Another aspect of stage two involves keeping records of what you're doing. So it's useful to keep a list of things like where you got a text from. If it's from a website, then what was the address? What's the URL? You might also want to keep records of who created the text and when it was created, especially if that information isn't obvious from the text itself or from your storage system.

When it comes to our student essays, our notional student essays, you might want to record information like whether the essay was written in the classroom or as homework, for example. If some students did the same topic, you could make a note of that. You might also want to make a note of whether the student is male or female, or what their first language is. However, it's only worth noting these things if they are relevant to your research questions or if you think they might reasonably be relevant to somebody else who might use your data at a later date. There's no point in spending a lot of time recording information about gender, for example, if you're not interesting comparing male female differences, and you don't think anybody will be.

OK. Obtaining permissions. The first aspect of corpus building involves getting permissions. If you're collecting data from your own students, then it's important to get permission from them in order to do so. Additionally, of course, if the students are aged under 18, then it's also necessary to get permission from their parents or guardians, as they are not viewed as being capable of giving informed consent, at least in Britain.

Getting permission can sometimes be a long process. One way around it is to send out letters to parents and explain that the students' essays will be used for the purposes of your research only and that the student's name won't be used or revealed to anybody. If they don't want this to be done, they should contact you to have their essays removed from the corpus.

If you're only going to get the data for your own purposes, then I think that would be enough. However, if you want to share your data with others, then it would be much more sensible to get the parents to sign a copyright release form. If you're getting spoken data for a corpus of conversation, for example, then as with any other sort of research, you need to get permission. You can't just tape record people in secret. That's unethical. As I've just mentioned, you should also anonymise people's identities. So as well as copyright, you may have to consider issues of ethics here.

So we're going to spend a little bit longer talking about the fourth step, text capture and the types of text that you could get your hands on. So each type of text brings with it specific issues. As well as size, one of the biggest questions that corpus builders have is, where do you get your texts from? And there are a number of options available to us. One possible source is to word process text by hand. Type them in.

So, for example, you may have got students' handwritten or typed essays in paper form, and you need to convert them to electronic form by typing them into a word processor, like Word, for example, by hand. The advantage of this is that you have a lot of scope for creating corpora of interesting text types. However, the disadvantage is it takes a very long time, especially if you're doing it all by yourself, and it's also error prone. You might make mistakes.

So when I worked on a corpus of Hindi, we have paid transcribers to type out all of the conversations we had, and even with a large team of people, it took a very long time to do it, and it was extremely expensive. So you should only really consider doing this if you're building a small corpus, and you don't mind doing the typing, or you can get someone to help you, or if you're phenomenally rich, for example. You might consider doing this. But very few people are.

If you're building a spoken corpus, then you don't really have much choice, however. So that's something to bear in mind. You might want to consider using something like the spoken section of the BNC instead or another spoken corpus if you can find one, because producing transcriptions of spoken material is of necessity, time-consuming, and, hence, expensive.

Another possibility is that you could use a corpus of scripts, for example, taken from film, television, or theatre. However, here you need to bear in mind that this isn't really spoken language, but something which we usually call written to be spoken language. And there are often big differences between spoken language and written to be spoken language. So, for example, in the BNC, people say the word "yeah" about 8,000 times per million words. But if we look at scripted comedy programmes, we find the people say "yeah" about half as much. So scripted language tends to be cleaned up or simplified. It often loses a lot of the discourse markers, hesitations, repetitions, false starts, mistakes, et cetera, that are very common to spoken language and, indeed, are often the things that we want to study. That's not to say that we can't use a corpus of scripts, just that we need to be careful about claiming that it represents spoken language.

Now, if you're converting printed data to electronic form, then another possibility is to run it through a

scanner with optical character recognition software. For most people, this is probably a lot quicker than keying the data in by hand, although it's generally not 100% accurate or even close to that. The print quality of the document is likely to have an impact on the accuracy of the output, and the data will probably need to be spell checked and also hand corrected for errors, which, in the worst case, can actually be to a lengthier process than just typing it in by yourself.

