5. More about collections

We have seen how FLAX can present language activities to students to enhance their language learning. It also provides different ways of looking at the information in the document collection itself. We saw some of these in Chapter 2, like searching and browsing. Now we take a more extensive look at what FLAX collections can provide.

This chapter uses a different and more extensive document collection: the Arts and Humanities collection of the British Academic Written English corpus, which contains 700 pieces of high-standard student writing in the area of arts and humanities. This includes the disciplines of Archaeology, Classics, Comparative American Studies, English, History, Linguistics, and Philosophy. Texts are categorized into 13 genre types: case studies, critiques, design specifications, empathetic writing, essays, exercises, explanations, literature surveys, methodology descriptions, narratives, problem questions, proposals and research reports. These classifications constitute “metadata” for the collection, which Greenstone, the digital library software that underpins FLAX, is capable of incorporating and using for browsing. You can see the collection at the FLAX website.5

It is also possible to create your own collection, based on your own documents. Teachers often want to use their own linguistic material in their courses. FLAX makes it easy to build small collections out of any documents – including images and audio – that you have available electronically. There are other ways of building large collections, but we won’t describe those here. Greenstone is capable of handling vast collections of all sorts of information.

Teachers may want to create their own collections of documents for reasons other than language learning. After all, libraries have always been central to education. It is surprising, and perhaps a little disappointing, that digital libraries do not usually seem to play a central role in computer systems for educational support. FLAX allows you to organize readings and provide students with integrated access to them. If appropriate, they can also serve as a source of material for language learning.

5 http://flax.nzdl.org/greenstone3/you?act=fp&sa=collAbout&c=BAWEAH&if=flax
5.1 Exploring articles

Here is the main page of the Arts and Humanities collection. Users can seek documents, sentences, paragraphs and collocations containing particular words or phrases; they can browse documents by genre and discipline, collocations by word, academic words, and lexical bundles.

The Browse by Genre button has been clicked, which brings up (on the left) the 13 genre types mentioned above. The first type has been selected, which brings up (on the right) the titles of all articles classified as critique. Users can select a different discipline (the Browse by Discipline button), or search for articles containing a particular word (the Search button). These are standard Greens tone facilities.

Also shown is an article in the collection, entitled “Corpus Linguistics and Information and Persuasion,” which can be accessed by selecting Browse by Genre and then exercise, or alternatively by selecting Browse by Discipline and then Linguistics.

Main page of the Arts and Humanities collection

A sample article in the collection
5.2 Wordlist view

The Wordlist view lets users analyze the range of vocabulary used in the article. In the accompanying image, the academic words, taken from a standard 570-word list used in language teaching, are highlighted in blue. These include individuals, conclusion, evidence, analysing, context, analysis, and so on. Clicking a highlighted word leads to a page that shows all sentences in the collection containing that word. The second illustration shows the result of clicking the word analysis. As well as showing the sentences (668 of them), the system suggests, at the top that the user may also be interested in related words: analyze, analyze, analytical, analytic, analysis, analyst, analytically. Clicking the little green arrow shown after each sentence pops up the paragraph in which that sentence appears, shown in a gray box.

The proportion of academic words in this article (13%) is given beside the dropdown box above the text in the upper image (showing “academic words”). This box provides other options: the most frequent 1000 and 2000 words, also taken from standard wordlists; other words; and keywords. By changing this, users can easily ascertain that 59% of the words in this article are in the top 1000 words, which is increased by a further 5% by extending this to 2000 words.

The keyword view shows keywords in the article. For this article, it highlights the words linguistics, concordances, phrases, corpus, persuade, prosody, and so on. In keyword view, a slider appears beside the dropdown that the user can manipulate to reveal more words. Moving it to the right makes the system less selective, highlighting more words; conversely, moving it to the left reduces the number of highlighted words. At the very left end only one keyword (“CSA”, an acronym for Child Support Agency) is given, while at the right end all content words are displayed.

Keywords are calculated by a heuristic method based on word frequency. For each content word, a score is calculated that reflects how important the word is to the document, based on the number of times it occurs in the document (which increases the score) and the number of times it occurs in the collection as a whole (which decreases it). For example, the word “CSA” receives the highest score for this article, because it occurs 11 times and does not occur in any other article.
Collocations, highlighted in the Collocation view

<table>
<thead>
<tr>
<th>Original</th>
<th>adjective</th>
<th>adverb</th>
<th>preposition</th>
<th>verb</th>
</tr>
</thead>
</table>
| Applied linguistics is the study of language in the real world – how it is really used by individuals and throughout society. Linguists find conclusions about language use by studying examples of it, by collecting evidence and analyzing it. One way of doing this is by examining concordance texts, which provide multiple examples of words or phrases in their context in a particular corpus. Critical Discourse Analysis allows the linguist to study the relationship between linguistic choices and effects (Cook 2003). It enables one to see the way language is manipulated in order to persuade both these areas of applied linguistics were investigated, and should provide interesting results. Task One. The field of corpus linguistics concerns words or phrases in their context. One can find the ways in which a word or phrase is used by searching corpora – vast banks of language data that relate to how words are used in reality. The phrase used in itself negatively does not entail negative connotations. However in its context, it is clear that it is used by others to relate to the context. The Times newspaper for example contains sentences such as this sentence. Each tab shows collocations starting with that word type; for example, the adjective panel hosts collocations starting with an adjective.

