Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English[1]

Christine Kamprath & Eric Adolphson
Caterpillar, Inc.
{kamprck,adolpej}@cat.com

Teruko Mitamura & Eric Nyberg
Carnegie Mellon University
{teruko,ehn}@cs.cmu.edu

1. Introduction

Caterpillar Inc., a heavy equipment manufacturing company headquartered in Peoria IL, supports world-wide distribution of a large number of products and parts. Each Caterpillar product integrates several complex subsystems (engine, hydraulic system, drive system, implements, electrical, etc.) for which a variety of technical documents must be produced (operations and maintenance, testing and adjusting, disassembly and assembly, specifications, etc.). To support consistent, high-quality authoring and translation of these documents from English into a variety of target languages, Caterpillar uses Caterpillar Technical English (CTE), a controlled English system developed in conjunction with Carnegie Mellon University's Center for Machine Translation (CMT) and Carnegie Group Incorporated (CGI).

This paper describes the characteristics of CTE, the effort required to develop CTE and the benefits observed so far, what we learned from the development process, remaining challenges, and plans for the future.

2. A Precursor to CTE: Caterpillar Fundamental English

CTE was not the first controlled English deployed at Caterpillar. In the 1970's, Caterpillar utilized a different controlled English approach for technical authoring and international distribution of documentation. This approach to writing was called Caterpillar Fundamental English (CFE), and involved the use of an extremely limited vocabulary and grammar. CFE was intended as a form of English as a Second Language for non-English speakers, who would be able to read the service manuals written in CFE after some basic training. This approach was intended to eliminate the need to translate service manuals.

CFE was designed around the basic English sentence patterns that could be learned in

[1]Proceedings of the Second International Workshop on Controlled Language Applications (CLAW '98), Pittsburgh, PA, May 1998.

an introductory ESL class. The initial version of CFE, designed in 1972, had a vocabulary of about 850 terms. The intent was to use an illustrated parts book and to heavily illustrate the service manuals so that the simplified CFE text could be followed by the service technician. Caterpillar employed CTE for slightly over ten years.

Caterpillar abandoned the CFE approach for a number of key reasons:

- The complexity of CAT equipment was expanding rapidly, especially in the areas of high-pressure hydraulics and electronics. The limited vocabulary was simply insufficient in these areas.
- Many service technicians in developing countries, especially in South America, were not learning CFE. Their limited formal education made the teaching of any foreign language and reading skills difficult. At the same time, high turnover of service technicians (as high as 25%) made CFE training very difficult.
- The increasing complexity of CFE created additional training requirements and increasing the costs of training.
- Growing markets in cultures that do not use the Roman character set made the introduction of CFE in these markets very difficult.
- For many cultures it is a point of national and cultural pride to have service literature translated; translated material was therefore recognized as an important marketing tool.
- Perhaps most importantly, the basic guidelines of CFE were not enforceable in the English documents produced. While there was a simplification (and a general improvement) in the writing of the English technical authors, very few documents, especially later in the program, were compliant with the CFE guidelines. So, while the documents were labeled as "CFE compliant," many users who had taken the CFE instruction were not able to read them. Beyond extensive editing and proofing of the English output by human editors, there was no means of enforcing compliance.

Because of all of these factors, CFE was discontinued in 1982[2]. In the mid 1980s, the first word processors were introduced into the Caterpillar writing environment, and by the late 1980's the rapid development of hardware and software technology led Caterpillar to re-examine their composition and publication procedures, with a view to gaining more automation and control over authoring. Improving the quality and reducing the cost of translation was also a consideration. It was determined that an enforceable controlled English was now possible, given advances in language parsing technology. An enforceable

[2]Caterpillar Fundamental English continued to appear outside of Caterpillar. In the late 1970's, it was marketed as the International Language for Service and Maintenance (ILSAM) and in a slightly different form as Basic 800. Research was continued by the Communications Studies Unit of the University of Wales Institute of Science and Technology. Unfortunately, Caterpillar has not followed the development of CFE derivatives since the early 1980s; however, we still continue to receive inquiries concerning CFE.

controlled English would enable a much higher degree of compliance than was available in the 1970's. These developments led to the design of the next generation of controlled English at Caterpillar: Caterpillar Technical English (CTE).

