نموذج ماركوف والذخيرة اللغوية

دراسة لغوية تطبيقية

أ.م.د. تامر سعد إبراهيم خضر

أستاذ علم اللغة المساعد – كلية الآداب والعلوم الإنسانية-جامعة قناة السويس

ملخص:

زادت في الآونة الأخيرة أهمية تطبيقات معالجة اللغة الطبيعية (NLP) خاصة التي تستخدم النهج الإحصائي، ويعد نموذج ماركوف المخفي (HMM) من أكثر التطبيقات المستخدمة من نماذج التعليم الآلي؛ وهو نموذج احتمالي يتنبأ بالحالة المستقبلية بناءً على الحالة الحالية وليست الماضية؛ وبالتالي يُستفاد منه في معرفة التغير في النماذج اللغوية قبل فترة تاريخية معينة وبعدها، ويمكننا في كل عام نشر قوائم الكلمات والتعبيرات اللغوية الجديدة، وكذلك يُستخدم في تصنيف النص، والترجمة الآلية، وتحليل المشاعر.

ولكن المعالجة الإحصائية للغة لن تتم إلا ببناء ذخيرة لغوية، نأمل أن تتحد مؤسساتنا القومية العربية لإنجازها.

الكلمات المفتاحية:

الذخائر اللغوية — اللغة والاقتصاد – نموذج ماركوف المخفي – نظرية بايز – برنامج بايثون.

Abstract

Recently, the importance of Natural Language Processing (NLP) applications has increased, especially those that use the statistical approach. The Hidden Markov Model (HMM) is one of the most used applications of machine learning models; it is a probabilistic model that predicts the future state based on the current state, not the past; therefore, it is useful in knowing the change in linguistic models before and after a certain historical period. Every year, we can publish lists of new words and linguistic expressions, and it is also used in text classification, machine translation, and sentiment analysis.

However, statistical processing of language will not be done without building a linguistic repertoire, which we hope our Arab national institutions will unite to accomplish.

Keywords:

Linguistics – Linguistics – Language and Economics – Hidden Markov Model – Bayes Theorem – Python Program

مقدمة:

تمرُّ اللغة بمراحل حياة أي كائن حي، يولد، وينمو، وينضج، وتتلاشى بعض خصائصه، وتظهر أخرى. وبعض اللغات تقوى، كما يقوى كائن حي إذا توافرت الظروف لذلك، وبعضها يندثر إذا لم نأخذ بأسباب القوة والبقاء. ويكون البقاء للأقوى في معترك الحياة. واللغة في وقتنا المعاصر إن لم تقوَ بالعلم صارت تراثًا مكتوبًا دون أن تكون واقعًا معاصرًا مستعملاً.

وتحذر المنظمة الدولية للتربية والثقافة والعلوم "اليونسكو" المجتمع الدولي من خطورة انقراض عدد من اللغات الأمّ، الأمر الذي دعاها إلى تخصيص اليوم العالمي للاحتفال باللغة الأم في الحادي والعشرين من شهر فبراير من كل عام، كي تقوم المجتمعات بالحفاظ على لغاتها عنوانًا لشخصياتها، ورمزًا لذاتياتها الثقافية، في ضوء ما اعتمدته المنظمة الدولية من الأخذ بالتعدد الثقافي والتنوع اللغوي(١).

وإن كان نفرٌ من المفكرين يرون أن اللغة العربية ستبقى في المستقبل محافظة على كيانها، ولن تعرف الأفول والانقراض مادام القرآن الكريم حارسًا لها، ومحافظًا عليها – ومع احترامي لهذه الآراء – فإنها تتحدث عن لغة عبادة ارتبطت بالقرآن الكريم، وربما لا يفهمون ما يقرءون كما في البلاد الإسلامية غير الناطقة بالعربية.

وإنما أنظر إلى اللغة العربية بوصفها لغة حياة ترتبط بسنة من سنن الكون في سقوط اللغة، كما قد أشار إليها "ابن حزم" في كتاب "الإحكام"؛ عندما قال: "إن اللغة يسقط أكثرها ويبطل، بسقوط دولة أهلها ودخول غيرهم عليهم في مساكنهم، أو تنقلهم عن ديارهم، واختلاطهم بغيرهم، فإنما يقيد لغة الأمة وعلومها وأخبارها قوة دولتها، ونشاط أهلها وفراغهم، وأما من تلفت دولتهم، وغلب عليهم عدوهم، واشتغلوا بالخوف والحاجة والذل وخدمة أعدائهم، فمضمون منهم موت الخاطر، وبما كان ذلك سببًا لذهاب لغتهم، ونسيان أنسابهم وأخبارهم، وبيود علومهم، هذا موجود

(١) محمود أحمد السيد، مستقبل اللغة العربية ومتطلبات العصر القادم، مجلة مجمع اللغة العربية، دمشق، ٢٠١٢، المجلد (٨٧)، الجزء (١)، ص٧.

بالمشاهدة، ومعلوم بالعقل والضرورة"[١]. والواقع أن ما هو موجود بالمشاهدة يدل على أن ثمة استبعادًا للعربية وتهميشًا لها في العملية التعليمية في معظم جامعات الوطن العربي.

كما أنَّ إتقان الأجنبية شرط للتعيين في القطاع الخاص، وفي المؤسسات الخدمية والسياحية في أغلب بقاع الوطن العربي، ولم تشتمل شروط التعيين على إتقان اللغة العربية. وما هو موجود بالمشاهدة أيضًا أن ثمة غيابًا للعربية على ألسنة معظم ممثلي الدول العربية في المحافل الدولية، مع أن العربية معتمدة لغةً رسميةً في هذه المحافل.

وإذا ظلت الأمور تسير على هذا المنوال فإن مستقبل العربية في خطر، ولا يكفي أن تكون لغة عبادة، بل نريدها لغة الحياة في جميع جوانبها وميادينها[٢].

أضف إلى ذلك اهتمام كثير من لغات العالم بالمعالجة الآلية لمستوياتها، حتى صارت بعض اللغات لا نتوقعها صاحبة ترتيب متقدم في الانتشار كاللغة الصينية التي تحتل المرتبة الثانية انتشارًا على مستوى العالم؛ وذلك يرجع "لاهتمام الصينيين على المستوى المؤسسي والأكاديمي والفردي بحل مشكلات اللغة الصينية بشكل متكامل وتطبيقي في إطار علوم المعالجة الآلية للغات الطبيعية"[٣].

ولا يصعب على المراقب الواعي ملاحظة تأخر مستوى المعالجة الآلية للغة العربية على عدة مستويات مقارنة باللغات الأخرى، واللغة – أي لغة – من أكثر الظواهر ديناميكية في حياة البشر، فهي تتطور وتتغير بمرور الوقت، ويعد تحليل هذه التغيرات أمرًا بالغ الأهمية لفهم كيفية تأثير العوامل الاجتماعية والثقافية والسياسية على اللغة. فإن لم تستعن المؤسسات العربية الراعية للثقافة واللغة بهذه

(١) ابن حزم، أبو محمد علي بن أحمد بن سعيد، الإحكام في أصول الأحكام، تحقيق: أحمد محمد شاكر، دار الآفاق الجديدة، بيروت، ج٢، ص٩٦.

(٢) محمود أحمد السيد، مستقبل اللغة العربية ومتطلبات العصر القادم، ص٥.

(٣) جانغ جنغ، مقدمة في علم اللغة الحاسوبي والترجمة الآلية، ترجمة: هشام موسى المالكي، المركز القومي للترجمة، مصر، ٢٠٢٣، ص٧.

التطبيقات التكنولوجية للتحليل صارت عربية الحياة في وادٍ وعربية الكتب في وادٍ آخر.

إن التكامل والاتحاد بين كلٍّ من اللغة والتكنولوجيا، وتبادل المنفعة فيما بينهما، أدى إلى ظهور تطبيقات تكنولوجية كبيرة الحجم؛ مثل: استرجاع المعلومات، والتمييز الآلي للأصوات اللغوية، فضلاً عن التقنيات صغيرة الحجم؛ مثل المعاجم الإلكترونية، ومحررات النصوص، وتقنيات إرسال الرسائل الإلكترونية واستقبالها، فالاثنان لا ينفصلان(١).

وفي هذا السياق، يعد نموذج ماركوف الكامن Hidden Markov Model (HMM) أداة فعالة لتحليل التغير اللغوي، ويساعد هذا النموذج على تحديد الأنماط والتغيرات الجديدة في اللغة، ويساعد على تحسين جودة الترجمة والتكيف مع التغيرات اللغوية الحديثة، في عالم سريع التغير، لكي تظل اللغة إحدى الأدوات الأساسية لفهم التفاعل الإنساني والتطور الاجتماعي.

واستنادًا إلى ما تحتاجه لغتنا العربية إلى مزيد اهتمام بالمعالجة الآلية؛ فإننا بحاجة إلى ما تسعى إليه كل الدراسات اللغوية التطبيقية الحديثة وهو علم الذخائر اللغوية(٢)، والتطبيقات الإحصائية عليه.

ويسعى البحث إلى تدعيم بعض النقاط المعرفية والإجرائية التي تحتاج إلى مزيد من المناقشة في إطار عنوان البحث من خلال ثلاثة محاور:

أولاً: مستقبل اللغة العربية في ظل التغيرات العالمية وتنبؤ نموذج ماركوف.

ثانيًا: حاجتنا إلى بناء ذخيرة لغوية عربية ومدى الإفادة منها.

ثالثًا: توظيف الإحصاء في علم الذخائر اللغوية (ماركوف نموذجًا).

---

(١) السابق، ص١٩.

(٢) علم الذخائر اللغوية: علم متخصص قائم بذاته له منهجياته وآلياته من حيث طرق جمع المادة اللغوية وتهيئتها وترميزها وإدارتها لأغراض بحثية مختلفة.

انظر: هشام موسى المالكي، إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا – اللغتان العربية والصينية نموذجًا – مجلة أواصر، المركز القومي للترجمة، مصر، ج٢، أبريل ٢٠٠٩.

١-١ مشكلة البحث:

دراسة مشكلة تراجع انتشار اللغة العربية والبحث عن أسباب المشكلة وحلها علميًا وتقنيًا وبلغة العالم الاقتصادية والتكنولوجية بعيدًا عن التعصب الديني والقومي.

١-٢ حدود البحث:

يركز البحث على مدى انتشار بعض اللغات وتراجع انتشار اللغة العربية بناءً على تقارير من هيئات عالمية، وحاجتنا إلى بناء ذخيرة لغوية عربية تكون في متناول ذلك العالم من خلال صور الكتاب الإلكتروني وأشكاله المتعددة، ومن خلال الذخيرة اللغوية نستطيع إجراء التطبيقات الإحصائية والتقنية عليها كالكلمات الأكثر تكرارًا والكلمات الجديدة والتغيرات اللغوية وغير ذلك، ومن هذه التطبيقات نموذج ماركوف الكامن (HMM).

