توظيف المدوّنات اللّغوية المُحوسبة في اختبار وتقييم المحلّلات الصّرفية
- محلل الخليل الصّرفي ومدوّنة NEMLAR - نموذجاً

Employing Computerised Linguistic Corpora in Testing and Evaluating Morphological Analyzers - AlKhalil Morphological Analyzer and NEMLAR Corpus - as a Model.

جلال ثامر*
جامعة عبد الرحمان ميرة – بجاية (الجزائر)
مخبر التأويل وتحليل الخطاب
Djalal.tameur@univ-bejaia.dz

تسعديت لحول
جامعة عبد الرحمان ميرة – بجاية (الجزائر)
مخبر التأويل وتحليل الخطاب
tassadit.lahouel@univ-bejaia.dz

تاريخ الإرسال: 2024/03/17	تاريخ القبول:2024/08/09

الملخص:

من مُنطلق فلسفة الحوسبة القائمة على تكامل وتعاضُد العمل بين النّظم والبرمجيات الموجّهة لمعالجة اللّغات الطبيعية آلياً، والمصادر اللّغوية المُهيّئة حاسوبياً، وفق المبدأ المعروف في عالم الحوسبة بتدوير المصادر المتاحة،من أجل خلق نفعيّة تبادليّة تجمع هذه الموارد البيانية (المعيارية) بتلك النّظم والأدوات المعلوماتية دونما اعتبار لطبيعتها أو نهجها أو حتى للمجال الذي طوّرت لأجله أو تُوظف فيه. حيث تَظهر براغماتية هذا المبدأ في نزوعه إلى توسيع وتنويع نطاقات الإفادة من هذه الموارد النَّصية المُحوسبة، بتوظيفها إمّا لرفع كفاءة بعض التطبيقات الحاسوبية وتحسين مُخرجاتها، أو لاختبارها وتقييم عملها وأدائها. عندها تصبح المدوّنات المحوسبة (corpus) عُمدة المصادر اللّغوية توظيفاً واستخداماً ضمن إجراءات الاختبار والتّدريب، للكثير من النّظم والبرمجيات وفي مقدمتها المحلّلات الصّرفية، لتقييم أدائها وقياس نسبة وثوقيّتها على مستويات عدّة: -التّحليل -التّشكيل -الوَسم...الخ، واختبار جاهزيّتها في تقديم تحليلات صحيحة ووُسوم دقيقة للمُدخلات، قياساً على ما هُيّئت ووُسمت به قبلاً هذه المدوّنات.

الكلمات المفتاحية:

- الحوسبة - المصادر اللّغوية - المدوّنات اللّغوية المُحوسبة - المحلّلات الصّرفية - اختبار وتقييم المحلّلات.

Abstract :

Within the realm of computing philosophy, which emphasizes the synergy between systems and software for automated natural language processing and computationally ready linguistic resources, a key principle is the efficient recycling of resources. This approach fosters a reciprocal benefit, merging standardized data resources with diverse informational systems and tools, irrespective of their origin or purpose. This principle's pragmatism is evident in its aim to broaden and diversify the use of these digital textual resources. They are employed to boost the effectiveness of computer applications and their outcomes or to assess and improve their functionality. Central to this are Computed corpora, pivotal for testing and training various systems, notably morphological analyzers.

*المؤلف المرسل: جلال ثامر.

These are assessed for performance and reliability in analysis, formation, and tagging, ensuring their capacity to deliver accurate and precise analyses and tags, reflecting the previously prepared and annotated content in these corpora.

Keywords:
Computing, Linguistic Resources, Computed Linguistic Corpora, Morphological Analyzers, Testing and Evaluation of Analyzers.

مقدمة:

تحتاج المعالجة الآلية للّغات الطبيعية ومن بينها اللّغة العربية تساوقاً مع مبدأ النّمذجة؛ إلى نماذج مُحوسبة (models) وإلى نُظم تحليل عالية الأداء، تعكس المنطق الطبيعي لهذه اللّغات وتراعي -قدر الإمكان- التمثيل الواقعي لظواهرها اللّغوية في مستوياتها البنائية (الصّوتية والصّرفية التركيبية والدّلالية والتّداولية) تحقيقاً لمقتضيات الدّقة والكفاءة المرجوّة من هذه النّماذج والنّظم وتحصيلاً للمأمول من حوسبة اللّغات ومعالجتها آليـاً.

إلّا أنّه -وفي الوقت ذاته- تحتاج أغلب هذه الأدوات والنّظم لأداء مهامها إلى **بيانات نصيّة**، ومصادر لغوية **مُهيّئة حاسوبياً**، تُعرف حديثاً بـ: (**المدوّنات اللّغوية الحاسوبية corpus**) والتي استحالت مؤخراً إلى وقود خام لهذه الأدوات، بالنّظر إلى مصادرها ومادّتها (الكمية والنوعية) والمغذّي الرّئيس للكثير من النّظم الحاسوبية بالبيانات المعيارية، والأدلّة النّصية على الظواهر اللّغوية المستقاة من النّصوص ذات التمثيل الحقيقي للواقع اللغوي، على مستويات عِدّة (الفصيح، العامي اللّهجي، المكتوب، المنطوق...الخ). خصوصاً إذا هُيّئت هذه المصادر حاسوبياً بشكل ثري ودقيق دعماً لمتطلبات المعالجة الآلية للّغات الطبيعية. إذ تُرتّب اليوم هذه المدوّنات وفق زاوية بحثية راهنة -إلى جانب زوايا عدّة- على أنها إحدى أكثر "**المنهجيات الاختبارية**" حضوراً وتوظيفاً؛ باعتمادها أداةً فعّالة ضمن إجراءات الاختبار والتّدريب، أو التّقييم وتقييس أداء الكثير من أدوات التحليل الحاسوبي للنّصوص، وعلى رأسها المحلّلات اللّغوية والصّرفية تحديداً، حيث وَجّهت الرّؤى التطبيقية المتنوّعة والخيارات الوظيفية المميّزة التي تُتيحها المدوّنات، إلى توظيفها في مجالات بحثية واسعة وبينيّة (لسانية وحاسوبية). فمن بين الخيارات الوظيفية التي تكفلها المدوّنات المُحوسبة والمشكّلَة تشكيلاً تاماً، والموسّمة بشتّى أنواع الوُسوم، هو استدعاؤها وتوظيفها بوصفها **مجموعة اختبار**[1] **Test collection** لقياس دقّة وكفاءة المحلّلات الصّرفية قيد التجريب أو التّطوير (محلل الخليل الصرفي نموذجاً AlKhalil Morpho Sys 1.0). ولمعرفة درجة وُثوقيّة هذه المحلّلات في تقديمها تحليلات

صحيحة للمُدخلات، وتحديدها بدقّة وُسوم الكلمات، وكذا التحقّق من نَجاعة مقاربات التّشكيل المنتهجة فيها.

وعليه جاءت هذه الدراسة لتسلّط الضوء على واحد من أشهر المحلّلات الصّرفية الموجّهة للّغة العربية،وهو(**برنامج الخليل الصّرفي**AlKhalil Morpho Sys).ومحاولة التّكشيف عن أهم الخطوات المنتهجة من قِبل مُطوّريه للتّحقق من درجة كفاءته وموثوقيّته على مستويات عدّة بدءً من التحليل الصّرفي وانتهاءً بالتّشكيل الآلي للكلمات والنّصوص. بتوظيف مدوّنة لغوية(**أداةً للاختبار**) هي مدوّنة **نملار**(Corpus NEMLAR)أو(شبكة موارد اللّغة الأورومتوسطية). وتقديم الأرقام والنّسب الدّقيقة المترتبة عن هذا الاختبار مع نتائج التّقييم في كلّ مستوى من تلك المستويات.

وعليه يُحاول المقال الإجابة عن إشكالية رئيسة هي: - كيف تمّ اختبار عمل وأداء محلّل الخليل الصرفي AlKhalil Morpho Sys 1.0 استناداً على مدونة NEMLAR؟ وما مدى شمولية المدونة(**أداة الاختبار**) لتغطية مستويات وإجراءات التقييم لمنظومة المحـلّل؟

وللإحاطة المثـلى بالموضوع وللإجـابة عن إشكاليّته؛ استدعت الدّراسة تأسيس البحث على أسئلة فرعية أخرى انبثقت عن إشكاليته الرئيسة وتمثلت في: - مـا خصائص مدونة (NEMLAR) التي رُصدت بوصفها مجموعة اختبار لمحلّل الخليل الصّرفي؟ - ومـا المواصفات والخيارات اللّغوية المشتملة عليها والمُؤهلة لها أداةً لاختبـار المحلّل؟ - مـا أهم المستويات التي أُختبر محلّل الخليل فيـها؟ وما النّتائج التي أفرزها الاختبـار والتقييـم؟

معتمدين في هذه الدراسة على **المنهج الوصفي** الأنسب لطبيعة الموضوع.

