مجلة كلية الآداب. جامعة الإسكندرية ، المجلد ٧٢ ، العدد ١١٠، اكتوبر ٢٠٢٢

نظام للتدقيق الإملائي للغة العربية للشبكة العنكبوتية باستخدام معجم حاسوبي وقوانين إملائية وصوتية

د. صلاح راشد الناجم

كلية الآداب، قسم اللغة العربية

جامعة الكويت

salah.alnajem@ku.edu.kw

مستخلص

يقدم هذا البحث نظاما لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية (Web Spell Checker) قمنا بتصميمه باستخدام نظام WebSpellChecker Engine. نظامنا متاح للمستخدمين على شكل خدمة ويب سحابية (Cloud Web Service) يمكن دمجها مع أي موقع أو تطبيق متاح على الشبكة العنكبوتية، كما يمكن دمجه مع تطبيقات الأجهزة الذكية وذلك من خلال واجهة لبرمجة التطبيقات (Application Programming Interface) تتيح التدقيق الإملائي لنصوص اللغة العربية المدخلة إلى مواقع وتطبيقات الشبكة العنكبوتية وتطبيقات الأجهزة الذكية. يستطيع نظامنا التعامل مع نسبة كبيرة من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة (Modern Standard Arabic) بشكل خاص باستخدام معجم حاسوبي (Lexicon). صُمم هذا المعجم باستخدام قائمة كلمات ضخمة (Word List) مفتوحة المصدر (Open Source). بُنِيَت هذه القائمة باستخدام قاعدة بيانات معجمية (Lexical Database) مفتوحة المصدر مخصصة للتحليل الصرفي (Morphological Analysis) للأسماء والأفعال العربية صُمِّمَت باستخدام تقنية الآلات منتهية الحالات (Finite State Automata). تحتوي قائمة الكلمات المذكورة على الصيغ الصرفية والاشتقاقية (Inflected and Derived Forms) المحتملة لكلمات اللغة العربية الفصحى (على سبيل المثال: كَتَبَ، ويكتبان، كتبوا، فسيكتبن، كاتِبة، للكاتِبَين، المكتوب). كما تم تزويد النظام بالقدرة على إعادة ترتيب (Re-Ranking) مقترحات التصحيح الآلي الناتجة من تطبيق خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) المستخدمة في التصحيح الحاسوبي الآلي للأخطاء الإملائية من خلال إعطاء الأولوية لإظهار مقترحات التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules). استُخدِم المعجم الحاسوبي والقوانين الإملائية والصوتية السياقية المذكورة لتزويد النظام بالمعرفة اللغوية التي تمكنه من اكتشاف وتصحيح الأخطاء الإملائية في نصوص اللغة العربية الفصحى المدخَلة إلى مواقع الشبكة العنكبوتية.

الكلمات المفتاحية: علم اللغة الحاسوبي، المعالجة الحاسوبية للغة العربية، التدقيق الإملائي، علم اللغة التطبيقي، المعجم الحاسوبي.

1. تمهيد

1.1. مشكلة البحث

تُعد الأخطاء الإملائية مشكلة تواجه مستخدمي اللغة العربية مثل الخطأ في كتابة الهمزة في بداية الكلمة والخطأ في كتابة الأصوات الانزلاقية[١] (Glides) في نهاية الكلمة. فهذه الأخطاء منتشرة بين الناطقين باللغة العربية بشكل عام ومنتشرة أيضاً بين المتعلمين والمثقفين منهم. تؤثر هذه المشكلة بشكل كبير على مجال تقنية المعلومات وأنظمتها التي تتعامل مع اللغة العربية. في هذا السياق، في أنظمة قواعد البيانات التي تستخدمها المواقع التفاعلية على الشبكة العنكبوتية، قد تحتوي قاعدة بيانات يستخدمها موقع تفاعلي لقاموس عربي – إنجليزي على صيغة عربية غير صحيحة إملائيا مثل صيغة "اكل" (دون كتابة لهمزة القطع في بداية الفعل) كقيمة للحقل النصي (Text Field) الذي يحوي الترجمة المقابلة للحقل النصي الذي يحتوي على صيغة الماضي البسيط للفعل الإنجليزي "Ate" المصرفة من الفعل "Eat". عندما يقوم المستخدم بإدخال الصيغة الصحيحة للفعل (أكل) في استفساره المدخل إلى قاعدة البيانات لاسترجاع الفعل الإنجليزي المقابل له "Ate" فلن يحصل على أي نتيجة، وذلك بسبب عدم وجود حقل في قاعدة البيانات يحمل الصيغة الصحيحة "أكل" (مع همزة القطع). حيث يوجد حقل نصي في قاعدة البيانات يحتوي على الصيغة الخاطئة "اكل" التي تمثل الترجمة المقابلة للفعل "Ate" وهي صيغة تختلف عن الصيغة التي أدخلها المستخدم في حرفها الأول، لذلك يعتبر الحاسوب الكلمتين "أكل" و "اكل" كلمتين مختلفتين. تواجه هذه المشكلة مستخدمي ومصممي أنظمة الحاسوب الذين يستخدمون أنظمة إدارة قواعد البيانات (Database Management Systems) التي تتعامل مع البيانات النصية العربية، وكذلك تواجه مستخدمي ومصممي المواقع الإلكترونية على الشبكة العنكبوتية التي تعتمد في عملها على استرجاع البيانات من قواعد البيانات كمحركات البحث (Search Engines) وأنظمة استرجاع المعلومات (Information Retrieval). تمثل هذه المشكلة مصدرا للإرباك والخطأ في عمليات البحث عن المعلومات وسببا في عدم القدرة على الوصول إلى المعلومات الصحيحة بشكل دقيق. فكثيرا ما تُخَزَّن السجلات في قواعد البيانات باستخدام تهجئة خاطئة لهمزة القطع وهمزة الوصل في بداية الكلمة على سبيل المثال، إلا أن المستخدم لا يعرف أن هذه السجلات الخاطئة موجودة في قاعدة البيانات لأنه يقوم بالبحث عن المعلومات في قاعدة البيانات باستخدام التهجئة الصحيحة للكلمة. من جهة أخرى، هناك أخطاء أخرى يقع فيها مستخدمو اللغة العربية عند استخدام تلك الأنظمة بسبب كتابة الكلمة العربية الفصحى بنفس طريقة نطقها باللهجة المحلية. على سبيل المثال، قد يقوم المستخدم بكتابة كلمة "خسائر" في استفساره المدخل إلى قاعدة البيانات على شكل "خساير" المستخدمة في اللهجة الكويتية، وهو ما يؤدي إلى عدم حصوله على النتيجة الصحيحة للاستفسار إذا كانت الكلمة مخزنة في قاعدة البيانات بشكلها الصحيح في العربية الفصحى (خسائر).

في هذا السياق هنالك ندرة في أنظمة التدقيق الإملائي المخصصة لاكتشاف وتصحيح الأخطاء الإملائية في النصوص المدخلة باللغة العربية إلى مواقع الشبكة العنكبوتية (Web Spell Checking Systems) وهي أنظمة تختلف في طريقة عملها واستخداماتها عن أنظمة التدقيق الإملائي المدمجة مع برمجيات معالجة الكلمات (Word Processors) المعروفة مثل نظام Microsoft Word. حيث إن أغلب الأنظمة التجارية الخاصة بالتدقيق الإملائي للشبكة العنكبوتية تعاني من عدم شمول معجمها الحاسوبي على كمية كافية من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة بشكل خاص. كما أن أنظمة التدقيق

[١] الصوت الانزلاقي (Glide) (ويعرف أيضاً بالصوت شبه الصائت) هو صوت يُنطق كما تُنطق الصوائت ويصنف كصوت صامت (الخولي ، 1982).

الإملائي المتقدمة المخصصة لاكتشاف وتصحيح الأخطاء الإملائية في النصوص المدخلة باللغة العربية إلى مواقع الشبكة العنكبوتية والتي تستخدمها الشركات العالمية الكبرى مثل أنظمة شركة مايكروسوفت وجوجل غير متاحة تجاريا لاستخدام مواقع الشبكة العنكبوتية الأخرى، حيث إنها أنظمة مخصصة فقط لاستخدامات تلك الشركات الكبرى وليست متاحة لجهات أخرى. هنا تأتي أهمية العمل الذي نعرضه في هذا البحث لسد هذه الفجوة من خلال تقديم نظام للتدقيق الإملائي للغة العربية للشبكة العنكبوتية يستخدم معجما حاسوبيا (Lexicon) ضخما يحتوي على ٩ ملايين كلمة من كلمات اللغة العربية الفصحى. يحتوي المعجم الحاسوبي لنظامنا على الصيغ الصرفية والاشتقاقية (Inflected and Derived Forms) المحتملة لكلمات اللغة العربية الفصحى (على سبيل المثال: كَتَبَ، ويكتبان، كتبوا، فسيكتبن، كاتِبة، للكاتِبَين، المكتوب). كما يستطيع نظامنا أيضا اقتراح التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules).

