Controlled English for Knowledge-Based MT: Experience with the KANT System

Teruko Mitamura
Eric H. Nyberg, 3rd
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213
teruko@cs.cmu.edu ehn@cs.cmu.edu

Abstract

In this paper, we describe the design and deployment of KANT Controlled English (KCE) for knowledge-based machine translation in the KANT system. KCE combines three kinds of constraints: constraints on the lexicon; constraints on the complexity of sentences; and the use of generalized markup language. We describe how each of these types of language control are utilized in the implementation of a typical KANT application. The principles described are not specific to knowledge-based MT, and can be applied in the design of controlled languages for any kind of MT application.

1 Introduction

In recent years, more emphasis has been given to the usefulness of controlled source languages in MT (cf. Adriaens and Schreurs, 1992 and the references cited there). In this paper, we focus on the use of controlled input languages in the KANT translation system (Mitamura, et al., 1991). Controlled English is used to improve the clarity of expression in the source text and to improve the quality of the MT analysis phase. We describe the design and deployment of KANT Controlled English (KCE) for knowledge-based machine translation in the KANT System. KANT has been applied to the domains of electric utility management (ESTRATO) and heavy equipment manuals (CATALYST). KCE combines three kinds of constraints:

- **Constraints on the Lexicon.** In order to reduce lexical ambiguity and complexity, constraints are placed on the source vocabulary.
- **Constraints on the Complexity of Sentences.** To limit parsing complexity during source analysis, the types of input sentences are limited to those necessary for concise technical authoring.

- **Use of Generalized Markup Language.** The use of generalized markup language (SGML) supports definition of complicated domain terminology and phrasal constructions without increasing the ambiguity of the analysis grammar.

Once the language has been defined and the data files constructed, the language may be embedded into a system for on-line document authoring which supports these activities:

- **Vocabulary Checking.** The input text is checked to ensure that it conforms to constraints on vocabulary; otherwise, the system helps the author to select alternative vocabulary.

- **Grammar Checking.** The input text is checked to ensure that it conforms to constraints on grammar; otherwise, the system prompts the author to re-write his sentence.

- **Interactive Disambiguation.** If ambiguities arise during grammar checking, the system may ask the author to choose among competing analyses, encoding those choices for later use during translation.

In the remainder of this paper, we describe how each of these types of language control is carried out in a typical KANT application. Although applied to knowledge-based MT in KANT, the principles of KCE are not specific to knowledge-based MT, and can be applied in the design of controlled languages for any kind of MT application. Then we discuss some global issues for controlled language design.

2 Controlled Vocabulary

A key element in controlling a source language is to restrict the authoring of texts such that only a pre-defined vocabulary is utilized. In order to define a controlled vocabulary for a particular application domain, pre-existing documents are analyzed as an initial source of vocabulary. This initial vocabulary is further refined as the domain meanings of each term are encoded, and emerging lexical classes begin to collect domain-specific closed-class items. It is inevitable that each domain will contain a set of ambiguous terms (words for which the same root/POS pair has more than one semantic assignment), so we have also designed a method for disambiguation of lexical items in the input which supports interactive disambiguation by the author.

2.1 Corpus Analysis and Vocabulary Extraction

The first step in defining a domain vocabulary is to extract as many terms as possible from pre-existing on-line documentation. In the case of the CATALYST

```
((:ROOT "account for")
 (:POS V)
 (:CONCEPT *A-ACCOUNT-FOR)
 (:TYPE :PHRASE)
 (:SYN-FEATURES (PREFER-PHRASE +))
 (:CLASS AGENT+PATIENT)
 (:NOTE (:SENSE "to furnish a justifying analysis or explanation"
                "This dipstick is used to account for variations in
                 engine installations."
                :INPHRASE
                "account for"))
 (:COMMENT "sholm: 'account' never occurs alone, made CTE minus
           sholm: corrected root to 'account for'")
 (:ACTION  :NEW)
 (:FREQUENCY 16   19)
 (:UPDATED  (41  2   15 4  12  1992)   "sholm"))
```

Figure 1: **Example Lexicon Entry**

project (Mitamura, et al., 1993), about 50 megabytes of existing corpus were used to extract a domain vocabulary for heavy equipment documentation. The steps taken to construct a lexicon from the source corpus are as follows:

1. *Automatic Deformatting of the Existing Corpus.* The existing corpus is processed by a. set of programs which remove and/or canonicalize the formatting codes used in the source documents.

