[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

A Language Checker of Controlled Language and its Integration in a Documentation and Translation Workflow

Ingrid Almqvist
Technical Information
Scania CV AB
Södertälje
Sweden
ingrid.almqvist@scania.com

Anna Sågvall Hein
Department of Linguistics
Uppsala University
Uppsala
Sweden
anna@ling.uu.se

Introduction

In this paper we present Scania Checker, a web-based language checker for Swedish automotive service literature. It has been developed by the Department of Linguistics at Uppsala University in co-operation with the After Sales department Technical Information at Scania, Sodertalje, Sweden. In order to ensure translation consistency and quality, technical writers at Technical Information use it to check grammar and vocabulary in the source document before it is being translated.

We will show what kind of responses Scania Checker offers the writer, and demonstrate some examples of the rules behind the responses. We will also show, what the writer is expected to do in order to integrate the use of Scania Checker into a documentation and translation work-flow.

The Scania Checker consists of a word checker and a grammar checker. A fundamental part of the checker is the Scania Checker Lexical Database. The contents and characteristics of the database will be presented, and how it is used by the author, as well as by the database administrator. The error coverage of the grammar checker will be out-lined and error type frequencies resulting from the training of the checker will be given. Some technical aspects of the checker will also be presented including the basic operation of the grammar checker.

Finally, some possibilities of future developments will be sketched.

Background

Most information products concerning repair and maintenance of Scania trucks and buses are produced at the After Sales department Technical Information in Sodertalje Sweden. Today the media ranges from fiche, printed matters, software to Internet publication. All information is produced in Swedish by approx. 20 in-house technical writers and from time to time by consultants. Information products are then translated into English and from English into the remaining eight target languages.

In mid 1990, it became obvious, that it was necessary to provide technical writers with source language support in order to ensure translation consistency and quality. As a result, a prototype of a Swedish language checker was developed and installed at Scania for evaluation in 1997 (Sågvall Hein et al. 1997). The basis for the controlled vocabulary of the checker was a corpus of some 200,000 words. It was found, however, that the vocabulary was too small, and work on a major extension of it was initiated. The basis for the extension was a corpus of some 1,8 million words (Tiedemann 1998.). As a result, a new version of the checker, Scania Checker, was developed. It was introduced at Scania in August 2000 as a tool for Swedish grammar and vocabulary checking. In connection with the extension of the vocabulary, a major development of the language checking technology was carried out, and the checker was made available on the web. Experiences and achievements made in the SCARRIE[1] project played an important role in the development of Scania Checker.

Documentation and Translation at Scania

The documentation process at After Sales Technical Information is through the Product Development process closely interlinked with the activities of Design Departments and Repair Method Department. The PD process starts by stating a market demand of a product or vehicle function. Once pre-studies are finalised, a development phase starts, which, via several decision points, where the future of the project continuously is evaluated, leads to an implementation stage. At the implementation stage, all information necessary to produce After Sales documentation should be available. After Sales departments are, however, involved already early in the development phase to provide designers with the maintainability perspective of a new component.

Technical documentation at Scania at large is written in Swedish as well as English. Internally, Swedish is often used, but English is used to communicate within Scania worldwide. Furthermore, suppliers' documentation is often in English, sometimes even translated from German to English. However, technical writers at After Sales Technical Information write the repair and maintenance documentation in Swedish, which has been considered the most efficient way in the long run, since Swedish is their mother tongue.

As shown in Figure 1, the workflow for the writing and translation process at After Sales department Technical Information is the following. After finalising a text, the writer runs it through Scania Checker, which marks grammatical and vocabulary mistakes, and words not found in the lexical database, i.e. candidates for new words. The writer then corrects the source document according to the suggestions made by the checker. New words are sent to a log file to be attended to by the database administrator.

[1] See http://www.scarrie.com/ for general information about the project and http://stp.ling.uu.se/~ljo/scarrie-pub/ for a test version of the resulting Swedish SCARRIE pilot.

Now, the document is sent to a selection of departments for technical approval. This is a process, which takes 1-2 weeks depending on the information product. People reading the document may insert comments on a specific word in the database. Such comments are accessible to the database administrator, who also updates the database in the mean time with logged words, that have been found correct, and with rejected words providing them with replacements.

The technical writer adjusts the source document according to comments received, and then runs it through Scania Checker again. After any remaining grammatical or vocabulary inconsistencies have been corrected, the source document is now ready for translation.

Translation has been out-sourced at Scania since approx. 1994. The present supplier, an UK translation company, translates the source document etc. from Swedish into English, and from English into the remaining target languages. In translating a translation memory as well as multilingual terminology resources are used interactively.

After formats and language quality have been checked, translated target language documents are delivered to Scania, where a product administrator, after having checked files for possible mistakes, handles the media production and distribution.

In the documentation process as well as the translation process, much work has focused on the importance of writers and translators having terminological resources to their help. At the moment, there are two resources, which presently are separate databases, but which we hope to combine in the future, Termlex and Scania Checker Database.

Termlex is a Scania multilingual database, where denominations and definitions of new articles are stored. It contains 10 languages, including, of course, Swedish and English. A denomination committee, which consists of people from different areas in Scania, makes all decisions about new Swedish words and their English equivalents. After an English translator has validated the English term, the 8 target language equivalents are created on the basis of the definition and the English term.

The Scania Checker Database contains all Swedish words and segments that were found in the repair and maintenance literature during the period 1995-2000. Segments are categorised morpho-syntactically and classified as accepted or rejected. If rejected, a replacement is offered. The database provides excellent opportunities to search for different kinds of data.

Translators have access to Termlex, as well as a reference terminological material provided by Scania, and translators' own entries. Furthermore, glossaries with frequently occurring general words, phrases and titles have been fixed.

In conclusion, much work has been spent on establishing terminological resources. Writers as well as translators are obliged to use these. In a sense, this may be felt as restricting their freedom to choose the words they find most appropriate in a certain context. By providing writers with the possibility to insert comments on words in Scania Checker Database, and translators with a routine how to implement a change

of a target language term, we hope to have established a way to create and maintain a high-quality multilingual terminology in a controlled manner.

