مجلة اللسانيات العربية
The Arabic Linguistics Journal
مجلة علمية محكمة تصدر عن مركز الملك عبدالله بن عبدالعزيز للتخطيط والسياسات اللغوية، الرياض، المملكة العربية السعودية
مركز الملك عبدالله بن عبدالعزيز للتخطيط و السياسات اللغوية
King Abdullah Bin Abdulaziz Center For Language Planning and Policies
ISSN: 1658-7421

مجلة اللسانيات العربية، العدد 14، جمادى الآخرة، 1443هـ/يناير، 2022م

التشابه والاختلاف في الوحدات المعجمية بين عربية التراث والعربية المعاصرة

أفراح بنت عبد العزيز التميمي

قسم الإعداد اللغوي، معهد تعليم العربية، جامعة الإمام محمد بن سعود الإسلامية، الرياض، السعودية

توثيق البحث APA Citation:

التميمي، أفراح. (2022). التشابه والاختلاف في الوحدات المعجمية بين عربية التراث والعربية المعاصرة. *مجلة اللسانيات العربية*، *14*، 81-98.

تاريخ الإرسال:1442/08/16

تاريخ القبول: 1442/10/12

Submission Date: 29/03/2021

Acceptance Date: 27/05/2021

الملخص

تتنزل هذه الورقة في مجال معالجة اللغات الطبيعية، وتطبق منهج التجميع الذي يعد منهجا في تعلم الآلة غير الموجه، والمستفاد منه أيضا في تعلم الآلة الموجه لتحسين جودة التصنيف. وأكشف فيها عن التشابه والاختلاف في الوحدات المعجمية بين العربية التراثية، والعربية المعاصرة. وتجيب هذه الورقة عن السؤال الآتي: هل ثمة تشابه معجمي في الوحدات المعجمية بين العربيتين التراثية والمعاصرة؟ وقد اعتمدت في ذلك على مدونتين جمعتهما، مثلت الأولى العربية التراثية، فيما مثلت الثانية العربية المعاصرة. وقد حسبت بُعديْ التشابه والاختلاف باستعمال خوارزمية جاكارد. وكشفت النتائج عن حجم الاختلاف الكبير في الوحدات المعجمية بين العربيتين، وعن تشاركهما وتشابههما بدرجة قدرها 0.31، مما يعني النمو المعجمي الهائل من جانب العربية المعاصرة.

الكلمات المفتاحية: تعلم الآلة - جاكارد – العربية التراثية – العربية المعاصرة -خوارزميات التشابه الاختلاف.

Abstract

Similarity and dissimilarity in the lexical units between heritage and contemporary Arabic

This paper belongs to the field of NLP and applies the clustering method. The latter is based on unsupervised ML, which improves supervised ML in terms of classification systems. I revealed the similarity and dissimilarity between heritage and contemporary Arabic at the level of lexical units. This paper answers the following question. Is there a lexical similarity in the lexical units between the heritage and the contemporary Arabic? I relied on two collected corpora. The former represented heritage Arabic while the latter represented contemporary Arabic. I calculate the similarity and dissimilarity adopting Jaccard algorithm. The results showed the similarity between the two corpora with the score of 0.31. This indicates that there was a significant growth in the contemporary lexicon.

Keywords: Machine Learning - Jaccard - Heritage Arabic - Contemporary Arabic - Similarity and dissimilarity Algorithms.

1. المقدمة

أدى التضخم في كم البيانات بأنواعها إلى تطوير أدوات وتقنيات تساهم في استخراج المعارف من تلك البيانات بسهولة ويسر (Hand & Adams، 2014). والتنقيب في البيانات النصية Text Mining بوصفها أحد أنواع البيانات هو إحدى التقنيات الحديثة التي تنطوي على عمليات استخراج المعارف المفيدة واستكشافها من النصوص المنظمة وغير المنظمة (Aggarwal

Zhai &، 2012، ص ص 2-3). وتزودنا تقنيات التنقيب في النصوص بالإجابات عن العديد من الأسئلة المتعلقة بالنصوص في وقت قياسي، خاصة تلك الأسئلة التي كان من غير الممكن أن يجاب عنها من خلال الأساليب الإحصائية التقليدية.

ويُجرى التنقيب في البيانات اللغوية بالبحث عن المصادر. وهذه المصادر تكون في صيغ رقمية متعددة الامتدادات، كأن تكون في ملفات نصية حاسوبية، مثل word وpdf، أو تكون في صفحات الويب على شبكة الإنترنت html، أو غيرها. وقد سهل لنا الحصول على بيانات هذه الورقة، توفر البيانات اللغوية في الصيغ الرقمية word وtext وhtml القابلة للمعالجة المباشرة التي لا تتطلب تحويل النصوص من مكتوب يدوي إلى مكتوب آليا.

ومن المهام المهمة والمفيدة في التنقيب في النصوص مهمة التجميع clustering التي تعد منهجا من مناهج تعلم الآلة غير الموجه، والتي يستفاد منها أيضا في تعلم الآلة الموجه لتحسين جودة التصنيف. وتُعرف مهمة التجميع بأنها إيجاد مجموعة من الكيانات المتشابهة في قاعدة البيانات (Aggarwal & Zhai، 2012، ص 78)، سواء كان ذلك على مستوى الحرف، أو الكلمة، أو الجملة، أو الفقرة، أو النص.

وفي مجال التنقيب في البيانات النصية، تُعد حسابات التشابه بين النصوص عنصرا مهما في مهام متعددة في مجال معالجة اللغات الطبيعية (Su & Seoung، 2017). وقد تكون أوجه التشابه والاختلاف في النصوص بين الكلمات، أو بين الجمل، أو بين الفقرات، أو غير ذلك. كما يمكن أن تتشابه هذه العناصر في النصوص تشابها معجميا أو تشابها دلاليا. ويعتمد التشابه المعجمي على التطابق بين أحرف الوحدات المعجمية، كأن تتطابق كلمة (رئيس) في النص الأول مع كلمة (رئيس) في النص الثاني. فيما يعتمد التشابه الدلالي على التشابه في المعنى، كأن تتشابه كلمة (الرئيس) مع (حاكم)، وعبارة (تغادر الطائرة صباحا) مع (تقلع الطائرة الساعة 11:00) في الدلالة.

وهذه الورقة تجيب عن السؤال الآتي: هل ثمة تشابه معجمي بين الوحدات المعجمية في العربيتين التراثية والمعاصرة؟ وتنطلق من مفهوم الوحدة المعجمية lexical unit في النظرية المعجمية الحديثة التي ترى أن الوحدة المعجمية قد تكون كلمة مفردة، نحو: كتاب، أو جزءًا من كلمة، نحو: عبد، أو سلسلة غير مفصولة من الكلمات، نحو: فأسقيناكموه (ابن مراد، 2006).

2. الأعمال السابقة

يستعمل الباحثون عددا من الخوارزميات التي تقيس التشابه بين اللغات ذات الفصيلة الواحدة، أو بين اللغات واللهجات، أو بين اللهجات. ويمكن تصنيف هذه الخوارزميات حسب التمثيل اللغوي المراد بحثه. فهناك الخوارزميات التي تعمل على مستوى الحرف character-based similarity، وهناك ما يعمل منها على مستوى الكلمة أو النص term-based similarity.

