

Scaling Up an MT Prototype for Industrial Use - Databases and Data Flow

Anna Sgvall Hein*, Eva Forsbom†, Jrg Tiedemann†,
Per Weijnitz†, Ingrid Almqvist†, Leif-Jran Olsson†, Sten Thaning†

Uppsala University
Department of Linguistics
Box 527, SE-751 20 Uppsala, Sweden
†{evafo, joerg, perweij, ljo, sten}@stp.ling.uu.se
‡ingrid.almqvist@scania.com
*anna@ling.uu.se

Abstract

In a cooperative project between Uppsala University, the bus and truck manufacturing company Scania CV AB, and the translation company Explicon AB, issues of scaling up the transfer-based machine translation prototype MULTRA for industrial use is being investigated. The project is limited to one domain, automotive service literature, and one translation direction, Swedish to English, but issues concerning the change of domain, translation direction and language pair are also considered. Three focal points of the project work have been the design and implementation of the new MATS system, including the redesign, porting and integration of MULTRA, the redesign and implementation of the dictionaries of the language modules as a lexical database, and the scaling up of the dictionaries and the grammars. The system is currently trained on a corpus of aligned bitexts from the automotive service domain. The coverage of the lexical data is almost complete, and validated by professional translators, but the grammars are still limited. Despite the incomplete state of the grammars, the system already translates more than a third of the segments in the corpus. Preliminary evaluations of system performance and coverage have been made, and further development of evaluation methods and metrics are in progress.

1. Introduction

The MATS project¹ is a cooperative project between the Department of Linguistics at Uppsala University, the translation and documentation company Explicon AB (earlier Translator Teknikinformation AB), and Scania CV AB. It aims at the development of a high-quality machine translation system for industrial use based on the research prototype MULTRA (Sgvall Hein, 1997). Demo versions of MULTRA translates from Swedish to English and German.

The project is a follow-up of a pilot study, in which the steps to be taken from research prototype to industrial system were investigated. Issues concerning the reverse translation direction, and the change of domain were also considered. For funding reasons, however, the goals of the main project had to be restricted to one translation direction (Swedish to English), and one domain (automotive service literature).

The three focal points of the project work have been the design and implementation of the MATS system, including the redesign, porting and integration of MULTRA, the design and implementation of a lexical database for multilingual data, the redesign and implementation of the dictionaries of the language module as a lexical database, and the scaling up of the dictionaries and grammars of the language module.

The presentation of the project work is centred round these issues, starting by a short description of the original version of MULTRA and the changes that were called for by the MATS system and the scaling up effort.

2. Background

MULTRA is short for Multilingual Support for Translation and Writing. It is strictly modular transfer-based sys-

tem. In addition to the classical components of a transfer-based system for analysis, transfer, and generation, MULTRA comprises a preference module that is responsible for ordering competing analysis structures in a preferred order (Sgvall Hein, 1994).

MULTRA was primarily intended for high-quality translation from Swedish to English and German within limited domains. Prior to the MATS project, MULTRA has mainly been used for research and teaching purposes.

Analysis is carried out by means of a chart parser, Uppsala Chart Processor, UCP (Sgvall Hein, 1983). It handles dictionary search, morphological analysis, and syntactic analysis in a uniform manner. Grammars and dictionaries in UCP are formulated in a procedural formalism (Ahrenberg, 1984; Sgvall Hein, 1984b; Dahllf, 1989).

The transfer and generation components of MULTRA are based on unification. Generation, in addition, includes concatenation and morphological processing. Transfer and generation rules are expressed in PATR-like formalisms that were specifically developed for MULTRA (Beskow, 1993; Beskow, 1997a; Beskow, 1997b). Lexical and grammatical transfer rules are formulated in the same formalism, facilitating the transfer of lexical units in context. The transfer rules are partially ordered; a more specific rule has precedence over a less specific one, i.e. a lexical transfer rule taking a context larger than the word itself into account will have precedence over a lexical rule out of context. The lexical translation rules are formulated in the MULTRA transfer formalism. The transfer and generation components of MULTRA are written in Prolog. The original version of MULTRA runs on an AIX Unix server.

3. System Enhancement and Design

The MATS platform is a machine translation core into which the new version of MULTRA has been integrated.

