MT Summit Catamaran Resort Hotel San Diego
29 Oct - 2 Nov 1997

Using MT in a Corporate Setting

Christine Kamprath, Ph.D.
Senior Computational Linguist

CATERPILLAR®

1. How did Caterpillar make the decision to use MT? Why did we choose CMU/s KANT system?
2. What was the investment (time, money) to make it work?
3. What are the benefits reaped so far?
4. What are some unresolved issues?
5. What are our plans for the future?

1. How did Caterpillar make the decision to use MT? Why did we choose CMU's KANT system?

- We needed a system that could handle the huge volume of writing and translating that we do
  - Documentation for 350 current products
  - Documentation for old products still in use (average product age is 17 years)
  - Multiple document types with different writing styles and standards
  - Accuracy of documents regulated by law
  - Worldwide manufacturing and distribution of products

- **We wanted controlled English input and software to control it.**
  CTE (Caterpillar Technical English)

For ENGLISH

- **New Authoring system was designed to exploit reusable "information elements" (IEs)**
- **Need to promote consistency to provide standard "look and feel" within manuals**
  - **IEs derived from different sources**
  - **Varied authoring staff**
- **Needed controlled English software sophisticated enough to allow writing of complex technical material**

For TRANSLATIONS

- **To improve accuracy and consistency of MT output**
- **To improve consistency and cost effectiveness of manual translation**
- **To keep post-editing to a minimum**

- **We needed to translate our technical documentation into multiple languages**
  - **KANT's Interlingua architecture appealed to that need.**
- **We were developing a complex new Authoring system that an MT software provider must accommodate**
  - **We wanted to use SGML for "lights out" publishing**
    - **For both paper publishing and electronic delivery of documentation**
    - **For both English and world-wide target languages**
  - **We needed SGML markup *within* sentences, so grammar had to:**
    - **Recognize the markup tags in their context**
    - **Interpret the markup variably according to function**
    - **Arrange the markup grammatically in the target language output**

- New Authoring system (cont.)
  - CTE and AMT systems had to interface smoothly with the other Authoring tools
    - Complex electronic file management systems(FMS)
    - Many references per IE to external objects, e.g., titles, warnings, graphics, tables, lists
    - References to external objects are inside SGML markup, imbedded in the text

★ The CMU developers have worked with us to customize the KANT to meet these requirements.

- We wanted assistance in managing the translation software development at a distance
  - We contracted with Carnegie Group, Inc. (CGI), a management and development company with linguistic expertise.
  - CGI was a co-developer of software with CMU and developed
    - The Language Editor software
    - The usage examples offered by the software to authors
    - Controller routines to interface with the KANT Analyzer
  - CGI subcontracted with CMU to:
    - Develop the linguistic engine (KANT) to meet CAT's linguistic and system requirements
    - Interface with the Lnaguage Editor software
  - CGI prepared and delivered the initial CTE training and training documentation for authors

2. What was the investment (time, money) to make it work?

- **Launched Project in November 1991**
- **CTE, Target Languages, and Authoring system were developed in parallel**

CTE

- Development
- Maintenance
- Training

AMT

- Development
- Evaluation
- Limits of controllability of input to AMT
- Acceptance of French AMT
- Maintenance
- Training

Authoring

- Development

- **CTE Development**
  - Personnel needed at Caterpillar for Development, Pilot, Training (1992-1997)
    - Averaged about 5 full-time equivalent employees over each of 5 years
    - Linguists, pilot authors, trainers, and mentors
  - Terminology
    - Corpus analysis to extract terms
    - Screening of candidate terms
    - Writing guidelines devised
    - Ambiguous terminology identified
    - Accepted domain meanings assigned
    - Usage examples for Authoring interface written
  - Grammar
    - SGML: Sentence-internal as well as formatting markup
    - Authoring Pilot to resolve DTD, grammar, and terminology issues
    - Usage examples honed for author interface in CTE software

■ CTE Maintenance

- Developed problem report software and process for author requests of terms and grammar
- Developed in-house linguistic expertise to do terminology screening prior to addition
- Developed in-house expertise for CTE lexical maintenance to speed vocabulary updates
  - CTE Language Editor software
  - CTE terminology addition
- Developed software and processes for electronic review of CTE work of authors
- Ongoing effort to maintain integrity of CTE writing standards

■ CTE Training

- Presented bi-monthly Brown Bag seminars for a year in advance of CTE training, 1994
- Prepared and administered CTE training to authors, beginning 1995
- Newsletter -- *CTE Author* -- keeps authors informed of enhancements, reminds them of key authoring principles

■ AMT Development

- Translated CTE terms
  - CMT created translation software to provide context, translation history and definitions
  - Located and trained qualified translators, experienced in Caterpillar terminology
  - Developed translation management software and processes
- Drafted AMT linguistic specifications
- Developed process to elicit requirements and approval from Cat Target Language experts

■ Evaluation of AMT System

- Explored a variety of means to evaluate the success of the AMT system
  - "percent" of correct translated output?
  - sufficient coverage of items in requirements specifications?
  - sufficient level of increase in translator productivity?
- Needed a means to factor out the effects of Authoring-induced variables on the evaluation:
  - Quality of CTE input
  - Accuracy of term translation
  - Level of post-editing done by translator
  - Level of translator experience
  - Level of Caterpillar domain expertise of translator
  - Interference of Authoring problems, difficult to identify those that come from AMT

★ Determined that increased productivity was key measure of AMT system

★ AMT output cannot be evaluated independently of quality of CTE input.

