A tribute to W. Nelson Francis and Henry Kučera

Introduction

Charles Meyer
University of Massachusetts at Boston

The 16th ICAME Conference was held at New College on the downtown campus of the University of Toronto from May 24–28, 1995, and was co-hosted by me and Ian Lancashire and Carol Percy of the University of Toronto (see the report elsewhere in this issue). Because this was the first ICAME conference to be held in North America, we thought that the conference would be a good occasion to honour Nelson Francis and Henry Kučera for the important contributions they have made to the field of corpus linguistics.

The papers that are included below represent some of the contributions to the special session honouring the two pioneers in computer corpus research. They demonstrate quite dramatically the profound influence that the Brown Corpus has had on the creation of new corpora and on the development of tools to analyze these corpora. Initially, the Brown Corpus spawned the creation of the Lancaster-Oslo/Bergen (LOB) Corpus, a corpus that permitted the comparison of written British and American English and that revealed the important information on language variation that can be obtained by studying corpora. But the influence of the Brown Corpus did not stop here. Since the Brown Corpus showed that it was possible to encode in computerized form a million words of written English, the next step was to do the same with a corpus of spoken English. As a consequence, the spoken texts maintained in printed form at the Survey of English Usage at University College London were computerized, resulting in the creation at Lund University of the London-Lund Corpus. With the compilation of the Brown, LOB, and London-Lund corpora, it was not long before we saw the creation of the first diachronic corpus of English, the Helsinki Corpus.

In addition to influencing the development of new corpora, the Brown Corpus set the standard for the analysis of corpora. The Brown Corpus was the first corpus to be lexically tagged, and the tagging routine developed to tag the corpus, TAGGIT, served as the starting-point for CLAWS, the program used to tag the LOB Corpus. Tagged corpora have greatly eased the task of analyzing corpora, and although numerous tagging programs now exist, all owe their existence and conceptual design to TAGGIT. And we all, as corpus linguists, owe a great debt of gratitude to Nelson Francis and Henry Kučera, whose efforts in the 1960s to create the Brown Corpus have made possible all of the work we are doing in 1996.

From Brown to LOB

Stig Johansson
University of Oslo

The Brown Corpus marks the beginning of the age of computer-aided corpus research. Thanks to the generosity of its compilers, it has been used by thousands of students all over the world as a source of data or as a means of exploring the ways in which computers can be employed in language research.

Not least, the work of the pioneers, W. Nelson Francis and Henry Kučera, has set an example for other corpus initiatives. This is seen most clearly in the efforts to compile corpora closely modelled on the Brown Corpus.

The LOB Corpus

The first of the Brown Corpus clones is the LOB Corpus. The initiative to compile a British English counterpart of the Brown Corpus was taken by Geoffrey Leech at the University of Lancaster. This is what he had to say about it at the first ICAME conference in Bergen in 1979:

> About seven or eight years ago I wrote to Nelson Francis, who at that time had already completed his Brown Corpus, and I said

> "Wouldn't it be a jolly good idea if somebody did a parallel corpus for British English?" ... I remember Nelson Francis was extremely friendly and helpful. He gave us all the information so that we could learn from his work, and the last thing he said to us was "Rather you than me. I wouldn't do it myself, but I send you my best wishes."

After a great deal of work had been done at Lancaster, the project was taken over and finished in Norway, through cooperation between the University of Oslo and the Norwegian Computing Centre for the Humanities at Bergen. This is how the corpus got its name: the *L*ancaster-*O*slo/*B*ergen Corpus.

Compiling the LOB Corpus was no easy task, in spite of the excellent example set by the Brown Corpus. One difficult problem, which had threatened to stop the whole project, was the copyright issue. This led indirectly to the beginning of the *I*nternational *C*omputer *A*rchive of *M*odern *E*nglish (ICAME).

