مشروع جمع المدونات النصية الخاصة بالنصوص الأكاديمية في اللغة العربية

الدكتورة حسلينا حسان

والدكتور محمد فهام محمد غالب

قسم اللغة العربية وآدابها

الجامعة الإسلامية العالمية بماليزيا

ملخص البحث

تمثل هذه الورقة تقريراً عن تطوير مشروعٍ[1] لجمع المدونات النصية الخاصة بالنصوص الأكاديمية في اللغة العربية في إحدى الجامعات الحكومية الماليزية المسمى بـ"المدونات النصية العربية للجامعة الإسلامية العالمية بماليزيا". ويقدر حجم هذه المدونات بحوالي أربعة عشر مليون كلمة منقسمة إلى مجالين هما الدراسات العربية والدراسات الإسلامية، مستمدة من الأوراق العلمية والأبحاث المنشورة في مجلات محكمة أو مؤتمرات ورسائل الماجستير والدكتوراه. تهدف المجموعة لأن تكون مرجعاً للبحث والتطوير اللغوي ولتدريس اللغة العربية وتعليمها وتعلمها.

المقدمة:

يجمع هذا المشروع المنشورات الأكاديمية باللغة العربية تحت مسمى "المدونات النصية العربية للجامعة الإسلامية العالمية بماليزيا"[٢]. تهدف المدونات بشكل عام إلى توفير الأدلة النصية عند وصف استخدام اللغة العربية في ماليزيا وتحديداً لمساعدة غير الناطقين باللغة العربية في تكوين رؤية أوسع عن كيفية استخدام اللغة العربية في المجالات الأكاديمية.

ابتدأ تطوير هذه المدونات في أغسطس ٢٠٠٩ بمواردها المستمدة من مؤسسة عريقة هي الجامعة الإسلامية العالمية بماليزيا المعروفة بالمحافظة على مستوى رفيع من اللغة العربية في منشوراتها. إن النصوص المشمولة في هذه المدونات تعتمد على اللغة العربية الفصحى الحديثة؛ وتأتي من خمسة أقسام: هي مكتبة الجامعة ومركز إدارة الأبحاث وقسم اللغة العربية وآدابها وقسم الفقه وأصوله وقسم القرآن والسنة؛ إذ منحت هذه الأقسام الإذن لتخزين كمية ضخمة من بياناتهم النصية الأساسية، وقامت بتوفير النصوص بشكل إلكتروني. وتمّ الحصول على الإذن بالنشر من أصحاب الحقوق عبر الاجتماع معهم والطلب الرسمي من ملاك قواعد البيانات. تم منح الإذن لاستخدام قواعد البيانات مباشرةً نظراً لأهمية مشروع المدونات النصية كأداة لدعم عملية التدريس والتعلم.

تحتوي المدونات على حوالي أربعة عشر مليون كلمة وهي موزعة إلى مجالين: الدراسات العربية والدراسات الإسلامية. تم تحديد الاختيار في هذين المجالين؛ لأن استخدام العربية كوسيط للتدريس في الجامعة المعنية محدود في الأقسام التي تدرس اللغة العربية وآدابها أو الدراسات الإسلامية فقط، مما يحدد محتويات المدونات بما أنتجته هذه الأقسام. ولهذا السبب تم اعتبار توفر النصوص باللغة العربية وبشكل إلكتروني شرطاً أساسياً لدخول المدونات.

إن النصوص المختارة لهذا المشروع تخاطب الفئة الأكاديمية باتباع المعايير والضوابط المحددة من الجامعة. فالأطروحات الجامعية وأوراق المؤتمرات تتم باتباع

معايير متجانسة ذات أقسام متشابهة كأن تبدأ بملخص تليه مقدمة والدراسات السابقة ومنهجية البحث وأسئلته ونتائج وخلاصة موضوع بحثي معين. وتغطي المقالات المنشورة في مجلات محكمة مجالات أوسع من الأطروحات وأوراق المؤتمرات لاحتوائها مقدمات ونقد تحليلي وتقارير عن الأنشطة الأكاديمية ومراجعات لكتب. تضيف هذه المجالات المتعددة ألواناً أوسع لهذه المدونات خاصة عندما يتم إجراء بحث تجريبي أو عملي عن النسق والأنماط اللغوية[٣].

١. أهمية الدراسة وفوائدها

هناك قائمة طويلة من المجالات اللغوية التي يمكن معالجتها بمنهج يعتمد على المدونات النصية مثل مجال الدراسات المعجمية واللغويات الاجتماعية والنظرية اللغوية واللغويات الحاسوبية والقواعد النحوية والقواميس ودراسة الأساليب وتحليل الأصوات واكتساب الأطفال للغة وعلم اللغة النفسي واللغويات التطبيقية وتصميم المواد الدراسية واختبارات اللغة وعلم الإملاء[٤]. وستكون هذه المدونات مرجعاً غنياً للمتعلمين لتعلم كيفية تقديم الأفكار والاقتراحات والتقارير والنقد والأسئلة بأساليب مختلفة باستخدام مجموعة متنوعة من الألفاظ، حيث توفر المدونات التي نحن بصددها نماذج أو قوالب يمكن لجميع المستخدمين تطبيقها في كتاباتهم.

