corpus processing tool. In Proceedings of the International Conference on Asian Language Processing (pp. 67–70). Urumqi, China.
- Al-Thubaity, A. O. A. (2015). 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49, 721–751.



- Ancarno, C. (2020). Corpus-assisted discourse studies. In A. De Fina & A. Georgakopoulou (Eds.), The Cambridge handbook of discourse studies. Cambridge University Press.
- Badawi, E. M., Carter, M. G., & Gully, A. (2004). Modern written Arabic: A comprehensive grammar. London: Routledge.
- Baker, P. (2006). Using corpora in discourse analysis. London: Continuum.
- Baker, P. (2010). Sociolinguistics and corpus linguistics. Edinburgh University Press.
- Baker, P., Hardie, A., & McEnery, T. (2006). A glossary of corpus linguistics. Edinburgh University Press.
- Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., & Suchomel, V. (2013). arTenTen: A new, vast corpus for Arabic. In Proceedings of WACL, 20.
- Biber, D. (2006). University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.
- Biber, D. (2011). Corpus linguistics and the study of literature: Back to the future? Scientific Study of Literature, 1(1), 15–23.
- Biber, D., & Conrad, S. (2004). Corpus-based comparisons of registers. In C. Coffin, A. Hewings, & K. O'Halloran (Eds.), Applying English grammar: Functional and corpus approaches (pp. 40–56). London: Arnold.
- Biber, D., & Conrad, S. (2009). Register, genre, and style. Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge University Press.
- Carter, R. (2010). Methodologies for stylistic analysis: Practices and pedagogies. Language and Style, 55–68.
- Celce-Murcia, M., Larsen-Freeman, D., & Williams, H. A. (1983). The grammar book: An ESL/EFL teacher's course. Rowley, MA: Newbury House.
- Conrad, S. M., & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. Lexicographica, 20, 56–71.

- Cuddon, J. A. (2013). A dictionary of literary terms and literary theory. Oxford: John Wiley and Sons.
- Fairclough, N. (2003). Analysing discourse: Textual analysis for social research. London: Routledge.
- Fischer-Starcke, B. (2010). Corpus linguistics in literary analysis: Jane Austen and her contemporaries. London: Bloomsbury Publishing.
- Freihat, A. A., Bella, G., Mubarak, H., & Giunchiglia, F. (2018). A single-model approach for Arabic segmentation, POS tagging, and named entity recognition. In Proceedings of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 1–8). Algeria: IEEE.
- Green, C. (2017). Introducing the Corpus of the Canon of Western Literature: A corpus for culturomics and stylistics. *Language and Literature*, 26(4), 282–299.
- Hafissatou, K. A. N. E. (2021). A contrastive study of reflexive verbs in English and French. *Journal of English Language Teaching and Applied Linguistics*, 3(4), 25–31.
- Holes, C. (2004). Modern Arabic: Structures, functions and varieties. Washington, DC: Georgetown University Press.
- Hunston, S. (2002). Corpora and applied linguistics. Cambridge University Press.
- Hunston, S. (2010). How can a corpus be used to explore patterns? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 140–154). London: Routledge.
- Hussein, K. S. (2015). The potentialities of corpus-based techniques for analysing literature. *Al-Ma’mun College Journal*, (25), 348–367.
- Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.
- Hyland, K., & Jiang, F. (2018). Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics*, 23(4), 383–407.



- Jeyanthi, E., Zalina, K. M., Afida, A., & Arbaayah, A. T. (2018). The hunger games: Linguistic modality as reflector of point of view in a dystopian novel. *Pertanika Journal of Social Sciences & Humanities*, 26(4), 2395–2415.
- Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 387–400). London: Routledge.
- Jucker, A. H. (2021). Features of orality in the language of fiction: A corpus-based investigation. *Language and Literature*, 30(4), 341–360.
- Kestemont, M. (2014). Function words in authorship attribution: From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 59–66).
- Leech, G. N., & Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. Pearson Education.
- Liu, W. (2020). A corpus-based study of semantic prosody and semantic preference of “seem”. *Open Journal of Social Sciences*, 8(5), 282–291.
- Mahlberg, M. (2007a). Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1), 1–31.
- Mahlberg, M. (2007b). Corpus stylistics: Bridging the gap between linguistic and literary studies. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis*, 8 (pp. 219–246). London: Continuum.
- Mahlberg, M. (2012). The corpus stylistic analysis of fiction—or the fiction of corpus stylistics? In M. Huber & J. Mukherjee (Eds.), *Corpus linguistics and variation in English* (pp. 77–95). Amsterdam: Rodopi.
- Mahlberg, M. (2013). *Corpus stylistics and Dickens’s fiction*. London: Routledge.
- Mahlberg, M. (2014). Corpus stylistics. In M. Burke (Ed.), *The Routledge handbook of stylistics* (pp. 396–410). London: Routledge.

