One obvious thing that you might want to do, of course, is build your own corpus and look at different sorts of data. Now as some of you are teachers, I believe, one interesting project could be for you to look at a corpus of student writing, or if your very ambitious student speech that you might collect in some small-scale study yourself. Obviously, the data that you would collect wouldn't be as large as the millions of words in available so-called learner corpora, corpora of learner writing and speech, which you'll find out about later in the course. But you would have more control over what went into the corpus. And you would, indeed, be much more familiar with the content of it, having collected that material yourself.

So let's imagine that you want to build your own corpus. Now, Kennedy says that there are five stages which go into compiling or creating a corpus. There are the design, planning a storage system, obtaining permissions, text capture, and markup stages. We'll look at each of these in turn.

So perhaps the most important stage of all is the very first one, design. Because without a solid design, everything else is likely to go wrong. When we design a corpus, we need to bear in mind what it is going to be used to do. What research questions are we going to ask of that corpus? And what other corpora are we intending to compare it to?

So questions might be asked such as, for example, are we going to look at speech or writing? We need to know that before we build the corpus. Are we going to look at different time periods and contrast those? Again, we need to know that. If we want to look at historical change, we need to build a corpus to enable that.

So a broad range of questions we might ask. But one question is undoubtedly in your mind. How big will the corpus have to be?

Well, it is, of course, an important question and one that's often asked. But there is no single, standard answer. To a large extent, it very much depends on a number of factors, such as how restrictive, for example, the genre of language that you intend to study is.

So, for example, Shalom carried out a study of a corpus of personal adverts, the sort of thing you find, I believe, in newspapers. Personal adverts tend to be very short. And they're also quite repetitive. They

use a very restricted form of language. So because of this Shalom didn't need to collect a lot of data. Even with only about 20,000 words, she was able to start to identify the patterns of language used across different adverts.

Now some studies have used learner data, as I was talking about before. Student essays, for example, have been collected. And they tend to be quite small documents in terms of word count. For example, in a study of American and Polish students by Lenko-Szymanska, she collected Polish and American essays, essays written by Polish students writing in English, 79 of those, and 80 essays written by American students writing in English.

Again, the reason why she could get away without having very much data-- in fact, just short of 160 essays-- was because everyone answered the same essay question and she only had two sets of data, American and Polish essays. On the other hand, a corpus like the British National Corpus, which you'll be finding out much more about later on in this MOOC but which you've heard about already, aims to be representative of written and spoken language in lots of different genres and contexts. So a reference corpus such as that needs to be a lot larger. So that's why the BNC is about 100 million words in size.

With that said, as you know, the Brown family of corpora, which we've also mentioned so far, only have a million words each. And that seems to be large enough to carry out comparative work across different language varieties and, indeed, through time. With that said, the Brown family are only written texts and they don't cover all forms of writing. So even they are restricted.

As well as the genre or language type that we want to analyse, we also need to bear in mind what we want to do with the corpus. A general point is that if we're focusing on the rare linguistic feature, we need a lot of data. And if we're focusing on a relatively common linguistic feature, a frequent, then we can get away with having a much smaller corpus. You'll find lots of examples anyway.

For example, let's compare the frequency of two features in two different corpora. One corpus is the British National Corpus. The other is a very small set of data from a large corpus called the Longman Learner Corpus. In this case, it's essays from the intermediate learners living in Hong Kong, so just a small extracted data from the Longman Learner Corpus written by students who have an intermediate level of English living in Hong Kong and writing an essay in English. That corpus is only about 62,000 words in size, that corpus crafted from the Longman Learner Corpus.

OK. Let's look at the frequencies of two words, a common word and a rare word. The common word is "because." And the rare word is "hereof." In the table, you can see that "because" occurs more often in the small part of the learner corpus than "hereof" occurs in the large reference corpus.

So as long as you're focusing on relatively common language features, it's OK to use smaller amounts of data. You can provide sufficient examples to analyse. And of course, sometimes you have to settle with what you can get.

To give an example of this, a few years ago, I was involved in a research project. We wanted to collect a corpus of language used in languages in India, languages such as Hindi, for example. This was going to be a large corpus, around the same size as the BNC.

However, one issue that arose, which didn't rise when building the British National Corpus, was that at the time it was a lot more difficult to get hold of electronic texts in Hindi. We ended up using a smaller number of sources, mainly from news-based websites, because that was the data that was available to us. It was a sort of pragmatic sanction at that moment.

So we have to take into account, say, for example, the time and money we have available, as well as the data that's available. And sometimes, yes, we need to compromise. So although it's good to try to be as representative as possible, sometimes we also have to accept that there are real-world limitations on what we can actually do.