١-٣ الهدف من البحث:

ينطلق البحث من أهمية علم الذخائر اللغوية الذي يصب بإمكاناته في جميع التخصصات اللغوية الحديثة ومن ثم الاستفادة من التطبيقات الإحصائية عليها؛ خاصة فيما يتصل بمستقبل المعجم العربي.

١-٤ منهجية البحث:

يتبع البحث منهجية وصفية تحليلية فيما يتعلق بمدى الإفادة من علم الذخائر اللغوية والتطبيقات الإحصائية عليه.

١-٥ الدراسات السابقة:

هناك العديد من الدراسات التي تناولت استخدام نموذج ماركوف الكامن في خدمة اللغة العربية؛ ولكنها ركزت على جانب المعالجة الآلية من تخصصات وكليات تطبيقية، ويحاول هذا البحث إيجاد سبيل لمزيد من الدمج بين الجانب النظري التحليلي والوصفي والمعياري للغة من جهة، والجانب التقني الإحصائي من جهة أخرى؛ وهذه بعض الدراسات السابقة التي أفدت منها إحصائيا:

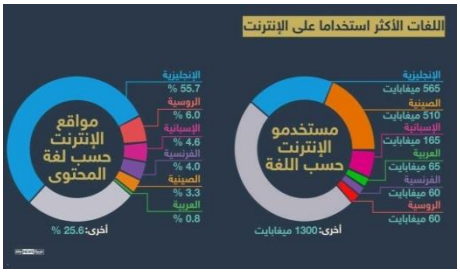
١. أزهري، نور مصطفى، "استخدام طوريات ماركوف المخفية في التعرف على الصوت والرموز"، جامعة تشرين، كلية العلوم، قسم الرياضات، ٢٠١٧.
٢. الرزوق، وائل، "خوارزميات تعرف الكلام آليًا"، رسالة ماجستير مقدمة إلى مجلس المعهد العالي للعلوم التطبيقية والتكنولوجية، قسم الاتصالات، الجامعة السورية، ٢٠١٧.
٣. الصوفي، رنا بشار، "استخدام سلاسل ماركوف المخفية في تمييز حروف العلة في اللغة الإنجليزية"، رسالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل، ٢٠٠٥.
4. Suleiman, Dima, Awajan,Arafat, Al Etaiwi, Wael, " The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey", published by Elsevier B.V., 2017.
5. A. F. Alajmi, E. M. Suad and M. H. Abdalla, "Hidden Markov Model Based Arabic Morphological Analyzer", Communication and Electronics Department, Faculty of Engineering, Helwan University, Egypt, 2011.
6. Al-Anziand F., AbuZeina D., A SURVEY OF MARKOV CHAIN MODELS IN LINGUISTICS APPLICATIONS, 10.5121/csit.2016.61305.
7. El-Hajar A, Hajar M, Zreik K (2010). A System for Evaluation of Arabic Root Extraction Methods. fifth international Conference on Internet and Web Applications and Services.
8. Duh K., Kirchhoff k. POST tagging of dialectal Arabic: a minimally supervised approach, In Proceedings of the acl workshop on computational approaches to semitic languages, pp. 55-62. Association for Computational Linguistics, 2005.
9. Al Shamsi F., Guessoum A., A Hidden Markov Model –Based POST Tagger for Arabic, Journées internationales d'Analyse statistique des Données Textuelles, 2006.
10. ElHadj, Y.O.M., I.A. AlSughayeir, A.M. Khorsi, A.M. Alansari, 2009. Morphology analysis of the Holy Quran: An indexed Quran text database (in Arabic). Proceeding of the 5th International Conference on Computer Sciences Practice in Arabic, Rabat, Morocco, May 2009.

11. Yahya O. Mohamed Elhadj(2009),” Statistical Part-of-Speech Tagger for Traditional Arabic Texts”, Journal of Computer Science 5 (11): 794-800.
12. Albared, M., Omar, N., Ab Aziz, M.J., and Nazri, M.2010. Automatic part of speech tagging for Arabic: An experiment using bigram hidden markov model. Lecture Notes Comput. Sci. Springer, 6401: 361-370. DOI: 10.1007/978-3-642-16248-0_52.
13. Albared M.,Omar N.,Juzaiddin Ab AzizMohd. Improving Arabic Part-of-Speech Tagging through Morphological Analysis, 2011.
14. Albared M., Al-Moslmi T., Omar N., Al-Shabi A., Mutaher Ba-Alwi F., PROBABILISTIC ARABIC PART OF SPEECH TAGGER WITH UNKNOWN WORDS HANDLING, Journal of Theoretical and Applied Information Technology, 31st August 2016. Vol.90. No.2

أولاً: مستقبل اللغة العربية في ظل التغيرات العالمية وتنبؤ نموذج ماركوف:

مما يدل على تأثير الاقتصاد والهيمنة السياسية على مدى انتشار اللغات على مستوى العالم؛ ما أظهرته العديد من الإنفوجراف حول اللغات الأكثر انتشارًا ، من حيث عدد السكان الناطقين بتلك اللغات كلغة أم، أو لغة ثانية، إضافة إلى تصنيف استخدامها عبر الإنترنت، وكذلك نسب استخدامها في محتوى المواقع الإلكترونية.

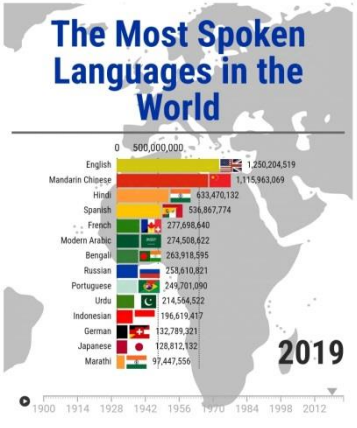
في عام ٢٠١٧ أعدت شبكة "سكاي نيوز" الإخبارية إنفوجراف، كشفت فيه أن اللغات العشر الأولى في العالم؛ هي: اللغة الإنجليزية وينطق بها ٢٥% من سكان العالم؛ أي نحو ١.٨ مليار نسمة، ولغة الماندرين (الصينية) يتحدث بها ١٨% من سكان العالم؛ أي مليار نسمة تقريبًا، واللغة الهندية يتحدث بها ١١.٥% من سكان العالم، واللغة العربية يتحدث بها ٦.٦% من سكان العالم، فيما يتحدث باللغة الإسبانية ٦.٥% من سكان العالم، واللغة الروسية يتحدث بها ٣.٩٥% من سكان العالم، وكذلك اللغة البرتغالية يتحدث بها ٣.٢٦% من سكان العالم، واللغة البنغالية ٣.١٩% من سكان العالم، واللغة الفرنسية ٣.٠٥% من سكان العالم، واللغة الألمانية ٢.٧٧% من سكان العالم.

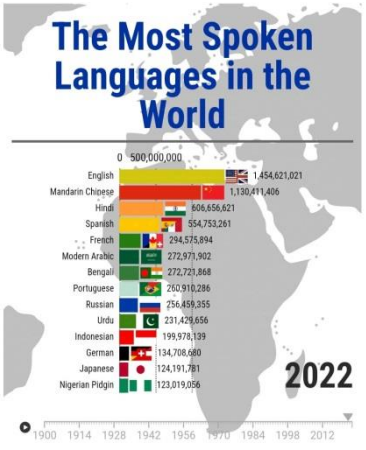


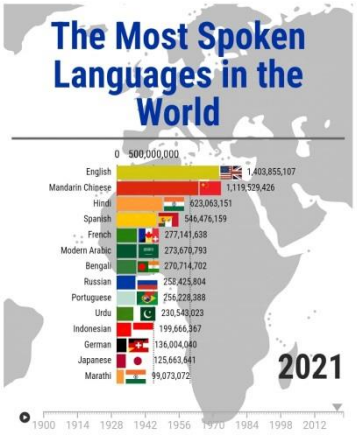
وقد حققت اللغة الصينية نموًا كبيرًا، وربما لم يكن متوقعًا بنسبة ١٢٠% بين عامي ٢٠٠٠ و٢٠١٠ أي خلال عشر سنوات فقط.

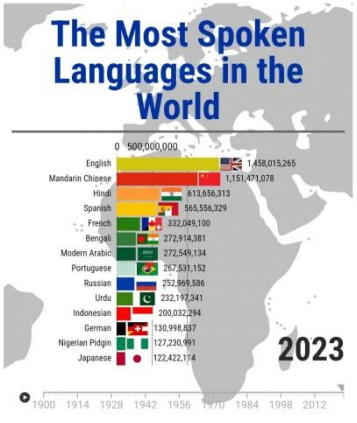
ومما يوضح مدى تأثر اللغة بالتغيرات العالمية ؛ وأنها أداة للاقتصاد والتكنولوجيا ، يلاحظ أن مستخدمي الإنترنت باللغة العربية عام ٢٠١٧ نسبتهم ٠،٨% فقط على مستوى العالم ، وهي نسبة ضئيلة جدا .

ومن خلال موقع (statisticsanddata.org) المهتم بانتشار اللغات في العالم والذي رصدها منذ عام ١٩٠٠ ميلاديًا حتى وقتنا هذا، نأخذ منه ترتيب انتشار اللغات خلال السنوات الخمس الماضية، ويتبين منه انتقال مرتبة انتشار اللغة العربية من المرتبة الرابعة ٢٠١٧ إلى المرتبة السابعة ٢٠٢٣.

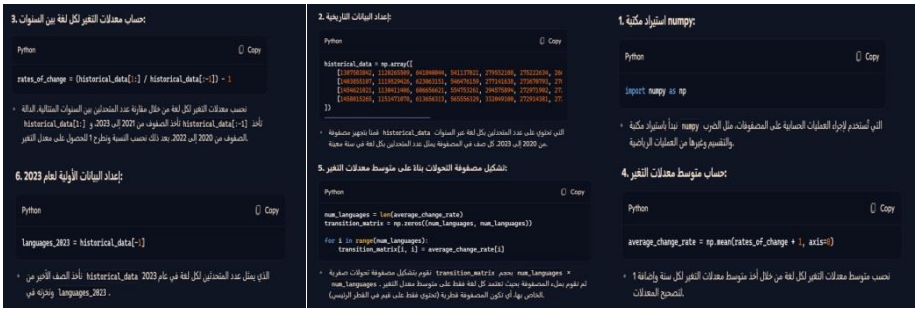




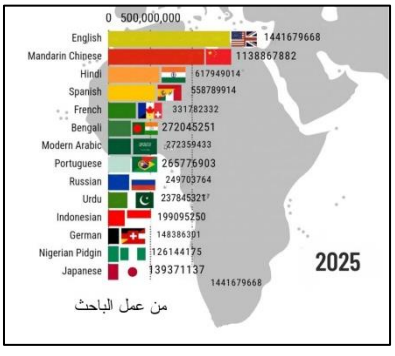


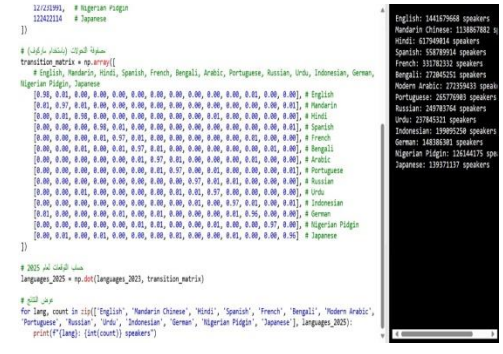


وباستخدام نموذج ماركوف من خلال برنامج بايثون يمكننا التنبؤ بمرتبة اللغة العربية نهاية ٢٠٢٤ وبداية عام ٢٠٢٥ والذي يعتمد على الحالة الحالية من خلال الخطوات التالية:



وكانت النتيجة بقاء اللغة العربية في المرتبة السابعة خلال عام ٢٠٢٥، مع منافسة اللغة البرتغالية صاحبة المرتبة الثامنة لها خلال الأعوام المقبلة .





