Text Analysis and Corpus Linguistics for English Language Teachers

Robert Buckmaster

An English Ideas Monograph
Text Analysis and Corpus Linguistics for English Language Teachers

Robert Buckmaster

An English Ideas Monograph
Text Analysis and Corpus Linguistics for English Language Teachers
by Robert A. Buckmaster
Independent Corpus Linguistics Researcher

Published by English Ideas
www.englishideas.org
First Published in 2015
10 9 8 7 6 5 4 3 2 1
© Robert A. Buckmaster 2015

Note that the text includes extracts from an article by Robert Buckmaster: Creating a key word list published in English Teaching Professional Issue 38 May 2005.

Cover Photograph: Autumn Leaves in Tallinn Old Town
© Robert A. Buckmaster 2005

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.
“Trust the text.”
John Sinclair
1933-2007

“You shall know a word by the company it keeps.”
John R. Firth
1890 - 1960

“Words are like autumn leaves. They may be beautiful but it is the tree which gives them life and meaning.”
Robert Buckmaster
2015
Introduction

The text and lexis are central, or should be central, to language learning. Texts are the most efficient way of communicating meaningful language to learners. Words need co-text and contexts to be meaningful and texts provide these.

Texts can range from one word to .... however many you want really, but short texts are better, as it is easier to cope with a short text in detail in a lesson than a long text, and it's better to use a short text to its full potential than to look superficially at a longer text.

This short guide will introduce you to ideas and tools which will help you to understand texts, and show you how to build and analyse a corpus which you can use to help your learners.

Text Analysis

In particular we will look at three levels of text analysis; at word frequency and range analysis tools; tools to compare texts; and quick ways to extract concordance lines from texts.

Building and Exploiting a Corpus

It’s relatively straightforward to build a micro-corpus and to run simple analyses on it to extract useful information for you and your learners. This guide will show you how to become a corpus linguist in simple steps.
Why Text Analysis?

It is important to really know a text before you teach it/with it – otherwise how do you know what can be done with, what the learners can learn from it and what it is that makes it special and worth doing with your learners?

Each text should be subject to a pedagogical analysis and a linguistic analysis. The pedagogical analysis should follow the linguistic one. The linguistic analysis will be explored here – how can you get to really know a text beyond reading it and ‘understanding it’? Reading and understanding a text are not enough, in themselves, for you to be able to say you ‘know’ a text sufficiently to use it for teaching.

Levels of Text Analysis

Teachers often look at a text and think about it in terms of lexis and grammar. Indeed, a text is a collection of words [and grammar if you make such a distinction] but that is not all: there are other levels of analysis to consider.

First take a text:

A crook tried to escape a police search by hiding in a carpet. Stanislaw Dabrowski, 38, held up a cosmetics shop and then escaped to his aunt's flat in Warsaw, Poland. The thief rolled himself up in a giant rug and propped himself up against a balcony wall when the police arrived to search the flat. He was only found when a detective went out onto the balcony for a cigarette and noticed the carpet was moving. Dabrowski is now facing a 12 year prison sentence for armed robbery.

The three diagrams below show different levels of analysis with a short commentary.

Level 1 Grammar

The Text Features 1 diagram shows the grammar of the short text marked. As you can see there is a lot of grammar in this short text. Most teachers think in terms of verb grammar when you mention ‘grammar’ but a look at this text shows other equally important features:
noun phrases
prepositions
prepositional phrases
conjunctions
adverbs of various kinds
To use this text just to focus on verb grammar would be a waste. The text should and can be dealt with in all its aspects – that’s the beauty of a short text. If learners beyond a certain point are not reminded of aspects of grammar regularly they will never come to master it. If you don’t deal with articles in this text, when will you do so? If not these prepositions now, when? If not the noun phrases here, which and when? Some grammar will be the main focus; other grammar will be mentioned in passing but all should be dealt with.

**Level 2 Topic and Comment, and Lexis**

**Text Features 2** shows the topic/comment [or theme/rheme] structure of the text. In this text each sentence begins with the same topic, which was established at the beginning of the text: "A crook".

