



UNIVERSITY OF LEEDS

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 الفيحي

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 برز الدور 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 لتعليمي 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 لتعليمي اللغة العربية، هذه المنهجية تتضمن مجموعة من الموارد والمعايير والأدوات التي تم تطويرها لمشروع المدونة اللغوية لتعليمي اللغة العربية Arabic Learner Corpus.

من الموارد التي تقدمها الدراسة مراجعةً علميةً 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 لتعليمي اللغة العربية، وهي أكبر مدونة مبنية وفق منهجية علمية لتعليمي 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 أنواع 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 المعايير 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 أداةً أخرى للبحث في نصوص المدونة على شبكة المعلومات الدولية "الإنترنت" (www.alcsearch.com)، وهذه الأداة تمكّن المستخدم من البحث في نصوص المدونة وفق مجموعة من المحددات، مع السماح له بتنزيل هذه النصوص على أكثر من هيئة (MP3، PDF، XML، وTXT).

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 لأراء المناقشين وبعض خبراء المدونات اللغوية في الأطروحة ومشروع المدونة

” تهتم أطروحة عبدالله الفيبي ببناء المدونة اللغوية لتعلمي اللغة العربية التي تعد مدونة حديثة في مجال المدونات اللغوية واللغويات الحاسوبية؛ قيّمت هذه الأطروحة المتميزة المدونات القائمة لتعلمي 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 الفيقي 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 لتعليمي اللغة العربية هي مدونة حديثة في مجالها، ومن المتوقع أن تسلط الضوء على جوانب جديدة في اللغة المرحلية لتعليمي 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 مكن عبدالله الفيافي 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 يكلف 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 لتعليمي اللغة العربية، إلا أن حجمها المحدود قلل من فائدتها، ولذا لم يجد الباحثون مصدراً لغوياً موسوماً بشكل شامل ودقيق إلى أن أنشئت المدونة اللغوية لتعليمي اللغة العربية؛ فحجم المدونة والوسم التفصيلي الذي قام به عبدالله الفيافي جعل من المدونة اللغوية لتعليمي اللغة العربية مورداً لغوياً ذا أهمية بالغة، وله أثر كبير على تقنيات حوسبة اللغة العربية (مثل تصحيح الأخطاء اللغوية)، وأود أن أحيي الجهد المبذول في هذا المشروع، وأن أدمع 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 المدونة اللغوية لتعلمي اللغة العربية مصممةً بشكل متميز، وأن نصوصها مجموعةً 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 المدونة مبنيةً 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

Publications	viii
Acknowledgements	xi
Abstract.....	xii
Contents	xiv
List of Tables	xxiv
List of Figures.....	xxviii
List of Abbreviations	xxxiv
Part I Introduction and Literature Review	1
1 Introduction.....	2
1.1 <i>Corpus</i> and <i>Learner Corpora</i>	3
1.1.1 The Term <i>Corpus</i>	3
1.1.2 <i>Learner Corpora</i>	3
1.2 Importance of Learner Corpora	4
1.3 Motivation and Aim.....	5
1.4 Objectives	6
1.5 Thesis Contributions	7
1.6 Structure and Scope of the ALC Project.....	10
1.7 ALC Participants.....	12
1.8 Thesis Outline	13
2 Literature Review and Related Work.....	17
2.1 Introduction.....	18
2.2 Literature Review of Learner Corpora.....	18
2.2.1 Purpose	27
2.2.2 Sizes.....	30

Contents

2.2.3	Target Language	35
2.2.4	Data Availability.....	37
2.2.5	Learners' Nativeness	39
2.2.6	Learners' Proficiency Level	40
2.2.7	Learners' First Language.....	42
2.2.8	Material Mode	43
2.2.9	Material Genre	45
2.2.10	Task Type	47
2.2.11	Data Annotation.....	48
2.3	Recommended Design Criteria to Develop a New Learner Corpus	53
2.3.1	Corpus Purpose.....	53
2.3.2	Corpus Size.....	54
2.3.3	Target Language	54
2.3.4	Availability	54
2.3.5	Learners' Nativeness	55
2.3.6	Learners' Proficiency Level	56
2.3.7	Learners' First Language.....	56
2.3.8	Material Mode	56
2.3.9	Material Genre	57
2.3.10	Task Type	57
2.3.11	Data Annotation.....	57
2.4	Related Work: Arabic Learner Corpora.....	57
2.4.1	Pilot Arabic Learner Corpus (Abuhakema <i>et al.</i> , 2009).....	58

