جامعة الملك سعود
كلية الآداب
مجلة الآداب
ردمد (ورقي): ٣٦٢٠-١٠١٨
ردمد (النشر الإلكتروني): ٨٣٣٩-١٦٥٨

King Saud University
College of Arts
Journal of Arts
ISSN (Paper):1018-3620
ISSN (Electronic):1658-8339

مجلة الآداب، مج (٣٤)، ع (١)، ص ص ٤٥-٥٨، جامعة الملك سعود، الرياض (٢٠٢١م/ ١٤٤٣هـ)
Journal of Arts, Vol. 34 (1), pp 45-58, © King Saud University, Riyadh (2021 /1443H.)

المصادر الصريحة المميزة في العربية: دراسة لغوية حاسوبية

أفراح عبد العزيز التميمي

أستاذ اللغويات الحاسوبية المساعد، قسم الإعداد اللغوي، معهد تعليم العربية، جامعة الإمام محمد بن سعود الإسلامية

(قدم للنشر في ٢١/ ٦/ ١٤٤٢هـ، وقبل للنشر في ٢٨/ ٢/ ١٤٤٣هـ)

الكلمات المفتاحية: درجة الكلمة، تكرار الكلمة، الكلمات المميزة، طول الكلمة، المصادر، مقياس درجة الكلمة.

ملخص البحث: تستهدف هذه الورقة استخلاص المصادر الصريحة المميزة بأنواعها من العربية التراثية والعربية المعاصرة باستعمال إحدى خوارزميات تعلم الآلة غير الموجه، وهي خوارزمية الاستخلاص الآلي السريع للكلمات المميزة RAKE، ثم تنظر فيما تتميز به المصادر المميزة في العربية التراثية عن المصادر المميزة في العربية المعاصرة انطلاقا من أطوالها، مفترضة نمو طول المصدر في العربية المعاصرة. وقد استخلصت أولا جميع الكلمات المميزة في كل من العربية التراثية والعربية المعاصرة، وحذف ما اشتركت به العربيتين من كلمات مميزة، ثم جذّعت بقية الكلمات المميزة بالاستعانة بمقطع آلي، وفصلت المصادر عن بقية الكلمات باستعمال محلل نحوي. وفي المرحلة الأخيرة صنفت المصادر المميزة حسب أطوال حروفها؛ للوقوف على ما يميز كل عربية عن الأخرى في المصادر. وقد أظهرت قائمتا المصادر المميزة اختلافا واضحا تجلى في زيادة متوسط طول الكلمة في المصادر المميزة في العربية المعاصرة، الذي فُسر بحدوث تطورات اجتماعية في تلك الفترة تستدعي زيادات في الكلمات تكون عادة إلصاقية؛ لعدم قدرة المصادر ذات الأطوال القصيرة على التعبير عن المفاهيم المستحدثة.

Gerundivization in Arabic: A computational Linguistic Study

Afrah Abdul Aziz Al-Tamimi

Assistant Professor of Computational Linguistics, Department of Language Preparation, Institute for Teaching Arabic, Imam Muhammad bin Saud Islamic University

(Received: 21/ 6/1442 H, Accepted for publication 28/ 2/1443 H)

Keywords: Word degree, gerundivization, word frequency, keywords, word length, gerunds, score metric degree.

Abstract. This paper aims to extract keywords classed as gerunds from the heritage and contemporary Arabic language. The Rapid Automatic Keyword Extraction algorithm (RAKE algorithm), which is based on automated machine learning, was used for such purpose in Python. It then considers what distinguishes the gerund-classed keywords in heritage Arabic from the gerund-classed keywords in contemporary Arabic based on the gerund lengths, assuming the growth of gerund lengths in contemporary Arabic. I first extracted all keywords from the heritage and contemporary Arabic language. The keywords intersected between the two eras were removed. Afterwards, I tokenized and tagged keywords using a part of speech tagger to define gerunds. Finally, I categorized the gerund-classed according to the character length. This was to show what distinguishes gerund-classed keywords in heritage Arabic from gerund-classed keywords in contemporary Arabic. The two gerund-classed keyword lists showed that word length average increased in the contemporary gerund-classed keywords, due to social developments that generated new concepts developed and formulated by new suffixes. Also, short gerund-classed keywords cannot express the new concepts.

تعرف الكلمات keywords أو العبارات المميزة keyphrase بأنها سلسلة حرفية تتألف من كلمة واحدة أو أكثر، وتستخلص من نص أو مجموعة من النصوص، وتزود بمعلومات مهمة ووصفية عن محتوى النص، والكلمات الأكثر صلة بالنص (Rose, Engel, Cramer, & Cowley, 2010). هذه الكلمات أو العبارات المميزة قد تكون بمثابة ملخص أو وصف لبيانات هذا النص أو تلك النصوص، وقد تفيد في تطبيقات تقنية مختلفة، كمحركات البحث retrieval engines، وواجهات التصفح browsing interfaces، والتنقيب في النصوص (Sarkar, 2013). كما تفيد أيضا في بناء المكانز والقواميس (Kosovac, Vanier, & Froese, 2000)، ومعرفة سمات النصوص المدروسة، والكشف عن الفروقات اللغوية الجغرافية في وعاء أو موضوع معين، وانتقاء الكلمات التي ينبغي التركيز عليها في الدراسات اللغوية (الثبيتي، ٢٠١٧).

إن عملية استخلاص الكلمات أو العبارات المميزة يدويا عملية تستهلك الوقت والجهد والمال، ولذلك من غير الممكن استخلاصها يدويا من النصوص الضخمة. ومن ثم تأتي أهمية الاستخلاص الآلي فيما يقدمه من دقة وسرعة في استخلاص الكلمات أو العبارات المميزة.

وتوجد في تعلم الآلة مناهج موجهة supervised لاستخلاص الكلمات المميزة، وأخرى غير موجهة unsupervised. وتحظى المناهج غير الموجهة بالاهتمام؛ لأنها لا تعتمد على لغة معينة، ولا تحتاج إلى بيانات تدريب (أي: بيانات استخرجت منها الكلمات المميزة يدويا) يترتب عليها وجود مشكلات ذاتية فضلا عن استهلاكها للوقت والمال. أما المناهج الموجهة (Caragea, Bulgarov, Godea, & Gollapalli, 2014; Kim, Medelyan, Kan, & Baldwin, 2013; Meng, Zhao, Han, He, Brusilovsky, & Chi, 2017) فهي أكثر دقة بسبب بنائها لنماذج أقوى وأدق من نماذج المناهج غير الموجهة حسب دراسات سابقة قارنت بينها.

