Arabic Learner Corpus (ALC) v2

—A New Written and Spoken Corpus of Arabic Learners—

Abdullah ALFAIFI
University of Leeds, UK

Eric ATWELL
University of Leeds, UK

Hedaya IBRAHEEM
Al-Imam Muhammad Ibn Saud Islamic University, KSA

Abstract

Arabic learner corpora have not received enough attention, particularly for learning Arabic as a second language (in Arabic speaking countries). Based on the literature, there are a few projects are developing Arabic learner corpora, of which most are not freely available for users or researchers. In addition to that they are intended to assist in the language acquisition of Arabic as a foreign language (collected from learners studying Arabic in non-Arabic speaking countries). The present paper aims to introduce the Arabic Learner Corpus. It is being developed at Leeds University, and comprises of 282,732 words, collected from learners of Arabic in Saudi Arabia. The corpus includes written and spoken data produced by 942 students, from 67 different nationalities studying at pre-university and university levels. The paper focuses on two angles of this corpus; the design criteria and the content. The design criteria of the ALC discuss the target language, the participants, the corpus size, the materials included, the method of data collection, the metadata of the corpus materials and contributors, and text distribution. The second part, ALC content, is illustrated based on 26 elements representing the corpus metadata. The goal of the ALC is to provide an open-source of data for some linguistic research areas related to Arabic language learning and teaching. So, the corpus data is available for download in TXT and XML formats, hand-written sheets which are in PDF format as well as the audio recordings which are available in MP3 format.

Keywords

Arabic, Learner, Corpus, Language, Materials, Data, Text

I Introduction

Learner corpora are increasingly being used in some linguistic research areas such as Language Teaching and Learning, Applied Linguistics, Lexicography, etc. as well as for other purposes such as Error Analysis, Learners' Improvement Monitoring, Language Materials Designing, Contrastive Inter-language Analysis, Building Learners' Dictionaries and Common Errors Dictionaries. However, a lack of developing freely available learner corpora for Arabic may interpret the shortage of research on Arabic in the linguistic research areas aforementioned. This paper introduces the open-source ALC project, which includes 282,732 words of written and spoken materials collected from Arabic learners in Saudi Arabia. The main goal of this project is for the ALC to be used as a data source to serve as a research tool for the Arabic language, particularly for learning and teaching Arabic as a second language in Arabic-speaking countries.

II Literature Review

The literature confirms that the Arabic learner corpora have not received enough attention, as a quite few projects of developing Arabic learner corpora can be found, which are intended for Arabic as a foreign language (collected from learners studying Arabic in non-Arabic speaking countries), such as the Pilot Arabic Learner Corpus (Abuhakema *et al.*, 2008), Corpus of L2 Arabic Malaysians learners (Hassan & Daud, 2011), and Arabic Learners Written Corpus (Farwaneh & Tamimi, 2012). A brief review of these corpora is given below.

2.1 Pilot Arabic Learner Corpus (PALC)

According to Abuhakema *et al.* (2008), this corpus is designed to include around 9.000 words of Arabic written materials produced by American native speakers of English who learn Arabic as a foreign language (AFL). Two levels were included, intermediate (3818 words) and advanced (4741 words). Some of the learners' texts were written while the learners were studying Arabic in the United States, whilst others were produced when they went to study abroad in Arab countries. A tagset of error annotation was developed – adopting FRIDA (Granger, 2003a) – to mark-up the learners' errors.

2.2 Corpus of Malaysians Arabic Learners

This written corpus was mainly designed to investigate the misuse of Arabic conjunctions among learners. It includes approximately 240,000 words, produced by 60 university students (97% of them were Malaysians) during their first and second year of their Arabic major degree at the Department of Arabic Language and Literature, International Islamic University Malaysia. The corpus materials include descriptive and comparative essays produced on the computer using Microsoft Word without any help from native speakers (Hassan & Daud, 2011). It seems that the previous two corpora are not freely available for public use and it was not specified otherwise by the corpora developers.

2.3 Arabic Learners Written Corpus (ALWC)

Materials of the Arabic Learners Written Corpus were produced by non-native Arabic speakers from the USA 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 in the first phase is to offer a data source for hypothesis testing and for developing teaching materials, in the second phase the corpus is intended to be tagged for errors (morphological, syntactic and orthographic errors) as well as the characteristics of each level. ALWC is available for download in PDF format files (Farwaneh & Tamimi, 2012).

