Bringing Controlled Language Support to the Desktop

Drs Michiel de Koning

Abstract

This paper discusses the work being carried out at Cap Volmac on tools and services involving controlled languages and MT pre-editing. It focuses on the planning involved in implementing projects for authoring environments, based on experiences of the last 5 years.

It describes the steps to be taken for a successful implementation. These steps involve typical activities such as analysis, design, product selection, development, and integration, both on the IT and linguistic side.

The main focus of the paper will be on the following aspects: language specification (company specific, industry specific, application specific, generic), active intervention/ author support during document writing/editing and organisational requirements.

This paper will conclude a brief discussion of expected development, in terms of links to other automated (sub)systems, such as mark-up languages (structure of the document), document Management and Workflow Management (status of document), PDM (external data)

Drs Michiel de Koning

Mr. M. C. de Koning is project manager at Cap Volmac in the Netherlands. For the last three years, he has been involved in the development of tools and services based on the application of controlled languages for MT projects. Prior to that, he was active as a IT consultant on various projects.

Mr. de Koning has a Masters degree in Linguistics at the Utrecht University.

Cap Volmac

Active Documentation is a collective term for the consultancy, services and tools provided by Cap Volmac which support the production process of documentation. Active Documentation is part of Cap Volmac Team Support Technology B.V., which incorporates document production, document usage (retrieval, publication), document storage (DMS), groupware (Lotus Notes), and workflow automation (DIS/WFM).

Active Documentation focuses on content management and, more specifically, language- and text technology. Automated correction and machine translation systems, text structuring (SGML), terminology management and document conversion are a few examples.

Cap Volmac in the Netherlands is a member of the Cap Gemini Group, a world-wide IT consultancy and services firm.

Drs Michiel de Koning,
Cap Volmac Active Documentation
Daltonlaan 400
Pobox 2575, 3500GN Utrecht, The Netherlands

Tel: +31 30 2527127, Fax: +31 30 2527045
E-mail: MCdeKoning@CapVolmac.nl

Introduction

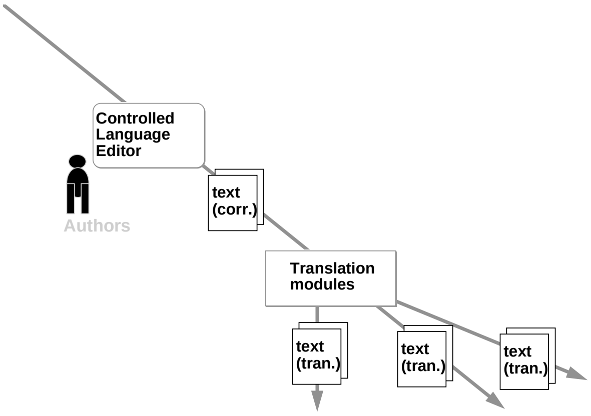
For the purpose of this paper, we will use a rather practical definition of controlled language. In terms of what we want to achieve with language control, a controlled language is the specification of a language that will improve processing of information (be it a human or machine) later on. Thus, a controlled language will most likely be implemented for the purpose of eliminating interpretation problems. As such, it can be seen as a kind of sieve, that allows only a subset of a natural language.

Interpretation problems may occur in different situations, real time (speech) or written (documentation). The typical effect of interpretation problems is that the recipient will check back with the originator to make sure he understood properly. In speech this normally is not a problem, but in delayed communication, this will take much more effort.

Controlled languages have been around now for a number of decades. Probably the best known example is Simplified English, used in the aerospace industry as a means to improve the 'understandability' of maintenance manuals, particularly for non-native audiences. This means that certain types of sentence structures, phrases or words are not allowed in maintenance manuals, because they may be understood incorrectly.

If we take this approach a step further, we could also use language control to improve performance of machine translation systems. Similar interpretation problems that, in this case, determine the performance of MT systems, may be reduced by providing input written in controlled language. This has been done for existing MT systems (Xerox Multinational Customized English), and is done for new systems that are optimized for the combination of controlled language input and machine translation.

When we look at using controlled language as a preparatory phase for using MT systems, we can draw the following figure:



Traditionally, MT systems produced output that had to be post-edited. When used in a specific domain and tuned to this domain, the amount of post-editing becomes less. We can further improve the performance of the MT system by pre-editing the input (according to specification of a controlled language). Considering that pre-editing only occurs once (in the source language), and post-editing must be done for all target languages, the extra effort of pre-editing may very well pay off.

Controlled language specification

The principle factor that must be taken into account for the specification of any controlled language is the purpose. Why do we want to control the expressive power and variation of language? If we look at Simplified English (or similar writing rules found in text books), some constructions are not approved so that people better or more quickly understand a written message. However, for MT systems, these same constructions may not cause any problems. Take, for example, the passive construction. In Simplified English, the passive may not be used in text segments that deal with procedures. An MT system does generally not have difficulty translating a passive construction.