¹<http://stp.ling.uu.se/mats>

Processing in the MATS system proceeds in a number of distinct steps from an SGML version of the source document to an SGML version of the target document via MULTRA. Even though the system was primarily made to process SGML documents, the modular design of the system allows for other front-ends and back-ends.

The platform itself does not offer a user interface although a reference interface has been implemented for testing and demonstration purposes². Most parts of the system existed as separate components before the project started, others were created from scratch. See further (Weijnitz, 2002).

3.1. Architecture

A starting point in the design was the modular nature of the MULTRA machine translation system. From the strict modularity of MATS follows that every step in the translation is handled by a stand-alone program. The modularity makes the system compositional, and it is possible to test parts of the pipe as easy as to test the complete system. It is easily adapted to new tasks, simply by removing, reordering or adding new modules. Examples of such modifications are replacing the parser or transfer module, or replacing the front-end and back-end in order to process other document formats.

The modules are sequentially connected using a unidirectional data pipe. The output of one module is the input of the next. The traffic is multiplexed, meaning that many different data channels may be transported in a shared pipe. This means that even though the traffic flows through the modules in sequential order, it is possible for modules to communicate privately with any module downstream without interference from the intermediate modules. All data communication is text based for transparency and traceability. It is possible to see exactly what happens to data as it is run through the pipe, before and after each module. A protocol specifies how a module communicates with other modules. It is quite simple and not bound to a certain programming language.

3.1.1. MULTRA Revisited

In the MATS project a light C-version of the parser (Weijnitz, 1999) has been substituted for the original Lisp version. The light version is limited to syntactic processing, and so is the new generation process. Thus, morphological analysis and generation have to be handled outside MULTRA. MULTRA has been linked to two monolingual lexical databases for the source and target language, respectively. The lexical databases deliver full form lexical representations in terms of wordforms, lemmas, and linguistic codes. Further, the lexical units of the two databases have been linked to each other via translation relations. In this way, an intermediary translation database has been created, see further 3.2. below. For each lexeme, there is only one translation relation in the database, a default translation. Default translations are retrieved from the database in connection with the dictionary search in the analysis phase. Alternative translations take the context into account are expressed in the MULTRA transfer formalism. This struc-

turing of the lexical data implies a major simplification and speeding-up of the system.

The two monolingual databases have a uniform structure to facilitate a future change of translation direction. See further 3.2.

The database format implies a restructuring of the lexical material as compared to the original version of MULTRA. It also implies a re-design and re-implementation of MULTRA itself. In particular, the input to the system will be a list of lemmas and linguistic codes that are retrieved from the database. The codes will be expanded to feature structures before parsing proper starts. The lemmas carry along information about their associated lexemes (distinct senses) in the source and the target language, respectively.

The work to adapt MULTRA to MATS included porting it from AIX to GNU/Linux. The old GUI was outdated and removed, and the code revised to fit the latest version of the Prolog system.

3.1.2. The MATS Modules

Each step in the translation process is handled by one module. Deciding on what should be handled by how many modules was primarily a question of looking at what software was available for reuse. Some modules have simple tasks and others have complex tasks. The tasks of transfer and generation are handled by one program, MULTRA, as it was already constructed that way. However, MULTRA is itself modular, offering similar kind of traceability as the other modules.

- extracting the text segments from the SGML-format and passing them
- on segment-wise, tokenising the segments, retrieving the lemmas
- and morpho-syntactic codes of the tokens from the source data base,
- creating a list structure of lemmas and codes, parsing the list
- structure and generating a syntactic feature structure,
- transferring the feature structure, generating a string of lemmas
- and morpho-syntactic features, substituting codes for the features,
- retrieving the words from the target database, generating a string
- of words, finalising the string with capital letter and a proper
- placement of signs of punctuation, and transforming the SGML output
- with XSLT for presentation.

Further processing is possible by attaching optional modules. An SGML rendering module has been developed

²<http://stp.ling.uu.se/~perwej/mats>

to present the documents in a graphical environment, in order to facilitate a comparison between the source document and the translated target document(Olsson, 2002).

There is also a module for computing the recall of the translation, see 5.2.