٢. مشكلة البحث

تم بذل جهود حثيثة لجمع مدونات عربية مستمدة من مجموعات متنوعة من الوسائل والموارد إمّا على نطاق واسع للمدونات اللغوية أو على نطاق مبادرات صغيرة بحسب احتياجات المستخدمين النهائيين. تأتي مصادر المدوناتالموجودة حالياً من الشبكة العنكبوتية (مدونات نصية بك ولتر العربية (Buck Walter 1986- 2003) المدونات النصية ليوفان (Leuven 1990-2004)، وكلارا (Clara1997) والمدونات النصية العربية العلمية العامة ٢٠٠٤ والمدونات النصية للغة العربية الفصحى ٢٠٠٤، ومن الراديو والتلفزيون (برودكاست نيوز سبيتش ٢٠٠٠)، ومن وكالة الأنباء (مجموعة وكالة

الأنباء العربية ١٩٩٤)، والمدونات النصية الصحفية "النهار" ٢٠٠١، والمدونات النصية الصحفية "الحياة" ٢٠٠٢ والمدونات النصية جيجاورد (Gigaword 2002) ومن المتحدثين باللغة العربية بوصفها لغة أم (مجموعة كول فريند ١٩٩٥)، ومن المجلات والروايات (مدونة نجميجن ١٩٩٦(Nijmegen)). يقدم الجدول التالي (١) وصفاً مختصراً للمدونات العربية الموجودة كما أدرجتها سوليتي (Sulaiti)[٥].

جدول (١): وصف للمدونات النصية العربية الموجودة (لطيفة سليتي ٢٠٠٦)

اسم المدونات النصية	المصدر	الوسيلة	الحجم	الهدف	المادة
المدونات النصية العربية بك ولتر (Buckwalter) 1986-2003	تيم بك ولتر	مكتوبة	٢,٥-٣ بليون مليون كلمة	معجم وتحليل صرفي ومشكل آلي	الموارد العامة في الشبكة العنكبوتية
المدونات النصية ليوفان (Leuven 1990-2004)	الجامعة الكاثولوكية ليوفان، بلجيكا	مكتوبة ومنطوقة	٣ مليون كلمة ٧٠٠،٠٠٠ منطوقة	عربي-هولندي هولندي-عربي قاموس المتعلم	مصادر الإنترنت راديو وتلفزيون كتب المدرسة الابتدائية
المدونات النصية لوكالة الأنباء العربية (١٩٩٤)	جامعة فنزويلا إل. دي. سي	مكتوبة	٨٠ مليون كلمة	التعليم وتطور التكنولوجيا	وكالة الصحافة الفرنسية، (زينو) وكالة الأنباء ومطبعة الأمة
المدونات النصية كول فريند (CALLFRIEND 1995)	جامعة فنزويلا إل. دي. سي	حواري	٦٠ حوارا هاتفياً	تطوير تكنولوجيا التعريف باللغة	الناطقون بالعربية "المصرين"
المدونات النصية نجميجن (Nijmegen 1996)	جامعة نجميجن	مكتوبة	أكثر من مليوني كلمة	قاموس:عربي-هولندي هولندي-عربي	مجلات وروايات
المدونات النصية كول هوم (CALLHOME 1997)	جامعة فنزويلا ا ل دي سي	حواري	١٢٠ حواراً هاتفياً	التعرف على الكلام من خطوط الهاتف	الناطقون بالعربية "المصرين"
كلارا (CLARA 1997)	جامعة تشارلز (براغ)	مكتوبة	٥٠ مليون كلمة	لغرض المعجم	الدوريات، الكتب، مصادر الإنترنت من ١٩٧٥-إلى الوقت الحاضر
مصر ١٩٩٩	جامعة جون هوبكينس	مكتوبة	غير معروف	إ م تي	المدونة النصية الموازية من القرآن باللغة الإنجليزية والعربية