2.1. فرضية البحث

يمكن تصميم نظام لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية (Web Spell Checker) يستطيع التعامل مع نسبة كبيرة من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة (Modern Standard Arabic) بشكل خاص باستخدام معجم حاسوبي (Lexicon). يقوم النظام باكتشاف وتصحيح الأخطاء الإملائية من خلال تطبيق خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) المستخدمة في مجال علم الحاسوب في التصحيح الحاسوبي الآلي للأخطاء الإملائية للغات العالمية.

كما يمكن تزويد النظام بالقدرة على إعادة ترتيب (Re-Ranking) مقترحات التصحيح الآلي الناتجة من تطبيق خوارزمية مسافة تحرير ليفينستين من خلال إعطاء الأولوية لإظهار مقترحات التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules).لإنجاز هذا النظام، يمكن استخدام المعجم الحاسوبي المذكور والقوانين الإملائية والصوتية السياقية لتزويد النظام بالمعرفة اللغوية التي تمكنه من اكتشاف وتصحيح الأخطاء الإملائية في نصوص اللغة العربية الفصحى في مواقع الشبكة العنكبوتية.

3.1. الدراسات السابقة

تعاني ساحة البحث العلمي من ندرة شديد في الأبحاث المكتوبة باللغة العربية والتي تتناول موضوع أنظمة اكتشاف وتصحيح الأخطاء الإملائية للغة العربية. لعل البحث الوحيد الذي يمكن أن نتحدث عنه في سياق الأبحاث المكتوبة باللغة العربية هو بحث كتبه الدكتور هيثم زينهم مرسي والذي قدم فيه دراسة وصفية نقدية لأنظمة التدقيق الإملائي للغة العربية. حيث اشتمل البحث على فصلين، قدم الأول منهما دراسة وصفية للمدقق اللغوي الحاسوبي تناول فيها أدوات المدقق الحاسوبي كخيارات التدقيق التلقائي، كما تناول تصحيح الأخطاء الإملائية في برامج Microsoft Word. أما الفصل الثاني فقد قدم دراسة نقدية للمدقق الإملائي الحاسوبي تناولت جوانب منها معجم المدقق الإملائي وخيارات التدقيق التلقائي. كما تطرق البحث إلى أهمية الدراسات البينية وضرورة الاستفادة من

معطياتها في مجال التدقيق الإملائي الحاسوبي، وضرورة حوسبة المعجم اللغوي العربي بشكل يسمح بتصميم التطبيقات الحاسوبية من خلاله بشكل صحيح[٢].

قدمت دراسات أخرى منشورة باللغة الإنجليزية نماذج لأنظمة حاسوبية يمكن استخدامها لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية لكنها لم تركز على اكتشاف وتصحيح الأخطاء الإملائية لتطبيقات الشبكة العنكبوتية. حيث قدمت إحدى الدراسات نموذجا حاسوبيا قائما على استخدام تتابع الحروف (Character-based Tri-gram Model) لتزويد الحاسوب بالمعرفة اللغوية المتعلقة بمجموعات الحروف المتتابعة (Character Clusters) المسموح باستخدامها في كلمات اللغة العربية، وهو ما يوفر طريقة جديدة للكشف عن الأخطاء الإملائية في الكلمات المكتوبة باللغة العربية. كما استخدمت الدراسة تقنية الآلات منتهية الحالات (Finite State Automata)[٣] لتحديد مدى التشابه بين الكلمات المدخلة بشكل خاطئ والكلمات التي يمكن اقتراحها باعتبارها كلمات صحيحة[٤]. من جهة أخرى، قدمت دراستان نظاما لاكتشاف وتصحيح الأخطاء الإملائية استُخدمت فيه أيضا تقنية الآلات منتهية الحالات. وقد استخدمت الدراسة نموذجا للغة (Language Model) صُمِّم عن طريق تحليل نسبة الأخطاء الموجودة في مصادر بيانات نصية مخزنة في مدونات نصية حاسوبية (Corpora) واختيار مجموعة البيانات المناسبة (Optimal Sub-dataset) لتدريب النظام عليها باستخدام تقنية تعلم الآلة (Machine Learning) وذلك لتمكين النظام من اكتشاف وتصحيح الأخطاء الإملائية في نصوص أخرى[٥]. كما قدمت دراسة أخرى نظاما لاكتشاف وتصحيح الأخطاء الإملائية باستخدام تقنية مقارنة جدول من سلاسل الحروف المتتابعة (N-Grams Scores) المستخرجة من مدونات نصية (Corpora) مع الكلمات المدخلة إلى النظام، حيث تُستخدم سلسلة الحروف المتتابعة لأغراض المقارنة للتحقق مما إذا كان من المحتمل أن يكون تتابع الحروف الموجود في الكلمة المدخلة إلى النظام تتابعا صحيحا مستخدما في اللغة[٦]. إضافة إلى ذلك، قدمت

---

[٢] مرسي، هيثم زينهم. "المدقق اللغوي الحاسوبي: دراسة نقدية." *مجلة كلية دار العلوم*، كلية دار العلوم، جامعة القاهرة، مصر، ٢٠١٥، العدد ٨٢ ، الصفحات ٥١٩ - ٥٨٦.

[٣] لمعرفة المزيد عن استخدام الآلات منتهية الحالات في المعالجة الحاسوبية للغة الطبيعية، يمكنكم الرجوع إلى:

Roche, E. & Schabes, Y. "Introduction." *Finite-state language processing*, edited by Roche, E. & Schabes, Y., MIT Press, Cambridge, 1997, pp. 1-66.

[4] Shaalan, Khaled, et al. "Arabic word generation and modelling for spell checking." *LREC*, 2012, pp. 719-725.

[5] Attia, Mohammed, et al. "Arabic spelling error detection and correction." *Natural Language Engineering*, 2016, 22.5, pp. 751-773.

Attia, Mohammed, et al. "Improved spelling error detection and correction for arabic." *Proceedings of COLING 2012*, 2012, Posters, pp. 103-112.

[6] Muaidi, Hasan, and Rasha Al-Tarawneh. "Towards arabic spell-checker based on N-grams scores." *International Journal of Computer Applications*, 2012, 53.3.

دراسة أخرى توصيفا وتصنيفا للأخطاء الإملائية الشائعة التي قد تحدث عند كتابة كلمة عربية وهي تلك الأخطاء التي تواجهها أنظمة التدقيق الإملائي للغة العربية[٧]. كما قدمت دراسة أخرى محاولة لتصميم نظام لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية بناء على المعرفة اللغوية التي اكتسبها النظام من خلال استخدام تقنية تعلم الآلة (Supervised Machine Learning) وذلك لتدريب النظام على نصوص مخزنة في مدونة نصية حاسوبية موسومة (Annotated Corpus) مخصصة لتعلم الأخطاء الإملائية وهي مدونة تحوي جملا تشتمل على أخطاء إملائية مع وسم كل خطأ بالتصحيح المناسب له [٨]. كذلك قدمت دراسة أخرى نموذجا حاسوبيا للتعامل مع الأخطاء الإملائية التي يتسبب بها تقليب الحروف وتبادل أماكنها (Character Permutation Errors)، حيث ذكرت الدراسة أن معظم الأخطاء الإملائية في كتابة الكلمات العربية هي أخطاء يتسبب بها تقليب الحروف وتبادل أماكنها حيث تمثل نسبة هذا النوع من الأخطاء ٦٥٪ من إجمالي الأخطاء الإملائية[٩]. تناولت دراسة أخرى أيضاً منهجا حاسوبيا يستخدم التحليل الصرفي الحاسوبي (Morphological Analysis) المبني على استخدام معجم حاسوبي لجذوع الكلمات (Stem Lexicon)[١٠]. من جهة أخرى، قدمت دراسة أخرى خوارزمية (Algorithm) لمطابقة سلاسل الحروف العربية (Arabic String Matching) التي تُكَوِّن الكلمات وهي خوارزمية تأخذ في الاعتبار السمات الفريدة للغة العربية ومستويات التشابه المختلفة بين الأحرف العربية مثل التشابه الصوتي وتشابه شكل الحروف بالإضافة إلى الأخطاء المتعلقة بالضغط على أزرار لوحة المفاتيح[١١]. كما قدمت دراسة أخرى نظاما للتدقيق الإملائي للغة العربية يستخدم مجموعة متسلسلة من المناهج الحاسوبية والتي تشتمل على المنهج المبني على المعجم الحاسوبي والمنهج المبنى على القوانين اللغوية والمنهج المبني على التحليل الإحصائي[١٢]. كما تناولت دراسة أخرى التحديات التي تواجه أنظمة التدقيق الإملائي للغة العربية وتحدثت عن الخصائص التي تميز اللغة العربية والتي

---

[7] Shaalan, Khaled, Amin Allam, and Abdallah Gomah. "Towards automatic spell checking for Arabic." Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), 2003, Cairo, Egypt.

