Controlled Languages in Technical Documentation

Pim van der Eijk

Cap Gemini Advanced Technology Services

Abstract

Controlled Languages (CLs) are increasingly adopted in technical documentation to support operation and maintenance of complex technical systems, in particular by non-native users. This paper provides a background on the application domain of technical documentation, the use of CLs in this area, and the use of Natural Language Processing to support creation and machine processing of CL source texts. We report on the AECMA Simplified English (SE) standard and experience gained in implementing authoring support for SE. The paper outlines the application of CLs to support for document translation, and concludes with a discussion of trends and research issues in development of CL processing systems.[1]

1 Introduction

Controlled Languages (CLs) are explicitly defined subsets or variants of language that are constructed for use in particular environments and are aimed at specific purposes.[2] They are used mainly for the authoring of technical documentation. The first CL for technical documentation, Caterpillar Fundamental English (CFE) was developed in the 1960s to improve comprehensibility for non-native users (Caterpillar is a manufacturer of heavy equipment). It was even hoped that CFE would altogether obviate the need for translation at Caterpillar (Kamprath et al., 1998). CLs have maintained a strong association with the objective of reducing costs or time requirements of translation.

As a result of increased quality requirements and economic globalization, CLs are increasingly used in technical documentation, and there is an increasing interest in using NLP to support the creation of comprehensible CL source texts. Apart from reducing time or costs of translation, a CL for technical documentation can therefore be designed to improve comprehensibility of technical documentation.

In this paper, we present CLs used in the application area of technical documentation as a challenge for the field of language engineering, and want to draw attention to the topic of authoring support and computer analysis of technical documentation as an interesting source of research topics for computational linguistics research.

[1] Part of the content of this paper appeared in abridged form in the February 1998 issue of the *Elsnews* newsletter.

[2] A useful on-line resource on Controlled Languages is the CL homepage at http://www-uilots.let.ruu.nl/www/Controlled-languages/.

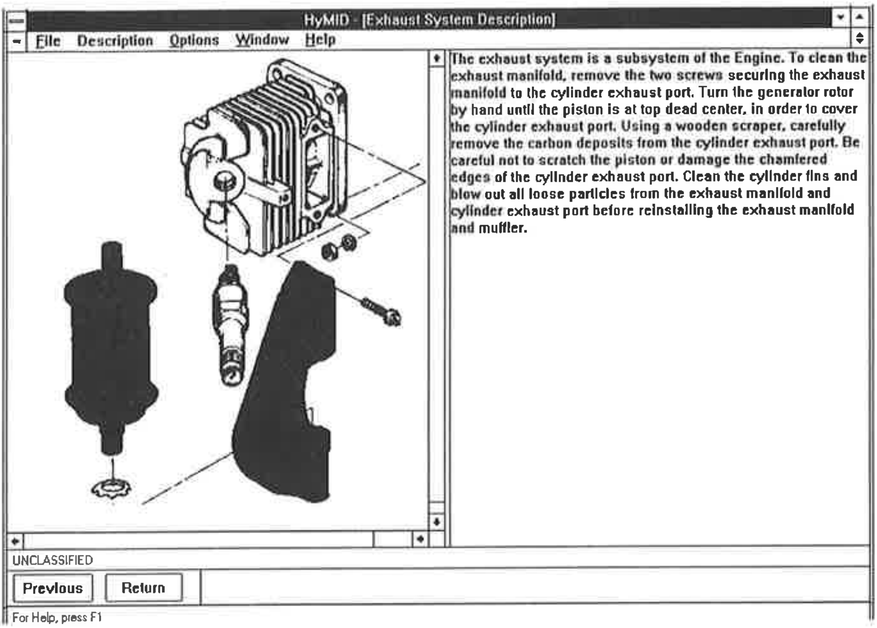
This paper is structured as follows. Section 2 provides some background on the application domain and discusses some of the linguistic properties of technical documentation in relation to the more general topic of sublanguages. In section 3, we turn to CLs, focusing on a particularly well-known CL, the aerospace industry's AECMA Simplified English standard. The SE lexicon and "writing rules" will be briefly described, as well as the actual benefits of SE for aircraft maintenance. Section 4 provides an overview of NLP support for the authoring of CLs in general. It describes a number of usage scenarios for these tools, as well as functional requirements that have to be addressed. Section 5 discusses some of the experience gained in the implementation of a specific authoring support system for Simplified English. Section 6 describes how CLs can support human, machine-aided and automatic translation. Section 7 discusses some trends and open issues in research and development on CL authoring systems. Section 8 summarizes the main conclusions of the paper.

2 Technical documentation

When technicians perform operational tasks, maintenance procedures and fault diagnostics on complex technical systems, they need recourse to technical documentation. The quality of such documentation is critical. If the documentation is inaccurate, incomplete, or hard to understand, the system's Mean Time To Repair and the number of Incorrect Parts Replacements will increase. More seriously, damage to expensive equipment and injury to humans may result, which could lead to costly liability claims against the supplier. This is one area where CLs are playing an increasingly important role: the use of a CL in the authoring of technical documentation improves its quality and comprehensibility.