The previous review of the Arabic learner corpora reveals that there is a lack of 'freely available' learner corpora, particularly those that may have been collected from Arabic speaking countries. The present project aims to fill this gap by compiling a systematic Arabic learner corpus to be used in research on Arabic language learning and teaching. In addition to the previous three corpora, the design criteria for the ALC was based on reviewing the *Centre for English Corpus Linguistics* (CECL) list of learner corpora (Granger & Dumont, 2012) in order to identify the best practice in such projects, the next section explores the design criteria of the ALC.

III Design Criteria of the ALC

It is believed that a smaller homogeneous corpus which features a high quality design is far more valuable than a larger corpus (Granger, 1993). Therefore, a specific criteria had to be designed for the ALC data, as it is intended to consist of a number of sub-corpora (*native* vs. *non-native speakers*, *males* vs. *females*, *pre-university* vs. *university*, and *written* vs. *spoken*). The following areas were given consideration when looking at the design criteria for the ALC: the target language, its participants, the corpus size, the materials included, the method of data collection, the metadata of the corpus materials and the contributors, and the text distribution.

3.1 Target Language

There were two essential reasons behind choosing Arabic as a target language for the learner corpus. Firstly works of two of the researchers in the field of teaching Arabic as a second language (ASL). Such a corpus will be a much welcomed and valuable resource for Arabic learning and teaching research and it could be an essential tool for use by teachers assisting them in the teaching process itself. The second reason is due to the absence of such a project where no such compilation of an Arabic learner corpus exists in the Arab world. This may well reveal some significant aspects of Arabic learning and teaching in Arabic speaking countries compared to the situation of Arabic learning and teaching in non-Arabic speaking countries.

Based on the researcher's experience of teaching Arabic to native speakers and Arabic as a second language to non-native learners, it can be argued 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 Arabic language targeted to be included in the ALC is the same as 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 (NAS) are learning Arabic as a part of their curriculum to improve their written Arabic. Non-native Arabic-speaking learners (NNAS) 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, more details about the contributors is covered in the next section.

3.2 Participants

One of the best practices in learner corpora is to have a balance between the production of both native and non-native learners (see for example these corpora; Hammarberg, 2010; Heuboeck *et al.*, 2008; O'Donnell & Römer, 2009). In terms of the mother tongue backgrounds of non-native learners, the best choice was to have learners who spoke various first languages, as 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).

The International Corpus of Learner English (Granger, 2003b) is one of the standard learner corpora, the general classification of learners levels includes secondary school and university. The same classification was used in the ALC, however the first level was named *Pre-university*, as it includes two parallel groups of learners, NAS learning at secondary schools and NNAS learning Arabic at institutions who teach Arabic as a second language. Both of these groups are counted as a pre-university, as it is the level they have to achieve before continuing their study at a university (Table 1). The second level, *University*, is for both undergraduate and postgraduate learners of those specialising in the same target language, Arabic. Each of these major levels, *Pre-university* and *University*, are broken up into an appropriate number of sub-categories based on levels used in their institutions such as the year of study following the British Academic Written English (BAWE) corpus (Heuboeck *et al.*, 2008) which is based on the learners' year of study as a level indicator.

Table 1 Levels of the learners who contributed to the ALC

Level	NAS	NNAS
Pre-university	Learning at secondary schools	Learning Arabic at institutions where Arabic is taught as a second language
University	Learning Arabic at institutions of teaching Arabic as a second language	

3.3 Corpus Size

Sinclair (2005) believes that it is not a significant factor about how large or small the corpus is, so there is no maximum size, but 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, it can be argued 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 (Granger, 2003b). However, Pravec (2002) emphasises the need to adequately represent the learner's language in a corpus, though the amount of time and effort involved in this meticulous process of compiling a learner corpus is very time consuming. He concludes by stating that despite there being no uniformity in the size of the corpora, each corpus was built to address the needs of its developers.

With respect to the ALC, it is intended (1) for it to be annotated with linguistic features and errors, and (2) will include a spoken part with transcriptions; these two processes are highly laborious and time-consuming. Additionally, (3) it is a unique and timed project (PhD project). These three elements consequently limited the size of current version of the corpus. In the design criteria for the ALC, the intended size was 200,000 words. Granger (2003a) argues that "[a] corpus of 200,000 words is big in the [Second Language Acquisition] 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).

