



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

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Abdullah Yahya G. Alfaifi to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

© 2015 The University of Leeds and Abdullah Yahya G. Alfaifi

Publications

Chapters 1 to 7 of this thesis are based on jointly-authored publications. The candidate is the principal author of all original contributions presented in these papers, the co-authors acted in an advisory capacity, providing feedback, general guidance and comments.

Chapter 1

The work in Chapter 1 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah and Eric Atwell (2013). Potential Uses of the Arabic Learner Corpus. In proceedings of *the Leeds Language, Linguistics and Translation PGR Conference 2013*. Leeds, UK.

Chapter 2

The work in Chapter 2 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah, Eric Atwell and Claire Brierley (under review). Learner Corpora: Present and Future, design criteria for creating a new learner corpus. *Applied Linguistics*.

Chapter 3

The work in Chapter 3 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah, Eric Atwell and Hidayah Ibraheem (2014). Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In Ishikawa, Shin'ichiro (Ed.), *Learner Corpus Studies in Asia and the World, Papers from LCSAW2014* (Vol. 2, pp. 77–89), School of Language & Communication. Kobe University, Japan.

Alfaifi, Abdullah and Eric Atwell (2013). Arabic Learner Corpus v1: A New Resource for Arabic Language Research. In proceedings of *the Second Workshop on Arabic Corpus Linguistics (WACL 2)*, Lancaster University, UK.

Alfaifi, Abdullah, and Eric Atwell (2014). Arabic Learner Corpus: A New Resource for Arabic Language Research. In the proceedings of *the 7th Saudi Students Conference*, 1-2 February 2014, Edinburgh, UK.

Chapter 4

The work in Chapter 4 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah and Eric Atwell (2012). المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء اللغوية "Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors" (in Arabic). In proceedings of *the 8th International Computing Conference in Arabic (ICCA 2012)*, 26 - 28 January 2012, Cairo, Egypt.

Alfaifi, Abdullah and Eric Atwell (2013). Arabic Learner Corpus: Texts Transcription and Files Format. In proceedings of *the International Conference on Corpus Linguistics (CORPORA 2013)*, St. Petersburg, Russia.

Chapter 5

The work in Chapter 5 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah, Eric Atwell and Ghazi Abuhakema (2013) Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In: *Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference (GSCL 2013)*, 25-27 September 2013, Darmstadt, Germany. Springer, (9) 14 - 22.

Alfaifi, Abdullah and Eric Atwell (2014). An Evaluation of the Arabic Error Tagset v2. *The American Association for Corpus Linguistics conference (AACL 2014)*. 26-28 September 2014, Flagstaff, USA.

Alfaifi, Abdullah and Eric Atwell (2015). Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. *The 8th Saudi Students Conference*, 31 January – 1 February 2015, London, UK.

Chapter 6

The work in Chapter 6 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah and Eric Atwell (2014). Tools for Searching and Analysing Arabic Corpora: an Evaluation Study. *BAAL / Cambridge University Press Applied Linguistics*, 14 Jun 2014. Leeds Metropolitan University, UK.

Alfaifi, Abdullah and Eric Atwell (accepted). Comparative Evaluation of Tools for Arabic Corpora Search and Analysis. *International Journal of Speech Technology (IJST)*.

Chapter 7

The work in Chapter 7 of the thesis has appeared in publication as follows:

Alfaifi, Abdullah and Eric Atwell (2014). Arabic Learner Corpus and Its Potential Role in Teaching Arabic to Non-Native Speakers. *The 7th Biennial IVACS conference*, 19 - 21 Jun 2014. Newcastle, UK.

Alfaifi, Abdullah (2015). Learner Corpora. In: Alosaimi, S, (Ed.) المدونات اللغوية: "Arabic Corpus: Development and Analysis Approaches" (in Arabic). King Abdullah bin Abdulaziz International Center for Arabic Language Service, KSA.

Acknowledgements

First and foremost, I praise *Allāh* (GOD) for His bounty and blessings and for providing me with health, patience, strength, wellbeing and skills to write this thesis and to finish one of the most important studies in my life.

I would like to thank my supervisor Dr Eric Atwell for supervising me during the last three years and for his advice, comments, guidance and encouragement on almost every aspect of the study. Thank you very much for the continuous encouragement that allowed me to publish most of my original contributions.

My deepest gratitude goes to my parents, for everything they have provided, and to my family (my wife Mohsinah and my children Fatin and Fanan), to whom this thesis is dedicated, for their love, patience and unfailing support. I would also like to express my joy at the birth of my daughter Tasneem who came into the world a few months before the viva.

My sincere gratitude and thanks are also directed to colleagues, friends, Arabic NLP group members and members of the research community who gave me invaluable advice, encouragement and support, and for the great seminars we enjoyed. These include the following people: Nizar Habash, Ghazi Abuhakema, Hidayah Ibraheem, Majdi Sawalha, Claire Brierley, Wajdi Zaghouni, Amal Alsaif, Eshrag Refae, Hussain Aljahdali, Sammer Alrehaili, Ahmad Alzahrani, Abdulaziz Albatli, Ibrahim Alzamil, Mohammad Alqahtani, Abdulrahman Alosaimy, Ayman Alghamdi, Saleh Alosaimy, Jaber Asiri, Ali Alhakam, Ahmad Alqarni, Ahmad Alotaibi, Ahmad Alshaiban, Fahad Alkhalaf and Adel Alfaifi.

I would like to acknowledge the hard work of the numerous volunteers (language learners, data collectors, evaluators, annotators and collaborators) who contributed their time and effort to this project. I also owe special thanks to the colleagues who worked very hard on data transcription of the Arabic Learner Corpus.

I would also like to acknowledge Al Imam Mohammad Ibn Saud Islamic University for providing me with the scholarship to carry out my studies in the UK.

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.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

﴿وَقُلْ رَبِّ زِدْنِي عِلْمًا﴾

“...and say, My Lord, increase me in knowledge”

The Quran, *Surat Ṭāhā*, verse (20:114)

Contents

Publications.....	iii
Acknowledgements.....	vi
Abstract.....	vii
Contents	ix
List of Tables	xix
List of Figures.....	xxiii
List of Abbreviations	xxix
Part I Introduction and Literature Review	1
1 Introduction.....	2
1.1 <i>Corpus and 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

Contents

2.2.1	Purpose	27
2.2.2	Sizes	30
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

Contents

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

Contents

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

Contents

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
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

Contents

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

Contents

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

Contents

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

List of Tables

Table 1.1: Phases of Developing the ALC with links to the thesis chapters	11
Table 1.2: The ALC participants.....	12
Table 2.1: Aspects covered in the review with percentage of the data not available	19
Table 2.2: Learner corpora reviewed with their references	20
Table 2.3: Calculations of corpora sizes	32
Table 2.4: Genres used in learner corpora	46
Table 2.5: Task types used in learner corpora.....	47
Table 2.6: Examples of learner corpora annotation	49
Table 2.7: Summary of the existing Arabic learner corpora.....	61
Table 3.1: Summary of ALC files available for download.....	74
Table 3.2: Summary of the design criteria used in the ALC.....	79
Table 3.3: Metadata elements used in the ALC	82
Table 3.4: Distribution of nationalities in the ALC	84
Table 3.5: Distribution of mother tongues in the NNS part of the ALC.....	85
Table 3.6: Levels of the learners who contributed to the ALC.....	87
Table 3.7: Word distribution based on general level, level of study, and year/semester.....	90
Table 3.8: Word distribution based on institutions from where the ALC data was collected.....	91
Table 3.9: Average length of the ALC texts based on some key factors	98
Table 3.10: Summary of the variables used in the ALC metadata.....	98
Table 4.1: Summary of the data collection procedures.....	109

List of Tables

Table 4.2: Standards followed in transcription with authentic examples from the corpus texts	112
Table 4.3: Consistency between transcribers of ALC v1	115
Table 4.4: Second test of consistency between transcribers of ALC v1	115
Table 4.5: Final test of consistency in ALC v1	116
Table 4.6: Consistency between transcribers of ALC v2.....	116
Table 4.7: Second test of consistency between transcribers of ALC v2.....	116
Table 4.8: Final test of consistency in ALC v2.....	117
Table 4.9: Aspects that are marked up in audio recording transcriptions.....	117
Table 4.10: Consistency between transcribers of spoken materials in ALC v2.....	119
Table 4.11: Classification features of the corpus files	124
Table 4.12: Example of corpus files naming method	127
Table 5.1: Results of task 1 of annotation speed by hand and using CETAr.....	154
Table 5.2: Samples used to test the auto-tagging feature.....	155
Table 5.3: Results of testing the auto-tagging feature.....	156
Table 5.4: Error taxonomies in some Arabic studies	158
Table 5.5: Error Tagset of Arabic (ETAr)	159
Table 5.6: Annotators who participated in the first evaluation of the ETAr.....	161
Table 5.7: Annotating comparison between ARIDA and ETAr.....	162
Table 5.8: Evaluators who participated in the first refinement of the ETAr.....	164
Table 5.9: Second version of the ETAr.....	165
Table 5.10: Annotators who participated in the second evaluation of the ETAr....	166
Table 5.11: Examples from the first list with its questionnaire	167
Table 5.12: Inter-annotator agreement in both lists of the second evaluation	168

List of Tables

Table 5.13: Inter-annotator agreement in both lists of the second evaluation	168
Table 5.14: The potential impact of training on the ease of finding the tags.....	169
Table 5.15: Responses to the final questionnaire.....	170
Table 5.16: Third version of the ETAr.....	172
Table 5.17: Annotators who participated in the third evaluation of the ETAr	174
Table 5.18: Distribution of the tags' use and agreement by the annotators.....	177
Table 5.19: Inter-annotator agreement in the third evaluation.....	179
Table 6.1: Benchmark score of the Khawas tool	192
Table 6.2: Benchmark score of the aConCorde tool.....	193
Table 6.3: Benchmark score of the AntConc tool.....	194
Table 6.4: Benchmark score of the WordSmith Tools.....	195
Table 6.5: Benchmark score of the Sketch Engine web tool	197
Table 6.6: Benchmark score for IntelliText Corpus Queries	198
Table 6.7: Comparison of the tools included in this evaluation.....	199
Table 6.8: Number of files available for each format in the ALC	212
Table 6.9: Number of results returned for each query on the reference tools compared to the ALC Search Tool	217
Table 6.10: Formulas used to compute precision, recall, and F-measure.....	217
Table 6.11: Evaluation of the normal search on the ALC Search Tool	218
Table 6.12: Evaluation of the <i>Separate Words</i> option on the ALC Search Tool....	219
Table 6.13: Evaluators of the ALC Search Tool.....	223
Table 6.14: Summary of the evaluators' responses to the questionnaire about the ALC Search Tool	223
Table 7.1: Three hierarchical degrees of level indicators in the ALC	239

List of Tables

Table 7.2: Concordances of the word “بالنسبة” <i>binnisba</i> ‘regarding’	240
Table 7.3: The 10 most common errors in a 10,000-word sample of the ALC	241
Table 7.4: The 10 most common errors based on the nativeness factor	242
Table 8.1: Example of the suggested annotation for the ALC	250

List of Figures

Figure 1.1: Structure of the thesis	14
Figure 2.1: Purposes of compiling the learner corpora	28
Figure 2.2: Percentages of corpora created for public purposes	29
Figure 2.3: Sizes of all textual corpora based on w/t sizes	33
Figure 2.4: Sizes of textual corpora with 4 million w/t or less	33
Figure 2.5: Number of textual corpora with 1 million w/t or less.....	34
Figure 2.6: Number of spoken corpora based on length (hours)	35
Figure 2.7: Learner corpora distribution based on target languages included	36
Figure 2.8: Target languages in learner corpora	37
Figure 2.9: Availability of learner corpora	38
Figure 2.10: Data of native and non-native speakers.....	40
Figure 2.11: Number of corpora based on proficiency levels included	41
Figure 2.12: Proficiency levels distribution	41
Figure 2.13: Corpora with various L1s vs. sole L1	42
Figure 2.14: First languages in learner corpora	43
Figure 2.15: Materials modes in learner corpora	45
Figure 2.16: Number of genres included in learner corpora	46
Figure 2.17: Number of task types included in learner corpora.....	48
Figure 2.18: Learner corpora tagging.....	49
Figure 2.19: Example of annotation from the Japanese Learner of English Corpus	50

List of Figures

Figure 2.20: Example of annotation from the Czech as a Second/Foreign Language Corpus	51
Figure 2.21: Example of annotation from the Foreign Language Examination Corpus	52
Figure 2.22: Types of annotation used in learner corpora	53
Figure 3.1: Illustration of the XML structure.....	73
Figure 3.2: Word distribution based on nativeness of the learners	76
Figure 3.3: Word distribution based on age ranges of the learners.....	83
Figure 3.4: Word distribution based on gender of the learners.....	84
Figure 3.5: Word distribution based on nativeness of the learners	86
Figure 3.6: Word distribution based on general level of the learners	88
Figure 3.7: Word distribution based on level of study of the learners.....	89
Figure 3.8: Word distribution based on year/semester of the learners.....	89
Figure 3.9: Locations of the Saudi cities from which the ALC data was collected	94
Figure 3.10: Word distribution of the ALC based on the text mode.....	96
Figure 3.11: Word distribution of the ALC based on the text medium	97
Figure 3.12: Lengths of the ALC texts.....	97
Figure 3.13: Projects that have used the ALC.....	101
Figure 3.14: Google Analytics map showing locations of ALC visitors	104
Figure 4.1: Instructions for Tasks 1 and 2 of the hand-written materials	109
Figure 4.2: Online form for data collection	110
Figure 4.3: Example of a text with its transcription.....	114
Figure 4.4: The ALC database with the entity-relationship diagram.....	121

List of Figures

Figure 4.5: Example of a text stored in the ALC database	122
Figure 4.6: Example of metadata stored in the ALC database.....	123
Figure 4.7: Three methods for generating files for the entire ALC	124
Figure 4.8: Custom file generation in the ALC database.....	125
Figure 4.9: Processes of the files generation function	126
Figure 5.1: Example of annotating Arabic text using the WebAnno2 tool.....	133
Figure 5.2: Example of annotating Arabic text using GATE	133
Figure 5.3: Example of annotating Arabic text using the Content Annotation Tool.....	134
Figure 5.4: Example of text tokenisation	137
Figure 5.5: Example of tokens separated from each other by a tab space	137
Figure 5.6: Example of error annotated with two error types	137
Figure 5.7: Plain text with inline annotation.....	138
Figure 5.8: Plain text with stand-off annotation by tokens	139
Figure 5.9: XML with inline annotation	139
Figure 5.10: XML with stand-off annotation by tokens	139
Figure 5.11: DTD model for XML files containing metadata and inline annotation.....	140
Figure 5.12: DTD model for XML files containing metadata and stand-off annotation by tokens	142
Figure 5.13: The main interface of the CETAr.....	143
Figure 5.14: Sample code of the tokenisation process.....	144
Figure 5.15: Example of a text tokenised by CETAr.....	146

List of Figures

Figure 5.16: Tagging multiple errors using the smart-selection feature in the CETAr.....	147
Figure 5.17: Sample code of the smart-selection feature	148
Figure 5.18: Steps of using the auto-tagging function	150
Figure 5.19: Sample code of the Auto-tagging function.....	152
Figure 5.20: Editing the list of tagged tokens	152
Figure 5.21: Example of a final output of the annotation in CETAr	153
Figure 5.22: Annotators' responses to the question about easiness of finding the tags	169
Figure 5.23: Example of the error annotation method in the third evaluation	175
Figure 5.24: Extracting the tags used by each annotator in the third evaluation	176
Figure 5.25: Differences in the distribution of tags use and agreement.....	178
Figure 6.1: A message from Notepad about the file encoding.....	188
Figure 6.2: Khawas Shows Arabic words in a right-to-left order	191
Figure 6.3: Some Arabic words were missed from concordances when Khawas was run on Windows.....	192
Figure 6.4: Frequency and concordances in aConCorde	193
Figure 6.5: Columns of Arabic concordances in AntConc were shown in the opposite direction.....	194
Figure 6.6: Diacritics do not appear in their correct positions in WordSmith Tools	195
Figure 6.7: Sketch Engine removed the diacritics when normalising the texts	196
Figure 6.8: Diacritics displayed correctly in IntelliText Corpus Queries	198
Figure 6.9: Example of determinants of the ALC in Sketch Engine	201

List of Figures

Figure 6.10: Search determinants on the website of the Michigan Corpus of Upper-level Student Papers.....	201
Figure 6.11: English interface of the main page of the ALC Search Tool.....	203
Figure 6.12: Updating the number of texts available based on the determinants selected.....	204
Figure 6.13: Example of a determinant with a numerical range value	205
Figure 6.14: Example of a determinant with a multi-selection list.....	205
Figure 6.15: Example of a determinant with only two options.....	206
Figure 6.16: Results section on the ALC Search Tool.....	207
Figure 6.17: Results with and without using the <i>Separate Words</i> choice.....	208
Figure 6.18: Sending a query to the ALC database	208
Figure 6.19: Showing or hiding the results based on the query response.....	210
Figure 6.20: Architecture of the search function in the ALC Search Tool.....	211
Figure 6.21: Architecture of the download function on the ALC Search Tool.....	213
Figure 6.22: The confusion matrix aspects and elements	215
Figure 6.23: Precision, recall, and F-measure of the normal search on the ALC Search Tool	219
Figure 6.24: Precision, recall, and F-measure of the <i>Separate Words</i> option on the ALC Search Tool	220
Figure 6.25: Map showing locations of the ALC Search Tool visitors	228
Figure A.1: Example of plain text file with no metadata.....	255
Figure A.2: Example of plain text file with Arabic metadata.....	255
Figure A.3: Example of plain text file with English metadata.....	256
Figure A.4: Example of XML file with Arabic metadata.....	257

List of Figures

Figure A.5: Example of XML file with English metadata.....	257
Figure A.6: Example of handwritten text in PDF file format	258
Figure C.1: An overview about the ALC project in the data collection questionnaire	261
Figure C.2: The consent form to take part in the ALC project	262
Figure C.3: The learner's profile questionnaire used in ALC.....	263
Figure C.4: The text's data questionnaire used in ALC.....	264
Figure C.5: Task 1 in the ALC questionnaire	265
Figure C.6: Task 2 in the ALC questionnaire	266
Figure G.1: Extended Code of the Search Function of the ALC Search Tool.....	329

List of Abbreviations

The abbreviations used in this thesis are listed in the following table with their meanings. The table also shows the page on which each abbreviation is defined.

Abbreviation	Meaning	Page
ALC	Arabic Learner Corpus	2
ARIDA	Arabic Interlanguage Database	135
CETAr	Computer-aided Error Annotation Tool for Arabic	9
DTD	Document Type Definition	71
ETAr	Error Tagset of Arabic	7
ETMAr	Error Tagging Manual for Arabic	9
FRIDA	French Interlanguage Database	134
GF	Grammatical Function	250
L1	First Language	4
L2	Second Language	3
MP3	MPEG-2 Audio Layer III	71
MSA	Modern Standard Arabic	70
NLP	Natural Language Processing	5
NS/NNS	Native Speaker/Non-Native Speakers	4
OCR	Optical Character Recognition	5
PDF	Portable Document Format	71
PoS	Part-of-Speech	49
SLA/FLT	Second Language Acquisition/Foreign Language Teaching	3
TXT	Plain text format	60
W/T	Words/Tokens	31
XML	Extensible Markup Language	71

Part I

Introduction and Literature Review

Summary of Part I

This part presents in Chapter 1 the theoretical framework of the research. It begins with definition of the terms corpus and learner corpora, an introduction to the importance of learner corpora, their uses in some relevant linguistic domains and computational applications, the motivation behind this thesis and its objective toward the development of the Arabic Learner Corpus. Chapter 1 concludes with the presentation of the study's novel contributions and description of the project participants and the thesis outline. Chapter 2 consists of a comprehensive review of the learner corpora domain and recommended guidelines for creating a new learner corpus on which the Arabic Learner Corpus was developed. The chapter also reviews related works, Arabic learner corpora, to justify the need for creating a new Arabic learner corpus.

1 Introduction

Chapter Summary

This chapter starts with defining the terms corpus and learner corpora. The chapter proceeds by highlighting the importance of learner corpora and summarising their uses in some relevant linguistic domains such as contrastive interlanguage analysis, error analysis, and teaching materials development, as well as in computational applications such as error correction systems, native language identification models, and optical character recognition applications. The chapter describes the motivation behind this thesis and its objective toward the development of the Arabic learner corpus. The chapter then provides details about the study's novel contributions including resources, proposed standards, and tools. The concluding sections present an overview of the structure and scope of the Arabic Learner Corpus (ALC) project, which is distributed in three main phases, before providing a description of the project participants and the thesis outline.

1.1 *Corpus and Learner Corpora*

This section presents the definition of the term *corpus* in general and some further definitions that focus on particular aspects. Then it defines *learner corpora* as a specialised type.

1.1.1 The Term *Corpus*

The term *corpus* (singular form of *corpora*¹) refers to an electronic collection of authentic texts or speeches produced by language speakers and stored in a machine-readable format (Jurafsky and Martin, 2009; Kennedy, 1998; McEnery, 2003; Nesselhauf, 2004; Nugues, 2006; Sinclair, 1996; Wynne, 2005). Researchers have made attempts to provide more specific definitions of *corpus*. Nesselhauf (2004), for example, argues that the corpus should be intended for general use, not merely for one specific study or even a limited number of studies. Sinclair (2005) demonstrates more concern for the design criteria, issues of representativeness, and the main role that a corpus plays. He defines a *corpus* as “a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research” (p 16). McEnery *et al.* (2006) point out that a corpus should be a principled collection of texts, which differs from a random collection of texts. Thus, a principled corpus can be defined as “a collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety” (p 5).

1.1.2 *Learner Corpora*

“Granger (2008) explains that “[l]earner corpus research is a fairly young but highly dynamic branch of corpus linguistics, which began to emerge as a discipline in its own right in the late 1980’s/early 1990’s” (p 259). *Learner corpus* is a specialised type of corpora, and Granger (2002) defines *learner corpora* as “electronic collections of authentic FL/SL [Foreign Language/Second Language (L2)] textual data assembled according to explicit design criteria for a particular SLA/FLT

¹ McEnery (2003) pointed out that *corpuses* is perfectly acceptable as a plural form of *corpus*.

[Second Language Acquisition/Foreign Language Teaching] purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance” (p 7).

Given the fact that 20% of learner corpora reviewed in this study includes data from both native speakers (NS) and non-native speakers (NNS), it can be noticed that Granger's definition emphasises the importance of data collected from FL/SL learners, and ignores data produced by native speakers in language learning contexts. Therefore, we can define learner corpora as electronic collections of authentic data (e.g. texts, speeches or videos) produced in a language learning context by NS and/or NNS according to explicit design criteria and stored in a machine-readable format.

The contribution of learner corpora – since their appearance a few decades ago – has focussed on second language acquisition in particular. However, researchers in other domains have started exploiting this valuable resource due to its potential uses. The next section highlights the importance of learner corpora by presenting an overview of their uses.

1.2 Importance of Learner Corpora

The number of learner corpora has noticeably grown in the last decade, which highlights the role they play in linguistic and computational research and the valuable data resource they can provide.

Researchers in the field of linguistic research frequently use learner corpora for Contrastive Interlanguage Analysis, which enables researchers to observe a wide range of instances of underuse, overuse, and misuse of various aspects of the learner language at different levels: lexis, discourse, and syntax (Granger, 2003b). Analysing errors also enables researchers and educators to understand the interlanguage errors caused by First Language (L1) transfer, learning strategies, and overgeneralization of L1 rules. Learner corpora were – and still are – used to compile or improve learner dictionary contents, particularly by identifying the most common errors learners make, and then providing dictionary users with more details at the end of relevant entries. These errors may take place in words, phrases, or language structures, along with the ways in which a word or an expression can be used correctly and incorrectly (Granger, 2003b; Nesselhauf, 2004). Also, error-

tagged learner corpora are useful resources to measure the extent to which learners can improve their performance in various aspects of the target language (Buttery and Caines, 2012; Nesselhauf, 2004). Analysing learners' errors may function as a beneficial basis for pedagogical purposes such as creating instructional teaching materials. It can, for instance, help in developing materials that are more appropriate to learners' proficiency levels and in line with their linguistic strengths and weaknesses.

With respect to computational applications, learner corpora can be utilised for different purposes. Developers of error correction systems, for example, use learner corpora, which include error annotation, to train their systems to detect and correct errors. They also perform experiments to test their models on raw data from learner corpora, as this approach gives authentic evaluation of such applications. Language identification systems are another example of applications that benefit from learner corpora. The aim of such applications is to infer the native language of an author based on texts written in a second language (Malmasi and Dras, 2014). Finally, learner corpora that contain original hand-written texts with their transcription in a computerised format can be used as a training dataset in the research and development of Optical Character Recognition (OCR) systems.

1.3 Motivation and Aim

Recent research developments in, and uses of, learner corpora were the main inspiration behind this research. These uses have allowed this type of corpora to play a growing role in some linguistic and computational research areas such as language teaching and learning and Natural Language Processing (NLP). Additionally, the lack of a well-designed Arabic learner corpus increases the importance of creating such a resource, which may encourage researchers to conduct more studies in the aforementioned research areas.

The aim of the project is to develop an open-source Arabic learner corpus and a system for Arabic error annotation to be used as a valuable resource for research on language teaching and learning as well as NLP. Using original scientific research, we focus on the question of how to create a methodology for developing a learner corpus based on the best practice in the field.

1.4 Objectives

In order to achieve the study aim, the researcher defined a number of objectives as following:

1. To review the learner corpora existing under specific criteria

This comprehensive review under 11 categories (corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation) will allow us to have an idea about the best practice in this field and to shape our design criteria of the ALC project.

2. To create a guide for developing a new learner corpus

This guidance is based on the review of previous work. It focuses on the eleven aspects of corpus design criteria in order to serve as open-source standards for developing new learner corpora and also to improve and/or expand the current corpora.

3. To collect data for the Arabic Learner Corpus based on its design criteria

The ALC is developed to be a resource for research on Arabic teaching and learning as well as Arabic NLP. Based on the guidance for developing a new learner corpus, the target size is 200,000 words (written and spoken), to be produced by learners of Arabic (native and non-native speakers) from various first language backgrounds and nationalities.

4. To develop an error tagset for Arabic

This includes developing error taxonomy for the most frequent errors in Arabic learners' production. It also includes a tagset designed for annotating those errors. Iterated evaluations will be performed on this tagset by a number of Arabic experts and annotators in order to provide the target users with easy-to-understand categories and types of errors. Additionally, a manual will be developed describing how to annotate Arabic texts for errors using the error tagset.

5. To develop a computer-aided error annotation tool for Arabic

This computer-aided error annotation tool is intended to be developed based on the error tagset of Arabic as a part of the ALC project. It will include some

automated features that can facilitate the annotation process and increase the consistency of error annotation more than purely manual annotation.

6. To develop a search tool based on the ALC metadata

This tool will be developed to enable users to search the ALC data based on a number of determinants (the ALC metadata). The corpus design criteria include metadata elements such as “Age”, “Gender”, “Mother tongue”, “Text mode”, “Place of writing”, etc. Those metadata elements will be utilised as determinants to search any sub-corpus of the ALC, or download the source files in different formats (TXT, XML, PDF, and MP3).

1.5 Thesis Contributions

The study presents a novel set of resources, proposed standards, and tools that contribute to Arabic NLP as well as Arabic linguistics. The following list classifies the contributions into the three dimensions.

A. Resources

1. Arabic Learner Corpus (ALC)

The ALC is a standard resource for research on Arabic teaching and learning as well as Arabic NLP. It includes 282,732 words and 1585 materials (written and spoken) produced by 942 students from 67 nationalities and 66 different L1 backgrounds. Based on our examination of the literature, we are confident that the ALC is the largest learner corpus for Arabic, the first Arabic learner corpus that comprises data from both native Arabic speakers and non-native Arabic speakers, and the first Arabic learner corpus for Arabic as a Second Language (ASL¹) collected from the Arab world.

2. Error Tagset of Arabic (ETAr)

The Error Tagset of Arabic is a part of the ALC project. It includes an error taxonomy which is designed based on a number of studies that investigated the most frequent errors in Arabic learners’ production. Additionally, it includes a

¹ The term *Second Language (SL)* usually refers in Applied Linguistics to the situation where learners can be exposed to the target language outside of the classroom, learning English in the UK for instance, while *Foreign Language (FL)* means that learners have less chance to be exposed to the target language (e.g. learning French in Saudi Arabia) (see for example Littlewood, 1984).

tagset designed for annotating errors in Arabic covering 29 types of errors under five broad categories. Seven annotators and two evaluators performed – in groups – iterated evaluations on this tagset, the ETAr was improved after each evaluation. The aim of the ETAr is to annotate errors in the ALC as well as for further Arabic learner corpora, particularly those for Arabic language teaching and learning purposes. It is available to researchers as an open source¹. It provides the target users with easy-to-understand categories and types of errors.

3. Review of the learner corpora domain

We published online² a summary review of 159 previous works (learner corpora) in order to create an easy-access and open source for the best practice in this field. Developers of new similar projects and learner corpora users can benefit from this source in their research.

B. Proposed standards

4. Guidance on design criteria for learner corpus

We created a guide for developing a new learner corpus based on a review of previous work. It focuses on 11 aspects of corpus design criteria, such as purpose, size, target language, availability, learners' nativeness, materials mode, data annotation, etc. Our aim is that these criteria will serve as open-source standards for developing new learner corpora. The guide can also be utilised to improve and/or expand the current corpora.

5. Proposed standards for transcribing Arabic hand-written texts

Given that the Arabic language has its own writing system, which includes for example different types of *Hamza* (ء)³, diacritics (short vowels), and characters with dots above or below, and that most of the ALC data are hand-written texts, we created specific standards for converting those texts into a computerised format in order to achieve the highest possible level of consistency in the transcription process. These standards cover cases such as when there is an

¹ This source can be accessed from:

http://www.comp.leeds.ac.uk/scayga/Error_Tagset_for_Arabic_Learner_Corpora.html

² This source can be accessed from:

http://www.comp.leeds.ac.uk/scayga/learner_corpora_summary.html

³ *Hamza* is consonant, glottal stop, it has specific rules for spelling that depend on its vocalic context. *Hamza* is written above or below specific letter forms (أ، إ، ؤ، ئ، ة)، and it has a stand-alone form as well (ء), see Habash, 2010; Samy and Samy, 2014.

overlap between two hand-written characters that cannot be transcribed together, when the writer used an unclear form of a character, or when a writer forgot a character's dots.

6. Error Tagging Manual for Arabic (ETMAr)

We developed this manual to describe how to annotate Arabic texts for errors. It is based on the final revised version of the ETAr. The ETMAr contains two main parts: The first defines each error type in the Arabic Error Tagset with examples of those errors and how they can be corrected. The second shows how annotators can deal with ambiguous instances and select the most appropriate tags.

C. Tools

7. Computer-aided Error Annotation Tool for Arabic (CETAr)

We developed a new tool for computer-aided error annotation in the ALC. It is based on the ETAr and includes some automated features such as the Smart-Selection function and the Auto-Tagging function. The Smart-Selection function finds similar errors and annotates them in a single step with no need to repeat the annotation process with each error. The Auto-Tagging function is similar to translation memories as it recognises the tokens that have been manually annotated and stores them into a database; subsequently, similar errors in other texts can be detected and annotated automatically. Using this tool increases the consistency of error annotation more than purely manual annotation.

8. ALC Search Tool

We established the ALC Search Tool¹ to enable users to search the ALC based on a number of determinants. The corpus design criteria include 26 metadata elements such as “Age”, “Gender”, “Mother tongue”, “Text mode”, “Place of writing”, etc. We structured the tool so that users can utilise those metadata elements as determinants to search any sub-corpus or download the source files in different formats (TXT, XML, PDF, and MP3).

To sum up, the thesis presents a number of resources, proposed standards, and tools developed for the ALC project. However, the main contribution of the thesis is not only the description of these components but also the detailed and original methodology that this thesis presents for developing a new learner corpus. The

¹ This tool can be accessed from: <http://www.alcsearch.com>

combination of the aforementioned resources, standards, and tools represents this new methodology.

1.6 Structure and Scope of the ALC Project

As described in the project aim, it is to develop an open-source Arabic learner corpus and a system for Arabic error annotation as valuable resources for research on Arabic NLP and Arabic teaching. The project includes some fundamental components such as the corpus data, the guidance on criteria for designing a learner corpus, and the ALC Search Tool. The system of Arabic error annotation consists of an error taxonomy, error tagset, error tagging manual, and computer-aided error annotation tool. We developed these resources, standards, and tools through three main phases which will be described in this section.

Design criteria are important for building a corpus. In order to follow the best practices, the first phase of the ALC was to review the literature which includes 159 learner corpora around the world. The review covered 11 aspects: corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation. The review provided us with a comprehensive view of the domain and helped us to create a review-based guide on design criteria for a new learner corpus. The design criteria of the ALC corpus were selected based on this guide and the ALC objectives. At this stage, we formed the theoretical basis of the project, and then we began to work on the practical phases.

The second stage was devoted to building the corpus and developing the required tools and standards. This step included creating tools for data collection, standards for converting the data into a computerised format, and a database for managing the corpus data. In this phase, we used the tools and standards to build the corpus.

During the third phase, we developed subsequent tools. These tools include a function to generate the corpus files automatically from the database in different formats, an error annotation tool with a tagset and manual for tagging Arabic errors, and a website for searching the ALC using the corpus metadata as determinants. Table 1.1 summarises these three main phases of developing the ALC and links each phase to the thesis chapters.

Table 1.1: Phases of Developing the ALC with links to the thesis chapters

Phase	Thesis chapter
1 Forming the theoretical basis of the project	
• Reviewing the literature (159 previous learner corpora) and related work (Arabic learner corpora)	2
• Developing guidance on design criteria of new learner corpora	2
• Defining the design criteria of the ALC	3
2 Developing tools and standards for building the corpus	
• Tools for data collection	4
• Standards for converting the data into a computerised format	4
• Database for managing the corpus data	4
3 Developing the subsequent tools	
• Function for generating the corpus files from the database in different formats	4
• Error annotation tool with a tagset and manual for tagging Arabic errors	5
• Website for searching the ALC using the corpus metadata as determinants	6

As seen from Table 1.1, the scope of the ALC project covers three pre-determined phases, (i) designing the corpus based on standard criteria which were derived from reviewing a large number of previous works, (ii) collecting the corpus materials using well-designed tools and developing a suitable database to manage these materials after they had been converted into an electronic format, and (iii) enabling users to benefit from the corpus data by generating the corpus files in different final formats and allowing users to search the corpus online.

The project scope does not include conducting a corpus-based study to exemplify the ALC use for three reasons. First, the benefits and value of using learner corpora in research are already proved through the studies conducted in this field. (Katja Markert, personal communication, 15 May 2014). Second, we designed the corpus to be an open source for relevant research areas; however, providing an example of

corpus use may lead researchers to conclude that its use is restricted, or at least more suitable, to a single research area. Finally, focussing on the three phases aforementioned allowed us to work further on the ALC tools such as the error annotation tool and the ALC Search Tool.

During these three phases of developing the ALC, around 1000 people contributed to the project. The following section describes those participants.

1.7 ALC Participants

The project was able to recruit 998 participants including language learners, data collectors, evaluators, annotators, and collaborators from more than 30 educational institutions in Saudi Arabia and the UK. Apart from the language learners, the other participants (i.e. data collectors, evaluators, annotators, and collaborators) included teachers of Arabic as a second language, secondary school teachers, university faculty (e.g. deans, vice deans, departments heads, and academic staff) and others. Table 1.2 illustrates the number of people based on their contribution to the project¹.

Table 1.2: The ALC participants

Number	Participation type
942	Arabic language learners (699 males and 243 females)
19	Data collectors (11 males and 8 females)
12	Evaluators (12 males)
7	Annotators (7 males)
18	Collaborators who facilitate the data collection from the learners (16 males and 2 females)

Each of the language learners signed a consent form which stated that the data collected would be published and used in relevant future research. The education in Saudi Arabia is made to single gender classes; that is, males and females do not mix. Therefore, it would have been impossible for a male researcher to enter a female school or university during their operational hours, making it necessary to recruit a

¹ More details about the ALC participants are available on the project website:
<http://www.arabiclearnercorpus.com/#!/corpus-team-en/c13uv>

number of female representatives to collect the required data. All representatives, male ($N = 11$) and female ($N = 8$), signed consent forms to confirm that all materials they collected would be kept securely until they were submitted to the researcher after the collection process. The form specified that the representatives would not keep any part of the data in any medium, and would not share any information they might know about the learners or their materials with any third party. The researcher also obtained permission from the institution from which the corpus data was collected to meet students and collect the corpus materials. Regarding the evaluators and annotators, their work was done either on anonymous data or different parts of the project, such as the error tagset and its manual, that did not contain any private information; thus, no consent forms were needed for them.

Most of the participants were interested in the Arabic language (i.e. researcher, teachers or specialists in Arabic). This was a significantly helpful factor, as they were all motivated to contribute to this project due to its importance to the research on Arabic. As a result, they were not paid for their participation, with the exception of some gifts (usually books) that were given to those learners who participated in all written and spoken tasks required for the project.

1.8 Thesis Outline

This thesis is divided into eight chapters under four parts as shown in Figure 1.1.

Part I: Introduction and Literature Review

Chapter 1: Introduction

Chapter 2: Literature Review and Related Work

Part II: Arabic Learner Corpus

Chapter 3: Design and Content

Chapter 4: Collecting and Managing the ALC Data

Part III: ALC Tools

Chapter 5: Computer-Aided Error Annotation Tool for Arabic

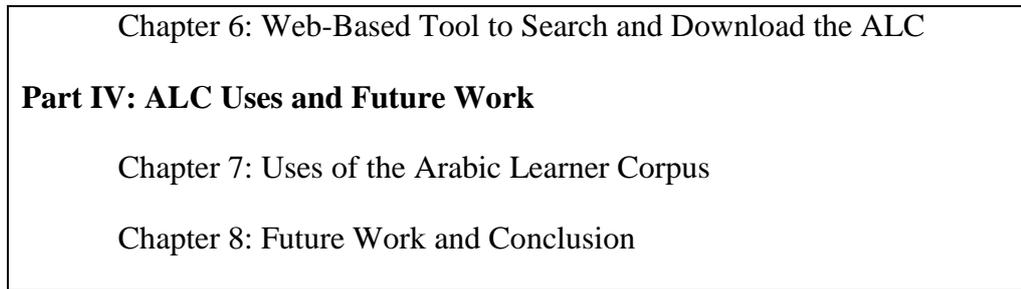


Figure 1.1: Structure of the thesis

- **Part I** includes the introductory information in Chapter 1 and the literature review with a focus on related work in Chapter 2.
 - **Chapter 1** provides an introduction and defines the terms *corpus* and *learner corpora*. It highlights the importance of learner corpora and summarises their uses in some relevant linguistic and computational domains. The chapter describes the motivation behind this thesis and its objective with details about the novel contributions including resources, proposed standards, and tools. It also gives an overview of the structure and scope of the ALC project and concludes by presenting the thesis outline.
 - **Chapter 2** provides a review of 159 learner corpora under 11 categories to derive design criteria for developing new learner corpora. It provides a quantitative view of the domain and concludes by recommending guidelines for creating a new learner corpus based on the analysis results. Additionally, this chapter reviews existing Arabic learner corpora and illustrates the contribution of the ALC project compared to those related corpora.
- **Part II** focuses on two aspects of the ALC: its design and content in Chapter 3 and how the corpus data was collected and managed using the ALC database in Chapter 4.
 - **Chapter 3** describes in detail the design and content of the ALC. It discusses the 11 design criteria on which the corpus development was based. The discussion of each criterion starts with an overview of relevant literature followed by the target design for the ALC and the final results achieved. The chapter also describes the corpus metadata, as the ALC has 26 elements of

- metadata related to learners and their texts. The ALC content is described regarding each of those elements.
- **Chapter 4** describes how the ALC data and metadata were collected using a questionnaire and guideline designed for this purpose. It describes also how the hand-written and spoken data was converted into an electronic form and how the consistency between transcribers was measured. The description in this chapter covers the design of a database to manage the ALC data and to automate generating the corpus files in different formats.
 - **Part III** describes two tools created as a part of the ALC project: the Computer-aided Error Annotation Tool for Arabic in Chapter 5 and the ALC Search Tool in Chapter 6.
 - **Chapter 5** describes the Computer-aided Error Annotation Tool for Arabic that we developed mainly to assist in annotating Arabic errors consistently in learner corpora. This chapter also describes the Error Tagset of Arabic with details on how it was evaluated by seven annotators and two evaluators to refine it from the first version to the third one. The chapter concludes with a discussion of the Error Tagging Manual for Arabic.
 - **Chapter 6** introduces the ALC Search Tool, a free-access, web-based concordancing tool. The chapter describes how the corpus metadata was used as determinants in order to enable users to search the ALC or a subset of its data or to download the source files of any sub-corpus based on those determinants. It also shows different types of evaluations for this tool.
 - **Part IV** highlights the uses of the ALC in various research areas in Chapter 7. It describes some plans that have been made for future work and discusses the conclusions drawn from this experimental work in Chapter 8.
 - **Chapter 7** describes examples of those projects that have used the ALC for different purposes, such as error detection and correction, error annotation guidelines, native language identification, applied linguistics research, and Arabic teaching and learning activities. The chapter also explores potential uses of the ALC in further research areas such as automatic Arabic

readability assessment, OCR, teaching materials development, and Arabic learner dictionaries.

- **Chapter 8** This chapter summarises the contributions of the thesis including a number of resources, proposed standards, and tools that contribute to Arabic NLP and Arabic Linguistics. It also describes some plans that have been made for future work on each component of the ALC project. The chapter discusses the challenges we faced and limitations still requiring more work before it discusses the conclusions drawn from this experimental work.

2 Literature Review and Related Work

Chapter Summary

This chapter provides a comprehensive review of 159 learner corpora under 11 categories (corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation). This review provides a quantitative view of the domain and concludes by recommending guidelines for creating a new learner corpus based on the analysis results. We used these guidelines as a basis to create the ALC. Additionally, the chapter presents a review of related work in the form of a number of existing Arabic learner corpora. The chapter discusses the rationale for creating the ALC, followed by a comparison of the existing Arabic learner corpora and the current project, ALC, in order to highlight the contribution of the latter. The comparison is based on the 11 design criteria discussed in the literature review.

2.1 Introduction

This chapter provides a review of the learner corpora domain by covering 11 aspects in a list of 159 corpora. Based on this review, recommended design criteria to develop a new learner corpus are presented. The chapter also reviews related work which include existing Arabic learner corpora.

2.2 Literature Review of Learner Corpora

At the first stage of developing the ALC, we collected data about existing learner corpora to get an idea about best practice in this kind of project. The review covered 159 learner corpora, which gives a picture about the general trend of research in the area and leads to a data-based prediction about the future. This review may not cover the whole research field of learner corpora; there may be corpora of which we are unaware and which are not covered in this review. However, the included corpora ($N = 159$) may represent the majority or at least a representative sample that enables us to generalise the results on the learner corpora field.

Since the appearance of learner corpora a few decades ago, a number of studies and surveys have investigated the state of the art of this field such as those by Pravec (2002), Granger (2004), Nesselhauf (2004), Wen (2006), Granger *et al.* (2013), Díaz-Negrillo and Thompson (2013), and Granger and Dumont (2014). However, this review is intended to include all current corpora in order to provide a quantitative view of the domain, which might be helpful in visualising the state of art of this domain. This approach may enable us to benefit from the best practice in our current project; furthermore, other researchers in learner corpora may benefit from this review in their current or planned projects.

In terms of the analysis approach, the current review presents a quantitative analysis of several aspects using the data available about those corpora. Further qualitative information is added when possible, but with an attempt not to restrict the findings of either type of analysis to specific interpretations. Such an approach may provide a different view on the data we know about learner corpora and help in monitoring the whole picture of this field. The review aims to cover 11 aspects: corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation.

In 2014, Granger and Dumont (2014) produced the Centre for English Corpus Linguistics (CECL) list of learner corpora around the world. We used the CECL list as our primary source of learner corpora due to the large number of corpora it contains; however, the list contains incomplete information for some corpora. Therefore, we searched for the original resource for each corpus to be used as a main reference. The original resources also provided further details about the corpora, such as the purpose of building each corpus and the types of annotation if the corpus includes one or more. If no reference was found, we referred to the CECL list as the only reference we had. One advantage of this list is that it contains links to a large number of references. Nevertheless, we faced challenges in finding all the references needed and finding the details required for each corpus. As a result, we found references for 137 corpora and had to rely on the CECL list for the other 22. Despite those references, information about some corpora was still not available, as illustrated in Table 2.1.

Table 2.1: Aspects covered in the review with percentage of the data not available

Aspect	Data not available
1. Corpus purpose	33%
2. Corpus size	33%
3. Target language	0%
4. Availability	32%
5. Learners' nativeness	0%
6. Learners' proficiency level	33%
7. Learners' first language	0%
8. Materials mode	0%
9. Materials genre	69%
10. Task type	3%
11. Data annotation	49%

Under each section, those corpora with unavailable data were excluded, so the analysis covers only corpora for which we were able to access information. For instance, the analysis of data annotation reflects 51% of the 159 learner corpora we reviewed. Table 2.2 includes a list of the 159 learner corpora and references from which we were able to obtain the information.

Table 2.2: Learner corpora reviewed with their references

No	Corpus	Reference
1.	Arabic Learner Corpus	
2.	Arabic Learners Written Corpus	Farwaneh and Tamimi (2012)
3.	Malaysian Corpus of Arabic Learners	Hassan and Daud (2011)
4.	The Pilot Arabic Learner Corpus	Abuhakema <i>et al.</i> (2008)
5.	The Learner Corpus of Arabic Spelling Correction	Alkanhal <i>et al.</i> (2012)
6.	The Czech as a Second/Foreign Language Corpus	Hana <i>et al.</i> (2010)
7.	The Learner Corpus Dutch as a Foreign Language	Granger and Dumont (2014)
8.	The ANGLISH Corpus	Hirst and Tortel (2010) Tortel (2008) Tortel and Hirst (2008)
9.	Asao Kojiro's Learner Corpus Data	Granger and Dumont (2014)
10.	The Barcelona English Language Corpus	Diez-Bedmar (2009)
11.	The Bilingual Corpus of Chinese English Learners	Wen (2006)
12.	The Br-ICLE corpus	Berber Sardinha (2002)
13.	The British Academic Written English Corpus	Heuboeck <i>et al.</i> (2008)
14.	The BUiD Arab Learner Corpus	Randall and Groom (2009)
15.	The Cambridge Learner Corpus	Cambridge University (2012)
16.	The Corpus of Academic Learner English	Callies and Zaytseva (2011a) Callies and Zaytseva (2011b) Callies <i>et al.</i> (2012)
17.	The Corpus of English Essays Written by Asian University Students	Ishikawa (2010)
18.	The Chinese Academic Written English Corpus	Lee and Chen (2009)
19.	The Chinese Learner English Corpus	Shichun and Huizhong (2012) Wen (2006)
20.	The City University Corpus of Academic Spoken English	Lee and Flowerdew (2012)
21.	The Cologne-Hanover Advanced Learner Corpus	Römer (2007)
22.	The College Learners' Spoken English Corpus	Wen (2006)

2 – Literature Review and Related Work

23. The Corpus Archive of Learner English in Sabah/Sarawak
Arshad (2004)
Botley (2012)
Botley and Dillah (2007)
24. The Corpus of Young Learner Interlanguage
Housen (2002)
Leacock *et al.* (2010)
25. The Eastern European English learner corpus
Granger and Dumont (2014)
26. The English of Malaysian School Students corpus
Arshad (2004)
Botley (2012)
Botley and Dillah (2007)
27. The English Speech Corpus of Chinese Learners
Hua *et al.* (2008)
28. The EVA Corpus of Norwegian School English
Hasselgren (1997)
Hasselgren (2007)
29. The GICLE corpus
Axelsson and Hahn (2001)
Granger and Dumont (2014)
30. The Giessen-Long Beach Chaplin Corpus
Jucker *et al.* (2005)
31. The Hong Kong University of Science and Technology Learner Corpus
Milton and Nandini (1994)
Pravec (2002)
32. The Indianapolis Business Learner Corpus
Connor (2012)
Connor *et al.* (1995)
33. The International Corpus of Crosslinguistic Interlanguage
Tono (2012b)
Tono (2012a)
34. The International Corpus Network of Asian Learners of English
Granger and Dumont (2014)
Ishikawa (2010)
Paulasto and Meriläinen (2012)
35. The International Corpus of Learner English
Granger (1993)
Granger (2003b)
Granger *et al.* (2010)
36. The International Teaching Assistants corpus
Thorne *et al.* (2008)
37. The ISLE Speech Corpus
Menzel *et al.* (2000)
38. The Israeli Learner Corpus of Written English
Waldman (2005)
39. The Japanese English as a Foreign Language Learner Corpus
Tono (2011)
40. The Janus Pannonius University Corpus
Pravec (2002)
41. Lancaster Corpus of Academic Written English
Banerjee and Franceschina (2012)

- | | |
|--|--|
| | Nesi (2008) |
| 42. The LeaP Corpus: Learning Prosody in a Foreign Language | Gut (2012) |
| 43. The Learner Corpus of English for Business Communication | Lan (2002) |
| 44. The Learner Corpus of Essays and Reports | Sengupta (2002) |
| 45. The Learners' Corpus of Reading Texts | Herment <i>et al.</i> (2010) |
| 46. The LONGDALE: LONGitudinal DATabase of Learner English | Meunier <i>et al.</i> (2010) |
| 47. The Longman Learner Corpus | Longman Corpus Network (2012) |
| 48. The Louvain International Database of Spoken English Interlanguage | Granger <i>et al.</i> (2012)
Kilimci (2014) |
| 49. The Malaysian Corpus of Learner English | Botley (2012) |
| 50. The Michigan Corpus of Academic Spoken English | Simpson <i>et al.</i> (2002) |
| 51. The Michigan Corpus of Upper-level Student Papers | O'Donnell and Römer (2009a)
O'Donnell and Römer (2009b) |
| 52. The Montclair Electronic Language Database | Fitzpatrick and Seegmiller (2001)
Fitzpatrick and Seegmiller (2004)
Fitzpatrick and Milton (2012)
Pravec (2002) |
| 53. The Multimedia Adult ESL Learner Corpus | Stephen <i>et al.</i> (2012) |
| 54. The Neungyule Interlanguage Corpus of Korean Learners of English | Granger and Dumont (2014)
Kwon (2009) |
| 55. The Japanese Learner of English Corpus | Izumi <i>et al.</i> (2004)
Tono (2008) |
| 56. The NUS Corpus of Learner English | Dahlmeier <i>et al.</i> (2013) |
| 57. The PELCRA Learner English Corpus | Pęzik (2012) |
| 58. The PICLE corpus | Kprzemek (2007) |
| 59. The Qatar Learner Corpus | Granger and Dumont (2014) |
| 60. The Québec Learner Corpus | Cobb (2003) |
| 61. The Romanian Corpus of Learner English | Granger and Dumont (2014) |

- | | | |
|-----|--|--|
| 62. | The Russian Learner Translator Corpus | Sosnina (2014) |
| 63. | The Santiago University Learner of English Corpus | Diez-Bedmar (2009) |
| 64. | The Scientext English Learner Corpus | Osborne <i>et al.</i> (2012) |
| 65. | The Seoul National University Korean-speaking English Learner Corpus | Kwon (2009) |
| 66. | The SILS Learner Corpus of English | Granger and Dumont (2014)
Muehleisen (2007) |
| 67. | The Soochow Colber Student Corpus | Chen (2000) |
| 68. | The Spoken and Written English Corpus of Chinese Learners | Wen (2006) |
| 69. | The Taiwanese Corpus of Learner English | Shih (2000) |
| 70. | The Taiwanese learner academic writing corpus | Granger and Dumont (2014) |
| 71. | The TELEC Secondary Learner Corpus | Pravec (2002) |
| 72. | The Telecollaborative Learner Corpus of English and German | Belz and Vyatkina (2005) |
| 73. | The Tswana Learner English Corpus | Van Rooy (2009) |
| 74. | The Uppsala Student English Corpus | Axelsson and Berglund (2002) |
| 75. | The UPF Learner Translation Corpus | Granger and Dumont (2014) |
| 76. | The UPV Learner Corpus | Granger and Dumont (2014)
O'Donnell (2010) |
| 77. | The Varieties of English for Specific Purposes Database Learner Corpus | Paquot <i>et al.</i> (2009) |
| 78. | The Written Corpus of Learner English | Mendikoetxea <i>et al.</i> (2008)
Rollinson and Mendikoetxea (2008) |
| 79. | The Yonsei English Learner Corpus | Granger and Dumont (2014) |
| 80. | The Korean Learner Corpus | Lee (2007) |
| 81. | The Estonian Interlanguage Corpus of Tallinn University | Eslon <i>et al.</i> (2012) |
| 82. | The International Corpus of Learner Finnish | Jantunen (2010) |
| 83. | The Cypriot Learner Corpus of French | Granger and Dumont (2014) |
| 84. | The COREIL Corpus | Delais-Roussarie and Yoo (2010) |
| 85. | The Dire Autrement Corpus | Hamel and Milicevic (2007) |
| 86. | The French Interlanguage Database | Granger (2003a) |

2 – Literature Review and Related Work

87. The French Learner Language Oral Corpora: Linguistic Development Corpus Myles and Mitchell (2012)
88. The French Learner Language Oral Corpora: Progression in Foreign Language Learning Myles and Mitchell (2012)
89. The French Learner Language Oral Corpora: Young Learners Corpus Myles and Mitchell (2012)
90. The French Learner Language Oral Corpora: Newcastle Corpus Myles and Mitchell (2012)
91. The French Learner Language Oral Corpora: Brussels Corpus Myles and Mitchell (2012)
92. The French Learner Language Oral Corpora: Reading Corpus Chambers and Richards (1995)
93. The French Learner Language Oral Corpora: LANGSNAP Myles and Mitchell (2012)
94. The InterFra Corpus Bartning (2011)
95. The “Interphonologie du Français Contemporain” Corpus Detey and Kawaguchi (2008)
96. The Learner Corpus French Granger and Dumont (2014)
97. The Lund CEFLE Corpus Ågren (2009)
98. The University of the West Indies Learner Corpus Peters (2009)
99. The Lexicon of Spoken Italian by Foreigners Corpus Gallina (2010)
100. The AleSKO corpus Zinsmeister and Breckle (2012)
101. Analyzing Discourse Strategies: A Computer Learner Corpus Granger and Dumont (2014)
102. The Corpus of Learner German Maden-Weinberger (2013)
103. The FALKO Corpus Granger and Dumont (2014)
Reznicek *et al.* (2012)
104. The KOLIPSI Corpus Granger and Dumont (2014)
105. The LeaP Corpus Gut (2012)
106. The LeKo Corpus Lüdeling *et al.* (2009)
107. The LINCS Corpus Granger and Dumont (2014)
108. The Telecollaborative Learner Corpus of English and German Granger and Dumont (2014)
109. The Langman Corpus Granger and Dumont (2014)
110. Corpus parlato di italiano L2 Spina *et al.* (2012)

- | | |
|---|---|
| 111. The KOLIPSI Corpus | Granger and Dumont (2014) |
| 112. The VALICO Italian Learner Corpus | Barbera and Corino (2003) |
| 113. The Korean Learner Corpus | Lee <i>et al.</i> (2009) |
| 114. The Norwegian Second Language Corpus | Tenfjord <i>et al.</i> (2006) |
| 115. The PIKUST pilot Learner Corpus | Stritar (2009) |
| 116. The Anglia Polytechnic University Learner Spanish Corpus | Granger and Dumont (2014) |
| 117. The Aprescrivlov Corpus | Granger and Dumont (2014) |
| 118. The Corpus Escrito del Español L2 | Lozano (2009) |
| 119. The Corpus of Taiwanese Learners of Spanish | Cheng <i>et al.</i> (2012)
Lu (2010) |
| 120. The DIAZ Corpus | TalkBank (2012) |
| 121. The Japanese Learner Corpus of Spanish | Granger and Dumont (2014) |
| 122. Spanish Learner Language Oral Corpus | Dominguez <i>et al.</i> (2010)
Mitchell <i>et al.</i> (2008) |
| 123. The ASU Corpus | Hammarberg (2010) |
| 124. The European Science Foundation Second Language Database | Max Planck Institute for Psycholinguistics (2012) |
| 125. The Foreign Language Examination Corpus | Granger and Dumont (2014) |
| 126. The MeLLANGE Learner Translator Corpus | Kübler (2007) |
| 127. The MiLC Corpus | Andreu <i>et al.</i> (2010)
O'Donnell <i>et al.</i> (2009) |
| 128. The USP Multilingual Learner Corpus | Tagnin (2006) |
| 129. The Padova Learner Corpus | Dalziel and Helm (2008) |
| 130. The PAROLE Corpus | Hilton (2008) |
| 131. The PolyU Learner English Corpus | Bilbow <i>et al.</i> (2004) |
| 132. The Learner Journals corpus | Xunfeng (2004) |
| 133. The corpus of English Written Interlanguage | Diez-Bedmar (2009)
Lightbound (2005) |
| 134. The Barcelona Age Factor Corpus | Diez-Bedmar (2009) |
| 135. The MADRID Corpus | Diez-Bedmar (2009) |
| 136. The ENO International Corpus of Student English | Paulasto and Meriläinen (2012) |
| 137. The Louvain International Database of Spoken English Interlanguage | Paulasto and Meriläinen (2012) |

- | | |
|--|---|
| 138. The Learner Corpus of Written Spanish | Rocha (2014) |
| 139. The Spanish Corpus of Italian Learners | Bailini (2013) |
| 140. The Bilingual Speech Corpus for French and German Language Learners | Fauth <i>et al.</i> (2014) |
| 141. The KoKo L1 Learner Corpus | Abel <i>et al.</i> (2014) |
| 142. The Advanced Learner English Corpus | Granger and Dumont (2014) |
| 143. The BATMAT Corpus | Lindgrén (2012a) |
| 144. The EFL Teacher Corpus | Kwon and Lee (2014) |
| 145. The ETS Corpus of Non-Native Written English | Blanchard <i>et al.</i> (2014) |
| 146. The Gachon Learner Corpus | Price (2013) |
| 147. The Lang-8 Learner Corpora | Komachi <i>et al.</i> (2013) |
| 148. The Learner Corpus of Engineering Abstracts | Tan <i>et al.</i> (2011) |
| 149. The Malaysian Corpus of Students' Argumentative Writing | Granger and Dumont (2014) |
| 150. The Non-native Spanish Corpus of English | Díaz-Negrillo (2012) |
| 151. The Young Learner Corpus of English | Granger and Dumont (2014) |
| 152. The Linguistic Basis of the Common European Framework for L2 English and L2 Finnish | Martin (2009) |
| 153. The Paths in Second Language Acquisition | Martin (2013) |
| 154. The Advanced Finnish Learner Corpus | Granger and Dumont (2014)
Siitonen and Ivaska (2008) |
| 155. The Finnish National Foreign Language Certificate Corpus | Granger and Dumont (2014)
Maijanen and Lammervo (2014) |
| 156. The Gaelic Adult Proficiency Corpus | Granger and Dumont (2014)
Maolalaih and Carty (2014a) |
| 157. The Spanish Learner Oral Corpus | Maolalaih and Carty (2014b) |
| 158. The University of Toronto Romance Phonetics Database | Colantoni and Steele (2004) |
| 159. The LONGLEX Project | Lindgrén (2012b) |
-

2.2.1 Purpose

Specifying the corpus purpose is usually the first step in its building process, as the design criteria should be based on and compatible with the corpus purpose. Therefore, purposes of learner corpora investigated in this section may explain some of the findings mentioned in the later sections.

2.2.1.1 Purposes Classification

Of the 159 corpora reviewed, a sizeable number (52) did not explicitly state the purpose for which they had been compiled. Purposes of the other corpora (107) were classified into two main categories: *public purposes* (for those corpora intended to be used under broad aspects of research or by a wide audience of users) and *specific purposes* (for those corpora intended to be used for investigating specific aspects or by a particular group of users).

Some stated purposes were difficult to assign to either category; however, we classified each one into the category that most closely matched our understanding of the purpose of the corpus.

Deciding whether a corpus is for public or specific purposes may affect its design criteria and content as well. The classification shows that 81 corpora (76%) were developed to be used for public purposes and 26 corpora (24%) were designed for specific purposes (Figure 2.1). This finding suggests a high interest in developing learner corpora that serve a large audience and can be used for various purposes, in addition to a longer lifetime of usability. This understanding does not negate the significant role of those corpora designed for specific purposes that have special characteristics in their design and content such as “business” or “translation” in data type and “professionals” or “immigrants” in terms of learners.

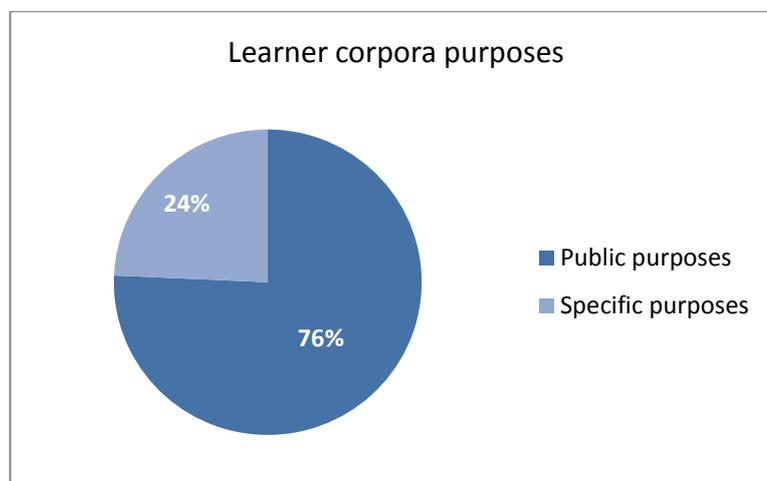


Figure 2.1: Purposes of compiling the learner corpora

We examined the corpora developed for public purposes in more detail and found that they were created for these purposes:

- Language learning/teaching,
- Interlanguage analysis,
- Materials development,
- Comparative analysis,
- Error analysis,
- Progress monitoring,
- Computer-Assisted Language Learning,
- NLP,
- Descriptive analysis,
- Translation, and
- Commercial use.

A corpus may include one or more of those purposes. For instance, the purpose of the International Corpus of Learner English is “to make use of advances in applied linguistics and computer technology to effect a thorough investigation of the interlanguage of the foreign language learner” (Granger, 1993: 57). The Japanese Learner of English Corpus (Izumi *et al.*, 2004; Tono, 2008) was designed to enable teachers and researchers to use the data for “second language acquisition research, syllabus and material design, or the development of computerized pedagogical tools, by combining it with NLP (Natural Language Processing) technology” (Izumi *et al.*, 2004: 120). Hammarberg (2010) developed the ASU Corpus to document “the

language of individual learners longitudinally at set intervals along a common time scale, so that it is possible to trace and compare stages of development within and between individuals” (p 3); it is also intended for comparisons of learner and native language production.

Figure 2.2 illustrates that “language learning and teaching” was included in 34 learner corpora. This finding is highly consistent with the definition of learner corpora mentioned previously: “[c]omputer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose” (Granger, 2002: 7). The next five purposes were mentioned in a number of corpora ranging between 9 and 14, while the remaining purposes were included in 4 corpora or less.

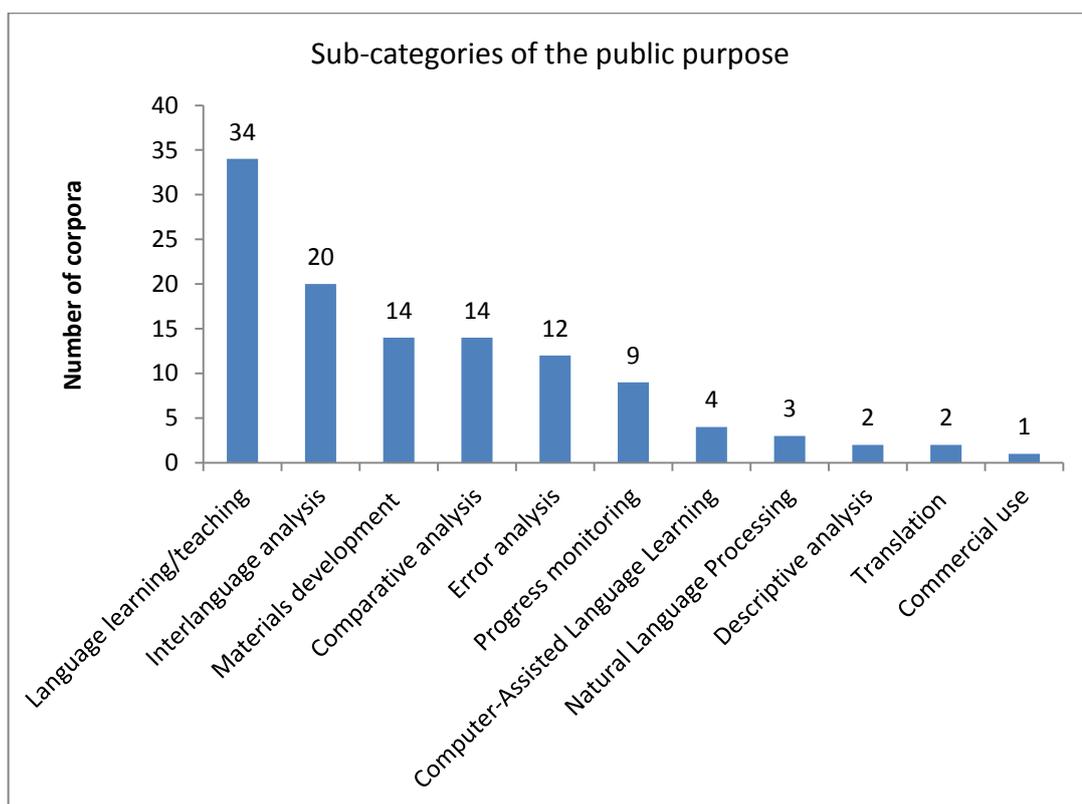


Figure 2.2: Percentages of corpora created for public purposes

In terms of the corpora with specific purposes, examples of these purposes include to examine the role of age and hours of learning, to train and test the spoken language education system, to understand the lexico-grammatical, phraseological and phonetic competence, to record lexical uses/acquisition, to improve classroom

management of content, and to describe the characteristics of contemporary academic speech. Most of these purposes were identified by 1% or 2% of the learner corpora, such as the LeaP Corpus: Learning Prosody in a Foreign Language (Gut, 2012) for description of non-native prosody, the Bilingual Speech Corpus for French and German Language Learners (Fauth *et al.*, 2014) for segmental and prosodic aspects, and the ISLE Speech Corpus (Menzel *et al.*, 2000) for training and testing the spoken language education system.

2.2.1.2 Longitudinal Learner Corpora

More than a decade ago, Granger (2002) stated that “[t]here are very few longitudinal corpora, i.e. corpora which cover the evolution of learner use. The reason is simple: such corpora are very difficult to compile as they require a learner population to be followed for months or, preferably, years” (p 11). At present, only 17 learner corpora among those 159 corpora reviewed utilised longitudinal data. Nine out of those 17 corpora were designed for public purposes, “progress monitoring” in particular. An example of a learner corpus with longitudinal data is the LONGDALE: LONGitudinal DAtabase of Learner English (Meunier *et al.*, 2010) which was designed “to build a large longitudinal database of learner English containing data from learners from a wide range of mother tongue backgrounds and thereby contribute to filling a major gap in corpus-based SLA studies” (Meunier *et al.*, 2010). Another example is the InterFra Corpus (Bartning, 2011) that was designed “to promote research in the field of French L2 second language acquisition in a developmental, interactional and variationist perspective” (Bartning, 2011). Among those longitudinal corpora, we found one which was for a specific purpose, to investigate “the role played by the age at which bilingual students begin their instruction in English as well as the hours of English classes received” (Diez-Bedmar, 2009: 922). For those corpora where purpose was not explicitly stated, “progress monitoring” seemed to be the most likely purpose.

2.2.2 Sizes

It seems that learner corpora sizes were adequate a decade ago. Granger (2003b), for example, argues that “[a] corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples but minute in the corpus linguistics field at large, where recourse to mega-corpora of several hundred million

words has become the norm rather than the exception” (p 465). She also notes that “learner corpora tend to be rather large, which is a major asset in terms of representativeness of the data and generalizability of the results” (Granger, 2004: 125). Sinclair (2005) believes that size is not a significant factor, so there is no maximum corpus size, and the minimum size of a corpus relies on two factors: “(a) the kind of query that is anticipated from users and (b) the methodology they use to study the data” (p 10). In addition, Granger (2003b) argues that learner corpora cannot be simply assessed by the number of words compared with large general corpora, but the factor equally important is the number of learners contributing. Pravec (2002) states that corpora have no uniform size because each corpus was built to address the needs of its developers. However, he emphasises the need to adequately represent the learner’s language in a corpus, though this meticulous process of compiling a learner corpus is very time-consuming.

The size of written corpora is usually measured by the number of words/tokens (w/t), whereas spoken corpora are measured by either the number of hours in the case of audio recordings or the number of w/t in the case of transcription. The current review of corpora sizes considers written data and transcripts of spoken corpora as one *textual* type (analysed based on the w/t number), while audio data is analysed separately (based on number of hours). We included only those corpora with known sizes; specifically, we evaluated 96 corpora with a w/t size and 16 corpora with a duration size, 112 in total. The total size of these 96 corpora with textual data is 134,547,037 w/t with an average of 1,401,532 w/t. The total size of the 16 oral corpora is 4,695 hours with an average of 293 hours. We used these numbers to estimate the total size of the entire 131 textual corpora and 34 oral corpora as following:

$$\text{Learner corpora with textual data} = \frac{134,547,037}{96} \times 131 = 183,600,644 \text{ w/t}$$

$$\text{Learner corpora with oral data} = \frac{4,695}{16} \times 34 = 9,976 \text{ hours}$$

It should be taken into account how valid the estimation can be, as the missing sizes in oral corpora represent 53%, while they represent only 27% in the textual type (see Table 2.3 for more detail). Therefore, it is important to emphasise that the actual sizes may differ largely from these estimated totals, which only give an estimate of statistics in existing corpora.

Table 2.3: Calculations of corpora sizes

	Textual data (w/t)	Oral data (hours)
No of corpora with known sizes	96	16
Total length	134,547,037	4695
Highest length	25,000,000	3600
Lowest length	9000	3
Average length	1,401,532	293
No of corpora with unknown sizes	35	18
Estimated length of corpora with unknown size	49,053,607	5281
Total no of corpora	131	34
Estimated length of all corpora	183,600,644	9976

A closer look at the sizes of learner corpora is given in the following two sections, which include only those corpora with known sizes (96 textual and 16 oral) in order to have a more accurate analysis about the sizes.

2.2.2.1 Textual Data

Textual data is predominant in learner corpora. Table 2.3 shows 131 corpora with textual data and 34 with oral data. The analyses of textual corpora data sizes reveal that the majority are concentrated in the smaller size groups. For instance, Figure 2.3 shows that most textual corpora tend to be 4 million w/t or less. Examples include the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010) with 3,700,000 w/t, the Michigan Corpus of Upper-level Student Papers (O'Donnell and Römer, 2009a, 2009b) with 2,600,000 w/t, and the ENO International Corpus of Student English (Paulasto and Meriläinen, 2012) with 2,250,000 w/t.

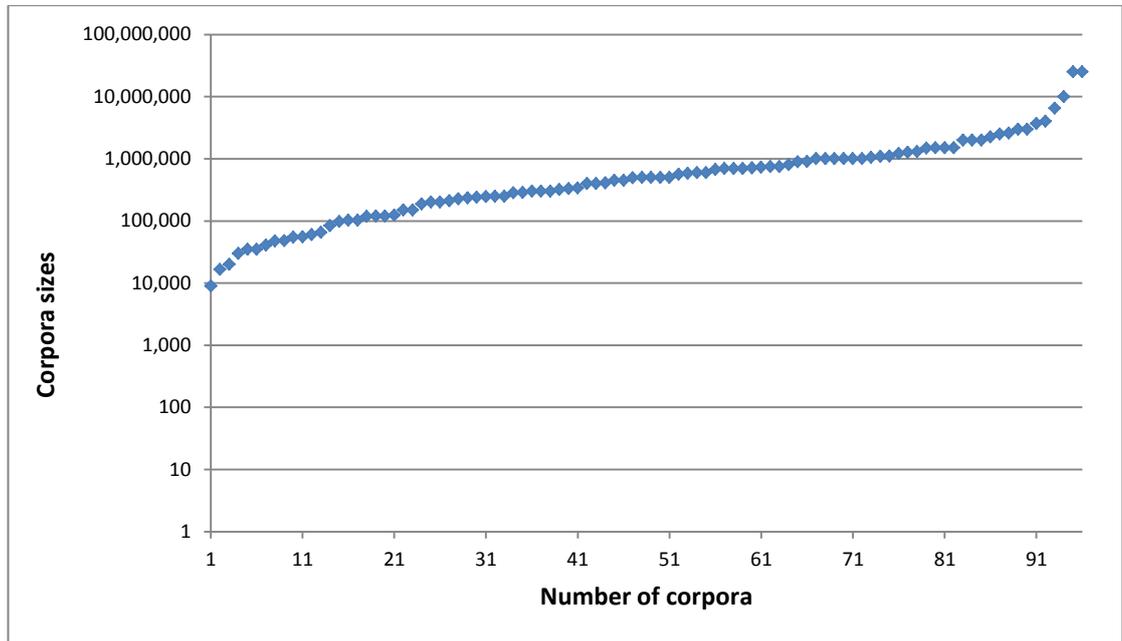


Figure 2.3: Sizes of all textual corpora based on w/t sizes

Figure 2.4 presents a closer look at this group (4 million w/t or less). The figure reveals that the majority lie at the bottom (1 million w/t or less). For example, the Seoul National University Korean-speaking English Learner Corpus (Kwon, 2009) contains 899,505 w/t, the Written Corpus of Learner English (Rollinson and Mendikoetxea, 2008) consists of 750,000 w/t, and the Taiwanese Corpus of Learner English (Shih, 2000) includes 730,000 w/t.

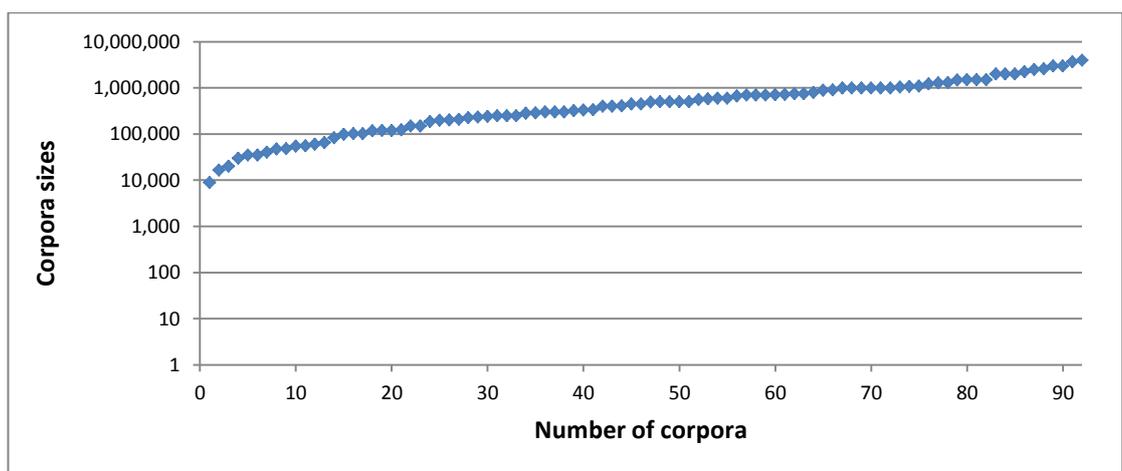


Figure 2.4: Sizes of textual corpora with 4 million w/t or less

Figure 2.5 gives a further focus on this specific group of 1 million or less. The figure shows that the highest number of corpora is again concentrated in the bottom group (200,000 w/t or less). Examples of this group include the Corpus of English Essays Written by Asian University Students (Ishikawa, 2010) with 200,000 w/t, the EVA Corpus of Norwegian School English (Hasselgren, 1997, 2007) with 102,343 w/t, and the Learner Corpus of English for Business Communication (Lan, 2002) with 117,500 w/t.

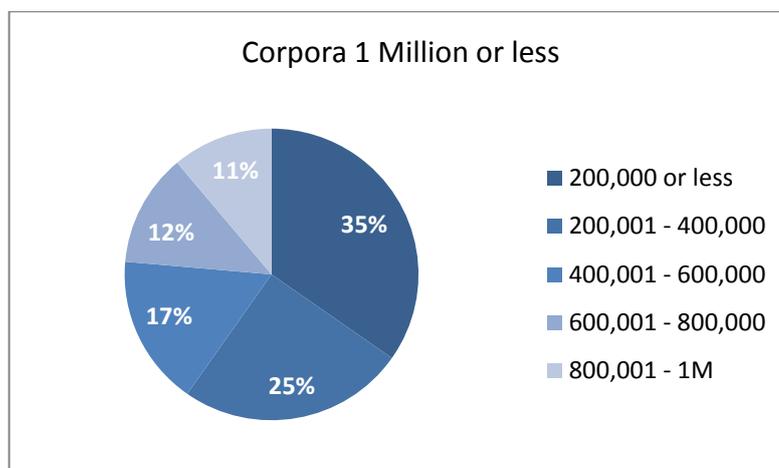


Figure 2.5: Number of textual corpora with 1 million w/t or less

2.2.2.2 Oral Data

A number of researchers (Branbrook 1996, Kennedy 1998, Thompson 2005) highlight the difficulties in compiling spoken corpora. Learner corpora are not unique in terms of these difficulties, as the proportion of spoken data is still much less than written data (see Section 2.2.8 for more details about materials mode, both written and spoken). Upon reviewing the sizes of 16 out of the 34 learner corpora, we found that 9 corpora (56%) are 50 hours or less (Figure 2.6), and 7 out of those 9 contain between 3 and 20 hours. For example, the ISLE Speech Corpus (Menzel *et al.*, 2000) contains 18 hours, the Spanish Learner Oral Corpus (Maolalagh and Carty, 2014b) contains 14 hours, and the LeaP Corpus: Learning Prosody in a Foreign Language (Gut, 2012) contains 12 hours.

The number of learner corpora that include oral data was not large enough to gain a deeper insight into their clusters as we did with the textual corpora. Sizes may

increase when the need for transcription is minimised or even dispensable by using new techniques of processing, analysing, and probably searching audio files directly.

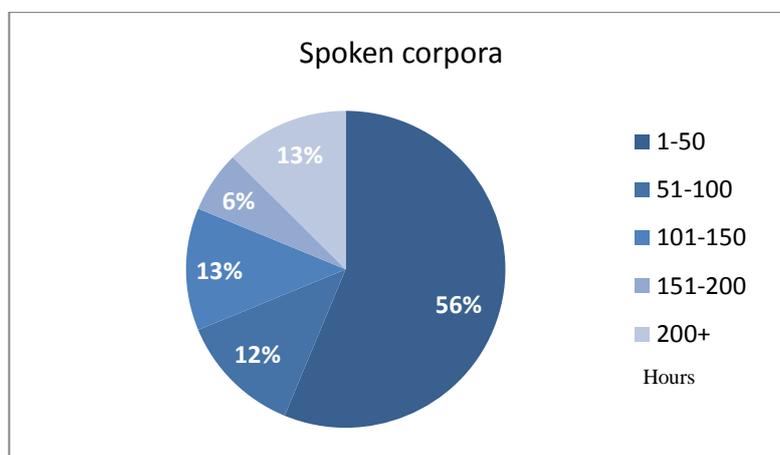


Figure 2.6: Number of spoken corpora based on length (hours)

The difficulty in collecting data from specific people (language learners) may lead the corpus developer to minimise the size of his corpus especially in the first versions. However, it can be expected that the continuous improvement in the techniques of collecting, computerising, and annotating corpora as well as the growing interest in using larger learner corpora may lead some existing corpora to expand as well as new large ones to emerge, particularly for general research purposes.

2.2.3 Target Language

The target language refers to the language used to produce the corpus materials, which is the language to be investigated. Some corpora include more than one language; however, the majority (90%) contain data of a single language (Figure 2.7). This finding may indicate that studies tend to be within one language rather than across languages. Several corpora can be used to undertake interlanguage studies, but it is important to ensure they include comparable materials. One of the options that developers use is to create a comparable learner corpus that includes similar materials of multiple target languages. This type represents 6% of the existing learner corpora. Corpora involving multiple languages are beneficial when researchers need to investigate the effect of learners' L1 on second or foreign language acquisition, particularly if the corpus contributors share the same L1. Some

corpora of this type exist, such as the Foreign Language Examination Corpus (Bański and Gozdawa-Gołębiowski, 2010) which includes data of three target languages, English, German, and French, produced by students sharing one L1, Polish. The creators anticipate that this corpora “will allow for measuring the influence of the Polish language on the acquisition of target-language structures” (Bański and Gozdawa-Gołębiowski, 2010: 56).

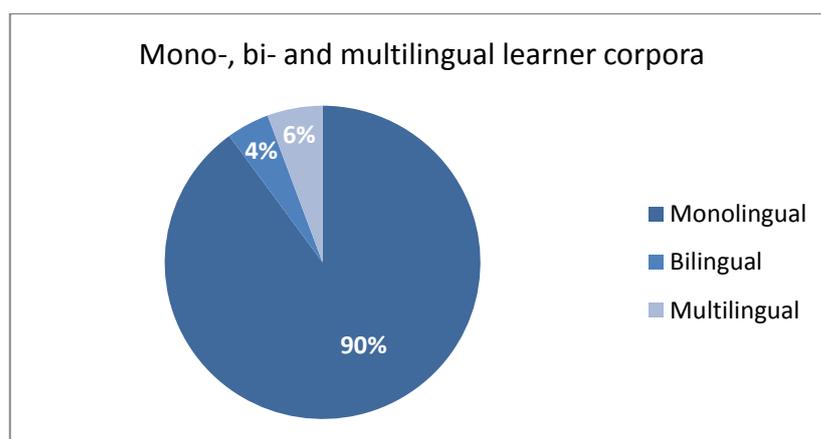


Figure 2.7: Learner corpora distribution based on target languages included

Figure 2.8 shows that 20 languages were targeted by the 159 learner corpora reviewed. The figure also shows how many times each language was targeted (without distinguishing between monolingual, bilingual, and multilingual corpora). The remarkable point we can see in Figure 2.8 is that “English clearly dominates the learner corpus scene” (Granger, 2008: 262) with more than 90 corpora. In fact, the significance of the discrepancy is clear when comparing English with French, the second most prevalent language, which is included in only 21 corpora. Among the 20 languages identified, 11 were targeted only once. This distribution of targeted languages may suggest the extent of the spread of teaching each language around the world and, consequently, the amount of research being conducted on them. In theory, languages being taught more may have more research in different aspects of learning and teaching, and thus have more learner corpora.

We expect that English, as an international language, may continue to dominate the field of learner corpora. However, many more languages might be targeted in the future to develop necessary resources that would allow researchers to conduct corpus-based studies in language learning and teaching as well as some other

relevant domains such as NLP, computer-assisted language learning, and automatic language correction. Additionally, the rapid progress in the tools used to collect, digitise, organise, annotate, distribute, and analyse the data may help researchers to develop language resources for their own languages with less effort than in the past.

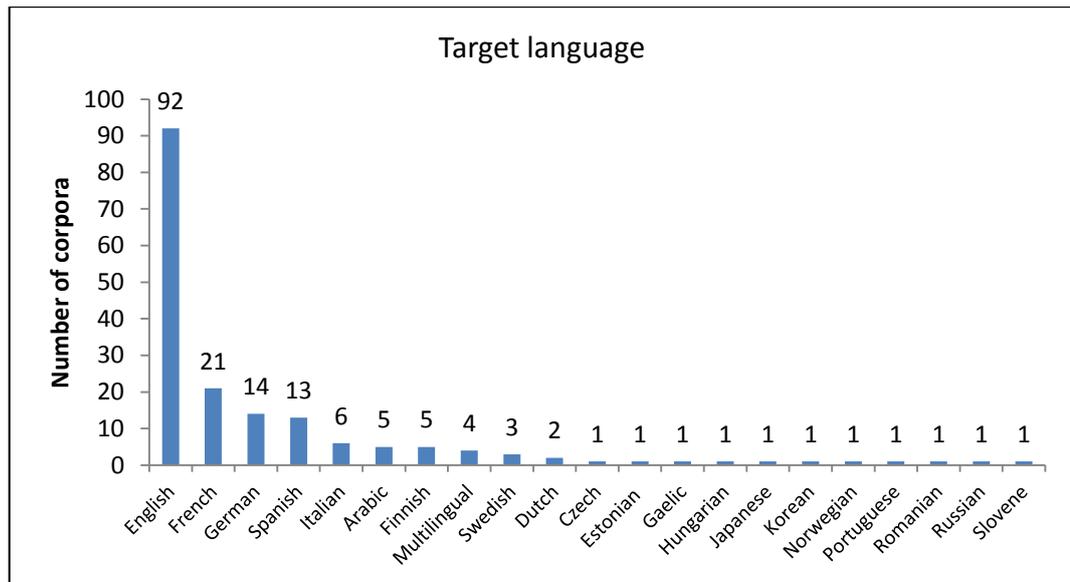


Figure 2.8: Target languages in learner corpora

2.2.4 Data Availability

We classified learner corpora into three main types. The first category contains those corpora that are freely available online for search or download, including those that are ready and intended to be publicly available. This category includes 66 corpora, representing the highest percentage (61%). For example, the Michigan Corpus of Upper-level Student Papers (O'Donnell and Römer, 2009a, 2009b) is searchable online, the data of the Arabic Learners Written Corpus (Farwaneh and Tamimi, 2012) is available for download, and the Spanish Learner Language Oral Corpus (Dominguez *et al.*, 2010; Mitchell *et al.*, 2008) is both searchable online and has data available for download.

The second category includes 29 learner corpora (27%) that are restricted to a specific research community whose members must input a username and password to receive access, such as the Chinese Learner English Corpus (Shichun and Huizhong, 2012; Wen, 2006), or that have paid access. The International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010) is an example of a

corpus with paid access, as it is distributed on CD-ROM via an online purchase order.

The third category includes 13 corpora (12%) still under development at the time of preparing the final updated version of this review (Figure 2.9). The Pilot Arabic Learner Corpus (Abuhakema *et al.*, 2008) is an example of this type. We do not know whether access to a given corpus in the third category will be free or restricted, making these corpora unsuitable for the present analysis.

