OK. It's an enormous pleasure to introduce you to Professor Geoffrey Leech, professor emeritus in the Department of Linguistics and English Language here at Lancaster University, and a long-standing corpus linguist. When did you first encounter corpora, Geoff?

It was way back in the '60s, actually. At that time I was in University College London.

OK. So Bloomsbury.

Bloomsbury, actually. And that is where corpora-- can I use that word, "corpora"?

Yes. Plural for corpus.

Yes, the plural, corpora. OK. That was in the-- by the way, I'm not in favor of corpora. It's just it's caught on.

I know. It's caught on. So this sort of non-plural that's caught on and we can't escape from it now.

I'm afraid so. So anyway, in those days, no corpus existed in the modern sense at all. But my senior colleague-- I can't call him my supervisor because he wasn't-- was Randolph Quirk.

OK. That would be Lord Quirk?

Now he's Lord Quirk. That's right. And he's still going strong at the age of 93, I think it is. Anyway in those days, of course, he was a youngish, bright and bushy tailed professor of about 50 I think. And so one of his great projects that he started at UCL was to build a corpus of modern English.

Where did he get the idea from?

I don't really know. It was in the air because if you go back a few years before 1960s, that was the era of American structuralism. American linguists were very keen upon building up linguistic description from the base, as it were, from the empirical base. Get your data, and then you can work out your grammar. Which is the opposite of what happened later with Chomsky. He started with the grander structures, and then he worked down to the data. So this bottom-up philosophy, it was very fashionable in the 1940s and '50s.

Of course, it meant that any language you were going to study, you really had to build a corpus around. And for the American structuralists, that was often a matter of going out into the field to the Native American tribes and collecting data from them, and writing it all down, and then transcribing it. And so nobody, I think, before that time, before Quirk, had really thought of making a fairly representative corpus of the present-day English language.

And did he put it on computer immediately or was that a later innovation? Because a lot of the American structuralist work was sort of field notes, as I'd think of them.

That's right.

It's all paper-based.

That's exactly. And so Randolph Quirk was in that sort of pre-computer age, really. I suppose in 1960, when he began, there were rather primitive corpora around, they were-- sorry, I mean primitive computers around. But they were more or less restricted to--

[INTERPOSING VOICES]

The scientists and the mathematicians. Basically they were number-crunching machines in the way people regarded them in those days. So getting them to work with language was not an easy task.

However, to start with Randolph Quirk, he was very advanced in his technology with recording speech. He had a lot of recorded, impromptu, discussions of spoken conversations and do on and so on.

Was it reel-to-reel tape or something like that?

Yes it was reel-to-reel tape. And then of course he had some researchers who worked with him, people like Jan Svartvik and Sidney Greenbaum, who were busy listening and transcribing this spoken language. So in that way, he wasn't any kind of technophobe. But all this data was collected, and transcribed, and put into enormous metal cabinets which were written like that because they had to be fireproof.

You certainly don't want to burn your only paper copy.

Exactly. I mean your corpus was there, in the room, in big metal cabinets. And so it wasn't until the '70s,

really, that that corpus at UCL was computerised.

Right. And that was a case of typing it up again essentially. Sort of typing it into a computer or something.

I suppose it was. Yes. That was Jan Svartvik who did most of that. But anyway, yes. Of course, a lot of the data was written, and so we started to use that corpus in the late '60s when we wrote our first sort of major grammar of English, Quirk et al, number one. That's to say, *A Grammar of Contemporary English.* Yes, so at that stage already you could say the corpus-based grammar was getting off the ground.

And it was something. When I think of that period that these survey in numbers, they were a sort of stella nursery of linguists. Because quite a lot of important linguists actually worked out there. I mean, there's you, Jan Svartvik, Sidney Greenbaum.

Yeah.

I think David Crystal.

Yes. David Crystal was also in the survey at that stage.

I think Norman Fairclough.

Norman Fairclough, yes.

Not that a lot of people would associate Norman with that type of work nowadays. But he did start off there.

Well, he was a student. And he was one of my earliest students, actually. And it's through Norman that I eventually moved from London to Lancaster.

Great. Oh, tell us that story. How did you get to be here?

Norman Fairclough was a senior student, I would think I would call him. He was still working on his BA when I was there. And when he graduated, he got a job in this new-- brand new almost-- university in the far, distant north of England called Lancaster. And many of us hadn't of it, or if we'd heard of it, we got confused with Manchester. That was the general feeling down south.