3.4 Materials Included

Determining the criteria of materials included in a corpus is one of the major steps in corpus building, it includes the mode of the text: speech or writing, and the type of text: a book, a journal, a notice or a letter, besides other criteria (Sinclair, 2005). An essential criteria which enables researchers to avoid any distortion in the results of their studies is to collect similar data, "this means that the essays must be written by learners at a similar level under the same conditions and on similar topics" (Granger, 1993: 61). Mode and genre will be discussed in order to justify the selections for the ALC.

3.4.1 Text Mode

Kennedy (1998) argues that most corpus-based grammatical and lexical studies of English have so far been based on written-corpora analysis, whereas spoken language represents the the most common mode or use of language. Leech (1997) has expressed the same concern about the dominance of written corpora, he suggested that a corpus should contain at least as much spoken materials as written materials" (p 17). Having a looked at the CECL list of 117 learner corpora (Granger & Dumont, 2012) it reveals that the case is still the same, the written corpora covers two thirds of the entire number whilst the spoken corpora proportion is 24%. The reset corpora (which are about 10%) have both written and spoken materials.

Compared to written language, some difficulties lie behind the lack of spoken corpora, as dealing with spoken language involves 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). However, some relatively new insights into the essential nature of language use can be explored through spoken language corpora (Kennedy, 1998). Thus, considering the difficulties spoken corpora may have and the benefits of such language, the ALC was designed to contain a small proportion of spoken language. The first version (v1) of the ALC includes only written materials modelled after the BAWE corpus (Heuboeck *et al.*, 2008). In the current version (v2) of the ALC, speech elements consisting of more than 3 hours' worth of audio, along with transcriptions is included, representing 7% of the corpus.

3.4.2 Text Genre

With respect to genres, "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 the corpora reviewed. For instance, in the ASU (Andraspråkets strukturutveckling "Structural development of the second language") corpus, some genres exist such as picture story, narration, description and discussion (for written texts); also narration of picture stories, description and discussion of photos, interviews and discussion of newspaper articles (for speech) (Hammarberg, 2010). We decided to choose two common genres in the learner corpora, narrative and discussion. For the written part, learners had to write narrative or discussion essays under two specific titles which were likely to suit both native and non-native learners of Arabic, entitled "*a vacation trip*" for the narrative and "*my study interest*" for the discussion type. The same genres and titles were chosen for the spoken sub-corpus, as it was thought that it may enable researchers to undertake comparisons between these two modes of materials (written and spoken).

3.5 Data Collection

The corpus data was not taken from materials that learners had produced previously; instead, some tasks were designed to collect the data from the Arabic learners. The written part of the ALC includes two tasks and the participants had the choice to do either one of them or both. Each of these tasks involved similar topics (narrative: a vacation trip, and discussion: my study interest), the difference was that the first task was timed (40 minutes for each text) and the learners were not allowed to consult any language references while writing their essays such as dictionaries, grammar books, etc. Students in the second task were asked to write their essays about the same topics as in task 1 but were asked to complete this as homework. They were allowed two days to complete the homework and were granted the opportunity to use any language references they wanted to, this was done to enable them to improve their writing before submitting their work as well as to allow them enough time within which to complete the homework. The first task was also used to collect the oral data, learners had a 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. The original hand-written sheets and audio files were transcribed into an electronic text format, so the corpus includes three types of materials; (1) the hand-written sheets scanned in PDF format, (2) the audio files as MP3 files and (3) the data collected electronically by the online form with transcripts of the hand-written sheets and audio recordings. The third type has been stored into a database which can generate the texts in TXT and XML files.

Two guides were created describing the ALC data capture method of written and spoken data, which have been designed to be followed by both researcher and research assistants. As for the tasks, a paper form was created and distributed to learners in schools and departments where there were no computer labs. For this type of form a post-process was required to transcribe the texts learners produced into a computerised format. An online equivalent form was also created for use in schools and departments that had labs, so learners' texts could be included in the corpus without the need to carry out the transcribing process. Data collection of the ALC took four months to conclude, one month for version 1 at the end of 2012 and a further 3 months for version 2 in 2013.

