



اللُّغَوِيَّاتُ وَاللُّغَةُ الْعَرَبِيَّةُ

وَالْمَعْرِفَةُ الَّلُّغِيَّةُ الْعَرَبِيَّةُ

Journal of Computational Linguistics & Arabic Language Processing



الكشف عن سمية النصوص وقياسها في نماذج توليد
النصوص العربية في سياق أنظمة البحث والتوصية

المصادر المعجمية العربية مدونةً محوسبةً: تحليلها، توسيمها،
دمجها لبناء مصدر معجمي دلالي موسّع مقروء آلياً

المعايير العالمية المستخدمة في صناعة المعاجم الرقمية:
تحليل مقارنة بين LMF و TEI

تعزيز البيانات النصية من أجل التعامل مع المفردات غير المرصودة
في النماذج اللغوية في مهام معالجة اللغة العربية المعاصرة

«مدار»: موارد حاسوبية ولهجات العربية

تحديد اللغات الأصلية لمُتعلِّمي اللغة العربية غير الناطقين بها

السنة الأولى - العدد الثاني
ربيع الثاني ١٤٤٦ هـ - أكتوبر ٢٠٢٤ م

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



اللغويات والحسابية والمعاصرة الجاهلية الإسلامية العربية

Journal of Computational Linguistics & Arabic Language Processing

www.ksaa.gov.sa

مجلة علمية مُحكمة تصدر عن مجمع الملك سلمان العالمي للغة العربية





المشرف العام على المجلة

أ.د. عبدالله بن صالح الوشمي

هيئة التحرير

د. سارة الحمود

د. عريب العويشق

د. عبدالعزيز المهيوبي

د. أيمن الغامدي

إدارة التحرير

أ.د. منصور الغامدي

رئيس التحرير

الهيئة الاستشارية

أ.د. إيريك أتويل

أ.د. نزار حبش

أ.د. خالد فؤاد شعلان

أ.د. صلاح راشد الناجم

أ.د. عبد المحسن عبيد الثبيتي

أ.د. محسن عبد الرزاق رشوان

أ.د. مصطفى عرسان جرار

د. عبد الرحمن العصيمي

مدير التحرير

أ. هاجر القبيشي

أمين المجلة



◆ قواعد النشر بالمجلة:

سياسات النشر:

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

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



Journal information:

◆ بيانات المجلة:

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

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

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

الرياض - المملكة العربية السعودية Riyadh- Saudi Arabia

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

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

تعليمات عامة:

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

٢. يُرسل نموذج تعهد من المؤلّف بأنّ الورقة لم يسبق نشرها، وأنها غير مُقدّمة للنشر، ولن تُقدّم للنشر في جهة أخرى.

٣. يجب أن تحتوي الورقة على ما يلي:

١-٣- صفحة أولى فيها عنوان الورقة، واسم المؤلّف أو المؤلّفين، وجهة عملهم، وعناوينهم، وتاريخ تقديم الطلب بنشر الورقة.

٢-٣- الكلمات المفتاحية.

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

٤-٣- المصادر والمراجع.

٥-٣- الملاحق -إن وجدت-.

٤. 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.

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

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 متوسطًا بين حدوده بحيث لا يتجاوز عنوان الشكل سطرين فقط.
- في حال وجود تعليق مصاحب على الشكل أو الصورة يجب دمجها 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 «...»، ويُكتب بحجم ١٢ للنص العربي و١٠ للنص الإنجليزي، وينتهي بتوثيق النص على النحو التالي: (العائلة، السنة: الصفحة).

مواصفات كتابة الحواشي:

- تُدرج الحواشي في نهاية البحث، على أن تُرقم ترقيمًا آليًا حسب تسلسل ورودها في النص.
- أن يكتب البحث بلغة عربية سليمة.
- أن يكون البحث أو الدراسة في أحد مجالات تعليم العربية.

- أن يعتمد الباحث الأسس العلمية في خطواته كافة، وأن يتبع قواعد التوثيق العلمي المستخدمة بالمجلة وهي طريقة جمعية علم النفس الأمريكية APA.
- أن تكون عدد كلمات البحث ما بين ٥٠٠٠ إلى ١٠,٠٠٠ كلمة.
- ألا يسبق نشر البحث في أي وعاء علمي - مالم ترهيئة التحرير خلاف ذلك -.
- أن يعد الباحث ملخصاً لبحثه في حدود ٢٠٠ كلمة باللغة العربية.
- كما ترحب المجلة بالاستعراض العلمي للكتب ذات العلاقة باهتمامات المجلة.

◆ صفة المجلة:

مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية (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 للغة 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 اللغة، وتشمل: اللغويات الحاسوبية (Computational Linguistics)، والمعالجة الآلية للغة العربية (Arabic Language Processing) وما يتبعها من تخصصات متصلة 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 ودقة المعلومات والاستنتاجات.

محتويات العدد

١٢

الكشف عن سمية النصوص وقياسها في نماذج توليد النصوص العربية في سياق أنظمة البحث والتوصية

فوزي حراق

٣٩

المصادر المعجمية العربية مدونةً محوسبةً: تحليلها، توسيمها، دمجها لبناء مصدر معجمي دلالي موسّع
مقروء آليًا

عبد العاطي إبراهيم هوارى

٧٣

المعايير العالمية المستخدمة في صناعة المعاجم الرقمية: تحليل مقارنة بين LMF و TEI

عمرو الجندي

٩٢

تعزيز البيانات النصية من أجل التعامل مع المفردات غير المرصودة في النماذج اللغوية في مهام معالجة اللغة العربية المعاصرة

هدى زيتون - سامح الأنصاري

١١٠

«مدار»: موارد حاسوبية اللهجات العربية

هدى بوعمر - نزار حبش - سلام خليفة - فضل الإرياني

١٣٤

تحديد اللغات الأصلية لمتعلمي اللغة العربية غير الناطقين بها

مراد عباس - محمد أمين شراقي - محمد مدياني

كلمة التحرير

The Journal of Computational Linguistics and Arabic Language Processing (JCLALP) is pleased to present to researchers and its readers its second issue hoping that they will find in it what satisfies some of their curiosity to learn about the latest developments in scientific research in the fields of the JCLALP. Aiming that its content will motivate them for further research and studies to advance the Arabic language processing.

This issue coincides with the Academy's Third Conference, Arabic Language Computing and Linguistic Data Enrichment. It contains a range of research papers and studies presented at the conference, including several focused on topics related to the link between the vocabulary of Arabic dialects and the Modern Standard Arabic. This field is rich in diversity all linguistic levels, starting with the phonetic to the semantic level. The issue includes the results of research related to programming, developing, labeling, and analyzing dictionaries. The Arabic dictionary needs more efforts to document, preserve, and make it available to users. Therefore, they can benefit from it by employing it in their daily uses. It needs to be available for researchers and developers to improve its performance and enrich its content.

As JCLALP progresses by publishing its second issue, it looks forward to receiving more of your research results. By submitting your manuscript, you will be serving the Arabic language, and making it a strong competitor to international languages in its presence at the level of many applications, most notably artificial intelligence. Ultimately, you will be serving Arabic speakers by enhancing their scientific presence in their daily work.

يسر مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية أن تقدم للباحثين والمهتمين عددها الثاني على أمل أن يجدوا فيه ما يشبع شئنا من فضولهم للاطلاع على آخر مستجدات البحوث العلمية في مجالات المجلة، وأن يكون محتواه محفزاً لهم لمزيد من البحوث والدراسات للنهوض باللغة العربية تقنياً.

ويصدر هذا العدد تزامناً مع مؤتمر المجمع الثالث (حوسبة اللغة العربية وإثراء البيانات اللغوية). ويضم عدداً من البحوث والدراسات التي نوقشت في المؤتمر منها ما يتعلق بالربط بين مفردات اللهجات العربية واللغة العربية الفصحى هذا الميدان الخصب خصوبة تعدد وتنوع اللهجات العربية وما يتعلق بها على جميع مستويات اللغة، من المستوى الصوتي إلى الدلالي. ومنها ما يتعلق بالمعاجم برمجة وبناء وتوسيعاً وتحليلاً، المعجم العربي الذي لا يزال يحتاج إلى مزيد من الجهود لتوثيقه وحفظه وإتاحته للمستخدمين ليتمكنوا من الاستفادة منه وتوظيفه في استخداماتهم اليومية، وإتاحته للباحثين والمطورين لتحسين أدائه وإثراء محتواه.

وهي تقدم لكم عددها الثاني، تتطلع المجلة إلى نشر المزيد من نتائج بحوثكم فيما يخدم اللغة العربية، مما يجعلها منافساً قوياً للغات العالمية في حضورها على مستوى التطبيقات التقنية المتعددة وعلى رأسها الذكاء الاصطناعي، فيما يخدم المتحدث العربي في نهاية المطاف ويعزز من حضوره العلمي في حياته الخاصة وحياته العملية.

الكشف عن سمية النصوص وقياسها في نماذج توليد النصوص العربية في سياق أنظمة البحث والتوصية*

Toxicity Detection of Arabic Generative Models in Search and Recommendation Systems*

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

فوزي حراق

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

توثيق البحث: APA Citation

حراق، فوزي. (2024). الكشف عن سمية النصوص وقياسها في نماذج توليد النصوص العربية في سياق أنظمة البحث والتوصية. مجلة اللسانيات العربية، (٢).

Submission Date: 25/07/2024, Acceptance Date: 19/09/2024

تاريخ الإرسال: ٢٥/٧/٢٠٢٤ / تاريخ القبول: ١٩/٩/٢٠٢٤ م

Abstract

Pre-training large language models on vast amounts of web-scraped texts is a current trend in natural language processing. While the resulting models are capable of generating convincing text, they also reproduce harmful social issues such as bias and toxicity. This paper explores expressions of toxicity in Arabic text generation. Analyses will be performed on samples produced by the generative model AraGPT2 (a GPT-2 fine-tuned for Arabic). An Arabic Toxicity Classifier (ATC) based on new transformer model AraBERT (a BERT fine-tuned for Arabic) will be used to capture the toxicity of an Arabic-generated sentence or text. For the development of this classifier, a dataset will be crowd-sourced, cleaned, and independently annotated. AraGPT2 will be used to generate more toxic descriptions from the standard prompts. Our Toxicity Detection Model (TDM) will be based on the combination of the two transformers (AraGPT2-AraBERT) models. Our proposed model has achieved very encouraging results by reaching an accuracy percentage of 81%.

Keywords: Toxicity detection, Arabic Generative Models, Transformers, Arabic Natural Language Understanding (NLU), Search, Recommendation Systems.

الملخص

يعد التدريب المسبق لنماذج اللغات الكبيرة على كميات هائلة من النصوص المحذوفة على الويب هو الاتجاه الحالي في معالجة اللغة الطبيعية. وفي حين أن النماذج الناتجة قادرة على توليد نص مقنع، فإنها تنتج أيضا إشكاليات اجتماعية ضارة ومن أهمها التحيز وسمية النصوص. يستكشف هذا البحث التعبيرات الدالة على السمية في توليد النص العربي. ستجرى التحليلات على العينات التي ينتجها النموذج التوليدي AraGPT2. وسيستخدم مصنف لاستكشاف سمية النص العربي (Toxicity Classifier) استنادًا إلى نموذج المحول الجديد Ara-BERT لالتقاط سمية الجملة أو النص المنشأ باللغة العربية. من أجل تطوير هذا المصنف، ستجمع مجموعة البيانات من مصادر جماعية وتنظف ويجري توسيمها بشكل مستقل. سيستخدم AraGPT2 لإنشاء أوصاف أكثر سمية من المطالبات القياسية. سيعتمد نموذج كشف السمية (TDM) الخاص بنا على مزيج من نمودجي المحولات (AraGPT2-AraBERT). لقد حقق نموذجنا المقترح نتائج مشجعة للغاية حيث وصل إلى نسبة دقة بلغت 81%.

الكلمات المفتاحية: الكشف عن السمية، نماذج التوليد العربية، المحولات، فهم اللغة الطبيعية العربية (NLU)، البحث، أنظمة التوصية.



١. المقدمة

حاليًا، يعد الذكاء الاصطناعي (AI) والتعلم الآلي (ML) جزءًا لا يتجزأ من حياتنا اليومية. نحن نعلم بشكل كبير على الخوارزميات الذكية لاتخاذ العديد من القرارات نيابة عنا، من تغذية الأخبار التي نتلقاها، إلى تصفية البريد العشوائي، بالإضافة إلى الإعلانات، والترجمة اللغوية، وتصنيف الصور، إلى التحكم في حركة المرور، وتحديد من له الأولوية في الحصول على القرض، والقائمة تطول. إن التحيزات البشرية لا مفر منها بسبب الطبيعة البشرية، أما التحيز الخوارزمي فيجب أن يكون قابلاً للإدارة. حيث إن أنظمة الذكاء الاصطناعي هي في الأساس أنظمة ذكية قادرة على محاكاة الذكاء البشري في أداء المهام الموكلة إليها. الأنظمة البرمجية المستقلة هي أيضاً أنظمة ذكاء اصطناعي تتمتع بالاستقلالية الكاملة في تشغيلها. وعليه، فإنها قادرة على اتخاذ قرارات معقدة دون تدخل بشري. يبدو أن أنظمة الذكاء الاصطناعي أصبحت عناوين رئيسية دون أي أسباب مقنعة أو واضحة، وذلك بسبب التحيز، بعد سلسلة من المشاكل التي تسببت فيها هذه الأنظمة. كانت البداية التقريبية لهذه السلسلة من الأخبار في عام ٢٠١٧، عندما اعتذرت فيسبوك عن خطأ في خدمة ترجمة المنشورات التي تسببت في اعتقال الشرطة الإسرائيلية لرجل فلسطيني بعد أن نشر منشوراً بالعربية يحتوي على كلمة "صباح الخير" على صفحته على فيسبوك، وترجمت خدمة فيسبوك الجملة إلى "اقتلوهم". تميل أنظمة الذكاء الاصطناعي أيضاً إلى ربط كلمات معينة بشكل غير عادل بمجموعة محددة من المواضيع، وهذا هو السبب في أن FaceApp اعتذرت عن بناء ذكاء اصطناعي كان عنصرياً يربط ميزة الجاذبية بالبياض. عندما ينقر المستخدم على زر الجاذبية في FaceApp لجعل الصورة جذابة، تلقائياً يجعل التطبيق الوجه في الصورة أبيضاً. وعليه، خلق الانطباع بأن الوجه الأبيض فقط هو الجذاب بغض النظر عن لون البشرة الآخر (Hern, 2017). في هذا السياق، حذر Caliskan et al. (2017)، "أن الذكاء الاصطناعي لديه القدرة على تعزيز التحيزات القائمة لأن الخوارزميات، وعلى عكس البشر، قد لا تكون مجهزة للتصدي بوعي للتحيزات المكتسبة". وبالإشارة إلى نتائج (Zhou, 2021)، يتضح أن نسبة "الرئيس التنفيذي" من الإناث هي فقط ١١٪ في صور Google، بينما في الولايات المتحدة هناك

أكثر من ٢٧٪ من الرؤساء التنفيذيين من الإناث، ونفس الملاحظات لصور "الأطباء"، التي يغلب عليها الذكور؛ في حين أن معظم صور "الممرضات" هي للإناث. في الواقع، هذه البيانات لا تعكس الواقع. مثال آخر يمكننا معالجته، هو الخطأ في دقة نظام COMPAS في مجال الجريمة، حيث عثر على أن هذا النظام متساهل جداً فيما يتعلق بالبيض وقاسي جداً فيما يتعلق بالسود. وفي (Pasquale, 2015)، اتضح أن خوارزميات التعلم الآلي هي "صناديق سوداء" محمية ببروتوكولات سرية صناعية وحمايات قانونية، بالإضافة إلى التمييز المتعمد من قبل الشركات الكبيرة. وعليه، يصبح التمييز ضد التحيز غير مرئي، بينما يصبح تقليله شبه مستحيل. وبينما تركز معظم الأبحاث بشكل أساسي على الكشف عن التحيزات في الخوارزميات أو قياسها، القليل فقط من الأعمال -حتى الآن- كرس الاهتمام لفهم أسباب التحيز المكتشف في الخوارزميات والنماذج التوليدية. من ناحية أخرى، أظهرت العديد من الدراسات الحديثة أنه على الرغم من أن خوارزميات التوصية تحقق دقة تنبؤية جيدة وعالية، فإن العديد منها يعاني من التحيزات المجتمعية. حللت هذه التأثيرات على نطاق واسع فيما يتعلق باللغة الإنجليزية، ولكن الدراسات للغة العربية تكاد تكون غير موجودة. في هذا العمل، نقترح إطار عمل قائم على التعلم بالنقل للكشف عن التحيز في نموذج توليد اللغة العربية. كما نقترح مصنعاً جديداً للكشف عن التحيز مبني على AraBERT (Antoun et al., 2020A)، والذي يسمح بفحص التحيز في بيانات النصوص الاجتماعية. هنا، نركز على قياس وتقليل التحيزات التي تعيد إنتاجها نماذج اللغة واسعة النطاق المستخدمة في توليد النصوص العربية (Antoun et al., 2020B) AraGPT2. يعتبر تأثير نماذج اللغة الكبيرة نقطة اهتمام خاصة. دافعنا هو التعامل مع هذا الموضوع بشكل محدد للغة العربية. تركز معظم الأبحاث -حتى الآن- في التحيز في توليد اللغة الطبيعية على اللغة الإنجليزية (Sheng et al., 2021)، على الرغم من وجود نماذج مثل GPT-2 باللغة الألمانية (Kraft, 2021)، كما أنه بالنسبة لـ GPT-3 (Brown et al., 2020)، فإن المؤسسات الأوروبية تعمل على تطوير نظيراتها متعددة اللغات. ونتيجة لذلك، فإن البحث حول تقييم وتقليل التحيزات في نماذج اللغة العربية متأخر. فكرة مصنف التحيز تشبه فكرة مصنف المشاعر. ومع ذلك، وعلى عكس المشاعر، لا



يلتقط التقدير قطبية الجملة بصورة عامة، بل التصور الاجتماعي هو الذي يعززه. ونتيجة لذلك، كان أحد أهداف هذه الورقة هو تطوير مصنف تحيز عربي. لتحقيق هذا الهدف، جرى تجميع مجموعة بيانات مخصصة، وتنظيفها، وتحديدتها. بقية هذه الورقة منظمة على النحو التالي: يقدم القسم ٢ نظرة عامة على الأبحاث ذات الصلة بقياس وتقليل التحيز في توليد اللغة الطبيعية. في القسم ٣، نقترح نهجاً مفاهيمياً للكشف عن التحيز في نماذج التوليد العربية. وخصص القسم ٤ لعرض نتائج التجارب المختلفة. ويستخلص القسم ٥ ما جاء في الورقة.

٢. مراجعة الأدبيات

أُجريت العديد من الدراسات في مجال قياس وتخفيف التحيز في معالجة اللغة الطبيعية. في الفقرات التالية، سنعرض سلسلة من الدراسات والأبحاث العلمية في هذا المجال.

قال Zhou (2021) إن "التحيزات المرتبطة واجهت تحديات في اكتشاف اللغة السامة وتطوير المصنفات، وما يجعل الأمر صعباً هو دقة العدالة". لحل هذه المشكلة، جرى إنشاء بعض الأساليب وتطبيقها لاكتشاف السمية، مع التركيز على المعجم واللهجات. كانت هذه الأساليب محدودة القدرة، لذا اقترحت استراتيجية أخرى لتصحيحها، حيث قللت هذه الاستراتيجية من السمية في اللهجات. في النهاية، اكتشف المؤلف أن إزالة التحيز من نموذج معد على بيانات سامة ليس بنفس نجاح إعادة تصنيف البيانات لإزالة التحيزات.

ناقش Liang et al. (2020) تأثير تقنيات التعلم الآلي واستخدامها في حياتنا الحالية في سياقات مختلفة. كما ذكروا حالات واقعية مختلفة من التحيز، حيث يمكن أن تعرض النماذج المدربة مسبقاً تحيزات خطيرة جداً، تشمل التحيز الجنسي والديني والعنصري، والعديد من الأنماط الأخرى من التحيز. وحددوا مصادر التحيز بدقة قبل إنشاء معايير ومقاييس جديدة أو التفكير فيها. باستخدام هذه التقنيات، قدموا استراتيجيات

لتخفيف التحيزات الاجتماعية أثناء إنتاج النصوص. وقد أظهرت نتائجهم بعد التقييم البشري والتجريبي فعالية تجاربهم في تخفيف التحيز.

اقترح عمل Xu et al. (2021) بعض التقنيات لتخفيف توليد نماذج اللغة السامة. أولاً، قللوا من فائدة نموذج اللغة المستخدم على النصوص من قبل المجموعات المهمشة. وعلى وجه الخصوص، أجروا تقييمات آلية وبشرية لجودة توليد النصوص عندما يكون النموذج مشروطاً بمدخلات من لهجات ومجموعات معرفية مختلفة. لاحظوا أن النموذج أصبح أكثر حساسية لتغيرات التوزيع، خاصة على لغة المجموعات المهمشة. والنتيجة أنهم وجدوا أن هذه الانهيارات ناتجة عن طرق إزالة السمية.

درّس Vaidya et al. (2020) مشكلة النماذج التي تعلمت بشكل غير صحيح ربط التعليقات غير السامة بالتعليقات السامة. كان هدفهم تقليل التحيز المرتبط بالهوية في التعليقات السامة. استخدموا نموذج تعلم متعدد المهام مع طبقة انتباه وحققوا دقة جيدة بنسبة ٧١٪ مقارنة بالنماذج المتطورة الأخرى. ونتيجة لذلك، وجدوا أن نماذج التعلم متعدد المهام تحسن بشكل كبير من أداء التصنيف عندما تكون التعليقات مرتبطة بالهويات التالية: المثليين، المسلمين، السود، واليهود.

عالجت Dathathri et al. (2019) مشكلة صعوبة التحكم في سمات نماذج اللغة المعتمدة على المحولات الكبيرة في النماذج اللغوية التوليدية. وناقشوا فكرة استخدام نموذج "التوصيل والتشغيل". وأكدوا أن نموذج التوصيل والتشغيل أكثر مرونة وسهولة في بناء السمات.

غطت ورقة Sheng et al. (2021) العديد من جوانب نماذج توليد اللغة الطبيعية وتطوراتها السريعة. إذ يمكن أن تولد هذه النماذج نصوصاً بطلاقة وكفاءة، ولكن الباحثين قالوا إن التحيز يمكن أن يظهر في جميع مراحل بناء هذه البرمجيات المعقدة من نشر الأنظمة، إلى التقييم، إلى فك الشيفرة، إلى البيانات، إلى بنية النموذج.

قدم Kraft (2021) العديد من الخوارزميات لاكتشاف وتخفيف التحيز في توليد اللغة الألمانية باستخدام نسخ من GPT-2 و GPT-3. يركز عمل Gururangan et al. (2020) على نماذج اللغة المدربة مسبقاً على نصوص، حيث يتم نقل الخبرة من نموذج



معين للاستفادة منه في مجال مهمة مستهدفة. أكدوا أن البيانات غير المصنفة تحسن الأداء حتى بعد التدريب المسبق المتكيف مع المجال. كما أظهروا أن توسيع النصوص باستخدام استراتيجيات بسيطة لاختيار البيانات هو بديل فعال، خاصة عندما قد تكون الموارد للتدريب المسبق المتكيف مع المجال غير متاحة.

قدمت Schick et al. (2021) تقنية جديدة تسعى "التشخيص الذاتي" لفك الشيفرة النصية. تقلل هذه الطريقة من احتمال إنتاج نصوص سامة. لا تتطلب تقنية إزالة التحيز الذاتي أي بيانات تدريبية أو قائمة كلمات مصممة يدوياً.

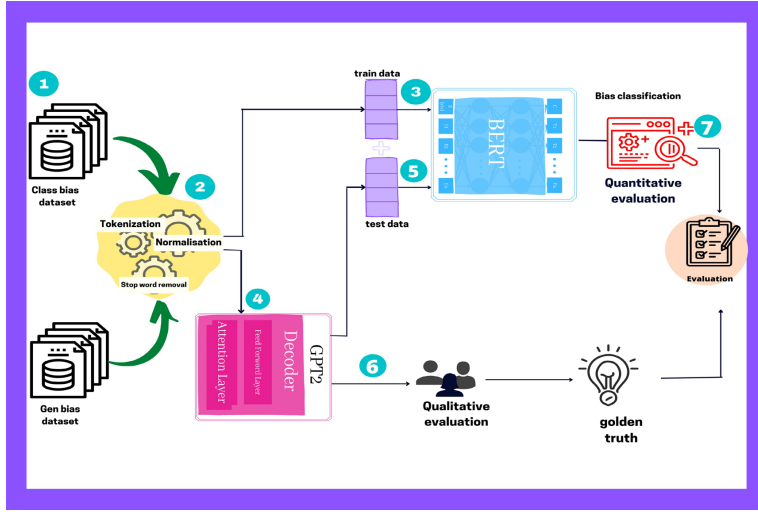
بناءً على هذه الدراسات، يمكننا استنتاج الحاجة الماسة لتطوير تقنيات وأدوات فعالة لاكتشاف وتقليل التحيزات في نماذج اللغة العربية. تتمثل الخطوات المستقبلية في دمج أفضل الممارسات من مختلف الدراسات، وتطوير أدوات قياس وتحليل متقدمة تساعد في تحسين أداء النماذج وتقليل التحيزات الضارة.

٣. بنية النموذج

تستعرض هذه الفقرة الخطوات المتبعة للتحقيق في المساهمات الرئيسية لدراستنا. يتكون النموذج المطور الموضح في الشكل ١ من سبع خطوات رئيسية:

- ٣,١. جمع مجموعة البيانات.
- ٣,٢. تنظيف مجموعة البيانات.
- ٣,٣. تدريب البيانات المجمعة باستخدام AraBERT.
- ٣,٤. توليد التغريدات باستخدام AraGPT2.
- ٣,٥. تصنيف التغريدات المولدة باستخدام نموذج الكشف عن السمية (TCM).
- ٣,٦. التقييم النوعي للنتائج.
- ٣,٧. التقييم الكمي للنتائج.

الشكل (١)
نموذج الهندسة المعمارية المقترح



٥. التجارب والنتائج

أُجريت تجارب مختلفة باستخدام النموذج المقترح لدينا. تستعرض هذه الفقرة تفاصيل عملنا، مع توضيح المقارنات بين التجارب باستخدام الجداول والرسوم البيانية.

٥.١. إعداد التجربة

استناداً إلى بنية النموذج التي سبق تقديمها في القسم السابق، استخدمنا لغة البرمجة Python ومكتباتها القوية لبناء الوحدات المختلفة لنظامنا القادر على التنبؤ بما إذا كانت الجملة منحازة أم لا. يتكون نظامنا من أربع (٤) وحدات رئيسية: معالج البيانات، المصنف القائم على AraBERT، مولد AraGPT2، ووحدة التقييم. قسمنا مجموعة البيانات إلى جزئين. الجزء الأول سيُخصص لمرحلة التوليد (حوالي ٢٠٠ تغريدة). أما الجزء الثاني من



مجموعة البيانات، فسيفصل عشوائياً إلى جزئين، مجموعة التدريب التي تتكون من ٧٠٪
من مجموعة البيانات الأولية و ٣٠٪ لمجموعة الاختبار.

٥,٢ أسئلة البحث

لتحليل النتائج، نركز بشكل رئيسي على الأسئلة البحثية التالية:

- RQ-1: ما مدى فعالية نموذجنا تحت تقنيات التوازن البياني المختلفة؟
- RQ-2: ما مدى فعالية نموذجنا من خلال ضبط بعض معاملات الضبط (hyperparameters) لنموذجنا؟
- RQ-3: ما مدى فعالية نموذجنا في التقييم النوعي؟
- RQ-4: ما هو تأثير نموذج الكشف عن التحيز الخاص بنا على أنظمة البحث والتوصية؟

٥,٣ مجموعة البيانات

استخدمنا مجموعة بيانات "تحليل خطاب الكراهية متعدد اللغات والجوانب" المتاحة على GitHub والتي نشرت في ٢٧ سبتمبر ٢٠٢١. وتتكون من أكثر من ١٣,٠٠٠ معرف تغريدة تتعلق بتصنيف تفصيلي مع سمات متقاطعة، جُمعت في ٤ ملفات CSV. استخدمنا الملف المسمى "ar_dataset.csv" الذي يحتوي على ٣,٣٥٣ تغريدة باللغة العربية. كما هو موضح في الجدول ١، تعرض السمة "الهدف" الحالات التي تحتوي فيها التغريدة على ما يؤدي معناه إلى الإهانات أو التمييز ضد الناس. هذا يمثل تحيزاً في خمس حالات مختلفة وهي الأصل، الانتماء الديني، الجنس، التوجه الجنسي، الاحتياجات الخاصة أو غيرها.

الجدول (١)

وصف مجموعة البيانات

السمة	التسمية	الإحصائية
	العرق	877
	الجنس	548
الهدف	الدين	145
	أخرى	1782
	الإثنيات	1

٥,٤. معالجة البيانات المسبقة

تعد خطوة معالجة البيانات المسبقة ضرورية لأي نموذج ناجح. تهدف إلى إعداد البيانات لاستغلالها اللاحق. بعد اختيار مجموعة البيانات الرئيسية لدينا، بدأنا خطوة التنظيم على البيانات المحددة. أولاً استوردنا مكتبة 'NLTK' لهذه العمليات، ولكل وظيفة أنشأنا دالة لحلها. لدينا في المجلد ست (٦) وظائف:

- إزالة علامات الترقيم: مثل \، ؟ () : ! ستزال.
 - إزالة الأحرف المتكررة: باستخدام مكتبة "re" وإزالة كل تكرار للأحرف أكثر من مرة متتالية.
 - إزالة التشكيل: إزالة التشكيل العربي مثل 'فَتْحَة'، 'ضَمَّة'، 'سُكُون' باستخدام مكتبة 're' ثم تحويل النتيجة إلى سلسلة نصية.
 - إزالة الأرقام: حذف جميع الأرقام العربية والأرقام الرقمية داخل البيانات.
 - إزالة العلامات: إزالة العلامات من الجمل المختلفة باستخدام مكتبة 're' مثل "@" و "#" والرموز التعبيرية و "@url" والرموز والرموز التصويرية.
 - تطبيع العربية: استبدال "ة" ب"ه" و "أ" ب"إ" و "ا" ب"إ" على سبيل المثال.
- بعد تطبيق عمليات المعالجة المسبقة المختلفة، حصلنا على مجموعة البيانات جاهزة للخطوات القادمة في بنية نموذجنا. تجدر الإشارة إلى أننا احتفظنا سابقاً بحوالي ٦٪ من بياناتنا لاستخدامها في مرحلة التوليد.



٥,٥. التجربة الأولى: (RQ-1) تقنيات موازنة البيانات المختلفة)

هذه التجربة تقدم النتائج التجريبية لأول سؤال بحثي مطروح في القسم ٤,٢. لقد اقترحنا تكوين نموذج يعتمد على البنية الأساسية لـ AraBERT. وتوصف معاملات الضبط لهذه البنية في الجدول ٢ أدناه.

الجدول (٢)

معاملات ضبط التجربة الأولى

المعامل	القيمة
الحد الأقصى للطول	64
معدل التعلم	0.00006
حجم الدفعة	6
عدد العصور	4

كما هو موضح في الجدول ١، يمكننا بسهولة ملاحظة عدم توازن كبير في الفئات في مجموعة البيانات، مما سيؤثر بالتأكيد على أداء النظام. لحل مشكلة البيانات غير المتوازنة، اخترنا أربع (٤) تكوينات مختلفة لمجموعتنا.

البيانات الأصلية:

بعد إزالة ٦٪ من البيانات لخطوة التوليد، احتفظنا بـ ٣١٥٣ تغريدة، قسمناها إلى ٢٢٠٧ تغريدة (٧٠٪) لبيانات التدريب، تتألف من ١٧٢٠ جملة متحيزة و٤٨٧ جملة غير متحيزة. واستُخدمت ٣٠٪ المتبقية من بياناتنا لأغراض الاختبار.

البيانات المتوازنة:

أردنا معرفة ما سيحدث إذا غيرنا توزيع بياناتنا. ونظرًا لأن بياناتنا الأولية لم تكن متوازنة، فقد قررنا إجراء نفس التجارب على نسخ جديدة من البيانات. وطبقنا ثلاث تقنيات موازنة مختلفة على بياناتنا الأولية.