ICAME

In February 1977, a small group of people met in Oslo to discuss the copyright issue as well as corpus work in general. Geoffrey Leech came from Lancaster with a suitcaseful of corpus texts. The other participants were: Nelson Francis, who was then guest professor at the University of Trondheim, Jan Svartvik, who was working on the London-Lund Corpus, Jostein Hauge, director of the Norwegian Computing Centre for the Humanities, Arthur O. Sandved, chairman of the English Department at the University of Oslo, and myself.

The outcome of the meeting was a document announcing the beginning of ICAME. I quote a passage from the text:

> The undersigned, meeting in Oslo in February 1977, have informally established the nucleus of an International Computer Archive of Modern English (ICAME). The primary purposes of the organization will be:
>
> (1) collecting and distributing information on English language material available for computer processing;
>
> (2) collecting and distributing information on linguistic research completed or in progress on the material;

(3) compiling an archive of corpuses to be located at the University of Bergen, from where copies of the material could be obtained at cost.

One of the main aims in establishing the organization is to make possible and encourage the coordination of research effort and avoid duplication of research.

The document announcing the establishment of ICAME was circulated to scholars active in the field, and it was used to support applications for permission to include texts in the LOB Corpus.

The first ICAME conference

After the LOB Corpus had been completed, the next task was to tag the corpus so that it could be used more efficiently for linguistic studies. A symposium on grammatical tagging was held in Bergen in March, 1979. There were some 30-40 participants, including: Jan Aarts, Sture Allén (the present Secretary of the Swedish Academy), Alvar Ellegård, Geoffrey Leech, Willem Meijs, Randolph Quirk, Jan Svartvik, and both Nelson Francis and Henry Kučera.

On the main program of the symposium there were a number of papers on grammatical tagging. Alvar Ellegård presented his detailed system of manual tagging used for parts of the Brown Corpus, Jan Aarts described the Nijmegen system, and Nelson Francis and Henry Kučera spoke about their automatic word class tagging system. I cannot resist quoting some remarks which Nelson Francis made in passing and which caused considerable merriment (most of the talks and discussion from the conference are preserved on audio tapes recorded by Knut Hofland):

I think you all are probably very familiar with the Fulton County Grand Jury. These are almost as famous first words as 'In the beginning was the word'. In that connection I had a startling experience the other day. I happened to be listening in a sort of desultory way to a news broadcast, and they were talking about the investigation – which is a very popular thing in the United States these days – of the peanut business of Billy Carter, the brother of the President, and they said at present this would not be brought before the Fulton County Grand Jury. I almost exploded, and I realized of course that Fulton County is the county where the city of Atlanta is, which is where the Carters come from, and

> a very appropriate legal organization to look into the affairs of the brother of the President would be the Fulton County Grand Jury.

The most tangible result of the symposium was the promise, extracted by Geoffrey Leech in exchange for a couple of bottles of wine, that the tagged Brown Corpus would be put at our disposal in our work on the tagging of the LOB Corpus.

Before I move on to this, let me just say that nobody knew at the time that the symposium in 1979 would be the start of a whole series of ICAME conferences. A second conference was held in Bergen in 1981. One of the participants was Magnus Ljung, who undertook to organize a conference in Stockholm the following year. This was the start of the regular ICAME conferences, which have been arranged annually since then.

CLAWS

The availability of the tagged Brown Corpus was of crucial importance for the tagging of the LOB Corpus, although this project opted for a probabilistic rather than a rule-based approach to tagging and disambiguation (an exciting new idea originating from Geoffrey Leech). The tagged Brown Corpus provided the first probabilities for tag combinations in the tagging suite which later came to be known as CLAWS (Constituent-Likelihood Automatic Word-Tagging System).

The rest of the story is well-known, I assume. Here I would just like to stress again the importance of the work of our two pioneers. To them we owe not only the Brown Corpus; they are the ones who gave the impetus to English computer corpus work. Their generosity in making the Brown Corpus freely available for research provided the model for ICAME.

When Nelson Francis and Henry Kučera started their corpus work, corpora were not the height of fashion. Now that “corpora are becoming mainstream”, to quote Jan Svartvik (1996: 3), it is important to remember that it is not wise just to follow the stream. We must recognize that restricting language study to corpora may be as questionable as ignoring corpora. Perhaps the best lesson we can learn from our pioneers is the value of having an independent mind – and the courage to go against the stream.