[8] Hassan, Youssef, Mohamed Aly, and Amir Atiya. "Arabic spelling correction using supervised learning." *arXiv preprint arXiv:1409.8309,* 2014.

[9] H. Gueddah and A. Yousfi, "The impact of arabic inter-character proximity and similarity on spell-checking." *2013 8th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2013, pp. 1-4.

[10] Hamza, Bakkali, et al. "For an independent spell-checking system from the Arabic language vocabulary." *(IJACSA) International Journal of Advanced Computer Science and Applications* , 2014, 5.1.

[11] Ghafour, Hesham H. Abdel, Ali El-Bastawissy, and Abdel Fattah A. Heggazy. "AEDA: Arabic edit distance algorithm Towards a new approach for Arabic name matching." *The 2011 International Conference on Computer Engineering & Systems*, 2011, IEEE, pp. 307-311.

[12] Mars, Mourad. "Toward a Robust Spell Checker for Arabic Text." *International Conference on Computational Science and Its Applications*, 2016, Springer, Cham, pp. 312-322.

تختلف عن خصائص اللغات الأخرى. كما استعرضت الدراسة الأعمال الحالية التي أنجزت في مجال التدقيق الإملائي للغة العربية، حيث قسمت الدراسة هذه الأعمال إلى فئات وفقاً للتقنيات المستخدمة فيها[١٣]. كذلك قدَّمتُ في بحث سابق منهجا حاسوبيا استخدمتُ فيه تقنية الآلات منتهية (Finite State Automata) للتعامل مع أربعة أنواع أساسية من التغيرات الإملائية (Orthographic Variations) في اللغة العربية والتي كثيرا ما تتسبب في أخطاء إملائية في كتابة الأفعال وهي: التغيرات في كتابة همزة القطع في بداية للفعل، التغيرات في كتابة همزة القطع في وسط للفعل، التغيرات في كتابة همزة القطع في نهاية للفعل، والتغيرات في كتابة حروف العلة في نهاية للفعل. يتناول المنهج ٤٢ نوعا من تلك التغيرات ويحدد التعميمات (Generalizations) والقوانين التي تحكم مثل هذه التغيرات استنادا إلى التحليل المقطعي (Syllabification) لبنية الكلمة. يمكن استخدام هذا المنهج الحاسوبي لتمكين أنظمة اكتشاف وتصحيح الأخطاء الإملائية من التعامل مع الأخطاء الإملائية المتعلقة بالتغيرات في كتابة همزة القطع في بداية للفعل، التغيرات في كتابة همزة القطع في وسط للفعل، التغيرات في كتابة همزة القطع في نهاية للفعل، والتغيرات في كتابة حروف العلة في نهاية للفعل[١٤].

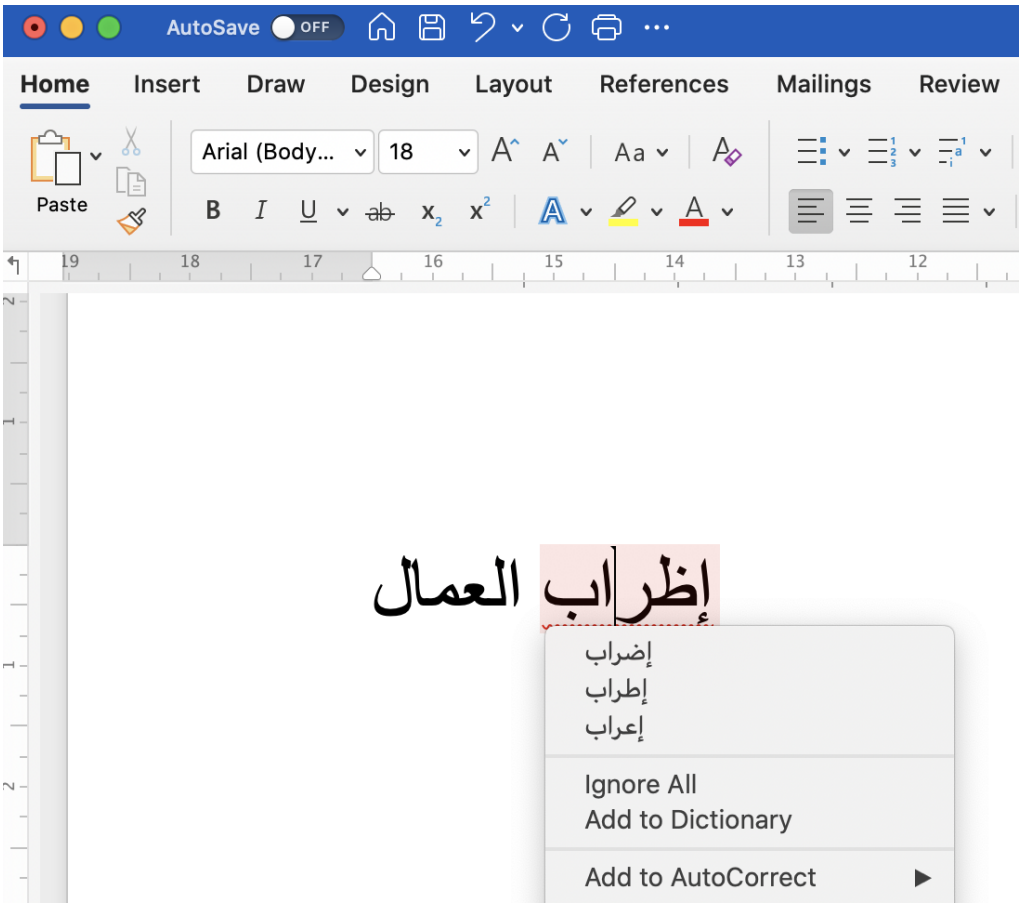
4.1. أنظمة التدقيق الإملائي

أنظمة التدقيق الإملائي هي برامج حاسوبية تقوم باكتشاف الأخطاء الإملائية في الكلمات واقتراح الكلمات الصحيحة عن طريق الرجوع إلى معجم حاسوبي مدمج للتأكد من صحة الكلمة. تستطيع هذه الأنظمة التعرف على الأخطاء الإملائية الشائعة واقتراح البدائل الصحيحة لها باستخدام قوانين صوتية وإملائية. تُستخدم هذه الأنظمة باعتبارها مكونا من مكونات برامج معالجة الكلمات (Word Processing) لأجهزة الحاسوب الشخصية مثل المدقق الإملائي المدمج في برنامج Microsoft Word. كما تُستخدم تلك الأنظمة في التدقيق الإملائي للكلمات المدخلة إلى الأجهزة الذكية كالكلمات المستخدمة في الرسائل القصيرة ورسائل البريد الإلكتروني كما هو الحال في المدقق الإملائي المدمج في نظام التشغيل iOS المستخدم في أجهزة iPhone و iPad الذكية. كما تُستخدم أنظمة التدقيق الإملائي في تدقيق النصوص المدخلة إلى مواقع وتطبيقات الشبكة العنكبوتية (Web Applications). من جهة أخرى، يمكن استخدام أنظمة التدقيق الإملائي أيضا في التصحيح الآلي للاستفسارات (Queries) التي يدخلها المستخدم إلى التطبيقات التفاعلية في الشبكة العنكبوتية (Interactive Web Applications). في هذا السياق، تُستخدم أنظمة التدقيق الإملائي الخاصة بالشبكة العنكبوتية باعتبارها مكونا من مكونات أنظمة استرجاع المعلومات (Information Retrieval). في هذه الأنظمة يقوم الحاسوب بمعالجة استفسار (Query) يدخله المستخدم للبحث عن بيانات موجودة في وثائق أو صفحات على الشبكة العنكبوتية أو على أنظمة إدارة الوثائق (Document Management Systems)، ثم يقوم الحاسوب بالتعرف على الوثائق أو الصفحات التي تحوي الكلمات المفتاحية (Keywords) الموجودة في