وإن كان هناك الكثير من الآراء والمقالات التي تنكر الأرقام السابقة خلال السنوات العشر الماضية قبل ظهور هذه الحقائق، وتقارنها بعدد سكان الدول العربية؛ والأمر لا يتحمل عصبية ، وسفينة الثقافة اللغوية العربية تكاد أن ينفلت منها زمام القيادة في أوطانها، ففي تقرير لمركز دعم اتخاذ القرار والتابع لمجلس الوزراء المصري عام ٢٠٢٢؛ "أن عدد المدارس الدولية ارتفع من ١٦٨ مدرسة عام ٢٠١١ إلى ٧٨٥ عام ٢٠٢٠، مع الاحتفال بافتتاح ٢٠ مدرسة عام ٢٠٢١، وأن عدد الطلاب التربويين ٢٠٢٠-٢٠٢١ بلغ ٢٥ مليون منهم ٢.٥ بالمدارس الخاصة[١]، ولو استثنينا من ٢.٥ مليون طالب في المدارس الخاصة، نصف مليون فقط للمدارس الدولية، قد أصابتهم الغربة عن لغتهم العربية.

(١) مركز دعم اتخاذ القرار، مجلس الوزراء، مصر، ١٦ يونيو ٢٠٢٢، مقال بعنوان: "الوجاهة الاجتماعية وانتشار ثقافة المدارس الدولية"، أماني عاطف.

إن اللغة أداة حياة بكافة مظاهرها السياسية والاقتصادية والاجتماعية وغيرها، وكلما قويت تلك المظاهر قويت اللغة؛ ولذا صارت اللغة في وقتنا المعاصر أداة من أدوات التعبير عن رقي الأمم اقتصاديًا وتكنولوجيًا؛ فاللغة الصينية على سبيل المثال هي اللغة الثانية الأكثر انتشارًا على مستوى العالم – كما يتضح من البيانات السابقة – وبقيت في المرتبة الثانية منفردة به على مدار أكثر من خمسة عشر عامًا ونصف، بل وتنافس اللغة الأولى انتشارًا على مستوى العالم وهي الإنجليزية في السنوات المقبلة.

وفضلاً عن النمو الاقتصادي والتكنولوجي للصين الذي أدى إلى انتشار لغتها، فلم تعد اللغة الصينية تلك اللغة الصعبة التي تمتلك مفردات عبارة عن مجموعة هائلة من الرسوم التي يستحيل فهمها أحيانًا، وإنما أصبحت الآن تمتلك ذخائر لغوية ضخمة نجحت في معالجة قضايا شائكة أصعب ما تكون مقارنة بما تعانيه اللغات الأخرى، "فالصينيون نجحوا في توطين علم الذخائر اللغوية[١] وتوظيفه بشكل عملي، وأنتجوا من خلاله التطبيقات الحاسوبية التي تعبر عن هويتهم الذاتية في حل الإشكاليات ذات الخصوصية اللغوية، وعلى رأسها التمييز الآلي لحدود الكلمات في النصوص التحريرية المكتوبة باللغة الصينية، أو التمييز الآلي للغة التحريرية سواء المكتوبة بحروف مطبعية أو بخط اليد وأنظمة إدخال اللغة إلى الحاسب الآلي، أو إدارة الاستعلامات المتقدمة باللغة الصينية عبر الشبكة أو التواصل بين العنصر البشري والآلة، وما إلى ذلك؛ الأمر الذي مكَّن الصينيين من إضافة بصمة واضحة المعالم في الوعاء المعرفي لعلم الذخائر اللغوية"[٢].

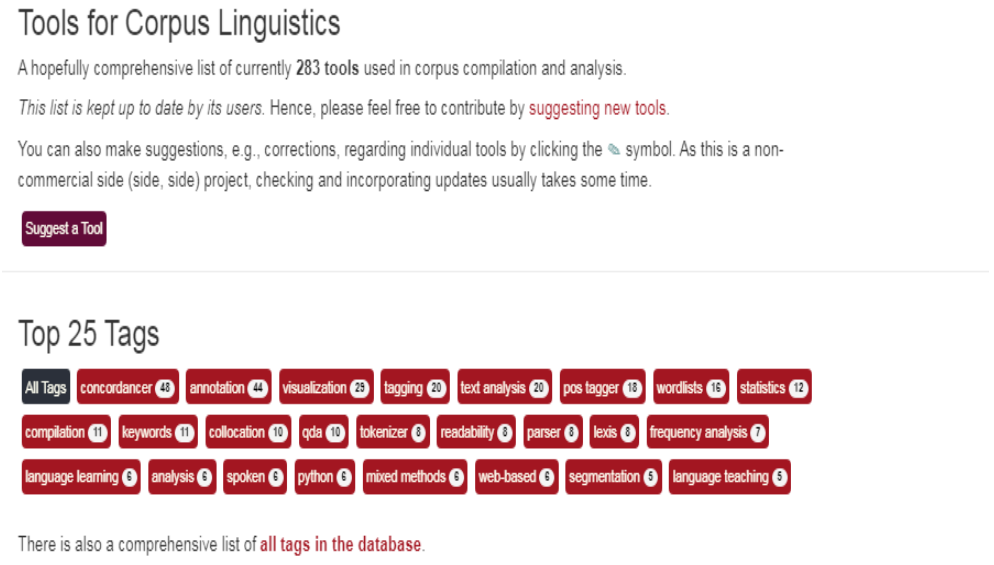
ولعل ما سبق ذكره عن اللغة الصينية يوضح لنا أسباب تراجع انتشار اللغة العربية، وارتفاع انتشار لغات أخرى ليست بلغة عبادة ولا بلغة هوية دينية إسلامية كلغتنا العربية.

(١) سبق تعريف علم الذخائر اللغوية في المقدمة.

(٢) خوانغ تشانغ نينغ، علم الذخائر اللغوية، ترجمة: هشام موسى المالكي، المركز القومي للترجمة، مصر، الطبعة الأولى، ٢٠١٦ ، ص٣٤

ولم تنتشر اللغة الإنجليزية لأنها لغة اقتصاد وهيمنة سياسية وحسب، بل لأنها لغة تعاملت مع التكنولوجيا وواكبت تلك الحركة المتجددة والمستمرة لكل جديد علمي في كل يوم، ولسنا بصدد الحديث عن تاريخ الذخائر اللغوية الإنجليزية، بما تحتويه من ملايين الكلمات ومئات التطبيقات(١).

ويكفي لكل باحث أن يكتب على محركات البحث جوجل على سبيل المثال: Tools for Corpus Linguistics لتظهر له هذه القائمة:



ثانيًا– حاجتنا إلى بناء ذخيرة لغوية عربية ومدى الإفادة منها:

إن البحث في العلوم اللغوية أو الأدبية لا يمكن أن تستقيم له مادته دون الاعتماد على ذخيرة لغوية، وأن إغفال الباحث لتقنيات تهيئة الذخائر اللغوية

(١) انظر: فصل: "تعريف بالذخائر اللغوية خارج الصين" من كتاب: علم الذخائر اللغوية" خوانخ تشانغ، ص١٠١.

الإلكترونية وبنائها من شأنه أن يقلص مهارات العمل البحثي لديه، لأن النصوص التي سيطبق عليها البحث ستكون ضئيلة الحجم والاستشهادات شديدة التواضع تأثرًا بمحدودية قدرته الذاتية على القراءة والتحليل، وفي معظم الأحيان ستغلق أمام الباحث العديد من الآفاق البحثية، فلن يستطيع على سبيل المثال دراسة مجموعة كاملة من أعمال أديب معين، أو دراسة الخصائص التركيبية للمصطلحات في نصوص تنتمي إلى عدد من المجالات أو دراسة المفردات الأكثر استخدامًا في لغة من اللغات، وغير ذلك من الأبحاث التي لا يمكن أن يستقر لها مسارها البحثي دون بناء ذخائر لغوية إلكترونية مناسبة الحجم والاعتماد بصورة أساسية على التقنيات الحاسوبية في تحليل المادة اللغوية واستخراج الشواهد التي تقود لنتائج بحثية تطبيقية بصورة دقيقة يعتد بها وتكون قادرة على خدمة المجتمع(١).

ولابد من تغيير طريقة جمع المادة اللغوية والتي تعتمد على معاجم السابقين، وهي الطريقة التي ظلت سائدة حتى العصر الحديث، دون محاولة أخذ مادة المعجم من مادة حية تم جمعها من خلال النصوص(٢).

وربما كان عذر المعجميين العرب المعاصرين في عدم اللجوء إلى الجمع الميداني صعوبة العمل من ناحية، وضخامة حجم المادة من ناحية أخرى، مما يجعل التعامل مع ملايين الكلمات والبطاقات أمرًا مستحيلاً. ولم يعد هذا العذر مقبولاً الآن بعد استخدام الحواسيب والماسحات البصرية وإمكانية التعامل اليومي مع ملايين الكلمات والاقتباسات.

وإذا كان أهم ما يميز المعجم القديم (أو الطريقة القديمة في جمع مادة المعجم) احتواءه على كثير من الاستعمالات التي لا تحيا إلا عن طريق الانتقال

(١) هشام موسى المالكي، إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا، مقال بمجلة أواصر، المركز القومي للترجمة ، القاهرة ، ٢٠٠٩، ج٢، ص٥٤.

(٢) وإن تم المزج بين الطريقتين في "المعجم العربي الأساسي"، وفي عدد من المعاجم الثنائية اللغة (عربية فرنسية، عربية ألمانية، عربية إنجليزية)، ولكن تم المزج بصورة فعالة في معاجم الأستاذ الدكتور/أحمد مختار عمر-رحمه الله -؛ مثل: معجم اللغة العربية المعاصرة، والمكنز الكبير، والمعجم الموسوعي لألفاظ القرآن الكريم وقراءاته.

من معجم إلى معجم[١]، فإن أهم ما يميز المعجم الحديث (أو الطريقة الحديثة في جمع مادة المعجم) احتواؤه على كثير من الاستعمالات التي تحيا خارج المعجم، وتتردد في النصوص الحية[٢].