There are other possible patterns for topic/comment, for example:

**Topic 1 + Comment on Topic 1**

Comment on **Topic 1** becomes **Topic 2 + Comment on Topic 2**

Text, both spoken and written, is structured in these topic + comment ways. A topic is introduced at the beginning of a sentence and then a comment is added about it. Topics can be brought to the beginning of sentences (using inversion), or to the start of paragraphs, in order to given them prominence, and the order of topics in a text is a key component of coherence. Learners rarely read enough text to imbibe the possible patterns through the brute force of exposure. The patterns need to be explicitly explored. Ask your learners to identify the sentence topics and analyse the sequence of topics and comments.

**Text Features 2** also shows some of the lexical chains and fields which are in the text.

- carpet + rug + carpet
- flat + flat
- balcony + balcony
- police search + the police arrived to search + detective
- held up a shop + armed robbery
Lexical chains and fields also help with coherence and cohesion. They help tie the text together. A text is about something and we expect a particular set of words related to that topic to be in the text and we expect variations in expression – not just the same words repeated throughout the text.
Level 3 Information

Text Features 3 shows the structure of information in the text. Topic and comment is a sentence level analysis but the information-focus develops throughout/within the sentence. A comment is often made up of more than one piece of information. ‘Tom is a cat’ is a simple sentence and has limited information. ‘A crook tried to escape a police search by hiding in a carpet’ has more information, and the focus changes from ‘crook’ to ‘search’ to ‘carpet’, and information is given about each of the foci.

The way information is organised within sentences is another key part of coherence. Learners should be asked to analyse what the key foci are in a sentence and what information is given about each: a + police + search.

Not only that but information structure leads us back into grammar [verb and noun phrases etc.] and the role of reference devices [for coherence and cohesion again]. How are the noun and verb phrases constructed around the key foci? How and why are articles used? How do the verb phrases connect the foci of the sentence?
Dealing with a Text

A text should be dealt with globally first, perhaps with comprehension or who, what, why, where, when, how questions for a text like this. Then the topics and comments should be highlighted and studied – why this sequence? Then the foci and information should be identified; then the grammar of noun phrases can be

Note: F8 and F11 include and expand upon earlier Foci: F7 and F4 respectively.

Key
F = Focus;
F1 = first focus = Topic of first sentence.

[brackets F1] enclose connected information about a topic e.g. [a police [search F2]] shows that 'a police' is information about the 'search'.

[brackets] without an F are just information.

Text Features 3

[A [crook F1]] [tried to escape] [a police [search F2]]
[by hiding in] [a [carpet F3]].

[Stanislaw Dabrowski F1], [38], [held up] [a cosmetics shop F4] [and] [then escaped to] [his aunt's [flat F5]]
in] [Warsaw F6, [Poland]].

[The [thief F1]] [rolled himself up in] [a giant [rug F3]]
[and] [leaned himself up against] [a balcony [wall F7]]
[when] [the [police F2]] [arrived to search] [the [flat F5]].

[He F1] [was only found] [when] [a [detective F2]]
[went out onto] [the [balcony F8]] [for] [a [cigarette F9]] [and] [noticed] [the [carpet F3]] [was moving].

[Dabrowski F1] is now facing] [a 12 year prison sentence F10] [for] [armed robbery F11].
explored and how the verbs tie noun phrases together and how prepositional phrases add meaning.

Unfortunately, these three levels of analysis cannot be done with electronic tools. It needs a teacher and teacher knowledge to do this, which is not a bad thing; in fact it’s a good thing. The machines haven’t won yet.

There are things which you can ask a computer to do with a text though and these will be explored in the next section.

Text Analysis Tools

The Compleat Lexical Tutor is the go-to site for these analyses: www.lextutor.ca

The site is divided up into sections for learners, researchers and teachers. All parts are worth exploring. We’ll look at three tools: Vocabprofile, Text Lex Compare and Concordance.

Using the Vocabprofile tool you can analyse a text in terms of the lexical frequency and range. This is important because research into vocabulary size and coverage has shown that the most common words in the language account for most text. The table below shows that the most common two thousand words
cover 81.3% of text. After learning these the learner is left with a huge task of learning thousands more words to achieve significant gains in coverage.

