

السنة الأولى - العدد الأول
شوال ١٤٤٥ هـ - أبريل ٢٠٢٤ م

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



اللُّغَوِيَّاتُ وَاللُّغَةُ الْعَرَبِيَّةُ وَالْمُحَاسَنَاتُ الْإِلَهِيَّةُ وَاللُّغَةُ الْعَرَبِيَّةُ

Journal of Computational Linguistics & Arabic Language Processing

www.ksaa.gov.sa

مجلة علمية مُحَكَّمة تصدر عن مجمع الملك سلمان العالمي للغة العربية





المشرف العام على المجلة

أ.د. عبدالله بن صالح الوشمي

إدارة التحرير

أ.د. عبد الملك سلمان

رئيس التحرير

د. عبد الرحمن العصيمي

مدير التحرير

أ. هاجر القبيشي

أمين المجلة

هيئة التحرير

أ.د. منصور الغامدي

د. سارة الحمود

د. عريب العويشق

د. عبدالعزيز المهيوي

د. أيمن الغامدي

الهيئة الاستشارية

أ.د. إيريك أتويل

أ.د. نزار حبش

أ.د. خالد فؤاد شعلان

أ.د. صلاح راشد الناجم

أ.د. عبد المحسن عبيد الثبيتي

أ.د. محسن عبد الرزاق رشوان

أ.د. مصطفى عرسان جرار

أ.د. هند سليمان الخليفة

◆ قواعد النشر بالمجلة:

سياسات النشر:

تنشر مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية البحوث الأصلية التي لم يسبق نشرها، والمكتوبة باللغة العربية أو الإنجليزية، والدراسات والمختصات العلمية المتخصصة في مجال حوسبة اللغة، كاللغويات 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. أما البحوث المقبولة مبدئياً، فتُرسل إلى اثنين من المحكمين لتقييمها سريعاً، وتبذل هيئة التحرير طاقتها في التواصل مع المحكمين للوصول إلى قرار بشأن كل بحث مقدّم في أسرع وقت ممكن، ويُخطر الباحث بنتيجة التحكيم. وفي حالة قبول الورقة للنشر، فإنّ كافة حقوق النشر تؤوّل للمجلة، ولا يجوز نشرها في أي منفذ نشر آخر ورقياً أو إلكترونياً، دون إذن كتابي من رئيس هيئة تحرير المجلة. وتُرسل نسختان ورقيتان مجانيتان من المجلة إلى المؤلف عند نشر ورقته.



Journal information:

◆ بيانات المجلة:

مجلة اللغويات 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
ردمك النسخة الورقية: ٩٤٧٥-١٦٥٨	ISSN paper version: 1658-9475
ردمك 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:

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

٣-٢- الكلمات المفتاحية.

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

٣-٤- المصادر والمراجع.

٣-٥- 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، وتقسّم 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) للنصّ العربي، والخط (Times New Roman) للنصّ الأجنبي وبحجم 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.
- تُرتّب المصادر والمراجع أبجديًا 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.
- تُرقّم 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.

الصُور والأشكال:

- تُدرج الصور والأشكال بعد الفقرة التي ذكرت 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.
- تُرقّم الصور والأشكال، ويوضع الرقم بين قوسين مسبوقة بكلمة «شكل». مثال: شكل (١).

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

مواصفات كتابة المعادلات:

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

مواصفات كتابة الاقتباس:

- إذا كان النص المقتبس ٣ أسطر أو أقل، فيجب أن يوضع بين علامتي تنصيص «...»، ويليه توثيق النص على النحو التالي: (العائلة، السنة: الصفحة).
- إذا كان النص أكثر من ٣ أسطر، فيجب أن يُميز النص المقتبس بكتابته في منتصف الصفحة، وتُترك من بداية السطر ونهايته هامش 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 - ما لم ترهئة 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:1658-9475)، 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. وللإطلاع 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

نحو بناء بنك المصطلحات العربية

١٥

جمال الزواين - ادريس نملي - عبد الحميد الجيهاد - كريم بوزويح - ابتهاج مصطفى

المدونات ومنزلتها في اللسانيات التطبيقية

٣٦

وليد أحمد العناتي

نظام بلقيس

٧٩

أحمد ميلود خرصي

التعرف على الكلام العربي على أساس المقطع

باستخدام Wav2Vec

٩١

إبراهيم عبد العال - محمد عبد الواحد - مصطفى الشافعي

كلمة المجمع

ضمن جهود المملكة العربية السعودية في رفع مكانة اللغة العربية، ودعم حضورها العالمي والعلمي؛ يأتي تأسيس مجمع الملك سلمان العالمي للغة العربية، ومواصلة السير في دعمها، والحفاظ عليها، ويحظى المجمع بإشراف سمو وزير الثقافة، ورئاسة مجلس أمناء المجمع، الأمير/ بدر بن عبد الله بن فرحان آل سعود - حفظه الله - وينطلق في تنفيذ برامجه وفعالياته الثقافية داخل العالم وخارجه، والعمل الوثيق مع المنظمات الدولية والمؤسسات اللغوية.

وينشط المجمع - ضمن برامجه العالمية - في مسارات عديدة، ولكنه يولي العمق اللغوي والمعرفي أولوية دائمة؛ فبه يشتدُّ عود المعرفة، ويقوى نشاطها، ومن أبرز اهتماماته: العمل في دعم الحراك العلمي اللغوي، ومن ذلك: إنشاء المجلات العلمية النوعية؛ فقد أطلق المجمع ثلاث مجلاتٍ علمية جديدة تساند المجلات والمجالات السابقة في خدمة اللغة العربية، وهي: (مجلة اللغويات الحاسوبية، والمعالجة الآلية للغة العربية)، و(مجلة علوم اللغة العربية)، و(مجلة اختبارات اللغة).

وتأتي أهمية (مجلة اللغويات الحاسوبية، والمعالجة الآلية للغة العربية) لتواكب مشروع المجمع الكبير في تمكين اللغة العربية في مجال: التقنية، والذكاء الاصطناعي، وإثراء البيانات اللغوية، ولندرة البحوث النوعية العلمية باللغة العربية في هذا السياق، وللحاجة المستمرة إلى زيادة البحث والمعرفة في حقل حوسبة اللغات، واللغة العربية على وجه الخصوص؛ حيث إن اللغة العربية

تمثّل - بترائها اللغوي الغني، واستخدامها العالمي الواسع النطاق - تحديات وفرصاً للعلماء، والباحثين، والمتخصصين بهذا المجال.

وستكون انطلاقة (مجلة اللغويات الحاسوبية، والمعالجة الآلية للغة العربية) - بإذن الله - محفزاً إلى حراك علمي وبحثي - في سياق حوسبة اللغة - يخدم اللغة العربية، وهنا أشيد بجهود هيئة التحرير ورئيسها، وأشكر الهيئة الاستشارية على تعاونهم، ويسعدني وزملائي فريق المجلة استقبال المشاركات العلمية من العلماء والباحثين؛ لإثراء هذا الحقل العلمي المتخصص.

الأمين العام

المشرف العام على المجلة

أ.د. عبد الله بن صالح الوشمي

كلمة التحرير

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

عزيزي القارئ / القارئة - عزيزي الباحث / الباحثة :

السلام عليكم ورحمة الله وبركاته

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

تهدف المجلة إلى استقطاب الأبحاث والدراسات العلمية التي تركز على تطبيقات الحوسبة في فهم وتحليل اللُّغة الطّبيعية، واستعراض أحدث التقنيات والأدوات المستخدمة في مجال الحوسبة اللُّغوية، إضافةً إلى استعراض شامل للتطورات الحديثة في التفاعل بين اللُّغة والحوسبة. كما تهدف إلى تسليط الضوء على التحديات والفرص المستقبلية في مجال الحوسبة اللُّغوية. وتسعى إلى تعزيز التفاعل والتبادل العلمي بين الباحثين والمهتمين بمجال الحوسبة اللُّغوية، وتشجيع التطبيقات العمليّة للأبحاث في هذا المجال الهام.

نأمل أن يكون إطلاق العدد الأول من المجلة نقطة تحوّل مهمة في مجال البحث العلمي والتطور التقني في المعالجة الآليّة للُّغات الطّبيعية، وفي مقدمتها اللُّغة العربيّة، وأن تكون المجلة مصدراً ثرياً للمهتمين بالحوسبة اللُّغوية، وأن

تُسهم في دفع عجلة التقدم والابتكار في هذا المجال الحيوي؛ إذ تعتبر الحوسبة اللغوية مجاًلاً متعدد التخصصات يجمع بين علوم الحاسب واللغويات، ويهدف إلى فهم وتحليل اللغة الطبيعية باستخدام التقنيات الحاسوبية.

نحن نفخر بتقديم هذا العدد الأول كخطوة أولى نحو تعزيز التفاعل والتبادل المعرفي في مجال الحوسبة اللغوية، وبداية طموحة لمسيرة مستمرة من الإصدارات المتميزة. كما نأمل أن تكون هذه المجلة منبراً للابتكار والإبداع وإضافة قيمة لمجتمع الحوسبة اللغوية.

تحتوي المجلة في باكورة أعدادها أربعة أبحاث مميزة تتناول جوانب متعددة ومتخصصة في صميم اهتمامات المجلة. البحث الأول يتناول فكرة بناء بنك للمصطلحات العربية من خلال تصريف الجذور العربية والاستفادة من الكلمات المتاحة. أمّا البحث الثاني فإنه يتحدث عن المدونات اللغوية وأهميتها في اللسانيات التطبيقية وأوجه الاستفادة منها. كما يعطي البحث الثالث وصفاً مميزاً عن نظام بلقيس الذي يعالج نواح متعددة من اللغة العربية باستخدام الذكاء الاصطناعي. والبحث الرابع -وهو باللغة الإنجليزية- يقدم طريقة مبتكرة للتعرف على الكلام العربي بدقة عالية.

نودّ أن نشكر جميع الباحثين والمحكمين الذين ساهموا في إثراء هذا العدد بمشاركاتهم القيّمة، ونتمنى منهم ومن غيرهم من قادة هذا العلم المزيد من المشاركة الفاعلة؛ إثراء لهذه المجلة الناشئة، وتوسيعاً لنطاقها، كما نرجو منهم التعريف بها في مكائهم البحث ومنابر العلم من الجامعات والمعامل العلمية وغيرها.

نتمنى لك -عزيزي القارئ، عزيزي القارئة- رحلة ممتعة ومفيدة مع صفحات هذه المجلة، ونتطلع إلى الاستماع إلى آرائك وملاحظاتك حول محتوى المجلة، وسُبل تطويرها. ويسعدنا مشاركتك وتفاعلِكَ المثمر معنا في الأعداد القادمة.

هيئة تحرير المجلة

نحو بناء بنك للمصطلحات العربية Towards Building a Bank of Arabic Terms

جمال الزواين^أ ادريس نملي^ب عبد الحميد الجيهاد^ج كريم بوزوبع^د ابتهاج مصطفى^{هـ}

^أ كلية الآداب والعلوم الإنسانية، جامعة محمد الخامس بالرباط، المغرب
^ب معهد الدراسات الأفريقية والأورومتوسطية والأيبيرية الأمريكية، جامعة محمد الخامس بالرباط، المغرب
^ج معهد الدراسات والأبحاث للتعريب، جامعة محمد الخامس بالرباط، المغرب
^د المدرسة المحمدية للمهندسين، جامعة محمد الخامس بالرباط، المغرب
^{هـ} كلية علوم الحاسب وتكنولوجيا المعلومات، جامعة السودان للعلوم والتكنولوجيا، الخرطوم، السودان

توثيق البحث APA Citation:

الزواين، ج.، نملي، ا.، الجيهاد، ع.، بوزوبع، ك.، مصطفى، ا. (٢٠٢٤). نحو بناء بنك للمصطلحات العربية. مجلة اللغويات الحاسوبية ومعالجة الآلية للغة العربية، العدد ١، ١٥-٣٥.

Submission Date: 04/07/2023
Acceptance Date: 14/08/2023

تاريخ الإرسال: ٢٠٢٣/٠٧/٠٤
تاريخ القبول: ٢٠٢٣/٠٨/١٤

Abstract:

This paper demonstrates the process of generating all possible Arabic terms from trilateral roots. Initially, we systematically generate all the trilateral roots considering the Arabic phonetic system. In the second stage supported linguistic justifications, we compile (363) nominal patterns and (15) verbal patterns. Finally, the generation process involves applying these roots to patterns resulting in a list of (6,784,448) Arabic lexical entries.

الملخص:

نعرض في هذا المقال مراحل توليد جميع المصطلحات الممكنة في اللغة العربية انطلاقاً من الجذور الثلاثية. نشعر بداية في بسط طريقة التوليد الكامل لجميع الجذور الثلاثية اعتماداً على القواعد التي تحكم النظام الصوتي العربي؛ للحصول على الجذور المقبولة صوتياً وغير المستعملة. وفي مرحلة ثانية، نقوم بمحصّر جميع الأوزان الاسمية والأوزان الفعلية. بالنسبة لأوزان الأفعال الثلاثية، فالتفق عليه (١٥) وزناً. أما الأوزان الاسمية، فقد تتبعنا أبنية الأسماء الثلاثية المجردة والمزيدة في المصادر اللغوية قديمها وحديثها، وحاولنا أن نوحّد ما جاء فيها على اختلاف اللغويين في عددها وتصنيفها. وبلغ استقصاؤنا لأوزان الأبنية الاسمية (٣٦٣) وزناً، أدرجنا فيها أوزان الأسماء الجامدة والمشتقة والمصادر. وأخيراً، تمت عملية التوليد عبر إلياس جذور صيغ الأوزان للحصول على حوالي (٦,٧٨٤,٤٤٨) دخلة معجمية.

Keywords:

roots, patterns, derivation, generation, terms.

الكلمات المفتاحية:

جذور، أوزان، توليد، اشتقاق، مصطلحات

المقدمة:

تتطور اللغات وتتوسع باستمرار؛ لمواكبة التطورات التي تحدث في العلوم والتكنولوجيا وكل نواحي الحياة، وقد أثبتت بعض الدراسات أن حوالي ٥٠٪ من الحصيلة اللغوية للدول المتقدمة تكنولوجياً عبارة عن كلمات ومصطلحات مستحدثة (دويدري ٢٠١٠). وتسعى اللغة العربية بكونها واحدة من أعرق اللغات وأكثرها انتشاراً إلى مواكبة هذا التطور والتعبير عن المفاهيم الجديدة بكلماتها الخاصة.

في الواقع، معظم المصطلحات تظهر باللغة الإنجليزية، وعملية نقلها وإدراجها في اللغة العربية ليست عملية بسيطة؛ نظراً لطبيعة الاختلاف بين اللغتين، فاللغة الإنجليزية لغة إصاقية، بينما تتميز العربية بكونها لغة اشتقاقية، وينهج اللغويون العرب طرقاً متعددة لإيجاد نظائر عربية للمصطلحات الجديدة، مثل: استخدام مصطلح موجود مسبقاً وتوسيع معناه للإشارة إلى المصطلح الجديد، مثل: «سيارة»، والتعريب، مثل: «تلفاز»، والنحت، مثل: «برماني»، والاشتقاق، مثل: «حاسوب». وعندما تفضل هذه الأساليب؛ يعتمد مستعمل اللغة إلى نقل المصطلحات الجديدة نقلاً صوتياً، مثل: «هيدروكسي».

في كثير من الأحيان، يتعذر وجود مصطلح قديم يناسب المعنى الجديد، فيلجأ اللغويون إلى استخدام طرق أخرى، مثل: النحت والتعريب. وبالرغم من أن هذه الوسائل تلعب دوراً هاماً في إدراج المصطلحات في اللغة العربية، إلا أنها قد تنتج مصطلحات لا تخضع لقواعد البناء الصوتي في اللغة العربية، فوزن مصطلح «هيدروكسي» مثلاً هو «فعلولي»، وهذا الوزن لا يوجد في اللغة العربية، وزيادة مثل هذه الكلمات قد يؤدي إلى فقدان اللغة لبعض سماتها وخصائصها، ويبقى الخيار الأفضل لتوسيع اللغة هو الاشتقاق وإدراج المصطلحات (Al-Shbiel 2017).

وقد دعا بعض اللغويين إلى استخدام الاشتقاق بصورة أساسية في توسيع اللغة عن طريق توليد جذور جديدة، ومن ثم اشتقاق المصطلحات المناسبة من هذه الجذور (Al-foadi 2018). كما دعا آخرون إلى إنشاء بنك لغوي يحتوي على جميع الجذور العربية الممكنة بجانب جميع المصطلحات التي يمكن توليدها انطلاقاً من هذه الجذور؛ لتصبح متاحة للاستخدام عند الحاجة (عباس ١٩٩٨).

وفي حين لا يتجاوز اليوم مجموع المصطلحات في القواميس العربية (166.000) مصطلح، فقد تداولت بعض المنصات الرقمية رسائل وصورًا ومقالات وزخارف لغوية بمناسبة اليوم العالمي للغة العربية، تفيد بأن عدد الكلمات الممكنة في اللغة العربية يفوق (١٢) مليون كلمة؛ قامت هذه الإحصائية على تقليب الحروف تقليباً رياضياً أعمى؛ فتبين أن الرباعي والخماسي يمثلان النسبة العظمى من الجذور إحصائياً؛ لاحتوائها على عدد أكبر من الحروف مقارنة بالجذور الثلاثية.

ولكن هذه الإحصائية تفتقر للدقة والموضوعية، إذ تشير الدراسات إلى أن نسبة الجذور الثلاثية هي ٧٠٪ من العدد الكلي للجذور المستعملة، مما يؤكد أن اللغة العربية ثلاثية الأصول (Mustapha and Bouzoubaa 2023)، بالإضافة إلى أن هذا التقليب الرياضي الأعمى -الذي أنتج ١٢ مليون كلمة- لم يضع في الاعتبار ما إذا كان الجذر مهماً، أو مستعملاً، وما إذا كان مقبولاً، أو مرفوضاً لغوياً.

يمثل هذا العمل جزءاً من مشاريع المعالجة الآلية للغة، ويشير لاتجاه الحالي في هذا المجال إلى أن النهج الأكثر كفاءة يتعلق باستخدام التعلم العميق؛ نظراً للدرجة العالية من الدقة التي تتيح هذه التقنية

الحصول عليها. بالرغم من ذلك، ونظرًا لطبيعة مشروعا، لا يُنصح بهذا النهج؛ لأنه يعتمد أساسًا على وجود مجموعة بيانات ضخمة، لكن ليس لدينا حاليًا أي مجموعة بيانات من هذا القبيل، ومن ناحية أخرى، يمكن التعلم العميق من توقع أو إنشاء بيانات جديدة بدرجة معينة من الاحتمالية، لا تصل في معظم الأحوال إلى ١٠٠٪، لكن مشروعا يمكن من إنشاء جميع البيانات الممكنة بدرجة عالية من الدقة، وبالتالي، فإن النهج الأنسب لمشروعا يعتمد على توظيف المعرفة اللسانية المتاحة واستغلالها رقميًا.

نخلص بعد هذه المقدمة إلى أن الإشكالية الأساسية تكمن في الصعوبة التي تواجهها اللغة العربية في إيجاد العناصر الاصطلاحية لترجمة المصطلحات الإنجليزية الجديدة، ووفقًا لطبيعة اللغة العربية، نعتقد أن أفضل طريقة هي استخدام الميزة التوليدية والاشتقاقية؛ لسد الحاجة المذكورة، وكما سنفصل لاحقًا في الدراسات السابقة، فقد اتبع بعض الباحثين الاتجاه المقترح، لكنهم لم يأخذوا بعين الاعتبار جميع القيود والضوابط التي تفرضها اللغة العربية، أو أن مشاريعهم لم تبلغ غايتها من حيث الشمول.

في هذا الصدد، يهدف مشروعا إلى تطوير بنك للمصطلحات العربية يحتوي على جميع العناصر الاصطلاحية التي تستطيع اللغة العربية إنتاجها، ويكون بمثابة دليل يعود إليه المتخصصون لاختيار مصطلح يتوافق مع مصطلح إنجليزي جديد؛ للحصول على هذا المورد، هيأنا في عمل سابق معجمًا للجذور الثلاثية الممكنة في اللغة العربية (Mustafa and Bouzoubaa 2023)، وأنشأنا معجمًا للأوزان العربية الممكنة، ثم قمنا بتطبيق الأوزان على الجذور -يطلق عليه أيضًا إلباس الجذور للأوزان- لتوليد جميع المصطلحات العربية الممكنة المنبثقة من الثلاثي، علمًا بأن الثلاثي هو الأكثر شيوعًا في العربية، كما سنوضح لاحقًا.

لأجل تقديم نظرة شاملة عن مراحل بناء بنك للمصطلحات العربية الممكنة، يُنظم بقية المقال على النحو التالي:

يستعرض القسم الموالي بعض المشاريع والدراسات التي سبق إنجازها في هذا الميدان، فيما يتضمن القسم الثالث وصفًا عامًا للمشروع، تناولنا فيه أيضًا مختلف القضايا التي ترتبط بالموارد اللسانية الضرورية لإنشاء بنك للمصطلحات العربية من جذور وأوزان، أما القسم الرابع، فيعرض عملية التوليد لجميع المصطلحات العربية الممكنة المنبثقة من الثلاثي، واختتم المقال بتقديم خلاصة المشروع مع ذكر بعض الآفاق المستقبلية.

١- الدراسات السابقة:

اهتم اللغويون منذ القدم بصناعة المعاجم بحصر المصطلحات الموجودة في اللغة، ودراسة الخصائص التي تميزها عن غيرها من اللغات، والتي يجب وضعها في الاعتبار عند نقل المصطلحات الجديدة للعربية.

وتواجه اللغة العربية اليوم بعض التحديات التي تؤثر على سلامتها، وأهم تحد هو إضافة المصطلحات الجديدة من اللغات الأخرى، وقد تناول العديد من الباحثين واللغويين هذه الظاهرة؛ حيث درس بعضهم الطرق المستخدمة لإدراج المصطلحات الجديدة للغة العربية، والمقارنة بينها؛ لمعرفة أفضلها، بينما دعا آخرون إلى توليد مصطلحات جديدة، كما عمل بعض الباحثين الحاسوبيين في مجال معالجة اللغات الطبيعية على إنشاء بعض الموارد التي تساعد اللغويين على تعريب المصطلحات، ونستعرض بعض هذه الدراسات.

أول عالم عربي لاحظ الفجوة بين الجذور الممكنة والجذور المستعملة هو الخليل بن أحمد (ت ١٧٠هـ) الذي أطلق على هذه الظاهرة اسم «المهمل» (غير المستعمل)، وأوضح أنها ناجمة عن قيود صوتية مرتبطة أساساً بقرب مخارج الأصوات المكونة لهذه الجذور (كشلي ١٩٩٦)، لكن علة الاستئصال النطقي ليست مبرراً كافياً للمجموعات غير المستعملة؛ لأن معظمها ليس فيه صعوبة في النطق، ولذلك حاول كثير من اللغويين بعد الخليل سبر هذه الظاهرة؛ لاكتشاف أسباب التركيبات غير المستعملة، وزاد ابن دريد (ت ٣٢١هـ) على الخليل سبباً آخر هو تنافر صفات الحروف (ابن دريد ١٩٨٧)، وهي علة صوتية محضة، أما ابن جني (ت ٣٩٢هـ) فقد حاول إيجاد مبررات أخرى، فذكر مثلاً عدم الحاجة إلى جميع الجذور لتأدية كل المعاني، وكذلك عدم الانسجام بين الأصوات التي تشكل الجذر والمعنى المقصود (ابن جني ١٩٩٣).

شهدت سبعينيات القرن الماضي أول استخدام للحاسوب في اللسانيات العربية، عندما أجرى موسى علي حلمي (موسى ١٩٧٨) دراسة إحصائية لجذور معجم الصحاح بهدف الكشف عن بعض الظواهر اللغوية، وخصص الفوزان (الفوزان ١٩٨٩) بحثاً لدراسة وتعداد وتلخيص التركيبات المستحيلة من الكتب القديمة، جمع فيه أكثر من ٨٠ قاعدة صوتية، وصحح بعض القواعد التي تناولها الخليل بن أحمد.

وقدم رمضان المجراب (Elmgrab 2011) دراسة عن طُرُق إدراج المصطلحات الجديدة في اللغة العربية، بين من خلالها أن الاشتقاق هو أسلم طريقة استخدمها العلماء القدامى والمحدثون؛ لإدخال مصطلحات جديدة للغة، إذ إنه يحافظ على العلاقة الدلالية بين المصطلح والجذر، كما أنه يقوم بإنشاء المصطلحات وفقاً للأوزان العربية المستخدمة في الاشتقاق؛ لذا فهو يسمح بتوسيع اللغة مع المحافظة على سماتها وخصائصها.

وربما لا يتوفر أحياناً جذر عربي يوافق المصطلح الجديد، فيُعتمد إلى نقل هذا المصطلح حرفياً (نسخ الكلمة الأجنبية بالحروف العربية)، في هذه الحالة، دعا بعض اللغويين إلى توسيع الاشتقاق عن طريق استغلال الجذور العربية غير المستعملة (المهملة)، وتطبيق الوزن المناسب عليها؛ لتدل على المصطلح بدلاً عن النقل الحرفي (عباس ١٩٩٨)؛ لذلك كان لا بد من استغلال جميع الجذور العربية الممكنة في توليد المصطلحات الجديدة.

في هذا السياق، أنشأ حجازي (Hegazi 2016) معجماً يحتوي على جميع الجذور الثلاثية، وذلك بتطبيق التقليلات الرياضية على الحروف العربية، فحصل على (٢١,٩٥٢) جذراً، وبعد استبعاد الجذور التي تتألف من ثلاثة حروف متكررة تبقى (٢١,٩٢٤) جذراً. ثم قام بتطبيق عدد من الأوزان على هذه الجذور؛ للحصول على المصطلحات. يؤخذ على هذه الدراسة أنها لم تضع في الاعتبار قيود التأليف الصوتي في العربية، إذ إن هناك بعض الحروف التي لا يمكن أن تجتمع في جذر واحد حسب النظام الصوتي للغة العربية.

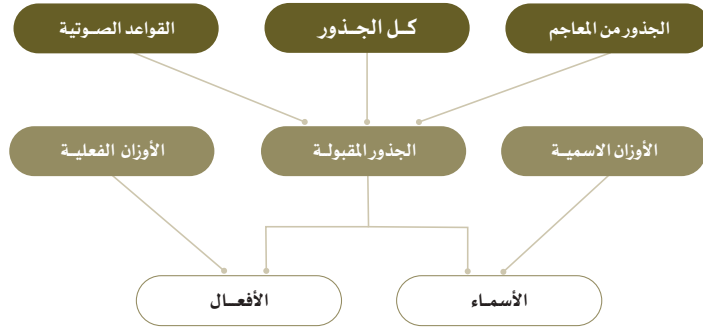
وهكذا، يبدو أن إشكالية وجود قاعدة مصطلحات كاملة؛ للمساعدة في العثور على مصطلحات عربية جديدة، لم يتم تناولها بصورة كاملة في أي من الأعمال السابقة، حيث اكتفى البعض بتسليط الضوء على المشكلة، أو اقتراح بعض الأفكار، في حين لم يتجاوز البعض الآخر وضع قوائم غير مكتملة يمكن أن تساهم في بناء هذا البنك للمصطلحات.

انطلاقاً مما سبق، يتبين ضرورة إنشاء مورد رقمي عربي متكامل، يحتوي على جميع المصطلحات العربية الممكنة، يوضع رهن الباحثين في شتى المجالات، ويمكن الرجوع إليه واستغلاله في إيجاد المفردات وتوسيع اللغة، عبر انتقاء مصطلحات عربية تناسب المفاهيم والمصطلحات المستحدثة في لغات أخرى خاصة الإنجليزية.

١- بنك المصطلحات:

١-٢- وصف المشروع:

يهدف مشروعنا «بنك المصطلحات العربية» إلى توليد جميع المصطلحات الممكنة في اللغة العربية، انطلاقاً من الجذور الثلاثية، مع تقديم أكبر قدر من المعلومات المتعلقة بكل مصطلح جديد، وتتمثل المنهجية في جرد مدخلات معجم الجذور الثلاثية، ثم إلbasها جميع الأوزان الاسمية والفعلية الممكنة، بالإضافة إلى ذلك، يتم رسم كل وزن؛ لتكون المصطلحات المولدة من أسماء وأفعال مصحوبة بمعلومات إضافية تيسر عملية انتقائها؛ للاستخدام في المستقبل (انظر الشكل ١).



شكل ١: منهجية بناء بنك المصطلحات

١-٢- معجم الجذور:

في عمل سابق (Mustafa and Bouzoubaa 2023)، شرعنا في التوليد الكامل لجميع الجذور الثلاثية، وتصنيفها وفقاً لسلامة تكوينها الصوتي، أي كونها محظورة أو ممكنة صوتياً، وكونها مستعملة بالفعل أو مهملة، وكونها ممكنة صوتياً ولم يتم استعمالها بعد. وركزنا على الجذور الثلاثية؛ لأنها تمثل أكثر من ٧٠٪ من الجذور المستخدمة حتى اليوم، والموجودة في القواميس المعتمدة في المشروع (الوسيط، الصحاح، لسان العرب، تاج العروس، المعاصر).

جدول ١: قواعد النظام الصوتي العربي

```
<Rules>
  <Rules_Category cat_id="0" value="composed of identical letters">
    <Rule id="0" letter1="أ" letter2="أ" letter3="أ"></Rule>
    <Rule id="1" letter1="ب" letter2="ب" letter3="ب"></Rule>
    ...
  </Rules_Category>
  <Rules_Category cat_id="1" value="start with two identical letters">
    <Rule id="28" letter1="أ" letter2="أ"></Rule>
    <Rule id="29" letter1="ب" letter2="ب"></Rule>
    ...
  </Rules_Category>
  <Rules_Category cat_id="2" value="can't be together">
    <Rule id="56" letter1="أ" letter2="ع"></Rule>
    <Rule id="57" letter1="ب" letter2="ظ"></Rule>
    ...
  </Rules_Category>
  <Rules_Category cat_id="3" value="can't be followed by">
    <Rule id="101" letter1="أ" letter2="ه"></Rule>
    <Rule id="102" letter1="ب" letter2="ك"></Rule>
    ...
  </Rules_Category>
</Rules>
```

انطلاقاً من الوحدات الأولى (المستوى الأول من الشكل ١)، قمنا بإنشاء ثلاثة موارد، يحتوي المورد الأول على جميع القواعد التي تحكم النظام الصوتي في اللغة العربية (انظر الجدول ١)، وتنقسم هذه القواعد إلى أربع فئات، تحتوي الفئتان الأولىتان على قواعد صوتية تنطبق على جميع الأحرف، وهي أن الجذر يجب ألا يتكون من ثلاثة أحرف متطابقة «composed_of_identical_letters» ويجب ألا يبدأ بحرفين متكررين «start_with_identical_letters» مثل «أأأ» و«ببب» على التوالي، في حين تحتوي الفئتان الأخيرتان: «can't_be_together» و«can't_be_followed_by»، على القواعد التي تمنع التواجد المشترك لبعض الأحرف في الجذر الواحد، على سبيل المثال، لا يمكن للجذر أن يحتوي على الحرفين «ت» و«ظ»، بغض النظر عن ترتيبهما، ولا يمكن أن يتبع الحرف «د» بالحرف «ت» في أي جذر، بينما الحرف «ت» يمكن أن يتبعه الحرف «د» كما في «وتد». لذلك، تنتمي هذه القاعدة إلى فئة «can't_be_followed_by»، حيث يمكن دمج الأحرف في جذر واحد فقط بترتيب معين.

وقد قمنا باستخلاص القواعد الصوتية من خلال استقراءنا لطبيعة الجذور في مصادر لغوية متعددة، وقمنا أيضاً باستخراج الجذور من المعاجم الخمسة السابقة الذكر (انظر الجدول ٢).

جدول ٢: المعاجم العربية

```
<Lexicons>
  <lexicon id="1" name="taj_alarous"></lexicon>
  <lexicon id="2" name="lisan_al_arab"></lexicon>
  <lexicon id="3" name="alsahah"></lexicon>
  <lexicon id="4" name="al_wassit"></lexicon>
  <lexicon id="5" name="al_moassir"></lexicon>
</Lexicons>
```

أخيرًا، يحتوي المورد الثالث على جميع الجذور الممكنة المولدة بطريقة تقليبية رياضية (انظر الجدول ٣).

جدول ٣: الجذور الممكنة في العربية

```
<All_Roots>
  <root id="1" root="أأأ" ></root>
  <root id="2" root="أأب" ></root>
  <root id="3" root="أأت" ></root>
  ...
  <root id="7544" root="ر ع ع" ></root>
  <root id="7545" root="ر ع ش" ></root>
  <root id="7546" root="ر ع ص" ></root>
  <root id="7547" root="ر ع ض" ></root>
  ...
  <root id="21949" root="ي ي ن" ></root>
  <root id="21950" root="ي ي ه" ></root>
  <root id="21951" root="ي ي و" ></root>
  <root id="21952" root="ي ي ي" ></root>
</All_Roots>
```

حسب المنهجية المسطرة في الشكل ١، مكنتنا هذه الموارد الثلاثة من استخلاص مورد الجذور الثلاثية التي يمكن الاعتماد عليها لتوليد المصطلحات الجديدة والمفصلة، حسب الإحصائيات الآتي ذكرها.

يخلص توليف الأحرف الثمانية والعشرين من الأبجدية العربية إلى إنتاج (٢١,٩٥٢) جذرًا مكونًا من ثلاثة أحرف، ويمكن تقسيم التوليفات إلى فئتين رئيسيتين: مقبولة صوتيًا، وعددها (١٣,٤١٠). ومرفوضة صوتيًا، وعددها (٨,٥٤٢)، وكل فئة من هذه الفئات، مقسمة إلى مستعملة وغير مستعملة في المعاجم، وينتج عن هذا أربع فئات: الفئة المستعملة المقبولة صوتيًا وعددها (٨,٠٢٧)، والفئة غير المستعملة المقبولة صوتيًا، وعددها (٥,٣٨٣)، والفئة المستعملة المرفوضة صوتيًا وعددها (٣٩٩)، والفئة غير المستعملة المرفوضة صوتيًا، وعددها (٨,١٤٣)، ويقدم الجدول ٤ إحصائيات لكل فئة من هذه الفئات.

جدول ٤: إحصائيات عن الجذور الثلاثية

عدد التوليفات الرياضية (٢١,٩٥٢)			
عدد الجذور المقبولة صوتيًا (١٣,٤١٠)		عدد الجذور المرفوضة صوتيًا (٨,٥٤٢)	
غير مستعملة (٥,٣٨٣)	مستعملة (٨,٠٢٧)	غير مستعملة (٨,١٤٣)	مستعملة (٣٩٩)
(٨,٤٢٦)			

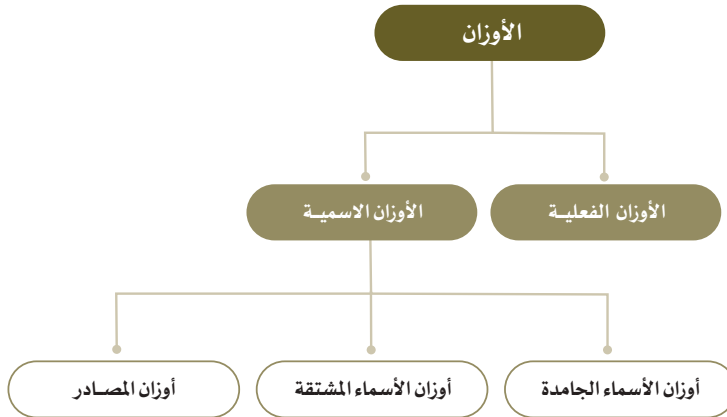
هكذا، فإن اللغة العربية تستخدم $8426 = 399 + 8027$ صيغة جذرية فقط، في حين تظل مجموعة واسعة من الجذور غير مستعملة (٥,٣٨٣)، ويمكن استخدامها؛ لتوسيع اللغة، ويمثل هذا المجموع (١٣,٨٠٩) الجذور المقبولة لغويًا (انظر السطر الثاني من الشكل ١).

وبالتالي، لإغناء اللغة العربية بمصطلحات جديدة، يمكن أولاً: استخدام جميع الجذور البالغ عددها (٨,٤٢٦)، وتطبيق الأوزان التي لم تطبق بعد، على سبيل المثال، بعض المصطلحات التي يمكن اشتقاقها من الجذر "ر ش ع" هي "راشع"، "مرشوع" و"رشعة"، ثانياً: استخدام فئة (٥,٣٨٣) جذراً المتبقية التي لم تستخدم بعد، وتطبيق جميع الأوزان عليها.

١-١- الأوزان:

قسم علماء النحو في اللغة العربية الكلمة إلى ثلاثة أقسام هي: الاسم، والفعل، والحرف. وبالتالي، هناك الأوزان الفعلية

والأوزان الاسمية، على اعتبار أن الحروف ليس لها أوزان. وأوزان الأفعال الثلاثية المتفق عليها خمسة عشرون، سيأتي بيانها أدناه، أما الأوزان الاسمية، فتتنوع بتنوع الفئات التي تنتمي إليها، ويمكن اختزال هذه الفئات في الأسماء الجامدة، والمشتقة، والمصادر (انظر الشكل ٢).



شكل ٢: التصنيف المقترح للأوزان

علاوة على ذلك، ومن أجل إغناء المصطلحات المولدة بمعلومات إضافية، فقد تم توسيمها باستخدام جزء الكلام (Part Of Speech) الخاص بالوزن المستعمل في عملية التوليد الخاصة بالأوزان الاسمية، بالإضافة إلى ذلك، قمنا بتعليق وجود كل وزن عن طريق إضافة مثال من المصطلحات المستعملة في المعاجم، ونعرض فيما يلي جميع التفاصيل اللغوية لهذه العناصر.

١-٣- الأوزان الفعلية:

الفعل هو اللفظ الذي يدل على حدث مقترن بزمانه، وأوزانه تختلف بحسب بنيته (ثلاثي أو رباعي، مجرد أو مزيد) وزمانه (ماضي، مضارع، أمر) وبناؤه (معلوم، مجهول).