ويمر استخلاص الكلمات المميزة في الأنظمة غير الموجهة عادة بعدد من الخطوات، هي كالتالي (Hasan & Ng, 2014):

١.اختيار الوحدات المعجمية المرشحة candidate lexical units بناءً على أسلوب معين، كاستبعاد ما يعرف بكلمات التوقف أو اختيار قسم محدد من أقسام الكلام.

٢.ترتيب الوحدات المعجمية المرشحة.

٣.تشكيل الكلمات المميزة باختيارها من أعلى الكلمات مرتبة، أو عن طريق اختيار عبارة ذات مرتبة عالية النتيجة أو الأجزاء التي لديها درجة عالية.

وأبرز المناهج المستعملة في استخلاص الكلمات المميزة في الأنظمة غير الموجهة منهج الترتيب المعتمد على التمثيل البيانيgraph-based method. والفكرة الأساسية في هذا المنهج هي إنشاء تمثيل بياني من نص يتضمن عناصر أو عُقدا nodes تمثل المرشَّحات من النص وتسمى رؤوس vertices، وترتبط فيما بينها بخطوط تسمى edges. والهدف الرئيس في ذلك هو ترتيب هذه العقد بالاعتماد على هذا المنهج (McHugh, 1989, p.5).

واللغة العربية مثل غيرها من اللغات الطبيعية تتطور تطورا حتميا في كل مستوياتها اللغوية متفاوتة في حجم التغيير الذي يحدث في كل مستوى. وعلى مستوى المعجم كشفت نتائج دراسة التميمي (قيد النشر) باستعمال خوارزمية Jaccard عن حجم الاختلاف الكبير في الوحدات المعجمية بين العربية التراثية والعربية المعاصرة، وعن تشارك العربيتين وتشابههما بدرجة قدرها ٠,٣١، وهذا ما يؤكد النمو المعجمي الهائل من جانب العربية المعاصرة الذي له أسبابه اللغوية والاجتماعية والاقتصادية والسياسية والحضارية، وغيرها. وتهدف هذه الورقة إلى استخلاص المصادر المميزة من العربية التراثية من جهة، والعربية المعاصرة من جهة أخرى باستعمال خوارزمية RAKE، بوصفها إحدى الخوارزميات الدقيقة والسريعة الأداء التي لا تعتمد على لغة

معينة، وتعمل على البيانات الضخمة، ولا تحتاج إلى بيانات تدريب، وذلك للوصول إلى قائمة المصادر المميزة فقط في كل من العربيتين، والنظر فيما تتميز به تلك المصادر المميزة في العربية التراثية عن المصادر المميزة في العربية المعاصرة انطلاقا من أطوالها، مع افتراضنا نمو طول المصدر في العربية المعاصرة بوصفه أحد مظاهر التطور اللغوي في اللغات الطبيعية على المستوى المعجمي.

خوارزمية الاستخلاص الآلي السريع للكلمات المميِّزة RAKE

من الخوارزميات الحديثة المعتمدة على منهج الترتيب المعتمد على التمثيل البياني graph-based method خوارزمية الاستخلاص الآلي السريع للكلمات المميِّزة RAKE (Rose et. al., 2010). وقد بُنيت خوارزمية ريك RAKE بالاعتماد على ما لوحظ من أن الكلمات المميزة تضم كلمات مفيدة متعددة الكلمات تسمى كلمات المحتوى، ولكنها بلا علامات ترقيم وبلا كلمات وظيفية. ولذا، ستكون الكلمات المميزة keywords في نص يتعلق بأعراض كورونا مثلا: " التهاب الحلق"، "حمى"، "تعب"، فيما لن تعد "أوجاع في الجسم" كلمة مميزة لاحتوائها على كلمة وظيفية "في" مضافة لقائمة الكلمات المستبعدة stopwords. وتعمل الخوارزمية على النص بعد تهيئة النص بتقسيمه إلى مصفوفة من الكلمات المفرقة بعدد من المحددات delimiters كعلامات الترقيم والمسافات، ثم تقسم الكلمات إلى سلاسل من الكلمات المتجاورة باستعمال قائمة الكلمات المستبعدة stopwords. وهكذا تعد كل سلسلة كلمات مرشحات candidate keyword.

ولتفصيل ما سبق سننظر في النص القصير التالي مثالا:

"عرض من أعراض كورونا"

أولا: يفرق النص إلى كلمات ليصبح في مصفوفة كما يلي:

["عرض"، "من"، "أعراض"، "كورونا"]

ثانيا: ترتب الكلمات في سلاسل من الكلمات بعد استبعاد قائمة الكلمات المستبعدة. ومن ثم تستبعد كلمة "من" الوظيفية؛ لوجودها في قائمة كلمات الاستبعاد. ومع كل تخطي لكلمة مستبعدة يصبح لدينا كلمة مميزة مرشحة. فالمرشحات في النص السابق ستكون:

["عرض"، "أعراض كورونا"]

وبعد أن يهيأ النص تشغل الخوارزمية لحساب قيمة score كل كلمة في قائمة الكلمات المميزة المرشحة. ويحسب ذلك بالمعادلة التالية:

درجة الكلمة degree| تكرار الكلمة frequency

ويقصد بتكرار الكلمة عدد مرات ظهورها في قائمة الكلمات المميزة المرشحة، ومن ثم ستظهر حسب النص السابق:

تكرار "عرض"= ١

تكرار "أعراض"= ١

تكرار "كورونا"= ١

أما درجة الكلمة فتفسيرها ينطوي على شيء من نظرية المخطط البسيطة simple graph theory في علم الرياضيات. فدرجة الكلمة هنا تشبه درجة العقدة node في المخطط. فلو مثلنا ما سبق في مخطط، سنرسم مخططا غير موجه ونجعل كل كلمة محتوى عقدة. ثم نربط أي عقدتين معا تظهران متصلتين في الكلمات المميزة المرشحة. ويعني زيادة ارتباطات العقدة (أي زادت درجتها degree) تكرار ورودها وحدوثها مع كلمات مميزة أطول من كلمة واحدة. وهكذا فإن درجة كل كلمة تمثل مدى تكرار ورودها المتصاحب مع الكلمات الأخرى في الكلمات المميزة المرشحة. وللوصول إلى درجة الكلمة degree يحسب عدد مرات ورود الكلمة المقصودة في المرشحات بما في ذلك الكلمة المقصودة نفسها. ولذلك ستكون درجة كلمات النص السابق كما يلي:

درجة "عرض"= ١

درجة "أعراض"= ٢

درجة "كورونا"= ٢

ويظهر أن درجة كلمة "أعراض" أعلى من درجة كلمة "عرض" لأن عرض لم ترد في الكلمات المرشحة سوى مرة واحدة مستقلة، فيما وردت كلمة "أعراض" مرتين: مرة بوصفها مستقلة ومرة بوصفها واردة في كلمة مميزة مرشحة. والجدير بالذكر أن الكلمة قد تزيد درجتها مع زيادة طول الكلمات المرشحة حتى وإن لم تكن واردة في النص سوى مرة واحدة. فدرجة كلمة "أعراض" في النص: "عرض من أعراض كورونا المتحور"، بإضافة كلمة واحدة، ستكون: ٣. وهكذا، تشير درجة الكلمة الأعلى أيضا إلى أن الكلمة قد تكون ظاهرة في كلمة مميزة مرشحة طويلة السلسلة وليس لكثرة ورودها وحسب. وتعتمد خوارزمية RAKE على مقياس درجة الكلمة الذي يمكن الحصول عليه بالمعادلة السابقة الذكر:

مقياس درجة الكلمة score metric = درجة الكلمة degree| تكرار الكلمة frequency

حيث درجة الكلمة في البسط، وتكرار الكلمة في المقام. ودرجة الكلمة degree تكون عالية حين تتكرر الكلمة كثيرا مع الكلمات الأخرى وحين ترد في كلمات مميزة مرشحة طويلة. ويكون تكرار الكلمة frequency حين تتكرر الكلمة كثيرا، ولكن بصرف النظر عن مكان ظهورها. إذن مقياس درجة الكلمة في RAKE لا يعتد بالكلمات ذات التكرارات العالية التي لا ترد في كلمات مميزة مرشحة طويلة، ويفضل الكلمات التي غالبا ما ترد في الكلمات المميزة المرشحة.

وبذلك تضيف الخوارزمية لكل كلمة مميزة مرشحة مقياس درجة الكلمات المبنية منها لإيجاد درجة الكلمة المميزة المرشحة. ثم تأخذ الثلث الأول من المرشحات الحاصلة على أعلى مقياس في قائمة المرشحات كقائمة نهائية للكلمات المميزة المستخلصة. وفي مثالنا السابق ستكون مقاييس درجة الكلمة، كالتالي:

مقياس درجة "عرض" = ١\١ = ١

مقياس درجة "أعراض"= ٢\١ = ٢

مقياس درجة "كورونا"= ٢\١= ٢

وهكذا يكون مقياس درجة الكلمة "عرض" = ١، فيما مقياس درجة "أعراض كورونا" = (مقياس درجة "أعراض" +مقياس درجة "كورونا") = ٤. فتكون الكلمة المميزة المستخلصة من الجملة السابقة "عرض من أعراض كورونا"، والأعلى درجة هي" أعراض كورونا".

لقد بنيت هذه الخوارزمية في الأساس وقيمت على نصوص إنجليزية (Rose, et. al., 2010). والدراسات التي أجرت هذه الخوارزمية على بياناتها طبقتها على لغات أخرى غير العربية. فقد أشار (Sandul & Mikhailova, 2018) إلى فعالية الخوارزمية على النصوص الروسية. وخلصا إلى أن هذه الخوارزمية أكثر دقة من PAKE عند تقليل عدد العبارات المختارة. وجرب (Siddiqi & Sharan, 2018) هذه الخوارزمية على الهندية. وحيث لا يوجد قائمة لكلمات الإيقاف الهندية، أعدا هذه القائمة، وقدما نتائج الخوارزمية على النصوص الهندية، ثم اقترحا عدة نماذج لتحسين فعالية الخوارزمية على تلك النصوص. لقد تميزت خوارزمية RAKE باستعمالها لعمليات حسابية أساسية وبسيطة، وهذا ما يجعل النظام أسرع عند تعامله مع المستندات النصية الكبيرة الحجم، فتستخلص كلمات مميزة أكثر بأخطاء أقل. فمقارنة مثلا بخوارزمية TextRank حسب دراسة (Rose et. al., 2010) حُسب إجمالي الوقت الذي تستغرقه الخوارزميتان لاستخلاص الكلمات المميزة، فوجدوا أن RAKE استخلصت الكلمات المميزة من ٥٠٠ ملخص في ١٦٠ مللي ثانية، فيما استخلصت TextRank الكلمات المميزة في ١٠٠٢ مللي ثانية أي أكثر من ٦ أضعاف زمن الاستخلاص في RAKE. أما من حيث الدقة، فقد أثبتت أيضا دراسة (Rose et. al., 2010) تفوق خورازمية RAKE على مقياس f مقارنة بخوارزمية TextRank وخوارزميات

التعلم الموجه supervised learning (N gram with tag – Pattern with tag –NP chunks with tag).

البيانات والتطبيق

جمع البيانات وتهيئتها

جمعت بيانات الدراسة من مصادر متعددة متاحة على شبكة الإنترنت، جاءت في أوعية متنوعة، وموضوعات مختلفة، وبطريقة غير منظمة لا تراعي سوى الزمن الذي كتبت فيه النصوص. وجاءت بيانات الدراسة في مجموعتين مفصولتين حسب الزمن انطلاقا من رؤية فرستيغ (٢٠٠٣) في تقسيمه لعصور العربية الفصحى. فتضمنت المجموعة الأولى النصوص التراثية التي يمتد زمنها من العصر الجاهلي، وحتى عام ١٢١٤هـ، وسأسميها مدونة العربية التراثية، وتتضمن ١٣٠ مليون كلمة تقريبا. فيما تضمنت المجموعة الثانية النصوص المعاصرة التي يمتد تاريخها من عام ١٢١٤هـ وحتى الآن، وسأسميها مدونة العربية المعاصرة، وتتضمن تقريبا ١٥٥ مليون كلمة. وقد استبعد في الجمع ما هو مقتبس في النصوص المعاصرة من النصوص التراثية، نحو: شروح المعلقات، وكتب الحديث والتفسير والفقه، واللغة والخطب والمقالات الدينية. ويرجع هذا المنهج غير المنظم في الجمع إلى غياب الأوعية المتشارَكة مع الزمن، وعدم توفر المحتوى العربي اللذين يعينان على الالتزام بإطار نموذجي تتشارك فيه العربية التراثية مع العربية المعاصرة؛ لجمع محتوى متجانس من العربيتين.

