

User's Guide

To the search website of the Arabic Learner Corpus

www.alcsearch.com

The screenshot shows the Arabic Learner Corpus search website interface. At the top, there is a header with the ALC logo and the text "ARABIC LEARNER CORPUS" and "المجموعة اللغوية للتعلم العربية". Below the header, there is a navigation bar with "عربي" and a home icon. The main content area is divided into several sections:

- Left sidebar:** Contains information about the corpus size (1585 texts), download options (Plain text with no metadata, Plain text with Arabic metadata, Plain text with English metadata, XML with Arabic metadata, XML with English metadata, Hand written sheets in PDF, Audio recordings in MP3), and search determinants (AGE, GENDER, NATIONALITY, MOTHER TONGUE, NATIVENESS, NUMBER OF LANGUAGES SPOKE, NUMBER OF YEARS LEARNING ARABIC).
- Top right:** A welcome message and a link to the user guide.
- Search bar:** A text input field with a "Search" button.
- Search result:** A table showing search results. The table has two columns: "Text ID" and "Concordance". The results are sorted by relevance, and the first 16 results are shown out of a total of 1139 results.

Text ID	Concordance
S001_T1_M_PRE_NNAS_W_C	الرحلة إلى القرية لزيارة ذوي
S002_T1_M_PRE_NNAS_W_C	ناولنا الطيور وحسبنا مسؤلوا الرحلة مجموعات حتى تكون الرحلة
S002_T1_M_PRE_NNAS_W_C	الرحلة مجموعات حتى تكون الرحلة منظمة بعد ذلك ركبا
S002_T1_M_PRE_NNAS_W_C	دعاء السفر وصحبا مسؤول الرحلة نمانح معده وصنبا في
S002_T1_M_PRE_NNAS_W_C	الله تعالى فكانت هذه الرحلة مشاركة لأثنا رحلة الطاعة
S002_T1_M_PRE_NNAS_W_C	أذكر كثيرا من تفاصيل الرحلة لعسل الوب
S004_T1_M_PRE_NNAS_W_C	وتطلت من المشاركين في الرحلة إلى اليوم الموجود وركت
S005_T1_M_PRE_NNAS_W_C	الرحلة إلى المدينة المنورة ذهبت

Version 1.0

ABDULLAH ALFAIFI

2015

Table of contents

1	Introduction	2
2	Definitions	3
3	Benefits of using determinants	4
4	Table of determinants	5
5	Website sections	7
5.1	'Sign up' and 'Log in' page.....	8
5.1.1	Sign up.....	8
5.1.2	Log in	9
5.1.3	Reset password	10
5.2	Search page	12
5.2.1	Determinants	13
5.2.2	Search.....	16
5.2.3	Results	17
5.2.4	File downloads	20
6	Conclusion.....	23
7	Appendix.....	24
7.1	File types included in the corpus	24
7.2	Examples of the corpus files	25

1 Introduction

This guide presents an overview of the first version of the search website of the Arabic Learner Corpus (ALC), with an illustration of how to use it and to take advantage of its functions.

The search website aims to enable users to search the ALC based on the determinants, as the corpus design criteria include 26 determinants that can be utilised to conduct comparisons between different groups of learners or texts. Additionally, the user can download a subset of the corpus data (sub-corpus) based on those determinants.

The search website was created independently from the ALC main website (<http://www.arabiclearnercorpus.com>); the latter contains details about the corpus, developers, publications and other information.

2 Definitions

This section presents some terms and indicates what they mean in this guide.

Corpus:

This indicates the Arabic Learner Corpus (ALC); it contains a collection of written essays and spoken recordings categorised under two topics: 'A vacation trip' (narrative) and 'My study interest' (discussion) by learners of Arabic in Saudi Arabia in 2012 and 2013. The corpus includes 282,732 words, 386,571 tokens, 29,627 types and 1,585 materials. It was produced by 942 students, of 67 nationalities and from 66 different L1 backgrounds, studying at pre-university and university levels. The average length of a text is 178 words. Version 2.0 of the ALC contains raw data that includes three parts: transcriptions of handwriting (76%); writing done on a computer (17%); and transcriptions of audio recordings (7%).

Search website:

This is the website created to search the ALC; it can be accessed from <http://www.alcsearch.com>.

Determinants:

They are a number of learners'/texts' features that can be selected to search a subset of the corpus data (sub-corpus), such as 'Age', 'Gender', 'Mother tongue', 'Text mode', 'Place of writing', etc. The corpus has 26 determinants: 12 related to the learners and 14 related to the texts.

Sub-corpus:

A sub-corpus means — as mentioned in the determinants above — a subset of data selected based on the determinants in order to search this sub-corpus separately. For example, the part of the data written by male learners can be a sub-corpus. Also, the part produced by a particular nationality is a sub-corpus, and so on.

3 Benefits of using determinants

1. To search any sub-corpus based on the determinants required.

Examples: searching the sub-corpus of non-native speakers of Arabic only or the sub-corpus of spoken data, and so on.

2. To compare the results of two sub-corpora (two comparable groups).

Examples: comparing the non-native speakers of Arabic to those native speakers, or comparing the learners at the pre-university level to those at the university level.

3. To download a specific part of the corpus in different formats (TXT, XML, PDF, MP3).

Examples: texts produced by the native speakers of Arabic or texts written by hand by a specific nationality, etc., can be downloaded separately.

4 Table of determinants

Learners' determinants

	Determinant	Comparisons
1	Age	To compare different groups of learners based on their ages (from 16 to 42)
2	Gender	To compare males to females
3	Nationality	To compare learners from 66 nationalities
4	Mother tongue	To compare learners from 67 L1s
5	Nativeness	To compare non-native speakers of Arabic to native speakers
6	No. of languages spoken	To compare different groups of learners based on the number of languages they speak (from 1 to 10)
7	No. of years learning Arabic	To compare different groups of learners based on the number of years they spent learning Arabic (from 0 to 21)
8	No. of years in Arab countries	To compare different groups of learners based on the number of years they spent in Arab countries (from 0 to 19)
9	General level of study	To compare learners from two general levels (pre-university and university)
10	Programme of study	To compare learners from five programmes of study: <ul style="list-style-type: none"> • General language course • Diploma language course • Secondary school • Bachelor's • Master's
11	Year or semester	To compare different groups of learners based on their year or semester when the texts were written
12	Educational institution	To compare learners from 25 different educational Institutions (university, language institute and secondary school)

Texts' determinants

	Determinant	Comparisons
1	Text genre	To compare two types of genres (narrative and discussion)
2	Place of writing	To compare two places of writing (in class and at home)
3	Year	To compare texts produced in two years, 2012 and 2013
4	Country	All the current texts were collected from Saudi Arabia; when more texts are collected from different countries, these countries will be added for comparison
5	City	To compare texts collected from 8 different cities
6	Timing	To compare the timed texts (about an hour) to those untimed (one day or more)
7	References use	To compare texts produced using references to those produced with no references
8	Grammar books use	To compare texts produced using grammar books to those produced without
9	Monolingual dictionaries use	To compare texts produced using monolingual dictionaries to those produced without
10	Bilingual dictionaries use	To compare texts produced using bilingual dictionaries to those produced without
11	Other references use	To compare texts produced using other references to those produced without
12	Text mode	To compare two text modes (written and spoken)
13	Text medium	To compare three text media <ul style="list-style-type: none"> • written by hand • written on computer • recorded interview
14	Text length	To compare different groups of texts based on their lengths (number of words)