- Mahlberg, M., Biber, D., & Reppen, R. (2015). Literary style and literary texts. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 346–361). Cambridge University Press.
- Mahlberg, M., Stockwell, P., Joode, J. D., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 11, 433–463.
- Mahlberg, M., Wiegand, V., Stockwell, P., & Hennessey, A. (2019). Speech-bundles in the 19th-century English novel. *Language and Literature*, 28(4), 326–353.
- Mahmood, R., Parveen, S., Shah, S. K., & Batool, A. (2013). A corpus driven comparative study of representation of 'man' and 'woman' in Pakistani and British English fictions. *Research on Humanities and Social Sciences*, 3(11), 95–105.
- Maiwald, P. (2011). Exploring a corpus of George Macdonald's fiction. *North Wind: A Journal of George MacDonald Studies*, 30(5), 50–84.
- McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3–13). London: Routledge.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McIntyre, D., & Walker, B. (2019). *Corpus stylistics: Theory and practice*. Edinburgh University Press.
- Mejri, A. (2022). The distribution of nafs in Modern Standard Arabic and Classical Arabic: A corpus-based study. *Saudi Journal of Language Studies*, 2(2), 84–106.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.
- Mikhailov, M., & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: A guide for research*. London: Routledge.



- Newman, D. L. (2013). The Arabic literary. In J. Owens (Ed.), The Oxford handbook of Arabic linguistics (pp. 472–494). Oxford University Press.
- Paterson, L. (2018). “You can just give those documents to myself”: Untriggered reflexive pronouns in 21st century spoken British English. In V. Brezina, R. Love, & K. Aijmer (Eds.), Corpus approaches to contemporary British speech: Sociolinguistic studies of the spoken (BNC2014) (pp. 235–255). London: Routledge.
- Senthok Singh, A. K., & Ang, L. H. (2019). Lexical bundles in selected children’s fiction: A corpus-based analysis. Journal of Creative Practices in Language Learning and Teaching (CPLT), 7(2), 12–28.
- Shalaby, M. (2016). Orality in fiction dialogue: A discourse analysis and corpus-assisted study in English and Arabic novels. Logos, 10(10), 203–252.
- Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. Functions of Language, 22(3), 362–399.
- Stefanowitsch, A. (2020). Corpus linguistics: A guide to the methodology. The Hague: Language Science Press.
- Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. Language and Literature, 14, 5–24.
- Van Mol, M. (2003). Variation in Modern Standard Arabic in radio news broadcasts: A synchronic descriptive investigation into the use of complementary particles. Leuven: Peeters Publishers.
- Versteegh, K. (2014). Arabic language. Edinburgh University Press.
- Wijitsopon, R. (2013). Corpus-based study of the style in Jane Austen’s novels. Manusya Journal of Humanities, 16, 41–64.
- Wright, D. (2021). Corpus approaches to forensic linguistics. In M. Coulthard, A. May, & R. Sousa-Silva (Eds.), The Routledge handbook of forensic linguistics (pp. 611–627). London: Routledge.

- Wynne, M. (2005). Stylistics: Corpus approaches. In K. Brown (Ed.), Encyclopaedia of language and linguistics. Oxford: Elsevier.
- Yelena Mkhitarian, & Loretta Bazikyan. (2016). On structure and semantics of reflexive constructions in English. Linguistics and Literature Studies, 4(6), 428–432. <https://doi.org/10.13189/lis.2016.040606>
- Zhou, Q., & Zhang, L. (2020). Corpus-based approach to explore the semantic prosody of synonym: A case study of “Lead to” and “Result in”. US-China Foreign Language, 18(8), 247–252.



نبذة تعريفية عن الباحث

Author Biodata Tareq Rubaye Alfraid, Associate Professor in Linguistics at the Department of Linguistics, College of Arabic Language and Humanities, Islamic University of Madinah. He received his PhD degree in 2017 from Exeter University (UK). His research interests include General Linguistics, Functional Linguistics, Corpus Linguistics, and Discourse Analysis.	بيانات الباحث: طارق بن ربيع خلف الفريدي، أستاذ مشارك في اللسانيات في قسم اللغويات بكلية اللغة العربية والدراسات الإنسانية في الجامعة الإسلامية بالمدينة المنورة، المملكة العربية السعودية، حاصل على درجة الدكتوراه في اللغويات الحديثة من جامعة اكستر عام ٢٠١٧، تدور اهتماماته البحثية حول اللسانيات العامة واللسانيات الوظيفية والمدونات اللغوية الحاسوبية وتحليل الخطاب.
معرف أوركيد (ORCID) : 5415-8962-0002-0000	
Email: t.alfraid@iu.edu.sa	



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