Language Control and Machine Translation

Anna Sågvall Hein
Department of Linguistics
Uppsala University, Box 513
S 751 20 Uppsala, Sweden
anna.sagvall_hein@ling.uu.se

Abstract. This paper describes ongoing work on the integration of a machine-translation software, Multra, into the multilingual document processing environment of Scania CV AB. Multra is a prototype of a modular, transfer-based MT-system with Swedish as its source language. It handles translation into English and German, based on a common analysis structure. In order to guarantee consistency in the original as well as in the translated versions of the documents, a controlled language, ScaniaSwedish, is defined. Also, a language checker for this language is developed. The core of the checker is the analysis component of Multra. The checker will provide two kinds of results, i.e. a controlled version of the text and the text as a sequence of grammatical structures. These structures can then be forwarded to transfer and further generation. In other words, checking the text means taking the first step in the translation process. The checker is developed in parallel with the definition of ScaniaSwedish.

1 Introduction

Work on the Multra system (Multilingual Support for Translation and Writing) was initiated as a pure research enterprise. It was seen as a natural and interesting follow-up of a Swedish parsing project resulting in a parser for Swedish, Sve.Ucp [Sågvall Hein 1983, Dahllöf 1989]. Thus Swedish was to be the source language of Multra. Another characteristic feature of Multra was the orientation towards a multilingual system. In order to give the project a clear focus, we looked around for an organisation where large amounts of Swedish text were systematically translated into several languages. The Swedish truck and bus manufacturer Scania CV AB in Södertälje is such a company. An informal co-operation was established, and Scania provided us with multilingual text (work descriptions in original and translated versions) on which the behaviour of Multra could be modelled. In 1993 a demonstrator was available that could handle translation from Swedish to English and German [Sågvall Hein 1996]. A major step to be taken then was to extend the coverage of the dictionaries and the grammars of Multra to a realistic and useful size. Preliminary studies of a multilingual corpus of maintenance manuals showed that, in spite of the fact that Scania maintains high demands on language quality, quite a few inconsistencies in the source as well as in the target versions could be found. This lead to the conclusion, that the source language had to be controlled, and work on the definition of such a language, ScaniaSwedish, was initiated. Efforts devoted to optimising the analyser, the heaviest part of the translation process, should pay off in a multilingual setting. Controlling the source language is, certainly, a fundamental part in such an optimisation process.

2 Multilingual Document Production at Scania

The documentation of truck maintenance at Scania Sv AB is extensive. In 1996 the production of text went up to 6,000 pages. To this should be added the already existing documentation, which

consists of 7,000 pages. The documentation is written by technical writers at Scania in Swedish, and is translated in its full versions into seven languages at the moment: English, German, Dutch, French, Italian, Spanish, and Finnish. Parts are also translated into Norwegian, Danish, and Portuguese. Quality in the maintenance documentation is an important competitive factor on the market. It should be consistent, correct, and easy to understand. Scania has therefore decided to use Swedish, the mother tongue of the technical writers, as the source language in the translation process. In doing so, Scania strongly believes that the quality of the translation is firmly grounded.

3 Basic Approach

According to our approach, multilingual translation should be based on a controlled source language, maintained by means of a language checker[1]. The grammar checker should fully cover the controlled language and guarantee a text in conformity with the specification of the controlled language. It should base its work on full parsing, generating grammatical structures that can be forwarded to the transfer and generation components. With this approach, the first, and heaviest, step of the translation process will be taken by the language checker, and there will be a firm ground for translation. Defining transfer (and generation) rules for the target languages implies a standardisation of them too. We base the implementation of this approach on the Multra system.

4 Multra

Multra is a transfer-based machine-translation system, with three main components, an analyser, a transfer component, and a generation component. In addition, there is a separate component ordering the analysis alternatives by preference before passing them on to the transfer component. Preference is expressed by means of linguistic rules defined over feature structures.

Transfer is implemented as unification of feature structures. Generation, in addition, involves concatenation. Also in the analysis, unification plays an important role. Thus we may say that Multra is a unification-based machine-translation system. Transfer rules are expressed in a PATR-like formalism, and there is no formal difference between lexical and structural transfer rules [Beskow 1993]. Also for the formulation of syntactic generation rules a PATR-like formalism has been defined. Morphological generation rules are formulated in a PROLOG like style.