5 Website sections

'Sign up' and 'Log in' page

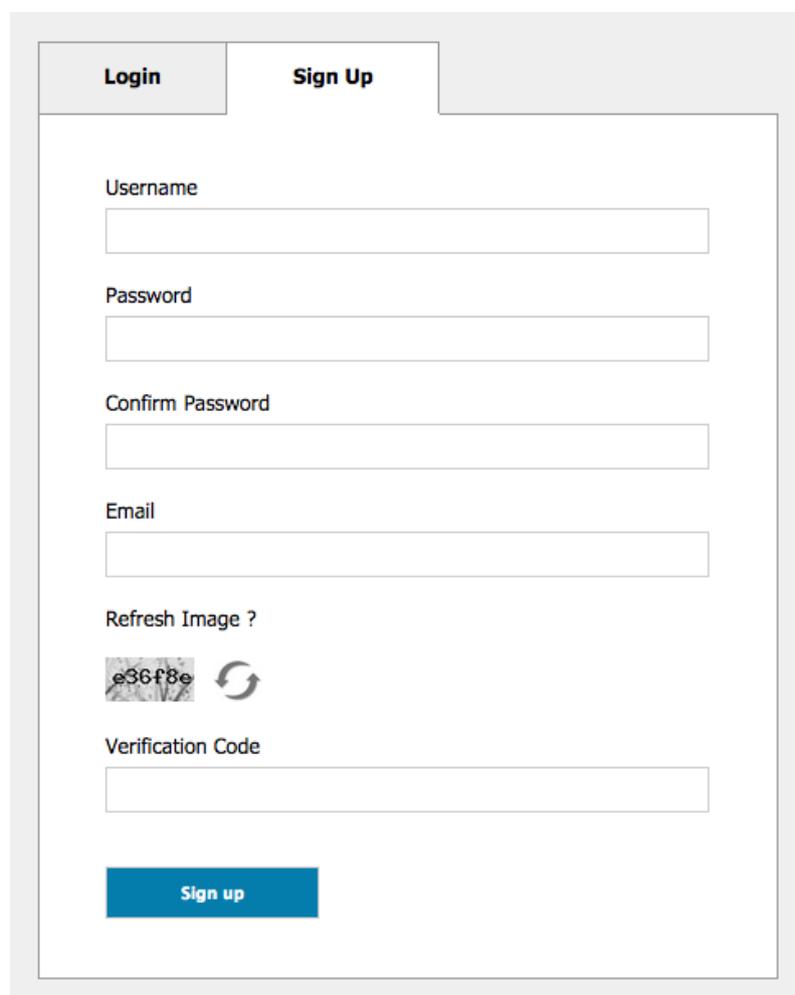
- Sign up
- Log in
- Password reset

Search page

- Determinants
- Search
- Results
- File downloads

5.1 'Sign up' and 'Log in' page

5.1.1 Sign up



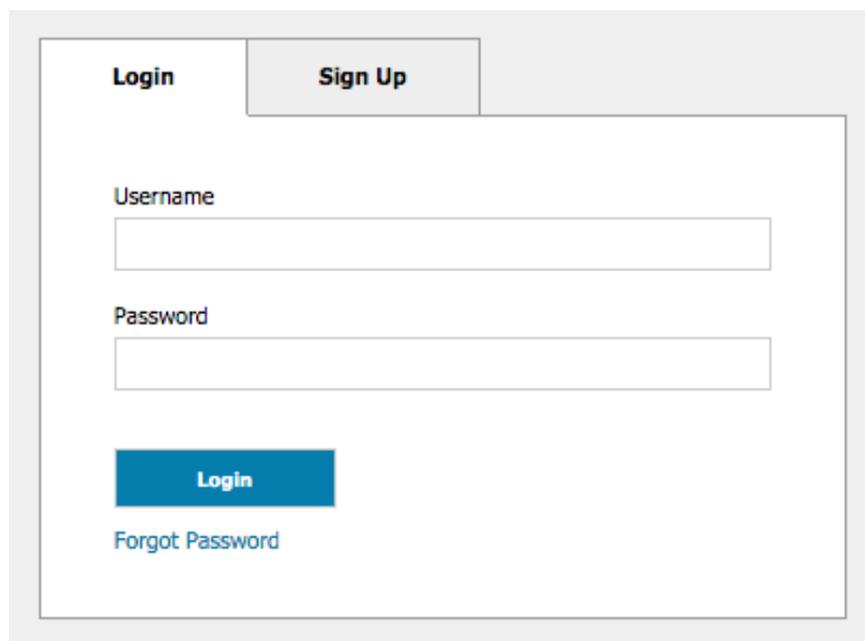
The screenshot shows a web form for signing up. At the top, there are two tabs: 'Login' and 'Sign Up', with 'Sign Up' being the active tab. The form contains the following fields and elements:

- Username:** A text input field.
- Password:** A text input field.
- Confirm Password:** A text input field.
- Email:** A text input field.
- Refresh Image ?**: A label above a small image containing the alphanumeric code 'e36f8e' and a circular refresh icon.
- Verification Code:** A text input field.
- Sign up:** A blue button at the bottom of the form.

This page is used to create a new user account. It is important to enter a valid email in case of forgetting the password. After creating a user account, the user will need to enter his username and password each time he logs in to the website. Latin characters should be used in both the username and password. Using other types of characters may cause an error.

5.1.2 Log in

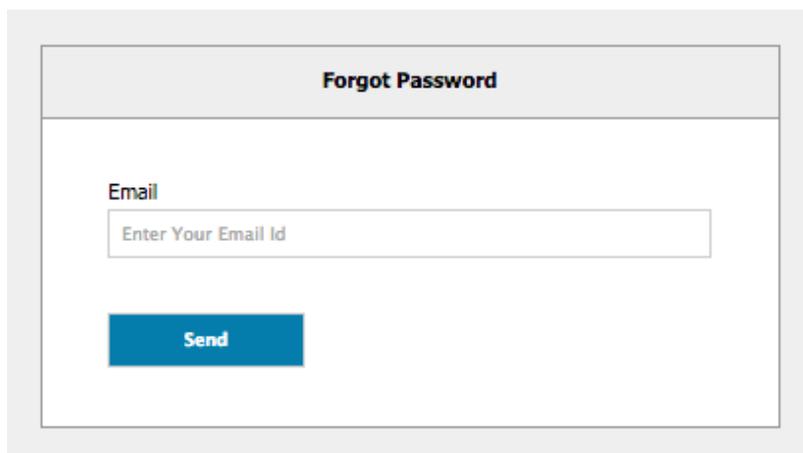
The user can log in to the website by entering the username and password correctly into the following form.

The image shows a login form interface. At the top, there are two tabs: 'Login' (which is active) and 'Sign Up'. Below the tabs, there are two input fields: 'Username' and 'Password'. Below the 'Password' field, there is a blue 'Login' button and a link labeled 'Forgot Password'.

After logging in, the user will be redirected to the main page of searching the corpus data. If the log in information is not correct, the user will be warned about this. When the password is forgotten, it can be reset by clicking on '[Forgot Password](#)' at the bottom of the form.

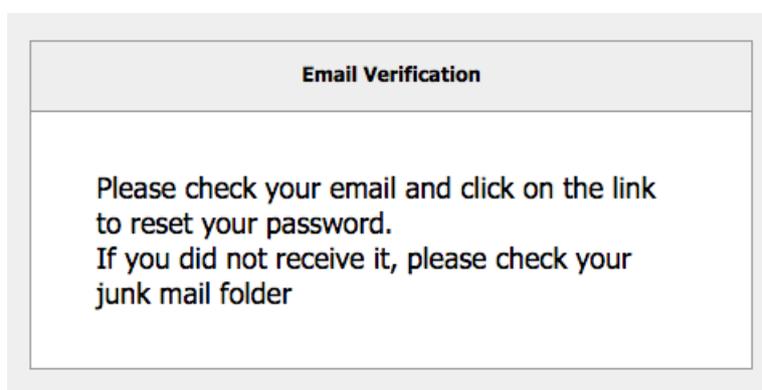
5.1.3 Reset password

If the user forgot his password, it can be reset using the '**Forgot Password**' link at the bottom of the Log in form. Clicking on this link shows the following window, which asks the user to enter his email that he used when he signed up.



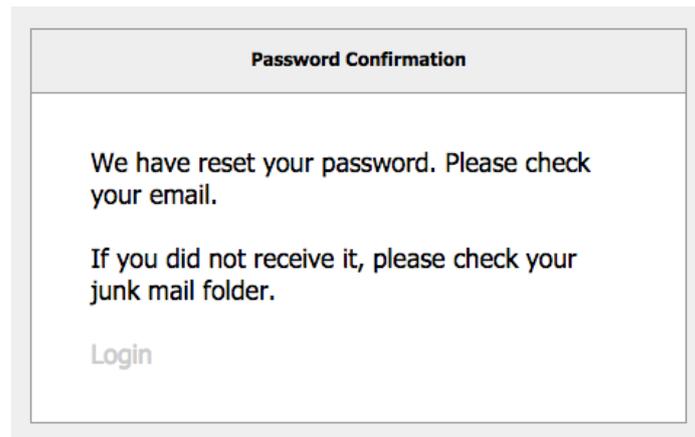
The screenshot shows a web form titled "Forgot Password". It contains a text input field labeled "Email" with the placeholder text "Enter Your Email Id". Below the input field is a blue button labeled "Send".

