The KANT Translation System

From R&D to Large-Scale Deployment[1]

Eric Nyberg, Carnegie Mellon University
Christine Kamprath, Caterpillar, Inc
Teruko Mitamura, Carnegie Mellon University

KANT is a knowledge-based translation system for multi-lingual information dissemination in pre-defined technical domains. KANT has been deployed for French and Spanish translations at Caterpillar, Inc. In this article, we give a brief technical summary, followed by a discussion of the user's perspective and experience with the system. We conclude with a discussion of ongoing work and future plans for the KANT software.

Overview of the KANT System

The KANT system (Knowledge-based, Accurate Natural-language Translation) is a machine translation system for automatic translation in technical domains, developed at Carnegie Mellon University. KANT is designed to allow the user to define the terminology for specific domains, where the meanings of terms are limited to those relevant to a particular context (e.g., heavy machinery, computer equipment, etc.); as a result, KANT can achieve better translations than general-purpose translation systems when applied in specific technical domains.

Full details concerning the design and implementation of KANT are beyond the scope of this article. The main characteristics of the system include the following:

- KANT supports the use of controlled input language with explicit conformance checking during text authoring. The same grammar is used during conformance checking and machine translation;
- KANT's Interlingua architecture supports efficient multi-lingual translation;
- KANT is designed to provide Automatic Machine Translation (AMT) with minimal post-editing;
- KANT is a Knowledge-Based Machine Translation (KBMT) system, making use of domain-specific lexicons, grammars and mapping rules for each language, to improve overall translations quality in each domain;
- The KANT design supports the use of SGML tagging within sentences, and SGML is treated as an integral part of the text to be translated.

For more details regarding the design and implementation of KANT, the reader is referred to the full set of KANT publications, available via the World-Wide Web from the KANT home page:

`http://www.lti.cs.cmu.edu/Research/KANT`

[1] Published in the Localization Industry Standards Association (LISA) Newsletter, Vol 2:1, March 1998, ISSN 1420-3693.

How did Caterpillar make the decision to use KANT?

Caterpillar required a system to handle a large volume of authoring and translation, for 350 current products as well as old products still in use (average product age is 17 years). Authoring encompasses multiple document types, each with different writing styles and standards. If the source and target documents are not accurate, there are potential legal liabilities; this is important because document production must support worldwide manufacturing and distribution of products.
Caterpillar was also very interested in controlled English authoring and compliance checking; as a result, we created a specific controlled English called Caterpillar Technical English (CTE).

CTE supports a rich subset of the syntactic structures of English and a large Caterpillar-specific technical lexicon. The basic CTE text is the information element (IE), a reusable component that can be shared among multiple documents. In addition to the obvious benefits of reusability, CTE also promotes consistency in writing, which results in a standard "look and feel" from manual to manual. To make CTE successful, Caterpillar required conformance checking software that was restrictive enough to promote consistent authoring, while flexible enough to allow writing of complex technical material.

CTE also has benefits for both manual and automatic machine translation (AMT). Consistent source authoring improves the accuracy and consistency of machine translation, and also improves the consistency and cost effectiveness of manual translation. The use of a controlled source language helps to keep post-editing of MT output to a minimum. Since Caterpillar requires translation to multiple target languages, KANT's interlingual architecture was appealing.

Caterpillar has committed to the use of SGML document markup, to promote efficient publishing for both paper and electronic delivery of CTE and target language documents. To support SGML most effectively, the compliance checker and translation software accomplish the following:

- Recognize and interpret markup tags inside sentences, without stripping them out;

- Interpret markup tags variably, according to function;

- Arrange markup tags properly in the target language output.

Caterpillar also required software that could meet various integration criteria:

- The CTE checker and AMT software must to interface smoothly with other authoring tools;

- The software must handle many embedded references to external objects (titles, warnings, graphics, tables, lists, etc.).

Because of the flexible, modular design of the KANT system, it was possible for us to meet all of these requirements by developing the appropriate lexicons, grammars, and interface protocols that were necessary.

The vocabulary and grammar, and conformance checking software were integrated into a user interface called the Language Environment (LE), which was constructed by Carnegie Group, Inc. The

LE software interacts with the KANT system as a co-process, while managing the interaction with the user and the SGML text editing environment (Arbortext).

What did it take to deploy KANT at Caterpillar?

The KANT application for Caterpillar was begun in November 1991, with the development of CTE, the Language Environment, and target language translation engines taking place in parallel. The cost to develop CTE involved both initial development (vocabulary and grammar definition, implementation), and also training authors to use CTE effectively. The ongoing cost of CTE maintenance involves both incremental vocabulary update (as new Caterpillar products are introduced, requiring new documentation) and grammar refinement (as new document types are introduced or existing documents are changed).