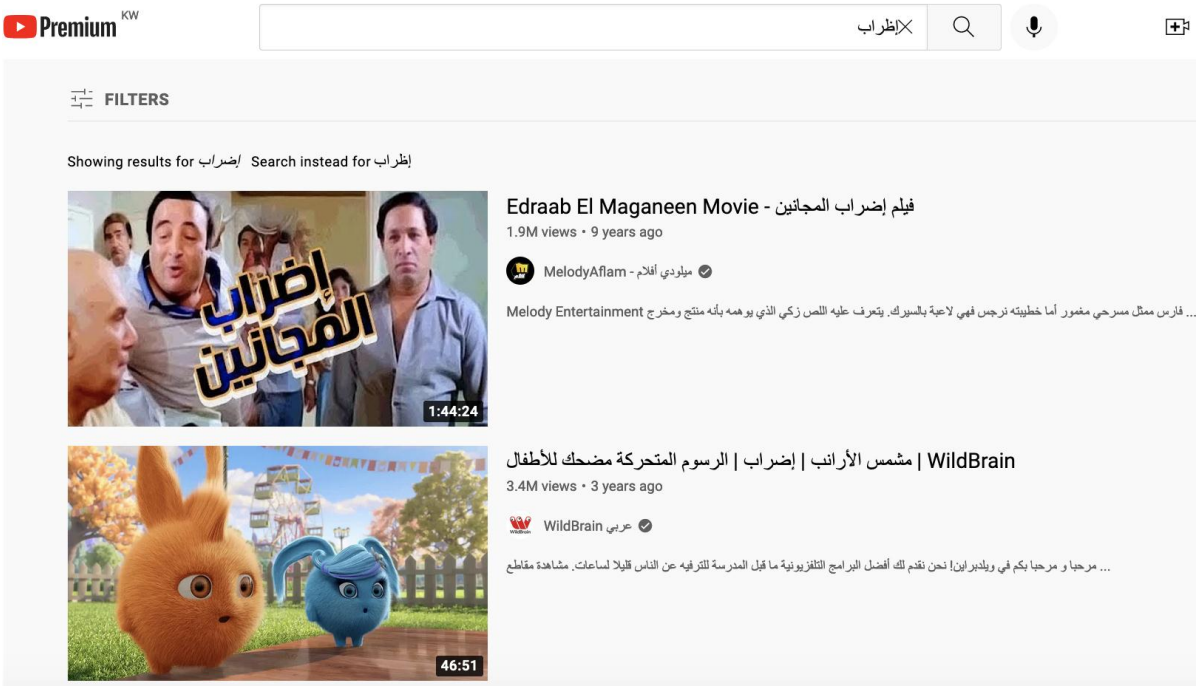
[13] Saty, Ahmed Abdalrhman, Karim Bouzoubaa Bouzoubaa, and Aouragh Si Lhoussain. "Survey of Arabic Checker Techniques." *Journal of Engineering and Computer Science (JECS)*, 2020, 21.1, pp. 34-41.

[14] Alnajem, Salah. "A computational approach to the variations in Arabic verbal orthography." *Computer Speech & Language*, 2005, 19.3, pp. 275-299.

الاستفسار واسترجاعها وعرضها للمستخدم. يعتمد الحاسوب في تحديد الوثائق أو الصفحات المطلوبة على مستوى التشابه بين كلمات الاستفسار ومحتوي النص الموجود في الوثائق أو الصفحات التي يبحث فيها لتحديد ما إذا كانت هذه الوثائق أو الصفحات هي المطلوبة في الاستفسار. يتم ذلك من خلال واجهة استخدام (User Interface) تقوم باستقبال الاستفسار الذي يُدخله المستخدم ثم استرجاع الوثائق أو الصفحات ذات الصلة بالاستفسار، وترتيبها وفقا لدرجة الارتباط بذلك الاستفسار. في هذا السياق، يمكن أن تقوم هذه الواجهة بتحسين دقة الاستفسار الذي يُدخِله المستخدم عن طريق نظام تدقيق إملائي مدمج لاكتشاف الأخطاء الإملائية في الاستفسارات والتصحيح الآلي (Automatic Correction) للكلمات المدخَلة فيها وذلك من خلال عرض النتائج المقابلة للصيغة الصحيحة لغويا للكلمة التي أدخلها المستخدم في استفساره بشكل خاطئ إلى النظام. على سبيل المثال، عندما يقوم المستخدم بإدخال الاستفسار الخاطئ "إظراب العمال" فإن واجهة الاستخدام لنظام استرجاع المعلومات تقوم آليا بتصحيح الاستفسار وتحويله إلى الصيغة الصحيحة إملائيا "إضراب العمال" واسترجاع الوثائق والصفحات المرتبطة بهذه الصيغة الصحيحة. من أمثلة أنظمة استرجاع المعلومات المستخدمة على الشبكة العنكبوتية محركات البحث (Search Engines) مثل محرك البحث Google ومحرك البحث المدمج في منصة You Tube.



شكل رقم (1): مثال على استخدام التدقيق الإملائي في برنامج Microsoft Word



شكل رقم (2): مثال على اكتشاف الأخطاء الإملائية في الاستفسارات والتصحيح الآلي (Automatic Correction) للكلمات المدخَلة فيها في منصة You Tube

5.1. طريقة عمل أنظمة التدقيق الإملائي

يقوم نظام التدقيق الإملائي بمسح النص المدخل إليه وتحديد الكلمات الموجودة فيه، بعد ذلك يقوم النظام بمقارنة كل كلمة في النص بالكلمات الصحيحة لغويا الموجودة في معجمه الحاسوبي[١٥]. في هذا السياق، يرتكز عمل نظام التدقيق الإملائي على فحص الكلمات في ذلك النص والتعرف على الكلمات الصحيحة لغويا منها وتحديد الكلمات التي تحتوي على أخطاء إملائية (Misspelled) وهي تلك الكلمات التي لا وجود لها في المعجم الحاسوبي. إذا تعرف النظام على كلمة تحتوي على أخطاء إملائية، فإنه يقوم بالرجوع إلى المعجم الحاسوبي المدمج ليقترح كلمة بديلة أو أكثر من الكلمات الصحيحة القريبة من الكلمة الخاطئة باعتبارها التهجئة الصحيحة المقترحة التي يمكن استخدامها بدلا من تلك الكلمة التي تحتوي على أخطاء إملائية. لتحديد أقرب الكلمات الصحيحة إلى الكلمة الخاطئة المدخلة إلى النظام، تستخدم أنظمة التدقيق الإملائي – ومنها نظام WebSpellChecker Engine الذي نستخدمه في هذا البحث – خوارزمية حاسوبية تعرف باسم خوارزمية مسافة التحرير (Edit Distance Algorithm). يُعرَف الحد الأدنى من مسافة التحرير بين سلسلتين من الحروف (Strings)[١٦] على أنه الحد الأدنى من عدد عمليات التحرير (مثل عمليات الإدراج والحذف والاستبدال) اللازمة لتحويل سلسلة من الحروف إلى سلسة أخرى. في هذه الخوارزمية، يقاس الاختلاف

[15] Bhaire, Vibhakti V., et al. "Spell checker." International Journal of Scientific and Research Publications , 2015, 5.4 , pp. 5-7.

[١٦] تجب الإشارة هنا إلى أن مصطلح سلسلة الحروف (String) يُستخدم في علم الحاسوب للدلالة على الكلمات.

(ويعرف حاسوبيا بالمسافة Distance) بين كلمتين (سلسلتين من الحروف) عن طريق حساب عدد عمليات التحرير (Edit Operations) اللازمة لتطبيقها على كلمة منهما لتتحول هذه الكلمة إلى الكلمة الأخرى. عمليات التحرير التي تُطبق على الكلمات (سلاسل الحروف) هي إدراج الحروف (Character Insertion) وحذف الحروف (Character Deletion) واستبدال الحروف (Character Substitution). في هذا السياق، تُستخدم قيم محددة تقابل هذه العمليات. على سبيل المثال، تُستخدم القيمة 1 للإدراج والحذف والقيمة 2 للاستبدال، حيث تُعتبر عملية الاستبدال عمليتين مدمجتين وهما الحذف ثم الإدراج. بعد ذلك يتم حساب مجموع قيم عمليات التحرير، حيث يُعرف هذا الحساب لمجموع قيم العمليات بمصطلح مسافة تحرير ليفينستين [١٧] (Levenshtein Edit Distance). مسافة تحرير ليفينستين بين كلمتين (سلسلتين من الحروف) هي الحد الأدنى من عدد عمليات التحرير (Single-Character Edits) وهي عمليات الإدراج والحذف والاستبدال والمطلوب تطبيقها على كلمة معينة لتغييرها إلى كلمة أخرى.