After entering the email and clicking on '**Send**', a link will be sent to reset the password and the following message will be shown.

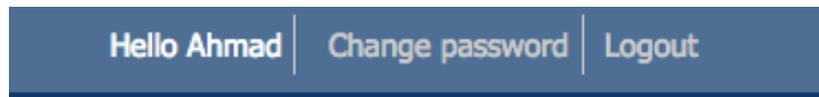


The screenshot shows a message box titled "Email Verification". The text inside reads: "Please check your email and click on the link to reset your password. If you did not receive it, please check your junk mail folder".

By clicking on the reset link sent by email, a new password will be sent to the email address, and the following message will be shown.



The user now can use the new password to log in again and then change the password from the link “[Change password](#)” at the top of the page.



5.2 Search page

The search page consists of four main sections, as follows.

The screenshot shows the Arabic Learner Corpus search website interface. The page is divided into four main sections, each indicated by a red bracket and a label:

- Files download:** Located at the top left, it displays the total number of corpus texts (1585) and the number of texts based on the current selection (1585). It offers options to download these texts in various formats: Plain text with no metadata, Plain text with Arabic metadata, Plain text with English metadata, XML with Arabic metadata, XML with English metadata, Hand written sheets in PDF, and Audio recordings in MP3. There are 'Download' and 'Clear' buttons.
- Search:** Located at the top right, it features a search bar with the text 'الرحلة' and a 'Search' button. Below the search bar, there is a 'Separate Words' checkbox.
- Determinants:** Located on the left side, it contains a 'Search Determinants' section with a 'Clear All Determinants' button. Below this, there is a list of 25 determinants, each with a dropdown arrow: AGE, GENDER, NATIONALITY, MOTHER TONGUE, NATIVENESS, NUMBER OF LANGUAGES SPOKE, NUMBER OF YEARS LEARNING ARABIC, NUMBER OF YEARS SPENT IN ARABIC COUNTRIES, GENERAL LEVEL OF EDUCATION, LEVEL OF STUDY, YEAR OR SEMESTER, EDUCATIONAL INSTITUTION, TEXT GENRE, PLACE OF WRITING, YEAR OF WRITING, COUNTRY OF WRITING, CITY OF WRITING, TIMING, REFERENCES USE, GRAMMAR BOOKS USE, MONOLINGUAL DICTIONARIES USE, BILINGUAL DICTIONARIES USE, OTHER REFERENCES USE, TEXT MODE, TEXT MEDIUM, and TEXT LENGTH.
- Results:** Located on the right side, it displays the search results. At the top, it says 'Welcome to the Arabic Learner Corpus search website' and '[USER GUIDE]'. Below this, it shows 'Search result' and '1-16 of total 1139 results'. A table with two columns, 'Text ID' and 'Concordance', lists the search results. The first row is S001_T1_M_PRE_NNAS_W_C with the concordance 'الرحلة إلى القرية لزيارة ذوي'. Below the table, there is a 'Page 1 of 72' indicator and navigation arrows. At the bottom, there is a preview of the selected result S015_T1_M_Pre_NNAS_W_C, showing a snippet of text from the corpus.

These sections will be illustrated in the following order:
Determinants, Search, Results and File downloads.

5.2.1 Determinants

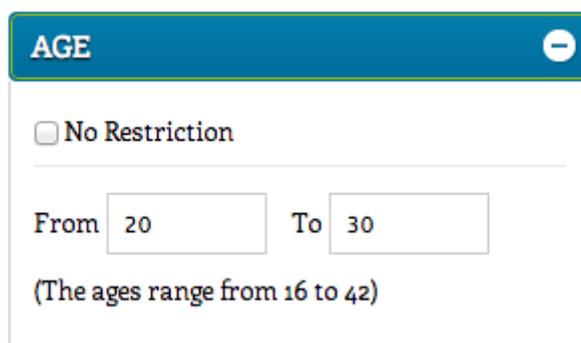
The user can search any sub-corpus by selecting different values for the determinants (see the determinants table on pages 4 and 5).

The determinants can be classified into three types:

1. Numerical range

This type requests two values, the maximum and minimum of the range.

Example: 'Age'; the user can select a range of learners' ages between 20 and 30 years.



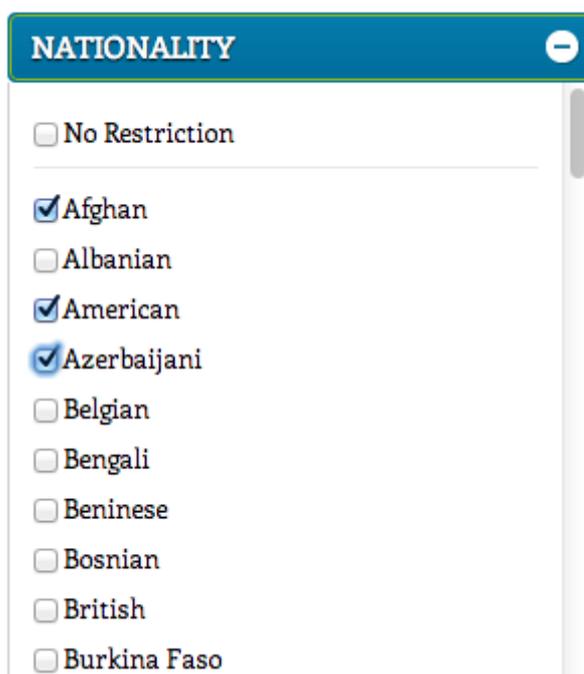
The screenshot shows a search filter for 'AGE'. At the top, there is a blue header with the text 'AGE' and a minus sign icon. Below the header, there is a checkbox labeled 'No Restriction'. Underneath, there are two input fields: 'From' with the value '20' and 'To' with the value '30'. At the bottom of the filter, there is a note in parentheses: '(The ages range from 16 to 42)'.

This type of determinant accepts only values in the Arabic numeral system (1, 2, 3, 4, 5, 6, 7, 8, 9 and 0).

2. Multi-selections list

One or more options can be selected from this type of list.

Example: 'Nationality'; the user can select one or more nationalities to search the sub-corpus of those learners only.



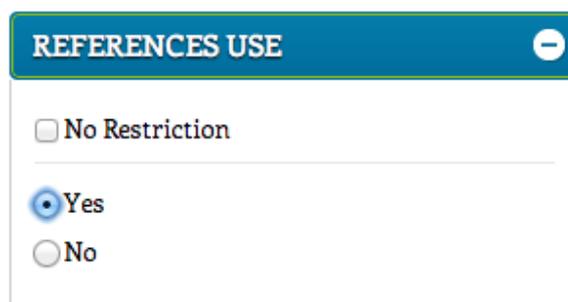
NATIONALITY

- No Restriction
- Afghan
- Albanian
- American
- Azerbaijani
- Belgian
- Bengali
- Beninese
- Bosnian
- British
- Burkina Faso

3. Two-options list ('yes' or 'no')

Only one choice can be selected from this type of list.

Example: 'References Use'; the user can select whether texts were produced using any language references by choosing 'Yes', or whether no references were used by selecting 'No'.



REFERENCES USE

- No Restriction
- Yes
- No

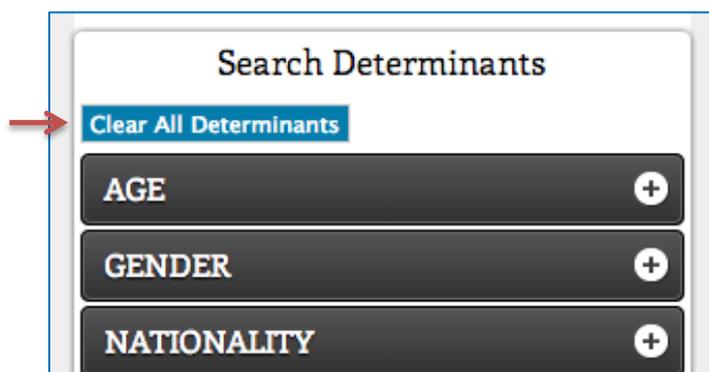
Clear the determinants

The values of any determinant can be cleared by clicking on 'No Restriction' at the top of options list of each determinant; doing this will reset the value of the selected determinant only.