<table>
<thead>
<tr>
<th>Different words</th>
<th>Percent of word tokens in average text</th>
</tr>
</thead>
<tbody>
<tr>
<td>86,741</td>
<td>100%</td>
</tr>
<tr>
<td>43,831</td>
<td>99.0</td>
</tr>
<tr>
<td>12,448</td>
<td>95.0</td>
</tr>
<tr>
<td>5,000</td>
<td>89.4</td>
</tr>
<tr>
<td>4,000</td>
<td>87.6</td>
</tr>
<tr>
<td>3,000</td>
<td>85.2</td>
</tr>
<tr>
<td>2,000</td>
<td>81.3</td>
</tr>
<tr>
<td>1,000</td>
<td>71.1</td>
</tr>
<tr>
<td>10</td>
<td>23.7</td>
</tr>
</tbody>
</table>

Table 1: Text/Word coverage (Carroll, Davies and Richman, 1971)

Lui and Nation (1985) suggests that guessing word meaning from context is only effective when the person already knows about 95% of the words of a text. To achieve this level of knowledge students would have to learn another 10,500 words. Clearly this is an almost impossible task for most learners.

The alternative is to choose a more restricted set of words from a particular genre of English. This will give high coverage within that specific genre. One attempt to do this is the Academic Word List (AWL). The AWL was developed by Averil Coxhead (2000) and is a set of words which are found in a wide variety of academic texts. It consists of 750 word families. Combined with the first 2000 words, the AWL provides over 90% coverage of a wide variety of academic texts. So university students who know these words will have fewer problems reading academic texts and be more effective at guessing unknown words.

We’ll look at how to create such word lists later but now we’ll look at the basic analysis of a text in terms of word frequency.

The Classic analysis on Compleat Lexical Tutor analyses a text in terms of which words [and how many] are in the 1st 1000 most frequent words in English [K-1]; the 2nd most frequent 1000 words [K2] and the AWL. Part of the graphical output is the colourful analysis below.
K1 words are in blue and make up 67% of the text. K2 are in green and make up a further 15%. AWL words, in yellow, are just over 1% and off-list words [in red] account for the remainder – almost 17%. This view and analysis will give you an idea of the level of difficulty of a text and the density of ‘difficult’ words. The off-list count is quite high so I ran the same text through the Complete version of this analysis which has word lists up to the 25,000 most common words, as derived by an analysis of extremely large corpora. The analysis in Screenshot 3 below shows that all but a handful of the off-list words above words occur in K-4 and K-5 and what we are left with are proper nouns like Warsaw and Poland and the name of the thief and these are ‘knowable’, that is the leaner can recognise them as names.
A word of warning though – frequency list are only as good as the corpora on which they are based and, as the words ‘airplane’ and ‘airport’ do not seem to occur in K-25 lists used here, there are some weaknesses in these lists.

**Paul Nation’s Programs**

http://www.victoria.ac.nz/lals/about/staff/paul-nation

The program on the Compleat Lexical Tutor above is based on Paul Nation’s work. You can go to Nation’s website and download the original programs and many other things, like vocabulary tests, wordlists etc.

<table>
<thead>
<tr>
<th>Range Program Files</th>
<th>File size</th>
<th>File type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range program with GSL/AWL list</td>
<td>521 KB</td>
<td>ZIP</td>
</tr>
<tr>
<td>Range program with British National Corpus list</td>
<td>701 KB</td>
<td>ZIP</td>
</tr>
<tr>
<td>Range program with BNC/COLCA lists</td>
<td>590 KB</td>
<td>ZIP</td>
</tr>
<tr>
<td>Information on the BNC/COLCA lists</td>
<td>156 KB</td>
<td>PDF</td>
</tr>
<tr>
<td>Headwords of the first 10,000 words</td>
<td>40 KB</td>
<td>ZIP</td>
</tr>
</tbody>
</table>

Download the **Vocabulary Resource Booklet**. This contains the Vocabulary Levels Test, bilingual 1000 and 2000 versions of the Vocabulary Levels Test, the survival vocabulary in various languages, and other tests and resources.

<table>
<thead>
<tr>
<th>Vocabulary Resource Booklet</th>
<th>File size</th>
<th>File type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary Resource Booklet</td>
<td>1,215 KB</td>
<td>ZIP</td>
</tr>
</tbody>
</table>

**Screenshot 4: Paul Nation Downloads**

Nation’s versions of the range programs can be used to analyse sets of texts, rather than just one as on the Compleat Lexical Tutor, but the output is in black and white. The interface is very simple and it comes with straightforward instructions.
The key benefit of using this Range program is that you can create your own wordlist and use it instead of the AWL, for instance, and see which words in your list occur in the text[s]. Imagine creating a specialised list and analysing a text with it. You’d be able to tell at a glance how much of the text is covered by your list.