2. *Automatic Creation of a Word Corpus.* All occurrences of inflected forms are counted and merged into a corpus of word occurrences by a statistical program.

3. *Automatic Creation of a Sentence Corpus.* All of the sentences which appear in the corpus are indexed by the words that appear in them, in order to support further analysis, including KWIC (Key Word in Context) access to the corpus.

4. *Creation of the Initial Word and Phrase Lexicons.* In order to produce an initial Lexicon, a lexicon building program uses a pre-existing tagged corpus (e.g., the tagged Brown Corpus (Francis & Kuĉera, 1982)) as a resource for part-of-speech information, in conjunction with a source language morphological analyzer. The initial Lexicon contains a part of speech marker for each root form found in the Word Corpus. In order to produce an initial

Phrasal Lexicon, a phrase-finding program uses both the initial Lexicon and the Sentence Corpus as resources.

5. *Human Refinement of the Lexicons.* An example of a finished lexicon entry is shown in Figure 1. The `:ROOT, :POS, : CONCEPT, : SYN-FEATURES` and `: FREQUENCY` fields are created automatically with default values. Subsequently, the lexicographer browses occurrences of the word in the Sentence Corpus using a KWIC browser, and refines the default values and also adds a definition and examples to the `: SENSE` field. These are not intended for use by the system, but are provided as a resource for future human readers of the lexicon.

2.2 Domain Technical Vocabulary

There are three broad categories of technical vocabulary to be considered in defining a Controlled English:

- *Technical Phrases.* In a given domain, there are likely to be several phrases whose meaning is difficult to recover unless the phrase is stored in the lexicon as a single unit. Such phrases include noun phrases whose meaning cannot be derived compositionally, such as *oil pan* when we assume that the word *pan* has no separate domain meaning. Phrasal verb-particle constructions such as *abide by* (cf. Figure 1) are also easier to analyze if taken as a unit.

  It is also the case that large numbers of technical noun phrases which might be compositionally analyzed can be more efficient to analyze during parsing if they too are represented as single units of lexical meaning. In the case of the KANT application for Caterpillar, there are about 50,000 domain phrases encoded in the lexicon (Mitamura, et al., 1993).

- *Technical Words.* In a typical domain, there are many single symbols which have a special meaning in the domain and are not found in other kinds of text. For example, technical documentation generally contains symbols such as acronyms (e.g., *Programmable Electronic Engine Control (PEEC))* and abbreviations (e.g., *foot pounds (ft-lb)}.* A given domain may also require types of lexical items that are particular to that domain (for example, a class of words denoting wire colors, or a class of words denoting labels on machine controls). Each class of technical words must be identified and filled in, generally with participation from the customer's terminology experts.

- *Technical Symbols.* Any special use of numbers, numerals, units of measure, letters of the alphabet, etc. must be specified and encoded in the lexicon as well.

2.3 Encoding the Meanings of Vocabulary Items

One important feature of the KANT Controlled English is that is explicitly encodes a set of domain meanings for each term in the lexicon. In knowledge-based systems like KANT, this meaning is encoded as pointer to a domain concept frame, and is used to access the domain knowledge base during source text analysis. Even in systems that do not utilize semantic processing, encoding domain meanings during lexicon creation helps to identify potentially difficult terms for translation. When defining a controlled English for a new domain, these three steps are taken:

- *Limit Meaning Per Word/Part-of-Speech Pair.* Wherever possible, the lexicon should encode a single meaning (domain concept) for each word/part-of-speech pair. This helps dramatically to reduce the amount of ambiguity in the source text, which in turn reduces the complexity of source analysis by an appreciable amount (Baker, et al., 1994).

- *Encode Meanings Using Synonyms.* Whenever a lexical item has more than one potential meaning in the domain, first an attempt is made to "split up" the meanings by finding separate, synonymous terms to encode them. Terms which are "split" in this manner are subsequently marked in the lexicon, so that it is possible to determine for any given word whether it has an alternate meaning which is encoded by a different term in the domain. This information can be used in support of on-line vocabulary checking (cf. Section 5.1).