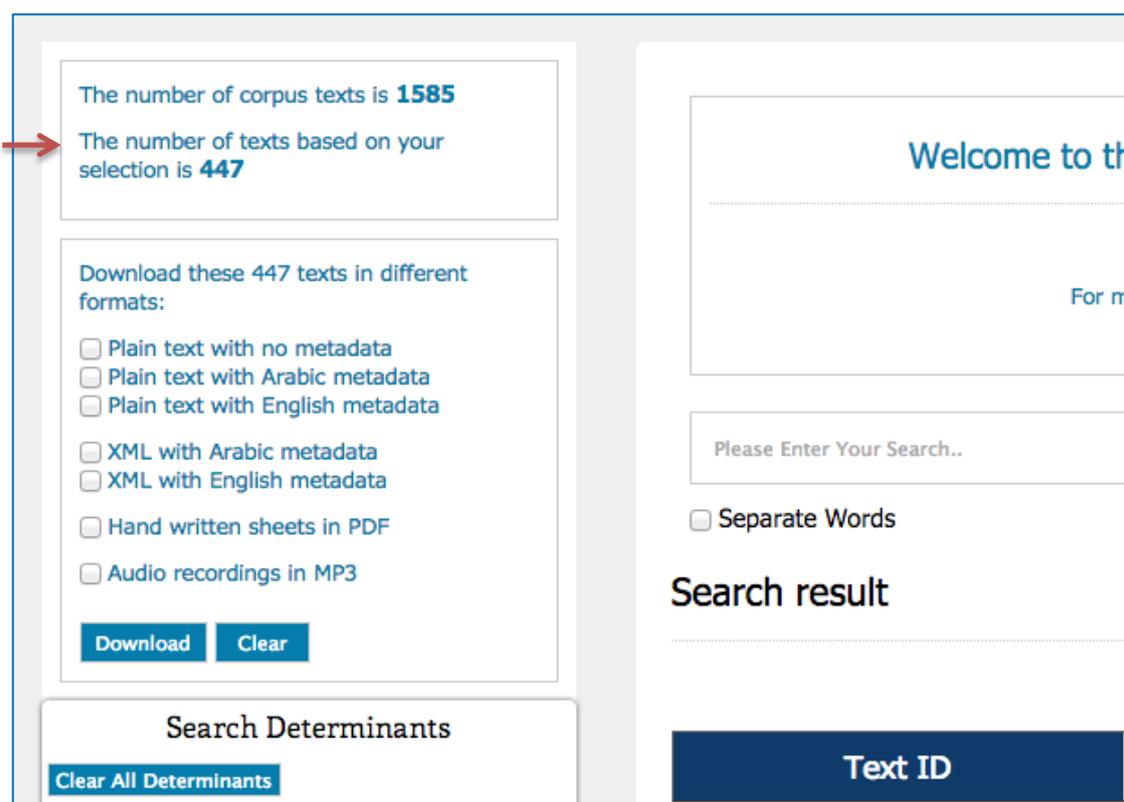


No Restriction

To clear the values of determinants all at once, the user can click on 'Clear All Determinants' at the top of the determinants list.



By selecting any option from the determinants, the number of texts available based on the new selection will be shown above the files-download section.



5.2.2 Search

The search section contains a text box for entering a search word and a “Search” button for showing the results.



The screenshot shows a search interface with a text input field containing the placeholder text "Please Enter Your Search..". To the right of the input field is a blue button labeled "Search". Below the input field is a checkbox labeled "Separate Words".

Please notice that the current version of the site cannot process more than one form of a word; if two or more forms are entered, no results will appear. The capability of processing more than one form will be added to one of the future versions.

When searching for a word such as 'كيف' [*how* in English], the results will include all examples where the search word appears, whether as independent word or the word with prefixes and/or suffixes, as shown below.



The screenshot shows search results for the word 'كيف' (how). The results are displayed in a list of six sentences, with the word 'كيف' highlighted in yellow in each sentence:

- لأنه شيء أستمتع بسماعها وكيف يكون القيام بها، أتمنى
- عن الحج وأما قرأت كيفية الحج وشروط الحج لأكون
- ذلك قراءة شروط العمرة، وكيفيتها وصفتها، واستعددت الطعام والملابس
- الحج كنت أتعلم عن كيفية أداء الحج بطريقة الصحيحة
- الشبكة الدولية لكي أعرف كيفية أداء الحج بدقة حتى
- راشد وعليكم السلام ماجد كيف حالك راشد انا بخير

The choice ‘Separate Words’ can be selected, which exists under the search box to show only those examples that include the search word independently. Once it is selected, all results with prefixes and/or suffixes will be excluded, and will look like the following:

الأولى إلى مكة المكرمة كيف منظر الحرم؟ هل الناس
فيبقى عند الأسئلة المهمة كيف أنال هذا وذاك؟ وبما
ما رأته قبله مثلاً كيف تزيّن الفيروز على الجدار
فاستشار الملك مع الوزراء كيف تبقى المحبة بعد وفاتها؟
البيئة ولهجات أهل مصر كيف تنطق في اللغة العربية
على نزول القرآن الكريم كيف نزل قديماً حتى وصل

5.2.3 Results

The results section consists of some sub-sections, as follows:

The screenshot shows a search results page with the following components and annotations:

- Number of results:** A bracket on the left points to the text "1-16 of total 1139 results" at the top of the results table.
- Moving between the results' pages:** A bracket on the left points to the pagination controls at the bottom of the table, showing "Page 1 of 72" and navigation arrows.
- Concordances with the ID of each instance:** A bracket on the right points to the "Text ID" column of the table.
- Print/Download the results:** A bracket on the right points to the print and download icons at the top right of the results table.
- The full text of an instance:** A bracket on the right points to the expanded view of the text for instance ID "S015_T1_M_Pre_NNAS_W_C" at the bottom of the page.

Text ID	Concordance
S001_T1_M_PRE_NNAS_W_C	الرحلة إلى القرية لزيارة ذوى
S002_T1_M_PRE_NNAS_W_C	تناولنا العطور وقسمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة
S002_T1_M_PRE_NNAS_W_C	الرحلة مجموعات حتى تكون الرحلة منظمة بعد ذلك ركنا
S002_T1_M_PRE_NNAS_W_C	دعاء السفر ونصائح مسؤول الرحلة نصائح مفيدة ومضنا في
S002_T1_M_PRE_NNAS_W_C	الله تعالى فكانت هذه الرحلة مشاركة لأنها رحلة الطاعة
S002_T1_M_PRE_NNAS_W_C	أذكر كثيرا من تفاصيل الرحلة لضيق الوقت
S004_T1_M_PRE_NNAS_W_C	وتعلت من المشاركين في الرحلة أنى اليوم الموعود وركبت
S005_T1_M_PRE_NNAS_W_C	الرحلة إلى المدينة المنورة ذهبت
S005_T1_M_PRE_NNAS_W_C	مع أصحابي أعجبتني هذه الرحلة مدينة رسول صلى الله
S005_T3_M_PRE_NNAS_W_H	الرحلة إلى المدينة المنورة ذهبت
S010_T1_M_PRE_NNAS_W_C	الرحلة إلى بلدي أنا طالب
S010_T1_M_PRE_NNAS_W_C	وتفلة وقد تعبنا في الرحلة ولكن عندما وصلنا إلى
S011_T1_M_PRE_NNAS_W_C	في ذاك الأسبوع أخذت الرحلة عشر ساعات وعندما وصلنا
S015_T1_M_PRE_NNAS_W_C	الرحلة إلى مكة المكرمة والمدينة
S015_T3_M_PRE_NNAS_W_H	الرحلة إلى مكة المكرمة والمدينة
S017_T1_M_PRE_NNAS_W_C	الرحلة إلى الحرمين بسم الله

Page 1 of 72

S015_T1_M_Pre_NNAS_W_C

الرحلة إلى مكة المكرمة والمدينة المنورة في يوم اثنان من أسبوع ماضى، سافرت إلى مكة المكرمة في الساعة ثلاثة والنصف بعد الصلاة العسر مع الأصدقاء، ركنا الحافلة كلمة غير معروفة، وصلنا إلى الميقات في الساعة الثامن ونصف ليلاً ثم لبسنا صلبنا الصلاة المفترت وأنعمنا، في الميقات قصر بعد ذلك لبسنا الأملانس الحرام، فلنا لبك عمرة ثم وجدنا إلى مكة المكرمة وصلنا إلى مكة المكرمة في الساعة الثانية صباح قبل وصل ذهبنا إلى المطعم ليكل الطعام العشاء، وجدنا إلى البيت الحرام لصلاة الصبح، ثم عمرة بعد ذلك، وجدنا إلى المدينة المنورة من مكة إلى المدينة فصبنا سنة الساعة المدينة المنورة فيها المسجد النبوي

The top part shows the total number of results and how many are displayed in the page.