بمعنى آخر، تُعد سلسلة الحروف (الكلمة) **س** على مسافة **ع** من سلسلة الحروف (الكلمة) **ص** إذا استطعنا جعل **س** سلسلة مطابقة لـ **ص** باستخدام تتابع مكون من عدد **ع** من عمليات:

١- إدراج أحرف مفردة في **ص**

٢- أو حذف أحرف مفردة من **ص**

٣- أو استبدال أحرف مفردة في **ص** [١٨]

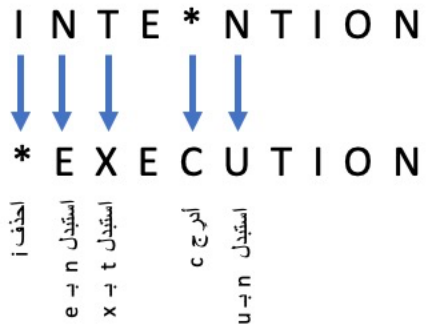
بناء على ما سبق، لكي يقوم نظام التدقيق الإملائي بالتأكد من صحة كلمة أُدخِلَت إليه، يقوم النظام أولا بالبحث عن هذه الكلمة في معجمه الحاسوبي، فإذا وجدها فسيعتبرها كلمة صحيحة؛ أما إذا لم يجدها فإنه سيقوم بحساب مسافة تحرير ليفينستين بين هذه الكلمة المدخلة والكلمات المشابهة لها (القريبة منها) في المعجم، حيث يقوم النظام بحساب الحد الأدنى لقيمة مسافة تحرير ليفينستين بين هذه الكلمة المدخلة وتلك الكلمات المشابهة، ليقوم في النهاية باختيار الكلمات المشابهة التي تحتاج إلى أدنى قيمة لمسافة تحرير ليفينستين ويقترحها للمستخدم باعتبارها صيغا صحيحة بدلا من الصيغة (الكلمة) الخاطئة التي تم إدخالها. على سبيل المثال، قيمة مسافة تحرير ليفينستين بين كلمتي "intention" و "execution" ، هي 8 (احذف i ، استبدل n ب e ، استبدل t ب x ، أدرِج c ، استبدل n ب u)[١٩].

---

[17] Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet physics doklady, 1966, Vol. 10. No. 8, pp. 707-710.

[18] Alnajem, Salah. "A computational approach to the variations in Arabic verbal orthography." *Computer Speech & Language*, 2005, 19.3, pp. 275-299.

[19] Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing.* Upper Saddle River, NJ, Prentice Hall, 2008.



شكل رقم (3): تطبيق خوارزمية مسافة تحرير ليفينستين على كلمتي "intention" و "execution"

الحد الأدنى لقيمة مسافة تحرير ليفينستين (Minimum Edit Distance) بين الكلمة المدخلة والكلمة المشابهة لها لا يمكن التنبؤ به. في هذا السياق، قد تتطلب بعض الأخطاء الإملائية عملية تحرير (Edit Operation) واحدة، ويُطلق على هذا النوع من الأخطاء "خطأ إملائي أحادي السبب" (Single Error Misspelling). في إطار هذا النوع من الأخطاء، تناولت دراسة مبكرة[20] أنواع الأخطاء الإملائية التي تواجه أنظمة التدقيق الإملائي، حيث ذكرت أن ٨٠٪ من الكلمات التي تحتوي على أخطاء إملائية تشترك في أن سبب الخطأ الإملائي فيها هو أحد الأخطاء الآتية:

١- الإدراج (Insertion): مثل كتابة كلمة the على شكل ther

٢- الحذف (Deletion): مثل كتابة كلمة the على شكل th

٣- الاستبدال (Substitution): مثل كتابة كلمة the على شكل thw

٤- التحويل (Transposition): مثل كتابة كلمة the على شكل hte

من جهة أخرى، فإن هنالك عددا من الأخطاء الإملائية لا تنتج عن أخطاء أحادية، وهو ما يؤدي إلى زيادة الحد الأدنى لقيمة مسافة تحرير ليفينستين بين الكلمة المدخلة والكلمة المشابهة.

2. نظام التدقيق الإملائي للغة العربية للشبكة العنكبوتية

باستخدام نظام WebSpellChecker Engine قمنا بتصميم نظام لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية (Web Spell Checker). يتوافر نظامنا للمستخدمين على شكل خدمة ويب سحابية (Cloud Web Service) يمكن دمجها مع أي موقع أو تطبيق متاح على الشبكة العنكبوتية، كما يمكن دمجه مع تطبيقات الأجهزة الذكية وذلك من خلال واجهة لبرمجة التطبيقات (Application Programming Interface) تتيح التدقيق الإملائي لنصوص اللغة العربية

[20] Damerau, Fred J. "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, 1964, 7.3, pp. 171-176.

المدخلة إلى مواقع وتطبيقات الشبكة العنكبوتية وتطبيقات الأجهزة الذكية. يمكن أيضا استخدام هذه الواجهة للتدقيق الإملائي لأنظمة قواعد البيانات ومحركات البحث (Search Engines)، كما يمكن كذلك استخدام هذه الواجهة في التصحيح الآلي للاستفسارات (Queries) التي يدخلها المستخدم إلى التطبيقات التفاعلية في الشبكة العنكبوتية (Interactive Web Applications). في هذا السياق، يمكن استخدام نظامنا باعتباره مكونا من مكونات أنظمة استرجاع المعلومات (Information Retrieval) للتعرف على الاستفسارات المدخلة بشكل خاطئ والتصحيح الآلي لتلك للاستفسارات. من جهة أخرى، يمكن أيضا الاستفادة من نظامنا في تطبيقات تحليل النصوص (Text Analytics) للتعامل مع الأخطاء الإملائية في نصوص اللغة العربية.

يستطيع نظامنا التعامل مع نسبة كبيرة من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة (Modern Standard Arabic) بشكل خاص باستخدام معجم حاسوبي ضخم يحوي ٩ ملايين كلمة من كلمات اللغة العربية الفصحى. يحتوي المعجم على الصيغ الصرفية والاشتقاقية (Inflected and Derived Forms) المحتملة لكلمات اللغة العربية (على سبيل المثال: كَتَبَ، ويكتبان، كتبوا، فسيكتبن، كاتِبة، للكاتِبَين، المكتوب). كما تم تزويد النظام بالقدرة على إعادة ترتيب (Re-Ranking) مقترحات التصحيح الآلي الناتجة من تطبيق خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) المستخدمة في التصحيح الحاسوبي الآلي للأخطاء الإملائية من خلال إعطاء الأولوية لإظهار مقترحات التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules). يمكن استخدام المعجم الحاسوبي والقوانين الإملائية والصوتية السياقية المذكورة لتزويد النظام بالمعرفة اللغوية التي تمكنه من اكتشاف وتصحيح الأخطاء الإملائية في نصوص اللغة العربية الفصحى في مواقع الشبكة العنكبوتية.

1.2. بناء معجم حاسوبي

تمثل مصادر البيانات المعجمية مكونا هاما في أغلب تطبيقات المعالجة الحاسوبية للغة الطبيعية (Natural Language Processing) بشكل عام وتطبيقات تحليل النصوص (Text Analytics) بشكل خاص مثل التطبيقات التي تهتم بانتزاع المعلومات (Information Extraction) من خلال انتزاع أهم الكلمات والعبارات المستخدمة في نصوص معينة، والتطبيقات التي تهتم باكتشاف العلاقات بين الكلمات في النصوص. كذلك تُستخدَم مصادر البيانات المعجمية في تطبيقات تحليل النصوص المتعلقة بتصنيف النصوص (Text Classification) وتحليل المزاج العام (Sentiment Analysis). من جهة أخرى يَعتمد تصميم أنظمة التدقيق الإملائي للنصوص (Spell Checkers) على قوائم الكلمات (Word Lists) لتكوين المعجم الحاسوبي لتلك الأنظمة من أجل اكتشاف وتصحيح الأخطاء الإملائية في الكلمات.

