معجم جامعة الملك سعود للغة العربية واللهجة السعودية (KSUSC)

الوصف:

تم بناء هذا المعجم ضمن مشروع بحثي بعنوان "إطار لتعدين وسائل الإعلام العربية واستخلاص المعلومات باستخدام التمثيل الدلالي المتصل للكلمات" مدعوم من مبادرة التعاون الدولي بوزارة التعليم وكنت الباحث الرئيس للمشروع. التعاون كان مع ا. د. محسن رشوان و ا. د. شريف عبدو من جامعة القاهرة.

يعد معجم جامعة الملك سعود للهجة السعودية (KSUSC) أحدث وأكبر معجم للغة العامية السعودية، متضمنًا أكثر من بليون كلمة. حيث يحتوي المعجم على +26 مليون كلمة فريدة و+ 184 مليون جملة وما يقارب بليون كلمة من اللغة العربية الفصحى الحديثة. كما يحتوي على + 6 مليون كلمة فريدة و+ 14 مليون جملة و+126 مليون كلمة من اللغة العامية السعودية. وعلى الرغم من أن العدد الإجمالي لكلمات اللغة العامية في معجم (KSUSC) أقل من العدد الإجمالي لكلمات اللغة العربية الفصحى، إلا أن حجم معجم اللغة العامية لايزال أكبر بكثير من معاجم اللغة العامية الأخرى للهجة السعودية. إضافة إلى ذلك، وعند مقارنة عدد الكلمات الفريدة بين اللغتين، يمكن ملاحظة أن عدد كلمات اللغة العربية الفصحى الحديثة تمثل 56 ٪ من العدد الإجمالي للكلمات الفريدة؛ لذلك لا يزال معجم (KSUSC) غني بمفردات اللغة العامية.

ومن الجدير بالذكر أنه تم بناء معجم (KSUSC) بناء على عدة معايير، ومنها:

- التاريخ: يشمل معجم (KSUSC) بعض البيانات المأخوذة من معاجم موجودة مسبقًا (حتى عام 2010) بالإضافة إلى بيانات جديد تم كتابتها بحلول نهاية عام 2020.
- المصادر: تم جمع النصوص الحديثة من خمسة مصادر مختلفة، تشمل المعاجم الموجودة مسبقًا، والمواقع الإلكترونية، ومنصات وسائط التواصل الاجتماعي.
- المجالات: تعددت مجالات البيانات المخزّنة في معجم (KSUSC) لتشمل أكثر من 26 مجالًا.

بداء من معيار التاريخ، فإن المعجم يركز على المصطلحات التي تم استخدامها في السنوات الحديثة وبالأخص بين عام 2018م وعام 2020م لحصر المفردات الجديدة التي تعكس آخر الأحداث ومنها جائحة الكرونا-19. حيث مثّلت عدد الكلمات الفريدة المكتوبة في عام 2020 أكثر من 50% من بقية البيانات بعدد يقارب 8 مليون كلمة فريدة، ومن بعدها بيانات عام 2018 بنسبة تقارب الـ 25%، ومن ثمة بقية السنوات حتى عام 2010.

أما فيما يخص معيار تنوع المصادر، فقد تم جمع معاجم سابقة ودمجها مع بيانات جمعت عبر الإنترنت من مصادر جديدة ومتنوعة، وذلك من خلال استخدام الفيسبوك (Facebook) واليوتوب (YouTube) والتويتر (Twitter) بالإضافة إلى مواقع الويب الأخرى. حيث تم جمع أكثر من 102 مليون كلمة باللغة العامية من اليوتوب، و +10 مليون كلمة من التويتر، و + 2 مليون كلمة من مواقع الويب الأخرى. وإضافة إلى ذلك، تم حصر + 900 مليون كلمة من اللغة العربية الفصحى الحديثة بناء على نظام المعالجة المقترح لمعاجم البيانات الموجودة مسبقًا، مما يشير إلى أن النصوص قد تم تنظيفها وتوحيدها بالفعل. كما تم جمع +32 مليون كلمة من اللغة العربية الفصحى الحديثة بناء على صفحات الويب و + 20 مليون كلمة جمعت من الفيسبوك.

ويُعد تنوع المجال معيارًا آخر يجب تسليط الضوء عليه فيما يتعلق بعدد الكلمات الفريدة. حيث أن المجال الذي يحتوي على أكبر عدد من الكلمات الفريدة هو المجال العام (بنسبة 34٪) وذلك لأنه يحتوي على نصوص لا تركز على موضوع معين. بعد ذلك، يأتي مجال التمثيل في المرتبة الثانية، ويشكل 19٪ من جميع الكلمات الفريدة، ومن ثم مجال الأخبار في المرتبة الثالثة بحوالي 7٪ من جميع الكلمات الفريدة، بينما تأتي بقية الـ 23 مجال بعد ذلك بنسب متشابهة ليحصل في النهاية مجالي الموسيقى (0.02٪) والتاريخ (0.03٪) على أقل عدد من الكلمات الفريدة.

وختامًا، من الضرورة التنويه إلى أنه نظرا لأن مصادر البيانات التي تم جمعها للمعجم كانت من مصادر متاحة عبر الإنترنت، فكان البعض منها عليه حقوق نشر نشطة، مثل الصحف والمجلات والكتب والتغريدات والمواقع الإلكترونية. وبالتالي، تم اتخاذ الإجراءات التالية في هذا المشروع: (1) توفير

المعلومات الببليوغرافية حول محتوى المجموعة. (2) تكون معاينة النص الكامل مقيدة وغير متاحة للجمهور؛ و (3) لا يتم توزيع النص الذي تم جمعه ويمكن استخدامه محليًا فقط لأغراض البحث. سيتم استخدام المعجم وفقًا للقيود المذكورة سابقًا، ولأنها مخصصة لأغراض البحث، فهي متوافقة مع قانون حقوق النشر السعودي الحالي.

لمزيد من المعلومات عن البحث، يمكن الاطلاع على الورقة المنشورة التالية والمرفقة:

Elgibreen, H., Faisal, M., Al Sulaiman, M., Abdou, S., Mekhtiche, M.A., Moussa, A.M., Alohali, Y.A., Abdul, W., Muhammad, G., Rashwan, M. and Algabri, M., 2021. An incremental approach to corpus design and construction: application to a large contemporary saudi corpus. IEEE Access, 9, pp.88405-88428.