Reference

Svartvik, Jan. 1996. Corpora are becoming mainstream. In: J. Thomas and M. Short (eds.), *Using corpora for language research*. 3–26.

Grammatical annotation

Jan Aarts
University of Nijmegen

Every year I quote to my students an anecdote that Nelson Francis tells in one of his articles, and which I now want to quote to you:

> In 1962, when I was in the early stages of collecting the Brown Standard Corpus of American English, I met Professor Robert Lees at a linguistic conference. In response to his query about my current interests, I said that I had a grant from the U.S. Office of Education to compile a million-word corpus of present-day American English for computer use. He looked at me in amazement an asked, "Why in the world are you doing that?" I said something about finding out the true facts about English grammar. I have never forgotten his reply: "That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text."
>
> (Francis, 1979:110)

This little story is a vivid illustration of the clash between two linguistic methodologies around the middle of this century – the one empirical and based on observation, the other mentalistic and based on intuition and introspection. And because this was in the early sixties, when TG was the dominant, if not the only respectable linguistic paradigm, it also shows what Nelson Francis and Henry Kučera were up against when they launched the project that produced the Brown Corpus. To carry out such a project at that time required courage, conviction and a good deal of obstinacy, three qualities that are typical of the true pioneer.

But it was not only in the compilation of the first computer corpus ever that Francis and Kučera were pioneers; they also played a pioneering role in the field of corpus annotation. Their team were the first to develop a tagger that was suitable to tag a large corpus. TAGGIT was a rule-based tagger, operating with a tagset of 86 tags and consisting of more than 3000 rules. It was used to tag the Brown Corpus. In the wake of the Brown Corpus came its sister – or daughter – the LOB Corpus (younger in years of existence, but of the same internal age). The advent of the LOB Corpus was the second landmark in the history of English corpus linguistics; a British companion to the Brown Corpus, it was the second corpus which was to set a standard for English studies. However, its annotation was of even greater importance, because its impact was not limited to English linguistics. The CLAWS tagger developed for the annotation of LOB was a major step forward in the art of tagging in general, in that it introduced the use of a matrix of collocational probabilities as a means of disambiguation. Thus the statistical tagger was born. Together, TAGGIT and CLAWS have pretty much set the scene for the present state of the art in tagging. There are a great many taggers around today, but basically they belong either to the class of rule-based or that of statistical taggers. Perhaps the rule-based tagger never got a proper chance to prove what it could do, because you might say it was overtaken by the stochastic tagger. Stochastic taggers do indeed have two great advantages over rule-based taggers: 1. whereas rule-based taggers take a long time and a lot of work to make, stochastic taggers do not; 2. stochastic taggers can be (re)-trained for different domains and even for different languages. Stochastic taggers come in two kinds: they are either Markov-modelled taggers or Hidden Markov taggers. The drawback of a Markov-modelled tagger, as compared to the other type, is that it needs a training corpus tagged with the same tag-set for which the tagger is made; a Hidden Markov tagger, on the other hand, can be largely trained on an untagged corpus.

In tagging, then, we have come a long way since the Brown Corpus was tagged. If we want to tag a corpus, we can choose from several ways of doing it, each of which will be reasonably successful. But if we want to take annotation one step further and go on to parsing, the landscape looks a bit bleaker. As recently as 1993, Ezra Black even called the state of the art in parsing unrestricted English ‘deplorable’ (Black *et al.* 1993:2). That is a bit of an exaggeration, but it is true that progress in the area of parsing has been less than it was expected

to be some ten or even five years ago. Corpora have been, are being or are hopefully going to be analyzed by means of manual, rule-based or probabilistic parsing, or by hybrid parsers combining two or even all three of these methods. The use of neural networks is still in an experimental stage.

Nowadays, manual parsing usually involves some sort of computational support, but basically it is still the linguist who provides the analysis. It is significant and illustrative of the state of the art that the majority of the fully parsed corpora of English that are available at the moment, were produced largely manually.

