Webtran Tools for In-company Language Support

1. Introduction

Webtran tools for authoring and translating domain specific texts can make the multilingual text production in a company more efficient and less expensive. The tools have been in production use since spring 2000 for checking and translating product article texts of a specific domain, namely an in-company language in sales catalogues of a mail-order company. Webtran tools have been developed by VTT Information Technology.

Use experiences have shown that an automatic translation process is faster than phrase-lexicon assisted manual translation, if an in-company language model is created to control and support the language used within the company. Company benefits from defining of such a model as

- source texts will have a uniform and explicit way of expressing information.
- source texts with systematic use of terminology and linguistic structures can be analysed and translated automatically with minimal post-editing.
- company needs to maintain only one pivot version of the source text facilitating a variety of ways for multilingual publishing.

According to the use experiences of Webtran tools, post-editing of automatic translation results only needs a minimal amount of human resources, especially when compared to entire process of manual translation. Depending on the degree of control exerted on the original pivot documents, the post-editing may be totally avoided. This enables a cost-effective way of providing multilingual views to text bases on WWW-services, like e-commerce systems. An in-company language model also constitutes central knowledge property in the company, when authoring and translation tools are embedded into the everyday document production process.

Webtran tools include a Language Modelling Tool, an Authoring Tool, and a Translation Engine. Language Modelling Tool is used for creating the in-company language model. Authoring Tool controls the use of terminology and linguistic structures according to the in-company language model. In the final phase, multilingual publishing is realised by using the Translation Engine to translate texts according to the translation rules defined with Language Modelling Tool. Translation Engine can be seamlessly integrated to a WWW-service or be used as a back-office tool integrated to the document production process.

In this paper, we first present the Webtran system for multilingual publishing and characterise the in-company language properties, that ensure the benefits of Webtran system. Then we detail how the modelling of an in-company language is carried out and how the language model is interconnected with the authoring process of pivot documents and multilingual publishing. In the end we summarise the above with use experiences of Webtran tools.

2. Webtran System for Authoring and Translating Controlled Sublanguage

Webtran is a generic authoring and translation software for multilingual publishing of domain specific texts. Webtran publishing process consists of 1) definition of an in-company

language model, 2) authoring of a pivot text in one base language in a controlled manner, and 3) automatic translation of the pivot text into multiple target languages.

The Webtran approach to multilingual publishing uses one base language, the in-company language, for maintaining of all source texts. Source texts need to be produced only as one pivot version, which then is a basis for automatic translations in several target languages for several situations. In the current use, translated texts are re-generated and re-used multiple times, e.g. as a printed sales catalogue, as varying collections of products in different forms of advertising or as translations in an on-line sales catalogue.

Webtran is a language independent system with no limits on the number of different languages. The basic way of use is that there is one source language which then is translated into a desired collection of target languages. Webtran design is based on the assumption that the processed language is restricted to a specific domain. This way it is possible to take full advantage of the natural restrictions of domain specificity and automatic translation can be accurately based on a limited set of pre-defined translation relations and linguistic structures (compared to automatic translation of general language).

Domain specific texts are characterised by a restricted domain and semantic context and therefore also a restricted use of terminology. These kinds of restrictions contribute to a simplified language, or 'domain specific sublanguage', with minimal ambiguities on lexical or syntactic level (Lehrberger 1982). A specific domain provides the semantic context for correct interpretation of possible ambiguous terms or constructions and that is why domain specific texts are especially suitable for automatic processing, since there exists only a limited number of possible translations that can be deduced on the basis of the domain context. Besides the semantic restrictions, domain specific texts often include elements and expressions that are repeated in the same form over and over again. In sales catalogues there are, for example, lists of products, product codes, colours, measures, and so on, always repeated in the same form.

Sublanguages naturally include restrictions to enable an efficient communication in some field of expertise, with precise and exact expression to avoid any ambiguities. These "natural" sublanguages usually give good results in automatic translation due to the restrictions set by the practical needs that arise from the situation where the sublanguage is used. To increase the efficiency of text production, sublanguages can be artificially controlled to minimise linguistic variety in texts and at the same time to maximise accuracy and clarity. Controlled languages have been widely used to improve the quality of text production, as the texts become more precise, readable and easy to maintain and update. Most importantly, controlling of the source texts enables automatic text analysis and translation (Arnold 1994, van der Eijk 1996, Hutchins 1992).

