Controlled languages are becoming crucial

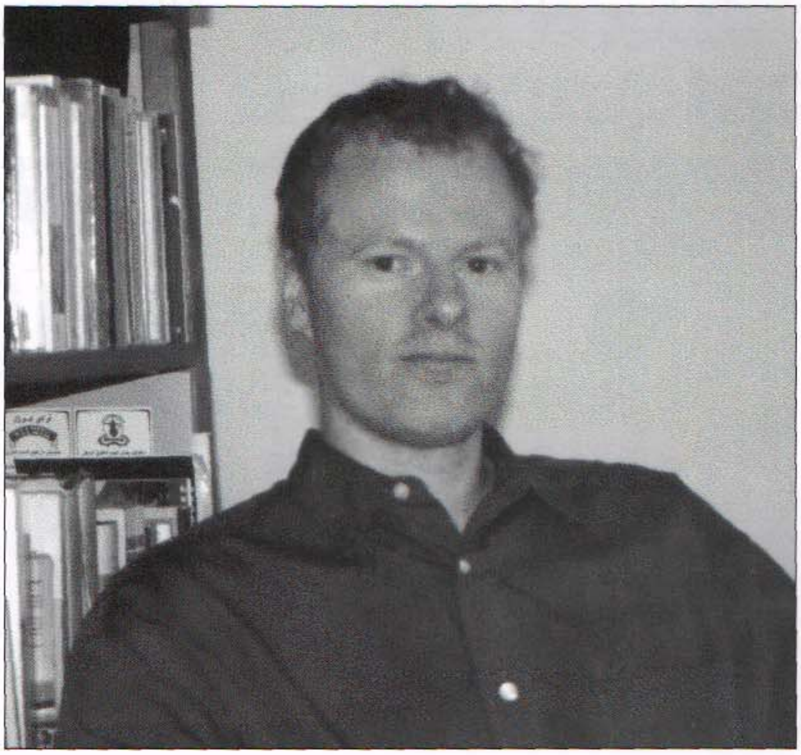
Report on the CLAW 98 conference

by
Andrew
Bredenkamp

Controlled languages (CLs) are increasingly seen as a crucial aspect of document authoring in a multilingual environment, especially where source files are enormous, may need to be translated into many target languages, and have an extended life.

The second Workshop in Controlled Language Applications (CLAW 98) was held in May at the Language Technologies Institute of Carnegie Mellon University in Pittsburgh, following on from CLAW 96 in Leuven. As with CLAW 96, the workshop brought together some big, long-term, American users and developers (Boeing, Caterpillar, CMU LTI, etc.) with existing, and potential, users and developers from Europe.

The workshop was really serving two different audiences, those who were interested in the integration of CLs in industrial scenarios, and those who were looking for, or offering software to assist in this process. Many companies are getting to the stage where they appreciate the potential importance of CL in the domain of large-scale technical documentation, but are unsure of how to get it up and running. Talks by representatives from Boeing, Aerospatiale, Caterpillar, General Motors, etc. gave a good impression of what was involved in an organisation adopting CL, as well as some of the benefits. There were also a number of companies (Lant NV and Cap Gemini for instance) offering generic, customisable language checking software applications.



A further distinction could be made according to the target of the CL texts. The main distinction between the Boeing and the Caterpillar approaches is that the former has as its aim the preparation of texts which can be easily read by users (i.e. engineers fixing airplanes), the latter seeks to ease the process of translating documents with machine translation (MT).

Interest in software support for the introduction of controlled languages has increased markedly

The interest in software support for the introduction of controlled languages has increased markedly, and this showed in the slightly more application oriented content of this year's workshop.

The workshop proper opened with a talk from Karen Hassen of Boeing, describing her experiences in training editors to use Simplified English at Boeing. She stressed how important (and often difficult) it is to convince authors and editors of the value of controlled language conformance, and the need for software tools to help them get used to writing in a CL. The talk thus served to contextualise many of the other talks and presentations, inasmuch as it was clear that implementing CL *without* software tools is a much more daunting task.

There followed an introduction by Willem-Olaf Huijsen to the key concepts in Controlled Language and an overview of existing work in the area of software support for CL checking.

The workshop presentations were a mix of industrial users relating their experiences and developers presenting their systems or system designs.

Arendse Bernthe from IBM presented

EasyEnglish, a software application which was designed principally to help with the preparation of texts for MT. IBM are integrating the language checker with their documentation process and use it mainly as a step to ensure that subsequent machine translation is more effective. Unfortunately, while the talk was very interesting and many of the insights might be picked up by others, the software is not currently available outside IBM.

Uus Knops from LANT NV presented LANT's work on integrating a language checker within their suite of translation technology applications (see feature article in the February 1998 issue of *Language Today*). Basically, the idea was to use the machine translation engine in their LANT®MARK product (based on METAL) and to write translation rules from Standard English to Controlled English (they claim the approach will transfer easily to other languages currently covered by the MT system). The "translation approach" to conformance checking has the advantage (assuming you have a good MT system already) that even rules requiring rather complete linguistic analysis can be checked, such as the disambiguation of different uses of *to*.

Caterpillar are, with Boeing, perhaps the most committed and experienced industrial user of controlled languages - they have been using their own controlled Englishes for over 25 years. So Christine Kamprath's description of the current Caterpillar system, developed with Carnegie Mellon University and Carnegie Group Inc., was of great interest to all. The project has built what amounts to the state-of-the-art in language checking for MT in a state-of-the-art controlled technical authoring domain. For anyone building systems there were plenty of good ideas, derived from a careful 25 person-year development period. Those thinking of implementing some controlled authoring in their own organisations will have seen what is involved in really doing the job properly. Not everyone will be able to persuade their bosses to spend that kind of money!