As an example of the text type and application domain of technical documentation, the system displayed in Figure 1 shows an Interactive Electronic Technical Manual for a generator developed for a defense application.[3] It is representative of the complexity of technical information conveyed by technical information systems.

[3] This screen display was generated using the HyMID research IETM application (Anderson, 1996). For background information on the IETM application and standards, see Harvey (1997) and van der Eijk (1998).



The exhaust system is a subsystem of the Engine. To clean the exhaust manifold, remove the two screws securing the exhaust manifold to the cylinder exhaust port. Turn the generator rotor by hand until the piston is at top dead center, in order to cover the cylinder exhaust port. Using a wooden scraper, carefully remove the carbon deposits from the cylinder exhaust port. Be careful not to scratch the piston or damage the chamfered edges of the cylinder exhaust port. Clean the cylinder fins and blow out all loose particles from the exhaust manifold and cylinder exhaust port before reinstalling the exhaust manifold and muffler.

Figure 1: Interactive Electronic Technical Manual.

Apart from the general objective of quality assurance of source documentation, a second objective of CL is to improve the comprehensibility of technical documentation for non-native users and to support the documentation translation process. As a result of economic globalization, companies in many sectors are selling their products into a global, and therefore multilingual customer base. In the case of high-tech products, such as medical equipment, military aircraft, or computer software, some vendors find that many (sometimes, the majority of) target users have insufficient command of technical English to correctly interpret complex procedural information. Availability of documentation in the local language is becoming a government requirement in a growing number of markets, in particular in the European Union. However, translation of documentation is not always an option, be it for reasons of costs or time-to-market.

Technical documentation can be viewed as an instance of the more general concept of sublanguage. Sublanguages have been studied extensively for applications in information extraction and machine translation. An extensive bibliography exists on the subject (Kittredge and Lehrberger, 1982; Grishman and Kittredge, 1986). While the term is most often used to refer to a language used in a restricted subject matter domain, in practice the linguistic properties of sublanguage text are also determined by various factors other than subject matter. One of these factors is the target user group, e.g. documents for experts differ from documents written for general users. A second factor is the communicative purpose of a document. A document can be written to provide descriptive (background) information, to specify certain required functionality, or to provide procedural information. These different purposes will result in different writing styles. There may also be general contextual restrictions that determine how text is written, such as the difference between on-line reading versus off-line reading. Even within a single relatively narrow domain, it is therefore sometimes possible to distinguish a variety of sublanguages. These sublanguages then differ on one of these other parameters, e.g. the difference between local weather reports versus regional weather synopses discussed in Kittredge (1982).

Although it is possible to view sublanguages as subsets, or specializations, of general language, sublanguages are best seen as independent syntactic systems, that have their own internal consistency. For example, in an analysis of a telegraphic sublanguage (Fitzpatrick et al., 1986), it is claimed that a sublanguage-specific constraint (where verbs apparently can either be intransitive or transitive, but never both) accounts for sublanguage-specific non-ambiguity of constructions involving gapping, non-copula passives, and middles. By taking advantage of this constraint, a simpler and more accurate grammar results than would be possible if the sublanguage grammar were defined in reference to, or as a more restricted variant of, the "standard" language grammar.

In some cases, a fairly direct encoding of syntactic structure in terms of domain-specific categories has been proposed. For instance, one paper refers to categories like "organ/cell" (for "liver") and "lipid" (for "cholesterol") in the sublanguage of lipoprotein kinetics. Subcategorization is then expressed in these categories (e.g. the "synthesize" operator combines with an "organ/cell" argument and a "lipid" argument) rather than in terms of syntactic categories (Sager, 1986).

Most frequently, sublanguages are contrasted with unrestricted, general language in terms of the reduced variety of grammatical and stylistic constructions, and the limited variety in words and word meanings that occur in sublanguage documents. As a sublanguage, the technical documentation of some complex industrial products indeed stands out against general language documents because of its combination of an unusual volume — tens, sometimes even hundreds of thousands of pages — with a highly repetitive character. For example, part of a computer software manual we have worked with contained 75,195 word tokens, but only 2874 word types, not even accounting for variation due to inflectional morphology. Beyond the lexical level, this document was also highly repetitive at the level of NPs and even entire sentences.

The reduced lexical and grammatical variety, and the ambiguity reduction that results from domain restrictions have often lead to optimism on the machine-processability of sublanguages. Indeed, the best-known successful application of machine translation, TAUM METEO, is a sublanguage application. A recent application of METEO technology produced near-perfect (93.2%) real-time weather translations and increased human translator efficiency by up to 800% (Chandaloux and Grimaila, 1996).