2.4.2	Malaysian Corpus of Arabic Learners (Hassan and Daud, 2011).....	58
2.4.3	Arabic Learners Written Corpus (Farwaneh and Tamimi, 2012).....	59
2.4.4	Learner Corpus of Arabic Spelling Correction (Alkanhal <i>et al.</i> , 2012).....	59
2.5	Rationale for Developing the Arabic Learner Corpus	60
2.6	The ALC's Contribution Compared to the Existing Arabic Learner Corpora	62
2.7	Conclusion	64
	Part II Arabic Learner Corpus	66
3	ALC Design and Content	67
3.1	Introduction.....	68
3.2	ALC: Design Criteria and Content	68
3.2.1	Purpose	68
3.2.2	Size	69
3.2.3	Target Language	69
3.2.4	Data Availability.....	70
3.2.5	Learners' Nativeness	76
3.2.6	Learners' Proficiency Level	76
3.2.7	Learners' First Language.....	77
3.2.8	Material Mode	77
3.2.9	Material Genre	78
3.2.10	Task Type	78
3.2.11	Data Annotation.....	78

3.2.12 Summary of the ALC Design	79
3.3 ALC Metadata: Design and Content.....	80
3.3.1 Age.....	82
3.3.2 Gender	83
3.3.3 Nationality	84
3.3.4 Mother Tongue	85
3.3.5 Nativeness.....	86
3.3.6 Number of Languages Spoken	86
3.3.7 Number of Years Learning Arabic	86
3.3.8 Number of Years Spent in Arabic Countries.....	87
3.3.9 General Level of Education.....	87
3.3.10 Level of Study	88
3.3.11 Year/Semester.....	89
3.3.12 Educational Institution.....	90
3.3.13 Text Genre	91
3.3.14 Where Produced	92
3.3.15 Year of Production.....	92
3.3.16 Country of Production	92
3.3.17 City of Production	93
3.3.18 Timing	94
3.3.19 Use of References.....	95
3.3.20 Grammar Book Use	95
3.3.21 Monolingual Dictionary Use	95
3.3.22 Bilingual Dictionary Use	95

3.3.23	Other References Use	96
3.3.24	Text Mode.....	96
3.3.25	Text Medium	96
3.3.26	Text Length.....	97
3.3.27	Summary of the ALC Metadata.....	98
3.4	Corpus Evaluation.....	100
3.4.1	Projects That Have Used the ALC	100
3.4.2	Specialists' Feedback	101
3.4.3	Downloads from the ALC Website	104
3.5	Conclusion	104
4	Collecting and Managing the ALC Data	106
4.1	Introduction.....	107
4.2	Collecting the ALC Data	107
4.3	Collecting the ALC Metadata	111
4.4	Computerising the ALC	111
4.4.1	Transcribing Hand-Written Data.....	111
4.4.2	Consistency of Hand-Written Data.....	115
4.4.3	Transcribing Spoken Data	117
4.4.4	Consistency of Spoken Data.....	118
4.5	ALC Database	119
4.5.1	Data Storing	122
4.5.2	File Generation Function	123
4.6	File Naming	126
4.7	Conclusion	127

Part III ALC Tools.....	129
5 Computer-Aided Error Annotation Tool for Arabic	130
5.1 Introduction.....	131
5.2 Background.....	132
5.2.1 Annotation Tools	132
5.2.2 Error Annotation Tagsets and Manuals	134
5.3 The Computer-Aided Error Annotation Tool for Arabic (CETAr)	136
5.3.1 Annotation Standards	136
5.3.2 Design.....	142
5.3.3 Tokenisation	143
5.3.4 Manual Error Tagging	146
5.3.5 Smart Selection.....	146
5.3.6 Auto Tagging.....	148
5.3.7 Further Features	152
5.3.8 Evaluation.....	153
5.4 Error Tagset of Arabic (ETAr)	156
5.4.1 Error Categories and Types	158
5.5 First Evaluation: Comparison of Two Tagsets	160
5.5.1 Sample and Annotators.....	160
5.5.2 Task and Training.....	161
5.5.3 Results	162
5.5.4 Limitations and Suggestions.....	162
5.6 Second Evaluation: Inter-Annotator Agreement Measurement.....	163
5.6.1 Sample	163