إن أحد أهم استخدامات منهجية الذخائر اللغوية في الدراسات اللغوية استخراج البيانات اللغوية التجريبية الأكثر شيوعًا وتقديمها للعاملين في مجال البحث اللغوي؛ ومنها:

[١] الذخائر اللغوية وتطبيقاتها في الدراسات المتعلقة بعلم المفردات:

إن تاريخ اعتماد مؤلفي المعاجم على البيانات اللغوية الواقعية واستخدامهم لها في مؤلفاتهم المعجمية يسبق ظهور علم الذخائر اللغوية، وعلى سبيل المثال، سبق أن استخدم العالم صمويل جونسون (Samuel Johnson) الجمل الواردة في الأعمال الأدبية في تأليف معجمه، وفي القرن التاسع عشر، استخدم معجم أكسفورد للغة الإنجليزية (Oxford English Dictionary) بطاقات الاستشهاد (Citation Slips) لدراسة الاستخدامات المختلفة للكلمات وشرحها، ومازالت طريقة جمع الاستشهادات اللغوية مستمرة حتى الآن[٣]. إلا أن ظهور الذخائر اللغوية وما صاحبها من منهجيات قد غير من أسلوب استقراء مؤلفي المعاجم واللغويين للحقائق اللغوية[٤]؛ فالذخائر اللغوية في الوقت الراهن تعني أن مؤلفي المعاجم بإمكانهم الجلوس أمام إحدى شاشات الحواسيب الإلكترونية، وفي ما لا يزيد عن عدة ثوانٍ يمكنهم استخراج الأمثلة الكاملة التي تمثل الاستخدام الحقيقي للكلمة أو تعبيرة لغوية في نصوص يتعدى حجمها مليون كلمة[٥]، **وعلى سبيل المثال:**

(١) سماها بعضهم الكلمات الأشباح Ghost Words.

(٢) أحمد مختار عمر، صناعة المعجم الحديث، عالم الكتب، القاهرة، الطبعة الثانية، ٢٠٠٩، ص٧٦-٧٧ بتصرف.

(٣) وقد قمت بهذه الطريقة أثناء عملي في فريق معجم "اللغة العربية المعاصرة" برئاسة أ.د. أحمد مختار عمر.

(٤) خوانغ تشانغ نينغ، علم الذخائر اللغوية، ترجمة: هشام المالكي، ص٢٨١.

(٥) السابق، ص٢٨٢.

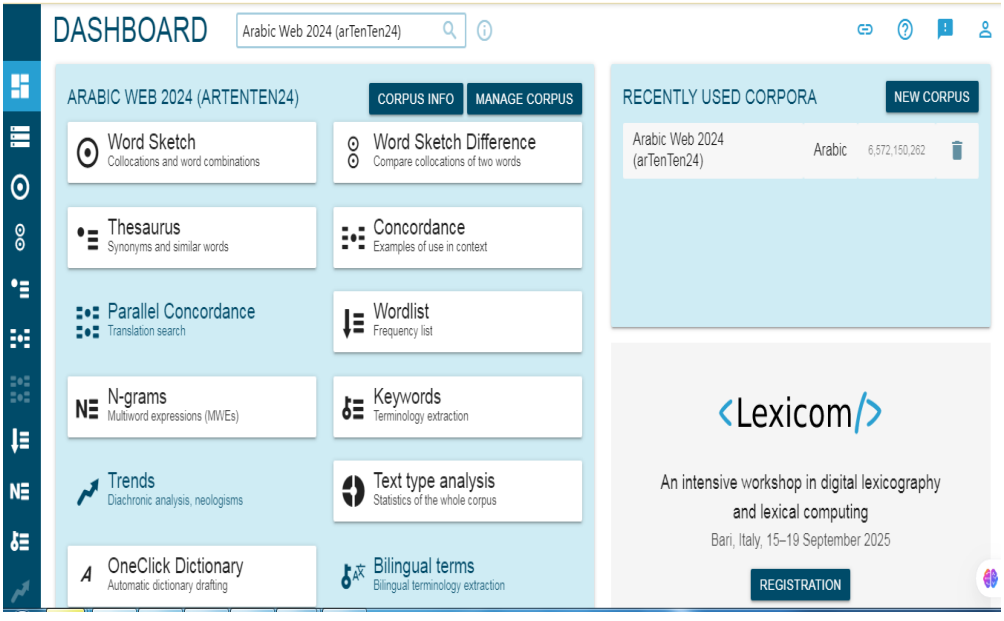
(أ) الاستعلام الإحصائي عن الكلمات داخل الذخيرة اللغوية:

عادة ما تلجأ الذخائر اللغوية المميكنة إلى أسلوب الاستعلام الإحصائي السياقي عن الكلماتconcordance لتقديم المعلومات الإحصائية المتعلقة بالسياقات التي تظهر فيها كلمة معينة داخل متن الذخيرة. وتُسجِّل موقع الكلمة موضع البحث في كل مرة ظهرت فيها داخل الذخيرة، وبناءً على ذلك يمكن تقديم المعلومات السياقية المتعلقة بتلك الكلمة. وهذه المعلومات يمكن أن تظهر مباشرة على شاشة الحاسب أو يتم حفظها في ملف معين . وهذا الملف الذي يتم حفظه يُطلق عليه اسم ملف الإحصاء السياقي للكلمات file concordance .

وقبل عمل استعلام إحصائي عن سياقات الكلمات، تكون هناك حاجة لبناء فهرس لكل كلمة من كلمات الذخيرة، يُسَجَّل في هذا الفهرس موقع هذه الكلمة داخل النص في كل مرة من مرات ورودها، بالإضافة إلى إمكانية تقديم إحصائية عن معدل ظهور هذه الكلمة داخل الذخيرة بأكملها.

كما يمكن الاستعلام عن كلمة مفتاحية داخل السياق in Word Key context والذي يطلق عليه اختصارا KWIC ، وفي ذلك الاستعلام تظهر الكلمة المُسْتَعلم عنها في منتصف كل سطر ، وقبلها وبعدها مسافة ، ويلي كل مسافة منهما سياق نصي بعدد من الكلمات يمكن التحكم في طوله ، كما يمكن تعديل طول السياق المصاحب للكلمة من جهة اليسار وجهة اليمين حسب الحاجة .

وهذه صورة واجهة شاشة لذخيرة لغوية أجنبية https://www.sketchengine.eu/ لعدد من اللغات منها العربية، توضح لنا إمكانيات البحث داخلها.



وقد قمت بالتطبيق على عبارة "ابتسامة هادئة" لأنها مستعارة من الترجمات الأجنبية وليست موجودة في المعاجم العربية القديمة مما يظهر مدى أهمية المسح الحاسوبي لكلمات اللغة العربية المعاصرة على أرض الواقع.

(ب) تعابير غربية في العربية المعاصرة من المسح الحاسوبي:

تتأثر اللغة بالزمان والمكان والظواهر الاجتماعية، فهي متطورة متجددة، وقد خضعت العربية لسنة التطور، فتنوعت أساليبها، فماتت فيها ألفاظ وجدت أخرى.

وقد جدت فيها أساليب كثيرة لم تكن إلا وليدة الترجمة، هذه الأساليب غريبة عن العربية، فهي بنت ظروف وأحوال اجتماعية لم توجد في هذا الشرق العربي، غير أن العربية وهي السمحة السهلة، اللينة، الطيعة لم تتنكر لهذه الأساليب، فقد قبلها الاستعمال وراضها حتى توهم القارئ وهو يقرأ صحيفته اليومية أن الذي يقرؤه عربي لم يعتوره الدخيل(١).

وفي اعتقادي أن معجم اللغة المعاصرة للأستاذ الدكتور أحمد مختار عمر(٢)، اعتمد على أكبر قاعدة بيانات في جمع مادته في العصر الحديث، وقد صدر في عام ٢٠٠٨م، ولكن بعد ٢٠٠٨م هناك الكثير من التعابير والأساليب التي ظهرت في حياتنا اليومية، وبالتالي لم تدون في معجم اللغة العربية المعاصرة؛ مثل:

- التعبير الإنجليزي: Calm Smile.

  ونقول: ابتسامة هادئة.
- والتعبير الإنجليزي: He represents public opinion

  ونقول: هو يمثل الرأي العام.
- والتعبير الإنجليزي: he played his last card

  ونقول: لعب ورقته الأخيرة.
- والتعبير الإنجليزي: to poison the public opinion

  ونقول: يسمم الرأي العام.
- والتعبير الإنجليزي: the world conscience

(١) إبراهيم السامرائي، معجم ودراسة في العربية المعاصرة، لبنان ناشرون، ط١ ، ٢٠٠٠، ص١، ٢ بتصرف.

(٢) أحمد مختار عمر، معجم اللغة العربية المعاصرة، عالم الكتب، القاهرة، ط١، ٢٠٠٨، انظر المقدمة: مصادر التحرير، ومصادر المادة المسحية من ص٣٣ إلى ص٤٨.

ونقول: الضمير العالمي[١].

وتعابير أخرى ظهرت عن غير طريق الترجمة في حياتنا المعاصرة، مثل: "أراح عقله"، و"التضبيب" بمعنى: التعتيم، و"الذكاء الاصطناعي"، وهو من أشهر المصطلحات المتداولة الآن، ورغم ذلك لم نذكر في أحدث المعاجم العربية إصدارًا كمعجم اللغة العربية المعاصرة؛ لأن المصطلح زاد انتشاره بعد عام إصدار المعجم ٢٠٠٨م، والمصطلحات كل يوم في جديد نظرًا لجديد الاختراعات العلمية.

(ج) استخراج أبنية جديدة في العربية المعاصرة من المسح الحاسوبي:

منها ما هو موجود في معجم اللغة العربية المعاصرة؛ مثل: أكسد، بهرج، تكتك، مكيج[٢].

ومنها ما ليس موجودًا؛ مثل:

- "مكنن، ومعناه: أخضع الشيء للماكنة".
- هرطق: فِعْلٌ أُخِذَ من "الهرطقة"؛ أي: الإلحاد.
- موسق؛ فعل جديد أخذ من الاسم "موسيقى"، ويعني جعل الكلمة أو الكلام النص موسيقيًا.
- طنش: بمعنى: تعمد عدم الاهتمام.

وأتمنى في هذا الشأن أن نعتني بـ: "معجم تيمور الكبير في الألفاظ العامية" لـ: أحمد تيمور[٣]؛ والذي ورد فيه الكثير من الأبنية الحديثة وكشف عن أصولها التاريخية، وأهملتها معاجمنا الحديثة لأن مبدءها الأخذ عن السابقين، باستثناء ما تميز به معجم اللغة العربية المعاصرة للدكتور/ مختار عمر.

(١) اقتبستها من: د. إبراهيم السامرائي، معجم ودراسة في العربية المعاصرة، ص٣-٦.

(٢) انظر: إبراهيم السامرائي، مرجع سابق، ص٢٥-٤٠.