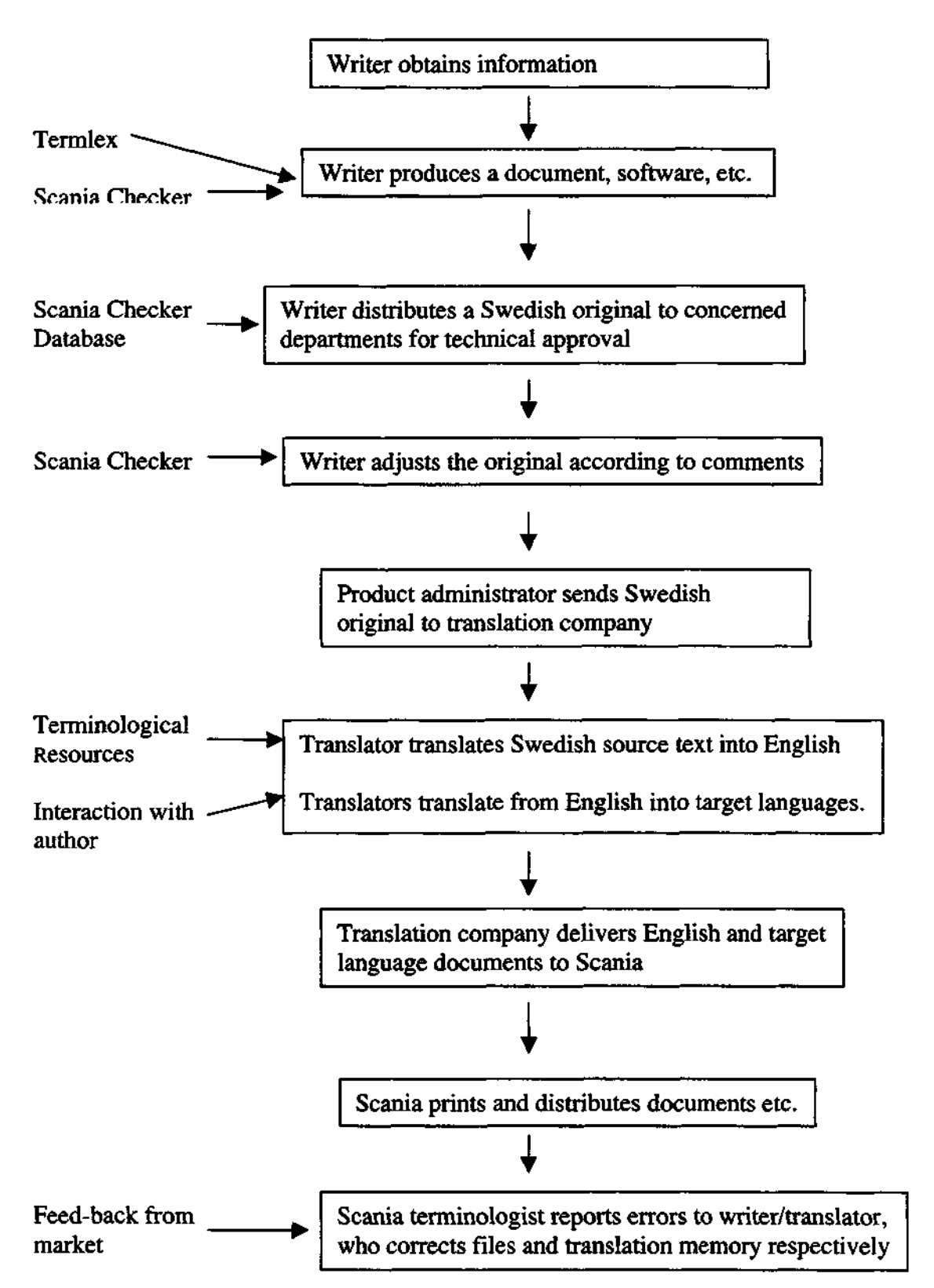

Figure 1. Documentation and translation process at After Sales Technical Information at Scania

Scania Checker

Scania Checker is a language checker for automotive service literature in Swedish. It comprises a word checker and a grammar checker. A fundamental part of the word checker is a lexical database. Scania Checker is a web-based application, which makes it platform independent, and this has proved to be an advantage in a department using a variety of formats.

When running a text through Scania Checker, responses from the checker are highlighted by different colours: red, yellow and green.