للفعل الثلاثي المجرد ستة أبنية، (ثلاثة أوزان في صيغة الماضي؛ لأن فاءه ولامه متحركتان بالفتح دائماً، وتبقى عينه متحركة بالفتح أو الضم أو الكسر). وهي على النحو التالي:

جدول ٥: أبنية الفعل الثلاثي المجرد

فَعَلَ - يَفْعُلُ	فَعَّلَ - يُفَعِّلُ	فَعَّلَ - يُفَعِّلُ	فَعَّلَ - يُفَعِّلُ	فَعَّلَ - يُفَعِّلُ	فَعَّلَ - يُفَعِّلُ
-------------------	---------------------	---------------------	---------------------	---------------------	---------------------

للفعل الثلاثي المزيد بحرف واحد ثلاثة أبنية:

جدول ٦: أبنية الفعل الثلاثي المزيد بحرف

أَفْعَلَ - يُفَعِّلُ	فَاعَلَ - يُفَاعِلُ	فَعَّلَ - يُفَعِّلُ
----------------------	---------------------	---------------------

للفعل الثلاثي المزيد بحرفين خمسة أبنية:

جدول ٧: أبنية الفعل الثلاثي المزيد بحرفين

افْعَلَّ - يَفْعِلُّ	افْتَعَلَ - يَفْتَعِلُ	انْفَعَلَ - يَنْفَعِلُ	تَفَاعَلَ - يَتَفَاعَلُ	تَمَعَّلَ - يَتَمَعَّلُ
----------------------	------------------------	------------------------	-------------------------	-------------------------

للفعل الثلاثي المزيد بثلاثة أحرف أربعة أبنية:

جدول ٨: أبنية الفعل الثلاثي المزيد بثلاثة أحرف

افْعُولَ - يَفْعُولُ	افْعُوعَلَ - يَفْعُوعِلُ	افْعَالًا - يَفْعَالُ	اسْتَفْعَلَ - يَسْتَفْعِلُ
----------------------	--------------------------	-----------------------	----------------------------

قمنا بتخزين هذه الأوزان في معجم يخص الأوزان الفعلية كما يبين الجدول ٩، فمثلاً، وزن «إفْعُولَ» من الأوزان الفعلية ('pos'=4) ومثاله: (ex) هو «إخْشَوْشَن».

جدول ٩: الأوزان الفعلية

```
<Verbal_Patterns>
  <Pattern id='365' vow='فَعَلَ' >
    <case pos='4' ex='ضَرْبٌ' />
  </Pattern>
  <Pattern id='366' vow='فَعَّلَ' >
    <case pos='4' ex='خَيْرٌ' />
  </Pattern>
  ...
  <Pattern id='378' vow='افْعُوعَلَ' >
    <case pos='4' ex='إخْشَوْشَن' />
  </Pattern>
  <Pattern id='379' vow='افْعُولَ' >
    <case pos='4' ex='إقْلُودٌ' />
  </Pattern>
</Verbal_Patterns>
```

٢-٣-٤- الأوزان الاسمية:

يخضع تكوين الأسماء العربية لقوالب صرفية محددة، تتوالى فيها القطع الصامتة والصائتة على أنماط وأشكال معينة لتكوين المصطلحات، ويتغير المعنى المعجمي لكل مصطلح بتغير الوحدات القطعية، وهذا المعنى هو الحمولة الدلالية المشتركة بين جميع المصطلحات المشتقة من الأصل الواحد (بوهاس ١٩٩٣)، في حين تختلف المعاني الصرفية الكبرى باختلاف الأبنية الصرفية، ومنذ بدايات التفكير اللغوي العربي، بحث علماء العربية في مظاهر إطراد الأبنية الاسمية، واجتهدوا في جمعها واستقصائها دون الوصول إلى اتفاق حول عددها (ابن القطاع ١٩٩٩).

وتصنف أبنية الأسماء باعتبار تجردها أو زيادتها، أي باعتبار أصالة الحروف المكونة لها، فما تكون منها من حروف أصلية فقط سمي بالمجرد، وما زيد على حروفه الأصلية ما ليس في أصل وضعه سمي بالزيد، وقد يصل عدد الأحرف الزائدة في البناء الواحد إلى أربعة أحرف.

فالأبنية المجردة تكون إما ثلاثية، أو رباعية، أو خماسية، ولا نجد في العربية بناءً مكوناً من أكثر من خمس صوامت أصلية (ابن القطاع ١٩٩٩):

- الأبنية الثلاثية: وعددها منطقياً اثنا عشر وزناً؛ لأن الصامت الأول قد يأتي على ثلاث صور، هي: الفتح أو الضم أو الكسر، والصامت الثاني قد يأتي على أربع صور، هي: الفتح، أو الضم، أو الكسر، أو السكون، لكن اللسان العربي لم يوظف منها إلا عشرة أوزان هي: فَعْل، فُعْل، فَعِلْ، فَعُلْ، فَعَّلْ، فُوعِلْ، وفُعِّلْ، واختلف في الوزنين المتبقيين، وهما: فُعَلْ وفِعْل.
- الأبنية الرباعية: المتفق عليه فيها خمسة أوزان، هي: فَعْلَلْ، فُعِلِلْ، فُعْلَلْ، فَعْلَلْ، فُعْلَلْ، وأضاف ابن عصفور على هذه الأوزان: فُعْلِيلْ، وأضاف الأخفش: فُعْلَلْ (الخطيب ٢٠٠٣).
- الأبنية الخماسية: المتفق عليه فيها أربعة أوزان، هي: فَعْلَلْ، فُعْلَلْ، فَعْلَلْ، فُعْلَلْ.

وإذا كان من السهل حصر أوزان الأسماء المجردة؛ لقلتها والاتفاق النسيي للعلماء حولها، فإن الإحاطة بأوزان الأسماء المزيدة عملية شاقة ومعقدة؛ لما ذكرناه قبلاً من كثرتها واختلاف اللغويين حول عددها.

ولا شك أن ارتباط بعض التحديدات وغموضها أحياناً يجعل مهمة تصنيف الأوزان عسيرة ومعقدة، لذلك اقتصرنا في هذه المرحلة من البحث بوضع الأوزان الاسمية تحت فئاتها الكبرى: جامدة - مشتقة - مصادر، وجعلنا تحت فئة المشتقات الفروع السبعة المتداولة في كتب الصرف، ووضعنا المصادر في فئة مستقلة: لتفادي النقاش التقليدي حول أسبقية الفعل أو المصدر في الاشتقاق، أي اعتبار المصدر من المشتقات حسب رأي مدرسة الكوفة، أو اعتباره أصلاً للمشتقات كما يقول البصريون، وهو ما يعني أن يكون اسماً مشتقاً أو جامداً، وهو نقاش لا يتناسب وموضوع البحث.

وعلى الرغم من أن الجزء الأعظم من الأسماء المشتقة أخذ من أسماء العاني، فإن هذا الميمنح المتكلم العربي من الاشتقاق من أسماء الذات أيضًا متى دعت الحاجة إلى ذلك، والواقع أن اللغة العربية لم تحظر الاشتقاق من أسماء الذات على الرغم من جمودها، وفي التاريخ من الأمثلة والوقائع ما يشذ عن الحصر، ففي كتاب المخصص لابن سيده باب أسماء: أفعال الضرب المشتقة من أسماء الأعضاء،

ذكر فيه أمثلة كثيرة لكلمات اشتقت من أسماء الذوات الدالة على أعضاء الجسد كالرأس والصدر والأذن، يقول: «رأسه رأسه رأساً أصبت رأسه، دمعته أدمغه ضربت دماغه، أذنته أصبت أذنه» (ابن سيده ١٩٩٦). ومن ذلك أيضاً ما يذكره الرضي في شرحه على الشافية، أن الشيء إذا كثّر بالمكان وكان اسمه جامداً صيغ منه على وزن مفعلة؛ للدلالة على هذه الكثرة، مثل قولنا: مأسدة ومذأبة، أي كثيرة الأسود والذئاب (الاستراباذي ١٩٧٥).

كما أن اللغويين في حديثهم عن الاشتقاق وقواعده، لا يوظفون تعابير قطعية بمنع أو ضرورة الاشتقاق من نوع معين، من ذلك قولهم: إن الاشتقاق «بتفادى الكلمات المجمدة» (قباوة ١٩٧١)، أو قولهم: «يندر الاشتقاق من أسماء الأجناس المحسوسة». يقول شعبان صلاح: «والأصل في المشتق أن يؤخذ من أسماء الأجناس المعنوية، ويندر الاشتقاق من أسماء الأجناس المحسوسة مثل: أورقت الأشجار (من الورق)، وأسبعت الأرض (من السبع)، وعقربت الفتاة صدغها (من العقرب)، وفلفت الطعام (من الفلفل)، ونرجست الدواء (من النرجس)» (شعبان صلاح ٢٠١٠). وجاء في المتع لابن عصفور: «أصعب الاشتقاق وأدقه في أسماء الأجناس؛ لأنها أسماء أول أوقعت على مسمياتها، من غير أن تكون منقولة من شيء» (ابن عصفور ٢٠٢٢).

إجمالاً، فقد تتبعنا أبنية الأسماء الثلاثية بنوعها المجردة والمزيدة في المصادر اللغوية قديمها وحديثها، وفي كتب النحو والصرف، وحاولنا أن نوجد ما جاء فيها على اختلاف اللغويين في عدها وتصنيفها. وقد بلغ استقصاؤنا لأوزان هذه الأبنية (٣٦٣) وزناً، أدرجنا فيها أوزان الأسماء الجامدة، والمشتقة، والمصادر، وهي مفصلة كالآتي.

١-٢-٣-٤ أوزان الأسماء المشتقة:

صنّف اللغويون المشتقات تصنيفاً دلاليّاً، وحصروا هذه الدلالات في سبعة أصناف، هي: اسم الفاعل، مبالغة اسم الفاعل، اسم المفعول، اسم المكان والزمان، اسم الآلة، اسم التفضيل، والصفة المشبهة.

- **اسم الفاعل:** اسم مشتق من الفعل المتصرف المبني للمعلوم، للدلالة على من قام بالفعل، أو من وقع منه الفعل. ويكثر اسم الفاعل في الأفعال المتعدية، ويقل في اللازم منها، ويصاغ اسم الفاعل من الفعل الثلاثي المجرد على وزن (فاعل)، ومن غير الثلاثي المجرد على وزن مضارعه مع إبدال حرف المضارعة ميماً مضمومة، وكسر ما قبل الآخر، وأوزانه هي (مُفْعَلٌ، مُفَاعِلٌ، مُفْعِلٌ، مُتَفَاعِلٌ، مُتَفَاعِلٌ، مُتَفَعِّلٌ، مُفْتَعِّلٌ، مُفْعَلٌ، مُسْتَفْعِلٌ، مُفْعَالٌ، مُفْعُوْعِلٌ، مُفْعُوْلٌ).

- **مبالغة اسم الفاعل:** اسم مشتق يدل على الكثرة والمبالغة في صفة حدوث الفعل، وله صيغة قياسية واحدة، هي (فَعَالٌ)، وتصاغ من الفعل الثلاثي المجرد اللازم أو المتعدي، والصيغ العشر التالية سماعية، وهي (فَعُوْلٌ، فَعِيلٌ، فَعِلٌ، فَعِيلٌ، مَفْعِيلٌ، مَفْعَالٌ، فَعْلَةٌ، فَعَالَةٌ، فَاعُوْلٌ، فَاعِلَةٌ).

- **اسم المفعول:** اسم مشتق من الفعل المتصرف المبني للمجهول؛ للدلالة على معنى مجرد غير لازم، وعلى الذي وقع عليه هذا المعنى، ويشق اسم المفعول من الفعل المتصرف، سواء كان لازماً أم متعدياً، ويصاغ اسم المفعول من الفعل الثلاثي المجرد على وزن (مَفْعُوْلٌ)، ومن غير الثلاثي المجرد على وزن الفعل المضارع المبني للمجهول، مع إبدال حرف المضارعة ميماً مضمومة،

وفتح ما قبل آخره، وأوزانه هي (مُفَعَّل، مُقَاعَل، مُفَعَّل، مُتَفَعَّل، مُتَقَاعَل، مُنْفَعَّل، مُفَعَّل، مُسْتَفَعَّل، مُفَعَّل، مُفَعَّل، مُفَعَّل).

- **اسما الزمان والمكان:** اسمان مشتقان يؤخذان من الفعل؛ للدلالة على زمان الحدث ومكانه، ويصاغ اسما الزمان والمكان من الفعل الثلاثي المجرد على صيغة (مَفْعَل) إذا كان الفعل صحيح اللام مكسور العين في المضارع، أو كان معتل الفاء، وقد يؤنث اسم المكان إذا أريد به البقعة، ويصاغ اسما الزمان والمكان من الفعل الثلاثي المجرد على صيغة (مَفْعَل) إذا كان الفعل الثلاثي معتل اللام، أو كانت عينه في المضارع مضمومة، أو كانت عينه في المضارع مفتوحة -شريطة ألا يكون مثلاً واوياً-، ويصاغ اسما الزمان والمكان من غير الثلاثي المجرد على وزن اسم المفعول أي وزن الفعل المضارع مع إبدال حرف المضارعة ميماً مضمومة، وفتح ما قبل آخره، وأوزانه هي (مُفَعَّل، مُقَاعَل، مُفَعَّل، مُتَفَعَّل، مُتَقَاعَل، مُنْفَعَّل، مُفَعَّل، مُسْتَفَعَّل، مُفَعَّل، مُفَعَّل، مُفَعَّل).

- **اسم الآلة:** اسم مشتق من الفعل الثلاثي المجرد، المتصرف، المتعدي، أو من مصدره؛ للدلالة على الأداة التي تستخدم في إيجاد معنى ذلك المصدر، وقد يشق من الفعل الثلاثي اللازم، ومن غير الثلاثي المجرد، وقد يكون من الأسماء الجامدة. وأوزان اسم الآلة القياسية سبعة، وهي (مَفْعَل، مَفْعَال، مَفْعَلَة، فَعَالَة، فَعَال، فاعِلَة، فاعِل). قال القدماء بقياسية الأوزان الثلاثة الأولى، وأقر مجمع اللغة العربية بالقاهرة بقياسية الأوزان الأربعة الباقية، أما الأوزان السماعية فهي (فاعِلَة، فَعَال، مَفْعَل، مَفْعَلَة، مُفَعَّل).

- **اسم التفضيل:** اسم مشتق من الفعل الثلاثي المجرد؛ للدلالة على أن شيئين اشتركا في صفة، وزاد أحدهما على الآخر فيها. وله وزن واحد، هو (أَفْعَل) ومؤنثه (فُعَلَى).

- **الصفة المشبهة:** اسم مشتق من فعل لازم متصرف أو من مصدره؛ للدلالة على ثبوت صفة لصاحبها ثبوتاً عاماً، ويغلب على الفعل الثلاثي المجرد عند الاشتقاق أن يكون من أحد البابين: فَعِل وفَعَل اللازمين، وتشق الصفة المشبهة من الوزن (فَعِل) على الأوزان التالية: (أَفْعَل مؤنثه فَعَلَاء)، (فَعْلان مؤنثه فَعَلَى وفَعْلانَة)، (فَعِل)، (فَعِيل)، (فَعِل)، (فُعَل)، (فُعَل). وتشق الصفة المشبهة من الوزن (فُعَل) على أربعة أوزان مطردة: (فُعَل)، (فَعَال)، (فَعُول)، (فُعَل)، وهذه الصفة يستوي فيها المذكر والمؤنث مفرداً ومثنى وجمعاً. وتشق الصفة المشبهة من الوزن (فَعَل)، وهو أندر أفعالها، على الأوزان التالية: (فَعِيل)، (فَعِيل)، (أَفْعَل مؤنثه فَعَلَاء).

٢-٣-٢ - أوزان المصادر:

المصدر: اسم يدل على حدث مجرد من الزمن، متضمناً أحرف فعله لفظاً، وينقسم إلى عدة أقسام منها:

- **المصدر الأصلي:** المصدر الذي اشتقت منه الأفعال والمشتقات. وأوزان المصدر الأصلي من الفعل الثلاثي المجرد غير قياسية، وإنما الغالب فيها السماع، وهي كثيرة ومختلفة، والمعول عليه في معرفتها كتب المعاجم واللغة، وقد حصرناها فيما يأتي: (تَفَعَّل، تَفَعَّل، تَفَعَّلَة، عَلَة، عَلَة).

[illegible]

- **المصدر الميمي:** اسم يدل على المصدر الأصلي، غير أنه يبدأ بـ **يـ**م زائدة مفتوحة، ويصاغ المصدر الميمي من الفعل الثلاثي المجرد على وزن (مَفْعَل) في الحالات التالية:
- إذا كان الفعل أجوفاً يائئياً مكسور العين في المضارع.
 - إذا كان الفعل مثلاً واوياً صحيح اللام تسقط فاؤه في المضارع.
- وفيما عدا النوعين السابقين فإن المصدر الميمي يشتق على وزن (مَفْعَل)، ويصاغ من الفعل الثلاثي المزيد على وزن المضارع المبني للمجهول، مع إبدال حرف المضارعة ميماً مضمومة، وفتح ما قبل آخره، وأوزانه هي (مُفَعَّل، مُفاعِل، مُفَعَّل، مُتَفَعَّل، مُتفاعِل، مُنْفَعَل، مُنْفَعَل، مُفَعَّل، مُسْتَفَعَّل، مُفَعَّل، مُفَعَّوَعَل، مُفَعَّوَعَل).
- **مصدر المرة:** مصدر يصاغ: للدلالة على أن الفعل حدث مرة واحدة، ويصاغ من الفعل الثلاثي على وزن (فَعَّلَة)، ويجمع على (فَعَّلَات)، إلا إذا كان عين الفعل معتلأً، فعندها تجمع على (فَعَّلَات)، ويصاغ مصدر المرة من الفعل الثلاثي المزيد بزيادة تاء في آخر المصدر الأصلي، وأوزانه هي (إفْعَالَة، تَفْعِيلَة، مُفاعَلَة، انْفِعالَة، افْعِعالَة، افْعِلالَة، تَفاعَلَة، تَفْعَلَة، اسْتِفْعالَة، افْعِيعالَة، افْعِوالَة، افْعِلالَة).
- **مصدر الهيئة:** مصدر يصاغ: للدلالة على صفة الحدث عند وقوعه، ويصاغ من الفعل الثلاثي المجرد على وزن (فَعْلَة)، ويصاغ مصدر الهيئة من الفعل الثلاثي المزيد بوصف المصدر الأصلي أو إضافته أو الإضافة إليه.

٣-٢-٢- أوزان الأسماء الجامدة:

ذكرنا قبلاً أن الاسم إما جامد، أو مشتق، أو مصدر، والجامد نوعان:

- اسم ذات: وهو الاسم الدال على معنى قائم بنفسه، مثل: رجل.
 - اسم معنى: وهو نوعان، اسم معنى مصدر، وهو ما دل على حدث مجرد من الزمان، مثل: «فهم» و«علم»، واسم معنى غير مصدر، مثل: «فكرة» و«سبب».
- ويدخل تحت الجامد الأسماء المبنية، مثل: الضمائر، وأسماء الإشارة، والأسماء الموصولة، لكنها خارجة لاشتغالنا؛ لأنها لا تقبل التصريف.

وأوزان هذا النوع من الأسماء كثيرة، وقد تشترك أوزانه مع أوزان المشتق أحياناً، مثل: جَبَل على وزن «فَعَلَ»، وكَوَّكَب على وزن «فَوَعَلَ»... إلخ؛ ولذا تصعب الإحاطة بالاسم الجامد في جميع تجلياته الثلاثية والرباعية وغيرها. وقد اقتصرنا في هذا الجزء من البحث على الثلاثي.

جدول ١٠: الأوزان الاسمية

```
<Nominal_Patterns>
  <Pattern id='116' vow='أفاعِل' >
    <case pos='1a' ex='أحامر' />
    <case pos='1b' ex='أحامر' />
  </Pattern> <Pattern id='1' vow='أفَعَلَ' >
    <case pos='2g' ex='أُبْغَلَ' />
    <case pos='2h' ex='أُبْكَم' />
  </Pattern>
  <Pattern id='2' vow='أفعوْلة' >
    <case pos='2f' ex='أزْجوخة' />
  </Pattern>
  <Pattern id='3' vow='إفعال' >
    <case pos='3a' ex='إبْداع' />
  </Pattern>
  <Pattern id='4' vow='إفعالة' >
    <case pos='3c' ex='إسْجادة' />
  </Pattern>
  ...
</Nominal_Patterns>
```

قمنا بتخزين كل هذه الأوزان في معجم يخص الأوزان الاسمية، كما يبين الجدول رقم ١٠، فوزن «أفاعِل» هو من الأوزان الجامدة (pos='1a') ومثاله (ex) هو «أحامر».

٤-٢- توليد المداخل المعجمية:

خاصية التوليد في اللغة العربية هي إنتاج مجموعة من المداخل، عن طريق إجراء عمليات محددة على مجموعة أولية، بمعنى أن آلية التوليد الصرفي تمكننا من توليد المداخل المعجمية المتعلقة بجذر معين، عن طريق وضع الجذر في مجموعة من القوالب التي تسمى الصيغ الصرفية أو الأوزان، وهذا ما يصطلح عليه بنظام الاشتقاق، والاشتقاق هو أداة لغوية توظف الجانب القالب للصرف العربي؛ لإثراء المفردات من خلال تكوين المداخل المعجمية، انطلاقاً من تطبيق الوزن على الجذر.

تتمثل عملية التوليد في تطبيق مجموعة من الأوزان على الجذور؛ لتوليد مداخلها المعجمية. فعلى سبيل المثال، قمنا بتوليد المداخل المعجمية «كَتَبَ، كَتَبْتُ، كَتَبْنَا، كَاتَبَ، كَاتَبْتُ، كَاتَبْنَا، كَاتَبْنَا» عبر تطبيق الأوزان «فَعَلَ، فَعَالٌ، فَاعِلٌ، مَفْعَلٌ، مَفْعَالٌ، فَاعِلٌ» على الجذر «ك ت ب».

يحتوي معجم الجذور على (١٣,٨٠٩) مداخل قابلة للاستعمال، ويتألف معجم الأوزان من (٣٧٨) مدخلاً (انظر الجدول ١١)، وفقاً لذلك، يجب أن نحصل على معجم يتكون من (٥,٢١٩,٨٠٢) مدخل معجمي أو مصطلح، (وهو ناتج ١٣,٨٠٩ * ٣٧٨). ولكن؛ لإمكانية اتصاف الوزن بأكثر من جزء واحد من الكلام (أي أن يشمل فئات كلمية مختلفة)، فقد خلصت هذه العملية إلى توليد (٦,٧٨٤,٤٤٨) مدخلاً (انظر الجدول ١٢). فمثلاً، عند تطبيق الوزن «فَعُول» على الجذر «ك ت ب» نحصل على مدخلين معجميين: «كُتِبَ» الأولى: مبالغة اسم فاعل، والثانية: مصدر أصلي. وتجدر الإشارة إلى أنه يمكن تنزيل جميع المعاجم المطورة من موقع فريقنا على الشبكة^(١).

جدول ١١: معجم الأوزان

```
<Patterns_lexicon>
  <Nominal_Patterns>
    <Pattern id='1' vow='أفعل' >
      <case pos='2g' ex='أُنْجِلْ' />
      <case pos='2h' ex='أُنْكُمْ' />
    </Pattern>
    <Pattern id='2' vow='أفَعُولَة' >
      <case pos='2f' ex='أَزْجُوحة' />
    </Pattern>
    ...
  </Nominal_Patterns>
  <Verbal_Patterns>
    <Pattern id='365' vow='فَعَلَ' >
      <case pos='4' ex='ضَرْب' />
    </Pattern>
    <Pattern id='366' vow='فَعِلَ' >
      <case pos='4' ex='خَيْر' />
    </Pattern>
    ...
  </Verbal_Patterns>
  <POSS>
    <POS id='1a' type='جامد' desc='اسم ذات' />
    <POS id='1b' type='جامد' desc='اسم معنى' />
    ...
    <POS id='4' type='فعل' desc='فعل' />
  </POSS>
</Patterns_lexicon>
```

(1) http://arabic.emi.ac.ma/alelm/?page_id=273/#Lexicon

جدول ١٢: معجم المداخل المعجمية

```
<Lemmas_Lexicon>
  <root id="29" >
    ...
  </root>
  ...
  <root id="16522" >
    <lemma pos="2g" pattern="1" >أكتب</lemma>
    <lemma pos="2f" pattern="2" >أكتوبة</lemma>
    <lemma pos="3a" pattern="6" >استكتاب</lemma>
    <lemma pos="3a" pattern="18" >انكتاب</lemma>
    <lemma pos="3a" pattern="20" >تكتاب</lemma>
    <lemma pos="3a" pattern="23" >تكتاب</lemma>
    <lemma pos="3a" pattern="24" >تكتب</lemma>
    <lemma pos="2a" pattern="31" >كاتب</lemma>
    <lemma pos="2c" pattern="33" >كاتوب</lemma>
    <lemma pos="2c" pattern="36" >كتاب</lemma>
    <lemma pos="2f" pattern="38" >كتاب</lemma>
    <lemma pos="3a" pattern="70" >كتيبة</lemma>
    <lemma pos="2c" pattern="71" >كتوب</lemma>
    <lemma pos="3a" pattern="74" >كتوبة</lemma>
    <lemma pos="2a" pattern="82" >مكتاب</lemma>
    <lemma pos="2b" pattern="83" >مكتب</lemma>
    <lemma pos="2b" pattern="85" >مستكتب</lemma>
    <lemma pos="2b" pattern="87" >مكتاب</lemma>
    <lemma pos="3a" pattern="89" >مكاتبة</lemma>
    <lemma pos="2d" pattern="94" >مكتب</lemma>
    <lemma pos="2d" pattern="103" >مكتبة</lemma>
    <lemma pos="1e" pattern="339" >مكتابة</lemma>
    <lemma pos="4" pattern="365" >كتب</lemma>
    <lemma pos="4" pattern="369" >كاتب</lemma>
    <lemma pos="4" pattern="376" >استكتب</lemma>
    ...
  </root>
  <root id="16523" >
    ...
  </root>
</Lemmas_Lexicon>
```

٣- خاتمة وآفاق مستقبلية:

تناولنا في هذا المقال إمكانية إثراء وتوسيع المعجم العربي بطريقة آلية، وذلك من خلال القيام بمجرد شامل لجميع الجذور المحتملة والممكنة من جهة، واستقصاء لجميع الصيغ الاسمية والفعلية الثلاثية التي يعرفها اللسان العربي من جهة أخرى، ويمكن اختزال عملية التوليد الآلي في إلباس كل جذر من لأئحة جذور جميع الأوزان الثلاثية المثبتة في العربية.

لقد خلص البحث إلى إمكانية توليد ما يقارب (٦,٧٨٤,٤٤٨) مصطلحاً، انطلاقاً من (١٣,٨٠٩) جذرو (٣٧٨) وزناً، ولا شك أن القيود التي تفرضها سلامة تكوين الكلمة في العربية قد قلص كثيراً عدد المفردات النهائية.

إن فكرة اختزال بناء الكلمة العربية في عملية انطباق أو تقاطع وزن وجذريست وليدة اليوم، فقد حاول العديد من اللسانيين المحدثين، خاصة من المستشرقين (انظر مثلاً: Cantineau 1950 و Cohen 1970)، تفسير النسق الصرفي في العربية من خلالها، إلا أن تطبيق هذه العملية آلياً ما يزال في حاجة إلى المزيد من البحث والتجريب، فقد تبين لنا من خلال هذا المشروع أن هذه الإمكانيات ليست معقدة تقنياً ما دامت تفترض وجود لأئحة مسبقة للجذور الممكنة، وأخرى للأوزان المثبتة في العربية؛ ولذا فإننا نعد هذا العمل بمثابة لبنة صغرى في صرح المعجم الاصطلاحي العربي، ندعو من خلاله جميع المهتمين بحقل المعالجة الآلية للغة العربية إلى المساهمة في تطويره وتوسيع مداه؛ ليشمل في وقت قريب الفئات اللفظية الأخرى، رباعية وخماسية.

كما تجدر الإشارة إلى أن التوسيم الدلالي للكلمات المولدة آلياً ما يزال بحاجة إلى المزيد من البحث والدراسة، ولا شك أن الربط الآلي بين الأوزان والحقول الدلالية سيساهم بفاعلية في تصنيف المفردات المولدة، وتيسير عملية بحث المختصين الذين ينتمون لمجال معين عن المصطلح المناسب لمجال تخصصهم.

٤- المراجع العربية:

- الصقلي، ابن القطاع (١٩٩٩). أبنية الأسماء والأفعال والمصادر. (تحقيق أحمد محمد عبد الدايم). مطبعة دار الكتب المصرية. القاهرة.
- ابن سيده، أبو الحسن علي بن إسماعيل (١٩٩٦). المخصص. (مكتب التحقيق بدار إحياء التراث العربي). (ط١). مؤسسة التاريخ العربي. بيروت.
- الإشبيلي، ابن عصفور (٢٠٢٢). الممتع الكبير في علم التصريف. (تحقيق خليل إبراهيم خليل). دار الكتب العلمية. بيروت.
- بديع يعقوب، إميل (١٩٩٣). معجم الأوزان الصرفية. عالم الكتب. بيروت.
- التهانوي، محمد بن علي (١٩٩٦). كشاف اصطلاحات الفنون. (تحقيق علي دحروج). مكتبة لبنان ناشرون.

- بوهاس، جورج (١٩٩٣). تصورات مختلفة عن الصرافة العربية. (ترجمة جمال الزواين). مجلة مصطلحيات. العدد الثاني عشر: ١٣٦-١٥١.
- الأسترايادي، رضي الدين محمد بن الحسن (١٩٧٥). شرح شافية ابن الحاجب. (تحقيق محمد نور الحسن ومحمد الزفزاف ومحمد محي الدين عبد الحميد). دار الكتب العلمية. بيروت.
- صلاح حسين، شعبان (٢٠١٠). تصريف الأسماء في اللغة العربية. دار غريب للطباعة والنشر.
- الخطيب، محمد عبد اللطيف (٢٠٠٣). المستقصى في علم التصريف. دار العروبة للنشر والتوزيع. الكويت.
- قباوة، فخر الدين (١٩٧١). تصريف الأسماء والأفعال. مطبعة دار المعارف. مصر.
- عباس، حسن (١٩٩٨). خصائص الحروف العربية ومعانيها. (ط١). اتحاد الكتاب العرب.
- دويدري، رجاء وحيد (٢٠١٠). المصطلح العلمي في اللغة العربية عمقه التراثي وبعده المعاصر. (ط١). دار الفكر المعاصر.
- كشلي، حكمت (١٩٩٦). كتاب العين للخليل بن أحمد الفراهيدي دراسة وتحليل ونقد. (ط١). دار الكتب العلمية.
- الفراهيدي، الخليل بن أحمد (١٩٩٦). كتاب العين. (تحقيق كشلي حكمت). دار الكتب العلمية. بيروت.
- ابن دريد، أبو بكر محمد بن الحسن (١٩٨٧). جمهرة اللغة. (تحقيق رمزي منير بعلبكي). (ط١). دار العلم.
- ابن جني، أبو الفتح عثمان (١٩٩٣). سر صناعة الإعراب. (تحقيق هنداوي حسن). (ط٢). دار القلم. دمشق.
- موسى، علي حلمي (١٩٧٨). دراسة إحصائية لجذور معجم الصحاح باستخدام الكمبيوتر. (ط١). الهيئة المصرية العامة للكتاب. القاهرة.
- الفوزان، عبد الرحمن إبراهيم (١٩٨٩). الاستيعاب في اللغة العربية الفصحى. دراسة صوتية. جامعة جلاسكو. اسكتلندا.

٥- المراجع الأجنبية:

- Al-Shbiel, A. (2017). Arabization and Its Effect on the Arabic Language. Journal of Language Teaching and Research, 8(3), 469-475.
- Al-foadi, R. (2018). Derivation as the Main Way of Adapting New Terms to Arabic. Modern Journal of Language Teaching Methods (MJLTM), 8(3), 194-99.

- Cantineau, J (1950). La notion de "schème" et son altération dans diverses langues sémitiques. *Semita*, 3, 78-83.
- Cohen, D. (1970). *Etudes de linguistique sémitique et arabe*, Mouton : La Haye, Paris.
- Elmgrab, R. (2011). Methods of Creating and Introducing New Terms in Arabic Contributions from English-Arabic Translation. *IPEDR-International Proceedings of Economics Development and Research*, 26, 491–501.
- Elmgrab, R. (2016). The Creation of Terminology in Arabic. *American International Journal of Contemporary Research*, 6(2), 75–85.
- Hegazi, M. (2016). An Approach for Arabic Root Generating and Lexicon Development. *International Journal of Computer Science and Network (IJCSNS)*, 16(1), 9-15.
- Mustapha, E., Bouzoubaa, K. (2023). A Bi-Gram Approach for an Exhaustive Arabic Trilateral Roots Lexicon. *Languages* 8(83). <https://doi.org/10.3390/languages8010083>.

١- نبذة تعريفية عن الباحثين:

AUTHOR BIODATA:

Jamal EZ-ZOUAINE is an Associate Professor of phonetics and phonology in the Department of Arabic Language and Literature, College of Arts and Humanities at Mohammed V University in Morocco. Prof EZ-ZOUAINE received his PhD degree in General and Arabic Linguistics (2009) from Sidi Mohamed Ben Abdellah University. His research interests include Arabic phonetics and phonology, natural language processing (NLP), and translation.

بيانات الباحث:

جمال الزواين، أستاذ مؤهل في الصوتيات والصواتة في قسم اللغة العربية وآدابها بكلية الآداب والعلوم الإنسانية في جامعة محمد الخامس بالمغرب. حاصل على درجة الدكتوراه في اللسانيات العامة واللسانيات العربية من جامعة سيدي محمد بن عبد الله عام ٢٠٠٩. تدوراهتماماته البحثية حول الدرس الصوتي العربي، والمعالجة الآلية للغات الطبيعية، والترجمة.

<https://orcid.org/0009-0001-8035-4081>

معرف أوركيد (ORCID) : 

Email: jamal.ezzouaine@gmail.com

AUTHOR BIODATA:

Driss NAMLY is an Assistant Professor of Computer Science at Mohammed V University in Rabat, Morocco. Prof NAMLY received his PhD degree in Computer Science from Mohammed V University in 2020. His research interests include Artificial Intelligence, especially Natural Language Processing.

بيانات الباحث :

ادريس نملي، أستاذ مساعد في علوم الحاسوب بجامعة محمد الخامس بالرباط، المغرب. حاصل على درجة الدكتوراه في علوم الحاسوب من جامعة محمد الخامس عام ٢٠٢٠. تدور اهتماماتي البحثية حول الذكاء الاصطناعي وخاصة معالجة اللغات الطبيعية.

<https://orcid.org/0000-0002-6535-5027>

معرف أوركيد (ORCID) : ID

Email: namly_driss@yahoo.fr

AUTHOR BIODATA:

Abdelhamid El Jihad is a Full Professor of computer science in the Department of Arabic Language Processing, College of The Institute of Studies and Research for Arabization, Mohammed V University. Prof El Jihad received his PhD degree in 2017 from Mohammed V University. His research interests are in Natural Language Processing.

بيانات الباحث :

عبد الحميد الجيهاد أستاذ محاضر في علوم الحاسوب في قسم المعالجة الآلية للغة العربية بمعهد الدراسات والأبحاث للتعريب بجامعة محمد الخامس بالرباط بالمغرب. حاصل على درجة الدكتوراه في اللسانيات الحاسوبية من جامعة محمد الخامس عام ٢٠١٧. تدور اهتماماته البحثية حول المعالجة الآلية للغات الطبيعية.

<https://orcid.org/0009-0001-3460-5652>

معرف أوركيد (ORCID) : ID

Email: jihad.hamid@gmail.com

AUTHOR BIODATA:

Karim Bouzoubaa is a Full Professor of computer science in the Department of Computer Science, Mohammadia School of Engineers, Mohammed V University in Rabat. Prof Bouzoubaa received his PhD degree in 1998 from Laval University. His research interests include Artificial Intelligence, Data Science, Natural Language Processing and Computational Linguistics.

بيانات الباحث :

كريم بوزويح، أستاذ محاضر في علوم الحاسوب في قسم علوم الحاسوب بالمدرسة المحمدية للمهندسين بجامعة محمد الخامس بالرباط بالمغرب. حاصل على درجة الدكتوراه في علوم الحاسوب من جامعة لافال عام ١٩٩٨. تدور اهتماماته البحثية حول الذكاء الاصطناعي وعلوم البيانات ومعالجة اللغات الطبيعية واللغويات الحاسوبية.

<https://orcid.org/0000-0002-5387-1980>

معرف أوركيد (ORCID) : ID



Email: karim.bouzoubaa@emi.ac.ma

AUTHOR BIODATA:

Ebtihal Mustapha is a PhD Student of Computer Science, College of Computer Science and Information Technology, Sudan University of Science and Technology. Her research interests include Arabic Natural Language Processing.

بيانات الباحث :

ابتihal مصطفى طالبة دكتوراه في علوم الحاسب، كلية علوم الحاسب وتكنولوجيا المعلومات، جامعة السودان للعلوم والتكنولوجيا. تدور اهتماماتها البحثية حول المعالجة الآلية للغة العربية.

<https://orcid.org/0000-0002-9511-1541>



معرف أوركيڊ (ORCID) :

Email: ebtihal99@hotmail.com

المدونات ومنزلتها في اللسانيات التطبيقية Corpora in Applied linguistics

وليد أحمد العناني

قسم اللغة العربية، كلية العلوم التربوية والآداب / الأونروا، عمان، الأردن

توثيق البحث APA Citation:

العناني، وليد. (٢٠٢٤). المدونات ومنزلتها في اللسانيات التطبيقية. مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية، العدد ١، ٣٦ - ٧٨.

Submission Date: 16/07/2023

Acceptance Date: 10/04/2023

تاريخ الإرسال: ٢٠٢٣/٠٧/١٦

تاريخ القبول: ٢٠٢٣/٠٤/١٠

Abstract:

The current research focuses on how Corpus Linguistics benefits from corpora as an advanced methodological tool, contributing to enhancing goals and maximizing the feasibility of various applications in Applied Linguistics concerning sample size and linguistic materials.

This research comprises three parts. The first part is methodological. The second part provides general introductions to corpora, outlining their objectives; and connections to language planning and policies. Additionally, it addresses the debate regarding whether corpora represent a method or a tool. The third part explores how Corpus Linguistics benefits from corpora and offers ten, comprehensive application examples within Language Teaching, Critical Discourse Analysis, Forensic Linguistics, and Classroom Discourse.

This representative study aims to explore the new aspects of implementing corpora in Applied Linguistics. It seeks to provide fresh insights, applications, ideas, and explanations about corpora, while avoiding repetition. Adhering to the methodological principles of Applied Linguistics and Language Planning is a core focus of this research.

Keywords:

applied linguistics, Processing, Language Planning, Discourse Analysis, Language teaching

المخلص:

يعتني هذا البحث بوجوه انتفاع اللسانيات التطبيقية من المدونات ولسانياتها بوصفها أداة منهجية متطورة تساهم في تعزيز غايات اللسانيات التطبيقية، وتُعظّم جُذوى تطبيقاتها المتنوعة بناءً على حجم العينات والمواد اللغوية المعتمد عليها.

وقد جاء البحث في ثلاثة مباحث:

الأول: تضمن مشكلة البحث وأسئلته وأهدافه، والدراسات السابقة.