In rule-based parsing, a full and linguistically respectable analysis still requires a considerable amount of pre- and postediting. In comparison with manual parsing, the human input is less, and, perhaps more importantly, is mainly limited to selection and correction. Manual parsing is linguist-oriented, while rule-based parsing is machine-oriented. It is probably still the most popular approach; at a workshop held at the University of Limerick in 1995, where ten parsing systems were compared, all systems but one were rule-based, while only one was probabilistic in nature.

The results achieved so far in probabilistic parsing with respect to full analysis of unrestricted language vary a great deal and are not yet impressive. The best account of what has been achieved so far is still to be found in Black *et al.*, *Statistically-driven computer grammars of English: The IBM/Lancaster approach*, 1993. On the basis of preliminary performance results, the book concludes that the approach is 'promising'. The great difference with the other two approaches is that such systems do not operate on the basis of linguistic knowledge, but on the basis of statistical knowledge, so that the analysis process itself does not provide any linguistic insights.

This has been a thumbnail sketch of the state of the art in corpus annotation. What is the role and the position of the ICAME community within this area of research? For it is clear that for quite a few years now it has not been the sole concern of corpus linguists anymore; we now find ourselves working within the much wider field of Natural Language Processing, which is not only populated by linguists (computational or otherwise) but also by computer scientists, machine translators and researchers in artificial intelligence. There can be little doubt, I think, about the leading role that ICAME played in the early days, both in its concern with the creation of corpora as large repositories of language data and in its pioneering work in the field of corpus annotation,

where the impetus of early taggers like TAGGIT and CLAWS has had a lasting effect. Today a community like ICAME probably constitutes a minority within the larger NLP community, but its role is a very special one; its first concern is not with practical applications, but with the furtherance of the study of the English language. Its objectives are therefore entirely linguistic in nature. For that reason a community like ours has a special task. Not only should we see to it that resources are created for the study of the English language, but we should also put a special emphasis on the linguistic nature of the tools that are used to create these resources. For it is only when these tools are based on a linguistic foundation that they can be used as vehicles for the hypotheses that we want to formulate about the English language and that we are able to use our corpora as test beds for these hypotheses. For, to return to Nelson Francis' article and quote him once again: what we want is "to find out the true facts about English grammar".

References

Black, E., R. Garside, G. Leech. 1993. *Statistically-driven computer grammars of English: The IBM/Lancaster approach*. Amsterdam: Rodopi.

Francis, W. Nelson. 1979. Problems of assembling and computerizing large corpora. In Bergenholtz and Schaeder (eds), *Empirische Textwissenschaft*, Königstein: Scriptor, 110-123.

Grammar and corpus linguistics

Gunnel Tottie
University of Zürich

I represent the corpus users, and I have been asked to speak about grammar and corpus linguistics.

I would like to start by telling you a story a friend once told me when I had complained to him about my computer applications becoming so rapidly obsolete. I will tell it in his version. I later found out that it was based on a science fiction story from 1944 by A. E. van Vogt, "Far Centaurus", and that it had changed a bit in my friend's memory.

The story is about an astronaut (that word didn't exist in 1944, but that is what the man was) who was about to leave the earth for the first voyage into outer space. The launching of the spaceship was a big event: the president was there, a large crowd of people, and a huge band playing patriotic tunes. After all the speeches, the performance of the band, and the cheers of the crowd, the spaceship was launched and went into orbit, and the solitary astronaut was projected into outer space. He stayed in orbit for several years, until he reached his destination, the star Centaurus, a never-visited, uninhabited celestial body, and the spaceship landed according to the carefully laid-out plans. When the astronaut got out of the spaceship, he was surprised, however; There were people there, on this supposedly uninhabited planet. There was a crowd to cheer him and greet him, there was a band to play heartening tunes to welcome him, and there was a president ready to give a speech. That wasn't the biggest surprise, however: The biggest surprise was that it was the same crowd, the same band, and the same president who had seen him off many years ago. They had started much later than the solitary astronaut, but they had arrived earlier.