1- المدوّنـات اللّغويـة المُحوسبـة (دلالة المصطلح والمفهوم):

تستدعي المنهجية قبل الاسترسال في عرض التنوّعات المصطلحية والمفاهيمية لـ:(**المدوّنـة اللّغويـة المُحوسبة**) ضرورة تنزيلها أولاً ضمن السّياق المعرفي الحاضن لها، وإحالتها على المجال العلمي المتموضعة فيه، المعـروف حديثاً[2] بـ: (**لسانيات المُدوّنات** Corpus Linguistics)، أو (لسانيات المُتـون) كما يرجّح البعض تسميتها. ذلك أنّ ورود مصطلح «(المدونة) في سياق لسانيات المدوّنات اللّغوية يقصد بها المدونّة اللّغوية المُحوسبة»[3]

فإذا عمدنا إلى محاولة تقديم تعريف لـ:(لسانيات المدوّنات) تَلوح أمامنا مفارقات قد تصل إلى حدّ الجدل والتّنازع، حول ماهية هذا المجال المعرفي، كما تتراص أمامنا تعريفات وتنوّعات مفاهيمية كثيرة،

حاولت كل منها مقاربة هذا المجال العلمي من زاوية معينة، ومردّ ذلك -في تقديرنا- إلى تلك المرونة التي يتصف بها هذا الحقل اللّغوي الحديث نسبياً في التمدّد على مجالات بحثية كثيرة لغوية وغير لغوية، نتيجة اقترانه بالمستجد التِّقاني الحاسوبي وتطبيقاته، يُضاف إليها التركيبة التعدّدية لمناحيه الوظيفية؛ بدليل انشطار الأغراض البحثية المتوسّلة له، أو القائمة على مادّته ومصادره المعيارية لبلورة نتائج أكثر واقعية ومصداقية. مؤكّدةً قدرة هذا الحقل على تشكيل وصياغة منهجيات البحث في كثير من الجوانب اللغوية وغيرها.

فإذا اعتبر البعض لسانيات المدونات«من العلوم الحديثة التي أحدثت تغييراً منهجياً في دراسة اللغة»[4] يصرّ نفرٌ غير قليل من الباحثين على عدّها بمثابة (**المنهجية**) التي تنساب إلى أغلب العلوم اللغوية، من خلال تَشكّلها في صورة مقاربة منهجية أصبحت تجنح إليها معظم الدراسات والبحوث اللّسانية. ويترتب عن هذا التّصور أو المفهوم أي (**مقاربة منهجية**) إخراج لسانيات المدوّنات عن دائرة العلم والاستقلالية العلمية، بحصرها والانزواء بها في نطاق منهجية بحث لغوية تتكئ على نصوص اللّغة الطبيعية لدراسة الظواهر اللّغوية في مستويات مختلفة.

كما نقف في سياق تصنيفها بـ (المنهجية) على فريق يُفرّعها بحسب تنوّع خيارتها الوظيفية -كما أسلفنا- إلى منهجيات وتحديداً (**منهجية اختباريّة Empirical**)[5] يلجأ إليها الباحثون اليوم بصورة لافتة، ومتزايدة باعتبارها «مصدراً حيوياً للبيانات التي يتمّ تحليلها إذا ما أُريد إجراء بحث أو تقييم في عدد من التخصصات»[6] ولا يقف الأمر عند حدّ اختبار أو تقييم النّظم المُحوسبة، بل يتعدّاه إلى اختبار النّظريات اللّغوية ورهانات تطبيق مبادئها عملياً فهي وإن لم«تعتبر نظرية في اللّغة إلا أنّها تسهم في إدراك نتائج اختبارية دقيقة وموثّقة، وبذلك تسهم لسانيات المدوّنات في بناء النّظريات وتدقيقها والتّثبت منها وتطويرها على المدى البعيد»[7] وهذا المنحى أو التصنيف (منهجية اختبارية) هو الذي اخترناه وركّزنا عليه ليتطابق ويتوافق تماماً مع الغرض من هذه الدراسة[8]

لكن ما يمكن الاطمئنان والركون إليه؛ هو اعتبار لسانيات المدوّنات بالنّظر إلى توسّعاتها وتمدّداتها في العقود الأخيرة على مجالات بحثية كثيرة في الدراسات اللّغوية التطبيقية والبينيّة؛ أنّها بمثابة (**حقل لساني**)[9] حديث يتدرّج باستمرار نحو النّضوج والاكتمال.

ومن التعريفات التي حاولت صياغة مفهوم للسانيات المدوّنات انطلاقاً من المصطلح نفسه، بتفكيك تركيبة شِقّيه لتقريب الدلالات والإحالات المفهومية المنوطة به، التعريف الذي ألفيناه عند الباحث (**عبد المحسن الثبيتي**) الذي عمد إلى التّفصيل في بعض محمولات هذا الحقل انطلاقاً من

حَدّيه (لسانيات) و(المدوّنات) حيث يقول:«للسانيات المدوّنات شِقّان، يعتمد ثانيهما على أولهما، ولا قيمة لأولهما بلا ثانيهما. فالشقّ الأول إحصائي، أما الثاني فتحليل كيفي لبيانات الشق الأول ضمن إطار نظري معين يجعل من هذه البيانات مؤثّرة ضمن مجالها، وقد يتعدّاه إلى غيرها»[10] وعلى نفس النّسق -تقريباً- نهج الباحث (**رضا الكشو**) في تبرير استعماله مصطلح (لسانيات المدوّنات) استناداً على الدلالة المضمرة لشقيّ المصطلح؛ الأول (لسانيات) معلّلاً ذلك بأنّ «مصطلح لسانيات يضمر المعالجة الحاسوبية للمدوّنة وكذلك الدراسة الوصفية لأنظمة الكم الهائل من النصوص المخزنة، ثم إنّ النصوص المجمّعة في المدونة يحكمها قصدٌ بحثي يسعى إلى تحليل اللغة الفعلية...، وإذا ما انتفى هذا الهدف صارت المدوّنة أرشيفاً لكم هائل من النّصوص لا غير»[11] أمّا عن دلالة ورود شقّه الثاني بصيغة الجمع (المدوّنات) Copora بدلاً من لسانيات المدوّنة وذلك «لتنوّعها فالمدوّنة تكون عامّـة أو خاصـة أو طلابيـة، ثمّ إنّ منهج جمع النّصوص يزيدها تفريعـاً»[12]

ويربط الباحث (**محمود إسماعيل صالح**) في سياق إشارته إلى الخلفية التي عزّزت مكانة لسانيات المدوّنات في البحث اللّغوي عموماً إلى عامل التِّقانة والحاسوب مؤكداً ذلك بقوله: «غير أنّ أحدث وأهم مجال لَعبه الحاسوب في خدمة البحث اللّغوي هو مجال المدوّنات اللّغوية Corpora»[13] وهو ما يتّفق في جزئيّة كبيرة منه مع الباحث (**عبد الله الفيفي**) إلّا أنّ هذا الأخير يربط تقدّم لسانيات المدوّنات ويقصره على التّطور والطّفرة الحاصلة في النّظم والتطبيقات المُحوسبة على وجه التحديد، حين يصرّح بأنّ لسانيات المدوّنات «لم تصبح طريقة فَعّالة في دراسة اللّغة أو تحليلها إلّا بعد بضعة عقود عندما تطورت التّقنيات الحاسوبية وسهّلت للباحثين الاستفادة من هذه المدوّنات»[14] وهو الرأي الذي نُشاطره ونتّفق معه إلى حدّ بعيد. غير أنّ هذا الكلام يحيلنا في المقابل إلى التّنويه والتّذكير بأنّ فكرة المدوّنات اللّغوية ليست جديدة العهد، أو أنهّا من بدائع التكنولوجيا ومُخرجاتها. فالمدوّنة قد وُجدت منذ آلاف السّنين، كما أنّ اعتماد البحث اللّغوي عليها كان قائماً منذ عقود خلت، إلّا أنّ التكنولوجيات الحديثة بأدواتها وتطبيقاتها قد أعادت بعثها وسهّلت الوصول إليها، للقيام بمهام كثيرة في البحث اللّغوي، كما قُوبلت باهتمام متزايد من المجتمع البحثي على مواردها ومصادرها البيانية والمعيارية الهامّة والغنّية.

2- تعريف المدوّنة اللّغوية Corpus (المعنى اللغوي والاصطلاحي):

تعترضنا في سياق الحديث عن المدوّنات اللّغوية مجموعة من التّنوعات المصطلحية، هذه التّنوعات وإن كانت مرهونة باصطلاح واستعمال كل باحث إلّا أنهّا تُفضي إلى المفهوم ذاته، إذ نجد من يطلق عليها: (**المدوّنة اللّغوية**) أو(**المدوّنة النّصية**)، وهناك من يصطلح عليها بـ: (**الذّخيرة اللّغوية**)[15] وأحياناً (**الذّخيرة النصية**)، وينعتها آخر بـ: (**المكنز**)...الخ. إلّا أنّ أشهرها استعمالاً وتداولاً بين أهل الاختصاص هو (**المدوّنة اللّغوية**).