البيانات المتوازنة يدويًا:

كما ذكرنا سابقًا، تحتوي بياناتنا على ٣١٥٣ تغريدة، ٧١٥ منها تغريدات عادية بدون أي نوع من التحيز مقارنة بـ ٢٤٣٨ تغريدة مصنفة على أنها متحيزة. لذلك، قررنا الاحتفاظ بنفس الحجم لمجموعتي التحيز والعادية. أزلنا ١٧٢٣ تغريدة من المجموعة الثانية واحتفظنا بـ ٧١٥ تغريدة عادية و ٧١٥ تغريدة متحيزة. قُسمت بياناتنا الجديدة المتوازنة بعد ذلك إلى ٧٠٪ للتدريب و ٣٠٪ للاختبار.

البيانات المتوازنة باستخدام SMOT:

SMOT هي طريقة تستخدم لجعل البيانات متوازنة عن طريق أخذ الفئة الأقل وجعلها مساوية للفئة الأكبر (Brownlee, 2021). استخدمنا تقنية SMOT مع المعلمة "استراتيجية أخذ العينات = الأقلية". وقد ضُخمت بيانات الفئة الأقلية من ٥١٢ إلى ١٦٩٥.

البيانات الموسعة:

يصف توسيع البيانات مجموعة من الخوارزميات التي تنشئ بيانات تركيبية من مجموعات البيانات المتاحة. تحتوي هذه البيانات التركيبية عادةً على تغييرات صغيرة في البيانات التي تتطلب أن تكون تنبؤات النموذج غير متغيرة. في حالة توسيع النصوص لدينا، حاولنا تطبيق توسيع البيانات من خلال التوليد. بياناتنا تحتوي على تغريدات متحيزة بحجم ضعف البيانات العادية. استخدمنا التغريدات العادية كمدخلات لخوارزمية AraGPT2 لتوليد الضعف باستخدام طولين مختلفين للإخراج، ٦٤ و ١٢٨.



أضفنا التغريدات الجديدة الناتجة إلى المجموعة الأصلية من التغريدات العادية، بحيث تحتوي مجموعتنا الجديدة على نفس عدد التغريدات المتميزة والتغريدات العادية.

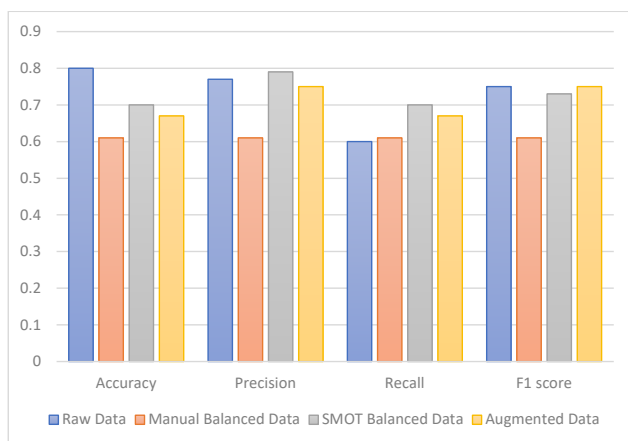
الجدول (٣)

نتائج التجربة الأولى

تقنية الموازنة	الدقة	الدقة المعيارية	الاسترجاع	F1
البيانات الخام	0.80	0.77	0.60	0.75
البيانات المتوازنة يدوياً	0.61	0.61	0.61	0.61
البيانات المتوازنة باستخدام SMOT	0.70	0.79	0.70	0.73
البيانات الموسعة	0.67	0.75	0.67	0.75

الشكل (٢)

أداء التجربة الأولى



كما هو موضح في الشكل ٢، حصل تكوين نموذجنا باستخدام التوزيع الأصلي للبيانات على أفضل درجات الدقة و **Macro F1** أظهرت الدقة وكذلك الاسترجاع تحسناً في

تقنيات SMOT و Augmentation من هذه التجربة، حصلنا على درجة F1 بلغت ٧٥٪
للتقنية الأولى، و ٦١٪ للتقنية الثانية، و ٧٣٪ لتقنية SMOT ، و ٧٥٪ لتقنية
Augmentation في التجربة التالية، سنضيف بعض التنظيم إلى نموذجنا ونرى ما إذا
كانت مقاييس الأداء ستتحسن.

٥,٦ التجربة الثانية: (RQ-2 معاملات الضبط المختلفة)
في هذه التجربة، نقدم النتائج التجريبية للسؤال البحثي الثاني Q2. اخترنا تكوينات
مختلفة باستخدام ٤ قيم لمعامل الضبط "Max length" ، وقيمتين لمعامل الضبط
"Batch size" ، وقيمتين لمعامل الضبط "Epochs" . معاملات الضبط المستخدمة
لتجميع وتشغيل النموذج موضحة في الجدول ٤.

الجدول (٤)

معاملات ضبط التجربة الثانية

المعامل	القيمة
الحد الأقصى للطول	64 / 128 / 256 / 512
معدل التعلم	0.00006
حجم الدفعة	6 / 32
عدد العصور	4 / 10

يوضح الجدول ٥ النتائج التجريبية لنموذجنا باستخدام المعلمات التالية
Max_length = 512 و Batch_size = 6 و Epochs = 4.



الجدول (٥)

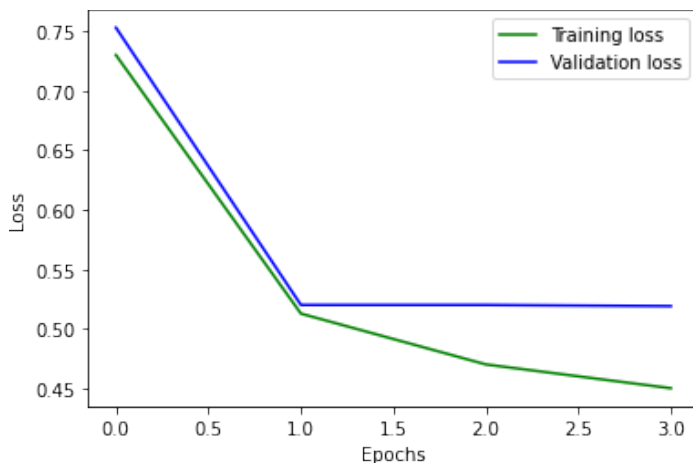
نتائج التجربة الثانية

الحد الأقصى للطول	الدقة	الدقة المعيارية	الاسترجاع	F1
64	0.80	0.77	0.60	0.75
128	0.78	0.72	0.77	0.69
256	0.78	0.74	0.78	0.74
512	0.81	0.78	0.81	0.78

الشكل ٣ أدناه يظهر تقريباً خسارة التدريب والتحقق التي حصل عليها نموذجنا بناءً على معاملات ضبط التجربة الثانية.

الشكل (٣)

خسارة التدريب والتحقق لأفضل معاملات الضبط



نتائج هذه التجربة توضح لنا بوضوح قدرة نماذج التعلم التحويلي على اكتشاف التحيز في النصوص العربية بدقة عالية. من خلال هذه التجربة، يمكننا أيضاً أن ندرك أهمية تأثير المعلمات الفائقة على أداء هذه المصنفات. نموذج AraBERT باستخدام معامل ضبط $\text{Max_Length}=512$ تفوق بصورة كبيرة على جميع النماذج الأخرى.

٥,٧ التجربة الثالثة: (RQ-3 التقييم النوعي لعملية التوليد)

تقدم هذه التجربة النتائج التجريبية للسؤال البحثي الثالث RQ3 المطروح في القسم ٤,٢. اقترحنا تكوين نموذج يعتمد على البنية الأساسية لـ AraGPT2. الحد الأقصى لطول النصوص المولدة هو ٨٠ كلمة. ولإستخدام عملية التوليد، استخدمنا مجموعة ٦٪ التي تم الاحتفاظ بها في البداية، وهي حوالي ٢٠٠ تغريدة. يوضح الجدول ٦ معاملات ضبط نموذج التوليد لدينا.

الجدول (٦)

معامل ضبط التجربة الثالثة

القيمة	معامل الضبط
PT	إرجاع الموتر
80	إدخال الحد الأقصى للطول
True	تخطي الرموز الخاصة
80	إخراج الحد الأقصى للطول

تقييم النموذج باستخدام Rouge و Bleu

BLEU (BiLingual Evaluation Understudy)

قياس Bleu يهتم بمدى ظهور (الكلمات) أو (n-grams) في الملخصات المولدة بواسطة

الألة في الملخصات المرجعية البشرية. (Papineni et al. (2002).

BLEU Corpus Score



يقارن مستنداً مرشحاً واحداً بوثيقة مرجعية متعددة الجمل. يختلف عن متوسط درجات BLEU لكل جملة، حيث يحسب النتيجة بجمع البسط والمقام لكل زوج من الفرضيات والمرجع قبل القسمة.

ROUGE (Recall Oriented Understudy for Gisting Evaluation)

يتضمن مقاييس لتحديد جودة الملخص تلقائياً من خلال مقارنته بملخصات أخرى (مثالية) أنشأت بواسطة البشر. Lin, (2004). تقيس المقاييس عدد الوحدات المتداخلة مثل n_gram ، وتسلسلات الكلمات، وأزواج الكلمات بين الملخص الذي أنشئ بواسطة الكمبيوتر ليحسب تقييمه والملخصات المثالية التي أنشأها البشر. يعرض الجدول ٧ نتائج تقييم النموذج باستخدام BLEU و ROUGE. يمكننا ملاحظة أن نظامنا أعطى نتائج مقبولة.

الجدول (٧)

إحصائيات تقييم ROUGE و BLEU

المقياس	القيمة
BLEU Score	1.107090
BLUE-1 Score	0.069522
ROUGE-1 Score "r"	0.008020
ROUGE-1 Score "p"	0.018894
ROUGE-1 Score "f"	0.010824

تقييم التوسيم البشري

الدافع

نحن مهتمون بدراسة كيفية تأثير اختلاف التوسيم على تقييم نموذجنا. في هذا البحث، نركز مهمة التوسيم على التحيز المستهدف لتسهيل المقارنة مع التقييم الآلي.

النهج

نجري تحليلاً نوعياً محدوداً لفهم مدى قدرة البشر على اكتشاف التحيز في النصوص المولدة. في هذه التجربة، تُوسَّع كل من العينات البالغ عددها ٢٠٠ بواسطة مؤسِّمين لتحديد التحيز المستهدف في التغريدات المولدة. أجرينا دراسة لتجربة المستخدم لنظامنا مع مؤسِّمين (E1, E2)، حيث يقيم المشاركون النظام على أنه منحاز أو غير منحاز لكل نص مولد. في التوسيم، قد لا يتفق المؤسِّمون دائماً: فقد يكون بعضهم أكثر تشدداً في أحكامهم من الآخرين. ناقش E1 و E2 النتائج التي حصلوا عليها من نظامنا لحل أي خلافات حتى الوصول إلى توافق.

كابا

في سياق الاتفاق بين المؤسِّمين، يستخدم معامل كابا لكوهين (Wang & Xia, 2019) لقياس موثوقية التوسيم البيني. يمثل معامل كابا في اختبار التكرار مدى الاتفاق بين مجموعتين من البيانات التي تم الحصول عليها في مناسبتين أو أكثر. نظمنا نتائج المؤسِّمين في الجدول ٨ أدناه.

الجدول (٨)

نتائج التوسيم

مؤسِّم ٢: غائب أو ١	مؤسِّم ٢: حاضر أو ٠	المجموع
مؤسِّم ١: غائب أو ١	A=75	B=11
مؤسِّم ١: حاضر أو ٠	C=32	D=82
المجموع	A+C=107	B+D=93
		A+B+C+D=200

اتفق المؤسِّمون على أن ٧٥ تغريدة منحازة و ٨٢ تغريدة عادية، لذا اتفقوا على ١٥٧ تغريدة واختلفوا في ٤٣ تغريدة بين العادية والمنحازة من إجمالي ٢٠٠ تغريدة. نسبة الاتفاق هي $p_0=0.79$ $p_0=0.79$ محسوبة بالصيغة التالية :



$$P_0 = \frac{A \text{ expected value} + D \text{ expected}}{(A+B+C+D)} * 100\%$$

(1)

لا يمكننا اعتبار هذه النسبة مرجعاً؛ لأن أحد المؤسّمين أو كلاهما قد يقيم التغريدة عن طريق الصدفة. تأخذ قيمة كابا هذا الاحتمال في الاعتبار. يكون حسابها بناءً على الاتفاق الملحوظ p_0p_0 مقارنة بمدى توقع هذا الاتفاق بالصدفة ويسمى الاتفاق المتوقع $p_e=0.27p_e=0.27p_e=0.27$.

يُحسب معامل كابا النهائي بالصيغة التالية:

$$Kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{0.79 - 0.27}{1 - 0.27} = 0.72 \quad (2)$$

التقييم الآلي

تنفذ خطوة التقييم الآلي تقنياً قبل المقارنة بين المؤسّمين وتوقعات النموذج. حالياً، وبعد مرحلة تدريب النموذج، استخدمنا التغريدات العربية المولدة بوصفها مدخلات لاختبار النموذج. طلبنا من النظام التنبؤ بالتحيز في التغريدات المولدة بناءً على ما تعلمه النموذج من التجارب السابقة. يُستخدم نظامنا كـ "مُصنف التحيز" لاكتشاف التحيز المقدم من النموذج التوليدي. تُستخدم نتائج مرحلة التنبؤ مع نتائج المؤسّمين السابقة (الحقيقة الأرضية) لاستكمال عملية التقييم الشاملة.

التقييم النهائي

تسلط هذه المرحلة الضوء على كل حالة بين التنبؤ ونتائج المؤسّمين. يقدم الجدول ٩ نتائج التقييم النهائي للسؤال البحثي الثالث. اتفق النظام مع المؤسّمين في ٩٦ تغريدة، ٦٩ منها منحازة بينما الباقي تغريدات عادية. نسبة الاتفاق هي ٤٨.٠%.

الجدول (٩)

إحصائيات التقييم النهائي

التغريدة العادية	التغريدة المنحازة	
115	85	مؤسّم ١
95	105	مؤسّم ٢
43	157	النظام

النتائج للتجربة الثالثة

يمكننا الآن معرفة مدى توافق المؤسّمين في تصويتهم من نتيجة كبا. حيث تبلغ ٧٢٪، مما يشير إلى أن مستوى توافقهم كبير وأنهم موثوقون في [٦١٪-٨٠٪] من البيانات. اتفق النظام مع المؤسّمين في ٩٦ تغريدة، ٦٩ منها منحازة بينما الباقي تغريدات عادية. نسبة الاتفاق مع النظام هي ٤٨٪، مما يعني أن المؤسّمين لا يزالون يتناقضون مع النظام في أكثر من ٥٠٪ من النتائج. هذا يعني أيضاً أن مجال دراسة التحيز واكتشافه وتخفيفه لا يزال يحتاج إلى جهود كبيرة، خاصة فيما يتعلق باللغات منخفضة الموارد مثل اللغة العربية واللهجات العربية.

٥,٨. تجربة : (4-RQ ما هو تأثير نموذج اكتشاف التحيز على أنظمة البحث والتوصية؟)

مقدمة

تقدم هذه التجربة نتائج البحث الرابع RQ4 المشار إليه في القسم ٤,٢. أجرينا تجربة تقييم بشري لتقييم تأثير نموذج اكتشاف التحيز لدينا على أنظمة البحث والتوصية.

المنهجية

أجرينا اختياراً عشوائياً لـ ١٠ تغريدات من مجموعة البيانات الاختبارية لدينا، ومن ثم ولّدنا ١٠ تغريدات جديدة باستخدام نموذج AraGPT2. طلبنا من المقيمين E1 و E2 ذوي المعرفة في مجال تحليل المشاعر وخطاب الكراهية تقييم تأثير التحيز الاجتماعي في



التغريدات المولدة استناداً إلى النتائج التي أرجعها محرك بحث Google كنظام بحث،
ونظام توصيات YouTube كنظام توصية. يعرض الجدول ١٠ نتائج تقييم E1 و E2.

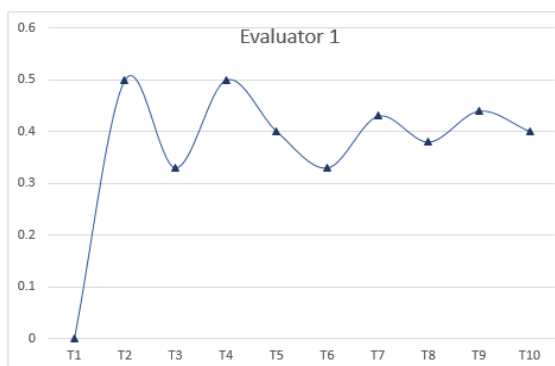
الجدول (١٠)

نتائج التقييم $P@K$ و $E1$ و $E2$

P@K	E2	P@K	E1	تغريدة مولدة
1.00	B	0.00	B	T1
1.00	B	0.50	NB	T2
0.67	NB	0.33	NB	T3
0.50	NB	0.50	B	T4
0.40	NB	0.40	NB	T5
0.33	NB	0.33	NB	T6
0.29	NB	0.43	B	T7
0.38	B	0.38	NB	T8
0.44	B	0.44	B	T9
0.40	NB	0.40	NB	T10

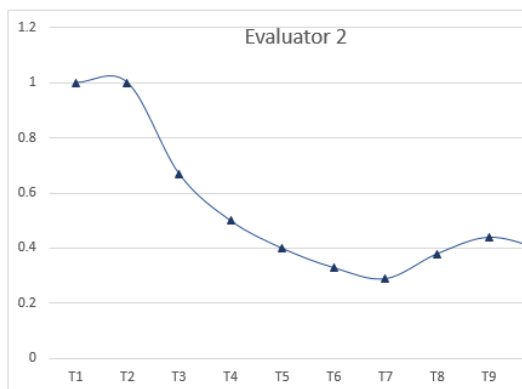
الشكل (٤)

نتائج الأداء لـ E1



الشكل (٥)

نتائج الأداء لـ E2



في الجزء الثاني من هذه التجربة، اخترنا عشوائياً استفسارين Q1 و Q2 مع جميع الإجابات المرتبطة بهما لعرض تأثير استخدام التغيرات المولدة كاستفسارات لكل من محرك بحث Google ونظام توصيات YouTube. تظهر نتائج هذه الدراسة في الجدول ١١.

الجدول (١١)

Q1 و Q2 المستخدمة في عملية التقييم

الاستفسار	المعنى	النص
Q1	العربية	انا اوافقك بخصوص السوريين العراقيين اما بخصوص السعودي مو بحاجة انه يقعد بالكويت بدو يزور الكويت لا يهم اني بس حاطط لك رقم او رقم انا اتزوج عراقي من سوري ولا
	الإنجليزية	I agree with you about the Iraqi Syrians. As for the Saudi, I don't need him to stay in Kuwait. He wants to visit



Kuwait. It doesn't matter that I just gave you a number or a number. I'm marrying an Iraqi to a Syrian.		
السعوديه شعبيها شعب الخيم بول البعير يستهزا الناس ويرمونهم بالحصى والنفايات من كل مكان ويرمونهم بالجرار ثم من كان يقول والله لولا فضل الله علينا لو تعلمون ما	العربية	Q2
Saudi Arabia, its people, the people of the tents, the urine of the camels, mock the people, and throw pebbles and rubbish at them from everywhere, and throw them at them with tractors.	الإنجليزية	

الجدول (١٢)

نتائج التقييم/المتحيز Q1 و Q2 على Google و YouTube

عدد النتائج المتحيزة على محرك بحث Google	عدد النتائج المتحيزة على نظام توصيات YouTube	استفسار
3	8	Q1
0	8	Q2

استخدمنا مصنف التحيز لدينا لاكتشاف الكلمات الرئيسية المتحيزة. أزلنا هذه الكلمات من Q1 و Q2 ثم شغلنا استفسارات جديدة على محرك بحث Google ونظام توصيات YouTube. تظهر نتائج هذه الدراسة في الجدول ١٢.

الجدول (١٣)

نتائج التقييم غير المتحيز لـ Q1 و Q2 على YouTube و Google

استفسار	الكلمات العربية المتحيزة المزالة	الترجمة الإنجليزية	عدد النتائج غير المتحيزة على محرك بحث Google	عدد النتائج غير المتحيزة على نظام توصيات YouTube
Q1	السوريين، العراقيين، يتزوج	Iraqis, Syrians, Get married	10	8
Q2	بول البعير، يستهزأ الناس، النفائيات	Camel's urine, make fun of people, Waste	302	12

النتائج

من الجدول ١١ لاحظنا تحسناً بمقدار +Q1 لـ ٧ من حيث النتائج التي أرجعها محرك بحث Google بالنسبة لـ Q2، رأينا تحسناً بمقدار +٤ للنتائج التي أرجعها نظام توصيات YouTube، بينما كان هناك تحسن بنسبة +٣.٢ في نتائج محرك بحث Google. تشير هذه النتائج إلى أن نموذج اكتشاف التحيز لدينا فعال في اكتشاف الكلمات المفتاحية الخاصة بخطاب الكراهية والتي لها تأثير مباشر على نتائج أنظمة البحث والتوصية. هذا يؤكد أن التحيز الاجتماعي يمكن أن يؤثر بشكل محتمل على قرارات أنظمة البحث والتوصية، مما يتماشى مع العديد من الدراسات السابقة التي تشير إلى مخاطر الظلم والتحيز الاجتماعي في البيانات المستخدمة لتدريب واختبار تلك الأنظمة.

٦. الخاتمة

في هذه الورقة، نقترح نموذجاً جديداً قائماً على التعلم بالنقل يتكون من سبع خطوات رئيسية. صممنا مصنف تقدير أو نموذج اكتشاف التحيز بناءً على نسخة معدلة من AraBERT لتحقيق هدف اكتشاف ما إذا كانت الجملة العربية أو التغريدة متحيزة أم لا.



يُبنى نموذج التوليد العربي باستخدام AraGPT2 للسماح لنا بتوليد مجموعة من الجمل باستخدام تغريدات عربية عادية وغير متحيزة بوصفها مدخلات. اختبرنا قدرة AraGPT2 على توليد أنواع مختلفة من القوالب النمطية المتحيزة. للحصول على أفضل نتيجة مرتبطة باكتشاف التحيز باستخدام AraBERT ، أجرينا عدة تجارب بتغيير معلمات مختلفة. حققنا أفضل نتيجة، والتي كانت قريبة من ٨١٪ لمقياس الدقة، باستخدام المعلمات التالية Max_Length=512 ، batch_size=6 ، و Epochs=4 في خطوة التوليد، تتكون تجربتنا من توظيف اثنين من المعلقين لمقارنة تقييماهم ونتائج عملهم على عملية تقييم النموذج التوليدي. قسنا الاتفاق بينهم باستخدام مقياس كابا ثم قارنا النتائج مع تنبؤات AraBERT. في النهاية، أسفرت الطرق الموصوفة في هذه الورقة عن نتائج واعدة، كما هو موضح في القسم الأخير. ومع ذلك، يمكننا النظر في تحسين نظامنا من خلال تدريب النموذج على مجموعة بيانات أكبر، وإجراء المزيد من التجارب. بالإضافة إلى ذلك، يمكننا الاستثمار في استخدام نموذجنا لخدمة تطبيقات معالجة اللغة الطبيعية الأخرى باللغة العربية.

المراجع

- Hern, A. (2017, October 24). Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*. <https://www.theguardian.com>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. <https://doi.org/10.1126/science.aal4230>
- Zhou, X. (2021). *Challenges in automated debiasing for toxic language detection* [Unpublished Master's thesis]. University of Washington.
- Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
- Antoun, W., Baly, F., & Hajj, H. (2020). *Arabert: Transformer-based model for Arabic language understanding*. *arXiv Preprint arXiv:2003.00104*.
- Antoun, W., Baly, F., & Hajj, H. (2020). *Aragpt2: Pre-trained transformer for Arabic language generation*. *arXiv Preprint arXiv:2012.15520*.
- Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. *arXiv Preprint arXiv:2105.04054*.
- Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Liang, P. P., Wu, C., Morency, L. P., & Salakhutdinov, R. (2021 July 18-24). *Towards understanding and mitigating social biases in language models*. In the 38th International Conference on Machine Learning, [Virtual].
- Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., & Klein, D. (2021). Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*.
- Vaidya, A., Mai, F., & Ning, Y. (2020, May). Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 683-693).



- Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., ... & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.
- Kraft, A. (2021). *Triggering Models: Measuring and Mitigating Bias in German Language Generation* [Unpublished Master's thesis], University of Hamburg.
- Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Schick, T., Udupa, S., & Schütze, H. (2021). Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9, 1408-1424. https://doi.org/10.1162/tacl_a_00437
- Brownlee, J. (March, 17, 2021). SMOTE for Imbalanced Classification with Python. *Machine Learning Mastery*. <https://machinelearningmastery.com>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002 July). *Bleu: A method for automatic evaluation of machine translation*. In the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.
- Lin, C. Y. (2004 July). *Rouge: A package for automatic evaluation of summaries*. In Text Summarization Branches Out, Barcelona, Spain.
- Wang, J., & Xia, B. (2019 August 13-15). *Relationships of Cohen's Kappa, sensitivity, and specificity for unbiased annotations*. In the 2019 4th International Conference on Biomedical Signal and Image Processin, Chengdu, China.

نبذة تعريفية عن الباحث

Author Biodata	بيانات الباحث
Fouzi Harrag is an Associate 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.	فوزي حراق، أستاذ مشارك 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، معالجة اللغات الطبيعية وأخلاقيات الذكاء الاصطناعي.

معرف أوركيد id : 7146-8348-0002-0000

Email: fouzi.harrag@univ-setif.dz

المصادر المعجمية العربية مدونة محوسبة: تحليلها، توسيمها، دمجها

بناء مصدر معجمي دلالي موسع مقروء آلياً

Arabic Lexical Resources as a Corpus: Parsing, Annotation, and Integration into a

Unified and Comprehensive Machine-Readable Lexical Semantic Database

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



عبد العاطي إبراهيم هوري

Datalex4ai, San Jose, CA, USA

توثيق البحث: APA Citation

هوري، عبد العاطي إبراهيم. (2024). المصادر المعجمية العربية مدونة محوسبة: تحليلها، توسيمها، دمجها لبناء مصدر معجمي دلالي موسع مقروء آلياً، مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية، (٢).

Submission Date: 30/08/2024, Acceptance Date: 19/09/2024

تاريخ الإرسال: 2024/08/30 / تاريخ القبول: ٢٠٢٤/٩/١٩ م

Abstract

This paper presents a comprehensive approach to constructing a unified Machine-Readable Dictionary (MRD) by leveraging printed Arabic lexical resources, including dictionaries, thesauri, and semantic dictionaries. These resources are invaluable for natural language processing (NLP) tasks, yet they remain underutilized due to their printed form. Our methodology involves parsing, annotating, and merging various lexical sources into a single, coherent MRD, designed for effective utilization in computational applications. The process includes syntactic and semantic parsing to capture linguistic nuances and ensure consistency across entries. We employ automated and manual annotation techniques to address discrepancies and enrich the lexical data. This unified resource not only bridges the gap between different lexical formats but also offers a powerful tool for tasks such as word-sense disambiguation, semantic analysis, and machine translation. The paper demonstrates the feasibility and value of transforming printed Arabic lexical resources into an accessible, standardized digital format for both linguistic research and practical NLP applications.

Keywords: MRD, Arabic Lexicography, Lexical Semantic, Dictionary Parsing, Lexical Integration, Thesaurus.

الملخص:

موضوع الدراسة وصف منهجية لسانية حاسوبية لتحليل متن المصادر المعجمية التقليدية العربية؛ لاستخلاص أنماط المعلومات المعجمية الدلالية الصريحة والضمنية في متن النص المعجمي العربي؛ ويشمل ذلك أنماط المعلومات الصرفية (الاشتقاقية والتصريفية)، والتركيبية (النحوية والإعرابية)، والمعجمية (المداخل المفردة، والمداخل المركبة، والمتصاحبات اللفظية)، والدلالية (العلاقات الدلالية، والحقول الدلالية)، والاستعمالية، والتأنيلية، وتوسيمها بالمعلومات الضرورية، واستكمال المعلومات التي يمكن استنتاجها، ثم دمج المعلومات المستخلصة، ووضع هذه المعلومات في مورد معجمي دلالي مقروء آلياً، يمكن توظيفه في بناء معجم لغوي ومكثّر دلالي موسّع. وتنطلق الدراسة من فرضية مؤداها أن الموارد المعجمية العربية التقليدية ثروة لغوية ولسانية تفتقر إلى الجدولة الحاسوبية، وتنقصها بعض أنماط المعلومات المعجمية الحديثة. ويمكن تحويل مادة هذه الموارد المعجمية إلى صيغة محوسبة وتوظيفها في المجال الحاسوبي من جهة، وفي تطوير أجيال من الموارد المعجمية أكثر حداثة، وأكفاً لسانياً، تحقيقاً لمتطلبات المستعمل البشري، والآلة في الوقت ذاته؛ وذلك باستخلاص الأنماط الصرفية والتركيبية والدلالية والمعجمية الفارة في المتن المعجمي التقليدي.

الكلمات المفتاحية: معجم مقروء آلياً، معجمية عربية، اطرادات معجمية دلالية، توسيم، بناء المصادر المعجمية.

المقدمة

تصف الدراسة منهجية معالجة عدد من الموارد المعجمية العربية التقليدية؛ والتعامل معها بوصفها مدونة محوسبة، ومعالجتها معالجة حاسوبية بتحليل نصوصها، وتوسيم مداخلها؛ تمييزاً لمعلوماتها الضرورية، ثم دمج هذه الموارد المعجمية -على اختلاف طبيعة معلوماتها، وأنماطها، ومنهجياتها المعجمية-؛ لبناء هيكلية مصدر معجمي حاسوبي متكامل؛ يفيد من الثروة اللغوية والمعجمية المنجزة، وتُعرض فيه المعلومات في صورة متكاملة، وتقدم فيها الاطرادات المعجمية الدلالية لبنية المعجم العربي تقديمًا صريحًا مُجدولًا.

وينطلق العمل من قناعتين، الأولى: أن الموارد المعجمية العربية ثرية، ومتنوعة، من حيث أنماطها المعجمية (معاجم عامة، ومعاجم موضوعات، ومكانز معجمية، ومعاجم قطاعية)، ومن حيث طبيعة المعلومات التي تقدمها. وأما الثانية: فأن هذه الموارد المعجمية العربية -رغم ثراء مادتها وعمق معلوماتها- لا تلتزم بنية معجمية مطردة؛ فالبنية الصغرى تفتقر -في كثير من الحالات- إلى الاتساق والانتظام في تقديم المعلومات المعجمية الدلالية، إضافة إلى أن كثيرًا من الاطرادات المعجمية الدلالية الموجودة في بنية المعجم العربي لا تُعرض بطريقة صريحة متسقة، بل توجد في صورة ضمنية يمكن استنباطها، وعرضها في صورة صريحة مطردة، تُجَلِّي البنية المعجمية الدلالية للمعجم العربي.

وتتلخص إجراءات العمل في خطوات منهجية تتضمن: أ) اختيار عدد من الموارد المعجمية العربية تمثل أصنافًا متنوعة من أنماط الموارد المعجمية، ومعالجتها معالجة أولية. ب) تحليل النص المعجمي للتوصل إلى أنماط المعلومات المعجمية الدلالية التي تقدمها للمداخل المعجمية الرئيسة والفرعية، إضافة إلى الحقول الدلالية، والمتصاحبات اللفظية؛ لتحويلها إلى مصادر معجمية مقروءة آليًا. ج) توسيم هذه الموارد المعجمية لاستكمال المعلومات الضرورية التي قد تكون ناقصة في كل نمط من الأنماط المعجمية.



(د) استخلاص أنماط المعلومات الدلالية والعلاقات والاطرادات المعجمية الدلالية في متن هذه الموارد المعجمية. (هـ) دمج هذه الموارد المعجمية للتوصل إلى مصدر معجمي متكامل، تقدّم معلوماته في صورة مطردة، ومتسقة، ومتكاملة، لتشمل -إلى جانب المعلومات الأساسية- القيود الانتقائية، والحقول الدلالية، والعلاقات الدلالية، والمتلازمات اللفظية، والمتصاحبات اللفظية.