وتعمل مقاييس التشابه المعجمي على تتابع من السلاسل الحرفية على مستوى الحرف، وعلى مدونات نصية على مستوى الكلمة. ويوضح الجدول (1) أبرز الخوارزميات المستعملة حسب التمثيل اللغوي الهدف، وإن كانت الدراسات التي أجريت على العربية واستعمالاتها قليلة في هذا المجال.

جدول 1

الخوارزميات الأكثر شيوعا لقياس أوجه التشابه بين النصوص

المستوى	الخوارزمية	الوصف
الحرف	Longest Common SubString النص المشترك الأطول	تحسب أطول تسلسل جزئي مشترك بين نصين، نحو: "ون"، في النصين: "مستنيرون" و"قادمون" (Manning, et. al.، 2008).
	Levenshtein distance مسافة ليفنشتاين	تحسب الحد الأدنى من عمليات التعديل اللازم تطبيقها على أحد النصين، سواء كانت حذفا أو استبدالا أو إدخالا؛ لتحويل النص إلى النص الآخر (Levenshtein، 1965).
	N-gram models نماذج التتابع	تحسب التشابه والاختلاف بين نصين بتقسيم عدد N-gram المشتركة بين النصين على عدد N-gram الكلي في كلا النصين (Kondrak، 2005).
	Dynamic programming البرمجة الديناميكية	تقيس مقدار التشابه بين نصين بإيجاد أفضل تحذية تغطي كامل النصين أو جزءا منها (Needleman & Wunsch، 1970؛ Smith & Waterman، 1981).
الكلمة أو النص	Vector space models نماذج التمثيل الشعاعي	يحسب بتمثيل النص في فضاء يكون لكل كلمة فيه بعد تابع لعدد مرات تكرار الكلمة في النص (Kumar, et. al. ، 2012).
	Cosine Similarity تشابه جيب التمام	تقيس زاوية جيب التمام بوصفها مؤشرا للتشابه بين متجهين (Bhattacharyya، 1946).
	Divergence Distance مسافة الاختلاف	تقيس الاختلاف بين التوزيعات الاحتمالية، ومنها: Kullback-Leibler distance، وHellinger، وManhattan distance، وغيرها (Ali & Silvey، 1966).
	Jaccard similarity تشابه جاكارد	تقيس مقدار التشابه بين نصين، بتقسيم عدد العناصر المتشابهة على عدد الكلمات الفريدة فيهما (Niwattanakulm et. al.، 2013).
	Latent Semantic Indexing فهرسة الالفاظ الدلالية	تقيس الكلمات المتقاربة في المعنى، والمتكررة في مواضع متشابهة في النص (Landauer, et. al.، 1998).

وقد كشفت دراسة المجيول Almujaiwel (2020)، باستعمال عدد من الأساليب التحليلية، عن وجود اختلاف معجمي دال إحصائيا بين المعاجم القديمة والحديثة، فضلا عن حسابها لحجم الاختلاف بين المعاجم العربية الحديثة، ومدونة arTenTen العربية[1]. وقد خلصت الدراسة إلى بعد المعاجم العربية كثيرا عن واقع الاستعمال الطبيعي للغة العربية. وفصل

الباحث نقاط الاختلاف بين ثلاث مراحل. أظهر في المرحلة الأولى التشابه الكبير بين المعاجم العربية المعاصرة والمعاجم العربية القديمة، مستنتجا من هذا التوضيح عدم وجود دلالة على تطور المعاجم العربية المعاصرة، وعدم وجود ما يشير إلى مداخل معجمية جديدة، تتلاءم مع طبيعة ظهور كم هائل من الكلمات العربية المولدة والجديدة في العصر الحديث. وكشف الباحث في المرحلة الثانية عن مدى تطابق الكلمات العربية واختلافها بين المعاجم العربية، ومليون كلمة معاصرة (قائمة كلمات arTenTen بدون تكرار)، كما كشف عن مدى ذاك التطابق والاختلاف بين المعاجم العربية المعاصرة والاستعمال العربي المعاصر المتمثل في قائمة مدونة arTenTen، وخلص إلى أن الاختلاف شبه كامل بينهما وبنسبة قدرت بـ 0.99. ولمح الباحث إلى أن هذه المعاجم العربية المعاصرة لم تستعن بالاستعمال اللغوي المعاصر، بقدر استعانتها بالمداخل المعجمية في المعاجم العربية القديمة. كما لمح إلى أن مقارنة المعاجم العربية المعاصرة بأي وعاء لغوي عربي معاصر متعدد السجلات كالعربية الاقتصادية والسياسية والصناعية، سيظهر ما يقارب هذه النسبة من الاختلاف. وأوصى بأهمية تنويع المداخل المعجمية في أي معجم عربي معاصر بأكثر الكلمات ذات المحتوى تكرارا من كل مجال وسجل وعاء لغوي (على سبيل المثال: الكتالوجات الصناعية والتسويقية).

وتختلف هذه الورقة عن دراسة كويك وآخرين .Kwaik et. al (2018) التي نظرت في العلاقة المعجمية بين العربية الفصحى الحديثة واللهجات العربية في أكثر من منطقة عربية. وهي تركز على قياس المسافة المعجمية بين الفصحى الحديثة واللهجات العربية باستعمال تقنيات وأدوات معالجة اللغات الطبيعية والمدونات النصية. وقد أُجريت الدراسة بقياس التداخل وأوجه التشابه والاختلاف بين العربية الفصحى الحديثة واللهجات العربية باستخدام مصفوفات الاختلاف المختلفة نحو: نموذج التمثيل الشعاعي(VSM) Vector Space Model القائم على توزيع الكلمات في البيانات، ونموذج الفهرسة الدلالية الكامن(LSI) Latent Semantic Indexing القائم على تحليل النصوص من أجل تمثيل المفاهيم في البيانات، ومسافة هلينقر (HD) Hellinger Distance التي تقيس الفرق بين توزيعين احتماليين. كما قاست الدراسة التداخل بين مفردات جميع اللهجات، مستعملة جاكارد، ومعامل الارتباط. وقد اعتمدت على عدد من المدونات، هي: مدونة اللهجات العربية المتوازية (PADIC) Parallel Arabic Dialect Corpus، والمدونة المتعددة اللهجات Multi-Dialect corpus، ومدونة اللهجة الشامية (SDC) Shami Dialect Corpus، ومدونة ملفات تحرير نصوص ويكيبيديا Wiki-Docs. وهي تضم العربية المعاصرة، واللهجات الشامية، واللهجة المصرية، ولهجات شمال أفريقيا. وقد دللت أغلب المقاييس المستعملة على أن اللهجات الشامية هي أقرب للعربية الحديثة، فيما كانت لهجات شمال أفريقيا أبعدها عنها. وبينت الدراسة أيضا درجة التشابه والتداخل اللغوي بين اللهجات الشامية الذي يكاد لا يظهر فيه التمييز على المستوى الكتابي دون وجود معلومات فونولوجية أو علامات تشكيل.