This story struck me as an apt metaphor for many phenomena in modern technology, but as especially pertinent to computerized corpus linguistics, where things have been moving fast in recent years. I certainly often had the feeling that if had started some research projects a little later, with bigger corpora and better hardware, I might have reached my goals a little sooner, and I think many have felt the same way. Perhaps things have been moving faster in the area of grammar research simply because it started a little later than other branches of corpus linguistics. It is an area which I don't think the founding fathers of computerized corpus linguistics, W. Nelson Francis and Henry Kucera, had in mind when they started their pioneering work on the Brown Corpus and produced their ground-breaking works on word frequencies in American English.

As we all know, however, computerized corpora have proved extremely useful for all kinds of linguistic research, not least grammatical research. It is fairly straightforward to search computerized corpora for grammatical phenomena that can be tied in some way to a lexical item. A very good example is Ingegerd Bäcklund's 1984 dissertation *Conjunction-headed abbreviated clauses*, based entirely on examples from the Brown Corpus, with comparisons between abbreviated and non-abbreviated clauses headed by a closed set of conjunctions, *if, then, when*, etc., which obviously

provided ideal search objects. One need only glance at the ICAME bibliography to find large numbers of examples of excellent grammatical studies based on this principle.

The real crunch comes when you are looking for something that isn't even in the corpus, i.e. deleted or zero elements. How do you find a hole in the corpus by a lexical search? How do you find competing structures which will enable you to study variation between sentences like (1) with the complementizer *that*, and (2), where the complementizer is missing?

(1) I know *that* he left

(2) I know *Ø* he left

Rissanen (1991) solved the problem elegantly by reducing it to a lexical search for structures containing the most common verbs that tend to be followed by clausal objects and permit variation between surface and zero complementizers, such as *know* and *hope*. Essentially the same method was adopted by Biber and Finegan in their 1995 diachronic study of the same problem in the ARCHER Corpus. Although structures corresponding to (3)-(4) from a sample of present-day English would thus have been missed, it is probable that they represent only a small minority of the zero complementizer structures and that their omission does not seriously skew the results. And, in principle, it would of course be possible to expand the lexical searches to a large number of less common verbs which usually seem to be followed by *that*, such as *note, imply, posit, maintain, point out, acknowledge*, to find out if and when they take a zero complementizer.

(3) But in order to explain them, I *believe Ø we* do not have to resort to notions of alienation, male control... (Cameron 1992:187)

(4) De Beauvoir ... fell into the opposite error of *assuming Ø there* was no meaning outside of the rigid definitions people gave her. (Cameron 1992:190)

If it is thus sometimes possible to reduce searches for zeroes to lexical searches, this is not always the case. Tagged corpora can of course be of great help here but they do not automatically provide solutions. A particularly vexed problem is the one of finding zero relative markers in texts, as the surrounding elements are highly variable. Even a pattern

search using a tagged corpus is likely to yield very disappointing results, as I will demonstrate below, using made-up examples for ease of exposition.

An easy type to pick out from a tagged corpus is (5), where the subject of the relative clause is a definite noun phrase:

(5) He bought *the camera Ø the salesman* recommended.
DEF NP DEF NP

This kind of sequence is highly likely to contain a zero relative, but unfortunately it is also rare (cf. Tottie 1995:212)

With an indefinite NP as the subject of the relative clause, the computer is likely to go wrong, as can be shown by means of (6):

(6) He bought *the camera Ø salesmen* recommended.
DEF NP INDEF NP VERB

The sequence *camera salesmen* is more likely to be a compound noun than an antecedent plus subject of the following relative clause (cf. Tottie 1995:212). As has been noticed by many researchers, the favoured subject in relative clauses with zero relatives is a personal pronoun, as in (7):

(7) He bought *the camera I* recommended.
DEF NP PERS PRON VERB

However, as appears from (8), searches for sequences of definite NP plus personal pronoun can also net irrelevant examples:

(8) While he bought *the camera I bought some film*
DEF NP PERS PRON VERB NP

Unfortunately, sentences like (8), where the problem is obviously caused by the object NP following the verb of the relative clause, cannot be eliminated from a computer search, as this would mean excluding examples like (9), where the postverbal NP is an adverbial:

(9) He bought *the camera Ø I recommended last week.*

DEF NP PERS PRON VERB NP

Because of problems such as those sketched here, it is easy to understand that automatic parsers have not been successful in dealing with zero relatives – as far as I know, no entirely automatic parser has yet solved this problem.