Excluding the third category, we can see that the number of freely available learner corpora is more than double those restricted even though access to the largest two learner corpora is restricted. These two corpora are the Hong Kong University of Science and Technology Learner Corpus (Milton and Nandini, 1994; Pravec, 2002) and the Cambridge Learner Corpus (Cambridge University, 2012) with 25 million w/t in each. The tendency to make learner corpora freely available is consistent with that tendency (mentioned in Section 2.2.1) to develop corpora for public purposes to allow a wider audience of researchers to re-use the data for their own purposes.

Different file formats, such as TXT and XML, are used for written learner corpora available for download, while MP3 and WAV are the most commonly used file formats for spoken corpora. Typically, each corpus file contains a single written or spoken text either with or without its metadata and annotation.

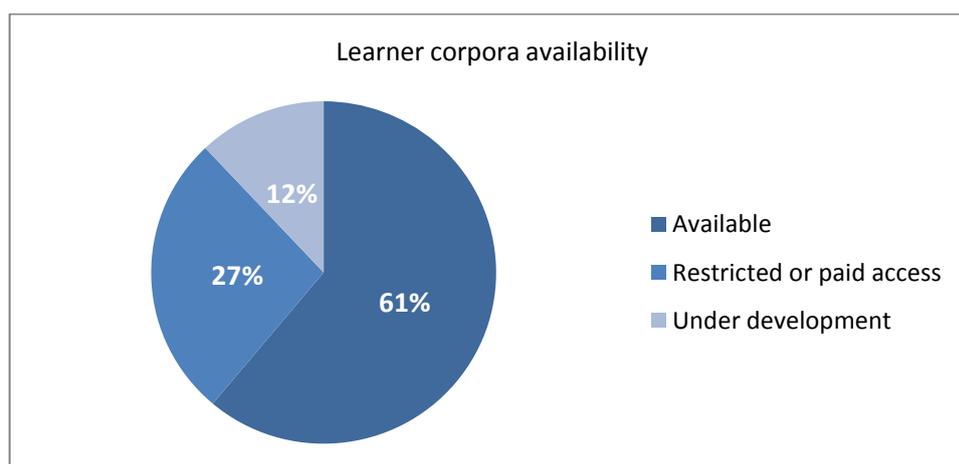


Figure 2.9: Availability of learner corpora

2.2.5 Learners' Nativeness

Based on Granger's (2002) definition of *learner corpora*, those corpora are usually designed for SLA/FLT purpose. As a result, we expected to see that most of them contain data from NNS of the target language with much less focus on those including NS data. Figure 2.10 illustrates that 124 learner corpora (78%) include data from only NNS such as the Uppsala Student English Corpus (Berglund and Axelsson, 2012), the NUS Corpus of Learner English (Dahlmeier *et al.*, 2013), and the Learner Corpus of Essays and Reports (Sengupta, 2002). We found 32 corpora (20%) with data from both NS and NNS, which is mostly for comparative purposes. Examples of this type include the ASU Corpus (Hammarberg, 2010), the Corpus of English Essays Written by Asian University Students (Ishikawa, 2010) and the ENGLISH Corpus (Hirst and Tortel, 2010; Tortel, 2008; Tortel and Hirst, 2008).

A few corpora (2%) contain data from only NS, such as the Learner Corpus of Arabic Spelling Correction (Alkanhal *et al.*, 2012) and the KoKo L1 Learner Corpus (Abel *et al.*, 2014). Presumably, this type includes L1 learners while they were learning more about their first language. The purposes of such corpora may include investigating language use, errors, and monitoring progress of the native speakers while learning. The reason behind this very small number of NS learner corpora may lie in the belief that learner corpora are based on the nativeness factor regardless of the context of data production; as a result, when corpus content is produced by native speakers in a language learning context, it is considered as a "general corpus of NS" and not a "learner corpus of NS". Thoday (2007), for example, believes that "learner corpora focus specifically on language produced by L2 learners" (p 146); this belief is based on the aforementioned *learner corpora* (Granger, 2002). Another possibility appears in relying on a general corpus of native speakers (as a native comparable corpus) when undertaking comparisons between the language of native and non-native speakers, even though it is clear that the data was not produced in a learning context. Such comparisons may simply mean that researchers see no need to build a particular learner corpus of NS while many easy, accessible, and free general corpora of NS exist.

In terms of those existing learner corpora that combine NS and NNS data, it is not clear whether the NS part was produced in a learning context, as obtaining this information would require more investigation into those parts.

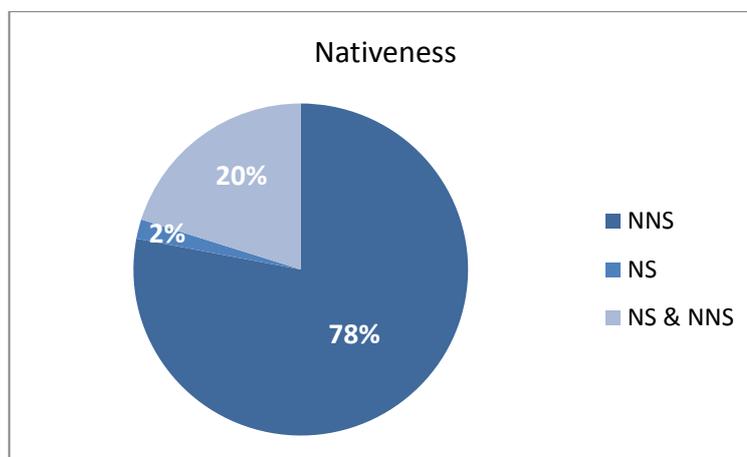


Figure 2.10: Data of native and non-native speakers

2.2.6 Learners' Proficiency Level

Proficiency levels in most learner corpora are described as “Beginning”, “Intermediate”, and “Advanced”. Some corpora, however, prefer to use the Common European Framework of Reference (Council of Europe, 2001), which includes three equivalent levels: A, B, and C. In this section, we excluded 52 corpora which use different level indicators that are not comparable with the three levels aforementioned (e.g. they indicate learner proficiency based on education level, degree, etc.) or for which we were unable to access information about learners' proficiency level.

Of the remaining 107 learner corpora, 58 corpora (54%) include all three levels (Beginning, Intermediate, and Advanced) as illustrated in Figure 2.11. This type includes, for example, the Spanish Learner Language Oral Corpus (Dominguez *et al.*, 2010; Mitchell *et al.*, 2008), the LeaP Corpus: Learning Prosody in a Foreign Language (Gut, 2012), and the Estonian Interlanguage Corpus of Tallinn University (Eslon *et al.*, 2012). This finding may indicate an interest in the kind of studies that include comparative analysis between learners from different levels.

We also identified a second category of corpora that include Intermediate and Advanced levels, such as the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), followed by those that identify Advanced alone, e.g. the Learner Corpus of English for Business Communication (Lan, 2002). This finding reveals the importance of the Advanced level in learner corpora. The Intermediate level also has some importance, particularly when it appears alongside

other levels. The Beginning level received the least attention among those three levels of proficiency.

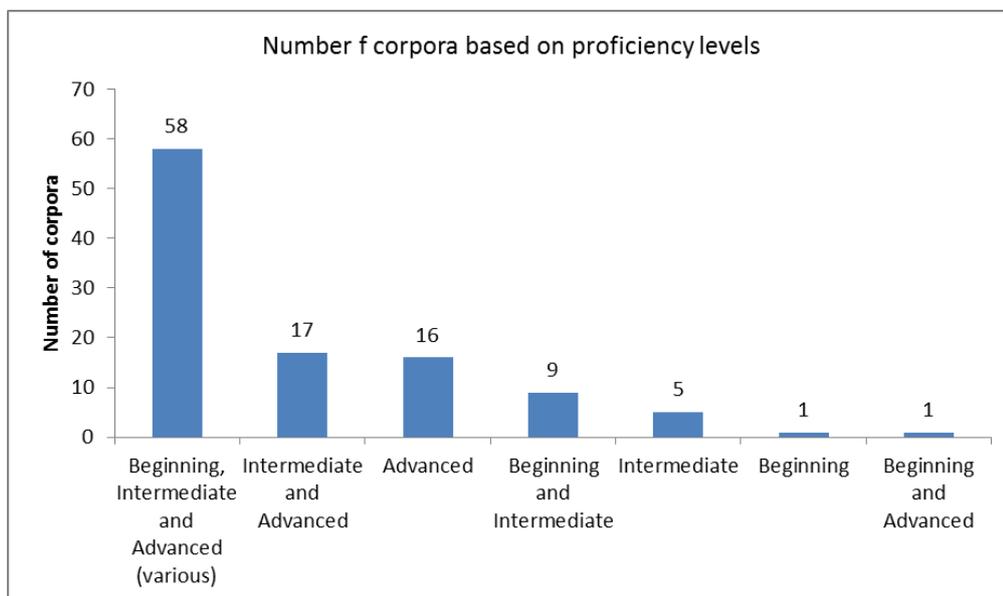


Figure 2.11: Number of corpora based on proficiency levels included

Considering these levels separately (i.e. by calculating how many times each level is included in a learner corpus regardless of whether it appears with other levels) reveals a relative balance, but the Advanced and Intermediate levels are still more prevalent (Figure 2.12).

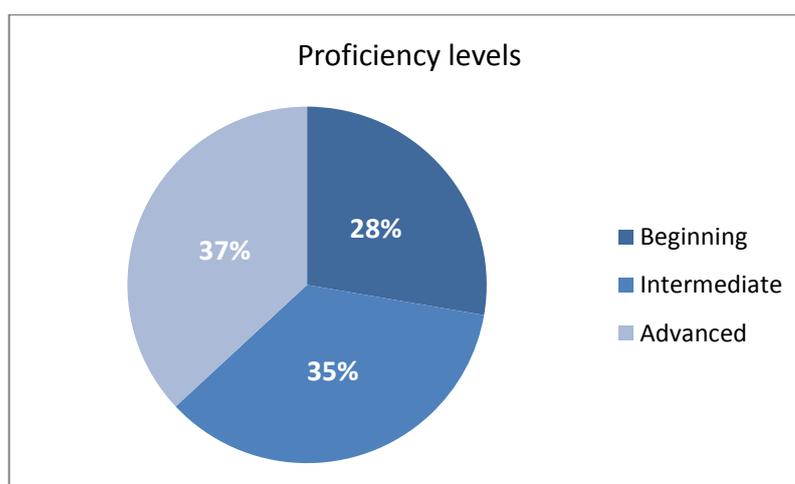


Figure 2.12: Proficiency levels distribution

2.2.7 Learners' First Language

It was not possible to show a comprehensive distribution of the first languages of the existing corpora. Many of them declare that they include students from various L1s but do not list those languages. Thus, we had to classify the corpora into two main categories. The first category includes those that have various L1s (89 corpora, 56%), and the second includes those with a single L1 (70 corpora, 44%) as seen in Figure 2.13. In Learners' Proficiency Level section, we noted an indication of interest in comparative studies; thus, it is not surprising in the current section to see that 89 learner corpora include various first languages, which highlights a similar interest in comparisons but between learners from different L1s in this case.

Examples of those corpora that have various L1s include the Corpus of Academic Learner English (Callies and Zaytseva, 2011a, 2011b; Callies *et al.*, 2012), the Giessen-Long Beach Chaplin Corpus (Jucker *et al.*, 2005), and the Indianapolis Business Learner Corpus (Connor, 2012; Connor *et al.*, 1995). In contrast, examples of those corpora that contain data from a sole L1 include the Japanese English as a Foreign Language Learner Corpus (Tono, 2011) with Japanese L1 learners, the Learners' Corpus of Reading Texts (Herment *et al.*, 2010) with French L1 learners, and the Learner Corpus of Essays and Reports (Sengupta, 2002) with Chinese L1 learners.

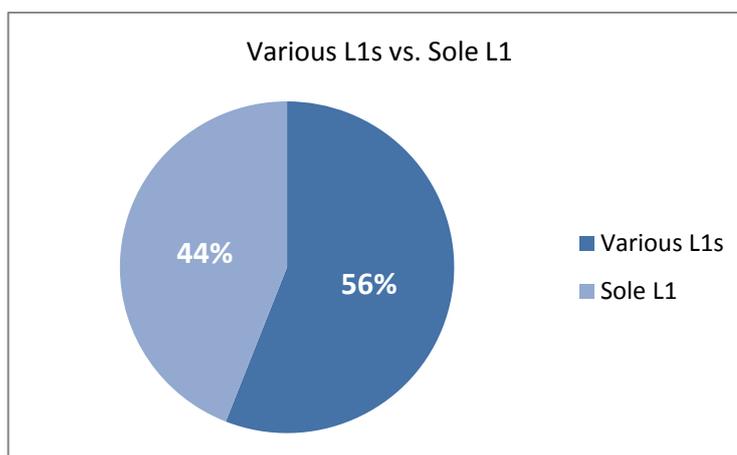


Figure 2.13: Corpora with various L1s vs. sole L1

In terms of those corpora focussing on a sole first language, Chinese-speaking students received the highest attention with 14 corpora including the Hong Kong University of Science and Technology Learner Corpus (Milton and Nandini, 1994; Pravec, 2002) which contains 25 million w/t of written data, the Spoken and Written

English Corpus of Chinese Learners (Wen, 2006) which contains 4 million w/t of written and spoken materials, and the NUS Corpus of Learner English (Dahlmeier *et al.*, 2013) which includes 1M w/t of written data. Aside from those concerning Chinese, the number of corpora focussing on a single first language ranges between 1 and 5 per language (Figure 2.14).

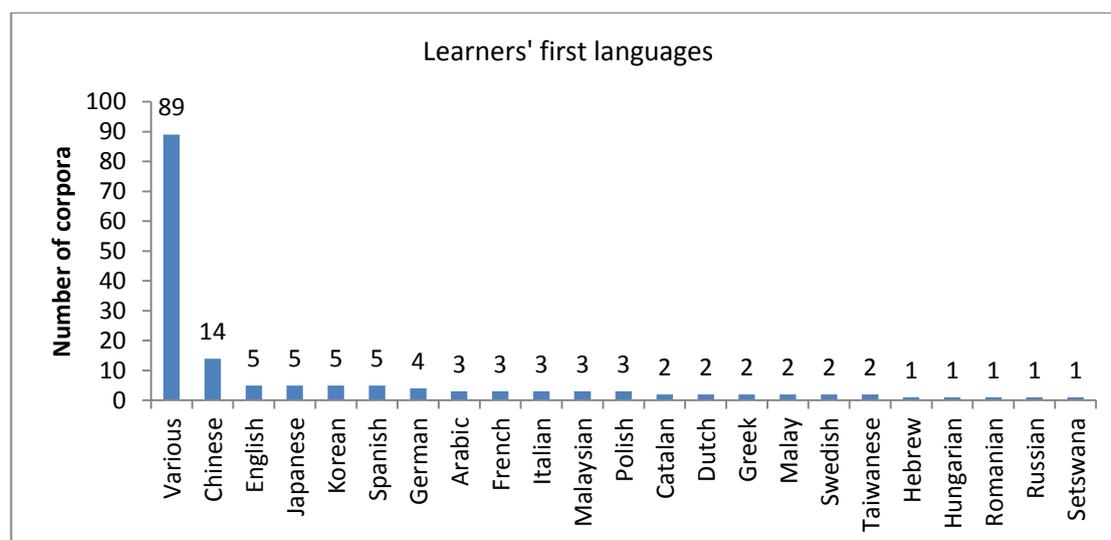


Figure 2.14: First languages in learner corpora

2.2.8 Material Mode

The term *materials mode* refers to whether the language originates in speech or writing (Sinclair, 2005). Compiling a corpus of hand-written texts is somewhat similar to compiling an oral corpus as both may need equivalent processes, particularly the step of converting data into a textual computerised format. Using tools for processing spoken data such as ELAN (Hellwig, 2014), Praat (Boersma and Weenink, 2014), Anvil (Kipp, 2001), EXMARaLDA (Schmidt and Wörner, 2009), and others¹ allows annotation to be added to the audio files directly with no essential need for the transcription process. However, McEnery (2003) highlights the benefits of building a spoken corpus that combines sound recordings and orthographic transcription; specifically, doing so enables the retrieval of words from the transcription and inspection of the original acoustic context in which the word was produced.

¹ See a list of this kind of processing software in Pustejovsky and Stubbs (2013).

Figure 2.15 reveals that two-thirds (66%) of the learner corpora include solely written data. This category includes the largest two learner corpora, the Longman Learner Corpus (Longman Corpus Network, 2012) and the Hong Kong University of Science and Technology Learner Corpus (Milton and Nandini, 1994; Pravec, 2002), each of which contains 25 million words. The learner corpora include solely spoken data represent 26%. Examples of this type are the COREIL Corpus (Delais-Roussarie and Yoo, 2010) and the French Learner Language Oral Corpora (Myles and Mitchell, 2012). Only 7% of learner corpora contain both modes, written and spoken, e.g. the ASU Corpus (Hammarberg, 2010) and the Santiago University Learner of English Corpus (Diez-Bedmar, 2009). Our findings revealed one remarkable multimodal corpus that includes written, spoken, and video data, the Multimedia Adult ESL Learner Corpus (Stephen *et al.*, 2012). The multimodal type could be able to provide more details about the learner language.

In line with our findings, Kennedy (1998) observes that most corpus-based grammatical and lexical studies of English have so far been based on written-corpora analysis but notes that spoken language represents the most common mode of language. Expressing the same concern about the dominance of written corpora Leech (1997) suggests that a corpus should “contain at least as many spoken materials as written materials” (p 17). Mauranen (2007) also suggests that “[w]hen we seek to capture language patterns in the process of ongoing change, the best data can be expected from spoken corpora rather than written, because speech is more sensitive to new trends” (p 41).

Compared with their written language counterparts, researchers creating spoken language corpora may encounter some difficulties, for example dealing with extra processes such as audio recording, converting these recordings into a written form, and sometimes annotating this written form for phonetic and prosodic features. These additional processes are laborious, time-consuming, and expensive to undertake (Branbrook, 1996; Kennedy, 1998; Thompson, 2005), which may help explain the lack of spoken corpora. However, some relatively new insights into the essential nature of language use can be explored only through spoken language corpora (Kennedy, 1998).

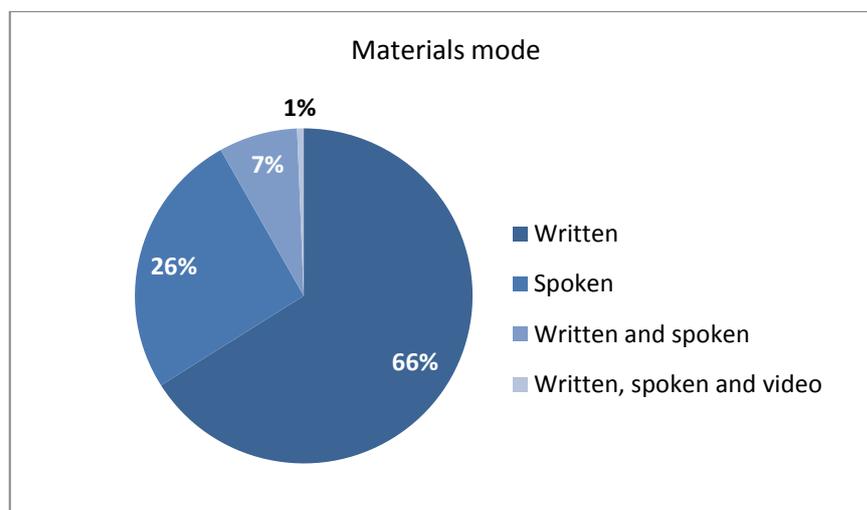


Figure 2.15: Materials modes in learner corpora

2.2.9 Material Genre

When building a corpus, “the question of what genres to include is not straightforward. There is, for example, no comprehensive taxonomy of genres from which to select” (Kennedy, 1998). However, some insights can be derived from reviewing existing corpora. We encountered some difficulties in ensuring that all genres used in learner corpora were distinguished properly, but our findings suggested 14 genres (see Table 2.4), of which Argumentative, Narrative, and Descriptive materials were the most used respectively. For instance, the Scientext English Learner Corpus (Osborne *et al.*, 2012) includes Argumentative materials; the Multilingual Learner Corpus (Tagnin, 2006) contains Argumentative and Narrative Essays; the French Interlanguage Database (Granger, 2003a) comprises Argumentative, Descriptive, and Narrative data; the Lund CEFLE Corpus (Ågren, 2009) includes Descriptive and Narrative materials; and the Taiwanese Corpus of Learner English (Shih, 2000) contains four genres: Argumentative, Narrative, Descriptive, and Expository.

Table 2.4: Genres used in learner corpora

Genre	No of corpora
1. Argumentative	30
2. Narrative	23
3. Descriptive	21
4. Discussion	5
5. Expositive	3
6. Journalistic	3
7. Informative	2
8. Administrative	1
9. Explanation	1
10. Injunctive	1
11. Legal	1
12. Persuasive	1
13. Reflective	1
14. Technical	1

The number of genres ranges from one to four in each corpus, as Figure 2.16 illustrates, with most learner corpora including one or two genres.

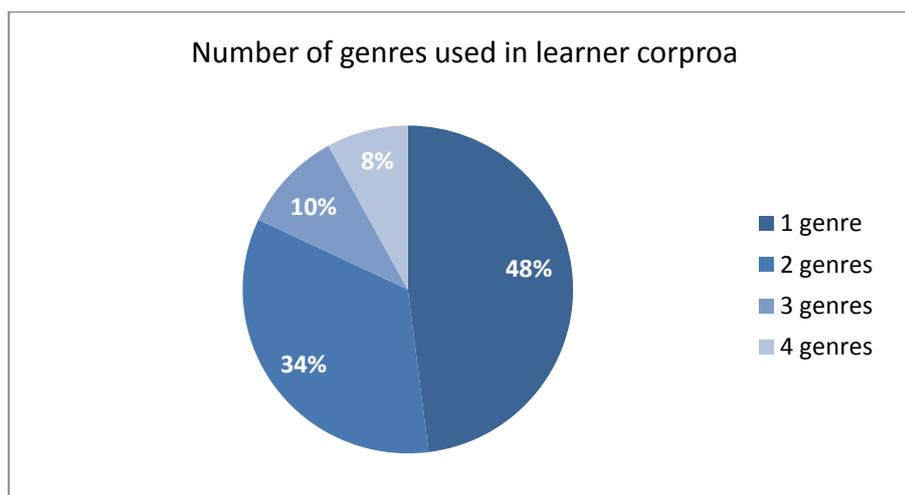


Figure 2.16: Number of genres included in learner corpora

2.2.10 Task Type

With respect to task types, we counted and listed (see Table 2.5) all labels used to indicate task type in learner corpora on the assumption that the corpus developers had their own specifications for using each of these labels, even though some may indicate similar types (e.g. Speech, Oral task, and Talk). The task types list suggests that Essays are the most preferable in written tasks, and Interviews in those spoken. The next most common types are Test and Exam, which can be either written or spoken, followed by Letter and then the other less common types.

For example, the Cologne-Hanover Advanced Learner Corpus (Römer, 2007) is a written corpus that used essays as the sole task type. In contrast, the Corpus of Young Learner Interlanguage (Housen, 2002; Leacock *et al.*, 2010) is a spoken corpus that used only interviews to collect its data. Three task types (i.e. Essays, Interviews, and Tests) were used to collect the data in the Czech as a Second/Foreign Language Corpus (Hana *et al.*, 2010), which is a written and spoken corpus.

Table 2.5: Task types used in learner corpora

1. Essays	77	18. Interaction	3	35. Application letter	1
2. Interview	24	19. Mail/Email	3	36. Curriculum Vitae	1
3. Test	17	20. Role-play	3	37. Debate	1
4. Exam	16	21. Presentation	3	38. Fax	1
5. Letter	11	22. Questions and answers	3	39. Imitation	1
6. Conversation	8	23. Sentences	3	40. Instruction	1
7. Reading	8	24. Speech	3	41. Language class	1
8. Story	8	25. Telling	3	42. Lecture	1
9. Report	7	26. Word list	3	43. Memo	1
10. Composition	6	27. Dialogue	2	44. Newspaper	1
11. Summary	6	28. Diaries	2	45. Recount	1
12. Assignment	5	29. Exercises	2	46. Repeat	1
13. Dissertation	4	30. Monologue	2	47. Review	1
14. Paper	4	31. Oral task	2	48. Social networking	1
15. Thesis	4	32. Proposal	2	49. Talk	1
16. Translation	4	33. Resume	2	50. Teaching	1
17. Abstract	3	34. Communication	2	51. Tutorial	1

Of the learner corpora examined, 58% included a sole task type. For example, the Corpus Escrito del Español L2 (Lozano, 2009) includes Compositions, the Korean Learner Corpus (Lee *et al.*, 2009) contains Assignments, and the Russian Learner Translator Corpus (Sosnina, 2014) consists of Translations. Using a single type to collect the data enables researchers to avoid any distortion in the results of their studies, though doing so prevents any comparative analysis in terms of task type. In contrast to those corpora which rely on a single task type, some corpora used four, five, or seven types to collect their data. For example, the MiLC Corpus (Andreu *et al.*, 2010; O'Donnell *et al.*, 2009) used seven task types: Letters, Summaries, Curriculum Vitae, Essays, Reports, Translations, and Communication.

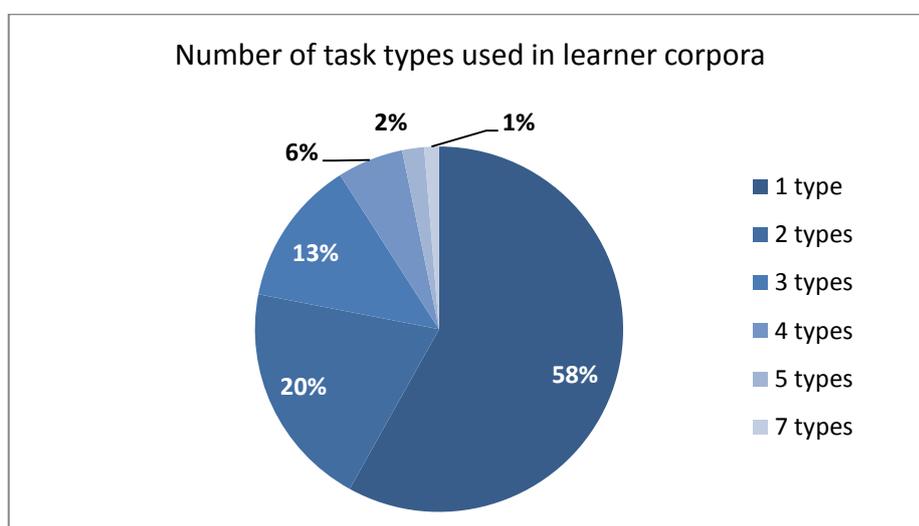


Figure 2.17: Number of task types included in learner corpora

2.2.11 Data Annotation

For around half of those learner corpora we reviewed, we were not able to determine whether they include any type of annotation. However, of the corpora that did address annotation, 82% were tagged with one or more types of annotation, and 18% included raw data only (Figure 2.18).

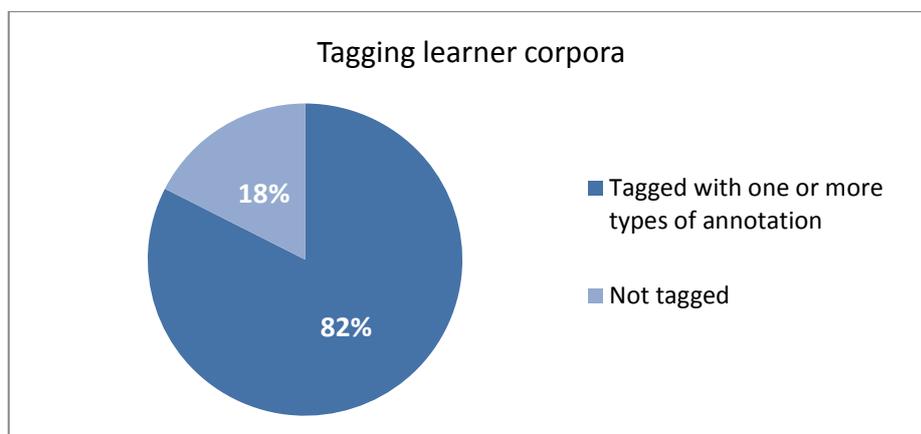


Figure 2.18: Learner corpora tagging

Table 2.6 shows examples of learner corpora and the annotations they include.

Table 2.6: Examples of learner corpora annotation

Corpus	Type of annotation
The Japanese Learner of English Corpus (Izumi <i>et al.</i> , 2004; Tono, 2008)	Spoken phenomena (see example in Figure 2.19)
The Norwegian Second Language Corpus (Tenfjord <i>et al.</i> , 2006)	Part-of-Speech (PoS), morpho-syntactic features, and errors
The KoKo L1 Learner Corpus (Abel <i>et al.</i> , 2014)	Lemma, graphical arrangement, PoS, and error
The Czech as a Second/Foreign Language Corpus (Hana <i>et al.</i> , 2010)	Errors and structural features (see example in Figure 2.20)
The Foreign Language Examination Corpus (Bański and Gozdawa-Gołębiowski, 2010)	Grammatical and error tagging (see example in Figure 2.21)

```
<head version="1.3">
  <date>1999-12-16</date>
  <sex>female</sex>
  <age></age>
  <country>Japan</country>
  <overseas></overseas>
  <category></category>
  <step>1.5</step>
  <TOEIC>765</TOEIC>
  <TOEFL></TOEFL>
```

2 – Literature Review and Related Work

```
<other_tests></other_tests>
<SST_level>6</SST_level>
<SST_task2>restaurant</SST_task2>
<SST_task3>train_advanced</SST_task3>
<SST_task4>department store</SST_task4>
</head>

...

<stage2>
  <task>
    <A>I see. O K. Now, let me show you the first picture. Please describe this
    picture.</A>
    <B>O K. <F>Er</F> <R>this is a</R> this is a <.></.> room in a hotel. And
    <.></.> <F>oh</F> sorry, it's not. Yeah, I think it's a restaurant. And there are
    three tables, <R>and</R> and there are three couples and <SC>two server</SC> two
    <R>waiter</R> waiter are serving. And <R>in the</R> in the middle of the restaurant,
    the couple is <F>er</F> drinking wine. And <F>err</F> the man is <.></.> testing the
    wine and saying something to the waiter. Maybe he is sommelier. And <R>he</R> he show
    the bottle to the man. I guess he is explaining something. And <F>er</F> the couple,
    <F>er</F> they dressed very nicely. <CO><R>And</R> <.></.> <F>mhmm</F> <R>and</R>
    <.></.> <R>and</R> <F>well</F> and</CO>. <.></.></B>
  </task>
  <followup>
    <A>O K.</A>
    <B>O K?</B>
    <A>O K. Thank you very much. <F>Er</F> how do you spend time with your
    husband?</A>
    <B><.></.> You mean, in our free time?</B>
    <A><F>Mhmm</F>.</A>
    <B><F>Er</F> like this? <.></.> <F>Well</F> <F>er</F> <R>I</R> I sometimes
    eating out with my husband. But we don't get dressed like this. <nvs>laughter</nvs>
    <.></.></B>
    <A>Can you compare the restaurant you often go to to this picture?</A>
    <B><nvs>laughter</nvs> It's very different from restaurant to we often go. We
    often go to a kind of family style restaurant <.></.> such as Denny's or Skylark. So
    I wish I could <SC>go like</SC> go to a nice restaurant like this.</B>
    <A><F>Er</F> what is good about family-type restaurant?</A>
    <B><F>Well</F> <SC>fir</SC> at first, it's very cheap and they served very
    quickly. And, <F>er</F> most of the cases, <F>er</F> that kind of restaurant is in
    suburb, so <SC>people are very</SC> <F>er</F> people can go there very easily. I
    think they are good point of family-type restaurant.</B>
  </followup>
</stage2>
```

Figure 2.19: Example of annotation from the Japanese Learner of English Corpus

```
<?xml version="1.0" encoding="UTF-8"?>
<adata xmlns="http://utkl.cuni.cz/czesl/">
  <head>
    <schema href="adata_schema.xml" />
    <references>
      <ref id="w" name="wdata" href="r049.w.xml" />
    </references>
  </head>
  <doc id="a-r049-d1" lowerdoc.rf="w#w-r049-d1">
    ...
    <para id="a-r049-d1p2" lowerpara.rf="w#w-r049-d1p2">
    ...
```

```

    <s id="a-r049-d1p2s5">
      <w id="a-r049-d1p2w50">
        <token>Bál</token>
      </w>
      <w id="a-r049-d1p2w51">
        <token>jsem</token>
      </w>
      <w id="a-r049-d1p2w52">
        <token>se</token>
      </w>
      ...
    </s>
    ...
    <edge id="a-r049-d1p2e54">
      <from>w#w-r049-d1p2w46</from>
      <to>a-r049-d1p2w50</to>
      <error>
        <tag>unk</tag>
      </error>
    </edge>
    <edge id="a-r049-d1p2e55">
      <from>w#w-r049-d1p2w47</from>
      <to>a-r049-d1p2w51</to>
    </edge>
    ...
  </para>
  ...
</doc>
</adata>

```

Figure 2.20: Example of annotation from the Czech as a Second/Foreign Language Corpus

Grammatical layer

a. CLAWS c5

```

<s xml:id="morph_1.1-s">
  <seg ana="PNP"
    corresp="segm.xml#_1.15.1-seg"/>
  <seg ana="VM0"
    corresp="segm.xml#_1.15.2.2.1-seg"/>
  <seg ana="VVI"
    corresp="segm.xml#_1.15.3-seg"/>
  <seg ana="PNP"
    corresp="segm.xml#_1.15.4-seg"/>
  <seg ana="?"
    corresp="segm.xml#_1.15.5-seg"/>
</s>

```

b. CLAWS c7

```

<s xml:id="morph_1.1-s">
  <seg ana="PPIS1"
    corresp="segm.xml#_1.15.1-seg"/>
  <seg ana="VM"
    corresp="#segm.xml#_1.15.2.2.1-seg"/>
  <seg ana="VVI"
    corresp="#segm.xml#_1.15.3-seg"/>
  <seg ana="PPH01"
    corresp="segm.xml#_1.15.4-seg"/>

```

```

<seg ana="?"
  corresp="#segm.xml_1.15.5-seg"/>
</s>

Error-identification layer
<spanGrp resp="#bansp"
  type="gram" n="art">
  <span from="#segm.xml_1.1.1-seg"
    to="segm.xml#_1.1.1-seg"
    cert="high"
    rend="add">the $1</span>
  <span from="segm.xml#_1.5.7-seg"
    to="segm.xml#_1.5.7-seg"
    cert="high" rend="del"/>
</spanGrp>
<spanGrp resp="#bansp"
  type="gram" n="w/o">
  <span from="segm.xml#_1.15.1-seg"
    to="segm.xml#_1.15.2-seg"
    cert="high"
    rend="change">$2 $1</span>
</spanGrp>

```

Figure 2.21: Example of annotation from the Foreign Language Examination Corpus

A deeper look at the tagged corpora shows a high interest in three types of annotations, starting with error annotation which assists in achieving one of the main corpora purposes, error analysis (Figure 2.22). The second is PoS, which is commonly used in corpora in general. The remarkable development in PoS tagging tool facilitates this type of annotation particularly for the most widely spoken languages. The developers of the corpora used a number of tools to add the PoS annotation to the texts, such as CLAWS (Garside, 1987, 1996; Garside and Smith, 1997; Leech *et al.*, 1994) in the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), or to speech, such as Praat (Boersma and Weenink, 2014) in the ENGLISH Corpus (Hirst and Tortel, 2010; Tortel, 2008; Tortel and Hirst, 2008). The third type of annotation is used to tag the structural features (e.g., titles, sections, headings, paragraphs, questions, examples, etc.). This type of tagging helps researchers for different functions, such as analysing specific parts/styles of the target language. One of the widely used markup languages for annotating the structural features is XML. It was used, for example, in the British Academic Written English Corpus (Heuboeck *et al.*, 2008), the ASU Corpus (Hammarberg, 2010), and the Michigan Corpus of Upper-level Student Papers (O'Donnell and Römer, 2009a, 2009b).

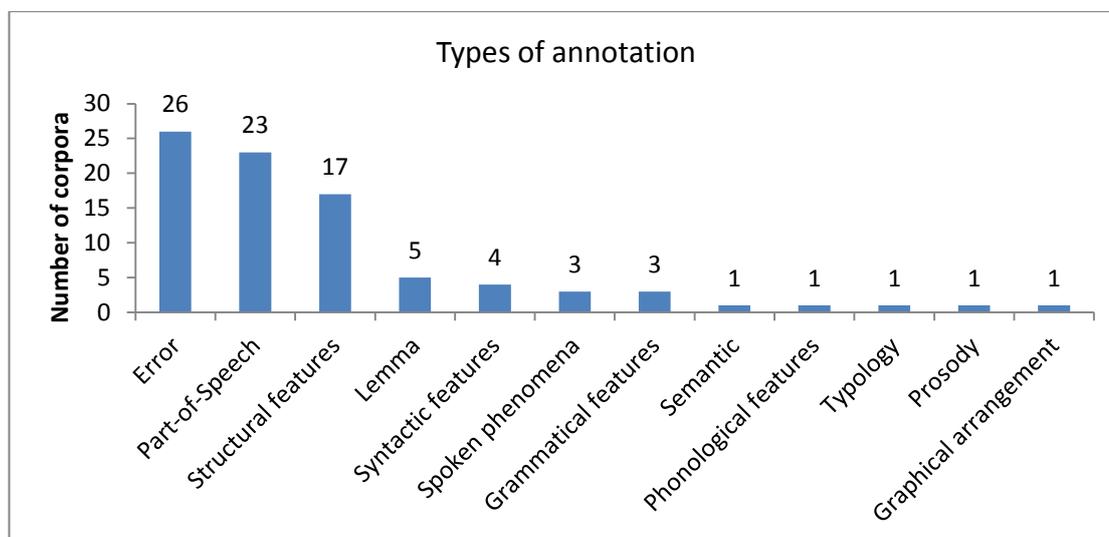


Figure 2.22: Types of annotation used in learner corpora

To sum up, the review covered 11 aspects: corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation. The review provided us with a comprehensive view of the general trends in the domain and helped us to create review-based guidance on design criteria for a new learner corpus which is presented in the following section.

2.3 Recommended Design Criteria to Develop a New Learner Corpus

This section highlights the choices available to learner corpora developers to use in the design criteria of their new corpora. We based our recommendations on the options that received more attention in our review of 159 existing corpora.

2.3.1 Corpus Purpose

Our review of learner corpora literature showed that 76% of learner corpora were created for public purposes while 24% were designed for specific purposes. We recommend to consider a public purpose when developing a new learner corpus, as it (i) serves a large audience, (ii) can be used for various studies, and (iii) may have a longer lifetime of usability. Developing a longitudinal corpus is also worth

considering, particularly when “monitoring learners’ progress” is one of the corpus purposes.

2.3.2 Corpus Size

Corpus size is a controversial issue in corpus development. However, our review revealed that a large number of the learner corpora have a small amount of data (200,000 w/t or less). This size can be utilised as a minimum level when developing a new learner corpus, though a larger corpus size is preferable.

In terms of oral data, our review revealed that 9 out of 16 spoken corpora include 50 hours or less, and 7 of those 9 contain between 3 and 20 hours in length. This finding indicates that up to 20 hours can be considered a starting level, while closer to 50 is a good level to achieve.

2.3.3 Target Language

In general, the language a corpus targets does not rely on what is predominant in the field; rather, the decision is based on the needs of the corpus developers. However, in terms of the number of languages, our findings showed that the standard practice is to develop a corpus with a sole target language; specifically, 90% of the learner corpora examined are monolingual.

Although developing a multilingual corpus with similar materials for each language might take a longer time and present some difficulties, this type is highly useful for some research areas, such as measuring the influence of L1 on the acquisition of target languages. Our findings indicated that it is important for the learners involved in such a project to share the same L1, especially if the corpus is not large enough to represent several L1s alongside the several target languages.

2.3.4 Availability

The number of freely available learner corpora (66 corpora, 61%) is more than double those restricted (29 corpora, 27%). This interest in making the data of learner corpora publicly available is consistent with the tendency to develop corpora for public purposes, as it allows a wider audience of researchers to re-use the data for further research, which serves the target language ultimately.

It is recommended for those corpora that are intended to be freely available and include multimodal data to offer the same free access to the data modes, e.g. hand-written texts and audio and video recording along with their transcriptions. This free access allows users to examine the primary sources instead of relying on the transcriptions, which may be significant due to the different natures of these modes. File formats such as .txt and .xml are recommended for those written learner corpora which tend to be available for download, and .mp3 and .wav for those spoken. Additionally, devoting a single file for each written or spoken text is most common either with or without its metadata and annotation.

With respect to user accessibility to free corpora, the user might be allowed to search the corpus data online with no access to its source files. The corpus in this case needs to be uploaded to one of the corpora search tools existing online. In some cases, one option is to create a search website to suit the properties of the corpus. Another option is to give the user access to the source files of the corpus to be downloaded; this can be under a particular license such as the GNU General Public License¹ or the Creative Commons copyright licenses². Registration might be required for any of these types of access to free corpora.

2.3.5 Learners' Nativeness

The majority of learner corpora (78%) contain data from non-native speakers of the target language, which may be the standard for developing a new learner corpus. However, if one purpose for the corpus is to allow users to conduct comparative analysis between NS and NNS, it is recommended to consider collecting data from NS as well. This approach allows the development of a corpus with similar and comparable materials. Additionally, it is recommended for the NS to be in a language learning context in order to unify the contexts of production of both learners. If this approach is not possible, relying on a general corpus of NS might be the alternative option for such comparative studies.

¹ The General Public License can be accessed from: <https://www.gnu.org/copyleft/gpl.html>

² The Creative Commons copyright licenses can be accessed from: <http://creativecommons.org>

2.3.6 Learners' Proficiency Level

If conducting comparative analysis between learners from different levels is one of the aims of building the corpus, it is recommended to collect data from all levels (e.g. Beginning, Intermediate, and Advanced). Our review revealed that this arrangement is present in 54% of learner corpora, making it arguably a standard practice compared to the other approaches. If analysing the language of beginners might be difficult, then intermediate and advanced levels, or even advanced level only, may be sufficient.

2.3.7 Learners' First Language

The literature review revealed that 56% of learner corpora include data of learners from various L1 backgrounds, whereas learners represent a single mother tongue in each corpus of the other corpora (44%). This relatively even division between the two approaches suggests that selecting either various L1s or a sole L1 in a corpus can be based on whether the designers are interested in conducting comparative analysis between learners from different L1s. The decision to use a single or various L1 backgrounds may be based on whether a corpus is designed with a target language group in mind or if the designers have access to particular language learners.

2.3.8 Material Mode

Written mode exists purely in 66% of learner corpora, while the spoken mode represents 26% and they are combined in 7%. This indicates that compiling a corpus with a single mode is the standard in 92% of learner corpora, and then the corpus aim plays the most significant role in selecting the materials mode, written or spoken. Given that speech is more sensitive to new language changes, as Mauranen (2007) indicates, combining spoken and written materials in a learner corpus could provide valuable opportunities for performing comparative analyses between those two data modes. As another choice, building a multimodal corpus with sound and video recordings, and orthographic transcriptions can be beneficial for depth analysis as McEnery (2003) suggests. A combination of multimodal materials in a learner corpus provides insights into learner needs in different contexts.

2.3.9 Material Genre

Most learner corpora tend to include one or two genres. The literature reviewed revealed that argumentative, narrative, and descriptive materials are the most used. The findings indicated that designing a corpus that focuses on a single genre is preferable, as 48% of learner corpora fall under this type unless there is a need to compare the learners' production of different genres. In terms of which genre to include, we recommend considering the familiarity of the learners in their learning environment; specifically, choosing a genre with which they are familiar may help them to produce more natural data.

2.3.10 Task Type

Essays dominate the task types used to collect written data of learner corpora, followed by interviews which are usually used for spoken corpora. The frequency of using those two types gives developers of learner corpora standard tools for both written and spoken data. Tests and exams are also commonly used and can be a good option for those who want to collect written and spoken data using a single task type.

2.3.11 Data Annotation

Most learner corpora that addressed annotation (82%) are tagged with one or more types of annotation. This practice reflects the importance of annotating data, which adds more value to the corpus data and consequently enables researchers to perform in-depth analysis. Errors, part-of-speech, and structural features are respectively the most popular types of annotation in learner corpora. The corpus aim may help in determining the types of annotation required; however, adding further types of annotation to the corpus will increase the value of corpus data (e.g., lemmas, syntactic and grammatical features, spoken phenomenon, etc.).

2.4 Related Work: Arabic Learner Corpora

The field of learner corpora is about 25 years old, with Arabic learner corpora emerge up more recently. This section presents a review of the small number of existing Arabic learner corpora. It is followed by a comparison between them and

the ALC in order to highlight the ALC's contributions. The comparison is based on the 11 design criteria discussed in the previous section.

2.4.1 Pilot Arabic Learner Corpus (Abuhakema *et al.*, 2009)

In developing the Pilot Arabic Learner Corpus, Abuhakema *et al.* (2009) aimed to collect a small learner corpus of Arabic, to develop a tagset for error annotation of Arabic learner data, to tag the data for errors, and to perform simple computer-aided error analysis. According to Abuhakema *et al.* (2009), the Pilot Arabic Learner Corpus includes about 9000 words of written Arabic materials produced by American native speakers of English who learn Arabic as a Foreign Language. Two levels were included, Intermediate (3818 tokens) and Advanced (4741 tokens). Abuhakema *et al.* (2009) used the guidelines of the American Council on the Teaching of Foreign Languages (ACFTL, 2012) to classify written texts into the Intermediate and Advanced levels. The texts of some of the learners were written while the learners were studying Arabic in the United States, while others were produced when the learners went to study abroad in Arab countries.

Abuhakema *et al.* (2009) stated that the data was available online¹, but at the time of writing it was not possible to access the website, suggesting a broken or out-of-date link. The errors of learners were tagged using a tagset for error annotation developed by adopting the French Interlanguage Database tagset (Granger, 2003a). It was not clear from the paper whether the error tagging was conducted manually, automatically, or semi-automatically (computer-assisted error annotation). However, Abuhakema *et al.* (2008) described a plan to include a pull-down menu of tags at each level to speed the annotation. This note indicates a semi-automatic process to mark up the errors of the learners. Further, they discussed a plan to reconstruct the texts by correcting all the mistakes and tagging the corpus for parts of speech, which will enable researchers to perform further morphological and syntactic analyses.

2.4.2 Malaysian Corpus of Arabic Learners (Hassan and Daud, 2011)

Hassan and Daud (2011) designed the Malaysian Corpus of Arabic Learners primarily to give accurate descriptions of Arabic conjunctions used among learners

¹ From: <http://chss.montclair.edu/~feldmana/publications/flairs21-data/>

of Arabic, to investigate the misuse of Arabic conjunctions among learners, and to see how certain combinations of words were preferred by learners. The corpus includes approximately 87,500 words, produced by Malaysian advanced learners of Arabic during the first and second years of their Arabic major degree programme, Department of Arabic Language and Literature at International Islamic University Malaysia. The corpus materials include around 250 descriptive and comparative essays produced on computers using Microsoft Word without any help from native speakers. The corpus is not accessible online, but there is a plan to upload the entire corpus into the Arabic Concordancer, which can be accessed online¹ (Haslina, personal communication, 15 September 2014; Hassan and Ghalib, 2013). The corpus consists of raw data without any type of annotation.

2.4.3 Arabic Learners Written Corpus (Farwaneh and Tamimi, 2012)

Farwaneh and Tamimi (2012) designed the Arabic Learners Written Corpus to serve as a source of empirical data for hypothesis testing, as well as a resource for developing materials for teaching Arabic. Materials used by the Arabic Learners Written Corpus were produced by non-native Arabic speakers from the United States and were collected over a period of 15 years. This corpus includes around 35,000 words covering three levels (Beginner, Intermediate, And Advanced), and three text genres (Descriptive, Narrative, and Instructional). It was developed over two phases. The aim of the first phase was to offer a source of raw data, and the aim of the second phase was for the corpus to be tagged. The raw data of the Arabic Learners Written Corpus is available for download in PDF files². The future work includes annotating the corpus for the errors and features of each level.

2.4.4 Learner Corpus of Arabic Spelling Correction (Alkanhal *et al.*, 2012)

Alkanhal *et al.* (2012) stated that the aim of compiling the Learner Corpus of Arabic Spelling Correction was to build and test a system developed to automatically correct misspelled words in Arabic texts. The corpus consists of 65,000 words that

¹ The Arabic Concordancer is accessed from: <http://efolio.iium.edu.my/arabicconcordancer>

² The files can be downloaded from: <http://l2arabiccorpus.cercll.arizona.edu/?q=allFiles>

were manually revised for spelling to annotate all misspelled words. This data covers diverse essays written by students studying at two universities.

“These essays were handwritten, and were manually converted to an electronic copy by data entry persons. The test data has two sources of errors; the actual misspelled words by the students and the generated mistakes during the data entry process” (Alkanhal *et al.*, 2012: 2118).

The corpus available for download contains two versions¹. The first, which is in plain text files, is not tagged. The second, in which errors are manually corrected, is available as a Microsoft Access database in MDB file format.

2.5 Rationale for Developing the Arabic Learner Corpus

The examination of the Arabic learner corpora details reveals that their sizes are small in comparison to those of some other widely spoken languages, such as English or French. The Pilot Arabic Learner Corpus, for example, covers 9000 words, and the other corpora are less than 100,000 words. Although size is a controversial issue in corpus development, the corpus size plays a significant role in terms of representativeness. In addition, size is important in some cases, for example when generalising the results of a corpus-based study on the population of language learners.

Availability is another important point, as two of the Arabic learner corpora are not available for search or download; additionally, the Arabic Learners Written Corpus is available only in PDF format, whereas the plain text format (TXT) is preferable for corpus data more than binary encoding formats such as PDF (Wynne, 2005). Only the Learner Corpus of Arabic Spelling Correction provides its data in plain text and a database. However, as its purpose is for Arabic NLP, the data covers only native speakers of Arabic, which may not be appropriate data to use when researching Arabic learning and teaching as a second language.

¹ The files can be downloaded from: http://cri.kacst.edu.sa/Resources/TST_DB.rar

The third point is the materials mode, as the existing Arabic learner corpora cover only written materials, with no spoken data counterpart. A number of researchers (e.g. Leech, 1997; Kennedy, 1998) note the significance of including spoken language even in a small percentage of the corpus because spoken language represents the most common mode of language.

These points highlight the need for creating an Arabic learner corpus that takes research needs into consideration during its design. Table 2.7 presents a summary for the existing Arabic learner corpora based on the 11 design criteria discussed in the literature review. The next section will highlight the contributions of the ALC in comparison to the reviewed Arabic learner corpora.

Table 2.7: Summary of the existing Arabic learner corpora

Design criterion	Pilot Arabic Learner Corpus	Malaysian Corpus of Arabic Learners	Arabic Learners Written Corpus	Learner Corpus of Arabic Spelling Correction
Purpose	Computer-aided Error Analysis	Interlanguage analysis	Arabic language teaching	To develop a spell-checker system for Arabic language
Size	9000 words	87,500 words	approximately 35,000 w/t	65,000 words
Target language	Arabic	Arabic	Arabic	Arabic
Availability	Not available, the link is out of date or broken	Not available, but intended to be searchable online	Available to download in PDF file format	Available to download
Learners' nativeness	Non-native speakers of Arabic	Non-native speakers of Arabic	Non-native speakers of Arabic	Native speakers of Arabic
Learners' proficiency level	Intermediate and advanced	Advanced	Beginner, intermediate, and advanced	N/A
Learners' first language	English	Malaysian	Not specified	Arabic
Material mode	Written	Written	Written	Written
Material genre	Not specified	descriptive and comparative	Descriptive, narrative, and instructional	Various

Task type	Essay	Essay	Essay	Essay
Data annotation	Tagged for errors	Not tagged	Not tagged	Errors are manually corrected

2.6 The ALC's Contribution Compared to the Existing Arabic Learner Corpora

The ALC's contribution compared to the existing Arabic learner corpora can be highlighted through the following points:

Purpose: The purposes of these corpora show that they are designed for public use, either Arabic language teaching or Arabic NLP. The ALC is to be used for both purposes: Arabic language teaching and Arabic NLP.

Size: The sizes of the Arabic learner corpora are relatively small, ranging between 9000 and 87,500 words. The ALC is designed to include at least 200,000 words. The current version (v2) includes 282,732 words (386,583 tokens/lexical items and 29,625 types).

Target language: Arabic is the target language in the data of the existing Arabic learner corpora, which is the case of the ALC as well.

Availability: Two of the existing Arabic learner corpora are available for download, one in PDF format and the other in plain text files and as a Microsoft Access database. The ALC data is available in four formats (PDF, MP3, TXT, and XML) based on the nature of the data. Specifically, users can download a PDF for the hand-written texts, an MP3 for the audio recordings, and plain text and XML for the electronic texts and transcriptions of the hand-written texts and audio recordings.

Learners' nativeness: Three corpora, which were developed for Arabic language teaching, include data produced by non-native speakers of Arabic, while the corpus that was designed for Arabic NLP purposes includes data by native speakers of Arabic. The ALC is designed to include a balance between the data of native and non-native speakers of Arabic. Speakers of both types are learning or specialising in the Arabic language.

Learners' proficiency level: The corpora differ in this criterion. Specifically, the Arabic Learners Written Corpus covers three levels (Beginner, Intermediate, and

Advanced), the Pilot Arabic Learner Corpus covers two levels (Intermediate and Advanced), and the Malaysian Corpus of Arabic Learners covers only Advanced learners. The proficiency level criterion is not applicable to the Learner Corpus of Arabic Spelling Correction, as its data is produced by native speakers of Arabic. The ALC is developed to cover two levels of Arabic learners in the current version: Intermediate and Advanced. In future versions, data from the Beginner level will be included as well.

Learners' first language: Each existing Arabic corpus includes learners from one first language, e.g. English in the Pilot Arabic Learner Corpus, Malaysian in the Malaysian Corpus of Arabic Learners, and Arabic in the Learner Corpus of Arabic Spelling Correction. It seems that the Arabic Learners Written Corpus includes learners from various first languages. The ALC is designed to include learners from various first languages. The current version includes writings from learners with 66 different mother tongues, which allows users to conduct comparative studies on those groups.

Material mode: Each existing Arabic learner corpus covers only written data, while the ALC is developed to include two materials modes: written and spoken.

Material genre: Existing Arabic learner corpora include different materials genres such as Descriptive, Comparative, Narrative, and Instructional. The ALC will focus on two genres which are commonly used in learner corpora: Narrative and Discussion.

Task type: As all Arabic learner corpora include written data, the essay is used to collect their data. The ALC is designed to use the essay for written data and the interview for spoken data, which are the most commonly used task types in learner corpora.

Data annotation: The Pilot Arabic Learner Corpus is tagged for errors but is not available, while errors in the Learner Corpus of Arabic Spelling Correction are corrected without tagging them for the error type. Data of the other corpora is not tagged. The ALC is designed to include error tags using a novel error tagset created for the ALC. The error tagging will also include suggested corrections for those errors in order to reconstruct the corpus data.

The design of the ALC with its contents are discussed in detail in the next chapter.

2.7 Conclusion

This chapter presented a review of 159 learner corpora to derive design criteria for developing new learner corpora or expanding corpora already in existence. A number of previous studies and surveys have investigated this field; however, this review was intended to include all current corpora in order to provide a quantitative view of the domain. We investigated the corpora in 11 categories: corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation.

This analysis revealed several trends in existing learner corpora. For instance, a third of learner corpora were developed to be used for language learning and teaching. The investigated corpora target 20 languages, and English is included in more than 90 of them. Fifty-six percent of language corpora include data of learners from various L1s. For those that focus on a single L1, Chinese speaking learners receive the highest attention. In terms of materials, most learner corpora tend to include one or two genres. Argumentative, narrative, and descriptive prose are the most-used genres. More than half of learner corpora include a sole task type; specifically, essays are preferred for written tasks and interviews for spoken. The findings illustrate that 82% of the learner corpora that addressed annotation are tagged with one or more types of annotation, and error tagging is the most popular.

Following the review, we offered recommended guidelines for creating a new learner corpus based on the analysis of the learner corpora field. These guidelines were the basis of building the ALC, and also can be utilised to improve and/or expand the current corpora or even when undertaking a study in this field.

Additionally, the chapter presented a review of related work in the form of the existing Arabic learner corpora. We discussed the rationale of creating the ALC, followed by a comparison between the existing Arabic learner corpora and the current project, the Arabic Learner Corpus, in order to highlight the contribution of the latter. Our comparison was based on the 11 design criteria discussed in the literature review.

The ALC was developed based on the guidelines we derived from reviewing the literature in this chapter. The existing Arabic learner corpora were also considered in order to justify the creation of the ALC. The following part of the thesis (Part II)

describes the design and content of the ALC in Chapter 3, and the methodology of data collection and management in Chapter 4.

Part II

Arabic Learner Corpus

Summary of Part II

This part discusses in Chapter 3 the design criteria and content of the ALC followed by the design and content of the ALC metadata elements. It also presents an overview of projects that have used the ALC. Chapter 4 describes the methodology for collecting and managing the ALC data. The description covers the questionnaire and guidelines for data collection, the standards for converting the hand-written texts and spoken materials into an electronic form, the method followed to measure the consistency between transcribers, the ALC database, the function of files generation, and the method for naming the ALC files.

3 ALC Design and Content

Chapter Summary

This chapter describes the 11 design criteria on which the ALC was developed. For each criterion, the description starts by referring to the relevant literature review, and then discussing the targeted ALC design and the content that was achieved. In addition to the design criteria, the ALC was developed with 26 variables of metadata. The chapter describes those metadata elements in terms of the target design and the content achieved for each element. The last section of this chapter highlights the increasing interest in using the ALC data by discussing the projects that have used the corpus, the comments that have been received from a number of specialists, and the downloads from the ALC website.

3.1 Introduction

It is believed that a smaller homogeneous corpus that features a high quality design is far more valuable than a larger corpus (Granger, 1993). Therefore, specific design criteria had to be defined for the ALC based on the recommended guidelines described in Section 2.3. In addition, the design of the ALC includes 26 variables as metadata elements, 12 for the learner and 14 for the text. The following sections describe the design and content of the ALC and its metadata.

3.2 ALC: Design Criteria and Content

The ALC data was collected during two stages: pilot (version 1 [v1]) and main (version 2 [v2]). The content of the second version absorbed v1. The design criteria of the corpus were defined to be achieved at the end of the second stage. This section will discuss the 11 design criteria: the corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation. Each of those criteria will be linked to the previous work discussed in the literature review, and the target design and achieved content will be described.

3.2.1 Purpose

The purposes of learner corpora were classified in the literature review under two main categories: public and specific purposes. The majority of corpora (81 out of 107) have public purposes, which suggests a high interest in developing learner corpora that serve a large audience and can be used for various purposes. Thus, the ALC follows the general trend and is meant for public use; specifically, it falls into the category of those corpora intended to be used under broad aspects of research or by a wide audience of users. The main goal of the ALC is to create a dataset to serve as a resource for research in Arabic NLP and Arabic language teaching. From its first version, the ALC has achieved this goal, as researchers have used it for both Arabic NLP (e.g. error detection and correction tools, evaluating the existing Arabic analysers, and native language identification systems) and Arabic language teaching (e.g. applied linguistics studies and data-driven Arabic learning activities). Examples of the works that have used the corpus are summarised in the corpus evaluation Section 3.4 and described in detail in Chapter 7.

3.2.2 Size

Learner corpora projects typically comprise less than one million w/t with the majority centring on the size of 200,000 w/t or less, as seen in the literature review. Additionally, Granger (2003a) argues that “[a] corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples but minute in the corpus linguistics field at large, where recourse to mega-corpora of several hundred million words has become the norm rather than the exception” (p 465). With respect to the ALC as a PhD project, the intended size at this stage (v2) was 200,000 words.

The ALC data was collected and entered into a database in which the corpus size was counted automatically by a short programming code the researcher added. The code calculated words on the basis that any set of characters between spaces was considered one word. Spaces in this sense included normal spaces, tabulator spaces, or new-line breaks. Based on this definition, the total amount of words the corpus includes is 282,732 in v2 (31,272 words in v1). After separating off all clitics – including clitic pronouns, prepositions, and conjunctions – using the Stanford Word Segmenter (Monroe *et al.*, 2014), the corpus data consists of 386,583 tokens (lexical items) and 29,625 types¹. The final number of words exceeded the target because only 17% of the corpus data was collected in an electronic format, while 83% had to be entered into the computer after the collection process (76% hand-written texts and 7% spoken data). The researcher had three months to collect the data of the second version of the ALC in Saudi Arabia, but this period did not include entering the data into the computer. As a result, the researcher did not know what the final size would be. This uncertainty in the total size led the researcher to collect more data to ensure that the target size was reached.

3.2.3 Target Language

Although bilingual and multilingual corpora can be used for comparative studies, the literature review showed that 90% of learner corpora are monolingual. The current corpus project was designed to be monolingual following the norm in the learner corpora domain. In terms of the target language, this element usually does not rely on what is predominant in the field; instead, the language is determined by the needs of the corpus developers. There were two essential reasons behind choosing Arabic

¹ A token is “an occurrence in text of a word from a language vocabulary”, while a type is “a word in a language vocabulary, as opposed to its specific occurrence in text” (Mitkov, 2003).

as a target language for the learner corpus. Firstly, the researcher teaches Arabic and works in the field of Arabic computational linguistics. The second reason is due to the absence of such a project; that is, no such compilation of an Arabic learner corpus exists with the specified design criteria.

The researcher's experience of teaching Arabic has shown that the field of teaching the Arabic language in Saudi Arabia is dominated by Modern Standard Arabic (MSA). However, this form is sometimes combined with other forms (classical Arabic or colloquial Arabic) in a small percentage. Thus, the class of the Arabic language targeted to be included in the ALC is the same as that which is taught to the corpus contributors with no concentration on a particular form. As for the context of learning Arabic, native Arabic-speaking students (NS) are learning Arabic as a part of their curriculum to improve their written Arabic. Non-native Arabic-speaking learners (NNS) are learning Arabic as a second language in order to continue their studies at Saudi universities. The corpus includes contributions from both of these groups of learners.

3.2.4 Data Availability

The review of learner corpora literature showed that those corpora publicly available online for search or download represent the highest percentage among the other types (61%). Additionally, this type is more than twice as common as those that have restricted or paid access (27%). Given that the ALC is intended to be an open-source of data for research on the Arabic language, the most appropriate choice was to make the ALC data freely available for download under the *Creative Commons Attribution-NonCommercial 4.0 International License*¹ and in a number of file formats (TXT, XML, PDF, and MP3). In addition, it is also available for online search using some tools that have different features. Such diversity in the corpus availability may serve a wider audience of users. Details about the choices to provide the information for download and for online search are provided in the following two sections.

¹ A summary of the license can be accessed from: <http://creativecommons.org/licenses/by-nc/4.0/legalcode>

3.2.4.1 For Download

Four file formats are available to the ALC users¹: plain text (TXT), Extensible Markup Language (XML), Portable Document Format (PDF), and MPEG-2 Audio Layer III (MP3). This section gives more details about the corpus files in these formats.

1. TXT format contains plain text without formatting such as font type, size, or colour. This format is preferable for corpus data more than binary encoding formats such as PDF, RTF, and Word, especially with generic tools (Wynne, 2005). Such files can be read and edited with any text editor, such as Notepad on Windows. Additionally, Arabic text in a plain text format is readable by most corpora analysis tools, such as Khawas (Althubaity *et al.*, 2013, 2014), aConCorde (Roberts, 2014; Roberts *et al.*, 2006), AntConc (Anthony, 2005, 2014a, 2014b), WordSmith Tools (Scott, 2008, 2012), and Sketch Engine (Kilgarriff, 2014; Kilgarriff *et al.*, 2004). ALC data is available in the plain text format encoded in UTF-16 with three choices: (i) plain text with no metadata (only the text with its title), (ii) plain text with Arabic metadata, or (iii) plain text with English metadata; see examples of these file formats in Appendix A.1. The metadata includes information about the author (e.g. age, gender, nationality, mother tongue, level of study, etc.) and about the text (e.g. genre, text mode: written or spoken, length, place of writing, etc.). Adding this type of information to the files enables researchers to identify characteristics of the text and its producer, which adds more depth to the data analysis.
2. The second option is to download the ALC files in XML, which was selected because XML is becoming the standard for representing annotation data (Pustejovsky & Stubbs, 2013). It defines a set of rules for encoding documents in a format that is both human-readable and machine-readable². Some corpus tools use this format to give the user more choices while still allowing the data to be searched efficiently. The XML files of the ALC were validated against Document Type Definition (DTD), which is described in the annotation standards section (5.3.1).

¹ These formats can be downloaded from: <http://www.arabiclearnercorpus.com>, <http://www.alcsearch.com>, or from the Linguistic Data Consortium (LDC): <https://catalog.ldc.upenn.edu/LDC2015S10> or <http://www.islm.org/resources/568-308-670-444-7/>.

² Wikipedia definition, <http://en.wikipedia.org/wiki/XML>

“A DTD is a set of declarations containing the basic building blocks that allow an XML document to be validated [...] The DTD defines what the structure of an XML document will be by defining what tags will be used inside the document and what attributes those tags will have. By having a DTD, the XML in a file can be validated to ensure that the formatting is correct” (Pustejovsky & Stubbs, 2013: 68).

The DTD was automatically added to the beginning of each XML file as a part of automating the corpus file generation process. The ALC offers two choices for XML files encoded in UTF-16: (i) XML with Arabic metadata and (ii) XML with English metadata; see examples of these file formats in Appendix A.2.

<code><doc ID="S004_T2_M_Pre_NNAS_W_C"></code>	Beginning of the document with its ID
<code><header></code>	Beginning of the header
<code><learner_profile></code>	Beginning of the learner information
<code><age>24</age></code>	Age
<code><gender>Male</gender></code>	Gender
<code><nationality>Ugandan</nationality></code>	Nationality
<code><mothertongue>Ugandan</mothertongue></code>	Mother tongue
<code><nativeness>NNAS</nativeness></code>	Nativeness
<code><No_languages_spoken>4</No_languages_spoken></code>	Number of languages spoken
<code><No_years_learning_Arabic>14</No_years_learning_Arabic></code>	Number of years learning Arabic
<code><No_years_Arabic_countries>2</No_years_Arabic_countries></code>	Number of years spent in Arabic countries
<code><general_level>Pre-university</general_level></code>	General level of education
<code><level_study>Diploma course</level_study></code>	Level of study
<code><year_or_semester>Second semester</year_or_semester></code>	Year/Semester
<code><educational_institution>Arabic Inst. at Imam Uni</educational_institution></code>	Educational institution
<code></learner_profile></code>	End of the learner information
<code><text_profile></code>	Beginning of the text information
<code><genre>Discussion</genre></code>	Text genre
<code><where>In class</where></code>	Where produced
<code><year>2012</year></code>	Year of production
<code><country>Saudi Arabia</country></code>	Country of production
<code><city>Riyadh</city></code>	City of production
<code><timed>Yes</timed></code>	Timed or not timed task
<code><ref_used>No</ref_used></code>	References use
<code><grammar_ref_used>No</grammar_ref_used></code>	Grammar book use
<code><mono_dic_used>No</mono_dic_used></code>	Monolingual dictionary use
<code><bi_dic_used>No</bi_dic_used></code>	Bilingual dictionary use
<code><other_ref_sed>No</other_ref_sed></code>	Other references use
<code><mode>Written</mode></code>	Text mode
<code><medium>Written by hand</medium></code>	Text medium
<code><length>100</length></code>	Text length
<code></text_profile></code>	End of the text information
<code></header></code>	End of the header
<code><text></code>	Beginning of the text part
<code><title>تخصصي العلمي مستقبلاً</title></code>	The text title
<code><text_body></code>	The text body
مسيرتي في حياتي وخاصة من جهة الأكاديمية صعبة التصور لدى البعض لما فيها من أفكار للتتبع في التخصصات فأتمنى أن أكون عالماً ملماً في الشريعة وكذلك طبيباً مسلماً ملماً في الطب، ذلك لأن بلدي لم تفتح انتقلاً تاماً لعلماء الشريعة الإسلامية لذلك ينبغي لي أن أتناول التخصص المقبول من قبل الدولة مثل الطب حتى ما إن دخلت الحكومة كالتبيب الماهر طبقت الشريعة التي درست لنصرة دين الله تعالى.	
هذا باختصار غيائي، فيبقى عند الأسئلة المهمة: كيف أنال هذا وذاك؟ وبما أبدا؟ ومن أين أدرس هذا وذاك؟ فهذه وغيرها من الأسئلة كثيرة، فأسأل الله أن يبارك في عملي ويرزقني الإخلاص فيه.	
<code></text_body></code>	
<code></text></code>	End of the text part
<code></doc></code>	End of the Document

Figure 3.1: Illustration of the XML structure

3. PDF is “a file format for representing documents in a manner independent of the application software, hardware, and operating system used to create them and of the output device on which they are to be displayed or printed” (Adobe Systems Incorporated, 2006: 33). It was used in the ALC for the hand-written texts after they had been scanned. PDF was used rather than an image format, as a text written on more than one page can be presented in a single multi-page PDF document.

4. The MP3 format was established by the Moving Picture Experts Group (MPEG; Fraunhofer Institute for Integrated Circuits IIS, 2015). MP3 is an audio-coding format for digital audio. It uses a form of lossy data compression technologies that make it possible to create smaller files (Thompson, 2005). Due to the small size of MP3 files and their quality, this format is commonly used in spoken corpora such as the French Learner Language Oral Corpora (Myles & Mitchell, 2012), the Spanish Learner Language Oral Corpus (Mitchell *et al.*, 2008), The PAROLE corpus (Hilton, 2008), and the Spanish Learner Oral Corpus (Maolalaigh & Carty, 2014b). Thus, it was used in the ALC for the learners' audio recordings. Only audio files of those learners who granted permission to publish their recordings are available, and the total length of these recordings is 3 hours, 22 minutes, and 59 seconds.

Table 3.1 shows a summary of the four file formats available for download.

Table 3.1: Summary of ALC files available for download

Format		
TXT (encoded in UTF-16)	Data type included	- Electronic written texts (17% of ALC) - Transcription of hand-written texts (76% of ALC) - Transcription of audio recordings (7% of ALC)
	Options available	1. Plain text with no metadata (1585 files) 2. Plain text with Arabic metadata (1585 files) 3. Plain text with English metadata (1585 files)
XML (encoded in UTF-16)	Data type included	- Electronic written texts (17% of ALC) - Transcription of hand-written texts (76% of ALC) - Transcription of audio recordings (7% of ALC)
	Options available	1. XML with Arabic metadata (1585 files) 2. XML with English metadata (1585 files)
PDF	Data type included	Hand-written sheets (76% of ALC)
	Options available	Scanned sheets in PDF files (1257 files)
MP3	Data type included	Audio recordings (7% of ALC)
	Options available	MP3 files (52 files = 3 hours, 22 minutes, and 59 seconds)

3.2.4.2 For Online Search

The ALC is available for online search via three tools: ALCsearch, Sketch Engine, and arabiCorpus. ALCsearch uses the ALC metadata as determinants to search any subset of the data. Sketch Engine has advanced functions for analysing corpora, but it requires paid access; for this reason, arabiCorpus was selected as a free-access choice with less sophisticated functions. The following points offer more information about these tools.

1. The ALCsearch¹ is a free-access, web-based tool developed specifically for the ALC. It provides a basic concordancing function which enables users to search the entire corpus or any subset of the corpus data by using the ALC metadata as determinants. For instance, the user can search the sub-corpus of spoken data by selecting the option “Spoken” from the determinant “Text Mode”. Chapter 6 provides details about this tool.
2. The Sketch Engine² (Kilgarriff, 2014; Kilgarriff *et al.*, 2004) is a commercial web-based tool for corpus analysis. Along with the general features of Sketch Engine (e.g. concordance, word lists, key words, collocation, and corpus comparison), it has some unique features; for example, the Word Sketches feature provides summaries of a word’s grammatical and collocational behaviour, while Word Sketch Difference compares and contrasts words visually. Adding the ALC data to Sketch Engine enables users to utilise the advanced functions of this tool in searching the ALC. The ALC version on Sketch Engine is tokenised and tagged for PoS using the Stanford Arabic Parser (Green & Manning, 2010).
3. The free-access, web-based tool arabiCorpus³ (Parkinson, 2015) “provides a fairly effective search mechanism in which the user specifies whether the search term is a noun, adjective, adverb, or verb. The search term is then expanded morphologically according to its inflectional category, and all appropriate prefixes and suffixes are added. Results (hits) are displayed in concordance format, and statistics are provided on the search term’s collocates and its distribution over various corpora” (Buckwalter & Parkinson, 2013). The ALC data was added to arabiCorpus in order to allow users to utilise its free access and search functions.

¹ ALCsearch can be accessed from: <http://www.alcsearch.com>

² The Sketch Engine tool can be accessed from: <http://www.sketchengine.co.uk>

³ The arabiCorpus tool can be accessed from: <http://arabicorpus.byu.edu>

3.2.5 Learners' Nativeness

Reviewing the literature revealed that most learner corpora contain data from NNS of the target language. However, about 20% have data from both NS and NNS which is mostly for comparative purposes. As previously described, the ALC is intended for public purposes. Enabling users to conduct comparative studies may serve this purpose. Therefore, the ALC was designed to include data from both NS and NNS.

One of the best practices in learner corpora covering NS and NNS is to have a balance between the productions of these two groups (see for example Hammarberg, 2010; Heuboeck *et al.*, 2008; O'Donnell & Römer, 2009a, 2009b). Thus, the ALC was designed to have 50% of the corpus data for each group (NS: 100,000 words and NNS: 100,000 words). The actual data collected from both groups was at the target percentages projected in v1 (NS = 50%, 15,741 vs. NNS = 50%, 15,531), and close to the target in v2 (Figure 3.2). The number of words included in v2 is greater than the target established in the design criteria for the reason explained in the corpus size section (3.2.2).

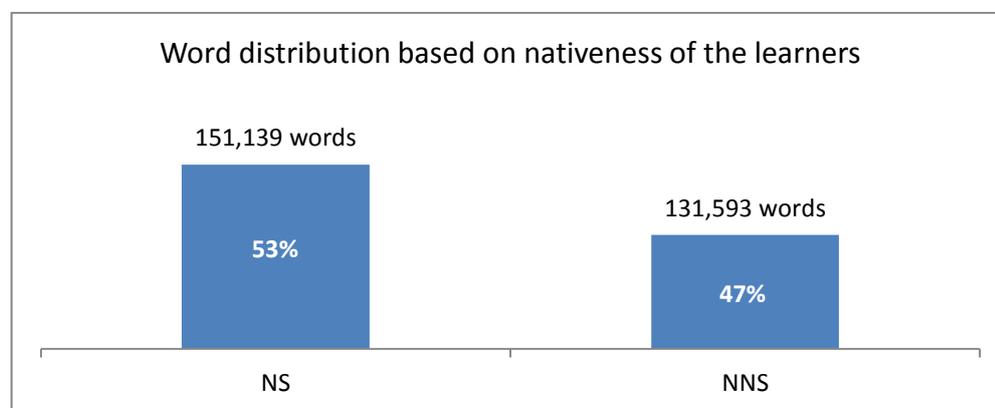


Figure 3.2: Word distribution based on nativeness of the learners

3.2.6 Learners' Proficiency Level

The literature review revealed a relative balance among the Beginning, Intermediate, and Advanced levels included in learner corpora (Beginning 28%, Intermediate 35%, and Advanced 37%). However, due to the limited time devoted to data collection, the researcher decided to include only advanced and intermediate levels in the current version modelling after one of the standard corpora, the International

Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010). The low language proficiency of beginning learners may require further care and time to collect data, as the researcher needs to ensure that tasks are well-explained and understood. There is a plan to include data from beginners in a future version of the ALC. It is important to highlight that this criterion applies only to the NNS learners, as native speakers cannot be classified on the basis of proficiency level of their mother language, Arabic. In addition, since the non-native learners are divided into levels of study that represent their levels of proficiency as determined by the institutions, this classification was used as a proficiency level indicator in the ALC. In the first version of the ALC, 23.14% of the total size was from NNS at the intermediate level, and 26.53% from the advanced. The second version of the ALC contains 28.13% from NNS at the intermediate level and 18.48% from the advanced.

3.2.7 Learners' First Language

An examination of the literature revealed that 56% of existing learner corpora include data from learners with various mother tongue backgrounds. Additionally, institutions teaching Arabic as a second language in Saudi Arabia have no focus on learners speaking a specific first language. One institution teaches Arabic to learners from 43 different mother tongue backgrounds (Alfaifi, 2011). Thus, the best choice was to have data from learners who spoke various first languages. The first version of the ALC covered 26 different mother tongue representations. The second version contains 66 L1s; 65 of them are spoken by the NNS learners while the Arabic language is the L1 of all the NS learners.

3.2.8 Material Mode

Although researchers have noted the importance of having balanced data in terms of their mode (Kennedy, 1998; Leech, 1997), reviewing the existing learner corpora showed that 66% include written data, 26% contain spoken data, 7% contain both modes, and 1% contain a multimodal corpus with written, spoken, and video data. Considering both the difficulties and benefits of building spoken corpora (Branbrook, 1996; Kennedy, 1998; Leech, 1997; McEnery, 2003; Thompson, 2005), the ALC was designed to contain 180,000 words (90%) of written data and 20,000 words (10%) of spoken language. The first version of the ALC included only written data (31,272 word). The second version, which also contains the content of v1,

consists of 263,045 words (93%) of written data and 19,687 words (7%) of speech data equalling more than three hours of audio along with transcriptions.

3.2.9 Material Genre

With respect to materials genre in learner corpora, reviewing the literature revealed that (i) argumentative, narrative, and descriptive materials were the most used respectively followed by discussion, and that (ii) learner corpora tend to include one or two genres. As the ALC includes various participants in terms of age, first language, nationality, nativeness, proficiency, and educational level, two genres were chosen, narrative and discussion (50% for each), in order to give the learners a variety of options that are likely to suit their preferences. From the researcher's Arabic teaching experience to both L1 and L2 Arabic speakers, the argumentative genre is not as common as discussion in teaching Arabic writing, so the latter was used instead. The narrative genre covered 66% in the first version and forms 67% of the v2 ALC content, while discussion was 34% in v1 and makes up 33% in v2. It seems that the learners enjoy writing in the narrative genre, as their production size was twice that of the discussion genre in both versions of the ALC.

3.2.10 Task Type

Reviewing the learner corpora literature showed that the essay was the most preferable task type in written tasks and interviews in those spoken. In addition, the literature review revealed that more than half of learner corpora used a single task type to collect their data, while 20% used two types and 13% used three types. The ALC uses two task types: essay for writing and interview for speaking. The tools used to collect the data will be discussed in Chapter 4 with more details about those task types. In the ALC data, the tasks followed the materials mode, so v1 of the ALC included only essays since it covered only written data; in contrast, 93% of the v2 content is written essays and 7% consists of spoken interviews.

3.2.11 Data Annotation

As seen in the literature review, 82% of the learner corpora are tagged with one or more types of annotation. Errors, PoS, and structural features are respectively the most popular types of annotation in learner corpora. The lack of an error tagset

appropriate for annotating Arabic errors led to the development of a new one to be used for the ALC and for any Arabic learner corpora. The entire ALC is targeted to be annotated for errors and PoS as well as marked up for structural features (titles and paragraphs).

Due to the time that was needed to develop the Error Tagset of Arabic (described in detail in Chapter 5), a sample of 10,000 words (3.5%) was annotated for errors in the second version of the ALC to illustrate the error annotation method. The current version (v2) of the ALC was entirely tagged for PoS using the Stanford Arabic Parser (Green & Manning, 2010). Another copy was also tagged for PoS but using the MADAMIRA tool (Pasha *et al.*, 2014). Both tools, the Stanford Arabic Parser and MADAMIRA, are among those commonly used for Arabic PoS tagging. In terms of structural features, the ALC database was programmed to mark them up automatically; consequently, the whole corpus was fully marked up for these features.

3.2.12 Summary of the ALC Design

Table 3.2 summarises the ALC design criteria including (where applicable) the target and the content of the current version (v2) of the ALC data.

Table 3.2: Summary of the design criteria used in the ALC

	Design criteria	Target and current content
1	Purpose	Public purpose: to create a data source to serve as a resource for research in Arabic NLP and Arabic language teaching
2	Size	Target: 200,000 words Current: 282,732 words
3	Target Language	Arabic
4	Data Availability	The ALC is designed to be freely available: 1. For download: in a number of file formats (TXT, XML, PDF, and MP3) 2. For online search: on some different tools

5	Learners' Nativeness	<p>Target: NS 100,000 (50%) and NNS 100,000 words (50%)</p> <p>Current: NS 151,139 (53%) and NNS 131,593 words (47%)</p>
6	Learners' Proficiency Level	<p>Target: to collect 25% of the total corpus from the intermediate level and 25% from the advanced level of NNS</p> <p>Current: 28.13% from the intermediate level and 18.48% from the advanced level (of NNS)</p>
7	Learners' First Language	<p>Target: to have data from learners who spoke various first languages</p> <p>Current: 66 different mother tongue representations</p>
8	Materials Mode	<p>Target: written 180,000 words (90%) and spoken 20,000 words (10%)</p> <p>Current: written 263,045 words (93%) and spoken 19,687 words (7%)</p>
9	Materials Genre	<p>Target: narrative 50% and discussion 50%</p> <p>Current: narrative 67% and discussion 33%</p>
10	Task Type	<p>Target: essay 90% and interview 10%</p> <p>Current: essay 93% and interview 7%</p>
11	Annotation	<p>Target: the entire corpus to be annotated for errors, tagged for PoS, and marked up for structural features</p> <p>Current: 10,000 words (3.5%) are annotated for errors, and 282,732 words (100%) are tagged for PoS and marked up for structural features</p>

3.3 ALC Metadata: Design and Content

Burnard (2005) defines metadata as “data about data” (p 30). Metadata is the information that describes the corpus data, which may be referred to as documenting

the corpus data (Granger, 2002). Burnard (2005) illustrates the importance of having this metadata as a part of the corpus.

“It is no exaggeration to say that without metadata, corpus linguistics would be virtually impossible. Why? Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language. A typical corpus analysis will therefore gather together many examples of linguistic usage, each taken out of the context in which it originally occurred, like a laboratory specimen. Metadata restores and specifies that context, thus enabling us to relate the specimen to its original habitat” (Burnard, 2005).

With respect to which variables should be documented by the metadata in learner corpora, Granger (2002) classifies them into two main categories, learner and task variables.

“Full details about these variables must be recorded for each text.... This documentation will enable researchers to compile subcorpora which match a set of predefined attributes and effect interesting comparisons, for example between spoken and written productions from the same learner population or between similar-type learners from different mother tongue backgrounds” (Granger, 2002: 10).

The ALC was designed to include a number of metadata variables which characterise features of the learners and texts such as “age”, “gender”, “mother tongue”, “text mode”, “place of writing”, etc. These features can be used as determinants to search any subset of the corpus data or to conduct comparisons between different groups of learners or texts. The corpus contains 26 metadata variables: 12 related to the learners and 14 related to the texts (Table 3.3).

Table 3.3: Metadata elements used in the ALC

Learner variables	Text variables
1. Age	1. Text genre
2. Gender	2. Where produced
3. Nationality	3. Year of production
4. Mother tongue	4. Country of production
5. Nativeness	5. City of production
6. Number of languages spoken	6. Timing
7. Number of years learning Arabic	7. References use
8. Number of years spent in Arabic countries	8. Grammar book use
9. General level of education	9. Monolingual dictionary use
10. Level of study	10. Bilingual dictionary use
11. Year/Semester	11. Other references use
12. Educational institution	12. Text mode
	13. Text medium
	14. Text length

As some of the corpus design criteria such as learners' nativeness, materials mode, and genre are also included as metadata variables, only a summary of their details will be mentioned here.

3.3.1 Age

Age is usually used to compare different groups of learners to investigate the effect of age on their language learning. Because including participants under 16 years of age would require further ethical considerations and because data was collected from various educational institutions, the minimum age in the ALC design was 16 with no maximum age.

Learners whose materials are included in the ALC (v2) range in age from 16 to 42; however, the majority were between 16 and 25. Figure 3.3 shows the word distribution of each learner group based on age.

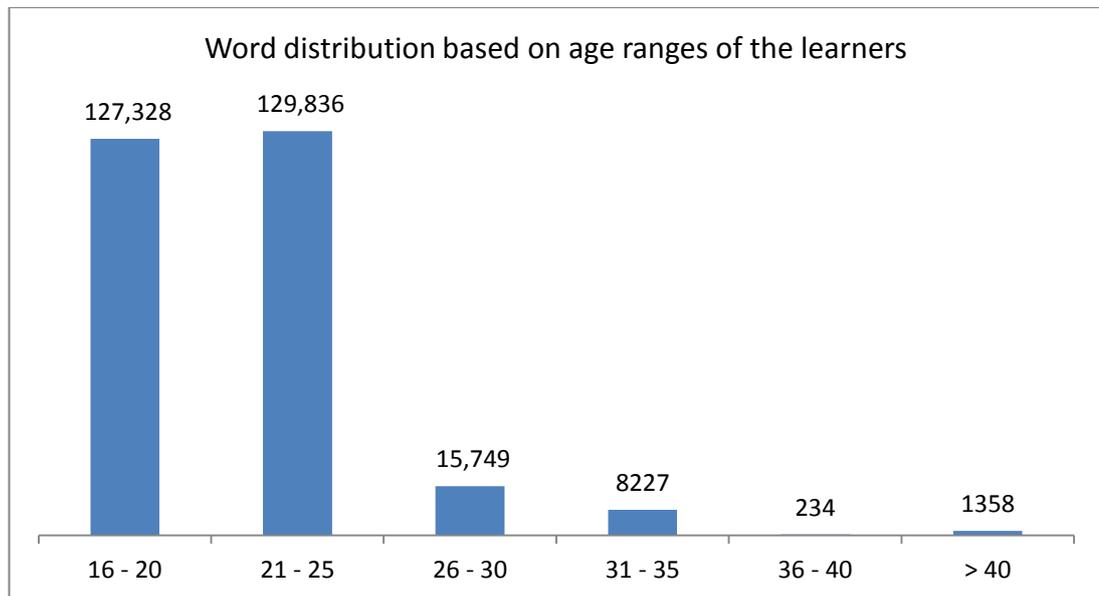


Figure 3.3: Word distribution based on age ranges of the learners

3.3.2 Gender

Special considerations were given to the gender variable of the learners' metadata because in Saudi Arabia, apart from pre-school establishments, all other education delivery is made to single gender classes; that is, males and females do not mix. Segregation of the genders in education is a relatively standardised practice. Therefore, it would have been impossible for a male researcher to enter a female school or university during their operational hours, making it necessary to recruit a number of female representatives to collect the required data from the female educational institutions. As a result of this restriction, the portion devoted to data concerning females in the ALC design was 20%. In terms of the current version of the ALC (v2), two-thirds of the data was produced by 699 male learners whilst 33% was produced by 243 female students (Figure 3.4). The data produced by females was collected by 8 representatives from 18 female educational institutions in Saudi.