- *Encode Truly Ambiguous Terms for Interactive Disambiguation.* When a term simply must carry more than one meaning in the domain, either because of customer requirements or because there is no synonym available for the additional meanings, these meanings must be encoded in separate lexical entries for the same word/part-of-speech pair. If more than one such entry is activated for a given lexical item during source text analysis, then the resulting output structure will be ambiguous (there will be more than one meaning analyzed for the sentence). In this case, lexical disambiguation must be performed to further narrow the meaning to just the meaning intended by author (cf. Sections 2.5 and 5.3).

2.4 Types of Lexical Constraints

In addition to restricting the meaning of domain terms, the controlled English may also pose constraints in other areas of the vocabulary as well. Aspects of vocabulary which are commonly restricted in KANT applications include:

- *Orthography.* Whenever possible, the spelling, capitalization, hyphenation and use of slash in domain terms should be consistently specified.

- *Functional Words.* Rules concerning determiners, pronouns, reflexives, quantifiers, and conjunctions must be specified. Wherever possible, the use of pronouns and conjunctions should be limited, since they increase the potential ambiguity of syntactic analysis.

- *Modal Verbs.* The senses of modal verbs, modals, and their interactions with negation must be clearly specified and taught to the authors in order to increase accurate use of these words during authoring.

- *Participial Forms.* The use of participial forms (such as *-ing* and *-ed)* should be restricted. For example, *-ing* should not be used in subordinate constructions (e.g., *When starting the engine...);* structures like these should be re-written to include an explicit subject (e.g., *When you start the engine* ...). The *-ed* form should not be used to introduce a relative clause without explicit use of a relative pronoun; these reduced relative clauses (e.g., *the pumps mounted to the pump drive)* should be rewritten to explicitly use a relative pronoun (e.g., *the, pumps that are mounted to the pump drive).*

2.5 Resolving Lexical Ambiguity

The KANT Controlled English supports the use of special SGML tags to annotate words in the input text. These annotations capture a choice of meaning when a particular word is potentially ambiguous. For example, suppose the lexical item `("rip",V)` has two domain meanings:

```
(*A-RIP-1 "To create a gash or slit in a piece of fabric")
(*A-RIP-2 "To break a section of pavement into large chunks
using a ripper attachment")
```

A sentence containing *rip* may be annotated in order to indicate which meaning is desired, for example:

```
"Do not rip <means text='rip' val='*A-RIP-l'> in a downhill
direction."
```

When means tags are supported in the controlled English, then the analysis grammar can be written to take advantage of them, potentially reducing the number of syntactic analyses when it is possible to have the authors insert these tags interactively (cf. Section 5.3).

3 Controlled Grammar

When analyzing a corpus of technical documents, especially those associated with assembly, use and maintenance of machinery, one finds that the range of English constructions required for effective authoring is not large. It is often preferable to adopt a set of rules for technical writing which improve and standardize the readability of texts, even if the texts are not translated. If the grammatical constraints on the source text are formally specified and satisfied during authoring, then a machine translation system may take advantage of the less complex, less ambiguous texts which result, generally producing better-quality output.

There are two general types of grammar restrictions; those that place constraints on the formation of complex phrases in Controlled English, and those that place constraints on the structure of sentences.

3.1 Phrase-Level Constraints

- *Verb Particles.* English contains many verb-particle combinations, where a verb is combined with a preposition, adverb, or other part of speech. Particles which are part of phrasal verbs are often ambiguous with prepositions, and a controlled English should limit this ambiguity by recommending that verb-particle combinations be rewritten whenever possible. This can usually be accomplished by choosing a single-word verb instead (for example, *turn on* can be rewritten using *start).*

9 *Coordination of Verb Phrases.* Coordination of single Vs or VPs is not recommended for controlled English, since the arguments and modifiers of verbs conjoined in this manner may be ambiguous. These constructions are to be authored using conjunction of full sentences; for example, *Extend and retract the cylinders* is re-written as *Extend the cylinders and retract the cylinders.*

- *Conjoined Prepositional Phrases.* Authors are encouraged to repeat the preposition in conjoined constructions where appropriate. It is important to distinguish the scope in phrases like *5 cubic meters of concrete and sand,* which could mean either 5 cubic meters of mixture or 5 cubic meters of each material.

