

promoting access to White Rose research papers



Universities of Leeds, Sheffield and York
<http://eprints.whiterose.ac.uk/>

This is a paper from the Second Workshop on **Arabic Corpus Linguistics**.

White Rose Research Online URL for this paper:

<http://eprints.whiterose.ac.uk/75470/>

Published paper:

Alfaifi, AYG and Atwell, ES *Arabic Learner Corpus v1: A New Resource for Arabic Language Research*. In: Second Workshop on Arabic Corpus Linguistics, 22 July 2013, Lancaster, UK.

Arabic Learner Corpus v1: A New Resource for Arabic Language Research

Abdullah Alfaifi

University of Leeds

scaayga@leeds.ac.uk

Eric Atwell

University of Leeds

e.s.atwell@leeds.ac.uk

1 Introduction

Recent developments in learner corpora (LC) have highlighted the growing role they play in language teaching and learning. Learner corpora can provide teachers, learners, second language acquisition researchers, lexicographers, language materials writers, etc., with a valuable data resource. For instance, a corpus-based analysis may reveal highly important issues of different aspects of language use by learners, and such results may enable pedagogical materials designers to develop better teaching materials considering the strengths and weaknesses of students, which could lead towards more effective acquisition of the language (Nesselhauf, 2004). Another example showing the relationship between learner corpora and the lexical area, is that monolingual learners' dictionaries (e.g. Longman Essential Activator) take advantage of learner corpora in order to provide learners with help boxes warning them against typical errors (Granger, 2003; Tono, 2009).

2 Background

“English clearly dominates the learner corpus scene” (Granger, 2008: 262), as more than a half of learner corpora on the CECL list¹ (the Centre for English Corpus Linguistics) were devoted to learners of English. We can find a sole project developing an Arabic learner corpus (Abuhakema et al., 2008), *the Pilot Arabic Learner Corpus*. Another Arabic learner corpus –not included on the CECL list– has been compiled by Farwanah and Tamimi (2012), *Arabic Learners Written Corpus*.

3 Arabic Learner Corpus

This paper introduces the first version of the Arabic Learner Corpus (ALC), which comprises a collection of texts written by learners of Arabic in Saudi Arabia. The corpus covers two types of students, non-native Arabic speakers (NNAS) learning Arabic as a second language (ASL) for academic purpose (AAP), and native Arabic

speaking students (NAS) learning to improve their written Arabic. Both groups are males at pre-university level.

4 Design criteria and contents

Design criteria of the ALC were based on a review of a large number of learner corpora in order to identify the best practice in this field such as the Pilot Arabic Learner Corpus (Abuhakema et al., 2008), Arabic Learners Written Corpus (Farwanah & Tamimi, 2012), the British Academic Written English (BAWE) corpus (Heuboeck et al., 2008), the ASU corpus (Hammarberg, 2010), the LONGitudinal DAtabase of Learner English (Meunier et al., 2010), the Michigan Corpus of Upper-level Student Papers (O'Donnell & Römer, 2009a, 2009b), the International Corpus of Learner English (Granger, 2003), the learner corpus of Czech (Hana et al., 2010), and others. These criteria include the determination of corpus contributors, materials included, corpus size, method of data collecting, and metadata. Based on the corpora reviewed, the corpus being developed in this project is the first Arabic learner corpus including both NAS and NNAS, and Arabic as a second language (the existing Arabic learner corpora were devoted for Arabic as a foreign language²).

The current version of ALC has been captured in November and December 2012, and it includes a total of 31272 words, 215 written texts (narrative and discussion) produced by 92 students from 24 nationalities and 26 different L1 backgrounds. 181 texts (84%) were written in class (timed essays), while 34 (16%) produced at home (untimed essays). Average length of the texts is 145 words. 95% of the texts were hand-written, so they had to be transcribed into a suitable computerised form. All identity information (e.g. names, contacts, dates of birth, etc.) have been removed from transcriptions.

	No of students	No of texts	No of words
NNAS	38 41%	105 49%	15531 50%
NAS	54 59%	110 51%	15741 50%

Table1: NNAS vs. NAS in ALC

5 Files format

Two types of non-annotated files have been

¹ The list is accessed from: <http://www.uclouvain.be/en-cecl-lcworld.html>

² 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 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) (Littlewood, 1984).

included: txt, and XML. They are available to download (with the original hand-written sheets) online¹. All corpus files were named in a method which indicates the basic characteristics of the text and its author (e.g. *S038_T2_M_Pre_NNAS_W_C*). They are in order: student identifier number, text number, author gender, level of study, nativeness, text mode, and place of text production.

6 Using the corpus in linguistic research

By annotating a small set of data (6 texts, 1488 tokens), 156 errors were detected and corrected. The most frequent errors were in the category of punctuation, particularly in "Punctuation missing" and "Punctuation confusion" respectively (see Appendix A). Such error analysis reveals the importance of paying more attention to punctuations in Arabic language teaching. This also helps designers of pedagogical materials by indicating the aspects of language which need to be included with more consideration.