3.2. MatsLex - a Multilingual Lexical Database

The MatsLex database is designed to provide a flexible and coherent environment for storing and managing multilingual lexical data, and for linking them bi-lingually. The internal structure of the lexicon is based on a relational database model. The database can be queried and updated via transparent database 'views' in web-based interfaces. MatsLex is the central store of all the lexical data available for the translation process, and from this "run-time lexica" such as bilingual link lexica are compiled. For consistency, modifications are allowed in the central database only whereas runtime lexica are strictly read-only. It facilitates porting, updating and maintenance of the dictionaries, and the future extension of the system for translation from English to Swedish. Prior to the running of the system, the dictionary is compiled.

The Database Structure The lexical database comprises a set of entities with morphological, syntactic, and semantic information with appropriate relations between them. The relational structure of a monolingual part of the database is shown in figure 1.

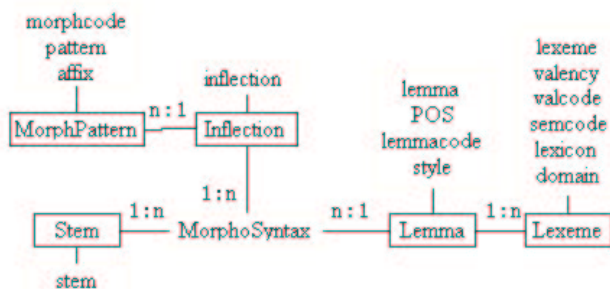


Figure 1: Monolingual database structure.

Morphosyntactic and semantic information is commonly expressed by feature structures. In the MATS database, compressed, compositional codes are used as short-cuts for feature structures. Codes are defined for expressing morphosyntactic features (morphcode), lemma-specific features (lemmacode), semantic features (semcode), and valency related features (valcode).

Surface wordforms are not included explicitly in the database. MatsLex keeps inflectional patterns instead and surface words are derived from these patterns and their technical stems. The crucial point of this approach is to define accurate paradigms and to correctly link lexical entries to appropriate paradigms. Generalised patterns may not be suitable for all languages but in the worst (but most unlikely) case each entry would have its own paradigm. The advantage of this approach is to make updating the database easier. All the possible surface forms are included implicitly when a lemma enters the database. The morphological paradigms in MatsLex are labeled by representative words

and their inflectional patterns are defined in table 'Morph-Pattern' by regular-expressions. The 'pattern' field specifies a regular expression to be matched against the technical stem and the 'affix' field holds the modification to be made in the creation of the wordform. In many cases (in Swedish and English) this simply means concatenating appropriate suffixes with the technical stem.

Another distinctive feature of the database is the possibility to use regular expressions as technical stems that match classes of similar tokens with the same morphosyntactic and semantic features. Constructions with a general pattern are, e.g., dates, time-expressions, and numbers. Some examples are given in table 1.

stem	examples
<code>([0-9]+),?([0-9]*)\(\%\)</code>	50,5% ; 99%
<code>([0-9]*1):a</code>	1:a; 261:a
<code>([0-9]+):e</code>	9:e; 764:e
<code>([0-9]{2})\(/([0-9]{2})\)/([0-9]{2})</code>	01/03/04

Table 1: Token classes defined by regular expressions.

The MatsLex database stores each table from the monolingual lexicon with a language prefix. Hereby, data for additional languages can be added easily to the database. The advantage of keeping several languages in parallel in one central database is the possibility to link them together. To accomplish this, MatsLex allows the establishment of bilingual links between lexemes from different languages. The structure of such links within the relational framework is shown in figure 2.

4. Scaling Up the Language Resources

4.1. Defining the domain

The coverage of the lexical data should be complete with regard to the automotive service domain. The domain was set by a corpus of service literature documents provided by Scania. This corpus, the MATS corpus, comprises Swedish source documents with English and German translations. The corpus has been split into text segments, i.e. sentences, head lines and other kinds of word sequences or words with an independent status in the text³; further, translation links have been established between the text segments language pair wise. The MATS corpus has been divided into two sub corpora, one for training and one for evaluation. The Swedish training corpus comprises roughly 40,000 current words and the evaluation corpus approximately 10,000 words.

4.2. Dictionaries

4.2.1. Augmenting the Swedish dictionary

The starting point for the Swedish dictionary was Scania Swedish, a vocabulary of some 24,000 lemmas extracted from previous corpora (Almqvist and Sagvall Hein, 1996; Almqvist and Sagvall Hein, 2000).⁴ It was developed as a basic resource for Scania Checker, a language

³The segmentation of the text into text segments is based on the SGML mark up of the text. Each text segment has its own id number.

