

promoting access to White Rose research papers



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

This is the author's version of a Proceedings Paper presented at the
International Conference on Corpus Linguistics (CORPORA-2013)

Alfaifi, AYG and Atwell ES, *Arabic Learner Corpus: Texts Transcription and Files Format*. In: UNSPECIFIED the International Conference on Corpus Linguistics (CORPORA-2013), St. Petersburg, Russia. (In Press)

White Rose Research Online URL for this paper:

<http://eprints.whiterose.ac.uk/id/eprint/75554>

Abdullah Alfaiji, Eric Atwell

ARABIC LEARNER CORPUS: TEXTS TRANSCRIPTION AND FILES FORMAT

Abstract. This paper introduces standards used for transcribing texts of the Arabic Learner Corpus (ALC) from the hand-written sheets into an electronic format. It describes the transcription process which was performed by three transcribers, and the measurement conducted for keeping consistency in transcription. The paper concludes with a description of the corpus file produced based on the electronic format.

1. Introduction

For searching and analysing corpora efficiently, they need to be in a standardised format. Corpus linguists tend to use plain text format which is readable by most of the language processing tools, and subsequently handle tags of mark-up languages such as XML. Some learner corpora, however, contains hand-written texts which required further work to convert them into an electronic form. Transcribing such texts with no standards –specifically by more than one– may yield differences in the final production, as many things 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., 2010¹; Thompson, 2005²).

¹ Pastor-i-Gadea, M., Toselli, A. H., Casacuberta, F., and Vidal, E. (2010, 23-26 Aug. 2010). A Bi-modal Handwritten Text Corpus: Baseline Results. In the proceedings of the 20th International Conference on Pattern Recognition (ICPR), 2010

² Thompson, Paul. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 59-70). Oxford: Oxbow Books

2. ALC¹

The transcription was performed on the first version of the Arabic Learner Corpus², which comprises a collection of texts written by learners of Arabic in Saudi Arabia. The corpus covers two types of students, native Arabic speaking students (NAS), and non-native Arabic speakers (NNAS). Both groups are males at pre-university level. ALC 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). In addition to part-of-speech annotation, the corpus is intended to be tagged for errors using error-tagset developed for Arabic learner corpora³.

3. Standards of texts transcription

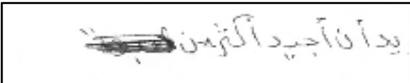
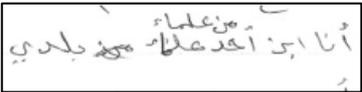
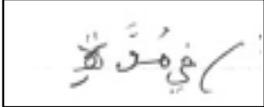
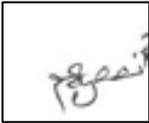
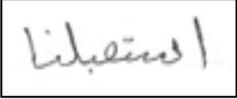
Three transcribers, the researcher and two volunteering colleagues (C1 and C2), performed the transcription upon standards they agreed on, as no standard practice was found for transcribing Arabic from hand-written into computerised form. Most of these standards had been extracted by the researcher in 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 transcription process when they come across uncertain points. The transcription standards list in Table 1.

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

² Alfaihi, Abdullah, and Atwell, Eric. (2013). Arabic Learner Corpus v1: A New Resource for Arabic Language Research. The Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK.

³ Alfaihi, Abdullah, and Atwell, Eric. (2012). المدونات اللغوية لمتعلمي اللغة العربية: نظام لتصنيف وترميز الأخطاء 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.

Table 1. Standards followed in transcription with authentic examples from the corpus texts

Standard followed by example with reference to its sheet	
1. Any struck-out texts should be excluded.	 <p style="text-align: center;"><i>S001_T2_M_Pre_NNAS_W_C</i></p>
2. If there is a correction above a non-struck out word, correction form is the transcribed.	 <p style="text-align: center;"><i>S005_T4_M_Pre_NNAS_W_H</i></p>
3. When there is a doubtful form of a character, the closer to the correct form is the transcribed. For instance, the author here wrote "هـ" which looks also "هـ", the correct form is "هـ", which has been thus transcribed.	 <p style="text-align: center;"><i>S005_T4_M_Pre_NNAS_W_H</i></p>
4. If there is an overlap between hand-written characters, which cannot be transcribed, the closest possible form is selected. The example word here can be transcribed as "نصصهم".	 <p style="text-align: center;"><i>S005_T4_M_Pre_NNAS_W_H</i></p>
5. If a writer forgot to add a character's dot(s) whether above or below, it should be transcribed as it is written by the learner, unless it is not possible, for example if there is no equivalent character on computer. The example here is transcribed as "استقبلنا".	 <p style="text-align: center;"><i>S006_T1_M_Pre_NNAS_W_C</i></p>
6. Inserting a new line (paragraph) only when it is clear. For instance, if there is a clear space at the end of a line (whether there is a period or not), also if there is a clear	

space at the beginning of the new line with a period at the end of the previous paragraph. Other instances, such as ending a line with period but with no clear space at the end or at the beginning of the new line, are considered as a single paragraph.

