[From: *CLAW '96*, Universiteit Leuven]

Controlled English Advantages

for Translated and Original English Documents

Phil Hayes

Steve Maxwell

Linda Schmandt

Carnegie Group Inc

5 PPG Place

Pittsburgh, PA 15222, USA

(hayes@cgi.com, maxwell@cgi.com, schmandt@cgi.com)

Abstract

The Controlled English we have developed for Caterpillar is designed to reduce translation costs by facilitating machine translation. It is also preferred by English-speaking readers of the resulting documents. The approach to machine translation and the resulting design of the Controlled English were optimized to Caterpillar's environment and requirements, including a very high document and translation volume. The basic framework of the Caterpillar solution can be adapted to other, more typical, sets of documentation requirements. We describe an adaptation we are making for Diebold, based on machine-assisted human translation.

1. Introduction

There is considerable current interest in controlling the English (or other language) used to author technical documentation. See Adriaens and Schreurs (1992) or Huijsen's Controlled Languages Homepage (http://wwwots.let.ruu.nk/Controlled-languages) for relevant references. The kinds of control usually considered include restricting the vocabulary used in a document, restricting the allowable meanings of particular words or phrases, restricting the kinds of syntactic constructions that may be used, and restricting the overall complexity of sentences. A collection of restrictions of these kinds is said to define a Controlled English (CE). Particular CEs have been developed for writing particular kinds of documents, typically specific to a particular company or industry. For instance, we have been involved in developing a CE called Caterpillar Technical English (CTE) that is appropriate for writing operation, maintenance, and service manuals for Caterpillar heavy equipment. See Pym (1990) or AECMA (1989) for other examples.

The current interest in CE stems from a growing consensus that CE documents can be translated more easily and hence at a lower cost and faster than non-controlled documents. Moreover, CE documents may also have readability advantages over non-controlled documents in the original

English. These readability advantages are harder to quantify and hence use as economic justification for CE, but we believe they are nevertheless real and significant. CEs are sometimes designed to improve English readability of documents that are not intended to be translated, but rather read by non-native speakers. Simplified English (AECMA, 1989), a CE developed by aircraft manufacturers for maintenance documentation, is an example. CEs developed for this purpose tend to have smaller vocabularies than CEs developed to reduce translation costs, but are otherwise similar.

This paper describes some of our experiences in developing CTE for Caterpillar and the kinds of advantages that it provides. We go on to discuss the use of controlled language to gain similar advantages in other settings, including some work we are doing for Diebold Inc.. Caterpillar's documentation requirements are unusual in several ways, including the high volume of documents that they produce, the large number of languages required for translation, the centrality and homogeneity of its technical writing staff, and the strong commitment Caterpillar has made to a high degree of translation automation. The other settings for CE that we discuss may fit better with the needs of a much broader range of technical documentation organizations.

2. Caterpillar Technical English

Caterpillar Technical English (CTE) was developed in the context of a major effort by Caterpillar Inc. to improve and modernize its delivery of service information. Caterpillar's re-engineering effort in this area is collectively known as the Service Information System (SIS). Elements of SIS include:

- SGML-tagged source documents
- on-line as well as hardcopy delivery of all documentation
- automatic formatting of both hardcopy and on-line versions of target documents from a single source
- using Information Elements (IEs) as the unit of writing. IEs are small coherent standalone units of information, for example, the instructions for changing fuses or for cleaning and replacing an oil filter. Writing in them provides for maximum reuse of writing.
- greatly expanded translation of documentation in up to 35 languages.

Given the large volume of Caterpillar documentation (over 100,000 new pages each year), the last of these bullets implied the need for heavy automation of the translation process. To meet its needs, Caterpillar selected some new translation technology, called KANT (Mitamura et al., 1991), from the Center for Machine Translation (CMT) of Carnegie-Mellon University. KANT is designed to produce very high accuracy translation, but requires that its input language be strictly controlled for both vocabulary and grammar. The degree of control required and the level of detail involved dictates the need for an interactive checker to help Caterpillar authors write within the controls. The KANT approach also requires the compilation and encoding of a large amount of language and domain-specific knowledge (Mitamura et al., 1993). KANT is designed to leverage this knowledge to resolve the kinds of ambiguities that cause problems for MT and hence produce highly accurate translations.