⁴See also <http://stp.ling.uu.se/~corpora/scania/>

checker that was built in a previous co-operation project between Scania and Uppsala University (Almqvist and Sgvall Hein, 2000). The vocabulary of the checker was filtered from a corpus of some 1,8 million words (Tiedemann, 1999). Initially, only wordforms that had actually occurred in the corpus were included. However, the running of the checker demonstrated that there was a need to fill in the wordform gaps. This was done by means of a wordform generation program that was developed for the purpose (Karlsson and Thaning, 2001). The Scania vocabulary includes approved words and non-approved words with replacements. It is assumed that the translation system should operate on text that has been run through the checker, and thus the non-approved words were deleted from the Swedish dictionary. As a result, the number of lemmas in the dictionary decreased by a couple of thousand words to 20,883. In the original version of MULTRA the number of Swedish lemmas was limited to 369. As regards the number of lexical units the dictionary should be complete. What is missing, however, is information about valency frames and semantic features, that will be needed in the analysis phase. Two studies were devoted to these tasks. They are completed and reported in (Thaning, 2001) and (Hellstrm, 2001), respectively. The material that was generated in the two studies will be included in the lexical database. As regards the valency issue, see further 4.3.1.

4.2.2. The Swedish-English Translation Dictionary

The translation dictionary for MULTRA has been scaled up from 59 lexemes in the prototype to 7,043 lexemes in the new lexical database (Forsbom, 2002b). The translation equivalents for a subset of the units in the Swedish dictionary were previously automatically extracted from the Scania corpus by means of a statistically based word aligner, the Uppsala Word Aligner (Tiedemann, 1999). This raw material was also lemmatised (heuristically, in the case of English), looked up in existing domain dictionaries of varying reliability, and preliminary evaluated, so that obvious errors were removed (Lfving, 2001).

Validation of Translation Dictionary During the MATS project, 8,216 of the remaining 14,128 entries (10,510 Swedish lemmas) were manually reviewed by professional translators, using translation memories of the Scania corpus for reference in tricky cases. In the material, one Swedish lemma could be linked to more than one suggested English equivalent, but in the translation dictionary, there should be only one translation relation, a default translation, for each Swedish entry. Alternative translations were to be expressed as contextual rules in the MULTRA transfer formalism.

For each entry the translators should either accept or reject the translation. If they rejected a translation, they were to replace it with an acceptable one. If there were two or more translations for a Swedish entry, they were supposed to choose one translation as the default translation, based on frequency or dictionary information. (Rejected and changed alternatives can later be extracted and entered as not acceptable alternatives in an English checker, in order to reduce unnecessary variation and increase consistency.) For some entries with alternative translations,

however, more than one candidate were acceptable, but in different contexts. The most general or frequent alternative was chosen as the default, and contexts were given for the others, as they were to be entered as contextual rules in the grammar. The result is shown in Table 2.

Step	Entries
For validation	8,216
Defaults back	7,053
Changed defaults	1,196
Alternative equiv.	43

Table 2: Validation of entries

Rejections and changes were made based on, *inter alia*, the following principles: normalisation, writing conventions, controlled language candidacy, corrections of translation and linking errors, generalisations, and specifications. See examples in Table 3.

Approximately 300 of the new default equivalents had been changed to another part of speech than the Swedish lemma, or to a phrasal expression, etc., and have consequently not been added to the dictionary, but will be handled by contextual rules instead. The rest have been assigned a lexeme number each. As we only have checked default translations for the direction Swedish to English yet, some English words are linked to more than one Swedish lexeme. By checking the other direction too, we can find candidates for controlled language on the Swedish side also. For example, *adjustable pliers* is linked to both *polygrip* and *polygriptång*, both of which refer to the same tool. For some of the alternative translations, the same side effect could be seen. When we looked in the MATS corpus at the contexts for the Swedish word *drev* with the default translation *gear wheel* and the contextual translation *pinion*, we found no occurrences of the relation *drev-gear wheel*, 7 occurrences of *drev-pinion*, and 73 occurrences of the new relation *pinjong-pinion*.