- *Using the Determiner in Noun Phrases.* In full sentences, the use of determiners in noun phrases is strongly recommended, since they make the referential nature of the noun they modify more precise. This in turn supports better quality translation.

- *Nominal Compounding.* In general, nominal compounding is not allowed unless it is licensed by domain rules which allow specific types of nominal compounding (e.g.. wire colors, component names/modifiers, etc.). This reduces the ambiguity that would result if arbitrary noun-noun compounding were allowed.

- *Quantifiers and Partitives.* These may not appear alone, and must modify a nominal head. For example, *Repeat these steps until none are left* can be more precisely written as *Repeat these steps until no bolts are left* when that is the intended meaning.

3.2 Sentence-Level Constraints

- *Coordinate Conjunction of Sentences.* In controlled English, it is recommended that the two parts of a conjoined sentence be of the same type. Sentence types should not be mixed in sentential conjunction, since a conjunction of different sentence types is difficult for a source analyzer to interpret. These constructions can be rewritten by choosing two sentences of the same type.

- *Clauses Introduced By Subordinate Conjunctions.* Both clauses in complex sentences using subordinate conjunctions must contain a subject and a verb; if the subordinate conjunction is removed, the subordinate clause should be able to stand alone as a simple sentence. Reduced clauses without subjects (e.g., *after installing the gear)* should be rewritten to include an explicit subject (e.g., *after you install the gear).*

- *Adjoined Elliptical Modifiers.* The use of ellipsis should be ruled out whenever possible in controlled English, since it introduces potential ambiguity in ellipsis resolution. However, some elliptical phrases (e.g., *if necessary, if equipped)* may be required. These should be explicitly specified as a closed class in controlled English, so that the source analyzer can treat them as special cases.

- *Relative Clauses.* Relative clauses can be added to independent clauses to form complex sentences. In controlled English, relative clauses should always be introduced by the relative pronouns *that* or *which.* Relative clauses contain a gapped argument which is coreferential with the element they modify. In unrestricted English, this gap can be in the subject position of the relative clause, or in the object position of the relative clause. A third type of relative clause is introduced by a "complex relative expression" such as *with which* or *for whom.* The gap can be said to be in the object position of a PP in this type of relative clause. KANT Controlled English applications typically support subject relative clauses, but not object or complex relative clauses.

- *WH-Questions.* A given controlled English application for technical documentation may or may not require support for WH-questions, depending on the domain. Whenever possible, the use of WH-questions is avoided, since deriving the long-distance dependencies between WH-words and their original, gapped position complicates syntactic analysis. Whenever possible, WH-questions should be rephrased as direct questions (for example, using *do* or *be).*

- *Punctuation.* The rules for consistent, unambiguous use of comma, colon, semicolon, quotation marks, and parentheses as inter- and intra-sentential punctuation should be clearly stated in the controlled English specification.

3.3 Resolving Structural Ambiguity

Occasionally, source sentences are truly ambiguous in the domain, in cases where there is more than one meaning for the sentence even when all the constraints of controlled grammar are met. The KANT Controlled English supports the insertion of SGML tags within sentences in order to indicate the desired choice among ambiguous structures. For example, in the sentence *Secure the gear with twelve rivets* the PP *with twelve rivets* could modify either *secure* or *gear.* The attach SGML tag can be used to indicate the desired attachment:

```
Secure the gear with <attach head='secure'  modi='with'>
twelve rivets.
```

Once the controlled English supports annotations of this type, it is possible to implement interactive disambiguation of the source text by the author (cf. Section 5.3).

4 Text Markup

In recent years, there has been much emphasis on the use of SGML and similar generalized markup languages for document production. KANT Controlled English supports the use of SGML tagging, and in doing so takes advantage of several positive features of SGML which reduce the complexity of source text analysis.