(٣) أحمد تيمور، معجم تيمور الكبير في الألفاظ العامية، تحقيق: د. حسين نصار؛ دار الكتب والوثائق القومية، القاهرة، الطبعة الثانية، ١٤٢٣هـ/٢٠٠٢م.

[٢] الذخائر اللغوية والنحو:

قبل ثمانينيات القرن العشرين، كانت الدراسات اللغوية التجريبية تضطر إلى الاعتماد بصورة أساسية على أساليب التحليل الثابتة، وكان هذا النوع من الدراسات يقدم وصفًا دقيقًا للمنظومة النحوية للغة، ولكن النتائج كان من الصعب أن ترصد معدلات التكرار الأعلى والأقل بشكل موضوعي، ومع ظهور الذخائر اللغوية المرمزة على مستوى تركيب الجملة، والتطور المستمر لأدوات البحث داخل الذخائر اللغوية، أصبح من السهل إجراء التحليل الكمي (Quantative Analysis) للظواهر النحوية بشكل أكبر مما سبق.

إن التحليل الكمي للظواهر النحوية على أقل تقدير يقدم للباحثين أفضل نماذج الاستخدام النحوي لتلك الظواهر، بالإضافة إلى كل درجات التحول التي تحدث وما إلى ذلك من معلومات، وهذه المعلومات لا تفيد فقط في فهم القواعد النحوية للغة ما، بل تفيد أيضًا في دراسة أوجه الاختلاف بين اللغات بعضها البعض، وفي مجال تعليم اللغات(١).

وعلى الرغم من كثرة ما تحويه المكتبة العربية من مؤلفات لغوية تتناول التراكيب النحوية إلا أن معظمها يقف عند فترة زمنية معينة لا تتجاوز القرن العاشر الهجري، مما استبعد من المؤلفات اللغوية مئات من التراكيب التي جدت بعد ذلك(٢).

ومن أهم المجهودات في الاستشهاد اللغوي المعاصر، "معجم الصواب اللغوي" للأستاذ الدكتور أحمد مختار عمر، والذي يكشف عنوانه الفرعي عن الهدف من تأليفه وهو: "دليل المثقف العربي"، وقد التزم فيه بما يأتي – على سبيل المثال –:

١. الاقتصار في المادة المعروضة على ما يشيع في لغة العصر الحديث على ألسنة المثقفين وفي كتاباتهم سواء استخلصناه بأنفسنا من لغة الإعلام، وكتابات الأدباء، أو وجدناه مذكورًا في دراسات السابقين.

(١) خوانج تشانغ نينغ، علم الذخائر اللغوية، ترجمة: هشام المالكي، ص٢٨٤.

(٢) انظر: شوقي ضيف، المدارس النحوية، ط٦، دار المعارف، القاهرة، ص٣٥٥.

٢. فتح باب الاستشهاد حتى يومنا هذا، وهو ما سبق أن طبقه مجمعنا اللغوي في معاجمه، وبذلك فتح الباب أمام الجميع لتخطي الحدود الزمانية والمكانية التي أقيمت خطأ بين عصور اللغة المختلفة.

وإذا كان مجمع اللغة العربية بالقاهرة قد توقف زمنيًا عند الثمانينيات من القرن العشرين، فإن معجمنا قد استوعب ما شاع في لغة العصر من الكلمات والاستعمالات التي خلا منها المعجم الوسيط، وإذا كان المعجم الوسيط يستشهد – على استحياء – بعدد محدود من المولدين والمعاصرين فقد فتحنا في معجمنا الباب على مصراعيه؛ ولذا نجد فيه أسماء، مثل: طه حسين، والعقاد، ومحمود تيمور، وتوفيق الحكيم، وأبي القاسم الشيباني، وميخائيل نعيمة، والطيب الصالح، ونجيب محفوظ... وغيرهم من المعاصرين، سواء كانوا أمواتًا أو أحياء.

٣. إجازة استعمال اللفظ على غير استعمال العرب مادام جاريًا على أقيستهم من مجاز واشتقاق، وتوسيع الدلالة وغيرها، وقديمًا قال ابن جني: "للإنسان أن يرتجل من المذاهب ما يدعو إليه القياس ما لم يلو بنص"(١).

وقال: "لو أن إنسانًا استعمل لغة قليلة عند العرب لم يكن مخطئًا لكلام العرب، لكنه يكون مخطئًا لأجود اللغتين"(٢).

٤. التوسع في فكرة التجمعات الحرة، والاختيارات الأسلوبية التي سمحت بتحريك الكلمات من مواقعها دون التزام بترتيب معين ما لم يكن هناك نص نحوي يعارض ذلك، وبناء عليه صوبنا تقديم الظرف "فقط" على متعلقه في مثل قولنا: "ليس فقط على المستوى المحلي"، **وأهم النقاط في مشروع المعجم الضخم، ما ذكره الدكتور مختار:**

(١) ابن جني ، أبو الفتح عثمان ، الخصائص، تحقيق : محمد علي النجار ، الهيئة المصرية العامة للكتاب ، الطبعة الخامسة ، ٢٠٠١م ، (١٨٩/١).

(٢) السابق، (١٢/٢).

٥. إذا كانت المعاجم السابقة قد ظهرت في شكل ورقي فقط، فقد حرصنا على تقديم هذا المعجم في شكلين: أحدهما ورقي، والأخر إلكتروني، وتتميز النسخة الإلكترونية باحتواء كل مدخل على مصادره اللغوية التي رجعنا إليها، بالإضافة إلى الإمكانات الهائلة في استدعاء المعلومة المطلوبة بسرعة(١).

ومما جاء في المعجم على المستوى الصرفي والنحوي، تصويبه ما كان مرفوضًا عند علماء العربية؛ مثل:

- كل عام وأنتم بخير؛ وهي: مرفوضة عند بعضهم؛ لأن الواو مقحمة بين المبتدأ و الخبر، وقد أجازه مجمع اللغة المصري على أن يكون "كل عام" مبتدأ حذف خبره، والتقدير: كل عام مقبل وأنتم بخير، والواو حالية، والجملة بعدها حال(٢).
- "أرجو الانتباه لاسيما وأن الأمر مهم"؛ وهي: مرفوضة عند بعضهم لمجيء الجملة بعد "لاسيما" مقترنة بالواو، وهو أسلوب غير عربي، لكن بعض النحويين أجازه على استعمال "لاسيما" بمعنى "خصوصًا"، فيؤتى بعدها بالحال مفردة، أو جملة مقترنة بالواو كما في المثال(٣).

وغير ذلك الكثير من الاستعمالات المعاصرة للتراكيب، ولكن معجم الصواب اللغوي صدر عام ٢٠٠٨م، فهل صدرت بعدها قرارات مجمعية بخصوص الاستعمالات اللغوية المعاصرة، وإن ظهر – وأعلم ذلك – فأين هو من الذخيرة اللغوية وقاعدة البيانات؟!

[٣] الذخائر اللغوية وعلم الدلالة:

تؤدي الذخائر اللغوية دورًا مهمًا في خدمة علم الدلالة، حيث يبرز دورها في إمداد علم الدلالة بشروح موضوعية تعتمد على أسلوب ديناميكي يتغير حسب

(١) أحمد مختار عمر، معجم الصواب اللغوي "دليل المثقف العربي"، عالم الكتب، القاهرة، ٢٠٠٨، الطبعة الأولى، باختصار من ص (ب) إلى ص (د).
(٢) السابق، ص٦٢٢.
(٣) السابق، ص٦٣٢.

طبيعة التغيرات اللغوية، ويتمثل أول دور مهم للذخائر اللغوية في علم الدلالة في إمكانية حصر المعاني الإضافية للكلمات بشكل موضوعي وفقًا للواقع اللغوي(١).

فمنذ القدم والكلمات تتطور معانيها، كتطور معنى كلمة "قطار" من موكب إبل إلى هذه الوسيلة المعروفة للنقل العام.

وحديثًا لاحظتُ الشباب يستخدمون كلمة "حبيبي" بمعنى: شكرًا.

ولا يخفى علينا عندما نقول عن فلان: عصفورة؛ بمعنى: نقَّال للكلام.

[٤] الذخائر اللغوية وتعليم اللغات:

إن الذخائر اللغوية تعتبر مصادر مهمة للأمثلة في عملية تعليم اللغة، لأن الدارسين للغة الثانية يحتاجون إلى الجمل والمفردات في الواقع اللغوي، ولابد من مراعاة اختلاف البيئة بمختلف أشكالها بين المجتمعات اللغوية.

ففي الإندونيسية مثلاً نجد الكثير من أنواع الفاكهة الاستوائية والتي لا توجد في البيئة العربية؛ وبالتالي ليس أمامنا إلا أن نترجمها ترجمة حرفية.

ولابد من مراعاة السياق والواقع الاجتماعي والثقافي؛ مثل:

- مستشفى (في العربية) تترجم إلى بيت المريض في الإندونيسية.
- المطب الصناعي يترجم إلى الإندونيسية بـ: البوليس النائم.
- ليلة الدخلة، تترجم إلى الإندونيسية بالليلة الأولى(٢).

ثالثًا– توظيف الإحصاء في علم الذخائر اللغوية (ماركوف نموذجًا):

تعتبر الذخائر اللغوية مصدرًا مهمًا لإجراء التحليل الكمي للغة، إلا أن استخدام الإحصاء الكمي في علم الذخائر اللغوية ليس بالأمر الذي يمكن إجراؤه بسهولة داخل المادة اللغوية الممثلة لمتن الذخيرة، والتقنيات الإحصائية المستخدمة في هذه

(١) خوانغ تشانغ نينغ، علم الذخائر اللغوية، ص٢٨٦.

(٢) بناء على زياراتي كأستاذ زائر في عدد من الجامعات الإندونيسية؛ وتبرز الحاجة إلى عدد من المعاجم ثنائية اللغة تراعي التنوع الثقافي بين اللغتين .

الحالة لا تقتصر على إجراء التحليل الرياضي للبيانات فحسب، بل يمكن استخدامها أيضًا في تفسير العلاقة بين كل من أسلوب الكتابة والتركيب اللغوي(١).

الهدف من استخدام تقنية الإحصاء في علم الذخائر اللغوية:

يمكننا عن طريق الإحصاء تنفيذ بعض التطبيقات، على سبيل المثال، يمكننا بعد تحليل المادة اللغوية المستخدمة في شبكة الإنترنت:

١. أن نتوصل بسهولة إلى معرفة التغيرات التي تحدث في الاستخدام اللغوي؛ من خلال مقارنة التغير في النماذج اللغوية قبل وبعد فترة تاريخية معينة. ومن هذه المقارنة يمكننا بسهولة التعرف على التغيرات التي تحدث في اللغة.
٢. يمكننا في كل عام نشر قوائم الكلمات والتعبيرات اللغوية الجديدة التي تنشأ في اللغة؛ وهذه من التطبيقات المبدئية والبسيطة التي يمكن أن تتم لدراسة حالات التغير اللغوي.
٣. يمكننا من اكتشاف ما يطلق عليه "المبادئ الجديدة للمصاحبة بين الكلمات"؛ من خلال مقارنة التغير في درجة الاقتران بين الكلمات.
٤. يستفاد من ذلك في أعمال الترجمة الكبيرة، حيث تتم معالجة المصاحبات اللغوية (التراكيب الجديدة) التي لا تظهر كثيرًا في المعاجم أو قوائم المصطلحات من خلال المترجم أو من خلال أحد المترجمين ذوي الخبرة قبل البدء في أعمال الترجمة(٢).

