Graphic Online Language Diagnostic (GOLD)

Manual Version 1.1
December 2011

Center for Advanced Language Proficiency Education and Research
The Pennsylvania State University

http://calper.la.psu.edu/corpus.php

Creating and Editing a Corpus in GOLD

1. Creating a new corpus in GOLD

To create your own corpus in GOLD, you need to first add a new corpus to the system and then upload documents to the corpus.

1.1 Adding a new corpus

To add a new corpus to GOLD, first click "Corpora" in the left panel. Then click "Add Corpus". Fill out all the fields in the form that pertain to the corpus you want to create, including

- Title: Give your corpus a meaningful title.
- Description: Briefly describe the sources, content, size, etc. of your corpus.
- Language: Select the language of your corpus from the drop-down list; if it is not on the list, select "Other" and then enter the language.
- Category: This can be ignored at the moment.
- Access: If you do not wish to share your corpus with anyone, select "Private"; if you wish to make your corpus public, select "Public"; if you wish to give access to authorized users only, select "Shared". To authorize a user, search for him/her by entering his/her username, first name, or last name, and then assign him/her appropriate access rights. You may authorize him/her to view, add documents to, edit, or delete your corpus by checking the appropriate boxes.
- Citations: Specify how you would like other users to cite your corpus.

After you have filled out all the fields, click "Submit". Your corpus has been added to GOLD, although it does not contain any document yet.

1.2 Uploading documents to a new corpus

There are two ways to upload documents to a new corpus.

1.2.1 Uploading documents using guided XML creation

If you have no prior experience with XML, we recommend that you begin with the guided XML creation function first. First click "Create New Document" to the right of "By Guided Wizard", under "Add Documents". Then click "Add A field" and provide a name for the field. A field is a (preferably one-word) name of a variable related to the learner or the text, e.g., name (name of the student), gender (gender of the student), rating (holistic rating of the essay), etc. Add as many fields as you deem necessary, but only include informative fields. After you are done with adding fields, click "Enter Documents" to enter your documents one by one. For each document,

enter a brief value for each field (e.g., John for name, Male for gender, B for rating, etc.), and then type or copy and paste the actual text of the document in the “Content” box. Click “Add Another Document” to add more documents. After you have added all your documents, click “Save Document to Corpora” to directly upload the documents to the corpus. If you wish to save a copy of the XML file generated by the system, you may click “Generate XML” to download it.

1.2.1 Creating an XML file using a text editor

If you are not familiar with editing your own XML file, we recommend that you first generate an XML file using the “Guided XML creation” function with at least two documents, and use that file to familiarize yourself with the format.

Edit your file with WordPad, NotePad, Emacs, etc., but not Microsoft Office Word, following the template below.

```
<corpus>
    <document>
        <metadata name="key1">VALUE</metadata>
        <metadata name="key2">VALUE</metadata>
        <content>
            <![CDATA[
                CONTENT
            ]]>
        </content>
    </document>
</corpus>
```

The file should begin with a <corpus> tag and end with a </corpus> tag. In between this pair of tags, you can have as many documents as necessary. Each document is enclosed in a pair of <document> and </document> tags. It is important to note that all documents come before the </corpus> tag, which indicates the end of the corpus. Within each document, you first have a few lines that specify the values of the metadata fields. In each metadata specification line, “key” represents the name of a field, and “VALUE” specifies the value of that field in this document. Two examples are given below, where the fields are “name” and “gender”, respectively, and the values are John and Male, respectively.

<metadata name=”name”>John</metadata>
<metadata name=”gender”>Male</metadata>

The content of the document is enclosed in a pair of <content> and </content> tags. The actual text of the document appears after <![CDATA[and before]]>, or in other

words, you should replace “CONTENT” in the template with the actual text of the document.

After you have typed in all your documents, you should name your file with the .xml suffix (e.g., MyCorpus.xml) and save it with Unicode encoding. For example, to save the file appropriately in WordPad, click “Save as”, then change the name of the file to something like MyCorpus.xml and change the “Save as type” box to Unicode Text Document.

To upload your own XML file to a new corpus, click “Browse” to the right of “By File Upload”, under “Add Documents”, to locate your document, and then click “Upload”.

2. Editing an existing corpus

To edit an existing corpus, first click “Corpora” to view the list of corpora you have access to. You can then do any of the following.

2.1 Deleting an existing corpus

Click the “Delete” button to the right of the corpus you wish to delete.

2.2 Editing general information and permissions about an existing corpus

Click on the title of the corpus or the “Edit” button to the right of the corpus you wish to edit. To edit general information or permissions, make changes directly in the forms and then click “Save Corpus”.

2.3 Adding documents to an existing corpus

Click on the title of the corpus or the “Edit” button to the right of the corpus you wish to edit. Then add documents to the corpus in the same way as you add documents to a new corpus (see Section 1.2 above for details).

2.4 Deleting documents from your corpus

Click on the number in the “Documents” column that corresponds to the corpus you wish to edit. You should see a list of documents in your corpus. You may view a specific document by clicking “View”. Check the boxes of the documents you wish to delete, and then click “delete”.

