Mahrajan Arabi
Proceeding International Conference
in Arabic Festival (INCAFA) 2026

Digital Renaissance of Al-Dhad
Strategies for Enhancing Arabic Content in Intelligent Systems

Dr. Rabii Chakir
University of Hassan II Casablanca Kingdom of Morocco

Abstract

This paper explores mechanisms to bridge the digital gap facing the Arabic language within generative artificial intelligence environments. The research problem lies in the scarcity of high-quality Arabic datasets required for training Large Language Models (LLMs), compounded by the structural complexity of Arabic, such as the absence of diacritics and complex morphological derivation. The study aims to propose applied strategies to transition from surface statistical processing to deep semantic understanding. The paper concludes that it is necessary to build semantic ontologies and massive tagged corpora through interdisciplinary collaboration between linguists and engineers to reinforce digital linguistic sovereignty.

Keyword: Generative AI; Computational Linguistics; Large Language Models (LLMs); Arabic Digital Content; Digital Sovereignty.

مستخلص البحث

تستكشف هذه الورقة العلمية آليات سد الفجوة الرقمية للغة العربية في بيئات الذكاء الاصطناعي التوليدي. تكمن المشكلة البحثية في ندرة البيانات العربية النوعية الموجهة لتدريب النماذج اللغوية الكبيرة(LLMs) ، فضلًا عن التعقيد البنيوي للغة العربية المتمثل في غياب الحركات التشكيلية والاشتقاق المتشعب. وتهدف الدراسة إلى وضع إستراتيجيات تطبيقية للانتقال من المعالجة الإحصائية البسيطة إلى الفهم الدلالي التام. وتوصلت الورقة إلى ضرورة بناء أنطولوجيات دلالية ومدونات لغوية ضخمة وموسومة (Tagged Corpora)بالتكامل بين اللسانيين والمهندسين لتعزيز السيادة الرقمية اللغوية.

كلمات أساسية: الذكاء الاصطناعي التوليدي، اللسانيات الحاسوبية، النماذج اللغوية الكبيرة(LLMs) ، المحتوى الرقمي العربي، السيادة الرقمية.

Methods (منهج البحث)

المبحث الأول: الإطار العام للدراسة

أولًا: مقدمة الدراسة

شهد العقد الحالي تحولًا جذرياً في مسيرة التطور التكنولوجي، حيث انتقلت البشرية من عصر الحوسبة التقليدية إلى عصر الذكاء الاصطناعي التوليدي والأنظمة الخبيرة، وهو ما فرض واقعاً أصبحت فيه اللغة الطبيعية هي واجهة التعامل الأساسية بين الإنسان

والآلة. وفي ظل هذا المشهد المتسارع، تواجه اللغة العربية تحدياً مصيرياً يتمثل في مدى قدرتها على مواكبة هذه الثورة الرقمية وحجز مكانة رائدة لها في خوارزميات الذكاء الاصطناعي ونماذجه اللغوية الكبيرة.

إن تعزيز مكانة اللغة العربية في الأنظمة الذكية لم يعد ترفاً فكرياً، بل هو ضرورة ملحة لتحقيق السيادة الرقمية وحماية الهوية الثقافية للأمة. ورغم أن اللغة العربية تصنف ضمن أكثر اللغات انتشاراً واستخداماً في العالم، إلا أن المحتوى الرقمي النوعي الموجه لتدريب الآلة لا يزال يعاني من فجوة كمية وكيفية عميقة مقارنة باللغات العالمية الأخرى. هذه الفجوة تؤدي مباشرة إلى ضعف فهم الآلة للخصوصية الصرفية والدلالية للغة العربية.

تأتي هذه الورقة البحثية لترصد واقع "نهضة الضاد الرقمية"، مستعرضة التحديات البنيوية والتقنية التي تقف عائقاً أمام حوسبة العربية العميقة. كما تسعى الورقة إلى رسم خارطة طريق عملية تتجاوز مجرد التشخيص إلى تقديم استراتيجيات تطبيقية لتعزيز المحتوى العربي؛ بدءاً من بناء المدونات اللغوية الضخمة، ووصولًا إلى دعم النماذج اللغوية الوطنية التي تثبت قدرة العربية على القيادة في عصر الآلة.

