Controlled Language and Machine Translation

Uus Knops & Bart Depoortere
LANT NV
Contact address: uus.knops@lant.be
bart.depoortere@lant.be

Abstract

In this paper we discuss how design issues in machine translation have influenced the development of our controlled language checker, and how customer-specific requirements to our controlled language checker have, in turn, influenced and stimulated further developments in machine translation. It is shown, more particularly, how grammar and software development, undertaken for our controlled language application, have opened new perspectives to machine translation. Future developments are aimed at further enhancing the robustness of our technologies based on a strictly modular, but hybrid approach.

1. Introduction

1.1 MT: LANT®MARK™

LANT®MARK™ is a second generation machine translation (MT) system with clearly separated analysis, transfer and generation modules for its different language pairs. The transfer approach is syntax-based as well as semantic with lexical and case frame semantics being central in transfer. The LANT®MARK™ system is sold to industrial companies, to governmental organizations and to large and medium-sized translation offices. It is most successfully put into use in combination with Eurolang® *Optimizer*™, LANT's translation memory (TM). Both technologies are marketed as separate products and as an integrated solution to organizations with high translation volumes.

1.2 CL: LANT®MASTER™

LANT®MASTER™ is a controlled language (CL) checker, the design of which is based on SECC (A Simplified English Grammar and Style Checker/Corrector), a controlled language application developed in the context of the LRE-2 project with the same name (Adriaens 1996a). The tool developed in SECC aims to check technical English documentation in the field of telecommunication. Some of the SECC results are now being re-implemented in CASL (Controlled Automotive Service Language), a project that LANT carries out for General Motors and which, among other objectives, aims at developing a conformance checker for the automotive industry (Means and Godden 1996). Some SECC ideas have also been re-used in the LANT internal AECMA project, aimed at developing a controlled English checker for technical documentation in the aeronautic industry.

By now, LANT has implemented a whole library of controlled English rules, with each rule being tagged with the appropriate customer label. The AECMA conformance checker is purchased as a standard LANT product. However, even with regard to the AECMA rule set (AECMA 1995), we offer customer-specific rule adaptations, as we have learned by now that the various aeronautic industries all have their own interpretations of the AECMA rules. Such a customization is currently taking place for Airbus Industries in the context of DocSTEP, an LE-III project.

In all our CL projects, the conformance checker is conceived as a special language pair (English to Controlled English) within the LANT MT environment. Its input is an English sentence which is first analyzed by the same English grammar as is used for translation into German, French and Spanish. Conformance checking itself is effected during the transfer and generation phase. The output of the checker is a dual object: the input sentence annotated with diagnosis labels for all rule violations detected, and an automatically rewritten version of the input string with proposed corrections for some of the diagnosed violations. The restricted vocabularies are implemented as an English-English transfer lexicon. Conformant terms have a target language form which is identical to the source language form. Non-conformant terms are mapped into conformant alternatives.

Up till now, all rules and vocabularies have been specified and implemented for English. A fair amount of our CL rules could be easily extended to other languages, such as German, French, Spanish and Dutch, but we have not yet implemented controlled grammars and lexicons for languages other than English. Our library of controlled English rules is extensive enough to serve different application needs. In some applications the conformance checker is primarily conceived as a post-editing tool for native and non-native writers in a monolingual English document development environment (e.g., SECC, DocSTEP). In other applications the emphasis lies more on translation. Here, the checker is primarily used as a pre-editing tool for translation either by humans or by machine (e.g., CASL).

2. An MT approach to CL

In designing the CL checker as a particular MT application, a lot of our existing MT modules could be re-used, such as our converters for separating lay-out from textual information, our software for extracting and re-inserting translation/checking units, and our client-server architecture for batch-level job-processing.

More far-reaching was our decision to re-use the English analysis module including the English monolingual lexicon already available for MT language pairs, such as English-German and English-Spanish. With this decision the best possible checker results could be obtained at the lowest possible development costs. Indeed, an intelligent, robust and high-quality conformance checker requires a complete in-depth analysis of the English input (cf. Nyberg and Mitamura 1996). By re-using the English analysis module, the quality and performance of our checker could be expected to go far beyond simple string matching and word counting procedures as used in some other checkers.

Consider the CL rule pertaining to the restrictive use of passives. It is possible for a tool based on shallow linguistics to recognize passives, but it is more difficult for such a tool to rephrase a passive sentence into an active and to propose this as a revision to the human editor. This is no problem for LANT®MASTER™, provided the logical subject is expressed in the sentence.

E.g., - The program *can be started* by the administrator from the main screen.
+ The administrator *can start* the program from the main screen.