وخطنا أن نقف في النهر.
" الحمد لله وصلنا - قال أبي - خرجنا من السيارة

Clear space at the end of previous line

جاء يوم جديد... ذهبنا إلى الجبل لترفع بناها.
وجدنا من شجر الذي يؤمل إلى ارتفاع الجبل - في الطريق

No clear space at the end of previous line

S003_T1_M_Pre_NNAS_W_C

7. Any identity information (e.g., learner's name, contacts, postal address, emails, etc.) should be replaced with #معلومة شخصية محذوفة# (personal information deleted). Other non-personal information can be left such as class, name of school, city, country, religion, culture, etc.

بأبي (Personal information deleted)
طوبوي

S014_T4_M_Pre_NNAS_W_H

8. Any shape, illustration, or ornamentation drawn by the learner on the sheet is excluded.

☆ رحلتني إلى اسطنبول ☆

S026_T1_M_Pre_NNAS_W_C

9. Text with no titles are given "النص بدون عنوان" (text with no title) in the title field.

Title:
... الحمد لله لأنني كتبت أريد أن أكل دراستي العربية

S030_T2_M_Pre_NNAS_W_C

10. Any text format is excluded such as underline words or sentences.

التفصيص - العنيزة -

S009_T2_M_Pre_NNAS_W_C

11. Unknown words or phrases are replaced with "#كلمة غير معروفة#" (unknown word), or "#عبارة غير معروفة#" (unknown phrase). The example here is transcribed as "الحافلة في" "#كلمة غير معروفة#، وصلنا

الحافلة في البتج، وطننا

S015_T1_M_Pre_NNAS_W_C

All texts have been transcribed into a database with hiding any identity information (Fig. 1). In addition, the transcription assistants were not allowed to access the learners' profiles, they only have the hand-written sheets.

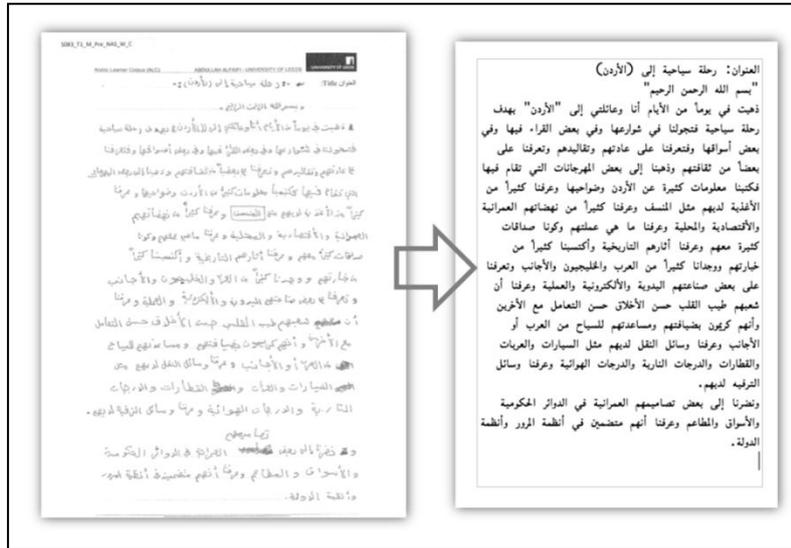


Fig. 1. Example of a hand-written text with its transcription

The assistants were given 10 sheets each time to be transcribed within 5 days (2 sheets per day). The transcription process started at the beginning of December 2012 (in parallel with the second month of data collection) and lasted about 46 days. One of the assistants (C1) withdrew after performing 20 transcriptions (10 days), the second

(C2) transcribed 60 texts (46 days including 16 days delay). The rest 135 texts were transcribed by the researchers in 14 days (Table 2).

Table 2. Time of texts transcription

	C1	C2	Researcher
No of texts transcribed	20	60	135
No of days	10	46	14
Average (sheet per day)	2	1.3	9.6
Total average of the transcribers	3.9		
Total days of transcription process	46		

4. Consistency measurement

In order to ensure that there is a consistency in transcription, the researcher with both of the assistants, after discussing the transcription standards, transcribed one text (*S011_T1_M_Pre_NNAS_W_C*), then the method that has been followed to measure the consistency between each two transcribers was number of agreements divided by number of word in the text (120). It yielded a percentage from which the average was extracted. The result showed an average of 93%, as illustrated in Table 3.

Table 3. Consistency between transcribers

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

*R = the researcher

This consistency measurement was performed again (on the text *S009_T1_M_Pre_NNAS_W_C*) after the first 10 transcriptions of C1 and C2, the result revealed an improvement by 5%, as Table 4 shown.

Table 4. Second test of consistency between transcribers

	C1 and C2	C1 and R	C2 and R
No of similarities (from 132)	128	129	131
percentage	97%	98%	99%
Average		98%	

A final test was between C2 and the researcher (on the text *S003_T3_M_Pre_NNAS_W_H*), at the end of December (about two weeks before the transcription completed), which illustrated that the consistency is still at 98% (Table 5).

Table 5. Final test of consistency

	C2 and R
No of similarities (from 104)	102
percentage	98%

5. Files format

Three types of non-annotated files have been generated after the transcription process: (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 (see the example in Fig. 3). 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. 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.