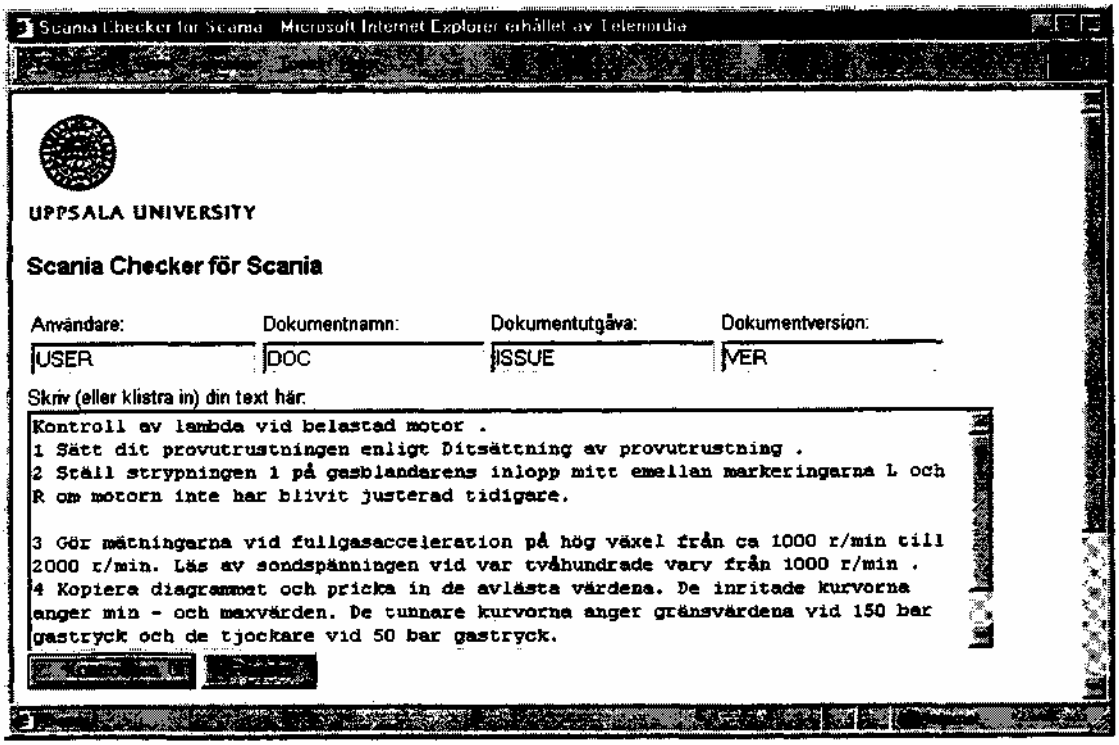

Figure 2. Opening page of Scania Checker with an example text.

Red marking
Lexical errors, where a replacement is offered, are marked red, e.g. if any of the words *fotkontakt, golvkontakt, fotströmställare* or *fotomkopplare* is used, it will be marked red and the replacement *golvströmstättare* will be offered. The rejected words all indicate the same switch situated in the floor, which is controlled by foot, but the Swedish denominations use 'foot' and 'floor' alternatively, combined with three different words meaning 'switch'.

When no appropriate replacement has been found, the author is prompted to rephrase the sentence. This comment occurs, for example, when we know, that a specific term is known to be difficult to translate, e.g. 'chansbyta', which means that a mechanic replaces a part by another, although he is not sure this will help. This verb clearly needs a clarifying sentence in order to be correctly translated.

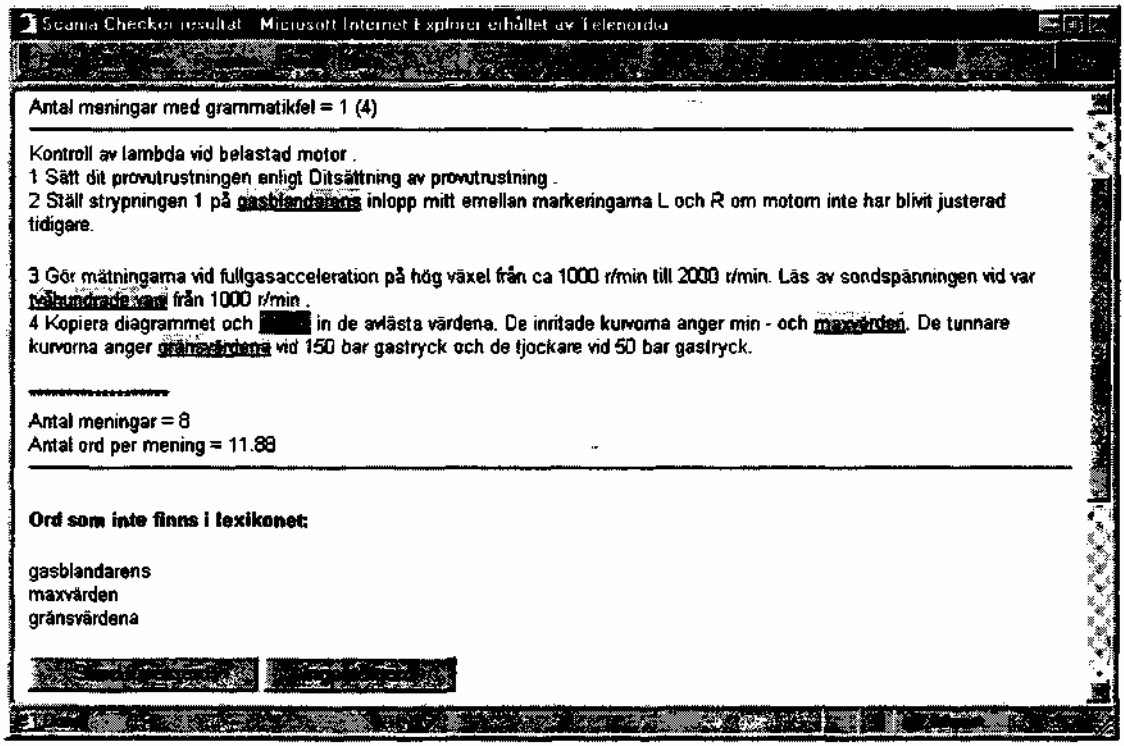

Figure 3. The result of the checking.

Yellow marking
Segments not recognised by the checker are marked yellow. It may be a truly new word, i.e. a word that is used for the first time as a denomination, but also a missing inflectional form of a lemma which is already present in the database. Alphanumeric expressions, tags etc. are also unknown to the checker, unless they have been inserted into the database.

If the writer just wants to test-run his document for errors, and he doesn't want the yellow-marked words to be logged, he presses the burton No action (Ingen åtgärd) at the end of the checked document protocol. Otherwise, he presses Send for action (Sänd för åtgärd) and the words are logged in the file for the database administrator to deal with.

Green marking
Grammatical errors are marked green, e.g. if the adjective predicative is in the wrong number or in the wrong gender. (The coloured fragment is here delimited by #.):

#Fjäderbromscylindrarna är kombinerad# med membrandel för färdbromsdelen.
'The spring brake chambers are combined[2] with a diaphragm part for the service brake part'.

Gäller inte om #fordonet är utrustad# med EBS.
'Does not apply, if the vehicle is equipped[3] with EBS.'

[2] In English the participle is not marked for number as in Swedish.
[3] In English the participle is not marked for gender as in Swedish.