Another example is offered by the CL rule, stating that "*in order to*" should be used instead of "*to*" to introduce a purpose clause. In order to disambiguate correctly between the subclause conjunction "*to*", the infinitive marker "*to*", the preposition "*to*" and the verbal adjunct "*to*", a complete structural analysis of the sentence is needed. Such a distinction is reliably made by LANT®MASTER™, and hence the proposed correction is restricted to real purpose clauses.

E.g., + The program offers the possibility *to revise the text.*
+ The program allows the user *to revise the text.*
- The user selects the 'batch' or 'interactive' option *to revise the text.*
+ The user selects the 'batch' or 'interactive' option *in order to revise the text.*

A further advantage in re-using the English analysis lies in the fact that a complete consistency between checker results and MT results can be obtained. Such a consistency is very important in applications where the conformance checker is conceived as a pre-processing step to MT. Any CL grammar aims to improve the readability and translatability of written texts. Most CL rules will enhance both properties. A good example is the rule which restricts the use of pronouns to the second person singular (*you*). Other rules affect readability only and do not have any implications for translation quality. For instance, the CL rule on the restrictive use of passives does not influence translatability, as passives are not known to cause translation problems in the Western European languages.

Still other rules do affect translation quality, but only in an indirect way. An example is the CL rule stating that a sentence should not exceed 20 or 25 words (in procedural and descriptive writing respectively). Its beneficial effect primarily lies in the fact that shorter sentences generally make for better translations. They do so, not because of some specific grammatical property implied by the CL rule as such, but mainly because shorter sentences suffer less from the combinatorial effect of local ambiguities, and because they tend to contain less ellipses typical of conjunctions.

Finally, some rules directly affect translatability. Particularly in such cases consistency of analysis becomes a very important issue. An obvious example relating to translatability by machines is the CL rule stating that only approved words may be used in a text. If an English word is not known to the MT system, it will not be translated, and the overall translation quality will be impaired. A less obvious example is the CL rule prescribing that an adverbial subclause has to be separated from the main clause by a comma. In this rule the punctuation mark is assumed to act as a local disambiguator. However, the English analysis module is based on exactly this same assumption, with the effect that an adverbial subclause is usually not recognized if the comma is missing. Here, we are faced with the paradox of our self-imposed consistency: the parser used for translation is at the same time required to correctly diagnose and repair its own limitations. Some ways to deal with this paradox are described in the next section.

3. A CL approach to MT

From the very beginning it was clear that further development of our CL application would not be possible without changes to the LANT®MARK™ MT system. These modifications pertain to the following issues:

- The system should be able to cope with ungrammatical and semi-grammatical English input.
- Text structure information should be used to cover some CL rules.
- The link between the surface input sentence and the analysis tree should be preserved for error localization.
- The tool should allow for interactive checking in addition to batch-based processing.

By implementing these features for the CL application, the MT system was affected in two ways. On the one hand, most features proved useful for MT and were therefore taken over, be it for a different usage. On the other hand, new MT development was stimulated, further enhancing the system quality in CL-MT applications.

3.1 Grammar Extensions

In those applications where the conformance checker is conceived as a post-editing tool for document development, especially by non-native writers of English, some important extensions to the English analysis had to be made. Initially, both our MT and CL systems could only cope with correct grammatical input. An additional pre-editing tool, the pattern matcher, was, and still is, offered to clear incorrect input from spelling mistakes, inconsistent term usage and grammatical errors. However, in order to be able to offer support to non-native English writers ungrammatical structures need to be recognized and analyzed as well. This was done in the SECC project by both relaxing certain conditions on existing rules and by introducing new rules to the analysis module. This has far-reaching consequences for the system: On the one hand, the search space for analysis is enlarged, which, if not done with care, entails unwanted side-effects, such as combinatorial explosions and an increase in misinterpretations. But, on the other hand, if done with care, it enlarges the scope and coverage of the grammar both for conformance checking and MT applications. Indeed, from a user point of view the distinctions between non-controlled, semi-grammatical and ungrammatical English are irrelevant in terms of their effect on readability and translatability. Future developments at LANT will therefore concentrate on implementing intelligent solutions to the problem of incorrect input.

3.2 Grammar Reductions

In those applications where the conformance checker is conceived as a pre-editing tool for MT it had some development consequences for the MT components as well, in this case English-French. Controlled English input implied the possibility to restrict the application of some analysis and transfer rules. The motivation for this development lay in the assumption that knowing a given document to be conformant with a set of rules would allow us to achieve a better machine translation quality in a shorter period of time. However, the analysis grammar functionality should never limit itself to CL conformant structures only, because other types of text are translated also using the same English analysis module. Instead of implementing a distinct subgrammar for CL conformant input, we preferred to augment the grammar so that it is capable of dealing with either type of input. The device needed here was modeled on the concept of a switch triggering one of two operation modes of the same grammar, i.e. normal mode and controlled English mode.