Nevertheless, the complexity of many sublanguages is more similar to general language than to the weather reports that are translated by METEO. As an example, in lexical analysis of the computer software manual referred to earlier, we analysed lexical word sense variation using a general-purpose bilingual lexicon and a (domain-specific) glossary as reference information. While many general-language word senses never occurred in the manual, the glossary often distinguished additional (domain-specific) senses of terms, as well as more specific subsenses that do not exist in the general language but did occur in the document. Such sense distinctions would also show up in translation.

Some commercial MT systems offer a facility where a hierarchy of dictionaries can be defined based on subject matter distinctions. When translating a particular document from a particular domain using one of these systems, domain-specific senses take precedence over general language senses. Unfortunately, this facility would be insufficient for the computer manual corpus, where it is not the case that domain-specific senses always suppress general senses when both a general sense and a domain-specific sense exist, and where there exists sense ambiguity even within the domain. This means that lexical ambiguity in general is an issue for this sublanguage as it is for unrestricted language, unless additional constraints are enforced.

At other levels than the lexical level, the same point can and has been made. As is well-known, the TAUM group could not reproduce the success of METEO in a follow-up project (AVIATION) for aerospace documentation. The structural ambiguity of, for instance, complex noun sequences could not be resolved correctly in an automatic fashion.

3 CLs for technical documentation

Where "sublanguage" is a descriptive term that can be applied to a collection of texts on the basis of certain observed properties, the term "controlled language" is a prescriptive term, which is associated with properties that are to be enforced during document creation.[4] Many CLs in use in the technical documentation application area can be viewed as "controlled sublanguages", as they are generally defined as additions to and selections of existing sublanguage practice (van der Eijk et al., 1996).

Nowadays, the concept of CL is adopted in some form by hundreds of companies and organizations, ranging from the use of (often rather informal) company-internal guidelines for technical writing and lists of preferred and unallowed terminology, to a number of professional document authoring systems that use full parsing to enforce validation of an application-specific CL grammar. Apart from English, controlled variants of French, German and Swedish have been defined (CLAW, 1996; CLAW, 1998).

However, most applications of CLs are unknown outside the sites at which they are used. There seem to be three reasons for this. First of all, technical documentation, in particular at the level of lexicon and terminology, is inherently

[4] Some artificial languages used for database access, command and control applications and requirements specification are referred to as CLs as well. However, these languages are very different from the controlled sublanguages used in technical documentation.

domain-specific, or even specific to a particular company, type of product, and target user. This reduces the portability of a CL to other domains.[5] Furthermore, many companies that develop and use CLs see no need to disclose to outsiders what, to them, represents proprietary knowledge and experience, and may in some cases offer them a competitive edge. Finally, the general NLP research community seems to view the technical documentation domain as a somewhat uninteresting application area, rather than recognize it as the source of research challenges it is to those who are more familiar with it.

There are some exceptions to the limited familiarity of CL. The best-known of these is the AECMA Simplified English (SE) standard. An initiative of the European Association of Aerospace Industries (AECMA) and the Aerospace Industries Association (AIA), Simplified English was developed as a controlled English for (mainly) procedural texts in aerospace maintenance manuals. Its latest revision, which dates from 1995 and can be ordered from the AECMA office, refers to an intention "to help the users of English-language documentation understand what they read". SE is now an international aerospace industry standard. Although the quote shows that SE was not designed to be a language for pre-editing MT, SE does in fact address many of the complexities that resulted in the failure of the TAUM AVIATION project.

SE consists of a controlled general vocabulary and a set of "writing rules". The general vocabulary consists of over a thousand Approved Forms, which can (with some exceptions) be used in one sense and in one part of speech only. The SE dictionary also lists a larger number of unapproved forms, and their correct replacements in the set of Approved Forms. An example of the distinction is provided in Table 1, which shows an approved form, viz. the prepositional use of "about", and an unapproved form, "abnormality", for which "defect" should be used. Note the use of capitalization in the SE guide to encode the approval status of vocabulary. The core SE vocabulary can be extended with Technical Names (nouns and adjectives), drawn from 20 subject-matter specific categories (e.g. "location on aircraft", "damage term", "medical term"), and Manufacturing Processes (verbs), of which there are six types (e.g. "remove material" and "change shape").

[5] This lack of portability is most easily demonstrated at the lexical level. As an example, we compared a core set of 1219 approved lexical forms from the Simplified English dictionary to a large (1.2 million word) computer software manual corpus. From this SE core vocabulary, 457 items did not occur at all in the software manual corpus.

Table 1: Sample SE dictionary items.

