The Importance of Reference Corpora

Geoffrey Leech
Lancaster University

Thank you very much for your introduction and I want to thank also the organizers of this conference for very kindly inviting me. I do not speak Basque, I do not speak Spanish, and I have to excuse myself for speaking only in English today.

My topic is *The Importance of a Reference Corpus*. There are various ideas of a Reference Corpus. The one which is most popular is the one on the screen here, which is 'A Reference Corpus is designed to provide comprehensive information about the language'. Also this notion of 'reference' conveys other things. It has to be a general Corpus of wide coverage of the language, and hopefully it will be treated by its user community as some kind of 'standard' for the language. I don't mean standard in the normative sense, although in fact if a reference corpus is successful, it will become accepted and recognized as a source of a standard by the user community and even more widely by the language community. So in that sense it is a very important type of corpus.

Another way to describe a reference corpus is as a benchmark, or a yardstick, something that people can regard as a standard of comparison, to compare with some other variety of the language or some other language. You can compare some data with a reference corpus, and know that you are comparing it with the same thing as other people.

I wasn't quite sure what people thought a reference corpus was when I was invited to give this talk, so I looked it up on the web. And the first pages I found on the web were from the EAGLES initiative of the European Community. In 1996, John Sinclair, a British corpus linguist, and a colleague of his, provided a typology of corpora, a classification of different types of corpora, and I thought it might be interesting to look at that now. But first of all, this notion of reference corpus, I think, has been promoted particularly by the European Union. And recently, in the last maybe ten years or fifteen years, different teams in different language communities all around Europe, and of course beyond Europe in other parts of the world, have decided to establish, and develop, a reference corpus for their language. And this idea has been promoted, I think, particularly successfully here, in Spain. Yesterday I was extremely impressed by the presentations of those who were talking about the CREA, the corpus of Spanish, the CORGA, the corpus of Galician, and the CTILC, the corpus of Catalan. Really, I think, these movements of establishing and developing reference corpora have been advanced more successfully in Spain than anywhere else in the world that I know. So I feel I am perhaps, as we say in English, 'carrying coals to Newcastle', talking about something which is particularly well known in this country.

Referring to different types of corpora, we have first of all then the **reference corpus**, and Sinclair also distinguishes it from a **special corpus**, a corpus which is for the specialist type of the language, a specialist variety of the language. Perhaps it might be a dialect, or it might be teenage language, or children's language, or the language of some very special type of language, such as computer manuals. All these have been developed on various occasions and are clearly distinguished from a reference corpus, because of their limitation to a particular variety. But if one has a reference corpus, these special corpora can be compared with it, to provide a kind of standard of comparison.

And Sinclair also invented the term *monitor corpus*, which is a very important idea, and again I notice that this has really taken off here in Spain with the CREA corpus and other corpora. The idea is that it is not sufficient just to build the corpus and then finish, and then sit back. You really need to go on monitoring the language as it develops, as it changes. What new words are coming to use, what old words disappear, how the grammatical structures evolve, morphology evolves, and so on. So, a monitor corpus is a corpus which exists in time: every year, or maybe every month, or even every week, new material is added to the corpus to keep track of how the language is developing.

The next term is not quite such a good idea: an **opportunistic corpus**. All people that have worked in developing corpora know what it means to choose the material that is easiest to obtain, to take what you can get, and... So very-very large corpora, for English for example we have the Cobuild Corpus and the Bank of English corpus, although they can be described as reference corpora because they try to cover a wide range of the language, at the same time they accept different kinds of material which are comparatively easy to obtain. Newspapers on the whole are easy to obtain comparing with other varieties. Up to recently, people in the USA have used the *Wall Street Journal* for corpus studies, because that was easy to obtain – they behaved as if the whole of the English language could be found in the *Wall Street Journal*. Nowadays people are gathering corpus data from the Internet. And so, if you collect the corpus in this way, making use of what you can get rather easily, it is not a bad idea, but on the other hand it is not quite the same thing as trying to seek a very balanced and recent sample of the whole language, which is what we are aiming at with a reference corpus.