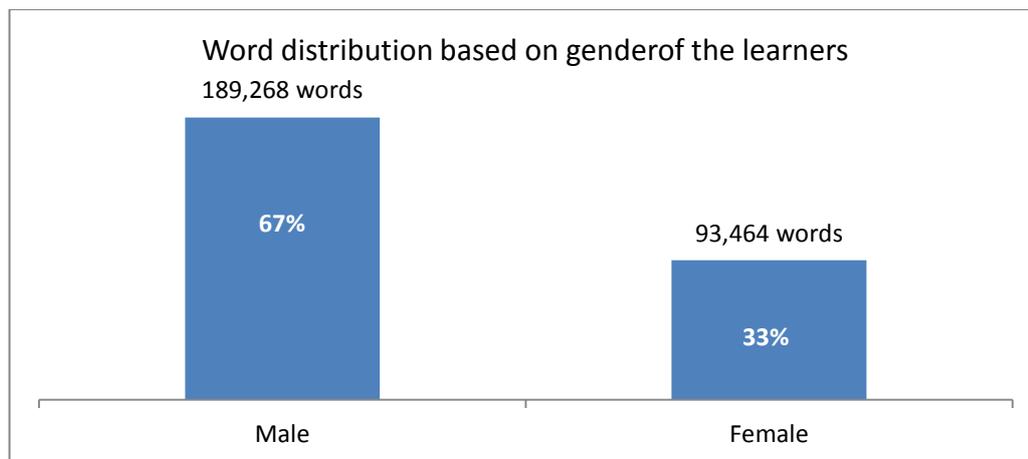


Figure 3.4: Word distribution based on gender of the learners

3.3.3 Nationality

The ALC design does not focus on a specific nationality; thus, the participants represented 67 different countries (Table 3.4). Participants from Saudi Arabia made up 49.38% of the corpus, as most learners in the NS part of the ALC were from Saudi Arabia.

Table 3.4: Distribution of nationalities in the ALC

1. Saudi	49.38%	24. Bengali	0.77%	47. Gambian	0.26%
2. Chinese	3.79%	25. Beninese	0.75%	48. Togolese	0.25%
3. Filipino	3.47%	26. Egyptian	0.73%	49. Canadian	0.21%
4. Guinean	3.16%	27. British	0.60%	50. Polish	0.16%
5. Indian	2.74%	28. French	0.60%	51. Albanian	0.15%
6. Nigerian	2.38%	29. Comorian	0.58%	52. Ukrainian	0.14%
7. Thai	2.17%	30. Somali	0.57%	53. Italian	0.10%
8. Nepalese	1.98%	31. Azerbaijani	0.50%	54. Ugandan	0.09%
9. Malian	1.96%	32. USA	0.46%	55. Kosovar	0.09%
10. Afghan	1.52%	33. Jordanian	0.45%	56. Montenegro	0.08%
11. Djibouti	1.47%	34. Indonesian	0.45%	57. Liberian	0.07%
12. Serbian	1.35%	35. Cambodian	0.41%	58. Central African	0.06%
13. Ivorian	1.34%	36. Senegalese	0.39%	59. Burundi	0.06%
14. Pakistani	1.26%	37. South Korean	0.38%	60. German	0.06%
15. Sri Lankan	1.26%	38. Kyrgyz	0.37%	61. Macedonian	0.05%

16. Burkina Faso	1.13%	39. Niger	0.37%	62. Belgian	0.04%
17. Ghanaian	1.12%	40. Kenyan	0.36%	63. Mongolian	0.04%
18. Syrian	1.09%	41. Turkish	0.33%	64. Lebanese	0.04%
19. Yemeni	1.06%	42. Palestinian	0.32%	65. Ethiopian	0.03%
20. Tajik	0.99%	43. Sudanese	0.32%	66. Kazakh	0.03%
21. Sierra Leonean	0.99%	44. Bosnian	0.32%	67. Dutch	0.02%
22. Malaysian	0.97%	45. Tanzanian	0.31%		
23. Russian	0.77%	46. Uzbek	0.30%		

3.3.4 Mother Tongue

Similar to nationalities, the ALC was designed to include students from various L1 backgrounds. The current version of the corpus (v2) contains 66 different mother tongue representations; specifically, the NNS learners spoke 65 different L1s while all of the NS learners spoke Arabic as their L1; see Table 3.5 for the distribution of L1s within the NNS part of the corpus.

Table 3.5: Distribution of mother tongues in the NNS part of the ALC

Urdu	9.38%	Hausa	1.56%	Kalibugan	0.38%
Chinese	8.41%	Mandinka	1.55%	Polish	0.35%
Somali	5.15%	Uzbek	1.26%	Zarma	0.35%
Malay	5.08%	Manga	1.21%	Susu	0.31%
French	4.52%	Swahili	1.18%	Portuguese	0.30%
English	4.39%	Dagomba	1.15%	Madurese	0.27%
Fulani	4.23%	Tajik	1.08%	Italian	0.22%
Yoruba	3.64%	Comorian	1.06%	Tatar	0.22%
Bosnian	3.15%	Yakan	1.01%	Ugandan	0.20%
Anko	3.13%	Filipino	1.01%	Ingush	0.18%
Bengali	2.91%	Maranao	0.91%	Kotokoli	0.16%
Tamil	2.70%	Cambodian	0.89%	Afar	0.16%
Moore	2.44%	Azerbaijani	0.84%	Modnaka	0.15%
Thai	2.29%	Korean	0.82%	Sango	0.14%
Persian	2.20%	Turkish	0.77%	Kurdish	0.13%
Maguindanao	2.10%	Nepali	0.76%	Malayalam	0.13%
Tagalog	1.75%	Indonesian	0.68%	Mongolian	0.08%

Beninese	1.73%	Albanian	0.68%	Amharic	0.07%
Russian	1.67%	Wolof	0.64%	Jola	0.06%
Soninke	1.64%	Indian	0.50%	Kazakh	0.06%
Bambara	1.62%	Kyrgyz	0.46%	Dutch	0.03%
Pashto	1.61%	Serbian	0.43%		

3.3.5 Nativeness

The learners' nativeness was one of the corpus design criteria and also one of the metadata variables. The data collected from the NS learners was 151,139 words (53%), while NNS learners produced 131,593 words (47%). The close percentages enable researchers to conduct comparative analyses between these two groups.

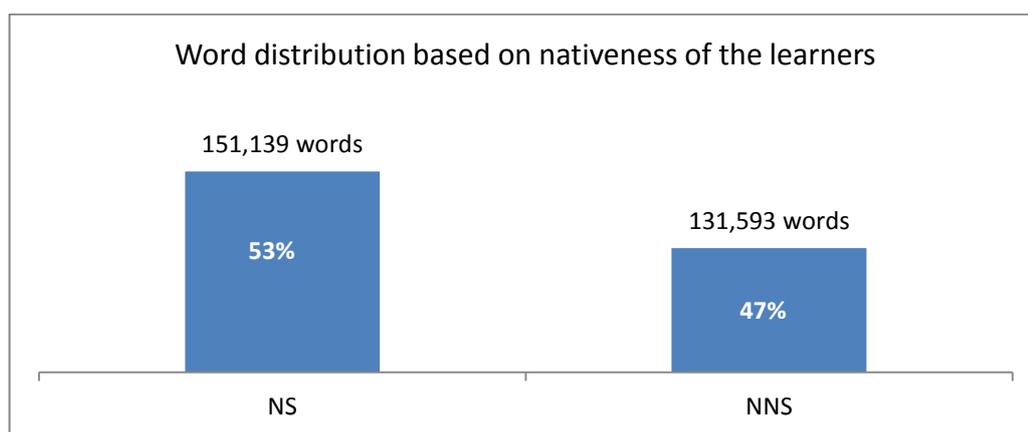


Figure 3.5: Word distribution based on nativeness of the learners

3.3.6 Number of Languages Spoken

Having this element as a metadata variable allows researchers to compare different groups of learners based on how many languages they speak, and to investigate whether this number plays a role in language learning. In the ALC, the number of languages spoken by each learner ranged from 1 to 10 in the case of NNS, while NS learners spoke between 1 and 4 languages.

3.3.7 Number of Years Learning Arabic

Similarly to the previous variable, researchers are able to compare different groups of learners based on how many years they have spent learning Arabic, and to

investigate the role this variable may play in learning Arabic. In terms of the ALC content, learners spent between a few months (indicated as 0 years in the corpus) and 19 years in their acquisition of Arabic since they arrived in Saudi Arabia. The native Arabic speakers were excluded from this category.

3.3.8 Number of Years Spent in Arabic Countries

This variable has the same function as the previous two. Specifically, it assists researchers in conducting comparisons between different groups of learners based on how many years they spent in Arab countries and whether this experience may affect their learning of Arabic. The ALC content indicates that the number of years an individual had spent in an Arabic-speaking country ranged from a few months (indicated as 0 years in the corpus) to 21 years. NS were also excluded from this category. In the corpus's questionnaire, the questions about this item and the previous one were allocated to NNS.

3.3.9 General Level of Education

The International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), a well-designed learner corpus, classifies learners' education levels into secondary school and university. The same classification was used in the ALC, although the first level was named pre-university because it included two parallel groups of learners, NS learning at secondary schools and NNS learning Arabic at institutions that teach Arabic as a second language. Both of these groups are counted as pre-university because they have to master this level before continuing their study at a university. The second level, university, is for both undergraduate and postgraduate students specialising in the same target language, Arabic (Table 3.6).

Table 3.6: Levels of the learners who contributed to the ALC

Level	NS	NNS
Pre-university	Learning at secondary schools	Learning Arabic at institutions where Arabic is taught as a second language
University	Undergraduate and postgraduate students (NS and NNS) specialising in Arabic	

In the design of the ALC, more focus was placed on the pre-university level because a greater number of learners could be recruited from this level. The target was for 140,000 words (70%) to be collected from learners at the pre-university level and 60,000 words (30%) from learners at the university level. The percentage of the ALC data was 80% for pre-university and 20% for university learners (Figure 3.6), though the target number of words was larger in the former level and near the target in the latter.

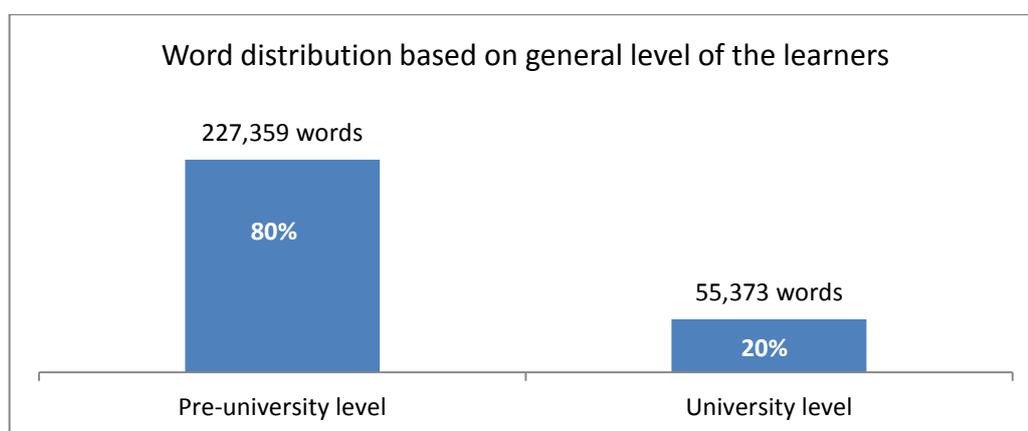


Figure 3.6: Word distribution based on general level of the learners

3.3.10 Level of Study

The ALC includes five levels of study: secondary school (37%), general language course (28%), diploma programme which is an advanced language course (15%), bachelor degree (BA, 13%), and master degree (MA, 7%). Learners from both the BA and MA levels were majoring in Arabic. See Figure 3.7 for the number of words included in the ALC for each level.

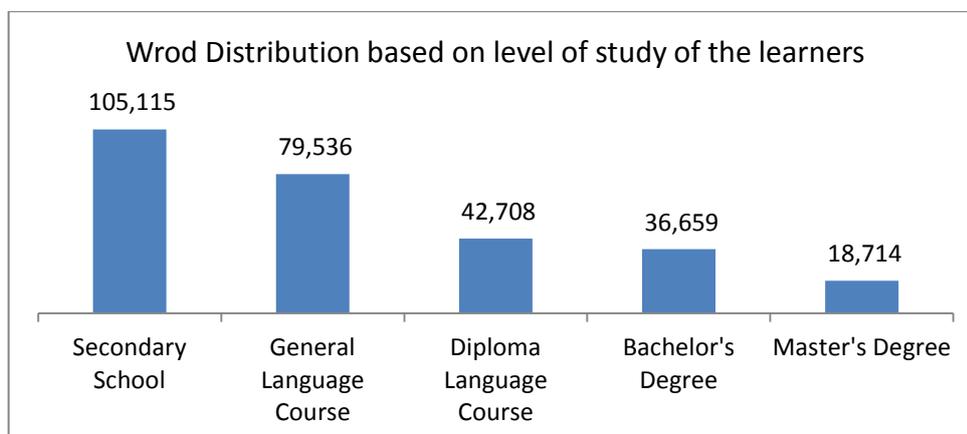


Figure 3.7: Word distribution based on level of study of the learners

3.3.11 Year/Semester

Each of the major levels, pre-university and university, was broken up into an appropriate number of sub-categories based on the levels (i.e. year or semester) used in their institutions. The designation of these sub-categories followed the British Academic Written English Corpus (Heuboeck *et al.*, 2008) which divides learners based on their year of study as a level indicator. The level of study was represented by a range of three years for the secondary school students (1st = 12.4%, 2nd = 9.5%, and 3rd = 15.28%) and eight semesters for the other groups: general and diploma language courses, BA, and MA (1st = 19.03%, 2nd = 3.84%, 3rd = 10.39%, 4th = 21.86%, 5th = 4.25%, 6th = 1.47%, 7th = 1.58%, and 8th = 0.41%); see Figure 3.8 for the word distribution in the ALC.

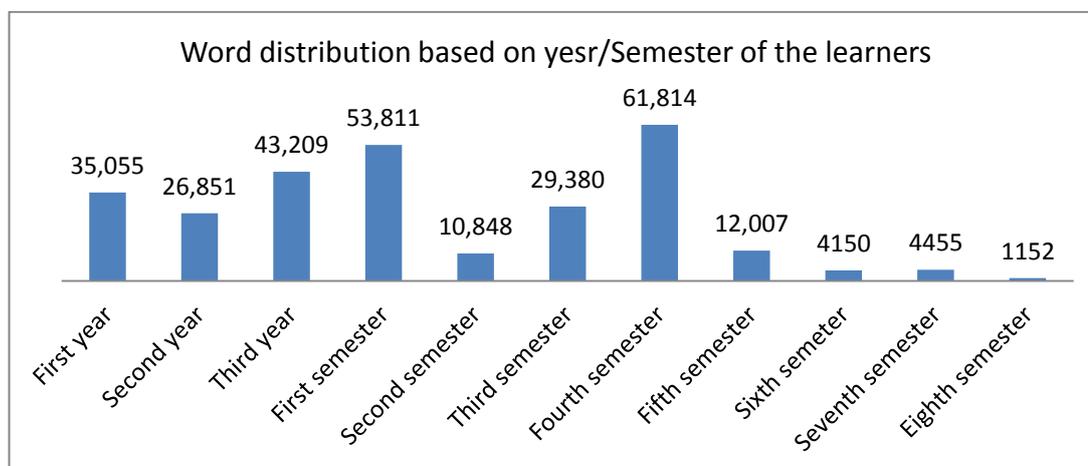


Figure 3.8: Word distribution based on year/semester of the learners

Table 3.7 illustrates the word distribution based on the previous three hierarchical levels combined together (general level, level of study, and year/semester).

Table 3.7: Word distribution based on general level, level of study, and year/semester

General level	Level of study	Year/Semester	No. of words	Percentage of the ALC
Pre-university	Secondary School	1 st year	35,055	12.40%
		2 nd year	26,851	9.50%
		3 rd year	43,209	15.28%
	General Language Course	3 rd semester	24,874	8.80%
		4 th semester	54,662	19.33%
	Diploma Language Course	1 st semester	24,465	8.65%
		2 nd semester	10,760	3.81%
		3 rd semester	3022	1.07%
		4 th semester	4461	1.58%
	University	Bachelor degree	1 st semester	10,632
2 nd semester			88	0.03%
3 rd semester			1484	0.52%
4 th semester			2691	0.95%
5 th semester			12,007	4.25%
6 th semester			4150	1.47%
7 th semester			4455	1.58%
8 th semester			1152	0.41%
Master degree		1 st semester	18,714	6.62%
Total			282,732	100.00%

3.3.12 Educational Institution

The ALC was designed to include various educational institutions, i.e. secondary schools, language institutions, and universities. In the current version, the participants were affiliated to 25 institutions. Table 3.8 shows how many words were collected from each institution alongside their percentage of the ALC.

Table 3.8: Word distribution based on institutions from where the ALC data was collected

Institute	No. of words	Percentage of the ALC
1 <i>Arabic Institute at Al-Imam University</i>	95,655	33.83%
2 <i>Alshura Secondary School for Boys in Riyadh</i>	28,799	10.19%
3 <i>Arabic College at Imam University</i>	24,330	8.61%
4 <i>Arabic Institute At PNU</i>	17,297	6.12%
5 <i>Capital Model Institute</i>	16,341	5.78%
6 <i>Arabic Institute at KSU</i>	14,960	5.29%
7 <i>Arabic Department at PNU</i>	13,571	4.80%
8 <i>The Sixth Secondary School for Girls in Qatif</i>	13,356	4.72%
9 <i>Arabic Institute at Umm Al-Qura University</i>	11,804	4.17%
10 <i>The Scientific Institute in Alkharj</i>	9124	3.23%
11 <i>The Third Secondary School for Boys in Riyadh</i>	7714	2.73%
12 <i>The Second Secondary School for Girls in Jesh</i>	5624	1.99%
13 <i>The Fourth Secondary School for Girls in Qatif</i>	4296	1.52%
14 <i>The Eighth Secondary School for Girls in Qatif</i>	4121	1.46%
15 <i>The Forty-Ninth Secondary School for Girls in Jeddah</i>	3555	1.26%
16 <i>The First Secondary School for Girls in Qatif</i>	2896	1.02%
17 <i>The Second Secondary School for Girls in Mahayil Asir</i>	2205	0.78%
18 <i>The Twenty-Third Secondary School for Girls in Hafr Albatin</i>	1680	0.59%
19 <i>The Thirty-Three Secondary School for Girls in Riyadh</i>	1558	0.55%
20 <i>The Twenty-Ninth Secondary School for Girls in Jeddah</i>	1493	0.53%
21 <i>The Forty-Eighth Secondary School for Girls in Jeddah</i>	654	0.23%
22 <i>The Twenty-First Secondary School for Girls in Jeddah</i>	556	0.20%
23 <i>The Fifty-Eighth Secondary School for Girls in Jeddah</i>	417	0.15%
24 <i>The Eighty-Fourth Secondary School for Girls in Jeddah</i>	379	0.13%
25 <i>The Eighth Secondary School for Girls in Jeddah</i>	347	0.12%

3.3.13 Text Genre

The ALC was designed to cover two text genres, narrative and discussion. The corpus content consists of 67% narrative texts and 33% discussion texts. This variable was explained in detail under the corpus design criteria (2.2.9).

3.3.14 Where Produced

This variable identifies two types of texts: those produced in class and at home. A text written in class may differ from one written at home, as the learner could have further sources of assistance at home. Comparing texts written in these two places may reveal some insights about the learner's language. By including this variable, the ALC follows some standard learner corpora such as the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), the Spoken and Written English Corpus of Chinese Learners (Wen, 2006), and the Montclair Electronic Language Database (Eileen & Milton, 2012; Fitzpatrick & Seegmiller, 2001; Pravec, 2002). Learners were allowed to choose to write their texts in class (62% of the ALC data) or at home (31%). However, all the audio recordings were produced in class (7%). The form explaining the at-home assignment was distributed to the same students who completed the in-class assignment. The fact that 62% of the corpus was written in class indicates that learners seem to be more motivated while performing in-class tasks.

3.3.15 Year of Production

The researcher conducted two field trips to collect the corpus data from learners in Saudi Arabia. During the first trip in November and December 2012, data for version 1 of the ALC was collected. The data gathered on this trip represents 12% of the ALC content, as it was a pilot study to collect about 10% and to explore the processes needed for developing the entire corpus. Data for version 2 was collected during the second trip from 15 August to 15 November 2013. The data collected in this trip forms 88% of the final content. Because the amount of data collected over a three-month period was much greater than that in the pilot study, more preparation was necessary for the second trip.

3.3.16 Country of Production

This variable is usually used by international learner corpora such as the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010). The current version of the ALC includes data from a sole country, Saudi Arabia. This variable was added to the corpus metadata for future expansion. The researcher plans for the corpus to cover learning Arabic in other Arabic-speaking countries, as well as in non-Arabic-speaking countries. This variable allows

researchers to undertake comparisons between learners of these countries individually or in groups, e.g. Arabic-speaking countries vs. non-Arabic speaking countries.

3.3.17 City of Production

Similarly to the previous variable, knowing the city of production may enable researchers to investigate whether there are any differences in the language use of learners within those cities. This variable is especially useful in large countries such as Saudi Arabia which has many dialects and accents that could affect the learner's language. The ALC was designed to include data from different regions of Saudi Arabia, namely the centre (Riyadh and Alkharj), north (Hafr Albatin), south (Mahayil Asir), east (Alqatif and Aljesh), and west (Makkah and Jeddah). In terms of data gathered, the current version of the ALC data was collected from eight cities, Riyadh (77%), Alqatif (9%), Makkah (4%), Jeddah (3%), Alkharj (3%), Aljesh (2%), Hafr Albatin (1%), and Mahayil Asir (1%). The map in Figure 3.9 illustrates the locations of the cities from which the ALC data was collected. Most of the data was collected from Riyadh, as it contains the highest number of schools, language institutions, and universities compared to the other cities.

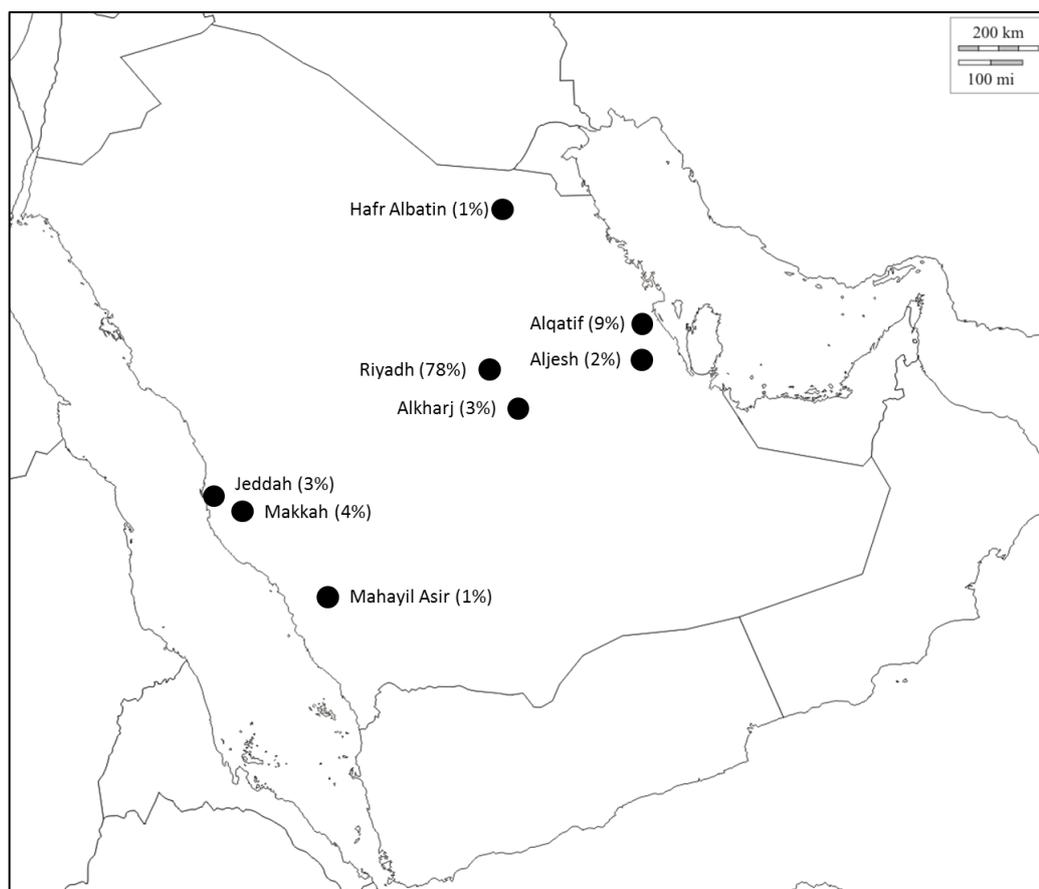


Figure 3.9: Locations of the Saudi cities from which the ALC data was collected¹

3.3.18 Timing

Including timed writing and untimed writing is a standard practice in developing learner corpora; see for example the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), the Montclair Electronic Language Database (Eileen & Milton, 2012; Fitzpatrick & Seegmiller, 2001; Pravec, 2002), the Chinese Learner English Corpus (Shichun, 2012; Wen, 2006), the Corpus Archive of Learner English in Sabah/Sarawak (Arshad, 2004; Botley & Dillah, 2007; Botley, 2012), the Hong Kong University of Science & Technology learner corpus (Milton & Nandini, 1994; Pravec, 2002), the Spoken and Written English Corpus of Chinese Learners (Wen, 2006), and the TELEC Secondary Learner Corpus (Pravec, 2002). Timing in the learner corpora aforementioned is usually based on the location of the material being produced; specifically, the materials

¹ This free map of Saudi Arabia was obtained from http://d-maps.com/carte.php?num_car=31&lang=en under the terms and conditions of use

produced in class are timed and those produced at home are not timed. For the ALC v2, 69% of the essays were timed (in class), and 31% were untimed (at home).

3.3.19 Use of References

This variable was modelled after the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010) and indicates whether any reference source was used by the learner in his or her writing. References include four main sources: (i) grammar books, (ii) monolingual dictionaries, (iii) bilingual dictionaries, or (iv) other references (e.g. the Internet, newspapers, radio, TV, etc.). Each source type is represented by an independent variable for those who need to conduct more specific analysis. Learners were allowed to use those references in their writing, which may enable researchers to investigate the possible effect of using such references on learners' language. In the ALC, references were used in 5% of the corpus data. Learners used the aforementioned source types as described in the following four variables.

3.3.20 Grammar Book Use

Under the larger category of “References Use”, this variable is devoted to one type of reference that learners may use in writing, grammar books. Using grammar books enables learners to improve the structure of their writing and to avoid grammatical errors. Grammar books were used in 2% of the ALC data.

3.3.21 Monolingual Dictionary Use

Because rapid technological developments have allowed electronic dictionaries to be used on portable devices such as smart phones, the researcher expected monolingual dictionaries to be used by both native and non-native Arabic-speaking students. However, only NNS learners used monolingual dictionaries which were used in 1% of the ALC data.

3.3.22 Bilingual Dictionary Use

Only NNS students used bilingual dictionaries to help in translating the vocabulary they wanted to use in their writing or to learn about the use or forms of those words. Bilingual dictionaries were used in 2% of the corpus.

3.3.23 Other References Use

The category of other references includes any linguistic references that learners may use except grammar books, monolingual dictionaries, and bilingual dictionaries, as they were considered as independent variables. For example, the Internet, newspapers, radio, and TV are counted as other references. Learners were advised to use other references not as sources of information for their writing but to help improve the linguistic aspects of their writing such as vocabulary, grammar, and style. In total, 2% of the ALC texts were produced with the use of other references.

3.3.24 Text Mode

As described in the ALC design criteria, the plan was to collect a total of 200,000 words, divided into 180,000 words (90%) of written text and 20,000 words (10%) of spoken data. The current version of the corpus (v2) includes 282,732 words in total, with 263,045 words of written text and 19,687 words of transcriptions in the spoken part. The original audio recordings consist of 3 hours, 22 minutes, and 59 seconds of speech.

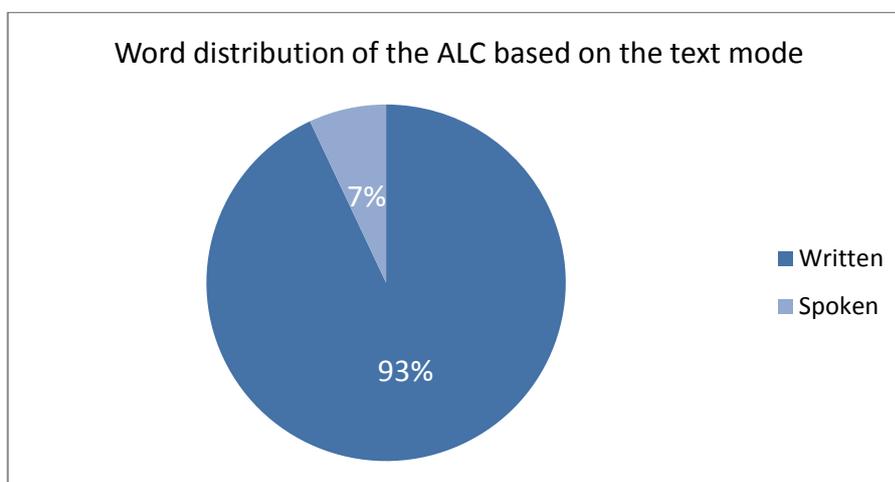


Figure 3.10: Word distribution of the ALC based on the text mode

3.3.25 Text Medium

The corpus includes two mediums of written data, text produced by hand (208,355 words) and text produced on a computer (54,690 words). Auditory data was collected in the form of recorded interviews only (19,687 words).

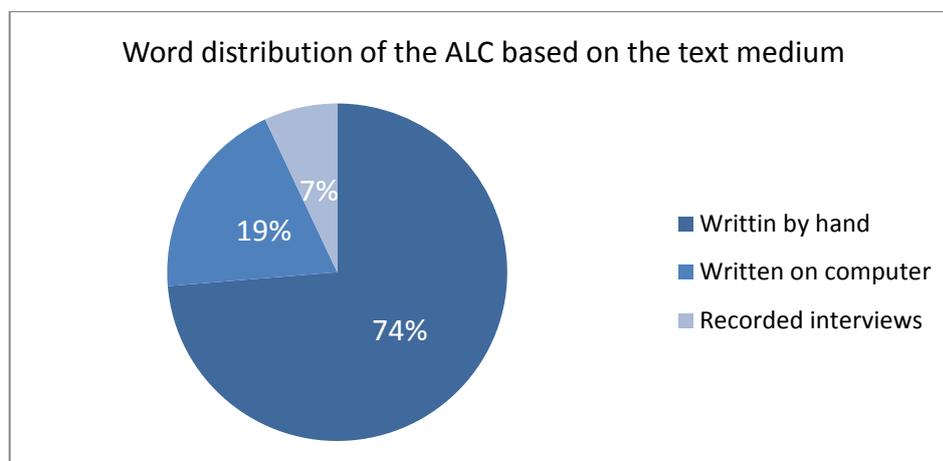


Figure 3.11: Word distribution of the ALC based on the text medium

3.3.26 Text Length

The ALC includes 1585 texts (written texts and transcriptions of spoken data). Participants were asked to produce about 500 words as an average length for each text. However, the lengths of texts included in the ALC v2 varied considerably from one sentence (3 words) in the shortest to 7298 words in the longest. Although the shortest texts may not be full essays, the researcher included them in the ALC for an authentic representation of the learners' productions. There are six texts representing the longest with 1000 words or more, and seven texts representing the shortest with 10 words or less (see Figure 3.12). The average length of the texts in the ALC is 178 words. Table 3.9 lists more length averages based on some factors that may help researchers to conduct further analysis to investigate reasons behind the differences in these averages.

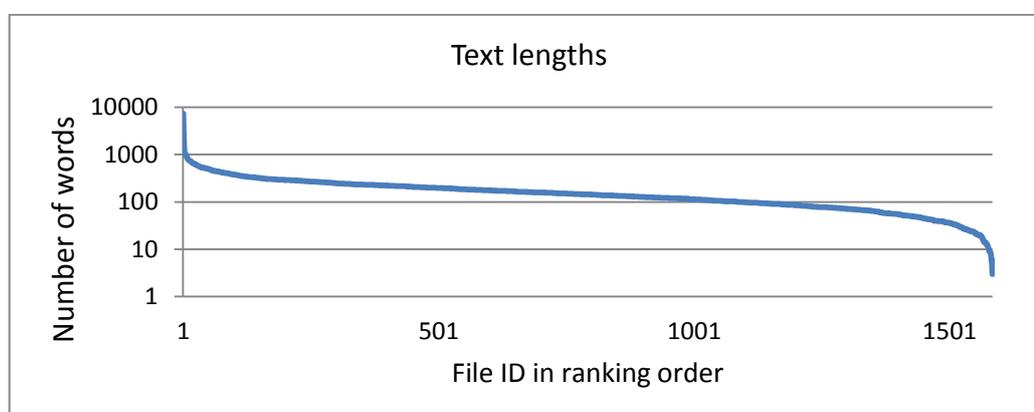


Figure 3.12: Lengths of the ALC texts

Table 3.9: Average length of the ALC texts based on some key factors

Factor	Average length	
Learners' gender	Males = 166	Females = 209
Learners' nativeness	NS = 191	NNS = 166
Learners' general level of education	Pre-university = 164	University = 283
Place of production	In class = 163	At home = 227
Text genre	Narratives = 205	Discussions = 145
Text mode	Written = 172	Spoken = 334

3.3.27 Summary of the ALC Metadata

Table 3.10 summarises the metadata variables with the values they contain and the percentages they represent in v2 of the ALC data.

Table 3.10: Summary of the variables used in the ALC metadata

1. Variable = Age	Values = range from 16 to 42
2. Variable = Gender	Values = Male (67%), Female (33%)
3. Variable = Nationality	Values = 67 nationalities
4. Variable = Mother tongue	Values = 66 first languages
5. Variable = Nativeness	Values = Native (53%), Non-native (47%)
6. Variable = Number of languages spoken	Values = range from 0 to 10
7. Variable = Number of years learning Arabic	Values = range from 0 to 19 years
8. Variable = Number of years spent in Arabic countries	Values = range from 0 to 21
9. Variable = General level of education	Values = Pre-university (80%), University (20%)
10. Variable = Level of study	Values = Secondary school (37%), General language course (28%),

Diploma language course (15%), BA (13%), MA (7%)

11. Variable = Year/Semester

Values = 1st year (12.4%), 2nd year (9.5%), 3rd year (15.28%), 1st semester (19.03%), 2nd semester (3.84%), 3rd semester (10.39%), 4th semester (21.86%), 5th semester (4.25%), 6th semester (1.47%), 7th semester (1.58%), 8th semester (0.41%)

12. Variable = Educational institution

Values = 25 educational institutions

13. Variable = Text genre

Values = Narrative (67%), Discussion (33%)

14. Variable = Where produced

Values = In class (69%), At home (31%)

15. Variable = Year of production

Values = 2012 (12%), 2013 (88%)

16. Variable = Country of production

Values = Saudi Arabia (100%)

17. Variable = City of production

Values = Riyadh (77%), Alqatif (9%), Makkah (4%), Jeddah (3%), Alkharj (3%), Aljesh (2%), Hafr Albatin (1%), Mahayil Asir (1%)

18. Variable = Timing

Values = Timed (69%), Not timed (31%)

19. Variable = References use

Values = Yes (5%), No (95%)

20. Variable = Grammar book use

Values = Yes (2%), No (98%)

21. Variable = Monolingual dictionaries use

Values = Yes (1%), No (99%)

22. Variable = Bilingual dictionaries use

Values = Yes (2%), No (98%)

23. Variable = Other references use

Values = Yes (2%), No (98%)

24. Variable = Text mode

Values = Written (93%), Spoken (7%)

25. Variable = Text medium

Values = Written by hand (74%), Written on computer (19%), Interview

recorded (7%)

26. Variable = Text length

Values = range from 3 to 7298 words

3.4 Corpus Evaluation

In this section, the Arabic Learner Corpus will be evaluated on its impact (i.e. works that have used the ALC), feedback from some specialists in computation and corpus linguistics, and the download rate from the corpus website which may support the extent of the corpus use.

3.4.1 Projects That Have Used the ALC

The ALC has been used for different purposes and applications that are described in detail in Chapter 7 and are listed here in order to highlight the ALC's impact. The ALC has been used for the following purposes and applications:

- Error detection and correction tools (Farra *et al.*, 2014; Obeid *et al.*, 2013);
- Error annotation guidelines (Zaghouani *et al.*, 2014);
- Native language identification systems (Malmasi & Dras, 2014);
- A training workshop on Arabic teaching (Alharthi, 2015);
- Evaluating robustness of the main existing Arabic analysers (Alosaimy, Alfaifi and Alghamdi, forthcoming);
- Applied linguistics studies including:
 - Alshaiban's (undertaking) PhD thesis started in 2014,
 - Alshehri's (undertaking) PhD thesis started in 2015,
 - Alqawsi's (personal communication, 1 April 2015) study on Arabic word frequency,
 - Alharthi's (personal communication, 13 April 2015) study of the influence of using corpora on Arabic learners' motivation; and
- Data-driven Arabic learning (Refaee, personal communication, 22 February 2015; Isma'il, personal communication, 4 April 2015).

These examples reveal that the use of the ALC has increased from its first release (v1) in March 2013 (1 work) to the time of writing in April 2015 (6 works); the second version was released in February 2014 (see Figure 3.13). The starting date was used to represent those works in progress.

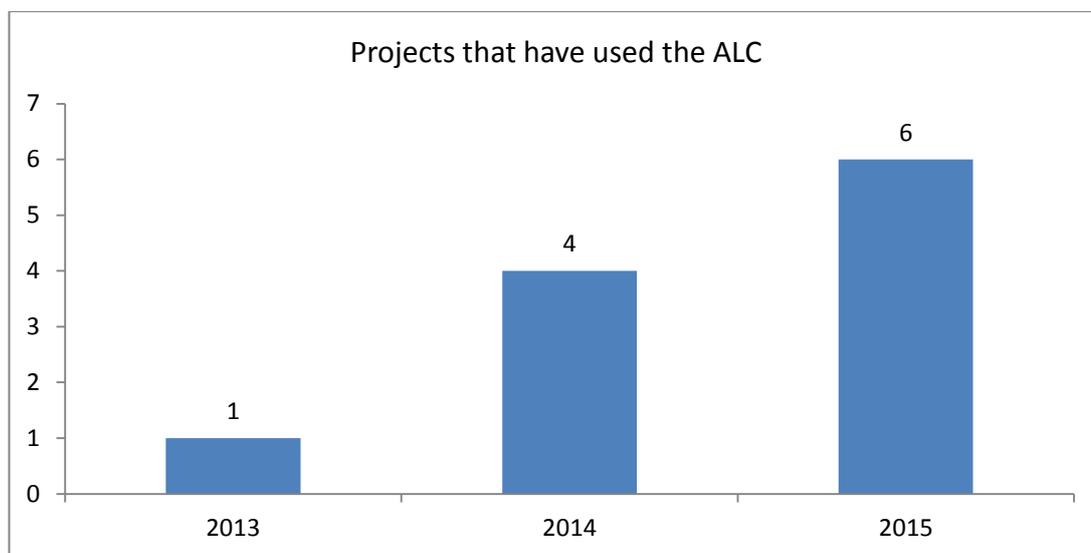


Figure 3.13: Projects that have used the ALC

Based on Figure 3.13, it can be expected that the ALC will be used in more work in the future, particularly once it has been entirely annotated for errors which is expected to be completed within two to three years based on the proposal suggested to complete the work¹. This finding makes it even more important to continue working on the additions and improvements to the ALC which are described in the future work section in Chapter 8.

3.4.2 Specialists' Feedback

A number of specialists in natural language processing and corpus linguistics were asked to provide general comments as feedback on the Arabic Learner Corpus project (e.g. on the design, content, uses, etc.). Their responses were valuable and positive, as illustrated in the following examples:

- **Professor Shin Ishikawa, School of Languages and Communication, Kobe University, Japan**

“The ALC is a brand-new learner corpus and it is expected to shed a new light on analysis of interlanguage of learners of Arabic.

Considering that there have been almost no freely available Arabic corpora to date, its academic value and contribution cannot be overestimated.

¹ See a copy of the project proposal on:

http://www.comp.leeds.ac.uk/scayga/Alfaifi_annotation_grant.pdf

Carefully analyzing the designs of major existing corpora and their potential drawbacks, Abdullah Alfaifi and his team have established detailed protocols to collect spoken and written data, which I think leads to high reliability of the data collected in ALC.

As one of the researchers in the field of learner corpus studies, I would like to congratulate on the compilation of the ALC project”.

- **Professor Nizar Habash, Computer Science, New York University Abu Dhabi, United Arab Emirates**

“Much of the research in natural language processing / computational linguistics is driven by resources: corpora, treebanks, and other sorts of annotated data. These valuable data treasures are costly and time consuming to build and need to be developed with care to maximize their utility for different researchers.

Arabic has been gaining a lot of interest in the last decade, but up to the time of the creation of the ALC, there has not been a large scale carefully annotated resource for Arabic learners. There were some early important efforts of course, but their small size limited their usability.

The collected corpus size and detailed annotations done by Mr. Alfaifi make the ALC an important resource that will influence a lot of work on Arabic technology (e.g. text correction). I applaud his effort and support extending the resource even further”.

- **Professor James Dickins, School of Arabic, Middle Eastern and East Asian Studies, University of Leeds, UK**

“Abdullah Alfaifi’s Arabic Learner Corpus is a corpus of written – and some spoken – materials produced by learners of Arabic with a large range of different first languages.

The corpus is very good for error analysis among learners of Arabic, because it allows for identification of errors according to numerous specific categories. The corpus will be particularly useful not only for Arabic L2 error analysis researchers but anyone working on problems in Arabic teaching and learning”.

- **Professor Yukio Tono, Graduate School of Global Studies, Tokyo University of Foreign Studies, Japan**

“I found the ALC very well designed and systematically collected. Especially I liked the idea of collecting data from both pre-university and university students as well as native vs non-native, which makes a unique, interesting comparison across subcorpora. They also provide very specific metadata, showing that the corpus compilation has been carefully done”.

- **Ali Hakami, Arabic Language Institute, Al Imam Mohammad Ibn Saud Islamic University, Saudi Arabia**

“We have been waiting for a long time for a corpus design such as this one for Arabic learners. Undeniably we (as Arabic language specialists) are late into our research and services regarding teaching and learning Arabic Language, as L1 or L2. No one can question how much benefit we can gather from the Arabic Learner Corpus.

Linguistics and Applied Linguistics researchers have lots of ideas and lots of research projects, which rely heavily on such a corpus. For instance:

- *Designing books and materials for teaching and learning Arabic for specific purposes.*
- *Creating tests to examine strategies used by Arabic L2 learners.*
- *Structuring frequency dictionaries of Arabic for learners and teachers.*

The current corpus is well organised, easy to follow and is used by scholars for different research aspects and purposes. We can only congratulate Mr. Abdullah and his supervisor Dr Eric Atwell on this great achievement, and we wish them more creativity and success”.

- **Ayman Alghamdi, Arabic Language Institute, Umm Al-Qura University, Saudi Arabia**

“You put a lot of effort into this remarkable and unique project to service learning and teaching Arabic as a second language.

This project leads me to be optimistic about the future of research on Arabic Applied Linguistics”.

3.4.3 Downloads from the ALC Website

Statistics from the ALC website show that 5845 unique visitors from 108 countries across the world performed a total of 16,251 downloads of the website resources from 5 February 2014 to 5 February 2015. Those downloads include the corpus files, publications, the ETAr and its manual ETMAr. Figure 3.14 shows a world map of users of the ALC with higher numbers of users shaded in darker blue.

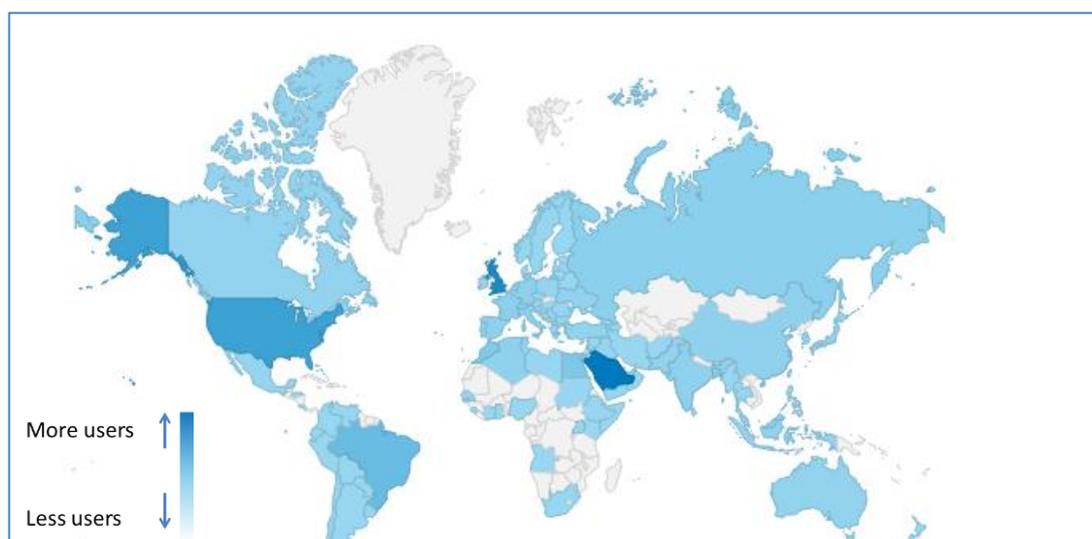


Figure 3.14: Google Analytics map showing locations of ALC visitors¹

3.5 Conclusion

The ALC was developed based on 11 design criteria: the corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation. This chapter describes those design criteria and links each criterion to the relevant literature review before discussing the ALC design target and the content achieved in both versions of the ALC (v1 and v2). In addition to those criteria, the ALC was designed to include 26 elements as metadata variables, 12 for the learner and 14 for the text that the learner wrote. The chapter describes those metadata elements in terms of the design target and the content achieved for each element. The last section in this chapter highlights the increasing interest in using the ALC data

¹ The map was obtained from the free service Google Analytics (<https://www.google.com/intl/en/analytics>) on 5 February 2015.

through (i) the projects which have used the corpus, (ii) the comments received from a number of specialists, and (iii) the downloads from the ALC website. They all give positive feedback about the project and its use.

4 Collecting and Managing the ALC Data

Chapter Summary

This chapter describes the method of collecting and managing the ALC data. The description covers the questionnaire and guidelines that were designed to collect the corpus data. It also covers the process of converting the hand-written texts and spoken materials into an electronic form according to specific standards created for transcribing the hand-written data. The chapter presents the method followed to measure the consistency between transcribers of both written and spoken data. It also describes the database which was developed to store and manage the ALC data, as well as to generate the corpus files automatically in different formats (TXT and XML) using a file generation function. The chapter concludes by illustrating the method of naming the ALC files which reflects the basic characteristics of the text and its author.

4.1 Introduction

The corpus data was not taken from previously existing materials; instead, a particular methodology was designed to carefully collect and manage the corpus data. This methodology includes (i) designing tasks and a questionnaire with guidelines to be followed for this process, (ii) defining the standards for converting the hand-written texts and spoken materials into an electronic form, (iii) measuring the consistency between transcribers of both written and spoken data, and (iv) creating a database to store and manage the ALC data and generate different types of files automatically. The methodology including all these processes is described in the following sections.

4.2 Collecting the ALC Data

The ALC contains three types of media: materials written by hand, texts written on a computer, and spoken data. As a result, three versions of the questionnaire were used. All three included the same questions, but the design was different in order to suit each medium. All the instruments used to collect the corpus data were in two languages, Arabic as the target language and English as an international language.

Guidelines were created to clarify the steps the researcher (or his representative) followed for collecting the ALC data (Appendix B). Data collection involved one main session that was repeated with each group of students, typically representing one class, at each educational institution. During this sole session, which was expected to last for about 2 hours, a questionnaire was distributed and procedures were explained to the participants. The questionnaire consists of five parts (Appendix C) as follows:

1. Brief outline of the project, the benefit, the procedures of data collection, and participation in the research.
2. Consent form in which the participant agrees that (i) he or she has read and understood the information explaining the research project and has had the opportunity to ask questions about it, (2) he or she will take part voluntarily in the research project, and (3) the data collected will be published and used in relevant future research.
3. Learner and task metadata (information about the participant and the task being performed).

4. Task 1 which includes writing two texts (narrative and discussion) in class.
5. Task 2 which includes writing two texts (narrative and discussion) at home.

After the researcher introduced the research, learners were allowed to ask any question about the research, its purposes, or their participation before signing the form. Then the first task was distributed with an explanation on how to complete it. In the last part of the session, Task 1 was collected from the learners and Task 2 was distributed to be performed at home. The participants had the choice to do either one or both of the tasks. Each task involved similar topics (narrative: a vacation trip, and discussion: my study interest), but the first task was timed (40 minutes for each text) and the learners were not allowed to consult any language references (e.g. dictionaries, grammar books) while writing their essays. Students completing the second task were asked to write essays at home about the same topics as in task 1. They were allowed two days to complete the homework and were granted the opportunity to use any language references they selected. The use of references was intended to enable them to improve their writing before submitting their work. Figure 4.1 shows the instructions for both tasks, and Table 4.1 illustrates the procedures followed in each session of data collection.

Task 1 Instructions

(In class)

First text: write a narrative essay about a vacation trip providing as many details as you can about this trip.

Second text: write a discussion essay about your study interest providing as many details as you can and also your future plans to continue your study and to work in this field.

Time: 40 minutes for each text.

Place: in class.

Language references: during this task you are NOT allowed to use any reference tools such as dictionaries or grammar books.

Medium of writing: writing these texts is by hand on the sheets provided by the researcher; two pages are provided for each text, and you can ask for more if needed.

Task 2 Instructions	
(At home)	
First text: write a narrative essay about a vacation trip providing as many details as you can about this trip.	
Second text: write a discussion essay about your study interest providing as many details as you can and also your future plans to continue your study and to work in this field.	
Time: one to two days.	
Place: at home.	
Language references: during this task you are allowed to use any reference tools such as dictionaries or grammar books.	
Medium of writing: writing this text is by hand on the sheets provided by the researcher; two pages are provided for each text, and you can use more if needed.	

Figure 4.1: Instructions for Tasks 1 and 2 of the hand-written materials

Table 4.1: Summary of the data collection procedures

Procedure	Description	Time (estimated)
Introduction	<ul style="list-style-type: none"> - To introduce the research purposes, benefits, and methods of participation, and to answer questions that learners may ask. - To distribute the participant consent form to be signed by the learners. 	30 minutes
Task 1	To write narrative and discussion compositions in class about topics provided (<i>A Vacation Trip</i> for the narration genre and <i>My Study Interest</i> for the discussion), with no use of references.	No more than 40 minutes for each composition
Task 2	To explain the second task, which is to write narrative and discussion compositions on the same topics at home, where the use of references is allowed.	10 minutes

An additional online copy of the questionnaire was created by the researcher using Google Forms¹ – in Arabic and English as the paper version – to collect texts in an electronic format (Figure 4.2). This questionnaire includes the same content as the paper form, and it was used in schools and departments that allowed the researcher – or his representatives – to use computer laboratories. In these situations, learners’ texts were included in the corpus database without the need to carry out the transcribing process.

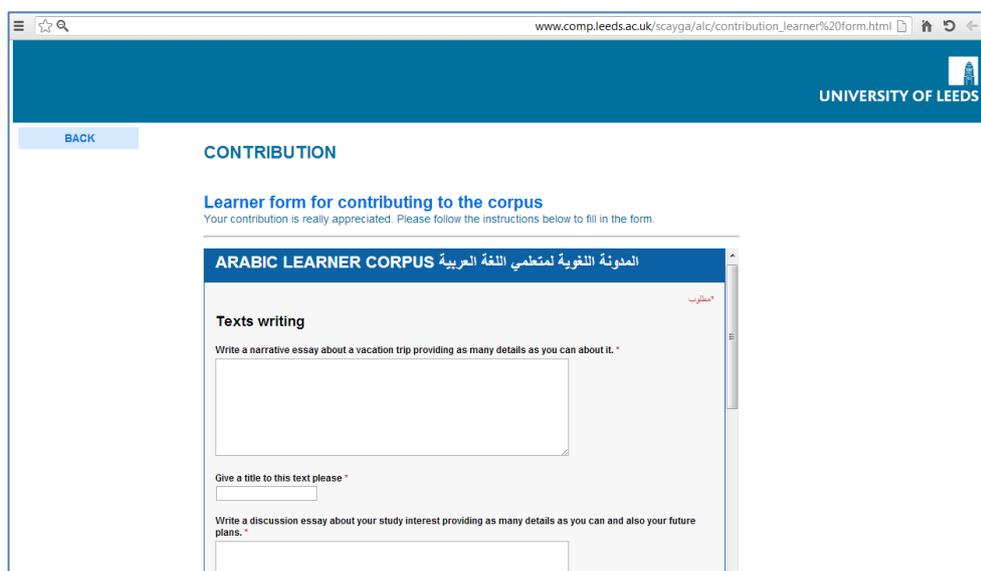
The image shows a screenshot of a web browser displaying an online form. The browser's address bar shows the URL 'www.comp.leeds.ac.uk/scayga/alc/contribution_learner%20form.html'. The page header includes the 'UNIVERSITY OF LEEDS' logo and a 'BACK' button. The main heading is 'CONTRIBUTION', followed by the sub-heading 'Learner form for contributing to the corpus'. Below this, there is a note: 'Your contribution is really appreciated. Please follow the instructions below to fill in the form.' The form itself is titled 'ARABIC LEARNER CORPUS' and 'المجموعة اللغوية لمتعلمي اللغة العربية'. It contains two main sections: 'Texts writing' and 'Give a title to this text please'. The 'Texts writing' section has a large text area and a prompt: 'Write a narrative essay about a vacation trip providing as many details as you can about it.' The 'Give a title to this text please' section has a smaller text area and a prompt: 'Write a discussion essay about your study interest providing as many details as you can and also your future plans.'

Figure 4.2: Online form for data collection

The first task of the written texts was also used to collect the oral data. One to three participants were selected for each recording session. The same procedures were followed as those for the written materials; however, the learners were asked to talk about their topics orally. Learners had the same limited amount of time to give a talk about their chosen topic without the use of any language references. All talks were recorded as MP3 files. Due to some differences in recording conditions, one of the researcher’s representatives collecting the oral data from the female participants was not able to use the corpus devices that produce 44100 Hz 2-channel files, so she used a different device which yielded 16000 Hz 1-channel files in 11 recordings out of 52.

¹ <https://docs.google.com/forms>

4.3 Collecting the ALC Metadata

The learner profile questionnaire of the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010) was used to collect the metadata for the ALC by making some modifications in order to suit the corpus purposes. The form, for example, was split into two separate sheets, a learner profile and text data, because a learner may produce more than one piece of text. Those questions about the learner's relatives were omitted such as father's mother tongue, mother's mother tongue, etc. In total, 26 elements were collected as the corpus metadata, 12 related to the learner and 14 associated with the text.

4.4 Computerising the ALC

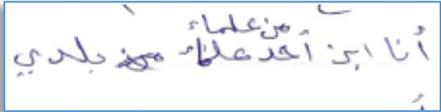
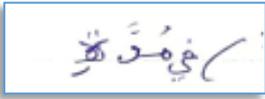
Those corpora containing hand-written texts and spoken materials required further work to convert them into an electronic form such as the plain text format which is readable by most language processing tools, and subsequently to handle tags of mark-up languages such as XML. Transcribing such hand-written and audio materials with no standards, specifically by more than one transcriber, yielded differences in the final production, as many items may be omitted or added during the transcription process and thus distort the results of the corpus analysis (see for example Pastor-i-Gadea *et al.*, 2010; Thompson, 2005). For the converting process, the researcher developed and used standards, which are described below.

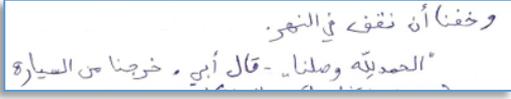
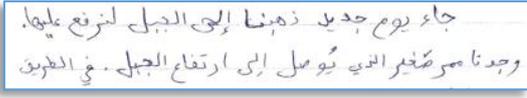
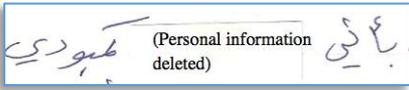
4.4.1 Transcribing Hand-Written Data

As most of the ALC data is derived from hand-written texts and no standard practice was found for transcribing Arabic from hand-written into computerised form, the researcher created specific standards in order to achieve a high level of consistency in transcription. Those standards address matters such as how to handle an overlap between two hand-written characters that cannot be transcribed together, a doubtful form of a character, or forgetting a character's dots. Three transcribers, the researcher and two volunteering colleagues (C1 and C2) who work as teachers of Arabic to NNS learners at Al-Imam University, performed the transcription based on a number of agreed-upon standards. Most of these standards had been extracted by the researcher in advance by reading the hand-written texts in order to identify issues that may cause dissimilarity in transcription. The standards were also revised

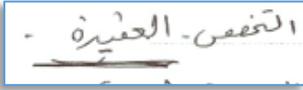
by transcribers prior to the task, and additional reviews were conducted throughout the transcription process when they come across uncertain points. The transcription standards are listed in Table 4.2.

Table 4.2: Standards followed in transcription with authentic examples from the corpus texts

Standard	Example with reference to its sheet
Any struck-out texts should be excluded.	 <p>S001_T2_M_Pre_NNAS_W_C</p>
If there is a correction above a non-struck out word, the corrected form is transcribed.	 <p>S005_T4_M_Pre_NNAS_W_H</p>
When there is a doubtful form of a character, the form closest to the correct form is transcribed. For instance, the author here wrote “هـ” which looks somewhat like “ة”. The correct form is “ة”, which has thus been transcribed.	 <p>S005_T4_M_Pre_NNAS_W_H</p>
If there is an overlap between handwritten characters, which cannot be transcribed, the closest possible form is selected. The example word here can be transcribed as “نصصهم”.	 <p>S005_T4_M_Pre_NNAS_W_H</p>
If a writer forgot to add a character’s dot(s) whether above or below, it should be transcribed as written by the learner, unless it is not possible (e.g. if there is no equivalent character on the computer). The example here is	 <p>S006_T1_M_Pre_NNAS_W_C</p>

transcribed as “استقبلنا”.	
A new line (paragraph) should be inserted only when the learner has clearly done so. Examples include if there is a clear space at the end of a line (whether there is a period or not) or if there is a clear space at the beginning of a new line with a period at the end of the previous paragraph. Other instances, such as ending a line with a period but with no clear space at the end or at the beginning of the new line, are considered as a single paragraph.	 <p>Clear space at the end of previous line</p>  <p>No clear space at the end of previous line</p> <p>S003_T1_M_Pre_NNAS_W_C</p>
Any identifying information (e.g. learner’s name, contacts, postal address, emails, etc.), which were replaced in the PDF sheet with “personal information deleted”, should be transcribed as “معلومة #” in the computerised text. Other non-personal information can be left such as class, name of school, city, country, religion, culture, etc.	 <p>S014_T4_M_Pre_NNAS_W_H</p>
Any shape, illustration, or ornamentation drawn by the learner on the sheet is excluded.	 <p>S026_T1_M_Pre_NNAS_W_C</p>
Texts with no titles are given “النص بدون عنوان” in the title field.	 <p>S030_T2_M_Pre_NNAS_W_C</p>

4 – Collecting and Managing the ALC Data

<p>Any text format is excluded such as underlined words or sentences.</p>	 <p>S009_T2_M_Pre_NNAS_W_C</p>
<p>Unknown words or phrases are replaced with 'unknown word', or 'unknown phrase'. The example here is transcribed as</p> <p>”الحافلة في #كلمة غير معروفة##، وصلنا“</p>	 <p>S015_T1_M_Pre_NNAS_W_C</p>

All identifying information was removed from texts before they were transcribed and added to the database. In addition, the transcription assistants had access only to the hand-written sheets and were not allowed to access the learners' profiles.

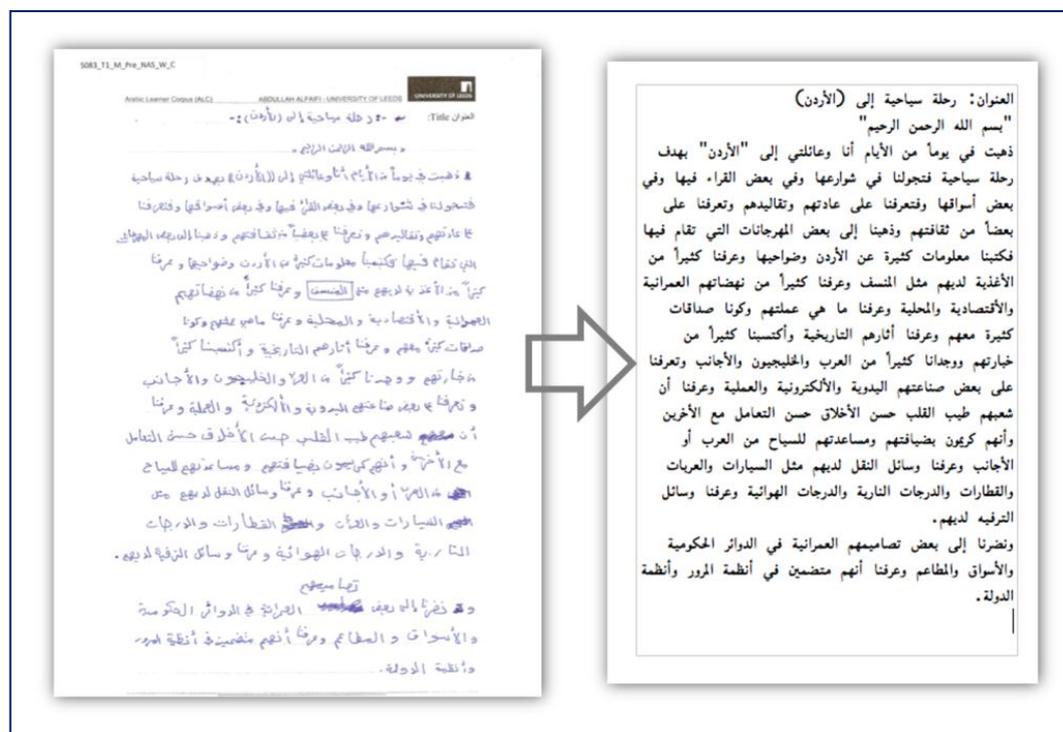


Figure 4.3: Example of a text with its transcription

4.4.2 Consistency of Hand-Written Data

In order to ensure consistency in transcribing version 1 of the ALC, the researcher and both assistants discussed the transcription standards before transcribing one text (*S011_T1_M_Pre_NNAS_W_C*). Then the consistency was measured between each pair of transcribers by dividing *the number of agreements on the total number of words in the text (120)*. This equation yielded a percentage from which the average was extracted for all pairs. The result showed an average of 93%, as illustrated in Table 4.3.

Table 4.3: Consistency between transcribers of ALC v1

	C1 & C2	C1 & R*	C2 & R
No. of similarities (from 120)	110	114	109
Percentage	92%	95%	91%
Average	93%		

*R = the researcher

After discussing the differences, this consistency measurement was performed again on a different text (*S009_T1_M_Pre_NNAS_W_C*). The result revealed an improvement by 5%, as Table 4.4 shows.

Table 4.4: Second test of consistency between transcribers of ALC v1

	C1 & C2	C1 & R	C2 & R
No. of similarities (from 132)	128	129	131
Percentage	97%	98%	99%
Average	98%		

A final test was conducted between C2 and the researcher (on the text *S003_T3_M_Pre_NNAS_W_H*) after assistant C1 withdrew. The consistency in this test was still at 98% (Table 4.5).

Table 4.5: Final test of consistency in ALC v1

	C2 & R
No. of similarities (from 104)	102
Percentage	98%

Four transcribers participated in version 2 of the ALC. The researcher was joined by three volunteering colleagues: C2, who participated in transcribing version 1 of the corpus, and C3 and C4 who, like C2 work as teachers of Arabic to NNS learners at Al-Imam University. After discussing the transcription standards, the researcher and assistants transcribed the text *S575_T1_M_Pre_NAS_W_C* (244 words). Then the consistency was measured between each pair of transcribers, from which the average was extracted. The result showed an average of 95%, as illustrated in Table 4.6.

Table 4.6: Consistency between transcribers of ALC v2

	C2 & C3	C2 & C4	C2 & R	C3 & C4	C3 & R	C4 & R
No. of similarities (from 224)	222	207	215	206	220	202
Percentage	99%	92%	96%	92%	98%	90%
Average	95%					

The differences were discussed, and the consistency measurement was performed again on the text *S579_T1_M_Pre_NAS_W_C* (354 words). The result revealed an improvement by 2% as Table 4.7 shows.

Table 4.7: Second test of consistency between transcribers of ALC v2

	C2 & C3	C2 & C4	C2 & R	C3 & C4	C3 & R	C4 & R
No. of similarities (from 354)	346	346	341	347	350	340
Percentage	98%	98%	96%	98%	99%	96%
Average	97%					

A final test was performed on the text *S656_T1_F_Uni_NAS_W_H* (377 words). The consistency in this test was improved by 1%, which resulted in an average of 98% agreement between the transcribers (Table 4.8). This result is the same as the final result of the consistency measurement in ALC v1.

Table 4.8: Final test of consistency in ALC v2

	C2 & C3	C2 & C4	C2 & R	C3 & C4	C3 & R	C4 & R
No. of similarities (from 377)	372	369	362	372	370	373
Percentage	99%	98%	96%	99%	98%	99%
Average	98%					

4.4.3 Transcribing Spoken Data

The Quick Rich Transcription Specification for Arabic Broadcast Data (Linguistic Data Consortium, 2008) was used to transcribe audio recordings. Aspects marked up in this process include, for example, punctuation, filled pauses and hesitation sounds, partial words, and mispronounced words. Table 4.9 shows examples of those aspects marked up.

Table 4.9: Aspects that are marked up in audio recording transcriptions

Examples from the ALC + text code	
Punctuation	
<i>Period (end-of-sentence mark-up for statement)</i>	
لدرجة أن يوجعني رأسي من كثر ما أفكر بها.	S942_T1_M_Uni_NAS_S_C
الحمد لله أفهم بشكل جميل.	S938_T2_F_Uni_NNAS_S_C
<i>Question mark (end-of-sentence mark-up for question)</i>	
لماذا اخترت هذا التخصص؟	S940_T1_M_Pre_NNAS_S_C
هل أفكر فيها؟	S942_T1_M_Uni_NAS_S_C
<i>Double dash (end-of-sentence mark-up for incomplete)</i>	
فيها سعوديين وفيها بحرينيين وفيها إمارات وفيها--	S935_T1_F_Uni_NAS_S_C
يعني ما حسيت إني حبل ابتعدت عن السعودية	

لأن لأني كنت أتمنى --	S937_T1_F_Uni_NNAS_S_C
أدعو الله دائماً أن أذهب في الحج	
<i>Comma (sentence-internal, used to aid readability)</i>	
هناك طالبات كثيرات يقولون إن هناك صعوبة، لم أجد	S939_T1_F_Uni_NNAS_S_C
صعوبة أبداً	
رحلتنا عن الطريق جزر القمر إلى صنعاء، وعندما	S936_T1_F_Uni_NNAS_S_C
وصلنا إلى صنعاء	
Filled pauses and hesitation sounds	
م (M sound)	S939_T1_F_Uni_NNAS_S_C
ي (E sound)	S940_T1_M_Pre_NNAS_S_C
Partial words (- dash)	
سنكون بإذن الله أ- أي أتطور بلغتي العربية	S938_T2_F_Uni_NNAS_S_C
وعندما ي- يأتي الأمر التي الذي يتعلق بالشرعية	S941_T1_M_Pre_NNAS_S_C
Mispronounced words (+ plus sign)	
من+الحرم إلى الفندق	S937_T1_F_Uni_NNAS_S_C
لكن الحمد لله مع رؤية كعبة والطواف الحمد لله ذهب كل	S929_T1_F_Pre_NNAS_S_C
+المشقة	

4.4.4 Consistency of Spoken Data

Similarly to the hand-written texts, the researcher and one of the assistants (C2) transcribed all audio recordings into the database. All identifying information was replaced with a beep sound in the audio recordings, and with #معلومة شخصية محذوفة# ‘personal information deleted’ in the transcriptions before they were added to the database.

The consistency in transcriptions was measured using the same method as that employed for the hand-written texts. Both the researcher and C2 transcribed the text *S939_T1_F_Uni_NNAS_S_C* (206 words) which yielded a percentage of 88%. In the second test, the text *S930_T1_F_Pre_NNAS_S_C* (219 words) was transcribed which showed a slightly higher result (90%). The consistency in transcribing the third text, *S928_T2_F_Pre_NNAS_S_C* (301 words), was improved by 4%, resulting

in a final consistency rate of 94% between transcribers of spoken materials (Table 4.10). The fact that, unlike written data, spoken data has no form may have added more difficulty to the transcription process and consequently reflected on the final result of consistency between transcribers, which was less than what was achieved in transcribing the written data.

Table 4.10: Consistency between transcribers of spoken materials in ALC v2

		C2 & R
Test 1	No. of similarities in first test (from 206)	182
	Percentage	88%
Test 2	No. of similarities in second test (from 219)	198
	Percentage	90%
Test 3	No. of similarities in third test (from 301)	282
	Percentage	94%

4.5 ALC Database

Corpora are often archived in various file formats (e.g. TXT, PDF, XML, DOC), and “XML is usually considered to be a more appropriate file format for long-term preservation, because it is an open international standard defined by the World Wide Web Consortium (W3C)” (Wynne, 2005). Other corpora, however, use databases to archive their content. A relational database provides multi-faceted benefits when storing, managing, and searching corpora (Davies, 2005). One of the benefits of this method is to automate the generation of the corpus content in different file formats to match the purposes of the target users. The International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010), for instance, uses a database which provides users with a built-in concordancer. Other corpora use databases to manage multipurpose searches of their large content, such as the Corpus del Español (Davies, 2005) and KACST Arabic Corpus (Althubaity, 2014). Such databases enable users to analyse the corpus using concordances, frequency words lists, and frequency of n-grams, in addition to allowing a large amount of annotation to be added and utilised in a corpus (Davies, 2005).

Given the fact that the corpus is not very large (it includes 1585 materials), a Microsoft Access database was a good option in this stage, as it can be designed quickly and managed easily for such size of data. The database was created by the researcher to store and manage the content of the ALC. The corpus data are stored in a main table where each record (row) represents the data of a single text with its metadata. Further tables for entities such as nationalities, mother tongues, and educational institutions were created and linked to the main table to easily manage those entities separately. Figure 4.4 shows the database with the entity-relationship diagram, the left and right sides present the English and Arabic translations of the same information.

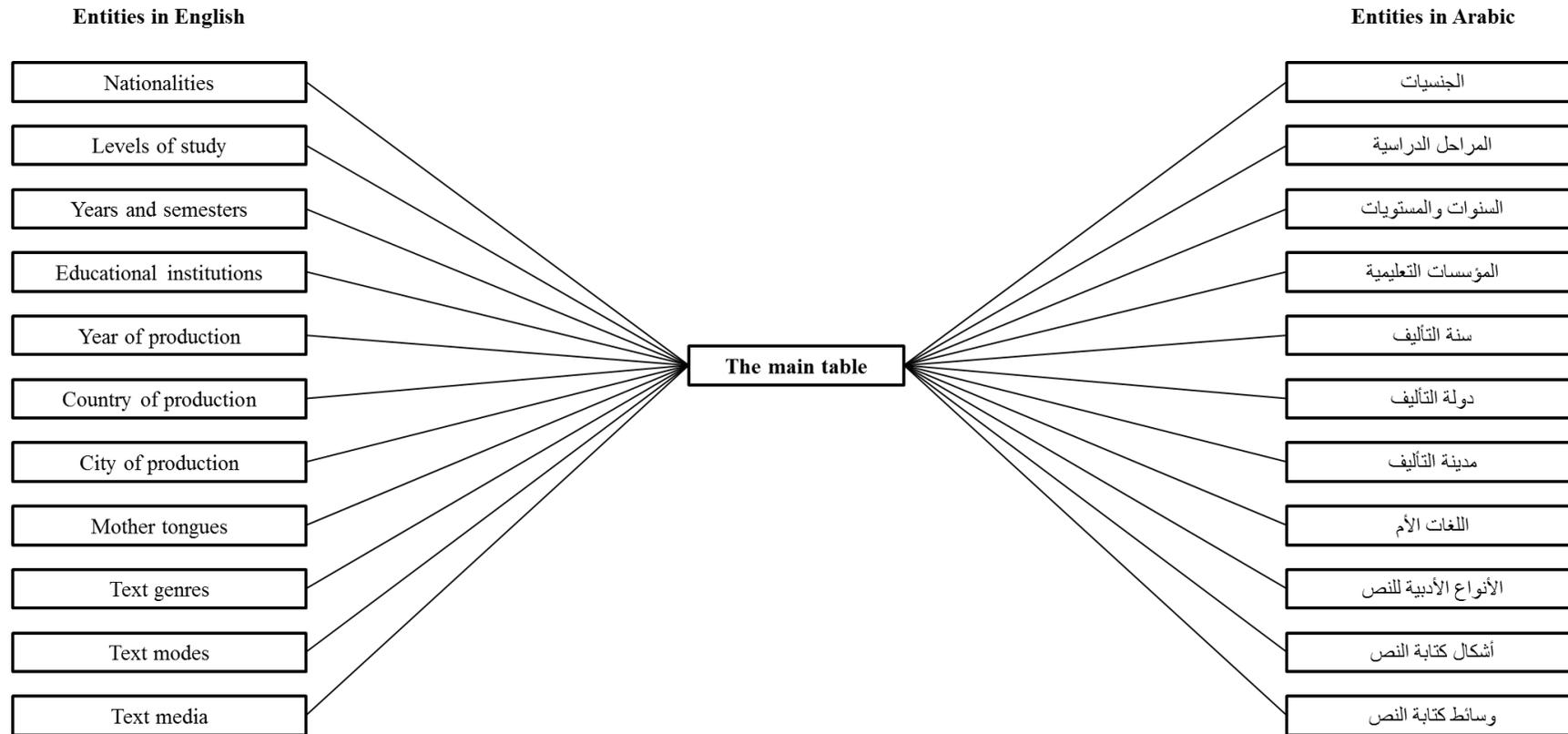


Figure 4.4: The ALC database with the entity-relationship diagram

4.5.1 Data Storing

Data gathered from the learners are in three different forms: written on a computer, written by hand, and audio recordings. The first type was directly stored into the database, whereas the hand-written texts and audio recordings were transcribed into electronic texts before being stored in the database. The metadata were entered to the database manually by the researcher and double-checked by him and an assistant colleague to ensure that nothing was missed or incorrect.

Data entered into the database includes the raw text, its title, its identification code, and 26 elements representing the metadata of the text. Some of these elements are numerical and others are textual; the textual elements are recorded in both English and Arabic.

The screenshot shows a web-based data entry interface. At the top, there are three tabs: 'Text', 'Learner Metadata', and 'Text Metadata'. Below the tabs, there is a red instruction: 'انقر نقرًا مزدوجًا على الأيقونة لفتح ملف PDF الأصلي Double click on the icon to open the original PDF file'. To the right of this instruction is an Adobe Acrobat icon. Below the instruction, there are two input fields: 'Text code_رمز النص' with the value 'S001_T1_M_Pre_NNS_W_C' and 'Text title_عنوان النص' with the value 'الرحلة إلى القرية لزيارة نوي القري'. Below these fields is a section labeled 'Raw text_النص دون ترميز' containing a paragraph of Arabic text:

اعتدت الذهاب إلى قريتي في الإجازات الصيفية الموجودة في غرب بلدي المسمى ببوركينا فاسو. تسمى قريتي ساغابتفا، قمت برحلة إليها في الإجازة الصيفية الماضية وكانت من أمتع الرحلات التي رحلت إليها وأفضلها، فبعد أن قررت الرحيل إليها، اتصلت بمن فيها من سادة القوم وكبارهم، فسروا وفرحوا؛ لما سمعوا أنني ما يسر أفندتهم ولما طال الالتقاء بيننا لسبب طلب العلم والسير في الأرض فازددت همّة في المسير لما سمعت إقداماً وحباً لأهل القرية تجاهي فخرجت في طريقي الخامس عشرة من رمضان متجهاً إلى القرية، فوصلت إليها في يومه، فأكرموني وطبخوا لي طعاماً شهياً طأقت إليه قلبي قبل ذوقي ثم قدموني إماماً، فصليت بهم الظهر وقلت بعدها: فلما استيقظت للعصر وصليت جاؤوني سائلين عن أهل المدينة ومن تركتهم في المدينة ثم سألوني عن السعودية وأهلها وكيف وجدتها ومن فيها، فأجبتهم بما تيسر لي أن أجيبه. ثم عقدوا لي مجلساً أدرّس فيه القرآن وأعطوا فيه الناس بما تيسر لي وما تعلمته من العقيدة الصحيحة من أساتذتي الفضلاء فيقبت على هذا أكثر من حين لحثي عدت إلى المدينة.

Figure 4.5: Example of a text stored in the ALC database

The screenshot displays a web interface for entering learner metadata. At the top, there are tabs for 'Text', 'Learner Metadata', and 'Text Metadata'. Below these, a text input field contains the code 'S001_T1_M_Pre_NNS_W_C'. The main section is titled 'Learner Profile' and contains several fields:

Age_العمر	20
Gender_الجنس	ذكر / Male
Nationality_الجنسية	بوركينا فاسو / Burkina Faso
Mother tongue_اللغة الأم	المورية / Moore
Nativeness_أصل اللغة	نلطق بخير العربية / NNAS
No of languages spoken_عدد اللغات التي يتحدثها	4
No of years learning Arabic_عدد سنوات تعلم اللغة العربية	14
No of years in Arab countries_عدد السنوات في بلدان عربية	3
General level_المستوى العام	ما قبل الجامعة / Pre-university
Level of study_المرحلة الدراسية	برنامج الدبلوم / Diploma course
Year/Semester_السنة أو المستوى	المستوى الثاني / Second semester
Educational institution_المؤسسة التعليمية	معهد اللغة بجامعة الإمام / Arabic Inst. at Imam Uni

Figure 4.6: Example of metadata stored in the ALC database

4.5.2 File Generation Function

A file-generation function was built as a part of the ALC database to generate the corpus files into five formats using a control form created for this purpose. The file generation process starts with retrieving all fields of one record, which represents a text with its metadata, from the database. Then the function constructs five formats from this record, including adding the appropriate tags to the XML format. Those five formats are: (i) text format with no metadata, (ii) text format with Arabic metadata, (ii) text format with English metadata, (iv) XML format with Arabic metadata, and (v) XML format with English metadata; see examples of these five files in Appendix A. In the second step, the database ensures that directories selected by the user, which will be used to save the generated files, exist; otherwise, it creates them. Finally, based on the five formats created in the first step, the corpus files can be generated in one of three ways (Figure 4.7): one file for the entire corpus, separate files (one file for each text), or separate, classified files based on predetermined features (Table 4.11) in which each group of texts is stored in a classifying folder. Producing such classified files simplifies searching and analysing the corpus contents, and more features can be added in the future.

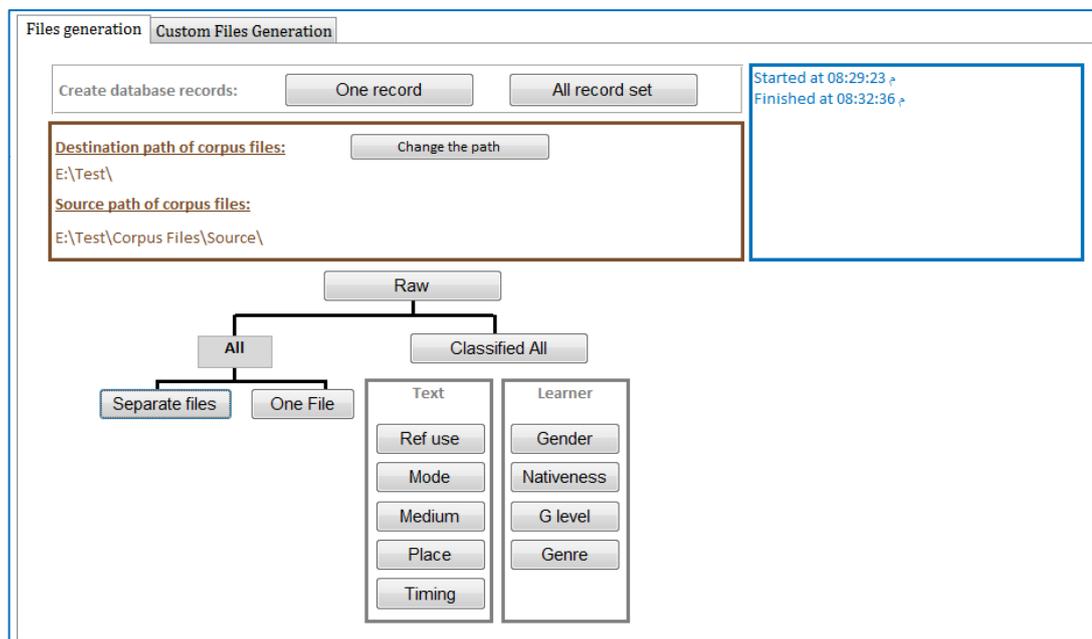


Figure 4.7: Three methods for generating files for the entire ALC

Table 4.11: Classification features of the corpus files

Based on	Feature	Classification
Learners	Nativeness	Native speakers vs. Non-native speakers
	Gender	Males vs. Females
	General level	Pre-university vs. University
Texts	Mode	Written vs. Spoken
	Medium	By hand vs. On computer
	Genre	Narrative vs. Discussion
	Place	In class vs. At home
	References	Ref. used vs. Ref. unused
	Timing	Timed vs. Untimed

The previous function generates data for the entire corpus. However, an additional function was developed to generate custom files based on specific conditions (Figure 4.8), for instance those texts written by hand, in class, by female learners, in Riyadh. Figure 4.9 illustrates the processes of the file-generation function.

The screenshot shows a web-based interface for generating custom files from a database. It is titled 'Files generation' and 'Custom Files Generation'. The interface is divided into two main columns of filters. The left column includes: Age (From 16 To 42), Gender (Female), Nationality (All), Mother Tongue (All), Nativeness (All), General Level (All), Level of Study (All), Year/Semester (All), Educational Institution (All), No of languages spoken (From 1 To 10), No of years learning Arabic (From 0 To 19), and No of years living in Arab countries (From 0 To 21). The right column includes: Genre (All), Mode (All), Medium (Written by hand), Place (In class), City (Riyadh), Country (All), Year (All), Timing (All), Ref. Use (All), Grammar Book Use (All), Mono Dic Use (All), Bi Doc Use (All), Other Ref. Use (All), and Length (From 0 To 10000). On the right side, there is a 'Number of files' button, a 'Calculate the number of files that can be generated' button, a 'Done.' status, a message '63 of 1585 records were found', and a 'Files Generate' button.

Figure 4.8: Custom file generation in the ALC database

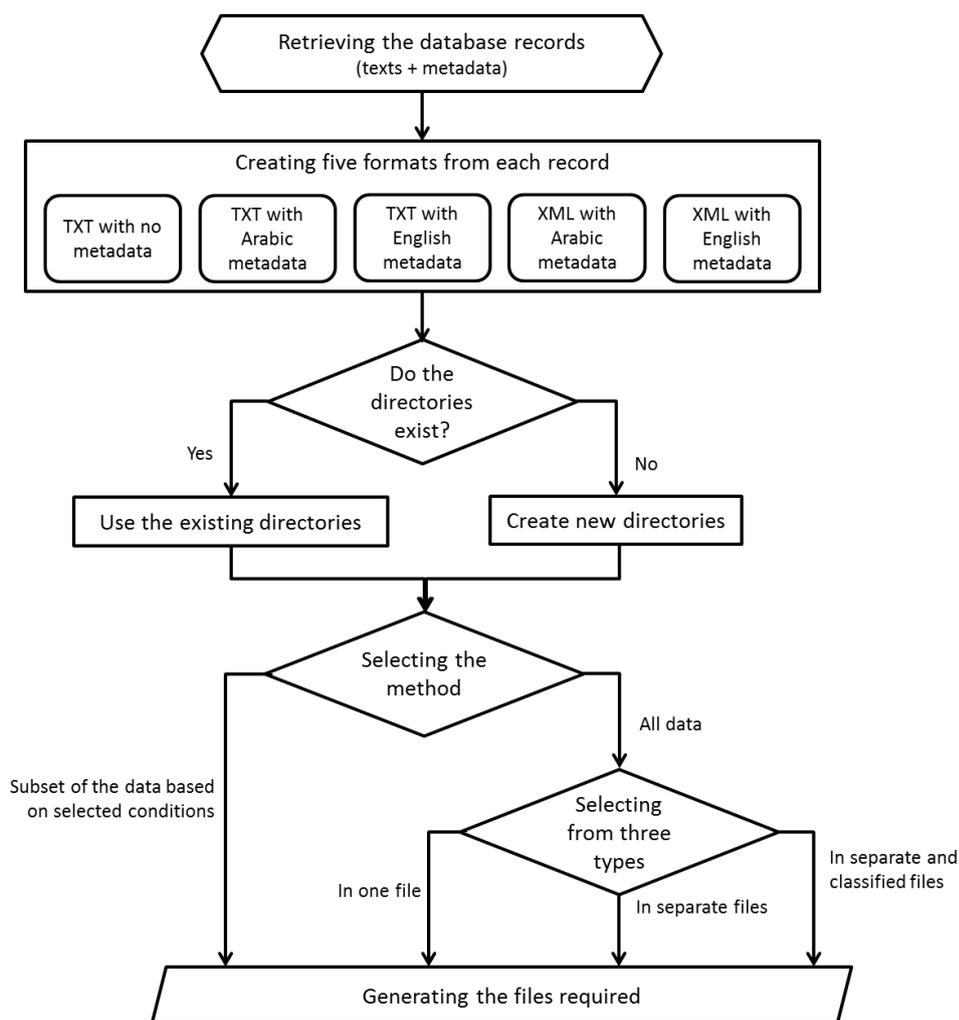


Figure 4.9: Processes of the files generation function

4.6 File Naming

All files were named following a method that indicates the basic characteristics of the text and its author. A name consists of seven parts separated by the underscore mark (_). The seven parts are:

1. The student identifier number (S102);
2. The number of the text written by the same student ID number (e.g. the label “S012_T1” indicates the first text written by student number 12);
3. The learner’s gender: male (M) or female (F);
4. The learner’s level of study: pre-university (Pre) or university (Uni);

5. The writer's nativeness: native Arabic speaker (NAS) or non-native Arabic speaker (NNAS);
6. The mode of the text: written (W) or spoken (S); and
7. The place of text production: in class (C) or at home (H).