2.5 Editing documents in your corpus

Click on the number in the “Documents” column that corresponds to the corpus you

wish to edit. You should see a list of documents in your corpus. To view the content of a document, click on the “Preview” button  to the right of the document. To see details of or edit a document, click on the “Details” button  of the document. To edit the value of a metadata field, click “Edit” to the right of the field you wish edit, make all necessary changes, then click “Save”. To edit the text of the document, click “Edit” to the very right of “Contents”, make all necessary textual changes, click “Save”, and then click “Refresh to update stats”.

Corpus Search in GOLD

1. Selecting corpora

Log on to GOLD and click "Search CORPORA" on the top navigation bar. You will see a list of corpora that you own or have access to. Check the boxes to the left of the corpora that you wish to search, and click "Continue".

2. Word count

After selecting corpora, you can click "See word count" to see a word list created for the corpora you selected. By default, the word list is organized in descending order of frequency. Click "Frequency" to rearrange the list in ascending order. Alternatively, you can also click "Word" to sort the list alphabetically. To view the entire list on the same page, click "View All". To view results on shorter pages, click "View Paged". The word list can be saved as plain text file (ASCII format) by clicking "Save as text".

3. Corpus statistics

After selecting corpora, you can click "See statistics" to view statistical information, including mean sentence length, average word length, and type/token ratio, both for the entire corpus (or corpora) and for each individual document in the selected corpus (or corpora).

4. Searching corpora

To search keyword(s) in selected corpora, you need to specify your searching criteria as follows.

Keyword(s):

Type the keyword(s) you wish to search in the "Keyword(s)" field, e.g., 'book', and specify the context window, i.e., the number of words you wish to see before and after the keyword(s) (5-30 words). To search for all inflections of a keyword, e.g., 'books', 'booking', 'booked', etc., attach the wildcard to the end of the keyword, e.g., "book*".

Surrounding words

You can specify up to 3 words that must occur within your context window. You can also specify the position in which each of these words must occur relative to the keyword(s), i.e., R5, R4, R3, R3, R2, R1 and L1, L2, L3, L4, L5, where R5 means "five words to the right of the keyword(s)", etc.

Associated word

In this field, you can specify a word that is associated with the keyword(s). The associated word could occur either to the left or to the right of the keyword(s).

Conditions

Conditions can be used to specify which documents in the corpora you wish to search. For example, you may wish to search documents written by male students only. Click "+" to start choosing your options. You can either specify one condition or multiple conditions using the logic operators AND and OR. The options you have will solely depend on the information encoded in your corpora. Once you have specified your search criteria, click "Search" to see the results.

5. Sorting results

The results are displayed in a keyword in context (KWIC) format and can be sorted according to the words occurring to the left or right of the keyword(s) using the "Sort by" fields provided.

6. View original text

If you wish to view the original text containing a specific occurrence of the keyword(s), click "View" to the right of the concordance line. A pop-up window will show the original text file.

7. Search statistics

Click "Statistics" to view a statistical analysis of the search results, including mean sentence length, average word length, type/token ratio, number of keyword(s) matches, total word count, and normalized frequency of the keyword(s) (Frequency / 100 words) both for the entire corpus (or corpora) and for each individual document in the selected corpus (or corpora) that satisfy your search condition.

Collocation Search in GOLD

The search collocations function allows you to retrieve collocations (otherwise referred to as lexical bundles, n-grams, or multiword expressions) containing a keyword from one or more corpora.

1. Selecting Corpora

Log on to GOLD and click "Search COLLOCATIONS" on the top navigation bar. You will see a list of corpora that you own or have access to. Check the boxes to the left of the corpora that you wish to search, and click "Continue".

2. Search Collocation

Before searching collocations in the selected corpora, you need to specify your searching criteria as follows.

Contains:

In this field, type the keyword that you wish to search, e.g., 'book'. If you wish to also include inflected forms of the word, e.g., 'books', 'booking', 'booked', etc., check "match lemmatized collocations".

Occurs at least:

In this field, specify the minimum frequency of occurrence for the collocations to be displayed. Collocations with a lower frequency will not be displayed.

Collocation length:

Specify the length of the collocation (in number of words) from the drop-down list. For example, to search for all three-word sequences (or trigrams), simply select 3. If the number selected is greater than 1, you can also specify what word should occur in what position of the collocation.

Conditions:

Conditions can be used to specify which documents in the corpora you wish to search. For example, you may wish to search documents written by male students only. Click "+" to start choosing your options. You can either specify one condition or multiple conditions using the logic operators AND and OR. The options you have depend on the information encoded in your corpora. Once you have specified your search criteria, click "Search" to see the results.

3. Sorting results

Search results can be sorted alphabetically or by frequency.

Sorting by frequency

By default, search results appear in descending order of frequency. Click “Frequency” to rearrange the results in ascending order.

Sorting alphabetically

To sort the results alphabetically, click “Word”.

4. View all vs. View paged

To view all results on the same page, click “View All”. To view results on shorter pages, click “View Paged”.

5. Save results

Collocation search results can be saved as a plain text file (ASCII format) by clicking “Save as text”.