من أجل بناء معجم حاسوبي لنظامنا الخاص بالتدقيق الإملائي للغة العربية للشبكة العنكبوتية، قمنا باستخدام قائمة كلمات (Word List) مفتوحة المصدر (Open Source) مكونة من ٩ ملايين كلمة باللغة العربية الفصحى. تحتوي قائمة الكلمات المذكورة على الصيغ الصرفية والاشتقاقية (Inflected and Derived Forms) المحتملة لكلمات اللغة العربية. تم توليد هذه

الكلمات باستخدام نظام AraComLex (Arabic Computer Lexicon) [٢١] وهو عبارة عن قاعدة بيانات معجمية (Lexical Database) مفتوحة المصدر مخصصة للتحليل الصرفي (Morphological Analysis) للأسماء والأفعال العربية صُمِّمَت باستخدام تقنية الآلات منتهية الحالات (Finite State Automata). تُستخدم تقنية الآلات الثنائية منتهية الحالات بشكل واسع في مجال علم الحاسوب وعلم اللغة الحاسوبي والمعالجة الحاسوبية للغة الطبيعية. بُنِيَت قاعدة بيانات AraComLex المعجمية من مدونة نصية حاسوبية (Corpus) ضخمة وهي مدونة مكونة من 1,089,111,204 كلمة، حيث استُخدِمت تقنيات تعلم الآلة (Machine Learning) وأدوات الوسم الآلي للكلمات (Annotation) وأدوات أخرى لاستخلاص وتنقية المعرفة المعجمية المتعلقة بالسمات الصرف-نحوية (Morpho-syntactic Features) وأشكال التصريفات (Inflection Paradigms) للصيغ الصرفية الأساسية (Lemmas) في هذه القاعدة[٢٢]. تحوي هذه القاعدة المعجمية – إلى الكلمات الشائعة في اللغة العربية الفصحى – الكلمات المستخدمة في العصر الحديث من قِبَل المتحدثين والكتاب الذي يستخدمون اللغة العربية الفصحى في الكتب والصحف ووسائل الإعلام الحديثة الأخرى كالمواقع الإخبارية. كما تحوي كلمات متداولة في مجالات السياسة والعلوم إضافة إلى المصطلحات المسكوكة الحديثة (Coined Terms).

2.2. قوانين إملائية وصوتية سياقية

كما ذكرنا سابقا، قمنا بتزويد نظامنا بالقدرة على إعادة ترتيب (Re-Ranking) مقترحات التصحيح الآلي الناتجة من تطبيق خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) المستخدمة في التصحيح الحاسوبي الآلي للأخطاء الإملائية من خلال إعطاء الأولوية لإظهار مقترحات التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية

---

[٢١] لمعرفة المزيد عن قاعدة بيانات AraComLex المعجمية، يمكنكم الرجوع إلى:

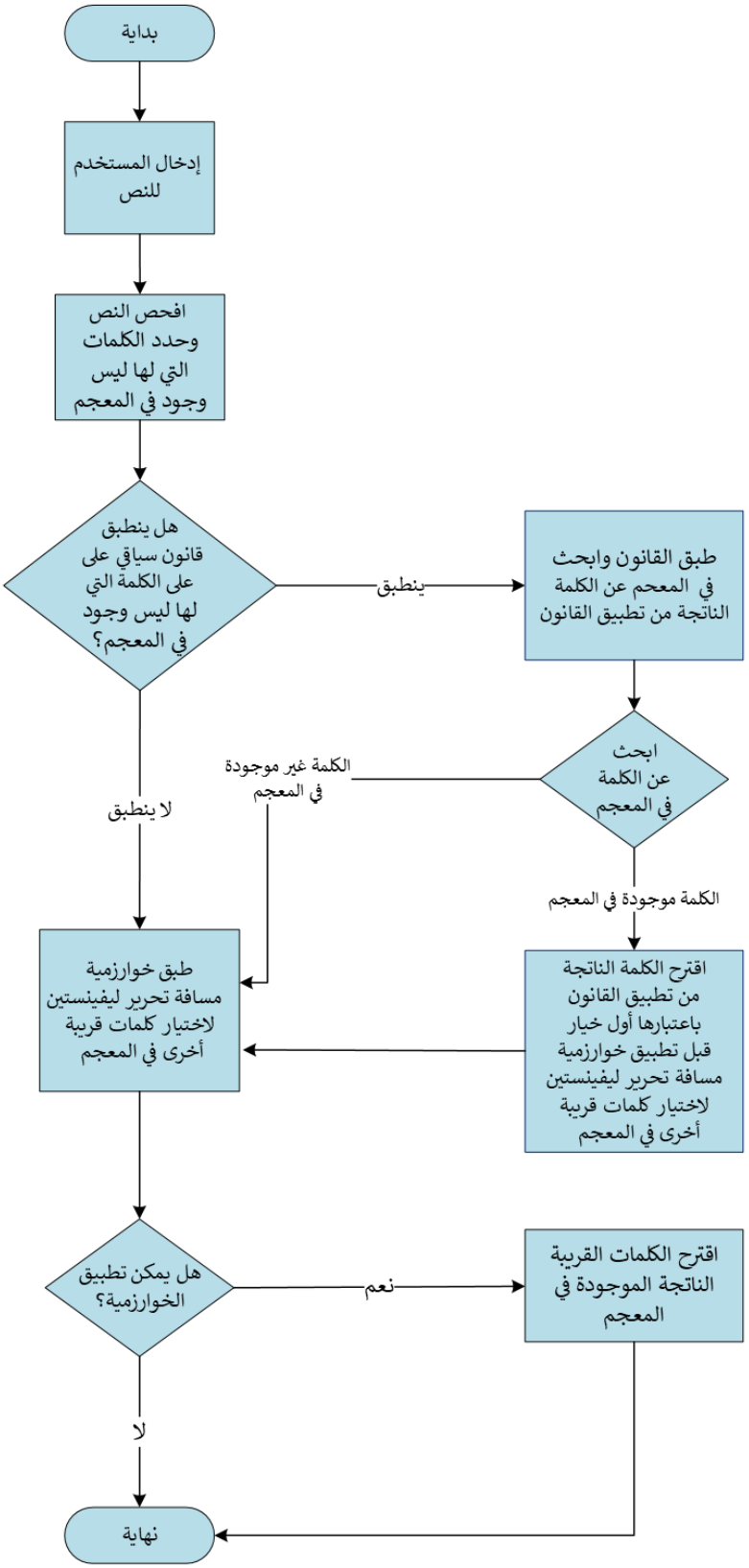
Attia, Mohammed, et al. "A lexical database for modern standard Arabic interoperable with a finite state morphological transducer." *International Workshop on Systems and Frameworks for Computational Morphology*, 2011, Springer, Berlin, Heidelberg, pp. 98-118.

Attia, Mohammed, et al. "A corpus-based finite-state morphological toolkit for contemporary Arabic." *Journal of Logic and Computation*, 2014, 24.2, pp. 455-472.

[٢٢] تتكون المدونة الحاسوبية النصية من 925,461,707 كلمة جُمِعَت من مدونة Arabic Gigword Corpus (Fifth Edition) و ١٦٣,٦٤٩,٤٩٧ كلمة جُمِعَت من موقع الجزيرة (www.aljazeera.net) الإخباري. لمعرفة المزيد عن مدونة Arabic Gigword Corpus، يمكن الرجوع إلى:

University of Pennsylvania. "Arabic Gigaword Fifth Edition." *Linguistic Data Consortium*, 2011, https://catalog.ldc.upenn.edu/LDC2011T11. Accessed 4 December 2021.

وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules). تغطي هذه القوانين الأخطاء التي تتسبب بها التغييرات الصوتية الخاطئة (Phonological Alternations) مثل تغيير الصيغة الصحيحة "تضاعف" إلى الصيغة الخاطئة "تظاعف". كما تتعامل هذه القوانين أيضا مع الأخطاء الكتابية (Orthographic Typos)مثل تغيير الصيغة الصحيحة "المعلمة" إلى الصيغة الخاطئة " المعلمه". إضافة إلى ذلك، تأخذ هذه القوانين بعين الاعتبار الأخطاء الإملائية والصوتية الشائعة والتي قد يتسبب بها تأثير اللهجات العامية مثل تغيير الصيغة الصحيحة "خسائر" إلى "خسساير". من خلال تطبيق هذه القوانين، يقوم النظام باقتراح الكلمة الصحيحة التي تقابل الكلمة الخاطئة بناء على سياق القوانين. تجدر الإشارة إلى أن النظام يقوم بتطبيق القوانين السياقية المذكورة – في حال مطابقة الكلمة الخاطئة لسياق قانون معين – قبل استخدام خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) للبحث عن أقرب كلمات في المعجم تقابل الكلمة الخاطئة المدخلة. يعرض الشكل التالي مخططا (Flow Chart) يبين مسار عمل نظامنا الخاص باكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية:



شكل رقم (4): مخطط (Flow Chart) يبين مسار عمل نظامنا الخاص باكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية

كما هو مبين في الشكل، بعد إدخال المستخدم للنص، يفحص النظام هذا النص ويحدد الكلمات التي ليس لها وجود في المعجم (وهي الكلمات الخاطئة). بعد ذلك يتأكد النظام من إمكانية تطبيق قانون صوتي وإملائي سياقي على الكلمة التي ليس لها وجود في المعجم. هنا يكون لدى النظام مساران فرعيان:

المسار الأول: إذا وجد النظام قانونا سياقيا ينطبق سياقه على الكلمة الخاطئة، فإنه يقوم بتطبيق هذا القانون ثم يبحث عن الكلمة الناتجة من تطبيق هذا القانون في المعجم. إذا وجد النظام الكلمة الناتجة من تطبيق القانون في المعجم فإنه يقوم باقتراح الكلمة الناتجة من تطبيق هذا القانون باعتبارها أول خيار للتصحيح الآلي قبل أن يقوم بتطبيق خوارزمية مسافة تحرير ليفينستين لاختيار كلمات قريبة أخرى. بعدها يقوم بتطبيق خوارزمية مسافة تحرير ليفينستين لاختيار كلمات قريبة أخرى في المعجم ليقترحها على المستخدم وينتهي عمله. أما إذا كانت الكلمة الناتجة من تطبيق القانون السياقي غير موجودة في المعجم، فإنه لا يقترحها على المستخدم، بل يذهب إلى تطبيق خوارزمية مسافة تحرير ليفينستين لاختيار كلمات قريبة أخرى في المعجم ليقترحها على المستخدم وينتهي عمله.