Alternative transfer rules are applied according to specificity; a specific rule takes precedence over a general one. The specificity principle also governs the application of alternative generation rules. The linguistic preference rules along with the specificity principle of the transfer and generation processes constitute the Multra preference machinery. The MT system as a whole, as well as its components, can be tuned to present the best alternative only, or the complete set of alternatives in the preferred order. For the design and testing of translation rules, a special environment, Multra Developer's Tool, MDT [Beskow 1992], has been built. In this environment each component can be tested independently. In specific, MDT provides rich tracing facilities.

[1] In implementing this approach, we find much inspiration in the achievements made in the Kant system [Mitamura & Nyberg 1995] where such a strategy has been applied successfully.

An example of translation in Multra

Input: Sätt upp växellådan i universalstativ: [Put the gearbox on universal stand.]

Result of the parsing process:

```
[PHR.CAT : CL
TYPE : IMP
REG : [V1.LEM : SATTA.VB
PRED.LEM : SATTA.VB+UPP.PL]
PRED : [VERB : [LEX : SATTA.VB+UPP.PL.1
VSURF : +
INFF : IMP]
        OBJ.DIR : [PHR.CAT : NP
                   NUMB : SING
                   GENDER : UTR
                   CASE : BASIC
                   DEF : DEF
                   DF : [HEAD : [LEX : VAXELLADA.NN.1
                                 WORD.CAT : NOUN]]]
        OBJ.LOC : [PREP : [WORD.CAT : PREP
                   LEX : I1.PP.1]
                   PHR.CAT : PP
                   POBJ : [PHR.CAT : NP
                   NUMB : SING
                   GENDER : NEUTR
                   CASE : BASIC
                   DEF : INDEF
                   DF : [HEAD : [LEX : UNIVERSALSTATIV.NN.X
                                 WORD.CAT : NOUN]]]]
SUBJ : 2ND]
SEP : [WORD.CAT : SEP
       LEX : STOP.SR.0]]
```

Result of the transfer process (first alternative of two):

```
[SEP : [WORD.CAT : SEP
        LEX : STOP.SR.O]
 PRED : [VERB : [LEX : SET_UP.VB.0
         VSURF : +
         INFF : IMP]
         SUBJ : 2ND
         OBJ.LOC : [PHR.CAT : PP
                    PREP : [LEX : 0N.PP.0]
                    POBJ : [DF : [HEAD : [LEX : UNIVERSAL_STAND.NN.0
                                  WORD.CAT : NOUN]]
         DEF : INDEF
         NUMB : SING
         PHR.CAT : NP]]
         OBJ.DIR : [DF : [HEAD : [LEX : GEARBOX.NN.0
                          WORD.CAT : NOUN]]
 DEF : DEF
 NUMB : SING
 PHR.CAT : NP]]
 TYPE : IMP
 PHR.CAT : CL]
```

Trace of the transfer process (first trace level):

```
Applying Rule SEP
Applying Rule PRED
Applying Rule VERB
        Applying Rule SATTA.UPP
    Applying Rule SUBJ
        Applying Rule IMP.SUBJ
    Applying Rule OBJ.LOC
        Applying Rule I_UNIVERSALSTATIV
            Applying Rule DF
                Applying Rule HEAD
            Applying Rule UNIVERSALSTATIV
    Applying Rule CASE
    Applying Rule GENDER
    Applying Rule DEF
    Applying Rule NUMB
    Applying Rule PHR.CAT
    Applying Rule OBJ.DIR
    Applying Rule DF
            Applying Rule HEAD
        Applying Rule VAXELLADA
Applying Rule CASE
Applying Rule GENDER
Applying Rule DEF
Applying Rule NUMB
Applying Rule PHR.CAT
Applying Rule TYPE
Applying Rule PHR.CAT
Applying Rule REG
Success!
```

Result of the generation (first alternative):

```
Set up gearbox on universal stand.
```

Trace of the generation process (first trace level):

```
Applying Rule CL.IMP.OBJ.DIR.OBJ.LOC
    Applying Rule HEAD
    Applying Rule PP
        Applying Rule NP1a
Success!
```

4.1 The Multra Analyser

The analyser of Multra is Sve.Ucp, a chart parser generating grammatical descriptions in terms of attribute value structures. Sve.Ucp uses a procedural formalism, and rule invocation is triggered from the grammar and the dictionaries. The same formalism is used both in the dictionary and in the grammar. This allows for the implementation of a flexible rule invocation strategy mixing top-down and bottom-up rule invocation.

Dictionary-search, morphological analysis, and syntactic analysis are handled in a common chart framework, and processing proceeds task by task. A unique start rule in the grammar specifies (for each application) what rule(s) should be applied to get the process going. The start-rule used in Multra specifies two kinds of processing, i.e. dictionary search and the application of sentence rules. The dictionary search rule will lead to the recognition of words and phrases. For instance, at the recognition of a nominal stem, a noun rule is triggered, which in its turn invokes an np-rule, if the morphological analysis of the noun succeeds.

