Arabic Learner Corpus

A New Resource for Arabic Language Research

Abdullah Alfaifi (scayga@leeds.ac.uk), Eric Atwell (e.s.atwell@leeds.ac.uk)

University of Leeds

Introduction

Recently, learner corpora have been used increasingly for several purposes in terms of language learning and teaching (e.g. language materials design, learner dictionaries development, learners' errors analysis, etc.). The present paper introduces the first version of the Arabic Learner Corpus (ALC) and shows details about the corpus content, files format, and the corpus website. The corpus was compiled in Saudi Arabia and the data was captured in November and December 2012. The participants were asked to write narrative and discussion texts. The current version includes 215 texts produced by 92 learners of Arabic. The corpus is intended to be annotated with linguistic features in order to enable researchers and teachers analysing the learners' production.

Content

The first version of the corpus has been captured in November and December 2012 from five groups of students. It includes a total of **31272** words, and **92** students (from **24** nationalities and **26** different L1 backgrounds). The participants produced **215** written texts (narrative and discussion). Detailed information about learners and texts in this version is illustrated in Table 3 below.

Gender	Male
Nativeness	NAS and NNAS
General Level	Pre-university
Number of students	92
Number of Nationalities	24
Number of L1 languages	26
Age	16 – 28
Number of texts	215
Number of words	31272
Average of text length	145

Table 1: Summary of the corpus content

Nativeness	No of students	No of texts	No of words
NNAS	38 (41%)	105 (49%)	15531 (50%)
NAS	54 (59%)	110 (51%)	15741 (50%)
Total	92	215	31272

Table 2: NAS vs. NNAS

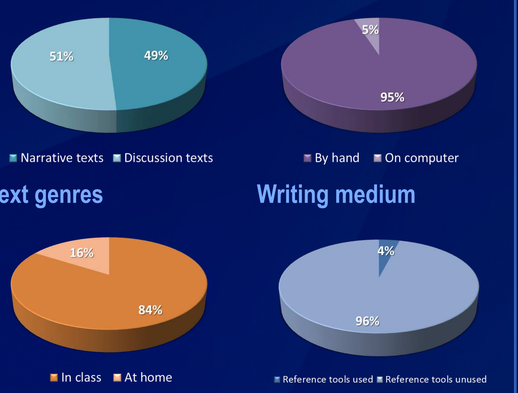
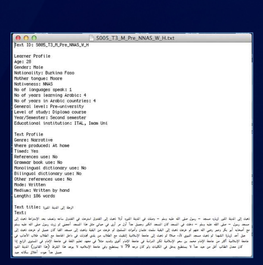


Figure 1: some properties of the corpus texts

Files format

Three types of non-annotated files have been generated: (1) with no header, (2) with metadata header in Arabic, (3) and in English. Files with header are available in two formats, txt and XML. The metadata information enables researchers to identify characteristics of text and its producer in each transcription. The original hand-written sheets are also available after they have been scanned and saved into PDF-format files.



Text file with English metadata header

Text file with Arabic metadata header

XML file with English metadata header

XML file with Arabic metadata header

Figure 2: File types of ALC

All corpus files were named in a method which indicates the basic characteristics of the text and its author (e.g. *S102_T1_M_Pre_NNAS_W_C*).

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

Table 3: Naming ALC files

Corpus website

A website has been created in order to make the corpus files publicly available for download. Beside the information about the corpus on this website, the texts can be downloaded in different types of text and XML files. At the same, they can be downloaded all in one ZIP file, or based on nine classifications. The corpus website also contains a page devoted for collecting data in further versions. This contribution form was created in Arabic as the target language with English translation as an International Language. Additional page was created to include an evaluation questionnaire which will be added after uploading the annotated part of the corpus.

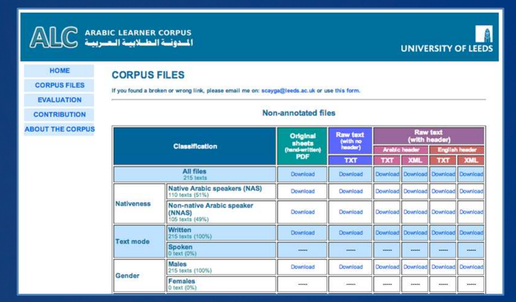


Figure 3: ALC website

http://www.comp.leeds.ac.uk/scayga/alc

Further work

As a next stage, the 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. Annotation of errors will be performed using a detailed error-type tagset, 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

Alfaifi, Abdullah, and Atwell, Eric. (2012). 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

Alfaifi, Abdullah and Atwell, Eric. (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., Atwell, E. and Abuhakema, G. (2013). Error Annotation of the Arabic Learner Corpus: A New Error Tagset. In proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology. Darmstadt, Germany.