Table 4.12 shows an example of a corpus file including the seven name sections with their description.

Table 4.12: Example of corpus files naming method

File name	102_	T1_	M_	Pre_	NNAS_	W_	C
Description	Student number	Text number	Gender (Male)	Level of study (Pre-university)	Nativeness (Non-native Arabic speaker)	Written text	Text produced in class

4.7 Conclusion

This chapter describes how the ALC data was collected and managed. It starts with a description of the questionnaire designed to collect the corpus data. The corpus materials were collected using guidelines that were created to clarify the steps and procedures that the researcher followed in each session of data collection. The chapter also illustrates the questionnaire that was adapted from the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010) and used to collect the corpus metadata (26 pieces of information about the learners and their productions).

As the ALC contains hand-written texts and spoken materials, further work was required to convert them into an electronic form. Specific standards were created for

transcribing the hand-written data, while the Quick Rich Transcription Specification for Arabic Broadcast Data was used to convert the spoken recordings. The chapter describes the method used to measure the consistency between transcribers of both data modes and discusses the results.

The chapter also describes the database developed to store and manage the ALC data. The database was also designed to automatically generate the corpus files in different formats (TXT and XML). The chapter illustrates the steps of the file-generation function which produces the entire corpus in three ways: the entire corpus in one file, each text in a separate file, or separate and classified files based on particular features. A further function was also developed for generating custom groups of files (sub-corpora) based on specific conditions (metadata elements). The chapter concludes by illustrating the ALC file-naming method, which indicates the basic characteristics of the text and its author.

Part III

ALC Tools

Summary of Part III

This part describes two tools that were created as part of the ALC project and the error annotation system. The first tool is the Computer-aided Error annotation Tool for Arabic (CETAr), which was developed mainly to assist in annotating Arabic errors consistently in learner corpora. The creation of this tool involved the development of the Error Tagset of Arabic (ETAr) and the Error Tagging Manual for Arabic (ETMAr), which are also described in this part.

The second tool is the free-access, web-based concordance, the ALC Search Tool. It provides users with two basic functions: searching the corpus or any subset of its data based on a number of determinants, and downloading the corpus files or a subset of its files in different formats (TXT, XML, PDF, and MP3) based on the same determinants.

5 Computer-Aided Error Annotation Tool for Arabic

Chapter Summary

This chapter highlights the need to develop a new tool for annotating errors of Arabic with an appropriate taxonomy of Arabic errors. The tool developed for this project, the CETAr, was designed based on the annotation standards defined for the ALC project to standardise the format of the annotated files. The CETAr includes a number of features to facilitate the annotation process such as text tokenisation, manual tagging, smart-selection, and auto tagging.

As a basic part of the CETAr and the ALC project in general, an error taxonomy was developed to be used for annotating errors in Arabic. The ETAr contains in the most recent version (v3) 29 error types divided into 5 broad categories. Seven annotators (including the researcher) and two evaluators performed three experiments on this tagset to measure several factors: (i) the extent to which the ETAr can be understood and compared against another tagset, (ii) the inter-annotator agreement, (iii) the value of training the annotators, (iv) the distribution of the ETAr tags on a sample of the ALC, and (v) the value of using the ETMar. The ETMar was developed specifically to serve two main functions: first, to explain the errors in the ETAr with examples and, second, to provide users with rules to follow for selecting the appropriate tags in error annotation.

5.1 Introduction

The benefits of learner error annotation are multi-faceted and extend to fields such as contrastive interlanguage analysis, learner dictionary making, second language acquisition, and designing pedagogical materials. Contrastive interlanguage analysis is still one of the most frequently used approaches for analysing errors in a learner corpus, as it enables researchers to observe a wide range of instances of underuse, overuse, and misuse of various aspects of the learner language at different levels: lexis, discourse, and syntax (Granger, 2003b). Analysing errors also enables researchers and educators to understand the interlanguage errors caused by L1 transfer, learning strategies, and overgeneralisation of L1 rules. Learner corpora are used to compile or improve learner dictionary contents, particularly by identifying the most common errors learners make and then providing dictionary users with more details at the end of relevant entries. These errors are indicated in words, phrases, or language structures, along with the ways in which a word or an expression can be used correctly and incorrectly (Granger, 2003b; Nesselhauf, 2004).

Error-annotated learner corpora are useful resources to measure the extent to which learners can improve their performance in various aspects of the target language (Buttery & Caines, 2012; Nesselhauf, 2004). Compilers of longitudinal learner corpora usually include this goal in their aims. Examples of these include the LONGDALE project: LONGitudinal DAtabase of Learner English (Meunier *et al.*, 2010), Barcelona Age Factor (Diez-Bedmar, 2009), and the ASU corpus (Hammarberg, 2010). Finally, analysing learners' errors may be beneficial for pedagogical purposes such as instructional teaching material development. It can, for instance, help in developing materials that are more appropriate to learners' proficiency levels and in line with their linguistic strengths and weaknesses.

As seen in the literature review, learner corpora tend to be tagged with one or more types of annotation. Linguistic errors, including describing, classifying, or correcting them, have received the most attention among other types of annotation such as PoS. This substantial use of error annotation assists in achieving one of the main purposes in learner corpora, error analysis. Granger (2008) believes that more research should be devoted to the error annotation of a learner corpus. Thus, this project involved the development of a basic tool (the Computer-Aided Error Annotation Tool for Arabic [CETAr]) with an error tagset (the Error Tagset of Arabic [ETAr]) and its manual

(the Error Tagging Manual for Arabic [ETMAr]) to annotate errors in Arabic texts and Arabic learner corpora in particular. This chapter is devoted to a discussion of this tool and tagset.

5.2 Background

This section explores tools used for error annotation and their suitability for Arabic script. It also gives an overview of the existing tagsets and guidelines for Arabic error annotation and why it is important to create a new tool and tagset for Arabic error annotation.

5.2.1 Annotation Tools

Researchers have developed several tools to annotate texts, not just for errors but also for PoS, lemma, dependency, and other matters. However, these tools encounter some problems in handling Arabic. WebAnno2 (Eckart de Castilho *et al.*, 2014; Yimam *et al.*, 2014; Yimam *et al.*, 2013), for example, shows Arabic words with many cases of overlapping words (Figure 5.1); in these overlapped cases, selecting tokens accurately is difficult. Another problem in this tool is that, when a token is annotated, the tag appears over another token, which seems to be an error in indexing the token positions. This latter problem happens also when using GATE (Cunningham *et al.*, 2011; Cunningham *et al.*, 2013), which is an open-source tool for different functions such as web mining, information extraction, language processing, and semantic annotation. Annotators may face another problem in the high level of training required to use GATE to annotate corpus errors.

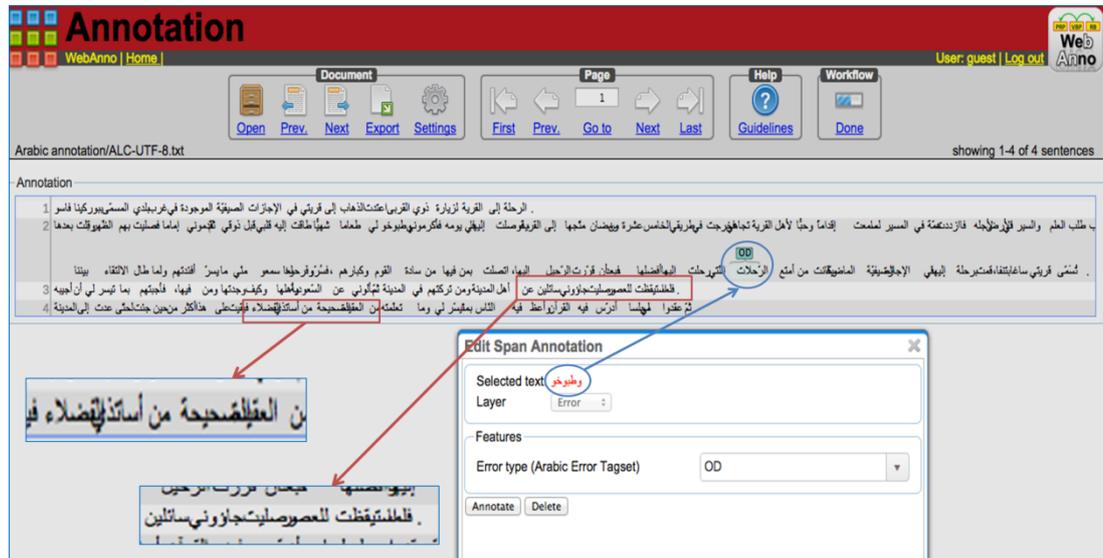


Figure 5.1: Example of annotating Arabic text using the WebAnno2 tool

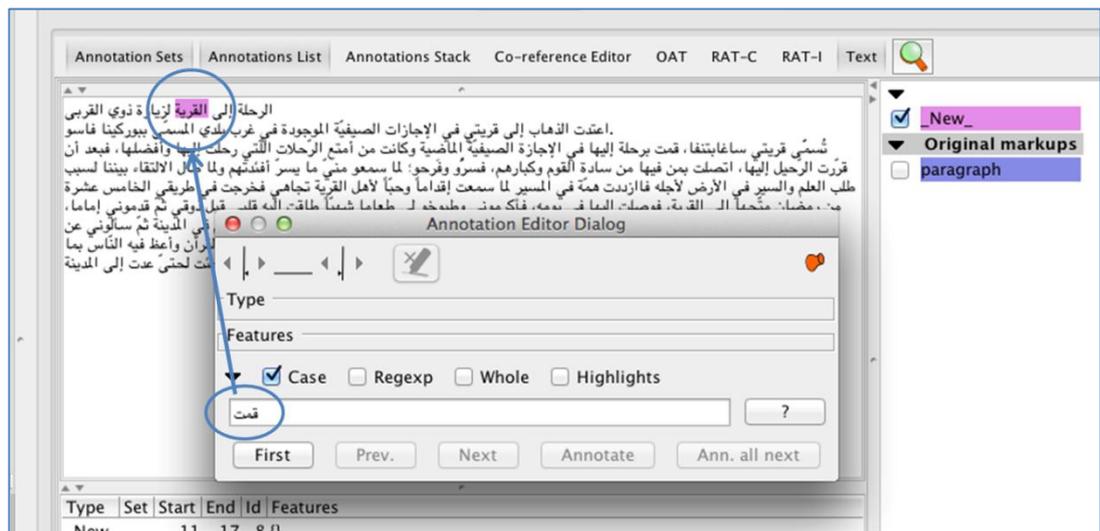


Figure 5.2: Example of annotating Arabic text using GATE

The Content Annotation Tool (Bartalesi Lenzi *et al.*, 2012) is another example of the existing annotation tools. However, the main problem in using this tool is that words are shown in the opposite direction, left-to-right, while Arabic is a right-to-left written language. Additionally, the tool seems to have a problem with showing the annotation boundaries of Arabic tokens, as it leaves off part of the highlighting (Figure 5.3). TextAE editor (Kim *et al.*, 2013) is an open-source web application for annotation. However, the main problem the researcher faced with this tool was that, after several attempts to open an Arabic text in both UTF-8 and UTF-16 formats, the text area within the tool was empty, leaving the researcher unable to see the file contents.

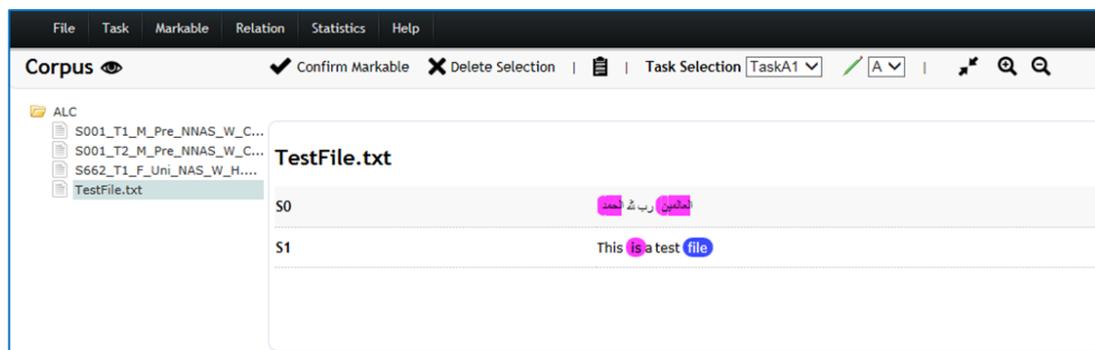


Figure 5.3: Example of annotating Arabic text using the Content Annotation Tool

A famous tool created particularly to annotate learner corpora for errors is the Université catholique de Louvain Error Editor software (Hutchinson, 1996) which uses a taxonomy of English errors tagset (Dagneaux *et al.*, 1996).

“The English ‘error toolkit’ contains a comprehensive error tagging manual (Dagneaux et al. 2008) which explains each of the 50-plus error tags, and the Université catholique de Louvain Error Editor (UCLEE) software which helps with the insertion of the error tags and the corrections in the data” (Centre for English Corpus Linguistics, 2010).

The Université catholique de Louvain Error Editor has been invaluable to English corpora, but no Arabic counterpart exists; thus, there is a need to develop a new tool for Arabic with an appropriate taxonomy of Arabic errors.

5.2.2 Error Annotation Tagsets and Manuals

Learner corpora may include errors made by the language learners. Given the fact that “current spelling and grammar-checking programs are not capable of detecting, let alone correcting, the majority of these errors, error annotation is the only solution for the time being” (Granger, 2003b: 542). For this reason, researchers have created several error annotation tagsets and manuals such as the Error Tagging Manual (Dagneaux *et al.*, 1996), Cambridge Error Coding (Nicholls, 2003), the French Interlanguage Database (FRIDA) Error Tagset (Granger, 2003a), the Japanese Learner English (JLE) Corpus Error Tagset (Izumi *et al.*, 2005), and the Learner

Corpus Annotation Manual of the Learner Corpus Development Corpus (Sigott & Dobrić, 2014).

With respect to Arabic, researchers have created the Arabic Interlanguage Database (ARIDA) tagset (Abuhakema *et al.*, 2008, 2009) and the Qatar Arabic Language Bank (QALB) Guidelines (Wajdi Zaghouni *et al.*, 2014). The ARIDA is the sole error tagset specifically created for Arabic learner corpora, and it is based on the FRIDA Error Tagset (Granger, 2003a). This adaptation from a French tagset, however, rendered some classification inconsistency with traditional Arabic linguistics dominating the curriculums of teaching Arabic in Saudi Arabia. For example, in traditional Arabic, grammatical and syntactic errors are combined under one category called either grammar or syntax; in the ARIDA tagset, these are two different error categories. In addition, a number of the categories in the FRIDA-derived tagset have a literal translation into Arabic with no clarification of what they linguistically or practically mean, which renders them vague. Examples include *Adjective Complementation* “متمة الصفة”, *Noun Complementation* “متمة الاسم”, and *Verb Complementation* “متمة الفعل”. Further, most of the morphological categories describe the error place and not the type. The sole exception is *Inflection Confusion* “الخلط في التصريف”, which describes an essential morphological error in Arabic learner production. In the *Form/Spelling* category, Abuhakema lists important error types, like *Hamza* “الهمزة” (ء) and *Tanwīn* “التنوين” (َِِّّ)¹, but neglects some others, like *tā’ mutatarriḥa* “التاء المتطرفة” (ت،ة،ة)², *’alif mutatarriḥa* “الألف المتطرفة” (ا،ى،ى)³, and *’alif fāriqa* “الألف الفارقة” (وا)⁴. Additionally, no manual has been published to explain how Arabic errors should be annotated by this tagset. It seems that the FRIDA manual is expected to be used, but doing so may result in Arab users facing challenges in applying the guidelines to Arabic when it was originally designed for French.

¹ *Tanwīn* is an extra “n” sound at the end of a word, but not an original character. It is written as double diacritic marks and pronounced only when continuing to the next word, however it is omitted when stopping.

² *Tā’ Mutatarriḥa* comes at the end of the words. It has two forms, opened “*Maftūḥa*” (ت), or closed “*Marbūṭa*” (ة).

³ *’alif Mutatarriḥa* comes at the end of the word in two forms: similar to *’alif* (ا) which is called *Mamdūda* (ا), and similar to *Yā’* (ي) that is called *Maqṣūra* (ى).

⁴ *’alif Fāriqa* is an *’alif* (ا) character that is added after *wāw ’alḡamā’a*, the plural pronoun (وا), to indicate that this *wāw* is not a part of the word root but is *wāw ’alḡamā’a*, the plural pronoun (وا).

In contrast to this adapted tagset, the QALB Guidelines (Wajdi Zaghouani *et al.*, 2014) form an error annotation manual specifically created for Arabic text corrections in the QALB project. The guidelines classify errors into six categories: spelling, punctuation, lexical, morphology, syntax, and dialect; however, the manual does not contain a tagset for annotating those Arabic errors. It includes information about how to use the project annotation tool and details about possible errors with examples and rules of the Arabic language (spelling, punctuation, etc.). Thus, the inadequacies of these two tools make it necessary to develop an error tagset complete with an error tagging manual.

This overview of problems in the tools and tagset existing for Arabic error annotation highlights the importance of creating a new tool and tagset with consistent guidelines that together can be useful resources for annotating Arabic learner corpora for errors.

5.3 The Computer-Aided Error Annotation Tool for Arabic (CETAr)

The problems with handling Arabic using the existing tools of annotation indicate a need to develop a new tagging tool for errors in Arabic. This tool was designed based on the annotation standards, i.e. requirements the researcher specified in order to standardise the format of the annotation files. The main purpose of this tool is to facilitate the manual tagging by enabling annotators to use the ETAr on Arabic texts by assigning a tag indicating the error type to each linguistic error. Further purposes include increasing the consistency in error annotation and automating a part of the tagging process. The following sections present the annotation standards before describing and evaluating the features of the CETAr.

5.3.1 Annotation Standards

The annotation standards are a set of requirements developed in order to standardise the format of the annotation files. The steps of this process are described below.

1. A text is tokenised as a pre-annotation process. This segments each token and locates it in a separate line. For instance, the phrase “بعدها أتيت هنا أقمت أول رحلة،” (*After I came here, I started the first journey*), which is taken from the text *S938_T1_F_Uni_NNAS_S_C*, is tokenised as follows:

بعدهما
أتيت
هنا
أقمت
أول
رحلة
,

Figure 5.4: Example of text tokenisation

2. Each token with an error requires three annotations: the error tag describing its type, the error form, and the suggested correction.
3. The token and annotations are separated from each other by a tab space.

والمروءة OT والمروه والمروءة

Figure 5.5: Example of tokens separated from each other by a tab space¹

4. More than one tag can be assigned to a token with a plus sign between the tags (e.g. OH+OM).

اعطيه OH+OM اعطيه أعطيته

Figure 5.6: Example of error annotated with two error types²

5. Each tag is assigned to only one token at a time. If two consecutive tokens have the same error, each token is tagged separately using the same tag. Errors covering multiple words, phrases, or sentences, such as style errors, are excluded in this stage of the project to avoid problems of overlapping mark-ups, particularly in XML file structure. The next stages will include conducting more research about this issue to select the most appropriate method for marking up the overlap cases, and then this method will be applied to the ALC data.

Following those standards helped the researcher standardise the format of the output files, and enabled the generation of two types of files structure, Inline Annotation and Stand-off Annotation by Tokens in order to provide the corpus users with various options. The two file structures are based on the literature review as follows:

¹ The OT tag indicates the error: *Redundant character(s)*.

² The OH and OM tags indicate the errors: *Hamza* and *Missing character(s)*.

“The phrase ‘inline annotation’ refers to the annotation XML tags being present in the text that is being annotated, and physically surrounding the extent that the tag refers to” (Pustejovsky & Stubbs, 2013: 94).

“One method that is sometimes used for stand-off annotation is tokenizing (i.e., separating) the text input and giving each token a number. The tokenization process is usually based on whitespace and punctuation” (Pustejovsky & Stubbs, 2013: 96).

Those two methods were adapted in two file formats, plain text and XML. This resulted in four options to the corpus users: (i) plain text with inline annotation (Figure 5.7), (ii) plain text with stand-off annotation by tokens (Figure 5.8), (iii) XML with inline annotation (Figure 5.9), and (iv) XML with stand-off annotation by tokens (Figure 5.10).

Text ID: S037_T1_M_Pre_NNAS_W_C

رحلتي إلى بلدي خلال إجازة الحج
لما وصلت إلى السعودية في هذا الفصل كنت عازماً على<OR-على/على> تأجيل الفصل ورجوعي إلى بلدي<PM-بلدي/
بلدي>؛ لأن زوجتي كان<XG-كان/كانت> على وشك الولادة، ولم يكن وعها<OR-وعها/معها> أحد، لكنني فوجئت بخبر
محزن وهو أنه لا يمكن تأجيل الدبلون،<OR-الدبلون،/الدبلون،> إما أن أمسحه<SW-أمسحه/ألغيه> كله فأرجع في
السنة القادمة للكلية، وإما أن أسافر وأرجع سريعاً. وأيضاً أخبرت بأن زوجتي مقبولة بجامعة الأميرة نورة
بداية من هذا الفصل، ففرحت وحرزنت في نفس الوقت. سبب الفرح كان أنه سوف يتحقق ما كنت أتمنى منذ عشر
سنوات، وذلك أن أذهب يوماً ما إلى السعودية لدراسة العلوم الإسلامية وأن أتزوج وأحضر زوجتي معي فندرس معاً،
وسبب حزني كان معرفتي بأنه من الصعب جداً أن أرجع الآن إلى بلدي وأن زوجتي ستواجه صعوبات الولادة وألمها
بنفسها ولن أكون معاً حتى أساعدها في ما أستطيع مساعدتها، وقد كنت وعدتها قبل مجيئي أنني سأرجع بسرعة وأن
لا تحزن. كنت أشعر كأنني خنتها، لكن ماذا أفعل؟ قدر الله وما شاء فعل. بعد أن تيدرن<OC-تيدرت/تيدرت> جيداً
جميع الاختيارات المتاحة بين يدي، كتبت الخطاب<XF-الخطاب/خطاباً> أطلب فيه من مجلس المعهد أن يسمحوا لي
بالسفر إلى بلدي لإحضار زوجتي، وبينت حالي بالتفصيل. ثم وافقوا على أن يسمحوا لي ذلك<XM-ذلك/بذلك> خلال
إجازة الحج، وأن يزيد<MI-يزيد/أزيد> بأسبوع قبل الإجازة ويومين بعدها، وكذلك فعلت.

Figure 5.7: Plain text with inline annotation

كنت			
عازماً			
على	OR	على	على
تأجيل			
الفصل			
ورجوعي			
إلى			
بلدي	PM	بلدي	بلدي؛
لأن			
زوجتي			
كان	XG	كان	كانت
على			
وشك			
الولادة			
,			
ولم			
يكن			
وعها	OR	وعها	معها

أحد
,
لكنني
فوجئت

Figure 5.8: Plain text with stand-off annotation by tokens

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc ID="S037_T1_M_Pre_NNAS_W_C">
  <text>
    <title>رحلتي إلى بلدي خلال إجازة الحج
    </title>
    <p id=1>
      <t n=16 ErrTag="OR" ErrForm="عأى" CorrForm="على">عأى</t>
      <t n=21 ErrTag="PM" ErrForm="بلدي" CorrForm="بلدي">بلدي</t>
      <t n=24 ErrTag="XG" ErrForm="كان" CorrForm="كانت">كان</t>
      <t n=31 ErrTag="OR" ErrForm="وعها" CorrForm="معها">وعها</t>
      <t n=43 ErrTag="OR" ErrForm="الديلون" CorrForm="الديلون">الديلون</t>
      <t n=46 ErrTag="SW" ErrForm="أمسحه" CorrForm="ألغيه">أمسحه</t>
      <t n=49 ErrTag="OR" ErrForm="فأرجع" CorrForm="فأرجع">فأرجع</t>
      <t n=52 ErrTag="OR" ErrForm="مقبولة" CorrForm="مقبولة">مقبولة</t>
      <t n=55 ErrTag="OR" ErrForm="الفرح" CorrForm="الفرح">الفرح</t>
      <t n=58 ErrTag="OR" ErrForm="السعودية" CorrForm="السعودية">السعودية</t>
      <t n=61 ErrTag="OR" ErrForm="كان" CorrForm="كانت">كان</t>
      <t n=64 ErrTag="OR" ErrForm="الولادة" CorrForm="الولادة">الولادة</t>
      <t n=67 ErrTag="OR" ErrForm="وعدتها" CorrForm="وعدتها">وعدتها</t>
    </p>
  </text>
</doc>
```

Figure 5.9: XML with inline annotation

```
<?xml version="1.0" encoding="UTF-8" ?>
<doc ID="S037_T1_M_Pre_NNAS_W_C">
  <text>
    <title>
      <t n=1>رحلتي</t>
      <t n=2>إلى</t>
      <t n=3>بلدي</t>
      <t n=4>خلال</t>
      <t n=5>إجازة</t>
      <t n=6>الحج</t>
    </title>
    <p id=1>
      <t n=7>لما</t>
      <t n=8>وصلت</t>
      <t n=9>إلى</t>
      <t n=10>السعودية</t>
      <t n=11>في</t>
      <t n=12>هذا</t>
      <t n=13>الفصل</t>
      <t n=14>كنت</t>
      <t n=15>عازما</t>
      <t n=16 ErrTag="OR" ErrForm="عأى" CorrForm="على">عأى</t>
      <t n=17>تأجيل</t>
      <t n=18>الفصل</t>
      <t n=19>ورجوعي</t>
      <t n=20>إلى</t>
      <t n=21 ErrTag="PM" ErrForm="بلدي" CorrForm="بلدي">بلدي</t>
      <t n=22>لأن</t>
      <t n=23>زوجتي</t>
      <t n=24 ErrTag="XG" ErrForm="كان" CorrForm="كانت">كان</t>
      <t n=25>على</t>
      <t n=26>وشك</t>
      <t n=27>الولادة</t>
      <t n=28>ولم</t>
      <t n=29>يكن</t>
    </p>
  </text>
</doc>
```

Figure 5.10: XML with stand-off annotation by tokens

The following model of DTD was used to validate the structure of XML files covering the metadata and inline annotation (Figure 5.11).

```

<!DOCTYPE doc [
<!ELEMENT doc (header,text)>
<!ATTLIST doc ID ID #REQUIRED >
<!ELEMENT header (learner_profile,text_profile)>
<!ELEMENT learner_profile
(age,gender,nationality,mothertongue,nativeness,No_languages_spoken,
No_years_learning_Arabic,No_years_Arabic_countries,general_level,lev
el_study,year_or_semester,educational_institution)>
<!ELEMENT age (#PCDATA)>
<!ELEMENT gender (#PCDATA)>
<!ELEMENT nationality (#PCDATA)>
<!ELEMENT mothertongue (#PCDATA)>
<!ELEMENT nativeness (#PCDATA)>
<!ELEMENT No_languages_spoken (#PCDATA)>
<!ELEMENT No_years_learning_Arabic (#PCDATA)>
<!ELEMENT No_years_Arabic_countries (#PCDATA)>
<!ELEMENT general_level (#PCDATA)>
<!ELEMENT level_study (#PCDATA)>
<!ELEMENT year_or_semester (#PCDATA)>
<!ELEMENT educational_institution (#PCDATA)>
<!ELEMENT text_profile
(genre,where,year,country,city,timed,ref_used,grammar_ref_used,mono_
dic_used,bi_dic_used,other_ref_used,mode,medium,length)>
<!ELEMENT genre (#PCDATA)>
<!ELEMENT where (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT country (#PCDATA)>
<!ELEMENT city (#PCDATA)>
<!ELEMENT timed (#PCDATA)>
<!ELEMENT ref_used (#PCDATA)>
<!ELEMENT grammar_ref_used (#PCDATA)>
<!ELEMENT mono_dic_used (#PCDATA)>
<!ELEMENT bi_dic_used (#PCDATA)>
<!ELEMENT other_ref_used (#PCDATA)>
<!ELEMENT mode (#PCDATA)>
<!ELEMENT medium (#PCDATA)>
<!ELEMENT length (#PCDATA)>
<!ELEMENT text (title,text_body)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT text_body (#PCDATA)>
]>

```

Figure 5.11: DTD model for XML files containing metadata and inline annotation

The same model of DTD but with further additions was used to validate the structure of XML files containing the metadata and stand-off annotation by tokens (Figure 5.12).

```

<!DOCTYPE doc [
<!ELEMENT doc (header?,text)>
<!ATTLIST doc ID ID #REQUIRED >
<!ELEMENT header (learner_profile,text_profile)>
<!ELEMENT learner_profile
(age,gender,nationality,mothertongue,nativeness,No_languages_spoken,
No_years_learning_Arabic,No_years_Arabic_countries,general_level,lev
el_study,year_or_semester,educational_institution)>
<!ELEMENT age (#PCDATA)>
<!ELEMENT gender (#PCDATA)>
<!ELEMENT nationality (#PCDATA)>
<!ELEMENT mothertongue (#PCDATA)>
<!ELEMENT nativeness (#PCDATA)>
<!ELEMENT No_languages_spoken (#PCDATA)>
<!ELEMENT No_years_learning_Arabic (#PCDATA)>
<!ELEMENT No_years_Arabic_countries (#PCDATA)>
<!ELEMENT general_level (#PCDATA)>
<!ELEMENT level_study (#PCDATA)>
<!ELEMENT year_or_semester (#PCDATA)>
<!ELEMENT educational_institution (#PCDATA)>
<!ELEMENT text_profile
(genre,where,year,country,city,timed,ref_used,grammar_ref_used,mono_
dic_used,bi_dic_used,other_ref_used,mode,medium,length)>
<!ELEMENT genre (#PCDATA)>
<!ELEMENT where (#PCDATA)>
<!ELEMENT year (#PCDATA)>
<!ELEMENT country (#PCDATA)>
<!ELEMENT city (#PCDATA)>
<!ELEMENT timed (#PCDATA)>
<!ELEMENT ref_used (#PCDATA)>
<!ELEMENT grammar_ref_used (#PCDATA)>
<!ELEMENT mono_dic_used (#PCDATA)>
<!ELEMENT bi_dic_used (#PCDATA)>
<!ELEMENT other_ref_used (#PCDATA)>
<!ELEMENT mode (#PCDATA)>
<!ELEMENT medium (#PCDATA)>
<!ELEMENT length (#PCDATA)>
<!ELEMENT text (title,p+)>
<!ELEMENT title (t*)>
<!ATTLIST t
n CDATA #REQUIRED
ErrTag CDATA #IMPLIED
ErrForm CDATA #IMPLIED

```

```
CorrForm CDATA #IMPLIED
>
<!ELEMENT t (#PCDATA)>
<!ELEMENT p (t+)>
<!ATTLIST p
id CDATA #REQUIRED
ErrTag CDATA #IMPLIED
ErrForm CDATA #IMPLIED
CorrForm CDATA #IMPLIED
>
]>
```

Figure 5.12: DTD model for XML files containing metadata and stand-off annotation by tokens

5.3.2 Design

Based on the annotation standards specified for the ALC files, the tagging tool CETAr was developed. With the CETAr, a user can (i) annotate each error with a tag indicating the error type, (ii) specify the error form, and (ii) suggest a corrected form based on the annotator’s experience. This tool was developed to be used in three phases in order to make annotation faster and more consistent. The aim of the first phase is to enable the user to select and tag the corpus tokens manually based on particular error categories and types; this phase includes a tokenisation process prior to the annotation. The second phase aims to avoid inconsistency in one text, so when a word is selected by the user, all similar words in the text are identified, allowing the user to add the same tag to them all in one tagging step. The third phase aims to automate a part of the tagging process by adapting the translation memory approach (Arthern, 1978, 1981; Kay, 1980). Arthern (1981) described the translation memory approach as following:

“It must in fact be possible to produce a programme which would enable the word processor to ‘remember’ whether any part of a new text typed into it had already been translated, and to fetch this part, together with the translation which had already been made, and display it on the screen or print it out, automatically. ... In effect, we should be operating an electronic ‘cut and stick’ process which would, according to my calculations, save at least 15 per cent of the time which translators now employ in effectively producing translations” (Arthern, 1981: 318).

Adapting this method allows using already tagged words as a source for tagging the same words automatically in new texts.

This tool is integrated in the ALC database on Microsoft Access – using the Visual Basic for Applications (VBA) language – in order to facilitate the retrieval of corpus texts before annotating and re-generating them after the annotation in four formats as described in the annotation standards. A number of features are included in the CETAr such as tokenisation, manual tagging, smart-selection, auto tagging, and others, all of which are described in the following sections. Additionally, the CETAr interface provides Arabic translations for the English interface shown in Figure 5.13.

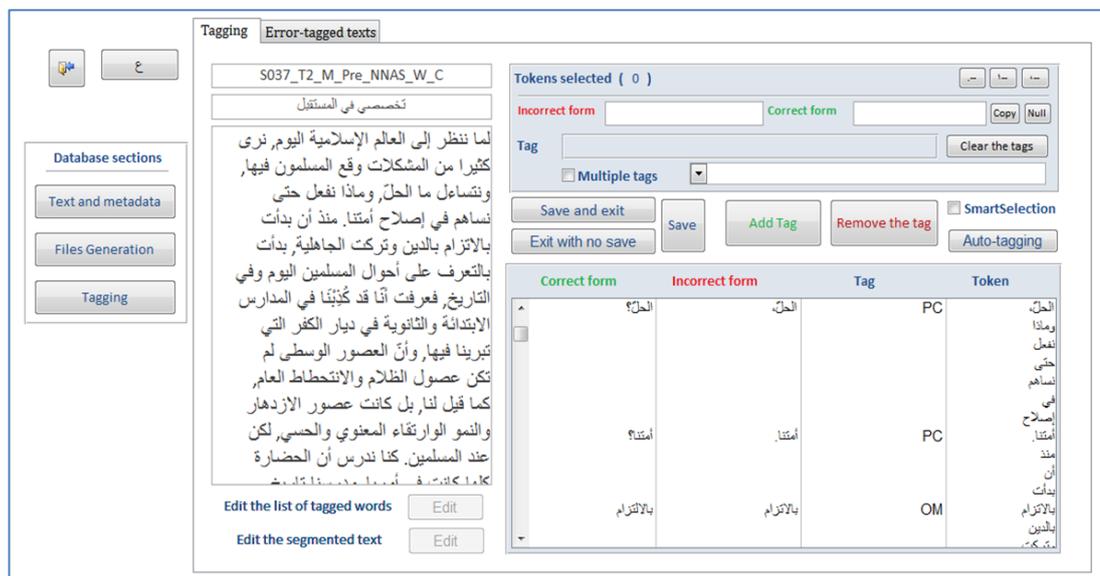


Figure 5.13: The main interface of the CETAr

5.3.3 Tokenisation

The text tokenisation process helps in segmenting the text into separate tokens in order to make it easier for the annotator to attach the tags to those tokens which include errors. The tokenisation function replaces spaces in the text with new line breaks with segmenting punctuations from the words. It also adds the structural features around each part of the text such as the title (<title> and </title>) and paragraphs with their numbers (<p n=1> and </p>). See sample code of the tokenisation process in Figure 5.14.

```

TxtStruc.ReadingOrder = 1
TxtStruc.TextAlign = 1

If TXTtitle.Value <> "" Then
SplTit = Replace(TXTtitle.Value, " ", vbCrLf)
TxtStruc.Value = "<title>" & vbCrLf & SplTit & vbCrLf & "</title>"
Else
TxtStruc.Value = "<title>" & vbCrLf & "</title>"
End If

TextArray() = Split(TXTraw.Value, vbCrLf)
ArrLen = UBound(TextArray)

For i = 0 To ArrLen

TextArray(i) = Replace(TextArray(i), " ", vbCrLf)
TextArray(i) = Replace(TextArray(i), ".", vbCrLf & ".")
TextArray(i) = Replace(TextArray(i), ";", vbCrLf & ";")
TextArray(i) = Replace(TextArray(i), "°", vbCrLf & "°")
TextArray(i) = Replace(TextArray(i), "¿", vbCrLf & "¿")
TextArray(i) = Replace(TextArray(i), "!", vbCrLf & "!")
TextArray(i) = Replace(TextArray(i), "/", vbCrLf & "/")
TextArray(i) = Replace(TextArray(i), "@", vbCrLf & "@")
TextArray(i) = Replace(TextArray(i), "#", vbCrLf & "#")
TextArray(i) = Replace(TextArray(i), "$", vbCrLf & "$")
TextArray(i) = Replace(TextArray(i), "%", vbCrLf & "%")
TextArray(i) = Replace(TextArray(i), "*", vbCrLf & "*")
TextArray(i) = Replace(TextArray(i), ")", vbCrLf & ")")
TextArray(i) = Replace(TextArray(i), "(", vbCrLf & "(")
TextArray(i) = Replace(TextArray(i), "?", vbCrLf & "?")
TextArray(i) = Replace(TextArray(i), ",", vbCrLf & ",")
TextArray(i) = Replace(TextArray(i), "'", vbCrLf & "'")
TextArray(i) = Replace(TextArray(i), "\", vbCrLf & "\")
TextArray(i) = Replace(TextArray(i), ">", vbCrLf & ">")
TextArray(i) = Replace(TextArray(i), "<", vbCrLf & "<")
TextArray(i) = Replace(TextArray(i), "'", vbCrLf & "'")
TextArray(i) = Replace(TextArray(i), "=", vbCrLf & "=")
TextArray(i) = Replace(TextArray(i), "+", vbCrLf & "+")
TextArray(i) = Replace(TextArray(i), "-", vbCrLf & "-")
TextArray(i) = Replace(TextArray(i), "_", vbCrLf & "_")
TextArray(i) = Replace(TextArray(i), "]", vbCrLf & "]")
TextArray(i) = Replace(TextArray(i), "[", vbCrLf & "[")
TextArray(i) = Replace(TextArray(i), "}", vbCrLf & "}")
TextArray(i) = Replace(TextArray(i), "{", vbCrLf & "{")
TextArray(i) = Replace(TextArray(i), ";", vbCrLf & ";")
TextArray(i) = Replace(TextArray(i), ":", vbCrLf & ":")
TextArray(i) = Replace(TextArray(i), "|", vbCrLf & "|")
TextArray(i) = Replace(TextArray(i), Trim("''''"), "")

TextArray(i) = "<p n=" & i + 1 & ">" & vbCrLf & TextArray(i) & vbCrLf & "</p>"
TxtStruc.Value = TxtStruc.Value & vbCrLf & TextArray(i)

Next

```

Figure 5.14: Sample code of the tokenisation process

Figure 5.15 shows the final result of the text S938_T1_F_Uni_NNAS_S_C in XML format (UTF-16 coding) after it has been tokenised by the CETAr.

```
<doc ID="S938_T1_F_Uni_NNAS_S_C">
  <text>
    <title>
      <t n="1">قصة</t>
    </title>
    <p id="1">
      <t n="2">بسم</t>
      <t n="3">لله</t>
    </p>
    <p id="2">
      <t n="4">السلام</t>
      <t n="5">عليكم</t>
      <t n="6">ورحمة</t>
      <t n="7">لله</t>
    </p>
    <p id="3">
      <t n="8">أ</t>
      <t n="9">ـ</t>
      <t n="10">اسمي</t>
      <t n="11">#</t>
      <t n="12">معلومة</t>
      <t n="13">شخصية</t>
      <t n="14">محدوفة</t>
      <t n="15">#</t>
      <t n="16">،</t>
      <t n="17">أ</t>
      <t n="18">ـ</t>
      <t n="19">جامعة</t>
      <t n="20">الأميرة</t>
      <t n="21">نورة</t>
      <t n="22">،</t>
      <t n="23">في</t>
      <t n="24">كلية</t>
      <t n="25">اللغة</t>
      <t n="26">العربية</t>
      <t n="27">آداب</t>
      <t n="28">،</t>
      <t n="29">بقسم</t>
      <t n="30">اللغة</t>
      <t n="31">العربية</t>
      <t n="32">،</t>
      <t n="33">مستوى</t>
      <t n="34">الثالث</t>
      <t n="35">،</t>
    </p>
  </text>
</doc>
```

```
<t n="36">أ</t>  
<t n="37">--</t>  
<t n="38">بعدهما</t>  
<t n="39">أتيث</t>  
<t n="40">هنا</t>  
<t n="41">أقيمت</t>  
<t n="42">أول</t>  
<t n="43">رحلة</t>  
<t n="44">،</t>  
<t n="45">أ</t>  
<t n="46">--</t>
```

Figure 5.15: Example of a text tokenised by CETAr

5.3.4 Manual Error Tagging

Error tagging is the fundamental function for which this tool was developed, as the main purpose of annotating errors using the CETAr is to standardise the format of the output files. This tool enables users to assign one or more tags to any token including an error. Additionally, the user can suggest a correct form to the error. Based on the annotation the user adds using the CETAr, the annotated text can be generated in a number of standard file formats as explained in the annotation standards section.

5.3.5 Smart Selection

The aim of this feature is to avoid inconsistency when working on a text. To achieve this aim, when the user selects an error, the smart selection feature identifies all similar error forms in the text, allowing the user to assign the same tag to them all in one tagging step with no need to repeat the annotation process with each error. This function can be enabled or disabled based on the user's choice (Figure 5.16). For instance, if a token requires a further tag, such as for missing punctuation, the smart selection feature should be disabled; otherwise, all similar tokens will be incorrectly tagged with the same error type.


```

        End If
    End If

    If ListTkns.ItemData(i) = ItemSelectedData Then
        If ListTkns.Selected(i) = False Then ListTkns.Selected(i) = True

        If iCount = 0 Then
            'When select one item

            If ListTkns.ItemsSelected.Count = 1 Then
                ItemContent = ListTkns.ItemData(i)
                ItemIndexSaved = ListTkns.ListIndex
                iCount = iCount + 1
                LBLItemIndex.Caption = i

            Else
                ItemContent = ItemContent & ", " & ListTkns.ItemData(i)
                ItemIndexSaved = ItemIndex & ListTkns.ListIndex
                iCount = iCount + 1
                LBLItemIndex.Caption = LBLItemIndex.Caption & ", " & i
            End If

            TXTIncorrectForm.Value = ListTkns.ItemData(i)
            LBLLastSelItemIndex.Caption = ListTkns.ListIndex
            LBLTag.Caption = ListTkns.Column(1, ItemSelectedIndex)
            TXTCorrectForm.Value = ListTkns.Column(3, ItemSelectedIndex)

        Else
            'When select more than one item
            ItemContent = ItemContent & ", " & ListTkns.ItemData(i)
            ItemIndexSaved = ItemIndex & ", " & ListTkns.ListIndex
            iCount = iCount + 1
            LBLItemIndex.Caption = LBLItemIndex.Caption & ", " & i
            TXTIncorrectForm.Value = ListTkns.ItemData(i)
            LBLLastSelItemIndex.Caption = ListTkns.ListIndex
            LBLTag.Caption = ListTkns.Column(1, ItemSelectedIndex)
            TXTCorrectForm.Value = ListTkns.Column(3, ItemSelectedIndex)
        End If
    End If

Next

LBLTknsSelected.Caption = ItemContent
LBLNoItems.Caption = iCount

```

Figure 5.17: Sample code of the smart-selection feature

5.3.6 Auto Tagging

The auto-tagging feature adapts the translation memories approach (Arthern, 1978, 1981; Kay, 1980) in order to automate a part of the tagging process. Specifically, all tokens that have been tagged in previous annotation processes are stored and used as a source for automatically tagging the same words in further texts. Using the auto-

tagging feature is optional to the user; however, users choosing to employ this feature are encouraged to do so before any manual tagging for two reasons. First, following this order makes it easy to check the errors tagged automatically and correct any possible wrong annotations. Second, doing so ensures that any tags added manually later will not be replaced by the auto-tagging function.

To use the auto-tagging function, the user starts by clicking on the auto-tagging button, which causes each token in the text to be compared to the table of pre-tagged tokens. If a given token is found, it is tagged automatically. Tokens that do not appear in the table require manual tagging if they include any error type. The second step is for the annotator to complete any manual tagging. When the annotator finishes and saves the annotated data to the database, the third step updates the table of pre-tagged tokens to include all new words that have been tagged manually and do not currently exist in the pre-tagged list of tokens (Figure 5.18). It is important to mention that, although all cases of tagged tokens are saved to the list of pre-tagged tokens, the auto-tagging feature annotates only those errors that lie under the first category of the ETAr, orthography, where errors depend on word form. The context must be analysed for the other categories. For instance, errors under the morphological category may need a morphological analysis to ensure that all contexts where the token appears are incorrect cases.

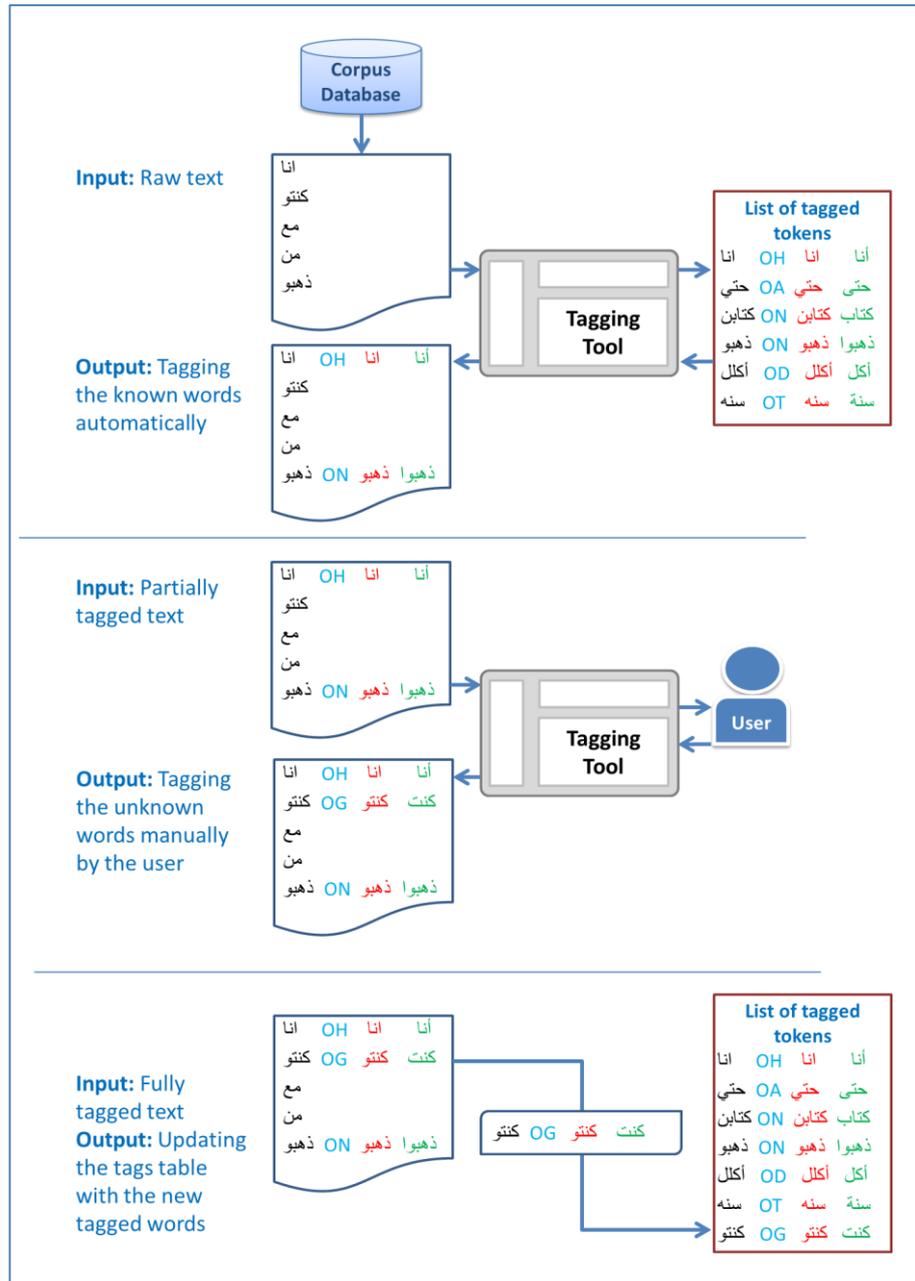


Figure 5.18: Steps of using the auto-tagging function

The auto-tagging feature starts by retrieving the list of pre-tagged tokens. If there is a token in this list tagged with any error type under the category *Orthographic*, which starts by the symbol "O", it will be compared to the tokens in the text, and when a similar token is found it will be tagged with the same values of *Tag*, *Incorrect form* and *Correct form*. See sample code of the auto-tagging feature in Figure 5.19.

After completing the manual tagging by the annotator, the auto-tagging feature updates the table of pre-tagged tokens by adding all new tokens that have been tagged manually and do not exist in the pre-tagged list of tokens.

```
For i = 0 To ToknsNumInListBox
    GlobalTknsArray(i, 0) = ""
    GlobalTknsArray(i, 1) = ""
    GlobalTknsArray(i, 2) = ""
    GlobalTknsArray(i, 3) = ""
Next
For i = 0 To ToknsNumInListBox
    If ListTkns.Column(0, i) <> nul Then GlobalTknsArray(i, 0) = ListTkns.Column(0, i)
    If ListTkns.Column(1, i) <> nul Then GlobalTknsArray(i, 1) = ListTkns.Column(1, i)
    If ListTkns.Column(2, i) <> nul Then GlobalTknsArray(i, 2) = ListTkns.Column(2, i)
    If ListTkns.Column(3, i) <> nul Then GlobalTknsArray(i, 3) = ListTkns.Column(3, i)
Next
If ListPreTagged.ListCount <> 0 Then 'if the list is not empty then do the process of auto-tagging
    For i = 0 To ListPreTagged.ListCount - 1
        For o = 0 To ToknsNumInListBox
            If ListPreTagged.ItemData(i) = GlobalTknsArray(o, 0) Then
                TagCat = ListPreTagged.Column(1, i)
                OrthoCat = InStr(1, TagCat, "O")
                If OrthoCat = 1 Then 'The Tag category is Orthography
                    GlobalTknsArray(o, 1) = ListPreTagged.Column(1, i)
                    GlobalTknsArray(o, 2) = ListPreTagged.Column(2, i)
                    GlobalTknsArray(o, 3) = ListPreTagged.Column(3, i)
                End If
            End If
        Next
    Next
End If
For intCounter = 0 To ToknsNumInListBox
ListTkns.RemoveItem 0
```