خطاب بث الأخبار ٢٠٠٠	جامعة فنزويلا إل دي سي	منطوقة	أكثر من ١٠٠ بث إخباري	التعرف على الخطاب	الأخبار المذاعة من راديو صوت أمريكا
المدونات النصية ٢٠٠٠	جمعية نايمقين سونتل،آي تي بالتنسيق مع جامعة ليون	مكتوبة	١٠ مليون كلمة	معجم، بحث عام معالجة اللغة الطبيعية	غير معروف
المدونات النصية النهار ٢٠٠١	إي. ال. آر. إي	مكتوبة	١٤٠ مليون كلمة	البحث العام	جريدة النهار (لبنان)
المدونات النصية الحياة ٢٠٠٢	إي. ال. آر. إي	مكتوبة	١٨,٦ مليون كلمة	هندسية اللغة واسترجاع المعلومات	جريدة الحياة (لبنان)
أرابيك جيجا وورد (Gigaword 2002)	جامعة فنزويلا إل دي سي	مكتوبة	حوالي ٤٠٠ مليون كلمة	معالجة اللغة الطبيعية، واسترجاع المعلومات ونمذجة اللغة	وكالة الصحافة الفرنسية وكالة الحياة للأنباء وكالة النهار للأنباء وكالة زينو للأنباء
المدونات النصية المتوازية من الإنجليزي للعربي ٢٠٠٣	جامعة الكويت	مكتوبة	٣ مليون كلمة	التعلم والترجمة المعجم	منشورات من المجلس القومي الكويتي
المدونات النصية العلمية العربية العامة ٢٠٠٤	يو.إم.آي.إس.تي، بريطانيا	مكتوبة	١,٦ مليون كلمة	بحث التحليل اللفظي	www.kisr.edu.kw
المدونات النصية العربية الفصحى ٢٠٠٤	يو.إم.آي.إس.تي، بريطانيا	مكتوبة	٥ مليون كلمة	بحث التحليل اللفظي	www.muhaoldith.org & www.alwaraq.com
المدونات النصية سوتيتل (SOTETEL)	سونيتل، آي تي تونس	مكتوبة	٨ مليون كلمة	صناعة المعاجم	الآداب، المواد الصحفية والأكاديمية
المدونات النصية اللغوية ٢٠٠٤	يو.إم.آي.إس.تي، بريطانيا	مكتوبة	١١,٥ مليون كلمة ٢,٥ عربية	ترجمة	تكنولوجيا المعلومات – المواقع المتخصصة – نظم حاسوب وبرمجيات الإنترنت هلب ون بوك
المدونات النصية العربية المعاصرة (سي سي إي) ٢٠٠٤	جامعة ليدز	مكتوبة ومنطوقة	حوالي ١ مليون كلمة	تدريس العربية للأجانب	المواقع والصحف الإلكترونية
داريا بابليون لفانتين للخطاب العربي والنصوص Babylon Levantine (Darpa Arabic Speech and Transcripts 2005)	جامعة فنزويلا إل دي سي	منطوقة	حوالي ٢٠٠٠ مكالمة هاتفية	آلة الترجمة وتعريف الخطاب ونظم الحوار المنطوق	فيشرستايل لمجموعة الحديث الهاتفي

يبدو واضحاً من القائمة المذكورة وجود جهود كبيرة لجمع المدونات النصية العربية، ولكن الوصول للمدونات محدود وتوفرها مادة جاهزة للتدريس محدود جداً أيضاً[٦]. وفي حالة أرابيك جيجا وورد (٢٠٠٢) وداربا بابليون لفانتين لتحليل الصوت العربي والنصوص (٢٠٠٥) وبرودكاست نيوز سبيتش (٢٠٠٠) والمدونات النصية كول هوم (١٩٩٧) والمدونات النصية كول فريند (١٩٩٥) والمدونات النصية لوكالة الأنباء العربية (١٩٩٤) فإن الوصول للمدونات مقتصر على الاستخدام الأكاديمي والمنظمات البحثية والأعضاء المسجلين في جمعية البيانات اللغوية (إل. دي. سي) في بنسلفانيا والجمعية الأوروبية لمصادر اللغة (إي. إل. آر. إي) في باريس[٧].

وهناك أيضاً برمجيات متاحة مجاناً عبر الإنترنت لإنشاء مدونات نصية (اصنعها بنفسك) مثل تيكستات (TextStat) المتاحة باللغات الآتية: الإنجليزية والألمانية والهولندية والبرتغالية والفرنسية والجاليكية، ولكن البرنامج لا يدعم اللغة العربية.

وبالعكس، هناك مدونات نصية إنجليزية متاحة للعموم وسهلة الاستخدام من كل أنحاء العالم عبر الشبكة العنكبوتية مثل: مدونات براون (Brown)، والمدونات العالمية الإنجليزية، ومدونات أكسفورد الإنجليزية، والمدونات الأسكتلندية للنصوص والكلام. كما توجد مدونات متاحة بلغات أخرى مثل: رسائل العمارنة (للأكادية والمصرية... إلخ)، ومدونات بجنخان (المدونات الفارسية)، والمدونات الكرواتية الوطنية، ومدونات الهمشري (فارسية)، ومدونات اليوم الفارسية، والمدونات الروسية الوطنية، وقاموس غريكاي (اليونانية القديمة)، والمدونات النصية النيوآشورية.