4.1 The Role of Markup in Controlled English

Use of SGML markup in controlled English text improves the quality of both the source and target text in the following ways:

- *Formalizing Document Structure.* A typical SGML implementation specifies tags to be used to mark paragraphs, lists of bulleted or enumerated items, titles and headings, tables, etc. When document context is tagged with SGML, it can be used as another source of information during analysis.

- *Limit Complexity of Analyzing Domain Vocabulary.* When SGML is used to identify items that fall into the same semantic class (e.g., part numbers, serial numbers, model names), these items need not be explicitly represented in the lexicon, allowing significant reduction in the size of the lexicon in a large technical domain with lots of component identifiers.

- *Reduce Lexical Ambiguity.* Symbols such as integers or alphanumerics, which might be ambiguous when untagged, are unambiguous when tagged.

- *Simplify Analysis of Domain-Specific Constructions.* When a technical domain requires that complicated sequences of numeric identifiers, modifiers, and component names be analyzed as noun phrases, the use of SGML tags can dramatically reduce the complexity of source analysis. Instead of allowing arbitrary composition of numbers and modifiers using unrestricted, recursive grammar rules, specific sequences of tagged elements may be introduced as right-hand-sides of grammar rules. For example, consider this general rule:

```
<NP>  <=   (<ALPHANUMERIC> <ADJP>  <NP> <NP>)
```

Although this rule could be used to parse complex phrases like *QA3556 upper control arm group,* it would also admit many other phrases as well. When fired in a context where it is not required, this rule would produce additional ambiguous analyses. This rule can be contrasted with a more domain-specific rule which uses SGML-tagged constituents:

```
<NP> <= (<PART-NUMBER> <PART-MODIFIER> <PART-GROUP> <GROUP-TYPE>)
```

This rule is sufficiently narrow that it will only fire in contexts where it is required.

4.2 Markup Examples

The following are some examples of SGML tagging conventions which improve the quality of the source text and should be considered for controlled English:

- *Callouts.* Integers which refer to arrow labels in schematic diagrams should be tagged, so they will not be confused with numeric quantifiers.

- *Special Forms.* Special phrases, such as chemical formulas, dates, addresses, and letter/number identifiers should be tagged and parsed with special grammar rules.

- *Measurement Expressions.* Compound expressions of measure should be tagged to reduce parsing complexity, for example:

```
<measure><metric>42.931 &plusmn;   0.013 mm</metric>
          <english>1.6902 &plusmn;  .0005 inch</english>
</measure>
```

Specific grammar rules which parse the open/close tags in nested constructions like this one guarantee that they will fire only in desired contexts, limiting ambiguity.

5 On-Line Controlled Authoring

In order to deploy controlled English for production authoring of technical text, an on-line system must be created for interactive checking of texts. This ensures that texts conform to the desired vocabulary and grammar constraints. An on-line authoring system can also support interactive disambiguation of lexical and structural ambiguities in the text. When problems are found, the author is asked to either rewrite parts of the sentence (with some help from the system) or answer questions about the sentence (to eliminate ambiguity). The result is a text which meets the constraints of controlled English, and encodes a single chosen meaning for each ambiguous lexical item or PP attachment.

5.1 Vocabulary Checking

Once a controlled English vocabulary has been specified, it can be built into a vocabulary checking tool for on-line use by the author. For example, CATALYST, the KANT application for Caterpillar, is combined with an authoring workstation environment called ClearCheck, developed by Carnegie Group, which checks that the vocabulary in each sentence conforms to the controlled vocabulary. The vocabulary checker uses information about synonyms and ambiguous terms to notify the author when his use of a term may not be appropriate, and attempts to offer alternatives whenever possible. Documents do not conform to controlled English until they pass vocabulary checking.

5.2 Grammar Checking

The ClearCheck tool also performs grammar checking. The controlled grammar is built into a grammar checking component, which uses the same parsing engine as the source text analyzer. This grammar checker parses each sentence in the source text to determine if a valid analysis can be found. If no analysis can be produced, then the sentence does not conform to controlled English and must be rewritten.

5.3 Interactive Disambiguation

If more than one valid analysis is found for a sentence during grammar checking, the grammar checker will indicate whether a lexical ambiguity or a structural ambiguity is the cause. The ClearCheck tool then queries the author interactively, providing a choice of meanings for the word in question (lexical ambiguity) or the structure in question (PP attachment ambiguity). ClearCheck then inserts an SGML tag into the sentence which captures this choice (cf. Sections 2.5 and 3.3 for a description of these tags).