It’s also (relatively) simple to create your own wordlist. Here’s how:

All you do is quickly construct a micro-corpus of a specialised topic from the Internet and produce a key word list using the Nation program. Unfortunately, this is only available for a PC. You will also need some computer skills and the most important are being able to cut and paste text, save files as different kinds of files, being able to convert table to text and text to table and to easily delete rows and columns in tables.

The first step is to create your micro-corpus. My latest corpus is two and a half million words but yours need not be so large, even 10-20,000 words will give you useful information. The easiest way is to go to the Internet and find texts there.

As an example let’s imagine creating a corpus about ‘cricket’. You can google for specialist websites about cricket and go to the BBC to get news reports. Think
about your students – do they want to read news stories about cricket? If so, then your corpus should include news reports.

You need to save the texts you are interested in as .txt files, not as .doc files. This is because these programs analyse plain text, nothing with tables or fancy formatting. The best way to do this is to cut and paste the text you want into a text file rather than saving the whole web page because then you’ll have a lot of text you don’t want like advertisements etc. Find as many texts as you can about cricket and save them on your desktop in a folder. Long texts are better. The Nation Range program can process 32 texts at one time so aim to get 30 texts.

Transfer all your texts into the same folder as the Range program you have downloaded as this will make it easier to find them.

Open the Range32 program by clicking on the icon. Click File and Open and select all the texts you want to analyse. Then click File and Save and enter a file name for the results of the analysis. Then on the screen below Process Files click on Basewrd3 to remove the tick. This is because we want to run our texts against the Basewrd1 (the most common 1000 words) and Basewrd2 (the next most common 1000 words). This will tell us which words are not in these. Also click on Sort by Freq and this will tell us how common the words are.

Now click on Process Files to run the program. This should be very quick. Close the program and open your results file which will be a ‘_range.txt’ file. This will be quite large. Go down the file until you see a table which will tell you the percentage of words found in word list one and two and ‘not in the lists’. Have a look at what percentage is not in the lists. These are the words we will use to create our key word list.

Scroll down the file until you reach the section called Types Not Found in Any List. It will be a list of Type with a Range, Freq, F1 and F2 etc. Delete everything apart from this list. Also delete the last part of the file after ‘time taken was’.

Now comes the tricky part. Save this .txt file as a word document. Then open it and convert the list into a table. Delete the F1, F2 etc. columns. Then go through
and delete the rows with words which are proper nouns (all the names) or fragments of words.

The next step is to review the other words. You will have to decide which to keep in your list and which not. Use the range information and frequency information to help you. The range tells you how many texts the words occurred in. If you had ten texts and the range was 10 then it is a widely used word in your corpus. Also look at the frequency. If it occurs once in one text then perhaps it is not an important word. If it occurs at least once in all texts then perhaps it is a useful word. Go through and evaluate all the words and then when you are finished delete all the columns apart from the Type column. Change this column back into text.

Now you have your key list of words which do not occur in the 1st 2000 most common words of English. It is a good idea to sort the words alphabetically at this stage by selecting all the words and then choosing Sort from the drop down Table menu.

You can leave this as your Key Word List or you can extend it a bit by adding all the members of the word families. ‘Cricket’ should be one of your words but you want to make sure that ‘cricketing’ and ‘cricketer’ are there as well. Similarly, you should have the word ‘bowler’ in your list. Now check and add ‘bowl’, ‘bowled’ and ‘bowling’.

You could stop at this stage and just give your list of words to your students or use the list as a reference for you to highlight important words or create exercises but you could make it into a word list you can use with the Nation Range program. Here’s how.

Your words should already be in capitals, if not change them all to capitals. Then after each word you need to add one space and a zero. Family members should be indented with a tab. Your list should look like this:

```
BOWL 0
  BOWLER 0
  BOWLING 0
```
Go to the Nation program folder and rename Basewrd3 as ‘Basewrd3original’. Save your word list file as a .txt file. Call it Basewrd3.txt and move it to the vocabulary program folder if it is not there already. When you run the program again it will use your list to analyse the texts. Run the Range program again with one of your corpus texts, making sure that Basewrd3 is ticked (not like last time), and see what coverage your list gives. If you have about 10% coverage then your list is a very good key word list for that topic.