Now another aspect of design involves the size of the individual files in your corpus. One thing to think about is how to balance all of the contributions so that no single person or source contributes too much towards the corpus and is unfairly represented. For example, if you're building this corpus of learner English from a school that you might be working in, imagine that you have three classes in your school. And you've got the teachers to help collect essays from each of the classes.

Each class has about 30 students. However, in each class, students have produced 15 essays. Each student produced 15 essays. Whereas in Classes 2 2 and 3 the other classes, each student has only produced about five essays each. Do you take all of the essays? Or do you try to create a more balanced corpus by disregarding some of the essays that was produced in Class 1, where they were producing many more essays per student?

There are arguments for both options. On the one hand, we want a lot of data. On the other, we want to build a balanced corpus.

What I'd probably be tempted to do here is to take all of the essays, but devise some sort of annotation scheme so that each student has a reference number. Then all of the essays from Student 1 in Class 1 can be tagged with the same tag so that we know that those essays come from that student. And in that way, if we found that a particular word or phrase has been used a lot in different essays, we could check the dispersion across the essays and see if it was just because of one or two students from a class, rather than being representative of the whole corpus. So it gives us a way, if we see lots of examples of something, just to double check it isn't because of one source which has produced lots of data, and within that, this thing that appears frequent is only frequent in that usually large volume of data.

Another problem may involve the fact that the essays might be of different size. If you look at the last column of the table here, you can see that the students in Class 1 and Class 2 have written essays of roughly a similar length. However, the students in Class 3 have written longer essays. So if we include their essays, then again we risk spoiling that simple balance in the corpus.

One way around this is just to take samples of the data. So we might only want to take a 300-word sample from each of the Class 3 essays and throw away the rest of the data. This can be done sometimes in building reference corpora. So for example, with the BNC and Brown family of corpora that we've looked at, they tend to only include samples of text rather than whole texts. And certainly in the Brown family of corpora, that is with the express intention of getting roughly equally similar sized samples.

Now using samples is OK if we aren't bothered about analysing the overall structure of the essay or text, but we instead are interested more in the grammatical or lexical features of that text. However, if we're going to do detailed analyses of textual structures, it can be very difficult to sustain an argument to do that, i.e., take samples.

Another point to bear in mind if we do have to decide to use samples rather than whole texts is that we should try and take samples from different parts of the text. So for example, with Class 3, we have 32 students who wrote eight essays each. That's 256 essays.

One thing we could do is to take the first 300 words from each of these 256 essays. However, that's potentially going to lead to problems because we'd end up with a corpus that's heavily focused on the beginnings of essays. So we might find lots of cases of things like, "In this essay, I'm going to." Words do tend to associate themselves with beginnings, middles, and ends. And if you sample from beginnings, middles, or ends, you will begin to skew your data towards the features of beginnings, middles, or ends.

The same would apply if we only took the ends of essays, therefore. Then we'd probably find a high frequency of phrases like "in conclusion," "to conclude." So we don't want to do this, necessarily, in our data unless we are unusually interested in beginnings, middles, and ends.

What's the solution? Well, even before you start to collect your corpus data, it's worthwhile thinking a little bit about the issues of size and representativeness. As with a lot of research, it's often very useful to carry out a pilot study first. If you like, it's just the same in corpus building. Try piloting it. So start off with trying to gather a few files to see how easy they are to get a hold of and convert to electronic form.

What about the second factor to take into account, planning a storage system and keeping records? This involves deciding what the file system you're going to use will look like. Again, let's think about our corpus of student essays, this pretend corpus that we've collected. Imagine you've got 600 essays from different students. How would you store these articles on your computer? Well, one way to do it is simply put all the essays in one big file, one after another, a gigantic Word file, for example, and save it as something like essay.doc.

However, I'd be very careful about storing your corpus like that. It's often wiser to store each essay as a separate file by itself because that allows you to make comparisons between the essays in a much more easy fashion. We may also want to compare different age groups against each other. And it's harder to do that if you've just got one big lump of data that you're working with.

So instead, it might be better to store each text in a separate file. I'm storing my files as text-only documents, rather than documents, per se, because Word documents tend to insert lots of invisible things into text that make the files appear on the screen as they do. We just want plain text files, typically, for what we're doing in corpus linguistics.

As well as that, you could also use a file system which allows you to include similar types of files

together. So in the example on my handout that I will have on the website for you, I've created a filing system where articles are stored according to the age groups of the students. So you can see that I've got three folders for ages 8, 10, and 12.

Then in each folder, I've given each essay a label that has three numbers. The first number is the age of the student. The second number is the number of the student. And the third number is the essay number. So 12.1.1 means that it's the first essay written by a 12-year-old whose ID number is 1, while 12.2.5 means the fifth essay written by 12-year-old whose ID number is 2.