أما دراسة حرات وآخرين .Harrat et. al (2015) فهي تختلف أيضا عن أهداف هذه الورقة، وتشبه سابقتها في دراستها للهجات عربية، ومقارنتها باللغة العربية الفصحى الحديثة. وتقدم هذه الدراسة مدونة اللهجات العربية المتوازية PADIC التي تتضمن خمس لهجات: لهجتين جزائريتين من مدينتي الجزائر وعنابة، واللهجة التونسية، واللهجة الشامية الفلسطينية، واللهجة السورية. وقد قدم الباحثون دراسة تحليلية لغوية لتلك المدونة، باستعمال المقارنات بين كل لهجة، والعربية الفصحى الحديثة. وخلال ذلك كشفوا عن الكلمات الأكثر شيوعا في كل لهجة دون إشارة لأسباب عدم حذفهم للكلمات الوظيفية التي غالبا ما تختلف بين كل لهجة ولهجة، وحسبوا النسبة المئوية للوحدات المعجمية المشتركة على مستوى النص

الواحد، وعلى مستوى الجملة؛ للتأكيد على وجود علاقة بين اللهجات، والعربية الفصحى الحديثة، كما يظهر الجدول (2). كما قاسوا الاختلاف اللغوي باستعمال خوارزمية هلينقر Hellinger، للكشف عن اللهجة الأقرب للعربية الفصحى الحديثة. وأظهرت النتائج أن اللهجة الفلسطينية هي أقرب اللهجات إلى العربية الفصحى الحديثة، وتليها اللهجة التونسية والسورية، فيما جاءت اللهجتان الجزائريتان الأكثر بعدا، كما يظهر في الجدول (3). وقد توقع الباحثون هذه النتائج مستدلين بقرب التونسيين للجزائريين جغرافيا أكثر من الفلسطينيين والسوريين. كما أن مقاييس الاختلاف تشير إلى أن اللهجات الجزائرية متقاربة فيما بينها، وكذلك اللهجات الشامية على الجانب الآخر. ومما عابه الباحثون على نصوص الدراسة ترجمتها يدويا من المحادثات الجزائرية إلى العربية الفصحى واللهجات الأخرى من قبل متحدث واحد في كل لهجة، مما قد ينتج عنه بعض الانحياز في الدراسة.

جدول 2

نسب التشابه بين اللهجات في الكلمات الأكثر شيوعا

	نسب التشابه المئوية				
المدونات	اللهجة الجزائرية	اللهجة العنابية	اللهجة التونسية	اللهجة السورية	اللهجة الفلسطينية
اللهجة الجزائرية	_	73.62	35.43	24.16	25.43
اللهجة العنابية	72.86	_	34.25	23.59	25.00
اللهجة التونسية	31.10	30.38	_	29.79	33.49
اللهجة السورية	21.01	20.73	29.52	_	44.00
اللهجة الفلسطينية	24.79	24.63	37.2	49.33	_

جدول 3

نسب تشابه اللهجات بالعربية الفصحى الحديثة في الكلمات الأكثر شيوعا

اللهجة	اللهجة الجزائرية	اللهجة العنابية	اللهجة التونسية	اللهجة السورية	اللهجة الفلسطينية
النسبة المئوية	21.18	21.07	37.6	37.36	51.68

وفي دراسة أخرى، يقارن أبو نصر Abu Nasser (2015) بين اللهجة الخليجية، واللهجة الشامية، واللهجة المصرية، واللهجة المغربية، ولم يكن دقيقا عندما ضم العربية الحديثة إلى هذه اللهجات باعتبارها لهجة لا لغة. وقد كانت مقارنته بين هذه التنوعات اللغوية بالوقوف على الاختلافات على مستوى الكلمة ومستوى الصوت. وكان اعتماده على قائمة سوادش Swadesh ومفهوم الكلمات غير المتشابهة في الأصل لقياس مقدار الاختلافات اللغوية بين تلك التنوعات اللغوية. وحيث إن قائمة سوادش قائمة صوتية، لا معجمية، فقد جمع الباحث البيانات من متحدثين اثنين من جنس الذكور في كل لهجة. وعُدلت قائمة سوادش مع قائمة من العربية الحديثة بالاعتماد على قاموسين حديثين ثنائيي اللغة (عربي- إنجليزي)، وهما: المورد والقاموس الحديث. وقد زُودت كل كلمة بجملة ترد فيها الكلمة في سياقها؛ لاستبعاد حدوث الغموض المعجمي. ومن ثم قِيس

الاختلاف بين التنوعات اللغوية على أساس نسبة الكلمات غير المتشابهة في الأصل في قائمة سوادش لكلمات العربية الفصحى الحديثة، بالإضافة إلى استعمال خوارزمية لفينشتاين Levenshtein لقياس الاختلاف بين المواد المعجمية على المستوى الصوتي، بالاعتماد على النظام الكتابي العالمي IPA لكلمات قائمة سوادش. وقد خلص الباحث إلى أن أقرب اللهجات إلى العربية الفصحى الحديثة هي اللهجة الخليجية واللهجة الشامية، وتليها اللهجة المصرية، فيما تأتي اللهجة المغربية أبعدها عن العربية الفصحى الحديثة. ولعل من أهم الحدود المهمة في هذه الدراسة هو الكيفية التي جمعت بها البيانات، حيث إن المتحدثين وجنسهم ومكانهم الجغرافي محدود بمتحدثين اثنين من الذكور لكل لهجة، والقاموسان الحديثان المستعملان في الدراسة لترجمة قائمة سوادش لمقابلاتها العربية الفصحى الحديثة قد أُلِّفا من قبل مؤلفَيْن شاميين، مما قد يجعل أن ثمة انحيازًا في العربية الحديثة للهجة الشامية. ومع أن الهدف كان هو قياس الاختلاف المعجمي، استعملت الدراسة التمثيل الصوتي للكلمات الأمر الذي قد يكشف أيضا اختلافات خفية غير معجمية.

3. خوارزمية جاكارد Jaccard

من أهم التطورات التي يشهدها مجال البحث العلمي في معالجة اللغة العربية، استعمال خوارزميات تعلم الآلة في مجال التنقيب في النصوص العربية. وتعد خوارزمية جاكارد Jaccard إحدى خوارزميات التشابه التي تعمل على بيانات غير موسمة unlabeled، وتقيس أوجه التشابه المعجمي بين النصوص. ولا تعتمد هذه الخوارزمية على مصفوفة كلمات المستند -term document matrix، بل تقيس التشابه بين مجموعتين قياسا مباشرا (Aggarwal & Zhai، 2012، ص 80)، فتكشف عن التطابق السَّلسَلي للوحدة المعجمية الواحدة التي تنفصل عن الوحدة المعجمية السابقة واللاحقة بمسافة واحدة في النظم الخطّي الكتابي.. وتُعرّف بأنها حجم التقاطع مقسومًا على حجم الاتحاد بين المجموعتين. وتحسب بالصيغة الآتية:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

حيث A تشير إلى المجموعة النصية الأولى، فيما تشير B إلى المجموعة النصية الثانية. ويشير الرمز ∩ إلى ما تتشارك فيه المجموعتان من الكلمات، وهو المقصود بحجم التقاطع. أما الرمز ∪فيشير إلى مجموع الكلمات في المجموعتين دون تكرار (أي جميع ما ورد في المجموعة الأولى، مع جميع ما ورد في المجموعة الثانية على ألا يرد ما تشاركتا فيه أكثر من مرة)، وهو المقصود بحجم الاتحاد بين النصين.

