

Controlled Language Technology in Multilingual User Interfaces

Aarno Lehtola, Catherine Bounsaythip, and Jarno Tenni

VTT Information Technology

P.O.Box 1201, FIN-02044 VTT, Finland.

Phone: +358 9 456 6032. Fax: +358 9 456 6027.

Email: {aarno.lehtola, catherine.bounsaythip,
jarno.tenni}@vtt.fi
<http://www.vtt.fi/tte>

Abstract. This paper studies how controlled language (CL) technology can be used to build multilingual user interfaces to information services on the WWW. Controlled languages are domain specific sublanguages that resemble human languages but have limited vocabulary and restricted syntax. Control is used to minimise ambiguities in the texts and enable their automatic processing. For instance, automatic information extraction and machine translation are difficult tasks when dealing with unrestricted natural languages. Language control enables practical and cost-effective solutions to these tasks.

In our paper we first outline the state of the art in CL technology. After that we consider the use of CL in fully automatic translation of contents of a monolingual text database and in information retrieval from a multilingual information base.

1 INTRODUCTION

The Webtran project aims at providing tools for building multilingual WWW services. These tools are based on the idea of using controlled languages instead of unconstrained natural languages. A controlled language is a language limited to a specific semantic domain, with a specifically selected vocabulary and simplified syntax. One of the central goals in the design of controlled languages is to avoid the many ambiguities that are present in ordinary natural language texts. These ambiguities usually cause most of the problems for machine translation and automatic language interpretation. It has been found in practice that controlled languages also reduce misunderstandings of human readers and this way intensify the human-computer interaction.

An example of a limited language is the one used to describe properties of women's clothes and accessories in postal sales catalogues. However, texts of advertising style that rely upon the readers' imagination, are beyond the scope of language control. Another common example of a controlled language is the language in weather reports.

General purpose machine translation systems always need human resources to produce reliable high-quality output. However, by controlling the expressions of the original texts reliable *fully automatic machine translation (FAMT)* can be reached. Moreover, information resources produced in controlled languages are applicable for automatic interpretation, e.g., in context of data mining and knowledge discovery. Also information retrieval command languages could benefit of controlled language technology and bring their users new flexibility by positioning themselves between formal languages and human languages.

2 CONTROLLED LANGUAGE TECHNOLOGY HELPS BUILD MULTILINGUAL INFORMATION SYSTEMS

The dictionaries in general purpose MT systems may be very large and still miss the domain-specific definitions needed. On the other hand, a large vocabulary causes extra troubles (like polysemy, homographs), and it is difficult for a general purpose system to choose the right interpretations. Word interference may lead to total changes in the meaning of the interpreted text. Obstacles stem also from the inability of systems to interpret syntactic construction correctly. Successful interpretation and translation often needs deeper understanding of the semantics of the input, unless the domain of the texts is very narrow. Altogether, fully automatic high-quality translations are beyond the scope of general purpose MT systems. In fact, the general purpose systems usually necessitate post-editing where human translators check and correct the results. Human involvement can also be made at the pre-editing stage in order to remove word ambiguities and simplify difficult syntactic constructs. For this purpose, language control is designed.

The use of controlled language aims at adapting the source text to syntax and vocabulary which the system can deal with accuracy. This can be done when the domain of the texts is restricted.

Controlled languages are simplified versions of natural languages. Simplifying is done both at word and grammar level. The main goals are to remove ambiguities, simplify sentence structures and so make automatic processing easier, but at the same time retain the readability of original texts. Below are examples of rules, that have been provided for technical authors to help in writing in AECMA Simplified English [dE]:

RULE: 2.1 Do not make noun clusters of more than three nouns.

RULE: 4.1 Keep to one topic per sentence.

RULE: 5.2 Write only one instruction per sentence.

RULE: 6.1 Keep sentences in descriptive writing as short as possible.

RULE: 9.3 Use the Dictionary correctly to get the correct words meanings, and parts of speech.

A use example: "Close" is a verb (and not an adverb).

WRITE: Do not go near the landing gear if ...

NOT: Do not go close to the landing gear if ...