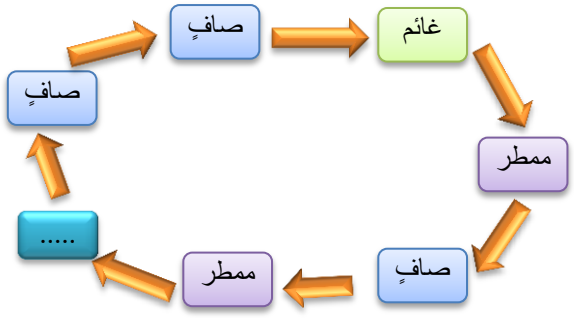
ويمكن تطبيق الأهداف السابقة عن طريق نموذج ماركوف الكامن (HMM).

تعريف نموذج ماركوف (HMM):

تنسب سلاسل ماركوف إلى اسم مكتشفها العالم الروسي أندريا ماركوف عام (١٩٠٧) الذي وضع المفاهيم الأساسية لسلسلة ماركوف(٣).

(١) خانغ تشانغ نينغ، علم الذخائر اللغوية، ص٢٠٥.
(٢) جانغ جنغ، مقدمة في علم اللغة الحاسوبي، بتصرف، ص٥٣-٥٤.
(٣) إنصاف جاسم، مقدرات بيز لبعض نماذج ماركوف المخفية مع التطبيق، جامعة كربلاء، ٢٠٢٣، ص٢٣.

وقبل أن نفهم نموذج ماركوف الكامن، علينا أن نفهم ما هي الحالة وما هو نموذج التوليد، الحالة يمكن استخدامها في وصف بعض الخصائص التي تنطبق على أي شيء في الوقت الراهن؛ على سبيل المثال، الإضاءة في إشارات المرور من المؤكد أنها تحمل إحدى الألوان بين الأحمر، والأصفر، والأخضر. أما حالات الطقس التي تتسم بالتعقيد، فمن الممكن أن تتراوح بين إحدى حالات ثلاث، هي: صافٍ، وغائم، وممطر.



الصورة البسيطة لنموذج ماركوف

ومن خلال مراقبة هاتين الظاهرتين لمدة معينة من الوقت، يمكننا ملاحظة أن هناك فارقًا جوهريًّا بين اللمبات المضيئة لإشارة المرور، وبين تسلسل حالات الطقس، فإشارات المرور عبارة عن نظام يتم التحكم فيه بشكل صارم، وعلى ذلك يعتبر نظامًا توليديًا محددًا، فإذا أضاءت إحدى الإشارات، فمن المؤكد أنه بعد وقت محدد ستضيئ اللمبة التالية التي تم تحديدها مسبقًا. أما تغيرات الطقس فتختلف اختلافًا كليًا، لأنها تحدث لنمط توليدي غير محدد، وعلى الرغم من أننا قد نستطيع التنبؤ بطقس اليوم استنادًا إلى يوم سابق، أو يومين سابقين، أو عدد (ن) من الأيام حيث (ن>٢)، ولكن هذا التوقع يظل نوعًا من التخمين، يتراوح بين احتمالات (Probability) ثلاثة من احتمالات الطقس، ولا يمكن أن يكون ذلك الاحتمال مؤكدًا.

وبالنسبة لمعالجة اللغات الطبيعية، يمكننا أن ننظر إلى جميع العلامات اللغوية على أن كلاً منها يمثل حالة، كما يمكننا أن ننظر إلى مجموعة كلمات متتالية على

أنها تسلسل لتلك الحالات، والأمر تمامًا كما تحدثنا عن النشرة الجوية، فعلى فرض أن أحدهم أخفى الكلمة الأخيرة، وطلب منك أن تخمن ما تلك الكلمة فمن المؤكد أنك ستعتمد على عددٍ من الكلمات السابقة في التوصل إلى الحكم.

إن الطريقة التي يقوم بها الحاسب الآلي في التوقع تتشابه تمامًا مع ما سبق. ففي البداية "نُخبر" الحاسب الآلي بمعلومات لغوية كافية، وكلما زاد حجم المعلومات كانت النتيجة أفضل.

فإذا كانت الكلمات التي نريد توقعها تنتمي إلى مجال لغوي محدد، فمن المؤكد أن المعلومات اللغوية التي يتم إدخالها ينبغي أن تنتمي إلى هذا المجال، أما إذا لم يكن المجال محددًا فينبغي أن تتصف المعلومات اللغوية التي يتم إدخالها بالتوازن، وغالبًا ما نسمي تلك المعلومات اللغوية باسم ذخيرة (Corpus)[١].

تطبيقات نموذج ماركوف HMM في الذخائر اللغوية:

نفترض أن لدينا سلسلة من الكلمات $W_1, W_2, W_3, \ldots . W_t$، ونريد توصيف تلك السلسلة من حيث الأنواع النحوية لها $C_1, C_2, C_3, \ldots . C_r$، ونظرًا إلى انتشار ظاهرة اللبس اللغوي في تمييز الأنواع النحوية للكلمات، فمن الممكن أن يقابل السلسلة الواحدة من الكلمات عدة سلاسل من أنواع الكلمات، هذا بالإضافة إلى أن سلسلة أنواع الكلمات التي نريد الحصول عليها ستجعل قيمة المعادلة: $Prob\ (C_1, C_2, C_3, \ldots . C_r \mid W_1, W_2, W_3, \ldots . W_t)$ تعادل قيمة أكبر سلسلة من الأنواع النحوية للكمات.

وباستخدام قانون بايز[٢] Bayes للاحتمالات يمكننا كتابة المعادلة السابقة بالصيغة التالية:

$$\frac{PROB(W_1, W_2, W_3, \ldots W_t | C_1, C_2, C_3, \ldots C_r)\ PROB\ (C_1, C_2, C_3, \ldots C_r)}{P(W_1, W_2, W_3, \ldots . W_r)}$$

(١) جانغ جنغ، مقدمة في علم اللغة الحاسوبي، ص٤٩-٥١.

(٢) قانون بايز: هو إحدى النتائج المهمة لنظرية الاحتمالات، ويقوم بحساب التوزيع الاحتمالي الشرطي للمتغير العشوائي A بمعلومة المتغير العشوائي B.

حيث يطلق على $PROB\ (\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \dots . \mathrm{W}_t | \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_r)$ اسم معادلة المعلومات المتعلقة بالمفردات، ويطلق على $PROB\ (\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_r)$ النموذج اللغوي، ونظرًا إلى ثبات المقام بالنسبة إلى سلاسل الكلمات المتساوية فإن المعادلة السابقة يمكن اختصارها إلى المعادلة التالية التي تحسب أكبر سلسلة من أنواع الكلمات:

$$PROB\ (\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \dots . \mathrm{W}_t) | (\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t) \times PROB\ (\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t)$$

ويمكننا وضع **مستوى** أعلى من الفروض للمعادلة السابقة: إن احتمال ورود الكلمة الحالية يتم التوصل إليه من النوع النحوي للكلمة، والنوع النحوي لهذه الكلمة مرتبط فقط بنوع الكلمة السابقة لها، وفي النهاية يتم التعبير عن الموضوع بالكامل من خلال المعادلة التالية:

$$T^* = \arg\max p(W_1)\ p(W_1|C_1) p(C_1) \prod_{i=2}^{T} p(C_1 | C_1, C_2 \dots . C_{i-1}) p(W_1|C_1)$$

حيث تشير T^* إلى سلسلة الكلمات التي يتم ترميزها في النهاية، وتشير $p(\dots)$ إلى احتمال(١):

من المعادلة السابقة يمكننا أن نستخرج المستوى الأول والثاني من نموذج ماركوف الكامن HMM؛ حيث إن المستوى الأول لنموذج HMM يعبر عن أن النوع النحوي للكلمة الحالية لا يرتبط إلا بنوع الكلمة السابقة عليها، وتكون المعادلة بالتفصيل كما يلي:

$$T^* \arg\max_{C_1, C_2 \dots . C_r} p(C_1)\ p(W_1|C_1) p(C_1) \prod_{i=2}^{T} p(C_1 | C_1, C_2 \dots . C_{i-1}) p(W_1|C_1)$$

(١) يعبر عن نماذج ماركوف المخفية بالصيغة $(\mathrm{A}, \mathrm{B}, \pi)$، إذ إن A هي مصفوفة احتمال انتقالية الحالة، B هي مصفوفة احتمالية رابطة بين الحالات المخفية والمشاهدات، Π هي متجه توزيع الحالة الابتدائية.
إنصاف جاسم، مقدرات بايز، ص٢٥.

حيث تعبر $p(C_1|C_{i+1})$ عن احتمال تغير الحالة في نموذج HMM وتعبر $p(W_1|C_1)$ عن احتمال توليد الكلمات(١).

C_{i+1} : تمثل الحالة المستقبلية.

C_i : تمثل الحالة الحالية.

C_{i-1} : تمثل الحالة السابقة.

أمثلة توضيحية:

المثال الأول:

نفترض أن لدينا سلسلة من الكلمات $[\mathrm{W}_1, \mathrm{W}_2, \dots \dots \mathrm{W}_t]$

المطلوب توصيف السلسلة من حيث الأنواع النحوية $(\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t)$.

مع مراعاة "ظاهرة اللبس اللغوي".