الثاني: مقدّمات كَلِّية عَرَفَتْ بلسانيات المدونات وغاياتها ووجودها من تعالقيها بالتخطيط اللغوي ورسم السياسات اللغوية. وتوقّف عند الجدل حول المدونات: أمنهج هي أم أداة؟ أما الثالث فوجوه من انتفاع اللسانيات التطبيقية بالمدونات؛ إذ قدّم عشرة تطبيقات في تعليم اللغة والتحليل النقدي للخطاب والخطاب الصفي واللسانيات الجنائية، وتطبيقاً متكاملاً.

وهذا البحث تمثيليّ يحاول تلمّس وجوه جديدة من تطبيقات المدونات في اللسانيات التطبيقية أو النّظر إليها من زوايا جديدة متنوعة؛ إذ حاول قُدّر الطاقة ألا يكون تكراراً وإنما يقدّم رؤى وتطبيقات وأفكاراً وتفسيرات جديدة. ويحْكُم هذا البحث ضوابط منهجية تتمثل روح اللسانيات التطبيقية وتصدر عن أدواتها التنفيذية « التخطيط اللغوي ».

الكلمات المفتاحية:

لسانيات المدونات، اللسانيات التطبيقية، المعالجة، التخطيط اللغوي، تحليل الخطاب، تعليم اللغة

المقدمة:

١-١- مُشكلة البحث:

تتمثل مشكلة البحث في بيان منزلة المدونات ولسانياتها في اللسانيات التطبيقية، ولا سيما في منطلقاتها الجوهرية التي تعنى بحل المشكلات اللغوية وتشابكاتها الاقتصادية والاجتماعية والسياسية والتربوية؛ فنتائج دراسات لسانيات المدونات لا جدوى منها ولا قيمة لها إن لم تكن وقود اللسانيات التطبيقية التي تمكنها من حل المشكلات والتحول عن الأوضاع اللغوية غير المرغوبة؛ ولهذا سعت الدراسة إلى بيان وجوه من استثمار المدونات في إنجاز أهداف اللسانيات التطبيقية ومهامها.

١-٢- أهداف البحث:

يهدف هذا البحث إلى:

- بيان منزلة لسانيات المدونات في التخطيط اللغوي ورسم السياسات اللغوية الرشيدة.
- بيان منزلة لسانيات المدونات في مجالات اللسانيات التطبيقية المختلفة.
- جسر الفجوة بين المستخلصات النظرية لللسانيات المدونات ووجوه استثمارها في اللسانيات التطبيقية ومجالاتها.
- تأكيد منزلة المدونات ولسانياتها في البحث اللساني؛ للوصول إلى نتائج أدق وأعمق بالنظر إلى حجم عينات الاستعمال اللغوي المدروسة.
- تأكيد مركزية المدونات في دراسة مجمل النشاط الاجتماعي البشري، بناء على اللغة الوظيفية التي يستعملها الناطقون باللغة في المدرسة والمحكمة والغرف الصفيّة والبحث العلمي وغرف التحقيق... إلخ.

١-٣- أسئلة البحث:

يسعى هذا البحث إلى الإجابة عن الأسئلة الآتية:

- ما وجوه افتراق لسانيات المدونات عن البحث اللساني التقليدي؟
- كيف تُسهم المدونات ولسانياتها في التخطيط اللغوي ورسم السياسات اللغوية؟
- ما أهمية نتائج المعالجات اللغوية للمدونات في اللسانيات التطبيقية ومجالاتها؟
- كيف نستفيد من المدونات ومعالجاتها في تنفيذ مهام اللسانيات التطبيقية وحل المشكلات اللغوية والتربوية والاجتماعية؟

١-٤- دراسات سابقة^(١):

ما تزال لسانيات المدونات في نشأتها الأولى، ولا سيما عند محاولة توظيفها في مجالات اللسانيات التطبيقية؛ ولذلك عني البحث بانتقاء دراسات سابقة ذات أبعاد تطبيقية متنوعة، ومنها:

- دراسة الفيضي (٢٠٢٣):

وهي دراسة فنية الطابع؛ إذ ركزت على كيفية وسم المدونات اللغوية وتهيئتها للمعالجة الحاسوبية، بدمج ثلاثة أبعاد متكاملة هي: وسم المفردات، وهيكّل النص، والبيانات الوصفية، مطبقة على اللغة العربية، بما ينتهي إلى تعظيم جدوى المدونات ومعالجتها في البحث اللساني، ولا سيما عندما توجّه نحو مجالات اللسانيات التطبيقية المتنوعة.

- دراسة بدرية العنزي (٢٠٢١):

انطلقت هذه الدراسة من لسانيات المدونات لدراسة التطور الدلالي للمتلازمات واحتمالات تشكل المتلازمات اللغوية بالنظر في حساب تواترها عبر الزمن، وانتهت الدراسة إلى تأكيد أهمية المدونات ولسانياتها في تتبع التطور التاريخي للدلالات ومراقبة تشكل المتلازمات اللغوية بأثر من الزمن وبناء على حساب كمي تُقدّمه تواترات كل متلازمة.

- دراسة العصيمي والثبتي (٢٠٢١):

هدفت الدراسة إلى التعرف على طبيعة الكتابة العلمية المتمثلة في المقترحات البحثية المقدمة إلى قسم علم اللغة التطبيقي في جامعة الإمام، وقد انطلقت من المدونات اللغوية بدراسة الكلمات المنة الأكثر تواتراً في هذه المقترحات البحثية، ودراسة الكلمات المميزة لها كلها، ولكل قسم من أقسامها، إضافة إلى دراسة المتتابعات اللفظية الأشيع، وتعرف المفردات الأكاديمية المستخدمة فيها، وقد أكدت الدراسة أهمية المفردات الأكاديمية في الدلالة على نوع النص وتخصّصه الأكاديمي العام والدقيق بالنظر في نسبة تواتر المفردات بالمقارنة مع حجم النص الكلي، وأكدت الدراسة أهمية المدونات في تدريس اللغة لأغراض خاصة ولا سيما الأغراض الأكاديمية.

ومن الدراسات الأجنبية اخترت:

- دراسة جيرلند ماونتر (٢٠١٤):

وقد ركزت الدراسة على جوانب استثمار المدونات ولسانياتها في التحليل النقدي للخطاب حين يدرس خطابات سياسية أو اجتماعية موجّهة حزبياً أو سياسياً؛ فقد ركّز على كيفية تصوّر المجتمع أو فئات محددة منه لمفردات مثل (عاطل وكادح) والتصورات الاجتماعية التي تثيرهما وتسهم في رسم معالم كل مفردة لدى المجتمع أو فئة منه.

وقد أكدت الدراسة أهمية لسانيات المدونات في التحليل النقدي للخطاب من زاوية هامة وهي تقريب التحليل النقدي للخطاب إلى الموضوعية وإبعاده عن التحليل الذاتي المجرّد من أدلة الواقع اللغوي الحقيقي.

- دراسة ستيفانوفتش (٢٠٢٠):

وهذا كتاب تأسيس في لسانيات المدونات؛ فهو يتناول تعريف المدونات والمعالجة اللغوية وأدواتها وأهمية توظيفها في الدرس اللساني المعاصر، ثم يتناول وجوهاً من تطبيقاتها النحوية

والصرفية والبلاغية المتنوعة، وتبرز قيمة الكتاب في جوانبه التطبيقية الفعلية التي أنجزها ووجوه استثمارها.

ولاشك أن دراستي هذه تتقاطع نسبياً مع الدراسات المَسوقة هنا من حيث العناية بالجوانب النظرية والتأسيسية لللسانيات المدونات؛ على أنها تفترق في عنايتها بمركزية لسانيات المدونات في رسم السياسات اللغوية والتخطيط اللغوي من حيث هي منتهى غاية اللسانيات التطبيقية، وسعيها لحل الإشكالات اللغوية وأثارها السياسية والاجتماعية والاقتصادية، كما ارتادت دراستي هذه فضاءات مُتَخَيِّلَة تُبْصِرُ الباحثين الشباب بمجالات ما تزال بحاجة إلى بحث ودراسة؛ لخدمة العربية ومجتمعاتها، انطلاقاً من شواهد الواقع اللغوي، وبمنهجية مستوعبة هي لسانيات المدونات.

٢- مقدّمات كُليّة

٢-١- التعريف والماهيّة:

المُدَوَّنَاتُ اللغوية جَمْعٌ من النصوص اللغوية المَهَيَّأة للمعالجة الحاسوبية؛ فهذه المدونة قد بُنِيَتْ لتحقيق أهداف محددة^(١)، ورُوِّدَتْ بأدوات تحليل حاسوبية تُصَمِّنُ تحقيق تلك الأهداف، والمُدَوَّنَاتُ مرنة مرونة كبيرة من حيث حجمها وأغراضها وأدوات معالجتها. ولعلَّ تَقَدُّمُ تقنيات الحوسبة والتخزين والاسترجاع يُجَعِّلُ كُلَّ مادة مطبوعة مادة خامة صالحة لبناء مدونة لغوية؛ فكل صحيفة ورقية يمكنها بناء مدوّنتها اللغوية الخاصة، وكلُّ مجلة ثقافية، مثل: مجلة العربي يمكنها بناء مدوّنتها اللغوية الخاصة حتى إذا ما أرادت دراستها وتحليلها كان ذلك أمراً ميسوراً^(٢).

وتنتسب لسانيات المدونات (المتون) في مجالها العام إلى اللسانيات الحاسوبية؛ إذ ترصد مواد لغوية ضخمة وتخزنها وتهيئها للمعالجة الحاسوبية؛ فُصِّدَ استثمارها في مجالات متعددة أغلبها مجالات تربوية وتعليمية تتمثل في صناعة المعاجم، وبناء المواد التعليمية والكتب المنهجية، والمقارنة بين الخطاب المكتوب والخطاب المنطوق، وتعليم اللغة الثانية، وقد تتجاوز ذلك إلى التحليل النقدي للخطاب واللسانيات القضائية واللسانيات الجنائية، والدراسات الاجتماعية والسياسية كدراسة اتجاهات المجتمعات نحو سياسة أو حدث ما.

وأعمدة لسانيات المدونات أربعة:

١. مادة لغوية مجموعة وفق معايير محدّدة تحكمها أهداف المدونة وغاياتها.
٢. التخزين الحاسوبي.
٣. برامج المعالجة الآلية المصمّمة لغايات معينة.
٤. المنتج الذي تحصل عليه من معالجة المادة اللغوية.

ولاشك بأن الأهداف المقصودة من بناء المدونة الخاصة هي التي تتحكم في الأعمدة الأربعة السابقة.

ولما كانت المدونات مدفوعة بأهداف وبواعث مُسبَّقة فلا بد أن تنتهي المعالجات الحاسوبية للمتن اللغوي (النصوص) وبياناتها إلى جوانب لسانية تطبيقية متنوعة؛ إذ يمكن رصد المتلازمات اللفظية أو المفردات والتراكيب الشائعة أو تحليل أخطاء متعلمي اللغة الثانية أو الوقوف على جوانب التحيز في خطابات أحد الأحزاب... إلخ.

٢-٢- المدونات.. منهج أم أداة؟

ثمة جدال بين اللسانيين والمشتغلين بلسانيات المدونات مفاده: هل المدونات منهج لساني جديد أم أداة بحثية مُساندة؟

يُغلبُ على ظني مذهب القائلين: إنها ليست منهجاً لسانياً مستقلاً بذاته؛ وإنما هي أداة مساندة تُيسِّر البحث اللساني وتزيد موثوقية نتائجه، وتُعظِّم جدواه وفائدته؛ ومرجع ذلك إلى أن:

- بناء المدونات وعملها محكوم في أغلبه بفكرة أو نظرية أو تصور لساني معيَّن للظاهرة اللغوية، أو سلوك بعض عناصرها؛ فتكون المعالجة الحاسوبية قاصدة إلى خدمة تلك الفكرة وتعزيزها أو إبطالها، انطلاقاً من واقع لغوي واسع وممتد يتجاوز الاستقراء الناقص البسيط إلى استقراء واسع يصل ملايين الكلمات وآلاف النصوص أحياناً.
- المدونات في أصلها وصفية، ترصد اللغة كما هي في واقع الاستعمال دون تدخُّل؛ ولكن مُشيِّد هذه المدونة يوظفها وفق توجهات منهجية متنوعة؛ فلا تخرج المعالجة الحاسوبية للمادة اللغوية عن مناهج البحث المعتادة (الوصفي والتاريخي) وما يندرج تحتها من تفرعات منهجية متنوعة (التقاليبي والتاريخي المقارن والوصفي الإحصائي والمعياري... إلخ)؛ فهذه المناهج إنما يحددها الباحث المشتغل على المدونة، أو تحددها أهداف المدونة المرسومة من الأول.
- نصوص المدونات ومحتوياتها تعالج بالطريقة ذاتها التي تعالج بها النصوص التقليدية خارج إطار المعالجة الحاسوبية؛ فالباحثون ينهجون أحد منهجين: المنهج الكمي أو المنهج النوعي (الكيفي)، وقد يمزج الباحث بينهما فتكون البيانات الكمية سبيلاً إلى التفسيرات النوعية، أو العكس، حين يلجأ الباحث إلى دعم فرضياته النوعية بتحليل كمي إحصائي.

وعلى ذلك، فإنه يمكننا النظر في مدونة الصحافة العربية بمقاربات منهجية متعددة كما هو الحال في الدرس اللساني بعيناته التقليدية المحدودة؛

١. فقد يرغب باحث في تتبع التطور التاريخي لتركيب نحوي معيَّن في الصحافة العربية منذ مئة عام؛ مثل: أسلوب البناء للمجهول؛ فيكون منهجه تاريخياً لا غير.
٢. وقد يرغب باحث آخر في بناء قائمة بالمفردات الشائعة في الصحافة العربية بنظرة وصفية محضة.
٣. وقد ينظر ثالث في هذه المدونة من منطلق معياري صوابي يحدد الأخطاء اللغوية الشائعة في الصحافة العربية المعاصرة؛ فيكون الرصد إما وصفيّاً محدداً أو تاريخياً تتبعياً.

٤. وقد يُقارن رابع نصوص هذه المدونة بنسختها المترجمة إلى الإنجليزية؛ للوقوف على وجوه التعبير المتباينة في اللغتين.

وهكذا يتبدى لنا أن المدونات ولسانيات المدونات إنما هي تطور في النظر المنهجي يكافئ تطوير أدوات البحث العلمي وانتقالها من النظر البسيط محدود العينات، إلى توظيف الحاسوب واستثمار قدراته في تحصيل نتائج أجود وأضبط وأدق وأسرع؛ ليس إلا!

٣-٢- المدونات والدرس اللغوي:

تعدّ لسانيات المدونات منعطفًا منهجيًا مهمًا في الدراسات اللسانية؛ ذلك أنها:

١. انتصار للمنهج الوصفي؛ فهي نقلت نوعيّة في النظر اللغوي من حيث تأكيده دراسة اللغة في واقع الاستعمال، وكما تتجلى في النصوص الأصلية التي ينتجها الناطقون باللغة.
٢. نقلت الدرس اللساني من النظر في عدد محدود من النصوص إلى معالجة عدد ضخم قد يصل إلى ملايين النصوص؛ بما يُسلم إلى نتائج أدق وأضبط في التحليل اللساني الحديث.
٣. وسّعت مدى الدرس اللساني ليشمل أممية كبرى على المحورين الأفقي والعمودي.

وإذا كانت هذه الفوائد عامّة وتنطبق على جميع مجالات الدرس اللغوي فإن لكل فرع خصوصيّة مستقلة اجتناها من المدونات اللغوية، ونستخذ مثالاً من استثمار المدونات في تنمية كفاية متعلمي اللغة الثانية المعجمية.

الكفاية المعجمية هي المحور الذي تدور عليه كفاية متعلمي اللغة الثانية التواصلية؛ إذ لا يمكن التواصل مع الآخرين دون حصيلة لغوية كبيرة ومتنوعة؛ فهذه الحصيلة يمكن للمتعلم التعبير عن نفسه وبناء جمل وتراكيب مفيدة، وبهذه المفردات يمكنه الاستماع وفهم الآخرين، والتفاعل مع الناطقين باللغة الثانية.

وليس هنا محل إفاضة في تطوير تعليم المفردات وبناء الكفاية المعجمية في اللغة الثانية؛ فهي مبثوثة في كثير من المراجع والدراسات وقوائم المفردات الشائعة... إلخ، على أن دخول المدونات اللغوية مجال تعليم اللغات الأجنبية أحدث تحولات كمّية ونوعيّة ومنهجية مهمة في بناء الكفاية المعجمية، وتمثلاتها في المهارات اللغوية وعناصر اللغة التي تؤلّف الكفاية التواصلية.

لقد أسهمت المدونات اللغوية في^(٤):

١. تقديم شواهد واقعية متنوعة لاستعمال المفردات في سياقات متباينة.
٢. ردّ المفردات إلى أصولها وجذورها.
٣. تصنيف المفردات المتضمنة تصنيفات متنوعة، حسب الشيوخ، والحقل الدلالي، ومجال الاستخدام.
٤. تقديم المعاني المختلفة للمفردة الواحدة (المشترك).

٥. بيان علاقة الكلمة بغيرها من المفردات ولا سيما المتلازمات اللفظية، والوحدات المعجمية المختلفة.
٦. التمييز بين المعاني اللغوية، والمعاني الاصطلاحية.
٧. التمييز بين المعاني الاصطلاحية بناء على مجال الاستعمال (الفاعل في النحو، والفاعل في القانون).
٨. خصائص المفردة النحوية والصرفية والكتابية (الفصائل النحوية والصرفية، تصنيف الأفعال حسب الزمن واللزوم... إلخ، والعائلة المعجمية (الاشتقاقية) للمفردة، وتأثير ذلك في امتداد الجملة الأفقي.
٩. سلوك المفردات في الخطاب بوصفها أداة لبناء الخطاب وتماسكه.
١٠. الفرق بين استعمال المفردات في الخطاب المكتوب والخطاب المنطوق.

ولاشك في أن المدونة تقدم صورة واقعية لسلوك المفردات في الاستعمال اللغوي، وتُبين عن دورها في بنية اللغة واستعمالها، وتقدم خدماتٍ جليّةً متنوعةً في مجال تعليم اللغة الثانية، منها ما تقدمه لمعلم اللغة الثانية؛ فهي تزوّده بمعلومات هائلة تساعد على بناء النشاطات والتدريبات الإثرائية المصاحبة للكتاب المقرر، وتتمثل هذه المساعدة في^(٥):

- بناء قائمة مفردات شائعة تمثل اللغة المتعلمة تمثيلاً حقيقياً وواقعياً.
- بناء قوائم العلاقات الدلالية التي تتجاوز المفردة إلى التراكيب.
- بناء معجم مرحلي يتضمن مفردات مناسبة لمستوى المتعلمين، ويستدرك على مفردات الكتاب مفرداتٍ تتناسب والمحتوى التعليمي للمادة المقررة.
- تحصيل مواقف واقعية وشواهد حيّة لاستعمال المفردات، ونقل الطلبة من مرحلة الاستقبال إلى الإنتاج.
- بيان أثر المفردات في بناء النص وتماسكه.
- توجيه الطلبة نحو الإستراتيجيات المناسبة لتعلم المفردات الجديدة.
- بناء خطة تعليمية متكاملة لكيفية تعليم المفردات من حيث: الأساليب، والإستراتيجيات، والتدريبات المناسبة، وأسس تقديم معنى على آخر.

وبلخص (شُميت) أثر المدونات في الدرس المعجمي بقوله: "إن مشاريع المدونات اللغوية أسهمت في تحوّل الدرس المُعجمي من

الاعتناء بالكلمة وحدها إلى العناية بالعناقيد والتراكيب الدلالية، والتحول من معاملتها بوصفها وحداتٍ منعزلة إلى النظر إليها بوصفها عنصراً مندمجاً في بنية الخطاب/ النص"^(٦).

٤-٢- المدونات والتخطيط اللغوي:

يُمْكِنُ القول إن لسانيات المدونات أداة مهمة في رسم السياسات اللغوية وتعديلها وتنفيذها في صورة التخطيط اللغوي؛ ذلك أن المعالجات الآلية للنصوص على سُرْعَتِها ودَقَّتِها وشمولها تنتهي إلى بيانات لغوية محددة، حركتها أهداف مُسَبَّقة يغلب عليها الطابع التعليمي والتربوي.

ولاشك في أن بناء المدونات وإخضاعها للمعالجات الآلية لن يتوقف عند تحصيل النتائج من حيث الشيوخ أو الندرة أو التطور التاريخي أو تحديد جوانب القصور في جانب ما؛ وإنما يُنظر إلى هذه النتائج بوصفها منطلقاً لمرحلة لاحقة هي جوهر اللسانيات التطبيقية؛ فمعلوم أن عناية اللسانيات التطبيقية تنصرف إلى حل مشكلة قائمة أو معالجتها أو التحول من وضع لغوي غير مرغوب إلى وضعي لغوي آخر.

وهكذا فإن جوهر (الحل) أو (التحول) في اللسانيات التطبيقية يتمثل في توظيف هذه البيانات والنتائج؛ لرسم سياسة لغوية عامة، تُم تنفيذ هذه السياسة العامة بالتخطيط اللغوي المنظم.

ومن الأمثلة على ذلك:

مثال ١:

نفترض أن المملكة العربية السعودية قد بَنَتْ مُدَوَّنَةً لتعليمي العربية غير الناطقين بها ضَمَّتْ آلاف النصوص التي أنتجها الطلبة، وانتهت في فرز أدق إلى تحليل أخطاء المتعلمين تحليلاً دقيقاً على المستويات كلها ثم فرزتها حسب اللغة الأم؛ فهل تتوقف عند تحصيل هذه البيانات؟

إنَّ طبيعة توظيف لسانيات المدونات في اللسانيات التطبيقية تفرض حلاً هذه المشكلات والتحول إلى وضع أَفْضَلَ بتقديم حلول ناجعة لهذه الأخطاء؛ وبناء على ذلك، يمكن للمملكة رسم خطة فرعية تنضوي تحت إطار "التخطيط اللغوي لنشر العربية وتطوير تعليمها" الذي ينبثق أصلاً من سياستها العامة لنشر العربية والإسلام.

ويمكن أن تُنفَّذ هذه الخطة وفق المسارات التالية:

- تحديد الأخطاء الشائعة المشتركة بين المتعلمين، على اختلاف خلفياتهم اللغوية، ومعالجتها معالجة إضافية في مناهج تعليم العربية لغة ثانية.
- تحديد الأخطاء التي يرتكبها متعلمو العربية لغة ثانية من خلفية لغوية محدّدة، ومعالجتها معالجة إضافية مركّزة في مناهج تعليم العربية لغة ثانية، وإعداد مواد إثرائية للتعليم الذاتي؛ مثل: تعليم العربية للكوريين أو الصينيين... إلخ.

وهكذا تصبح البيانات الناجمة عن معالجة النصوص منطلقاً لحل المشكلة أو التحول عن الوضع غير المرغوب إلى وضع آخر أفضل، وتغدو أساساً متيناً لتقييم نتائج التخطيط اللغوي وإجراءاته ومدى نجاح السياسة اللغوية المرسومة ونجاح أدواتها التنفيذية.

مثال ٢:

رأت إحدى الجامعات أن تبني معجمًا سياقيًا للمصطلحات السياسية الشائعة في الصحافة العربية المعاصرة؛ فأنشأت مدونة من عدد كبير من النصوص الصحفية الممتدة لخمس سنين، ثم انتهت إلى فرز المصطلحات السياسية الشائعة في الصحافة العربية، وانتهى هذا الفرز إلى بناء المعجم السياقي للمصطلحات السياسية انطلاقًا من المدونة، على أن بناء المعجم كان هدفًا مرحليًا من أهداف بناء المدونة، ويترتب عليه هدف أبعد ينبثق من "التخطيط اللغوي لنشر العربية وتطوير تعليمها لأغراض خاصة"؛ إذ تأسس هذه المصطلحات وتعريفاتها وسياقاتها؛ لبناء مادة تعليمية لتعليم العربية لأغراض سياسية ودبلوماسية على هدي من توظيف المدونات وتناجها في توجيه تعليم العربية لأغراض خاصة توجيهاً علمياً رشيداً. ثم تخضع هذه المواد التعليمية للتقييم والمتابعة للحكم على مدى استيعاب المدونة وقدرتها على تمثيل المصطلح السياسي والدبلوماسي المعاصر، ثم يُنظر في هذه المواد مجدداً وتعزز بمداخل أو معاني وسياقات جديدة للمداخل القديمة.

ونخلص مما تقدم إلى أن لسانيات المدونات أساس متين لتحقيق جوهر اللسانيات التطبيقية وفلسفتها وبمزيد من الثقة والاطمئنان الذي توفره المدونات ومعالجتها النصية والإحصائية المتنوعة.

٣- وجوه من انتفاع اللسانيات التطبيقية بالمدونات:

لعلّ مما يمكن قوله إن اللسانيين نجحوا في استثمار المدونات وتعظيم دورها وصولاً إلى لسانيات المدونات، وأمكن لهم تطويرها؛ لتكون أداة طيّعة ومُعينة في مجالات اللسانيات التطبيقية، بدءاً بالنظر وانتهاءً بالتطبيق، ومن هذه المجالات: تعليم اللغة الأم، وتعليم اللغة الثانية، والصناعة المعجمية، وتحليل الخطاب، واللسانيات الجنائية، والأسلوبية النقدية، والترجمة... إلخ^(٧).

ولما كانت هذه التطبيقات ووجوه الانتفاع متداخلة ومتقاطعة بين علوم ومجالات متنوعة، مثل: تعليم اللغة الثانية وتحليل الخطاب، فإنني سأكتفي بالإطار العام ومن ثمّ أفصل في المتن.

٣-١- مَثَلٌ من وظيفية اللغة^(٨):

إن منتهى مقصد الوظيفية الوقوف على واقع استعمال اللغة في مناحي الحياة المختلفة؛ فتكون اللغة وسيلة للتواصل

بالمعنى الاجتماعي الواسع، أما الوظيفية بالمعنى اللساني التطبيقي فهي: تعليم قواعد اللغة وعناصرها كما هي في واقع الاستعمال، لا كما وضعها النحاة واللغويون. ولا شك بأن لسانيات المدونات تلتقي مع الوظيفية في توجهها الوصفي الواقعي حين ترصد واقع اللغة في الاستعمال الجاري على فترات من الزمن، وأبسط القول في المسألة على النحو الآتي:

يقتضينا الوقوف على وصف وظيفي معاصر لأحد أبواب النحو العربي، وليكن (الاستثناء) مثلاً، في مرحلتين أساسيتين:

الأولى: أن نبني مدونة خاصة بمصادر مختارة بعناية من مصادر النحو العربي وامتدادها على تراخ من الزمن حتى هذا اليوم، وستنتهي معالجة هذه المصادر وتحليلها إلى:

- رصد قواعد باب الاستثناء كما وردت عند النحاة واللغويين القدماء والمعاصرين.
- تصنيف القواعد وفق شيوعها لدى العلماء.
- بناء وصف كلي جامع لقواعد الباب.

الثانية: استثمار إحدى المدونات العربية الضخمة المعاصرة التي تضم نصوصاً متنوعة في فترات زمنية ممتدة، ولابد أن تنتهي معالجة هذه النصوص والبحث فيها إلى:

- تعيين جميع أمثلة أسلوب الاستثناء الواردة في المدونة.
- ترتيب أمثلة أسلوب الاستثناء وفق شيوعها.

وستنتهي العمليات المشتركة على مدونة القواعد النحوية والمدونة اللغوية (مدونة النصوص وشواهدا) إلى تحديد القواعد النحوية نادرة الاستعمال أو عديمته، وهكذا نخلص إلى صورة وظيفية رشيقة لباب الاستثناء تقتصر على القواعد اللغوية المستعملة فحسب، وهي مفروزة بحسب الشيوع ومؤيدة بشواهد واقعية معاصرة كما وردت في الاستعمال الواقعي لدى أبناء اللغة دون تكلف أو تمحل!

ولاشك أن عمليات المعالجة وأدوات المدونة المتنوعة ستكشف لنا جوانب من الاستعمال الواقعي وموارد هذا الأسلوب في العربية المعاصرة؛ ومن أهم هذه الجوانب، الشواهد التي خرقت قواعد أسلوب الاستثناء المستقرة في العربية ولا سيما "الاستغناء بالمعنى عن اللفظ"؛ وذلك أنه شاع:

صَوَّتَ أعضاء المجلس **باستثناء واحد**. والأصل: صَوَّتَ أعضاء المجلس **إلا واحداً**.

أغلقت الأردنُّ حدودها **باستثناء الجنوبية**. والأصل: أغلقت الأردنُّ حدودها **إلا حدودها الجنوبية**.

وهكذا فإن لسانيات المدونات قد أسهمت في ترشيح باب الاستثناء في النحو العربي، وجعله في قسمين: ينسب الأول إلى النحو العلمي الموضوع للباحثين وطلبة العلم والمنشغلين بالدرس اللغوي، ويتجه الثاني إلى نحو تعليمي رشيقي يُقدَّم للمتعلمين من أبناء اللغة عامة وغير المتخصصين في النظر اللغوي وفلسفته وصورته التاريخية. على أن استثمار لسانيات المدونات حتى هذه المرحلة ما يزال نظرياً ينتظر التطبيق والإسهام في جهود التخطيط اللغوي وتنفيذ سياسة تيسير العربية ووظيفتها؛ وإنما يكون ذلك باللسانيات التربوية مساراً للسانيات التطبيقية، ويتمثل ذلك في بناء مواد تعليمية قائمة على واقع المدونة وموارد استعمال الباب عند أبناء اللغة، وبذا يُقدَّم باب الاستثناء في صورته الوظيفية الرشيقة مسنوداً بأمثلة من الاستعمال الواقعي للعربية^(٩).

٣-٢- مَثَلٌ مِنْ عِلْمِ الدَّلَالَةِ، التَّرَادُفِ:

نالت قضية الترادف في اللغة العربية حظاً وافراً من الدراسة والمناقشة في الدرس اللساني العربي، ولا سيما اللسانيات التراثية؛ فأكثر علماء العربية على وقوع الترادف بدليل عدّه الطريق الأول والأولى في تفسير المعاني وبيانها؛ فأغلب المعاجم العربية تعتمد على الشرح بالمرادف. وقليل من العلماء لا يرون الترادف وارداً في العربية، وإنما يفسرون المعنى بما يقاربه ويوافقه في كثير من صفاته ومحتواه الدلالي.

ومع أن لدينا حدساً يراجع النظر في آراء الفريقين إلا أنه يَظَلُّ حَدْسًا تقريبياً، وإن كان كافياً عند تشومسكي، بأن الترادف يقع جزئياً وليس كلياً؛ وإنما يستند هذا الحدس على تجربة واقعية مرجعها سياقات قريبة نستدعيها عند الحاجة الأنبية المتعجلة. ويظل الحكم قائماً إلى أن نَظْفَر بدراسة أشمل وأوسع تُصَدِّق هذا الحدس؛ فهل يمكن للمدونات تصديق حدسنا؟

إن غاية القصد من التساؤل السابق القول: إنّه يمكننا أن نختبر فرضية الترادف في العربية^(١)، بالنظر في المترادفات السياقات الواردة في مدونتنا التي اعتمدناها.

وهذا مثال من المترادفات المشهورة في العربية: مُعَلِّمٌ، مُدَرِّسٌ، أَسْتَاذٌ، مُحَاضِرٌ، دُكْتُورٌ؛ فإنها جميعاً على معنى مشترك عام.

ولكن ما رأي السياق وواقع الاستعمال؟

لاشك بأن واقع الاستعمال في حدوده الدنيا سيكشف لنا عن أمثلة وسياقات متنوعة، منها:

- مُعَلِّمٌ المدرسة بحاجة إلى تحسين ظروفه.
 - أَسْتَاذُ اللُّغَةِ الْعَرَبِيَّةِ يمكنه تدريس مادة التربية الإسلامية.
 - الأستاذ الجامعي يعاني صعوبات في الترقية.
 - دكتور الجامعة يجب أن يتمتع بكفايات معرفية متنوعة.
 - يعمل عبد الله محاضراً متفرغاً في الجامعة.
- وهذه نماذج سياقية بسيطة يفرزها البحث الأولي في المدونات، ولتحقيق هدفنا من التثبت من مدى الترادف وانطباقه على المفردات المقترحة؛ فإنه يمكن لنا وبشكل يدوي الآن محاولة تبديل مواضع المفردات في السياقات السابقة، وتحكيم حدسنا في وقوع الترادف أو امتناعه؛ فهل يصح أن نقول:

- أستاذ المدرسة بحاجة إلى تحسين ظروفه.
- أستاذ الجامعة يجب أن يتمتع بكفايات معرفية متنوعة.
- يعمل عبد الله أستاذاً متفرغاً في الجامعة.
- عَمِلْتُ مُعَلِّماً متفرغاً في الجامعة.
- دكتور المدرسة بحاجة إلى تحسين ظروفه؟

والظاهر أن الحدس يمكنه تقبل الجمل الأولى والثانية والثالثة، ولا يتقبل الرابعة والخامسة؛ وهذا يعني أن الترادف نسبي بين هذه الكلمات، وأنه لا يستقيم في سياقات معينة ولا سيما إن اقترنت بكلمة (مدرسة) أو (جامعة).

وهكذا يمكن أن تمضي المدونة الذكيّة وأدواتها في فرز السياقات وتحديدتها، ثم تقوم أداة أخرى بإجراء تباديلٍ موقعيةٍ سياقيةٍ اعتماداً على سياقات المدونة نفسها ومن ثَمَّ التوصل إلى مقدار الترادف النسبي بين هذه الشواهد؛ فقد يكون الترادف بنسبة ٥٠٪ أو ٨٠٪... إلخ.

ويسعدنا بعد هذا التوثق العلمي من مدى الترادف بين الكلمات المتقاربة معنى؛ أن نستثمر هذه النتائج المحصّلة في سياق تعليم اللغة لأبنائها بدءاً من المراحل الإعدادية بمزيد من تدريبات الحدس التي تُمكن ابن اللغة من تنمية إحساسه اللغوي وسليقته وحساسيته السياقية للمترادفات؛ فيكون من السهل عليه التمييز بين الترادف المعجمي في المعجم (خارج السياق) والاستعمال السياقي وخصوصياته؛ فيسهل عليه لاحقاً معرفة أن كلمة (دكتور) تعني (معلم) في المعجم على التقريب، ولكنها تختص بكلمة (جامعة) وما يقترن بها من مفردات؛ فلا يستطيع ابن اللغة استعمال كلمة (معلم) مكان (دكتور) وهكذا.

ولعل خير إستراتيجية؛ لتوجيه الحساسية اللغوية نحو الحساسية السياقية الوظيفية، تكون:

- أن يبني المعلم مدونته السياقية الخاصة المحدودة من الكتاب المقرر، ويدير عليها أمثله وتداريبه التبادلية الخاصة.
- أن يوظف المدونة على نحو مباشر، بمباشرتها مع المتعلمين؛ لاستكشاف فرضياتهم البسيطة الخاصة.

وأحسب أنه يمكن استثمار هاتين الإستراتيجيتين في تعليم اللغة الثانية مع فارق في التنفيذ ومدى التعرّض؛ فتطبيق الإستراتيجية الأولى مع متعلمي اللغة الثانية يقتضي توجيهها مباشراً نحو الفرق السياقي^(١١) بين المترادفات المعجمية، أما الإستراتيجية الثانية فتصلح أيضاً لتنمية الحدس باللغة الثانية، وبناء فرضيات حول التبادل الموقعي السياقي لأشباه المترادفات حين يبني المتعلم فرضياته الخاصة ثم يبدأ تطبيقها واختبارها في التدريبات المناسبة التي أعدها المعلم، مثل: تدريبات اختيار المرادف المناسب للسياق، وإعادة الصياغة باستخدام المرادف الأقرب، والإنشاء الكلّي الذي يكشف عن مدى تمثّل المتعلم للفروق السياقية الوظيفية بين المفردات حين يُنشئ نصّاً^(١٢).

١-١- مَثَلٌ من تحليل الخطاب الصفي وأنماط التفاعل^(١٣):

يعتني الخطاب الصفي بأنماط التفاعل وعملياته داخل الصف، ويحظى الخطاب الصفي بقدر أكبر من العناية في صفوف تعليم اللغة الثانية؛ نظراً لموضوع اشتغاله، فهو يُعلّم اللّغة باللّغة نفسها بعفوية وتلقائية غالباً، وقد يكون عن قَصْدٍ وبيّنة، ولا شك بأن دراسة الخطاب الصفي وما ينطوي عليه من كفايات إدارة الصف (توزيع الأدوار) وطرق التدريس (الصرحة والضمنية) وأنماط التفاعل (العفوية والمخططة)؛ تُسهم إسهاماً كبيراً في إدارة عملية تعليم اللغة الثانية، وتوجيه اكتسابها عن بيّنة وقَصْدٍ، ومع ما يكتسبه الخطاب الصفي من أهمية كبرى في تيسير عملية التعليم والتعلم وتسييرها، فإن الدراسة العلمية الدقيقة لهذا الخطاب ما تزال تُقَصّر عن الحد الأدنى المؤمل منها في اللسانيات التربوية الغربية،

أما في تعليم العربية فلا يكاد هذا الموضوع يُذكر في الدرس العلمي، فكيف يمكن أن تُسهم لسانيات المدونات في دراسات الخطاب الصفي وتوجيه تعليم اللغة توجيهاً رشيداً قاصداً؟

وأحرص بالقول إن ثمة عواملَ تتحكم في الخطاب الصفي وتوجهه؛ فمنها:

- الكتاب التعليمي وما يتصل به من محتوى معرفي ولغوي، وحجم الكتاب، ومدى ملائمة حجم المحتوى للوقت المخصص للتعليم.
- طرق التدريس وإستراتيجيات التعلم التي يفرضها الكتاب.
- طبيعة المهارة اللغوية والمحتوى المعرفي (عناصر اللغة).
- مستوى كفاية المتعلمين.
- أنماط شخصيات المتعلمين.
- شخصية المدرس بجوانبها المختلفة، ولا سيما كفاياته المعرفية والمهنية.