3.6 Metadata

The ICLE (the International Corpus of Learner English) corpus's learner profile questionnaire (Granger, 1993) 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, as it was felt that learners may produce more than one piece of text. Some questions about personal details were omitted such as father's mother tongue or mother's mother tongue, etc. The corpus metadata includes 26 elements, 12 related to the learner and 14 associated with the text (see table 2). More details about these elements can be found in the ALC Content section.

Table 2 Metadata elements used in ALC

Learner metadata	Text metadata
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. Timed or not timed task
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

3.7 Text Distribution

In the design criteria for the corpus, texts were distributed based on four variables, nativeness of the learner (native vs. non-native), general level of education (pre-university vs. university), gender (males vs. females), and materials mode (written vs. spoken), see Table 3. Well-designed learner corpora, like ASU (Hammarberg, 2010), BAWE (Heuboeck *et al.*, 2008), and MICUSP (O'Donnell & Römer, 2009), include a balance of contributions from both native and non-native learners. The ALC aims for the same target. However, more focus was placed on the pre-university level (70%) because of the number of learners had the potential of being recruited from this level. Special considerations were given to the gender component of the research due to the fact that in Saudi Arabia, apart from pre-school establishments, all other education delivery is made to single gender classes, i.e. 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 and thus a female representative was essential to collect the required data, hence the portion devoted to data concerning females was 20%. With respect to the materials mode due to the fact that the spoken corpus is a laborious and time consuming, 10% of the ALC was devoted to the oral data.

Table 3 Distribution of words based on 4 variables in the ALC design criteria*

Learners' nativeness	Learners' general level of education	Learners' gender	Materials mode
Native 50% 100,000 words	**Pre-university** 70% 140,000 words	**Male** 80% 160,000 words	**Written** 90% 180,000 words
Non-native 50% 100,000 words	**University** 30% 60,000 words	**Female** 20% 40,000 words	**Spoken** 10% 20,000 words

* The actual data collected in each variable is shown in section IV

IV ALC content

As described, the ALC is a project of comprising of a collection of written and spoken materials from learners of Arabic in Saudi Arabia. The corpus includes 282732 words, 1585 materials (written and spoken), produced by 942 students from various first language (L1) backgrounds. In this section, the corpus content will be illustrated in more detail based on the metadata elements.

4.1 Learners Metadata

Age: Ages of the learners ranged between 16 and 42, however the majority were between 16 and 25 (Fig. 1).

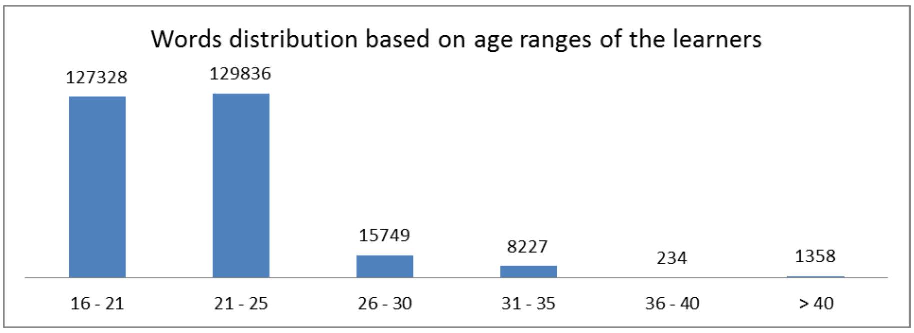

Fig. 1 Words distribution based on age ranges of the learners

Gender: Although two thirds of the participants were males whilst 33% were females (Fig. 2), the data collected from both was more than what had been projected in the design criteria, (Male: 160,000 and Female: 40,000 words).

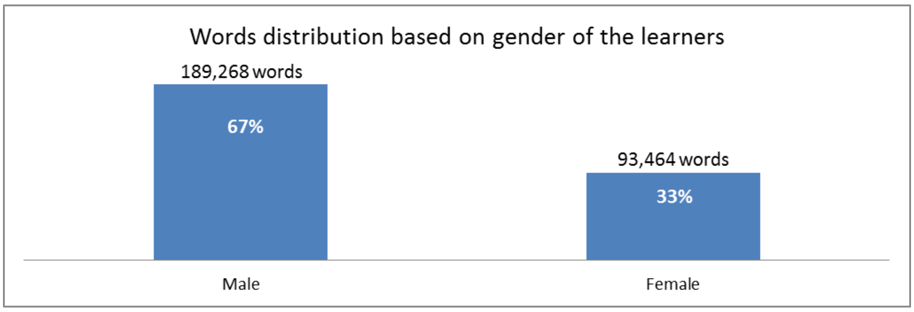

Fig. 2 Words distribution based on gender of the learners

Nationality: Nationalities of the students represented 67 different countries.