1-16 of total 1139 results

All these results can be printed or downloaded in an Excel file format (.xls) using the print and download buttons.



In the concordance part, all results are shown by highlighting the search word in a different colour. The part on the left side shows the text IDs of each example.

Text ID	Concordance
S001_T1_M_PRE_NNAS_W_C	الرحلة إلى القرية لزيارة ذوي
S002_T1_M_PRE_NNAS_W_C	تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة
S002_T1_M_PRE_NNAS_W_C	الرحلة مجموعات حتى تكون الرحلة منظمة بعد ذلك ركنا
S002_T1_M_PRE_NNAS_W_C	دعاء السفر ونصحنا مسؤول الرحلة نصائح مفيدة ومضيا في
S002_T1_M_PRE_NNAS_W_C	الله تعالى فكانت هذه الرحلة مباركة لأنها رحلة الطاعة
S002_T1_M_PRE_NNAS_W_C	أذكر كثيرا من تفاصيل الرحلة لصيق الوقت
S004_T1_M_PRE_NNAS_W_C	وجُعِلت من المشاركين في الرحلة أتى اليوم الموعود وركبت
S005_T1_M_PRE_NNAS_W_C	الرحلة إلى المدينة المنورة ذهبت
S005_T1_M_PRE_NNAS_W_C	مع أصحابه أعجبتني هذه الرحلة مدينة رسول صلى الله
S005_T3_M_PRE_NNAS_W_H	الرحلة إلى المدينة المنورة ذهبت
S010_T1_M_PRE_NNAS_W_C	الرحلة إلى بلدي أنا طالب
S010_T1_M_PRE_NNAS_W_C	وثقيلة وقد تعبنا في الرحلة ولكن عندما وصلنا إلى
S011_T1_M_PRE_NNAS_W_C	في ذاك الأسبوع أخذت الرحلة عشر ساعات وعندما وصلنا
S015_T1_M_PRE_NNAS_W_C	الرحلة إلى مكة المكرمة والمدينة
S015_T3_M_PRE_NNAS_W_H	الرحلة إلى مكة المكرمة والمدينة
S017_T1_M_PRE_NNAS_W_C	الرحلة إلى الحرمين بسم الله

The user can move between the results pages via the buttons located under the concordance part.

Page 4 of 72

The full text of any example can be displayed by clicking on the highlighted search word in any example; a text box will appear at the bottom of the page showing the full text and highlighting all the word matches.

S060_T1_M_Pre_NAS_W_C

رحلة إلى بريطانيا بسم الله الرحمن الرحيم سوف أقسمها إلى ثثة أقسام 1 قبل **الرحلة** 2 أثناء **الرحلة** 3 بعد **الرحلة** وأبدأ أولاً بما قبل **الرحلة** عندما دعاني والدي ليخبرني بأنه رتب أمور السفر وكانت السفارة لي وحدي فوجأت فكرت خفت قليلاً من الغربة ومسألة إذا رجعت ولم أستفد شيئاً فسيأبيني ضميري أني خذلت والدي ولكن تشجعت ووافقت وقرب وقت **الرحلة** الدراسية في بلاد الغرب سرعان ما جاء وقت إقلاع الطائرة وكانت **الرحلة** شبه طويلة 6 ساعات متواصلة ولكن الوقت يمر مرور البرق حتى وإن كان في هذي الفترة من سيطرة لعب أو للمعاناة فهو سريع عموماً عندما سمعت نداء الوصول إلى بلادهم هنا كانت لحظة تراكم الأفكار ولكن سيطرة على فكرة بذل المجهود للإستفادة من هذه **الرحلة** في أمور الدراسة خاصة بعد الوصول وإجراء الأمور الأزرمة لسماح لي بالدخول إلى البلاد أوقفت سائق الأجرة وأعطيتني العنوان وعندما وصلت إلى المكان الذي فالعنوان وهو منزل لعائلة أجنبية اتفق والدي معهم أن أمكت عندهم فترة يقانني فالبلد مقابلاً النقود فالبلد التالي وهو بداية دخولي المعهد وتكمية العائلة

Changing the determinants' values after the search

The determinants' values can be changed after the search; doing this will be reflected in the number of texts and results will be shown automatically, as they will be updated based on the new values of the determinants.

For example, searching the word 'الرحلة' [*the journey* in English], in the 1585 texts all, returns 1,139 results, but when selecting 'Native Arabic speaker' from the determinant 'Nativityness', the number of texts will decrease to 790 and the results to 457.

The number of corpus texts is **1585**
The number of texts based on your selection is **790**

Search Determinants

Clear All Determinants

AGE +

GENDER +

NATIONALITY +

MOTHER TONGUE +

NATIVENESS -

No Restriction

Native Arabic speaker

Non-native Arabic speaker

1-16 of total 457 results

Text ID	Concordance
S039_T1_M_PRE_NAS_W_C	يكروني بسنوات، وفي بداية الرحلة كنتُ تسير بسرعة عالية
S039_T3_M_PRE_NAS_W_H	الرحلة إلى المدينة بسم الله
S041_T1_M_PRE_NAS_W_C	هون علينا لعب هذه الرحلة الطويلة، وكان الناس عند
S043_T1_M_PRE_NAS_W_C	ما يلزمنا في هذه الرحلة وكانت أمي وأخواني يجوزون
S043_T1_M_PRE_NAS_W_C	المنزل مُتعبين في هذه الرحلة الجميلة الممتعة الشاقة وفي
S044_T1_M_PRE_NAS_W_C	إلى الرياض العاصمة وانتهت الرحلة
S052_T1_M_PRE_NAS_W_C	كتاب الله اهتقد أن الرحلة لا تكفيها صفحة واننبن
S054_T1_M_PRE_NAS_W_C	متاع ومستلزمات السفر واستغرقت الرحلة 4 ساعات فيها عدة

5.2.4 File downloads

The current website enables its user to download any subset of the corpus data — using the determinants — in different formats, as follows (see the appendix for an explanation of these formats and examples of the corpus files):

Format	Files number
TXT files with no metadata	1585
TXT files with Arabic metadata	1585
TXT files with English metadata	1585
XML files with Arabic metadata	1585
XML files with English metadata	1585
Original handwritten sheets in PDF	1257
Audio recordings in MP3	52

Once the determinants' values are changed, the available files will be updated based on the new selection; the 'Download' button then can be clicked after selecting the format(s) required.