Entries not sent to the translators included simple copies, entries marked as general domain, and translation shifts, such as different parts of speech (*certifierad* (adj.) vs. *certification* (noun)), different verb frames (*han heter John* vs. *he is called John*⁵), and phrasal translations. Most copies are taken care of by regular expressions (see 3.2.), but some need to be added. Translations shifts and phrasal translations are (or will be) handled by contextual rules.

4.2.3. The English Generation Dictionary

The generation dictionary for MULTRA has been scaled up from 184 lexemes in the prototype to 7,280 lexemes in the new lexical database. The new entries are derived from the English part of the validated translation dictionary. By changing the translation direction, we have run the English target documents of the MATS corpus through the first three modules of the system to see how well the new dictionary covers them, and got a coverage of 89% of 63,870 tokens.

⁵In Swedish, the verb *heter* has no passive voice.

Type	Swedish	English, original	English, changed
Normalisation	ändskoning lyftbälg systemkrasch böja lång TC-kommunikation	end ferrules air bellow system crashing bent longer TC Communication	end ferrule air bellows system crash bend long TC communication
Conventions	centreringsverktyg AC-förångare vänster	centring tool AC evaporator left hand	centering tool A/C evaporator left-hand
Controlled language	splitknapp skjutmått mätarenhet bromsvibration snöslunga splinesmedbringare	split button slide caliper gauge assy brake vibration snow canon splined driver	splitter button sliding caliper gauge assembly brake judder snowblower splined companion flange
Corrections	kolstoff axellutning måttband radialdäck	sulphur axle inclination tape cold radial	carbon kingpin inclination measuring tape radial tyre
Generalisation	momentstyrning kyla mellanrum	retarder torque control cold weather certain clearance	torque control cold (noun) interval
Specification	lägesjustering	positional adjustment programme	positional adjustment

Table 3: Types of changes made by translators during validation.

In order to move the English generation dictionary into the new lexical database (see 3.2.), we had to define a set of English-specific morphosyntactic codes and inflectional paradigms, analogous to those in the Swedish database (Forsbom, 2002a). The most frequent inflectional paradigm, for example, is labelled DOG and is mapped to four morphosyntactic codes representing four wordforms (*dog*, *dog's*, *dogs*, and *dogs'*).

As one project aim was to facilitate a reversal of the translation direction, the codes were supposed to serve both analysis and generation needs. When used in the analysis module, a code is expanded to a feature structure containing all morphosyntactic information needed in the analysis (cf. Figure 2). In the generation module, the procedure is the reverse: a feature structure corresponding to a wordform is conflated into a corresponding code, and the actual wordform retrieved from the dictionary. This two-fold purpose sometimes means balancing between under- and overspecification of features with regard to the two modules, as not all information which is used for analysis is used for generation, and vice versa.

$$\left[\begin{array}{l} \text{lem} : [\text{sym} : \text{dog.nn}] \\ \text{word.cat} : \text{NOUN} \\ \text{number} : \text{sing} \\ \text{case} : \text{basic} \end{array} \right]$$

Figure 2: FS for the code NNSB and lemma *dog*.nn.

4.3. Grammars

There are three grammars in the system: a Swedish grammar, a translation grammar, and an English grammar. The scaling up of the grammars is in progress. The corpus-based training of the grammars has, so far, been limited to a mini training text of 53 text segments of varying complexity. It is part of the MATS corpus and constitutes the first part of a fairly large document that by Scania was considered to be representative of the text type.

The MATS platform provides an excellent basis for training the grammars in tandem. Typically, the training corpus is processed via the MATS platform, and the first text segment that is not fully parsed (marked by green) in the output is manually analysed; in particular, the chart of the segment is investigated, and the problems are taken care of. When an appropriate parse is generated, the text is processed again, and the translation is generated or there are shortcomings in the transfer or generation phases, and, if so, the results are inspected and the grammars are accordingly developed.

4.3.1. The Swedish grammar

The Swedish grammar covers declarative clauses, imperative clauses, infinitive clauses, subjunctive clauses, adjective phrases, adverbial phrases, nominal phrases and nominal groups, prepositional phrases, quantifier phrases, and participial verb phrases. By nominal groups we refer to indefinite nominal expressions that typically appear as headings, list elements, tables cells and similar text segments. Verb phrases are generally analysed at the clause level only, and not as independent units. In the course of

the training, the variety of clause and phrase types that are covered by the grammar has increased, and the rules have been fine-tuned to the contexts in the training text.