3. Characteristics of CTE

CFE employed under 1000 terms; many of these had broad semantic scope and it was assumed that they would be disambiguated in context by the human reader. In contrast, there are around 70,000 CTE terms, of narrow semantic scope, which are designed to be unambiguous to both the human reader and the checking software. The design and implementation of CTE are intended to provide the following characteristic benefits of controlled language authoring:

- Controlled input to CAT's Automated Machine Translation (AMT) system, in order to improve translation quality and reduce the cost of manual translation to minimal postediting.
- Standard terminology use and writing style for all English technical documentation:
  - A centrally-located software environment for use by a authoring group containing 80-150 authors;
  - Full linguistic analysis of each input sentence;
  - Sentence-by-sentence (interactive) analysis of each document;
  - A user interface which provides on-line terminology definitions and usage information to the author;
  - A controlled terminology inventory (now ca. 70,000 terms);
  - Controlled use of terminology (interactive disambiguation), for both English consistency / accuracy and also for more accurate machine translation;
  - Controlled sentence structures which nevertheless support the required grammatical complexity for writing in the given domain.

When integrated as part of CAT's overall document authoring process, the CTE system was designed to meet the following requirements:

- The CTE system should exploit the modular creation of documents as a collection of reusable "information elements" (IEs).
- The CTE system should promote consistency and provide a standard "look and feel" among manuals. Consistency should be supported, even when new documents are assembled from existing IEs taken from different sources, and despite varying expertise / experience levels among authoring staff.

- The CTE system should support integrated use of SGML tagging, for both paper publishing and electronic delivery of multilingual documentation. SGML markup is used heavily within sentences, so the CTE grammar must recognize markup tags in their context, interpret markup variably according to function, and arrange the markup grammatically in the target language output.
- The CTE system should interface smoothly with the other Authoring tools. IEs are managed via a complex electronic file management system (FMS). Each IE may contain several embedded references IE to reusable external objects (e.g., titles, warnings, graphics, tables, etc.).
- The CTE system must handle the large volume and variety of technical documentation produced. This includes documentation for 350 current products, as well as documentation for old products still in use. Documentation must be available to support products for the life of the product; some products are still in use since the 1930's. Each product is associated with multiple document types, each with its own writing style and standard.

4. Effort to Develop and Implement CTE

The CTE development effort was launched in November 1991. The CTE definition, authoring software, and machine translation (AMT) system were developed in parallel. Main categories of effort included CTE development, maintenance, and training, along with development of required document type specifications and the authoring process itself.

4.1. CTE Development

The personnel required at Caterpillar for CTE development, pilot, and training (from 1992-1997) averaged about 5 full-time equivalent employees per year. Required personnel included linguists, pilot authors, trainers and mentors.

CTE development included the following main activities:

- **Terminology Development**
  - *Extract terms via corpus analysis.* A 50-megabyte corpus of existing documentation was analyzed with phrase-finding programs written at CMU. The result was a rough list of about 120,000 candidate technical terms. Since the part-of-speech tagging used by the phrase-finding program utilized general, error-prone sources (e.g., the Brown Corpus), it was assumed that the rough terminology list would contain many candidates that were not valid for the CAT domain.

- *Screen candidate terms.* The candidate term list was reviewed by CAT terminology experts to eliminate terms that were not valid for CTE, reducing the initial set of terms down to about 65,000 CTE technical terms.
- *Devise writing guidelines.* The CMU and CAT teams worked together to produce author guidelines for the proper use of CTE terminology.
- *Identify ambiguous terminology.* CAT identified which of the terms in the CTE vocabulary were likely to have more than one meaning per part-of-speech (e.g., "valve" has 7 separate meanings in CTE depending on the type of valve).
- *Assign approved domain meanings.* For terms which had some domain-approved meanings as well as out-of-domain meanings, the domain-approved meanings were selected and out-of-domain meanings were marked in the lexicon for use in on-line vocabulary checking.
- *Write usage examples for authoring interface.* For each CTE term (and its out-of-domain meanings, if present), how-to and how-not-to usage examples were written for use in the on-line Authoring system.

- **Grammar Development**.
  - *Define CTE grammar rules*. CMU developed a set of grammar rules and writing guidelines (usage examples) for CTE. These grammar rules define a subset of English that is complex enough to permit efficient technical authoring, while limiting the set of available linguistic constructions to promote consistency and accurate translation[3].
  - *Define SGML DTDs for CTE*. CAT devised an SGML Document Type Definition (DTD) for each document type, to define the use of sentence-internal and document-structure markup.
  - *Implement CTE grammar checker*. The grammar rules and SGML DTD were implemented in an application of the KANT Analyzer (Mitamura, et al., 1991) for use in a grammar-checking application.
  - *CTE Pilot Test and Refinements*. The grammar checker was integrated and tested with the prototype Authoring system in an Authoring pilot phase, in order to resolve outstanding DTD, grammar, and terminology issues. As a result of the pilot phase, the grammar was refined and the usage examples in the authoring interface were improved.