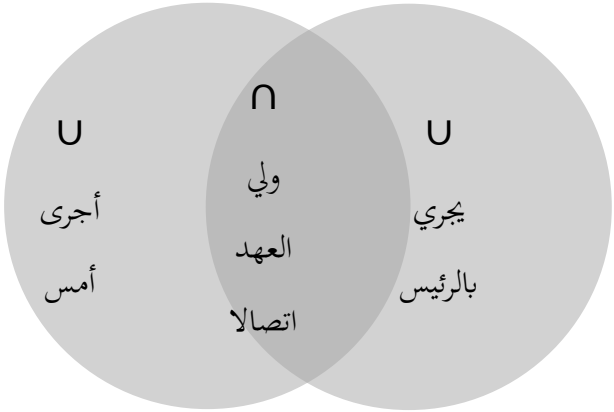
وهكذا يُستعمل جاكارد في تحديد التشابه بين نصين أو مدونتين، ويقيس مدى تشابههما واختلافهما وفق تصنيف عناصرهما المراد قياس الشبه بينها، فقد يقيس الشبه بين الكلمات، أو بين العبارات، أو بين الجمل. وتعني قيمة جاكارد (1) أن التداخل بين الكلمات حدث في كل الكلمات، فيما تعني القيمة (صفر) عدم وجود كلمات مشتركة بين النصين. وهذا يدل على أن القيمة كلما زادت كانت نسبة التشابه أعلى، وكلما انخفضت اختلف النصان. والشكل (1) يوضح ببساطة فكرة الخوارزمية الرياضية، للجملتين الآتيتين:

الجملة A: |ولي العهد يجري اتصالا بالرئيس|

الجملة B: |أجرى ولي العهد اتصالا أمس|

شكل 1

مخطط فِن Venn لمفهوم خوارزمية جاكارد



إذن سيكون معامل جاكارد للجملتين (A, B)

$$\frac{\{'ولي','العهد','يجري','اتصالا','بالرئيس'\}\cap\{'أجرى','ولي','العهد','اتصالا','أمس'\}}{\{'ولي','العهد','يجري','اتصالا','بالرئيس'\}\cup\{'أجرى','ولي','العهد','اتصالا','أمس'\}}=$$

$$\frac{\{'ولي','العهد','اتصالا'\}}{\{'ولي','العهد','يجري','أجرى','اتصالا','بالرئيس','أمس'\}}=$$

$$0.428 = \frac{3}{7}$$

ويشير معامل جاكارد 0.428 إلى أن الجملتين متشابهتان بنسبة 42.8%، ومختلفتان بنسبة 57.2%.

وقد استعملت دراسة مركز اعتدال (2020) مقياس جاكارد في التحليل المعجمي لخطاب داعش. وكشفت به عن مدى التشابه والاختلاف بين خطب مرئية ومسموعة نشرت بين عامي 2014-2019 لقادة داعش، من أجل فهم أدق وأعمق للتوجهات المتطرفة التي يتشارك بها قادة هذه الجماعة. كما استعرضت دراسة سانقر وواورين Sanger & Warin (2019) نتائج مؤشرات التشابه والاختلاف التي استعملا فيها جاكارد على 1597 بيانًا سياسيًّا من 27 دولة أوروبية منذ عام 1945. وهي دراسة تقع في مجال علم السياسة المقارن. وطبقت باستعمال لغة الآر R.

4. البيانات: التحليل والنتائج

4. 1. التهيئة

جرى جمع النصوص انطلاقا من رؤية فرستيغ (2003) في تقسيمه لعصور العربية الفصحى[2]، فجعلتها على امتداد أزمنتها تقع في نوعين. يتضمن النوع الأول النصوص التراثية التي يمتد زمنها من العصر الجاهلي، وحتى عام 1214هـ والنوع الثاني يتضمن

النصوص المعاصرة التي يمتد تاريخها من عام 1215هـ وحتى الآن. وقد كانت آلية جمعي للنصوص غير منظمة، لكنها تراعي زمنيا هذا التقسيم، فضلا عن شموليتها لأوعية ومجالات متعددة للعربية، واستبعاد ما هو مقتبس في النصوص المعاصرة من النصوص التراثية، نحو: شروح المعلقات، وكتب الحديث والتفسير والفقه، واللغة والخطب والمقالات الدينية. واستفدت في ذلك من المحتوى العربي المتاح، وخصوصا المكتبة الشاملة، ومدونات صحفية متنوعة الأزمنة والأمكنة والموضوعات، ورسائل جامعية من صنوف العلوم والمعارف، وروايات عربية من مختلف الدول العربية. وقد راعيت ألا يزيد حجم كل نوع نصي عن 1 قيقا؛ ليسهل التعامل معه حسب الإمكانات الحاسوبية المتاحة، فكانت بتكراراتها ما يقارب 150 مليون كلمة في النوع الأول، و120 مليون في النوع الثاني.

لقد كانت النصوص المجموعة ذات صيغ وترميزات مختلفة، بسبب تنوع المصادر والأصول. وحيث إن الأدوات المستعملة تتطلب مني التعامل مع نوعين من الترميز، حولت جميع الملفات إلى صيغة txt بالترميز ANSI وUTF، إذ يتطلب البايثون استعمال ترميز ANSI، فيما يتطلب غواص والمشذب العربي الصيغة UTF. فأصبحت لدي مجموعة ملفات من النصوص التراثية، ومجموعة ملفات من النصوص المعاصرة. وقد استعملت لمعالجة نصوصها المشذب العربي، لحذف الرموز، والأرقام، والتشكيل، وعلامات الترقيم، والكلمات، والحروف غير العربية، وعلامات التطويل، والمسافات المتكررة. كما استعملت أداة غواص للحصول على قائمة التكرارات لكلمات كل نوع تمهيدا لتنقيحها وفحصها، فلم يكن الهدف من الحصول على تكراراتها إلا تسهيل مراجعة قوائم الكلمات في كل نوع بتجميعها، والاستفادة من خصائصها في التنقيح. وبمراجعة لقائمة الكلمات في العربيتين، عثرت على كلمات لم ترد إلا مرة واحدة، ومعظمها كان لكلمات فارسية أو كردية، وكلمات وجمل وفقرات متصلة بدون مسافات. فاستعنت بخواص تطبيق نوتباد++، لحذف أي كلمة تتضمن الحروف الفارسية والكردية، كما استعنت بدالة الأكسل ()len الخاصة بأطوال الكلمات، لحذف الكلمات والجمل والفقرات التي لم تفصل بالمسافات بتحديد ما يفوق طوله عن أطوال الكلمات العربية. وقد كانت أطول الكلمات العربية السليمة طباعيا في العربية القديمة ذات 14 حرفا، ومنها: والإسكندرانيون، وأطولها في العربية الحديثة كلمات ذات 17 حرفا، ومنها: البارابسيكيولوجيا. فما زاد طوله عن هذين الطولين حذفته، ثم راجعت القوائم مرة أخرى لحذف ما هو أقل من هذا الطول، ولم يفصل فيه بين الكلمات بمسافة، نحو: عبدالله، غيرأنها، علىأنها ... إلخ. وبعد المعالجة والتنقيح وحذف التكرارات، أبقيت على 1,080,977 كلمة تماما في كل مجموعة. ثم دمجت ملفات كل نوع في ملف مستقل خاص بها بالترميز ANSI تمهيدا للتحليل الكمي، انظر الجدول (4).