Keyword (part of speech)	*Assigned Meaning (USE)*	*APPROVED EXAMPLE*	*Not Acceptable*
abnormality (n)	DEFECT (TN)	EXAMINE THE CANOPY SEAL FOR DEFECTS	Inspect the canopy seal for abnormalities
ABOUT	"Concerned" with. NOTE: For other meanings, USE: APPROXIMATELY, AROUND	FOR DATA ABOUT THE LOCATION OF CIRCUIT BREAKERS, REFER TO THE WIRING LIST	

The grammatical part of the SE guidelines consists of a set of writing rules. Nine categories of rules are distinguished. This classification is partly based on syntactic distinctions, viz. "words", "noun phrases", "verbs", "sentences", "punctuation", and "writing practices". It is also based on linguistic differences between different document component types: "procedures", "descriptive writing", "warnings and cautions". The 57 writing rules in the standard demonstrate a greatly varying complexity. Five examples of these are given in Table 2.

Table 2: Sample SE writing rules.

Identifier	*Rule*
5.1	"Keep procedural sentences as short as possible (20 words maximum)"
1.2	"Use approved words from the Dictionary only as the part of speech given"
2.3	"When appropriate, use an article or a demonstrative adjective before a noun"
6.8	"Present new and complex information slowly"
7.3	"If necessary, add a brief explanation to a warning or caution to give a clear idea of the possible risk".

The first three rules require different levels of linguistic analysis. Rule 5.1 requires only an ability to count sentence length, which could be readily implemented using pattern matching. Rule 1.2 requires part of speech disambiguation, another issue that has received much attention in the literature and can nowadays be considered to be a standard technique. Rule 2.3 requires a much more sophisticated syntactic analysis, as it needs detailed knowledge of article placement rules in English syntax. The latter two rules demonstrate that SE, although presented as a standard for the "aerospace language", is more strongly focused on the general issue of technical communication rather than on linguistics per se. These rules are concerned with discourse and pragmatics of technical written language, rather than adherence to a particular grammar.

Despite its informal character, SE is by far the most well-researched CL in practical use. An advantage of this fact is that the benefits SE was intended to bring have been investigated relatively well. In a recent study (Chervak et al., 1996), SE has been subjected to a fairly rigorous evaluation of the extent to which the objectives of improved comprehensibility were met. One of the encouraging results of this study is that non-native Aircraft Maintenance Technicians (AMTs) were found to perform complex procedures significantly better if they had been given an SE version of the documentation instead of the earlier non-SE version. Clearly, the case of complex instructions that are to be performed by non-native users is one of the situations where SE is most relevant. Given the safety-critical nature of procedural information in aerospace documentation and the considerable cost of aerospace equipment, these results are important. Compliance to SE, and investment in SE implementations within an organization, are not just a matter of meeting industry standards, but represent an investment for which a business case can be made.

Furthermore, Holmback et al. (1996) reports on a case study where the benefit of increased comprehensibility of SE source text was found to carry over in translation to a related target language, Spanish.[6]

4 Supporting CL authoring

Given the benefits of SE, and of CLs used in similar conditions, as well as SE's status as an aerospace industry standard, there is a genuine need in industry for NLP tools to support the process of authoring CL-compliant documentation. These tools can be designed and used for three similar, but distinct, usage scenarios:

- Quality Assurance departments are concerned with the extent to which a company meets its legal and contractual requirements. Just as the use of SGML implies a need to validate documentation against an aerospace industry standard Document Type Definition (e.g. in the aerospace industry, there are the AECMA 1000D and the MIL-PRF-87269 DTDs), aerospace industries perceive a need for *validation* of the documentation against the SE standard.
- Another use for support tools for CLs is in Computer-Based Training (CBT). SE is designed for increased ease of comprehension, but authoring SE is significantly more difficult than authoring general technical English. Experience at Alcatel Telecom showed that CL authoring can take up to 20% more time than authoring uncontrolled text (Goyvaerts, 1996). Anecdotally, a manager of a technical documentation unit in a former Netherlands-based aerospace industry company claims that the company's native English authors were actually *less* SE proficient than their native

[6] Translation to Chinese and Japanese did not reveal any significant differences between SE and non-SE input, however.

Dutch staff. A CL tool can be used to help technical writers familiarize themselves with the constraints of the CL.

- Finally, on-line checking, presentation of alternative phrasings, and diagnostic explanations of errors can support authors in production authoring. They will at this stage increasingly use the tool for confirmation purposes only.

From a computational linguistics point of view, CLs have two important properties that make them interesting candidates for NLP. These properties follow from CLs being restricted variants of sublanguages:

- The enforced coverage bounds of a CL indicate that a potentially high level of recall is achievable, as compared to applications for unrestricted text that are inherently sensitive to e.g. creative use of language or to unknown words. Even proper nouns like part names can often be retrieved from Part Catalogs stored on-line in Product Data Management (PDM) or Enterprise Resource Planning (ERP) systems.
- The objective of ambiguity resolution means that CL processing should in principle be able to achieve exceptionally high levels of precision.