ومهما يكن من تأثير هذه العوامل في الخطاب الصفي وما ينطوي عليه من إدارة أنماط التفاعل فإن توجيه المدونات ولسانياتها للإسهام في دراسة الخطاب الصفي ينبغي أن يقوم على مدونة ناجزة قابلة للمعالجة واستخلاص النتائج منها على النحو المرتجى، وهذا يعني أن بناء تحليل الخطاب الصفي في صفوف تعليم العربية لغة ثانية يقتضي منا البدء من الصفر؛ وإنما يكون هذا على وفق المراحل التالية:

المرحلة الأولى: مرحلة ما قبل بناء المدونة:

وهي مرحلة تحضيرية وتجريبية في الوقت نفسه، وتتضمن عمليات التحضير هذه:

- أ. تسجيل دروس واقعية من دروس تعليم اللغة العربية لغة ثانية في موضوعات ومحاوَر متنوعة بين (مهارات أساسية، ومهارات فرعية) وعناصر اللغة (الأصوات والمفردات...).
- ب. تفريغ النصوص المسجلة وتحويلها إلى نصوص مكتوبة تُرمز ببطاقة خاصة تتضمن بيانات تعريفية مثل: موضوع الدرس (قراءة)، ومستوى الطلبة (مستوى رابع)، والمدرس، والطلّاب (عددهم، جنسهم) ... إلخ^(١٤).
- ت. بناء نموذج خاص لتحليل الخطاب الصفي وفق النماذج المتوافرة في دراسات التفاعل الصفي: النظرية الاجتماعية الثقافية أو نظرية لونغ... إلخ.
- ث. تحليل نماذج متنوعة من الخطاب الصفي وفق النموذج المعتمد؛ لتكون أمثلة يتدرب عليها النظام ويبني خبرته بناء على النماذج الجاهزة.

المرحلة الثانية: مرحلة بناء المدونة:

تقوم هذه المرحلة بشكل مباشر على عمليات الإدخال وتهيئة المدونة للعمل، ويشمل:

- أ. إدخال نموذج تحليل الخطاب الصفي المعتمد مُدَيلاً ببطاقة توصيف النص وبياناته.
 - ب. إدخال نماذج محللة يدوياً إلى المدونة؛ ليتمكن النظام من اعتياد التحليل وبناء افتراضاته وتطويرها مع اطراد النماذج المحللة.
 - ت. إدخال أدوات التحليل التي ستحاكي النموذج المحلّل؛ فمن المتوقع أن تكون أدوات المدونة قادرة على فرز النصوص حسب المهارة والموضوع ومدة الدّرس ومستوى المتعلمين... إلخ، وقادرة على تحليل أنماط التفاعل بناء على مدخلات تقريبية تخضع للمراجعة الدورية؛ فمثلاً، يمكن فرز نص الدرس بناءً على الترميزات الخاصة بكلام المعلم وكلام الطلاب: م / المعلم. ط ١: أحد الطلاب. ط ٢: طالب ثانٍ. ط ٣: طالب ثالث. ويمكن بناءً على هذا الفرز تحليل نمط التفاعل القصير عندما يتضمن الدرس أسئلة:
 - س م: سؤال المعلم.
 - س ط ١: سؤال طالب.
 - س ط ٢: سؤال آخر من طالب ثانٍ.
- وهكذا تُرَمَز مثل هذه الأسئلة وتحدّد وظيفتها في سياق الدّرس.
- ويمكن تحديد مساهمة الأستاذ ومساهمة الطلبة الآخرين من التفاعل بتحليل حجم النصوص.

المرحلة الثالثة: مرحلة بناء المدونة وتزويدها بالنصوص:

غاية هذه المرحلة تزويد المدونة بأكبر عدد ممكن من الدروس الواقعية في تعليم العربية لغة ثانية والمُفَرَّغَة كتابياً والمزوَّدة ببطاقات بيانات وصفية بُنِيَتْ سابقاً، ولعل من الضروري أن تكون المدونة مرنة قابلة للإضافة متى توفرت نصوص جديدة بما يزيد العينات الخاضعة للتحليل، وصولاً إلى تعميمات أدقّ وأصْبَطَ.

المرحلة الرابعة: مرحلة تشغيل المدونة وإعمالها:

في هذه المرحلة تكون المدونة جاهزة، فتضم نصوص الدروس كاملة (متن المدونة)، وأدوات اشتغال المدونة وتحليل متنها، مثل: نموذج التحليل المعتمد، ونماذج مُحَلَّلَة وفق النموذج المعتمد، والترميزات والتحشيات المتنوعة... إلخ.

المرحلة الخامسة: مرحلة التجريب والمراجعة:

غاية هذه المرحلة تجريب أدوات المدونة، ومدى فاعليتها في تحقيق هدفها العام المتمثل في دراسة الخطاب الصفي وأنماط التفاعل في صفوف تعليم العربية لغة ثانية، وما يتصل بهذا الهدف من أهداف فرعية أخرى تمثلها الأسئلة التالية، تمثيلاً لا حصراً:

- ما أنماط التفاعل الشائعة في صفوف تعليم العربية لغة ثانية؟
- ما مدى توزيع التفاعل الصفي بين المعلم والمتعلم؟

- ما أثر المهارة في اختلاف أنماط التفاعل الصفي؟
- ما أثر موضوع الدرس في اختلاف أنماط التفاعل الصفي؟
- ما مدى استخدام اللغة الوسيطة في الخطاب الصفي؟
- ما أنواع المدرسين من حيث مدى عنايتهم بالتفاعل الصفي؟
- ما أنماط الكفاية الإستراتيجية لدى الطلبة؟
- ما أنماط التفاعل المعدّل التي يقدمها المعلم؟

ولاشك بأن دقة الإجابة عن هذه الأسئلة معتمدة على حجم متن المدونة، ودقة أدوات التحليل المعتمدة، وقد قدّمنا بالقول إن المدونة ولسانياتها أداة فاعلة في التخطيط اللغوي، وتحقيق غايات اللسانيات التطبيقية في حلّ مشكلة أو التحوّل نحو وضع أفضل؛ وعلى هذا، فإن المراحل الخمس المتقدمة إنما تتعلق بدور المدونة ولسانياتها المتمثّل في تقديم المعلومات والبيانات والتفاصيل المطلوبة لحل مشكلة ما مهما كانت هذه المشكلة^(١٥)، دون أن تقدم حلولاً لها؛ فالبيانات والتحليلات التي قدمتها لنا لسانيات المدونات في دراسة الخطاب الصفي هي المرتكز الرئيس الذي ستبني عليه اللسانيات التطبيقية وأداتها التخطيط اللغوي لتشخيص واقع الخطاب الصفي في صفوف تعليم العربية لغة ثانية، ومن ثمّ وضع الخطط التنفيذية اللازمة لتعزيز جوانب القوة، ومعالجة مشكلات الخطاب الصفي، ومسالك المعلمين في التدريس.

وهذان مثالان توضيحيان^(١٦):

مثال ١:

ذكرنا أن من الأهداف الفرعية لهذه المدونة الإجابة عن السؤال: ما مدى توزيع التفاعل الصفي بين المعلم والمتعلم؟

وقد أظهر التحليل النتائج التالية:

المهارة	خطاب المعلم	خطاب الطلبة
التحدّث والمحادثة	٥٠٪	٥٠٪

بالنظر إلى طبيعة المهارة وأهدافها العامّة والخاصة فمن الواضح أن هناك مشكلة عامّة في المجال؛ ذلك أن مهارة التحدث وما يداخلها من المحادثة والمناقشة تقتضي دوراً أقلّ من المعلم، ودوراً أعظم للطلبة، ولعلّ البحث أكثر في خلفيات هذه النتيجة وما يتصل بها من متغيرات أخرى يفضي إلى مشكلة في المنهاج المعتمد نفسه؛ فلعله لا يركز على مهارة التحدث والمحادثة والمناقشة، ولعل البرنامج نفسه لا يعتني كثيراً بمهارة المحادثة، فإن كانت هذه النتيجة المحصّلة متعلقة بالبرامج القائمة على التحدث (دورات تحدّث) فالمشكلة أكبر وأعظم!

مثال ٢:

قدّرنا أن من الأهداف الفرعية لهذه المدونة وعملياتها الإجابة عن السؤال: ما أنماط التفاعل المعدّل التي يقدمها المعلّم؟

وبعد التحليل والبحث، تبين لنا ما يلي:

نمط التفاعل	نسبة التوظيف والاستخدام
إعادة الصياغة	٨٠٪
إعادة السؤال بطريقة مختلفة	٥٪
إبطاء سرعة الكلام	٣٪
التصويب المباشر	٤٠٪
.....	٠٪

إنّ هذه النتائج تُظهر بوضوح لاشك فيه أن ثمة مشكلة لدى معلمي اللغة العربية لغة ثانية، تتمثل في عدم الإحاطة بإستراتيجيات التفاعل الصفّي، وكيفية مساعدة المتعلم على تعديل أدائه وتطويره نحو الكفايات المرومة، وأنّ هناك وجهة نظر خاطئة سائدة هي: تعديل الأخطاء أنيّا، وتوجيه مباشر، وعدم اللجوء إلى أنماط أخرى يمكن للمتعلم أن يفهمها من طبيعة التفاعل نفسه، وبمزيد من التحليل الدقيق يتكشف لنا، مثلاً، أنّ مرّد المشكلة إلى عدة أسباب، نقدرها بما يلي:

- غياب تأهيل المعلم تأهيلاً لسانياً منهجياً في كيفية إدارة التفاعل الصفّي وأنماطه الكفيلة بمساعدة الطالب وترقية كفاياته دون إحراج.
- خلوّ برامج تدريب المعلمين من دروس وتطبيقات في إدارة الخطاب الصفّي وتنويع أنماط التفاعل.
- ضعف خلفية المعلمين اللسانية.

وهكذا فإنّ تحصيل مثل هذه النتائج من لسانيات المدونات ووضعها بين يديّ اللساني التطبيقية خطوة مهمة في إصلاح الواقع وحلّ مشكلاته: فلا يمكن تطوير أداء المعلمين وترقية كفاياتهم المعرفية والمنهجية والتطبيقية، دون تشخيص الواقع تشخيصاً مبنياً على تنوع كبير من الدروس الواقعية، ولا يمكن تطوير هذا الأداء دون تحديث برامج تدريب المعلمين وتضمينها كل جديد معرفي، ولا سيما نظريات التفاعل وإدارة الصف اللغوي ونظريات الخطاب الصفّي في صفوف اللغة الثانية. كما تقودنا مثل هذه البيانات والوقائع إلى إعادة النظر في منهجية تأليف كتب العربية لغة ثانية، والتدقيق في الأوزان النسبية للمهارات والمعارف اللغوية بناء على الأهداف التي صمّم الكتاب أو محتوى الدورة لها.

وبذا تقدّم لسانيات المدونات حصيلة جهدها خالصاً على طبق من ذهب للسانيات التطبيقية، ورأسمي السياسات اللغوية، ومنفذي التخطيط اللغوي.

٤-٣- مثل من تحليل الخطاب الصفيّ المقارن:

قدّمنا القول بإمكانية إسهام المدونات ولسانياتها في ضبط عملية التعليم والتعلّم، بالنظر في أنماط التفاعل وطرق إدارة الصّف وأساليب التدريس المتنوعة. وتتناول هنا طريقاً آخر في توظيف لسانيات المدونات في مراقبة تعليم العلوم والمعارف والمهارات المختلفة؛ قصداً إلى استثمار الجوانب الإيجابية وتعديل المسالك الضعيفة وترقيتها، وتتناول هنا جانباً محدداً من الخطاب الصفي وأنماط التفاعل التي تجري داخل الصفوف التعليمية العربية بشكل عام^(١٧)، وغايتنا مقارنة الخطاب الصفي وأنماط التفاعل السائدة في تعليم المواد الإنسانية (التربية الإسلامية، والعلوم الاجتماعية، والفنون... إلخ) وخصائص الخطاب وأنماط التفاعل في تعليم المواد العلمية (العلوم، والرياضيات، والأحياء... إلخ).

ولاشك بأن هذه الغاية توجّه محتوى المدونة وعملها؛ فيقوم متن المدونة على دروس حقيقية متنوعة المجالات، وموزعة بين الإنساني والعلمي، ومفرّغة كتابياً؛ لتكون صالحة للإدخال والمعالجة والتحليل وفق نموذج تحليل التفاعل الصفيّ المعتمد، وبذلك ينتهي بنا التحليل إلى نتائج متنوعة بناءً على غايتنا الرئيسة من التحليل؛ فقد نتمكن من:

- تحديد الخصائص العامة للخطاب الصفي في الإنسانيات.
- تحديد الخصائص العامة للخطاب الصفي في العلوم.
- تحديد أنماط التفاعل الصفي في الإنسانيات.
- تحديد أنماط التفاعل الصفي في العلوم.
- تبين علاقة الموضوع بأساليب التدريس المتنوعة.
- تبين تأثير الموضوع بأنواع التفاعل الصفي.
- تبين الفروق بين أنماط التفاعل في دروس الإنسانيات في ما بينها.
- تبين الفروق بين أنماط التفاعل في دروس العليّيات في ما بينها.

ولاشك بأن أدوات التحليل التي تنتهي بنا إلى هذه النتائج والمستخلصات قادرة على إجراء المقارنات بين خصائص الخطاب الصفي وأنماط التفاعل المستخدمة في الإنسانيات والعليّيات، بما ينتهي إلى تمييز خطاب كل منها، ومن ناحية أخرى فإن هذه المقارنة قد تكشف عن ضعف أنماط محدّدة في الإنسانيات وشيوعها في العلوم، مع أنه يمكن توظيفها في الإنسانيات دون مشكلات جوهرية؛ كالتفكير الناقد وحل المشكلات... إلخ. وقد تكشف المقارنة عن أنماط جوهرية للتفاعل الصفي في الإنسانيات يغلب عليها السلبية من ناحية تدريسية؛ فقد يغلب على التفاعل هيمنة المعلم ومركزيته، وتهميش مشاركة الطلاب وإضعاف دورهم في التعلّم.

وبهذا تقدم لسانيات المدونات عصارة عملها للسانيين التطبيقيين؛ للنظر في هذه الخلاصات واستثمارها في دراسة الخطاب التعليمي وأنماط التفاعل وطرق التدريس الموظفة في مدارس إحدى الدول العربية، والانطلاق من هذه النتائج لـ:

- إعادة النظر في برامج إعداد المعلمين، وتطويرها بما يتوافق مع المعايير العالمية.
- تقييم طرائق التدريس، وفرزها بين قوي وضعيف، وإعادة النظر فيها.
- بناء المناهج على نحو يمكّن من تفعيل خصائص الخطاب الصفي، وأنماط التفاعل التي تخدم الإنسانيات من ناحية، وتلك التي تخدم العِلْمِيَّات من ناحية ثانية، وتلك التي تعزز كليهما معاً.
- تمكين المتعلمين من ممارسة إستراتيجيات التعلم المناسبة، وإعادة تشكيل الخطاب الصفي المهيم السائد (سؤال من المعلم ثم جواب من الطالب، أو معلومة من المعلم ثم أسئلة من المعلم وإجابات من الطلاب).

٥-٣- مَثَلٌ من التحليل النقدي للخطاب^(٨١):

يجمع التحليل النقدي للخطاب بين تحليل لغة الخطاب والكشف عن خلفياتها الخفية؛ وبذا يكون منهجاً في تفكيك الاستعمال اللغوي والنصي، مستقصياً ما خفي فيه من هيمنة وتحيّز وتسلّط واستعلاء وعنصرية... إلخ. ولا تتوقف مقاصده عند فضح المحرّكات الفكرية للبنية اللغوية والمفردات والأساليب والنصوص، ولكنه يمكن أن يكون كاشفاً لرؤى اجتماعية واقتصادية مُهيمنة تُستغلّ في اللغة.

ويبدو أن المدونات تقدّم خدمات عظمى لهذا التحليل النقدي؛ فهي تمكّن من الوقوف على عدد كبير من الممارسات اللغوية الخطائية المنبئة بسلوك اجتماعي أو اقتصادي ما، بالنظر إلى ضخامة عدد النصوص والممارسات اللغوية المجموعة؛ وهذا بلا شك أداة كاشفة أصدق من استطلاعات الرأي الموجهة والمعتمدة على عدد محدود من العينات مقارنة بالمجتمع أو الفئة المقصودة من التحليل. ولعلّ المدونات ولسانياتها تُبَرِّئ التحليل النقدي للخطاب من كثير من ذاتيّته بالنظر إلى توسيع دائرة المتون التي يعتمد عليها.

وتشبه هذه المهمة، على اختلاف أهدافها وأدوات تحليلها، النتائج التي تُصدرها محركات البحث العالمية حين تشير إلى أكثر الكلمات تكراراً في البحث في دولة محددة؛ فقد اعتادت شركة غوغل إصدار تقارير سنوية تكشف عن الموضوعات الأكثر بحثاً في دولة معينة؛ فقد كان موضوع الحرب الروسية الأوكرانية، مثلاً، هو الأكثر شيوعاً في سنة ٢٠٢٢. وهكذا يمكن كشف اتجاهات الخطاب السياسي والاجتماعي والثقافي في بلد ما، انطلاقاً من مدونة وسائل الإعلام المختلفة، والبحث في نصوصها المتنوعة، ومعجمها المتداول في فترة زمنية ما.

مَثَلٌ من الخطاب الاجتماعي السياسي:

لنتخذ مثلاً يجمع بين السياسة والاجتماع ونَفْصُحُ اللغة المستعملة^(٨٢)؛ وهو موضوع انتشار خطاب الكراهية نحو اللاجئين في العالم، وفي مناطق محدّدة تعيّننا؛ فهل يُتَلَقَّى اللاجئون دائماً بخطاب كراهية وعنصرية بغضضة أم يُتَلَقَّون بخطابات أخرى؟

إنّ الناظر في مواقع التواصل الاجتماعي الحاضنة لفئات المجتمع ووسائل الإعلام الرسمية وغير الرسمية العبّرة عنه سرعان ما يتوقّف عند خطابات مليئة بالكراهية ضد اللاجئين في معظم أنحاء العالم^(٨٣)، ولا شك بأن خطاب الكراهية هذا تحمّلُهُ مفرداتٌ مُستَقَلَّةٌ بذاتها، وتراكيبٌ متنوعة (تركيب وصفي أو إضافي أو جَرِّي) وأساليب لغوية تحمل بين ثناياها تحريضاً أو كراهية أو معارضة أو ترحيباً

وتقبلاً... إلخ، وجملٌ ممتدة تتراوح في طولها انتهاءً بنصوص كاملة. ومعلوم أنَّ للمفردات (العامّة والمصطلحات) دوراً عظيماً في بناء الخطاب ومضمونه، بما يتجلى من حضورها في النصوص التي تعكس الخطاب وتمثله^(٢٦)؛ فعلى ذلك تتكاثر الحقول الدلالية التي تدور في فلك مشاعر الكراهية (الكراهية والبغضاء والحقْد... إلخ) وما يكتنفها من مصطلحات (الرفض السياسي) وعدم تقبُّل الآخر، وما يتقاطع مع هذا الحقل من مفردات الحقل الدلالي المضاد: الحب والاحترام والتعاطف والتكافل والإنسانية... إلخ؛ فإنها تجلُّ طبيعياً لثنائية ضدية لأي قضية!

إن النظر في مواقع التواصل الاجتماعي على اختلاف أصحابها (رسمية / غير رسمية، مثقفون / عامة الناس... إلخ) وبناء مدونة ضخمة تتضمن آلاف النصوص وملايين المفردات والتراكيب التي تتضمن أنواعاً مختلفة من النصوص الأصيلة والتعليقات والردود والمناقشات والرسوم الكاريكاتورية التي تتناول موضوع اللاجئين وقضاياهم المتنوعة، سيمكننا من الوقوف على وجهات نظر الناس والحكومات في العالم ومواقفهم من قضية اللاجئين.

إن تصميم أدوات بحث دقيقة يمكن الحكومات أو المنظمات الدولية المعنية باللاجئين من التوصل إلى وجهة نظر واقعية متباينة يصعب تحصيلها باستطلاعات الرأي العاجلة أو محدودة العينات؛ ويمكن لأدوات المعالجة أن تكشف لنا عن اتجاهات الناس نحو اللاجئين من زوايا متعددة؛ ولعل أفضل ما تبدأ به هذه الدراسات، البحث في الجمل التي وردت فيها كلمة اللاجئين؛ فقد يُظهرُ لنا جملًا مثل^(٢٧):

- مفوضية اللاجئين تُصدرُ تقريرها السنوي.
- مسؤول ملف اللاجئين في الأمم المتحدة يصل عمان.
- كم يبلغ عدد اللاجئين السوريين في الأردن؟
- ما مصير اللاجئين السوريين في الأردن؟
- خطاب «إعادة اللاجئين السوريين» لبلدهم يتصاعد في الأردن.
- إقرار خطة أمنية لمخيمات اللاجئين السوريين في الأردن.

وهكذا يكون الفرز الأولي للعبارات والجمل التي تضمنت كلمة (اللاجئين) عمومًا، ثم يمكننا البحث عن متلازمات الكلمة التي تصفها من ناحية اليسار: اللاجئين السوريون، اللاجئين الفلسطينيون، اللاجئين السودانيون، اللاجئين الأكران (الأوكرانيون)... إلخ. وكذا القول في تراكيب أخرى ملازمة لها: اللاجئين في تركيا، اللاجئين في لبنان، اللاجئين في إثيوبيا... إلخ. وحتى الآن، فإن عملية المعالجة عامّة، قصدت إلى تضييق نطاق البحث وحصره في إطار جُملي، ومن هنا تكون نقطة الانطلاق لتحديد خطاب الكراهية نحو اللاجئين بعامّة، أو اللاجئين السوريين تحديداً، ويمكن البدء بتحديد الحقول الدلالية المتعلقة بالموضوع؛ على النحو التالي:

- المفردات والتراكيب التي تنطوي على معاني تعبر عن مشاعر الكراهية والبغضاء والعدائية بناءً على جذورها وانطلاقاً من معجم المدونة: كره، بغض، عدا، حقر، سفل، سوء، هجم، شتم، سب، غضب، هدف... إلخ.

- المفردات (الجدور) التي تنطوي على موقف معارض للاجئين، مثل: عارض، رفض، منع، حارب... إلخ.
 - المفردات والتراكيب التي تحمل موقفًا صريحًا ومباشرًا، مثل: أَدَان، أنكر، شَجَب، عبَّر عن استيائه (امتعاضه، أسفه... إلخ)، مثل: أَدَانُ الإساءة للاجئين، شَجَبَ حَمَلَات الكراهية ضد اللاجئين على مواقع التواصل الاجتماعي.
 - المفردات التي تنطوي على معانٍ مضادة للكراهية والبغضاء مسبقة بأدوات نفي: لا يُحِبُّ، لا يَقْبَلُ، لا يُريد، لَنْ يَسْمَحَ، لَنْ يوافقَ، لَمْ يَرعَ مشاعرَ، لَمْ يَسْتَقْبَلْ، لَمْ يُحَسِّنْ مُعَامَلَةً.
 - الجملة الاسمية المنفية: لَيْسَ مرحبًا به، لا لجوءَ بَعْدَ اليَوْم... إلخ.
- وبالنظر إلى عمليات الفرز المتعددة والمتنوعة بين مفردة وتركيب وجملة ممتدة، يمكن الحصول على جمل محددة تتضمن الحقول الدلالية المقصودة؛ مثل:
- **اللاجئون السوريون يعانون من الكراهية في بعض المدن التركية.**
 - **خطاب الكراهية يتعالى ضد اللاجئين في باريس.**
 - **الانتخابات التركية تكشف حجم الإساءات والكراهية التي يعاني منها اللاجئون السوريون.**
 - **حملة كليدار أوغلو توجَّع مشاعر الكراهية والبغضاء تجاه اللاجئين السوريين.**
 - **تصاعد وتيرة العدوان على اللاجئين السوريين في لبنان.**
 - **لا يتعاطف مناصرو كليدار أوغلو مع اللاجئين السوريين.**
 - **خطاب المعارضة التركية لم يرعَ مشاعر اللاجئين السوريين.**
 - **الأردن لم يستقبل إلا اللاجئين السوريين الذين قدموا مع عائلاتهم.**
 - **الأردن لن يسمح بعودة اللاجئين السوريين إلى بلدهم دون ضمانات.**
 - **ألمانيا تفتح أبوابها لاستقبال اللاجئين الأوكران.**
 - **رومانيا تستعد لاستقبال مليون لاجئ أوكراني.**
 - **الاتحاد الأوروبي يفتح ذراعيه لاستقبال لاجئي أوكرانيا.**
 - **تزايد اللاجئين السودانيين في إثيوبيا.**
 - **اللاجئون السودانيون لا يجدون مقومات الحياة الأساسية على الحدود...**
 - **ألمانيا لن تسمح باستغلال اللاجئين الأوكرانيين على أراضيها.**

ومع مزيد من البحث والتحليل النقدي الاجتماعي يمكن التوصل إلى أسباب الكراهية والعنصرية تجاه اللاجئين؛ كأن تكون أسبابًا اقتصادية متعلقة بفرص العمل كما هو في تركيا ولبنان، وقد تكون أسبابًا

قومية كما هو الحال في معظم مناطق جنوب تركيا التي تقطنها غالبية كردية، وقد تكون أسباباً سياسية متعلقة بخلافات حدودية... إلخ.

ويمكن للمدونات وأدوات تحليلها أن تَمَدَّنَا بمقارنات بين حجم خطاب الكراهية وتفاوته بين بلد وآخر، وعقد مقارنات دقيقة إحصائياً بين خطاب الكراهية وخطاب الترحيب والتعاطف في بلاد أخرى، وعلى صعيد آخر، فإنه يمكننا توظيف نتائج هذه الدراسة وأمثالها في دراسات التحليل النقدي المقارن للخطاب؛ إذ يسهل مقارنة المشاعر المتناقضة تجاه اللاجئين بعامة، كما يمكن إجراء دراسات خطاب مقارنة تكشف عن مفارقات السياسة الأوروبية في التعامل مع اللاجئين العرب والمسلمين من ناحية واللاجئين الأوكرانيين من ناحية أخرى؛ وهذا خطاب تجلّى على نحو ظاهر وعنصري لدى كثير من الساسة وصناع القرار والرأي في أوروبا!

مَثَلٌ ثانٍ من الأكاديمية:

يتجلى هذا المثال في أنّ كثيراً من الجامعات الغربية وبعض الجامعات العربية، توفّر مساحة للطلاب وأولياء أمورهم للتعبير عن آرائهم في سياسات الجامعة وتعليماتها، وطرائق التدريس، والتخصصات المختلفة، وتعتمد هذه الجامعات على تحليل هذه الآراء وتحويلها إلى بيانات تعتمد عليها في تقييمها الذاتي وتطوير برامجها وتجويد مخرجات التعلم بناءً على نتائج تحليل تلك البيانات؛ وبذا تسهم هذه المدونات في توفير فرصة للجامعة؛ لترى نفسها بأعين متلقي الخدمة وذويهم، بما ينتهي إلى توجيه سياسة الجامعة ورسم خططها المستقبلية بعين بصيرة، ورؤى متنوعة كاشفة، ولا شك في أن مؤسسات الاعتماد الأكاديمي والجودة تُعَدُّ مثل هذه الاستطلاعات (المدونات) متطلباً من متطلبات ضمان الجودة وتحسين مخرجات المؤسسات الأكاديمية وتطويرها على نحو دائم.

وقد يستقيم لنا أن نربط المثل الثاني بالأول على وجه التقريب، انطلاقاً من مفهوم السياحة التعليمية أو الاستثمار التعليمي؛ فكثير من بلدان العالم تعتمد على سمعتها الأكاديمية وسمعة جامعاتها العلمية في استقطاب الطلبة للدراسة فيها ولا سيما في برامج الدراسات العليا. إنّه يمكن لوزارة التعليم العالي أن توجّه جامعاتها نحو تقييم دائم وإعادة تقييم من زاوية استقطاب الطلبة الأجانب، وأثر سياسات الجامعة وتعليماتها ومسالك الطلاب والأساندة فيها ومواقفهم السياسية، وتوجهاتهم نحو الطلبة الأجانب وطريقة التعامل معهم، وسبل استبقائهم حفاظاً على سمعة الجامعة وما يتأتى من هؤلاء الطلبة من عوائد اقتصادية تكون ضخمة في كثير من البلدان.

والظاهر أن غاية التحليل النقدي للخطاب أبعد من النقد فحسب؛ فهي تتجاوز إلى إحداث التغيير في المجتمعات والاتجاهات السلبية (التهميش والكراهية والعنصرية ومعاداة الأجانب والأقليات... إلخ) والتحوّل نحو العدالة والإنصاف واحترام الآخر وتقبُّله؛ وبذا يصبح التحليل النقدي للخطاب أداة تغيير على مستوى الخطاب، بوصفه فكراً موجّهاً للممارسات الاجتماعية المتعالية والتهميشية والعنصرية، ولاشك بأن جميع الحكومات ومصادر التشريعية وأدواتها التنفيذية (الجامعات والوزارات والمؤسسات الاجتماعية والدينية والقضائية ومجلس النواب وديوان المظالم... إلخ) بحاجة دائمة ومتجددة للتحليل النقدي المستقل للخطاب؛ لإحداث التغيير وديمومته.

مَثَلٌ ثَالِثٌ من تعليم اللغات الأجنبية (٢٣):

ربما يستطيع التحليل النقدي للخطاب المبني على المدونات الكشف عن وجوه التحيز والعنصرية والخط من قيمة الآخر عند النظر في محتوى المناهج الدراسية والمواد التعليمية المقدّمة للمهاجرين في بلد ما، انطلاقًا من مضامين تلك النصوص من ناحية أو من ناحية طبيعة الخطاب الصفي الذي يمارسه المعلم، وكيفية توجيهه للمحتوى التعليمي؛ فقد يكشف تحليل المحتوى مثلًا، عن:

- تركيز المحتوى والخطاب الصفي على مَهْنٍ لا تحظى بكثير من الاحترام والتقدير في مجتمع ما، بالنظر إلى أنها مَهْنٌ يدوية.
- تركيز المحتوى على مخاطبة المتعلمين بوصفهم عائلة على المجتمع، وأن تعلم لغة المَهْجَر هو الأداة الأساسية لبسّ كفافهم والتخفيف عن كاهل الدولة.
- تغيب المحتوى لنماذج رفيعة من المهاجرين الناجحين الذين حققوا نجاحات متميزة في مجالات مختلفة.
- ربط التعليم الجامعي الراقى بلغة المَهْجَر، والتقليل من قيمة اللغة الأم، وتشويه دورها في حياة المهاجرين.

وهكذا تقدّم المدونات ولسانياتها أدلةً صريحة للتحليل النقدي للخطاب على رؤية تلك المناهج للمهاجرين بوصفهم عبئًا على المجتمع، وأنه لا بد من التخلص من هذا العبء بأي وسيلة مهما كانت دونيةً في السلم الاجتماعي والاقتصادي. وعلى اللسانيات التطبيقية التي تعتمد على هذه المنهجية تقديم حلولها التعليمية والاجتماعية والاقتصادية العادلة لهؤلاء المهاجرين بوصفهم إضافاتٌ جديدةٌ تحمّد التنوع اللغوي والثقافي في المجتمع، وتسهم في فتح آفاق جديدة له؛ وإنما يكون ذلك برسم سياسات تربوية ولغوية جديدة تُعظّم قيم العدالة والإنصاف والمساواة والتنوع الثقافي، ويمكن أن تتمثل هذه السياسات في:

- إعداد معلمي الأقليات واللاجئين إعدادًا يرسخ حقوق الإنسان اللغوية والاجتماعية والاقتصادية؛ إعدادًا يجمع بين المعرفة العلمية بالنظريات الثقافية والاجتماعية والأنثروبولوجية من ناحية، والمعرفة التربوية التطبيقية من ناحية أخرى.
- بناء مناهج تعليمية تُعلي من شأن التنوع اللغوي والثقافي، وتقدّم اللغة الثانية بوصفها أداة حرةً ومتاحةً بالتساوي للجميع للترقي الثقافي والاجتماعي والاقتصادي؛ فتكون (الإنجليزية) أداةً للمساواة وليست أداةً للتمييز والفرقة، وتعلي من شأن اللغة الأم بوصفها كنزًا ثقافيًا يُسهم في تجانس المجتمع وانسجامه.

٦-٣- مَثَلٌ من التحليل النصي المقارن بين اللغة الأم واللغة الثانية:

تَقْصِدُ مدوّنة كهذه إلى الإجابة عن سؤال عريض مفاده: ما مدى اقتراب نصوص متعلم اللغة الثانية من نصوص الناطق الأصلي المكتوبة؟ أو: هل يمكن لمتعلم اللغة الثانية إنتاج نصوص مكتوبة تضاهي نصوص الناطق الأصلي؟

لا بد من الإشارة هنا إلى أن دراسة إنتاج متعلمي اللغة الثانية المكتوب وعثراته ما يزال نادرًا في اللسانيات التطبيقية العربية؛ إذ ركزت دراسات تحليل الأخطاء على الأخطاء المنعزلة (الإملائية،

والصرفية، والصوتية والنحوية... إلخ) ونُدرت الدراسات التي توقفت عند النص الكامل، وما يتعرّف فيه المتعلمون من بناء النص وأعرافه وأساليبه^(٢٤).

إن إخفاق متعلم اللغة الثانية في تمثّل عناصر الكفاية الخطابية وإنجازها في خطابه المكتوب يورّثه أخطاء ظاهرة يسهّل تعرّفها وتمييزها، وهي أعطاب في الخطاب تنبئ من يطالعها بأنها أخطاء متعلّم أجنبي، وقد عرّفت هذه الأخطاء بـ «لكنة الخطاب» «Discourse Accent» ومن هذه الأخطاء^(٢٥):

١. اختلال في بنية الخطاب وتنظيم وحداته الداخلية، وكيفية الانتقال من وحدة خطابية إلى أخرى.
٢. أخطاء في استعمال الروابط اللغوية الشكلية (حروف العطف، حروف الجر).
٣. أخطاء في تقسيم الفقرات ودلالاتها على الوحدات الخطابية.
٤. أخطاء لغوية متعددة منها: استعمال مفردات غير مناسبة للسياق، واستخدام التصريف غير المناسب للمفردة (الفعل بدلاً من المصدر)، أخطاء في العدد... إلخ. وأكثر هذه الأخطاء في المفردات، ولكنها تؤدي إلى أعطاب في الخطاب.

ولطالما كانت عشرات الخطابات المكتوب علامات فارقة في إنتاج غير الناطقين بالعربية، ومتعلمي اللغات الثانية؛ إذ يصعب عليهم التخلص من ملامحها مهما بلغوا من مستوى الإتقان والكفاية! فهل يمكن أن تسعفنا لسانيات المدونات في مقارنة مستوى اقتراب متعلم العربية من مستوى الناطق الأصلي؟ وكيف يكون ذلك؟

لتحقيق المقارنة النصّية (النصوص المقارنة) بين نصوص أبناء لغة ما، ونصوص متعلميها لغة ثانية، فإن ذلك يقتضي بناء مدونتين مختلفتين أول الأمر؛ فالمدونة الأولى تتضمن نصوصاً متنوعة أنتجها ناطقون بالعربية -مثلاً- على اختلاف مستويات كفاياتهم وتحصيلهم من العربية، والثانية مدونة متعلمي العربية لغة ثانية، وهي مدونة تحتوي على نصوص متنوعة الشكل والغرض والمستوى أنتجها متعلمو العربية في سياق تعلّم مهارات الكتابة والتعبير. ثم ندمج المدونتين في مدونة واحدة دون خلط محتواهما:

المدونة العامّة	
مدونة الناطقين بالعربية	مدونة متعلمي العربية لغة ثانية
مقالات	مقالات ^(٢٦)
رسائل رسمية	رسائل رسمية
رسائل ودّيّة: شُخصيّة	رسائل ودّيّة: شُخصيّة
يوميّات	يوميّات
قصة قصيرة	قصة قصيرة

وبتصميم أدوات مقارنة ومضاهاة على المستوى النصي (التماسك والانسجام) بتفاصيلهما الدقيقة، يمكن إنتاج مقارنات متنوعة؛ فيمكن أن تفضي لنا مثل هذه المقارنات إلى نتائج، مثل:

مجال المقارنة (المقالة)	الناطق بالعربية	متعلم العربية للغة ثانية
الإحالة الضميرية	٪١٠٠	٪٩٠
مؤشرات الخطاب والربط بين الفقرات: لهذا، لذلك، وهكذا، وبناء على ما تقدم	٪٩٦	٪٦٠
استخدام المفردات في السياق المناسب	٪١٠٠	٪٩٠
استخدام المشتق المناسب	٪٩٩	٪٧٥

وتظهر المقارنة الافتراضية السابقة أن هناك مشكلاتٍ جوهريّة في بناء نصوص متعلمي العربية لغة ثانية، ولا سيما مشكلات التماسك والانسجام النصي، وتمضي بنا المقارنة إلى مزيد من الأمثلة الواقعية التفصيلية من النصوص المخزّنة الكاشفة عن مدى التقارب بين نصوص أبناء اللغة، ونصوص متعلميها لغة ثانية؛ فقد تنتهي المقارنات إلى نتائج، مثل:

١. ثمة تقارب شديد بين رسائل الناطقين بالعربية الشخصية / الودية ورسائل متعلمي العربية لغة ثانية.
٢. ثمة تقارب جيد بين رسائل الناطقين بالعربية الرسمية، ورسائل متعلمي العربية لغة ثانية.
٣. ثمة بون شاسع بين الكتابة النقدية لدى أبناء العربية، ومتعلميها لغة ثانية.
٤. ثمة بون شاسع بين يوميات الناطقين بالعربية، ومتعلميها لغة ثانية.

إن هذه النتائج العامّة لا بد أن تنطوي على مقارنات أدق، تكشف عن وجوه الافتراق بين النصين؛ فقد يرجع سبب الفرق

في الرسائل الودية المكتوبة إلى أسباب معجمية، تتمثل في عدم تمكّن المتعلمين من التمييز بين الخطاب الرسمي والودي؛ فهم لا يميزون متى يستخدمون (عزيزي، صديقي العزيز، أخي العزيز، أختي الحبيبة) وعبارات المجاملة المتنوعة التي تدلّ على طبيعة العلاقة الرابطة بين المتخاطبين (سعادة، معالي، رئيس الجامعة الموقر... إلخ)، وقد يرجع الاختلاف الكبير بين يوميات الناطقين بالعربية ويوميات متعلميها لغة ثانية إلى ضعف التمكن من أدوات السرد (الفعل الماضي، كان، ضمائر التكلم)... إلخ.

وتضعنا هذه النتائج والبيانات أمام تحدّد كبير طالما شغل المشتغلين بنظريات اكتساب اللغة الثانية وتعليمها: جدوى تدريس اللغة الثانية وإمكانية بلوغ المتعلم مستوى يضاهاى الناطق الأصلي، ويظهر أن تحصيل نتائج تقريبية تُعزّز كثيراً من فرضيات تأثير التعليم وأنماطه المتنوعة في ترقية كفايات المتعلمين وتطويرها؛ فمثلاً يمكن:

- تعزيز فرضية الفترات الحرجة واختلافها بين مستويات اللغة؛ فقد تؤكد النتائج صعوبة الوصول إلى تمكّن صوتي يضاهي الناطق باللغة، ولكنه يمكن له أن يتمكّن من النظام النحوي بل يتفوق على أبناء اللغة أنفسهم.
 - تثبيت مقولات التعرّض المكثّف ومنزله في تنمية التفاعل الصفيّ ومن ثمّ تطوير الأداءات (المُخرجات).
 - تأييد مفهوم «كنة الخطاب» التي تُشبه الكنة الصوتية؛ فقد يحتفظ متعلم اللغة الثانية بأخطاء في نصوصه المكتوبة أبد الدهر؛ أمانة على أنه ناطق أجنبيّ.
- وبالعودة إلى ما تقدّم من نتائج، فإنّه تتعرّز لدى اللسانيين التطبيقيين ضرورة بناء مناهج خطابية التوجيه^(٢٧)؛ تلك المناهج التي تعتني بالأنواع النصيّة وأعرافها وتقاليدها الإنشائية، ومظاهر تماسكها الشكلي والمضموني.