The moral of the story is of course not that certain grammatical problems cannot be solved even by means of computerized corpus linguistics, but that thanks to ever-larger corpora and faster and better-equipped computers, talented linguists with enough computer skills (or vice versa) now have a real chance of getting at least very close to total recall and precision in finding zero elements and solving elusive linguistic problems.

We are grateful that Nelson Francis and Henry Kucera launched linguistics into computational space when they did.

References

Bäcklund, Ingegerd. 1984. *Conjunction-headed abbreviated clauses*. Studia Anglistica Upsaliensia 50. Stockholm: Almqvist & Wiksell.

Cameron, Deborah. 1992. *Feminism and linguistic theory*. Second ed. New York: St. Martin's Press.

Finegan, Edward and Douglas Biber. 1995. That and zero complementizers in late Modern English: Exploring Archer from 1650-1990. In: Bas Aarts and Charles F. Meyer (ed.). *The verb in contemporary English*. Cambridge: Cambridge University Press, 241–257.

Rissanen, Matti. 1991. On the history of *that*/zero as object clause links in English. In: Karin Aijmer and Bengt Altenberg (ed.). *English corpus linguistics*, 272–289.

Tottie, Gunnel. 1995. The man Ø I love: An analysis of factors favouring zero relatives in written British and American English. In: Gunnel Melchers and Beatrice Warren (eds.). *Studies in anglistics*. Stockholm Studies in English 85. Stockholm: Almqvist and Wiksell, 201–215.

van Vogt, A.E. 1973 (1994). Far Centaurus. In: Robert Silverberg (ed.). *Deep space*. Nashville, Camden, New York: Thomas Nelson Inc., 160–183.

Historical corpora

Matti Rissanen
University of Helsinki

Dear Nelson, dear Henry,

In his excellent plenary paper at the Nobel Symposium on Corpus Linguistics, in Stockholm in 1991, Charles Fillmore elaborated on the distinction between corpus linguists and armchair linguists. He gave due credit to both types of scholars, and duly emphasized that both groups need the work and ideas of the other. While I very much appreciate Fillmore's broadminded classification, I would like to point out that if all linguists were historical linguists, or at least had a fair idea of the aims and methods of historical linguistics, the division between corpus linguists and armchair linguists would never have appeared. The student of the history of the language simply lacks the tools an armchair linguist can rely on: native speaker intuition and personal knowledge (real or assumed) of what is grammatical and acceptable in the language form (s)he is studying. Historical linguists have to rely on a corpus, either in the old sense of the word, that is, a text or a selection of texts to provide empirical evidence, or in the new one, that is, a computerized version of the same. And if historical linguists try to practise armchair linguistics, their only way of doing so is to rely on somebody else's earlier corpus work, or on the more refined outcome of corpus work, such as dictionaries.

We need go no further back in time than the mid-1980s to find the beginnings of historical or diachronic corpus linguistics in the ICAME world. This was when my late colleague Ossi Ihalainen and I attended the ICAME Conference for the first time, at Windermere. At that memorable Lake District meeting, Ossi and I did our best to convince our colleagues that the M in the acronym ICAME need not only mean 'modern' in the sense of 'present-day', but that it could also mean 'modern' in the sense 'Shakespearean' – and why not even 'medieval'. I can never forget the generosity and cordiality of the founding fathers and mothers of ICAME, including Nelson and Henry, as they readily accepted a number of weird language historians in their group. Furthermore, they emphatically encouraged us to go on in our crazy plan to create a large (by the standards of the 1980s) corpus of one thousand years of English texts. I can remember Nelson asking me whether we had included his 1942 edition of the Middle English Vices and Virtues

in our corpus – we had. And I can remember talking with Henry about the possibility of adapting his spelling checker for the purpose of lemmatizing the word indexes of the Helsinki Corpus.