Basically, phrase constituents are invoked bottom-up and sentence rules are invoked top-down.

5 Defining ScaniaSwedish

ScaniaSwedish will be defined with regard to vocabulary, phraseology, grammar, punctuation, and general writing conventions. It will be based on an examination of the unrestricted Swedish to be found in a corpus of up-to-date maintenance text (15,000 pages). On this language, systematic restrictions will be imposed, with the aim of eliminating unnecessary linguistic variation while keeping the required expressive power.

For this purpose, a multilingual corpus of 80 documents in seven languages has been established, the Scania corpus (Table 1). This corpus comprises the full documentation of the new truck model 4, as of January 1996. It is representative of the linguistic style of unrestricted Swedish that was used at the company prior to the definition of ScaniaSwedish.

Language	files	words	bytes
Swedish	80	172259	7792597
Dutch	80	216424	8072128
English	80	220827	7886082
Finnish	80	148348	7833990
French	80	244239	8156457
German	80	186293	8004331
Italian	80	228631	8127121
Spanish	80	250730	8090916
total	640	1667751	63963622

Table 1. The Aligned Scania Corpus at June 10, 1996 [Tjong Kim Sang 1996]

5.1 Vocabulary

In order to determine the vocabulary of ScaniaSwedish, we made an investigation of the words in the Swedish Scania corpus (Table 2).

In total, the corpus comprises 206,900 tokens, 22,646 types and 9,184 lemmas that were approved in the first run. To get at the lemmas we analysed the word types by means of the Multra analyser (morphological processing only), and made a rough lemmatising based on these findings. Table 2 also shows the distribution of the vocabulary over the sub-dictionaries of the analyser, i.e. a core dictionary, a general dictionary, Scania terms, and other Scania

Dictionary	Tokens	%	Types	%	Lemmas
General	109,207	53	6,567	29	3,183
Scania	65,948	32	11,486	51	6,001
Numerical	28,542	14	3,284	15	-
Minus	2,137	1	940	4	135
Zeros	1,071	0.5	369	2	-
Total	206,900	99.5	22,646	101	9,319

Table 2. Swedish tokens, types, and lemmas in the Scania corpus [Starbäck 1996]

words. In addition, we have collected a couple of hundreds of so-called minus-words (types and lemmas), i.e. non-approved words [Almqvist & Sågvall Hein 1996], and entered them into the minus-dictionary of the Scania Checker. By Zeros in Table 2 we refer to various kinds of misspellings. As can be seen, the approved vocabulary of ScaniaSwedish will not exceed 9,184 lemmas.

5.2 Phraseology and Grammar

Phraseology is an important aspect of a controlled language. In specific, the expressions of noun valency relations are found to cause troubles and lead to inconsistencies in the formulation of the source language. Currently, these issues are systematically investigated at Scania.

As a preparation for a systematic study of the grammatical structure in the corpus, the text was split into sentence like segments that are to function as translation segments in the translation process. The most typical translation segment is the sentence. However, also headers (major and minor), list elements, list element labels, and table cells have a fairly independent status in the text and should be treated as translation segments in their own right. In order to recognise them, we have to use typographical information in the documents. Consequently, a software has been developed that converts the Framemaker version of the documents into TEI Lite SGML [Tjong Kim Sang 1996]. The SGML version of the documents provides a basis the segmentation into sentences and sentence like segments. Based on this segmentation, statistics about sentence lengths in the corpus has been calculated. Further, the corpus has been accordingly aligned, and as a result of the alignment process we get a very useful material for finding lexical translation equivalents. Work in this direction is also on its way. Meanwhile, we observe that even though this seems to be a very repetitive lund of text, not more than 35% of the sentence pair types (source and target sentences) appear more than once in the Swedish - German aligned sub-corpus.

6 A Language Checker for ScaniaSwedish

The language checker should cover all the aspects characterising ScaniaSwedish. The core of the checker will be the Multra analyser. The Multra analyser has no means for handling incorrect input. Everything that deviates from the specifications of the grammar and the dictionaries will cause the parsing process to stop, be it on the lexical, the morphological, or the syntactic level. The Scania checker, on the other hand, must be capable of handling deviations from ScaniaSwedish at all these levels.