مَثَلٌ من اللسانيات الجنائية:

تعنى اللسانيات الجنائية والقضائية بدراسة اللغة بوصفها وسيلة التواصل في النصوص القانونية والجنائية؛ قصدًا إلى دراسة ما يكمن في هذه النصوص من عُوارق قانوني أو تحيز من ناحية، ودراسة اللغة بوصفها حدثًا تواصلًا تفاعليًا بين أطراف الاختصاص القضائي من ناحية أخرى، ومن ناحية ثالثة، تدرس اللغة بذاتها حين تكون دليلاً جرميًا أو دليل إدانة أو تبرئة.

كيف تسهم لسانيات المدونات في اللسانيات الجنائية^(٢٨)؟

تتفاوت أهمية لسانيات المدونات في اللسانيات القضائية والجنائية على وفق الحالة القضائية والأدلة المتوافرة: فتعظم أهميتها عندما يتعلق الأمر بمدونات النصوص القانونية التي تحتكم إليها دولة ما في شؤونها القانونية في شتى مناحي الحياة (السياسة والاقتصاد والمجتمع... إلخ). فإذا كانت المملكة العربية السعودية -مثلاً- قد أنشأت مدونتها القانونية الخاصة في جميع مجالات الحياة؛ سهّل على الباحثين الوصول إلى جميع قوانين العقوبات المعمول بها في المملكة بدءًا من تأسيس المملكة وتنظيم شؤونها القانونية والقضائية إلى هذا اليوم. وبذلك، يمكن تتبع التطورات التي طرأت على القوانين والأنظمة والتعليمات منذ وضعها الأول حتى يومنا هذا؛ فيكون دور المدونة في هذه المعالجة منحصراً في تتبع التطور التاريخي للقوانين وتفاعلها مع معطيات التطور الزمني والتحديث؛ وعلى هذا، فإنه يمكن معرفة التاريخ الدقيق لدخول «قانون عقوبات الجرائم الإلكترونية» المملكة بدقة متناهية، ويمكن أيضاً تتبع التطورات والتعديلات المتصلة بهذا الجانب من وسائل التواصل الاجتماعي؛ كتعديلات قوانين استخدام وسائل التواصل الاجتماعي وعلاقتها بأمن المجتمع وسلامته وتجريم استخدامها لنشر الفكر المتطرف أو الفكر الظلامي... إلخ.

ثم تكون هذه المدونة نفسها صالحة لاستخراج بيانات قانونية فرعية متعددة بالبحث والفهرسة والاسترجاع؛ يمكن -مثلاً- تحصيل بيانات دقيقة حول اللغة العربية في القوانين والتشريعات السعودية، في التعليم والاقتصاد والترجمة ولافات المحال التجارية، والتواصل الرسمي، ومراسم الدولة... إلخ.

وكذا القول عندما تبني دولة ما مدونة الأحكام القضائية في جرائم القتل وملفاتها الكاملة، فإنه يمكن للباحثين والمشتغلين بالقضاء والإدارة القضائية واللسانيات القضائية من:

- مقارنة الأحكام القضائية أو النظر فيها؛ للاستئناس بطبيعة الأحكام السابقة في قضايا مشابهة.
- استخلاص طبيعة المسوغات التي قدمتها هيئة المحكمة؛ لتسويغ حكمها وتخفيفه أو تغليظه.
- احتراسات المحكمة عند إصدار الحكم وما اطمأنت إليه من الأدلة فسوغ اطمئنان ضمير المحكمة للحكم.
- كيفية مقارنة المحكمة للقضايا وتكييفها القانوني ووجوه الشبه أو الاختلاف في الأحكام القضائية لجرائم متشابهة ومسوغات ذلك الاختلاف.

ويمكن استثمار هذه المدونة الخاصة بجرائم القتل في جوانب أخرى تسهم في تعزيز أمن المجتمع وحمايته من الجريمة؛ مثل:

- خُصِرَ جنس الجاني والضحية قد يُسْهِمُ في تقديم صورة لظواهر اجتماعية محدّدة في ذلك المجتمع؛ فإذا كانت نسبة الضحايا النساء ٦٠٪ فإن هذا يعطي مؤشراً على عُنفِ أساسه الجنس (ضد المرأة) في ذلك المجتمع.
- حصر العلاقة بين الجاني والضحية؛ مما يُسْهِمُ في تقديم تصوّر عن أسباب هذا العُنف المُفضي إلى القتل؛ فقد تكشف المدونات وعملياتها عن أنّ العلاقة المهيمنة هي علاقة الزوجية بنسبة ٦٥٪ مثلاً، وأن ٣٠٪ هي علاقة إخوة. وظاهر أن علاقة الزوجية تفسّر سبب العنف ضد المرأة إلى حد بعيد وكذا علاقة الإخوة.
- فإذا أضفنا إلى متغيري (الجنس، والعلاقة) متغير (سبب الجريمة)؛ فإن هذا قد يكشف عن تصور واضح لمثل هذه الجرائم: جريمة ضد الزوجة؛ لخلافات زوجية أو خلافات مع أهلها أو الشك في سلوكها أو مسالك الرجل المنحرفة (السُّكر والمخدّرات... إلخ)، وجريمة ضد الأخت تصنف على أنها جريمة شرف أو جريمة متعلقة بالميراث أو لخلافات شخصية... إلخ. أما إن كانت المجني عليها أمّاً فيغلب أن تكون أسبابها: طمع في الميراث أو علاقتها بزوجة الابن القليل أو أبناء زوجها من أم أخرى... إلخ.

وفي سياق منفصل فإنّ النظر في المدونة من حيث (جنس الجاني، وعمره) قد يكشف لنا مساراً ما في مجتمع ما؛ فإذا انتهت المدونة إلى أن نسبة مرتكبي جرائم القتل من الذكور الذين تقل أعمارهم عن ٣٠ عاماً كانت ٣٥٪ فإن هذا مؤشر اجتماعي خطير يؤكّد أن نسبة عالية من المجرمين شباب، وهي ظاهرة اجتماعية جنائية جديرة بالدراسة والعناية.

ولاشك في أنّ هذه الصورة التخيلية التي رسمتها مدونات قضائية محدّدة، غالباً ما تُسْهِمُ إسهاماً عظيماً في رؤية المجتمع رؤية كليّة واقعية، وتقدم لعلماء الاجتماع ولا سيما المتخصصين في علم اجتماع الجريمة والمخططين الاجتماعيين مادة دقيقة مسنودة بأدلة وإحصاءات تمكنهم من رسم تدابير وقائية وتنموية يستند عليها أصحاب القرار من المؤسسات المجتمعية والدوائر الأمنية المختصة بالأسرة... إلخ.

بل لعلّ دراسة لغوية دقيقة للمفاتيح هذه الجرائم تكشف لنا دور اللغة المباشر في هذه الجرائم؛ وذلك حين تكون اللغة هي دليل الجريمة بحدّ ذاته؛ كأن يكون سبب القتل:

- الدّم والتحقيق والشتم.
- التشهير عبر وسائل التواصل الاجتماعي.
- شتم الذات الإلهية والدين والأعراض.
- شهادة الزور واليمين الكاذبة.
- استخدام الألفاظ والعبارات العنصرية والتكفيرية.
- رسائل نصّية متبادلة بين أحد طرفي القضية وطرف ثالث مثلاً؛ فقد يكون سبب الجريمة رسالة نصّية تتضمن تهديدات، أو عبارة موحية بعلاقة غرامية، أو تخطيط لعمل ما.

فهذه مقاربات يمكن للباحثين في اللسانيات القضائية التوقّف عندها من نواح متعددة، مثل: الجوانب اللغوية لنصوص الأحكام ومسوغاتها وأدلتها، ومرافعات الطرفين وقدرتهما على المحاججة ومن ثمّ الإقناع والظفر بالنصر القانوني.

ولا شك في أن النظر في النصوص القضائية على تنوعها مجال مهم تتعاظم فيه أدوار المدونات، ولكن قد يغيب دور لسانيات المدونات إن كان الدليل في القضية نصّاً واحداً، وليس ثمة مشتبه به، ولكن أهميتها تزداد إذا كان هناك شكوك قويّة في شخص واحد أو عدّة أشخاص، وأمكن الحصول على نصوص أخرى للمقارنة؛ عندها تساعد المدونة في تحديد أصل التأليف ونسبته إلى متهم بعينه.

ولعلّ من الوجوه المستجدة التي يمكن للسانيات الجنائية الاستناد على لسانيات المدونات فيها، ما يتصل بدراسات تحليل خطابات الإرهاب والجماعات المتطرفة؛ إذ يمكن بناء مدونات خاصة بهذه الجماعات وتحليل خطاباتاتها تحليلًا دقيقًا يطوّق معجمها وتراكيبها وأساليبها اللغوية الشائعة وموضوعاتها وأفكارها؛ لتكون هذه التحليلات مرجعاً تفيء إليه الحكومات لتشخيص أصول التأليف، والتوثّق من مصادر نصوص جديدة كرسائل التهديد والبرامج الترويجية.

اللسانيات الجنائية في المجال الأكاديمي (الانتحال):

لعلّ الانتحال هو أكثر القضايا التي تشغل العلماء والباحثين والأكاديميين منذ زمن طويل؛ على أن هذا الانشغال تحوّل إلى قلق مقيم يجرّكه تحوّل الباحثين والمخترعين والعلماء والأكاديميين من السطو على جهودهم وإنتاجهم الفكري. وقد تفهّمت كثير من الدول هذا القلق، ونظرت إليه بوصفه قلقاً دولياً يستحق الطمأنينة، وأن هؤلاء الناس بحاجة إلى من يحمي إنتاجهم؛ فُسرّعت قوانين «حماية الملكية الفكرية» في كثير من دول العالم.

ومع هذه الحماية القانونية والتشريعية، إلا أن انفتاح العالم وتساقط الحدود على الشبكة زاد من حدّة هذه الانتهاكات والسطو على أفكار الآخرين، وتحصيل منافع منها، مثل: الدرجات العلمية والجوائز والمنح... إلخ. ولعلّ هذا الوضع دفع بكثير من المؤسسات الأكاديمية ولا سيما الجامعات والمراكز البحثية

إلى البحث عن وسائل للتثبت من أصالة الأطاريح والكتب والبحوث، والتوثق من نسب الاقتباس المسموح بها؛ فسعت جاهدة إلى بناء برامج لفحص «الاستلال» والتوثق من نسبة النصوص إلى أصحابها، وصار هذا الإجراء تقليداً وعرفاً في الجامعات ومنها الجامعات العربية؛ إذ تخضع النسخة الخام من الأطروحة للفحص والتثبت.

ومما انتهى إليه مَبْلَغُ عِلْمِي دراسة حالة أوجزها في ما يأتي:

الانتحال، دراسة حالة^(٢٩):

نشأت هذه الدراسة عندما استشعرت هيئة القبول وخدمات التسجيل في الجامعات البريطانية (UCAS) أن ثمة تشابهات كبيرة في «المقالة الذاتية» التي يقدمها الطلبة الراغبون في الدراسة في بريطانيا؛ وفيها يقدم الطالب معلومات ذاتية موجزة تبين دوافعه وغايته وأسباب اختياره للدراسة في بريطانيا في جامعة محددة وفي تخصص بعينه، وقد غلبَ على ظنّ وحدة التدقيق ومتابعة الطلبات أن الطلبة ينقلون أو يلجأون إلى قوالب جاهزة منشورة على مواقع أكاديمية أو تجارية على الشبكة؛ فما يكون منهم إلا «القَصُّ واللَّصْقُ» مع محاولات خجولة للخداع والتضليل بعبارات أو مفردات محدودة.

وقد رأت هذه الهيئة أنها بحاجة إلى نظام الكتروني (مُؤنَّمت) يمكنه فحص جميع نصوص الطلبة المتقدمين وإفادتهم

ومقارنتها بنصوص قديمة ترجع إلى سنتين من قَبْل، كما يمكنه مقارنة هذه النصوص بمصادر مفتوحة ومتاحة من الشبكة في المجال نفسه؛ وهذا يعني النظر في ٦٠٠,٠٠٠ نص على وجه التقريب، قُدِّمَتْ عام ٢٠٠٩، منها ٤٠,٠٠٠ (أربعون ألف) مقالة استُقبِلَتْ في اليوم الأخير! وقد استقر رأي الهيئة على أنها بحاجة إلى برنامج حاسوبي يمكنه التعرف على وجوه التشابه بين النصوص على مستوى الجملة، وأن ينجز البرنامج مقارنة ١٢٥ مقالة في عشر دقائق؛ بما يضمن التزام الهيئة بالموعد المضروب لمعالجة الطلبات دون أي تأخير.

وتحقيقاً لفكرة العمل هذه عملت الهيئة على:

١. تفرغ النصوص (مقالات الطلبة) لثلاث سنوات خَلَوْنَ، اعتماداً على النسخة الإلكترونية الخام المحفوظة لدى الهيئة، وهي نسخة قابلة للمعالجة.
٢. فحص طول النص؛ فمعظم المتقدمين استفادوا غاية الاستفادة من حجم الحروف المتاح (أربعة آلاف حرف) فكانت النصوص بالطول نفسه على وجه التقريب. وقد كان من المفيد استثمار هذا الطول للانطلاق من دراسات شيوخ المفردات، وتوزيعها في النصوص المتساوية الطول، بما يسهّل تعرّف المفردات المتوقعة.
٣. تغذية النظام بالنصوص الموجودة أصلاً من سنوات سابقة؛ لمعالجة النصوص الجديدة في حزمة مكونة من ١٢٥ نصاً؛ ضمناً لنسق محدّد من تدفق البيانات الذي يمكن للنظام بناؤه.

لقد كانت غاية الهيئة من هذا العمل التوصل إلى خلاصة واضحة وبسيطة وسريعة لجميع الطلبات التي تُستَقْبَلُ في أي وقت، ومعرفة طبيعة التشابهات بين المواد المُدْخَلَة (المقالات الذاتية) ونص واحد مقارن في وثيقة مفردة. وقد صُمِّمَ التحليل بناءً على دراسات مألوفة في شيوع المفردات، في النصوص الإنجليزية المتشابهة بالنظر إلى أن ٥٠٪ من المفردات وظيفية، و٥٠٪ معجمية؛ وعلى ذلك، فإن كل جملة تتألف من عشرين كلمة لا بد أن تكون عشر منها كلمات معجمية. وبناءً على هذه الاعتبارات واعتبارات توزيعية أخرى، كانت الجملة هي الوحدة التي سيتم العمل عليه، وكما هو معلوم؛ فإن أي دراسة لمعالجة قضية الانتحال «السراقات العلمية»^(٣٠)، لا بد أن تعتمد على مصادر ضخمة للنصوص؛ لإنجاز مضاهاة أدق وأعمق؛ إذ كلما زاد المحتوى المنشور زادت القدرة على كشف وجوه التشابه بين النصوص!

وبالعودة إلى الدراسة التي نتحدث عنها فإن تعرّف المصدر المنحول عنه مرتبط بنصوص السنوات السابقة المُدْخَلَة إلى النظام. وبناءً على التحليل والمقارنة حُدِّدَت ثلاثة مصادر للتشابه ظَهَرَت بألوان مختلفة؛ وبذلك يمكن للناظر إلى الشاشة معرفة العبارة المنحولة والمصدر الأصلي الذي أخذت منه، ولما كان مصمّم البرنامج إنسياً فقد فَكَّرَ كما يُفَكِّرُ الْبَشَرُ منطلقاً من فكرة أن الحاسوب قارئ غيبي ويمكن الاحتمال عليه بتغيير كلمة أو تقديم وتأخير... إلخ! وحققاً ذلك؛ فقد تعرّف البرنامج على درجات كبيرة من التشابه مع اختلافات بسيطة في الكتابة أو زمن الفعل وغير ذلك من محاولات الخداع والتضليل^(٣١)! وحتى إنجاز هذا البحث عام ٢٠١، فإن البرنامج أستخدم على مدى سنتين، وتعرّف على ٢٦ ألف حالة؛ لتقييمها، وهي نسبة ضئيلة تمثل أقل من ٣٪ من عدد المتقدمين سنوياً^(٣٢).

الظاهر لنا أن ثمة علاقة طردية بين مدى التشابه الذي يمكن الكشف عنه من ناحية، وتطوير عمليات المعالجة وتنويعها من ناحية أخرى؛ فزيادة سرعة المعالجة ستفضي إلى الكشف عن نسبة أعلى من التشابه، كما أن تطوير أدوات المعالجة نوعياً ستفضي إلى مزيد من القدرة على كشف عمليات التحايل أو محاولات تضليل الحاسوب، فمثلاً، إذا اقتدر المعالج على ربط المفردات بمرادفاتها؛ فإنه سيسهل عليه كشف عمليات التحايل السطحية القائمة على استبدال كلمة بمرادفها؛ كأن يفهم الحاسوب أن كلمة (سُمْعَة) مرادفة لكلمة (مَكَانَة) و(مَنْزِلَة) في قولنا: إن لهذه الجامعة مكانة / منزلة / سُمْعَة مُحَرَّمَة في بريطانيا والعالم.

ولاشكّ عندي بأن هذا البرنامج قد تطوّر كثيراً الآن وصار بإمكانه النظر في نصوص ترجع إلى مدة أطول: عشر سنوات خَلَّتْ، ومعالجتها على نحو أسرع: ٣٥٠ مقالة في عشر دقائق -مثلاً-، وغير ذلك من التحسّن والتطوير كمّاً ونوعاً.

٨-٣- قَتْلٌ مِنَ اللسانيات النفسية:

لعلّ إسهامات المدونات ولسانياتها في اللسانيات النفسية التطبيقية ما تزال في بدايتها نظراً لكون البحث اللساني النفسي بحاجة إلى مراقبة بشرية حثيثة وتبصُّراً دقيقاً بالظواهر اللسانية، ولا سيما إن تعلّق الأمر بالأطفال، واكتساب اللغة الأم. ومع ما تقدم فإنه يتوقّع لللسانيات النفسية التطبيقية مكانة مستقبلية طيبة بمعونة لسانيات المدونات وأدواتها في معالجة اللغة؛ بل لعل طبيعة اشتغال اللسانيات

النفسية تفرض ابتكار أدوات معالجة جديدة ومناسبة للبحث اللساني النفسي تُسَجِّلُ وتَرَصِّدُ وتُسْتَنْتِجُ؛ فكيف يمكن أن تُسهِمَ لسانيات المدونات في اللسانيات النفسية التطبيقية؟

لَمَّا كانت عناية اللسانيات النفسية تنصرف في جوهرها إلى قضايا اكتساب اللغة وتعلمها وتمثيلاتها الذهنية؛ فإنه يبدو مناسباً اتخاذ مثال من اكتساب أبناء اللغة لغتهم الأم ومضاهاتها بتعلم نظرائهم أطفال اللغات الأخرى اللغة نفسها؛ كأن نقارن اكتساب أطفال العرب لغتهم الأم بتعلم أطفال آخرين أجنب للعربية؛ فكيف نفيذ من هذا الاكتساب في تعليم الأطفال الأجانب اللغة العربية؟

تقتضينا هذه المقارنة والمقاربة بناء مدونة مستقلة للغة الطفل العربي تكون منطلقاً أساسياً لدعم عملية تعليم منظّمة ودقيقة للأطفال غير الناطقين بالعربية أو الأطفال العرب خارج أوطانهم، وتهدف هذه المدونة التأسيسية إلى رَصْد أداء الأطفال العرب في عدد من البلاد العربية؛ لتحصيل معلومات متنوعة. ولا شك بأن قوام هذه المدونة سينطلق من ذخيرة الطفل اليومية التي يكتسبها من المنزل مُعَزَّزَةً بالفصحى منها الذي يتلقاه في المدرسة، فيكون رصد الأداءات جامعاً بين العامية القريبة من الفصحى والفصحى التي يتلقاها الطفل في المدرسة مع سنوات التعلم الأولى (حتى الصف السادس مثلاً). وبناءً على ذلك، ستتألف المدونة من المفردات والتراكيب اليومية التي يتفاعل بها الطفل (تلقياً وإنتاجاً) في بيته، ومع أصدقائه، وفي صفّه، وفي كتبه التعليمية، وسيفضي بنا هذا الرصد ثم المعالجة الحاسوبية إلى نتاج متنوعة منها:

- رَصْد المفردات والتراكيب الشائعة وتواترها وسياقاتها الوظيفية (البيت، المدرسة، الكتاب التعليمي).
- ترتيب السياقات الأشيع وفق تكراراتها.
- مدى التعرُّض وسياقاته؛ فتواتر المفردات يكشف عن مدى التعرُّض لكل مفردة وسياقاتها، وتصنيفها وفق نطاق التفاعل والتعرُّض: استقبالي وإنتاجي.
- بناء المتلازمات اللفظية وأنواع هذه المتلازمات ودرجات تواترها.
- تحديد الإستراتيجيات الذاتية في بناء الكفاية اللغوية في أثناء التفاعل الصفي أو مع الأقران: اللعب، حديث النفس، حوارات عائلية، القراءة منفرداً... إلخ.
- الكشف عن ترتيب اكتساب المفردات والتراكيب وأولوياته.
- الكشف عن مدى إسهام الكتاب التعليمي في اكتساب المفردات والتراكيب والأنماط اللغوية المتنوعة.

وكما ذكرنا في غير مكان من هذا البحث؛ فإن هذه النتاج إنما هي مادة خام لا بد من استثمارها في مجالات اللسانيات التطبيقية^(٣٣)، والمجال المناسب هنا هو تعليم العربية للأطفال لغة ثانية، ولا سيما في بناء مناهج العربية لغة ثانية للأطفال، وأحسب أن هذا الإسهام يتمثل في:

- بناء تصوّر منهجي محدّد لكيفية استثمار التعرُّض المكثّف في الكتاب التعليمي؛ وذلك تعويضاً للأطفال عن التعرُّض الطبيعي الذي يفقدونه في البيئات غير العربية.

- بناء تصور منهجي لبناء الكتاب والمواد التعليمية، يقوم على التنوع في أساليب تقديم المحتوى والتمرينات المبنية عليه.
- بناء تصور منهجي يُساعد الأطفال على توظيف استراتيجيات تعلُّم معيَّنة، تمكّنهم من بناء قدراتهم الذاتية؛ للتقدم المستمر في التعلُّم، ومباشرة مصادر المعرفة المناسبة لأعمارهم.
- بناء تصور منهجي دقيق، يستثمر النظريات اللسانية النفسية في رسم مخططات مناسبة لإدارة التعلُّم الصفي، بما يضمن التفاعل المتنوع، ويعلي من شأن الممارسة اللغوية النشيطة.
- بناء برامج ممارسة اجتماعية متنوعة خارج الفصول الدراسية، بما يعزّز فرص الاكتساب، كما في برامج الانغماس اللغوي أو الاختلاط مع عائلات عربية... إلخ.

وأحتسب بالقول إن هذا التصوّر لن يجعل الأطفال متعلمي العربية لغة ثانية بمستوى الأطفال العرب؛ فثمة فرق يظل حاضراً بين البيئة الطبيعية والبيئة المصنوعة لتعلم اللغة، ولكنّه سيرقى بكفائتهم ويمدّهم باستراتيجيات تعلُّم مناسبة تُقصر مدة التعلُّم وتزيد مدى الممارسة وتنوعه.

٩-٣- مثل من البحث اللساني:

ناقش عدد من اللسانيين والمشتغلين بالمدونات ولسانياتها دور المدونات في البحث العلمي اللساني؛ وكان مدار البحث على اتجاهين يمثلان اتجاهات البحث اللساني؛ أما الأول: فيتمثل في إمكانية أن تكون المدونات أساساً لاستنباط النظريات اللسانية أو الرؤى اللسانية بالنظر في الأداءات اللغوية الفعلية التي تتضمنها المدونة، وأما الثاني: فيتمثل في استقراء المادة التي تحتويها المدونة؛ للثبوت من فرضية لغوية مُسبّقة؛ فتكون وظيفة المدونة إما تصديق الفرضية ودعمها وتعزيزها بإضافات جديدة، أو إبطال الفرضية وإضعافها بمزيد من الأدلة المبنوثة في المدونة.

ولعلّ الاتجاه الأول أضعف من الثاني، ويكاد أن يقتصر على النّظر الدلالي والتلازم اللفظي في المستوى الأفقي، وهو ما تعارف عليه لسانيو المدونات بنظرية الفيرثيين الجدد؛ إذ استثمرت المدونات وأدواتها لمزيد من البحث في الدلالة السياقية بالنظر في المتلازمات وقضايا دلالية أخرى، ولا يحضرنى -بعد النظر- فرضيات أو نظريات جديدة انبثقت من معالجة المدونات أو أسفرت عنها معالجة متون المدونات على تعدد أنواعها.

على أنني سأفترض رغبة بعض المشتغلين باللسانيات النفسية التطبيقية في النظر إلى التأثير المتبادل بين اللغة الأم واللغة الثانية على نحو مختلف؛ لاختبار الفرضية السائدة أو الاعتقاد النمطي السائد بأن اللغة الثانية تؤثر سلباً على اللغة الأم في سنّ الطفولة؛ فكأنما دماغ الطفل إناء فارغ ينبغي أن يمتلئ باللغة الأم؛ فإنّ تعلّم الطفل لغة ثانية نازعت هذه اللغة، اللغة الأم واحتلت بعض مواضعها في الدماغ! لقيت هذه الصورة النمطية رفضاً لدى بعض اللسانيين النفسيين المشتغلين باكتساب اللغة، وتمثيلات ثنائي اللغة الذهنية؛ وأبرز هؤلاء (Istvan Keskcs) صاحب نظرية «اللغة الثنائية»: (Dual Language)؛ إذ يرى أن اللغتين تشتركان في مستقر معرفي ذهني واحد، ومن ثمّ فإنّ كلتا اللغتين

تؤثر إحداهما في الأخرى، وأن اللغة الثانية تؤثر إيجابياً في اللغة الأم؛ وقدّم هو وغيره من الباحثين أدلة وشواهد على تأثيرات إيجابية للغة الثانية في اللغة الأم، ومستصفى هذا التأثير، في الغالب وضمن ظروف محددة، أن الطفل ثنائي اللغة يتفوق على أحادي اللغة بأثر من اللغة الثانية^(٣٦)؛ فكيف تُسهم لسانيات المدونات في إثبات هذا التأثير الإيجابي، وتقديم دعم إجرائي وميداني لتفوق ثنائي اللغة على أحاديها؟

ينبغي لنا لتحقيق أهداف مدونة (الطفل ثنائي اللغة):

- إدخال بيانات نظامي اللغتين: الأصوات، والنظام الصرفي، والنظام النحوي، والأساليب اللغوية... إلخ؛ ليكون مرجع التحليل.
 - النظر في الدراسات السابقة المحدودة التي حاولت تبين أثر اللغة الثانية في لغة الطفل الأم؛ وذلك للاستفادة من أدوات البحث التي استخدمتها.
 - تصميم بطاقة بيانات لكل نص تتضمن: عُمر الطفل، واللغة الأم، واللغة الثانية، ومدى التعرّض... إلخ.
 - نموذج للأخطاء المتوقعة في لغة الطفل الأم.
 - بناء نموذج التحليل المعتمد أو نماذج تحليلية متنوعة، مثل: النظام الصوتي، النظام النحوي، المفردات والتراكيب، الرسم والإملاء... إلخ.
 - إدخال نماذج محللة جاهزة يدوياً، أنجزها لسانيون متخصصون؛ ليسهل على النظام بناء افتراضاته والتدرب عليها.
 - بناء مدونة نصية من أداءات الطلاب الكتابية والشفوية.
 - تفعيل أدوات التحليل وإعمالها في نصوص المدونة.
- ولعلّ نتائج التحليل تظهر لنا تأثيرات متفاوتة، ونظراً لظروف متنوعة؛ فقد تُظهر النتائج:
- أن هناك تأثيراً سلبياً للغة الثانية في مهارات الكتابة.
 - أن هناك تأثيراً إيجابياً واضحاً في طريقة تعبير الطفل عن أفكاره وبطرق تجمع بين اللغتين.
 - أن اللغة الثانية تُيسّر عمليات القراءة وفهم النص.

وقد يُظهر مزيد من النّظر في هذه المستخلصات، وربطها بمتغيرات بطاقة النص؛ تفسيرات مهمة، مثل:

- إن تأثير مهارات كتابة الطفل سلباً باللغة الثانية مرجعه إلى ثنائية جائرة تحتفي باللغة الثانية على حساب اللغة الأم؛ وبذلك فإنّ تعرّض هذا الطفل (الصّبي) تعرّضاً مكثّفاً للغة الثانية؛ هيّمن على أعراف الكتابة باللغة الأم وتقاليدها؛ فتعالت كتابته باللغة الثانية، وتهاوت باللغة الأم.
- إن التأثير الإيجابي في مهارات التعبير الشفوي والكتابي مرجعه إلى تعرّض متوازن بين اللغتين، ومحتوى لغوي معرفي متوازن أيضاً؛ يسر للطفل أن يرى العالم بطرق مختلفة وبأساليب تعبير وتفكير متعددة؛ جعلته يتفوق على أحادي اللغة الذي يرى العالم بمرآة واحدة هي اللغة الأم.

- إن تفوق الطلبة الأتراك في القراءة وفهم النص متصل بالمفردات التركيبية ذات الأصل العربي؛ فهذا التشابه أثر تأثيراً إيجابياً كبيراً في قدرتهم على فهم النصوص وتوظيف إستراتيجيات التخمين القائمة على الجذر وتخمينه ومقارنته.

وظاهر لنا أن هذه النتائج تدعم فرضية «الكفاية المتعددة» وفرضية «اللغة الثانية» والتأثير الإيجابي للغة الثانية في اللغة الأم ضمن ظروف محدّدة على التقريب؛ أهمها: الثنائية المتوازنة والتعرض المتوازن، والتشابه بين اللغة الثانية واللغة الأم، ولا سيما في النظام الكتابي والمفردات.

وأما الاتجاه الثاني، فهو الأغلب بالنظر إلى حضور النظريات والفرضيات، وحاجتها إلى مزيد من البحث والتوسيع أو إعادة النظر بمزيد من التعميم أو التخصيص أو التقييد... إلخ. ولعله يحسن أن نتخذ مثلاً من نظرية «الفترة الحرجة». ومعلوم أن العلماء لم يتفقوا على رأي واحد واضح في الموضوع؛ فبعضهم يؤكّد وجودها، وبعض هؤلاء ينقسمون على أنفسهم بين مؤيد تأييداً كاملاً لوجودها، ومؤيد جزئي يراها تتحقق في عناصر اللغة على تفاوت؛ فثمة، عندهم، فترة حرجة لاكتساب النظام الصوتي، وأخرى للنظام النحوي، وثالثة للنظام الصرفي... إلخ. وينكر آخرون فكرة الفترة الحرجة، ويستدلون لذلك بأدلة من النحو الكلي نظرياً، وتطبيقاته عملياً.

ولعله يُمكن الاستفادة من مدونة متعلمي اللغة الثانية الناجزة سلفاً^(٢٥)، ولا سيما إن كانت المدونة مقتصرة على من تجاوزوا الفترة الحرجة (١٨ عاماً فما فوق). ولا شك بأن العمل على هذه المدونة سيكشف لنا أدلة إيجابية وأدلة سلبية من أداءات المتعلمين؛ فالأدلة الإيجابية تُضعف الفرضية؛ نظراً لأن المتعلم تجاوز العمر المحدّد ونجح في اكتساب خصائص وعناصر لغوية مهمة، وأما الأدلة السالبة؛ فإنها تُقوّي الفرضية من حيث إن المتعلم لم يتمكّن من اكتساب بعض أنظمة اللغة وعناصرها اكتساباً مقنعاً ولا سيما إن كانت العناصر التي قُصّر عنها المتعلم جوهرية وتأسيسية كالنظام الصوتي.

على أنه يمكن لنا البحث عن مزيد من الاطمئنان ببناء مدونة أخرى تقتصر على أداء متعلمي اللغة الثانية نفسها من الأطفال (حتى سن ١٦ سنة مثلاً). ولا شك بأن دراسة مدونة الأطفال ستكون أكثر واقعية وموضوعية لاختبار فرضية الفترة الحرجة؛ لأنها ستتركز على أداء الأطفال في فترة النشاط الذهني العالي المدعومة بمقولة شائعة مفادها: «إن الطفل أقدر على اكتساب اللغة الثانية في طفولته، وإنه قادر على اكتساب أكثر من لغة مع لغته الأم».

وتنتهي بنا المقارنة بين المدونتين إلى نتائج تدعم الفرضية أو تفنّدها كلياً أو جزئياً، وهي:

- تفوق الأطفال على البالغين في اكتساب لكمة اللغة الثانية بنسب كبيرة واضحة يؤكّد أن اكتساب الطفل لنظام اللغة الثانية الصوتي يكون أجودّ وأكفأ من الراشدين، وهذا يعني مزيداً من شواهد وجود فترة حرجة لاكتساب النظام الصوتي.
- وعي متعلم اللغة الراشد بالنظام النحوي والتركيب للغة الثانية وتمكّنه منه يُضعف فرضية الفترة الحرجة للنظام النحوي والتركيب من ناحية، ويُعزّز القول بالفرق بين الاكتساب الضمني والتعلّم المقصود من ناحية أخرى.

بل لعلَّ المقارنة الدقيقة تكشف لنا عن فروقات أدقَّ بين الأطفال والراشدين؛ كاكْتِسَابِ الأبعاد التداولية أو الخصائص الصوتية الدقيقة (المخارج والألفونات...).

١٠-٣- مَثَلٌ من التطبيق المتكامل:

بناءً على ما تقدّم كله، فإنه لا مجال للمجادلة في أن لسانيات المدونات صورة متقدمة لمركزية اللسانيات الحاسوبية في العلوم البينية المعاصرة؛ إذ تمثّل هذه المدونات ومعالجاتها القلب الذي يغذي مجالات اللسانيات التطبيقية الأخرى بالمحتوى وأدوات المعالجة؛ فيقدم خلاصة معالجاته ناجزة للعلوم اللسانية الأخرى؛ فقد يعتمد عليها اعتماداً كلياً، وقد يحتاج إلى معالجات إضافية؛ لتخصيص المخرجات اللازمة لكل فرع.

وفيما يلي فَصْلُ بيانٍ وتوضيح:

لقد أظهر النظر أن أبهى تطبيقات لسانيات المدونات يُنَجَرُ في مجال تعلم اللغة الثانية واكتسابها؛ إذ يمكن بناء مدونة لغة المتعلّم بصورتها الخام قبل أن تخضع للمعالجة الحاسوبية، وهكذا تكون لغة المتعلم الحافظة التي تُخَزَّنُ فيها نصوص المتعلمين التي تنهي للمعالجة الآلية، وتفضي بنا هذه المعالجة الآلية إلى مدونات فرعية متعددة تتخذ منها مدونة أخطاء المتعلمين مثلاً لتطبيقاً.

إن تحليل أخطاء المتعلمين يقوم على تحليل حصيلة كبيرة من إنتاج الطلاب، ولا سيما إنتاجهم الكتابي المضمّن في المدونة، وينتهي بنا تحليل هذه الأخطاء إلى مجموعة من المخرجات، منها:

- رصد أخطاء متعلمي لغة أجنبية ما، على اختلاف جنسياتهم وخلفياتهم اللغوية.
- تصنيف الأخطاء بالنظر في المستوى اللغوي الذي وقعت فيه (صوتي، صرفي، تركيبي، نصّي... إلخ).
- المقارنة بين هذه الأخطاء المرصودة؛ للتوصل إلى الأخطاء العامة المشتركة بين جميع المتعلمين، والأخطاء الأقل عمومية.
- الرصد الدقيق للأخطاء التي يرتكبها أبناء خلفية لغوية محدّدة.

والظاهر لنا من الأول أن البيانات المتوافرة بعد تحليل الأخطاء، يمكن توظيفها في مجالات لسانية متعددة، يقع بعضها في إطار تعليم اللغة الثانية واكتسابها، مثل: بناء المناهج وإعداد الاختبارات، وبعضها يقع في مجالات لسانية تطبيقية أخرى.

وهذا مثال مؤلّف يُقدّم صورة تؤكد مركزية لسانيات المدونات في اللسانيات التطبيقية، بوصفها مركزاً أساسياً في اتخاذ القرارات، ومن ثمّ تحقيق مبتغى اللسانيات التطبيقية في أن تقدم حلولاً للمشكلات اللغوية، والتحول عن الأوضاع غير المرغوبة، وينطلق هذا المثال من نتائج تحليل أخطاء متعلمي اللغة الثانية.

سيفضي بناء مدونات فرعية تحتوي أخطاء المتعلمين إلى فرز الأخطاء وتصنيفها بين العام المشترك وما يختص به متعلمون من خلفيات لغوية معينة، ولا شك بأن الأخطاء الخاصة غير المشتركة، تسهم إسهاماً كبيراً في استثمار هذه المدونة في اللسانيات الجنائية، ولا سيما في قضايا الإرهاب الدولي؛ فمثلاً، إذا

بُنِيَ تركيباً مدونة تستنفذ جميع الأخطاء التي يرتكبها متعلمو العربية الأتراك؛ فإن هذا سيسهل عليها تحديد إن كان مرتكب جريمة ما، عربياً أم تركياً أم من بلاد أخرى على التقريب. وتزداد الثقة بنتائج عمل هذه المدونات، ما تعاضم عدد النصوص التي اشتملت عليها؛ إضافة إلى تقدّم برامج المعالجة الحاسوبية، ومهارة اللسانيين المشتغلين بالقضية، واحترافهم التحليل اللساني بأبعاده المتنوعة.

فإذا تجاوزنا المدونات الضيقة إلى مدونات أوسع؛ كان لنا سند قوي من نتائج تحليل المدونات في الفصل في قضايا الإرهاب الدولي؛ فإذا كانت مدونة أخطاء المتعلمين تتجاوز بلدًا واحدًا وجنسية واحدة، وصممت لها برامج تحليل حاسوبي دقيقة، أمكن لنا تعيين جنسية المجرم على وجه التقريب.