If your students know the word families of the most common words in English and the words in your list then they should have fewer problems reading texts on the topic of cricket. Perhaps you could even teach your students how to create their own key words lists.

**The Compleat Lexical Tutor: Text Comparisons**

The next analysis at the Compleat Lexical Tutor we will look at is the Text Lex Compare where you can compare two texts and find out how much lexis they share. Just paste two texts into the interface page. The output below in *Screenshot 6* is of a comparison of two versions of the “Crook in a Carpet” text. There are 9 words which are not in the first text. All the shared words are underlined.
This kind of tool gives you the opportunity to create a set of texts which share varying proportions of lexis – an invaluable tool.

The Compleat Lexical Tutor: Text Concordance

At the Compleat Lexical Tutor it is also very easy to create a concordance from a whole text, or collections of texts. Just copy and paste your text into the interface, press the button and the program will concordance every word in your text.
Concordance Lines from Corpora

You might want concordance lines from other sources and you can go to publically accessible corpora like the Corpus of Contemporary American English, put in your word [in the screen shot below: sex] and see the word in all [or a selection of] the contexts it occurs in in the corpus. This presents a massive amount of data in a concise way. You can see at a glance the collocations of the word and make a selection of the ones which you think are most important for your learners.

![Screenshot 8: Sex Concordance](image)

The British National Corpus (BNC) corpus will give you a random selection of collocations from the corpus but these will be in complete sentence contexts and some learners might find less intimidating than the kind of concordance lines above.

This leads us into a consideration of corpus linguistics.

**What is Corpus Linguistics?**

Corpus linguistics is the study of collections of texts.
There are different types of corpora:

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>written:</td>
<td>A collection of written texts: the easiest corpus to collect.</td>
</tr>
<tr>
<td>spoken:</td>
<td>A collection of recordings and their transcripts. Fiendishly difficult and time consuming to produce.</td>
</tr>
<tr>
<td>multi-media:</td>
<td>Video recordings and associate transcripts. Even more tricky than spoken corpora.</td>
</tr>
<tr>
<td>historical:</td>
<td>A collection of old texts.</td>
</tr>
<tr>
<td>learner:</td>
<td>A collection of texts produced by learners e.g. essays; exam writing</td>
</tr>
<tr>
<td>parallel:</td>
<td>A corpus in two languages with texts running side by side.</td>
</tr>
<tr>
<td>synchronic:</td>
<td>A corpus which is focused on language at one period of time</td>
</tr>
<tr>
<td>diachronic:</td>
<td>A corpus which is used to study language as it changes over time.</td>
</tr>
</tbody>
</table>

Table 2: Types of corpus

A key thing to remember is that: one text is not a corpus.

Publicly available corpora include:

- British National Corpus:  [http://www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/)
- Corpus of Contemporary American English:  [http://corpus.byu.edu/coca/](http://corpus.byu.edu/coca/)

**Building a Micro Corpus**

A micro corpus is a small corpus, not like the huge general corpus like COCA with 450 million words.

Building a micro corpus is easy.

1. Choose a topic, say ‘basketball’.
2. Find a set of texts on the topic. Ideally find pdfs or copy webpage texts as text files; if pdfs then convert them to text files.
3. Save the texts on your computer in a folder.
4. There you are; you have a corpus.

There are, though, some issues to consider about your corpus, not least because
you might have broken some laws.

**Representativeness**: is your collection of texts representative of that part of the language? If you want a corpus of basketball texts, don’t have any texts about football; but also try to have a set of texts from a variety of sources, unless of course you are studying something very specific like ‘basketball reports on CNN’. Representativeness is an ideal and difficult to realise. You should try to have a corpus which is as representative as you can manage.

**Balance**: is your corpus balanced? If you have a corpus of sports texts, are the number of texts and the word count in these texts balanced so that one text or one sport is not over-represented in your corpus?

You should carefully consider the questions of **copyright** and **ethics**. Are your texts in the public domain? Or are they subject to copyright? Can you legally keep texts on your computer? Some researchers claim they can avoid copyright issues by never publishing their corpus as a collection of texts. Some researchers just list the sources for their corpus and if you wanted to replicate their research you would have to assemble the same corpus. You should carefully investigate the copyright laws where your live and act accordingly.

Ethical questions are rarely considered in corpus linguistics. Is it ethical for you to study the texts you have collected?

**Case 1**: If you want to study the language of a publically viewable and accessible internet discussion which everyone can see and contribute to, is it ethical to copy the forum posts and analyse them?