Scania Checker Lexical Database

The database holds some 40.000 segments (word types and phrases) and approx. 13,000 lemmas. Raw materials for the controlled vocabulary were filtered out from two corpora, the initial one (1995) comprising some 220,000 word, the second one (1998) more than 1,6 million words. As can be seen from Table l, the total number of general dictionary segments amounts to 12,791, constituting 30.2 % of the total vocabulary in terms of word types and phrases. General dictionary lemmas constitute 28.9 % out of the total amount of lemmas.

	General	Specific	Total
one-word units	11,411	28,824	40,235
multi-word units	1,380 '	759	2,139
Total	*12,791*	*29,583*	42,374
(lemmas)[4]	(3,803)	(9,355)	(13,158)

Table l. Domains and word units

The lexical database is categorised according to wordform, stem, inflection, lemma, word class (POS), morpho-syntactic code, approved (N) or rejected style (R), replacement, lexical domain and information product (domain). BTI refers to the location of the part/term in the vehicle, but this category is not used at present.

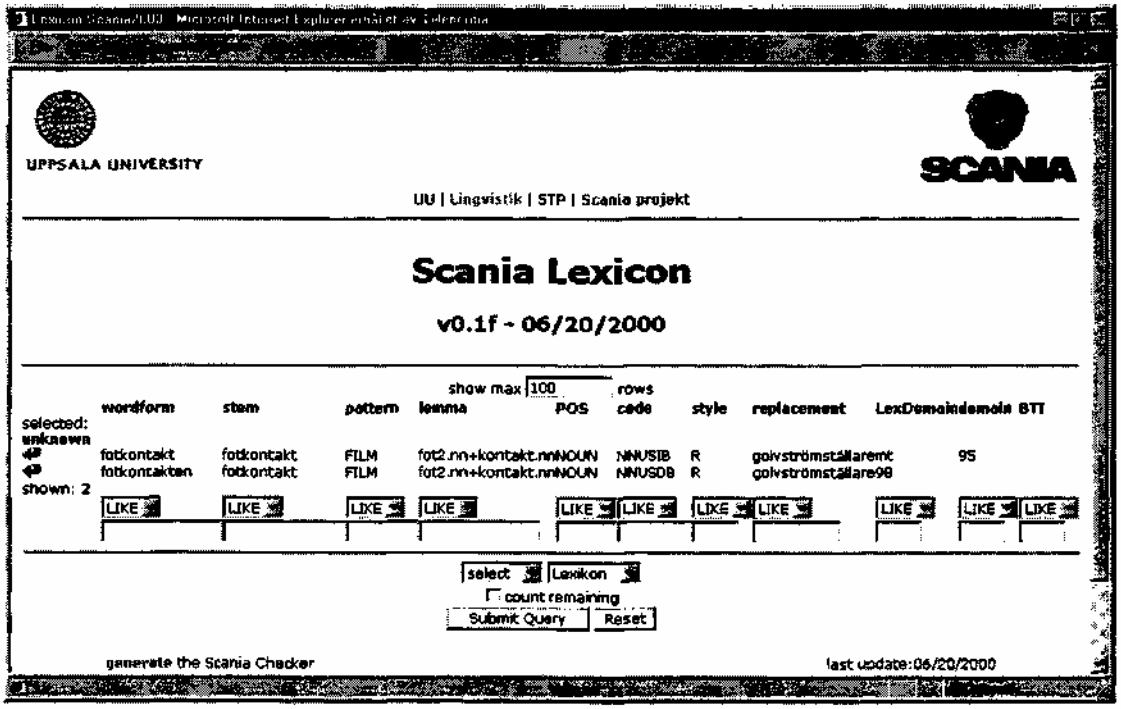

Figure 4. Lexicon categories in Scania Checker Database.

A segment, which is to be inserted into the database must be classified regarding to lemma, syntactic code, style and replacement in order for the checker to be able to respond. The remaining fields in the database are not compulsory, but they provide

[4] Multi-word units are not included.

information that makes the database very useful from a linguistic point of view, and makes it possible to develop it in different directions. Besides from looking up words, users may, for example, want to check what adjectives are used in the Scania literature, what words are rejected etc. Furthermore, it is also possible for writers to make a comment on a specific word in the database.