Caterpillar engaged Carnegie Group and CMT to develop and deploy a combined authoring and translation system based on the CMT KANT technology. The system contains the following major components:

- **Caterpillar Technical English:** a Controlled English designed to express Caterpillar's service information in a way that meets the requirements of the translation technology for accurate translation. CTE includes the following components:
  - several thousand individual words, both general and technical, most of which are restricted to one meaning.
  - several tens of thousands of technical phrases, with only one unambiguous meaning each.
  - a collection of syntactic rules designed to allow freedom of expression while minimizing problems for machine translation. Restrictions include limitations on conjunction use, elimination of pronouns, limitations on subordinate clauses, and several others. The net effect is to promote simple, direct language that is very appropriate for technical writing.

- **An interactive checker** for CTE, called ClearCheck that tells Caterpillar authors whether what they write conforms to CTE and helps them make it conform if it does not. See Adriaens (1995) and Hoard et al. (1992) for other examples of CE checkers. ClearCheck can also identify constructions that would be ambiguous to the machine translation system and then ask the author to choose between the alternative meanings that the translation system would find. For example, the title "Clean Filters" could be a noun phrase or an imperative sentence. To do this accurately, ClearCheck uses the analysis component of the machine translation system described below to do its grammar checking. Any sentence that is passed by the checker is thus guaranteed parseable by the MT system. ClearCheck is implemented as an addition to the Arbortext Adept editor, which has been adopted by Caterpillar as its authoring environment for SIS. Authors can perform checking on their documents in SGML source form and make appropriate corrections without leaving the Arbortext environment. ClearCheck is currently in production use within the Caterpillar environment.
- **A machine translation system,** called AMT, based on the KANT technology. AMT translates English SGML source documents into SGML source documents in other languages which can then be automatically formatted for hardcopy and online publication. This avoids the expensive reformatting effort often associated with document translation. AMT is currently under development for six languages: French, Spanish, German, Italian, Portugese, and Russian. French is currently undergoing field testing at Caterpillar and is expected to be in production by mid year.

Although we must wait until full deployment for a final answer, the current experience at Caterpillar suggests that CTE will indeed serve its intended function in substantially reducing translation costs by permitting the use of the AMT translation system. However, to be workable, CTE also needed to meet two other goals:

- not have a major negative impact on author productivity
- be acceptable to English readers

There is a significant productivity hit involved in the production use of CTE and ClearCheck by Caterpillar authors. However, given the benefits of improved grammar checking, spell checking and consistent use of vocabulary, Caterpillar believes that the impact remains within acceptable bounds for the authoring group.

To answer the question of acceptability to English readers, CGI and Caterpillar conducted some focus testing with two American Caterpillar dealers. The testing had the dealers compare older non-CTE documents with versions rewritten into CTE. The results were strongly positive, suggesting that the dealers in fact preferred the CTE versions of the documents. Examples of points made by the dealers were:

- The relatively simpler sentences in the CTE versions were easier to understand than the original. For example, the focus groups preferred the CTE

  > "This category indicates that an alternator is malfunctioning. If the indicator comes on, drive the machine to a convenient stopping place. Investigate the cause and determine the solution."

  to

  > "If this indicator, which indicates that the alternator is malfunctioning, comes on, drive the machine to a convenient stopping place and investigate the cause to determine the solution."

- Bulletted lists are easier to understand than lists embedded in long sentences.
- The lack of pronouns made it easier to skim for information, which is the primary way they use the documents. In that mode of use, having each sentence self-contained in meaning is a significant advantage.

This experience with the dealers underscores the importance for both technical writers and translators of writing with the target audience in mind. What is the "best" English (or French, Spanish, etc.) by some abstract standards of elegance or style may not be the best for a specific purpose, such as reference use.

3. Applicability of the Caterpillar Model

Since the use of CE at Caterpillar has been successful, we have naturally considered how the Caterpillar model can be applied to other technical documentation groups. We have found that there are several characteristics of the Caterpillar environment that make direct transfer difficult:

- Caterpillar documentation volume is unusually high. Caterpillar produces over 100,000 pages of new documentation every year and wishes to translate it into many different languages. While this is not in itself an impediment to transfer of the model, it is a major factor in determining economic feasibility. The development of CTE and the corresponding knowledge required to drive the AMT translation system involved large upfront costs. It is hard to.justify these costs at volume levels significantly lower than Caterpillar's.
- Caterpillar authors are unusually homogeneous. All Caterpillar technical authoring is done in one building in East Peoria, Illinois, by one department of approximately 150 authors, all using identical hardware and software tools. Such uniformity is the exception, rather than the rule in technical writing groups. A more typical situation is multiple locations, multiple management, and multiple hardware and software environments. Such diversity makes it much harder to agree on and then get leverage out of a CE approach.
- Caterpillar has a corporate commitment to re-engineer its process for the creation and distribution of service information. This allowed it to optimize the overall process with corporate goals in mind, including putting an additional load on the authors by requiring

them to conform to CTE and balancing it against the potential payback of getting a much greater reduction of effort during translation. More frequently, the cost of translation is borne by a component of a company (frequently marketing) that is organizationally distant from the technical publications group. This means that there is no incentive for the tech pubs group to increase its costs by introducing CE because it does not pay the translations costs and hence would receive no benefit from their reduction. Moreover, the place where the decision would have to be made to balance the benefit against the cost is typically too high in the company for the possibility of improvement to receive the necessary degree of notice.

These factors combine to make the Caterpillar situation unique and mean that the combined CTE/AMT model we have developed is probably not directly applicable elsewhere. However, we believe that elements of the model are much more widely applicable and can provide significant and practically attainable value to a wide variety of corporations.

4. Adapting the Caterpillar Model to other Companies

We have explored the deployment of CE with several other companies besides Caterpillar. Given our successful experience at Caterpillar, we naturally started with the Caterpillar model and tried to see how it would fit the needs of each of the other companies. Two main themes emerged during these explorations:

- The amount of effort required to develop and deploy a CE system had to be much lower than for Caterpillar. None of the other companies came even close to matching the documentation and translation volume of Caterpillar, and hence had a much lower potential payback in the form of translation costs to reduce. This meant that cost of development had to be much less. It also meant that the time required to develop had to be less than the several years involved for Caterpillar. Longer time frames made the return on investment unattractive because it took too long before the benefits of lower translation costs were obtained.
- The CE needed to be looser than CTE or the consequences for not adhering to it needed to be less severe than for Caterpillar. None of the other companies had a centralized authoring group like Caterpillar's. Authors were distributed organizationally and/or geographically. This made it very hard for the companies to sign up to a requirement for all authors to adhere to a single tight CE standard. In the Caterpillar model, failure to adhere to CTE is likely to result in a missing or incorrect machine translation for the offending sentence, thus eliminating the payback. Since they could not guarantee full compliance to whatever CE was developed, the other companies needed there to be partial benefits for partial compliance to CE.

These two themes were universal. A third theme that occurred with some of the companies we looked at concerned the uniformity of document representation. A surprisingly large number of the companies we worked with had made a corporate commitment to SGML as a document representation, thus following the Caterpillar model in this area. Most of the others had made a corporate commitment to some desktop publishing (DTP) system such as Word, WordPerfect, Interleaf, or Frame. A few did not have a corporate standard and did not expect to establish one. Multiple DTP standards make it very difficult to provide an interactive checker following the Caterpillar model because of all the different pieces of software that would need to be interfaced with. On the other hand a single DTP standard would only require the development of a checker

that interfaced with the DTP system analogously to the way in which the Caterpillar checker interfaces with Arbortext Adept SGML editor.

There is more than one way to adapt the Caterpillar model to meet the requirements we found at these other companies. In the remainder of this section, we describe one specific adaptation that we have devised for Diebold Inc. and are now in the process of piloting for them. We believe that this adaptation could provide value to a large number of other companies at an economically justifiable cost and hope to use it in the future to do so.

The solution we are piloting for Diebold follows the essence of the Caterpillar model. It is based on a Controlled English. It includes an interactive checker to help authors conform to that CE. The use of CE reduces the cost and time of translation. Following are the adaptations we have made to the Caterpillar implementation of that model.