Where the normal mode does not make any specific assumptions about the input, the controlled English mode allows the grammar to exclude certain (non-conformant) interpretations. With the controlled English mode switched on, the grammar accordingly behaves as a subgrammar, which may be expected to result in improving both translation quality and processing speed. Translation quality will profit from CL biased disambiguation, while the exclusion of non-conformant structures reduces the search space for analysis, and thus enhances translation speed.

The CL switch as implemented so far in our system has proven its beneficial effect in the following areas, further illustrated below:

- Filtering during morphological analysis
- Semantic disambiguation
- Lexical structural disambiguation
- Grammatical structural disambiguation
- SGML-label coded information.

During morphological analysis the CL switch allows straightforward disambiguation for those lexical elements for which less conformant readings are available than in the English lexicon as such. The basis for this disambiguation is the CL rule that only conformant terms may be used, so that all other lexical interpretations can be ruled out when the controlled English switch is on. Technically, this implied changing the system software of the morphological analysis component, so that it filters out all lexical analyses for which no corresponding entry exists in the controlled English lexicon. As a result, less lexical readings are produced by the morphological analysis, and the parse chart is initialized with less combinatory possibilities. The performance of the parser improves due to a reduction of the search space. But more importantly, only conformant lexical readings are produced, so that the analysis provides a more reliable basis for translation, resulting in an overall improvement of translation quality.

In some cases, the CL restrictions do not exclude a given lexical element in the input, but rather a specific usage or meaning of it. For example, we have implemented a CL rule stipulating that the subordinating conjunction "*since*" is acceptable when used in its temporal meaning (roughly synonymous with *from the time that*), but not when it indicates a reason (*because*). This proves to be relevant information in two ways. Firstly, it influences the analysis of the grammatical structure in which the lexical element occurs. Secondly, the difference in meaning is reflected in the translation itself. Thus, the French translation of "*since*" is either "*depuis que*" (temporal) or "*puisque*" (reason). The implementation of the CL switch makes use of a restriction in the English-French transfer lexicon, explicitly excluding the non-conformant translation "*puisque*" when the switch is active.

E.g., + The program runs on the Win NT 4.0 platform *since version 3.0 was released.*
- The program runs on the Win NT 4.0 platform *since it is 32bit compatible.*

Conformance of a single lexical element may also imply that it can only occur in a specific grammatical context. A typical example is the subordinating conjunction "*to*". It is allowed in the context of a complement clause, but it is not considered conformant when used as a conjunction of purpose.

E.g., + The menu allows the user *to select the next item.*
- The user should click left *to select the next item.*

The possible context of a subclause introduced by "*to*" can therefore be structurally disambiguated on the basis of the CL switch. Note that the example sentences may seem unambiguous enough to a human reader, but the system's bottom-up parser will have to maintain two alternative structures in either case. When the parser algorithm can rely on the CL switch, its search space is reduced and the chance of producing an incorrect translation diminishes accordingly.

An example of structural ambiguity where the CL switch contributes directly to disambiguation is the rule which states that a preposition needs to be repeated for all conjoined members in a prepositional phrase (PP). Given a conjunction of a noun phrase (NP) containing an embedded PP and another NP, there are two possible structural analyses. Either the conjunction is situated at the top level of the construction, or else the conjunction involves the NP inside the embedded PP.

E.g., + The owner of the gun and the detective
- The owner of the gun and the bullets
+ The owner of the gun and *of* the bullets

The CL switch forces the conjunction at the top level as the only possible analysis. This will prove to be an asset, especially when translating into French, because French normally requires that the prepositions "*par*", "*de*" and "*à*" are repeated in a conjunction. Accordingly, the French translation will have to reflect the structural difference between the example sentences: "Le propriétaire du fusil et le détective" vs. "Le propriétaire du fusil et des balles".

The situation where SGML labels can influence the parser is implied by the rule which states that all items of a bulleted list should be of the same grammatical category. Here, the CL switch would lead the parser to reject an analysis of a bulleted item which does not comply with regard to its grammatical category indicated by means of SGML labels. For example, the fact that "*transition functions*" occurs in an enumeration of noun phrases forces the parser to leave out its usually preferred sentence analysis in favor of a nominal interpretation.

The last example shows how text layout information can affect the analysis in general, regardless of a CL switch. Many CL rules are text structure specific, which led us to develop a device to pass over text structure information from converter to grammar (see also Adriaens 1996b). The information that *"transition function"* constitutes a title or heading has the same effect of ruling out a sentence analysis and favoring a nominal interpretation. In this way, both the CL and MT grammars are enhanced because text structure information can be used to disambiguate.