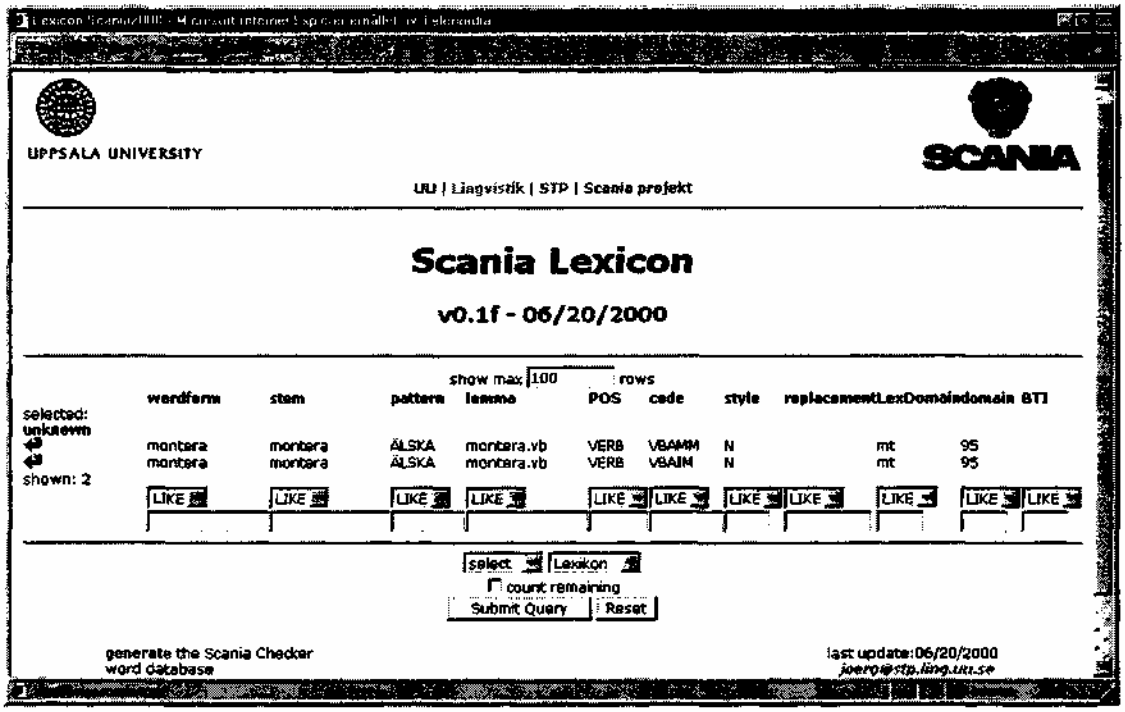

Figure 5. Asking for the comment field in the Scania Checker Database.

By clicking on the pen symbol to the left of the word 'montera' (mount, assemble), the user gets access to a comment field in the database. The fields 'wordform' and 'date' are automatically filled in, whereas the writer must fill in 'author' and 'comment'.

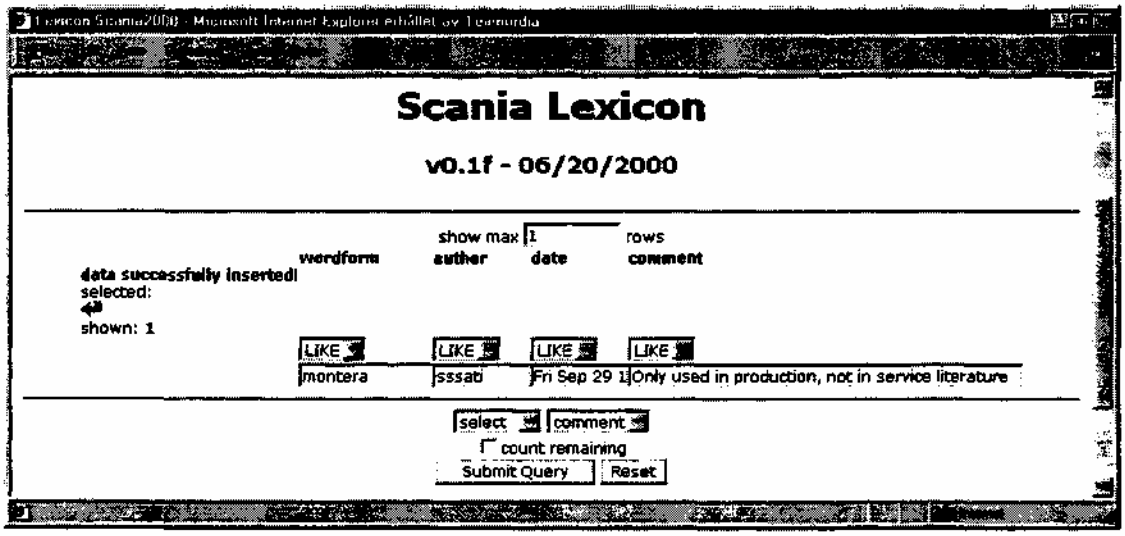

Figure 6. The comment field in the lexicon database.

This comment is then accessible for the database administrator, who, also can filter out comments made by a specific person or on a specific date or point of time, in order to make necessary modifications of the vocabulary.

Considerations when defining the controlled vocabulary

When defining the Scania controlled vocabulary, the goal has been a standard Swedish, with few restrictions in the general dictionary, following general recommendations regarding writing conventions and grammar. In selecting a preferred verb or noun inflection, the preferred variant according to the Swedish normative lexicon, Svenska Akademiens Ordlista, has been chosen.

Regarding the domain specific vocabulary, important goals have been to:

limit the use of synonyms
- limit the use of words that do not translate easily
- limit the use of words that are difficult to understand
- achieve consistency between compounds or derivations of the same stem

At present, 2830 segments have been rejected, most of them have been identified as synonyms, and given a replacement in the database. In 88 instances the comment 'Rephrase' is given, which means there is no clear replacement candidate on a word level.

Words that are difficult to translate correctly are for example colloquialisms like 'mek' (mech) instead of 'mekaniker' (mechanic). Some words are more or less culture-specific, which is the case with 'landsväg', which in the Nordic countries implies a road of high quality, where it's possible to drive quite fast without changing gear. Most other European languages tend to translate it by a word meaning a country-side road, a B-road.

Compounds in technical literature tend to get very long in Swedish, and therefore they are often shortened. In that process some information is usually lost, e.g. 'fullvarvsreglering' (control of full rear) actually means 'reglering av varvtal för full gas' (control of engine speed at maximum acceleration).

It has also been important to assure, that, for example, compounds and derivations are consistently formed, e.g. '*kall*startsfunktion' *(cold* start function), *kall*startsmängd' *(cold* start amount) but '*köld*startstreglage' *(cold* start control) were found in the material, which brought about a change of '*köld*startreglage' to *kall*startsreglage'.

Besides from general language conventions, translatability and understandability reasons, frequency ratings have been considered, when choosing between two or several candidates. If no other argument was valid, the most frequent term in the Scania literature was chosen.

Error types covered by the grammar checker

Primarily, the grammar checker should account for a correct language according to Swedish standards. It should also impose additional constraints specific to the controlled Scania language. The error types have been filtered out from a corpus of roughly 150 documents corresponding to some 450.000 words. The same corpus was used as a training corpus. The starting point for filtering out deviations from standard Swedish was the grammar checker of the SCARRIE prototype for Swedish and the error types that it targets (Sagvall Hein 1999a). The resulting set of error types represents a subset of the SCARRIE error types. Scania specific constraints were proposed by the language consultant at the company. It is assumed that their number will increase substantially when the machine translation aspects are focused.