Boeing are still developing their controlled languages as well. Richard Wojcik told us about Boeing Technical English (BTE) which is being developed from AECMA SE with the aim of providing a controlled language which could be used for all kinds of technical documentation. One of the main tasks is to extend the dictionary of AECMA SE to a much broader technical domain without losing the control over terminological consistency and the unnecessary use of synonyms.

Further ideas on improving SE (Simplified English) were presented by Isobel Heald, as the result of joint work with Rémi Zajac. The work focused on improving the usability of noun cluster rules in SE (e.g. no noun clusters of more than three nouns), which writers are known to find difficult.

The field of controlled language is still very much limited to English, so it was encouraging to hear about some research projects developing software for controlling other languages.

One research project showing real results was the German nationally-funded MULTILINT project carried out by IAI in Saarbrücken, in co-operation with BMW. The work was interesting not only because it focused on a language other than English, but also because it did not make use of large-scale processing machinery available to many of the other systems (those of LANT, IBM and CMU for instance). Working with quite simple computational machinery, they have been able to identify an impressive array of complex linguistic errors or stylistic infelicities.

Satoru Ikehara presented a paper giving an insight into the kind of research going on in Japan in the domain of CL. The work focused on preparing texts for Japanese-English MT, by means of rules for re-writing problematic structures automatically (apparently Japanese authors have been very resistant to the idea of controlled language). The automatic re-writing process was integrated with the ALT Japanese to English MT system, and evaluation showed a 20% improvement on the results from the system.

Research projects are developing software for controlling languages other than English

Aerospatiale were represented not just in terms of their work with AECMA SE, but also, as a French company, with moves towards controlled French. Kathy Barthe from Aerospatiale presented the work of GIFAS (the French Aerospace Industries Group) in developing a controlled French based on Simplified English. The aim was to allow French authors to write in "Français Rationalisé" and have the text translated reliably into SE. Veronika Lux, also from Aerospatiale, presented a detailed study of the types of rules used in controlled languages with the aim of designating sets of rewriting rules which might form modules in language checking application.

Jarno Tenni presented an interesting and original research project, Webtran by VTT Finland, for maintaining a text resource (such as sales information) in a controlled language to facilitate its translation. In this case, texts are stored in a controlled Swedish for real-time translation into Finnish and, later in the project it is hoped, into Estonian and other languages.

Each day also included a panel session, where a number of experienced CLers talked about their experiences and their views of the future.

On the first day, the formal session ended with a panel discussion on "Standardisation and Acceptance of Controlled Language" - chaired by Rémi Zajac from New Mexico State University, with Arendse Bernth from IBM, Linda Schmandt from Carnegie Group, and Eric Adolphson from Caterpillar. The discussion centred around how to encourage *writers* to accept the use of CLs, rather than the users, and what organisational factors were important, for example how the integration with existing documentation workflows should be achieved. Relatively little was said about standardisation, it was noted that there was very little in the way of standards in the world of CL, except as far as the CLs themselves are a standardising influence. The panel pointed out the problem of standardising the process of editing text for CL conformancy, since different editors tended to produce different conformant text.

Very little work has been done to get user feedback

It was also noted that very little work had been done to get real user (i.e. reader) feedback on the benefits of CL conformant documentation.

The second day's panel looked at how AECMA Simplified English was coping after ten years of use in the aerospace industry. The general feeling seemed to be that, while there was still a lot of work to be done in the domain of gaining acceptance for CLs, generally it had proved its worth. Again the issue of evaluation raised its head, and the panel agreed that it would be really motivating for writers having to learn controlled authoring, if they knew that the texts they were writing would be really easier for users to work with.

The "MT approach" to Controlled Language checking has also been adopted by CoGenTex Inc. Richard Kittredge presented a processing engine which can offer paraphrases (i.e. corrections) of texts on the basis of quite abstract linguistic representations.

The conference ended with an upbeat talk by Kurt Godden from General Motors. GM's Controlled Automotive Service Language (CASL) is part of a large-scale re-organisation of their documentation workflow. CL conformance will be integrated as part of a complex multilingual documentation system using automated translation technology (translation memory and machine translation) from LANT NV in Belgium. We will have to wait and see how this works, since they will not be ready to start the pilot until the end of this year. However it was encouraging to see yet another "big player" taking controlled languages seriously.

So what was missing from the workshop? Well, the workshop really reflected the state of the market in CL applications. As a developer of CL checking technology, I would have to say that the market is still a little young, and the generic customisable "industrial strength" applications are still thin on the ground, especially if your source language is not English. What is also missing from the field is a strong tradition of evaluation (with the notable exception of the work in Japan) - Emmanuelle Rodier presented an interesting paper outlining the difficulties in evaluating checkers when both the input text and the internal workings of the checker were confidential.

However this is definitely technology with a future, as large companies realise the value of consistency, simplicity and translatability in documentation. ■

INFORMATION

For more information on CLAW 98, see:
http://www.lti.cs.cmu.edu/CLAW98

For details of future events, etc. contact Eric Nyberg as:
ehn@cs.cmu.edu

Andrew Bredenkamp can be contacted as:
andrewb@dfki.de

His home page is at:
http://www.dfki.de/~andrewb