المسار الثاني: إذا لم يجد النظام قانونا سياقيا ينطبق سياقه على الكلمة الخاطئة، فإنه يذهب مباشرة إلى تطبيق خوارزمية مسافة تحرير ليفينستين لاختيار كلمات قريبة أخرى في المعجم ليقترحها على المستخدم وينتهي عمله.

أما إذا لم ينجح النظام في المسارين المذكورين في اختيار كلمات قريبة أخرى في المعجم ليقترحها على المستخدم فإن عمل النظام ينتهي بعدم اقتراح صيغ صحيحة للكلمة الخاطئة بسبب عدم وجود كلمات قريبة منها في المعجم.

كما يشرح الجدول التالي بشكل مبسط طريقة عمل بعض القوانين الإملائية والصوتية السياقية التي تحدثنا عنها. حيث يعرض العمود الأول الحرف أو تتابع الأحرف (Character Sequence) الموجود في الكلمة المدخَلة إلى النظام بشكل خاطئ، ويبين العمود الثاني الحرف أو تتابع الأحرف الصحيح الذي يجب على النظام اقتراحه باعتباره تصحيحا للخطأ الإملائي بشرط وجود الكلمة المقترحة في معجم النظام، أما العمود الثالث فيعرض السياق (Context) الذي يأتي فيه الحرف أو تتابع الأحرف في الكلمة المدخَلة إلى النظام بشكل خاطئ. أخيرا يعرض العمود الأخير مثالا على الأخطاء الإملائية التي يتناولها كل قانون.

جدول رقم (1): شرح مبسط لطريقة عمل بعض القوانين الإملائية والصوتية السياقية

أمثلة	الحرف الصحيح / تتابع الحروف الصحيح	موقع الحرف / تتابع الحروف المكتوب بشكل خاطئ في الكلمة المُدخَلة	الحرف / تتابع الحروف المكتوب بشكل خاطئ
اذا ← إذا اسماعيل ← إسماعيل ارادة ← إرادة	إ	بداية الكلمة [٢٣]	ا

[٢٣] كذلك يأخذ النظام بعين الاعتبار عند تعامله من الأخطاء في الحروف التي تأتي في بداية الكلمة وجود حرف متصل ببداية الكلمة كما في "فاذا" التي تصحح إلى "فإذا".

اهانة ← إهانة			
ارسل ← أرسل	أ	بداية الكلمة	ا
إمرأة ← امرأة إستخرج ← استخرج	ا	بداية الكلمة	إ
أامن ← آمن أازر ← آزر أادم ← آدم	آ	بداية الكلمة	أا
مرءاة ← مرآة	آ	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	ءا
قرأان ← قرآن	آ	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	أا
خسايير ← خسائر[٢٤] نايم ← نائم قايم ← قائم مايل ← مائل	ئ	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	ي
موذي ← مؤذي[٢٥]	ؤ	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	و
عبائة ← عباءة	اء	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	ائ
سيأة ← سيئة	يئ	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	يأ
زيير ← زئير	ئير	ليس في بداية أو نهاية الكلمة (وسط الكلمة)	يير
يكافيء ← يكافئ	ئ	نهاية الكلمة	يء
المعلمه ← المعلمة	ة	نهاية الكلمة	ه
اللة ← الله الإلة ← الإله	ه	نهاية الكلمة	ة
رعات ← رعاة	ة	نهاية الكلمة	ت

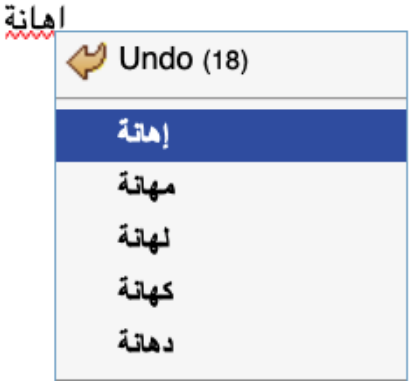
[٢٤] هذا النوع من الأخطاء سببه التأثر باللهجة الكويتية والتي تشتمل على تغيير صوتي متمثل بتحويل صوت الهمزة إلى /ي/ في وسط الكلمة.

[٢٥] هذا النوع من الأخطاء سببه التأثر باللهجة الكويتية والتي تشتمل على تغيير صوتي متمثل بتحويل صوت الهمزة إلى /و/ في وسط الكلمة.

عندكي ← عندك كتابكي ← كتابك	ك	نهاية الكلمة	كي
تفاهمتي ← تفاهمت استخدمتي ← استخدمت	ت	نهاية الكلمة	تي
استعلا ← استعلى	ى	نهاية الكلمة	ا
ظغينة ← ضغينة مستظعف ← مستضعف تظاعف ← تضاعف	ض	في أي موقع	ظ
انتضار ← انتظار	ظ	في أي موقع	ض

3. اختبار نظام التدقيق الإملائي للشبكة العنكبوتية

تبين الشاشات الآتية تطبيقا فعليا لنظامنا الخاص باكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية. حيث تعرض الشاشات كلمات أدخلناها إلى النظام بشكل خاطئ ويظهر لنا أن النظام قد نجح في اكتشاف أن هذه الكلمات غير صحيحة. كما نلاحظ أنه اقترح علينا الكلمات الصحيحة التي يمكن اختيار أحدها باعتباره بديلا صحيحا للكلمة الخاطئة بناء على المعرفة اللغوية التي اكتسبها نظامنا من المعجم الحاسوبي وبناء على القوانين الإملائية والصوتية السياقية.



شكل رقم (5): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "اهانة"

شكل رقم (6): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "إستخرجتم"

شكل رقم (7): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "خسايره"

شكل رقم (8): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "الموزي"

شكل رقم (9): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "عبائات"

شكل رقم (10): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "للمعلمه"

شكل رقم (11): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "كتابكي"

شكل رقم (12): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "استخدمتي"

شكل رقم (13): مثال على استخدام نظامنا لاكتشاف وتصحيح الأخطاء في الإملائية في كلمة "للمستظعف"

4. الخلاصة

قدم هذا البحث نظاما لاكتشاف وتصحيح الأخطاء الإملائية للغة العربية للشبكة العنكبوتية (Web Spell Checker) يمكنه التعامل مع نسبة كبيرة من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة بشكل خاص باستخدام معجم حاسوبي. صُمم هذا المعجم باستخدام قائمة كلمات ضخمة (Word List) مفتوحة المصدر (Open Source). بُنِيَت هذه القائمة باستخدام قاعدة بيانات معجمية (Lexical Database) مفتوحة المصدر مخصصة للتحليل الصرفي (Morphological Analysis) للأسماء والأفعال العربية صُمِّمَت باستخدام تقنية الآلات منتهية الحالات (Finite State Automata). تحتوي قائمة الكلمات المذكورة على الصيغ الصرفية والاشتقاقية (Inflected and Derived Forms) المحتملة لكلمات اللغة العربية الفصحى.

أثبت البحث أنه يمكننا بناء معجم حاسوبي لنظام التدقيق الإملائي باستخدام بيانات معجمية تم جمعها من مدونة نصية حاسوبية (Corpus) ضخمة وهي مدونة مكونة من ١,٠٨٩,١١١,٢٠٤ كلمة. كما أثبت البحث بأنه يمكن تزويد النظام بالقدرة على إعادة ترتيب (Re-Ranking) مقترحات التصحيح الآلي الناتجة من تطبيق خوارزمية مسافة تحرير ليفينستين (Levenshtein Edit Distance Algorithm) المستخدمة في التصحيح الحاسوبي الآلي للأخطاء الإملائية من خلال إعطاء الأولوية لإظهار مقترحات التصحيح الآلي للأخطاء الإملائية الشائعة لدى مستخدمي اللغة العربية وذلك باستخدام قوانين إملائية وصوتية سياقية (Context Sensitive Orthographic and Phonological Rules). حيث استُخدِم المعجم الحاسوبي والقوانين الإملائية والصوتية السياقية المذكورة لتزويد النظام بالمعرفة اللغوية التي تمكنه من اكتشاف وتصحيح الأخطاء الإملائية في نصوص اللغة العربية الفصحى في مواقع الشبكة العنكبوتية.