- **Translation Method:** Translation is performed by human translators assisted by translation tools, including online glossaries and translation memory, rather than by machine translation and postediting. This is a radical change to the Caterpillar model, but one that allows us to make the CE considerably simpler and more flexible. We can do this because human translators can deal with much greater variation in the CE than can the AMT system used for Caterpillar. Humans are able to use their deeper understanding of document context and subject matter to resolve potential ambiguities that would be well beyond AMT. It also allows us to avoid compiling the detailed language and domain knowledge that is necessary for AMT, and which was a very large part of the upfront cost for Caterpillar. The corresponding downside is that the potential for savings is less than in the Caterpillar model. Nevertheless, we believe that use of CE with machine-assisted human translation in this solution will reduce translation costs by around 25%.
- **Vocabulary:** The vocabulary framework is very similar to Caterpillar's. The primary difference is that it is not fully comprehensive, allowing the author more flexibility in new terms and in example uses of the product concerned in a wide and inherently unpredictable range of areas. Specifically, it contains:

    - *technical terms for objects:* typically noun phrases with domain-specific meanings, such as "access control" or "personnel profile", but including some single nouns with restricted technical meanings. For example, "button" is used only with the technical meaning of 'a mouse-sensitive object on a graphical user interface', and "reader" may be used only to refer to a device that reads electronically encoded identification cards. This group forms the bulk of the vocabulary and is as comprehensive as possible.
    - *other technical terms:* typically verbs and adjectives used for actions and properties. There is frequently a greater potential for ambiguity in these terms than those describing technical objects. We try to choose the terms in the CE to minimize the potential for such ambiguity, and where possible select just one of the possible meanings as approved. We also provide a definition and usage examples for each of the terms. For example, the verb "maintain" may be defined only to mean 'to keep in good condition through regular attention', as in the sentence "This chapter presents instructions for maintaining the IBM 437X.". To indicate that "maintain" should not be used with the meaning 'to continue', we would include a usage example noting that the sentence "The machine maintains the alarm during the programmed elapsed time." should be rewritten using "keeps" instead of "maintains".

- *more general terms:* typically verbs used in more general senses, such as "provide", "increase", and "decide". We handle ambiguity similarly to the other terms.
- *excluded terms:* these terms are not actually in the CE, but might well have been. They are terms that relate to the technical domain and which authors have used in earlier, non-CE documentation. We define synonyms for them from among the terms that are included. For example, non-CE documentation for a Diebold product sometimes refers to a part by the engineering part name, and sometimes by a more common name used by customers. During CE development writers agree to standardize on one of the two names.

All the terms, both included and excluded, are represented in a relational database for easy maintenance and publication. This database will be used to generate the English side of the online glossaries that will be used by the human translators. The definitions and usage examples available to the authors will also be available to the translators through this database.

- **Grammar:** This is much less comprehensive than in the Caterpillar model. In at least the initial implementation, it will be based on gross measures of sentence complexity, including sentence length, and numbers of commas, prepositions, and conjunctions, supplemented by restrictions on some locally checkable grammatical phenomena, such as passives and *-ing* verbs. The basic philosophy is that if an author can avoid complex sentences and particularly tricky constructions, then human translators will be well able to deal with whatever else is used.
- **Interactive Checker:** This component is similar in concept to the one that was used with Caterpillar. As with Caterpillar, on request of the author it makes a pass through the current document, indicating CE issues for the author to attend to. More specifically:

  - *Vocabulary clicking:* unlike Caterpillar, it does not have a complete vocabulary that the author must stay within. So it indicates which vocabulary items are approved and which are explicitly excluded from the CE, suggesting alternatives for those explicitly excluded. It also identifies terms that are ambiguous or have the potential for ambiguity even though only one meaning is approved. For example, an author who writes "...a message appears at the bottom of the screen" may be shown a message indicating that "appear" will be interpreted to mean that the message comes into view at that moment; if the writer intended only "There is a message at the bottom of the screen.", the sentence should be rewritten. It is left up to the author to decide whether unknown items are legitimate extensions of the CE or should be replaced by terms in the CE. Legitimate extensions include non-domain terms included to illustrate real-world examples as well as function words such as prepositions, articles, and conjunctions.
  - *Grammatical Checking:* as with Caterpillar, the grammatical component of the checker identifies sentences with grammatical issues. The primary difference is that the grammatical checks do not involve a full parse of each sentence. We rely on the author to determine whether the sentence is grammatical English and just look for the kinds of complexity that are hard to translate. For example, the following sentence would be flagged:

"The objectives of implementing this system include creating a cost-effective process to generate cards that requires minimal clerical effort, which includes encoding the magnetic stripe and entering data into a centralized system."

- *System Environment:* this is similar to Caterpillar in that the text is represented in SGML and the checker is designed to work appropriately with the SGML tags in place. In addition, Diebold has chosen Arbortext Adept as its SGML editor, so like Caterpillar, the checker is implemented as an extension to Adept. However, the Diebold checker is implemented on a PC, whereas Caterpillar operates in a Unix environment.

