An interview with Hilde Hasselgård

University of Oslo, Norway
13:30-15:15, 27 June, 2018

（**XLX:** Xiuling Xu; **HH:** Hilde Hasselgård）

XLX: When did you initially know about corpora and corpus linguistics?

HH: I had Stig Johansson as my teacher already when I was a master student in the late 1980s. So I knew about corpora at that stage and learned how to search in a corpus. But I didn't use corpora for my MA thesis although I did use methods from corpus linguistics only because it was manual. I didn't own a computer when I was an MA student, and we had only two computers that MA students could use and of course no Internet so it was quite difficult to use corpora in practice. But when I started my PhD, I started using corpora.

XLX: Who has the greatest impact on you regarding corpus linguistics? And in what ways?

HH: That is definitely Stig Johansson, who was my teacher and supervisor both for my MA and PhD, and later he was a colleague and a friend. He was one of the pioneers in English corpus linguistics overall, so having him close by was very important. He took me to my first ICAME conference in my first year as a PhD student in 1990 when the ICAME conference was still very small. So he also introduced me to corpus linguists.

XLX: What do you think are the most fascinating aspects of corpus linguistics?

HH: I think it's great to have authentic data that we can access very easily.

For people who study a language that isn't their first language, it's a great advantage to investigate authentic language instead of relying on intuition because people generally have poor intuitions. Sometimes you find really surprising things that you wouldn't have thought of yourself. When you've carried out a corpus linguistics investigation, it should be replicable if you have specified how you searched and how you analysed and sorted the material. So it is good scientific practice as well. We can also use corpora in teaching. When corpus linguistics methods are used in a simple manner, students can do the same kind of things as we do as researchers. I think that's an excellent thing about corpus linguistics. You can do it at several levels of expertise.

XLX: How do you see the synthesis of linguistic theory and corpus linguistics?

HH: I do not think that corpus linguistics is a theory and I suppose most people do not think that. In corpus linguistics, you need a linguistic theory in order to interpret your findings and perhaps even to define your research questions. I think several theories go very well with corpus linguistics, but they would tend to be functionally oriented theories that aim to account for language use rather than theories that aim to model how language is produced. I think that you do need linguistic theory in order to make sense of corpus linguistics and particularly now that the technical bits are very easy compared to when I got to corpus linguistics. Often people spent much time and energy on the technical aspects and were not so interested in theory. They were so fascinated with the fact that they could look through a million words in a few hours instead of collecting examples for years.

XLX: Are there contrastive linguistic issues that can be significantly better addressed with the assistance of corpus linguistics?

HH: What contrastive linguistics gained from the multilingual corpora was a much better empirical basis for carrying out what both Stig Johansson and Carl James refer to as systematic comparison of two languages. Previously I suppose people had compared grammars, and to some extent vocabularies, but it was much more difficult to know what to compare, particularly beyond

rather trivial topics such as how you form a question in English compared with Norwegian, Spanish or Chinese. You can do much more sophisticated things with a corpus. You can study translation correspondences, which translation scholars obviously had done, but they had been looking at one text at a time because that's what you can do without a corpus.

Another thing is that parallel corpora help to define a *tertium comparationis* (TC) for contrastive analysis. TC is about a background of sameness against which two linguistic items can be compared, and this is something that is built into a bidirectional corpus model such as the English-Norwegian Parallel Corpus. This corpus has originals in both English and Norwegian and translations into the other language. We know that translators try their best to preserve the meaning and function of the original, so we can regard the translation relation as a TC. You can check translation correspondences in both directions and be fairly certain that you are comparing like with like. That is a huge advantage.

XLX: Right. Before the use of corpora, it was very difficult to define TC precisely. People had to rely on intuition.

HH: Obviously people could rely on dictionary equivalents if they were comparing vocabularies. What else? Grammatical functions I suppose. For example, you could decide to study modality by comparing the modal auxiliaries in two languages. But with a bidirectional translation corpus, you can search for a modal auxiliary, for example *should,* to see how it is reflected in the other language. The translations may be not just modal auxiliaries but also adverbs, adjectives and nouns, so you get a much broader picture of modal meaning and how it is expressed across languages.

XLX: To what extent can corpus linguistics enhance our understanding of some theoretical linguistic issues?

HH: I think the example of modality is a very good illustration, because it is a semantic field that contains much more than the modal auxiliaries. I'm sure there are other fields like that and where languages can have or prefer

different resources to express the same kind of meaning. You get that even with languages that are typologically rather similar, like the Germanic languages. But I imagine that if you want to look at time reference for example, and you compare English with Chinese, which doesn't have tense, you will find a spread of devices in Chinese that you can use instead of verb morphology to express present time reference. Sometimes there will be no marker because people just know that you are talking about the present time. The main thing is that the English tense system would correspond to something else in Chinese, and the parallel corpus can help you see what that is.