مثال ١:

إذا أفضت برامج التحليل الحاسوبي إلى أن متعلمي العربية لغة ثانية من خلفية لغوية محددة يتفردون بأخطاء لا تظهر عند غيرهم، ولا سيما في النصوص المكتوبة، كما في الجدول الآتي^(٣٦):

الجنسية	الأخطاء التي تفرّدت بها
أ	قلب (ظ) (ز) نطقًا وكتابة. قلب (ح) (خ) نطقًا وكتابة. قلب (ع) همزة نطقًا وكتابة.
ب	كتابة الحركات القصيرة؛ طويلة. تعريف أسماء الأعلام. تقديم الصفة على الموصوف.

إن هذه النتائج تكون مفيدة إلى حد بعيد في تحديد جنسية مرتكب جريمة تهديد إرهابي أو ما شابه، وإذا انتهينا إلى عمليات حصر دقيقة، صار ثمة يُسرُّ في تقدير جنسية المتهَم أو الظَّنِّين.

ولعلّ مثل هذه النتائج الدقيقة تكون أداة لبناء اختبارات لغوية دقيقة تركز على أخطاء المتعلمين في اختبارات تحديد المستوى؛ ليكون الاختبار موجّهًا لقياس أمرين: أولهما: دقة نتائج التحليل التي توصلت إليها المعالجة من ناحية، وثانيهما: توجيه جزء مهم من الاختبار إلى الأخطاء الشائعة والصعوبات المتوقعة.

ويمكننا الانطلاق من النتائج نفسها التي أفضى إليها تحليل الأخطاء؛ للكشف عن مؤلفي النصوص ومنشئيه في طلبات القبول والاختبارات وسواها؛ فإذا افترضنا أن إحدى الجامعات العربية قد هيأت ضمن طلبات القبول في برامج اللغة العربية أن يكتب الطالب نصًّا متوسطًا في حدود ألف كلمة يوضّح فيها أسباب رغبته بتعلم العربية في الإمارات، ثم عالجَتْ هذه النصوص المكتوبة وفق برنامج معالجة مخصص لقياس مستوى المتعلم في اللغة العربية (أول، ثان، متقدّم)، والتوثّق من مطابقة بياناته التعريفية التي أدلى بها ومستواه الفعلي، ثُمَّ ظهر للجامعة، مثلاً:

- أن الطالب (س) لم تظهر لديه أي أخطاء من الأخطاء الشائعة بين أبناء بلده.
- أن مجموعة من الطلبة الذين يتحدثون اللغة الفرنسية لغة أم؛ أظهروا تمكُّنًا واضحًا من مهارات الكتابة، ولم تظهر لديهم أي أخطاء، ولا سيما تلك التي تشيع بين الفرنسيين؛ كأنما كاتب النص عربي متمكّن.

إنّ تدقيق النظر في هذه النصوص وتحليلاتها يفضي إلى ظنٍّ كبير، مفاده: أن كاتب النص شخصٌ عربيٌّ ساعدَ هذا الطالب لسبب ما، ولا سيما إن تعلّق الأمر بأعطاب الخطاب وإستراتيجيات كتابة النص العربي؛ ذلك أن «لكنة الخطاب» تظل ملازمة للمتعلم الأجنبي زمنًا طويلاً، لعله يستمر أبداً الدهر!

٤- خاتمة:

ارتاد هذا البحث أفاقاً متنوعة في قضايا المدونات ولسانياتها؛ إذ قدّم في جميع جوانبه أفكاراً طريفةً تصلح أن تكون موجّهات بحثية للباحثين وطلبة الدراسات العليا، في الدراسات اللغوية والاجتماعية والقانونية والثقافية، ولعلها كانت موجّهات لطيفاً للمشتغلين بالتحليل النقدي للخطاب في الثقافة العربية؛ أولئك الراغبين في إحداث تغيير نوعي في مجتمعاتهم وتوجهاتها السياسية والاجتماعية والاقتصادية؛ فكانت بذلك جولة افتراضية في مسالك ذهن اللغوي وإمكانية استثمار المدونات ولسانياتها في تحويل الافتراضي إلى واقعي.

وقد أكّد البحث نظرياً وتطبيقياً أن المدونات أداة مركزية مساندة للسانيات التطبيقية؛ إذ تُعِينها في الوصول إلى تعميمات أدقّ بناء على عينات ضخمة من الأداءات اللغوية الواقعية، وتؤسس لسانيات المدونات لعلاقة شراكة جوهرية ثلاثية الأضلاع: المدونات ولسانياتها، واللسانيات التطبيقية ومجالاتها، والسياسة اللغوية والتخطيط اللغوي؛ وذلك كله ينتهي إلى إنجاز آمال اللسانيات التطبيقية وأحلامها بحل المشكلات اللغوية والتحوّل عن الأوضاع اللغوية والاجتماعية والثقافية والسياسية غير المرغوبة.

وتوصل البحث إلى النتائج التالية:

- المدونات أداة منهجية مُساندة، تُسهّم في الوصول إلى تعميمات أدقّ في البحث اللساني؛ نظراً لحجم عينات البحث والمواد اللغوية الأصيلة المدروسة.
- لسانيات المدونات مرتكز منهجي في التخطيط اللغوي، ورسم السياسات اللغوية الرشيدة؛ إذ تمده ببيانات ومعلومات دقيقة تُيسّر عمل المخططين، وراسمي سياسات الدول اللغوية.
- لسانيات المدونات علم نظري يمثّل مرتكزاً أساسياً في اللسانيات التطبيقية، وطموحاتها في حل المشكلات اللغوية وما يترتب عليها من مشكلات تربوية واجتماعية واقتصادية... إلخ.
- تسهم لسانيات المدونات في تعميم جدوى اللسانيات التطبيقية، ومنزلتها في المجتمع المعاصر؛ فهي تُسهّم في تقديم رؤى نافعة في التحليل النقدي للخطاب، وتحليل أخطاء متعلمي العربية لغة ثانية... إلخ.

٥- الهوامش:

١. تحيَّرت هذه الدراسات وفق معياري الحدائث والتنوع.
٢. ينظر: العصيمي، صالح (٢٠١٥). المدونات اللغوية بناؤها وطرائق الإفادة منها. ط١. مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية، الرياض.
٣. العناني، وليد. (٢٠١٧). المدونات وتحليل الخطاب، مجلة أفكار، وزارة الثقافة الأردنية، العدد ٣٣٦.
٤. هذه الفوائد من استخلاص الباحث، وبعضها ورد بمعناه في كتاب:
Jeanne McCarten, (2007). p3. ولتفاصيل إضافية عن استثمار المدونات في تعليم مفردات اللغة الأجنبية، انظر: (2000). Schmitt. N, (2007). R. Carter.
٥. العناني، وليد. (٢٠٠٩). مفردات العربية دراسة لسانية تطبيقية في تعليمها للناطقين بغيرها. المؤتمر العالمي لتعليم اللغة العربية لغير الناطقين بها. معهد اللغة العربية. جامعة الملك سعود. ٢-٣/١١، ص: ٥٢٣.
٦. Schmitt. N, (2000). p89.
٧. لمزيد من المعلومات والشروحات ينظر: العصيمي، صالح. (٢٠١٨). لسانيات المتون: قضايا أساسية في التأصيل والتطبيق والمنهج. ط١. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية. الرياض.
٨. هذا المثال مستوحى من بحث أستاذنا المرحوم نهاد الموسى «النحو العربي بين النظرية والاستعمال.. مثل من باب الاستثناء».
٩. ينطبق مفهوم الاستعمال الواقعي على القرآن الكريم والحديث النبوي الشريف، وجميع الشواهد الجارية على القواعد الشائعة، مهما يكن النص قديماً.
١٠. الترادف ظاهرة لغوية كلية؛ فما يصدق على العربية يصدق على غيرها من اللغات.
١١. وفق مفهوم التعلم المباشر (الشكلي) ونظرية الملاحظة في تعلم اللغة الثانية.
١٢. ليس هذا موضع تفصيل لكيفية تحقق ذلك لدى متعلم اللغة الثانية، بناء على نظريات اكتساب اللغة الثانية؛ فالتعرض المكثف المتنوع يساعد المتعلم على بناء فرضياته حول استخدام المفردات، وكلما زاد التعرض وتنوع أعاد المتعلم بناء فرضياته وتثبيتها.
١٣. هذا موضوع يبيني مشترك بين تحليل الخطاب واكتساب اللغة وتعليمها. ولمزيد من المعلومات حول الخطاب الصفي ينظر:
١٤. Walsh.S. (2006), Investigating Classroom Discourse.
١٥. هذه البيانات هي معلومات (التَّحْشِيَّة والتوسيم) لاحقاً.

١٦. مفهوم حل المشكلة في اللسانيات التطبيقية يعني وجود وضع لغوي غير مرغوب فيه، ولا بد من التحول عنه؛ فعدم معرفة الفرد بلغة أجنبية مشكلة يمكن حلها بتعليم هذا الفرد اللغة الأجنبية، وحاجة لغة ما، إلى مصطلحات جديدة مشكلة أخرى بحاجة إلى حل وتحول نحو وضع آخر أفضل... إلخ.
١٧. مثالان افتراضيان، ولكنهما متوقعان إن أجري البحث وفق المثال المضروب.
١٨. يمكن أن ينطبق هذا البحث على أي لغة وأي خطاب صفي في أي مجتمع.
١٩. مرادف المصطلح (Critical Discourse Analysis) ومختصره CDA. وقد أثرت هذا المرادف على (تحليل الخطاب النقدي)؛ هرباً من الالتباس بالتحليل المعني بالخطاب النقدي. ومن أمثله: ماونتر، جيرلند. (٢٠١٤). كيف تضيف المدونات اللغوية إلى التحليل النقدي للخطاب، فوداك، روث وماير، ميشيل (محرر) مناهج التحليل النقدي للخطاب، ترجمة: حسام فرج وعزة شبل محمد. ط١. المركز القومي للترجمة، سلسلة العلوم الاجتماعية للباحثين الكتاب ٢٠٨٨، القاهرة.
٢٠. وظّف جيرلند ماونتر المدونات في بحثه «كيف تضيف المدونات اللغوية إلى التحليل النقدي للخطاب» لدراسة كلمتي (عاطل، كادح) في الصحف البريطانية للنظر في دلالات الكلمتين الاجتماعية والاقتصادية في المجتمع، ينظر: من أمثله في العربية: بحث سلطان الميجول «المعالجة الآلية للصحف العربية: تحليل الأنماط الخطابية بمناهج BCU» ينظر: الميجول، سلطان. (٢٠١٦). لغويات المدونة الحاسوبية. ط١. مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية، الرياض.
٢١. غالباً ما تظهر هذه الخطابات في الانتخابات، وقد بدا هذا الموضوع متجاذباً في الانتخابات التركية الأخيرة ٢٠٢٣، حين كان موضوع اللاجئين السوريين محوراً مركزياً ووعداً قاطعاً للمعارضة!
٢٢. مزيد من المعلومات عن منزلة المفردات في الخطاب، ومنزلتها في دلالة الخطاب، ينظر: العناتي، وليد. (٢٠١٨). الخطاب والتعليم. دراسات في تحليل الخطاب وتعليم العربية للناطقين بغيرها، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.
٢٣. هذه بعض نتائج بحث (غوغل) عن كلمة (اللاجئين).
٢٤. مَثَلٌ جَيِّدٌ في كتاب براين بارتلج، ص: ٢٤٣، وثمة أمثلة من مجالات أخرى في الفصل التاسع.
٢٥. الثوابية، محمد. (٢٠١٥). عثرات الخطاب المكتوب عند غير الناطقين بالعربية، رسالة ماجستير. عمادة البحث العلمي، جامعة البترا. عمان.
٢٦. Jack C. Richards and Richard Schmidt, (2002), Longman Dictionary of Language Teaching & Applied linguistics, p:261
٢٧. هذا تمثيل لبعض أنواع الفنون الكتابية التي ينتجها الناطقون بالعربية ومتعلموها لغة ثانية.
٢٨. ينظر: العناتي، وليد. (٢٠١٧). المدونات وتحليل الخطاب، مجلة أفكار، وزارة الثقافة الأردنية، العدد ٣٣٦.

٢٩. مزيد من المعلومات في:

Janet Cotterill, how to use corpus linguistics in forensic linguistics, in: Anne O'Keeffe and Michael McCarthy (edit). (2010), The Routledge Handbook of Corpus Linguistics, pp:587-59.

٣٠. تفاصيل دراسة الحالة هذه وبيانات أخرى في بحث:

(David Wools): Computational Forensic linguistics, Searching for similarity in large, specialized Corpora, in: Malcolm Coulthard and Alison Johnson (edi). (2010), The Routledge Handbook of Forensic Linguistics, Routledge, pp:576-590.

والحقيقة أن كثيراً من المكاتب تقوم بهذه الأعمال في هذه الأيام إضافة إلى خدمات أخرى!

٣١. يتساءل كثيرون عن تفوق برامج الكشف عن السرقات والانتحال باللغة الإنجليزية عليها في اللغة العربية؛ ولا شك أن مرجع ذلك إلى حجم المصادر المنشورة بالإنجليزية ويمكن الاعتماد عليها في المقارنة، وهذا يعني أن ضعف المحتوى العربي المنشور على الشبكة هو السبب في ذلك، وكثيراً ما تُكتشف سرقات رسائل ماجستير ودكتوراه بعد سنوات طويلة مع أن الرسالة خضعت للكشف في برامج مثل (تيرن إت إن)، والسبب هو أن الرسالة الأصلية المسروقة ليست منشورة أو متاحة على الشبكة، ولذلك لن يتمكن البرنامج من الإحالة عليها.

٣٢. كُنّا في إحدى الجامعات الأردنية نستخدم برنامج (تيرن إت إن)؛ للنظر في بحوث الزملاء المقدمة للترقية؛ للتثبت من أصالتها وأنها لا تتضمن أكثر من ٢٠٪ من الاقتباسات، وكان من نقائص ذلك البرنامج أيامها عدم تنبيه المبرمجين إلى قضايا جوهريّة في البحث العلمي، ومنها: المراجع؛ فقد كان يشير إلى المراجع بأنها منقولة؛ وكما هو معروف فإن المراجع منشورة واردة في معظم البحوث التي تبحث في الموضوع نفسه أو المجال العام نفسه؛ فمثلاً، كتب الجاحظ كانت تظهر في بحوث النقد القديم، وكان البرنامج يدلّ عليها بأنها انتحال! ومن ذلك أيضاً، أنه كان يعدّ النصوص الشعرية أو الآيات القرآنية موضع التحليل والنظر نصوصاً منقولة! ولذلك لم يكن مستغرباً أن تظهر نسبة الاقتباس أحياناً ٦٠ أو ٧٠٪، ولست متيقناً كيف صار الحال الآن، ولكن لا شك بأن زيادة المحتوى العربي العلمي المنشور يزيد من فعالية هذه البرامج.

٣٣. ليس لدي أي معلومات عن هذا البرنامج؛ إذ لم يُذكر اسمه في البحث، ولكن بلا شك يمثل نواة مبكرة لبرامج كشف الانتحال.

٣٤. لا شك بأنه يمكن استثمار هذه النتائج في وجوه متنوعة من البحث اللساني النفسي التطبيقي؛ وذلك بتفحص فرضيات ورؤى في اكتساب اللغة الأم وتطورها وما يتعلق بذلك من الفرضيات المشهورة أو فرضيات جديدة ما تزال بحاجة إلى دعم بحثي.

٣٥. لتفاصيل أكثر عن هذا الموضوع، ينظر: كِسْكَس، استفان. (٢٠١٦). «تأثير اللغة الأجنبية في اللغة الأم». ترجمة: وليد العناتي. مجلة تبين، المركز العربي للأبحاث ودراسة السياسات، العدد ١٨ / المجلد الخامس.

٣٦. ينظر: الفيضي، عبدالله، تحليل الأخطاء في المدونة اللغوية لتعلمي اللغة العربية، المجول، سلطان. لغويات المدونة الحاسوبية.

٣٧. ثَمَّ بيانٌ عن تعالق لسانيات المدونات واللسانيات الجنائية في: العناني، وليد. (٢٠٢٣). «مقدمة في اللسانيات الجنائية؛ رؤية لسانية تطبيقية». المجلة العربية للعلوم الإنسانية بجامعة الكويت. العدد ١٦٤.

٦- المراجع العربية:

بارتلج، براين. (٢٠١٨). تحليل الخطاب، ترجمة: عبد الرحمن الفهد. ط١. منشورات مركز الترجمة بجامعة الملك سعود، الرياض.

بيكر، بول. (١٤٣٥). مناهج المتنون في اللسانيات، ليا ليتوسيليتي (محرر)، مناهج البحث في اللسانيات، ترجمة: صالح العصيمي. ط١، معهد الملك عبد الله للترجمة والتعريب بجامعة الإمام محمد بن سعود الإسلامية، ص: ١٧٧-٢١٩.

التكروري، أيمن. (٢٠١٨). المدونات اللغوية ودورها في معالجة النصوص العربية. ط١. مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية. الرياض.

الثوابية، محمد. (٢٠١٥). عثرات الخطاب المكتوب عند غير الناطقين بالعربية، رسالة ماجستير. عمادة البحث العلمي، جامعة البترا. عمان.

ستابس، مايكل. المدونات اللغوية. ديفز، ألين وإلدر، كاثرين (محررون) المرجح في اللغويات التطبيقية، ترجمة: ماجد الحمد وحسين عبيدات. منشورات جامعة الملك سعود. ص ١٠٣-١٢٨.

الشهري، علي غرامة. (٢٠٢٢). لسانيات المدونات: تعريف ومجالات الإفادة. المجلة العربية للعلوم ونشر الأبحاث، المجلد ٨، العدد ٣٠. ص: ٧٧-٨٦.

صالح، محمود إسماعيل. (٢٠١٢). الحاسوب والبحث اللغوي، المدونات اللغوية نموذجًا. ط١. منشورات كرسى بحث صحيفة الجزيرة للدراسات اللغوية الحديثة. جامعة الأميرة نورة بنت عبد الرحمن. الرياض.

العصيمي، صالح. (٢٠١٥). المدونات اللغوية بناؤها وطرائق الإفادة منها. ط١. مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية، الرياض.

العصيمي، صالح. (٢٠١٨). لسانيات المتنون: قضايا أساسية في التأصيل والتطبيق والمنهج. ط١. مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية. الرياض.

العصيمي، صالح والثبيتي، عبد المحسن. (٢٠٢١). الكتابة العلمية المتخصصة في المقترحات البحثية لعلم اللغة التطبيقي: دراسة قائمة على المدونات. مجلة جامعة الشارقة للعلوم الإنسانية والاجتماعية، المجلد ١٨، العدد ٢.

العناتي، وليد. (٢٠٠٩). مفردات العربية دراسة لسانية تطبيقية في تعليمها للناطقين بغيرها. المؤتمر العالمي لتعليم اللغة العربية لغير الناطقين بها. معهد اللغة العربية. جامعة الملك سعود. ١١/٣، ص: ٥٢٣.

العناتي، وليد. (٢٠١٧). المدونات وتحليل الخطاب، مجلة أفكار، وزارة الثقافة الأردنية، العدد ٣٣٦.

العناتي، وليد. (٢٠١٨). الخطاب والتعليم. دراسات في تحليل الخطاب وتعليم العربية للناطقين بغيرها، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

العناتي، وليد. (٢٠٢٣). «مقدمة في اللسانيات الجنائية: رؤية لسانية تطبيقية». المجلة العربية للعلوم الإنسانية بجامعة الكويت. العدد ١٦٤.

العنزي، بدرية بنت براك. (٢٠٢٠). «المعالجة الآلية للتطور الدلالي وفق لغويات المدونة المحوسبة دراسة دلالية حاسوبية». مجلة اللسانيات العربية، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، العدد ١١.

الفيضي، عبد الله. (٢٠٢٣). وسم المدونات اللغوية: المفهوم والمجالات. مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية. المجلد الثامن والثلاثون، العدد الأول، ص: ٢٧٥-٣١٦.

كسكس، استفان. (٢٠١٦). «تأثير اللغة الأجنبية في اللغة الأم». ترجمة: وليد العناتي. مجلة تبين، المركز العربي للأبحاث ودراسة السياسات، العدد ١٨/ المجلد الخامس.

الكشور، رضا. (٢٠٢٢). «توظيف لسانيات المدونات في صناعة المعاجم لغير الناطقين بالعربية». مجلة اللسانيات العربية، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، العدد ١٥/ ١٥٨ - ١٩٤.

كولتارد، كالوم وآخرون. (٢٠١٩). مقدمة إلى علم اللغة الجنائي: اللغة في علم الأدلة، ط٢، ترجمة عبد الرحمن ابن عبد العزيز القرشي، ط١، مركز النشر العلمي، جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية.

ماونتر، جيرلند. (٢٠١٤). كيف تضيف المدونات اللغوية إلى التحليل النقدي للخطاب، فوداك، روث وماير، ميشيل (محرر) مناهج التحليل النقدي للخطاب، ترجمة: حسام فرج وعزة شبل محمد. ط١. المركز القومي للترجمة، سلسلة العلوم الاجتماعية للباحثين الكتاب ٢٠٨٨، القاهرة.

المجيول، سلطان. (٢٠١٦). لغويات المدونة الحاسوبية. ط١. مركز الملك عبد الله بن عبد العزيز لخدمة اللغة العربية، الرياض.

مكزي، توني، وهاردي، أندرو. (٢٠١٦). لغويات المدونة الحاسوبية، ترجمة: سلطان بن ناصر المجيول. ط١. منشورات مركز الترجمة بجامعة الملك سعود. الرياض.

وود، ديفد. (٢٠١٨). مقدمة في أسس المتلازمات اللغوية، ترجمة: عبد الله بن علي الأسمرى. ط١. دار جامعة الملك سعود للنشر، الرياض.



v- المراجع الأجنبية:

- Agali Paquot, Stefan Th. Gries, (2021). A Practical Handbook of Corpus Linguistics, Springer Nature.
- Bethany Gray and Douglas Biber, Corpus approaches the study of discourse, in Ken Hyland and Brian Paltridge. (2011). Continuum Companion to Discourse Analysis, Continuum International Publishing Group, New York NY 10038, pp: 138-152.
- Carter, R. (2007). Vocabulary, Applied linguistic Perspectives, Routledge, London.
- David Woolls, Computational Forensic linguistics, Searching for similarity in large, specialized Corpora, in: Malcolm Coulthard and Alison Johnson (edit). (2010), The Routledge Handbook of Forensic Linguistics, Routledge, pp:576-590.
- Jack C. Richards and Richard Schmidt, (2002), Longman Dictionary of Language Teaching & Applied linguistics.
- Janet Cotterill, how to use corpus linguistics in forensic linguistics, in: Anne O'Keeffe and Michael McCarthy (edit). (2010), The Routledge Handbook of Corpus Linguistics, pp:587-590.
- Marie-Noëlle Guillot. (2002), Corpus-based work and discourse analysis in FL pedagogy: a reassessment, System 30 (2002) 15–32, www.elsevier.com/locate/system.
- McCarten, J. (2007). Teaching Vocabulary, Lessons from the Corpus, Lessons for the Classroom, Cambridge university press. UK.
- Michael McCarthy. (2006), Explorations in Corpus Linguistics, Cambridge university press.
- Schmitt. (2000), Vocabulary in Language Teaching, Cambridge University Press, USA
- Stefanowitsch, Anatol. 2020. Corpus linguistics: A guide to the methodology (Textbooks in Language Sciences 7). Berlin: Language Science Press
- Susan M. Conrad, Kimberly R. Levelle. (2008), Corpus Linguistics and Second Language Instruction, in Spolsky and Hult 2008), The Handbook of Educational Linguistics, Blackwell Publishing,
- TONY MCENERY. (2011), What Corpora Can Offer in Language Teaching and Learning, in: Eli Hinkel,(2011), Handbook of Research in Second Language Teaching and Learning Volume II, Routledge, New York.
- Walsh.S. (2006), Investigating Classroom Discourse, Routledge, New York.

٨- نبذة تعريفية عن الباحث أو الباحثين:

AUTHOR BIODATA: Waleed Alanati is a Full Professor of Applied linguistics in the Department of Arabic, College of Arts, Jordan University. He received his PhD degree in 20000 from University of Jordan. His research interests include Applied linguistics and teaching Arabic as a second language	بيانات الباحث : وليد العناتي، أستاذ في (اللسانيات التطبيقية) في (قسم اللغة العربية) (بكلية الآداب) في جامعة العلوم التربوية والآداب / الأونروا بـ (الأردن). حصل على درجة الدكتوراه في اللسانيات الاجتماعية التطبيقية من الجامعة الأردنية عام ٢٠٠٠. تدور اهتماماته البحثية حول اللسانيات التطبيقية عامة وتعليم العربية لغة ثانية تخصصياً.
--	---

<https://orcid.org/0009-0000-3855-6707>

معرف أوركيد (ORCID) : 

Email: anati_waleed@hotmail.com

نِظَامٌ بَلْقِيسَ: نِظَامُ اكْتِشَافِ وَتَصْحِيحِ اللَّحْنِ فِي الْكَلَامِ الْعَرَبِيِّ الْمُنْطَوِقِ

Belqis: A System to Detect and Correct Audible Grammar Mistakes in Spoken Arabic

أحمد ميلود خرصي

جامعة الإمام محمد بن سعود الإسلامية، الرياض، المملكة العربية السعودية

توثيق البحث APA Citation:

خرصي، أحمد. (٢٠٢٤). نِظَامٌ بَلْقِيسَ: نِظَامُ اكْتِشَافِ وَتَصْحِيحِ اللَّحْنِ فِي الْكَلَامِ الْعَرَبِيِّ الْمُنْطَوِقِ. *اللغويات الحاسوبية والمعالجة الآلية للغة العربية*، العدد ١، ٧٩-٩٠.

Submission Date: 26/05/1445 AH

تاريخ الإرسال: ١٤٤٥/٠٥/٢٦هـ

Acceptance Date: 24/06/1445 AH

تاريخ القبول: ١٤٤٥/٠٦/٢٤هـ

Abstract:

Speech as writing is bound by the language rules to ensure a common interpretation. Although Arabic grammar has been studied and documented since the eighth century, research in Arabic language processing seem to miss much of it. Unlike English, many of the grammar mistakes in Arabic are audible rather than visible. A sentence like "كتب الطالب الدرس" should be pronounced «kataba ttaAlibo ddarsa» "The student wrote the lesson": the translation of "كَتَبَ الطَّالِبُ الدَّرْسَ", might also be misinterpreted "the lesson wrote the student" if mispronounced "كَتَبَ الطَّالِبُ الدَّرْسَ" (kataba ttaAliba ddarso). A single diacritic switched the meaning. In this article, we show how we addressed the challenge of audible grammar mistakes. To the best of our knowledge, no approach addresses the Arabic grammar checking to the depth we do. Tests on the grammar checker show 93%+ accuracy 99% recall.

Keywords:

NLP, Arabic language, Diacritics, Automated grammar checking...

الملخص:

في كل لغة، للكتابة والكلام قواعد تضمن الفهم الصحيح لما يقصده الكاتب والمتكلم. ورغم أن مكتبة النحو العربي قد لا يذانيها ما كتب في قواعد أي لغة أخرى، لا تزال الأبحاث في حوسبة اللغة العربية تغفل جوانب مهمة من خصوصيات اللغة العربية. على خلاف لغات أخرى، كثير من الأخطاء النحوية في اللغة العربية لا يُقرأ، بل يُسمع. جملة بسيطة مثل «كتب الطالب الدرس» يذون ضبطاً للحركات - كما هو الغالب في الكتابة العربية - قد تُنطق «كَتَبَ الطَّالِبُ الدَّرْسَ». مثل هذه الأخطاء لا تميز إلا بالسمع. في هذا البحث، نسلط الضوء على تصميم نظام حاسوبي لاكتشاف وتصحيح هذا النوع من الأخطاء. هذا النظام، نفدناه - بحمد الله - في منصة بلقيس التي تضم عدداً من الأنظمة الحاسوبية التي تعتمد الذكاء الاصطناعي لمعالجة المحتوى العربي. نتائج التجارب أظهرت دقة تصل إلى ٩٣٪ واسترجاعاً وصل إلى ٩٩٪.

الكلمات المفتاحية:

المعالجة الآلية للغات الحقيقية، معالجة المحتوى العربي، حوسبة اللغة، الذكاء الاصطناعي، التصحيح الآلي للأخطاء النحوية.

المقدمة:

إنَّ الأبحاثَ في مجال أنظِمةِ التَّصحيحِ الآليِّ للأخطاءِ النُّحويَّةِ والإملائيَّةِ جاريٌّ مُنذُ عُمُودِ (singh & singh, 2020). عندمَا يُفْضَى الخطأُ الإملائيُّ إلى تَرْكِيبِ كَلِمَةٍ لَيْسَتْ مِنَ الْقَامُوسِ، يَكُونُ أَدَاءُ مِثْلِ هَذِهِ الْأَنْظِمةِ دَقِيقًا. وَلَا غَرَابَةَ؛ إِذْ إِنَّ أَفْضَلَ مَا يَقُومُ بِهِ الْحَاسِبُ الْآلِيُّ الْمَقَارَنَةُ الدَّقِيقَةُ بَيْنَ قِيَمَتَيْنِ. تُصْبِحُ الْأُمُورُ أَكْثَرَ تَعْقِيدًا جِئْنَ نَقْلَبُ الْكَلِمَةَ إِلَى أُخْرَى مِنَ نَفْسِ الْقَامُوسِ، مِثْلَ كِتَابَةِ "إِنْ شَاءَ اللَّهُ" خَطَأً "إِنْشَاءَ اللَّهِ". فِي مِثْلِ هَذِهِ الْحَالَاتِ لَا يُمْكِنُ الْإِعْتِمَادُ عَلَى خَوَارِزِمِيَّاتِ الْمَقَارَنَةِ وَالبَحْثِ الْبَسِيطَةِ. تَبَعْدُ الشُّقَّةُ أَكْثَرَ، إِذَا كَانَتْ الْأَخْطَاءُ نُحُويَّةً غَيْرَ إملائيَّةٍ مِنْ مِثْلِ "رَأَيْتُ أَخُو زَيْدٍ". الْأَبْحَاطُ الَّتِي أَطْلَعْنَا عَلَيْهَا إِلَى تَارِيخِ كِتَابَةِ هَذَا الْبَحْثِ فِي مَجَالِ التَّصْحِيحِ الْآلِيِّ لِلأَخْطَاءِ النُّحُويَّةِ فِي النَّصِّ الْعَرَبِيِّ، كُلُّهَا عَدَا وَاحِدًا تَغْفِلُ الْحَرَكَاتِ بِالْكَلْبِيَّةِ، رَغْمَ أَنَّ عُلَمَاءَ النُّحُوِّ إِنَّمَا وَضَعُوهُ لِجَفْظِ اللِّسَانِ مِنَ اللَّحْنِ. وَلَا شَكَّ أَنَّ أَكْثَرَ اللَّحْنِ لَا يُرْسَمُ.

فِي الْفُقْرَةِ التَّالِيَةِ، نَذْكُرُ عِدَّةً مِنَ الْأَبْحَاطِ الَّتِي نُشِرَتْ فِي مَجَالِ التَّصْحِيحِ الْآلِيِّ لِلأَخْطَاءِ الإملائيَّةِ وَالنُّحُويَّةِ. وَقَدْ اقْتَصَرْنَا عَلَى الَّتِي تُخَصُّ النَّصَّ الْعَرَبِيَّ. فِي الْفُقْرَةِ الَّتِي تَلِيهَا، نُسَلِّطُ الضُّوءَ عَلَى خَصَائِصَ فِي اللُّغَةِ الْعَرَبِيَّةِ تَجْعَلُ الْأَنْظِمةَ الْمُصَمَّمةَ لِلْغَايَةِ أُخْرَى غَيْرَ مُرْضِيَةٍ الْأَدَاءِ فِي النَّصِّ الْعَرَبِيِّ. أَمَّا الْفُقْرَةُ الرَّابِعَةُ فَتُخَصِّصُهَا لِتَفْصِيلِ هَيْكَلِ (نِظَامِ بَلْقَيْسِ) الَّذِي يُحَاوِلُ مَلَأَ هَذَا الْفَرَاغَ، ثُمَّ مُنَاقَشَةِ لِنَتَاجِ التَّجَارِبِ فِي الْفُقْرَةِ الْخَامِسَةِ.

I-1- الْأَبْحَاطُ ذَاتُ الْعَلَاَقَةِ:

حَتَّى الْخَبِيرُ اللُّغَوِيُّ مِنَ الْبَشَرِ، قَدْ تَفَوُّهُ أخطاءُ إملائيَّةٌ أَوْ نُحُويَّةٌ عِنْدَ تَدْقِيقِ نَصِّ عَرَبِيٍّ (Alothman & Alsaman, 2020) يُعْطِي هَذَا تَصَوُّرًا عَن مَدَى تَعْقِيدِ نِظَامِ آيٍّ مُصَمَّمٍ لِلْقِيَامِ بِنَفْسِ الْمُهْمَةِ. لَمْ يَنْهِنِ هَذَا الْمُخْتَصِّصِينَ عَنِ الْبَحْثِ عَن تَقْنِيَّاتٍ تُحَاكِي الْخَبِيرَ الْبَشَرِيَّ فِي تَدْقِيقِ النُّصُوصِ. فِي هَذِهِ الْفُقْرَةِ، نَعْرِضُ عِدَّةً مِنَ مُحَاوَلَاتِ تَصْمِيمِ مُدَقِّقٍ إملائيٍّ وَ/أَوْ نُحُويٍّ لِلنَّصِّ الْعَرَبِيِّ (singh, 2020 Azmi, Et.al.2019; Attia, Et.al. 2015;) لَيْسَ الْهَدَفُ اسْتِيعَابَ كُلِّ تِلْكَ الْأَبْحَاطِ، بَلْ لَمْ نُخَصِّرْ إِلَّا مَا كَانَ مُعْنِيًا بِالْأَخْطَاءِ الإملائيَّةِ وَالنُّحُويَّةِ الَّتِي تُؤَوَّلُ إِلَى تَرْكِيبِ كَلِمَاتٍ مِنَ الْقَامُوسِ الْعَرَبِيِّ. وَإِلَّا؛ فَإِنَّ اكْتِشَافَ الْأَخْطَاءِ الَّتِي تُفْضِي إِلَى تَرَكَيبِ غَرِيبَةٍ عَنِ الْقَامُوسِ الْعَرَبِيِّ، سَهْلٌ بَسِيطًا.

فِي الْبَحْثِ الْمَوْسُومِ "مُقَارَبٌ لِتَحْلِيلِ وَتَصْحِيحِ الْأَخْطَاءِ الإملائيَّةِ لِلْمُتَعَلِّمِينَ غَيْرِ النَّاطِقِينَ بِاللُّغَةِ الْعَرَبِيَّةِ" (An approach for analyzing and correcting spelling errors for non-native Arabic learners) (Shaalán & Fahmy, 2010) يَذْكُرُ الْبَاحِثَانِ أَنَّهُمَا اعْتَمَدَا الْمُحَلَّلَ الصَّرْفِيَّ (بُوكْ وَلْتِر) (Buckwalter, 2004)؛ لِاِكْتِشَافِ التَّرَاكِيِبِ الإملائيَّةِ الدَّخِيلَةِ. كَمَا يَذْكُرَانِ أَنَّهُمَا اسْتَعْمَلَا عِدَّةً مِنَ الْمَعَايِيرِ الدَّلَالِيَّةِ لِاِكْتِشَافِ الْأَخْطَاءِ الَّتِي أَفْضَتْ إِلَى كَلِمَاتٍ مِنَ الْقَامُوسِ الْعَرَبِيِّ. أَمَّا لِتَصْحِيحِ الْأَخْطَاءِ، فَيَقُولَانِ إِنَّهُمَا اعْتَمَدَا آليَّةَ حِسَابِ الْمَسَافَةِ الإملائيَّةِ (Deza & Deza, 2012) وَقَوَاعِدَ مُنْطَقِيَّةً (Sabry, 2023). رَغْمَ أَنَّ الْبَاحِثَيْنِ يَرَيَانِ أَنَّ مُقَارَبَتَهُمَا تُعَالِجُ الْأَخْطَاءَ الْمُتَعَلِّقَةَ بِضُبُطِ الْحَرَكَاتِ، إِلَّا أَنَّ الْفَاحِصَ يَتَبَيَّنُ لَهُ أَنَّهُمَا يَقْصِدَانِ

بهذا النوع من الأخطاء، زلة إدراج حرف بدل الحركة بسبب الالتباس الذي يسببه نطق الحركة. مثل رسم نون بدل التنوين في كلمة "كتب" خطأ "كتبن". حسب الباحثين فإن التجارب أثبتت أن مقارنتهما تكشف ٩٥,٣١٪ من الأخطاء، وتصيب في ٩٦,٨٢٪ من حالات الاكتشاف. أما في التصحيح، فيقولان إن مقارنتهما تمكنت من تصحيح ٨٧,٧٠٪ من الأخطاء. وكانت التصويبات التي افترختها مقارنتهما في ٩٥,١٣٪ من الحالات صحيحة. في بحث آخر (Azmi, Almutery and Aboalsamh, 2019)، يقرر صاحبه أن أداء المقارنة سألقة الذكر أفضل في حالات الأخطاء التي نتج عنها تراكم ليست في القاموس. ولا شك أن هذا هو المتوقع.

زبيبي وابن أحمد (٢٠١٣) في بحثهما الموسوم "اكتشاف الأخطاء الدلالية في النص العربي" (Detection of semantic errors in Arabic texts)، درسا إمكانية الاكتشاف الآلي للأخطاء الدلالية البحتة، حين يكون التركيب غير مستهجن، لا إملائي، ولا نحوي. وقد حاولا نمذجة السياق النصي معتمدين على الإحصائيات. ثم دربا النظام الذي صمما على أكثر من ست وعشرين ألف جملة. في مرحلة التجربة، يقيس النظام مناسبة ألفاظ الجملة المكتوبة للسياق. فإن تطابق السياق والألفاظ، وإلا أغشيت الجملة غير رصينة. يذكر الباحثان أن تجاربهما أثبتت أن نظامهما يصيب في ٩٠٪ من الحالات، وتمكن من اكتشاف ٨٣٪ من الأخطاء. وقد أجريا تجاربهما على جمل في مجال الإقتصاد. أدخلت عليها يدويا أخطاء دلالية. يجدر بالذكر أن الباحثين استعملوا تقنية تعدد الوكلاء في بيئة موزعة الأنظمة؛ وذلك لسببين: الأول: تسريع عملية المعالجة، والثاني: توزيع نماذج إحصائية مختلفة على أكثر من وكيل.

