[From: *CLAW '96*, Universiteit Leuven]

Defining ScaniaSwedish - a Controlled Language for Truck Maintenance

Ingrid Almqvist
Scania CV AB
Sodertalje
Sweden

Anna Sågvall Hein
Department of Linguistics
Uppsala University
Sweden

Abstract

An approach to integrated multilingual document production is proposed. The basic idea of this approach is to use the analyzer of a modular, transfer-based machine translation system as the core of a language checker. The checker generates grammatical structures to be forwarded to the transfer and generation components for the various target languages. A precondition for such an approach is a controlled source language. The source language in focus of this presentation, is ScaniaSwedish, to be defined via a standardization of the language presently used by Scania in their truck maintenance documents. Here we concentrate on the identification of the vocabulary of current ScaniaSwedish and present the results that we achieved so far. In parallel with the inventory of the vocabulary, the competence of the language checker is developed.

Background

The documentation of truck maintenance at Scania CV AB is extensive. In 1996 the production of text will amount to 6,000 pages. To this should be added the already existing documentation, which consists of approx. 7,000 pages. The documentation is written by technical writers at Scania in Swedish, and it 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. Needless to say truck maintenance documentation is a vital part in the *After Sales Service* and its quality is an important competitive factor on the market.

Truck maintenance documentation at Scania is at the moment going through a rapid process of change, which aims at the storage of information in a central data base, from where the desired information at the moment (e.g. the retrieval of a spare part; the analysis of a fault code) is fetched. This also effects the work routines of the technical writers, and will to some extent fragmentize their work. This is one of many reasons why it is essential to Scania that the language of the documentation is 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.

In view of the fact that the documentation as well as the number of translation languages will increase in the period to come, Scania has reinforced its co-operation with the Department of Linguistics at Uppsala University, UU, aiming at the development of a computerized translation support (Almqvist & Sågvall Hein 1995).

An Approach to Integrated Multilingual Document Production

A fundamental step in the co-operation project between Scania and UU is a standardization of the source language into what will be called *ScaniaSwedish* and the development of a language checker modeled for that language. ScaniaSwedish will be defined with regard to vocabulary, grammar, punctuation, and general writing conventions, and the Scania language checker will cover all these aspects.

The Scania language checker will make use of a positive and a negative grammar, a positive and a negative dictionary. The negative parts of the language description will include foreseen mistakes and give recommendations for corrections. The checker will be based on the analyzer of the Multra Machine Translation System (Sågvall Hein 1995). It is a prototype of a translation engine with demo-versions for the translation of car maintenance manuals from Swedish to English and German. Multra has a strictly modular architecture, and the analyzer generates structural descriptions which can, immediately or at a later state, be forwarded to the transfer and generation components. For each target language there are separate transfer and generation components, whereas the analysis is one and the same regardless of target language. Consequently, efforts devoted to optimizing the analyzer, being the heaviest part of the translation process, will pay off in a multilingual setting. Controlling the source language is, certainly, a fundamental part in such an optimization process.

In our approach, the language checking part will be identical to the analysis part of the translation process. In other words, when the source document has been checked, the first step in the translation process has been taken and the resulting structures can be forwarded for transfer to the various target languages. Defining transfer (and generation) rules for the target languages implies a standardization of them too.

Identifying the Vocabulary of ScaniaSwedish.0

The unrestricted Swedish currently used in the service documents, *ScaniaSwedish.0,* will be identified via an examination of the language in an extensive corpus of up-to-date maintenance text (15,000 pages). On this language, systematic restrictions will then be imposed, which aim at eliminating unnecessary linguistic variation while keeping the expressive power that is required. The resulting controlled language will be referred to as *ScaniaSwedish.*

A methodology for the identification of the vocabulary of ScaniaSwedish is worked out in an on-going pilot study. The study comprises 40 documents (984 pages; 89,112 current words), which constitute half of the Scania documentation on the new truck (104) introduced in October 1995. Consequently, they are representative of the current linguistic style of ScaniaSwedish.0.

After due conversion from FrameMaker to ASCII, the text corpus was processed by a text handling system developed at UU (Dahlqvist 1994), and, in addition to various kinds of quantitative data, an index of simple and phrasal words, basically, particle verbs, was created. According to this index, the text consists of 86,517 tokens and 10,491 types, see Table 1 (below). The numerical tokens (including hybrids of digits, letters, and special characters) amount to 14,348 and the numerical types to 1,335.