The number of corpus texts is **1585**

The number of texts based on your selection is **1138**

Download these 1138 texts in different formats:

Plain text with no metadata

Plain text with Arabic metadata

Plain text with English metadata

XML with Arabic metadata

XML with English metadata

Hand written sheets in PDF

Audio recordings in MP3

Search Determinants

AGE +

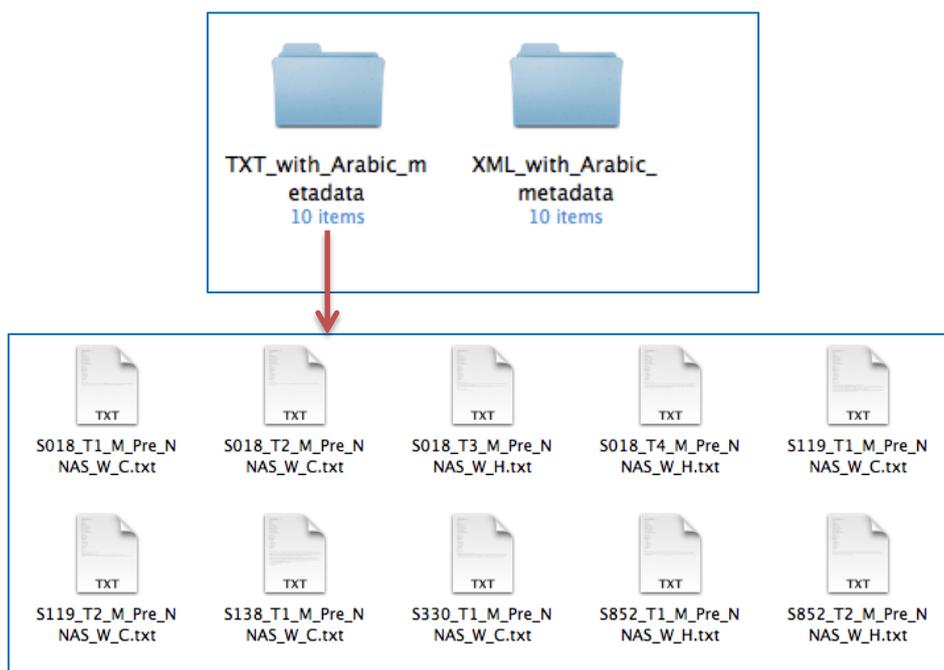
GENDER -

No Restriction

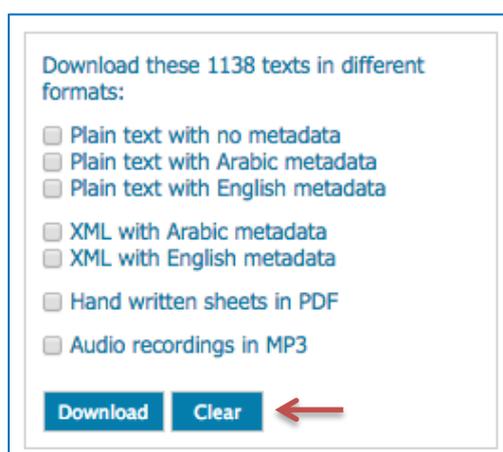
Male

Female

The files will be downloaded in one ZIP file containing subfolders, and each folder includes the files of one format of those selected.

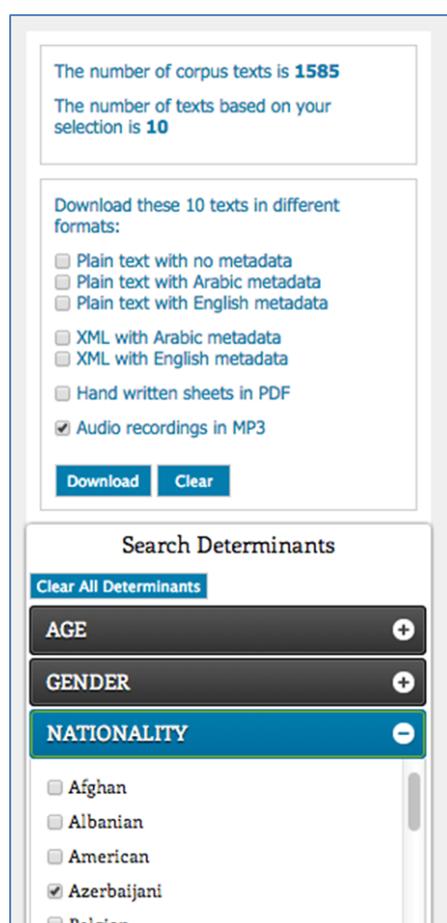


The 'Clear' button — in the files download section — clears all selections of the files' formats to facilitate making another selection.



Notes:

- One of the file formats should be selected, at the very least, to begin the download process.
- The PDF and MP3 formats have larger files than the other formats and, consequently, take more time to be prepared before starting the download process.
- The PDF and MP3 formats have fewer files than the other formats, i.e. some texts may not have files in these two formats. For example, when selecting 'Azerbaijani' from the 'Nationality' determinant, 'Audio recordings in MP3' from the files download section, and clicking on the 'Download' button, a message will appear indicating that there are no files to download for this selection. This occurs because there are no MP3 files for this selection, even though 10 texts can be downloaded in any of the other formats.



The number of corpus texts is **1585**
The number of texts based on your selection is **10**

Download these 10 texts in different formats:

- Plain text with no metadata
- Plain text with Arabic metadata
- Plain text with English metadata
- XML with Arabic metadata
- XML with English metadata
- Hand written sheets in PDF
- Audio recordings in MP3

Download **Clear**

Search Determinants

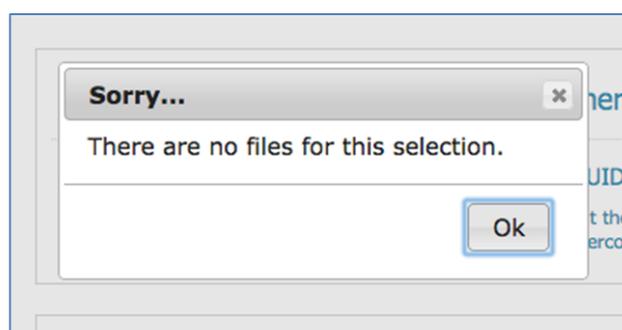
Clear All Determinants

AGE +

GENDER +

NATIONALITY -

- Afghan
- Albanian
- American
- Azerbaijani
- Belgian



6 Conclusion

This guide was created to illustrate the first version of the search website of the Arabic Learner Corpus. It aims to assist users in searching the corpus or a subset of its data as well as in downloading the files of any sub-corpus selected based on a number of determinants that have been explained in this guide.

If you have found a problem in the website or if you have an idea about improving it, please contact us at the following email:

Arabic.learner.corpus@gmail.com

Thank you.

7 Appendix

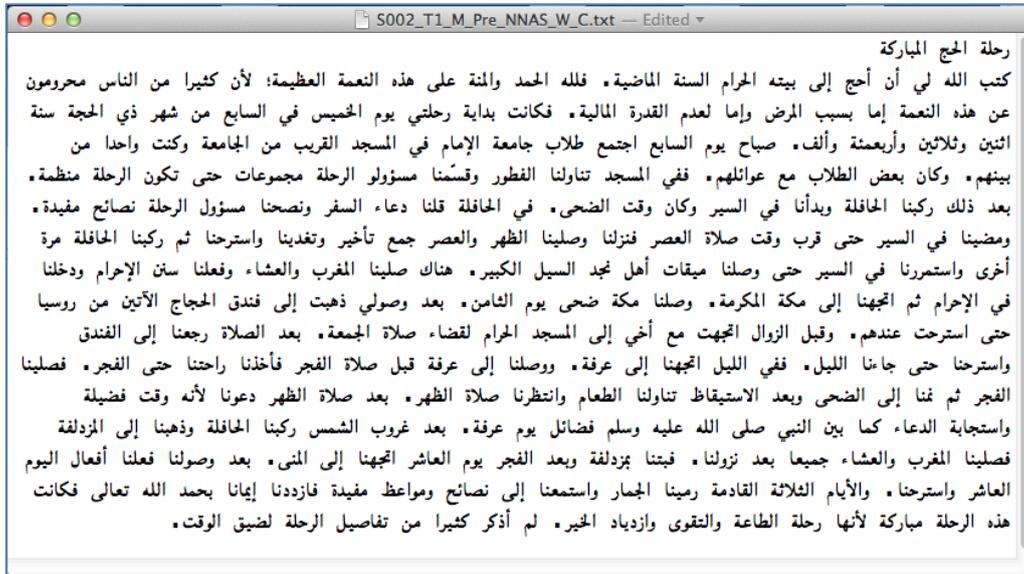
7.1 File types included in the corpus

This section gives an overview of the four file formats available to users of the corpus website (TXT, XML, PDF, MP3).