4.2. CTE Maintenance

Support for the ongoing use of CTE in day-to-day production required the following activities:

[3] More details regarding design issues and examples are discussed in (Mitamura and Nyberg, 1995; Nyberg and Mitamura, 1996).

- *Problem Reporting*. CAT developed in-house problem reporting software and a process for author requests for new terms and grammar updates. CGI and CMU provided a problem report database, built using the GNATS software (Osier, 1993), which is used for problem reporting, tracking, and resolution.

- *Terminology Screening and Update*. CAT developed in-house linguistic expertise to do terminology screening prior to requesting CTE vocabulary additions. Since vocabulary requests come from variety of sources (authors) with different levels of expertise, this step is essential to avoid the addition of redundant and/or unnecessary terms. CAT also developed in-house expertise for CTE lexical maintenance to speed vocabulary updates.

- *Software Maintenance*. The maintenance of the CTE Language Editor software (originally developed by CGI) was brought in-house at CAT during 1997.

- *Process Monitoring and Quality Control*. CAT developed software and processes for electronic (on-line) review of completed CTE documents by CTE editors. This step is seen as essential, both as part of the mentoring process for new authors and as part of a broader effort to maintain the integrity of CTE writing standards over time.

4.3. CTE Training

In order to prepare authors for CTE training, CAT presented bi-monthly 'Brown Bag' lunch-time seminars, for one year in advance of CTE training (1994). Beginning in 1995, CAT prepared and administered CTE training to authors. Periodic updates to training materials are distributed at user communicator meetings. CAT also distributes regular editions of "CTE Author", an internal publication which documents updates to CTE and various CTE writing tips.

4.4. Authoring Process Development

To support fully-integrated SGML authoring, CAT needed to clarify the number, purpose, and content of the required set of document types. A dozen document types were identified in all. For each document type, CAT determined the internal document structure, and developed and codified the SGML markup (DTD) for English and 15 other languages. CAT also developed general stylistic guidelines for authoring.

To manage the creation, update, checking and translation of CTE IEs, CAT developed an on-line File Management System (FMS). The FMS currently includes about 60 integrated pieces of software. A workflow management tool was also created to provide support for the document production chain, which involves passing tasks (IEs) through a series of different steps, all of which are undertaken by different members of the authoring / translations staff.

5. Benefits from CTE

The benefits realized from CTE authoring include:

- Increased consistency of English writing and terminology because of CTE.
- Increasing ability to reuse Information Elements. Consistent writing and terminology use allows IEs to be reused across product lines, leading to increases in production efficiency for technical manuals.
- Heightened awareness of language-related issues at Caterpillar. The requirement to write according to a standard which would be checked sentence-by-sentence has brought renewed attention to a set of issues which are essential for high quality documentation:
  - The value of writing guidelines and terminology management;
  - What it takes to standardize the authoring process;
  - Development of authoring standards;
  - The level of system support required for authoring;
  - The amount of training required for effective high-quality authoring;
  - The high level of personnel/skills required (authors and translators, as well as terminology experts, lexicographers, and system maintainers).

6. What We've Learned

After developing the initial CTE system and testing it in pilot and production use, we made a set of refinements to the initial design in order to achieve a better fit with CAT's requirements (as observed through real use):

- *Domain Ambiguity*. Our initial conception of CTE was that each (root, part-of-speech) lexical entry would have only one possible semantic interpretation in the domain. The goal was to eliminate the need for interactive disambiguation of ambiguous terminology. However, a complete analysis of the domain determined that were many (1000+) terms in the domain which require more than one semantic interpretation, depending on context (e.g., ("charge", V) can be interpreted either as charging an electrical element such as a battery, or pressurizing a gas container such as an ether cylinder). Support for these multiple meanings is essential for accurate machine translation, so we extended the design of the grammar checker to include an interactive disambiguation module. When ambiguous terms are identified, the system asks the author to specify the intended meaning, which is then preserved in the input text using an unobtrusive SGML marker (Mitamura and Nyberg, 1995).

- *Incremental IE Updates*. When reusing one IE in several different documents, it is often necessary to make only minor incremental revisions. However, the original CTE system required that the entire IE be re-checked sentence-by-sentence if a change was made to a previously-checked IE. To avoid re-CTEing the whole IE when an incremental change is made, a region marking mechanism was added to the LE software to mark areas of text that are CTE-approved. For rapid IE re-use, the region marking mechanism was coupled with a Batch Checker application, which runs on modified IEs and returns a list of sentences which do not pass CTE or require disambiguation.