تمثل صعوبة استخدام المدونات المتاحة عبر الإنترنت حجزاً كبيراً في ترويج هذه المدونات وتطبيقها في تدريس وتعلم اللغة العربية. يهدف هذا المشروع أن يتبوأ اللغة العربية في مجال اللغويات الحاسوبية وذلك بتوفير المدونات النصية الأكاديمية في اللغة العربية في الأسواق. وجدير بالذكر أن المدونات النصية العربية المعاصرة (سي. سي. أي) المتوفرة مجاناً عبر (الإنترنت) مستمدة من الصحف والمجلات والإذاعة والتلفاز ومواقع (الإنترنت)[٨].

٣. دواعي إنشاء المدونات النصية العربية الأكاديمية

أجري العديد من الدراسات حول إيجابيات المدونات النصية للتدريس والتعلم مما وفر أدلة تطبيقية عن استعمالات اللغة؛ فاستخدام المدونات يجعل اكتشاف استعمالات اللغة أمراً واقعياً للطلاب حيث يصبحون محللين لغويين أكثر فاعلية واستقلالاً[٩]. وعليه؛ فيجب تعريف الطلاب بأنماط نموذجية لاستعمالات اللغة عبر عينات أكاديمية من مصادر متنوعة تعينهم على قيامهم بالكتابة الأكاديمية. وهذا ضروري لأنهم يتعرضون للنصوص الأكاديمية ويكتبون ويقرؤون هذا النوع من الكتابات يومياً[١٠]. هذا ما يبرر إنشاء المدونات الأكاديمية العربية لتخدم الأغراض المذكورة أعلاه حيث تتكون هذه المدونات من الكتابات الأكاديمية. ومن المعروف أن هذا النوع من الكتابات معتمد، كما أنها تتعرض للتحكيم والتقويم- إما من قبل المشرفين أو المدققين أو المحكمين- من جهة المحتوى والأسلوب واللغة؛ الخاصية التي تميزها عن المدونات العادية.

وعند مقارنة المدونات الأكاديمية مع تلك المستمدة من شبكة المعلومات العالمية أو ما يطلق عليه مصطلح المدونات الافتراضية، فإن الباحثين قد عبروا عن قلقهم تجاهها لأسباب عديدة؛ منها لصدقها وثباتها[١١]؛ والعثور على مواد من المدونات النصية عبر (الإنترنت) أمر صعب[١٢]؛ كما لا تخضع المدونات

الافتراضية لعملية تحرير وتدقيق. بالإضافة إلى ذلك، فإن مدونات الشبكة العنكبوتية (المدونات الافتراضية) "غير المنقحة" كما وعبرها- كيلقاريف [١٣] (Kilgarriff) وجرافنستيت (Grefenstette 2003) حيث لا تراعى فيها خواص اللغة مثل القواعد والتهجئة والتناسق من قبل المؤلفين. ومن الملاحظ أن المدونات الافتراضية لا تمر بعملية التنقيح كما تنشر فيها الصور الفاضحة، وذلك بسبب إنشاء هذه المواقع الإلكترونية بغرض التسلية ودون الخضوع لأي قوانين لغوية. وبالمثل، فإن جمع المدونات النصية من الانترنت يستغرق الكثير من الوقت كما يعد تضييعاً للوقت ما لم يستخدم الجامع للمدونات المصدر نفسه في المستقبل[١٤].

يتبين مما سبق ضرورة جمع المدونات النصية المنقحة الصحيحة التي تناسب احتياجات المعلمين والطلبة، التي يمكن استخدامها في الأنشطة الصفية والتي يمكن تطبيقها كذلك في البحوث اللغوية وتطويرها. وعليه؛ فيقوم هذا المشروع بجمع المدونات النصية الأكاديمية مما يعطي نماذج وأدلة عن الاستخدام الأكاديمي للغة المكتوبة تلبيةً للحاجات المذكورة آنفاً.

٤. آليات التنفيذ

لا بد لأي مصمم للمدونات النصية أن يأخذ بالحسبان القطاع المستهدف من المدونة[١٥]. فالقطاع المستهدف في هذه المدونات هو النصوص المنتجة وليست النصوص المستقبَلَة؛ حيث تتمثل الأولى في النصوص المخطوطة والمكتوبة أما الأخيرة فالنصوص المسموعة والمقروءة.

ولغرض الحصول على عينة نموذجية من القطاع المستهدف، تم تعريف وحدة النماذج وإطارها[١٦]. إن وحدة العينات لهذه المدونات النصية هي النصوص العربية الأكاديمية المنشورة في ماليزيا والمنتجة من قبل الجامعة الإسلامية العالمية

بماليزيا من عام ٢٠٠٠ حتى وقتنا الحاضر، وإطارها مجموعة الأبحاث والأوراق العلمية المنشورة في مؤتمرات ومجلات محكمة من الجامعة الإسلامية العالمية بماليزيا.