Outside the current scope of the grammar are coordinated clauses, some types of infinitive clauses, some types of subjunctive clauses, and some types of valency rules. The extension of the set of valency rules has been in focus of the scaling up effort in the project. The scaling up of the set of valency rules is based on the complete training corpus. Work on this issue is still in progress.

Verb valency Verb valency, i.e. the potential of the verb to participate in constructions with other members of the clause is a primary issue in scaling up the grammar. The analysis of a clause is based on access to a valency rule that is associated with the finite verb. If there is no such rule, the analysis of the clause will fail. In other words, completeness with regard to valency rules is decisive for the coverage of the grammar.

In MULTRA, valency rules are formulated in the procedural UCP formalism and part of the grammar. They are referred to as verb action rules, VA-rules. VA-rules are defined in terms of grammatical relations, e.g.

```
defrule va.plundra {
  (act,  obj.dir /
   pass, (agent // continue)) }
```

The rule in the example defines the standard transitive rule. It is named by a typical representative *plundra* [pillage]. The rule prescribes that in the active form (diathesis), the verb takes an obligatory direct object, while in the passive form, it takes an optional agent.

In the original version of MULTRA, the association between a verb lexeme and its VA-rule is defined in the UCP lexeme base. In the industrial version, it will be provided by the lexical database (see 3.2.).

Verb valency appears to be domain dependent. Thus rather than consulting a standard Swedish dictionary, we decided to analyse the verbs in the MATS corpus with regard to valency relations. A total of 680 lemmas (including particle verbs), and 689 lexemes (distinctive senses) were identified (Thaning, 2001).

Prior to the analysis of the valency relations, each context was (manually) rewritten into a basic form, a declarative main clause in the active form. The basic forms were investigated with regard to complements and adjuncts. A syntactic valency model was developed. According to this model, the valency of a verb is expressed by means of a syntactic frame and a set of semantic codes. The syntactic frame accounts for the complements of the verb, and the semantic codes for the adjuncts. For instance, NP _ NP is the syntactic frame for a standard transitive verb. Semantic codes were defined for adjuncts expressing *means*, *degree*, *measure*, *location*, *manner*, *time*. A total of 105 different frames were defined. More than a third of the verbs, however, were analysed as standard transitive verbs, 64 were analysed as intransitive verbs, and 28 as transitive or intransitive (e.g. *bromsa* [brake]). The remaining verbs are distributed over 102 different frames. 68 frames have only one representative each.

The syntactic valency frames along with the semantic codes will be stored in the lexical database (see 3.2.). Meanwhile, links between the verb lexemes and the VA-rules are defined in the grammar. Currently, 354 such links have been established, implying that links are missing for 335 verb lexemes. Test runs with the current version of the grammar show that more than 25% of the parsing failures in the training corpus are due to valency problems.

Before the syntactic valency frames and the semantic codes can be used by the parser, they need to be reformulated in the UCP formalism, and grammatical relations corresponding to the phrase categories have to be established. This work is in progress. As a matter of fact, a total of 113 VA-rules have been defined including the 33 rules in the original version of the grammar. However, the original rules are more fine-grained than the new rules. In particular, they take into account not only complements but also adjuncts. In other words, two valency models have to be adjusted to each other. The old model will be adjusted to the new one, the primary reason being that the new model will be more transparent and facilitate future work on the lexical database.

4.3.2. The transfer grammar

The transfer grammar of the original version of MULTRA comprises 87 rules, versus 181 in the new version. The rules are of 6 basic types, see table 4.

Type	Old	New
Copy a structure	24	54
Delete a feature	4	19
Transfer a structure, no change	57	90
Transfer a structure, with a change	6	10
Transfer a structure, Implying lexical disambiguation	1	9

Table 4: Rules in the transfer grammar.

A structure preserving transfer rule may include a shift of the value of an individual feature as well as the deletion of a source feature not required by the target language.

4.3.3. The generation grammar

The generation grammar has been scaled up from 100 rules to 166. The rules give a direct characterisation of the syntactic surface structure. So far, there is no provision for the inclusion of valency relations, or for the expression of optional or repeated constituents. Thus it is to be foreseen, that in the further development of the grammars, in tandem, the number of rules in the generation grammar will increase faster than in the analysis grammar.