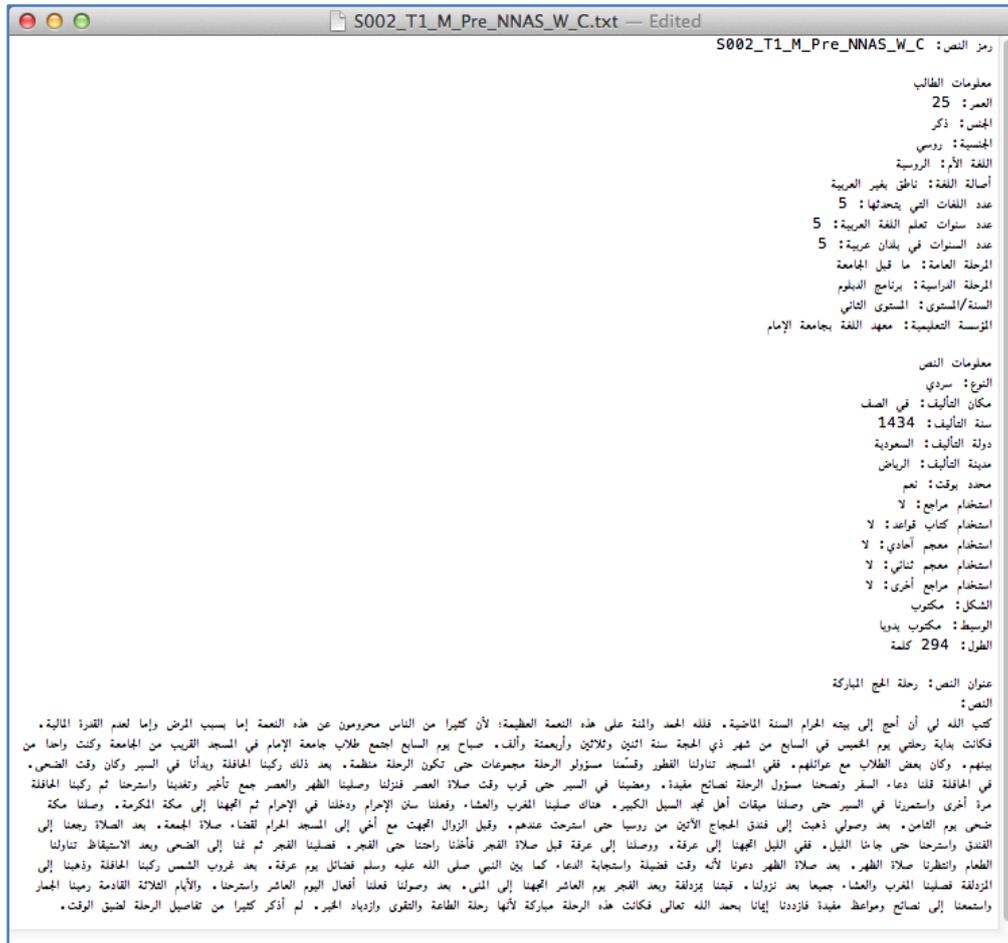
1. **TXT file:** This file type contains plain text without formatting (such as font type, size or colour). Such files can be read and edited with any text editors, such as Notepad on Windows. The corpus provides TXT files with three types of content; the first contains text with no data about the author or the text itself; the second includes, in addition to the text, data about the author (e.g. age, gender, nationality, mother tongue, level of study, etc.) and the text (e.g. genre, text mode: written or spoken, length, place of writing, etc.) in Arabic; the third provides the same data about the author and the text, but in English.
2. **XML file:** This file type is encoded using Extensible Markup Language (XML) that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. XML files were used for the corpus, as some corpus tools use this format for more choices and the efficient search of the data.
3. **PDF file:** The Portable Document Format (PDF) is a file format used to present documents in a manner independent of the application software and hardware, and the operating systems. It was used in the corpus for handwritten text after it had been scanned
4. **MP3 file:** This file format, which was designed by the Moving Picture Experts Group (MPEG), is an audio-coding format for digital audio; it uses a form of lossy data compression for the transfer and playback of music on most digital audio players. This format was used in the corpus for the learners' audio recordings.

7.2 Examples of the corpus files

1. Text file with no metadata



2. Text file with Arabic metadata



3. Text file with English metadata

S002_T1_M_Pre_NNAS_W_C.txt -- Edited

Text ID: S002_T1_M_Pre_NNAS_W_C

Learner Profile
 Age: 25
 Gender: Male
 Nationality: Russian
 Mother tongue: Russian
 Nativeness: NNAS
 No of languages speak: 5
 No of years learning Arabic: 5
 No of years in Arabic countries: 5
 General level: Pre-university
 Level of study: Diploma course
 Year/Semester: Second semester
 Educational institution: Arabic Inst. at Imam Uni

Text Profile
 Genre: Narrative
 Where produced: In class
 Year of production: 2012
 Country of production: Saudi Arabia
 City of production: Riyadh
 Timed: Yes
 References use: No
 Grammar book use: No
 Monolingual dictionary use: No
 Bilingual dictionary use: No
 Other references use: No
 Mode: Written
 Medium: Written by hand
 Length: 294 words

Text title: رحلة الحج المباركة
Text:
 كتب الله لي أن أجد إلى بيته الحرام السنة الماضية، فله الحمد والمثني على هذه النعمة العظيمة؛ لأن كثيراً من الناس محرومون عن هذه النعمة إما بسبب المرض وإما لعدم القدرة المالية. فكانت بداية رحلتي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعين وألفاً. صباح يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكنت واحداً من بينهم. وكان بعض الطلاب مع عوائلهم. ففي المسجد تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة. بعد ذلك ركبنا الحافلة وبدأنا في السير وكان وقت الضحى، في الحافلة قلنا دعاء السفر ونصحتنا مسؤول الرحلة بتصائح مفيدة. ومضينا في السير حتى قرب وقت صلاة العصر فنزلنا وصلينا الظهر والعصر جمع تأخير وتدبينا واسترحنا ثم ركبنا الحافلة مرة أخرى واستمرنا في السير حتى وصلنا ميقات أهل نجد السيل الكبير. هناك وصلينا المغرب والعشاء. وعلنا سن الإحرام ودخلنا في الإحرام ثم الجهننا إلى مكة المكرمة. وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الآتين من روسيا حتى استرحت عندهم. وقبل الزوال جهجت مع أخي إلى المسجد الحرام لقضاء صلاة الجمعة. بعد الصلاة رجعنا إلى الفندق واسترحنا حتى جانا الليل. ففي الليل جهجت إلى عرفة. ووصلنا إلى عرفة قبل صلاة الفجر فأخذنا راحتنا حتى الفجر. فصلبنا الفجر ثم قمنا إلى الضحى وبعد الاستيقاظ تناولنا الطعام وانتظرنا صلاة الظهر. بعد صلاة الظهر دعونا أنه وقت فضيلة واستجابة الدعاء كما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس ركبنا الحافلة وذهبتنا إلى المؤدفة فصلبنا المغرب والعشاء جميعاً بعد نزولنا. فبينما يزدلفنا وبعد الفجر يوم العاشر جهجتنا إلى النبي. بعد وصولنا فعلنا أعمال اليوم العاشر واسترحنا. والأيام الثلاثة القادمة ربيتنا الجمار واستمعنا إلى تصائح ومواعظ مفيدة فأزدنا إيماناً بحمد الله تعالى فكانت هذه الرحلة مباركة لأنها رحلة الطاعة والتقوى وازدياد الخير. لم أذكر كثيراً من تفاصيل الرحلة لضيق الوقت.