The Brown Corpus, the creation of Nelson and Henry, was a source of inspiration for our team when we started compiling the Helsinki Corpus. And at this conference, in this city, it is appropriate to emphasize the value of another source of inspiration, the Toronto Corpus of Old English. The Old English samples of the Helsinki Corpus are derived from the magnificent Toronto database.

When the Helsinki Corpus was completed, around 1990, I used to introduce it by saying that it was the biggest, the best and the most beautiful English long-time-span corpus because it was the only one so far in existence. Now, fortunately, the situation is much better, with many major corpus projects in full swing. The Archer Corpus, compiled by Doug Biber and Ed Finegan begins roughly where the Helsinki Corpus ends, so that the entire history of English, from the beginnings to the present day, will be covered. Louis Milic's Century of English Prose Corpus covers most of the little-explored 18th century. A chronological continuation of the Toronto Old English Corpus will be provided by ICAMET, the Middle English corpus in preparation at Innsbruck, under the leadership of Manfred Markus. We are looking forward to the completion of Sue Wright and Jonathan Hope's Cambridge-Leeds Corpus of the seventeenth and eighteenth centuries.

Of the corpora more focused by genre, David Denison's 19th century letter corpus is now completed. Terttu Nevalainen and Helena Raumolin-Brunberg's Corpus of Early English Correspondence (Helsinki), Josef Schmied's Lampeter Corpus of pamphlets (Chemnitz) and Udo Fries's Zen Corpus of newspapers (Zurich) are important and promising projects.

As to the regional historical corpora, Anneli Meurman-Solin's great achievement, the Corpus of Older Scots (Helsinki), is now in use. The American and Irish corpora by Merja Kytö (Uppsala and Helsinki) and Raymond Hickey (Essen) are in preparation.

One of the most acute problems in compiling historical corpora has been the question of whether they can be adequately tagged and parsed. Again, the Brown and LOB corpora have set us an example, but we have been painfully aware of the inadequacy of modern automatic or semiautomatic techniques when applied to the text material of the past. Thanks to the efforts of our colleagues both in the United States and Europe, however, even these problems are being solved. The Penn-Helsinki Parsed Corpus of Middle English Texts, compiled by Anthony Kroch

and Ann Taylor, can now be used by scholars all over the world, and an international team directed by Susan Pintzuk is equipping the Old English section of the Helsinki Corpus with glosses and grammatical and syntactic coding (the Brooklyn-Geneva-Amsterdam- Helsinki Parsed Corpus of Old English). Merja Kytö and Atro Voutilainen are solving the problems of tagging and parsing the Early Modern part of our corpus. Another international team working in this field is formed by Anneli Meurman-Solin (Helsinki) and Keith Williamson (Edinburgh).

What about the future of the compilation of diachronic corpora?

It seems that this is a particularly flourishing and dynamic field of corpus linguistics today. It is possible, however, that the time of multi-genre, multi-period general corpora of the type of the Helsinki Corpus will fairly soon be over, and that scholars will concentrate on corpora specified by genre or sociolinguistic parameters. We will no doubt have diachronic corpora of the language of science or law, of speech-based texts, of women's writings, etc. Perhaps one of the most interesting new developments would be the compilation of a series of regional historical corpora, to supplement the Scots, Irish and American ones now either completed or in preparation. What about a diachronic ICE, an International Corpus of the History of English?

Dear Nelson, dear Henry, in compiling a corpus, just as in climbing Mount Everest, the real achievement is to be the first. When it has been done once, it is easy to do it again. Without your pioneering work, we certainly would not be where we are now in compiling, using and developing historical corpora and in giving new life and new impetus to the study of the history of the English language. Thank you!