الطريقة المتبعة في البحث الموسوم "تدقيق إملائي دلالي باستعمال روابط القوة" (Arabic semantic spell checking based on power links) (Rokaya, 2015) اعتمدتها المقارنة السابقة (Zibiri & Ben Ahmed, 2013)، غير أن الباحث في هذه، استعمل معياراً سماً: روابط القوة (Power links) لقياس مدى ارتباط لفظ بآخر، ولفظ بنص. بالإضافة إلى جدول الالتباسات الشائعة لافتراح تصويبات. يقول الباحث إن تجاربه أثبتت أن مقارنته قادرة على اكتشاف ٨٥,٥٧٪ من الأخطاء، وتصيب في ٩٤,٣٥٪ من الحالات. حسب الباحث، فإن تجاربه كانت على ما يقارب خمساً وثلاثين ألف جملة، أدخل عليها أقل من ألف وخمسمائة خطأ دلالي.

في البحث الموسوم "مدقق إملائي للنص العربي، حساس للسياق باستعمال كلمات السياق ونموذج لغوي إحصائي" (Context-Sensitive Arabic Spell Checker Using Context Words and N-Gram Language Models)، (Al-Jefri & Mahmoud, 2013)، استعمل الباحثان ٧٠٪ من مدونة مؤلفة من أكثر من واحد وأربعين مليون كلمة لبناء نموذج لغوي إحصائي س-وحدة (n-grams). كما بنيا جدولاً من ثمان وعشرين التباس شائع معتمدين على أخطاء نظام للتعرف المرئي على الأحرف (OCR)، وأخطاء متكلمين باللغة العربية غير عرب. بعد التدريب، تستخدم هذه المقارنة النموذج اللغوي لاكتشاف الأخطاء، في حين يستخدم جدول الالتباسات في افتراح التصويبات. يقول الباحثان إن دقة التصويبات وصلت في التجارب إلى ٩٥,٩٪.

حَسَبَ عِلْمُنَا، فَإِنَّ الْمُقَارَبَةَ الْوَحِيدَةَ الَّتِي تَنْظُرُ فِي حَرَكَاتِ الْحُرُوفِ وَلَوْ جُزْئِيًّا، هِيَ الْمُنْشُورَةُ فِي بَحْثٍ بِعَنْوَانٍ "مَدَقَّقُ لِلنَّحْوِ الْعَرَبِيِّ بِاسْتِعْمَالِ قَوَاعِدِ التَّبَعِيَّةِ" (An Arabic Grammar Auditor Based on "Dependency Grammar") (Alothman & Alsaman. 2020). الْبَاحِثَانِ بَنَيَا مَجْمُوعَةً قَوَاعِدَ تَبَعِيَّةٍ (Hays. 1964; Owens. 1983)، مُسْتَقْبَاةً مِنْ قَوَاعِدِ النَّحْوِ الْعَرَبِيِّ. تَسْتَخْرِجُ هَذِهِ الْمُقَارَبَةُ خَصَائِصَ الْكَلِمَاتِ عِبْرَ تَوْظِيفِ الْمُحَلِّلِ الصَّرْفِيِّ (مَا دَامِيرَا (Pasha et al. 2014)، (MADAMIRA). اعْتِمَادًا عَلَى تِلْكَ الْخَصَائِصِ، ثُبُنِيَ خَارِطَةُ تَعْلُقي (Dependency Graph). حَيْثُ إِنَّ قَوَاعِدَ التَّعْلُقي سَالِفَةَ الذِّكْرِ مُوسُومَةٌ بِنَهَائِيَّاتِ الْكَلِمَاتِ الْمُفْتَرَضَةِ، يَتِمُّ التَّأَكُّدُ مِنْ مُطَابَقَةِ نِهَائِيَةِ الْكَلِمَةِ لِمَا هُوَ مُفْتَرَضٌ. هَذِهِ الْمُقَارَبَةُ تَفْتَرِضُ أَنَّ النَّصَّ خَالٍ مِنَ الْأَخْطَاءِ الْإِمْلَانِيَّةِ. حَسَبَ الْبَاحِثَيْنِ؛ فَإِنَّ نَتَاجَ التَّجَارِبِ عَلَى مَجْمُوعَةٍ مِنْ ٧٥٢ جُمْلَةً أُثْبِتَتْ أَنَّ مُقَارَبَتَهُمَا اكْتَسَفَتْ ٩٤,٤٢٪ مِنَ الْأَخْطَاءِ وَكَانَتْ مُصِيبَةً فِي ٨٦,٩٦٪ مِنَ الْحَالَاتِ.

فِي بَحْثٍ أَقْدَمَ (Shaalan & Khaled. 2010; Shaalan. 2005)، اعْتَمَدَ صَاحِبُهُ عَلَى التَّبَزُّجَةِ الْمُنَظَّمَةِ لِاِكْتِسَافِ الْأَخْطَاءِ النَّحْوِيَّةِ. قَاعِدَتُهُ الْمَعْرِفِيَّةُ كَانَتْ تَصْنِيفًا يَدَوِيًّا لِأَكْثَرِ الْأَخْطَاءِ النَّحْوِيَّةِ شُبُوعًا. اسْتَعْمَلَهَا لِإِنْدَاءِ ١٦٢ قَاعِدَةً نَحْوِيَّةً. حَسَبَ الْبَاحِثِ؛ فَإِنَّ تَجَارِبَهُ الَّتِي أَجْرَاهَا عَلَى مِثْلَةِ جُمْلَةٍ شَارَكَ فِي بِنَائِهَا مُحْتَضُونَ مِنْ خَارِجِ فَرِيقِهِ؛ أُثْبِتَتْ أَنَّ النِّظَامَ الَّذِي سَمَّاهُ (GramCheck) كَانَ دَقِيقًا فِي ٩٦٪ مِنَ الْحَالَاتِ.

هَلْ تَكْفِي مِنْهَجِيَّةُ اِكْتِسَافِ الْأَخْطَاءِ فِي اللُّغَاتِ الْآخَرَى؟

كَثِيرًا مَا يَخْتَصِرُ الْبَاحِثُونَ فِي مَجَالِ حَوْسَبَةِ اللُّغَةِ الْعَرَبِيَّةِ، مُعَالَجَةَ التَّحْدِيَّاتِ الْمُتَعَلِّقَةِ بِهَذَا الْمَجَالِ فِي تَكْيِيفِ مُقَارَبَاتٍ صُمِّمَتْ لِتَحْدِيَّاتٍ مُشَابِهَةٍ فِي لُغَاتٍ أُخْرَى، خَاصَّةً الْإِنْجِلِيزِيَّةِ. هَذِهِ الْمُمَارَسَةُ تَنْبِيْهِ عَلَى افْتِرَاضِ عَدَمِ وُجُودِ اخْتِلَافٍ بَيْنَ اللُّغَتَيْنِ يَمْنَعُ مِنْ اعْتِمَادِ مُقَارَبَةٍ صُمِّمَتْ لِإِحْدَاهُمَا فِي مُعَالَجَةِ نَصٍّ فِي اللُّغَةِ الْآخَرَى.

وَأِنْ كُنَّا فِي هَذَا الْبَحْثِ بِصَدَدِ تَقْرِيرِ مُقَارَبَةٍ لِمُعَالَجَةِ تَحَدٍّ مُتَعَلِّقٍ بِالتَّدْقِيقِ اللَّغَوِيِّ الْأَيْ، إِلَّا أَنَّنَا مِنْ ثُبُتٍ أَنَّ خَصَائِصَ اللُّغَةِ الْعَرَبِيَّةِ قَدْ تَغَيَّرَتْ تَمَامًا زَاوِيَةَ النَّظَرِ وَنَوْعَ الْمُقَارَبَةِ الْمَدْرُوسَةِ فِي حِلِّ تَحَدٍّ مُتَعَلِّقٍ بِحَوْسَبَةِ اللُّغَةِ.

وَقَدْ نَلْخُصُ الْإِضَافَاتِ الْعِلْمِيَّةَ لِهَذَا الْبَحْثِ فِيمَا يَلِي:

١. عَرَضُ آلِيَّةٍ جَدِيدَةٍ لِمُعَالَجَةِ أَخْطَاءِ نَحْوِيَّةٍ خَاصَّةً بِاللُّغَةِ الْعَرَبِيَّةِ، لَمْ تُدْرَسْ مِنْ قَبْلُ -حَسَبَ عِلْمُنَا-.
٢. (مُسْتَمْعٌ بِلَقِيْسٍ). هُوَ أَخَذَ أَنْظَمَةً (مَجْمُوعَةً بِلَقِيْسٍ). يَتَعَرَّفُ عَلَى الْكَلَامِ الْعَرَبِيِّ الْمُنطَوِقِ؛ لِيُوَلِّدَ النَّصَّ الَّذِي سَمِعَهُ مَعَ كَامِلِ الْحَرَكَاتِ. وَحَسَبَ عِلْمُنَا؛ لَا وُجُودَ لِنِظَامِ اسْتِمَاعٍ يُوَلِّدُ الْحَرَكَاتِ.
٣. مَدَوْنَةُ صَوْتِيَّةٌ عَالِيَةِ الْجَوْدَةِ، مُؤَلَّفَةٌ مِنْ أَكْثَرِ مِنْ ثَلَاثِ عَشْرَةِ أَلْفِ مَقْطَعٍ صَوْتِيٍّ عَالِيِ الْجَوْدَةِ، يُنْطَقُ صَحِيحًا، مُرْفَقٌ بِالنَّصِّ الْعَرَبِيِّ كَامِلِ التَّشْكِيلِ. بَنَيْنَا هَذِهِ الْمَدَوْنَةَ بِاسْتِعْمَالِ (نِظَامِ بِلَقِيْسٍ)، الَّذِي يُوفِّرُ أَدَاةَ تَوْظُفِ الذِّكَاءِ الْإِصْطِنَاعِيِّ لِتَحْوِيلِ التَّسْجِيلَاتِ الصَّوْتِيَّةِ إِلَى مَدَوْنَاتٍ كَامِلَةِ التَّشْكِيلِ. وَلَا تَوْجَدُ -حَسَبَ أَطْلَاعِنَا- مَدَوْنَةُ صَوْتِيَّةٌ عَرَبِيَّةٌ مُشَابِهَةٌ مُنْشُورَةٌ.

٤. إثبات أن أهمية ضبط الحركات في النص العربي المكتوب لا تقتصر على التوليد الصوتي والتحليل الدلالي للمحتوى العربي [١٨]. بل إن دقة الكثير من التطبيقات في مجال حوسبة اللغة العربية تتأثر تأثراً كبيراً بضبط حركات النص.

في اللغات المسماة اللغات الجرمانية (Germanic Languages) مثل الإنجليزية (Konig & Auwera, 2013)، تُكتب الكلمة كأملة الحركات مثل كلمة "Arabic" والتي تنطق "أريك". جملة مثل "you read the Arabic book" والتي تعني "أنت تقرأ الكتاب العربي" وتنطق "يوريد ذ أريك بك"، لن يضر عارفاً باللغة الإنجليزية أن تكتب نفس الجملة "yu read th arbc buk" إذ ما أن يحاول نطقها، حتى تتضح له الحركات التي حذف أو استبدلت. الكتابة الأخيرة تُعد خطأ فادحاً من زاوية نظر أكاديمية. ولن تجد معلماً لغة إنجليزية يدرب تلامذته على مثل هذه الممارسة. أما اللغة العربية فشائع جداً كتابة نصها بلا ضبط للحركات. بل إن نقاط الحروف لم تكن تكتب عند العرب ولا عرفت قبل الإسلام (اللوحي، ١٩٥٥). ومع هذا، لا يجد الفصحى باللغة العربية عناء في نطق الكلام وفهم معناه، وإن كُتب بلا حركات. لم تكتب الحركات إذن؟ جواب هذا السؤال هو مريب الفرس ومركز هذا البحث.

حين يقرأ العربي الفصحى نصاً خالياً من الحركات، يستعمل معرفته اللغوية التي تسمى السليقة؛ لترتيب حركات لا يراها على النص. وحيث إن دقة هذا العمل الذهني مرتبطة بقوة سليقة القارئ، وأحياناً بدرايته بسباق الواقع الموصوف في النص، قد يخطئ في تقدير الحركة. وقد لا يتجاوز أثر خطأ في تقدير الحركة مجرد لحن يستهجنه سامع فصيح. بيد أن كثيراً من الأخطاء قد يحيل المعنى إلى ضده. كخطأ في قراءة قوله تعالى "إنما يحشى الله من عباده العلماء".

أدوات التدقيق الإملائي والنحوي كذلك التي استعرضناها في الفقرة السابقة، تنبني كلها على تحليل الرسم (النص المكتوب).

ولاشك أن كثيراً من الأخطاء يمكن التنبه له بمجرد معاينة النص، ومثل "كنت أول المشركون" و"رايت أبو يزيد". غير أن كثيراً من الأخطاء النحوية يسمع ولا يقرأ، وهو المتعلق بنطق الحركات.

في الفقرة التالية، نستعرض نظام اكتشاف الأخطاء النحوية المنطوقة غير المكتوبة. هذا النظام واحد من (مجموعة بلقيس) التي تضم عدداً من الأنظمة الحاسوبية التي تعتمد تقنيات الذكاء الاصطناعي الحديثة لمعالجة المحتوى العربي.

٢- بلقيس: مكتشف اللحن:

بلقيس اسم مستعار من اسم ملكة سبأ (ابن كثير، ٢٠١٥) لمجموعة أنظمة حاسوبية تعتمد على تقنيات الذكاء الاصطناعي الحديثة لمعالجة المحتوى العربي. عدد من تلك الأنظمة متاح عبر موقع شبكي (https://belqis.intelgx.com)، ومتاح كذلك عبر مكتبة برمجية (Buna, 2021) (API) (Biehl, 2016).

أَحَدُ أَنْظِمَةِ مَجْمُوعَةٍ بِلْقَيْسٍ، نِظَامُ التَّدْقِيقِ النَّحْوِيِّ فِي الْكَلَامِ الْمُنطَوِقِ. وَهُوَ الَّذِي صَمَّمْنَاهُ لِسَدِّ
الْفُجُوءِ الَّتِي اسْتَعْرَضْنَاهَا فِيمَا سَبَقَ مِنَ الْمَقَرَّاتِ.

يَسْتَمِعُ هَذَا النِّظَامُ لِنُطْقِ جُمْلَةٍ عَرَبِيَّةٍ لِيُحَاوِلَ اكْتِشَافَ مَوَاضِعِ اللَّحْنِ فِيهَا إِنْ وُجِدَتْ.

نِظَامُ اكْتِشَافِ اللَّحْنِ مُؤَلَّفٌ مِنْ نِظَامَيْنِ:

١. نِظَامُ الْإِسْتِمَاعِ (التَّعَرُّفُ الْأَلْيُّ عَلَى الْكَلَامِ الْمُنطَوِقِ).

٢. نِظَامُ اكْتِشَافِ اللَّحْنِ فِي النَّصِّ الْمَكْتُوبِ.

٢-١- بِلْقَيْسُ: نِظَامُ الْإِسْتِمَاعِ:

هُوَ أَحَدُ أَنْظِمَةِ مَجْمُوعَةٍ بِلْقَيْسٍ. يَتَعَرَّفُ عَلَى الْكَلَامِ الْعَرَبِيِّ الْمُنطَوِقِ؛ لِيُولِّدَ النَّصَّ الَّذِي سَمِعَهُ مَعَ
كَامِلِ الْحَرَكَاتِ. بِنَاءَ هَذَا النِّظَامِ تَطَلَّبَ مِنَّا بِنَاءَ مَدُونَةٍ صَوْتِيَّةٍ عَرَبِيَّةٍ كَامِلَةٍ التَّشْكِيلِ بِجُودَةٍ عَالِيَةٍ وَبِحَجْمٍ
كَافٍ، وَلَا عَلِمْنَا بِوُجُودِ مِثْلِ هَذِهِ الْمَدُونَةِ.

وَقَدْ سَلَكْنَا لِبِنَاءِ هَذِهِ الْمَدُونَةِ مَنَهْجًا تَدْرِيجِيًّا:

١. عَمَلْنَا عَلَى تَقْسِيمِ كُتُبِ صَوْتِيَّةٍ إِلَى مَقَاطِعِ صَوْتِيَّةٍ قَاصِرَةٍ أَلْيَا.

٢. بَنَيْنَا نُسْخَةَ أَوَّلِيَّةٍ مِنْ نِظَامِ الْإِسْتِمَاعِ، وَدَرَّبْنَاهُ عَلَى عَدَدٍ قَلِيلٍ نَسْبِيًّا مِنْ مَقَاطِعِ صَوْتِيَّةٍ مُخْتَلِفَةٍ.

٣. اسْتَعْدَمْنَا هَذِهِ النُّسْخَةَ الْأَوَّلِيَّةَ كَأَدَاةٍ مَسَاعِدَةٍ عَلَى التَّعَرُّفِ عَلَى الْمَقَاطِعِ الصَّوْتِيَّةِ الَّتِي بَيَّنَّتْ
خِلَالَ الْمَرَحَلَةِ الْأُولَى. حَيْثُ إِنَّ هَذِهِ النُّسْخَةَ مِنْ نِظَامِ الْإِسْتِمَاعِ لَمْ يَتِمَّ تَدْرِيبُهَا بِشَكْلِ كَافٍ،
كَانَتْ أَخْطَاؤُهَا كَثِيرَةً. لِذَا، أَصَفْنَا لِنِظَامِ الْإِسْتِمَاعِ نِظَامَ مُطَابَقَةِ النَّصِّ. هَذَا النِّظَامُ يُقَارِنُ
بِشَكْلِ ذِكِّي نُسْخَةَ نَصِيَّةٍ مِنَ الْكِتَابِ الصَّوْتِيِّ وَمَا يُولِّدُهُ نِظَامُ الْإِسْتِمَاعِ؛ لِيَسْتَخْرِجَ أَقْرَبَ
نَصٍّ لِمَا فِي الْمَقْطَعِ الصَّوْتِيِّ. يُجَدُّرُ بِالذِّكْرِ أَنَّ النُّسْخَةَ النَّصِيَّةَ مِنَ الْكِتَابِ بِهَا اخْتِلَافَاتٌ كَثِيرَةٌ،
وَلِضْمَانِ جُودَةٍ عَالِيَةٍ، اضْطَرَرْنَا لِبِنَاءِ نِظَامٍ حَاسُوِيٍّ يُسَاعِدُ عَلَى الْمُرَاجَعَةِ السَّرِيعَةِ لِلنَّصِّ
الْمُولَّدِ مِنْ نِظَامِ الْإِسْتِمَاعِ.

٤. بَعْدَ أَنْ بَنَيْنَا خِلَالَ الْمَرَحَلَةِ السَّابِقَةِ مَدُونَةً مُكَوَّنَةً مِنْ أَكْثَرِ مِنْ خَمْسَةِ آلَافِ مَقْطَعِ صَوْتِيٍّ.
دَرَّبْنَا نِظَامَ الْإِسْتِمَاعِ عَلَى هَذِهِ الْمَدُونَةِ، آمِلِينَ أَنْ تَزِيدَ دَقَّتُهُ فِي التَّعَرُّفِ عَلَى الْمَقَاطِعِ الصَّوْتِيَّةِ
الْمُسَابَهَةِ، وَفِعْلًا، قَلَّتْ أَخْطَاءُ النِّظَامِ كَثِيرًا، وَلِلَّهِ الْحَمْدُ.

٥. كَرَّرْنَا الْمَرَّاجِلَ ٣، ١ و ٤ حَتَّى تَمَكَّنَّا مِنْ بِنَاءِ مَدُونَةٍ بِهَا أَكْثَرُ مِنْ ثَلَاثَةِ عَشَرَ أَلْفَ مَقْطَعِ صَوْتِيٍّ.
مَجْمُوعُ مَدُونَتِهَا إِخَذَ عَشْرَةَ سَاعَةً، وَمُعَدَّلُهَا ثَلَاثُ ثَوَانٍ تَقْرِيبًا.

بِنَفْسِ مَنَهْجِ بِنَاءِ الْمَدُونَةِ هَذَا، بَنَيْنَا نِظَامَ بِنَاءِ الْمَدُونَاتِ الصَّوْتِيَّةِ الْعَرَبِيَّةِ، وَأَصَفْنَاهُ إِلَى (مَجْمُوعَةٍ
بِلْقَيْسٍ). وَهُوَ مِنَ الْأَنْظِمَةِ الْمَتَاحَةِ لِلْبَاحِثِينَ عِبْرَ الرَّابِطِ (<https://belqis.intelgx.com>).

٢-٢- نظام اكتشاف اللحن:

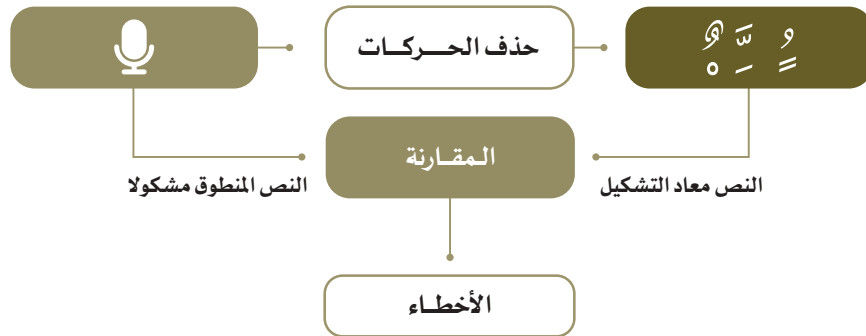
يَعْتَمِدُ هَذَا النِّظَامُ عَلَى نِظَامِ التَّشْكِيلِ الْآلِيِّ، الَّذِي هُوَ أَخَذَ أَنْظِمَةَ (مَجْمُوعَةَ بَلْقَيْس). يُحْلُلُ هَذَا الْأَخِيرُ نَصًّا عَرَبِيًّا لِيُولِّدَ نَفْسَ النَّصِّ مَشْكُولًا تَشْكِيلًا كَامِلًا.

نَوَافُةً نُمُوذَجَ هَذَا النِّظَامُ، مُحَوَّلٌ بِمَشَقَرٍ عَضْبُونِيٍّ (encoder only transformer)، (Clark. et.al. 2022)، يَضُمُّ ١٣٢ مَلْيُونًا مِغْيَارٍ (trainable parameter). يَخْتَلِفُ عَنْ غَيْرِهِ فِي أَنَّهُ لَا يَحْتَاجُ إِلَى مُمَيِّزِ كَلِمَاتٍ (tokenizer)، (Tunstall. 2022)، لِتَدْرِيبِهِ، عَمَلْنَا عَلَى بِنَاءِ مَدَوْنَةِ نَصِّ مَشْكُولٍ بِالْكَامِلِ. تَضُمُّ هَذِهِ الْمَدَوْنَةُ أَكْثَرُ مِنْ ٩٤ مَلْيُونًا كَلِمَةً، وَمَا يَقَارِبُ الْبِلْيُونِ وَالنِّصْفِ زَمْرًا. مِنْ ضَمَمِهَا نُصُوصُ الْمَدَوْنَةِ الشَّاعِعَةِ "تَشْكِيلَةً" (Zerrouki & Balla.2017).

التَّجَارِبُ عَلَى نَصِّ مُخْتَلِفٍ يَحْتَوِي ٨٢٨٨ كَلِمَةً، أَظْهَرَتْ دَقَّةَ تَسَاوِي ٩٨,٠٧٪. وَهُوَ النِّظَامُ نَفْسُهُ الَّذِي اسْتَعْمَلْنَاهُ لِتَشْكِيلِ نَصِّ هَذَا الْبَحْثِ مَعَ تَصْحِيحٍ يَدَوِّيٍّ لِلْأَخْطَاءِ.

٢-٣- النِّظَامُ: مِنَ الْأَلْفِ إِلَى الْيَاءِ:

يُعَدِّي نِظَامُ الْإِسْتِمَاعِ نِظَامَ اكْتِشَافِ اللَّحْنِ بِالنَّصِّ الَّذِي يُولِّدُهُ مِمَّا يَتَعَرَّفُ عَلَيْهِ مِنَ الْكَلَامِ الْمُنْطَوِقِ. بَعْدَ نَزْعِ حَرَكَاتِ الْكَلِمِ، يُوَجَّهُ إِلَى نِظَامِ التَّشْكِيلِ الْآلِيِّ؛ لِيُولِّدَ حَرَكَاتِ الْكَلِمِ مُجَدَّدًا؛ لِيُقَارَنَ نِظَامُ اكْتِشَافِ اللَّحْنِ بَيْنَ الْحَرَكَاتِ الَّتِي وَلَدَهَا نِظَامُ التَّشْكِيلِ الْآلِيِّ وَالْحَرَكَاتِ فِي النَّصِّ الَّذِي وَلَدَهُ نِظَامُ الْإِسْتِمَاعِ. إِذَا اخْتَلَفَتْ الْحَرَكَةُ الْمُنْطَوِقَةُ عَنِ الْحَرَكَةِ الْمَتَوَقَّعَةِ عَدَّهَا خَطَأً.



شكل ١: نظام اكتشاف وتصحيح اللحن في الكلام العربي

جدول ١: خوارزمية اكتشاف اللحن:

١	ج م = استماع (ص) # الجملة المنطوقة حسب ما يولده نظام الاستماع، ص الصوت
٢	ج = نزع الحركات (ع)، # ج هي ع، منزوعة التشكيل.
٣	ج ت = تشكيل (ع) # ع الجملة ج مشكولة عبر نظام بلفيس للتشكيل.
٤	من أجل كل كلمة ن في ع، والكلمة نفسها في ج ت.
٥	إذا كانت حركة إعراب (ن)، <> حركة إعراب (ن).
٦	تعيين الكلمة خطأ.
٧	اقترح حركات (ن) تصويبا.

٣- نتائج التجارب:

نظام اكتشاف اللحن مؤلف من نظامين، كل منهما قد يعثره الخطأ:

١. نظام الاستماع: الذي يحول الكلام المنطوق إلى نص مشكول كما نطقت الحركات.
٢. نظام التشكيل الآلي: الذي يعيد تشكيل الكلمات في النص الذي يولده نظام الاستماع.

في مثل هذه الحالات قد يسلك تقييم الأداء مسلكين: أحدهما أو كليهما:

١. التقييم العام (extrinsic)، (Mitkov. 2022): يقيس أداء النظام باعتبار صندوق أسود، بغض النظر عن أي عنصر (نظام الاستماع، نظام التشكيل) تسبب في الخطأ.

٢. التقييم الخاص (intrinsic)، (Mitkov. 2022): يقيس أداء العنصر محل الاهتمام (نظام التشكيل) معزولا عن أي عنصر آخر. في مثل هذا القياس، يزود العنصر محل الاهتمام بما قد يرد من مداخل سائلة من الخطأ الذي قد تتسبب فيها العناصر الأخرى.

فيما نحن بصدد، العنصر محل الاهتمام، هو نظام التشكيل الذي يعيد توليد حركات الكلام. هذا النظام أثبتت التجارب أن دقته وصلت إلى ٩٨,٠٧٪ في توليد التشكيل عموما.

لتقييم نظام اكتشاف اللحن بصورة منعزلة عن أداء نظام الاستماع، بنينا مدونة من ٤٧٣ جملة، وما يزيد عن أربع عشرة ألف كلمة. بنينا برنامجا لتوليد انتقائي لأخطاء الإعراب المقتصر على الحركات في آخر الكلمة بنسبة ٣٪.

ثُمَّ قَارَنًا بَيْنَ مُخْرَجَاتِ نِظَامِ التَّشْكِيلِ عَلَى نَفْسِ الْجَمَلِ بَعْدَ نَزْعِ التَّشْكِيلِ مِنْهَا، وَالتَّشْكِيلِ الْأَصْلِيِّ، وَالْأَخْطَاءِ. مَعَ مُرَاعَاةِ التَّالِي فِي الْمُقَارَنَةِ بَيْنَ كَلِمَتَيْنِ مَشْكُولَتَيْنِ:

٣. إِذَا كَانَتْ إِحْدَى الْحَرَكَتَيْنِ غَيْرَ مَوْجُودَةٍ، افْتَرَضْنَا التَّسَاوِي، وَمِثْل: "كَتَبْتُ"، "كَتَبْتُ" حَيْثُ حُذِفَ السُّكُونُ مِنْ حَرْفِ الْبَاءِ.

٤. إِذَا كَانَ الْحَرْفُ مُشَدَّدًا دُونَ ظُهُورِ الْحَرْكَةِ، وَالْحَرْفُ نَفْسُهُ فِي الْكَلِمَةِ الْأُخْرَى مُشَدَّدًا، لَكِنْ مَعَ الْحَرْكَةِ، كَذَلِكَ افْتَرَضْنَا التَّسَاوِي. مِثْلُ "الْعَمِّ" وَ"الْعَمِّ" حَيْثُ حُذِفَتِ الصَّمَّةُ مِنَ الْمِيمِ.

٥. إِذَا اخْتَلَفَتْ حَرَكَةُ الْإِعْرَابِ، عُذَّتِ الْكَلِمَةُ خَطَأً. أَمَّا إِذَا اتَّفَقَتْ حَرَكَتَا الْإِعْرَابِ وَاخْتَلَفَتْ حَرَكَتَا حَرْفٍ آخَرَ وَسَطَ الْكَلِمَةِ، لَمْ تُعَدَّ خَطَأً. إِذْ إِنَّ مَحَلَّ اهْتِمَامِنَا هُوَ اللَّحْنُ فِي إِعْرَابِ الْكَلِمَةِ، وَهُوَ الْمُتَعَلِّقُ بِالنَّسْقِ النَّحْوِيِّ.

بَلَغَتْ نِسْبَةُ الصَّوَابِ (Precision) ٩٣,٢٨٪ فِي جِلْدٍ بَلَغَتْ نِسْبَةُ الاسْتِرْجَاعِ (Recall) ٩٠,٠١٪. لَيْسَ غَرِيبًا أَنْ تَكُونَ نِسْبَةُ الصَّوَابِ أَقَلَّ مِنْ نِسْبَةِ الاسْتِرْجَاعِ. ذَلِكَ أَنَّ نِسْبَةَ الصَّوَابِ مُتَعَلِّقَةٌ بِمَدَى تَطَابُقِ حَرَكَةِ الْإِعْرَابِ

الْمُولَدَةِ مِنْ نِظَامِ التَّشْكِيلِ، وَحَرَكَةِ الْإِعْرَابِ الْأَصْلِيَّةِ فِي النَّصِّ التَّجْرِيئِيِّ. أَمَّا نِسْبَةُ الاسْتِرْجَاعِ، فَمُتَعَلِّقَةٌ بِاخْتِلَافِ الْحَرَكَةِ الْمُولَدَةِ مِنْ نِظَامِ التَّشْكِيلِ عَنْ حَرَكَةِ الْإِعْرَابِ الْخَطَأِ.

٤- خاتمة:

فِي هَذَا الْبَحْثِ، سَلَطْنَا الضُّوءَ عَلَى جَانِبٍ مُهِمٍّ مِنْ حَوْسَبَةِ اللُّغَةِ الْعَرَبِيَّةِ، وَبَيَّنَّا أَنَّ تَقْنِيَّاتِ حَوْسَبَةِ اللُّغَةِ الْعَرَبِيَّةِ خُصَبَةٌ،

وَيُمْكِنُ أَنْ تَنْطَوِي عَلَى تَحْدِيَّاتٍ تَحْتَصُّ بِاللُّغَةِ الْعَرَبِيَّةِ فَقَطْ. وَذَلِكَ مِنْ خِلَالِ تَفْصِيلِ مُقَارِنَةٍ لِحَلِّ تَحَدٍّ لَا يَنْطَبِقُ عَلَى اللُّغَاتِ الْأَكْثَرِجُحْنَا فِي مَجَالِ حَوْسَبَةِ اللُّغَةِ، كَالْإِنْجِلِيزِيَّةِ. مُعَالَجَةُ تَحْدِيٍّ اكْتِشَافِ اللَّحْنِ فِي الْكَلَامِ الْعَرَبِيِّ أَبْرَزَ تَحْدِيَّاتٍ أُخْرَى مُتَعَلِّقَةٌ بِالْبَيِّنَاتِ، وَالْيَتَةِ التَّقْيِيمِ. عَمَلْنَا فِي هَذَا الْبَحْثِ عَلَى إِيجَادِ حُلُولٍ عَمَلِيَّةٍ قَابِلَةٍ لِلتَّطْبِيقِ وَبِإِدَاءِ عَالٍ. نَعْتَقِدُ أَنَّ التَّحْدِيَّ الْأَسَاسَ فِي حَوْسَبَةِ اللُّغَةِ الْعَرَبِيَّةِ هُوَ الْبَيِّنَاتُ، كَمَا وَكُنْهَا. إِنَّ التَّطْبِيقَاتِ الْحَدِيثَةَ، مِثْلُ: النَّمَاذِجِ اللُّغَوِيَّةِ الْعَمَلَاءَةِ، نَهْمَةٌ مِنْ حَيْثُ بَيِّنَاتِ التَّدْرِيبِ. وَلَعَلَّ أَيْ نَظَرَةً مُسْتَقْبَلِيَّةً فِي مَجَالِ حَوْسَبَةِ اللُّغَةِ، لَا يَبْدُ وَأَنْ تَتَّخِذَ مِنْ تَجَاوُزِ عَائِقِ وَفَرَةِ الْبَيِّنَاتِ الْمُنَاسِبَةِ أَوْلَوِيَّةً. وَهُوَ مَا نَعْمَلُ عَلَيْهِ مِنْ خِلَالِ أَنْظِمَةٍ قَادِرَةٍ عَلَى الْإِسْهَامِ فِي تَوْليِدِ وَتَدْقِيقِ بَيِّنَاتِ الْمُحْتَوَى الْعَرَبِيِّ.

المراجع العربية:

- أبو الطيب، اللغوي. (١٩٥٥) مراتب النحويين. مصر: مكتبة نهضة مصر.
ابن كثير، إسماعيل. (٢٠١٥) البداية والنهاية. مصر: دار الكتب.

المراجع الأجنبية:

- Alothman, A., & Alsaman, A. (2020). An Arabic grammar auditor based on dependency grammar. *Advances in Human-Computer Interaction*, 2020, 1-10.
- Azmi, A. M., Almutery, M. N., & Aboalsamh, H. A. (2019). Real-word errors in Arabic texts: A better algorithm for detection and correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8), 1308-1320.
- Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Lrec* (Vol. 14, No. 2014, pp. 1094-1101).
- Belqis System Accessed: Nov. 29, 2023. [Online]. Available: <https://belqis.intelgx.com/>
- Buckwalter, T. (2004). Buckwalter arabic morphological analyzer version 2.0. linguistic data consortium, university of pennsylvania, 2002. ldc cat alog no. Ldc2004l02. Technical report.
- Zribi, C. B. O., & Ahmed, M. B. (2013). Detection of semantic errors in Arabic texts. *Artificial Intelligence*, 195, 249-264.
- Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, 40(4), 511-525.
- Konig, E., & Van der Auwera, J. (2013). *The Germanic languages*. Routledge.
- Sabry, F. (2023). *Rule Based System: Fundamentals and Applications*. One Billion Knowledgeable.
- Clark, J. H., Garrette, D., Turc, I., & Wieting, J. (2022). Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10, 73-91.

- Owens, J. (1983). On Getting a Head: A Problem in Dependency Grammar.
- Shaalán, K. F. (2005). Arabic GramCheck: A grammar checker for Arabic. *Software: Practice and Experience*, 35(7), 643-665.
- K. Shaalan and A. Fahmy, "An approach for analyzing and correcting spelling errors for non-native Arabic learners," Apr. 2010, pp. 1–7.
- Shaalán, K. (2010). Rule-based approach in Arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3), 11-19.
- Tunstall, L., Von Werra, L., & Wolf, T. (2022). Natural language processing with transformers. "O'Reilly Media, Inc."
- Al-Jefri, M. M., & Mahmoud, S. A. (2013). Context-sensitive Arabic spell checker using context words and n-gram language models. In 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (pp. 258-263). IEEE.
- Attia, M., Pecina, P., Samih, Y., Shaalan, K., & Van Genabith, J. (2016). Arabic spelling error detection and correction. *Natural Language Engineering*, 22(5), 751-773.
- Biehl, M. (2016). *RESTful Api Design*. API-University Press.
- Deza, M. m. & Deza, E. (2012). *Encyclopedia of Distances*. Springer Science & Business Media.
- Rokaya, M. (2015). Arabic semantic spell checking based on power links. *International Information Institute (Tokyo). Information*, 18(11), 4749-4770.
- Mitkov, R. (Ed.). (2022). *The Oxford handbook of computational linguistics*. Oxford University Press.
- Buna, S. (2021). *GraphQL in Action*. Simon and Schuster.
- Singh, S., & Singh, S. (2020). Systematic review of spell-checkers for highly inflectional languages. *Artificial Intelligence Review*, 53, 4051-4092.
- Zerrouki, T., & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in brief*, 11, 147-151.

٨- نبذة تعريفية عن الباحث أو الباحثين:

AUTHOR BIODATA:	بيانات الباحث :
Ahmed Mailoud Khorsi is an Associate Professor of Artificial Intelligence in Al Imam Mohamed Ibn Saud Islamic University. Dr. Ahmed received his PhD degree in 2007 from Djillali Liabes University. His research interests include .Deep Natural Language Processing	أحمد ميلود خرصي، أستاذ مشارك في الذكاء الاصطناعي في جامعة الإمام محمد بن سعود الإسلامية بالملكة العربية السعودية حاصل على درجة الدكتوراه في الذكاء الاصطناعي من جامعة جيلالي اليابس بالجزائر عام ٢٠٠٧. تدور اهتماماته البحثية حول حوسبة اللغة العربية ومعالجة المحتوى العربي بالذكاء الاصطناعي الحديث.

0009-0003-6960-7054

معرف أوركيد (ORCID) : 

Email: amakhorsi@imamu.edu.sa | ahmed@intelgx.com

Syllable-Based Arabic Speech Recognition Using Wav2Vec

التعرف على الكلام العربي على أساس المقطع باستخدام Wav2Vec

إبراهيم عبد العال^أ محمد عبد الواحد^ب مصطفى الشافعي^ج

^أ قسم الاتصالات وتكنولوجيا المعلومات، كلية الهندسة، جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر
^ب قسم الاتصالات وتكنولوجيا المعلومات، كلية الهندسة، جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر
^ج قسم الاتصالات وتكنولوجيا المعلومات، كلية الهندسة، جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر

APA Citation:

Abdalaal, I., Abdelwahed, M., & Elshafei, M. (2024). Syllable-Based Arabic Speech Recognition Using Wav2Vec. *Journal of Computational Linguistics and Arabic Language Processing*, 1, 91-109

Submission Date: 01/08/2023

Acceptance Date: 23/10/2023

توثيق البحث:

عبد العال، إبراهيم، عبد الواحد، مصطفى، (٢٠٢٤). التعرف على الكلام العربي على أساس المقطع باستخدام Wav2Vec. *مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية*، العدد ٩١، ٩١-١٠٩.