4. XML file with Arabic metadata

```
<?xml version="1.0"?>
<!--Arabic Learner Corpus_v2_2014-->
<!DOCTYPE doc>
- <doc ID="S002_T1_M_Pre_NNAS_W_C">
  - <header>
    - <learner_profile>
      - <age>25</age>
      - <gender>ذكر</gender>
      - <nationality>روسى</nationality>
      - <mothertongue>لروسية</mothertongue>
      - <nativeness>ناطق بفر العربية</nativeness>
      - <No_languages_spoken>5</No_languages_spoken>
      - <No_years_learning_Arabic>5</No_years_learning_Arabic>
      - <No_years_Arabic_countries>5</No_years_Arabic_countries>
      - <general_level>كما قبل الجامعة</general_level>
      - <level_study>برنامج البليج</level_study>
      - <year_or_semester>الستوى الثاني</year_or_semester>
      - <educational_institution>ميدقة لغة بجامعة الإمام</educational_institution>
    - <learner_profile>
  - <text_profile>
    - <genre>سردى</genre>
    - <where>فى مكة</where>
    - <year>1434</year>
    - <country>السعودية</country>
    - <city>رياض</city>
    - <timed>نعم</timed>
    - <ref_used>لا</ref_used>
    - <grammar_ref_used>لا</grammar_ref_used>
    - <mono_dic_used>لا</mono_dic_used>
    - <bi_dic_used>لا</bi_dic_used>
    - <other_ref_used>لا</other_ref_used>
    - <mode>مكتوب</mode>
    - <medium>مكتوب يدويا</medium>
    - <length>294</length>
  - </text_profile>
  - </header>
  - <text>
    - <title>رحلة الحج المباركة</title>
    - <text_body>
      كتب الله لي أن أجد إلى بيته الحرام السنة الماضية، فله الحمد والمثني على هذه النعمة العظيمة؛ لأن كثيراً من الناس محرومون عن هذه النعمة إما بسبب المرض وإما لعدم القدرة المالية. فكانت بداية رحلتي يوم الخميس في السابع من شهر ذي الحجة سنة اثنين وثلاثين وأربعين وألفاً. صباح يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكنت واحداً من بينهم. وكان بعض الطلاب مع عوائلهم. ففي المسجد تناولنا الفطور وقسّمنا مسؤولو الرحلة مجموعات حتى تكون الرحلة منظمة. بعد ذلك ركبنا الحافلة وبدأنا في السير وكان وقت الضحى، في الحافلة قلنا دعاء السفر ونصحتنا مسؤول الرحلة بتصائح مفيدة. ومضينا في السير حتى قرب وقت صلاة العصر فنزلنا وصلينا الظهر والعصر جمع تأخير وتدبينا واسترحنا ثم ركبنا الحافلة مرة أخرى واستمرنا في السير حتى وصلنا ميقات أهل نجد السيل الكبير. هناك وصلينا المغرب والعشاء. وعلنا سن الإحرام ودخلنا في الإحرام ثم الجهننا إلى مكة المكرمة. وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الآتين من روسيا حتى استرحت عندهم. وقبل الزوال جهجت مع أخي إلى المسجد الحرام لقضاء صلاة الجمعة. بعد الصلاة رجعنا إلى الفندق واسترحنا حتى جانا الليل. ففي الليل جهجت إلى عرفة. ووصلنا إلى عرفة قبل صلاة الفجر فأخذنا راحتنا حتى الفجر. فصلبنا الفجر ثم قمنا إلى الضحى وبعد الاستيقاظ تناولنا الطعام وانتظرنا صلاة الظهر. بعد صلاة الظهر دعونا أنه وقت فضيلة واستجابة الدعاء كما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس ركبنا الحافلة وذهبتنا إلى المؤدفة فصلبنا المغرب والعشاء جميعاً بعد نزولنا. فبينما يزدلفنا وبعد الفجر يوم العاشر جهجتنا إلى النبي. بعد وصولنا فعلنا أعمال اليوم العاشر واسترحنا. والأيام الثلاثة القادمة ربيتنا الجمار واستمعنا إلى تصائح ومواعظ مفيدة فأزدنا إيماناً بحمد الله تعالى فكانت هذه الرحلة مباركة لأنها رحلة الطاعة والتقوى وازدياد الخير. لم أذكر كثيراً من تفاصيل الرحلة لضيق الوقت.
    - </text_body>
  - </text>
  - </doc>
```

5. XML file with English metadata

```
<?xml version="1.0"?>
<!--Arabic Learner Corpus_v2_2014-->
<!DOCTYPE doc>
- <doc ID="S002_T1_M_Pre_NNAS_W_C">
  - <header>
    - <learner_profile>
      <age>25</age>
      <gender>Male</gender>
      <nationality>Russian</nationality>
      <mother_tongue>Russian</mother_tongue>
      <nativeness>NNAS</nativeness>
      <No_languages_spoken>5</No_languages_spoken>
      <No_years_learning_Arabic>5</No_years_learning_Arabic>
      <No_years_Arabic_countries>5</No_years_Arabic_countries>
      <general_level>Pre-university</general_level>
      <level_study>Diploma course</level_study>
      <year_or_semester>Second semester</year_or_semester>
      <educational_institution>Arabic Inst. at Imam Uni</educational_institution>
    - </learner_profile>
    - <text_profile>
      <genre>Narrative</genre>
      <where>In class</where>
      <year>2012</year>
      <country>Saudi Arabia</country>
      <city>Riyadh</city>
      <timed>Yes</timed>
      <ref_used>No</ref_used>
      <grammar_ref_used>No</grammar_ref_used>
      <mono_dic_used>No</mono_dic_used>
      <bi_dic_used>No</bi_dic_used>
      <other_ref_used>No</other_ref_used>
      <mode>Written</mode>
      <medium>Written by hand</medium>
      <length>294</length>
    - </text_profile>
  - </header>
  - <text>
    <title>رحلة فتح المدينة</title>
    <text_body>
      كتب
      الله لي أن أجد إلى بيتك الحرم السنة الماضية، فلتة الحمد والمنة على هذه النعمة العظيمة؛ لأن كثيرا من الناس محرومون عن هذه النعمة إما بسبب المرض وإما لعدم القدرة المالية. فالتقت بداية رحلتي يوم السابع من شهر ذي الحجة سنة اثنين واثنين وأحمد سبحان يوم السابع اجتمع طلاب جامعة الإمام في المسجد القريب من الجامعة وكانت واحدة من بيوتهم وكان يحض الطلاب مع عوائلهم. ففى المسجد تشاركنا في تناول الطعام وشكنا مسألتنا الرحلة بمجموعات حتى تكثرت الرحلة فتمتد، بعد ذلك رأينا الحافلة وبدأنا في السير وكان وقت الضحى في الحافلة كنا نداء سفر ونصعدنا مسألتنا الرحلة فتمتد فبدأنا في السير حتى قرب وقت صلاة العصر فتركنا وصعدنا الظهر والعصر جميع تأخير وتعدينا واسترخينا ثم رأينا الحافلة مرة أخرى واستمرنا في السير حتى وصلنا بيوت أهل نجد المسيل الكبير. هناك صعدنا العرب والقطار ولقد سألنا الأجرم ونحن في الأجرم ثم ذهبنا إلى مكة المكرمة، وصلنا مكة ضحى يوم الثامن. بعد وصولي ذهبت إلى فندق الحجاج الأثين من روسيا حتى استرحيت عندهم. وقبل الزوال توجهت مع أخي إلى المسجد الحرام لتأدية صلاة الجمعة. بعد الصلاة رجعا إلى الفندق واسترخينا حتى جاءنا الليل. ففى الليل توجهنا إلى عرفة، ووصلنا إلى عرفة قبل صلاة الظهر فأخذنا راحة حتى نلتج، فوصلنا الظهر ثم نمنا إلى الضحى يوم الاستيقاظ تناولنا الطعام وتكثرتنا صلاة الظهر. بعد صلاة الظهر عدنا لأنه وقت الخطبة واستجابة الدعاء مما بين النبي صلى الله عليه وسلم فضائل يوم عرفة. بعد غروب الشمس رأينا الحافلة وأخذنا في الدائرتنا فوصلنا المغرب وبعثنا جميعا بعد تزويتنا، فبدأنا بدائرتنا وبعد الظهر يوم العاشر توجهنا إلى العشر. بعد وصولنا فبدأنا أفعال اليوم العاشر واسترخينا، والإيام الثلاثة القادمة رأينا الجمار واستمعنا إلى لصالح وبعنا صلاة مفيدة فزينا إيما بعدد الله تعالى فبدأت هذه الرحلة مباركة لأبنا رحمة الطاعة والتقوى والزياد الخير. ثم
    </text_body>
  - </text>
</doc>
```

6. PDF file of handwritten text

