The Doctor is *In*

Language Engineering gets a powerful new advocate

Tony Hartley

Tony Hartley was recently appointed to the first British chair in Language Engineering at the Language Centre at the University of Brighton, where he lectures in computational linguistics and machine translation. He is also an associate researcher at the Information Technology Research Institute (ITRI), which is the main locus for research in language engineering at the University, with a special emphasis on information-design problems. Tony has also worked extensively as a French/English interpreter and translator. Sharon Denness asked him to explain this innovative discipline of language engineering and evaluate its practical potential.

Language International: *What is language engineering and what led to your interest in it?*

Tony Hartley: Let me quote from the European Commission's LingLink publication for a succinct description: "Language engineering applies knowledge of language to the development of computer systems that can recognize, understand, interpret, and generate human language in all its forms."

What the quote does not say is that these systems should be designed in conjunction with a thorough analysis of the needs of the people who will work with them. If a major goal is to use these systems to bring more people into the communication loop, then those same people need to be represented in the design loop too.

My own connection with this field goes back to my undergraduate days at the University of Salford, where the ethos at the time was application of language knowledge to real-world problems like translation, terminology, and interpreting.

Professor Frank Knowles (now Professor of Applied Language at Aston University) encouraged me to start programming, in Fortran. And then as a post-graduate at Besançon I met Yves Gentilhomme, who was working in MT and introduced me to algorithmic approaches to natural-language processing. And finally I am grateful to the late Professor Robert-Léon Wagner for introducing me to the political lexicology group at Saint-Cloud, an experience which convinced me of the immense value of corpora for linguistic research.

By temperament, I am extremely curious about the differences between languages and am interested in trying to uncover plausible rules underlying certain aspects of language behavior and model them in a program.

During my career I have worked extensively as a translator and interpreter and

> **"To my knowledge Brighton is the only university in the UK to use MT for translating communications with actual and potential research partners."**

have faced a number of language problems now being addressed by researchers in language engineering. My own experience as a language professional has been an important influence in my view of the language-engineering field.

LI: *What aspects of language engineering particularly interest you now that you are at Brighton?*

Tony Hartley: As a researcher, I have been working in natural-language generation, specifically the generation of multilingual software manuals. Under the leadership of Professor Donia Scott, the ITRI has gained an international reputation for excellence in this field. We are

the lead partners in Agile, a three-year project (funded by the EU) just beginning with colleagues in Prague, Moscow, Sofia, and Saarbrücken to generate instructional texts in Bulgarian, Czech, and Russian.

My other main research interest is controlled language. I am particularly interested in the creation of guidelines for the design of controlled languages for human languages other than English.

As a user of language technology, I am focused on the quality of MT output. To my knowledge Brighton is the only university in the UK to use MT for translating communications with actual and potential research partners in France. We are currently using Systran (Fre-Eng/Eng-Fre). The aim of this initiative is to integrate MT into the university's information-gathering and communication process. At the same time I intend to educate people to have reasonable expectations of what this technology can do. I am currently setting up an infrastructure to evaluate the MT systems we are using, including T1 (Ger-Eng/Eng-Ger). This exercise will include involvement from undergraduate students in the Applied Language course at Brighton. The results of our pilot studies have been very positive.

LI: *What is your teaching agenda as Professor of Language Engineering?*

Tony Hartley: My intention is to pursue my research activities while actively maintaining contact with professional multilingual information activities. I feel it is essential to stay in touch with the real needs of industry and education.

At the Language Centre, I am involved in an internal project titled "Composer" which has been running for the past two years. The goal is to support the authoring of scientific texts in English by non-native writers. At the moment we have a working prototype, and my objective is to find partners interested in developing this into a commercial product.

I also have teaching responsibilities. At the moment I am co-developing Master's modules in "multilingual information

design" which will be available from October 1998. This will be. I believe. the first course of its kind in Europe to bring together the subjects of software localization, information design, and language engineering. The modules will be offered on the basis of a week's intensive study with preparatory and follow-up activities. They are intended for students and participants from business and industry,

"A number of MT suppliers claim to be addressing the issue of ill-formed input such as Web pages... but I personally think they are grossly underestimating the scale of the problem."

from the UK as well as Europe. What is distinctive about this course is that it will bring together project teams of linguists and software engineers.