جدول 4

إحصاءات نصوص العربية التراثية، والعربية المعاصرة

	عدد الكلمات الفعلي	عدد الكلمات النوعي	عدد الكلمات بعد المعالجة
العربية التراثية	146,355,435	2,825,225	1,080,977
العربية المعاصرة	119,500,938	3,614,672	1,080,977

4. 2. التحليل الكمي

من أجل الإجابة عن سؤال الدراسة، كانت الخوارزمية المناسبة لطبيعة البيانات الخام وحجمها هي خوارزمية جاكارد. وتعتمد خوارزمية جاكارد على متغيرين هما: intersection، وunion، حيث يشير الأول إلى الكلمات التي تشترك فيها المجموعتان، ويرمز له رياضيا بالرمز ∩، فيما يشير الثاني إلى الكلمات التي تؤلف المجموعتين من دون تكرار، ويرمز له بالرمز ∪. وباستعمال البايثون نفذت خوارزمية جاكارد بالكود الموضح في الشكل (2) على الملفين، فاستخرجت قيمة هذين المتغيرين، وأظهرت النتيجة تشارك العربية التراثية والعربية المعاصرة في 511,122 كلمة فقط، فيما تميزت كل عربية بـ 569,855 كلمة. وهذا يعني أن قيمة ∩في العربيتين 511,122، وقيمة ∪ 1,650,832 كلمة، أي مجموع الكلمات في العربيتين دون تكراراتها. أما معامل التشابه جاكارد، فقد أشار إلى وجود تشابه بين العربيتين درجته 0,31. كما يشير إلى وجود اختلاف معجمي بين العربية التراثية، والعربية المعاصرة بما قدره 0,69. وهذا يعني أن حجم الاختلاف أكثر من حجم التشابه بما يزيد عن الضِعف.

ومن خلال مخطط فن Venn في الشكل (3)، يتضح مقدار هذا التشابه المتمثل في المنطقة البيضاء بين الدائرتين، والمقدر بدرجة 0,31، كما يتضح الاختلاف المتمثل في الجزء الرمادي في كلا العربيتين، والمقدر بدرجة 0,69.

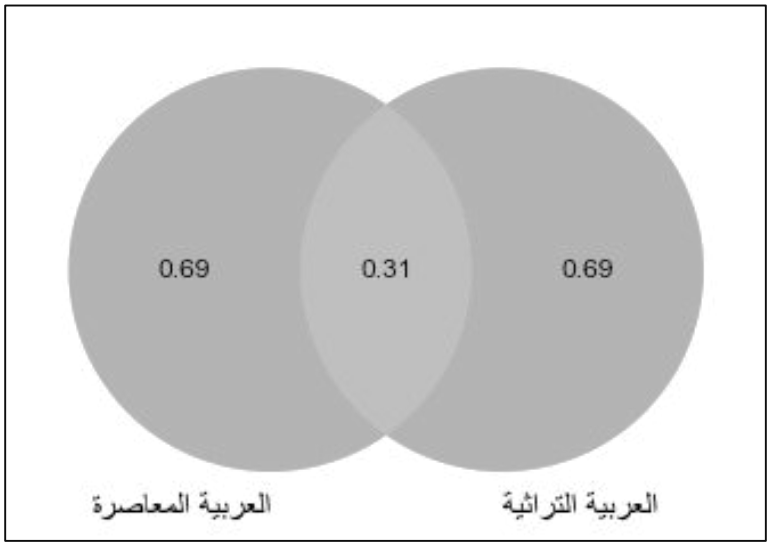
شكل 2

كود خوارزمية جاكارد بلغة البايثون

```
oldArabic = open('C:\\...\\oldArabic.txt')
modernArabic = open('C:\\..modernArabic.txt')
oldArabic = oldArabic.read()
modernArabic = modernArabic.read()
def Jaccard_Similarity(oldArabic, modernArabic):
    words_doc1 = set(oldArabic.lower().split())
    words_doc2 = set(modernArabic.lower().split())
    intersection = words_doc1.intersection(words_doc2)
    union = words_doc1.union(words_doc2)
    print(len(intersection))
    print(len(union))
    return float(len(intersection)) / len(union)
Jaccard_Similarity(oldArabic, modernArabic)
```

شكل 3

مخطط فن Venn لتمثيل جاكارد بين العربية التراثية والعربية المعاصرة



4. 3. التحليل النوعي

لقد كشف لنا تطبيق خوارزمية جاكارد على بيانات الورقة الخام ثلاث قوائم كلمات، هي:

أ- قائمة الكلمات التي تنتمي للعربية التراثية فقط، وهي 569,855 كلمة، انظر الجدول (5).
ب- قائمة الكلمات التي تنتمي للعربية المعاصرة فقط، وهي 569,855 كلمة، انظر الجدول (6).
ج- قائمة الكلمات التي تشترك فيها العربيتان، وهي 511,122 كلمة، انظر الجدول (7).

ويبدو من ملاحظة بعض الأمثلة في تلك القوائم، أن ثمة ظواهر لغوية كانت من أسباب هذا التشابه، أو هذا الاختلاف المعجميَّيْن بين العربيتين. وهي أسباب قد ينفيها أو يثبتها البحث اليدوي المعمق في هذه القوائم التي يتعذر الكشف عن بعض خصائصها آليا بأدوات التحليل اللغوية المتوفرة، كالمحللات الصرفية، والموسمات بأقسام الكلام، بسبب غياب التشكيل، والسياق، والإملاء الصحيح أحيانا. ومع ذلك، حاولت الوقوف من خلال بعض الأمثلة على بعض منها، كان قد بدا واضحا منذ معالجة البيانات، وحتى مرحلة استخلاص القوائم، وهي كما يأتي:

أ- يظهر الاعتناء من جانب العربية التراثية بمواقع الهمزات الصحيحة، فيما تكثر في العربية المعاصرة الكلمات التي لا تضع الهمزة في الموقع الصحيح، فتؤثر في حجم الاختلاف عن العربية التراثية. ومن تلك الأمثلة: (إبليسك، ابليسك) و(باطراحهم، بإطراحهم)، كما يظهر في الجدولين (5) و(6).

جدول 5

بعض الكلمات المختارة من قائمة الكلمات التي تنتمي للعربية التراثية فقط

تنكسا	خاضبي	تعوطت	امتحاق	كرمانيها	معتلقي	متربيا	والمسترقي
تنكسفان	فتواطأنا	رهصناه	امتحاش	معتلفة	يمانهم	متذهبا	باطراحهم
تنكسه	مثبوتها	هزلين	ثنيتين	عنكباء	طاجنة	امتحاص	دارعات
تنكسها	المعاظم	الأبهجان	اقتقطوا	ففضضتهن	يمانهم	امتحاش	وضيفان
تنكسهم	والتحثيث	واللائث	الحرقفتين	الظليف	الإمبرزيانا	الإمبروزيانا	اختلعها
تنكشحن	المعاظم	واللئمة	مسوديها	استسفهت	والقبيقول	ضميلة	قلقمان
بغتهم	التخطف	واللائب	تأهبن	تغنمين	فتوعيده	شعانينه	متراشق
تنكشفون	المنعمم	خنسير	تأنوا	معتمره	المتضايف	فعتقتا	إبليسك