Webtran takes advantage of both the natural restrictions of a sublanguage and, in addition, also the possibility to control the source language. In other words, it is possible to make a description of a naturally restricted sublanguage and define the way it is to be translated. Furthermore, it is possible to control the language in the authoring phase which even more increases the exactness and enables an efficient automatic analysis and translation of the source text.

Another advantage in using Webtran is that it places no specific requirements on the language to be authored and translated (other than the restricted nature of domain specificity). The whole Webtran system can be adapted to the individual sublanguage of the company and it is

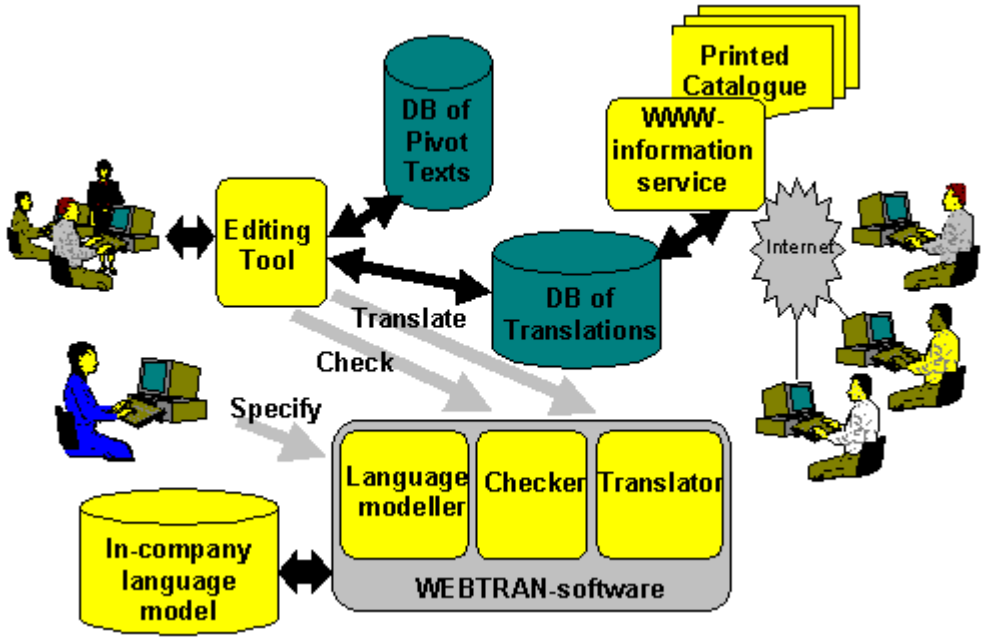
possible to adjust the level of control and exploitation of automatic translation according to the document production process of the company.

2.1. Generic Architecture

Integration of the Webtran system to in-company text production depends on the asserted goals. First, there is the choice of exerting strict control over the in-company language which enables fully automatic and accurate machine translation. This is the case when on-line translation is needed in WWW-services. On the other hand, there is the case where Webtran is used as a human-assisted machine translation tool. In this case, the in-company language can be less restricted and Webtran is used as an back-office tool.

Webtran system consists of three tools:

1. Language Modelling Tool: definition of the in-company language model (described in 3).
2. Authoring Tool: embedded in the text production process to control that texts are written according to the in-company language model.
3. Translation Engine: translates authored texts automatically, also embedded in the text processing program.



3. In-company Language Model

In-company language model is the basis for the text production. A language model includes 1) multilingual lexicon, 2) writing rules describing the allowed structures of the in-company language, 3) checking rules to control lexicon and writing, and 4) translation rules defining the relations between the source language and target language constructions.

Different sections of the language model are interdependent and therefore it is necessary that the modelling work has a responsible manager who ensures the consistency between rules and definitions. This is necessary also because working phases of the modelling are integrated and

simultaneous, which means that the modelling work can not be divided into clearly separate sections. Training of the responsible main user consists of learning the every day use of Language Modelling Tool and comprehending how its operation is interconnected with other tools in the system.