7 Future work

As a next stage, the entire corpus will be annotated for errors, and word-tagged with morphological tags to identify part of speech and certain grammatical sub-categories. Additionally, the correct form will be reconstructed by correcting the mistakes (Appendix B). Annotation of errors will be performed using a detailed error-type tagset (Appendix A), which has been developed for Arabic learner corpora in general and to be used in the present corpus in particular (Alfaifi & Atwell, 2012). In future, further versions will be issued including more materials (written and spoken), different genders (male and female), and different levels of study (pre-university and university).

References

Abuhakema, Ghazi, Feldman, Anna, and Fitzpatrick, Eileen. (2008). *Annotating an Arabic Learner Corpus for Error*. In the proceedings of the International Conference on Language Resources and Evaluation (LREC 2008), 26 May - 1 June 2008, Marrakech, Morocco

Alfaifi, Abdullah, and Atwell, Eric. (2012). المدونات اللغوية "Arabic Learner Corpora (ALC): A Taxonomy of Coding Errors". In the proceedings of the 8th International Computing Conference in Arabic (ICCA 2012) 26-28 December 2012, Cairo, Egypt

Farwaneh, S, and Tamimi, M. (2012). Arabic Learners Written Corpus: A Resource for Research and

Learning. Retrieved 2 September, 2012, from the the University of Arizona, the Center for Educational Resources in Culture, Language and Literacy web site: <http://l2arabiccorpus.cercll.arizona.edu/?q=homepage>

- Granger, Sylviane. (2003). 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, Sylviane. (2008). Learner Corpora. In A. Ludeling and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 259-275). Berlin: Walter de Gruyter.
- 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 University: Department of Linguistics.
- Hana, J. , Rosen, A. , Škodová, S. , and B., Štindlová. (2010). *Error-tagged learner corpus of Czech*. In the proceedings of the Fourth Linguistic Annotation Workshop, Uppsala. Association for Computational Linguistics
- Heuboeck, A., Holmes, J., and Nesi, H. (2008). The BAWE Corpus Manual. Retrieved 24 July 2012, from: http://www.reading.ac.uk/AcaDepts/ll/app_ling/interanl/bawe/BAWE.documentation.pdf
- Littlewood, William. (1984). *Foreign and Second Language Learning: Language Acquisition Research and Its Implications for the Classroom*. Cambridge Cambridge University Press.
- Meunier, F., Granger, S., Littré, D. , and Paquot, M. . (2010). The LONGDALE (Longitudinal Database of Learner English). Retrieved 14 September, 2012, from the Université Catholique de Louvain, Centre for English Corpus Linguistics web site: <http://www.uclouvain.be/en-cecl-longdale.html>
- Nesselhauf, Nadja. (2004). Learner Corpora and Their Potential in Language Teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 125-152). Amsterdam & Philadelphia: Benjamins
- O'Donnell, M. B. , and Römer, U. (2009a). *From student hard drive to web corpus: The design, compilation, annotation and online distribution of the MICUSP corpus*. A poster present at ICAME 30, 27-31 May 2009, Lancaster University, UK.
- O'Donnell, M. B. , and Römer, U. (2009b). Michigan Corpus of Upper-level Student Papers. Retrieved 27 July 2012, from: <http://micusp.elicorpora.info/>
- Tono, Yukio. (2009). The potential of learner corpora for pedagogical lexicography. In V. B. Y. Ooi, A. Pakir, I. S. Talib and P. K. W. Ten (Eds.), *Perspectives in Lexicography: Asia and beyond* (pp. 105-115). Tel Aviv: K DICTIONARIES LTD

¹ It can be accessed from: <http://www.comp.leeds.ac.uk/scayga/alc>

Appendix A

Table2: Error tag set developed for Arabic learner corpora and their frequency in small set of data