Clearly, we can see that for different purposes, we can specify different types and levels of control. For this paper, we will take a closer look at MT systems, and the way in which language control can affect the performance of such systems.

Language requirements

Until now, MT systems are used in two ways: human aided machine translation, and machine aided human translation. The first is represented by systems that provide rough translations that must be post-edited, the second by systems that provide functions such as dictionary look-up or translation memory to a translator. Combinations of the two types of systems in one system have also been designed and implemented, such as a translation memory with a translation engine for unknown phrases.

Language control can be used to improve such types of translation systems. Let us take a look at one example. An MT system that would have to translate the sentence “Remove panels and open doors.” into French will most likely produce at least two translations: (1) “Enlever panneaux et ouvrir portes” or (2) “Enlever panneaux et portes ouvertes ”. If we make the input sentence a bit more specific, we can solve this problem. Assume we have a rule in our controlled language that we must use determiners whenever possible. We would then write our English sentence as (1) “Remove the panels and open the doors” or (2) “Remove the panels and the open doors.” Now, when we feed one of these alternatives to an MT system, we will get a proper translation.

When we want to specify a controlled language, we need to find out which grammatical and lexical phenomena must be controlled. This is not an easy task. We must also make sure that the communication style of the documents for which the language is targeted, is suitable for language control. For example, legislative texts often contain ambiguities to allow for interpretation by a judge. Clearly, this type of language usage should not be controlled. For that reason, most applications of controlled language are found in industrial documentation, where the primary objective is to pass information unambiguously. Examples are manuals, help texts, course material, et cetera.

The specification of a controlled language can be done in several ways. We can consider to create a controlled version of each existing natural language. But again, when we look at the areas where we want to apply controlled languages, a more focused approach can be more beneficial. Document types, as mentioned above, often comply to a sort of sublanguage, because of the nature of the document. On-line help text has certain characteristics, so have maintenance manuals. Furthermore, each type of industry (automotive, aerospace, IT) has its own language characteristics. At Cap Volmac, it is our belief that we should try to specify controlled language for specific applications, be it to improve communication or improve automated processing (or both).

The use of a controlled language has a large impact on the way documents are created (see also below). Because of this, we must make sure that the extra effort will be worth while. Improved performance of MT systems will reduce the amount of work done in the post-editing phase. This ‘balancing’ of pre-editing versus post-editing is important in specifying the rigorousness of the controlled language. For example, MT systems typically have problems with PP-attachment[1]. A control rule might be to not use any PPs at all. This is however hardly attainable in most types of communication. We have to consider whether to reduce the input (pre-editing) or leave it for the post-editing phase.

Specification in steps

If we choose to specify specific controlled languages for specific applications, we have to go through a number of steps to arrive at such a specification. Below the activities are discussed briefly as they are implemented in projects at Cap Volmac. As before, we do not include any aspects relating to readability, we focus mainly on the translatability. Readability is of course one of the implementation issues of controlled language that must be considered since the documents are to be published in the controlled source language, but is outside the scope of this paper.

The objective of these steps is to come to a controlled language specification that can be implemented in an automated system.

Purpose

We must first determine the purpose for which we want to use controlled language. Which target languages are involved (that affect the type of control in the source language), and which type of MT system —if any, language control can also be used to facilitate manual translation, particularly Western to Asian— will be used.

Document analysis

The next step is to get a good idea of the sublanguage as it is used in existing documentation (on paper or electronic). Three aspects are particularly important: lexicon (terminology), grammar, and style. In a representative set of documents, we first look at specific terminology used. This could be just nouns, but also expressions (phrases or verb clusters) that are only found in these documents.

Next, we look at grammatical constructions. All sentence structures are documented in global terms, e.g. S-V-O-(A), A-S-V-O, NP-V-O. Finally, in this phase, the type of constructions that are preferred (active over passive, adverb placement, etc.) are marked separately.

[1] e.g. He saw the girl with binoculars. Who has the binoculars?

These findings are discussed with the organization to come to an understanding on the language usage as it occurs (the sublanguage).

Analysis (descriptive) grammar

From the complete set as specified in the previous phase, we must arrive at what is presumably a smaller set that is tuned to the input requirements of processes to follow. Here, we much check the result of the previous phase with the requirements of the MT system. At Cap Volmac, we also develop translation modules which makes it possible to very precisely select the kind of constructions or terms that cause problems. When developing controlled languages for other MT systems, finding out the shortcomings is not a trivial matter. Some manufacturers even claim their system cannot be analyzed to that effect. In that case, general phenomena should be tested in the translation module to check whether they are processed correctly or should be avoided.