Modelling work requires both linguistic and translation expertise as the definition of writing rules and translation rules includes syntactic and semantic analysis of the languages to be authored and translated. Definition of terminology and its translation requires also deep knowledge of the company domain and the sublanguage to be modelled. Creating of a language model from the scratch is a full-time job even for an expert on linguistics. Depending on the time invested in modelling work and the level of requirements set by the company language, it takes a couple of months to build up a functioning foundation for the language model. It can then be edited and adapted along the implementing of the rest of the Webtran system.

Depending on the background and expertise of the people to be trained, the introduction to the system takes a full-time attention of the main user from one to a couple of weeks. This is a lot more than it takes to introduce the system to the translators and other users who do not have to deal with the language model as they only need to be able to use the Webtran system embedded in the company programs. A basic introduction gives the capability to start using the system, detailed and deeper knowledge comes along, as the language model is developed and the tools are fully adapted to the company process.

3.1. Language Modelling Rules in ALE-formalism

In the language model, writing rules, checking rules and translation rules are presented in form of ALE (Augmented Lexical Entries). ALEs are multidirectional entries with equal, non-directed language excerpts on desired linguistic levels. Entries can describe the linguistic information in one, two or more languages. In an entry, each language is represented in its own section. Below is an example of an ALE, for technical and detailed description of the ALE-formalism, see Lehtola et al. (1999).

Example 1. Translation rule concerning a phrase with product name and the material that the product is made of with the material percentage. For example: 'trousers of 100% cotton' (en), 'housut 100% puuvillaa' (fi), 'byxa av 100% bomull' (se).

```
[Cloth.material.001
[se (A){product} av tag_percentage(X) (B){material}]
[fi (A){nom} tag_percentage(X) (B){ptv}]
[en (A){nom} tag_percentage(X) (B){nom}]
```

The advantage of the ALE-formalism is that there are no restrictions to the contents of ALE-rules. Depending on the needs of the company and the requirements that the sublanguage places, writing rules can be applied to single surface expressions, for example some idiomatic expressions that are often repeated in the company texts or, in addition to the surface level, ALEs can also represent more abstract phenomena. Phrases and sentences that recur in same semantic or syntactic form can be generalised and represented by an ALE that describes all the phrases and sentences with same construction. This way, there is no need to separately describe all possible surface forms in a language, but instead, similar constructions are grouped and presented in one general ALE. Different levels of generalisation can be

combined in one and same ALE, and the rules can refer to other rules to cover more complex structures. Generalisation and recursive rules make it easier to describe all kinds of sublanguages, and they save both the amount of modelling work and number of rules in one language model.

The generalisation of linguistic expressions can be done on many different levels. Constituents can be generalised, e.g., according to their semantic role in the language or by their grammatical or syntactic status (Lehtola et al. 1999). Semantic classification is a practical way of distinguishing phrases and sentences with same syntactic structure but different semantic meaning or translation. Use of semantic features is also easy to adapt, as there is no need for learning some formal language to describe the constructions, merely the understanding of the language and its semantic structure.

3.1.1. Writing Rules

Writing rules are a description of the sublanguage used in the company text production. The idea is to describe the allowed sentence and phrase structures in the language as simply as possible. This is possible due to the nature of controlled sublanguage, where terminology and textual expressions are exact and the range of different interpretation alternatives is limited. The role of restricted sublanguage is highlighted here, as it is obvious that an entire general language description would be impossible to accomplish.

Writing rules contribute to the exact way to express information in texts. Besides the aesthetic value of consistent expression, it is then easier to create translation rules that relate only to these allowed sentence structures. Otherwise there would have to be just as many translation rules as there would be ways of saying the same thing by different authors. By comparing authored text to these writing rules and with help of the checking rules, the Authoring Tool can perform the language check to ensure that the text is written according to the language model.