- **المدوّنة لغة**: تعني المدوّنة في اللّاتينية "Bady" أي الجسد أو المتن، دلالة عن الكيان أو الجسد اللغوي (Bady of text). أمّا **اصطلاحاً**: فيعرّفها الباحث (**إسماعيل صالح**) بقوله: «لعلّ أبسط تعريف للمدوّنة اللّغوية هو: مجموعة من النّصوص اللّغوية الشّفوية أو المكتوبة الموثّقة (من حيث المصدر والتّاريخ والنّوع كحد أدنى)»[16] ويستخدم مصطلح المدوّنات اللّغوية دلالة على «أي رصيد ضخم من النّصوص، المكتوبة أو المنطوقة أو كلتيهما، التي يتمّ تجميعها بطريقة عشوائية أو منظّمة من مصادر النّصوص المختلفة»[17] إلّا أنّ الباحث يحيل معنى المدوّنة المُحوسبة ويقصره على (**لسانيات المدوّنات**) فينبّه إلى «أنّ «المدوّنة» في سياق لسانيات المدوّنات اللّغوية يُقصد بها المدوّنة اللّغوية المُحوسبة، أي المخزّنة رقمياً في الحاسوب، لذلك نجد أنّ البعض يتحدثون عن لسانيات المدوّنات باسم لسانيات المدونات الإلكترونية Electronic corpus linguistics»[18]

وتعني المدوّنة الحاسوبية أيضاً «مجموع معطيات لغوية مكتوبة أو سمعية-بصرية تُستقى من خطابات يُنتجها متكلمون حقيقيون في تبادلات اجتماعية، ووقع اختيارها وتنظيمها حسب معايير لسانية وخارج لسانية لتكون عيّنة من اللّغة»[19] إذن فالمدوّنات بهذا المعنى وبعد اختزانها رقمياً في الحاسوب تصبح «تحوي نصوصاً تعكس الاستعمال الحقيقي أو الواقعي authentic للّغة في شكلٍ مقروء آلياً machine-readable تؤخذ عيّنة ممثلة لمجال معين، أو لأوعية معلومات بعينها، كالكتب، أو الدوريات العلمية، أو الصحف، أو المراجع...الخ. وقد يلحق بهذه النّصوص ترميزٌ marking-up بإضافة حقول ميتاداتا[20]، أو تحشيةٌ annotation أو وسمٌ tagging»[21]

وفي تعريف آخر لـ (**عبد الله الفيفي**) يقول عن المدوّنة اللّغوية إنّها:«مجموعة حاسوبية من البيانات النّصية الواقعية التي جُمعت وفقاً لمعايير تصميم محدّدة، بغرض تحليل اللّغة أو جزء منها ودراستها، مع وَسم هذه النّصوص بطريقة معيارية متجانسة، وتوثيق أصلها ومصدر الحصول عليها»[22] وبهذا فإنّ الباحث قد ضَمَّن هذا التعريف مجموعةً من المحدِّدات التّوصيفية للمدوّنة؛ فبالإضافة إلى كونها بيانات نصيّة حاسوبية يُشترط فيها الواقعية؛ وهي وُرودها في سياق لغوي طبيعي غير مصطنع أو مرتّبٍ

له سلفاً. وأن تخضع في تصميمها إلى معايير محدّدة ومتفق عليها هي بمثابة عناصر أساسية في بناء المدوّنات على اختلاف أنواعها. كما أنّ عدّها مجموعة حاسوبية مفاده حسب الباحث؛ أن تكون آلية حفظها وتخزينها في الحاسب بصيغة تسمح بقراءتها آلياً، وبالتالي أخرج عن المدوّنة المُحوسبة كلّ أشكال المستندات النّصية الممسوحة ضوئياً في صيغة صور، لأنّ النّصوص بهذه الصيغة لا يمكن اعتبارها أو إدراجها ضمن المدوّنة اللّغوية المُحوسبة.

3- أنواع المدوّنات اللّغوية المُحوسبة:

تنقسم المدوّنات اللّغوية الحاسوبية بدلالة النّوع، والغرض، والعدد، وبحسب طريقة البناء والتّصميم إلى أنواع عدّة. فهناك من يقسّمها تبعاً لطريقة معالجة نصوصها إلى: مدونة لغوية ذات تحشية (Annotated corpus)، أو مدونة لغوية مُرمّزة (Marked-up corpus)، ومدوّنة لغوية خام (Raw corpus). وهناك من يقسّمها تبعاً للغات نصوصها إلى: مدوّنات لغوية أحادية، ومدوّنات لغوية مقارنة، ومدوّنات لغوية متوازية. غير أنّ هناك تقسيماً آخر مبني على الغرض من استخدام المدوّنة حيث يفرّعها إلى نوعين رئيسيين: **مدوّنات لغوية بحثية** Research Corpora، و**مدوّنات لغوية اختبارية** Test corpora[23] وهذا القسم الأخير هو الذي يعنينا منها جميعاً تساوقاً مع منحى هذه الدراسة وتحديداً ذلك النّوع المخصّص والموجّه لإجراءات الاختبار والتقييم. مع الإشارة إلى أنّ تقسيم المدوّنات وفق هذه الاعتبارات يبقى غير موحّد ويختلف من باحث لآخر.

1-3 مدوّنات لغوية بحثية Research Corpora:

وهي «عبارةٌ عن رصيد من النّصوص الأصلية التي تستخدم في إجراء تجارب بحثية من أجل تطوير المعرفة؛ حيث تستخدم كونها قاعدة للتّحليل الفكري بصفتها مستودعاً للغة الطبيعية»[24] حيث يتفرّع هذا النوع إلى أربعة أصناف أخرى هي:- مدونة لغوية عامة – مدونة لغوية متخصصة – مدونة لغوية تاريخية أو تعاقبية أو راصدة – مدونة لغوية تعليمية.

2-3 مدوّنات لغوية اختباريّة Test corpora:

وهذا النّوع هو«عبارةٌ عن رصيد من النّصوص الأصلية أو المُخَلَّقة Invented التي تُستخدم في اختبار، أو تجريب، أو تقييم أو تقييس الأداء. وتسمى أيضاً مجموعات الاسترجاع التجريبية، أو مجموعات الاختبار test suits/Collection»[25]

إجمالاً وإضافة إلى النّوعين السابقين هناك تقسيم آخر للمدوّنات يتحدّد من خلاله نوع المدوّنة على حسب المعايير المصنّفة لها، وأشهر هذه المعايير التّصنيفية نجملها اختصاراً في: - **معيار عدد اللغات -معيار الزمن -معيار الحجم size - معيار تهيئة وترميز المدوّنة - معيار المستوى اللّغوي -معيار طبيعة النّصوص - معيار هيئة النّصوص**... وغيرها، إلّا أنّه يتعيّن على مصمّم المدوّنة مراعاة هذه المعايير والالتزام بها عند بناء مدوّنته، بما يتوافق مع المنهجية وكذا الأهداف والأغراض البحثية الموَجَّهة إليها المدوّنــة.

04- أشهر المدوّنات اللّغوية الحاسوبية العربية:

01- المدوّنة اللّغوية العربية الدولية International Corpus Of Arabic (ICA)

02- المدوّنة العربية للقرآن الكريم The Quranic Arabic Corpus

03- المدوّنة اللّغوية العربية لمدينة الملك عبد العزيز للعلوم والتقنية King Abdulaziz Arabic Corpus (KACST) City For Science and Technology.

04- المدوّنة اللّغوية لمتعلمي اللغة العربية Arabic Learner Corpus

05- المدوّنة اللّغوية التاريخية للجامعة الأردنية Historical Arabic Corpus

06- مدوّنة عربي كوربص Arabic Corpus

07- مدوّنة **نـملار NEMLAR** أو (المشروع الأورومتوسطي).

05- المدوّنة اللّغويـة في خدمـة المعالجة الآليـة للّغات الطبيعيـة:

يرى الباحث (**هشام موسى المالكي**) أنّ المدوّنات أو الذخائر اللّغوية هي «الرّكيزة الأساسية لفروع علم اللغة التطبيقي بمعناه الحديث، الذي يرصد الأداء اللغوي الواقعي؛ حيث يرسي قواعد جمع المواد اللّغوية الطبيعية ومنهجيات تهيئتهـا، وترميزهـا لخدمة أغراض بحثية مختلفة وهو من العلوم البينيّة التي تنطلق من علم اللّغة التطبيقي، وتتداخل مع نظريّات علم الإحصاء كمنهجية لرصد الظواهر اللّغوية، ومع تطبيقات علم اللّغة الحاسوبي كأدوات للمعالجة»[26] وباعتبار المدوّنات موارد لغوية في صورة حاسوبية فهي من وجهة نظر الباحث (**محسن رشوان**) بمثابة الرّكيزة الأساسية لبناء وتطوير أدوات المعالجة الآلية للّغات الطبيعية، وتمثّل هذه الموارد حسب وصف الباحث ضابطاً معيارياً يمكن الاسترشاد به في وصف واقع اللّغة بمستوياتها المتعدّدة، كما يَنظر إليها أيضاً على أنّها

وسيلة لتقويم أدوات المعالجة الآلية للغات[27] وهذا التّصور للمدوّنات أي (وسيلة لتقويم أدوات المعالجة الآلية للغات) هو ما يتقاطع مع فكـرة ومضمون هذه الدراسة، ويسند الغرض البحثي القائمة عليـه، من خلال اختبـار كفاءة وأداء بعض نُظم وأدوات المعالجة الآلية للنّصوص. والتي تأتي في مقدّمتها المحلّلات الصّـرفية، من خلال توظيف المدوّنات المُحوسبة/الحاسوبية لاختبار وتقييم عمل وجـاهزيّـة هذه المحلّلات على مستويات عـدّة.

ولكي تخدم المدوّنات اللّغوية مجال المعالجة الآلية لابدّ أن تتحقّق فيها مجموعة من الشّروط والضّوابط أهمهـا (**الوَسْم** Tagging/ **وَسْم المدوّنات**).

06- وَسْـم المدوّنـات اللّغويـة:

إنّ المدوّنات اللّغوية في ذاتها قد لا تقدّم شيئاً أكثر من كونها وعاءً يَختزِن نُصوص اللّغة، بوصفها بياناتٍ خام. ثم تُشغّل على هذه البيانات لاحقاً بعض البرمجيات « بمقدور هذه البرمجيات إعادة تنظيم وترتيب هذه البيانات. ومن ثمّ يتم عمل مجموعةٍ من التحليلات الإحصائية النّظمية والدّلالية عليها؛ بهدف فحص كل مصطلحٍ أو كلمةٍ مفتاحيةٍ من حيث سِماتها أو تركيبتها المعرفية واللّغوية، وما يرتبط بها من كلماتٍ تسبقها أو تلحقها»[28] وهو ما يحيلنا على مفهوم (**الوَسْـم** Tagging).

