
# Building the Arabic Learner Corpus and a System for Arabic Error Annotation

Abdullah Yahya G. Alfaifi

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

May 2015

# بناء المدونة اللغوية لمتعلمي اللغة العربية مع نظام لوسم الأخطاء اللغوية

عبدالله بن يحيى الفيفي

قدمت هذه الأطروحة وفقاً لمتطلبات الحصول على درجة الدكتوراه في الفلسفة

**جامعة ليدز**

**قسم الحاسب الآلي**

مايو ٢٠١٥م

## ملخص الأطروحة باللغة العربية

لا يخفى على الباحثين في مجال المدونات اللغوية ذلك التطور الملموس في الآونة الأخيرة لمدونات المتعلمين، فمن خلاله بَرزَ الدورُ المتنامي الذي يلعبه هذا النوع في المجالات البحثية اللغوية والحاسوبية، ومنها تعليم اللغة، وتحليل الأخطاء اللغوية، وبناء معاجم الطلاب، ومعالجة اللغة الطبيعية، والكشف عن الأخطاء اللغوية وتصحيحها آلياً، وغير ذلك من المجالات؛ لكن الافتقار إلى مدونة لمتعلمي اللغة العربية مصممة وفق معايير دقيقة أدى إلى قصور في الدراسات التطبيقية المعتمدة على هذه المدونات في المجالات البحثية آنفة الذكر؛ ولذا تهدف هذه الأطروحة إلى تقديم منهجية جديدة وأصيلة لبناء مدونة معيارية لمتعلمي اللغة العربية، هذه المنهجية تتضمن مجموعة من الموارد والمعايير والأدوات التي تم تطويرها لمشروع المدونة اللغوية لمتعلمي اللغة العربية Arabic Learner Corpus.

من الموارد التي تقدمها الدراسة مراجعةٌ علميةٌ لأكثر من مئة وخمسين مدونة من مدونات المتعلمين الدولية، وهي مبنية على مجموعة من المقاييس التي تمثل أسس التصميم لهذه المدونات، حيث تقدم هذه المراجعة رؤية شاملة لهذا المجال، إضافة إلى التوجهات الحديثة في مدونات المتعلمين؛ ومن الموارد أيضاً المواد المكتوبة والمنطوقة التي تقدمها المدونة اللغوية لمتعلمي اللغة العربية، وهي أكبر مدونة مبنية وفق منهجية علمية لمتعلمي اللغة العربية حسب المراجعة العلمية التي أجراها الباحث لمدونات المتعلمين العربية؛ وتشمل الموارد أيضاً تصنيفاً جديداً للأخطاء اللغوية Error Tagset of Arabic، صُمم لوسم الأخطاء في نصوص متعلمي العربية، وهو يغطي خمسة مجالات عامة (الأخطاء الإملائية، والصرفية، والنحوية، والدلالية، وأخطاء علامات الترقيم)، ويقع تحت هذه المجالات أنوع فرعية تصل في مجموعها إلى تسعة وعشرين نوعاً.

تقدم الأطروحة عدداً من المعايير، ومنها الدليل الإرشادي لمعايير بناء مدونات المتعلمين، والذي بُني على مراجعة دقيقة لأكثر من مئة وخمسين مدونة من هذا النوع، ويركز هذا الدليل على أحد عشر معياراً لتصميم مدونات المتعلمين؛ وتشمل المعاييرُ كذلك منهجيةً لتحويل النصوص التي حررها الطلاب إلى شكل حاسوبي مع المحافظة على أعلى قدر من الاتساق بينها.

أما الأدوات التي يقدمها البحث، فتشمل أداةً لوسم الأخطاء اللغوية بمساعدة الحاسب Computer-Assisted Error Annotation، والتي توفر مجموعة من الوظائف العملية للمساعدة على وسم الأخطاء، مثل وظيفة التحديد الذكي Smart Selection، ووظيفة الوسم الآلي Auto Tagging؛ كما يقدم المشروع أداة أخرى للبحث في نصوص المدونة على شبكة المعلومات الدولية "الإنترنت" (www.alcsearch.com)، وهذه الأداة تمكِّن المستخدم من البحث في نصوص المدونة وفق مجموعة من المحددات، مع السماح له بتنزيل هذه النصوص على أكثر من هيئة (TXT، وXML، وPDF، وMP3).