This same principle underlies the European Commission Socrates project we are leading, with the universities of Bayonne (France) and San Sebastian (Basque country, Spain). The idea is that translation students at one site work with software-engineering students at another on the localization of interfaces and documentation into two or three languages. In this context we plan to use Corel Catalyst, a new tool which combines an environment for interactive translation, translation memory, and project-management facilities. We will also be looking to GMS's Spanish/English version of T1 when it is available. [See *LI* 9.4 for a review.]

LI: What are the key challenges facing researchers in language engineering?

Tony Hartley: There are a number of issues, including dialogue translation and coping with ill-formed input e.g.. Web pages/Internet-based communication. A number of suppliers claimed to be addressing this issue at the recent MT Summit, but I personally think they are grossly underestimating the scale of the problem. If expectations are hyped, they will inevitably be dashed. "Sensible" MT

vendors are steering clear of this market for the time being, concentrating on the Intranet instead. Clearly, it is a potentially huge market, but a great deal of fundamental research into the nature of less formal communication is needed, especially in multilingual environments.

Another big challenge is the development of sharable resources, especially those that capture the knowledge necessary for accurate understanding of human language.

LI: Why have language-engineering applications had such a hard time breaking into business or public awareness?

Tony Hartley: The main reason has been the lack of publicity for both research and language-technology products. Also, most people have only recently gained access to a computer in their work environment. Public expectations of software interfaces are now very high and some of the earlier MT systems had rather daunting interfaces.

I think small and mid-market firms have too long been excluded from the research agenda, although various government agencies in Europe have paid lip service to involving them in language engineering. If language-engineering technology is to be taken up by smaller users, then the conditions of their involvement in projects must take realistic account of their circumstances.

This situation is changing. A big PR effort is beginning in Europe with support from the European Union for projects like Euromap (see *LI* 9.1), which is an encouraging start. There is also the MLIS (Multilingual Information Society) initiative which stresses user-friendliness in information management throughout society and the combatting of social exclusion.

One all-important factor is that LE products are coming down in price and increasing in usability and effectiveness; products such as Systran and T1 are now available at a reasonable price.

LI: What is your dream language-engineering application?

Tony Hartley: The dream application that I recently dangled in front of the audience at my inaugural lecture in October '97 was the "babel fish." According to *The Hitchhiker's Guide to the Galaxy*, when inserted into your ear, the "babel fish" enables you to understand any human language. If we forget fantasy and think of dreams that can be made true, then I would nominate any application that can be of real use to

people with physical or mental conditions that impair their ability to communicate through language. I also favor applications that enhance users' linguistic knowledge and skills by supporting them rather than replacing them in tasks like multilingual multimedia design.

LI: Do you share the feeling that current models can no longer deliver, and that language engineering is heading for a paradigm shift?

I do not accept that current models are bankrupt, though they may be impoverished. In general, there are rule-based models and statistics-based models, and much competition between the proponents of each. In part-of-speech tagging, for example, the statistics camp was claiming an upper limit of 97-percent accuracy, when along came a rule-based parser from Helsinki achieving 99.7 percent.

One avenue being explored in MT, at least, is to build hybrid engines that do rule-based and statistics-based processing in parallel. When they are done or time-out, whichever is sooner, another module of the system has the task of choosing the best translation from available outputs, possibly combining fragments from the two processes. These come with scores for "goodness" to guide the choice.

"One significant shift in practice in language engineering would be to ensure that every project team included a specialist in user-centered design."

In the meantime, there is a move which is already technically feasible but which might nevertheless represent a significant shift in practice, and that would be to ensure that every language-engineering project team included a specialist in user-centered design.

Contact Tony Hartley at afh@bton.ac.uk
Sharon Denness is a Research Analyst working with Equipe Consortium in the UK.
You can reach her at: sharon.denness@equipe.co.uk