Collocations for the word connotation

<table>
<thead>
<tr>
<th>Connotation Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexual connotations</td>
<td>14</td>
</tr>
<tr>
<td>Negative connotations</td>
<td>10</td>
</tr>
<tr>
<td>Different connotations</td>
<td>7</td>
</tr>
<tr>
<td>Social connotations</td>
<td>5</td>
</tr>
<tr>
<td>Pejorative connotations</td>
<td>4</td>
</tr>
</tbody>
</table>

5.3 Collocation view

Once an article has been reached, the collocation view is accessed by clicking one of the tabbed panels, which are described below:

- **verb + noun**
- **adjective + noun**
- **noun + noun**
- **noun + of + noun**
- **preposition + noun**

Each tab shows collocations starting with that word type; for example, the adjective panel hosts collocations starting with an adjective.

Collocations are highlighted in the text, to help students notice them and study their context. In the example shown, collocations related to the subject of the article, corpus linguistics – linguistic choice, negative connotation, unpleasant connotation, abstract nouns, negative prosody, neutral company, persuasive language, interpersonal function, attributive function, linguistic interaction – stand out from the surrounding text. The collocation negative connotations has been clicked to bring up a superimposed popup with two small icons. The underlined words here, negative and connotations, are hyperlinked to entries that involve those words in an external collocation database, built from all the written text in the British National Corpus.

Users can easily study further collocations related to these two words. For example, clicking connotation generates a further popup, shown below, that lists sexual connotations, different connotations, social connotations, pejorative connotations, etc., along with their frequency in that corpus. Likewise, clicking negative would show collocations such as negative attitudes, negative side, negative aspects.

The panel shown helps users learn about the word connotation. They can see samples of these collocations in context by clicking them here, which shows relevant extracts from the British National Corpus – for example, clicking sexual connotations brings up ten sentences that use this phrase.

Returning to the top image, clicking the first icon in the popup (the “web” icon) generates a page containing sample texts retrieved by a live Web search. The system connects to a search engine, uses the collocation as a phrase query, retrieves sample texts, parses them, and displays extracts that contain the collocation. However, the text, being extracted from individual Web pages, is often unclean, incomplete and repetitive. The function of the second icon in the popup (the “cherries” icon) is explained in Section 5.5.
5.4 Searching and browsing collocations

Users can investigate collocations by searching or browsing for partners of a particular word. As well as the standard search function of locating articles, sentences, or paragraphs that contain a particular word or words, the collection’s “Search” button allows users to search collocations in the collection. The accompanying image shows the result for the word notion, which returns nearly 300 collocations. The first four can be seen here, along with their contexts. They are grouped under tabs that reflect the syntactic roles of the associated word or words, Verb, Adjective, Noun, Preposition, and Noun + of (shown). The dominant pattern for our chosen word is notion + of (half the total of 300), as in notion of performance, notion of syllable, notion of time. The next most dominant is verb + notion (nearly a third of the total): solidifies the notion of, renders the notion of, advocate the notion of, and so on.

Collocations can be browsed as well as searched, using the Collocations button on the collection’s home page and article pages (shown beside Exploring articles, above). Then, an alphabetic selector leads to the word in question. Clicking the letter n, followed by the word notion, obtains the collocations shown opposite.

Exploring collocations associated with the word notion
5.5 Cherry picking

Collocation knowledge is difficult to acquire simply because there is so much of it. Native speakers carry hundreds of thousands of lexical chunks in their heads, ready to draw upon in order to produce fluent, accurate and meaningful language. This presents language learners with a daunting challenge.

To help them, learners are often encouraged to collect their favorite collocations for possible use in later writing. FLAX provides a facility for “picking” favorite collocations and saving them in a kind of notebook. We use the metaphor of cherries because, like collocations, they are tasty fruit that come in small groups. Cherry picking is initiated by clicking the “cherries” icon that follows collocations shown in earlier illustrations. Students can pick collocations from an article, or from search results or browsing results pages.

This icon beside a collocation adds it to the user’s cherry basket.

In the illustration here, the collocation negative connotation has been chosen. In the upper image it is being added to the student’s personal cherry basket. If desired, the basket can be organized into categories, and new categories can be defined when picking a collocation. Or, if the user prefers, the basket can be left unorganized.