ثانيًا: مشكلة الدراسة

تكمن مشكلة الدراسة في وجود فجوة رقمية ونوعية عميقة بين مكانة اللغة العربية الحضارية وبين حجم تمثيلها الفعلي في أنظمة الذكاء الاصطناعي الحديثة. ورغم الجهود المبذولة، تواجه حوسبة العربية تحديات جوهرية تتلخص في النقاط التالية:

- ضعف المحتوى التدريبي: قلة المحتوى العربي النوعي الصالح لتدريب نماذج الذكاء الاصطناعي مقارنة بلغات أخرى كالإنجليزية.
- التعقيد البنيوي للعربية: تفرّد العربية بظواهر كغياب الحركات التشكيلية (اللبس الدلالي) والاشتقاق المتشعب.
- التبعية الرقمية: الاعتماد الواسع على نماذج أجنبية مترجمة ينتج مخرجات ركيكة أو مشوهة ثقافياً.

ثالثًا: أهداف الدراسة

تسعى هذه الورقة البحثية إلى تحقيق الأهداف الاستراتيجية التالية:
- تشخيص الواقع الرقمي: تقييم أداء النماذج اللغوية الحالية في التعامل مع الخصوصية الدلالية للغة العربية.
- تحديد الفجوات التقنية: حصر العوائق اللسانية والبرمجية التي تحول دون الفهم التوليدي العميق للغة الضاد.
- اقتراح حلول هندسية: وضع أطر منهجية لبناء أنطولوجيات دلالية ومدونات لغوية ضخمة وموسومة.
- رسم خارطة طريق للسيادة الرقمية: تقديم استراتيجيات للمؤسسات الأكاديمية والتقنية للانتقال بالعربية من مرحلة الاستهلاك إلى مرحلة الإنتاج.

رابعًا: أهمية الدراسة

· الأهمية العلمية: تكمن في تقديم إطار نظري يدمج بين علم اللسانيات العربي التقليدي وبين تقنيات الذكاء الاصطناعي الحديثة لتأسيس فقه لغة رقمي حقيقي.
· الأهمية التطبيقية: تكمن في توفير استراتيجيات قابلة للتنفيذ لبناء مستودعات البيانات التي تحتاجها الشركات والمراكز البحثية لتطوير تطبيقات ذكية ناطقة بالعربية.

خامسًا: حدود الدراسة

· الحدود الموضوعية: تقتصر الدراسة على بحث استراتيجيات تعزيز المحتوى العربي بالفصحى المعاصرة في نماذج الذكاء الاصطناعي التوليدي، مع التركيز المحدود على إشكالية اللهجات المحلية.
· الحدود الزمانية: تم الاعتماد على الأبحاث والدراسات المرجعية الصادرة في الفترة الممتدة حتى عام 2022 لرصد أحدث التطورات التقنية.

***

المبحث الثاني: واقع اللغة العربية في عصر الذكاء الاصطناعي
المطلب الأول: حجم وجودة المحتوى الرقمي الموجه للآلة

يعيش العالم طفرة غير مسبوقة في الاعتماد على البيانات الضخمة لتغذية خوارزميات التعلم العميق. وفي خضم هذا السباق، يبرز تباين حاد بين مكانة اللغة العربية الوجدانية وواقعها الرقمي الفعلي:

· قصور حجم البيانات التدريبية: تشير التقارير المتخصصة إلى أن المحتوى العربي الرقمي المنظم الصالح لتدريب النماذج اللغوية الكبيرة (LLMs) لا يتعدى نسبة ضئيلة جداً مقارنة بالمحتوى التدريبي باللغة الإنجليزية.
· أزمة ندرة البيانات النظيفة: تعاني أغلب المدونات المتاحة آلياً من الخلط بين الفصحى واللهجات العامية الدارجة، مما يخلق عائقاً أمام النماذج الذكية في توليد نصوص ذات سبك لغوي احترافي.

المطلب الثاني: تقييم أداء النماذج اللغوية في فهم السياق والثقافة

يتمحور التحدي الحقيقي في نقل الآلة من مجرد التخمين الإحصائي للألفاظ إلى مستوى الإدراك المعرفي والدلالي:

· تحدي اللبس الدلالي: ينتج عن غياب الحركات التشكيلية في الكتابة العربية لبس دلالي كبير للآلة (مثل الفرق بين عَلَم، عِلْم، عُلِم).