تاريخ الإرسال: ٠١/٠٨/٢٠٢٣

تاريخ القبول: ٢٣/١٠/٢٠٢٣

Abstract:

This paper introduces an innovative Arabic speech recognition method that relies on Arabic syllables, a 5-gram language model, and the Wav2Vec-2 architecture. It starts by segmenting Arabic speech into syllables, improving accuracy by handling the language's complexity. The method is rigorously evaluated, showing a significant boost in system performance. To enhance accuracy further, a 5-gram language model is used to address linguistic nuances. The Wav2Vec-2 architecture, known for its robust acoustic representations, is employed as the acoustic model. Fine-tuning on a smaller labelled dataset improves Arabic speech recognition, making it resilient to pronunciation variations and noise. This unique combination yields impressive Word Error Rates (WER) of 0.06624 with Tashkeel and 0.05959 without Tashkeel, demonstrating its effectiveness.

Keywords:

Wav2Vec- WER- Language model – Syllables - Pronunciation rules

الملخص:

تقدم هذه الورقة طريقة مبتكرة للتعرف على الكلام العربي، بالاستفادة من الخصائص الفريدة للمقاطع الصوتية العربية، ونموذج لغة ٥ جرام، بالإضافة إلى نموذج Wav2Vec-2 الحديث. تبدأ الطريقة بتقسيم الكلام العربي إلى مقاطع صوتية يتم تكوينها برمجياً باستعمال قواعد نطق اللغة العربية، وتتميز بقدرتها على وصف التفاصيل الصوتية الدقيقة. ولزيادة تحسين الدقة، تم استخدام نموذج لغوي ٥ جرام للمقاطع الصوتية والذي يأخذ في الاعتبار الظواهر اللغوية المتأصلة في اللغة العربية، بما في ذلك الاختلافات الكبيرة على مستوى الكلمات والتبعيات السياقية. كما تم استعمال نموذج Wav2Vec كنموذج صوتي، وهو معروف بتمثيلاته الصوتية القوية التي تم تعلمها من خلال التدريب المسبق غير الخاضع للإشراف على كميات كبيرة من بيانات الكلام غير المسماة. وتم ضبط النموذج بدقة على مجموعة بيانات أصغر، مما عزز قدرته على التعرف على الكلام العربي. قد حقق هذا المزيج الجديد من التقنيات معدل خطأ في الكلمات بلغ ٠.٠٦٦ مع تشكيل النص، و٠.٠٥٩٦ بدون تشكيل النص، مما يدل على فعالية هذا النهج.

الكلمات المفتاحية:

نموذج لغة - المقاطع الصوتية - قواعد النطق

1. Introduction

Speech recognition technology has become an integral aspect of our daily lives, playing a critical role in powering interactive education and training, virtual assistants, speech transcription, speech translation, games, and various voice-enabled applications. Meeting the escalating demand for multilingual and dialect-specific speech recognition systems necessitates the development of robust solutions tailored to languages with intricate phonetics and morphology. Arabic, being one of the world's most widely spoken languages, spoken by 300 million, and the official language of 22 countries, presents distinctive challenges due to its complex morphological structure and diverse phonetic variations across different regions and dialects.

Conventional methods for Arabic speech recognition predominantly rely on phoneme-based models, facing challenges in capturing the subtle phonetic nuances and contextual dependencies inherent in the language. Consequently, the demand for more sophisticated techniques capable of effectively handling the intricacies of Arabic pronunciation and morphology has grown substantially. One notable issue in recognizing Modern Standard Arabic (MSA) speech involves contextual phoneme alteration in both cross-word pronunciation and intra-word pronunciation variation. These contextual variations alter the phonetic spelling of words, resulting in a high word error rate (WER). In a study by (AbuZeina et al., 2011), the authors presented a knowledge-based approach based on modelling cross-word pronunciation variation by expanding the phonetic dictionary and corpus transcription. They tested the method on a speech corpus of 5.4 hours of Modern Standard Arabic. The WER was reduced by 2.3% below the baseline system.

In data-driven end-to-end speech recognition approaches, the recognition model is provided with the speech utterances and its corresponding text. The problem here is that the phonetic representation of the actual utterance is not produced by a simple grapheme-to-phoneme relationship, but produced by a network of pronunciation rules. For example, a letter is written in the text, but may not be pronounced depending on the context, e.g. the word "Alshams" is pronounced "Ash-shams" without letter L, and introducing a second phoneme "sh". As a sample of common substitution of letters; the letter N in "man laho" is uttered "mal-laho" with phoneme N substituted by phoneme L. Similarly, the phoneme N in the word "Anbar" is replaced by M and pronounced "Ambar". However, these variations follow up certain pronunciation rules.

To overcome these problems, the end-to-end approach requires enormous data sets and complex deep learning models to capture the latent pronunciation rules, hopefully by a large number of training examples.



Our proposed approach is to apply Arabic pronunciation rules to produce syllable strings reflecting the actual utterances. There are two advantages to this approach. The first is that the syllables represent un-interrupted sound unit. The second advantage is that every uttered phoneme is written and every unuttered phoneme is removed from the corresponding syllable script, resulting in a one-to-one grapheme to phoneme mapping. This approach could help to use smaller datasets and more efficient models for the same WER.

In the proposed syllable-based approach the text is converted to syllables using the Arabic pronunciation rules, (Alghamdi et al., 2004). If foreign names or words are present in the text, exceptions may occur leading to unknown syllables, or could be forced to follow the Arabic syllable patterns. For example, "street" could be uttered as "es te reet".

The paper introduces an innovative method for Arabic speech recognition that utilizes syllables, a linguistically motivated unit, in combination with a 5-gram language model and the advanced Wav2Vec-2 architecture (Baevski et al., 2020). Embracing a syllable-based approach enables us to surpass the limitations of phoneme-based systems, providing a more resilient and precise solution for Arabic speech recognition. Syllables, natural linguistic units in many languages, offer several advantages for Arabic speech recognition. Firstly, they provide a concise representation of phonetic data, bridging the gap between phonemes and more extensive linguistic components like words and phrases. This quality allows for a more detailed analysis of speech, capturing crucial phonetic variations necessary for precise recognition. Secondly, syllables in Arabic closely relate to morphological units, making them suitable for handling the language's complex word structure and root-based derivational patterns. Leveraging syllables can contribute to a more linguistically informed and contextually aware speech recognition system.

To further enhance the accuracy and robustness of the proposed system, an integrated 5-gram language model captures the statistical dependencies between successive syllables, effectively addressing contextual variations and dependencies in Arabic speech. Considering sequences of five syllables enables the language model to exploit long-range dependencies that often influence Arabic word formation and pronunciation, leading to improved recognition performance.

As the acoustic model, the Wav2Vec-2 architecture is adopted, showcasing remarkable success in various speech recognition tasks. Wav2Vec-2 utilizes unsupervised pre-training on vast amounts of unlabelled speech data, allowing it to learn powerful acoustic representations without the need for aligned transcriptions. The pre-trained model is then fine-tuned on a smaller labelled dataset, specifically tailored to the complexities of Arabic speech. Integrating Wav2Vec-2 enhances the system's robustness against variations in pronunciation, dialects, and ambient noise, making it well-suited for real-world applications.

This paper aims to evaluate the proposed approach by conducting extensive experiments on a benchmark Arabic speech dataset. The results showcase the advantages of employing syllables alongside the 5-gram language model and Wav2Vec-2, demonstrating the system's capability to achieve state-of-the-art performance in Arabic speech recognition.

In summary, this paper introduces a novel approach to Arabic speech recognition that leverages syllables as a fundamental linguistic unit, combined with a 5-gram language model and the powerful Wav2Vec-2 architecture. Anticipating that this blend of linguistic awareness and cutting-edge deep learning techniques will lead to significant advancements in Arabic speech recognition technology, paving the way for more accurate, contextually aware, and robust systems in the future. The next section offers a brief introduction to Arabic syllables, followed by a literature review on the state of the art of speech recognition for the Arabic language and the use of syllable-based speech recognition. Section 3.1 describes the dataset, followed by the model description in Section 3.2. Section 3.3 covers the Language Model and Decoding, followed by the Results and Discussion. Finally, the paper concludes with the Conclusion section.

2. Literature review

2.1 Introduction to Arabic syllables

Arabic is a Semitic language, and it is one of the oldest languages in the world. It is the 5th widely used language nowadays (Sayed, 2015). Modern Standard Arabic has 28 consonants and 6 vowels, 3 short and three longs (Alghamdi et al., 2004). A consonant is a basic speech sound, which is produced when the airflow is partly or fully obstructed at one or more points along the vocal tract. Vowels, on the other hand, are produced when the vocal tract is open, and the vocal cords vibrate. Vowels have high energy and distinct formant patterns (Huang et al., 2001). A syllable is a cluster of basic sound units, which is pronounced as a single complete unbroken sound. Words consist of one or more syllables. Arabic is a syllable-based language. In Arabic language a syllable must begin with a consonant followed by a vowel and contains a single vowel, short or long. Modern Standard Arabic MSA Language is based on 5 syllable structures or classes: CV, CVV, CVC, CVVC, and CVCC. Where C stands for a consonant, V stands for a short vowel, and VV stands for a long vowel. The CVCC is a rare syllable and could occur in a speech at the end of an utterance e.g., فجر، شهر، أمر. and does not appear in MSA diacritized text. No more than two consecutive consonants could occur in the Arabic language (Elshafei, 1991). The Arabic syllable never begins with two consonants. No single phoneme syllable exists in Arabic. All Arabic syllables must contain at least one vowel. Also, Arabic vowels cannot be initials and can occur either between two consonants or the final in a syllable.

Arabic syllables are categorized as either short or long, with the CV type being short, while all others fall into the long category. These syllables can also be categorized as open or closed, where an open syllable concludes with a vowel, and a closed syllable ends with a consonant. In Arabic, a syllable nucleus always comprises a vowel, and the number of syllables in a word corresponds to the number of vowels present within it. Syllables play crucial roles in articulation, cross-articulation, speech prosody, and in the pronunciation of allophones of phonemes (Elshafei, 1991). Furthermore, the structure of syllables holds significant importance in Arabic poetry and the musicality of the language.

A statistical analysis was performed using a dataset of Modern Standard Arabic (MSA) text complete with diacritics, containing 349,981 characters, 39,217 words, and 4,574 lines. This text was collected from written Arabic news articles covering topics in sports and economics.

The syllables were generated utilizing Arabic pronunciation rules via a program outlined in (Alghamdi et al., 2004). In total, there are 101,958 syllables, comprising 2,498 unique syllables within the dataset. Table 1 illustrates each syllable structure alongside its respective percentage concerning the total syllables. Columns 3 and 4 present the most and least repeated syllables. Notably, the most prevalent syllable is CV, while the least common is CVCC. In instances where the last word in a sentence concludes with CVC CV, it is articulated with a stop instead of ending with CV, resulting in a closing CVCC.

Referring to Table 1, it can be deduced that, on average, a diacriticized word consists of 8.9 characters. The approximate number of characters per syllable is 3.4, with an average of 2.66 syllables per word. However, it was observed that these statistics were influenced solely by the text, including an end-case diacritic. If the final consonant is not vocalized to mirror actual speech, more instances of CVCC may be observed. These statistics might also vary based on text domains and the size of the text corpus.

Table 1: General Syllables Percentages.

Syllables	%	Most Repeated	Least repeated
CV	42.66957	وَ = 3139	أ = 1
CVC	29.76593	تِلْ = 1476	دُضْ = 1*
CVV	23.1783	مَا = 1025	مَلَا = 1*
CVVC	4.385217	لَاَزْ = 202	وِينْ = 1*
CVCC	0.000981	أَمْرْ = 1	أَمْرْ = 1

* Indicates that there are other syllables having the same number of occurrences.

Theoretically speaking, the possible number of CV cases is (28 consonants X 3 vowels) = 84 ; for CVC = $28 \times 3 \times 28 = 2352$, and CVV=84, CVVC=2352, a total of 4872. However, many of them may not appear in the MSA.

2.2 Related works

The motivation behind using syllables comes from recent research on syllable-based recognition.

Syllable-based speech recognition was proposed and evaluated for several languages, (Jasmin et al., 2022), (Fantaye et al., 2019), (Aslyan, 2011), (Majewski, 2008), such as Malayalam, Chaha, Turkish, and Polish low resource languages, as well for Mandarin Chinese (Zhou et al., 2018). The syllable role was also investigated in studies of human perception (Kirchhoff, 1996), (Wu et al., 1998), which demonstrate the central role of the syllable played in human perception and generation of speech. One important factor that supports the use of syllables as the acoustic unit for recognition is the relative insulation of syllables from pronunciation variations arising from the addition and deletion of phonemes as well as co-articulation.

In the following literature review, we present selected methodologies that have been used for different languages for speech recognition. Developing a speech recognition system with a language like Arabic presents several challenges. The first set of challenges relates to the different dialects of Arabic. While Arabic has millions of native speakers, each area has its own dialect of Arabic. Consequently, the number of datasets available for some of these is few while other dialects do not even have datasets available for them (Alsayadi et al., 2022). In addition, some dialects lack a well-defined orthographic system like the Egyptian dialect (Ali et al., 2017). These issues combined led to a lack of performance of the developed Arabic Speech Recognition (ASR) systems compared to what has been achieved in other languages. Furthermore, the absence of the diacritized text (with tashkeel) is a hindering factor in developing high-quality speech recognition systems. The studies developing ASR systems using a non-diacritized version of Arabic usually leads to limited accuracy because non-diacritized Arabic words may have different pronunciations and meaning depending on the context. Finally, the recent work in (Alsayadi et al., 2022) found that focusing on one dialect resulted in better accuracy than developing a multi-dialect system. An approach was demonstrated by students at Zewail City: University of Science and Technology to create an ASR pipeline for Egyptian Arabic dialects by adapting two models and comparing them. These models are bidirectional RNN and unidirectional RNN

(Mahrous et al., 2021). They used the MGB-3 data set to train their two models. They built their first model by using a gated recurrent unit (GRU), with additive attention. The second model was a bidirectional GRU with an attention-based layer. They were able to achieve a loss for unidirectional GRU of 0.3473 and a loss of 0.4599 for bidirectional GRU (Mahrous et al., 2021).

Authors (Alsayadi et al., 2021) propose an end-to-end model for the Arabic ASR system in which they used long short-term memory networks (LSTM). The proposed model makes use of LSTM in learning long-term dependencies. In addition, they use CNN to further extract useful information from the input. Is obvious from the proposed model architecture in Figure 1.

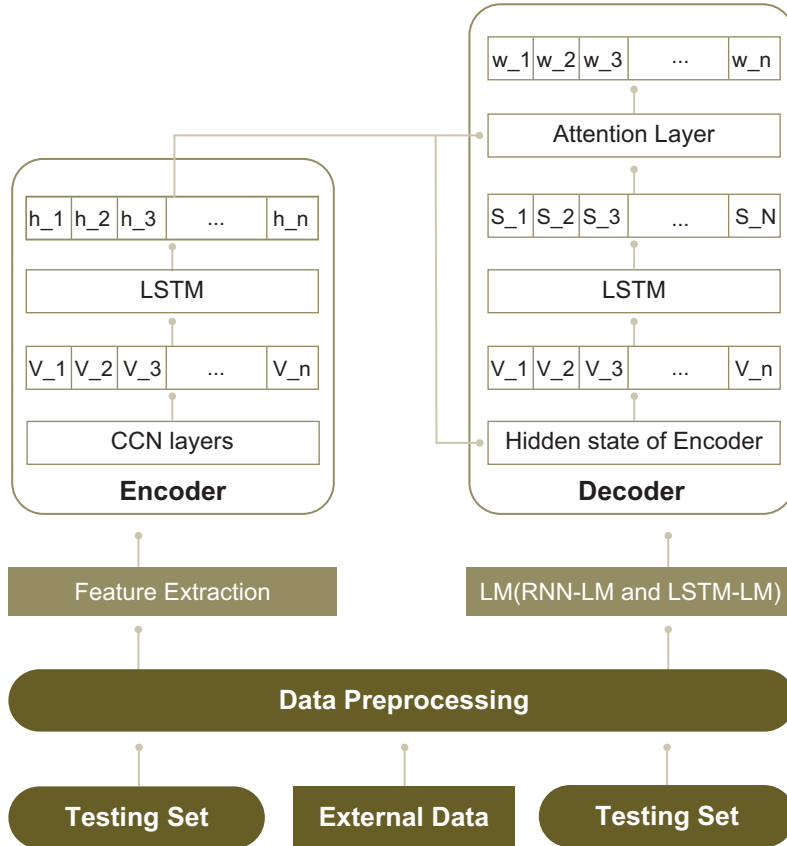


Figure 1: model architecture (Alsayadi et al., 2021)

The first step is pre-processing the data in the corpus, language model and lexicons. Then LSTM and CNN are used for the encoder. The encoder role is to build the acoustic model with CNN to prepare the input as vectors and send them to LSTM layers. For the decoder part, the output from the encoder is received and LSTM with attention decodes the encoder output and obtains the predication. The model was tested in Standard Arabic Single Speaker Corpus (SASSC) and obtained WER 28.48 % . Although the accuracy and WER are not impressive, they give us more insight into the model and architectures used.

Automatic Speech Recognition (ASR) has made remarkable strides in recent times, achieving levels of accuracy comparable to human transcribers, prompting discussions about whether machines have attained human-level performance. Previous research in this domain has primarily focused on the English language, using modular Hidden Markov Model-deep neural network (HMM-DNN) systems (Hussein et al., 2022). They aimed to fill the gap by conducting an extensive benchmarking study on Arabic speech recognition, encompassing end-to-end Transformer ASR, modular HMM-DNN ASR, and human speech recognition (HSR) across various Arabic dialects. They explore the performance of end-to-end Transformer ASR for Arabic and its dialects, achieving notable Word Error Rates (WER) of 12.5%, 27.5%, and 33.8% for the MGB2, MGB3, and MGB5 challenges, respectively. However, despite the impressive advancements in ASR for Arabic, the study concludes that human performance remains notably superior to machine performance. The authors highlight an absolute WER gap of 3.5% on average, underscoring the continued challenges in achieving human-level recognition accuracy for Arabic speech.

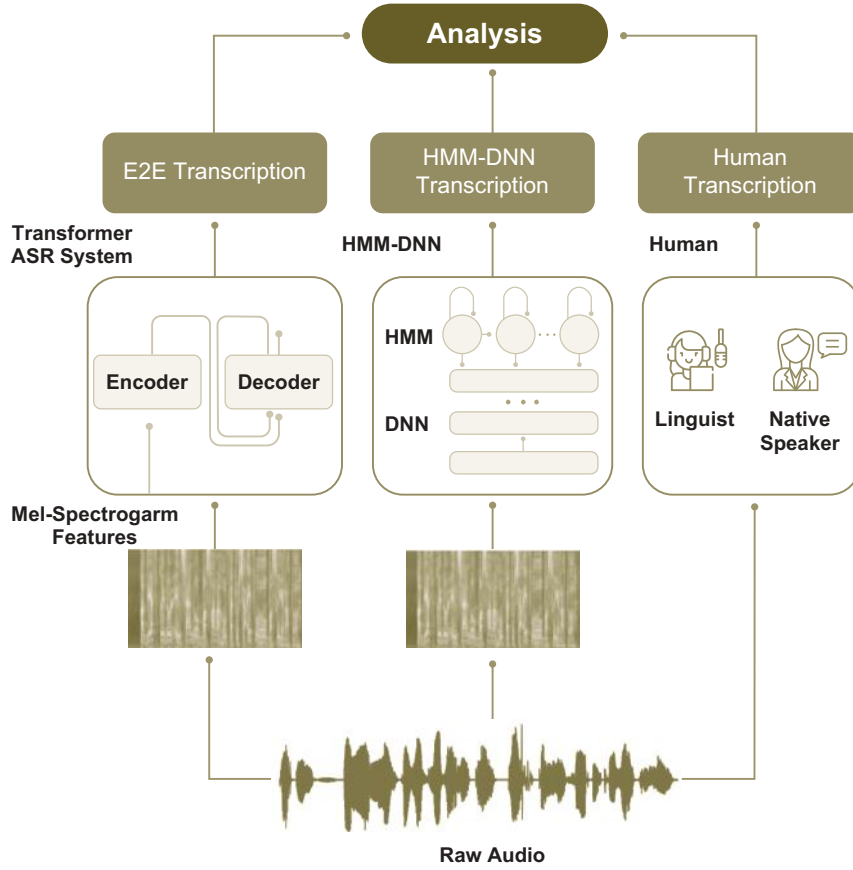


Figure 2: Core Architecture of the paper (Hussein et al., 2022)

The performance of well-trained speech recognition systems, based on high-quality full bandwidth speech data, often suffers when applied in real-world scenarios, especially in telephone speech recognition. The limited bandwidth of transmission channels in telephone communications poses significant challenges for accurate recognition. (Azmi & Tolba, 2008) investigated telephone speech recognition of Egyptian Arabic speech using syllables as acoustic units. The authors designed a speaker-independent Hidden Markov Models (HMMs)-based speech recognition system using the (HTK) tool kit. The database used for both training and testing comprised recordings from forty-four Egyptian speakers. They found that syllable-based speech recognition has better accuracy in noisy environments over conventional phoneme-based speech recognition.

One of the advantages of using syllables lies in their longer time frame, typically encompassing two or more phones. This extended duration offers a more parsimonious framework for modelling pronunciation variations in spontaneous speech, contributing to the system's robustness and adaptability to real-world environments.

Additionally, syllable-based recognition requires a relatively smaller number of acoustic units, resulting in faster processing times compared to word-based recognition. This efficiency makes syllable-based ASR systems more suitable for real-time applications and more computationally efficient.

3. Methodology

3.1. The Arabic News Corpus Dataset

The primary dataset used in this research, known as the Arabic News Corpus, comprises 15 hours of WAV recordings with diacritized Modern Standard Arabic (MSA) transcriptions. Originally, this corpus was approximately 5.7 hours long, developed through a grant from King Abdulaziz City of Science and Technology (KACST) in 2008. Over time, it has expanded and undergone refinement through grants from various public, institutional, and private sources. Additionally, it now includes fully diacritized word-level transcriptions. Crucially, the Arabic News Corpus was created using automated software that converts text into syllables. While this method helped generate a substantial volume of data necessary for our task, it's important to note that the automated syllabification process may have introduced errors due to issues in the diacritized text or the network of pronunciation rules. We estimate these errors to account for approximately 3% of the dataset. Despite this potential source of inaccuracies, our model has yielded significant results, demonstrating its robustness in handling minor inconsistencies within the training data. The dataset was divided into training, validation, and testing sets. The training set, comprising 12.5 hours of recordings, constituted the largest portion, while 1 hour was assigned to the validation set, and the remaining 1.5 hours were allocated to the testing set.

The dataset's size significantly influences the methodology, with smaller datasets possibly requiring more specific approaches. Our dataset, with varied sound file durations, provides a sturdy foundation for our study. The average duration of each sound file is 6 seconds, ranging from a minimum of 0.6 seconds to a maximum of 22 seconds. Initially, discrepancies were identified in the sampling sizes across the sound files, leading to conflicts during the training phase. These inconsistencies were subsequently rectified during pre-processing, ensuring uniformity across the dataset and streamlining the training process.



3.2. Model

The Wav2Vec 2.0 model serves as a comprehensive architecture for speech recognition tasks. It incorporates a multi-layer convolutional feature encoder, which ingests raw audio input and produces latent speech representations. These representations are then fed to a Transformer that develops context-aware encodings capturing information from the entire sequence. A unique aspect of Wav2Vec 2.0 lies in its use of a quantization module, which discretizes the feature encoder's output into a finite set of speech representations using product quantization. This model has the capability to capture dependencies over the entire sequence of latent representations end-to-end, providing robust speech recognition capabilities. The design of the Wav2Vec 2.0 model facilitates efficient processing of raw audio data, successfully capturing important features of the speech that prove beneficial for a range of downstream tasks.

3.3. Fine Tuning

Wav2Vec 2.0 offers numerous pre-trained checkpoints for fine-tuning, including a multilingual variant, XLSR, trained across 56,000 hours and 53 languages. For our Arabic syllable recognition task, we fine-tuned this model by minimizing the Connectionist Temporal Classification (CTC) loss, enhancing the model's effectiveness. During the fine-tuning process, we employed a specially tailored vocabulary from our text data, consisting of 3086 tokens representing the unique syllables in the dataset, including padding, unknown, and word delimiter tokens. This vocabulary effectively encapsulates the diverse elements of the speech data. We utilized the wav2vec tokenizer and feature extractor to process the data. The tokenizer converted our data into a token format interpretable by the wav2vec model, while the feature extractor transformed the raw audio data into suitable features for the model's input. Notably, the first component of XLSR, a stack of Convolution Neural Networks (CNN) layers for extracting acoustically meaningful but contextually independent features, required no further fine-tuning. This combination of a custom vocabulary and wav2vec-specific tools, alongside effective pre-training utilization, facilitated efficient fine-tuning of the model for Arabic syllable recognition. To bolster the model's robustness and generalization capabilities during fine-tuning, we incorporated data augmentation techniques. These techniques included Gaussian Noise injection with an amplitude range of 0.001 to 0.012, Time Stretch with a rate between 0.8 and 1.25, and Pitch Shift with semitone alteration ranging from -4 to 4.

The fine-tuning was conducted under the following hyper parameters:

Table 2: used hyper parameters.

Hyperparameter	Value
Space augment	True
Attention Dropout	0.001
Hidden Dropout	0.008
Feature Proj Dropout	0.0
Mask Time Prob	0.1
Laydrop	0.0001
Per Device Train Batch Size	2
Gradient Accumulation Steps	8
Num Train Epochs	20
Learning Rate	1e-5
Eval Steps	500

The combination of effective utilization of pre-training, strategic data augmentation, and carefully selected hyper parameters, enabled efficient fine-tuning of the model for Arabic syllable recognition.

3.4. Language models and decoding

Our system incorporates a 5-gram language model (LM) trained on our training transcriptions as well as a large corpus of specifically curated Arabic syllables. This strategy enables effective capture of the statistical properties of Arabic syllable sequences, which is crucial for our task. The hyper parameters for the language model were optimized based on the performance on our validation set. For the test performance measurement, we employed a beam search decoder with a beam size of 1500, ensuring a comprehensive search over potential outputs.

4. Results

Our model was evaluated under various conditions, as shown in Table 3, to investigate the impact of different components and conditions on its performance.

Table 3: Experiments results

Experiment	WER
Model without Data Augmentation	0.175
Model with Data Augmentation	0.125
Model without 5-gram Language Model (LM)	0.125
Model with 5-gram Language Model (LM)	0.066

The use of data augmentation techniques significantly improved performance, with the Word Error Rate (WER) decreasing from 0.175 to 0.125. Furthermore, the incorporation of the 5-gram Language Model (LM) resulted in a further reduction of the WER to 0.066. The evaluation of the model's performance on text with and without Tashkeel provides insight into the model's ability to handle Arabic text in different forms.

Our final model, which incorporates data augmentation techniques and the 5-gram Language Model (LM), was evaluated under various conditions, as shown in Table 4.

Table 4: Results with different metrics

Metric	With Tashkeel	Without Tashkeel
WER	0.06624	0.05959
CER *	0.03044	0.03994
MER	0.06577	0.05916
WIL	0.11201	0.09928
WIP	0.88798	0.90071

*Notice for CER: With Tashkeel, CER count the letter with Tashkeel as one character

In this table, WER stands for Word Error Rate, CER for Character Error Rate, MER for Match Error Rate, WIL for Word Information Lost, and WIP for Word Information Preserved. The results suggest that our model performs effectively in Arabic syllable recognition regardless of the presence or absence of Tashkeel.

Below are samples transcription based on test set:

Table 5: transcription sample

Reference	Prediction
مِمَّ مَاعَ مَلَعَ تَوِي دَعِ وَأَدِسِ يَاجِي تِنْتُ قَدْ دَرْبِ زَبَعِ تِ مِلَّ يَارَاتِنِ وَتْ لَآتِ مَءِ تِ مِلِّي وُنْ دُوْلَازْ	مِمَّ مَاعَ مَلَعَ تَوِي دَعِ وَأَدِسِ يَاجِي تِنْتُ قَدْ دَرْبِ زَبَعِ تِ مِلَّ يَارَاتِنِ وَتْ لَآتِ مَءِ تِ مِلِّي وُنْ دُوْلَازْ
وَإِثْنِيْنِ وَسْتَقِيْنِ فُلْمِ عَهْ	وَأَثْنِيْنِ وَبِسْتَقِيْنِ فُلْمِ عَهْ

5. Discussion

This work presents an effective approach to syllable-based Arabic speech recognition using a model that incorporates data augmentation techniques and a 5-gram language model. However, two key points emerge for future enhancements:

- **Dataset Errors:** The Arabic News Corpus, prepared via automated software, contains an estimated 3% error rate. Refining this process could improve data accuracy and potentially enhance model performance.
- **Language Model Complexity:** While our 5-gram language model contributed significantly to performance, it is conceivable that a more complex model, such as a transformer-based one, could yield improved results with an increased data volume.

In order to showcase the competitiveness of our approach, we conducted a further comparison with existing models. Specifically, we assessed the performance of our two-step model, employing the wav2vec model and a 5-gram language model for syllable recognition, followed by the mT5 model for syllable-to-word conversion. This was compared with Google's API, which offers a direct speech-to-text conversion service. For a fair omparison, we evaluated both systems on the first 100 audio samples from the test split of the Common Voice Arabic dataset ("common_voice · Datasets at Hugging Face"). This comparison is done without Tashkeel, so it can give a better indication relative to Google API. The mT5 model is a highly versatile pre-trained text-to-text transformer with extensive multilingual capabilities. This model is essentially a multilingual adaptation of T5, having undergone pre-training on a dataset encompassing 101 languages. Consequently, it is a commendable choice for text-to-text tasks. The mT5- -small model achieved accuracy for the Arabic Language reaching 65.2% (Xue et al., 2020).



Table 6: Comparison with Google API.

metric	Our pipeline	Google API
WER	0.412	0.459
CER	0.135	0.127
MER	0.409	0.446
wil	0.630	0.680
wip	0.369	0.319

Despite training on a significantly smaller dataset of approximately 15 hours of speech, our model outperforms Google's API in key metrics such as WER, MER, and WIP. However, Google's API shows marginally better CER, areas our model could potentially improve upon. The disparity in training data highlights our model's potential and efficiency, suggesting further enhancements with larger, more diverse training datasets.

To foster transparency and reproducibility in research, our models are publicly available on the Hugging Face model hub. The model trained with a 5-gram language model can be accessed with the Syllables to text model in the last two references.

6. Conclusion

We introduce a groundbreaking approach to Arabic speech recognition by seamlessly integrating three fundamental elements: syllable-based segmentation, a robust 5-gram language model, and the cutting-edge Wav2Vec-2 architecture. Our work, founded on meticulous experimentation and rigorous evaluation, showcases the unparalleled effectiveness of this innovative method.

Syllable-based segmentation serves as the cornerstone of our approach, allowing us to navigate the intricate morphological structure of the Arabic language. By capturing nuanced phonetic details, this approach substantially enhances recognition accuracy, opening avenues for the creation of more reliable and practical Arabic speech recognition systems.

In addition, the seamless integration of a 5-gram language model significantly enhances our system's ability to handle the complex nuances present in Arabic. Addressing word-level variations and contextual dependencies, this linguistic modelling elevates overall system performance, ensuring precise recognition even in the most challenging linguistic scenarios.

The adoption of the Wav2Vec-2 architecture as our acoustic model further fortifies the system's resilience. Leveraging unsupervised pre-training on extensive, unlabelled speech data and fine-tuning on meticulously curated labelled datasets tailored for Arabic speech recognition, our system acquires robust acoustic representations. These representations empower the system to adeptly handle variations in pronunciation and ambient noise, ensuring reliability across diverse settings.

Our exhaustive experimentation culminated in remarkable Word Error Rates (WER) of 0.06624 with Tashkeel and 0.05959 without Tashkeel, underscoring the exceptional performance and reliability of our proposed approach.

To conclude, our paper's contributions are dual-fold: introducing a novel methodology harnessing the innate power of Arabic syllables, a resilient 5-gram language model, and the state-of-the-art Wav2Vec-2 architecture for speech recognition and validating its efficacy through comprehensive evaluation. We firmly believe that this research will catalyse a transformative shift in Arabic speech recognition, inspiring further advancements in the development of robust, accurate systems not only for Arabic but also for other complex languages.

7. References

- AbuZeina, D., Al-Khatib, W., Elshafei, M., & Al-Muhtaseb, H. (2011). Cross-word Arabic pronunciation variation modelling for speech recognition. *International Journal of Speech Technology*, 14(3), 227–236. <https://doi.org/10.1007/s10772-011-9098-0>
- Alghamdi, M., Almuhtasib, H., & Elshafei, M. (2004). Arabic phonological rules. *King Saud University Journal: Computer Sciences and Information*, 16, 1-25.
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449-12460.
- Sayed, F. (2015.). A few surprising facts about the Arabic language. British Council. <https://www.britishcouncil.org/voices-magazine/surprising-facts-about-arabic-language>
- Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.
- Elshafei, M. (1991). Toward an Arabic text-to-speech system. *The Arabian Journal for Science and Engineering*, 16(4B), 565-583.

- Jasmin, S., Samuel, A. A., & Rajan, R. (2022). A study on conventional and syllable-based approaches for automatic speech recognition in Malayalam. *Sādhana*, 47(4). <https://doi.org/10.1007/s12046-022-02058-z>
- Fantaye, T. G., Yu, J., & Hailu, T. T. (2019, December). Syllable-based Speech Recognition for a Very Low-Resource Language, Chaha. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 415-420).
- Aşlyan, R. (2011). Syllable Based Speech Recognition. *Speech Technologies*, 263.
- Majewski, P. (2008). Syllable based language model for large vocabulary continuous speech recognition of polish. In *Text, Speech and Dialogue: 11th International Conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008. Proceedings 11* (pp. 397-401). Springer Berlin Heidelberg.
- Zhou, S., Dong, L., Xu, S., & Xu, B. (2018). Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. *arXiv preprint arXiv:1804.10752*.
- Kirchhoff, K. (1996, October). Syllable-level desynchronisation of phonetic features for speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2274-2276). IEEE.
- Wu, S. L., Kingsbury, E. D., Morgan, N., & Greenberg, S. (1998, May). Incorporating information from syllable-length time scales into automatic speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98* (Cat. No. 98CH36181) (Vol. 2, pp. 721-724). IEEE.
- Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Alotaibi, B., & Fayed, Z. T. (2022). Deep investigation of the recent advances in dialectal Arabic speech recognition. *IEEE Access*, 10, 57063-57079.
- Ali, A., Vogel, S., & Renals, S. (2017, December). Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 316-322). IEEE.
- Mahrous, M., Noursir, A., Wael, A., & Elshafei, M. (2021). Speech recognition using Attention Based Bidirectional RNNs for Egyptian Arabic dialect. *CIE556, Term Paper*.
- Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., & Fayed, Z. T. (2021). Arabic speech recognition using end-to-end deep learning. *IET Signal Processing*, 15(8), 521-534

- Hussein, A., Watanabe, S., & Ali, A. (2022). Arabic speech recognition by end-to-end, modular systems and human. *Computer Speech & Language*, 71, 101272.
- Azmi, M. M., & Tolba, H. (2008, July). Syllable-based automatic Arabic speech recognition in a noisy environment. In *2008 International Conference on Audio, Language and Image Processing* (pp. 1436-1441). IEEE.
- Common_voice · datasets at hugging face. common_voice · Datasets at Hugging Face. (2021). https://huggingface.co/datasets/common_voice/viewer/ar/test (accessed Jul. 30, 2023).
- Hugging face – the AI community building the future. Hugging Face – The AI community building the future. (2023.).
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- https://huggingface.co/IbrahimSalah/Arabic_speech_Syllables_recognition_Using_Wav2vec2/settings (accessed Jul. 30, 2023).
- Hugging face – the AI community building the future. Hugging Face – The AI community building the future. (2023). https://huggingface.co/IbrahimSalah/Arabic_Syllables_to_text_Converter_Using_MT5/settings (accessed Jul. 30, 2023).

٨- نبذة تعريفية عن الباحث أو الباحثين:

AUTHOR BIODATA:	بيانات الباحث :
Ibrahim Abdalaal is a recent graduate student in the Department of Communications and Information Engineering, College of Engineering, University of Science and Technology in Zewail City. His research interests include Speech Recognition, Deep Learning and Using Artificial Intelligence with Arabic Language.	إبراهيم عبدالعال، طالب حديث التخرج في (قسم الاتصالات وتكنولوجيا المعلومات) (كلية الهندسة) في جامعة العلوم والتكنولوجيا بمدينة زويل ب (مصر). تدور اهتماماته البحثية حول تعرف على الكلام، تعلم متعمق، استعمال الذكاء الصناعي مع اللغة العربية.

(0937-5398-0007-0009)



معرف أوركيد (ORCID) :

Email: s-ibrahimsaad@zewailcity.edu.eg

AUTHOR BIODATA: Mohamed Abdelwahed is a recent graduate student in the Department of Communication and Information Engineering, College of Engineering, University of Science and Technology in Zewail City. His research interests include Speech Recognition, Deep Learning and Using Artificial Intelligence with .Arabic Language	بيانات الباحث: محمد عبد الواحد، طالب حديث التخرج في (قسم الاتصالات وتكنولوجيا المعلومات) (كلية الهندسة) في جامعة العلوم والتكنولوجيا بمدينة زويل (مصر). تدور اهتماماته البحثية حول تعرف على الكلام، تعلم متعمق، استعمال الذكاء الصناعي مع اللغة العربية.
--	---

(ORCID: 0009-0003-1679-1556)

معرف أوركيڍ (ORCID) : ID

Email: s-mohamedabdelwahed@zewailcity.edu.eg

AUTHOR BIODATA: Moustafa Elshafei is a Full Professor and Director of the Department of Communications and Information Engineering, College of Engineering, and acting director of the School of Computational Sciences and Artificial Intelligence, University of Science and Technology in Zewail City. He obtained his Ph.D. from McGill University, Canada, in 1982. His research interests include Arabic Speech Recognition and Artificial Intelligence applications in Robotics and Industrial .Automation	بيانات الباحث: مصطفى الشافعي، أستاذ / مدير برنامج هندسة الاتصالات وتكنولوجيا المعلومات و المدير المكلف لكلية العلوم الحاسوبية والذكاء الاصطناعي بمدينة زويل: جامعة العلوم والتكنولوجيا بجمهورية مصر العربية . حصل على درجة الدكتوراة من جامعة مكجيل بكندا عام ١٩٨٢ وتدور اهتماماته البحثية حول استعمال الذكاء الصناعي في اللغة العربية والروبوتات وفي التطبيقات الصناعية.
--	--

(ORCID: 0000-0002-8205-666X)

معرف أوركيڍ (ORCID) : ID

Email: Moelshafei@zewailcity.edu.eg