5. Preliminary Evaluation

A preliminary evaluation of the system includes system performance and recall (number of translated segments). Future evaluations will also be made regarding the quality of the translations, and the maintainability of the system (see 7.).

5.1. System Performance

The system performance is measured in CPU seconds, counting both system calls and each program's processing time. Total performance for the whole MATS corpus is 26 segments/second, or 137 words/second, on a AMD Duron 750MHz processor. Transfer and generation take most of the time, and will probably take even longer as more and more segments pass the analysis module (see Table 5).

Module	CPU secs
SGML extr.	2.67
Tokeniser	1.74
Lex. lookup	32.79
Parser	89.73
Transf. & gen.	210.92
Code comp.	9.17
Lex. lookup	4.97
Finish	8.92

Table 5: Performance evaluation (CPU seconds).

5.2. Coverage

Given the modularity of the system, it is easy to perform a glassbox evaluation of each module's contribution to the total recall, by simply adding an evaluation module at the end of the pipe. Such a module is currently under development, and preliminary results have been collected for the MATS corpus (see Tables 6 and 7).

These results are primarily aimed for developing purposes, as logfiles on missing words and grammar rules can be used as a basis for updating the dictionaries and grammars, but some of them can also be used for comparing the system to other systems.

In an ideal world, there should be no missing words in the Swedish dictionary, since all source documents are supposed to adhere to the controlled language (see 4.2.1.). In reality, however, this is not true, and we have to revise the source documents accordingly by running them through the checker. This has not been done for all documents yet.

The number of words reported missing from the translation dictionary is rather large. However, as this number is measured before the disambiguation process in the analysis module, many missing words actually correspond to rejected alternatives. The ideal place to measure this would be after the analysis, but as not all segments pass this module yet, the current place of measuring yields the best information for development purposes. As it is now, words reported as missing from the English dictionary are mostly disambiguated words missing from the translation dictionary, as the analysis module copies the Swedish word if there is no default translation. The overall recall of the target dictionary could instead be measured by reversing the translation dictionary, as described in 4.2.3..

The evaluation module distinguishes between partial parse and no parses for development purposes. Partial parses are caused by lacking coverage of the grammar, while no parses are caused by processing errors, e.g. by calls to non-existent rules.

Concept evaluated	Training Total	Unique	Eval. Total	Unique
Tokens	44,193	6,842	11,150	2,431
Segments	7,412	4,734	1,773	1,190
Words not in source dict.	190	86	286	166
Words not in transl. dict.	10,363	1,802	2,000	519
Words not in target dict.	818	252	168	64
No target code	77	39	5	5
Partially parsed seg.	2,900	2,302	755	591
Not parsed segments	1,293	1,134	369	329
Not transf. segments	278	39	24	14
Not gen. segments	84	60	31	9
Translated segments	2,857	1,199	594	247
Fully transl. segments	2,748	1,118	586	229

Table 6: Module evaluation.

Segments	Training Total	Unique	Evaluation Total	Unique
Translated	38.5	25.3	33.5	20.8
Fully translated	37.1	23.6	33.1	19.2

Table 7: System evaluation (% recall).

Segments passing the generation module are passed on in their translated form, possibly with markups of copies of Swedish words, English base forms of words with a faulty feature structure or code, or words missing in the English dictionary. In the evaluation module, these are reported as translated segments. Segments passing all modules with no markups are reported as fully translated, and are included in the figure for translated segments.

The segments counted as translated are primarily the short simple ones (see 4.3.1.). This fact is not reflected in the results of the evaluation module yet, but a count of tokens per segment will be added shortly, to measure recall also in the context of translated tokens.

6. Conclusion

The aim of the project has been achieved, and a system for machine translation, the MATS system, has been de-

signed and implemented. The system translates automotive service literature from Swedish to English. Though no formal evaluation of the output quality has been made, manual inspection shows that the translations are of high quality and on a par with the manual translations used as evaluation standard.

The coverage of the lexical data is complete with regard to the automotive domain. The coverage of the grammars, however, is still limited. In spite of this, the system already translates more than a third of the segments in the corpus.

7. On-going and future work

Continued evaluation of the system will be performed, with a special focus on the output quality and the development of an evaluation methodology. User support for the definition of grammar rules and lexical units for new domains and language pairs will be emphasised. Work on the grammars is in progress. Extending the system with a translation memory represents another line of development. The translation memory should be built from scratch and based solely on translations generated by the system.