So far, we have a specification of ScaniaSwedish only as regards its vocabulary, and this specification has been built into the checker. Thus we have a dictionary of approved words, plus words, and a dictionary of non-approved words, minus words. The dictionary of minus words comprises anticipated violations of the norm that has been defined, i.e. *AChäfte* [AC-booklet] instead of *AC-häfte.* When a minus word appears the checker presents the approved alternative. It also considers spelling errors (performance errors) such as *ansluning* [connection] instead of *anslutning* etc. A spelling error is a string that is not found in any of the dictionaries. When such an unrecognised string appears, the checker presents it as a spelling error candidate for the user to react to and goes on to find the next word. Dictionary search is based on morphological analysis and the morphological grammar accounts for the recognition of non-approved inflectional forms such as *medbringarn* [the driver] instead of *medbringaren.*

On the syntactic level, the checker recognises phrase constituents including some types of grammatical errors (e.g. violations of gender, number, and species agreement in NPs), and also some types of deviations from ScaniaSwedish as regards valency (so far, basically, post attributes). Thus, in its current version, the checker bases its operation on partial parsing, recording the errors that are found in the chart. The chart is then checked for errors, and an error report is generated. When the checker has been fully implemented, grammar rules at sentence level will be applied top-down as a complement to the bottom-up rule invocation. Sve.Ucp with its procedural formalism has been found well suited for this kind of relaxed processing.

A first version of the checker will be installed at Scania in June 1997 and evaluated by the technical writers during a period of three months.

7 Conclusions

A full implementation of our approach will have implications on the organisation of the multilingual document production at the user site. The main task of the technical writer today is to provide a source document to be printed and translated. According to our approach, the technical writer, supported by a language checker, will be in charge, not only of the production of the source document, but also of the first step in the translation process, the analysis of the text. The technical writer has to respond to the messages produced by the checker and react accordingly, hereby helping the analyser along at the same time as he will get the required guidance in his own work. The translator's work starts with the application of the transfer component of the translation system. In specific, the translator supervises the operation of the transfer, and subsequently, generation component. He may also be in charge of updating the dictionaries and the grammars of the translation system.

Once the grammars and the dictionaries of the analyser/checker and the translation modules have been developed and well tested, we anticipate that the technique with a translation memory will be a useful complement.

References

[Almqvist & Sågvall Hein 1996] Almqvist, Ingrid & and Anna Sågvall Hein. 1996. Defining Scania Swedish - A controlled language for Truck Maintenance. In: *Proceedings from the First International Workshop on Controlled Language Applications,* Leuven, Belgium, pp. 159 - 165.

[Beskow 1992] Beskow, Björn, 1992. *Transfer Tool on UNIX: An Introduction.* Dept. of Linguistics. Uppsala University.

[Beskow 1993] Beskow, Björn. 1993. *Unification Based Transfer.* RUUL 24. Dept. of Linguistics. Uppsala University.

[Dahllöf 1989] Dahllöf, Mats. 1989. *Satslösning i en lexikonorienterad parser for svenska.* [Parsing with a Lexicon-Oriented Parser for Swedish]. Master's thesis. Gothenburg University.

[Dahllöf & Sågvall Hein 1989] Dahllöf, Mats & Anna Sågvall Hein. 1989. Procedural Frameworks. In Sågvall Hein et al. (eds.) *Studies of Grammars, Formalisms, and Parsers. A Report from the Project Grammar Models for Natural Language Processing.* Dept. of Computational Linguistics, University of Gothenburg, pp. 3-7.

[Mitamura & Nyberg 1995] Mitamura, Teruko & Eric H. Nyberg 1995. Controlled English for Knowledge-Based MT: Experience with the KANT System In: Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation, Leuven, Belgium.

[Sågvall Hein 1983] Sågvall Hein, Anna. 1983. *A Parser for Swedish. Status Report for Sve.Ucp.* Report No. UCDL-R-83-1. Center for Computational Linguistics. Uppsala University.

[Sågvall Hein 1996] Sågvall Hein, Anna. 1995. Preference Mechanisms of the Multra Machine Translation System. In: Hall Partee, Barbara & Peter Sgall (eds.) *Discourse and Meaning. Papers in Honor of Eva Hajičová.* John Benjamin's Publishing Company, Amsterdam/Philadelphia, pp. 321 - 333.

[Starbäck 1996] Starbäck, Per. 1996. *Definition av Scaniasvenska. Arbetspaket 7.* [Definition of ScaniaSwedish. Workpackage 7]. Dept. of Linguistics. Uppsala University.

[Tjong Kim Sang 1996] Tjong Kim Sang, Erik. 1996. *Converting the Scania Framemaker Documents to TEI SGML.* Dept. of Linguistics. Uppsala University.