وقد استعنت في تنقيح البيانات بالمشذب العربي[١]، فحذفت الرموز والتشكيل والكلمات والحروف اللاتينية وعلامات التطويل والمسافات المتكررة. واستعنت ببرنامج النوت باد++ ++Notepad لحذف الكلمات الفارسية والكردية التي تتخلل بعض النصوص العربية. وقد أفدت من تطبيق غواص[٢] في استعراض كلمات المدونة، وأفدت بعد ذلك من دالة ()len في برنامج الأكسل في تحديد ما يفوق طوله من قائمة غواص عن ٢٠ حرف وحذفه من المدونة؛ للتخلص من الكلمات التي أزيلت من بينها المسافات عن طريق تغيير الترميز للنصوص أو أخطاء من المصدر، نحو: عبدالله، غيرأنها، عليأنها....إلخ.. وأخيرا حفظت النصوص في ملفين بصيغة txt وبالترميز ANSI تمهيدا للعمل عليها.

وحيث تقوم خوارزمية RAKE في تحديدها للكلمات المرشحة على قائمة الكلمات المستبعدة stopwords، أعددت بعد تهيئة البيانات قائمة بالكلمات المستبعدة. وقد ضمنتها الكلمات الوظيفية في العربية بكل أحوالها التصريفية والكتابية، فأضفت مثلا حرف الجر (إلى) بأحواله التصريفية والكتابية: إلى – الى – إليهم – اليهم - إليها- وإليهم، فإليهم- أإليهم، وهكذا...، وفاقت القائمة الألفي كلمة، والوصول إليها متاح في صفحتي على القت هوب GitHub. أما ما يتعلق بالمحددات delimiters فقد ضمنتها علامات الترقيم، والأرقام.

التطبيق

اعتمدت هذه الورقة خوارزمية ريك RAKE لاستخلاص الكلمات المميزة من مدونة العربية التراثية، ثم من مدونة العربية المعاصرة باستعمال البايثون واستدعاء مكتبة rake_nltk. وتبدأ الخوارزمية العمل على كل مدونة باستخلاص الكلمات المميزة من كل مدونة من خلال تحليل نصوصها إلى مجموعة من الكلمات المرشحات candidate keywords. فتقسم أولا النص إلى مصفوفة من الكلمات

(١) إحدى الأدوات التي طورها المركز الوطني لتقنية الذكاء الاصطناعي والبيانات الضخمة في مدينة الملك عبد العزيز للعلوم والتقنية، واستعملتها في المعالجة القبلية (preprocessing) لنصوص المدونة العربية: https://sourceforge.net/projects/ghawwasv4/files/Almoshatheb_Alarabi%28V_4.0.0%29_09_12_18.jar/download

(٢) أداة طورت أيضا من المركز الوطني لتقنية الذكاء الاصطناعي والبيانات الضخمة في مدينة الملك عبد العزيز للعلوم والتقنية لمعالجة المدونات العربية، انظر:

Al-Thubaity, A.; Khan, M.; Al-Mazrua, M., & Al-Mousa, M. (2013- Aug.). *New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool*. 2013 International Conference on Asian Language Processing, Urumqi, China.

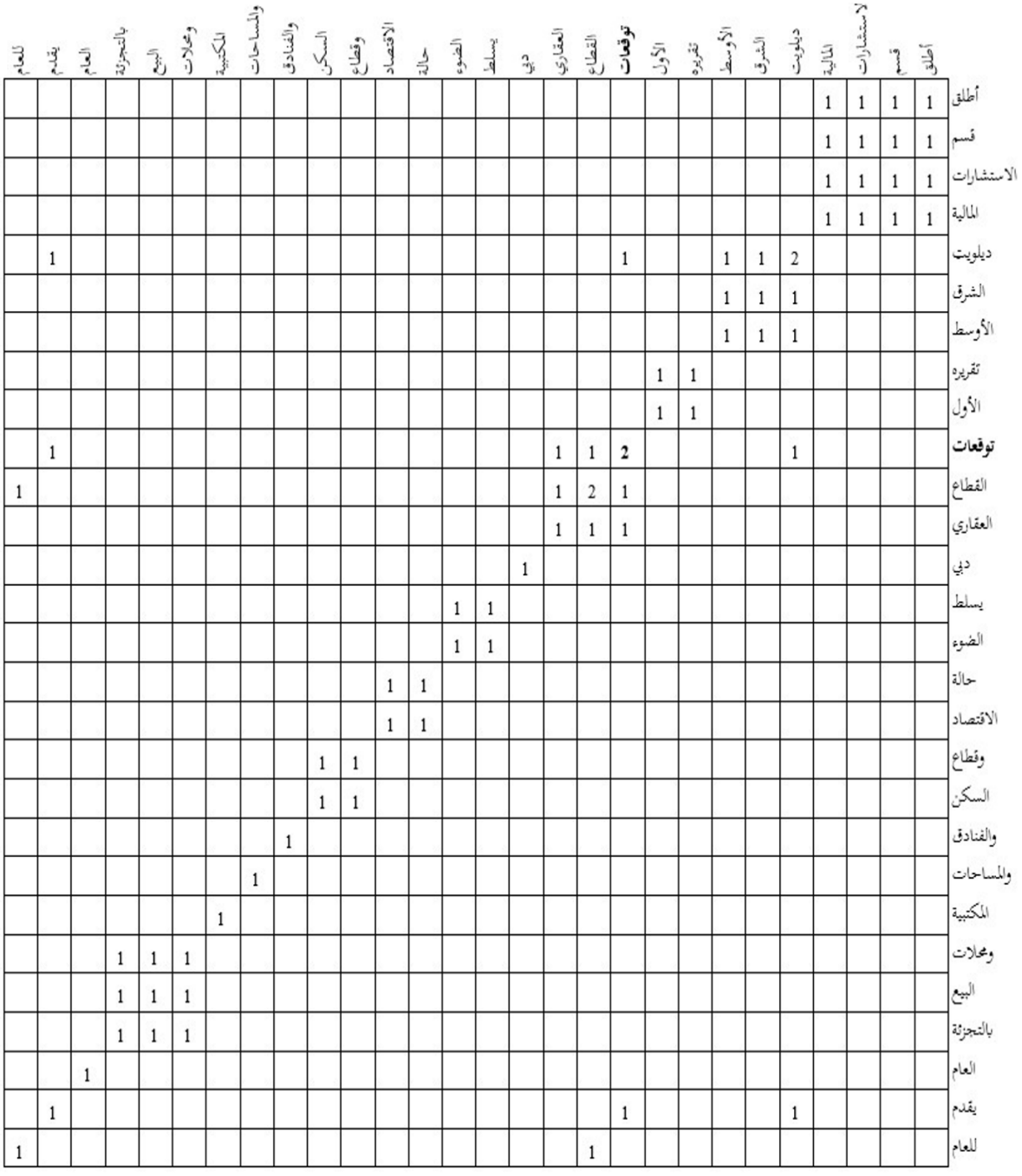
باستعمال المحددات المعينة (علامات الترقيم والأرقام). ثم تقسم هذه المصفوفة إلى سلسلة من الكلمات المتجاورة المفصولة بالمحددات والكلمات المستبعدة المحددة في القائمة. ونظرا لضعف استعمال علامات الترقيم في بعض النصوص، وغيابها في نصوص أخرى ستتشكل لدينا كلمات مرشحة طويلة لاعتماد الخوارزمية في ترشيحها على الكلمات الوظيفية المستبعدة فقط. فضبط أقصى حد لطول الكلمة المرشحة على أن يكون من كلمتين.