3.3 Input String Restitution

The conformance checker must preserve the link between the surface input sentence and the analysis tree in order to attach its diagnosis information correctly to the relevant parts of the input sentence. In the regular translation cycle, the analysis tree does not preserve this link: lexical elements are featurized, moved around and deleted, gapped elements are re-inserted in view of translation. For controlled English, we needed a tree structure on top of the original untouched input sentence. Therefore, all missing or moved input string elements are restored by using a table containing all partial representations created during the parsing process, and ordered by starting and ending indices w.r.t. the original input string. The restitution of the input string is to a large extent a heuristic process, and the final combination, structure and input string, deviates from the canonical LANT®MARK™ analysis tree. Roughly, it recursively descends the tree representation and checks the feature on the nodes that holds the key to the table containing the partial representations. Wherever the heuristic encounters "holes" between nodes, it restores the most likely missing partial representation; wherever it finds nodes whose indices fall outside the scope of the dominating node, it removes them. It also handles cases of raising and lowering constituents in a general fashion: if a node has already been attached higher in the tree than at a proposed restoration point, it is not restored; if it is present, but lower in the tree, it is restored, and its copy removed.

The procedure that restores the original input string can be re-used in the translation cycle when the overall sentence analysis failed. The LANT®MARK™ system has a fail-soft mechanism, called phrasal analysis, putting together the useful sentence parts for which an analysis could be found. This mechanism sometimes produces awkward translation results, particularly if drastic structural changes have taken place. In those cases the transfer phase should start from a combination of partial tree structures and the original input string. This should improve the translation results, even if one has to resort to simple transfer for all elements not covered by a node.

3.4 Interactive Checking

Another CL specific development relates to the introduction of interactive checking, next to batch checking. From a user point of view it should be possible to re-submit an individual sentence to the checker after it has been edited and it is deemed ready for re-insertion into the output text (cf. Adraens 1996a, Hayes et al. 1996, Hoard et al. 1992). Interactive checking is offered as a standard LANT®MASTER™ feature. However, it has become part of our MT environment as well, and has proven to be particularly useful in environments where the MT system is used as a tool for information acquisition. We think here of browser-based systems, where real-time word, phrase and sentence translation to and from several languages is provided as a non-native-language reading aid for web pages.

4. A TM approach to CL

Future investigations and experiments will be related to the use of a TM component in checking. It is well known that the use of an object-oriented database in which information objects are stored and linked to their translations constitute a serious structural and organizational change for large industries, but one that is worth the effort as it turns out that a lot of time and costs can be saved in the documentation development and translation process (see Godden 1998). The gain lies in the re-usability of information objects as compared to whole documents. When an information object that is linked to its translation is re-used, a new translation need not be provided and hence time and costs can be saved both in document development and in translation.

Even with such an organization of the document development process it pays to use a translation memory for all complete and fuzzy matches on sentence level. The advantage of a translation memory over machine translation is that it stores translations which have already been validated and post-edited by humans. Hence time and money can be saved by reducing human revision tasks to an absolute minimum. In the same vein, human revision tasks can be reduced by integrating an English-Controlled English translation memory into the conformance checker. Indeed, the sentence revisions proposed by the memory can be assumed to be more correct and complete than those proposed by the checker

In a further step, we would even think of re-designing the parser as a hybrid combination of both the grammar-based and the TM-based approach. Such a hybrid approach would allow us to test for subsentences instead of testing for complete sentences only. We believe that testing for complete or fuzzy matches with subsentences (combined with grammar) could be still more efficient than the approach taken now.

References

Adriaens, G. (1996a), SECC - A Simplified English Grammar and Style Checker/Corrector. Final Report. SNI, Liège 1996.

Adriaens, G. (1996b), SECC: Using Text Structure Information to Improve Checker Quality and Coverage. In *CLAW 96, Proceedings of the First International Workshop on Controlled Language Applications*, Leuven 1996, 226 - 232.

AECMA, The European Association of Aerospace Industries (1995), AECMA Simplified English. A Guide for the Preparation of Aircraft Maintenance Documentation in the International Aerospace Maintenance Language. AECMA Document PSC-85-16598 Issue 1, Brussels, 1995.

Godden, K. (1998), Controlling the Business Environment for Controlled Language. (This volume).

Hayes, P., S. Maxwell and L. Schmandt (1996), Controlled English Advantages for Translated and Original English Documents. In *CLAW 96, Proceedings of the First International Workshop on Controlled Language Applications*, Leuven 1996, 84 - 92.

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

Means, L. and K. Godden (1996), The Controlled Automotive Service Language (CASL) Project. In *CLAW 96, Proceedings of the First International Workshop on Controlled Language Applications*, Leuven 1996, 106-114.