· الخصوصية الثقافية والاجتماعية: تفشل النماذج العالمية المترجمة غالباً في إدراك البنى المجازية والكنايات والتعبيرات الاصطلاحية الأصيلة، منتجة نصوصاً ذات سمة أجنبية واضحة تُخل بالهوية الثقافية للمتلقي.

المطلب الثالث: النماذج اللغوية العربية الوطنية الرائدة

رغم التحديات الكبيرة شهدت الأعوام القليلة الماضية حراكاً تقنياً عربياً رائداً لكسر الهيمنة الأجنبية وبناء نماذج تأسيسية مبنية من الصفر لتناسب البيئة العربية:

· نموذج Jais 2: أطلقته شركة إنسيبتيون ومعهد النماذج التأسيسية بجامعة محمد بن زايد للذكاء الاصطناعي، كنموذج مفتوح الوزن مبني على أكبر قاعدة بيانات عربية المنشأ.

· نموذج عَلاّم (ALlaM): طورته الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) كأول نموذج توليدي وطني يفهم المحادثات العربية بدقة متناهية.

· نموذج Falcon-Arabic: المعتمد على بيئة تدريب ثرية قادرة على فهم الوثائق الطويلة وتنوع اللهجات بطول سياق يصل إلى 20 ألف رمز.

***

المبحث الثالث: التحديات البنيوية والتقنية في حوسبة العربية

المطلب الأول: غياب الحركات التشكيلية ومشكلة اللبس الدلالي

تتميز الكتابة العربية في الغالب بسقوط الحركات الإعرابية (التشكيل)، وهو ما يمثل تحدياً بارزًا أمام الفهم الدلالي للأنظمة الذكية:

· تعدد القراءات الدلالية: يؤدي خلو السلسلة الحرفية من الحركات إلى إمكانية قراءة الكلمة الواحدة بأوجه متعددة ومعانٍ متناقضة (مثل: عَلَم، عِلْم، عُلِم)، وهو ما يعرف باللبس المعجمي أو السياقي.

· عجز المعالجة الإحصائية: تكتفي النماذج اللغوية التقليدية بالتنبؤ الإحصائي للكلمات دون إدراك للمعنى العميق، مما يتطلب إدماج المحللات النحوية والسياقية لرفع كفاءة الآلة في استنباط الحركات تلقائياً لضبط الدلالة.

المطلب الثاني: ثراء الاشتقاق الصرفي كفرصة وعائق

تعد اللغة العربية من فئة اللغات الطبيعية الأكثر تعقيداً والأغنى صرفياً لاعتمادها الكلي على البنية الاشتقاقية:

· توليد الكلمات من الجذور: تتيح الطبيعة الاشتقاقية للعربية توليد مئات الكلمات ذات المعاني المتنوعة من جذر ثلاثي واحد عبر إدخال الأوزان واللواحق.

· عائق النمذجة الحاسوبية: يشكل هذا التنوع الهائل صعوبة بالغة في حوسبة القواعد المعجمية والاشتقاقية، حيث تصبح النماذج المستندة إلى اللغات الهندوأوروبية عاجزة عن استيعاب الخصوصية الصرفية والاشتقاقية للغة الضاد.

المطلب الثالث: ندرة المدونات اللغوية الضخمة والموسومة

تعتمد كفاءة أي نموذج للذكاء الاصطناعي التوليدي بالدرجة الأولى على جودة البيانات والذخائر المستخدمة في تدريبه:
· فجوة المحتوى الموسوم (Tagged Corpora): تعاني الساحة الرقمية العربية من ندرة المستودعات اللغوية المفتوحة التي تم وسم نصوصها صرفياً ودلالياً ونحوياً بدقة.
· ضعف التطوير التشاركي: تتطلب صناعة هذه المدونات الضخمة تعاوناً وثيقاً ومتعدد التخصصات بين اللسانيين والمبرمجين لتغذية الآلة ببيانات تمثل الوجه الحقيقي والتراثي السليم للغة دون تشويه عامي أو ترجمة ركيكة.