As regards the types (excluding the numerical expressions), we want to find the answers to the following questions:

How many lemmas do they represent?
What kinds of lemmas do they represent?
What lemmas are covered by the Multra dictionaries?
What lemmas are outside the scope of the Multra dictionaries?

Table 1: Tokens and types in the pilot corpus (most phrasal verbs regarded)

	Tokens	%	Types	%
simple	69,666	80	8,819	84
phrasal	2,5.03	3	337	3
numerical	14,348	17	1,335	13
Total	86,517	100	10,491	10

To answer the questions formulated above, we set the Multra analyzer at work. For each type whose stem is present in its dictionaries, it generates a word description including information about lemma, form, inflectional type, dictionary stem, and origin. Origin refers to the dictionary in which the word was found. When the stem of a word is missing in the dictionaries, the analysis fails and no description is provided.

Examples of word descriptions generated by Multra (with frequency in parentheses):

"arbetstakten" (1)
Origin GD: arbetstakt.nn
(* = (LEM = ARBETSTAKT.NN
INFL = PATTERN.EXIL
DIC.STEM = ARBETSTAKT
NUMB = SING
WORD.CAT=NOUN
GENDER = UTR
FORM = DEF
CASE = BASIC))

"arbetstemperatur" (3)

Origin Multerm: arbetstemperatur.nn
(* = (LEM = ARBETSTEMPERATUR.NN
INFL = PATTERN.FILM
DIC.STEM = ARBETSTEMPERATUR
NUMB = SING
WORD.CAT=NOUN
GENDER = UTR
FORM = INDEF
CASE = BASIC))

"arbetstemperaturen" (1)

Origin Multerm: arbetstemperatur.nn
(* = (LEM = ARBETSTEMPERATUR.NN
INFL = PATTERN.FILM
DIC.STEM = ARBETSTEMPERATUR
NUMB = SING
WORD.CAT = NOUN
GENDER = UTR
FORM = DEF
CASE = BASIC))

Multra operates with four different dictionaries:

Core Vocabulary Dictionary, *CV*
General Dictionary (excl. CV), *GD*
Core Scania Terminology, *Termlex*
General Scania Terminology, *Multerm* (excl. Termlex)

CV comprises roughly 2,000 simple and 4,000 phrasal units, and GD approx. 60,000 simple units. *Termlex* is Scania's terminological database consisting of 2,000 terms and denominations. All new components and articles, which are parts of products, are given a name, which is listed in the data base together with equivalents in six languages. This vocabulary constitutes the technical core terminology of Scania Swedish. Termlex was included in the enlarged terminology of *Multerm*, done by UU and Scania in 1994 (Brännström & Dahlqvist 1994). Multerm comprises approx. 3.500 entries and the equivalents in six languages.

With the analyses produced by Multra as a basis, we may estimate the number of lemmas in the pilot corpus and their lexical origin, see Table 2.

Table 2: Lemmas in the pilot corpus and their lexical origin
Numerical expressions are excluded.

Origin	Lemmas	%
Termlex	345	6
Multerm excl. Termlex	1,369	17
CV	758	12
GD	1,414	23
Particle verbs	248	4
Unanalyzed	2,306	38
Total	6,095	100

As can be seen from the table, 35% of the lemmas belong to the general vocabulary (CV + GD) and 23% to the terminology (Termlex + Multerm). As regards the particle verbs, the majority of them, eventhough not all, belong to the general vocabulary. 2,713 types, corresponding to 2,306 lemmas, remain unanalyzed. The proportion of the general vocabulary is still larger if we look at the textual frequency, the tokens, see Table 3.

Table 3: Tokens in the pilot corpus and their lexical origin. Numerical expressions are excluded.

Origin	Tokens	%
Termlex	6,969	10
Multerm excl. Termlex	11,652	16
CV	18,621	49
GD	7,138	10
Particle verbs	248	4
Unanalyzed	8,804	12
Total	72,169	100

Unanalyzed words

The word types that remain unanalyzed by Multra are outside the scope of its dictionaries. Among them we find misspellings, e.g. *ansluning (anslutning) 'connector',* acronyms, e.g. *ACL (Automatic Chassis Lubrication),* type words e.g. *GR (gearbox),* English words e.g. *brake,* phrasal words, e.g. *API GL-5 (lubrication test),* and simple, termlike words e.g. *accelerationskurvan 'acceleration curve'.*

The unanalyzed words have been examined manually in their contexts (concordances), and in Table 4 we present some quantitative facts about them.