But anyway, Norman, a pioneering experience, started working in the English department here at Lancaster, which was then obviously very new. And it hadn't really got much of a track record at all in research. And I so when I came up, partly I think it was through Norman informing me of the job, you know.

Dear Geoff, have you seen this-- type of thing?

And so I came up here. Yes, so we sat around a table, about five of us who were specialising in modern English. And thought, what should we do? Lancaster is a new university, and can we do anything to put it on the map? And so a suggestion I made at that stage was, why not imitate this newfangled Brown Corpus.

Which is the 1 million words from Brown University.

That's exactly so. American English.

All computerized.

Exactly, yes. And that was the first computerised corpus of English, actually. And so we got a copy of that corpus. We were in touch with the main compiler of that corpus.

Henry Francis?

Nelson Francis.

Nelson Francis. And Henry Kucera, got them mixed up there.

And so we thought we'd take it up. And Norman was very keen. And I was keen. And we didn't really realise what a dreadful time we were going to have over the next 10 years.

A big task.

The biggest task was actually the copyright problem we ran into. We were trying to negotiate copyright with the British publishers for about 500 texts. Eventually I sort gave up the ghost. I said, I can't do this anymore. We've run out of money. We've got rather primitive computing resources. We can't get permission from the publishers, or the agents, or the authors. Let's give it up.

But at that stage, an angel from on high appeared in the form of Stig Johansson.

From Oslo?

Yes. And he was actually originally he was a student of Jan Svartvik's in Lund, in Sweden. So he was already baptised into the corpus studies. And he happened to come to Lancaster on a Fulbright Fellowship or something for a year. And I loved that. He had the corpus bug, fortunately for us. He went back to Scandinavia and decided that he would take the LOB, the Lancaster corpus, as it was then, to make this 1 million-word word corpus which would be a matching corpus for the Brown Corpus in American. So we would then have two comparable corpora that we could--

Compare and contrast.

Compare and contrast. Yes. And Nelson Francis himself was actually instrumental in making us work towards that goal. Because when I said to him, we're thinking of doing a corpus. I sent him a letter. There were no emails or anything in those days. I sent him a letter saying, well, we're thinking of building a corpus like your Brown Corpus in British English. Do you have any recommendations or thoughts about it?

And so he said, well, for heaven's sake, make it as close as possible a match for what we've done for American English. And of course that was a very sensible piece of advice. Because it did make it possible, the whole story that I was talking about in my last interview about comparison of corpora across different regions, and different generations.

That's marvellous. So that gives us the L and the O of LOB because we have Lancaster and Oslo.

Yes.

And the B is for Bergen.

Bergen. That's right. Of course that's another Norwegian university. And when Stig Johansson got back to Norway, he was looking for somebody to collaborate with him. Got some quite generous grant from the Norwegian Scientific Foundation.

But he needed somebody with a computer expertise. And so he found it in someone called Knut Hofland in Bergen University. They have a kind of research center there on language. And so he was

able to tap into that and make a good collaboration between the two Norwegian universities. Which really made a big difference to the way the corpora had developed.

So Norwegian funding helped there. Did you get any funding at the British end when you started?

Yes. A little bit. But it wasn't enough, as it turned out. I mean, this is one of the reasons I decided to-- I wanted to give up in about 1976. The reason was-- I'm sorry. To begin with, we got a nice little grant from the publishers, Longman.

Right.

Yes. Because they had been in close liaison with Randolph Quirk, [INAUDIBLE], and I'd had books published by them and so forth. And so they were quite anxious to do the work. I think the amount of money was 3,200 pounds.

A princely sum in those days.

Sounds really like peanuts today. It was important, truly.

Yes, absolutely.

We soon used that money, and really we just kept going hand to mouth. But again, because of the help Stig got in Norway, we were able to complete the corpus. And from then, you know, things went up because we then got more money from British sources. I think it was the Social Science Research Council and bodies like that who got some money to do the annotation of the corpus.

Right. But that was the next major step, really. Moving on to doing graphical annotation, to LOB, parts of speech annotation.

That's right. yes.

Could you tell us the story of that? Because it's interesting in that that entailed the development of one of the earliest and still most accurate parts of speech taggers, CLAWS.