I just briefly mention the other types of corpus here, because they are also important for the developing field of corpus linguistics. Recently people have put a lot of effort into developing parallel corpora, which are corpora for translation, where you have one set of texts in language A, another set of texts in language B, which are translations of the texts in language A. And so you can develop corpora of many different languages being compared very precisely through these translations – **parallel corpora**, as they are called.

Use of **comparable corpora** is not quite the same thing, but again it involves comparing different types of data. 'Comparable corpora' means that you have more than one corpus designed on the same principles, using exactly the same principles of selection, the same principles of corpus design are employed, so that you can then compare these two varieties through the corpora.

When I first began corpus linguistics, which is now thirty two years ago, we already had, as Professor Rojo mentioned, the Brown Corpus, which had been built by Nelson Francis and his colleagues in America, and we wanted to build a corpus like that, but of British English. And Nelson Francis advised me to be very careful to make the selection exactly the same as the selection of texts in the Brown Corpus, so that it would become possible to compare those two varieties of language,

American English and British English, through these two corpora – although in those days, of course, it was very limited, one million words each, and it was written language only.

And more recently, some scholars in Freiburg, in Germany, have developed new versions, newer versions of those two corpora, the Brown and the LOB (Lancaster-Osle/Bergen) corpus, as the British one tends to be called. So these are called Frown and FLOB, because they have the Freiburg elements, and they are exactly the same in their design and principles of selection as the earlier corpora of Brown and LOB. And so we can make a kind of limited diachronic comparison to see how the language has changed during this thirty-year period, and again make another comparison between British and American English. So that is the useful idea of comparable corpus.

And now we can go on to mention other kinds of corpora. I think one type of corpus I should have put on the screen, but I forgot, that is a **diachronic corpus**. Of course here in Spain you have been developing such diachronic corpora. For English, my own language, there are many many different diachronic corpora which have been developed, amongst which is the Helsinki Corpus of English Texts.

All these ideas of corpora seem to be different and yet I would argue that a reference corpus probably combines most of them if it is well designed. and let's say a reference corpus can contain special corpora – it can be used for special varieties of the language, because it contains many different varieties of the language and they can be easily compared with one another; one can select particular varieties to study within the general framework of the reference corpus.

And also, although it is not the case with my own association with the British National Corpus, in the Spanish corpora which we discussed yesterday, and that is to say the CREA and CORGA, you have a diachronic element built in, and also of course in the Catalan corpus there is a diachronic element, because you subdivide the corpus into various periods, so that you can make that type of comparison.

Is a reference corpus an *opportunistic corpus*? Well, I think all corpus linguists that are honest have to agree that they have to be opportunistic in the way that they collect data. They cannot choose exactly the data that they would like to have, and so, inevitably reference corpora do contain an opportunistic element.

A reference corpus can also be designed to continue in time - it can become a monitor corpus. So reference corpora can be many things to many people.

I say here that to be truly valuable a reference corpus has to be accepted as a *de facto* standard for the language and users have to feel that it somehow represents the language. Maybe conceptually it's like a standard reference dictionary; in Spain there is the Academy Dictionary, for Basque Batua´s Academy Dictionary, and these become accepted as some kind of reference for all users, and particularly perhaps for scholars and teachers, but also for the community at large.

In my abstract, although I will repeat this, I argue that although it is impossible actually to define a representative corpus in our present state of knowledge, nevertheless the important point is that you try to make it as representative as you can, and so that it will be accepted by the community of speakers of the language, and particularly those studying the language, as some kind of standard reference point.

Who uses a reference corpus? Well, of course there are different types of users and here I give four different types of users: obviously scholars and researchers; lexicographers and grammarians and those who build reference books and materials of the language, for example frequency dictionaries; and also teachers and learners of the language would benefit from availability of the reference corpus. Even if learners do not use it themselves, those who are developing language materials will want to use the reference corpus, those who are developing teaching materials of various kinds.