3.1.2. Terminology Definition

Language model also includes a lexicon of the terminology to be used in the company texts and translations in multiple target languages. When the language model is created all the way from beginning, the terminology definition can be based on general bilingual lexicon that is defined and completed with domain specific terminology. Depending on the target languages, there might occur some troubles in finding exact translation for each source term. Though, definition of the terms and their translations in Webtran confirms to the idea that the terminology used in the company texts is controlled and reduced, so that each term should have the most exact and concise translation as possible. If there is no explicit solution to a problem with multiple translation alternatives, there is always the possibility of using the semantic classes and ALEs to distinguish the use cases with different meanings form each other.

3.1.3. Checking Rules

Correctness of the terminology and phrase structures is checked by comparing the source text with the writing rules and by executing checking rules. Checking rules are ALEs with sections for incorrect and correct constructions. Replacement of the invalid expression can be done automatically by the Authoring Tool or by presenting some repair instruction for the author to fulfil manually. Like writing rules, checking rules can also cover different levels of abstraction. One checking rule can, for example, include a list of forbidden terms in a surface form and a correct term to replace the invalid ones. Checking rules can also identify longer expressions, phrases and sentences, if there are some specific invalid constructions that are often used and that could cause ambiguity and misinterpretation.

3.1.4. Translation Rules

A set of translation rules is defined to match the description of the in-company language and to conduct the automatic translation of the authored texts. Translation rules can, like the other ALEs, describe terms, phrases and sentences on different levels of abstraction. In an entry, each language has its own section and the sections represent corresponding text excerpts seen as translation equivalents. Language sections are equal and the translation can be generated in all directions, simply by defining one of the languages as the source language and the others as target languages. Choice of source language can be changed within the Webtran system, without touching the actual translation rules. A simple translation rule can, e.g., present a list of corresponding surface form terms or constructions in each language. Again, with help of semantic classes, specific types of terms or a larger set of different surface constructions can be generalised and translated according to one generalised rule.

4. Webtran in the Translation Process at Ellos

Webtran tools have been in commercial use since spring 2000 in the mail-order company Ellos Postimyynti Oy which is a Finnish subsidiary of the Swedish mail-order company Ellos. At Ellos Postimyynti Oy, Webtran tools have been used in checking and translating product articles from Swedish into Finnish.

The main emphasis in using Webtran has been on translation of sales catalogues that are originally produced in Swedish. Catalogues are translated into Finnish both for printed sales catalogues and on-line catalogues available on the Internet. Besides the role of the translation engine, Webtran has been a practical tool for terminology management in form of bilingual lexicon and checking rules that conduct and control the use of different terminological variants.

Company texts at Ellos are produced originally in Swedish and translations are then made in several countries into their local language. Each country has its own, independent translation process and besides translation, the original texts are modified to correspond the expectations of local market and customers. This way, Ellos text production process is ideal for use of Webtran tools as there is only one source language which then is translated into multiple target languages and for multiple ways of publishing the texts.

Product articles in Ellos catalogues are short text descriptions with basic facts about the product properties, materials, colours, parts on the product, and so on. Terminology in the catalogue texts is restricted. According to the product group, information is presented with domain terminology that most effectively transmits the information to the customer. There is a limited space for each product article, so expressions must be as short and exact as possible. Same format is repeated in presenting sizes, colours, materials, product codes, and other recurring elements throughout the catalogues.

The basic goal in using Webtran tools at Ellos Postimyynti Oy has been rationalisation of the translation process. Previously, Finnish translations were made with a little help of manually collected batch replacement lexicon that took a lot of translators time in both updating and using the lexicon as a help in the pre-edit phase of manual translation. The translation phase itself took also a lot of time as the translations were done manually in a desktop publishing program by professional translators. More time was spent as the product articles were translated several times all the way from beginning when articles were placed in different locations, e.g. printed catalogues and web pages. This way, the translation work and management was done multiple times with different results, and the original source text was split up, not only in different languages, but also in multiple versions in each language.

Webtran system has speeded up the translation process in the Finnish subsidiary of Ellos, as the original Swedish source text is translated only once and then distributed into various use locations without any overlapping processing. Source text is automatically checked with the help of checking rules in the pre-edit phase and this way the possible ambiguities in the source text have already been eliminated before automatic translation. Since Webtran has so far been used only in the translation from Swedish to Finnish, the pre-edit phase has been done by the Finnish translators in Finland. Implemented in the whole text production and translation process at Ellos, the language checking will be done already in the authoring phase of the original source texts in Swedish.