Split compounds
Swedish compounds differ from e.g. English compounds in that they are written without any intervening space. A frequent error type in the Scania documentation is the splitting of the compounds by the inclusion of spaces and the deletion of hyphens. Four different cases are distinguished, primarily motivated by the specific style of the Scania documentation.

Violation of agreement in the NP
Swedish NPs are complex in comparison with English. In particular, there are strict rules for the order between the attributive determiners and the agreement relations between them and the head noun. Two cases are distinguished: violation of number agreement and violation of gender agreement.

Wrong species in the NP
Species, as opposed to gender and number, does not follow the same agreement rules as do gender and number. For instance, there are definite determiners that agree with the adjective but take a head noun in the indefinite form.

Violation of agreement between the subject and the predicate
Two cases are distinguished, i.e. violation of number agreement and violation of gender agreement. The picture is complicated by the fact that Swedish may have inverted word order at sentence level.

Errors in the VP
Two main types are considered, i.e. wrong form of the main verb after modal verbs and after temporal auxiliary verbs, and missing infinitive marker after certain verbs.

Errors after preposition
Two types of errors after preposition are considered, i.e. basic case of a pronoun instead of the oblique case, and missing infinitive marker before a verb.

Redundant word
This error type applies to expressions that are found to be unnecessarily verbose, i.e. *ett styck uttag* 'one piece of socket' instead of *ett uttag* 'one socket'. The noun is superfluous in this specific context but appropriate in others. This is a problem that cannot be handled by the word checker.

Wrong word
For instance, in the Scania documentation there is some confusion as regards the use of kl*amma* 'clam'. It is used both as a noun and as a verb. The recommendation is to use the synonym *klamma* for the noun and *klamma* for the verb. The checker has to distinguish between the two uses and give an alarm for kl*amma* as a noun. This is a problem that cannot be handled by the word checker.

Grammar error frequencies

Error type	**No of hits**	**Correct diagnosis**	**Incorr diagnosis**	**Proportion of hits**
S-P agreement numb	1	1	-	0,26%
S-P agreement gender	4	4	-	1,06%
NP agreement numb	23	23	-	6,07%
NP agreement gender	66	63	3	17,41 %
NP wrong species	23	23	-	6,07 %
Split comp /wrong case	8	8	-	2,11 %
Wrong pronoun case after prep	1	1	-	0,26 %
Wrong verb form after modal	60	60	-	15.83 %
Wrong verb from after temp aux	1	1	-	0,26%
Missing infinitive mark after verb	12	12	-	3,17 %
Missing infinitive mark after prep	2	2	-	0,53 %
Superfluous infinitive mark after verb	1	1	-	0,26 %
Split comp /wrong spec after genitive	3	3	-	0,79 %
Split comp, def, hyphen missing	21	21	-	5,54%
Split comp, indef, hyphen missing	146	146	-	38,52 %
Wrong word	3	3		0,79 %
Superfluous word	2	2	-	0,53 %
	377	374	3	

Table 2. Error type frequencies in the training corpus.

As can be seen from Table 2, the dominating error type is split compounds (four different sub types). After that come NP agreement errors and errors in the VP.

By correct diagnosis we understand correct in relation to the grammar specification; the diagnosis may, however, be one of two or more possible interpretations of the error segment and differ from the intention of the writer. For instance, *#med andra hand#* 'with other hand' may be diagnosed as a number error with the correction *med andra hander* 'with other hands'. It may also be diagnosed as a species error with the correction *med andra handen* 'with the other hand'. The species error diagnosis should be preferred in this context. However, when choosing between a number error diagnosis and a species error diagnosis, the checker selects the number error diagnosis. The choice is based on data from the SCARRIE project directed towards

newspaper text. For the Scania documentation, it seems it should be the other way round.

An interesting incorrect diagnosis (appearing three times) was observed in the training corpus. The text segment *någon #annan tecken# inställning* 'any #other character# setting' is diagnosed as a gender agreement error. This is in accordance with the grammar specification for the isolated segment *#annan tecken#, annan* being in the uter gender and *tecken* in the neuter. However, this diagnosis is, in fact, a consequence of an erroneously split compound. The proper formulation would be *teckeninställning*, a compound in the uter gender. The technical writers are informed about cases of this kind and recommended to use their language intuition when an error is signaled but the diagnosis seems inappropriate (Almqvist & Sågvall 2000).

The grammar checker is based on partial parsing implying that we have to count on false alarms as well as missing hits. It was trained in four rounds until it gave no false alarms on the training corpus of 450,000 words. In other words, for the training corpus the precision is close to 100 %, taking into account the incorrect diagnosis. Systematic validation of the checker on a proper validation corpus remains to be done. It should include *recall* as well, i.e. the missing-hit aspect. For this purpose reference data (a gold standard) have to be manually created.
The evaluation will be based on prior reference data as was the case for the evaluation of word alignment in the parallel corpus project PLUG (Sågvall Hein 2000, Ahrenberg et al. 2000).

Technical aspects

Scania Checker applies to plain text. It has a simple architecture. Basically, it consists of a word checker and a grammar checker.

The word checker

The word checker includes

- a tokeniser
- a sentence splitter
- a dictionary search function
- a code assembler

The tokeniser is of a traditional kind.

The sentence splitter is somewhat adapted to the Scania documentation text type:

1. It treats a line as a sentence, if it is shorter than 40 characters.
2. It treats the tab sign to be a sentence delimiter.
3. It treats three spaces and more as a sentence delimiter.

The dictionary search function has the following interesting features.

1. It handles the recognition of multi-word units that are stored in the dictionary; dictionary here is to be understood as the lexical database mentioned above, compiled into an efficient search format.
2. In the plain text there is no encoding difference between a hyphen denoting end-of-line syllabification and a hyphen that is part of a compound. Both cases are accounted for in the dictionary search and may be illustrated by the following example:

Detta WABCO ABS/TC-	'This WABCO ABS/TC-
system har en styr-	system has a control-
enhet av en ny generation	unit of a new generation'

   ABS/TC-system is a Swedish compound and *styrenhet* is another one. They are both in the dictionary.