Lastly I mention software engineers, which have been a very important group of users recently. For example it is necessary to develop different types of software for particular applications in information technology; we develop speech synthesis devices, speech recognition devices, maybe we have information retrieval software, software for automatic abstracting, for data mining people sometimes call it, and for machine- assisted translation. All these software developers of course need to have access to some large and authoritative body of data for them to work on.

And I would like to emphasise that users are not really a passive bunch of people. They are not people who just use the corpus and don't communicate with anybody else about their use. No. There tends to be a user community and the users can contribute to the corpus. Maybe after the main corpus has been completed, there are a lot of things which have to be done: the corpus has to be popularised, software has to be developed, user problems have to be sorted out, results of research have to disseminated..., and these can be fed back, so that the users of the corpus can benefit from developments that are made by other users. I would like to imagine members of the user community as communicating amongst themselves, and also feeding back into the corpus as an institution their own research and their own projects.

But one area is a big problem of course, and that is this notion of representativeness, which always crops up when people discuss corpora: how can a corpus somehow be taken to be representative of the language.

I said in my abstract that the problem with representativeness is that any corpus is a finite sample of an infinite population – I am not using the word 'population' here in its human sense, human population, but the population of linguistic events which take place in a particular language – I think you can say they are infinite. I am not a mathematician, so I cannot really define infinity, but the general idea I have is that we never reach the end of this set of linguistic events, there are always new ones coming along that we haven't seen yet, and here there are a couple of very interesting cases mentioned yesterday. First of all, Miriam Urkia talked about the universe of texts in Basque in the twentieth century, and she came up with this remarkably precise figure of, I think it was 769.000.000 words or something like that, and this is a very large population, but it is a finite population, so maybe this is a special case, where we have a finite population from which to sample. But of course this is not really looking at the living language, this is drawing a line - saying 'up to the end of the twentieth century there is this amount of data', so there are still more data coming along in the twenty-first century, which we have to keep track of and make use of. So even in this case, you might say there is an infinite population coming along year by year.

Another interesting case that was raised yesterday was Guillermo Rojo's case of CREA. He gave us one the slides in which he showed that however big you make the corpus, if you add another 50 million words, another 50 million words..., you always find that the number of *hapaxes* is the same proportion of the whole set of word types in the corpus. These are words which only occur once in the corpus, and they

are always the same proportion of the whole vocabulary. So, however big your corpus gets, according to this model, the vocabulary increases at the same rate – you will still find new words that you haven't already found in the previous corpus. So this is the idea. Maybe the language itself is infinite in the lexical sense of the number of new items which you will find in it.

How to make a reference corpus as representative as possible? I have already said that this is not an exact science, but at the same time we can make a rational attempt to make the corpus as representative as we can. First of all we decide on what is sometimes called the sampling frame, the universe or the population of linguistic events from which you are sampling. In the case of a reference corpus, this sampling frame is usually very broad indeed. Of course it may contain both the written language and the spoken language, it may cover a period of time – the BNC actually covered about thirty years in time, but it was not truly diachronic, but there is usually some limit I think to the extent of time from which one samples text – historically one only goes back so far and so on. Even there then there is some kind of limitation implied by a corpus of the current language. And also there may be other limitations. In the BNC we excluded children's language, in the sense that we sampled for only speakers over the age fifteen; but actually this turned out to be erroneous, because all kinds of children actually came into the corpus through speaking in the room where the recordings took place, so in fact there are children's utterances in the corpus.

And of course one might exclude other categories like foreign speakers of the language. Maybe one should try to sample only for native speakers. So these are the kinds of things one has to think about when considering a sampling frame.