After source text checking, a whole sales catalogue is automatically translated into Finnish in a fraction of the time that human translators have spent in manual translation of individual product articles. In addition, compared to the manual translation process at Ellos, Webtran requires a short post-edit phase of the automatically translated texts, as the translation results have to be checked and possibly corrected before publishing. The whole process still takes considerably less time than a fully manual translation process of sales catalogues.

5. Integration in practice

Webtran system is embedded on a server with real-time connection to the publishing programs of Ellos Postimyynti Oy. Translators execute the automatic translation directly from their own desktop programs without having any visible connection to the Webtran interface. Only the responsible main user of the Webtran system has access to the Webtran interface and operations between the Webtran tools.

Text production process begins, as the Swedish source text catalogues are produced in Sweden with Quark XPress program for desktop publishing. Catalogue pictures and text elements are combined in same files. These files are then delivered to target countries for translation. At Ellos Postimyynti Oy, text elements are automatically separated from the catalogue files and sent to Webtran for automatic checking and translation. The original text

elements in source files are automatically replaced with the translated text and this way the translator deals with whole catalogue files instead of separate text elements. In the company desktop publishing program the catalogue files are then proof-read and possibly post-edited before they are sent to be printed or generated as an on-line sales catalogue.

Webtran implementation started with installation of Webtran tools and refinement of the language model, with its basic foundation specified by VTT to be locally completed for in-company use of Ellos Postimyynti Oy. As the language model existed in a preliminary form before the implementation, time was saved in the modelling work and most resources were assigned to training of the staff at the translation department.

One of the company translators at Ellos Postimyynti Oy was trained to be the responsible main user of Webtran system. After getting a general insight to the basic architecture of Webtran in a few days, the main user focused on adopting and developing the translation rules. This has been a continuing process as changes come along and the language model needs to be evolving towards extensive coverage between the language description and the translation rules.

Terminology definition had also been initiated by VTT before the actual implementation phase, though the refinement of the lexicon practically needed to be done term by term as the main user added the final domain knowledge to specify term interpretations. This corresponds to the basic assumption that terminology definition could be based on general bilingual lexicon which then is completed and refined with the specific domain knowledge and sublanguage features.

Since spring 2000, terminology refinement has taken approximately two or three weeks of main users work. After the Webtran system had been fully implemented and adopted to the company translation process, which was done by the end of the summer, the main user has continued adding new terms and refining the existing lexicon. Updating of the lexicon is done in a centralised way once a week with the main user working for an hour to add new terms and editing the existing terms and translations.

After the system is in everyday use, updating mostly concerns the terminology as new terms are added in the lexicon or the translations need to be changed. Translation rules might need some adjusting if there is a fundamental change in the company sublanguage, and then of course the original writing rules should also be altered, but this is not a task to be done weekly. Controlling of the source text authoring should restrict the need to make changes in the language model or especially, in the translation rules.

5.1. Use Experiences

Compared to the situation before implementing of Webtran system, automatic translation has speeded up the translation process considerably. First phase in the translation process is initiating the automatic language checking and translation on the translators desktop. The translated documents return to the desktop automatically and another working phase is then the post-editing of translated catalogue files, which does not usually take a long time compared to the time that would be spent in a manual translation process.

Time needed for the actual automatic text processing and translation depends on the size and quality of the processed material. During the different business seasons of the year, Ellos Postimyynti Oy publishes catalogues of different size and content, catalogues with emphasis on pictures and visual image or catalogues with emphasis on textual information and wide range of products. In the domain of Ellos Postimyynti Oy, Webtran has longest traditions in translating product articles of clothing products and textiles. These subdomains have gone through the longest period of refinement, both on terminology and translation rules, and they naturally give better and faster results than the other product categories that have been added to the language model later on, during the implementation of Webtran.