When the system is complete, we believe the next effect of these changes from the Caterpillar model will be a controlled English authoring system that:

- will reduce translations cost by around 25%.
- will be comfortable for authors to write in. Specifically, it will provide sufficient support to allow a willing, motivated author to conform to the CE, while still allowing ad hoc extensions as the need arises.
- will be straightforward for Diebold to maintain and extend.

The above scenario represents one adaptation of the Caterpillar model. Others are possible. In particular, another promising adaptation would be to use postedited machine translation by one of the commercial translation systems in place of human translation. The vocabulary database would provide the English side of a local dictionary for the MT system. We believe that this could yield even greater savings in overall translation costs, particularly if coupled with translation memory tools, although the initial set up costs would be somewhat greater.

5. Conclusion

The combined CE authoring and translation system that we have developed for Caterpillar will, when fully deployed, allow Caterpillar to translate large volumes of technical information at a much lower cost than by manual methods. The approach taken involves a large amount of upfront work to compile language and domain knowledge and is difficult to cost justify for smaller document and translation volumes. We have described an adaptation of the approach that we are in the process of implementing for Diebold. This adaptation involves much lower costs up front and places less demand on the authors. It does not deliver as great a degree of savings as the Caterpillar model, but on balance is much easier to cost justify at Diebold's level of document and translation volume. This level is much more typical than Caterpillar's exceptionally high volume.

In the case of Caterpillar, we found that English-speaking users of the documents actually preferred the CE form over the original uncontrolled English documents. They found them more readable and easier to use as reference materials. We have not yet gathered any corresponding user feedback for Diebold documents, but have no reason to assume that we will not get the same kind of favorable response as occurred for Caterpillar.

Based on our work with Caterpillar and Diebold, we believe that CE approaches can help control translation costs, while enhancing readability and will enjoy increasingly widespread use and acceptance. We conclude with a list of issues to be resolved in achieving wider deployment of CE, including:

- **Standardization:** Currently CE development is highly customized. How can it be made less so? What can be shared at the grammatical level (probably a great deal) and at the lexical level (probably much less)?
- **Maintenance:** The technical subject matter of most CEs is in a constant state of flux with new products and components being introduced continually. A CE, particularly, the lexical part must be kept up to date with that change without invalidating the meaning choices or violating consistency of word use. Effective methodologies are required for this maintenance activity.
- **Degree of control:** Greater control can reduce ambiguity and hence make human or machine translation faster and cheaper. However, greater control is more expensive to set up and is less likely to be complied with by some authoring groups. Every CE requires a balance between these two competing forces. The answers may be different for different situations.

The benefits of successfully resolving these issues and thereby producing widespread deployment of CE will be large. Carnegie Group is committed to working these issues in the context of particular CE deployments and expects to make significant contributions to their resolution.

References

Adriaens, G. Simplified English Grammar and Style Corrector in an MT Framework: The LRE SECC Project. In *Proceedings of the 16th Conference on Translating and the Computer,* pp. 78-88, 1995.

Adriaens, G. and Schreurs, D. From COGRAM to ALCOGRAM: Toward a Controlled English Grammar Checker. In *Proceedings of the 15th Int. Conf. on Computational Linguistics (COLING '92),* pp. 595-601, 1992.

AECMA. A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language. Change 5, AECMA, Paris, 1989.

Hoard, J. E., Wojcik, R. H., and Holzhauser, K. C. An Automated Grammar and Style Checker for Writers of Simplified English. In O'Brian, P. and Williams, N. (eds.), *Computers and Writing: State of the Art.* pp. 278-296, Intellect Books, Oxford, 1992.

Mitamura, T., Nyberg, E. H., and Carbonell, J. G. An Efficient Interlingua Translation System for Multi-lingual Document Production. In *Proceedings of the Third Machine Translation Summit.* 1992.

Mitamura, T., Nyberg, E. H., and Carbonell, J. G. Automated Corpus Analysis and the Acquisition of Large, Multi-Lingual Knowledge Bases for MT. In *Proceedings of TMI-93,* 1993.

Pym, P. J. Pre-Editing and the Use of Simplified Writing for MT: An Engineer's Experience of Operating an MT System. In Mayorcas, P. (ed.) *Translating and the Computer 10: The Translation Environment 10 years On.* pp. 80-95, Aslib, 1990.