Potential Uses of the Arabic Learner Corpus

Abdullah Alfaifi
(PhD student)
University of Leeds
scayga@leeds.ac.uk

Eric Atwell
(Supervisor)
University of Leeds
e.s.atwell@leeds.ac.uk

1 Introduction

This paper presents the potential uses of the first version of the Arabic Learner Corpus (ALC), which comprises a collection of texts written by learners of Arabic in Saudi Arabia. The original source of the corpus (hand-written sheets) is available online alongside with their transcription (plain text format)[1]. The corpus is also intended to be lemmatised and annotated with linguistic features, including Part-of-Speech and grammatical functions tags, and mark-up of errors with their corrections.

2 Design criteria and contents of the Arabic Learner Corpus

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 (Abuhakema *et al.*, 2008; Farwaneh & Tamimi, 2012; Granger, 2003; Heuboeck *et al.*, 2008, and others) and others. These criteria include corpus contributors, materials included, corpus size, method of data collecting, and metadata.

The current version of ALC (Alfaifi & Atwell, forthcoming) 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. 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.

Table 1: Table1: NNAS vs. NAS in ALC

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

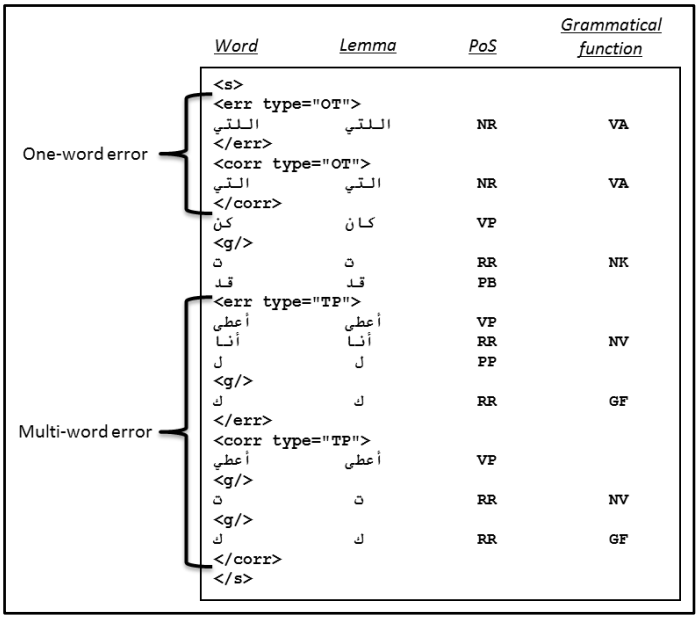

Figure1: Example of annotated text prepared for Sketch Engine

[1] The corpus can be accessed from: http://www.comp.leeds.ac.uk/scayga/alc/index.html

3 Using the corpus in linguistic research

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.

3.1 Contrastive Interlanguage Analysis (CIA)

CIA is still one of the most frequently used approaches for analysing 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, 2003). Analysing errors will also enable researchers and educators understand the interlanguage errors caused by L1 transfer, learning strategies and overgeneralization of L1 rules.

3.2 Learner dictionary making

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 provide 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, 2003; Nesselhauf, 2004).

3.3 Second Language Acquisition

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 & Caines, 2012; Nesselhauf, 2004). Longitudinal learner corpora usually involve such goal in their compilation purposes. 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).

3.4 Designing pedagogical materials

Analysing learners' errors may function as a beneficial basis for pedagogical purposes such as creating instructional teaching materials 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.

3.5 Optical Character Recognition (OCR)

Finally, the corpus can be used as a training data set in research of Optical Character Recognition (OCR), as it contains hand-written texts and their transcription in a computerised format.

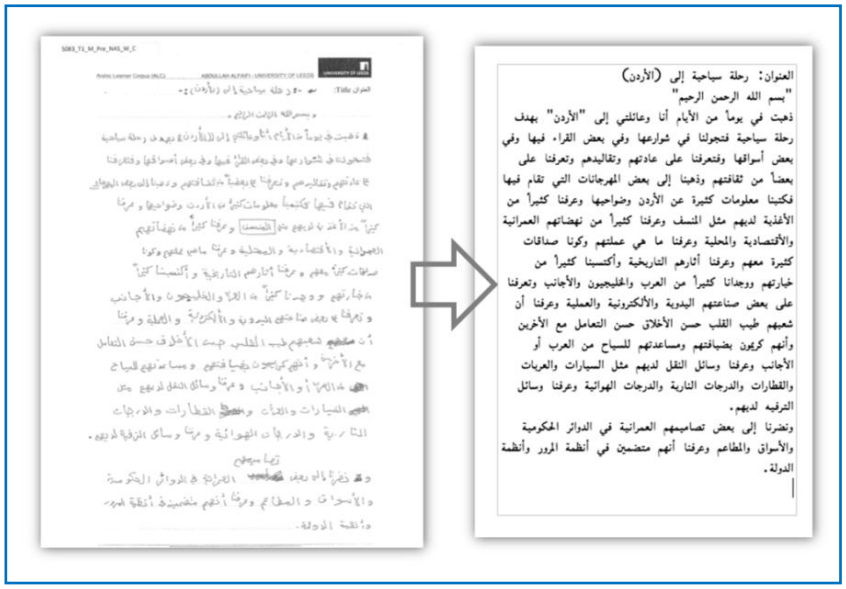

Figure2: Example of a hand-written text with its transcription

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

Buttery, P, and Caines, A. (2012). Normalising Frequency Counts to Account for 'opportunity of use' in Learner Corpora. In Y. Tono, Y. Kawaguchi and M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 187-204). Amsterdam: John Benjamins

Diez-Bedmar, M. B. (2009). Written Learner Corpora by Spanish Students of English: an overview. In P. C. Gómez and A. S. Pére (Eds.), *A Survey on Corpus-based Research. Proceedings of the AELINCO Conference* (pp. 920-933). Murcia: Asociación Española de Lingüística del Corpus

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.

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.

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/internal/bawe/BAWE.documentation.pdf

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