Discussions concerning a commercialisation of the system are in progress.

8. Acknowledgements

The research reported herein was supported in part by VINNOVA (Swedish Agency for Innovation Systems), contract no. 341-2001-04917, Scania CV AB, and Explicon AB.

9. References

- Lars Ahrenberg. 1984. De grammatiska beskrivningarna i SVE.UCP [the grammatical descriptions in SVE.UCP]. In Sågvalld Hein (Sågvalld Hein, 1984a), pages 1-13.
- Ingrid Almqvist and Anna Sägvalld Hein. 1996. Defining ScaniaSwedish - a controlled language for truck maintenance. In *Proceedings of the 1st International Workshop on Controlled Language Applications (CLAW'96)*, pages 159-164, Centre for Computational Linguistics, Katholieke Universiteit, Leuven.
- Ingrid Almqvist and Anna Sägvalld Hein. 2000. A language checker of controlled language and its integration in a documentation and translation workflow. In *Proceedings of the Twenty-Second International Conference on Translating and the Computer*, Translating and the Computer 22, Aslib/IMI, London, November 16-17.
- Björn Beskow. 1993. Unification-based transfer in machine translation. Reports from the Department of Linguistics, RUUL #24, Uppsala University.
- Björn Beskow, 1997a. *Generation in MULTRA*. Uppsala University. Department of Linguistics.
- Björn Beskow, 1997b. *Morphology in MULTRA*. Uppsala University. Department of Linguistics.
- Mats Dahllöf. 1989. En lexikonorienterad parser för svenska [A lexicon-oriented parser for Swedish]. Master's thesis, Gothenburg University.
- Eva Forsbom. 2002a. Scaling up the generation lexicon for Multra. Project report.
- Eva Forsbom. 2002b. Scaling up the transfer lexicon for Multra. Project report.
- Jan Hellström. 2001. Semantiska särdrag för substantiv i Multra [semantic features for nouns in Multra]. Project report.
- Stina Karlsson and Sten Thaning. 2001. Automatisk generering av ordformer till Scania-databasen [Automatic generation of wordforms for the Scania database]. Project report.
- Camilla Löfling. 2001. Att skapa ett lemalexikon för manuell och maskinell översättning [To create a lemma lexicon for manual and automatic translation]. Master's thesis, Uppsala University.
- L-J Olsson. 2002. Rendering of MATS SGML documents with XSL transformations. Working report.
- Anna Sägvalld Hein. 1983. A parser for Swedish. status report for SVE.UCP, February. UC DL-R-83-1, Center for Computational Linguistics, Uppsala University.
- Anna Sägvalld Hein, editor. 1984a. *Föredrag vid De nordiska datalingvistikdagarna 1983 [Talks at the Nordic computational linguistics' days 1983]*, UC DL-R-84-1, Center for Computational Linguistics, Uppsala University.
- Anna Sägvalld Hein. 1984b. Regelaktivering i en parser för svenska (SVE.UCP) [Rule-activation in a parser for Swedish (SVE.UCP)]. In Sägvalld Hein 1983 (Sägvalld Hein, 1984a), pages 187-199.
- Anna Sägvalld Hein. 1994. Preferences and linguistic choices in the Multra machine translation system. In Robert Eklund, editor, *Proceedings of '9:e Nordiska Datalingvistikdagarna' (NODALIDA'93)*, pages 267-276, Department of Linguistics. Stockholm University. Stockholm, June 3-5.
- Anna Sägvalld Hein. 1997. Language control and machine translation. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'97)*, St. John's College, Santa Fe, New Mexico, July 23-25.
- Sten Thaning. 2001. Verbvalenser i teknisk text. En fallstudie. [Verb valency in technical texts. A case study]. Master's thesis, Uppsala University. (<http://stp.ling.uu.se/educa/thesis/arch/2001-010.pdf>).
- Jörg Tiedemann. 1999. Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (NODALIDA'99)*, University of Trondheim, Norway.
- Per Weijnitz. 1999. Uppsala chart parser light. System documentation. In Anna Sägvalld Hein, editor, *Chart-Based Grammar Checking in SCARRIE*, number 12 in Working Papers in Computational Linguistics & Language Engineering. Uppsala University.
- Per Weijnitz. 2002. MATS - a platform for machine translation. Project report.