ويشير الجدول (١) إلى الكلمات المرشحة بالترتيب الذي حللت عليه من عينة نصية في مدونة العربية المعاصرة. فالكلمة المرشحة (حالة الاقتصاد) تبدأ من الكلمة المستبعدة (على) وتنتهي بالفاصلة. والكلمة المرشحة التي تليها (قطاع السكن) تبدأ من الفاصلة وتنتهي بالفاصلة.

الجدول (١) الكلمات المرشحة في عينة من مدونة العربية المعاصرة

النص الخام	الكلمات المرشحة
أطلق قسم الاستشارات المالية في ديلويت الشرق الأوسط، تقريره الأول حول توقعات القطاع العقاري في دبي، والذي يسلط الضوء على حالة الاقتصاد، وقطاع السكن، والفنادق، والمساحات المكتبية، ومحلات البيع بالتجزئة، في العام ٢٠١٤. كما يقدم توقعات ديلويت لهذا القطاع للعام ٢٠١٥.	أطلق قسم الاستشارات المالية – ديلويت الشرق الأوسط – تقريره الأول – توقعات القطاع العقاري – دبي– يسلط الضوء – حالة الاقتصاد – وقطاع السكن – والفنادق – والمساحات المكتبية – ومحلات البيع بالتجزئة – العام – يقدم توقعات ديلويت – القطاع للعام

وبعد تحديد الكلمات المرشحة واكتمال التمثيل البياني للمتصاحبات كما هو موضح في الشكل (١)، يُحسب مقياس الدرجة لكل كلمة محتوى مرشحة بناء على درجة الكلمة degree (د) وتكرار الكلمة frequency (ت)، وذلك بقسمة (د) على (ت)، كما يوضح الشكل (٢). فمثلا في الشكل (١) يمثل رقم (٢) المربع الذي تلتقي فيه كلمة (توقعات) تكرار الكلمة (ت) في النص السابق، وهذه الكلمة أيضا وردت متصاحبة في كل مرة مرة واحدة مع الكلمات التالية (ديلويت – القطاع – العقاري – يقدم)، وتمثل الأرقام في الصف الذي ترد فيه كلمة (توقعات) نقاط الالتقاء بينها وبين هذه الكلمات، ومجموع هذه الأرقام هو درجة كلمة (توقعات).



الشكل (١). التمثيل البياني للكلمات المتصاحبة في عينة من مدونة العربية المعاصرة

	اطلق	قسم	الاستشارات	المالية	ديلويت	الشرق	الاوسط	تقريره	الاول	توقعات	القطاع	العقاري	دبي	يسلط	الضوء	حالة	الاقتصاد	وقطاع	السكن	والفنادق	والمساحات	المكتبية	ومحلات	البيع	بالتجزئة	العام	يقدم	للعام
د	4	4	4	4	6	3	3	2	2	6	5	3	1	2	2	2	2	2	2	1	1	1	3	3	1	3	3	2
ت	1	1	1	1	2	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
د/ت	4	4	4	4	3	3	3	2	2	3	2.5	3	1	2	2	2	2	2	2	1	1	1	3	3	1	3	3	2

الشكل (٢). مقاييس درجة الكلمات المحسوبة من التمثيل البياني للكلمات المتصاحبة في عينة من مدونة العربية المعاصرة

تقسم درجة الكلمة (٦) على تكرار الكلمة (٢)، فتعطينا مقياس درجة الكلمة score وهو ٣، كما هو واضح في الشكل (٢). وفي الحالات التي نستهدف فيها استخلاص العبارات المميزة لا الكلمات المميزة، فإن مقياس الدرجة score للعبارات المميزة يحسب بجمع مقاييس الدرجة لكل كلمة أو عنصر في العبارة المرشحة، ولنقل عبارة: (توقعات القطاع العقاري)، فتكون كالتالي:

مقياس درجة (توقعات) ويساوي ٣ + مقياس درجة (القطاع) ويساوي ٥ , ٢ + مقياس درجة (العقاري) ويساوي ٣ = ٨ , ٥.

ويعرض الشكل (٣) مقاييس درجات المرشحات بأطوالها المختلفة مرتبة من الأعلى للأدنى. وحيث نستهدف في هذه الورقة الكلمات المميزة ذات الوحدات المعجمية المفردة، خُصص عمل الخوارزمية على المرشحات ذات الطول (٢)، للوصول إلى الكلمات المميزة المفردة التي ترد متجاورة مع كلمات مرشحة من كلمتين فقط، كما حدد منها ما مقياس درجته أعلى من ١، تجنبا للكلمات التي لا ترد إلا مرة واحدة دون تجاور مع كلمات أخرى.

وبعد تخصيص الخوارزمية بالباراميترات الهدف، أجريت على المدونتين، فأظهرت أولا الكلمات المرشحة ذات الطول (٢) في كل مدونة، مرتبة حسب درجاتها من الأعلى للأدنى، ثم الكلمات المرشحة ذات الطول (١) التي وردت في الكلمات المرشحة ذات الطول (٢)، كما في الشكل (٣).

```
('أطلق قسم الاستشارات المالية', 16.0)
('يقدم توقعات ديلويت', 9.0)
('ومحلات البيع بالتجزئة', 9.0)
('ديلويت الشرق الأوسط', 9.0)
('توقعات القطاع العقاري', 8.5)
('القطاع للعام', 4.5)
('يسلط الضوء', 4.0)
('وقطاع السكن', 4.0)
('والمساحات المكتبية', 4.0)
('حالة الاقتصاد', 4.0)
('تقريره الأول', 4.0)
('والفنادق', 1.0)
('دبي', 1.0)
('العام', 1.0)
```