- **Limits on Controllability of Input to AMT**
  - CTE terminology (the basis of AMT terminology) is too complex to control completely
    - New terms that have not yet entered the CTE vocabulary cause gaps in output
    - Domain is too complex for lexicographers to anticipate all the ways authors use words
  - Adherence to CTE principles by authors is variable.
    - Authors may misuse words that are then mistranslated by AMT and must be post-edited
    - Sentence structures are ambiguous and complex in ways that authors cannot anticipate
      -- so AMT may translate an unintended interpretation of the sentence
      -- so AMT may fail on analyzing and translating the sentence
    - Word use is very complex in Caterpillar's technical domain
      -- Lexical selection rules and disambiguation techniques don't keep up with the variety
  - Authoring system indeterminacies cause need for translator review of AMT output
    - References to documentation in the text may or may not require translation
    - Words on decals (e.g.) are translated in some countries and not in others, so neither system nor author can always correctly anticipate how to render them
    - Other formatting or technical indeterminacies may require translator review

- **Acceptance of French AMT System**
  - Determined that increased productivity is the key purpose of AMT system
    - Measures of accuracy that do not correlate to productivity gain are misleading and counterproductive
  - Conducted a Usability Lab for French AMT system to determine productivity gain
  - Verified that AMT increased productivity over manual translation
    - AMT system accepted for production translation use
- **Acceptance of Spanish AMT System**
  - Three-phase Refinement process prior to Acceptance of Spanish system for use in Production Translation environment:
    - Compared post-edited Spanish with AMT output sentence by sentence
    - Improved lexical selection and term translations
    - Piloted system in production environment
- **AMT Maintenance**
  - Developed software and process for ease of review of outpus and semi-automatic AMT problem reporting
  - Developed a process for updating AMT terminology as new CTE terms enter the system
  - Developed a means for feeding AMT terminology concerns into maintenance of CTE terminology (e.g., need for additional terminology disambiguation)
  - Developed a multi-lingual database to keep AMT terms accessible for manual translators

- **AMT Training**
  - Overview of AMT system components and processes to clarify sources of problems
  - Training on use of FMS system tools, editing short-cuts, etc.
  - Post-editing training:
    - Amounts to development of *Caterpillar Technical Spanish, Caterpillar Technical French*, etc.

- **Authoring Development**
  - Clarified number, purpose and content of various document types
    - 15 document types
  - Determined new structure of documents
  - Developed and codified SGML markup patterns (DTD) for English and 35 other languages
  - Developed general authoring stylistic guidelines
  - Developed electronic File Management System (FMS)
    - currently about 60 integrated pieces of software
  - Developed data for FMS tables
  - Had text of standardized external objects translated
  - Determined standard keyboards and fonts for languages
  - Developed means of integrating work of external vendor translators into Authoring system
  - Designed and developed workflow and software to support it
  - Developed Translation Memory Tool to increase reuse of previously translated sentences

3. What are the benefits reaped so far?

- Increased consistency of English writing and terminology because of CTE
- Reuse of IEs speeds production of technical documentation
- AMT systems offer significant productivity gain over manual translation
- AMT promotes terminology consistency and serves as training ground for new translators in terminology & writing style
- Heightened awareness of language-related issues at Caterpillar:
  - Value of writing and terminology management
  - What it takes to standardize
  - Definition of standard
  - Agreement among users of the system
  - Technical and financial support
  - Training
  - Need for highly qualified work force: writers, translators, terminologists, and system maintainers

4. What are some unresolved issues?

- Management Issues
  - It is difficult to devise an acceptable metric for evaluation of usefulness of MT system
  - 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 is difficult for managers to assess the validity of translators' judgments
    - On quality and productivity evaluations
    - On amount and nature of post-editing needed to maintain publishable quality

- Terminology work is critical
  - It is difficult to find qualified people to do terminology work well
    - English terminology expert
      --Must create definitions & classes for terms
      -- Needs domain knowledge, linguistic training, and knowledge of transl. issues
    - Translation terminology expert
      -- Needs domain knowledge, translation experience, some linguistic knowledge, understanding of how AMT system uses the terms
  - Costs of upkeep are continuous and unpredictable given frequent vocabulary additions to reflect new products and processes documented
  - Time needed for terminology work in development and production is often underestimated
  - Difficult to train and maintain consistency standards among offsite translators

5. What are our plans for the future?

- Bring maintenance of the CTE terminology and CTE editing software in-house
- Improve translation terminology management
  - Across all Caterpillar translation languages
  - With our vendor translators (website terminology depository)
- Expand the use of AMT as Authoring expands to include more of the translators' work
- Support the development of tools to perform more maintenance functions in-house
- Determine optimal use of Translation Memory and AMT per document type and writing style
  - Relative effectiveness of TM and AMT varies by complexity and ambiguity of the input
- Continue to expand use of our product-line-specific ambiguity reduction techniques