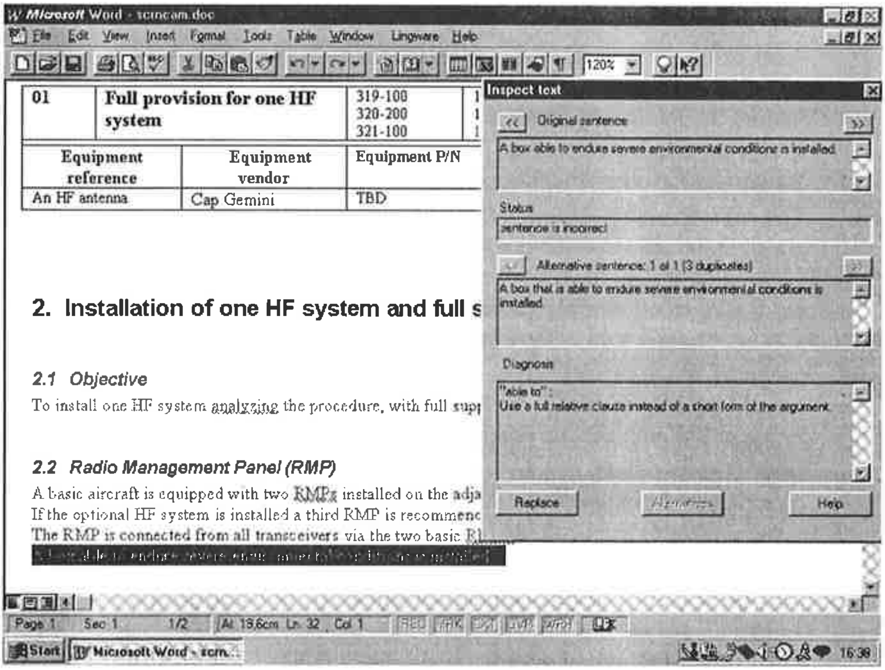
Several large companies have developed CL support tools using in-house staff. There are also a small number of companies in the computer services industry, including Cap Gemini, that offer CL support tools on a per-project basis. These tools in general aim to perform all or a subset of the following functions:

- linguistic analysis of CL text (as discussed for the rules in Table 1, the required "depth" of analysis varies greatly); this is a prerequisite for the following functions;
- generation of useful *critiques* to authors;
- general morpho-syntactic and spelling correction;
- support for interactive *transformation* of general sublanguage expressions into the CL (this feature has the advantage of being well appreciated by technical documentation authors, but it has the drawback of requiring a fairly complete and accurate syntactic analysis).

To the authors, CL checking and correction is to be offered as one additional, integrated, type of functionality in their familiar authoring environment. Professional authoring systems generally offer extensive support for customization and integration of additional functionality, which can be used to add linguistic functionality to the product.

To be able to offer transformation (correction) of input text, a correction system also has to model both uncontrolled author input and its mapping to correct CL expressions, in addition to implementing a grammar and a lexicon for the CL. Current CL authoring systems typically adopt the MT Transfer model or a simpler transduction grammar approach to implement this functionality.

CL systems differ in their treatment of ambiguity. In the case where a particular critique only applies to one interpretation of an input sentence (e.g. it is incorrect in one reading, but correct in another), it can be argued that a reportoriented (batch) system should suppress the critique.[7] In a system that offers corrections, ambiguity arises as well, because the paraphrase relation in general is many-to-many. An interactive system can offer the user a list of alternatives from which (s)he can choose. Each alternative will be associated with its own (set of) paraphrases and associated critiques. Figure 2 shows a screen example of an interactive correction session of an aerospace document using an integration of Cap Gemini's Controlled English correction software with the Microsoft Word editor.


Microsoft Word - scmcam.doc
File Edit View Insert Format Tools Table Window Lingware Help
120%
Inspect text
<<
>>
Replace
Alternatives
Help
Page 1 Sec 1 1/2 At 19,6cm Ln 32 Col 1
REC TRK EXT OVR WPH
Start
Microsoft Word - scm...
16:38

01	Full provision for one HF system	319-100 320-200 321-100	1 1 1

Equipment reference	Equipment vendor	Equipment P/N
An HF antenna	Cap Gemini	TBD

2. Installation of one HF system and full s

2.1 Objective

To install one HF system analyzing the procedure, with full sup

2.2 Radio Management Panel (RMP)

A basic aircraft is equipped with two RMPs installed on the adja
If the optional HF system is installed a third RMP is recommenc
The RMP is connected from all transceivers via the two basic R

Original sentence

A box able to endure severe environmental conditions is installed.

Status

sentence is incorrect

Alternative sentence: 1 of 1 (3 duplicates)

A box that is able to endure severe environmental conditions is installed.

Diagnosis

"able to":
Use a full relative clause instead of a short form of the argument.

Figure 2: Interactive CL authoring session.

This Controlled English authoring system is based on a grammar for Controlled English written in a feature constraint-based formalism. The formalism was designed for on-line correction and checking, and offers a compact notation for both incorrect (but correctable) input language expressions and the corresponding CL "output" expressions. The compiled grammar is interpreted at runtime by an LR-parser for ambiguous extended context-free grammars. As the system is to be used interactively during authoring, acceptable run-time performance on standard hardware and software platforms is very important for this application.