ممكن أن يقابلها

فالسلسلة الواحدة من الكلمات ⟵ **عدة سلاسل من الأنواع النحوية**

السلسلة التي ستحصل عليها ستجعل:

أكبر سلسلة من الأنواع النحوية للكلمات ⟵

$PROB\ (\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \dots . \mathrm{W}_t)|(\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t)$

قانون بايز (الاحتمال المشروط):

$$PROB\ (\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t / \mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \dots . \mathrm{W}_t) = PROB\ (\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \dots . \mathrm{W}_t / \mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \dots . \mathrm{C}_t) \times PROB\ (\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3)$$

النموذج اللغوي **معادلة المعلومات المتعلقة بالمفردات**

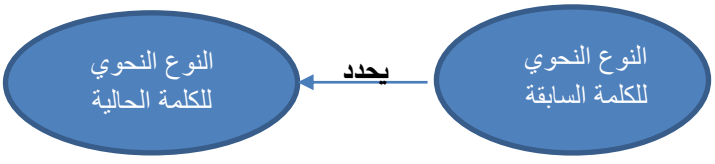
(١) خوانغ تشانغ نينغ، علم الذخائر اللغوية، ترجمة : أ.د/ هشام المالكي ، ص٢٢٤–٢٢٦.(بتصرُّف بسيط مع توضيح إلى مايشير إليه كل رمز باجتهاد من الباحث).

$$T = \arg Max\, p(W_1)\, p(W_1|C_1) p(C_1) \prod_{i=2}^{T} (C_i, C_1, C_2, \ldots . C_{i-1})\, p(W_i|C_i)$$

(ضرب كل القيم ضمن نطاق تسلسلي) Capital Pi

T = سلسلة الكلمات التي يتم ترميزها في الذخائر اللغوية

المستوى الأول:

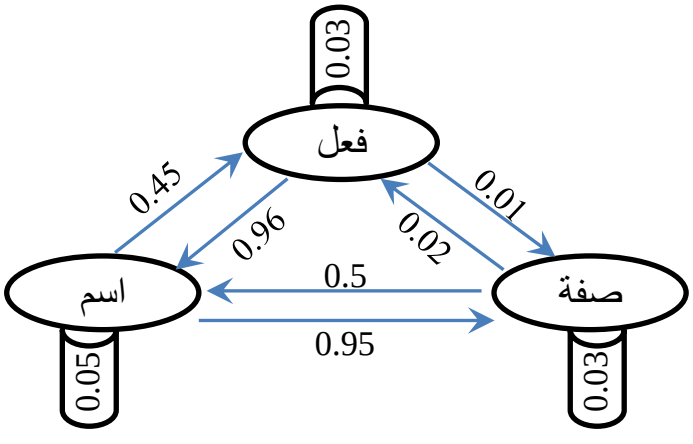


$$T = \arg \max_{C_1, C_2 \ldots . C_t} p(C_1)\, p(W_1|C_1) \prod_{i=2}^{t} p(C_i\, , C_1, C_2 \ldots . C_{i-1})\, p(W_1|C_1)$$

احتمال توليد كلمات

المثال الثاني:

نفترض أن لدينا ثلاث حالات (فعل، اسم، صفة)؛ احتمالية الانتقال من الفعل إلى الصفة هي ٠.٠١، ومن الصفة إلى الفعل هي ٠.٠٢، فتكون مصفوفة الانتقال من حالة إلى أخرى في الشكل التالي:



نموذج ماركوف بثلاث حالات

احتمالية الانتقال (مصفوفة الانتقال)[1]:

Probability of Transition (Transition Matrix)

الحالة	الحالة المستقبلية			
		فعل	اسم	صفة
الحالة السابقة/ الحالية	فعل	P(فعل/فعل) = 0.03	P(اسم/فعل) = 0.96	P(صفة/فعل) = 0.01
	اسم	P(فعل/اسم) = 0.45	P(اسم/اسم) = 0.05	P(صفة/اسم) = 0.05
	صفة	P(فعل/صفة) = 0.02	P(اسم/صفة) = 0.95	P(صفة/صفة) = 0.03

نماذج تطبيقية:

أولاً- المحلل الصرفي:

تعد معالجة اللغة الطبيعية واحدة من أكثر المواضيع اهتمامًا في مجال الحاسب والذكاء الاصطناعي، ومعالجة اللغة الطبيعية تتم إما باستخدام القواعد المعيارية أو بالنهج الإحصائي؛ ونحن مع هذه الثورة المعلوماتية نحتاج إلى المزج بين الأمرين:

واللغة العربية تحتاج في معالجتها إلى الأمرين؛ لأن معظم كلمات اللغة العربية من وحدات متصلة عبارة عن بدايات وجذوع وجذور ولواحق.

البادئة + (الجذع ← الجذر) + اللاحقة

مثال: كلمة: "يتعارفون"

البادئة + الجذع (الساق) + اللاحقة

ي + تعارف + ون

↓

الجذر (ع ر ف)

(1) Dima, p. 242.(سبق توثيق المرجع في الدراسات السابقة)

وصف المشكلة:

أحتاج تصنيف النص لاستخدامه في الترجمة الآلية وتحليل المشاعر.

حل المشكلة:

١. تقسيم الجملة إلى كلمات.

٢. يعتمد تصنيف النص ومعالجته بشكل أساسي على إيجاد سمات معينة للكلمة، ومن أهم سمات الكلمة هو جذع الكلمة؛ والذي يمكن تحقيقه بإزالة البادئات واللواحق من الكلمة.

يمثل جذع الكلمة (تعارف) في المثال السابق؛ أحد النماذج الاشتقاقية المأخوذة من الجذر، الذي يحمل إلى جانب معناه المعجمي المعنى الصرفي، أو معنى الصيغة المتولد عن وضع حروف الكلمة الأصلية في قالب معين اصطلح الصرفيون على تسميته بالميزان الصرفي(١).

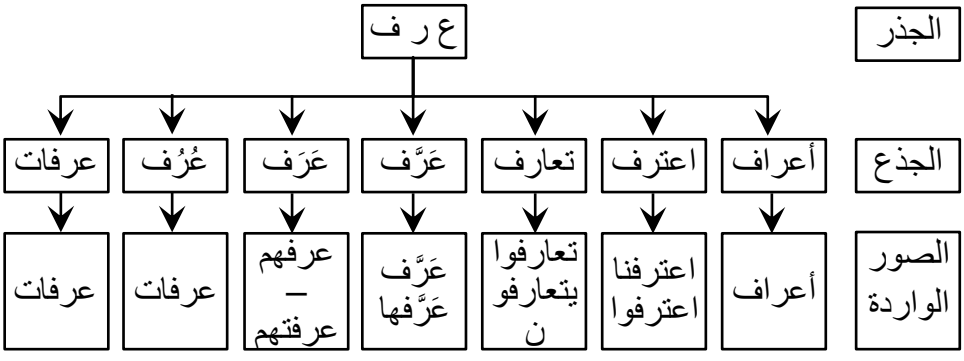
٣. تبلغ أوزان اللغة العربية ٣٠٠ وزن تقريبًا؛ وهذا يتطلب منا وجود ذخيرة لغوية تسمح بمعالجة الكلمات؛ وبالتالي الجذع (تعارف) يستدعي الجذوع المشتركة معه في الجذر، وأن نعالجها بصورة متتابعة باعتبارها أفرادًا في أسرة واحدة يحمل كل فرد فيها ملامح مشتركة تجمع بينها؛ لأن المعنى الكلي للكلمة لا يتضح إلا بمجموع معنييها المعجمي والصرفي.

٤. عن طريق HMM يتم طرح جميع الجذور الممكنة، مع استخدام الملاحظة البشرية للحصول على الصحيح والمجرد من البدايات واللواحق سواء أكان ثلاثيًا أو غيره، ويتم أخذ موضع الكلمة في النص في الاعتبار فالجذر هو الذي يحمل المعنى الأساسي للكلمة.

(١) أحمد مختار عمر، المعجم الموسوعي لألفاظ القرآن الكريم،عالم الكتب ، القاهرة ، ط١ ، ٢٠٠٢ ص٣٨.

٥. تتم هذه العملية حاسوبيًا بشكل أساسي عن طريق ترميز أجزاء الكلام باستخدام HMM.

٦. يتم استخدام HMM لتمثيل بنية الجمل العربية على شكل هرمي، مثال ذلك[١]:



٧. في قوله تعالى: (فَإِذَا أَفَضْتُم مِّنْ عَرَفَاتٍ فَاذْكُرُوا اللَّهَ)[٢]، إذا تم تقسيم الجملة إلى كلمات وتم ترميزها وإدخالها في ذخيرة لغوية لمعرفة المصاحبات اللفظية، لا يمكن أن نضع إلا كلمة "عرفات" في الآية بالرغم من وجود جذوع أخرى لنفس الجذر.

٨. رغم ظهور دراسات سابقة تناولت الشكل الهرمي للمحلل الصرفي عن طريق HMM انظر: (ديما، ٢٠١٧) و(العجمي، ٢٠١١)، إلا أن الأستاذ الدكتور أحمد مختار عمر في المعجم الموسوعي لألفاظ القرآن الكريم سبق الجميع بفريقه المعجمي من اللغويين والبرمجيين وتم إخراج المعجم عام٢٠٠٢م، وما نحتاجه الآن أن نكمل المسيرة، وأن نبدأ من حيث انتهى هؤلاء.

(١) مختار، المعجم الموسوعي، بتصرف، ص٣٤.
(٢) سورة البقرة، آية [١٩٨].

ثانيًا– معرفة الكلمات الأكثر تكرارًا:

هذه بعض التعبيرات والكلمات التي قمت باستقصاء عدد مرات ورودها عبر شبكة الإنترنت من خلال ذخيرة Sketch engine، وطبقت عليها نظرية ماركوف من خلال برنامج بايثون لتوقع عدد مرات ورودها في ٢٠٢٥م.

[١] دموع التماسيح:

رغم أنه تعبير جديد على العربية كما سبق في هذا البحث؛ إلا أن توقع مرات ورودها في ٢٠٢٥ في تناقص شديد حسب نظرية ماركوف.

العام	٢٠٢٢	٢٠٢٣	٢٠٢٤	المجموع = ٤٨٦
عدد مرات الورود	٢٦٨	١١٩	٩٩	

المتوقع في عام ٢٠٢٥ حسب نظرية ماركوف ٦٣ مرة فقط بفاقد نسبة مئوية –٣٦,٢٠.

```python
import numpy as np

# إعداد البيانات التاريخية للغات على مدار السنوات من 2020 إلى 2023
historical_data = np.array([[268],[119],[99]])

# حساب معدلات التغير لكل لغة بين السنوات
rates_of_change = (historical_data[1:] / historical_data[:-1]) - 1

# حساب متوسط معدلات التغير
average_change_rate = np.mean(rates_of_change + 1, axis=0)

# تشكيل مصفوفة التحولات بناءً على متوسط معدلات التغير
num_languages = len(average_change_rate)
transition_matrix = np.zeros((num_languages, num_languages))

# ملء مصفوفة التحولات حيث كل لغة تعتمد فقط على معدلات التغير الخاصة بها (لنموذج مبسط)
for i in range(num_languages):
    transition_matrix[i, i] = average_change_rate[i]

# إعداد البيانات الأولية لعام 2023
languages_2023 = historical_data[-1]
```

```
English: 63 speakers, -36.20% increase
```

[٢] ابتسامة هادئة:

من التعبيرات الجديدة أيضًا كما سبق.