كما بين البحث التحدي الذي يواجه مستخدمي اللغة العربية بسبب ندرة أنظمة التدقيق الإملائي المخصصة لاكتشاف وتصحيح الأخطاء الإملائية في النصوص المدخلة باللغة العربية إلى مواقع الشبكة العنكبوتية (Web Spell Checking Systems) وهي أنظمة تختلف في طريقة عملها واستخداماتها عن أنظمة التدقيق الإملائي المدمجة مع برمجيات معالجة الكلمات (Word Processors) المعروفة مثل نظام Microsoft Word. حيث إن أغلب الأنظمة التجارية الخاصة بالتدقيق الإملائي للشبكة العنكبوتية تعاني من عدم شمول معجمها الحاسوبي على كمية كافية من الكلمات التي تغطي العربية الفصحى بشكل عام والعربية الفصحى الحديثة بشكل خاص. كما أن أنظمة التدقيق الإملائي المتقدمة المخصصة لاكتشاف وتصحيح الأخطاء الإملائية في النصوص المدخلة باللغة العربية إلى مواقع الشبكة العنكبوتية والتي تستخدمها الشركات العالمية الكبرى مثل أنظمة شركة مايكروسوفت وجوجل غير متاحة تجاريا لاستخدام مواقع الشبكة العنكبوتية الأخرى، حيث إنها أنظمة مخصصة فقط لاستخدامات تلك الشركات الكبرى وليست متاحة لجهات أخرى. هنا تأتي أهمية هذا النظام الذي عرضناه في هذا البحث لسد هذه الفجوة من خلال تقديم نظام للتدقيق الإملائي للغة العربية للشبكة العنكبوتية يستخدم معجما حاسوبيا (Lexicon) ضخما يحتوي على ٩ ملايين كلمة من كلمات اللغة العربية الفصحى. كذلك تتمثل فائدة النظام أيضا في إتاحته للمستخدمين على شكل خدمة ويب سحابية (Cloud Web Service) يمكن دمجها مع أي موقع أو تطبيق متاح على الشبكة العنكبوتية، كما يمكن دمجه مع تطبيقات الأجهزة الذكية وذلك من خلال واجهة لبرمجة التطبيقات (Application Programming Interface) تتيح التدقيق الإملائي لنصوص اللغة العربية المدخلة إلى مواقع وتطبيقات الشبكة العنكبوتية وتطبيقات الأجهزة الذكية.

المراجع

الخولي، محمد. *معجم علم اللغة النظري*. مكتبة لبنان، بيروت، 1982.

مرسي، هيثم زينهم. "المدقق اللغوي الحاسوبي: دراسة نقدية." *مجلة كلية دار العلوم*، كلية دار العلوم، جامعة القاهرة، مصر، ٢٠١٥، العدد ٨٢ ، الصفحات ٥١٩ – ٥٨٦.

Alnajem, Salah. "A computational approach to the variations in Arabic verbal orthography." *Computer Speech & Language*, 2005, 19.3, pp. 275-299.

Attia, Mohammed, et al. "A corpus-based finite-state morphological toolkit for contemporary Arabic." *Journal of Logic and Computation*, 2014, 24.2, pp. 455-472.

Attia, Mohammed, et al. "A lexical database for modern standard Arabic interoperable with a finite state morphological transducer." *International Workshop on Systems and Frameworks for Computational Morphology*, 2011, Springer, Berlin, Heidelberg, pp. 98-118.

Attia, Mohammed, et al. "Arabic spelling error detection and correction." *Natural Language Engineering*, 2016, 22.5, pp. 751-773.

Attia, Mohammed, et al. "Improved spelling error detection and correction for arabic." *Proceedings of COLING 2012*, 2012, Posters, pp. 103-112.

Bhaire, Vibhakti V., et al. "Spell checker." *International Journal of Scientific and Research Publications* , 2015, 5.4 , pp. 5-7.

Damerau, Fred J. "A technique for computer detection and correction of spelling errors." *Communications of the ACM*, 1964, 7.3, pp. 171-176.

Ghafour, Hesham H. Abdel, Ali El-Bastawissy, and Abdel Fattah A. Heggazy. "AEDA: Arabic edit distance algorithm Towards a new approach for Arabic name matching." *The 2011 International Conference on Computer Engineering & Systems*, 2011, IEEE, pp. 307-311.

H. Gueddah and A. Yousfi, "The impact of arabic inter-character proximity and similarity on spell-checking." *2013 8th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 2013, pp. 1-4.

Hamza, Bakkali, et al. "For an independent spell-checking system from the Arabic language vocabulary." *(IJACSA) International Journal of Advanced Computer Science and Applications* , 2014, 5.1.

Hassan, Youssef, Mohamed Aly, and Amir Atiya. "Arabic spelling correction using supervised learning." *arXiv preprint arXiv:1409.8309,* 2014.

Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Upper Saddle River, NJ, Prentice Hall, 2008.

Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet physics doklady, 1966, Vol. 10. No. 8, pp. 707-710.

Mars, Mourad. "Toward a Robust Spell Checker for Arabic Text." *International Conference on Computational Science and Its Applications*, 2016, Springer, Cham, pp. 312-322.

Muaidi, Hasan, and Rasha Al-Tarawneh. "Towards arabic spell-checker based on N-grams scores." *International Journal of Computer Applications*, 2012, 53.3.

Roche, E. & Schabes, Y. "Introduction." *Finite-state language processing*, edited by Roche, E. & Schabes, Y., MIT Press, Cambridge, 1997, pp. 1-66.

Saty, Ahmed Abdalrhman, Karim Bouzoubaa Bouzoubaa, and Aouragh Si Lhoussain. "Survey of Arabic Checker Techniques." *Journal of Engineering and Computer Science (JECS)*, 2020, 21.1, pp. 34-41.

Shaalan, Khaled, Amin Allam, and Abdallah Gomah. "Towards automatic spell checking for Arabic." *Proceedings of the 4th Conference on Language Engineering*, Egyptian Society of Language Engineering (ELSE), 2003, Cairo, Egypt.

Shaalan, Khaled, et al. "Arabic word generation and modelling for spell checking." *LREC*, 2012, pp. 719-725.

University of Pennsylvania. "Arabic Gigaword Fifth Edition." *Linguistic Data Consortium*, 2011, https://catalog.ldc.upenn.edu/LDC2011T11. Accessed 4 December 2021.

مجلة كلية الآداب. جامعة الإسكندرية ، المجلد ٧٢ ، العدد ١١٠، اكتوبر ٢٠٢٢

A spell-checking system for the Arabic language for the World Wide Web using a computational lexicon and orthographic and phonological rules

Dr. Salah Alnajem

College of Arts, Arabic Department
Kuwait University
salah.alnajem@ku.edu.kw

Abstract

This paper presents a system for detecting and correcting spelling errors for the Arabic language for the World Wide Web (Web Spell Checker) that we designed using the WebSpellChecker Engine system. Our system is available to users in the form of a Cloud Web Service that can be integrated with any website or application available on the World Wide Web, and it can also be integrated with smart device applications through an Application Programming Interface (API) that allows spell checking of Arabic texts entered to web applications and smart device applications. Our system can handle a large percentage of words covering Standard Arabic in general and Modern Standard Arabic in particular using a computational lexicon. This lexicon is built using a huge open source word list. This list was built using an open source Lexical Database dedicated to morphological analysis of Arabic nouns and verbs, which was designed using Finite State Automata technique. The mentioned word list contains the possible inflected and derived forms of Standard Arabic words (examples: كَتَبَ، ويكتبان، كتبوا، فسيكتبن، كاتِبة، للكاتِبَين، المكتوب). The system was also provided with the ability to re-rank the automatic correction suggestions resulting from the application of the Levenshtein Edit Distance Algorithm used in the automatic computer correction of spelling errors by giving priority to showing the automatic correction suggestions for common spelling errors among Arabic language users using context sensitive orthographic and phonological rules. The computational lexicon and the mentioned context sensitive orthographic and phonological rules are used to provide the system with linguistic knowledge that enables it to detect and correct spelling errors in Standard Arabic texts entered to web applications.

Keywords: Computational Linguistics, Arabic Language Processing, Spell-Checking, Applied Linguistics, Computational Lexicon.