As an example, a smaller Christmas catalogue with 134 pages consists of more text and products than an average season catalogue and it can therefore be seen as 200 pages of normal Ellos catalogue text. The entire catalogue is translated as a whole and the translation process includes a translation phase and the file transfer between the server and the translators desktop, where time of transfer depends on the amount of text elements in the catalogue files. Approximately, the whole catalogue is processed in an hour after which it is ready for post-editing and proof-reading. Time and effort used in the post-editing depend on how well controlled the source text has been. The only parts that actually need post-editing are usually longer sentences with other information than restricted facts about the product contents. These "advertising" elements are always re-formulated manually to fit the local style in the product articles and automatic translation would not provide the best possible result for that kind of purpose.

An additional advantage in using Webtran is that it eliminates the need of controlling the lay-out of the text, especially in the post-editing phase. When editing is done manually in a desktop publishing program lots of time is spent in clicking and activating the text boxes and in constant checking of paragraph marks, spacing and capital letters and so on. Every time a translator edits the text manually, there is a risk that some characters or product codes are changed unintentionally. Texts translated in the Webtran system have retained the original lay-out and therefore the translator does not have to edit, for example, the product codes or colour and size lists. These parts do not need any post-editing as they are the simplest elements to be translated automatically, as well as all the other recurring elements in a controlled sublanguage.

The possibility to perform a language check in the authoring phase of text production process contributes to production of texts that are consistent and unambiguous, which already saves time when separate pre-editing in all of the target languages becomes unnecessary. The automatic translation process is more efficient and faster than manual translation, as time is not wasted in manual processing of recurring simple constructions.

6. Conclusions

Considering the whole translation process at Ellos Postimyynti Oy, Webtran system is a cost-effective investment. It is fully adaptable to the company needs already in the implementing phase. Centralised translation into all target languages becomes possible if language models and translation capabilities are created in all target languages. This way, only one pivot source text is needed as the translation can be generated automatically for each publishing case. Webtran provides the possibility to real-time translation of product articles, which radically

diminishes the time space between the initial authoring of a product article and its immediate publishing in the target language to a customer via on-line service on the company web site.

Ellos Postimyynti Oy has been using the Webtran tools in the translation process from Swedish to Finnish, and the results have been satisfying. Automatic translation has speeded up the whole translation process. The results have given evidence of usefulness of this methodology and thus the work continues. In the future, concentration will be in adding new languages and integrating Webtran deeper into the text authoring process.

7. Acknowledgements

We would like to thank Antti Mälkönen, René Holm, Irma Sinerkari and Kirsi Villo-Moen at Ellos Postimyynti Oy for support and co-operation. We also thank Eeva Palosuo, Maarit Virtanen and Veli Kahrala from TietoEnator and Matti Sihto from Tekes Technology Development Centre in Finland. We also want to acknowledge Professor Seppo Linnainmaa and our colleagues Kuldar Taveter and Juha Sorva at VTT Information Technology.

8. Bibliography

Arnold, D., L. Balkan, R. Lee Humphreys, S. Meijer and L. Sadler (1994): *Machine Translation. An Introductory Guide*. NCC Blackwell Ltd., Oxford.

van der Eijk, Pim, Michiel de Koning and Gert van der Steen (1996): Controlled language correction and translation. In: *Proceedings of the First International Workshop on Controlled Langiage Applications*, CLAW 1996.

Hutchins, W. John, Harold L. Somers (1992): *An Introduction to Machine Translation*. Academic Press, London.

Lehrberger, John (1982): Automatic Translation and the Concept of Sublanguage. In: Kittridge, Richard & John Lehrberger (ed.). *Sublanguage. Studies of Language in Restricted Domains*. de Gruyter, Berlin.

Lehtola, Aarno, J. Tenni, C. Bounsaythip, K. Jaaranen (1999): WEBTRAN: A Controlled Language Machine Translation System for Building Multilingual Services on Internet. In: Machine Translation Summit VII '99 (MTSUMMIT '99), September 13-17, 1999, Singapore.

http://www.vtt.fi/tte/projects/webtran/doc/index.html

Kristiina Jaaranen, Aarno Lehtola, Jarno Tenni and Catherine Bounsaythip
VTT Information Technology
P.O.Box 1201, FIN-02044 VTT, Finland
E-mail: kristiina.jaaranen@vtt.fi