The cost to develop the target language translation involved initial development (CTE vocabulary translation, grammar definition and implementation), plus evaluation and refinement in conjunction with use by translator/post-editors.

The main challenges in deployment of machine translation included limitations on the controllability of the source language (translators are accustomed to "rewriting the source" for stylistic reasons), as well as the general issue of stylistic quality (sentence-by-sentence machine translations are typically understandable, and may not require post-editing; but translators often choose to rewrite to improve quality, even when not strictly necessary). The ongoing cost of maintenance includes the cost of updating the translated vocabulary when new CTE terms are created, as well as occasional changes to the target grammars when CTE grammar changes are implemented.

In order to develop the Language Environment, Caterpillar undertook development, pilot, and training phases from 1992 to 1997, involving about 5 full-time equivalent employees over each of the 5 years. Participants at Caterpillar included linguists, pilot authors, trainers, mentors and system developers.

Development Details

- *Terminology development*: Corpus analysis programs were written to extract candidate terms from 50M of existing documentation. These candidate terms were screened by authors for inclusion in CTE. The same development corpus was analyzed by linguists in order to develop CTE writing guidelines. Ambiguous terms were identified, and accepted domain meanings were assigned to each term. Finally, usage examples were constructed for each term, for inclusion in the Language Environment interface as a lookup aid to authors.

- *Grammar development*: SGML tags were designed for sentence-internal (constituent-level) markup, as well as formatting markup (document level). Once an initial SGML Document Type Definition (DTD) was constructed, an authoring pilot phrase was conducted to resolve DTD, grammar, and terminology issues; the usage examples were refined based on author feedback.

- *CTE Maintenance*: Caterpillar developed in-house problem report software and a process for authors to request both terminology and grammar updates; this required that Caterpillar develop in-house linguistic expertise to perform terminology screening prior to update. Caterpillar also developed software and processes for electronic review of each author's work; these steps were deemed essential to maintain the integrity of CTE writing standards.

- *CTE Training*: During 1994, authors were acclimated to the new authoring paradigm during informal bi-monthly seminars, for one year in advance of CTE training. Formal CTE training was prepared and given to authors starting in 1995.

- *AMT Development*: CTE terms were translated to each target language, using a specialized vocabulary translations editor (VTE) developed at CMU. The VTE was designed to provide context, translation history and definitions for each term. Caterpillar located and trained qualified translators, experienced with Caterpillar's technical terminology. Linguistic specifications were written for each target language, and a process was developed to elicit requirements and approval from the Caterpillar language experts.

- *AMT Evaluation*: We explored several different evaluation criteria for target translations (percentage of sentences correctly translated; sufficient coverage of items in requirements specifications; sufficient level of increase in overall translator productivity). One challenge was to develop a method for factoring out the effects of author-induced (rather than AMT-induced) phenomena which could impact the evaluation; the quality of the input text can vary widely, and the well-known "garbage in, garbage out" principle was found to hold when input text quality was marginal. Other factors affecting the evaluation included the accuracy of term translation (AMT is only as good as the human-provided terminology resources); the level of post-editing done by the translator (despite training in "minimal post-editing", many translators continue to post-edit texts to human levels of stylistic quality, even when unnecessary); and the overall level of translator experience.

- *Limits on Controllability of Input to AMT*: The CTE terminology (the basis of AMT terminology) is too complex to control completely; most difficult to control are common terms which are highly ambiguous in the Caterpillar domain (for example, the word "valve" has at least nine different meanings, depending on the type of valve the role it plays in the vehicle system it is found in). Despite careful analysis during terminology development, the Caterpillar domain is too complex for lexicographers to anticipate all the ways authors use words. Although the Language Environment performs a CTE conformance check, adherence to CTE principles by authors is variable, as authors may misuse words that are then mistranslated by AMT and must be post-edited. In some instances (e.g., deadline pressure), authors may bypass CTE checking completely; this may result in publication of inconsistent English documents, and the use of uncontrolled input with AMT (which usually increases the post-editing effort).

- *Acceptance of French AMT system*: Following our experience with several evaluation methods (mentioned above), we determined that increased post-editor productivity was the ultimate goal of the AMT system. Indeed, any measure of "accuracy" which does not correlate with productivity gain is potentially misleading and counterproductive. Caterpillar conducted a formal usability study for the KANT French AMT system to determine productivity gain, and verified that AMT increased productivity over manual translation by at least a factor of 2 to 1. As a result, the KANT French system was accepted for production translation use in 1996.