١,١. الاطرادات المعجمية الدلالية

تقدم الموارد المعجمية معلومات تتعلق بالوحدة المعجمية، مفردة أو مركبة، وتباين فيما بينها، إن كان بحسب أصنافها (معاجم لفظية، أو معاجم دلالية)، أو بحسب منهجياتها المعجمية؛ جمعاً وتحريراً للمادة المعجمية. والحديث عن اطرادات معجمية؛ يعني وجود تصور لبنية معجمية دلالية لمعجم اللغة، ينبغي للبحث أن يعمل على رصد أنماط المداخل المعجمية؛ على اختلافها وتنوعها شكلاً ومحتوى، ثم رصد المعلومات المقدمة لكل نمط، وتصنيفها بحسب طبيعتها، ثم تجلية العلاقات الدلالية، بين هذه المداخل المعجمية، التي تتميز فيها المقولات المعجمية من حيث أنماط المداخل، أو أصناف المعلومات المقدمة. ويمكن أن نتصور نوعين لأشكال المعلومات المعجمية الدلالية التي يتضمنها المتن المعجمي العربي من حيث إمكانية التعرف عليها، لدى التحليل، كما يلي:

(أ) أنماط صريحة: وهي أنماط المعلومات المنصوص عليها نصّاً في منهجية المعجم، والتي يعبر عنها بمصطلحات محددة، أو مشار إليها برموز أو مختصرات متفق عليها. وهذه الأنماط يمكن التعرف عليها من خلال التحليل اليدوي والتحليل الآلي Parsing.

(ب) أنماط ضمنية: وهي المعلومات المعجمية التي لم تُذكر صراحة في النص المعجمي؛ وتستوجب تحليلاً أعمق للتوصل إليها وتمييزها وتصنيفها. ويستعان على ذلك بتحليل لغة الشروح والتعريفات لرصد هذا الصنف من الاطرادات في هذه اللغة، وما تنطوي عليه من

إشارات لطبيعة المحتوى المعجمي الدلالي للوحدة المعجمية المشروحة، أو سلوكها التركيبي، أو المتصاحبات اللفظية لهذا المدخل مع هذا المعنى، وذلك الاستخدام. ويهدف هذا العمل إلى تحقيق ما يلي:

- تحليل مادة المعجم العربي التقليدي: لاستخلاص الأنماط الصرفية والمعجمية والتركيبية والدلالية التي قد يكون لها أهمية في بناء مصدر معجمي دلالي حاسوبي للغة العربية؛ لتحويل متن المعجم التقليدي (بالتطبيق على عدد من المواد المعجمية ممثلة لثلاثة أصناف من الموارد المعجمية العربية) إلى قاعدة بيانات معجمية دلالية موسعة.
- توسيم الموارد المعجمية: بإضافة عدة مستويات من المعلومات المعجمية الدلالية التي نراها ضرورية لسد الثغرات الموجودة في بعض الموارد المعجمية التقليدية؛ للإسهام في إعداد مادة مورد معجمي دلالي حاسوبي موسع، يحقق المواصفات المعجمية الحديثة.
- دمج مخرجات التحليل المعجمي: لبناء قاعدة بيانات معجمية دلالية معمقة للغة العربية الفصحى؛ وذلك للوصول إلى قائمة موحدة من المداخل المعجمية، ثم دمج المعلومات الوصفة لكل مدخل، والمعلومات المعجمية الدلالية لكل معنى.

١,٢. المنهجية والإجراءات

تتلخص إجراءات العمل في خطوات منهجية تتضمن:

- بناء المدونة النصية المعجمية لمتن الموارد المعجمية العربية: باختيار عدد من الموارد المعجمية العربية التي تمثل أصنافاً متنوعة من أصناف الموارد المعجمية، ومعالجتها معالجة أولية، ووضعها في صيغة مناسبة.
- التحليل Parsing: ويشمل تحليل النصوص المعجمية لتحديد مكونات البنية الصغرى لكل مدخل، بتحديدته وتمييزه؛ ويشمل ذلك:



• تحليل مكونات البنية الكبرى Macrostructure للمداخل المعجمية مفردةً ومركبة.

• تحليل مكونات البنية الصغرى Microstructure لكل مدخل معجمي: ويشمل فصل المعاني، والتعريفات، والشروح وتمييزها، وتحديد المتصاحبات اللفظية لكل معنى، ثم الأمثلة.

- التوسيم Annotation:

توسم المداخل المعجمية ببعض الوسوم؛ استكمالاً للمعلومات التي قد تكون ناقصة في بعض الموارد المعجمية المعتمدة في المدونة النصية المعجمية، على أهميتها في مراحل الدمج بين الموارد المعجمية.

ثم إضافة المعلومات الضرورية، لاستكمال المعلومات المعجمية الدلالية، وتتميم توصيف المداخل والمعاني تمهيداً لدمج هذه المداخل والمعاني بين المعاجم.

- الدمج Merging:

ويكون بدمج مخرجات التحليل بين الموارد المعجمية المختلفة، ويشمل ذلك:

• دمج المداخل: لبناء قائمة المداخل الرئيسة (وحدات معجمية مفردة، ومركبة)، واستكمال معلوماتها الواصفة (الصرفية والهجائية والنطقية والتأنيلية) المتاحة.

• دمج المعاني: لجمع كل المعلومات المتعلقة بمعنى في مجموعة واحدة: المعنى، وما جُمع له من معلومات تركيبية، ومتصاحبات لفظية، وأمثلة، إضافة إلى معلومات استعماله؛ إن وجدت.

ولعل أحد أهم أهداف LMF هو اقتراح صيغة حاسوبية تساعد في استكمال المعلومات بين الموارد المعجمية، والتمكين من دمج عدد من الموارد المعجمية المختلفة في مورد معجمي واحد متكامل، تقدّم معلوماته في صورة مطردة، ومتسقة، ومتكاملة (Francopoulo 2013)، لتشمل -إلى جانب المعلومات الأساسية- القيود

الانتقائية، والحقول الدلالية، والعلاقات الدلالية، والمتلازمات اللفظية، والمتصاحبات اللفظية.

١,٣. أهمية العمل

- منهجيتا التحليل الحاسوبي **parsing**، والتوسيم **Annotation**، تستهدف استخلاص المعلومات الصريحة والضمنية، واستكمال المعلومات من الموارد المعجمية المختلفة، واستكمال الثغرات، وتوحيد لغة المورد المعجمي الشامل عن طريق التوسيم.
- قاعدة البيانات المعجمية الدلالية ستكون مصدراً يضاف إلى الموارد المعجمية العربية الحاسوبية، باحتوائها المعلومات والمعارف الموجودة في المعاجم التقليدية التي حُوّلت إلى صورة منتظمة مجدولة، قابلة للقراءة آلياً وبشرياً.
- أنها منهجية تربط المعجمية التقليدية بالعمل المعجمي الدلالي الحاسوبي؛ فلا تركز للقطعة بينهما؛ فتؤدي إلى تطوير العمل المعجمي الحاسوبي.
- أن المنهجية المقترحة توفر وقتاً وجهداً كبيرين مقارنة بالموارد المعجمية الدلالية الحاسوبية التي تنجز ابتداءً دون اعتماد على معاجم عمل عليها لغويون ومعمجون متخصصون.

١,٤. الأعمال المشابهة

سبق هذا العمل أعمالاً استهدفت تحليل النص المعجمي، بغرض تحويله إلى قاعدة بيانات معجمية (هواري، عبدالعاطي. ٢٠٠٨)، أو إلى معجم مقروء آلياً Machine Readable Dictionary (MRD)، (Khemakhem et al, 2009) ولكنها أعمال استهدفت معجماً أو صنفاً معجمياً محدداً، واكتفت بتحليل النص المعجمي تحليلاً شكلياً اعتماداً على الرموز والكلمات المفتاحية والمعلومات الموردة في المتن المعجمي بصورة صريحة. أما العمل الذي تصفه ورقتنا البحثية فقد ركز على استنتاج المعلومات المعجمية الدلالية الضمنية في النصوص المعجمية، إلى جانب المعلومات الصريحة. (ب) أضاف



مستوى للتوسيم الصرفي المعجمي والتركيبى والدلالي. ج) اعتمد أنماطاً متنوعة من الموارد المعجمية. د) أورد منهجية الدمج المقترحة.

١,٥. بنية الورقة البحثية

تتكون الورقة البحثية بعد هذا التمهيد من أربعة أقسام، القسم الأول للمدونة النصية المعجمية المعتمدة، والقسم الثاني لمنهجية تحليل النص المعجمي، والقسم الثالث لمنهجية توسيم المتن المعجمي، أما القسم الرابع فلمنهجية دمج مخرجات التحليل في مورد معجمي واحد.

المدونة النصية المعجمية

في هذا القسم نعرض لأصناف الموارد المعجمية العربية؛ بالتركيز على الأصناف المعتمدة في مدونة الدراسة. ونوضح أهم إشكالاتها من منظور معجمي، ثم نستعرض أهم أنماط المعلومات المعجمية الدلالية المقدمة في هذه الأصناف المعجمية الدلالية.

٢,١. أصناف الموارد المعجمية العربية ومدونة العمل

تمثل أصناف الموارد المعجمية مجموعة مناهج معجمية متنوعة؛ فهي تعرض أنماطاً متباينة من المعلومات المعجمية (معجم لفظي، معجم معان، مكنز مترادفات ومتضادات...) كما تختلف معاجم الصنف الواحد في منهجية المعالجة؛ من حيث طبيعة مصادرها المعتمدة، والمنهجية المعجمية تنظيراً وتطبيقاً، ونوع المعلومات وعمقها، ولغة عرضها، ومنهجية تنظيمها، وطريقة تنظيم البنية الكبرى (مثل اختيار المداخل الكبرى)، والبنية الصغرى.

ولقد اختيرت ثلاثة أصناف من الموارد المعجمية تمثل أهم اتجاهات الصناعة المعجمية، كما يلي:

١) معاجم ألفاظ: وهي المعاجم التي تبني مداخلها (بنيتها الكبرى) انطلاقاً من اللفظ أو الكلمة، ليعرض لكل وحدة معجمية (بسيطة أو مركبة) معلوماتها الواصفة: الصرفية،

والنحوية، والدلالية، والاستعمالية. واختير ممثلاً لهذا الصنف ثلاثة معاجم هي: المعجم الوسيط، ومعجم اللغة العربية المعاصرة، ومعجم الغني.

٢) معاجم المعاني: وهي الموارد المعجمية التي تنظم بنيتها الكبرى انطلاقاً من المعنى، فتورد الألفاظ تالية للمعاني. وهي في ذلك تفترض منهجية لتنظيم المعاني تنظيمًا منطقيًا تؤسس به البنى الكبرى، ويمثل هذا الصنف معجم الإفصاح في فقه اللغة، ويمكن اعتداد كتاب فقه اللغة وسر العربية، ضمن هذا الصنف؛ لكونه ينظم مادته انطلاقاً من المعنى تحت عناوين موضوعية.

٣) المكنز: وهي الموارد المعجمية المخصصة لعرض مترادفات اللغة ومتضاداتها. وتمثلها في المدونة: المكنز الكبير، ومكنز اللغة العربية المعاصرة.

ويمكن إيراد بعض الملاحظات على الأعمال المعجمية التقليدية منها:

- غياب الأسس التي يبني عليها فصل المعاني، خصوصاً في التعامل مع الوحدة المعجمية متعددة الكلمات كالأفعال العبارية، والمتلازمات اللفظية، إضافة إلى التعامل مع المتصاحبات اللفظية، والخلط بينها وبين المداخل متعددة الكلمات.
- الخلط بين المشترك اللفظي وتعدد المعنى.
- نقص المعلومات وعدم الاطراد.
- مشكلات تحقيق الاستيعاب.

٢,٢. أنماط المعلومات المعجمية الدلالية في الموارد المعجمية

في هذا القسم نقدم وصفاً سريعاً لأنماط المعلومات المعجمية الدلالية موضوع الدراسة. وفيه نقسم الاطرادات المعجمية الدلالية التي يمكن استخلاصها مباشرة، أو استنتاجها من لغة الموارد المعجمية، إلى مجموعتين رئيسيتين: الأولى مجموعة متعلقة بالمداخل المعجمية بوصفها معرفات، ويشمل المعلومات الاشتقاقية، والتصريفية، والقسم الكلامي، والثانية مجموعة المعلومات المتعلقة بالمعاني والمحتوى المعجمي الدلالي.



٢,٢,١ المعلومات المتعلقة بالمعرفات / المداخل المعجمية

(١) المعلومات الصرفية:

أ. الاشتقاق:

- الجذر:

يستخدم الجذر اللغوي رأساً للأسرة الاشتقاقية في معظم المعاجم اللفظية؛ فيُجمع تحته مشتقات الأسرة الواحدة، أفعالاً، وصفات، ومشتقات اسمية. ويقدم الجذر في الموارد المعجمية العربية اللفظية باعتباره وحدة لبناء البنية الكبرى للمعجم (كما في المعجم الوسيط، ومعظم المعاجم العربية)، أو بالنص عليه بصورة مباشرة بعد المدخل (كما في معجم الغني)، أو بالدمج بين الطريقتين (كما فعل معجم اللغة العربية المعاصرة) الذي اتخذ من الجذر أساساً لتنظيم البنية الكبرى، وأورد الجذر بعد كل مدخل.

- الصيغة الصرفية:

تقدم بعض المصادر المعجمية معلومة الوزن الصرفي؛ عندما يكون ذلك ضرورياً؛ لغاية معجمية دلالية، أو للمساعدة في ضبط نطق الكلمة (معلومات النطق). ولكن تذخر التعريفات والشروح المعجمية بالمعلومات المتعلقة بالصيغة الصرفية، وتظهر عند إرادة التفريق بين المعاني المختلفة للصيغ الصرفية، وتبين الفروق المعجمية الدلالية بين وزن وآخر، والعلاقات البينية بين صيغتين صرفيتين.

- إيراد الأسرة الاشتقاقية للمدخل المعجمي وتحديد المقولة الصرفية المعجمية:

الأسرة الاشتقاقية لمدخل هي كل المشتقات التي تشترك في جذر لغوي واحد، ونقصد بالمقولة الصرفية المعجمية: تصنيف المشتقات الاسمية بحسب طبيعتها الصرفية المعجمية. مثل: اسم آلة، اسم تفضيل، اسم زمان، اسم فاعل، اسم مؤنث منسوب، اسم مَرَّة، اسم مفعول، اسم المكان، الاسم المنسوب، اسم الهيئة، التصغير، الصفة الثابتة

للمفعول، الصفة المشبهة، صيغة المؤنث لاسم الفاعل، صيغة المؤنث لاسم المفعول، صيغة المبالغة، المصدر، المصدر الصناعي، المصدر الميبي.

ب. التصريف Inflection

وتشمل معلومات التصريف: عرض تصريفات المدخل؛ لاستيعاب كل الصيغ التصريفية التي تكون

الليكسيم Lexeme (الوحدة المعجمية بتصريفاتها الممكنة) وتشمل -ضمن ما تشمل- الجوانب التالية:

- تصريفات الفعل: معلومات الزمن (ماض ومضارع، وأمر)، ومعلومات إسناد الأفعال للضمائر التي يلتزم فيها قواعد خاصة عند الإسناد، مثل المعتل والمضعف والمهموز.

- تصريفات الاسم: ويتضمن معلومات العدد (الإفراد والتثنية والجمع)، والنوع (التذكير والتأنيث).

٢) المعلومات الهجائية:

وتركز المعاجم على عرض المعلومات الهجائية الخاصة بظاهرتين هجائيتين: - الكلمات البصرية أو الكلمات التي لها قواعد خاصة تخالف منهجية الهجاء المعروف: مثل الكلمات التي بها حروف تنطق ولا تكتب، والكلمات التي بها حروف تكتب ولا تنطق.

- الكلمات ذات التنوعات الهجائية المقبولة: مثل الكلمات المعربة التي تكتب بأكثر من طريقة (جوجل - غوغل - قوقل...)، وبعض حالات رسم الهمزة التي ترسم بأكثر من طريقة (شؤون/شئون).

٣) المعلومات النطقية:



تحاول الموارد المعجمية العربية على ضبط نطق المداخل من خلال التشكيل أو النص أو بتحديد الوزن الصرفي، أو إعطاء مثال مشابه للكلمة المدخل.

٤) المعلومات التأصيلية Etymology:

تعرض الموارد المعجمية بعض المعلومات المتعلقة بأصل الكلمة (معرّبة، أو دخيلة)، وتحديد أصل بعض المعاني (المولّدة). إضافة إلى التأصيل لبعض الكلمات المنحوتة أو المشتقة من أصول أخرى.

٥) المعلومات المعجمية:

توظف المعلومات المعجمية في تحديد المداخل وتصنيفها، وتمييز المداخل الرئيسية من المداخل الفرعية، والتمييز بين متعدد المعنى (Polysemy) والمشارك اللفظي (Homonymy)، وتحرير المتصاحبات اللفظية (Collocation).

٦) معلومات الإعراب:

مثل: الإشارة إلى المداخل الممنوعة من الصرف، وبعض حالات البناء، والإعراب الفرعي... إلخ.

٧) التأثيل Etymology:

التأثيل هو توضيح أصل الكلمة: عربية، معربة، أجنبية (دخيلة)، مفصّحة... إلخ، وقد يشمل ذلك المعنى.

٢,٢,٢ المعلومات المتعلقة بالمعاني المحتوى المعجمي الدلالي

١) المعاني:

يعد المعنى الوحدة الثانية في التحليل، بعد تحديد المدخل المعجمي، ويستهدف العمل فصل المعاني وترقيمها، عند تعدد المعنى للمدخل المعجمي الواحد. كما يعد الأساس الذي تنسب إليه البنية الحملية والسلوك التركيبي للوحدة المعجمية، وبه تدخل

الوحدة المعجمية في العلاقات الدلالية المختلفة، وتنتمي لحقل دلالي بعينه، وإليه تضاف معلومة الاستعمال (usage).

٢) المعلومات النحوية Grammatical information:

- التعدي واللزوم Transitivity.
- البنية الحملية Argument Structure: هي البنية النحوية التركيبية الممكنة للمدخل المعجمي، وتعني تمثيلاً وصياغة للسلوك التركيبي للفعل أو ما يشبهه. ، وتحدد طبيعة الفاعل وطبيعة المفاعيل الممكنة لكل بنية حملية. علماً بأن البنية الحملية مرتبطة بالمعنى الواحد؛ لا بالمدخل المعجمي كله.
- القيود الانتقائية / التفضيلية.

٣) العلاقات الدلالية Semantic Relation:

تُظهر الموارد المعجمية العلاقات الدلالية، مثل الترادف والتضاد، والتضاد، وعلاقة الجزء بالكل، والكل بالجزء...إلخ. والموارد المعجمية العربية، إذ تورد هذه المعلومات؛ فإنها توردها في صورة غير نظامية، ولكن يمكن استنتاجها من نصوص التعريفات المعجمية والشروح.

٤) الحقول الدلالية Semantic Field:

لا تنصّ معظم الموارد المعجمية على معلومة الحقل الدلالي للمداخل المعجمية؛ غير أن الاطرادات القارة في التعريفات والشروح المعجمية يمكن توظيفها في تصنيف المداخل بحسب حقولها الدلالية، وجميع المداخل التي يضمها حقل دلالي ما.

٥) الاستعمال Usage:

معلومات الاستعمال لا تورد في الموارد المعجمية العربية إيراداً منتظماً، لكنها متواجدة في بعض المواضع؛ لتحديد مستوى الاستعمال للمدخل المعجمي/المعنى، وتحدد المستوى الذهني للكلمة، وربما طبيعة الموقف الذي يمكن أن تقال فيه الكلمة.



(٦) الأمثلة:

تعامل الأمثلة باعتبارها مدونة فرعية للعمل، مربوطة بالمعاني التي قصد التمثيل عليها بهذه الأمثلة.

٢,٣. أنماط الاطرادات المعجمية الدلالية في معاجم المعاني

تنظم معاجم المعاني مادتها انطلاقاً من المحتوى الدلالي للوحدات المعجمية؛ ولكنها تقدم شروحاً -وإن كانت مختصرة- لكل مدخل من المداخل التي تندرج تحت الحقل الدلالي؛ فتبدأ بتحديد الباب (الحقل الدلالي الأوسع) ثم تأخذ في التفرع حتى تصل إلى قائمة المداخل المعجمية التي تندرج تحت كل حقل دلالي فرعي؛ مما يجعل من الممكن إعادة بناء المورد المعجمي بناء لفظياً دلاليّاً. ومن ذلك أيضاً اطرادات الأنماط المعجمية الدلالية في معاجم المعاني.

ويمكن استخلاص المعلومات المعجمية والدلالية والتركيبية من متن كتاب فقه اللغة والكتب المشابهة للظواهر الآتية: المتصاحبات اللفظية **Collocation**، القيود الانتقائية **Selectional Restrictions**، المتلازمات اللفظية **Multiword expressions**، الفروق الدلالية، العلاقات الدلالية **Semantic Relations**، الحقول الدلالية **Semantic Fields**، التعريفات، المعلومات الصرفية.

٢,٤. أنماط الاطرادات المعجمية الدلالية في المكانز

تورد المكانز قوائم المترادفات (والمضادات). وتعد المصدر الأهم في رصد العلاقات الدلالية، ويستعان بها في استكمال هذا النمط من المعلومات الدلالية في الموارد المعجمية اللفظية؛ فالعلاقات الدلالية تقدم فيه بصورة غير متسقة، أو تقدم فيها ضمن الشرح أو التعريف.

تحليل النص المعجمي العربي

غرض تحليل النص المعجمي (Parsing) إلى وحداته الأساسية المكونة، وتعيين أصنافها؛ هو تحويل النص السردي إلى بنية مجدولة تصلح لبناء معجم مقروء آلياً (Machine Readable Dictionary). ولقد اعتمدنا منهجيتين للتحليل الأولى آلية، والثانية يدوية. وكانت غاية التحليل أ) استخراج قائمة المداخل المعجمية؛ مفردة ومتعددة الكلمات.

ب) استخراج المعلومات الخاصة بكل مدخل معجمي. ج) فصل معاني متعدد المعنى. د) تمييز المعلومات المتعلقة بكل معنى من المعاني، بما في ذلك الأمثلة والمتصاحبات اللفظية المتعلقة بكل معنى.

٣,١ قائمة المداخل المعجمية: البنية الكبرى

- تمييز المشترك اللفظي (مدخل معجمية منفصلة)، ومتعدد المعنى (مدخل واحد معان مختلفة).

- تحديد معايير فصل المداخل المعجمية؛ مفرداً ومتعدد الكلمات:

- تمييز المتصاحبات اللفظية عن المتلازمات اللفظية.
- تمييز المتصاحبات اللفظية عن الأمثلة العادية (معالجة مشكلة تقييد المطلق، وإطلاق المقيد في الشروح).
- تمييز الأفعال العبارية؛ حيث إن الموارد المعجمية العربية لا تكاد تميز بين الأفعال العبارية، والأفعال الاعتيادية^(١).

٣,٢ تحليل البنية الصغرى

تحليل النص المعجمي يقتضي تمييز عناصره، ومداخله المعجمية، ومعلومات المعجمية الدلالية، وتمييزها. وغاية تحليل النص المعجمي (Parsing) هي تحديد وحداته



الأساسية المكونة، وتعين أصنافها؛ بغرض تحويل هذا النص السردى إلى بنية مُجدولة تصلح لبناء معجم مقروء آلياً (Machine Readable Dictionary). ونتبع منهجية للتحليل تمزج بين التحليل الآلي (للمعلومات الصريحة). والتحليل البشري اليدوي (للمعلومات الضمنية).

وتختلف وحدات التحليل باختلاف طبيعة الموارد المعجمية المحللة، فتحليل المعجم اللفظي يشمل وحدات التحليل التالية:

- المدخل المعجمي (المفرد، والمركب)، ومعلوماته الصرفية، والتأيلية، والنطقية، والهجائية.
- المعنى المفرد للمدخل المعجمي، ومعلوماته التركيبية والدلالية والاستعمالية، وأمثله.

أما تحليل معاجم المعاني (والمكانز) فيستهدف ما يلي:

- فصل المعاني وترقيمها لكل مدخل، ولكل مصدر معجمي definition parsing.
- تحديد الحقول الدلالية، وحقولها الفرعية.
- قوائم المترادفات والمتضادات.
- تحديد الوحدات المعجمية، ومعلوماتها الصرفية والنحوية، والدلالية والاستعمالية.

٣,٢,١ أدوات التحليل

- يفيد التحليل الآلي للنص المعجمي من منهجية التحرير المتبعة في كل معجم: لتعيين حدود المدخل المعجمي (بدايته ونهايته)، وتحديد أنماط المعلومات المعجمية الدلالية. ومن بين العلامات والصور التي تفيد في عملية التحليل الآلي، ما يلي:
- علامات الترقيم والأقواس بأنواعها: (،)، [،]، «، »، "، "، {، }، *... إلخ.
 - الرموز والمختصرات المستخدمة في توضيح النص المعجمي.

أما التحليل اليدوي؛ فيكون بجمع الأنماط المتشابهة والمتكررة في لغة الشروح والتعريفات، من خلال رصد التكرار الإحصائي للكلمات (N-gram)، على مستوى الكلمة (unigram) أو الكلمتين (bigram) أو الكلمات الثلاث (trigram).

٣,٢,٢ إجراءاته

يمكن تلخيص المستهدف من إجراء تحليل النص المعجمي في الخطوات التالية:

(١) استخلاص قائمة المداخل:

انطلاقاً من المعاجم اللفظية؛ تحدّد المداخل البسيطة والمداخل المركبة. وتشمل قائمة المداخل كل كلمة مفردة لها شرح أو تعريف، أو وردت تحت حقل دلالي، أو وردت في علاقة دلالية في الموارد المعجمية. وفي سبيل توصيف قائمة المداخل نضيف المستويات الآتية: استكمال الملفات الصرفية: المعلومات الاشتقاقية والتصريفية، والقسم الكلامي، والهجائي، والنطقي لكل مدخل معجمي.

(٢) فصل المشتركات اللفظية (Homonymy): بإعطاء كل مشترك لفظي مدخلاً معجمياً مستقلاً في البنية الكبرى للمعجم.

(٣) تحديد قائمة المداخل المركبة: قائمة الوحدات المعجمية متعددة الكلمات.

(٤) فصل معاني المداخل متعددة المعنى (Polysemy): فصل المعاني/تحليل المعاني/تحليل التعريفات: بحيث نخلص إلى تعريف مفرد واحد لكل معنى.

(٥) استكمال المعلومات التركيبية لكل معنى.

(٦) جمع معلومات الاستعمال الموزدة لكل معنى، إن وجدت.



٣,٣. تحليل معجم الألفاظ

٣,٣,١ البنية الكبرى: قائمة المداخل المعجمية

البنية الكبرى لمعجم الألفاظ تتضمن أكثر من حلقة: الأولى هي حلقة الجذور، وتحت كل جذر يوجد عدد من المداخل. الوحدة الأساسية للبنية الكبرى هي الكلمة المفردة، وتُجمّع الكلمات تحت جذورها، كما أن كل مدخل بسيط يضم تحته المداخل الفرعية متعددة الكلمات. ويفترض في البنية الكبرى أن تضم القائمة المعجمية للمداخل المعجمية المفردة، ومتعددة الكلمات.

٣,٣,٢ البنية الصغرى: أنماط المعلومات المعجمية الدلالية

وتتضمن أنماط المعلومات المختلفة المقدمة لكل مدخل معجمي، والمعاني، والاستعمالات المرتبطة بكل مدخل.

٣,٤. تحليل معجم المعاني

- البنية الكبرى: هيكلية المعجم: الحقل الدلالي وتفرعاته.
- قائمة مداخل كل حقل فرعي.
- البنية الصغرى: المعلومات المقدمة لكل مدخل معجمي.

٣,٥. تحليل المكنز

- البنية الكبرى: قائمة المترادفات والمتضادات.
- البنية الصغرى: بنية قائمة المترادفات أو المتضادات، والمعلومات الصرفية، والتركيبية المصاحبة (كالقسم الكلامي لكل قائمة... إلخ).
- الرموز والاختصارات.
- التقويس.
- الميثلة لغة/المصطلحات الشارحة.

- لغة التعريف.
- الأمثلة.
- تسييق كلمة المدخل.

التوسيم واستكمال المعلومات المعجمية الدلالية

يفترض في المعلومات المعجمية الدلالية التي توردها الموارد المعجمية الاكتمال والوضوح، غير أن تأمل واقع التطبيق المعجمي يظهر لنا خلاف ذلك؛ فبعض المعلومات قد تقدم صريحة في مورد معجم دون آخر، وقد تقدم تقديمًا غير مطرد، أو متسق؛ لذلك افترضنا إضافة مستوى لتوسيم المداخل بالمعلومات الضرورية استكمالاً لتوصيفها، وإعداداً لها لمرحلة الدمج التي تطلب جوانب مختلفة من المعلومات؛ لإنجاز الدمج بصورة أدق.

والتوسيم (Annotation) هو إضافة معلومات لغوية للوحدات المعجمية أو لمعانيها؛ نستهدف منه تحقيق غايتين ضمن المشروع: الأولى هو استكمال المعلومات الضرورية لبناء مورد معجمي موسع. والثانية المساعدة في إنجاز إجراء الدمج، إذ إن الدمج قد يحتاج لعدد من المعلومات؛ للتأكد من تطابق المداخل المعجمية والمعاني المدمجة.

٤,١. منهجية التوسيم

يُلْتزم في إجراءات التوسيم الإفادة من المعلومات الضمنية التي توردها الموارد المعجمية، واستكمال المعلومات الناقصة من الموارد الأخرى داخل المدونة النصية المعجمية المعتمدة. كما يلتزم باعتماد دليل توجيهي مرجعي لكل إجراء توسيمي؛ يوثق للإجراء، ويوضح المقصود بكل وسم في قائمة الوسوم (Annotation Tag-set) التي أعدت لها أدلة توجيهية لإدارة عملية التوسيم. وتشمل هذه المرحلة توسيم المداخل، وتوسيم المعاني، وتوسيم الأمثلة.



٤,٢. أنواع التوسيم

تشمل مرحلة التوسيم عدة مستويات، يتوجه بعضها إلى المداخل المعجمية؛ مثل التوسيم الصرفي، والمعجمي، والنطقي، والهجائي، والتأنيلي. كما يتوجه بعضها إلى المعنى ومتعلقاته في البنية الصغرى. مثل التوسيم التركيبي، والدلالي، وتوسيم الأمثلة.

٤,٢,١ التوسيم الصرفي

(١) أقسام الكلام (Part of Speech)

لا تلتزم معظم موارد المدونة المعتمدة بإيراد معلومة أقسام الكلام. ومن يوردها من هذه المصادر؛ يوردها في صورة مقتضبة مكتفياً بأقسام الفعل، أو الاسم، أو الأداة، أو يكتفي بالدلالة الضمنية؛ كحديث معجم اللغة العربية المعاصرة عن قسم (الفعل) "...ولا ينصّ على ذكر النوع أمامه. ويُكتفى بأن المعلومات الصرفية الخاصة به تكشف عنه، كالفعل المضارع الذي يلي المدخل الفعلي مباشرة" (عمر، أحمد مختار. ٢٠٠٨: ص. ١٣)^(١)، إضافة إلى عدم تعيين مفهوم بعض الأقسام، كما في تعامل معجم اللغة العربية المعاصرة مع قسم الأداة، وقد حددت مقدمة المعجم مفهوم الأداة بأنها: "كلمات اكتسبت دلالة جديدة بعيدة عن الدلالة اللغوية لألفاظها، وتشمل حروف الهجاء وجميع حروف الجرّ، وأدوات الاستفهام، والأسماء الموصولة، وأسماء الإشارة، وأدوات الشرط، والظروف، وأسماء الأفعال. كما اشتملت على بعض الأفعال الجامدة، مثل: عسى... إلخ" (عمر، أحمد مختار. ٢٠٠٨)، ولا يخفى الخلط بين أقسام الكلام المختلفة تحت مصطلح الأداة، واعتداد حروف الهجاء أدوات. في مقابل إعطاء الاسم تفاصيل تصنيف (المفرد، والمتنّ، والجمع) وهي من معلومات التصنيف لا أقسام الكلام الرئيسية.

وفي هذا المشروع نتبنى قائمة أقسام الكلام التالية: الاسم Noun (ويشمل -أيضاً- الأصناف الفرعية: اسم موصول، اسم استفهام، اسم إشارة...)، والصفة Adjective،

والفعل Verb، وحرف الجر Preposition، والأداة Particle، والظرف Adverb، والعلم Proper-noun، والاختصار Abbreviation.

لا تقدم معلومات أقسام الكلام في المعاجم العربية التقليدية (المعتمدة في هذه المدونة)، وإن قدمت فتكون غير كاملة، أو غير مطردة. إضافة إلى أن قوائم أقسام الكلام المستخدمة غير مكتملة؛ فتكتفي بالتصنيف إلى اسم، وفعل، وحرف (أداة).

وتعتمد الدراسة إجراء توسيم المداخل بإضافة وسوم أقسام الكلام إلى المداخل المفردة (البسيطة) وتشمل القائمة المعتمدة أقسام الكلام التالية: (فعل، صفة، اسم، علم)؛ للقوائم المفتوحة، إضافة إلى القوائم المغلقة: (أداة، اسم موصول، اسم إشارة...).