3. It recognises formal expressions that are not explicitly in the dictionary but may be recognised as regular expressions. The regular expressions themselves are explicitly stored in the database.

Regular expression	Example	Comment
[0-9] [0-9{3} [0-9]{3}	1253942	Chassi No.
[01][0-9]:[0-9]{2}\-[0-9]{2}	18:02-01	WSM
[01][0-9]\-[0-9][0-9] [01][0-9] [0123][0-9]	18-000526	TI _____
[A-Z]{2,3}[0-9]{3,4}[A-Z]	AM840D	Type of axle

Table 3. Examples of regular expressions and their instances.

4. It assembles information about lemmas and morpho-syntactic codes sentencewise and builds a list structure that is forwarded to the grammar checker. For instance,

   Gäller inte omfordonet är utrustad med EBS. 'Does not apply if the vehicle is equipped with EBS.'

   —>

   (#(gällal.vb VBAPM)#(inte.ab ABX)#(om.sn SNO)#(fordon.nn NNNSDB)#(vara4.vb VBAPC)#(utrusta.vb PCPUSIB)#(med.pp PRN)#(EBS.pm PMXBA)(PUNC)))

   The list structure reflects lexical ambiguities in the dictionary, which is normally the case (even though not in the simple example above). If a sequence of word may be interpreted as a multi-word unit according to the dictionary, such an interpretation is preferred to a compositional one.

The Grammar checker

Grammar checking in Scania Checker is handled by a software ScarCheck (Starback 1999) that was developed in the SCARRIE project. ScarCheck consists of a chart parser (Carlsson 1981; Sagvall Hein 1983,1987) and an error reporting function.

The checker performs partial parsing based on an augmented phrase structure grammar (Sågvall Hein 1999a). It builds a chart from the list structure delivered by the word checker. For an illustration, see Figure 7.

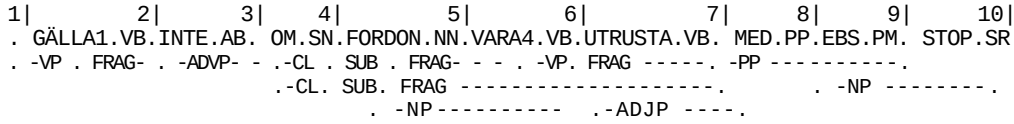

Figure 7. An illustrative chart structure[5]

The graphical representation of the chart in Figure 7 is simplified in two respects. First, the edges are labeled with lemma or phrase category, only. As a matter of fact, the labels are complex feature structures. Secondly, the chart, being an active chart, comprises not only inactive edges (those presented in the figure) but also active edges representing grammar rules and incomplete feature structures. The active edges are not shown in the figure.

An error is reported, i.e. a violation of the Subject-Predicate gender agreement constraint:

```
INTERVALL: 3,7
ERROR: gpagna03:  "Fel genus på adjektivet i
predikatsfyllnaden" 'Wrong gender of the adjective in the
predicative'
```

`gpagna03` is the error type code (see Wedbjer Rambell 1999). The error was found in segment 3 to 7, and in Fig. 8 we show the feature structure of the edge representing this segment.

[5] The chart corresponds to the list structure: (#(gällal.vb VBAPM)#(inte.ab ABX)#(om.sn SNO)#(fordon.nn NNNSDB)#(vara4.vb VBAPC)#(utrusta.vb PCPUSIB)#(med.pp PRN)#(EBS.pm PMXBA)(PUNC)))

```
3-7    Creator: 140
         Features: (* =
                    (START = 3
                    END = 7
                    PHR.CAT = CL.SUB.FRAG
                    SUBJU = OM.SN
                    SUBJ =
                        (START = 4
                        END = 5
                        PHR.CAT = NP
                        GENDER = NEUTR
                        CASE = BASIC
                        DEF = DEF
                        HEAD =
                                (FORM - <* SUBJ DEF>
                                NUMB = SING
                                WORD.CAT = NOUN
                                GENDER = <* SUBJ GENDER>
                                LEM = FORDON.NN)
                                NUMB = <* SUBJ HEAD NUMB>)
                    PRED =
                        (START = 5
                        END = 7
                        INFF = FIN
                        PHR.CAT = VP.FRAG
                        HEAD =
                                (LEM = VARA4.VB
                                WORD.CAT = VERB
                                VERB.TYPE = COP
                                TENSE = PRES
                                DIAT = ACT
                            COMPL =
                                (HEAD =
                                    (PART.TYPE = PAST
                                    WORD.CAT = PART
                                    LEM = UTRUSTA.VB)
                                    PHR.CAT = ADJP)
                                A-FORM = -
                                NUMB - SING
                                GENDER = UTR
                                FORM = INDEF))
                ERR = (1 = GPAGNA03)))
```

Figure 8. An example of a feature structure with an error code; error related features are highlighted for the sake of illustration.

Running Scania Checker and Updating the Database

The immediate and spontaneous reactions from writers, when starting to use Scania Checker was, that it is fast, easy to use, flexible because there are no format constraints, and that the database makes it possible to make refined searches for lexical items. In a normal production flow, the database is updated regularly by the database administrator at Scania using the log file.

The log file displays the document's ID, the writer name or ID, the issue number, the version number and the date when the word form was logged. This data makes it possible for the database administrator to access the context of the word, and, if necessary, to contact the writer to discuss the specifics of a word. By adding 1 to 2 weeks[6], which is the time allocated for other departments to comment the document, to the log date, the database administrator knows, when the wordform needs to be available in the database for a re-run of the document.

At the time of writing, however, it is not possible to evaluate how well the use of Scania Checker can be integrated into the work-flow of documentation and translation, because it has only been used since August. A critical resourcing point will occur, when Scania Checker is used at a maximum during a production peak. Then, writers' time-schedules for getting a source text ready at the agreed dead-line set for translation will be even more strained, and the time for the database administrator to update the database before a re-run of the document may be decreased.