تم أخذ العينات بطريقة المسح الشامل إذ يأخذ المشروع جميع المصادر المتاحة والنصوص المتوفرة، شريطة أن تكون مكتوبة بالعربية. وقسمت جميع وحدات العينات ضمن إطار العينات إلى ثلاثة أصناف أساسية: الأطروحات الجامعية، وأوراق المؤتمرات، والأبحاث المنشورة في مجلات محكمة.

وبالرغم من أن المدونات مكونة من نصوص مكتوبة فقط، فقد راعى تصميمها قضايا التمثيل الملائم والتصنيف والحجم والعينات النموذجية.

٤.١ تصنيفات المدونات

لأجل المحافظة على أكبر كمية من النصوص، صنفت البيانات إلى مجالين: دراسات اللغة العربية، والدراسات الإسلامية، التي قسمت بدورها إلى ثلاث وسائط: أولاً الأطروحات لمرحلتي الماجستير والدكتوراه، ثانياً أوراق المؤتمرات العالمية، ثالثاً الأبحاث العلمية المنشورة في مجلات محكمة والمنتجة من الجامعة نفسها. ونظراً لطبيعة المدونات، فقد تم اعتماد النصوص الأكاديمية حصرياً. ولذلك فإن منطق اختيار البيانات وتصنيفها غير قابل للتطبيق في هذا المشروع كما اقترحه بايبر (Biber)[١٧]، الجدول (٢) يصف تصنيفات المدونات.

تقدم هذه المجموعات نماذج مثالية للأنماط الموجودة في المدونة كما اقترح بونيلي.(Bonelli) [١٨]

نوع النص	مصنفات تقسيم النص	مصنفات النص	مجال النصوص

اللغويات، والآداب، وتدريس اللغة العربية، والقرآن والحديث، ومقارنة الأديان، والفقه، والتفسير	الماجستير والدكتوراه	الأطروحات	الدراسات العربية والدراسات الإسلامية
اللغويات، والآداب، وتدريس اللغة العربية	مؤتمرات الدراسات العربية	الأوراق العملية للمؤتمرات العالمية	
الفقه والأصول، والقرآن والسنة	مؤتمرات الدراسات الإسلامية		
	مقدمة، والمقالات، والرأي والنقد، ومراجعات الكتب، وتقارير المؤتمرات والندوات، وملخص الرسائل العلمية، وملخص الكتب	المجلات المحكمة	

الجدول (٢) تصنيفات المدونات

ويكمن سبب حصر نطاق المدونات في أقسام اللغة العربية والدراسات الإسلامية في محدودية استخدام اللغة العربية في الجامعة حيث لا تقدم تخصصاتها أو المواد التي تدرس بها إلا في كلية معارف الوحي والعلوم الإنسانية، وتحديداً في الأقسام الأربعة المذكورة، وبالتالي ينحصر سبب اختيار هذين النطاقين بتوافر المواد والنصوص العربية فيهما.

وعلى الرغم من كتابة النصوص جميعها باللغة العربية، فقد كان من المهم أن يتم تقسيمها إلى مجالين: دراسات إسلامية تحوي أبحاثاً وشواهد قرآنية وحديثية، ودراسات لغوية فنية في طبعها.

٤.٢ حجم المدونات النصية

ويبلغ حجم النصوص (٥) خمسة ملايين كلمة تقريباً. ومن حيث عدد كلمات النصوص المجموعة، فقد هيمنت الأبحاث المنشورة في مجلات محكمة على كل التصنيفات الأخرى، حيث يوجد ١,٤٣٧,٤٧٢ كلمة بمعدل ٥٠٠٠ كلمة لكل مقال.

تصنيفات النصوص	تصنيفات تبويب النصوص	عدد النصوص في كل تصنيف	متوسط الكلمات في كل فئة	عدد الكلمات
الأطروحات	الماجستير	٤٦	٢٦،٠٠٠	١١٩١٦٥٦
	الدكتوراه	٩	٦٠،٠٠٠	٥٦٨٨٩٧
أوراق المؤتمرات	مؤتمرات الدراسات العربية	٢٠١	٤٦٠٠	٩٥٣١٣٣
	مؤتمرات الدراسات الإسلامية	١٥٦	٤٦٠٠	٩٥٢٨٠١
المجلات المحكمة	-	٢٩٠	٥٠٠٠	١٤٣٧٤٧٢
المجموع				٤١٥١١٥٨

جدول (٣) حجم النصوص

٣. التمثيلية في المدونات

إن المدونات التي تمثل مجتمع العينة بشكل جيد يمكن أن تمد الباحث بمعلومات وفيرة عن اللغة[١٩]، ولذلك فإنه من الضروري لجمع البيانات التي تمثل الاستعمال الصحيح للغة[٢٠].