جدول 6

بعض الكلمات المختارة من قائمة الكلمات التي تنتمي للعربية الحديثة فقط

المتبسمة	كشكول	وسنوليه	للملاريا
استرتش	وشعوبيته	الكارتيه	أرينبورغ
يهملني	بالمراكز	احترستم	وبغاياكم
لمقترحه	الأخروفي	الكارتون	للإقليمية
فيثيرها	الشيكية	الآسيوى	جوازئها
بكيلو	لكهربائي	بالأخونة	بافاداتهم
بكينج	وإمتاعه	الغزلتان	ومترتباتها
يموكي	مدرساتي	بالمراقص	الآسيوي
الإلكتروميكانيكي	سوتيونيوس	تصدقينى	سيختلق
دوسلوروف	منشوراته	لحضاراتهم	سيختلف
البتروكيماويات	بالترسبات	بالأدبين	تصدقيهم
والفضوليون	بالترسانة	بإطراحهم	بمنظماتهم
الانتركونتيننتال	ستستغرب	ستستغرق	والعابرة
البتركيماويات	المتبسطة	تصدقيها	مسافرتان
والكوليكالسيفيرول	لأكاديميات	الكارتيل	جارندنيير
السوسيوديمغرافية	الثانوى	إستيعابها	ابليسك

جدول 7

بعض الكلمات المختارة من قائمة الكلمات التي تشترك فيها العربيتين

موصوفا	ودوام	ليدري	ويبذلوا	الإسرائيليات	كنبراس	تخويفها	تأرز
وعابد	فاعتجر	لتلك	قادران	الاسترابادي	بالخلاخيل	بالإغاثة	والاحتشام
وطغت	الرضاعية	موازنه	بنيات	الاسترابادي	فعليته	بالمحراث	تبجحت
يكتموه	يتخول	غازيهم	أنجدت	الإسرائيليات	البركة	أتعقل	المعصومة
بمشهده	رأوك	البجر	إرزي	يخزنا	خارق	والعيزارة	وكرس
شوكك	وحجبته	نايم	أبصراه	إلاك	مفؤود	والصحاف	فكلامنا
وشغلا	فيفرضون	واقتيد	كيفك	السقاط	وانخفاضها	عراصهم	لعقدي
ودوام	كالرعد	تتمرغ	كارها	ولصالح	لربكم	كنثيل	الثانوي

ب- ثمة خلط في العربية المعاصرة يؤثر في حجم الاختلاف والتشابه معا. إذ ينتج عنه تكرار ورود الكلمة نفسها بأكثر من شكل في العربية المعاصرة، فتتطابق مع كلمات أخرى في العربية التراثية ربما لا تكون ذات صلة بها. هذا الخلط يظهر في الأسلوب الكتابي، ولعل أوضحه الخلط بين الياء والألف المقصورة، ومنشؤه النصوص المصرية التي تنحاز للكتابة العربية الأصولية. وقد ترتب على ذلك محاولة تنقيط الألف المقصورة (ى) بعد انتشار الحواسيب، وتخصيص محرف للياء، ومحرف للألف المقصورة في لوحة المفاتيح، ومن ثم تنقيط الألف المقصورة في مواضع ليست صحيحة. فنجد مثلا كلمة: (الثانوى) في العربية المعاصرة (انظر الجدول (6))، وهي كلمة تشترك بها مع العربية التراثية بشكلها الكتابي (الثانوي)، فوجودها -وأمثالها- في العربية المعاصرة زاد من حجم الاختلاف بين العربيتين، رغم تشاركهما إياها (انظر الجدول (7)).

ج- يلاحظ على المستوى المعجمي في قائمتي الاختلاف ظهور نفس الكلمة بإملاء مختلف في القائمة الواحدة، نحو: (الإمبرزيانا، الإمبروزيانا) في العربية التراثية (انظر الجدول 5)، ونحو: (البتروكيماويات، البتركيماويات) في العربية المعاصرة (انظر الجدول 6). وقد نقلل حجم الاختلاف بين المجموعتين بمعالجة الأعلام، والمصطلحات، وغيرها من الأسماء ذات الدلالات الخاصة، والمكتوبة عادة بأساليب إملائية غير ثابتة، كما في المثالين السابقين. ومن الجانب الآخر، نجد في قائمة الكلمات المشتركة (الجدول 7) بعض أسماء الأعلام وتصريفاتها قد وردت بأكثر من شكل إملائي، ومن ذلك: (الاسترابادي-الاسترابادي)، و(الإسرائيليات-الإسرائيليات)، والاعتناء بمعالجتها سيقلل من حجم التشابه أيضا.

د- تتشابه العربية التراثية مع العربية المعاصرة في استعمال بعض المورفيمات الصرفية استعمالا يخالف الاستعمال المعياري لها في العربية، وذلك بتحويلها من مورفيمات حرة إلى مورفيمات مقيدة، نحو كاف الخطاب في: (إلاك -كيفك)، إذ في القياس يقال: (إلا إياك -كيف أنت)، بفصل مورفيم الضمير عن الأداتين، انظر الجدول (6).

وبنظرة عامة في القوائم الكاملة التي استخلصتها الخوارزمية من بياناتنا، لوحظ ما يأتي:

Æ زيادة الانتشار الصرفي لبعض الجذور دون غيرها في قائمة الكلمات المشتركة كاملة. فالانتشار الصرفي فيها للجذور: (علم- سلم - صدق) يفوق الانتشار الصرفي لجذور، مثل: (درس – فرح – حزن). ويبين الجدول (8) حجم هذا الانتشار الصرفي في قائمة الكلمات المشتركة، والقائمتين الأخريين بالاستعانة بخاصية البحث في برنامج الإكسل. ويلاحظ أيضا أن حجم الانتشار في قائمة الكلمات المشتركة متسق مع حجمه في القائمتين الخاصة بكل نوع. أي أن هذا الانتشار في الجذور نفسها تتشارك فيه كل القوائم بالزيادة والنقصان نفسه، فيما عدا جذر (درس) الذي بدا في العربية المعاصرة أكثر، وتلك ظاهرة جديرة بالدراسة والبحث في العربية المعاصرة.

جدول 8

حجم الانتشار الصرفي لبعض الجذور في العربية

الجذر	حجم انتشاره الصرفي في قائمة:		
	الكلمات المشتركة	العربية التراثية	العربية المعاصرة
علم	914	521	687
سلم	561	500	602
صدق	478	397	402
درس	193	106	393
فرح	209	109	226
حزن	167	85	121

ب- ساهم التنوع في أوعية ومجالات وموضوعات العربية المعاصرة في وجود الاختلاف الكبير بين العربيتين، فالعربية التراثية مثلا تخلو من الصحف، ومن مجالات وموضوعات التقنية، وهي في المقابل تزخر بالمحتوى الديني واللغوي أكثر من المحتوى المعاصر، وما ورد منه في المعاصر معظمه مقتبس منها. ففي الجدول (6) تظهر مصطلحات عديدة ذات صلة بالعلوم الحديثة، نحو: (البتركيماويات، السوسيوديمغرافية)، فيما تغيب مثل هذه المصطلحات عن العربية التراثية، ونجد مصطلح دينيا وحيدا، وهو (الإسرائيليات) بين ما اقتبس من قائمة ما تشترك فيه العربيتان (انظر الجدول (7)).