أسهم في نجاح هذه المشروع – بعد فضل الله تعالى – أكثر من ٩٩٠ شخصاً من أكثر من ٣٠ مؤسسة تعليمية في المملكة العربية السعودية والمملكة المتحدة، أغلبهم من متعلمي اللغة العربية، ومنهم من أسهم في جمع نصوص المدونة، أو في وسم الأخطاء في عينة من النصوص، أو في تقويم منهجية الوسم، أو قدم إسهاماً كان له دور ملموس في نجاح هذا المشروع، ولهؤلاء جميعاً وافر الشكر والعرفان.

وإن مما يسلط الضوء على أهمية هذا المشروع ذلك الاستخدام الواسع لنصوص المدونة منذ إنشائها في عدد من المشاريع البحثية النظرية والتطبيقية، ومنها مجموعة من الدراسات في مجال اكتشاف الأخطاء وتصحيحها آلياً، وتقييم المحللات الصرفية للغة العربية، واكتشاف اللغة الأم لكاتب النص، وكذلك مجموعة من الدراسات في مجال اللغويات التطبيقية، والتعليم الموجه بالبيانات Data-Driven Learning، وغيرها.

## ترجمة لآراء المناقشين وبعض خبراء المدونات اللغوية في الأطروحة ومشروع المدونة

> تهتم أطروحة عبدالله الفيفي ببناء المدونة اللغوية لمتعلمي اللغة العربية التي تعد مدونة حديثة في مجال المدونات اللغوية واللغويات الحاسوبية؛ قيّمت هذه الأطروحة المتميزة المدونات القائمة لمتعلمي اللغة العربية إضافة إلى عدد كبير من مدونات المتعلمين الدولية المبنية للغات الأخرى، وقد قدمت نموذجاً متقدماً وفريداً في ذلك؛ وهذا يدل على تطبيقات وآثار كبيرة لمشروع المدونة في مجال تعليم اللغة العربية وتعلمها

**Internal examiner of the thesis**
**Prof. Janet Watson**
*School of Languages Cultures & Societies*
*University of Leeds*
*United Kingdom*

> لقد قدم عبد الله الفيفي أطروحة استثنائية رفعت من سقف المنافسة في مجال البحث العلمي، وهذه الأطروحة لم تمهد الطريق للبحث في مدونات المتعلمين العربية فحسب، بل تجاوزت ذلك إلى مجال البحث في المعالجة الآلية للغة العربية بشكل عام، وقد كان لاجتهاده واهتمامه بالتفاصيل دور واضح في هذه الأطروحة المتميزة، وأود أن أوصي أي باحث دكتوراه، أو طالب مستجد في مجال المعالجة الآلية للغة العربية بأهمية قراءة هذه الأطروحة

**External examiner of Alfaifi thesis**
**Dr. William Teahan**
*School of Computer Science*
*University of Wales at Bangor*
*United Kingdom*

"إن المدونة اللغوية لمتعلمي اللغة العربية هي مدونة حديثة في مجالها، ومن المتوقع أن تسلط الضوء على جوانب جديدة في اللغة المرحلية لمتعلمي العربية، ويبرز دور هذا المشروع عند النظر إلى القصور الكبير في مدونات المتعلمين العربية الأخرى، حيث إنه لا يمكن إغفال القيمة الأكاديمية والإسهام العلمي لهذه المدونة في مجال البحث العلمي؛ إن التحليل الدقيق لتصاميم أغلب مدونات المتعلمين حول العالم مع مقارنة نقاط القوة والضعف قد مكّن عبدالله الفيفي من إنشاء إجراءات مفصلة ودقيقة لجمع مواد المدونة المكتوبة والمنطوقة، وهذا في اعتقادي أدى إلى موثوقية عالية في البيانات التي جُمعت في المدونة، وباعتباري أحد الباحثين في مجال مدونات المتعلمين أود أن أقدم التهنئة على إتمام بناء هذا المشروع"