But if we want to achieve the best results in terms of *representativeness* or comprehensive sampling of the language, we have to aim at the three factors I mention below: **diversity**, a full sampling of the varieties of the language, as wide a range as possible of the varieties of the language. **Balance**, a very difficult notion; I define it as follows: 'the subsamples or the subcorpora of different language varieties must in some sense be proportionate to their *importance* in the language' - importance is the difficult word there. And **size -** I mention size in brackets at the end, because although people mention the size of the corpus as if it's the most important feature to indicate its status as a reference corpus, I would argue that size must take a back seat, because diversity and balance are what we are truly aiming for. And if we sample for diversity and balance, inevitably we will get a rather large corpus – necessarily a a reference corpus will tend to be rather large. But of course, whether it is 20 million, 100 million, 1000 million..., this is not really part of the definition of a reference corpus. But if we have sampled for diversity and balance, then size is an important thing in the sense that *more data is always better*. As Bob Mercer used to say to me when we were working on this with IBM, *there's no data like more data*. So you can accept that or not, as you wish.

So why is balance needed? Well, because we have to include sufficient samples of all the variant forms of the language, so we have to cover as wide a range as possible, and some experience of the past shows that existing corpora of English in particular have not been particularly well balanced. Just one example: there is a famous corpus called the London Lund Corpus, the first electronic corpus of spoken English to be developed in the seventies and eighties. And this was in the days when people used very heavy tape-recorders to record, these reel-to-reel tape recorders to record speech, and the machines were very heavy. They didn't want to carry them about too much. So, actually they recorded most of this in University College London, a particular College, and they tended to record students and professors, and a lot of dialogues were about the University syllabus and things like

that. So in this sense it was not a very balanced corpus. But this is a kind of bias that can easily enter into a corpus, even though we want to make it as balanced as possible.

Now I am going to just focus a little bit on *diversity*. The set of factors here, which I have listed for diversity, have been adapted from an article, a very well known article by Douglas Biber on representativeness in a corpus. And he present these various parameters as the most important ones. I am not suggesting that this is necessarily the very best classification, but it just gives you an idea of the kinds parameters, the kinds of factors of language variety that we have to somehow take into account.

First of all, the **primary channel**. Is it spoken data or written data? And then we can go on to consider the **secondary channel**s like, is it scripted? that is to say, spoken language which is already written before it is performed? E-mail is an interesting category, because although it is written language, it is a very special kind of written language which contains many features of the spoken language, and so, differences of media, in this case an electronic medium, also cause differences in the language.

Format is also a secondary factor - whether materials for example in written data are published or not published - books and journals make different types of format... Ephemera, which occur in various corpora, are again, a rather miscellaneous group of documents which are neither books nor general publications – they are documents of short-lived interest like office documents, labels on bottles, public notices, invoices….

Then the **setting**. Is it an institutional setting? Is it a private setting? Is it public to certain degree? For example you could call this a public event that we are taking part in now, but is not so public as some others like a written newspaper or a TV broadcast.

Addressor. Who is the producer of the text or of the dialogue? What is their sex? Are they male or female? What is their age? Age grouping can be important for spoken language in particular. What is their social class? These are what you might call demographic categories, considered when sampling particularly for spoken language.

Now let me think about the receiver of the message, the **addressee** or the audience. The reading public perhaps in the case of a written published text. But also we have to consider this factor for speech. Is it a monologue? Is it one person speaking? Or is the channel a two-way channel? Is it a dialogue? Is it face-to-face or by telephone? And so on.

Factuality. In written materials of course we frequently make a distinction between factual and fictional material. British newspapers these days are moving from the factual towards the fictional variety.

Communicative function. Is the text persuasive? Is it an instructional text? Is it narrative? Is it exposition of ideas? These categories are not very well defined, but I think they are important to try to sample for.

And **domain** or **thematic categories**. I think there are various names which are used for these types of categories. What domain of human activity or human knowledge does this text belong to? Is it science? Is it leisure? Is it humour? Is it law? religion? And so on.