ينقل(**عبد الله الفيفي**) عن (**ليتش**Leech) تعريفه لعملية وسم المدوّنات بأنّها: إضافة معلومات لغوية تفسيرية إلى مجموعة إلكترونية من البيانات اللّغوية المكتوبة أو المنطوقة، كما يُعرّف الوُسوم - باعتبارها المنتج النهائي لعملية الوسم- بأنها الرموز المرتبطة بالتمثيل الإلكتروني لمواد اللّغة[29].
ويزيد (**الفيفي**) على تعريف (ليتش Leech) تعريفاً خاصاً به للوسم Tagging بأنّه «إضافة علامات - نصّية وغير نصّية- إلى نُصوص المدوّنة اللّغوية؛ لإثرائها بمعلومات إضافية تزيد من فائدتها، أو تسهّل البحث فيها وتحليل نُصوصها»[30]

فالوَسم إجمالاً هو إضافة حُزمة من المعلومات اللّغوية (الصّرفية أو النّحوية أو المعجمية) وغير اللّغوية (أرقام ورموز) إلى المدوّنة، تتخذ أحياناً صورة مفردات أو جملاً أو شُروحات لغويّة، وتكون أيضاً على هيئة أرقام واختصارات، أو رموز دلالية يفهمها واضع الوَسم والحاسوب، لأنّها تُلقَّن إلى الحاسوب في هيئة صيغٍ رياضية واختصارات رمزية مُعرّفة مسبقاً في شكل جداول و رموز تعريفية.

6-1- أنــواع الوُســوم:

إنّ من بين أنواع المدوّنات التي ذكرنا سابقاً، ذلك النّوع أو الصّنف الذي يُقسَّم بحسب طريقة معالجة نصوص المدوّنة نفسها. والقصد بطريقة معالجة نصوصها هو **وَسْم** هذه النّصوص أو **تحشيتها**. ويترتّب عن هذا التصنيف مجموعة من المدوّنات ذكرنا منها: المدونة الخالية من الوَسم أو الخام Raw corpus، وهي التسمية الأشهر وقد تسمى بالمدونة الصّافية أو الخالصة Pure corpusوبعد إدخال الوسم عليها فإنها تسمّى بالمدوّنة الموسُومة Tagged corpus أو المرمّزةMarked up corpus أو ذات تحشية Annotated corpus[31]

مع التنويه مبدئياً إلى ما يتعلق بمصطلح التّحشية Annotation القريب من الوسم Tagging والذي غالباً ما يُقرن به أو يُتداول بالدلالة نفسها، غير أنّ في عرف الحاسوبيين فرقٌ صريح بينهما، ومعيار التّفريق بين الآليتين هو(وضوح الدّلالة). **فالتحشية**: تعرّف«بأنّها مجموعة من التّحليلات والمعالجات اللّغوية التي تتمّ على النّصوص بهدف إضفاء توصيف دقيق عليها، ومن الممكن أن تتمّ تحشية المدوّنات اللّغوية في عدّة مستويات وبأشكال مختلفة»[32] فعلى المستوى الصّرفي مثلاً «من الممكن أن تتمّ التّحشية للسّوابق prefixes واللّواحق suffixes، والجذور roots، والجذوع stems (تحشية صرفية)»[33] وبالتّالي تصبح التّحشية إضافة معلومات لغوية مفهومة وواضحة الدلالة بشكل مباشر، فهي إذن لا تحتاج إلى جدول يفسر دلالتها مثلما هو الحال بالنسبة إلى الوسم.

نستنتج على ضوء ما سبق أنّ إثراء المدوّنات بهذا الكمّ من المعطيات والمعلومات الإضافية سيزيد من قيمتها وفائدتها، كما يزيد من مرونة عمليات البحث، التحليل والإحصاء عبرها، للباحث أو الحاسوب على السواء، وأنّ المدوّنات المَوسّمة و المحشّاة بدقّة والمشكّلة تشكيلاً تاماً و وافياً، تُعدّ من منطلق هذه الدراسة بمثابة (**آلية معيارية مُتجانسة**) هامة، وأداة (**تقييم موضوعيّة**) يَلجأ إليها - بصفة دائمة- الباحثون في مجال اللسانيات الحاسوبية، والمشتغلون في حقل المعلوماتية من مطوّري النّظم والبرامج ذات الأساس اللّغوي كالمحلّلات وعلى رأسها المحلّلات الصّرفية، لاختبار وتقييم عملها وأدائها، نظراً لعناية هذه الأنظمة -المحلّلات- بعمليّات التّحليل والتّشكيل والوَسم. والتي هي من أهم المعطيات اللّغوية التي بُنيت المحلّلات لغرض تقديمها وتوضيحها. ناهيك عن مجمل المعلومات الصّرفية وكذا الصّرف- نحوية التي تعطيها عن بنية الكلمة المحلّلة.

وللتّدليل على دور المدوّنة المُحوسبة في معالجة وتحليل مستويات اللّغة، المستوى الصّرفي نموذجاً (مثلما يعالجه هذا المقال) فإنّها تضمن لدراسة اللّغة في جانبها الصّرفي دراسةً وصفية أكثر واقعية، لمعرفة خصائص المستوى المورفولوجي في اللّغة المستهدفة، باعتبارها تُهيئ للحاسوبي الإطار النّظري والنّموذج الذي سيتبنّاه في بناء وتطوير محلّله الصّرفي، لافتين النظر- في الوقت ذاته- إلى أنّ اختيار

المدوّنة لا بدّ وأن يتوافق مع أهداف المعالجة الآلية وكذا مستوى اللغة المستهدفة، لأنّ تحديد مستوى لغة المدوّنة هو شرطٌ أساسي في التّعامل مع المحلّل؛ فإذا كان المحلّل قد بُني لاستهداف لغة عربية تراثية أو كلاسيكية فلا بدّ من اختيار مدوّنة لغوية كلاسيكية، أمّا إذا كان المحلّل يستهدف لغة عربية فصحى حديثة(MSA) فيتعيّن اختيار مدوّنة مصمّمة من نصوص هذه اللّغة، والأمر نفسه مُنطبق على إجراءات الاختبار لهذه المحلّلات لأنّه لابدّ من التّوافق بين مستوى لغة المدونة (أداة الاختبار) ومعجم المحلّل الصّرفي وقواعد معطياته على حسب ما بيّنا.

وبالرجوع إلى ملخّص هذه الدراسة وتحديداً عند إشارتنا إلى قضية "**المنفعة التبادلية**" التي تجمع المدوّنات المُحوسبة بوصفها موارد بيانية بالمحلّلات كونها نُظم وأدوات حاسوبية. فإنّه من البديهي أن تكفُل المدوّنات المُحوسبة،المشكّلة والموسّمة بدقّة إمكانية اختبار وتقييم المحلّلات الصّرفية بناء على هذه المعطيات والخيارات، وبالمقابل فقد تولّت بعض المحلّلات الصّرفية -منذ ظهورها- مهمّة إضافة مجمل هذه المعلومات اللّغوية إلى المدوّنات التي كانت توسّم قبلاً يدوياً، ثمّ تحولت بها إلى ما يعرف بـ: وسم المدوّنات أو الذخائر آلياً. حيث«تستعمل في العادة المحلّلات الصّرفية التي يمكنها تحليل المفردات وإضافة الوُسوم الصّرفية المناسبة لها»[34] ذلك أنّ تطعيم المدوّنات بكلّ هذه المعطيات والمعلومات يدوياً يستنزف وقتاً وجهداً بشرياً كبيرين، حتّى وإن صغر حجم المدوّنة. أمّا بالنسبة للّغة العربية وبعد بروز مجموعة من المحلّلات الصّرفية في السنوات الأخيرة، فإنّ بعضها قد تولّى القيام بهذه المهام. وأشهر المحلّلات الصرفية المطوّرة للعربية نذكـــر:

01- محلّل **الخليل الصرفي**[35] Alkhalil Morpho Sys التابع لإدارة العلوم والبحث العلمي في المنظمة العربية للتربية والثقافة والعلوم (ALECSO)، بالتعاون مع جامعة محمد الأوّل بوجدة المغربية.

02- محلّل **باك والتر** Buck walter التابع لجامعة بنسلفانيا الأمريكية.

03- محلّل **مادااميرا** Madamira التابع لجامعة كولومبيا الأمريكية.

04- محلّل **سلمى** SALMA التابع لجامعة ليدز بالمملكة المتحدة.

وعلى ذكر المحلّلات الصّرفية نأتي إلى بيان هذه الأخيرة من خلال التّعريف بواحد من أشهر هذه المحلّلات الموجهة للّغة العربية وهو محلّل أو برنامج الخليل الصّرفي.