Correction (prescriptive) rules

Once the core specification of the controlled language has been determined, it must be considered whether specific correction rules for non-approved structures or words are included in the specification. This can never be a complete set of rules. Correction rules can be two-fold. Either they regulate the conversion of a non-approved form to an approved alternative, or they give hints on how to come to an approved alternative. Correction rules can occur for all three levels: style, grammar and lexicon (terminology).

Lexicon encoding

Based on the grammatical specification of the analysis grammar (the approved grammar) and correction rules, we can now start to build our lexicon. Lexical entries will require specific syntactic and semantic attributes to be able to check its validity. For example, if the word *display* may only be used as a noun, yet not as a verb, the lexicon must contain that information. Likewise, if agreement between subject and verb is required (which is very likely), we need number information for both nouns and verbs. The specification of the grammar will dictate the required attributes for the lexical entries.

To create a proper lexicon, all existing documentation must be analyzed. Apart from closed class items (prepositions, pronouns, etc.) at least all approved words must be encoded. Preferably, all non-approved words are included and marked as such, with a reference to an approved alternative. Compared to the work done for terminology management systems, all words are encoded, as opposed to primarily nouns and adjectives (typically 90% of terminology databases). Yet, from the point of view of a controlled language, less pragmatic information is stored, such as number of occurrences, context, parent-child relations, etc.)[2]

To make this work easier, at Cap Volmac we use concordance programs for this purpose (to determine the category of a word, a list of occurrences with a limited context is used). The number of entries required is roughly determined by two things:

[2] Unless of course you combine these two in one application

the type of application (approved words) and the proficiency of authors (non-approved words)

Testing

The specification must be tested. At Cap Volmac, we used a methodology similar to software development. Grammar rules must all be tested, to see whether their correction produces approved alternatives. A good practice here is to include examples for all rules, and use these examples to test. In the lexicon, the non-approved words must be checked. This can be done by generating short phrases from all entries in the lexicon and have these tested against the specification using the grammar. For example, assuming a noun phrase as a sentence is accepted in controlled language, you take all nouns from the lexicon and produce NPs by adding a determiner and closing with a period ('door' —> 'The door.').

All errors must be evaluated and corrected in an iterative process.

Implementation issues

Two major issues must be tackled when implementing a solution based on controlled language.

The first is that documentation in controlled language must be accepted by the audience, the users of the information. Again, we believe communication in controlled language can only be implemented for industrial documentation. In the case of Simplified English, it was a regulation for the entire industry to comply to this controlled language. In a more consumer oriented environment (e.g. consumer electronics) it may not always be possible to start using controlled language if this will affect the consumer's perception of the product in a negative sense.

The second is that the people involved in the document creation process must be willing and capable of accepting such a change in the way they work. In our experience, larger organizations with a number of professionally trained authors will make this transition much easier than smaller companies with only one or two authors. This is mostly due to the fact that larger organizations have already begun to implement Quality Assurance programs and are ready to accept the consequences (less 'freedom', procedures) of the introduction of controlled language.

When a controlled language is introduced in the authoring environment, the author must be ready to adopt the new way of working. This will involve training. Not only in working with new tools (see below), but also learning to write in controlled language. Part of this can be done in a training course, but the author will have to go through a learning curve. Basically, he/she has to learn a new language. It turns out only full time authors are able to come up to speed in writing in controlled language; it is not feasible for people who occasionally write documentation to learn to write in this way.

The implementation of controlled language as pre-editors for MT systems involves another issue. Apart from the fact that a company will have to think of which documents will be written in controlled language (to implement a certain style of communicating with its customers), it must also be decided which existing documentation must be rewritten. Typically, this is done on an on-demand basis. If further processing of the document (translation to a new language, changes or updates, etc.) requires conformance to the new controlled language standard, the document must be rewritten.

When controlled language is introduced in an organization, this will also introduce new tasks regarding the maintenance of resources dealing with the controlled language specification. Most importantly, the (company specific) terminology or more generally the lexicon (in case of automated controlled language support) must be kept up-to-date. This is a very important job, since it will manage the style of communication (consistency, preference, etc.) of an organization.

Other resources, such as grammars, when built right, tend not to change. Building them requires specialist linguistic knowledge not found in industrial organizations. So this remains the responsibility of the supplier of the tools (when relevant) or the regulatory body (in case of industry wide specifications).