The use of controlled languages is growing in multilingual information systems. There has been already two international workshops dedicated to their applications (CLAW'96 and CLAW'98). The approach has been successfully used to enhance the quality of translation [Kit87], and also readability, understandability, and maintainability of the original texts [WWWa, WWWb]. Examples include the TITUS system which was designed for storing and translating abstracts on textiles from French to English, German and Spanish [HS92]. In it syntactic structures were very limited in addition to some further restrictions. One-to-one mapping is also defined onto the equivalent structures in each supported language [HS92]. Also, the success of TAUM-METEO is explained by the restricted vocabulary and telegraphic style syntax used in the weather forecast bulletins [WK95]. Simplified English is used by AECMA [dE] and

others [AM95, DH96, SF96]. Currently Scania [WWWc] is implementing ScaniaSwedish for the preparation of truck maintenance manuals in a controlled Swedish [AH96, Hei97].

3 WEBTRAN SOFTWARE BRINGS CONTROLLED LANGUAGE IN WWW SERVICES

This section outlines Webtran Software, which we provide as a generic building block for embedding controlled language processing in multilingual WWW services.

Users who interact with Webtran Software can be classified in the following categories:

Controlled language specification designers, e.g. professional translators, who will use specification tools of Webtran Software to define, verify and test language specifications.

Contents editors, who will use Webtran Software to check syntactic and/or semantic admissibility of inputs while entering them to be ready for translation. This can be for instance editors of text databases or even end-users themselves.

End-users, are the people who read the translated texts through the WWW Information service systems which embeds Webtran Translator.

Webtran Software consists of two major parts: a specification part (*Webtran Modeller*) which is used by designers of a controlled language and a run-time part (*Webtran Translator*) used by editors and service end-users (see Figure 1). In between these two there are the controlled language specifications that are presented in *Augmented Lexical Entries (ALE)* formalism, which is described in [LTB98]. The grey arrows between the modules denote data flows while the black arrows mean request-response type of function invocations.

Modeller has basic text editing properties and special features for automatic alignment and synchronised viewing of sample translations, generating and testing specifications. Translator is used for checking the input source texts and for performing their translations. Given language specifications, Translator can check both semantic and syntactic correctness of input data. Translator can also be used online, separately from Modeller. It can be embedded into information services in order to provide seamless translations. Specifications include conceptual model and the syntactico-semantic language specifications stated as Augmented Lexical Entries.

4 AUTOMATIC TRANSLATION OF CONTROLLED LANGUAGES IN INFORMATION SERVICES

Many WWW sites contain language that is almost controlled. Only little extra effort would make their language fully controlled and bring the benefits of automatic processing. Such language can be found, e.g., in mail-order catalogues, in which the product descriptions often have a controlled nature.

We are currently testing the Webtran Software in providing multilingual views to product descriptions of women's clothes in a WWW-based mail-order catalogue. The

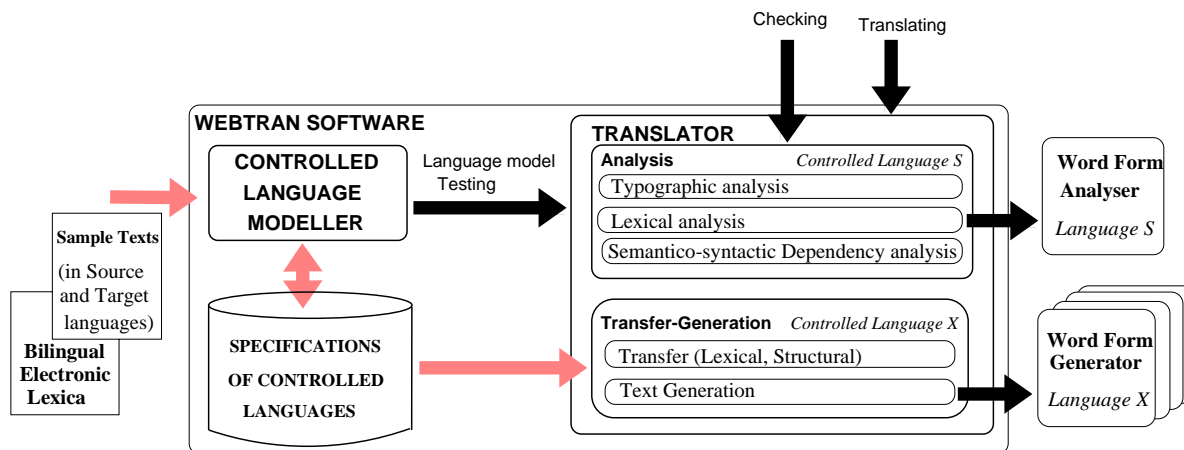


Figure 1: Webtran Software - a building block for multilingual WWW services.

product descriptions will be maintained in only one controlled language (sublanguage of Swedish). End-users get their translated descriptions through the Information Service in the language of their preference. The first target language is Finnish. Figure 2 shows the way controlled language processing is to be embedded in multilingual WWW catalogue service.