**Case 2**: If you have to create a profile to join such a discussion forum as in Case 1, is it then ethical to copy the forum posts without telling the posters what you are doing?

Your ethics will have to guide you in such cases. Whatever you decide it is very important to **anonymise your data** like recordings of people so that the people who said or wrote what they said or wrote cannot be identified.

**Noise** is another issue in corpus data. If you include a report in your corpus and the report has page numbers, chapter headings on each page and tables of data,
this kind of information will create noise in your analysis. The word ‘Africa’ may occur in a lot of tables and therefore appear to be very common. It may push more interesting words further down the frequency rankings. If you took out all the page numbers and table data and chapter page headings you would reduce the noise and make it easier to find the key words in the continuous text of the document.

A Corpus Creation Case Study

The table below [Table 3] shows a proposed corpus which focuses on intelligence texts.

<table>
<thead>
<tr>
<th>The Corpus</th>
<th>No. of Texts</th>
<th>Sample</th>
<th>Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-corpus 1A: US Intelligence Community</td>
<td>22</td>
<td>100% sample</td>
<td>370,655</td>
</tr>
<tr>
<td>Terrorism Reports/ Guides/Briefings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-corpus 1B: Private Sector Terrorism Reports</td>
<td>20</td>
<td>Convenience sample</td>
<td>401,367</td>
</tr>
<tr>
<td>Sub-corpus 2A: State Department Terrorism</td>
<td>13</td>
<td>100% sample</td>
<td>1,429,942</td>
</tr>
<tr>
<td>Reports</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-corpus 2B: State Department Cables [Secret]</td>
<td>332</td>
<td>Secret/Noforn Sample of 251,287 texts</td>
<td>396,122</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,826,064</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,598,086</td>
</tr>
</tbody>
</table>

Table 3: A Corpus Construction Case Study

Decisions and Issues Commentary

1. The corpus is a collection of US texts. A decision had to be made whether to study texts from one or more language varieties.

2. The second decision was about which aspect of ‘intelligence’ would be studied. It would have been perfectly possible to construct an intelligence and climate change corpus. Terrorism and counter-terrorism were chosen.

3. Sub-corpus 1A is a collection of 100% of the publically available US government reports on terrorism from the US intelligence community. As such it is completely representative. There are 22 texts and a total of just over 370,000 words.
4. Sub-corpus 1B is a collection of similar texts from privately funded think tanks in America; their agenda is to influence the US government. It is a **convenience sample** of the available reports and thus not as representative as 1A. There is a similar number of texts though and a similar word count.

5. Sub-corpus 2A is another **representative** sample of terrorism reports from the US State Department. The issue here is that the word count is 1,429,942, which is much higher than any other part of the corpus and so the corpus is unbalanced. A solution would be to take three of the 15 texts in this sub-corpus and just use those, with a word count of approximately 400,000. This would restore balance in the word count but would mean that one part of the corpus was heavily weighted in favour of only three texts, and some texts would not be included. Representativeness and balance are in conflict here.

6. Sub-corpus 2A is a sample of a set of 251,287 texts, totalling tens of millions of words. The decision was made to sample texts by searching for *Secret/Nofo* [No Foreigners] cables with the key words of terrorism and counter-terrorism. This search found 332 full cables totalling 396,122 words, a nice fit with the word totals of the other sub-corpora. These cables are shorter than the other texts in the corpus and it is debatable whether they should be studied together with the long reports of 1A, B and 2A.

7. There is also an **ethical question** surrounding the study of stolen data. These cables were leaked in an act of espionage by US Army Private Bradley Manning and it is illegal for US servicemen to access the Wikileaks site where they are held. The cables are [not now] secret and stolen – can they be studied? Are they now in de facto public domain?

8. There is a lot of noise associated with the texts. This includes tables of data in the reports and the headers in the cables [see **Screenshot 9** below]. The texts should be tidied up, while carefully preserving the original files.

9. The easiest way to keep a record of your corpus is to create a separate document with data about the texts, including source, text types, topic, word count etc. It is fashionable to include metadata and the corpus texts together as .xml files but that’s not really necessary. Text files [.txt] and a separate corpus information file will be fine.
As you can see from the discussion in the case study above there are some questions to be asked and decisions to be made when constructing a corpus but if you keep your corpus straightforward these should not be big issues.