In general, the best types of text that respond to scanning are those which are published in what I would call a straightforward format. If you have a text like a magazine which has lots of columns and different font styles, then the software's recognition accuracy may really go down. So scanning may only be a tiny bit faster than typing in some of these cases.

Another option is to try and get text that already exist in electronic form. And this is a method that more and more people are turning to in order to build their corpus, and it makes a great deal of sense. It cuts out many of these processing problems I've talked about. Now, there are a number of ways you could approach this. One is to simply ask people you know if they'd give you their texts. So, for example, you could build a corpus of academic essays by getting essays from your friends. Or you could buy texts on a CD-ROM or DVD or something. For example, quite a few newspapers sell archives of their stories, either online or in CD or DVD format. And there's also a lot of commercially ready-made corpora that you can buy.

Another option for essays is to use essays that can be obtained from the internet, from essay banks, for example, although here, you need to be careful about copyright issues because those are usually commercial services. Or you could use all or part of an existing corpus. For example, the BNC has lots of different genres and sub-genres of language in it. If you wanted to study a specific type of language, just letters or courtroom speech, religious writing or sports broadcasting, then you could simply tell, if you like, your corpus search software to just look for that type of text. And later on in the course, you'll come across a software package called BNC which allows you to do precisely that.

In similar ways, the Brown family of corpora are divided into 15 genres. So you could work with just one or more of those. Another option would be to use the internet as a potential source of data. For example, there are quite a few text archives which you can browse or carry out searches on in order to get data. I've put a number of websites links on the screen here to different places where you can get text-based data. These include text archives of novels, such as the Oxford Text Archive includes things

like that, databases of new articles, such as LexisNexis, and links to news group discussion links, such as Google Groups. But this is only a tiny sample of what's available, and you can probably-- in fact, you probably know about other sites yourself or can find out about them very easily just by digging around using a web search engine.

Now, once you've found a relevant internet site, then it's a relatively simple matter of downloading the text from it. One way of doing this is to save the web page as text files. For example, here's a screenshot of a BBC News website which might be a good source of data if you wanted to collect a corpus of news articles. Now, one problem with websites is that they're written in code, which gives instructions about all of the formatting instructions of the page, just as, say for example, a Word document does that for your computer in the package Word.

This is what the underlying code of a website like that looks like. So if you were to save this page onto your computer using the Save As function, you'd get this. Now, this can be useful to an extent, because the codes don't actually interfere too much with the data. You can set the codes, or you can see them all at the start, and then we'll brace this and less that, et cetera. And you can see the signs there which delineate little bits of what probably looked to you like gobbledy gook, and you can see the words in between. So you could actually just decide to use those pages as they stand.

However, if you want to just simply save text pages like that and not have the code, then you can save the file in a text-only format, and this gives you this version of it. Sometimes, internet text isn't encoded in text format but is instead represented in other formats such as JPEG and GIF files. When text is represented, in fact, in what is a photograph of a textual page, then it will either need to be keyed in by hand or scanned in, and there's no point in simply saving the graphics file.

So going back to the original news web page, which we have here, we can look at that and say, OK, the word "news" there, for example, is actually a graphics file. You can see it in big letters at the top of the screen. It's not text at all. It's a picture, a photograph, if you like, of the word "news" in big letters. So if you save as text, you won't get that word saved within your corpus.

So one problem with saving the entire page from a website address is that we may end up with unwanted text such as menus, titles, or links to other pages. It may be easier to strip them all, clean them from the individual files once they've been saved as text. Or it could be quicker to simply copy and paste relevant parts of the page. So go in and actually highlight parts of the page, copy it, and then

copy it into a Word file or some such.

Now, some sites do indeed have versions of pages which are text only, so you can find these, then those save you a lot of trouble in terms of having to remove a lot of the unwanted stuff that we saw a couple of screens ago. If there are lots of pages to collect from a website or archive and the site itself is structured in a relatively accessible format with not much unwanted text, then it may save time to utilise a website copier such as HTTrack. There's a link to that easily findable on the internet. This such software allows you to download a website or part of a website from the internet to your PC and can be a good, time-saving device, although it doesn't work, it must be said, with all websites. That's something to experiment with.

Thank you.