The basket’s contents are shown when the user clicks My Cherry Basket, on the right of the control bar near the top of the screen. The bottom image shows a student’s cherry basket, displaying collocations that have been picked and placed in two categories: theory and role. The usual options are provided for organizing the basket: a cherry can be dragged into the “folder” icon beside another category, collocations can be deleted (with the “x” symbol), and category names can be altered (with the edit icon). The web icon on to the right of each collocation retrieves further examples from the web. A “print friendly” version of the basket can be generated, which is also suitable for sharing by email.

The cherry basket, and indeed FLAX’s entire collocation facility, is particularly useful when learning English in a certain domain, like geography, management, or in a genre like academic writing. The teacher creates a collection of documents in the domain, perhaps from Wikipedia or some other public source. Students study them, investigate the collocations they contain, and choose ones they like for use in future writing exercises. Instead of simply mastering words – pronunciation, forms, and meanings – students learn the contexts in which they are used, which helps them combine words into apt phrases, sentences and texts.

Even for general writing, collocations help students express themselves more aptly and precisely. For example, learners, especially lower level ones, tend to overuse common words like very because of their limited stock of adverb modifiers. FLAX enables them to find modifiers that are more suitable for use with particular adjectives or verbs, such as completely, physically, mentally, emotionally with the word exhausted and heavily, strongly, deeply, easily, unduly with the phrase influenced by.
5.6 Exploring words

Searching documents for particular words and phrases is one of the standard functions of digital libraries. However, ordinary library users seek information about the content of articles, whereas language learners want to know about how words are used. The same search mechanism can be applied, but the results need to be displayed in different ways.

FLAX lets users seek documents, paragraphs and sentences containing a particular word, along with its variants. Here we focus on sentence and paragraph searching. The upper image shows the first 9 of the 1800 sentences that contain the word suggest; sentences containing the inflected forms suggesting, suggested and suggests are also included. A lemma list containing about 15,000 entries is consulted to recognize inflected forms of a query word. Clicking the “arrow” icon at the end of a sentence pops up the paragraph that contains the sentence, showing its context.

Other terms derived from the query word, in this case the word suggestion, are given at the very top of the panel as another possible search. Although suggest has only one derived term, there are often several; for example, the query word importantly yields the derivatives importance, importantly, unimportant, and unimportance, which are sorted into descending order of frequency in the collection.

Search queries can contain more than one word, in which case sentences are returned that contain all the query terms. For phrase searching, a query can be enclosed by quotation marks; for example, “language learning theory” returns sentences containing this phrase, while language learning theory returns sentences containing these three words.

Users can explore academic words in the collection by clicking the Wordlist button on the collection’s home page and article pages (shown beside Exploring articles, above). This yields the screen shown at the bottom. The words can be sorted alphabetically or by frequency, as here; in either case, frequency in the collection is shown alongside the word.

The “cherries” icons link to the collocations associated with the word, yielding the display described earlier under Searching and browsing collocations. The “British flag” icon invokes a search of the British National Corpus, which allows users to examine how the word is used in different contexts. Note that in each of these cases all inflected forms of the word are also included in the search.
5.7 Exploring usage patterns of words

On the panel showing the result of searching for sentences containing the word suggest, just described, shows, near the top, a Group by pattern option, which is turned off by default. This allows users to study word usage by showing salient syntactic patterns that involve it.

Patterns are grouped by whether the word in question occurs near the beginning of sentences or in the middle, because these provide different views of the word’s usage. The upper illustration shows (in the header bar) that 532 patterns are found for the word suggest. They are separated into two tabs, At the beginning (166 patterns) and In the middle (lower illustration; 366 patterns).

The most common usage of the word suggest is to begin sentences of the form Aitchinson suggests that, which occurs – with different subjects – in many sentences (123). Clicking the pattern brings these up; they include Thomas suggests that, Archeologists suggested that, and Lenin also suggested that. These all exhibit the pattern: Noun + suggest + that-clause. The next most common sentence-initial patterns, This suggests that, They suggest that, Aitchinson suggests, It has been suggested that, It is suggested that, and so on, demonstrate that the most dominant usage of this word near the beginning of sentences is subject + suggest + that-clause. Note that there are two patterns that open with Aitchinson suggests; the second is different because omits the word that from the that-clause.

The lower illustration shows how suggest is used in the middle of sentences. The top two patterns are suggest + that-clause, while the third is suggest + noun phrase. The second example is distinguished as a different pattern because, being preceded by a comma, it begins a clause, as in Swinburne reinforced this idea, suggesting that people expect too much ....

The patterns that follow (from 71 to 11 occurrences) show the same trend. The that-clause is followed by patterns such as noun phrases (older adults, the number of), pronouns (we, it, he) and proper nouns (Jones). Noun phrase patterns include article + adjective + noun (suggests a test, suggest an annual activity) and article + adjective + noun + of (the origin of, the means of, particular methods of).