The formalism has also been applied to machine translation, in dedicated systems developed for software localization projects. These projects involved

[7] This point is made in Barthe et al. (1998).

authoring of CL source data, with subsequent machine translation to multiple target languages.

5 Implementing authoring support for SE

On the basis of experience gained in implementing systems for SE and other CLs, and evaluating their use in practical technical documentation systems, a number of observations can be made on SE and computer support for it, some of which generalize to other CLs.

SE has its origins in the technical documentation community, rather than the field of linguistics, and is intended to provide practical recommendations to technical publications departments in aerospace companies. As a result, the SE standard is, from a (computational) linguistic point of view, a rather informal standard. Interactive correction functionality as displayed in Figure 2 requires implementation of a grammar for SE. An SE grammar implementation, by necessity, is an interpretation of the informal SE specification.

There are many issues that are implicit or not discussed at all in the SE specification. For example, in implementing a relative clause module for our SE grammar, we did not find specific information in the standard on whether or not preposition stranding and omission of relative pronouns are allowed.[8] In the absence of a conformance test mechanism, different developers are bound to make incompatible design choices, and therefore will produce implementations that do not interoperate. In the lexicon, a similar lack of detail is the absence of subcategorization information (Humphreys, 1992). As various aerospace companies tend to interpret SE in different ways, SE tools need to be customized for different applications.

Furthermore, as argued in section 3, many rules in SE (e.g. 6.8 and 7.3 in Table 2) are rules of technical communication rather than rules of linguistics. Despite demand for validation of these rules from the SE user base, these rules cannot be machine-checked effectively at present. For these reasons, checkers and correctors for SE and similar CLs may best be viewed and used as aids for initial training and for on-line assistance during production authoring. We have found that looking at a CL checker as a validation tool can have unwanted “Pavlov” effects on some authors. By stressing their value as training and assistance tools, the higher-level goal of assisting authors to create understandable documentation is emphasized over the lower-level goal of getting the CL checker to accept all sentences in a particular document. With or without an SE checker or corrector, it is easy to produce “Bad SE”, and no CL tool can be used as a substitute for general writing skills, the necessary understanding of the information that is to be conveyed, general domain knowledge, or common sense.

[8] In developing our SE grammar, we assumed they are not and we implemented automatic correction of sentences without relative pronouns and with stranded prepositions. A similar position was taken in BTE, a recent extension of SE developed at Boeing (Wojcik et al., 1998).

CL designers and specification committees, as well as CL tool developers, would very much benefit from research by linguists and communications researchers that would help determine precisely which features of a CL contribute most to comprehensibility, and at which cost, in terms of authoring complexity. This research could also help resolve the inconsistencies that currently exist among various CL specifications. For instance, where SE rule 2.3 prefers articles or demonstrative adjectives before nouns, clause 3.3.3.b of MIL-PRF-87268A, a US defense specification for IETM content, states that when "procedural text is combined with graphics, [...] so long as the meaning is not altered or obscured: (1) Eliminate articles".

6 CLs and translation

As discussed previously, the multilingual customer base of its products motivated the early work on Controlled English at Caterpillar. At several sites, such as Perkins engines (Pym, 1988) and Xerox (Hutchins and Somers, 1992: 188), where (first generation) MT systems were used, pre-editing source text was found to reduce the need to post-edit MT system output, and thus overall translation costs. Controlled input and user interaction were unique design features of an operational MT system, TITUS, developed and used in the European textile industry in the 1970s and early 1980s (Hutchins and Somers, 1992; Kingscott, 1989). From the early TITUS system on, CLs have often been associated with MT, and with translation in general.

The Canadian weather reports translated by METEO can be argued to be the exception, rather than the rule, as far as the *inherent* suitability to machine processing is concerned. Language control can be viewed as an attempt to (artificially) increase the set of exceptions. In looking at the benefits of language control for translation in general, the following cases can be distinguished.

First of all, given the tiny fraction of the world's annual volume of translation of documents for publication in foreign languages that is currently performed using MT, it seems safe to say that the most obvious benefit that CLs bring to technical translation is the fact that comprehensibility of source documentation is as beneficial to human translators as to any other reader. Anecdotal evidence from companies specializing in technical documentation and translation seems to confirm that use of a CL prevents misinterpretations (which obviously may have dramatic consequences) and can help translators make significant time savings. In a comparison of human translation of SE and non-SE versions of maintenance procedure documents to three target languages (Holmback et al., 1996), the translations from SE source data were found to produce significantly higher ratings for style match and fewer minor omissions. Additionally, the Spanish translations from SE source data were also found to be rated significantly higher on accuracy and comprehensibility.