```

Next
For i = 0 To ToknsNumInListBox
ListTkns.AddItem GlobalTknsArray(i, 0) & ";" & GlobalTknsArray(i, 1) & ";" &
GlobalTknsArray(i, 2) & ";" & GlobalTknsArray(i, 3)
Next

```

Figure 5.19: Sample code of the Auto-tagging function

5.3.7 Further Features

The annotator is able to edit the list of tagged tokens manually using the feature *Edit the list of tagged words*. This feature is helpful in cases where the list includes any token that has been tagged incorrectly. The annotator may need to check the list to ensure that all orthographical errors it includes are authentic and can be used for the purpose of this feature (Figure 5.20).

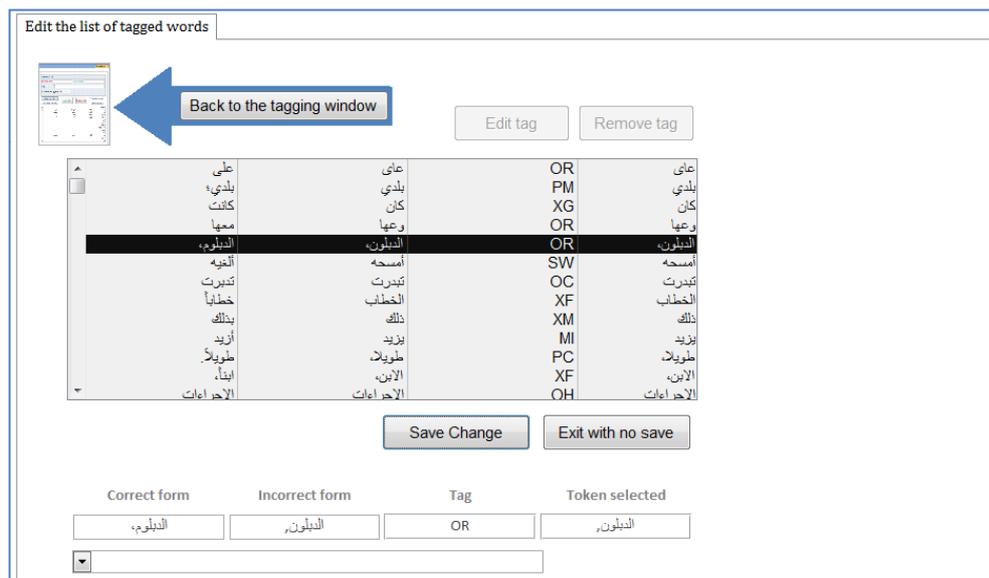


Figure 5.20: Editing the list of tagged tokens

In the same way, the segmentation of the text being annotated can be manually edited. If a token has been segmented incorrectly or the annotator recognises a need to split two tokens for any reason, the annotator can manually make these adjustments by using the feature *Edit the segmented text*. Any token that is split manually into two tokens will be read and annotated as two separate tokens in the future; likewise, any tokens manually combined will be read and annotated as a

single token. Additionally, the annotator can see the final output of the annotated text in the four formats before they are generated (Figure 5.21).

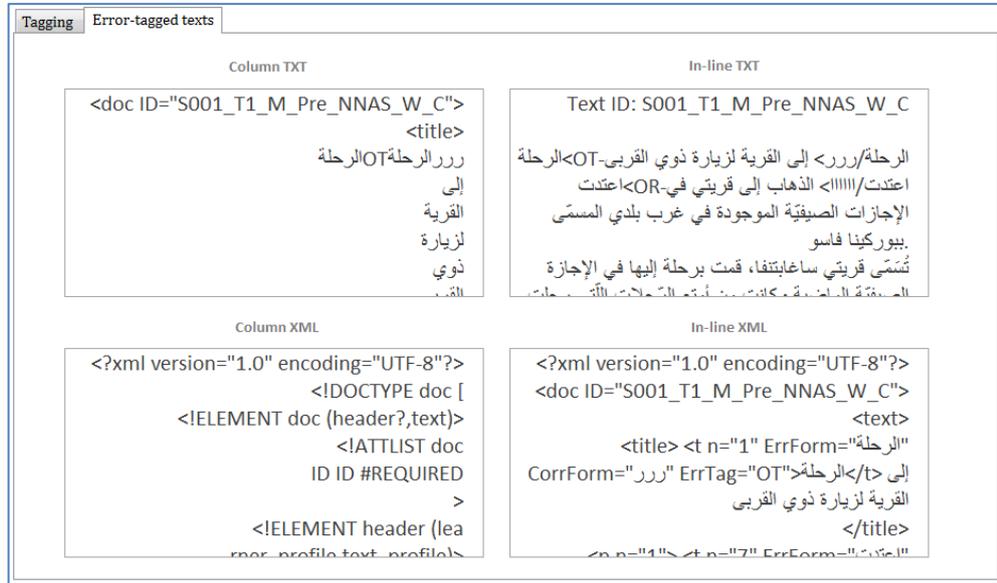


Figure 5.21: Example of a final output of the annotation in CETAr

5.3.8 Evaluation

To evaluate the consistency and speed of annotation by the CETAr, two annotators were asked to tag errors in a sample of five texts from the ALC data. Those annotators (indicated by T3 and T6) participated in some evaluation experiments with the ETAr (see Section 5.7.2 for more information about the annotators). Both annotators had the same sample and were asked to annotate errors twice. The first annotator (T3) was asked to annotate errors first on a paper copy and then using the CETAr the next day, while the second annotator performed the tasks in the opposite order to ensure that they were unable to familiarise themselves with the errors when switching from the hard copy to the CETAr or vice versa.

The consistency evaluation results revealed that the paper copy included the tag “NI”, which does not exist in the ETAr; it seems that the annotator confused two tags or misspelled a tag. However, the possibility of using non-existent tags was reduced to zero in the CETAr, as all tags are selected from a drop-down menu containing 29 error types under 5 categories. Another observation in terms of consistency is that some similar errors received different tags in the paper copy; for example, “وبرنامج” was first tagged with “XN” (syntactic error in number) and then

with “XG” (syntactic error in gender). Using the CETAr, the smart-selection feature helps in selecting and tagging all similar tokens with the same tag in one step.

With respect to speed, Table 5.1 illustrates how much time was taken for each text by each annotator. The table shows that using the CETAr was slightly faster than the paper annotation, with an average of 8.6 minutes for the CETAr compared to 9.15 minutes for the paper task. One possibility for this difference is the use of the smart-selection feature, which reduces the time needed for similar errors, as they can be selected and tagged as a single error. Another possibility is that the annotators spent extra time consulting the error tagset, which was on a separate sheet, for the paper annotation. In contrast, the tagset is hosted in a drop-down menu as a part of the CETAr; thus, annotators had no need to use any external reference.

Table 5.1: Results of task 1 of annotation speed by hand and using CETAr

Sample	Text Code	Text size (tokens)	Tagging time (minute)			
			By hand		By CETAr	
			T3	T6	T3	T6
1	S002_T1_M_Pre_NNAS_W_C	294	6.5	8	7	7.5
2	S323_T1_M_Pre_NNAS_W_C	269	12	13.5	11.5	11
3	S752_T1_M_Pre_NAS_W_C	259	5.5	5.5	5.5	6.5
4	S793_T2_F_Pre_NAS_W_H	232	6	6	5.5	6
5	S927_T2_F_Pre_NNAS_S_C	321	14.5	14	12.5	13
Average		275	8.9	9.4	8.4	8.8
			9.15		8.6	

In addition, a 10-fold cross-validation experiment was performed to evaluate the auto-tagging feature. This experiment used 10 samples from the ALC. Each sample contained two texts of approximately 1000 tokens, resulting in a total size of 10,031 tokens (Table 5.2).

Table 5.2: Samples used to test the auto-tagging feature

Sample	Text Code	Text size (tokens)	Sample size (tokens)
1	S793_T1_F_Pre_NAS_W_H	527	1095
	S799_T1_F_Pre_NAS_W_C	568	
2	S662_T1_F_Uni_NAS_W_H	561	971
	S938_T1_F_Uni_NNAS_S_C	410	
3	S785_T2_F_Pre_NAS_W_H	593	1072
	S931_T1_F_Pre_NNAS_S_C	479	
4	S498_T1_M_Uni_NAS_W_C	529	978
	S927_T1_F_Pre_NNAS_S_C	449	
5	S274_T1_F_Pre_NAS_W_H	521	931
	S505_T1_M_Uni_NNAS_W_C	410	
6	S496_T1_M_Uni_NAS_W_C	511	923
	S301_T1_M_Pre_NNAS_W_C	412	
7	S664_T1_F_Uni_NAS_W_H	544	963
	S038_T1_M_Pre_NNAS_W_C	419	
8	S037_T1_M_Pre_NNAS_W_C	593	1053
	S037_T2_M_Pre_NNAS_W_C	460	
9	S437_T1_M_Uni_NAS_W_C	571	1023
	S448_T1_M_Uni_NNAS_W_C	452	
10	S670_T1_F_Uni_NAS_W_H	514	1022
	S938_T2_F_Uni_NNAS_S_C	508	

The experiment was conducted 10 times. During each experiment, the orthographical errors in one of the samples were tagged by the auto-tagging feature using the annotation of the remaining nine samples, and the annotation was checked manually by the researcher. The percentage of correctness varied from 76% as the lowest achieved to 95% as the highest percentage, with an average of 88% (Table 5.3).

Table 5.3: Results of testing the auto-tagging feature

Iteration	Samples used	Sample tested	Sample size (tokens)	Instances		%
				Total	Correct	
1	S2 S3 S4 S5 S6 S7 S8 S9 S10	S1	1095	53	41	77%
2	S1 S3 S4 S5 S6 S7 S8 S9 S10	S2	971	81	75	93%
3	S1 S2 S4 S5 S6 S7 S8 S9 S10	S3	1072	81	74	91%
4	S1 S2 S3 S5 S6 S7 S8 S9 S10	S4	978	59	56	95%
5	S1 S2 S3 S4 S6 S7 S8 S9 S10	S5	931	40	36	90%
6	S1 S2 S3 S4 S5 S7 S8 S9 S10	S6	923	40	34	85%
7	S1 S2 S3 S4 S5 S6 S8 S9 S10	S7	963	48	45	94%
8	S1 S2 S3 S4 S5 S6 S7 S9 S10	S8	1053	34	26	76%
9	S1 S2 S3 S4 S5 S6 S7 S8 S10	S9	1023	37	31	84%
10	S1 S2 S3 S4 S5 S6 S7 S8 S9	S10	1022	106	101	95%
				Average	88%	

The inaccuracy in correcting orthographical errors using the auto-tagging feature were mostly centred on those situations in which a word was annotated with an orthographic error based on a specific context, saved to the list of pre-tagged tokens, and then applied to other cases in other contexts. For example, the word “إنه” was tagged as an orthographical error in the letter *Hamza* in a particular context where it was wrong; however, when we used the auto-tagging feature, nine cases of the word “إنه” in different contexts were tagged as an orthographical error in *Hamza*, while they were not errors in those contexts. This confusion also occurred with other words such as “أن” which was tagged as an error in *Hamza* six times.

The fact that these errors appeared in *Hamza* is significant. *Hamza* has specific rules in Arabic writing, but it seems to be complicated to learners. *Hamza* ranked as the second among the 10 most common errors found in a 10,000-word sample that was tagged for errors by three annotators (for more details about the sample, see Section 5.7.1; for more details about the most common errors in the ALC, see Section 7.3.4). A possible solution to reduce the cases annotated inaccurately, particularly those based on specific contexts, is to remove those tokens manually from the list of the pre-tagged tokens.

5.4 Error Tagset of Arabic (ETAr)

As previously discussed, the sole tagset existing for Arabic error annotation is the ARIDA tagset (Abuhakema *et al.*, 2009), which has a number of limitations. To

address this gap, a new error taxonomy was developed for this project based on the results of a number of error-analysis studies (Alaqeeli, 1995; Alateeq, 1992; Alhamad, 1994; Alosaili, 1985) as well as ARIDA itself. The reason for relying on the ARIDA tagset is that it includes two comprehensively well-described categories, style and punctuation. The other four studies investigate different types of errors in Arabic learner production using the bottom-up method where the authors analyse their own samples and then extract the corresponding error-type lists. These studies do not aim to develop an error-type tagset to be used for further projects such as learner corpora. Nonetheless, their error taxonomies are valid and adaptable since they include significant and comprehensive classes of learner errors. Furthermore, the texts from which these error types are derived are authentic, which adds to the validity of their taxonomies. The following is a brief overview of those studies:

- Alosaili (1985) investigates errors of Arabic learners in their spoken production. His list of errors consists of three main classes: phonological, syntactic, and lexical errors, with sub-types under each domain. Some of these types are included in the tagset proposed in this study, specifically those related to orthography, as they are well-formed and cover clearly significant types.
- Alateeq (1992) focusses on semantic errors and extracts a detailed list of them, which is adapted in the proposed tagset. Aside from these semantic errors, the study also lists several phono-orthographical, morphological, and syntactic types of errors.
- Alhamad (1994) focusses on the writing production of advanced level Arabic learners, and concludes with a list of error categories: phonological, orthographical, morphological, syntactic, and semantic errors. The most comprehensive errors are under orthography and syntax, which are added to the tagset created in this project.
- Alaqeeli (1995) examines learners' written errors in a particular type of sentence: a verbal sentence "الجملة الفعلية". This study, therefore, has a limited number of error types under two categories: morphological and syntactic. However, errors under the morphological category are deemed worthy of inclusion in the tagset suggested, due to their comprehensiveness.

Table 5.4: Error taxonomies in some Arabic studies

Alosaili	Alateeq	Alhamad	Alaqeeli
<ul style="list-style-type: none"> • Phonological errors • Syntactic errors • Lexical errors 	<ul style="list-style-type: none"> • Phono-orthographical errors • Morphological errors • Syntactic errors • Semantic errors 	<ul style="list-style-type: none"> • Syntactic errors • Morphological errors • Orthographical errors • Phonological errors • Semantic errors 	<ul style="list-style-type: none"> • Syntactic errors • Morphological errors

5.4.1 Error Categories and Types

This study aimed to develop a new error tagset that can provide users (e.g. researchers of Arabic, teachers, etc.) with easily understood broad classes or categories and comprehensive error types. The suggested taxonomy, ETAr, includes 37 types of errors, divided into 6 classes or categories: orthography, morphology, syntax, semantics, style, and punctuation. The ETAr has two levels of annotation in order to simplify its use and evaluation at this early stage of development. Each tag consists of two Arabic characters (with an equivalent tag in English). The first character in each tag indicates the error class or category, while the second symbolises the error type. For example, in the tag <OH>, the letter *O* indicates the error category, *Orthography*, while the letter *H* indicates the error type, *Hamza*, which lies under the category *Orthography*.

This taxonomy is flexible and can be modified based on studies, evaluations, or relevant results. In addition, end each category contains an item named “*Other [...] errors*”, which can handle any error that does not yet have a tag.

Table 5.5: Error Tagset of Arabic (ETAr)

Error Category	Error Type	Arabic tag	English tag
Orthography الإملاء 'l'imlā'	1. Hamza (هـ، أ، إ، و، ي، ئ، ؤ)	<إه>	<OH>
	2. Tā' Mutaṭarrifa (تاء المتطرفة (ة، ت))	<إة>	<OT>
	3. 'alif Mutaṭarrifa (ألف المتطرفة (ا، ي))	<إى>	<OA>
	4. 'alif Fāriqa (الألف الفارقة (كتبوا))	<إت>	<OW>
	5. Lām Šamsīya (اللام الشمسية (أطالب))	<إا>	
	6. Tanwīn (تنوين (وُؤ))	<إل>	<ON>
	7. Faṣl wa Waṣl (Conjunction) (الفصل والوصل)	<إو>	<OF>
	8. Shortening the long vowels (اوي) تقصير الصوائت الطويلة → وُؤ)	<إف>	<OS>
	9. Lengthening the short vowels (اوي) تطويل الصوائت القصيرة (وُؤ → وُؤ)	<إق>	<OG>
	10. Wrong order of word characters (الخطأ في ترتيب الحروف داخل الكلمة)	<إط>	<OC>
	11. Replacement in word character(s) (استبدال حرف أو أحرف من الكلمة)	<إس>	<OR>
	12. Redundant character(s) (وجود حرف أو أحرف زائدة)	<إز>	<OT>
	13. Missing character(s) (نقص حرف أو أحرف)	<إن>	<OM>
	14. Other orthographical errors (أخطاء إملائية أخرى)	<إخ>	<OO>
Morphology الصرف 'ssarf'	15. Word inflection (صيغة الكلمة)	<صص>	<MI>
	16. Verb tense (زمن الفعل)	<صز>	<MT>
	17. Other morphological errors (أخطاء صرفية أخرى)	<صخ>	<MO>
Syntax النحو 'nnaḥw'	18. Case/mood mark (الموقع الإعرابي أو علامة الإعراب)	<نب>	<XC>
	19. Definiteness (التعريف والتنكير)	<نع>	<XF>
	20. Gender (التذكير والتأنيث)	<نذ>	<XG>
	21. Number (singular, dual, and plural) (العدد (الإفراد والتثنية والجمع))	<نف>	<XN>
	22. Word(s) order (ترتيب المفردات داخل الجملة)	<نت>	<XR>
	23. Redundant word(s) (وجود كلمة أو كلمات زائدة)	<نز>	<XT>
	24. Missing word(s) (نقص كلمة أو كلمات)	<نن>	<XM>
	25. Other syntactic errors (أخطاء نحوية أخرى)	<نخ>	<XO>

Semantics الدلالة ' <i>ddalāla</i>	26. Word selection اختيار الكلمة المناسبة	<دب>	<SW>
	27. Phrase selection اختيار العبارة المناسبة	<دق>	<SP>
	28. Failure of expression to indicate the intended meaning قصور التعبير عن أداء المعنى المقصود	<دد>	<SM>
	29. Wrong context of citation from Quran or Hadith الاستشهاد بالكتاب والسنة في سياق خاطئ	<دس>	<SC>
	30. Other semantic errors أخطاء دلالية أخرى	<دخ>	<SO>
Style الأسلوب ' <i>uslūb</i>	31. Unclear style أسلوب غامض	<سغ>	<TU>
	32. Prosaic style أسلوب ركيك	<سض>	<TP>
	33. Other stylistic errors أخطاء أسلوبية أخرى	<سخ>	<TO>
Punctuation علامات الترقيم ' <i>alāmāt 't-tarqīm</i>	34. Punctuation confusion الخلط في علامات الترقيم	<تط>	<PC>
	35. Redundant punctuation علامة ترقيم زائدة	<تزر>	<PT>
	36. Missing punctuation علامة ترقيم مفقودة	<تن>	<PM>
	37. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<تخ>	<PO>

5.5 First Evaluation: Comparison of Two Tagsets

The main aim of this evaluation was to compare two tagsets for annotating errors in Arabic, the ARIDA tagset (Abuhakema *et al.*, 2009) and the ETAr (Table 5.5). The comparison was performed by measuring the inter-annotator agreement when using each tagset to annotate a sample of ALC texts for errors. Such measurement should provide valuable insights into the understandability and usability of the ETAr when compared to the ARIDA tagset.

5.5.1 Sample and Annotators

Two texts were selected randomly for this experiment from the first version of the ALC. The first text, *S003_T1_M_Pre_NNAS_W_C*, includes 107 words, while the second text, *S022_T2_M_Pre_NNAS_W_C*, includes 132 words. Two annotators (indicated by T1 and T2) participated in this experiment (see Table 5.6 below for more details).

Table 5.6: Annotators who participated in the first evaluation of the ETAr

	T1	T2
Qualifications	<ul style="list-style-type: none"> • First degree in Arabic and Islamic studies • Master degree in Applied Linguistics 	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Applied Linguistics
Experience in teaching Arabic	Teaching Arabic and Islamic culture to native and non-native Arabic speakers in Saudi Arabia for several years	Teaching Arabic to non-native Arabic speakers in Saudi Arabia for several years
Experience in error annotation	No previous experience	No previous experience

5.5.2 Task and Training

Each annotator was required to do two basic steps for each error in the experiment sample. First, the annotator was to underline any token including a clear error. Subsequently, the annotator was instructed to add the most appropriate tag that matched the error type using first the ARIDA tagset and then the ETAr. The annotators were able to complete this task on the same day due to the small sample given.

As the aim of the evaluation was to measure the extent to which the tagset could be understood and used by untrained users, the annotators received the tables of both tagsets with no training or explanation about the meaning or scope of the tags. The assumptions were that both error tagsets should be clear enough to both annotators and that both should be able to understand which tag was most appropriate to use for each error. This measurement may be sufficient to check whether a tagset can be independently understood against another tagset, bearing in mind that the differences between annotators may occur sometimes because of the annotator's view of the error type.

5.5.3 Results

The results show that T1 detected 80 errors, while T2 found 91, and they shared 42 errors. The comparison was performed on the 42 shared errors by calculating matched tags between T1 and T2 in each tagset. The evaluation used *Cohen’s Kappa* (Cohen, 1960), which measures the agreement of the assigned tags between two annotators and takes into consideration the possibilities of agreement by chance. The observed agreement when the annotators used the ARIDA tagset was 33%, resulting in a weighted *Cohen’s Kappa* value of $k = 0.292$ ($p < 0.001$). By using the ETAr, the observed agreement was 52%, resulting in a weighted *Cohen’s Kappa* value of $k = 0.468$ ($p < 0.001$). Although the ETAr achieved a higher score, it was still not perfect, which means that it needs more refinement and that more tests are still needed using other texts and more annotators.

Table 5.7: Annotating comparison between ARIDA and ETAr

Tagset	No. of matching tags (out of 42)	Percent*	<i>Cohen’s Kappa</i>	Sig.
ARIDA	14	33%	0.292	$p < 0.001$
ETAr	22	52%	0.468	$p < 0.001$

* Number of agreement cases divided by the total cases

After they completed the annotation task, the annotators received a short questionnaire with two main questions (Appendix E.1). In response to the question “Which taxonomy was more understandable? And why?”, both selected the ETAr because of the logical order of its items and its comprehensiveness. For the question “Which of them was quick and easy for annotating? And why?”, they both chose the ETAr, noting their belief that using the ETAr made it easier to select the proper tag and stating that the tags were clearer with no ambiguity or overlap.

5.5.4 Limitations and Suggestions

Determining whether a word/phrase was right or wrong was completely based on the annotator’s view. It was very likely that some differences in their decisions, particularly in some categories such as semantics and style, relate to the annotator’s degree of linguistic knowledge. The disagreements might have been minimised if annotators were given texts with errors that had been identified and were asked to

mark the appropriate tag on each error. This method was used in the next experiment to avoid such differences.

The scores achieved using the ETAr were not as high as expected, which might be because of the lack of training. Thus, for maximum accuracy, the tagset needs to be combined with a manual and the annotators need to be trained prior to performing the task.

5.6 Second Evaluation: Inter-Annotator Agreement Measurement

The aim of this experiment was to improve the understandability and usability of the ETAr by considering three steps. The first step was for two experts in the Arabic language to conduct a review of the tagset. The second step was to give the annotators texts with errors already identified by the researcher and one of the Arabic language experts who participated in reviewing the tagset; using this pre-identified text, the annotators were tasked with marking the appropriate tag on each error using the Error Tagging Manual for Arabic (ETMar) that explains all error types in the tagset with rules and examples of how to tag linguistic errors. The third step was to train the annotators during the experiment.

5.6.1 Sample

The sample used in the second evaluation consists of two lists of 100 varied sentences that contain errors. Errors in these lists were distributed equally among 28 error types existing in the ETAr (excluding the last type in each category, reserved for “other” such errors), which yielded three or four examples for each error type in each list.

5.6.2 Evaluators

Two Arabic language experts (indicated by E1 and E2) participated in this experiment. See Table 5.8 below for more details about these evaluators.

Table 5.8: Evaluators who participated in the first refinement of the ETAr

	E1	E2
Qualifications	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Arabic Linguistics 	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Arabic Morphology and Syntax • Undertaking a PhD degree in Arabic Syntax
Experience in teaching Arabic	Teaching Arabic to university students in Saudi Arabia for several years	Teaching Arabic to university students in Saudi Arabia for several years
Experience in error annotation	No previous experience	No previous experience

The evaluators were given the ETAr and asked to give suggestions based on their experience in five aspects: error types to be added, error types to be deleted, error types to be changed, error types to be integrated, and error types to be split. Their suggestions included moving the error *Faṣl wa Waṣl (Conjunction)* to the Semantics category. They recommended integrating the errors *Word selection* and *Phrase selection* into one error named *Word/phrase selection* and the errors *Unclear style* and *Prosaic style* into one error named *Unclear or weak style*. Additionally, they suggested removing the error *Failure of expression to indicate the intended meaning* as well as renaming some error types to be more specific. For example, *'alif Fāriqa* became *Confusion in 'alif Fāriqa*, and *Definiteness* was changed to *Agreement in definiteness*. Other changes can be seen in the second version of the ETAr in Table 5.9. This version includes 34 types of error, divided into 6 categories.

Table 5.9: Second version of the ETAr

Error Category	Error Type	Arabic tag	English tag
1. Orthography الإملاء 'l'imlā'	1. Hamza (هـ، ء، أ، إ، و، ئ، ث)	<إه>	<OH>
	2. Confusion in <i>Hā'</i> and <i>Tā'</i> <i>Mutaṭarrifatain</i> (هـ، ء، ت) الخلط في الهاء والتاء المتطرفتين	<إة>	<OT>
	3. Confusion in <i>'alif Mutaṭarrifa</i> (أ، ي) الخلط في الألف (أ، ي) المتطرفة	<إى>	<OA>
	4. Confusion in <i>'alif Fāriqa</i> (كتبوا) الخلط في الألف الفارقة (كتبوا)	<إت>	<OW>
	5. <i>Lām Samsīya</i> dropped (إسقاط اللام الشمسية (الطالب)	<إا>	
	6. Confusion between <i>Nūn</i> (ن) and <i>Tanwīn</i> (ٍٍٍ) الخلط بين النون والتنوين	<إل>	<ON>
	7. Shortening the long vowels (أوي) تقصير الصوائت الطويلة → (أوي) (ٍٍٍ)	<إف>	<OS>
	8. Lengthening the short vowels (أوي) تطويل الصوائت القصيرة (أوي → (ٍٍٍ)	<إق>	<OG>
	9. Wrong order of word characters (إط) الخطأ في ترتيب الحروف داخل الكلمة	<إط>	<OC>
	10. Replacement in word character(s) (إس) استبدال حرف أو أحرف من الكلمة	<إس>	<OR>
	11. Redundant character(s) (إز) حرف أو أحرف زائدة	<إز>	<OD>
	12. Missing character(s) (إن) حرف أو أحرف ناقصة	<إن>	<OM>
	13. Other orthographical errors (إخ) أخطاء إملائية أخرى	<إخ>	<OO>
2. Morphology الصرف 'ssarf	14. Word inflection (صص) صيغة الكلمة	<صص>	<MI>
	15. Verb tense (صز) زمن الفعل	<صز>	<MT>
	16. Other morphological errors (صخ) أخطاء صرفية أخرى	<صخ>	<MO>
3. Syntax النحو 'nnaḥw	17. Agreement in grammatical case (تب) المطابقة في الإعراب	<تب>	<XC>
	18. Agreement in definiteness (تع) المطابقة في التعريف والتنكير	<تع>	<XF>
	19. Agreement in gender (تذ) المطابقة في الجنس (التذكير والتأنيث)	<تذ>	<XG>
	20. Agreement in number (singular, dual, and plural) (تف) المطابقة في العدد (الإفراد والتثنية والجمع)	<تف>	<XN>
	21. Words order (نت) ترتيب المفردات داخل الجملة	<نت>	<XR>
	22. Redundant word(s) (نز) كلمة أو كلمات زائدة	<نز>	<XT>
	23. Missing word(s) (نن) كلمة أو كلمات ناقصة	<نن>	<XM>
24. Other syntactic errors (نخ) أخطاء نحوية أخرى	<نخ>	<XO>	
4. Semantics الدلالة 'ddalāla	25. Word/phrase selection (دب) اختيار الكلمة أو العبارة المناسبة	<دب>	<SW>
	26. <i>Faṣl wa Waṣl</i> (confusion in use/non-use conjunctions) (دف) الفصل والوصل (الخلط في استخدام أو عدم استخدام أدوات العطف)	<دف>	<SF>
	27. Wrong context of citation from Quran or Hadith (دس) الاستشهاد بالكتاب والسنة في سياق خاطئ	<دس>	<SC>
	28. Other semantic errors (دخ) أخطاء دلالية أخرى	<دخ>	<SO>
5. Style الأسلوب 'l'uslūb	29. Unclear or weak style (سغ) أسلوب غامض أو ركيك	<سغ>	<TU>
	30. Other stylistic errors (سخ) أخطاء أسلوبية أخرى	<سخ>	<TO>
6. Punctuation	31. Punctuation confusion (تط) الخلط في علامات الترقيم	<تط>	<PC>

علامات الترقيم	32. Redundant punctuation علامة ترقيم زائدة	<نز>	<PT>
'alāmāt 't-tarqīm	33. Missing punctuation علامة ترقيم مفقودة	<تن>	<PM>
	34. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<تخ>	<PO>

5.6.3 Annotators

Three annotators (indicated by T3, T4, and T5) participated in this experiment. See Table 5.10 below for more information about them.

Table 5.10: Annotators who participated in the second evaluation of the ETAr

	T3	T4	T5
Qualifications	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Applied Linguistics 	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Arabic Applied Linguistics • Undertaking a PhD degree in Applied Linguistics 	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Linguistics
Experience in teaching Arabic	Teaching Arabic to non-native Arabic speakers in Saudi Arabia for a few years	Teaching Arabic to non-native Arabic speakers in Saudi Arabia for several years	Teaching Arabic to university students in Saudi Arabia for several years
Experience in error annotation	No previous experience	No previous experience	No previous experience

The annotators' task included: (i) annotating the first list and completing the accompanying questionnaire (see example in Table 5.11), (ii) discussing the annotation of the first list with the researcher and completing a short training session on tagset use, (iii) annotating the second list and completing the accompanying questionnaire, and (iv) completing a final questionnaire about the whole task (see the task and questionnaires in Appendix E.2). Asking annotators to complete the

training after the first list allowed the researcher to distinguish the value of the training by measuring the difference between the annotations of both lists.

Table 5.11: Examples from the first list with its questionnaire

Example	Tag	Did you find the suitable tag easily?			
		Very easily found	Somewhat easily found	Found with difficulty	Not found
1	وضع أخي الصورة في جهازه المحمول				
2	وقتح النمر فمهو بشكل مخيف				
3	لم أحب الذهاب إلى في هناك				
4	قضينا فيها عدة أيام رائع				
5	يعلمون أبنائهم في مدارس				

After the annotator completed the first list, training began with a discussion about any difficulties or ambiguity in annotating the examples on the list. The discussion did not affect the annotations already made to the first list. The training session included examples for practical tagging of errors that seem to match more than one error category. Further information on how to deal with these errors is also included in the ETMAr.

5.6.4 Results

The tags of each list were converted into their numbers on the tagset list, from 1 to 34, and those cases that were untagged by the annotators were coded as 0. Inter-annotator agreement was measured between each pair of annotators using two methods. First, the number of observed agreement cases between two annotators was divided by the total examples (200), which yielded an average of 176 cases of agreement (88%) for all the pairs of annotators. The second method was to apply the *Cohen's Kappa* measure, which gave an average of $k = 0.877$ ($p < 0.001$) among all the pairs as well (Table 5.12). The level of agreement between T3 and T5 was higher than the others because T4 left 11 cases with no tags, which negatively affected T4's agreement with the other annotators.

Table 5.12: Inter-annotator agreement in both lists of the second evaluation

Annotators	No. of agreement cases (out of 200)	Percent*	<i>Cohen's Kappa</i>	Sig.
T3 & T4	173	87%	0.860	$p < 0.001$
T4 & T5	173	87%	0.860	$p < 0.001$
T3 & T5	183	92%	0.912	$p < 0.001$
Average	176	88%	0.877	

* Number of agreement cases divided by the total cases

Inter-annotator agreement was also measured between the annotators for each list. The results showed the clear positive influence of training. The average of agreement cases increased from 87 on the first list to 89 on the second list; in addition, the *Cohen's Kappa* measure increased from $k = 0.869$ to $k = 0.886$. The significance value remained stable at $p < 0.001$ for each pair of annotators (Table 5.13).

Table 5.13: Inter-annotator agreement in both lists of the second evaluation

List	Annotators	No. of agreement cases (out of 100)	Percent	<i>Cohen's Kappa</i>	Sig.
First list	T3 & T4	85	85%	0.844	$p < 0.001$
	T4 & T5	86	86%	0.855	$p < 0.001$
	T3 & T5	91	91%	0.907	$p < 0.001$
	Average	87	87%	0.869	
Second list	T3 & T4	88	88%	0.876	$p < 0.001$
	T4 & T5	87	87%	0.865	$p < 0.001$
	T3 & T5	92	92%	0.917	$p < 0.001$
	Average	89	89%	0.886	

The evaluation form asked the annotators to answer the question, “Did you find the suitable tag easily?” by selecting one of four responses after tagging each error. The responses showed that 94.5% of the tags were *Very easily found*, 3.3% were *Somewhat easily found*, 0.3% were *Found with difficulty*, and 1.8% were *Not found*.

All of the annotators selected the choice *Very easily found* (T3 = 187, T4 = 184, and T5 = 196) and *Somewhat easily found* (T3 = 13, T4 = 5, and T5 = 2). However, T4 selected *Not found* 11 times; similarly, T5 selected *Found with difficulty* twice (Figure 5.22).

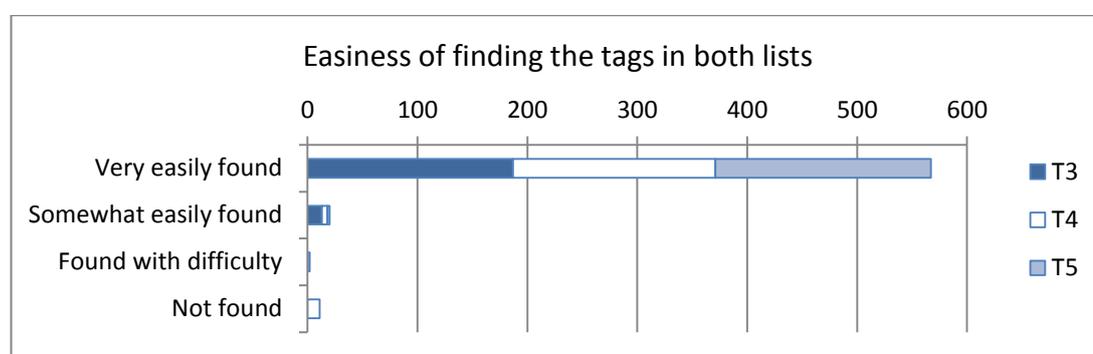


Figure 5.22: Annotators’ responses to the question about easiness of finding the tags

The easiness of finding the tags was also calculated for each list separately, which may reflect the effect of training. The results revealed that the percentage of those tags that were found too easily increased from 46.7% in the first list to 47.8% in the second. Percentages of option 2 and option 3 decreased, while only option 4 increased due to the responses of T4 (Table 5.14).

Table 5.14: The potential impact of training on the ease of finding the tags

	Very easily found	Somewhat easily found	Found with difficulty	Not found	Total
List 1	93.4%	4.6%	0.6%	1.4%	100.0%
List 2	95.6%	2.0%	0.0%	2.4%	100%
Average	94.5%	3.3%	0.3%	1.9%	100%

The final questionnaire about the task and the questionnaire itself showed highly positive responses as illustrated in Table 5.15. The questions are in bold and italic font, followed by the choices, and then the number of responses to each choice in the shaded cells.

Table 5.15: Responses to the final questionnaire

<i>1. Are the error labels clear and easily understood?</i>				
Appropriate and do not need more clarification	Need some clarification	Ambiguous and need to be fully clarified		
3	0	0		
<i>2. Is the division of error categories clear and understandable (6 categories)?</i>				
Yes	To some extent	No		
3	0	0		
<i>3. Is the division of error types clear and understandable (34 types)?</i>				
Yes	To some extent	No		
3	0	0		
<i>4. How easy and fast is selecting the suitable tag?</i>				
It can be selected easily and quickly	It requires some time to be selected	It requires a long time to be selected		
2	1	0		
<i>5. How suitable is the tagset in general for errors in Arabic?</i>				
It is OK	It requires some modifications	It is completely unsuitable		
3	0	0		
<i>6. Please provide your general opinion about this questionnaire.</i>				
It is OK	It requires some modifications	It is completely unsuitable		
3	0	0		
<i>7. What do you think about the methodology used to evaluate the error tagset in this questionnaire?</i>				
Excellent	Good	Acceptable	Poor	Unsuitable
3	0	0	0	0
<i>8. What do you think about the number of error examples used (200 examples)?</i>				
Excellent	Good	Acceptable	Poor	Unsuitable
2	1	0	0	0
<i>9. What do you think about the ease of finding errors tags (after tagging each error)?</i>				
Excellent	Good	Acceptable	Poor	Unsuitable

1	2	0	0	0
10. What do you think about the design of the “Error Tagging Manual for Arabic”?				
Excellent	Good	Acceptable	Poor	Unsuitable
3	0	0	0	0
11. What do you think about the comprehensiveness of the information in the “Error Tagging Manual for Arabic”?				
Excellent	Good	Acceptable	Poor	Unsuitable
3	0	0	0	0
12. What do you think about the clarity of the explanations in the “Error Tagging Manual for Arabic”?				
Excellent	Good	Acceptable	Poor	Unsuitable
3	0	0	0	0

The final questionnaire included a part for annotators to evaluate the tagset. Questions in this part were similar to those given to the evaluators (i.e. error types to be added, error types to be deleted, error types to be changed, error types to be integrated, and error types to be split). The annotators gave their comments after completing the second evaluation, so they were considered in the third version of the ETAr (see Section 5.7 for a discussion of the refinement of this version).

5.6.5 Limitations and Suggestions

This evaluation does not provide insight into the authentic distribution of the tagset on a corpus sample, as the annotators were given two lists of examples where errors were identified and equally distributed. Based on these limitations, the researcher decided to use a number of entire texts in the third experiment. In addition, errors in this sample will not be pre-defined in order to measure the distribution of the tagset from the annotators’ view. The third experiment is described in Section 5.7 below.

5.7 Third Evaluation: ETAr Distribution and Inter-Annotator Agreement

The primary aim of this experiment was to measure the distribution of the ETAr on a number of ALC texts instead of error examples. Errors in this sample were not pre-identified, which may help to measure the distribution based on the annotators’ error

identification. The second aim was to measure the inter-annotator agreement of version 3 of the ETAr which was refined based on the annotation standards and the annotators' suggestions.

5.7.1 Refining the Tagset

The ETAr was refined based on the annotators' suggestions in the second evaluation as well as the annotation standards that the researcher specified at this stage for standardising the format of the annotation files (the annotation standards have been described in Section 5.3.1). The refinement included removing those tags used for multi-word annotations such as *Word order*, *Wrong context of citation from Quran or Hadith*, and the entire category of *Style*. It also involved the modification of previously single- and multi-word annotations to cover only single words; for example, *Word/phrase selection* became *Word selection*, *Redundant word(s)* became *Redundant word*, and *Missing word(s)* became *Missing word*. The modifications also included adding *Yā'* to the type *Confusion in 'alif Mutatarrifa*, resulting in *Confusion in 'alif and Yā' Mutatarrifatain*. Additionally, 13 error types were renamed for more clarity. The third version of the ETAr is shown in Table 5.16.

Table 5.16: Third version of the ETAr

Error Category	Error Type	Arabic tag	English tag
1. Orthography الإملاء 'l'imlā'	1. Hamza (ء، أ، إ، و، ي، ث)	<إه>	<OH>
	2. Confusion in <i>Hā'</i> and <i>Tā'</i> <i>Mutatarrifatain</i> الخطأ في الهاء والتاء المتطرفتين (هـ، ع، ت)	<إة>	<OT>
	3. Confusion in <i>'alif</i> and <i>Yā'</i> <i>Mutatarrifatain</i> الخطأ في الألف والياء المتطرفتين (ا، ي، ي)	<إى>	<OA>
	4. Confusion in <i>'alif Fāriqa</i> (كتبوا)	<إت>	<OW>
	5. Confusion between <i>Nūn</i> (ن) and <i>Tanwīn</i> (ٍٍٍ) الخلط بين النون والتنوين	<إل>	<ON>
	6. Shortening the long vowels تقصير الصوائت الطويلة (اوي → ٍٍٍ)	<إف>	<OS>
	7. Lengthening the short vowels (اوي → ٍٍٍ) تطويل الصوائت القصيرة	<إق>	<OG>
	8. Wrong order of word characters الخطأ في ترتيب الحروف داخل الكلمة	<إط>	<OC>
	9. Replacement in word character(s) استبدال حرف أو أحرف من الكلمة	<إس>	<OR>

	10. Redundant character(s) زيادة حرف أو أكثر	<إز>	<OD>
	11. Missing character(s) نقص حرف أو أكثر	<إن>	<OM>
	12. Other orthographical errors أخطاء إملائية أخرى	<إخ>	<OO>
2. Morphology الصرف 'ssarf	13. Word inflection الخطأ في اختيار بنية الكلمة المناسبة	<صص>	<MI>
	14. Verb tense الخطأ في زمن الفعل	<صز>	<MT>
	15. Other morphological errors أخطاء صرفية أخرى	<صخ>	<MO>
3. Syntax النحو 'nnaḥw	16. Case الخطأ في الإعراب	<نص>	<XC>
	17. Definiteness الخطأ في التعريف والتذكير	<نع>	<XF>
	18. Gender الخطأ في الجنس (التذكير والتأنيث)	<نذ>	<XG>
	19. Number (singular, dual, and plural) الخطأ في العدد (الإفراد والتثنية والجمع)	<نف>	<XN>
	20. Redundant word كلمة زائدة	<نز>	<XT>
	21. Missing word كلمة ناقصة	<نن>	<XM>
	22. Other syntactic errors أخطاء نحوية أخرى	<نخ>	<XO>
4. Semantics الدلالة 'ddalāla	23. Word selection الخطأ في اختيار الكلمة المناسبة	<دب>	<SW>
	24. Faṣl wa Waṣl (confusion in use/non-use of conjunctions) الخطأ في الفصل والوصل (الخطأ في استخدام أدوات العطف)	<دف>	<SF>
	25. Other semantic errors أخطاء دلالية أخرى	<دخ>	<SO>
5. Punctuation علامات الترقيم 'alāmāt 't-tarqīm	26. Punctuation confusion علامة ترقيم خاطئة	<تط>	<PC>
	27. Redundant punctuation علامة ترقيم زائدة	<تز>	<PT>
	28. Missing punctuation علامة ترقيم مفقودة	<تن>	<PM>
	29. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<تخ>	<PO>

5.7.2 Sample and Annotators

The target size of the sample in the third experiment was 10,000 words. The larger sample size in comparison with the previous two experiments was intended to make it possible to measure the extent to which the error types in the ETAr are distributed on this sample. For this purpose, 20 texts were selected randomly among those texts having a length between 400 and 600 words, which totalled a sample of 10,031 words. Two annotators (T3 and T6) participated in this experiment in addition to the researcher. See Table 5.17 below for more details about the annotators.

Table 5.17: Annotators who participated in the third evaluation of the ETAr

	T3	T6
Qualifications	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Applied Linguistics 	<ul style="list-style-type: none"> • First degree in Arabic Linguistics • Master degree in Applied Linguistics
Experience in teaching Arabic	Teaching Arabic to non-native Arabic speakers in Saudi Arabia for a few years	Teaching Arabic to non-native Arabic speakers in Saudi Arabia for a few years
Experience in error annotation	Participated in the second evaluation	No previous experience

5.7.3 Task and Training

Each annotator was required to manually complete three basic steps for each error in the experiment sample:

1. Underline the token including an error,
2. Add the most appropriate tag that matched the error type using the ETAr and its manual, and
3. Suggest the correct form for each error.

For instance, with the word “الى”, which includes an error in *Hamza*, the annotator must underline it, assign the tag *OH* to it, and correct it to “إلى”; see an example of output in Figure 5.23. Due to the large sample, the annotators were allowed a few weeks to finish the annotation task.



Figure 5.23: Example of the error annotation method in the third evaluation

As a part of the experiment plan, a training session was conducted with each annotator in order to familiarise them with the method required for the annotation. At the beginning of the session, which lasted between 2 and 3 hours, each annotator received an explanation about the following points:

1. the purpose of this experiment;
2. the error types included in the ETAr;
3. the Error Tagging Manual for Arabic; and
4. an annotated example showing the form of the output expected.

Each annotator then was asked to do an error annotation test on a sample text, which was not from the 20 texts of the experiment sample. The annotator and the researcher discussed the annotation of this testing text both within and after the annotation process. The discussion primarily centred on how to select the most appropriate tag for each error following the rules in the tagging manual.

5.7.4 Distribution of the ETAr

For the first aim of this experiment, to measure the distribution of the ETAr, the analysis started by extracting the distribution of the error tags by each annotator independently (Figure 5.24). The average of this use revealed that the most used tags were *Missing punctuation* in the Punctuation category (397) and *Hamza* (338) in Orthography followed by *Word selection* in Semantics (126) and *Punctuation confusion* in Punctuation (119). In contrast, the least used tags were *Other errors in punctuation*, *Other semantic errors*, and *Other morphological errors* (no assignments for each), which may indicate that tags under those categories covered

all possible errors in the sample. The other categories may need more investigation, particularly Syntax as the type *Other syntactic errors* had an average use of 13. However, in most of these cases, the annotators explained that the error was in word order. The error type *Word order* was removed from the third version of the ETAR based on the annotation standards which do not cover multi-word annotation at this stage. It might be re-considered in later stages when adding further layers of annotation.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Origin	R	T3	T6				Tag	R	T3	T6		Total use	Average use	
2	رحلتني							OH	361	347	307		1015	338	
3	إلي							OT	92	78	80		250	83	
4	بذني							OA	12	7	7		26	9	
5	خاتل							OW	5	4	3		12	4	
6	إجازة							ON	2	1	1		4	1	
7	الحج							OS	3	8	0		11	4	
8	لغاً							OG	3	4	0		7	2	
9	وصلت							OC	18	10	9		37	12	
10	إلي							OR	67	36	47		150	50	
11	المسودة							OD	35	19	47		101	34	
12	هي							OM	68	41	53		162	54	
13	ها							OO	2	5	3		10	3	
14	الفصل							MI	27	42	28		97	32	
15	كنت							MT	14	6	4		24	8	
16	عازماً							MO	0	0	1		1	0	
17	عاني	OR		OR				XC	72	45	39		156	52	
18	تأجيل							XF	91	87	77		255	85	
19	الفصل							XG	55	47	34		136	45	
20	ورجوعي			MI				XN	14	18	12		44	15	
21	إلي							XT	103	105	76		284	95	
22	بذني	PM		PM		PM		XM	94	92	52		238	79	
23	بان							XO	11	21	8		40	13	
24	زوجتي							SW	143	127	108		378	126	
25	كان	XG		XG		XG		SF	59	52	93		204	68	
26	عطي							SO	0	0	0		0	0	
27	وتلك							PC	48	182	128		358	119	
28	الولادة			PC				PT	11	41	5		57	19	
29	ولم							PM	458	127	605		1190	397	
30	يكن							PO	0	0	0		0	0	
31	وعها	OR		OR		OR									
32	أعدت			PC											
33	لكني							TOTAL	1868	1552	1827		5247	1749	
34	فرجت														

Figure 5.24: Extracting the tags used by each annotator in the third evaluation

After analysing the inter-annotator agreement, the researcher extracted the distribution of only those tags which had been used with agreement either between two annotators (partial agreement) or all annotators (full agreement) on the same error (Table 5.18).

Table 5.18: Distribution of the tags' use and agreement by the annotators

Tag	Instances of use				Instances of Agreement (between 2 or 3 annotators)			
	R	T3	T6	Average	R	T3	T6	Average
OH	361	347	307	338	344	337	303	328
OT	92	78	80	83	88	75	79	81
OA	12	7	7	9	10	6	4	7
OW	5	4	3	4	5	4	3	4
ON	2	1	1	1	2	1	0	1
OS	3	8	0	4	3	7	0	3
OG	3	4	0	2	3	4	0	2
OC	18	10	9	12	13	9	9	10
OR	67	36	47	50	56	31	43	43
OD	35	19	47	34	28	18	36	27
OM	68	41	53	54	51	37	42	43
OO	2	5	3	3	1	5	3	3
MI	27	42	28	32	21	27	23	24
MT	14	6	4	8	7	5	2	5
MO	0	0	1	0	0	0	0	0
XC	72	45	39	52	53	40	33	42
XF	91	87	77	85	73	72	57	67
XG	55	47	34	45	39	38	27	35
XN	14	18	12	15	13	14	8	12
XT	103	105	76	95	60	53	38	50
XM	94	92	52	79	59	54	28	47
XO	11	21	8	13	6	10	8	8
SW	143	127	108	126	85	89	55	76
SF	59	52	93	68	41	28	40	36
SO	0	0	0	0	0	0	0	0
PC	48	182	128	119	40	74	83	66
PT	11	41	5	19	7	18	3	9
PM	458	127	605	397	336	110	333	260
PO	0	0	0	0	0	0	0	0
Total	1868	1552	1827	1749	1444	1166	1260	1290

The average of agreement was quite similar to the average of use in tags under the *Orthography* and *Morphology* categories. The possible interpretation of this finding is that errors under those categories were usually related to the word form; however, when the error was related to sentence structure and meaning (i.e. syntactic,

semantic, and punctuation errors), the annotators had different views. Consequently, the gap emerged between tag use and inter-annotator agreement (Figure 5.25).

The distribution of the ETAr tags may be fundamental material for a deeper linguistic investigation about the reasons behind those most common errors in Arabic. It may lead to some suggested solutions as well as different designs of teaching materials which focus on those solutions.

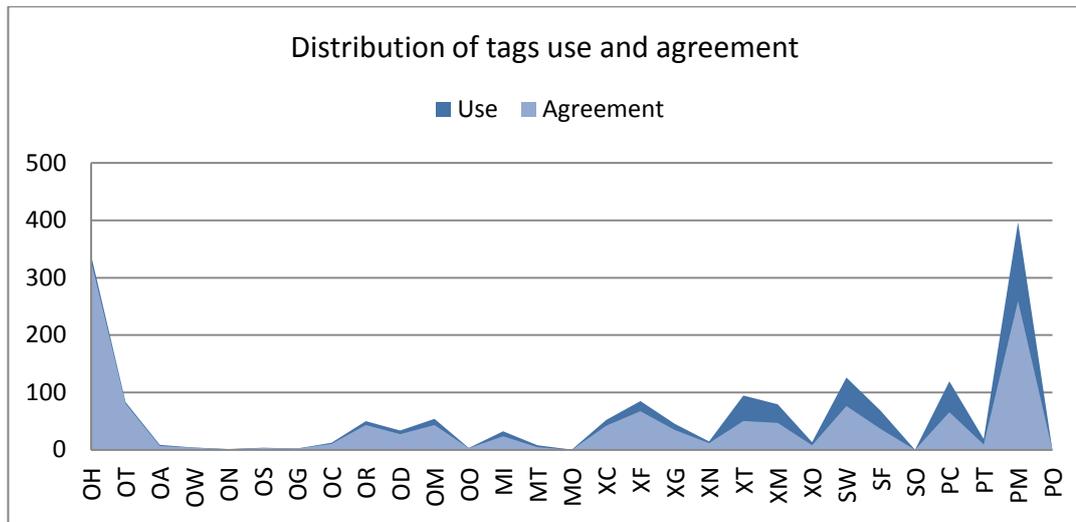


Figure 5.25: Differences in the distribution of tags use and agreement

5.7.5 Inter-Annotator Agreement

The second aim of the experiment was to measure the inter-annotator agreement of version 3 of the ETAr. The agreement was measured on those error cases that annotators detected similarly in order to know to what extent they assigned the same tag to these errors. The annotators R and T3 had detection agreement in 1061 errors; of these, they assigned the same tag to 908 errors and different tags to 153 errors. The observed agreement was 86%, resulting in a weighted *Cohen's Kappa* value of $k = 0.811$ ($p < 0.001$). The results showed that R and T6 had the highest number of error detection agreements (1153); similarly, they assigned the same tags in the highest number of cases (1023). The observed agreement was 89% with a weighted *Cohen's Kappa* value of $k = 0.842$ ($p < 0.001$). In contrast, T3 and T6 had the lowest number of error detection agreements (881); likewise, they assigned the same tags to the lowest number of cases (732). The observed agreement was 83% with a weighted *Cohen's Kappa* value of $k = 0.771$ ($p < 0.001$). The average weighted

Cohen's Kappa of the three groups (i.e. R & T3, T3 & T6, and R & T6) was $k = 0.808$.

In terms of those cases where all annotators identified the same errors (full agreement), the annotators had detection agreement in 757 errors, while they assigned the same tag to 624 of them. The observed agreement is 82% (Table 5.19).

It can be seen from the distribution of tags use and agreement (Table 5.18 and Figure 5.25) that the annotators' disputes is the most likely reason behind not achieving higher scores than which have been achieved, as most of the disagreements were in sentence structure and meaning where it is possible to have different views to the error nature. However, when the error was related to the word form, the agreement was very high, for example the agreement was 100% in the errors OW (Confusion in *'alif Fāriqa*), ON (Confusion between *Nūn* and *Tanwīn*), OG (Lengthening the short vowels); also 98% in OT (Confusion in *Hā'* and *Tā' Mutatarrifatain*) and 97% in the OH error (*Hamza*). See Table 5.18 for more details about the distribution of the tags agreement on the error types. Performing further investigation and improvement on the error categories that have less agreement may assist in achieving higher inter-annotator agreement results for these categories.

Table 5.19: Inter-annotator agreement in the third evaluation

	R & T3	T3 & T6	R & T6	All annotators
Agreement in errors detected	1061	881	1153	757
Agreement in tags assigned	908	732	1023	624
Observed agreement	86%	83%	89%	82%
Average		86%		
<i>Cohen's Kappa</i>	0.811 ($p < 0.001$)	0.771 ($p < 0.001$)	0.842 ($p < 0.001$)	
Average		0.808		

Comparing the results of inter-annotator agreement in the three evaluation experiments reveals the influence of some factors that may play a role in achieving higher results in the second ($k = 0.877$) and third evaluations ($k = 0.808$) compared to the first experiment ($k = 0.468$). The first factor is the training that the annotators received in the second and third experiments, which emphasises the importance of

such training in error annotation. The second factor may be the review of the ETAr by two experts in the Arabic language, as this may have helped in clarifying the error types to the tagset users. The third factor is the use of the Error Tagging Manual of Arabic which explains the errors in the Error Tagset of Arabic and provides rules to follow for selecting the appropriate tags. The Error Tagging Manual of Arabic is described and evaluated in the following section.

5.8 Error Tagging Manual for Arabic (ETMAr)

The main aim of developing the ETMAr is to provide users of the ETAr with clear instructions on how to identify errors and select the most appropriate tags for them. Such a manual can be used to enhance error tagset use and understandability. The evaluation of the ETAr showed that error tagging was more accurate using the instructions provided in the ETMAr.

As the ETMAr is intended to be used by a worldwide audience interested in the Arabic language, the manual includes all the information, instructions, rules, and examples in two languages: Arabic as the target language and English as the international language. Additionally, the English part includes phonetic descriptions of the Arabic examples using the DIN 31635 (Deutsches Institut für Normung, 2011) standard for the transliteration of the Arabic alphabet.

5.8.1 Purpose

The ETMAr performs two functions. It explains errors in the ETAr with examples, and it provides users with rules to follow for selecting the appropriate tags in error annotation. The ETMAr provides information about each error type in the third version of the ETAr. This information covers the definition of the error type, its scope, and forms of errors expected with the corrections suggested. An attempt has been made to accommodate all possible error forms under each type in order to provide their appropriate corrections.

The ETMAr provides a number of rules for error tagging. The aim of these rules is to help annotators to identify ambiguous instances and to select the most appropriate tags for these cases. One of these rules, for example, states that choosing an error category should be based on a specific order (except punctuation), starting from the highest level (Semantics) to the lowest level (Orthography). This rule was

established because testing the tagset showed that, when two categories are applicable to one error, the higher one is usually the most appropriate unless there is a clear reason for the opposite. The ETMAr presents some examples of exceptions with reasons and explanations as to why they are exceptions and how they are annotated.

5.8.2 Evaluation

As seen in the evaluation of the ETAr, the observed inter-annotator agreement was increased from 52% ($k = 0.468$) in the first evaluation to 88% ($k = 0.877$) in the second evaluation where the ETMAr was used for the first time, and to 86% ($k = 0.808$) in the third evaluation where the ETMAr was used for the second time. In addition, the second evaluation involved a questionnaire that posed three questions to the annotators about the ETMAr:

1. What do you think about the design of the “Error Tagging Manual for Arabic”?
2. What do you think about the comprehensiveness of the information in the “Error Tagging Manual for Arabic”?
3. What do you think about the clarity of the explanations in the “Error Tagging Manual for Arabic”?

The annotators all chose the highest ratings among five choices given: *Excellent*, *Good*, *Acceptable*, *Poor*, and *Unsuitable* (see Table 5.15 in Section 5.6.45.6.4).

5.9 Conclusion

This chapter describes three elements of the ALC project: the Computer-aided Error Annotation Tool for Arabic (CETAr), the Error Tagset of Arabic (ETAr), and the Error Tagging Manual for Arabic (ETMAr). The CETAr includes a number of features for facilitating the manual annotation process such as text tokenisation, smart-selection, and auto tagging. The evaluation of consistency and speed in the CETAr showed that the annotation time was reduced while consistency in annotation was increased when using this tool; based on the results, the smart-selection feature may play a role in this achievement. Additionally, evaluating the auto-tagging feature revealed accuracy levels between 76% and 95% with an average of 88%.

The ETAr was developed as an error taxonomy and tagset for tagging errors in Arabic texts. The third version of this tagset includes 29 types of errors under 5 categories. Two evaluators and seven annotators have evaluated the ETAr a total of three times for a number of purposes. The first purpose of the evaluations was to determine the extent to which the ETAr could be understood and usable against another tagset. The results of this evaluation showed that the ETAr achieved an observed agreement of 52% ($k = 0.468$) compared to 33% ($k = 0.292$) by the ARIDA tagset. The second purpose was to measure the inter-annotator agreement, and the results revealed that the observed agreement increased from 52% ($k = 0.468$) in the first evaluation to 88% ($k = 0.877$) in the second and 86% ($k = 0.808$) in the third. The third aim was to evaluate the value of training the annotators; while no training was given in the first evaluation, results of the second and third experiments emphasised the importance of such training in error annotation.

The fourth purpose was to measure the distribution of the ETAr tags on a sample of the ALC. *Missing punctuation* and *Hamza* were the most used tags, with an average use of 397 and 338 uses respectively. In contrast, the tags *Other errors in punctuation*, *Other semantic errors*, and *Other morphological errors* were not used at all. In categories such as Orthography and Morphology where errors usually relate to the word form, the average of tag agreement was quite similar to the average of tag use. However, a gap emerged between tag agreement and tag use under the Syntax, Semantics, and Punctuation categories where the annotators may have different views of the contexts.

The fifth goal was to measure the value of using the ETMar, which was developed for two main functions: to explain the error type and to establish the rules for how to select the appropriate tags in error annotation. The ETMar was used in the second and third evaluations of the ETAr, with the result that the observed inter-annotator agreement increased from the first evaluation to the second and third evaluations as mentioned above. Additionally, the annotators' responses to the questions about the ETMar in the second evaluation's questionnaire were highly positive, with all annotators selecting "Excellent" among the five scores in the rating scale (i.e. Excellent, Good, Acceptable, Poor, and Unsuitable) for all questions.

To sum up, nine people evaluated the CETAr, ETAr, and ETMar for annotating Arabic errors, and the results achieved in the experiments have been positive.

Additionally, these results highlight the value of these novel contributions that present the most comprehensive system for error annotation in Arabic.

6 Web-Based Tool to Search and Download the ALC

Chapter Summary

This chapter introduces the first version of a free-access, web-based tool developed for searching and downloading the ALC data. The tool was developed to help users search the ALC or a subset of its data and download the source files of any sub-corpus based on a number of determinants. It has an interface in Arabic and English including translations of labels and buttons, as well as the ability for the entire website layout to be right-to-left. In addition, a user guide was also created in both Arabic and English to give an overview of the tool and to illustrate its use. The dynamic functions of the ALC Search Tool allows the data to be retrieved and the results updated quickly. The database of the ALC Search Tool can be fed with additional corpus data in the future, which will be immediately available to the users for searching and downloading.

The accuracy of the output of the ALC Search Tool was evaluated based on two aspects: Recall and Precision. The accuracy was extracted based on the values of precision, recall, and F-measure of two types of searches: the normal search function and the Separate Words option. The evaluation shows that the normal search achieved a high value in terms of recall while the Separate Words option achieved a high value in terms of precision. Additionally, both options achieved a high result in F-measure. A number of specialists in computer science, linguistics, and applied linguistics have participated in further evaluation of this tool through a questionnaire. Their feedback was highly positive with valuable comments and suggestions to improve its functionality in the future. The website's statistics have also shown that the website received more than 50,000 visits in its first four months.

6.1 Introduction

Creating a corpus provides a valuable source of data for research. However, creating an analysis tool increases the usefulness level of the data source. Many analysis tools such as Khawas (Althubaity *et al.*, 2013, 2014), aConCorde (Roberts, 2014; Roberts *et al.*, 2006), AntConc (Anthony, 2005, 2014a, 2014b), and WordSmith Tools (Scott, 2008, 2012) focus on the statistical tests that can be done on the corpus data. However, few tools use the corpus metadata as determinants when searching the corpus such as Sketch Engine (Kilgarriff, 2014; Kilgarriff *et al.*, 2004).

For instance, the ALC corpus includes 26 elements in its metadata such as “age”, “nationality”, and “gender”. Searching a specific group of ages or nationalities, or comparing males to females, may require manually splitting the data based on the factors needed. The need to search the data based on more than one factor means more effort to consider those factors when splitting, uploading, and searching each sub-corpus.

To resolve this problem, the idea of creating the ALC Search Tool emerged. It uses the 26 elements of the ALC metadata as determinants to facilitate searching the corpus data or any sub-corpus. In addition, it enables users to download the source files of the corpus or a subset of those files in different formats (TXT, XML, PDF, and MP3), so those subsets can be used with external tools with no need for manual splitting. This chapter presents a description of the first version of this tool including its purpose, design, and functions (search and files download) and concludes with an evaluation.

6.2 Review of Tools for Searching and Analysing Arabic Corpora

A number of tools exist for searching and analysing Arabic corpora. Choosing a suitable tool for supporting Arabic seems to be difficult and requires a comparison between multiple tools, as their potentials and functions differ in terms of handling Arabic. This review attempts to present a fundamental comparative evaluation of six tools that are described as supporting multiple languages including Arabic. The purpose of this review is to evaluate those tools which allow searching and analysing Arabic corpora including the ALC.

The tools that are used for searching and analysing corpora generally provide some basic functions (e.g., frequent words and concordances), whereas some of these tools have more functions and statistics such as collocations, n-gram/clusters, keywords, etc. A number of these search and analysis tools are web-based, e.g., The Sketch Engine (Kilgarriff, 2014; Kilgarriff *et al.* 2004), IntelliText Corpus Queries (Sharoff, 2014; Wilson *et al.* 2010), so in order to use them, researchers need to remain online. Other tools are PC-based, so they can be downloaded on computers and used offline, such as the KACST Arabic Corpora Processing Tool “Khawas” (Althubaity *et al.* 2013, 2014), aConCorde (Roberts, 2014; Roberts *et al.* 2006), AntConc (Anthony, 2005, 2014a,b), WordSmith Tools (Scott, 2008, 2012).

The websites, manuals, or other resources of these tools indicate that Arabic is one of the languages supported; therefore, we included the newest versions of these tools in this review. Additionally, it seems that those tools aforementioned – both web-based and PC-based – handle written corpora only unlike auditory signals. However, similarly to handling written corpora, those tools may support searching transcriptions of spoken corpora including typed sequence of phonetic symbols or spoken syllables if they are in a written format.

Previous surveys have reviewed concordance tools but not specifically for Arabic corpora, for example Wiechmann and Fuhs (2006) reviewed ten corpus concordance programs tested on English corpora. Other surveys have covered Arabic text analysis resources, for example Atwell *et al.* (2004) reviewed a sample of tools for Arabic morphological analysis and part-of-speech tagging, machine-readable dictionaries, and corpus visualization tools as well as concordancing. Thus, there is need for a survey focused on Arabic corpus search and processing tools that support features of the Arabic language.

6.2.1 Method of Review

In this review, six tools designed to search and analyse corpora were selected to be evaluated against eight criteria. Each of these tools was evaluated separately against each benchmark. The evaluation was repeated, with the second one conducted two months after the first, on the same tool versions used in the first evaluation, in order to be sure that the criteria were properly covered. One of the tools was not available in the first evaluation, but the opportunity was taken to include it in the second. A

sample of Arabic corpus texts was used in two formats, UTF-8 and UTF-16. More details about the evaluation method appear in the following sections.

6.2.2 Tools Investigated

This review includes six tools:

1. The KACST (King Abdulaziz City for Science and Technology) Arabic Corpora Processing Tool “Khawas” 3.0 (Althubaity and Al-Mazrua, 2014; Althubaity *et al.* 2013)
2. aConCorde 0.4.3 (Roberts, 2014; Roberts *et al.*, 2006)
3. AntConc 3.4.0 (Anthony, 2005, 2014a,b)
4. WordSmith Tools 6.0 (Scott, 2008, 2012)
5. The Sketch Engine (Kilgarriff 2014; Kilgarriff *et al.*, 2004)
6. IntelliText Corpus Queries (Sharoff, 2014; Wilson *et al.*, 2010)

As mentioned previously, the tools selected were designed to support Arabic along with other languages.

6.2.3 Evaluation Criteria

Given the fact that functions of the tools examined here differ from one to the next, most of the criteria used were based on linguistic features, particularly those related to Arabic. While many benchmarks could be examined in an evaluation of these tools, eight points were selected that seemed to be the most essential criteria for searching and analysing Arabic corpora¹. Wiechmann and Fuhs (2006) reviewed ten corpus concordance programs; they mainly used general software evaluation criteria such as: platform, price, ease of installation, help, and performance. They also compared a range of functionalities, such as: input/output formats, text search, frequency and collocation outputs. However all but one of the systems they evaluated were developed for English text, and they did not investigate in detail how well the systems adapted to corpora in other languages such as Arabic. There was one exception: aConCorde was explicitly targeted at Arabic.

¹ Further criteria can be added in future evaluations, for example using Regular Expression and wildcards – which is supported by some of those tools – for searching Arabic corpora.

6.2.3.1 Reading Arabic Text Files in UTF-8 Format

This point examines whether the tools being tested are able to read Arabic text files in UTF-8 format and show the characters correctly. According to Burnard (2005), the Unicode Standard has three UTFs: UTF-16, UTF-8 and UTF-32 (in chronological order). He indicates that UTF-16 is known in Microsoft applications as “Unicode”, and demonstrates that UTF-8 is superior to the other two, as UTF-16 and UTF-32 are more complex architecturally than UTF-8. Burnard recommends using UTF-8 as a universal format for data exchange in Unicode, and for corpus construction.

6.2.3.2 Reading Arabic Text Files in UTF-16 Format

This is to examine whether the tools are able to read Arabic text files in UTF-16 format and show the characters correctly. UTF-16 is one of the formats Microsoft applications use to save files containing characters in Unicode format. Notepad is one application in particular upon which many people rely to create and save their corpus files. However, when a user tries to save a text including Arabic characters in different encoding formats such as ANSI, Notepad shows a message about how to keep the Unicode information with an advice to select one of the Unicode options (Figure 6.1). Thus, corpora tools may or may not be able to handle the UTF-16 encoding format besides the UTF-8 format that is most widely used in corpus construction. For this reason the ability to read Arabic characters in UTF-16 was included in this evaluation.

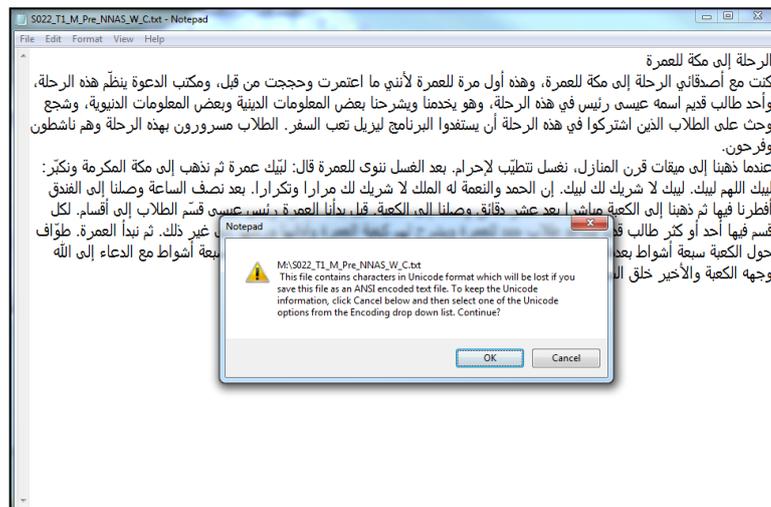


Figure 6.1: A message from Notepad about the file encoding

6.2.3.3 Displaying Diacritics Correctly

Diacritics are small symbols that optionally written above or below a letter “providing a more accurate indication about how a word is pronounced” (Samy and Samy, 2014). There are three types of diacritics, Vowel, Nunation and Shadda:

Vowel diacritics represent Arabic’s three short vowels, Fatha /a/ [ـَ], Damma /u/ [ـُ], Kasra /i/ [ـِ], and the absence of any vowel (no vowel) Sukūn [ـْ].

Nunation occurs only in final position in nominals (nouns, adjectives and adverbs). In addition to helping in the word pronunciation, they indicate indefiniteness as well, Faḥatān [ـً], Ḍammatān /u/ [ـٌ], Kasratān /i/ [ـٍ].

Shadda is a consonant doubling diacritic, it typically combines with a vowel or Nunation diacritic, [ـّ / ـّ].

See Habash (2010, pp 11-12) for more details about these three types and how they are written and pronounced.

The ability to show Arabic diacritics – if there are any – is tested under this point, e.g., “هِمَّةٌ”. Displaying diacritics might be essential in some cases, particularly with similar forms that cannot be distinguished if they have no diacritics, e.g., ذهب (past tense of the verb ‘to go’) and ذهبٌ (noun: ‘gold’).

6.2.3.4 Displaying Arabic Text in the Correct Direction (Right to Left)

As Arabic is written from right to left, the tools were examined to ascertain whether they can show Arabic text in the correct direction, particularly in concordances, where the contexts must also be ordered correctly.

6.2.3.5 Normalising Diacritics

This is to check if the tool is able to normalise the diacritics, so that the user has an option to search Arabic texts which include diacritics using a single word form in the query. For example, if a text includes the word “هِمَّةٌ” (with diacritics) and the word “هِمَّة” (without diacritics), is the user able to search for both using the single form “هِمَّة”? This is significant in searching Arabic corpora, as one form may have several sub-forms with diacritics. Unless the diacritics are normalised, the user may face difficulty in counting them, and accordingly in combining them into a single query.

6.2.3.6 Normalising *Hamza* “ء”

Normalising *Hamza* is similar to the previous benchmark. Here, we check to see whether the tool has the ability to normalise words that have *Hamza*, so the user has an option to search Arabic texts, which include *Hamza* using a single word form in the query. For example, if a text includes the word “إلى” (with *Hamza*) and the word “الى” (without *Hamza*), is the user able to search for both using the single form “الى”?

6.2.3.7 Providing Arabic User Interface

This is to determine whether these tools provide an Arabic user interface for Arabic users, as some researchers may not be able to use a tool should its interface be in a language different from their mother tongue, and thus cannot benefit from its functions.

6.2.3.8 Enabling Users to Upload or Open Their Arabic Personal Corpora

Researchers may desire to use particular Arabic corpora, or even build their own corpora from scratch and use some tools to search and analyse these resources. Therefore, the tools here are examined to see whether they accept external data files.

6.2.4 Evaluation Sample

The current evaluation was based on a sample from the ALC. We randomly selected 8 files from ALC, containing about 4000 words, to be used as a sample of our examination. The evaluation includes testing as to whether Arabic characters can be read in UFT-8 and UTF-16 formats, and since ALC files are already in UTF-16 format, we made an additional copy of the sample in UTF-8.

6.2.5 Khawas¹

The KACST (King Abdulaziz City for Science and Technology) Arabic Corpora Processing Tool “Khawas” (Alhubaity and Al-Mazrui, 2014; Alhubaity *et al.*, 2013) is an open-source tool that Abdulmohsen Alhubaity and his team at KACST developed specifically for processing Arabic language with an Arabic/English interface (Alhubaity and Al-Mazrui, 2014). It is free to download and can provide

¹ Khawas can be downloaded from: <http://sourceforge.net/projects/kacst-acptool>

analysis including frequency lists, concordance N-grams lexical patterns and corpora comparison. Khawas was developed using Java which means it can be run on many operating systems. The developers claim that this tool works with texts from all languages in principle, and it was tested on Arabic, English, and French (Althubaity and Al-Mazrua, 2014).

Khawas was able to read Arabic texts in UTF-8 format; however this was not the case with texts in UTF-16, as nothing readable was displayed. Khawas is set to remove diacritics by default in order to normalise the text, but they can be shown by changing the settings. Consequently, searching the data follows the diacritics settings; i.e. if the diacritics are shown, the search results will include those words that match the query word including its exact diacritics, and the same words with other diacritics will be excluded. Khawas displays words in the correct right to left orientation (Figure 6.2); however, some words or parts of words were missed from concordances when the tool was run on Microsoft Windows (Figure 6.3). All of the missing words appeared when Khawas was run on Mac OS X. This tool has an option to normalise *Hamza*, which enables both those words that have, or should have but are missing *Hamza*, to be included in the search results. Users need to be aware that *Hamza* normalisation means all *Hamzas* will be removed from the texts, so the query word should not include one, otherwise no results will be returned. Khawas has an Arabic/English interface, and this tool was developed to open external data, i.e., users are able to open their personal corpora on Khawas. This tool garnered 7 points out of 8 in the benchmark evaluation (Table 6.1).

الجدول	الملف	الكلمات_السابقة	الكلمة	الكلمات_اللاحقة
	/Users/Abdullah/...	في الاجازه الصيفيه الماضيه وكانت	من	امتم الرحلات اللتي رحلت اليها
	/Users/Abdullah/...	الرحيل اليها اتصلت بين فيها	من	ساده القوم وكبارهم فسرو وفرحو
	/Users/Abdullah/...	فخرجت في طريقى الخامس عشره	من	رمضان متجها الي القرية فوصلت
	/Users/Abdullah/...	بما تعلمت لي وما تعلمته	من	العقيدة الصحيحه من اساتذتي الفضلاء
	/Users/Abdullah/...	وما تعلمته من العقيدة الصحيحه	من	اساتذتي الفضلاء، فبقيت على هذا
	/Users/Abdullah/...	الفضلاء، فبقيت على هذا اكثر	من	حين لحتي عدت الي الدينيه

Figure 6.2: Khawas Shows Arabic words in a right-to-left order



Figure 6.3: Some Arabic words were missed from concordances when Khawas was run on Windows

Table 6.1: Benchmark score of the Khawas tool

Evaluation criteria								Score
1	2	3	4	5	6	7	8	7/8
Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	

6.2.6 aConCorde¹

aConCorde (Roberts, 2014; Roberts *et al.*, 2006) is a free tool which was created by Andrew Roberts in his spare time while he was a PhD student at Leeds University. It is relatively basic in comparison to the others included in this review, as it only provides users with concordances and a word frequency list. However, one of the distinctive features of aConCorde is that it is “[o]riginally developed for native Arabic concordance” (Roberts, 2014) in addition to that “the provision of an Arabic interface. Not only does this provide Arabic translations for all the menus, buttons etc., but even switches the entire application layout to right-to-left” (Roberts *et al.*, 2006, 6).

aConCorde was able to read Arabic texts in both UTF-8 and UTF-16 formats. It also correctly shows Arabic diacritics as well as words in a right-to-left direction (Figure 6.4). However, diacritics and *Hamza* cannot be normalised, so the search results will literally match the query word. aConCorde has an Arabic/English interface, and enables users to open their personal corpora. aConCorde achieved 6 points in this evaluation (Table 6.2).

¹ aConCorde can be downloaded from: <http://www.andy-roberts.net/coding/aconcorde>

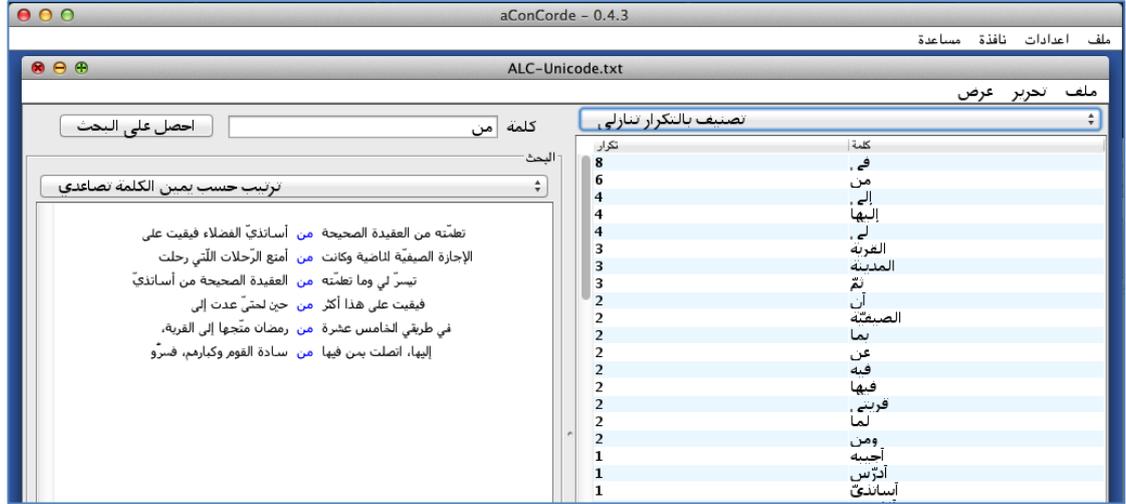


Figure 6.4: Frequency and concordances in aConCorde

Table 6.2: Benchmark score of the aConCorde tool

Evaluation criteria								Score
1	2	3	4	5	6	7	8	6/8
Yes	Yes	Yes	Yes	No	No	Yes	Yes	

6.2.7 AntConc¹

AntConc (Anthony, 2005, 2014a,b) is a free corpus analysis tool developed by Laurence Anthony, a professor in the faculty of science and engineering at Waseda University, Japan. AntConc provides users with concordances, clusters/n-grams, collocates, word list, and keyword list. This tool was “developed in Perl using ActiveState's PerlApp compiler to generate executables for the different operating systems” (Anthony, 2014b, 1). According to AntConc-discussion (2013), Anthony stated that “AntConc 3.2.4 and 3.3.5 were not designed to handle right-to-left languages”, while we evaluated the version 3.4.0 on which he stated that “[i]n the new version coming soon, the graphics engine supports right-to-left languages properly” (AntConc-discussion, 2013).

Although AntConc reads Arabic texts in UTF-8 and UTF-16 formats, it behaves unexpectedly when the user clicks on any of the text words. Diacritics were displayed within the texts; however, AntConc does not normalise diacritics or *Hamza*. Additionally, columns in the concordances screen were shown in the

¹ AntConc can be downloaded from: <http://www.antlab.sci.waseda.ac.jp/software.html>

opposite direction, as the right side should be the left and vice versa (Figure 6.5). AntConc does not provide an Arabic interface, only English is available. Users are able to open their corpora on this tool. AntConc was awarded four of eight points in this benchmark evaluation (Table 6.3).

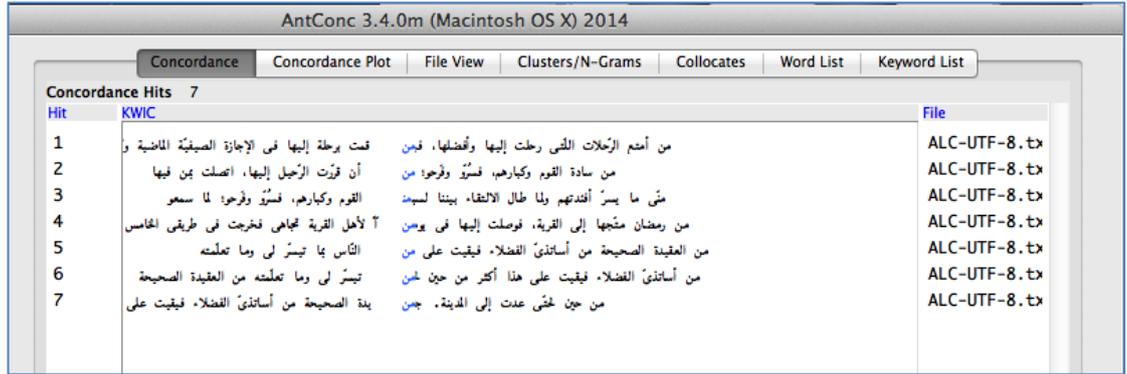


Figure 6.5: Columns of Arabic concordances in AntConc were shown in the opposite direction

Table 6.3: Benchmark score of the AntConc tool

Evaluation criteria								Score
1	2	3	4	5	6	7	8	
Yes	Yes	Yes	No	No	No	No	Yes	4/8

6.2.8 WordSmith Tools¹

WordSmith Tools (Scott, 2008, 2012) is a commercial project developed by Lexical Analysis Software Ltd. The user can download the complete package with no registration code, but it will run in demo mode which will only show a sample of the output. WS Tools are developed for use on Mac, Linux or Windows, with an emulator for Windows. These tools provide users with a word list, concordances, and keywords, and they support many languages, including Arabic. WordSmith Tools even has an Arabic manual²; however, the interface of these tools is only in English. “WordSmith Tools handles a good range of languages, ranging from Albanian to Zulu. Chinese, Japanese, Arabic etc. are handled in Unicode. You can

¹ WordSmith Tools can be downloaded from: <http://www.lexically.net/wordsmith>

² The manual can be accessed here:

http://www.lexically.net/wordsmith/step_by_step_Arabic6/index.html

view word lists, concordances, etc. in different languages at the same time.” (WordSmith Tools, 2013).

WordSmith Tools were able to read Arabic texts in both UTF-8 and UTF-16 formats, and they also display Arabic text correctly in the right-to-left direction. However, WordSmith Tools did not put the diacritics in their correct positions (Figure 6.6). Instead, they are put on small circles, e.g., َ, ُ, ِ or ِ. Diacritics and *Hamza* were not normalised in this tool, so similar words with differences in diacritics and/or *Hamza* will not be retrieved in the results. As mentioned above, WordSmith Tools do not have an Arabic interface, as the only language available is English. Users can open their corpora files on these tools. The evaluation resulted in 4 out of 8 points for WordSmith Tools (Table 6.4).

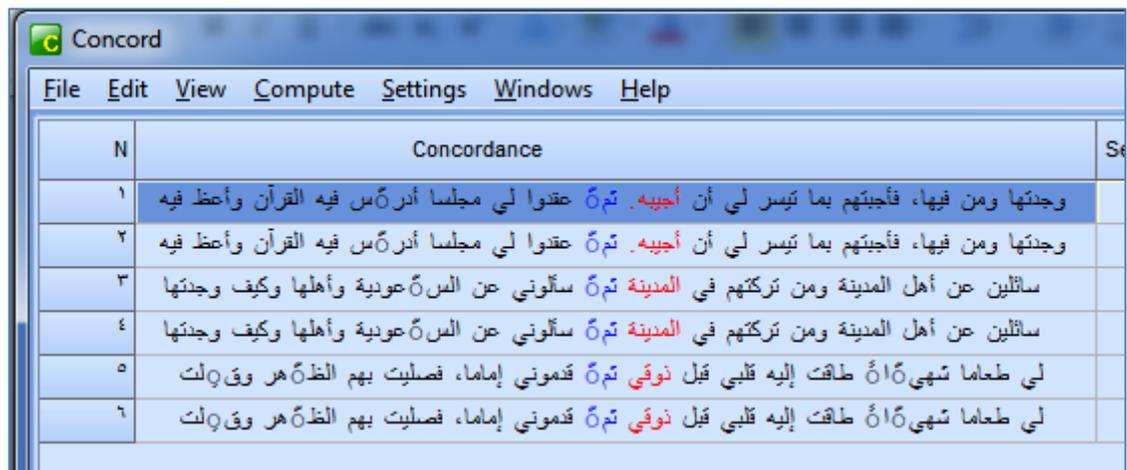


Figure 6.6: Diacritics do not appear in their correct positions in WordSmith Tools

Table 6.4: Benchmark score of the WordSmith Tools

Evaluation criteria								Score
1	2	3	4	5	6	7	8	4/8
Yes	Yes	No	Yes	No	No	No	Yes	

6.2.9 Sketch Engine¹

The Sketch Engine (Kilgarriff 2014; Kilgarriff *et al.*, 2004) is a commercial web-based tool for corpus analysis developed by Lexical Computing Ltd. In addition to

¹ Sketch Engine can be accessed from: <http://www.sketchengine.co.uk>

the corpora searching tool, the users are provided with corpora in many languages including Arabic. Arabic was included in the list of languages supported by Sketch Engine (Sketch Engine, 2014). Along with the usual features of such tools (e.g. concordance, word lists, key words, collocation, and corpus comparison), Sketch Engine has some unique features such as Word Sketches that provide summaries of a word's grammatical and collocational behaviour, Word Sketch Difference to compare and contrast words visually, and WebBootCat, which lets users create specialised corpora from the Web.

Sketch Engine correctly read Arabic texts in both UTF-8 and UTF-16 formats, and displayed Arabic texts in the proper right-to-left direction. Diacritics and *Hamza* were normalised when using the built-in Arabic Segmenter and Tagger (Figure 6.7), so researchers can use a single word form for those words with differences in diacritics and *Hamza*; however, the diacritics will not show throughout if they are normalised. The Sketch Engine interface can be used in several languages, but Arabic is not yet included. Sketch Engine provides users with a large number of corpora in many languages, and also accepts personal corpora via upload in several file formats. When it came to the criteria of this evaluation, Sketch Engine obtained 7 out of 8 possible points (Table 6.5).

<ul style="list-style-type: none"> Concordance Word List Word Sketch Thesaurus Find X Sketch-Diff Corpus Info ? Save View options KWIC Sentence Sort Left Right Node References Shuffle 	<p>Query 36,217.3) 18 من per million)</p> <p>file1832593 الي ها في الاجازة الصيفية الماضية و كانت من امتع الرحلات التي رحلت الي ها و افضل ها</p> <p>file1832593 , ف بعد ان قررت الرحيل الي ها , اتصلت ب من في ها من سادة القوم و كبار هم , ف سرو و</p> <p>file1832593 ان قررت الرحيل الي ها , اتصلت ب من في ها من سادة القوم و كبار هم , ف سرو و فرحو ; ل</p> <p>file1832593 القرية تجاهي فخرجت في طريقي الخامس عشرة من رمضان متج ها الي القرية , ف وصلت الي ها</p> <p>file1832593 العصر وصلت جازوني ساتلين عن اهل المدينة و من تركة هم في المدينة ثم سلوني عن السعودية</p> <p>file1832593 سلوني عن السعودية و اهل ها و كيف وجدت ها و من في ها , فاجبة هم ب ما تيسر ل ي ان اجيب</p> <p>file1832593 اعظ في ه الناس ب ما تيسر ل ي و ما تعلمت ه من العقيدة الصحيحة من اساتذي للفضلاء ف بقيت</p> <p>file1832593 تيسر ل ي و ما تعلمت ه من العقيدة الصحيحة من اساتذي للفضلاء ف بقيت على هذا اكثر من حين</p> <p>file1832593 الصحيحة من اساتذي للفضلاء ف بقيت على هذا اكثر من حين ل حتى عدت الي المدينة . الرحلة الي</p> <p>file1832592 الي ها في الاجازة الصيفية الماضية و كانت من امتع الرحلات التي رحلت الي ها و افضل ها</p> <p>file1832592 , ف بعد ان قررت الرحيل الي ها , اتصلت ب من في ها من سادة القوم و كبار هم , ف سرو و</p> <p>file1832592 ان قررت الرحيل الي ها , اتصلت ب من في ها من سادة القوم و كبار هم , ف سرو و فرحو ; ل</p> <p>file1832592 القرية تجاهي فخرجت في طريقي الخامس عشرة من رمضان متج ها الي القرية , ف وصلت الي ها</p> <p>file1832592 العصر وصلت جازوني ساتلين عن اهل المدينة و من تركة هم في المدينة ثم سلوني عن السعودية</p> <p>file1832592 سلوني عن السعودية و اهل ها و كيف وجدت ها و من في ها , فاجبة هم ب ما تيسر ل ي ان اجيب</p> <p>file1832592 اعظ في ه الناس ب ما تيسر ل ي و ما تعلمت ه من العقيدة الصحيحة من اساتذي للفضلاء ف بقيت</p> <p>file1832592 تيسر ل ي و ما تعلمت ه من العقيدة الصحيحة من اساتذي للفضلاء ف بقيت على هذا اكثر من حين</p> <p>file1832592 الصحيحة من اساتذي للفضلاء ف بقيت على هذا اكثر من حين ل حتى عدت الي المدينة . جنت</p>
---	--

Figure 6.7: Sketch Engine removed the diacritics when normalising the texts

Table 6.5: Benchmark score of the Sketch Engine web tool

Evaluation criteria								Score
1	2	3	4	5	6	7	8	7/8
Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	

6.2.10 IntelliText Corpus Queries¹

IntelliText Corpus Queries (Sharoff, 2014; Wilson *et al.*, 2010) is a web-based system developed by the Centre for Translation Studies (CTS) at the University of Leeds for the purpose of facilitating and enhancing teaching and research in various areas of the humanities. IntelliText provides a list of corpora of languages supported including Arabic (Sharoff, 2014), as well as a number of functions to search these corpora, such as concordances, collocations, affixes, compare frequencies, key words, and phrases.

IntelliText Corpus Queries enables users to upload their own corpora in several languages. Arabic is not one of them, although this tool includes some built-in Arabic corpora. Uploading UTF-8 and UTF-16 files of Arabic is unfortunately not supported, however. In the built-in Arabic corpora, Arabic texts were displayed in the correct direction, right to left, and diacritics were presented correctly (Figure 6.8), but diacritics and *Hamza* were not normalised, and the search results therefore do not include the query form that shows differences in diacritics or *Hamza*. The interface of IntelliText is available only in English. The score IntelliText achieved in this evaluation is 2 of 8 possible points (Table 6.6).

¹ IntelliText Corpus Queries can be accessed from: <http://smlc09.leeds.ac.uk/itb/htdocs/Query.html>

Figure 6.8: Diacritics displayed correctly in IntelliText Corpus Queries

Table 6.6: Benchmark score for IntelliText Corpus Queries

Evaluation criteria								Score
1	2	3	4	5	6	7	8	2/8
No	No	Yes	Yes	No	No	No	No	

6.2.11 Comparing the Results

Comparing all results of the evaluation reveals some significant points as follows:

1. Although none of the tools examined fulfilled all the evaluation criteria and achieved 8 points, three tools (Khawas, aConCorde and Sketch Engine), met more than 75% of the criteria and achieved the highest scores (Table 6.7).

Table 6.7: Comparison of the tools included in this evaluation

Evaluation criteria	PC-based tools				Web-based tools	
	Khawas	aConCorde	AntConc	WS Tools	Sketch Engine	IntelliText
1. Reading Arabic UTF-8 files	+	+	+	+	+	-
2. Reading Arabic UTF-16 files	-	+	+	+	+	-
3. Displaying Arabic diacritics	+	+	+	-	+	+
4. Arabic text in R-to-L direction	+	+	-	+	+	+
5. Normalising diacritics	+	-	-	-	+	-
6. Normalising <i>Hamza</i>	+	-	-	-	+	-
7. Providing Arabic interface	+	+	-	-	-	-
8. Arabic personal corpus	+	+	+	+	+	-
Score	7/8	6/8	4/8	4/8	7/8	2/8

2. The most significant commonalities that Khawas, aConCorde, and Sketch Engine share are that they paid more attention to the features of Arabic such as diacritics and *Hamza*, specifically in Khawas and Sketch Engine, which have the highest points (7 for each), and Arabic was one of the languages that these tools were developed for, Khawas and aConCorde in particular.

3. Khawas and aConCorde are PC-based software while Sketch Engine is a web-based tool. While there is no difference in terms of the basis of the tools (PC or web) with regard to handling Arabic language, taking Arabic features into consideration when developing these tools may help to make them more appropriate for Arabic corpora.

4. Both Khawas and Sketch Engine are strong competitors as tools for searching and analysing Arabic corpora. Khawas provides an Arabic interface which might be a significant factor to some users, while this was the only shortcoming in Sketch Engine. By contrast, Khawas reads only text files in the UTF-8 format, whereas Sketch Engine can read many types of data files (e.g., .doc, .docx, .html, .pdf, .ps, .tar.gz, .txt, .xml, .zip, and other formats). Sketch Engine can also download the content of a website and store it as a corpus, and text from any external source can

be pasted into the tool. Such flexibility helps when there is a need to use a diversity of data resources.

6.3 Using the ALC Metadata to Restrict the Search

Some online corpora allow the user to restrict the search to specific parts of the data based on some determinants. For instance, the search in the British National Corpus (Burnard, 2007) can be restricted to a specific text mode (written and spoken), time period (since 1990), or genre such as spoken, fiction, magazine, newspaper, and academic. The Michigan Corpus of Academic Spoken English (Simpson *et al.*, 2009) offers some determinants such as gender, age, academic position/role, nativeness, and first language. The Michigan Corpus of Upper-level Student Papers (O'Donnell & Römer, 2009a) allows the user to restrict the search to some features such as student level, nativeness, textual feature, paper type, and discipline.

However, few of those search tools allow users to upload their own corpora such as the commercial web-based tool Sketch Engine (Kilgarriff, 2014; Kilgarriff *et al.*, 2004) which allows configuring a number of sub-corpora based on pre-determined features. Based on the review of tools for searching and analysing Arabic corpora, the ALC was added to Sketch Engine with the configuration of all possible sub-corpora based on the 26 metadata elements; however, the researcher wanted the presentation of the determinants on the user interface to be more friendly and easy to use. For example, to search the ALC on Sketch Engine, users must be aware of the values of some determinants; that is, users must know which nationalities may be entered in the Nationality element, which L1s are included under Mother Tongue, and the names of institutions that can be given for the Educational Institution element (see Figure 6.9). As a result, there is a need for an external source listing those values; otherwise, these determinants might be useless.

Figure 6.9: Example of determinants of the ALC in Sketch Engine

To eliminate the need for an external list, the researcher decided to list the determinants' values so the user can select one or more of them. The Michigan Corpus of Upper-level Student Papers website uses this method (Figure 6.10). We contacted the developers of this corpus, the Michigan Corpus of Upper-level Student Papers (Ute Römer, personal communication, 2 July 2013), in order to adapt the interface to Arabic and host the ALC. However, they responded that there was no longer a corpus team as project funding ended in August 2011, so the corpus website is frozen with no prospect of further development. Thus, we decided to build our own website using the same friendly method for restricting the search to the determinants.

Paper ID	Title
BIO.G0.15.1	Invading the Territory of Invasives: The Dangers of Biotic Disturbance
BIO.G1.04.1	The Evolution of Terrestriality: A Look at the Factors that Drove Tetrapods to Move Onto Land
BIO.G3.03.1	Intracellular Electric Field Sensing using Nano-sized Voltmeters
BIO.G0.11.1	Exploring the Molecular Responses of Arabidopsis in Hypobaric Environments: Identifying Possible Targets for Genetic Engineering
BIO.G1.01.1	V. Cholerae: First Steps towards a Spatially Explicit Model
BIO.G1.07.1	Zebrafish and PGC mis-migration
BIO.G2.06.1	A Conserved Role of Cas-Spg System in Endoderm Specification during Early Vertebrate Development
BIO.G3.02.1	Linking scales to understand diversity
BIO.G0.01.1	The Ecology and Epidemiology of Plague
BIO.G0.02.1	Host-Parasite Interactions: On the Presumed Sympatric Speciation of Vidua
BIO.G0.02.2	Sensory Drive and Speciation
BIO.G0.02.3	Plant Pollination Systems: Evolutionary Trends in Generalization and Specialization
BIO.G0.02.4	Chromosomal Rearrangements, Recombination Suppression, and Speciation: A Review of Rieseberg 2001
BIO.G0.02.5	On the Origins of Man: Understanding the Last Two Million Years
BIO.G0.04.1	Fetal Endocrine System
BIO.G0.05.1	Mn (III) TPPS4: A Metallophorphyrin Used for Tumor Identification in MRI
BIO.G0.06.1	Global Reproductive Strategies of Tursiops and Stenella (Family Delphinidae)
BIO.G0.07.1	Complementation Between Histidine-Requiring Mutants of Saccharomyces Cerevisiae
BIO.G0.09.1	Nest Selection In Weaver Birds
BIO.G0.11.3	Fungal Eye Infections Due to ReNu MoistureLoc

Figure 6.10: Search determinants on the website of the Michigan Corpus of Upper-level Student Papers

6.4 Purpose

The aim of the ALC Search Tool is to enable users to search the corpus data based on a number of determinants and to download a subset of the corpus files (sub-corpus) based on those determinants. The ALC design criteria include a number of learner and text features that can be selected to search a sub-corpus, such as “age”, “gender”, “mother tongue”, “text mode”, and “place of writing”. The corpus has 26 features which are used as determinants on this tool. Using those determinants provides three main advantages. First, it allows users to search any sub-corpus based on the determinants required (e.g. searching the sub-corpus of non-native speakers of Arabic). Second, users may compare the results of two sub-corpora (two comparable groups such as learners at the pre-university level to those at the university level). Finally, users can download a subset of the corpus in different formats (TXT, XML, PDF, and MP3).

6.5 Design

The ALC Search Tool is a free-access, web-based tool, but registration is required to obtain this free access. The website of the ALC Search Tool (<http://www.alcsearch.com>) was created by the researcher – and hosted on a web hosting service paid by the researcher – independently from the ALC main website (<http://www.arabiclearnerscorpus.com>), which contains details about the corpus, developers, publications, and other information. The reason of developing a separate website for the search tool is that it is intended in future to be used not only for the ALC but as a generic search tool for further Arabic corpora as described in the future work in Chapter 8.

The screenshot displays the English interface of the Arabic Learner Corpus (ALC) Search Tool. The interface is divided into several sections:

- Header:** "ALC ARABIC LEARNER CORPUS" and "المحوثة اللغوية لطلبة اللغة العربية".
- Welcome Message:** "Welcome to the Arabic Learner Corpus search website" with a [USER GUIDE] link.
- Search Section:** A search bar labeled "الرجلة" and a "Search" button. Below it, a checkbox for "Separate Words" and the text "Search result".
- Search Results:** A table showing "1-16 of total 1139 results". The table has two columns: "Text ID" and "Concordance".

Text ID	Concordance
S001_T1_M_Pre_NNAS_W_C	الرجلة إلى القرية لزيارة دوي
S002_T1_M_Pre_NNAS_W_C	ناولنا العظمى وفضلاً مسؤولو الرجلة مجموعات حتى تكو الرجلة
S002_T1_M_Pre_NNAS_W_C	الرجلة مجموعات حتى تكو الرجلة منظمة بعد ذلك ركنا
S002_T1_M_Pre_NNAS_W_C	دعاء السفر وضماً مسؤول الرجلة صائح منبده وضماً في
S002_T1_M_Pre_NNAS_W_C	الله تعالى فكانت هذه الرجلة مشاركة لأنها رجلة الطاعة
S002_T1_M_Pre_NNAS_W_C	أذكر كثيراً من تفاصيل الرجلة لخص الوقت
S004_T1_M_Pre_NNAS_W_C	وخطبت عن المشاركين في الرجلة إلى اليوم الموعود وركبت
S005_T1_M_Pre_NNAS_W_C	الرجلة إلى المدينة الصورة ذهبت
S005_T1_M_Pre_NNAS_W_C	مع أصحابه أحتجني هذه الرجلة مدينة رسول صلى الله
S005_T3_M_Pre_NNAS_W_H	الرجلة إلى المدينة الصورة ذهبت
S010_T1_M_Pre_NNAS_W_C	الرجلة إلى بلدي أأ طالب
S010_T1_M_Pre_NNAS_W_C	وتخله وقد نصيا في الرجلة ولكن عندما وصلنا إلى
S011_T1_M_Pre_NNAS_W_C	في ذات الأنيوع أخذت الرجلة عشر ساعات وعندما وصلنا
S015_T1_M_Pre_NNAS_W_C	الرجلة إلى مكة المكرمة والمدينة
S015_T3_M_Pre_NNAS_W_H	الرجلة إلى مكة المكرمة والمدينة
S017_T1_M_Pre_NNAS_W_C	الرجلة إلى الحرمين بسم الله
- Search Determinants:** A list of filters including AGE, GENDER, NATIONALITY, MOTHER TONGUE, NATIVENESS, NUMBER OF LANGUAGES SPOKE, NUMBER OF YEARS LEARNING ARABIC, NUMBER OF YEARS SPENT IN ARABIC COUNTRIES, GENERAL LEVEL OF EDUCATION, LEVEL OF STUDY, YEAR OR SEMESTER, EDUCATIONAL INSTITUTION, TEXT GENRE, PLACE OF WRITING, YEAR OF WRITING, COUNTRY OF WRITING, CITY OF WRITING, TIMING, REFERENCES USE, GRAMMAR BOOKS USE, MONOLINGUAL DICTIONARIES USE, BILINGUAL DICTIONARIES USE, OTHER REFERENCES USE, TEXT MODE, TEXT MEDIUM, and TEXT LENGTH.
- File Download:** A section on the left indicating "The number of corpus texts is 1585" and "The number of texts based on your selection is 1585". It offers download options: Plain text with no metadata, Plain text with Arabic metadata, Plain text with English metadata, XML with Arabic metadata, XML with English metadata, Hand written sheets in PDF, and Audio recordings in MP3. There are "Download" and "Clear" buttons.
- Results Section:** A detailed view of the selected result (S015_T1_M_Pre_NNAS_W_C) showing the concordance text in Arabic: "الرجلة إلى مكة المكرمة والمدينة الصورة في يوم ألتا في يوم ألتا في يوم ألتا...".

Figure 6.11: English interface of the main page of the ALC Search Tool

The website consists of two pages: the login/sign up page in which the user can register and obtain free access to the tool, and the main page in which the user can search and download the corpus or any subset of its data (Figure 6.11). As the ALC Search Tool is intended to be used by a worldwide audience interested in the ALC, one of the distinctive features is the provision of the interface in two languages, Arabic and English. Importantly, the development of the two interfaces offers not

only translations for labels and buttons, but even switches the entire website layout to right-to-left. Additionally, the researcher created a user guide to present an overview of the ALC Search Tool and an illustration of how to use it and to take advantage of its functions. Similar to the website, this guide is available in two languages, Arabic¹ and English². A link to each copy is located on the interface matching its language.

When the user clicks on the title of any of the determinants, its values will appear for selection. The values can be cleared by clicking on “No Restriction” at the top of the list of options; doing this will reset the value of the selected determinant only. To clear the values of determinants all at once, the user can click on “Clear All Determinants” at the top of the determinants list. By selecting any option from the determinants, the number of texts available based on the new selection will be shown above the files download section (Figure 6.12).

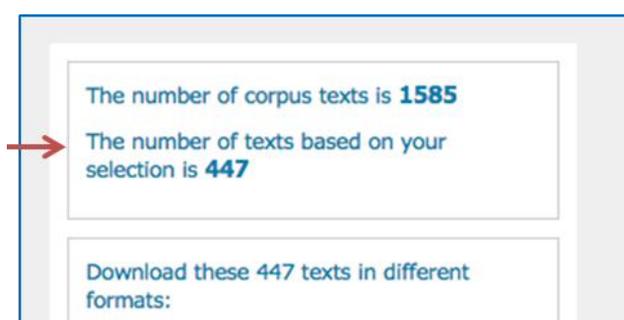


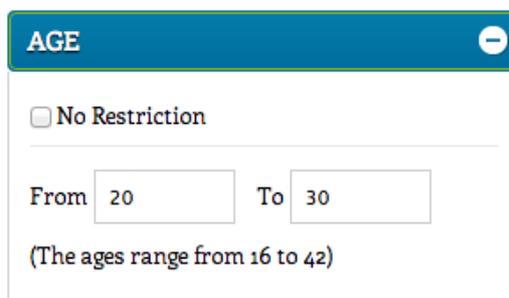
Figure 6.12: Updating the number of texts available based on the determinants selected

6.6 Determinant Types

The determinants on the website can be classified into three types. The first type includes those with a numerical range value. This type requests two values, the minimum and maximum of the range, and it accepts only values in the Arabic numeral system (1, 2, 3, 4, 5, 6, 7, 8, 9, and 0). For example, the user can select a range of learners’ ages between 20 and 30 years (Figure 6.13).

¹ The Arabic version can be accessed from:
http://www.alcsearch.com/ALCfiles/ALC_User_Guides/User_Guide_Ar.pdf

² The English version can be accessed from:
http://www.alcsearch.com/ALCfiles/ALC_User_Guides/User_Guide_En.pdf



AGE

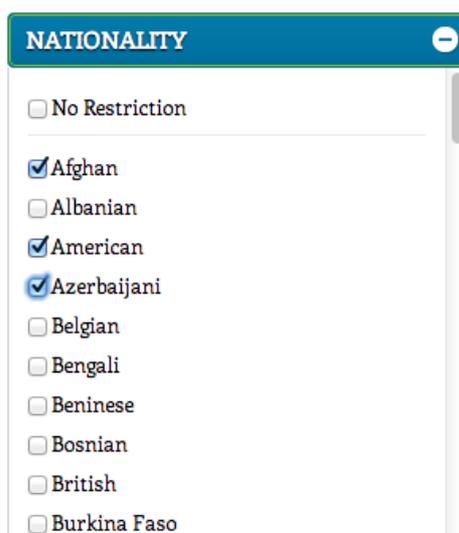
No Restriction

From To

(The ages range from 16 to 42)

Figure 6.13: Example of a determinant with a numerical range value

The second type is those determinants with a multi-selection list where user can select one or more options from this list. The user for example can select any number of nationalities to search the sub-corpus of learners belonging to those nationalities (Figure 6.14).



NATIONALITY

No Restriction

Afghan

Albanian

American

Azerbaijani

Belgian

Bengali

Beninese

Bosnian

British

Burkina Faso

Figure 6.14: Example of a determinant with a multi-selection list

The third type includes determinants with two options (“Yes” or “No”). Only one choice can be selected from this type of list. For instance, the user can select whether texts were produced using any language references by choosing “Yes” or “No” (Figure 6.15).



A screenshot of a web interface element titled "REFERENCES USE" in a blue header bar. Below the header, there are three radio button options: "No Restriction" (unchecked), "Yes" (checked), and "No" (unchecked). The "Yes" option is selected, indicated by a blue dot inside the radio button.

Figure 6.15: Example of a determinant with only two options

6.7 Functions

As mentioned in the tool purpose, the ALC Search Tool was designed to perform two main functions. The first is to enable users to search the corpus data based on a number of determinants, and the second is to enable them to download a subset of corpus files based on those determinants. Those two functions are described in the following subsections.

6.7.1 Searching the Corpus

The search function works as a basic concordancing tool, so users are able to search for a particular word. It retrieves all results matching the given search term along with their contexts (that is, four words on either side of the search term). The results section consists of some subsections as shown in Figure 6.16. All results are displayed with the search term highlighted in a different colour. The text ID of each example of the results is also retrieved and shown next to the example. Results can be printed using the print button or exported into Excel file format (.xls) using the download button. The full text of any example can be displayed by clicking on the highlighted word; a text box will appear at the bottom of the page showing the full text and highlighting all the matches.

Search result

1-16 of total 1139 results

Text ID	Concordance
S001_T1_M_PRE_NNAS_W_C	الرحلة إلى القرية لزيارة ذوي
S002_T1_M_PRE_NNAS_W_C	تناولنا العطور وقسمنا مسؤولو مجموعات حتى تكون الرحلة
S002_T1_M_PRE_NNAS_W_C	الرحلة مجموعات حتى تكون الرحلة منظمة بعد ذلك ركنا
S002_T1_M_PRE_NNAS_W_C	دعاء السفر ونصحنا مسؤول الرحلة بمناجح مفيدة ومضيا في
S002_T1_M_PRE_NNAS_W_C	الله تعالى فكانت هذه الرحلة مباركة لأنها رحلة الطاعة
S002_T1_M_PRE_NNAS_W_C	أذكر كثيرا من تفاصيل الرحلة لصق الوقت
S004_T1_M_PRE_NNAS_W_C	وتحطت من المشاركين في الرحلة أدى اليوم الموعود وركبت
S005_T1_M_PRE_NNAS_W_C	الرحلة إلى المدينة المنورة ذهبت
S005_T1_M_PRE_NNAS_W_C	مع أصحابه أعجبتني هذه الرحلة مدينة رسول صلى الله
S005_T3_M_PRE_NNAS_W_H	الرحلة إلى المدينة المنورة ذهبت
S010_T1_M_PRE_NNAS_W_C	الرحلة إلى بلدي أنا طالب
S010_T1_M_PRE_NNAS_W_C	ونقلة وقد نصنا في الرحلة ولكن عندما وصلنا إلى
S011_T1_M_PRE_NNAS_W_C	في ذاك الأسبوع أخذت الرحلة عشر ساعات وبعدها وصلنا
S015_T1_M_PRE_NNAS_W_C	الرحلة إلى مكة المكرمة والمدينة
S015_T3_M_PRE_NNAS_W_H	الرحلة إلى مكة المكرمة والمدينة
S017_T1_M_PRE_NNAS_W_C	الرحلة إلى الحرمي بسم الله

Page 1 of 72

S015_T1_M_Pre_NNAS_W_C

الرحلة إلى مكة المكرمة والمدينة المنورة في يوم اثنان من اسبوع ماضي، سافرت إلى مكة المكرمة في الساعة ثلاثة والنصف بعد الصلاة العسر مع الأصدقاء، ركنا الحافلة كلمة غير مفروقة، وصلنا إلى الصفات في الساعة الثامن ونصف ليلاً ثم لبسنا صلبنا الصلاة المغرب والعشاء في الصفات فصر بعد ذلك لبسنا الألباس الحرام فلما لبك عمرة تم وجدنا إلى مكة المكرمة وصلنا إلى مكة المكرمة في الساعة الثانية صباح قبل وصل ذهبنا إلى المطعم ليكل الطعام العشاء، وجدنا إلى البيت الحرام لصلاة الصبح، تم عمرة بعد ذلك، وجدنا إلى المدينة المنورة من مكة إلى المدينة فبصنا سنة الساعة المدينة المنورة فيها المسجد النبوي

Figure 6.16: Results section on the ALC Search Tool

When searching for a word such as “كيف” *kaīfa* ‘how’, the results will include all examples where the search term appears, whether as an independent word matching the search form “كيف” or with prefixes and/or suffixes such as “كيفية” *kaīfiya* ‘method’ and “كيفيتها” *kaīfiyatuha* ‘its method’ (see Figure 6.17 for an example). In the search box, users can select *Separate Words* to show only those examples that include the search word independently “كيف”. Once it is selected, all results with prefixes and/or suffixes (e.g. “كيفية” and “كيفيتها”) will be excluded, Figure 6.17 illustrates the results of the word “كيف” with and without selecting the choice *Separate Words*.

Without selecting the <i>Separate Words</i> choice	with selecting the <i>Separate Words</i> choice
لأنه شيء أستمتع بسماعها وكيف يكون القيام بها، أتمنى	الأولى إلى مكة المكرمة كيف منظر الحرم؟ هل الناس
عن الحج وأما فرأت كيفية الحج وشروط الحج لأكون	فيبقى عند الأسئلة المهمة كيف أنال هذا وذاك؟ وبما
ذلك فراءة شروط العمرة، وكيف بنتها وصفتها، واستعددت الطعام والملابس	ما وأنه قبله مثلاً كيف نرتب الغرور على الجدار
الحج كنت أنعلم عن كيفية أداء الحج بطريقة الصحيحة	فاستشار الملك مع الوزراء كيف تبقى المحبة بعد وفاتها؟
الشبكة الدولية لكي أعرف كيفية أداء الحج بدقه حتى	البيئة ولهجات أهل مصر كيف تنطق في اللغة العربية
راشد وعليكم السلام ماجد كيف حالك راشد أنا بخير	على نزول القرآن الكريم كيف نزل قديماً حتى وصل

Figure 6.17: Results with and without using the *Separate Words* choice

The mechanism of the website is dynamic, so the determinants' values can be changed after the search. Any changes made by the user will be reflected in the number of texts, and new results will be shown automatically as they will be updated based on the new values of the determinants.

In terms of the architecture of the search function, it starts once a determinant value is changed or the search button is clicked. The search function sends a query to the ALC database with the values of the determinants and the search term (Figure 6.18), and the results retrieved are stored in an array.

```
var dataString =
'search_txt1='+search_txt+'&fromAge='+fromAge+'&toAge='+toAge+'&gender='+gender+'&nat
ionality='+nationality+'&mother='+mother+'&nativeness='+nativeness + '&fromLangSpok='
+fromLangSpok + '&toLangSpok='+toLangSpok + '&fromYearLearnAr='
+fromYearLearnAr+'&toYearLearnAr='+toYearLearnAr+'&fromYearSpentAr='+fromYearSpentAr+
'&toYearSpentAr='+toYearSpentAr+'&genLevEdu='+genLevEdu+'&levStudy='+levStudy+'&yearS
em='+yearSem+'&eduInsti='+eduInsti+'&textGenre='+textGenre+'&placeWrite='+placeWrite+
'&yearWrite='+yearWrite+'&countWrite='+countWrite+'&cityWrite='+cityWrite+'&Timing='+
Timing+'&refUse='+refUse+'&grBookUse='+grBookUse+'&monoDict='+monoDict+'&bilDict='+bi
lDict+'&othRefUse='+othRefUse+'&textMode='+textMode+'&textMedium='+textMedium+'&fromT
ext='+fromText+'&toText='+toText+'&search_type='+search_type;

$.ajax({
  type: "POST",
  url: "<?php echo base_url(); ?>en/ajaxTextSearch",
  data: dataString,
  dataType: 'json',
  success: function(response)
```

Figure 6.18: Sending a query to the ALC database

If no text matches the query conditions (the determinants values and the search term), the tool shows zero in the number of texts available, hides the files download part, and clears all results from the results section. If there are results matching the query conditions, then the number of those texts will be shown as well as the download section (Figure 6.19). The final step before showing the results is to check if the Separate Words checkbox is selected; if it is, then the concordances will be sorted by excluding those matches with prefixes and/or suffixes.