6 Design Issues

In this section, we discuss a few of the more important issues in designing a controlled English.

6.1 Does Controlling the Source Text Really Help?

When controlled English is introduced, the number of parses per sentence can be reduced dramatically. If a general lexicon and grammar are used to parse specialized domain texts, then analyses may be assigned which are not appropriate in the domain.

We have experimented with the KANT analyzer in order to determine the positive effects of the controlled English mentioned above. We used a test suite of about 750 sentences (part of a development/regression test suite for one KANT application). The sentences in the test suite range in length from 1 word to over 25 words. When a constrained lexicon and grammar for the domain were utilized, along with disambiguation by the author, the average number of syntactic analyses dropped from 27.0 to 1.04. 95.6% of the sentences were assigned a single interlingua representation. Constraining the lexicon seems to achieve the largest reduction in the average number of parses per sentence. As expected, the best results are achieved when the system is run with constrained lexicon and grammar (Baker, et al., 1994).

6.2 Expressiveness vs. Complexity

If we assume that the *expressiveness* of a language is some measure of the variety of lexical and grammatical constructions it allows, then the more expressive a language is the more complex it will be to analyze during translation. In some cases, however, reducing the expressiveness of a language does not necessarily reduce the complexity of analysis. In systems where the vocabulary is extremely limited (as, for example, in the earlier Caterpillar Fundamental English), the authors may need to write long, convoluted sentence to express complicated meanings. In KANT Controlled English, the size of the vocabulary is not limited, and only those lexical or grammatical constructions which are unnecessarily complex are ruled out. The result is a language which is expressive enough to author technical documents, but limited in complexity such that high-quality translations can be achieved (Baker, et al., 1994).

6.3 Author Involvement vs. Post-Editing

An original goal in developing KANT Controlled English was to eliminate lexical ambiguity entirely. When this seemed impractical following domain analysis, it was decided to increase the amount of author involvement by introducing interactive disambiguation. Since the effect of ambiguity in the source text is reduced accuracy in the target text, increased post-editing is avoided when authors help to disambiguate the text. This is desirable in domains where the source language is translated to several target languages and increased cost of post-editing is prohibitive. In domains where there are fewer target languages, the other side of this trade-off might be explored, if the number of ambiguous terms and types of post-editing operations required allow cost-effective post-editing.

6.4 Judging Domain Suitability

Controlled English is not suitable for every domain where English is a source language. In particular, controlled English works well in domains with the following characteristics:

- *Centralized Authoring for Document Production.* When documents are authored at a centralized facility for document production, it is possible to control the style and content of the source text. This type of translation is referred to as translation for *dissemination.* When the documents to be translated are authored at multiple, remote sites (translation for *assimilation),* the domain is less amenable to a controlled English approach because there is less control over the source documents (unless, of course, the documents are checked and rewritten before translation).

- *Well-Defined Set of Highly-Trained Authors.* Although anyone can use an authoring tool to improve the quality of their source text, it seems that the best results are achieved by authors who receive comprehensive training and use controlled English on a daily basis.

- *Use of Controlled English Authoring Environment.* Although controlled English can be used simply as a set of guidelines for authors, uniform quality of authored text is maximized if the controlled English is closely integrated into the editing environment the authors use to create text.

- *Focused Technical Domain.* The success of controlled English relies heavily on the possibility of eliminating meanings for terms which are not necessary for the domain. The implication is that domains which require general, unrestricted use of terminology are less appropriate for controlled English.