٢) استكمال المعلومات الاشتقاقية:

- إضافة الجذر؛ إن لم يكن موجوداً في المورد المعجمي.
- إيراد أفراد الأسرة الاشتقاقية.
- إضافة الصيغة الصرفية للأفعال والمشتقات الفعلية.

٣) استكمال معلومات التصريف: لتتيمم مكونات الليكسيم Lexeme (تصريف الفعل، وتصريف الاسم، والصفة...)

٤) إضافة الأوزان الصرفية: لمداخل الأفعال ومشتقاتها ومصادرهما.

بعض الموارد المعجمية تورد -دون اطراد- أوزان بعض الأفعال، أو الصفات، أو الأسماء. وتستهدف المنهجية المقترحة تحديد الصيغ الصرفية لكل الأفعال والصفات، والأسماء المشتقة. أُضيف الصيغ الصرفية لكل الأفعال والصفات والأسماء؛ لتحقيق الغاية من رصد الاطرادات المعجمية الدلالية القارة في الصيغة الصرفية، وتصنيف الأفعال تصنيفاً معجمياً دلالياً. إلى جانب تحقيق بنية تحتية للعلاقات المعجمية بين الأفعال ومشتقاتها.



٤,٢,٢ استكمال معلومات الهجاء والنطق

(١) معلومات الهجاء: ويتضمن ذلك جانبيين، هما: كيفية كتابة المدخل المعجمي، واستكمال الصور الهجائية المقبولة.

(٢) معلومات النطق: ويتضمن ذلك جانبيين، هما: كيفية نطق المدخل المعجمي، واستكمال الصور النطقية المقبولة؛ إن وجدت.

٤,٢,٣ التوسيم النحوي الإعرابي

توسيم المداخل المعجمية التي لها خصائص إعرابية مهمة، مثل: معلومات البناء، والإعراب الفرعي، والممنوع من الصرف...إلخ.

٤,٢,٤ التوسيم المعجمي

يتضمن تصنيف المداخل المعجمية بحسب طبيعة الشكل، أو المحتوى الدلالي.

(١) التوسيم المعجمي للمداخل بحسب طبيعة شكلها:

(أ) المدخل المعجمي المفرد، ويشمل كلا من الكلمة (Word)، والمختصر

(Abbreviation)

(ب) المدخل المعجمي، متعدد الكلمات (Multi-word Expression)، كما تتفرع

المداخل المعجمية متعددة الكلمات إلى عدة أصناف فرعية، فتشمل ما يلي:

- التعبير الاصطلاحي (Idiom)، وتمييزه عن المتصاحبات اللفظية.

- الفعل العباري (Phrasal Verb)، وفصله عن الأفعال غير العبارية.

- الاسم المركب (Compound noun).

(٢) التوسيم المعجمي للمداخل بحسب طبيعة محتواها الدلالي:

ويتضمن توسيم المداخل المعجمية (أو المعاني عند التعدد الدلالي)، بأحد

التصنيفات التالية:

- (أ) الوحدة المعجمية المصطلحية: ويضاف هذا الوسم للمداخل المصطلحية المقصور استخدامها على مجال تقني وفي أو علمي بعينه.
- (ب) الوحدة المعجمية الموسوعية: وتشير إلى الوحدة المعجمية التي تعبر عن معلومة موسوعية كالمواضع الجغرافية، أو تصف حيواناً أو نباتاً، أو تشير إلى علم، أو حدث تاريخي، أو منتج فني، أو حضاري... إلخ.
- (ج) الوحدة المعجمية العامة: وهي الوحدة المعجمية التي تنتهي إلى اللغة والاستعمال العام؛ فلا تكون مصطلحاً مقصوراً في استخدامه على مجال تقني وفي أو علمي بعينه، أو وحدة موسوعية تشير إلى معلومة موسوعية.
- (٣) تمييز المتصاحبات اللفظية عن المتلازمات اللفظية.
- (٤) تمييز المتصاحبات اللفظية عن الأمثلة العادية (معالجة مشكلة تقييد المطلق، وإطلاق المقيد في الشروح).

٤,٢,٥ التوسيم النحوي التركيبي

(١) التعدي واللزوم

تنص بعض الموارد المعجمية على معلومة التعدي واللزوم للفعل، وقد تُقدّم هذه المعلومة ضمناً، بحيث يمكن استنتاجها. على أن معلومة التعدي واللزوم معلومة مهمة في توضيح السلوك التركيبي للفعل من جهة، وتمييز المعنى الذي يقدمه المورد المعجمي من جهة أخرى.

كما أن المعاجم التي تنص على التعدي واللزوم قد تخلط بين التعدي واللزوم للمعاني المختلفة. ويمكن أن نضرب لذلك مثلاً لمدخل الفعل (أبدع) من معجم الغني (أبو العزم، عبدالمغني. ٢٠١٣: ص. ٣٥):



جدول (١)

مدخل الفعل أَدْعَ

"أَدْعَ - [ب د ع]. (فعل رباعي لازم متعد بحرف). أَدْعْتُ، أَدْعُ، أُدْعُ، مصدر إِدْعَاً. ١. "أَدْعَ الصَّانِعُ": أَجَادَ، أَتَقَنَ. ٢. "أَدْعَ آلَهُ": أَنشَأَهَا، إِخْتَرَعَهَا. "أَدْعُوا حَوَاسِبَ جَدِيدَةً فِي عَالَمِ الْمَعْلُومَاتِيَّةِ". ٣. "أَدْعَ الشَّاعِرُ": جَدَّدَ، أَتَى بِالْبَدِيعِ مِنَ الْقَوْلِ. ٤. "أَدْعَتِ الطَّبِيعَةُ مَنَاطِرَ خَلَابَةٍ سُبْحَانَ خَالِقِهَا": أَوْجَدَتْ. ٥. "أَدْعَ الرَّجُلُ": أَتَى بِدَعَةٍ. ٦. "أَدْعَ بِهِ": أَهْمَلَهُ."

ففي هذا المثال نعيد توسيم المعاني كما يلي:

أَدْعَ [ب د ع]. (فعل: رباعي). أَدْعْتُ، أَدْعُ، أُدْعُ، مصدر إِدْعَاً.

١. "أَدْعَ الصَّانِعُ" (لازم): أَجَادَ، أَتَقَنَ.

٢. "أَدْعَ آلَهُ" (متعدٍ): أَنشَأَهَا، إِخْتَرَعَهَا. "أَدْعُوا حَوَاسِبَ جَدِيدَةً فِي عَالَمِ الْمَعْلُومَاتِيَّةِ".

٣. "أَدْعَ الشَّاعِرُ" (لازم): جَدَّدَ، أَتَى بِالْبَدِيعِ مِنَ الْقَوْلِ.

٤. "أَدْعَتِ الطَّبِيعَةُ مَنَاطِرَ خَلَابَةٍ سُبْحَانَ خَالِقِهَا" (متعدٍ): أَوْجَدَتْ.

٥. "أَدْعَ الرَّجُلُ" (لازم): أَتَى بِدَعَةٍ.

٦. "أَدْعَ بِهِ" (متعدٍ بحرف): أَهْمَلَهُ.

جدول (٢)

إعادة بناء المدخل: أبدع

	المدخل	متعدٍ/لازم	المعنى
١.	أَبْدَعَ الصَّانِعُ	لازم	أَجَادَ، أَتَقَنَ
٢.	أَبْدَعَ آلَهُ	متعدٍ	أَنْشَأَهَا، إخْتَرَعَهَا. "أَبْدَعُوا حَوَاسِبَ جَدِيدَةً فِي عَالَمِ الْمُعْلُومَاتِيَّةِ"
٣.	أَبْدَعَ الشَّاعِرُ	لازم	جَدَّدَ، أَتَى بِالْبَدِيعِ مِنَ الْقَوْلِ
٤.	أَبْدَعَتِ الطَّبِيعَةُ مَنَاطِرَ خَلَابَةٍ سُبْحَانَ خَالِقِهَا	متعدٍ	أَوْجَدَتْ
٥.	أَبْدَعَ الرَّجُلُ	لازم	أَتَى بِدُعَاةٍ
٦.	أَبْدَعَ بِهِ	متعدٍ بحرف	أَهْمَلَهُ

٢) السلوك التركيبي (البنية الحملية (Argument Structure)

لكل فعل سلوك تركيبي مرتبط بالمعنى؛ يحدد عدد الأركان التركيبية التي يتطلبها (الفاعل، المفاعيل...)، وأنواعها. ويعد توصيف البنية الحملية -إلى جانب معلومة التعدي والازم- من متممات تمثيل المعنى والمحتوى المعجمي الدلالي للفعل. وتتبنى معظم المعاجم العربية طريقة ضمنية في تمثيل البنية الحملية. ويأتي توسيم المعلومات التركيبية لتجلية السلوك التركيبي للفعل. ويتضمن ذلك تجريد الأبنية التركيبية وتحوسبها لأنماط مثل: فعل [فلان] [كذا].

ويكون التوسيم باختيار بنية من الأبنية التركيبية المجردة؛ فيمكن استخدامها لاحقاً في تحديد الأنماط التركيبية للأفعال في المعجم العربي.



(٣) المعلومات التركيبية للمداخل النحوية الوظيفية.

٤,٢,٦ التوسيم الدلالي

ويشمل التوسيم الدلالي الجوانب التالية:

- فصل معاني متعدد المعنى Polysemous.
- ربط المتصاحبات اللفظية Collocation، والقيود الانتقائية Selectional Restrictions بالمعاني المتعلقة بها.

- توسيم المعاني، بتصنيفها بحسب الحقول الدلالية التي تنتمي إليها:

المعاجم اللفظية العربية لا تصنف مداخلها المعجمية بحسب الحقول الدلالية، على خلاف معاجم المعاني التي تتأسس بنيتها على تحديد الحقول الدلالية، وحقولها الفرعية. تفترض الدراسة أنه يمكن تصنيف المداخل المعجمية بحسب الحقول الدلالية، من خلال استنباط الروابط الدلالية من متن لغة التعريفات المعجمية.

- إضافة العلاقات الدلالية:

تركز المكانز على العلاقات الدلالية: ترادفًا، وتضادًا، فالعلاقات الدلالية منصوص عليها، وهي الأساس في بناء هذا الصنف من الموارد المعجمية، كما أن المعاجم اللفظية تقدم كثيرًا من الشروح معتمدة الترادف أو التضاد؛ لتمثيل المعاني. على أن هذا المستوى من التوسيم يفيد من أصناف الموارد المعجمية المختلفة في استكمال بنية العلاقات الدلالية.

- الفروق الدلالية بين أشباه المترادفات:

تقدم المكانز العربية الترادف دون اعتناء كبير في التفريق بين المعاني المتقاربة دلاليًا. على أن بعض معاجم المعاني تهتم بالتفريق بين المعاني؛ فعلى سبيل المثال، قائمة المترادفات للفعل: "قطف [ق ط ف]" في المكنز العربي المعاصر، تشمل المترادفات التالية:

(جَنَى، اجْتَنَى، قَطَعَ، حَرَفَ، حَصَدَ)، دون تمييز بين دلالات كل منها. والمصدر الأساسي لمعلومات الفروق الدلالية في مدونتنا هو كتاب فقه اللغة. وفيما يلي مثال لمدخل منه: "حَرَّ اللحم، جَرَّ الصُوف، قَصَّ الشَّعْرَ، عَضَدَ الشَّجَرَ، قَضَبَ الكَرْمَ، قَطَفَ العِنَبَ، جَرَمَ النَّخْلَ، بَرَى القَلَمَ، فَلَحَّ الحديدَ، حَضَدَ النباتَ الرُّطْبَ، حَصَدَ النَّبَاتَ اليابسَ، قَطَعَ الثوبَ، جَابَ الجَيْبَ، قَدَّ السَّيْرَ، حَدَا النَّعْلَ، حَدَقَ الحَبْلَ". فنلاحظ الفروق الدلالية بين المترادفات، وكيف وضحتها المتصاحبات اللفظية، والقيود الانتقائية.

دمج مخرجات التحليل في مورد معجمي واحد

تظهر دقة نتائج تحليل نصوص الموارد المعجمية، في عملية دمج الموارد المعجمية في مورد واحد. ويعد دمج هذه الموارد الغاية النهائية للمشروع. ويسير الدمج في مراحل ثلاث: أ) دمج المعاجم اللفظية. ب) دمج معاجم المعاني. ج) دمج المعاجم اللفظية بمعاجم المعاني؛ لتحقيق فكرة دمج المعجم بالمكنز.

٥,١. دمج معاجم الألفاظ

ونستهدف من مرحلة دمج المعاجم اللفظية تحقيق ما يلي:

- دمج المداخل: بأن يكون لدينا قائمة موحدة شاملة بالمداخل المعجمية (المفردة، والمركبة)، مع معلوماتها الواصفة (القسم الكلامي، المعلومات الصرفية، والهجائية، والنطقية...إلخ).
- دمج المعاني: بأن نورد لكل معنى من معاني المدخل المعجمي المعلومات الدلالية والنحوية المتعلقة بهذا المعنى.

ويحقق الدمج الأهداف التالية:

- سد ثغرات الموارد المعجمية باستكمال المعلومات الناقصة من الموارد الأخرى.
- فصل المعاني من خلال مقارنة التعريفات والشرح.



- فصل متعدد المعنى عن المشترك اللفظي.
 - ترتيب المعاني على أسس نظرية.
 - توسيم عناصر النص المعجمي.
 - توسيم الشواهد والأمثلة وتصنيفها.
 - تصنيف المادة المعجمية.
 - استخلاص المادة الموسوعية.
 - استخلاص المادة المصطلحية.
 - تسهيل فحص لغة المعجمية العربية.
- أما عن آلية الدمج، فإننا نعتمد أصنافاً من المعلومات لكل نوع من أنواع الدمج، كما يلي:
- دمج المداخل: ونعتمد فيه كلاً من: المدخل (الليما Lemma)، الجذر، القسم الكلامي (POS).
 - دمج المعلومات الواصفة: وهي مصنفة في مراحل التحليل، ومستكملة في مرحلة التوسيم.
 - دمج المعاني: ونعتمد فيه كلاً من: البنية التركيبية، المتصاحبات اللفظية، القيود الانتقائية، المعنى، الأمثلة.
- ٥,٢. دمج المعاجم الدلالية
- بعد تحويل المعاجم الدلالية إلى قائمة من المداخل المعجمية مصحوبة بالتصنيفات والمعلومات المعجمية الدلالية المعطاة لها. توسم هذه المداخل بمنهجية التوسيم المتبعة، تجهيزاً لدمجها مع معلوماتها في قائمة المداخل الموحدة؛ بحيث تدمج معلومات الحقول الدلالية، والحقول الدلالية الفرعية، والمعلومات الدلالية الأخرى، لكل مدخل معجمي في القائمة.

٥,٣. دمج المكانز

يجعل لكل كلمة في قائمة المترادفات مدخلاً، ويعطي لها المعلومات التصنيفية المعطاة لرأس قائمة المترادفات، ثم توسيمها بالمعلومات الضرورية؛ لدمجها في القائمة المعجمية العامة.

وينتج عن دمج هذه الموارد المختلفة بناء قاعدة بيانات معجمية دلالية لها صورتا المعجم اللفظي، والمكانز الدلالي.

٥,٤. مخرجاتها المتوقعة

ويوضح الشكل (الملحق أ) المخرجات المتوقعة بعد عمليات التحليل والتوسيم والدمج.

الخاتمة

استهدفت الدراسة تطوير منهجية لمعالجة عدد من الموارد المعجمية العربية التقليدية، بالنظر إليها بوصفها مدونة محوسبة، ومعالجتها معالجة حاسوبية تحليلياً لنصوصها، وتوسيماً لمداخلها؛ واستكمالاً لمعلوماتها الضرورية؛ بغرض دمج الأنماط والمعلومات المعجمية الدلالية المستخلصة من هذه الموارد المعجمية؛ لبناء هيكلية مصدر معجمي دلالي حاسوبي شامل مقروء آلياً؛ يفيد من الثروة اللغوية والمعجمية المنجزة، وتُعرض فيه المعلومات في صورة متكاملة، وتقدم فيها الاطرادات المعجمية الدلالية لبنية المعجم العربي تقديمًا صريحاً مُجدّوًلاً.

انطلق العمل من التعامل مع المصادر المعجمية العربية من قناعة بأنها مصادر ثرية ومتنوعة من حيث أنماطها المعجمية (معاجم عامة، ومعاجم موضوعات، ومكانز



معجمية، ومعاجم قطاعية)، ومن حيث طبيعة المعلومات التي تقدمها. واقتناعاً بأن هذه المصادر المعجمية العربية -رغم ثراء مادتها وعمق معلوماتها - لا تلتزم بنية معجمية مطردة؛ فالبنية الصغرى تفتقر، في كثير من الحالات، إلى الاتساق والانتظام في تقديم المعلومات المعجمية الدلالية؛ إضافة إلى أن كثيراً من الاطرادات المعجمية الدلالية الموجودة في بنية المعجم العربي لا تُعرض بطريقة صريحة متسقة، بل تعرض في صورة ضمنية يمكن استنباطها، وعرضها في صورة صريحة مطردة تُجَلِّي البنية المعجمية الدلالية للمعجم العربي.

ويمكن تلخيص إجراءات العمل في الخطوات المنهجية التالية: أ) اختيار عدد من الموارد المعجمية العربية التي تمثل أصنافاً متنوعة من أنماط الموارد المعجمية، ومعالجتها معالجة أولية. ب) تحليل النص المعجمي للتوصل إلى أنماط المعلومات المعجمية الدلالية التي تقدمها للمداخل المعجمية الرئيسة والفرعية؛ لتحويلها إلى مصادر معجمية مقروءة آلياً. ج) توسيم هذه الموارد المعجمية؛ لاستكمال المعلومات المعجمية الدلالية التي قد تكون ناقصة في كل نمط من الأنماط المعجمية. د) استخلاص أنماط المعلومات الدلالية والعلاقات والاطرادات المعجمية الدلالية في متن هذه الموارد المعجمية. هـ) دمج هذه الموارد المعجمية؛ للتوصل إلى مصدر معجمي متكامل، تقدّم معلوماته في صورة مطردة، ومتسقة، ومتكاملة لتشمل -إلى جانب المعلومات الأساسية- القيود الانتقائية، والحقول الدلالية، والعلاقات الدلالية، والمتلازمات اللفظية، والمتصاحبات اللفظية.

٦. الهوامش

- (١) أمثلة على أعمال انتقدت معالجة المعجم العربي لعلاقة الفعل بحرف الجر؛ ينظر: (مصلوح، سعد، ٢٠٠٠).
- (٢) أورد معجم اللغة العربية المعاصرة في مقدمته عن أقسام الكلام ما يلي: "يصنف المعجم مداخله ويقسمها إلى خمسة أنواع كالتالي: أولها: الفعل... ثانيها: الاسم المفرد، وينصّ بعده على النوع هكذا: [مفرد]، ثالثها: الاسم المثنى، وينصّ بعده على النوع هكذا: [مثنى]، رابعها: الاسم الجمع، وينصّ بعده على النوع هكذا: [جمع]، وخامسها: الكلمات الوظيفية..." (عمر، أحمد مختار، ٢٠٠٨: ص ٩)



المراجع العربية

- أبو العزم، عبد الغني. (2013). *معجم الغني الزاهر*. مؤسسة الغني للنشر.
- الثعالبي، أبو منصور. (٢٠٠٩). *فقه اللغة وسر العربية*. (يحي مراد، محقق). مؤسسة المختار.
- صيني، محمود إسماعيل، وعبد العزيز، ناصف مصطفى، وسليمان، مصطفى أحمد. (١٩٩٣). *المكنز العربي المعاصر*. مكتبة لبنان ناشرون.
- عمر، أحمد مختار، وآخرون. (٢٠٠٠). *المكنز الكبير معجم شامل للمجالات والمتراقات والمتضادات*. سطور.
- عمر، أحمد مختار. (٢٠٠٣). نظرة في معجمين حديثين للمتراقات: المكنز العربي المعاصر والمكنز الكبير. *مجلة مجمع اللغة العربية*. ٧٨ (٤) صص ٨٦٧-٨٩٤.
- عمر، أحمد مختار، وآخرون. (٢٠٠٨). *معجم اللغة العربية المعاصرة*. عالم الكتب.
- مجمع اللغة العربية بالقاهرة (١٩٨٥)، *المُعْجَم الوسيط*. مجمع اللغة العربية.
- مصلوح، سعد. (٢٠٠٠). *تعليق الجار والمجرور في العربية من منظور نحوي معجمي*. ضمن: حلي هليل (٢٠٠٠). *بحوث الندوة الدولية للمعاجم اللغوية والمختصة*. جامعة الكويت.
- موسى، حسين يوسف والصعيدى، عبد الفتاح. (١٩٦٤). *الإفصاح في فقه اللغة*. دار الفكر العربي.
- هواري، عبد العاطي إبراهيم. (٢٠٠٨). *لغة التعريف في المعجم العربي العام الحديث إشكالية الصياغة والمحتوى، رؤية لغوية حاسوبية*. (رسالة دكتوراه غير منشورة). جامعة القاهرة.
- الجيزة، مصر
- اليازجي، إبراهيم. (١٩٨٥). *نجعة الرائد وشرعة الوارد في المتراقات والمتوارد*. مكتبة لبنان.

المراجع الأجنبية

- Francopoulo, Gil (Editor). (2013). *LMF Lexical Markup Framework*. Wiley-Iste.
- Khemakhem, A., Elleuch, I., Gargouri, B. & BenHamadou, A. (2009 April 22–23). *Towards an automatic conversion approach of editorial Arabic dictionaries into LMF-ISO 24613 standardized model*. In The 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt.
- Kilgariff, A., Yallop, C. (2000 May). *What's in a thesaurus?* In Proceedings of the 2nd conference on language resources and evaluation (LREC'00), Athens, Greece.

نبذة تعريفية عن الباحث أو الباحثين

بيانات الباحث	Author Biodata
عبد العاطي إبراهيم هوارى، باحث في مجال اللسانيات الحاسوبية، حاصل على درجة الدكتوراه في اللسانيات الحاسوبية من جامعة القاهرة عام ٢٠٠٨، تدور اهتماماته البحثية حول حوسبة المعجم العربي، والمعجمية الدلالية، والصرف، والأنماط النحوية التركيبية، وبناء المدونات النصية المحوسبة، ولغة الأدب.	Author Biodata Abdelati Hawwari is a researcher in the field of computational linguistics. He holds a PhD in computational linguistics from Cairo University, awarded in 2008. His research interests focus on the computational processing of the Arabic lexicon, semantic lexicography, morphology, syntactic patterns, the construction of computerized text corpora, and the language of literature.

معرف أوركيد (ORCID) : 0000-0002-9135-8205

Email: abhawwari@gmail.com



ملحق أ:



المعايير العالمية المستخدمة في صناعة المعاجم الرقمية: تحليل مقارنة بين LMF و TEI*

The Global Standards used in Digital Lexicography:

A Comparative Analysis between LMF and TEI

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

عمر الجندي

مجمع اللغة العربية، القاهرة، مصر

توثيق البحث APA Citation:

الجندي، عمرو. (2024). المعايير العالمية المستخدمة في صناعة المعاجم الرقمية، تحليل مقارنة بين LMF و TEI. مجلة اللغويات الحاسوبية ومعالجة الآلية للغة العربية، (٢).

Submission Date: 10/3/2024, Acceptance Date: 24/9/2024

تاريخ الإرسال: ٢٠٢٤/٣/١٠ / تاريخ القبول: ٢٠٢٤/٩/٢٤ م

Abstract

Despite the fact that the Lexical Markup Framework (LMF) is the most widely recognized global standard, and it is the standard announced for building the Riyadh dictionary, there is another standard for digital lexicons within the Text Encoding Initiative (TEI). This research paper aims to introduce both standards, explore the key differences between them, and discuss the main challenges the researcher believes may arise when applying them to Arabic lexicons. The paper also seeks to establish the criteria that may assist in selecting the appropriate standard for developing an Arabic digital lexicon.

Keywords: Digital Lexicons, LMF, TEI, Natural Language Processing, Lexical Modeling

الملخص

رغم أن إطار الترميز المعجمي Lexical Markup Framework (LMF) هو المعيار العالمي الأشهر، وهو نفسه المعيار الذي أعلن عن صناعة معجم الرياض وفقه؛ فإن ثمة معياراً آخر للمعاجم الرقمية ضمن مبادرة ترميز النصوص Text Encoding Initiative (TEI). وهذه الورقة البحثية هي محاولة تستهدف التعريف بهذين المعيارين، واستعراض الفروق الأساسية بينهما، ومناقشة أبرز التحديات التي يرى الباحث أنها تواجه تطبيقهما على المعاجم العربية، وتحديد الأسس التي يعتقد الباحث أنها قد تساعد في اختيار المعيار المناسب عند صناعة معجم رقمي عربي.

الكلمات المفتاحية: المعاجم الرقمية، LMF، TEI، معالجة اللغة الطبيعية، النمذجة المعجمية.

١. مقدمة

يشكل الانتقال نحو الأساليب الرقمية الخطوة الأحدث في تاريخ صناعة المعاجم؛ حيث لم تعد التكنولوجيا تستخدم فقط لإنتاج المعاجم المطبوعة، بل غدت هناك أعداد متزايدة من المعاجم الرقمية دون أن يكون لها نسخ مطبوعة. وأصبحت تكنولوجيا المعلومات والاتصالات منتشرة في كل مراحل صناعة المعاجم الحديثة؛ حيث يعتمد كل من اكتساب البيانات وتمثيل المعرفة اللغوية على نحو كبير على هذه التكنولوجيا. بالإضافة إلى ذلك، فقد تحول صانعو المعاجم من الأساليب التقليدية إلى أساليب تعتمد على موارد رقمية متقدمة مثل المدونات اللغوية (Budin, Majewski, & Mörrth, 2024). وربما يعود ذلك إلى أسباب متعددة، منها: سهولة الاستخدام، وإمكانية التحديث المستمر، وانخفاض التكلفة، فضلاً عن الخصائص المتقدمة التي تتيحها المعاجم الرقمية لمستخدميها، وما يبنى على ذلك كله من إسهامات كبيرة في تطوير الأدوات والتطبيقات المختلفة، بالإضافة إلى الحفاظ على اللغة، وإثراء بياناتها، وتعلمها، ونشر المعرفة بوساطتها.

وقد أدت رقمنة المعاجم المطبوعة إلى إبراز ما بها من فجوات طالما تحدث عنها الباحثون. وإلى كشف النقاب عن ضرورة وجود معيار مشترك ينظم رقميتها، ويساعد على ضمان جودتها واستدامة تحديثها وتطويرها، وتهيئتها للمعالجة الآلية، وإتاحة وصول المستخدمين إليها، وسهولة استخدامهم إياها. بالإضافة إلى توافق هذه المعاجم بعضها مع بعض وتكاملها وتيسير تبادلها بين المنصات المختلفة وربما دمجها، والاستفادة منها على نطاق واسع؛ خاصة وأن تحسين إنتاج المعاجم الرقمية وصيانتها وتوسيعها من الجوانب الحاسمة التي تؤثر على تقنيات معالجة اللغة الطبيعية (ISO, 2019). والمقصود بوضع معيار مشترك للمعاجم الرقمية، هو «وضع مواصفات ينبغي تحقيقها في النص المعجمي على مستوى بنيته الكبرى والصغرى، شكلاً ومحتوى، وذلك



لتنميط لغة التمثيل المعجمي؛ بحيث يسهل التعامل معها حاسوبياً عند إرادة معالجة اللغة آلياً» (هوارى وآخرون، ٢٠١٩: ٣٥).

وتهدف هذه الورقة البحثية إلى تحديد المعيار الأنسب لاستخدامه في صناعة المعاجم الرقمية العربية، من خلال إجراء تحليل غير معمق لأوجه الاتفاق والاختلاف بين إطار الترميز المعجمي LMF ومبادرة ترميز النصوص TEI. كما تتناول التحديات التي تواجه تطبيق كل من المعيارين على المعاجم الرقمية العربية، وتقدم توصيات مبنية على تقييم -حاول توخي الدقة- لكيفية توافق كل معيار مع احتياجات المعاجم الرقمية في السياق العربي.

٢. الدراسات السابقة

على الرغم من وجود الكثير من الدراسات التي تناولت كلاً من إطار الترميز المعجمي LMF ومبادرة ترميز النصوص TEI، فإن هذه الورقة البحثية - في حدود علم الباحث - تتميز بكونها الأولى التي تقارن بين المعيارين، وتستعرض التحديات التي تواجه تطبيقهما على المعاجم الرقمية العربية، وتقدم التوصيات حول كيفية اختيار المعيار الأنسب لصناعة معجم رقمي عربي.

ويعد بحث (Salmon-Alt, Akrou, & Romary, 2005) مرجعاً أساسياً في تطوير الموارد المعجمية للغة العربية باستخدام معيار LMF؛ حيث يبرز أهمية تكييفه مع الخصائص اللغوية الفريدة للعربية.

ومهما يكن من أمر، فإن تقييم الدراسات السابقة في هذا الموضوع يتطلب دراسة تفصيلية لكل حالة على حدة، وهو ما يمكن أن يكون موضوعاً لمزيد من البحث في المستقبل.

٣. معايير صناعة المعاجم الرقمية

إن المعاجم الورقية لا تخلو من معايير ترتب عناصرها ترتيباً يعين مستخدمها على الاستفادة منها؛ لكن هذه المعايير تتنوع وتتعدد، وربما تتضارب بحسب الأهداف المرجو الاستفادة منها على نحو أو آخر؛ لذلك تصبح عملية ترميز البيانات والمعلومات في تلك المعاجم عند رقمتها أمراً بالغة الأهمية (TEI Consortium, 2024). ولذلك كان هناك عدد من المعايير التي هدفت إلى ترميز البيانات والمعلومات المعجمية، منها:

- GENELEX: هو مشروع أوروبي، هدف إلى وضع نموذج عام للمعاجم الرقمية، وتصميم أدوات برمجية وتطويرها وإدارتها، وتطبيق هذا النموذج وتلك الأدوات، وبناء معاجم رقمية كاملة النطاق (Antoni-Lay, Francopoulo, & Zaysser, 1994).
- EDR (Electronic Dictionary Research and Development Association): وهو مشروع ياباني يهدف إلى تطوير معجم إلكتروني شامل يمكن استخدامه في تطبيقات معالجة اللغة الطبيعية.

٣،١ إطار الترميز المعجمي (LMF) Lexical Markup Framework

ويعد إطار الترميز المعجمي LMF هو المعيار الأشهر لصناعة المعاجم الرقمية، وهو نموذج مرجعي تجريدي يوفر إطار عمل مشتركاً وموحداً لبناء الموارد المعجمية بما في ذلك المعاجم الرقمية؛ حيث يضمن ترميز المعلومات اللغوية بطريقة تتيح تخزينها وإعادة استخدامها في تطبيقات مختلفة ولأغراض مختلفة، بما في ذلك تبادلها و/أو دمجها بسهولة بين/مع الأنظمة والقواميس والمعاجم الإلكترونية لتشكيل موارد إلكترونية عالمية واسعة النطاق، كما يوفر تمثيلاً مشتركاً وموحداً للحالات المعجمية، بما في ذلك الجوانب الصرفية والتركيبية والدلالية. (Lexical Markup Framework, (ISO, 2019)، (هوارى وآخرون، ٢٠١٩).

كانت بداية هذا الإطار في عام ٢٠٠٣م، حين قدم الوفد الأمريكي اقتراحاً بشأنه إلى المنظمة العالمية للمعايير ISO، تبعه -في العام نفسه- اقتراح من الوفد الفرنسي لنموذج



بيانات مخصص لمعاجم معالجة اللغة الطبيعية. وفي العام التالي قررت اللجنة الفنية 37 (TC37) تكليف لجنتها الفرعية ٤ (SC4) بمشروع مشترك لتطوير LMF بمشاركة ٦٠ محرراً (Lexical Markup Framework, 2024).

كانت الخطوة الأولى في تطوير LMF هي تصميم إطار عام بناءً على الميزات العامة للمعاجم الحالية، وتطوير مصطلحات متسقة لوصف مكونات هذه المعاجم. أعقب ذلك التصميم الفعلي لنموذج شامل يمثل جميع المعاجم بالتفصيل (Lexical Markup Framework, 2024).

وقد ساهم فريق كبير من الخبراء والمحررين في تقديم مجموعة واسعة من المتطلبات والحلول؛ للتوصل إلى توافق في الآراء بشأن تصميم LMF بما يغطي أنواعاً كثيرة من المعاجم، كما أولى علم الصرف اهتماماً خاصاً بغية توفير آليات قوية لمعالجة التحديات المتعلقة به في عدة لغات (Lexical Markup Framework, 2024).