```
var show_data='<table width="100%" border="0" cellspacing="0"
cellpadding="0" class="tblRes1">'+

'<tr>'+

'<th>Text ID</th>'+

'<th>Concordance</th>'+

'</tr>'+

'<tr>'+

'<td colspan="2" style="border-right:0px; padding:0px;">'+

'<table width="100%" border="0" cellspacing="0" cellpadding="0">';

if(response != null)
{
if(response['title'] !='')
{
for(i=0;i<response['title'].length; i++)
{
show_data += response['title'][i];
}
show_data += '</table>'+

'</td>'+

'</tr>'+

'</table>';

$('#search_data').html(show_data);

$('#print_id').show();
```

```
        $('#download_id').show();
    }
    else
    {
        show_data += '<tr>'+
Here</td>'+
                '<td colspan="2" align="center">No Records
                '</tr>';
        show_data += '</table>'+
                '</td>'+
                '</tr>'+
                '</table>';

        $('#search_data').html(show_data);
        $('#print_id').hide();
        $('#download_id').hide();
    }

    $('#paginationBx').html('<div
id="test">'+response["pagination"]+'</div>');

    $('#search_rows').html(response['total_rows']);
    $('#search_rows2').html(response['total_rows']);
    $('#no_of_rows').html(response['no_of_results']);
    $('#ajaxLoaderDiv').hide();
    ajaxSearch_paging();
```

Figure 6.19: Showing or hiding the results based on the query response

See an extended sample code of the search function in Appendix G. Figure 6.20 illustrates the architecture of the searching function.

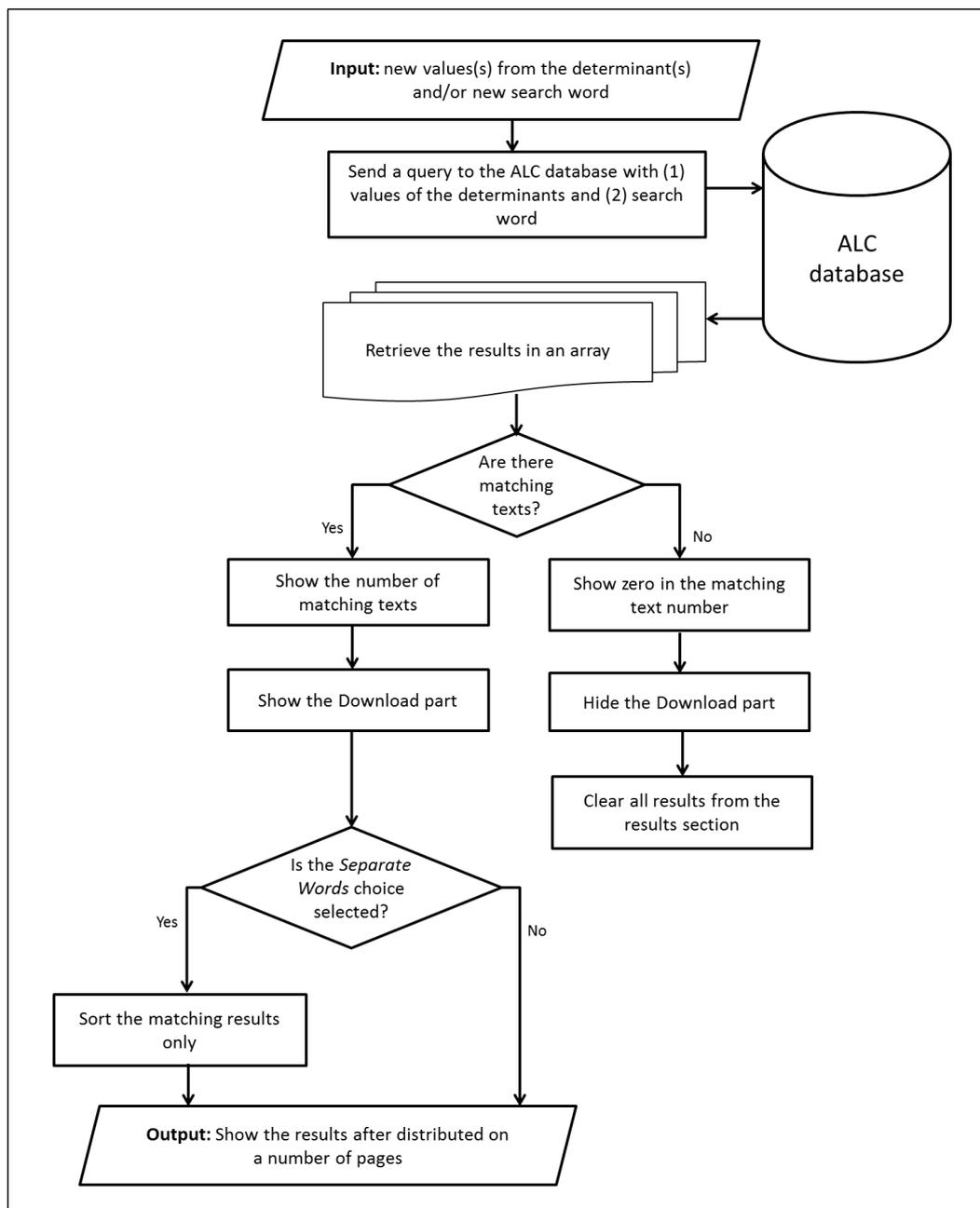


Figure 6.20: Architecture of the search function in the ALC Search Tool

6.7.2 Downloading the Corpus Files

One of this tool's aims is to enable users to download any subset of the corpus files using the determinants in different formats (Table 6.8). The number of files available depends on the determinants values, and this number is updated based on any changes in the determinants values.

Table 6.8: Number of files available for each format in the ALC

Format	No of files
TXT files with no metadata	1585
TXT files with Arabic metadata	1585
TXT files with English metadata	1585
XML files with Arabic metadata	1585
XML files with English metadata	1585
Original hand-written sheets in PDF	1257
Audio recordings in MP3	52

The PDF and MP3 formats have fewer files than the other formats, i.e. some texts may not have files in these two formats. For example, when selecting “Azerbaijani” from the “Nationality” determinant, “Audio recordings in MP3” from the files download section, and clicking on the “Download” button, a message will appear indicating that there are no files to download for this selection. This occurs because there are no MP3 files for this selection, even though 10 texts from Azerbaijani learners can be downloaded in any of the other formats.

The architecture of the download function includes three main steps before sending the files to the user. The first step is to retrieve the texts’ IDs from the array of the searching function; this list of IDs includes only those texts matching the query conditions. The second step is to retrieve the file formats selected by the user among the seven formats available in the download section. The third step is to retrieve those files matching the results of step 1 and step 2. The files are then compressed into one ZIP file containing subfolders, each of which includes the files of one format of those selected. Finally, the ZIP file is sent to the user for downloading (Figure 6.21).

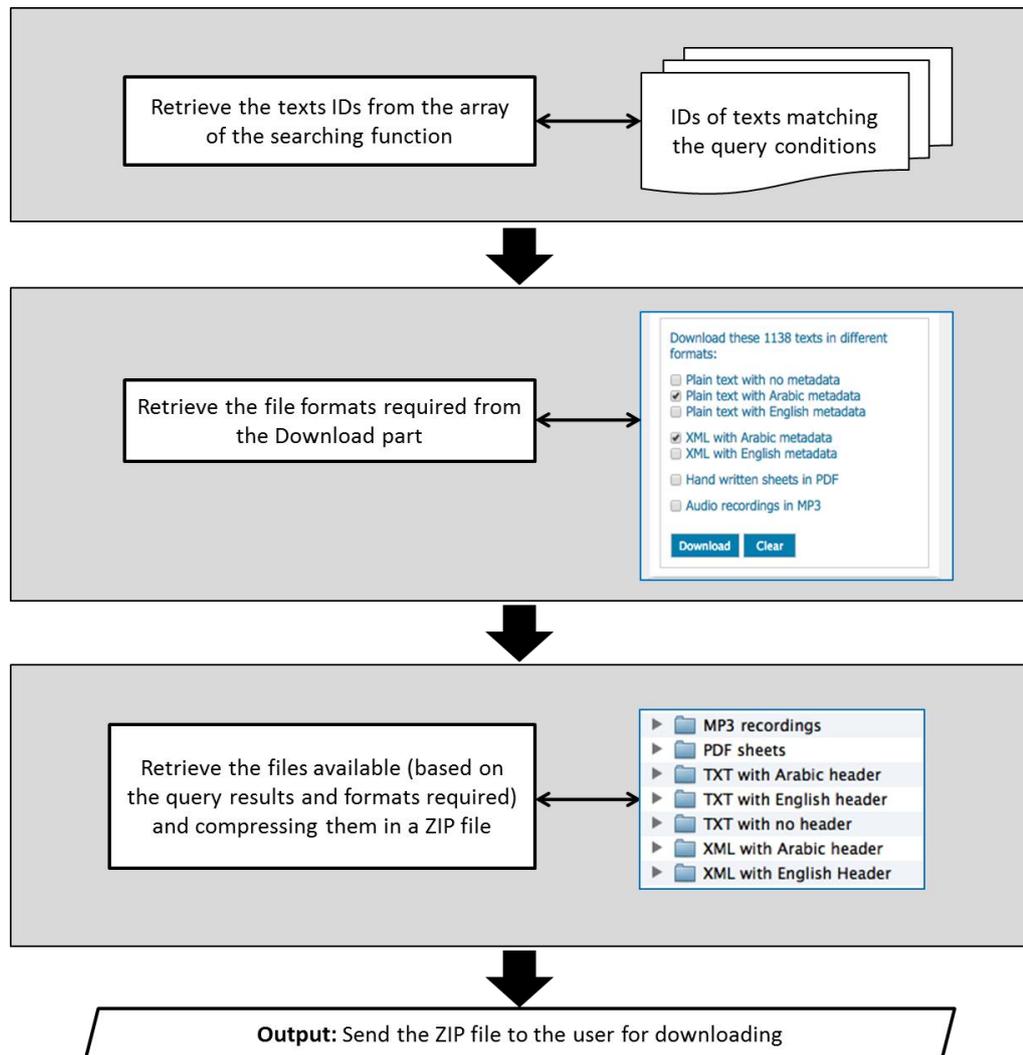


Figure 6.21: Architecture of the download function on the ALC Search Tool

6.8 Evaluation

The ALC Search Tool was evaluated in three dimensions: (i) the accuracy of the search results which presents a technical view of the ability and limits of the search function, (ii) the views of a number of specialists in some computational and linguistic research areas, and (iii) the number of website visits, which gives an indication of the extent of its use.

6.8.1 Evaluating the Output of the ALC Search Tool

In evaluating the outputs of the ALC Search Tool, the focus was to evaluate the accuracy of retrieving a query string. The search tool includes two choices: normal

search, which returns all matches including the query string with or without prefixes and suffixes, and the *Separate Words* option, which returns only those matches that have no prefixes or suffixes. Each of these options was evaluated separately. Two aspects for measuring the accuracy of the ALC Search Tool were investigated: *Recall*, equates to whether or not the query string is found, and *Precision*, equates to whether or not the string retrieved is relevant to the query.

These two aspects define the elements of the confusion matrix used to calculate the accuracy of the ALC Search Tool outputs. The confusion matrix contained four elements: true positive, true negative, false positive, and false negative. According to the observations of the ALC Search Tool outputs, these elements are defined as:

- True Positive (TP): True and applicable; the case is relevant to the query and retrieved to the output correctly.
- True Negative (TN): True but not applicable; the case is not relevant and not retrieved.
- False Negative (FN): False retrieving of a relevant case; the case is relevant but not retrieved.
- False Positive (FP): False retrieving of a non-relevant case; the case is not relevant but is retrieved as relevant.

Using this confusion matrix allowed the researcher to classify the output into four categories:

1. Relevant strings and retrieved as relevant: this category represents those strings retrieved by the ALC Search Tool as relevant results to the query. For example, strings such “وقتاً” *Waqtan* ‘a time’ and “الوقت” *’alwaqt* ‘the time’ contain the query string “وقت” *Waq̣t* ‘time’; thus, they are relevant results to the query and retrieved as relevant.
2. Non-relevant strings and not retrieved: this category indicates cases not relevant to the query and not retrieved in the output results, which includes all strings in the corpus other than those retrieved and those relevant.
3. Relevant strings but not retrieved: this category includes strings relevant to the query, but that were not retrieved by the ALC Search Tool. For example, the string “أكبر” *’akbar* ‘greater’ was not retrieved through the query “أكبر” *’akbar* ‘greater’ because of the difference between those two strings in the way of writing the first character, with *Hamza* above it (i) in the query and without it in the non-retrieved strings (l).

4. Non-relevant strings and retrieved as relevant: this category represents strings retrieved as relevant results to the query when they were in fact not. For instance, the string “أهل” *ahl* ‘family’ was retrieved for the query “هل” *hal* ‘question particle’, as the latter string is a part of the former, but they are irrelevant.

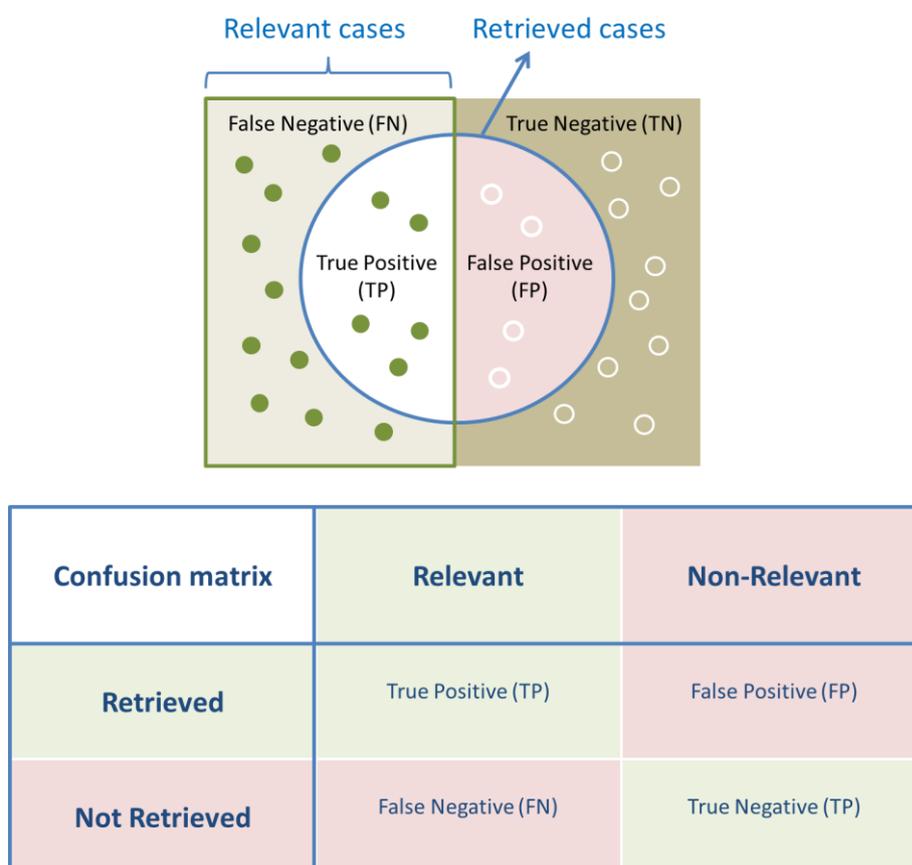


Figure 6.22: The confusion matrix aspects and elements

The results retrieved from the ALC Search Tool were sorted into either relevant (TP) or non-relevant (FP). The classification was performed manually on the following basis: if the retrieved word shared the same lemma with the query string, it was deemed relevant; otherwise, it was not relevant.

However, checking those relevant but not retrieved cases (FN) manually was difficult due to the large amount of data not retrieved, so a reference tool was used to check whether any relevant cases were not retrieved. In particular, arabiCorpus (Parkinson, 2015) and Sketch Engine (Kilgarriff, 2014; Kilgarriff *et al.*, 2004), on which the ALC is searchable, were used for this task. On arabiCorpus, the user can search for a string of characters where all words including the string will be

retrieved. Because this approach is similar to the normal search function on the ALC Search Tool, the results from arabiCorpus were used as an indicator of possibly relevant cases to be compared with the results of the ALC Search Tool. For the *Separate Words* option on the ALC Search Tool, results from Sketch Engine were used to indicate possible relevant cases because the search on Sketch Engine is based on a tokenised version of the ALC where the exact tokens are retrieved. However, the clear difference between Sketch Engine and the *Separate Words* option is that the former returns the query token regardless of whether it has prefixes and/or suffixes in its original form, while the *Separate Words* option on the ALC Search Tool returns only those that have no prefixes and/or suffixes in their original forms. Although this difference created a gap between the retrieved results in some queries, the results of Sketch Engine can be seen as a typical target to which the *Separate Words* option on the ALC Search Tool needs to achieve in future.

Therefore, the FN value is the number of results of the reference tool minus the number of results of the ALC Search Tool. For instance, a query for the string “عن” ‘an ‘about’ on arabiCorpus returns 3925 results, and the normal search on the ALC Search Tool returns 3906 results, which means there are 19 possible relevant cases that were not retrieved. The same query on Sketch Engine returns 1210 results, while the *Separate Words* option on the ALC Search Tool returns 1007 results, indicating that there are 203 possible relevant cases that were not retrieved.

Finally, the TN value is the total number of ALC words (282,732) minus all other categories: TP, FP, and FN. This gives the total number of cases which are not relevant and were not retrieved.

The sample of query strings was selected from the 1000 most frequent words in the ALC. One word was randomly selected from each 100, generating 10 words to be searched using the ALC Search Tool.

Table 6.9: Number of results returned for each query on the reference tools compared to the ALC Search Tool

Query string	arabiCorpus	ALC Search Tool (normal search)	Sketch Engine	ALC Search Tool (Separate Words)
عن 'an 'about'	3925	3906	1210	1007
وقت waqt 'time'	592	590	271	231
خاصة ḥāṣṣa 'special'	154	150	131	77
هل hal 'question particle'	724	714	105	94
قضيت qaḍāitu 'I spent'	79	77	73	50
أكبر 'akbar 'greater'	101	77	65	50
العطلة 'al'uṭla 'the holiday'	56	56	56	54
الدراسات 'addirāsāt 'the studies'	44	43	46	36
واجهت wāḡahtu 'I faced'	101	101	42	26
نعود na 'ūd 'we return'	36	36	35	27

Precision, recall, and F-measure are the most frequent measures for information retrieval effectiveness. Precision represents the relevant fraction of the returned results, while recall is the returned fraction of those relevant results, and F-measure (F_1 score) is the weighted harmonic mean of precision and recall (Manning & Raghavan, 2008). Table 6.10 illustrates formulas used for the computation of precision, recall, and F-measure.

Table 6.10: Formulas used to compute precision, recall, and F-measure

Measure	Formulas
Precision	$\mathbf{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of retrieved items}} = \frac{TP}{TP + FN}$
Recall	$\mathbf{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}} = \frac{TP}{TP + FP}$
F-measure	$\mathbf{F - measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2PR}{P + R}$

As explained above, the ALC Search Tool was evaluated using a sample of 10 queries extracted from the 1000 most frequent words of the ALC data. The evaluation covered the output of the normal search function on the ALC Search Tool as well as the *Separate Words* option on the same tool. The results of those queries were evaluated using the results of the same queries from the reference tools previously described (i.e. arabiCorpus as a reference for the normal search and Sketch Engine as a reference for the *Separate Words* option). A confusion matrix was defined to compute two aspects (recall and precision) with four elements: true positive, true negative, false negative, and false positive. The computation of precision, recall, and F-measure was performed based on this confusion matrix. The results of the normal search function are shown in Table 6.11, which illustrates the confusion matrix of each query. It also shows the values of the measures: precision, recall, and F-measure with their average for all queries. The results of the *Separate Words* option are shown in Table 6.12

Table 6.11: Evaluation of the normal search on the ALC Search Tool

Query word	TP	TN	FN	FP	Precision	Recall	F1-score
عن 'an 'about'	1523	281,846	19	2383	38.99%	98.77%	55.91%
وقت waqt 'time'	583	282,492	2	7	98.81%	99.66%	99.23%
خاصة ḥāṣṣa 'special'	150	282,578	4	0	100.00%	97.40%	98.68%
هل hal 'question particle'	57	282,620	10	657	7.98%	85.07%	14.60%
قضيت qaḍāitu 'I spent'	77	282,653	2	0	100.00%	97.47%	98.72%
أكبر 'akbar 'greater'	77	282,631	24	0	100.00%	76.24%	86.52%
العطلة 'al'utla 'the holiday'	56	282,676	0	0	100.00%	100.00%	100.00%
الدراسات 'addirāsāt 'the studies'	43	282,686	1	0	100.00%	97.73%	98.85%
واجهت wāḡahtu 'I faced'	100	282,631	0	1	99.01%	100.00%	99.50%
نعود na'ūd 'we return'	35	282,696	0	1	97.22%	100.00%	98.59%
Average					84.20%	95.23%	85.06%

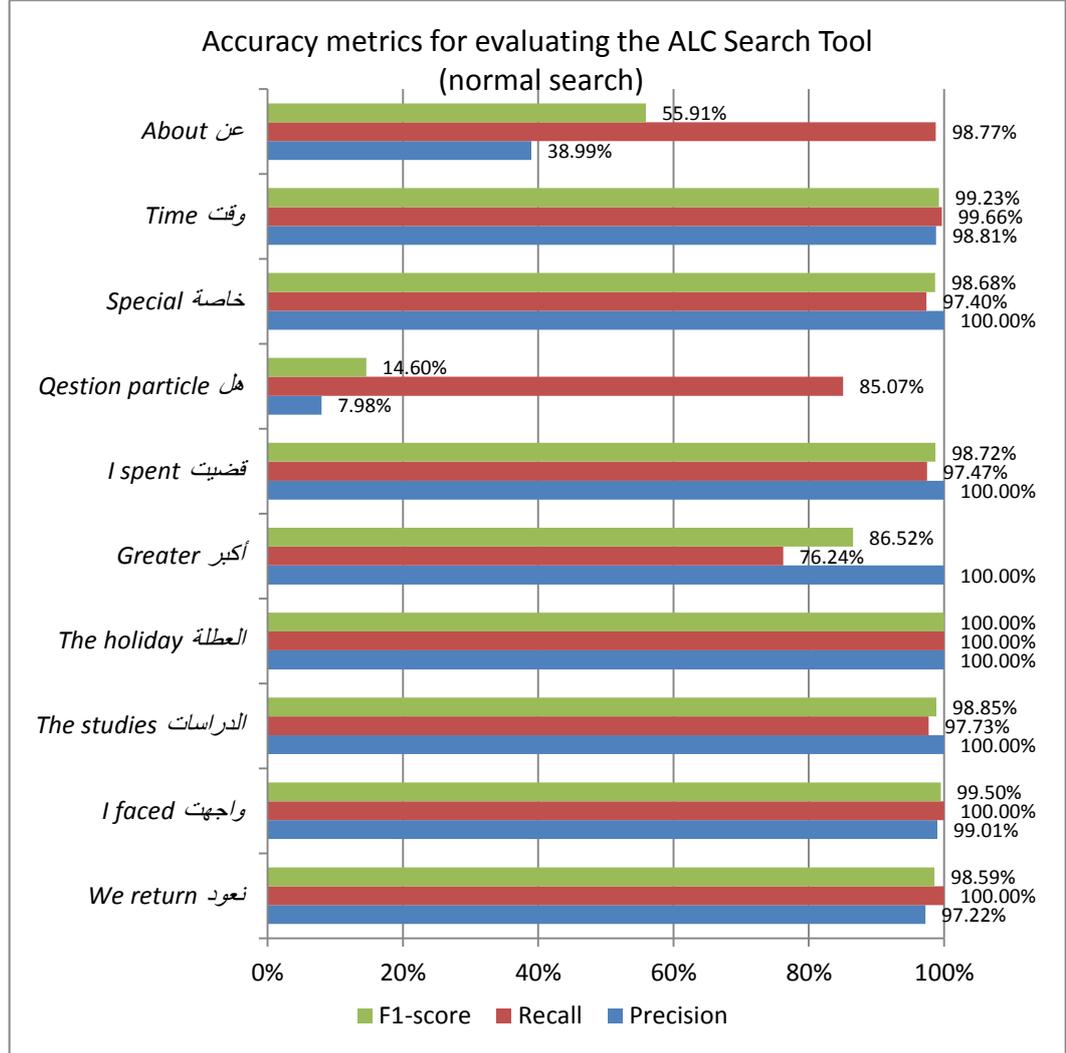


Figure 6.23: Precision, recall, and F-measure of the normal search on the ALC Search Tool

Table 6.12: Evaluation of the *Separate Words* option on the ALC Search Tool

Query word	TP	TN	FN	FP	Precision	Recall	F1-score
عن 'an 'about'	1007	281,522	203	0	100.00%	83.22%	90.84%
وقت waqt 'time'	231	282,461	40	0	100.00%	85.24%	92.03%
خاصة ḥāṣṣa 'special'	77	282,601	54	0	100.00%	58.78%	74.04%
هل hal 'question particle'	93	282,627	11	1	98.94%	89.42%	93.94%
قضيت qaḍāitu 'I spent'	50	282,659	23	0	100.00%	68.49%	81.30%
أكبر 'akbar 'greater'	50	282,667	15	0	100.00%	76.92%	86.96%
العطلة 'al'uṭla 'the holiday'	54	282,676	2	0	100.00%	96.43%	98.18%
الدراسات 'addirāsāt 'the studies'	36	282,686	10	0	100.00%	78.26%	87.80%

6 – Web-Based Tool to Search and Download the ALC

واجهت <i>wāḡahtu</i> 'I faced'	26	282,690	16	0	100.00%	61.90%	76.47%
نعود <i>na'ūd</i> 'we return'	27	282,697	8	0	100.00%	77.14%	87.10%
Average					99.89%	77.58%	86.87%

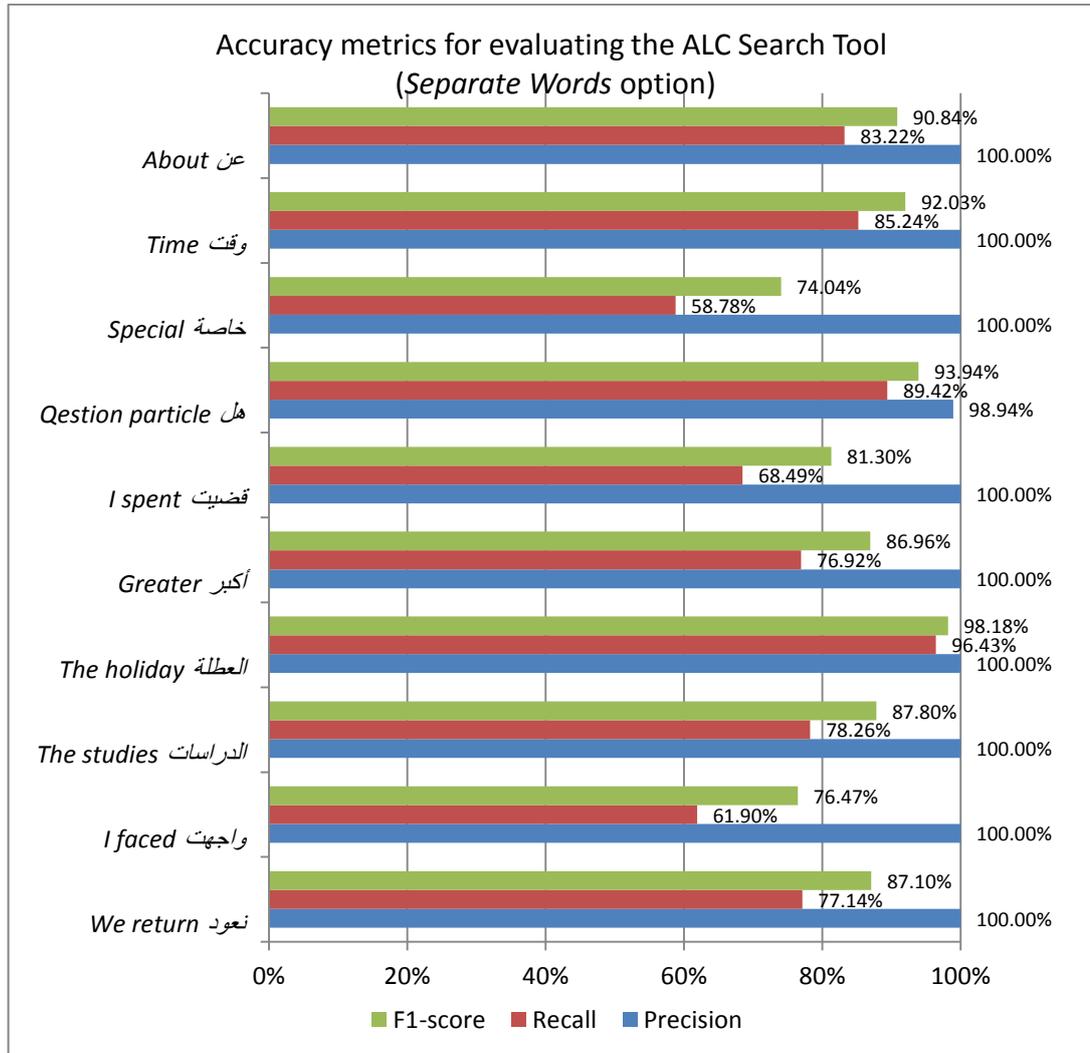


Figure 6.24: Precision, recall, and F-measure of the *Separate Words* option on the ALC Search Tool

The evaluation shows that both types of search (normal search and *Separate Words*) achieved similar average values of F-measure, 85.06% for the former and 86.87% for the latter. However, they achieved different results in terms of the precision and recall measures, as the normal search shows a higher score in recall (95.23%) than the *Separate Words* search (77.58%). In contrast, the *Separate Words* search shows

99.89% in precision, which is higher than the results achieved by the normal search (84.20%).

The normal search function depends on the existence of the query string without conditions in terms of prefixes and suffixes and even if the string is a part of another word, which may explain its high accuracy in the recall measure. However, some words appeared in different forms, e.g. “أكبر” *'akbar* ‘greater’ appeared in the incorrect form “اكبر” without the *Hamza* sign above the first character “ا”; in this case, all cases of the latter form were not retrieved which resulted 76.24% in the recall value of the word “أكبر”. Another example is the word “خاصة” *ḥāṣṣa* ‘special’ that did not return the forms “خاصتي”¹, “خاصته”, “خاصتا” and “فخاصتنا”, as the character *Tā’ Marbūṭa Mutatarrifa* “ة” was not existing in these forms while it is there in the query form “خاصة”, this resulted 97.40% in the recall value. A normalisation process of *Hamza, Tā’ Marbūṭa Mutatarrifa* and other signs such as diacritics may contribute to resolving this problem, as all types of *Hamza, Tā’ Marbūṭa Mutatarrifa*, diacritics etc. can return to a unified form which can achieve a higher value of retrieval.

The precision value of the normal search, on the other hand, was affected by the short strings “عن” *'an* ‘about’ and “هل” *hal* ‘question particle’. They include a small number of characters that can be a part of any other strings. This is the reason behind retrieving irrelevant strings which gave low values of precision in these cases. For example, the string “هل” is a part of the word “لاهل”, “يسهل”, “جهلاء” and “سهلا” where all were retrieved but they are irrelevant to the search form. This resulted in a low precision for the short word “هل” 7.98%. The other queries with longer strings achieved high values of precision ranging between 97.22% and 100.00%. For instance, the word “العطلة” *'al'uṭla* ‘the holiday’ shows 100.00% in both precision and recall. A suggested solution for returning a high value of precision is mentioned below after discussing results of the Separate Words option.

The *Separate Words* option returns only those results exactly matching the query string without any difference in the string form including the existence of prefixes and suffixes. This restriction makes all results retrieved relevant to the query, except one case where the query “هل” *hal* ‘question particle’ returned the word “والاهل” that included an incorrect space between the word characters; consequently, the second

¹ The characters related to the query form are in black colour, and other characters are in red.

part of this word matched the query string. Based on these findings, the *Separate Words* option achieved a high value in the precision measure (99.89%). On the other hand, the requirement of an exact match between the query string and the retrieved results in this type of search resulting in the engine's failure to retrieve a number of relevant results; for example, the *Separate Words* option failed to identify 41.22% of instances of the word "خاصة" *ḥāṣṣa* 'special' e.g. "الخاصة" and "والخاصة", and 38.10% of the word "واجهت" *wāḡahtu* 'I faced' e.g. "واجهتي" and "فواجهت".

As a suggested solution for returning high values for the recall and precision measures in both: the normal search and the *Separate Words* option, the ALC Search Tool needs further development to adapt a lemmatised and PoS-tokenised version of the ALC in the search function. This may assist in retrieving more relevant results, as if the search can find all relevant tokens and distinguish them from the prefixes and suffixes (e.g. the token "خاصة"), it can retrieve all matches regardless the different forms of *Tā' Marbūṭa Mutatarrifa* "ة" (e.g. "خاصة", "خاصة" and "خاصت") and whether the word have prefixes and/or suffixes or not (e.g. "الخاصة", "والخاصة" and "وخاصة").

To sum up, the normal search shows 95.23% in recall and 84.20% in precision. Conversely, the *Separate Words* search achieved 99.89% in precision and 77.58% in recall. Those results showed similar values in F-measure (85.06% for the former and 86.87% for the latter). Developing the ALC Search Tool to operate a normalised, lemmatised and PoS-tokenised version of the ALC may assist in achieving higher values in the recall and precision measures. Finally, those results of the ALC Search Tool are not intended to be compared to other tools, as this tool was designed only for this study.

6.8.2 Specialists' Views

In order to evaluate the ALC Search Tool based on the views of specialists, a short questionnaire was distributed to nine researchers from different universities in Saudi Arabia and the UK specialising in the different research areas of computer science, computer-assisted mobile learning, linguistics, and applied linguistics. Seven of them responded to the questionnaire (Table 6.13).

Table 6.13: Evaluators of the ALC Search Tool

#	Research area	University
1	Computer Science	Taibah University (work) –University of Leeds (study)
2	Computer Science	University of Leeds (study)
3	Computer Science	Jazan University (work) – Heriot-Watt University (study)
4	Computer-Assisted Mobile Learning	Al-Imam University (work) – University of Liverpool (study)
5	Linguistics	University of Leeds (study)
6	Applied Linguistics	Al-Imam University (work) – York University (study)
7	Applied Linguistics	Al-Imam University (work)

The questionnaire included eight questions, mostly open-ended, that asked the respondents about the advantages and disadvantages of the website, using the determinants, using the download function, the user’s guide, and the ease and efficiency of the searching functionality in both versions, Arabic and English. Table 6.14 shows the evaluators’ responses to each question in the questionnaire.

Table 6.14: Summary of the evaluators’ responses to the questionnaire about the ALC Search Tool

Q1. What are the main features you liked in the website?

Responses summary:

1. The diversity of available alternatives to obtain the corpus or parts of it
2. The well-designed and attractive user interface with perfectly chosen colours
3. Ajax supporting
4. Updating fields according to the provided query
5. The ease of searching the corpus
6. The high speed of retrieving the results
7. Having Arabic and English interfaces
8. Using the determinants to search a part of the corpus data
9. The instructions of the user’s guide are clear as well as the illustrations
10. The diversity of file formats for download
11. Highlighting the search word in results
12. Data diversity and richness

Quotations:

- *“It is designed in an organised and clear way to any user, so he does not need previous knowledge in searching corpora”.*
- *“The easy use of this tool enables researchers in linguistics, language learning and acquisition in particular to benefit from its data”.*
- *“Using the determinants saves much time, especially when searching a specific group of ages or gender for example”.*
- *“Highlighting the search word in the results is very positive. It helps the researcher [have] more focus on the target word and the context. It also helps researchers and language learners to study particular words among their structures”.*
- *“It is a rich and trusted resource, from which researchers can obtain language learning data, written and spoken, from different ages and nationalities”.*
- *“This tool tempts researchers to consult the inspiring corpus from which they can draw new ideas for their studies”.*

Q2. What are the main shortcomings (improvement points) that should be considered in the future?

Responses summary:

1. Enabling the user to search strings of words in addition to a single word
2. Adding a part-of-speech determinant with a tagged version of the corpus

Quotations:

- *“Users may need to search for a sentence or phrase, so it is worthy to work on this feature”.*
- *“In design, logos of Leeds and Al-Imam universities are smaller than the other components on the website”.*
- *“I can expect from the effort put on this project that it will be a destination for researchers of Arabic Language Acquisition, so it is very useful to have a Part-*

of-Speech tagged version with a determinant next to the search box to select the word type. For example, one of the distinctive topics that can be studied using such feature is the use of the particles ‘في’ ‘in’ and ‘على’ ‘on’ by Chinese learners of Arabic between 20-30 years old’.

- *“I have not seen any cons worth mentioning”.*
- *“I cannot see any shortcomings”.*

Q3. Do you think it is useful to use the determinants for searching the corpus?

And why?

- *“Yes, because they help to remove irrelevant information from the retrieved information”.*
- *“Yes, it is useful, as the determinants are consistent with the metadata of the corpus”.*
- *“This large number of precise determinants is a very positive point. From my experience with Arabic corpora I have not seen such effort on such a large number of determinants”.*
- *“Yes, because they can be used to focus on a particular part of the data or to undertake a comparative analysis”.*
- *“Yes without doubt, and as a researcher in corpora of Arabic and English, I think it is a creative mechanism. They are a substantial factor to search and analyse differences in the data based on these various determinants”.*
- *“(1) it should be of great importance for promoting the corpus among users with interest in particular parts of the corpus, (2) making the corpus searchable with possibility to download the search results will give this corpus a great advantage among other corpus (if any), (3) it works perfectly when switching from a determination to another”.*

Q4. In general, how easy and efficient it is to search the corpus using this website?

- *“It is easy and efficient”.*
- *“It is excellent in the current stage; it may need more improvement in future”.*

especially when adding more data”.

- *“Easy, flexible and fast. In addition, existence of the determinants would significantly reduce the time needed for the analysis and assists in achieving various research aims”.*
- *“By testing the website, I found that it was easy to use, and help to concentrate on any part of the learners’ language”.*
- *“The use of the website is very easy even to non-specialists”.*
- *“Easy to use and seems to work perfectly”.*

Q5. Did you find any differences between the Arabic and English sites in terms of searching functionality?

- *“I did not notice any differences”.*
- *“No difference, and having English interface is good for those for whom Arabic is not their first language”.*
- *“I never found a difference”.*
- *“Most things I tested were on the Arabic version, but when I used the English version I found that there is no difference”.*
- *“I have used both versions, Arabic and English, and found no difference between them which is one of the website features”.*

Q6. To what extent do you think the files download function is useful, which enables the user to download a subset of the corpus files based on the determinants?

- *“Great idea”.*
- *“The method of downloading the files is good and facilitates the use of data by external tools”.*
- *“Very excellent, this will attract more people to use the corpus”.*
- *“Appropriate”.*
- *“Such function is important. It enables the user to download the files needed*

based on the determinants. It is clear and easy to use”.

- *“It works great and output saved into an easy to use/parse XML format”.*

Q7. Are the guidelines adequately clear to help in using the website? Do you have any feedback about them?

- *“Very helpful and clear for this stage”.*
- *“Yes, the guide is clear and includes all information needed to search the website. It would be more beneficial if the guide includes some information about the features of the file formats and their use; this would help the researcher in selecting the appropriate format which serves his research aim”.*
- *“Very clear with no complexities, any researcher with no computational background can go through it step by step to do any search”.*
- *“From my point of view, I think it is clear and helps in utilising the website”.*
- *“I have just had a quick look at it and seems informative and clear enough”.*
- *“I did not look at it yet”.*

Q8. Further comments

- *“Great work”.*
- *“I’m sure that lots of people will benefit from this work”.*
- *“Very great effort, I’m going to use this in my research on vocabulary”.*
- *“I found it really great and useful”.*
- *“I can summarise my final comment in a few words: the website is ready to use”.*

As seen from the responses of the evaluators, they provided highly positive feedback and valuable comments. The feedback and comments will be used to improve the functionality of the ALC Search Tool in the future development.

6.8.3 Website Visits

The hosting statistics of the ALC Search Tool showed that it received 51,932 visits from 25 November 2014 to 25 March 2015 (four months) from 75 countries around the world. The highest numbers of visits were from the UK (31,656), the United

States (4529), and Saudi Arabia (2308) respectively. These statistics may indicate high interest in using this tool to search and download the ALC, which adds more importance to the future improvements and features intended to be added to this tool.



Figure 6.25: Map showing locations of the ALC Search Tool visitors¹

6.9 Features and Limitations

One of the tool's features is that further materials collected and added to the ALC database of the website will be immediately searchable. Another feature is that the determinants values and number of options are all changeable to meet any future requirements. In terms of limitations, due to technical difficulties, the current version cannot process more than a single word in each query; if two or more forms are entered, no results will appear, as the search considers multiple-word queries as a single word, i.e. spaces between words are not read as spaces, while the corpus is tokenised based on spaces between words. This leads to no results matching the query form. The capability of processing more than one form will be added to one of the future versions.

¹ The map was obtained from the free service StatCounter (<http://www.statcounter.com>) on 25 March 2015.

6.10 Conclusion

This chapter illustrates the first version of the ALC Search Tool for the Arabic Learner Corpus, which was designed to assist users in searching the corpus or a subset of its data and to download the files of any sub-corpus based on a number of determinants. The ALC Search Tool was created to be a free-access, web-based tool. It has an interface in two languages, Arabic and English, with full translations of labels and buttons as well as the ability to switch the entire website layout to be right-to-left. This design may help a wider audience to benefit from the data of the ALC. Determinants used in this tool are classified into three types: determinants with a numerical range value, determinants with a multi-selection list, and determinants with two options (“Yes” or “No”). A user guide was also created to give an overview of the tool and an illustration of how to use it and to take advantage of its functions.

This chapter explains the mechanism of the functions of searching and downloading the corpus. The last section presents three types of evaluation: (i) evaluating the accuracy of the output, (ii) specialists’ feedback, and (iii) statistics of the website visits. The section detailing the evaluation of the accuracy of the ALC Search Tool’s output covers two aspects (i.e. recall and precision) with a confusion matrix containing four elements: true positive, true negative, false positive, and false negative. Those elements assisted in measuring the accuracy through precision, recall, and F-measure of two types of search: the normal search function and the *Separate Words* option. Evaluating the accuracy of the normal search revealed that it obtained 95.23% in recall and 84.20% in precision. In contrast, the *Separate Words* search achieved 99.89% in precision and 77.58% in recall. Those results showed similar values in F-measure: 85.06% for the normal search and 86.87% for the *Separate Words* search. Seven researchers in different specialties participated in the questionnaire and evaluated a number of aspects including the pros and cons of the website design and functionality, the utility of using the determinants in the search and download functions, and the user’s guide, in addition to other aspects. The evaluators provided very positive and valuable feedback, comments, and suggestions to improve its functionality in the next versions. In addition to the specialists’ evaluation, the website’s statistics show that the ALC Search Tool received more than 50,000 visits in the first four months, which reflects the level of interest in using this tool.

Part IV

ALC Uses and Future Work

Summary of Part IV

This part highlights the value of the ALC project through a number of works that have used the corpus in various research areas such as Arabic natural language processing, Arabic applied linguistics, Arabic linguistics, and data-driven Arabic learning. The potential uses of the ALC in further research areas are also explored including automatic Arabic readability assessment, OCR, teaching materials development, and Arabic learner dictionaries. After this exploration, this part summarises the ALC project's contributions, describes plans for future work on each component of the ALC project, and discusses challenges faced during the research before presenting the conclusion of this experimental work.

7 Uses of the Arabic Learner Corpus

Chapter Summary

This chapter describes examples of those projects that have used the ALC for different purposes such as error detection and correction, error annotation guidelines, native language identification, evaluating Arabic morphological analysers, and applied linguistics. The ALC was also used for Arabic teaching and learning activities including, for example, a workshop on teaching Arabic and some data-driven Arabic learning activities. The chapter also explores potential uses of the ALC in further research areas such as automatic Arabic readability assessment, OCR, teaching materials development, and Arabic learner dictionaries. These potential uses offer additional insight into how future researchers might use the ALC.

7.1 Introduction

As previously mentioned, the ALC was intended to be used in various computational and linguistic research areas. We used different strategies to publicise and disseminate the ALC, (i) by creating the ALC website¹ from which the users can download the corpus data and access more details about the ALC project, (ii) by creating the ALC Search Tool² which enables users to search and download any part of the corpus using a number of determinants, (iii) by uploading the ALC data to further tools, Sketch Engine and arabiCorpus which provide additional functions for searching and analysing the corpus, (iv) by publishing papers at a wide range of conferences in different disciplines, e.g. Arabic NLP, learner corpora, Applied Linguistics, Second Language Acquisition and Foreign Language Teaching, (v) by posting information about the ALC to the CORPORA, ARABIC-L and other discussion-lists, (vii) and by making some YouTube videos³. This has led to wide publicity, dissemination and re-use of the ALC resources.

This chapter describes relevant work that has used the ALC. It highlights select examples that have made use of the ALC, although several studies have cited the corpus as related work. The chapter also describes further uses in which the ALC can play a substantial role.

7.2 Projects That Have Used the ALC

The ALC has been used for different purposes and in various applications. For instance, researchers have used it for error detection and correction tools (Farra *et al.*, 2014; Obeid *et al.*, 2013), error annotation guidelines (Zaghouani *et al.*, 2014), native language identification (Malmasi & Dras, 2014), and evaluating Arabic morphological analysers (Alosaimy, Alfaifi and Alghamdi, forthcoming). Other researchers, such as Alshaiban and Alshehri, are currently using the ALC data as a sample for applied linguistics studies for their PhD theses. Additionally, the ALC has been the focus of some practical activities such as a workshop on teaching Arabic (Alharthi, 2015) and data-driven Arabic learning (Refae, personal communication, 22 February 2015; Isma'il, personal communication, 4 April 2015).

¹ The ALC website can be accessed from: www.arabiclearnercorpus.com

² The ALC search Tool can be accessed from : www.alcsearch.com

³ Those videos can be accessed from:

<https://www.youtube.com/channel/UCjJXbOzBA6cvglMNRaqltnw>

Following are more details about the various capacities in which the ALC has been used.

7.2.1 Error Detection and Correction

Linguistic errors are most likely to occur in language produced by learners, which makes learner corpora the most appropriate dataset for performing research in areas such as error detection and correction. The ALC provides an accurate and evaluated version of the ETAr (v3). This error tagset was applied to the ALC by annotating a part of the corpus data manually. When the annotation of the entire corpus data is completed, the ALC will provide a valuable source for training and testing error detection and correction systems. Additionally, the error annotation goes beyond classifying errors into spelling or grammatical, which is common in such systems; instead, it includes a wider classification of errors into five categories which are well-known by Arabic linguists: orthographical, morphological, syntactic, semantic, and punctuation errors. Each category includes a number of sub-type errors which assists in drawing a comprehensive picture of the most common errors made by Arabic learners.

The ALC was utilised in building a web-based, language-independent annotation framework used for manual correction of a large Arabic corpus (Obeid *et al.*, 2013). This framework provides interfaces for annotating text and managing the annotation process. It is able to speed up the annotation process by employing automated annotators to fix basic Arabic spelling errors.

Data of the ALC was also used in the development of the Generalised character-level Spelling Error Correction model (Farra *et al.*, 2014). This generalised discriminative model for spelling error correction targets character-level transformations and uses supervised learning to map input characters into output characters in context. This model learns common spelling error patterns automatically, without guidance of manually selected or language-specific constraints.

Those examples described above highlight the contribution of the ALC to error detection and correction systems.

7.2.2 Error Annotation Guidelines

The Arabic Learner Corpus includes 1585 authentic written and spoken samples of learner data. This authenticity enables researchers to develop their standards based on the ALC data.

The QALB Annotation Guidelines (Zaghouani *et al.*, 2014) is an example of such use. These guidelines consists of seven sections explaining a number of aspects such as annotation goals, text-specific annotation rules, various error categories with illustrated examples (more than 50 examples of errors with their corrections), and a reference summary for selected Arabic spelling rules. The ALC was utilised as a data source in preparing the QALB Annotation Guidelines (Zaghouani, personal communication, 2 April 2015).

7.2.3 Native Language Identification

The Arabic Learner Corpus covers 66 different first languages. This variety in L1s has encouraged some researchers to test their tools on the ALC data, for example those for predicting a writer's first language from his writing.

Malmasi and Dras (2014) used the ALC data in developing their native language identification application.

“[W]e present the first application of Native Language Identification (NLI) to Arabic learner data. NLI, the task of predicting a writer's first language from their writing in other languages has been mostly investigated with English data, but is now expanding to other languages. We use L2 texts from the newly released Arabic Learner Corpus and with a combination of three syntactic features (CFG production rules, Arabic function words and Part-of-Speech n-grams), we demonstrate that they are useful for this task. Our system achieves an accuracy of 41% against a baseline of 23%, providing the first evidence for classifier-based detection of language transfer effects in L2 Arabic. Such methods can be useful for studying language transfer, developing teaching materials tailored to students' native language and forensic linguistics” (Malmasi and Dras, 2014: 180).

7.2.4 Development of Robust Arabic Morphological Analyser and PoS-Tagger

A number of morphological analysers and PoS-taggers have been developed for Arabic, but are generally targeted and evaluated on well-formed, published MSA. Alosaimy, Alfaifi and Alghamdi (forthcoming) are using the ALC and a range of other Arabic corpus genres to evaluate robustness of the main existing Arabic analysers.

7.2.5 Applied Linguistics

The ALC contains written and spoken materials by Arabic learners, native and non-native speakers, from different ages, genders, nationalities, mother tongues, proficiency levels, and with different text genres, modes, mediums, and other production conditions. This diversity in the corpus data is a strong basis for conducting a variety of research in applied linguistics. Researchers are able to undertake different investigations and comparisons on vocabulary and the structures of learners' language using the ALC.

For instance, the corpus has inspired Alshaiban (in progress) to investigate the grammatical competence of learners of Arabic as a second language in his PhD study. Alshaiban aims to investigate grammatical structures that learners of Arabic use in order to identify the extent of grammatical competence in their language. This investigation uses the ALC data, the written texts produced by NNS learners in particular.

The ALC data also inspired Alshehri (in progress) to do his PhD thesis in applied linguistics on the topic of grammatical coherence and textual cohesion in the learner corpus of Arabic as a second language. The study aims to investigate the role that particles play in grammatical coherence and textual cohesion in Arabic as a second language. This covers some aspects such as which of those particles are used by Arabic learners, which are the most frequently used, and to what extent they are used correctly. Such a study might be a fundamental basis for creating pedagogical materials that can lead learners of Arabic as a second language towards more efficient use of those particles.

The corpus also led Alqawsi (personal communication, 1 April 2015) to start a joint research study on the most frequent words in some applications of social media. The ALC will be used to extract words which will be investigated in this study.

Additionally, the ALC was one of the elements that encouraged a research team to start their research on the influence of using corpora on Arabic learners' motivation (Alharthi, personal communication, 13 April 2015), where the Arabic Learner Corpus is used as one of the main samples along with other corpora in this study.

These examples described above highlight the contribution of the ALC to the Applied Linguistics domain.

7.2.6 Workshop on Teaching Arabic

As an indication of the high usability of the ALC for research, a workshop held by Maha Alharthi – at the Princess Nora Bint Abdulrahman University, Riyadh, 3 March 2015 – entitled “Applications of Using Arabic Corpus in Teaching Arabic as a Second Language” (Alharthi, 2015) explained those applications based on examples derived from the ALC. The workshop also highlighted the capabilities of the ALC for many research purposes. Specifically, the workshop recommended that the following research topics could be studied using the ALC (Alharthi, personal communication, 13 April 2015):

- Investigating the properties of written language of Arabic learners (non-native speakers of Arabic) compared to their spoken language in order to test the assumption of whether their spoken language is influenced by properties of the written language;
- Investigating instances of underuse, overuse, and misuse in the language of Arabic learners compared to native speakers in their vocabulary and structures;
- Studying the impact of the age factor in acquiring Arabic as a second language;
- As the ALC contains production of learners representing 66 different first languages, the role of first language on learning Arabic can be investigated to identify whether L1 is an assisting factor in learner Arabic, and whether similarities and differences between Arabic and those languages in some linguistic phenomena contribute positively or negatively to the learning process;

- Comparing different groups of learners based on years they spent in learning Arabic, which may answer the question of whether a longer period of learning Arabic indicates a higher proficiency level;
- Studying the linguistic errors that learners made, and whether the frequency of those errors differs based on a factor such as text genre (narrative texts vs. discussion texts);
- Investigating the influence of using language references and dictionaries on the writing level, by comparing texts where references and dictionaries were used to those where such references were not used; and
- Measuring the extent to which the place and timing factors may affect the text produced; for instance, researchers may investigate whether those texts which were written in class and during a specific time (about one hour) were of lesser quality than those written at home where learners had more time (one or two days) to complete their texts and consequently an opportunity to improve their writing.

This list of research topics recommended specifically to be conducted on the ALC offers additional insight into how future researchers might use the ALC.

7.2.7 Data-Driven Arabic Learning

Some Arabic language teachers who were interested in using the ALC in data-driven language learning activities have contacted the researcher. Johns and King (1991) define this type of language learning as:

“the use in the classroom of computer-generated concordances to get students to explore the regularities of patterning in the target language, and the development of activities and exercises based on concordance output” (p iii).

Refaee (personal communication, 22 February 2015) from Saudi Arabia, for example, used the ALC data to improve her students’ writing in Arabic. She developed pedagogical activities based on the ALC data where students had the opportunity to identify correct and incorrect structures of Arabic writing from those activities.

Similarly, Isma'il (personal communication, 4 April 2015) from Egypt started a learning project where students were able to use language resources such as the ALC data for further learning about vocabulary and structures of Arabic language. For instance, the learners were asked to use some of those structures derived from the ALC in their own writing.

7.3 Further Uses of the ALC

The ALC can be used in further research areas such as automatic readability research, OCR, teaching materials development, and Arabic learner dictionaries.

7.3.1 Automatic Arabic Readability Research

According to Altamimi *et al.* (2014), the term *text readability* refers to the ability of the reader to understand and comprehend a given text. Text readability systems are usually trained on a pre-graded set of texts. As examples of those datasets, Altamimi *et al.* (2014) have trained their system on more than 1196 Arabic texts in different subjects extracted from the Jordanian curriculum from first grade through tenth grade. Another example is Alkhalifa and Alajlan (2010), who relied on a corpus comprising 91 webpages written by students or adults across three levels: Kindergarten – Grade 2, Grade 3 – Grade 5, and Grade 6 – Grade 8. Additionally, Forsyth (2014) used the Defense Language Institute corpus which contains 179 documents ranked by the authors into five proficiency levels: 1, 1+, 2, 2+, and 3 from easiest to most difficult according to the Inter-agency Language Round table standard levels.

The ALC is a suitable resource for undertaking research on readability systems, as its data includes different types of grading from general levels to more specific levels. For example, the category addressing the general level of education classifies learners into two main levels: pre-university and university. The level of study category includes five grades: secondary school, general language course, diploma programme, bachelor degree, and master degree. The year/semester classification indicates the levels used in learners' institutions. Table 7.1 illustrates how those level indicators fit together in one scale with three hierarchical degrees of levels.

Table 7.1: Three hierarchical degrees of level indicators in the ALC

General level	Level of study	Year/Semester	
Pre-university	Secondary School	First year	
		Second year	
		Third year	
	General Language Course	Third semester	
		Fourth semester	
	Diploma Language Course	First semester	
		Second semester	
		Third semester	
		Fourth semester	
	University	Bachelor degree	First semester
			Second semester
			Third semester
Fourth semester			
Fifth semester			
Sixth semester			
Seventh semester			
Eighth semester			
Master degree		First semester	

The ALC data is graded using these three levels. Text readability systems can be trained based on any of those degrees.

7.3.2 Optical Character Recognition Systems

OCR is one of the applications that can benefit from using the ALC as training data. Three-quarters of the ALC (76%) texts are hand-written texts in PDF format, and their transcriptions are provided in computerised formats (TXT and XML). The availability of such data allows OCR systems to learn from authentic data which contains different types of handwritings in addition to different types of errors, which may lead OCR systems to achieve greater levels of accuracy.

7.3.3 Teaching Materials Development

Granger (1998) believes that the efficiency of language tools could be improved if teaching materials designers relied not only on data from authentic native speakers which gives information about what is typical, but also on authentic learner data,

which highlights what is difficult for learners in general and for specific groups of learners.

As an example of this, we extracted a number of concordances from a corpus of native Arabic speakers: the KACST Arabic Corpus (Althubaity, 2014) and the same number from a corpus of Arabic learners: the Arabic Learner Corpus (Table 7.2). Those concordances show the word “بالنسبة” *binnisba* ‘regarding’ with its contexts in both corpora. The table reveals that the typical prepositions following the word “بالنسبة” in the native corpus are “لـ” *li* ‘for’ and “إلى” *’ilā* ‘to’, while the learners used the preposition “في” *fi* ‘particle’ and the nouns “أهل” *’ahl* ‘people’ and “أسرة” *’usra* ‘family’. Designers of teaching materials can benefit from such an example to develop materials that help learners develop a more efficient use of the language vocabulary and structures.

Table 7.2: Concordances of the word “بالنسبة” *binnisba* ‘regarding’

From a native speaker corpus: <i>The KACST Arabic Corpus</i> (Althubaity, 2014)		
له شراً فيخرج من هذه الجهة عن كونه	بالنسبة	وأما العبد فقد يريد الشيء ويكون
إلى الزمن بعامة -الزمن المطلق- لا شك أنه	بالنسبة	وإن كانت مدته أو عمره طويل لكنه
لنا سنجري على وفق ما جرى هو عليه	بالنسبة	وفق ما تيسر لمؤلفها، والترتيب ينفع المتلقي لكن
للجهات المانحة للمساعدات و الممولة للبرامج	بالنسبة	و أصبحت المخاطب المفضل
للمنتجين أو المستهلكين كما إن المعرفة قد تلعب	بالنسبة	فرص الاختيار بين السلع والخدمات سواء
لمن بعدهم . فالصحابة رضي الله عنهم	بالنسبة	وهو قلة كلام السلف وعظيم فقههم
لتخريج الأحاديث فإن كان الحديث في الصحيحين	بالنسبة	ومن ثم اختيار القول الراجح في كل صورة . أما
إلى الفقه	بالنسبة	الذي هو بالنسبة إلى النحو كأصول الفقه
إلى الجيش الإسلامي فقد كان قليل العدد من	بالنسبة	مما أضعف عزيمة أفراده . أما
للحافلات السياحية فإن التحسن السياحي في	بالنسبة	وبالعادة يفضل هؤلاء السير على الأقدام . أما
From a learner corpus: <i>The Arabic Learner Corpus</i>		
لي بعيداً، ولكن هدفي وإرادتي وعزمي	بالنسبة	ولهذا أرى النجاح إلى الآن في هذا التخصص
للطعام والأجرة و عدة مشاكل أخرى	بالنسبة	مشكلة مع العائلة التي كانت تأويني
إلى التخصص الذي اخترته فهو التخصص العلمي	بالنسبة	اهتماماتي الدراسية هي كثيرة وعديدة ولكن
أهل بلدي، معظمهم وأكثرهم محتاجون إلى الدعوة	بالنسبة	وذلك أن الحاجة تدعو إليها
للمعهد سألتحق بكلية أصول الدين بإذن الله	بالنسبة	بعد دراسة اللغة العربية
لي فقد قمت بوصفها لك واتمنى أن	بالنسبة	هذه هي قصة حياتي وأجمل قصة
في اختياري له فليس له سبب	بالنسبة	و بإذن سيحقق حلمي الذي أريده، أما
الي، فلازلت استشعر ذلك الموقف	بالنسبة	فقد كانت من أكثر المواقف روحانيه
في كليات أخرى	بالنسبة	و ايضاً في كلية الشريعة إستفادة كثيرة
أسرتي هم يقولون أي التخصص أريد	بالنسبة	أو يرغب بهذا التخصص.

7.3.4 Arabic Learner Dictionaries

More recently, developers of learner dictionaries have utilised learner corpora to improve the contents of their dictionaries by warning learners against the most common errors at the end of relevant entries. These dictionaries also suggest the

ways in which a word or an expression can be used correctly (Granger, 2003b; Nesselhauf, 2004).

The ALC adopts a novel error taxonomy with a tagset that has been applied to a part of the ALC (10,000 words, 3.5% of the corpus data). When the entire corpus is being tagged for errors using this suggested tagset (within two to three years and by three annotators who have experience in teaching Arabic to both native and non-native speakers), the ALC will provide developers of Arabic learner dictionaries with substantial information about the most common errors in the language of Arabic learners, in addition to a classification of those errors under 6 major categories encompassing 29 error types. Table 7.3 lists the 10 most frequent errors in the annotated part of the ALC using the third version of the ETAr. Table 7.4 shows the same information but classified based on the nativeness factor (NNS vs. NS). The availability of information about the common errors in learners' language can also lead to the creation of a common error dictionary for Arabic in much the same way the *Longman Dictionary of Common Errors* (Turton & Heaton, 1996) functions for English learners.

Table 7.3: The 10 most common errors in a 10,000-word sample of the ALC

Error category	Error type	% *
Punctuation	Missing punctuation	23%
Orthography	<i>Hamza</i> (ء، ا، إ، ؤ، ئ، ة)	19%
Semantics	Word selection	7%
Punctuation	Punctuation confusion	7%
Syntax	Redundant word	5%
Syntax	Definiteness	5%
Orthography	Confusion in <i>Hā'</i> and <i>Tā' Mutatarrifatain</i>	5%
Syntax	Missing word	5%
Semantics	<i>Faṣl wa Waṣl</i> (confusion in use/non-use of conjunctions)	4%
Orthography	Missing character(s)	3%
		83%

* Percentage of the most common errors to the whole sample

Table 7.4: The 10 most common errors based on the nativeness factor

No	Non-native speakers			Native speakers			
	Error category	Error type	%	Error category	Error type	%	
1	Punctuation	Missing punctuation	18%	Orthography	<i>Hamza</i> (ء، أ، إ، ؤ، ئ، ة)	28%	
2	Syntax	Definiteness	12%	Punctuation	Missing punctuation	26%	
3	Semantics	Word selection	11%	Orthography	Confusion in <i>Hā'</i> and <i>Tā' Mutatarrifatain</i>	7%	
4	Syntax	Redundant word	10%	Punctuation	Punctuation confusion	6%	
5	Syntax	Missing word	8%	Semantics	Word selection	5%	
6	Punctuation	Punctuation confusion	8%	Syntax	Case	4%	
7	Syntax	Gender	5%	Semantics	<i>Faṣl wa waṣl</i> (confusion in use/non-use of conjunctions)	4%	
8	Semantics	<i>Faṣl wa waṣl</i> (confusion in use/non-use of conjunctions)	4%	Orthography	Missing character(s)	3%	
9	Orthography	<i>Hamza</i> (ء، أ، إ، ؤ، ئ، ة)	4%	Orthography	Replacement in word character(s)	3%	
10	Morphology	Word inflection	4%	Syntax	Redundant word	3%	
			84%				88%

7.4 Conclusion

This chapter illustrates various uses of the ALC by highlighting projects that have used the corpus data and describing further projects that might be able to utilise it for different purposes. Projects that have used the ALC include computational applications such as the web-based, language-independent annotation framework, the Generalised character-level Spelling Error Correction model, the QALB Annotation Guidelines, and the application of a native identification system to Arabic learner data. The ALC has also been used in applied linguistics research projects to investigate grammatical coherence and textual cohesion in the learner corpus of Arabic as a second language, also to study grammatical competence of learners of Arabic as a second language. The authors of both studies are currently conducting their PhD research degrees. Additionally, a research team has included the ALC in the sample for a new study entitled *Influence of Using Corpora on Arabic Learners' Motivation*. The ALC materials were also used as a sample for the workshop – at the Princess Nora Bint Abdulrahman University, Riyadh, 3 March 2015 – entitled *Applications of Using Arabic Corpus in Teaching Arabic as a*

Second Language. This workshop concluded by offering several recommendations for avenues of research using the ALC. Additionally, the ALC was used in some data-driven language learning activities in order to improve learners' writing in Arabic, and for further learning about vocabulary and structures of the Arabic language.

In terms of potential uses of the ALC in further research areas, the chapter explains how the corpus can be used for automatic Arabic readability research, as its data includes different types of grading from general levels to more specific levels. OCR systems may also benefit from the corpus data, particularly because 76% of the corpus data are hand-written texts which are available with their transcriptions in computerised formats. The chapter gives an example of how the ALC can be a basis for developing teaching materials for Arabic learners. Finally, the chapter describes how a part of the ALC has been annotated for errors using a novel error taxonomy which can be used in Arabic learner dictionaries to provide the users with valuable information about those errors. Through those projects that have used the ALC and the potential uses the chapter suggests, the capability of the ALC and the ways in which it can serve as a basis for many pioneer research subjects in the future are clear.

8 Future Work and Conclusion

Chapter Summary

This chapter summarises the contributions presented in this thesis including a number of resources, proposed standards, and tools that contribute to the domains of Arabic natural language processing and Arabic linguistics. It also summarises the evaluation of the ALC components and describes some plans that have been made for future work on those components, such as the Guide on Design Criteria for Learner Corpus, the Arabic Learner Corpus, the Computer-aided Error Annotation Tool for Arabic, the Error Tagset of Arabic, the Error Tagging Manual for Arabic, and the ALC Search Tool. The chapter discusses the challenges the researcher faced and then presents the conclusion of this experimental work.

8.1 Introduction

Learner corpora have become a popular area of research. Work presented in this thesis represents the first stages of the ALC project, which includes a number of resources, proposed standards, and tools that contribute to Arabic NLP and Arabic linguistics domains. This chapter summarises the contributions presented in this thesis and the evaluation of the ALC components. Continuation of this work by the researcher – and his institute at Al Imam Mohammad Ibn Saud Islamic University – is fundamental not only for improving the project but also for maintaining the usability of the corpus and its components to the highest possible level. Thus, this chapter discusses some plans that have been made for future work on each part of the ALC project.

8.2 Thesis Achievements

The primary aim of the current research was to develop an open-source Arabic learner corpus and a system for Arabic error annotation to be used as a valuable resource for research on language teaching and learning as well as NLP. Chapter 7 in this thesis described examples of those projects that have used the ALC for different purposes such as error detection and correction, error annotation guidelines, native language identification, evaluating Arabic morphological analysers, and applied linguistics. The ALC was also used for Arabic teaching and learning activities including, for example, a workshop on teaching Arabic and some data-driven Arabic learning activities. These uses and potential uses of the ALC – such as automatic Arabic readability assessment, OCR, teaching materials development, and Arabic learner dictionaries – give evidence that the study has achieved its aim.

The study objectives were achieved through a novel set of resources, proposed standards, and tools that contribute to the fields of Arabic NLP and Arabic linguistics. The following list explains how the study objectives were achieved:

1. To review the learner corpora existing under specific criteria

The thesis presents a comprehensive review of 159 previous works (learner corpora) under 11 categories (corpus purpose, size, target language, availability, learners' nativeness, learners' proficiency level, learners' first language, materials mode, materials genre, task type, and data annotation) that provide an idea about

the best practice in this field. Developers of new similar projects and learner corpora users can benefit from this source in their research.

2. To create a guide for developing a new learner corpus

We created a guide for developing a new learner corpus based on a review of previous work. It focuses on 11 aspects of corpus design criteria, such as purpose, size, target language, availability, learners' nativeness, materials mode, data annotation, etc. Our aim is that these criteria will serve as open-source standards for developing new learner corpora. The guide can also be utilised to improve and/or expand the current corpora.

3. To collect data for the Arabic Learner Corpus (ALC) based on its design criteria

The ALC is a standard resource for research on Arabic teaching and learning as well as Arabic NLP. It includes 282,732 words and 1585 materials (written and spoken) produced by 942 students from 67 nationalities and 66 different L1 backgrounds. Based on our examination of the literature, we are confident that the ALC is the largest learner corpus for Arabic, the first Arabic learner corpus that comprises both native Arabic speakers and non-native Arabic speakers, and the first Arabic learner corpus for Arabic as a Second Language collected from the Arab world.

4. To develop an error tagset for Arabic

The Error Tagset of Arabic (ETAr) includes an error taxonomy that was designed based on a number of studies that have investigated the most frequent errors in Arabic learners' production. Additionally, it includes a tagset designed for annotating errors in Arabic covering 29 error types under five broad categories. Seven annotators and two evaluators performed iterated evaluations on this tagset, and the ETAr was improved after each evaluation. The ETAr is intended to be a tool for annotating errors in the ALC as well as in further Arabic learner corpora, particularly those for Arabic language teaching and learning purposes. The ETAr is available to researchers as an open source. It provides target users with easy-to-understand categories and types of errors.

In addition to the ETAr, the Error Tagging Manual for Arabic (ETMAr) was developed to describe how to annotate Arabic texts for errors. It was based on the final revised version of the ETAr. The ETMAr contains two main parts. The first defines each error type in the ETAr with examples of those errors and how they

can be corrected. The second illustrates a method of how annotators can deal with ambiguous instances and select the most appropriate tags.

5. To develop a computer-aided error annotation tool for Arabic

A new tool was developed for computer-aided error annotation in the ALC. It was based on the ETAr and includes some automated features such as the smart-selection function, which finds similar errors and annotates them in a single step with no need to repeat the annotation process for each error, and the auto-tagging function, which is similar to translation memories as it recognises the tokens that have been manually annotated and stores them into a database so that similar errors in other texts can be detected and annotated automatically. Using this tool increases the consistency of error annotation over pure manual annotation.

6. To develop a search tool based the ALC metadata

The ALC Search Tool was established to enable users to search the ALC based on a number of determinants including 26 metadata elements such as “age”, “gender”, “mother tongue”, “text mode”, and “please of writing”. Those metadata elements were utilised as determinants to allow users to search any sub-corpus of the ALC based on the determinants selected and then to download any part of the corpus data (sub-corpus) based on those determinants and in different formats (TXT, XML, PDF, and MP3).

To sum up, this thesis presents a number of resources, tools, and proposed standards developed for the ALC project. However, the major contribution of the thesis is not only the description of these components but also the detailed and original methodology that this thesis presents for developing a new learner corpus. The combination of the aforementioned resources, standards, and tools represents this new methodology.

8.3 Evaluation

The ALC includes 282,732 words in 1585 materials (written and spoken) produced by 942 students from 67 nationalities with 66 different L1 backgrounds. It was evaluated through a number of examples of works that have used the ALC data. The evaluation shows an increasing interest from its first release in 2013 to the time of writing in 2015. Additionally, a questionnaire was used to gather feedback from specialists in related fields. The specialists’ comments about the corpus were highly

positive, which also highlights researchers' interest in using the ALC to conduct research on the Arabic language. This interest was also supported by more than 16,000 downloads from the ALC website over a 12-month period.

Seven annotators and two evaluators worked on the CETAr, ETAr, and ETMAR in order to evaluate their usefulness in annotating Arabic errors. The results achieved in the experiments were highly positive, as shown in Chapter 5.

The CETAr includes a number of features for facilitating the annotation process such as text tokenisation, smart-selection, auto tagging, and others. An evaluation of consistency and speed in the CETAr showed that the annotation time was reduced while the consistency in annotation was increased when using the CETAr in comparison to manually tagging errors; in particular, the smart-selection feature may have played a role in this achievement. Additionally, evaluating the auto-tagging feature revealed an accuracy level between 76% and 95% with an average of 88%.

The ETAr was developed as an error taxonomy for tagging errors in Arabic texts. The third version of this tagset includes 29 error types distributed under 5 categories. Seven annotators and two evaluators have evaluated the ETAr three times for a number of purposes. An evaluation of understandability and usability of the ETAr against the only other existing Arabic tagset, ARIDA, showed that the ETAr achieved an observed agreement rate higher than the ARIDA tagset. Results of the inter-annotator agreement revealed an increase in the results of the second and third experiments, which was due to the improvements that were made following the first evaluation. These improvements include refining the ETAr, creating the ETMAR, and adding training sessions.

The ETMAR was developed for two main functions: to explain the error types in the ETAr and to establish rules for how to select the appropriate tags in error annotation. The ETMAR was used in the second and third evaluations of the ETAr, with the result that the observed inter-annotator agreement increased from the first evaluation to the second and third evaluations.

The ALC Search Tool was designed to assist users in searching the corpus and downloading the files based on a number of determinants. Evaluating the accuracy of the output of this tool revealed that the normal search achieved 95.23% in recall and 84.20% in precision, whereas the *Separate Words* search achieved 99.89% in precision and 77.58% in recall. The F-measure was 85.06% for the normal search

and 86.87% for the *Separate Words* search option. The tool evaluators provided positive and valuable feedback, comments, and suggestions to improve its functionality. In addition, statistics from the website showed that the website received more than 50,000 visits in the first four months.

8.4 Future Work

This section describes future work on the guide on design criteria for learner corpus, the ALC, the CETAr, the ETAr and its manual ETMAr, and the ALC Search Tool.

8.4.1 Guide on Design Criteria for Learner Corpus

Developing the ALC based on this guide represented a practical application which may give researchers an illustration of the extent to which the guide can be used. Future development plans for this guide include the addition of more design criteria, which will be derived from an additional review of other aspects of existing learner corpora such as metadata, more details about file formats, and tools that can be used for each stage of building a corpus. In addition, the researcher will review other learner corpora to update the guide. The development of these design criteria will include issuing a detailed guide that adds to the theoretical information by offering practical steps on constructing a learner corpus based on each design criterion. In doing so, the ALC project is an authentic example that can be used to illustrate the practical aspects.

8.4.2 Arabic Learner Corpus

The first phase of the future work is to add more data to the ALC for two purposes. The first goal is to gather more data from learners with first languages that currently have low representation. The second aim is to achieve a greater balance between some comparable elements of the design criteria such as general level (pre-university vs. university), materials mode (written vs. spoken), materials genre (narrative vs. discussion), and task type (essay vs. interview). The size targeted in the next version of the corpus is 1,000,000 tokens where those elements can have balanced representations.

This phase would involve collecting data from Arabic learners at the beginning level as well, which is not represented in the current version of the ALC. An attempt will

be made to represent the three general levels defined by the Common European Framework of Reference (Council of Europe, 2001) in balance: beginner, intermediate, and advanced. In order to achieve this aim, one of the pre-collection steps will involve the administration of a proficiency test to classify learners into three groups of proficiency prior to the data collection process.

In terms of annotating, a part of the corpus is currently annotated (10,000 words, 3.5%). The researcher has applied for a grant from the King Abdullah Bin Abdulaziz International Center for Arabic Language Service to aid in the annotation of the corpus data. The grant proposal suggested three annotators to work on tagging the entire corpus. In addition to the layers currently exist, the annotation at this stage will add three further layers of annotation: (i) lemma, (ii) PoS, and (iii) Grammatical Function (GF) (see Table 8.1 for an example). The response has not yet been received from the centre.

Table 8.1: Example of the suggested annotation for the ALC

Token	Lemma	PoS	GF	Error Tag	Error Form	Correct Form
ف	ف	PC				
سُرّ	سر	VP				
و	و	RR	NV	OW	و	وا
و	و	PC				
فَرَح	فرح	VP				
و	و	RR	NV	OW	و	وا
؛	؛	UL		PT	؛	null
لما	لما	NV				
سمع	سمع	VP				
و	و	RR	NV	OW	و	وا
مَنْ	من	PP				
ي	ي	RR	GF			
ما	ما	NC	AO			
يسرّ	سر	VC				
أفندت	فؤاد	NQ	AO			

A later phase will aim to create an international version of the ALC. This version would contain parallel corpora following the ALC's design but using texts collected from other Arab countries, such as Egypt, Jordan, Kuwait, United Arab Emirates, Morocco, Sudan, and Lebanon. The Egyptian version may be first as some Egyptian researchers have expressed their interest in participating in this project. Creating these parallel corpora may lead to more comprehensive research on the language of Arabic learners in both the linguistic and computational domains. Additionally, this international corpus may attract more researchers to participate in the corpus development process, as evidenced by such collaboration in international learner corpora, such as the International Corpus of Learner English (Granger, 1993, 2003b; Granger *et al.*, 2010).

8.4.3 Computer-Aided Error Annotation Tool for Arabic (CETAr)

The researcher feels that it is important to perform further development on the CETAr to make it a web-based tool instead of a part of the ALC database as it is currently. Such an online tool for annotating errors in Arabic corpora/texts would be usable by a wider audience of annotators by allowing them to upload their own corpora and use the ETAr. This design would also allow a team of users to work on the same annotation project worldwide. The development needs to take into account the ability to handle Arabic scripts in different browsers and on different operating systems, as these problems were encountered when trying to use the existing online annotation tools.

In a further phase, a user might be allowed to define his or her own tagset to be used not only for error annotation but for further types of text annotation such as PoS, dependency, prosody, and anaphora. Adding this functionality would mean enabling the user to add more than one layer of annotation to the same text. In this phase, the ability to make multi-word annotations might be necessary in order to enable one tag to cover more than one token; consequently, the researcher may need to identify an appropriate methodology for dealing with cases of overlapping tags.

8.4.4 Error Tagset of Arabic (ETAr) and Its Manual (ETMAr)

For the ETAr, further layers can be added to some error types. These additional layers may enable users to conduct deeper error description and analysis. For instance, the error in *Hamza* has several forms based on its position in the word (beginning, middle, and end). At the beginning, it is either *Waṣl* (a phonemic glottal stop) or *Qat'* (a non-phonemic glottal stop) based on its morphological form. In the middle and end, it can be on '*lif* (ل), *Wāw* (و), *Yā'* (ي), *Nabira* (نـ), or on the line (ء). The placement depends on the diacritics of *Hamza* itself and the preceding character. These cases can be added as a further layer under the error type *Hamza*.

Another addition could be the re-introduction of error types covering multiple words such as the stylistic errors that were removed from the ETAr in order to avoid problems of overlapping mark-ups in the files structure. Once the most appropriate method for marking up structures of the corpus files has been determined, those multi-word errors will be represented in new versions of the ETAr.

In terms of the ETMAr, the upcoming version will include more linguistic rules/grammars of error types such as cases of '*alif Fāriqa* (كتبو), distinguishing between *Hā'* (هـ) and *Tā' Mutaṭarrifatain* (هـ), '*alif* (ع / ا) and *Yā' Mutaṭarrifatain* (ي), and *Nūn* (ن) and *Tanwīn* (نّ). The punctuation rules that are described in the current version serve as an example of such additions.

8.4.5 ALC Search Tool

Future work on the ALC Search Tool focusses primarily on three dimensions. The first goal is to improve the precision and recall of the search function by exploiting features such as tokenisation and lemmatisation which may enable the tool to provide high-quality results for the search query and consequently achieve higher precision and recall levels.

The second aim is to add more functions, statistical functions in particular. Making such an addition may involve extracting a list of word frequencies, either before or after any processing steps such as tokenisation and/or lemmatisation. Extracting the collocations from learners' language can be also added as a valuable function with some measures such as mutual information, likelihood ratios, *t* tests, and *z* tests. Other features to be added to this tool include the ability to search and analyse a corpus based on its annotation (e.g. errors and PoS). Although such functions exist

in corpus analysis tools like Sketch Engine, the researcher believes that combining those functions with the search determinants of the ALC Search Tool would result in a more user-friendly tool.

The third goal is to enable users to upload their own corpora to the ALC Search Tool. This feature would likely encourage researchers to develop further Arabic learner corpora and to benefit from the ALC Search Tool, which provides some distinctive features such as using determinants to search the corpus and download its source files, and providing an interface in Arabic with a right-to-left layout in addition to the English one.

To sum up, future work on the guide of design criteria for learner corpus, the ALC, the CETAr, the ETAr, ETMAr, and the ALC Search Tool may reduce the effort usually spent on designing, collecting, annotating, and analysing learner corpora, especially Arabic learner corpora. This future work will result in more benefits for researchers in the form of the resources, standards, tools, and the comprehensive methodology on creating standard learner corpora.

8.4.6 Further Applications of the ALC

An area for future work is to further investigate applications of the ALC. This will allow extending the uses of the ALC cited in Chapter 7.

8.4.7 Dissemination

A part of future work is to promote the ALC and its applications on a range of websites, portals, etc., to further disseminate the resources and results, and hence promote uptake and re-use of the ALC.

8.5 Challenges

During this study, the researcher faced a number of challenges that required rethinking approaches or redesigning experiments. One of the main challenges was the large number of participants needed to produce a reasonable size of data. Creating a large corpus requires the recruitment of more participants. An essential criteria in learner corpora which enables researchers to avoid any distortion in the results of their studies is to collect similar data, which “means that the essays must be written by learners at a similar level under the same conditions and on similar

topics” (Granger, 1993: 61). Therefore, collecting materials that learners had previously produced such as homework or assignments may not be suitable, as the conditions and topics will not be the same for all participants and consequently will lead to distortion in the results. Designing the corpus to include the smallest possible size – 200,000 words as Granger (2003b) suggests – was one possible solution for this challenge. Nevertheless, this study succeeded in recruiting 942 learners in addition to 50 participants who served as data collectors, evaluators, annotators, and collaborators.

It was not possible to start annotating the corpus for errors until completing the evaluation and revision of the ETAr. Reaching the most recent version of the ETAr (v3) required a combination of nine annotators and evaluators to perform three experiments of annotation and evaluation on the previous versions. The next step then was to apply for a grant in order to annotate the corpus for errors manually. The grant proposal requested three annotators who have experience in teaching Arabic to both native and non-native speakers. The proposal further suggested the use of the same methodology used in the third evaluation of the ETAr in order to achieve a high quality of inter-annotator agreement. However, the fund was not obtained within the timescale of the project, as the proposal is still under review¹.

8.6 Conclusion

This thesis presents an original methodology for developing the ALC including a combination of resources, proposed standards, and tools. This methodology may inspire new developers of not only learner corpora but further specialised corpora when building their own projects. The large number of contributors to this work included language learners, data collectors, evaluators, annotators, and collaborators from more than 30 educational institutions in Saudi Arabia and the UK. The use of the ALC in its first years and for multiple purposes highlights the significance of the planned future developments. We think that we are at the beginning of an exciting project for Arabic NLP and Arabic teaching.

¹ See a copy of the project proposal on:
http://www.comp.leeds.ac.uk/scayga/Alfaifi_annotation_grant.pdf

Appendix A

Examples of ALC File Formats

A.1 Plain text files

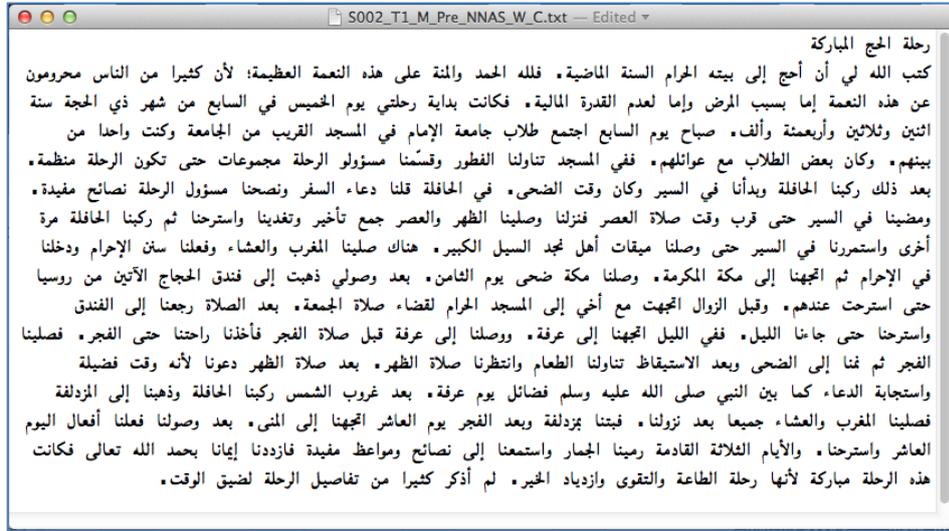


Figure A.1: Example of plain text file with no metadata

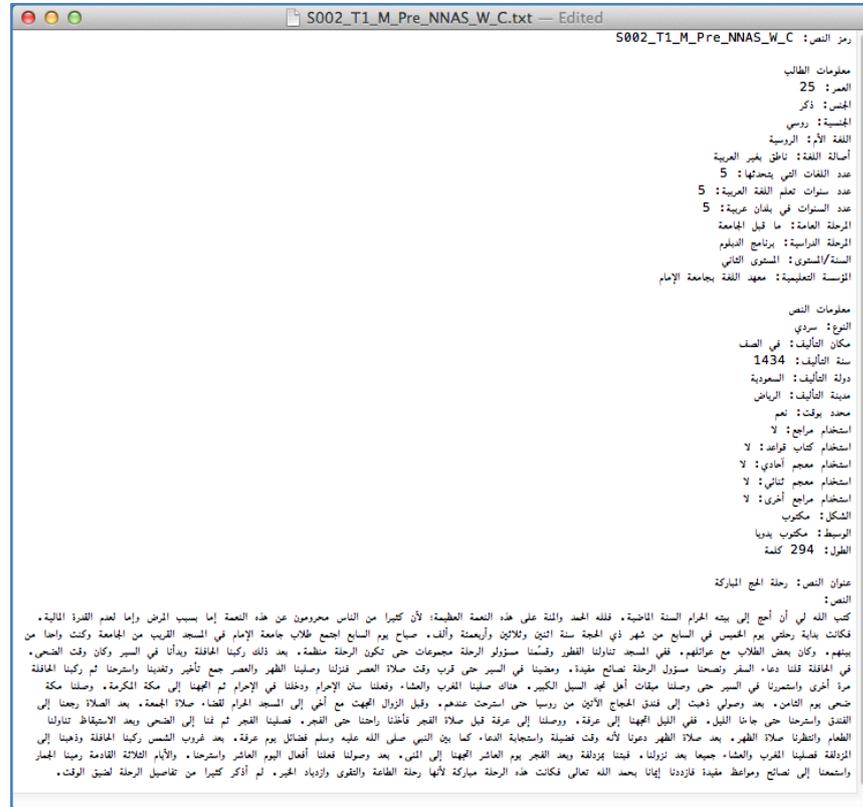


Figure A.2: Example of plain text file with Arabic metadata

Appendix A – Examples of ALC File Formats

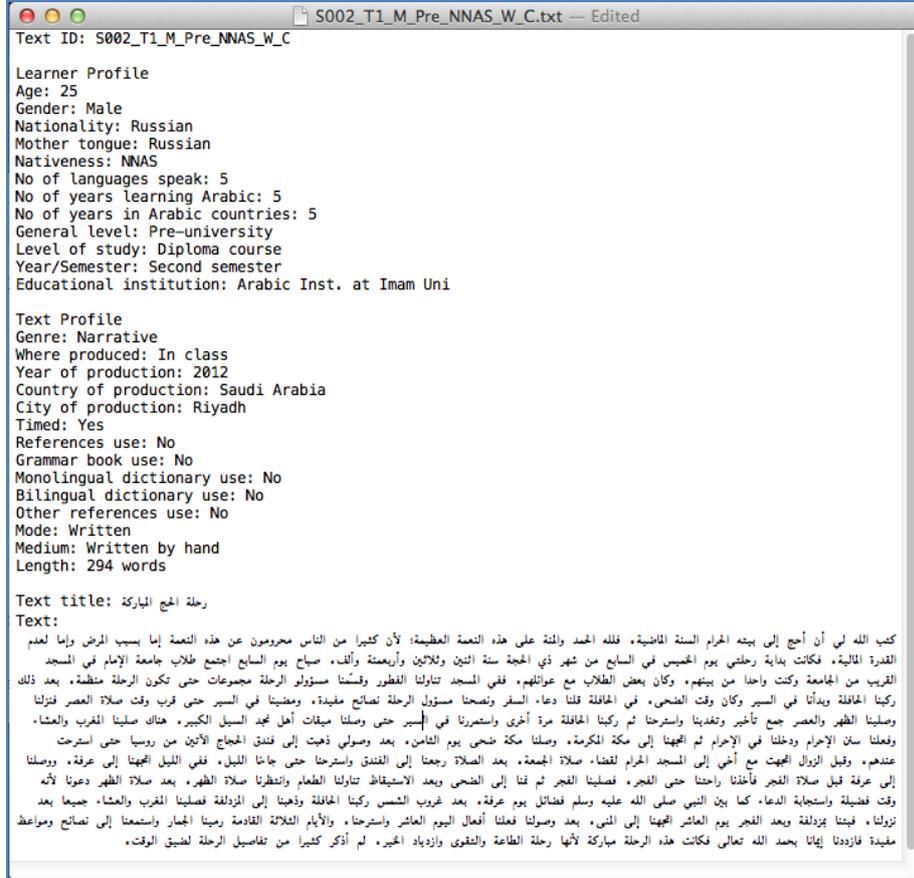


Figure A.3: Example of plain text file with English metadata

A.2 XML files

```
<?xml version="1.0"?>
<!--Arabic Learner Corpus_v2_2014-->
<!DOCTYPE doc>
- <doc ID="S002_T1_M_Pre_NNAS_W_C">
  - <header>
    - <learner_profile>
      <age>25</age>
      <gender>ذكر</gender>
      <nationality>روسي</nationality>
      <mothertongue>روسية</mothertongue>
      <nativeness>متعلم</nativeness>
      <No_languages_spoken>5</No_languages_spoken>
      <No_years_learning_Arabic>5</No_years_learning_Arabic>
      <No_years_Arabic_countries>5</No_years_Arabic_countries>
      <general_level>كما في الجامعة</general_level>
      <level_study>برنامج شهادتي</level_study>
      <year_or_semester>الترم الثاني</year_or_semester>
      <educational_institution>الجامعة الإسلامية</educational_institution>
    </learner_profile>
    - <text_profile>
      <genre>سوي</genre>
      <vwhere>في</vwhere>
      <year>1434</year>
      <country>السعودية</country>
      <city>الرياض</city>
      <timed>نعم</timed>
      <ref_used>ي</ref_used>
      <grammar_ref_used>لا</grammar_ref_used>
      <mono_dic_used>ي</mono_dic_used>
      <bi_dic_used>ي</bi_dic_used>
      <other_ref_used>ي</other_ref_used>
      <mode>مكتوب</mode>
      <medium>مكتوب ورقيا</medium>
      <length>294</length>
    </text_profile>
  </header>
  - <text>
    <title>رحلة الحج المباركة</title>
    <text_body>
      كتب
      الله لي أن أضع إلى بيته الحرم النبوي، فله الحمد والمنة على هذه التلمذة الحقيقية؛ لأن كثيرا من التلمذات محرومون عن هذه التلمذة إما بسبب المرض وإما لعدم القدرة المالية، ففككت يداه ريشي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعمائة وألف، صباح يوم السابع لعين طلاب جامعة الإمام في المسجد العربي من الجامعة وقت واحد من بينهم، وكان بعض الطلاب مع عائلاتهم، فلي المسجد ثلاثا نظفوا وأشكوا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة، بعد ذلك رغبنا الحافلة وبدأنا في السير وكان وقت الضحى، في الحافلة كنا دعاء السفر وإصمحا مسؤولو الرحلة تصاحف مغنوا، وبصحبنا في السير حتى قرب وقت صلاة العصر فزلتوا ومصلينا نظفوا والعصر جمع تأخير وتعدنا واسترحنا ثم رغبنا الحافلة مرة أخرى واستمرنا في السير حتى وصلت بيوت أهل نجد شيل التغيير هناك صلبنا المغرب والشاء وأخذنا سنن الإحرام وبخنا في الإحرام ثم جعلنا إلى مكة المكرمة، وصلنا مكة شحى يوم الثامن، بعد وصايي ذهبت إلى فندق الحجاج الأيمن من روسيا حتى استرحت عندهم، وأول الأرزاق التي جئت مع أخي إلى المسجد الحرام لشاء صلاة الجمعة، بعد الصلاة رجعا إلى الفندق واسترحنا حتى جاعا الليل، فلي الليل جعلنا إلى عرفة، ووصلنا إلى عرفة قبل صلاة الظهر فأخذنا راحنا حتى الظهر، فسلمنا الحجر ثم كنا إلى المشي بعد الاستراحة تناولنا الطعام وارتقنا صلاة الظهر، بعد صلاة الظهر دعنا ولنا وقت فشيئة واستجابة للدعاء ما بين النبي صلى الله عليه وسلم فشدنا يوم عرفة، بعد غروب الشمس رغبنا الحافلة وأهنا إلى المزدلفة فسلمنا المغرب والشاء جميعا بعد ترتيبنا، أينما بدركنا وبعد الظهر يوم العاشر جعلنا إلى المشى، بعد وصولنا أخذنا أفعال اليوم العاشر واسترحنا، والإيام الثلاثة القادمة رغبنا الجمار واستمعنا إلى تصاحف ومواظب مغنوا فزادنا أينما بعدد الله تعالى ففككت هذه الرحلة مباركة لأنها رحلة لطاعة والتقوى وإزدياد الخير، ثم
    </text_body>
  </text>
</doc>
```

Figure A.4: Example of XML file with Arabic metadata

