I've just talked about 'annotation', adding or making explicit information in a text. What do I mean by that?

I'm now going to give you a short look at what we might call the markup of a text. Do not get the impression that you have to do this very often yourselves, or even deeply understand what I'm about to talk about. I'm simply doing it to illustrate to you, in part, how dumb computers are. How much information we have to actually pack in to a text in order to get the computer to be able to allow us to ask some meaningful questions of that text.

And the texting question, I'm going to refer to it here but it's on the slide, is just some straightforward text, which you will probably look at and say, that's really quite unexceptional. What's difficult about analysing that? OK, well if I was a computer, of course, I don't have all of the knowledge of language and of reading text and the conventions of setting up text that you have. So almost anything that we might want to say about that text has to be made explicit to the computer.

Here, and for those of you who are people involved in annotating text with XML and all of these advanced languages, you're probably going to dislike this because I'm using a rather old fashioned way of marking up texts, but remember that this is just for purposes of illustration. I'm using a little system here, and I will use the mouse to point this out-- to indicate, for example, that this here is a headline, or the head of the text. And you do that by putting in these things in angle brackets, saying that this is a 'head', this is a main 'head'. This here indicates this is where the 'head' finishes.

Also you'll be able to see a byline in a newspaper text, but the computer can't. So you have to say this thing here is a 'head', which is a byline. And it finishes here. Oh, and by the way, this is a p. And it ends there. And p, of course, stands for paragraph. Now there it might be useful because say, for example, if you're interested in the language of newspaper headlines, you might want to limit your search for the word 'arrest', for example, to all examples occurring within this type of area, where you've clearly delimited it as a headline.

Now all of this might look quite alien. But in fact, in reality, in many of the word processing programmes you use, this type of annotation, or something analogous to it, is hanging around in the background hidden away from you. And it's what allows the computer to put things on the screen in a way that you

understand, let's say, for example, bold, or underlined, or a paragraph break, etcetera . So this type of coding is hidden away all the time.

We can do more. We can start to say, well, actually we think this is a sentence. And that is a sentence. And that is a sentence. And mark it up using this little marker here. We could, of course, add to the end of it backslash 's' to show that's where we think the sentence finishes.

But just for speed, purposes of illustration, I'm saying this is a sentence. It's number one sentence. This is a sentence. It's number two sentence. This is a sentence. It's number three sentence. I should probably called these 's' units, because of course something like "by Daniel John" is more of a fragment. But just let's say sentence for the moment. Again, for the purpose of illustration.

I can even do things like change characters, such as the quote marks, to sequences which indicate a quote is starting and ending. Why am I going to want to use that? Well, quote marks can be ambiguous. Sometimes it stands for inch in English. So I might want to disambiguate between where it's acting as a quote mark and acting as a marker of measurement.

I can also go through and indicate where punctuation is, for example, if I want to. And I can also go in and put in a little marker to indicate where I think the different words are. Now at this stage you're probably thinking, why on earth would I want to indicate where different words are?

Let's say, for example, if you have a sequence like 'can't', you might decide that you view 'can't' as one word. Or you might want to break it up and say 'ca' is a word and 'n't' is a word. But that's your choice. You'd be able to make that choice and code it in the data. And then if you've got the computer to search for words, it would understand that for you, 'ca' and 'n't' are words.

So you can make the choice and then the computer follows it. But it has no real intuitions other than perhaps white space, a blank following the word, that that delimits a word.

But also, usefully, if you've noted the words, we can start to get really interesting information into the data, which can help in really linguistically motivated searches of that data. And that is adding in some type of codes which indicates, in this case, the parts of speech of the word. Now later on in the course we'll be looking at how to get that type of parts of speech information in automatically. So for today, again, as with the rest of this particular video, just look at this by way of illustration.

What's happening here is I'm saying this word is an NN1, which is a mnemonic standing for 'singular common noun'. So I go through and pop in a little code for each of these words to indicate what parts of speech I think it is. I say I, in reality, of course, this is typically done automatically by a computer, as we'll hear later in the course. But for now the key point is by adding in explicitly all of that information, I'm able to undertake quite sophisticated linguistically motivated queries of the data.

So rather than just say - I want to look for 'arrest' with a blank before and after it-, I could say, for example, I'm interested in a sentence with the world arrest in it where that sentence is the header of some newspaper article, and where the word 'arrest' is a singular common noun. And I can just find those examples of it.

I could also say, for example, I'm interested in what word follows 'collapse', in this case, within a sentence. So it would say in this case, well actually nothing follows within a sentence. Because a new sentence starts there, rather than if we didn't have that type of mark up saying, well 'collapse' is followed by 'by'. So you can use that what we call mark up to your advantage to improve the quality of your searches through the data.