XLX: Do you think corpus linguistics has come of age as a discipline of its own?

HH: That's an interesting question, because in one way it has and in another way I think it's still developing. Maybe people felt it was more of a field when the discipline was young. Then people started to think that they study linguistics by means of corpora. So they came to regard corpus linguistics more as a method for doing linguistics. From what I've seen at recent ICAME conferences, where you can kind of monitor the developments, I think it's still a field but it's getting diversified. We now have subfields, such as corpus pragmatics or learner corpus research. The field diversifies and it also kind of sneaks into other disciplines. I mean there used to be a water-tight division between what we call theoretical linguistics and corpus linguistics, but now theoretical linguists use corpora to get material and corpus linguists use theory to interpret their material.

XLX: Can you comment on the corpus research at the University of Oslo from an insider's perspective? e.g. What are the main strands and characteristics of corpus research at the University of Oslo? What are the key contributions of corpus research at the University of Oslo to linguistics research in general?

HH: Our university has a Text Laboratory (in the Department of Linguistics and Scandinavian Studies) where they have developed corpora mainly of Norwegian but also other languages, and I'll come back to their work. It's

in the English Department that corpus linguistics has the longest history at the University of Oslo. That was again because of Stig Johansson who introduced corpora already at the end of 1970s. I think at that time only he used them. Stig's pioneering role started when he spent a sabbatical in Lancaster and met Geoffrey Leech, who was struggling to compile the corpus that was to become LOB (LOB is short for Lancaster, Oslo and Bergen). Stig offered to help with the project that seemed to be stopping before it was finished. Knut Hofland in Bergen agreed to do the technical stuff, and together they managed to complete the LOB Corpus. As you know, the LOB Corpus was only the second computerized corpus of English after Brown. These were the corpora that people used for rather a long time. Obviously they were extremely important in establishing corpus methodology. The second major contribution was also Stig Johansson's work: it was the multilingual development with the English-Norwegian Parallel Corpus. The idea was first presented at an ICAME conference in 1993. Already in the mid-1990s we had materials to work with and we had our first symposium in 1994 in Lund with colleagues from Sweden, Denmark and Finland. But the ENPC was the first parallel corpus that was made public and available to a wider circle of people than those who were directly involved in the project. It contributed a lot to parallel corpus methodology too. There have been many other corpus projects that have been based on the model of the ENPC. The first were the sister projects in Sweden and Finland. The Finnish corpus was I think only partly completed, and the Danes decided not to join, unfortunately. There are also similar corpus projects elsewhere with other language pairs, which relate directly to what was started here in the 1990s.

XLX: How do you view your own role in corpus research at the University of Oslo and your contribution to corpus linguistics beyond UiO and Norway?

HH: Within the English Department, I suppose that both Signe Ebeling and myself see ourselves as Stig's children in our teaching and research. We try to continue the tradition. Obviously the field has moved forward, so we can't do the same things as Stig did. We have to do something new. But we had a good starting point because Stig introduced us to his network of

corpus linguists. I think when we work with English in Norway or another country where English is a widely taught and very dominant second language, contrastive linguistics is a natural field to work within. Another is English as a foreign or second language. And this is something that we can do well in a non-English speaking country. One international role is that I have been on the board of ICAME and am now vice president in the Learner Corpus Association. Also, when you know people in the field, you are asked to be on committees and to review projects, articles, and books for example. So I guess that is another role: participating in the international community of corpus linguistics. I have published internationally, so there are people outside Norway who have read about my research. I have also edited a number of publications, including *Languages in Contrast*, which is an important journal for contrastive linguistics, and we get many corpus-based submissions.

XLX: You've talked about the history of corpus research at the University of Oslo. How do you see the future directions of corpus research at the University of Oslo?

HH: I hope it will continue. Of course we give courses on corpus linguistics as an obligatory part of our MA program in English linguistics because we believe it's a useful thing for people to know. Hopefully there will be staff here after my time as well who can continue the tradition. We also have colleagues who work with historical corpora and others who work with languages other than English. Our future directions will depend on what personnel we have and their research interests. But it would be great to continue with more contrastive research, with corpora that are better in the sense of newer and more comprehensive. There are very natural limitations because not everything is translated. Particularly from Norwegian into English you don't get a lot of published translation. We have been looking at ways of using comparable corpora in a more productive way. The main problem is the *tertium comparationis* I was talking about earlier, but maybe if we use comparable and translation corpora in combination with each other, then we might find a reliable method of comparing data from comparable corpora as well. I think that's an exciting avenue to pursue. Another thing that we could

study is the way that English is spreading in Norwegian society, outside of the educational institutions, more or less as a second language rather than a foreign language. I think it would be very interesting to look at English in workplaces for example. It will be different from both learner English and English as a lingua franca in the sense that the users of English won't be in a learning situation, and it will be within predominantly Norwegian contexts. It would be possible and very interesting to try to capture some of this development in corpora.

