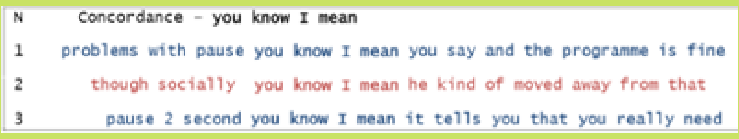
# How to use a corpus:

## A teacher educator's guide to supporting the creation of a corpus

Okay, so we are not talking about a making a multibillion-word general reference corpus here or going against the good sense of using the long list of robust ready-made corpora. But the experience of putting together a small specialized corpus can build teachers' trust in corpora and confidence in their skills to use them.

## 01

**Discuss distrust.** Do you really know what is in a corpus? Those who are new or newish to using corpora need to see that it is real language in use (or as close as we can get). We are right to be suspicious of bits of language taken out of their social context of use and presented in centre-aligned lists running down a page. Critical and contemplative human beings, the ideal candidates to be an English language teacher one would have thought (!), ought to be doing just that. Where do those words come from and who said them? How do I know, for sure, that 'no one' is more frequent than 'nobody' in academic writing? Part of the criticism (and downright distrust) of corpora is that they take language out of our familiar everyday communicative experiences. We don't recognize the chat in the coffee shop when we see it written down in snippets four words to the right and left of a search term, and newspapers just don't present words like that.



```
N      Concordance - you know I mean
1   problems with pause you know I mean you say and the programme is fine
2        though socially  you know I mean he kind of moved away from that
3          pause 2 second you know I mean it tells you that you really need
```

VS

So, why should new or experienced teachers trust that it is real and people really do use language in that way? A book is real. A video is real. But are concordance lines, er, real? There is much discussion on what any corpus can represent. One way of tackling those questions and doubts head on is through the activity of creating one's own corpus. It is important not to be overzealous about corpora or hide behind the tech. Guide teachers to their uses. Be open and ready to have good critical discussions.

## 02

**Identify the reasons for creating a corpus** and make sure you decide on one before you start. Here are our top five reasons:

- You can't find a corpus that matches your specific needs, interests, or what it is that you are interested in (it might be something subject specific).
- You need to examine your own set of texts, or your students.
- You want to understand the nuts and bolts of a corpus.
- You want to get more hands-on experience of working with a corpus.
- You want to carry out research on documents, materials, and other texts.

The chances are that when we approach creating a corpus with a specific reason for doing so, rather than just creating one for the sake of it, our corpus-creating experience will be a meaningful one with long-lasting effects. Teacher development time is precious- there is a danger that guidance and workshops on how to use corpora can be seen as the latest training on a trendy topic. In reality, the skills that are gained by creating a corpus can help to sustain professional development for many years thereafter.

# How to use a corpus:

## A teacher educator’s guide to supporting the creation of a corpus

### 03

**Help teachers build their corpus according to the** principles of corpus construction. We are a good number of decades into corpus studies. There are a number of accessible handbooks, publications and resource books which provide in-depth discussion on what makes for a principled collection of natural texts. Key to these discussions are concepts of representativeness (see #1), balance and sampling. At their most basic level these are concerned with: what texts do I include, how many, and how do I select? What textual and contextual information do I include with the text? Do I need to annotate? But a more tenable approach to engaging teachers with these concepts is to support their reading of such publications as:

- Corpus-based Language Studies by Tony McEnery, Richard Xiao, Yukio Tono (2006)
- The Routledge Handbook of Corpus Linguistics by Anne O’Keeffe and Michael McCarthy (2010)
- Exploring Corpus Linguistics: Language in Action by Winnie Cheng (2011)
- Corpus Linguistics for ELT by Ivor Timmis (2015)

This will ensure that the analytical skills and depth of knowledge about language that they develop through the activity of creating a corpus will sustain them throughout their careers.