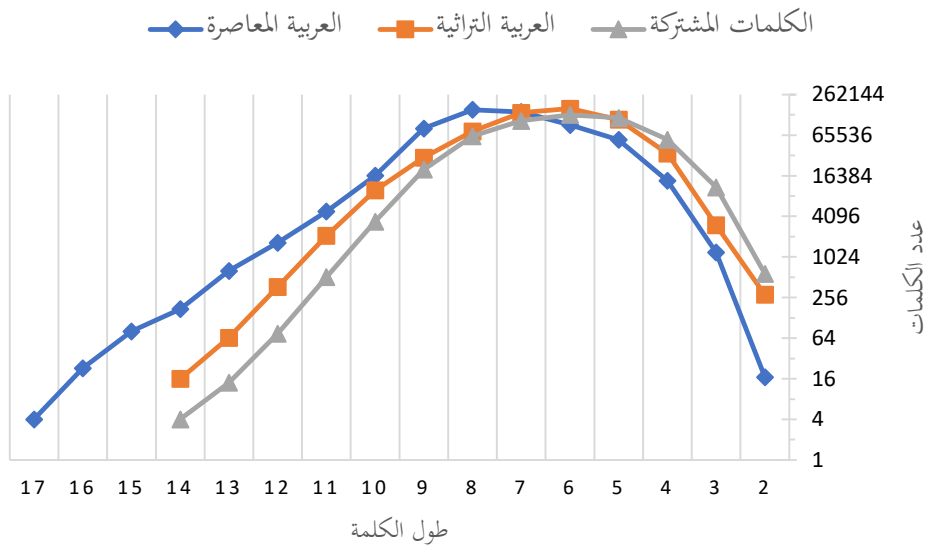
ج- رغم أن النقحرة Transliteration (النقل الصوتي للكلمات من لغة إلى لغة) ممارسةٌ قديمةٌ في العربية موجودةٌ في القرآن الكريم الذي استعار الكلمات الآرامية والحبشية والعبرية والفارسية والسريانية والرومية واليونانية والهندية... (السيوطي، ت. 911 هـ، ط. 2006، 3/934-974)، وبحثها اللغويون القدماء (انظر مثلا: ابن خلدون، ت. 808 هـ، ط. 1979، 1/811-816)، إلا أن وجودها في العربية المعاصرة أكثر، وفي أعلام ومصطلحات أطول، وبأساليب كتابية متعددة، انظر الجدول (6). ولا شك أن التنوع المعرفي الذي يزيد مع امتداد العربية زمنيا ومكانيا، وحاجة العربية للتكيف مع مصطلحات المعارف الأدبية والعلمية والتقنية، فضلا عن أسماء الأعلام الأجنبية الحديثة، كان له دور

بارز في الاختلاف بين العربيتين. فعلى الرغم من الجهود المبذولة في بناء القواميس المتخصصة للمصطلحات العلمية إلا أن التسارع في التطور العلمي، وقلة الوعي بأهمية توحيد كتابة المصطلح ساهما في زيادة حجم الاختلاف، ولعل مثالنا السابق (البتروكيماويات، البتركيماويات) شاهد على ذلك.

د- باستخلاص أطوال الكلمات في القوائم الثلاثة ينكشف لنا سبب آخر للاختلاف. لقد تبين أن العربية المعاصرة تضمنت كلمات بلغت في طولها 17 حرفا، فيما لم تزد الكلمات العربية التراثية عن 14 حرفا. غير أننا نشير هنا إلى أن هذه الزيادة في الطول لم تكن في صيغ صرفية معينة تطورت من صيغ كانت في العربية التراثية، أو في مورفيمات صرفية مستحدثة. إنما جاءت جميعُ هذه الكلمات التي تميزت بها العربية المعاصرة مصطلحاتٍ منقحرة وأعلامًا لا وزن لها، ومصادر صناعية منها، نحو: (البارابسيكيولوجيا، والسوسيواقتصادية، والأرسطوطاليسيون). ويظهر الشكل (4) بالخط الرمادي أن ما اشتركت فيه العربية التراثية مع المعاصرة لم يكن إلا فيما لم يتجاوز طوله 14 حرفا من الكلمات، وأن ما تجاوز الـ 14 حرفًا كان عددًا قليلًا قياسًا بالأطوال الأخرى من الكلمات. وهذا يؤكد أن الاختلاف في العربيتين لم يكن في الأعلام والمصطلحات المنقحرة، وحسب.

شكل 4

أطوال الكلمات بين العربية التراثية والعربية المعاصرة



5. الخاتمة

قدمت هذه الورقة نموذجا إحصائيا يتعلق بمدى التشابه والاختلاف المعجمي بين العربية التراثية والعربية المعاصرة. وأجابت عن السؤال: هل العربية التراثية تتشابه في وحداتها المعجمية مع العربية المعاصرة؟ فقد أظهرت النتائج بعد التجريب على مدونة

عربية تراثية ضمت 146,355,435 كلمة، ومدونة عربية معاصرة ضمت 119,500,938 كلمة، أن العربية التراثية والعربية المعاصرة تتشاركان في 511,122 كلمة فقط، فيما تميزت كل عربية بـ 569,855 كلمة. وأشار معامل التشابه جاكارد إلى أن نسبة التشابه بين العربيتين درجته 0,31، أي أن حجم الاختلاف المعجمي بين العربيتين أكبر من ضعف حجم التشابه، إذ جاء بدرجة 0,69.

إن نتائج هذه الورقة تعزز فهمنا للتنوع اللغوي، وللغات البشرية عموما. كما تثري النقاشات النظرية والأبحاث التطبيقية المتعلقة بصناعة المعاجم العربية. وتؤكد ما توصلت إليه أبحاث سابقة تتعلق بأهمية معالجة اللغة العربية الطبيعية بتخصيص أنظمة آلية لكل عربية على حدة، فقد أدت المحاولات السابقة في تطبيق الأدوات المدرَّبة على العربية الحديثة، وعلى العربية التراثية إلى الحصول على أداء منخفض في تحقيق النتائج المرجوة (Alrabiah, et al.، 2014؛ Alosaimy & Atwell، 2017؛ Emad، 2018؛ AbuZeina & Abdalbaset، 2019). وكل أساليب التصنيف classification تفترض شيئا من المعرفة عن البيانات. وقد نتج عن هذا العمل القائم على بيانات عربية غير منظمة نوعان من البيانات العربية المنظمة، وهي متوفرة على حساب قت هوب GitHub[3]. هذه البيانات عبارة عن ثلاث قوائم. الأولى قائمة كلمات تنتمي للعربية التراثية فقط، وهي 569,855 كلمة. والثانية قائمة كلمات تنتمي للعربية المعاصرة، وهي أيضا 569,855 كلمة. والثالثة قائمة كلمات تشترك فيها العربيتان، وتضم 511,122 كلمة.

ويمكن أن تدرس هذه القوائم لغويا، بتتبع أكثر وسائل الإنماء اللغوي التي استعملها ويستعملها الناطقون بالعربية من قياس واشتقاق ونحت وارتجال واقتراض، أو بناء القواميس للكلمات النادرة التي لم ترد سوى مرة واحدة. كما يمكن أن تفيد هذه القوائم الباحثين في مجال تعلم الآلة، والتنقيب في النصوص، ومعالجة اللغات الطبيعية. إذ يمكن أن تستعمل بوصفها بيانات تدريب وتطور بها نماذج تمكن من تصنيف النصوص العربية حسب نوعها (التراثي والمعاصر). أو قد يكون العمل عليها بتحشيتها لغويا، ومن ثم بناء نماذج لغوية في مستويات مختلفة، أو الاستعانة بها في بناء المدققات الإملائية، أو دمجها مع برامج معالجة النصوص، كبرنامج الوورد word للتصحيح الإملائي، أو الاستعانة بها في بناء أنظمة التعرف على الكيانات named-entity recognition.