وعلى مدار خمس سنوات من العمل والمناقشات تُؤَصِّل إلى نموذج UML متماسك،
وصف بأنه تجسيد لأحدث ما توصل إليه العلم في مجال معاجم معالجة اللغة الطبيعية (Lexical Markup Framework, 2024).

حتى الآن نُشرت ٦ معايير خاصة بـ LMF؛ هذه المعايير هي: النموذج الأساسي، ونموذج المعجم المقروء آلياً، والامتداد الاشتقاقي، وتسلسل TEI، وتسلسل تبادل القاعدة المعجمية، والتركيب والدلالة (ISO, 2019 : 2024).

٣,٢. مبادرة ترميز النصوص (Text Encoding Initiative (TEI

أما مبادرة ترميز النصوص TEI فهي مجموعة دولية تطور معياراً لتمثيل النصوص الإلكترونية بكافة أنواعها؛ حيث تقدم مجموعة من الإرشادات التي تحدد طرائق ترميز النصوص القابلة للقراءة آلياً؛ وخاصة في مجالات العلوم الإنسانية والعلوم الاجتماعية واللغويات. وتستخدم هذه الإرشادات على نطاق واسع منذ عام ١٩٩٤م، من المكتبات

والمتاحف ودور النشر والعلماء؛ لعرض النصوص لأغراض البحث والتعليم وغيرها. وتعد المعاجم جزءاً من أهداف تلك المبادرة (TEI Consortium, 2024).

٤. أوجه الاتفاق بين LMF و TEI

يرى الباحث أن ثمة أوجهاً للاتفاق بين المعيارين؛ لعل أبرزها: التعامل مع المعاجم المختلفة، وتغطية المعلومات المعجمية، والمرونة في تمثيل تلك المعلومات، ودعم الامتداد والتخصيص، والمراجع المتقاطعة، وتكييف اللغات المختلفة، والاستدامة والتطوير المستقبلي.

- التعامل مع المعاجم المختلفة:

يتناول كلا المعيارين جميع أنواع المعاجم المطبوعة والرقمية وما شابههما من مصادر تستخدم بوساطة برمجيات معالجة اللغة من جميع اللغات، بما في ذلك أحادية اللغة وثنائية اللغة ومتعددة اللغات، والصغيرة والكبيرة، والبسيطة والمعقدة (TEI Consortium, 2024).

- تغطية المعلومات المعجمية:

يغطي كلا المعيارين كل التفاصيل المعجمية؛ حيث يوفران حزمات ومكونات أساسية تضم فئات لتمثيل مختلف المعلومات المعجمية؛ بما في ذلك الشكل المكتوب والمنطوق، والمعلومات الصرفية، والنحوية، والدلالية، وأصول الكلمات، والأمثلة والسياقات، ومعلومات الاستخدام (TEI Consortium, 2024).

- المرونة في تمثيل المعلومات:

يتميز كلا المعيارين بمرونة كبيرة في تمثيل المعلومات المعجمية؛ فعلى سبيل المثال: يوفر LMF آليات لتوريث الفئات Class Inheritance، واختيار فئات البيانات المناسبة، أو تعريف فئات بيانات مخصصة لتلبية احتياجات التطبيقات المعجمية الخاصة. كما يتميز TEI بقدرته على التعامل مع أنواع متنوعة من المعلومات المعجمية بمرونة عالية، مع إمكانية ترميز البيانات النصية المعقدة.



شكل (١)

نموذج لتمثيل بعض البيانات المعجمية بوساطة LMF

```
<LexicalResource dtdVersion="15">
  <GlobalInformation>
    <feat att="languageCoding" val="ISO 639-3"/>
  </GlobalInformation>
  <Lexicon>
    <feat att="language" val="eng"/>
    <LexicalEntry>
      <feat att="partOfSpeech" val="commonNoun"/>
      <Lemma>
        <feat att="writtenForm" val="clergyman"/>
      </Lemma>
      <WordForm>
        <feat att="writtenForm" val="clergyman"/>
        <feat att="grammaticalNumber" val="singular"/>
      </WordForm>
      <WordForm>
        <feat att="writtenForm" val="clergymen"/>
        <feat att="grammaticalNumber" val="plural"/>
      </WordForm>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

هذا المثال بسيط نسبياً، بينما يمكن ل LMF تمثيل أوصاف لغوية أكثر تعقيداً، وبالتالي، يصبح وضع العلامات في XML أكثر تعقيداً بما يتناسب مع ذلك.

شكل (٢)

تمثيل بوساطة TEI مدخل يحتوي على معنيين

```
<entry>
  <sense n="1"/>
  <sense n="2"/>
</entry>
```

(TEI Consortium, 2024)

شكل (٣)

تمثيل بوساطة TEI لمدخل يحتوي على صيغتين؛ الأولى مزدوجة المعنى، والثانية متعددة المعاني

```
<entry>
  <hom n="1">
    <sense n="1">
<!-- ... -->
    </sense>
    <sense n="2">
<!-- ... -->
    </sense>
  </hom>
  <hom n="2">
    <sense n="1">
      <sense n="a">
<!-- ... -->
      </sense>
      <sense n="b">
<!-- ... -->
      </sense>
    </sense>
    <sense n="2">
<!-- ... -->
    </sense>
    <sense n="3">
<!-- ... -->
    </sense>
  </hom>
</entry>
```

(TEI Consortium, 2024)

- دعم التخصيص والامتداد:

يدعم كلا المعيارين التخصيص والامتداد؛ حيث يتيحان استخدام الامتدادات لإضافة وظائف جديدة أو تعديل النموذج الأساسي لتلبية متطلبات محددة. فعلى سبيل المثال يوفر LMF إمكانية تخصيص النموذج من خلال امتدادات مثل: الامتداد الخاص بعلم أصول الكلمات، بينما يوفر TEI إمكانيات واسعة للتخصيص من خلال مجموعة متنوعة من العناصر والسمات، مما يسمح بتوسيع استخدامه لدعم أنواع جديدة من المعلومات أو تكييفه لتمثيل المداخل غير النمطية.



- الإحالة والمراجع المتقاطعة:

يأخذ كلا المعيارين مسألة الإحالة والمراجع المتقاطعة بعين الاعتبار؛ حيث يوفران عناصر مثل CrossREF و <xr> للإشارة إلى معلومات معجمية مرتبطة عبر المداخل المختلفة، أو مداخل أخرى أو أجزاء من النص؛ مما يتيح مرونة أكبر في دمج المعلومات وترابطها عبر الموارد المختلفة، ومن ثم تعزيز التكامل والاتساق في تمثيل المعلومات المعجمية.

- تكييف اللغات المختلفة:

يوفر كلا المعيارين تكييفاً عالياً للغات المختلفة؛ فعلى سبيل المثال يتميز LMF بمرونته في التعامل مع لغات متعددة وأنظمة كتابة مختلفة بفضل استخدامه لمعايير مثل Unicode وترميزات اللغة، مما يجعله معياراً عالمياً لتطوير المعاجم الرقمية. وفي الوقت نفسه يدعم TEI ترميز النصوص من أي لغة بفضل مرونته، وهو مصمم للتعامل مع التحديات المتعددة للغات في السياقات الأكاديمية والبحثية.

- الاستدامة والتطوير المستقبلي:

يتميز كلا المعيارين باستدامة وتكيف مع التطورات؛ فـ LMF يحافظ على استدامته في معالجة اللغات الطبيعية من خلال التطوير المستمر من المنظمة الدولية للمعايير (ISO) والدعم الذي يحظى به من مجتمع اللسانيات الحاسوبية. كما يحظى TEI بدعم من مجتمع متعدد التخصصات وتطور مستمر من TEI Consortium، مما يضمن توافقه مع متطلبات البحث والنشر الإلكتروني.

٥. أوجه الاختلاف بين LMF و TEI

بعد استعراض أوجه الاتفاق بين المعيارين LMF و TEI، نتناول في هذا القسم أبرز الأوجه التي يختلف فيها كل منهما، والتي يعتقد الباحث أنها كالاتي: الإطار العام، والمعايير الأساسية، والتصميم الوحداتي، ونماذج الموارد المعجمية، والتوافق مع الأدوات الأخرى.

- الإطار العام:

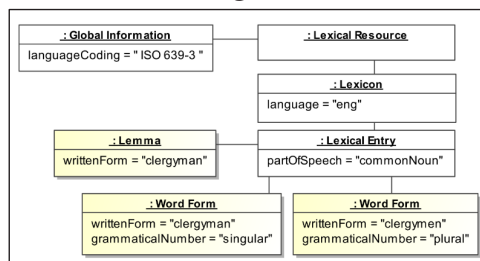
يوفر LMF نموذجاً معيارياً موحداً لإنشاء وإدارة الموارد المعجمية بمختلف أنواعها؛ بما يسهل تخزينها وإعادة استخدامها وتبادلها بين الأنظمة والتطبيقات المختلفة، وبخاصة تطبيقات معالجة اللغات الطبيعية NLP. (ISO, 2019) أما TEI فإنه يوفر نموذجاً معيارياً لتمثيل النصوص الأدبية والتاريخية واللغوية والمعجمية؛ وبخاصة المطبوعة منها؛ بما يسهل تحليل النصوص وفهمها عبر مجموعة واسعة من التطبيقات.

- المعايير الأساسية:

بالرغم من استخدام كلا المعيارين لغة التوصيف القابلة للتوسع XML؛ فإنهما يختلفان في المعايير الأساسية المستخدمة في ترميز المعاجم الرقمية؛ حيث يعتمد LMF على معايير صارمة مثل يونيكود Unicode؛ لضمان التوافق اللغوي، وترميز اللغة Language coding وترميز الكتابة Script coding؛ لضمان تمثيل دقيق ومتكامل للغات المختلفة. كما يستخدم لغة النمذجة الموحدة UML لتصميم النموذج المرجعي للترميز المعجمي وتوثيقه على نحو هيكلي ومنهجي؛ مما يوفر تصوراً واضحاً للعلاقات بين العناصر المعجمية المختلفة. أما TEI فإنه لا يتبع هذا النهج الصارم نفسه.

شكل (٤)

نموذج UML



شكل (٥)

تمثيل LMF لبعض المعلومات المعجمية بوساطة XML

```
<LexicalResource>
  <Lexicon language="ar">
    <LexicalEntry id="book_noun">
      <Lemma writtenForm="كتاب" />
      <WordForm writtenForm="كتاب" partOfSpeech="noun"/>
      <Sense id="sense1">
        <Definition>مصدر للقراءة والتعليم</Definition>
        <Example>أحب قراءة الكتاب.</Example>
      </Sense>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

- تصميم وحداتي Modular Design:

يُبنى LMF على تصميم وحداتي، حيث يُمثّل كل جانب من جوانب المعلومات المعجمية بوحدة مستقلة، وهذا يتيح مرونة كبيرة في إضافة أو تعديل أو استبدال الوحدات دون التأثير على النظام ككل، مما يجعله قابلاً للتوسع والتكيف مع مختلف الاحتياجات المعجمية. أما TEI فيعتمد على بنية هرمية تسمح بتجميع العناصر المتعددة تحت نظام، ولكنها تتطلب تكاملاً شاملاً للعناصر المختلفة بدلاً من العمل بوصفها وحدات مستقلة.

- نماذج الموارد المعجمية:

على عكس TEI فإن LMF يدعم نماذج الموارد المعجمية عبر توفير بنية معيارية تتيح تمثيل هذه النماذج المختلفة بطريقة متسقة وموحدة؛ مما يسهل تبادل المعلومات المعجمية ودمجها بين الأنظمة المختلفة. من أشهر هذه النماذج: Genelex، ومعيار EAGLES الدولي لهندسة اللغة (ISLE)، ونماذج المدخل المعجمي متعدد اللغات (MILE)،

وإرشادات مبادرة ترميز النصوص (TEI)، وOntolex، وتسلسل تبادل قاعدة اللغة (LBX)، إلى جانب نظام القاموس عبر الإنترنت للحكومة الأمريكية Wordscape، وغيرها (ISO, 2019).

- التوافق مع الأدوات الأخرى:

يتميز LMF بتوافق واسع مع أدوات معالجة اللغة الطبيعية NLP والتطبيقات المستخدمة فيها؛ أما TEI فإنه يتمتع بتوافق جيد مع مجموعة واسعة من أدوات معالجة النصوص والنشر الإلكتروني؛ مما يجعله مناسباً لمجالات العلوم الإنسانية واللغويات والنشر الرقمي، وأقل تخصصاً في التعامل مع الموارد المعجمية.

٦. أبرز التحديات التي تواجه تطبيق LMF وTEI على المعاجم العربية

يرى الباحث أن ثمة تحديات تواجه تطبيق المعايير العالمية LMF وTEI على المعاجم العربية، وأن أبرز هذه التحديات ما يتعلق بتنوع اللغة العربية، وتعقيداتها المعجمية، ونقص الموارد والخبرات الفنية المتخصصة.

- التنوع اللغوي واللهجات:

حيث يتطلب تطبيق تلك المعايير تكييفاً مع التنوع اللغوي الكبير، واللهجات المختلفة التي تنقسم به اللغة العربية، وهذا ربما يتطلب تخصيصات إضافية؛ لضمان التمثيل الدقيق والشامل للمعلومات المعجمية.



شكل (٦)

تمثيل TEI لبعض المعلومات المعجمية بواسطة XML

```
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>معجم عربي</title>
      </titleStmt>
      <publicationStmt>
        <p>Published by Example Publisher</p>
      </publicationStmt>
      <sourceDesc>
        <p>Electronic Dictionary</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <entry>
        <form>
          <orth>كتب</orth>

          </form>
          <gramGrp>
            <pos>noun</pos>
          </gramGrp>
          <sense>
            <def>مصدر للقراءة والتعليم</def>
            <cit>
              <quote>أحب قراءة الكتب</quote>
            </cit>
          </sense>
        </entry>
      </body>
    </text>
  </TEI>
```

- التعقيدات المعجمية:

تتطلب المعاجم العربية نماذج دقيقة وتخصيصات إضافية لتمثيل تعقيداتها الصرفية والنحوية والدلالية والطباعية والبنائية، وكل هذا يشكل تحدياً في بناء قواعد بيانات معجمية دقيقة وفعالة.

- نقص الموارد والخبرات الفنية:

مما لا شك فيه أن هناك نقصاً كبيراً في الموارد التقنية والخبرات المتخصصة لتطبيق المعايير العالمية على المعاجم العربية، وهذا يستلزم - بطبيعة الحال - تدريباً مخصصاً لفهم العناصر والسمات والتقنيات ذات الصلة التي توفرها هذه المعايير، ولضمان استخدامها على نحو يحقق التمثيل الدقيق والشامل للمعلومات المعجمية.

٧. أسس اختيار المعيار المناسب عند صناعة معجم رقمي عربي

في حدود علم الباحث، فإن معجم الرياض هو المعجم الرقمي العربي الوحيد الذي ذكر صراحة أنه بني وفقاً لمعيار LMF؛ (King Salman Global Academy for Arabic Language, 2024) في حين لم تنص معاجم أخرى على استخدامها معيار TEI.

وأياً يكن الأمر، فإن الباحث يعتقد أن ثمة عوامل ينبغي مراعاتها عند اختيار المعيار المناسب لصناعة معجم رقمي عربي، وهذه العوامل تتشابه إلى حد بعيد مع عوامل صناعة المعجم التقليدي. تشمل هذه العوامل: أهداف المعجم، ونوع المعجم، والتوافق مع الأدوات الحالية، والمرونة في التخصيص.



- أهداف المعجم الرقمي:

فإذا كان الهدف الرئيس أو حتى أحد الأهداف الأساسية هو معالجة اللغات الطبيعية وتطبيقات الذكاء الاصطناعي فإنه يفضل LMF. أما إذا كان الهدف الرئيس يتضمن ترميز النصوص الأدبية والتاريخية فإنه يفضل TEI.

- نوع المعجم:

يعد LMF أكثر مناسبة للمعاجم التي تتطلب تمثيلاً دقيقاً للمستويات اللغوية؛ في حين أن TEI مناسب أكثر لهذا النوع من المعاجم الذي يكون ضمن مواد أدبية أو ثقافية.

- التوافق مع الأدوات:

يتوافق LMF على نحو كبير مع أدوات NLP والأنظمة المعتمدة على المعايير العالمية؛ في حين يتوافق TEI مع أدوات معالجة النصوص والنشر الإلكتروني.

- المرونة في التخصيص:

يدعم LMF الامتدادات لتلبية متطلبات محددة؛ في حين يتمتع TEI بمرونة في ترميز النصوص المتنوعة ودعم المداخل غير النمطية.

٨. خاتمة

في ختام هذه الورقة يتضح أن كلاً من إطار الترميز المعجمي LMF ومبادرة ترميز النصوص TEI لهما دور مهم في تطوير المعاجم الرقمية، ويعتمد اختيار المعيار المناسب على أهداف المشروع ومتطلباته؛ حيث يبرز LMF بوصفه بنية معيارية توفر مرونة عالية ودعمًا لتطبيقات معالجة اللغة الطبيعية؛ بحيث يُنصح باستخدامه في المشاريع التي تتطلب تمثيلاً لغوياً دقيقاً ومتعدد المستويات. أما TEI فهو خيار مناسب لترميز المعلومات المعجمية في المشاريع ذات النصوص الأدبية والتاريخية المعقدة.

كما أظهر التحليل أن المعيارين يشتركان في بعض الخصائص، كالمرونة والتكيف مع لغات متعددة، ويختلفان في بعضها الآخر، كطرق تمثيل المعلومات وإدارتها.

وأخيراً، يوصي الباحث بإجراء أبحاث مستقبلية لتطبيق هذه المعايير في سياقات مختلفة ودراسة تأثيرها على جودة المعاجم الرقمية العربية، وبأن تحذو المجمع اللغوية حذو مجمع الملك سلمان العالمي للغة العربية في تبني أحد المعايير الدولية في صناعة المعجم الرقمي، وبأن يكون لهذه المعايير مكان ضمن المقررات الدراسية بأقسام اللغة العربية.

الهوامش:

(١) من هذه المعايير: EAGLES (Expert Advisory Group on Language Engineering Standards) مشروع أوروبي يهدف إلى وضع معايير موحدة في هندسة اللغة؛ لدعم تبادل البيانات والأدوات بين الباحثين. MULTEXT: مشروع يهدف إلى تطوير موارد لغوية وأدوات لدعم معالجة النصوص متعددة اللغات. PAROLE: جزء من مبادرة أكبر لتطوير موارد لغوية للغات الأوروبية، ويركز على إنشاء قواعد بيانات لغوية معيارية لتلك اللغات. SIMPLE: مشروع يهدف إلى إنشاء بنية معجمية موحدة للغات الأوروبية تعتمد على إطار عمل موحد لتحليل المعاني. ISLE (International Standards for Language Engineering) مشروع دولي يهدف إلى تطوير معايير هندسة اللغة لدعم تطوير وتبادل الموارد والأدوات اللغوية.



المراجع العربية

مجمع الملك سلمان العالمي للغة العربية. (٢٠٢٤). معجم الرياض للغة العربية المعاصرة.
<https://dictionary.ksaa.gov.sa/about>
هوارى، عبدالعاطي وآخرون. (٢٠١٩). الموارد المعجمية العربية الحاسوبية. في المعتز بالله السعيد،
ومحسن رشوان (محرران). *الموارد اللغوية الحاسوبية*. (ص ص. ١١-٥٠). مركز الملك عبد
الله الدولي لخدمة اللغة العربية.

المراجع الإنجليزية

Antoni-Lay, M.-H., Francopoulo, G., & Zaysser, L. (1994). A Generic Model for
Reusable Lexicons: The Genelex Project. *Literary and Linguistic Computing*,
9(1), 47-54.
Budin, G., Majewski, S., & Mörtz, K. (٢٠١٢). Creating Lexical Resources in TEI P5: A
Schema for Multi-purpose Digital Dictionaries. *Journal of the Text Encoding
Initiative [Online]*, 3. <https://doi.org/10.4000/jtei.522>
International Organization for Standardization. (2019). ISO 24613-1: Core Model.
Retrieved June 2024, from <https://www.iso.org/standard/82014.html>.
International Organization for Standardization. (2020). ISO 24613-2: Machine-
readable Dictionary (MRD) Model. Retrieved June 2024, from
<https://www.iso.org/standard/75407.html>.
International Organization for Standardization. (2021). ISO 24613-3: Etymological
Extension. Retrieved June 2024, from
<https://www.iso.org/standard/75410.html>.
International Organization for Standardization. (2021). ISO 24613-4: TEI Serialization.
Retrieved July 2024, from <https://www.iso.org/standard/75411.html>.

- International Organization for Standardization. (2022). ISO 24613-5: Lexical Base Exchange (LBX) Serialization. Retrieved July 2024, from <https://www.iso.org/standard/72099.html>.
- International Organization for Standardization. (2024). ISO 24613-6: Syntax and semantics. Retrieved July 2024, from <https://www.iso.org/standard/83180.html>.
- International Organization for Standardization. (٢٠٢٤). ISO TC 37: Terminology and other language and content resources. Retrieved June 2024, from <https://www.iso.org/committee/48104.html>
- Jarrar, M., & Hammouda, T. (٢٠٢٤ May 20-25). *Qabas: An Open-Source Arabic Lexicographic Database*. In the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia.
- Lexical Markup Framework. (٢٠٢٤). LMF — Lexical Markup Framework website. Retrieved June 2024, from <https://www.lexicalmarkupframework.org>
- Salmon-Alt, S., Akrou, A., & Romary, L. (2005, November). *Proposals for a Normalized Representation of Standard Arabic Full Form Lexica*. In International Conference on Machine Intelligence, Tozeur, Tunisia.
- TEI Consortium. (٢٠٢٤). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Retrieved July 2024, from <https://tei-c.org>
- TEI Consortium. (٢٠٢٤). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Retrieved July 2024, from <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>



نبذة تعريفية عن الباحث أو الباحثين

Author Biodata	بيانات الباحث
Amr El-Gendy is a Researcher in Academy of the Arabic Language in Cairo, Egypt. Dr. Amr received his PhD degree in 2024 from Ain Shams University. His research interests include Dictionaries, Artificial Intelligence and Arabic Language Computing.	عمرو الجندي، باحث بمجمع اللغة العربية بالقاهرة، مصر. حاصل على درجة الدكتوراه في اللسانيات الحاسوبية من جامعة عين شمس عام ٢٠٢٤. تدور اهتماماته البحثية حول المعاجم والذكاء الاصطناعي وحوسبة اللغة العربية.

معرف أوركيد (ORCID) ID : 0009-0008-8996-5904

Email: amr@arabicacademy.edu.eg

تعزير البيانات النصية من أجل التعامل مع المفردات غير المرصودة في النماذج اللغوية في مهام معالجة اللغة العربية المعاصرة^١

Textual Data Augmentation for Handling Out of Vocabulary (OOV) Words in Arabic Language Processing Tasks

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

هدى زيتون^أ سامح الأنصاري^ب

^أمعيد الدراسات اللغوية والترجمة، كلية الآداب، جامعة الإسكندرية بالتعاون مع الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحري، الإسكندرية، مصر.

^بقسم الصوتيات والمسانيات، كلية الآداب جامعة الإسكندرية، الإسكندرية، مصر.

توثيق البحث: APA Citation

زيتون، هدى، والأنصاري، سامح. (2024). تعزير البيانات النصية من أجل التعامل مع المفردات غير المرصودة في النماذج اللغوية في مهام معالجة اللغة العربية المعاصرة. مجلة اللغويات الحاسوبية ومعالجة الآلية للغة العربية، (٢).

Submission Date: 10/3/2024, Acceptance Date: 25/9/2024

تاريخ الإرسال: ٢٠٢٤/٣/١٠ م / تاريخ القبول: ٢٠٢٤/٩/٢٥ م

Abstract

The present study deals with out-of-vocabulary (OOV) words in Arabic Natural Language Processing and proposes solutions in light of applying some techniques of Textual Data Augmentation (TDA). TDA refers to small changes made to accessible text at the character, word, or sentence level to generate new, synthetic data that is subsequently inserted into data loaders to train the model. This application contributes to enriching and diversifying linguistic vocabularies in tasks such as machine translation, named entity recognition, and automatic text summarization. By applying the technique of TDA to handle out-of-vocabulary words, specifically in the task of machine translation, this study concludes that applying these techniques at the word level has helped provide alternative options for words, leading to the improvement of the capability and efficiency of models. Additionally, the performance of these models may vary depending on the techniques used.

Keywords: Augmentation, machine translation, out-of-vocabulary words

الملخص

تهدف هذه الدراسة إلى التعامل مع الكلمات غير المرصودة في مجال المعالجة الآلية للغة العربية، والحلول المقترحة في ضوء تطبيق التقنيات المختلفة لتعزير البيانات النصية. ويشير مصطلح تعزير البيانات النصية إلى التقنيات التي يمكن تنفيذها من أجل إجراء بعض التغييرات على العينة اللغوية المستخدمة في بناء نموذج لغوي معين أو نظام لغوي محدد، من أجل توليد ألفاظ وتراكيب لغوية جديدة تدرج مرة أخرى في أدوات تحليل البيانات لتدريب النموذج اللغوي. ويسهم هذا التطبيق في العمل على إثراء وتنوع المفردات اللغوية في مهام مثل الترجمة الآلية، والتعرف على الكيانات المسماة، وتلخيص النص آلياً. وبتطبيق تقنية تعزير البيانات النصية للتعامل مع الكلمات غير المرصودة وبالتحديد في مهمة الترجمة الآلية، خلصت هذه الدراسة إلى أن تطبيق هذه التقنيات على مستوى المفردات ساهم في توفير خيارات بديلة للكلمات أدت إلى تحسين قدرة وكفاءة النماذج اللغوية. أيضاً قد يختلف أداء هذه النماذج بناء على التقنيات المستخدمة.

الكلمات المفتاحية: تعزير، المفردات، غير المرصودة، الترجمة الآلية



١. المقدمة

لقد ساهمت التطورات الأخيرة في مجال الذكاء الاصطناعي في تحقيق الكثير من الإنجازات الملحوظة التي كانت قديماً تشكل تحدياً كبيراً للبشر. وأحد أهم هذه التحديات هو المعالجة الآلية للغات الطبيعية. تعد اللغة العربية إحدى أكثر لغات العالم انتشاراً، ورغم التطورات المتسارعة التي شهدتها اللغة العربية في مجال المعالجة الآلية وحوسبة اللغة، إلا أنها لا زالت تعد من اللغات ذات المصادر اللغوية المحدودة. فقد نجد أن هناك العديد من المفردات العربية غير المرصودة أو الموجودة عند تدريب الآلة، مما قد يعيق النموذج اللغوي عند إجراء عملية الاختبار، ومن ثم ينعكس ذلك سلباً على دقة المهام المطلوبة. فقد نجد العديد من مهام معالجة اللغة العربية التي تتطلب وجود مفردات لغوية متنوعة ومعاجم لغوية ثرية وذات فاعلية، مثل مهام الترجمة الآلية، والتعرف على الكيانات المسماة، وتحليل المشاعر، وتلخيص النص آلياً. لذا فإن التعامل مع مشكلة الكلمات غير المرصودة يعد أمراً مهماً لضمان دقة النماذج اللغوية والمهام المتبناة. ويشمل ذلك تحسين التغطية والدقة والأداء العام من خلال توفير مجموعة شاملة من الكلمات والخيارات البديلة للمفردات غير المرصودة.

يواجه الباحثون في مجال المعالجة الآلية للغة العربية الكثير من الكلمات غير المرصودة بسبب التعقيد الصرفي، واختلاف اللهجات، وثرأ المصطلحات التقنية التي تميز اللغة العربية عن غيرها من اللغات. ومن ثم جاء هدف هذه الدراسة التي تركز على تحليل متعمق للتحديات التي تواجهها المعالجة الآلية للغة العربية للتعامل مع الكلمات غير المرصودة، والحلول المقترحة للتعامل مع هذه التحديات في ضوء تطبيق التقنيات المختلفة لتعزيز البيانات النصية (Text Data Augmentation). ويشير مصطلح تعزيز البيانات النصية إلى التقنيات التي يمكن تنفيذها من أجل إجراء بعض التغييرات على العينة اللغوية المستخدمة في بناء نموذج لغوي معين أو نظام لغوي محدد (Feng et al., 2021)، والذي يمكن أن يحدث على مستوى الحرف أو الكلمة أو الجملة، من أجل توليد ألفاظ وتراكيب لغوية جديدة تدرج

مرة أخرى في أدوات تحليل البيانات لتدريب النموذج اللغوي. وتعتمد هذه الدراسة على اتباع آليات مختلفة لتطبيق عملي شامل لمناهج تعزيز البيانات النصية من خلال التعزيز الرمزي والتضميني على مستوى الكلمات والمفردات فقط، ومقارنة أداء النماذج مرة أخرى، وكذلك إبراز بعض أوجه التشابه والاختلاف المهمة بين هذه الآليات. ومن خلال تعزيز البيانات النصية الموجودة باستخدام هذه الآليات يمكننا توليد جمل جديدة ذات دلالات متشابهة بصيغ أو تراكيب مختلفة، مما يؤدي إلى تحسين أداء النماذج اللغوية ودقتها.

في القسم الثاني من هذه الدراسة سوف نقدم الإطار النظري، والذي يتضمن تعريفاً بتقنيات تعزيز البيانات النصية في مجال معالجة اللغة الطبيعية وتعلم الآلة، وسناقش هذا الإطار تعريف الترجمة الآلية وتحدياتها في ظل تعلم الآلة ومعالجة اللغة الطبيعية. وفي القسم الثالث من الدراسة سوف نناقش الإطار العملي، والذي استندت إليه الدراسة في تطبيق تعزيز البيانات النصية للتعامل مع الكلمات غير المرصودة في مهمة الترجمة الآلية. وأخيراً، سوف تقدم الدراسة ما أفضت إليه من نتائج في القسم الأخير أو الخاتمة.

٢. الإطار النظري

٢.١. تقنية تعزيز البيانات النصية

تعزيز البيانات النصية هو عملية مهمة في مجال معالجة اللغة الطبيعية وتعلم الآلة، حيث تهدف إلى زيادة مجموعات البيانات النصية المتاحة وتحسينها، وذلك لتحسين أداء النماذج اللغوية عند تدريبها. هناك العديد من التقنيات المستخدمة في تعزيز البيانات النصية، ومن بين هذه التقنيات الشائعة:

- استبدال الكلمات (Word Replacement) و يتضمن استبدال كلمات في النص بمترادفات لها أو كلمات مشابهة، وهذا يساهم في زيادة تنوع المفردات المستخدمة في تدريب النماذج (Fadaee et al., 2017).



- ترجمة الجمل (Back Translation) وهنا تترجم الجمل إلى لغات أخرى ثم تعاد للغة المصدر، مما يساهم في تنوع النصوص والتعرف على سياقات مختلفة (Sennrich et al., 2016).
 - إدراجات عشوائية (Random Insertions) وتتضمن هذه التقنية إدراج جمل أو كلمات عشوائية من مصادر خارجية إلى البيانات النصية، مما يجعل النموذج يتعلم من سياقات متنوعة (Wei & Zou, 2019).
 - تعويض الجمل بالترتيب (Sentence Shuffle) وهذه التقنية تعمل على تغيير ترتيب الجمل في النص، مما يساهم في تعزيز فهم النماذج للهياكل اللغوية المختلفة.
 - توليد النصوص (Text Generation) وتشمل تقنيات مثل إنشاء جمل جديدة باستخدام نماذج توليد اللغة الطبيعية.
 - استخدام تقنيات التضمين الكلمائية (Word Embeddings) (Wang & Yang, 2015) ونماذج اللغة (Language Models) وتعمل هاتان التقنيتان على توليد أو استبدال مرادفات للكلمات والعبارات (Wu et al., 2019).
- هذا ويمكن توظيف المزيد من التقنيات وفق احتياجات وأهداف المهمة المحددة في مجال معالجة اللغة الطبيعية.

٢.٢. الترجمة الآلية وتحدياتها في اللغة العربية في ظل تعلم الآلة ومعالجة اللغة

الطبيعية

يقصد بالترجمة الآلية تحويل نص من لغة إلى لغة أخرى باعتماد نظم آلية حاسوبية. وكعلم الحديث، يتقاطع علم الترجمة الآلية مع علوم عدة مثل: اللسانيات الحاسوبية، اللسانيات، الذكاء الاصطناعي (القريش، ٢٠٢٤). وتتجسد الترجمة الآلية في ترجمة نصوص أو جمل أو ألفاظ موزة بثلاث مراحل هي: التحليل، والتحويل، والتوليد. وكل مرحلة تنقسم إلى ثلاث مراحل أخرى جزئية لمعالجة الصرف والنحو والدلالة (القريش، ٢٠٢٤). لقد كانت أولى

محاولات تطوير برنامج ترجمة من الإنجليزية إلى العربية في نهاية السبعينيات، وقد احتوى النظام على مرحلتين: الأولى لتحليل اللغة، و الثانية لتوليد اللغة الهدف (خضر ، ٢٠١٨). و من أهم مشكلات الترجمة من الإنجليزية إلى العربية هي التوافق، وإعادة الترتيب بالصفات والأسماء، والتوافق بين الفعل وفاعله، والتوافق بين الضمائر والأسماء الموصولة، حيث إن ترتيب الكلمات في هذه المجالات الثلاثة مختلف بين العربية والإنجليزية (خضر ، ٢٠١٨).