### 04

**Advise care and caution when compiling.** In #1, we are reminded to ask ourselves: Where do those words come from? Who said them and who wrote them? Before including a text in a corpus, it is important to consider these questions in moral and economic terms too. Do you have permission to use the text? How do you know and how can you confirm it? Is it owned? Do you have written permission from the speaker in the café to record their conversation? Will texts really be anonymous? Guidance through ethical issues are crucial and it is important to consult your own institutional requirements, guidelines and advice. In addition, professional association recommendations on good practice are a great resource. See for example:  British Association for Applied Linguistics. 2016. "Recommendations on Good Practice in Applied Linguistics, 3rd edition". British Association for Applied Linguistics available here: https://www.baal.org.uk/who-we-are/resources/

### 05

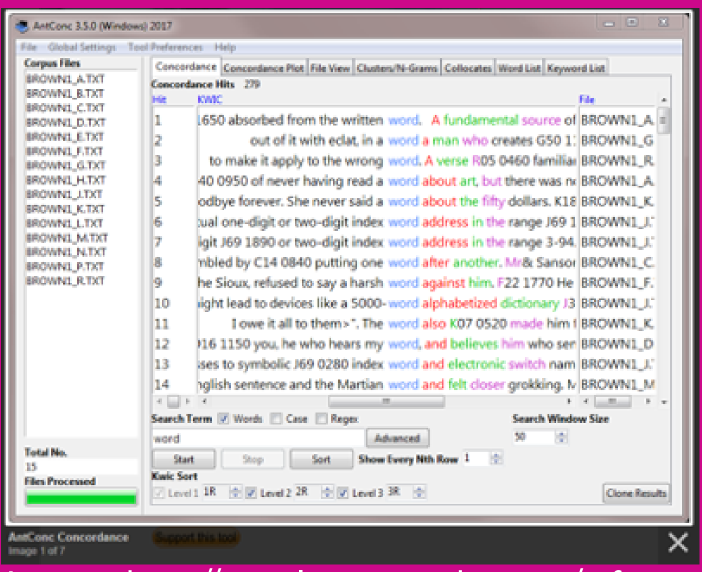
**Remember that reformatting is not just about layout or tech specs.** Okay, maybe mostly, but it is a useful reminder of #1. Preparing texts for inclusion in a corpus can take a bit of time. It is not always a simple copy and paste, though this is possible on some websites 😊.  Normally, it involves taking the texts from their original format, even if written, and converting them into something which can be processed by corpus software or websites (see further discussion of these in #6). We can think of a text as the result of somebody doing something with words usually for a particular purpose or the capturing of an instance of people making meaning in their lives. But, to put it bluntly, reformatting is about taking that text and all its meanings and converting it into a series of binary numbers to make it machine readable. This may seem harsh, but it is important to recognize, nonetheless. The software, technology and machinery will only process input it can read, and only then will it perform functions to compute frequency, collocates, clusters and distributions as a result of the instructions provided. Therefore, every character, formatting strike, and layout feature counts.

Making Best Use of Corpora for English Language Teaching

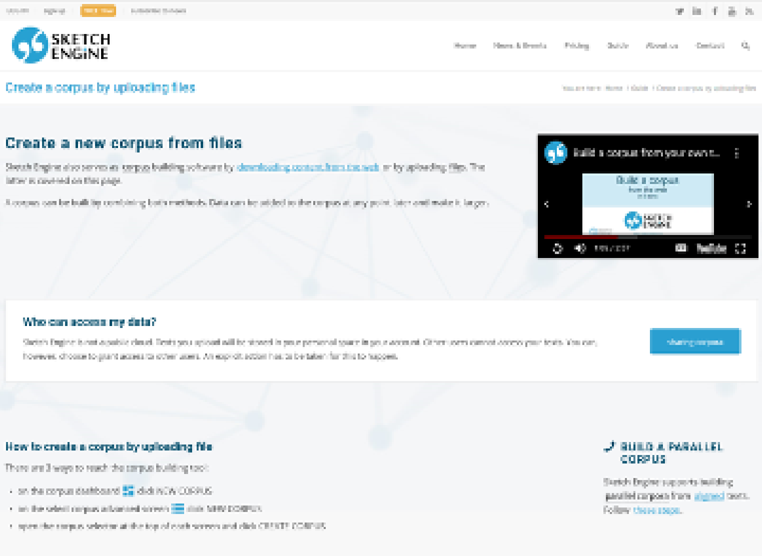
# How to use a corpus:

## A teacher educator's guide to supporting the creation of a corpus

**06** **Suggest software and websites.** You don't need to reinvent the wheel. There are several reliable and well-known corpus software tools and websites which will give you what you need to create and search a corpus. Some are free, offer a free trial, require you to purchase a one-off individual or site license, or pay a monthly/annual subscription fee. Below are screenshots of some examples:

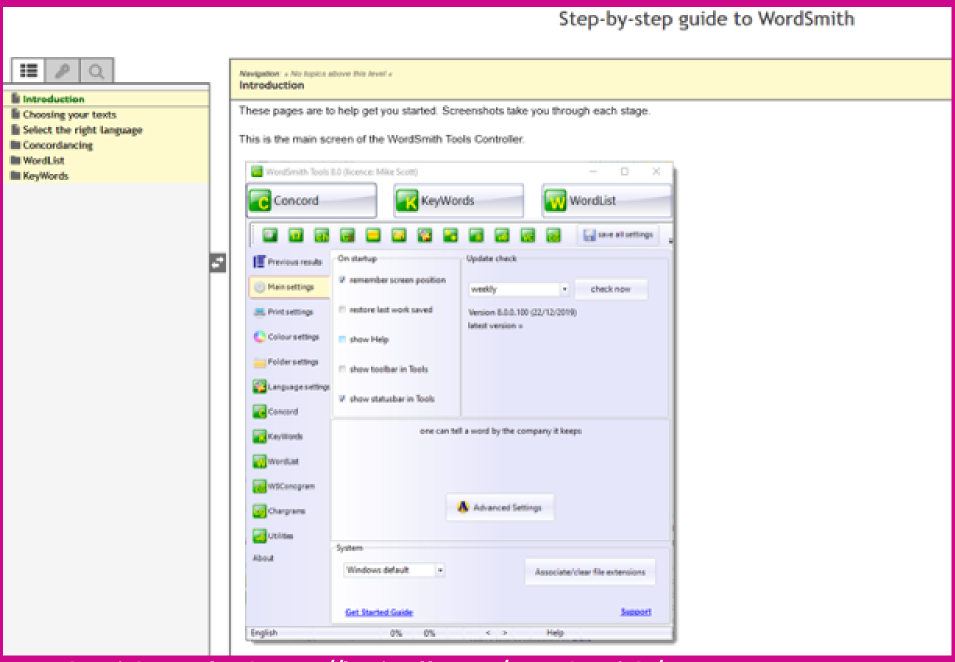


Antconc: https://www.laurenceanthony.net/software/antconc/



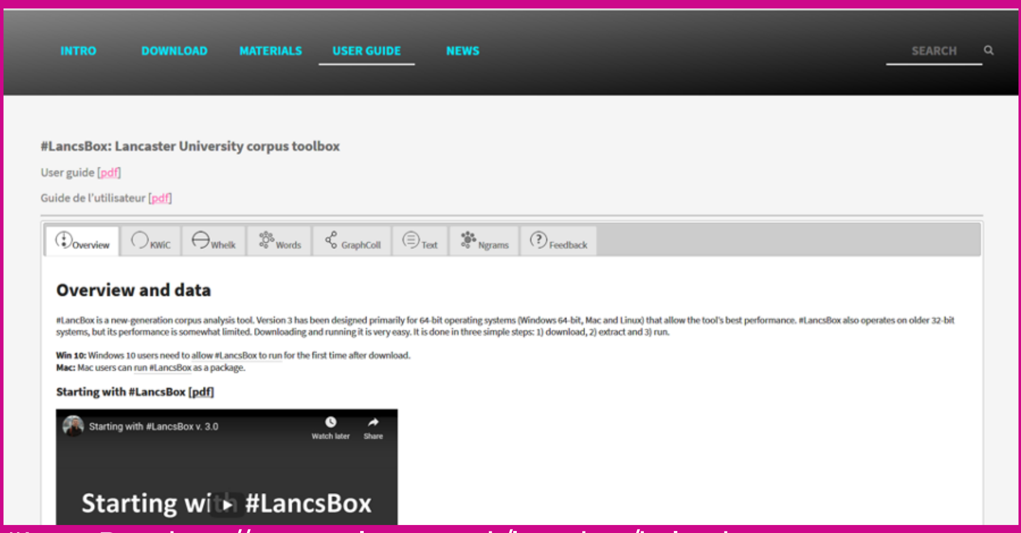
SKETCH ENGINE
Create a corpus by uploading files
Create a new corpus from files
Who can access my data?
How to create a corpus by uploading file
BUILD A PARALLEL CORPUS