7- التعريف ببرنامج الخليل الصّرفي (Alkhalil Morpho Sys):

برنامج الخليل الصّرفي هو محلّل حاسوبي مفتوح المصدرOpen source[36] موجّه للّغة العربية، تمّ تطويره والعمل عليه بمخبر البحث في الإعلاميات بجامعة محمد الأول بوجدة (المغرب)، تحت مظلّة المنظمة العربية للتربية والثقافة والعلوم (ألكسو)، وبالتّعاون مع مدينة الملك عبد العزيز للعلوم والتقنية بالمملكة العربية السعودية. يهدف هذا المحلل إلى تحليل الكلمات والنّصوص العربية مورفولوجياً. ويشيد مطورو البرنامج بالخوارزمية التي يشتغل وفقها المحلّل لاعتماديّتها بالدرجة الأولى على قوانين النحو والصرف في استخلاص وتعيين مجمل الصّفات والخصائص الصّرفية للكلمة العربية، بدءً بتقطيعها إلى «لبناتها الصّرفية الأساسية... قصد تحديد مجموعة من المعلومات الصّرفية المحتملة للكلمة»[37] وذلك لتحديد ما يلـــي:

- **نوع الكلمة**: اسم أو فعل أو حرف - **جذع الكلمة (ساقها)** - **جذرها** (في حالة الأسماء والأفعال) - **تحديد الزوائد التي تلحق بالكلمة** (السوابق واللواحق) - **وزن الكلمة** مشكولاً (في حالة الأسماء والأفعال) - **حالات التّشكيل** الممكنة للكلمة - **حالتها الإعرابية** (في حالة الأسماء والأفعال).

كما يعتمد المحلل أيضاً على مجموعة من قواعد المعطيات الضرورية التالية:[38]

- ✓ قاعدة معطيات السوابق واللواحق.
- ✓ قاعدة معطيات بالأدوات.
- ✓ قاعدة معطيات بأسماء الأعلام.
- ✓ قاعدة معطيات تضم الأوزان غير المشكولة.
- ✓ قاعدة معطيات تضم الأوزان المشكولة.
- ✓ قاعدة معطيات تضم الجذور العربية مرفقة بأوزان مشتقاتها.

8- اختبار برنامج الخليل الصّرفي AlKhalil Morpho Sys 1.0:

لاختبار وتقييم عمل وأداء محلّل الخليل AlKhalil Morpho Sys تم رصد مدوّنة **نملار (NEMLAR)** أداةً للاختبار، وتوظيفها لتغطية كافّة إجراءات التّقييم في منظومة المحلّل، وقد تركّزت إجراءات الاختبار أساساً حول ثلاثة مستويات وظيفة هامة في عمل المحلّل، وهذه المستويات هي بمثابة آليات رئيسة في عمليات التّحليل الصرفي الآلي عموماً، وهي ضرورية للمحلّل للوصول به إلى درجة أداء عالية من حيث: -**التحليل** Analyse أي (تعيين الخصائص الصّرفية للكلمة المدخلة) - تعيين **جذر** الكلمة Root - **تشكيل** الكلمة Diacritization. وهـذه المستويات هي:

1-**المستـوى الأول: اختبار (التّحليـل الصّـرفي):** تقييم تحليل الخليل AlKhalil Analyser 1.0

2-المستوى الثاني: اختبار (التّجذير): تقييم مستخرج جذور الخليل Alkhalil Root Extractor 1.0

3-**المستوى الثالث:اختبار (التّشكيـل)**: تقييم التشكيل الآلي للخليل AlKhalil Diacritizer 1.0

وسنحاول أن نعرض لهذه المستويات وأهم إجراءات الاختبار المطبقّة فيها بالتفصيل، على ضوء (عينات الاختبار) المُدخلة إلى المحلّل والمستقاة عشوائياً من نُصوص المدوّنة المستخدمة، وما ترتّب عن الاختبار من نتائج مُشفعة بالنّسب والأرقام في الشقّ التطبيقي لهـذه الدراسـة.

9- مجموعـة الاختبـار: مدوّنـة (نـمـلار NEMLAR)

مشروع نـملار "**NEMLAR**"[39]هو مشروع لغوي حاسوبي مموَّل من طرف الاتحاد الأوروبي، أُنشئ المشروع لرصد سيرورة اللّغة العربية على الصّعيد البحثي واللّهجي، وكلمة (**NEMLAR**) هي اختصار لـ: Network for Euro-Mediterranean Language Resources. والمشروع هو مدوّنة صغيرة نسبياً من النّصوص باللّغة العربية مشروحة من قبل شركة(RDI Egypt)نيابة عنNEMLAR Consortium التي تمتلك حقوق الملكية. تستهدف هذه المدوّنة اللّغة العربية الحديثة وتحتوي على (**نصف مليون**) 500.000 كلمة مشروحة «تم تصنيفها في **13** مجالاً مختلفاً(أخبار سياسية، نصوص إسلامية، عبارات شائعة، نصوص من نشرات الأخبار، والأدب العربي، والأخبار العامة والصحافة العلمية والصحافة الرياضية والنصوص القانونية، وشروح مواد المعجم)»[40]مقسّمة على **489** ملف(كما يظهره الجدول:01) وتغطي هذه النصوص فترة تمتد من 1990 إلى 2005. والمدوّنة هي إحدى النتائج العلمية لمشروع الشبكة الذي نُفّذ بين عامي 2003-2005، بهدف إنتاج ذخائر لغوية وأدوات تقنيّة أساسية للباحثين في اللغة العربية بطريقة منهجية معيارية، إضافة إلى دعم التعاون والشراكة بين الشركاء المؤهلين في منطقة حوض البحر المتوسط. حيث جمع المشروع **14** شريكاً من مختلف بلدان حوض البحر الأبيض المتوسط في إطار برنامج MED-Unco المدعوم من الاتحاد الأوروبي بين أطراف عربية وأوروبية عاملة في هذا المجال. وقد تلا المشروع مشروع تكميلي له بين 2008-2010. وهو مشروع (**ميدار Medar**) Mediterranean Arabic Language and Speech Technology

	المجالات أو (الميادين)	عدد الملفات	عدد الكلمات
01	مداخل المعجم	12	52000
02	الأدب العربي	24	30000
03	نصوص من الأخبار العامة	4	5500
04	عبارات شائعة	10	20000
05	معلومات عامة	159	100000
06	تجارة	18	56000
07	نصوص إسلامية	12	29000

08	نصوص قانونية	10	21000
09	مقابلات	6	8500
10	نقاشات سياسية	22	30000
11	أخبار سياسية	63	48000
12	صحافة علمية	51	50000
13	صحافة رياضية	98	50000
	المجمـــوع	489	500000

الجدول رقم (01): محتوى مدوّنة NEMLAR وتوزيعه حسب المجالات (الميادين)[41]

وقد وقع اختيار فريق تطوير المحلّل على مدوّنة NEMLAR لتوظيفها واستخدامها في كل إجراءات الاختبار والتدريب على المحلّل. وذلك لاعتبارات عائدة بالدرجة الأولى إلى مستوى لغة المدوّنة التي تستهدف اللّغة العربية الحديثة(MSA)، وهي نفسها اللّغة التي يستهدفها المحلّل، إضافة إلى ثراء المدوّنة والخيارات التي تُتيحها ما أهّلها لأن تكون أداة اختبار موضوعية ونوعية لمحلّل الخليل الصّرفي؛ بالنظر إلى حجم المدوّنة (**نصف مليون** كلمة مشروحة) ما يضمن تغطية وافية وشاملة لمستويات الاختبار والتقييم في المحلّل، كما تُوفّر NEMLAR أيضاً جملة من الخيارات اللّغوية للكلمة العربية خلافاً لبعض المدونات الأخرى. فهي تضمن مايلي:

- ✓ تشكيل الكلمة.
- ✓ ساق الكلمة (الجذع) stem.
- ✓ تعلق الزوائد بالجذع.
- ✓ الفئة النحوية للكلمة.
- ✓ مخطط الجذع المرتبط بالكلمة.

للعلم فإن مدونة NEMLAR متاحة في نمطين تام التّشكيل، وغير المشكّل.

10- مستويات اختباروتقييم محلّل الخليل الصّرفي:

10-1- المستوى الأوّل اختبار (التحليل الصّرفي): تقييم تحليل الخليل AlKhalil Analyser 1.0

يأتي اختبار محلّل الخليل من حيث التّحليل الصّرفي وتقييم آلية هذا التّحليل على رأس كلّ المستويات أو الخطوات، لأنّ باقي المستويات الأخرى أو الخطوات التّحليلية المتبقية ستترتّب بناء على هذه الخطوة المبدئية الهامة، وعلى مدى صحّة وشمولية ما يعرضه المحلّل أولاً من تحليل صرفي للكلمات أو الجمل. وعليه تأتي مرحلة اختبار المحلّل من حيث التحليل على رأس هذه المستويات .

- مـدوّنـة الاختـبـار:

لاختبار كفاءة برنامج الخليل في التّحليل الصّرفي للكلمات أو النّصوص، قام مطورو المحلّل باختباره على عيّنة من الكلمات التي تمّ أخذها عشوائياً من المدوّنة أو(**مجموعة الاختبار**) NEMLAR حيث ضَمّت المجموعة ما حجمه (**4955**كلمة)، بغية إجراء سلسلة التّحليلات الصّرفية عليها. وبإدخال عيّنة الكلمات هذه إلى المحلّل تبيّن أنّ برنامج الخليل لم يتمكّن من تحليل ما مجموعه (**297** كلمة) فقط من حجم مجموعة الاختبار.