الشكل (٣). الكلمات المرشحة ودرجة مقياس كل كلمة

وجميع ما ظهر من الكلمات المتميزة في هذا التخصيص المحدد بكلمة واحدة فقط في المدونتين تراوح مقياس درجته ما بين ٢ و١,٥. وقد نتج عن ذلك ٢١١,٩٣٤ كلمة مميزة من مدونة العربية التراثية، و٢٤٥,٠١٠ كلمة مميزة من مدونة العربية المعاصرة. وللوصول إلى الكلمات المميزة التي تنفرد بها كل مدونة عن الأخرى، استعين بخوارزمية جاكارد Jaccard لإجرائها على قائمتي الكلمات المميزة في كل مدونة. وخوارزمية جاكارد هي إحدى خوارزميات التشابه التي تعمل على بيانات غير موسمة unlabeled، وتقيس مباشرة أوجه التشابه المعجمي بين نصوص أو قوائم معجمية (Aggarwal & Zhai, 2012, p. 80). وتعني قيمة جاكارد ١ أن التداخل بين الكلمات المميزة حدث في كل الكلمات، فيما تعني القيم ٠ عدم وجود كلمات مميزة مشتركة بين المدونتين. وهذا يدل على أن القيمة كلما زادت كانت نسبة التشابه أعلى، وكلما انخفضت تباعدت القائمتان. وقد جاءت النتائج بتشابه القائمتين بدرجة 0.37 فقط وتباعدهما بدرجة ٠,٦٣. فاستبعدت الكلمات المميزة التي اشتركت فيها القائمتان وهي ١٢٣,٨٣٣ كلمة مميزة من كل مدونة من بين ٤٥٦,٩٤٤ كلمة مميزة (المجموع الكلي للقائمتين). فأصبحت مدونة العربية التراثية تختص بـ ٨٨,١٠١ كلمة مميزة، فيما اختصت العربية المعاصرة بـ ١٢١,١٧٧ كلمة مميزة. وحتى أقف على المصادر تحديدا من بين الكلمات المميزة في المدونتين، استعنت بمقطع وموسم التميمي (٢٠٢٠) النحوي الذي فصلت به ما التصق بهذه الكلمات من حروف عطف وجر وأدوات أخرى في كل قائمة، ثم حددت به القسم الكلامي (المصدر)، وراجعت النتائج يدويا للتصحيح، وحذف المتكرر من المصادر بعد التقطيع، وكذلك حذف المشترك بين القائمتين مرة أخرى باستعمال جاكارد. روجعت النتائج يدويا في المدونتين، فتوقفت على 2376 مصدر من بين الكلمات المميزة في مدونة العربية التراثية، و٣١٠٥ مصدر من بين الكلمات المميزة في مدونة العربية المعاصرة.

مناقشة النتائج

توصلت الخوارزمية إلى ٢٣٦٧ مصدرا مميزا من بين قائمة من ١,٠٨٧,٨٥٧ كلمة نوعية في المدونة العربية التراثية. و٣١٠٥ مصدرا مميزا من بين قائمة من ١,٥٥٠,٥٩٩ كلمة نوعية في المدونة العربية المعاصرة. وبعد استخلاص المصادر المميزة من كل مدونة، صنفت بناء على طول حروف المصدر، ثم حسب متوسط طول المصدر في كل مدونة، بالإضافة إلى حساب التكرارات النسبية للمصادر في قائمة المصادر حسب أطوالها؛ للنظر فيما يميز كل قائمة عن الأخرى، انظر الجدول (٢).

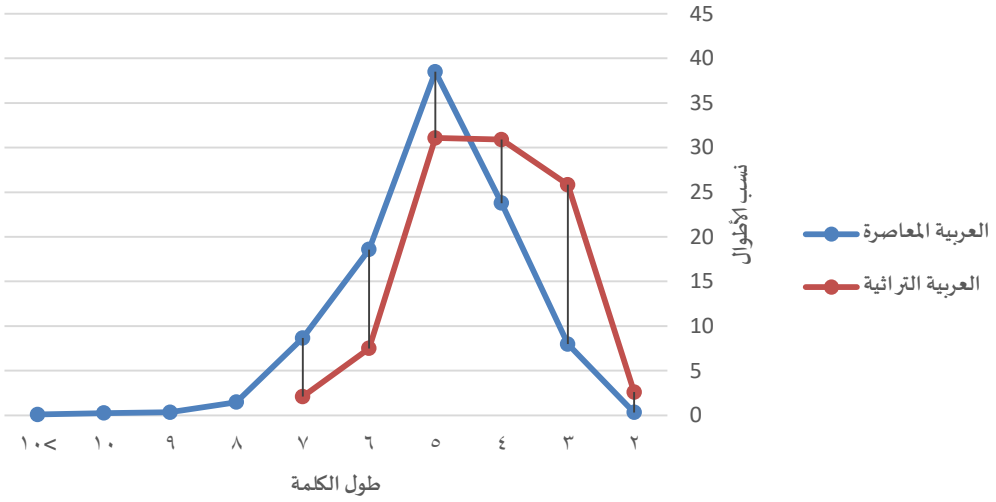
الجدول (٢) إحصاءات المصادر المميزة في المدونتين حسب طول الكلمة

طول المصدر المميز	عدد المصادر في مدونة العربية التراثية	نسبة الطول	عدد المصادر في مدونة العربية المعاصرة	نسبة الطول
مصادر ذات حرفين	62	2.609428	10	0.322061
مصادر ذات ٣ أحرف	614	25.84175	248	7.987118
مصادر ذات ٤ أحرف	734	30.89226	738	23.76812
مصادر ذات ٥ أحرف	738	31.06061	1195	38.48631
مصادر ذات ٦ أحرف	178	7.491582	577	18.58293
مصادر ذات ٧ أحرف	50	2.104377	269	8.663446
مصادر ذات ٨ أحرف	0	0	46	1.481481
مصادر ذات ٩ أحرف	0	0	11	0.354267

تابع الجدول (٢)

طول المصدر المميز	عدد المصادر في مدونة العربية التراثية	نسبة الطول	عدد المصادر في مدونة العربية المعاصرة	نسبة الطول
مصادر ذات ١٠ أحرف	0	0	8	0.257649
مصادر ذات ١١ حرفا وأكثر	0	0	3	0.096618
المجموع	2376	100	3105	100
متوسط طول المصدر	4.5		7	

ويلاحظ في الشكل (٤) أن مصادر العربية التراثية لا تقل عن حرفين، ولا تزيد عن ٧ أحرف، ولكنها إذا ما كانت بالطول ٣ و٤ و٥ تكون أكثر ورودا. وفيما يتشابه توزيع المصادر ذات الأطوال ٣ و٤ و٥ في المدونة التراثية، يظهر الفرق الواضح إذا ما قارناها بالمصادر ذات الستة والسبعة أحرف في نفس المدونة. ومقارنة بالعربية المعاصرة يظهر تميز واضح في ورود المصادر المميزة ذات الثلاثة والأربعة أحرف في عربية التراث أكثر من العربية المعاصرة، ويظهر الفرق الكبير في المصدر الثلاثي تحديدا.