Error Category مجال الخطأ	Error Type نوع الخطأ	Arabic tag الرمز العربي	English tag الرمز الإنجليزي	Error freq. in the test data
Orthography الإملاء 'l'imlā'	1. Hamza (ء، أ، إ، و، ي، ن)	<اه>	<OH>	5
	2. Tā' Mutadarrifa (تاء المتطرفة (ة، ت))	<اة>	<OT>	0
	3. 'alif Mutadarrifa (ألف المتطرفة (ا، ي))	<إي>	<OA>	1
	4. 'alif Fāriqa (ألف الفارقة (كتبتوا))	<إت>	<OW>	5
	5. Lām Šamsiya (اللام الشمسية (الطالب))	<لا>		0
	6. Tanwin (تتوين (قَوِّو))	<إل>	<ON>	0
	7. Fast wa Wasl (Conjunction) الفصل والوصل	<إو>	<OF>	9
	8. Shortening the long vowels تقصير الصوائت (أوي → قَوِّو)	<إف>	<OS>	2
	9. Lengthening the short vowels تطويل الصوائت (أوي → قَوِّو)	<إق>	<OG>	0
	10. Wrong order of word characters الخطأ في ترتيب الحروف داخل الكلمة	<إط>	<OC>	1
	11. Replacement in word character(s) استبدال حرف أو أحرف من الكلمة	<إس>	<OR>	6
	12. Character(s) redundant وجود حرف أو أحرف زائدة	<إز>	<OD>	6
	13. Character(s) missing وجود حرف أو أحرف ناقصة	<إن>	<OM>	3
	14. Other orthographical errors أخطاء إملائية أخرى	<إخ>	<OO>	0
Morphology الصرف 'ssarf'	15. Word inflection صيغة الكلمة	<حص>	<MI>	2
	16. Verb tense زمن الفعل	<صز>	<MT>	1
	17. Other morphological errors أخطاء صرفية أخرى	<صخ>	<MO>	0
Syntax النحو 'nnaḥw'	18. Case/Mood Mark الموقع الإعرابي أو علامة الإعراب	<نب>	<XC>	1
	19. Definiteness التعريف والتنكير	<نع>	<XF>	11
	20. Gender التذكير والتأنيث	<نذ>	<XG>	3
	21. Number (Singular, Dual and plural) العدد (الإفراد والتثنية والجمع)	<نف>	<XN>	0
	22. Word(s) order ترتيب المفردات داخل الجملة	<نت>	<XR>	1
	23. Word(s) redundant وجود كلمة أو كلمات زائدة	<نز>	<XT>	4
	24. Word(s) missing وجود كلمة أو كلمات ناقصة	<نن>	<XM>	9
25. Other syntactic errors أخطاء نحوية أخرى	<نخ>	<XO>	0	
Semantics الدلالة 'ddalāla'	26. Word selection اختيار الكلمة المناسبة	<دب>	<SW>	17
	27. Phrase selection اختيار العبارة المناسبة	<دق>	<SP>	1
	28. Failure of expression to indicate the intended meaning قصور التعبير عن أداء المعنى المقصود	<دد>	<SM>	2
	29. Wrong context of citation from Quran or Hadith الاستشهاد بالكتاب والسنة في سياق خاطئ	<دس>	<SC>	0
	30. Other semantic errors أخطاء دلالية أخرى	<دخ>	<SO>	0
Style الأسلوب 'l'uslūb'	31. Unclear style أسلوب غامض	<سغ>	<TU>	2
	32. Prosaic style أسلوب ركيك	<سض>	<TP>	7
	33. Other stylistic errors أخطاء أسلوبية أخرى	<سخ>	<TO>	0
Punctuation علامات الترقيم 'alāmāt 't-tarqīm'	34. Punctuation confusion الخلط في علامات الترقيم	<تط>	<PC>	21
	35. Punctuation redundant علامة ترقيم زائدة	<تنز>	<PT>	1
	36. Punctuation missing علامة ترقيم مفقودة	<تنن>	<PM>	35
	37. Other errors in punctuation أخطاء أخرى في علامات الترقيم	<تنخ>	<PO>	0

Appendix B

Examples of annotated texts

One-word error	<pre> <err type="OT" errform="اللتى " crrform="التي "> <w id="1">اللتى <t id="1" token="اللتى " lemma="التي " pos="NR" fun="VA"></t> </w> </err> </pre>
Multi-word error	<pre> <w id="2">كنت <t id="2" token="كن" lemma="كان" pos="VP"></t> <t id="3" token="ت" lemma="ت" pos="RR" fun="NK"></t> </w> <w id="3">قد <t id="4" token="قد" lemma="قد" pos="PB"></t> </w> <err type="TP" errform="أنا لك أعطي " crrform="أعطيتك "> <w id="4">أعطي <t id="5" token="أعطي " lemma="أعطي " pos="VP"></t> </w> <w id="5">أنا <t id="6" token="أنا " lemma="أنا " pos="NP" fun="NV"></t> </w> <w id="6">لك <t id="7" token="ل" lemma="ل" pos="PP"></t> <t id="8" token="ك" lemma="ك" pos="RR" fun="GF"></t> </w> </err> </pre>

Figure1: Example of annotated text in XML format

	<i>Word</i>	<i>Lemma</i>	<i>PoS</i>	<i>Grammatical function</i>
One-word error	<s>			
	<err type="OT">			
	اللتى	اللتى	NR	VA
	</err>			
Multi-word error	<corr type="OT">			
	التي	التي	NR	VA
	</corr>			
	كن	كان	VP	
	<g/>			
	ت	ت	RR	NK
	قد	قد	PB	
	<err type="TP">			
	أعطي	أعطي	VP	
	أنا	أنا	RR	NV
	ل	ل	PP	
	<g/>			
ك	ك	RR	GF	
</err>				
<corr type="TP">				
أعطي	أعطي	VP		
<g/>				
ت	ت	RR	NV	
<g/>				
ك	ك	RR	GF	
</corr>				
</s>				

Figure2: Annotated text prepared for Sketch Engine