***

المبحث الرابع: استراتيجيات تعزيز المحتوى العربي في الأنظمة الذكية
المطلب الأول: رقمنة التراث اللغوي وتطوير أنظمة التعرف الضوئي (OCR)

تمتلك الأمة العربية أضخم تراث علمي وأدبي مخطوط ومطبوع، إلا أن الجزء الأكبر منه لا يزال غائباً عن العالم الرقمي:
· تحويل المخطوطات إلى نصوص حية: تكمن الخطوة الأولى في توليف تقنيات الرؤية الحاسوبية والتعرف الضوئي المتقدمة على الحروف (OCR) لتحويل ملايين الصفحات التراثية إلى نصوص رقمية قابلة للبحث والتدريب الآلي.
· إثراء البيئة التدريبية: رقمنة هذا النتاج تضمن تغذية نماذج الذكاء الاصطناعي بلغة عربية فصحى رفيعة المستوى، مما يرفع من جودة التوليد اللغوي ويبعده عن الركاكة المترجمة.

المطلب الثاني: بناء الأنطولوجيات الدلالية للانتقال من الإحصاء إلى الفهم

للانتقال بنماذج الذكاء الاصطناعي من مرحلة "التنبؤ الإحصائي" السطحي إلى مرحلة "الإدراك المعرفي"، نقترح الآتي:

· تمثيل المعرفة والأنطولوجيا: بناء شجر دلالي وشبكات مفاهيمية تربط الكلمات العربية ببعضها بناءً على علاقات الترادف، والتضاد، والجزء من الكل.
· فك اللبس الدلالي السياقي: يتيح بناء الأنطولوجيات للآلة فهم السياق الثقافي والمجازي التراثي، مما يساعد في حل مشكلة غياب الحركات التشكيلية وتحديد المعنى المراد بدقة متناهية.

المطلب الثالث: دعم مبادرات المصادر المفتوحة والتعاون التشاركي

إن صناعة نماذج لغوية ذكية فائقة الجودة تتطلب جهداً يفوق قدرة المؤسسات الفردية، ويحتاج إلى بيئة تشاركية عربية موسعة:

· بناء المدونات اللغوية الموسومة (Tagged Corpora): إطلاق مشاريع عربية مفتوحة المصدر يشارك فيها اللغويون والتقنيون لوسم النصوص نحوياً وصرفياً ودلالياً.
· توطين السيادة الرقمية: دعم المبادرات التشاركية الإقليمية – على غرار المشاريع الناجحة مثل مشروع Aya العالمي – لضمان بقاء البيانات والسيادة الرقمية في نطاق الأمة العربية.

***

المبحث الخامس: الخاتمة والنتائج والتوصيات
أولًا: خاتمة الورقة

يُخلص من خلال ما سبق إلى أن الثورة الرقمية المعاصرة المتمثلة في الذكاء الاصطناعي التوليدي، تفرض على اللغة العربية تحديات مصيرية لضمان بقائها كلغة عالمية للحضارة الإنسانية. وقد أثبتنا أن الحل لا يكمن في استهلاك الحلول المستوردة المترجمة، بل في بناء أنظمة لغوية أصيلة تنطلق من عمق البنية الصرفية والدلالية الفريدة للغة الضاد.

ثانيًا: نتائج الدراسة

أسفرت الدراسة التحليلية عن مجموعة من النتائج المحورية، أبرزها:
· تفوق النماذج الأصيلة: أثبتت النماذج المدربة على بيانات عربية المنشأ تفوقاً هائلًا في فهم السياق الثقافي مقارنة بالنماذج الأجنبية المترجمة.
· مركزية البيانات النظيفة: تمثل ندرة المدونات اللغوية الموسومة بدقة العائق الأكبر أمام وصول الآلة إلى الفهم الدلالي التام.
· حتمية التكامل المعرفي: لا يمكن لمهندسي الذكاء الاصطناعي بناء نماذج متطورة دون الاعتماد الكلي على القواعد الصرفية والنحوية التي يضعها علماء اللسانيات.

ثالثًا: توصيات الدراسة

بناءً على ما تقدم من مباحث وتحليلات، فإنه يوصى بتبني الاستراتيجيات التالية:
أ. توصيات للمؤسسات الأكاديمية والجامعات
· إطلاق تخصصات بينية تجمع كليات الآداب وكليات الهندسة لإنتاج جيل متخصص في "اللسانيات الحاسوبية".
· رقمنة الأطاريح والبحوث وإلزام الباحثين بنشر إنتاجهم العلمي في مستودعات رقمية مفتوحة المصدر لتغذية الآلة.

