Interview with Eric Nyberg

Professor Eric Nyberg, of Carnegie-Mellon University (CMU), is one of the world's leading experts and most original thinkers in the field of natural language processing. How did it come about? IJLD's Bob Clark caught up with him at the Controlled Language conference in Seattle, and put the question to him.

Bob: So, Eric, tell me how you came to be involved in a field as esoteric as Natural Language Processing

Eric: I'm an Assistant Professor in the School of Computer Science at Carnegie Mellon University. I've been at Carnegie Mellon since 1986. I did my graduate work there and then I became a member of the faculty in January 1993. My interest in Natural Language Processing really stems from my first encounter with symbolic processing when I was an undergraduate at Boston University and even before that when I was a student in High School. I studied French quite seriously, in fact I studied French so well that I was exempted from taking any foreign languages in college because I replaced the language requirement with my SAT scores. Even though I was interested in computer science in general, I was not really interested in numeric processing. But I took a course in artificial intelligence where we did symbolic processing. We did a little bit of language processing then but I really decided that I wanted to work on symbolic processing and artificial intelligence. Then I was fortunate enough after I graduated from college to get a job working at GTE Laboratories in Waltham, Massachusetts at a time when they were starting a project on natural language interfaces for databases.

Bob: What year was this?

Eric: That would have been 1983. So I worked at GTE from 1983 to1986 and the focus there was on coming up with a tool that would allow senior managers who were thinking about various strategic directions for the company to access all of the data in the company databases without having to wait for a week for somebody to write an SQL query and then run some kind of an SQL job against the database. So I became involved with writing parsers, doing semantic interpretation and then essentially translating natural language queries into database queries. Interestingly enough, at that time, Jaime Carbonell, who is currently my boss at CMU -he's the director of the Language Technologies Institute – was a consultant at GTE and that's how I met him. So when I told him I was interested in going to graduate school, he said, "Well, you really should consider coming to Carnegie Mellon".

So in the fall of '86 I drove out from Boston to Pittsburg thinking that I would just stay at Carnegie Mellon long enough to get a PhD and then move back again. But you know, that was almost 14 years ago and I'm still in Pittsburg!

As a graduate student I worked on a variety of things related to knowledge-based machine translation. We did a system that did bi-directional English-Japanese translation for IBM PC manuals. That was kind of interesting. We also did some fundamental work on algorithms for natural language generation that I got involved in. While I was still a graduate student, this would have been about the fall of 1990, I think, we got a visit from some folks from Caterpillar. They were very interested in evaluating all of the current technologies for machine translation and they were very interested in the work that we were doing. Not only because we were working on knowledge-based MT, which achieves higher accuracy than transfer-based MT, but also because we were very willing to customise the system for their domain, their vocabulary, their language.

Most of the high-profile efforts that had gone on up till that time were really focussed more on "let's solve the general problem with machine translation, let's translate any text". So, if you look at all of the Japanese systems that were built in the 1980s, they were all trying to be sort of the ultimate translation engine. We looked at this and we realised that everything we learnt from the work we did with IBM led us to believe that you really needed to focus the vocabulary and the grammar of the systems if you really wanted to boost the accuracy. So we were already thinking in that direction and it was very fortuitous that Caterpillar came along. That was exactly the kind of thing they were looking at and then we raised this issue with them of, "Well, what if we control the text? You know, what if we limit the kinds of writing that you do so that it improves the accuracy even more?" And they said, "That's great and we don't even care so much about the translation benefits. We want to standardise and improve the documentation even for the English reader". Because they had a lot of issues with lack of consistency.

Not all the folks who write their documentation are formally trained as technical writers. Many of them have actually been at Caterpillar for many years as mechanics and in other capacities and they've grown into this type of job. So this is work that I started with Teruko Mitamura, who is also now a faculty member at Carnegie Mellon. We sat down and we wrote a system that could handle 19 sentences

from one chapter of one operating maintenance manual for them. Then they came back and took a look at that – we were just doing the analysis at that time – and they were very happy with that.

Then they said, "OK, let's do a prototype". So we built a prototype system in 1991 for English to French and the rest, as they say, is history. We got into a full scale development effort with them and over the last 10 years we've developed controlled language authoring, controlled language checking and they're now using translation into French, German and Spanish as part of their daily production system.

Bob: Did you have any idea that it was going to develop into something this big when it started?

Eric: I think we sensed at that time that there was really this niche that hadn't been filled. Nobody was really focussing on this idea of controlling the vocabulary or writing an MT system just for a particular domain because everybody was trying to solve the general problem. We realised right away that you could really do much better translations if you took those limitations into account. If you'd asked me back then if I'd still be working for Caterpillar in the year 2000, I'd probably have been pretty surprised if you'd told me yes I would be but I think in the end things have worked pretty much along the lines that we anticipated. I think we underestimated the complexity of the domain and we underestimated the linguistic complexity. We didn't appreciate all of the issues that are involved in changing the process of authoring, changing the process of translation, getting input and the buy-in from the people who use the software, whether they're authors or translators. We just learned a tremendous amount from having gone through this experience with them. But I think that initial spark of the idea that we had, that this was a niche that could really be exploited because no-one was working in that area, that was a good intuition and the devil was in the details, as they say, in terms of actually getting the system built.

Bob: So before you arrived on the scene what kind of things was CMU involved in?

Eric: CMU had always been strong in Language Processing and Natural Language Processing and that whole branch of Artificial Intelligence. So Jaime Carbonell, who's the Director of the LTI and was also my thesis advisor at CMU, came to CMU in 1979 after he finished his PhD at Yale and he immediately brought a lot of research momentum in this area in terms of language understanding and interfaces to databases. He had worked on systems in this area and then he wrote this really seminal paper in 1979 with two of his colleagues from Yale on Knowledge-based Machine Translation and how theoretically it would be able to do a better job by doing a deeper semantic analysis. I would point to that as being the thing that probably motivated everything that's come afterwards at CMU. Of course, Tomita was a graduate student at CMU and then he became a faculty member. He's very well known for his development of various parsing algorithms, some of which are still in use today, not only on our project but also on other projects at CMU.

Since that time things have grown quite a bit through the work that we did for industry, also work that we've done for the government. We have a variety of projects on Machine Translation now, not solely limited to technical translation but also in areas of speech to speech translation, translation for rapid assimilation of texts in an information gathering sort of an application. So the work that we've done for Caterpillar is only one piece of the larger picture. I think that Carnegie Mellon is probably one of the strongest places where this research is done today. ■