Nowadays, some vendors of MT software that target the corporate technical documentation segment of the translation industry[9] strongly recommend the use of CL for source language data. The reason is that they find they can reduce the overall MT post-editing effort by "tuning" the MT system (in particular, its dictionaries) to the source CL. The use of a CL is therefore seen to reduce the overall "Cost of Ownership" of the MT system. This advantage increases if there are multiple target languages.

In a dedicated system, the CL authoring environment and the MT system can be developed in combination. This greatly simplifies the design and implementation of the MT system, and will in general result in better translation quality. Such dedicated systems tend to be more expensive to develop, however.

The same complementary benefits reported by MT system vendors are reported by vendors of products for terminology management and Translation Memory (TM). TM systems are productivity tools for human translators that operate by finding fuzzy matches and generating proposal translations based on previous translations and terminological data. The lexical and syntactic standardization that a CL brings to source documentation increases the hit rate of these systems, and thus overall human translation efficiency (Janssen et al., 1996; Brockmann, 1997).

7 Trends and issues in CL processing

There are many unresolved issues in machine processing of CL that need to be addressed to improve the performance and usefulness of CL systems. First of all, in terms of the level of analysis, most production CL checkers tend to be limited to morphological and syntactic analysis. Some limited experiments have been performed in the area of semantic analysis, on word sense disambiguation (Boeing) and on resolution of anaphoric references (Cap Gemini). These systems seem to operate largely at the level of individual sentences, whereas comprehensibility of documentation obviously depends heavily on inter-sentential relations such as coherence (or the more computationally tractable concept, cohesion). It would certainly be worthwhile to explore whether results from computational analysis of discourse and text analysis could be applied to CL analysis.

In CL checking, as in the more general field of grammar checking and correction, precision (accuracy) is vital and much more important than recall (coverage). This is because users tend to rapidly loose confidence in the system if it produces bad critiques or corrections, whereas they are more willing to accept that the system cannot provide critiques, let alone corrections, for arbitrary incorrect input expressions. As a result, CL systems tend to be rather conservative in the type of checks, or corrections, that they implement. Often, violations of the more complex CL constraints are much easier to detect than to correct. Nevertheless, increasing the coverage of CL systems and their ability to propose

[9] Logos and Systran company presentations at the Machine Translation Summit VI, San Diego, California, USA, 1997.

high-quality paraphrases is important to gain end-user acceptance. This is especially important in cases where large volumes of legacy documentation have to be rewritten to comply with the CL standard.

The paraphrase relation between incorrect and correct SE is not limited to relations between lexical items within a single syntactic category. The SE guides contains many cases reminiscent of the so-called "complex transfer" cases discussed in the Machine Translation literature. These include nouns in noun compounds that are converted to prepositional phrases, gerunds and past participles that are rewritten as relative clauses, verbs that are rephrased as constructions involving support verbs, passives that become active sentences with expressed subjects, and anaphoric references that are replaced by (copies of) the referenced phrases. Unfortunately, these rules, unless artificially restricted, tend to overgenerate when implemented as corrections. In some cases, this results in large numbers of (at times rather implausible) paraphrase proposals. Therefore, there is a need to be able to rank paraphrase proposals according to a plausibility metric.

Apart from reducing precision, these complex paraphrases are also non-trivial to implement and tend to have a serious impact on performance. In systems that are intended for interactive production use, run-time performance is important and coverage extensions need careful consideration.

Reduction of ambiguity and syntactic complexity are important objectives in the design of CLs. The restrictions on the CL that are needed to improve machine-processability often go well beyond the restrictions needed to produce comprehensible, readable documentation for human users. In practice, certain domains are inherently too complex, even at the lexical level, to allow for the drastic reduction (or even elimination) of ambiguity that is required for an application such as high-quality fully automatic MT. For instance, in the Caterpillar domain there are reportedly seven senses for the general term "*valve*" that have distinct target language translations.[10] A complete elimination of this ambiguity was considered not to be feasible. The engineering solution adopted in this and other CL projects was to obtain disambiguation information from authors, and to store this information with the source data using an SGML encoding. In subsequent stages this encoding can be used by the MT or TM application for target language vocabulary selection. The encoding is also independently useful, as it can also be used by an IETM system for glossary creation and hyperlink generation. It seems likely that future CL authoring systems will extend this approach to other types of ambiguity, including attachment of PPs, non-finite clauses and coordinated structures, and referential ambiguity.

[10] Caterpillar presentation at the MT Summit VI, San Diego, California, USA.

8 Summary and discussion

With the growing complexity of industrial systems, and with increased globalization, the readability and comprehensibility of the associated technical documentation that are needed to correctly operate and maintain these systems becomes more and more important. SE has been shown to contribute to these objectives, and as such presents a successful instance that will further reinforce the interest in CLs and the demand for CL authoring support in industry.