Sketch Engine: https://www.sketchengine.eu/guide/create-corpus-from-files/



Step-by-step guide to WordSmith
Introduction
Choosing your texts
Select the right language
Concordancing
WordList
KeyWords
These pages are to help get you started. Screenshots take you through each stage.
This is the main screen of the WordSmith Tools Controller.
Concord
KeyWords
WordList
Get Started Guide
Support

Wordsmith Tools : https://lexically.net/wordsmith/



#LancsBox : http://corpora.lancs.ac.uk/lancsbox/help.php

Consider all your options and work out what will be best for you and your students. A key factor to consider is how much guidance there is for the user. This may vary in form – from explanatory texts to video demonstrations. Think also about any extras and updates that are made available.

@AislingOBoyle1 @vanderviana

# How to use a corpus:

## A teacher educator's guide to supporting the creation of a corpus

## 07

**Remember to refresh the basics.** With a focus on compiling, gathering and uploading texts, it is possible that the reason for creating the corpus in the first instance has slipped into distant memory. Refreshing the fundamental tools of a corpus exploration is a good way to kick-start the action of corpus exploration.

Hello, remember me?

- Concordance: this tool will display all the examples of a chosen word or phrase in context. It will give access to information about collocates, dispersion plots and cluster analyses.
- Word list: this tool generates a list of all the words contained in a file. It can generate these in order of frequency and it can identify which words frequently cluster with other words. Lists from different files can be compared.
- Key word: this tool allows the development of a list of words which are unusually high frequency in comparison with a designated norm (i.e. from a general reference corpus). This can be used to generate a way to characterise a particular text or genre.

## 08

**Prepare to have beliefs challenged.** Once we start to work with real language data, certain tightly held beliefs and intuitions can be challenged by what we find. Some of these will be our knowledge and beliefs about language and language teaching. Corpus investigations tend to dissolve the usually neat pedagogic distinctions we make between grammar, vocabulary, pragmatics and so on. Access to varieties of English leads us to the question- whose English are we teaching? testing? producing materials with? And once we start working with real language data our social and cultural beliefs may also be tested. For example, do you think this statement is true or false?:

**Female speakers don't use taboo language in academic contexts.**

(Note: a quick taboo word search of MICASE will help you with this: https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple).

The point is - be prepared: it is not a question of if these challenges happen, but when.

## 09

**Don't stop! Analysing, sharing, adding, deleting, teaching . . .**

One of the reasons why corpus building activity is time well spent is that a corpus is a very reusable object. Although the original plan may have been to create a list of key words in a specific subject area, a multitude of other further investigations with the same corpus are now possible. Sharing a self-made corpus with other teachers, colleagues and students can be very rewarding and lead to many and varied learning opportunities. Showing others how to use a corpus is a good way to develop their knowledge and skills but it also strengthens the teacher's. Adding more or different files (and deleting if appropriate) extends the use of the corpus. The more that can be done with the same data, the better - it really does reward the time and effort of corpus creation.

# How to use a corpus:

## A teacher educator's guide to supporting the creation of a corpus

10

**Evaluate the experience.** Okay, so it has been a challenge. But the activity of creating a corpus helps to settle some of the research-related insecurities that teachers often talk about. It is important to be honest about the cost and benefits of the activity. And think of what these are not just in the immediate short term, but also in the longer term.

Prompt these kinds of questions:

- Which skills have you gained and can you name them?
- What do you know now that you didn't before?
- What will you do with those skills and knowledge in future practice? What do/did you struggle with?
- How do you think you can overcome challenges?
- Ultimately, and speaking honestly - for you, for your students, for your professional practice, for your career - was it worth it?