I mentioned in connection with a previous question that there is a Text Laboratory at the University of Oslo, where people work mainly with Norwegian and other Scandinavian languages. An interesting thing that they've done is to collect spoken data from various kinds of linguistic communities. So they have developed corpora of spoken language that are very well suited for studying variation, such as dialect corpora. They have also compiled corpora for lexicographical purposes. Hopefully they will continue to develop ways of preparing and displaying spoken material for corpus use.

XLX: What are the main strands of Scandinavian Corpus Linguistics?

HH: I'll focus on English corpus linguistics in the Scandinavian countries Norway, Sweden, Finland and Denmark. Sweden has been very important, with their very early pioneer Jan Svartvik, who was one of the authors of the Quirk *et al*. grammars. He got interested in corpus linguistics at University College London where he and Quirk developed the London Lund Corpus, which was a pioneering step in spoken corpus research. The collection and much of the annotation took place in London but the computerization I think was mainly in Lund. The Swedes were very early to do research on spoken English and they had a research project about differences between speech and writing in the 1980s. The Finns have done tremendous work in historical corpus linguistics. Matti Rissanen was behind the first computerized historical corpus of English, the Helsinki Corpus of English Texts. That was pioneering work. His team had to find ways of representing Old and Middle English in computer readable format and also ways of standardizing texts so they would be searchable. Later, the Helsinki group has developed a lot

of different historical corpora, for instance specialized corpora of letters, medical texts and so on. Unfortunately, Denmark doesn't seem to have a strong tradition of English corpus linguistics, but there used to be people who were interested in corpora of English for specific purposes. To simplify, maybe the contributions from Scandinavia are specialized branches of corpus linguistics, so spoken English in Lund, multilingual corpora here and historical corpora in Finland.

XLX: Could you refer me to some of the pioneering and leading Scandinavian corpus linguists in these branches of linguistics?

HH: In terms of the spoken language, I already mentioned Jan Svartvik. There are also some other people from Sweden, for example Bengt Altenberg, Karin Aijmer, Gunnel Tottie and Anna-Brita Stenström. All these people have now retired, but Karin Aijmer is still very active. She has also done work in the new field of corpus pragmatics, and I would say she has contributed to developing new methods and using corpora in new ways. The Norwegian corpus linguist Gisle Andersen has also done important work in corpus pragmatics.

XLX: How about the leading corpus linguists in contrastive linguistics?

HH: We can't get around Stig Johansson. Jarle Ebeling and Signe Ebeling worked with him on technical and practical matters. They've been much more hands-on with the corpus development than me. I wasn't directly involved in the work, although I was in the department at that time and was part of the ENPC research group in a way. But of course I have had the advantage of using the corpus in my research. I should also mention Cathrine Fabricius-Hansen, who has worked with the Oslo Multilingual Corpus, mainly with a focus on German.

In Sweden, Bengt Altenberg and Karin Aijmer have done very important contrastive work, and they have also studied learner language. Jan Svartvik was never part of the parallel corpus project, but has been influential for English Corpus Linguistics overall, and he had a research group in the 1970s

that worked with learner language analysis and contrastive topics. They used a lot of empirical material, but it was not computerized. I think this early empirical work, without corpora, was part of the reason why it was so easy to get people in Lund interested in contrastive corpus linguistics, because they were already interested in similar research questions. And they saw that a parallel corpus was an excellent way of studying language contrasts. A fun-fact about the project that produced both the English-Norwegian and the English-Swedish Parallel Corpus is that Stig Johansson, Bengt Altenberg and Karin Aijmer were old friends from Lund, and that was a very good starting point for successful cooperation.

XLX: Could you name some historical corpus linguists? Is it mainly because of Matti Rissanen that Finland has a strong tradition of corpus-based historical linguistics?

HH: Matti Rissanen is obviously an important name. He developed the Helsinki Corpus and managed to get a large research group together. Many of these people are also important historical linguists, such as Terttu Nevalainen and Irma Taavitsainen. Merja Kytö was Matti's research assistant on the Helsinki Corpus and she later moved to Uppsala in Sweden. We also have historical corpus linguists in Oslo, for example Kristin Bech has had two projects on Old English, and Gjertrud Stenbrenden has done important research in historical phonology based on corpora compiled in Edinburgh.

XLX: Could you share some stories with regard to the interaction of Scandinavian Corpus Linguistics with corpus linguistics in other regions, for example, with Granger's team in Belgium, and how you came into cooperation with them?