Mother Tongue: The ALC contains 66 different mother tongue representations, 65 of them are spoken by the non-native Arabic-speaking learners and the Arabic language is the L1 of all of the Arabic native speakers.

Nativeness: The ALC design criteria indicate 50% of the corpus data for NAS and the same for NNAS. The actual data collected from both groups was still close to the target percentages projected (Fig. 3), and the number of words was more than what had been selected in the design criteria, (NAS: 100,000 and NNAS: 100,000 words).

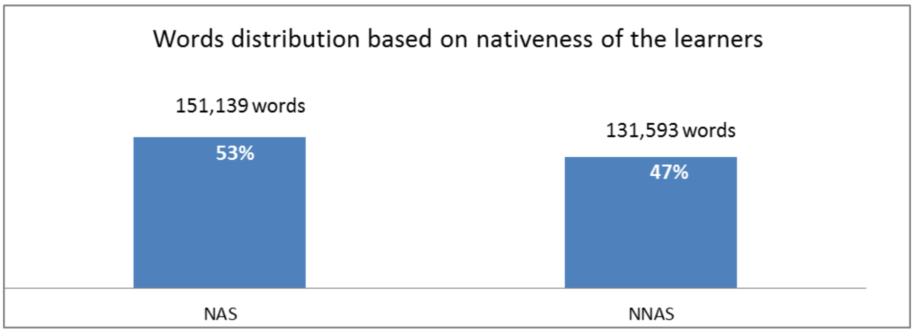

Fig. 3 Words distribution based on nativeness of the learners

Number of Languages Spoken: the number of languages spoken by each learner raged from 1 to 10 in the case of NNAS, while NAS spoke between 1 and 4 languages.

Number of Years Learning Arabic: Learners spent between a few months (indicated as 0 years in the corpus) and 19 years in their acquisition of Arabic. The native Arabic speakers were excluded from this category.

Number of Years Spent in Arabic Countries: 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 right up to 21 years. NAS were also excluded from this category. In the corpus's learner profile questionnaire, this item and the previous one were allocated to NNAS.

General Level: As explained in the design criteria of the ALC, 70% (140,000 words) of the data was devoted to learners from pre-university level and 30% (60,000 words) were from university level. However, the percentage of the actual data was 80% for Pre-university and 20% for University, though the number of words was larger in the former level, and near the target in the latter (Fig. 4).

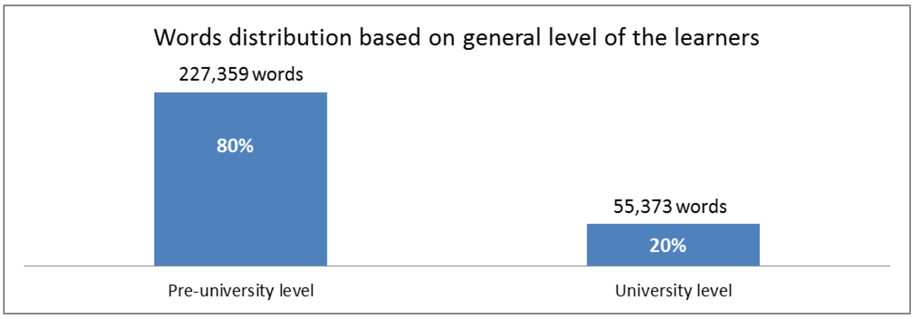

Fig. 4 Words distribution based on general level of the learners

Level of Study: The ALC includes five levels of study; secondary school, general language course, diploma programme (advanced language course), Bachelors degree and Masters. Learners of both the BA and MA were majoring in Arabic.

Year/Semester: The level of study was represented by a range of three years for the secondary school students and eight semesters for the other groups (general and advanced language courses, BA and MA).

Educational Institution: The corpus participants were affiliated to twenty-five educational institutions, i.e. secondary schools and universities.