Possible Development Directions

A possible direction for further development would be to develop a Scania Checker for English and implement it at Scania worldwide. This would ensure consistency and quality of technical documentation produced, taking into consideration that technical information is partly written in English today at Scania by people, whose native language is not English. Controlled English would produce a convenient platform for further translation to other target languages.

Another direction is to develop a full-fledged machine translation system based on the platform provided by Scania Checker. As a result of previous research at the Department of Linguistics a machine translation prototype, Multra (Multilingual Support for Translation and Writing) was developed (Sågvall Hein 1994). Multra is directed towards the automotive industry, and test materials for Multra were provided by Scania. Multra translates from Swedish to English or German. It is a modular transfer-based machine translation system, and the parsing machinery used by Scania Checker is entirely based on the Multra parser. A main task in the further development of the Multra prototype is the extension of the translation dictionary. Lexical data for the translation dictionary may be extracted from the multilingual corpus material that have been delivered by Scania and compiled at the Department of Linguistics. The primary tool for this purpose will be the PLUG Word Aligner, PWA (Sågvall Hein 2000, Tiedemann 1999). A pilot project aiming at a specification of the resources that would be needed for developing Multra into a commercial product has at the time of writing just been approved by NUTEK (Swedish National Board for Industrial and Technical Development), see further Sågvall Hein (2000b).

[6] The time varies depending on the type of information product.

References

Ahrenberg, L., Merkel, M., Sågvall Hein, A., Tiedemann, J., 2000. Evaluation of Word Alignment Systems. In *Proceedings of Second International Conference on Language Resources and Evaluation, LREC 2000, Athens, Greece, 31 May - 2 June 2000.*

Almqvist, I. & Sågvall Hein, A. 1996. "Defining ScaniaSwedish - A Controlled Language for Truck Maintenance." In: *Proceedings of the First International Workshop on Controlled Language Applications.* Centre for Computational Linguistics. Katholieke Universiteit Leuven.

Almqvist, I. & Sågvall Hein, A.2000. "Scania Checker. Användarmanual för skribenter." 'Scania Checker. User Manual for Technical Writers.' Scania, Södertälje.

Carlsson, M. "Uppsala Chart Parser 2: System documentation." Technical Report UCDL-R-81-1. Uppsala University, Center for Computational Linguistics.

Olsson, L.-J. 2000. Systemdokumentation för Scania Checker och Scania Lex. 'System documentation of Scania Checker and Scania Lex'., Department of Linguistics, Uppsala University, Uppsala.

Sågvall Hein, A. 1983. "A Parser for Swedish. Status report for sve.ucp." Technical Report UCDL-R-81-1. Uppsala University, Center for Computational Linguistics.

Sågvall Hein, A. 1987. 'Parsing by means of Uppsala Chart Processor (UCP). In L. Bolc (ed.) *Natural Language Parsing Systems, pp. 203-266.* Springer, Berlin, Heidelberg.

Sågvall Hein, A. 1994. "Preferences and Linguistic Choices in the Multra Machine Translation System." In R. Eklund (ed.) *Proceedings of the '9:e Nordiska Datalingvistikdagarna*'. Stockholm.

Sågvall Hein, A. 1997. "Language Control and Machine Translation." In: *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation.* St. John's College, Santa Fe, New Mexico.

Sågvall Hein, A. 1998. "A Chart-Based Framework for Grammar Checking. Initial Studies." In *Nodalida '98. Proceedings of the 11th Nordic Conference on Computational Linguistics.* Copenhagen, Denmark, pp.68-80.

Sågvall Hein, A. 1999a. "A grammar checking module for Swedish." In A. Sågvall Hein (ed.) *Chart-based Grammar Checking. Working Papers in Computational Linguistics & Language Engineering 12.* Uppsala University, Department of Linguistics. ISSN-1401-923X.

Såagvall Hein, A. 1999b. "The PLUG-project: Parallel Corpora in Linköping, Uppsala, Göteborg. Aims and achievements." In A. Sågvall Hein (ed.) *Working Papers in Computational Linguistics & Language Engineering 16.* Uppsala University, Department of Linguistics. ISSN-1401-923X.

Sågvall, A. 1999c. "A Grammar Checking Module for Swedish." In A. Sågvall Hein (ed.) *Chart-based Grammar Checking. Working Papers in Computational Linguistics & Language Engineering 12.* Uppsala University, Department of Linguistics. ISSN-1401-923X.

Sågvall Hein, A. 2000a. "PLUG. Parallel Corpora in Linköping, Uppsala, Göteborg. 1998-04-01 - 2000-03-31. Final Report. Uppsala University, Department of Linguistics.

Sågvall Hein, A. 2000b. "Projektplan för ett regelbaserat maskinöversättningssystem för svenska. Förstudie 2000." 'Project Plan for a Rule-based Machine Translation System for Swedish. Pilot Study 2000'. Uppsala University, Department of Linguistics.

Sågvall Hein, A., Almqvist, I. & Starbäck, P. 1997. "Scania Swedish - A Basis for Multilingual Machine Translation." In: *Translating and the Computer 19. Papers from the Aslib conference held on 13 & 14 November 1997.* London.

Starbäck, P. 1999. "ScarCheck - A Software for Word and Grammar Checking". In A. Sågvall Hein (ed.) *Chart-based Grammar Checking. Working Papers in Computational Linguistics & Language Engineering 12.* Uppsala University, Department of Linguistics. ISSN-1401-923X.

Tiedemann, J. 1998. "Compiling the Scania 1998 Corpus." Uppsala University. Department of Linguistics.

Tiedemann, J., 1999. "Word Alignment Step by Step". In *Proceedings of the 12th Nordic Conference on Computational Linguistics*, 1999, Technical University of Trondheim. Department of Linguistics.

Wedbjer Rambell, O. 1999. "Error Typology for Automatic Proof-reading." In A. Sågvall Hein (ed.) *Working Papers in Computational Linguistics & Language Engineering 4.* Uppsala University, Department of Linguistics. ISSN-1401-923X.