These are well understood, but the big problem is how to decide how much of material from each category. And this is where the notion of balance comes in. We should if possible employ some kind of random sampling method. It sounds a bit scary, random sampling, but I think a very simple idea really - that you tried to avoid bias in the way you select text. For example in fiction, you don't decide to select all the best literary writers, because you want also to have popular fiction and various types of detective fiction, which are often read by very large numbers of people. So, what we did in the case of the LOB corpus which I was working on – we got a big bibliography, the British National Bibliography for a particular year, and we randomly sampled, so that we could select exactly the texts which the random sampling had turned up. This made a lot of work, because then we had problems chasing down these texts, finding copies of them if we could. We had to spend a lot of time in the national library - the British Library, as it is called. So this type of sampling has a penalty, but one should like to do it if possible.

Perhaps it is useful here to make a distinction between **selection criteria** and classification criteria.

Selection criteria. We use various parameters of language variation as selection criteria, if we actually build them into the sampling of the corpus, build them into the design of the corpus – for example, such a proportion of fiction texts, and such a proportion of non-fiction.

But once we have collected the data, we may want to classify the texts in other ways which have not been closely monitored, not being selective criteria. We like to keep track of these other factors. One example I remember in the BNC was the category of ethnicity - in the UK, as you can imagine, there is a very ethnically mixed language community. And so, we didn't actually sample.... we didn't actually say we'll have X percent of this and Y percent of that ethnic group... We didn't actually build this into the selection process, but we classified the texts, including of course spoken transcriptions by such characteristics as these.

So the classification criteria are built into the header information which we get with the corpus, so that people can search and select for those factors, even though they were not actually part of the selection criteria.

Now, this difficult notion of **balance**. I am not sure whether what I have to say here is very scientific... I would say, to begin with, it is easier to agree that a corpus is an unbalanced one, than that it is balanced. As I said, people on the whole find it very difficult to achieve balance, and they maybe have to accept that a large quantity of texts are rarely available. Newspapers may form a very large proportion of your corpus, but then you might have to think 'oh, maybe newspapers are not that important'. Some corpora may have up to eighty per cent newspapers and you have to feel that it's surely too much, isn't it? Like you know, because here is a Basque newspaper, which Miriam showed me this morning, and I'm sure this is very important for the Basque language, and so with different languages you may finds that newspapers are more or less important – but rarely so important as to be more than 30 per cent of the corpus.

OK. How do we measure this elusive concept of balance? One measure is in terms of importance, but what is importance? Perhaps we should define importance in objective terms as the quantity of individual-to-individual linguistic communication that take place for this or that variety.

I see three factors here. The amount of linguistic material produced. You can measure how many column inches of newspapers are printed every day, if you

really try. It is a lot of work and you might find somebody has already calculated it for other purposes.

How many addressors? How many people are there who receive the message? Well, again, for newspapers you might be able to discover that from the circulation, but for text messaging and to conversation it is almost impossible. For example, for Internet usage. How many addressors are there for Internet? Maybe you have to study the number of hits.

Popular fiction. Yes, you can probably find out the number of copies sold.

But all this is extremely difficult to calculate and very often it's impossible to collect this kind of data.

So these three factors... You could think of adding them together, X plus Y, plus Z, and you get some idea of how important this text is the language community.

If you could find that out, but in practice, for a large proportion of the cases you can't discover it. You can say that certain varieties of the language are very different from others. For example, in conversation there is a very large amount of linguistic material, a very large number of addressors, and a very large number of addressees. But each piece of discourse is likely to have only one or two addressees. So in this sense, 10000 words of a conversation are less important than 10000 words of a newspaper text.

We tried to sample conversation in the British National Corpus by an interesting method. This was conducted by Longman, the publishers. They hired – what is it called - a market research firm who were used to doing demographic sampling of the population, and so this market research firm identified different people around the country, sampled them according to their age, according to sex, according to social class, and according to geographical origin. These are the kinds of factors which market research people always sample for. And so, we identified a number of informants –they were actually called respondents- and these people were given a walkman and they just walked around, or went about their everyday life, at home, at the office, and they still carried this walkman. And whenever they engaged in any kind of linguistic activity, they switched it on. And so in this way we got a very large corpus of in a way a very representative set of speakers of the language. But it was limited, because it was so difficult to collect and transcribe the data. This was the biggest task.