4.2 Texts Metadata

Text genre: two genres were covered in the ALC, *Narrative* (67% of the ALC content) and *Discussion* (33%).

Where produced: 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%).

Year of production: The first version of the ALC was collected in 2012 (12%), whereas the v2 data was collected in 2013 (88%).

Country of production: The entire corpus data was produced in Saudi Arabia.

City of production: The ALC data was collected from eight cities, Riyadh (78%), Alqatif (9%), Makkah (4%), Jeddah (3%), Alkharj (3%), Aljesh (2%), Hafr Albatin (1%)and Mahayil Asir (1%).

Timed: Timing was based on the location of the material being produced, all the materials produced in class were timed (69%), those produced at home were not timed (31%).

References use: References were used in 5% of the corpus data, this includes the following four types:

Grammar book use: Grammar books were used in 2% of the corpus data.

Monolingual dictionary use: Monolingual dictionaries were used in 1% of the ALC.

Bilingual dictionary use: Bilingual dictionaries were used in 2% of the corpus.

Other references use: 2% of the corpus includes the use of other references.

Text mode: 93% of ALC is written data, whilst 7% is spoken material.

Text medium: The corpus includes two mediums of written data, text produced by hand (76%) and text produced on a computer (17%). Auditory data was collected in one medium only as recorded interviews (7%).

Text length: The average length of the texts in the ALC is 178 words. More details about the average length based on some factors are listed in Table 4.

Table 4 Average length of the ALC texts based on some key factors

Factor		
Learners' gender	Males 166	Females 209
Learners' nativeness	NAS 191	NNAS 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

V Corpus Availability

As the ALC is an open source project, the entire data of the corpus is available to download from the following website <http://www.arabiclearnercorpus.com>. It is available in different formats as shown in Table 5. In addition, the user has the choice to download the whole corpus in one file (TXT or XML format), or to have each text in a separate file, 1585 files exist in the current version.

Table 5: File formats available on the ALC website

Data Mode	Type	Format available
Written	Hand-written sheets	PDF
	Transcripts of the hand-written sheets	TXT and XML (with Arabic and English metadata)
Spoken	Audio files	MP3
	Transcripts of the audio files	TXT and XML (with Arabic and English metadata)

VI Conclusion and Further Work

The paper has explored the project of compiling the Arabic Learner Corpus, which is available for public use by Arabic language researchers. A number of relevant learner corpora were reviewed. The paper also showed the ALC design criteria which covered the target language, participants, corpus size, materials included, method and tasks used for data collection, as well as metadata of both the corpus materials and contributors. Details about the content of the current version of the ALC were presented based on 26 elements representing the corpus metadata. Further work has been done to annotate the corpus, e.g. a second version of the error tagset for Arabic corpora with an error tagging manual. Additionally, a computer-aided tagging tool is currently being developed.

References

Abuhakema, G., Feldman, A., & Fitzpatrick, E. (2008). *Annotating an Arabic learner corpus for error. The proceedings of the International Conference on Language Resources and Evaluation (LREC 2008).* May 26 – June 1, Marrakech, Morocco.

Alfaifi, A. (2011). *The attitude of ASL learners in Saudi Arabia towards printed and electronic dictionaries.* (Unpublished master's thesis), University of Essex.

Branbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.

Farwaneh, S., & Tamimi, M. (2012). Arabic learners written corpus: A resource for research and learning. Retrieved September 2, 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, 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: Rodopi

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., & Dumont, A. (2012). *Learner corpora around the world.* Retrieved 17 July 2012, from: the Université Catholique de Louvain, Centre for English Corpus Linguistics Web site: http://www.uclouvain.be/en-cecl-lcworld.html

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.

Hassan, H., & Daud, N. (2011). *Corpus analysis of conjunctions: Arabic learners' difficulties with collocations*. Paper presented at the Workshop on Arabic Corpus Linguistics (WACL), 11th -12th

April 2011, Lancaster University, UK. http://ucrel.lancs.ac.uk/wacl/slides-HASSAN-DAUD.pdf

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

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). London: Longman

O'Donnell, M. B. , & Römer, U. (2009). Michigan corpus of upper-level student papers. Retrieved 27 July 2012, from: http://micusp.elicorpora.info/

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal, 26*: 81-114.

Sinclair, J. (2005). Corpus and text - Basic principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books

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