6.5 Remaining Challenges

An interesting phenomenon arises during the process of building a checker for a controlled language. A sentence may have more than one possible syntactic analysis, but only one of the analyses conforms to Controlled English. Even if the author intended the “incorrect” reading of the sentence, it will still pass with the “correct” analysis. In such cases, the translation output will be for the “correct” analysis of the sentence, rather than the intended meaning of the sentence, leading to accuracy errors in the output. For example, the system will appear to accept conjoined noun phrases with a gap in one NP, when in fact it has analyzed the sentence without introducing a gap:

Input Phrase: “top(N,ADJ) and bottom gaps”
Analysis: “top(N) and [bottom gaps]”
Author’s Intended Use: *”top(ADJ) e_i and bottom gaps$_i$”

Another issue arises with rules that restrict the attachment of adverbs to verbs only, when the author intends that the adverb modify a following preposition:

Input Sentence: “Do not stand directly under a hoist.”
Analysis: “Do not [stand directly] under a hoist.”
Author’s Intended Use: *”Do not stand [directly under] a hoist.”

Sentences which pass grammar checking in manner unintended by the author usually result in incorrect translations (except in the infrequent cases where the translations of the two analyses are string invariant). We are presently investigating two ways in which these sentences might be addressed:

- By placing tighter semantic restrictions on the conjuncts; for example, one could eliminate readings like `″top(N) and [bottom gaps]″` by stating in the domain model that “top” and “gap” cannot be conjoined because they are of different semantic types.
- By asking the author to confirm the system’s analysis when checking potentially problematic conjunctions. If the analysis made by the system is not what the author intended, the sentence must be rewritten.

7 Conclusion

In this paper, we have discussed principles of controlled English design that we have developed during work on KANT Controlled English. A version of this controlled English, called CTE (Caterpillar Technical English), has been deployed in CATALYST, a KANT application for heavy machinery, and is currently supported by Carnegie Group’s ClearCheck authoring tool. The CATALYST/ClearCheck system supports a technical vocabulary of 60,000 words and phrases, and has been deployed at Caterpillar for technical authoring and translation. Although still in the tuning phase, the first language pair (English-French) has achieved 90% accuracy on controlled texts using a strict evaluation methodology (Nyberg, et al., 1994). More details concerning the implementation can be found in (Mitamura et al., 1993; Baker et al., 1994).

The results of our work indicate that the effort taken to develop a controlled input language for translation can certainly improve the quality of source and target text when the domain is suitable for a controlled language implementation. The techniques we have shared here are not specific to KANT, and can be used with any translation system that uses a lexicon and a grammar in source analysis, and/or SGML tags as text markup.

Acknowledgements

A large number of people at the Center of Machine Translation have contributed to the creation of Controlled English, including Kathy Baker, Nicholas Brownlow, Jaime Carbonell, Alex Franz, Sue Holm, Kathi Iannamico, Bob Igo, Marion Kee, Daniela Lonsdale, Deryle Lonsdale, Venkatesh Narayan and other members of the CATALYST project. We are also indebted to our colleagues at Caterpillar Inc. and the Carnegie Group for their collaboration.

References

[1] Adriaens, G. and D. Schreurs (1992). “From COGRAM to ALCOGRAM: Toward a Controlled English Grammar Checker,” *Proceedings of COLING-92.*

[2] Baker, K., A. Franz, P. Jordan, T. Mitamura and E. Nyberg (1994). “Coping With Ambiguity in a Large-Scale Machine Translation System,” *Proceedings of CO LING-94.*

[3] DeMauro, P. and M. J. Russo (1984). “Computer Assisted Translation at XEROX Corporation,” *Proceedings of the 25th Annual Conference of the American Translators Association,* New York, NY, September 19-23.

[4] Francis, W. and H. Kuĉera (1982). *Frequency Analysis of English Usage,* Boston, MA: Houghton Mifflin.

[5] Goodman, K. and S. Nirenburg (eds.) (1991). *A Case Study in Knowledge-Based Machine Translation,* San Mateo, CA: Morgan Kaufmann.

[6] Grishman. R. and R. Kittredge (eds.) (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing,* Hillsdale, NJ: Lawrence Erlbaum.

[7] 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.

[8] Mitamura, T, E. Nyberg and J. Carbonell (1993). “Automated Corpus Analysis and the Acquisition of Large, Multi-Lingual Knowledge Bases for MT,” *Proceedings of TMI-93.*

[9] Nyberg, E., T. Mitamura and J. Carbonell (1994). “Evaluation Metrics for Knowledge-Based Machine Translation,” *Proceedings of COLING-94,* Kyoto, Japan, August 5-9.