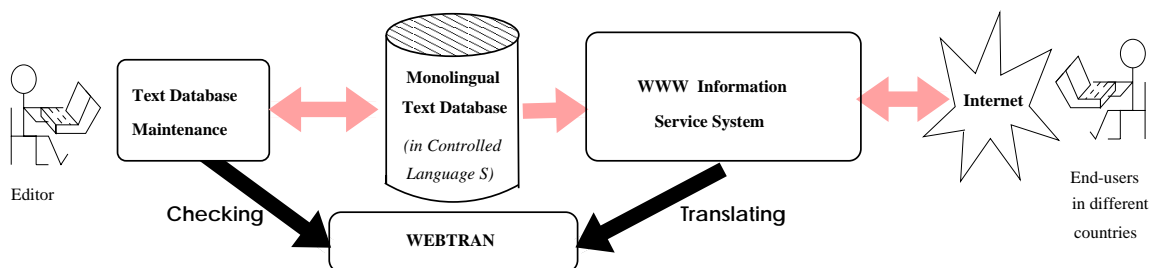


Figure 2: Webtran Software in a multilingual catalogue service.

The editor who maintains the text database of product descriptions uses Webtran Translator to check the syntactic and semantic admissibility of the input product descriptions. The check result is returned to the database maintenance program and the editor may be required to do some modifications to the text before it is accepted and stored into the database.

The end-user accesses the text database of product descriptions through the WWW information service system. Webtran Translator which is embedded into the service process gets the product description and translates it into the user's language.

5 CONTROLLED LANGUAGES IN INFORMATION RETRIEVAL

The use of controlled languages will also be tested for cross-language information retrieval in WWW environment. The user can make a query in one language to search for texts from different databases maintained in different languages. It is essential that these databases deal with the same quite narrow domain. In this case the query language will be controlled consisting of terms and expressions from the domain. It will be designed

specifically for the target domain. The query is first translated by Webtran Translator before directing the search to the databases. The found results are partially translated (e.g. headers only) by Webtran, into the language in which the user has chosen to view the texts.

Texts databases in piloting belong to the legislative domain. The first part of the project deal with taxation and Value Added Tax laws in Europe. The objective is to make queries e.g. in Finnish or Swedish and output results in English. Example texts will be provided by the Finnish and Swedish Parliaments. The potential user groups are broad covering, e.g., local and central governments, industrial lawyers, libraries, ordinary citizens.

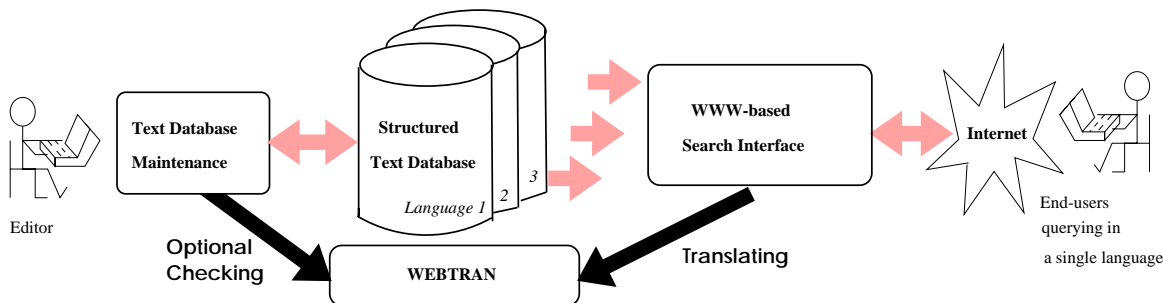


Figure 3: Application of Webtran system in multilingual information retrieval from multilingual databases.