```
<?xml version="1.0"?>
<!--Arabic Learner Corpus_v2_2014-->
<!DOCTYPE doc>
- <doc ID="S002_T1_M_Pre_NNAS_W_C">
  - <header>
    - <learner_profile>
      <age>25</age>
      <gender>Male</gender>
      <nationality>Russian</nationality>
      <mothertongue>Russian</mothertongue>
      <nativeness>NNAS</nativeness>
      <No_languages_spoken>5</No_languages_spoken>
      <No_years_learning_Arabic>5</No_years_learning_Arabic>
      <No_years_Arabic_countries>5</No_years_Arabic_countries>
      <general_level>Pre-university</general_level>
      <level_study>Diploma course</level_study>
      <year_or_semester>Second semester</year_or_semester>
      <educational_institution>Arabic Inst. at Imam Uni</educational_institution>
    </learner_profile>
    - <text_profile>
      <genre>Narrative</genre>
      <vwhere>In class</vwhere>
      <year>2012</year>
      <country>Saudi Arabia</country>
      <city>Riyadh</city>
      <timed>Yes</timed>
      <ref_used>No</ref_used>
      <grammar_ref_used>No</grammar_ref_used>
      <mono_dic_used>No</mono_dic_used>
      <bi_dic_used>No</bi_dic_used>
      <other_ref_used>No</other_ref_used>
      <mode>Written</mode>
      <medium>Written by hand</medium>
      <length>294</length>
    </text_profile>
  </header>
  - <text>
    <title>رحلة الحج المباركة</title>
    <text_body>
      كتب
      الله لي أن أضع إلى بيته الحرم النبوي، فله الحمد والمنة على هذه التلمذة الحقيقية؛ لأن كثيرا من التلمذات محرومون عن هذه التلمذة إما بسبب المرض وإما لعدم القدرة المالية، ففككت يداه ريشي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعمائة وألف، صباح يوم السابع لعين طلاب جامعة الإمام في المسجد العربي من الجامعة وقت واحد من بينهم، وكان بعض الطلاب مع عائلاتهم، فلي المسجد ثلاثا نظفوا وأشكوا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة، بعد ذلك رغبنا الحافلة وبدأنا في السير وكان وقت الضحى، في الحافلة كنا دعاء السفر وإصمحا مسؤولو الرحلة تصاحف مغنوا، وبصحبنا في السير حتى قرب وقت صلاة العصر فزلتوا ومصلينا نظفوا والعصر جمع تأخير وتعدنا واسترحنا ثم رغبنا الحافلة مرة أخرى واستمرنا في السير حتى وصلت بيوت أهل نجد شيل التغيير هناك صلبنا المغرب والشاء وأخذنا سنن الإحرام وبخنا في الإحرام ثم جعلنا إلى مكة المكرمة، وصلنا مكة شحى يوم الثامن، بعد وصايي ذهبت إلى فندق الحجاج الأيمن من روسيا حتى استرحت عندهم، وأول الأرزاق التي جئت مع أخي إلى المسجد الحرام لشاء صلاة الجمعة، بعد الصلاة رجعا إلى الفندق واسترحنا حتى جاعا الليل، فلي الليل جعلنا إلى عرفة، ووصلنا إلى عرفة قبل صلاة الظهر فأخذنا راحنا حتى المغرب، فسلمنا الحجر ثم كنا إلى المشي بعد الاستراحة تناولنا الطعام وارتقنا صلاة الظهر، بعد صلاة الظهر دعنا ولنا وقت فشيئة واستجابة للدعاء ما بين النبي صلى الله عليه وسلم فشدنا يوم عرفة، بعد غروب الشمس رغبنا الحافلة وأهنا إلى المزدلفة فسلمنا المغرب والشاء جميعا بعد ترتيبنا، أينما بدركنا وبعد الظهر يوم العاشر جعلنا إلى المشى، بعد وصولنا أخذنا أفعال اليوم العاشر واسترحنا، والإيام الثلاثة القادمة رغبنا الجمار واستمعنا إلى تصاحف ومواظب مغنوا فزادنا أينما بعدد الله تعالى ففككت هذه الرحلة مباركة لأنها رحلة لطاعة والتقوى وإزدياد الخير، ثم
    </text_body>
  </text>
</doc>
```

Figure A.5: Example of XML file with English metadata

A.3 PDF files

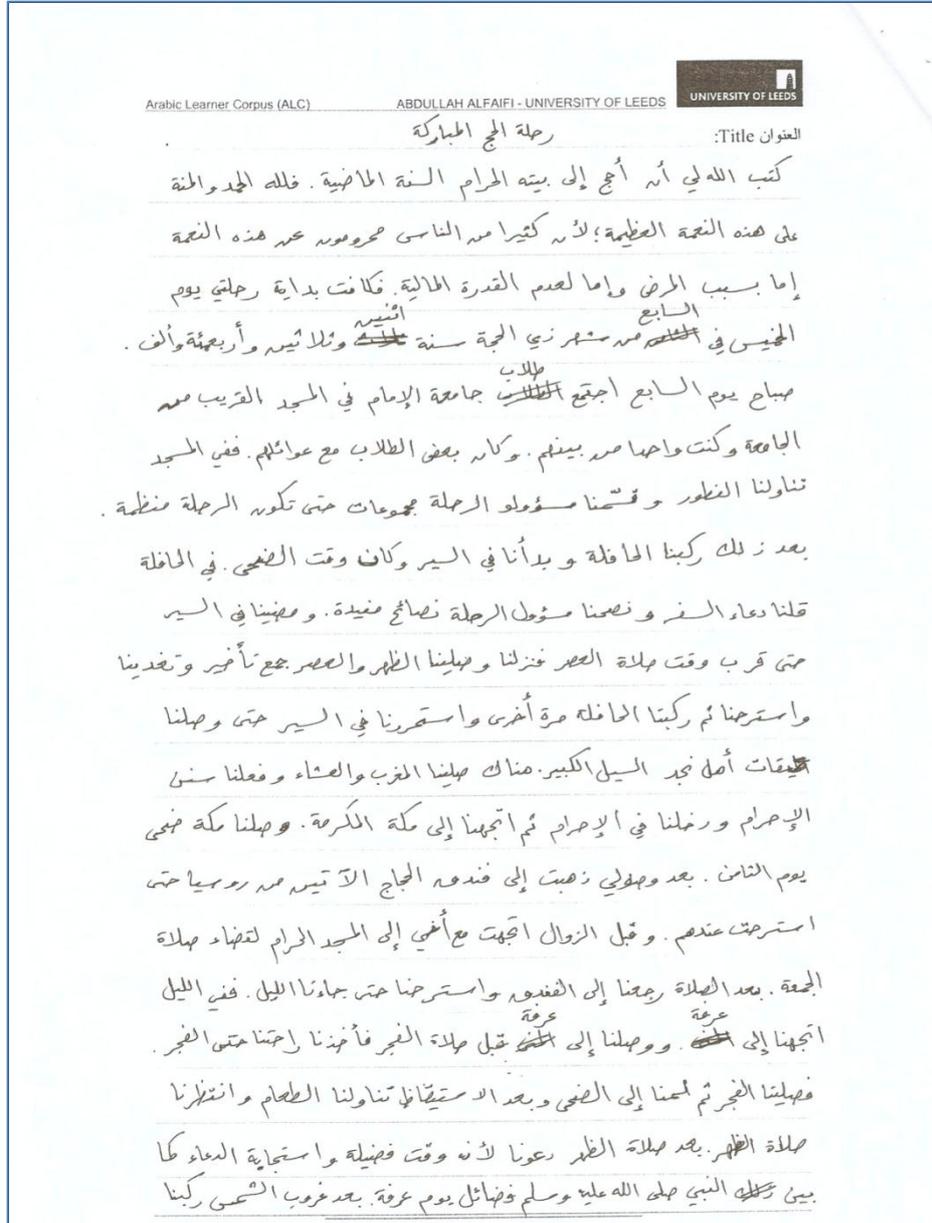


Figure A.6: Example of handwritten text in PDF file format

Appendix B

The Guide for Data Collection

1. Introduction

This guide is to clarify the steps the researcher (or his representative) will follow for collecting data for the Arabic Learner Corpus. The corpus will consist of written and spoken materials, produced by native and non-native speaking-Arabic learners, males and females, from Pre-university and University levels.

2. Collecting the Data

Following the outlines of the corpus, there will be one main session for data collection which is repeated with each group of students at every educational institution.

2.1. Session

During the sole session, which expected to last for about 2 hours, the researcher (or his representative) will introduce the research purposes, benefits, and methods of participation with clarifying that:

1. the participation is fully voluntary,
2. a participant is free to withdraw at any time, and
3. a participant's materials will be used in the corpus for research purposes.

The learners will be allowed to ask any question about the research, its purposes, or their participation.

2.2. Tasks

Two tasks will be distributed to the participants with clear explanation in advance about the tasks and how to complete them.

2.2.1. First task: two timed- compositions in class (40 minutes for each)

The first task is timed and carried out with no prior preparation. Learners in this task will be asked to write two narrative and discussion essays – in Arabic – about the topics presented in 40 minutes for each with no use of language references such as dictionaries or grammar books.

2.2.2. Second task: two take-home compositions

For the second task, the participants will be required to write the same narrative and discussion essays at home, but with an ability to use reference tools such as dictionaries or grammar books, as this task is untimed and prior preparation is allowed. They will be required to bring the essays in the next day or the day after.

2.3. Topics of Writing

The writing tasks include two topics lie under two different genres, *a vacation trip* (narration) and *my study interest* (discussion).

3. Summary of the data collecting procedures

Procedure	Description	Time (estimated)
Introduction	<ul style="list-style-type: none"> - To introduce the research purposes, benefits, methods of participation and answering questions that learners may ask. - To Distribute the participant consent form to be signed by the learners. 	30 minutes
Task 1	To write narrative and discussion compositions, in class, about topics provided (<i>A Vacation Trip</i> for the narration genre and <i>My Study Interest</i> for the discussion), with no prior preparation.	No more than 40 minutes for each composition
Task 2	To Explain the second task which to write narrative and discussion compositions under the same topics, at home, with prior preparation.	10 minutes

Appendix C

The Paper Copy of ALC Questionnaire

Arabic Learner Corpus (ALC) ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

Overview about the ALC project

Title of project: Building a corpus of learner Arabic with Part-of-Speech Tagging and Error Annotation

Brief outline of project

The project aims to compile a corpus of Arabic learner, a representative collection of texts written (and speech transcribed) by learners of Arabic. The corpus will cover both learners of Arabic as a second or foreign language, and native Arabic speaking students learning to improve their written Arabic. The corpus will be annotated with linguistic features, including Part-of-Speech tags and mark-up of errors, to enable diagnostic patterns to be identified. This stage of project is devoted for a first collection of written texts as an initial version.

The benefit of this study

The corpus will be used for research purposes on Arabic language learning and teaching. It will also help designers of language materials to develop better learning materials, dictionaries, language applications, and textbooks for Arabic learning and teaching.

Procedures

Participants in this first version of the project will be asked to involve in two writing tasks as following:

1. To write narrative and discussion compositions, in class, about two topics provided, with no prior preparation
2. To write the same narrative and discussion compositions, at home, with prior preparation

Participation

Participation in the project is voluntary, and you are free to withdraw at any time. Data will be anonymous and your identity will not be revealed when we publish the findings of our research.

Figure C.1: An overview about the ALC project in the data collection questionnaire

Arabic Learner Corpus (ALC) _____ ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

FORM OF CONSENT TO TAKE PART IN A RESEARCH PROJECT

Please tick the three statements below:

I confirm that I have read and understand the information explaining the research project and I have had the opportunity to ask questions about the project.

I agree for the data collected from me to be published and used in relevant future research.

I agree to take part voluntarily in the research project.

Name of participant	
Participant's signature	
Date	
Name of lead researcher or person taking consent	
Signature	
Date	

Figure C.2: The consent form to take part in the ALC project

Arabic Learner Corpus (ALC) _____ ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

Arabic Learner Corpus

Learner Profile

Personal details

1. Name: _____

2. Contact number (optional): _____

3. E-MAIL (optional): _____

4. Age: _____

5. Gender Male Female

6. Nationality: _____

7. Native language: _____

Educational details

1. Current level of study _____

2. Current year/semester of study _____

3. Name of educational institution _____

4. The total number of years of learning Arabic ()

5. The period you stayed in Arabic-speaking countries ()

Figure C.3: The learner's profile questionnaire used in ALC

Arabic Learner Corpus (ALC) _____ ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

Arabic Learner Corpus

Text Profile

1. Text code (by the researcher):

2. Text topic (by the researcher):

3. Text title:

4. Writing date / /

5. Is the text timed? Yes No

6. Have you used any reference tool from the following:

Grammar books

Monolingual dictionary

Bilingual dictionary

Other references (please specify) _____

7. Medium of writing : Computer By hand

8. Text length (words number):

Figure C.4: The text's data questionnaire used in ALC

Arabic Learner Corpus (ALC) _____ ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

The tasks of collecting the Arabic Learner Corpus data

Task 1 – first text

Task: write a narrative essay about a vacation trip providing as many details as you can about this trip.

Time: 40 minutes

Place: in class

Language references: during this task you are NOT allowed to use any reference tools such as dictionaries or grammar books

Medium of writing: writing these texts is by hand on the sheets provided by the researcher, two pages are provided for each text and you can ask for more if needed

Task 1 – second text

Task: write a discussion essay about your study interest providing as many details as you can and also your future plans to continue your study and to work in this field.

Time: 40 minutes

Place: in class

Language references: during this task you are NOT allowed to use any reference tools such as dictionaries or grammar books

Medium of writing: writing these texts is by hand on the sheets provided by the researcher, two pages are provided for each text and you can ask for more if needed

1

Figure C.5: Task 1 in the ALC questionnaire

Arabic Learner Corpus (ALC) _____ ABDULLAH ALFAIFI - UNIVERSITY OF LEEDS 

The tasks of collecting the Arabic Learner Corpus data

Task 2 – first text

Task: write a narrative essay about a vacation trip providing as many details as you can about this trip.

Time: one to two days

Place: at home

Language references: during this task you are allowed to use any reference tools such as dictionaries or grammar books

Medium of writing: writing this text is by hand on the sheets provided by the researcher, two pages are provided for each text and you can use more if needed

Task 2 – second text

Task: write a discussion essay about your study interest providing as many details as you can and also your future plans to continue your study and to work in this field.

Time: one to two days

Place: at home

Language references: during this task you are allowed to use any reference tools such as dictionaries or grammar books

Medium of writing: writing this text is by hand on the sheets provided by the researcher, two pages are provided for each text and you can use more if needed

Figure C.6: Task 2 in the ALC questionnaire

Appendix D

The Questionnaires That Used to Evaluate the ETAr

D.1 First evaluation questionnaire

Dear Annotator,

Thank you for participating in the annotation task, please answer the following questions:

=====

Name:

Degree:

Qualification:

Major:

=====

- In general, which tagset was easier and faster when annotating:

() First

() Second

() About the same

Why?

Appendix D – The Questionnaires That Used to Evaluate the ETAr

- Which of the tagsets was more understandable?

() First

() Second

() About the same

Why?

- Which error types need to be added?

- Which error types need to be deleted?

- Which error types need to be changed?

- Which error types need to be integrated?

- Which error types need to be split?

Thank you for your cooperation..

D.2 Second Evaluation Questionnaire



LEEDS UNIVERSITY – SCHOOL OF COMPUTING

Evaluating the Error Tagset of Arabic

[A practical annotating task and questionnaire]

Abdullah Alfaifi

2013

I agree to participate in the task of evaluating the error tagset of Arabic

Name:

Major:

Position:

Degree:

Signature:

Date:

Dear annotator,

Thank you for participating in the task of evaluating the error tagset of Arabic. Your evaluation will significantly contribute in improving this tagset and its method of use. This file of evaluation consists of six sections as following:

1. About the annotation task
2. The tag-set of errors in Arabic
3. Examples of error annotation
4. Errors required to be annotated
5. Questionnaire of evaluation and comments
6. About this questionnaire

Please read the annotating method carefully, and then do the task as accurate as possible, as explained in the instructions.

Thank you for your kind cooperation,

Abdullah Alfaifi

About the annotation task

You are required to assign the suitable error tag – from the error tagset – to each error of those listed below. Also you have to select how much easy it was to choose the appropriate tag for each error.

You can use either the Arabic or English tag. If you think the word/sentence never include any error, please put a tick sign (✓) instead of an error tag.

Please have a look at the annotated example to be aware of how to annotate the errors.

Prior to the annotation process, you are advised to read the “Error Tagging Manual for Arabic”, as this guideline shows the method of how to use the error tagset. It aims to lead annotators to the best way to selecting tags that properly match error in Arabic texts.

Appendix D – The Questionnaires That Used to Evaluate the ETAr

The tag-set of errors in Arabic

Error Category	Error Type	Arabic tag	English tag
1. Orthography الإملاء ' <i>imlā'</i>	1.1. Hamza (هـ، ء، أ، إ، ؤ، ئ، نْ)	<إه>	<OH>
	1.2. Confusion in <i>Hā'</i> and <i>Tā'</i> <i>Mutaṭarrifatain</i> (هـ، تـ)	<إه>	<OT>
	1.3. Confusion in <i>'alīf Mutaṭarrifa</i> (أ، ي)	<إي>	<OA>
	1.4. Confusion in <i>'alīf Fāriqa</i> (كتبا)	<إت>	<OW>
	1.5. <i>Lām Šamsīya</i> dropped (الطالِب)	<إا>	
	1.6. Confusion between <i>Nūn</i> (ن) and <i>Tanwīn</i> (وُؤُ)	<إل>	<ON>
	1.7. Shortening the long vowels (اوي → وُؤُ)	<إف>	<OS>
	1.8. Lengthening the short vowels (اوي → وُؤُ)	<إق>	<OG>
	1.9. Wrong order of word characters (الخطأ في ترتيب الحروف داخل الكلمة)	<إط>	<OC>
	1.10. Replacement in word character(s) (استبدال حرف أو أحرف من الكلمة)	<إس>	<OR>
	1.11. Character(s) redundant (حرف أو أحرف زائدة)	<إز>	<OD>
	1.12. Character(s) missing (حرف أو أحرف ناقصة)	<إن>	<OM>
	1.13. Other orthographical errors (أخطاء إملائية أخرى)	<إخ>	<OO>
2. Morphology الصرف ' <i>ssarf</i>	2.1. Word inflection (صيغة الكلمة)	<حصص>	<MI>
	2.2. Verb tense (زمن الفعل)	<حصز>	<MT>
	2.3. Other morphological errors (أخطاء صرفية أخرى)	<صصخ>	<MO>
3. Syntax النحو ' <i>nmaḥw</i>	3.1. Agreement in grammatical case (المطابقة في الإعراب)	<ننب>	<XC>
	3.2. Agreement in definiteness (المطابقة في التعريف والتذكير)	<ننع>	<XF>
	3.3. Agreement in gender (المطابقة في الجنس (التذكير والتأنيث))	<ننذ>	<XG>
	3.4. Agreement in number (singular, dual and plural) (المطابقة في العدد (الإفراد والتثنية والجمع))	<ننف>	<XN>
	3.5. Words order (ترتيب المفردات داخل الجملة)	<ننت>	<XR>
	3.6. Word(s) redundant (كلمة أو كلمات زائدة)	<ننز>	<XT>
	3.7. Word(s) missing (كلمة أو كلمات ناقصة)	<ننن>	<XM>
	3.8. Other syntactic errors (أخطاء نحوية أخرى)	<ننخ>	<XO>
4. Semantics الدلالة ' <i>ddalāla</i>	4.1. Word/phrase selection (اختيار الكلمة أو العبارة المناسبة)	<دب>	<SW>
	4.2. <i>Faṣl wa waṣl</i> (confusion in use/non-use conjunctions) (الفصل والوصل (الخلط في استخدام أو عدم استخدام أدوات العطف))	<دف>	<SF>
	4.3. Wrong context of citation from Quran or Hadith (الاستشهاد بالكتاب والسنة في سياق خاطئ)	<دس>	<SC>
	4.4. Other semantic errors (أخطاء دلالية أخرى)	<دخ>	<SO>
5. Style الأسلوب ' <i>l'ustūb</i>	5.1. Unclear or weak style (أسلوب غامض أو ركيك)	<سع>	<TU>
	5.2. Other stylistic errors (أخطاء أسلوبية أخرى)	<سسخ>	<TO>
6. Punctuation علامات الترقيم ' <i>alāmāt 't-tarqīm</i>	6.1. Punctuation confusion (الخلط في علامات الترقيم)	<تط>	<PC>
	6.2. Punctuation redundant (علامة ترقيم زائدة)	<تنز>	<PT>
	6.3. Punctuation missing (علامة ترقيم مفقودة)	<تنن>	<PM>
	6.4. Other errors in punctuation (أخطاء أخرى في علامات الترقيم)	<تنخ>	<PO>

Examples of error annotation

These are examples of all error types in the tagset (excluding last type in each category)

No	Example	Suggested correction	Tag
1	لم اعلم أن هذا الأمر ضروري	لم أعلم أن هذا الأمر ضروري	OH
2	كانت تلك المعلومات غريبة	كانت تلك المعلومات غريبة	OT
3	جميع المدن و القرى كانت تعج بالسكان	جميع المدن و القرى كانت تعج بالسكان	OA
4	وكان الحجاج قد رجعو من رحلة الحج	وكان الحجاج قد رجعوا من رحلة الحج	OW
5	فنظرنا جميعاً من انافذة	فنظرنا جميعاً من النافذة	OL
6	كانت الحرارة مرتفعة جداً	كانت الحرارة مرتفعة جداً	ON
7	صعوبة الطريق منعتني من مرافقتكم	صعوبة الطريق منعتني من مرافقتكم	OS
8	وصل الضيف فرحبت به في بيتي	وصل الضيف فرحبت به في بيتي	OG
9	من واجبات المسلم أن يتفقه في دينه	من واجبات المسلم أن يتفقه في دينه	OC
10	جهزنا أمتعتنا و انطلقنا مباشرة إلى وجهتنا	جهزنا أمتعتنا و انطلقنا مباشرة إلى وجهتنا	OR
11	عندما كنا راجعين إلى بيوتنا	عندما كنا راجعين إلى بيوتنا	OD
12	وطلب مني أن أستلقي على ظهري	وطلب مني أن أستلقي على ظهري	OM
14	ووقف التلميذون يتفرون على المشهد	ووقف التلاميذ يتفرون على المشهد	MI
15	فلما لقيته أسلمت عليه بحرارة وسألته عن حاله	فلما لقيته أسلمت عليه بحرارة وسألته عن حاله	MT
17	كان للمسلمين تراث علمي عظيم	كان للمسلمين تراث علمي عظيم	XC
18	وعند النهر الكبير قابلنا أصدقاءنا	وعند النهر الكبير قابلنا أصدقاءنا	XF
19	فجريت جرياً سريعاً لأصل إليه	فجريت جرياً سريعاً لأصل إليه	XG
20	بقي الطلاب يستمعون إلى شرح المعلم حتى النهاية	بقي الطلاب يستمعون إلى شرح المعلم حتى النهاية	XN
21	لقيت أخي عند باب الفصل ودخلنا سوياً	لقيت أخي عند باب الفصل ودخلنا سوياً	XR
22	توقفنا عند أحد المطاعم	توقفنا عند أحد المطاعم	XT
23	انتهى الدرس وخرجنا من الفصل مسرعين	انتهى الدرس وخرجنا من الفصل مسرعين	XM
25	وعندها وصلنا إلى مرتفع الجبل	وعندها وصلنا إلى قمة الجبل	SW
26	عندما لقيت زميلي الجديد فحييته ورحبت به	عندما لقيت زميلي الجديد فحييته ورحبت به	SF
27	وقد قال تعالى في الزكاة : "قد أفلح من زكاها "	الآية لا تدل على الموضوع المذكور	SC
29	فانطلقنا متوجهين إلى مكة المكرمة، وعندما وصلنا مكة المكرمة أدينا العمرة	فانطلقنا متوجهين إلى مكة المكرمة، وعندما وصلنا إليها أدينا العمرة	TU
31	من مختلف الأجناس، والديانات، والثقافات.	من مختلف الأجناس، والديانات، والثقافات.	PC
32	ولم أكن أتوقع أن أجد كل هذا العدد من الناس	ولم أكن أتوقع أن أجد كل هذا العدد من الناس	PT
33	ثم قال لي: "أنت طالب جيد"	ثم قال لي: "أنت طالب جيد"	PM

Errors required to be annotated

First List

Please tag the following list of errors then discuss it with the researcher

No.	Example	Tag	Did you find the suitable tag easily?			
			Very easily found	Somewhat easily found	Found with difficulty	Not found
1	وضع أخي الصورة في جهازه المحمول					
2	وفتح النمر فمهو بشكل مخيف					
3	لم أحب الذهاب إلى في هناك					
4	قضينا فيها عدة أيام رائع					
5	يعلمون أبنائهم في مدارس					
6	أوجب الله الصوم على المؤمنين، قال تعالى: " قل هذه سبيلي أدعو إلى الله "					
7	وصلنا مكة وذهبنا مباشرة إلى مسجد الحرام					
8	وضعت فيه قليلان من الماء					
9	أقبلو جميعاً للسلام عليه					
10	علينا أن نذاكر درس اليوم درس الأمس					
11	تكبيرت الإحرام					
12	الكبير، والصغير، وكذلك، أنواع أخرى.					
13	وبعد أن تحدثنا قليلاً سألوني عن أهل المدينة ومن تركتهم في المدينة ، ثم سألوني...					
14	ألم يأت إليك أحد					
15	ووضع يدهو على موضع الجرح					
16	ثم قررنا أن نودي إجازتنا في دولة مجاورة					
17	وهل يعني؛ عنك هذا؟					
18	في أغلب الاماكن من حولنا					
19	فسألته: و يدريك أنه هو؟					
20	وأخذنا الاماكننا المخصصة لنا في الحافلة					
21	وجريت حنا نهاية الملعب					
22	وقد كانت هذه الرحلة جميلات ورائعة بالفعل					
23	وقد مكث هناك؛ لفترات طويلة غير معلومة.					
24	فوقف اطلاب يشاهدون					
25	وقد كان رحمه الله- يحب مجالسة العلماء					
26	في تمام ساعة الثامنة مساءً					
27	عندما نظرت من انافذة					
28	وقد قام منظمو الحفل					
29	وإن من نعم الله لنا أن جعلنا مسلمين					
30	وعندما التقينا فسلمنا على بعضنا					
31	وفي الغد سافرنا صباحاً، فتجهزوا					
32	رأيت اشمس مشرقة					
33	كانت مفاجآت رائعة					
34	الذي اعطاني درساً					
35	لذا أحب دائماً أزور أن ذلك المكان					
36	فلقبته وأسلم عليه ثم تحدثنا قليلاً					

Appendix D – The Questionnaires That Used to Evaluate the ETAr

37	لأداء العمرة في مكة				
38	وكان الخبر يتحدث عن الرئيس أوباما				
39	فسألته: ألك عندي حاجة؟				
40	ثم ذهبنا الزيارة إلى المكان يصنع القرآن الكريم				
41	على الآباء إحسان تربية أبنائهم، قال تعالى: " فهل عسيتم أن توليتم أن تفسدوا في الأرض وتقطعوا أرحامكم "				
42	نسأل الله التوفيق، والحمد لله رب العالمين				
43	ثم هبطت اطائرة بسلام				
44	على شاشات الحاسب				
45	أما الطالبات المجتهدة فقد نجحن بتفوق				
46	كان عدد المشاركون أكثر من خمسين				
47	و عندمى وصلنا إلى مقر الرحلة				
48	في الصباح كان جو جميلاً والسماء صافية				
49	اجبرتني الظروف على السفر قبل زملائي				
50	أما أخي وزوجته فقد ذهبتا في اليوم التالي				
51	أما الطلاب فقد ذهبوا لمقابلة معلمهم				
52	قد أمسك السمكة بيديهي جميعاً				
53	في كل أنحاء المعمورة				
54	فسلمت على زملائي الطلاب				
55	فلما وصلت الرسالة و فتحتها بسرعة				
56	لما له من أثر كبير				
57	سأخصص في الطب لأني أهواه منذ الصغر				
58	استيقظت في الصباح الباكر				
59	أنهيت قراءة الكتاب وأضعه جانباً ومضيت				
60	هلا أسديتني خدمة منن فضلك				
61	مع أخوتي وأخوتي جمعياً				
62	ثم رأيت سيارتين قادمة				
63	تتسعد للذهاب من الصباح الباكر				
64	فأخذت معي أكتبة كثيرة لأقرأ فيها				
65	شكراً لك؛ فلقد منحتني ثقة كبيرة				
66	المغرب والعساء				
67	على المسلم أن يرعى حق جيرانه، قال صلى الله عليه وسلم: " من حسن إسلام المرء تركه ما لا يعنيه "				
68	خذ مني جواب سؤالك مفصلاً				
69	وعبرنا تلك المتنوعات المانية				
70	ولما وصلنا إلى قمة ذلك الجل المرتفع				
71	يا للعجب؟				
72	صلة الرحم واجبة بين المسلمين، قال صلى الله عليه وسلم: " تبسمك في وجه أخيك صدقة "				
73	ورميناه بالجحارة حتى ذهب				
74	وبقينا معهم عدة أيام لنستفيد من علمهم				
75	وبعد أن انتهينا أداء العمرة				
76	وقد استغرقت الرحلة خسة أيام متواصلة				
77	ورسمت عليها عدداً من الدوائر الملونة				
78	يا لجمال المكان ؟				
79	كلما سمعت كلمة أكررها حتى رسخت				
80	فلبيت معلموا المعهد وقتاً أطول				
81	هذا هو الطالب أعطني الكتاب بالأمس				

Appendix D – The Questionnaires That Used to Evaluate the ETAr

82	ومنها تلك الساجد التي تبنى على الطرقات				
83	فوافقت عليه عليه مرغماً				
84	فسلمت على النبي وأبو بكر وعمر				
85	هلا أعطيتن كأساً من الماء				
86	وعند الساعة الخامسات والنصف انطلقنا				
87	ابن القيم (ت ٧٥١هـ) الذي يقول				
88	وكانت رائحة العطر يفوح من أرجاء المكان				
89	هذا أخيك قادم				
90	ومن هذه الأخطاء ترك السنن الرواتب				
91	وضعت الأوراق درج في الطاولة				
92	هؤلاء المسافرون الذين وصل من رحلة الحج				
93	فمنهم القائم. ومنهم القاعد، ومنهم من غادرنا				
94	لحقت بزملائي الطلاب، وعندما وصلت إلى زملائي الطلاب فرحت جداً				
95	ألم نتفق على ـ نجعل لنا لقاءً مستمراً				
96	أخي الصغير يدرس الرابع الصف في				
97	وهل يشك أحد على ذلك؟				
98	وكانت الساعة تشير على إلى السادسة والنصف				
99	ومن هذا الذي معك.				
100	فألقي خطبته مقيماً أمام الناس على المنبر				

Second List

Please tag the following list of errors after discussing the first list with the researcher

No.	Example	Tag	Did you find the suitable tag easily?			
			Very easily found	Somewhat easily found	Found with difficulty	Not found
1	في يوم الواحد والاثنتين والثلاثاء					
2	في الغرفة الواسعة					
3	ابن رجل كبير					
4	عندما أعطيتة كتابه					
5	ومررنا في بقرب أحد المنتزهات					
6	في تمام الساعة الثالث عصراً					
7	فطلبت منه عدة أمور					
8	ولا تتسى أن تأخذني معك					
9	ثم تلى ما تيسر له من القرآن					
10	من نواقض الوضوء أكل لحم الإبل، قال تعالى: "أفلا ينظرون إلى الإبل كيف خلقت"					
11	فتعجب لذلك أشد العجب					
12	ومد يده علينا مصافحاً					
13	وفي أحد اليوم ذهبنا لزيارة مصنع الكسوة					
14	وبقينا في المنى أيام التشريق					
15	وعندما حانت السابعة الساعة غادرنا المكان					
16	وكانت هذه نهاية القصة ؛					

Appendix D – The Questionnaires That Used to Evaluate the ETAr

17	وصلينا المغرب والعشاء جمعاً <u>قصرأ</u>				
18	ثم استلم الطلاب <u>المتفوق</u> شهاداتهم				
19	فلما <u>وصلو</u>				
20	ولما وصلنا إلى <u>الجبل قمة</u> رأينا مناظر خلابة				
21	فسمعت منه ما نصه "اغتنم هذه الفرصة"				
22	فارتطمت <u>اسيارة</u> بالجدار				
23	إذ أقبل علينا <u>رجلن</u> طاعن في السن				
24	في البيت <u>الكبير</u>				
25	فقابلت زملائي <u>ووقفن</u> نتحدث قليلاً				
26	أحب قراءة الكتب <u>وأحب</u> كتابة المقالات.				
27	يجب صون اللسان عن الكذب، قال تعالى: " <u>ولساناً</u> <u>وشفتين</u> "				
28	فقلت: له حين سألني: لا علم لي بهذا				
29	<u>فيفيقت</u> علي هذا أكثر من يوم				
30	وترددت <u>أن أذهب</u> معه				
31	العقيدة <u>الصحيحة</u>				
32	وهو واضح كما <u>ترا</u>				
33	كنجوم السماء <u>ادنيا</u>				
34	كوضع الدواء على <u>جرحن</u> نازف				
35	و ماء النهر <u>يجر</u> من أمامنا				
36	ولما عرفنا مكان الغابة الجميلة، ذهبنا إلى <u>الغابة</u> <u>الجميلة</u> واستمتعنا بمناظرها				
37	لم يأت أحد إلى تلك <u>الوليمة</u> الكبيرة				
38	فقال لنا <u>أذهبوا</u> من هذا الاتجاه.				
39	وعندما، وصلت إليهم				
40	ولطالب العلم <u>خاصه</u>				
41	وصعد فوق <u>سح</u> المبنى				
42	لكني <u>الزددت</u> ثقة عندما سألته				
43	الأخت <u>الصغرا</u>				
44	فقمنا وأشعلنا <u>انار</u> للتدفئة				
45	بعد ذلك اتجهنا إلى مشعر عرفة، ووصلنا إلى <u>مشعر عرفة</u> قبل صلاة الفجر				
46	وكان عدد <u>التلميذين</u> كبيراً جداً				
47	ثم <u>رحعنا</u> إلى الفندق				
48	كان المنظر <u>في في</u> قمة الجبل رائعاً				
49	ثم <u>هوا</u> إلى أسفل الوادي				
50	ألم تذاكر <u>دوسك</u> جيداً				
51	توقفنا <u>إلى</u> أحد المطاعم				
52	فلما توقفنا <u>يذهب</u> أحدنا لشراء بقية الأغراض				
53	<u>وبدئنا</u> نجهز بضائعنا				
54	صلاة <u>العسر</u>				
55	فجاء <u>المسلمين</u> من كل القرى القريبة				
56	فقابلت <u>رجلين صالحان</u>				
57	غرف الفندق نظيفة و <u>واسع</u>				
58	صلينا الظهر والعصر <u>مجموع</u> وقصرأ				
59	<u>وفرحو</u> لما سمعوا مني				
60	فسلمت <u>عليهي</u> سلاماً حاراً				
61	فرفض الطلاب أن <u>يذهبون</u> من مكانهم				
62	وبعد صلاة <u>عصر</u> أكملنا طريقنا				
63	كم هي <u>مرتفع</u> هذه الجبال				

Appendix D – The Questionnaires That Used to Evaluate the ETAr

64	فقال لنا؛ تعلموا العلم				
65	رأيت الكرم مادح والبخل مذموم				
66	وعندما وصلنا إلى الفندق				
67	كان مراقبوا الامتحان				
68	هلا أعرتني ورقة لاكت عليها الدرس				
69	الفرصة التي أكون فيها				
70	خرجت مع مع أخي نتمشى				
71	تلك الأبراج التي يوجد في وسط المدينة				
72	ذهبنا إلى الجبل لنرفع على أعلى مكانه				
73	فوصلنا ونحن متعب من طول الطريق				
74	خرج المعلم الفصل مسرعاً				
75	اليوم نستاق معاً				
76	وإبعادهم من البدع و خرافات شركية				
77	وشربت الماء الذي الكأس				
78	كثير من الناس محرومون عن هذه النعمة				
79	فلما اقتربت منه لم رأيت شيئاً				
80	البلاد الاسلامية				
81	الحمد لله الصلاة والسلام على رسول الله				
82	لم أستعد لمثل الحدث الهام				
83	وما رأيت مثل هذا الهكف				
84	وكان قدماً من جهة زمزم بئر				
85	يجب على الداعية أن يبين الحلال والحرام للناس، قال صلى الله عليه وسلم: " الحلال بين والحرام بين "				
86	وصلى معنا كثيرين من الرجال				
87	باب من أبواب الجنة				
88	ذهب إليه أخي تبعته مباشرة				
89	لبثنا هناك طويلاً وقتاً إلى أن مللنا				
90	كانت الرحلة في أحد الأيام الصيف				
91	ثم سلمت على الطلاب هم يدرسون معي				
92	وأرجو من الله تيسير في تعليم الشرع لنا				
93	كما كان يفعل نبيينا. وكما كان يفعل أصحابه.				
94	كما هم دون فرقن يذكر				
95	فأبتسم عندما أكن سعيداً				
96	توضأنا صلينا ثم أكملنا طريقنا				
97	أهذا؟ الكتاب لك؟				
98	فكنت أقرأ الدرس وهو ردد ورائي				
99	بعد الانتها من المعهد				
100	هل أنت متأكد فأجابني : نعم.				

Questionnaire of Evaluation and Comments

Are the error labels clear and easily understood?

- Appropriate and do not need more clarification
- Need some clarification
- Ambiguous and they need to be fully clarified

Is the division of error categories clear and understandable (6 categories)?

- Yes
- To some extent
- No

Is the division of error types clear and understandable (34 types)?

- Yes
- To some extent
- No

How easy and fast is selecting the suitable tag?

- It can be selected easily and quickly
- It requires some time to be selected
- It requires a long time to be selected

How suitable is the tagset in general for errors in Arabic?

- It is OK
- It requires some modifications
- It is completely unsuitable

Suggestions about the error types

Error types should be added

Error types should be deleted

Error types should be integrated in one type

Error types should be spitted into different types

Labels of error types should be changed

Final suggestions about the error tagset types

Advantages of the error tagset

Disadvantages of the error tagset

How the tagset can be improved?

About this Questionnaire

Please provide your general opinion about this questionnaire

- OK
- Requires some modifications
- Unsuitable

What do you think about the methodology used to evaluate the error tagset in this questionnaire?

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

What do you think about the number used of error examples used (200 examples)?

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

What do you think about evaluating the ease of finding errors tags (after tagging each error)?

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

What do you think about the “Error Tagging Manual for Arabic” in its:

Design

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

Comprehensiveness of the information

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

Clarity of the explanations

- Excellent
- Good
- Acceptable
- Poor
- Unsuitable

Consent for the evaluation to be included on the corpus website

Dear Evaluator,

Did you give the permission to put some or all of this evaluation on the corpus website?

Please tick (✓) one of the following options.

- I give consent for some or all of this evaluation to be included on the ALC website - including my name
- I give consent for some or all of this evaluation to be included on the ALC website - NOT including my name
- I don't give consent for this evaluation to be included on the ALC website

Thank you for your kind cooperation.

Appendix E
The Error Tagging Manual for Arabic
(ETMAr)



UNIVERSITY OF LEEDS

دليل ترميز الأخطاء
للغة العربية
[الإصدار الثاني]

عبدالله الفيفي

ERROR TAGGING MANUAL
FOR ARABIC
[VERSION 2]

ABDULLAH ALFAIFI

2014

١٤٣٥هـ

1. مقدمة Introduction

بسم الله الرحمن الرحيم

يشرح هذا الدليل المختصر طريقة استخدام النسخة الثالثة من جدول ترميز الأخطاء العربية المصمم خصيصاً للمدونات اللغوية العربية. حيث يهدف إلى إرشاد المُرمِّز إلى الطريقة الصحيحة لاختيار الرموز المناسبة للأخطاء بشكل دقيق يوافق أنواع الأخطاء في النصوص العربية.

يركز هذا الدليل بشكل أكبر على الجوانب اللغوية للترميز، فيشرح كل نوع من أنواع الخطأ مع أمثلة مناسبة لمزيد من التوضيح. كما يزود المُرمِّز ببعض النقاط والقواعد الهامة التي يجب اتباعها عند عملية الترميز، مع شرحه لعدد من حالات التداخل المحتملة وطريقة التعامل معها. أما بالنسبة للجوانب التطبيقية مثل مكان وضع الرمز، وطريقة ظهور الرمز في النص، والشكل النهائي للنص بعد الترميز، فهي متروكة للمستخدم نفسه يضعها حسب تصميم مدونته اللغوية.

يفترض الدليل وجود معرفة جيدة لدى المُرمِّز بأساسيات اللغويات العربية، كقواعد الإملاء، والصرف، والنحو؛ أو على الأقل بالمصادر التي يمكنه الحصول على هذه المعلومات منها، ولذلك لا يتطرق لتعريف أو شرح مثل هذه الأساسيات اللغوية.

This guide shows how to use the third version of the Error Tagset of Arabic (ETAr), which has been particularly developed for Arabic corpora. It aims to show annotators the best ways to select tags that properly match errors in Arabic texts.

The main focus of this manual is on linguistic aspects, so it gives details about each error type, with appropriate examples for more clarification. The guide draws the annotator's attention to important points and rules that should be followed in the annotation process. Some possible instances of overlap and how to deal with them are also explained. In terms of applied issues, such as where the tag should be put, how the tag will appear, and the final format of the tagged text, all of these have been left to the user to be based on his corpus design.

It is assumed that the annotator has an adequate knowledge of Arabic language basics, such as orthographical, morphological, and grammatical rules, or at least the resources from which he can access such information, so this manual does not include detailed definitions and explanations about all of these basics.

2. Error Tagset جدول رموز الأخطاء

(6 categories, 30 error types - 29 نوعاً - 6 مجالات)

Error Category مجال الخطأ	Error Type نوع الخطأ	Tag الرمز
1. Orthography الإملاء 'l'imlā'	1. Hamza (ء، أ، إ، و، ي، ن) الخطأ في الهمزة	<OH>
	2. Confusion in Hā' and Tā' Mutaṭarrifatain (ت، هـ، ع، ت) الخطأ في الهاء والتاء المتطرفتين	<OT>
	3. Confusion in 'alif and Yā' Mutaṭarrifatain (ا، ي، ي) الخطأ في الألف والياء المتطرفتين	<OA>
	4. Confusion in 'alif Fāriqa (كتبوا) الخطأ في الألف الفارقة	<OW>
	5. Confusion between Nūn (ن) and Tanwīn (ّوّو) الخلط بين النون والتنوين	<ON>
	6. Shortening the long vowels (اوي → ّوّو) تقصير الصوائت الطويلة	<OS>
	7. Lengthening the short vowels (اوي → ّوّو) تطويل الصوائت القصيرة	<OG>
	8. Wrong order of word characters الخطأ في ترتيب الحروف داخل الكلمة	<OC>
	9. Replacement in word character(s) استبدال حرف أو أحرف من الكلمة	<OR>
	10. Redundant character(s) زيادة حرف أو أكثر	<OD>
	11. Missing character(s) نقص حرف أو أكثر	<OM>
	12. Other orthographical errors أخطاء إملائية أخرى	<OO>
2. Morphology الصرف 'ssarf'	13. Word inflection الخطأ في اختيار بنية الكلمة المناسبة	<MI>
	14. Verb tense الخطأ في زمن الفعل	<MT>
	15. Other morphological errors أخطاء صرفية أخرى	<MO>
3. Syntax النحو 'maḥw'	16. Case الخطأ في الإعراب	<XC>
	17. Definiteness الخطأ في التعريف والتنكير	<XF>
	18. Gender الخطأ في الجنس (التذكير والتأنيث)	<XG>
	19. Number (singular, dual and plural) الخطأ في العدد (الإفراد والتثنية والجمع)	<XN>
	20. Redundant word كلمة زائدة	<XT>
	21. Missing word كلمة ناقصة	<XM>
	22. Other syntactic errors أخطاء نحوية أخرى	<XO>
4. Semantics الدلالة 'ddalāla'	23. Word selection الخطأ في اختيار الكلمة المناسبة	<SW>
	24. Faṣl wa waṣl (confusion in use/non-use of conjunctions) الخطأ في الفصل والوصل (الخطأ في استخدام أدوات العطف)	<SF>
	25. Other semantic errors أخطاء دلالية أخرى	<SO>
5. Punctuation علامات الترقيم 'alāmāt 't-tarqīm'	26. Punctuation confusion علامة ترقيم خاطئة	<PC>
	27. Redundant punctuation علامة ترقيم زائدة	<PT>
	28. Missing punctuation علامة ترقيم مفقودة	<PM>
	29. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<PO>

3. Error-types explanation شرح أنواع الخطأ 3.

3.1 Orthography الإملاء [’imlā’]

1. Hamza (أ، إ، و، ئ، ؤ)

الهمزة

[OH – ه]

للهمزة عدة حالات حسب موضعها في الكلمة (أولها، وسطها، آخرها).
 • ففي أولها تكون إما همزة وصل أو قطع، وذلك يعتمد على البنية الصرفية للكلمة.
 • وفي وسطها وآخرها إما أن تكتب على ألف (أ) أو على واو (و) أو على ياء (ي) أو على نبرة (ئ) أو على السطر (ء)، وهذا يعتمد على حركة الهمزة وحركة الحرف الذي قبلها.
 ولأن شرح هذه القواعد يأخذ حيزاً كبيراً من هذا الدليل، فليس هذا مكاناً لإيرادها، ويكفي القول بأن المرمز يحتاج إلى الإلمام بقواعد الهمزة بشكل جيد.

Hamza has several forms based on its position in the word (beginning, middle, and end).

- At the beginning, it is either *Waṣl* or *Qati’* based on the morphological form of the word.
- In the middle and end, it can be on *’lif* (أ), *Wāw* (و), *Yā’* (ي), *Nabira* (ئ), or on the line (ء). This depends on the diacritics (short vowels) of *Hamza* itself and the preceding character.

Explaining the rules of *Hamza* may take up lots of space in this guide, which is not appropriate, so what can be said here is that it is important to choose annotators who have a solid knowledge of *Hamza* rules.

<p><i>Hamza</i> errors are identified as the following:</p> <p>1. <i>Hamza</i> confusion (put in the wrong place) <u>Example:</u> <u>سَأَل</u> (correct form: سَأَل <i>Sa’ala</i> [asked])</p> <p>2. <i>Hamza</i> redundant <u>Example:</u> <u>أَلْبَيْت</u> (correct form: البيت <i>albaiṭ</i> [home])</p> <p>3. <i>Hamza</i> missing <u>Example:</u> <u>أحمد</u> (correct form: أحمد <i>aḥmad</i> [Ahmad])</p>	<p>هذا النوع يشمل جميع أخطاء الهمزة كما يلي:</p> <p>١. الخلط في كتابة الهمزة، أي كتابتها في غير موضعها الصحيح <u>مثل:</u> <u>سَأَل (والصحيح سَأَل)</u></p> <p>٢. همزة زائدة <u>مثل:</u> <u>أَلْبَيْت (والصحيح البيت)</u></p> <p>٣. همزة مفقودة <u>مثل:</u> <u>أحمد (والصحيح أحمد)</u></p>
--	---

2. Confusion in *Hā'* and *Tā' Mutaṭarrifatain* (هـ، ء، ت)

الخطأ في الهاء والتاء المتطرفتين

[ة] – OT

الهاء والتاء المتطرفتان هما اللتان تأتيان في آخر الكلمة، وللهاء شكل واحد (هـ)، بينما تأتي التاء مفتوحةً (ت) أو مربوطةً (ة).

Hā' and *Tā' Mutaṭarrifatain* usually come at the end of words. *Hā'* has one form (هـ), while *Tā'* can be opened "*Maftūha*" (ت), or closed "*Marbūṭa*" (ة).

<p>Two types of errors:</p> <p>1. Confusion between <i>Hā' Mutaṭarrifa</i> (هـ) and <i>Tā' Marbūṭa Mutaṭarrifa</i> (ة).</p> <p><u>Example:</u></p> <p>(1) <u>المدرسه</u> (correct form: المدرسة 'almadrasa [the school])</p> <p>(2) <u>انتباه</u> (correct form: انتباه 'intibāh [attention])</p> <p>2. Confusion between <i>Tā' Marbūṭa</i> (ة) and <i>Tā' Maftūha</i> (ت) <i>Mutaṭarrifatain</i>.</p> <p><u>Example:</u></p> <p>(1) <u>غابات</u> (correct form: غابات <i>gābāt</i> [forests])</p> <p>(2) <u>نافذت</u> (correct form: نافذة <i>nāfiḍa</i> [window])</p>	<p>يشمل هذا النوع صنفين من الخطأ:</p> <p>١. الخلط بين الهاء المتطرفة (هـ) والتاء المربوطة المتطرفة (ة)</p> <p><u>مثل:</u></p> <p>(١) <u>المدرسه</u> (والصحيح المدرسة)</p> <p>(٢) <u>انتباه</u> (والصحيح انتباه)</p> <p>٢. الخلط بين التاء المربوطة (ة) والتاء المفتوحة (ت) المتطرفتين</p> <p><u>مثل:</u></p> <p>(١) <u>غابات</u> (والصحيح غابات)</p> <p>(٢) <u>نافذت</u> (والصحيح نافذة البيت)</p>
--	--

3. Confusion in '*alif* and *Yā' Mutaṭarrifatain* (ي، ا، ي)

الخطأ في الألف والياء المتطرفتين

[ا، ي] – OA

الألف والياء المتطرفين تأتيان في آخر الكلمة، فالألف إما أن تكون ممدودة (ا) أو مقصورة (ي)، أما الياء فتأتي فلها شكل واحد (ي).

'*alif* and *Yā' Mutaṭarrifatain* come at the end of a word; the '*alif* comes in two forms:

Mamdūda (ا) or *Maqṣūra* (ي), while the *Yā'* comes in one form (ي).

<p>Two types of errors:</p> <p>1. Confusion between '<i>alif Mutaṭarrifa</i> <i>Mamdūda</i> (ا) and '<i>alif Mutaṭarrifa</i> <i>Maqṣūra</i> (ي).</p> <p><u>Example:</u></p> <p>(1) <u>أنا</u> (correct form: أتى <i>atā</i> [came])</p> <p>(2) <u>سما</u> (correct form: سما <i>samā</i> [soar])</p>	<p>يشمل هذا النوع صنفين من الخطأ:</p> <p>الأول: الخلط بين الألفين الممدودة (ا) والمقصورة (ي)</p> <p><u>مثل:</u></p> <p>(١) <u>أنا</u> (والصحيح أتى)</p> <p>(٢) <u>سما</u> (والصحيح سما)</p> <p>الثاني: الخلط بين الألف المقصورة (ي) والياء (ي)</p> <p><u>مثل:</u></p>
--	---

<p>2. Confusion between <i>'alif Mutaṭarrifa Maqṣūra</i> (ى) and <i>Yā' Mutaṭarrifa</i> (ي).</p> <p><u>Example:</u></p> <p>(1) <u>القاضي</u> (correct form: القاضي <i>qāḍī</i> [the judge])</p> <p>(2) <u>الأقصى</u> (correct form: الأقصى <i>'al'aqṣā</i> [a proper noun, the name of the famous mosque in Palestine])</p>	<p>(١) <u>القاضي</u> (والصحيح القاضي)</p> <p>(٢) <u>الأقصى</u> (والصحيح الأقصى)</p>
---	---

4. Confusion in *'alif Fāriqa* (كتبوا)

الخطأ في الألف الفارقة

[OW – إت]

ألف التفريق أو الفارقة: هي ألف تزداد بعد واو الجماعة، للتنبية أن الواو ليست أصلية في الفعل بل هي واو الجماعة (ضمير).
قاعدة الألف الفارقة: إن كانت الواو ضميراً فتكتب الألف (مثل: لم يكتبوا)، وإن كانت حرفاً فلا تكتب (مثل: ا. يرجو ٢. معلمو الطلاب).

'alif Fāriqa is an *'alif* (ا) that is added after *wāw 'alġamā'a*, the plural pronoun (و), to indicate that this *wāw* is not a part of the word root, but is *wāw 'alġamā'a*, the plural pronoun (و). The rule of *'alif Fāriqa*: If the character *wāw* is a pronoun, the *'alif Fāriqa* should be added (e.g. لم يكتبوا *lam yaktubū* [did not write]), but if the *wāw* is the last character of the word root, the *'alif Fāriqa* should not be added (e.g. 1. يرجو *yarġū* [hope] 2. معلمو الطلاب *mu'allimū 'aṭṭullāb* [students' teachers]).

<p>Two types of errors:</p> <p>1. Adding <i>'alif Fāriqa</i> where it should not be added.</p> <p><u>Example:</u></p> <p>مسلمو أفريقيا (correct form: مسلموا أفريقيا <i>muslimū 'afriqyā</i> [Muslims of Africa])</p> <p>2. Omitting <i>'alif Fāriqa</i> where it should be added.</p> <p><u>Example:</u></p> <p>لم يذهبوا (correct form: لم يذهبوا <i>lam yaḍhabū</i> [they did not go])</p>	<p>هذا النوع يشمل الخلط في حالتين: الأولى: كتابة الألف الفارقة في غير موضعها <u>مثل:</u> <u>مسلموا</u> أفريقيا (والصحيح مسلمو أفريقيا)</p> <p>الثانية: إسقاط الألف الفارقة من موضعها <u>مثل:</u> لم يذهبوا (والصحيح لم يذهبوا)</p>
---	--

5. Confusion between *Nūn* (ن) and *Tanwīn* (ٍٍٍ)

الخلط بين النون والتنوين

[ON – إن]

النون المقصود هنا عبارة عن حرف أصلي من حروف الكلمة (نحو: مؤمن)، ويكتب نوناً في آخرها، ويلفظ في حال الوصل أو الوقف.

Appendix E – The Error Tagging Manual for Arabic (ETMAr)

أما التتوين فهو عبارة عن نون زائدة في آخر الاسم لفظاً لا كتابةً، ولذا تكتب على شكل حركة مضعفة: فتحتين (قرأت كتاباً مفيداً) أو ضمتين (هذا بابٌ واسعٌ) أو كسرتين (مررتُ ببيتٍ كبيرٍ)، ولفظها يكون في حال الوصل مع الكلمة التي تليها فقط، أما في الوقف فلا تنطق (رجلٌ كريم). ويحدثُ الخطأ في عدم التفريق بينها وبين حرف النون فيكتبُ التتوين نوناً.

The *nūn* (ن) intended here is one of the original word characters (e.g. مؤمن *mu'min* [believer]), which is written as *nūn* (ن) at the end of a word and is pronounced similarly, whether stopping at the end of this word or continuing to the next word.

The *Tanwīn* is an extra sound, like *nūn* at the end of a word, but not an original character. It is written as double diacritic marks (ٍٍ), double *fatha* َ (e.g. قرأت كتاباً مفيداً *qara'tu kitāban mufīdan* [I read a useful book]), double *damma* ُ (e.g. هذا بابٌ واسعٌ *hādā bābun wāsi'un* [This is a wide door]), or double *kasra* ِ (e.g. مررتُ ببيتٍ كبيرٍ *marartu bibaītin kabīrin* [I pass a big house]). The *Tanwīn* is pronounced only when continuing to the next word, but it is omitted when stopping (e.g. رجلٌ كريمٍ *rağulun karīm* [generous man]).

An error occurs when it is not distinguished between the *nūn* at the end of a word and the *Tanwīn*, so the *Tanwīn* may be written as *nūn*.

<p>This error occurs when one of the <i>Tanwīn</i> forms (ٍٍ) is written as a <i>nūn</i> (ن). <u>Example:</u> ثوبٌ جديدٍ (correct form: ثوبٌ جديدٍ <i>taūbun ġadīd</i> [a new dress])</p>	<p>هذا النوع يختص بكتابة النون مكان التتوين <u>مثل:</u> ثوبين جديد (والصحيح ثوبٌ جديد)</p>
--	--

6. Shortening the long vowels

(ٍٍ → اوي) تقصير الصوائت الطويلة

[إف - OS]

الصوائت الطويلة هي حروف العلة: الألف (ا) والواو (و) والياء (ي). وتقصيرها يكون بكتابتها حركاتٍ بدل الحروف (فتحة بدل الألف، ضمة بدل الواو، كسرة بدل الياء).

The long vowels are: 'alif (ا), Wāw (و) and Yā' (ي). Shortening those long vowels is done by replacing them with short vowels using *Fatha* (َ) instead of 'alif (ا), *damma* (ُ) instead of Wāw (و), and *Kasra* (ِ) instead of Yā' (ي).

<p>Three types of errors: 1. Writing the 'alif (ا) as <i>Fatha</i> (َ) <u>Example:</u> أوقاتٍ (correct form: أوقاتٍ <i>awqāt</i> [times]) 2. Writing the Wāw (و) as <i>damma</i> (ُ) <u>Example:</u> محامُنٍ (correct form: محامونٍ <i>Muhāmūn</i> [lawyers])</p>	<p>يشمل هذا النوع من الأخطاء ١- كتابة الألف فتحة <u>مثل:</u> أوقتٍ (والصحيح أوقات) ٢- كتابة الواو ضمة <u>مثل:</u> محامُنٍ (والصحيح محامون) ٣- كتابة الياء كسرة</p>
---	--

<p>3. Writing the <i>Yā'</i> (ي) as <i>Kasra</i> (ِ)</p> <p><u>Example:</u> عميق (correct form: عميق 'amīq [deep])</p>	<p>مثل: عميق (والصحيح عميق)</p>
--	-------------------------------------

7. Lengthening the short vowels

اوي (→ ِوُ) تطويل الصوائت القصيرة

[OG - إق]

الصوائت القصيرة هي الحركات: الفتحة (ـِ) والضممة (ـُ) والكسرة (ـِ). وهذا الخطأ عكس الخطأ السابق تماماً، فتطويل الصوائت القصير يكون بكتابتها حروفاً، حيث تكتب الفتحة ألفاً، والضممة واواً، والكسرة ياءً.

The short vowels are: *Fatha* (َ), *ḍamma* (ُ) and *Kasra* (ِ). Lengthening those short vowels is the opposite of the previous error; the short vowels are replaced with long vowels, using 'alif (ا) instead of *Fatha* (َ), *Wāw* (و) instead of *ḍamma* (ُ), and *Yā'* (ي) instead of *Kasra* (ِ).

<p>Three types of errors:</p> <p>1. Writing the <i>Fatha</i> (َ) as 'alif (ا)</p> <p><u>Example:</u> عندكا (correct form: عندك 'indaka [you have])</p> <p>2. Writing the <i>ḍamma</i> (ُ) as <i>Wāw</i> (و)</p> <p><u>Example:</u> عندهو (correct form: عنده 'indahu [he has])</p> <p>3. Writing the <i>Kasra</i> (ِ) as <i>Yā'</i> (ي)</p> <p><u>Example:</u> بيهي (correct form: به 'bihi [with it])</p>	<p>يشمل هذا النوع من الأخطاء</p> <p>٨. كتابة الفتحة ألفاً</p> <p>مثل: عندكا (والصحيح عندك)</p> <p>٢. كتابة الضمة واواً</p> <p>مثل: عندهو (والصحيح عنده)</p> <p>٣. كتابة الكسرة ياءً</p> <p>مثل: بيهي (والصحيح به)</p>
--	---

8. Wrong order of word characters

الخطأ في ترتيب الحروف داخل الكلمة

[OC - إط]

يقع هذا الخطأ غالباً بسبب السرعة في الكتابة خصوصاً على الحاسب الآلي، ولذا يعتبر أحياناً من الأخطاء الطباعية أو المطبعية. ويقصد به هنا أن تكون جميع حروف الكلمة مكتملة دون زيادة أو نقص، لكنها في غير مكانها الصحيح.

This error usually occurs because of speed-typing on computer keyboards (typos). This error occurs when the word characters are all present, but in the wrong order.

<p>This error type occurs when the word characters are in the wrong order.</p> <p><u>Example:</u></p>	<p>يشمل هذا النوع وجود حرف أو أكثر من أحرف الكلمة في غير مكانها الصحيح</p> <p>مثل: استغفر (والصحيح استغفر)</p>
---	--

<p><u>استغفر</u> (correct form: استغفر 'istağfara [ask forgiveness])</p>	
---	--

9. Replacement in word character(s)

استبدال حرف أو أحرف من الكلمة

[OR – إس]

يقع هذا الخطأ كسابقه بسبب السرعة في الكتابة. ويقصد به هنا أن يوجد حرف (أو أكثر) من خارج الكلمة مكان أحد (أو بعض) حروفها مع بقاء عدد الأحرف صحيحاً.

This error usually occurs – like the previous error – because of speed-typing on computer keyboards (typos). This error occurs when one or more characters of a word is/are replaced with one or more other characters, and the number of the word characters is still correct.

<p>This error type occurs when one or more characters of a word is/are replaced. <u>Example:</u> <u>امتنع</u> (correct form: امتنع 'imtana'a [refrain]) one character was replaced <u>Example:</u> يقلل من السكر (correct form: يقلل من السكر <i>Yuqallil min 'assukkar</i> [reduce the amount of sugar]) two characters were replaced <u>Example:</u> <u>الاعتمادات</u> (correct form: الاعتمادات <i>'al'i'timadāt</i> [funds]) three characters were replaced</p>	<p>يشمل هذا النوع استبدال حرف أو أكثر من أحرف الكلمة <u>مثل:</u> <u>امتنع</u> (<u>والصحيح</u> امتنع) استبدال حرف واحد <u>يقيب</u> من السكر (<u>والصحيح</u> يقلل من السكر) استبدال حرفين <u>الاعتمادات</u> (<u>والصحيح</u> الاعتمادات) استبدال ثلاثة أحرف، وهكذا..</p>
--	---

10. Redundant character(s)

حرف أو أحرف زائدة

[OD – إز]

هذا الخطأ يحدث عندما توجد أحرف الكلمة كاملة إضافة إلى حرف أو أحرف زائدة، سواء من أحرف الكلمة نفسها أو من خارجها.

This error occurs when all characters of a word are present, in addition to further characters either from those used in the word or from others.

<p>This error type includes repeating one or more of the word's characters, or adding further characters. <u>Example:</u> <u>كتبتت</u> (correct form: كتبت <i>katabtu</i> [I wrote])</p>	<p>يشمل هذا النوع من الأخطاء زيادة حرف (أو أكثر) في الكلمة سواء من جنس أحرفها أو من غير أحرفها <u>مثل:</u> <u>كتبتت</u> (<u>والصحيح</u> كتبت) زيادة حرف واحد من نفس أحرف الكلمة</p>
---	---

<p>one character of the word was repeated <u>Example:</u> المستوصفات (correct form: المستوصفات) <i>‘almustawsafāt</i> [the health centres] two characters were added</p>	<p>المستوصفات (والصحيح المستوصفات) زيادة حرفين من خارج أحرف الكلمة</p>
---	--

11. Missing character(s)

حرف أو أحرف ناقصة

[OM – إن]

هذا الخطأ يحدث عند إسقاط أحد أو بعض أحرف الكلمة.

This error occurs when one or more characters of a word are omitted.

<p>This error type includes omitting one or more of the word’s characters. <u>Example:</u> الحاسب الآلي (correct form: الحاسب الآلي) <i>‘alḥāsib_‘al’ālī</i> [the computer] one character of the word was omitted <u>Example:</u> المرضى في (correct form: المرضى في) <i>‘almarḍā fī ‘almustašfayāt</i> [the hospitals] two characters were omitted</p>	<p>يشمل هذا النوع سقوط حرف أو أكثر من الكلمة <u>مثل:</u> الحاب الآلي (والصحيح الحاسب الآلي) سقوط حرف واحد المرضى في (والصحيح المستشفيات) سقوط حرفين</p>
---	--

12. Other orthographical errors

أخطاء إملائية أخرى

[OO – إخ]

<p>All uncatagorisable error types should be placed here, and when there is a group of errors that can be separated, a new error type can be created.</p>	<p>تحت هذا النوع يوضع كل خطأ لا تشمله الأنواع السابقة، وعند وجود مجموعة من الأخطاء التي تمثل نوعاً جديداً واضح المعالم، فيمكن إنشاء بند جديد خاص به.</p>
---	--

3.2 Morphology [’ssarf]

13. Word inflection

الخطأ في اختيار بنية الكلمة المناسبة

[MI – صص]

المقصود ببنية الكلمة الصيغة الصرفية لها. اختلاف البنية الصرفية للكلمة قد يسبب عدة أنواع من الخطأ، بعضها صرفي وبعضها نحوي كما يلي:

- قد تختلف الصيغة الصرفية للفعل مما يسبب اختلافاً في زمنه، وهذا يمثل الخطأ الصرفي في زمن الفعل.
- قد تختلف الصيغة الصرفية مما يسبب عدم التطابق إما في العدد أو الجنس أو التعريف أو الإعراب، وهذا تمثله الأخطاء النحوية الخاصة بالتطابق في هذه النقاط الأربع.

Appendix E – The Error Tagging Manual for Arabic (ETMAr)

- قد تختلف الصيغة الصرفية دون أن تسبب أياً من الأخطاء السابقة، فيكون خطأ صرفياً عاماً تحت هذا النوع.

Word inflection is the morphological form of the word. Using a different form of a word may cause morphological or syntactic errors as follows:

- Using a different form of a verb may lead to an incorrect verb tense. This error is represented by the error: verb tense.
- Using a different form may lead to disagreement in number, gender, definiteness, or case. These errors are represented by four error types in the Syntax section.
- Using a different form may lead to an error, but not one covered by any of those already mentioned. This can be considered a general morphological error, classified under the current error type.

<p>This error type includes using an incorrect morphological form, but not in verb tense or classified in the Syntax section.</p> <p><u>Example:</u> كتابة <u>الدرس</u> (correct form: <u>اكتتاب</u> <u>الدرس</u> <i>kitābata ‘addars</i> [to write the lesson])</p>	<p>يشمل هذا النوع استخدام بنية صرفية خاطئة، مع عدم ارتباطه بزمن الفعل أو الأخطاء النحوية</p> <p><u>مثل:</u> <u>اكتتاب</u> <u>الدرس</u> (<u>والصحيح</u> كتابة <u>الدرس</u>)</p>
---	--

14. Verb tense

الخطأ في زمن الفعل

[MT – صز]

الأفعال في اللغة العربية تدل على أزمنة ثلاثة: ماضٍ وحاضر ومستقبل، حسب صيغتها الصرفية. ولا بد للأفعال أن تراعي الزمن الذي يتحدث عنه النص، فإذا تم استخدام صيغة مختلفة (حاضر بدل الماضي، أو الماضي بدل المستقبل مثلاً) فقد يختل المعنى المقصود.

Verbs in Arabic indicate three tenses (past, present and imperative) based on their forms. The verb used should be consistent with the context in terms of its tense. Using different verb tenses may lead to different meanings.

<p>This error type includes using an incorrect verb tense.</p> <p><u>Example:</u> غداً سنشتري (correct form: <u>اشترينا</u> الملابس <i>ḡadan sanaštari ‘almalābis</i> [tomorrow, we will buy the clothes]) The past tense (<u>اشترينا</u> [bought]) was used in a sentence about the future, the correct form of the verb is (<u>سنشتري</u> [will buy]), which indicates the future</p>	<p>يشمل هذا النوع بنية صرفية تدل على زمن خاطئ</p> <p><u>مثل:</u> غداً <u>اشترينا</u> الملابس (<u>والصحيح</u> غداً سنشتري الملابس) تم استعمال صيغة الفعل الماضي (اشترينا) في جملة تتحدث عن المستقبل، والصحيح استخدام الفعل الدال على نفس الزمن (سنشتري)</p>
--	--

15. Other morphological errors

أخطاء صرفية أخرى

[صخ - MO]

All uncatagorisable error types should be placed here, and when there is a group of errors that can be separated, a new error type can be created.

تحت هذا النوع يوضع كل خطأ لا تشمله الأنواع السابقة، وعند وجود مجموعة من الأخطاء التي تمثل نوعاً جديداً واضح المعالم، فيمكن إنشاء بند جديد خاص به.

3.3 Syntax النحو [’nnaḥw]

16. Case

الخطأ في الإعراب

[XB - نب]

يقصد بالخطأ في الإعراب تغيير حالة الكلمة عما يفترض أن تكون في السياق: مرفوعة أو منصوبة أو مجرورة أو مجزومة.

The error in word case occurs when a word is in the wrong case from which it should be based on the context: Nominative, Genitive, Accusative, etc.

This error type includes the error in a word case.

Example:

على الحاضرين (**correct form:** على الحاضرون
'alā 'alḥāḍirīn [on the attendees])

يشمل هذا النوع الخطأ في إعراب الكلمة

مثل:
على الحاضرون (والصحيح على الحاضرين)

17. Definiteness

الخطأ في التعريف والتنكير

[XF - نع]

المعارف سبعة أنواع هي: العلم، والضمير، والاسم الموصول، واسم الإشارة، والمعرب ب(ال)، والمضاف إلى إحدى المعارف السابقة، والمنادى المقصود تعيينه بالنداء. فإذا لم تكن الكلمة من هذه الأنواع فهي نكرة. يقصد بالخطأ في التعريف والتنكير أن تكون الكلمة مخالفة لكلمة أخرى يلزم تطابقهما، أو مخالفة لما يفترض أن يكون عليه السياق.

Types of definite nouns are: proper noun, pronoun, relative noun, demonstrative noun, definite noun with a definite article 'al (ال), indefinite noun added before a definite noun from the above, and the noun addressed by Yā (يا). Other types of nouns are indefinite.

The error in definiteness occurs when there is no agreement between two definite or indefinite nouns when there should be, or between the noun used and what it should be in the context.

This error type occurs:

2. When there is no agreement between two definite or indefinite nouns when there

يشمل هذا النوع:

٢. التطابق في التعريف والتنكير بين كلمتين أو أكثر في الجملة

<p>should be</p> <p><u>Example:</u> اشترت السيارة الحمراء (correct form: اشترت السيارة الحمراء <i>'ištarāitu 'alssayyārata 'alḥamrā'</i> [I bought the red car])</p> <p>The word (السيارة [the car]) is definite, while its adjective (الحمراء [red]) is not, so a definite adjective should be used (الحمراء [the red]).</p> <p>2. When there is no agreement between the noun used and what it should be in the context</p> <p><u>Example:</u> عندما وصلنا كان الجو جميلاً (correct form: عندما وصلنا كان جو جميلاً <i>'indamā wasalnā kān 'alġawwu ġamīlan</i> [when we arrived, the weather was nice])</p> <p>The word (جو [weather]) is indefinite, while in this context, it should be definite (الجو [the weather]).</p>	<p><u>مثل:</u> اشترت السيارة الحمراء (والصحيح اشترت السيارة الحمراء) خطأ في تطابق التعريف والتذكير بين الصفة والموصوف</p> <p>٢. خطأ في تعريف وتكثير الكلمة مع ما يفترض أن يكون عليه الكلام</p> <p><u>مثل:</u> عندما وصلنا كان جو جميلاً (والصحيح عندما وصلنا كان الجو جميلاً)</p>
---	---

18. Gender

الخطأ في الجنس (التذكير والتأنيث)

[XG – نذ]

التطابق في الجنس يكون لازماً أحياناً بين كلمتين أو أكثر في الجملة، أو بين الكلمة مع ما يفترض أن يكون عليه الكلام.

ولذا فإن عدم وجود التطابق يسبب خللاً في بنية الجملة.

Agreement in gender is sometimes required. This can be either between two or more words, or between a word and its context. Disagreement in these cases may cause an error in sentence structure.

<p>This error type includes:</p> <p>2. When there is no agreement in gender between two nouns when there should be</p> <p><u>Example:</u> اشترت كتاباً جديداً (correct form: اشترت كتاباً جديداً <i>'ištarāitu kitāban ġadīdan</i> [I bought a new book])</p> <p>The word (كتاباً [book]) is masculine, while its adjective (جديداً [new-F]) is feminine, so a masculine adjective should be used (جديداً [new-M]).</p> <p>2. When there is a no agreement in gender between a noun used and what it should be in the context</p> <p><u>Example:</u></p>	<p>يشمل هذا النوع:</p> <p>٢. التطابق في الجنس بين كلمتين أو أكثر في الجملة</p> <p><u>مثل:</u> اشترت كتاباً جديداً (والصحيح اشترت كتاباً جديداً) خطأ في تطابق الجنس بين الصفة والموصوف حيث استخدمت صفة مؤنثة مع موصوف مذكر، والصحيح استخدام صفة مذكرة</p> <p>٢. التطابق في الجنس بين كلمة مع ما يفترض أن يكون عليه الكلام</p> <p><u>مثل:</u> أعطت المعلمة الطالب كتابها (والصحيح</p>
---	---

<p>أعطت المعلمة (correct form: أعطت المعلمة الطالب كتابها أعطت المعلمة الطالبة كتابها) <i>a'tat 'almu'allimatu 'alṭālibata kitābahā</i> [the teacher gives the book to her student]) The word (الطالب [student]) is masculine, while in this context, it should be feminine (الطالبة [student- F]).</p>	<p>أعطت المعلمة الطالبة كتابها)</p>
--	-------------------------------------

19. Number (singular, dual and plural)

الخطأ في العدد (الإفراد والتثنية والجمع)

[XN – نف]

التطابق في العدد يكون لازماً أحياناً بين كلمتين أو أكثر في الجملة، أو بين الكلمة مع ما يفترض أن يكون عليه الكلام. ولذا فإن عدم وجود التطابق يسبب خللاً في بنية الجملة.

Agreement in number is sometimes required. This can be either between two or more words, or between a word and its context. Disagreement in these cases may cause an error in sentence structure.

<p>This error type includes:</p> <p>2. When there is no agreement in number between two nouns when there should be</p> <p><u>Example:</u> اشترت ثياباً بيضاً (correct form: اشترت ثياباً أبيضاً) <i>'ištaraītu ṭiyāban bīḍan</i> [I bought white clothes]) The word (ثياباً [clothes]) is plural, while its adjective (أبيض [white-S]) is singular, so a plural adjective should be used (أبيضاً [white-P]).</p> <p>2. When there is no agreement in number between a noun used and what it should be in the context</p> <p><u>Example:</u> تشرق الشمس قبل الساعة السادسة (correct form: تشرق الشمس قبل الساعة السادسة) <i>tušriqu 'alššamsu qabla 'alsā'a</i> <i>A'lsādisa</i> [the sun rises before six o'clock]) The word (الشمس [sun-P]) is plural, while in this context, it should be singular (الشمس [sun-S]).</p>	<p>يشمل هذا النوع:</p> <p>٢. التطابق في العدد بين كلمتين أو أكثر في الجملة</p> <p><u>مثل:</u> اشترت ثياباً أبيضاً (والصحيح اشترت ثياباً بيضاً) خطأ في تطابق العدد بين الصفة والموصوف</p> <p>٢. التطابق في العدد بين كلمة مع ما يفترض أن يكون عليه الكلام</p> <p><u>مثل:</u> تشرق الشمس قبل الساعة السادسة (والصحيح تشرق الشمس قبل الساعة السادسة)</p>
--	---

20. Redundant word

كلمة زائدة

[XT – نز]

يحدث هذا الخطأ غالباً عند تكرار كلمة أكثر من مرة مع عدم الحاجة للكلمة المكررة. كما يحدث كذلك عند وجود كلمة زائدة عن حاجة الجملة، مما يسبب خللاً في بنيتها النحوية.

Appendix E – The Error Tagging Manual for Arabic (ETMAr)

This error usually occurs when a word is repeated more than once, and there is no need for the second word. It also occurs when there is a redundant word that causes an error in sentence structure.

<p>This error type includes:</p> <p>2. Repeating the word more than once when not necessary</p> <p><u>Example:</u> <i>wasaltu</i> وصلت إلى بيتي (correct form: وصلت إلى بيتي) <i>'ilā 'albaytī</i> [I've arrived at my house]) The word (إلى [at]) was repeated twice.</p> <p>2. When there is a redundant word that causes an error in the sentence structure</p> <p><u>Example:</u> ذهبت إلى المدرسة (correct form: ذهبت إلى عند المدرسة) <i>dahabtu 'ilā 'almadrasa</i> [I went to school]) The word (عند [at]) was redundant.</p>	<p>يشمل هذا النوع: ٢. تكرار الكلمة أكثر من مرة داخل الجملة <u>مثل:</u> وصلت إلى إلى بيتي (والصحيح وصلت إلى بيتي)</p> <p>٢. وجود كلمة أو كلمات زائدة عن البنية الصحيحة للجملة <u>مثل:</u> ذهبت إلى عند المدرسة (والصحيح ذهبت إلى المدرسة)</p>
---	---

21. Missing word

كلمة ناقصة

[XM - نن]

هذا الخطأ يحدث عند سقوط كلمة من الجملة مما يسبب خللاً في بنيتها النحوية.

This error occurs when a word is omitted from the sentence, which leads to an error in sentence structure.

<p>This error type occurs when a word is omitted from the sentence.</p> <p><u>Example:</u> تحدثنا عن المشروع (correct form: تحدثنا المشروع) <i>taḥaddatnā 'an 'almašrū'</i> [we talked about the project]) The word (عن [about]) was omitted.</p>	<p>يشمل هذا النوع سقوط كلمة أو أكثر من الجملة <u>مثل:</u> تحدثنا المشروع (والصحيح تحدثنا عن المشروع)</p>
---	--

22. Other syntactic errors

أخطاء نحوية أخرى

[XO - نخ]

<p>All uncatagorisable error types should be placed here, and when there is a group of errors that can be separated, a new type</p>	<p>تحت هذا النوع يوضع كل خطأ لا تشمله الأنواع السابقة، وعند وجود مجموعة من الأخطاء التي تمثل نوعاً جديداً واضح المعالم، فيمكن إنشاء بند جديد خاص</p>
---	--

can be created.

به.

3.4. Semantics [’ddalāla]

23. Word selection

الخطأ في اختيار الكلمة المناسبة

[SW – دب]

الاختيار المناسب للكلمات يعتمد على معرفة معانيها المعجمية، وكذلك السياقات المناسبة لاستخدامها، ومن هنا ينشأ الخطأ الدلالي في عدم اختيار الكلمة المناسبة للسياق الذي ورت فيه.

Appropriate selection of a word depends on its lexical meaning and the suitable contexts when it can be used. A semantic error may arise when the word selected is inappropriate for the context.

This error type includes selecting a word that is inappropriate for the context.

Example:

[I miss him; we were together for a long time] اشتقت له، فقد طال الالتقاء بيننا (correct form: اشتقت له، فقد طال البعد بيننا) *’ištiqtu lahu faqad ṭāla ‘albu’du baynanā* [we have been away for a long time]

The word [البعد] [away] is more appropriate for the context from the word [الالتقاء] [together]

يشمل هذا النوع عدم الدقة في اختيار المفردة المناسبة للسياق

مثل:

اشتقت له، فقد طال الالتقاء بيننا (والصحيح اشتقت له، فقد طال البعد بيننا) الخطأ في اختيار الكلمة المناسبة

24. Faṣl wa waṣl (confusion in use/non-use conjunctions)

الفصل والوصل (سوء استخدام أدوات العطف)

[SF – دف]

الفصل والوصل أحد العناصر الهامة في توضيح معاني الجمل، وعند الوصل تستخدم حروف العطف غالباً، وتحذف عند الفصل، وقد يؤدي ربط الجمل أحياناً – أو فصلها عن بعضها – إلى معنى خاطئ، وهو المقصود بالخطأ في الفصل والوصل.

Faṣl wa waṣl (conjunctions) are important factors for clarifying sentence meanings.

Conjunctions are usually used for connecting sentences (*waṣl*), and they are omitted for disconnecting (*Faṣl*). Sometimes, using conjunctions – or omitting them – may lead to an unintended meaning, which falls under this error type.

This error type includes:

1. Using a conjunction where it is not appropriate

Example:

عندما رأيت مشغولاً فلم أكلمه إلى أن كلمني (correct form: عندما رأيت مشغولاً لم أكلمه إلى أن كلمني) *’indamā ra’aytuhu maṣḡulan lam ‘ukallimhu ‘ilā ‘an kallamanī* [when I saw he was busy, I did not talk to him till he started talking to

يشمل هذا النوع:

١. وصل الجمل في موضع يفترض فيه الفصل

مثل:

عندما رأيت مشغولاً فلم أكلمه إلى أن كلمني (والصحيح عندما رأيت مشغولاً لم أكلمه إلى أن كلمني)

<p>me]) The conjunction (ف [and]) was not appropriate in this context.</p> <p>2. Omitting a conjunction where it is appropriate Example: لقيت زميلي في الطريق، فصحبته إلى المدرسة (correct form: لقيت زميلي في الطريق، صحبته إلى المدرسة <i>laqītu zamīlī fī 'alṭarīq faṣaḥibtuhu 'ilā 'almadrasa</i> [I met my classmate on my way and accompanied him to the school]) The conjunction (ف [and]) was more appropriate in this context.</p>	<p>٢. فصل الجمل في موضع يفترض فيه الوصل مثل: لقيت زميلي في الطريق، صحبته إلى المدرسة (والصحيح لقيت زميلي في الطريق، فصحبته إلى المدرسة)</p>
--	---

25. Other semantic errors

أخطاء دلالية أخرى

[دخ - SO]

<p>All uncatagorisable error types should be placed here, and when there is a group of errors that can be separated, a new error type can be created.</p>	<p>تحت هذا النوع يوضع كل خطأ لا تشمله الأنواع السابقة، وعند وجود مجموعة من الأخطاء التي تمثل نوعاً جديداً واضح المعالم، فيمكن إنشاء بند جديد خاص به.</p>
---	--

3.5. Punctuation علامات الترقيم ['alāmāt 't-tarqīm]

26. Punctuation confusion

علامة ترقيم خاطئة

[تظ - PC]

علامات الترقيم هي علامات تستخدم عند كتابة النصوص لتسهيل قراءة النص، وبيان أماكن الوقف مثل نهاية الجمل والفقرات، ومن هذه العلامات:
الفاصلة (،): تستخدم عدة مواضع منها الفصل بين الجمل القصيرة، وبين أقسام الشيء، وبعد المنادى.
الفاصلة المنقوطة (؛): تستخدم بين جملتين أحدهما سبب للأخرى، أو بين مجموعتين أو أكثر من العبارات أو الجمل القصيرة.
النقطة (.): تستخدم في نهاية الجمل التامة، وفي نهاية الفقرات.
علامة الاستفهام (?): تستخدم بعد الجمل الاستفهامية.
علامة التعجب (!): تستخدم بعد الجمل التعجبية، والانفعالية.
النقطتان الرأسيتان (:): تستخدم قبل القول، أو المثال، أو التعريف؛ وبين الشيء وأقسامه.
القوسان (()): تستخدم حول الجمل التفسيرية ونحوها.
علامة التنصيص أو الاقتباس (" "): تستخدم للكلام المنقول بنصه.
الشرطة (-): تستخدم حول الجمل الاعتراضية.
علامة الحذف (...): تستخدم عند الاستغناء عن بعض الكلام.

Punctuation includes those marks used when writing to facilitate text readability and to clarify the stop or pause positions in sentence flow, such as the end of a sentence or a paragraph. The following are examples of punctuation:

Appendix E – The Error Tagging Manual for Arabic (ETMAr)

Comma (،): used in a number of situations, e.g., after dependent clauses, between items in a series, and after the noun addressed by *Yā* (يا).

Semicolon (;): used between two sentences when one of them is the reason for the other, or between two clauses, or short sentences.

Full stop or period (.): used at the end of complete sentences or paragraphs.

Question mark (?): used after questions.

Exclamation point (!): used after strong emotions.

Colon (:): used before speech, examples, definitions, or parts.

Brackets (()): used around interpretation sentences or related items.

Quotation marks (""): used for direct quotes.

Dash (-): used for adding emphasis or an interruption.

Ellipsis (...): used for omission of words.

<p>This error type includes using incorrect punctuation. <u>Example:</u> من أنت؟ (correct form: من أنت؟) [who are you?] A comma was used where a question mark should have been used.</p>	<p>يشمل هذا النوع حدوث خطأ في استخدام علامة الترقيم المناسبة، بحيث يتم استخدام علامة في موضع علامة أخرى. <u>مثل:</u> من أنت؟ (والصحيح من أنت؟)</p>
---	--

27. Redundant punctuation

علامة ترقيم زائدة

[PT – تز]

<p>This error type includes using punctuation when none should be used. <u>Example:</u> وكيف نجوتم؟ (correct form: وكيف نجوتم؟) <i>wakayfa naġawtum</i> [and how did you survive?] The first question mark is redundant.</p>	<p>يشمل هذا النوع استخدام علامة ترقيم في موضع لا يفترض أن توجد فيه أي علامة. <u>مثل:</u> وكيف؟ نجوتم؟ (والصحيح وكيف نجوتم؟)</p>
--	---

28. Missing punctuation

علامة ترقيم مفقودة

[PM – تن]

<p>This error type includes omitting punctuation when it should be used <u>Example:</u> (correct form: وعندها غادرت البيت وأقفلت الباب وركبت السيارة. وعندها غادرت البيت، وأقفلت الباب، وركبت السيارة. <i>wa'indaha gādartu 'albayit wa'aqfaltu 'albāb</i></p>	<p>يشمل هذا النوع سقوط علامة ترقيم من موضعها، بحيث يترك المكان خالياً دون علامة في حين يفترض وجودها. <u>مثل:</u> وعندها غادرت البيت وأقفلت الباب وركبت السيارة. (والصحيح وعندها غادرت البيت، وأقفلت الباب، وركبت السيارة.)</p>
--	--

<p><i>warakibtu 'alssayyāra</i> [and then I left the house, locked the door, and got into the car]) A comma was missed between the phrases.</p>	
---	--

29. Other errors in punctuation

أخطاء أخرى في علامات الترقيم

[تخ - PO]

<p>All uncatégorisable error types should be placed here, and when there is a group of errors that can be separated, a new error type can be created.</p>	<p>تحت هذا النوع يوضع كل خطأ لا تشمله الأنواع السابقة، وعند وجود مجموعة من الأخطاء التي تمثل نوعاً جديداً واضح المعالم، فيمكن إنشاء بند جديد خاص به.</p>
---	--

4. Method of error annotating طريقة ترميز الأخطاء

<p>Three things should be properly identified when tagging:</p> <ol style="list-style-type: none"> 1. Error form 2. Error category and type 3. Correcting form <p><u>Example:</u> Text: "وقد ركبنا السيارة الكبير، وذهبنا في نزهة" <i>Waqad rakibnā 'assayyāra 'alkabīr, waḡahabnā fī nuzha</i> [and we got into the big [gender: M] car [gender: F] and went on a journey]</p> <ol style="list-style-type: none"> 1. Error form: "الكبير" <i>'alkabīr</i> [big]+[gender: M] 2. Error category and type: Syntax, Agreement in gender. (The adjective "الكبير" <i>'alkabīr</i> [big]+[gender: M] did not have an agreement with the noun "السيارة" <i>'assayyāra</i> [car]+[gender: F]) 3. Correct form: "الكبيرة" <i>'alkabīra</i> [big]+[gender: F] 	<p>عند ترميز الخطأ لا بد من تحديد ثلاث نقاط بدقة:</p> <ol style="list-style-type: none"> ١. العبارة الخاطئة ٢. مجال الخطأ ونوعه (من خلال جدول الأخطاء) ٣. العبارة الصحيحة <p><u>مثال:</u> النص: "وقد ركبنا السيارة الكبير، وذهبنا في نزهة" ١. العبارة الخاطئة: "الكبير" ٢. مجال الخطأ ونوعه: خطأ نحوي، المطابقة في الجنس (لم تتطابق الصفة "الكبير" [مذكر] مع الموصوف "السيارة" [مؤنث]) ٣. العبارة الصحيحة: "الكبيرة"</p>
---	---

5. Rules of tagging and overlap instances

٥,١ القاعدة الأولى: الترميز يكون على أساس الكلمة الخاطئة، وليس الصحيحة (كما عند Dagneaux et al, 2005).

فمثلاً كلمة "أخذت" (والصحيح "أخذ") ترمز على أنها خاطئة إملائياً لوجود حرف زائد. وليس مقبولاً ترميز الكلمة الصحيحة "أخذت" على أن بها حرفاً ناقصاً عن الكلمة الخاطئة.

5.1 Rule 1: Tagging should be performed on the basis of the incorrect form, not the correct form (as in Dagneaux et al, 2005).

For example, the word "أخذت" 'ahḡada [took] (correct form: "أخذ" 'ahḡada) should be tagged with the orthographical error "Redundant character(s)". It is not acceptable to tag the correct form "أخذ" as it has a missing character compared to the incorrect one "أخذت".

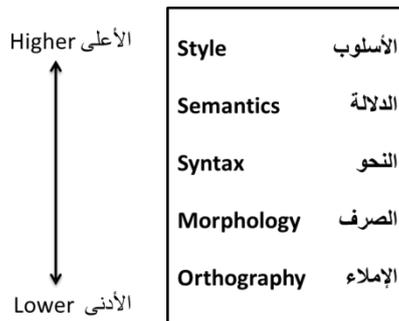
٥,٢ القاعدة الثانية: اختيار مجال الخطأ لا بد أن يكون حسب ترتيب محدد (ما عدا علامات الترقيم) ابتداءً من الأعلى (الدلالة) وانتهاءً بالأدنى (الإملاء) حسب ترتيب المستويات اللغوية، انظر الشكل أدناه.

أظهر اختبار جدول الترميز أنه عندما ينطبق اثنان من المجالات على خطأ واحد فإن المجال الأعلى هو الأنسب غالباً، إلا أن يكون هناك سبب واضح لعكس ذلك. علامات الترقيم لا تدخل في هذه القاعدة لأنها لا تتداخل مع المجالات الأخرى.

من المهم ملاحظة أنه ينبغي أن يكون تصحيح علامات الترقيم بعد تصحيح النص نفسه، حيث إنها مبنية على الشكل النهائي للنص لتؤدي دورها في تسهيل قراءته وفهمه.

5.2 Rule 2: Choosing an error category should be based on specific order (except punctuation), starting from the highest (style) to the lowest (orthography). See the figure below.

Testing the tagset showed that when two categories are applicable to one error, usually the higher one is the most appropriate, unless there is a clear reason for the opposite. The punctuation category is not included in this rule, as it does not overlap with other categories. It is **IMPORTANT** to notice that punctuation should be corrected after text corrections, as they depend on the final form of the text to make it more readable and understandable.



الأمثلة التالية تشرح هذه القاعدة بتفصيل أكثر:

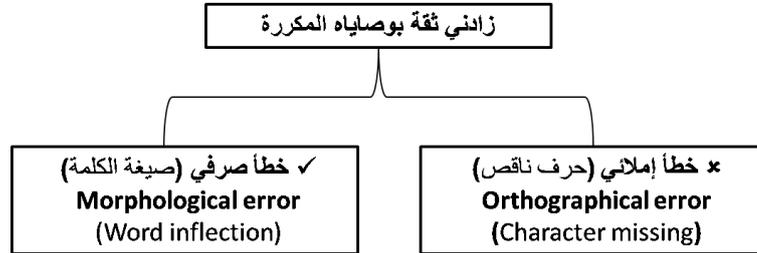
١, ٢, ٥ الخطأ الصرفي أعلى من الخطأ الإملائي

الخطأ في جملة "زادني ثقة بوصاياه المكررة" (والصحيح "المتكررة" – لأن صيغة المكررة تدل على الملل وعدم الفائدة، أما المتكررة فتدل على حرص الناصح بتكراره لنصيحته) يمكن أن يكون (١) إملائياً (حرف مفقود – "ت")، و (٢) صرفياً (صيغة الكلمة – "المتكررة")، ولكن نظراً لأن كلمة "المكررة" هي كلمة صحيحة إملائياً، فالاحتمال الأكبر أن الخطأ يرجع إلى اختيار الصيغة الصرفية المناسبة للسياق أكثر من كونه نسياً لحرف التاء. فالخطأ هنا صرفي وليس إملائي.

The following instances show more details about this rule:

5.2.1 The morphological error is higher than the orthographical

The error in the sentence "زادني ثقة بوصاياه المكررة" *zādānī ṭīqatan biwaṣāyāh 'almukarrara* [he made me more confident by his repeated advice] (correct form: "المتكررة" *'almutakarrira* – as the inflection "المكررة" *'almukarrara* shows that the advice was boring and useless, while "المتكررة" *'almutakarrira* indicates the adviser's concern by repeating the advice) can be (1) **orthographical** (character missing - "ت"), and (2) **morphological** (Word inflection - "المتكررة"). However, given that the word "المكررة" is a correct word orthographically, the error is more likely to relate to selecting the suitable inflection rather than missing the character "ت". So, the error here is morphological, not orthographical.



٢, ٥ الخطأ النحوي أعلى من الخطأ الصرفي

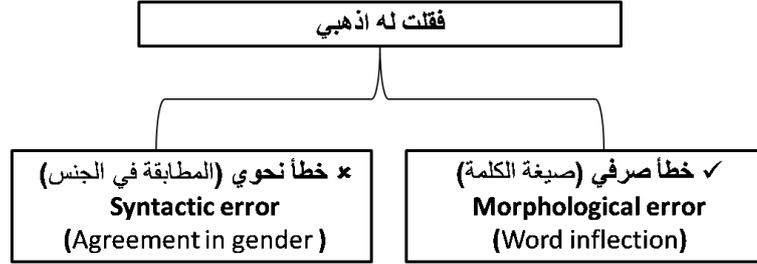
الخطأ في جملة "فقلت له اذهبي" (والصحيح "اذهب") يمكن أن يكون (١) صرفياً (صيغة الكلمة – اختار الطالب كلمة "اذهبي" مكان "اذهب")، و (٢) نحوياً (المطابقة في الجنس – لم يتطابق الفعل المؤنث "اذهبي" مع الضمير المذكر في "له")، ولكن نظراً لأن تحديد الصيغة الصحيحة للكلمة يبنى على المطابقة النحوية فإن الخطأ هنا نحوي. **ملاحظة:** بناء على المثال السابق يمكن القول بأن الأنواع الأربعة الأولى من الأخطاء النحوية تستلزم وجود اختلاف في الصيغة الصرفية، ومع ذلك فإن الخطأ يبقى نحوياً لأنه أكثر تحديداً من الخطأ الصرفي.

5.2.2 The syntactic error is higher than the morphological

The error in the sentence "فقلت له اذهبي" *faqltu lahu 'idhabī* [I told him (gender: M) to go (gender: F)] (correct form: "اذهب" *'idhab* [go (gender: M)]) can be (1) **morphological** (Word inflection – selecting the inflection "اذهبي" *'idhabī* [go (gender: F)] instead of "اذهب" *'idhab* [go (gender: M)]), and (2) **syntactic** (Agreement in gender – the verb form "اذهبي" *'idhabī* [go (gender: F)] does not agree with the pronoun "له" *lahu* [him (gender: M)]). However,

given that selecting the right form of the word is based on the syntactic agreement, the error here is syntactic.

Note: Based on the previous example, it can be said that the first four errors in the syntactic category have to include differences in the words' inflections, but such errors are still syntactic, as these four types are more specific than the morphological type.

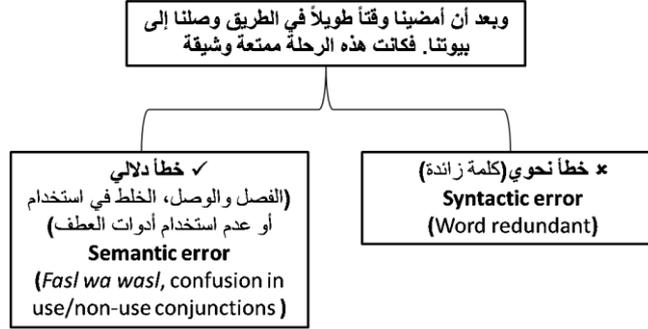


٣, ٢, ٥ الخطأ الدلالي أعلى من الخطأ النحوي

الخطأ في جملة "وبعد أن أمضينا وقتاً طويلاً في الطريق وصلنا إلى بيوتنا. فكانت هذه رحلتي الأولى لهذا المكان." (والصحيح "وبعد أن أمضينا وقتاً طويلاً في الطريق وصلنا إلى بيوتنا. كانت هذه رحلتي الأولى لهذا المكان" – بحذف حرف العطف للفصل بين الجملتين) يمكن أن يكون (١) نحويًا (كلمة زائدة – حرف العطف "ف")، و (٢) دلاليًا (الفصل والوصل، أي الخطأ في استخدام أدوات العطف – حيث تم هنا الوصل بين جملتين منفصلتين في المعنى). وحيث إن الغرض هنا من إضافة الكلمة "ف" هو الربط بين الجملتين، فالغالب أنه مقصود ولا يمكن اعتباره كلمة زائدة، أي يُستبعد أن تكون قد وُضعت عن طريق الخطأ. ولذا فإن الخطأ هنا دلالي وليس نحويًا.

5.2.3 The semantic error is higher than the syntactic

The error in the sentence "وبعد أن أمضينا وقتاً طويلاً في الطريق وصلنا إلى بيوتنا. فكانت هذه الرحلة ممتعة" (and after we spent a long time on the road, we arrived at our homes, so the trip was exciting) (**correct form:** "وبعد أن أمضينا وقتاً طويلاً في الطريق وصلنا إلى بيوتنا. فكانت هذه الرحلة ممتعة وشيقة" *waba'da 'an 'amdaynā waqtan ṭawilan fī 'aṭṭarīq waṣalnā 'ilā buyutinā. fakānat hādīhi 'arriḥla mumti'a* [and after we spent a long time on the road, we arrived at our homes, so the trip was exciting]) (**correct form:** "وبعد أن أمضينا وقتاً طويلاً في الطريق وصلنا إلى بيوتنا. فكانت هذه الرحلة ممتعة وشيقة" *waba'da 'an 'amdaynā waqtan ṭawilan fī 'aṭṭarīq waṣalnā 'ilā buyutinā. kānat hādīhi 'arriḥla mumti'a* [and after we spent a long time on the road, we arrived at our homes. The trip was exciting] – the conjunction should be deleted to separate the two sentences) can be (1) **syntactic** (Redundant word – the conjunction "ف" *fa* [so]), and (2) **semantic** (*Faṣl wa waṣl*, confusion in use/non-use of conjunctions – in this case, two semantically independent sentences were connected by a conjunction). It is very likely that the purpose of adding the word "ف" here was to connect the two sentences, rather than as a redundancy – namely, the word "ف" was not added by mistake. So the error here is semantic, not syntactic.



٤, ٢, ٥ مثال على الاستثناءات

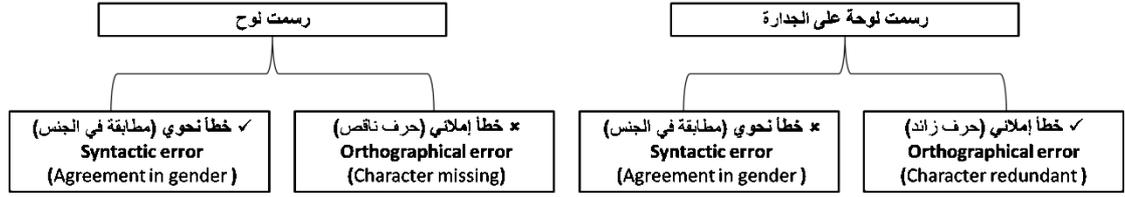
قد توجد بعض الاستثناءات من الأمثلة السابقة، ولكن ينبغي أن يكون السبب واضحاً. فمثلاً الخطأ في جملة "رسمت لوحة على الجدار" (والصحيح "الجدار") يمكن أن يكون (١) إملائياً (حرف زائد – التاء المربوطة "ة")، و (٢) نحوياً (المطابقة في الجنس – عدم وجود تطابق بين الكلمة المستخدمة "الجدار" وما هو مستخدم في العادة "الجدار"). ولأن كلمة "الجدار" لا يوجد لها مؤنث من الأصل حتى يتم اختيار الجنس المناسب منها لمطابقته، فليس الخطأ فيها نحوياً (المطابقة)، لكنه إملائى (وجود حرف زائد على الكلمة وهو التاء المربوطة).

هذا بخلاف الخطأ في جملة "رسمت لوح" (والصحيح "لوحة")، حيث إن الجملة الصحيحة في العادة أن ترسم "لوحة" وليس لوحاً، ولأن المذكر "لوح" والمؤنث "لوحة" صحيحين إملائياً، فالغالب أن الخطأ هنا يعود لاختيار الكلمة التي تتطابق في الجنس مع ما هو مستخدم، ولذا فالخطأ هنا نحوي وليس إملائياً.

5.2.5 Examples of exceptions

Some exceptions from the previous examples can exist, but there should be a clear reason. For instance, the error in the sentence "رسمت لوحة على الجدار" *rasamtu lawḥatan 'alā 'alġidāra* [I've drawn a painting on the wall (gender: F)] (**correct form**: "الجدار" *alġidār* [the wall (gender: M)]) can be (1) **orthographical** (Character redundant – *Tā Marbūṭa* "ة"), and (2) **syntactic** (Agreement in gender – no agreement between the gender of the word used here and what is used is usual). The word "الجدار" *'alġidār* [wall (gender: M)] has no feminine form to be able to select the correct form for agreement, so the error is not syntactic in gender agreement, but it is orthographical, as there is a redundant character, which is *Tā Marbūṭa* "ة".

In contrast, this is not the case in the sentence "رسمت لوح" *rasamtu lawḥ* [I've drawn a board] (**correct form**: "لوحة" *lawḥa* [painting]), as the usual sentence is to draw a "painting" not a "board." Given the fact that both "لوح" and "لوحة" are orthographically correct words, the error here is very likely to relate to selecting the form that agrees with what is usually used, so the error here is syntactic, not orthographic.



٥,٣ القاعدة الثالثة: أنواع الأخطاء الأكثر تحديداً تقدم على الأنواع العامة

مع أخذ القاعدة الأولى في الاعتبار، فإنه عندما ينطبق نوعان من الخطأ على كلمة واحدة، فإن الأكثر تحديداً أولى بالاختيار من الخطأ العام مالم يكن هناك سبب واضح لعكس ذلك.

مثال ذلك: الألف المحذوفة في "ذهبوا" (والصحيح "ذهبوا")، يمكن أن يعتبر حرفاً ناقصاً، أو من الخلط في الألف الفارقة (كلاهما خطأ إملائي). ولأن الخطأ في الألف الفارقة أكثر تحديداً؛ لأنه يدل على سياق خاص، فإنه الأكثر ملاءمة.

من الأمثلة كذلك الأنواع النحوية الأربعة المذكورة سابقاً في ٥,٢,٢

5.3 Rule 3: More specific types of errors should be preferred to general types

With taking Rule 1 into consideration, when two categories are applicable to one error, usually the more specific one is the most appropriate, unless there is a clear reason for the general.

For example, the missing 'alif (ا) in "ذهبوا" *dahabū* [they went] (**correct form**: "ذهبوا") can be classified as "missing character," or "confusion" in 'alif Fāriqa (both are orthographical errors). However, confusion in 'alif Fāriqa is a more specific error, as it indicates a particular context, so it is the most appropriate.

An additional example (the four syntactic types) has been mentioned in 5.2.2

ملاحظات إضافية

- هل هو خطأ بالفعل؟

ينصح بالتأكد من أن الخطأ المقصود ترميزه هو خطأ بالفعل ويستحق الترميز، وخصوصاً أخطاء الدلالة والأسلوب لغموضها أحياناً أكثر من غيرها.

Additional notes

- Is it a real error?

It is advisable to ensure that the word/phrase intended to be tagged is a real error and requires a tag, particularly semantic and stylistic errors, as they may be more ambiguous than other categories.

تأكد دائماً من اختيارك للرمز الأكثر ملاءمة

ينصح بعد ترميز أي خطأ بالمرور على جدول الأخطاء للتأكد من عدم وجود تداخل بين النوع المحدد والأنواع أخرى.

- Always ensure you select the most appropriate tag

It is advisable after tagging each error to go through the tagset to ensure that there is no overlap between the selected error type and the other types.

6. المُرْمِز The annotator

An annotator needs three qualities: 1. To be a specialist in the Arabic language (with at least his first degree in Arabic linguistics) 2. To have a good knowledge of the rules of Arabic (e.g., orthography, morphology, syntax, punctuation, etc.) 3. To have at least some experience in correcting errors made by students of Arabic, which can be helpful in identifying category and type of error.	ينبغي أن تتوفر فيمن يقوم بالترميز ثلاثة شروط: ١. أن يكون متخصصاً في اللغة العربية (أن يكون على الأقل قد حصل على الدرجة الجامعية في الدراسات اللغوية العربية). ٢. أن يُلمَّ بقواعد اللغة العربية شكل جيد (مثل القواعد الإملائية، والصرفية، والنحوية، وعلامات الترقيم، ونحوها). ٣. أن تكون لديه خبرة بسيطة على الأقل في تصحيح الأخطاء لدى طلاب اللغة العربية تساعده على تحديد نوع الخطأ ولأي مجال ينتمي.
---	---

8. المراجع References

- Alfaifi, A. and Atwell, E. (2012). المدونات اللغوية لمتعلمي اللغة العربية: نظامٌ لتصنيف وترميز الأخطاء. "Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors". In: the proceedings of the 8th International Computing Conference in Arabic (ICCA 2012), Cairo, Egypt.
- Alfaifi, A., Atwell, E. and Abuhakema, G. (2013). Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In: the proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology. Darmstadt, Germany.
- Atwell ES. (1982). *LOB Corpus Tagging Project Manual Postedit Handbook*. University of Lancaster.
- Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. and Thewissen, J. (2005). *Error tagging manual (Version 1.2)*.

Deutsches Institut für Normung. (2011). *Information und Dokumentation - Umschrift des arabischen Alphabets für die Sprachen Arabisch, Osmanisch-Türkisch, Persisch, Kurdisch, Urdu und Paschtu* [DIN 31635 information and documentation - Romanization of the Arabic alphabet for Arabic, Ottoman-Turkish, Persian, Kurdish, Urdu and Pushto]. Retrieved from <http://www.nabd.din.de/cmd?artid=140593750&bcrumblevel=1&contextid=nabd&subcommitteeid=54749615&level=tpl-art-detailansicht&committeeid=54738855&languageid=en>

Appendix F

The DIN 31635 Standard for the Transliteration of the Arabic Alphabet

No.	Arabic letter shape	DIN 31635
1.	أ	ʾ
2.	ب	b
3.	ت	t
4.	ث	ṭ
5.	ج	ǧ
6.	ح	ḥ
7.	خ	b
8.	د	d
9.	ذ	ḏ
10.	ر	r
11.	ز	z
12.	س	s
13.	ش	š
14.	ص	ṣ
15.	ض	ḏ
16.	ط	ṭ
17.	ظ	ẓ
18.	ع	ʿ
19.	غ	ǧ
20.	ف	f
21.	ق	q
22.	ك	k
23.	ل	l
24.	م	m
25.	ن	n
26.	ه	h
27.	و	w
28.	ي	y
29.	ـَ (Short Vowel)	a
30.	ـِ (Short Vowel)	u
31.	ـِـ (Short Vowel)	i
32.	ا (Long Vowel)	ā
33.	و (Long Vowel)	ū
34.	ي (Long Vowel)	ī

Appendix G

Extended Code of the ALC Search Function