ونظراً لأن هذا المشروع يجمع المدونات المتخصصة فإن تمثيل العينات يقاس بدرجة عالية بقرب المفردات في كل العينات، وهذا يخالف المدونات العامة التي تعتمد بقوة على العينات المستمدة من مجالات واسعة ومتعددة وبالتالي تكون مفرداتها عامة ومفتوحة[٢١].

ويعتبر نجاح المدونات بتمثيل كامل تنوعات اللغة أمراً ممتازاً، مع العلم بأن تصميم مدونة تمثل لغة ما أو جزءاً منها يعتبر أمراً مشكلاً لصعوبة معرفة "مدى تمثيل العينة للتنوع الموجود في اللغة"[٢٢].

وبالرغم من أن المصادر مقتصرة على المجال الأكاديمي، فإن تنوعها ملحوظ في المجالات المتنوعة التي تجمعها المدونات. وكما ذكر آنفاً فإن المصادر اشتقت من مجالين: الدراسات العربية والدراسات الإسلامية، مع العلم بأن لكل مجال أقساماً فرعية مختلفة، ففي الدراسات العربية وحدها يوجد مجالات منفصلة للتدريس وللآداب وللغة.

٤.٣ عينات المدونات النصية

تعرض أيّ مدونات كيفية استعمال اللغة في سياقها الطبيعي، ولهذا شملت هذه المدونات كافة محتويات الوثائق المشمولة ضمن العينات المختارة؛ مما يتيح مجالاً أوسع للدراسات اللغوية مقارنة بالمدونات المستندة على قطع مجتزأة من النصوص المجموعة[٢٣].

علاوة على ذلك، تم اعتبار دقة لغة العينة أكثر المعايير أهمية للاختيار في مدونات الجامعة الإسلامية العالمية بماليزيا. ولم يجر تدقيق لغوي بعد اختيار العينة لخضوع جميع النصوص لتمحيص لغوي وتدقيق تحريري قبل نشرها. فأي أطروحة علمية يجب أن يوافق عليها المشرف والقارئ الداخلي والقارئ الخارجي

قبل تقديمها للمناقشة (مع العلم أن سلامة اللغة والإملاء من شروط قبول الأطروحات). وهذا يؤكد وثوقية العينات المختارة من جهة دقتها اللغوية سواءً أكانت اللغة العربية هي اللغة الأم لكتّاب النصوص أم كانت اللغة الثانية. ويعتبر إدخال كتّاب -ممن لم تكن العربية لغتهم الأم في المدونة- أمراً إيجابياً يضمن تنوع اللغة المستخدمة.

وبالإضافة إلى ذلك فقد مرّت جميع نصوص المدونة بثلاث مراحل قبل أن تقبل كبيانات: تصفية البيانات، وتحويل البيانات، والتعرف على البيانات.

٤.٣.١. تصفية البيانات

أجريت عملية تصفية البيانات لضمان أن العينات ذات أهمية للتحليل اللغوي وتعلم اللغة. ولذلك تم تجريد المدونات من المحتوى والرموز غير ذات الأهمية للمعنى اللغوي. فعلى سبيل المثال تفرض الجامعة بنية محددة للرسائل الجامعية تشمل ملخص البحث باللغة الإنجليزية والعربية وقائمة المحتويات والموافقة واتفاقية حقوق النشر والإهداء والمراجع والفهرس وقائمة الجداول وقائمة الأشكال وقائمة الاختصارات والرموز المستخدمة وقاموس المصطلحات وصفحة الشكر؛ فتم الاستغناء عنها. وإن كان تاريخ النصوص يعود إلى التسعينيات عندما كانت الحواشي والهوامش أمراً لا غنى عنه، فإن هذه المعلومات الإضافية كذلك لا تعتبر جزءاً من بيانات المشروع.

وبما أن عينات المدونات النصية استهدفت النصوص العربية، فقد تم التخلص من الملخصات الإنجليزية مثل ما تم التخلص من قائمة الأشكال والاختصارات والرموز والمصطلحات التي لا تتمتع بأي دلالة نصية هامة. أما

بالنسبة إلى قائمة الكلمات فتم تجريدها لأنها تكرر ما يرد في متون النصوص المختارة.

٤.٣.٢. تحويل البيانات

تتكون النصوص المجموعة من نوعين من الملفات الإلكترونية: (مايكروسوفت وورد وبي دي إف PDF)، وقد تم تحويلها جميعاً إلى ملفات نصية لغرض التحليل اللغوي.

٤.٣.٣. تعريف البيانات

أعطيت جميع العينات اسماً وترويسةً على الصفحة الأولى من كل ملف لتسهيل عملية التعرف على البيانات وإدارة الملفات.