العام	٢٠٢٢	٢٠٢٣	٢٠٢٤	المجموع = ٣٨
عدد مرات الورود	٣٠	٥	٣	

المتوقع في عام ٢٠٢٥ حسب نظرية ماركوف #١# مرة واحدة فقط بفاقد نسبة مئوية –٦١,٦٧.

```python
import numpy as np

# إعداد البيانات التاريخية للغات على مدار السنوات من 2020 إلى 2023
historical_data = np.array([[30],[5],[3]])

# حساب معدلات التغير لكل لغة بين السنوات
rates_of_change = (historical_data[1:] / historical_data[:-1]) - 1

# حساب متوسط معدلات التغير
average_change_rate = np.mean(rates_of_change + 1, axis=0)

# تشكيل مصفوفة التحولات بناءً على متوسط معدلات التغير
num_languages = len(average_change_rate)
transition_matrix = np.zeros((num_languages, num_languages))

# ملء مصفوفة التحولات حيث كل لغة تعتمد فقط على معدلات التغير الخاصة بها (لنموذج مبسط)
for i in range(num_languages):
    transition_matrix[i, i] = average_change_rate[i]

# إعداد البيانات الأولية لعام 2023
languages_2023 = historical_data[-1]
```

```
English: 1 speakers, -61.67% increase
```

[٣] بلورة الفكرة:

العام	٢٠٢٢	٢٠٢٣	٢٠٢٤	المجموع = ٣٧
عدد مرات الورود	٢٨	٧	٢	

المتوقع في عام ٢٠٢٥ [صفر] بفاقد نسبة مئوية –٧٣,٢١.

```python
import numpy as np

# إعداد البيانات التاريخية للغات على مدار السنوات من 2020 إلى 2023
historical_data = np.array([[28],[7],[2]])

# حساب معدلات التغير لكل لغة بين السنوات
rates_of_change = (historical_data[1:] / historical_data[:-1]) - 1

# حساب متوسط معدلات التغير
average_change_rate = np.mean(rates_of_change + 1, axis=0)

# تشكيل مصفوفة التحولات بناءً على متوسط معدلات التغير
num_languages = len(average_change_rate)
transition_matrix = np.zeros((num_languages, num_languages))

# ملء مصفوفة التحولات حيث كل لغة تعتمد فقط على معدلات التغير الخاصة بها (لنموذج مبسط)
for i in range(num_languages):
    transition_matrix[i, i] = average_change_rate[i]

# إعداد البيانات الأولية لعام 2023
languages_2023 = historical_data[-1]
```

```
English: 0 speakers, -73.21% increase
```

[٤] دفع الثمن غاليًا:

العام	٢٠٢٢	٢٠٢٣	٢٠٢٤	المجموع = ٧٥٤
عدد مرات الورود	٤٠٧	٢٠٥	١٤٢	

المتوقع في عام ٢٠٢٥ [٨٤] مرة بفاقد نسبة مئوية –٤٠,١٨.

```python
import numpy as np

# إعداد البيانات التاريخية للغات على مدار السنوات من 2020 إلى 2023
historical_data = np.array([[407],[205],[142]])

# حساب معدلات التغير لكل لغة بين السنوات
rates_of_change = (historical_data[1:] / historical_data[:-1]) - 1

# حساب متوسط معدلات التغير
average_change_rate = np.mean(rates_of_change + 1, axis=0)

# تشكيل مصفوفة التحولات بناءً على متوسط معدلات التغير
num_languages = len(average_change_rate)
transition_matrix = np.zeros((num_languages, num_languages))

# ملء مصفوفة التحولات حيث كل لغة تعتمد فقط على معدلات التغير الخاصة بها (لنموذج مبسط)
for i in range(num_languages):
    transition_matrix[i, i] = average_change_rate[i]

# إعداد البيانات الأولية لعام 2023
languages_2023 = historical_data[-1]
```

```
English: 84 speakers, -40.18% increase
```

[٥] تأكسد:

العام	٢٠٢٢	٢٠٢٣	٢٠٢٤	المجموع = ٧٨١
عدد مرات الورود	٣٩٧	٢٢٠	١٦٤	

المتوقع في عام ٢٠٢٥ [١٠٦] مرة؛ بفاقد نسبة مئوية –٣٥,٢٩.

```python
import numpy as np

# إعداد البيانات التاريخية للغات على مدار السنوات من 2020 إلى 2023
historical_data = np.array([[397],[220],[164]])

# حساب معدلات التغير لكل لغة بين السنوات
rates_of_change = (historical_data[1:] / historical_data[:-1]) - 1

# حساب متوسط معدلات التغير
average_change_rate = np.mean(rates_of_change + 1, axis=0)

# تشكيل مصفوفة التحولات بناءً على متوسط معدلات التغير
num_languages = len(average_change_rate)
transition_matrix = np.zeros((num_languages, num_languages))

# ملء مصفوفة التحولات حيث كل لغة تعتمد فقط على معدلات التغير الخاصة بها (لنموذج مبسط)
for i in range(num_languages):
    transition_matrix[i, i] = average_change_rate[i]

# إعداد البيانات الأولية لعام 2023
languages_2023 = historical_data[-1]
```

```
English: 106 speakers, -35.02% increase
```

وهذه النتائج السلبية لا تعود في الأساس إلى الاستخدام اللغوي الشائع في أروقة العلم والكتب؛ وإنما تعكس تراجع استخدام اللغة العربية في الحياة العامة المعاصرة رغم أنها تعبيرات جديدة على العربية.

فهل لنا من وقفة؟! هل لنا أن تتحد المجامع اللغوية العربية لإنتاج أول ذخيرة لغوية عربية تعتمد على تقنيات الحاسوب والبرمجة والذكاء الاصطناعي؟

أرجو أن نرى ذلك في المستقبل القريب إن شاء الله، فلن تعدم الأمة من كوادرها على اختلاف التخصصات ولكننا فقط نحتاج إلى منظومة وتوحيد الجهود.

الخاتمة

توصلت الدراسة إلى ما يلي:

١. ضرورة البحث عن حلول علمية عملية تساعد على زيادة انتشار اللغة العربية؛ ومواكبة العصر في التقنيات الحديثة للمعالجة الآلية للغات الطبيعية.

٢. لا يمكن أن يستقيم البحث في العلوم اللغوية والأدبية في وقتنا المعاصر والمستقبلي دون الاعتماد على ذخيرة لغوية، وإغفال الباحث لتقنيات تهيئة الذخائر اللغوية الإلكترونية من شأنه أن يقلص مهارات العمل البحثي وما يترتب عليه من جمود بل تراجع انتشار لغتنا العربية.

٣. ينبغي على القائمين على أقسام اللغة العربية في شتى بقاع الوطن العربي والدول الإسلامية غير الناطقة بالعربية أن تدمج بين الجانب النظري للغة المعيارية والجانب التقني والتطبيقات الإحصائية للغة المعاصرة.

٤. لابد من التحرك السريع لتسجيل التغيرات اللغوية التي تحدث في الاستخدام اللغوي المعاصر؛ وهي غير موجودة في معاجمنا العربية.

٥. من خلال هذا التسجيل يمكننا بسهولة نشر قوائم الكلمات والتعبيرات اللغوية الجديدة التي تنشأ في اللغة.

٦. يستفاد من ذلك في الترجمة الآلية وتحليل النص وتحليل المشاعر.

٧. تهدف المعالجة الآلية للغة العربية إلى دراسة تقليل الأوزان المستخرجة من خلال تجميع الأوزان ذات المعنى الواحد.

٨. من خلال تدريسي للغة العربية لغير الناطقين بها، أتمنى إنشاء برنامج يكتب فيه الطالب المجال الدلالي فقط الذي يريد البحث فيه، ثم يكتب ما في عقله من كلمات عربية يعرفها لتتهيأ إليه بعد ذلك على الشاشة الموضوعات بأكملها مرتبة ومفهرسة بأبسط التعابير وأسهلها، ولا يتم ذلك إلا بذخيرة لغوية تتبناها مؤسسات قومية.

المصادر والمراجع:

١. إبراهيم السامرائي، معجم ودراسة في العربية المعاصرة، لبنان ناشرون، الطبعة الأولى، ٢٠٠٠.
٢. ابن جني، أبو الفتح عثمان ، الخصائص ، تحقيق : محمد علي النجار، الهيئة المصرية العامة للكتاب ، القاهرة ، الطبعة الخامسة ، ٢٠١١
٣. ابن حزم، أبو محمد علي بن أحمد بن سعيد، الإحكام في أصول الأحكام، تحقيق: أحمد محمد شاكر، دار الآفاق الجديدة، بيروت. ٢٠٠٧
٤. أحمد تيمور، معجم تيمور الكبير في الألفاظ العامية، تحقيق: د. حسين نصار؛ دار الكتب والوثائق القومية، القاهرة، الطبعة الثانية، ١٤٢٣هـ/٢٠٠٢م.
٥. أحمد مختار عمر، المعجم الموسوعي لألفاظ القرآن الكريم، عالم الكتب ، القاهرة ، الطبعة الأولى ٢٠٠٢م.
٦. أحمد مختار عمر، صناعة المعجم الحديث، عالم الكتب، القاهرة، الطبعة الثانية، ٢٠٠٩.
٧. أحمد مختار عمر، معجم الصواب اللغوي "دليل المثقف العربي"، عالم الكتب، القاهرة، الطبعة الأولى، ٢٠٠٨.
٨. أحمد مختار عمر، معجم اللغة العربية المعاصرة، عالم الكتب، القاهرة، الطبعة الأولى، ٢٠٠٨.
٩. إنصاف جاسم، مقدرات بيز لبعض نماذج ماركوف المخفية مع التطبيق، جامعة كربلاء، ٢٠٢٣.
١٠. جانغ جنغ، مقدمة في علم اللغة الحاسوبي والترجمة الآلية، ترجمة: هشام موسى المالكي، المركز القومي للترجمة، مصر، الطبعة الأولى،٢٠٢٣م.
١١. خوانغ تشانغ نينغ، علم الذخائر اللغوية، ترجمة: هشام موسى المالكي، المركز القومي للترجمة، مصر، الطبعة الأولى، ٢٠١٦م
١٢. شوقي ضيف، المدارس النحوية، دار المعارف، القاهرة ، الطبعة السابعة.
١٣. محمود أحمد السيد، مستقبل اللغة العربية ومتطلبات العصر القادم، مجلة مجمع اللغة العربية، دمشق، ٢٠١٢، المجلد (٨٧)، الجزء (١).
١٤. مركز دعم اتخاذ القرار، مجلس الوزراء، مصر، ١٦ يونيو ٢٠٢٢، مقال بعنوان: "الوجاهة الاجتماعية وانتشار ثقافة المدارس الدولية"، أماني عاطف.

١٥- هشام موسى المالكي، إشكاليات تهيئة الذخائر اللغوية وبنائها حاسوبيًا – اللغتان العربية والصينية نموذجًا – مجلة أواصر، المركز القومي للترجمة، مصر، ج٢، أبريل ٢٠٠٩.

16- Suleiman, Dima, Awajan,Arafat, Al Etaiwi, Wael, " The Use of Hidden Markov Model in Natural ARABIC Language Processing: a survey", published by Elsevier B.V., 2017.

17- A. F. Alajmi, E. M. Suad and M. H. Abdalla, "Hidden Markov Model Based Arabic Morphological Analyzer", Communication and Electronics Department, Faculty of Engineering, Helwan University, Egypt, 2011.