```
function ajaxSearch()
{
    $('#ajaxLoaderDiv').show();
    $('#show_title_full').hide();

    if($("#input[name='search_type']:checked").val())
    {
        var search_type =1;

    }else{

        var search_type =0;

    }

    $('#search_type_p').val(search_type);
    $('#search_type_d').val(search_type);

    var search_txt = $("#search_txt").val();

    $('#search_string').val(search_txt);
    $('#search_string_p').val(search_txt);

    var fromAge = $("#fromAge").val();
    var toAge = $("#toAge").val();

    if(fromAge !='' || toAge !='')
    {
        $("#ageRestriction").prop('checked', false);
    }
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
else
{
    $("#ageRestriction").prop('checked',true);
}

var fromLangSpok = $("#fromLangSpok").val();
var toLangSpok = $("#toLangSpok").val();

if(fromLangSpok !='' || toLangSpok !='')
{
    $("#numlangRestriction").prop('checked',false);
}
else
{
    $("#numlangRestriction").prop('checked',true);
}

var fromYearLearnAr = $("#fromYearLearnAr").val();
var toYearLearnAr = $("#toYearLearnAr").val();

if(fromYearLearnAr !='' || toYearLearnAr !='')
{
    $("#numYlearRestriction").prop('checked',false);
}
else
{
    $("#numYlearRestriction").prop('checked',true);
}

var fromYearSpentAr = $("#fromYearSpentAr").val();
var toYearSpentAr = $("#toYearSpentAr").val();
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
if(fromYearSpentAr !='' || toYearSpentAr !='')
{
    $("#sepntRestriction").prop('checked',false);
}
else
{
    $("#sepntRestriction").prop('checked',true);
}

var fromText = $("#fromText").val();
var toText = $("#toText").val();

if(fromText !='' || toText !='')
{
    $("#texLeRestriction").prop('checked',false);
}
else
{
    $("#texLeRestriction").prop('checked',true);
}

$('#fromage_d').val(fromAge);
$('#toage_d').val(toAge);
$('#fromLangSpok_d').val(fromLangSpok);
$('#toLangSpok_d').val(toLangSpok);
$('#fromYearLearnAr_d').val(fromYearLearnAr);
$('#toYearLearnAr_d').val(toYearLearnAr);
$('#fromYearSpentAr_d').val(fromYearSpentAr);
$('#toYearSpentAr_d').val(toYearSpentAr);
$('#fromText_d').val(fromText);
$('#toText_d').val(toText);
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
    $('#fromage_p').val(fromAge);
$('#toage_p').val(toAge);
    $('#fromLangSpok_p').val(fromLangSpok);
    $('#toLangSpok_p').val(toLangSpok);
    $('#fromYearLearnAr_p').val(fromYearLearnAr);
    $('#toYearLearnAr_p').val(toYearLearnAr);
    $('#fromYearSpentAr_p').val(fromYearSpentAr);
    $('#toYearSpentAr_p').val(toYearSpentAr);
    $('#fromText_p').val(fromText);
    $('#toText_p').val(toText);

    $('#fromage_dt').val(fromAge);
$('#toage_dt').val(toAge);
    $('#fromLangSpok_dt').val(fromLangSpok);
    $('#toLangSpok_dt').val(toLangSpok);
    $('#fromYearLearnAr_dt').val(fromYearLearnAr);
    $('#toYearLearnAr_dt').val(toYearLearnAr);
    $('#fromYearSpentAr_dt').val(fromYearSpentAr);
    $('#toYearSpentAr_dt').val(toYearSpentAr);
    $('#fromText_dt').val(fromText);
    $('#toText_dt').val(toText);

var Timing= $("input[name='Timing']:checked").val();

if(Timing == 1 || Timing == 0 )
{
    $('#timRestriction').attr('checked',false);
}
else
{
    $('#timRestriction').attr('checked',true);
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
}
$('#Timing_d').val(Timing);
$('#Timing_p').val(Timing);

$('#Timing_dt').val(Timing);

var refUse= $("input[name='refUse']:checked").val();

if(refUse == 1 || refUse == 0 )
{
    $('#refUseRestriction').attr('checked',false);
}
else
{
    $('#refUseRestriction').attr('checked',true);
}

$('#refUse_d').val(refUse);
$('#refUse_p').val(refUse);

$('#refUse_dt').val(refUse);

var grBookUse= $("input[name='grBookUse']:checked").val();

if(grBookUse == 1 || grBookUse == 0 )
{
    $('#geRestriction').attr('checked',false);
}
else
{
    $('#geRestriction').attr('checked',true);
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
    }

    $('#grBookUse_d').val(grBookUse);
    $('#grBookUse_p').val(grBookUse);

    $('#grBookUse_dt').val(grBookUse);

    var monoDict= $("input[name='monoDict']:checked").val();

    if(monoDict == 1 || monoDict == 0 )
    {
        $('#monoRestriction').attr('checked',false);
    }
    else
    {
        $('#monoRestriction').attr('checked',true);
    }

    $('#monoDict_d').val(monoDict);
    $('#monoDict_p').val(monoDict);

    $('#monoDict_dt').val(monoDict);

    var bilDict= $("input[name='bilDict']:checked").val();

    if(bilDict == 1 || bilDict == 0 )
    {
        $('#bilRestriction').attr('checked',false);
    }
    else
    {
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
        $("#bilRestriction").attr('checked',true);

    }

    $('#bilDict_d').val(bilDict);
    $('#bilDict_p').val(bilDict);

    $('#bilDict_dt').val(bilDict);

    var othRefUse= $("input[name='othRefUse']:checked").val();

    if(othRefUse == 1  || othRefUse == 0 )
    {
        $("#othrefRestriction").attr('checked',false);

    }
    else
    {
        $("#othrefRestriction").attr('checked',true);

    }

    $('#othRefUse_d').val(othRefUse);
    $('#othRefUse_p').val(othRefUse);

    $('#othRefUse_dt').val(othRefUse);

        var gender1= $("input[name='gender[]']");
        var gender = new Array();

    for (var i=0, iLen=gender1.length; i<iLen; i++) {
    if (gender1[i].checked) {
        gender.push(gender1[i].value);
    }
    }

    if(gender !='')
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
{
    $("#genderRestriction").attr('checked', false);
}
else
{
    $("#genderRestriction").attr('checked', true);
}

$('#gender_d').val (gender);
$('#gender_p').val (gender);

$('#gender_dt').val (gender);

var nationality1= $("input[name='nationality[]']");
var nationality = new Array();

for (var i=0, iLen=nationality1.length; i<iLen; i++) {
if (nationality1[i].checked) {
    nationality.push(nationality1[i].value);
}
}

if(nationality !='')
{
    $("#natRestriction").attr('checked', false);
}
else
{
    $("#natRestriction").attr('checked', true);
}

$('#nationality_d').val (nationality);
$('#nationality_p').val (nationality);
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
$('#nationality_dt').val(nationality);

    var mother1= $("input[name='mother[]']");
    var mother = new Array();

for (var i=0, iLen=mother1.length; i<iLen; i++) {
if (mother1[i].checked) {
    mother.push(mother1[i].value);
}
}

if(mother !='')
{
    $('#motRestriction').attr('checked',false);
}
else
{
    $('#motRestriction').attr('checked',true);
}

$('#mother_d').val(mother);
$('#mother_p').val(mother);

$('#mother_dt').val(mother);

    var nativeness1= $("input[name='nativeness[]']");
    var nativeness = new Array();

for (var i=0, iLen=nativeness1.length; i<iLen; i++) {
if (nativeness1[i].checked) {
    nativeness.push(nativeness1[i].value);
}
}
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
if(nativeness !='')
{
    $("#nativRestriction").attr('checked',false);
}
else
{
    $("#nativRestriction").attr('checked',true);
}

$('#nativeness_d').val(nativeness);
$('#nativeness_p').val(nativeness);

$('#nativeness_dt').val(nativeness);

var genLevEdu1= $("input[name='genLevEdu[]']");
var genLevEdu = new Array();

for (var i=0, iLen=genLevEdu1.length; i<iLen; i++) {
if (genLevEdu1[i].checked) {
    genLevEdu.push(genLevEdu1[i].value);
}
}

if(genLevEdu !='')
{
    $("#genLevEdRestriction").attr('checked',false);
}
else
{
    $("#genLevEdRestriction").attr('checked',true);
}
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
$('#genLevEdu_d').val(genLevEdu);
$('#genLevEdu_p').val(genLevEdu);

$('#genLevEdu_dt').val(genLevEdu);

    var levStudy1= $("input[name='levStudy[]']");
    var levStudy = new Array();

for (var i=0, iLen=levStudy1.length; i<iLen; i++) {
if (levStudy1[i].checked) {
    levStudy.push(levStudy1[i].value);
}
}

if(levStudy !='')
{
    $("#levStuRestriction").attr('checked',false);
}
else
{
    $("#levStuRestriction").attr('checked',true);
}

$('#levStudy_d').val(levStudy);
$('#levStudy_p').val(levStudy);

$('#levStudy_dt').val(levStudy);

    var yearSem1= $("input[name='yearSem[]']");
    var yearSem = new Array();

for (var i=0, iLen=yearSem1.length; i<iLen; i++) {
if (yearSem1[i].checked) {
    yearSem.push(yearSem1[i].value);
}
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
    }

    if(yearSem != '')
    {
        $("#yeSemRestriction").attr('checked',false);
    }
    else
    {
        $("#yeSemRestriction").attr('checked',true);
    }

    $('#yearSem_d').val(yearSem);
    $('#yearSem_p').val(yearSem);

    $('#yearSem_dt').val(yearSem);

    var eduInstil= $("input[name='eduInsti[]']");
    var eduInsti = new Array();

    for (var i=0, iLen=eduInstil.length; i<iLen; i++) {
        if (eduInstil[i].checked) {
            eduInsti.push(eduInstil[i].value);
        }
    }

    if(eduInsti != '')
    {
        $("#eduInsRestriction").attr('checked',false);
    }
    else
    {
        $("#eduInsRestriction").attr('checked',true);
    }
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
$('#eduInsti_d').val(eduInsti);
$('#eduInsti_p').val(eduInsti);

$('#eduInsti_dt').val(eduInsti);

    var textGenre1= $("input[name='textGenre[]']");
    var textGenre = new Array();

for (var i=0, iLen=textGenre1.length; i<iLen; i++) {
if (textGenre1[i].checked) {
    textGenre.push(textGenre1[i].value);
}
}

if(textGenre !='')
{
    $("#texGenRestriction").attr('checked',false);
}
else
{
    $("#texGenRestriction").attr('checked',true);
}

$('#textGenre_d').val(textGenre);
$('#textGenre_p').val(textGenre);

$('#textGenre_dt').val(textGenre);

    var placeWritel= $("input[name='placeWrite[]']");
    var placeWrite = new Array();

for (var i=0, iLen=placeWritel.length; i<iLen; i++) {
if (placeWritel[i].checked) {
    placeWrite.push(placeWritel[i].value);
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
}  
}  
  
if(placeWrite !='')  
{  
    $("#plaWriteRestriction").attr('checked',false);  
  
}  
else  
{  
    $("#plaWriteRestriction").attr('checked',true);  
  
}  
  
$('#placeWrite_d').val(placeWrite);  
$('#placeWrite_p').val(placeWrite);  
  
$('#placeWrite_dt').val(placeWrite);  
  
    var yearWrite1= $("input[name='yearWrite[]']");  
    var yearWrite = new Array();  
  
for (var i=0, iLen=yearWrite1.length; i<iLen; i++) {  
    if (yearWrite1[i].checked) {  
        yearWrite.push(yearWrite1[i].value);  
    }  
}  
  
if(yearWrite !='')  
{  
    $("#yeWriRestriction").attr('checked',false);  
  
}  
else  
{  
    $("#yeWriRestriction").attr('checked',true);  
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
    }

    $('#yearWrite_d').val(yearWrite);
    $('#yearWrite_p').val(yearWrite);

    $('#yearWrite_dt').val(yearWrite);

        var countWrite1= $("input[name='countWrite[]']");
        var countWrite = new Array();

for (var i=0, iLen=countWrite1.length; i<iLen; i++) {
if (countWrite1[i].checked) {
    countWrite.push(countWrite1[i].value);
}
}

if(countWrite !='')
{
    $("#countRestriction").attr('checked',false);
}
else
{
    $("#countRestriction").attr('checked',true);
}

$('#countWrite_d').val(countWrite);
$('#countWrite_p').val(countWrite);

$('#countWrite_dt').val(countWrite);

        var cityWrite1= $("input[name='cityWrite[]']");
        var cityWrite = new Array();

for (var i=0, iLen=cityWrite1.length; i<iLen; i++) {
if (cityWrite1[i].checked) {
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
        cityWrite.push(cityWrite1[i].value);
    }
}

if(cityWrite != '')
{
    $("#cityRestriction").attr('checked', false);
}
else
{
    $("#cityRestriction").attr('checked', true);
}

$('#cityWrite_d').val(cityWrite);
$('#cityWrite_p').val(cityWrite);

$('#cityWrite_dt').val(cityWrite);

var textModel= $("input[name='textMode[]']");
var textMode = new Array();

for (var i=0, iLen=textModel.length; i<iLen; i++) {
    if (textModel[i].checked) {
        textMode.push(textModel[i].value);
    }
}

if(textMode != '')
{
    $("#texMRestriction").attr('checked', false);
}
else
{
    $("#texMRestriction").attr('checked', true);
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
    }

    $('#textMode_d').val(textMode);
    $('#textMode_p').val(textMode);

    $('#textMode_dt').val(textMode);

        var textMedium1= $("input[name='textMedium[]']");
        var textMedium = new Array();

for (var i=0, iLen=textMedium1.length; i<iLen; i++) {
if (textMedium1[i].checked) {
    textMedium.push(textMedium1[i].value);
}
}

if(textMedium !='')
{
    $("#texMedRestriction").attr('checked',false);
}
else
{
    $("#texMedRestriction").attr('checked',true);
}

$('#textMedium_d').val(textMedium);
$('#textMedium_p').val(textMedium);

$('#textMedium_dt').val(textMedium);

var dataString =
'search_txt1='+search_txt+'&fromAge='+fromAge+'&toAge='+toAge+'&gender='+gender+'&nationality='+nationality+'&mother='+mother+'&nativeness='+nativeness + '&fromLangSpok=' +fromLangSpok +
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
'&toLangSpok='+toLangSpok +'&fromYearLearnAr='
+fromYearLearnAr+'&toYearLearnAr='+toYearLearnAr+'&fromYearSpentAr='
+fromYearSpentAr+'&toYearSpentAr='+toYearSpentAr+'&genLevEdu='+genLe
vEdu+'&levStudy='+levStudy+'&yearSem='+yearSem+'&eduInsti='+eduInsti
+'&textGenre='+textGenre+'&placeWrite='+placeWrite+'&yearWrite='+yea
rWrite+'&countWrite='+countWrite+'&cityWrite='+cityWrite+'&Timing='+
Timing+'&refUse='+refUse+'&grBookUse='+grBookUse+'&monoDict='+monoDi
ct+'&bilDict='+bilDict+'&othRefUse='+othRefUse+'&textMode='+textMode
+'&textMedium='+textMedium+'&fromText='+fromText+'&toText='+toText+'
&search_type='+search_type;

$.ajax({
  type: "POST",
  url: "<?php echo base_url(); ?>en/ajaxTextSearch",
  data: dataString,
  dataType: 'json',
  success: function(response)
  {

      var show_data='<table width="100%" border="0"
cellspacing="0" cellpadding="0" class="tblRes1">'+
          '<tr>'+
          '<th>Text ID</th>'+
          '<th>Concordance</th>'+
          '</tr>'+
          '<tr>'+
          '<td colspan="2" style="border-right:0px;
padding:0px;">'+
          '<table width="100%" border="0"
cellspacing="0" cellpadding="0">';

      if(response != null)
      {

          if(response['title'] != '')
          {
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
        for(i=0;i<response['title'].length; i++)
        {

                show_data += response['title'][i];

        }

        show_data += '</table>'+

                                '</td>'+
                                '</tr>'+
                                '</table>';

        $('#search_data').html(show_data);
        $('#print_id').show();
        $('#download_id').show();
    }

    else
    {
        show_data += '<tr>'+

                                '<td        colspan="2"
align="center">No Records Here</td>'+

                                '</tr>';
        show_data += '</table>'+

                                '</td>'+
                                '</tr>'+
                                '</table>';

        $('#search_data').html(show_data);
        $('#print_id').hide();
        $('#download_id').hide();
    }
}
```

Appendix G – Sample Code of the Search Function of the ALC Search Tool

```
                $('#paginationBx').html('<div  
id="test">'+response["pagination"]+'</div>');  
  
                $('#search_rows').html(response['total_rows']);  
  
                $('#search_rows2').html(response['total_rows']);  
  
                $('#no_of_rows').html(response['no_of_results']);  
  
                $('#ajaxLoaderDiv').hide();  
  
                ajaxSearch_paging();  
            }  
            else  
            {  
  
                location.reload();  
  
            }  
        }  
    });  
}
```

Figure G.1: Extended Code of the Search Function of the ALC Search Tool

References

- Abel, A., Glaznieks, A., Nicolas, L., & Stemle, E. W. (2014). KoKo: An L1 learner corpus for German. In: *Proceedings of the LREC 2014, International Conference on Language Resources and Evaluation* (pp. 2414–2421). Reykjavik, Iceland: European Language Resources Association.
- Abuhakema, G., Feldman, A., & Fitzpatrick, E. (2008). Annotating an Arabic learner corpus for error. In: *Proceedings of the LREC 2008, International Conference on Language Resources and Evaluation* (pp. 1347–1350). Marrakech, Morocco: European Language Resources Association.
- Abuhakema, G., Feldman, A., & Fitzpatrick, E. (2009). ARIDA: An Arabic interlanguage database and its applications: A pilot study. *Journal of the National Council of Less Commonly Taught Languages*, 7, 161–184.
- Adobe Systems Incorporated. (2006). *PDF reference, 6th ed., Adobe Portable Document Format, version 1.23*. Retrieved 12 March 2014 from http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf
- Ågren, M. (2009). *The Lund CEFLE corpus (Corpus Écrit de Français Langue Étrangère)*. Retrieved 17 September 2012 from <http://projekt.ht.lu.se/cefle/information>
- Alaqueeli, A. S. (1995). *تحليل الأخطاء في بعض أنماط الجملة الفعلية للغة العربية في الأداء الكتابي لدى دارسي المستوى المتقدم* [Error analysis in some verbal sentence patterns of Arabic in writing production of advanced-level learners] (MA thesis). Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
- Alateeq, Z. M. (1992). *تحليل الأخطاء الدلالية لدى دارسي اللغة العربية من غير الناطقين بها في مادة التعبير الكتابي* [Semantic errors analysis of non-native Arabic learners in writing] (MA thesis). Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
- Alfaifi, A. (2011). *The attitude of ASL learners in Saudi Arabia towards printed and electronic dictionaries*. (MA thesis), University of Essex.

References

- Alfaifi, A. and Atwell, E. (2012). المدونات اللغوية لمتعلمي اللغة العربية: نظامٌ لتصنيف وترميز الأخطاء اللغوية "Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors" (in Arabic). In proceedings of *the 8th International Computing Conference in Arabic (ICCA 2012)*, 26 - 28 January 2012, Cairo, Egypt.
- Alfaifi, A. and Atwell, E. (2013). Arabic Learner Corpus v1: A New Resource for Arabic Language Research. In proceedings of *the Second Workshop on Arabic Corpus Linguistics (WACL 2)*, Lancaster University, UK.
- Alfaifi, A. and Atwell, E. (2013). Arabic Learner Corpus: Texts Transcription and Files Format. In proceedings of *the International Conference on Corpus Linguistics (CORPORA 2013)*, St. Petersburg, Russia.
- Alfaifi, A. and Atwell, E. (2013). Potential Uses of the Arabic Learner Corpus. In proceedings of *the Leeds Language, Linguistics and Translation PGR Conference 2013*. Leeds, UK.
- Alfaifi, A., Atwell, E. and Abuhakema, G. (2013) Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In: *Language Processing and Knowledge in the Web, Lecture Notes in Computer Science. 25th International Conference (GSCL 2013)*, 25-27 September 2013, Darmstadt, Germany. Springer, (9) 14 - 22.
- Alfaifi, A. and Atwell, E. (2014). An Evaluation of the Arabic Error Tagset v2. *The American Association for Corpus Linguistics conference (AACL 2014)*. 26-28 September 2014, Flagstaff, USA.
- Alfaifi, A. and Atwell, E. (2014). Arabic Learner Corpus and Its Potential Role in Teaching Arabic to Non-Native Speakers. *The 7th Biennial IVACS conference*, 19 - 21 Jun 2014. Newcastle, UK.
- Alfaifi, A. and Atwell, E. (2014). Arabic Learner Corpus: A New Resource for Arabic Language Research. In the proceedings of *the 7th Saudi Students Conference*, 1-2 February 2014, Edinburgh, UK.

References

- Alfaifi, A. and Atwell, E. (2014). Tools for Searching and Analysing Arabic Corpora: an Evaluation Study. *BAAL / Cambridge University Press Applied Linguistics*, 14 Jun 2014. Leeds Metropolitan University, UK.
- Alfaifi, A., Atwell, E. and Ibraheem, H. (2014). Arabic Learner Corpus (ALC) v2: A New Written and Spoken Corpus of Arabic Learners. In Ishikawa, Shin'ichiro (Ed.), *Learner Corpus Studies in Asia and the World, Papers from LCSAW2014* (Vol. 2, pp. 77–89), School of Language & Communication. Kobe University, Japan.
- Alfaifi, A. and Atwell, E. (2015). Computer-Aided Error Annotation A New Tool for Annotating Arabic Error. *The 8th Saudi Students Conference*, 31 January – 1 February 2015, London, UK.
- Alfaifi, A. (2015). Learner Corpora. In: Alosaimi, S, (Ed.) المدونات اللغوية: بناؤها وطرائق “*Arabic Corpus: Development and Analysis Approaches*” (in Arabic). King Abdullah bin Abdulaziz International Center for Arabic Language Service, KSA.
- Alfaifi, A. and Atwell, E. (accepted). Comparative Evaluation of Tools for Arabic Corpora Search and Analysis. *International Journal of Speech Technology (IJST)*.
- Alfaifi, A., Atwell, E. and Brierley, C. (under review). Learner Corpora: Present and Future, design criteria for creating a new learner corpus. *Applied Linguistics*.
- Alhamad, M. M. (1994). تحليل أخطاء التعبير الكتابي لدى المستوى المتقدم من دارسي العربية غير الناطقين بها في جامعة الملك سعود [Writing errors analysis of advanced-level Arabic learners at King Saud University] (MA thesis). Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
- Alharthi, M. (2015, March). *Applications of using Arabic corpus in teaching Arabic as a second language workshop*. Workshop on Teaching Arabic. Princess Nora Bint Abdulrahman University, Riyadh, Saudi Arabia.

References

- Alkanhal, M., Al-Badrashiny, M., Alghamdi, M., & Al-Qabbany, A. (2012). Automatic stochastic Arabic spelling correction with emphasis on space insertions and deletions. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7), 2111–2122.
- Alkhalifa, H., & Alajlan, A. A. (2010). Automatic readability measurements of the Arabic text: An exploratory study. *The Arabian Journal for Science and Engineering*, 35(2C), 103–124.
- Alosaili, A. I. (1985). *دراسة الأخطاء الشائعة في الكلام لدى طلاب اللغة العربية الناطقين بلغات أخرى* [Common errors in speech production of non-native Arabic learners] (MA thesis). Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia.
- Alshaiban, A. (in preparation). *النضج النحوي لدى متعلمي العربية لغة ثانية* [Grammatical competence of Arabic learners as a second language] (Ph.D. thesis). Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh.
- Alshehri, D. (in preparation). *الترابط النحوي والتماسك النصي في المدونة اللغوية لمتعلمي اللغة العربية لغة ثانية* [Grammatical coherence and textual cohesion in the Arabic learner corpus as a second language] (Ph.D. thesis). Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh.
- Altamimi, A., Jaradat, M., Aljarrah, N., & Ghanim, S. (2014). AARI: Automatic Arabic readability index. *The International Arab Journal of Information Technology*, 11(4), 370–378.
- Althubaity, A., & Al-Mazrua, M. (2014). Khawas: Arabic Corpora Processing Tool USER GUIDE. Retrieved 6 April 2014 from <http://sourceforge.net/projects/kacst-acptool/files/?source=navbar>
- Althubaity, A., Khan, M., Al-Mazrua, M., & Almoussa, M. (2013). New language resources for Arabic: Corpus containing more than two million words and a corpus processing tool. In: *Proceedings of the IALP International Conference on Asian Language Processing, Urumqi* (pp. 67–70). Urumqi, China: IEEE.

References

- Althubaity, A., Khan, M., Al-Mazrua, M., & Almoussa, M. (2014). *KACST Arabic Corpora processing tool Khawas*. Retrieved 8 April 2014 from <http://kacst-acptool.sourceforge.net/>
- Althubaity, A. O. (2014). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*: 1–31.
- American Council on the Teaching of Foreign Languages. (2012). *The ACTFL proficiency guidelines*. Retrieved 22 March 2014 from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf
- Andreu, M. Á., Astor, A., Boquera, M., Macdonald, P., Montero, B., & Pérez, C. (2010). Analysing EFL learner output in the MiLC project: An error it's*, but which tag? In M. C. Campoy, B. Belles-Fortunato, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 167-179). London, UK: Continuum.
- AntConc-discussion. (2013). AntConc and Arabic Texts. Retrieved 20 September 2014 from <https://groups.google.com/d/msg/antconc/7v3TrtW2LiE/DySK9GIzPooJ>
- Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In: *Proceedings of the Professional Communication Conference* (pp. 729–737). Limerick, Ireland: IEEE.
- Anthony, L. (2014a). *AntConc* (Version 3.4.0) [Computer software]. Tokyo, Japan: Waseda University. Retrieved 11 January 2015 from <http://www.antlab.sci.waseda.ac.jp/>
- Anthony, L. (2014b). *AntConc 3.4.2 - Readme*: Tokyo, Japan: Waseda University. Retrieved 11 January 2015 from http://www.laurenceanthony.net/software/antconc341/AntConc_readme.pdf

References

- Arshad, A. (2004). Beyond concordance lines: Using concordances to investigating language development. *Internet Journal of e-Language Learning & Teaching*, 1(1), 43–51.
- Arthern, P. J. (1978). Machine Translation and Computerized Terminology Systems: A Translator's Viewpoint. In B.M. Snell (Ed.) *Translating and the Computer: Proceedings of a Seminar, London, 14th November 1978* (pp. 77–108). Amsterdam: North Holland.
- Arthern, P. J. (1981). Aids Unlimited: The Scope for Machine Aids in a Large Organization. In: *Aslib Proceedings*, (Vol. 33, pp. 309–319).
- Atwell, E.S.; Al-Sulaiti, L., Al-Osaimi, S., & Abu Shawar, B. A. (2004). A review of Arabic corpus analysis tools - un examen d'outils pour l'analyse de corpora Arabes. In B. Bel & I. Marlien (Eds.) *Proceedings of TALN04, XI Conference sur le Traitement Automatique des Langues Naturelles* (Vol. 2, pp. 229–234). Fez, Morocco: ATALA.
- Axelsson, M. W., & Hahn, A. (2001). The use of the progressive in Swedish and German advanced learner English - a corpus-based study. *ICAME Journal*, 25, 5–30.
- Axelsson, M. W., & Berglund, Y. (2002). The Uppsala student English corpus (USE): A multi-faceted resource for research and course development. In L. Borin (Ed.), *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora* (pp. 79–90). Amsterdam, the Netherlands: Rodopi.
- Bailini, S. (2013). SCIL: A Spanish corpus of Italian learners. *Procedia - Social and Behavioral Sciences*, 95, 542–549. doi:10.1016/j.sbspro.2013.10.680
- Banerjee, J., & Franceschina, F. (2012). *Lancaster corpus of academic written English (LANCAWE)*. Retrieved 13 September 2012 from <http://www.ling.lancs.ac.uk/activities/294/>

References

- Bański, P., & Gozdawa-Gołębiowski, R. (2010). Foreign language examination corpus for L2-learning studies. In: *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora* (pp. 56–64). Valletta, Malta: LREC.
- Barbera, M., & Corino, E. (2003). *VALICO - An Italian learner corpus*. Retrieved 4 October 2012 from http://www.valico.org/index_en.html
- Bartalesi, L. V., Moretti, G., & Sprugnoli, R. (2012). *CAT: The CELCT annotation tool*. In: *Proceedings of the LREC 2012, International Conference on Language Resources and Evaluation* (pp. 333–338). Istanbul, Turkey: European Language Resources Association.
- Bartning, I. (2011). *The InterFra project*. Retrieved 4 September 2012 from <http://www.fraitaklass.su.se/english/interfra>
- Belz, J., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: The case of German modal particles. *The Canadian Modern Language Review*, 62(1), 17–48.
- Berber Sardinha, T. (2002). *The Br-ICLE corpus (Brazilian component of ICLE)*. Retrieved 12 August 2012 from <http://www2.lael.pucsp.br/corpora/bricle>
- Berglund, Y., & Axelsson, M. W. (2012). *Uppsala student English corpus (USE)*. Retrieved 24 August 2012 from http://www.engelska.uu.se/Forskning/engelsk_sprakvetenskap/Forskningsomraden/Electronic_Resource_Projects/USE-Corpus/
- Bilbow, G., Greaves, C., Lee, S., Lan, L., & Cheung, R. (2004). *PolyU learner English corpus (PLEC)*. Retrieved 8 August 2012 from <http://langbank.engl.polyu.edu.hk/index1.html>
- Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). *ETS corpus of non-native written English*. Retrieved 15 December 2014 from <https://catalog.ldc.upenn.edu/LDC2014T06>

References

- Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* (Version 5.4.04) [Computer software]. Retrieved 15 January 2015 from <http://www.praat.or>
- Botley, S. P., & Dillah, D. (2007). Investigating spelling errors in a Malaysian learner corpus. *Malaysian Journal of ELT Research*, 3, 74–93.
- Botley, S. P. (2012, February). *Error tagging a Malaysian learner corpus: Pitfalls and rewards*. Paper presented at the First Asia Pacific Corpus Linguistics Conference. Abstract retrieved 20 October 2013 from <http://corpling.com/conf/prog.html>
- Branbrook, G. (1996). *Language and computers*. Edinburgh, Scotland: Edinburgh University Press.
- Buckwalter, T., & Parkinson, D. (2013). Modern lexicography. In J. Owens (Ed.), *The Oxford handbook of Arabic linguistics* (pp. 539–560). New York, NY: Oxford University Press.
- Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 30–46). Oxford, UK: Oxbow Books.
- Burnard, L. (2007). *Reference guide for the British national corpus (XML edition)*. Retrieved 6 September 2014 from <http://www.natcorp.ox.ac.uk/docs/URG/>
- Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for ‘opportunity of use’ in learner corpora. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and cross-linguistic perspectives in learner corpus research* (pp. 187–204). Amsterdam, the Netherlands: Benjamins.
- Callies, M., & Zaytseva, E. (2011a). The corpus of academic learner English (CALE): A new resource for the study of lexico-grammatical variation in advanced learner varieties. In H. Hedeland, T. Schmidt, & K. Wörner (Eds.), *Multilingual resources and multilingual applications (Hamburg working*

References

- papers in multilingualism b 96*) (pp. 51–56). Hamburg, Germany: Collaborative Research Centre.
- Callies, M., & Zaytseva, E. (2011b). *English for advanced learners: Linguists at Mainz University examine obstacles to native-like proficiency in foreign language acquisition*. Retrieved 7 August 2012 from <http://www.uni-mainz.de/eng/14369.php>
- Callies, M., Zaytseva, E., Kinne, A., Sperling, T., & Wiemeyer, L. (2012). *The corpus of academic learner English*. Retrieved 7 August 2012 from <http://www-user.uni-bremen.de/~callies/ALV.htm>
- Cambridge University. (2012). *Cambridge learner corpus*. Retrieved 7 August 2012 from http://www.cambridge.org/gb/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=en_GB
- Centre for English Corpus Linguistics. (2010). *Learner corpus research*. Retrieved 18 March 2013 from <http://www.uclouvain.be/en-169937.html>
- Chambers, F., & Richards, B. (1995). The “free conversation” and the assessment of oral proficiency. *Language Learning, 11*, 6–10.
- Chen, J. N. (2000). Adaptive word sense disambiguation using lexical knowledge in machine-readable dictionary. *Computational Linguistics and Chinese Language Processing, 5*(2), 1–42.
- Cheng, C.-C., Lu, H.-C., Huang, S.-M., Hsieh, C.-Y., Liu, G.-Z., Lu, W.-H., & Yang, K.-R. (2012). *The corpus of Taiwanese learners of Spanish*. Retrieved 5 October 2012 from <http://corpora.flld.ncku.edu.tw/>
- Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review, 59*(3), 393–423.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

References

- Colantoni, L., & Steele, J. (2004). *The University of Toronto romance phonetics database (RPD)*. Retrieved 16 December 2014 from <http://r1.chass.utoronto.ca/rpd/>
- Connor, U. (2012). *Indianapolis business learner corpus*. Retrieved 4 September 2012 from http://www.liberalarts.iupui.edu/icic/research/indianapolis_business_learner_corpus
- Connor, U., Davis, K. W., & De Rycker, T. (1995). Correctness and clarity in applying for overseas jobs: A cross-cultural analysis of U.S. and Flemish applications. *Text*, 15(4), 457–476.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. New York, NY: Cambridge University Press.
- Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., ... Peters, W. (2011). *Text processing with GATE (Version 6)*.
- Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2), e1002854. doi:1002810.1001371/journal.pcbi.1002854
- Dagneaux, E., Denness, S., Granger, S., & Meunier, F. (1996). *Error tagging manual (Version 1.1)*.
- Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. In: *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22–31). Atlanta, GA: Association for Computational Linguistics.
- Dalziel, F., & Helm, F. (2008). CMC and learner corpora: From focus on interaction to focus on form. *Computer Mediated Communication and Learning*:

References

- Research and Practice*. Retrieved 2 October 2012 from <http://www.eurocall-languages.org/sigs/cmc.html>
- Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, 10(3), 307–334. doi:10.1075/ijcl.10.3.02dav
- Delais-Roussarie, E., & Yoo, H. (2010). The COREIL corpus: A learner corpus designed for studying phrasal phonology and intonation. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 1-3 May 2010* (pp. 100–105). Poznan, Poland: Adam Mickiewicz University.
- Detey, S., & Kawaguchi, Y. (2008, December). *Interphonologie of French contemporary (IPFC): Automated data collection and Japanese learners*. Paper presented at the PFC Days: Phonology of Contemporary French: variation, interfaces, cognition, Paris, France.
- Deutsches Institut für Normung. (2011). *Information und Dokumentation - Umschrift des arabischen Alphabets für die Sprachen Arabisch, Osmanisch-Türkisch, Persisch, Kurdisch, Urdu und Paschtu* [DIN 31635 information and documentation - Romanization of the Arabic alphabet for Arabic, Ottoman-Turkish, Persian, Kurdish, Urdu and Pushto]. Retrieved 6 June 2012 from <http://www.nabd.din.de/cmd?artid=140593750&bcrumblevel=1&contextid=nabd&subcommitteeid=54749615&level=tpl-art-detailansicht&committeeid=54738855&languageid=en>
- Díaz-Negrillo, A. (2012). Learner corpora: The case of the NOSE corpus. *Systemics, Cybernetics and Informatics*, 10(1), 42-47.
- Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 9–30). Amsterdam, the Netherlands: Benjamins.

References

- Diez-Bedmar, M. B. (2009). Written learner corpora by Spanish students of English: An overview. In P. C. Gómez & A. S. Pére (Eds.), *A survey on corpus-based research. Proceedings of the AELINCO Conference* (pp. 920–933). Murcia, Spain: Asociación Española de Lingüística del Corpus.
- Dominguez, L., Mitchell, R., M., Florence, Tracy-Ventura, N., Arche, M. J., & Boardman, T. (2010). *Spanish learner language oral corpora (SPLLOC 2)*. Retrieved 6 October 2012 from <http://www.splloc.soton.ac.uk/splloc2/index.html>
- Eckart de Castilho, R., Biemann, C., Gurevych, I., & Yimam, S. M. (2014, October). *WebAnno: A flexible, web-based annotation tool for CLARIN*. Paper presented at the CLARIN Annual Conference, Soesterberg, the Netherlands.
- Eileen, F., & Milton, S. S. (2012). *The Montclair electronic language learners' database (MELD)*. Retrieved 16 September 2012 from <http://www.montclair.edu/chss/linguistics/department-research-projects/meld/>
- Eslon, P., Matsak, E., Kippar, O., Metslang, H., Mare, K., Rebas, V., ... Dovgan, E. (2012). *Estonian interlanguage corpus (EIC)*. Retrieved 2 October 2012 from http://evkk.tlu.ee/wwwdata/what_is_evk
- Farra, N., Tomeh, N., Rozovskaya, A., & Habash, N. (2014, June). Generalized character-level spelling error correction. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 161–167). Baltimore, Maryland: Association for Computational Linguistics.
- Farwaneh, S., & Tamimi, M. (2012). *Arabic learners written corpus: A resource for research and learning*. Retrieved 2 September 2012 from <http://l2arabiccorpus.cercll.arizona.edu/?q=homepage>
- Fauth, C., Bonneau, A., Zimmerer, F., Trouvain, J., Andreeva, B., Colotte, V., ... Möbius, B. (2014, May). Designing a bilingual speech corpus for French and German language learners: A two-step process. In: *Proceedings of the LREC 2014, International Conference on Language Resources and Evaluation* (pp.

References

- 1477–1482). Reykjavik, Iceland: European Language Resources Association.
- Fitzpatrick, E., & Seegmiller, M. S. (2001). The Montclair electronic language learner database. In: *Proceedings of the International Conference on Computing and Information Technologies* (pp. 369–375). Montclair, NJ: World Scientific.
- Fitzpatrick, E., & Seegmiller, M. S. (2004). The Montclair electronic language database project. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 223–237). New York, NY: Rodopi publisher.
- Fitzpatrick, E., & Seegmiller, M. S. (2012). *The Montclair electronic language learners' database (MELD)*. Retrieved 16 September 2012 from <http://www.montclair.edu/chss/linguistics/departement-research-projects/meld>
- Forsyth, J. N. (2014). *Automatic readability detection for modern standard Arabic*. (MA thesis). Brigham Young University, Provo, UT.
- Fraunhofer Institute for Integrated Circuits IIS. (2015). *The mp3 history*. Retrieved 17 March 2015 from http://www.mp3-history.com/en/the_story_of_mp3.html
- Gallina, F. (2010). The LIPS corpus (lexicon of spoken Italian by foreigners) and the acquisition of vocabulary by learners of Italian as L2. In: *Proceedings of the Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching* (Vol. 5, pp. 30–50). Lancaster, UK: Lancaster University.
- Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach* (pp. 30–41). London, UK: Longman.

References

- Garside, R. (1996). The robust tagging of unrestricted text: The BNC experience. In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in the honour of Geoffrey Leech* (pp. 167–180). London, UK: Longman.
- Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London, UK: Longman.
- Granger, S. (1993). The international corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis and exploitation* (pp. 57–69). Amsterdam, the Netherlands: Rodopi.
- Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London, UK: Longman.
- Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam, the Netherlands: Benjamins.
- Granger, S. (2003a). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3), 465–480.
- Granger, S. (2003b). The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538–546.
- Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123–145). Amsterdam, the Netherlands: Rodopi.

References

- Granger, S. (2008). Learner corpora. In A. Ludeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 259–275). Berlin, Germany: Walter de Gruyter.
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2010). *International corpus of learner English v2 (ICLE)*. Retrieved 21 June 2012 from <http://www.uclouvain.be/en-277586.html>
- Granger, S., & Dumont, A. (2014). *Learner corpora around the world*. Retrieved 16 December 2014 from <http://www.uclouvain.be/en-cecl-lcworld.html>
- Granger, S., Gilquin, G., & De Cock, S. (2012). *The Louvain international database of spoken English interlanguage (LINDSEI)*. Retrieved 14 September 2012 from <http://www.uclouvain.be/en-cecl-lindsei.html>
- Granger, S., Gilquin, G., & Meunier, F. (2013). *Twenty years of learner corpus research. looking back, moving ahead: Proceedings of the First Learner Corpus Research Conference*. Vol. 1. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.
- Green, S., & Manning, C. (2010). Better Arabic parsing: Baselines, evaluations, and analysis. In: *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 394–402). Stroudsburg, PA: Association for Computational Linguistics.
- Gut, U. (2012). The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In T. Schmidt & K. Worner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 3–23). Amsterdam, the Netherlands: Benjamins.
- Habash, N. (2010). Introduction to Arabic natural language processing. In G. Hirst (Ed.), *Synthesis Lectures on Human Language Technologies*. San Rafael, CA: Morgan and Claypool.
- Hamel, M.-J., & Milicevic, J. (2007). Analyse d'erreurs lexicales d'apprenants du FLS : démarche empirique pour l'élaboration d'un dictionnaire

References

- d'apprentissage [Analysis of lexical errors FSL learners: Empirical approach to the development of a training dictionary]. *Canadian Journal of Applied Linguistics*, 10(1), 25–45.
- Hammarberg, B. (2010). *Introduction to the ASU corpus, a longitudinal oral and written text corpus of adult learners' Swedish with a corresponding part from native Swedes*. Stockholm: Stockholm University, Department of Linguistics.
- Hana, J., Rosen, A., Škodová, S., & Štindlová, B. (2010). Error-tagged learner corpus of Czech. In: *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 11–19). Uppsala, Sweden: Uppsala University.
- Hassan, H., & Daud, N. M. (2011, April). *Corpus analysis of conjunctions: Arabic learners' difficulties with collocations*. Paper presented at the Workshop on Arabic Corpus Linguistics (WACL). Retrieved 25 September 2014 from <http://ucrel.lancs.ac.uk/wacl/slides-HASSAN-DAUD.pdf>
- Hassan, H., & Ghalib, M. (2013). مشروع جمع المدونات النصية الخاصة بالنصوص الأكاديمية في اللغة العربية [Project of collecting texts of academic corpora in Arabic]. *Journal of the Jordan Academy of Arabic*, 85, 57–77.
- Hasselgren, A. (1997). The EVA corpus of Norwegian school English. *ICAME Journal*, 21, 123–124.
- Hasselgren, A. (2007). *The EVA corpora of pupil language*. Retrieved 24 August 2012 from <http://www.hf.uib.no/i/Engelsk/EVA.html>
- Hellwig, B. (2014). *ELAN - Linguistic annotator*. Retrieved 12 December 2014 from <http://www.mpi.nl/corpus/html/elan/>
- Herment, S., Kerfelec, V., Leonarduzzi, L., & Turcsan, G. (2010). *A learners' corpus of reading texts*. Retrieved 13 August 2012 from http://sldr.ortolang.fr/voir_depot.php?id=15&lang=en&sip=1

References

- Heuboeck, A., Holmes, J., & Nesi, H. (2008). *The BAWE corpus manual*. Retrieved 24 July 2012 from http://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/bawe/BAWE.documentation.pdf
- Hilton, H. (2008). *Corpus PAROLE: Parallèle Oral en Langue Étrangère*. Retrieved 18 July 2012 from http://www.umr7023.cnrs.fr/sites/sfl/IMG/pdf/PAROLE_manual.pdf
- Hirst, D., & Tortel, A. (2010). *ANGLISH*. Retrieved 9 August 2012 from http://sldr.org/voir_depot.php?lang=en&id=731&prefix=sldr
- Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 77–116). Amsterdam, the Netherlands: Benjamins.
- Hua, C., Qiufang, W., & Aijun, L. (2008). A learner corpus ESCCL. In: *Proceedings of the Speech Prosody Conference* (pp. 155–158). Campinas, Brazil: ISCA.
- Hutchinson, J. (1996). *Université catholique de Louvain Error Editor (UCLEE) software*. Louvain-la-Neuve, Belgium: Centre for English Corpus Linguistics, Université Catholique de Louvain.
- Ishikawa, S. (2010). *The corpus of English essays written by Asian university students (CEE AUS)*. Retrieved 6 August 2012 from <http://language.sakura.ne.jp/s/ceeause.html>
- Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE corpus exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2), 119–125.
- Izumi, E., Uchimoto, K., & Isahara, H. (2005). Error annotation for corpus of Japanese learner English. In: *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora* (pp. 71–80). Jeju Island, Korea: Springer.

References

- Jantunen, J. (2010). *The international corpus of learner Finnish (ICLFI)*. Retrieved 2 October 2012 from <http://www oulu.fi/hutk/sutvi/oppijankieli/ICLFI/Yleinen/index.html>
- Johns, T., & King, P. (1991). Classroom concordancing. *English Language Research Journal*, 4, 1–13.
- Jucker, A. H., Müller, S., & Smith, S. (2005). *GLBCC (Giessen - Long Beach Chaplin corpus)*. Retrieved 3 September 2012 from <http://ota.oucs.ox.ac.uk/headers/2506.xml>
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Upper Saddle River, NJ: Prentice Hall.
- Kay, M. (1980). *The Proper Place of Men and Machines in Language Translation*. Research Report CSL-80-11. Palo Alto, CA: Xerox PARC. Reprinted in *Machine Translation* 12:3–23 (1997).
- Kennedy, G. (1998). *An introduction to corpus linguistics*. London, UK: Longman.
- Kilgarriff, A. (2014). Sketch Engine [Computer software]. Retrieved 6 April 2014 from <http://www.sketchengine.co.uk/>
- Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In: *Proceedings of the Euralex* (pp. 105–115). Lorient, France: Université de Bretagne-Sud.
- Kilimci, A. (2014). LINDSEI-TR: A new spoken corpus of advanced learners of English. *International Journal of Social Sciences and Education*, 4(2), 401–410.
- Kim, J.-D., Wang, Y., & Nakajima, S. (2013). *The TextAE editor: An embeddable visual editor of text annotation*. Retrieved 22 May 2014 from <http://textae.pubannotation.org>

References

- Kipp, M. (2001). Anvil - A generic annotation tool for multimodal dialogue. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1367–1370). Aalborg, Denmark: ISCA.
- Komachi, M., Nagata, M., & Matsumoto, Y. (2013). *The Lang-8 learner corpora*. Retrieved 15 December 2014 from <http://cl.naist.jp/nldata/lang-8/>
- Kprzemek, P. (2007). *About PICLE (Polish component of ICLE)*. Retrieved 17 September 2012 from http://ifa.amu.edu.pl/~ifaconc/blog/?page_id=60
- Kübler, N. (2007). *The MeLLANGE learner translator corpus (LTC)*. Retrieved 1 October 2012 from <http://mellange.eila.univ-paris-diderot.fr/>
- Kwon, H. (2009). The SNU Korean learner corpus of English: Compilation and application. *영어학연구 English Language and Linguistics*, 28, 203–228.
- Kwon, Y.-E., & Lee, E.-J. (2014). Lexical bundles in the Korean EFL teacher talk corpus: A comparison between non-native and native English teachers. *The Journal of Asia TEFL*, 11(3), 73–103.
- Lan, L. (2002). *Learner corpus of English for business communication*. Retrieved 8 August 2012 from <http://langbank.engl.polyu.edu.hk/index1.html>
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated grammatical error detection for language learners*. San Rafael, CA: Morgan & Claypool.
- Lee, D. J. (2007). *Corpora and the classroom: A computer-aided error analysis of Korean students' writing and the design and evaluation of data-driven learning materials* (Ph.D. thesis). University of Essex, UK.
- Lee, D., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, 18, 181–196.
- Lee, D., & Flowerdew, J. (2012). *A multimedia City University corpus of academic spoken English (CUCASE)*. Retrieved 12 August 2012 from <http://roweb.cityu.edu.hk/2008-2009/project/7002193P.htm>

References

- Lee, S.-H., Jang, S. B., & Seo, S.-K. (2009). Annotation of Korean learner corpora for particle error detection. *CALICO Journal*, 26(3), 529–544.
- Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British national corpus. In: *Proceedings of the 15th International Conference on Computational Linguistics* (Vol. 1, pp. 622–628). Kyoto, Japan: Association for Computational Linguistics.
- Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). London, UK: Longman.
- Lindgrén, S.-A. (2012a). *The BATMAT corpus*. Retrieved 15 December 2014 from <http://www.abo.fi/fakultet/Content/Document/document/31388>
- Lindgrén, S.-A. (2012b). *The LONGLEX project*. Retrieved 15 December 2014 from <http://www.abo.fi/fakultet/Content/Document/document/31388>
- Linguistic Data Consortium. (2008). *Quick rich transcription (QRTR) specification for Arabic broadcast data (Version 3)*. Retrieved 22 January 2013 from <https://catalog ldc.upenn.edu/docs/LDC2013T04/Arabic-XTransQRTR.V3.pdf>
- Littlewood, W. (1984). *Foreign and second language learning: Language acquisition research and its implications for the classroom*. Cambridge, UK: Cambridge University Press.
- Longman Corpus Network. (2012). *The Longman learner corpus*. Retrieved 8 July 2012 from <http://www.pearsonlongman.com/dictionaries/corpus/learners.html>
- Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In Bretones Callejas, C. M., Fernández Sánchez, J. F., Ibáñez Ibáñez, J. R., García Sánchez, M. E., Cortés de los Ríos, M. E., Ramiro, S. S., ... Márquez, B. C. (Eds.), *Applied linguistics now: Understanding language and mind / La Lingüística*

References

- Aplicada Hoy: Comprendiendo el Lenguaje y la Mente* (pp. 197–212). Almería, Spain: Universidad de Almería.
- Lu, H.-C. (2010). An annotated Taiwanese learners' corpus of Spanish, CATE. *Corpus Linguistics and Linguistic Theory*, 6(2), 125–311.
- Lüdeling, A., Briskina, E., Hantschel, J., Krüger, J., Sigrüst, S., & Spieler, U. (2009). *The LeKo corpus*. Retrieved 14 July 2012 from http://linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/lehre/alte_jahrgaenge/ws-2004/hs-phaenomene/pdf/LekoHandbuch.pdf
- Lightbound, P. M. (2005). *An analysis of interlanguage errors in synchronous/asynchronous intercultural communication exchanges* (Ph.D. thesis). Universitat de Valencia, Spain.
- Maden-Weinberger, U. (2013). *CLEG13 version 07-19-2013*. Retrieved 20 July 2012 from http://korpling.german.hu-berlin.de/public/CLEG13/CLEG13_documentation.pdf
- Maijanen, A., & Lammervo, T. (2014). *The Finnish national foreign language certificate corpus (YKI)*. Retrieved 16 December 2014 from http://yki-korpus.jyu.fi/index_eng.html
- Malmasi, S., & Dras, M. (2014). Arabic native language identification. In: *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language* (pp. 180–186). Doha, Qatar: Association for Computational Linguistics.
- Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press.
- Maolalaigh, R. Ó, & Carty, N. (2014a). *Comasan Labhairt ann an Gàidhlig (CLAG)*. Retrieved 16 December 2014 from <http://www.soillse.ac.uk/en/world-leading-resource-launched-for-assessing-and-increasing-gaelic-proficiency>
- Maolalaigh, R. Ó., & Carty, N. (2014b). *Spanish learner oral corpus*. Retrieved 16 December 2014 from http://cartago.llf.uam.es/corele/home_en.html

References

- Martin, M. (2009). *Linguistic basis of the common European framework for L2 English and L2 Finnish (CEFLING)*. Retrieved 16 December 2014 from <https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/paattyneet-hankkeet/cefling/en>
- Martin, M. (2013). *Paths in second language acquisition (TOPLING)*. Retrieved 16 December 2014 from <https://www.jyu.fi/hum/laitokset/kielet/tutkimus/hankkeet/topling/en>
- Mauranen, A. (2007). Investigating English as a lingua franca with a spoken corpus. In M. C. Campoy & M. J. Luzón (Eds.), *Spoken Corpora in Applied Linguistics* (Vol. 51, pp. 33–56). Berlin, Germany: Lang.
- Max Planck Institute for Psycholinguistics. (2012). *The ESF (European Science Foundation second language) database*. Retrieved 7 October 2012 from <http://www.mpi.nl/tg/lapp/esf/esf.html>
- McEnery, A. M. (2003). Corpus linguistics. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 448–463). Oxford, UK: Oxford University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Oxford, UK: Routledge.
- Mendikoetxea, A., Chocano, G., Jiménez, R., Lozano, C., Murcia, S., O'Donnell, M., & Rollinson, P. (2008). *The written corpus of learner English (WriCLE)*. Retrieved 1 October 2012 from <http://web.uam.es/woslac/WriCLE/>
- Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., & Souter, C. (2000). The ISLE corpus of non-native spoken English. In G. Maria (Ed.), *Proceedings of LREC 2000, Language Resources and Evaluation Conference* (Vol. 2, pp. 957–964). Athens, Greece: European Language Resources Association.

References

- Meunier, F., Granger, S., Littré, D., & Paquot, M. (2010). *The LONGDALE (longitudinal database of learner English)*. Retrieved 14 September 2012 from <http://www.uclouvain.be/en-cecl-longdale.html>
- Milton, J., & Nandini, C. (1994). Tagging the interlanguage of Chinese learners of English. In L. Flowerdew & A. K. K. Tong (Eds.), *Entering text* (pp. 127–143). Hong Kong, China: The Hong Kong University of Science and Technology.
- Mitchell, R., Myles, F., Dominguez, L., Marsden, E., Arche, M. J., & Boardman, T. (2008). *Spanish learner language oral corpora (SPLLOC 1)*. Retrieved 6 October 2012 from <http://www.splloc.soton.ac.uk/splloc1/index.html>
- Mitkov, R. (2003). *The Oxford handbook of computational linguistics*. New York, NY: Oxford University Press.
- Monroe, W., Green, S., & Manning, C. D. (2014). Word segmentation of informal Arabic with domain adaptation. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 206–211). Baltimore, Maryland: Association for Computational Linguistics.
- Muehleisen, V. (2007). *The SILS learner corpus of English*. Retrieved 27 September 2012 from <http://www.f.waseda.jp/vicky/learner/index.html>
- Myles, F., & Mitchell, R. (2012). *French learner language oral corpora*. Retrieved 12 September 2012 from <http://www.flloc.soton.ac.uk/index.html>
- Nesi, H. (2008). Corpora & EAP. In LSP: Interfacing language with other realms. In: *Proceedings of the 6th Languages for Specific Purposes International Seminar* (pp. 1–14). Johor Bahru, Malaysia: Universiti Teknologi Malaysia.
- Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–152). Amsterdam, the Netherlands: Benjamins.

References

- Nicholls, D. (2003). The Cambridge learner corpus - error coding and analysis for lexicography and ELT. In: *Proceedings of the Corpus Linguistics 2003 Conference* (Vol. 16, pp. 572–581). Lancaster, UK: Lancaster University.
- Nugues, P. M. (2006). *An introduction to language processing with Perl and Prolog*. Berlin, Germany: Springer-Verlag.
- O'Donnell, M., Murcia, S., García, R., Molina, C., Rollinson, P., MacDonald, P., ... Boquera, M. (2009). Exploring the proficiency of English learners: The TREACLE project. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Fifth Corpus Linguistics*. Liverpool, UK: University of Liverpool.
- O'Donnell, M. (2010, April). *Building learner English proficiency profiles using automatic syntactic analysis*. Paper presented at the AESLA, Vigo, Spain. Retrieved 11 May 2013 from <http://www.uam.es/proyectosinv/treacle/Publications/AESLA10-vigo.pdf>
- O'Donnell, M. B., & Römer, U. (2009a). *From student hard drive to web corpus: The design, compilation, annotation and online distribution of the MICUSP corpus*. A poster presented at ICAME 30, Lancaster University, UK. Retrieved 27 July 2012 from http://micusp.elicorpora.info/files/0000/0199/ICAME_MICUSP_poster_Matt_Ute--finalX.pdf
- O'Donnell, M. B., & Römer, U. (2009b). *Michigan corpus of upper-level student papers*. Retrieved 27 July 2012 from <http://micusp.elicorpora.info/>
- Obeid, O., Zaghouni, W., Mohit, B., Habash, N., Oflazer, K., & Tomeh, N. (2013). A web-based annotation framework for large-scale text correction. In: *Proceedings of the IJCNLP-2013 – Demo Session* (pp. 1–4). Nagoya, Japan: Toyohashi University of Technology.
- Osborne, J., Henderson, A., & Barr, R. (2012). *The Scientext English learner corpus*. Retrieved 26 September 2012 from <http://scientext.msh-alpes.fr/scientext-site-en/?article19>

References

- Paquot, M., de Cock, S., Granger, S., & Meunier, F. (2009). *The Varieties of English for Specific Purposes dAtabase (VESPA) learner corpus*. Retrieved 1 October 2012 from <http://www.uclouvain.be/en-258647.html>
- Parkinson, D. (2015). *arabiCorpus*. Retrieved 08 February 2015 from <http://arabicorpus.byu.edu>
- Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. E., Eskander, R., Habash, N., ... Roth, R. M. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: *Proceedings of the LREC 2014, International Conference on Language Resources and Evaluation* (pp. 1094–1101). Reykjavik, Iceland: European Language Resources Association.
- Pastor-i-Gadea, M., Toselli, A. H., Casacuberta, F., & Vidal, E. (2010). A bi-modal handwritten text corpus: baseline results. In: *Proceedings of the 20th International Conference on Pattern Recognition* (pp. 1933–1936). Istanbul, Turkey: IEEE.
- Paulasto, H., & Meriläinen, L. (2012, June). *New corpora in World Englishes and learner English*. The GlobE seminar 2012, University of Eastern Finland, Helsinki.
- Peters, H. (2009). Développement d'un corpus oral d'apprenants: Apport à la didactique du français langue étrangère dans les Caraïbes anglophones [Development of an oral corpus of learners: Contribution to teaching French as a foreign language in the English-speaking Caribbean]. *Cuadernos de Lingüística de la Universidad de Puerto Rico*, 2(2), 21–32.
- Pęzik, P. (2012). *The PELCRA learner English corpus (PLEC)*. Retrieved 26 September 2012 from <http://ia.uni.lodz.pl/plec/>
- Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal*, 26, 81–114.
- Price, N. (2013). *The Gachon learner corpus*. Retrieved 15 December 2014 from <http://koreanlearnercorpusblog.blogspot.be/p/corpus.html>

References

- Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning*. Sebastopol, CA: O'Reilly Media.
- Randall, M., & Groom, N. (2009). The BUiD Arab learner corpus: A resource for studying the acquisition of L2 English spelling. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Fifth Corpus Linguistics*. Liverpool, UK: University of Liverpool.
- Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., ... Andreas, T. (2012). *Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01*.
- Roberts, A. (2014). *aConCorde*. Retrieved 6 April 2014 from <http://www.andy-roberts.net/coding/aconcorde>
- Roberts, A., Al-Sulaiti, L., & Atwell, E. (2006). aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora*, 1(1), 39–60. doi:10.3366/cor.2006.1.1.39
- Rocha, C. F. (2014). A coleta de corpus de aprendizes: questões qualitativas em uma pesquisa sobre a escrita de aprendizes de língua espanhola [Collecting learner corpus: qualitative issues in a research about the writing of Spanish language learners]. *Linguistic Studies*, 42(1), 286–297.
- Rollinson, P., & Mendikoetxea, A. (2008). *The written corpus of Learner English (WriCLE)*. Retrieved 1 October 2012 from <http://web.uam.es/woslac/WriCLE/>
- Römer, U. (2007). Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals. In S. Volk-Birke & J. Lippert (Eds.), *Anglistentag 2006 Halle, Proceedings* (pp. 355–363). Trier, Germany: Wissenschaftlicher Verlag Trier.
- Samy, W., & Samy, L. (2014). *Basic Arabic: A Grammar and Workbook*. Routledge, London, UK.

References

- Schmidt, T., & Wörner, K. (2009). EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics*, 19(4), 565–582.
- Scott, M. (2008). Developing wordsmith. *International Journal of English Studies*, 8(1), 95–106.
- Scott, M. (2012). WordSmith Tools version 6, Liverpool: Lexical analysis software. Retrieved 10 August 2014 from <http://www.lexically.net/wordsmith>
- Sengupta, S. (2002). *Learner corpus of essays and reports*. Retrieved 3 October 2014 from <http://langbank.engl.polyu.edu.hk/index1.html>
- Sharoff, S. (2014). IntelliText Corpus Queries [Computer Software]. Retrieved 3 October 2014 from <http://corpus.leeds.ac.uk/itweb/htdocs/Query.html>
- Shichun, G. (2012). *Chinese learner English corpus (CLEC)*. Retrieved 8 August 2012 from <http://langbank.engl.polyu.edu.hk/corpus/clec.html>
- Shichun, G., & Huizhong, Y. (2012). *Chinese learner English corpus (CLEC)*. The PolyU Language Bank. Retrieved 8 August 2012 from <http://langbank.engl.polyu.edu.hk/corpus/clec.html>
- Shih, R. H.-H. (2000). Compiling Taiwanese learner corpus of English. *Computational Linguistics and Chinese Language Processing*, 5(2), 87–100.
- Sigott, G., & Dobrić, N. (2014). *Learner corpus annotation manual*. Retrieved 16 March 2015 from http://www.uni-klu.ac.at/iaa/downloads/Learner_Corpus_Annotation_Manual.pdf
- Siitonen, K., & Ivaska, I. (2008). *The advanced Finnish learner corpus (LAS2)*. Retrieved 16 December 2014 from <http://www.utu.fi/fi/yksikot/hum/yksikot/suomi-sgr/tutkimus/tutkimushankkeet/las2/Sivut/home.aspx>
- Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English*. Retrieved 16 September 2012 from <http://www.helsinki.fi/varieng/CoRD/corpora/MICASE/>

References

- Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2009). *The Michigan corpus of academic spoken English*. Ann Arbor, MI: The Regents of the University of Michigan. Retrieved 16 September 2012 from <http://micase.elicorpora.info/>
- Sinclair, J. (1996). *EAGLES. Preliminary recommendations on corpus typology*. Retrieved 11 April 2013 from <http://www.ilc.cnr.it/EAGLES/corpusstyp/corpusstyp.html>
- Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1–16). Oxford, UK: Oxbow Books.
- Sketch Engine. (2014). Overview of language integration in Sketch Engine. Retrieved 12 February 2015 from <https://www.sketchengine.co.uk/documentation/wiki/LanguagesOverview>
- Sosnina, E. (2014). *Russian learner translator corpus (RusLTC)*. Retrieved 29 September 2014 from <http://rus-ltc.org>
- Spina, S., Pazzaglia, S., & Perini, M. (2012). *Observatory on Italian and foreigners on Italian abroad*. Retrieved 4 October 2012 from <http://elearning.unistrapg.it/osservatorio/Corpora.html>
- Stephen, R., Harris, K., & Setzler, K. (2012). *Multimedia adult English learner corpus (MAELC)*. Retrieved 16 September 2012 from <http://www.labschool.pdx.edu/research/methods/maelc/intro.html>
- Stritar, M. (2009). Slovene as a foreign language: The pilot learner corpus perspective. *Slovene Linguistic Studies*, 7, 135–152.
- Tagnin, S. E. O. (2006). A multilingual learner corpus in Brazil. *Language and Computers*, 56(1), 195–202.
- TalkBank. (2012). *SLABank database guide*. Retrieved 5 October 2012 from <http://talkbank.org/manuals/SLABank.doc>

References

- Tan, H., Heng, C. S., Nadzimah, A., & Mashohor, S. (2011). *Learner corpus of engineering abstract (LCEA)*. Retrieved 15 December 2014 from <http://www.upmip.upm.edu.my/index.php?content=getfaculty&ipid=1977&ipdetailid=1096&projectlead=263&cluster=7&fac=10>
- Tenfjord, K., Meurer, P., & Hofland, K. (2006). The ASK corpus: A language learner corpus of Norwegian as a second language. In: *Proceedings of the LREC 2006, International Conference on Language Resources and Evaluation* (pp. 1821–1824). Genoa, Italy: European Language Resources Association.
- Thoday, E. (2007, June). *Issues in building learner corpora: An investigation into the acquisition of German passive constructions*. Paper presented at the 2nd Newcastle Postgraduate Conference in Theoretical and Applied Linguistics, Newcastle, UK.
- Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 59–70). Oxford, UK: Oxbow Books.
- Thorne, S., Reinhardt, J., & Golombek, P. (2008). Mediation as objectification in the development of professional academic discourse: A corpus-informed curricular innovation. In J. P. Lantolf & M. E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 256–284). London, UK: Equinox.
- Tono, Y. (2008). The role of oral L2 learner corpora in language teaching: The case of the NICT JLE corpus. In M. C. Campoy & M. J. Luzon (Eds.), *Spoken corpora in applied linguistics* (pp. 163–179). Bern, Switzerland: Lang.
- Tono, Y. (2011). *The JEFLL corpus project*. Retrieved 10 August 2012 from <http://jefll.corpuscobo.net/>
- Tono, Y. (2012a). *The international corpus of cross-linguistic interlanguage (ICCI)*. Retrieved 10 August 2012 from <http://tonolab.tufs.ac.jp/icci/index.jsp>

References

- Tono, Y. (2012b). International corpus of cross-linguistic interlanguage: Project overview and a case study on the acquisition of new verb co-occurrence patterns. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and cross-linguistic perspectives in learner corpus research* (pp. 27–46). Amsterdam, the Netherlands: Benjamins.
- Tortel, A. (2008). ANGLISH: Une base de données comparatives de l'anglais lu, répété et parlé en L1 & L2 [ANGLISH: Comparative database of read, repeated and spoken English in L1 and L2]. *Travaux Interdisciplinaires du Laboratoire Parole et Langage* [The TIPA Journal: Interdisciplinary Works on Speech and Language], 27, 111–122.
- Tortel, A., & Hirst, D. (2008). Rhythm and rhythmic variation in British English: Subjective and objective evaluation of French and native speakers. In P. A. Barbosa, S. Madureira & C. Reis (Eds.), *Proceedings of the Speech Prosody Conference 2008* (pp. 359–362). Campinas, Brazil: ISCA.
- Turton, N. D., & Heaton, J. B. (1996). *Longman dictionary of common errors*. Harlow, UK: Pearson Longman.
- Van Rooy, B. (2009). *Tswana learner English corpus*. Retrieved 9 August 2012 from <http://www.nwu.ac.za/ctext/data>
- Waldman, T. (2005, September). *The use of collocations in Israeli learners' written English*. Linking Up Contrastive and Learner Corpus Research workshop, University of Santiago de Compostela, A Coruña, Spain.
- Wen, Q. (2006, June). *Chinese learner corpora and second language research*. Paper presented at the 2006 International Symposium of Computer-Assisted Language Learning, Beijing, China. Retrieved 17 July 2012 from <http://call2006.fltrp.com/PPT/Keynote/Wen%20Qiufang.ppt>
- Wiechmann, D., & Fuhs, S. (2006). Concordancing software. *Corpus Linguistics and Linguistic Theory Journal*, 2(1), 107-127.

References

- Wilson, J., Hartley, A., Sharoff, S., & Stephenson, P. (2010). Advanced corpus solutions for humanities researchers. In: *Proceedings of PACLIC 24*. Sendai, Japan.
- WordSmith Tools. (2013). WordSmith Tools Manual. Retrieved 17 October 2014 from <http://www.lexically.net/downloads/version6/HTML/index.html?language.htm>
- Wynne, M. (2005). Archiving, distribution and preservation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17–29). Oxford, UK: Oxbow Books.
- Xunfeng, X. (2004). *The Learner Journals corpus*. Retrieved 8 August 2012 from <http://langbank.engl.polyu.edu.hk/index1.html>
- Yimam, S. M., Eckart de Castilho, R., Gurevych, I., & Biemann, C. (2014). Automatic annotation suggestions and custom annotation layers in WebAnno. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 91–96). Baltimore, MD: Association for Computational Linguistics.
- Yimam, S. M., Gurevych, I., Eckart de Castilho, R., & Biemann, C. (2013, August). *WebAnno: A flexible, web-based and visually supported system for distributed annotations*. The Annual Meeting of the Association for Computational Linguistics (ACL-2013), Sofia, Bulgaria.
- Zaghouani, W., Habash, N., & Mohit, B. (2014). *QALB Guidelines (pre-release version)*. Retrieved 8 January 2015 from http://nlp.qatar.cmu.edu/qalb/QALB-guidelines_0.90.pdf
- Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., ... Oflazer, K. (2014). Large-scale Arabic error annotation: Guidelines and framework. In: *Proceedings of the LREC 2014, International Conference on Language Resources and Evaluation* (pp. 2362–2369). Reykjavik, Iceland: European Language Resources Association.

References

- Zinsmeister, H., & Breckle, M. (2012). The ALeSKo learner corpus: Design–annotation–quantitative analyses. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 71–96). Amsterdam, the Netherlands: Benjamins.