Welcome to this course on corpus linguistics. Now I assume that most of you know what a corpus is, as you've chosen to take the course. But let's just go through some type of basic definition, anyway.

What is a corpus? Is it just a collection of any old words? Well, yes and no. It's certainly a collection of words. We need some type of very large, typically, collection of words stored on a computer, so that we can rapidly and reliably search through that collection of words. A very simple definition of a corpus is a hell of a lot of words stored on a computer. But we'll refine that definition shortly.

Is it some type of theory of language, then, you might be wondering to yourself. No, would be my very clear answer. It is not a theory of language. It is a methodology for approaching the study of language. It will allow us to approach language and describe it better, test out hypotheses, etcetera. So to that extent there is an interface with theories of language. So if you have some theory about how language works, you might be able to use a corpus, go to the corpus and see whether this theory works a lot with your data.

But the actual corpus linguistic methods are methods which sit alongside, and indeed, as I think we'll be saying throughout this course, can often fruitfully be used in combination with other methods in linguistics. So why use a corpus? You might be saying, 'oh dear, I might have made a mistake here'. Lots and lots of words, computers, oh, didn't think of doing that. I'd actually rather sit around thinking about language, reflect upon my usage, talk to one or two people about those, and then hypothesise on that basis.

Well I'd say there's some space for that in linguistics, but by and large, I think we should be looking at language as it is. And looking at large amounts of data to tell us about things which typically we find very hard to think about, and certainly to express using intuition alone, or even chatting to a few people about the use of language.

We see over and over again in corpus linguistics that we're unable to make observations about language on the basis of looking at very large collections of data, which really don't appeal to our intuitions at all. They add to them. They show us things that we're doing routinely on a daily basis, which we find it very hard to imagine, if you like, that we're doing. But we undoubtedly do. We'll see some examples of that later in this lecture when we first look at 'collocation'.

Corpora can also reveal instances of very rare or exceptional cases that we wouldn't get from looking at single texts. Or even, perhaps, an introspection. You might just not imagine that that's possible, but when you see in the corpus you think, 'actually, yes it is'.

Also, importantly, if we wanted to take this approach without using a computer to studying language, it would be terrifically difficult. For the purposes of the types of searches that I'm talking about, humans are incredibly slow, and indeed can be inaccurate. They make mistakes. We all make mistakes.

So if say, for example, as we'll be looking at later in this session, you looked at the nearly 100 million-word BNC, and tried to sit down and read it and use a pen and paper in order to do your analyses, well, it's highly unlikely you'd finish looking at one research question before you met the grave. And even if you did, you would then encounter the bitter veil of tears that you would encounter when you discover that you probably made some errors while you were doing it. The computer allows us to do it swiftly and reliably, insofar as the computer is able to search through that data. More on that in a moment.

So what type of corpora can we have? What are our criteria in building a corpus? Well I always say, when you're building a corpus or choosing your corpus to use, you should think of the purpose of the use of that corpus. Your research question, or hypothesis, if you like, crucially determines whether that corpus is going to be useful to you, if you're selecting it off the shelf, or what the corpus should look like if you're going to be building it.

But typically, it must be a large body of text. If you're dealing with half a page of data, you may as well analyse it by hand and eye. So typically, it must be a large body of text. But large, for these purposes, might be 30,000 or 40,000 words. Or it could be billions of words, depending upon your needs.

It must, in some broad sense, be representative of language or a genre of that language. So what do I mean by that? Well, it's got to be representative of some type of language so that you can start to measure it up against the research questions you have.

So if you're interested in the writings of Jane Austen, well you'll say, 'I want something which sort of represents the writings of Jane Austen'. So when you go through a corpus and you discover it's Charles Dickens, you can say, no not that. Then when you go to another corpus and you discover that it's most of the novels of Jane Austen, you can say, 'oh yes that'. So that's what I'm talking about there.

It must be in machine-readable form. The computer has to be able to search through looking for the words and extracting them and doing nice things to them, as we'll see later on, in order to give us information about their usage. It can't just be a scanned image of the text or a photograph of the text. The computer must be able to analyse the individual words. So like text files on computers, for example, that you've probably generated yourselves.

It can act as some type of nominal standard reference about what's typical in a language if the corpus is built to be broadly representative of that language. And the example we'll look at later on is the British National Corpus, which had the aim of being broadly representative of British English.

But it's also often annotated with additional linguistic information. I use the word additional toere, I should have really said, with linguistic information that makes explicit, with the codes that make explicit linguistic information within the text, such as grammatical codes. So for example, if I have a noun in the text and I say, I'm going to label this as a noun, you might argue that that doesn't make it a noun. It simply makes it explicit that people normally process that as a noun. So that's what we're talking about with this type of annotation that we might add to the text.