That's right. The Lancaster tagger, CLAWS, really began shortly after we completed the LOB corpus, that was the one that we had finished with the Norwegians. We though, well, what's the next stage? Obviously we want to do some sort grammatical analysis of this data. And, once again, fortunately, we

had our model in what they'd been doing at Brown University.

Right.

So some people in Brown under Nelson Francis and Henry Kucera had developed a very primitive grammatical tagger. I mean it was just an MA project.

Was that Greene and Rubin?

Yes, you remember them. Greene and Rubin.

The TAGGIT, I believe it was called.

They invented this new device called TAGGIT. And it's unbelievable now to think back to how things have developed since then. And so, they managed to do a tagging of their Brown Corpus. Each word, in other words, having a grammatical label, saying it's a noun, or an adjective, or what kind of pronoun it is, et cetera.

Yes, I think it was at a conference in Norway-- by the way, I should tell you we had set up this organisation called ICAME, International Computer Archive of Modern English. OK, that's what it was then. It changed its name slightly since. So that was the first organisation of any kind for corpus research, corpus development.

And so we're either the very first conference, or possibly the second conference. That we met there, and Nelson Francis and Henry Kucera came over from Brown University in the States. And they brought this wonderful mag tape. You know corpora, in those days, were stored on these enormous, heavy tapes, mag tapes. And very, very kindly and fortunately they lent us this corpus which had all the tags on it. And so we were then able to develop this method of using the corpus as a training corpus.

Yes.

I think nowadays, probably later called them training courses. In other words, it gave us all the statistical data we needed to develop our own tagger for British English.

Well, that's interesting you say "statistical," because I remember picking up your book, *The Computational Analysis of English,* in about 1987 or 1988. And up to that point, I really looked at logic

and things like that as a way of analysing language or writing programs to do it. I remember reading that book and thinking, this is amazing. They're using statistics to do it.

Yes.

Which nobody else seemed to be at the time. And I went back and read that book with enormous fascination. And thought it was a really strong idea and a great thing to do. And unlike a lot of the other NLP, or Natural Language Processing technology at the time, it actually worked.

Exactly.

[INTERPOSING VOICES]

So how did you come across the idea and decide to approach something like grammatical analysis by using numbers?

Yes. Well it was a matter of kind of trial and error, I suppose. The tagger, the first tagger of all, that's Greene and Rubin's TAGGIT, was about 70% success for that thing. And about 30% of the tags had to be disambiguated to use the technical term. Had to be resolved by hand.

And then, we had this little project. Roger Garside from computing, me from the linguistics, and there were two research associates who we could afford to employ on that project. One of them was called Ian Marshall. Enough And Ian Marshall was playing around with the data, and he said, well look, I've just discovered that if we do it this way, using this probabilistic algorithm, sort of stochastic methodology, we can get a much higher success rate. Up, way up in the 90s, something like 96%. And so that was a great breakthrough from our point of view.

Yeah.

And after, of course, that became almost a kind of recognized methodology.

Yeah. Continuing in most languages.

But there was also a kind of division between those who believed in the statistical methodology and those who believed in rule-driven methodology. And eventually I think people arrived at some sort of compromise. You know? For some things it's good to use rule-driven methods, for some purposes it's

good to use statistical methodology. The great trick is to, somehow, combine--

Two together.

A hybrid tagger. But that was really a hybrid tagger in the original form.

It was. Because there were rules, weren't there? We used to call it, I think inaccurately in some ways, we used to call it the idiom list and things like that. But when you looked through it, it was actually rules.

They were Exactly.

And then through the '80s, you worked quite a lot with IBM, as I recall.

That's right. Yes. When do you come on the scene?

Late '80s. We met up when I came up for a supervision with Jenny Thomas. And we were started to talk about this stuff because I'd just read that book. And I was terrifically interested in finding out more about it

And before that, you were an undergraduate.

I was an undergraduate, yeah. But we really didn't get exposed to that much then. We expose students to it now. But I remember there was a sort of a locked room in the department in which we knew things were going on, but we didn't quite know what. So it was really in the late '80s when I came.

You were working on the IMB then, weren't you?

Yeah.

I guess that's the stage where we really reached our zenith of our research efforts. Because by that time, it hadn't become routine all over the world. We were still, in a way, had our... we were still with our noses in front as it were. We felt we were still doing something exciting and new. And I think after that, in the '90s probably, we got overtaken by the immense efforts that the Americans put into corpus-based--

Tool development.