ب. توصيات لمراكز الأبحاث وشركات التقنية
· التركيز على دعم مشاريع حوسبة المعاجم لبناء الأنطولوجيات الدلالية التي تتيح للآلة فك اللبس الناتج عن غياب التشكيل.
· تكثيف الجهود في تطوير برمجيات قراءة المخ

طوطات العربية (OCR) لتحويل التراث إلى بيانات حية.

ج. توصيات لصناع القرار والمنظمات العربية

- تمويل نماذج لغوية وطنية ضخمة تضمن حماية الهوية والبيانات العربية من الهيمنة الخارجية.
- تأسيس منصة عربية موحدة لإنتاج وتدقيق المدونات اللغوية الضخمة المفتوحة المصدر.

الفهرس

- الملخص باللغة العربية والإنجليزية
- المبحث الأول: الإطار العام للدراسة
  - أولًا: مقدمة الدراسة
  - ثانيًا: مشكلة الدراسة
  - ثالثًا: أهداف الدراسة
  - رابعًا: أهمية الدراسة
  - خامسًا: حدود الدراسة
- المبحث الثاني: واقع اللغة العربية في عصر الذكاء الاصطناعي
  - المطلب الأول: حجم وجودة المحتوى الرقمي الموجه للآلة
  - المطلب الثاني: تقييم أداء النماذج اللغوية في فهم السياق والثقافة
  - المطلب الثالث: النماذج اللغوية العربية الوطنية الرائدة
- المبحث الثالث: التحديات البنيوية والتقنية في حوسبة العربية
  - المطلب الأول: غياب الحركات التشكيلية ومشكلة اللبس الدلالي
  - المطلب الثاني: ثراء الاشتقاق الصرفي كفرصة وعائق
  - المطلب الثالث: ندرة المدونات اللغوية الضخمة والموسومة
- المبحث الرابع: استراتيجيات تعزيز المحتوى العربي في الأنظمة الذكية
  - المطلب الأول: رقمنة التراث اللغوي وتطوير أنظمة التعرف الضوئي (OCR)
  - المطلب الثاني: بناء الأنطولوجيات الدلالية للانتقال من الإحصاء إلى الفهم
  - المطلب الثالث: دعم مبادرات المصادر المفتوحة والتعاون التشاركي
- المبحث الخامس: الخاتمة والنتائج والتوصيات
  - أولًا: خاتمة الورقة
  - ثانيًا: نتائج الدراسة
  - ثالثًا: توصيات الدراسة
- المراجع والمصادر

References (المراجع)

أولًا: الكتب والمطبوعات المؤسسية

- المعتز بالله السعيد (محرر)، 2022. العربية والذكاء الاصطناعي. منشورات مجمع الملك سلمان العالمي للغة العربية.
- محمود إسماعيل صالح، 2022. الحاسوب واللغة العربية تطبيقات عملية. منشورات مجمع الملك سلمان العالمي للغة العربية.
- سلطان بن ناصر المجيول، 2022. الإحصاء وتقنيات الذكاء الاصطناعي للبيانات اللغوية. منشورات مجمع الملك سلمان العالمي للغة العربية.
- أحمد روبي محمد، 2022. البنك الشجري: بناؤه وتوليفه في إطار تقنيات الذكاء الاصطناعي. منشورات مجمع الملك سلمان العالمي للغة العربية.

ثانيًا: الدراسات والأبحاث المحكمة (المنصات الرقمية)

- الميساوي، خليفة بن الهادي، 2022. "الذكاء الاصطناعي وحوسبة اللغة العربية: الواقع والآفاق". مجلة مدارات في اللغة والأدب، المنصة الجزائرية للمجلات العلمية (ASJP).
- جراح، وهيبة، 2022. "حوسبة اللغة العربية وتطبيقاتها في ضوء معطيات الذكاء الاصطناعي". المنصة الجزائرية للمجلات العلمية (ASJP).
- جامعة محمد بن زايد للذكاء الاصطناعي (MBZUAI)، (2024). أوراق عمل حول إطلاق نموذج جيس 2 ثنائي اللغة.
- الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا)، 2022. التقرير الفني لنموذج عَلاّم (ALlaM) التوليدي للغة العربية.