Organizations quite easily adopt new tasks, tools, and responsibilities, when they are directly related to the existing processes. In this perspective, changes to the document creation process can be implemented properly. However, when it comes to translation, they prefer to keep this as it was, in the hands of subcontractors. This means that the introduction of controlled language to improve translation efficiency, should not imply organizations must also start to use and manage MT systems themselves. A good way to deal with this is to offer a package deal. When an organization delivers its documentation in controlled language, specialized translation agencies can provide high quality translation with an improved efficiency (through the use of MT tools). The translation agency manages the MT system and has the capability to verify the quality of the output.

User support and interaction in authoring tools

When a company decides to implement a controlled language, a number of methods and/or tools can be used to help authors in their activities. We already mentioned training of authors, because they must learn the 'new' language.

The most elementary support is to have all related documentation (the controlled language specification, word lists, guide lines, do's and don't's) on paper available on the desktop.

The next stage would be to make this information available electronically.[3] Depending on the platform, this can be done in an on-line help, hyper-linked style. A major advantage here is that it is easier to manage updates to the specification. This type of support can exist independently of the actual authoring environment (word processor, DTP-package).

More advanced solutions can be provided, but these will require integration with the authoring environment. The most common application is a spell checker, that is tuned to the CL specifications. Obviously, this will only provide support on part of the specifications, namely the lexical requirements (most likely it will not be able to do this task properly when lexical specifications are context sensitive, c.f. use 'display' only as a noun, otherwise use 'show')

[3] One issue to be taken into account when starting to make things electronic, is distribution of resources. If an organization does not yet have a means of sharing resources (no network, people working away from the office), it generally is more difficult to make sure everyone is using the same (version of the) resources.

At Cap Volmac, we develop tools that can be best typified as grammar checkers. These are integrated in the authoring environment, and are able to give the author support for the lexical specifications, the grammatical specifications, and style. The author can use the functions of this tools in the document itself to check whether the text is in accordance with the specification. The author will get help and suggestions, if the text contains non-approved elements).

For this type of tool, it becomes more important to manage the resources properly. Again, in operational terms this often concerns primarily the lexicon, as this is subject to change, where the grammar typically is not. Authors should be able to continue there work, even if certain terms may be yet unknown (e.g. a new product name). But, when language control is used in an MT environment, any updates to the source language must first be verified against the possibilities of the MT system[4].

Future developments

When we have augmented the author's work place with tools as described above, it is tempting to take a look at what kind of information is available electronically and might be used to improve or add functions to support the authors. Below are some examples we might want to consider.

Apart from text, documents contain more information, such as layout and/or structure. This structure information can be used to improve grammar checking. For example, headers generally do not have determiners, whereas you might prefer determiners (for disambiguation, as we have seen above) in running text. You can then trigger certain rules, depending on type position in the document (the type of text).

We might also consider to make the authoring environment more 'interactive'. Consider we want to use a term that is ambiguous, yet we cannot rewrite it because it is accepted and preferred by the audience. We could ask the author to indicate which interpretation is meant, store the information (annotation) and use that again when the document is processed in an MT systems. Another possibility is to work with pronoun resolution --what does a pronoun refer to in a text-- this way. The problem of PP-attachment mentioned earlier could also be handled along similar lines. In simple terms, this would change the author's job from writing to storing knowledge.

Finally, we can use information available in other automated systems (CAD/CAM, MIS, DIS, EIS, STEP ..).In these systems, we might find information that could help the author verify more pragmatic aspects. Suppose we are able to distill and model information in or from these systems that tells us a bit about the world. We could then start checking on pragmatic correctness of documents. For example, an author may write: 'Attach module XYZ to unit ABC.'. If we are able to check, in the CAD/CAM specification, that indeed this module should be attached to that print (by

[4] This is not a trivial matter. Consider for example an MT system that cannot handle category switching, such as:

Dutch:	Hij wandelt snel (He walks fast)	Hij wandelt graag (He walks 'likingly')
English:	He walks fast	He likes to walk

Although we want to be able to use constructions on the left (in Dutch), we want to block some adverbs (such as 'graag'). This means we always have to make sure that we can translate what we allow.

design), we can signal this to the author, or issue a warning when it does not match the design.

Conclusion

Controlled language is a useful instrument in optimizing the efficiency of the translation process, especially if it uses MT systems. For the specification of a controlled language, we must look very closely at the requirements of the MT system, and we must make sure the specification is complete in the sense that is must be consistent and must properly describe the approved language. The authoring environment can be upgraded with additional tools that help the author in writing in controlled languages. The resources used by these tools will introduce new resource management activities, mainly in the area of terminology or the general lexicon.

Organizations are not anxious to take on the whole issue of translating their documents. They prefer to have this done by subcontractors. Therefor we must look for solutions where translation bureaus provide translation services that are in key with the use of controlled language by their customers.

The adoption of controlled language will make this relationship very smooth since its major objective is to eliminate interpretation problems.