الشكل (٤). توزيع أطوال المصادر في العربيتين

كما يلاحظ من خلال الشكل (٤) نشوء مصادر بالطول ٨ و٩ و١٠ وأكثر من ١٠ في العربية المعاصرة، وبروزها كمصادر مميزة. وتتميز هذه المصادر الأربعة (ذات الأطوال ٨ و٩ و١٠ وأكثر من ١٠) بقلة عددها من مجموع المصادر المميزة الأخرى، إذ لا تتجاوز ٢,٢٪. وهذا ما يتوافق مع قانون زيف للاختصارات[٣] Zipf's Law of Abbreviations الذي يشير إلى أنه كلما زاد طول الكلمة قل تكرارها (قارن بالجدول ٢ أيضا)، وما أثبتته دراسة الخولي التي أجراها على ٣٢٢ كلمة عربية (١٩٨٣)، ودراسة مليكا (٢٠١٨) التي استكشف فيها العلاقة بين متوسط طول الكلمة العربية والتوزيع التكراري لها من القرن الثامن الميلادي وحتى القرن العشرين. وتُظهر العربيتان في الشكل السابق ارتفاعا واضحا في المصادر المميزة الخماسية الطول يفوق المصادر الأخرى بأطوالها المختلفة.

وترى الدراسات المهتمة بطول الكلمة (Grzybek, 2007) أن ثمة جوانب مختلفة تسهم في تحديد متوسط طول الكلمة. ومن خلال الشكل (٤)، يظهر ارتباط بين طول

(٣) انتظام إحصائي يوجد في الأنظمة واللغات الطبيعية، ويزعم بأنه قاعدة عامة. انظر:

Zipf, G. K. (1935). The Psycho-Biology of Language. volume ix. Houghton, Mifflin, Oxford,England.

المصدر والزمن. ويتجلى أثر الزمن في طول الكلمة بوجود مصادر بطول ٨ أحرف وأكثر في العربية المعاصرة، فيما لا يظهر أي مصدر يتجاوز ٧ أحرف في العربية التراثية. والمصادر ذات الأطوال من ٢-٤ تضاعفت نسبة أطوالها عن نسبة أطوال المصادر في العربية المعاصرة، وينطبق العكس تماما على المصادر ذات الأطوال ٥ -٧. وقد أثر وجود المصادر الطويلة ذات الأطوال من ٨ وأكثر على متوسط طول المصدر في العربية المعاصرة الذي بلغ ٧. كما تزامن مع انخفاض في تكرارات المصادر الأخرى ذات الطولين ٢ – ٣.

وبالعودة لكامل قوائم المصادر المميزة في العربيتين (انظر إلى عينة منها في الجدول ٣ و٤)، نجد تميز الكلمات التي يزيد طولها عن ٧ أحرف في العربية المعاصرة بكونها جميعا مصادر صناعية، ما عدا مصدر مرة واحدا بالطول ٨ وهو (استطرافة). فأبرز ملامح الزيادة في الطول جاءت من لاحقة المصدر الصناعي (ياء النسب والتاء)، وبالتركيب المزجي أحيانا أخرى، كما في الكلمة: (جيوسياسية)، وبالاستعارة من لغات أخرى في الغالب، كما في: (ميتافيزيقية). كما نلحظ أيضا أن المصادر العربية الأصل تقف في أطوالها عند ٩ أحرف، وما زاد عن ذلك جاء مصدرا مستعارا من لغة أخرى، نحو: (ميكيافيلية – بيروقراطية).

الجدول (٣) عينة من المصادر المميزة في العربية التراثية بأطوالها

طول الكلمة	٢	٣	٤	٥	٦	٧
الكلمات	لك	صيت	هلاك	تشارط	اجتهاد	استقراض
	كد	صوغ	تبتل	توسيط	اعتزاء	استصغار
	قر	صمم	ثبور	تواري	انكفاف	استصحاب
	قذ	بين	ثبوت	إدغال	انتباذ	استيفاء
	فن	صلب	تشفي	تزرير	اعتياض	استطابة
	فك	بيع	نكاح	إعتاق	اعتياد	استحياء
	فص	وهن	توبة	تشاحن	اشتداد	استئذان
	فد	صفع	عدول	إسراج	مقامرة	استفتاح
	فت	صفح	نقمة	تسمير	تشديده	استتابة
	غي	وقص	تفضل	تسميت	انفكاك	استنصار

الجدول (٤) عينة من المصادر المميزة في العربية المعاصرة بأطوالها

الطول	٢	٣	٤	٥	٦	٧	٨	٩	١٠	١١ وأكثر
الكلمات	نب	ضحك	فكاك	تطوير	ابتعاث	استصراخ	انشقاقية	راديكالية	إمبراطورية	ميتافيزيقية
	طن	دعم	تنصل	تقعيد	اصطكاك	استعباط	احترازية	استمرارية	بيروقراطية	انثروبولوجية
	ضو	بوح	تنصت	تطبيق	افتكار	استظهار	اشتراكية	استشراقية	دراماتيكية	
	ضل	ولع	تسخط	تهريج	انتماء	استعباد	ارتكازية	دبلوماسية	ميكيافيلية	
	ضخ	دخل	صراع	تهريب	ترنيمة	استثارة	انضباطية	جيوسياسية	أرستقراطية	
	رض	قسم	صراخ	إشراق	تعليمة	استقطاع	لوجيستية	استرجاعية	ديموقراطية	
	رص	وعي	حصول	إشراف	اكتمال	استهداف	احتجاجية	استهلاكية	فرنكوفونية	
	ذب	وعظ	كلفة	تطبيب	امتهان	استقطاب	احتياطية	استعراضية	استراتيجية	
	حب	بطر	طموح	إشراط	اخضرار	استرحام	انطباعية	ثيوقراطية		
	أر	بطء	أناة	تضايف	انتظام	استشكال	انضباطية	تاريخانية		

إن أطوال المصادر المميزة في ارتباط زيادتها بالزمن تعكس اتجاهات التطور المجتمعي والتوسع الحضاري في تلك الفترة (من ١٢١٤ ه وحتى الآن). فالمصادر ذات الأطوال القصيرة محدودة، ولا يمكن أن تعبر عن جميع المفاهيم المستحدثة، وهذا ما أدى إلى النمو في متوسط طول المصدر في العربية المعاصرة الذي يتشكل عادة بإلصاق اللاحقتين (التاء– ياء النسب والتاء). وإذا ما نظرنا في أمثلة الجدول (٤) وفي الأطوال ٨–٩–١٠–١١ وأكثر، سنجد أن المصادر المميزة يجمعها تمثيلها لأولويات المجتمع في تلك الفترة (سياسة – اقتصاد)، وهي ما ساهمت في زيادة متوسط طول المصدر.