So, in practice, it was a very small sample of the everyday conversation of speakers of the language. I think it was about 150 respondents (who recorded over a thousand speakers) only. But at least there was an attempt there to be representative, using demographic criteria.

Of course is a very large amount of linguistic material is ordinary conversation. Some people estimate the 90% of all linguistic activity in any linguistic community is conversation, and it was really conversation that was obtained by this demographic method. There a very large number of addressors to sample from - everybody speaks in a conversation; there is not a restricted group of addressors like TV announcers or journalist reporters. So the numbers are very large for addressors and for addressees, and also a very large number of linguistic materials. So on this ground we would judge that conversation should be a very large component for balance in any corpus, but in practice of course it is very difficult to achieve this proportion. In the British National Corpus we ended up with just 10% of the corpus as spoken data and of that only 4% was this conversational data, simply because of

the expense and the trouble and the time of collecting that spoken data. But at the same time we have to say that we tried to achieve balance. At least we made an effort, and I think it is the attempt to achieve balance, which is all we can very often do when we are designing a corpus.

Right. So far we cannot measure the importance of a variety of language. I have showed the difficulty here. So we have to rely to some extent on subjective judgement. We have to make tactical decisions as well on what material is easy to obtain or to convert into electronic form. We may have all kinds of practical problems. One of the biggest is copyright, at least for the English speaking world. Publishers are very defensive of their rights and they will not allow people to use their data in a corpus (unless you pay a big fee, which people building corpora do not have). So you have to go down on your bended knees, or ask some very special favour, if you happen to know the director's nephew, or something like that. It is very difficult indeed to obtain copyright permission, and so, this is another tactical fact that we have to consider, how to choose a corpus for which very often you may have to reject half of your material because the copyright permission will not be granted.

In my abstract I suggest that there is another way of determining balance. Unfortunately I am short of time, so I will not be able to go into this. But the basic idea is that you achieve balance though analysis of an existing corpus. Once you have collected a roughly representative a corpus, then you can undertake analysis to discover which areas of the corpus are more homogeneous and less homogeneous - which elements of the corpus will have a bigger yield in terms of variation, which parts of the corpus have a bigger amount of variation than others, and you will want to take further samples in the areas where you do not have sufficient quantities to represent all linguistic variation. It is a rather difficult concept, but I can give an illustration, which is that in ...

Well, I think that probably it is better if I don't go into it, because it will take too much time...

But if you consider conversation, actually in spoken language, particularly in conversation, the amount of new vocabulary is relatively small the more that you sample, because people are very repetitive in the way that they use the lexicon in speech. But in writing, particularly in certain types of technical writing, you find data which is very rich in new vocabulary, in exploiting the lexicon of particular specialist areas. So you have rich vocabulary types of texts and less rich vocabulary types of texts.

By this argument, it is interesting that conversation is not so important, because people repeat the same kinds of expressions or the same kinds of vocabulary the whole time However many utterances there are, like 'thank you very much' or 'glad to meet you', you may meet thousands of those in conversation - but once you have a few of them –you have enough maybe. But in contrast, in other types of language, the more you sample, you will acquire a greater yield of new items, new vocabulary, new structures.

Well, I was going to talk a little bit about my own experience of the British National Corpus....

In my abstract I mistakenly say that this was the first reference corpus – at least, the first corpus of its kind, and therefore we could be excused at some of the failings of our attempt. It was a very difficult project to partake in, but I don't think it was the first, because yesterday Professor Rojo mentioned that the CREA corpus

began in the 1970's I think, and also in Sweden they had also been building a national corpus a way back in the seventies.