وبالنّظر إلى العدد (**297** كلمة) غير محلّلة فإنّه يعبّر عن عدد جدّ ضئيل بالنسبة إلى عدد كلمات المجموعة. هذا ما أوعز إلى القائمين على المحلّل بتتبّع هذه الكلمات في المدوّنة (عيّنة الاختبار) وفحص طبيعتها يدوياً كلمة بكلمة، ثم القيام بتصنيفها وفرزها حسب أهميتّها ضمن مجموعة الكلمات التي عجز عن تحليلها الخليل. وقد أسفرت عملية الفحص والتّصنيف عن النتائج الآتية كما يوضحه الجدول التـالي:

فئة الكلمات	عدد الكلمات غير المحلّلة	النسب المؤوية	أمثلة عن الكلمات غير محلّلة
كلمات منسوبة، ومصادر صناعية	73	%24.6	الكيميائية- التنموية-الصحراوية- الأهمية-استمرارية
كلمات دخيلة (أجنبية)	70	%23.6	أجاكس- التلفزيون- الديمقراطية- الفيلم- تكتيكات
أسماء جامدة	64	%21.5	التلميذات- العناصر- أساتذتنا- الوسائط
أسماء مشتقة	58	%19.5	جهودكم- الكبرى- مقاضاة- الانتماء- مراسلينا
أعداد	14	%4.7	الاثنى- ستين- خمسون- ملايين- مليارين
أفعال	07	%2.3	يبرخه- تأسست- رآها
كلمات غير صحيحة	05	%1.6	خلاااااص- الجغرفيت
أسماء علم	04	%1.3	القطاونة- الدليمي
التصغير	02	%0.6	الكويكبات- المصيطبة

الجدول رقم (02): عينة من الكلمات التي عجز برنامج الخليل الصّرفي عن تحليلها.[42]

إنّ الذي يمكن ملاحظته والتّعليق عليه، على ضوء معطيات الجدول وكذا النّسب المحصّلة مع عينة الاختبار؛ أنّ النسبة الأكبر من الكلمات التي شكّلت لَبساً للمحلّل قد بلغت نسبتها 24.6% وهذه النسبة اندرجت تحت فئة (كلمات منسوبة، ومصادر صناعية) وهو ما يستدعي بالدرجة الأولى العودة إلى قاعدة معطيات المحلّل، وكذا معجمه بالتحديد (ملف اللّواحق) قصد إثرائه وتحسين تغطيته بما يتناسب مع لحم لواحق الكلمات المنسوبة والمصادر الصناعية مع الكلمات.

أمّا نسبة 23.6% فقد مثلت ثاني أكبر نسبة من الكلمات التي عجز محلّل الخليل عن تحليلها، والتي تَبيّن بعد الفحص والتدقيق أنّها كلمات أجنبية أو دخيلة. وهذه الكلمات (الأجنبية) هي محل لبس حقيق ليس لمحلّل الخليل وفقط، وإنّما لعموم المحلّلات الصّرفية العربية الأخرى. وأنّ هذا الفصل (الكلمات الأجنبية) كما يسميه مطوّر برنامج الخليل «هو فصل مفتوح تزداد شروطه يومياً مما يجعل من الصعب إدارته في التحسينات المستقبلية»[43]

وبنسبة أقل تمثّل النسب 4.7% و2.3% فئة الأعداد والأفعال، وهي نسب جدّ ضئيلة نبّهت فريق التطوير للرجوع إلى المحلّل بغية تحسين قاعدة معطياته من أجل تغطية أفضل للأعداد والأفعال.

10-2- المستوى الثاني اختبار(التّجذير): تقييم مستخرج جـذور الخليل Alkhalil Root Extractor1.0

عملية "**التجذير**" كما هو معروف هي وصول المحلّل إلى جذر الكلمة المدخلة، ويعتبر التجذير آخر مرحلة يتوقف عندها التّحليل الصّرفي، وأعمق مستوى تصل إليه المحلّلات الصّرفية الموجهة للّغة العربية، خلافاً للمحلّلات الأخرى غير العربية التي يتوقف التحليل المورفولوجي فيها عند مستوى الساق أو الجذع stem.

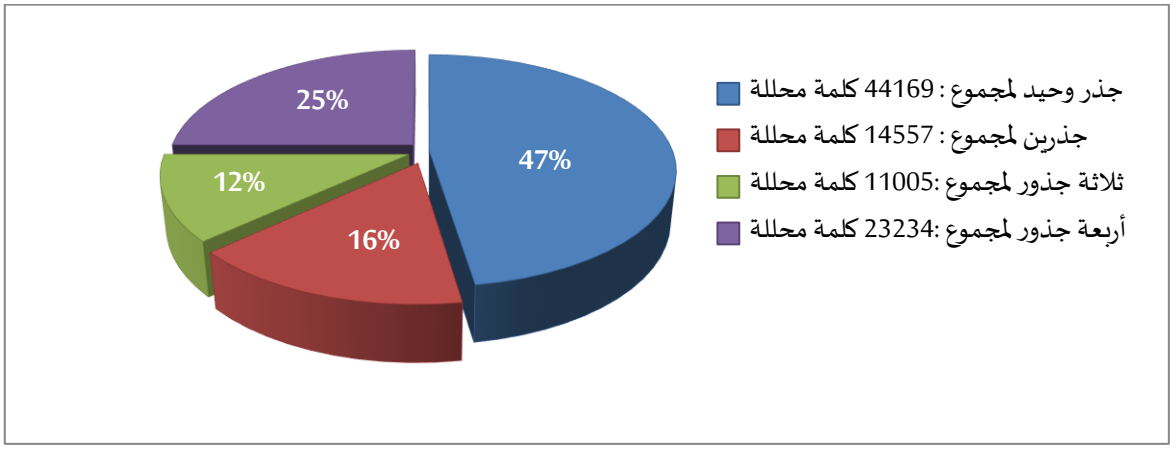
- مـدوّنـة الاختبـار:

لاختبار أداء وفاعلية محلّل الخليل في استخراج الجذور الصّحيحة والمحتملة للكلمات في مرحلتي التدريب والاختبار، تمّ استخدام المدوّنة نفسها (NEMLAR) من أجل اختبار آلية التّجذير في منظومة المحلّل. وبما أنّ المدوّنة (NEMLAR) لا تتوفّر ضمن خياراتها اللّغوية إلّا على (**جذع الكلمة** stem) -كما سبق وبيّنا- فقد عمل مطورو الخليل على إعادة شرح كلماتها على نسق (**الجذر** Root). حتى يتسنى اختبار مستخرج جذور الخليل Alkhalil Root Extractor. وقد تمّ تقسيم مجموعة الاختبار إلى قسمين:

-**القسم الأول**: ضمّ مجموعة التّدريب المكوّنة من 93% من المجموعة، و07% الباقية خُصّصت للاختبـار. بعدها تمّ تقييم مستخرج الجذور على 19% من مجموعة التدريب المكونة من (**92965** كلمة) وعلى النسبة الباقية من عناصر الاختبار أي (**38022** كلمة).

للعلم فقد كان اهتمام القائمين على محلل الخليل عند اختبار أدائه وكفاءته في استخراج الجذور منصباً حول مدى تقديم المحلّل لعدد الجذور المحتملة لكل كلمة سواءً داخل السياق أو خارجه.

وبالتالي فإنّهم قد لاحظوا أنّ المحلّل أعطى عدد جذور محتملة للكلمات تراوح ما بين الجذر الواحد إلى (04) جذور، أين نجح المحلّل في تعيين ما نسبته 47% من جذور (**44169** كلمة) محلّلة، و(04) جذور محتملة كأقصى حد لـ (**23234** كلمة) من مجموعة الاختبار. وهو مؤشر جدّ إيجابي على كفاءة التّجذيـر في المحلّل. ولتمثيل النسب المحصل عليها بيانيـاً نحصل على الرسـم التالـي:



الرسم البياني رقم (01): يوضح الجذور المحتملة المحصل عليها بعد التحليل بالنسب والأرقام[44]

بعد تعيين الجذور المحتملة تأتي المرحلة الثانية الأهم وهي اختبار كفاءة البرنامج (المحلّل) في تعيين الجذر الصّحيح للكلمة من بين تلك الجذور المحتملة المستخرجة في المرحلة الأولى. وهنا نجد أنّ المحلّل قد تمكّن من تحديد الجذر الصحيح لأكثر من (**91403**كلمة) أي ما نسبته 98.31% من كلمات مجموعة التدريب المكونة من 93%من المجموعة. كما تمكّن من تحديد 93.81% من الجذور الصحيحة لـ: 07% المتبقية من المجموعة المخصّصة للاختبـار. أي ما مجموعه (**35672** كلمة).

المرحلة الثانية من التحليل	عدد كلمات عيّنة الاختبار	عدد الجذور الصحيحة	النسبة المئوية
مجموعة التدريب	92965	91403	%98.31
مجموعة الاختبار	38022	35672	93.81%

الجدول رقم (03): يوضح كفاءة محلل الخليل في تعيين الجذور الصحيحة داخل السياق[45]

10-3- المستوى الثالث: اختبار المحـلّل من حيث (التّشكيل) تقييم مشّكل الخليل AlKhalil Diacritizer1.0

لاشكّ أنّ أكبر المشاكل والتحدّيات التي تواجه المعالجة الآلية للّغة العربية؛ هو غياب علامات التّشكيل في معظم نصوصها الحديثة، ويترتّب عن هذا الغياب معضلة كبيرة لأغلب البرمجيات المعاصرة التي تتعامل مع نُصوص العربية تحليلاً و توليداً، وبالتّالي تحتاج هذه النّظم والبرمجيات - في مقدمتها المحلّلات الصّرفية- إلى خوارزميات لاستعادة التّشكيل الصّحيح للكلمات العربية اعتمادا على

معارف ومقاربات مختلفة؛ بعضها يتكئ على المقاربة اللّغوية أي التّحليل الصّرفي والنحوي للنّصوص، البعض الآخر يعتمد على طرائق ومنهجيات إحصائية تستند بالدرجة الأولى على المدوّنات، أما بعضها الآخر فيعتمد **المقاربات الهجينة**[46] التي تجمع بين المنهجيتين السابقتين.