HH: Before Karin Aijmer retired, she organized several workshops in Gothenburg on both cross-linguistic and learner corpus research. That was one forum of contact and cooperation. We also cooperate with the University of Louvain. Apart from ICAME conferences, Sylviane Granger took part in the first symposium about parallel corpus research in Lund in 1994. She had started

compiling the ICLE Corpus then and was developing a method for studying learner corpora, which was also in a way contrastive. Now we think of cross-linguistic and learner language research as two different but related fields, but they used to be much more closely connected in people's minds. I have met up with people from Louvain from time to time because we have a lot of shared research interests in both contrastive linguistics and learner corpus research. In Louvain they have a parallel corpus similar to our ENPC, and on their initiative, or invitation, we now have Norwegian components of the learner corpora ICLE and VESPA. Of course once you get interested in learner corpus research and someone from Louvain is interested in cooperation with you, it's a big bonus, because the learner corpus research obviously comes from Louvain-la-Neuve and other people have followed, like the bidirectional parallel corpus comes from here and other people have followed. So I suppose we've always been interested in each other's work and that we can learn from each other.

XLX: What do you consider are some of the essential qualifications for a corpus linguist? e.g. Should a corpus linguist be able to do programming? How important is statistics for a corpus linguist?

HH: I think it's becoming more and more demanding to be a corpus linguist, because the number of things you need to know keeps being added to. I think the most important thing is still the linguistic part. As a corpus linguist who works with a foreign language, you have to be fluent in that language, because otherwise you can't tell what you see, you can't interpret the data. You need to know linguistics, so you know how to categorise things. In addition to those basic skills, you need to know how to handle the corpus and how to construct a search string that will get you the data that you need and not overlook a lot of relevant data. And you need to know corpus properly. For example, you need to know what's in your corpus. There are a surprising number of people who use a corpus but they do not know what they are looking at, whether it's a corpus of fiction, or whether the texts come from the 1930s or the 2000s. As seasoned corpus linguists we know that makes a difference. Diachronic studies have shown that age matters, and results can differ even across short time spans. One of the important

insights from for example Douglas Biber's work, is that genres differ in how they use language. People need to know these things when they use corpora. Because if you don't know what's in the corpus, you can't really use it. As I said, corpus linguistics is diversifying, so what skills you need to have depends on what you want to do. To answer your question "Should a corpus linguist be able to do programming?" Not necessarily, because if you are a corpus user rather than a corpus compiler or software developer, you don't need to do programming.

It's important to understand the quantitative side as well. Of course that part of corpus linguistics has developed immensely in the last few years, and the statistical methods have become more and more sophisticated. I mean it's not just that we now have statistical measures that are more precise than the percentages and normalized frequencies that people used before, but statistics is also becoming a separate specialization. I think it's important for a corpus linguist to have some understanding of statistics and the more you know the better. But I think few people can be a fully skilled statistician and a fully skilled linguist at the same time. It's important to find someone to cooperate with, or to know your limitations so that you don't say things that you don't have evidence for. Linguistics, not programming or statistics, is the main home ground of a corpus linguist. The other things are important but not as essential as the linguistic understanding.

XLX: Do you have any advice to young scholars who wish to do corpus research?

HH: I think for someone who is getting into corpus linguistics now, it is very important to get an overview of the field, and decide on your priorities. Because as the field is diversifying we need more and more skills. I think it's unlikely that you'll get all of those kinds of expertise in one head, so you need to cooperate with people who have the skills that you don't have. So I think I would advise people to find out who they can cooperate with and how.

XLX: Do young corpus linguists need to know the history of corpus linguistics?

HH: Maybe it is not essential, but it's good for them to know it. I think it is useful for people who have grown up with Internet and computers to understand how difficult it was for linguists to get their data without powerful computers, and also to appreciate the kind of work that goes into compiling a proper corpus. If you read an old corpus study that uses the LOB and the Brown corpora, which are one million words with various genres in them, and someone writes about genre variation on the basis of that, maybe you would say "Oh gosh, there is almost nothing in those corpora and they are from 1961. Why should I be interested or how can people say anything on the basis of so little data?" if you don't realize that at the beginning of corpus linguistics a million words was massive. It was really at the limit of what computers could hold and what you could search in, so it was a lot of data. It was much more adventurous to build a one-million-word corpus than it is to build a one-billion-word corpus now. Computer power and storage were very real limitations then and people had to think of clever ways of getting all that data into as little space as possible. I think it's important to understand that people who used small corpora and percentages rather than sophisticated statistics were not stupid. They were people who worked with whatever they had and perhaps even very adventurous pioneers who paved the way for what we can do now.

XLX: I totally agree. Thank you very much.