يحتوي اسم الملف على رقم تعريفي للملف واسم المؤلف والمصدر والنطاق وسنة النشر، باتباع الصيغة التالية:

رقم تعريف الملف/ المؤلف/ سنة النشر/ البيانات الفرعية/ النوع/ التصنيف الفرعي. مثال:

0001 HAZEM MOUHIEDIN2007 PHD_ISLAMIC_QURAN

وتم عرض معلومات إضافية في ترويسة كل ملف، تظهر العنوان والتخصص العام والتخصص الدقيق فوق اسم الملف.

تتبع الترويسة الصيغة التالية:

رقم تعريف الملف/ المؤلف/ سنة النشر/ البيانات الفرعية/ النوع/ التصنيف الفرعي. مثال:

حازم زكريا محي الدين/٢٠٠٧/ رسالة دكتوراه/ مكتبة الجامعة/ الجامعة الإسلامية العالمية بماليزيا/ الدراسات الإسلامية/٠٠٠١ / الحديث.

تـم إيجـاد مخـزون يحـوي تفاصـيل كـل العينـات ويعمـل بمثابـة أداة لإدارة المدونات بحفظ تغيرات حجمها، كأن يراقب عدد الكلمات في المدونات، ويعطي تصوراً كاملاً للمدونات. يتم تسجيل المخزون باستخدام مايكروسوفت إكسل وباتباع الصيغة التالية:

رقم تعريف الملف/ المؤلف/ اسم الملف/ العنوان/ السنة/ التخصص/ عدد الكلمات/ السنة/ المجال/ التخصص/ عدد الصفحات/ عدد الكلمات/ ملاحظات.

الخلاصة:

كانت عملية جمع المدونات النصية الخاصة بالنصوص الأكاديمية في اللغة العربية مجهدة واستغرقت عملاً ووقتاً طويلاً، ولكن النتائج كانت رائعة؛ فهذه المدونات أول مرجع من نوعه في ماليزيا لخدمة التدريس والتعلم والبحث في مجال اللغة العربية.

وبما أن المدونات مكونة من نصوص عربية أكاديمية فإنها لا توفر بحثاً عاماً عن التنوع في اللغة العربية ككل أو في مجالات سوى الدراسات العربية والإسلامية. والأمل معقود أن تشمل هذه المدونة في المستقبل القريب نصوصاً أكاديمية وغير أكاديمية متاحةً بصورة إلكترونية وغير إلكترونية في ماليزيا.

المصادر والمراجع

Atkins, S., Clear, J., Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1). UK: Oxford University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing,* 8: 1-15. doi: 10.1093/llc/8.4.243.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Biber, R. R. (2002). What Does Frequency Have to Do with Grammar Teaching. *Studies in Second Language Acquisation (24)* , 199-207.

Biber, S. C. (2001). Quantitave Corpus-Based Research: Much More Than Bean Counting. *Teachers of English to Speakers of Other Languages (TESOL) Quarterly* , 331-336.

Breyer, Y. (2009). *Learning and teaching with corpora: reflections by student teachers, Computer Assisted Language Learning.* Cambridge: Cambridge University Press.

Cermakova, W. T. (c2007). *Corpus Linguistics :a short introduction.* London and New York: Continuum.

Conrad, S. (1999). The Importance of corpus-based reasearch for language teachers. *System* , 1-18.

Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching, *The Internet TESL Journal*, 14(3).

Fox, G. (2001). Using corpus data in the classroom. In Tomlison, B. *Materials development in language teaching.* Cambridge: Cambridge University Press.

Hadley, G. (2002). An Introduction to Data-Driven Learning. *RELC Journal 33(2)* , 99-124.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge : Cambridge University Press.

Kilgarriff, A. , Gregory, G. (2003). Introduction to the Special Issue on Web as Corpus. *Computational Linguistics,* 29 (3).

Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching. Retrieved November 8, 2009, from *The Internet TESL Journal*, Vol. IX, No. 3. Retreived from

http://iteslj.org/Articles/Krieger-Corpus.html

Krishnamurthy, W. T. (2007). *Corpus Linguistics : Critical Concepts in Linguistics*. London: Routledge.

Leech, G. (1997). Teaching and Language Corpora. In S. F. Anne Wichmann (Ed.), *Teaching and Language Corpora,* (343- 360). United States of America: Longman.

McEnery, T., Andrew Wilson, A. (2001). *Corpus Linguistics.* Edinburgh: Edinburgh UP.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In M. S. Thomas, *Using Corpora for Language Research* (232-247). New York: Longman Group Limited.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford : Oxford University Press.

Sinclair, J. (2004). *Developing Linguistic Corpora:a Guide to Good Practice.* Retrieved May 4, 2009, available from

http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

Sinclair, J. (2004). How to use corpora in language teaching. Amsterdam: John Benjamins.