**The Reference Corpus**

To analyse a corpus you also need a reference corpus so you can compare your corpus with it. This should generally be bigger than your corpus but of the same type. For the corpus in the case study above the written component of an American English corpus would be appropriate. The Open American National Corpus mentioned above is suitable and freely downloadable. There is an issue with its contents though: it contains the 911 report into the intelligence failings before the September 11th attacks in New York. This conflicts, by being similar in topic to the texts in the corpus above, and would have to be removed from the reference corpus.

**Analysing a Corpus**

With electronic corpora you need a program to analyse them. The classic is Wordsmith Tools but you have to pay for that. There are a number of free
options though. Three are listed below.

**Free Programs**

Microconcord: [http://lexically.net/software/index.htm](http://lexically.net/software/index.htm)
AntConce: [http://www.laurenceanthony.net/software/antconc/](http://www.laurenceanthony.net/software/antconc/)
[in German]

Microconcord can be used to create concordance lines, while AntConce can also be used to create word lists and compute key word lists, the classic entry point into corpus analysis.

**Key Word Lists**

Key Word lists are lists of words which are more frequent in your corpus than they are in the reference corpus. The exact statistics behind this are not very important for us. As the words are more frequent they are likely to be more important in the texts in your corpus, so they are the most interesting words and the best ones to study. It’s easy to make a key word list for your corpus.

1. Make a word list from your corpus. ‘The’ will be the most common word in a text corpus.
2. Make a word list from your reference corpus. ‘The’ will also be the most common word in your reference text corpus.
3. Compare the word lists using your corpus analysis program: sit back while it creates a Key Word list.
4. ‘The’ will probably not be the most key word in your Key Word list.

Once you have a Key Word list you can start to explore the collocations and even colligations [the grammar] of the key words: where do the words appear in the text, in the paragraph, in the sentence? Are the verbs usually used in the active or passive? Are the words found in positive or negative contexts? A thousand and one questions can be answered with a corpus.

There you are. You are now a corpus linguist.
Concluding Thoughts

I think teachers should understand a text thoroughly before they teach it and that texts should be dealt with in the classroom at all levels of analysis – even more if you want to consider style and other aspects. Information structure, noun phrases, and coherence and cohesion are much neglected in teaching materials and if we do not draw the learners’ attention to these features of text they will be seen but not noticed; understood at some level, depending on the questions we ask about the text, but not really processed; half-forgotten, not learned. And that’s less than desirable.

Building a corpus can help you meet your learners’ needs because you can build a micro corpus to answer specific questions, like what language/words do my learners need to know to read [and write] music reviews or sports reports?

Can learners act as researchers? Yes, if they are interested enough and you help them. It could be a small project and need not take too long. Micro corpora can be very small and still answer important questions. Learner built corpora can help them meet their needs independently.

Should we use concordances in the classroom? Some books have been written on this very idea and some learners might find it useful – a collection of concordance lines can give a lot of information about a word but some learners might not like the fragmentary nature of the data. University learners might be more receptive.

What do we mean by knowing a word? This is a ‘big’ question and one for which Firth has a good answer; we know a word from the company it keeps: the collocations. How and when a word is used is different from its ‘pure’ meaning. Words ‘mean’ when they are used, not while sitting listed in a dictionary, or waiting in the mind to be recalled and used. Text analysis and corpus analysis can help you and your learners understand what a word really means by considering when and where and how and why it is used.
References


Robert Buckmaster presenting at the Latvian Association of Teachers of English Annual Conference in 2014.

Rob has over 24 years of experience as a teacher, trainer, examiner and consultant in Eastern and Central Europe, Central Asia, North Africa and the Middle East. He works as a lecturer at the University of Latvia as well as being a consultant for McIlwraith Education. He provides Language Audit services for companies and works as an independent corpus linguistics researcher. He regularly presents at conferences in the Baltic region and is on the Board of the Latvian Association of Teachers of English. He has also trained teachers for Macmillan and Cambridge University Press. He has written widely on English language issues with articles in the Guardian, EFL Gazette, Modern English Teacher, English Teaching Professional, the British Council PEP Newsletter, and was the first columnist for the British Council's Network News. In 2014 he published the Grammar of English Ideas, which is a reimagining of pedagogical grammar.

www.rbuckmaster.com
www.mcilwraith.org
www.englishideas.org