So what was special about our corpus is that we had a limited time to achieve this 100 million corpus. We promised 100 million and we had three years to do it, and so we had to work very very hard, and this is where we came upon problems ... Of course the mark-up, the use of SGML, a special mark-up, was extremely complex. We also undertook grammatical tagging, or morphosyntactic tagging of the texts, and so, the big problem really was that we soon ran out of a lot of resources. We ran out of time, we ran out of research people, we ran out of money. So this is a lesson perhaps for the future.

I was going to talk a little bit about the difficulties that we had, but I think, as time is getting short, I will just mention one or two things.

We made a smaller corpus called **the BNC Sampler Corpus**. It was like a mini-corpus sampled from the big corpus, consisting of one-million words of speech and one million words of writing. In this it is a better balanced corpus than the whole BNC! And I think this is an important point to make, that the big corpus is not just a big monolithic chunk of language. Of course it is very diverse, and also it can be used in very diverse ways. I have talked about the user community being able to make different uses of the corpus. There are so many uses of the corpus that you cannot possibly keep track of them. I think anybody who has built a corpus has the same experience that people use it for all kinds of purposes that you never imagined would be possible, as they thought of it after you had formed the corpus. This is a very common experience.

You see a corpus of this kind, a reference corpus, is a kind of resource for the whole language. Users themselves contribute their analyses and bring added value to the corpus through their own work. And this can include types of annotation. For the BNC, we already built in the morphosyntactic or part-of-speech tagging, as it is called, but this was not perfect by any means. We actually improved it and then reissued the corpus, because originally there was 3,5% error in the tagging, and then we lowered this level of error to 2%. So we were able to produce an improved version of the corpus around about 2000.

So there is that kind of annotation, but there are many different types of annotation. For example, syntactic annotation, semantic annotation, speech act annotation, anaphoric annaotation, discourse annotation, stylistic annotation, prosodic annotation, error annotation... One of the principles of corpus annotation is that the annotations should always be capable of being removed when you don't want it. So different types of analysis can produce different types of annotation, but people who don't want them don't have to use them. These do not have to be for the whole corpus, they can be of small sections of the corpus.

Well, I think it is time to conclude. So, what I have to say is this: there are many lessons for the future to be learned from the production of a large and varied corpus such as the BNC. There are also many pitfalls, many faults, many deficiencies which are bound to arise, and of course if we did the job again, God forbid, we would hope to correct some of those defects in the BNC. But in spite of all deficiencies and limitations, I would argue that the BNC was successful in coming to be regarded as a reference corpus, a kind of yardstick or benchmark or standard corpus which can be consulted by anyone interested in the developing state of the current language. The BNC has now been made widely available throughout the world. Hundreds of copies have been distributed on CD ROM, or people can consult and use the corpus on line from wherever they happen to be, and various kinds of sophisticated

software have been developed to enable people to retrieve information not only for lexical research but for morphological, syntactic, and other kinds of research. And so, people can use it to study all kinds of fields: pragmatics, sociolinguistics, discourse analysis, stylistics... You could not make a list of all the uses.

A corpus is a powerful resource for the future of the language. I would go further and say that a good reference corpus can become an institution just as an authoritative dictionary of the language becomes an institution. This does not mean that the corpus has to be perfect; far from it, people love to point out faults in a prestigious dictionary. Guillermo Rojo yesterday pointed out some mistake that they made in the Oxford English Dictionary, which is regarded as almost the standard reference dictionary of the English language, at least in my own country. Now, this pointing out the defects and faults of a big dictionary does not really damage their reputation, I think, because it is something which people actually enjoy doing. (It's a bit like making fun of the royal family.)

The BNC has its own user community, a collection of people who use the corpus, share knowledge about it and share interest in the research they carry out with this resource. And this develops so easily nowadays into an e-mail special-interest community.

So one of the spin-offs of the corpus is a development of the community of those who promote interest, understanding and research in the language.

Thank you very much.