Contents

5.6.2	Evaluators	163
5.6.3	Annotators	166
5.6.4	Results	167
5.6.5	Limitations and Suggestions.....	171
5.7	Third Evaluation: ETAr Distribution and Inter-Annotator Agreement	171
5.7.1	Refining the Tagset.....	172
5.7.2	Sample and Annotators.....	173
5.7.3	Task and Training.....	174
5.7.4	Distribution of the ETAr.....	175
5.7.5	Inter-Annotator Agreement	178
5.8	Error Tagging Manual for Arabic (ETMAr).....	180
5.8.1	Purpose	180
5.8.2	Evaluation.....	181
5.9	Conclusion	181
6	Web-Based Tool to Search and Download the ALC.....	184
6.1	Introduction.....	185
6.2	Review of Tools for Searching and Analysing Arabic Corpora	185
6.2.1	Method of Review	186
6.2.2	Tools Investigated.....	187
6.2.3	Evaluation Criteria.....	187
6.2.4	Evaluation Sample	190
6.2.5	Khawas	190
6.2.6	aConCorde	192
6.2.7	AntConc.....	193

6.2.8	WordSmith Tools	194
6.2.9	Sketch Engine	195
6.2.10	IntelliText Corpus Queries	197
6.2.11	Comparing the Results.....	198
6.3	Using the ALC Metadata to Restrict the Search.....	200
6.4	Purpose.....	202
6.5	Design	202
6.6	Determinant Types.....	204
6.7	Functions.....	206
6.7.1	Searching the Corpus.....	206
6.7.2	Downloading the Corpus Files	211
6.8	Evaluation	213
6.8.1	Evaluating the Output of the ALC Search Tool	213
6.8.2	Specialists' Views.....	222
6.8.3	Website Visits.....	227
6.9	Features and Limitations.....	228
6.10	Conclusion	229
	Part IV ALC Uses and Future Work	230
7	Uses of the Arabic Learner Corpus.....	231
7.1	Introduction.....	232
7.2	Projects That Have Used the ALC.....	232
7.2.1	Error Detection and Correction	233
7.2.2	Error Annotation Guidelines	234
7.2.3	Native Language Identification	234

7.2.4	Development of Robust Arabic Morphological Analyser and PoS-Tagger	235
7.2.5	Applied Linguistics.....	235
7.2.6	Workshop on Teaching Arabic.....	236
7.2.7	Data-Driven Arabic Learning.....	237
7.3	Further Uses of the ALC.....	238
7.3.1	Automatic Arabic Readability Research	238
7.3.2	Optical Character Recognition Systems	239
7.3.3	Teaching Materials Development.....	239
7.3.4	Arabic Learner Dictionaries	240
7.4	Conclusion	242
8	Future Work and Conclusion	244
8.1	Introduction.....	245
8.2	Thesis Achievements	245
8.3	Evaluation	247
8.4	Future Work.....	249
8.4.1	Guide on Design Criteria for Learner Corpus	249
8.4.2	Arabic Learner Corpus	249
8.4.3	Computer-Aided Error Annotation Tool for Arabic (CETAr) ...	251
8.4.4	Error Tagset of Arabic (ETAr) and Its Manual (ETMAr).....	252
8.4.5	ALC Search Tool.....	252
8.4.6	Further Applications of the ALC.....	253
8.4.7	Dissemination	253
8.5	Challenges.....	253

8.6 Conclusion	254
Appendix A Examples of ALC File Formats	255
A.1 Plain text files.....	255
A.2 XML files	257
A.3 PDF files	258
Appendix B The Guide for Data Collection	259
Appendix C The Paper Copy of ALC Questionnaire	261
Appendix D The Questionnaires That Used to Evaluate the ETAr	267
D.1 First evaluation questionnaire	267
D.2 Second Evaluation Questionnaire	269
Appendix E The Error Tagging Manual for Arabic (ETMAR)	283
Appendix F The DIN 31635 Standard for the Transliteration of the Arabic Alphabet	309
Appendix G Extended Code of the ALC Search Function	310
References	330