تَنْبني فكرة المشّكل الآلي لبرنامج الخليل AlKhalil Diacritizer 1.0 مبدئياً على الأوزان؛ أي تحديد الوزن المشكول للكلمة في الجملة بدلاً من البحث المباشر عن تشكيلها، وسبب اهتمام مطوري البرنامج بالأوزان عوض الكلمات مردّه حسب القائمين على المحلل إلى أمرين:

الأول: تلك العلاقات الدّقيقة التي تحكم تسلسل أوزان جمل اللّغة العربية، فباستقراء هذا التسلسل يمكن تحديد التّشكيل المناسب لهذه الأوزان ما يتيح للمشّكل عن طريق المطابقة معالجة أوزان جمل لم يتدرّب عليها مسبقاً.

الثاني: له علاقة بمدوّنة أو ذخيرة التدريب نفسها، إذ لا يمكنها أن تستوعب جميع أنماط جمل اللّغة العربية مهما كبر حجمها ما يشكّل عائقاً أمام المحلّل أثناء إجراءات الاختبار في التعامل مع جمل لم يسبق له التدرّب عليها. ولهذا سيشتغل البرنامج على مرحلتين:

-**الأولى**: تتمّ بناءً على نتيجة المعالجة الصّرفية للكلمات التي يحلّلها برنامج الخليل، وينتج عن هذه المعالجة حلولاً متعدّدة لكل كلمة، وبالتالي لائحة بالأوزان الممكنة لكل كلمة مشكولة ومصحوبة بسابقة ولاحقة الكلمة الخاصّين بكل وزن. أما الكلمات التي لا وزن لها فيتم التعامل معها بتعويض الوزن بنوع الكلمة في حالة الأدوات (الحروف، الضمائر، الظروف، أسماء الإشارة، أسماء الشرط، الأسماء الموصولة...) أمّا في حالة الأعلام فيعوّض الوزن بـ:(**اسم علم**).

- **المرحلة الثانية:** بعد مخرجات التحليل الصّرفي ومن أجل اختيار حل واحد على الأكثر لكل كلمة من الجملة. قام فريق التطوير«بإنشاء نموذج إحصائي لتسلسل أوزان الكلمات بالاعتماد على نماذج ماركوف الخفية[47]HMM حيث يتم اختيار الحل الأكثر رجحاناً لكل كلمة باستخدام خوارزمية Viterbi»[48] من أجل وضع التّشكيل الصحيح للكلمة داخل الجملة.

- مدوّنـة الاختبـار:

بخصوص المدوّنة الموظفة لاختبار وتقييم مشكّل الخليل AlKhalil Diacritizer فقد كانت عيّناتها أيضاً مستقاة عشوائياً من مجموعة الاختبار نفسها (NEMLAR)، باعتبار أنّ المدوّنة متاحة

بالصيغتين المشكولة وغير المشكولة، وكل عيّنة من عينات المجموعة كانت مكوّنة من 100 جملة، أي ما يربو عن 30.000 كلمة مشكولة[49] تشكيلاً تاماً.

وقد علّق الباحث (**المزروعي**) على النّتائج المحقّقة في اختبار برنامج الخليل، وكذا المقاربة المنتهجة للتشكيل الآلي Diacritizer بأنها جدّ مشجعة مقارنة بالحجم المحدود لذخيرة التدريب «فقد أدى اختبار البرنامج على نصوص غير مشكولة إلى نسبة نجاح بلغت 79.5% على مستوى الكلمات، و91.8% على مستوى الحروف»[50] وأنّ الأبحاث لا تزال جارية لتوفير حجم مدوّنة تدريب أكبر تتيح لمشكّل الخليل التعرّف على أكبر عدد ممكن من الأوزان المشكولة للجمل تطابقاً مع أنماط أخرى للجمل في العربية.

خـاتمـة:

يخلص هـذا المقـال في نهايـة مسـاره إلى جملـة من النتائج نجملها في النقـاط التاليـة:

- إنّ الحوسبة والمعالجة الآلية للغات الطبيعية في حاجة دائمة إلى مصادر وموارد لغوية رقميّة ومعيارية تُغذّي أغلب نظمها، وتُوَظّف لرفع كفاءتها أو لاختبار عملها وأدائها.

- تُعدّ المدونات الحاسوبية اليوم عُمدة المصادر اللغوية استدعاءً وتوظيفاً في نُظم الحوسبة وأدواتها والمغذّي الرئيس لها بالأدلة النّصية والبيانات المعيارية ذات التمثيل الحقيقي للواقع اللغوي.

- عطفاً على الزوايا والمناحي الوظيفية الكثيرة التي تكفلها المدونات اللغوية في مختلف المجالات البحثية التطبيقية والبينيّة، تلفت هذه الدراسة النّظر إلى زاوية وظيفية وبحثية جديدة للمدونة من خلال التركيز عليها بوصفها **أداةً اختبارية**، واستدعائها على أنها **مجموعة اختبار**لقياس دقّة وكفاءة أنظمة التّحليل الحاسوبي، أو المحلّلات الصّرفية قيد التجريب أو التطوير.

- تعدّ المدونات المُهيّئة حاسوبياً بشكل ثري ودقيق، بمثابة ركيزة أساسية لبناء وتطوير أدوات المعالجة الآلية للغات الطبيعية ومن بينها العربية، كما تُعدّ ضابطاً معيارياً يُسترشد به في وصف واقع اللغة، ووسيلة لاختبار وتقييم أدوات ونُظم المعالجة الآلية.

- رَصَدت هذه الدراسة -من بين أنواع المدونات الحاسوبية- ذلك النّوع المعروف بـ: **المدوّنات الاختبارية Test Corpora**.المُعوّل عليها في إجراءات الاختبار والتدريب أو التّقييم وتقييس الأداء للكثير من النّظم المُحوسبة، وعلى رأسها المحللات الصّرفية.

- نظراً لتزايد اهتمام المجتمع البحثي في السنوات الأخيرة بالذخائر والمدونات الحاسوبية لتوظيفها ضمن إجراءات المعالجة الآلية للّغة العربية، فقد تمّ بناء ذخائر ومدوّنات كمشاريع عربية أو بالشراكة مع دول أجنبية مهتمة باللغة العربية أشهرها: مدوّنة (**نمـلار NEMLAR**).

- لقد عمل القائمون على برنامج الخليل الصّرفي AlKhalil Morpho Sys 1.0 لاختباره، وتقييم أدائه بتوظيف مدوّنة (NEMLAR) **أداةً للاختبـار**، لاستهدافها اللغة العربية الحديثة توافقاً مع مستوى اللغة المستهدفة من المحلل، ولتوفّرها على جملة من الخيارات اللغوية التي تؤهلها لأغراض الاختبار.

- عَمَد مطورو محلل الخليل ضمن إجراءات التدريب والاختبار للمحلل، على اختباره وفق مستويات عدّة تركزت أهمها حول ثلاثة مستويات : 01- **التحليل** 02- **التجذير** 03- **التشكيل الآلي**.

- في كل مستوى من مستويات الاختبار والتقييم الثلاثة المذكورة قدّم محلل الخليل نتائج مشجعة. وهذه النتائج عكست حجم العمل **الحاسـوبي واللّغـوي** المبذول من طرف فريق التطويـر.

- يعتبر محلّل أو (برنامج الخليل الصّرفي) من المحلّلات الرائدة في مجال المعالجة الآلية للّغة العربية على المستوى المورفولوجي، وهو يلقى اهتماماً كبيراً من قِبـل المجتمع البحـثي .

الإحالات:

1- يُؤرخ لاستخدام مصطلح (**مجموعة الاختبار**) في مجال المعلوماتية إلى منتصف خمسينيات القرن الماضي، للدلالة حصرياً على قياس دقّة وفعالية النّظم المستخدمة في تجارب استرجاع المعلومات. ليُعمم بعدها المصطلح وبنفس الدلالة على كل الأدوات والنّظم قيد الاختبار والتّجريب في باقي المجالات الحاسوبية. للمزيد ينظر: الدكروري، أيمن، 2018، المدوّنات اللغوية ودورها في معالجة النّصوص العربية،ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، ص48.

2- المدوّنات اللّغوية قديمة ومعروفة تاريخياً، إلّا أنّ حداثة أساليب تناولها ومنهجيات دراستها، وكذا طرائق جمعها وتخزينها واسترجاعها، هو من أصبغ عليها طابع الجدّة والحداثة.

3- العصيمي، صالح بن فهد، وآخرون،2015، المدونات اللغوية العربية بناؤها وطرق الإفادة منها، مباحث لغوية، ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، ص21.

4- العصيمي، صالح بن فهد، وآخرون، المدونات اللغوية العربية بناؤها وطرق الإفادة منها، المرجع السابق، ص،07.

5- العصيمي، صالح بن فهد، وآخرون، المرجع نفسه، ص 186.

6- الدكروري، أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص38.

7- الكشو، رضا،جوان 2019، توظيف المدونات الحاسوبية في تأليف المواد التعليمية،المجلة العربية للتربية، المنظمة العربية للتربية والثقافة والعلوم، تونس، المجلد 38، العدد 01، ،ص13.