وفي هذا السياق، فإن إحدى المشكلات والتحديات التي تواجه الترجمة الآلية، خاصة في سياق اللغة العربية المعاصرة، هو التعامل مع الكلمات غير المرصودة. ومن أهم أسباب وجود كلمات غير مرصودة من قبل الآلة عند القيام بعملية الترجمة الآلية من الإنجليزية إلى العربية هو وجود مصطلحات جديدة، والتي قد لا تحتويها نماذج معالجة اللغة الطبيعية المدربة على مجموعة بيانات ثابتة في مفرداتها. أيضاً، وجود بعض الكلمات التي قد تحتوي على أخطاء إملائية أو مطبعية قد لا تتطابق مع أي مصطلحات معروفة في المفردات. وأخيراً، تصادف بعض المصطلحات غير الشائعة أو ذات التخصص، والتي قد لا تكون نماذج معالجة اللغة الطبيعية المدربة على نصوص عامة تعرضت لها من قبل. لذلك فإن استخدام تقنيات تعزيز البيانات النصية على مهام الترجمة الآلية، خاصة في سياق اللغة العربية قد يسهم وبشكل ملحوظ في تحسين دقة الترجمة، وذلك بتحديد معنى الكلمة بناءً على الكلمات المحيطة بها، مما يزيد من دقة الترجمة. وتعمل هذه التقنيات أيضاً على توسيع قاعدة البيانات اللغوية، وذلك بالتعرف على الكلمة وتقديم صيغ مختلفة منها، مثل الأسماء والأفعال. أخيراً، يساعد التطبيق الأشمل لتطبيقات تعزيز البيانات النصية في الترجمة الآلية على إيجاد الحلول لعدد من التحديات اللغوية التي تواجه المعالجة الآلية للغة العربية المعاصرة، على سبيل المثال، بعض الكلمات في اللغة العربية تحمل معاني متعددة بحسب السياق. مثلاً، كلمة "عقد" يمكن أن تعني اتفاقية أو يمكن أن تشير إلى عملية الربط. هذا التنوع قد يجعل عملية الترجمة أكثر تعقيداً، وقد تُترجم بشكل خاطئ إذا لم يكن هناك سياق كافٍ.



٣. الإطار العملي

في سياق الدراسة الحالية ولتحقيق أهدافها، وهي تطبيق تعزيز البيانات النصية على النصوص المترجمة آلياً من الإنجليزية إلى العربية على مستوى الكلمات أو المفردات فقط، فقد اعتمدت الدراسة الحالية على تقنية استبدال الكلمات حيث تستبدل كلمات في النص بمرادفات لها أو كلمات مشابهة أو ذات سياق مشابه. وفي الإطار ذاته، فقد اعتمدت الدراسة في تطبيق هذه التقنية على استبدال الكلمات باستخدام قاعدة بيانات لإعادة الصياغة وردنت (WordNet) (Miller, 1995) والتي يمكن استخدامها لتوليد بيانات جديدة مولدة. وتقوم قاعدة البيانات وردنت لإعادة الصياغة من مكتبة NLTK بإحضار قائمة من المرادفات لكلمة معينة، في هذه الحالة الكلمات غير المرصودة، ثم تراجع الكلمات في الجملة الأصلية، وتتحقق مما إذا كان لديها مرادفات في القاموس المخصص، وتبديلها إذا كانت متوفرة. إضافة إلى ذلك، فقد اعتمدت الدراسة الحالية على استخدام تقنيات التضمين الكلمائية (Word Embeddings)، والتي تعمل على استبدال الكلمات وفقاً للشبه الدلالي، مستنداً إلى خاصية تضمين الكلمات بتبديل الكلمات بأعلى مصطلحات قابلة للمقارنة. ويمكن تلخيص الخطوات والأدوات المستخدمة في هذه الدراسة إلى النقاط الآتية:

- استخدمت الدراسة ذخيرة البي بي سي (BBC) (Greene & Cunningham, 2006)، والتي تحتوي على مجموعة بيانات عامة مستمدة من أخبار بي بي سي، والتي تتكون من ٢,٢٢٥ وثيقة، مصنفة إلى خمس مجموعات: الأعمال، الترفيه، السياسة، الرياضة، التكنولوجيا، وتم اختيار عدد من النصوص من المجموعات الخمس اختياريًا عشوائيًا لتطبيق أهداف الدراسة.

- ترجمت الدراسة النصوص المختارة باستخدام نموذج المعرفة للترجمة الآلية من الإنجليزية للعربية (Marefa-Mt-En-Ar)^١ نموذج المعرفة للترجمة الآلية من الإنجليزية للعربية

- هو أول نماذج الترجمة الآلية التي تصدر تحت رعاية موسوعة المعرفة، ويتميز هذا النموذج بدعمه لحروف الأبجدية العربية الإضافية لتمييز الصوتيات الخاصة في اللغة الإنجليزية.
- عملت الدراسة بعد ذلك على تقييم النصوص المختارة والمترجمة، حيث وظفت مقاييس روج (ROUGE) و بلو (BLEU) لتقييم جودة الترجمة من اللغة الإنجليزية إلى اللغة العربية. وقد استخدمت هذه المقاييس كأدوات لتحديد فعالية الترجمة وجودتها.
 - في الخطوة التالية لتقييم النصوص المترجمة، طبق الباحثون تقنية تعزيز البيانات النصية على النصوص المترجمة آلياً من الإنجليزية إلى العربية للتعامل مع الكلمات غير المرصودة في النصوص المترجمة والتي تم رصدها والكشف عنها بواسطة الباحثين، حيث استبدلت كلمات في النص بمرادفات لها أو كلمات مشابهة أو ذات سياق مشابه باستخدام قاعدة البيانات ورد نت المستدعاة من مكتبة NLTK وتقنية التضمين الكلمائية، وذلك بإحضار قائمة من المرادفات لكلمة معينة، ثم تبديلها إذا كانت متوفرة.
 - في الخطوة الأخيرة، وللعمل على التأكد من كفاءة و تحسين جودة النصوص المترجمة بعد تطبيق تقنية تعزيز البيانات النصية على النصوص المترجمة آلياً، أعادت الدراسة توظيف مقاييس روج و بلو لتحديد فعالية تقنية تعزيز البيانات النصية في التعامل مع الكلمات غير المرصودة في النصوص المترجمة.



٤. التحليل والنتائج

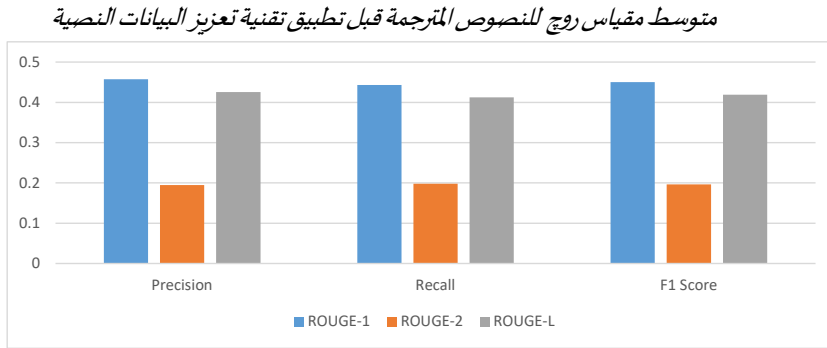
٤,١. قبل تطبيق تقنية تعزيز البيانات النصية

جدول (١)

مثال لنص مترجم من الإنجليزية للعربية باستخدام نموذج المعرفة للترجمة الآلية

Koubek suspended after drugs test: Stefan Koubek says he has been banned for three months by the International Tennis Federation (ITF) after testing positive for a banned substance. The world number 60 failed a routine drugs test at this year's French Open but now plans to lodge an appeal. Koubek believes an injection given to him by an Austrian doctor to treat a wrist injury is to blame for producing traces of the substance in his system. I have acted correctly," the 27-year-old Austrian said in a statement. Koubek, who defeated Britain's Greg Rusedski in the decisive rubber of the Davis Cup in September, is now set to miss the start of the season. He said, "A three-month ban would mean that I not only will miss the Australian Open, but also the Davis Cup in Australia.	ويقول كوبيك معلقاً بعد اختبار الأدوية، إن ستيفان كوبيك يقول إنه تم حظره لمدة ثلاثة أشهر من قبل الاتحاد الدولي للتنس بعد الاختبار موجباً لمادة محظورة، أما العالم رقم ٦٠ فقد فشل في اختبار عادي للعقاقير في هذا العام الفرنسي المفتوح ولكنه يخطط الآن لتقديم نداء، ويرى كوبيك أن حقنة قدمها له طبيب نمساوي لعلاج إصابة في الرسغ هي إلقاء اللوم على إنتاج آثار للمادة في نظامه، فقد تصرف بشكل صحيح، كما قال النمساوي ذو السبعة والعشرين عاماً في بيان، أما كوبيك الذي هزم البريطاني غريغ روستسكي في المطاط الحاسم من كأس ديفز في سبتمبر، فيتوسط الآن ليفوت بداية الموسم، وقال، إن حظراً لمدة ثلاثة أشهر يعني أنني لن أفوت الأسترالي المفتوح فحسب، بل أيضاً كأس ديفيس في أستراليا.
--	--

شكل (١)



بعد اختيار عدد من النصوص من المجموعات الخمس اختياريًا عشوائيًا للترجمة باستخدام نموذج المعرفة للترجمة الآلية من الإنجليزية للعربية (انظر جدول ١)، وظفت مقاييس روج و بلو لتقييم جودة النصوص المترجمة. من خلال النتائج الموضحة في (شكل ١)، يبين الرسم البياني قياسات قيم الدقة (Precision)، والاستدعاء (Recall)، ونقاط F1 لكل مقياس. وعند النظر إلى متوسط مقاييس النقاط المختلفة F1 في روج نجد أنها أعلاها في روج الأولي (ROUGE-1) وتمثل ٠.٤٥٠٢. والجدير بالذكر أن مقياس روج الأولي يقيس تشابه الكلمات بين النص المرجعي (النص الأصلي) والنص المترجم فيما يتعلق بالجملة الواحدة. أي إنه يحسب كمية الكلمات المشتركة بين الجملتين. أما روج الثانية (ROUGE-2) فتقيس كمية الكلمات المشتركة بين جملتين متتاليتين. أما روج إل (ROUGE-L) فتقيس تشابه الكلمات بين النص المرجعي والنص المترجم عبر الجمل بشكل متسلسل. وعند استخدام مقياس بلو لتقييم جودة الترجمة، وجد أن القياس هو ٠.٦٥٢٧. ويعتمد معدل بلو في تقييمه على المقارنة بين n-grams (تسلسلات متتالية من n عناصر مثل الكلمات أو الأحرف) بين الترجمة المرشحة والترجمات المرجعية.



ومن العوامل اللغوية المؤثرة على مقاييس روج و بلو للتقييم، وكما هو موضح في (جدول ١)، عدم توافق المفردات بين النص الأصلي والنص المترجم. فمثلاً، نجد المفردة "a" routine في النص المصدر قد ترجمت إلى "عادي"، بينما المفردة "appeal" قد ترجمت إلى "نداء"، و المفردة "system" في النص المصدر قد ترجمت إلى "نظام". فمع أن كلمات الترجمة يمكن أن تكون مناسبة في سياقات أخرى، إلا أنها قد لا تكون مناسبة تماماً في سياق الترجمة الحالي. ويمكن أن يؤثر عدم التوافق هذا في الاختيارات اللغوية على دقة الترجمة، ومن ثم على نتائج مقاييس التقييم المختلفة.

٤,٢. بعد تطبيق تقنية تعزيز البيانات النصية للتعامل مع الكلمات غير المرصودة

جدول (٢)

مثال لنص مترجم و نص معزز بيانات للتعامل مع الكلمات غير المرصودة باستخدام قاعدة البيانات

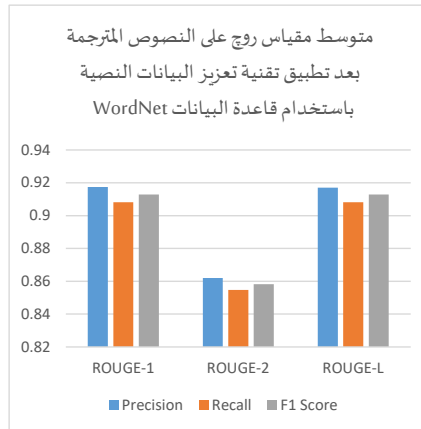
وردت

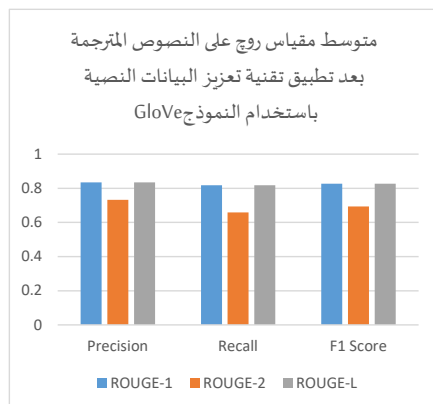
النص المترجم ألياً:	النص المعزز بياناتياً:
ويقول كوبيك معلقاً بعد اختبار الأدوية، إن ستيفان كوبيك يقول إنه تم حظره لمدة ثلاثة أشهر من قبل الاتحاد الدولي للتنس بعد الاختبار موجباً لمادة محظورة، أما اللاعب رقم ٦٠ فقد فشل في اختبار عادي للعقاقير في هذا العام الفرنسي المفتوح ولكنه يخطط الآن لتقديم نداء، ويرى كوبيك أن حقنة قدمها له طبيب نمساوي لعلاج اصابة في الرسغ هي إلقاء اللوم على إنتاج آثار للمادة في نظامه، فقد تصرف بشكل صحيح، كما قال النمساوي ذو السبعة والعشرين عاماً في بيان، أما كوبيك الذي هزم البريطاني كريك روستسكي في المطاط الحاسم من كأس ديفز في سبتمبر،	ويقول كوبيك معلقاً بعد اختبار الأدوية، إن ستيفان كوبيك يقول إنه تم حظره لمدة ثلاثة أشهر من قبل الاتحاد الدولي للتنس بعد الاختبار موجباً لمادة محظورة، أما اللاعب رقم ٦٠ فقد أخفق في اختبار اعتيادي للعقاقير في هذا العام الفرنسي المفتوح ولكنه يخطط الآن لتقديم نداء، ويرى كوبيك أن حقنة قدمها له طبيب نمساوي لعلاج اصابة في الرسغ هي رمي اللوم على إنتاج آثار للمادة في نظامه، فقد تصرف بشكل صحيح، كما قال النمساوي ذو السبعة والعشرين عاماً في بيان، أما كوبيك الذي هزم البريطاني كريك روستسكي في المطاط الحاسم من كأس ديفز في سبتمبر،

سبتمبر، فيتجهز الآن ليضيع بداية الموسم، وقال، إن حظراً لمدة ثلاثة أشهر يعني أنني لن أغيب عن الأسترالي المفتوح فحسب، بل أيضاً كأس ديفيس في أستراليا.	فيتوطد الآن ليفوت بداية الموسم، وقال، إن حظراً لمدة ثلاثة أشهر يعني أنني لن أفوت الأسترالي المفتوح فحسب، بل أيضاً كأس ديفيس في أستراليا.
--	---

شكل (2)

متوسط مقياس روج بعد تطبيق تقنية تعزيز البيانات النصية





كما هو موضح في (شكل ٢)، فإنه وباستخدام قاعدة بيانات ورد نت و تقنية التضمين الكلمائية باستخدام نموذج جلوف لتعزيز النصوص المترجمة آلياً للتعامل مع الكلمات غير المرصودة ومعالجتها في النصوص المترجمة آلياً، فقد وجد أن مقياس روج حقق ارتفاعاً ملحوظاً مقارنة بما قبل تطبيق تقنية تعزيز البيانات النصية. أيضاً، بمقارنة مقياس روج في التقنيتين المستخدمتين لتعزيز النصوص المترجمة، فإنه وجد أنه وباستخدام تقنية استبدال الكلمات بواسطة قاعدة البيانات ورد نت، فإن مقياس روج قد حقق ارتفاعاً في مقاييس النقاط المختلفة مقارنة باستخدام نموذج جلوف. بالنظر إلى متوسط مقاييس النقاط المختلفة F1 في روج، باستخدام تقنية استبدال الكلمات بواسطة قاعدة بيانات ورد نت، نجد أنها جاءت متساوية في روج الأولي و روج الثانية بنقاط ٠.٩١٢٨. وفي حالة استخدام نموذج جلوف، فقد وجد أيضاً تعادلاً بين متوسط مقاييس النقاط المختلفة F1 في روج الأولي و روج ال بنقاط ٠.٨٢٦٥.

أيضاً، وبواسطة قاعدة بيانات ورد نت، فقد تحقق استبدال لبعض المفردات التي ترجمت خارج السياق - انظر (جدول ٢) - مثل استبدال كلمة "المطاط" - والتي كانت ترجمة لكلمة "rubber" - بكلمة "اللقاء". مثال آخر لكلمة "فيتوطد" فقد تم استبدالها بكلمة "فيتجيز". وقد

بلو ارتفاعاً ملحوظاً بعد تطبيق تقنية تعزيز البيانات النصية مقارنة بما قبل تطبيق التقنية. ففي حالة استخدام تقنية استبدال الكلمات بواسطة قاعدة بيانات ورد نت، حقق مقياس بلو النقاط ٠.٠٩٤١٤. وجاء مقياس بلو في حالة استخدام نموذج جلوف بنقاط ٠.٠٩٣٦٤. والجدير بالذكر أن هناك بعض المفردات لم تستبدل في النص المعزز، مثل كلمات "نداء" و "نظام". ويمكن أن يرجع السبب في ذلك إلى أن هذه الكلمات غير المرصودة ليس لديها مرادفات متوفرة في قاعدة البيانات المستخدمة.

٥. الخاتمة

في سياق الدراسة الحالية، وتطبيق تقنيتين مختلفتين لتعزيز البيانات النصية للتعامل مع المفردات غير المرصودة في النصوص المترجمة آلياً من الإنجليزية إلى العربية على مستوى المفردات فقط، فقد خلصت الدراسة إلى أن التعامل مع الكلمات غير المرصودة في الترجمة الآلية هو تحدٍ كبير، ولكنه ضروري لتحسين جودة الترجمة. يتطلب هذا الأمر فهماً دقيقاً للغة، واستخدام تقنيات متقدمة أيضاً، والذي يمكن أن يؤدي إلى نتائج أفضل في تطبيقات الترجمة الآلية. وختاماً، توصي هذه الدراسة بتطبيق تقنيات تعزيز البيانات النصية على النصوص المترجمة آلياً على مستوى أشمل، وهو مستوى الجمل، فعند العمل على مستوى الكلمات فقط، وجدت صعوبة في استبدال كلمات معينة بسبب اختلاف البنيان التركيبي لبعض الجمل من النص المصدر إلى النص المترجم عربياً. مثال على ذلك: الجمل المبنية للمجهول والمبنية للمعلوم في اللغتين.



الهوامش:

¹ marefa-nlp/marefa-mt-en-ar. Huggingface. Co. Retrieved on May 25, 2024,
from [https://huggingface.co/marefa-nlp/marefa-mt-en-ar#%](https://huggingface.co/marefa-nlp/marefa-mt-en-ar#%#)

المراجع العربية

القريش، فطوم. (٢٠٢٤). صعوبات الترجمة الآلية في اللغة العربية المستوى التداولي نموذجاً.

المجلة المصرية لهندسة اللغة، عدد ١١ (١)

خضر، محمد زكي. (٢٠١٨). الترجمة الآلية في خدمة اللغة العربية. مجمع اللغة العربية الأردنية.

المراجع الأجنبية

Fadaee, M., Bisazza, A., & Monz, C. (2017 July). *Data Augmentation for Low-Resource Neural Machine Translation*. In the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada.

Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T. & Hovy, E. (2021). A survey of data augmentation approaches for NLP. In Zong, C., Xia, F., Li, W. & Navigli, R (Eds.) *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 968–988). Association for Computational Linguistics.

Greene, D., & Cunningham, P. (2006 June 25-29). *Practical solutions to the problem of diagonal dominance in kernel document clustering*. In the 23rd International Conference on Machine Learning - ICML '06, Pittsburgh, Pennsylvania, USA.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41. <https://doi.org/10.1145/219717.219748>

Pennington, J., Socher, R., & Manning, C. (2014 October 25-29). *Glove: Global vectors for word representation*. In the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.



- Sennrich, R., Haddow, B., & Birch, A. (2016 August 7-12). *Improving neural machine translation models with monolingual data*. In the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.
- Wang, W. Y., & Yang, D. (2015 September 17-21). *That's so annoying!!!: A lexical and frame semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets*. In the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal.
- Wei, J., & Zou, K. (2019 November 3-7). *EDA: Easy data augmentation techniques for boosting performance on text classification tasks*. In the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China.
- Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019 June 12–14). *Conditional BERT Contextual Augmentation*. In Computational Science- 19th International Conference on Computational Science (ICCS 2019), Faro, Portugal.

نبذة تعريفية عن الباحثين

Author Biodata	بيانات الباحث
Hoda Zaiton is currently pursuing the Master degree in Applied Linguistics, Institute of Applied Linguistics and Translation, Faculty of Arts, Alexandria University in collaboration with College of Language and Communication (Alexandria)-AASTMT. In addition to her MA studies in Applied Linguistics, Hoda has completed the Applied Data Science & Artificial Intelligence Training in March 2023, preceded by the completion of Machine Learning Training in April 2022. Her research is highly interdisciplinary, focused on Natural Language Processing methodologies applied to Arabic and English texts, with a particular interest in the disciplines of Text Classification and Text to Text Generation.	هدى زيتون، ماجستير اللغويات التطبيقية، معهد الدراسات اللغوية والترجمة، كلية الآداب، جامعة الإسكندرية بالتعاون مع الأكاديمية العربية للعلوم والتكنولوجيا والنقل البحري، الإسكندرية، مصر. بالإضافة إلى دراستها للمجستير في اللغويات التطبيقية، أتمت هدى التدريب في علوم البيانات التطبيقية والذكاء الاصطناعي في ٢٠٢٣، مسبقاً بإتمام التدريب في تعلم الآلة في ٢٠٢٢. تنضم أبحاثها بالتعددية، حيث تركز على منهجيات معالجة اللغة الطبيعية المطبقة على النصوص العربية والإنجليزية مثل تلخيص النصوص، تصنيف النصوص وتعزيز البيانات النصية.

معرف أوركيد (ORCID) 6220-7890-0008-0009

Email: hoda.zaitoon@alexu.edu.eg



Author Biodata	بيانات الباحث
Sameh Alansary is the Professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. His main areas of interest are concerned with computational linguistics and language engineering, corpus work, morphological analysis and generation, and building formal grammar. He established the Arabic Computational Linguistics Center in Bibliotheca Alexandrina. Prof. Alansary is the co-founder and Vice president of board of the Arabic Language Technology Center (ALTEC) in Cairo, a Non-Governmental Organization (NGO) aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He is a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland	سامح الأنصاري هو أستاذ اللغويات الحاسوبية ورئيس قسم الصوتيات واللغويات بكلية الآداب جامعة الإسكندرية. مصر. يرتبط اهتمامه باللغويات الحاسوبية وهندسة اللغة، وبناء المدونات اللغوية، والتحليل الصرفي والنحوي والدلالي. أسس مركز اللغويات الحاسوبية العربية بمكتبة الإسكندرية، كما أنه مؤسس مشارك ونائب رئيس مجلس إدارة مؤسسة مركز تقنيات اللغة العربية بالقاهرة (ALTEC)، مؤسسة غير ربحية تهدف إلى توفير الموارد الحاسوبية وبناء خارطة طريق لتقنيات اللغة العربية في مصر والشرق الأوسط. هو عضو في العديد من المنظمات العلمية: (١) الجمعية المصرية لهندسة اللغة - القاهرة، (٢) الجمعية اللغوية العربية - الولايات المتحدة الأمريكية، (٣) جمعية اللغويات الحاسوبية - الولايات المتحدة الأمريكية - أوروبا، (٤) مؤسسة لغة الشبكات الدلالية العالمية، الأمم المتحدة - سويسرا. وله أكثر من ٨٠ بحثاً منشوراً في المؤتمرات والمجلات العلمية المحلية والإقليمية والدولية. أشرف وناقش أكثر من ٦٠ رسالة علمية لمرحلي الماجستير والدكتوراه. وحصل على الجائزة التشجيعية وجائزة العلوم الحديثة والمتطورة لجامعة الإسكندرية.

معرف أوركيد (ORCID) : 0009-0007-2950-9555

E-mail: S.alansary@alexu.edu.eg

«مدار»: موارد حاسوبية للهجات العربية

«MADAR»:

Multi-Arabic Dialect Applications and Resources

المعرف الرقمي: <https://doi.org/10.60161/2521-001-002-005>

هدى بوعمر^١ نزار حبش^٢ سلام خليفة^٣ فضل الإيراني^٤

^١ قسم نظم المعلومات، جامعة كارنيجي ميلون قطر، الدوحة، قطر

^٢ مختبر كامل، قسم علوم الحاسوب، جامعة نيويورك أبو ظبي، أبو ظبي، الإمارات العربية المتحدة

^٣ قسم اللسانيات، جامعة ستوني بروك، نيويورك، الولايات المتحدة

^٤ قسم اللغويات العامة والحاسوبية، كلية العلوم الإنسانية، جامعة توينجن، توينجن، ألمانيا

توثيق البحث APA Citation

بوعمر، هدى، حبش، نزار، خليفة، سلام، والإيراني، فضل. (2024). مدار: موارد حاسوبية للهجات العربية. مجلة اللسانيات العربية، (٢).

Submission Date: 2 /01/2024, Acceptance Date:19/09/2024

تاريخ الإرسال: ٢٠٢٤/١/٢ / تاريخ القبول: ٢٠٢٤/٩/١٩ م

Abstract

This paper presents the results of the MADAR project, which aims to build scientific foundations for managing linguistic differences between Modern Standard Arabic (MSA) and its dialects, while developing specialized resources and applications. A key feature is its focus on 25 Arab cities, from the Maghrib to the Gulf. The project successfully developed: (a) the MADAR Corpus, with 140K dialectal sentences and translations in MSA, English, and French; (b) the MADAR Lexicon, with 47K dialectal entries; (c) the MADAR Twitter Corpus, containing nearly 300K tweets from 3K users; and (d) the MADAR CODA Corpus, which provides spelling corrections for dialectal Arabic.

Keywords: Arabic dialects, natural language processing, MADAR corpus, MADAR lexicon.

الملخص:

تقدم هذه الورقة البحثية نتائج مشروع (مدار) الذي عمّد إلى بناء أسس علمية لتحديد وتنظيم أساليب التعامل مع الاختلافات اللغوية بين العربية الفصحى ولهجاتها، وتطوير موارد وتطبيقات متخصصة لمعالجتها. الميزة الأساسية في هذا المشروع هي التركيز غير المسبوق على عدد كبير من المدن العربية، 25 مدينة من المحيط إلى الخليج. نركز في هذه الورقة على أربعة مواد نجح المشروع في تطويرها: أ- مكنز مدار للسفر: ويحتوي على 140,000 جملة من لهجات 25 مدينة، مقترنة بترجمات بالفصحى والإنجليزية والفرنسية. ب- معجم مدار: ويحتوي على أكثر من 47,000 مدخلاً من لهجات المدن الـ 25. ج- مكنز مدار تويتر: ويضم قرابة 300,000 تغريدة من 2,980 مستخدماً، مع تصنيف لهجاتهم يدوياً. د- مكنز مدار كودا: الذي يثري جزءاً من مكنز مدار بتصحيح إملائي على نهج كودا، كتابة واحدة للدارجة. **الكلمات المفتاحية:** اللهجات العربية، معالجة اللغة الطبيعية، مكنز مدار معجم مدار للهجات



١. المقدمة

اللغة العربية اليوم هي مجموعة من اللهجات التي ترتبط تاريخياً باللغة العربية الفصحى (Classical Arabic). اللغة العربية القياسية الحديثة^١ (Modern Standard Arabic) هي اللغة الرسمية المكتوبة المعترف بها للثقافة والإعلام والتعليم في جميع أنحاء العالم العربي. أما اللهجات المحكية (غير قياسية بالطبع) تختلف فيما بينها وبين الفصحى القياسية. الاختلافات تكون في المعجم والأصوات والصرف، وأقلها في النحو. ونضرب في هذا بعض الأمثلة، فيختلف مثلاً نطق صوت^٢ حرف القاف بين q/ في الفصحى القياسية واللهجة التونسية، وw/ (همزة) في اللهجة المصرية واللهجة الشامية، وg/ (جيم غير معطشة) في اللهجة الخليجية. تُكتب الفصحى القياسية بقواعد إملائية معروفة، أما اللهجات فلا قواعد تضبط إملاءها، فالمتحدثون بها إما أن يكتبوا ما ينطقون من أصوات مغايرة أو يرجعونها إلى أصولها الفصيحة أو كليهما، وبعضهم قد يستعمل حروفاً غير عربية للكتابة (العريزي أو الفرانكو). تختلف اللهجات كذلك في الصرف فمن أوجه الاختلافات حروف الاستقبال والمضارعة وأوزان الأفعال وغيرها. أما في النحو فالاختلافات طفيفة مقارنة بالاختلافات الصرفية، ولكن صيغ النفي تختلف في بعض اللهجات. الاختلافات المعجمية بين اللهجات كثيرة أيضاً، حيث تُستخدم كلمات مختلفة لنفس المعنى في مناطق مختلفة. تشكل هذه الاختلافات تحديات كبيرة للمعالجة الآلية للغة العربية، فالأدوات المطورة للفصحى القياسية أو اللهجات غير كافية لمعالجة تعقيدات اللهجات العربية.

في علوم اللسانيات تُدرّس اللهجات العربية في نطاق ما جُمع من البيانات القليلة. أما في سياق المعالجة الآلية للغة العربية، فقد حصلت الفصحى القياسية على النصيب الأكبر من الاهتمام، فالفصحى القياسية العديد من مجموعات البيانات المتوازنة والأحادية، والمجموعات غنية التوسيم كأشجار التحليل النحوي وأدوات متقدمة لتحليل الصرف والجمل وما إلى ذلك (Habash, 2010). كانت الجهود لإنشاء موارد للهجات العربية محدودة بعدد قليل من اللهجات الرئيسية (Habash et al., 2013). جاء بعض الباحثين بأدوات وموارد للهجات عن طريق ربطها بأدوات وموارد للفصحى القياسية (Chiang et al., 2006; Salloum and Habash, 2012). نجحت بعض الجهود بجمع البيانات

من خلال الحشود (crowdsourcing) لبناء موارد متخصصة للهجة بعينها، كاليانان المتوازية للترجمة (Zbib et al.,2012). هذا بالإضافة لجهود عديدة في تمييز اللهجات على المستوى الإقليمي (Elfardy, 2013; Zaidan and Callison-Burch, 2013; and Diab, 2013).

سعى مشروع «مدار» إلى تطوير مجموعة من الموارد لم تُعهد من قبل للغة العربية (بمعناها الجماعي الفصيح والعلمي) وإظهار قيمة هذه الموارد في الترجمة الآلية للهجات وتمييز اللهجات.³ نركز في هذه الورقة على أربعة موارد مصممة لخدمة التطبيقين الرئيسيين في المشروع، ولكن قيمتها تتجاوز هذه التطبيقات. الأول هو مكنز «مدار» للسفر الذي يحتوي على ١٤٠,٠٠٠ جملة مقسمة على خمس وعشرين مدينة من الرباط إلى مسقط، ومقارنة بترجمات إلى الفصحى القياسية والإنجليزية والفرنسية. الثاني هو معجم «مدار» الذي يحتوي على أكثر من ٤٧,٠٠٠ مدخلا من نفس المدن تضم أهم المفاهيم في مكنز «مدار» للسفر بشكل مختصر صرفيًّا مدعومًا بتمثيلات قياسية للأصوات والإملاء. صاحب إنشاء المعجم تطوير إرشادات للتوسيم نأمل أن يتبعها الآخرون في المستقبل لتوحيد الجهود. الثالث هو مكنز «مدار تويتر» الذي يضم قرابة ٣٠٠,٠٠٠ تغريدة من ٢,٩٨٠ مستخدماً مع تصنيف لهجاتهم يدونا.