Tool development. But anyway, we kept going. We're still ahead in some ways.

Of course, we don't switch to applications often, and use of That's, in many ways, what this course is about.

Yes.

That we're focusing on the many uses of corpora. Because you're right, when the likes of New Mexico computing labs are interested, building taggers, they have the person power to do it. But, yes, that was an interesting period in the '80s where it was technically, in the sense of tool development.

And I must just tell you about our relationship with IBM and Fred Jelinek. He was a great guru in this area.

I remember meeting him. He was quite terrifying.

He was quite terrifying.

He was a nice chap for quite terrifying.

Oh yes. And he had this broad Czech accent, although he had been living in America for a generation. And he really thumbed his nose at the linguistic fraternity in those years, Chomsky. He wasn't at all interested. He thought they were idiots, really.

Yes.

Give me enough data, and I can produce a much better machine translation algorithm than you can with as many rules as you'd like to build.

Of course he did famously-- he said that every time he sacked a linguist from his team the accuracy of the system improved.

I believe he did.

Whether we approve of that or believe it, we won't say.

And of course they were using it to develop speech recognition.

Absolutely. Harpy I think was his system from that or something like that.

And to begin with, they could only reproduce speech in written form if you separated each row by a little gap and then went on to the next row. So it's a little bit of a slow process, dictating your letters to a computer.

Word by word.

It was the beginning of something very, very, very important.

I remember going across to TJ Watson to give a talk. I met Fred Jelinek and [INAUDIBLE] John Lafferty and people like that. It was a very exciting moment to go over there because they'd been doing such interesting work for such a long time. It was probably one of the most challenging audiences I ever talked to because they were very, very bright people.

And then into the '90s, of course, there was another major corpus development with the British National Corpus.

That's right. Yes. That began, I suppose, the very first beginnings of it were in the late '80s. I remember being in a conference-- what was it called? BARR(?) Conference or something? It was in Greece, in a very lovely area of Greece, Halkidiki.

And for some reason, this chap, a rather impressive-looking chap from Oxford, who button-holed me. His name was Simon Murison-Bowie. I remember that. And he was developing a proposal to build a British national corpus. This was in about '90. The project actually took off in '91.

But there was a lot of, obviously, a lot of work developing the project. Which was, again, it was a kind of special time where this was possible because the British government, like many governments in those days, were trying to develop electronic wizardry in all kinds of directions, and particularly linguistic applications. For the first time using computers in linguistic applications like machine translation, or speech recognition. You know this is really considered an important industrial development.

Automated content analysis.

Exactly, yeah. So all these things were getting underway, and the British government decided it could afford to put forward this programme which would involve collaborating between industry and universities in this kind of research. In building a big, national corpus, which is now, of course, referred

to as the BNC. And it took about four years to finish the job. And it involved collaboration between Oxford University, Lancaster University, we were doing the annotation side. And also

Longman?

Longman. Yes Longman, the publishers were responsible for the spoken part of the BNC. And of course, as we've mentioned originally, OUP, Oxford University Press were the lead collaborator on this project, the investigator. Yes, and so somehow we managed to get this very large project off the ground. And all kinds of difficulties which I don't need to elaborate on, but eventually we finished it.

It was a major undertaking.

Yes, but, of course, having been involved, I can still see the flaws and thought, if only we'd had more time to do a proper job, et cetera. But still it has, I think it's proved its worth over the years.

It's 10 million words of spontaneous speech orthographically transcribed. It's still, I think-- I think-- the best resource for studying the English as it is spoken. To this day. That's good.

When people said, well, we need a new BNC. But then somehow I don't think it will ever happen. It will be too much money for somebody to give to that project.

We'll see. Maybe some ingenious fellow or lady somewhere who will think of a way of doing it. It would be great if they did.

Well I think, you know, and nowadays we have the great advantage of all that text out there on the web, and even transcribed speech. And large corpora like Mark Davies' Corpus of Contemporary American English

Absolutely. COCA.

They're all online now. So, you know it's not beyond the wit of the human race to dream up another--

That's right.

BNC perhaps.

Keep our fingers crossed. Along that hopeful note, shall we conclude? I've really enjoyed that

conversation. It's been a wonderful walk down memory lane.

Right. Thank you.

And thank you for the conversation, Geoff. Greatly appreciated.

Great. Great. Thank you also turning out this great.