الخاتمة

تبين نتائج الورقة أنه من الممكن تطبيق خوارزمية RAKE لاستخلاص الكلمات المميزة من النصوص العربية الكبيرة الحجم. فطبقت هذه الخوارزمية على مدونتين: الأولى تمثل العربية التراثية، والثانية تمثل العربية المعاصرة. وبالاستعانة بأدوات معالجة أخرى، توصلت إلى المصادر المميزة في المدونتين؛ للوقوف على ما يميز العربية التراثية عن العربية المعاصرة في هذا النوع الكلامي تحديدا. وحيث صنفت الورقة المصادر المميزة حسب أطوالها، كانت النتائج غير مفاجئة، فأبحاث اللسانيات الكمية التي اهتمت بدراسة طول الكلمة تاريخيا في أكثر من لغة (Bochkarev, Shevlyakova, & Solovyev, 2015; Chen & Liu, 2014)، ومنها العربية (الخولي، ١٩٨٣ ;Milička, 2018)، تشير إلى أن متوسط طول الكلمة ينمو مع الزمن، وقد كانت المصادر المميزة في العربية كذلك.

إن الملاحظات الذاتية على المصادر المميزة في هذه الورقة تستحق مزيدًا من التحليل الكمي يتجاوز نطاق هذه الورقة. ويمكن أن تستثمر قوائم المصادر المميزة التي استخلصت من المدونتين في تتبع التغيرات اللغوية الصوتية الصرفية التي تعرضت لها المصادر عبر عصور العربية. كما يمكن تطبيق الخوارزمية على أقسام الكلام الأخرى، والوقوف على التغيرات التي حدثت للعربية من خلالها.

المراجع العربية

التميمي، أفراح (٢٠٢٠). *نحو تحسين أداء نموذج التميمي للوسم النحوي الآلي للغة العربية. المجلة الدولية للسانيات العربية*، *٧* (١)، ٧٨–٦١

التميمي، أفراح (قيد النشر). *التشابه والاختلاف في الوحدات المعجمية بين عربية التراث والعربية المعاصرة. مجلة اللسانيات العربية*.

الثبيتي، عبد المحسن (٢٠١٧). *المعالجة الآلية للكلمات المميزة للمدونات*: قضايا تقنية.

المجيول، سلطان (محرر). *لغويات المدونة الحاسوبية: تحليلات تطبيقية على العربية الطبيعية*. (ص ص. ٩٢–١٣٥). الرياض: مركز الملك عبد الله الدولي لخدمة اللغة العربية.

الخولي، محمد (١٩٨٣). *العلاقة بين طول الكلمة وشيوعها في اللغة العربية. دراسات: مجلة كلية التربية*،*١* (٥)، ١١١–١٢٥.

فرستيغ، كيس (٢٠٠٣). *اللغة العـــربية: تاريخها ومستوياتها وتأثيرها*. ترجمة محمد الشرقاوي. مصر: المجلس الأعلى للثقافة. (العمل الأصلي دون تاريخ).

المراجع الأجنبية

Aggarwal, Ch. C., & Zhai, Ch. (2012). A Survey of Text Clustering Algorithms. In Ch. C. Aggarwal & Ch. Zhai (Eds.), *Mining text data* (pp. 77-128). New York: Springer-Verlag.

Al-Thubaity, A.; Khan, M.; Al-Mazrua, M., & Al-Mousa, M. (2013- Aug.). *New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool*. 2013 International Conference on Asian Language Processing, Urumqi, China.

Bochkarev, V. V., Shevlyakova, A. V., & Solovyev, V. D. (2015). The average word length dynamics as an indicator of cultural

changes in society. *Social Evolution & History, 14*(2), 153-175.

Caragea, C., Bulgarov, F. A., Godea, A., & Gollapalli, S. D. (2014). Citation-enhanced keyphrase extraction from research papers: A supervised approach. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *The 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1435–1446). Stroudsburg (USA): Association for Computational Linguistics.

Chen, H., & Liu, H. (2014). A diachronic study of Chinese word length distribution. *Glottometrics, 29*, 81-94.

Grzybek, P. (2007). History and Methodology of Word Length Studies. In: P. Grzybek (Ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues* (Vol. 31, pp.15- 90). Dordrecht: Springer.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1262–1273). Stroudsburg (USA): Association for Computational Linguistics.

Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2013). Automatic keyphrase extraction from scientific articles. *Language Resources and Evaluation*, *47*(3), 723–742.

Kosovac, B., Vanier, D. J., & Froese, T. M. (2000). Use of keyphrase extraction software for creation of an AEC/FM thesaurus. *Electronic Journal of Information Technology in Construction*, *5*, 25–36.

McHugh, J. A. M. (1989). *Algorithmic graph theory*. New York: Prentice Hall

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. In R. Barzilay & M-Y Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 582–592). Stroudsburg (USA): Association for Computational Linguistics.

Milička, J. (2018). Average Word Length from the Diachronic Perspective: The Case of Arabic, *Linguistic Frontiers*, *1*(2), 81-89.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In M. W. Berry & J. Kogan (Eds.), *Text Mining: Applications and Theory* (pp. 1–20). New Jersey: John Wiley and Sons.

Sandul, M. V., & Mikhailova, E. G. (2018). Keyword extraction from single Russian document. In Y. Litvinov, M. Akhin, B. Novikov & V. Itsykson (Eds.), *Proceedings of the Third Conference on Software Engineering and Information Management* (Vol. 2135, pp. 30–36). New York: Association for Computing Machinery.

Sarkar, K. (2013). Automatic single document text summarization using key concepts in documents. *Journal of Information Processing Systems*, *9*(4), 602-620.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. volume ix. Houghton, Mifflin, Oxford,England.