At the beginning the controlled query language will be defined by experts of legislative domain and computer specialists. Then end-users, who have now the role of entering input, use a WWW-based search interface to enter a query. Search engine communicates with Webtran Translator to check the input. If the input is not accepted, the user is asked to modify the query sentence. When the query is accepted, it is translated by Webtran Translator. Then the search engine sends translated query to the legislative databases and gets results in their original language. Before showing results to the end-user the search engine sends headers of the documents found to Webtran Translator, which translates them as much as possible. As those headings are not in any controlled language translation accuracy cannot be guaranteed, but as headings usually have quite simplified structures it can be assumed that translation quality is mostly acceptable. Then those (possibly partially) translated headings are shown with original form to the user in his/her own language. The user gets a good indicator of what documents are interesting and good for further study.

Controlled query languages require training of the users. We will test next year this approach with end-users of legislative databases. This will reveal how hard the learning is for the users and what are the gained benefits.

6 CONCLUSIONS

This paper described two ways to use controlled language technology in building multilingual WWW services targeted for wide user groups. This far we have been implementing this technology in a multilingual catalogue service. This work is still going on. During next winter we will also apply it in a multilingual information query service.

While developing the controlled language software some new research topics have risen concerning automatic language model acquisition, rule generalisation, language specification validation, and automatic estimation of final translation quality.

Acknowledgements. The authors would like to thank the partners of the project, namely the Technical Development Centre of Finland (TEKES), Tieto Corporation Ltd., and Ellos Ltd., all from Finland. Also, many thanks to Prof. Timo Honkela and Prof. Seppo Linnainmaa for commenting this paper.

REFERENCES

- [AH96] Ingrid Almqvist and Anna S. Hein. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. In *Proceedings of the 1st Int. Workshop on Controlled Language Applications, CLAW'96*, pages 159–164, KU Leuven, Belgium, 1996.
- [AM95] Geert Adriaens and Lieve Macken. Technological Evaluation of a Controlled Language Application: Precision, Recall and Convergence Tests for SECC. In *The Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 123–141, 1995.
- [dE] Pim Van der Eijk. Controlled Languages in Technical Documentation. Cap Gemini ATS, <http://www.uilots.let.ruu.nl/www/Controlled-languages/Doc/capgemini.html>.
- [DH96] Shona Douglas and Matthew Hurs. Controlled Language Support for Perkins Approved Clear English (PACE). In *Proceedings of the 1st Int. Workshop on Controlled Language Applications, CLAW'96*, pages 93–105, 1996.
- [Hei97] Anna S. Hein. Language Control and Machine Translation. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-97)*, Santa Fe, USA, 1997.
- [HS92] W. John Hutchins and Harolds L. Somers. *An Introduction to Machine Translation*. Academic Press, 1992.
- [Kit87] Richard I. Kittredge. The Significance of Sublanguage for Automatic Translation. In S. Nirenburg, editor, *Machine translation: Theoretical and methological issues*, Studies in Natural Language Processing, pages 59–67. Cambridge University Press, 1987.
- [LTB98] Aarno Lehtola, Jarno Tenni, and Catherine Bounsaythip. Definition of a Controlled Language Based on Augmented Lexical Entries. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW'98)*, pages 16–29, Pittsburg, Pennsylvania, USA, 21-22 May 1998. Language Technologies Institute, Carnegie Mellon University.
- [SF96] Rolf Schwitter and Norbert E. Fuchts. Attempto Controlled English - A Seemingly Informal Bridgehead in Formal Territory. In *Proc. poster session of JICSLP'96*, Bonn, Germany, September 1996.
- [WK95] Peter Whitelock and Kieran Kilby. *Linguistic and Computational Techniques in Machine Translation System Design*. UCL Press Limited, second edition, 1995. Harold Somers ed.
- [WWWa] ClearCheck Controlled English Checking. <http://www.cgi.com/web2/cis/newtext.html>. Carnegie Mellon Group, Inc.
- [WWWb] LANTMASTER. <http://www-uilots.let.ruu.nl/www/Controlled-languages/Doc/lant.html>. Authoring and checking software offered by LANT (Controlled Lnguage Organizations), can be combined with the MT system LANTMARK.
- [WWWc] Scania WWW Page. <http://strindberg.ling.uu.se/~ corpora/scania/>.