8- تناولت الكثير من الدراسات موضوع المدوّنات اللغوية وفق مناحي مختلفة وزوايا معينة، إلّا أنّ هذه الدراسة تعرض بجديّة لموضوع المدوّنات اللغوية الحاسوبية من زاوية جديدة كليةً؛ وذلك في سياق استثمارها كأداة اختبار وتدريب في أغلب نظم وأدوات الحوسبة. وكذا جنوح مقاربات التّحليل الصرفي إلى توظيفها وفق هذا المبدأ لقياس كفاءة المحلّلات الصرفية من جهة، والرّفع من مستوى أدائها من جهة أخرى.

9- يتوافق هذا مع ما ذهب إليه (**بول بيكر**) عندما عرّف لسانيات المدونات وعدّها بأنها **حقل لساني** يشمل تحليلاً لمجموعة كبيرة من النصوص، ينظر: بيكر، بول، 2014، مناهج المتون في اللسانيات، ضمن مناهج البحث في اللسانيات، تحرير ليا ليتوسيليتي، ترجمة، صالح العصيمي، جامعة الإمام محمد بن سعود الإسلامية معهد الملك عبد الله للترجمة والتعريب، الرياض، ص178.

10- المجيول، سلطان بن ناصر، وآخرون، 2016، لغويات المدونة الحاسوبية- تطبيقات تحليلية على العربية الطبيعية- ، ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، ص92.

11- الكشو، رضا، توظيف المدونات الحاسوبية في تأليف المواد التعليمية، المرجع السابق، ص18.

12- الكشو، رضا، المرجع نفسه،ص18.

13- العصيمي، صالح بن فهد، وآخرون، المدونات اللغوية العربية بناؤها وطرق الإفادة منها، المرجع السابق، ص19.

14- الفيفي، عبد الله بن يحيى، 2023، وسم المدونات اللغوية: المفهوم والمجالات، مجلة مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، المجلد 08، العدد01، ص278.

15- يُعزى مصطلح الذخيرة اللغوية إلى المرحوم (عبد الرحمن الحاج صالح) دلالة على المشروع الذي أطلقة والذي لا زال يعرف به.

16- العصيمي، صالح بن فهد، وآخرون، المدونات اللغوية العربية بناؤها وطرق الإفادة منها، المرجع السابق، ص19.

17- الدكروري، أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص25.

18- العصيمي، صالح بن فهد، وآخرون، المدونات اللغوية العربية بناؤها وطرق الإفادة منها، المرجع السابق، ص21.

19- الكشو، رضا، توظيف المدونات الحاسوبية في تأليف المواد التعليمية، المرجع السابق،ص 17.

20- تسمّى الميتاداتا (**Meta data**) أو البيانات الوصفية. ويطلق عليها أيضاً (ما وراء البيانات).

21- الدكروري، أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص25.

22- الفيفي،عبد الله بن يحيى،2023، وسم المدونات اللغوية: المفهوم والمجالات، المرجع السابق، ص278.

23- الفكرة التي ينبني عليها المقال تعرض لهذا النوع من المدونات تحديداً أي (**المدونة اللغوية الاختبارية**).

24- الدكروري،أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص57.

25- الدكروري، أيمن، المرجع نفسه، ص57.

26- المالكي، هشام موسى، جانفي2009، إشكاليات تهيئة الذخائر اللغوية، وبنائها حاسوبياً اللغتان العربية والصينية نموذجاً، مجلة علوم اللغة، دار غريب، مصر، المجلد 12،العدد01، ص08.

27- ينظر: رشوان، محسن، المعتز بالله، السعيد، وآخرون، 2019، الموارد اللغوية الحاسوبية، ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض، ص09.

28- الدكروري،أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص28.

29- ينظر: الفيفي، عبد الله بن يحيى، وسم المدونات اللغوية: المفهوم والمجالات، المرجع السابق، ص 280.

30- الفيفي، عبد الله بن يحيى، وسم المدونات اللغوية: المفهوم والمجالات، المرجع السابق، ص 287.

31- ينظر: الفيفي،عبد الله بن يحيى، المرجع نفسه، ص 281.

32- الدكروري، أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص 67.

33- الدكروري، أيمن، المدونات اللغوية ودورها في معالجة النصوص العربية، المرجع السابق، ص 67.

34- الفيفي، عبد الله بن يحيى، وسم المدونات اللغوية: المفهوم والمجالات، المرجع السابق، ص290.

35- محلل الخليل هو مجال التطبيق في هذه الدراسة.

36- يمكن تحميل النسخة المصدرية لبرنامج الخليل الصرفي من خلال الرابط التالي: http://sourceforge.net/projects/alkhalil

37- المزروعي، عز الدين، وآخرون، 2012، مقاربة صرفية إحصائية للتشكيل الآلي، مجلة Communications of the Arab Computer Society، المغرب، المجلد05، العدد01، ص02.

38- المزروعي، عز الدين، وآخرون، المرجع نفسه، ص02.

39- موقع المشروع على الويب: www.nemlar.org

40- حمادة، سلوى، المعالجة الآلية للغة العربية، جهود الحاضر وتحديات المستقبل، أغسطس 2008، مجلة لغة العصر، موقع المجلة على الرابط: https://archive.org/details/haggag_gmail_2009/page/n9/mode/2upk، شوهد بتاريخ: 05 ديسمبر2023، على الساعة: 10:26.

41 -Voir: Boudchiche, Mohamed, 2020, Toolkit Alkhalil pour l'analyse et la désambiguïsation morphologique des texts arabes, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc, p39.

42- Ould Abdallahi Ould Bebah, Mohamed, 2013, Contribution à l'analyse morpho-syntaxique de la langue Arabe et application à la voyellation automatique, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc, pp83-84.

43- Voir: Ould Abdallahi Ould Bebah, Mohamed, 2013, Contribution à l'analyse morpho-syntaxique de la langue Arabe et application à la voyellation automatique, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc, p80.

44- الرسم البياني من إعداد الباحث.

45- Voir: Ould Abdallahi Ould Bebah, Mohamed, 2013, Contribution à l'analyse morpho-syntaxique de la langue Arabe et application à la voyellation automatique, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc, p97.

46- المقاربة الهجينة (**مقاربة صرفية إحصائية للتّشكيل الآلي**) هي المقاربة أو المنهجية المعتمدة في برنامج الخليل الصرفي لتشكيل الكلمات والنصوص العربية آلياً.

47- **HMM** هو الرمز المختصر لنماذج ماركوف الخفية أو المخفية، وهي اختصار لعبارة : **Hidden Markov Models**

48- المزروعي، عز الدين، وآخرون، مقاربة صرفية إحصائية للتشكيل الآلي، المرجع السابق، ص 04-05.

49- ينظر: المزروعي، عز الدين، وآخرون، المرجع نفسه، ص 08.

50- المزروعي، عز الدين، وآخرون، المرجع نفسه، ص 08.

المراجع:

- بيكر، بول، 2014، مناهج المتون في اللسانيات ضمن: مناهج البحث في اللسانيات، تحرير ليا ليتوسيليتي، ترجمة، صالح العصيمي، جامعة الإمام محمد بن سعود الإسلامية معهد الملك عبد الله للترجمة والتعريب، الرياض.

- حمادة، سلوى، المعالجة الآلية للغة العربية، جهود الحاضر وتحديات المستقبل، أغسطس 2008، مجلة لغة العصر، موقع المجلة على الرابط: https://archive.org/details/haggag_gmail_2009/page/n9/mode/2upk.

- الدّكروري، أيمن، 2018، المدوّنات اللغوية ودورها في معالجة النّصوص العربية،ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

- رشوان، محسن، المعتز بالله، السعيد، وآخرون، 2019، الموارد اللغوية الحاسوبية،ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

- العصيمي، صالح بن فهد، وآخرون،2015، المدوّنات اللغوية العربية بناؤها وطرق الإفادة منها، مباحث لغوية، ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

- الفيفي، عبد الله بن يحيى، 2023، وسم المدونات اللغوية: المفهوم والمجالات، مجلة مؤتة للبحوث والدراسات، سلسلة العلوم الإنسانية والاجتماعية، المجلد 08 العدد01.

- الكشو، رضا، جوان 2019، توظيف المدونات الحاسوبية في تأليف المواد التعليمية،المجلة العربية للتربية، المنظمة العربية للتربية والثقافة والعلوم، تونس، المجلد38، العدد 01.

- المالكي، هشام موسى، جانفي2009، إشكاليات تهيئة الذخائر اللغوية، وبنائها حاسوبياً اللغتان العربية والصينية نموذجاً، مجلة علوم اللغة، دار غريب، مصر، المجلد 12،العدد01.

- المجيول، سلطان بن ناصر،وآخرون، 2016، لغويات المدونة الحاسوبية- تطبيقات تحليلية على العربية الطبيعية- ،ط1، مركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية، الرياض.

- المزروعي، عز الدين، وآخرون،2012، مقاربة صرفية إحصائية للتشكيل الآلي، مجلة Communications of the Arab Computer Society، المغرب، المجلد05، العدد01.
- Boudchiche, Mohamed, 2020, Toolkit Alkhalil pour l'analyse et la désambiguïsation morphologique des texts arabes, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc.
- Ould Abdallahi Ould Bebah, Mohamed, 2013, Contribution à l'analyse morpho-syntaxique de la langue Arabe et application à la voyellation automatique, Thèse doctorat, Université Mohamed Premier, Oujda, Maroc.