**Prof. Shin Ishikawa**
*School of Languages and Communication*
*Kobe University*
*Japan*

"أغلب الأبحاث في مجال معالجة اللغة الطبيعية واللغويات الحاسوبية تتأثر بشكل رئيس بالموارد اللغوية المتوفرة لها، ومنها المدونات اللغوية، والبنوك اللغوية الشجرية، وغيرها من أنواع البيانات الموسومة؛ وهذه الموارد ذات القيمة العالية يكلّف بناؤها الكثير وتستهلك وقتاً طويلاً، كما أن تطويرها يحتاج إلى دقة كبيرة لتحقيق أقصى قدر من الفائدة لأكبر عدد ممكن من المجالات البحثية؛ وفي العقد الماضي تزايد الاهتمام باللغويات الحاسوبية العربية بشكل كبير، ورغم وجود بعض الجهود لبناء مدونات لمتعلمي اللغة العربية، إلا أن حجمها المحدود قلل من فائدتها، ولذا لم يجد الباحثون مصدراً لغوياً موسوماً بشكل شامل ودقيق إلى أن أُنشئت المدونة اللغوية لمتعلمي اللغة العربية؛ فحجم المدونة والوسم التفصيلي الذي قام به عبدالله الفيفي جعل من المدونة اللغوية لمتعلمي اللغة العربية مورداً لغوياً ذا أهمية بالغة، وله أثر كبير على تقنيات حوسبة اللغة العربية (مثل تصحيح الأخطاء اللغوية)، وأود أن أحيّي الجهد المبذول في هذا المشروع، وأن أدعم توسعه في المستقبل"

**Dr. Nizar Habash**
*Computer Science*
*New York University Abu Dhabi*
*United Arab Emirates*

"تحوي المدونة اللغوية لمتعلمي اللغة العربية التي أنشأها عبدالله الفيفي مجموعة من المواد المكتوبة والمنطوقة والتي حررها متعلمو اللغة العربية الناطقون بعدد كبير من اللغات الأم المختلفة؛ وهذه المدونة مناسبة جداً لتحليل الأخطاء اللغوية لمتعلمي العربية، لأنها تسمح للباحثين بتمييز الأخطاء اللغوية وفقاً للعديد من الفئات المحددة بدقة، وهذه المدونة ليست مفيدة للباحثين في مجال تحليل أخطاء متعلمي اللغة العربية فحسب، بل لأي باحث في مجال تعليم اللغة العربية وتعلمها"

**Prof. James Dickins**
*School of Arabic, Middle Eastern and East Asian Studies*
*University of Leeds*
*United Kingdom*

لقد وجدتُ أن المدونةَ اللغوية لمتعلمي اللغة العربية مصممةٌ بشكل متميز، وأن نصوصَها مجموعةٌ بطريقة منهجية؛ وأكثر ما شدني في تصميمها أنها تجمع بين متعلمي اللغة العربية من المستويين الجامعي وما قبل الجامعي، وتجمع كذلك الناطقين بالعربية والناطقين بغيرها، وهذا ما جعلها مناسبة لكثير من الدراسات اللغوية المقارنة؛ إضافة إلى ذلك فإن المدونة توفر بيانات وصفية شاملة مما يدل على أن المدونةَ مبنيةٌ وفق تصميم دقيق

**Prof. Yukio Tono**
*Graduate School of Global Studies*
*Tokyo University of Foreign Studies*
*Japan*

# Abstract

Recent developments in learner corpora have highlighted the growing role they play in some linguistic and computational research areas such as language teaching and natural language processing. However, there is a lack of a well-designed Arabic learner corpus that can be used for studies in the aforementioned research areas.

This thesis aims to introduce a detailed and original methodology for developing a new learner corpus. This methodology which represents the major contribution of the thesis includes a combination of resources, proposed standards and tools developed for the Arabic Learner Corpus project. The resources include the *Arabic Learner Corpus*, which is the largest learner corpus for Arabic based on systematic design criteria. The resources also include the Error Tagset of Arabic that was designed for annotating errors in Arabic covering 29 types of errors under five broad categories.

The *Guide on Design Criteria for Learner Corpus* is an example of the proposed standards which was created based on a review of previous work. It focuses on 11 aspects of corpus design criteria. The tools include the *Computer-aided Error Annotation Tool for Arabic* that provides some functions facilitating error annotation such as the smart-selection function and the auto-tagging function. Additionally, the tools include the *ALC Search Tool* that is

developed to enable searching the ALC and downloading the source files based on a number of determinants.

The project was successfully able to recruit 992 people including language learners, data collectors, evaluators, annotators and collaborators from more than 30 educational institutions in Saudi Arabia and the UK. The data of the Arabic Learner Corpus was used in a number of projects for different purposes including error detection and correction, native language identification, Arabic analysers evaluation, applied linguistics studies and data-driven Arabic learning. The use of the ALC highlights the extent to which it is important to develop this project.

# Contents