استمر هذا الجهد لأكثر من ثلاث سنوات، وشارك فيه حوالي ٦٠ فردًا بأدوار مختلفة لإنشاء هذه المجموعة الفريدة من الموارد التي نأمل أن تمهد الطريق لمزيد من البحث في المعالجة الآلية للغة العربية بعموم لهجاتها.

٢. تمهيد لغوى وأعمال ذات صلة

تواجه اللغة العربية العديد من التحديات في مجال المعالجة الآلية، لأنها تجمع بين الصرف الغني والتشكيل الاختياري، والذي ينتج بسببه عدد ضخم من الكلمات الممكنة، كل منها بتشكيلات ومعانٍ مختلفة. فمثلاً نجد بين تصاريف الفعل "بَيَّنَّ": بَيَّنْتُ، بَيَّنْتَهَا، فَسَيَّبْتُوْنَهَا، إلخ. وفي نفس الوقت، نجد تحاليل عديدة للكلمة "بين" غير المشكلة: بَيْنٌ، بَيْنٌ، بَيْنٌ، ب "يُنْ" (ياباني). كما يشكل نقص المعايير الإملائية للهِجَات وتعدد أنواعها تحديات إضافية، مثل توفر مكان موسمة كافية لاستخدامها في تدريب أدوات



المعالجة الآلية، وهو أمر مكلف للغاية. في سياق العمل على المعالجة الآلية، تفتقر اللهجات العربية إلى الأدوات والموارد مقارنة بما هو متاح للفصحى القياسية، مع وجود القليل من المكنز المتوازنة بين اللهجات العربية وما يقابلها من الإنجليزية أو الفصحى القياسية أو لهجات عربية أخرى. ولا يمكن لأدوات الفصحى القياسية نمذجة اللهجات العربية بفعالية بسبب الاختلافات التي ذكرناها آنفاً، مثل عدم قدرة محلل صرفي للفصحى القياسية على تحليل ٤٠٪ من الأفعال في اللهجة الشامية (Habash and Rambow, 2006).

ركزت الأبحاث الحديثة على المعالجة الآلية للهجات العربية مدعومة بمكنز ومعجم جديدة أحادية ومتعددة اللهجات. من بين الجهود ذات الصلة مكنز متخصصة للهجة المصرية (Gadalla et al., 2014; Maamouri et al., 1997; al.), وأخرى للفلسطينية (Jarar et al., 2016) والتونسية (Zaidan et al., 2014; McNeil, 2018)، وكذلك مكنز غير متوازنة متعددة اللهجات (and Callison-Burch, 2011; Cotterell and Callison-Burch, 2014; Salama et al., 2014; 2019; Zaghrouani and Charfi, 2018; Alshargi et al., 2019). وهناك عدد من المكنز المتوازنة الصغيرة الحجم بين اللهجات مثل المكنز سباعي اللهجات المتوازي (Bouamor et al., 2014) والمكنز سداسي اللهجات المتوازي المعروف ببإدك (Meftouh et al., 2015) وقاعدة بيانات جامعة مانشستر (Matras et al., 2017)، مقارنة بمكنز أكبر نسبياً للهجات معينة مع اللغة الإنجليزية (Zbib et al., 2012). وأخيراً تفتقر اللهجات العربية أيضاً إلى المعاجم الحاسوبية مما دفع الجهود لبناء معاجم للهجات كاللهجة المصرية (Dibas et al., 2022) والفلسطينية (Diab et al., 2014; Kilany et al., 2014). للمزيد عن الأعمال ذات الصلة، يمكن الاطلاع على (Shoufan and Al-Ameri, 2015) و (Al-Ayyoub et al., 2017).

٣. موارد مشروع «مدار»

أنظمة المعالجة الآلية للغات الطبيعية تعتمد على البيانات اعتماداً كبيراً لتطويرها وتقييمها، ولذلك تساهم البيانات الغنية والموسمة بشكل كبير في تحسين هذه الأنظمة. ومع توفر الكثير من

البيانات الخام للهجات العربية مثل تغريدات تويتر/أكس والمدونات (blogs) وحتى ويكيبيديا لهجات (ويكيبيديا مصري ومغربي) فإنها جميعها تفتقر إلى التوسيمات الضرورية لتطبيق تمييز اللهجات وترجمتها آلياً. اللهجات العربية، كما ذكر سابقاً، تمثل تحدياً بسبب الإملاء غير المعياري والصرف الغني. لذلك وجب علينا أن نوازن بين البيانات المتخصصة التي يمكن أن تكون مفتعلة (مترجمة من مصادر غير عربية بهدف بناء مكانز متوازنة) والبيانات الطبيعية التي تتضمن تبايناً طبيعياً يعكس طبيعة المتحدثين (أي المكتوبة أساساً بنية التواصل بين متحدثي اللهجة مثل تغريدات تويتر/أكس). لمعالجة هذه الأمور طورنا أربعة موارد متخصصة نستعرضها أدناه.

٣,١. مكانز «مدار» للسفر

أنشأنا مكانز «مدار» للسفر، وهو أول مجموعة من الجمل المتوازنة التي تضم لهجات ٢٥ مدينة عربية، بالإضافة إلى الإنجليزية والفرنسية والفصحى القياسية. مصدر المكانز جمل مختارة من مكانز التعابير الأساسية للسفر (BTEC) (Takezawa et al., 2002)، وهو مكانز متعدد اللغات، بما فيها الفصحى القياسية، والفرنسية والإنجليزية. ويحتوي على جمل قصيرة جوارية متعلقة بالسفر والسياحة، مما يجعله مثالياً لتدريب وتقييم نماذج الترجمة الآلية بين اللغات واللهجات المستخدمة. اخترنا ٢٠,٠٠٠ جملة من مكانز BTEC وترجمت جميعها إلى لهجات خمس مدن مختارة: الدوحة وبيروت والقاهرة وتونس والرباط، لتشكل مع ما يقابلها في الفصحى القياسية «مكانز-٦». بالإضافة إلى ذلك، تُرجمت ٢,٠٠٠ جملة مختارة من «مكانز-٦» إلى لهجات ٢٠ مدينة أخرى، لتشكل «مكانز-٢٦». قام المترجمون، وجميعهم من ناطقي لهجات المدن المعنية بالسليقة، بترجمة الجمل من الإنجليزية أو الفرنسية إلى لهجاتهم، متجنين الفصحى القياسية كنقطة انطلاق لمنع التأثير اللغوي غير المقصود (Bouamor et al., 2014). تفاصيل الإرشادات المقدمة للمترجمين موجودة في (Bouamor et al., 2018). جدول ١ يحتوي على جمل متوازنة من مكانز-٢٦.



جدول (١)

أمثلة لجمل متوازية من ٢٥ مدينة عربية، والفصحى القياسية، والإنجليزية والفرنسية.

الفصحى القياسية			. أرغب في تأجير سيارة لمدة أسبوع
English			I'd like to hire a car for a week.
French			Je voudrais louer une voiture pour une semaine.
الخليج العربي	سلطنة عمان	مسقط	ابغا استنجر سيارة لمدة أسبوع.
الخليج العربي	المملكة العربية السعودية	جدة	أبا استاجر سيارة لمدة أسبوع.
الخليج العربي	المملكة العربية السعودية	الرياض	ابغى استأجر سيارة لاسبوع.
الخليج العربي	قطر	الدوحة	بغيت أأجر سيارة حق اسبوع.
خليج عدن	اليمن	صنعاء	اشتي استخدم سياره لمدة اسبوع.
العراق	العراق	بغداد	اريد استأجر سيارة هذا الأسبوع.
العراق	العراق	البصرة	اريد أأجر سيارة لمدة اسبوع.
العراق	العراق	الموصل	اغيد استاجر سياره لاسبوع.
بلاد الشام	سوريا	دمشق	بدي استأجر سيارة لمدة اسبوع.
بلاد الشام	سوريا	حلب	بدي أستأجر سيارة أسبوع.
بلاد الشام	لبنان	بيروت	بدي إستأجر سيارة عجمعة.
بلاد الشام	الأردن	عمان	بدي استأجر سيارة للأسبوع.

السلط	الأردن	بلاد الشام	بدي استأجر سيارة لأسبوع.
القدس	فلسطين	بلاد الشام	بدي أستأجر سيارة لأسبوع.
القاهرة	مصر	حوض النيل	عاوز أأجر عربية لمدة أسبوع.
أسوان	مصر	حوض النيل	كنت عايز أأجر عربية لمدة أسبوع.
الإسكندرية	مصر	حوض النيل	عاوز أأجر عربية لمدة أسبوع.
الخرطوم	السودان	حوض النيل	داير أأجر عربية لأسبوع.
طرابلس	ليبيا	المغرب العربي	نبي نأجر سياره لأسبوع.
بنغازي	ليبيا	المغرب العربي	نبي نأجر سيارة لأسبوع.
تونس	تونس	المغرب العربي	نحب نكري كرهبة ل جمعة من زمان.
صفاقس	تونس	المغرب العربي	نحب نكري كرهبة لمدة جمعة.
الجزائر العاصمة	الجزائر	المغرب العربي	راني حاب نكري سيارة لأسبوع.
الرباط	المغرب	المغرب العربي	بغيت نكري طوموبيلة لسيمانا.
فاس	المغرب	المغرب العربي	بغيت نكري سيارة لمدة أسبوع.



٣,٢. معجم «مدار»

بالإضافة إلى مكنز «مدار» للسفر، طورنا أوسع معجم متعدد اللهجات، يغطي لهجات ٢٥ مدينة. على النقيض من المكانز، تركز المعاجم على الكلمات منفردة وفي شكلها الصرفي المجرد، أو المدخل المعجمي (Lemma). نطلنا معجم «مدار» حول مفاهيم رئيسة مُعرّفة بثلاثية من الكلمات الإنجليزية، والفرنسية، والفصحى القياسية، كلها مستخرجة من مكنز BTEC باستخدام المحاذاة الآلية للكلمات (automatic word alignment). يقلل هذا النهج متعدد اللغات من الالتباس الناجم عن المعاني المختلفة للكلمة الواحدة. تفاصيل تحديد واستخراج تلك المفاهيم موجودة في (Bouamor et al., 2018). كل مدخل معجمي مرتبط بمفهوم معين يُعرّف بثلاثة جوانب: إملاء معياري على نهج كودا CODA، التمثيل الصوتي بنظام كافي CAPHI، والمدن التي تستخدم فيها الكلمات. اتبعنا مجموعة موحدة من الإرشادات للإملاء المعياري للهجات العربية مستمدة من (Habash et al., 2018). شمل إنشاء المعجم تعبئة تلقائية لبعض المداخل من المعاجم الموجودة، يليه التحقق اليدوي والتوسعة بواسطة ثلاثة عشر لغوياً من مناطق عربية مختلفة. يحتوي معجم «مدار» على ١,٠٤٥ مفهوماً، يغطي ٨٨٪ من المداخل المعجمية الإنجليزية لكلمات مكنز BTEC. المعجم النهائي يحتوي على ٤٦,٢٧٥ مدخلا. بعض أمثلة مداخل المعجم موجودة في الجدول ٢.

جدول (٢)

أمثلة للمداخل المعجمية المرتبطة بمفهوم (very, très, جدا)

كودا	CAPHI كافي	لهجات المدن
برشا	bar sha	تونس، صفاقس
بزاف	be z z a a f	الرباط، فاس، الجزائر العاصمة
بكل	bik kil	بنغازي
جدا	gid dan	القاهرة، الإسكندرية
جدا	jid dan	جدة، الخرطوم، الرياض

القااهرة، الإسكندرية، أسوان	kh a a l i s	خالص
الخرطوم	sh a d i i d	شديد
القااهرة، الإسكندرية	2 a w i	قوي
أسوان، صنعاء	g a w i	قوي
الإسكندرية، القااهرة	k i t i i r	كثير
بيروت، القدس، دمشق، حلب، عمان، فاس، الرباط	k t i i r	كثير
عمان، السلط	k t h i i r	كثير
الموصل	k t h i i g h	كثير
جدة، أسوان، الخرطوم	k a t i i r	كثير
الرياض، مسقط	k i t h i i r	كثير
البصرة، بغداد	k u l l i s h	كلش
الموصل، الدوحة	k e l l i s h	كلش
جدة	m a r r a	مرة
طرابلس	h a l b a	هلبة
مسقط	3 o o m	عوم
البصرة، بغداد	h w a a y a	هوايا
البصرة، بغداد، الدوحة	w a a y i d	واجد
بنغازي، طرابلس، الدوحة	w a a j i d	واجد
مسقط	w a a g i d	واجد

٣,٣. مكنز «مدار تويت»

يتفاضل مكنز «مدار تويت» على مكنز «مدار» للسفر المتحكم بتقديم بيانات طبيعية، أي إنها أقرب للهجة المحكية الطبيعية وليست مفتعلة، ولكنه أقل ثراءً؛ لأنه ليس متوازناً مع لغات أو لهجات



أخرى. يشمل هذا المكنز قرابة ٣٠٠,٠٠٠ تغريدة من ٢,٩٨٠ مستخدماً من ٢١ دولة عربية مع تصنيف لهجاتهم يدوياً. جمعنا التغريدات باستخدام ٢٥ وسماً ابتدائياً تتوافق مع دول جامعة الدول العربية الـ ٢٢ (مثل #Kuwait، #Egypt، #Algeria، إلخ.) بالإضافة إلى الوسوم الثلاثة التالية: #Arab، #ArabWorld، #ArabLeague. على نهج مستوحى من (Mubarak and Darwish, 2014; Zaghouani and Charfi., 2018). جُمع عدد متساوٍ من الحسابات الشخصية (١٧٥ * ٢٥ = ٤,٣٧٥) لكل وسم وصُنفت يدوياً بواسطة ثلاثة مقيمين عرب. بالنسبة لكل حساب شخصي لمستخدم مصنّف، يُحتفظ بأول ١٠٠ تغريدة متاحة وقت التجميع تستخدم كملف تعريف (user profile). فحص المقيّمون كل ملف تعريف وصنفوه بناءً على المعلومات المتوفرة فيه، مثل لهجة التغريدات ومواضيعها والهوية الوطنية للمستخدم والتي قد يعبر عنها باستخدام شعارات وصور عن بلده، وإذا لزم الأمر الاعتماد على معلومات من وسائل تواصل اجتماعي أخرى. قد يتم تصنيف بعض الملفات الشخصية على أنها غير شخصية، أو غير عربية، أو صعبة التخمين. على مجموعة مشتركة من ١٥٠ ملف تعريف، حقق المقيّمون متوسط درجة ارتباط كاي ٨٠,١٦٪ (Cohen, 1968)، مما يشير إلى مستوى توافق كبير.

بعض الأمثلة من مكنز مدار تويتر موجودة في الجدول ٣.

جدول (٣)

أمثلة من مكنز مدار تويتر

الدولة	المستخدم	التغريدة
الكويت	DrBehbehaniAM	وتقولون ليش الأطباء داشين تويتر
الكويت	DrBehbehaniAM	لا القافلة وقفت، ولا الكلاب تعبت.
الكويت	DrBehbehaniAM	مسالنور على قوم التويتير مساء ما يليق الا بالطيبين مساء الابتسامه من القلب
مصر	HederAshraf	الى عايز يجيب جون في الأهلى يجيب زى ما هو عايز بس الهدف يتحسب أو لا دي بتاعتنا

مصر	HederAshraf	يعني هي جت عالجون دة بس الدورى مليون بلاوى
مصر	HederAshraf	ههههههههههه مش عارف انت مكبر الموضوع كدة ليه انا حاسيت اننا هنحمر فلسطين
الجزائر	samykhalildz	اعطيهم فكرة علاش راك غير تنتقد ..
الجزائر	samykhalildz	صلاة المغرب رسميا سيصبح اسمها صلاة الجزائر فى السعودية
الجزائر	samykhalildz	تسريبات: نائب افلانى من بليدة قد يكون متهمًا فى قضية كوكابين العاصمة .

يشمل مكتب «مدار كودا» على ٢,٠٠٠ جملة من المدن الخمس الرئيسية في مكتب «مدار» للسفر من الجزء الموجود في مكتب-٢٦. تم تعديل هذه الجمل يدويًا لتتبع نهج الكتابة الواحدة للدرجة «كودا*» CODA (Habash et al, 2018). ستكون هذه البيانات مفيدة للباحثين الذين يعملون على ضبط نصوص اللهجات العربية، وهي مهمة ستساعد في تقليل تأثير التحدي الهائل الذي يواجه اللهجات العربية من حيث جودة الموارد. للمزيد من التفاصيل، يُرجى مراجعة (Eryani et al., 2020). بعض الأمثلة موضحة في الجدول ٤.

أمثلة من مكنز مدار كودا الجمل تقابل الجملة الفصحى:

«حسنًا، هل يمكنك أن تخبرها أنني اتصلت، من فضلك؟ سأتصل بها مرة أخرى فيما بعد.»

المدينة	الجملة الأصلية	الجملة المصححة حسب نظام كودا
بيروت	طبيب، فيك تقلا إني دقيت، عمول معروف؟ رح دقلا مرة ثاني.	طبيب، فيك تقول لها اني دقيت، اعمول معروف؟ رح دق الها مرة ثانية.
القاهرة	تمام . ممكن تقولها اني اتصلت، من فضلك؟ حاكمها ثاني بعدين .	تمام. ممكن تقول لها اني اتصلت، من فضلك؟ حاكمها ثاني بعدين.



الدوحة	انزين، ممكن تقول لها اني اتصلت، لو سمحت؟ باتصل لها مرة ثانية بعدين.	انزين، ممكن تقول لها اني اتصلت، لو سمحت؟ باتصل لها مرة ثانية بعدين.
تونس	باهي، تنجم تقول لها الي انا تكلمت، يعيشك.	باهي، تنجم تقول لها الي انا تكلمت، يعيشك.
الرباط	المهم، تقدر تقول لها بلي تاصلت، عافاك؟ غادي نعاود نعيط من بعد.	المهم، تقدر تقول لها بالي تاصلت، عافاك؟ غادي نعاود نعيط من بعد.

٤. تأثير مشروع «مدار»

هذا المشروع الذي انتهى قبل خمس سنوات كان له تأثير كبير على المعالجة الآلية للغة العربية، حيث حصلت الورقة الرئيسية على أكثر من ٢٠٠ اقتباس (وحصلت الأوراق المختلفة التي أنجزت في المشروع مجتمعةً على أكثر من ٥٠٠ اقتباسي). كان المشروع مؤثراً في العديد من المهام اللاحقة للغة العربية. فيما يلي، نستعرض التأثير على ثلاث مهام رئيسية

٤.١. تمييز اللهجة

وهو مهمة تصنيف لهجة مقطع معين من الكلام أو النص بأي حجم (أي كلمة أو جملة أو مستند) تلقائياً. تناول الباحثون (Salameh et al, 2018) من مشروع مدار مهمة تمييز اللهجات بدقة عالية تغطي ٢٥ مدينة بالإضافة إلى الفصحى القياسية. أظهرنا أنه في المتوسط يتطلب الأمر ٥٢ كلمة للوصول إلى تصنيف مثالي للهجة ١٦ كلمة للوصول إلى دقة بنسبة ٩٠٪. كانت هذه الجهود بمثابة الأساس لأول مسابقة أجريتها حول تمييز اللهجات جذبت ٢١ فريقاً من ١٥ دولة (Bouamor et al., 2019). تبعها عدة مسابقات مماثلة وهي سلسلة مسابقات «نادي» NADI (Abdul-Mageed et Mishra and Mujadia, 2021, 2022, 2023) كما تبعها أعمال أخرى مستقلة عن تلك المسابقات (Mishra and Mujadia, 2021, 2022, 2023; Olsen et al., 2023; Keleg et al., 2023; Abdelali et al., 2021; Baimukan et al., 2022; Olsen et al., 2023; Keleg et al., 2023; Keleg and Magdy, 2023).

٤,٢. الترجمة الآلية للهجات

اكتسبت الترجمة من وإلى اللهجات العربية زخمًا بعد إصدار مكنز «مدار». استهدفت عدة جهود أهداف ترجمة مختلفة، مثل الترجمة من اللهجة إلى اللهجة (Rassem, 2023)، والترجمة من اللهجة إلى الفصحى القياسية (Baniata et al., 2021; Slim et al., 2022; Baniata et al., 2022; Babaali et al., 2024) بالإضافة إلى عدة مسابقات (Abdul-Mageed et al., 2021, 2022, 2023; Atwany et al., 2024). والترجمة من اللهجة إلى الإنجليزية (Sajjad et al., 2020; Ko et al., 2021)، والترجمة من اللهجة إلى التركية (Alkheder et al., 2023).

٤,٣. ضبط النصوص

ضبط نصوص اللهجات العربية (text normalization) هي مهمة جديدة نسبيًا. أنشأ (Eryani et al., 2020) مكنز «مدار كودا» المذكور في القسم ٣,٤. ومؤخرا جاء (Alhafni et al., 2024) بأول نظام ضبط تلقائي يجمع بين تمييز اللهجات وضبط النصوص ويعتمد كلياً على موارد «مدار».

٥. الخاتمة

في الختام، قاد مشروع «مدار» تقدماً كبيراً في مجال المعالجة الآلية للغة العربية، من خلال تطوير وتوفير مجموعة شاملة من الموارد للهجات العربية المتعددة، وربطها ببعضها البعض وبالفصحى القياسية. يعالج مكنز «مدار» للسفر، والمعجم، ومكنز مدار تويتر، ومكنز مدار كودا فجوات حرجية في توافر البيانات، مما يعزز من دقة وقوة تطبيقات المعالجة الآلية للغة العربية. استخدمت موارد المشروع لتطوير الأبحاث في تمييز اللهجات، والترجمة الآلية، وضبط النصوص. «مدار» مبادرة رائدة تضع معايير جديدة لموارد اللهجات العربية، مما يمهّد الطريق لمزيد من التقدم في المجال، ويدعم مجموعة واسعة من الدراسات اللغوية والحاسوبية. يتجلى تأثير هذا العمل من خلال عدد الاقتباسات الكبير والتطور المستمر للمهام والمنهجيات ذات الصلة، مما يبرز مساهمته المستدامة في المعالجة الآلية للغة العربية.



شكر وتقدير

لم تكن نتائج هذا المشروع ممكنة دون الفريق الممتاز من الباحثين والمطورين واللغويين والمقيمين ومجموعة تزيد عن ٦٠ فرداً، معظمهم من الناطقين بالعربية بالسليقة من عدة دول حول العالم العربي. استفاد المشروع من التأزر مع مشروعين آخرين لتوسيم اللهجات في جامعتي نيويورك أبوظبي وكولومبيا لمصلحة المشاريع الثلاثة. وقد دُعم البحث الذي أجري في هذا المشروع والمُوضح في هذه الورقة بمنحة NPRP ٧-٢٩٠-١-٤٧ من صندوق قطر الوطني للبحث العلمي (عضو مؤسسة قطر). التصريحات الواردة هنا هي مسؤولية المؤلفين وحدهم.

الهوامش

^١ تُعرف كذلك بفصحى العصر Modern Standard Arabic، سنسميها الفصحى القياسية اختصاراً.

^٢ نستخدم الألفبائية الصوتية الدولية (IPA) لتمثيل الأصوات.

^٣ صفحة المشروع: <http://madar.camel-lab.com>



المراجع

- Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., and Darwish, K. (2021 April 19). *QADI: Arabic dialect identification in the wild*. In the Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Virtual).
- Abdul-Mageed, M., Elmadany, A., Zhang, C., Nagoudi, E. M. B., Bouamor, H., and Habash, N. (2023 December 7). *NADI 2023: The fourth nuanced Arabic dialect identification shared task*. In the ArabicNLP Conference 2023, Singapore (Hybrid).
- Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2021 April 19). *NADI 2021: The second nuanced Arabic dialect identification shared task*. In the Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Virtual).
- Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2022 December 8). *NADI 2022: The third nuanced Arabic dialect identification shared task*. In the seventh Arabic Natural Language Processing Workshop (WANLP), Abu Dhabi, United Arab Emirates (Hybrid).
- Al-Ayyoub, M., Nuseir, A., Alsmearat, K., Jararweh, Y., and Gupta, B. (2018). Deep Learning for Arabic NLP: A Survey. *Journal of Computational Science*, 26, 522-531. <https://doi.org/10.1016/j.jocs.2017.11.011>
- Alhafni, B., Al-Towaity, S., Fawzy, Z., Nassar, F., Eryani, F., Bouamor, H., and Habash, N. (2024 August 16). *Exploiting dialect identification in automatic dialectal text normalization*. In the Second Arabic Natural Language Processing Conference, Bangkok, Thailand.
- Alkheder, H., Bouamor, H., Habash, N., and Zengin, A. (2023 September 4-8). *Benchmarking dialectal Arabic-Turkish machine translation*. In Machine Translation Summit XIX, Macau SAR, China.
- Alshargi, F., Dibas, S., Alkhereyf, S., Faraj, R., Abdulkareem, B., Yagi, S., Kacha, O., Habash, N., and Rambow, O. (2019 August 1). *Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan*. In the Fourth Arabic Natural Language Processing Workshop, Florence, Italy.

- Atwany, H., Rabih, N., Mohammed, I., Waheed, A., and Raj, B. (2024 May 20-25). *OSACT 2024 task 2: Arabic dialect to MSA translation*. In the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation, Torino, Italia (Hybrid).
- Babaali, B., Salem, M., and Alharbe, N. R. (2024). Breaking language barriers with chatgpt: enhancing low-resource machine translation between Algerian Arabic and MSA. *International Journal of Information Technology*, pp. 1–10. <https://doi.org/10.1007/s41870-024-01926-7>
- Baimukan, N., Bouamor, H., and Habash, N. (2022 June 20-25). *Hierarchical aggregation of dialectal data for Arabic dialect identification*. In the 13th Conference on Language Resources and Evaluation (LREC 2022, Marseille, France).
- Baniata, L. H., Ampomah, I. K. E., and Park, S. (2021). A transformer-based neural machine translation model for Arabic dialects that utilizes subword units. *Sensors*, 21(19), 6509. <https://doi.org/10.3390/s21196509>
- Baniata, L. H., Kang, S., and Ampomah, I. K. E. (2022). A reverse positional encoding multi-head attention-based neural machine translation model for Arabic dialects. *Mathematics*, 10(19), 3666. <https://doi.org/10.3390/math10193666>
- Belinkov, Y. and Glass, J. (2016 December 12). *A Character-level Convolutional Neural Network for Distinguishing Similar Languages and Dialects*. In the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), Osaka, Japan.
- Bouamor, H., Habash, N., and Oflazer, K. (2014 May). *A Multidialectal Parallel Corpus of Arabic*. In the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.
- Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). *The MADAR Arabic Dialect Corpus and Lexicon*. In The International Conference on Language Resources and Evaluation Conference (LREC), Miyazaki, Japan.



- Bouamor, H., Hassan, S., and Habash, N. (2019 August 1). *The MADAR shared task on Arabic fine-grained dialect identification*. In The Fourth Arabic Natural Language Processing Workshop (WANLP 2019), Florence, Italy.
- Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. (2006 April 3-7). *Parsing Arabic Dialects*. In the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4), 213- 220. <https://doi.org/10.1037/h0026256>
- Cotterell, R. and Callison-Burch, C. (2014). A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC), pages 241–245, Reykjavik, Iceland.
- Diab, M., Hacioglu, K., & Jurafsky, D. (2007). *Automatic processing of modern standard Arabic text*. Arabic Computational Morphology: Knowledge-based and Empirical Methods, 159-179. Springer Netherlands.
- Diab, M. T., Al-Badrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., and Eskander, R. (2014 May). *Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon*. In the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.
- Dibas, S., Khairallah, C., Habash, N., Sadi, O. F., Sairafy, T., Sarabta, K., & Ardah, A. (2022). Maknuune: A Large Open Palestinian Arabic Lexicon. In Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid).
- Elfardy, H. and Diab, M. (2013 August 4-9). *Sentence Level Dialect Identification in Arabic*. In the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria.

- Eryani, F., Habash, N., Bouamor, H., and Khalifa, S. (2020 May 11-16). *A spelling correction corpus for multiple Arabic dialects*. In the Twelfth Language Resources and Evaluation Conference, Marseille, France.
- Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). *CALLHOME Egyptian Arabic transcripts LDC97T19* (Web Download). Linguistic Data Consortium. <https://doi.org/10.35111/39gj-5h87>
- Habash, N. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.
- Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018 May 7-12). *Unified guidelines and resources for Arabic dialect orthography*. In the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.
- Habash, N. and Rambow, O. (2006 July 17-21). *MAGEAD: A morphological analyzer and generator for the Arabic dialects*. In the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia.
- Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. (2013 June 9-14). *Morphological Analysis and Disambiguation for Dialectal Arabic*. In the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia.
- Jarrar, M., Habash, N., Alrimawi, F., Akra, D., and Zalmout, N. (2016). Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51, pp. 745-775.
- Keleg, A., Goldwater, S., and Magdy, W. (2023 December 6-10). *ALDi: Quantifying the Arabic level of dialectness of text*. In the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore.



- Keleg, A. and Magdy, W. (2023 December 7). *Arabic dialect identification under scrutiny: Limitations of single-label classification*. In The First Arabic Natural Language Processing Conference (ArabicNLP 2023), Singapore (Hybrid).
- Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (2002). *Egyptian Colloquial Arabic Lexicon* (Web Download). Linguistic Data Consortium. <https://doi.org/10.35111/v678-8018>
- Ko, W.-J., El-Kishky, A., Renduchintala, A., Chaudhary, V., Goyal, N., Guzmán, F., Fung, P., Koehn, P., and Diab, M. (2021 August 1-6). *Adapting high-resource NMT models to translate low-resource related languages without parallel data*. In the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online.
- Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., & Eskander, R. (2014, May). *Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development* (pp. 2348-2354). In the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.
- Masmoudi, A., Ellouze Khmekhem, M., Esteve, Y., Hadrich Belguith, L., and Habash, N. (2014 May). *A corpus and phonetic dictionary for Tunisian Arabic speech recognition*. In the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.
- Matras, Y. et al. (2017). Database of Arabic Dialects. The University of Manchester. <https://www.kratylos.org/~raphael/arabicProject/database-of-arabic-dialects/>.
- McNeil, K. (2018). Tunisian Arabic Corpus: Creating a Written Corpus of an 'Unwritten' Language. *Arabic Corpus Linguistics*, 30.
- Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015 October 30- November 1). *Machine translation experiments on PADIC: A parallel Arabic dialect corpus*. In the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China

- Mishra, P. and Mujadia, V. (2019 August). *Arabic dialect identification for travel and Twitter text*. In the Fourth Arabic Natural Language Processing Workshop, Florence, Italy.
- Mubarak, H. and Darwish, K. (2014 October 25). *Using Twitter to collect a multi-dialectal corpus of Arabic*. In the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar.
- Olsen, H., Touileb, S., and Velldal, E. (2023 December). *Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus*. In The First Arabic Natural Language Processing Conference (ArabicNLP 2023), Singapore (Hybrid).
- Rassem, M. (2023). *Neural machine translation of dialectal-dialectal Arabic*. University of Stuttgart. <https://d-nb.info/133154632X/34>
- Sajjad, H., Abdelali, A., Durrani, N., and Dalvi, F. (2020 December 8-13). *AraBench: Benchmarking dialectal Arabic-English machine translation*. In the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online).
- Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014 May). *YouDACC: the Youtube Dialectal Arabic Comment Corpus*. In the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.
- Salameh, M., Bouamor, H., and Habash, N. (2018 August 20-26). *Fine-Grained Arabic Dialect Identification*. In the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA.
- Salloum, W. and Habash, N. (2012 December). *Elissa: A Dialectal to Standard Arabic Machine Translation System*. In the International Conference on Computational Linguistics (COLING 2012), Mumbai, India.
- Shoufan, A. and Al-Ameri, S. (2015 July 30). *Natural language processing for dialectal Arabic: A survey*. In the Second Workshop on Arabic Natural Language Processing, Beijing, China.



- Slim, A., Melouah, A., Faghihi, U., and Sahib, K. (2022). Improving neural machine translation for low resource algerian dialect by transductive transfer learning strategy. *Arabian Journal for Science and Engineering*, 47(8), pp. 10411–10418.
- Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002 May). *Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world*. In the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands, Spain.
- Zaghouani, W. and Charfi, A. (2018 May). *ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification*. In the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.
- Zaidan, O. and Callison-Burch, C. (2014). Arabic Dialect Identification. *Computational Linguistics*, 40(1), pp. 171-202. https://doi.org/10.1162/COLI_a_00169
- Zaidan, O. F. and Callison-Burch, C. (2011 June 19-24). *Crowdsourcing Translation: Professional Quality from Non-Professionals*. In the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA.
- Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012 June 3-8). *Machine Translation of Arabic Dialects*. In the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montréal, Canada.