Although the sublanguage characteristics seem to make effective language processing of CL texts feasible, experience in implementing SE processing tools shows that full machine checking and correction of the SE standard, especially the part that is concerned with technical communication rather than lexical or syntactic restrictions, is beyond the state of the art. In the authoring phase, CL tools are therefore best viewed as support tools for training and on-line interactive assistance, rather than for validation. Additional constraints are to be imposed beyond the SE rules if objectives like improved machine-processability are to be met.

Apart from improving the quality of source documentation, language control also contributes indirectly to the overall productivity in a full documentation life cycle, by reducing human translation time, improving the hit rate of Translation Memory systems, or reducing post-edition efforts when using Machine Translation systems. The increased use of translation technology also creates a need for CL source data, because of the need to reduce post-editing costs, particularly in the case of multiple target languages. In general, there is an increasing tendency to pragmatic solutions, where CL is one ingredient of a broader organizational and technical solution framework. CL, and its support environment, is just one of the factors in the entire documentation production chain, in which judicious use of SGML encoding, Translation Memory, terminology management, Machine Translation and human intervention at various stages all combine for optimization of the full documentation process.

Acknowledgements

For comments on an earlier version of this paper, the author wishes to thank Joost de Jonge, Jacqueline van Wees, Dennis Janssen and an anonymous reviewer.

References

AECMA (1995). *A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language.* AECMA Document: PSC-85-16598. European Association of Aerospace Industries, Brussels.

Anderson, M. ed. (1996). *The Metafile for Interactive Documents. Application Guide and Draft Performance Specification for the Encoding of Interac-*

tive Documents. MID-2. Naval Surface Warfare Center, Carderock Division, Maryland.

Barthe, K., G. Bès, J. Escande, D. Pinna and E. Rodier (1998). Issues related to realistic evaluation of Controlled Language Checkers, in CLAW (1998), 134-144.

Brockmann, D. (1997). Controlled Language & Translation Memory Technology: A Perfect Match to Save Translation Costs. *Technical Communicators' Forum* (4), 10-11.

Chandaloux, J. and A. Grimaila (1996). "Specialized" Machine Translation, in *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, 206-211.

Chervak, S., C.G. Drury and J.P. Ouellette (1996). Field Evaluation of Simplified English for Aircraft Workcards, in *Proceedings of the 10th FAA/AAM Meeting on Human Factors in Aviation Maintenance and Inspection.* Alexandria, Virginia, January 1996.

CLAW (1996). *Proceedings of the First International Workshop on Controlled Language Applications (CLAW 96).* Centre for Computational Linguistics, Leuven, Belgium.

CLAW (1998). *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW 98).* Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Eijk, P. van der, M. de Koning and G. van der Steen (1996). Controlled Language Correction and Translation, in CLAW (1996), 64-73.

Eijk, P. van der (1998). IETM: Interactieve Hypermedia met SGML, in *Element* 4:1, 11-19.

Goyvaerts (1996). Controlled English, curse or blessing? A user's perspective, in CLAW (1996), 137-142.

Grishman, R. and R. Kittredge (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing.* Lawrence Erlbaum, Hillsdale, New Jersey.

Harvey, B. (1997). Interactive Electronic Technical Manuals. Paper presented at the ISO 10303 STEP Conference. Chester, England.

Holmback, H., S. Shubert and J. Spyridakis (1996). Issues in Conducting Empirical Evaluations of Controlled Languages, in CLAW (1996), 166-177.

Humphreys, L. (1992). The Simplified English Lexicon, in *Proceedings of EURALEX-92*, 353-364.

Hutchins, J. and H. Somers (1992). *An Introduction to Machine Translation.* Academic Press, London.

Janssen, G., G. Mark and B. Dobbert (1996). Simplified German. A Practical Approach to Documentation and Translation, in CLAW (1996), 150-158.

Kamprath, C., E. Adolphson, T. Mitamura and E. Nyberg (1998). Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English, in CLAW (1998), 51-61.

Kingscott, G. (1989). *Applications of Machine Translation.* Study for the Commission of the European Communities.

Kittredge, R. and J. Lehrberger (1982). *Sublanguage: Studies of Language in Restricted Semantic Domains.* De Gruyter, Berlin.

Kittredge, R. (1982). Variation and homogeneity of sublanguages, in Kittredge and Lehrberger (1982), 107-137.

MIL-PRF-87268A (1995) *Interactive Electronic Technical Manuals - General Content, Style, Format, And User-Interaction Requirements.* Performance specification, US Department of Defense.

Pym, P. (1988). Pre-editing and the use of simplified writing for MT; an engineer's experience of operating an MT system, in *Proceedings of the 10th ASLIB Conference on Translation and the Computer*, 80-96. ASLIB, London.

Sager, N. (1986). Sublanguage: Linguistic Phenomenon, Computational Tool, in Grishman and Kittredge (1986), 1-19.

Wojcik, R., H. Holmback and J. Hoard (1998). Boeing Technical English: An Extension of AECMA SE beyond the Aircraft Maintenance Domain, in CLAW (1998), 114-123.