- *Acceptance of Spanish AMT system:* To evaluate the KANT Spanish AMT system, Caterpillar chose not to do a formal usability study; instead, Caterpillar's Spanish expert, a Caterpillar linguist, and the CMU technical team worked together in a rapid feedback/refinement phase, over a period of several months. Documents were translated with the current version of the system, post-edited by the expert, and resulting problem areas were identified and fixed. This approach to alpha testing can be highly effective when a small team of highly-motivated individuals works together for rapid turn-around of new test systems. As a result of the successful refinements made to the initial Spanish system, the KANT Spanish system was accepted for production translation use in 1997.

What are the benefits reaped so far?

- Use of CTE has increased the consistency of English writing and terminology use;

- Production of technical manuals has increased;

- French AMT system is operational for translated document production, with at least a 2 to 1 productivity gain;

- Spanish AMT is operational for use in translated document production, and productivity evaluations are ongoing (the informal perception of the translators is that Spanish AMT requires very little post-editing);

- The awareness of language-related issues at Caterpillar has increased, including the value of writing and terminology management, standardization, training and support for authors.

What are some unresolved issues?

- It is difficult to devise a valid metric for evaluating the usefulness of an MT system which factors out other variables (such as author compliance, quality of terminology translation, knowledge of SGML, and training on the interface) which may affect the assessment of MT productivity;

- Translators are reluctant to back AMT. They would rather use their own terminology and translate their own way; they're reluctant to do minimal post-editing; it takes time to learn how to post-edit efficiently;

- Managers may find it difficult to assess the validity of translator's judgments on quality, productivity, and the amount and nature of post-editing required; this makes it potentially difficult to evaluate the benefits of AMT;

- CTE and target language terminology work are critical and time-consuming, and people who are qualified to do terminology work are difficult to find. The time needed for terminology work during development and production use is usually underestimated;

- CTE and target language terminology maintenance costs fluctuate, and can be difficult to plan for due to frequent vocabulary additions (for new products, etc.). When CTE terminology changes constantly, the target language vocabulary must be enhanced continuously to maintain consistency.

What are our plans for the future?

In order to address many of these challenges for MT deployment, the KANT development team at CMU has already begun a redesign and reimplementation of the KANT software. The new KANTOO system (KANT Object-Oriented) is designed to improve the efficiency of implementation and deployment of new KANT applications. The main features of KANTOO include:

- **Language Translations Database (LTD)**: A PC-based Oracle database and forms application for rapid development and efficient maintenance of target language terminology banks;

- **Lexicon Maintenance Tool (LMT)**: A PC-based Oracle database and forms application for rapid development and efficient maintenance of source language vocabulary (e.g., CTE terminology);

- **KANTOO Analyzer**: A reimplementation of the KANT analyzer, which is used for grammar checking and analysis during translation;

- **KANTOO Generator**: A reimplementation of the target language translations engine;

- **Knowledge Maintenance Tool (KMT)**: A graphical user interface which allows real-time browsing, editing, and incremental update of the knowledge sources used during analysis and generation (lexicon, grammars, domain model, mapping rules, etc.).

The overall goals of the KANTOO reimplementation include:

- Lowering the cost and time for terminology maintenance by providing better database management tools;

- Lowering the cost and time for system knowledge updates by providing better troubleshooting tools for the developer, as well as an improved modular design for the software itself (which promotes easier incremental update);

- Improving the general robustness and maintainability of the software by porting from Lisp to C++;

- Improving the portability of the software by reimplementing in C++ (we plan to support the system for Microsoft Windows as well as Unix in the future).

The LTD and Analyzer modules of KANTOO have already been implemented; the remainder of the KANTOO modules are scheduled for completion in 1998. Deployment of KANTOO modules at Caterpillar will begin during 1998.

At Caterpillar, ongoing deployment of the KANT system is focused on the following goals for the near future:

- Bring maintenance of the CTE terminology and CTE editing software in-house at Caterpillar;

- Implement the Spanish AMT system in production use during 1998;

- Improve translation terminology management;

- Begin the deployment of newly re-designed KANT software (KANTOO) as new modules become available in 1998.

Our experience thus far has demonstrated that KANT can have a significant positive impact on productivity. Nevertheless, many challenges remain in an environment with a complex set of products and document types, and where terminology is updated constantly. At the same time that the KANT application for Caterpillar has advanced the state of the art in deployed MT systems, it has also helped to set the research agenda for future work on new systems at CMU.

Eric Nyberg

Carnegie Mellon University

5000 Forbes Avenue
Pittsburgh, PA 15213

Phone: (412) 268-7281
FAX: (412) 268-6298
Email: ehn@cs.cmu.edu