نبذة تعريفية عن الباحث أو الباحثين

Author Biodata: Houda Bouamor is an Associate Teaching Professor in the Information Systems department at Carnegie Mellon University in Qatar (CMU-Q). She obtained her Ph.D. in Computational Linguistics from Paris-Sud University in 2012. Her research interests include machine translation, and Arabic natural language processing.	بيانات الباحث: هدى بوعمر، أستاذة تدريس في قسم نظم المعلومات بجامعة كارنيجي ميلون في قطر (CMU-Q). حصلت على درجة الدكتوراه في اللغويات الحاسوبية من جامعة باريس سود في عام ٢٠١٢. تشمل اهتماماتها البحثية الترجمة الآلية والمعالجة الآلية للغة العربية.
---	---

معرف أوركيد (ORCID):  0000-0001-8605-957X

Email: hbouamor@cmu.edu

Author Biodata: Nizar Habash is a Full Professor of Computer Science at New York University Abu Dhabi (NYUAD), and the director of the Computational Approaches to Modeling Language (CAMEL) Lab. He obtained his PhD in Computer Science from the University of Maryland College Park in 2003. His research interests include machine translation, morphological analysis, and computational modeling of Arabic and its dialects.	بيانات الباحث: نزار حبش، أستاذ علوم الحاسوب في جامعة نيويورك أبوظبي (NYUAD)، ومدير مختبر الأساليب الحاسوبية لنمذجة اللغة (CAMEL Lab) مختبر «كامل». حصل على درجة الدكتوراه في علوم الحاسوب من جامعة ميريلاند كوليدج بارك في عام ٢٠٠٣. وتشمل اهتماماته البحثية الترجمة الآلية، التحليل الصرفي، والنمذجة الحاسوبية للغة العربية ولهجاتها.
--	---

معرف أوركيد (ORCID):  0000-0002-1831-3457

Email: nizar.habash@nyu.edu



Author Biodata: Salam Khalifa is a PhD candidate in Linguistics with a focus on Computational Linguistics at Stony Brook University. Her research interests include modeling morphology for different Arabic dialects and Standard Arabic.	بيانات الباحث: سلام خليفة، طالبة دكتوراه في اللغويات مع تركيز على اللغويات الحاسوبية في جامعة ستوني بروك. تشمل اهتماماتها البحثية نمذجة الصرف لمختلف اللهجات العربية والفصحى.
--	---

معرف أوركيد (ORCID):  0000-0003-0049-3637

Email: salam.khalifa@stonybrook.edu

Author Biodata: Fadhl Eryani is a Masters of Arts student at the University of Tübingen focusing on Computational Linguistics. His research interests include natural language processing of Arabic and its dialects.	بيانات الباحث: فضل الإرياني، طالب ماجستير في الآداب في جامعة توبنغن يركز على اللغويات الحاسوبية. تشمل اهتماماته البحثية المعالجة الآلية للعربية ولهجاتها.
---	---

معرف أوركيد (ORCID):  0009-0007-1191-2215

Email: fadhl.al-eryani@student.uni-tuebingen.de

تحديد اللغات الأصلية لمتعلمي اللغة العربية غير الناطقين بها

Native Language Identification for Non-native Arabic Learners

المعرف الرقمي: <https://doi.org/10.60161/2521-001-002-006>

مراد عباس^١ محمد أمين شراقي^٢ محمد مدياني^٣

^١ المجلس الأعلى للغة العربية، الجزائر

^٢ قسم الرياضيات والإعلام الآلي، جامعة أحمد درايعية أدرار، الجزائر

^٣ قسم علوم الحاسوب، كلية تقنية المعلومات، جامعة الإمارات العربية المتحدة، العين

توثيق البحث APA Citation

عباس، مراد، وشراقي، محمد أمين، ومدياني، محمد. (2024). تحديد اللغات الأصلية لمتعلمي اللغة العربية غير الناطقين بها. مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية، (٢).

Submission Date:01/03/2024, Acceptance Date:19/09/2024

تاريخ الإرسال: 01/03/2024 م، تاريخ القبول: ١٩/٠٩/٢٠٢٤ م

Abstract

Native language identification (NLI) is the process of predicting people's mother tongue based on their writing in a language other than their mother tongue. NLI is a recent topic that draws researchers' attention, because of its impact on many other topics, such as author profiling, authorship identification, forensic linguistics, second language acquisition, educational development, marketing, and security. Many studies addressed this topic, especially with regard to learning English language. Despite the existence of studies on other languages, English remains the most fortunate, mainly due to the availability of data. It is worth noting that most of these works used linguistic features and machine learning methods that led to satisfactory results. This research deals with determining the mother tongue of Arabic language learners, where we adopted an experimental approach using three pre-trained models.

Keywords: Arabic, learning, classification, corpus, identification.

الملخص

يتمثل تحديد اللغة الأم في التنبؤ باللغة الأم لشخص من خلال كتابته بلغة غير لغته الأم أثناء تعلمه لها. وتعد موضوعًا حديثًا يشد انتباه الباحثين؛ لما له من أثر على موضوعات أخرى كثيرة، مثل تحديد سمات المؤلفين، والتأكد من نسبة المؤلفات لمؤلفيها، واللسانيات الجنائية، واكتساب اللغة الثانية، والتطور التعليمي، والتسويق والأمن. تناولت العديد من الدراسات هذا الموضوع، خصوصًا فيما يتعلق بتعلم اللغة الإنجليزية؛ فعلى الرغم من وجود دراسات حول لغات أخرى إلا أن الإنجليزية تبقى صاحبة الحظ الأوفر، ويرجع ذلك أساسًا إلى توفر البيانات. ويجدر التنبيه إلى أن أغلب هذه الأعمال استخدمت الخصائص اللغوية وطرائق التعلم الآلي التي أفضت إلى نتائج مرضية في أبحاث سابقة. يتناول هذا البحث تحديد اللغة الأم لمتعلمي اللغة العربية، حيث اعتمدنا المنهج التجريبي لتقييم أداء التعلم العميق في هذه المهمة؛ إذ استخدمنا ثلاثة نماذج من النماذج المدربة مسبقًا على اللغة العربية. **الكلمات المفتاحية:** العربية، التعلم، التصنيف، المدونة، التحديد.



١. المقدمة

إن تحديد اللغة الأم هو عملية آلية تهدف إلى تحديد اللغة الأصلية لشخص ما، بالاعتماد فقط على حديثه المكتوب أو المنطوق بلغة أخرى. وهي عملية تندرج ضمن حقل المعالجة الآلية للغة. وتتم أساساً بالاعتماد على تحليل الخصائص اللغوية، مثل: القاموس المستعمل، والنحو، والأساليب النمطية. ويبني كل ذلك انطلاقاً من فرضية أن اللغة الأم يمتد تأثيرها إلى تعلم اللغات الأخرى (Zampieri et al., 2017). يكتسب تحديد اللغة الأم أهميته من تطبيقاته المتعددة والمتنوعة، مثل التحقق من نسبة التأليف، وتوصيف المؤلفين، وتحليل سماتهم، والتسويق والأمن، وغير ذلك. ويكون التحقق من نسبة التأليف بالفصل في نسبة قطعة نصية مجهولة إلى مؤلف معروف، عن طريق مقارنتها بكتابات أخرى للمؤلف نفسه. ونشير هنا إلى أن هذا الاختبار لا يقوم إلا على مجموعة من المؤلفين، المعروفين والمحدد مسبقاً (Jarvis & Paquot, 2015). أما توصيف المؤلفين وتحليل سماتهم؛ فينطوي على التعرف على الخصائص السوسيو-نفسية للمؤلف. وتندرج تلك الخصائص ضمن صنفين، خصائص اجتماعية كالجنس والعمر وبلد المنشأ والمستوى التعليمي، وخصائص نفسية تتعلق بشخصية المؤلف، مثل الانفتاح والوعي والوفاء (Estival et al, 2007). وفي التسويق، يكون لتحديد اللغة الأم دور في الإعلانات على الشبكة ومستوى التشاركية مع العملاء والمحتوى الموجه (Chen et al, 2017). كما يساعد أمنياً على الكشف والتصدي لجزء كبير من التهديدات الأمنية (Malmasi & Dras, 2017). ترتكز أغلب أدبيات موضوع تحديد اللغة الأم على مزيج من المؤشرات اللغوية، مثل: تصنيف أقسام الكلام (Wong & Dras, 2011) وأنماط الحروف المتعاقبة المكونة للكلمات (Kulmizev et al., 2017)، والأخطاء الإملائية (Gebre et al., 2013) والتراكيب النحوية (Kyle et al., 2015). إذ تشكل هذه المؤشرات اللغوية أساساً تبني عليه طرائق التعلم الآلي (Tetreault et al, 2013)، مثل مصنف بايز الساذج (Naive Bayesian) وآلة المتجهات الداعمة (SVM). تهدف هذه الدراسة إلى إجراء مجموعة من التجارب؛ لاختبار أداء التعلم العميق في تحديد اللغة الأم لمتعلمي اللغة العربية، من وجهين اثنين، التصنيف الثنائي، والتصنيف متعدد الأصناف. وسنعمد في

ذلك على ثلاث نسخ من النموذج اللغوي الذي يعرف باسم (تمثيلات التشفير ثنائية الاتجاه من المحولات) وباللغة الإنجليزية (BERT) التي تعني (Bidirectional Encoder Representations from Transformer) وهي AraBERT V0.2، و CAMeLBERT، و ARBERT V2.

يتناول القسم التالي من هذه الورقة البحثية الأدبيات والدراسات السابقة المتعلقة بالموضوع. بينما توصف المنهجية المتبعة والبيانات المستعملة، في القسم الثالث. يتلو ذلك عرض ومناقشة نتائج التجارب التي قمنا بها، ونختتم هذه الورقة بأفاق البحث المستقبلي.

٢. الدراسات السابقة

على غرار جميع المواضيع التي تندرج تحت المعالجة الآلية للغات الطبيعية، ركزت الأبحاث في مجال تحديد اللغة الأم بصورة أساسية على تعلم اللغة الإنجليزية. كما أنجز في السنوات الأخيرة عدد من الدراسات على لغات أخرى كالصينية والنرويجية والبرتغالية والعربية. في هذه الفقرة، سنركز على الأعمال البحثية التي تخص بالتحديد تعلم اللغة العربية لغةً ثانية. قدم الباحثان (Malmasi & Dras, 2014) أول تطبيق للتعرف على اللغة الأم لمتعلمي اللغة العربية استناداً إلى نهج تصنيف متعدد الأصناف، خاضع للتعلم الموجه من خلال الجمع بين ثلاث خصائص نحوية (قواعد إنتاج CFG، والكلمات الوظيفية، والتركيب النحوية للكلمات؛ لإجراء تصنيف متعدد الأصناف، حيث استخدمت آلات المتجهات الداعمة، وقد حقق النظام دقة بلغت ٤١٪ على مدونة ALC. كما قام (Mechetti et al., 2020)، بدراسة تأثير التصنيف الآلي باستخدام بعض البيانات المستخرجة إحصائياً من مجموعة من المدونات للكشف عن اللغة الأم لمتعلمي اللغة العربية، حيث قاموا بدمج ثلاث خصائص نحوية، هي: أجزاء الكلام، والكلمات الوظيفية، وقواعد إنتاج القواعد النحوية الخالية من السياق. بالنسبة للتصنيف، استخدمت مكتبة آلة المتجهات الداعمة (LIBSVM2) والمدونة اللغوية لمتعلمي اللغة العربية في عمليتي التدريب والتقييم، حيث حقق نموذجهم دقة بلغت ٤٥٪. أما يونسكو (Ionescu, 2015)، فقد قام بدراسة تستند إلى أساليب تعلم الآلة التي تعمل على



مستوى الحرف، باستخدام نواة تعتمد على مسافة الرتبة المحلية (LRD). وكان التدريب والاختبار على مدونة (ALC)، إذ بلغت دقة النظام ٥٠,١٪.

٣. منهج الدراسة وإجراءاتها

أ. المدونة

استخدمنا في تجربتنا المدونة اللغوية لتعليم اللغة العربية (Alfaifi et al., 2014)، التي شكلت بالنسبة لنا وللدراسات التي تهتم باللسانيات الحاسوبية وتعلم اللغة العربية مصدراً قيماً، حيث تضم هذه المدونة مجموعة من النصوص المكتوبة والمنطوقة، أنتجها متعلمو اللغة العربية خلال عامي ٢٠١٢ و ٢٠١٣ بغرض استخدامها في البحث اللغوي وتطوير أدوات وطرق تعليم اللغة. وتشمل ١٥٨٥ نصاً، أي ما يعادل ٢٨٢,٧٣٢ كلمة؛ وكان عدد من ساهم في إنتاج هذه المدونة ٩٤٢ طالباً من ٦٧ جنسية، أما الخلفيات اللغوية التي يمثلونها فبلغ عددها ٦٦ لغة. بالإضافة إلى ذلك، تتضمن المدونة اللغوية لتعليم اللغة العربية ستة وعشرين متغيراً كعناصر بيانات وصفية، ١٢ للمتعلم و ١٤ للنص.

ب. النماذج

الهدف من بحثنا هو دراسة تأثير نماذج BERT العربية المختلفة المدربة مسبقاً فيما يتعلق بتحديد اللغة الأم. ولتحقيق ذلك، قمنا بضبط دقيق لثلاثة نماذج للتعليم العميق هي: AraBERTV0.2، ARBERTV2، CAMELBERT (Inoue et al, 2020; Abdul-Mageed et al, 2020; Antoun et al, 2021) باستخدام المدونة اللغوية لتعليم اللغة العربية (ALC). استُخدم كل نموذج لتنفيذ التصنيف الثنائي والتصنيف متعدد الأصناف. واختيرت هذه النماذج لعدة أسباب؛ من بينها تدريبها المسبق، خصوصاً على المدونات العربية واسعة النطاق، الأمر الذي يساعدها على كشف الفروق الدقيقة والتعقيدات في اللغة العربية. كما أنها أثبتت أداءً تنافسياً في العديد من مهام وتطبيقات المعالجة الآلية للغات الطبيعية، مثل تحليل المشاعر، وتحديد اللغة، والتعرف على أسماء الأعلام، وكشف الأخبار

المزيفة، وما إلى ذلك. وقد صُممت بنياتها وإجراءاتها التدريبية لتحقيق أحدث النتائج في مجموعة من المهام كفهم اللغة العربية، ومهام التصنيف أيضاً.

جدول ١: معاملات الضبط

المعاملات	القيمة
الدورات	05
الحزمة	08
معدل التعلم	4.87 e-5
وزن decay	0.01
Seed	20

جدول ٢: إعدادات النماذج

النموذج	ARBERT V2	CAMeLBERT	AraBERT V0.2
عدد المعاملات	163 M	108 M	136 M
عدد المفردات	100 K	30 K	64 K
المقسم	WordPiece	WordPiece	WordPiece
التسوية	NO	Yes	Yes
تصفية البيانات	No	No	No
حجم البيانات النصية	٢٤٣ GB	107 GB	٧٧ GB
عامل التكرار	/	10	10
تدريب الدورة	42	2	27
الخطوات	8 M	1 M	3 M



- غالبًا ما تأتي هذه النماذج بإصدارات مختلفة (كبيرة/أساسية) أو متغيرات (اللغة العربية الفصحى/اللهجات). كذلك، من المهم الإشارة إلى أنه على الرغم من أن النماذج الثلاثة قد طورت بناءً على البنية نفسها (BERT)، إلا أن هناك عددًا من الخصائص المميزة (الجدول ١). ومن أجل فهم أعمق لتقييم أداء النماذج فيما يتعلق بالتصنيفات المذكورة، نقدم تعريفًا مبسطًا للمفاهيم التالية:
- **الخطوة الشاملة:** تمثل إجمالي عدد خطوات التحسين التي أُخذت أثناء التدريب. وتعد وحدة قياس تزيد قيمتها في كل مرة يقوم فيها المحسن بتحديث معاملات النموذج (الأوزان والتحييزات) باستخدام حزمة واحدة من بيانات التدريب.
 - **خسارة التدريب:** تشير إلى متوسط الخسارة المحسوبة على جميع أمثلة التدريب، خلال دورة تدريب واحدة. وهو يمثل التناقض بين تنبؤات النموذج BERT (على سبيل المثال، تنبؤات نموذج اللغة المقنع أو تنبؤات الجملة التالية) والتسميات الفعلية في مجموعة التدريب.
 - **وقت تشغيل التدريب:** يشير إلى إجمالي الوقت المستغرق لتدريب نموذج BERT على بيانات التدريب.
 - **خسارة التقييم:** تعد خسارة التقييم، والمعروفة أيضًا بخسارة التحقق أو خسارة الاختبار، مقياسًا لأداء نموذج BERT في مجموعة بيانات منفصلة للتحقق/الاختبار.
 - **وقت تشغيل التقييم:** يشير إلى إجمالي الوقت المستغرق لتقييم أداء نموذج BERT في مجموعة بيانات التحقق أو الاختبار. فهو يقيس مدة عملية التقييم، بما في ذلك الوقت المستغرق في معالجة كل حزمة من البيانات، وحساب التنبؤات، والمهام الأخرى المحتملة مثل تسجيل المقاييس أو توليد تقارير التقييم.
 - **الدقة:** تقيس نسبة الحالات (الأمثلة) التي يتنبأ بها النموذج تنبأً صحيحًا (التصنيفات الإيجابية الصحيحة والسلبية) من بين جميع الحالات في مجموعة بيانات التقييم.

- الاستدعاء: المعروف كذلك باسم الحساسية، ويقاس نسبة التنبؤات الإيجابية الصحيحة من بين جميع الحالات الإيجابية الفعلية في مجموعة بيانات التقييم.
- مقياس إف ١: مقياس لدقة النموذج في التصنيف الثنائي عن طريق حساب المتوسط التوافقي للإحكام والاستدعاء. فهو يوازن بين الدقة والاستدعاء، ويوفر مقياساً واحداً يجمعهما.

ج. التصنيف الثنائي (Binary Classification)

التصنيف الثنائي مهمة أساسية في التعلم الآلي، حيث يكون الهدف التنبؤ بفئة واحدة فقط من بين فئتين محتملتين. من أجل ضبط النماذج، قمنا بتقسيم مجموعة البيانات إلى فئتين متميزتين. الفئة الأولى تحمل الرقم (١)، وهي تتعلق بالنصوص التي تكون اللغة الأم لكتابتها هي اللغة العربية. أما بقية النصوص التي تشكل الفئة الثانية فقد سميت (٠). يوضح الجدول ٣.٠ النتائج المتحصل عليها من خلال التجارب.

جدول ٣: نتائج التصنيف الثنائي

مقياس	ARBERTv2	Arabertv0.2	CAMeLBERT
الخطوة الشاملة	545	545	545
خسارة التدريب	0.1034	0.1165	0.0883
وقت تشغيل التدريب	208.5752	202.6301	192.4646
خسارة التقييم	0.2287	0.0945	0.1303
وقت تشغيل التقييم	2.7287	2.6542	2.657
الدقة	%94.37	%95.62	%96.51
الاستدعاء	%95.57	%96.84	%97.14
إف ١	%94.95	%96.21	%96.82



د. التصنيف متعدد الأصناف (Multi-class Classification)

التصنيف المتعدد يتضمن تصنيف البيانات إلى أكثر من مجموعتين أو فئتين، على عكس التصنيف الثنائي، حيث دُرّب النموذج متعدد الأصناف على التنبؤ بفئة واحدة من ثلاث فئات أو أكثر لعنصرٍ ما. في حالتنا، قمنا بإعداد تجربتين، الأولى تعتمد على اللغة الأم، والثانية على المستوى الدراسي.

التصنيف متعدد الأصناف بناءً على اللغة الأم: تحتوي المدونة على عدد معتبر من الخلفيات اللغوية، والتي يبلغ عددها ٦٦ لغة. إلا أنها غير متوازنة، لهذا ارتأينا اعتماد الخلفيات اللغوية المعبر عنها في الجدول ٤، والتي تضم عددًا أكبر من النصوص. كما يوضح الجدول ٥ النتائج المتحصل عليها من خلال التجارب.

جدول ٤: الخلفيات اللغوية المستعملة في التجارب

اللغة	عدد النصوص
العربية	790
الصينية	76
الأوردية	64
الملاوية	48
الفرنسية	44
الفلاينية	36
الإنجليزية	35

جدول ٥: نتائج التصنيف المتعدد بناءً على اللغة الأم (قيم إف ١)

مقياس	ARBERTv2	Arabertv0.2	CAMELBERT
الخطوة الشاملة	545	545	545
خسارة التدريب	0.5796	0.583	0.3727
وقت تشغيل التدريب	202.8736	193.526	188.2814

0.4653	0.5417	0.6075	خسارة التقييم
2.6231	2.6516	2.6367	وقت تشغيل التقييم
%64.74	%33.49	%28.67	الدقة
%61.74	%40.59	%35.91	الاستدعاء
%60.43	%36.45	%30.58	إف ١

التصنيف متعدد الأصناف بناءً على المستوى الدراسي: تصنف المدونة الطلبة المشاركين إلى خمس فئات وفقاً لمستواهم الدراسي هي: المرحلة الثانوية، ودورة اللغة العامة، وبرنامج الدبلوم (دورة لغة متقدمة)، ودرجتي البكالوريوس والماجستير. النتائج موضحة في الجدول ٦.

جدول ٦: نتائج التصنيف المتعدد بناءً على المستوى الدراسي (قيم إف ١)

CAMeLBERT	Arabertv0.2	ARBERTv2	المستوى الدراسي
%48.45	%30.23	%26.87	المرحلة الثانوية
%52.19	%35.77	%29.04	دورة اللغة العامة
%56.78	%39.39	%30.27	برنامج الدبلوم
%63.32	%43.61	%33.91	البكالوريوس
%66.45	%47.89	%36.46	الماجستير
%٥٧,٤٩	%٣٩,٣٧	%٣٠,٩١	متوسط إف ١

٣. تفسير النتائج

حققت النماذج الثلاثة نتائج جيدة فيما يخص التصنيف الثنائي إلا أن تلك المتعلقة بالتصنيف المتعدد تميزت بانخفاض محسوس في نجاعة النماذج، ويرجع ذلك إلى كون نماذج التعلم العميق مثل CAMeLBERT و ARABERT V02 و ARABERT V2 تحدث لها



اختلافات في الأداء عند تطبيقها على مهام التصنيف متعدد الأصناف، نتيجة لعدة أسباب:

- بنية النموذج وبيانات التدريب: هذه النماذج، كونها متغيرات من BERT (تمثيلات التشفير ثنائية الاتجاه من المحولات)، مصممة بشكل أساسي لالتقاط الأنماط المعقدة في النص من خلال تمثيلات عميقة ثنائية الاتجاه. ومع ذلك، تعتمد فعالية هذه النماذج بشكل كبير على جودة وتنوع بيانات التدريب. بالنسبة للغة العربية التي تحتوي على العديد من اللهجات والبنية الصرفية الغنية، قد لا يكون أداء النماذج المدربة على اللغة العربية الفصحى جيداً عند مواجهة الاختلافات المتعلقة باللهجة، ما لم تُضبط خاصةً على مجموعات بيانات متنوعة تتضمن مثل هذه الاختلافات.
- تعقيد التصنيف: يعد التصنيف متعدد الأصناف أكثر تعقيداً من التصنيف الثنائي. ففي التصنيف متعدد الأصناف، يجب على النموذج اختيار الفئة الصحيحة من بين عدة فئات محتملة؛ مما يزيد من فرصة الخطأ، خاصةً إذا كانت بعض الفئات ليست ممثلة بشكل كافٍ في بيانات التدريب. وتضاف إلى التصنيف متعدد الأصناف درجة أخرى من التعقيد؛ نظراً لأن كل عينة يمكن أن تنتمي إلى فئات متعددة في وقت واحد، مما يتطلب من النموذج فهم جميع الفئات القابلة للتطبيق والتنبؤ بها.
- عدم التوازن بين الفئات: في كثير من الأحيان، في الإعدادات متعددة الأصناف، تحتوي بعض الفئات على أمثلة أكثر بكثير من غيرها. يمكن أن يؤدي عدم التوازن هذا إلى نماذج منحازة نحو الفئات الأكثر تكراراً، مما يقلل من فعاليتها الإجمالية عبر الفئات الأقل تكراراً. وبالمثل، في إعدادات الأسماء المتعددة، قد تتكرر بعض الأسماء أكثر من غيرها؛ مما قد يؤدي إلى انحراف تنبؤات النموذج.
- الضبط الدقيق والتكيف: بينما يطبق تدريب نماذج مثل CAMELBERT و ARABERT V2 و ARABERT V2 مسبقاً على مجموعة كبيرة من النماذج، فإن

أدائها في مهام محددة، مثل التصنيف متعدد الأصناف، يمكن أن يتوقف على مدى جودة ضبطها. إذ يعد الضبط الدقيق على مجموعة بيانات خاصة بمهمة محددة أمراً بالغ الأهمية، ولكن بدون بيانات كافية خاصة بمهمة محددة، أو تنظيم مناسب؛ يمكن أن تتكيف النماذج بإفراط مع بيانات التدريب ويكون أداؤها ضعيفاً على البيانات غير المرئية.

٤. الخاتمة

أجرينا في هذه الورقة البحثية دراسة مقارنة لنماذج التعلم العميق في مجال تصنيف النصوص. استخدمنا لهذا الغرض ثلاثة نماذج: CAMeLBER و ARABERT V02 و ARABERT V2 والتي تستند إلى بنية BERT؛ درّبت هذه النماذج الثلاثة واختبرت وفقاً لنمطيّ التصنيف: الثنائي، ومتعدد الأصناف. وقد أظهرت نتائج التجارب التي قمنا بها، باستخدام المدونة اللغوية لتعليمي اللغة العربية (ALC)، أن أداء النماذج الثلاثة ممتاز بالنسبة للتصنيف الثنائي، حيث بلغت قيم إف ١: ٩٤,٩٥٪ و ٩٦,٢١٪ و ٩٦,٨٢٪ للنماذج الثلاثة: ARBERTv2، ARABERT V02 و CAMeLBER. على التوالي. بينما تناقصت قيمة إف ١ بالنسبة للتصنيف متعدد الأصناف، حيث حقق نموذج CAMeLBER أفضل نتيجة في حالتي التصنيف المتعدد بناءً على اللغة الأم، والتصنيف المتعدد بناءً على المستوى الدراسي (٦٠,٤٣٪ / ٥٧,٤٩٪). ولعل السبب الذي جعل أداء CAMeLBER أفضل من ARBERTv2 و ARABERT V02، هو أنه دُرّب على اللهجات العربية أيضاً؛ مما جعله أكثر تمثيلاً والتقائاً للظواهر اللهجية، خاصة تلك المتعلقة بأخطاء الكتابة الصوتية والصرفية والنحوية، والتي تشبه الأخطاء الموجودة في مدونات تعليمي اللغة العربية إلى حد كبير. ونخطط مستقبلاً للعمل على بناء مدونة أكبر حجماً، والعمل على إدخال ميزات لغوية أخرى، مثل التحليل النحوي، ووضع العلامات الصرفية؛ من أجل زيادة نجاعة النماذج.



المراجع الأجنبية:

- Zampieri, M., Ciobanu, A. M., & Dinu, L. P. (2017). Native language identification on text and speech. *arXiv preprint arXiv:1707.07182*.
- Mechti, S., Khoufi, N., & Hadrich Belguith, L. (2020). Improving native language identification model with syntactic features: Case of arabic. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 2* (pp. 202-211). Springer International Publishing.
- Malmasi, S., & Dras, M. (2017). Multilingual native language identification. *Natural Language Engineering*, 23(2), 163-215.
- Kyle, K., Crossley, S. A., & Kim, Y. J. (2015). Native language identification and writing proficiency. *International Journal of Learner Corpus Research*, 1(2), 187-209.
- Kulmizev, A., Blankers, B., Bjerva, J., Nissim, M., van Noord, G., Plank, B., & Wieling, M. (2017, September). The power of character n-grams in native language identification. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications* (pp. 382-389).
- Jarvis, S., & Paquot, M. (2015). Native language identification. *Cambridge handbook of learner corpus research*, 2015.
- Ionescu, R. T. (2015). A fast algorithm for Local Rank Distance: Application to Arabic native language identification. In *Neural Information Processing: 22nd International Conference, ICONIP 2015, Istanbul, Turkey, November 9-12, 2015, Proceedings, Part II* 22 (pp. 390-400). Springer International Publishing.
- Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

- Gebre, B. G., Zampieri, M., Wittenburg, P., & Heskes, T. (2013, June). Improving native language identification with tf-idf weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 216-223).
- Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007, September). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics* (Vol. 263, p. 272).
- Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2020). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. *arXiv preprint arXiv:2101.01785*.
- Wong, S. M. J., & Dras, M. (2011, July). Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1600-1610).
- Tetreault, J., Blanchard, D., & Cahill, A. (2013, June). A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 48-57).
- Alfaifi, A. Y. G., Atwell, E., & Hedaya, I. (2014, June). Arabic learner corpus (ALC) v2: a new written and spoken corpus of Arabic learners. In *Proceedings of Learner Corpus Studies in Asia and the World 2014* (Vol. 2, pp. 77-89). Kobe International Communication Center.
- Malmasi, S., & Dras, M. (2014, October). Arabic native language identification. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 180-186).
- Chen, L., Strapparava, C., & Nastase, V. (2017, July). Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 542-546).



نبذة تعريفية عن الباحث أو الباحثين

بيانات الباحث	AUTHOR BIODATA
مراد عباس، مدير البحث في الإلكترونيك (معالجة اللغات الطبيعية والإشارة) في (مركز البحث في اللغة العربية) ومدير الترجمة والتأليف والنشر (بالمجلس الأعلى للغة العربية) (بالجزائر). حاصل على درجة الدكتوراه في الإلكترونيك من المدرسة الوطنية متعددة التقنيات عام ٢٠٠٨م. تدور اهتماماته البحثية حول تطبيق التقنيات الحديثة في المجالات المتنوعة لحوسبة اللغة العربية ولهجاتها.	Mourad Abbas is research director in electronics (natural language and speech processing) at the Arabic language research center, and director of translation, writing and publishing at the High Council of Arabic Language in Algeria. Dr. Mourad Abbas received his PhD degree in Electronics from the National Polytechnic School in 2008. His research interests include the major issues related to natural language processing, in particular to Arabic language and its dialects.

معرف أوركيذ (ORCID): 0000-0002-8291-0862

Email: m_abbas04@yahoo.fr

بيانات الباحث	AUTHOR BIODATA
محمد أمين شرقي، أستاذ مساعد في (إعلام آلي) في (قسم الرياضيات والإعلام الآلي) (بكلية علوم المادة، الرياضيات والإعلام الآلي) في جامعة أحمد دراية (بالجزائر). حاصل على درجة الماجستير في الإعلام الآلي من المدرسة الوطنية العليا للإعلام الآلي عام ٢٠١٠. تدور اهتماماته البحثية حول المعالجة الآلية للغات الطبيعية.	Mohamed Amine Cheragui is an assistant professor of computer science in the department of mathematics and computer science, college of matter Sciences, mathematics and computer Science, Ahmed Draï University. Mr. Mohamed Amine Cheragui received his Magister degree in 2010 from Ecole Nationale Supérieure d'Informatique. His research interests include natural language processing.

معرف أوركيذ (ORCID): 0000-0003-2760-5547

Email: m_cheragui@univ-adrar.edu.dz

AUTHOR BIODATA	بيانات الباحث
Mohammed Mediani is an Assistant Professor of Computer Science in the Department of Computer science in the college of information technology, United Arab Emirates University. Dr. Mohammed Mediani received his PhD degree in 2017 from Karlsruhe Institute of Technology. His research interests include natural language processing, generative AI, and AI for language learning and teaching.	محمد مدياني، أستاذ مساعد في (علوم الكمبيوتر) في (قسم علوم الكمبيوتر)، (بكلية تكنولوجيا المعلومات) في جامعة الإمارات العربية المتحدة. حاصل على درجة الدكتوراه في عام ٢٠١٧ من معهد كارلسروه للتكنولوجيا. تشمل اهتماماته البحثية معالجة اللغات الطبيعية والذكاء الاصطناعي التوليدي والذكاء الاصطناعي لتعلم اللغة وتعليمها.

معرف أوركيد (ORCID): 0000-0003-1452-0560

Email: medmediani@gmail.com



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