Sulaiti, L., Atwell, E. (2006). The Design of a Corpus of Contemporary Arabic. Journal: *International Journal of Corpus Linguistics*, 2(7). Amsterdam: John Benjamins.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*: *Studies in Corpus Linguistics*. Amsterdam: John Benjamins

McEnery, T., Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T, R. X. (2006). *Corpus-Based Language Studies.* New York: Routledge.

Michael, W. (2006). Compiling Corpora for use as Translation Resources. *Translation Journal*, 10(1). Available from http://accurapid.com/journal/35corpus.htm

Wynne, M. (2004). *Developing Linguistic Corpora:a Guide to Good Practice.* Retrieved May 4, 2009, available from

http://icar.univ-lyon2.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf

Varantola, Krista (2003). "Translators and Disposable Corpora", in Michael, W. (2006). Compiling Corpora for use as Translation Resources. *Translation Journal*, 10(1). Available from http://accurapid.com/journal/35corpus.htm

Zanettin, F. (2002). DIY Corpora: The WWW and the Translator. In Belinda, M., Jonathan, H., Margherita, U. (eds.). *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras, Universidade do Porto, pp 239-248. Available from http://www.federicozanettin.net/DIYcorpora.htm.

الهوامش

١. حاز المشروع على الميدلية الذهبية في معرض الاختراعات والإبداعات الذي نظمته الجامعة الإسلامية العالمية بماليزيا في ٢١-٢٢ فبراير ٢٠١٢.

٢. المدونة النصيةcorpus تجمع corpora مدونات ويعرّفها (سينكلير، ١٩٩١) بمجموعات نصية لغوية طبيعية اختيرت لتميز حالة أو تنوع لغة ما. مصطلح "المدونة corpus" مشتق من كلمة لاتينية تعني (جسم)، لذلك فإن أي جسم أو مجموعة نصية تعتبر مدونة. كما يتم تخزين المدونات إلكترونياً لسهولة استرجاعها، وهذه خاصية تفيد التحليل اللغوي عبر برامج تحليل النصوص (مكانري وغيره، ٢٠٠١).

3 . Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

٤. نفسه.

5. Sulaiti, L., Atwell, E. (2006). The Design of a Corpus of Contemporary Arabic. Journal: *International Journal of Corpus Linguistics*, 2(7). Amsterdam: John Benjamins.

6 . Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers, Computer Assisted Language Learning. Cambridge: Cambridge University Press.

٧. سليتي، ٢٠٠٦.

8. Sulaiti, L., Atwell, E. (2006). The Design of a Corpus of Contemporary Arabic. Journal: *International Journal of Corpus Linguistics*, 2(7). Amsterdam: John Benjamins.

9 . Krieger, D. (2003). Corpus Linguistics: What It Is and How It Can Be Applied to Teaching, *The Internet TESL Journal*, 14(3)

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*: *Studies in Corpus Linguistics*. Amsterdam: John Benjamins.

10 . Sinclair, J. (2004). *Developing Linguistic Corpora:a Guide to Good Practice*. Retrieved May 4, 2009, available from http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

11 . Varantola, K. (2003). "Translators and Disposable Corpora", in Michael, W. (2006). Compiling Corpora for use as Translation Resources. *Translation Journal*, 10(1). Available from http://accurapid.com/journal/35corpus.htm
Zanettin, F. (2002). DIY Corpora: The WWW and the Translator. In Belinda, M., Jonathan, H., Margherita, U. (eds.). *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras, Universidade do Porto, pp 239-248. Available from http://www.federicozanettin.net/DIYcorpora.htm.

12 . Varantola, K (2003).

13 .Kilgarriff, A. , Gregory, G. (2003). Introduction to the Special Issue on Web as Corpus. *Computational Linguistics*, 29 (3).

14 . Zanettin, F. (2002). DIY Corpora: The WWW and the Translator. In Belinda, M., Jonathan, H., Margherita, U. (eds.). *Training the Language Services Provider for the New Millennium*, Porto: Faculdade de Letras, Universidade do Porto, pp 239-248. Available from 14 Biber, D. (1993).

15. http://www.federicozanettin.net/DIYcorpora.htm.

16 .McEnery, T., Wilson, A. (2001). *Corpus Linguistics*. Edinburgh: Edinburgh UP.

17 .Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8: 1-15. doi: 10.1093/llc/8.4.243.

18 . Tognini-Bonelli, E. (2001).

19 .Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

20. Sinclair, J. (2004). How to use corpora in language teaching. Amsterdam: John Benjamins.

21 . McEnery, T; R. X. (2006). *Corpus-Based Language Studies.* New York: Routledge.

22 .Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing,* 8: 1-15. doi: 10.1093/llc/8.4.243.

23 . Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford : Oxford University Press.