Table 4: Kinds of unanalyzed word types:

Misspellings	93	4.03
Acronyms/type words	33	1.43
English words	43	1.86
Phrasal words (technical)	134	5.8
Simple, term-like words	2,003	86.9
Total	2,306	100

Actions to be taken with the unanalyzed words

The misspellings are saved for inclusion in the "negative" dictionary of the language checker. These misspellings entries will comprise recommendations for corrections.

The acronyms/type words, and also the technical phrasal words, are more or less Scania specific. They will be referred to the ScaniaSwedish vocabulary and entered into Multra's term dictionaries.

The English words are found to be part of English phrasal expressions, such as *After Sales Market* and as such they will be represented in ScaniaSwedish and Multra.

The majority of the unanalyzed words (86,9%) are simple, termlike words that are candidates for ScaniaSwedish. However, among them we also find unnecessary spelling and inflection variation, such as *AChäfte* and *AC-häfte* 'ac-booklet', *skiljda* and *skilda* 'separated'. Such phenomena are typical cases for a standardization effort. As regards spelling, one variant is defined as standard (in this example *AC-häfte)* and included in the positive dictionaries of Multra, and the non-chosen variant *(AChäfte)* is entered into its negative dictionaries with information about the preferred spelling.

Concerning inflection, the situation is a little different. Here, three actions may have to be taken. First a choice has to be made as to what form is to be preferred. Then the non-accepted form has to be listed in the negative dictionary of Multra. Finally, the inflection grammar has to be restricted, accordingly. The inflection grammar of Multra (Sågvall Hein 1992) is designed for general Swedish, and it accepts variation to the same extent as Svensk Ordbok (1986) after which it has been modeled.

When decisions about spelling and inflection variation have been taken, the termlike words will be entered into Multra's term dictionaries. (Questions about synonymy will be postponed to a later state of the standardization process.) After that, the previously unanalyzed words will be processed by Multra making use of its negative as well as positive dictionaries. The word descriptions resulting from this processing will be merged with the previously analyzed words, and we can present a complete picture of the lemmas in the pilot corpus. As a side-effect, we also have a raw version of the Scania language checker, tuned for the vocabulary of the pilot corpus.

Conclusions and future work

Next we will apply Multra to an enlarged corpus, consisting of the remainder of the documentation of the new truck, i.e. adding one and a half times the size of the pilot corpus. In a final lemma identification step, the 4,000 pages of documentation of the previous truck model (3) will be processed. According to our estimations so far, the total number of lemmas of ScaniaSwedish will not exceed 10,000. It is still an open question how many synonyms there are in ScaniaSwedish.0, synonyms that should be eliminated in the standardized version of the documentation language. So far, we see no other way to answer questions about unmotivated synonymy than by manual inspection of a well-organized lemma material with contexts.

The standardization process outlined so far focuses on determining spelling, inflection and a set of lemmas. Another important aspect of the vocabulary concerns phraseology, in specific nominal collocations and nouns with prepositional post-attributes. In approaching these questions, we also take a step towards the standardization of the grammar of ScaniaSwedish.

A first version of ScaniaSwedish, defined with regard to vocabulary (set of lemmas, inflection, spelling and certain types of phraseology, see above), along with a tailor-made vocabulary checker is planned for the end of 1996. In 1997 work on the grammatical aspects of ScaniaSwedish will be initiated.

Literature

Almqvist, Ingrid & Anna Såagvall Hein, 1995. Scaniasvenska - ett kontrollerat språk för lastbilsunderhåll. 'ScaniaSwedish - a controlled language for truck maintenance'. Utkast till projektplan 1995-11-10. 'An outline of a project plan 1995-11-10'. Scania CV AB. Södertälje. [In Swedish]

Brännström, Berith & Bengt, Dahlqvist, 1994. Multerm - Projektrapport. Uppsala universitet. Institutionen för lingvistik. [In Swedish]

Dahlqvist, B. 1994. TSSA 2.0. A PC Program for Text segmentation and Sorting. Uppsala University. Department of Linguistics.

Sågvall Hein, Anna, 1992. From Natural to Formal Dictionaries. In: Proceedings of the 5th EURALEX International Conference on Lexicography. Tampere 1992. pp. 301 - 310.

Sågvall Hein, Anna, 1995. Preference Mechanisms of the Multra Machine Translation

System. In: Hall Partee, B. & P. Sgall (eds.) Meaning and Discourse. Festschrift fuer Eva Hajicova. J. Benjamin's Publishing Co. 12 p.
Svensk Ordbok. 'A Dictionary of Swedish'. 1986.