- *Human Factors during Disambiguation*. For highly-ambiguous terms, authors were asked to select from a large number of potential meanings during interactive disambiguation. This proved to be annoying when many of the potential meanings were not considered relevant in the given context (e.g., asking if “charge” means charging an ether cylinder, when the author is working on an electrical document). To relieve this burden on the author, we developed a set of subdomain codes, assigned to each IE by the author, which indicate the general topic area. These codes are used by the CTE system to mark the different ambiguous meanings for highly-ambiguous terms; when the author indicates subdomain information for an IE, only those meanings relevant to the subdomain are used.

- *Terminology Maintenance*.

  Keeping the CTE dictionary up to date with required terminology is a challenge. For some new products, there is very little time between the final engineering design and the time to “first ship” of the product. It is not always possible to add new terminology to the CTE system quickly enough so that the authors use only approved terminology in new documents. This leads to situations where documents are used for publication without CTE approval. Once authors are given the opportunity to bypass CTE approval when some terms are missing, they are tempted to bypass CTE approval completely for any IE containing a missing term. To increase CTE compliance, CAT introduced a means of automating the process of requesting a missing term, while temporarily "shielding" that term from CTE analysis, thus allowing approval of the rest of the sentence (and the rest of the IE) despite the presence of a missing term.

7. Challenges

The set of challenges we faced during CTE development are probably common to any large-scale implementation of controlled technical authoring, and include the following issues:

- CTE and Authoring (including SGML markup) were developed simultaneously, which occasioned some revision of early work in both parts of system when they were finally integrated.

- CTE terminology maintenance is an ongoing task, which includes control of terminology proliferation, removal of redundant terms, and screening of new terms requested by authors.
- Maintenance of usage examples is required. Every time the CTE grammar is improved in any way, the existing usage examples must be re-validated to ensure that they are still proper CTE.
- The CTE domain is too complex for lexicographers to anticipate all the ways authors use words; hence ambiguous phenomena cannot be defined in advance, and the lexicon and grammar must be extended through successive refinements after initial deployment of the system.
- The requirement for accurate machine translation is a driving force in representing semantic ambiguity during CTE terminology development. Terms that don't appear to be ambiguous during superficial review turn out to have several context-specific translations in different target languages, prompting a finer-grained (ambiguous) representation in the CTE lexicon for some terms.
- Adherence to CTE principles by authors is variable and sometimes difficult to enforce. Authors may use words in senses that are not approved, and sometimes authors select the wrong meaning choices for words during interactive disambiguation (both phenomena tend to degrade the AMT output). It is possible for authors to write CTE-approved sentences which are syntactically correct, but semantically incorrect or incomprehensible.
- Qualified people to do terminology work are difficult to find. The CTE terminology expert needs a combination of detailed domain knowledge, linguistic training, and some knowledge of the target language issues.

8. Plans for the Future

In order to address these challenges for further 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++ (KANTOO will be available 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:

- Continue to expand use of subdomain-specific ambiguity reduction techniques.
- Improve tools and processes to facilitate connections between types of terminology (engineering, parts-books, marketing).
- Support maintenance of the CTE terminology and Language Environment software in-house at Caterpillar;
- Continue to 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 CTE can have a significant positive impact on both authoring quality and translations productivity. Nevertheless, many challenges remain in an environment with a complex set of products and document types, and where terminology is updated constantly. The CTE application for Caterpillar has helped to advance the state of the art in controlled language systems, while simultaneously driving the research agenda for future work on new applications at CMU.

9. References

Mitamura, T. and E. Nyberg (1995). “Controlled English for Knowledge-Based MT: Experience with the KANT System”, " *Proceedings of TMI-95*.

Mitamura, T., E. Nyberg and J. Carbonell (1991). “An Efficient Interlingua Translation System for Multi-lingual Document Production,” *Proceedings of Machine Translation Summit III*, Washington, DC, July 2-4.

Nyberg, E., T. Mitamura and C. Kamprath (1998). “The KANT Translation System: From R&D to Large-Scale Deployment”, *LISA Newsletter*, Vol. 2:1, March.

Nyberg, E. and T. Mitamura (1996). “Controlled Language and Knowledge-Based Machine Translation: Principles and Practice”, *Proceedings of the First First International Workshop on Controlled Language Applications*.

Osier, J. (1993). *Keeping Track: Managing Messages with GNATS*, Cygnus Support, Version 3.2, October.

Note: For more information about the KANT system, visit the KANT project home page on the World-Wide Web:

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