وقد تعد هذه الورقة أساسا لبحوث أخرى مماثلة، إذ إن قياس التشابه بين الوحدات المعجمية نواة لخوارزميات التشابه الأخرى. فيمكن أن تثرى نتائج هذه الورقة بأبحاث تركز التشابه على المستوى الصوتي والصرفي والتركيبي والدلالي على مستوى الكلمات والجمل. كما يمكن أن يقاس تشابه الجمل بين العربيتين بالاعتماد على نصوص هذا العمل، وباستعمال إحدى الخوارزميات الخاصة بقياس التشابه بين الجمل. وحيث تتجاهل خوارزمية جاكارد تسلسل الكلمات وتكراراتها ودلالالتها؛ فمن المهم أن ننظر في الكلمات المتشابهة في المعنى الواجب التعامل معها على أنها كلمات متشابهة؛ ولذا أخطط للعمل مستقبلا على استعمال خوارزميات أخرى تكشف عن أوجه التشابه الدلالي بين عربية التراث والعربية المعاصرة.

الهوامش

1 هي إحدى مدونات TenTen Corpus المكونة من مجموعة مدونات بأكثر من 30 لغة، وقد جمعت نصوص TenTen العربية من الويب العربي، وبلغت 7,475,624,779 كلمة، وتهدف للوصول لأكثر من 10 بلايين كلمة.

2 يقسم فرستيغ العربية إلى العربية الفصحى الكلاسيكية الممتدة من (ما قبل الإسلام وحتى عام 183 هـ)، والعربية الوسيطة الممتدة من (184 - 1214هـ)، والعربية المعاصرة الممتدة من (1215هـ وحتى الآن).

3 https://github.com/AfrahAltamimi/Arabic_lexical_units

المراجع العربية

ابن خلدون، عبد الرحمن بن محمد. (ت. 805 هـ، ط. 1979). *مقدمة ابن خلدون* (ط 3)، تحقيق: علي عبد الواحد وافي. دار نهضة مصر، القاهرة.

ابن مراد، إبراهيم. (2006). الوحدة المعجمية بين الأفراد والتضام والتلازم. *مجلة الدراسات المعجمية*، *5*، 23-31.

السُّيوطي، أبو الفضل جلال الدين عبد الرحمن بن أبي بكر. (ت. 911 هـ، ط. 2006). *الإتقان في علوم القرآن* (ط 1). تحقيق: مركز الدراسات القرآنية، مجمع الملك فهد لطباعة المصحف الشريف، المدينة المنورة.

فرستيغ، كيس. (2003). *اللغة العربية: تاريخها ومستوياتها وتأثيرها* (ط 1)، ترجمة: محمد الشرقاوي. المجلس الأعلى للثقافة، القاهرة.

المراجع الأجنبية

Abunasser, M. (2015). *Computational measures of linguistic variation: A study of Arabic varieties* [Doctoral dissertation, University of Illinois at Urbana-Champaign]. University of Illinois at Urbana. http://hdl.handle.net/2142/78346

AbuZeina, D., & Abdalbaset, T. M. (2019). Exploring the Performance of Tagging for the Classical and the Modern Standard Arabic. *Advances in Fuzzy Systems*, 1-10. https://doi.org/10.1155/2019/6254649

Ali, S. M., & Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, *28*(1), 131-142.

Aggarwal, C.C., & Zhai, C. (2012). A Survey of Text Clustering Algorithms .In C.C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 77-128). Springer.

Al-Kabi, M. N., & Al-Sinjilawi, S. I. (2007). A comparative study of the efficiency of different measures to classify Arabic text. *University of Sharjah Journal of Pure and Applied Sciences*, *4*(2), 13-26.

Almujaiwel, S. (2020). An Overview of the Lexical Variation Between Arabic Lexicons and Natural Arabic Language. *International Journal of Arabic Linguistics*, *6* (1-2), 91-107.

Alosaimy, A., & Atwell, E. (2017). Tagging Classical Arabic Text using Available Morphological Analyzers and Part of Speech Taggers. *Journal for Language Technology and Computational Linguistics*. *(32)*1, 1-26.

Alrabiah, M., Al-Salman, A., Atwell, E., & Alhelewh, N. (2014). KSUCCA: A key to exploring Arabic historical linguistics. *International Journal of Computational Linguistics (IJCL)*. *5*(2), 27-36.

Hand, D.J., & Adams, N.M. (2015). Data Mining. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri & J.L. Teugels (Eds.), *Wiley StatsRef:Statistics Reference Online* (pp. 1-7). Wiley.

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaili, K. (2015). Cross-dialectal Arabic processing. In Gelbukh, A. (Ed.), *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 620-632). Springer.

Kondrak, G. (2005). N-gram similarity and distance. In Mariano C. & Gonzalo N. (Eds.), *International symposium on string processing and information retrieval* (pp.115-126). Springer.

Kumar, C. A., Radvansky, M., & Annapurna, J. (2012). Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval. *Cybernetics and Information Technologies*, *12*(1), 34-48.

Kwaik, K. A., Saad, M., Chatzikyriakidis, S., & Dobnika, S. (2018). A Lexical Distance Study of Arabic Dialects. *Procedia computer science*, *142*, 2-13.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletion of ones. *Problems of information Transmission*, *1*(1), 8-17.

Manning, C., Raghavan, P., & Schütze, H. (2008). Scoring, term weighting, and the vector space model. In Christopher D. M., Prabhakar R., & Hinrich S. (Eds.), *Introduction to Information Retrieval* (pp. 100-123). Cambridge University Press.

Mohamed, E. (2018). Morphological segmentation and part-of-speech tagging for the Arabic heritage. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *17*(3), 1-13.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, *48*(3), 443-453.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2014). Using of Jaccard coefficient for keywords similarity. In Sio-Iong A., Alan Hoi-Shou C., Hideki K., Li X. (Eds), *IAENG Transactions on Engineering Sciences: Special Issue of the International MultiConference of Engineers and Computer Scientists 2013 and World Congress on Engineering 2013* (pp. 380-384), CRC Press.

Sanger, W., & Warin, T. (2019). Dataset of Jaccard similarity indices from 1,597 European political manifestos across 27 countries (1945–2017). *Data in brief*, *24*, [paper n. 103907]. https://doi.org/10.1016/j.dib.2019.103907

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.

بيانات الباحث

أفراح عبد العزيز التميمي، أستاذ مساعد في اللغويات التطبيقية في قسم الإعداد اللغوي بمعهد تعليم العربية في جامعة الإمام محمد بن سعود الإسلامية بالسعودية. حاصلة على درجة الدكتوراه في اللغويات التطبيقية من جامعة الإمام محمد بن سعود الإسلامية عام 2019. وتدور اهتماماتها حول اللسانيات التطبيقية وتحديدًا اللسانيات الحاسوبية.

AUTHOR BIODATA

Afrah Altamimi is an Assistant Professor of Applied Linguistics (AL) in Arabic Language Teaching Institute, Imam Muhammad bin Saud Islamic University (IMSIU). Dr. Altamimi received her PhD degree in AL (2019) from (IMSIU). Her research interests include Applied Linguistics and Computational Linguistics.

معرف أوركيد(ORCID): 0000-0001-6295-0050

Email: aahaltamimi@imamu.edu.sa